A Nuffield Farming Scholarships Trust

Report

Award sponsored by

NFU Mutual Charitable Trust

Turning data into information: maximising the benefit of digital data technology

Robert Allen

April 2015

NUFFIELD UK

NUFFIELD FARMING SCHOLARSHIPS TRUST (UK)

TRAVEL AWARDS

"Nuffield" travel awards give a unique opportunity to stand back from your day to day occupation and to study a subject of interest to you. Academic qualifications are not essential but you will need to persuade the Selection Committee that you have the qualities to make the best use of an opportunity that is given to only a few – approximately 20 each year.

Awards are open to those who work in farming, growing, forestry, or otherwise in the countryside, and sometimes to those working in ancillary industries, or are in a position to influence those who do. You must be resident in the UK. The normal age range is 25 to 45 but at least one younger candidate each year will receive an Award. You must have spent at least 2 years working in a relevant industry in the UK. Pre- and post-graduate students are not eligible for an Award to support their studies.

The Nuffield Arden Award is unique in that there is no age restriction and the subject is set by the Selection Committee. An Arden Award is offered every 2 years.

Full details of all Awards can be seen on the Trust's website: www.nuffieldscholar.org. Application forms can be downloaded and only online submission is accepted.

Closing date for completed applications is the 31st July each year.

A Nuffield (UK) Farming Scholarships Trust Report

"Leading positive change in agriculture.
Inspiring passion and potential in people."

Date of report: July 2015

Title	Turning data into information: maximising the benefit of digital data technology
Scholar	Robert Allen
Sponsor	NFU Mutual Charitable Trust
Objectives of Study Tour	To investigate how agriculture realises the potential of new data technologies and to understand what barriers may prevent delivery of on-farm benefit
Countries Visited	UK USA Canada Russia Australia
Messages	The potential of data can only be realised through integration with appropriate intellect and skills that allow meaning and insight to be derived. The key data challenge is not collections, but to extract meaning from data. Data are inherently noisy and correlations within data do not necessarily indicate causation. The integrity of data should always be challenged before acceptance of the information it contains. A formal data strategy explicitly defines the relationship between people, process and the technology used to collect, collate and analyse data within a business. This architecture is central to ensuring that procurement of data tools or systems is within scope and fit for purpose.

Contents

DISCLAIMER

CONTACT DETAILS

Dr. Robert Allen
22 Rooks Street
Cottenham
Cambridgeshire
CB24 8RB

Telephone: 07816 791371
Email: robert.allen@greenvale.co.uk

Nuffield Farming Scholars are available to speak to NFU Branches, Agricultural Discussion Groups and similar organisations

Published by The Nuffield Farming Scholarships Trust
Southill Farmhouse, Staple Fitzpaine, Taunton TA3 5SH
Tel : 01460 234012
email : director@nuffieldscholar.org
www.nuffieldscholar.org

Acknowledgements

I wish to thank both the NFU Mutual Charitable Trust and the Nuffield Farming Scholarship Trust for the opportunity my Nuffield Farming Scholarship has provided me. I am very grateful to all the people and organisations who generously gave up their time for me during my study tour; this report is only possible because of the thought provoking conversations I was able to have with them. Finally, I must thank my wife, Hana, for the love, support and encouragement she has shown me during my Scholarship.

Chapter 1 - Personal introduction

During my Nuffield Farming Scholarship year I was appointed as the Research Manager for UK potato packers Greenvale AP. Taking this role has seen me return to working in agriculture for the first time in 15 years. My initial interest in agriculture and agronomy developed as summer help on the experimental potato plots at the Cambridge University Farm, where my father was Farm Director. During this time I was fortunate to undertake a summer's internship at PepsiCo's research farm in Wisconsin USA.

University saw a change of direction with an undergraduate degree in Physical Geography followed by a PhD in Palaeo-glaciology and Palaeo-climatology, both from the University of Bristol. Post university life took me to work as a Data Scientist for Landmark Information Group, the largest providers of geo-spatial and environmental data in the UK.

Whilst at Landmark I had been watching with interest from the sidelines the rise of ag-data. Being awarded a Nuffield Farming Scholarship has allowed me to interview leading lights in the field and become more knowledgeable about this increasingly important sector of the industry. Whilst my route into agriculture may be circuitous I hope that experiences and knowledge picked up along the way can make a positive contribution going forward.

I live near Cambridge with my wife Hana, a medical bioinformatician.

Figure 1.1. The author, Robert Allen

Chapter 2 - Background to study subject and countries visited

High profile acquisitions - such as Monsanto's $1bn purchase of the Climate Corp *[1]* (*see References on page 42*) - plus projections of the global ag-data market size reaching $20bn *[2],* have provided 'oxygen' to the hype surrounding the data sector in recent years and the promises of how it will revolutionise agricultural production *[3,4,5].* I have been fortunate to work for organisations that not only understand the value that can be generated through effective use of data, in both academic and commercial contexts, but also recognise the overheads, investment and commitment required. My Nuffield Farming project was proposed out of a wish to understand my perception of a gulf between some of the ag-data hype and my experiences working with data technologies.

To fully explore the topic of this report it was necessary to visit a mix of farming businesses, established and start-up ag-data companies, academic research groups, lobbyists and non-agricultural data expertise. This allowed insights from across the agricultural spectrum to be heard, as well as gaining knowledge from other sectors experiencing similar data challenges to agriculture. A summary of countries visited is listed in Table 1 below and a full list of all interviewees who have made this report possible is provided in Appendix 1 on page 44 .

The report is my personal interpretation of the current state of the agri-data sector, focused on crop production, within the countries visited during my study period. **The reader should note that data technologies are rapidly evolving and should view this report as a snapshot of the agri-data sector between mid-2014 and mid-2015.**

Date (month/year)	Country	Reason for Choice
January 2014 – October 2014	UK	Home to progressive farming businesses, innovative ag-data companies and world leading data expertise (e.g. Office of National Statistics)
March 2014	Australia	Leading research institutions and farming operations implementing data solutions
June 2014	Canada	Well established and advanced ag-data management companies and expanding precision agriculture sector
July 2014	USA	Start-up and disruptive technology companies being funded by venture capital investment. Farming businesses in transition from limited implementation of data solutions to advanced data systems. Active lobbying community.
September 2014	Russia	The challenging operating environment means data and data systems are an essential operational tool for effective farm and business operation.

Table 2.1 : Countries visited

Chapter 3 - Data and decision making

3.1. What is data and why do we use it?

Today 'data' are commonly used as a generic statement which often leads to misunderstanding and misuse of datasets with respect to what information or insight can, or cannot, be derived. Appreciating the difference between data types is relevant to both technical data analysis and commercial decision making.

> *Raw* data are unprocessed statements and form the foundation of objective information and insight.
>
> Through analysis and addition of human intelligence (either directly or indirectly) raw data are converted into *processed* data that provide insight and information.
>
> A particular type of processed data, relevant to agriculture, is *derived* data which are created through combining elements of different raw datasets together.
>
> An important class of data, often overlooked, is descriptive *metadata* that provides information on the content of datasets; these are vital for effective data management and sharing.

Information and insight are used to improve knowledge and understanding, which in a commercial business can then be used to support or justify decision making. Within this context it is important to appreciate that information or insight can be stratified by time: do the data relate to events in the past, present or future? (*see Table 3.1*). This theoretical framework describes why we collect data and reinforces the previous paragraph that 'data' are a more complex concept than frequently assumed. For example, measuring a particular crop metric can be achieved using different methodologies which should be selected dependent on purpose: a visual estimate of current crop canopy is probably suitable for a statement on "what is happening". In contrast, to predict "what will happen" requires a time series of good quality data describing crop canopy development and structured for repeat access by the user over time.

Arable farming business will create and consume data for legislative requirements, operational reporting (including financial) and agronomic understanding of crop performance. Software solutions covering these categories have existed for a considerable time; however, changes to on-farm practices and interest in agriculture from technologists are forcing change in the ag-data sector for incumbents, and attracting new market entrants. The long term decline of the agricultural labour force and consolidation of farming operations increasingly means decision makers are managing increasing acreages of land [*3*]. Traditional management techniques of frequent field visits and on-site management are becoming unsustainable. Data systems which provide the right data, to the right person, at the right time are important elements of agricultural decision making. The lack of appropriate data is often the root cause of failing to understand failures in agronomic performance and inefficient business processes.

From my interviews a near universal opinion was that the burden of legislative reporting will increase. If correct, meeting this will require data and data architecture to allow farming businesses to report the necessary information at appropriate timescales. Underpinning improved on-farm

application of data is the requirement to increase agricultural production to meet predicted population growth in an environment of diminishing resources, i.e. sustainable intensification [*2*]. Quantifying and measuring progress in the efficiency of agricultural production can only be achieved through the collection and analysis of appropriate agronomic data.

	Past	Present	Future
Information	What happened? *(reporting)*	What is happening? *(alerts)*	What will happen? (*extrapolation)*
Insight	How/why did it happen? (*modelling*)	What's the next best action? (*recommendation*)	What's the best or worst that can happen (*prediction*)

Table 3.1: Six major reasons for data collection. Adapted from [1].

Deciding what are the appropriate data solutions for a farming business is challenging. What data strategy optimises the internal business benefit from the data they collect whilst meeting the external, imposed, information obligations? The growing number of companies providing ag-data services coupled with the wide array of sensor technologies now available for collecting data are making this an increasingly complex decision and are explored in this report.

- Chapter Four provides a summary of the different agri-data markets and profiles example companies visited during my study tour.
- Lobbying organisations and open source projects have the potential to drive the effective use of data in agriculture and are profiled in Chapter Five.
- Chapter Six investigates the concept of 'Big Data' and agriculture.
- Data ownership, explored in Chapter Seven, is central to both a good business data strategy and also creating an industry-wide environment where data can be shared effectively.
- Chapter Eight studies why an enterprise data strategy is relevant for modern farming businesses.
- Future challenges to the ag-data sector are investigated in Chapter Nine.
- A list of interviewees who have made this report possible is provided in Appendix 1.

Chapter 4 – The data technology market place

4.1. Introduction

The ag-data market contains discrete sub-sectors focused on different aspects of agricultural production, which are summarised in this chapter.

4.2. Review of ag-data markets

4.2.1. Large company platforms

Large, global, agricultural businesses - e.g. machinery, fertiliser, seed manufacturers - have used IT and data technologies to support their core business activities for many years [*1,2*]. These systems have the advantage of scale of funding for development and ease of purchase for consumer. They can be provided as an integral part of the machinery, seed or ag-chem purchases of a farm. New data collection and management technologies are now realising the potential of collecting and aggregating data from across the user base of these companies. These data can be used in 'Big Data' analysis and provide services for optimum varietal selection, nutrient management and anonymised benchmarking, to name but a few.

The limitation of this market sector is that products are intrinsically restricted to the scope of the provider rather than the user. For example, a seed provider may offer a variety 'optimisation' service for a farm, based on historical climate and yield data; however, it will only locally optimise the best variety within their portfolio. This may not be the global optimum when evaluating all available varieties. Historically, data systems supplied by machinery manufacturers have been non-compatible and presented a substantial challenge for mixed fleet operations, though this is starting to change.

The financial power of global agribusinesses will ensure that their data products will remain a significant presence in the market and will provide a valuable service to growers who don't wish to evaluate alternative, independent, market options: an understandable position. However, I don't believe that they will reach the levels of market domination predicted by some. The scope of their data products and services are inherently self-limiting as they are designed to support sales of their primary products. This is confounded by currently low levels of data technology adoption which is providing the incentive for independent providers to develop alternative solutions and develop market share, in particular playing on the power of 'local'. Underpinning these trends is the mistrust amongst many growers about the true intentions of global agri-businesses and they remain wary of fully adopting the data services the latter provide. The recent change of strategy amongst some global agri-businesses from completely closed and proprietary to a more open approach allowing integration of third party systems reflects a realisation that, despite their size, a position of isolation is not advantageous.

4.2.2. Farm management software

Farm management software is the most mature ag-data sector. Desktop software products were developed in the 1990s to formalise the collection of field operation, field mapping, precision ag and financial data [*3,4,5*]. Whilst there is a large degree of commonality between the products offered in

this sector they can often be differentiated by their areas of expertise: for example, legislative due diligence, supply chain management, recording farm operations or financial control. Therefore, clearly defined user objectives are a necessity when evaluating products.

The sector is currently undergoing significant change reflecting changes in available technology and commercial requirements. Multi-user access, in-field data collection via mobile devices and seamless data sharing are making cloud based systems a sector standard. For many incumbents the pressing technical challenge is migrating existing desktop architecture to cloud architecture whilst simultaneously managing the wind-down of legacy systems actively used by customers. A clear challenge to the sector is delivering improved platform interoperability. Previous practices of file transfer via USB memory sticks are time consuming and poor data management. The requirement for users to manage data via multiple platforms is viewed as onerous. Technologies, discussed in subsequent chapters, are being developed to improve interoperability, and strategic commercial relationships are being developed to share content *[6].* Start-up businesses adopting the view that agriculture is 'open air manufacturing' are developing new generation software services into this market. Efficient, just-in-time manufacturing, relies on tight control of stock inventory and movement plus detailed knowledge of the true cost of production per unit. Many of the protocols developed in manufacturing are being adopted and adapted for agriculture *[7,8].*

There are interesting variations in the business models being used in this sector: from basic software vending, whereby a user is sold a licence and may receive basic training and support, through to complex franchise networks where the software provider uses a network of franchisees to promote and distribute the brand and product offering. In return the franchisees have access to a data management system to use with their clients. 'Coaches' for the Canadian Agri-Data Solution company are available across the agricultural production life cycle from in-field agronomy to trading on international commodity markets *[9].*

4.2.3. Data collection and management

A current challenge with machinery telematics and as-applied data is that many growers do not operate single manufacturer fleets. Existing proprietary manufacturer systems prevent collation of all their data into a single system. Also, there are growers who do not wish to use manufacturer platforms. An emerging market is exploiting data technologies in mobile data technologies, wireless data transfer and cloud data storage to develop agnostic platforms for collecting on-farm data *[10,11,12].* The basic concept is to extract data from the CANBus diagnostic port and wirelessly upload it to a cloud database. There is divergence in the commercial models being used to commercialise this technology. Farmobile are initially focused on the data collection and management. The hardware is leased to growers for a fee but their aim is to generate the bulk of their revenue from data trading, through commission charged against data sales made by growers. At the time of my interview 640 labs had a greater focus on developing analytical services on top of the data collection service. However, it is likely that their business model may have changed after their acquisition by Climate Corporation *[13].*

4.2.4. Precision agriculture

The principles of precision agriculture (PA) have been extensively investigated in previous Nuffield Farming Reports and won't be repeated here. Readers interested in PA are recommended the following reports *[14,15].* PA has been the most data-intensive sector of the ag-data market. It makes extensive use of remote sensing, machinery sensors and guidance, and soil mapping data sources. The traditional PA business is based on integrating different data sources and then providing analytical agronomic services based on information derived from the raw data, primarily for broadacre combinable crops.

The sector is being changed by emerging technology and development of new markets for historic and current PA data. Data management technologies, e.g. wireless data transfer, have transformed the capacity to collate in-field and as-applied data by bypassing cumbersome manual file-transfer procedures *[10,11,12].* Unmanned Aerial Vehicles (UAVs) have dramatically increased the spatial resolution at which fields can be observed *[16,17].* In Europe, the Sentinel Mission launched by the European Space Agency is now operational offering free real-time satellite data *[18].* This has the potential to radically alter the business model for many PA businesses. Historically, technology limited capacity for aggregating data from across the user base. The increasing ease with which data can now be aggregated, coupled with the large spatial extent and lengthy time series that now reside in the archive of established PA companies, offers the potential to gain new insights and understanding of agronomic performance across large geographical areas. However, there will be considerable cost associated with these analyses and a key for PA is continued delivery of quantifiable benefit on farm. Furthermore, generating data describing true costs of production (i.e. in- and ex-field costs and revenues) requires the continued integration of PA with farm management software. Efficient farming is the product of both precision management of resources in the field and optimising the cost of operations between farm and field.

4.2.5. Big Data companies

This market sector is still in its infancy and is currently going through a spin-up period of development and testing. In essence the technology companies in this space are looking to develop commercial applications from the richness of data now available. The technical aspects of 'Big Data' are discussed in Chapter 6. It should be noted that this sector is very broad, with start-ups developing solutions for micro-finance for growers in Africa, to risk management of global assets *[19].*

The most famous 'Big Data' company in agriculture is Climate Corp, founded as Weatherbill in 2006 by two ex-Google employees. It was acquired by Monsanto for $1bn in 2012 *[20].* The original business was selling event-based insurance to businesses dependent on suitable weather conditions (e.g. ski resorts, outdoor shows, farmers). Historical meteorological data across the US was analysed to calculate the probability of a specific weather event happening at a specific location (e.g. late season frost or extreme night-time heat). Customers could purchase insurance against the weather event and the policy paid out if the weather event occurred within the time-frame of the policy. Continuous feeds of real time weather data are received by the Climate Corp to provide details on current weather events. The Climate Corp have now focused solely on agriculture and are developing their analyses to provide agronomic information to clients.

4.2.6. In-house development

The ag-software described in the previous sections are mass market; they are designed to be used off the shelf by a broad user base. As such there is considerable scope for bespoke development by individual farming businesses. From my interviews the primary driver for bespoke development was a desire to improve specific elements of crop agronomy that were not adequately serviced by commercially available services. For example, Produce World's Soils for Life project was built to develop a company-scale soil information system to map, assess and monitor soil resources and quantify the impact of soil condition on their crop production. The Perry Brothers, based in Southern Alberta, Canada, grow processing potatoes for Frito-Lay and McCains. The surface topography of their irrigated pivots is sufficient to make effective irrigation across the pivot challenging; supplying adequate irrigation to the high areas without over-irrigating the low areas is very difficult. In conjunction with the Alberta Potato Lab they are developing an in-house variable rate irrigation program. This forms part of their 'Data Drive Agriculture' initiative whereby they are integrating data from different sources across their farming operation to improve knowledge of, and decision making on, their crops. The advantage of bespoke development is you can innovate and develop solutions tailored to your specific requirements with trusted partners. The disadvantage is the cost of design; development and maintenance are borne by the grower. Furthermore, the unit costs will tend to be higher as they cannot be distributed across a broad user base.

Chapter 5 – The role of pan-industry organisations

5.1. Introduction

The previous chapter described the range of different ag-data market sectors and profiled example companies operating in these sectors. This chapter profiles the role of industry lobby and special interest groups and their objectives in shaping the uptake and use of ag-data technologies.

5.2. The Ag Gateway

The Ag Gateway is a non-profit business consortia based in Washington DC *[1].* Its mission is to "promote, enable and expand eBusiness in agriculture" and their long term goal is "to become the recognised international source for enabling the use of information and communication technologies". It has a membership of over 200 companies and at the time of writing had active projects in ag-retail, software and service providers, crop nutrition, crop protection, feed, grain, seed and precision agriculture.

To date, much of the Ag Gateway's work has focused on improving e-connectivity between agricultural supply businesses rather than interfacing directly with growers. Their Ag Industry Identification System (AGIIS) directory is a database housing 4.8 million uniquely identified entities (e.g. business locations) and over 91,000 agricultural products (e.g. crop protection chemicals, seed and fertilizer) *[2].* These data can be used as the fundamental building block for efficient electronic interactions between agricultural supply companies.

Industry segment in the Ag Gateway is operated by an independent council which allows for independent prioritisation of projects. For example, the Precision Agriculture group launched the Standardized Precision Ag Data Exchange (SPADE) project in 2012 to "establish a framework of standards to simplify mixed-fleet field operations, regulatory compliance, crop insurance reporting, traceability, sustainability assessment and field or crop-scale revenue management". Much of this work is focused on creating standards which define the structure of the data, thus enabling automated data transfer (data standards are discussed in detail in Section 6.2.4). The Ag Gateway deliberately adopts a pragmatic approach and has a preference to build on standards that have previously been developed and will promote adoption from other industries where relevant standards already exist.

More recently, the Ag Gateway has become more involved in more direct grower issues. In July 2014 they published an open glossary of agricultural terms *[4].* Whilst the contents of the glossary may appear trivial to a farmer it is serving a useful purpose. Currently there is no standard definition of agricultural terms and data categories. Software engineers and designers working in agricultural data companies may often come from outside the industry and a recognised central resource that defines basic agricultural terminology will help minimise confusion. In November 2014 the Ag Gateway's Data Privacy and Security Committee published a white paper on data privacy best practice setting out what should be considered and potentially included in the data privacy policies of data service providers. The document clearly stated it was for educational purposes only, but by generating debate and moving the industry towards a more consistent and transparent set of policies around data privacy and terms of use, it has served a positive purpose.

5.3. Open Ag Data Alliance (OADA)

The Open Ag Data Alliance (OADA) was launched in March 2014 by the Open Ag Technology (OAT) Group, Purdue University, Indiana *[5].* The OADA is a consortia of 18 partners, including the Climate Corporation, Monsanto, Granular and the UK's Ag Space. The starting principle of the project is that **"**each farmer owns data generated or entered by the farmer**,** their employees or by machines performing activities on their farm" *[6]* and their aim is to develop open reference implementations of data storage and transfer mechanisms with security and privacy protocols. Technical details of how the OADA will achieve this are discussed in Section 6.2.4.

Traditional approaches to improving interoperability have tended to focus on fixed closed standards that attempt to enforce uniformity of data. This is achieved by implementing pre-defined structured frameworks but has the weakness of being inherently inflexible and will exclude non-conforming data. Organisations may continue to create non-conforming data for practical or commercial reasons: for example, if they have methods of creating data that are incompatible with the defined standard. While uniform representations of data are a laudable goal, organic growth of standards out of open implementations has proven far more effective in other industries than top-down control. For example, digital images can be stored in many "standard" formats (e.g. jpeg, png, gif) for which developer communities have created standard libraries. The OADA is focused on providing developers with relevant and useful libraries required to communicate and access any format of data between systems without tedious human intervention and repeated reinventions of the wheel.

The OADA is not a commercial organisation and will not have any products to sell nor endorse any technical solutions. Its output will be a set of open source libraries allowing interoperability between hardware and software that create and manage on-farm data. At the time of writing the OADA is still in its infancy which makes it difficult to draw conclusions on its impact on the industry. The adopted approach is logical and if successful will deliver a very useful set of tools to the North American ag-data sector. The challenge will be getting a critical mass of adoption in the industry whereby the OADA libraries become the *de facto* libraries of choice when handling the transfer of ag-data. From a UK perspective it is a project that should be monitored and consideration given to how it could be adopted in the UK.

5.4. American Farm Bureau Federation

The American Farm Bureau Federation (AFBF) is the overarching body of the local and state farm bureau network in the US. During 2014 it has taken an active role in defining policy around data ownership from the grower perspective and also providing workshops and online video tutorials to educate growers on what they should consider when contemplating the use of data in their operations *[6,7].* At this stage, whether or not you agree with the content of the data privacy policy of advice in the video tutorials is a moot point. The fact that the AFBF are engaging with their members on this topic is to be supported, as it should allow North American growers to tackle the issues of data ownership and data strategy, discussed in Chapter 7 and 8 respectively, with more confidence. However, care should be taken in the future to ensure that policies of lobbying groups remain flexible. The current AFBF approach is conservative, but this is not a universal stance amongst all producers.

5.5. Chapter summary

Improving interoperability of ag-data and education of growers in the developing data market will be significant challenges in the next decade. Commercial bias will always be an accusation, rightly or wrongly, levelled at initiatives from commercial companies to educate, or promote initiatives within, the customer base, which tends to limit the effectiveness of these campaigns. Whilst the OADA and Ag Gateway are funded by the ag-data industry, their funding is received from multiple sources and their objectives are pan-industry. Therefore, they have a powerful role to play in ensuring that meaningful progress is made to deliver real benefit to users. From a UK perspective it is encouraging to note the direct involvement of AgSpace in the OADA initiative. Whilst the impact of these initiatives may currently seem distant from the farm I believe they will have a significant, positive, impact in the next decade. It is important for the UK industry to actively cultivate and maintain an international perspective in this area and ensure awareness of, and where necessary access to, leading technologies and ideas regardless of geography.

Chapter 6 – Big Data and agriculture

6.1. Introduction

It is difficult to define when 'Big Data' started; analysts and researchers were publishing research notes around *c.* 2000 which explored the potential implications of modern data technologies (e.g. information-sensing mobile devices, remote sensing, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks) to create very large volumes of data *[6].* During the 2000s new technologies were developed that enabled companies, such as Google, to effectively store, analyse and retrieve very large and ever increasing volumes of data *[7].* Today 'Big Data' has become a buzzword that is used to cover many aspects of data management, in its broadest definition. This chapter will explore what is meant by 'Big Data' and how it might impact on agriculture.

6.2. What is Big Data?

6.2.1 The four 'V's

The most commonly used definition of 'Big Data' is the four 'V's:

- Volume
- Velocity
- Variety
- Veracity

Volume refers to the exponential creation of data (*Figure 4.1 overleaf);* the world's technological, per-capita, capacity to store information has roughly doubled every 40 months since the 1980s. This rate of expansion has been such that 90% of all the data in the world has been created within the last two years *[8].* **Velocity** refers to both the rapid pace of change in digital technologies and the increasing capacity for data systems to consume, process and analyse large volumes of data. Furthermore, user expectations of system performance have also increased. **Variety** refers to the fact that the capacity to create data via different technologies has increased. Efficient and effective integration of data from different sources is a non-trivial task. **Veracity** reflects the inherent uncertainty and noise[1] in data. An International Data Corporation (IDC) report in 2012 estimated that only 0.5% of all the data created is actually analysed. More startlingly still, they estimated that only *c.* 20% of the data could be identified, through appropriate tagging techniques, which is required to make the data analysable *[9].*

The explanation of 'Big Data' described in the previous paragraph is a global definition. Its relevance to a specific industry is predicated on the assumption that the industry in question has the same inherent structure and technical challenges as described by global statistics. In the case of agriculture it is far from clear that this assumption is valid.

[1] Noise in a dataset refers to the errors (random or systematic) that mask the true signal being recorded. These errors can be so large as to render the data meaningless but many data can be 'cleaned' or corrected prior to use. The danger is that users assume data are clean without investigating the impact of error on the recorded signal which can lead to erroneous interpretations.

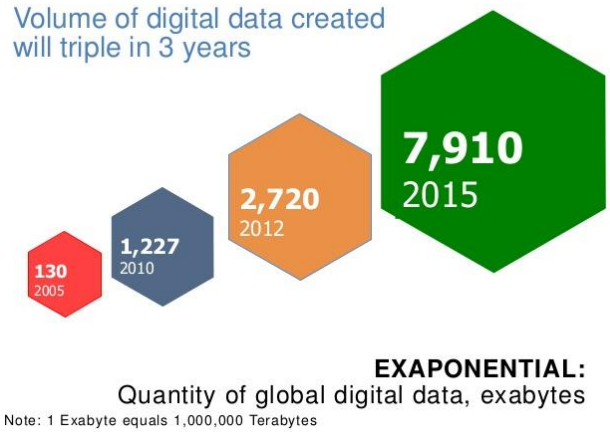

Figure 6.1: estimated volumes of global data, created per day [10]

6.2.2. Data volume and agriculture

There are a few agricultural companies, e.g. John Deere and Monsanto, who are dealing with very large volumes of data; the Climate Corporation data systems are running in excess of 50 terabytes of live data. This is made up of meteorological observations from 2.5 million locations, daily forecasts from major climate models and 150 billion soil observations. These are used to generate 10 trillion weather simulation data points used in the company's weather insurance pricing and risk analysis systems *[11].* However, these companies are exceptions to the norm.

At the farm level, it is my opinion that even the largest operations are not creating data volumes that necessitate new data storage technologies. Also, it should be considered if new 'Big Data' data storage technologies are appropriate for agriculture. Relational databases management systems (RBDMS) remain the most widely used technology for storing data. Initially developed in the 1970s, RBDMS are based around a formal architecture of related tables, each containing rows and columns and traditionally stored on a single server. Each row in a table contains attributes relating to a single entity. The advantage of RBDMS systems is that they are highly structured and very good at maintaining data integrity. However, there is very little flexibility in the structure of the database which must be created *a-priori* of the data being loaded; there are limitations of scale, and maintenance of the service is dependent on the server not failing. Hadoop, created by the Apache Foundation *[12]* is a widely used open source platform for very large volumes of data. It works by splitting data and processing across multiple machines (or servers) using the Hadoop Distributed File System (HFDS). There is then a framework (called MapReduce) used to retrieve data from across the HFDS. This technology allows for effective and efficient storage of very large data volumes, in particular unstructured data; i.e. data that does not have a formally defined relationship. If you conceptualise a data model for a farming business it becomes readily apparent that it would be a highly relational model; i.e. the majority of data and information is related to a single discrete entity: the field. RBDMS technology remains a highly appropriate solution for storing and management of agricultural data.

There is a developing concept of 'data density' (data volume x requirement to access) that better defines the management requirements of a specific data source. Storing large volumes of data that

only require infrequent access is relatively straightforward; data can be stored in a basic data warehouse infrastructure. In contrast, managing a smaller volume of data that requires continuous access is more difficult, owing to the requirement to ensure timely and consistent delivery of service.

6.2.3. Data velocity and agriculture

Since 2000, global stock exchanges have seen consistent growth of High Frequency Trading (HFT); these are trading strategies that use complex data analysis algorithms to exploit the technical capacity to execute trades at extremely short timescales (milli or micro-seconds). In 2012 half of all equity trades on Wall Street were HFT *[13].* This example highlights that there are powerful technologies capable of handling and acting on data at extremely short timescales. However, do they have relevance to agriculture? In my view, it is difficult to envisage an application in crop production where it would be necessary to take actions on these timescales; it is the "taking action" that is the critical component. For example, machinery telematics is one area where rapid polling of data from the device is useful; however actions are taken infrequently, only when the machine reports a problem. Furthermore, the timescale of action discussed here should not be confused with ensuring the timeliness of information which remains a significant issue.

6.2.4. Data variety and agriculture

The range of potential data sources available in modern agriculture: e.g. machinery telematics, remote sensing, field recordings and automatic environmental sensors, has created a genuine data management challenge for modern farming operations. In many cases this issue is exacerbated by the business trying to incorporate new and additional data sources into an incomplete or poorly functioning data management strategy. The following eloquent quote, from Stephen Few, summarises the far more prosaic 'more data than your business can handle' definition of 'Big Data' and has much more relevance at the farm level:

> *"We are overwhelmed by information, not because there is too much, but because we don't know how to tame it. Information lies stagnant in rapidly expanding pools as our ability to make sense of it and communicate it remains inert and largely without notice" [14]*

This definition is independent of the absolute volume of data created and reflects the relative ability of a business to effectively use the data it generates. For agriculture the issue of data variety is further complicated by a reliance on third party data providers and a growing range of different data markets, as described in Chapter 4. Solving the purely technical issue of managing and integrating multiple sources and formats of data must be carried out in parallel with consideration of commercial terms and conditions, data licencing and data ownership. The challenge of data variety is, I believe, the largest 'Big Data' issue in agriculture and should be the focus of the majority of the innovation and development in the ag-data market in the next decade.

In the majority of interviews with technology companies it was clear that they perceived having a direct relationship with the end user, i.e. the grower, as the value space for their business. However, from a consumer perspective this is unsustainable as it inevitably creates increasing numbers of

unconnected platforms. The structure of more mature data markets reflects this position. For example, the environmental and risk data market in UK has a very small number of companies with direct access to the end user. However, the products they offer contain data integrated from many different sources and suppliers. This pyramid structure with business-to-business commercial agreements enables the relevant data from suppliers to reach the end consumer even though the end consumer is only using a single product.

Strategic partnerships and projects launched during 2014 are an indication that potentially the ag-data market is starting to evolve into a market structured around a more pyramidal structure with fewer direct relationships with growers. This structure would also better reflect a universal theme in my visits, that of grower trust and preference for dealing with local suppliers. In 2014 Agri-Data Solution (Section 4.2.2) completed an integration with the MyJohnDeere.com platform which allowed data to be shared between the platforms; thus allowing the end user to access data from both systems from a single system *[15].* At the time of writing Agri-Data solutions are actively working towards providing integration with Trimble and Raven systems.

As described in Section 5.3 the OADA is an open standards software project aiming to create a 'data ecosystem' based on a suite of Application Programming Interfaces (APIs). An API is a protocol defining the structure of a piece of data, thus enabling two separate systems to be able to share data in a consistent manner. The term 'open' here refers to the published data transfer protocols and not the data content which remains owned and controlled by the grower. The ultimate aim of the project is, in their words, to "*remove the overwhelming walled gardens of incomplete data*"; i.e. enable communication and data transfer between data siloes that currently exist. Growers face the "production issue" yet there is no single data provider that covers the entirety of crop production. Figures 6.2 and 6.3 (on next page) are schematics of OADA's interpretation of the current and future data flows in the ag-data industry pre- and post-OADA, respectively. Furthermore, the use of open APIs in this manner will also allow data to be transferred between cloud providers if necessary. A set of open API standards that enabled data service providers to transfer data whilst maintaining user and data integrity will be beneficial to the industry.

6.2.5. Data veracity and agriculture

Data veracity is the second element of 'Big Data' highly relevant to agriculture. Adjunct Professor at New York University Kaiser Fung believes that the majority of challenges in 'Big Data' analytics are not about data volume, which in his opinion can be solved through the technologies of storage and processing capacity, but are related to how and why errors occur in the analytical output *[16].* He uses the OCCAM acronym to describe the main causes of data veracity:

(O) Observational
(C) Control (lack of)
(C) Complete (seemingly)
(A) Adapted
(M) Merged

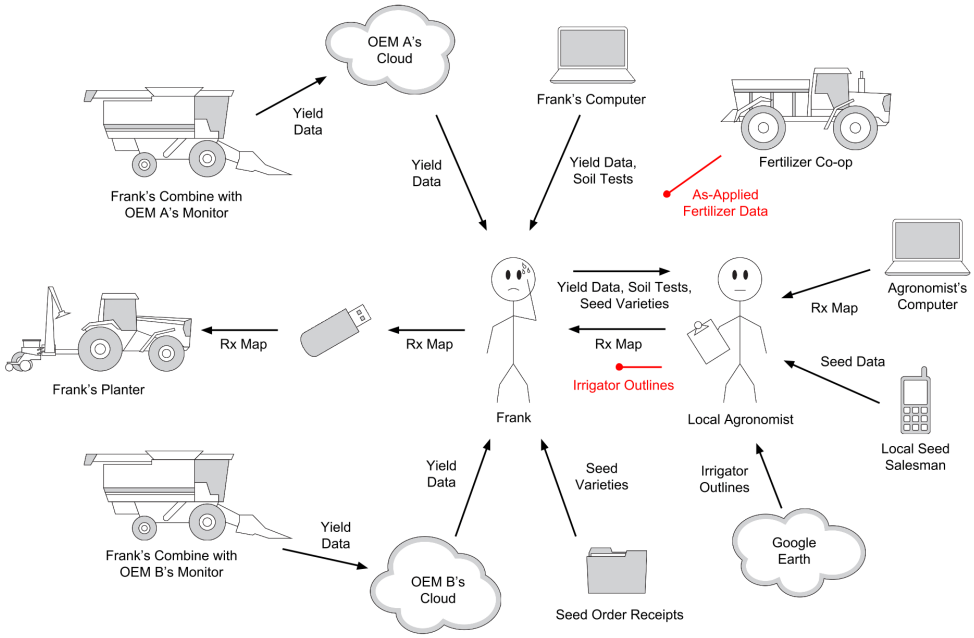

Figure 6.2: Current state of the ag data industry from a farmer's (Frank) point of view. Data in in red represents data that may not make it back to the Farmer. Image source: http://openag.io/principles/.

(If there is difficulty viewing this diagram increase the zoom level: Editor)

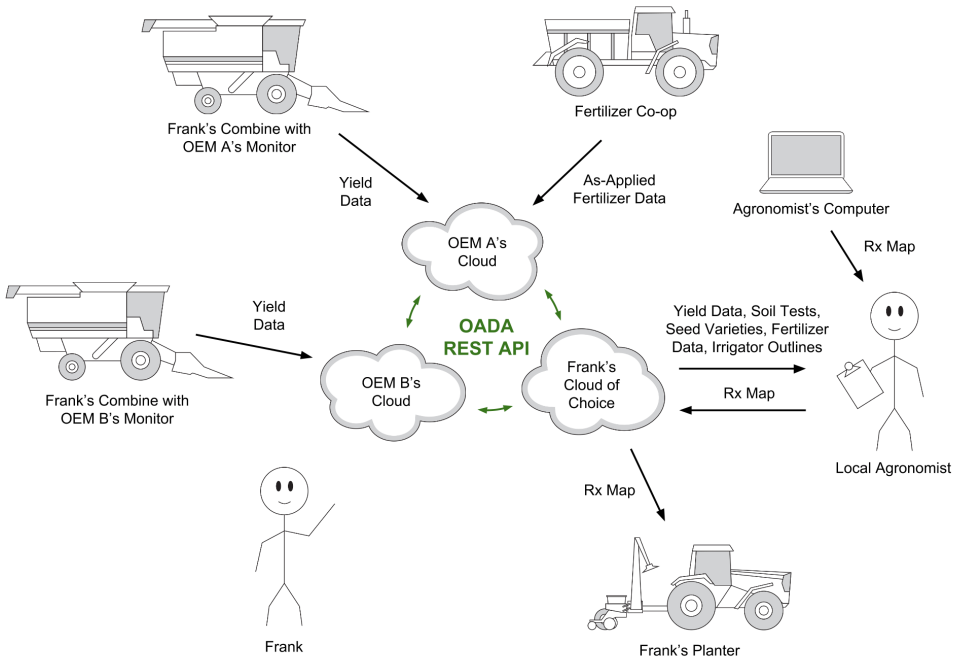

Figure 6.3: Farm data (Frank's) can be shared across multiple cloud storage networks through use of common OADA REST API. OEM = Original Equipment Manufacturer. Image source: http://openag.io/principles/.

Much of the data in agriculture is now generated from sensors or tracking devices. These devices continuously and indiscriminately create data without design: i.e. it is observational data. In contrast, formal experiments have an inherent purposeful design which focuses collection of data relevant to the question being investigated. The power of designed data collection is that it defines boundaries of what information can be imparted from the raw data. In contrast, the analytical boundaries of observational data are much more blurred, making out of scope analyses more likely.

Data control refers to the physical maintenance, calibration, technical validation and supportive metadata of sensors that ensure the data they generate are valid and fit for purpose. Sensors, like other devices, will have a finite lifespan where the recorded data are within defined tolerances. Without maintenance, data from sensors have the capacity to 'drift'; i.e. introduce systematic bias into the results or simply stop collecting data. Methods do exist for correcting systematic bias but are only effective when analysts recognise that the raw data contain a bias. Where the standards of physical and technical maintenance are not adequate cumulative errors will occur and be amplified when merging data from multiple sources. Calibrating yield sensors on harvesters is a classic example of why data control is important. Incorrectly or un-calibrated yield data from a harvester cannot be reliably used in an absolute comparison with yield data from other machines.

As discussed at the start of this chapter, volumes of data creation are unprecedented. In a perfect world this would increase the statistical certainty of analysis as sample sizes are increasing. However for the reasons outlined in the previous two paragraphs raw data will remain inherently noisy and will always contain errors. The scale of these errors will be proportional to the data volume; i.e. more data creates more false leads and blind alleys. This adds significant complication when searching for meaningful and predictable structures in data. The financial and physical costs of data cleaning are overlooked in data analysis projects, but in many cases they are the largest proportion of the project.

Adapting data refers to the increasing re-use and recycling of data and often for purposes unrelated to the original reason why the data have been collected. For example, variations in engine revs from a tractor whilst ploughing can be used as a proxy for soil resistance and used to identify areas of heavier ground across the field. Related to the issue of adapting data is the capacity to merge data from different sources. Errors are multiplicative through data; therefore, integrating errors from different sources will results in even larger error in the combined dataset which will make finding the signal in the data more challenging. Appreciating the limits of the data are crucial to ensuring that any re-using of data is delivering a valid and usable output.

6.3. Big Data analytics

In 2008 Wired, a technology magazine, carried the editorial "The end of theory: The Data Deluge Makes Scientific Method Obsolete" *[17]* which predicted that the rise of 'Big Data' analytics would cause the demise of conventional scientific theory. The basis of this argument is that scientific methods were developed when data collection was limited by technology and analytical and statistical methods were designed to accommodate this. Interpretation was achieved through explicit consideration of causal mechanisms that might explain relationships displayed in sparse data. Current technology is now able to record the whole population which fundamentally changes

the rules of data analysis. By having all the data, future progress will be made through identification of correlations in the data and understanding of causation is no longer a necessity; the data will speak for themselves. For reasons outlined in the previous section and case studies discussed below I believe this argument is flawed.

Data analytics company Dunnhumby provide the analytic power behind the highly successful Tesco Clubcard program. The initial analysis of Clubcard data by Tesco had proved unsuccessful, primarily because they had attempted to analyse all data being created. Dunnhumby adopted traditional statistical techniques to sample the raw data. Even with a sampling rate of 10% of the data their predictive results have an accuracy in excess of 95% [18]. Increasing the volume of data does not mean that 'small data' techniques become obsolete.

A famous example of 'Big Data' hubris is the Google Flu Trends (GFT) model published in the leading scientific journal Nature in 2009 [19]. Google in collaboration with the US Center for Disease Control (CDC) had identified significant correlations between search terms in a sample of 50 million historic internet search engine queries and flu rates recorded by the CDC. These correlations were used to develop a predictive model of flu rates in the US. During 2009 and 2010 the GFT model performed well and outperformed the in-house CDC model *(Figure 6.4).* However, in 2011 and 2013 the GFT model performance declined and eventually over-predicted flu rates in 100 out of 108 weeks *(Figure 6.4) [20].* The root cause of this decline was the use of correlations that had no causal mechanisms at a fundamental level; i.e. the search term in Google was not actually related to flu rates and the correlation was coincidental. Eventually the behaviour behind the search term being used in the model started to change and was not correlating with flu rates. Subsequently the model has undergone repeated iterations of refinement yet remains a substantial point of discussion amongst the 'Big Data' and academic communities *[21].*

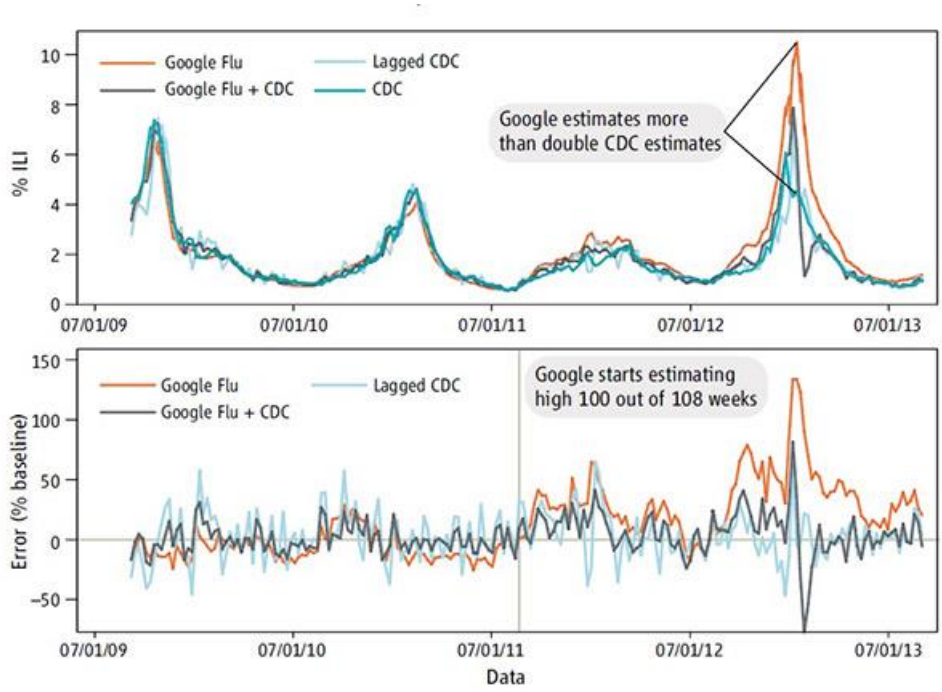

Figure 6.4: Performance of the Google Flu Model and data from the Center for Disease Control. The top panel shows the predicted rates of the population with influenza like illnesses (ILI). The bottom panel shows the error rate of the different models [20].

The unnerving element of this example is that it took two years for the model to break down and has more relevance to agriculture than is immediately obvious. The annual flu cycle being predicted by the GFT is the same timescale as agricultural production and the variable being predicted can be considered a 'yield': i.e. how many people have flu. This 'yield' is an outcome of the interaction between a large number of factors and will exhibit inter-season variation.

One of the potential applications of the GFT model was to use the predictive insight of the model to support the CDC in their operational management of resources for treatment of flu outbreaks in the US. This thought process is very similar to how growers and processors use yield prediction models (e.g. the Cambridge University Farm yield prediction model for potatoes) to manage crops and downstream supply chain logistics.

The issue with the GFT approach is that it can appear to work in the short term and may only break down after several years. In this scenario many users are likely to put the first poor performance down to the 'season' and will only start to challenge model efficacy after the second year of under-performance. In hindsight what seemed like a sensible investment in the technology has taken four years to demonstrate its inherent limitations. At best it delivered no improvement in understanding and knowledge of yield production and, at worst, may have caused incorrect decision making.

The three to four year timescale is relevant in a broader context. Many ag Big-Data start-ups are either in beta-testing or year one of their data collection. As such the volumes of data in their systems are, currently, relatively small and do not contain the element of time required to test the robustness of their algorithms over multiple seasons. The next decade could be a period of fluctuation as inherent weaknesses of analytical services being developed are uncovered.

The contrasting approach is to impose minimum data standards to use a system. This strategy has been adopted by Monsanto for entry into their FieldScripts platform, which requires a minimum of three years' data meeting specified criteria. For a grower making a strategic decision to join the scheme in 2014 this would mean having to collect three years of data, with associated costs, which would only allow entry into FieldScripts in 2017 *[22].* From a data perspective *a priori* definition of required data quality and standards will decrease the chances of "rubbish in, rubbish out" (data quality is discussed further in Section 9.5). But it does require users to make a long term commitment to a system, at their own expense, before realising any benefit.

What is clear is that whichever system users choose there is a common element of time; both approaches require the accumulation of a time series of data before genuine insight can be extracted.

6.4. Data visualisation

The purpose of data visualisation is the clear and efficient communication of information. Whilst this sounds simple it is a specialist data field and a very considerable skill and highly relevant to maximising the utility of ag-data.

Professor Edward Tufte, a pioneer of data visualisation in the 1970s defined the following principles for data visualisation:

"Graphical displays should:

1. *show the data*
2. *induce the viewer to think about the substance*
3. *avoid distorting what the data have to say*
4. *present many numbers in a small space*
5. *make large data sets coherent*
6. *encourage the eye to compare different pieces of data*
7. *reveal the data at several levels of detail*
8. *serve a clear purpose: description, exploration, tabulation or decoration*
9. *be integrated with the statistical and verbal descriptions of a data set." [23]*

Whilst this is an academic point of discussion good data visualisation offers clarification not simplification. The power of this should not be under estimated; it is not dumbing down. The field of data visualisation has been developing rapidly alongside the other aspects of 'Big Data' described in previous sections but appears by many to be undervalued. It is an aspect of data management where the ag-data sector can improve and should actively look to learn from external expertise.

In the UK one organisation that has to tackle the challenge of effectively representing large volumes of data to a non-specialist audience is the Data Visualisation Team at the Office of National Statistics (ONS). The team was founded in 2007 with the aim of improving access to complex data for non-specialist audiences and to separate the data analysis from the visualisation within the ONS. During my visit, it became apparent that a key element to their work is appreciating the importance of the difference between personal perceptions and reality. This is the concept of 'emotional innumeracy': the fact that our own personal perception of statistical certainty is influenced by our own bias, fears, prejudices and anecdotal experiences *[24].* In other words our intuitive statistician is inherently poor.

This has profound implications for agriculture, where data collection is often prioritised over data accessibility or visualisation. In this environment decision makers cannot consume the required information at the right time or in a usable format and by default rely on their personal perceptions and experience. From personal experience, a simple data visualisation of agronomic data describing commercial potato crops, supplied by the grower himself, was sufficient to demonstrate that his perceptions of the crop were substantially different from what had occurred. In this instance the data has been collected but no emphasis had been placed on presenting the data in a consumable format. At a practical level this matters because making in-season agronomic decisions based on a perception, rather than objective data, will often lead to the wrong decision being made and, furthermore, is likely to create confusion as to why the intended outcome did not occur.

In 2012 the Visualisation Team were asked by the UK government's Office for Standards in Education, Children's Services and Skills (OFSTED) to create an interactive tool displaying regional performance at key education milestones *[25].* The different elements considered when creating the visualisation are highly relevant to agricultural data:

- *Variable geographical scales*. A simple solution to the OFSTED request would have been to plot the school performance data against the boundary of each local authority on a map. The weakness is that local authorities are not the same geographical size. Visually results would be dominated by the largest local authorities, leaving users with a skewed interpretation of the data. On a map it would be very difficult for users to assess relative performance between non-contiguous local authorities. The solution was to plot the data by region in a dot or boxplot and a map is provided in the margin to provide the geographical context to the user *(Figure 6.5a and 6.5b).*
- *Displaying key stages*. The dataset included five education stages and the visualisation tool had to display results of all stages. These could be selected via the dropdown box at the top of the page. After selecting the key stage of interest the tool auto-ranks the regions based on the data selected, providing an interactive guide to the relative performance of each region.
- *Incorporating absolute values and relative performance*. It is possible to select an individual local authority from the list in the right hand margin which then highlights its overall position in the data. This simple touch provides a clear indication as to the relative performance of an individual point of interest *(Figure 6.5c).*

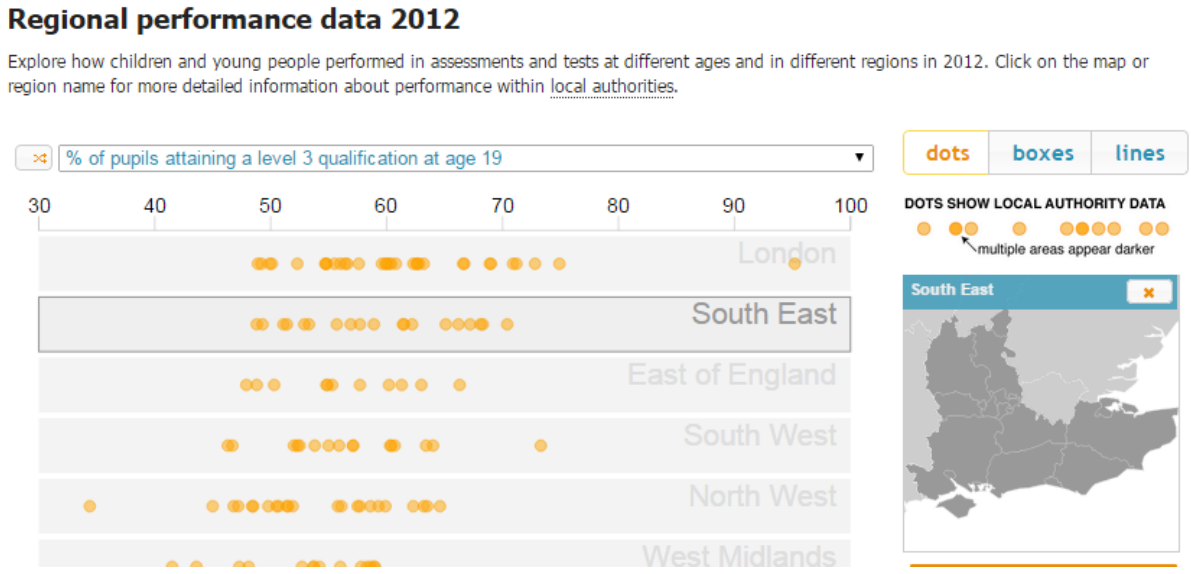

Figure 6.5(a). .

Dot plot (a) of OFSTED local authority educational data displayed by region.
Map in right hand margin provides the geographical context for the data.

The box-plot (b) is shown on next page.

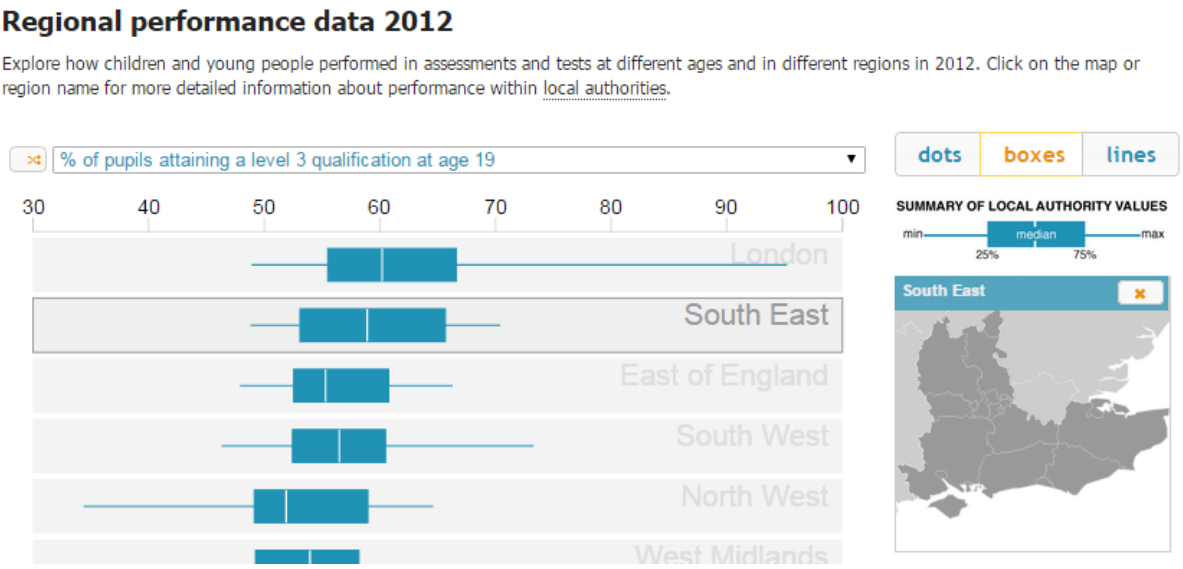

Figure 6.5(b):

Box-plot (b) of OFSTED local authority educational data displayed by region.
Map in right hand margin provides the geographical context for the data.

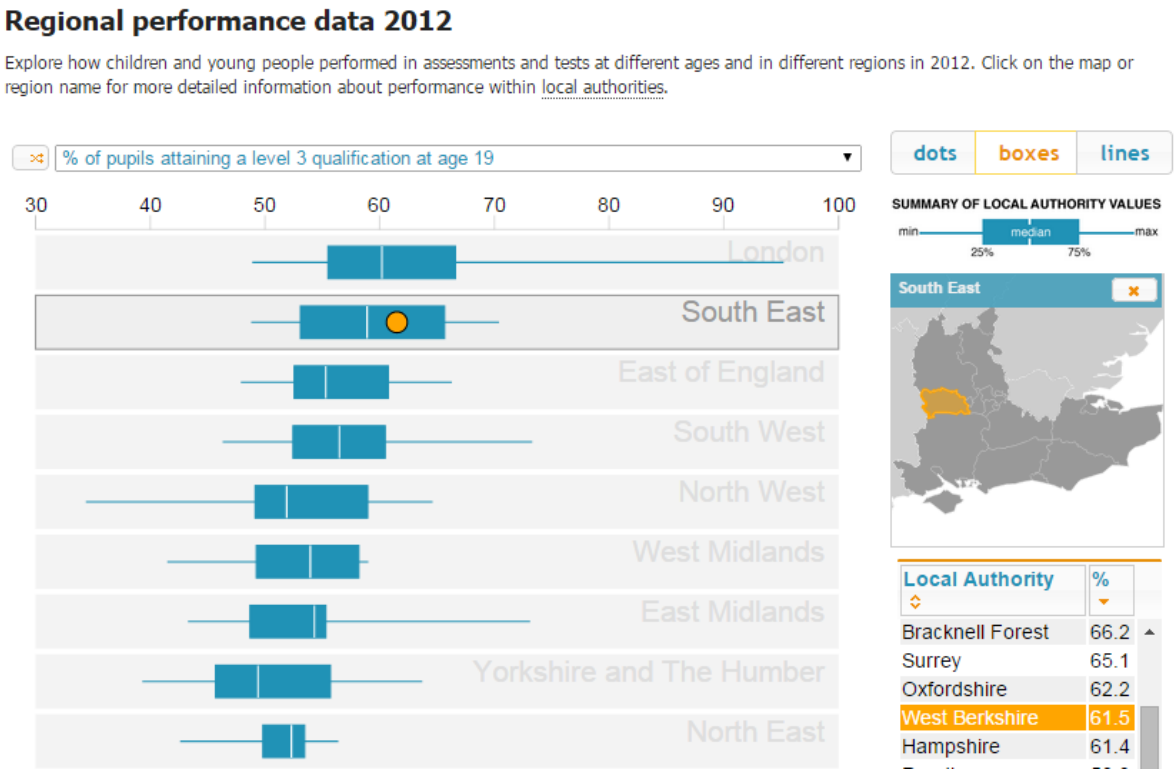

Figure 6.5(c): Relative performance of West Berkshire (orange dot) in the South East of England.

Structuring the data in this manner allows the user to engage directly with the core message of the data; i.e. the educational performance of local authorities at each key stage. Replacing region with farms, local authority with fields, and educational key stages with crop growth stages, a template such as this demonstrates how consideration of data visualisation can provide greater accessibility to agricultural data.

6.5. Chapter summary

This chapter has explored the different elements of 'Big Data'; the four V's, 'Big Data' analytics and data visualisation. Elements of 'Big Data' are often globally defined which do not necessarily translate to a specific industry. From my observations and discussions the future impact of 'Big Data' on agriculture is unclear; the following quote from Peter Thiel summarises the problem:

> *" … the problem is actually finding meaning within the data. It's to make big data small. That's actually the core challenge. It's not collecting more and more data" [26]*

If 'Big Data' is allowed to develop and mature beyond the current hype into concrete ideas focused on finding meaning in data, the increasing availability of technologies to collect, analyse and visualise data is an exciting prospect. Furthermore, it is clear working from first principles is a necessity, and understanding causation is still central to using data. In contrast, belief that progress will be achieved by just collecting more data will, I believe, inevitably lead to a perceived under-performance of 'Big Data' over time.

Chapter 7 – Data ownership and commercial implications

7.1. Introduction

Data ownership is currently a heavily debated and contentious aspect of the ag-data sector *[1,2,3]* yet is central to progress. Without a framework that allows growers to utilise ag-data technologies with confidence, they will remain under-utilised or at worse become redundant. The root cause of this debate is that the majority of growers are, and will remain, dependent on third party data service providers. As discussed in Section 6.2.5 there are very few growers with the scale, capacity or desire to develop their own data solutions. Growers are creating and storing proprietary data using third party systems. The primary objective of commercial data service providers is generating viable returns through monetising their intellectual property (IP); i.e. the systems and intelligence that allow growers to collect, analyse and use their data.

The crux of the debate is: how to create a viable market for the data service providers whilst protecting growers' rights to control their proprietary data? This chapter explores the different aspects of this debate. Definitions presented in this chapter are for illustrative purposes only and should not be considered as correct legal definitions. Furthermore, the opinions expressed within the chapter are my own and owing to the sensitivity of the topic case studies have been anonymised.

7.2. Current state of the debate

7.2.1 What do we mean by data protection?

The following three elements of data protection were commonly raised by growers in my interviews:

- Protection from loss of data
- Protection from malicious theft of data
- Protection from non-authorised re-use of data

From the perspective of protecting against losing data the increasing use of web based and cloud computing can only be viewed as progress from the disparate desktop based systems of the past. Data can now be aggregated and stored in central locations on platforms with robust maintenance and disaster recovery strategies.

Hacking and poor data security are the most likely cause of malicious theft of electronic data. Whilst the risk of being hacked can never be eliminated, in the same way that you can never eliminate someone breaking into the farm office, good data companies providing software solutions will have appropriate security protocols in place and should be willing to discuss them openly when requested. Non-authorised re-use of data is probably the biggest concern amongst growers, with focus on avoiding publication of personal or sensitive data, uncertainty around intentions of third parties who are re-using their data, and ensuring appropriate remuneration when it has been used.

7.2.2. Data ownership

To fully understand the issues of data ownership it must be appreciated that there are different types of data. These categorisations are not just semantics and have significance in both legal and commercial terms. In my view, they should be considered more explicitly in the current debate on data ownership which in reality is more nuanced than 'who owns the data?

Raw data are the un-analysed source data created by an organisation. In a farming context, these could be data created from machinery, sensors or scouting reports. Derived data are new datasets created by extracting information from raw data. In agriculture aggregated data are the most common form of derived data. Aggregated refers to the merging and degrading of raw data into a lower resolution output. Software providers offering services to many growers can, by default, generate archives of information on large acreages of land which have potential off-farm value when aggregated.

The ownership of the raw data was undisputed in all the ag-data companies visited during my tour – it is owned by the farmer. However, being owned by the farmer is different to preventing others from having access to the data which will be defined by the terms and conditions; this is discussed further in Section 7.4. It is important to also appreciate the right of the ag-service provider to protect the IP of their data systems, created to handle and store the raw data. This particularly applies to software companies whose sole assets are the IP in their software solutions. Therefore, it is common for software companies to allow customers ownership of their raw data but to retain ownership of the database architecture within which these data are stored. This matters because if a farmer decides to leave a software supplier and asks for their raw data the company may choose to provide the data in a format that does not match the database structure.

The ownership of derived data is where the debate becomes more ambiguous, which reflects the nature by which derived data is created. Derived data can only be created from raw data but the additional value is created by the input of intelligence; i.e. IP, which imparts new insights through manipulation of the raw data. Therefore, there is a requirement to share the value created from the derived data between the owner of the raw data and owner of the IP whose input was required to create the added value. From my interviews this partition of value is both legally hard to define and in some cases influenced by the market value of the derived data. There is no 'standard' splitting of value and it requires a commercial negotiation and contract between the involved parties.

7.3. Protecting data ownership

Treatment of personal data is governed by data privacy laws in the country of operation. Generically, personal data is defined as information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context. In some cases this will include agricultural specific information. Formal legal counsel will be required to clarify what is a working legal definition of personal data in any country of operation. However, data privacy laws provide a legal backstop for preventing misuse of personal data.

The customer ownership of raw data is defined in the terms and conditions of use. During my studies I could not find one contract that didn't state that the raw data was owned by the customer. The area for closer consideration is: what rights that gives you. For the majority of growers the best case

scenario would be that it affords you complete control over who has access to your raw data and where your raw data can be used. This right is enacted by the power to grant access permissions and the requirement for the software provider to control access via firewalls.

As an example, a food processor has a requirement to collect field data relating to "food safety" (e.g. last pesticide application relative to harvest date) and "sustainability" data (e.g. nitrogen and water use). The processor, acting as the account licence holder, uses a farm data management solution to collect the required data from the grower base. As the licence holder the processor can grant growers access to the account to record their data. The data in this account is owned by the processor. Growers, who might also be using the same system, have a licence to collect data across the rest of the farm. To avoid double entry of the same data the ag-data company can be instructed to allow shared access to the common data and restrict access, with a firewall, to all other data.

In other cases the terms and conditions could state that you own the raw data but the licence agreement allows the data service provider perpetual access to the data for purposes defined in the agreement: for example, creating aggregated data products. This type of licence is not illegal or necessarily immoral. There may well be scenarios where this type of licence is acceptable. For example, a grower does not want to invest time in the management of the farm data for external commercial development and is comfortable allowing the data service provider to undertake that work, in particular if it delivers a benefit back to the grower. Fundamental to this arrangement is transparency from the service provider and that the terms and conditions have been clearly understood by the grower. Rightly, or wrongly, a frequently voiced perception is that companies using data services to augment their primary business are "*only doing it to sell me something'*. This point of view stems, in part, from a combination of opaque intentions and a lack of knowledge around the content of the terms and conditions.

Details agreed in the terms and conditions are also central to handling derived data. As concluded in the previous section there is no standard for how to share the commercial revenue of derived data. This could be achieved through royalties, a flat fixed fee or split percentages of total revenue. The relative merits of each have to be assessed and negotiated on a case-by-case basis and there is a requirement for transparency to ensure each party has clarity on what has been agreed.

In the US the contributions of Ag Gateway and American Farm Bureau Federation, profiled in Chapter 5, have been helpful in promoting the debate on the ownership and sharing of data. The provision of knowledge to growers about terms and conditions are important first steps.

7.4. Data ownership in a commercial framework?

Conversation with growers highlighted that there is not a uniform view on the topic of data ownership; a whole range of views was expressed from "*I will never share any data"* through to willingness to be completely open with data if it benefited the business. Also the currently low rates of utilisation of ag-data mean there is considerable scope for growth and development of different business models differentiated by their approach to data ownership.

The theoretical range runs from one extreme of the customer owning and not sharing any data through to the other where the provider owns all rights to the data. In between are hybrid models

with increasing amounts of shared ownership of derived data. These models will be differentiated by cost to the grower. Closed control of all data can be achieved through two methods that come at a premium to the market; first, developing standalone in-house data systems with the associated development and maintenance costs, or second, using closed accounts on third party software systems.

For a software provider the latter approach means that your use of their product is the only method of generating revenue and alternative revenue streams cannot be exploited. As the proportion of data sharing increases, so the potential to generate revenue from alternative income streams increases. This reduces the reliance on generating income from the grower. The reduction in relative cost to the grower could be achieved through several possible routes: a reduction in the unit cost of an account to incentivise uptake; revenue sharing of downstream data sales; improved farm production through access to relevant benchmarking data; or reduced data management costs from automated transfer of data, for example. Deciding which data model is appropriate is a cost-benefit decision to be taken as part of a broader business data strategy, discussed in Chapter 8.

7.5. Chapter summary

The issue of data ownership is a more nuanced topic than much of the current debate. Explicit consideration must be given to raw data, derived data and protection of technical IP. The farming sector contains the whole spectrum of views on this topic, from a completely proprietary stance through to a willingness to openly share data. The more conservative end of the spectrum has been the most vocal participant in the current debate but does not represent a universal view. Fundamentally what is apparent is that there will not, nor should there be, a single data ownership solution. Growers will have to choose what approach suits their requirements. The ideal should be that this can be achieved in a market with sufficient choice and, crucially, availability of information.

Chapter 8 – The power of a data strategy

8.1. Introduction

An inescapable conclusion when collectively evaluating my visits and interviews is that data is an enterprise asset and the farm businesses that had embraced this concept had a far more effective relationship with their data. Treating data as an asset goes far beyond the debate over ownership and into designing effective strategies for integrating the relevant people, processes and technology that enable data to deliver commercial benefit. Like any asset, underinvestment and lack of strategic vision will, inevitably, lead to ineffective implementation and probably failure to deliver intended business benefits. This chapter reviews the central concepts and components that should be considered when designing and implementing an effective data strategy.

8.2. Does your business have a data problem?

Before any business can implement a data strategy it must develop awareness that current practices are not fit for purpose. The following 'symptoms' indicating poorly executed data strategy are taken form the report "Creating an enterprise data strategy" by Wayne Eckerson *[1].* In the author's words *"if your organisation exhibits any of these characteristics it is wasting time and money, jeopardising the success of major initiatives and forfeiting valuable opportunities to gain a competitive edge".* The list below is abbreviated to the symptoms which have the most relevance to farming businesses. The terms "business users" and "IT" are used generically.

- The business retains customers who cost more money than they generate. "Customers" could be replaced by "field" and the statement would still be valid.
- Meetings degenerate into arguments about whose spreadsheet is right
- Business users think buying tools will address data and system problems
- Business users don't trust IT to deliver applications or data in a timely manner
- There is no process for archiving data that must be retained or disposing of data that's no longer needed.

These symptoms of day-to-day business operations combine to form an overall picture of the data governance practised in a business. Tony Fischer, a leading expert on data quality and management, has created a data governance maturity model (*Table 8.1, on next page*). Whilst absolute numbers are hard to measure, from my observations the majority of farming businesses are Level 1 or 2 (Undisciplined or Reactive). A small number are operating at Level 3 (Proactive) and a select few have achieved Level 4 (Governed). Moving a business from the lower to higher levels is not straightforward and requires strategic vision and planning.

8.3. What is an enterprise data strategy?

Data strategy can be divided into two components:

- data governance, the strategic policy
- data management, the practical implementation of the policy.

Level		Description of business behaviour
1	**Undisciplined**	Inability to adapt to business changes Executives unaware of the cost of poor data Reactive, IT-driven projects often carried out by individuals Duplicate, inconsistent data No standards for cleaning or sharing data
2	**Reactive**	Short range projects Line of business influences IT projects Little cross functional collaboration High cost to maintain multiple applications Little executive oversight
3	**Proactive**	Data viewed as a strategic asset IT and business collaborate Data stewards maintain corporate data definitions and business rules Emphasis on preventing rather than fixing data quality problems
4	**Governed**	Business requirements drive IT projects Data projects funded appropriately Business processes are automated and repeatable Executives trust data based decisions

Table 8.1: Categorisation of data governance in businesses; adapted from Table 1 in [2].

The Master Data Management Institute defines data governance as "*the formal orchestration of people, process and technology to enable an organisation to leverage data as an enterprise asset*". The sequence of people, process and technology is deliberate. A successful strategy starts with an evaluation of the people with respect to having the correct blend of data skills, domain knowledge and management engagement to champion data quality. Process describes the how, what and why are data being collected and used *(Table 3.1).* It exploits the defined reasons why data need to be collected, used, and reused, across different business departments. Central to this is defining which personnel are responsible for each 'touch point' in the lifecycle of a piece of data, i.e. when human interaction with a piece of data is required, who is it and what do they have to do.

Once this theoretical data model has been created consideration can then be given to procurement of the technology required to deliver the relevant data. Data management is the practical implementation of the conceptual strategy. Like the strategy, personnel are key. Engagement from senior management is paramount to ensure that individuals responsible for key decision points feel empowered to maintain the integrity of the process and data. This approach is opposite to common current practice of having an ad-hoc procurement policy for data technology which is then shoehorned into pre-existing procedures.

What has been obvious from my farm visits and reading data strategy literature is that there is no 'one size fits all' data solution. Managing the data requirements of the 'production issue' faced by growers requires using multiple tools which can be operated in isolation, manually integrated or automatically integrated. Like physical tools, understanding which tool is right for the job is crucial.

In many cases the frequently heard complaint that the "software (or system) doesn't work" reflects user expectations being outside the scope of the software rather than flawed software. The heavy reliance of many growers on Excel spreadsheets to record, store and analyse their data is a good example. Spreadsheets are designed to hold tabular data for analysis. Excel is extremely good at doing this. However, it was not designed to be a data warehouse or data management system. It does not contain the in-built data controls of database software that ensure integrity, secure storage and accessibility of data. In this example, the frustrations listed by growers about the limitations of what they can do with their data are primarily artefacts of their out-of-scope expectations of Excel.

The final element of good data strategy is appreciating that this is a dynamic process. The evolution of technology, new understanding and changes in required skills mean that, over time, best practice will change. Periodic reviews of data processes and technology should be undertaken from first principles. The starting point for reviewing each element should be: why? Why do we collect this data and what purpose does it serve? If this question can be answered in the affirmative then review the procedures used to collect and integrate the data; can they be improved through new technology or more effective working practices? Where necessary, alterations should then be made to the process.

8.4. Case studies

Treating data as an enterprise asset implicitly implies that it can be assigned a defined financial value. In practice quantifying the net value can be challenging and will be different for compliance, operational and agronomic data. Data collected for ex-farm compliance or legislative purposes are, in effect, costs of production defined by the software licence and the man hours spent collecting and processing the required data. Operational and agronomic data at the point of creation are also costs of production. They are primarily documentary records of business operations and crop status, respectively, at that moment in time. As inputs to a season review and planning meeting they have potential to improve operational performance in the subsequent season. Defining the net value to the business is made complex by the fact that the cost and benefit of the data can be incurred and realised in different seasons.

A review of business solution requirements is one method that can help to define the gross and net value of data to the business. The purpose of the review is to formally document the requirements, objectives and constraints of all business processes (or subsets thereof depending on the scope of the review). A 'process' in this context is any function within the business where data are recorded. The review should deliver a definition of current requirements, analysis of the data flow required to meet the requirements, and the cost of core and optional products required by the process. This allows formal definition of both the costs and benefits to the business, from which net value can be determined. Furthermore it provides a benchmark against which alternative solutions can be evaluated. For large businesses this type of review is a time consuming process but ensures efficient and value-for-money solutions are adopted for each process.

Black Gold Farms *[3],* a large multi-state US potato producer, is undergoing this exact process under the IT director Bert Buckholder. Their aim is to deliver necessary capabilities with their off-the-shelf

procurement and also to document historical bespoke processes that have been developed. This will enable redesigns to be undertaken as appropriate.

Established in 2005, Black Earth Farms (BEF) operates a land bank of *c*. 275,000 ha in the Voronezh Oblask region of Russia *[4].* In 2011 the business instigated an operational review and strategic overhaul in the face of persistent trading loses and endemic operational inefficiencies. Whilst many of the challenges faced by the business were about fixing basic financial, operational and agronomic practices, the strategic approach adopted by the management team, under CEO Richard Warburton, provides a good case study of using data within a strategic business plan.

Their overarching commitment was that business operations should be based on robust decision making and science, which intrinsically places a requirement for relevant data at the heart of the strategy. Beneath this it was recognised that, to effectively measure progress, evaluation of financial and operational performance needed to be divorced.

At the operational level a complete review of available agronomic data was undertaken and concluded that current agronomic knowledge was not fit for purpose. This resulted in a research farm being created to generate the required agronomic data. A financial review of the economic performance by field and crop type led to a 35,000 ha reduction in planted acreage in 2014, removal of spring sown oilseed rape from the rotation, improved discipline on planting winter crops, and the transfer of lighter land into an irrigated land bank for vegetable production. In 2014 potato was planted on c. 1000ha of irrigated land for Frito-Lay.

Longer term plans are to expand the irrigated land bank to 16,000 ha to grow a mixed vegetable rotation. To tackle endemic inefficiencies in operational performance all machinery, including contractors' vehicles, is fitted with transponders linked to a central command centre which is monitored 24 hours a day (*Figure 8.1*). Using a system developed by Ukrainian company Cropio *[5]* all the farm operations are monitored and recorded through the system.

Improvements in operational and agronomic performance have been substantial during the last three years. The summer of 2014 was as hot and dry as 2010 when Russia suffered diminished yields owing to summer drought. During my visit yields reported from winter wheat were at or above budget in the majority of fields.

In rare cases external factors can create a market, which imposes an external value, for operational data. In 2007 the Alberta state government introduced legislation to force reduction in CO_2 emissions from primary industrial producers. Two options were available in the legislation: first, physically reduce emissions; or second, either invest in the development of new technologies or purchase carbon offsets. In 2009 the legislation protocols defined min-till farmland as a carbon sink and therefore eligible for payments under the legislation.

The current value of the scheme is worth C$1.50/ac and has seen c.C$30m returned to growers practising min-till farming in Alberta. The legislation has created an additional income stream for growers but this remains additional to, rather than the cause of, why the operational data were initially collected.

Figure 8.1: Central control room at Black Earth Farms.

8.6. Chapter summary

Data is a strategic business asset and like any asset maximising its commercial value requires strategic vision yet it can be difficult to define its net value. Whilst the cost of data collection and storage is relatively easy to define, the benefits are often realised elsewhere in the business and potentially in different financial years. Furthermore, quantifying the cost of bad data on business performance is extremely hard. Developing and delivering a data strategy is an involved and complex process, in particular for large businesses, as illustrated by the examples of Black Gold and Black Earth Farms. However, the principles these companies have adopted are independent of scale and technology used. The objective is to ensure maximisation of the data asset in the business.

Chapter 9 – Future developments and remaining challenges

9.1. Introduction

This chapter explores areas which, I believe, have the potential to limit the potential of agriculture to fully exploit data technology. They are based on personal observations made during my study tour.

9.2. Does the industry have the right data research agenda?

Randomised block experiments, pioneered by Fisher at Rothamstead Research Centre, have been the bedrock of advancing agronomic understanding over the last century. They allowed improvements in commercial production to be driven by rigorous statistical testing of hypotheses. They remain as relevant today as ever to progressing agronomic insight through academic and commercial research programmes, as they allow estimation of variance and reduce the impact of confounding factors on results.

However, these methods are not appropriate for analysing large observational datasets now being created. Observational datasets are by their very nature un-replicated, which means they fail the fundamentals of experimental design; i.e. randomisation, blocking and replication. Observational data frequently contain visual patterns that are often erroneously interpreted as significant. The power of observational data lies in their capacity to record phenomena and processes at many different spatial and temporal scales. The challenges of analysing un-replicated data have long been considered by other academic fields where replicated experimentation is not possible: for example, econometrics. These fields have created strong statistical methods for interpreting un-replicated data and have direct applications in agricultural research. Prior to his retirement from the University of California, Davis Richard E. Plant published "Spatial Data Analysis in Ecology and Agriculture Using R" *[1],* which he wrote, in his own words, as a piece of "disguised propaganda" to challenge the industry and research into thinking about how and why we quantify the variability in spatial relationships described in observational data.

Developing the research theme articulated by Richard Plant would be advantageous to both research and commercial production. In research, a robust analytical protocol that can generate research questions from on-farm observational data would be a powerful augmentation to orthodox agronomic research. Commercial agriculture, especially precision agriculture, is reliant on proprietary algorithms which can be impossible to test if no information is released by the company which owns the IP. Many of the central research questions in agro-ecology, such as understanding the biophysical and biochemical processes and relationships that determine crop production, are location specific. Precision agriculture tools represent these processes either implicitly or explicitly. Research focused on developing methods and applications, using spatial statistics, to test if the representation of relevant environmental processes is robust, would help to identify and define best practice.

Disappointingly, it appears that this potential is, currently, not widely recognised. At the 2014 International Society of Precision Agriculture (ISPA) conference, held in Sacramento, the overwhelming focus of papers was around output from sensor technologies. There was no discussion on analytical methods, improving observational data analysis or using observational data

to drive understanding in agronomic processes. The reasons behind this current state are complex. The orthodoxy of traditional trials in agricultural research is strong and it will take time for the community to recognise that alternative methods for creating insight have merit. Sensor technology has become a standalone research field in its own right, which prioritises technological development. Commercially there will be limited impetus, especially if the research might highlight weaknesses in products being sold. Finally, the industry has not previously required the statistical and analytical skillset, on a wide scale, necessary for this type of research or for use in a commercial context.

9.3. Does the industry have the right balance of data skills?

9.3.1. Digital immigrants versus digital natives

A point of view expressed on several occasions including the ISPA conference was the notion that progress required the replacement of so called 'digital immigrants' with the 'digital natives'. Digital immigrants are those born before the digital age; i.e. they have metaphorically immigrated into the digital era. Digital natives are those born and bought up in the digital era. The premise of this argument is that the younger generation have an intuitively better understanding of digital technology and are therefore better placed to exploit it.

From a data perspective this argument is fundamentally flawed. It relies on the assumption that understanding how to use a piece of digital hardware or software automatically means the user understands why they are doing it or can correctly interpret the generated data. This does not automatically follow. As Stephen Few eloquently puts it in the quote below, much of the intelligence required to extract meaning from data lies with the digital immigrants who have the learned experience and knowledge:

> *"Computers speed up the process of information handling but they don't tell us what the information means or how to communicate its meaning to decision makers. These skills are not intuitive; they rely largely on analysis and presentation skills that must be learnt". [2]*

It is worth remembering that data contained within a well-designed data system is only there because it represents an element considered important by the designer. It would be to the considerable detriment of agriculture if the collective knowledge and experience of the 'digital immigrants' is ignored. The central point to recognise is that their contribution is independent of technology; i.e. they probably have a better understanding of why the piece of data needs to be in the system. The real challenge for the industry is exposing this experience to the data scientists, analysts and managers sourced from the 'digital natives' who will be designing, building and eventually using the systems over the next ten to twenty years.

9.3.2. Attracting data skills into agriculture

The growth of data usage currently being experienced in agriculture is not a unique phenomenon. Many other business sectors are also undergoing similar structural changes from their own 'digital

revolutions'. A consequence of this is a rising demand for data-related skills and expertise across a broad range of economic sectors *[3, 4].* Furthermore, it should be remembered that increases in system efficiency and automation delivered to the end user are underpinned by skilled expertise in the design, maintenance and analysis of IT and data systems. These skills are increasingly generic and are not application-specific.

The core workforce in the majority of ag-data companies I visited had agricultural backgrounds, primarily through family history. The calibre of these staff is not being challenged but the capacity for agriculture as a sector to generate sufficient candidates, of the required calibre, to meet the growing requirements of the ag-data sector is. The available evidence would suggest that this will become increasingly difficult. As such, agriculture will have to learn to become more comfortable with recruiting talent external to the industry. This is made more challenging by having to compete in an increasingly commoditised market already facing shortages of suitably qualified candidates.

The suite of technical skills required to service data and information is broad, ranging from web design through database management and into analytics and 'Data Science'. Data scientists blend a mix of developmental skills, knowledge of mathematics and statistics, and relevant domain expertise *[5].* These skills are increasingly universal and no longer confined to specific industries or commercial sectors. Furthermore, projections for the growth of the data analytics sector in the UK predict *c.* 55,000 jobs being created per annum by 2020 *[4].* These projections indicate that the market for high grade data scientists and analysts will become highly competitive.

Therefore, it is important for agriculture, as a sector, to consider what it can do to ensure access to the required talent on an ongoing basis into the future. This process should evaluate both physical and promotional measures. Physical measures could include reviewing what strategic alliances should be made with academic institutions: for example, would it be sensible to establish formal links with university departments of informatics? Placing agricultural case studies on the undergraduate syllabus in these departments would expose the sector to data science undergraduates. *"I like the problems and the challenges created by my work"* was a standard response from employees with non-agricultural backgrounds working in the ag-tech companies I visited; i.e. using data to improve our knowledge of agricultural production is a stimulating intellectual exercise. Whilst this is not the only reason for external talent entering the industry it does seem to be a strong driving force and is one that agriculture should look to forcibly exploit.

Competitive remuneration through salary or, as is common among start-up companies, equity in the business, is the obvious way of attracting suitable data talent. The latter approach provides the incentive of significant capital gain should the company be acquired at a later date. Well-funded start-ups and new companies have adopted this approach, US companies in particular. They have recognised that the success of their business is dependent on investing in high-grade staff. However, the extent to which the broader industry values these skills and is prepared to invest in competitive remuneration remains unclear.

9.4. Defining the true value of data

An implicit assumption being made across agriculture is that farm data is intrinsically valuable. A coherent argument can be constructed to challenge this assumption. Data are only commercially

valuable if they are ‘useful’; i.e. by having possession of a dataset it enables an organisation to make decisions that make them more profitable. There are large volumes of ‘interesting’ data that are not ‘useful’. However, considerable amounts of ‘interesting’ data are sold as ‘useful’ to unprepared clients. Furthermore, the market value of useful data is not dependent on the cost of production, which is often forgotten by those who create data and have first-hand knowledge of the cost of production – they erroneously assume that data which have a high cost of production are automatically of high value. Best practice, from a data strategy perspective, would be to regularly test the intrinsic value of data used in the business - is the value of decisions taken using these data greater or smaller than the cost of data procurement or production? i.e. is this data ‘useful’ or just ‘interesting’.

Currently ‘useful’ data on-farm may only be ‘interesting’ ex-farm. This, I believe, partly reflects poor data management practices which reduce the capacity to derive a combined output of commercial value. Standard practice in good data management is unique identifiers which are allocated to each data entity and are used to manage the relationship between data elements. Unique identifiers are simple to implement within a well-designed standalone database but become highly challenging and prone to error when considered across multiple, separately designed systems or poorly managed systems. Systems developed separately will not have common unique identifiers and most likely the same identifier will reference a different data element in each system. From personal experience, the biggest single overhead in analysis of agricultural data from individual farming operations has been the creating of a definitive unique list of field identifiers from which the subsequent analysis can be driven. For many projects the cost of data cleansing will be prohibitively expensive.

Other commercial data sectors have developed solutions by generating datasets that contain unique identifiers of physical entities at a national scale. In the UK the national mapping agency, the Ordnance Survey, sells a product called AddressBase *[6]* which contains the location of all 28 million addresses in the Post Office’s Postcode Address File (PAF) *[7].* AddressBase allows users to relate any other data that can be associated to a property (e.g. sales information, estate agent listings, property attributes) to a nationally recognisable and consistent unique identifier for the property in question.

Figure 9.1. Sample of Ordnance Survey AddressPoint data, each numerical label (e.g. 8759912) is maintained as persistent unique identification for each address in the UK.

This means the data management overhead in integrating property-related data is not subject to the same issues of integrating agricultural data from separate systems with unrelated identifiers. The UK does have the Rural Land Register which "holds details of all registered land parcels in a digital format" *[8]* and is used by the Rural Payments Agency in their processing of Single Farm Payment and Environmental Stewardship Schemes in the UK. Quite rightly the raw data are not available for public release or commercial re-use as they contain personal data.

However, using AddressBase as a conceptual template it should be feasible to create a database of field location - the field centroid would suffice - and a unique identifier, whilst protecting personal data. A national field database allowing field data to be integrated at much lower cost than present would increase the inherent ex-farm value of data. Creating a national field database would be an involved process requiring consideration of licencing, commercial terms, revenue models and technical issues such as maintenance and update cycles. However, the upside from a data management and analytics perspective would be worthwhile.

The other factor which reduces the ex-farm value of agronomic data relates to the concepts of data veracity (see Section 6.2.5) which make deconstructing observational data back to agronomic first principles extremely difficult. This capacity is required to truly understand the observed signal in the data and from which meaningful decisions can be made. This issue is further confounded by the fact that agronomic decision making, even in precision agriculture, remains largely subjective. Recording subjective decision making as data is extremely difficult; however, the decisions made and subsequent actions are integral to interpreting the observed signal in the crop. When considering these factors at a regional or national scale it is clear that there is no simple solution for undertaking first principle analysis of agronomic data at these spatial scales.

Operational data offer greater opportunities, in the short term, for delivering value as an ex-farm asset. However, the 'revenue' generated ex-farm is not necessarily to be received directly as income. Benchmarking groups have proved themselves to be effective at comparing the relative costs of production amongst contributing members. The technology is there to allow expansion of the spatial scale and scope of benchmarking if desired.

Machinery manufacturers are interested in telematics data as it allows them to monitor machinery performance, identifying parts nearing the end of their working lifespan amongst other points. These data can be used to create products and services to promote loyalty amongst their customer base.

Food supply chain customers increasingly require operational data to demonstrate that the grower has met legislative or customer-specific due diligence obligations. These data are valuable to the grower's client but there is no additional remuneration to the grower ex-farm. All the value is contained within the contract between the grower and client.

9.5. Data quality

Dan Frieberg, CEO of Premier Crop Systems *[8],* titled his August 2014 column in Corn and Soybean Digest: 'Trust…but verify' *[9].* In this he argued that growers need to become more robust in checking the validity of the technology and data they use and create, respectively. There is an inherent assumption made by many growers that, by simply existing, data are 'correct'. This is

worrying and is a significant barrier to progress in the effective utilisation of data. Adopting the mantra 'trust, but verify' would be an advantageous step for any farming business.

Verification of data starts by defining how the data are to be used. This will define the limits within which the data are valid. This is important as data describing the same variable will have different limits depending on application. For example, in a formal field trial the tolerances of precision and accuracy are low. However in a commercial crop a directional trend might be sufficient. Once the envelope of operation has been defined then physical quality assurance (QA) steps can be established to screen data before use. A crucial step, often ignored, is defining what to do if data fail QA. This will reduce the risk of using poor quality data and maximise the window of opportunity to recollect data if necessary.

9.6. Open source technology

With the exception of the Soils for Life project at Produce World no other farming business I visited was making extensive use of open source software, which can be installed and used free of charge, most commonly under the terms and conditions of a General Public Licence (GPL) *[10].*

There are now robust open source tools for all data types and parts of the data lifecycle; these are now widely used in many commercial sectors and research environments [*11, 12].* The advantages of open source software are zero procurement costs and access to a large user community. The disadvantages are a requirement for re-training or up-skilling of existing staff and a lack of traditional technical support from the software vendor. However, technical support is available from commercial consultancies specialising in open source technologies.

Chapter 10 – Conclusions and Recommendations

Data alone is not going to revolutionise agriculture. The potential of data can only be realised through integration with appropriate intellect and skills that allow meaning and insight to be extracted. For this to be achieved the following are necessary:

- *Adoption of business data strategies.* A formal data strategy explicitly defines the relationship between people, process and technology used to collect, collate and analyse data within a business. This architecture is central to ensuring that procurement of data tools or systems is within scope and fit for purpose.

- *Adopting a culture of 'trust, but verify'.* Data are inherently noisy and correlations within data are not necessarily an indication of causation. The integrity of data should always be challenged before acceptance of the information they contain.

- *Appreciation that 'data' is a generic term*. 'Data' is a generic term to describe raw, derived, aggregated and metadata. Appreciating the difference between these is important: in particular in the context of understanding data ownership, terms and conditions or types of business models offered by data service providers.

- *Accepting that data technologies are agnostic to application or geography.* Agriculture is not alone in being challenged by new data technologies. Many research and business sectors are in a similar position. Globally, there is experience and expertise that can be exploited from many non-agricultural sectors. These are a valuable resource and should not be wasted.

- *Pan-industry initiatives*. There is scope for pan-industry or national strategic initiatives to support use and uptake of ag-data technologies. Ensuring agriculture is included in university data science courses will help attract relevant skills into the industry. Creating national reference datasets (e.g. national field dataset) for commercial use would drive effective development of the ag-data market.

Chapter 11 - After my Nuffield Farming study tour

In many ways the 'after my Nuffield Study Tour' started before I had finished! The theme of the 25th Annual Cambridge University Potato Growers Research Association (CUPGRA) conference, held in December 2014, was the role of new data technologies. I was asked by the organising committee to speak about the findings of my Nuffield study tour and the role of data technologies in agriculture. Two delegates at the CUPGRA conference were from CSS Farms, a large US potato producer, who subsequently invited me to address their annual company meeting on data strategy in Bakersfield, California, in January 2015.

During 2014 the Agri-Tech East was established to create an innovation hub in the east of England, to accelerate the application of research, generate opportunities for economic growth and create a competitive advantage for UK agricultural companies. An important component of the hub is its special interest groups (SIG). These are forums to allow companies and individuals in specific ag-technology areas to meet, discuss and form collaborations. One SIG is for Big Data and in January I accepted the position as the industrial co-chair of the group. The role of the co-chair is to help guide the objectives and content of events. During 2015 the Big Data SIG will be holding events on data strategy, data visualisation and the role of metadata data in data integration.

In September 2014 I accepted the position of Research Manager at Greenvale AP. The companies' research portfolio includes variety evaluation, field trials and collaborative projects with academic partners. Effective data management is integral to ensuring that the research programme is both effective and value for money. Many of the insights gained during my Nuffield Study Tour will be central to implementing effective data solutions. This process has already started with the design of a new database architecture for housing all variety and field trial data. This is allowing more rigorous analysis of these data.

For me, the relevance of my Nuffield Farming experience can be summed up by the two books I refer to most often in my work:

- Firstly, the British Potato Council's research review: Potato Agronomy: the agronomy of effective potato production (*Allen and Scott, 2001)*
- and secondly, Automated Data Collection with R – a practical guide to web scraping and text mining (*Munzert et al., 2014).*

In many ways our understanding of agronomy and the factors that affect it are well known. But the array of tools we now have at our disposal to collect and analyse agricultural data is large and expanding. Being able to determine the appropriate tool for the relevant agronomic job is not easy and requires the integration of agronomic and data expertise.

References

Chapter 2

[1] http://en.wikipedia.org/wiki/The_Climate_Corporation; retrieved 28/01/2015
[2] The Hale Group, 2013. '*Agricultural Big Data Overview'* Investing in Agriculture Symposium, New York City, December 3, 2013
[3] http://knowledge.wharton.upenn.edu/article/can-big-data-feed-world/; *retrieved 28/01/2015*
[4] http://www.bbc.com/news/business-26424338; *retrieved 28/01/2015*
[5] http://www.enterrasolutions.com/2014/09/big-datas-big-role-agriculture.html; *retrieved 28/01/2015*

Chapter 3

[1] Harris J.G., 2010 *'How to turn data into a strategic asset*'. The journal of high-performance business
[2] Foresight. The Future of Food and Farming (2011) Final Project Report. The Government Office for Science, London.
[3]http://webarchive.nationalarchives.gov.uk/20130123162956/http://www.defra.gov.uk/statistics/files/defra-stats-foodfarm-farmmanage-earnings-labour2011-20110610.pdf; retrieved 03/02/2015

Chapter 4

[1] https://myjohndeere.deere.com; retrieved 29/03/2015
[2] https://encirca.pioneer.com/landing/; retrieved 29/03/2015
[3] http://en.muddyboots.com/; retrieved 29/03/2015
[4] http://www.farmade.com/farmade-software-products/; retrieved 29/03/2015
[5] http://agconnections.com/; retrieved 29/03/2015
[6] http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aOEnCbyUpCuY; retrieved 29/03/2015
[7] http://www.granular.ag/; retrieved 29/03/2015
[8] http://www.conserviscorp.com/; retrieved 29/03/2015
[9] http://www.agritrend.com/services/data-management/service-overview.aspx; retrieved 29/03/2015
[10] https://farmobile.com/; retrieved 29/03/2015
[11] http://www.640labs.com/; retrieved 29/03/2015
[12] http://www.farmcommand.com/; retrieved 29/03/2015
[13]http://www.reuters.com/article/2014/12/08/ca-the-climate-corp-idUSnBw086353a+100+BSW20141208; retrieved 29/03/2015
[14] Precision Agriculture: In the Field and Beyond the Farm Gate, Davina Fillingham, 2014
[15] Precision Agriculture: how to realise the full potential, Andrew Williamson, July 2014
[16] http://precisionhawk.com/; retrieved 29/03/2015
[17] http://www.ursula-agriculture.com/; retrieved 29/03/2015
[18] https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/sentinel-1; retrieved 29/03/2015
[19] https://agvesto.com/home; retrieved 29/03/2015
[20] http://climate.com/; retrieved 29/03/2015

Chapter 5

[1] http://www.aggateway.org/; *retrieved 28/01/2015*
[2] http://www.aggateway.org/eConnectivity/AGIIS.aspx; *retrieved 28/01/2015*
[3] http://www.aggateway.org/eConnectivity/Projects/CurrentOngoing/SPADE.aspx; *retrieved 28/01/2015*
[4] http://agglossary.org/wiki/index.php/Main_Page; *retrieved 28/01/2015*
[5] http://openag.io/; *retrieved 28/01/2015*
[6] http://openag.io/about-us/principals-use-cases/; *retrieved 28/01/2015*
[7] http://www.fb.org/index.php?action=issues.bigdata; *retrieved 28/01/2015*

Chapter 6

[1]http://www.forbes.com/sites/bruceupbin/2013/10/02/monsanto-buys-climate-corp-for-930-million/; *retrieved 28/01/2015*
[2] http://knowledge.wharton.upenn.edu/article/can-big-data-feed-world/; ; *retrieved 28/01/2015*
[3] http://www.bbc.com/news/business-26424338; ; *retrieved 28/01/2015*
[4] http://www.enterrasolutions.com/2014/09/big-datas-big-role-agriculture.html; *retrieved 28/01/2015*
[5]http://www.cleantech.com/release/i3-quarterly-innovation-monitor-reports-trends-in-sustainable-innovation-from-third-quarter-of-2014/; *retrieved 28/01/2015*
[6]http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf; *retrieved 28/01/2015*

[7]https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/; *retrieved 28/01/2015*
[8] http://www.sciencedaily.com/releases/2013/05/130522085217.htm; *retrieved 28/01/2015*
[9] Figure 9 in "The digital universe in 2020: Big Data, Bigger Digital Shadow s, and Biggest Growth in the Far East" (Gantz and Reinsel).
[10] Slide 2 from "knowing when and what customers are buying based on travel data" by Tobias Wessels; Source: http://www.slideshare.net/DDMalliance/adara-presentation-ddm-may-16-14; retrieved 28/01/2015
[11] http://en.wikipedia.org/wiki/The_Climate_Corporation; retrieved 28/01/2015
[12] http://hadoop.apache.org/; retrieved 28/01/2015
[13] http://en.wikipedia.org/wiki/High-frequency_trading; retrieved 28/01/2015
[14] http://www.perceptualedge.com/; retrieved 28/01/2015
[15] http://www.agri-data.net/www/news/; retrieved 28/01/2015
[16] https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/; retrieved 28/01/2015
[17] http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory; retrieved 28/01/2015
[18] http://www.bbc.com/news/business-30095454; retrieved 28/01/2015
[19] Nature, Vol. 457, 19 February 2009, doi: 10.1038/nature07634
[20] Science, Vol. 343, 14 March 2014, doi: 10.1126/science.1248506
[21] http://gking.harvard.edu/files/gking/files/ssrn-id2408560_2.pdf; retrieved 05/02/2015
[22] http://news.monsanto.com/press-release/corporate/midwest-farmers-prepare-first-fieldscripts-harvest; retrieved 28/01/2015
[23] http://www2.cs.uregina.ca/~rbm/cs100/notes/spreadsheets/tufte_paper.html; retrieved 28/01/2015
[24] http://www.theguardian.com/public-leaders-network/2013/jul/15/deluded-public-sector-ipsos-mori; retrieved 28/01/2015
[25] http://www.ofsted.gov.uk/annualreport1213/regional-performance-2012; retrieved 12/12/2014
[26] http://www.vox.com/2014/11/14/7213833/peter-thiel-palantir-paypal; retrieved 28/01/2015

Chapter 7

[1] http://www.precisionag.com/opinion/joe-russo/data-privacy-ownership-in-precision-agriculture/; retrieved 28/01/2015
[2] http://www.fb.org/index.php?action=newsroom.news_article&id=188; retrieved 28/01/2015
[3] http://www.forbes.com/sites/emc/2014/07/08/who-owns-farmers-big-data/, retrieved 28/01/2015

Chapter 8

[1] Eckerson W., 2011. *'Creating an Enterprise Data Strategy; Managing data as a Corporate Asset'*
[2] Fischer T., 2009. *'Data Governance Part II: Maturity Models – A Path to Progress'*
[3] http://www.blackgoldpotato.com/; retrieved 28/01/2015
[4] http://www.blackearthfarming.com/about.html;; retrieved 28/01/2015
[5] https://cropio.com/; retrieved 28/01/2015

Chapter 9

[1] Plant R.E., 2012. *Spatial Data Analysis in Ecology and Agriculture Using R. 648pp,* ISBN 9781439819135
[2] http://www.perceptualedge.com/; retrieved 28/01/2015
[3] http://www.extension.harvard.edu/hub/blog/extension-blog/why-data-science-jobs-are-high-demand; retrieved 28/01/2015
[4] http://www.sas.com/en_gb/news/press-releases/2014/october/demand-big-data-skills-analytics.html; retrieved 28/01/2015
[5] http://www-01.ibm.com/software/data/infosphere/data-scientist/; retrieved 28/01/2015
[6] http://www.ordnancesurvey.co.uk/business-and-government/products/addressbase.html; retrieved 28/01/2015
[7] http://www.poweredbypaf.com/; retrieved 28/01/2015
[8] http://webarchive.nationalarchives.gov.uk/20140305104944/http://rpa.defra.gov.uk/rpa/index.nsf/0/57EB5CADAD0BFAD580256F72003D47AE?Opendocument; retrieved 28/01/2015
[9] http://cornandsoybeandigest.com/precision-ag/data-decisions-use-technology-verify-company-data; retrieved 01/04/2015
[10] http://www.gnu.org/licenses/; retrieved 28/01/2015
[11] http://www.mysql.com/customers/; retrieved 28/01/2015
[12] http://www.revolutionanalytics.com/companies-using-r; retrieved 28/01/2015

Appendix 1 – List of interviews

Technological

640 Labs – Chicago, Illinois, USA
Ag Space – Swindon, Wiltshire, UK
Agri Data Solutions – Calgary, Alberta, Canada
Agri Trend - Calgary, Alberta, Canada
Agrii, UK
Beyond Agronomy – Three Hills, Alberta, Canada
Conservis, Minneapolis, Minnesota, USA
Farmers Edge, Winnipeg, Manitoba, Canada
Farmobile, Kansas City, Kansas, USA (Telephone interview)
Granular, San Francisco, California, USA
John Deere Innovation Center – Des Moines, Iowa, USA
Muddy Boots software – Ross on Wye, Herefordshire, UK
Point Forward Solutions, St. Albert, Alberta, Canada
Premier Crop - Des Moines, Iowa, USA
SOYL – Swindon, Wiltshire, UK

Farm/Agronomy

Black Earth Farms – Voronezh, Voronezh Oblast**,** Russia
Black Gold Farms – Grand Forks, North Dakota, USA
Cobrey Farms - Ross-on-Wye, Herefordshire, UK
Harold Perry, Perry Brothers - Lethbridge, Alberta, Canada
Innisfail Growers – Innisfail, Alberta, Canada
Produce World – Peterborough, Cambridgeshire, UK
RD Offutt Company – Park Rapids, Minnesota, USA
Robert Salmon – Norfolk, UK
Spearhead International – Swaffham Prior, Cambridgeshire, UK
Sunrise Ag – Taber, Alberta, Canada
Tasteful Selections – California, USA

Academic

Professor Dennis Buckmaster, Open Agricultural Data Alliance (OADA) – Purdue University, West Lafayette, Indiana, USA
Professor Chris Rawlings - Rothamstead Research, Harpenden, Hertfordshire, UK
Professor Colin Adams, Centre for Informatics, University of Edinburgh (Telephone interview)
Professor Richard E. Plant - University of California, Davis (Telephone interview)
Richard Heath, University of Sydney, Sydney, Australia

Other

Ag Gateway, Washington DC, USA
American Farm Bureau Federation, Washington DC, USA
International Society of Precision Agriculture (ISPA) Conference – Sacramento, California, USA. July 20th-23rd July, 2014. Delegate
25th Cambridge University Potato Growers Research Association Annual Conference – Cambridge, UK. 16th-17th December 2014. Session speaker '*Making Sense of Big Data'*
Office of National Statistics (ONS) – Southampton, Hampshire, UK.
CCS Farms Annual Meeting – Bakersfield, California 28th-30th January 2015. Invited speaker

Executive summary

Today ‘data’ is commonly used as a generic statement, which often leads to misunderstanding and misuse of datasets with respect to what information or insight can, or cannot, be derived. Information and insight are necessary to improve knowledge and understanding used to support and justify decision making.

Appreciation of different data types is relevant to technical data analysis and commercial decision making. *Raw* data are unprocessed statements; through analysis they are converted into *processed* data which provide insight and information. An important class of data, often overlooked, is *metadata* that describe the content of datasets and are vital in data management.

The capacity of modern data technologies (e.g. mobile devices, remote sensing, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks) to create data has given rise to the concept of ‘Big Data’. However, ‘Big Data’ is commonly defined in global terms (Volume, Velocity, Variety and Veracity) which do not have equal relevance to agriculture. If ‘Big Data’ is allowed to develop and mature beyond the current hype and into concrete ideas focused on finding meaning in data, the increasing availability of technologies to collect, analyse and visualise data is an exciting prospect.

Concern about data ownership is widely debated. The root cause is because the majority of growers is dependent on third party data service providers. The crux of the debate is how to create a viable market for the data service providers whilst protecting growers’ rights to control their proprietary data? Explicit consideration must be given to raw data, derived data and protection of technical IP. Ownership of raw data was undisputed in all the ag-data companies; it belongs to the farmer. Ownership of derived data is more ambiguous, reflecting how derived data are created. Interestingly, the whole spectrum of views on data ownership, from being completely proprietary through to openly sharing data, were expressed during my interviews.

An inescapable conclusion from my study tour is that data is an enterprise asset; the farm businesses which had embraced this concept had a more effective relationship with their data. Treating data as an asset goes far beyond the debate over ownership and into designing effective strategies for integrating the relevant people, processes and technology that enable data to deliver commercial benefit. Like any asset, underinvestment and lack of strategic vision will, inevitably, lead to ineffective implementation and probably failure to deliver intended business benefits.

Existing challenges remain. Data is inherently noisy and as an industry we are far too accepting of data. A culture which challenged the integrity of data before acceptance of the information they contain would help minimise the impact of poor data. There is considerable scope for improving industry-wide skills in analysing un-replicated observational datasets and the industry should actively include the knowledge of ‘digital immigrants’ in the application of new data technologies.

Robert Allen