Perfect Information Conference 2014
Big Data Panel
Donald Roll Preparation Notes

What is Big Data?

- Tells you What not Why
- Things you can do at a large scale that you can't do on a smaller one
- It is applying math to a huge amount of data to infer probabilities
    - Goal is to extract new insights or create new forms of value
    - Finds correlations
    - Best known example of finding correlations is Amazon recommendations - now accounts for 1/3 of their sales
- Can be Messy data - More data trumps some and sometimes More trumps Better. "Simple models with lots of data are often better than more elaborate models based on less data. Better at looking for trends and forecasts.
- Predictive Analysis - foresee events before they happen
    - UPS - monitors its fleet of 60,000 vehicles to help predict when a vehicle needs maintenance
    - Rolls Royce - monitors its customers aircraft engine and by gathering all the data from all their engines, can better predict when an aircraft engine requires servicing. This Big Data application is now 70% of Rolls Royce business.
- Move from Linear Relationships to finding non-linear relationships
    - relatively new field
    - network analysis is an example of this
    - think of a cubist painting of a face - multi-faceted

What is Driving Big Data?

- Decrease in storage costs makes it easier to keep everything.
    - No longer do you need to use samples, you can now capture much more data, allowing you to dig deeper into the data.

- Increase in computer power
- Datafication - everything is now in digital format making it easier to record, analyze and re-organise data
    - Due to smart devises and GPS, lots of new types of data are now available about people and what they are doing allowing for new types of data capture which can then be analysed for insight
    - Barnes & Noble Nook - never before have we been able to capture how people read books - discovered many non-fiction books only read half way through.
    - Car traffic using auto manufacturer GPS data, can provide insight if a retailer is increasing or decreasing business based on traffic flow near the store.
- Digitization turbo charges Datafication
    - Google used its Google Book index to improve Google Translate by looking at books available in several languages and then looking for correlations between them. Did this over thousands of books. Makes it a big math problem, not a text problem.

Who has the Big Data?

- Google - one of the leaders in Big Data
- Amazon - a leader but uses it internally, does not make it available externally.
- Facebook
    - Gartner - by 2011, it had collected over 2.1 trillion pieces of "monetizable content" such as likes, posted material and comments.
    - Has not attempted to commercialise it yet
- Twitter - has a fire hose of comments it resells
- Governments
    - Open Data Institute promotes access to government data in the UK
    - data.gov in the US went from 47 datasets in 2009 to 450,000 datasets in 2012.

- Cell phone providers - Telefonica Digital Insights sells anonymous and aggregated subscriber location data.
- GPS providers- car movements
- Credit Card Providers - Visa and MC
- Companies
    - smart sensors on goods to track performance
    - tracking buying patterns - Supermarkets, Wal-Mart, Target, etc using loyalty programmes

New types of Data Aggregators

- Companies like Inrix who license GPS data from auto manufacturers - having more

Data Skills

- Terradata - data analytics firm
- Google
- New companies like Kaggle "solving business problems through predictive analytics"
- Rise of the Data Scientist - combines the skills of a statistician, software programmer, info graphics designer and story teller.
- Statisticians and machine learning people

Issues of Big Data

- Data Privacy
- Data Ownership and Data Governance
- Demise of Subject Matter Experts?
    - experience is less important as you can more easily build expert systems

Examples of Big Data

- Spam Filters are an example of big data

- Google Translate is an example of big data - treated translation as a math problem.
- Farecast - 2007, an early example
    - monitors flight prices being offered on the internet.
    - Collected over 200 billion price quotes
    - Predicts price movements for US flights
    - Makes the correct call over 75% of the time, saving users $40 per flight.
- Google Flu Trends
    - Takes searches related to flu
    - Then compared them to government statistics
    - Found it is a very good leading indicator to alert for new flu outbreaks.
- Pricestats
    - Gathers on-line prices from hundreds of retailers for millions of products in over 70 countries.
    - Better and faster indicator of inflation trends
    - Because of the vast amount of data collected, can also do much more deep dive analysis on different types of product inflation.
- Target - US retailer
    - Using Baby Gift Register data they looked at what people purchased in the 9 months before the due date
        - 3rd month - started buying unscented lotion
        - 4th month - started buying vitamin supplements like magnesium, calcium and zinc
    - Gathering all this info allowed them to create a pregnancy prediction score and then use the score for targeted marketing
- Airsage
    - captures 15 billion geo location data points a day from cell phone providers
        - can be used for real time traffic reports
        - Look at road use, where is more capacity required?
- Twitter - sells the fire house of tweets to third parties
    - Used for sentiment analysis for tweets on companies and products to drive trading models

Bibliography

- Big Data, A Revolution That Will Transform How We Live, Work and Think by Victor Mayer-Schonberger and Kenneth Cukier
- Various Economist and FT articles