Lecture 1: Intro to Applied Data Science

ADS/Jan 2017/Raul Santos-Rodriguez

Teaching staff

Tom Diethe
tom.diethe@bristol.ac.uk

Daniel Schien
Daniel.Schien@bristol.ac.uk

Peter Flach
cspaf@bristol.ac.uk

Niall Twomey
niall.twomey@bristol.ac.uk

Raul Santos-Rodriguez
enrsr@bristol.ac.uk

Prerequisites and objectives

Before you take the class, you should have some knowledge of ...

- Programming
- Machine Learning / Data Mining

At the end of this course, you should...

- Understand different Data Science techniques
- Be able to tackle real-world tasks with the appropriate Data Science tools
- Be more proficient at presenting and interpreting data to/for a (non-)technical audience
- Have practised teamwork and time management.

Coursework

Project: 90%

- Groups of 5.
- Report by the end of term.

Questionnaires: 10%

- Up to 5 forms.

Lectures: Wednesday, 10:00-12:00. Friday, 12:00-13:00 (Q&A).

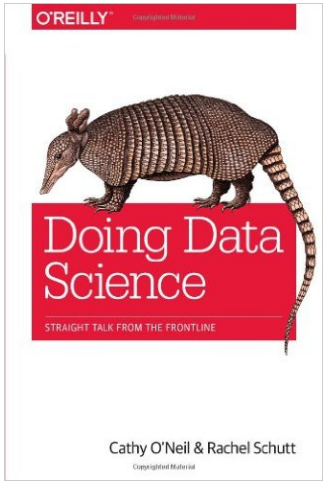
- Ask questions and give feedback!

References

Mining of Massive Datasets, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2011.
Principles of Data Mining, David J. Hand, Heikki Mannila and Padhraic Smyth, MIT Press, 2001.
Information Visualization, Colin Ware, Morgan Kaufmann, 2012.
The Visual Display of Quantitative Information, Edward Tufte, 2001.

Additional reading material in the form of research papers, blogs, articles, etc.

Have a look at ...



... Doing Data Science, Cathy O'Neil and Rachel Schutt (Ch. 1 and 2)

... Data Science: An Introduction, wikibooks

... Kdnuggets

... Kaggle

... Data Science Central

Outline

We will discuss:

- Why learn Data Science?
- What will you learn?

`https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century`

Data

- Numerical, categorical, or binary
- Text: emails, tweets, articles
- Records: user-level data, timestamped event data, log files
- Geo-based location data
- Network data
- Sensor data
- Images and video
- Audio and music

Big Data

The numbers:

- **48** - The hours of video uploaded to YouTube every minute, resulting in nearly 8 years of content every day.
- **7 Million** - The numbers of DVDs internet traffic information would fill EVERY hour. Side by side, theyd scale Mount Everest 95 times.
- **3 Billion** - The number of people who will be online in 2015, generating 8 zettabytes of data. (One zettabyte equals one sextillion bytes- thats twenty-one zeros!)
- **30 Billion** - Pieces of content shared on Facebook every day.
- **247 Billion** - The number of e-mail messages sent each day – up to 80% are spam.
- **90%** - Percentage of the worlds data created in the last 2 years.

Big Data

The numbers:

- Library of Congress **text** database of around 20 TB.
- Thirteen million **photographs**, even if compressed to a 1 MB JPG each, would be 13 TB.
- AT&T 323 TB, 1.9 trillion **phone call records**.
- 3.5 million **sound recordings**, which at one audio CD each, would be almost 2,000 TB.
- World of Warcraft utilizes 1.3 PB of storage to maintain its **game**.
- Avatar **movie** reported to have taken over 1 PB of local storage at Weta Digital for the rendering of the 3D CGI effects.
- **Google** processes 24 PB of data per day.
- **YouTube**: More video is uploaded in 60 days than all 3 major US networks created in 60 years. According to cisco, internet video will generate over 18 EB.

Big Data and Small Data

What is large?

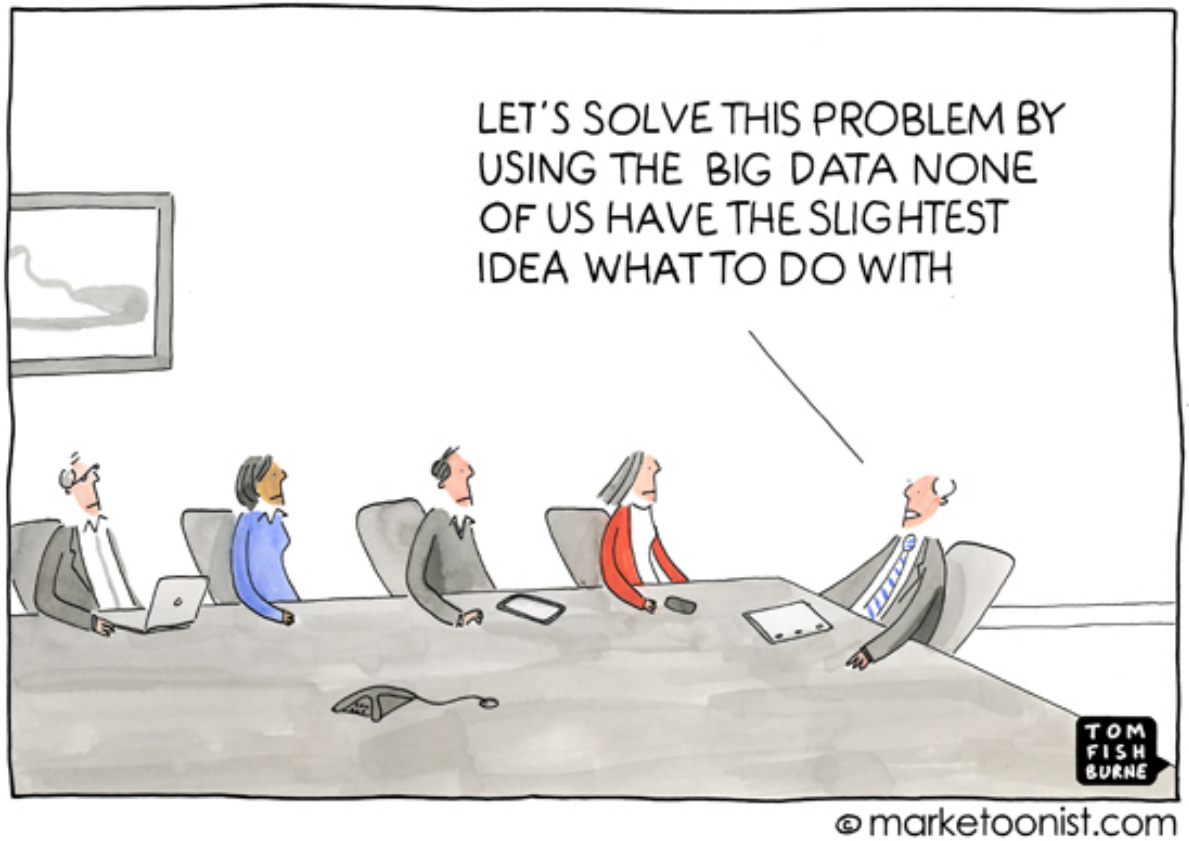
Large text dataset:

1,000,000 words in 1967
1,000,000,000,000 words in 2006

	Big Data	Small Data
Data Condition	Always unstructured, not ready for analysis, many relational database tables that need merged	Ready for analysis, flat file, no need for merging tables.
Location	Cloud, Offshore, SQL Server, etc.	Database, local PC
Data Size	Over 50K Variables, over 50K individuals, random samples, unstructured	File that is in a spreadsheet, that can be viewed on a few sheets of paper
Data Purpose	No intended purpose	Intended purpose for Data Collection

https://www.bbvaopenmind.com/en/small-data-vs-big-data-back-to-the-basics/

Data science



https://www.youtube.com/watch?v=dKHz9LbgRmo
https://www.youtube.com/watch?v=htNN-RtFb1Q

What is Data Science?

"Data science, also known as data-driven science, is an interdisciplinary field about scientific processes and systems to extract knowledge or insights from data in various forms." (Wikipedia)

"Data science is an advanced discipline, requiring proficiency in parallel processing, map-reduce computing, petabyte-sized noSQL databases, machine learning, advanced statistics and complexity science." (Data Science: An Introduction)

"Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies." (TechTarget)

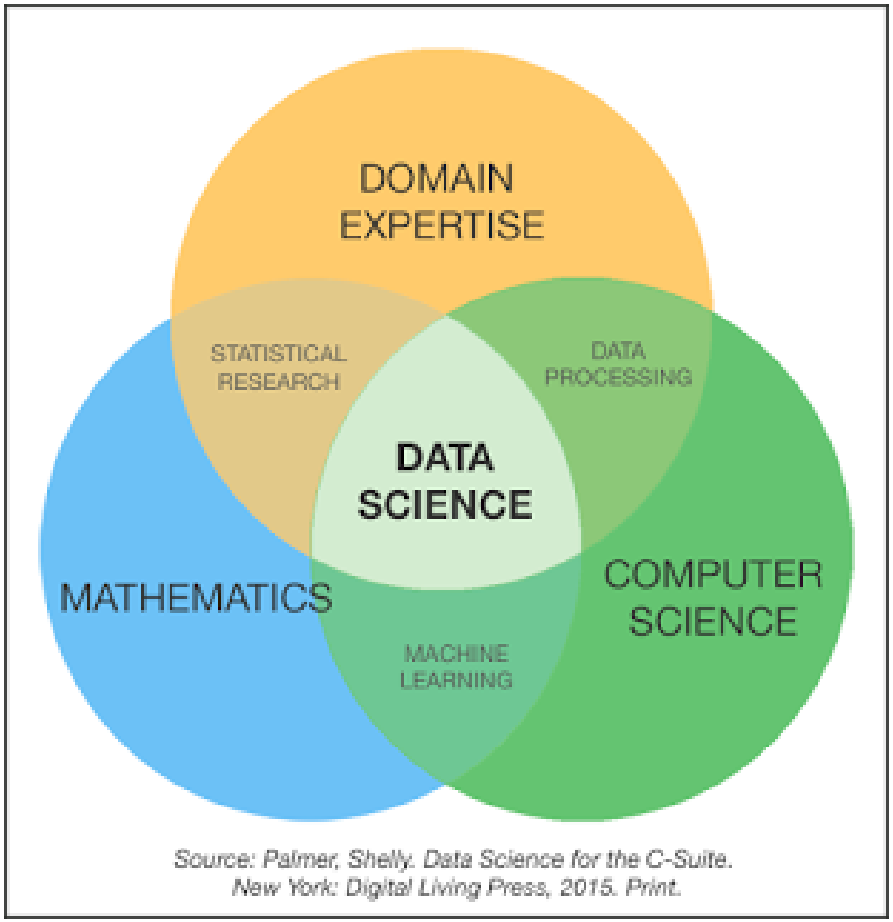
"Data Science: An action plan to expand the field of statistics." (William Cleveland, 2001)

"Data science, as it's practiced, is a blend of Red-Bull-fuelled hacking and espresso-inspired statistics. [...] Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible." (Mike Driscoll)

"Data science is an act of interpretation." (Riley Newman)

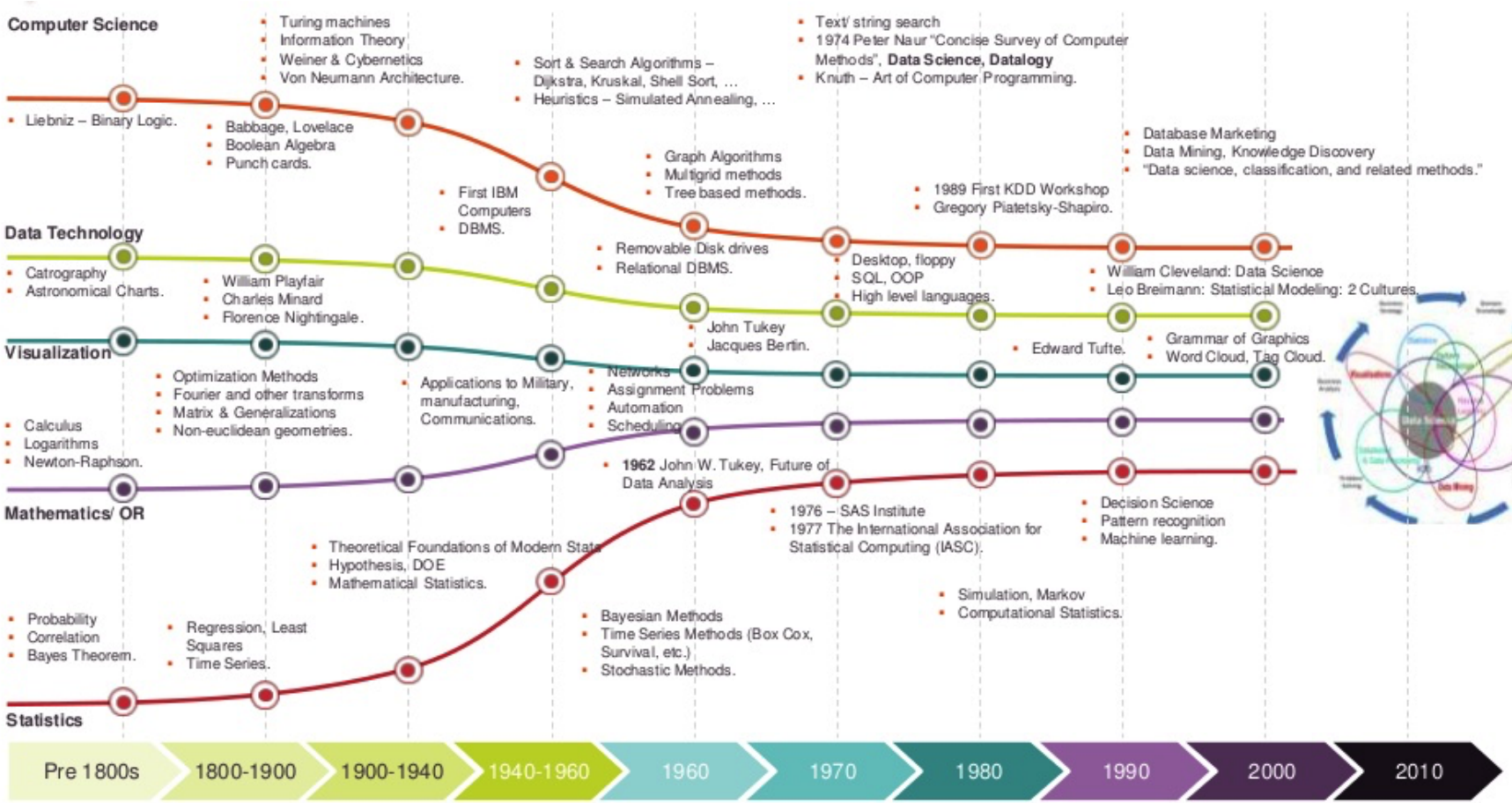
"There is no such thing as data science." (Robin Bloor)

What is Data Science?



http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html

A bit of history

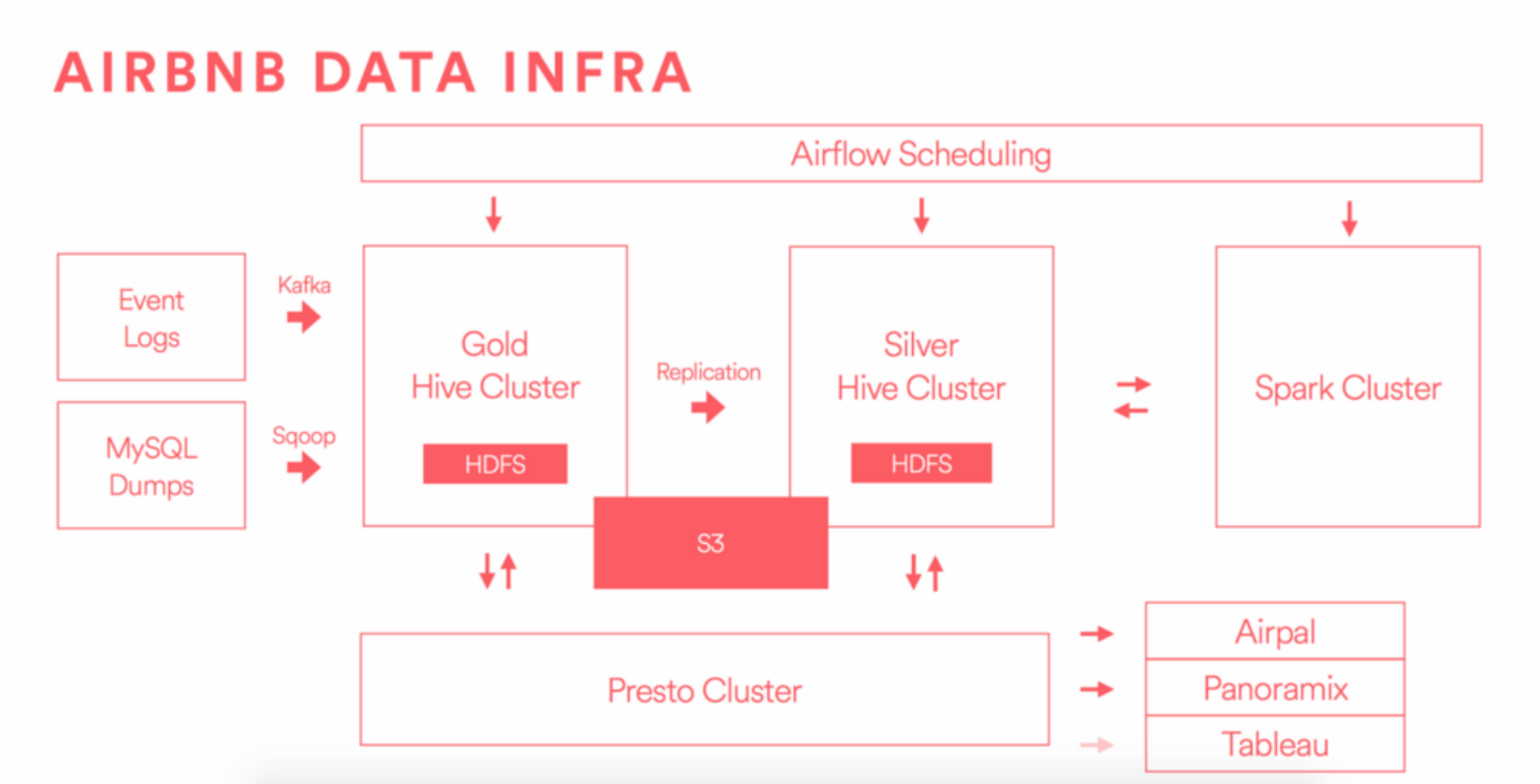


Impact of Big Data on analytics, M. Upadhyaya

An example

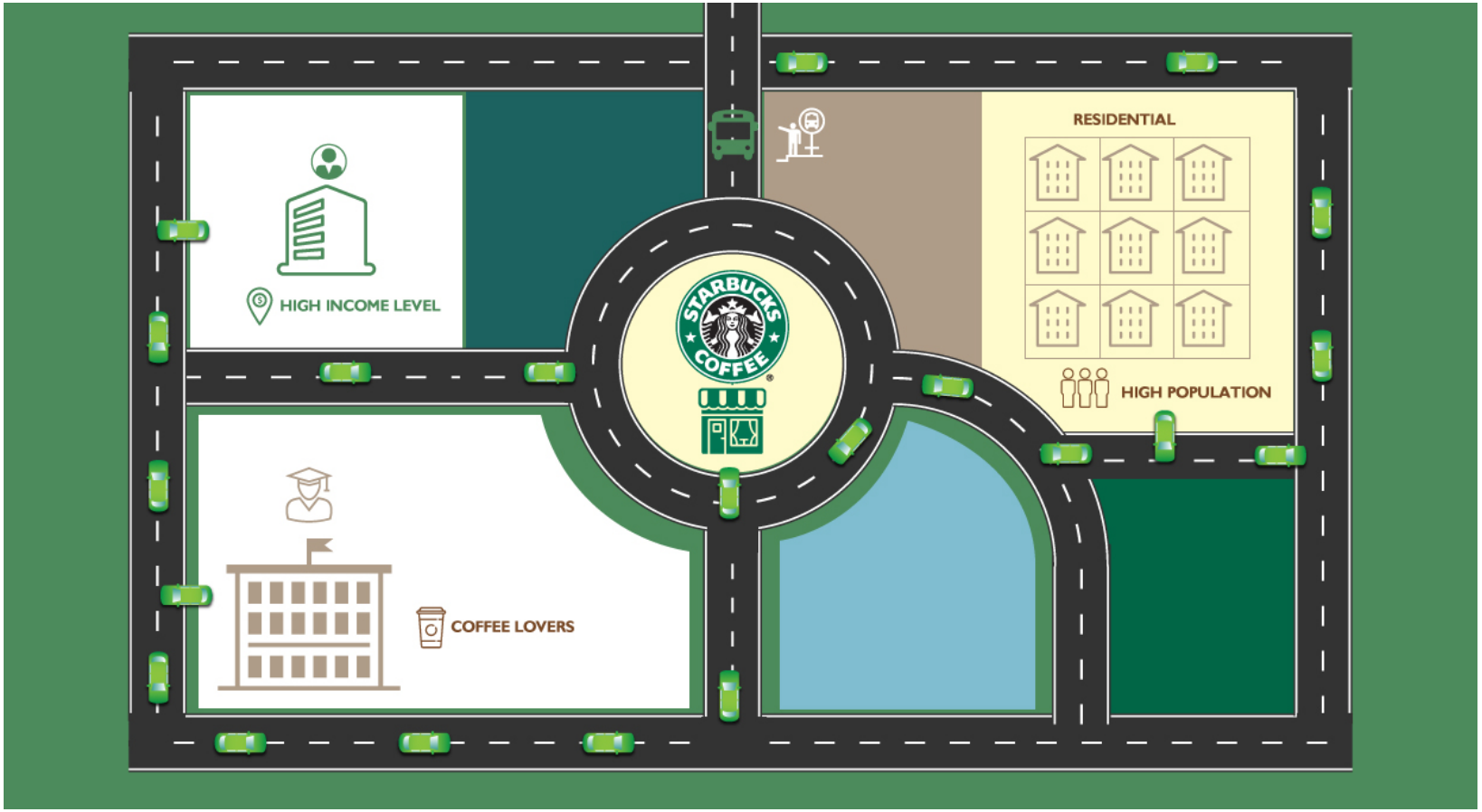
http://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/http://nerds.airbnb.com/

An example



https://medium.com/airbnb-engineering/data-Infrastructure-at-airbnb-8adfb34f169c#.18unc3j1q

Food

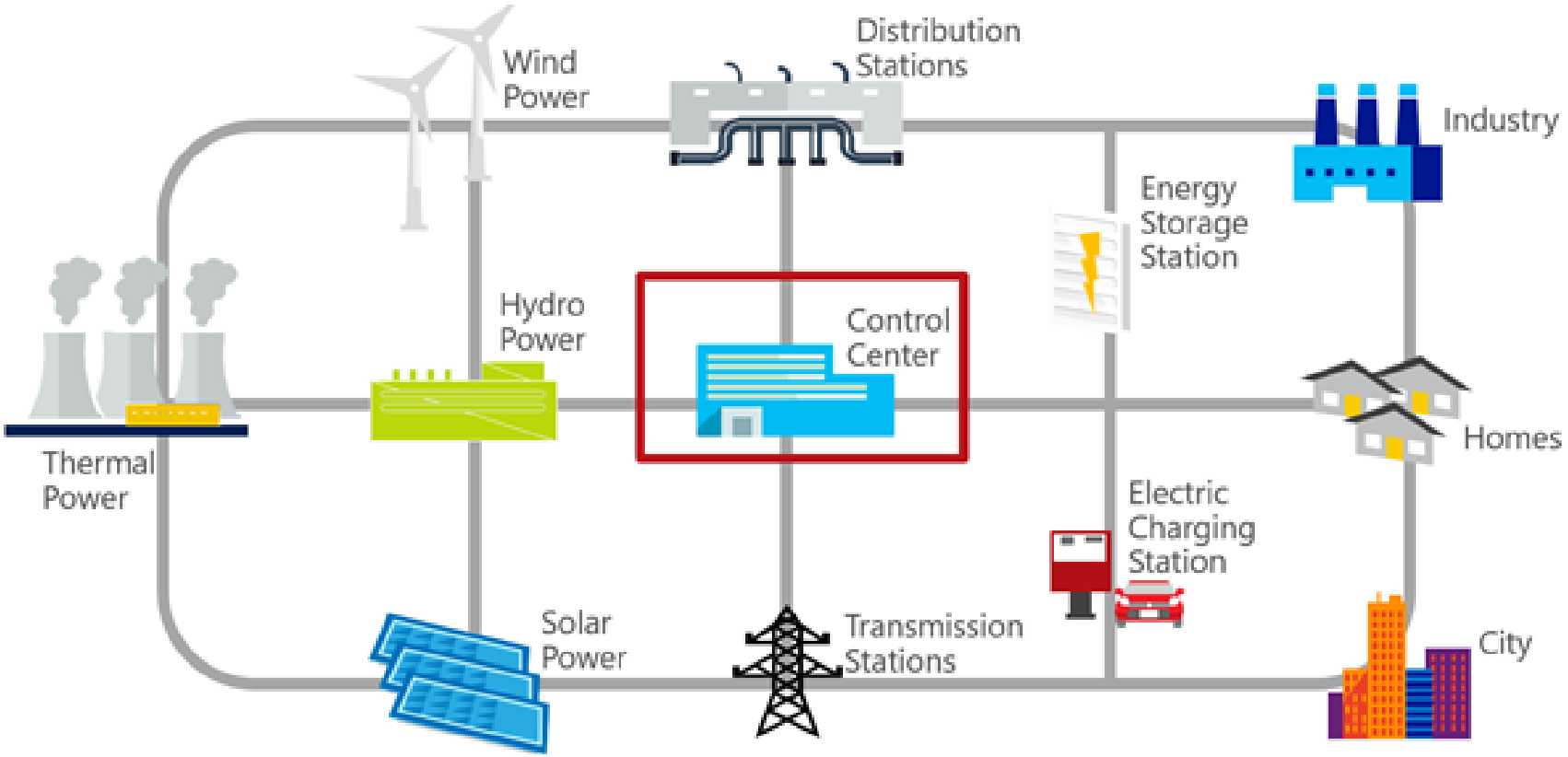


https://www.linkedin.com/pulse/starbucks-roasting-data-brewing-analytics-nigrah-bamb

Retail


amazon.com
Hello. Sign in to get personalized recommendations. New customer? Start here.
Your Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards
Shop All Departments
Search
Movies & TV
Movies & TV
Advanced Search
Browse Genres
New Releases
Bestsellers
DVD & Blu-Ray Deals
Jon and Kate Plus Ei8ht: The Complete Season 4 (6 DVD Set)
Starring: Jonathan Gosselin, Kate Gosselin Director: Jennifer Stocks Format: DVD
List Price: $49.98
Price: $29.99 & this item ships for FREE with Super Saver Shipping. Details
You Save: $19.99 (40%)
In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.
Customers Who Bought This Item Also Bought
Hanging Ceiling Fan
12" Footstool
Noose
(1) $94.99
(10) $36.99
(2) $19.99

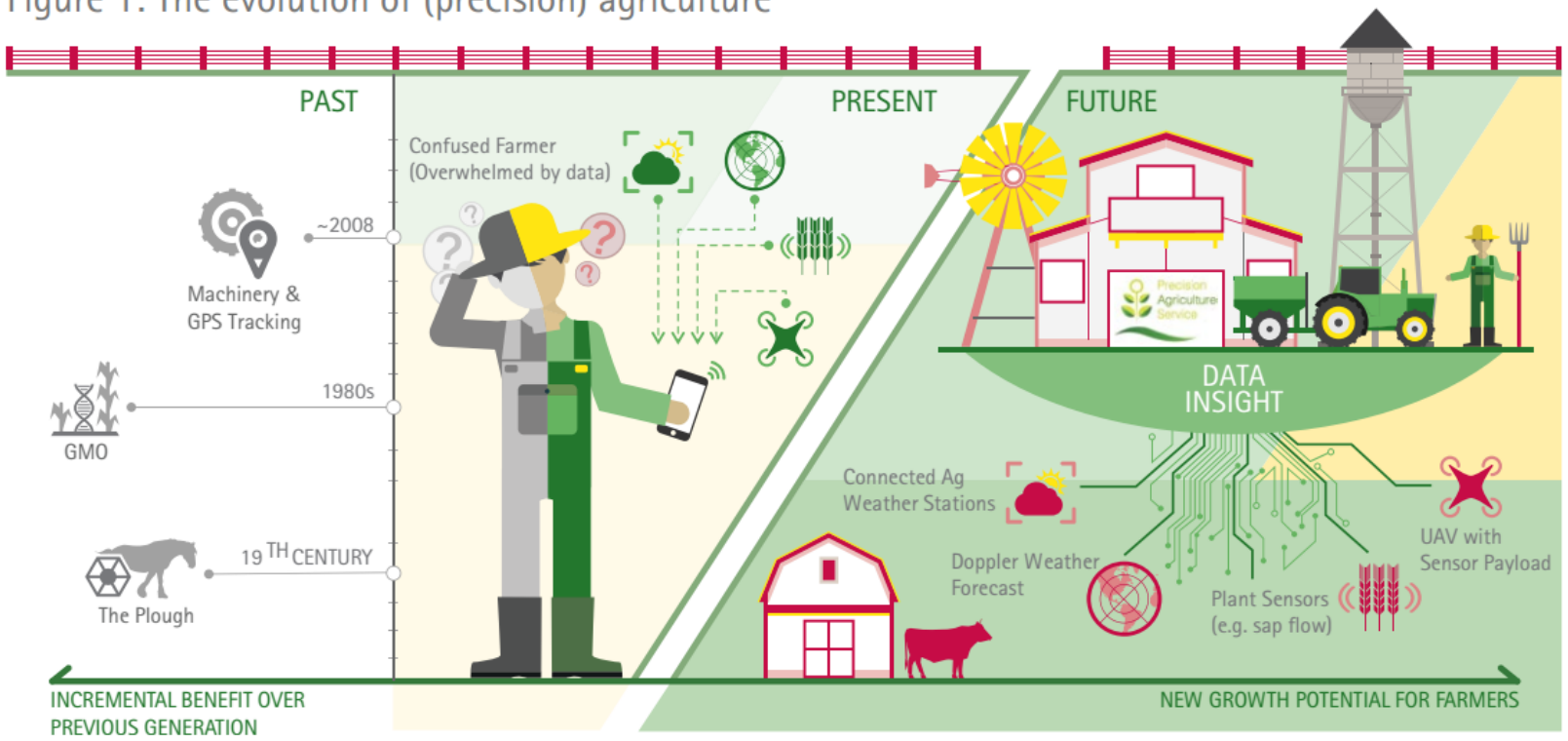
Energy and Logistics



Cortana Intelligence and ML Blog Team

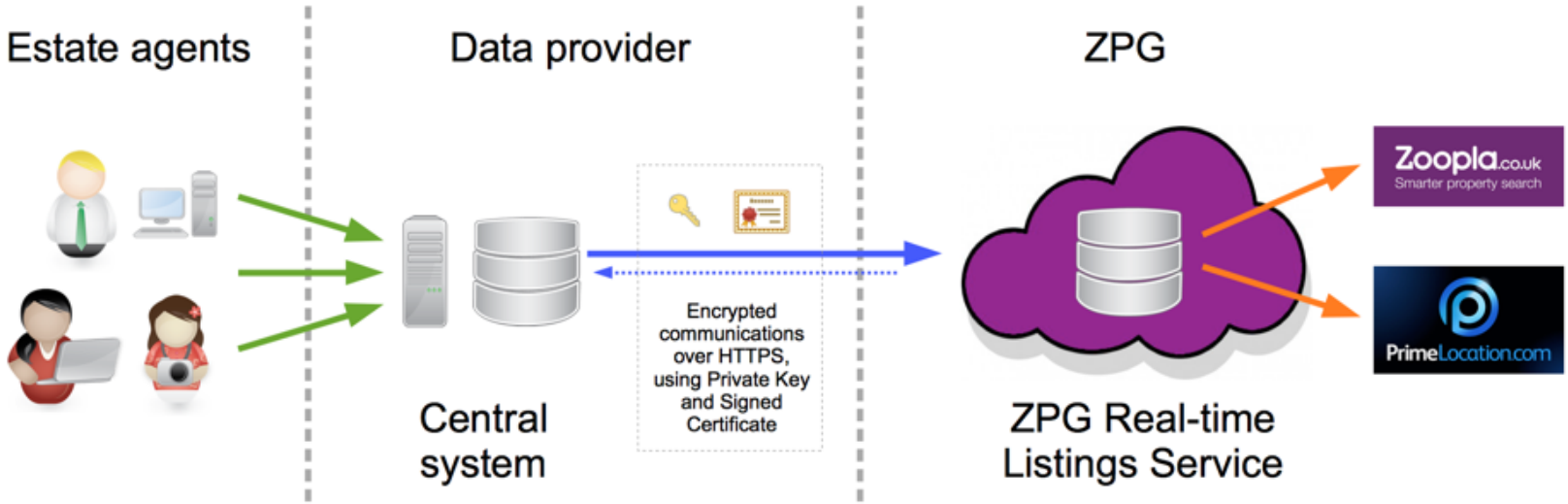
Agriculture

Figure 1: The evolution of (precision) agriculture



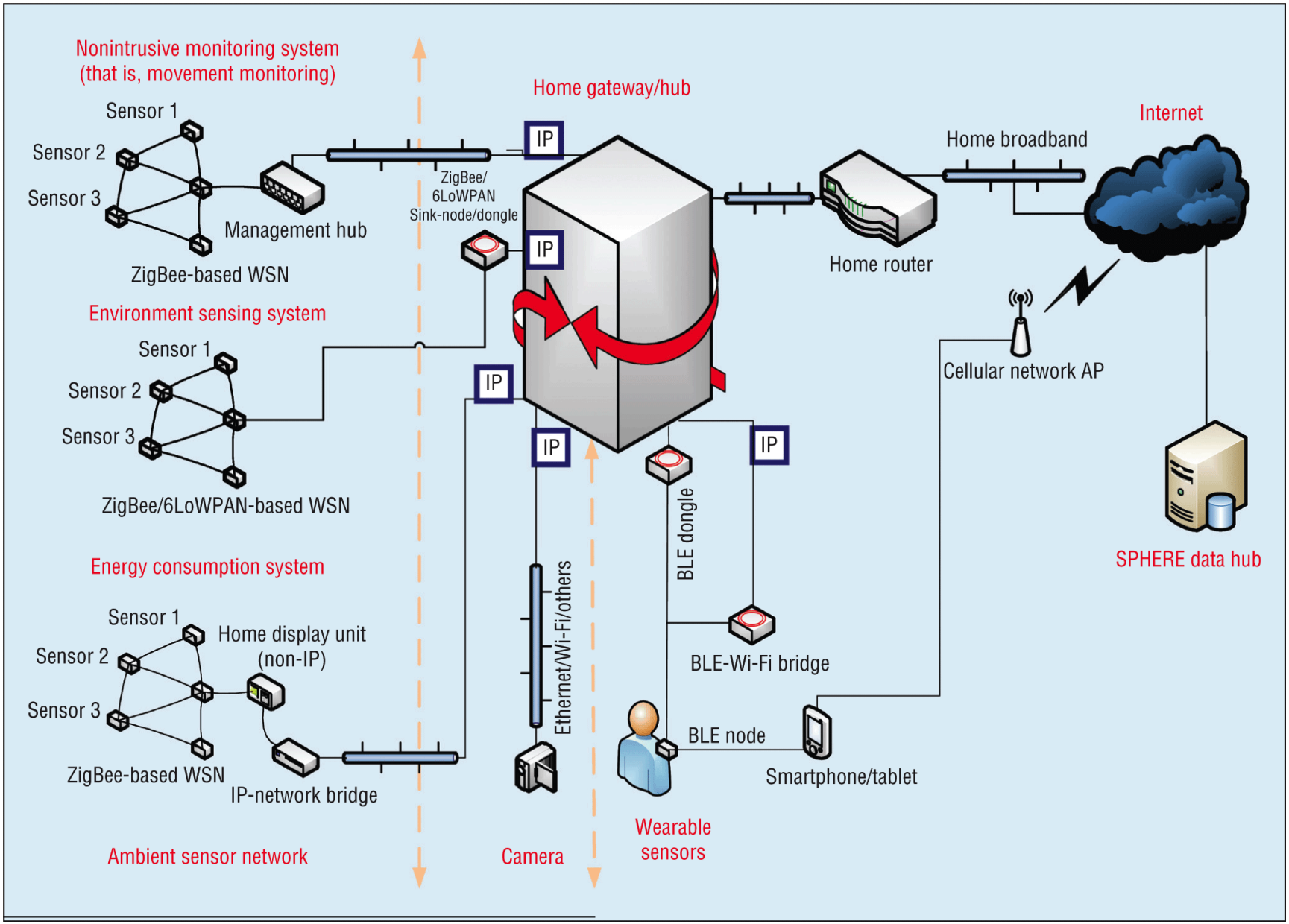
http://www.forbes.com/sites/kurtmarko/2015/08/25/precision-ag-cloud/2/

Housing



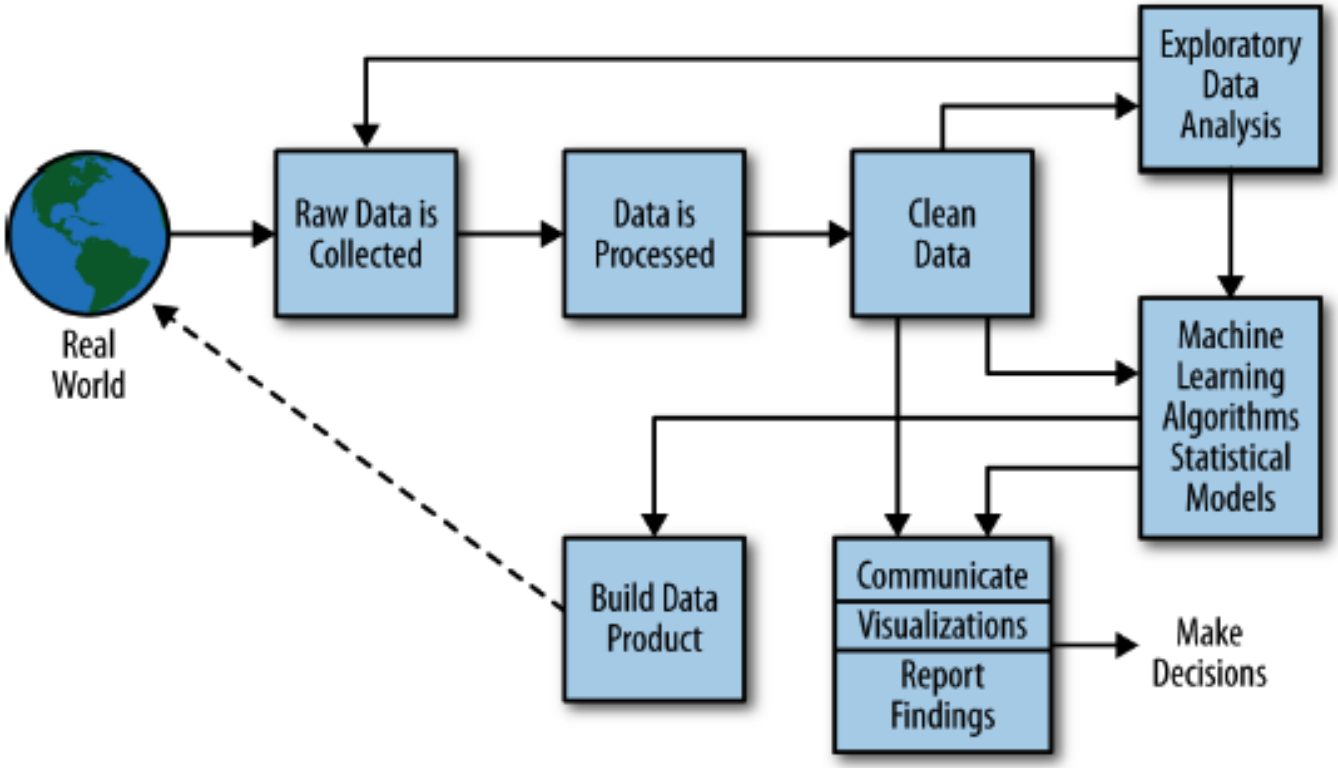
`ZooplaPropertyGroup`

Healthcare



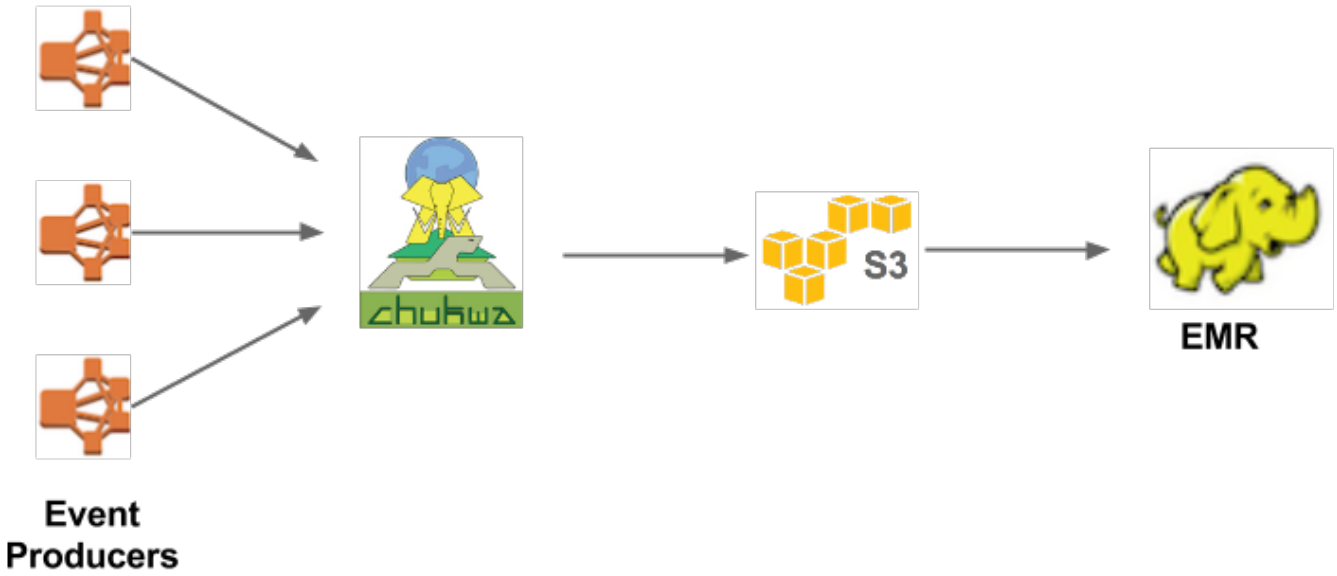
http://www.irc-sphere.ac.uk/

How do we tackle these tasks



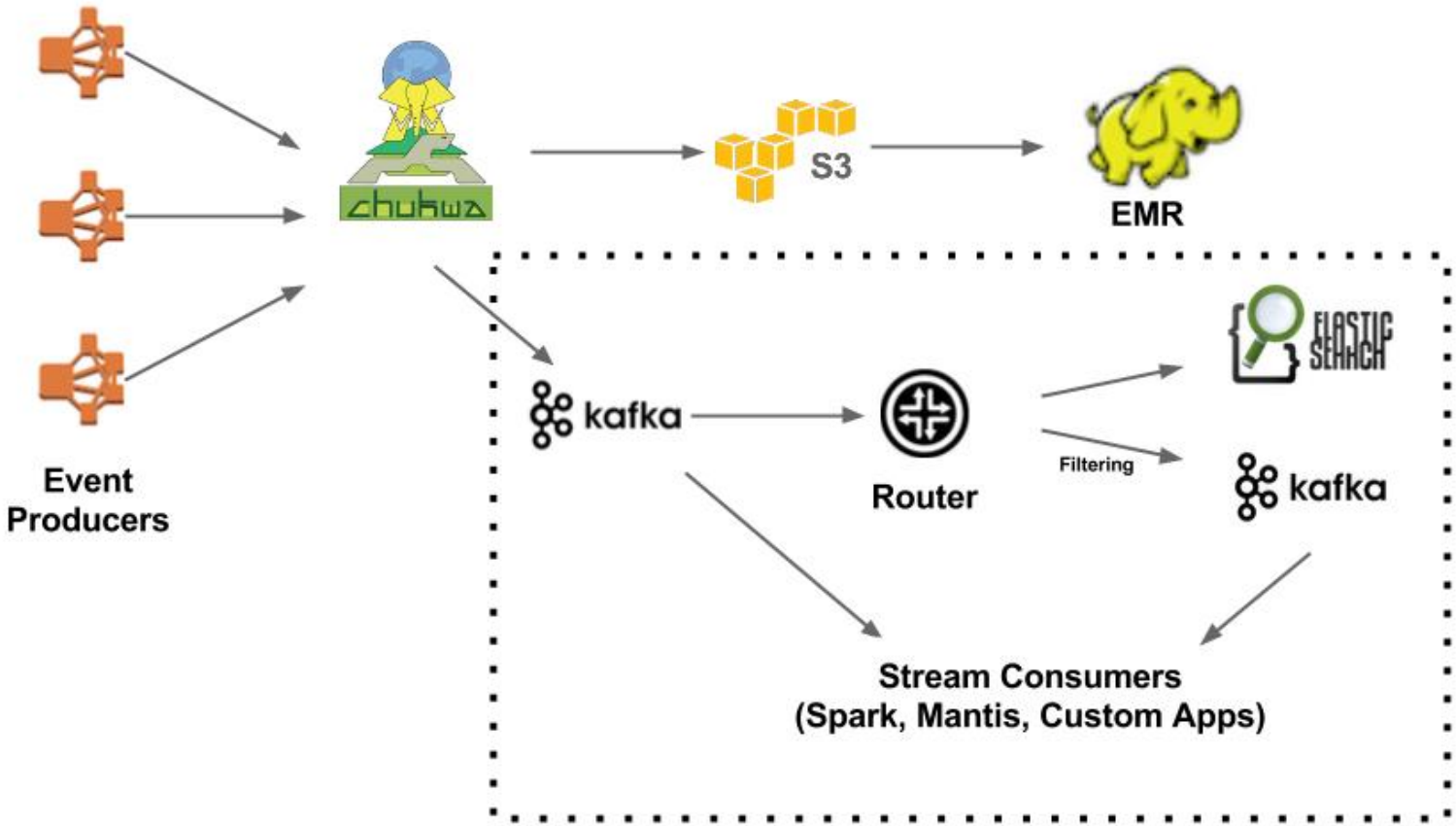
Doing data science, Cathy O'Neil and Rachel Schutt

Evolving



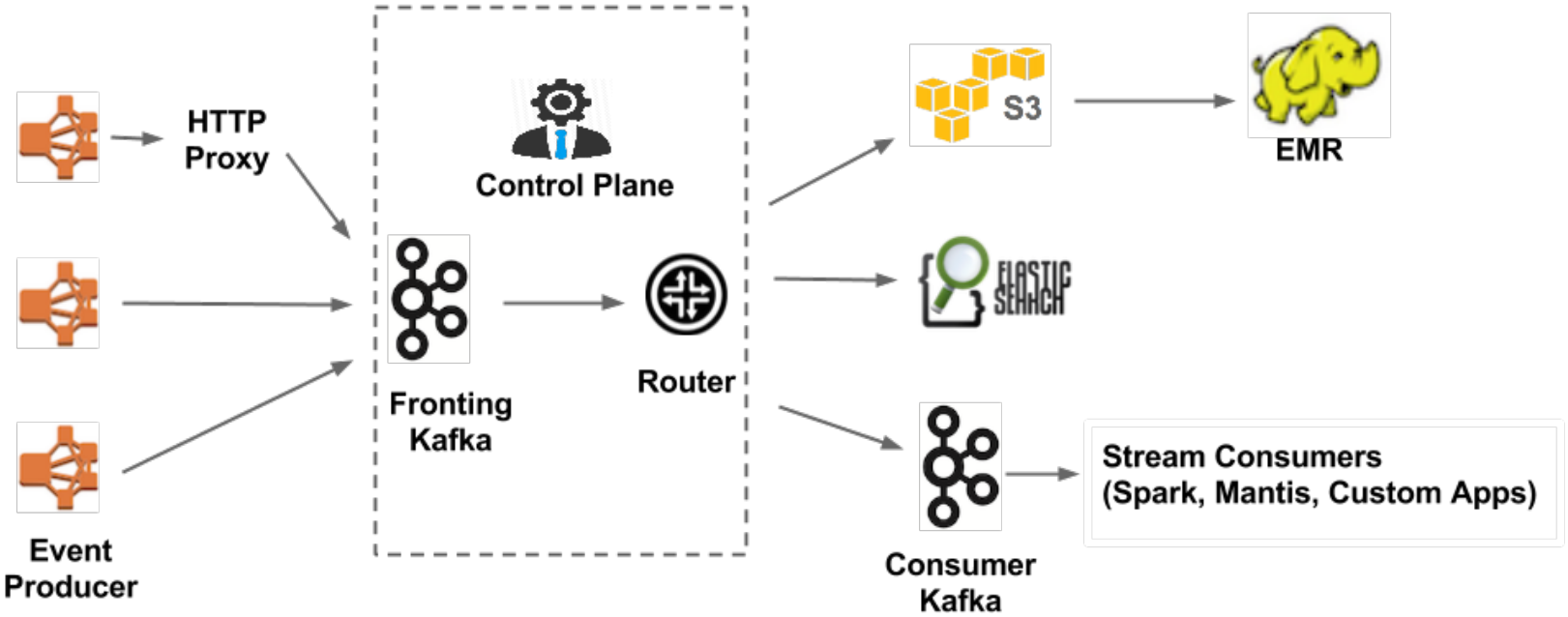
http://techblog.netflix.com/2016/02/evolution-of-netflix-data-pipeline.html

Evolving



http://techblog.netflix.com/2016/02/evolution-of-netflix-data-pipeline.html

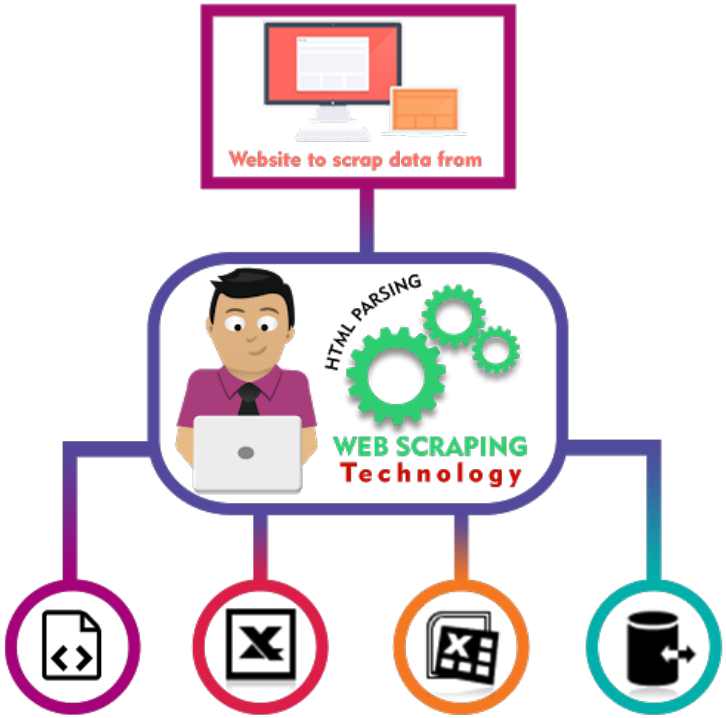
Evolving



http://techblog.netflix.com/2016/02/evolution-of-netflix-data-pipeline.html

Roadmap

- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Infrastructure
- Sharing & Privacy



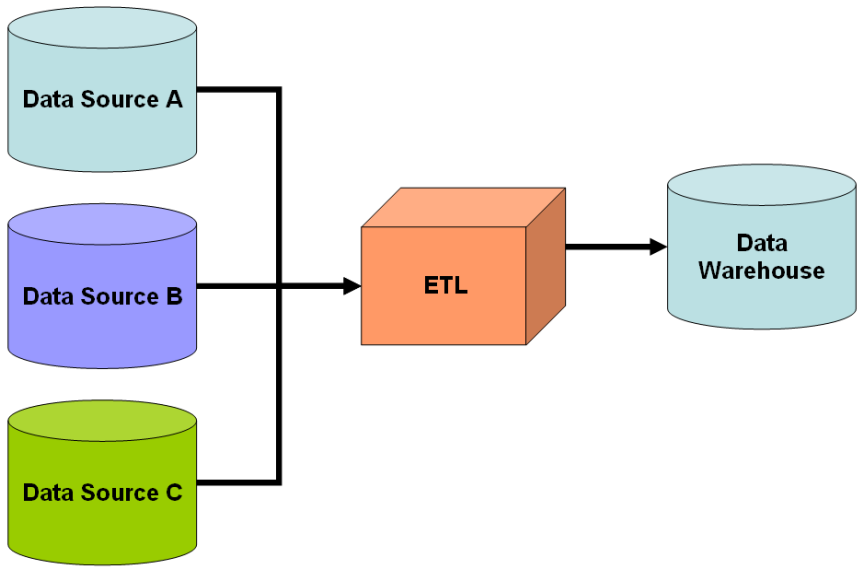
https://medium.com/@TheTopWeb/web-scrape-data-from-any-website-2dad9c332070#.g0pmetg90

Roadmap

- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Infrastructure
- Sharing & Privacy

Roadmap

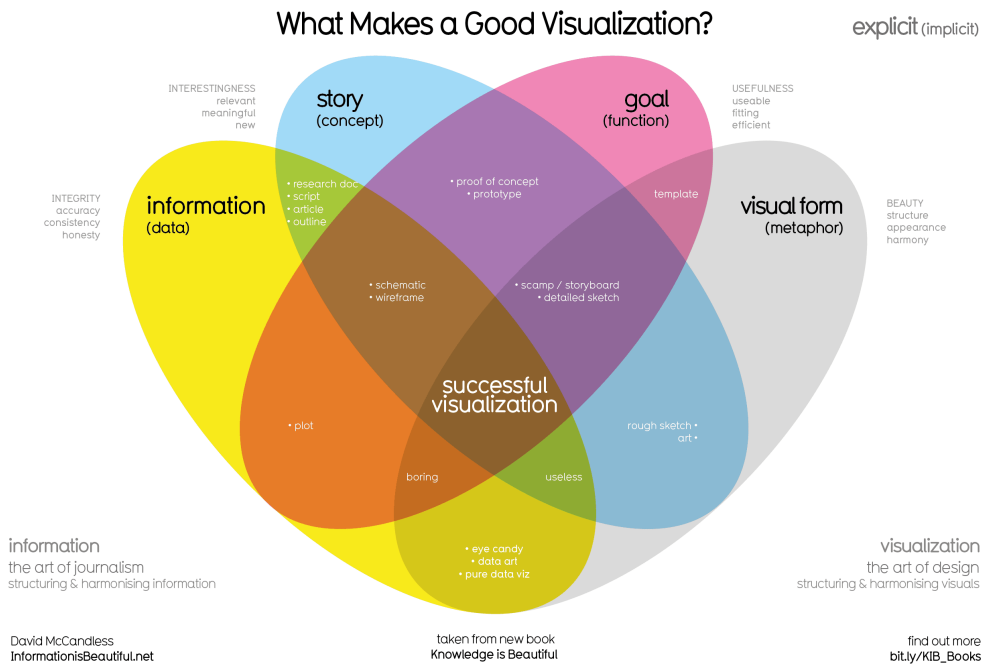
- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Infrastructure
- Sharing & Privacy



Wikipedia

Roadmap

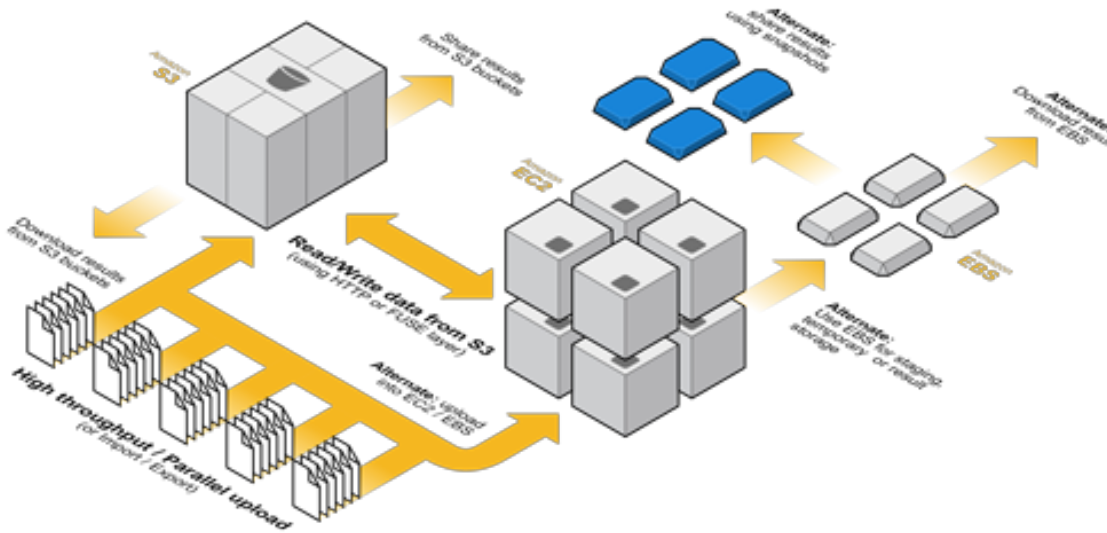
- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Infrastructure
- Sharing & Privacy



`http://www.informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization/`

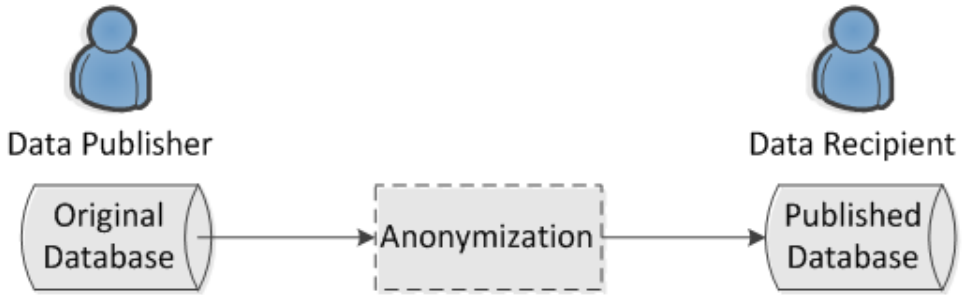
Roadmap

- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Infrastructure
- Sharing & Privacy



Roadmap

- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Infrastructure
- Sharing & Privacy



---

https://www.ericsson.com/research-blog/data-knowledge/preserving-privacy-big-data-world/

Summary so far

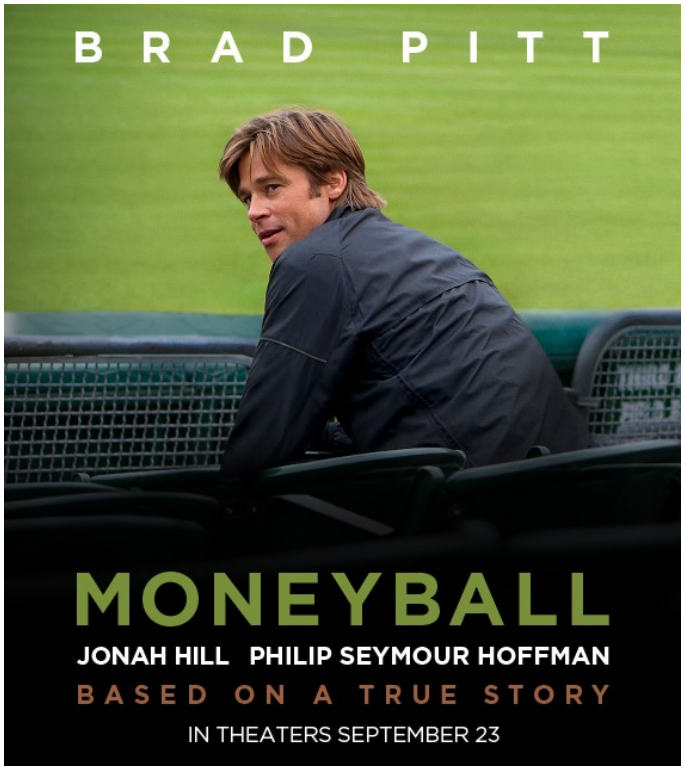
Applications of Data Science: high-impact, diverse

Challenges: computational/information complexity

Course plan

Next lecture

We will have a look at Data Ingress and Preprocessing!



`http://www.bbc.co.uk/programmes/b07lk6tj`