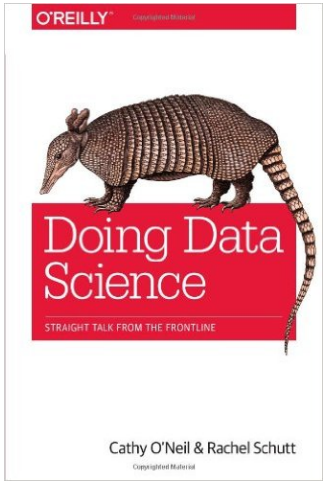
Introduction to Applied Data Science

ADS/Jan 2021/Raul Santos-Rodriguez

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 were 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 world's 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/

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)

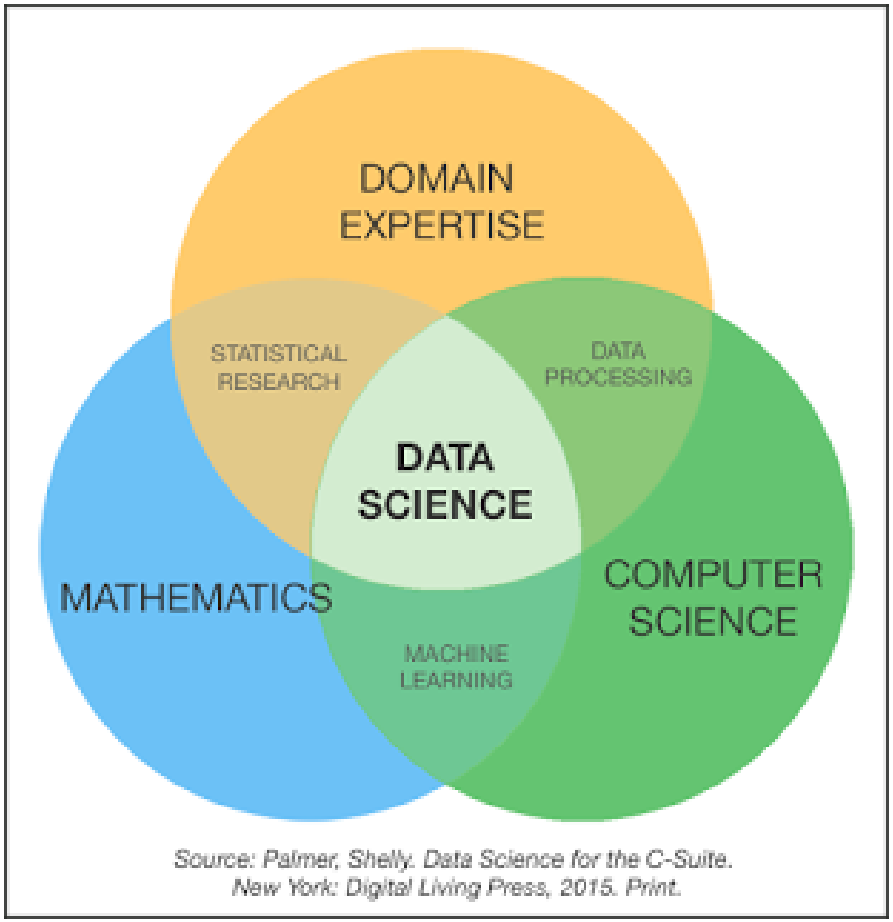
"Data Science: An action plan to expand the field of statistics." (William Cleveland, 2001)

"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)

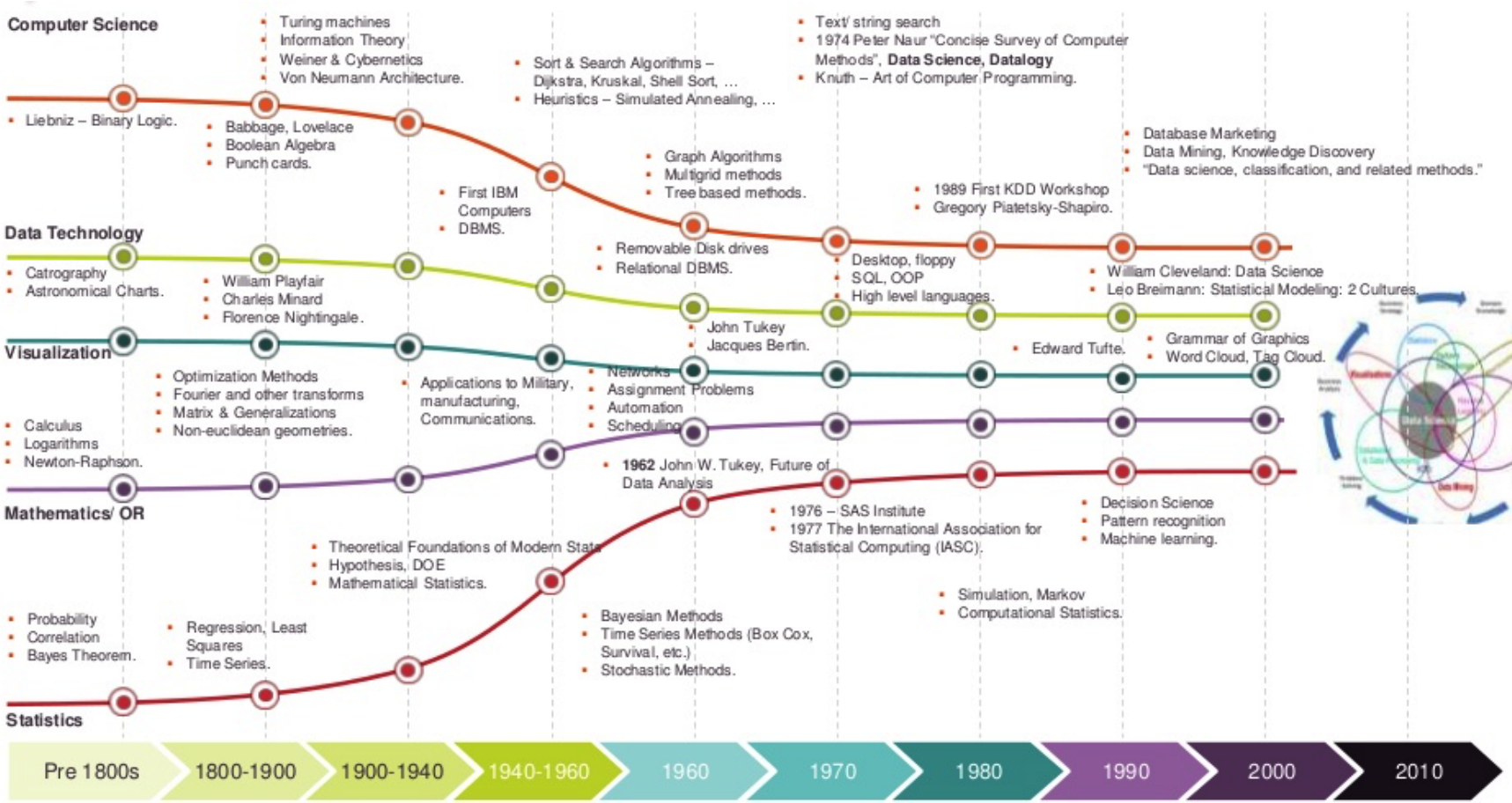
"There is no such thing as data science." (Robin Bloor)

What is Data Science?



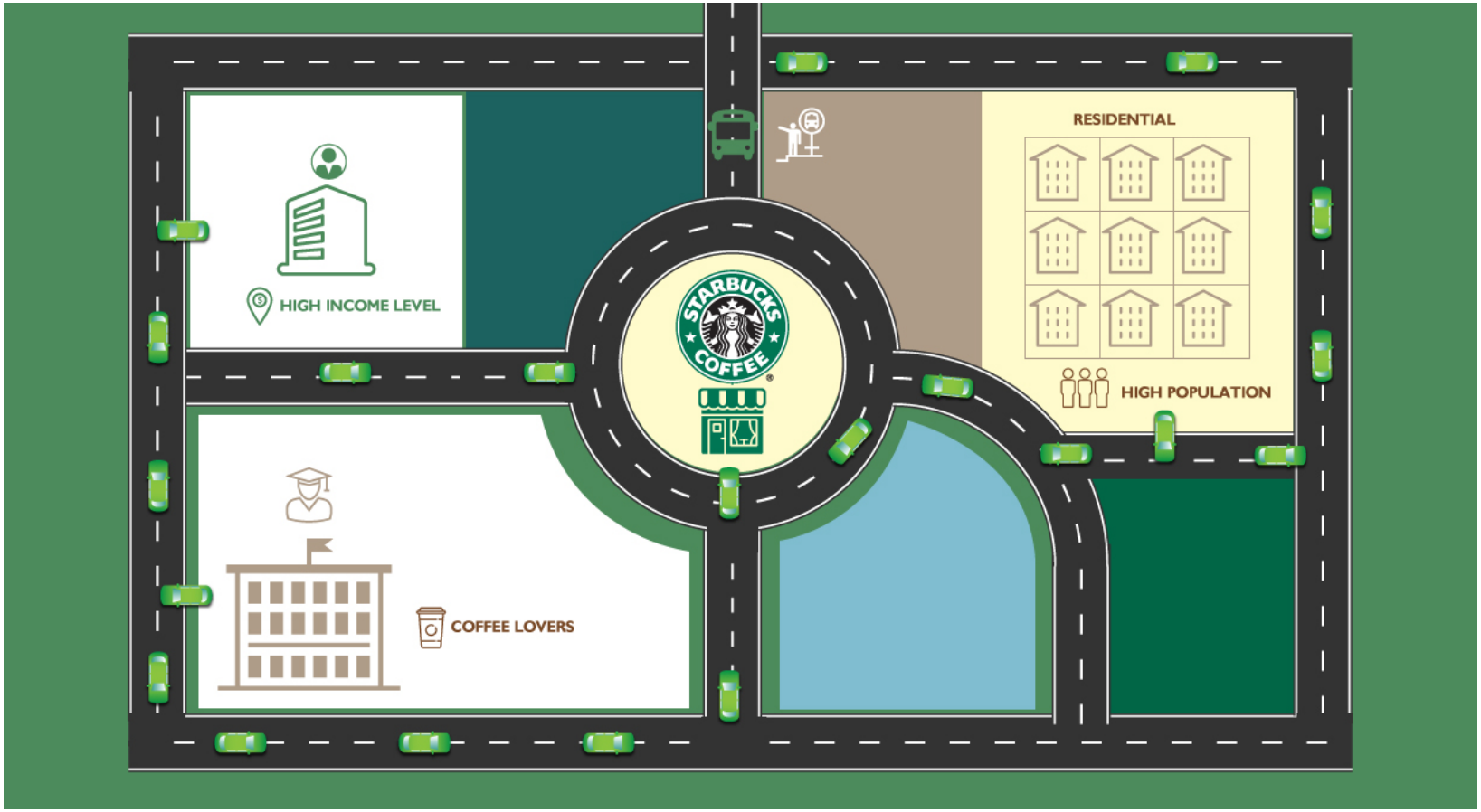
http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html

A bit of history



Impact of Big Data on analytics, M. Upadhyaya

Food



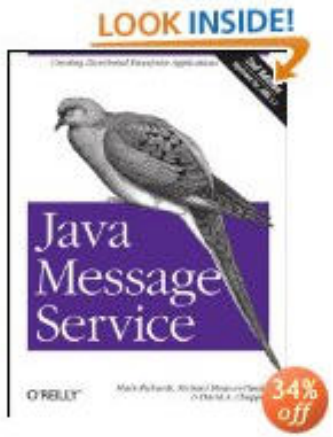
https://www.linkedin.com/pulse/starbucks-roasting-data-brewing-analytics-nigrah-bamb

Retail

amazon.com

Dear Amazon.com Customer,

As someone who has purchased or rated *Successful Affiliate Marketing for Merchants (Que-Consumer-Other)* by Shawn Collins, you might like to know that *Java Message Service* will be released on June 4, 2009. You can pre-order yours at a savings of $13.60 by following the link below.



Java Message Service
Mark Richards
List Price: ~~$39.99~~
Price: $26.39
You Save: $13.60 (34%)

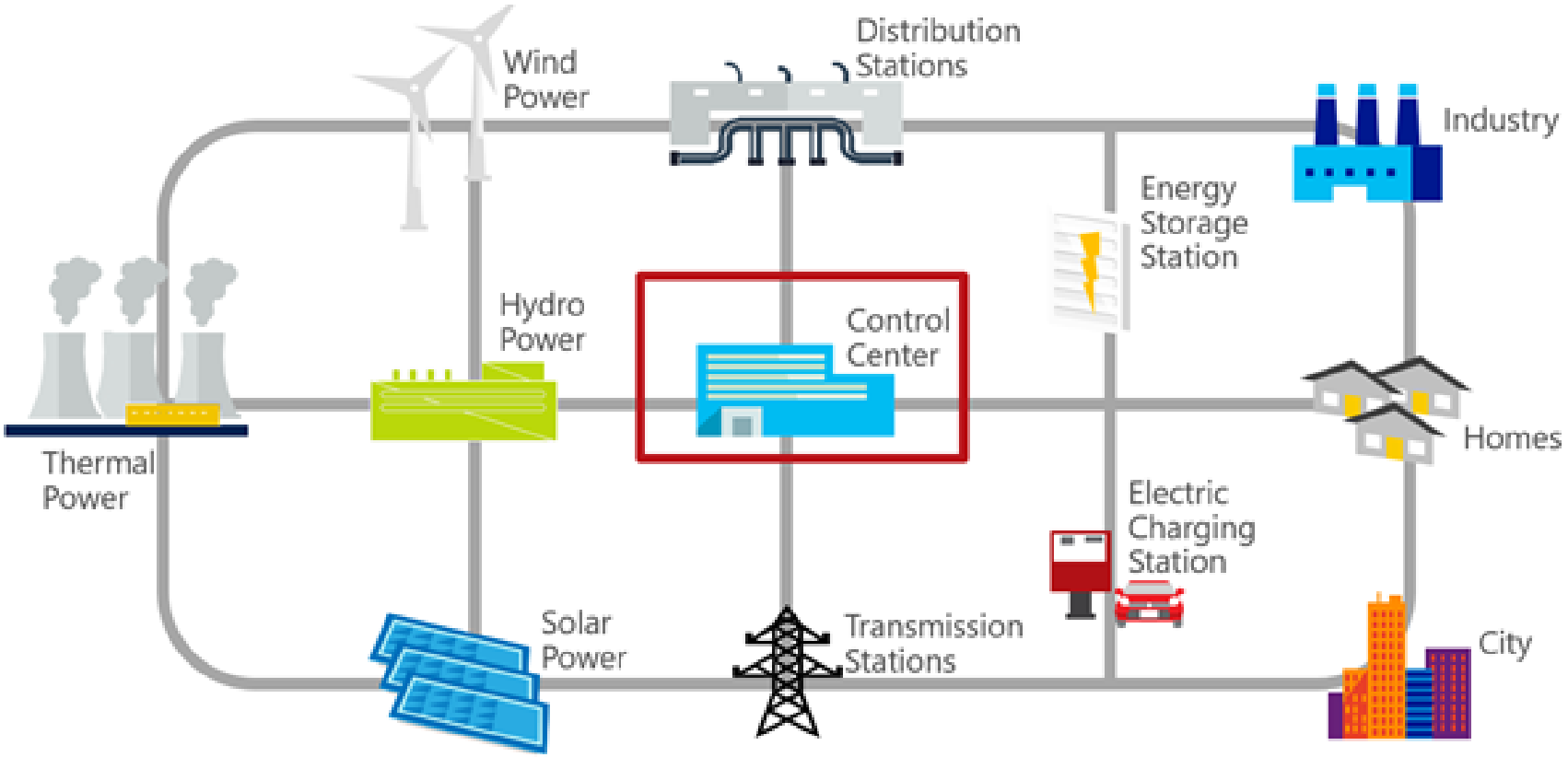
Release Date: June 4, 2009

Other Versions and Languages
Kindle Edition (Kindle Book)
Paperback

Pre-order this item today

https://www.flickr.com/photos/affiliate/3575395691

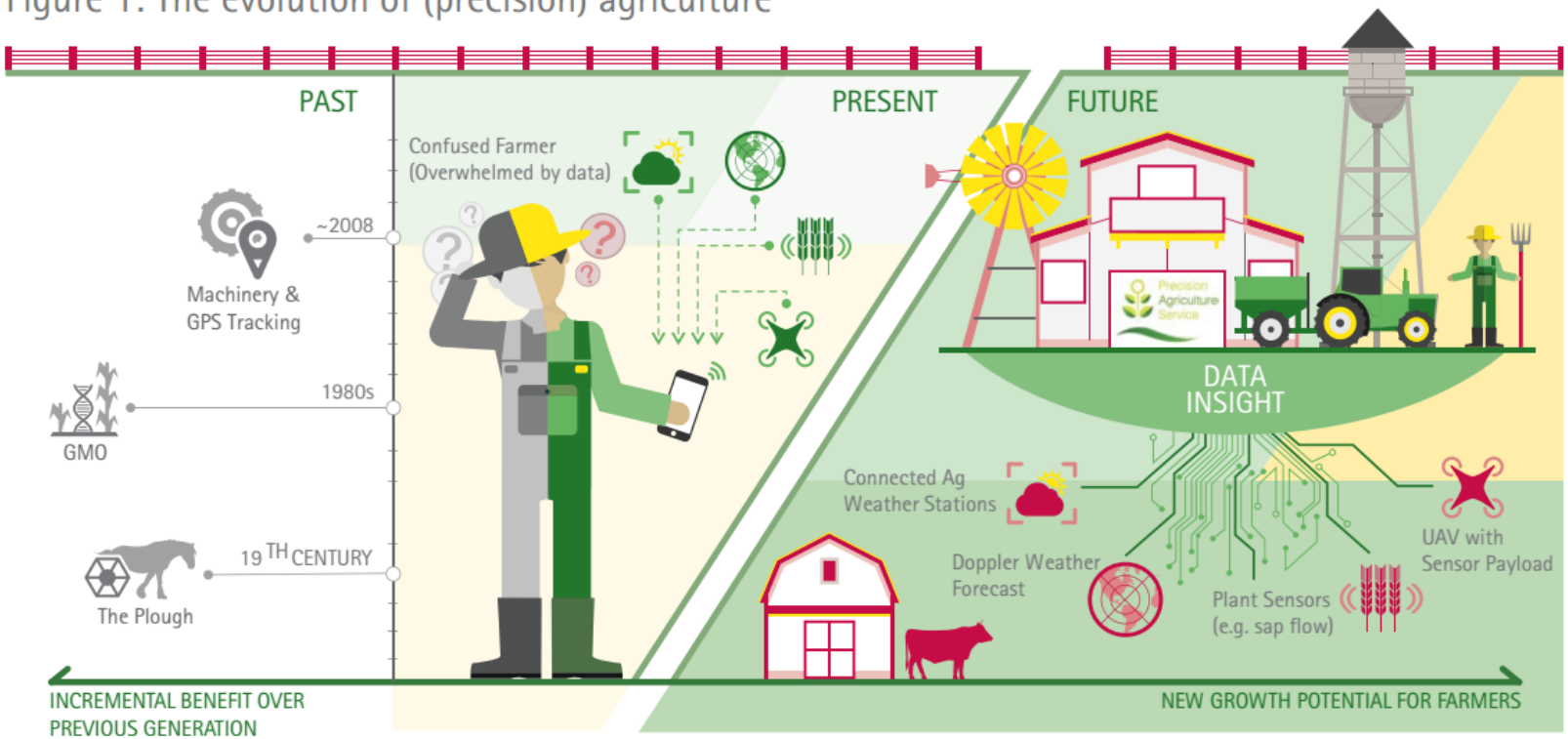
Energy and Logistics



Cortana Intelligence and ML Blog Team

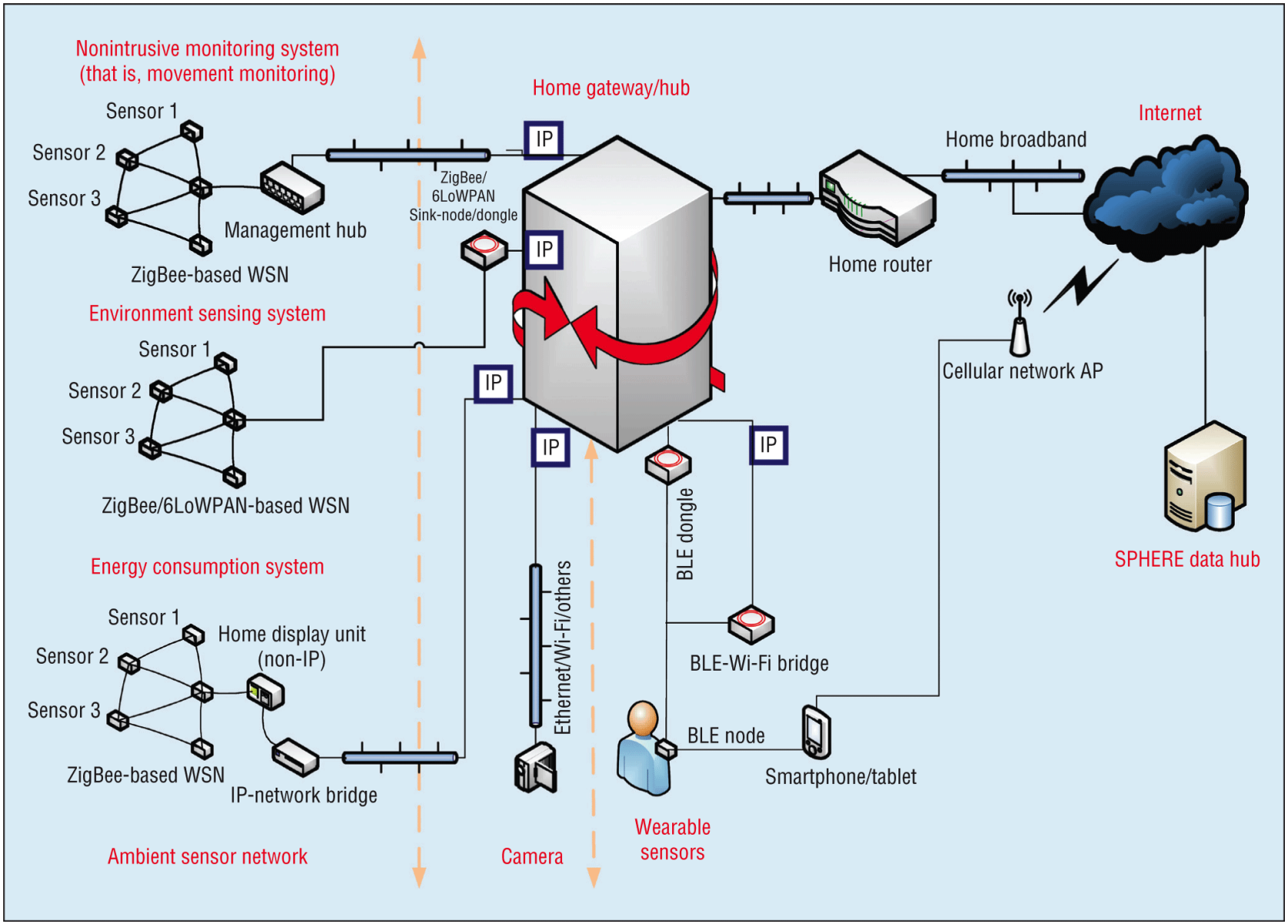
Agriculture

Figure 1: The evolution of (precision) agriculture



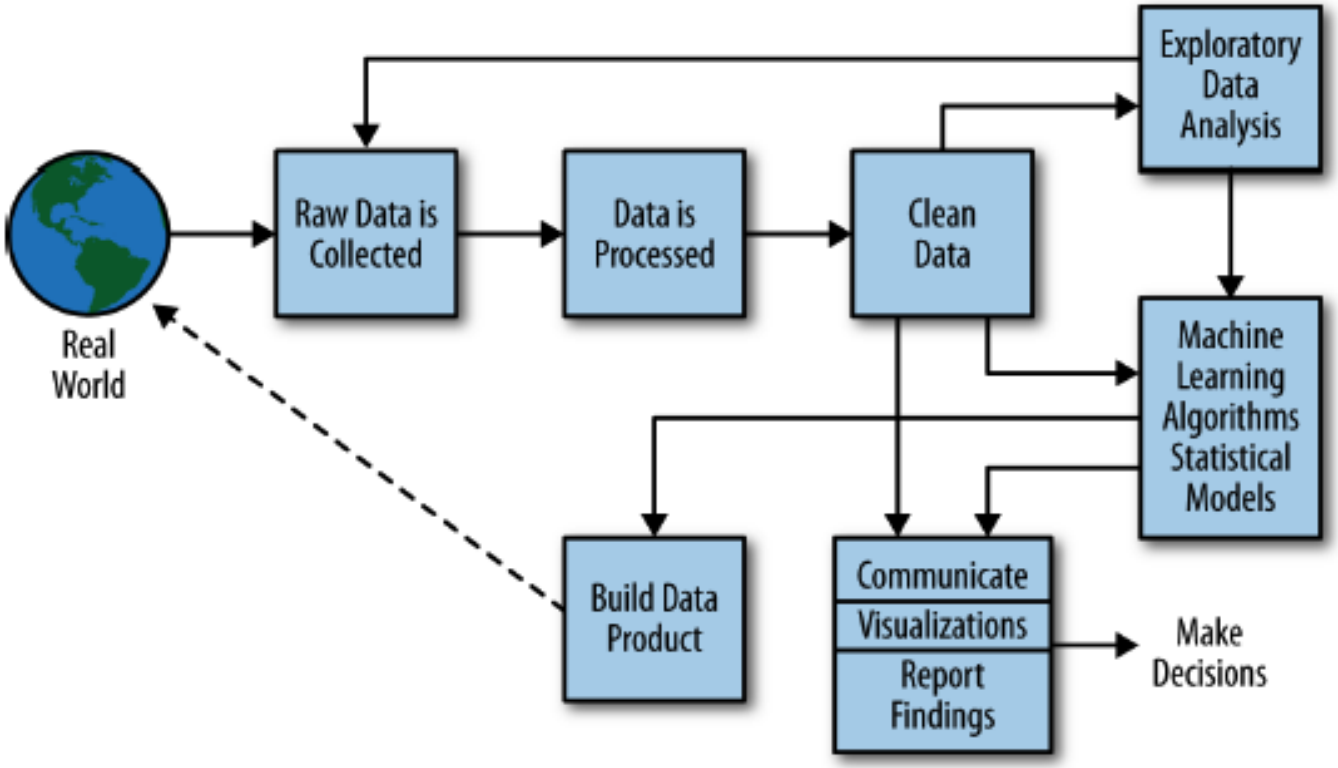
http://www.forbes.com/sites/kurtmarko/2015/08/25/precision-ag-cloud/2/

Healthcare



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

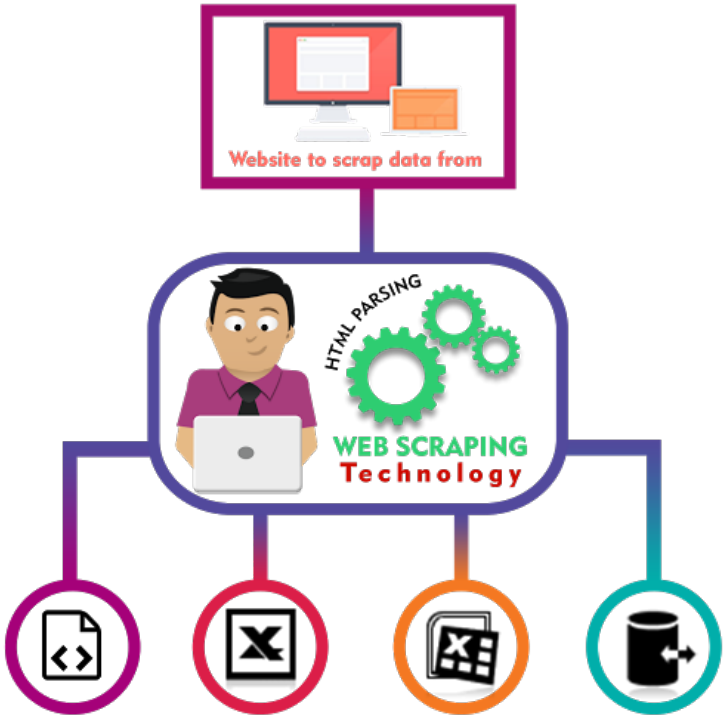
How do we tackle these tasks



Doing data science, Cathy O'Neil and Rachel Schutt

Roadmap

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

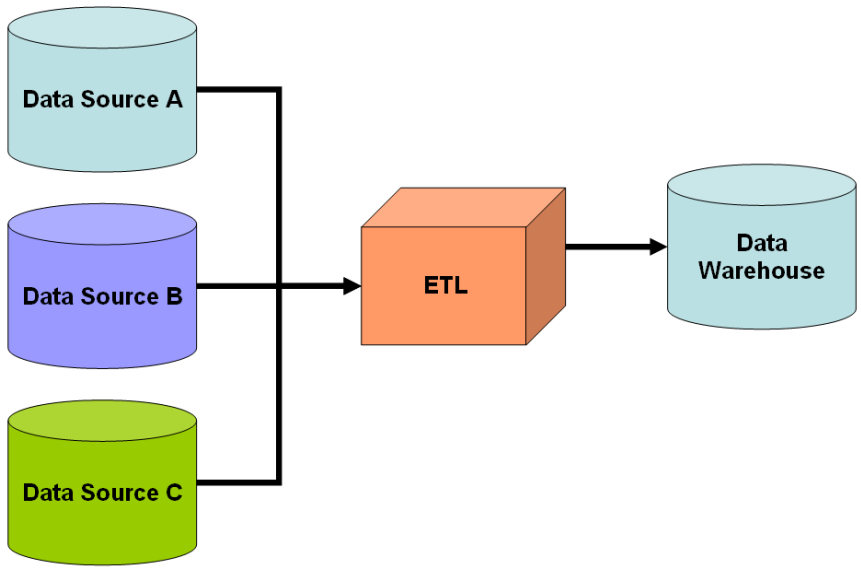


Roadmap

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

Roadmap

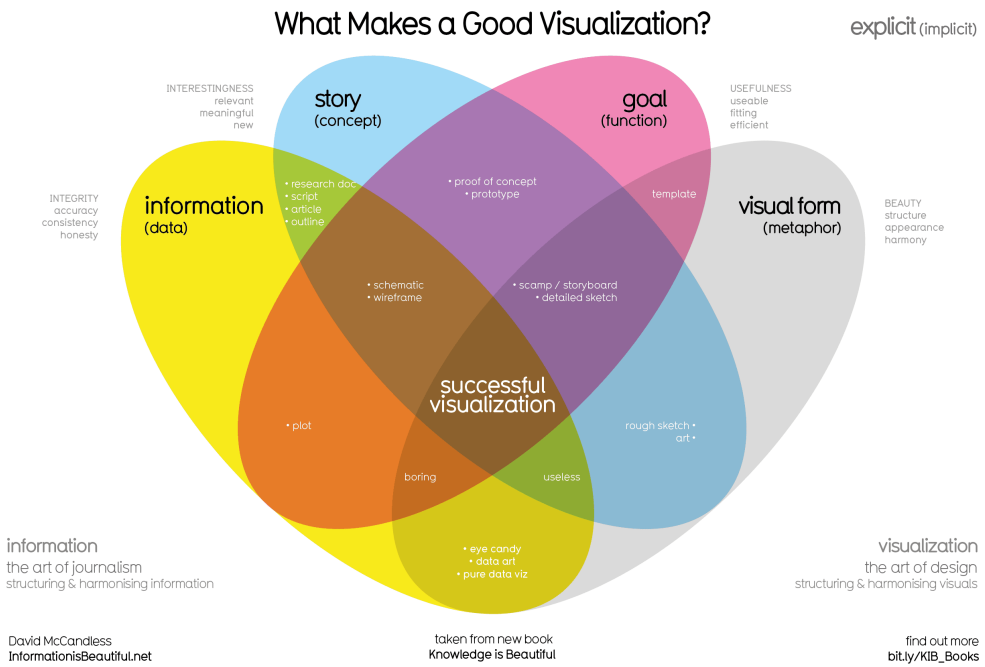
- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Deployment
- Sharing & Privacy



Wikipedia

Roadmap

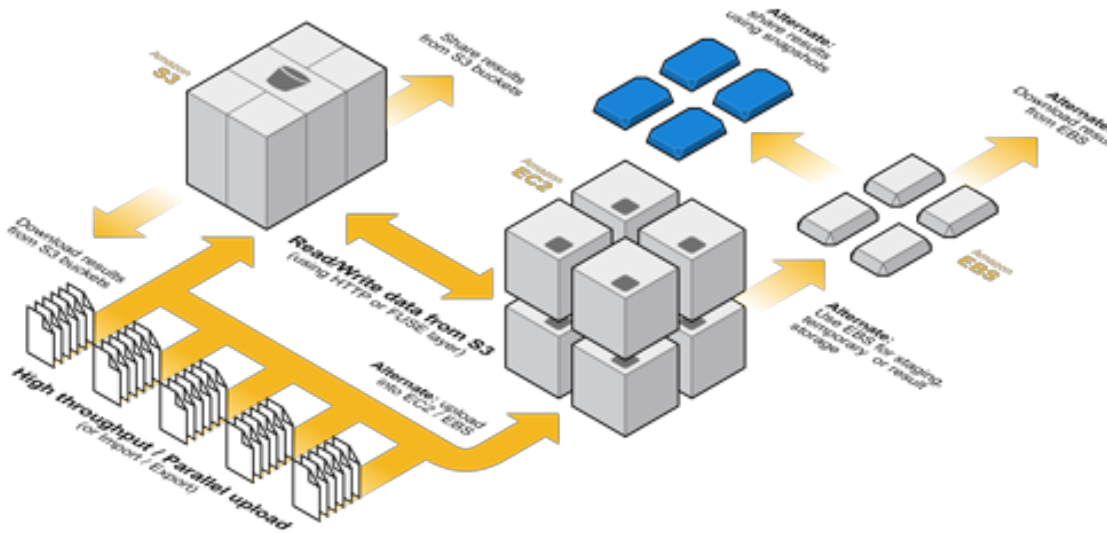
- Ingress & preprocessing
- Storage & management
- Transformation & Integration
- Exploration & Visualisation
- Deployment
- Sharing & Privacy



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

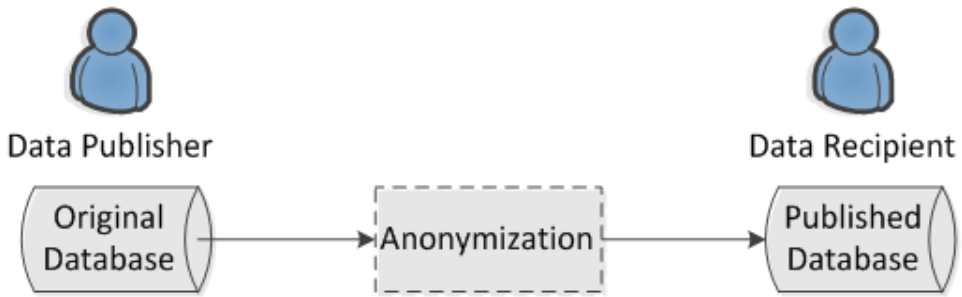
Roadmap

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



Roadmap

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



Summary so far

Applications of Data Science: high-impact, diverse

Challenges: computational/information complexity

Course plan