# Introduction to Data Science

Wray Buntine

http://topicmodels.org

Monash University

# Background

- material developed as part of an introductory masters unit at Monash University
- 6 modules over semester, 1 module in 2 weeks
- download the slides now:
  - links in the slides are active
  - there are some videos and readings I will recommend
- useful resources (blogs, news lists, magazines, *etc.*) at my blog
- please interrupt me with questions!

# Overview of Content

| | |
|---|---|
| **1.** | Data Science and Data in Society<br>overview and look at projects<br>(job) roles, and the impact |
| **2.** | Data Models in Organisations<br>data business models<br>application areas and case studies |
| **3.** | Data Types and Storage<br>characterising data and "big" data<br>data sources and case studies |
| **4.** | Data Resources, Processes, Standards and Tools<br>resources and standards; resources case studies |
| **5.** | Data Analysis Process<br>data analysis theory; data analysis process |
| **6.** | Data Curation and Management<br>issues in data management<br>data management frameworks |

# Outline

# What is Data Science

basic descriptions and history

# What is Data Science?

- contains the word “science” so cannot be a science; **NB.** this is an old joke ...

# What is Data Science?

- contains the word “science” so cannot be a science; **NB.** this is an old joke ...
- circular:

  *data science is what the data scientist does*

# What is Data Science?

- contains the word “science” so cannot be a science; **NB.** this is an old joke ...
- circular:

  *data science is what the data scientist does*

- less circular but a tiny bit more helpful:

  *data science is the technology of handling and extracting value from data*

# What is Data Science?

- contains the word “science” so cannot be a science; **NB.** this is an old joke ...
- circular:

  *data science is what the data scientist does*

- less circular but a tiny bit more helpful:

  *data science is the technology of handling and extracting value from data*

- narrow:

  *machine learning on big data*

# Machine Learning Definition

(well understood and agreed on)

Machine Learning is concerned with the development of algorithms and techniques that allow computers to *learn*.

- concerned with building computational artifacts
- but the underlying theory is statistics

# Why Machine Learning?

- Human expertise does not exist. *e.g.* Martian exploration.
- Humans cannot explain their expertise or reduce it to a ruleset, or their explanation is incomplete and needs tuning, *e.g.* speech recognition.
- Many solutions need to be adapted automatically *e.g.* user personalisation.
- Situation changing in time, *e.g.* junk email.
- There are large amounts of data *e.g.* discover astronomical objects.
- Humans are expensive to use for the work, *e.g.* zipcode recognition.

# Why Machine Learning?

you don't want to be this guy!

# Why Machine Learning?

- the information society
- information warfare
- information overload
- information access

**Exercise:** Google these to find out about them!

# Data Science Examples

- Google's spell checker and *translate*
- Amazon.com's *recommendation engine*
- *"saturated fat is not bad for you after all"*
- Microsoft's *Predictive Analytics for Traffic* from 2005

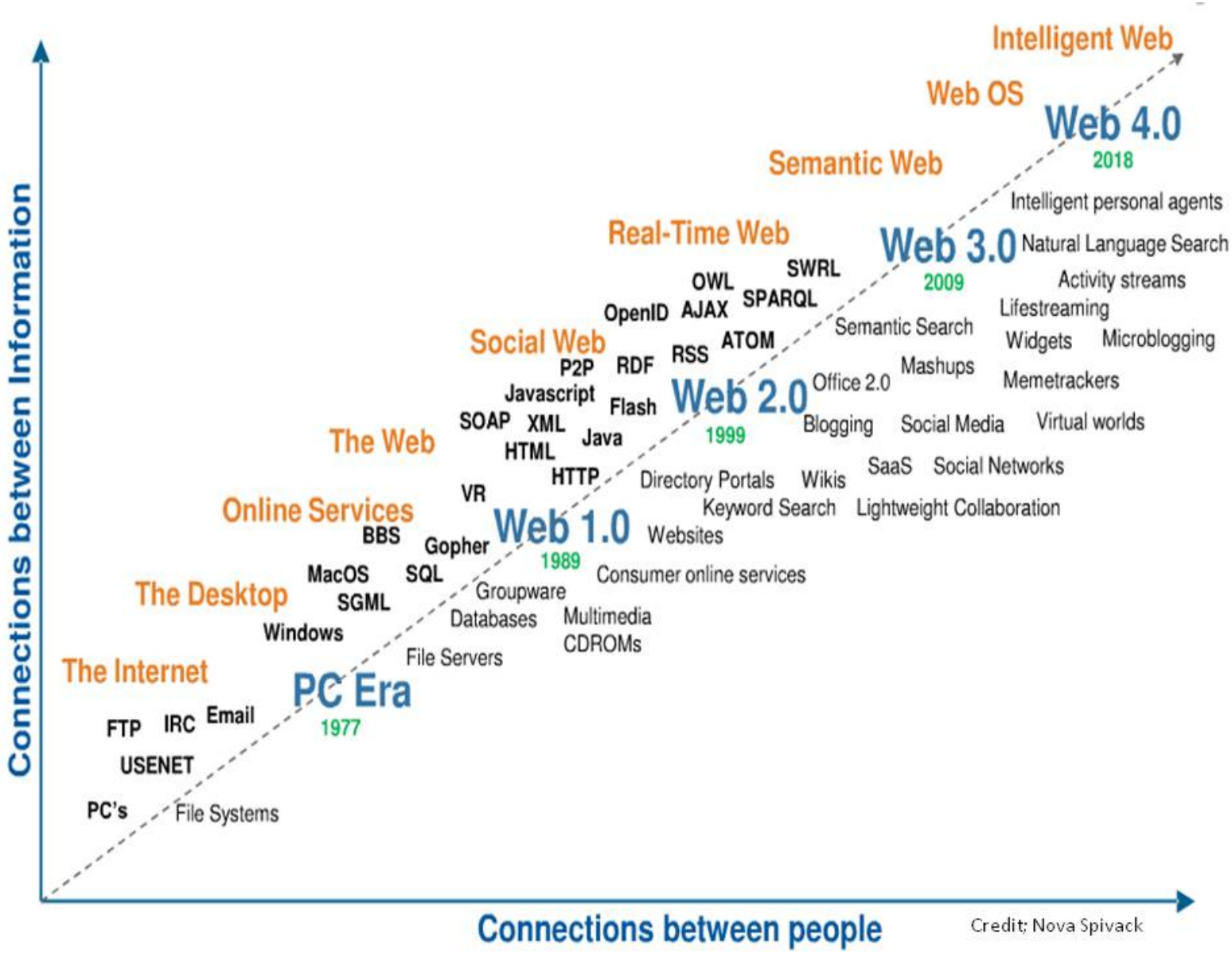
# Historical Context

*Wolfram Alpha: computable knowledge history*

*Cloud Infographic: Evolution Of Big Data*

# Web X.0

(credit: Nova Spivack)



# The Data Science Process

what happens in a Data Science project?

- illustrating the process
  - a quick walkthrough illustrating the steps
- the standard value chain
  - our model of the process

# The Data Science Process:
# Illustrating the Process

a quick walkthrough illustrating the steps

# The Data Science Process

- many different tasks come together to complete a Data Science project
- a data scientist should be familiar with most, but doesn't need to be an expert in all
- not all are labelled as Data Science
  - some from other field such as computer engineering, business, ...

Pitch your ideas.

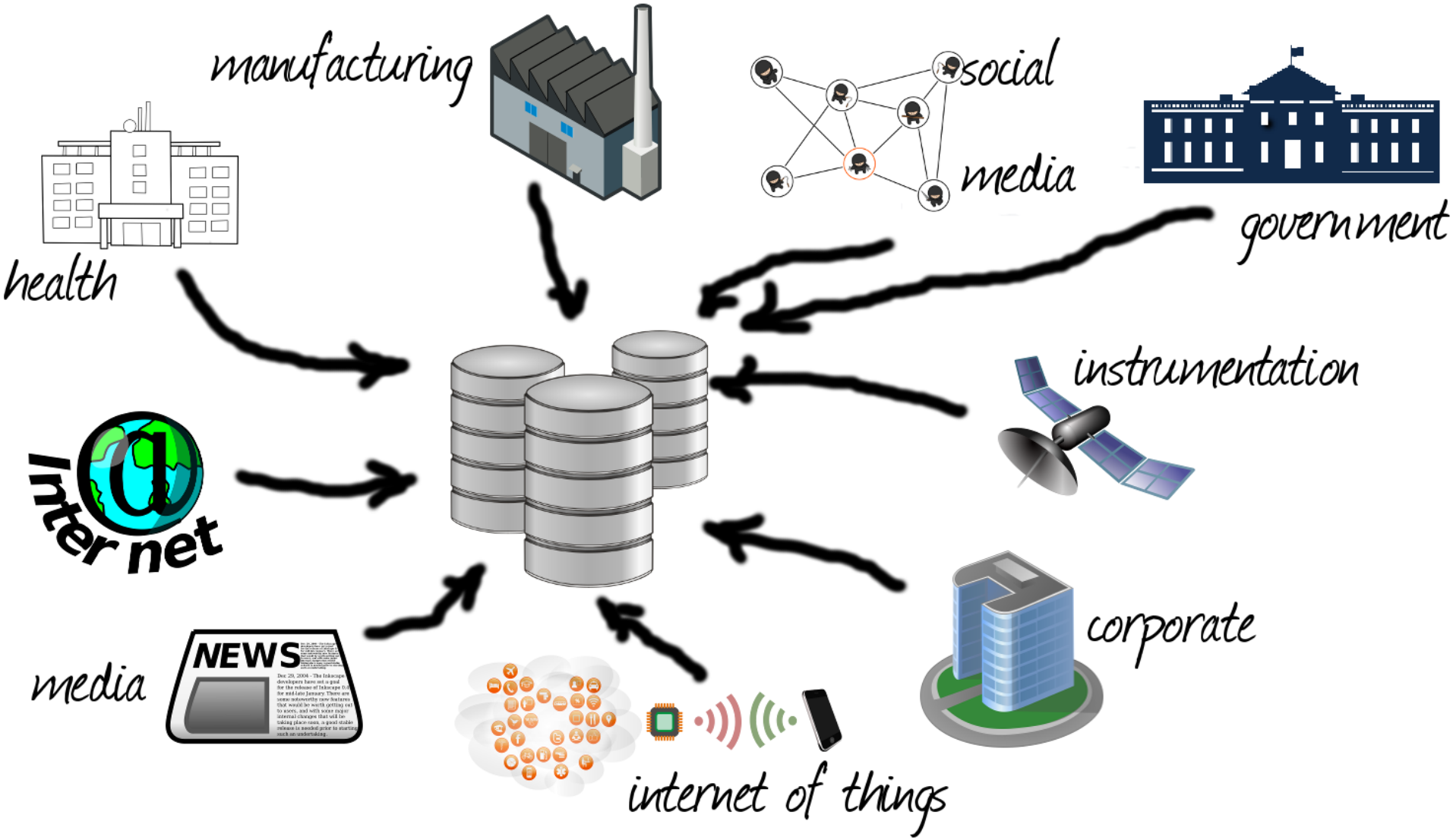
*"Young Business Man Holding a Tablet" by Pic Basement, CC-BY 2.0*


Researchers preparing to x-ray a patient.

*by Stephen Ausmus acquired from USDA ARS, public domain.*

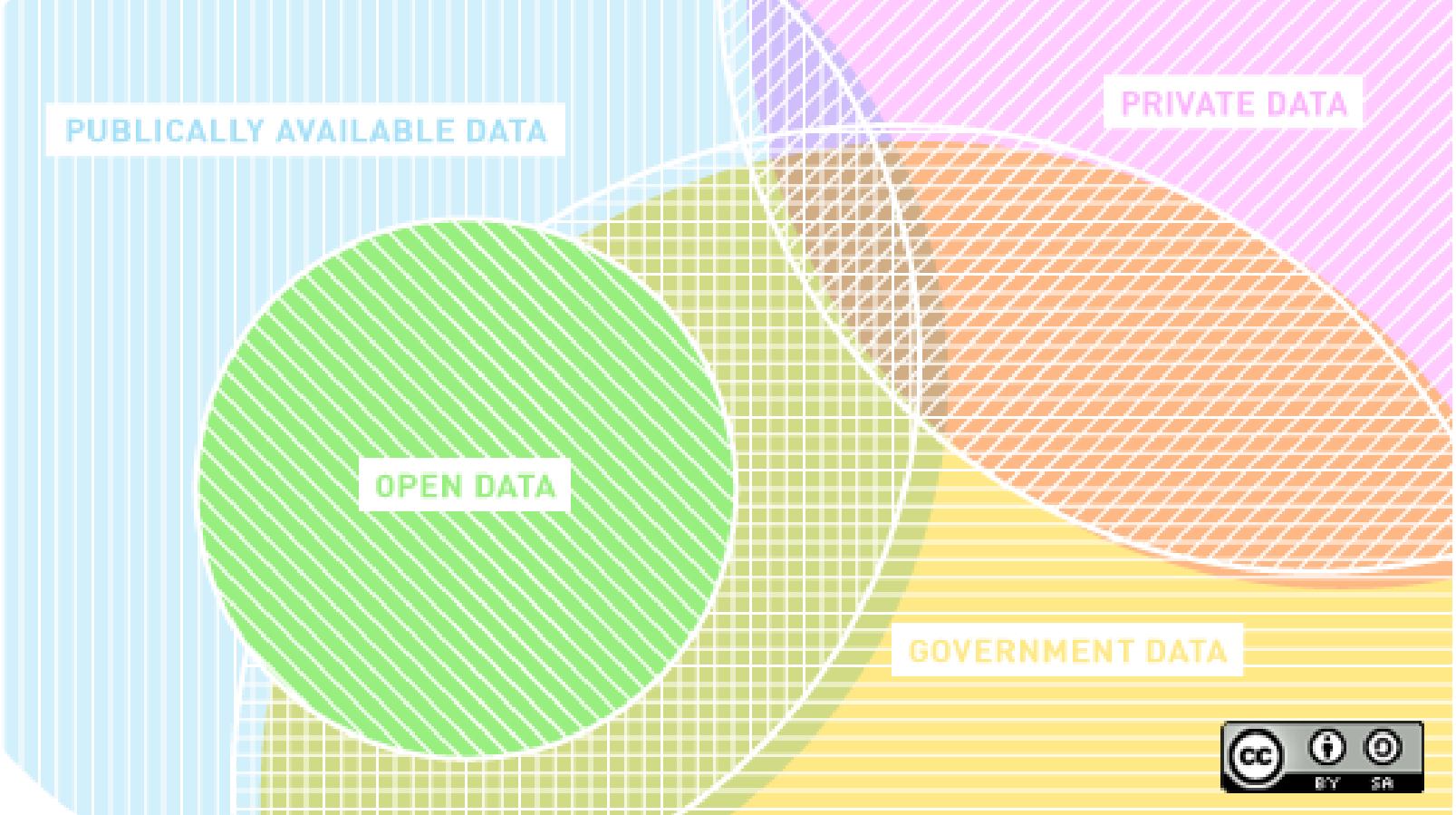
Scientists watch over data collected by the gravimeter and magnetometer instruments.

*by NASA/GSFC/Jefferson Beck, CC-BY 2.0*



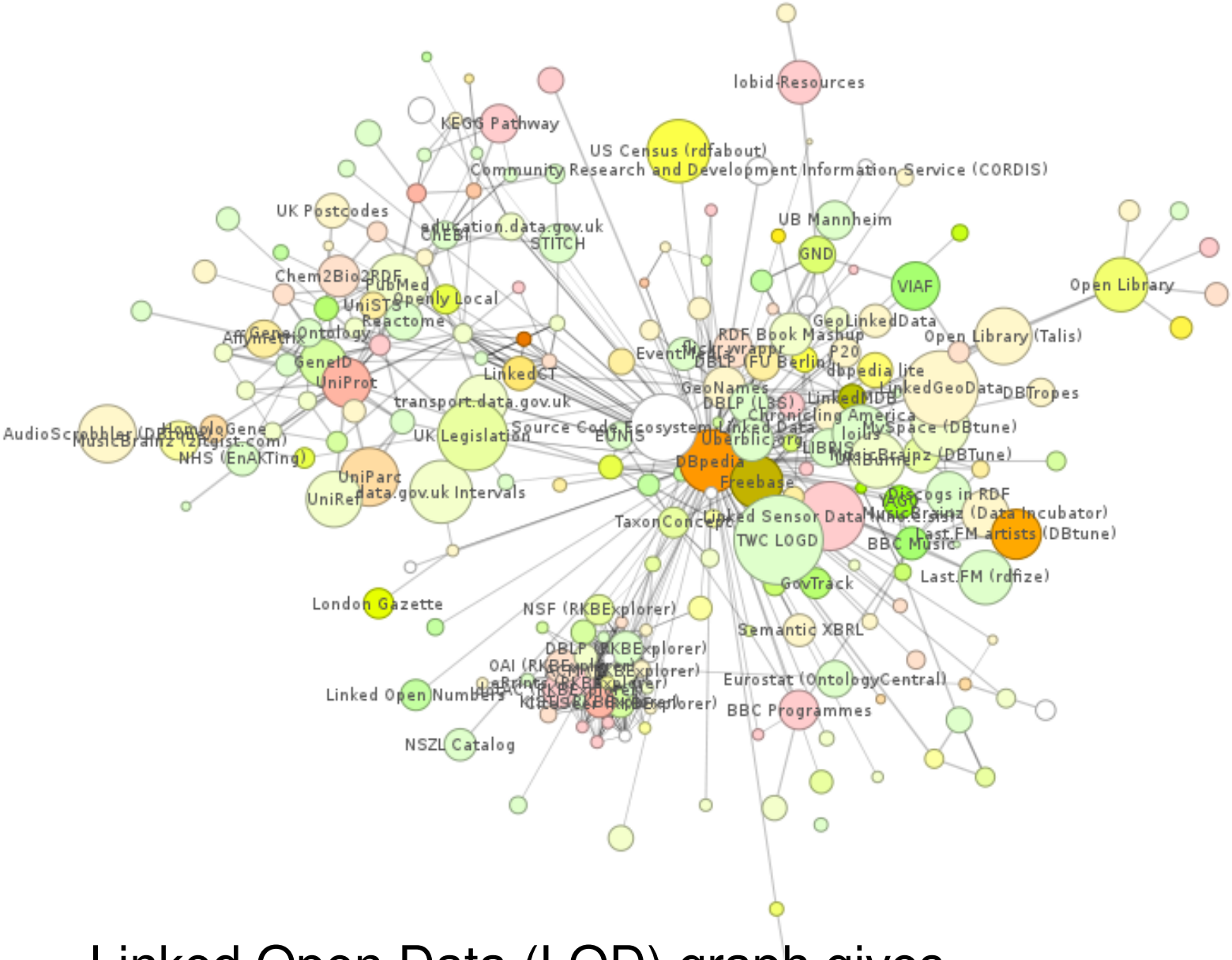

Data can be got from many sources.

*icons from by Openclipart.org, public domain*

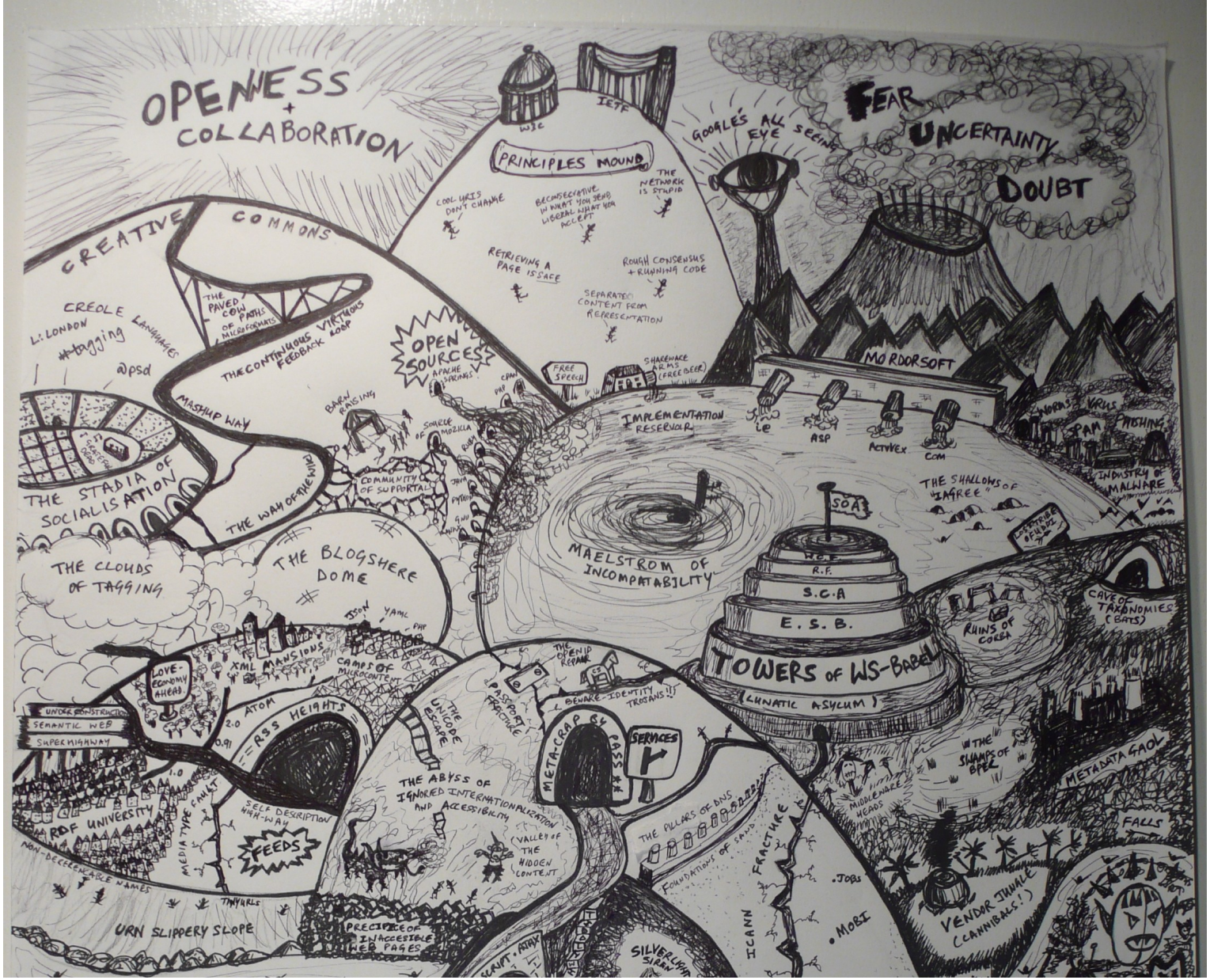



Some of the best data is Open Data.

*by Libby Levi for opensource.com, CC-BY-SA 2.0*

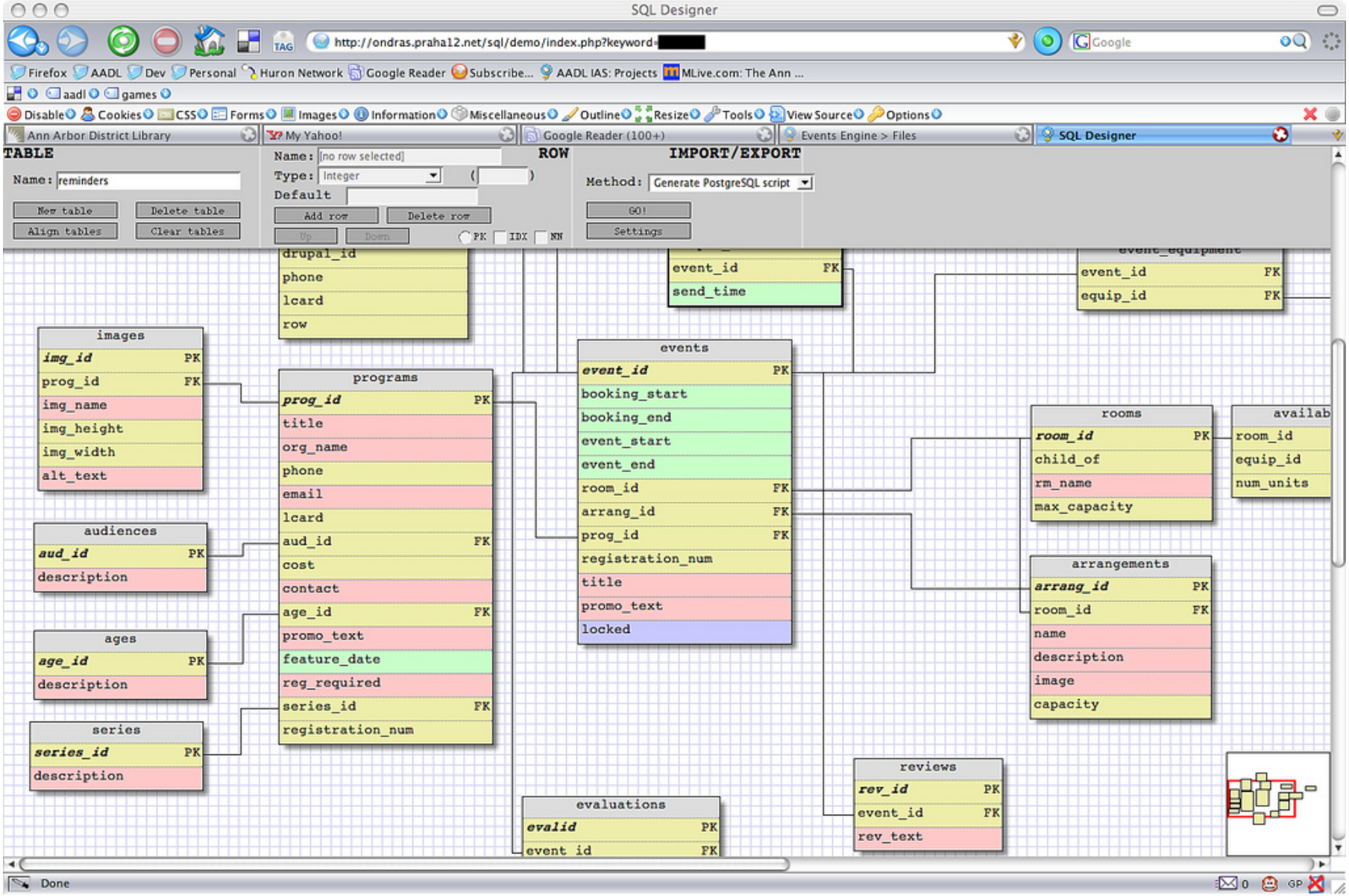



Linked Open Data (LOD) graph gives semantics.

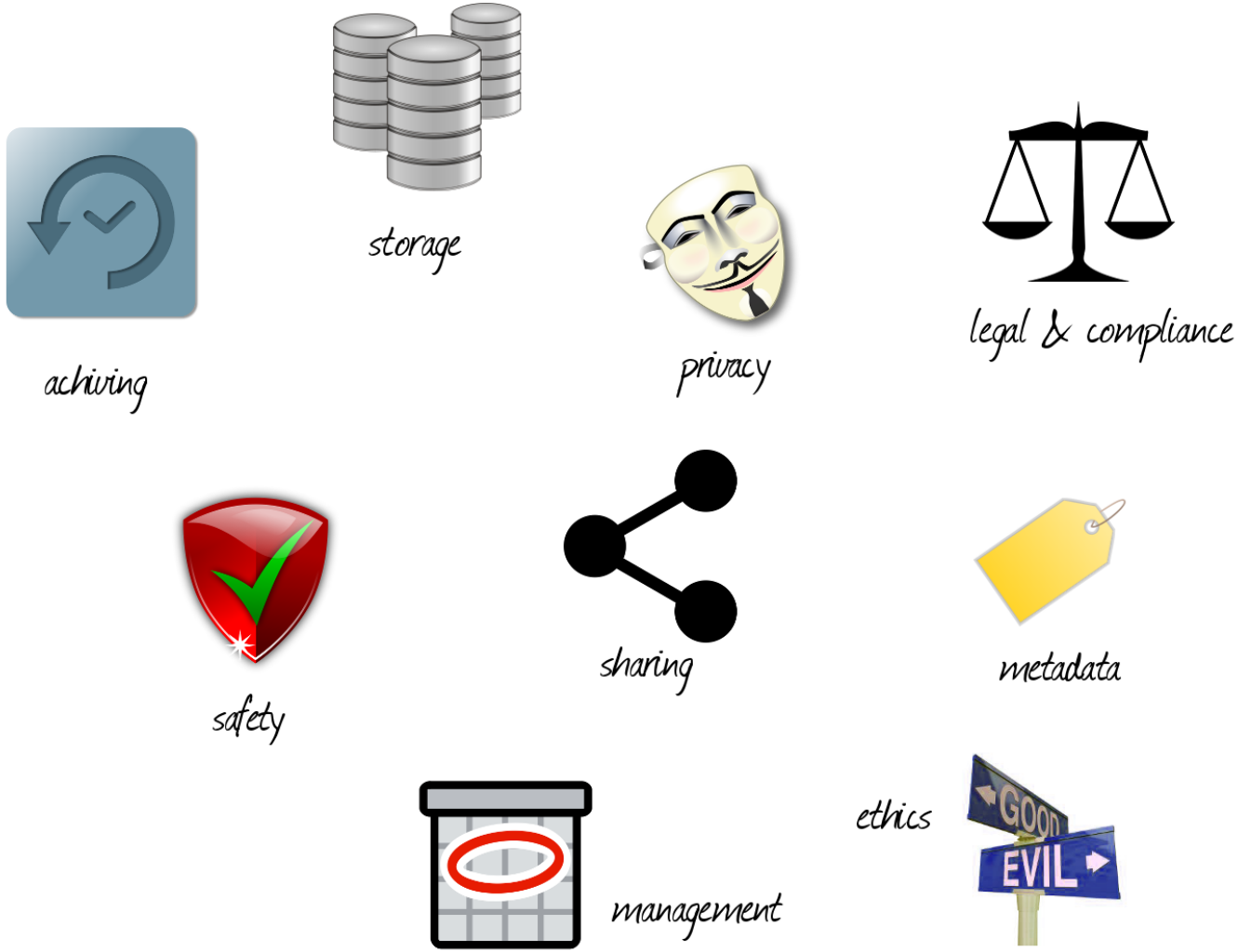



# Navigate data standards and formats

*"The Web is Agreement" cropped, by Paul Downey, CC-BY 2.0*




Understand the database schema.



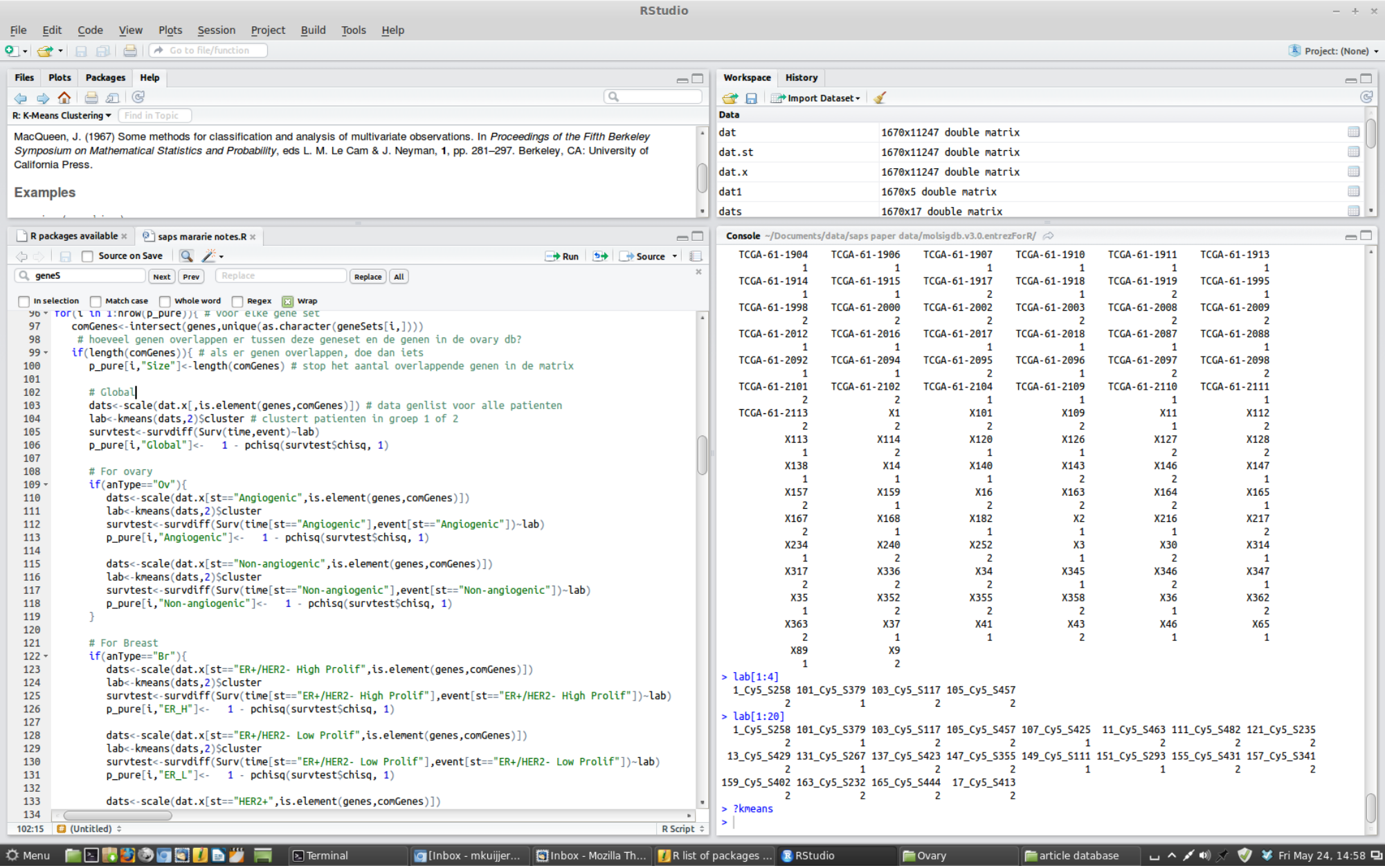

Governance cares for the data and its subjects.

*icons from by Openclipart.org, public domain;*

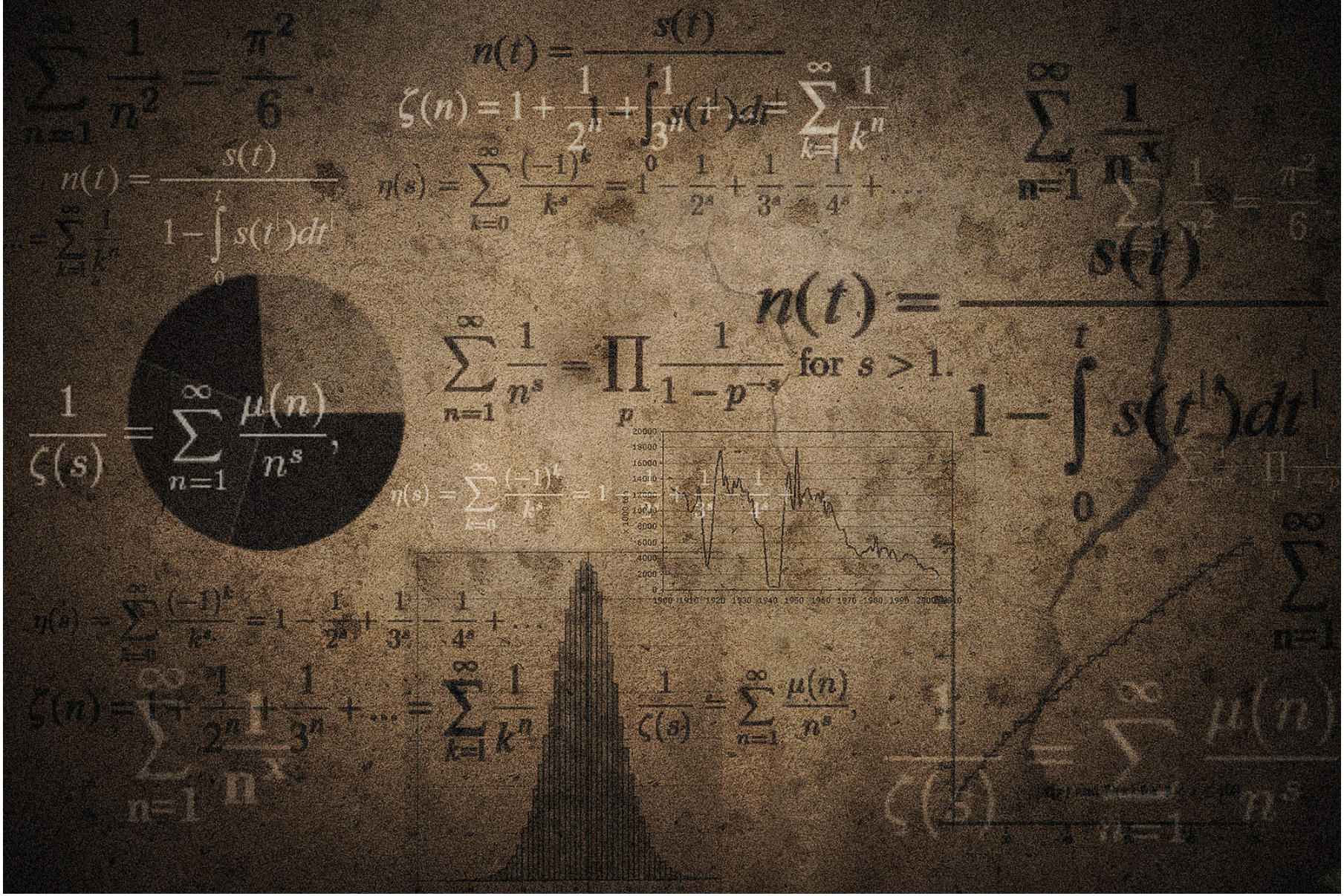
*Good and Evil by AJC ajcann.wordpress.com, CC-BY-SA 2.0*

Data engineers make the back-end work

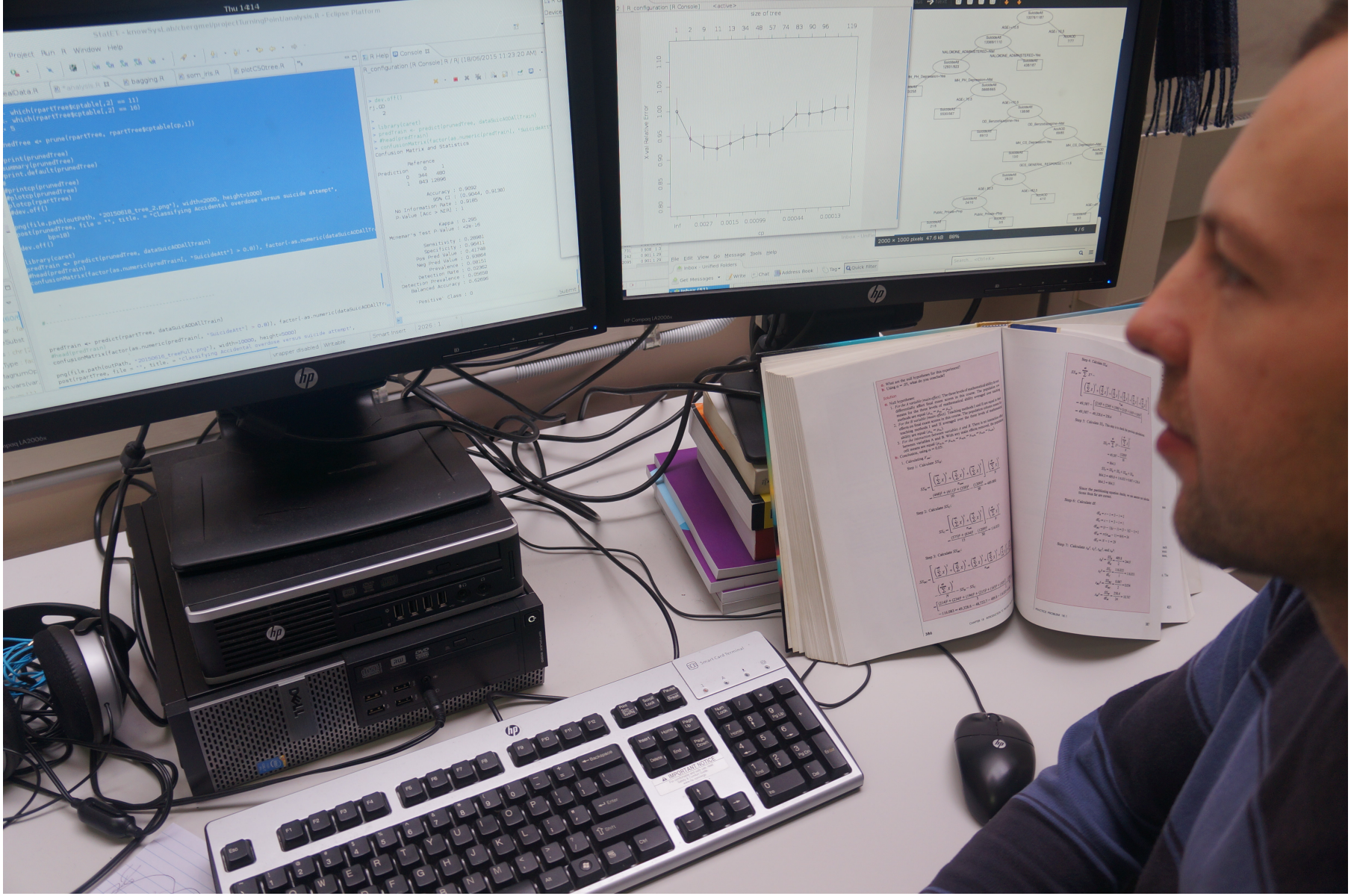
*by Intel Free Press, CC-BY 2.0*



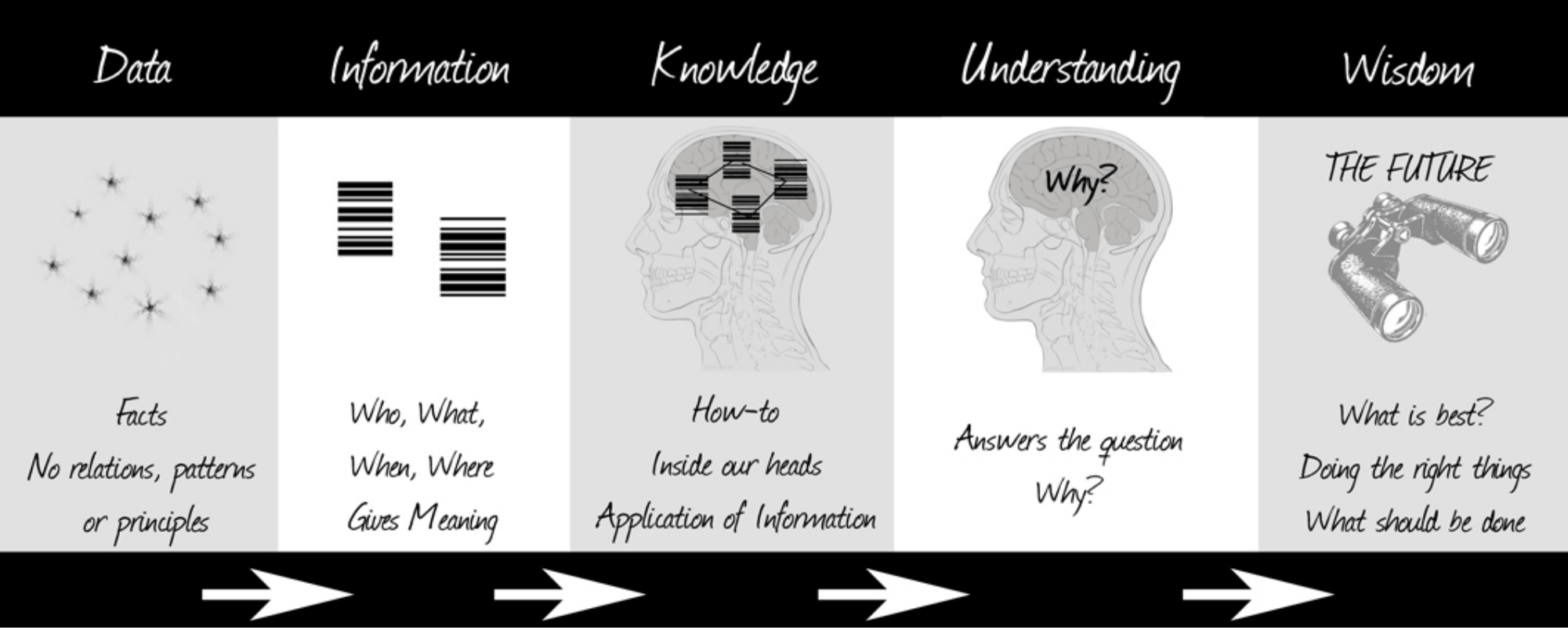

# Inspect and clean the data.



Propose a conceptual/mathematical/functional model.



Analyst builds models with his favorite tool.



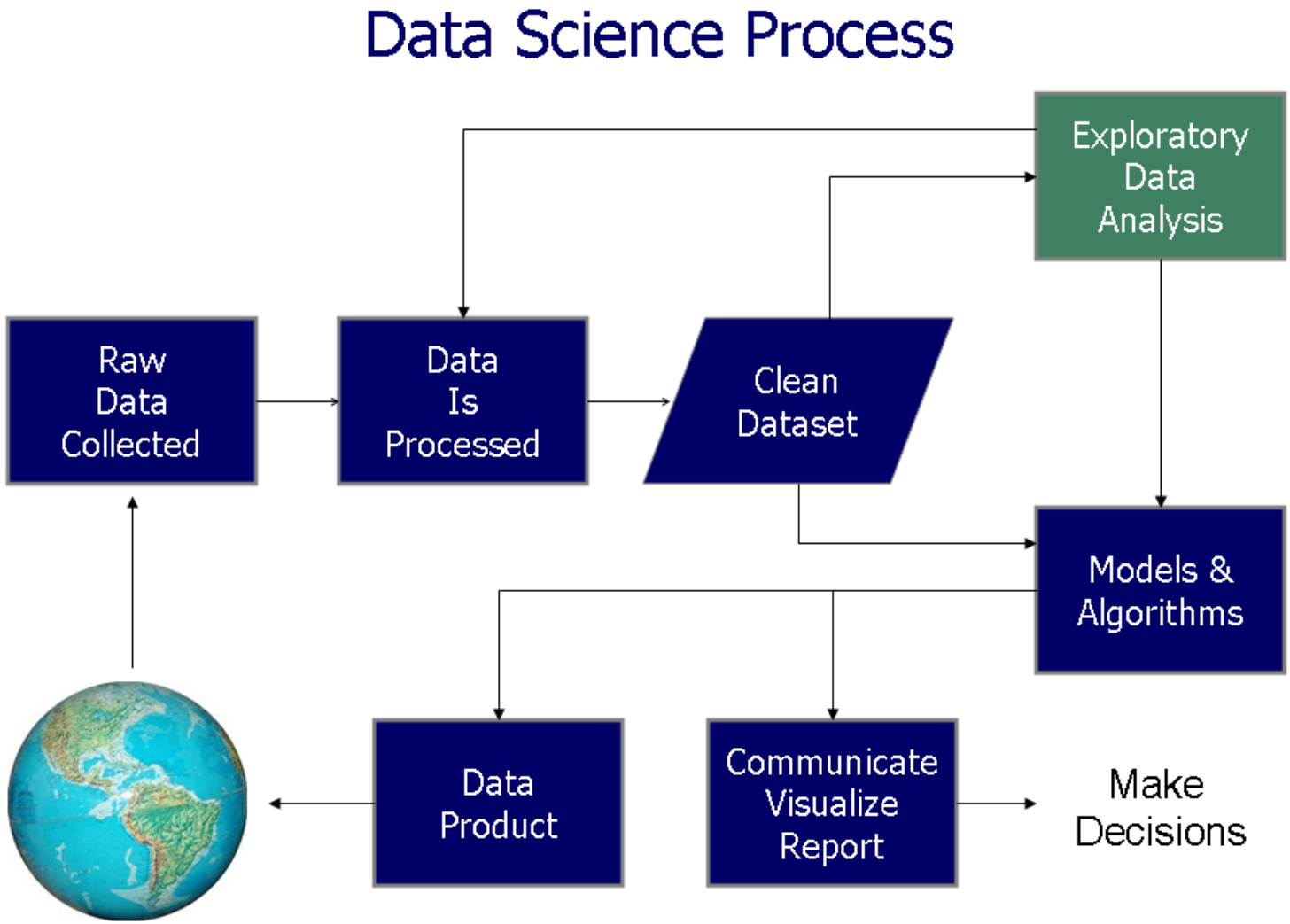

Analysis, statistics and/or machine learning works on the data.


Choose visualizations, many different options!

*"Visualization Matrix" cropped, by Lauren Manning, CC-BY 2.0*


Visualise data to interpret/present results.

*by Stephen Ausmus acquired from USDA ARS, public domain.*




Data science process flowchart.


Operationalization: putting the results to work.

*"Illustration of Strategy“ by Denis Fadeev, CC-BY-SA 3.0*

# The Data Science Process:
# A Proposed Value Chain

our model of the process

# Parts of a Data Science Project

**Collection:** getting the data

**Engineering:** storage and computational resources across full lifecycle

**Governance:** overall management of data across full lifecycle

**Wrangling:** data preprocessing, cleaning

**Analysis:** discovery (learning, visualisation, *etc.*)

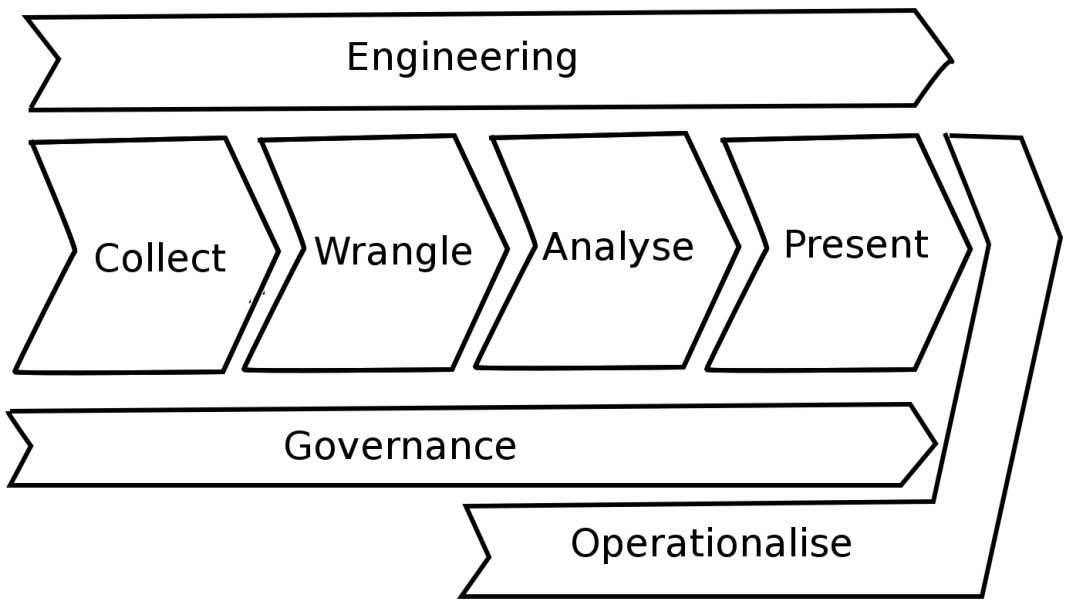
**Presentation:** arguing the case that the results are significant and useful

**Operationalisation:** putting the results to work, so as to gain benefits or value

We call this the Standard Value Chain.

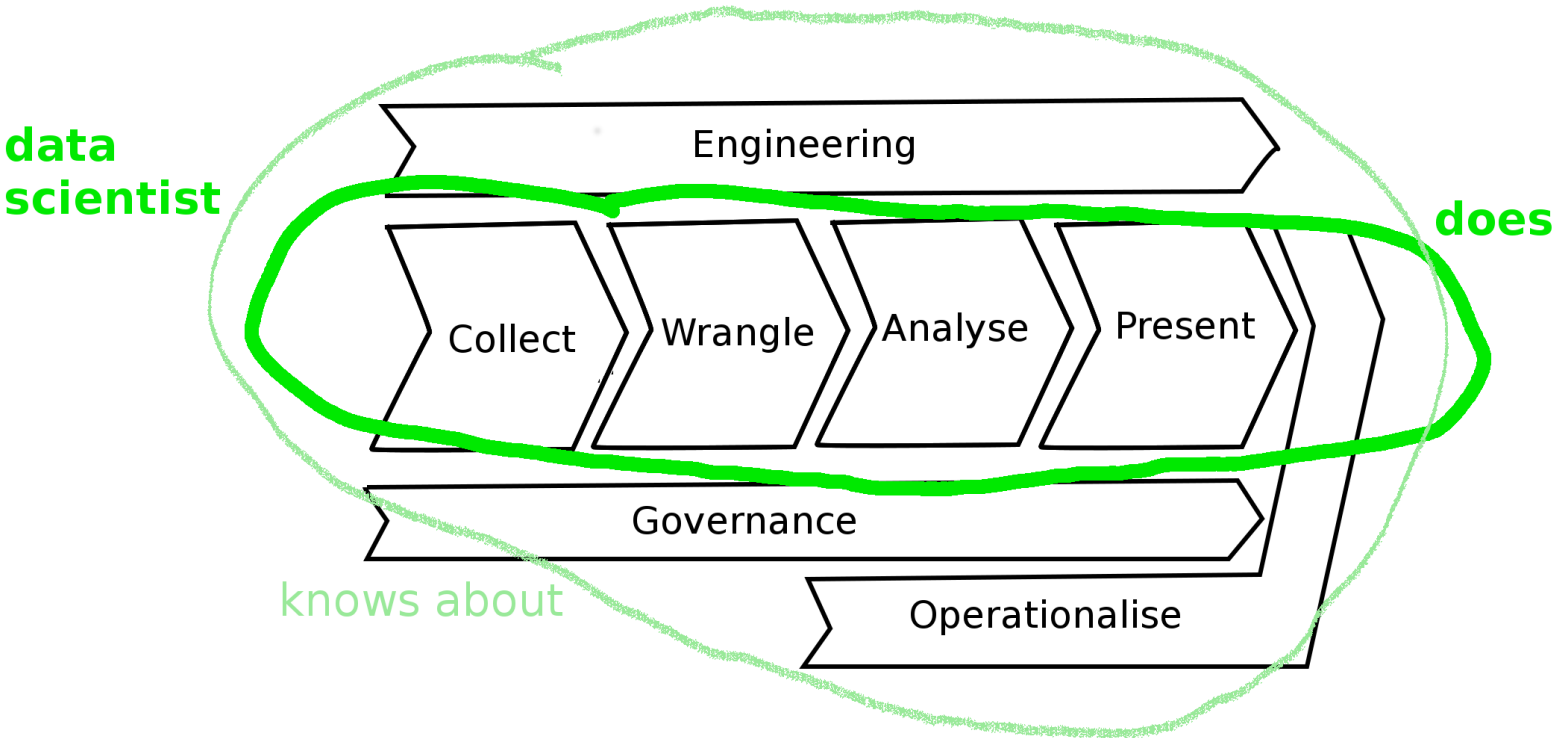
# The Value Chain

Collection: getting the data
Engineering: storage and computational resources
Governance: overall management of data
Wrangling: data preprocessing, cleaning
Analysis: discovery (learning, visualisation, *etc.*)
Presentation: arguing that results are significant and useful
Operationalisation: putting the results to work




we will refer to this throughout!

# Doing Data Science




Data scientist ::= addresses the data science process to extract meaning/value from data

# From *What is Data Science?*

A quote from *Jeff Hammerbacher*

> *... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization ...*

hypothesis test ::= statistical test to evaluate a simple claim

regression analysis ::= fitting a curve to real valued data

Hadoop ::= system for partitioning computation across a compute cluster

# Data Science Emerges

the beginnings of data science

# Related: Data Engineering

- building scalable systems for storage, processing data
- *e.g.* Amazon Web Services, Teradata, Hadoop, ...
- databases, distributed processing, datalakes, cloud computing, GPUs, wrangling, ...
- huge, continuous improvement ....

# Related: Data Analysis

- performing analysis and understanding results
- *e.g.* R, Tableau, Weka, Microsoft Azure Machine Learning, ...
- machine learning, computational statistics, visualisation, ...
- huge, continuous improvement ....

# Related: Data Management

- managing data through its lifecycle
- *e.g.* ANDS, Talend, Master Data Management, ...
- ethics, privacy, providence, curation, backup, governance, ...
- huge, continuous improvement ....

# Fits and Starts

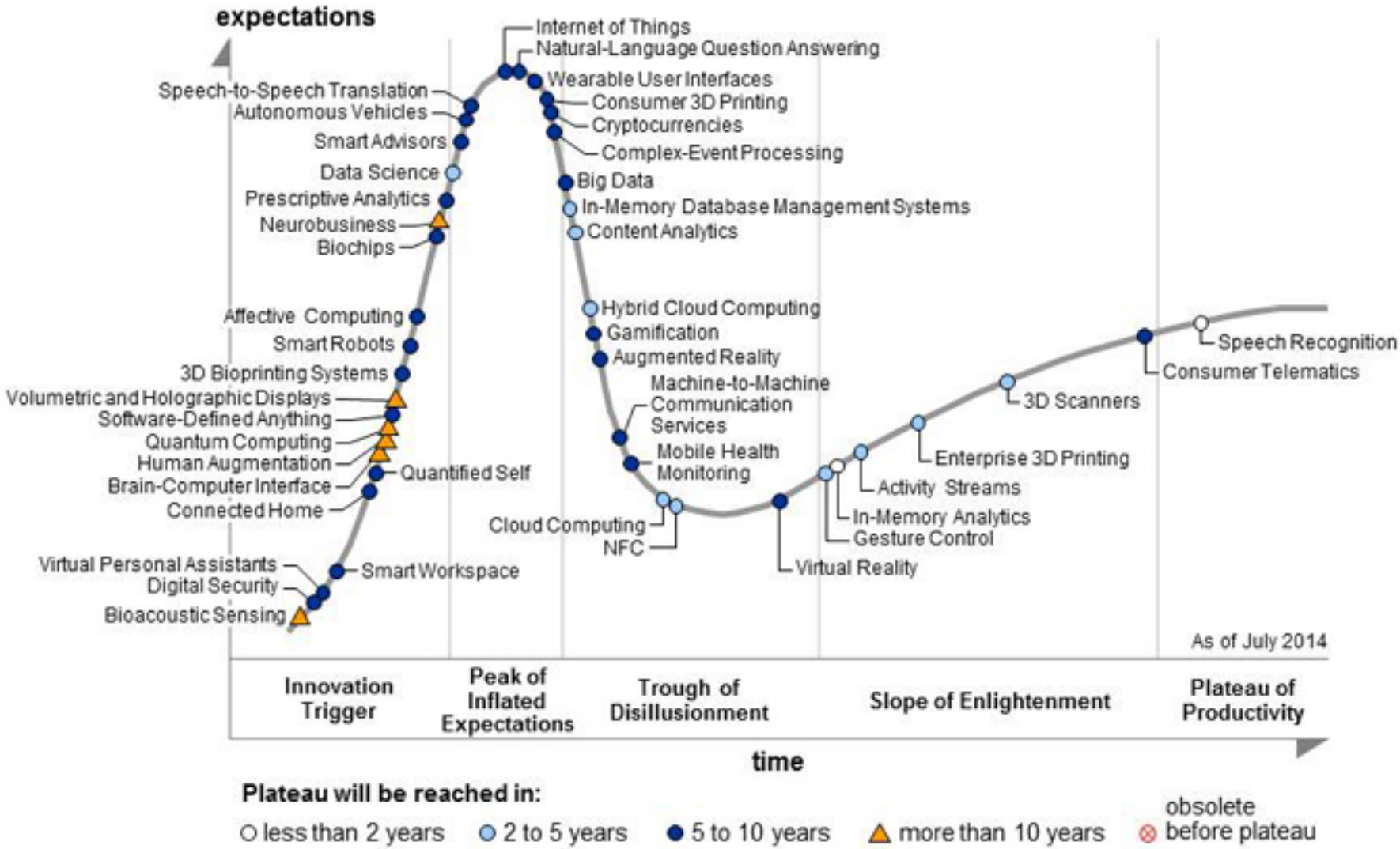
- Data Analysis (John Tukey) in 1962
- Expert Systems in the 1980's
- Machine Learning in the 1980's
- Data Mining in the 1990's
- see *Business Week's "Database Marketing"* cover story September 1994

# Data Science Emerges $\sim$2000

- data analysis came of age 1990's
- William Cleveland publishes in 2001 *"Data Science: An Action Plan for ... the field of Statistics"*
- data engineering came of age 2000's (Dot.Com boom)
- (digital) data management came of age 2000's (Dot.Com boom)
- the data/information society
- business pressure on decision making
- "data" as a valuable asset
- Dot.Com companies show the way

see also David Donoho's *"50 years of Data Science"* (PDF paper)

# Hype Cycle 2014



# Data Science Research Programs

- National Institute of Standards (NIST, in US) Big Data Working Group (2013-2015)
- US National Academy of Sciences' Committee on the Analysis of Massive Data (2013)
- Alan Turing Institute for Data Science at London's new Knowledge Quarter (near National Library, 2016-???)
- major growth in universities internationally

# Impact of Data Science

some examples of how data science is impacting others:

- your life in the cloud
  - datafication of you
- science and social good
  - scientific method holds true, but broadens technology

# Your Life on the Cloud

From *Year Zero: Our life timelines begin*

Our personal information is increasingly stored in the cloud (though perhaps behind firewalls): social life (Facebook), career (LinkedIn), search history (Google, *etc.*), health and medical (Fitbit, TBD), music (Apple), ...

Many, many advantages:

*e.g.* personal agents

- computerised support for health
- ...

But some disadvantages:

*e.g.* security and privacy breaches

- ...

# Your Life on the Cloud (cont.)

But

- corporate leakage to government (security, tax, *etc.*)
- what if you don't have rights to access/delete data?
- security and privacy breaches
- what if we've changed our ways?
- the department of pre-crime
- corporate mergers
- "the science is settled" and government mandates

# Data Science for Science

- fields like physics, bioinformatics and earth science used big data anyway
  - had their own independent data science revolution
- in other areas has raised the profile of data-driven science
- spurred on governments to develop cross-disciplinary programmes
  - *Alan Turing Institute for Data Science* in the UK
- has provided new data sources and tools for collecting data
  - crowd sourcing
  - social media
- allows for citizen/participatory science
  - *DataONE*

# Data Science for Social Good

Example:
*"Data, Predictions, and Decisions in Support of People and Society"*
by Eric Horvitz (Distinguished Scientist & Managing Director at Microsoft) see the final section of video 46:51-53:00 mins.

Interactive website *Aid Data* (making development finance data more accessible).

*Data Science for Social Good* movement training data scientists to support community and charity.

# Health Care Futurology

see “Big data – 2020 vision” talk by SAP manager John Schitka

- your stomach can be instrumented to assess contents, nutrients, *etc.*
- your bloodstream can be instrumented too assess insulin levels, *etc.*
- your “health” dashboard can be online and shared by your GP
- health management organisations (HMO) tying funding levels to patient care performance
- GP/HMO will know about your icecream/beer binge last night and you missing your morning run
- longitudinal studies feasible

# Outline

# Business Models

From Wikipedia:

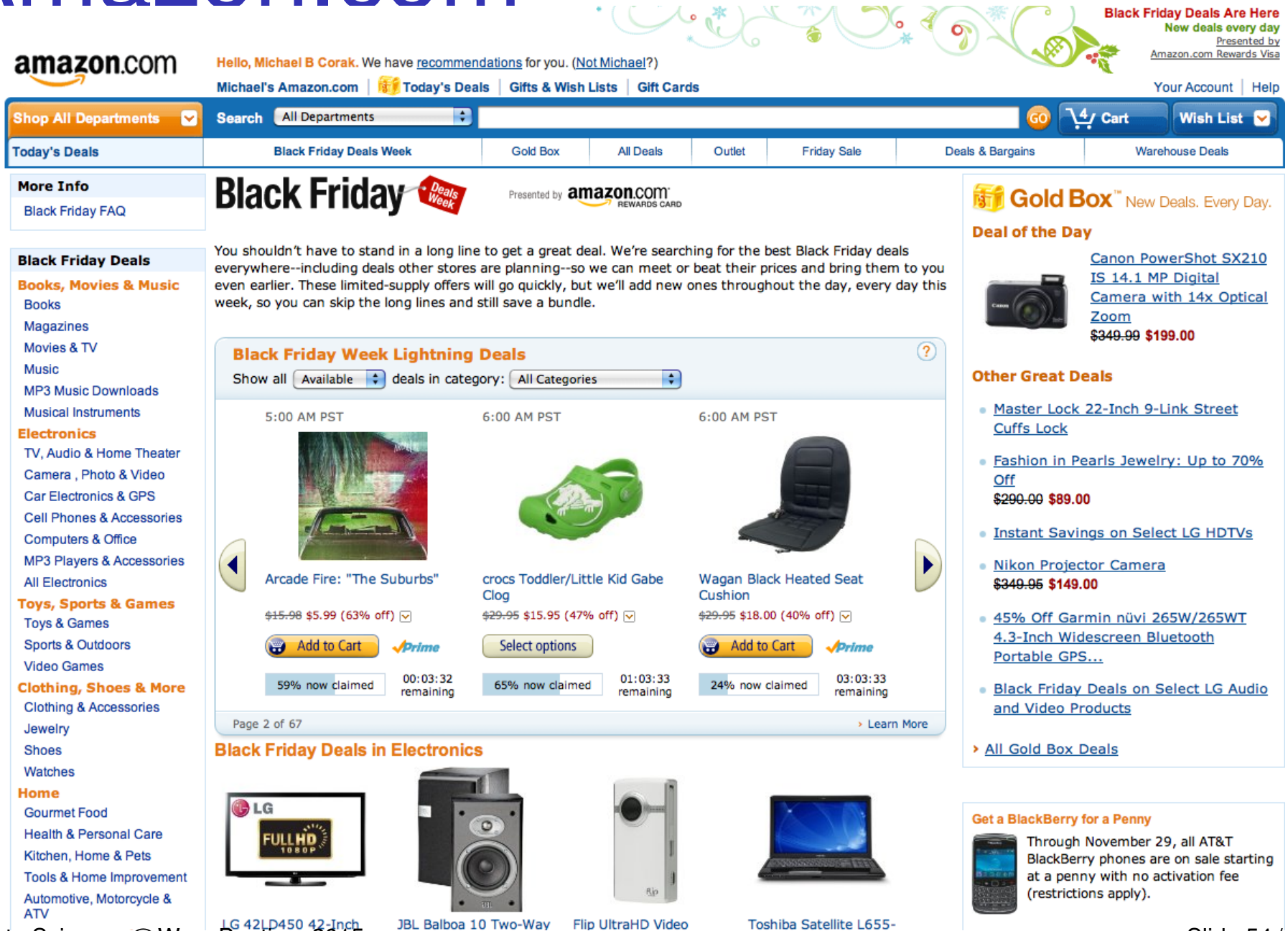
> *A business model describes the rationale of how an organization creates, delivers, and captures value, in economic, social, cultural or other contexts.*

Examples of general classes:

- retailer versus wholesaler
- luxury consumer products
- software vendor
- service provider

What kinds of businesses do we have operating in the Data Science world?

# Amazon.com


amazon.com
Hello, Michael B Corak. We have recommendations for you. (Not Michael?)
Michael's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards
Your Account | Help
Black Friday Deals Are Here
New deals every day
Presented by
Amazon.com Rewards Visa
Shop All Departments
Search
All Departments
GO
Cart
Wish List
Today's Deals
Black Friday Deals Week
Gold Box
All Deals
Outlet
Friday Sale
Deals & Bargains
Warehouse Deals
More Info
Black Friday FAQ
Black Friday Deals
Books, Movies & Music
Books
Magazines
Movies & TV
Music
MP3 Music Downloads
Musical Instruments
Electronics
TV, Audio & Home Theater
Camera , Photo & Video
Car Electronics & GPS
Cell Phones & Accessories
Computers & Office
MP3 Players & Accessories
All Electronics
Toys, Sports & Games
Toys & Games
Sports & Outdoors
Video Games
Clothing, Shoes & More
Clothing & Accessories
Jewelry
Shoes
Watches
Home
Gourmet Food
Health & Personal Care
Kitchen, Home & Pets
Tools & Home Improvement
Automotive, Motorcycle & ATV
Black Friday
Deals Week
Presented by amazon.com REWARDS CARD
You shouldn't have to stand in a long line to get a great deal. We're searching for the best Black Friday deals everywhere--including deals other stores are planning--so we can meet or beat their prices and bring them to you even earlier. These limited-supply offers will go quickly, but we'll add new ones throughout the day, every day this week, so you can skip the long lines and still save a bundle.
Black Friday Week Lightning Deals
Show all Available deals in category: All Categories
5:00 AM PST
Arcade Fire: "The Suburbs"
$15.98 $5.99 (63% off)
Add to Cart
Prime
59% now claimed
00:03:32 remaining
6:00 AM PST
crocs Toddler/Little Kid Gabe Clog
$29.95 $15.95 (47% off)
Select options
65% now claimed
01:03:33 remaining
6:00 AM PST
Wagan Black Heated Seat Cushion
$29.95 $18.00 (40% off)
Add to Cart
Prime
24% now claimed
03:03:33 remaining
Page 2 of 67
Learn More
Black Friday Deals in Electronics
LG 42LD450 42-Inch
JBL Balboa 10 Two-Way
Flip UltraHD Video
Toshiba Satellite L655-
Gold Box New Deals. Every Day.
Deal of the Day
Canon PowerShot SX210 IS 14.1 MP Digital Camera with 14x Optical Zoom
$349.99 $199.00
Other Great Deals
Master Lock 22-Inch 9-Link Street Cuffs Lock
Fashion in Pearls Jewelry: Up to 70% Off
$290.00 $89.00
Instant Savings on Select LG HDTVs
Nikon Projector Camera
$349.95 $149.00
45% Off Garmin nüvi 265W/265WT 4.3-Inch Widescreen Bluetooth Portable GPS...
Black Friday Deals on Select LG Audio and Video Products
All Gold Box Deals
Get a BlackBerry for a Penny
Through November 29, all AT&T BlackBerry phones are on sale starting at a penny with no activation fee (restrictions apply).

# Amazon.com


P-1 B52 500-670
P-1 B53 500-676
P-1 B54 500-676
P-1 B55 500-676
P-1 B56 500-676
P-1 B57 500-676
P-1 B58 500-676
P-1 B59 500-676
P-1 B60 500-676
P-1 B61 500-676
P-1 B62 500-670
P-1 B63 500-670
P-1 B64 500-670
P-1 B65
BODMAN 91

# Amazon.com

# Amazon.com (cont.)


- an assembly line for the retail industry, with support for embedded online retailers
- huge stock of books, DVDs, CDs, *etc.*,, easily searchable
- extensive customer reviews

# Amazon.com (cont.)

Information-based differentiation: satisfies customers by providing a differentiated service:

- superior information including reviews about products
- superior range

Information-based delivery network: they deliver information for others; retailers in the Amazon marketplace get:

- customers directed to them
- other retailers' support

# Data Business Models

information brokering service: buys and sells data/information for others.

Information-based differentiation: satisfies customers by providing a differentiated service built on the data/information.

Information-based delivery network: deliver data information for others.

*"What a Big-Data Business Model Looks Like"* by Ray Wang in the Harvard Business Review claims these are unique in the data world.

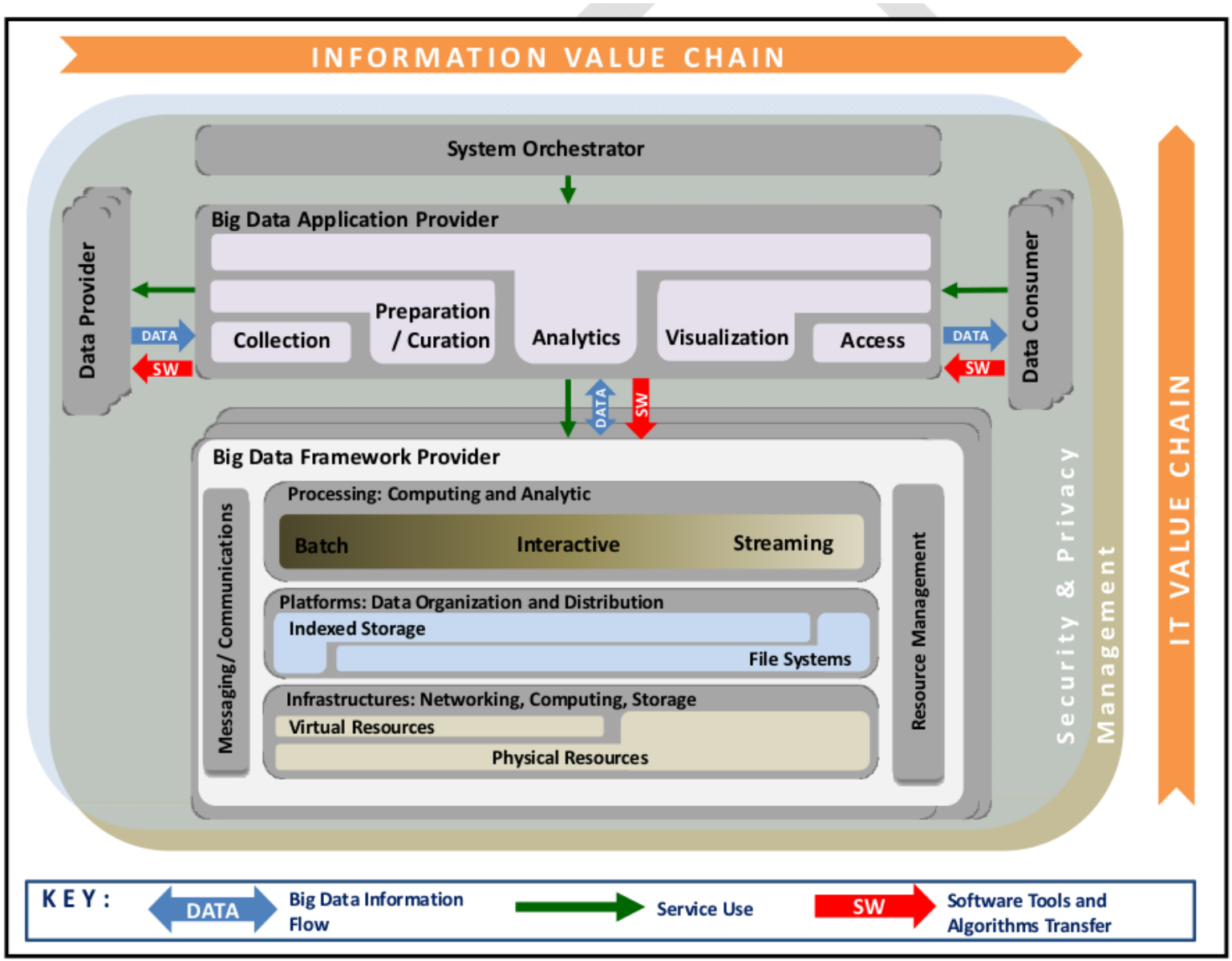
Data Science companies can pursue other business models, software as a service, consulting, CRM, *etc.*

# Data Providers

data provider ::= business selling the "data" it collects,
*e.g.*, Lexus-Nexus

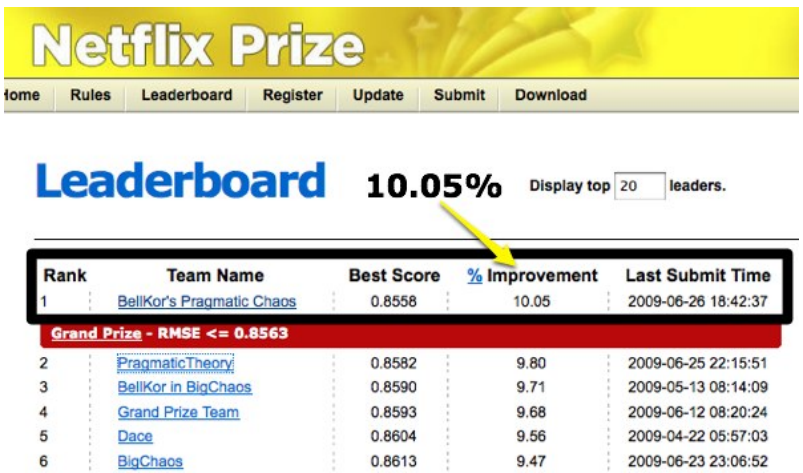
- this is a traditional business model, selling data not widgets
- so does not fit into Wang's categories (though is borderline "data broker")
- fasting growing segment of the IT industry post 2000 (see Evan Quinn's blog post on `Infochimps.com` April 2013 *"Is Big Data the Tail Wagging the Data Economy Dog?"*)
- some call this the data economy

# Value Chains



# Netflix: Example Case Study

- on demand internet streaming, and flat-rate DVD rental
- over 50 million subscribers in the US by 2014
- international market
- video recommendation!
- established the *Netflix Prize* in 2006-2009 as a crowdsourced way of testing out algorithms




*By Ivongala (Own work) [Public domain], via Wikimedia Commons*

# Netflix: Analysis

| | |
|---|---|
| data sources: | user rankings, user profiles |
| data volume: | (2012) 25 million users, 4 million rates/day, 3 million searches/day, video cloud stordage 2 petabytes |
| data velocity: | video titles change daily, rankings/ratings updated |
| data variety: | user rankings, user profiles, media properties |
| software: | Hadoop, Pig, Cassandra, Teradata |
| analytics: | personalised recommender system |
| processing: | analytic processing, streaming video |
| capabilities: | ratings and search per day, content delivery |
| security/privacy: | protect user data; digital rights |
| lifecycle: | continued ranking and updating |
| other: | mobile interface |

# Application Areas from MGI

The McKinsey Global Institute report on Big Data from 2011,
*"Big data: The next frontier for innovation, competition, and productivi*

1. Health
2. Government
3. Retail
4. Manufacturing
5. Location Technology

**NB.** What happened to Science? MGI is an industry organisation.

# Application Areas from NIST

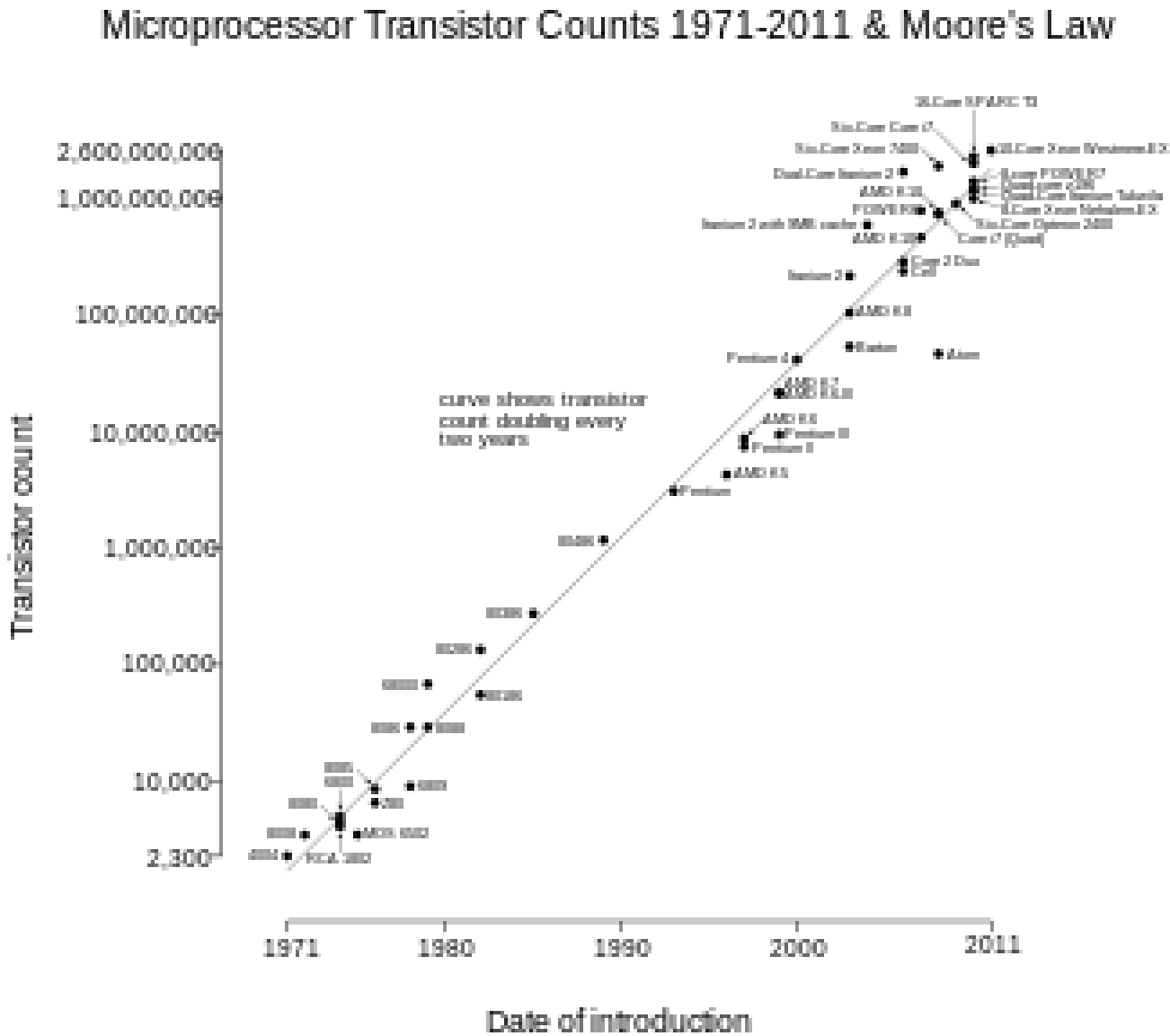
- government operation
- commercial
- defense
- healthcare and life sciences
- social media
- research infrastructure/ecosystem
- astronomy/physics
- earth science
- energy

# Outline

# Big Data

describing big data and its characterisation

# Moore's Law


Microprocessor Transistor Counts 1971-2011 & Moore's Law
Transistor count
2,600,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
2,300
curve shows transistor count doubling every two years
1971
1980
1990
2000
2011
Date of introduction

# Moore's Law

- stated to double every 2 years starting 1975
- transistor count translates to:
  - more memory
  - bigger CPUs
  - faster memory, CPUs (smaller==faster)
- pace currently slowing

# Big Data

From *Big data* on Wikipedia:

> *Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, ...*

- don't always ask why, can simply detect patterns
- a cost-free byproduct of digital interaction
- enabled by the cloud: affordability, extensibility, agility

# Big Data and "V"s

- 2001 Doug Laney produced report describing 3 V's:

  *"3-D Data Management: Controlling Data Volume, Velocity and Variety"*

- these characterise bigness, adequately
- other V's characterise problems with analysis and understanding

  Veracity: correctness, truth, *i.e.*. lack of ...
  Variability: change in meaning over time, *e.g.*, natural language

- other V's characterise aspirations

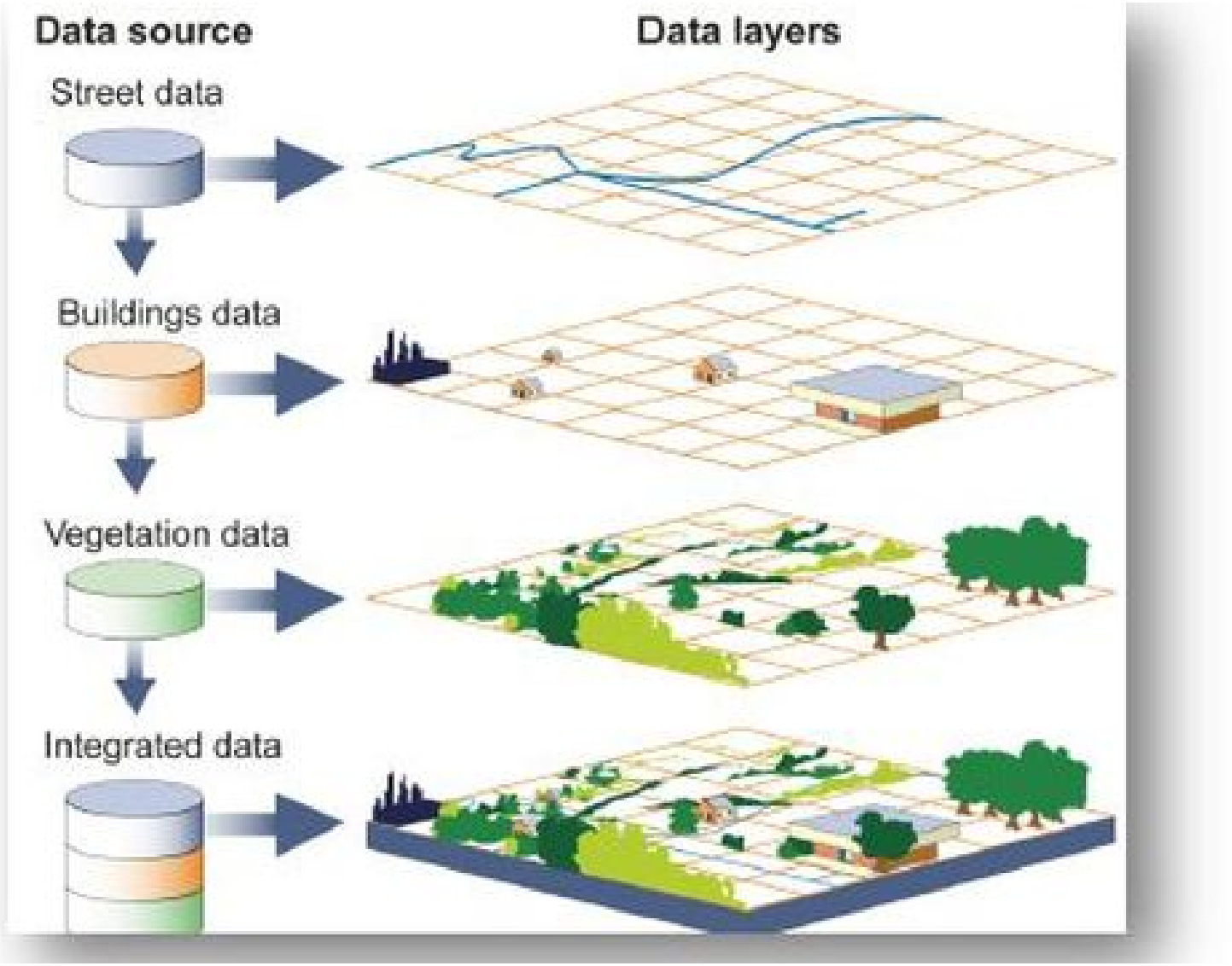
  Visualisation: one method for analysis
  Value: what we want to get out of the data

- think of any more? write a blog!

# Different Kinds of Data

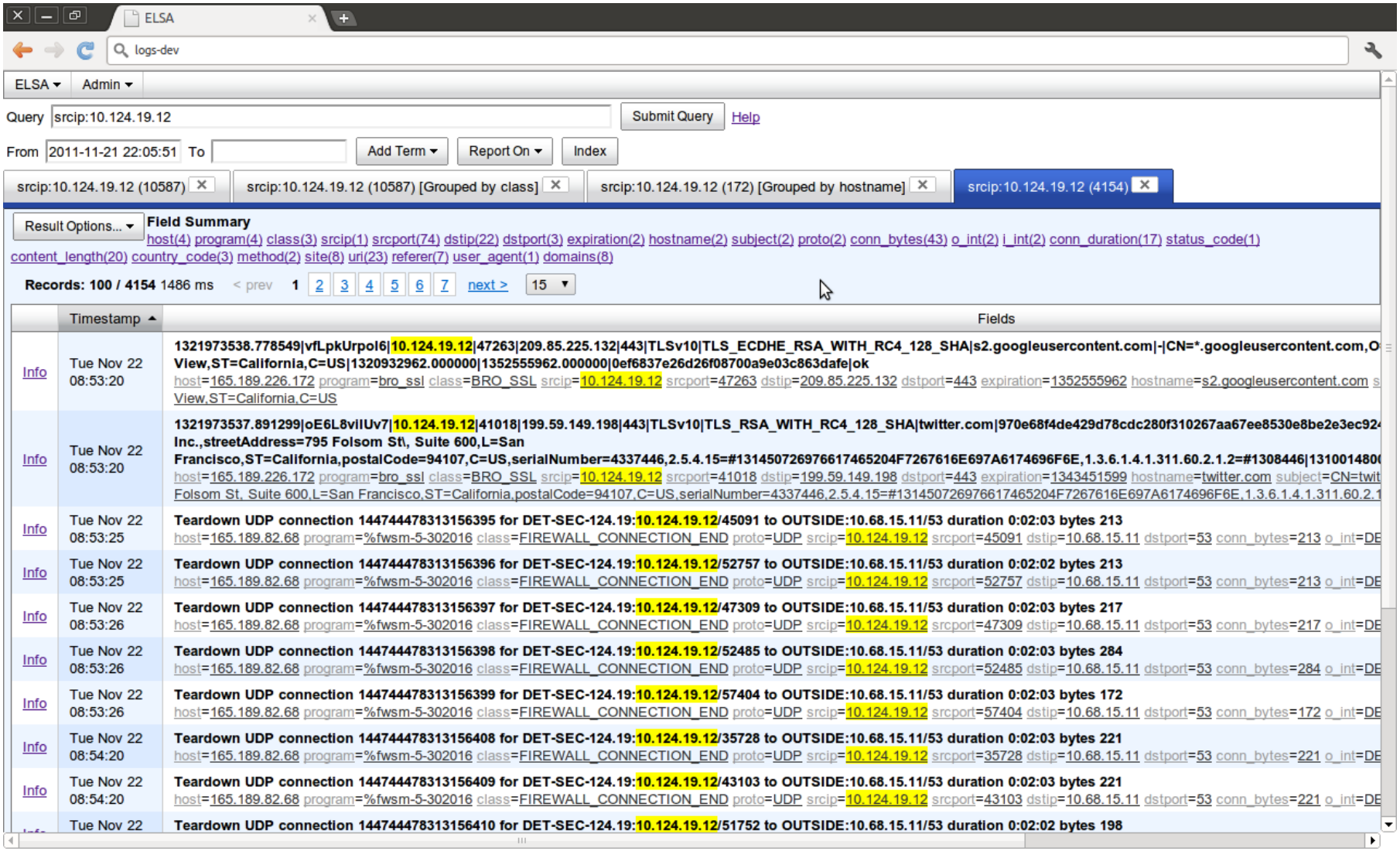
some examples

# Geospatial Data



# Linked Open Data: XML

# IP Connection Data



# Transactional Data

Trans

Entity: All Entity | Time: Default member: All Time Calendar | Account: All Account

| | Credit | Debit | Account | Entity | Transaction | Txn Date |
|---|---|---|---|---|---|---|
| All Transactions | $2,441,364.68 | $1,402,410.62 | | | | |
| # 501, BillPaymentCheck | $0.00 | $625.00 | Checking | Wheeler's Tile Etc. | BillPaymentCheck # 501 | 12/15/2012 |
| # none, Transfer | $0.00 | $500.00 | Savings | None | Transfer # none | 12/15/2012 |
| # none, ReceivePayment | $440.00 | $0.00 | Undeposited Funds | Roche, Diarmuid:Garage repairs | ReceivePayment # none | 12/15/2012 |
| # none, Bill | $0.00 | $670.00 | Accounts Payable | Keswick Insulation | Bill # none | 12/15/2012 |
| # 6236, PurchaseOrder | $0.00 | $65.00 | Purchase Orders | Daigle Lighting | PurchaseOrder # 6236 | 12/15/2012 |
| # 502, BillPaymentCheck | $0.00 | $640.92 | Checking | Daigle Lighting | BillPaymentCheck # 502 | 12/15/2012 |
| # 503, BillPaymentCheck | $0.00 | $754.50 | Checking | Patton Hardware Supplies | BillPaymentCheck # 503 | 12/15/2012 |
| # 1097, Invoice | $12,420.98 | $0.00 | Accounts Receivable | Robson, Darci:Robson Clinic | Invoice # 1097 | 12/15/2012 |
| # 504, BillPaymentCheck | $0.00 | $6,935.75 | Checking | Perry Windows & Doors | BillPaymentCheck # 504 | 12/15/2012 |
| # 505, BillPaymentCheck | $0.00 | $45.00 | Checking | Lew Plumbing | BillPaymentCheck # 505 | 12/15/2012 |
| # 12/03, Bill | $0.00 | $122.68 | Accounts Payable | Cal Gas & Electric | Bill # 12/03 | 12/15/2012 |
| # 506, BillPaymentCheck | $0.00 | $1,631.52 | Checking | East Bayshore Tool & Supply | BillPaymentCheck # 506 | 12/15/2012 |
| # 507, BillPaymentCheck | $0.00 | $1,358.00 | Checking | Timberloft Lumber | BillPaymentCheck # 507 | 12/15/2012 |
| # 508, BillPaymentCheck | $0.00 | $1,476.23 | Checking | East Bayshore Tool & Supply | BillPaymentCheck # 508 | 12/15/2012 |
| # 509, BillPaymentCheck | $0.00 | $450.00 | Checking | Hopkins Construction Rentals | BillPaymentCheck # 509 | 12/15/2012 |
| # 510, BillPaymentCheck | $0.00 | $896.00 | Checking | Timberloft Lumber | BillPaymentCheck # 510 | 12/15/2012 |
| # 511, BillPaymentCheck | $0.00 | $696.52 | Checking | East Bayshore Tool & Supply | BillPaymentCheck # 511 | 12/15/2012 |
| # 512, BillPaymentCheck | $0.00 | $400.00 | Checking | Patton Hardware Supplies | BillPaymentCheck # 512 | 12/15/2012 |
| # 513, BillPaymentCheck | $0.00 | $1,610.00 | Checking | Timberloft Lumber | BillPaymentCheck # 513 | 12/15/2012 |

# Twitter Data

**Brian D. Earp** @briandavidearp · 3h
Major publisher retracts 64 scientific papers in fake **peer review** outbreak - The Washington Post washingtonpost.com/news/morning-m…

6 4 View summary

**Christina Larson** @larsonchristina · 4h
Scientific publisher Springer retracts 64 papers - mostly by Chinese academics - for fake **peer review**: blogs.wsj.com/chinarealtime/… @feliciasonmez

3 4 View summary

**Felicia Sonmez** @feliciasonmez · 4h
A publisher has retracted 64 articles for fake peer reviews. Nearly all were from China. By me for @ChinaRealTime: on.wsj.com/1NQDVez

9 7 View summary

ALiEM

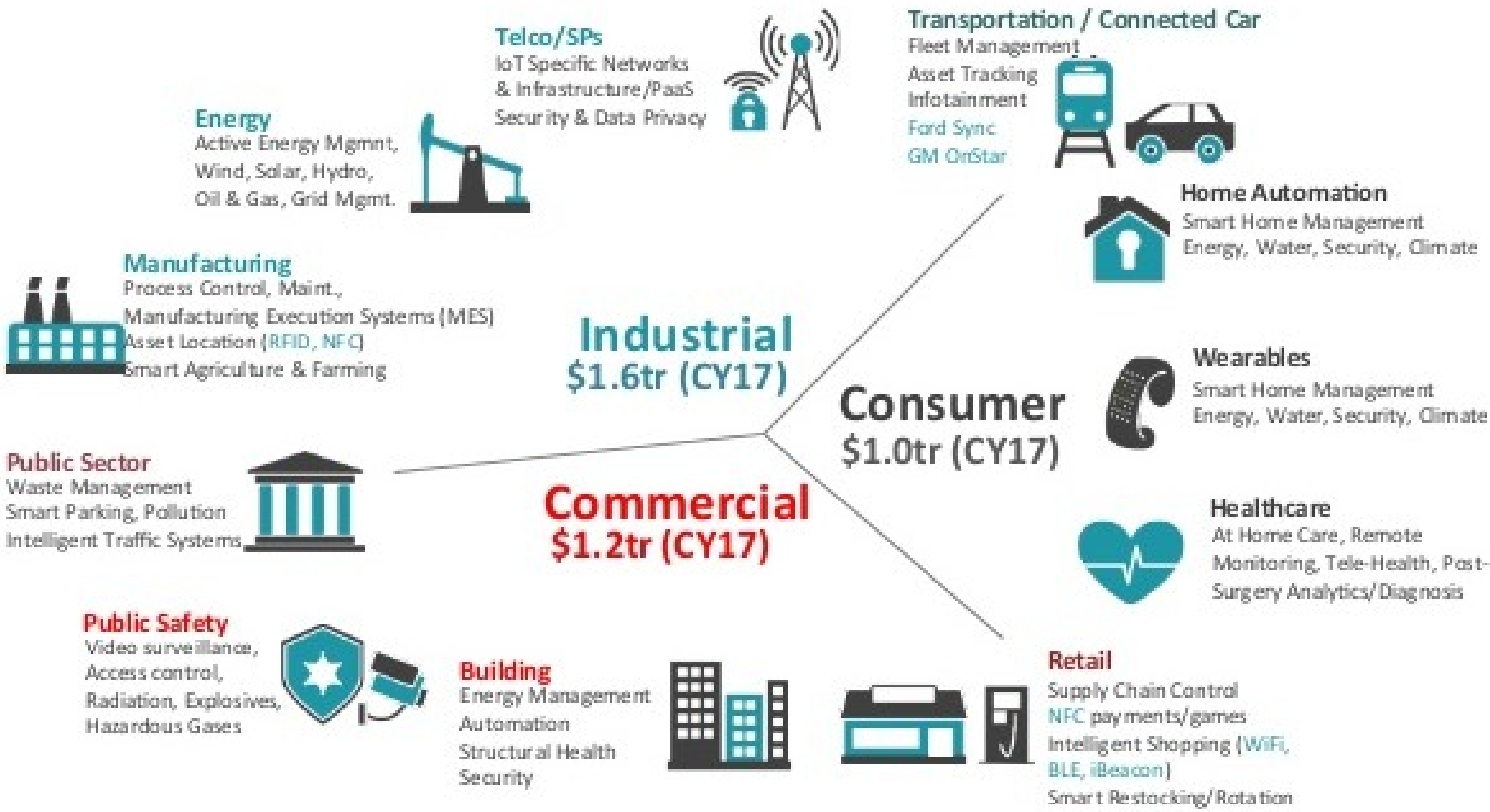
**Academic Life in EM** @ALiEMteam · 8h
CAPSULES module 2 is out! Pressors & Inotropes

By: @iEMPharmD & @DougEDPharm
**Peer-review**: @EMPharm & @DavidJuurlink

aliemu.com/courses/presso…

9 10

# Internet of Things Data




Source: IDC Internet of Things Spending Guide by Vertical Market 2014

# MetaData

metadata ::= structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource.

- is data about data
- a computer can process and interpret it

Descriptive: describes content for identification and retrieval
*e.g.* title, author

Structural: documents relationships and links
*e.g.* elements in XML, containers in MPEG

Administrative: helps to manage information
*e.g.* version number, archiving date, DRM

# Metadata Example

Let us look at examples to characterise the metadata:

- Australian Government
  *Digital Transformation Office, Service Standard* webpage
- *medical bibliographic data* in XML on PubMed,

  > *"Lower respiratory tract disorder hospitalizations among children born via elective early-term delivery"*

# Infographics on Data

- *"Data Science Matters"* from the datascience@berkeley Blog
- *"60 Seconds – Things That Happen On Internet Every 60 secs"* from GO-Gulf
- *"60 Seconds – Things That Happen Every 60 secs Part 2"* again

# Databases

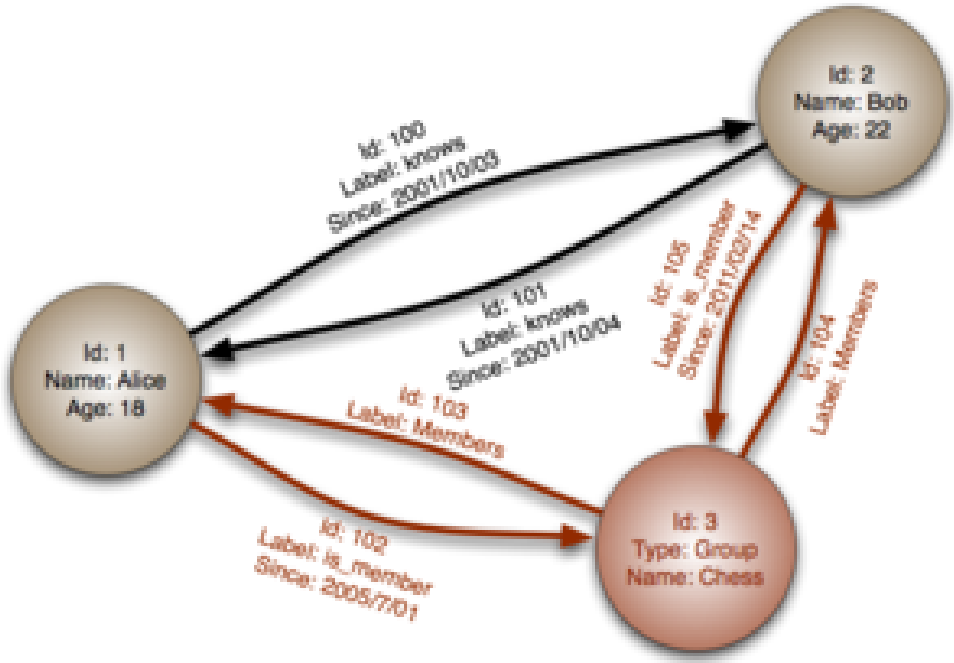
modern variations beyond SQL and RDBMS, and distributed systems

# JSON Example

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

- example from Wikipedia
- no fixed format
- semi-structured, key-value pairs, hierarchical
- “friendly” alternative to XML
- self-documenting structure
- example, *EventRegistry file*

# Graph Database Example




- example graph
- example content
  *FreeBase page for "Arnold Schwarzenegger"*
- example content format *FreeBase extract*
- stores graph, commonly as triples, `subject, verb, object`
- commonly used to store Linked Open Data

# Database Background Concepts

Many NoSQL and SQL DBs offer:

- large scale, distributed processing
- robustness achieved
- general query languages
- some notion of consistency

  *e.g.* "eventually" as nodes spread updates

# Beyond SQL Databases (NoSQL)

| Type | Examples | Notes |
|---|---|---|
| RDBMS | *MySQL*, *MSSQL Server* | SQL |
| Object DB | *Zope*, *Objectivity* | navigate network |
| Doc. DB | *MongoDB*, *CouchDB* | JSON like, Javascript like queries |
| key-val cache | *Memcached*, *Coherence* | in-memory |
| key-val store | *Aerospike*, *HyperDex* | not in-memory but highly optimised |
| tabular key-val | *Cassandra*, *HBase* | relational-like, "wide column store" |
| graph DB | *Neo4j*, *OrientDB* | RDF, SPARQL, |

# Beyond SQL Databases, cont.

- NoSQL databases offer a rich variety beyond traditional relational.
- Many target web applications.
- See blog post by Eric Knorr 19/11/2012 on `Infoworld.com`, *"The wild, crazy world of databases"*
- See blog post by Fabian Pascal 12/17/2015 on `AllAnalytics.com`, *"Data Fundamentals for Analysts: Documents and Databases"*.

# Overview: Processing

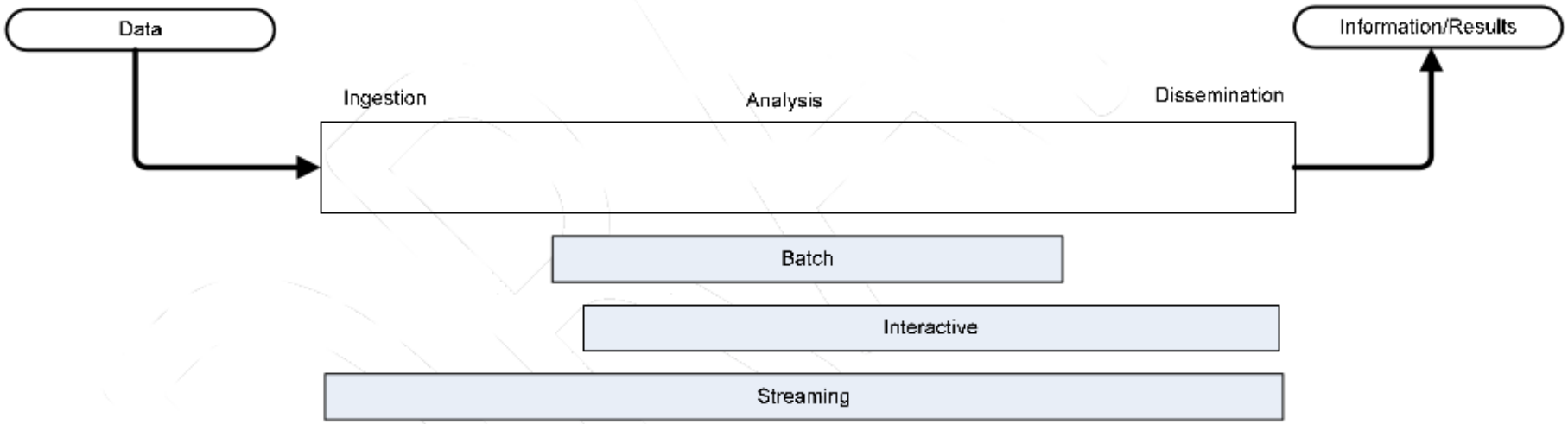



*Figure 5: Information Flow*

Interactive: bringing humans into the loop

Streaming: massive data streaming through system with little storage

Batch: data stored and analysed in large blocks, "batches," easier to develop and analyse

# Distributed Analytics

- legacy systems provide powerful statistical tools on the desktop
  - SAS, R, Matlab

  but often-times without distributed or multi-processor support
- supporting distributed/multi-processor computation requires special redesign of algorithms
- in-database analytics systems intended to support this

*e.g.* MADLib from Pivotal and MLLib from Spark integrates with their distributed SQL;

# Hadoop

- Java implementation of Map-Reduce developed by *Doug Cutting* while at Yahoo!
- architecture:

  Common: Java libraries and utilities

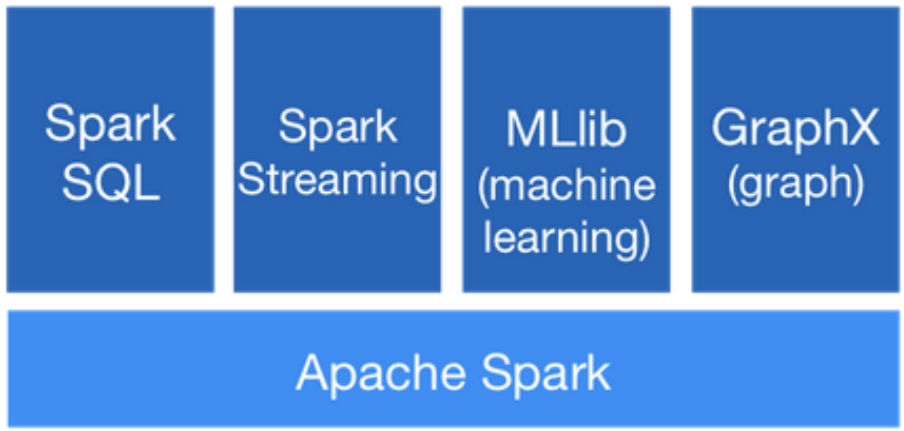
  YARN: job scheduling and cluster management

  Hadoop Distributed File System (HDFS™):

  MapReduce: core
- huge tool ecosystem
- well passed the peak of the hype curve

# Spark

- another (open source) Apache top-level project at *Apache Spark*
- developed at *AMPLab* at UC Berkeley
- builds on Hadoop infrastructure (HDFS, etc.)
- interfaces in Java, Scala, Python, R
- provides in-memory analytics
- works with some of the Hadoop ecosystem



# Applications

example application areas

# Case Study: Health Care Data

*When Health Care Gets a Healthy Dose of Data* "How Intermountain Healthcare is using data and analytics to transform patient care," June 25, 2015, Michael Fitzgerald

- 8000 word article in *Sloan Review MIT*
- behind "membership" (ask Wray if you want a copy)
- use of Electronic Health Records (EHR)
- *Intermountain Healthcare* has 22 hospitals and 185 clinics
- 2009 US government mandated "all health care providers adopt and demonstrate 'meaningful use' of EHR"
- promote data-driven decision making

# Health Care Data, cont.

- first computer support in 1985
- data quality and data gathering important

*e.g.* if you show physician their quality metrics are below average they say (1) “your data isn’t accurate” and (2) “my patients are different”

- need a common language for data across departments and hospitals
- big clinical teams (newborn, cardiovascular, ear nose & throat) have their own data manager and data analyst
- some nurses assigned to data recording roles
- approx. 10 analysts per hospital

# Health Care Data, cont.

- *e.g.* analysed lifestyle of diabetes patients with good blood sugar levels to understand factors and inform the team
- *e.g.* experimented with different practices in surgery (*e.g.* no personal clothing items in operating theatre), measure effect on infections after 6 months then keep/drop
- *e.g.* approx. US$40 million supply costs per hospital per year; analysed alternative items for use and cost to recommend which to use
- *e.g.* pull readings from patients' vital signs, sends an email alert telling patients at risk of heart failure
- *e.g.* give assessment of patient's likelihood of being readmitted to the hospital once released

# General Comments

- requires careful collection of standardised data
- embed analytics in processes rather than as an add-on, to allow feedback loops
- implementation pushback from staff over the effort
- analytics demonstrated improved patient care due to monitoring and improving processes
- main use is descriptive analytics not full data science
  - comparative evaluation of alternatives
  - data-driven process improvement
  - experimental evaluation of processes
- some predictive analytics
  - predicting readmittance
  - predicting heart failure

# Outline

# Getting Data

data sources and working with data

# NYC Data

NYC under Major Bloomberg embarked on a program to make the cities data accessible:

- *"How data and open government are transforming NYC"* in *Radar.O'Reilly*:

  > *"In God We Trust," tweeted New York City Mayor Mike Bloomberg this month. "Everyone else, bring data."*

- *Bloomberg signs NYC 'Open Data Policy'* into law, plans web portal for 2018," in *Engadget*
- *NYC Open Data portal*
- *City of Melbourne's open data platform*

# NYC Data, cont.

*"How we found the worst place to park in New York City"* is examples, and a discussion of the complexities of getting data out of NYC:

Map of road speed by day+time: GPS data for NYC cabs gives; data obtained via FOIL request, then made public by recipient

Danger spots for cycles: *NYPD crash data* obtained by daily download of PDF files followed by (non-trivial) extraction

Dirty waterways: *fecal coliform measurements on waterways* from Department of Environmental Protection's website; extracted from Excel sheets per site; each in a different format

Faulty road markings: parking tickets for fire-hydrants by location from *NYC Open Data portal* need to normalize the addresses supplied

# Traffic Prediction

see 7:40-11:06 on Clearflow in
*"Data, Predictions, and Decisions in Support of People and Society,"*
by Eric Horvitz

- forecasting traffic: blockages, clearing, surprising situations, alternate routes
- critical data:
  - GPS data on traffic flow
  - maps
  - incidents and events
  - weather
- see *Microsoft Introduces Tool for Avoiding Traffic Jams* in *NYT* 2008

# Democratization of Data

*"The New Data Republic: Not Quite a Democracy"* in MIT Sloan Review 2015

- from Hal Varian: "information that once was available to only a select few. . . available to everyone"
- from Robert Duffner: "finally puts crucial business information in the hands of those who need it"
- government and IT departments building data and infrastructure to allow sharing
  - *USA Open Gov Initiative*
- analytic tools, desktop and web-based, available to analyse it
- but people need the right skills

open data is all good and well, but people need to be able to use it too!

# Linked Open Data

LOD project started by
*Prof. Sir Tim Berners-Lee, OM, KBE, FRS, FREng, FRSA, DFBCS*.

- objects given a URI (like a URL)
- relationships between two objects can be represented as a triple, (subject, verb, object)
- relation itself is another URI
- data has an open license for use

*e.g.* *example on NYT*

- a *tutorial on LOD* by Tom Heath

# Wrangling

manipulating data to make it directly usable for analysis

# Wrangling Examples

- want the core news text, title, date, *etc.* off the following page *Apple's iPhone loses top spot to Android in Australia*
- want the text plus details from the PDF file *"Data Wrangling: The Challenging Journey from the Wild to the L*
- want all article titles from the *PUBMED results xml*
- want to digitize the text off a *scanned letter*
- want to extract all the sentences referring to Hillary Clinton in *a news article*

# Wrangling Examples, cont.

- your company has customer records in 4 different databases in different formats; you want a single standardised set of customer names and addresses
- convert addresses in your customer database into geographic latitude and longitude
- convert free text dates to standard format, *e.g.* “next Tuesday”, “2nd January 15”, “January 3 next year”, “3rd Friday in the month”, “03/31/15”, “31/03/15”
- recognise what values in your data are “unknown” or “illegal”

# Introduction to Resources
# Standards

support cooperation, reuse, common tools, *etc.*

# Example Standards

- metadata such as *Dublin Core*
- XML formats for sharing models, PMML (see below)
- standards for the data mining/science process, such as *CRISP-DM*
- health codes: disease and health problem codings *ICD-10*
- systematized nomenclature of medicine, clinical terms, *SNoMed-CT*

What other sorts of things might you have standards for?

# Data Science Process

- using our own “standard Data Science value chain” to describe the process
- *CRISP-DM* discussed previously
- statisticians sometimes use the term exploratory data analysis for part of the process
- while not a standard„ one can take this sort of specification to the extreme: see *“Data Science life-cycle”*

# The API Economy

- *The Application Economy: A New Model for IT* (CISCO)
- *The Application Economy Is Changing the Future of Business*
- *ProgrammableWeb API Category: Data*
- *Top 30 Predictive Analytics API*

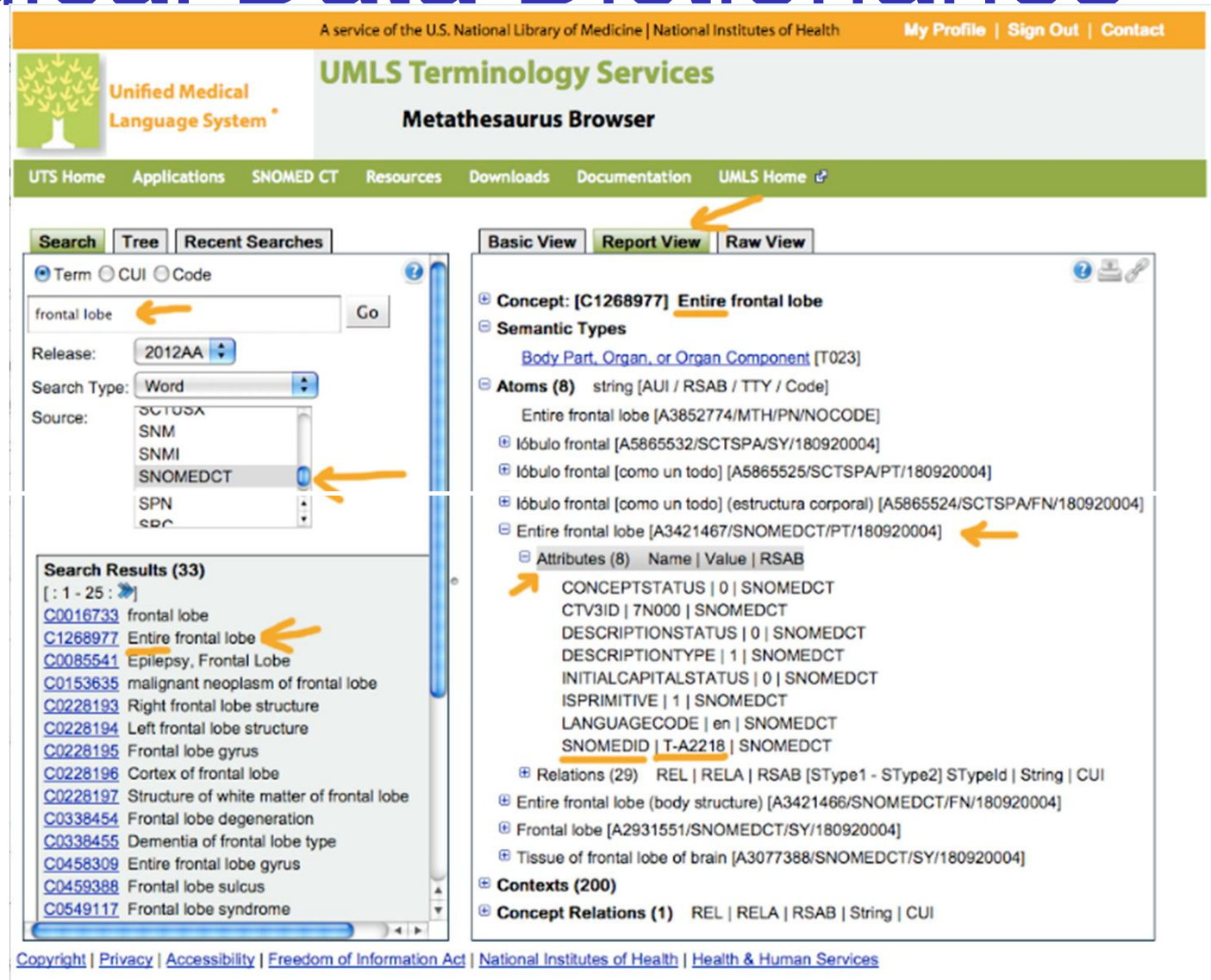
# Case Studies of Data and Standards

look at some examples of standardised data collections

# Freebase

- an example of a graph database we looked at earlier
- graph can be represented in RDF which is triples of URIs
- *Freebase*, now owned by Google, currently read-only and to be decommissioned soon
- used by others as a knowledge-base in knowledge language processing, *e.g.*, *TextRazor*, “extract meaning from your text”
- see also *DBpedia*

# Medical Data Dictionaries




*The Unified Medical Language System (UMLS)*

# Medical Data Dictionaries, cont.

ICD: the **I**nternational **C**lassification of **D**iseases

- used ... to classify diseases and other health problems ... on ... health and vital records

example: *Pneumonia due to Streptococcus pneumoniae*

# Publishing Repositories

- PUBMED, we have seen before
- *ACM Digital Library*
- Patent databases (for WIPO, USPTO, EPO, *etc.*), *e.g.*, *Global Patent Search Network*

# News and Event Registry

- collect news article globally, process and organise as events
- perform concept and event identification
- create a document database for inspection
- *Event Registry*
- sometimes news stored as *NewsML*

# Government Data

- US Government's *Data.GOV*
- *NYC Open Data*
- *Australia's Urban Intelligence Network (AURIN)*
- *BioGrid Australia*

# Outline

# Essential Viewing

- *"The wonderful and terrifying implications of computers that can learn"* at TED by Jeremy Howard
- *"The Unreasonable Effectiveness of Data"* lecture at Univ. of British Columbia by Peter Norvig
- *"Knowledge is Beautiful"* by David McCandless at the RSA
- *"The power of emotions: When big data meets emotion data"*, by Rana El Kaliouby

# Types of Data Analysis

*"Six types of analyses every data scientist should know"*, by Jeffrey Leek

- extends SAS's "analytic levels" with a some nuances
- introducing inference and causality

# Tools for the Data Analysis Process

popular software and prototyping

# Common Software

access: SQL, Hadoop, MS SQL Server, PIG, Spark

wrangling: common scripting languages (Python, Perl)

visualisation: Tableau, Matlab, Javascript+D3.js

statistical analysis: Weka, SAS, R

multi-purpose: Python, R, SAS, KNIME, RapidMiner

cloud-based: Azure ML (Microsoft), AWS ML (Amazon)

# Mapping Big Data

See *"Mapping Big Data: A Data-Driven Market Report"* by Russell Jurney, published by O'Reilly 2015. See Table 1-6.

| **Cluster** | **Company** |
|---|---|
| Old Data Platforms | IBM, Microsoft, Oracle, Dell, Netapp |
| Servers | Intel, SUSE, MSC Software, NVidia, Redline Trac |
| Analytic Tools | Tableau, Teradata, Informatica, Talend, Actian |
| New Data Platforms | Cloudera, Hortonworks, MapR, Datastax, Pivotal |
| Enterprise Software | HP, SAP, Cisco, VMWare, EMC |
| Cloud Computing | Amazon Web Svcs., Google, Rackspace, MarkLo |

Note the Enterprise Software segment developing good connections with all others, but already has strong connections with Old Data Platforms.

# Scripting Languages

see Wikipedia entry *scripting languages*:

- no formal or universally agreed definition
- often interpreted and are high-level programming languages
- automating tasks originally done one-by-one by hand
- also, extension language, control language

e.g. bash, Perl, Python, R, Matlab, ...

kinds: glue languages (connecting software components), GUI scripting, job control, macros, extensible languages, application specific, ...

- an *endless discussion on StackExchange*

# Rapid Prototyping

see Wikipedia entry *software prototyping*:

- software development for data science projects is often (almost) one-off ... get the results, but ensure it is reproducable
- not standard software engineering, not “waterfall model”, not “agile”
- little requirements analysis
- the results are tested, not the software and its full capability
- development speed and agility are important
- hence use of scripting languages

# Discussion: Python versus R

- both are free
- R developed by statisticians for statisticians, huge support for analysis
- Python by computer scientists for general use
- R is better for stand-alone analysis and exploration
- Python lets you integrate easier with other systems
- Python easier to learn and extend than R (better language)
- R has vectors and arrays as first class objects; similar to Matlab!
- R currently less scalable.

See *In data science, the R language is swallowing Python* by Matt Asay, recent blog in *Infoworld*.

# Scientific Method

is Data Science writ large, so what can we learn

# Scientific Method in Medicine

- *"How science goes wrong"* on *The Economist*, 2013
- *"Battling Bad Science"* a TED talk by Ben Goldacre, 2011
- *"The Truth Wears Off"* by Jonah Lehrer in *The New Yorker*, 2010
- *"Richard Smith: Time for science to be about truth rather* than careers" blog on *BMJ*, 2013
- *"Offline: What is medicine's 5 sigma?"* by Richard Horton on *The Lancet*, 2015
- *"The 10 stuff ups we all make when interpreting research"* by Will J Grant and Rod Lamberts in *The Conversation*, 2015.

Broadly:

- ~~industry coercion~~
- ~~academic games~~
- errors in application of scientific method

# Scientific Method in Medicine

Major applications errors are:

- misuse of significance testing
- correlation does not imply causation
- not checking/testing the true costs
- inadequate reproducibility *e.g.*, difficult to repeat
- selection bias

# Significance Testing Errors

**Significance chasing:** repeat many experiments until you get significance

- *"I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How."* by John Bohannon,

**In parallel:** multiple teams trying different experiments until one gets significance

**Ignoring negative results:** a variation on the above; similar to repeated testing until success

**The decline effect:** a variation on the above, as *some* negative results get recorded, eventually the original (flawed) positive result gets overturned

**Inadequate repeatability:** means subsequent teams cannot check your results, so you're inital inadequate significance testing doesn't get retested

# Significance Testing

- be careful with P-values and significance levels: use strong significance levels and don't "repeat until success"
- record negative results
- ensure repeatability by properly recording experimental methodology and data processing

# Error: Correlation versus Causation

- See *correlation does not imply causation* (Wikipedia).
- also *Hilarious Graphs*
- happens when medical experts use observational data to draw conclusions, *e.g.*, epidemiological data
- methods for testing/estimating causation from data is currently a research agenda in discovery science
- "intervention" is a basic part of *double blind trials* (a major experimental standard)

# Data Analysis Meta Case Studies
# What is Hard?

comments

# The Hardest Parts

See blog *"The hardest parts of data science"* by Yanir Seroussi 23rd Nov. 2015.

**Model fitting:** core statistics/machine learning not usually hard (*e.g.,* many use R as a black box for this)

**Data collection:** can be critical sometimes, but often more routine

**Data cleaning:** can be a lot of work, but often more routine

**Problem definition:** getting into the application and understanding the real problem can be hard

**Evaluation:** what is measured? should multiple evaluations be done? can be hard

**Ambiguity and uncertainty:** invariably these occur and we need to live with them; can be hard

# Outline

# Issues

presentation of basic issues

# Terminology

- Privacy is (for our purposes) having control over how one shares oneself with others.
  - e.g. closing the blinds in your living room
- Confidentiality is information privacy, how information about an individual is treated and shared.
  - e.g. excluding others from viewing your search terms or browse history
- Security is (for our purposes) the protection of data, preventing it from being improperly used
  - e.g. preventing hackers from stealing credit card data
- Ethics is (for our purposes) the moral handling of data (especially, other data about others)

# Regulations and Compliance

- Regulations devised by various government bodies: taxation, medical care, securities and investments, work health and safety, employment, corporate law.
- they need to check companies for their compliance
- Auditing

  *systematic and independent examination of books, accounts, documents and vouchers of an organization to ascertain how far they present a true and fair view*

- Regulatory compliance:

  *that organisations ensure that they are aware of and take steps to comply with relevant laws and regulations.*

- auditing data and records are a good source for Data Science

# Data Governance

Supporting and handling:

- ethics, confidentiality
- security
- regulatory compliance
- organisation policies
- organisation business outcomes

which may include handling the steps in the data science and/or big data value chain

# Data Management

managing to achieve governance, *etc.*

# Data Management

Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

# Data Management and Data Science

**medical informatics:** for predicting fungal infections from nursing notes, the team needs to abide by confidentiality and security

**internet advertising:** what implicit and explicit data is stored about a user

**retailing:** conduct market intelligence on new products; put together data from different divisions, brands

**predictive medical system:** implementation may need changing standard operating procedure for staff

# Contexts for Data Management

**Science:** reproducibility and credibility of scientific work, producing artifacts of knowledge, creating scientific data

**Business:** governance, compliance, information privacy, *etc.*

**Curation:** *e.g.* museums and libraries, preservation, maintenance, *etc.*

**Government:** a unique regulatory environment (*e.g.*, "transperancy"), archiving, FOIs, support data infrastructure, *etc.*

**Medicine:** significant privacy issues, conflicting corporate financial constraints, government regulations and furthering of medical science

# Digital Curation Centre

About:

> *The Digital Curation Centre (DCC) is a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management across the UK's higher education research community.*

See *"The DCC Curation Lifecycle Model"* by DCC (PDF)

# Australian Public Service

Background:

> *the creation, collection, management, use and disposal of agency data is governed by a number of legislative and regulatory requirements, government policies and plans*

- data needs to be authentic, accurate and reliable
- strong governance framework
- sensible risk management and a focus on information security, privacy management
- clear and transparent privacy policies and provide ethical leadership

# Conclusion


Data Science and Data in Society

Data Models in Organisations

Data Types and Storage

Data Resources, Processes, Standards and Tools

Data Analysis Process

Data Curation and Management