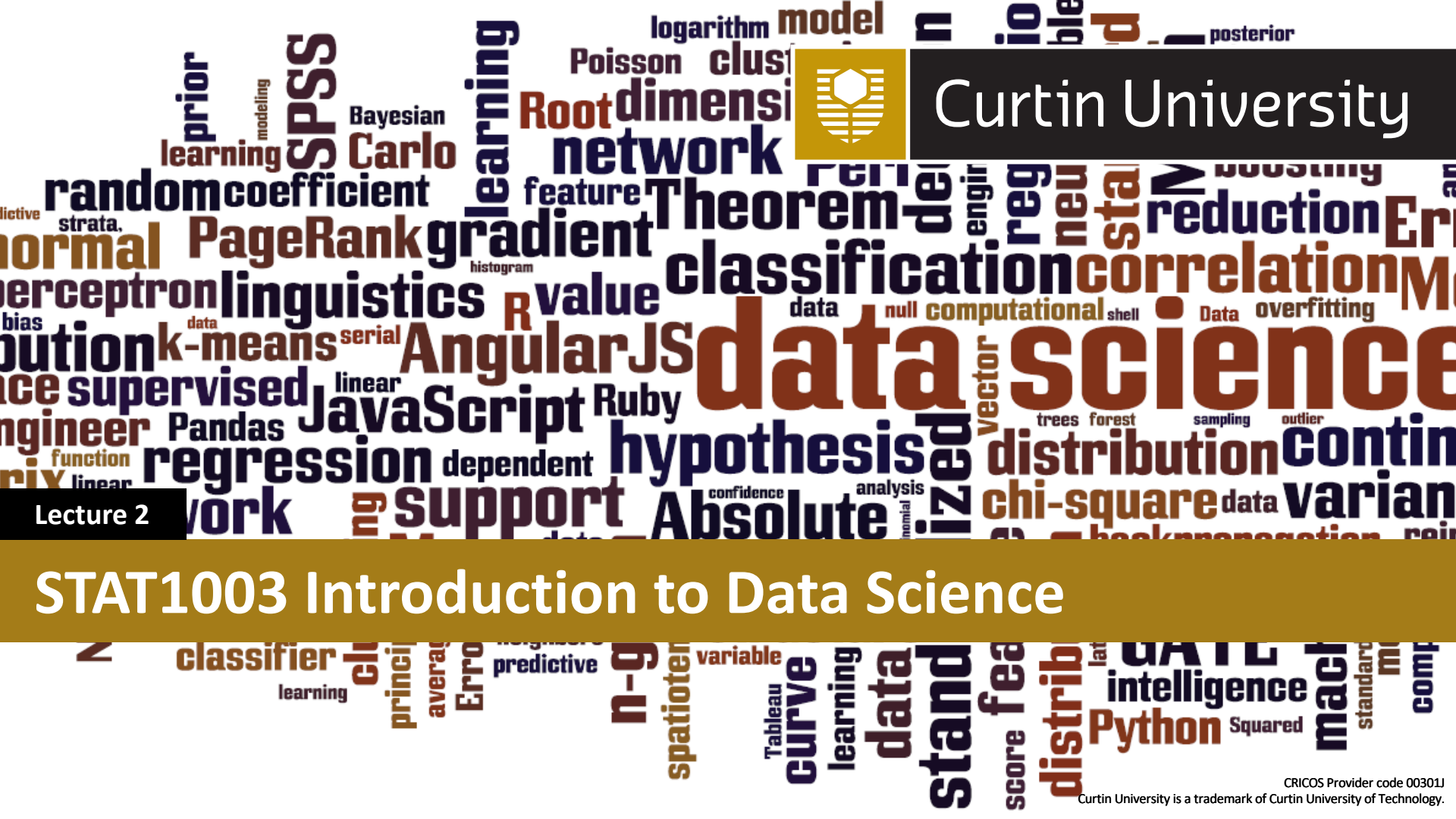


Outline

- Loose ends from Lecture 1
- Data
  - What is it?
  - Old and new sources of data
  - Big data and new problems

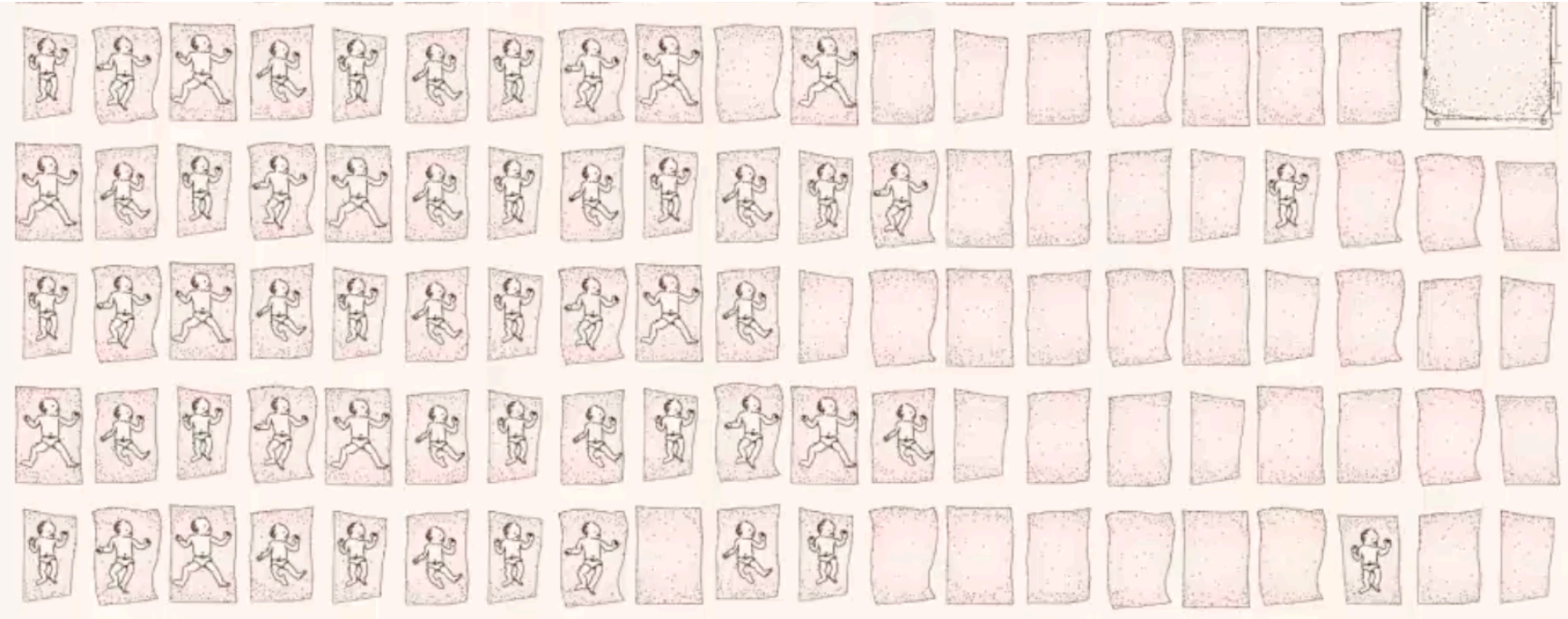
Characterizing data science

- Data is everywhere, and everyone thinks they need to derive useful information from it, but
  - No neat definition of data science – *yet!*
- New sources of data, ways of obtaining it, methods for analysing it, visualising and communicating results

Distinguishing features of data science

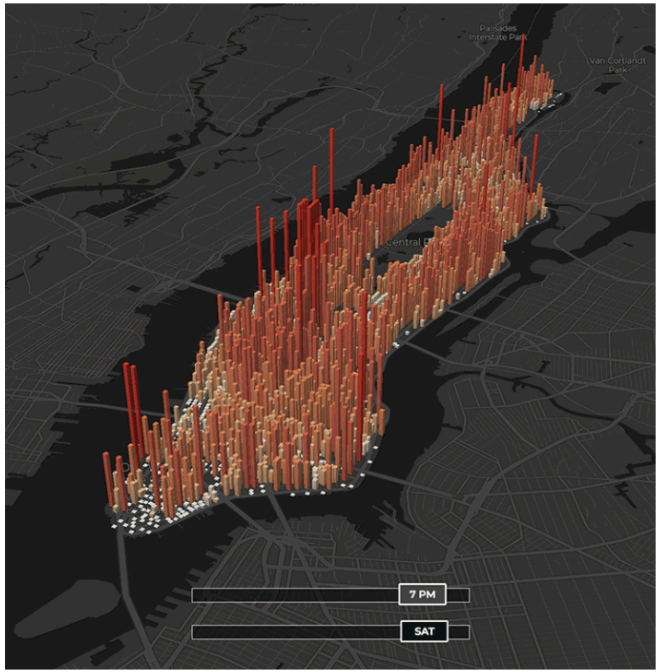
- ‘Large’ data sets
- Passively measured or observed, in contrast to active experimentation
- Focus on predictive models over models that seek to model the data generative process
- Emphasis of new ways of visualizing data, especially dynamic visualization
- Obtaining, analyzing, predicting, and visualizing in real-time

You decide Australia's population

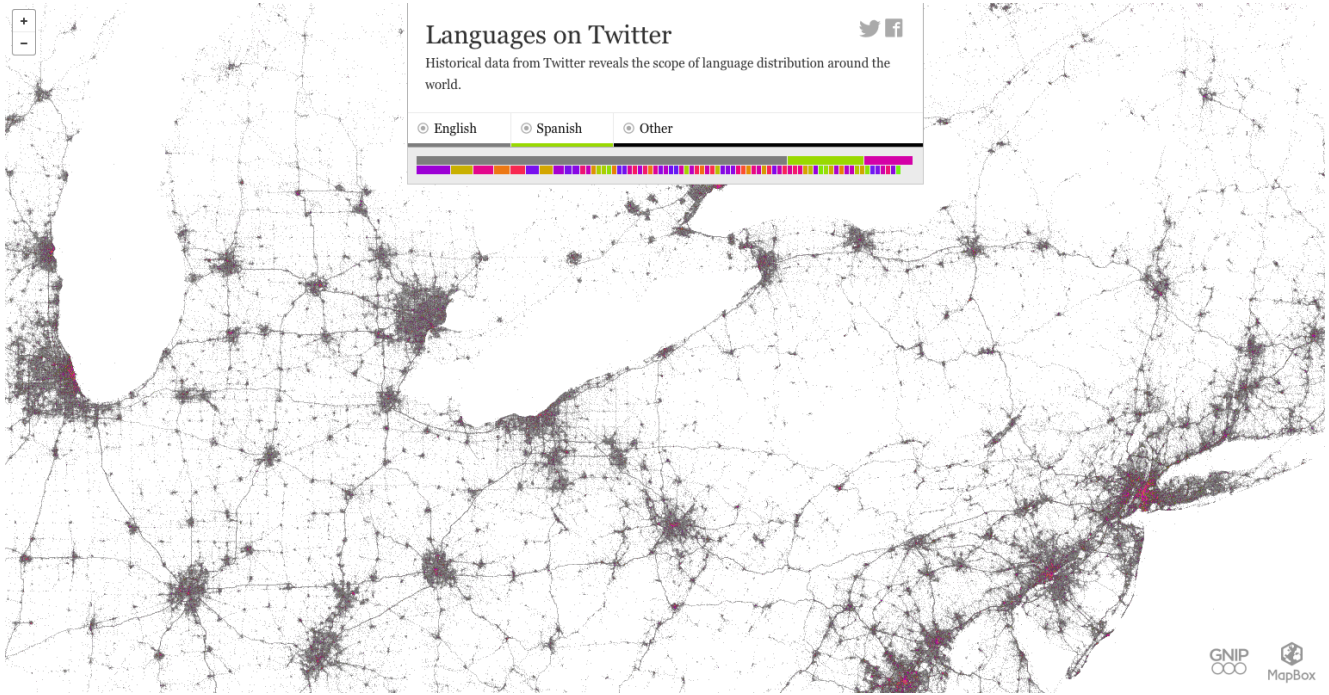


Curtin University

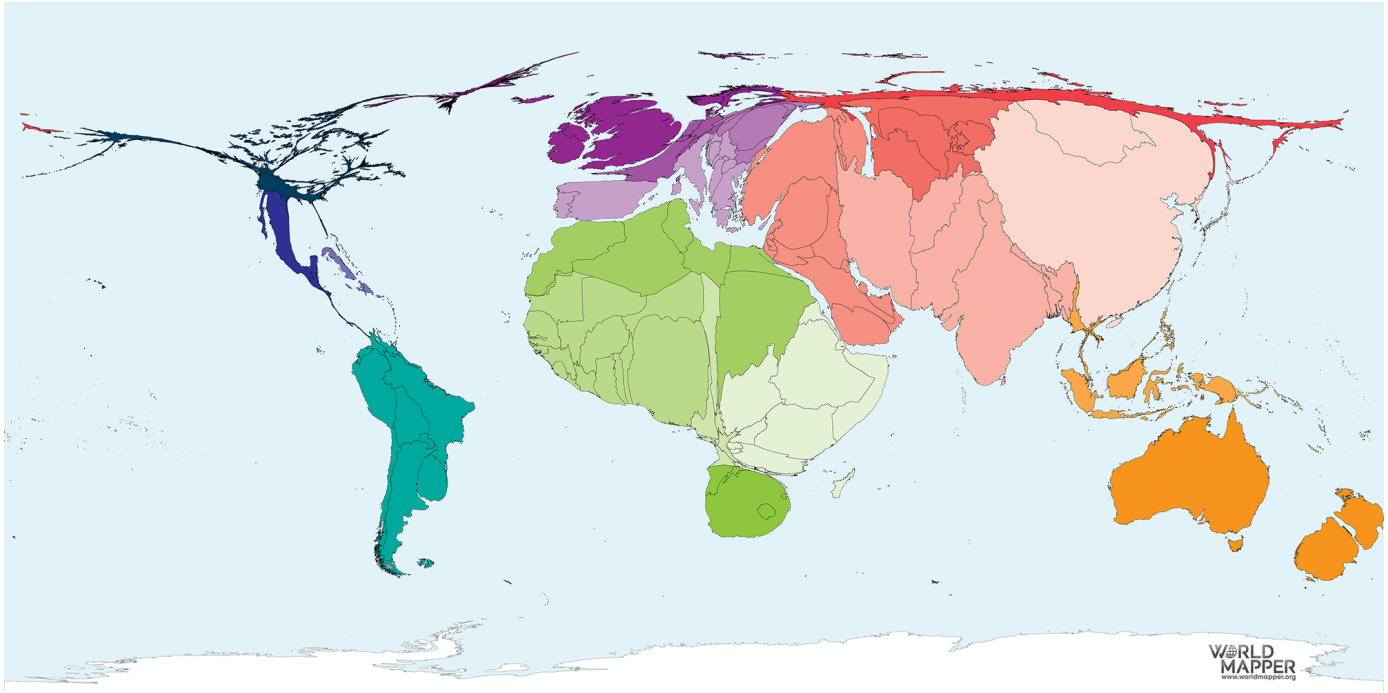
The city is alive: Manhattan, Hour-by-Hour



Language distribution from Twitter



Cartograms: number of sheep as a mapping variable



Teaching computers to draw manga and anime

MangaGAN

Teaching computers to draw new and original manga and anime faces with DCGANs

TD Follow

Dec 10, 2017 · 8 min read

Manga and anime faces generated with a model trained for 100 epochs

DATA

Data: what is it?

CURRENT ISSUE | ARCHIVE | SUBSCRIBER SERVICES | ABOUT GO BROWSE BY: TOPIC | AUTHOR

Welgos/Frederic Lewis/Hulton Archive/Getty

Why Data Is Never Raw

On the seductive myth of information free of human judgment

Nick Barrowman

A curious fact about our data-obsessed era is that we're often not entirely sure what we even mean by "data": Elementary particles of knowledge? Digital records? Pure information? Sometimes when we refer to "the data," we mean the results of an analysis or the evidence concerning a certain question. On other occasions we intend "data" to signify something like "reliable evidence," as in the saying "The plural of anecdote is not data."

In everyday usage, the term "data" is associated with a jumble of notions about information, science, and knowledge. Countless reports marvel at the astonishing volumes of data being produced and manipulated, the efficiencies and new opportunities this has made possible, and the myriad ways in which society is changing as a result. We speak of "raw" data and laud it for its independence from human judgment. On this basis, "data-driven" (or "evidence-based") decision-making is widely endorsed. Yet data's purported freedom from human subjectivity also seems to allow us to invest it with agency: "Let the data speak for itself," for "The data doesn't lie."

Out of this quizzical mix, it is perhaps unsurprising that near-magical thinking about data has emerged. In the 2015 book *Digital Destiny: How the New Age of Data Will Transform the Way We Work, Live, and Communicate*, Shawn DuBravac describes a collection of "properties of data" and expresses them in anthropomorphic terms. DuBravac, former chief economist at the Consumer Electronics Association and

This article appears in the **SUMMER/FALL 2018** Issue of The New Atlantis

PDF version
Printer-friendly
Buy this issue
E-mail this page

Related articles

- Tafari Mbadiwe, "**Algorithmic Injustice**," Winter 2018
- Nick Barrowman, "**Correlation, Causation, and Confusion**," Summer/Fall 2014

Related topics

- Technology and Culture
- Big Data

- ?

Curtin University

Data: what is it?

- We often speak of “raw” data and see it as independent of human judgement
  - (But is it?)
- This “independence” is the source of “data-driven” decision-making
- The purported freedom of data from human subjectivity invests it with agency: “Let the data speak for itself”, or “The data doesn’t lie”

Big Data – new paradigm for science?

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

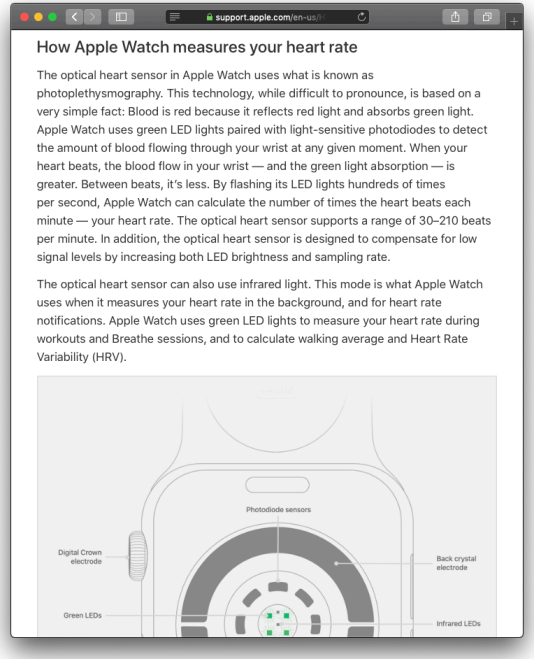
- "We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot"

Data: what is it?

- However, …
- How data are recorded and collected is the result of human decisions about what to measure, when and where to do so, and by what methods
- ”Raw” data signifies that no processing took place following data collection, but the term obscures the steps that took place before even this data was collected
- Barrowman (2018): “All data is cooked”
- True in scientific as well as sociological contexts

Example: measuring heart rate

How to check your heart rate



How Apple Watch measures your heart rate

The optical heart sensor in Apple Watch uses what is known as photoplethysmography. This technology, while difficult to pronounce, is based on a very simple fact: Blood is red because it reflects red light and absorbs green light. Apple Watch uses green LED lights paired with light-sensitive photodiodes to detect the amount of blood flowing through your wrist at any given moment. When your heart beats, the blood flow in your wrist — and the green light absorption — is greater. Between beats, it's less. By flashing its LED lights hundreds of times per second, Apple Watch can calculate the number of times the heart beats each minute — your heart rate. The optical heart sensor supports a range of 30–210 beats per minute. In addition, the optical heart sensor is designed to compensate for low signal levels by increasing both LED brightness and sampling rate.

The optical heart sensor can also use infrared light. This mode is what Apple Watch uses when it measures your heart rate in the background, and for heart rate notifications. Apple Watch uses green LED lights to measure your heart rate during workouts and Breathe sessions, and to calculate walking average and Heart Rate Variability (HRV).

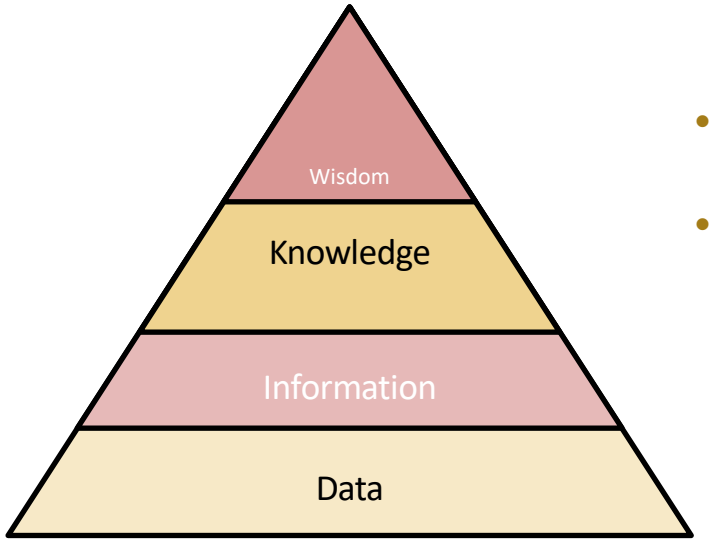
Unreliable data

- Data that is collected for one reason but is used for another
- Missing births
- Data that has been processed but is stored without any metadata about that processing
- … any other examples?

Data to knowledge

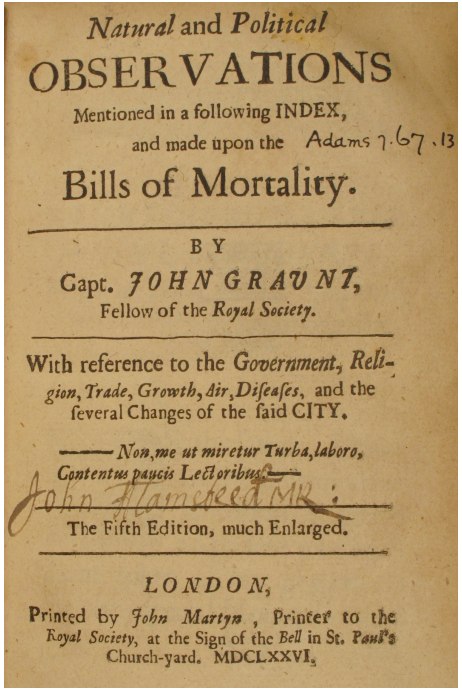
- “Information and knowledge are what we want, data is what we get.”
  - What’s data?
  - What’s information?
  - What’s knowledge?
- Can we generalize these terms beyond the specific example?

DIKW hierarchy (Rowley, 2007)



- ‘Data are discrete, objective facts or observations, which are unorganized and unprocessed, and do not convey any specific meaning’
- ‘Information is data which adds value to the understanding of a subject’
- ‘Knowledge is data and information that have been organized and processed to convey understanding, experience, accumulated learning, and expertise as they apply to a current problem or activity’

Observational and experimental data



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the Adams 7.67.13

Bills of Mortality.

BY

Gapt. JOHN GRAUNT,

Fellow of the *Royal Society*.

With reference to the *Government, Religion, Trade, Growth, Air, Diseases*, and the several Changes of the said CITY.

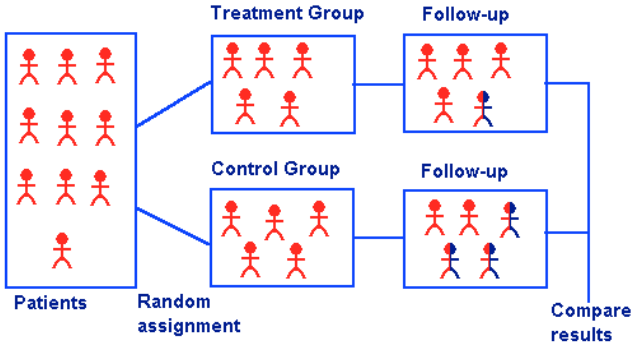
——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

John Flamsteed MR:

The Fifth Edition, much Enlarged.

LONDON,

Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the *Bell* in St. *Paul's* Church-yard. MDCLXXVI.



Observational and experimental data

- Experimental data: in properly conducted experiments, we can ascribe causality as in clinical trials, for example
  - Experiments conducted using a *sample* from a *population*
- Observational data: passively observed or measured; cannot ascribe causality

“correlation is not causation”

- Big data: we happen to have a **lot** more observational data now:
  - Velocity, variety, and volume

Data sources

www.kofax.com

Old and ‘new’ data

Traditional

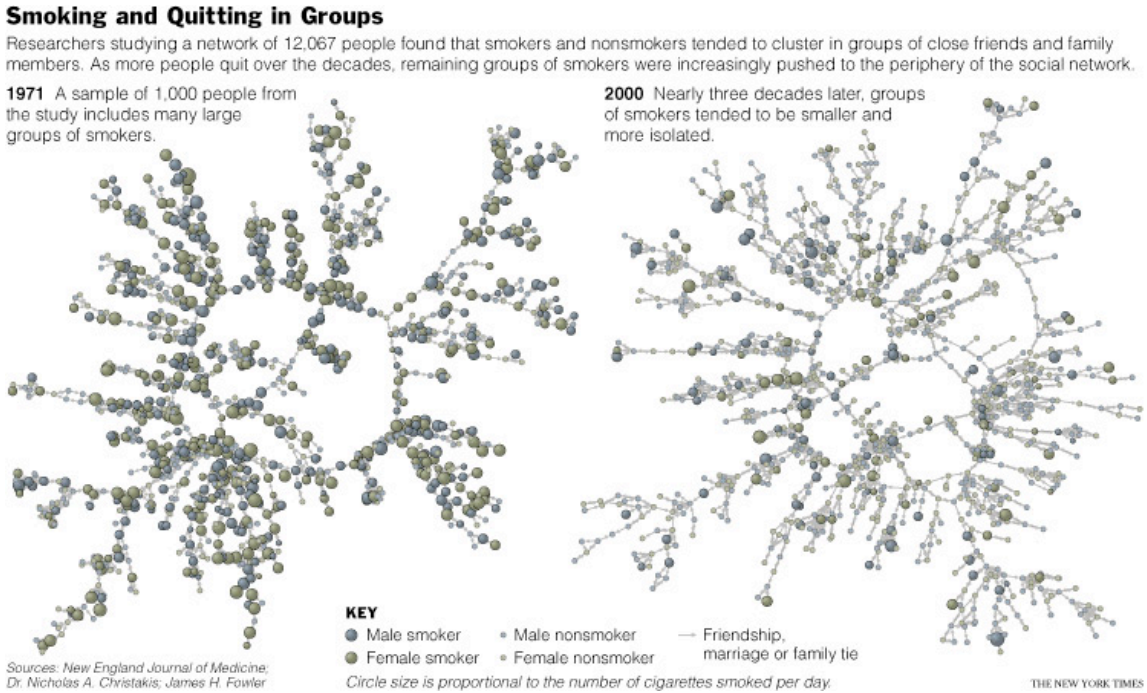
- ‘small’ datasets of numerical, categorical data

New

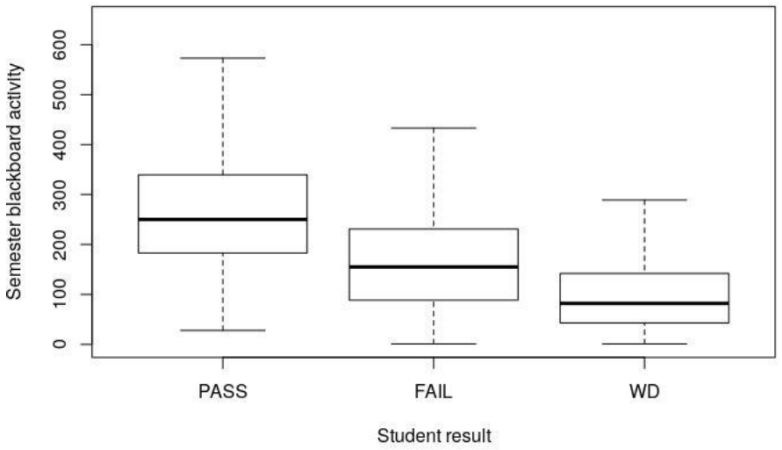
- Text: emails, tweets, online articles
- Records: user-level data, time-stamped event data
- Geo-based location data
- Network data
- Sensor data
- Images

Datafication: “the ability to render into data many aspects of the world that have never been quantified before” (Cukier and Mayer-Schoenberger, 2013)

Example: social networks as data



Example: BB activity as data



Total Item Accesses	Avg Mins per Item
32	15.4
2	17.8
2	1.0
1	17.5
2	9.3
1	1.1
1	0.0
1	0.6
1	1.8
1	9.5
4	64.7
1	0.0
2	10.6
1	17.9
1	0.4
3	0.3

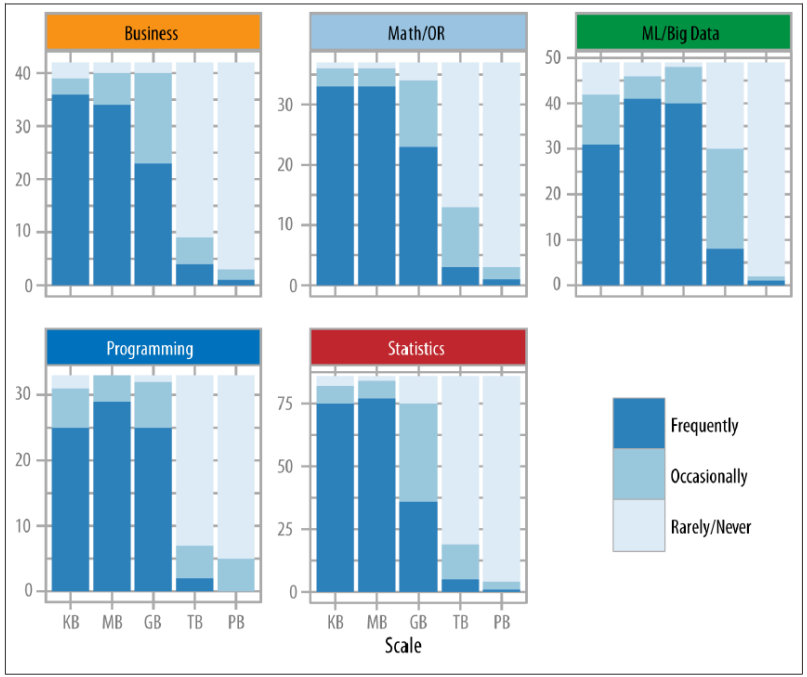
Problems in statistical inference for big data

- A lot of statistical theory was developed for drawing inferences about populations from relatively small sample sizes
- As n, the sample size, gets very large, small effects become 'statistically significant' – but so what?
- Nevertheless, having large n, **does not**, relieve us of the burden of understanding and evaluating the context in which the data was collected, and the limitations of the conclusions

The rise of big data means …

- Collecting and using a lot of data rather than small samples
- Accepting messiness in data
- Giving up on knowing causes and focusing on predictions
- $n =$all?
- (Cukier and Mayer-Schoenberger, 2012)

Big data – how big is ‘big’?



Some sources of data

Upcoming

This week

- Computer Lab: Introduction to R
- Workshop: Former student talks

Next week

- Elements of visualization
- Misleading visualizations

Task for next week

- Find an example of a misleading visualization
  - But please don’t type “misleading visualization” into Google!