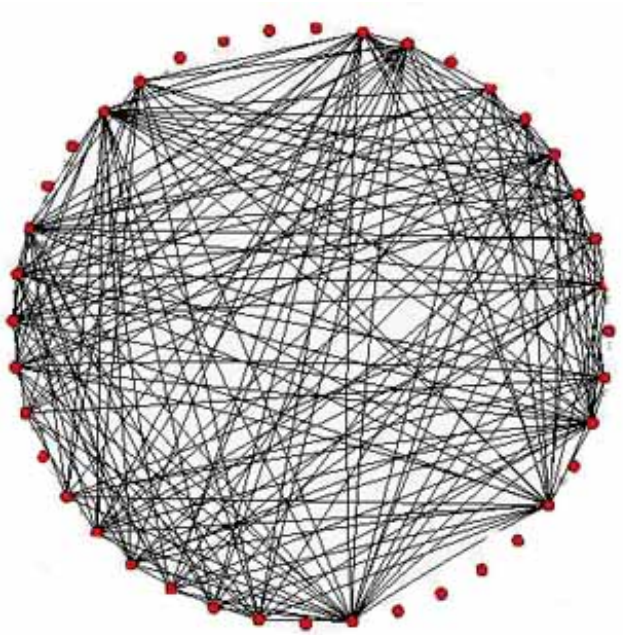
The Challenge of Big Data

Google™

Chris Budd



GRESHAM COLLEGE

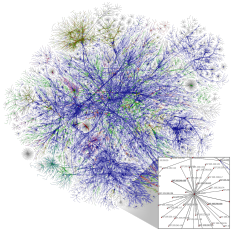
IMI

UNIVERSITY OF BATH

How Big is My Data?

Big Data is all around us!

Google



Anyone who has a Smart Phone, a LapTop or who uses Google is already interacting with Big Data.

We generate Big Data (all of which can be analysed) whenever we travel, use a credit card, turn on a light switch, see the doctor or even go to the shops.

TARGET

Big Data in medicine has been advertised as a means of curing many diseases including cancer and it is used to fight crime.

But .. Does the hype match the reality, and is it all good?

Bridget Jones's Baby features a mathematician who claims that he can find the perfect date by using an algorithm

Is this possible/good?

Is Big Data Watching You?

Who knows if you are pregnant?

After narrowing in on their customers' specific needs, Target's Mom and Baby sales skyrocketed

http://www.dailymail.co.uk/news/article-2102859/How-Target-knows-shoppers-pregnant--figured-teen-father-did.html#ixzz4ObG7CSr1

But .. Are we happy to lose our privacy in this way?

The UK Perspective

2013: David Willets, then UK minister for universities identified

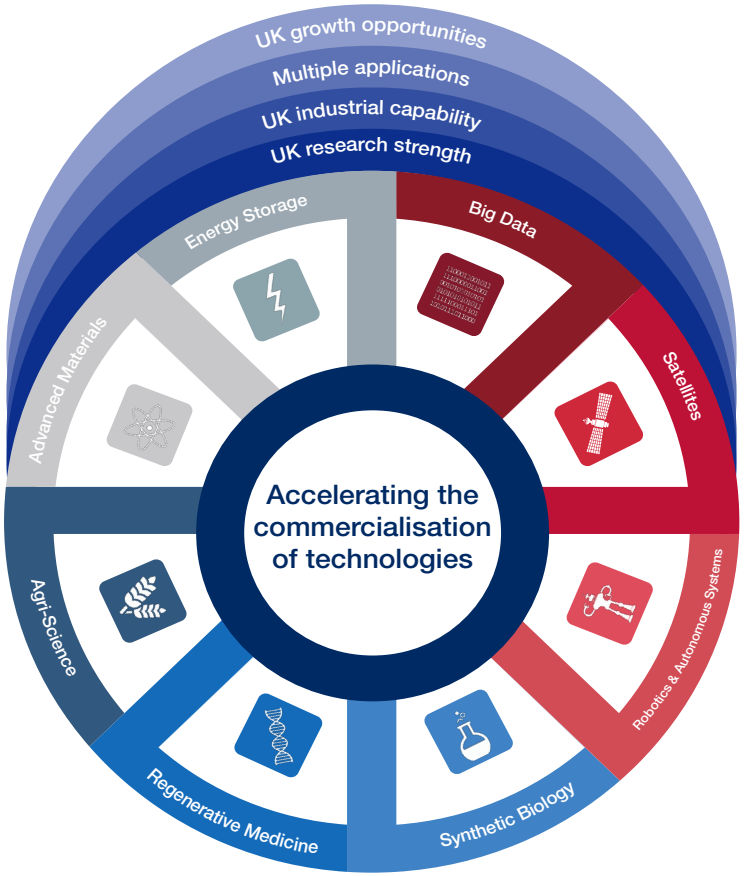
8 great technologies for investment

HM Government

Industrial Strategy

Eight Great Technologies

Technologies in which the UK is set to be a global leader



Accelerating the commercialisation of technologies

www.gov.uk/bis/industrial-strategy

#8great

UK list identified by the Policy Exchange think tank and the Technology Strategy Board in collaboration with research scientists.

Made it on the list if:

1. An important area of scientific advance
2. Already some existing capacity
3. Likely that new commercial technologies will arise from them
4. 'Some' popular support

Promise of an immediate £600M and then up to £1.5bn of new capital investment

On top of £4.6bn baseline science research funding

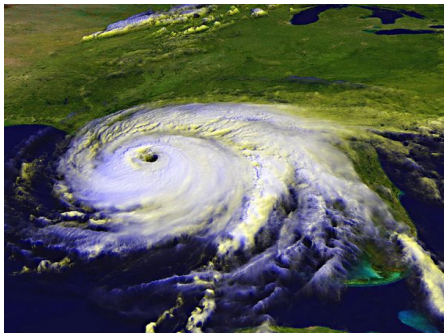
HM Government's Eight Great Technologies

1. Big Data
2. Satellites and Space
3. Robotics
4. Genomics and Synthetic Biology
5. Regenerative Medicine
6. Agri-Science
7. Advanced Materials and Nano-Technology
8. Energy and its Storage

Why is Big Data first on the list?

The key to the modern world is (lots of) information!

Google™



The rate at which we receive data and store has grown Incredibly in the last 100 years

Morse Code: 2 Bytes per second

Teleprinter: 10 Bytes per second

Modem: 1 Kilobyte per second

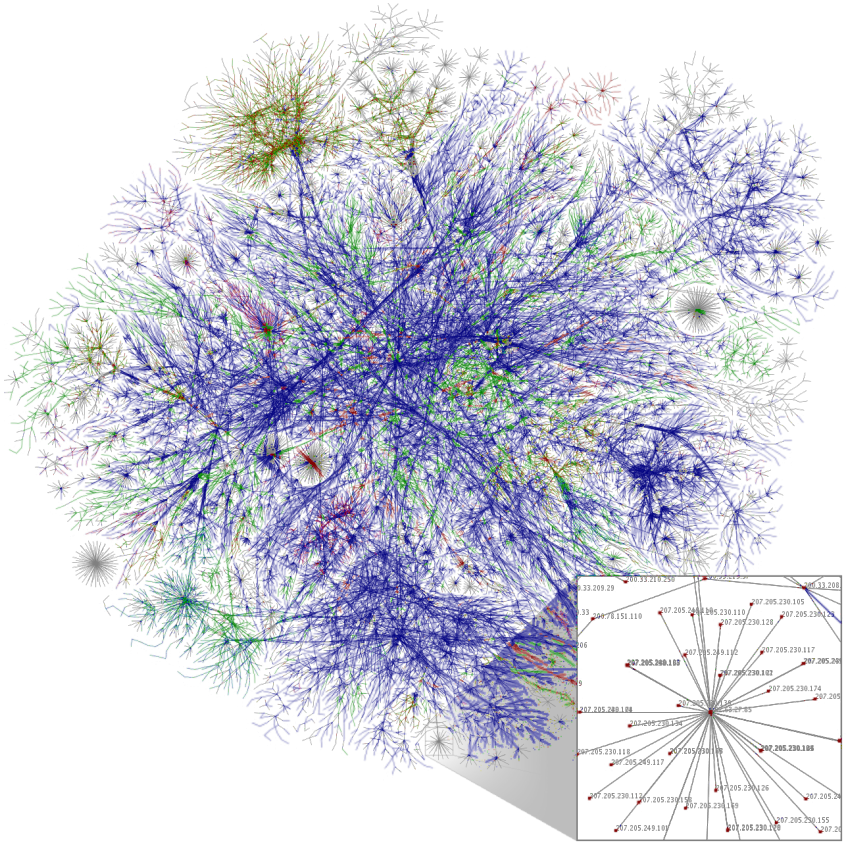
Modern data: 1 Gigabyte per second

Early computers: 1 kilo byte of memory

Now 1 Tera Byte of memory on a lap top

But this leads to challenging problems!!

For example, how do we control and search the internet for vaguely defined information?

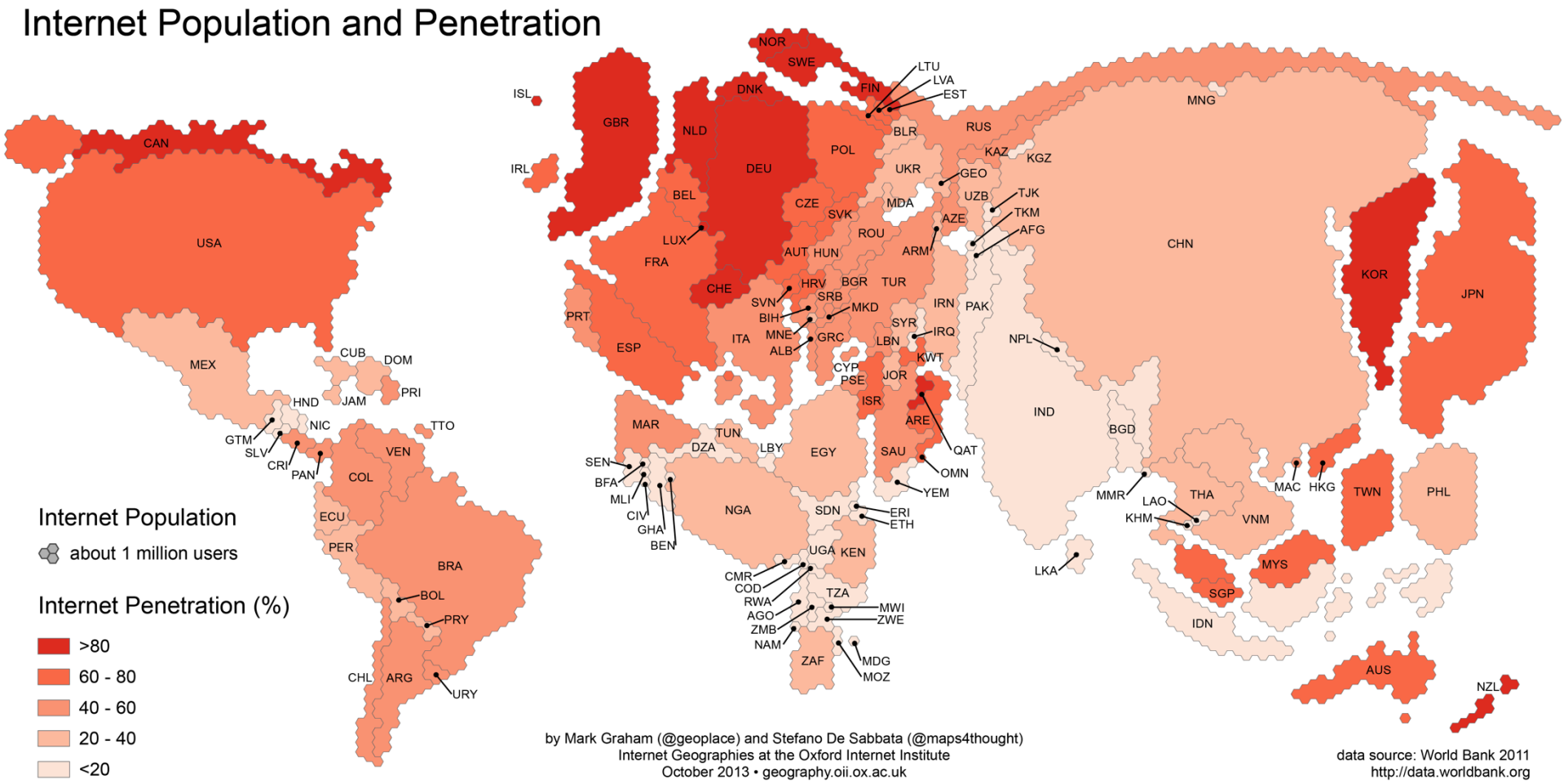


Over a Zetta bytes = 10^{21} bytes of information and growing fast.

The challenge is to derive value from signals buried in an avalanche of noise

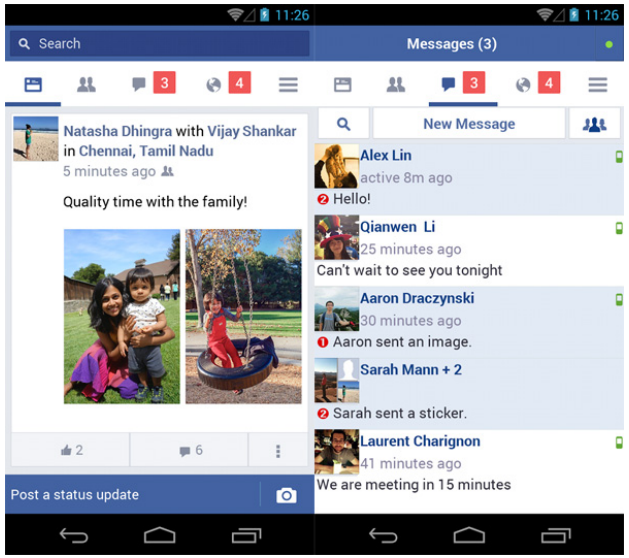
Where does big data come from?

The Internet



facebook

Launched in 2004



2 Billion Registered users

1.5 Billion active

Huge amount of data stored as pictures

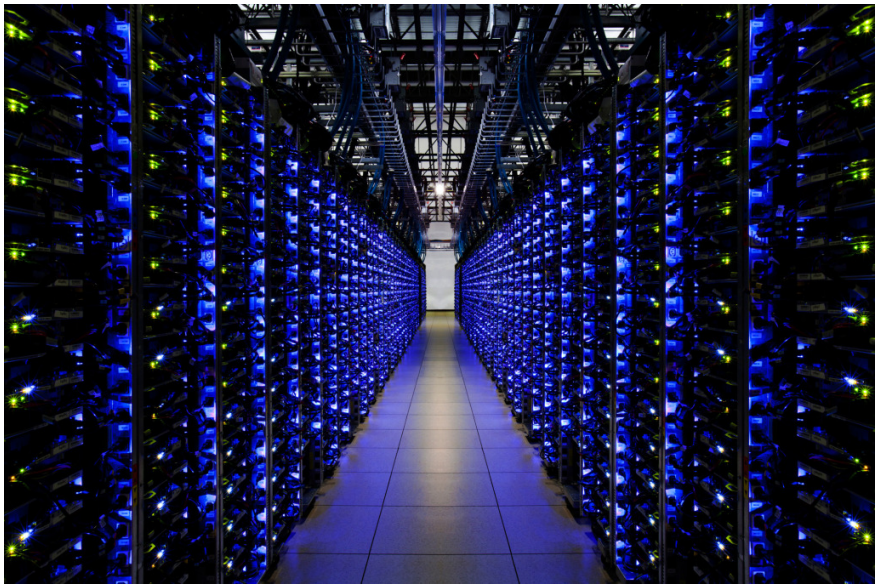
2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day

Google

Google holds somewhere around 10-15 Exabytes of data.

An Exabyte equals 1 million Terabytes

Equates to enough boxes of punch cards to fill up the entire region of New England to a depth of just under 3 miles.



It searches through this in seconds by looking for eigenvectors of large matrices!

Mobile Phones

Over 7 000 000 000 Phones in use in the world (more than people!!!!)

Over 25 000 000 000 000 000 000 possible conversations

3G then 4G now onto 5G

mm wavelength 70GHz

Data rates of several tens of Megabits per second should be supported for tens of thousands of users

1 Gigabit per second to be offered simultaneously to tens of workers on the same office floor

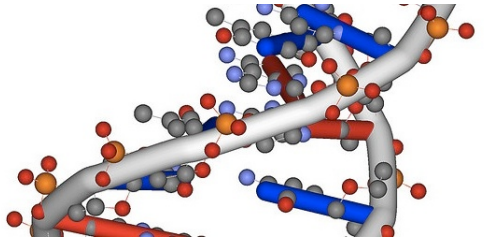
Several hundreds of thousands of simultaneous connections to be supported for massive sensor deployments

Brunel Mile

Smart Grid

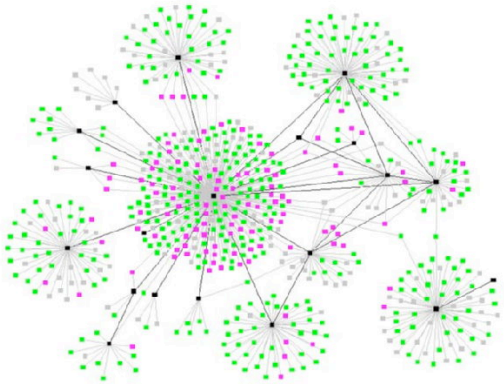
Constant data supplied of our energy usage

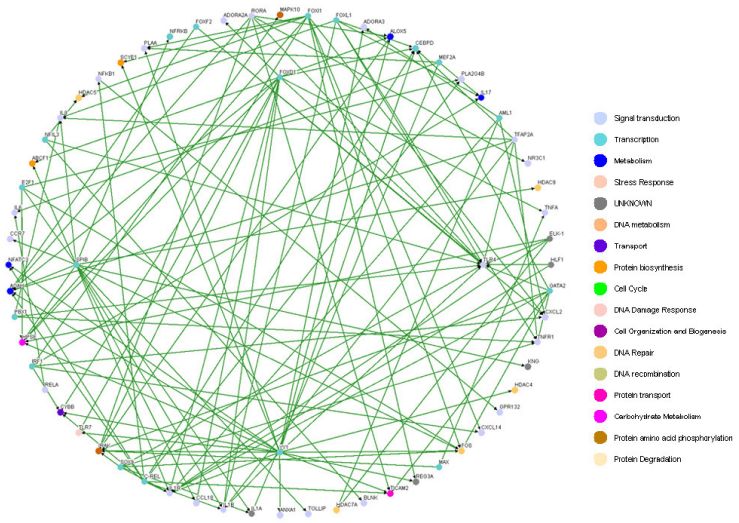
Genomics



Gene/protein networks help us to understand how genes interact

May help to cure cancer





What you might you do with the data

- Rank information from vast networks in web browsers such as Google
- Identify consumer preferences, loyalty or even sentiment and making personalised recommendations
- Eg. *Target* and pregnancy
- Model uncertainties in health trends for individual patients
- Monitor health in real time
- Optimise energy supply using smart data

Current challenges

1. Dealing with the size of the data!

Static – Exabytes at rest

Streaming – Gigabytes per second = Zettabytes per year

Mega = 10^6
Giga = 10^9
Tera = 10^{12}
Peta = 10^{15}
Exa = 10^{18}
Zetta = 10^{21}

2. Dealing with the nature of the data

Challenging data

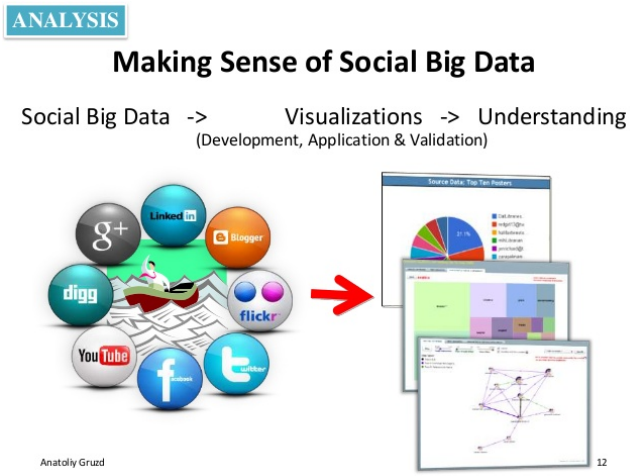
Garbled
Partial
Unreliable
Complex
Soft
Fast
Big

Novel data

Heterogeneous
Qualitative
Relational
Partial
Single-model
Streaming
Big

Challenge to:

Visualise
Speculate
Model
Understand
Experiment
Control



Future Challenges

1. Internet Of Things

Objects talking to each other
with no human intervention

2. Social challenges

Privacy
Ownership
Ethics
Over reliance on algorithms

The Mathematics of Big Data

The very scale of big data makes automation necessary and this, in turn, necessarily relies on mathematical algorithms.

Andreas Weigend former chief scientist at Amazon.com

"It's like an arms race to hire statisticians nowadays. Mathematicians are suddenly sexy."

Example 1. Weather forecasting

1 000 000 000 equations to solve and 1 000 000 data points

to produce a 5 day operational forecast every 6 hours

Forecast works by combining a

GOOD MATHEMATICAL MODEL

with

GOOD DATA

Big problem .. But not quite Big Data

Example 2: Tsunami Forecasting

Part 1. Modeling the tsunami

- Develop a mathematical model
- Take deep water measurements
- Fit the model to the measurements to predict size and speed of the tsunami

Part 2. Modeling the people

- Make a model of crowd motion as they run from the tsunami
- Record mobile phone and social media traffic
- Fit these together to advise emergency services

A true Big Data problem … much harder

How do modeling and Machine Learning work?

IDEA: Data set y of noisy responses to an input x

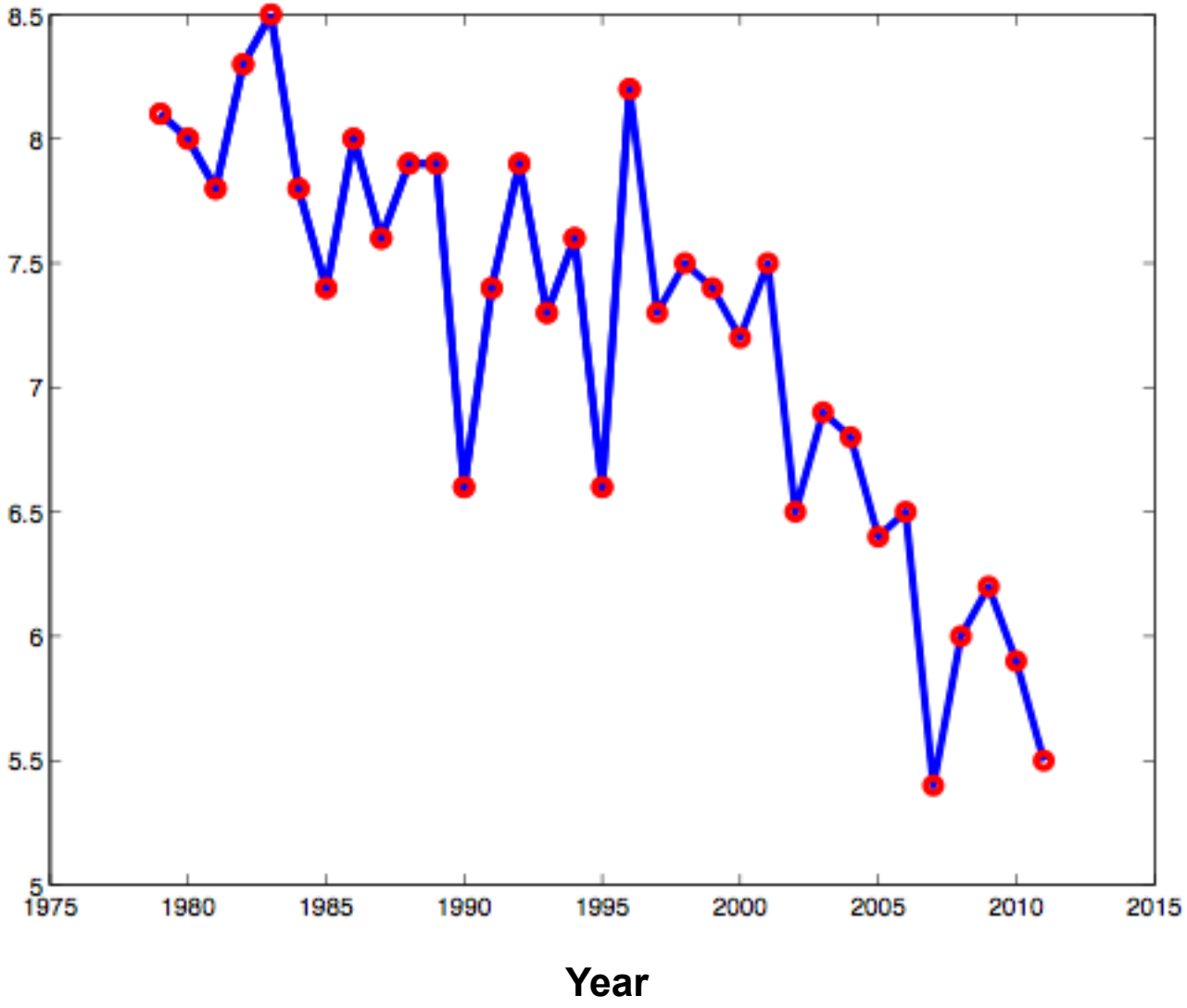
Try to fit a model which links the two

Simplest: Linear models

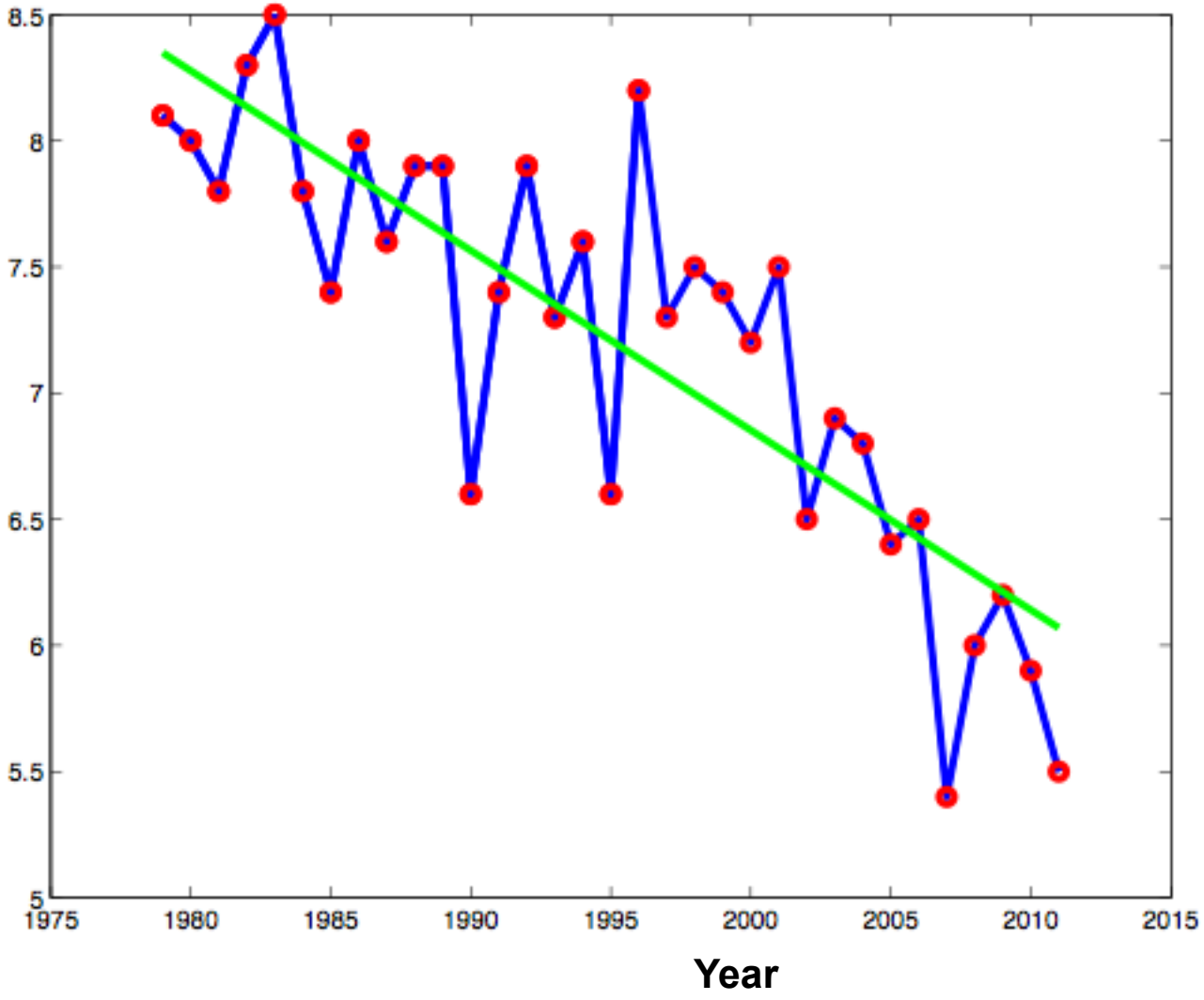
Eg. $$y_i = m\, x_i + c + noise$$

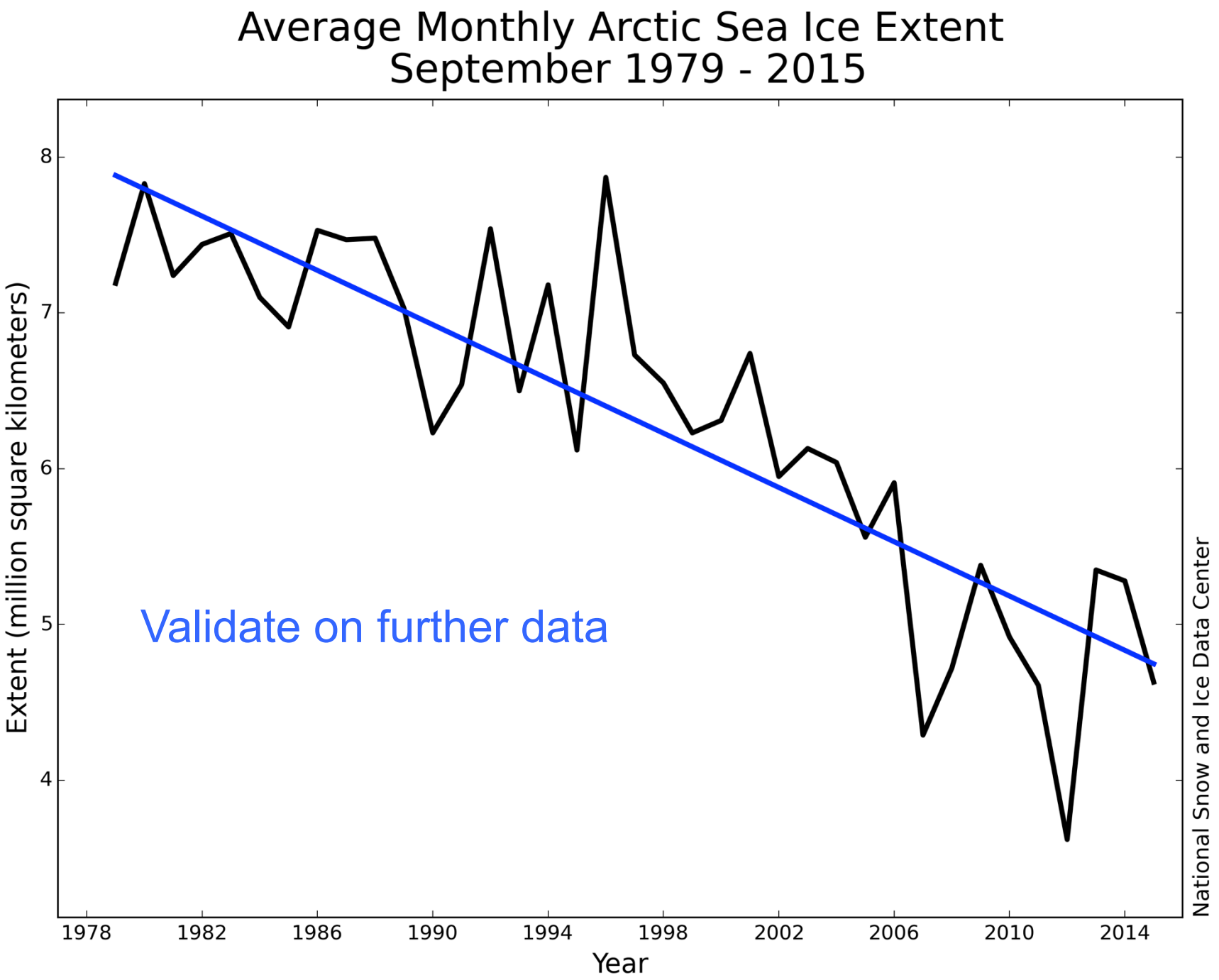
IDEA
1. Train on some data to find m and c
2. Validate model on further data
3. Predict future responses

Eg. Summer Arctic ice in millions of square km till 2011

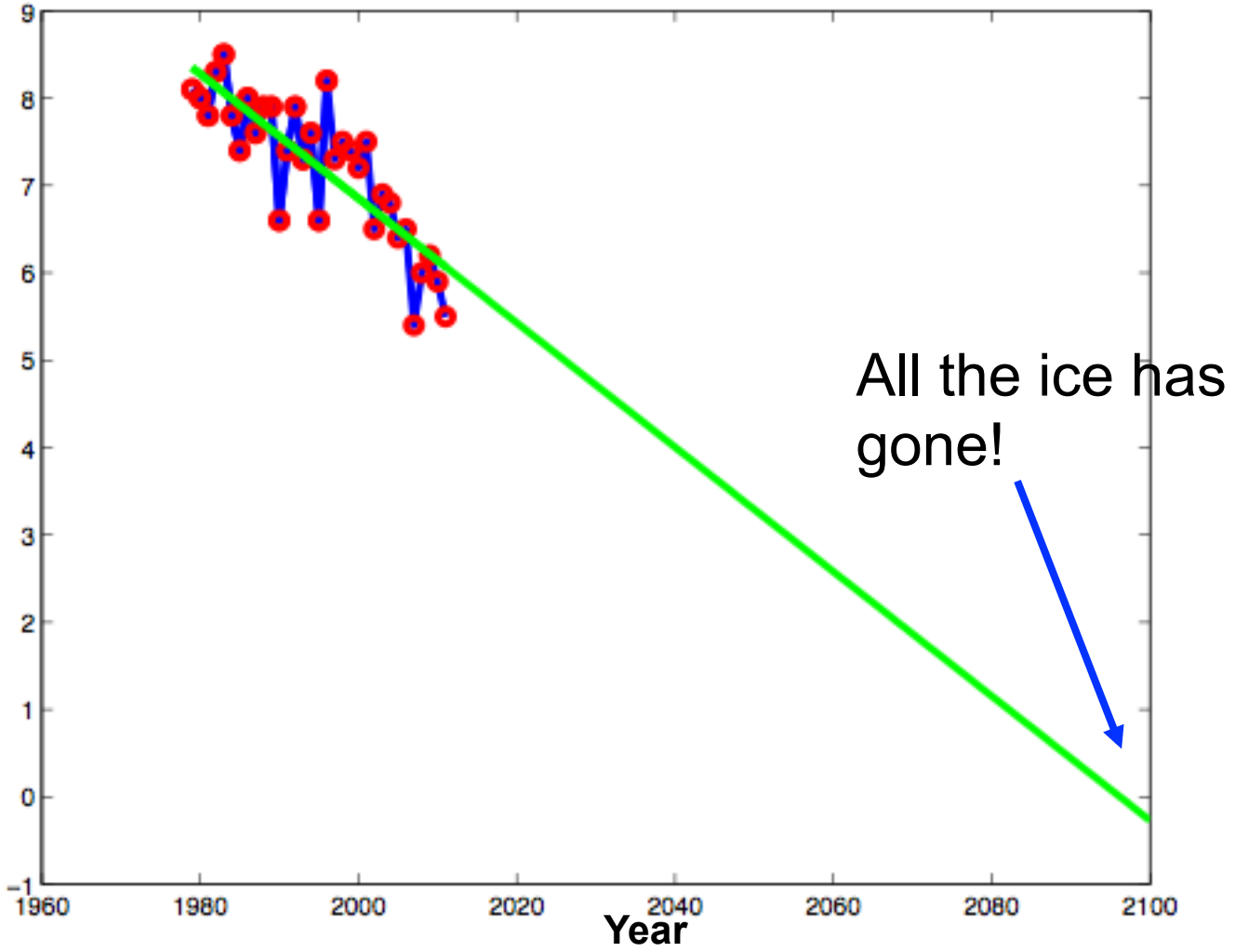


Best fit straight line

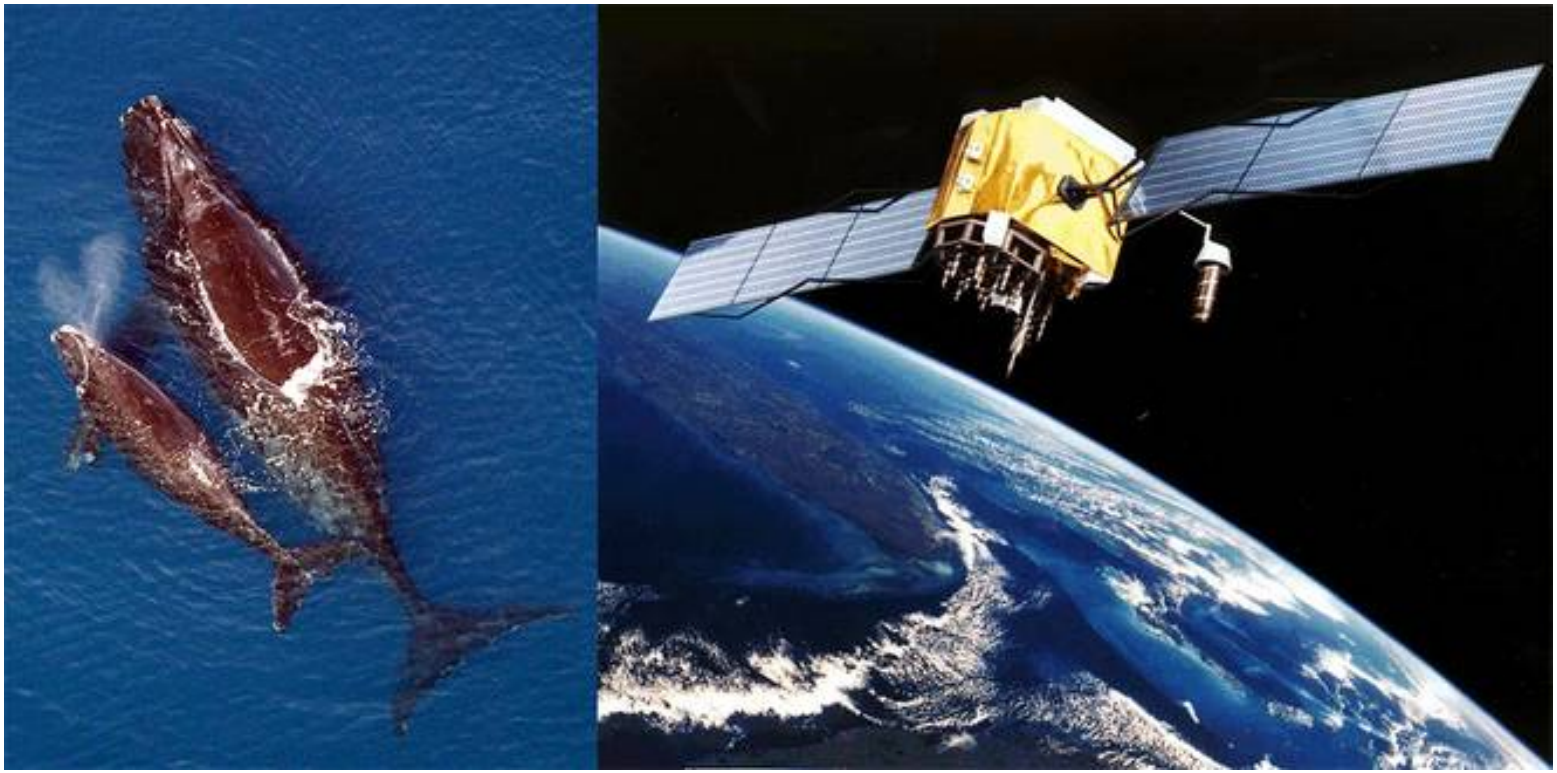



Average Monthly Arctic Sea Ice Extent
September 1979 - 2015
Extent (million square kilometers)
8
7
6
5
4
1978
1982
1986
1990
1994
1998
2002
2006
2010
2014
Year
Validate on further data
National Snow and Ice Data Center

Future prediction???



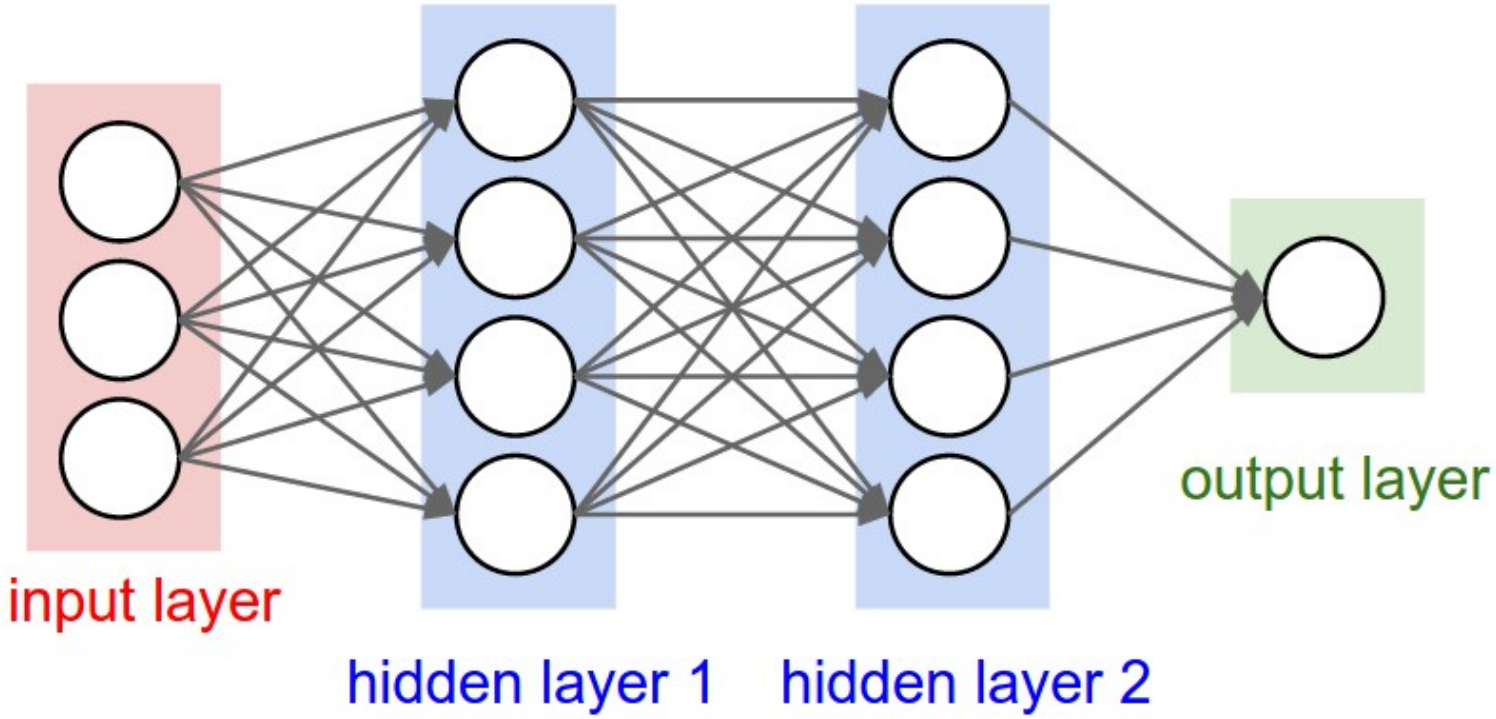
Kalman filters are used to apply linear models to streams of data to estimate the states of a system.



Very widely used in tracking, forecasting (and lots of other technology)

Machine Learning

Sophisticated Neural Nets are now fitted to data



Train, validate, predict

Eg. Health

Data: Lots of X-rays of patients with different medical states

Neural Net:

1. Finds a link between the X-rays and the states.
2. Can then be used to aid in a future diagnosis

Also used in other applications such as: face recognition, energy demand forecasting, government, fraud detection, management, sales forecasting and short range weather forecasting

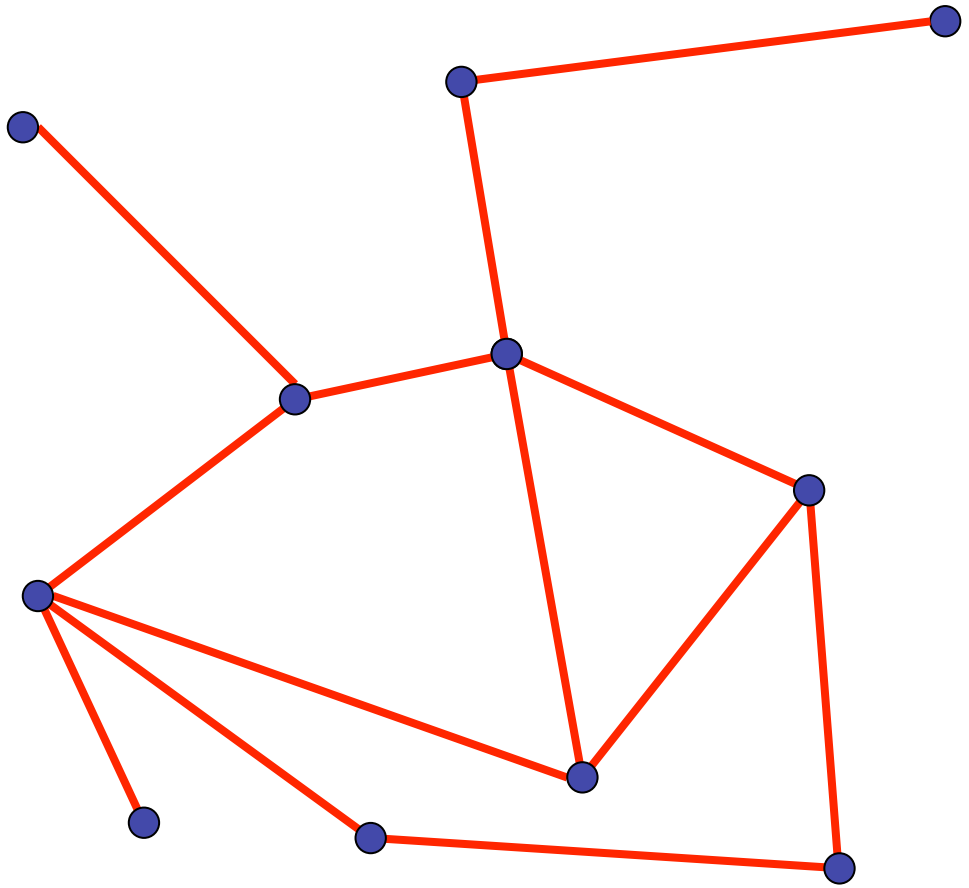
BUT big (Ethical) Issues

Lack of a link between the models in neural nets and reality

Should we leave decisions about people to neural nets?!!

Another vital mathematical tool for Big Data: Networks

These describe how things are connected



Node: eg. Computer or Person or Handset

Edge: eg. Web-link or Conversation or close proximity of two handsets

Some Networks

Social: Friendship, Sexual partners, FACEBOOK

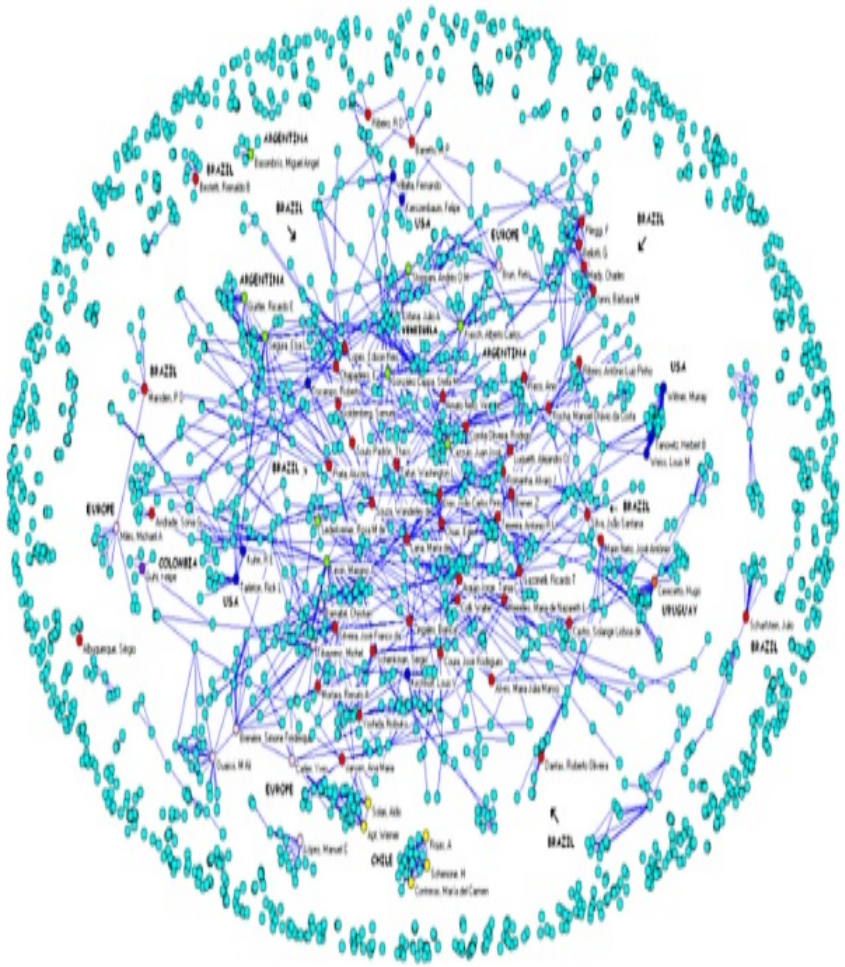
Organisational: Management, crime, Eurovision

Technological: World-wide-web, Internet, power grid

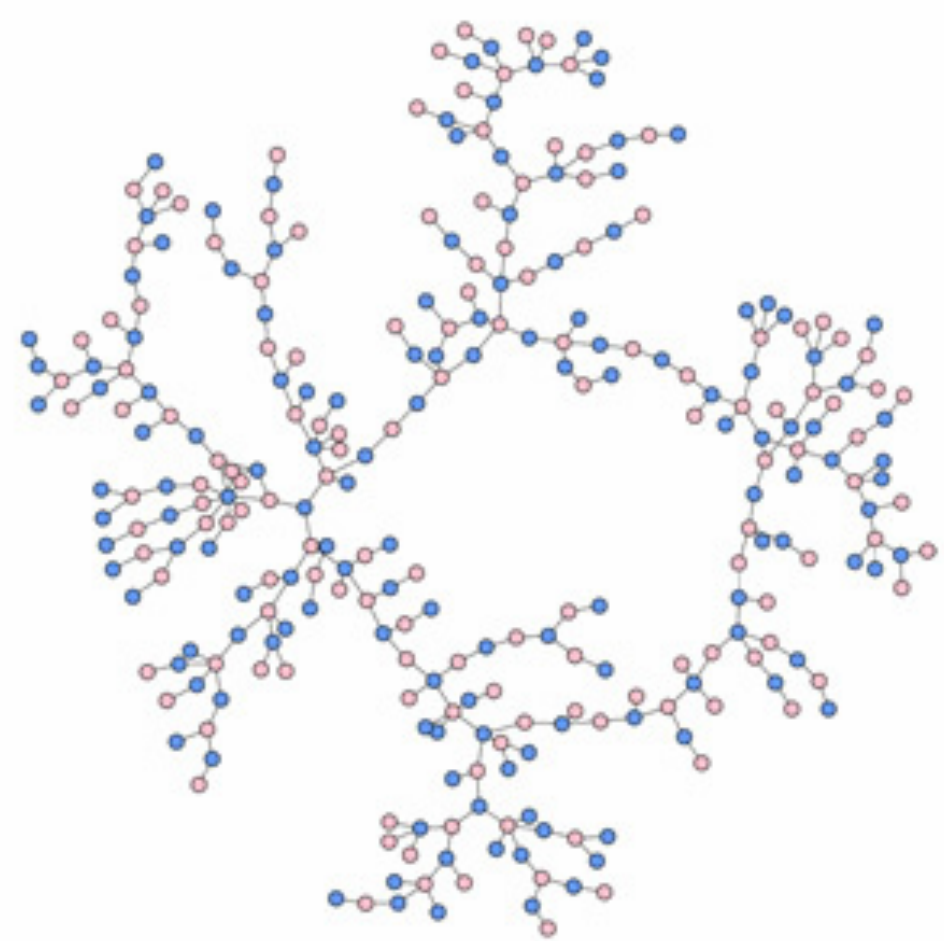
Information: DNA, Protein-Protein, Citations, word-of-mouth,

Transport: Airlines, London Underground

Ecological: Food chains, disease & infection



uthor's network (≥ 5 co-authorships) on Chagas disease in the *Medline* database (1940-2009).



Scientific collaborations

Sexual contacts

Network theory can help answer questions such as:

1. What is the distribution of the nodes and edges?

   What are popular websites, who are party animals?

2. How connected is the network: What is the shortest length of a path through the network?

   Efficient routing in the Internet, SatNav, rumour spreading, marketing, London Underground

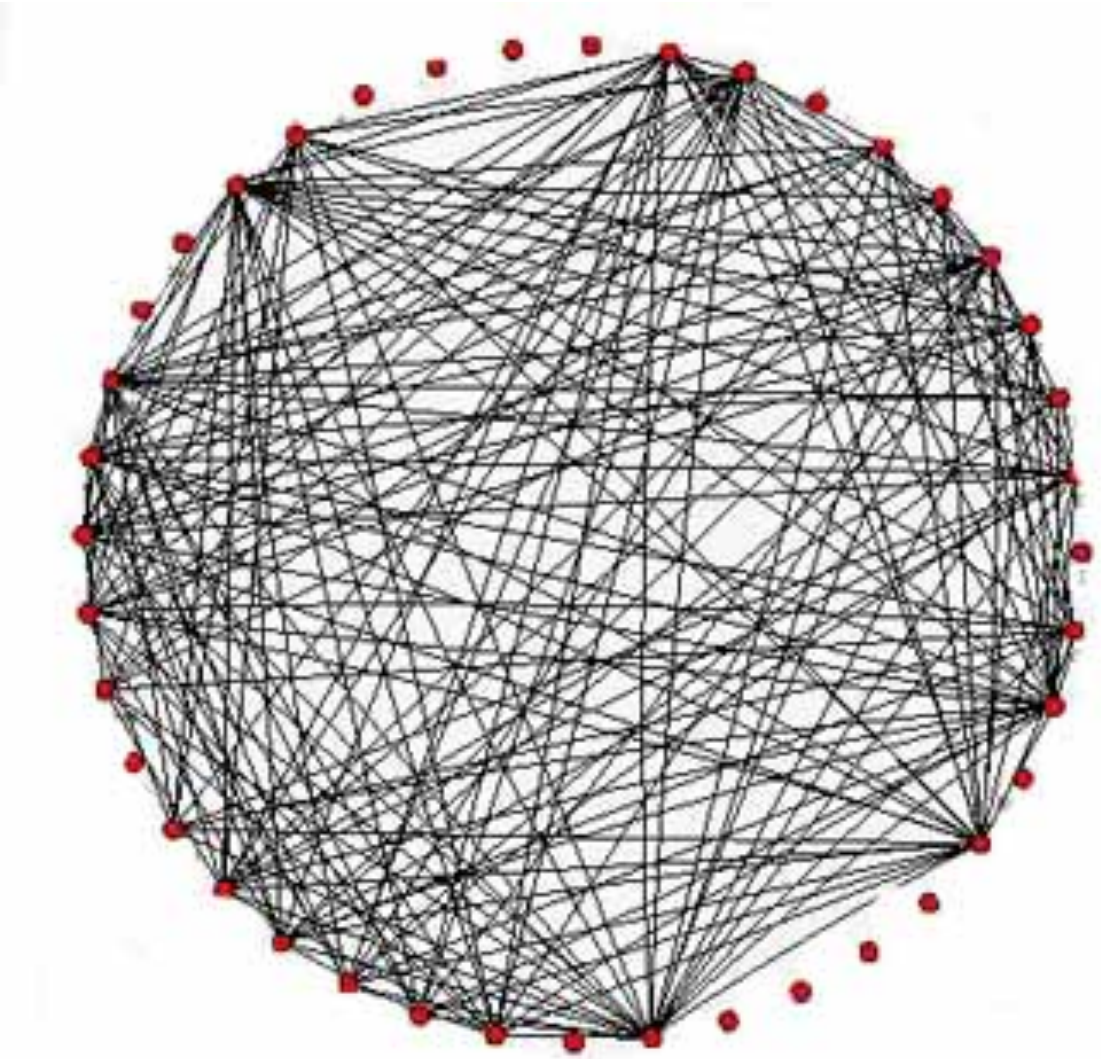
3. How resilient is the network to removing nodes or edges?

   Breaking a terrorist network, stopping an epidemic

4. What are the clusters in the network?

   Friendship groupings, organisation of the brain,

Voting patterns

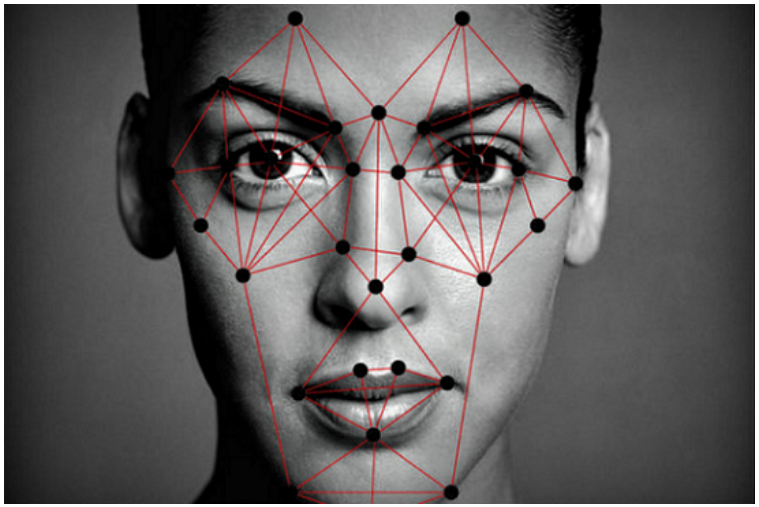


Can segment this network to find voting blocs

Yes .. It's all true!

Even more mathematical methods .. Should keep us busy!

- Optimal and dynamic sampling
- Compressed sensing, sparsity, L_1 approximation
- Probably approximately correct methodologies
- Uncertainty modelling & generalisation error bounds
- Trend tracking & novelty detection
- Context awareness
- Integration of multi-scale models
- Real-time forecasting
- Data integrity & provenance methods
- Visualization methods
- Data compression and visualistion
- Algebraic topology – persistent homology
- Tropical geometry – combinatorial skeleta
- Dimension reduction – machine learning
- Logic and reasoning
- Optimisation and decision
- Encrypted computation Number theory
- Quantum algorithmics



The Ethical Dimension

Big Data leads to big ethical issues

- Loss of privacy
- Attempt to work out how we are thinking
- Decisions made about us by algorithms which the users don't really understand.

Urgent need for mathematicians, computer scientists, lawyers, and policy makers to work together to address these issues

Turing Institute

The work of the Alan Turing Institute will enable knowledge and predictions to be extracted from large-scale and diverse digital data. It will bring together the best people, organisations and technologies in data science for the development of foundational theory, methodologies and algorithms. These will inform scientific and technological discoveries, create new business opportunities, accelerate solutions to global challenges, inform policy-making, and improve the environment, health and infrastructure of the world in an 'Age of Algorithms'.

Bath Institute for Mathematical Innovation

Institute for Mathematical Innovation helps industry solve complex problems using powerful mathematical and statistical methods. By simulating outcomes, modelling behaviour and using sophisticated data analysis, we deliver innovative solutions and exceptional results.

So .. Where is this all heading

In Conclusion

Challenges of Big Data will change our lives

Have already seen a huge growth in the last ten years

Will see even greater changes in the future

Mathematicians started the revolution and should be involved in its future … but only in collaboration with many others