estimand

Introduction to Data Science

Dr Gianluca Campanella

3rd May 2016

Contents

What?

What is Data Science?

A problem-solving approach based on the **scientific method**

What is Data Science?

Mathematics and Statistics/OR

Computing and Software Engineering

Visualisation and Communication Skills

Domain expertise

What does Data Science deal with?

Problems!

- Will our customers switch to our competitors?
- Where will the next fire strike in London?
- Who will develop cancer?

What does Data Science deal with?

Predictions?

- Can we **predict** how likely a customer is to switch to our competitors in the next year?
- Can we **predict** where the next fire will strike in London?
- Can we **predict** how likely a person is to develop cancer in the next ten years?

What does Data Science deal with?

Mechanisms?

- **Why** does a customer decide to switch to our competitors?
- **Why** are fires more likely to strike in Lambeth?
- **Why** do people develop cancer?

Why?

Why Data Science?

You have a **problem**...

Why Data Science?

You have a **problem**...

...and some **data**...

Why Data Science?

You have a **problem**...

...and some **data**...

Data Science will give you the tools to **ask a question** and **find an answer**

Data sources

Personal

- Many devices automatically generate data
- All devices will soon do the same (**Internet of Things**)
- You can do it too (**Quantified Self**)

Business

- Many business processes routinely collect data
- Can also design **experiments**

Government

- **Open Data**

Why care about data?

There’s never been so much data around!

Why care about data?

There’s never been so much data around!

- Is it exploited?
- Are users even aware that it could be exploited?

Who?

Who is a Data Scientist?

Someone who is...

- Curious
- Rigorous
- A good communicator
- A team player

Curiosity

That’s funny…

- Expect the unexpected
- …and investigate it!

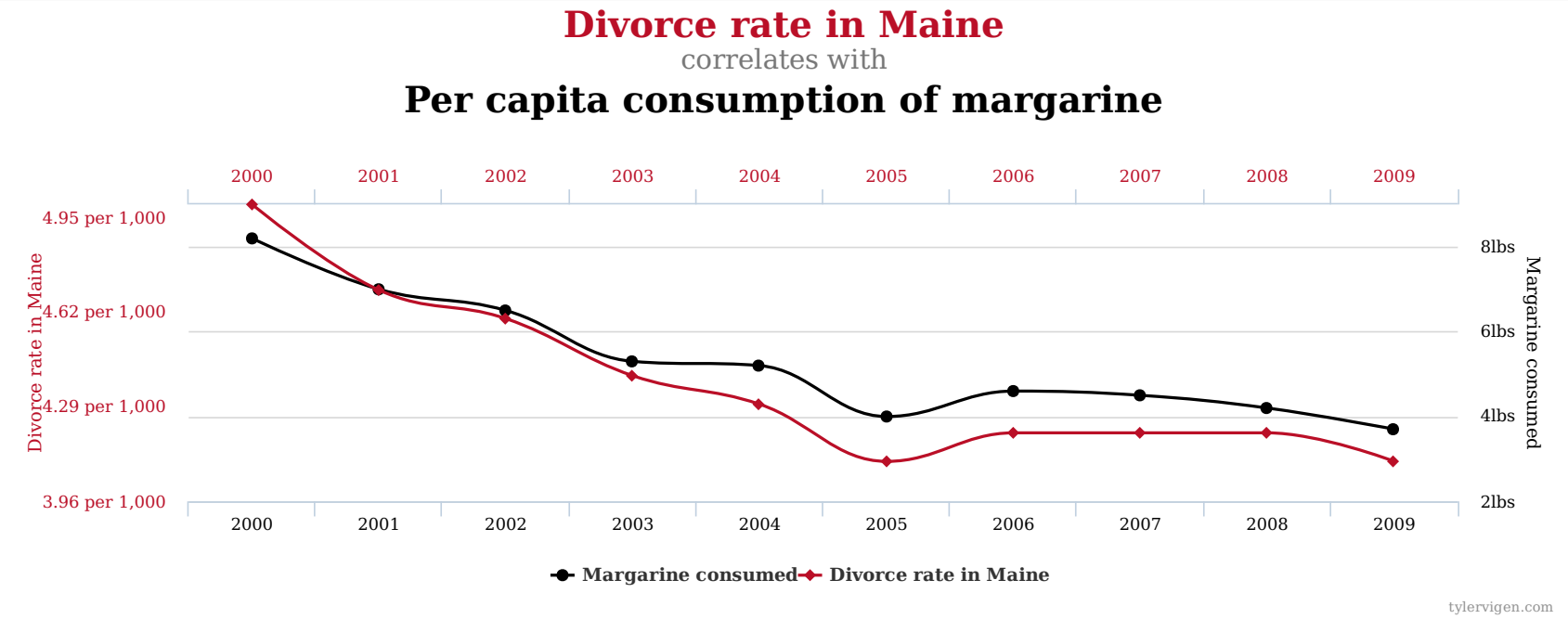
Rigour

Patterns are everywhere!

(Or do we just *see* them everywhere?)

- Beware of **confirmation bias**
- What is **signal**, what is **noise**?

Relevance



(More at `http://tylervigen.com/spurious-correlations`)

Relevance

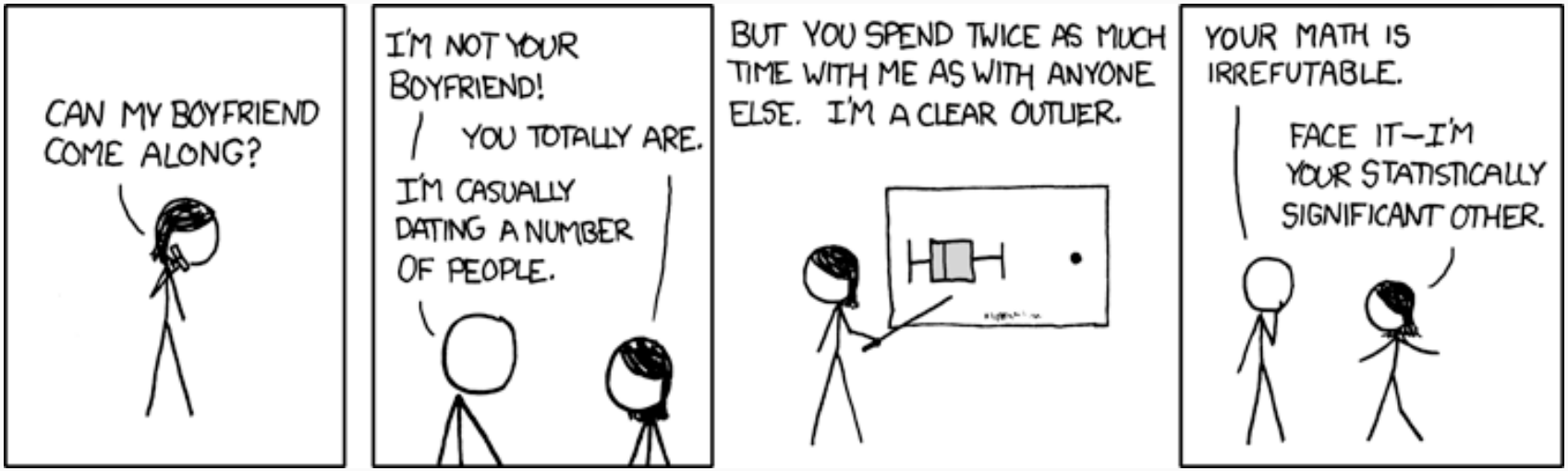
Curiosity will find the unexpected...

...rigour will confirm it...

...but does it matter?

('statistically significant' ≠ 'significant')

Relevance



http://xkcd.com/539/

Communication

The scientists say...

> *'...colorectal cancer risk was related to intake of fresh red meat (RR for 100 g/day increase = 1.17, 95% CI = 1.05–1.31) and processed meat (RR for 50 g/day increase = 1.18, 95% CI = 1.10–1.28)'*
>
> — *DSM Chan et al. (2011), PLOS ONE* ***6****(6)*

Communication

The media say...

> *'Processed meats — such as bacon, sausages and ham — do **cause** cancer, according to the World Health Organization (WHO)'*
>
> *— BBC News (26th October 2015)*

> *'New cancer alert over eating just ONE steak a week: eating just 10 oz of red meat can increase chance of bowel cancer by two-fifths'*
>
> *— Daily Mail (2nd November 2015)*

Communication

Tell a story!

- Most people don’t understand statistics
- Explain clearly and listen

Teamwork

Data Science is interdisciplinary:

- Mathematics and Statistics/OR
- Computing and Software Engineering
- Visualisation and Communication Skills
- Domain expertise

Only a team can be strong in all areas

How?

Different paradigms

Different paradigms

Statistics

- Uses the tools of **probability**
- Focus is on **inference**

Machine Learning

- Based on statistics and computer science
- Focus is on **prediction**

Artificial Intelligence

- Based on statistics and computer science
- Focus is on **adaptability** to a changing environment

Types of problems

Supervised

- Data is ‘labelled’ (**outcome**)
- Aim is to:
  - Identify ‘inputs’ assumed to cause the outcome (**independent variables** or **exogenous factors**)
  - Predict the outcome from some **predictors**

Unsupervised

- Data is ‘unlabelled’
- Aim is to unearth **latent structure**

Types of data

Qualitative

- Categorical
- **Nominal** or **ordinal** scale

Quantitative

- Numerical (**continuous** or **discrete**)
- **Interval** or **ratio** scale

Example: scales of measurement

Nominal

Sex, colour, ...

Ordinal

Ranks

Interval

Temperature

(there is no ‘zero’)

Ratio

Mass, length, duration, ...

(there is a ‘zero’)

Data Science solutions

	Categorical	Numerical
Supervised	Classification	Regression
Unsupervised	Clustering	Dimensionality reduction