Data Science 1

Tilburg University

Open Universiteit

Understanding Society

Setting the scene…

- **Across all disciplines, there is a need for more data scientists**
- **Data science is being touted as a ‘hot’ career choice for academics and practitioners alike!**
- **A recognized need for the use of power technological tools to explore the vast amounts of digital material**
- **Data-analysis methods are not just being employed in natural sciences, but also in law, economics, sociology, political sciences, history, and arts**

What is Data Science?

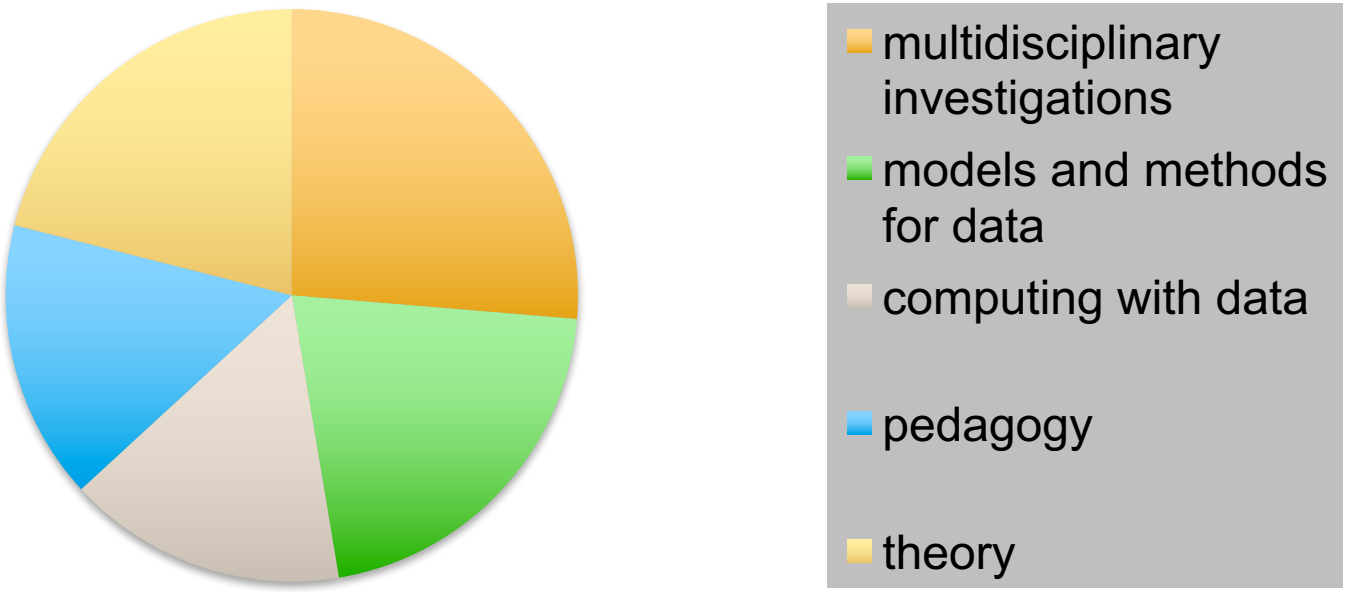
> *"The science of learning from data; it studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner."*
>
> (**David Donoho**, 2015)

The Beginning:

- 1962 - **John Tukey** in 'The Future of Data Analysis' introduced a new field, the result of 4 forces:
  1. The formal theories of statistics
  2. Accelerating developments in computers and display devices
  3. The challenge of more and ever larger bodies of data
  4. The emphasis on quantification in an ever wider variety of disciplines

What is Data Science?

- 2001 – **William Cleveland** in 'Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics'



What is Data Science?

- 2001 – **Leo Breiman** in 'Statistical Modeling: The Two Cultures':

  There are two goals in analyzing data:

Information Extraction (Inference)

Associating the response variables to the input variables

98% of academic statisticians

Prediction

Predicting what the responses are going to be to future input variables

2% of academic statisticians

Tilburg University

Open Universiteit

What is Data Science?

- COMMON TASK FRAMEWORK
  - A publicly available training dataset
  - Competitors whose is to derive a predictive model from the training data
  - A testing dataset on which the predictive models are objectively evaluated
- HACKATHONS

Tilburg University

Open Universiteit

The Full Scope of Data Science

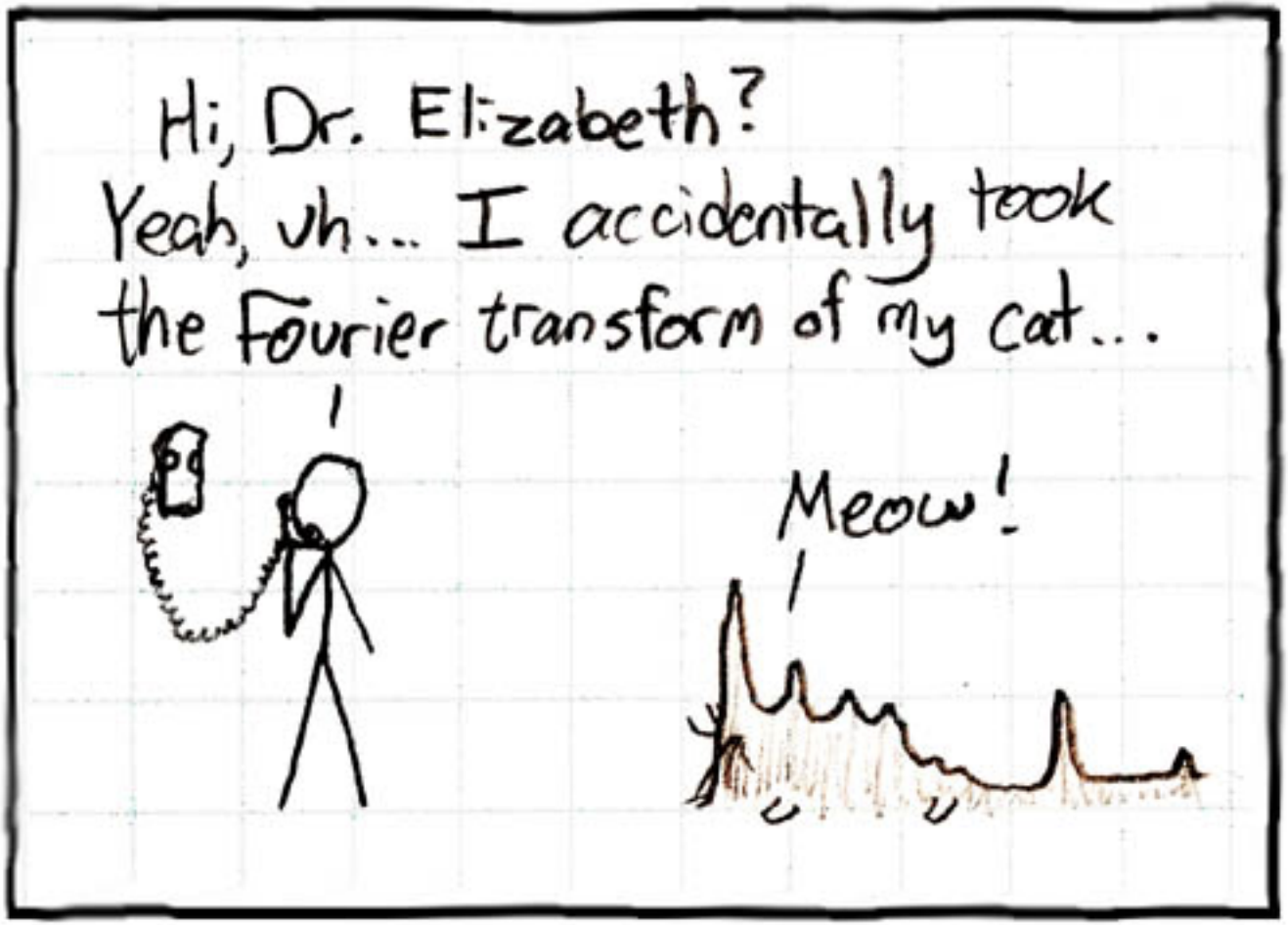
- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science

Data Exploration and Preparation

- 80% of each data scientific project is…

CLEANING...

Data Representation and Transformation



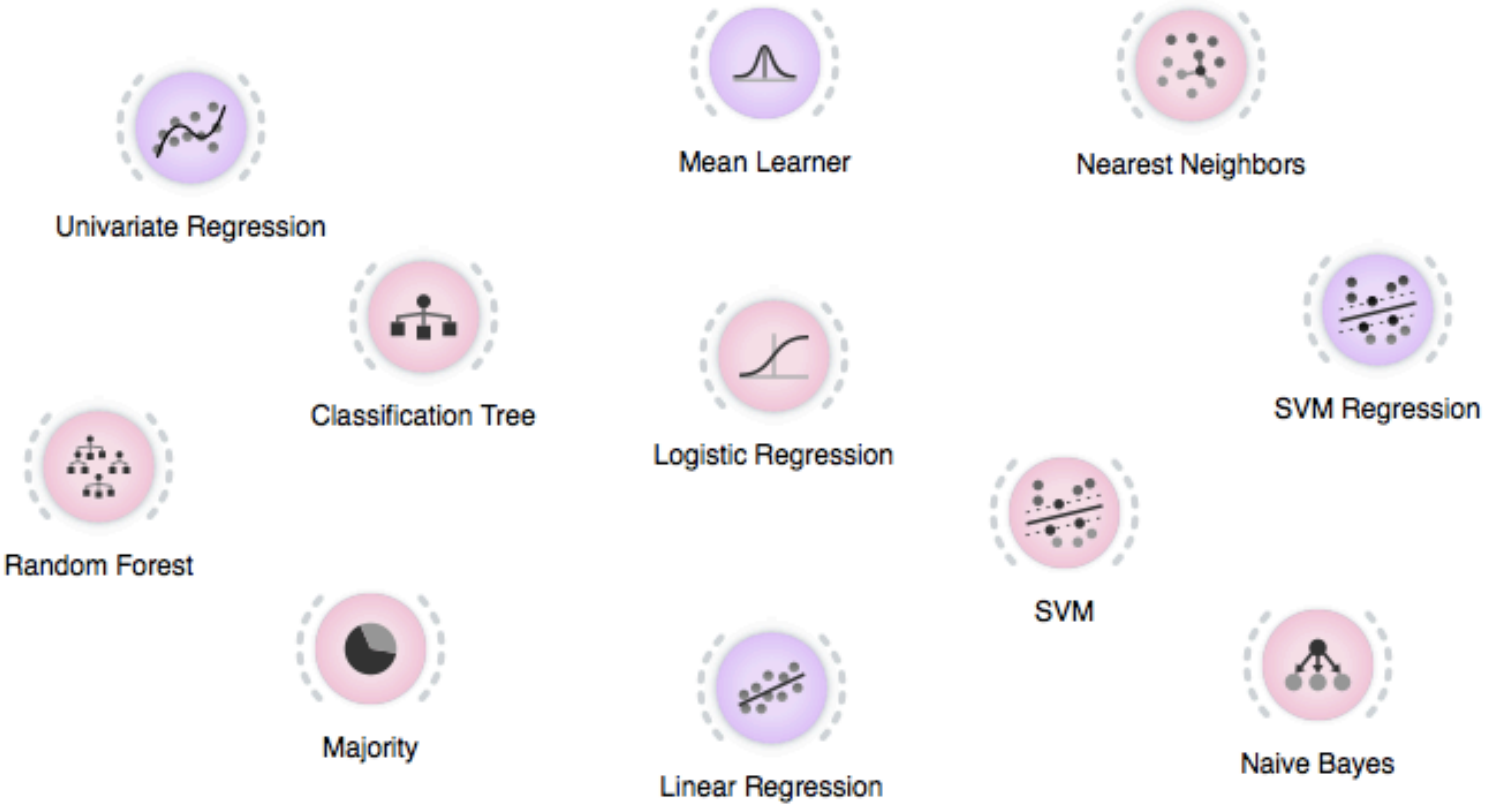
gizmodo.com

Computing with Data

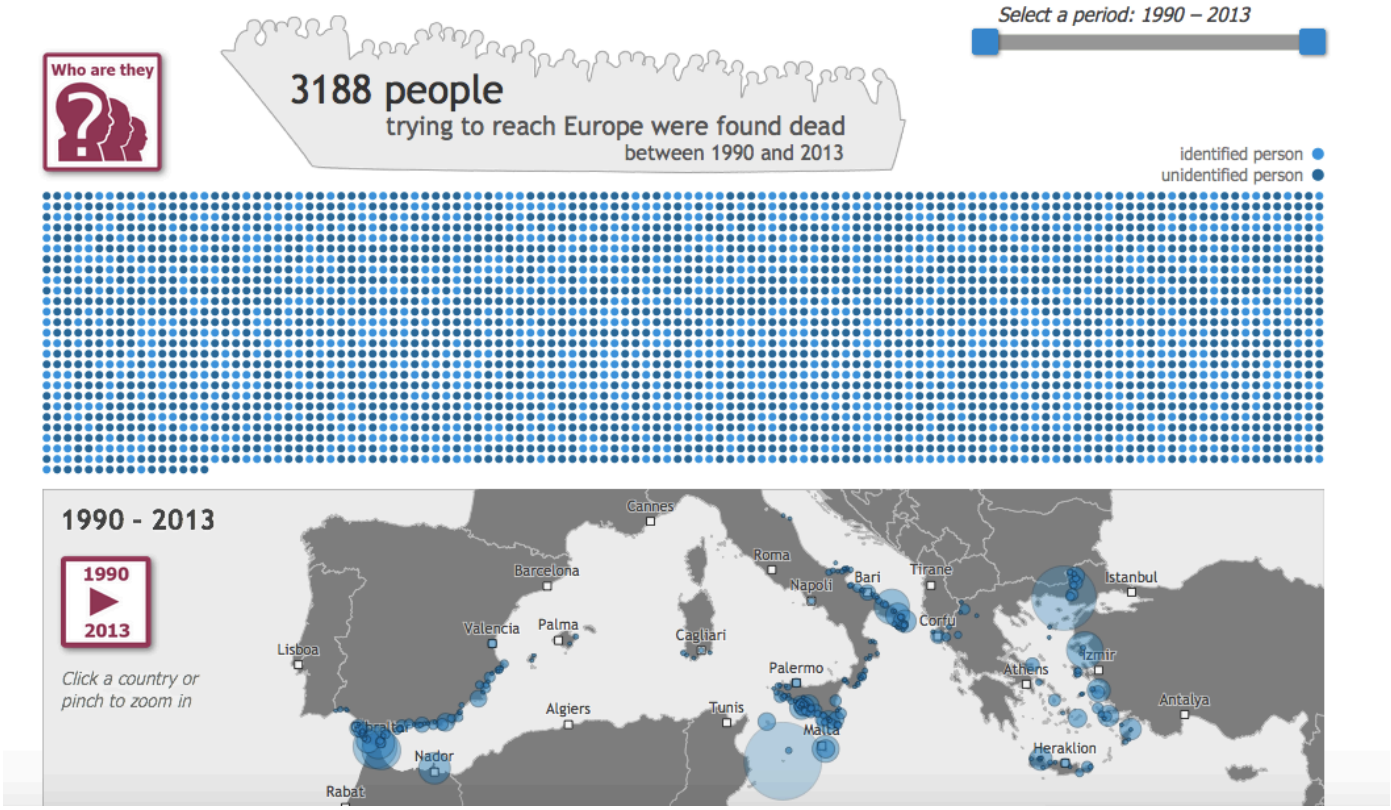
TILBURG UNIVERSITY

Open Universiteit

Data Modeling



Data Visualization and Presentation



Mirjam Leunissen (Dutch Data Design)
for www.boarderdeaths.org

Tilburg University

Open Universiteit

Science about Data Science

Home

I am a PhD student at Tilburg University. My research focuses on the detection and prevention of statistical errors and researcher degrees of freedom in psychological research.

Contact: m.b.nuijten@uvt.nl

"The prevalence of statistical reporting errors in psychology (1985-2013)" published at Behavior Research Methods

October 2015

In this paper we use the automated procedure "statcheck" to extract over 250.000 p-values from 30.000 psychology articles and check whether they are consistent.

We find that half of the articles contain at least one inconsistency, and 1 in 8 articles contains a gross inconsistency that affects the statistical conclusion. The prevalence of inconsistencies seems to be stable over time.

Tilburg University

Open Universiteit

Summary

- Data scientists use both ‘academic’ statistics for inference and machine learning for predictive modeling;
- An inherent part of their job is that they do not work with carefully curated data but with large messy data sets; therefore, a substantial amount of time is devoted to data cleaning and preparation;
- “From causation to correlation”: The quest for the best theory (hypothesis testing) has been replaced by data analysis (Let the data speak!) and predictive modeling.

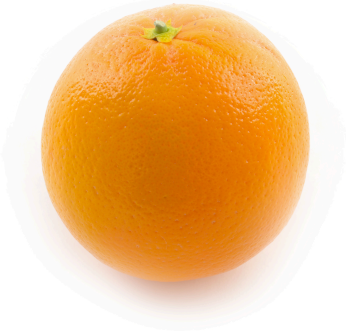
Typical Data Science Tasks

- Prediction

- Recognizing previously unseen things

Two Main Types of Prediction

- **Classification**
  - Discrete output
  - E.g. color, gender, truth, inequalities, class membership, ...



FRUIT?

"YES"

- **Regression**
  - Continuous output
  - E.g. temperature, salary, length, pressure, …

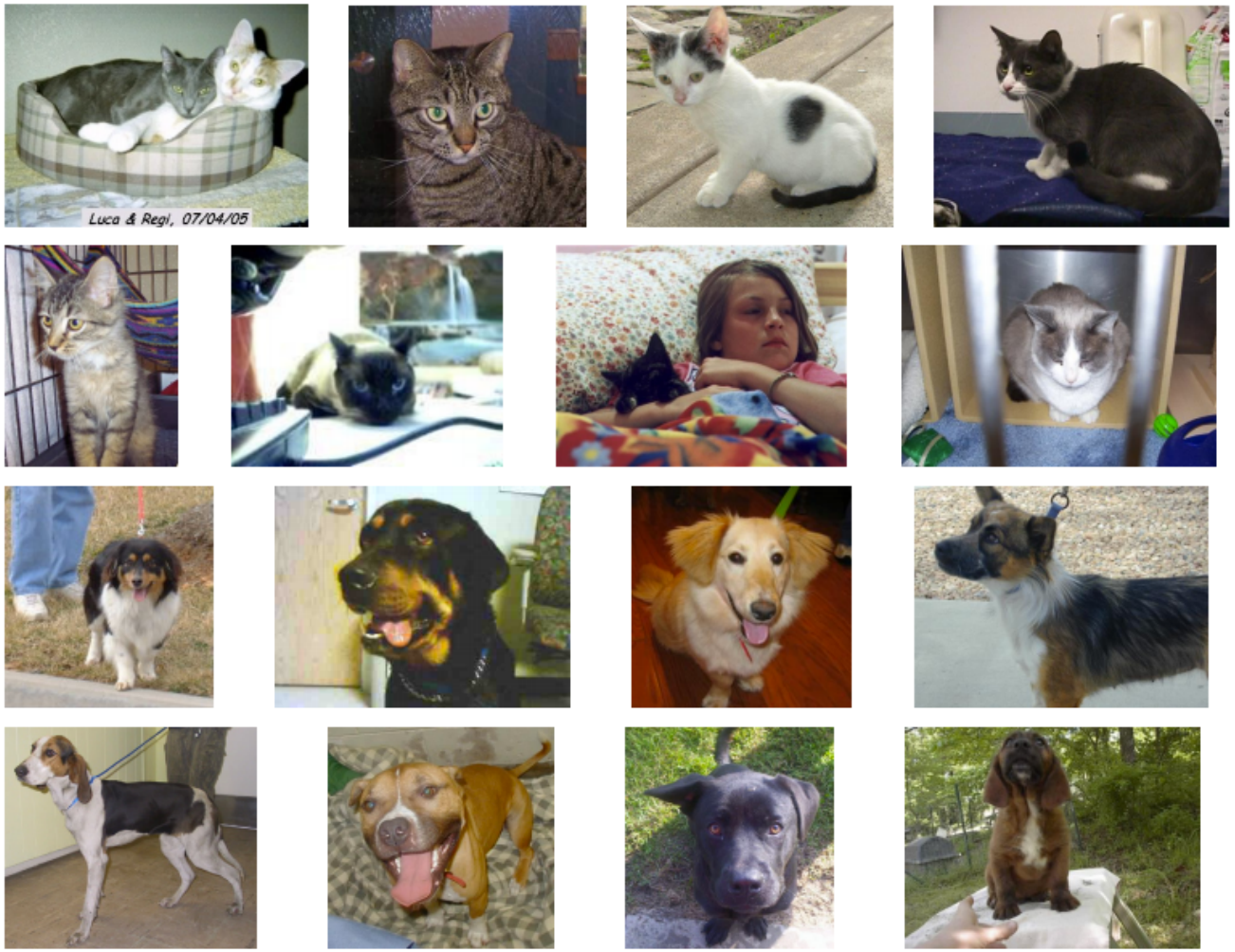
TEMPERATURE?

"36.5"

Binary Classification: DOG/NO DOG?

TILBURG UNIVERSITY

Binary Classification: DOG/CAT?



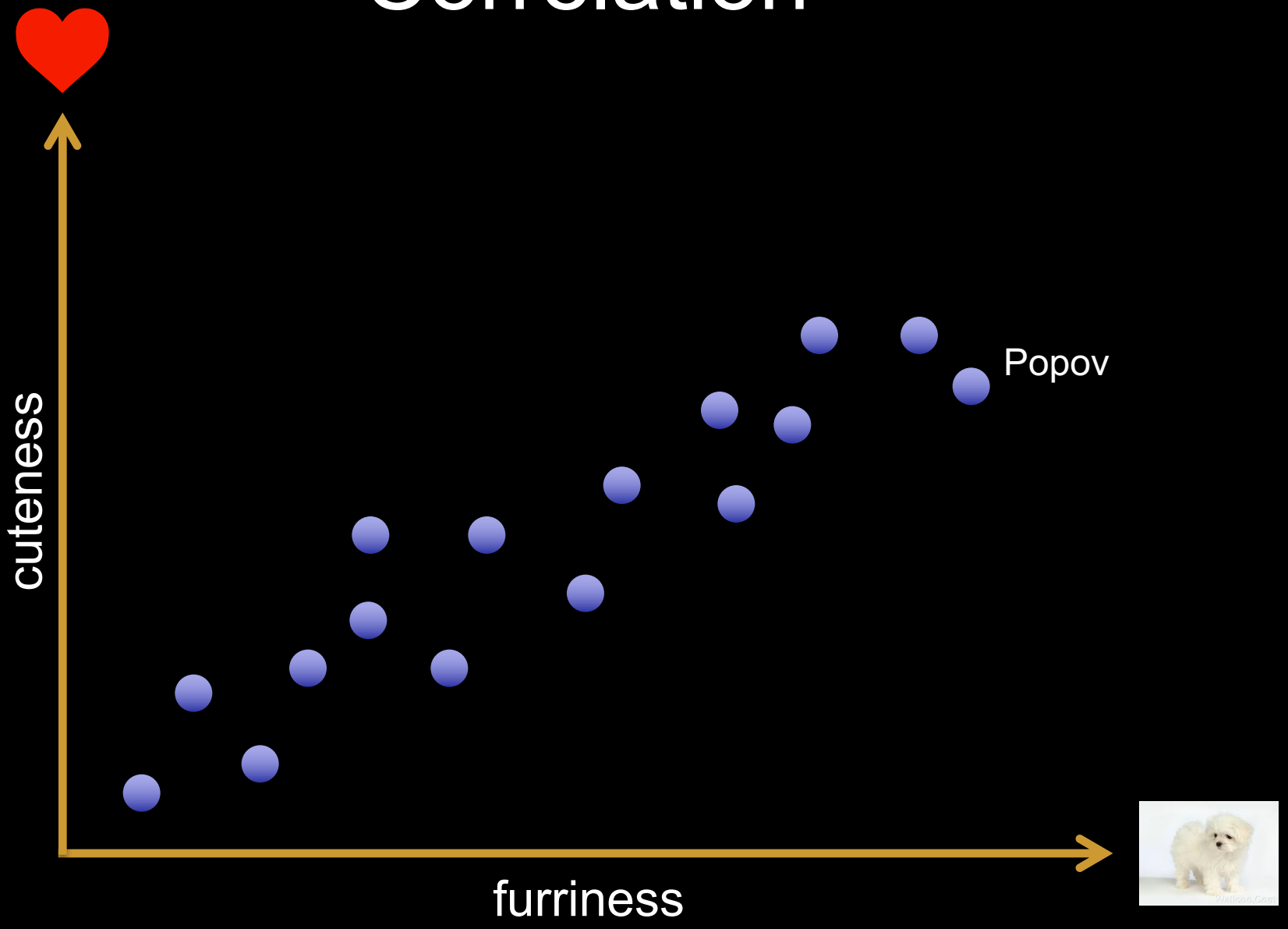
Regression: AGE?

TILBURG UNIVERSITY

Data Science Basic Concepts

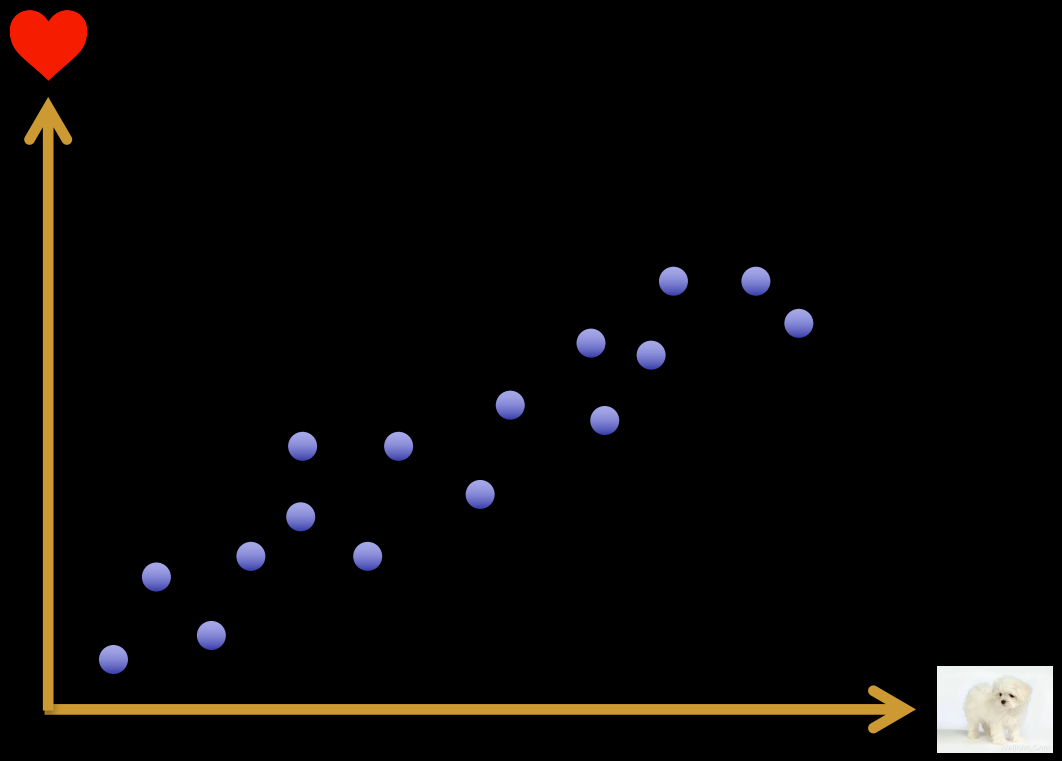
CONCEPT	EXAMPLE
• instances	“Popov”
• labels	“dog” or “4 years”
• features	color, height, furriness
• feature values	“white”, “0.3m”, “high”
• feature vector	(“white”,”0.3m”,”high”)

Correlation

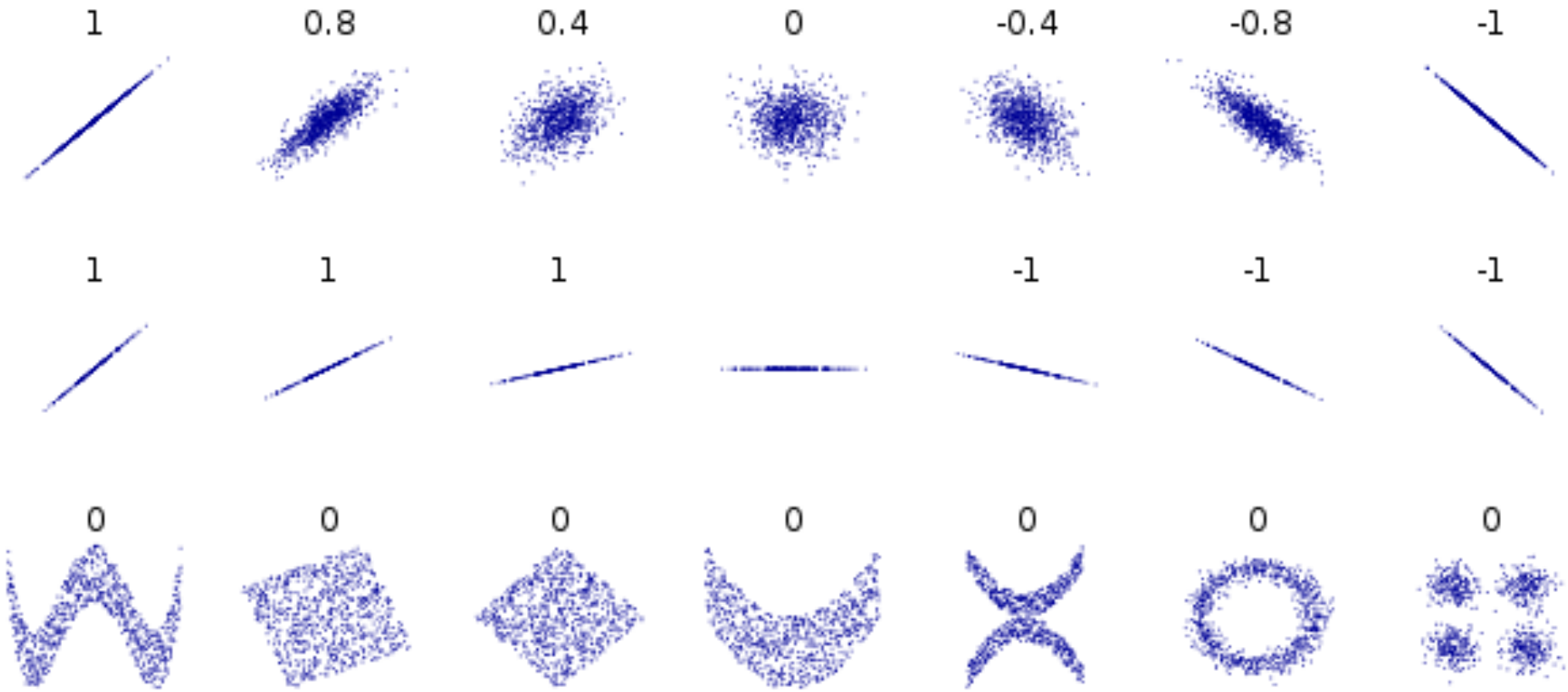


Correlation

- furriness "predicts" cuteness
- correlation may allow for prediction



Pearson's correlation coefficient a.k.a. "Pearson's r"



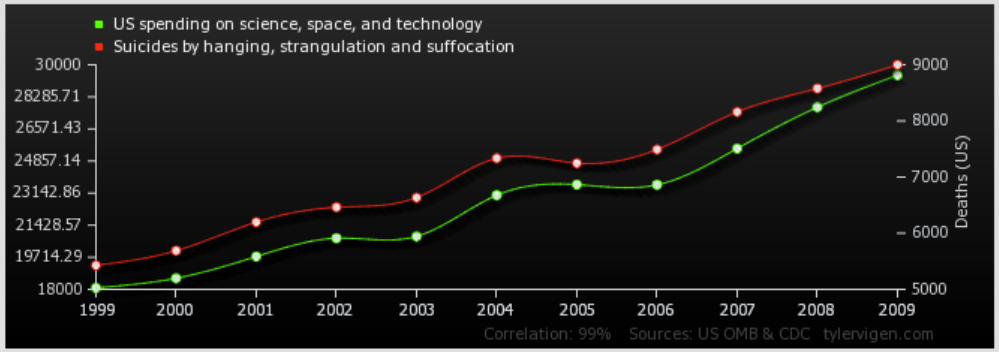
Big Data Hype: “The End of Theory”

Finding correlations in large data collections to make (scientific) discoveries.
“Theories not necessary anymore”

CORRELATION DOES NOT IMPLY CAUSATION

Spurious Correlations

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

TILBURG UNIVERSITY

Data Science: Prediction

- What is the goal?
- What are meaningful features?
- How can feature values be obtained?