Introduction to Data Science 1

Data Analysis

Pattern Recognition

Data Mining

Machine Learning

Data Analysis
the basics

DATA ANALYSIS IS
ABOUT PREDICTION
or
ABOUT RECOGNISING PREVIOUSLY UNSEEN THINGS

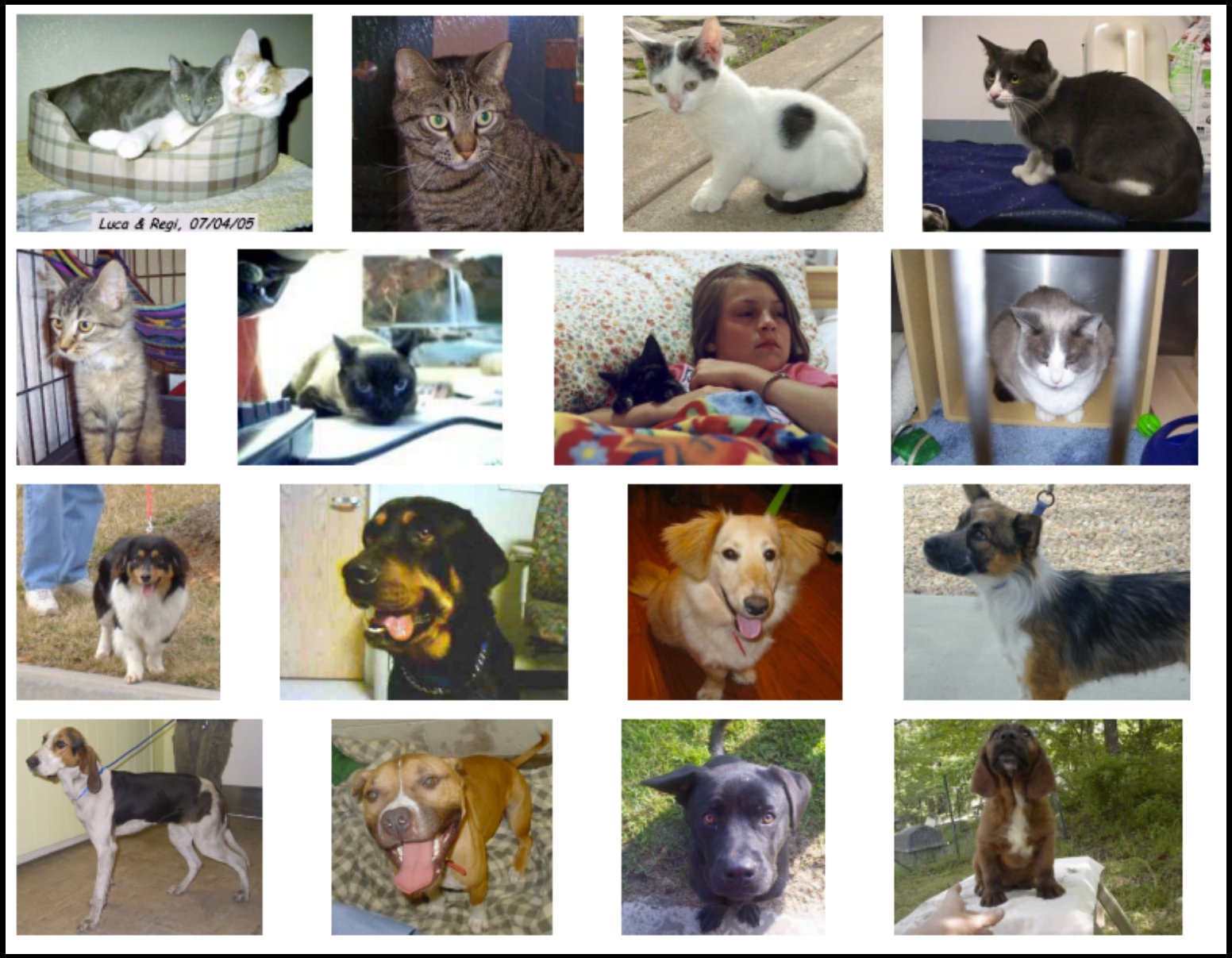
2 types of prediction

CLASSIFICATION

REGRESSION

CLASSIFICATION

DOG/NO DOG?



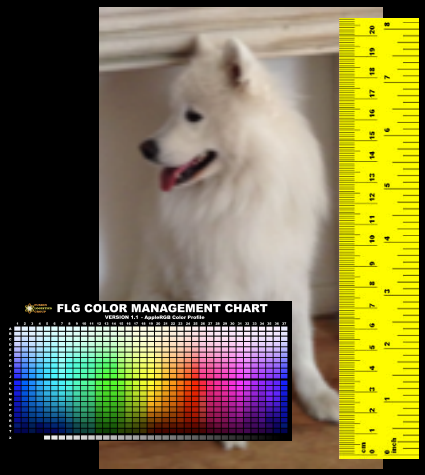
DOG/CAT?

REGRESSION

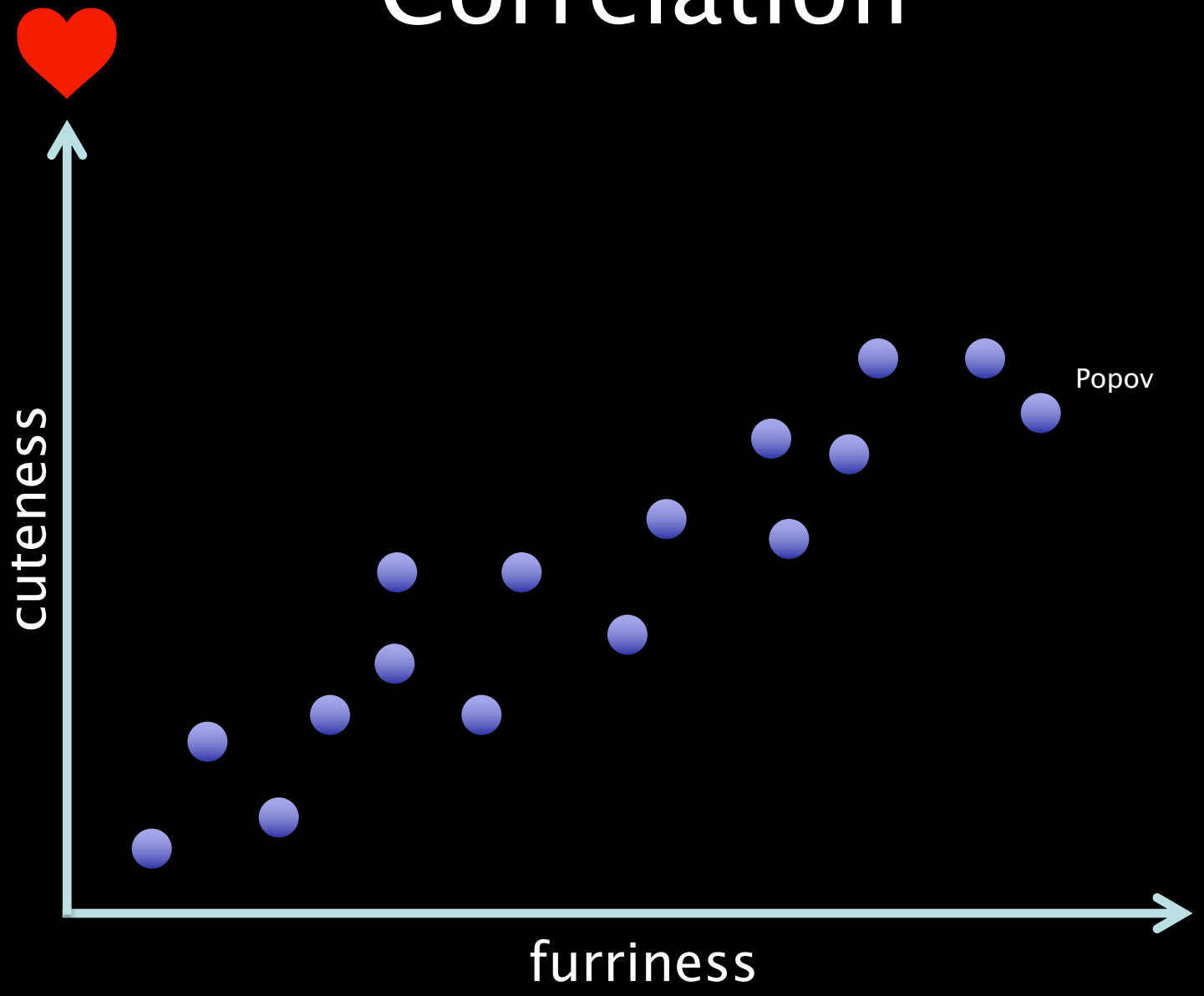
AGE?

Data Analysis concepts

- instances “Popov”
- labels “dog” “3 years”
- features color height furriness
- feature values “wit” “0.3m” “high”
- feature vector (“wit”, “0.3m”, “high”)

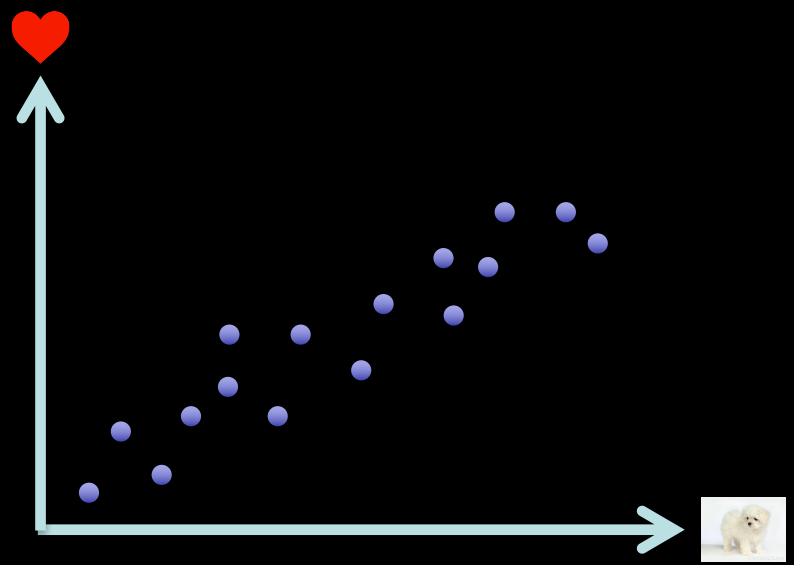


Correlation

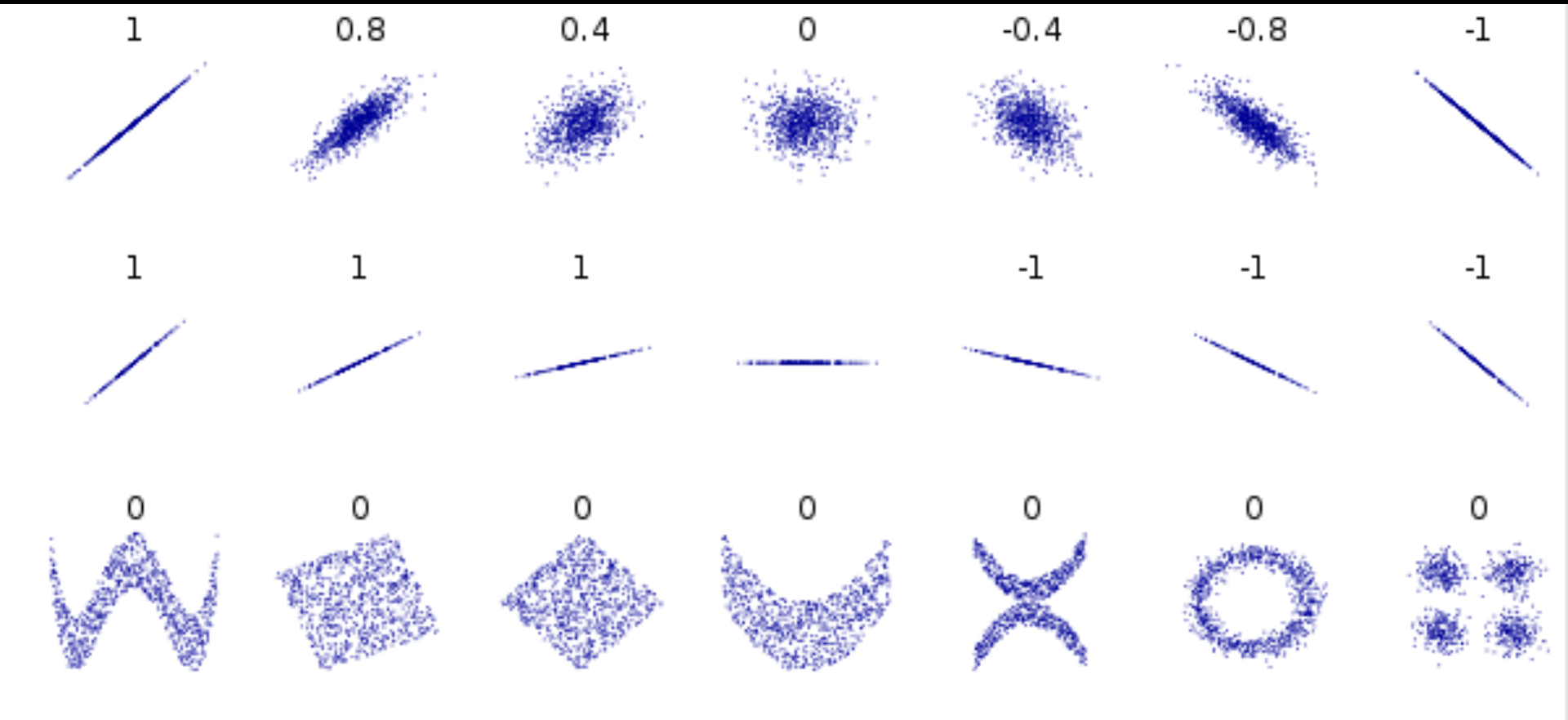


Correlation

- furriness predicts cuteness
- correlation allows for prediction



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



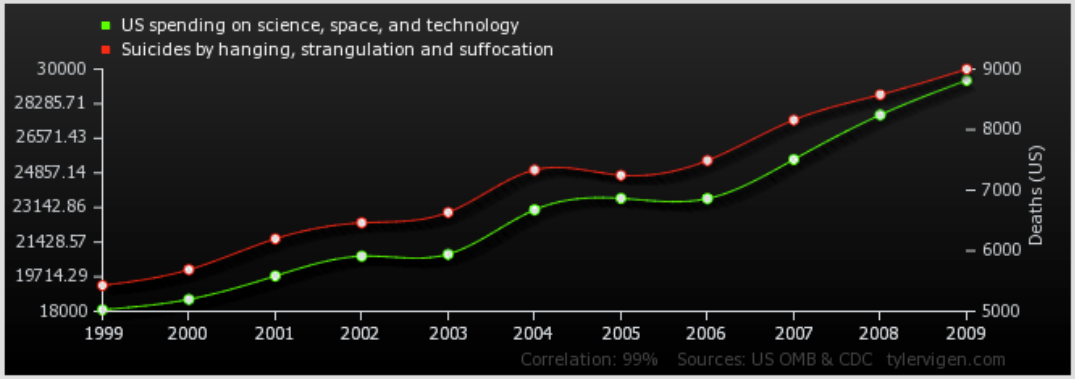
Background: Big Data, the impact on Science

THE END OF THEORY
Will the Data Deluge Makes the Scientific Method Obsolete? [6.30.08]
By Chris Anderson

(Big) Data Analysis

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

Data sets

Feature values are continuous
e.g., 0.1, 0.3, 0.4, 0.2, 0.7, …

Feature values are discrete
e.g., yes, no, male, small, …

Toy versus realistic datasets

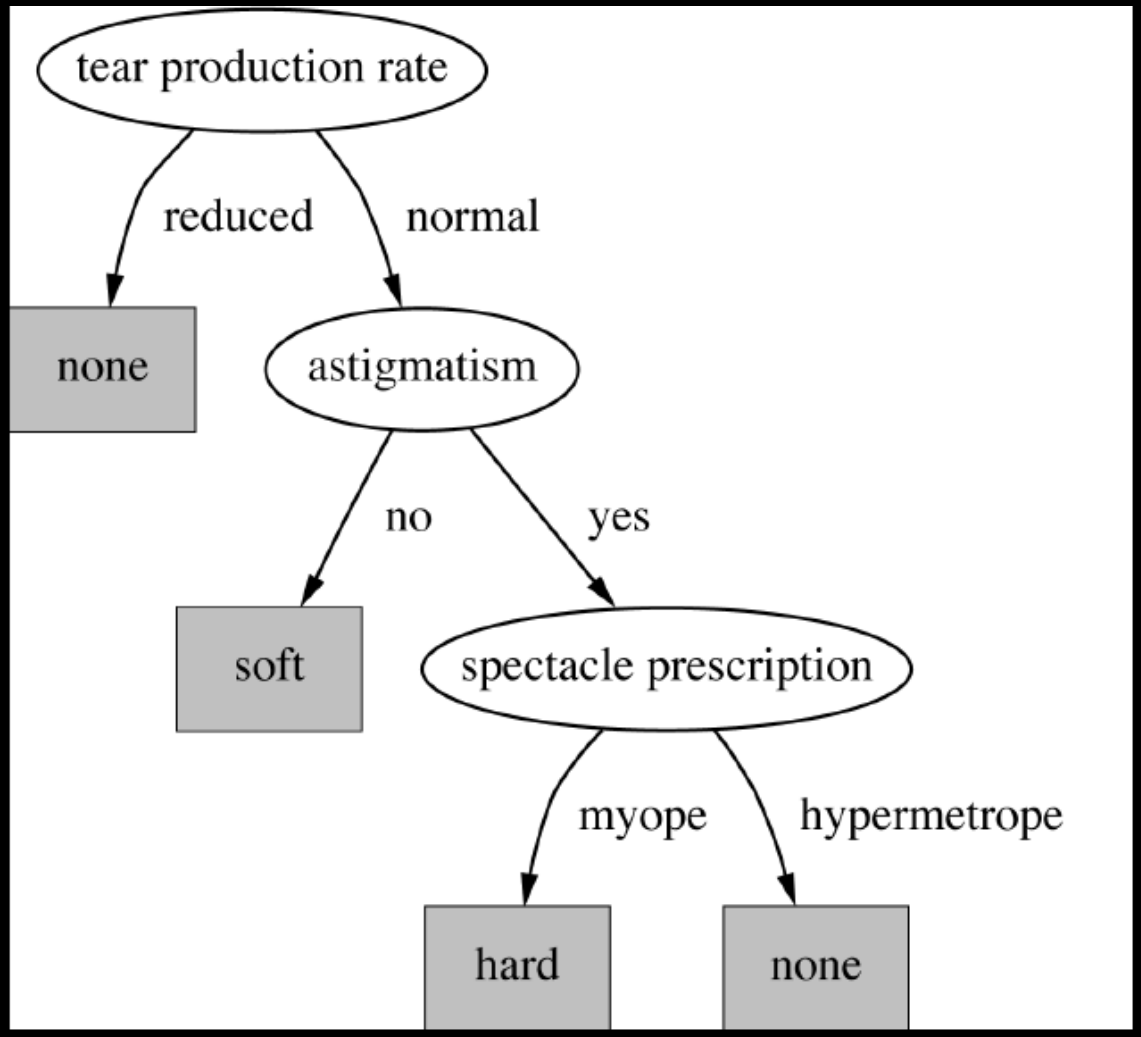
contact lenses (toy) dataset

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

Complete description in terms of rules

```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope
    and astigmatic = no  then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no
    and tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age young and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age = pre-presbyopic
    and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
```

Decision Tree



Description versus generalisation

The rules or decision tree provide a (complete) **description** of the contact lenses dataset. (A kind of summary.)

Depending on their structure, they may also **generalise** to novel cases (i.e., cases not included in the contact lenses dataset.

Prediction with Data Mining

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

Data Science tools

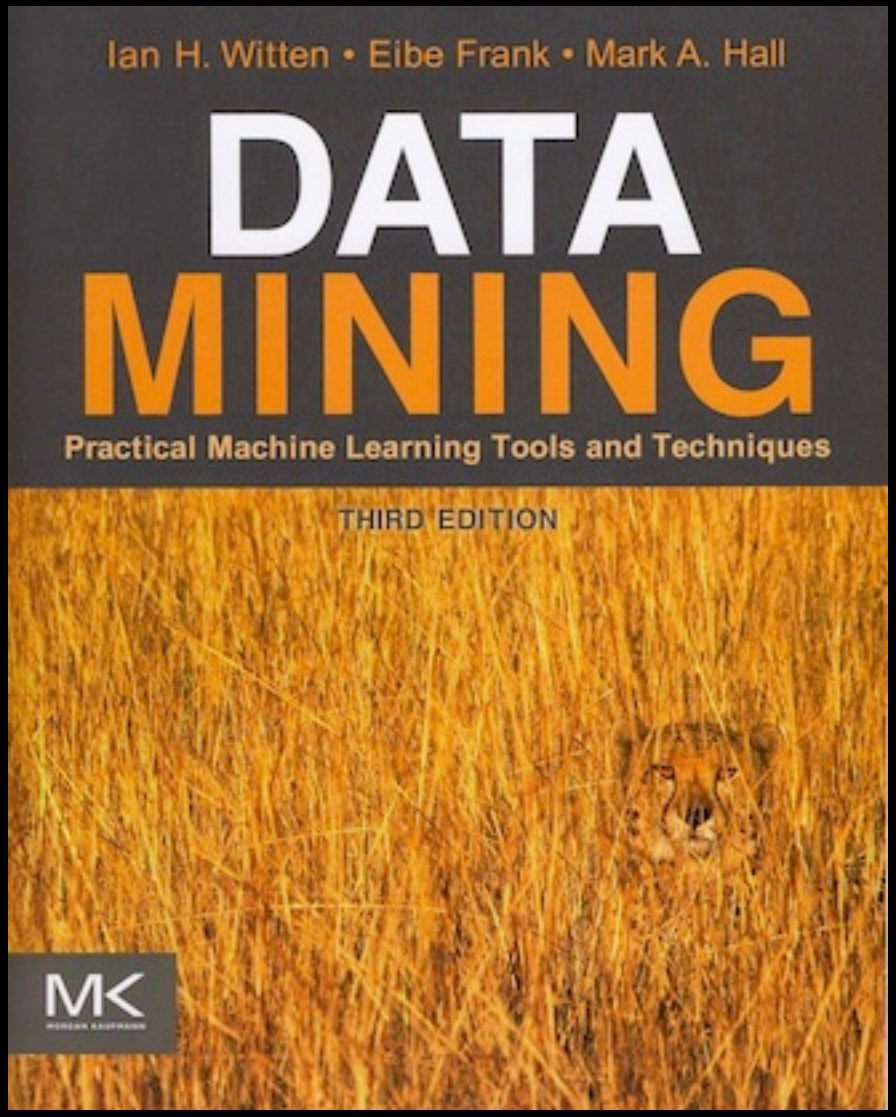
WEKA

Orange

Python

R

WEKA book


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Why WEKA?

Data Scientists work with (a.o.):

R (statisticians) or

Python (computer scientists)

WEKA and Orange are for beginners