Machine Learning & Artificial Intelligence for Data Scientists: Introduction

Ke Yuan
https://kyuanlab.org/
School of Computing Science

Who are we?

Dr Ke Yuan
Senior Lecturer

Machine Learning
and Computational
Biology

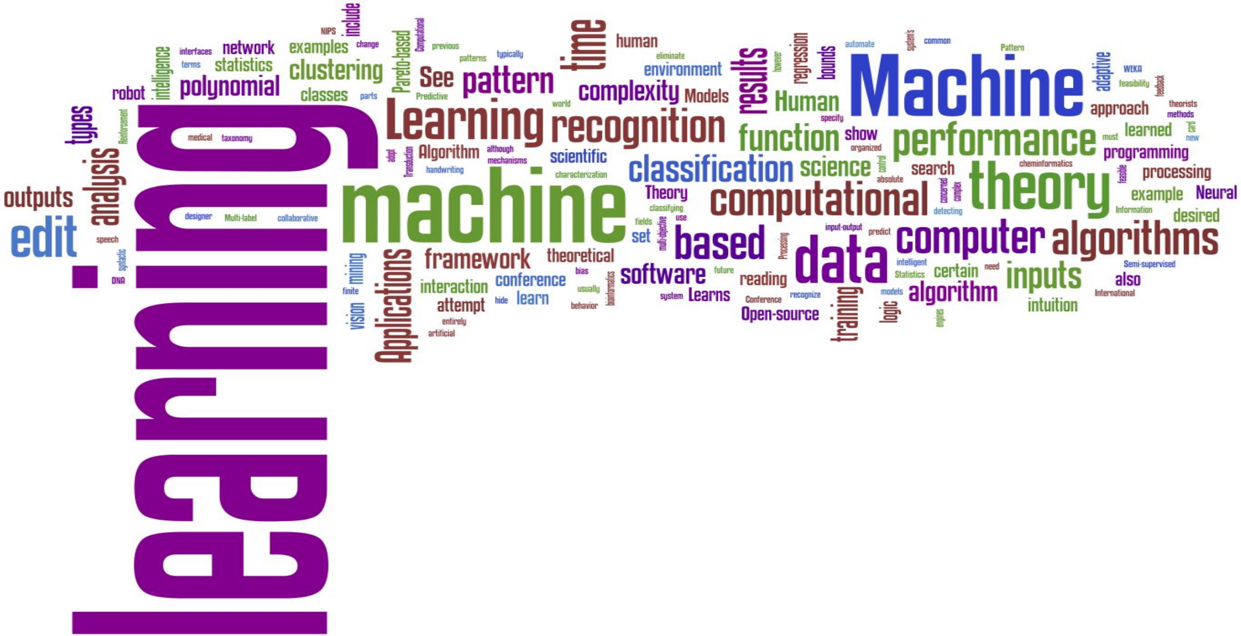
Dr Fani Deligianni
Senior Lecturer

Machine Learning
and Healthcare

Dr Tanaya Guha
Senior Lecturer

Machine Learning
and Human Machine
Interaction

What is Machine Learning?



What is Machine Learning?

- Machine learning starts with **data**.
- Observations of **objects**:
  - Observations of people (preferences, health, etc)
  - Observations of the world (images, sounds, etc)

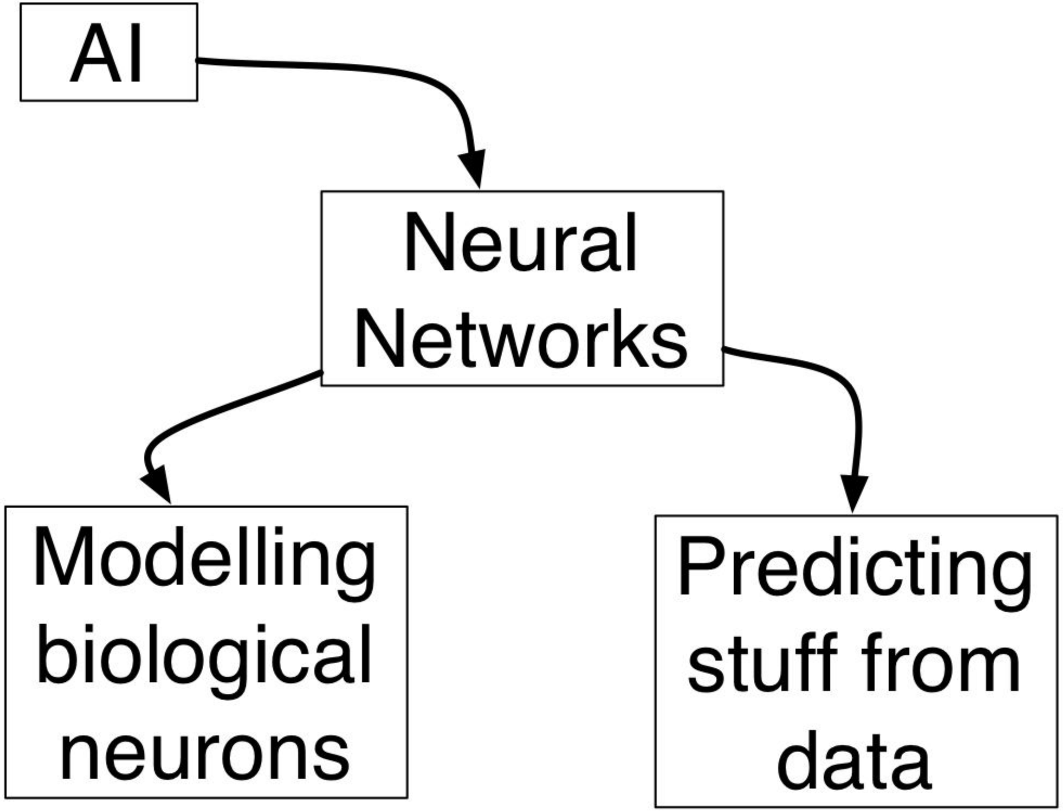
What is Machine Learning?

- Can we find **similar** objects?
- Can we make **predictions** about objects?
- Can we **learn** something about the objects?
- Can we **group** the objects?

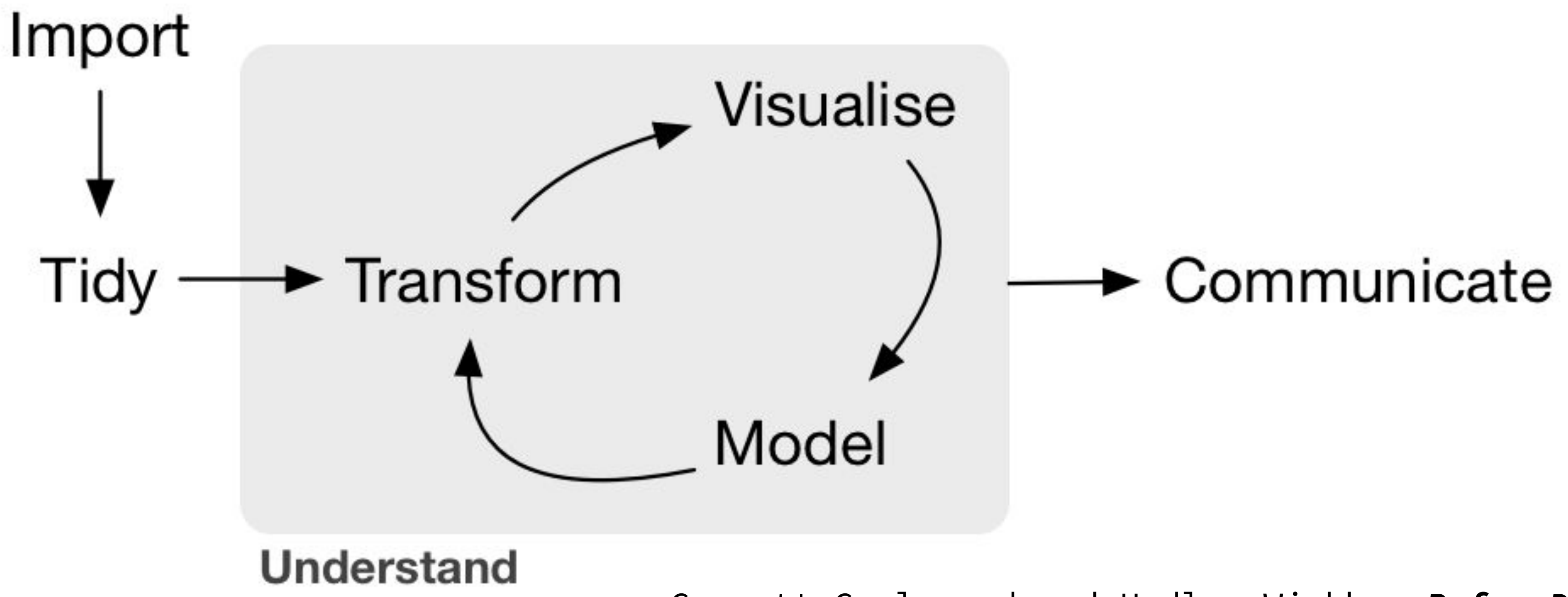
Algorithms

- Machine Learning could be thought of as an ever-growing set of algorithms.
- But the algorithms are hard to use.
- Many have to be tuned. It is important to understand them.

Where did it come from?



Where is ML in Data Science?

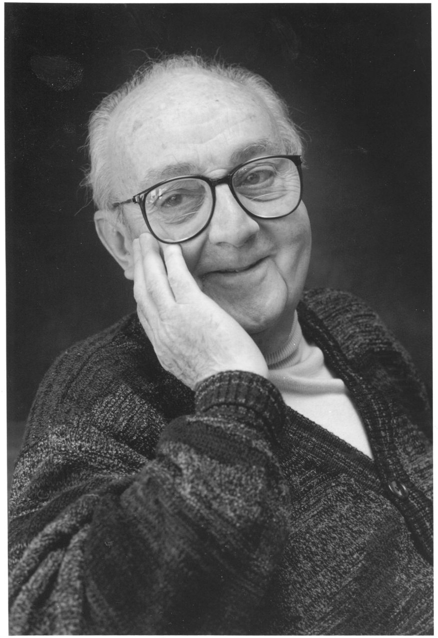


Garrett Grolemund and Hadley Wickham ***R for Data Science*** https://r4ds.had.co.nz/

What do we want to achieve?

“All models are wrong, but some are useful.”

George E. P. Box



Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an “ideal” gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

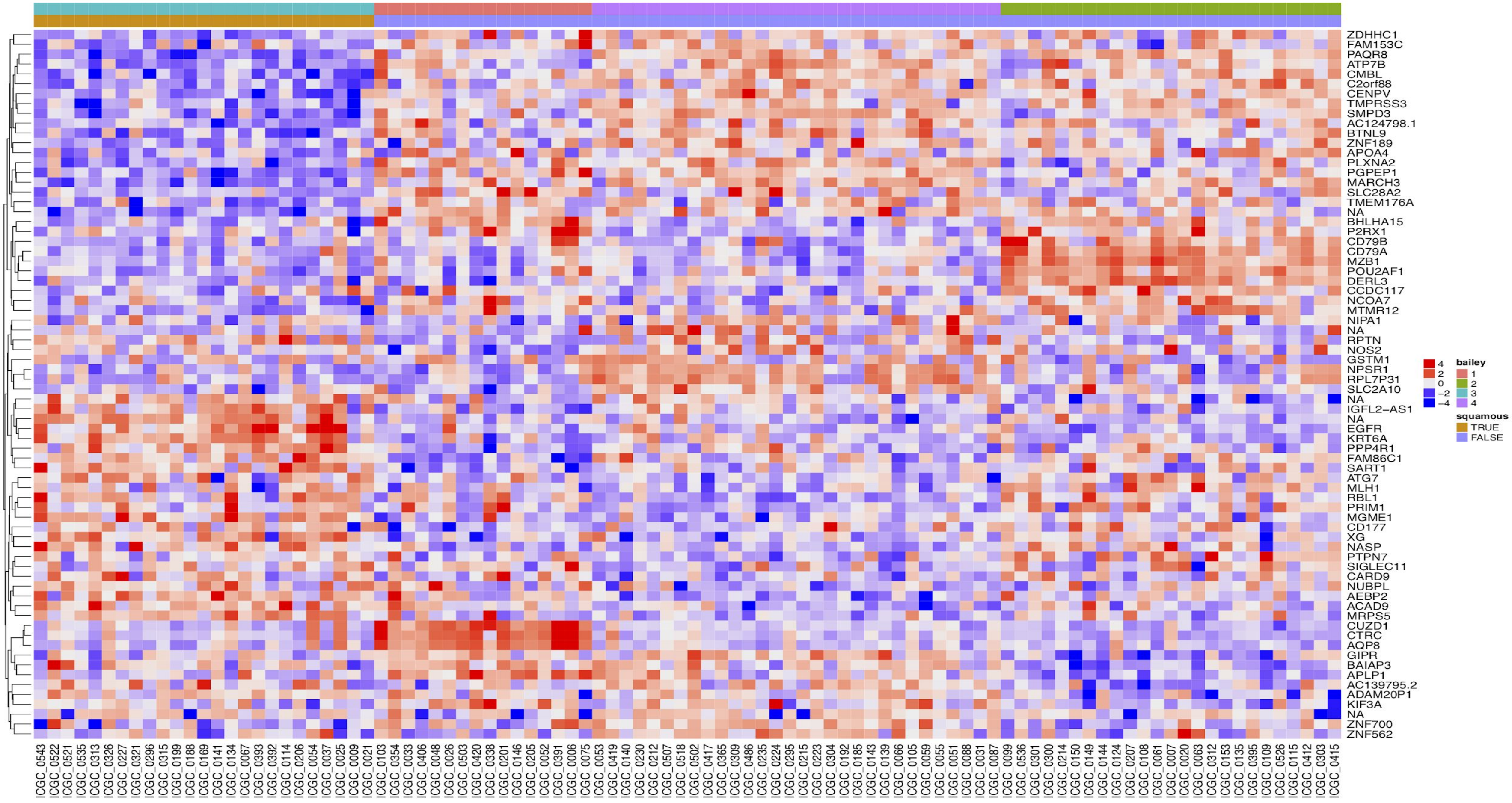
For such a model there is no need to ask the question “Is the model true?”. If “truth” is to be the “whole truth” the answer must be “No”. The only question of interest is “Is the model illuminating and useful?”.

What will you learn?

- Fundamental idea of ‘learning’ from data.
- Caveats: what people do wrong.
- Several common algorithms (that I think are important...not exhaustive).
- How to visualise results?
- How to apply and compare these algorithms in 5 case studies?

Who uses it? Google, Microsoft, Amazon, etc

Who uses it? Biotech/Pharmaceutical companies



Some examples within SoCS

- **Computational Biology/Bioinformatics**
  - Cancer Biology
  - Antimicrobial resistance
- **Information Retrieval**
  - Search & Recommendation Systems
  - Building conversational agent
- **Human Computer Interaction**
  - Speech recognition
  - Gesture recognition

What we'll cover

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

We will not cover:

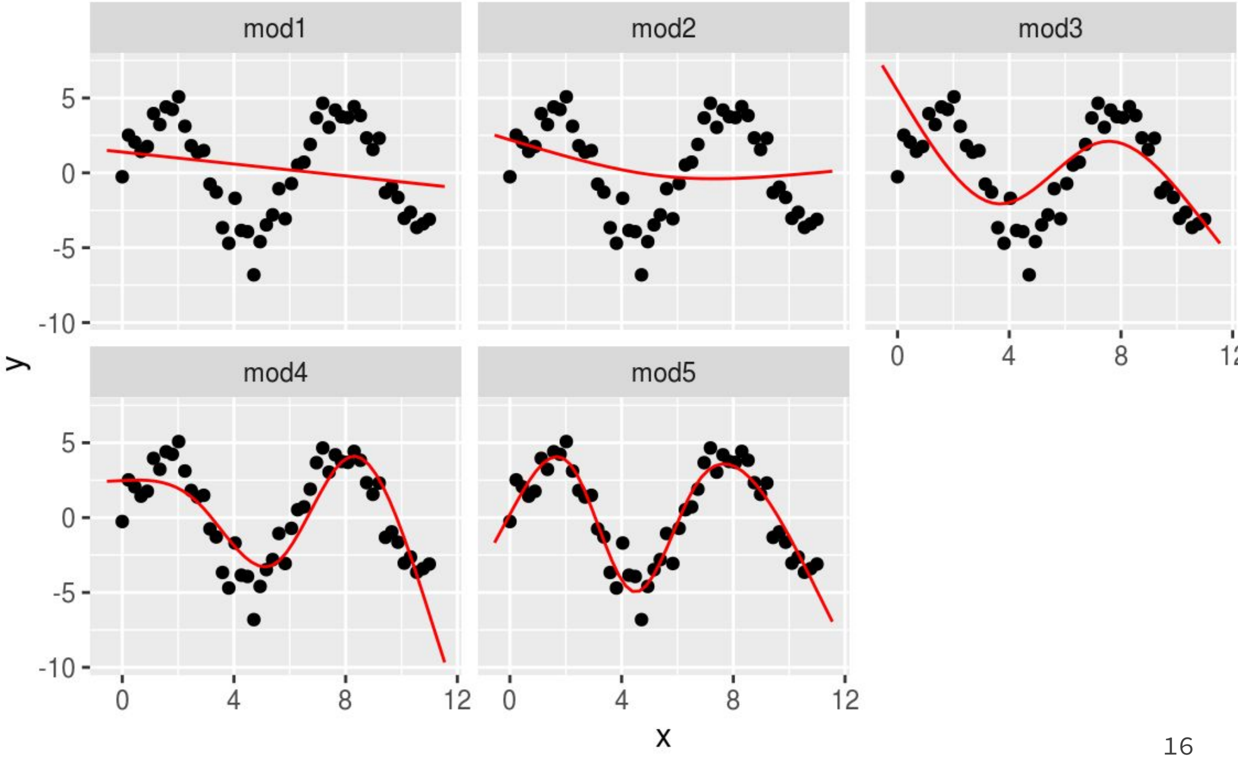
- Deep Learning

Course structure

- 6 introduction units
  - What is “learning from data”
  - Introduction to the problems of Regression, Classification, Clustering, and Projection.
  - Training, validation, and testing.
  - Performance metrics and the importance of robust baselines.
  - Common pitfalls.
  - Presentation of results.

Supervised Learning: Regression

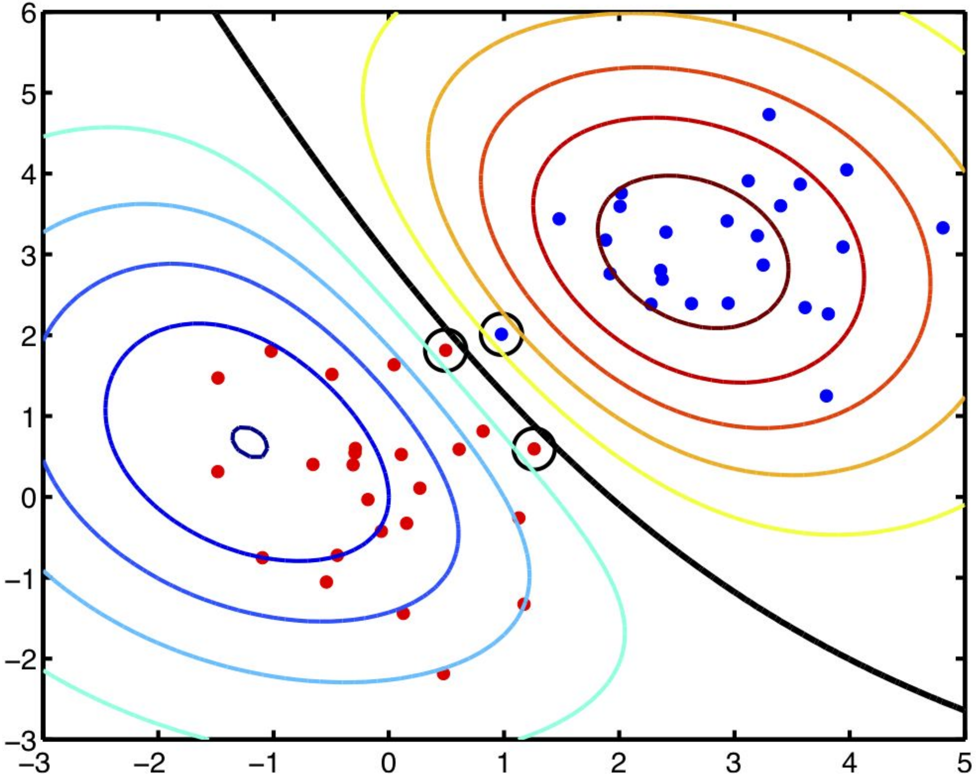
- Learning a continuous function from a set of examples.
- Example: Predicting stock prices (x might be time or some other variable of interest).



Supervised Learning: Classification

- Learning a rule that can separate objects of different types from one another

- Examples: Disease diagnosis, spam email detection.



Predicting skin cancers

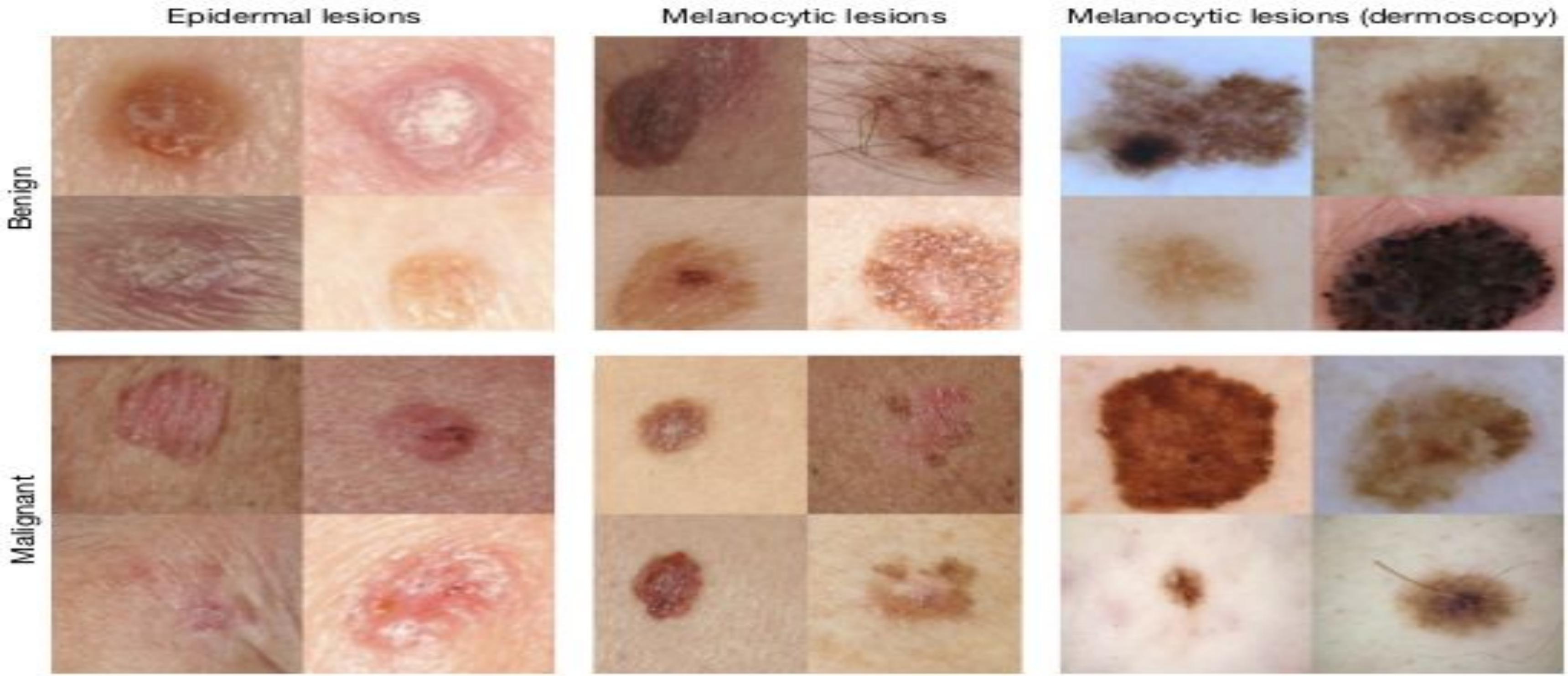
LETTER

doi:10.1038/nature21056

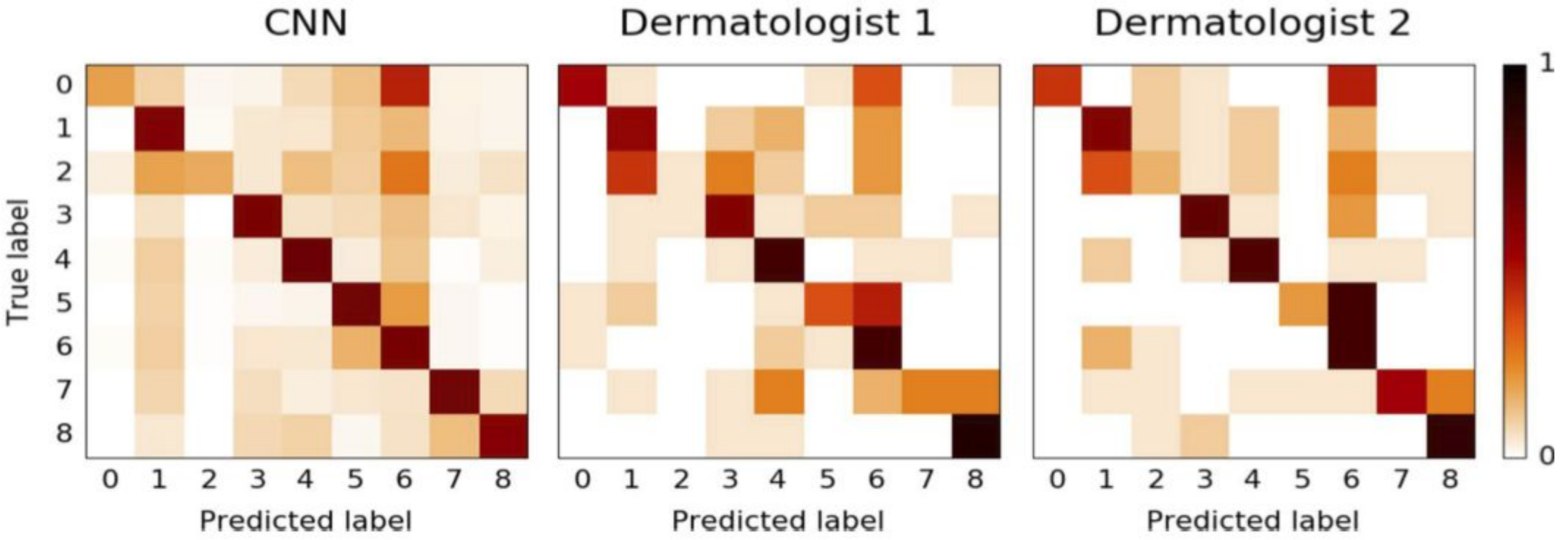
Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Predicting skin cancers

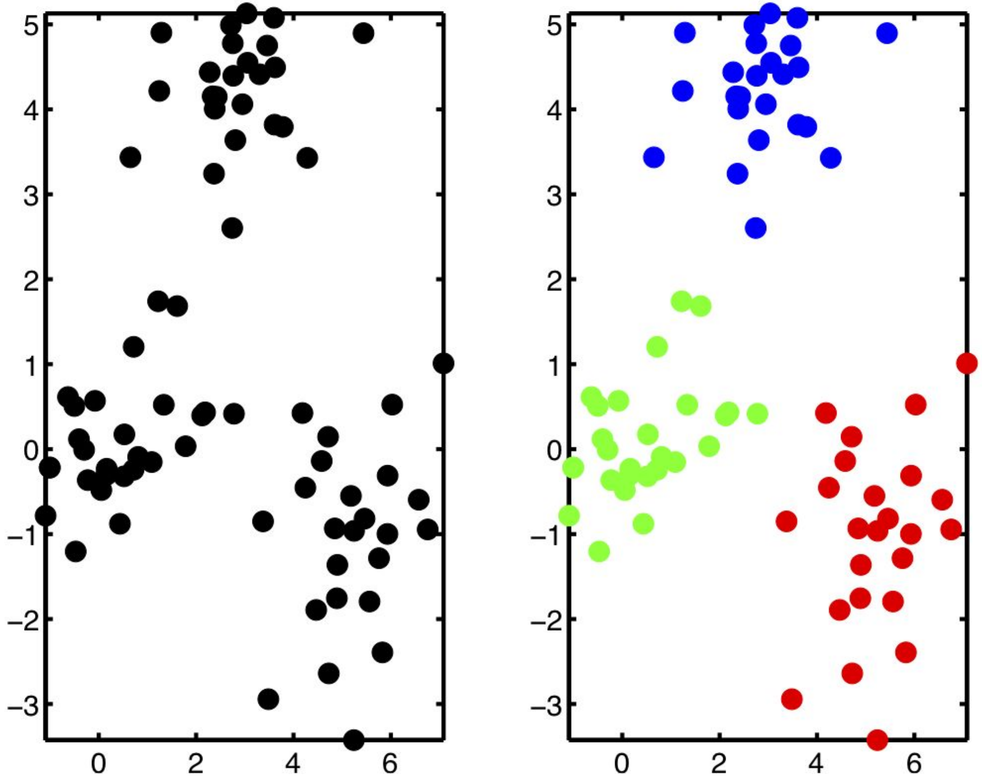


Predicting skin cancers

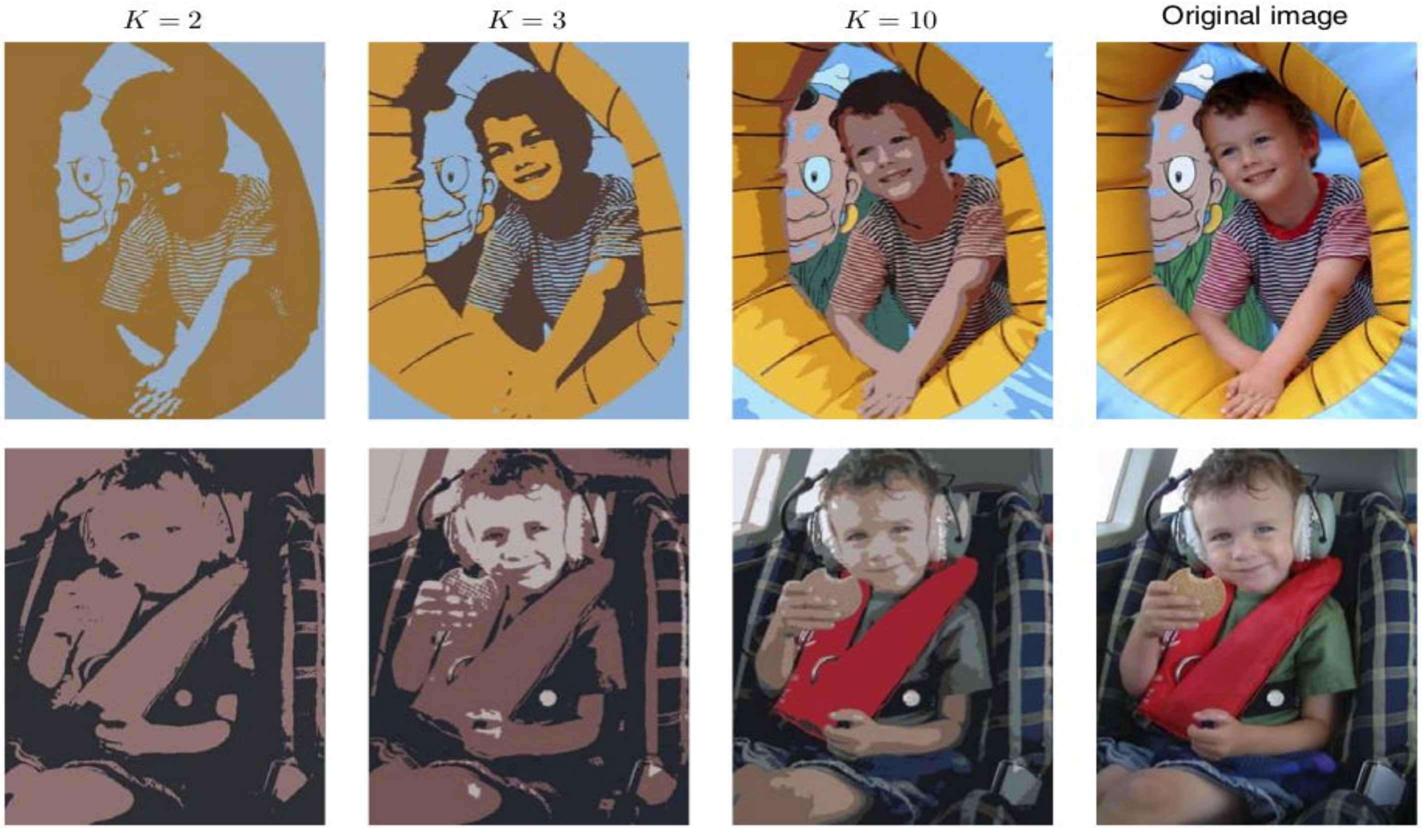


Unsupervised Learning: Clustering

- Finding groups of similar objects.
- Examples: People with similar ‘taste’, genes with similar function

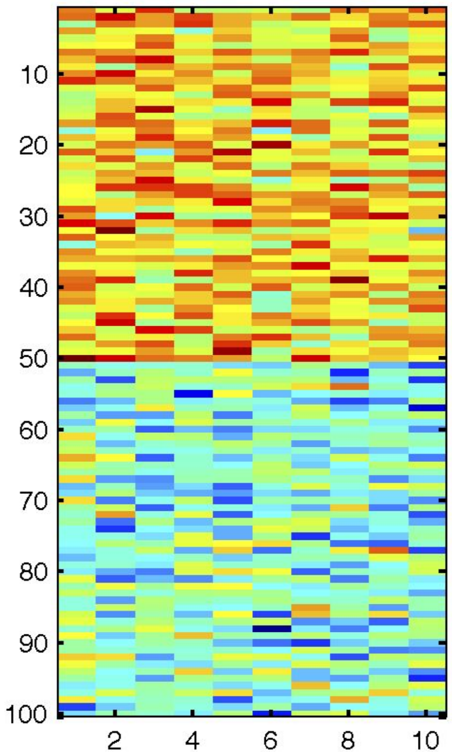


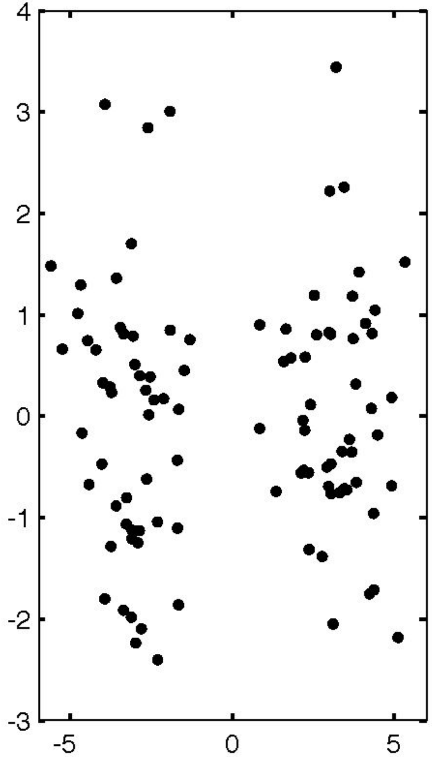
Clustering



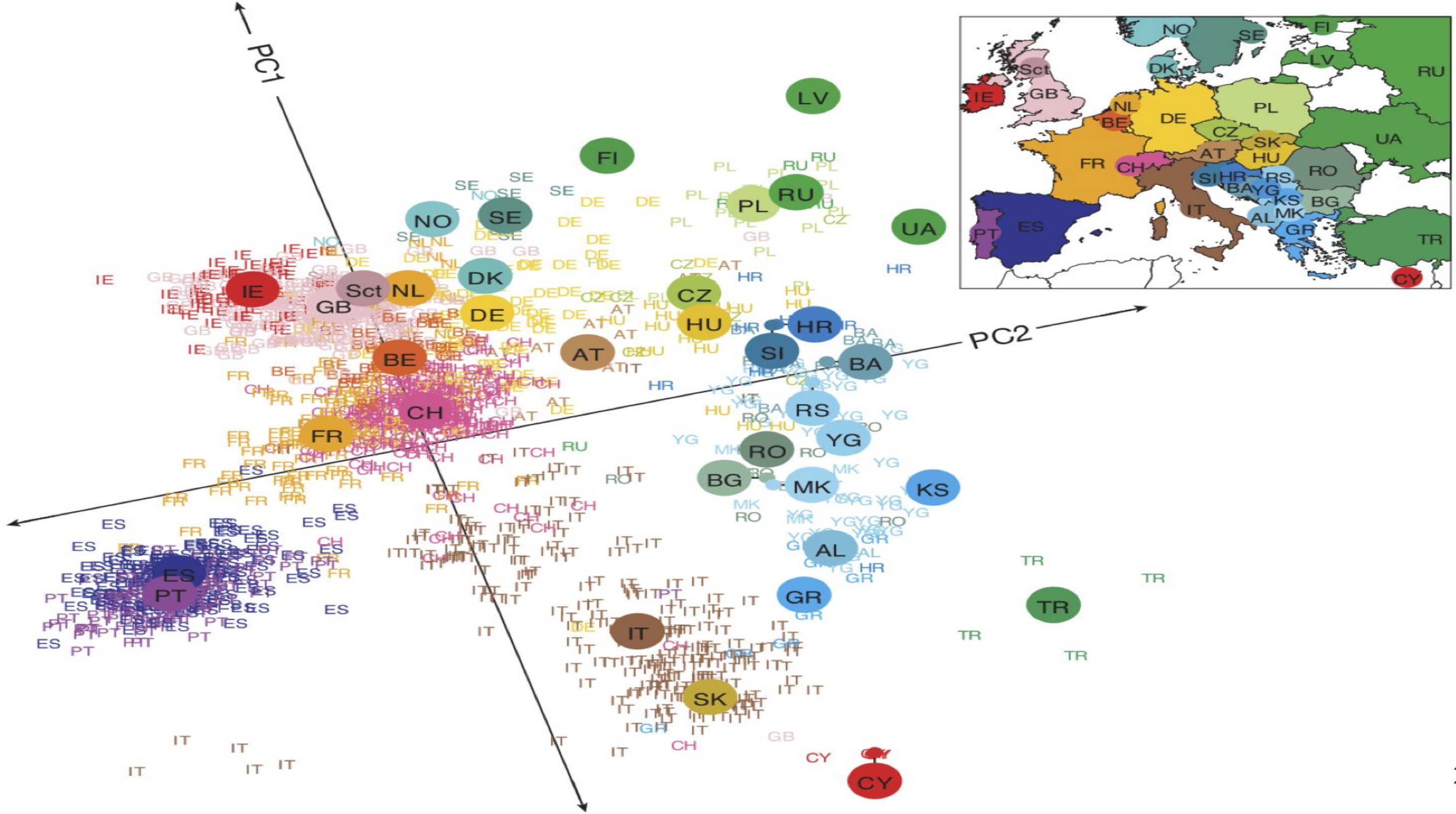
Unsupervised Learning: Projection

- Reducing the number of variables - e.g. from 10 to 2.
- Visualising complex data.





Novembre et al. (2008) doi:10.1038/nature07331



Course structure

- **4 units of case studies** covering different datasets, visualisation techniques, models and learning algorithms. Taught by multiple staff.

Guest lecturers for case studies

Dr Sebastian Stein
Research Associate

Machine Learning,
Close-Loop Data Science

Dr John Williamson
Senior Lecturer

Machine Learning,
Computational
Interaction

Case study

- A dataset and predictive / exploratory problem to be solved.
- An introduction to one or more algorithms.
- A practical session (1 hr in lab plus 1hr in students’ owen time).
- A wrap-up session .

Assessment

- Exam: 60%, 6 introduction units.
- Coursework: 40%, choose 2 out of 4 case studies.

Don’t panic! Have fun!