MANCHESTER 1824
The University of Manchester

Introduction to Applied Machine Learning

David Topping
david.topping@manchester.ac.uk

https://github.com/loftytopping

@lanky_manky

Proposed structure of talk

- What is it and how can we use it?

- Data dependencies
 - Quantity
 - Features
 - Fitting/testing

- Neural networks

- Todays exercises

Question: ‘How many groups of aerosols have we sampled?’
 - Clustering
 - Convolutional Neural Networks
 - Autoencoders and Network Optimisation

- What next?
 - Build your tools and software stack

What is it and how can we use it?

'Machine learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.'

https://blogs.nvidia.com/blog/2016/07/29/


MENU nature International journal of science

Subscribe Search

COMMENT · 30 JULY 2019

Three pitfalls to avoid in machine learning

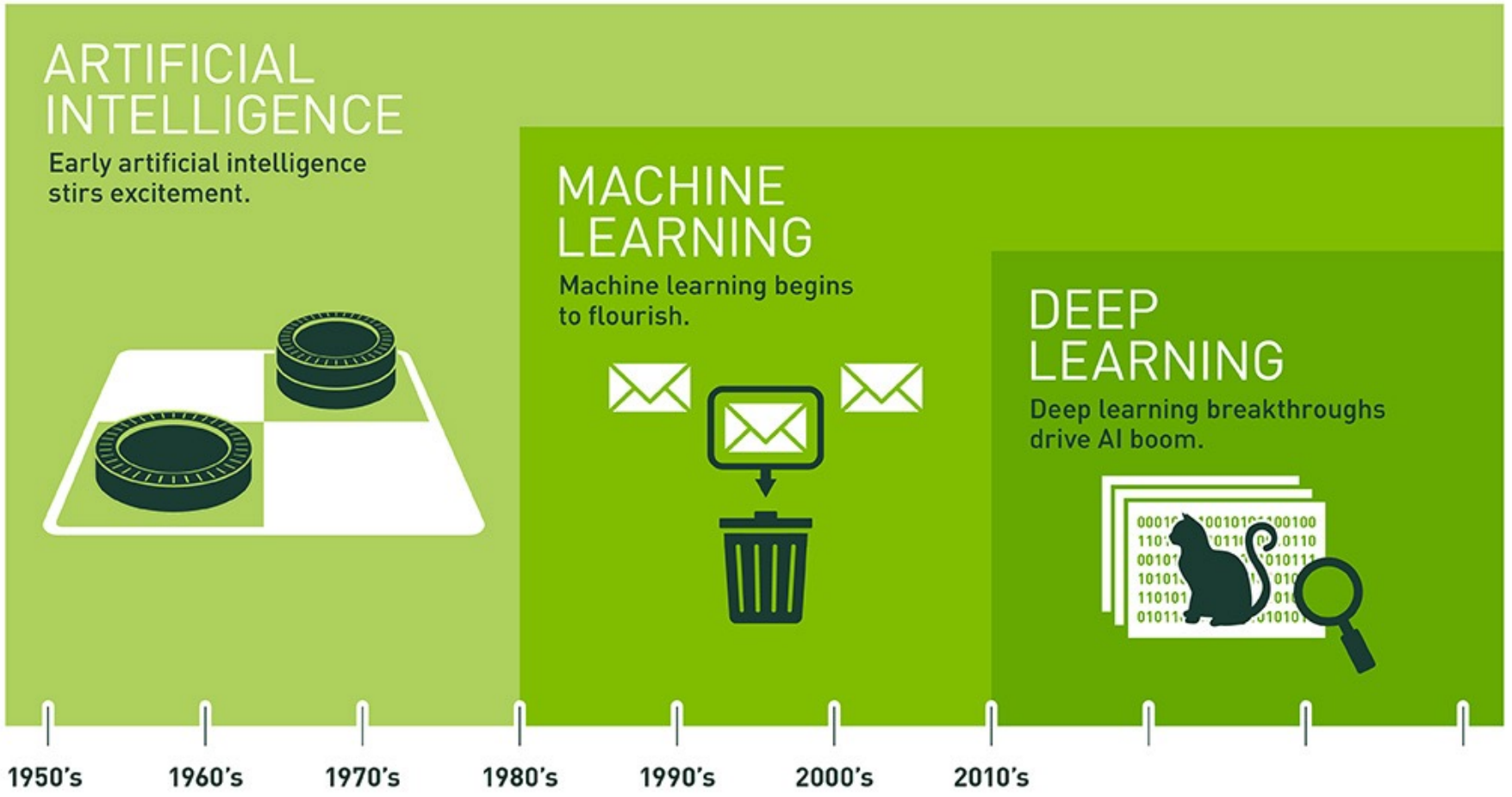
As scientists from myriad fields rush to perform algorithmic analyses, Google's Patrick Riley calls for clear standards in research and reporting.

Patrick Riley

'When a new piece of lab equipment arrives, we expect our lab mates to understand its functioning, how to calibrate it, how to detect errors and to know the limits of its capabilities. So, too, with machine learning. There is no magic involved, and the tools must be understood by those using them.'

MANCHESTER 1824
The University of Manchester

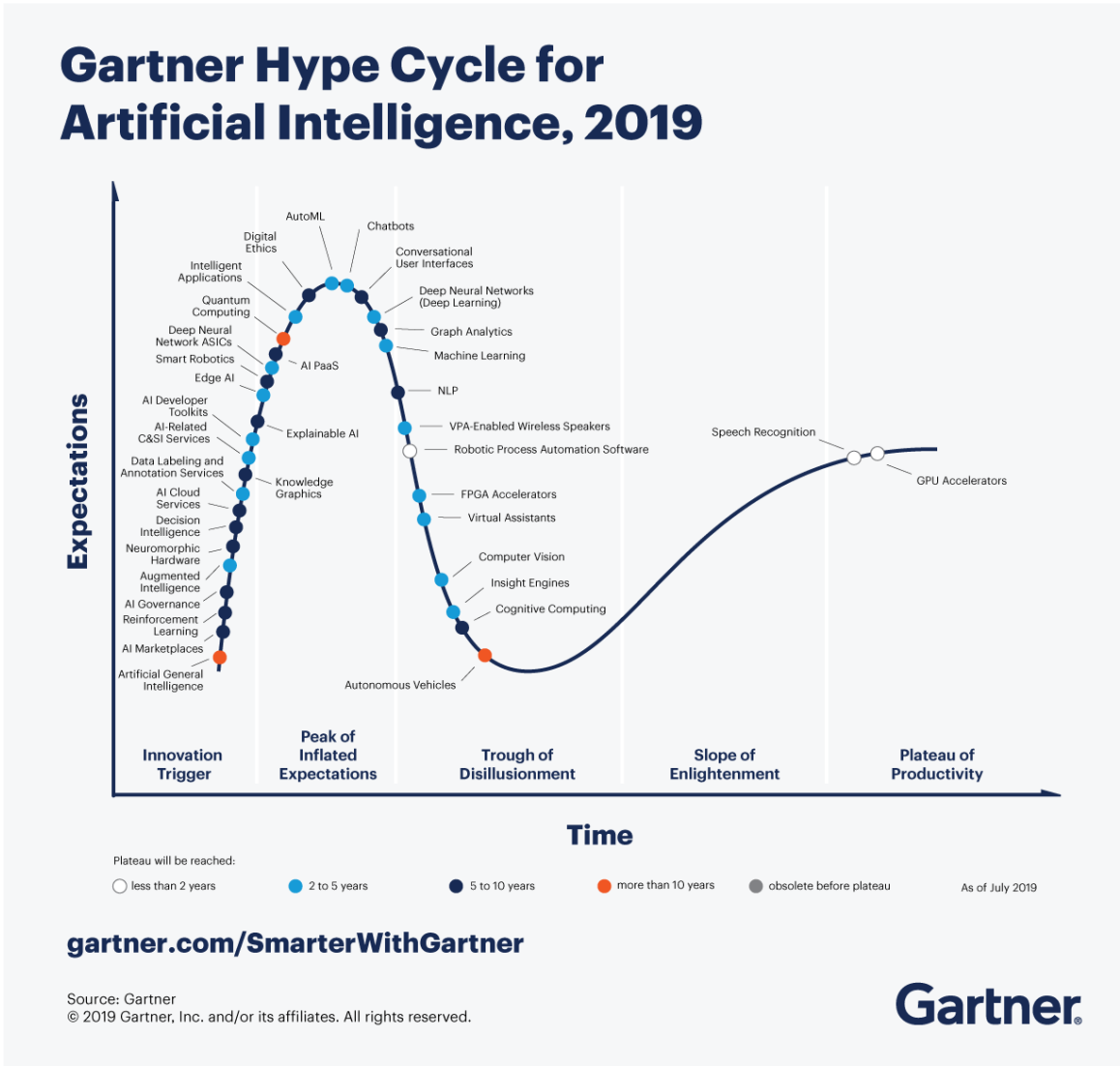
Long history behind current accessibility

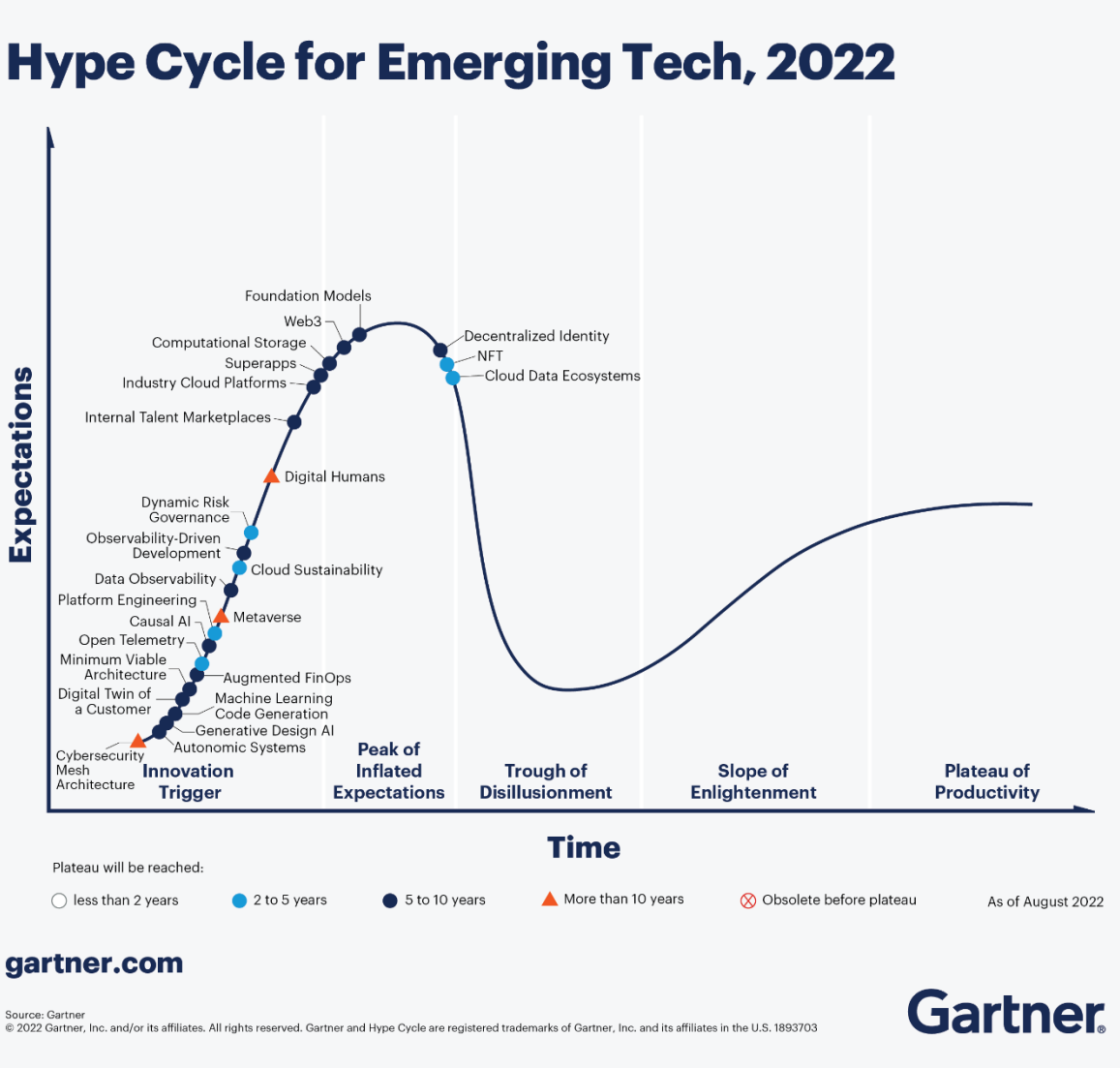


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://news.developer.nvidia.com/on-demand-webinar-deep-learning-demystified/

Constantly at/near the peak of expectation?

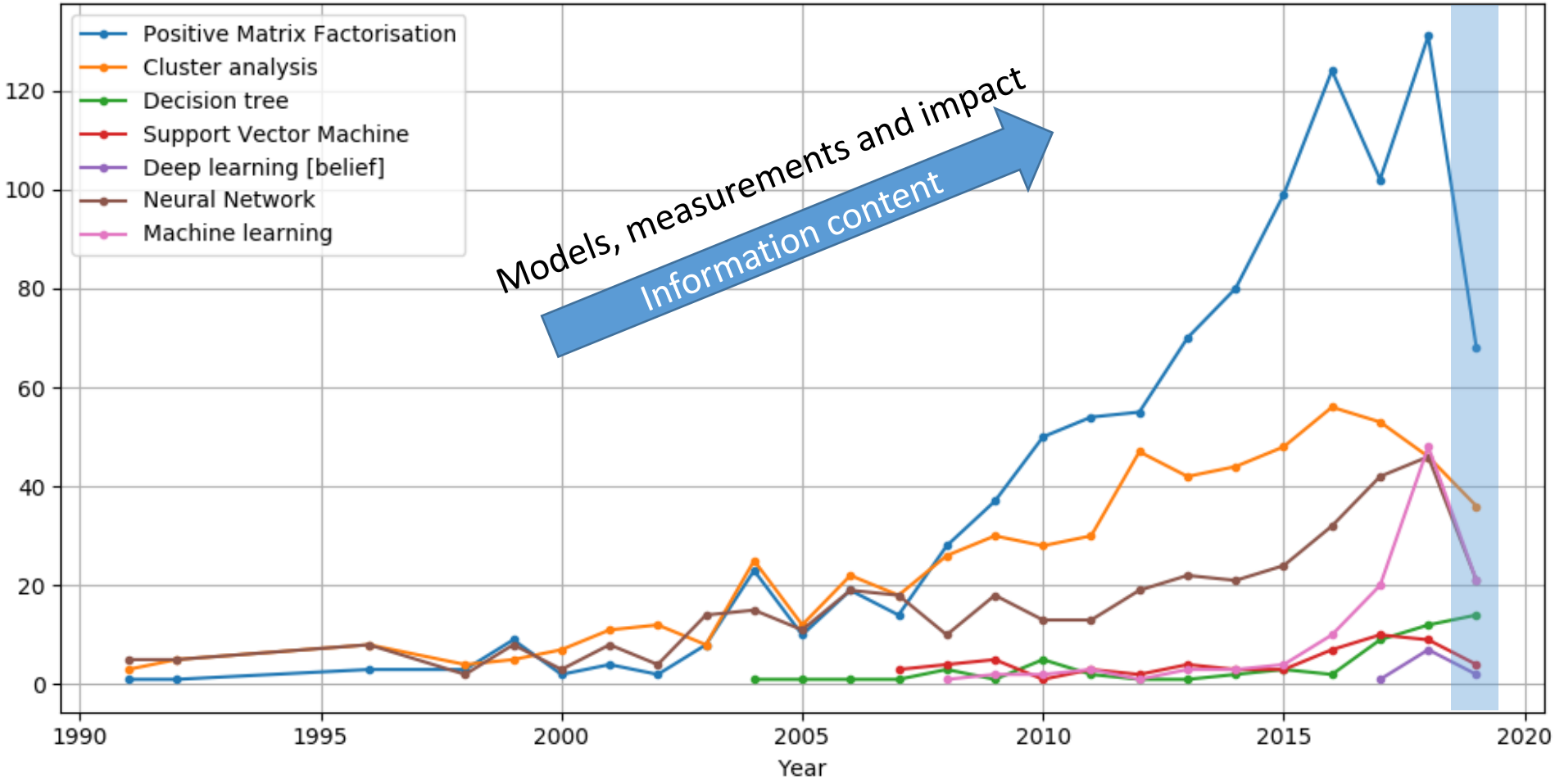




Emerging technologies for 2022 fit into three main themes: evolving/expanding immersive experiences, accelerated artificial intelligence automation, and optimized technologist delivery.

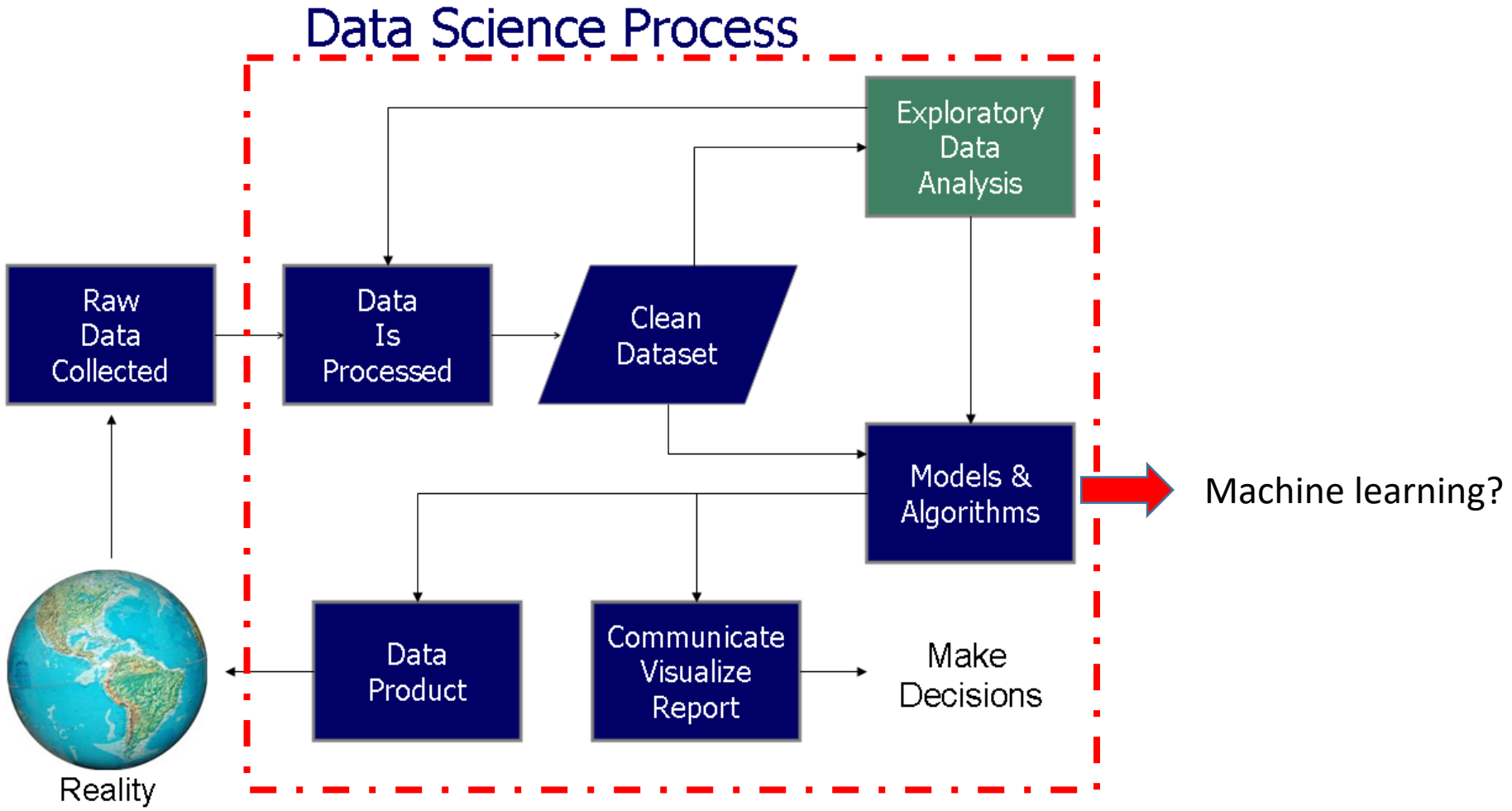
MANCHESTER 1824
The University of Manchester

Expected increase in publications.



MANCHESTER 1824
The University of Manchester

- What am I solving?
- What data do I have?
- Do we have any meta data?
- Do we need more dat
- Why cant we use existing tools?
- What ML methods might help?
- Who will use my results?

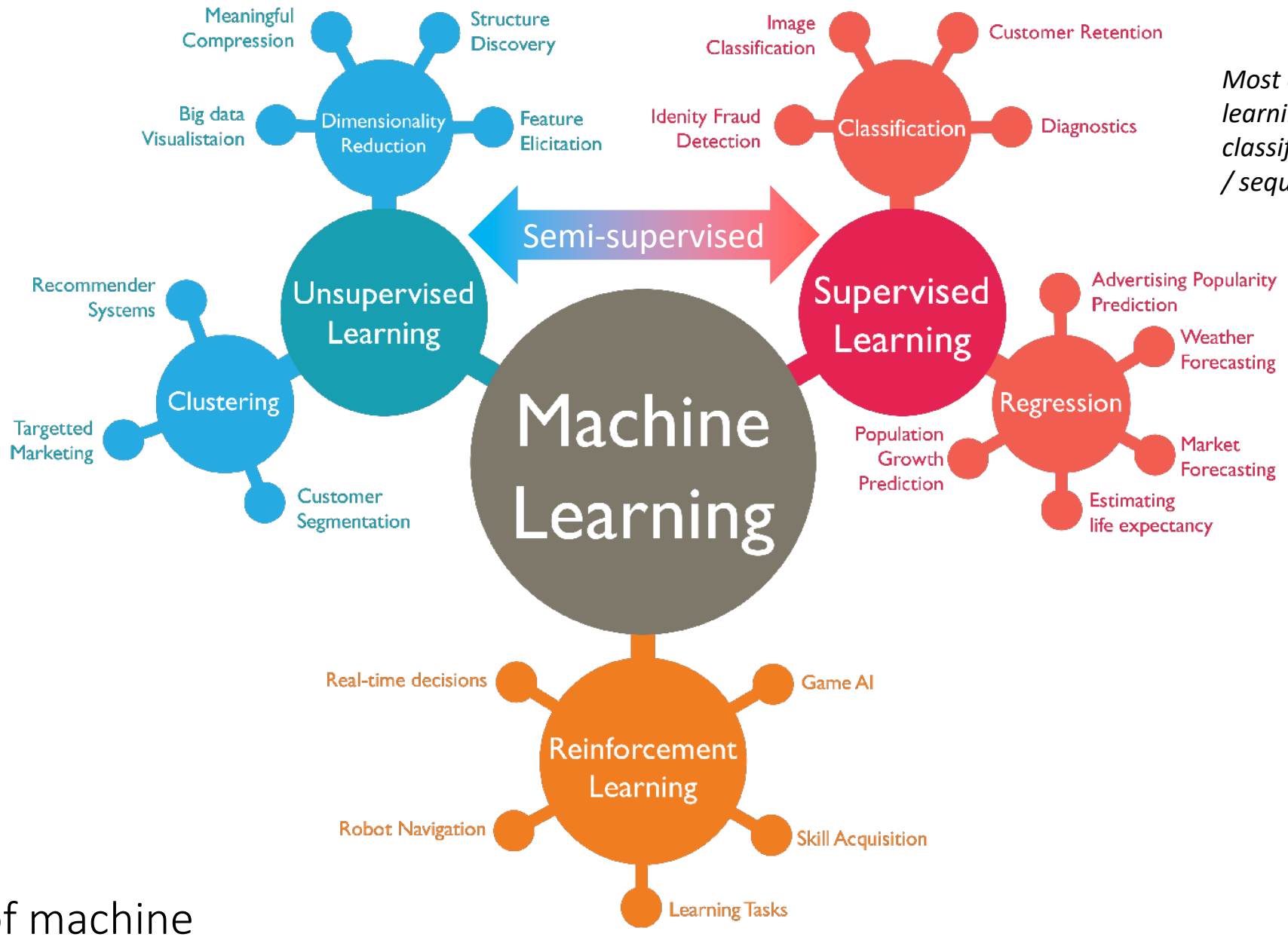


Doing Data Science, O'Reilly ISBN: 978-1-449-35865-5. ↑ Forbes-Gil Press-A Very Short History of Data Science-May 2013

MANCHESTER 1824
The University of Manchester

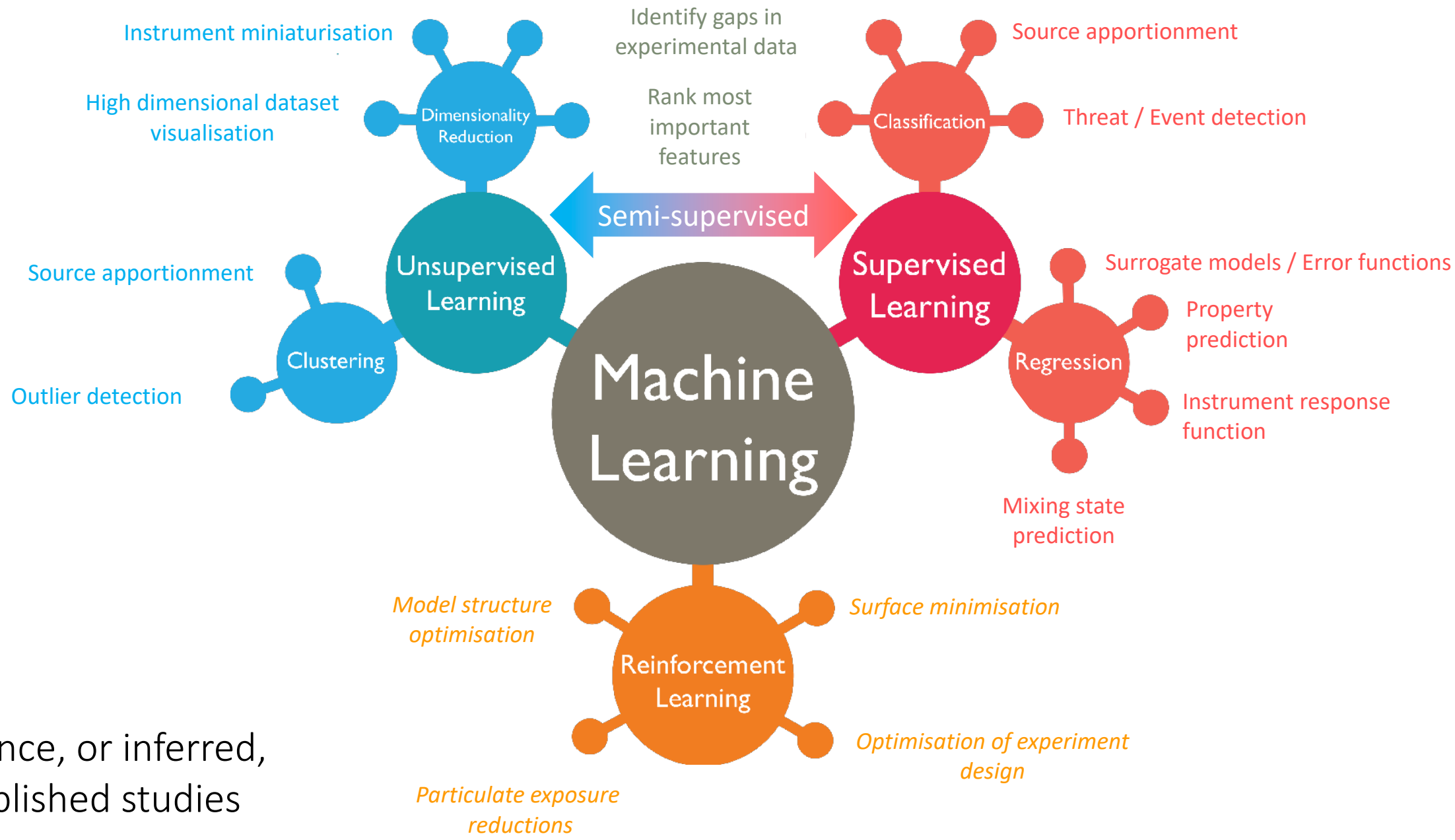
IGAC attendee journey

• What am I solving?	• Chemical Simulation	• Aerosol classification	• High-res pollutant maps
• What data do I have?	• Rates, initial concentrations	• Mass spec / scattering.	• Model output, remote sensing, single sites, land use, met fields
• Do we have any meta-data?	• Some…	• Some…	• Some…
• Do we need more data?	• Probably?	• Probably?	• Probably?
• Why cant we use existing tools?	• ODE solvers too slow / cant integrate in regional model	• No standardized method	• No standardized method / poorly documented
• What ML methods might help?	• Neural nets or any regression technique	• Time series clustering / computer vision	• Computer vision / Generative methods
• Who will use my results?	• Academics / large-scale modellers	• Academics	• Academics / Policy makers



Most examples of deep learning including classification and time series / sequence analysis

Broad areas of machine learning applications



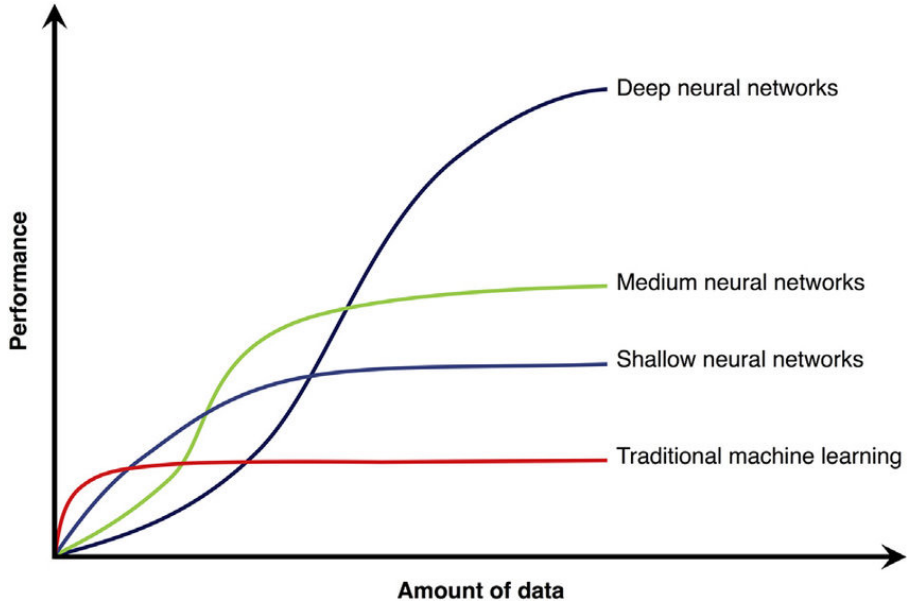
Evidence, or inferred, in published studies

MANCHESTER 1824
The University of Manchester

Data dependencies

How much data do I need?

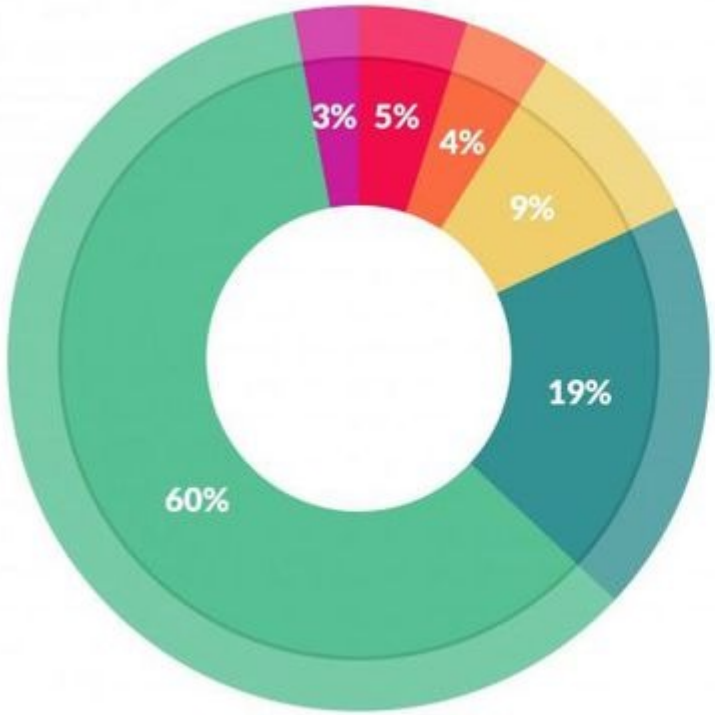
'…It has been well established both across industry and academia that for a given problem, with large enough data, very different algorithms perform virtually the same…'



More data enables us to try and minimise both bias and variance

What if the cost of accessing new data is simply too expensive or not easily resourced?
What if I don't have access to a massive GPU or lots of memory?

Data preparation takes up >90% of the machine learning pipeline



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

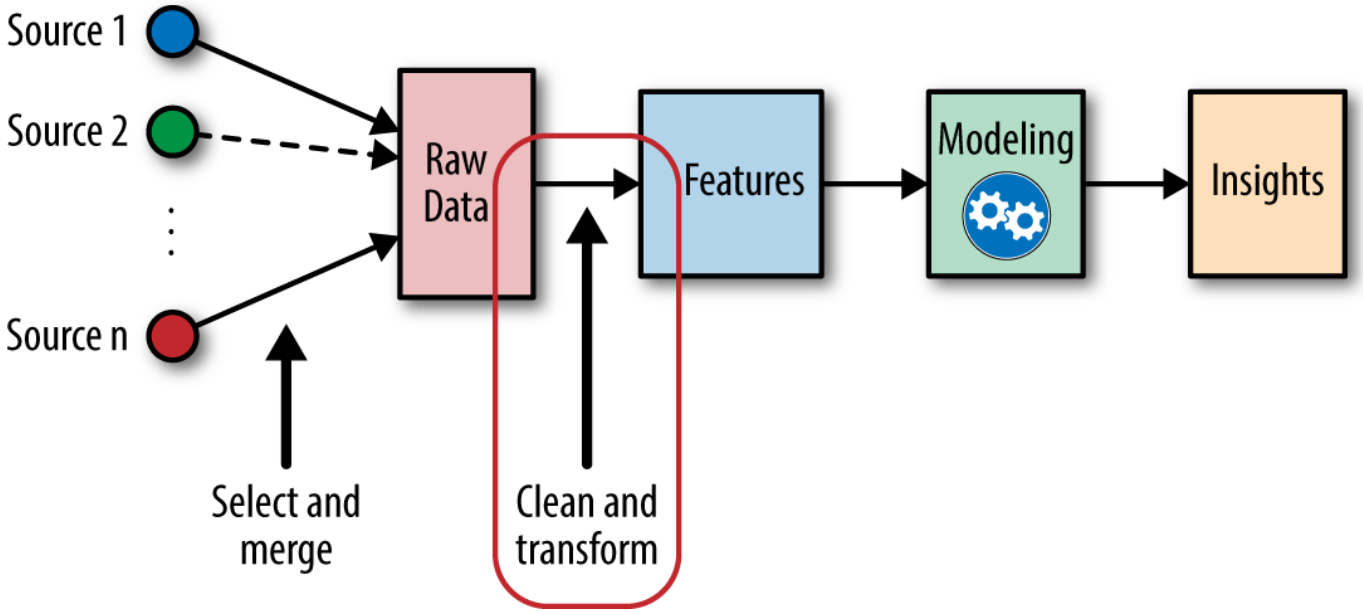
https://www.forbes.com/sites/gilpress/2016/03/23

'76% of data scientists view data preparation as the least enjoyable part of their work'

Domain expertise and calibration standards essential. Your jobs are safe!

Features
Feature engineering

Using domain knowledge of the data to create features that make machine learning algorithms work.



What do I need to predict PM2.5 ahead of time?
I have access to:
- Time series from one site
- Basic met data [pressure, temp, RH]

I could get access to:
- Boundary layer height
- NO2, O3,..
- land-use

Coming up with features is difficult, time-consuming, requires expert knowledge. ***"Applied machine learning" is basically feature engineering****. — Andrew Ng, Stanford University*

Using Interpretable algorithms

Is the algorithm extracting relationships that make sense?

Bryan N. Vu et al. **Developing an Advanced PM2.5 Exposure Model in Lima**, Peru. Remote Sens. 2019, 11(6), 641; https://doi.org/10.3390/rs11060641

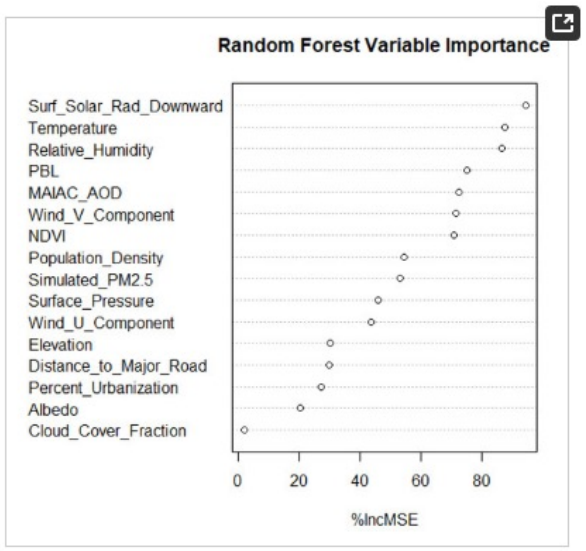


Figure 4. Importance of each variable in the random forest model by percent increase mean square prediction error (MSE).

Lamb, K. D.: **Classification of iron oxide aerosols by a single particle soot photometer using supervised machine learning**, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2019-106, in review, 2019.

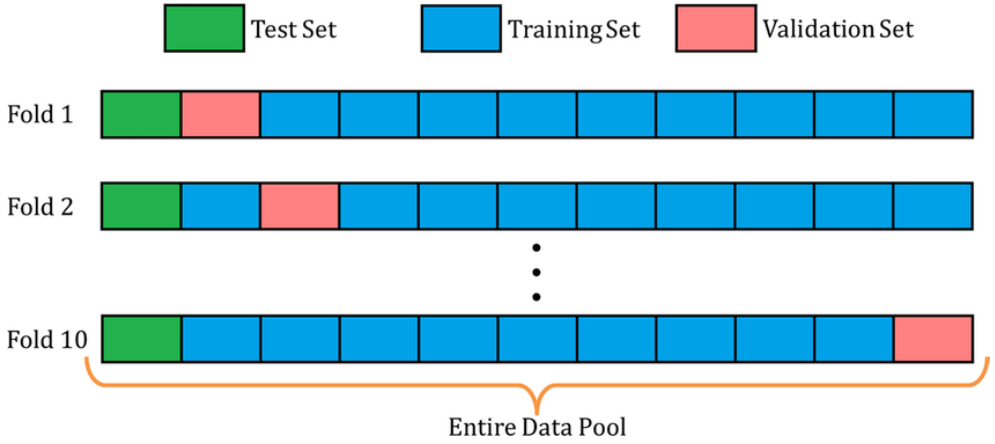
Table 3. Importance of different features. The relative importance of the different features for the optimal random forest for the 6-class and 3-class cases. All denotes that all 17 features were used in the algorithm, and reduced indicates that only a subset of features were included as input to the learning algorithm (11 features for the 6 class case and 9 features for the 3 class case). The top 5 most important features in each category are bolded.

Feature	All (6 classes)	Reduced (6 classes)	All (3 classes)	Reduced (3 classes)
Blue peak amplitude	**0.064**	**0.075**	**0.049**	**0.057**
Color ratio	**0.391**	**0.358**	**0.455**	**0.587**
Core scattering	0.032	0.044	**0.051**	0.019
Total scattering max.	0.021	•	0.014	•
Post incandescent scattering	**0.148**	**0.190**	**0.183**	**0.136**
Evaporation scattering size	**0.042**	**0.053**	**0.043**	0.037
Position sensitive wideness	0.018	•	0.011	•
Min. scattering before incandescence	0.023	•	0.020	•
Position sensitive trigger position	0.024	•	0.017	•
Scatter peak location	0.026	0.045	0.019	•
Saturation width	0.028	0.046	0.015	•
Incandescent start position	0.039	**0.054**	0.029	**0.057**
Evaporation point	0.035	0.052	0.025	**0.054**
Incandescent total length	0.030	0.036	0.023	0.028
Incandescent used length	**0.044**	0.047	0.026	0.025
Light on laser intensity	0.018	•	0.011	•
Width fraction from center	0.016	•	0.011	•

The easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models..[…]. – Christoph Molnar, https://christophm.github.io/interpretable-ml-book/

Fitting – testing
How do I stop my model overfitting?

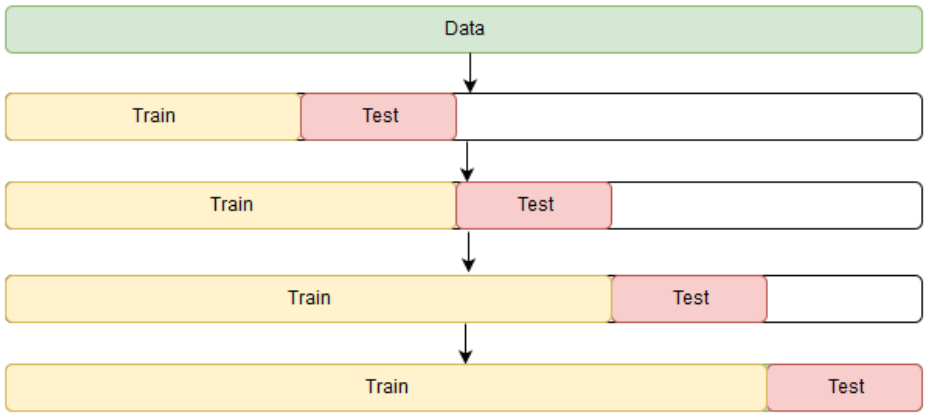
I want to evaluate how good my model is on data it hasn't seen…



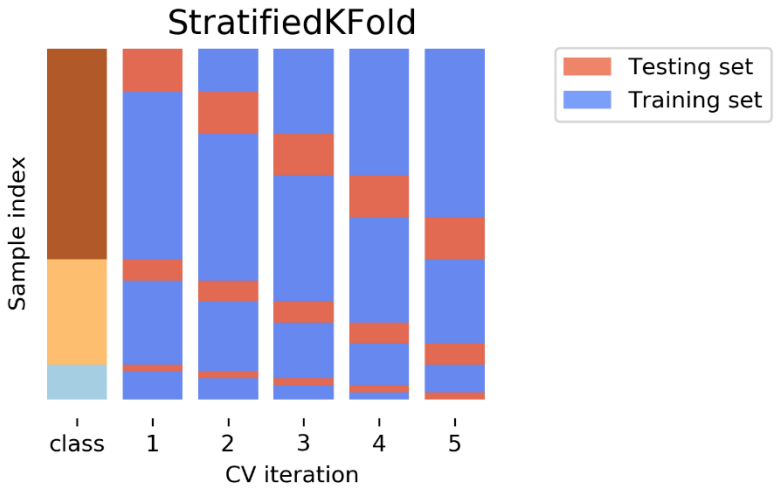
There are functions in popular ML package's that enable the user to specify

- How many folds
- test:train:validation split
- level of stratification

I want to predict ahead of time…

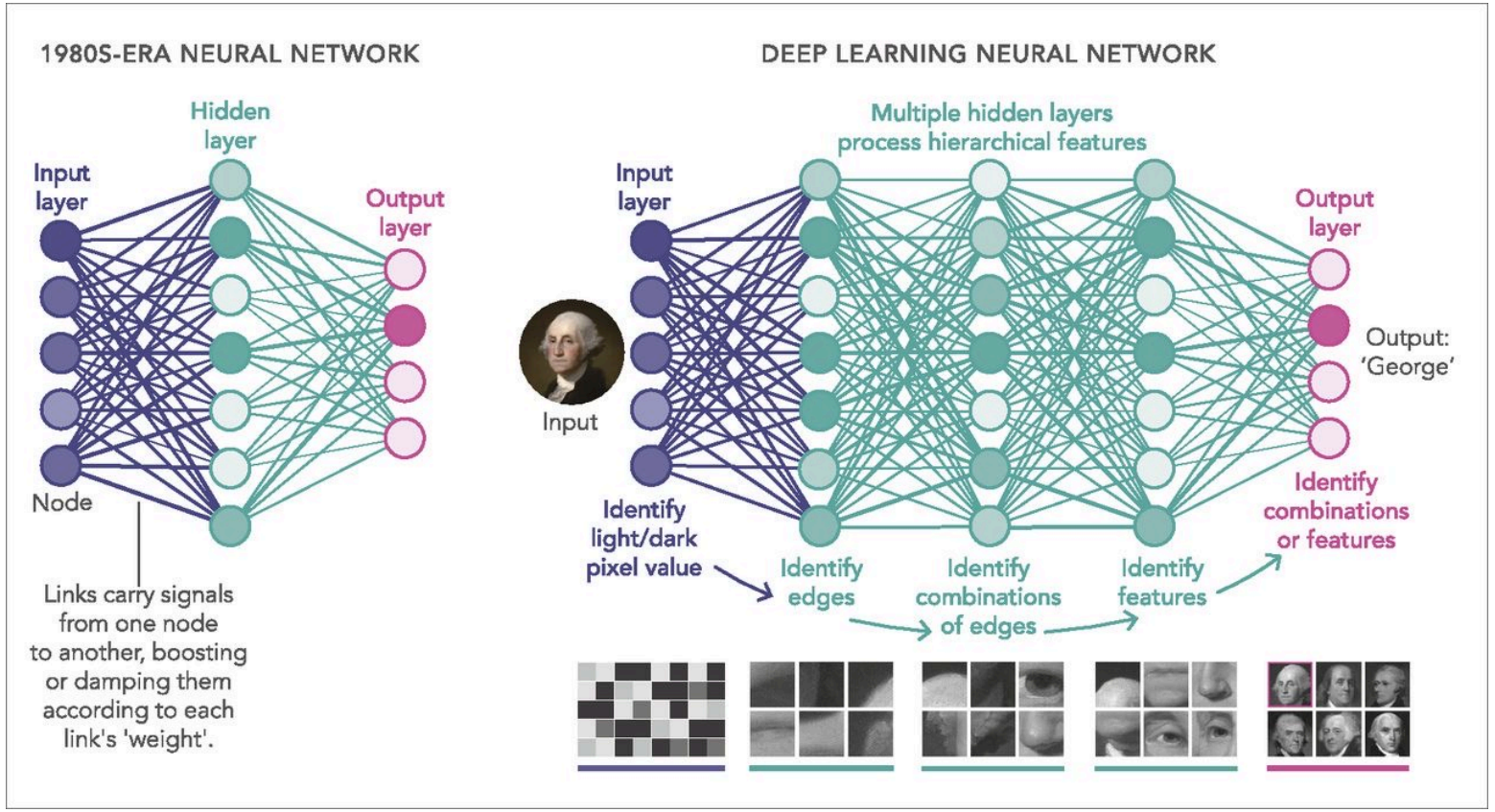


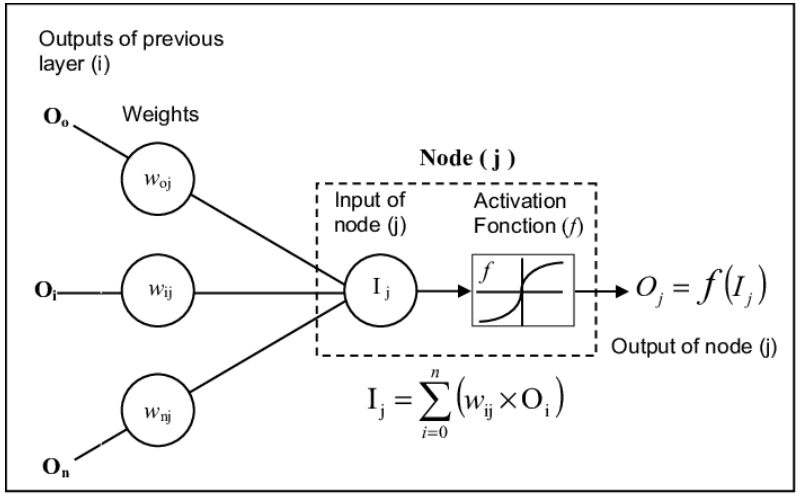
My data has imbalanced classes……



Neural networks

name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.





Activation Functions

Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$

tanh
$\tanh(x)$

ReLU
$\max(0, x)$

Leaky ReLU
$\max(0.1x, x)$

Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU
$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$

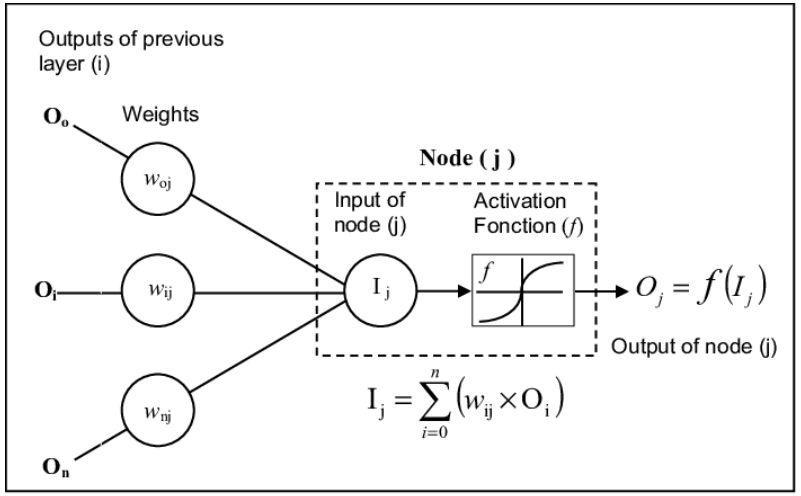
Neural networks

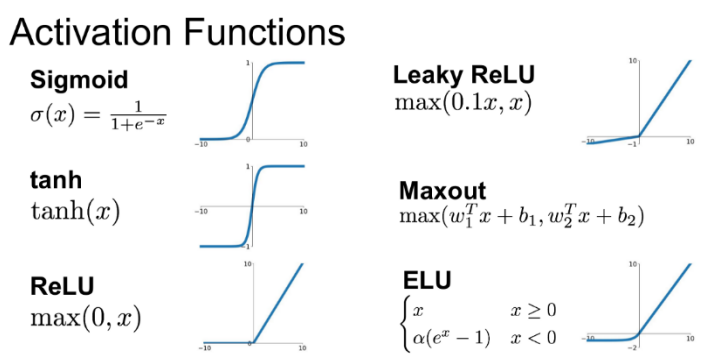
As you build your network you need to think/ask

How many layers do I need?
How many nodes?
Are all nodes connected?
What is the most appropriate activation function?
How do I define performance?
- there are multiple loss functions [mean squared error, binary cross entropy…]

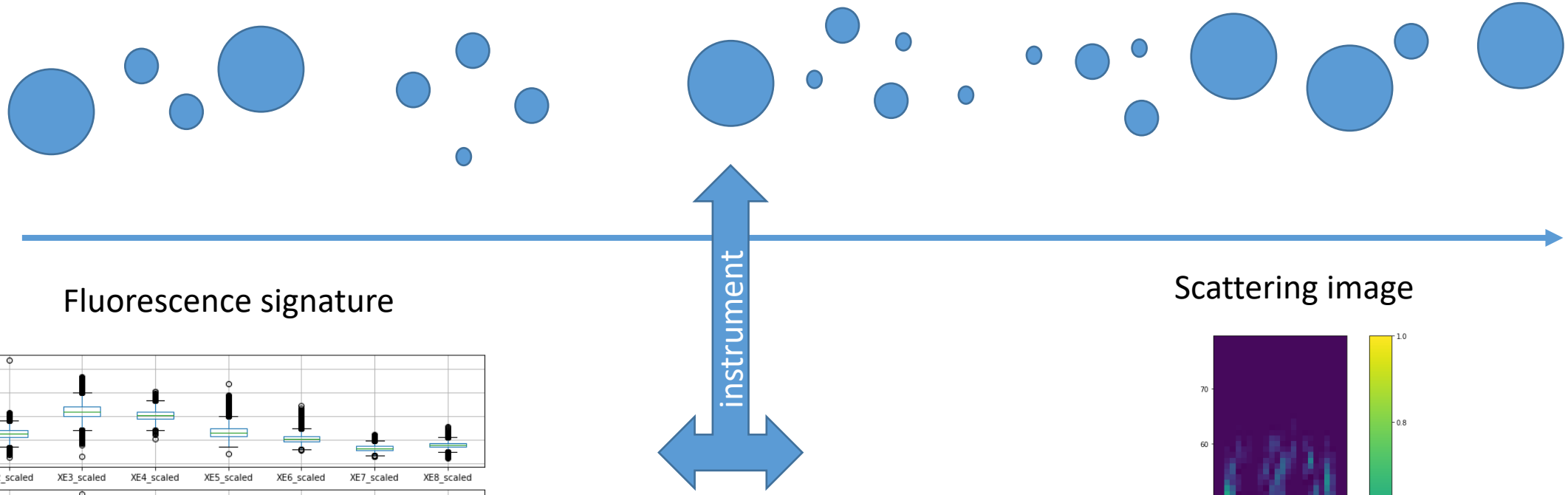
What happens when I modify my input data? Does that improve model performance?

There are ways to do all of the above. Best way is to start 'something'
- I will provide code on a toy example

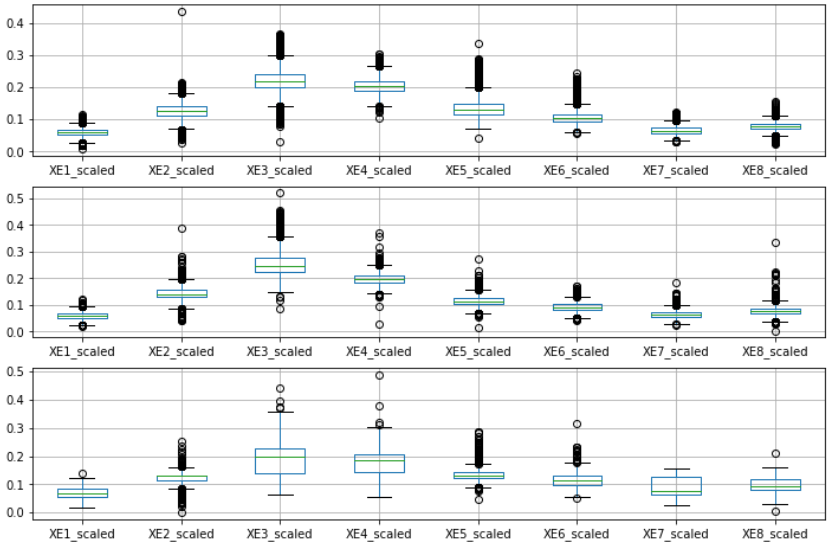




Todays exercises



Fluorescence signature



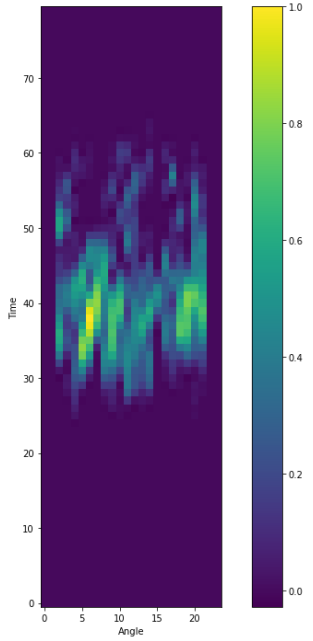
How many different types of particles do we detect?

What are they?

Are they important?

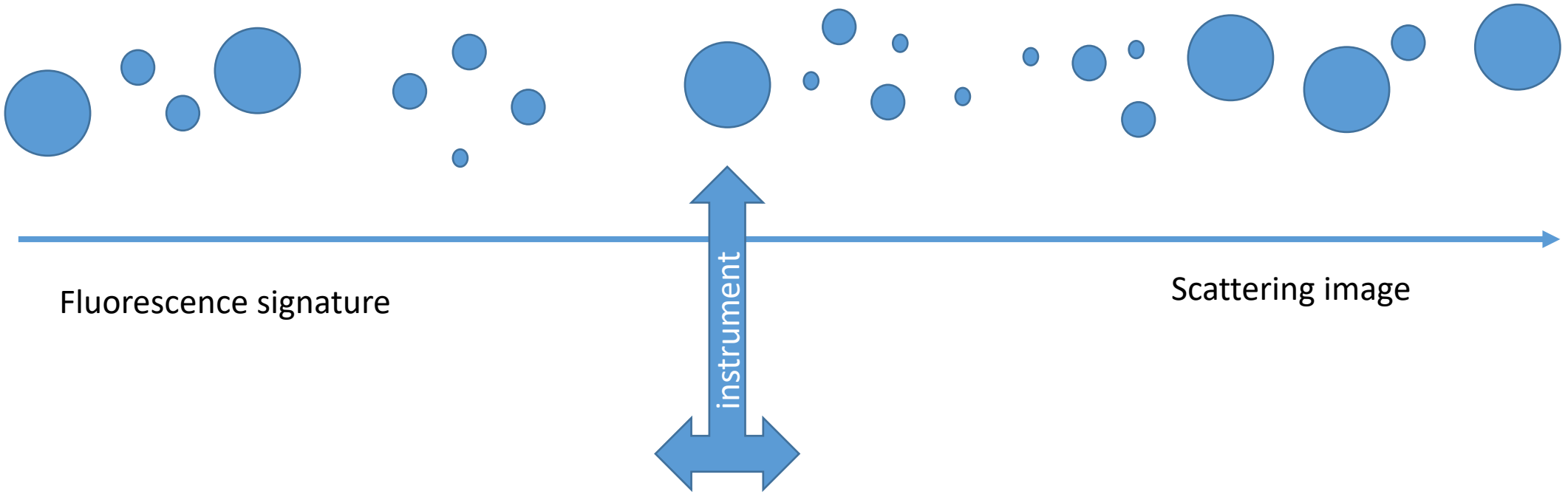
Why do the concentrations change?

Scattering image

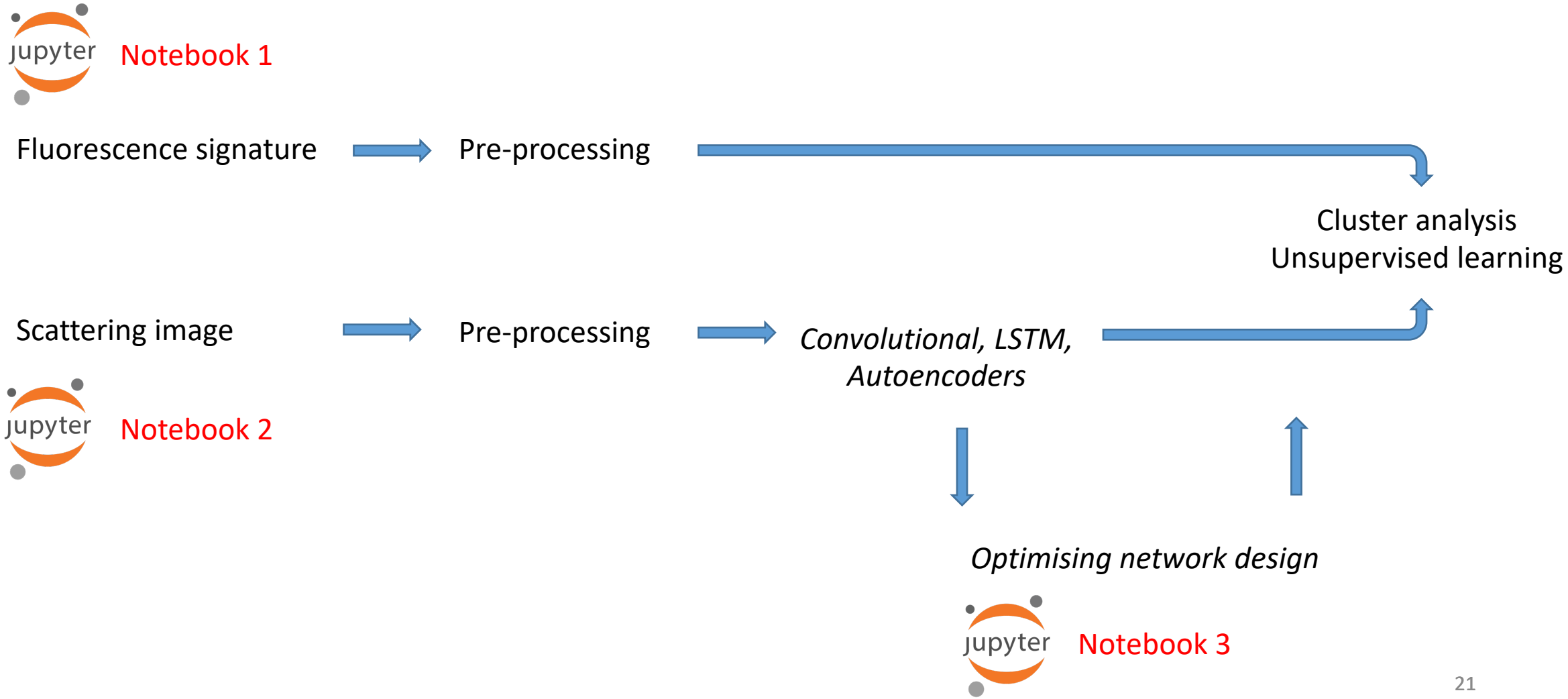


MANCHESTER 1824
The University of Manchester

Todays exercises



Our workflow….



Cluster Analysis

Group ‘similar’ members.

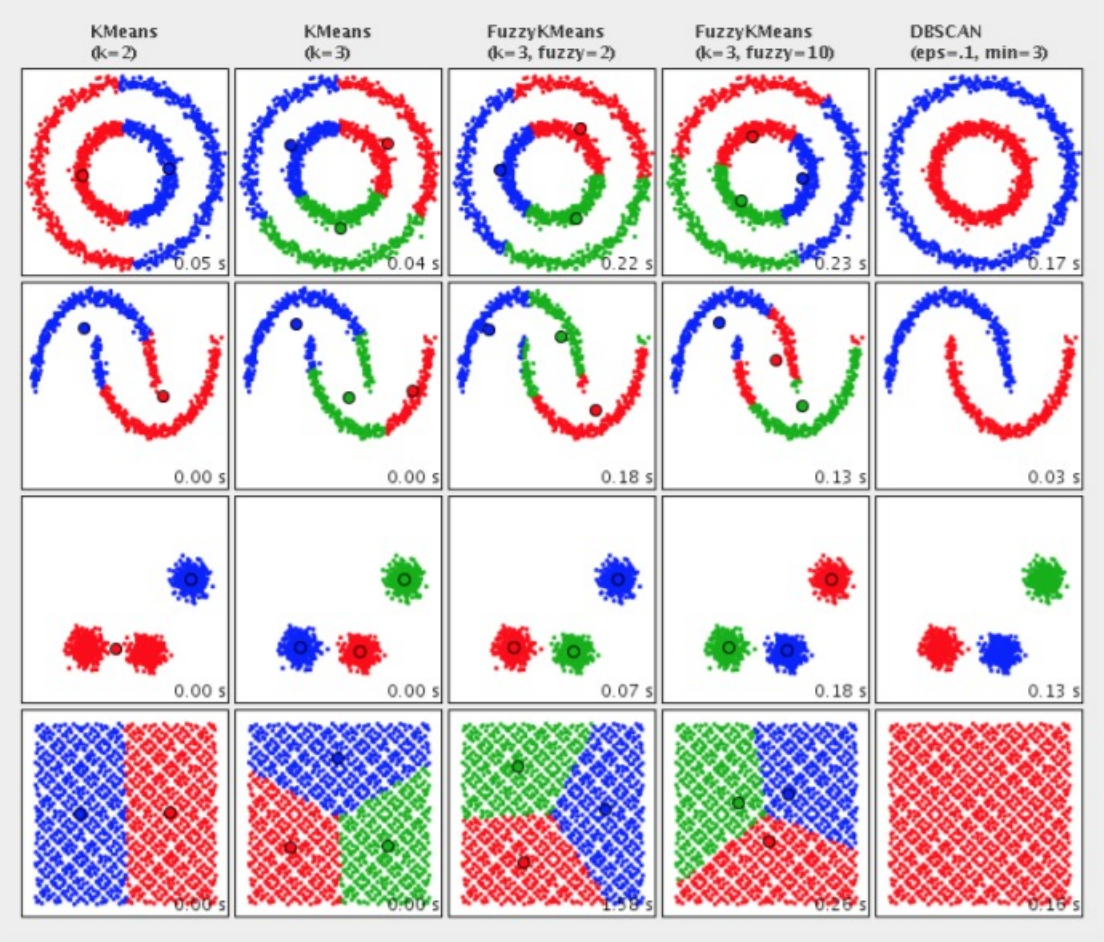
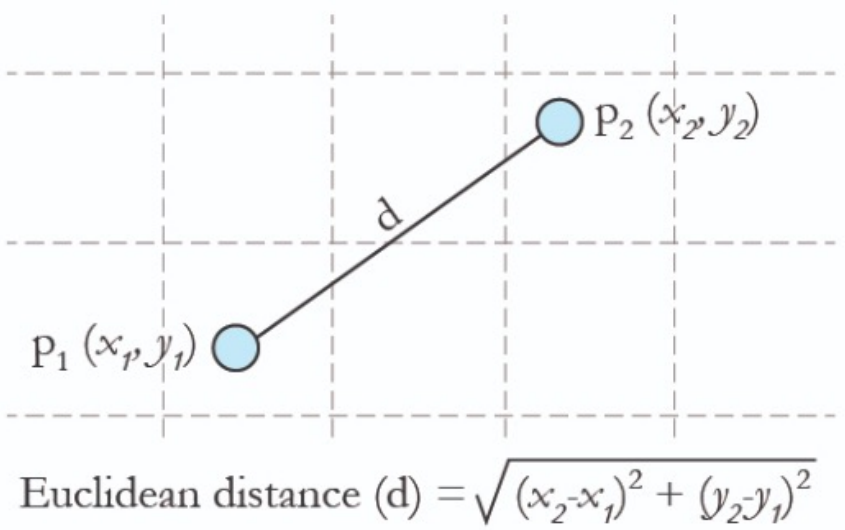


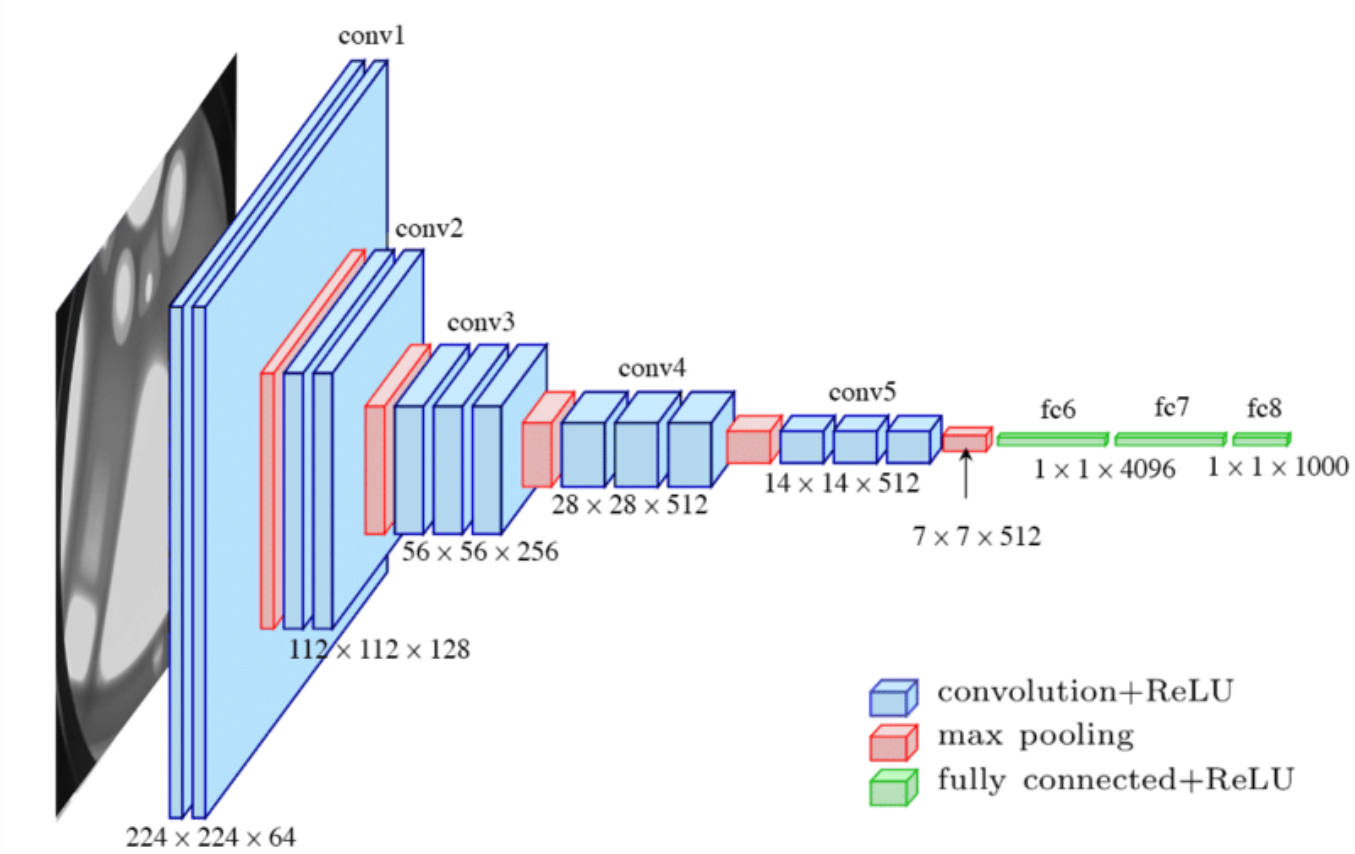
Image source: http://commons.apache.org/proper/commons-math/userguide/ml.html

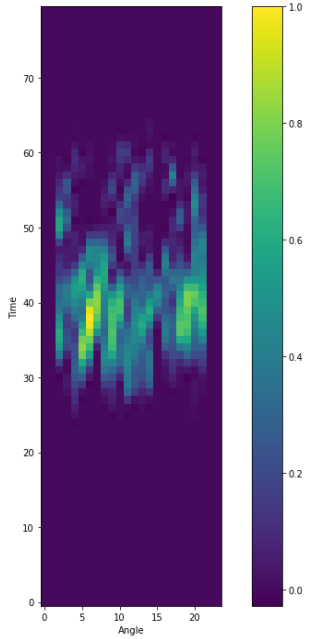
How do we set the number of groups?
How do we define a similarity ‘distance’?



MANCHESTER 1824
The University of Manchester

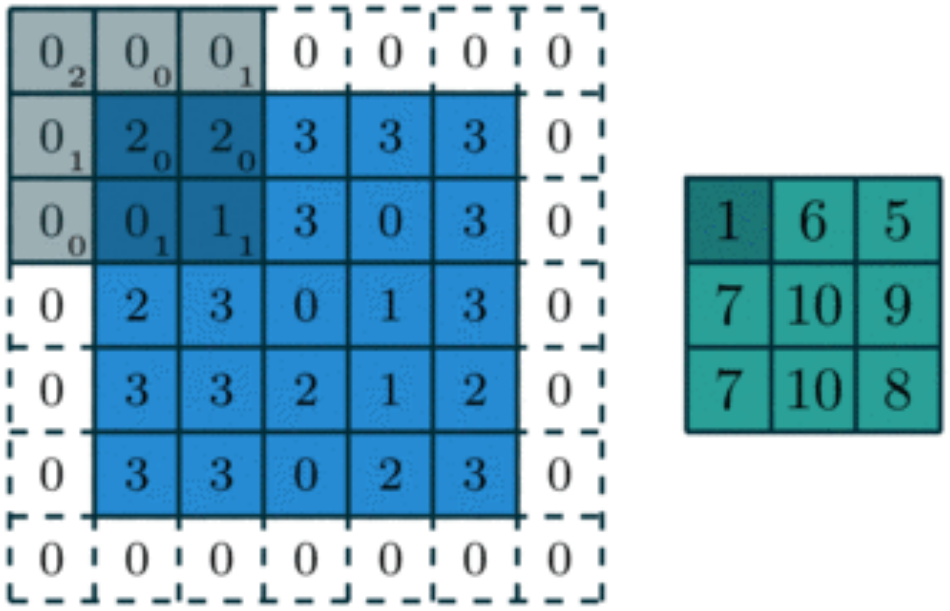
Convolutional Neural Networks [CNNs]

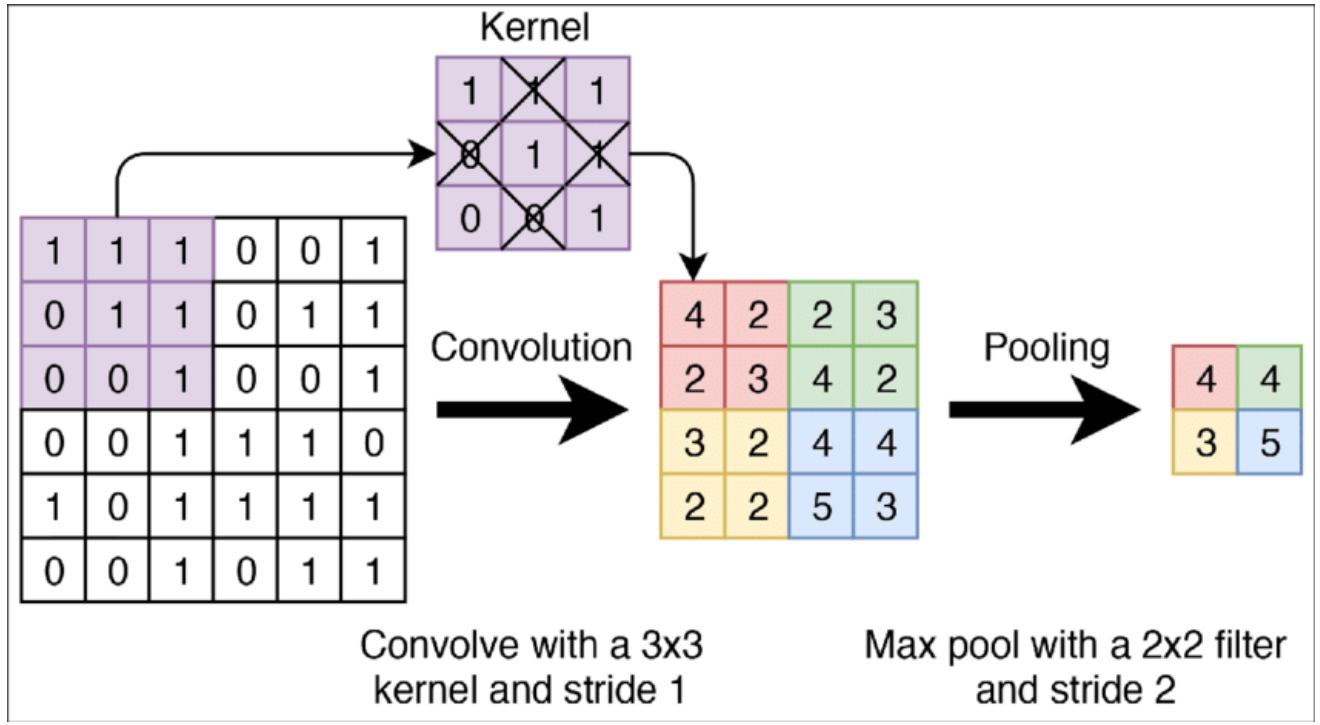




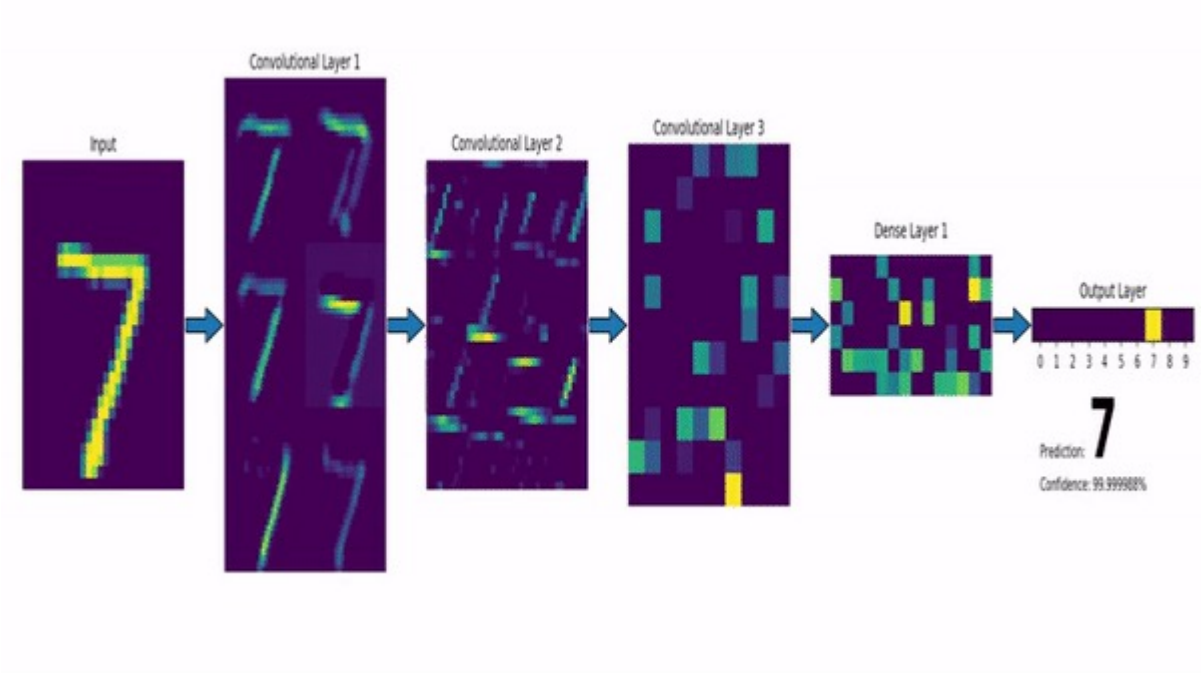
The standard VGG-16 network architecture

convolutional layer is a stack of feature maps where we have one feature map for each filter.
large datasets with many different object categories will require a large number of filters; each responsible for finding a pattern in the image.
More filters mean a bigger stack which means that the dimensionality of our convolutional layers can get quite large.

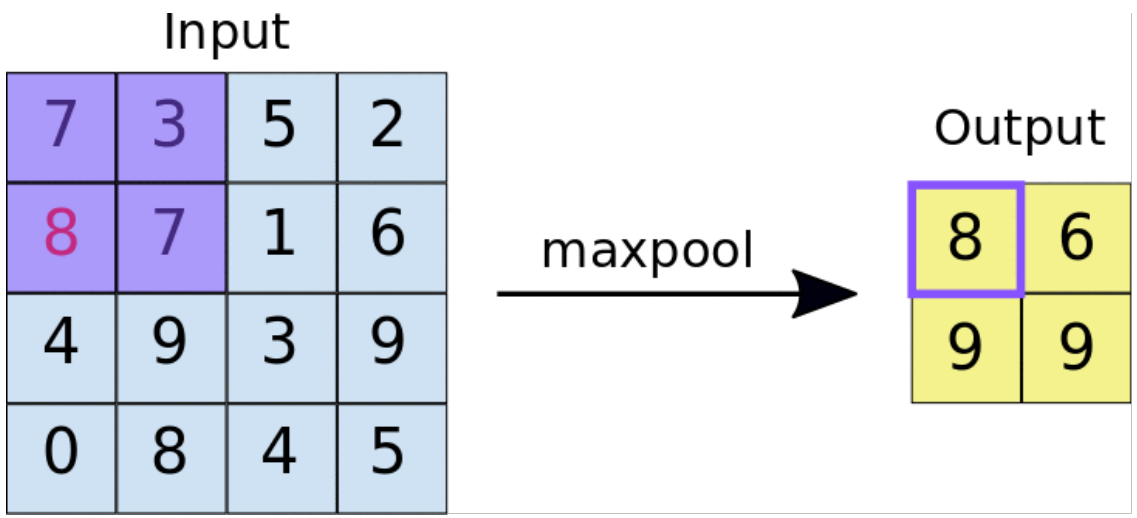


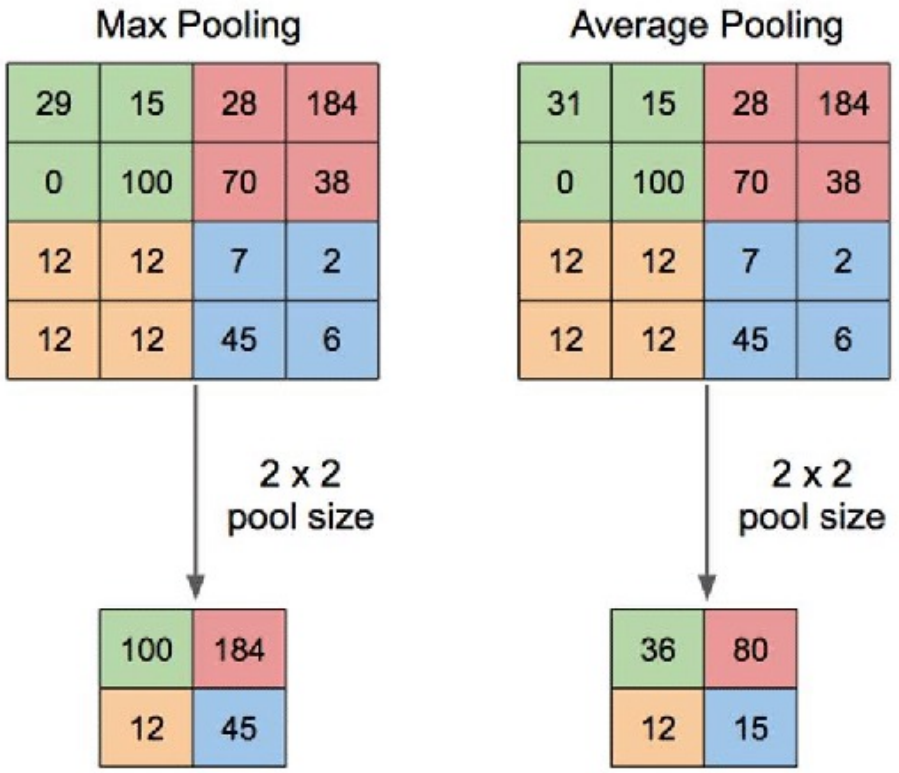


1	6	5
7	10	9
7	10	8



Higher dimensionality means we'll need to use more parameters which can lead to overfitting
We need a method for reducing this dimensionality. This is the role of pooling layers.



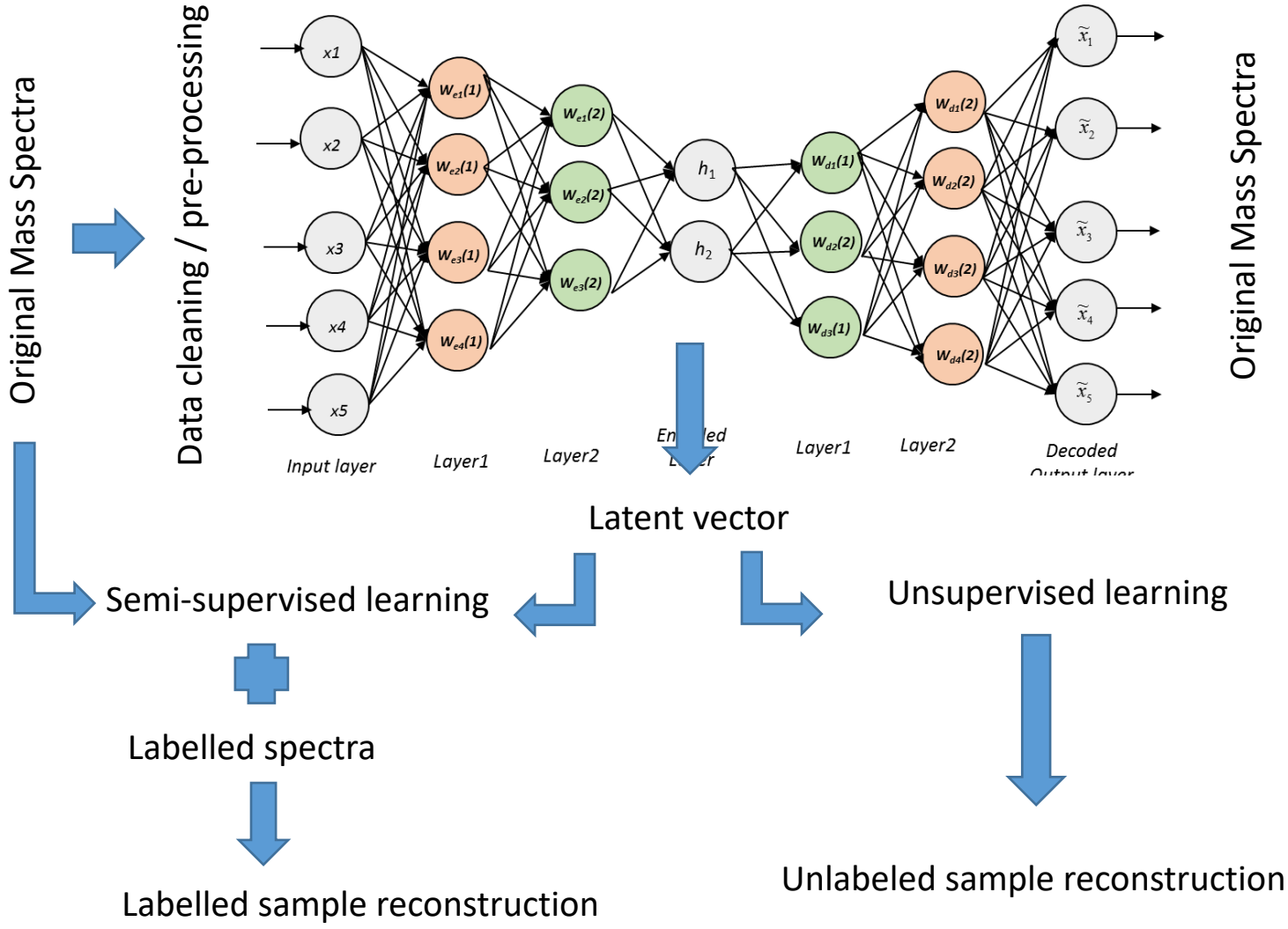


Autoencoders – generating a space to cluster

High number of dimensions problem for standard clustering techniques

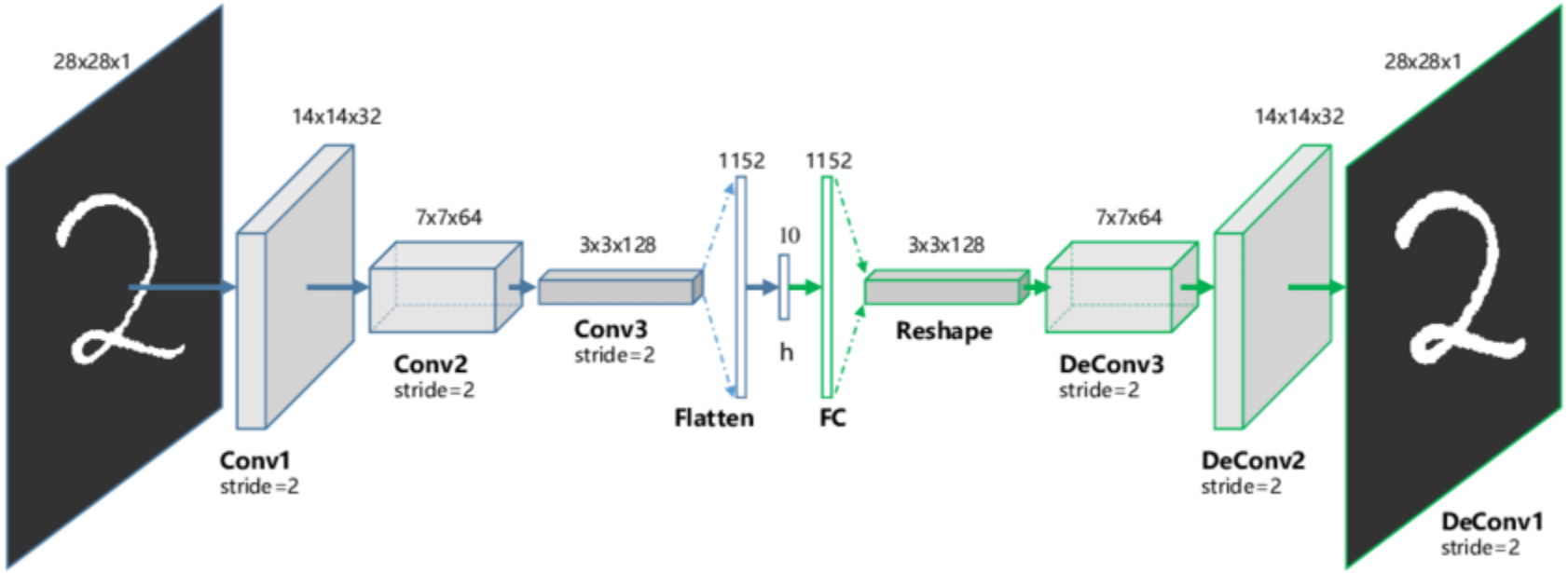
1) Dimension reduction techniques + standard clustering shown to be more stable

- What are distinct clusters?
- What about outliers?



MANCHESTER 1824
The University of Manchester

CNNs - Autoencoders

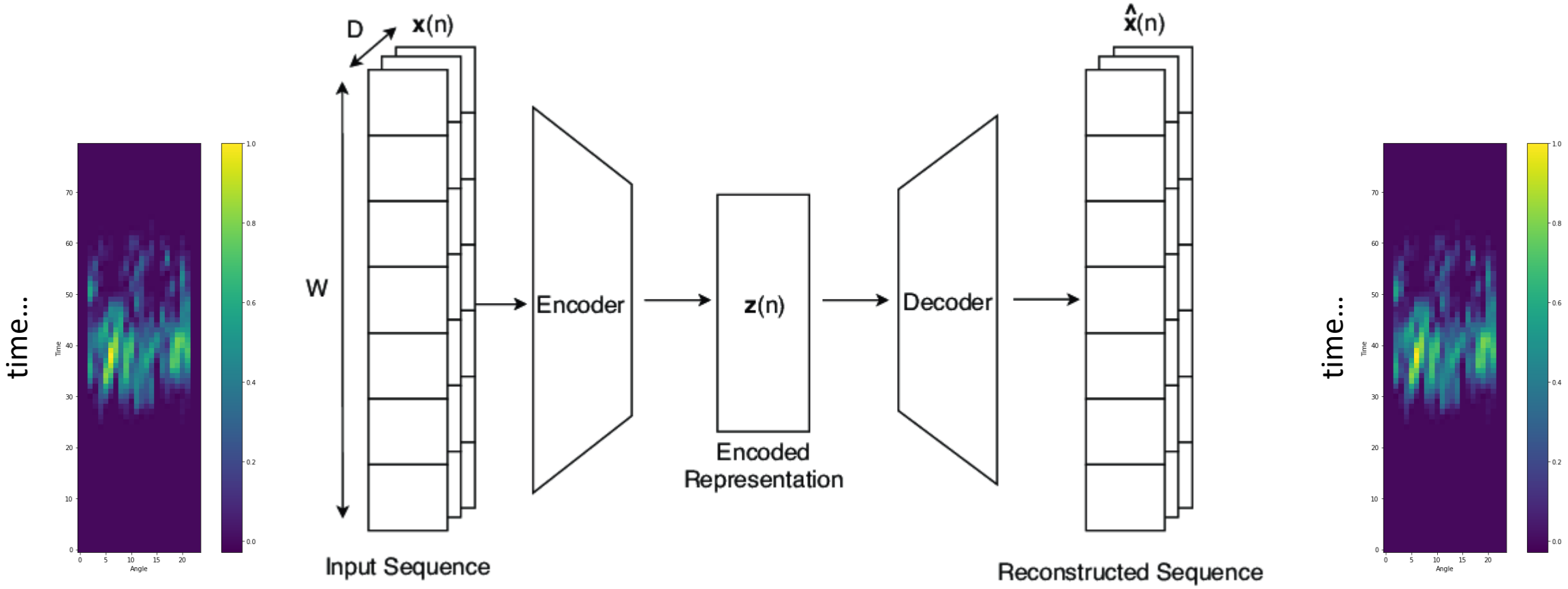


What does the latent space represent??

Id like us to look at this in Python code

MANCHESTER 1824
The University of Manchester

LSTM - Autoencoders

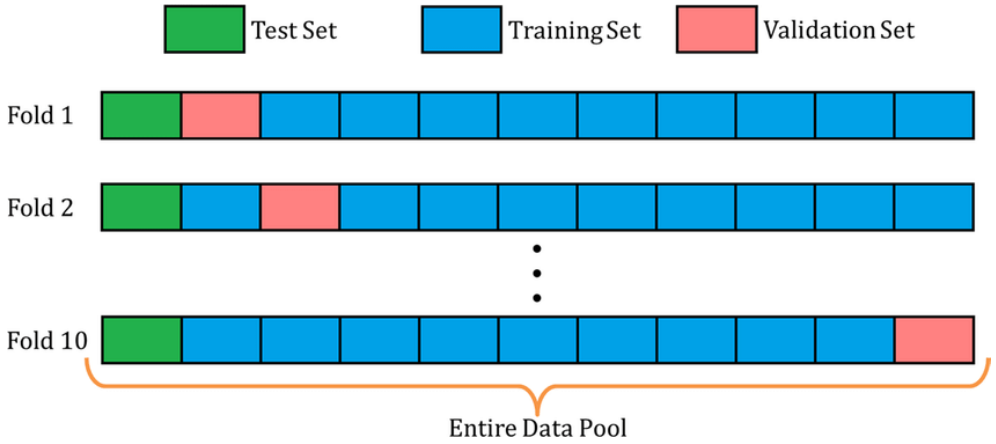


What does the latent space represent??

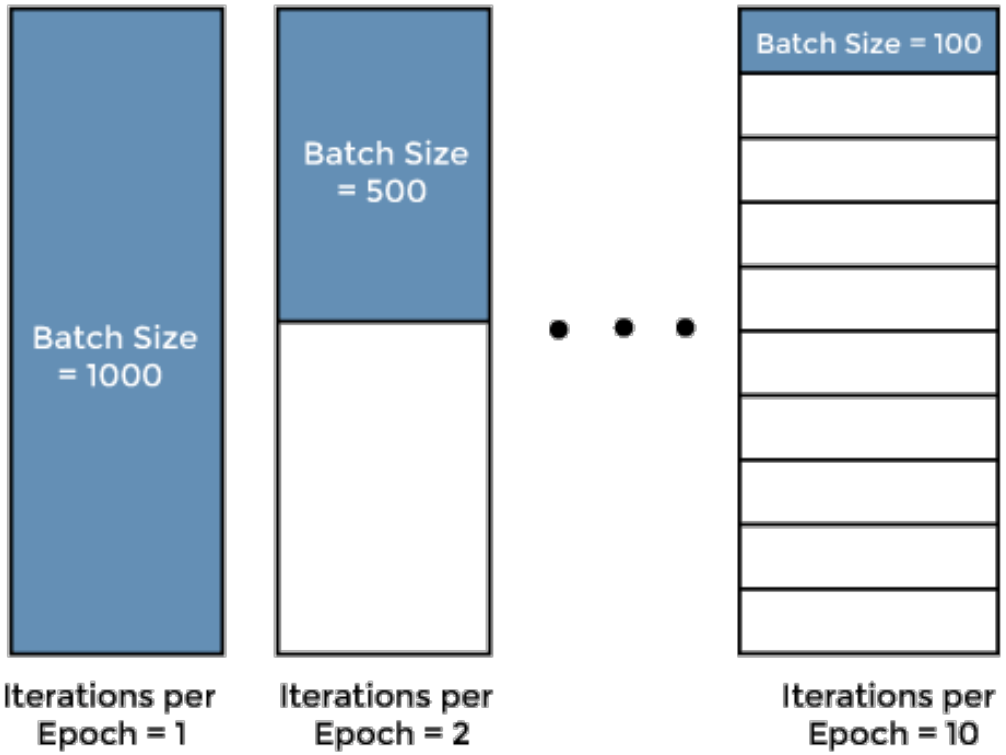
Id like us to look at this in Python code

Fitting – testing
How do we fit an entire neural network

I want to evaluate how good my model is on data it hasn't seen…



Epoch – one sweep over the entire dataset
Batch size – how many samples in a ubset
Iterations – number of times weights are updated



MANCHESTER 1824
The University of Manchester

What next?

Build your tools and software stack [Python]

Data

Pandas: https://pandas.pydata.org
Scikit learn: https://scikit-learn.org/stable/

Dask: https://www.dask.org

Machine learning

Scikit learn:https://scikit-learn.org/stable/

Keras:https://keras.io
PyTorch:https://pytorch.org

Digital CV

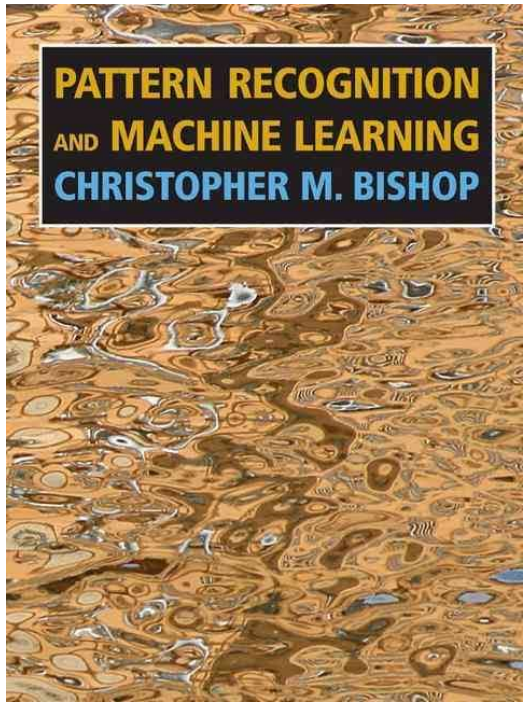
Github: https://github.com
Zenodo:https://zenodo.org

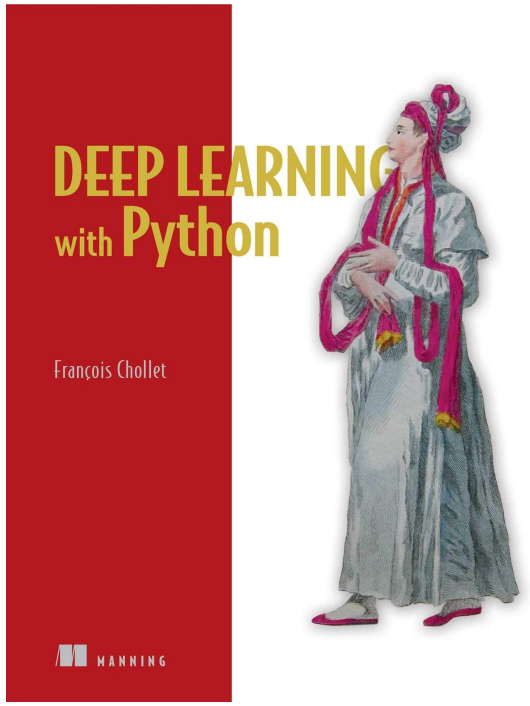
Cloud services

Google Colab: https://colab.research.google.com
Binder: https://binderhub.readthedocs.io/en/latest/

MANCHESTER 1824
The University of Manchester

Embrace the culture of AI





Machine Learning Glossay
https://developers.google.com/machine-learning/glossary

What does the future hold?

Do ‘pure’ ML representations have a limited shelf-life?

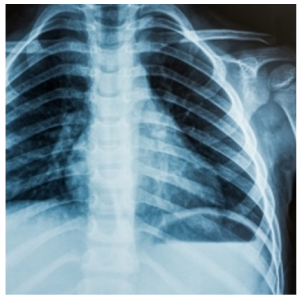
e.g. Pharmacokinetics/Pharmacodynamics

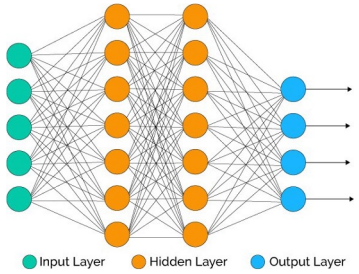
Patient Data
(e.g. X-ray scans)

Simulation parameters
(e.g. drug half-life)

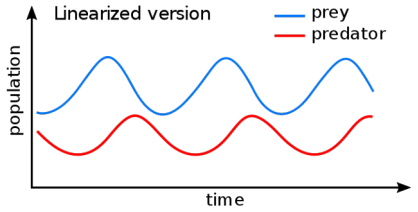
Treatment Outcome

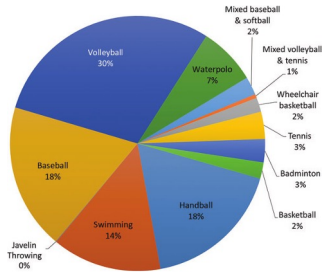
Neural Network





$\theta = 0.56$





Backpropagation

Slide c.o Mike Innes, Julia Computing
[mike@juliacomputing.com]

Zygote

The next step will be a hybrid modelling approach, coupling physical process models with the versatility of data-driven machine learning. ***Deep learning and process understanding for data-driven Earth system science.*** *Markus Reichstein et al. Nature 566, pages 195–204 (2019)*

Left with important questions around community ethos

Do we value the ‘slow and hard’ stuff as we should?

- Databases, calibration standards and meta-data.
- Renewed vigor in combining mechanistic and empirical model developments.

Training is key. Heterogenous issue. More than *.fit() .predict()*. Can we publish negative results? What platforms will we be using in 5+ years?

Is there a role of ‘big tech companies’?

Open source a great step – documentation needs to follow!!!

We might benefit from a community repository of relevant examples with relevant data

There are challenges, but there are many exciting opportunities!