

Introduction to machine learning

CASA0006: Data Science for Spatial Systems

Huanfa Chen

- 1 Introduction to Module
- 2 Supervised Machine Learning
- 3 Tree-based Methods
- 4 Artificial Neural Networks
- 5 Analysis Workflow
- 6 Panel Regression
- 7 Difference in Difference
- 8 Regression Discontinuity
- 9 Dimensionality Reduction
- 10 Spatial Clustering

Objectives

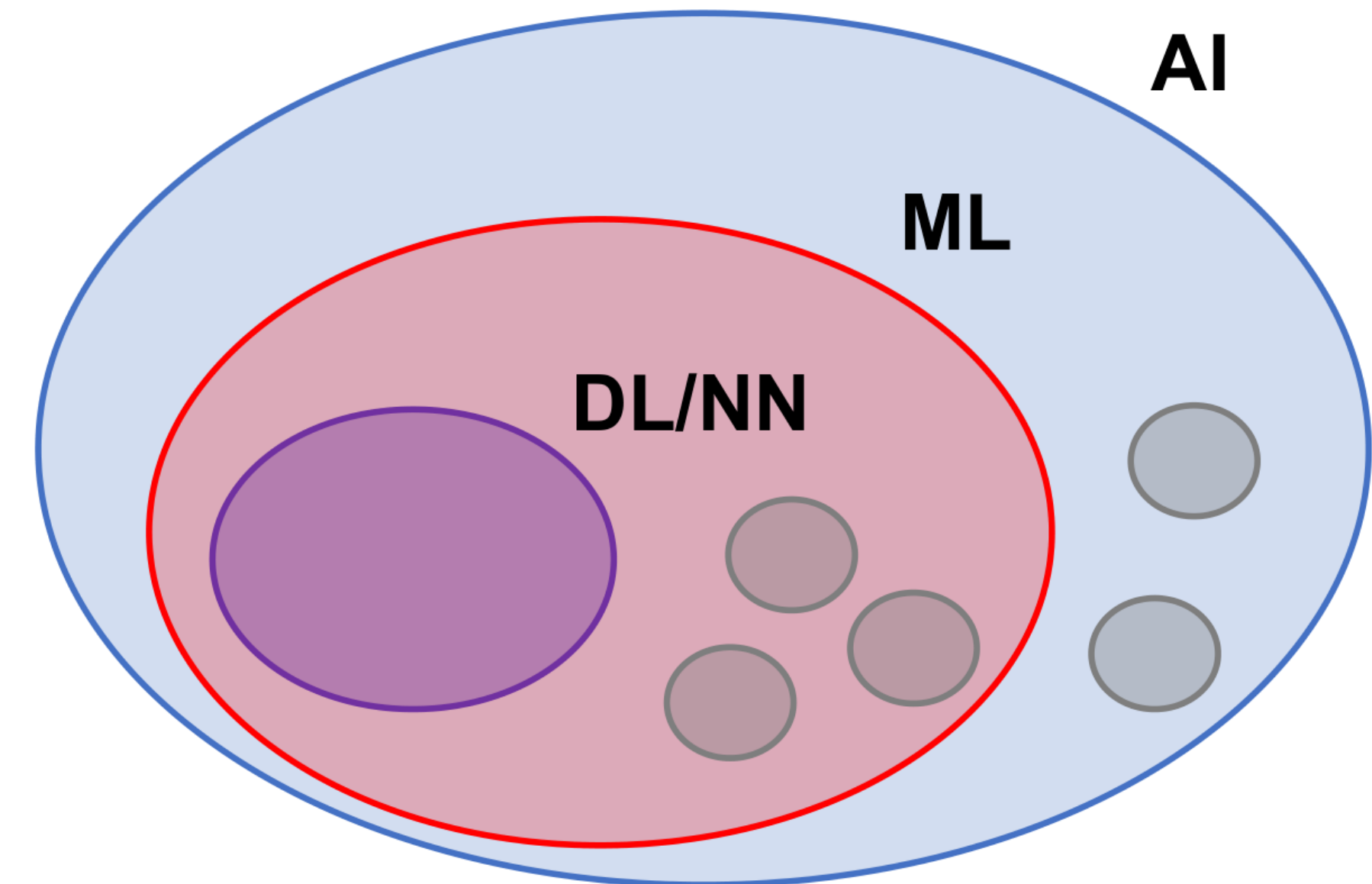
- Learn the basics and classification of machine learning
- Understand the differences between statistical methods and machine learning
- Understand several important theorems in machine learning and data science



Introduction to ML

ML is a subset of AI

- Machine learning (decision tree, random forest, k-means, etc.)
- Deep learning (deep neural networks)
- Others AI tools: graphical models, symbolic AI
- In this module, we don't distinguish ML/DL and consider NN as part of ML.



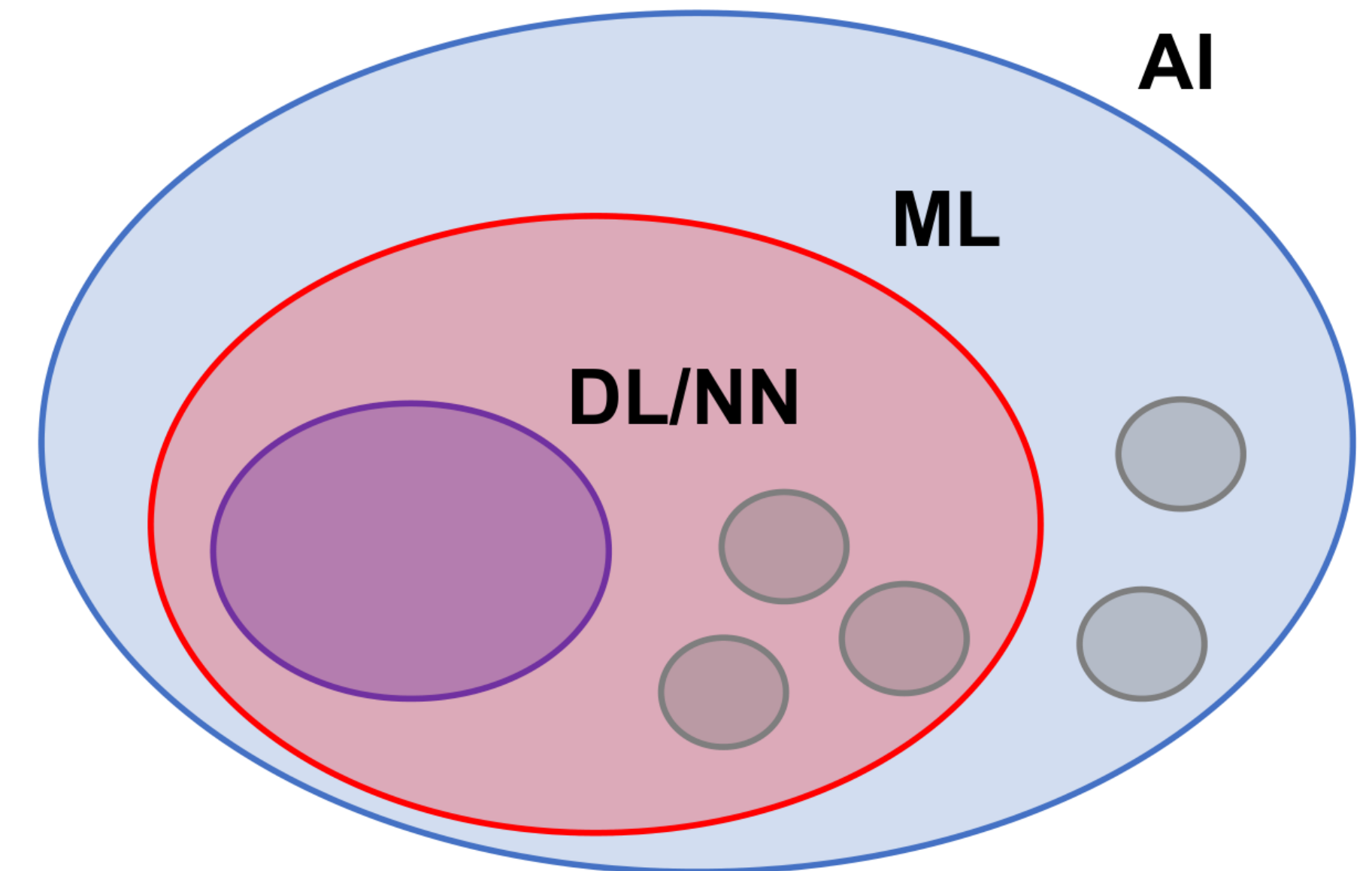
Definition of ML

Arthur Samuel (1959)

- (Machine learning is the) field of study that gives computers the ability to learn without being explicitly programmed.

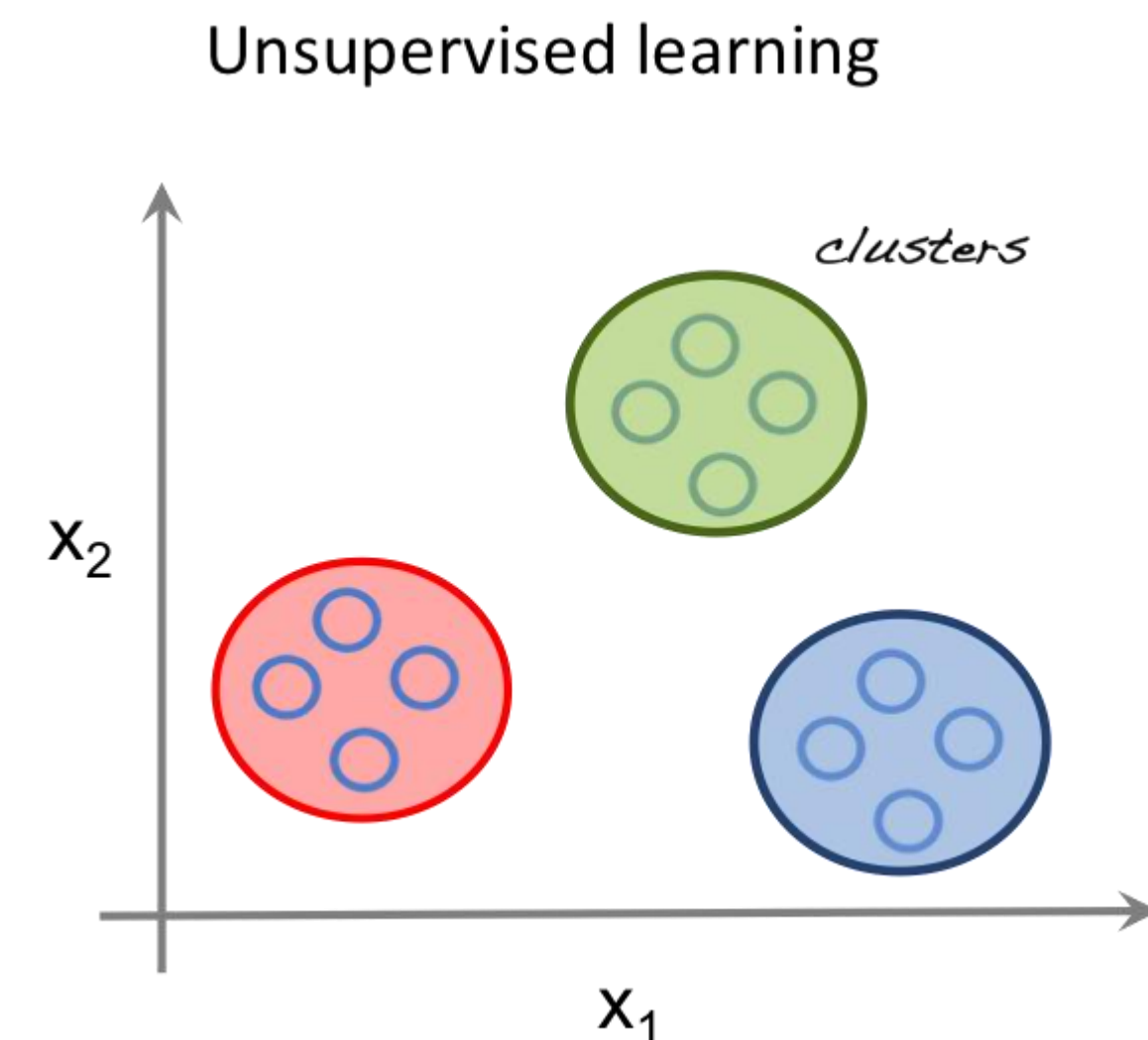
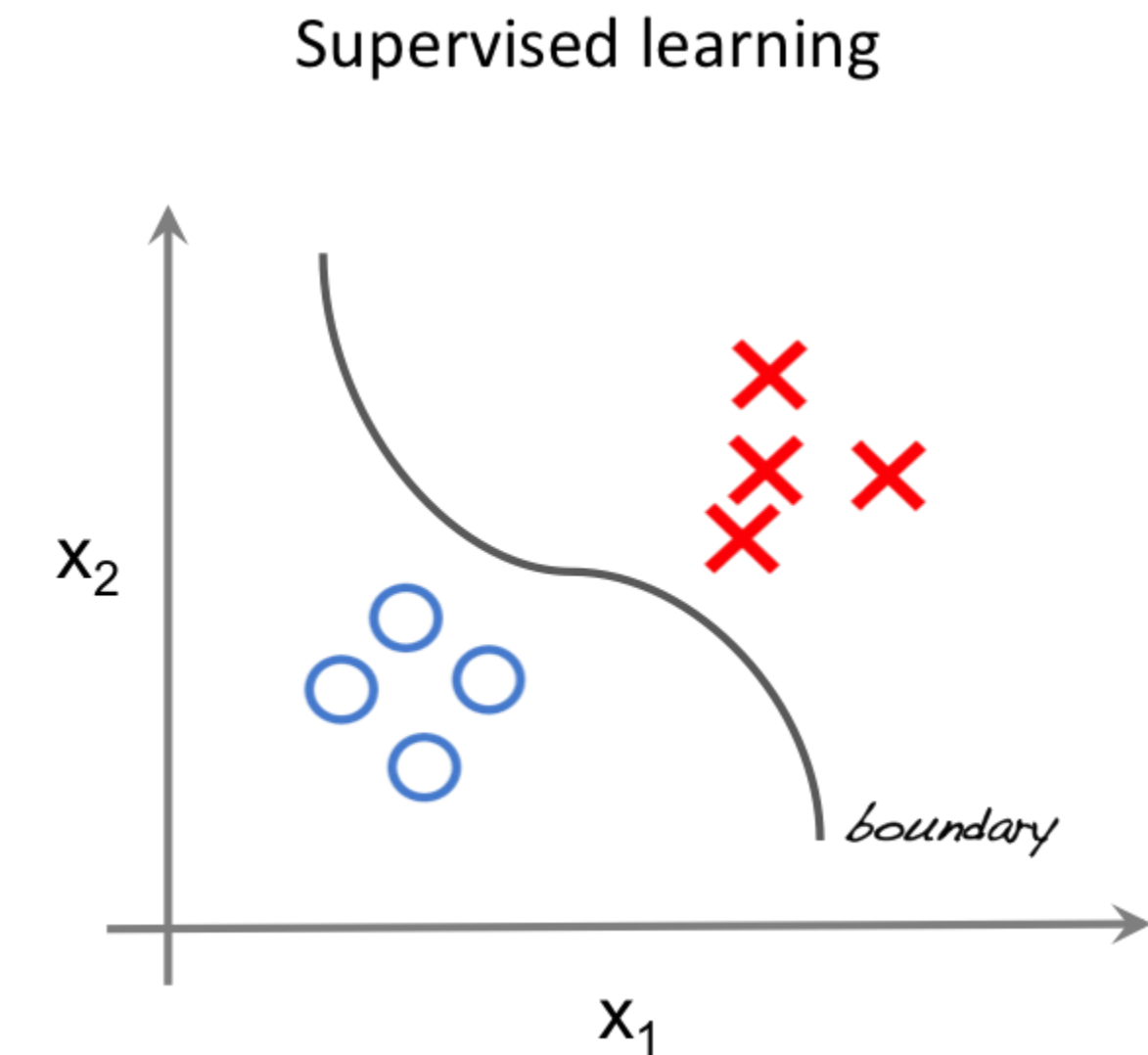
Tom Mitchell (1997)

- A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .

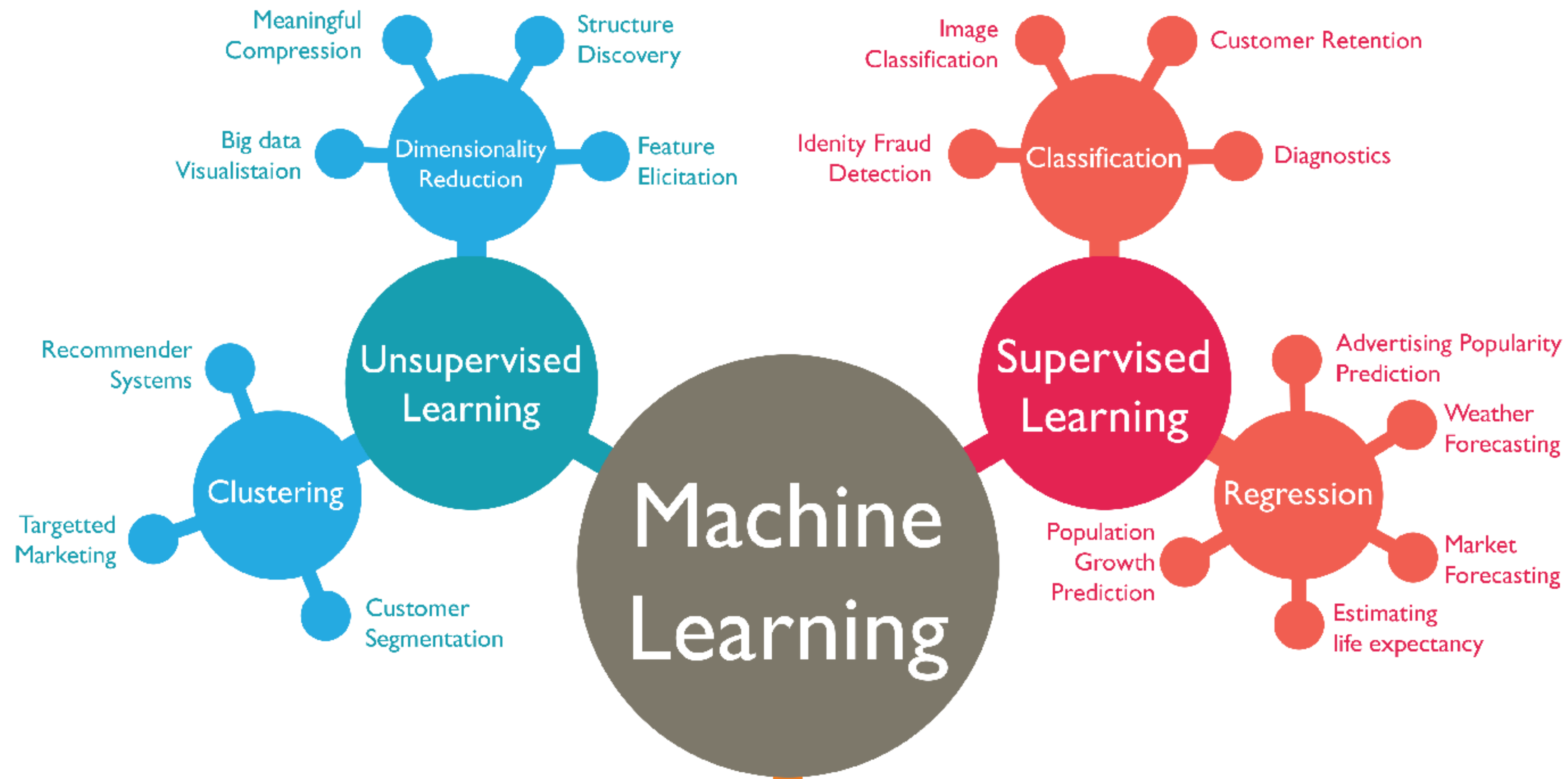


Three types of ML

- Supervised: Learn to predict output given input (with labelled data).
- Unsupervised: Discover internal representation/structure of input (without labelled data).
- Reinforcement: Learn actions to maximise payoff (via interactions with the environment).



Three types of ML



Go to **www.menti.com** and use the code ~~2515-0560~~

Putting linear regression into the framework of ML. Which type of ML will linear regression fall into?

☐

Supervised

☐

Unsupervised

☐

Reinforcement

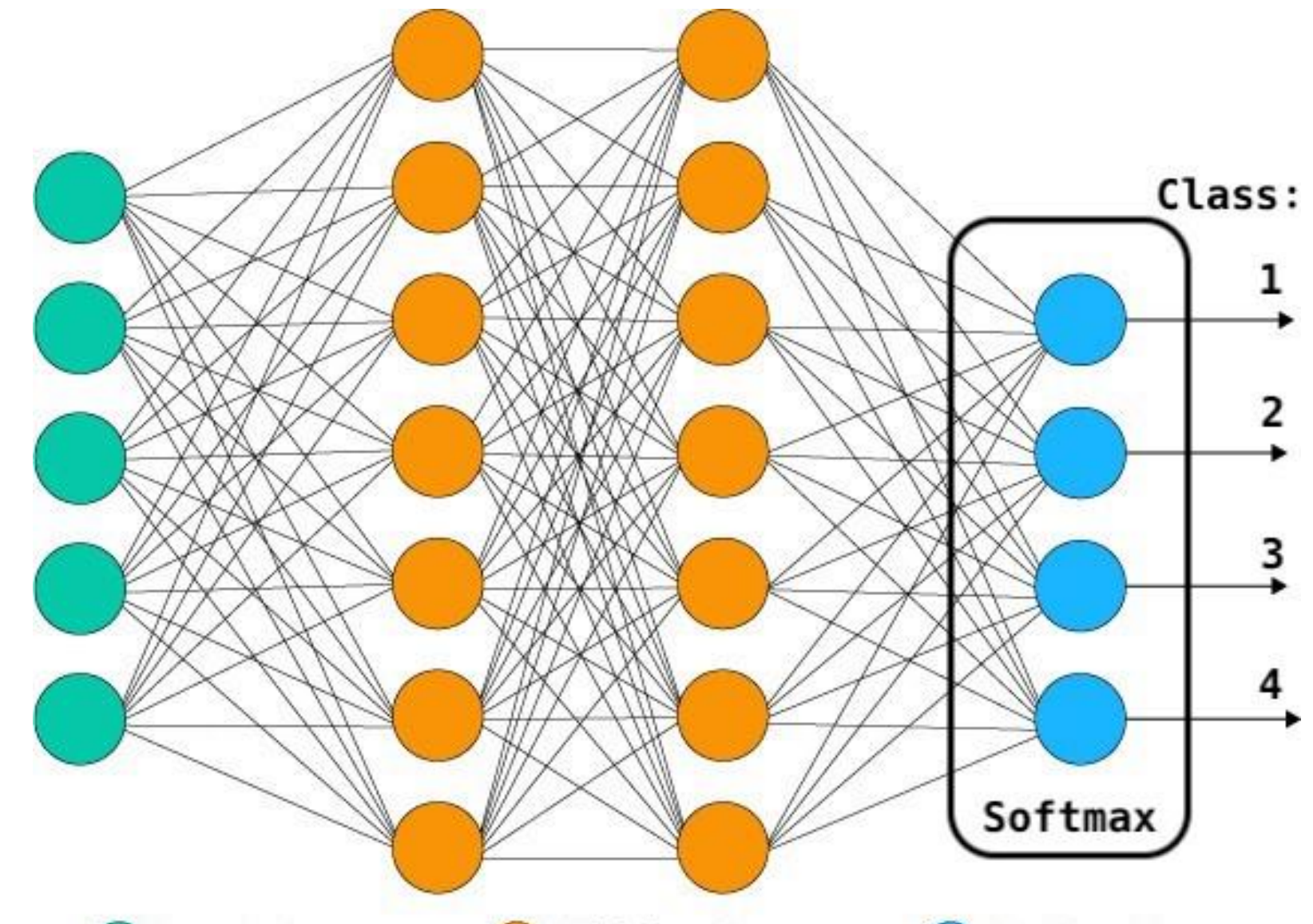
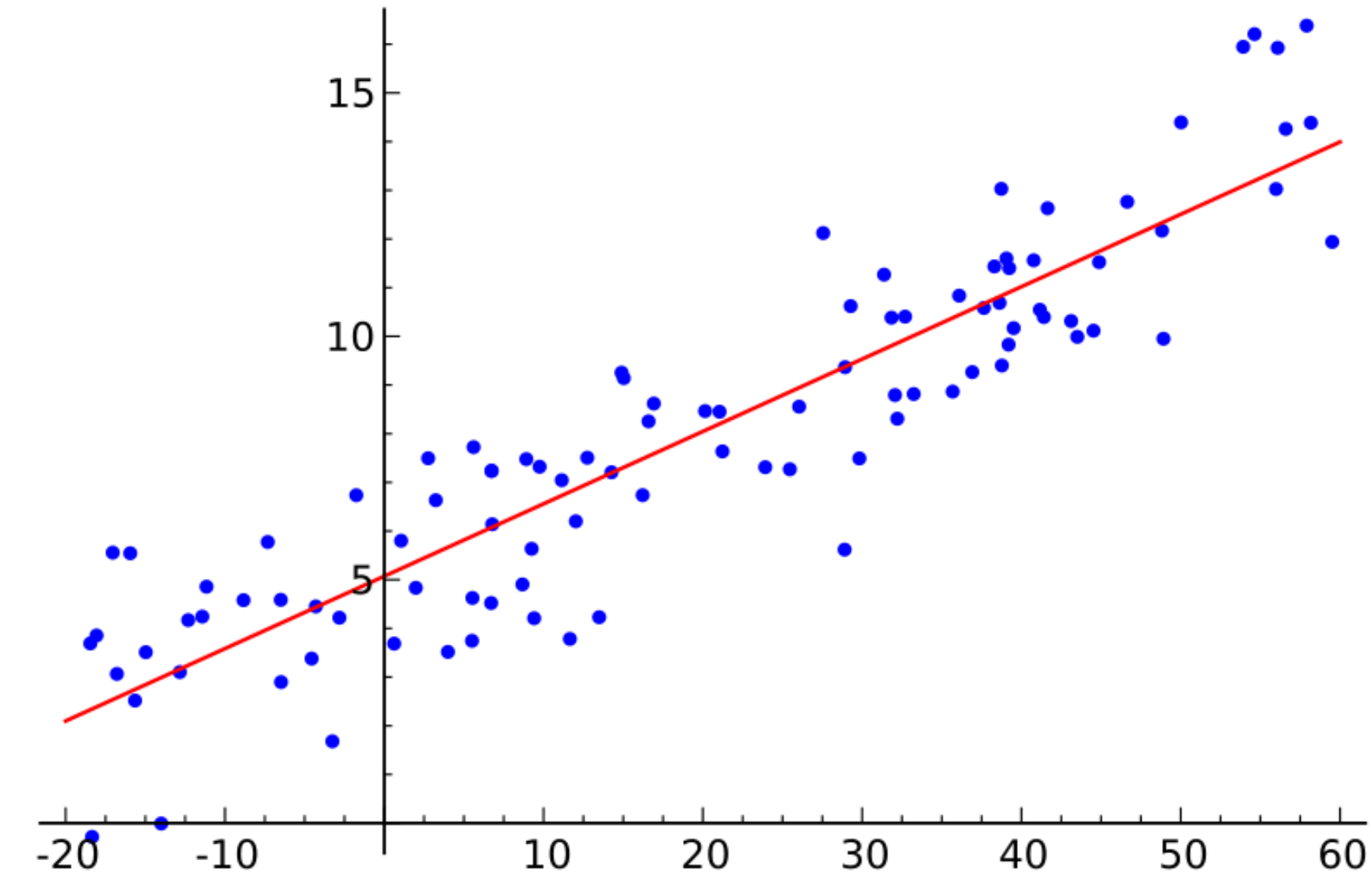
Submit

An aerial photograph of an airport tarmac and runways, overlaid with a semi-transparent dark grey layer. The image shows several aircraft parked at gates or on the tarmac, and multiple runways and taxiways. The text "Statistics versus ML" is centered in white.

Statistics versus ML

More specific questions

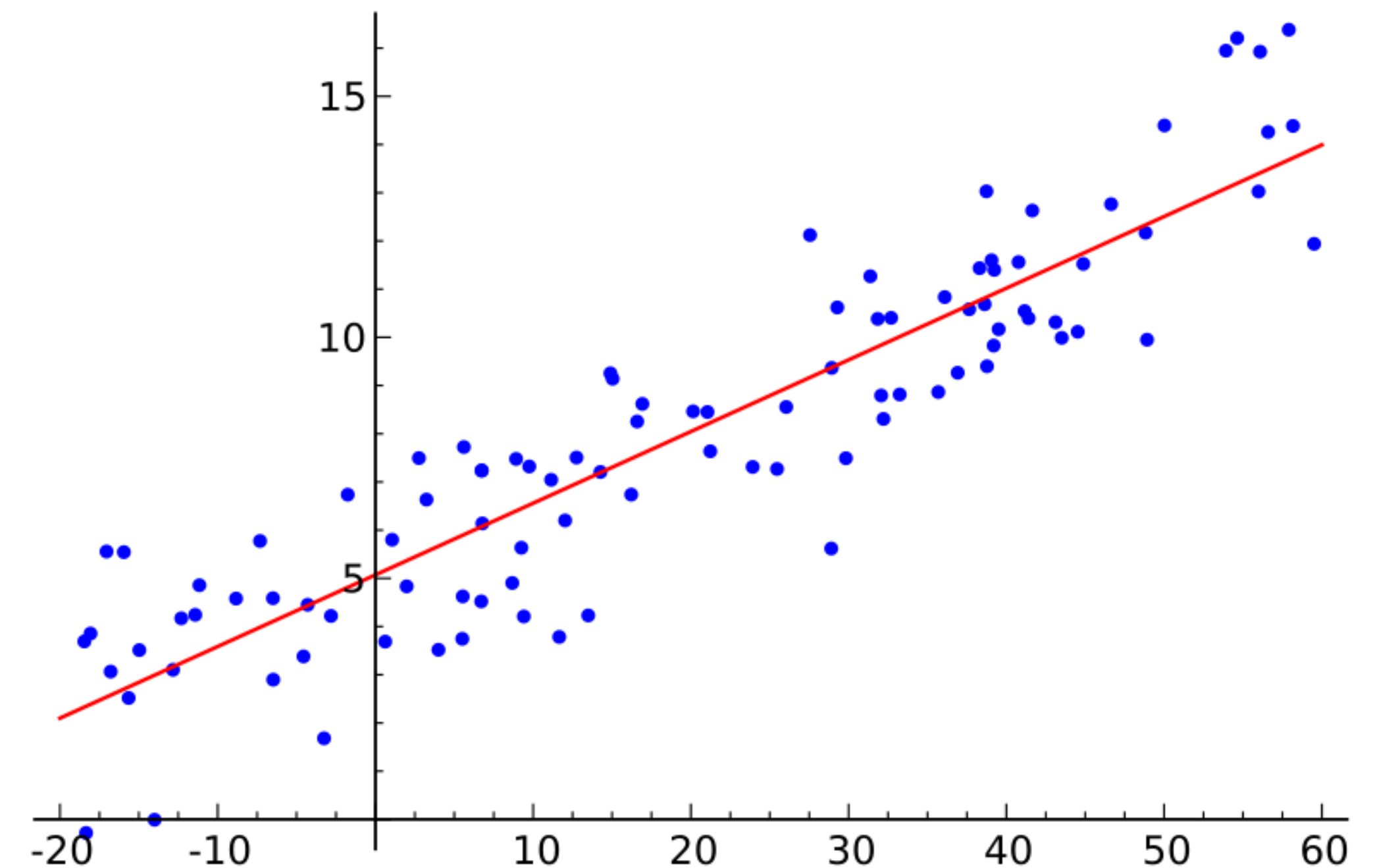
- Given advanced and emerging ML algorithms, are statistical models such as linear regression still important?
- What are the differences between statistical models and ML?



Statistics

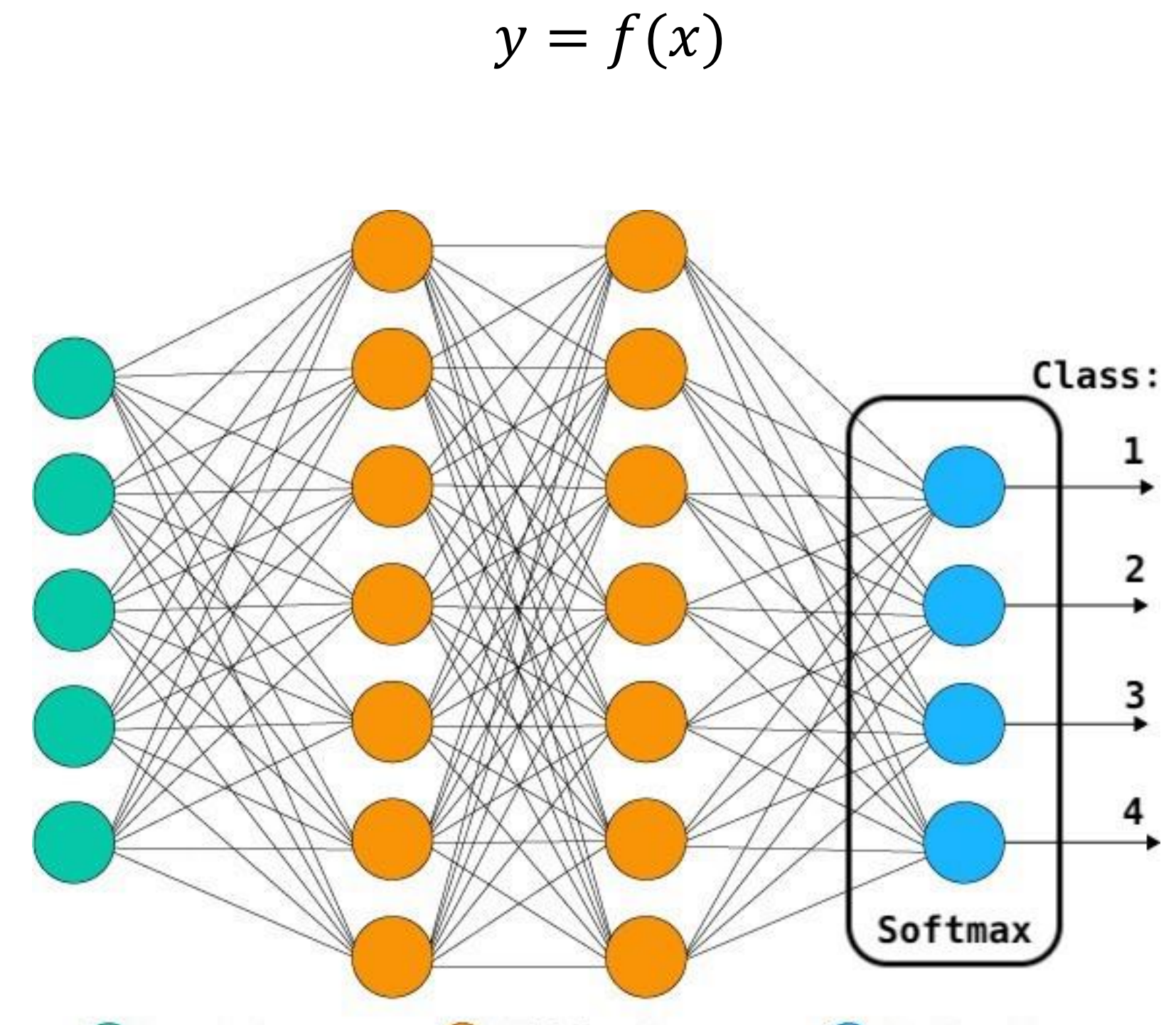
$$\hat{y}_i = \sum_k \beta_k x_{ik} + \beta_0$$

- Assumptions (Linear relationship, independent errors, normally distributed errors, equal variance of errors): need to test if assumptions hold true
- Need to check multicollinearity between variables
- Simple model structure
- Relatively low predictive accuracy, good interpretation



Machine learning

- Very few assumptions
- Complex model structure, difficult to understand why they work
- Require a lot of data
- High predictive accuracy and relatively low interpretation (black-box)



Major difference: purposes

- Statistical models are more used to estimate the (causal) relationship between variables, esp in social and economic research
- Does smoking lead to lung cancer?
- Does family background affect the level of education?
- ML models are good for predictions, esp when the data size is huge and predictive performance is the priority
- What is tomorrow's weather like?
- Can we automatically classify emails into spam and non-spam?



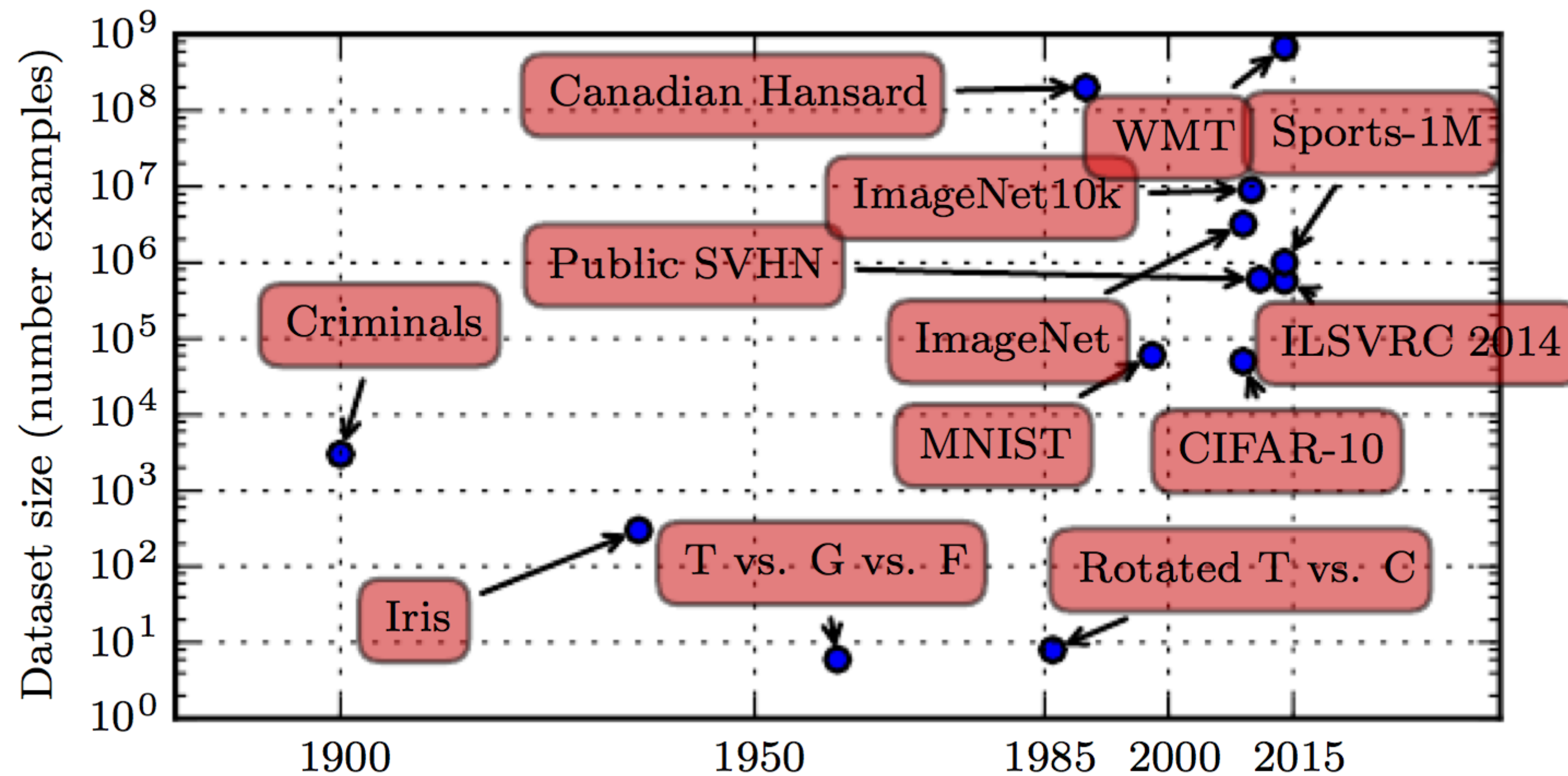
Theorem 1:

Good data is more important than algorithms

(similar to Garbage in, garbage out)

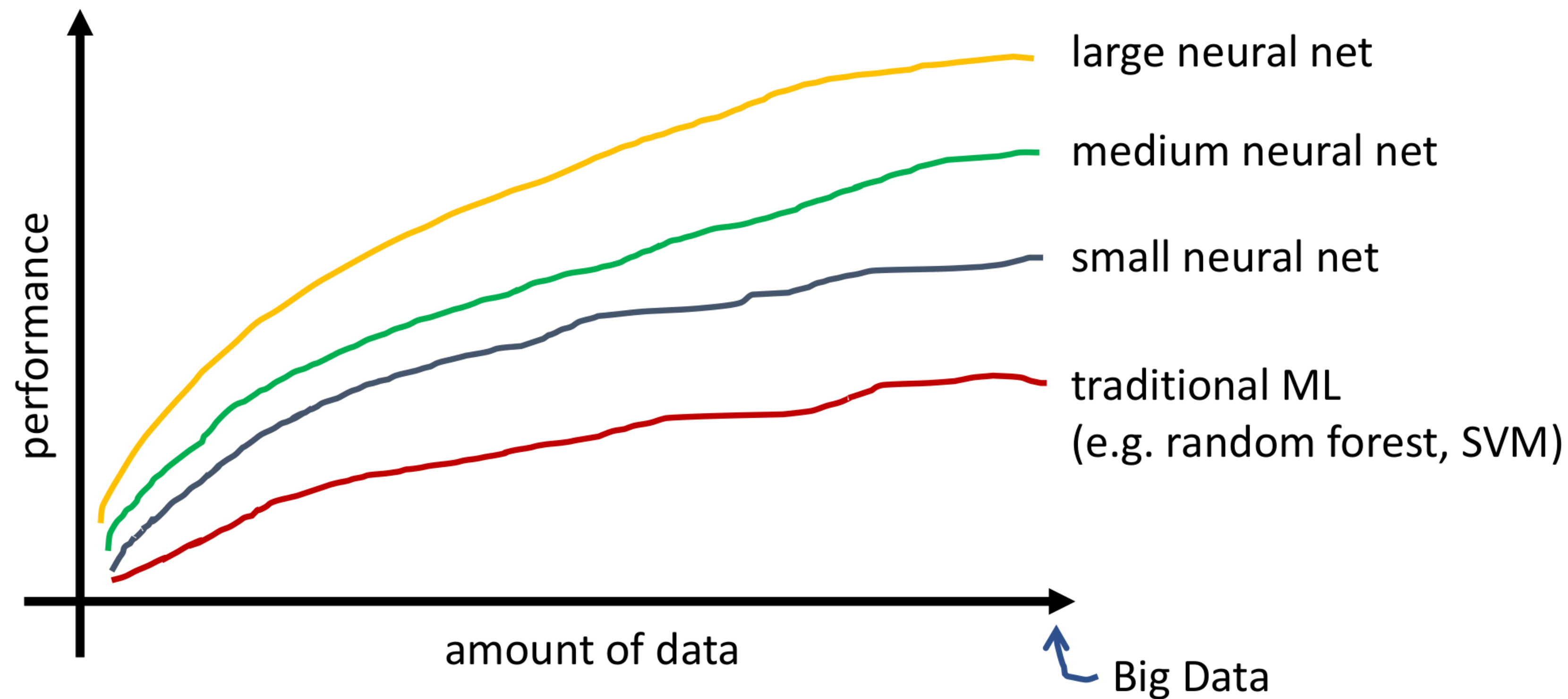
Good data = large size + high quality

Size of benchmark datasets



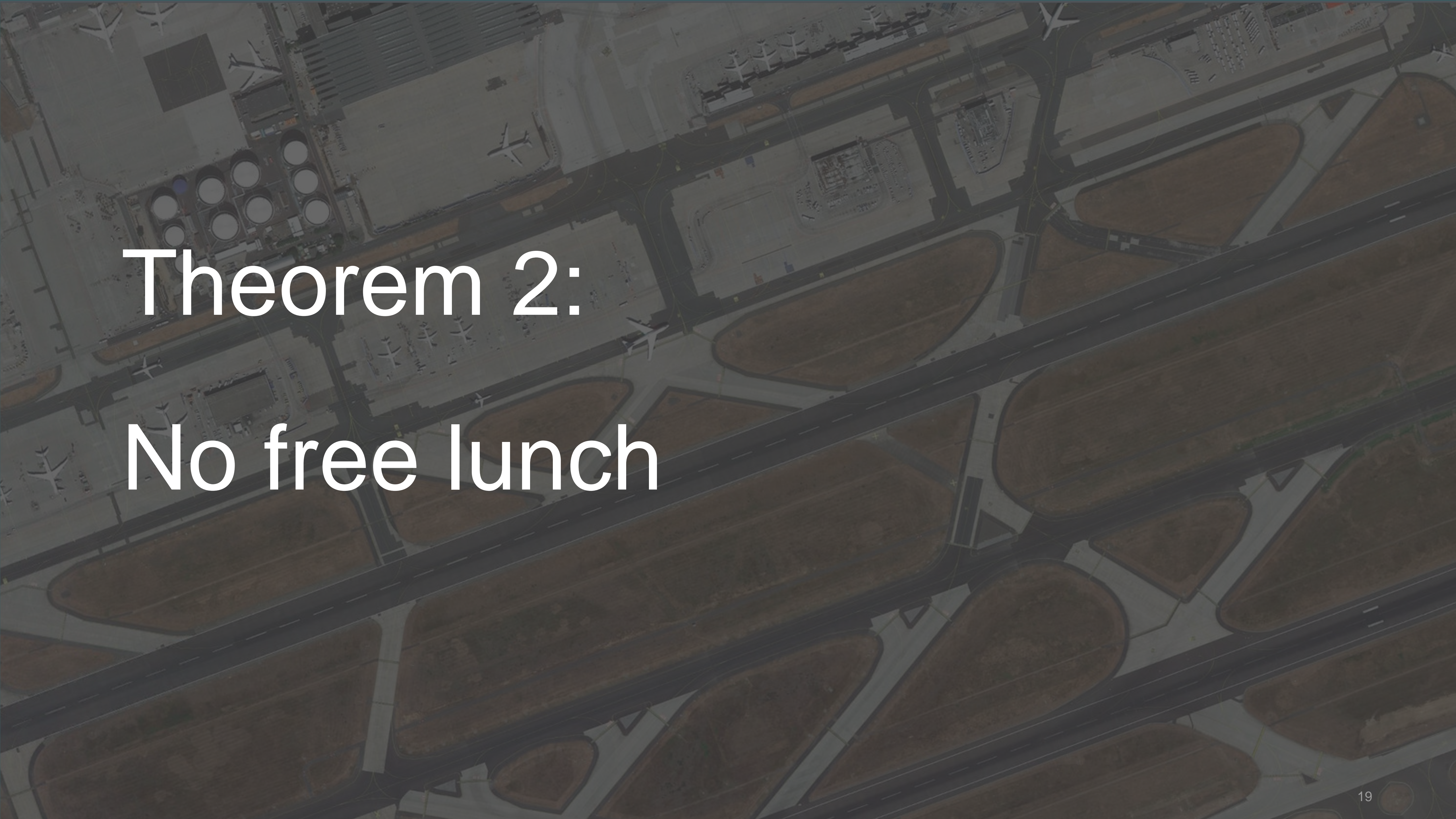
Source: <https://www.deeplearningbook.org/>

The performance of ML/DL increases rapidly with the size of the data



Data quality is important

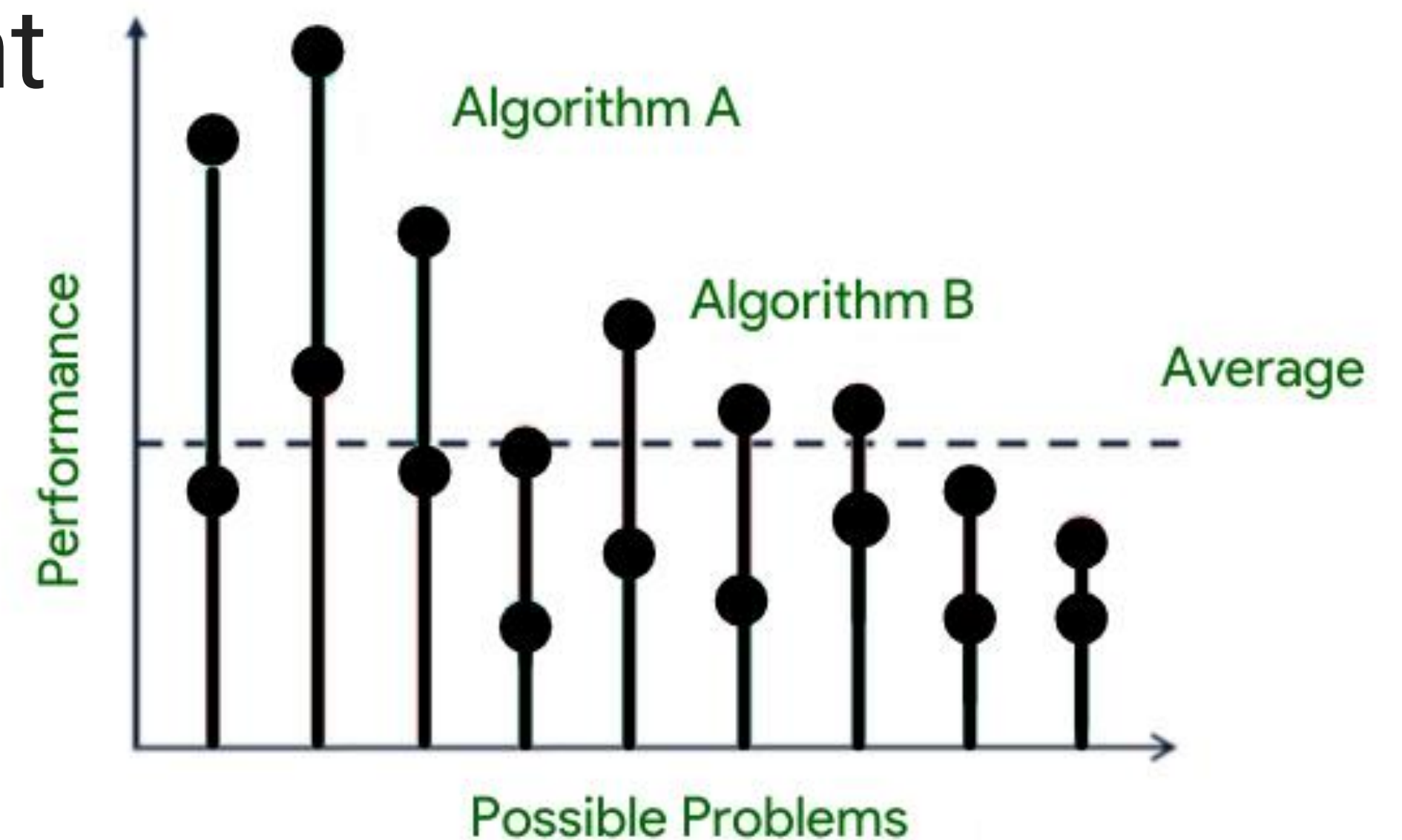
- Feature engineering: using domain knowledge to extract and transform features from raw data, in order to improve the performance of ML algorithms
 - Removing noisy or erroneous data
 - Dealing with missing data
 - Generating new features by combining existing ones
- Example of house price prediction: how can locations be used in ML methods?
 - Long/lat;
 - Distance to POIs (train stations/schools);
 - etc.
- What are features in ML? Characteristics, properties, attributes, explanatory variables.



Theorem 2: No free lunch

Sorry, no free lunch

- Essentially, all algorithms are equivalent when performance is averaged over all possible problems.
- There is no a priori model that is guaranteed to work best on all problems.
- Therefore, it is a matter of validating models empirically



Tree models vs. NN

- Which model is more competitive?
Depending on the data type
- For tabular data, tree models have higher predictive accuracy
- For image/text data, NNs are easier to use and have better performance

size of house (square feet)	# of bedrooms	price (1000\$)
523	1	115
645	1	0.001
708	unknown	210
1034	3	unknown
unknown	4	355
2545	unknown	440

Tabular data

I read the news today, oh boy
About a lucky man who made the grade
And though the news was rather sad
Well, I just had to laugh
I saw the photograph

He blew his mind out in a car
He didn't notice that the lights had changed
A crowd of people stood and stared
They'd seen his face before
Nobody was really sure if he was from the House of Lords

Unstructured data

Summary

- There are three types of ML
- The difference between supervised and unsupervised ML
- The difference between statistical models and ML. Choose the tool according to the purpose and domain
- Good data is more important than algorithms
- No free lunch

Workshop

- Weekly quiz on Moodle: please finish them before the workshop and we will discuss the quiz in the workshop
- Python notebooks for workshop: will be ready by 5pm Thursday
- See you in the workshop on Friday 1–3pm