

# CS-C1000 – Introduction to Artificial Intelligence

## Machine Learning

**Arno Solin**

March 12, 2021

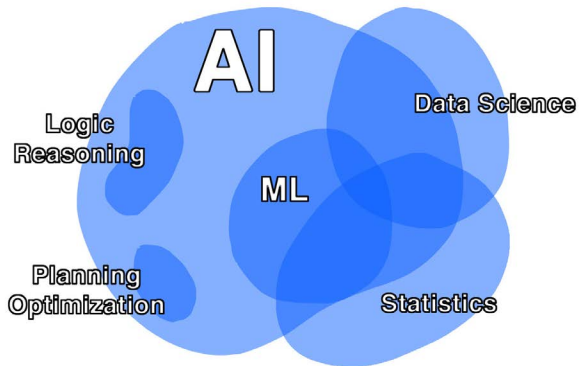
 @arnosolin

 arno.solin.fi

# Outline

- ▶ What is machine learning?
- ▶ Data
- ▶ Models
- ▶ Algorithms
- ▶ Examples of machine learning application areas

# AI and Machine Learning



# What is machine learning?

*“Can machines think?”*

is replaced with the question

*“Can machines do what we  
(as thinking entities) can do?”*

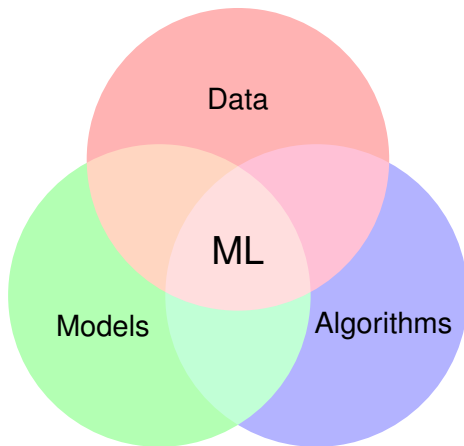
# What is machine learning?

- ▶ Intelligence involves learning—artificial or not.
- ▶ ML extracts high-level insights from raw data.
- ▶ One can say that ML is low-level AI.
- ▶ Or even in some cases **pragmatic AI**.

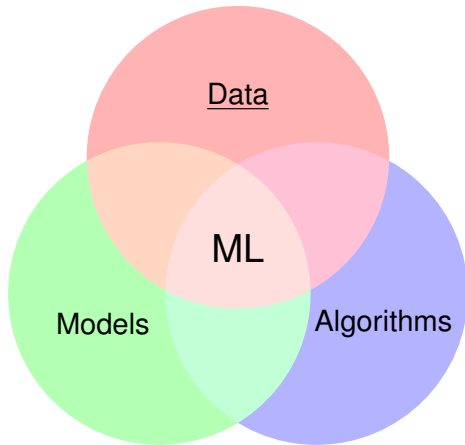
# What is machine learning?

- ▶ Computer programs (algorithms) which use **past data** to learn how to make **predictions** about future data.
- ▶ In ML learning is called **training**.
- ▶ Typically the learning/training phase is a separate stage, after which the method is put into use.

# Machine learning



# Data





# What is data?

- ▶ Basically anything that can be stored on a computer.

# What is data?



## Traditional data:

spreadsheets, database tables, ...



## Time-evolving:

logs, event sequences, transactions, ...



## Multimedia:

images, video, sound, music, ...



## Text:

news articles, books, tweets, web-searches, ...



## Location data:

coordinates, routes, geotags, ...



## Networks:

social networks, road networks, who-calls-whom, ...

# Representing data

- ▶ Numerical data is straight-forward.
- ▶ Text can be encoded, e.g., by ‘bag of words’:

vocabulary = {‘I’, ‘you’, ‘am’, ‘are’, ‘a’, ‘cat’, ‘dog’}

“I am a cat” = (1, 0, 1, 0, 1, 1, 0)

“You are a dog” = (0, 1, 0, 1, 1, 0, 1)






- ▶ Images are concatenated into arrays of numbers.

# Representing image data

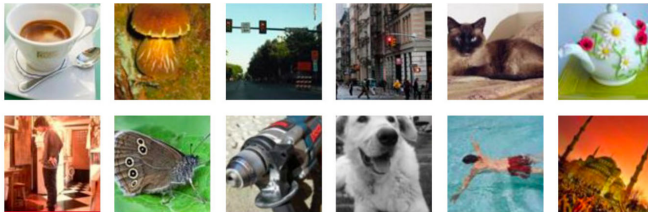
34	5	6	14	17	21	22	23	26	25	24	19	42	135	109	63
133	57	21	18	18	20	21	19	22	28	29	52	135	189	97	67
172	129	65	31	24	26	27	26	26	26	45	107	181	185	61	63
162	144	81	38	20	34	30	28	32	40	57	112	175	137	51	53
122	125	101	61	46	36	44	43	41	57	63	92	146	103	57	54
95	88	66	54	67	47	61	55	67	51	79	75	75	58	63	70
47	44	55	79	90	55	73	57	98	100	116	110	78	63	59	84
58	54	67	112	115	137	123	80	129	163	165	125	99	71	50	81
66	74	90	126	154	128	121	84	115	112	138	133	135	81	42	52
65	69	101	116	96	72	61	67	73	62	95	115	148	106	49	42
68	107	123	96	93	92	47	45	46	81	115	149	143	105	68	54
93	81	124	146	120	86	86	71	87	104	117	182	145	125	98	62
107	84	101	165	165	113	129	127	127	159	160	170	136	122	107	78
104	101	117	135	148	168	173	126	134	180	152	126	90	105	120	115
93	98	105	100	105	107	130	155	159	139	124	90	101	103	111	135
72	78	82	85	71	76	80	103	113	117	103	84	118	86	127	126

# Labeled data

- ▶ Each data point has an informative label.
- ▶ The labels go under various names: *tag*, *label*, *class*, ...
- ▶ Can be seen as input–output pairs.

Data	Label
	True
	True
	False
	True
	True

# Unlabeled data



Example images from the ImageNet database.

- ▶ 'Just data' without any additional information what it is.
- ▶ Easy to acquire, but can be chaotic.

# Data mining

- ▶ Data mining is the process of **discovering patterns** in large data sets.
- ▶ Involving methods at the intersection of **machine learning**, **statistics**, and **database systems**.
- ▶ “Knowledge discovery in databases”.

# Real vs. synthetic data

- ▶ Acquiring data samples can be tedious, dangerous, or expensive.
- ▶ Sometimes even impossible.
- ▶ Synthetic data can help in making the data more versatile.



Figure: Johnson-Roberson *et al.* Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks?





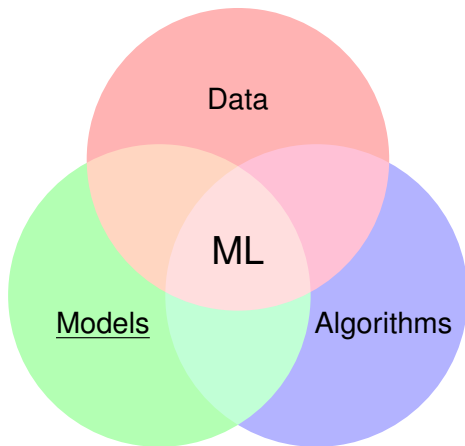
**Real domain**



**Simulated domain**

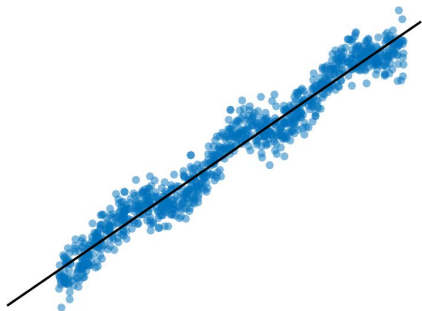
Simulation Training, Real Driving  
<https://wayve.ai/blog/sim2real>

# Models



# Models

- ▶ The model encapsulates our **belief** of what type of process could be generating the data.
- ▶ Models typically have some **parameters**.
- ▶ **Learning** (often) amounts to characterizing/finding those parameters.



$$y = ax + b$$

# It is just a model

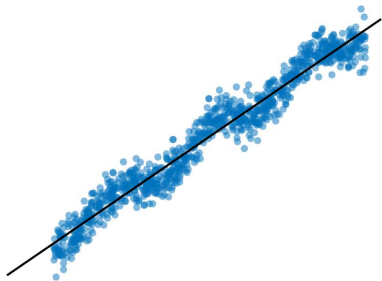
*“All models are wrong, but some are useful.”*

George Box

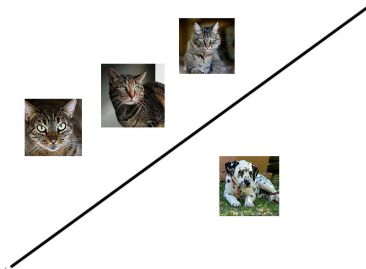
# Probabilistic models

- ▶ Related terms:  
Statistical inference / Probabilistic methods / Bayesian.
- ▶ Deals with quantities in terms of probability distributions
  - ▶ **Prior**: beliefs about the phenomenon.
  - ▶ **Posterior**: the probability (distribution) after the data has been taken into account.
- ▶ Important for **uncertainty quantification**.

# Regression vs. classification



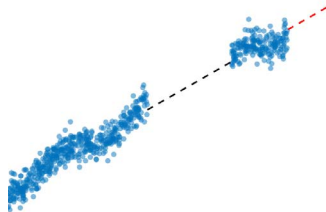
- ▶ Estimating relationships among variables.
- ▶ Given your height and weight: How old are you?



- ▶ Separating samples into classes.
- ▶ Given your height and weight: Is your age  $> 25$ ?

# Interpolation vs. extrapolation

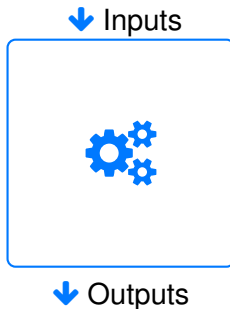
- ▶ **Interpolation:**  
Fill in the gaps.  
➔ Not that hard.
- ▶ **Extrapolation:**  
Come up with something new.  
➔ Clearly harder.



# Black vs. white box models



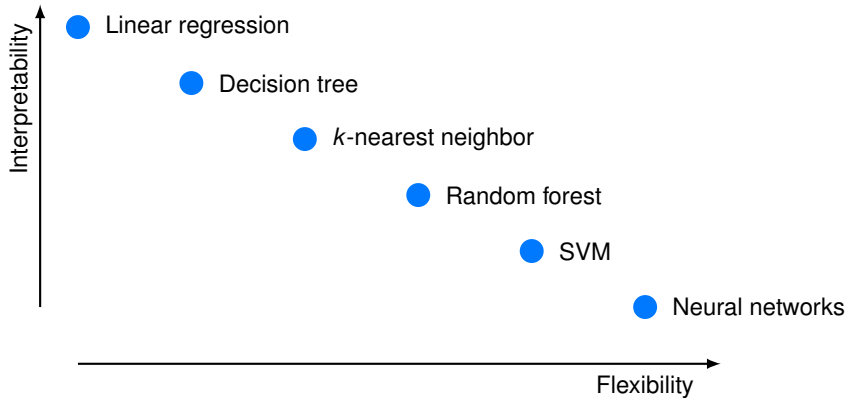
No visibility  
to what the model does.



Visibility  
to what the model does.

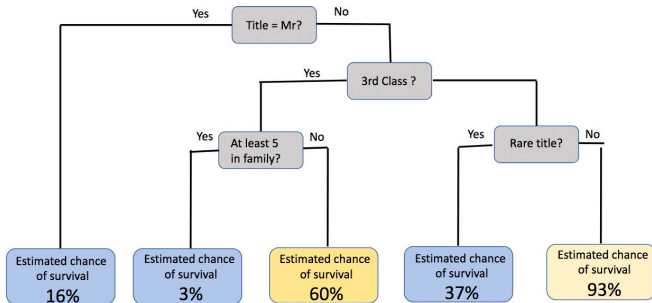


# Interpretability of different models



# Decision trees

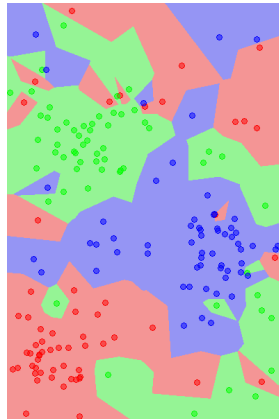
- ▶ Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.
- ▶ Called 'regression trees' if the target variable takes continuous values.



Titanic survival prediction from David Spiegelhalter's talk <https://videoken.com/embed/JBddhT9fius>

# $k$ -nearest neighbors

- ▶ The input consists of the  $k$  closest training examples in the feature space.
- ▶  $k$ -NN classification: An object is classified by a plurality vote of its neighbors.
- ▶  $k$ -NN regression: This value is the average of the values of its  $k$  nearest neighbors.



# Random forest

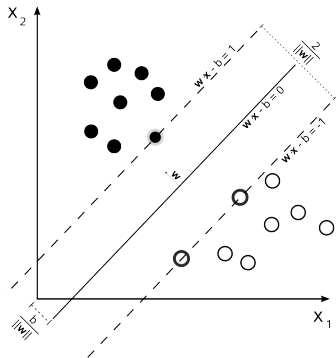
- ▶ An **ensemble method** constructing a multitude of decision trees.



- ▶ The prediction is then based on the mode of the classes (classification) or mean (regression) of the individual trees.
- ▶ Corrects for some bad tendencies in decision trees (see overfitting).

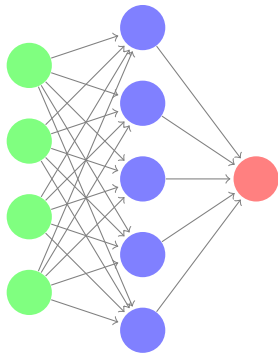
# Support-vector machines (SVM)

- ▶ A non-probabilistic binary linear classifier (also non-linear variants).
- ▶ Finds separating hyperplanes between classes.
- ▶ Practical and used a lot in applications.



# Neural networks

- ▶ Inspired by the biological neural networks (brains).
- ▶ Collection of connected units or nodes called artificial neurons.
- ▶ Have been very successful recently in certain type of tasks (vision, audio, *etc.*).



# Overfitting vs. underfitting

Important problems in machine learning:

- ▶ **Underfitting:**

Fit a (too) simple model to all the data.

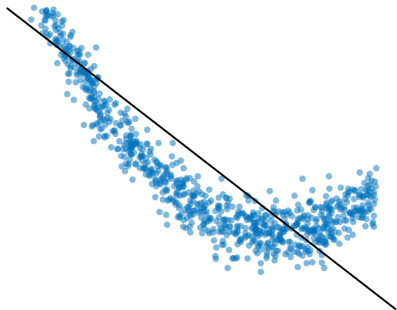
➔ Can't adapt to local differences.

- ▶ **Overfitting:**

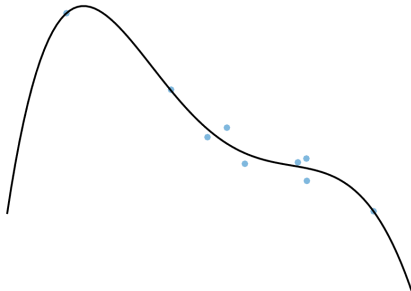
Fit a complex model to small local neighborhood.

➔ Too little data to be accurate.

# Underfitting vs. overfitting



Lots of data — simple model



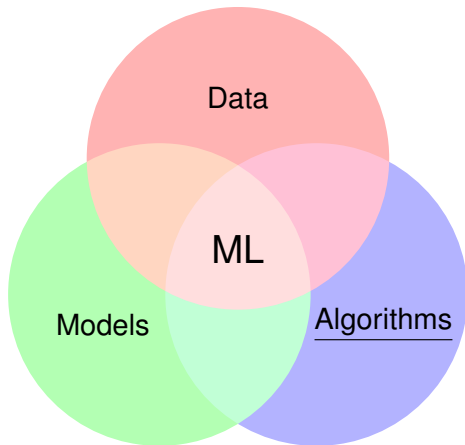
Little data — complicated model



# Training and testing with data

- ▶ If the **same data is used for training a model and testing** how well it works, there is an apparent risk of overfitting to the training data.
- ▶ This gives false confidence in how well the model generalizes to **unseen** data.
- ▶ A typical way to deal with this is to **hold out** some data from the data set to evaluate the performance of the model with.
- ▶ The procedure can also be repeated by splitting the data into subsets, and training and testing with different subsets at a time (**cross-validation**).

# Algorithms



# Algorithms

- ▶ An algorithm is an unambiguous **specification** of how to solve a specific problem.
- ▶ It is the **set of instructions** to go through to train a model (or evaluate predictions).
- ▶ We consider ‘algorithms’ in a broad sense, also covering **different ways of learning**.

# Supervised vs. unsupervised learning



Supervised  
learning



Unsupervised  
learning

# Supervised learning

- ▶ The algorithm is given **labeled data**, with known inputs and outputs.
- ▶ The algorithm builds a model that tries to capture the link between the inputs and outputs.
- ▶ **Example:** Given a labeled set of pictures of cats and dogs, learn to do binary classification for unseen cat/dog images.
- ▶ **Example:** Given a set of your height measurements of a child up to age 10, predict how tall the child will be as an adult.

# Unsupervised learning

- ▶ The algorithm is given samples of data, but **without labels**.
- ▶ The aim is to discover the structure of the data by grouping similar things together (**clustering**) or reducing the complexity of the data by mapping it to some lower-dimensional space.
- ▶ **Example:** Given a large set of face images, learn what a human face is.
- ▶ **Example:** Given a large set of hand drawn digits, learn to tell them apart.

# Active learning

- ▶ A special case of machine learning in which a learning algorithm is able to **interactively query** the user (or something else) to obtain desired outputs of data points.
- ▶ Well suited for cases where unlabeled data is not informative enough, but labeling is expensive.
- ▶ Can be seen as iterative supervised learning (or **semi-supervised** learning).
- ▶ **Example:** Pedestrian detection or customer feedback analysis.

# Feature learning

- ▶ A form of unsupervised (can also be supervised) ML that tries to discover the underlying features of data.
- ▶ Can be handy in getting a grasp of data, **simplifying** complicated data to something more understandable.
- ▶ **Dimensionality reduction** is often used in data visualization.
- ▶ Some common techniques include Principal component analysis (PCA), matrix factorizations, and autoencoders.

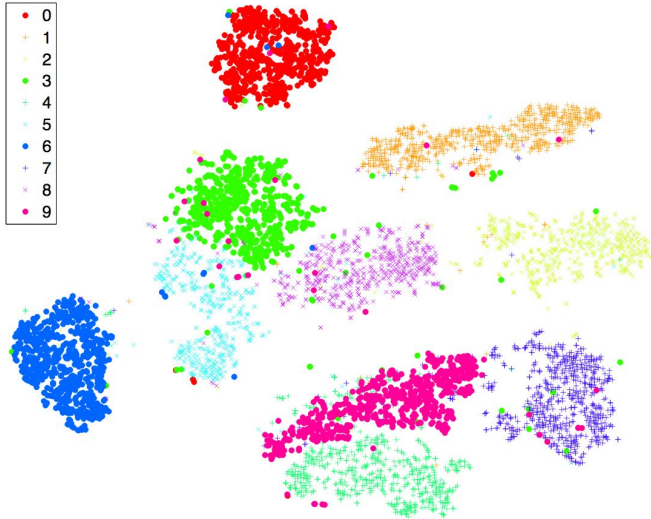


# Feature learning



From van der Maaten and Hinton (2008). Visualizing Data using t-SNE. JMLR.

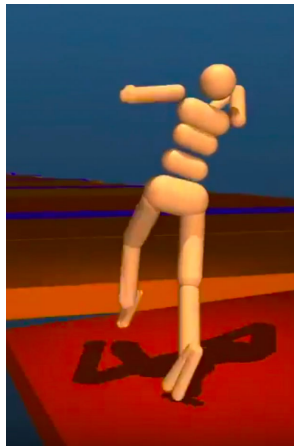
# Feature learning



From van der Maaten and Hinton (2008). Visualizing Data using t-SNE. JMLR.

# Reinforcement learning

- ▶ Concerned with how algorithms ought to take actions to maximize some cumulative reward.
- ▶ Typical uses cases where the task and environment are complicated, but some reward (and feedback) can be formulated.
- ▶ **Example:** Learning to walk, drive, grasp things.



# Example application areas of ML

- ▶ Agriculture
- ▶ Anatomy
- ▶ Adaptive websites
- ▶ Affective computing
- ▶ Bioinformatics
- ▶ Brain-machine interfaces
- ▶ Cheminformatics
- ▶ Computer Networks
- ▶ Computer vision
- ▶ Credit-card fraud detection
- ▶ Data quality
- ▶ DNA sequence classification
- ▶ Economics
- ▶ Financial market analysis
- ▶ General game playing
- ▶ Handwriting recognition
- ▶ Information retrieval
- ▶ Insurance
- ▶ Internet fraud detection
- ▶ Linguistics
- ▶ Machine learning control
- ▶ Machine perception
- ▶ Machine translation
- ▶ Marketing
- ▶ Medical diagnosis
- ▶ Medical imaging
- ▶ Natural language processing
- ▶ Online advertising
- ▶ Optimization
- ▶ Recommender systems
- ▶ Robot locomotion
- ▶ Search engines
- ▶ Sentiment analysis
- ▶ Sequence mining
- ▶ Software engineering
- ▶ Speech recognition
- ▶ Structural health monitoring
- ▶ Syntactic pattern recognition
- ▶ Telecommunication
- ▶ Theorem proving
- ▶ Time series forecasting
- ▶ User behavior analytics

# Anomaly detection

- ▶ Often coupled with data mining and known as **outlier detection**.
- ▶ Identification of rare events or observations which raise suspicions by differing significantly from the majority of the data.
- ▶ **Example:** Detecting credit card fraud by finding transactions that do not fit with the normal behavior of the user.
- ▶ **Example:** Identification of medical problems, structural defects, or errors in text.

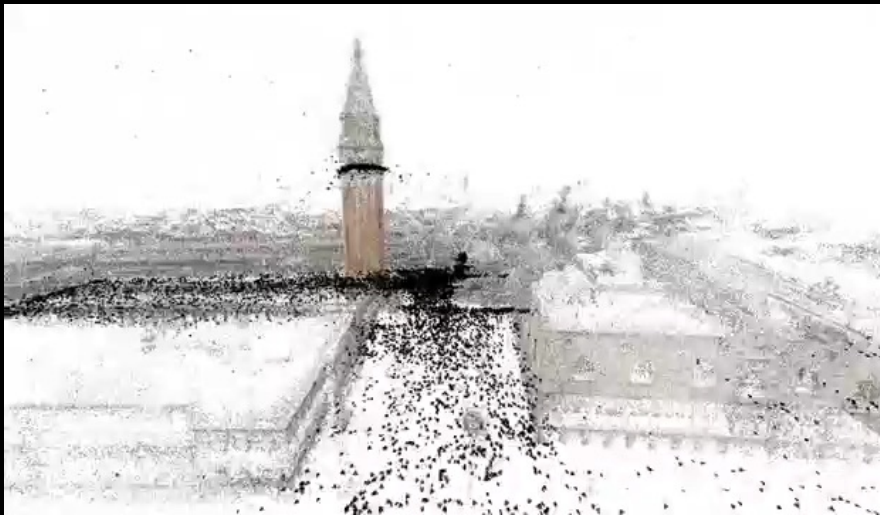
# Sentiment analysis

- ▶ Also known as **opinion mining** or **emotion AI**.
- ▶ Combination of natural language processing, text analysis, computational linguistics, and biometrics.
- ▶ Identify, extract, quantify, and study affective states and subjective information.
- ▶ **Example:** Analysis of customer reviews, responses to treatments, social networking platforms.



# Computer vision

- ▶ Humans see effortlessly, but there are billions and billions of neurons dedicated for this in our brain.
- ▶ Visual **perception** is needed for many everyday tasks.
- ▶ In many cases, easier to show things to an AI than trying to explain in other ways.



Reconstructing San Marco square in Venice from 13k tourist photos:  
<https://www.youtube.com/watch?v=y9zF97JL30A>

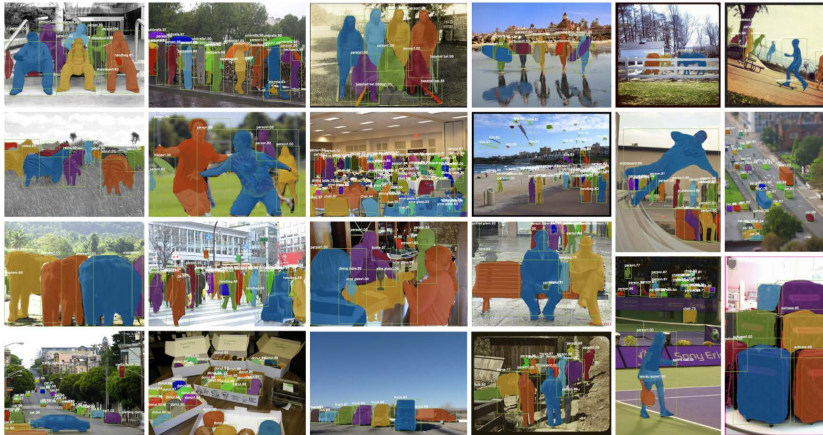


# State-of-the-art in 2008



Felzenszwalb *et al.* (2008). A discriminately trained, multi scale, deformable part mode. CVPR.

# State-of-the-art now



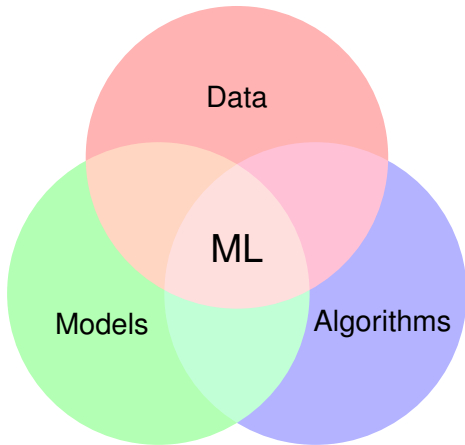
He *et al.* (2017). Mask R-CNN. ICCV.

# State-of-the-art now



He *et al.* (2017). Mask R-CNN. ICCV.

# Recap



**AI**