LMP 1210H: Basic Principles of Machine Learning in Biomedical Research

Bo Wang
AI Lead Scientist, PMCC, UHN
CIFAR AI Chair, Vector Institute
Assistant Professor, University of Toronto

Peter Munk Cardiac Centre

Administrative Details

1. Teaching Staff

Instructor: Bo Wang, **bowang.wang@utoronto.ca**
TA: Adamo Young, **adamo.young@mail.utoronto.ca**

2. Useful Information

Office Hours: Friday 10-11am (except reading week and holidays), Zoom
Piazza: the best way to ask questions!
Website: **https://lmp1210-uoft.github.io/2022/**

Administrative Details

3. Evaluations

Three assignments (**45%**) (Theory + Coding)

Term projects (**40%**) (2-3 students per group, details later)

In-class participation (**15%**)

More on Assignments

1. **Collaboration** on the assignments is **NOT** allowed!
   Each student is responsible for their own work. Discussion of assignments should be limited to clarification of the handout itself, and should not involve any sharing of pseudocode or code or simulation results. Violation of this policy is grounds for a semester grade of F, in accordance with university regulations.

2. **Assignments should be handed in by deadline.**

A late penalty of 10% per day will be assessed thereafter (up to 3 days, then submission is blocked.)

Extensions will be granted only in special situations, and you will need a Student Medical Certificate or a written request approved by the course coordinator at least one week before the due date.

More on Programming

1. We will use **Python only**!

   Python is the most popular programming language for machine learning. Tutorials about the basics of python will be provided. We will also have programming exercise at every assignment.

2. **Don’t be scared by Python!**

A real-life example:

Why this course?

1. One of the first graduate courses in medical departments about machine learning!

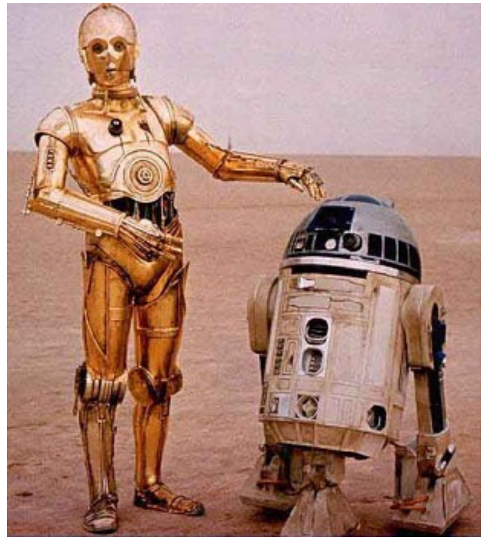
2. We actually code!

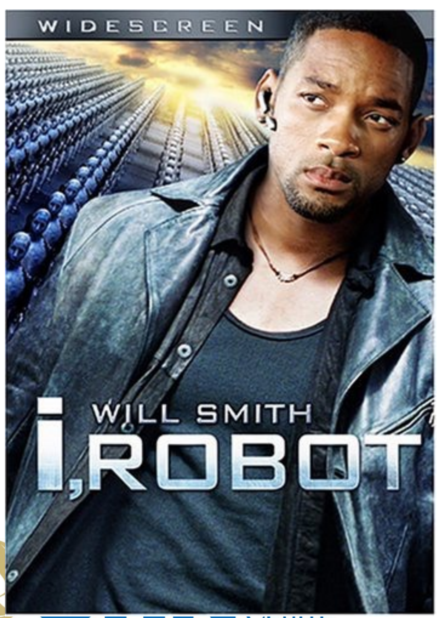
3. AI/ML is changing the way we perform research in medicine.

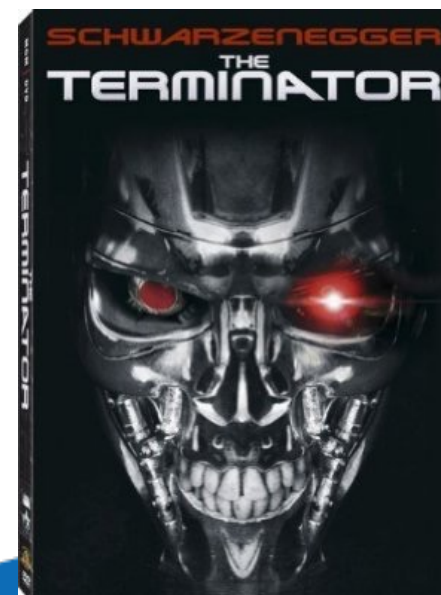
AI will not replace doctors, but doctors who use AI will replace those who don't. ---- some famous person.

AI will not replace PHDs, but PHDs who use AI will replace those who don't. ---- Bo Wang.

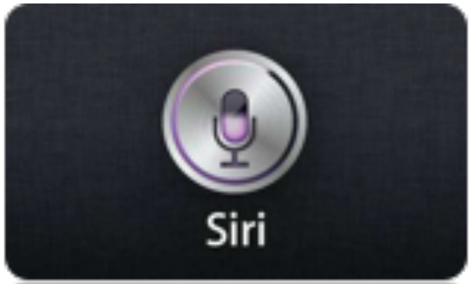
What is Artificial Intelligence (AI)



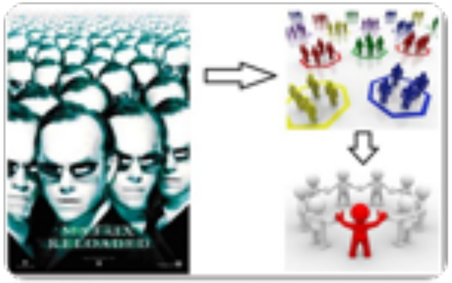




What is Artificial Intelligence (AI)



语音助理



个性化推荐

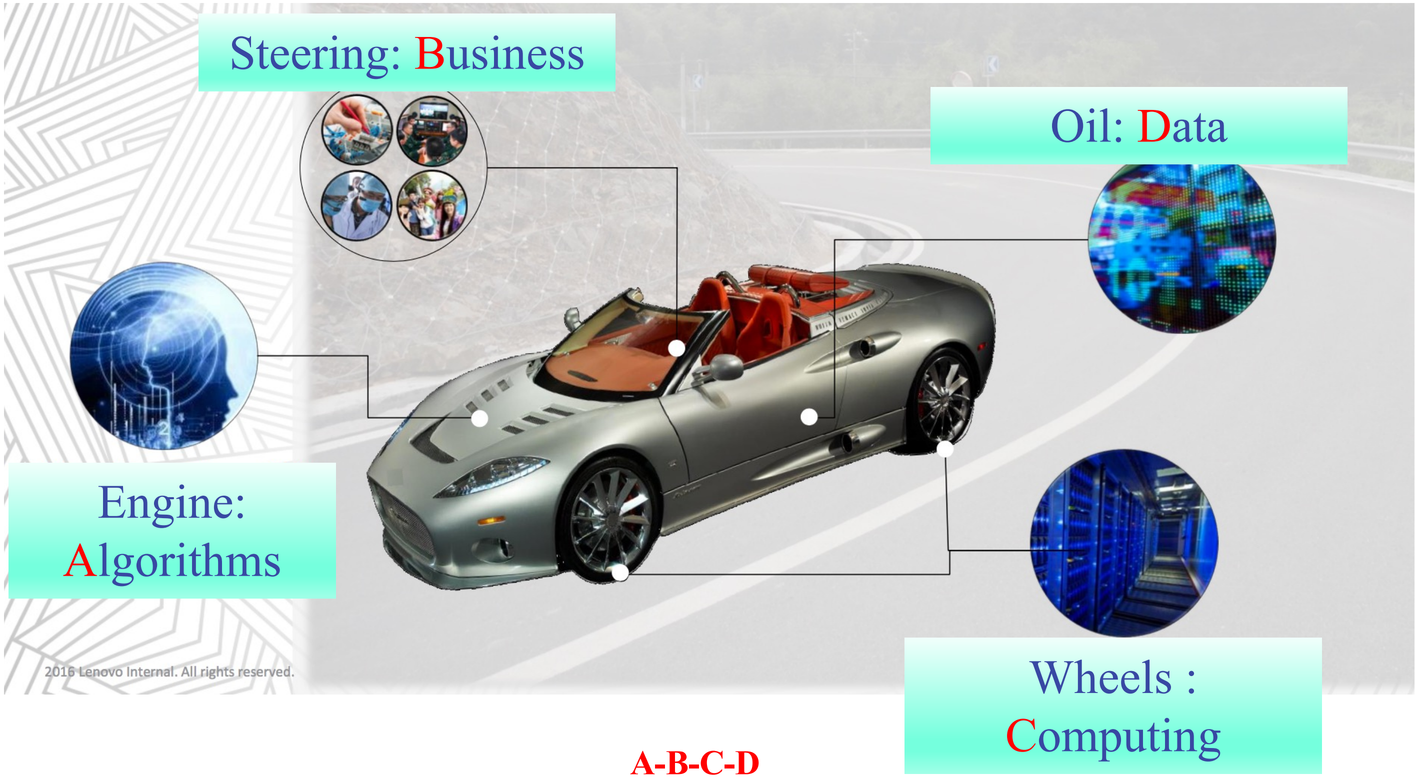
人脸识别

自动驾驶汽车

家庭助手机器人

工业机器人

What makes AI so successful?

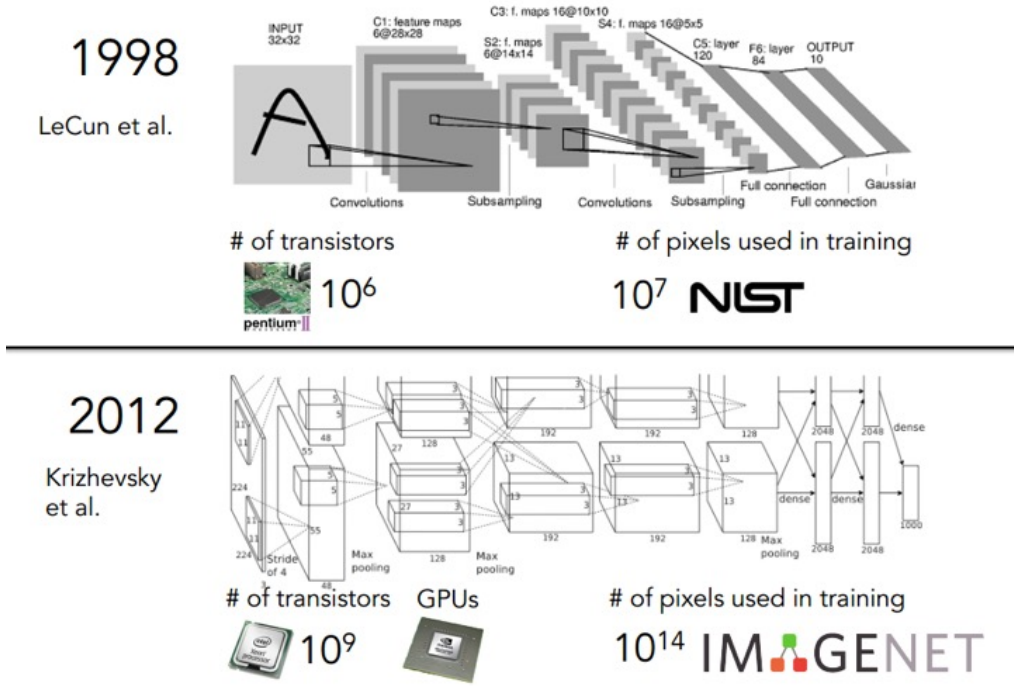


A-B-C-D

Peter Munk Cardiac Centre

A : Algorithms

Convolutional Neural Network (CNN) Images



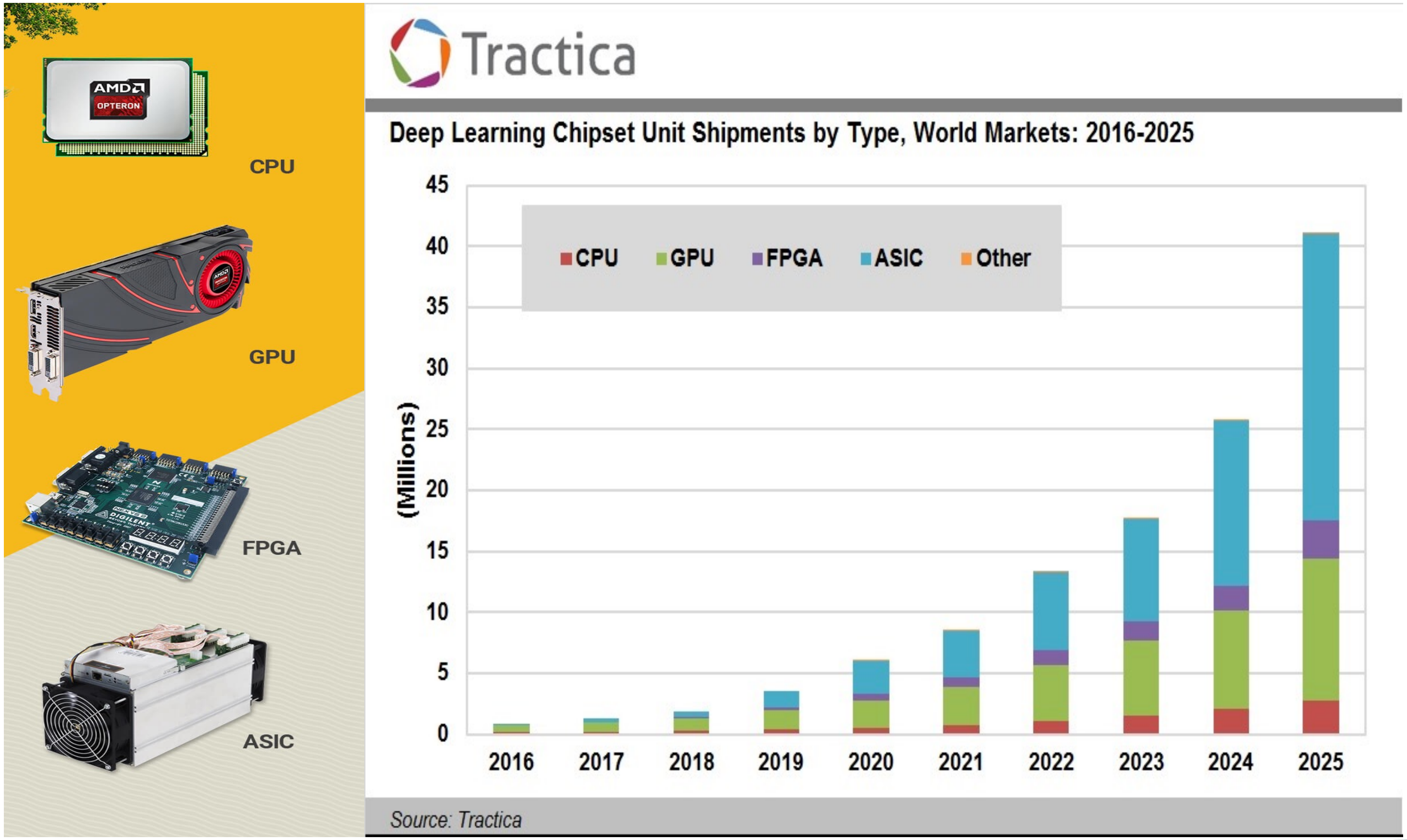
Very Deep Convolutional Neural Network (CNN)

Peter Munk Cardiac Centre

C : Computing

Peter Munk
Cardiac
Centre

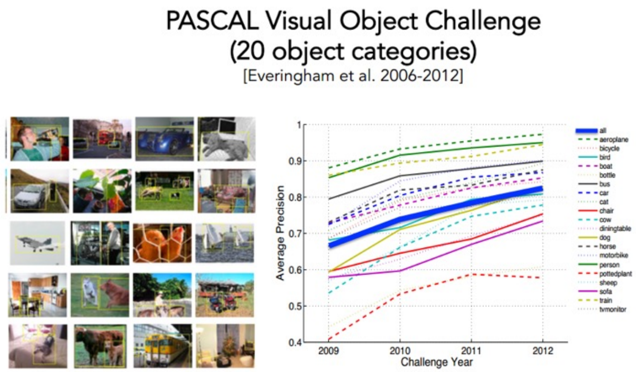
Computing Hardware for AI

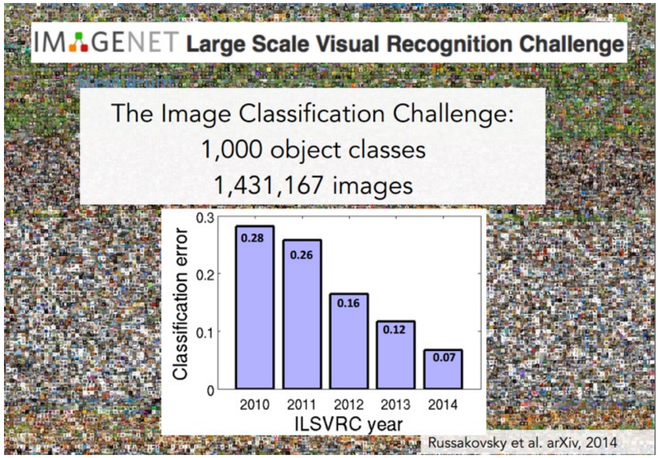


UHN Peter Munk Cardiac Centre

D : Data

Big Data in Computer Vision





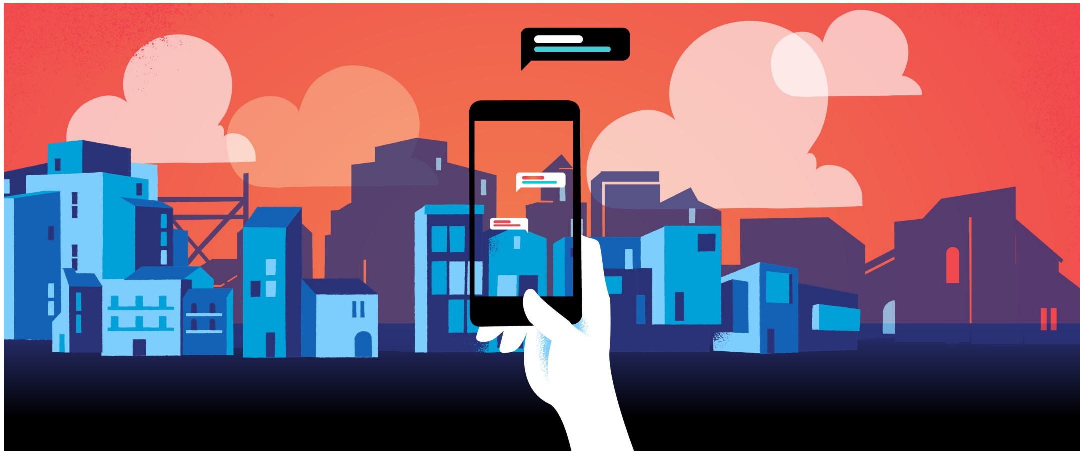
UHN Peter Munk Cardiac Centre

Crowd-sourcing Data Annotation

UHN Peter Munk Cardiac Centre

B : Business

AI + Computer Vision



Peter Munk Cardiac Centre

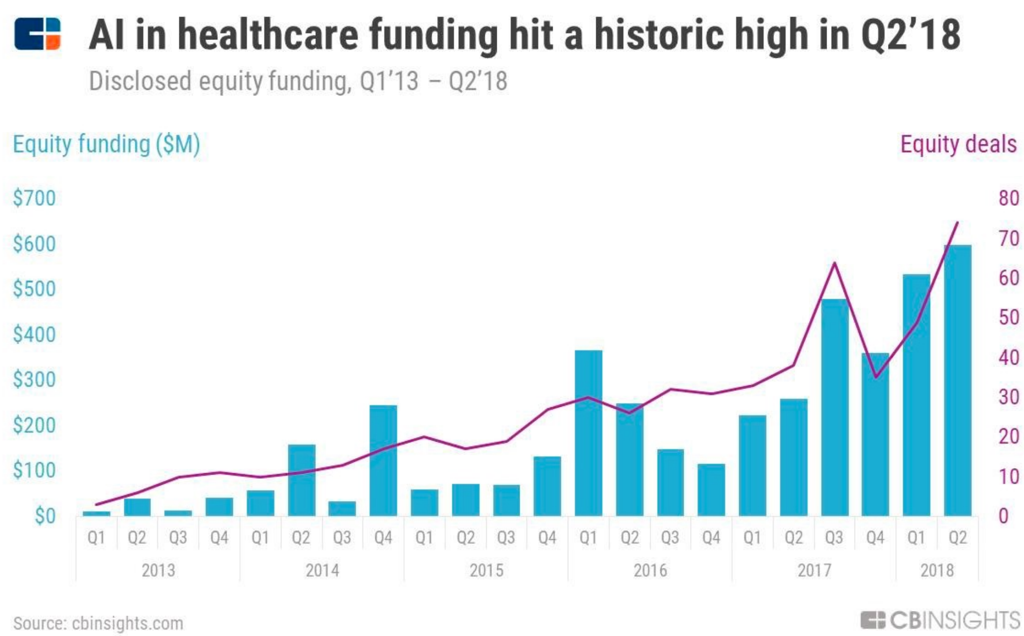
AI + Finance

AI + Manufacturing

AI + Retails

Peter Munk Cardiac Centre

New Era: AI + Health Care



UHN Peter Munk Cardiac Centre

Where are we now (AI + Healthcare)?

Where are we now (AI + Healthcare)?

Human driver monitors environment			System monitors environment		
0 No automation The absence of any assistive features such as adaptive cruise control.	1 Driver assistance Systems that help drivers maintain speed or stay in lane but leave the driver in control.	2 Partial automation The combination of automatic speed and steering control—for example, cruise control and lane keeping.	3 Conditional automation Automated systems that drive and monitor the environment but rely on a human driver for backup.	4 High automation Automated systems that do everything—no human backup required—but only in limited circumstances.	5 Full automation The true electronic chauffeur: retains full vehicle control, needs no human backup, and drives in all conditions.

BMW

Tesla

Baidu

Toyota

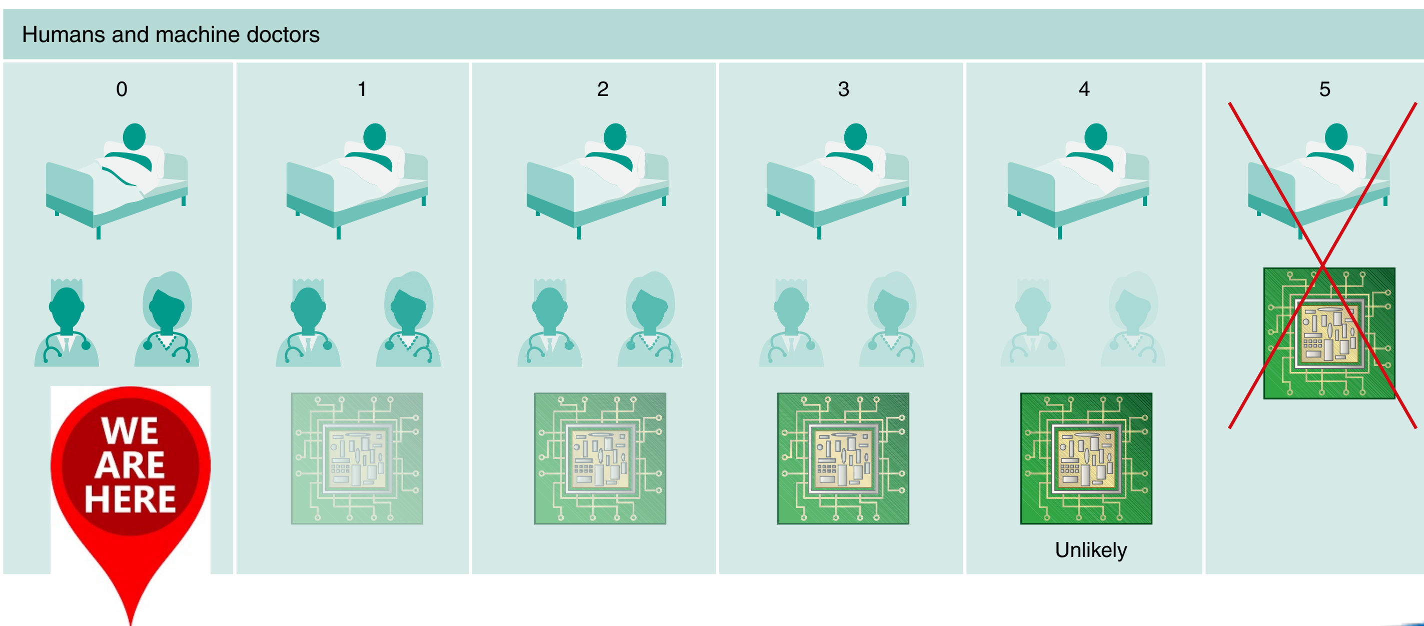
Waymo

Honda

Credit: Debbie Maizels/Springer Nature

Where are we now (AI + Healthcare)?

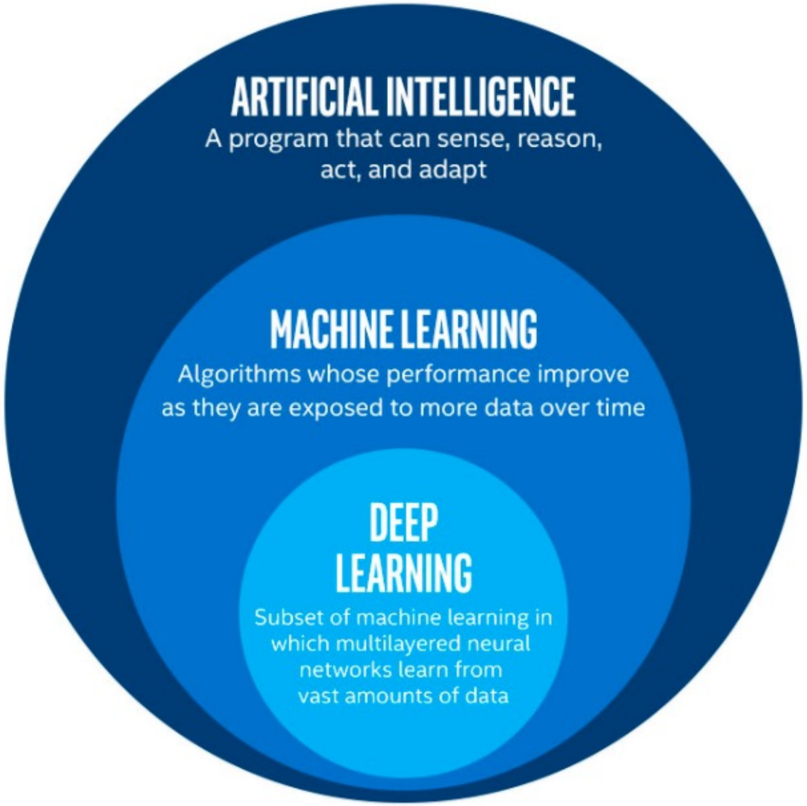
Human driver monitors environment			System monitors environment		
0 No automation	1 Driver assistance	2 Partial automation	3 Conditional automation	4 High automation	5 Full automation
The absence of any assistive features such as adaptive cruise control.	Systems that help drivers maintain speed or stay in lane but leave the driver in control.	The combination of automatic speed and steering control—for example, cruise control and lane keeping.	Automated systems that drive and monitor the environment but rely on a human driver for backup.	Automated systems that do everything—no human backup required—but only in limited circumstances.	The true electronic chauffeur: retains full vehicle control, needs no human backup, and drives in all conditions.



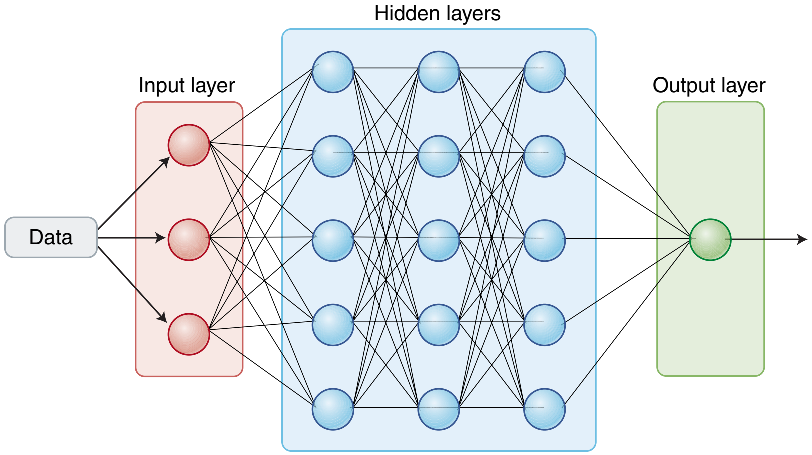
Credit: Debbie Maizels/Springer Nature

UHN Peter Munk Cardiac Centre

What is Artificial Intelligence (AI)



Deep Neural Network



Peter Munk Cardiac Centre

What is Machine Learning (ML)

MACHINE

LEARNING

Dictionary

Search for a word

ma·chine

/məˈSHēn/

See definitions in:

All | Mechanics | Politics

noun

an apparatus using or applying mechanical power and having several parts, each with a definite function and together performing a particular task.
"a fax machine"

What is Machine Learning (ML)

MACHINE

LEARNING

What is learning?

"The activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something."

Merriam Webster dictionary

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell

What is Machine Learning (ML)

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence.

----- From Wikipedia, the free encyclopedia

Relation to Human Learning

- Human learning is:
  - Very data efficient
  - An entire multitasking system (vision, language, motor control, etc.)
  - Takes at least a few years :)
- For serving specific purposes, machine learning doesn't have to look like human learning in the end.
- It may borrow ideas from biological systems, e.g., neural networks.
- It may perform better or worse than humans.

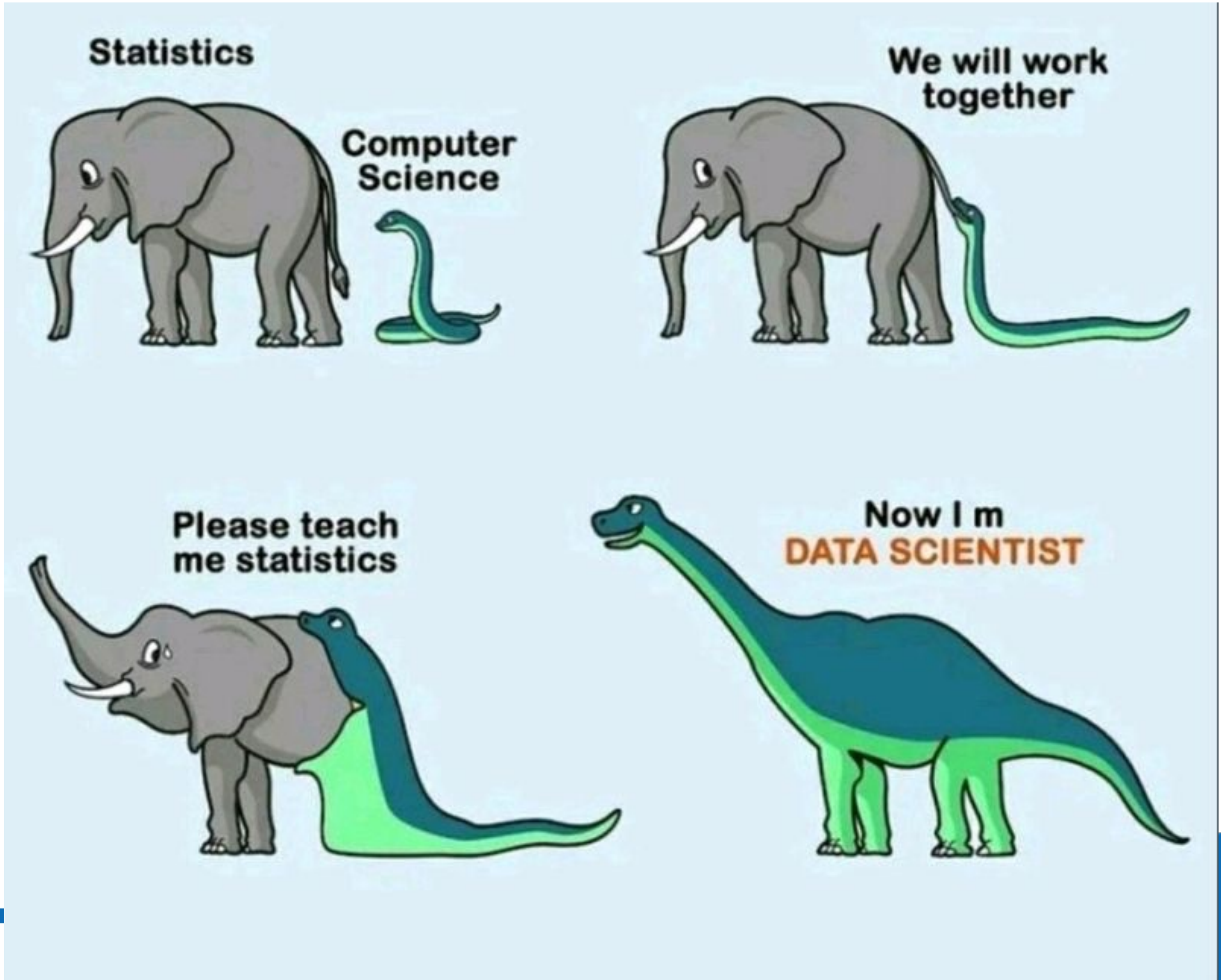
Relation to Statistics

- It is similar to statistics...
  - Both fields try to uncover patterns in data
  - Both fields draw heavily on calculus, probability, and linear algebra, and share many of the same core algorithms
- But it is not statistics!
  - Stats is more concerned with helping scientists and policymakers draw good conclusions; ML is more concerned with building autonomous agents
  - Stats puts more emphasis on interpretability and mathematical rigor; ML puts more emphasis on predictive performance, scalability, and autonomy

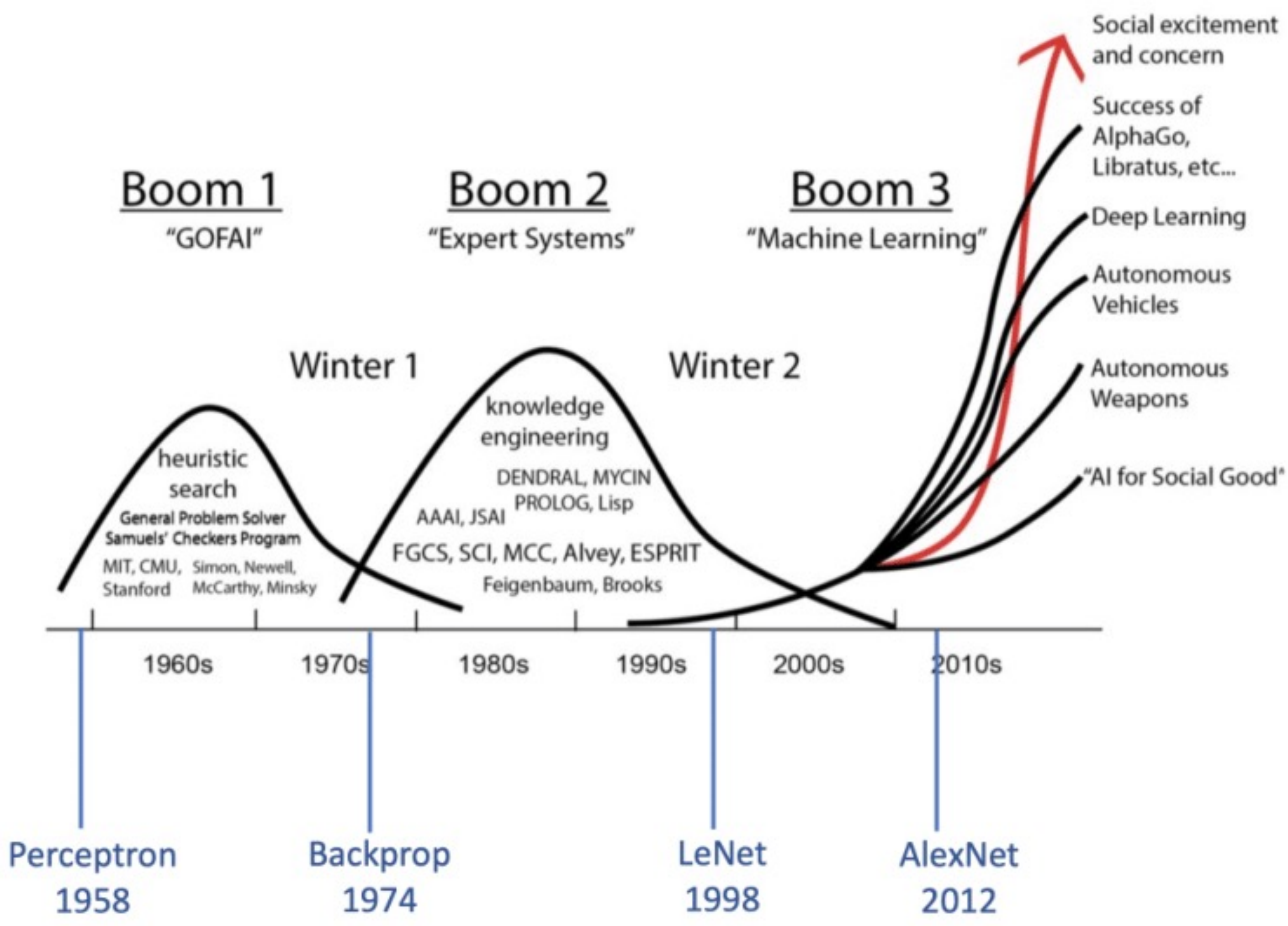
Relation to Statistics

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Relation to Statistics

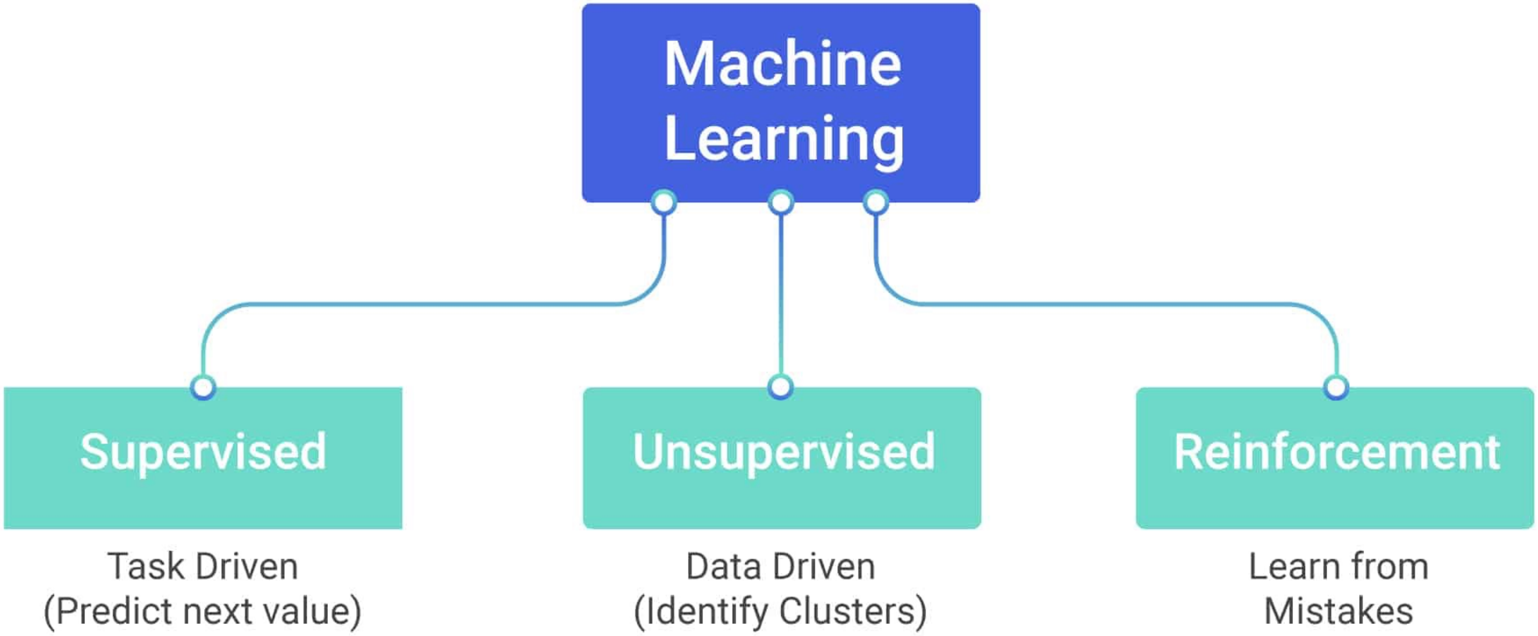


A brief history of Machine Learning



UHN Peter Munk Cardiac Centre

What is Machine Learning (ML)



Source: https://perfectial.com/blog/reinforcement-learning-applications/

How does AI/ML take over the world? Three Steps!

SUPERVISED LEARNING

UNSUPERVISED LEARNING

REINFORCEMENT LEARNING

A cake without the cherry is still a good cake!

How Much Information is the Machine Given during Learning?

Y. LeCun

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Why not jump straight to deep learning?

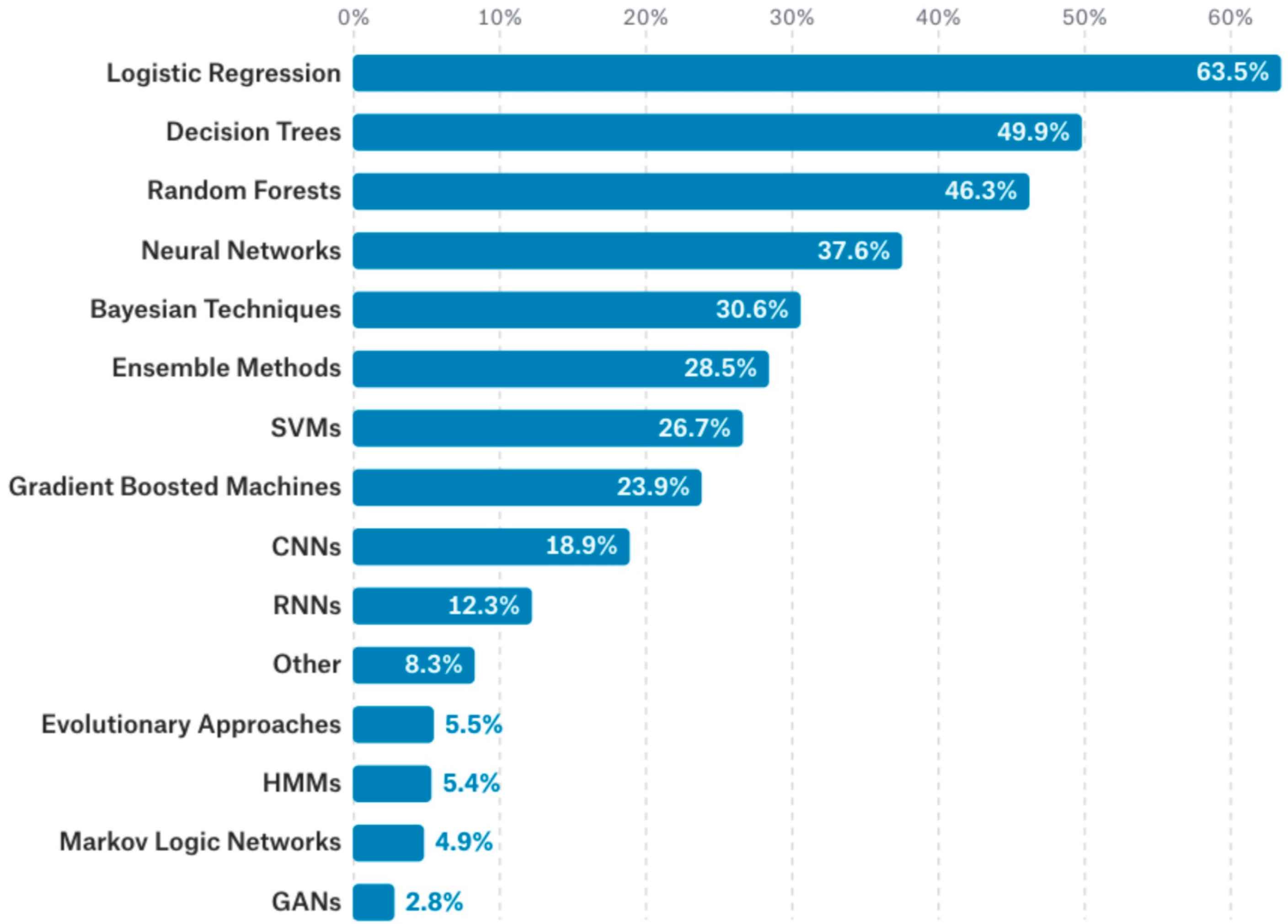
1. The principles you learn in this course will be essential to really understand deep learning.

2. The techniques in this course are still the first things to try for a new ML problem.

3. All models are wrong, but some are useful. --- George E. Box

Why not jump straight to deep learning?

2017 Kaggle survey of data science and ML practitioners: what data science methods do you use at work?



UHN

10-min Break

Next: Preliminaries and Nearest Neighbors

A typical ML workflow

ML workflow sketch:

1. Should I use ML on this problem?
   - Is there a pattern to detect?
   - Can I solve it analytically?
   - Do I have data?
2. Gather and organize data.
   - Preprocessing, cleaning, visualizing.
3. Establishing a baseline.
4. Choosing a model, loss, regularization, ...
5. Optimization (could be simple, could be a PhD!).
6. Hyperparameter search.
7. Analyze performance and mistakes, and iterate back to step 4 (or 2).

Supervised Learning---Basic Setup

We are going to focus on supervised learning for the next few lectures.

This means we are given a training set consisting of inputs and corresponding labels, e.g.

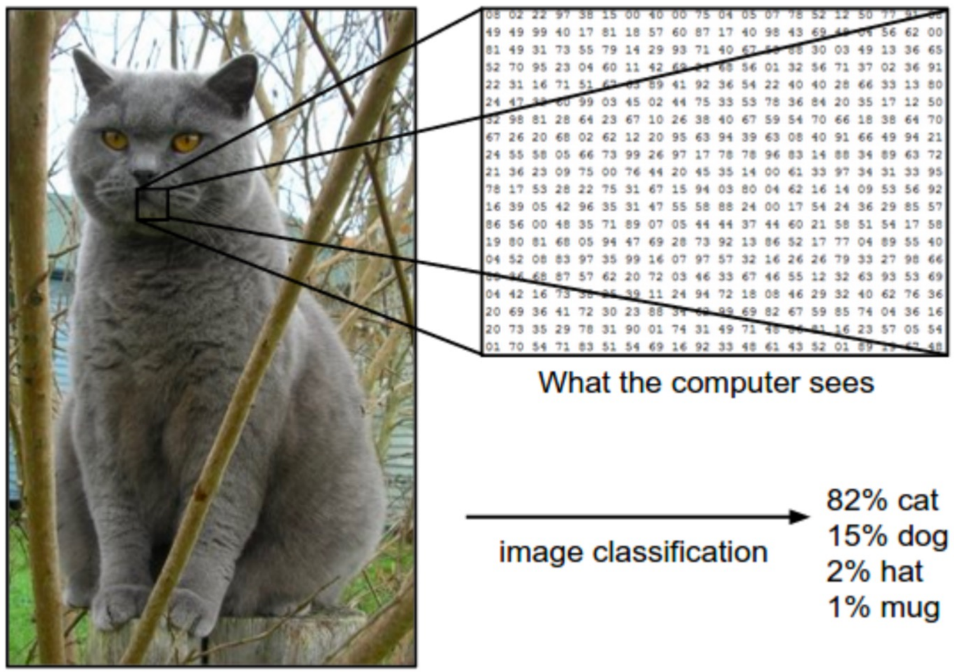
Task	Inputs	Labels
object recognition	image	object category
image captioning	image	caption
document classification	text	document category
speech-to-text	audio waveform	text
⋮	⋮	⋮

Input Vectors

- Machine learning algorithms need to handle lots of types of data: images, text, audio waveforms, credit card transactions, etc.
- Common strategy: represent the input as an input vector in $\mathbb{R}^d$
  - Representation = mapping to another space that is easy to manipulate
  - Vectors are a great representation since we can do linear algebra

Input Vectors

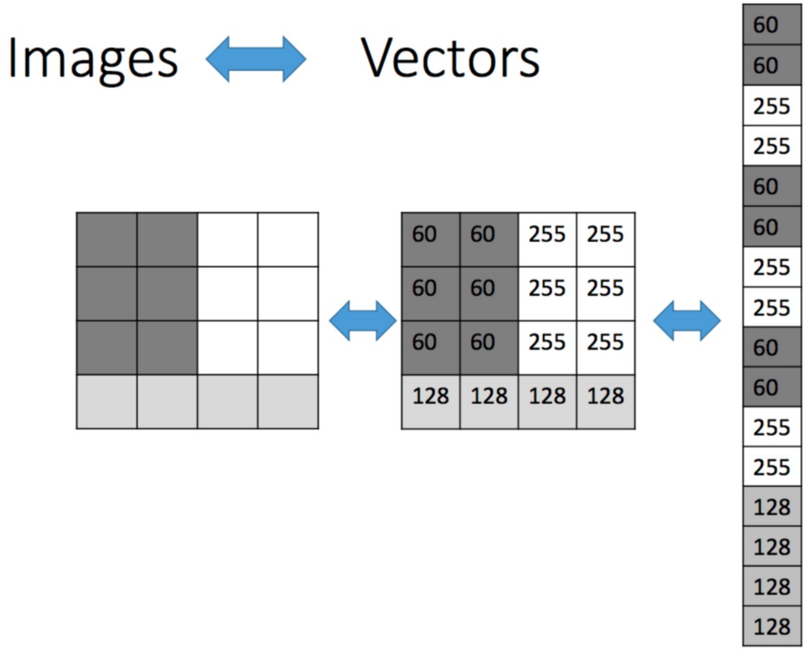
What an image looks like to the computer:



[Image credit: Andrej Karpathy]

Input Vectors

Can use raw pixels:



Can do much better if you compute a vector of meaningful features.

Input Vectors

- Mathematically, our training set consists of a collection of pairs of an input vector $\mathbf{x} \in \mathbb{R}^d$ and its corresponding target, or label, t
  - Regression: t is a real number (e.g. stock price)
  - Classification: t is an element of a discrete set $\{1, \ldots, C\}$
  - These days, t is often a highly structured object (e.g. image)
- Denote the training set $\{(\mathbf{x}^{(1)}, t^{(1)}), \ldots, (\mathbf{x}^{(N)}, t^{(N)})\}$
  - Note: these superscripts have nothing to do with exponentiation!

The very first supervised learning algorithm: **Nearest Neighbors**

- Suppose we're given a novel input vector $\mathbf{x}$ we'd like to classify.
- The idea: find the nearest input vector to $\mathbf{x}$ in the training set and copy its label.
- Can formalize "nearest" in terms of Euclidean distance

$$||\mathbf{x}^{(a)} - \mathbf{x}^{(b)}||_2 = \sqrt{\sum_{j=1}^{d} (x_j^{(a)} - x_j^{(b)})^2}$$

Algorithm:

1. Find example $(\mathbf{x}^*, t^*)$ (from the stored training set) closest to $\mathbf{x}$. That is:

$$\mathbf{x}^* = \underset{\mathbf{x}^{(i)} \in \text{train. set}}{\operatorname{argmin}} \ \text{distance}(\mathbf{x}^{(i)}, \mathbf{x})$$

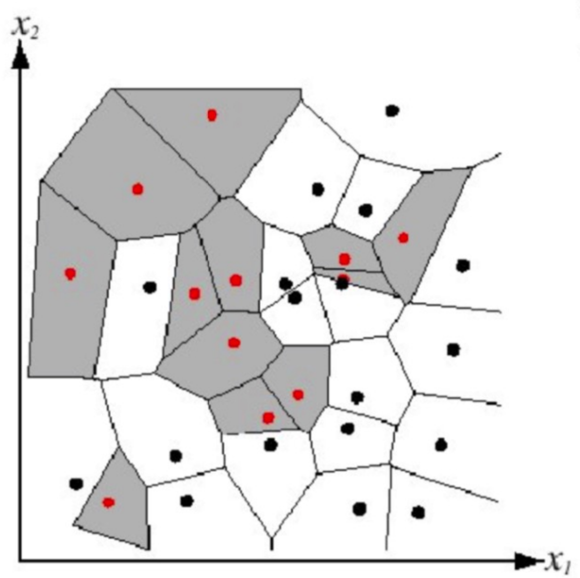
2. Output $y = t^*$

- Note: we do not need to compute the square root. Why?

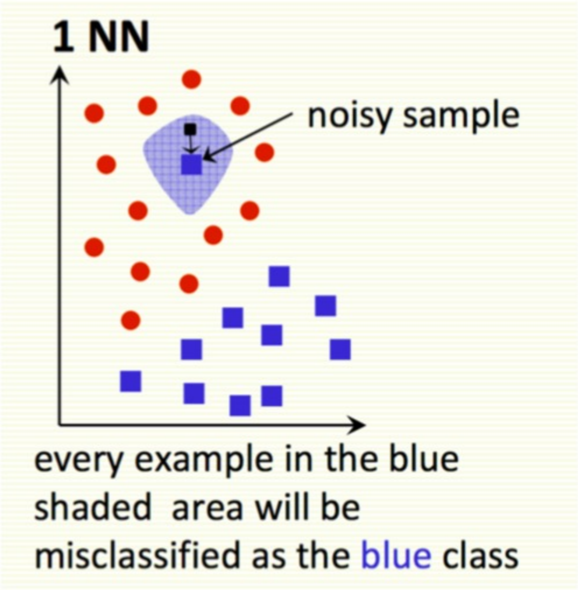
Nearest Neighbors: Decision Boundary

Decision boundary: the boundary between regions of input space assigned to different categories.

We can visualize the behaviour in the classification setting using a Voronoi diagram.



Nearest Neighbors: Pitfalls

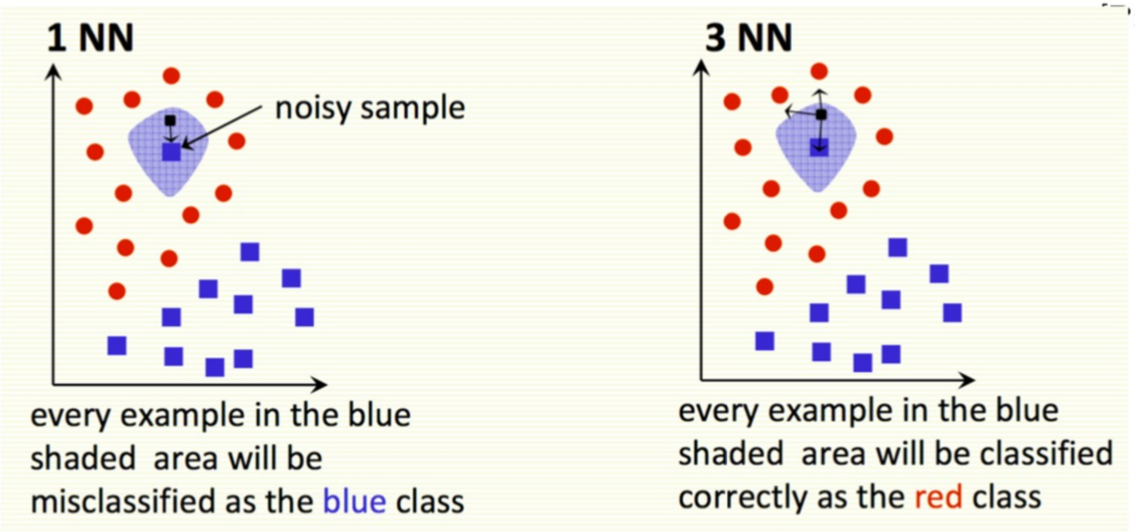


[Pic by Olga Veksler]

- Nearest neighbors sensitive to noise or mis-labeled data ("class noise"). Solution?

K - Nearest Neighbors (KNN)

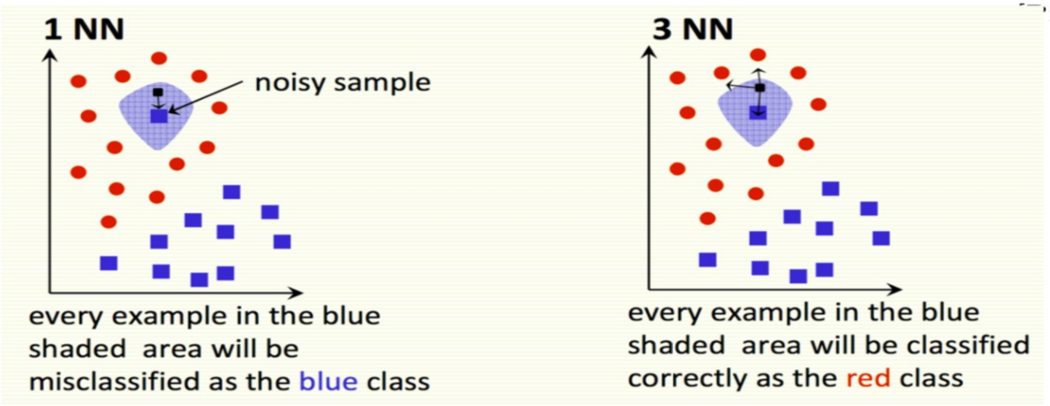
[Pic by Olga Veksler]



- Nearest neighbors sensitive to noise or mis-labeled data ("class noise"). Solution?
- Smooth by having k nearest neighbors vote

K - Nearest Neighbors (KNN)

[Pic by Olga Veksler]



- Nearest neighbors sensitive to noise or mis-labeled data ("class noise"). Solution?
- Smooth by having k nearest neighbors vote

Algorithm (kNN):

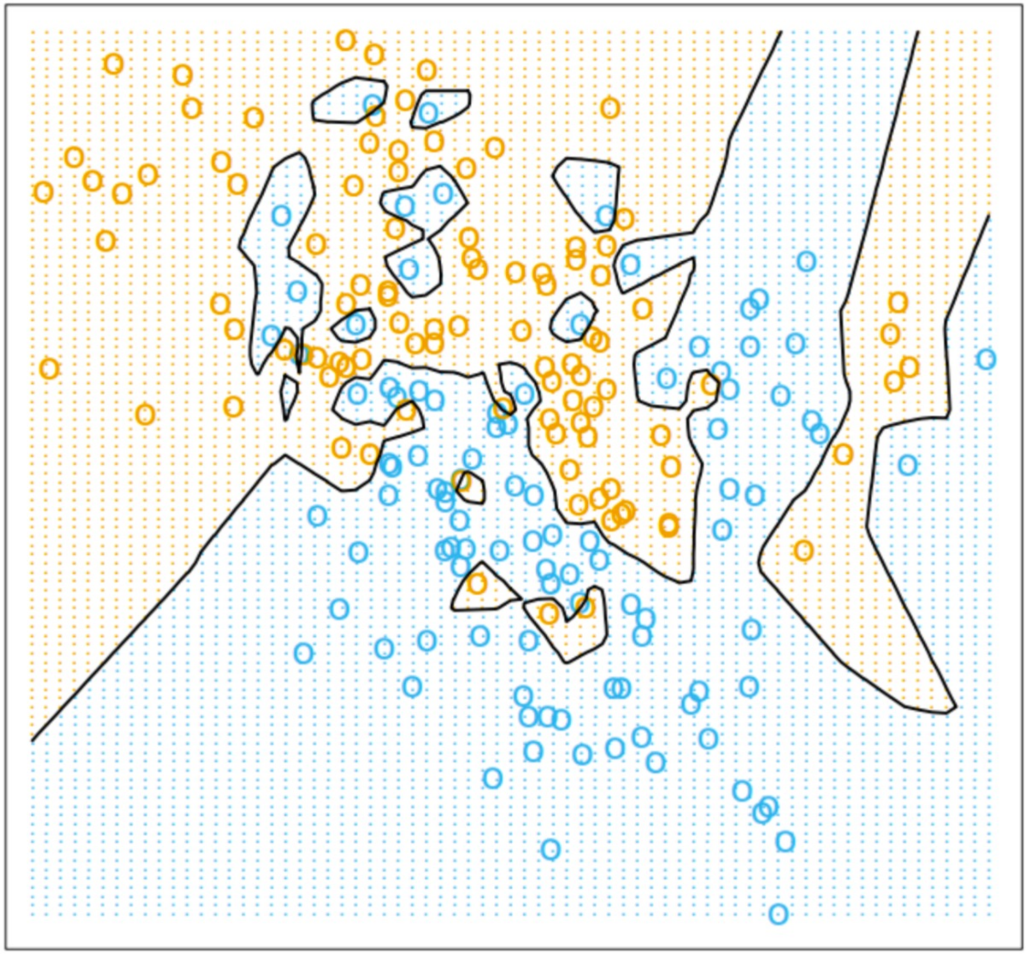
1. Find k examples $\{\mathbf{x}^{(i)}, t^{(i)}\}$ closest to the test instance $\mathbf{x}$
2. Classification output is majority class

$$y = \underset{t^{(z)}}{\operatorname{argmax}} \sum_{i=1}^{k} \mathbb{I}\{t^{(z)} = t^{(i)}\}$$

$\mathbb{I}\{\text{statement}\}$ is the identity function and is equal to one whenever the statement is true. We could also write this as $\delta(t^{(z)}, t^{(i)})$ with $\delta(a, b) = 1$ if $a = b$, 0 otherwise. $\mathbb{I}\{1\}$.

KNN Decision Boundaries

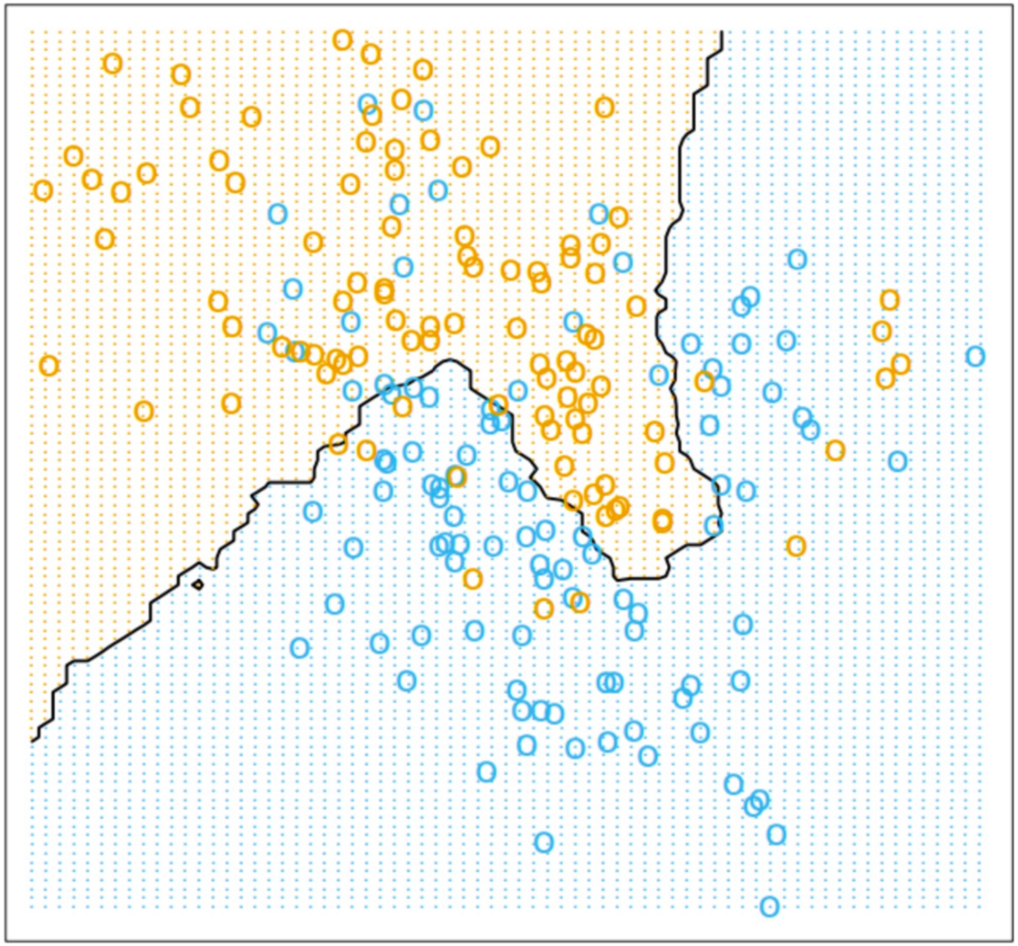
k=1



[Image credit: "The Elements of Statistical Learning"]

KNN Decision Boundaries

k=15



[Image credit: "The Elements of Statistical Learning"]

How to choose K?

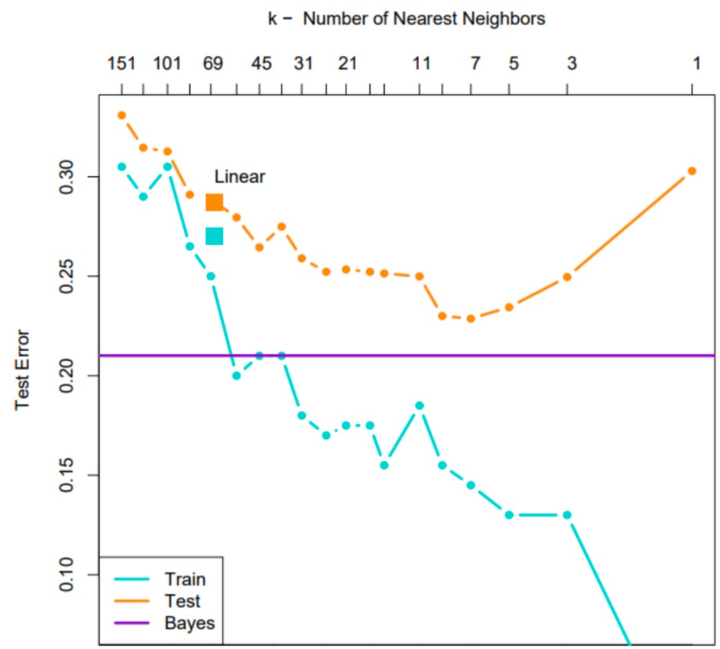
Tradeoffs in choosing k?

- Small k
  - Good at capturing fine-grained patterns
  - May overfit, i.e. be sensitive to random idiosyncrasies in the training data
- Large k
  - Makes stable predictions by averaging over lots of examples
  - May underfit, i.e. fail to capture important regularities
- Balancing k:
  - The optimal choice of k depends on the number of data points n.
  - Nice theoretical properties if $k \to \infty$ and $\frac{k}{n} \to 0$.
  - Rule of thumb: Choose $k = n^{\frac{2}{2+d}}$.
  - We explain an easier way to choose k using data.

UHN Peter Munk Cardiac Centre

How to choose K?

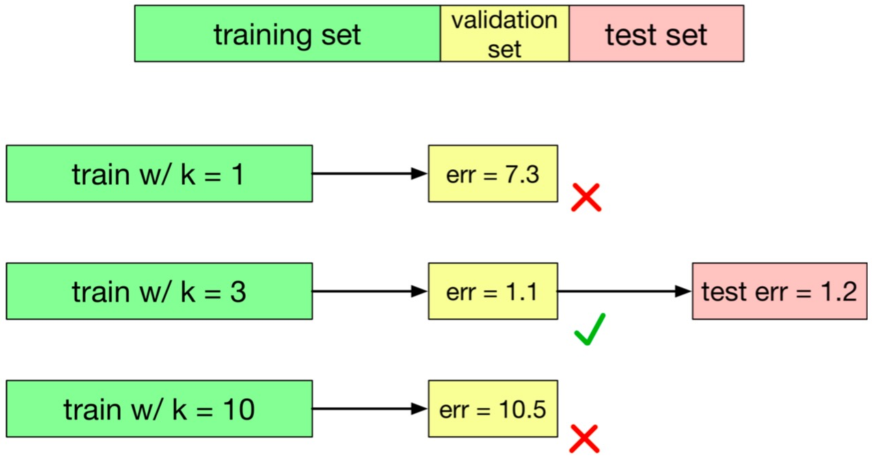
- We would like our algorithm to generalize to data it hasn't seen before.
- We can measure the generalization error (error rate on new examples) using a test set.



[Image credit: "The Elements of Statistical Learning"]

How to choose K?

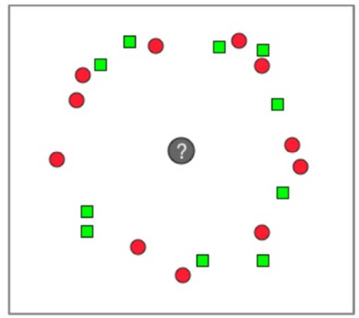
- k is an example of a hyperparameter, something we can't fit as part of the learning algorithm itself
- We can tune hyperparameters using a validation set:



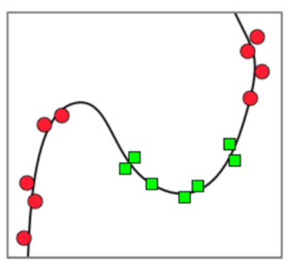
- The test set is used only at the very end, to measure the generalization performance of the final configuration.

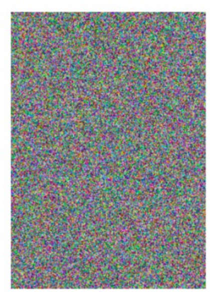
Pitfalls: Curse of Dimensionality

- In high dimensions, "most" points are approximately the same distance. (Homework question coming up...)



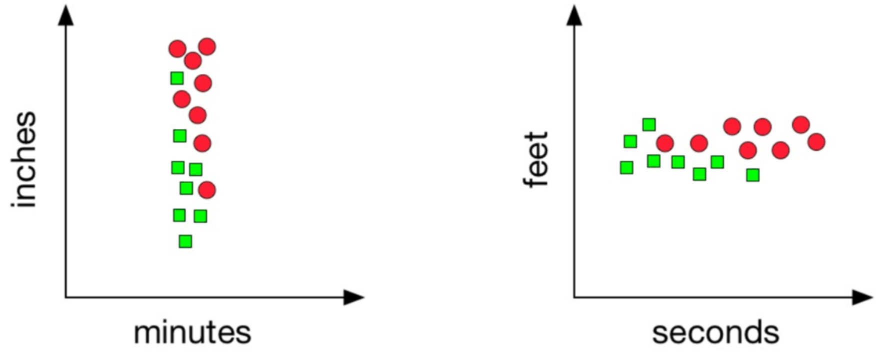
- Saving grace: some datasets (e.g. images) may have low intrinsic dimension, i.e. lie on or near a low-dimensional manifold. So nearest neighbors sometimes still works in high dimensions.





Pitfalls: Normalization

- Nearest neighbors can be sensitive to the ranges of different features.
- Often, the units are arbitrary:



- Simple fix: normalize each dimension to be zero mean and unit variance. I.e., compute the mean μ_j and standard deviation σ_j, and take

$$\tilde{x}_j = \frac{x_j - \mu_j}{\sigma_j}$$

- Caution: depending on the problem, the scale might be important!

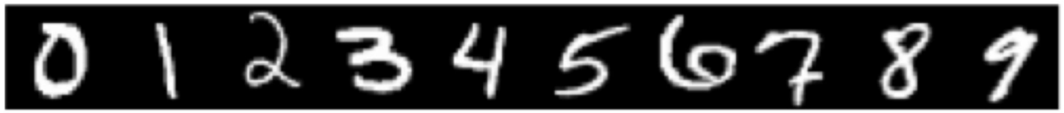
UHN Peter Munk Cardiac Centre

Pitfalls: Computational Cost

- Number of computations at training time: 0
- Number of computations at test time, per query (naïve algorithm)
  - Calculuate D-dimensional Euclidean distances with N data points: $\mathcal{O}(ND)$
  - Sort the distances: $\mathcal{O}(N \log N)$
- This must be done for *each* query, which is very expensive by the standards of a learning algorithm!
- Need to store the entire dataset in memory!
- Tons of work has gone into algorithms and data structures for efficient nearest neighbors with high dimensions and/or large datasets.

Pitfalls: Sensitive to similarity metrics

- Decent performance when lots of data

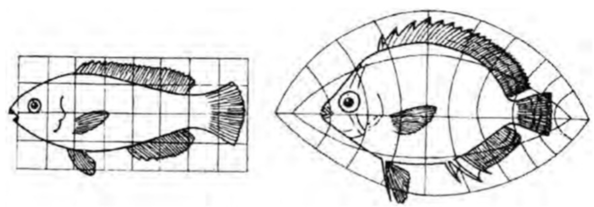


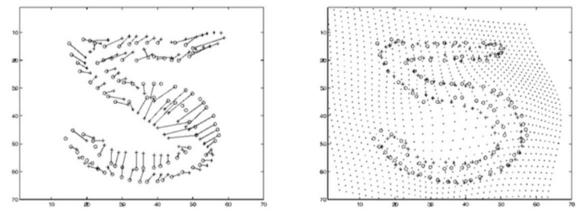
- Yann LeCunn – MNIST Digit Recognition
  - Handwritten digits
  - 28x28 pixel images: d = 784
  - 60,000 training samples
  - 10,000 test samples
- Nearest neighbour is competitive

	Test Error Rate (%)
Linear classifier (1-layer NN)	12.0
K-nearest-neighbors, Euclidean	5.0
K-nearest-neighbors, Euclidean, deskewed	2.4
K-NN, Tangent Distance, 16x16	1.1
K-NN, shape context matching	0.67
1000 RBF + linear classifier	3.6
SVM deg 4 polynomial	1.1
2-layer NN, 300 hidden units	4.7
2-layer NN, 300 HU, [deskewing]	1.6
LeNet-5, [distortions]	0.8
Boosted LeNet-4, [distortions]	0.7

Pitfalls: Sensitive to similarity metrics

- KNN can perform a lot better with a good similarity measure.
- Example: shape contexts for object recognition. In order to achieve invariance to image transformations, they tried to warp one image to match the other image.
  - Distance measure: average distance between corresponding points on *warped* images
- Achieved 0.63% error on MNIST, compared with 3% for Euclidean KNN.
- Competitive with conv nets at the time, but required careful engineering.





[Belongie, Malik, and Puzicha, 2002. Shape matching and object recognition using shape contexts.]

Conclusions

- Simple algorithm that does all its work at test time — in a sense, no learning!
- Can be used for regression too, which we encounter later.
- Can control the complexity by varying k
- Suffers from the Curse of Dimensionality
- Next time: decision trees, another approach to regression and classification

KNN in Healthcare

Digital Twin

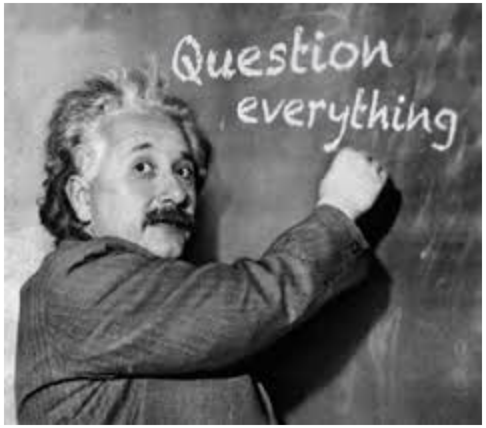
KNN in Healthcare

Predicting Cardiovascular Disease Using KNN

Source: https://towardsdatascience.com/predicting-cardiovascular-disease-using-k-nearest-neighbors-algorithm-614b0ecbf122

	Feature	Description
Categorical Data	age_days	| Factual Information | age in days | int (days)
	age_year	| Factual Information | age in years | int (days)
	height	| Factual Information | height | int (cm) |
	weight	| Factual Information | weight | float (kg) |
	ap_hi	| Systolic blood pressure | Examination Feature | int |
	ap_lo	|Diastolic blood pressure | Examination Featur | int |
Numerical Data	gender	| Factual Information |2:male, 1:female|
	cholesterol	Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
	gluc	Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
	smoke	Smoking |Subjective Feature | smoke | binary |
	alco	Alcohol intake | Subjective Feature | alco | binary |
	active	physical activity | Subjective Feature | active | binary |
	cardio	Presence or absence of cardiovascular disease | **Target Variable** | cardio | binary |
	id	| Factual Information |

UHN Peter Munk Cardiac Centre



UHN Peter Munk Cardiac Centre