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

Lecture 1 — Intro to ML in medicine, Nearest neighbours

Sana Tonekaboni

January 11, 2024

Teaching team

- **Instructors**
  - Sana Tonekaboni
  - Rahul Krishnan
- **TAs**
  - Fateme Pourghasem
  - Ahmadreza Attarpour

- Lectures: Thursdays 10:30-12:30
- Office hours: Wednesdays 9-10 am

Course overview

Lecture 1: Intro to ML in medicine, nearest neighbour classifier

Lecture 2: Tree-based classifier; Introduction to Python

Lecture 3: Linear methods for regression and classification; Evaluation methods

Lecture 4: ensemble-based methods; neural networks

Lecture 5: Supervised learning

Lecture 6: Unsupervised learning for clustering: K-means, Gaussian mixture models

Lecture 7: Unsupervised learning for clustering: auto-encoder, graph-based methods

Lecture 8,9: Guest lecturer

Lecture 10: Advanced deep learning methods for medical image analysis

Lecture 11,12: Final project presentations

Course deliverables

- [10%] Math diagnostics (Due next week!)
- [15%] Assignment 1 (Due Feb. 1)
- [15%] Assignment 2 (Due Feb. 15)
- [15%] Assignment 3 (Due March 1)
- [45%] Final project

Resources

- Lecture slides / Recommended reading
- Course webpage: https://sanatonek.github.io/lmp1210/
- Piazza: https://piazza.com/utoronto.ca/winter2024/lmp1210h
- Querqus

More on Assignments

1. Collaboration on the assignments is NOT allowed! Each student is responsible for their own work. Discussion of assignments should not involve any sharing of pseudocode or code or simulation results.
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.)
3. Chat-GPT or other language models can only be used as an aid! Read full policy on the course webpage.

Overview of today's lecture

- Administrative details
- Intro to Machine Learning in biomedical/healthcare research:
  - What is AI/ML? Why is it so popular now?
  - Importance of ML in healthcare and medicine
  - Examples of ML in healthcare
  - Challenges of ML for healthcare
- Break
- Supervised learning: Nearest neighbour

Why LMP1210?

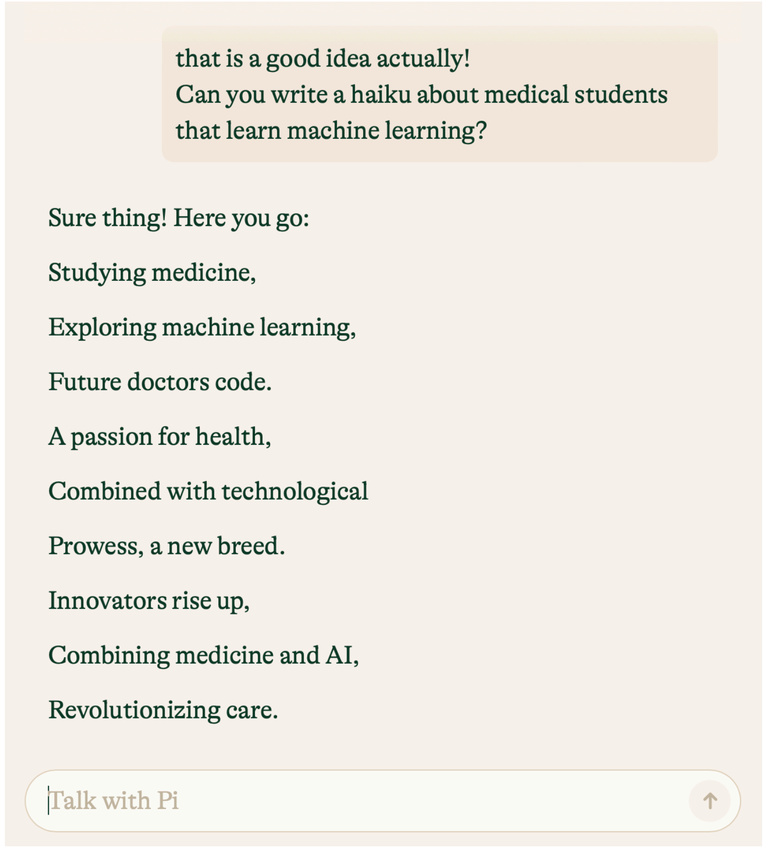
1. AI/ML is changing the way we perform research in medicine.
2. One of the first graduate courses in medical departments about machine learning!
3. We actually code!

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

What makes you excited about Machine Learning?

What is Artificial Intelligence (AI) and Machine Learning (ML)?

1. Autonomous driving
2. Language models
3. Generative photos
4. Personalized recommendations
5. Deep fake
6. So much more!

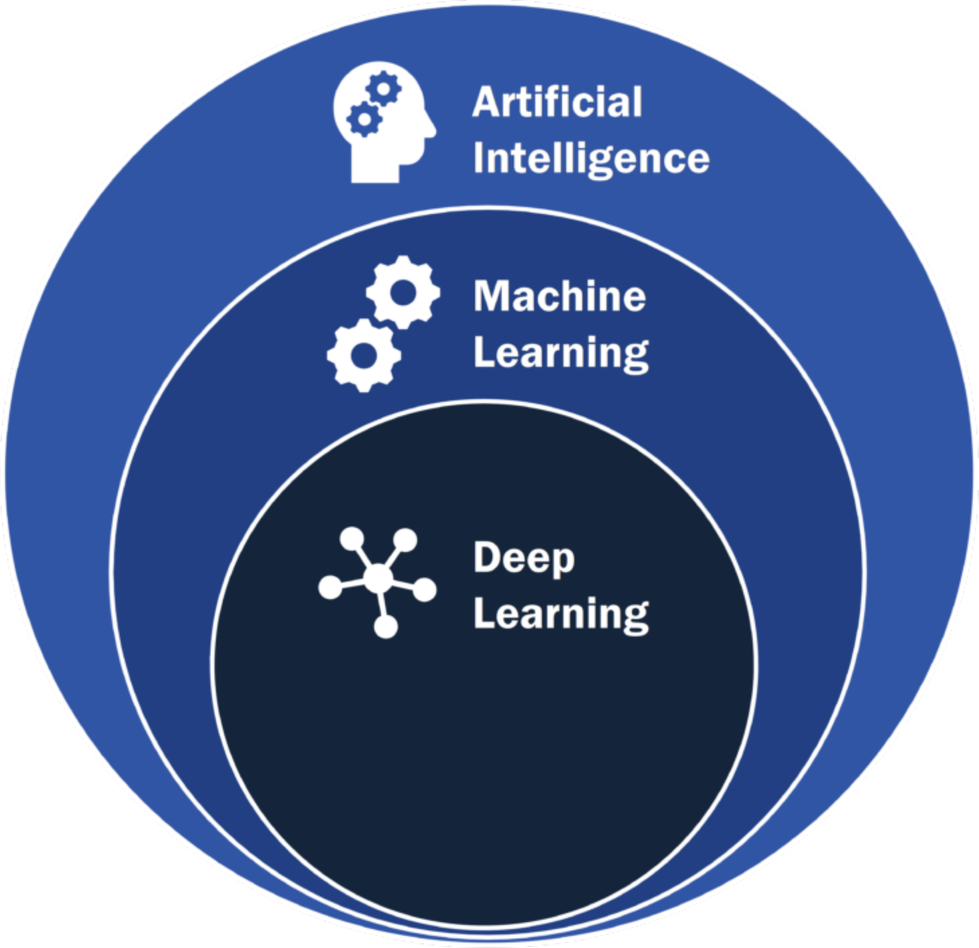


https://pi.ai/talk

https://legacy.midjourney.com/showcase/recent/

What is Artificial Intelligence (AI) and Machine Learning (ML)?

- **AI**: Any technique that enables computers to mimic human behaviour
- **ML**: Models with the ability to learn without explicitly being programmed
- **DL**: Machine learning based on neural networks



What makes AI so successful

Algorithms

Business applications

Data

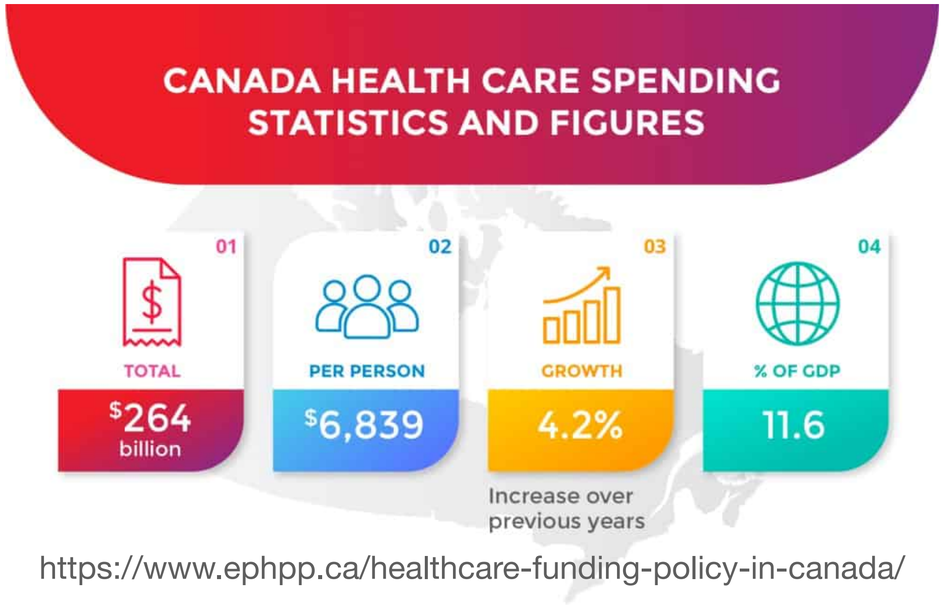
Compute

A-B-C-D

Why AI in healthcare is important?

- AI can take advantage of the increasing amount of digitized data
  - Improved diagnosis and treatment: diagnose diseases earlier and more accurately
  - Personalized medicine: tailored treatments to individual patients, based on their specific genetic and medical profiles.
  - Increased efficiency: automate routine tasks, freeing up healthcare professionals to focus on more complex tasks and patient care.

- Healthcare costs around the world are rising
- Expertise is limited and expensive
- Human inconsistencies (Bias and error)



https://www.ephpp.ca/healthcare-funding-policy-in-canada/

History of AI in healthcare

1978: Mycin expert system at Stanford

International Journal of Man-Machine Studies
Volume 10, Issue 3, May 1978, Pages 313-322

MYCIN: a knowledge-based consultation program for infectious disease diagnosis †

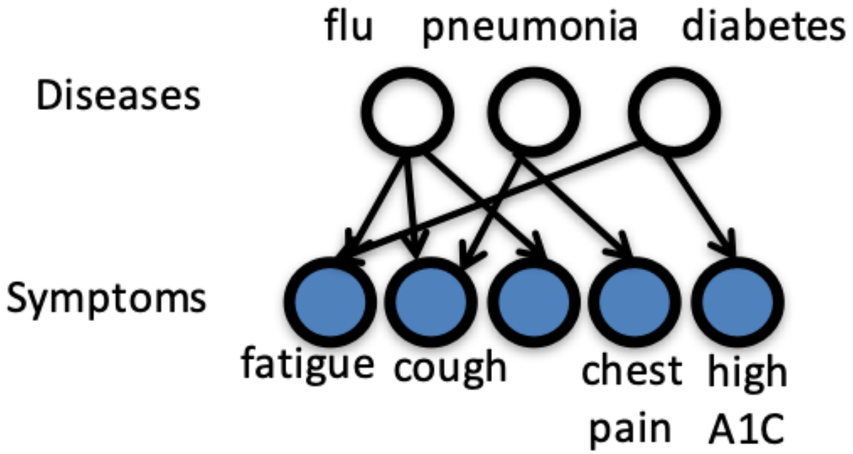
William van Melle

- Consultation system designed to assist physicians in the diagnosis and therapy selection for patients with bacterial infections based on symptoms and test results
- Used >500 prediction rules:
  - If A & B then predict pneumonia
- Worked better than specialists in blood infections

History of AI in healthcare

1986 : INTERNIST-1/QUICK MEDICAL REFERENCE (QMR) Project

- Automated diagnosis for general internal medicine
- Probabilistic model:
  - hundreds of disease variables,
  - thousands of symptom variables
  - >40000 directed edges between them



The creation of this model led to several advancements in probabilistic inference!

History of AI in healthcare

1990s: Neural networks in clinical medicine

- Used very few features to make predictions with
- Data collected by chart review

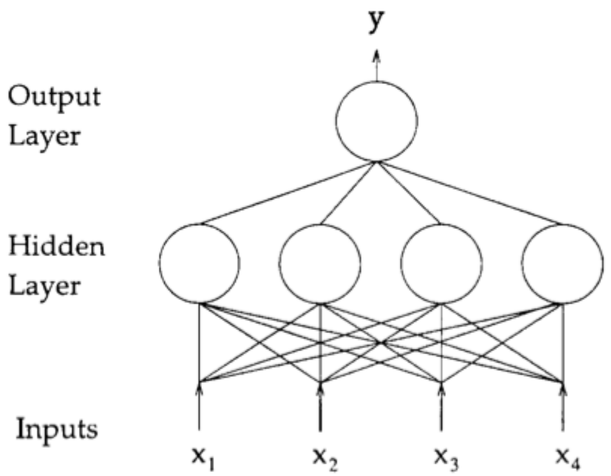


FIGURE 2. A multilayer perceptron. This is a two-layer perceptron with four inputs, four hidden units, and one output unit.

Did not generalize well to new places and difficult to fit into clinical workflow

Why AI in healthcare now?

So AI in healthcare is not new, but why is it so popular now?

- Before, models were not data driven, and mostly focused on domain knowledge.
- Large labeled open source data
- Advances in ML and large scale models!
  - GPT4/medical LLMs/self-supervised models of images/ text/sound/echocardiograms
- Localized compute power
- Digital health funding
- Industry interest

Why AI in healthcare now?

Better data

- Large datasets
  - AllofUs precision medicine initiative: deep phenotyping of 1 million people in the US
  - GEMINI dataset: Standardized clinical data from multiple hospitals in Ontario
  - MIMIC dataset: Critical care patient data
  - MedPix: Medical imaging dataset
  - Physionet
- Data standardization and digitization
  - FHIR, OHDSI

Why AI in healthcare now?

Advances in ML

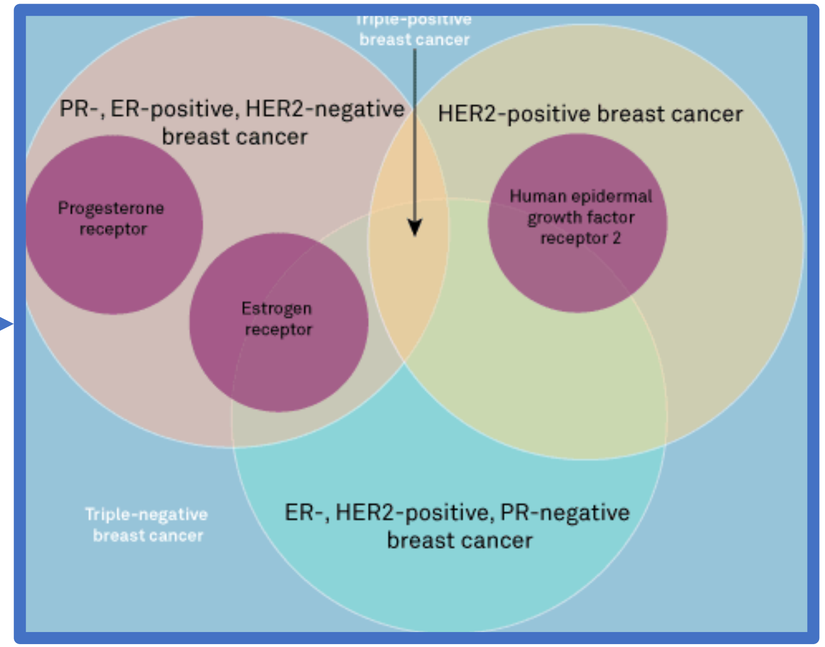
- 1990s – AI winter, but a productive one!
  - Markov Chain Monte Carlo
  - Variational Inference
  - Convolutional neural networks
  - Reinforcement learning
- 2000s – Vision and NLP started adopting ML models
- 2013: Imagenet – watershed moment for deep learning
- 2018-now:
  - Photorealistic GANs
  - GPT-3 can simulate text indistinguishable from text written by humans
  - Midjourney can create synthetic looking videos

What can we do with the data?

- For the right tasks, we could use the data to train ML models.
- **General recipe:**
  - Identify a problem, which if automated, can reduce the cost of a process or help clinicians complete a task better/faster/ with less error.
  - Program a model to automatically learn patterns from data
  - Use the model to automate task

There are different kinds of machine learning strategies we can make use of

Subtype discovery



Build clinical tools

ML strategies

Supervised learning

Patient features

Chest X-ray image

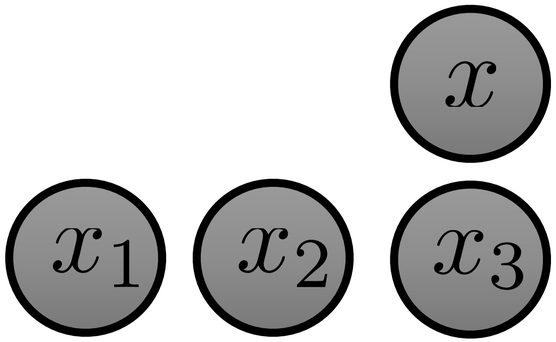
Time to readmission

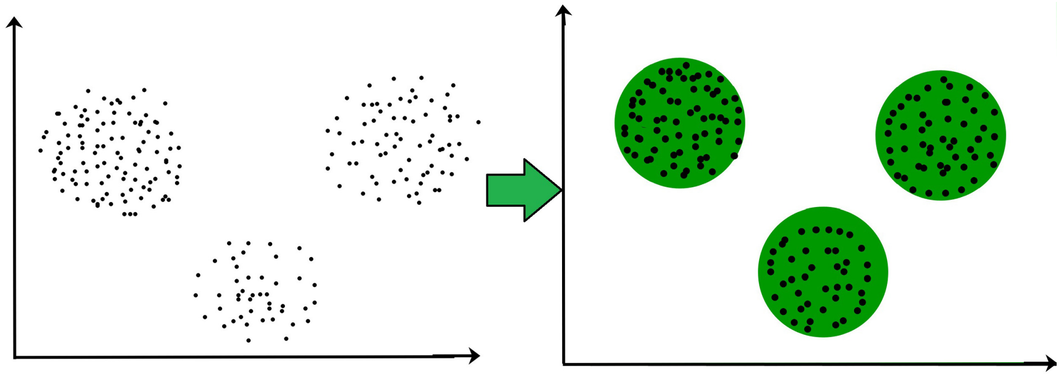
pneumonia/pneumothorax

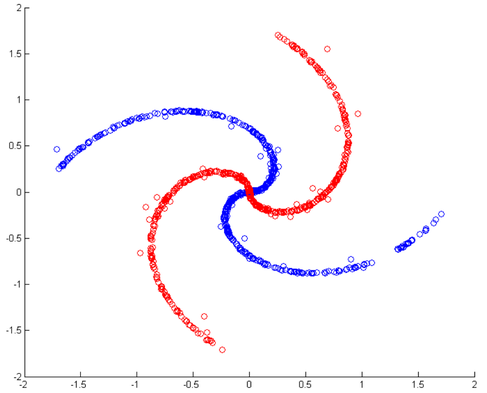
- Use labeled data to train the ML model (task-driven)
- **Examples**: Logistic regression, random forests, XGBoost, Deep neural networks

ML strategies

Unsupervised learning



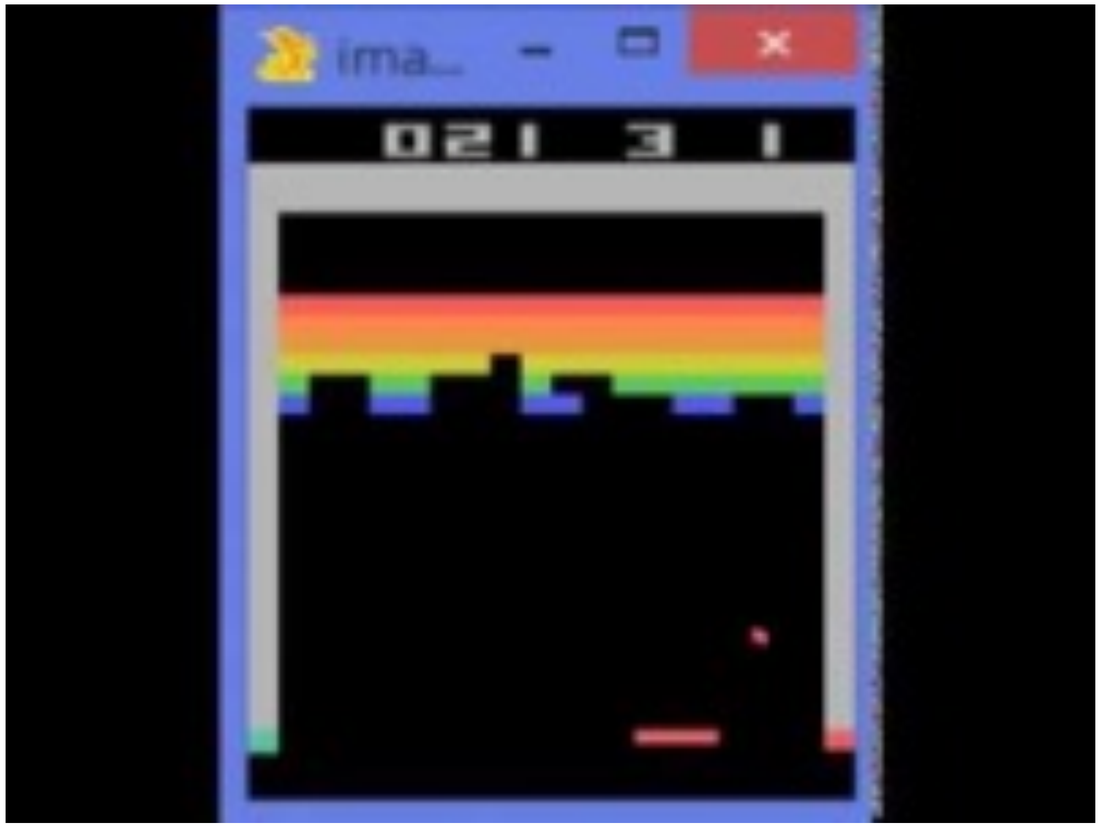


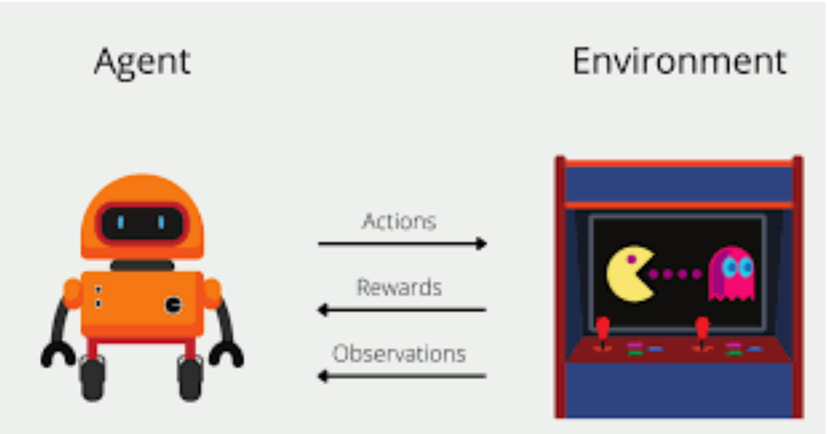


- Uncover insights about the data and validate with domain experts (data-driven)
- **Examples**: Nearest neighbours, latent factor models, hidden Markov models, variational autoencoders

ML strategies

Reinforcement Learning (RL)





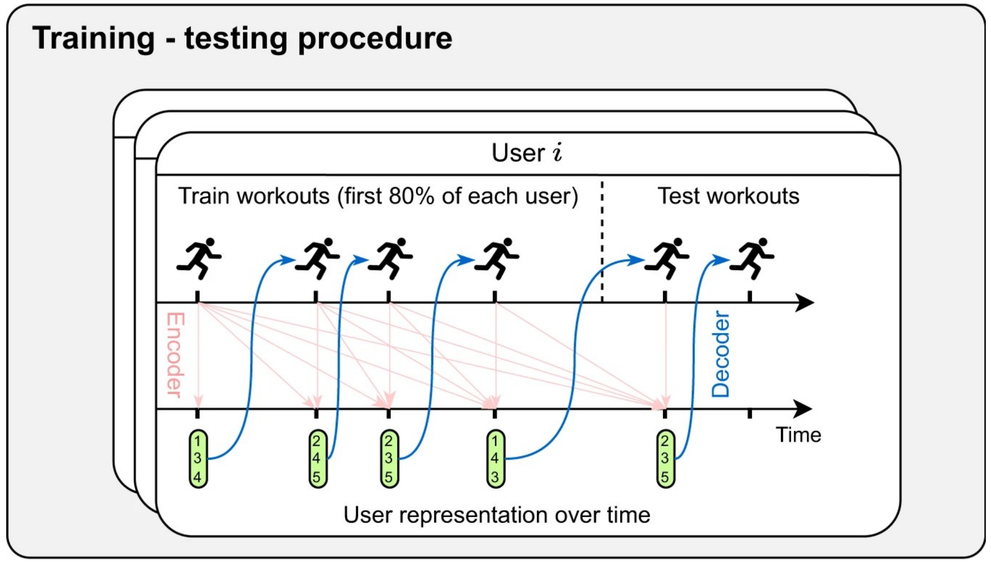
- On the left is an example of Deepmind's ATARI RL agent that learns to move the paddle at the bottom
- Can we use similar techniques for problems in healthcare such as developing strategies to treat people?

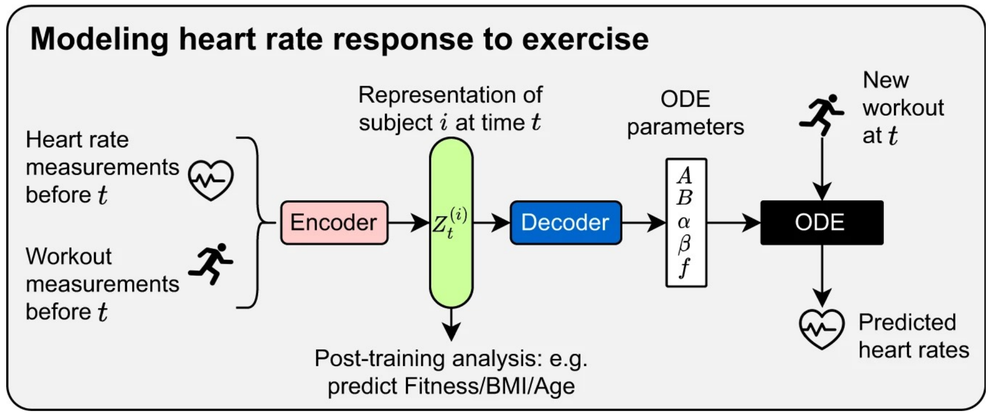
Challenge: Difficult to build good simulators of how the human body will react to drugs

Examples of ML application in healthcare

Modelling heart rate response to exercise

Using machine learning to model personalized cardiovascular response to exercise

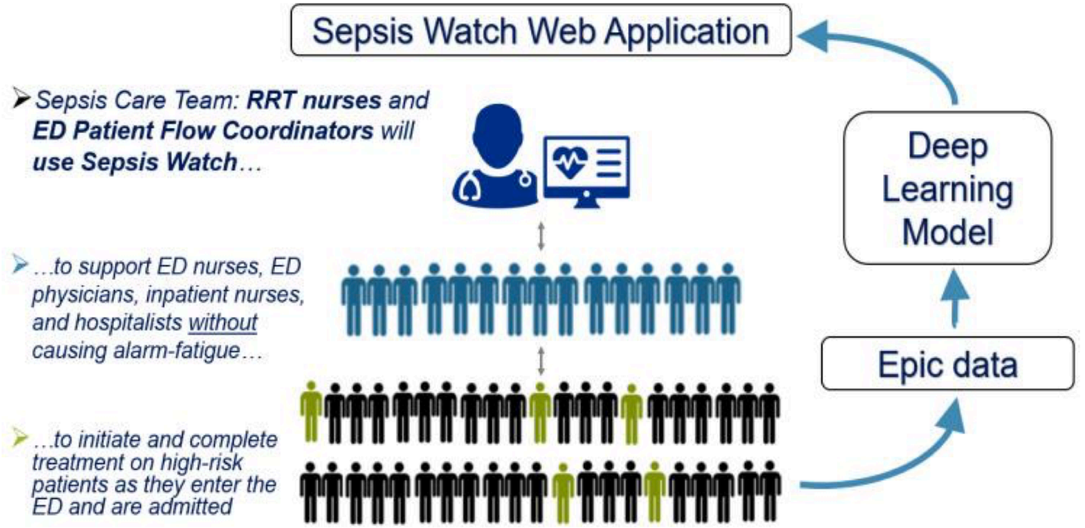




https://www.nature.com/articles/s41746-023-00926-4

Sepsis Watch

- **42,000+ inpatient encounters analyzed** at Duke Hospital over 14 months, **21.3%** with a sepsis event.
- **32+ million data points incorporated**: 25 million vital sign measurements, 2 million med admins, 5.2 million labs.
- **34** physiological variables (5 vitals, 29 labs).
  - At least one value for each vital in 99% of encounters.
  - Some labs rarely measured (2-4%), most measured 20-80% of the time.
- **35** baseline covariates (e.g. age, transfer status, comorbidities).
- **10** medication classes (antibiotics, opioids, heparins).

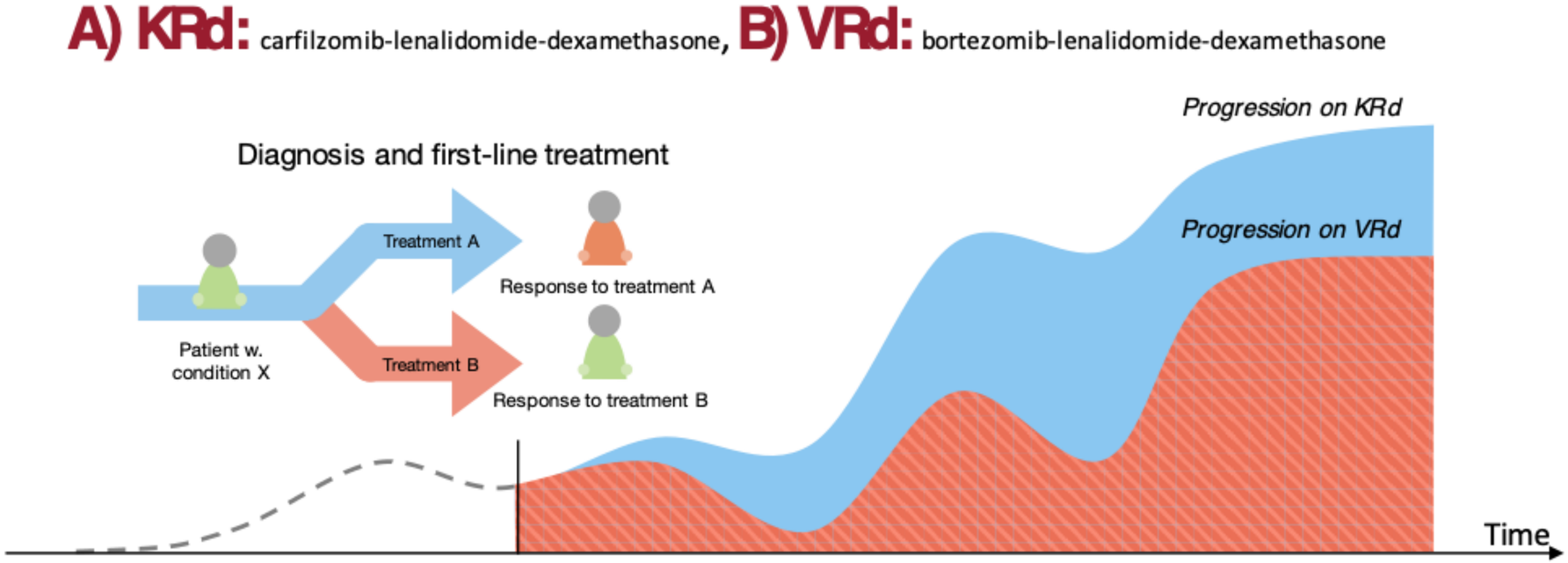


Source: https://dihi.org/wp-content/uploads/2020/02/Sepsis-Watch-One-Pager.pdf

Using machine learning to predict risk of developing Sepsis in patients admitted to the ICU

Precision oncology

Using machine learning to guide treatment decisions for cancers therapy



And many more applications…

- Drug discovery for faster, cheaper drug development pipelines
- Automating polyp detection in gasterointestinal diseases

- New and upcoming places for machine learning to have an impact in healthcare:
  - Microbiome
  - Liquid biopsies for cancer detection and tracking
  - Improving documentation burden for clinicians

So why ML isn’t everywhere in healthcare now?

Challenges for machine learning in healthcare

- Challenging risk/reward ratios
  - **Why**: In healthcare, clinicians make life or death decisions
  - What do we need:
    - Algorithm development should proceed with caution and care
    - Need **robust** algorithms with checks and balances
    - Algorithms need to be **fair** and **accountable**
- Labelled data is scarce
  - **Why**: Clinician time is expensive
  - Not all solutions are necessary, need to talk to stakeholders to find the ones that are worth solving
  - They may not be the problems you want to solve!

Challenges for machine learning in healthcare

- Patient populations are different:
  - **Why**: Everyone is unique and people from Mumbai display different clinical phenotypes than those in Toronto
  - What do we need:
    - New methods for transfer learning so that models generalize well across different hospitals

- Missingness
  - **Why:** We only go to the doctor/clinician/hospital when we are sick; hospital administrators may forget to annotate data, records can go missing
  - What do we need:
    - Machine learning models that can make robust predictions even when data is missing

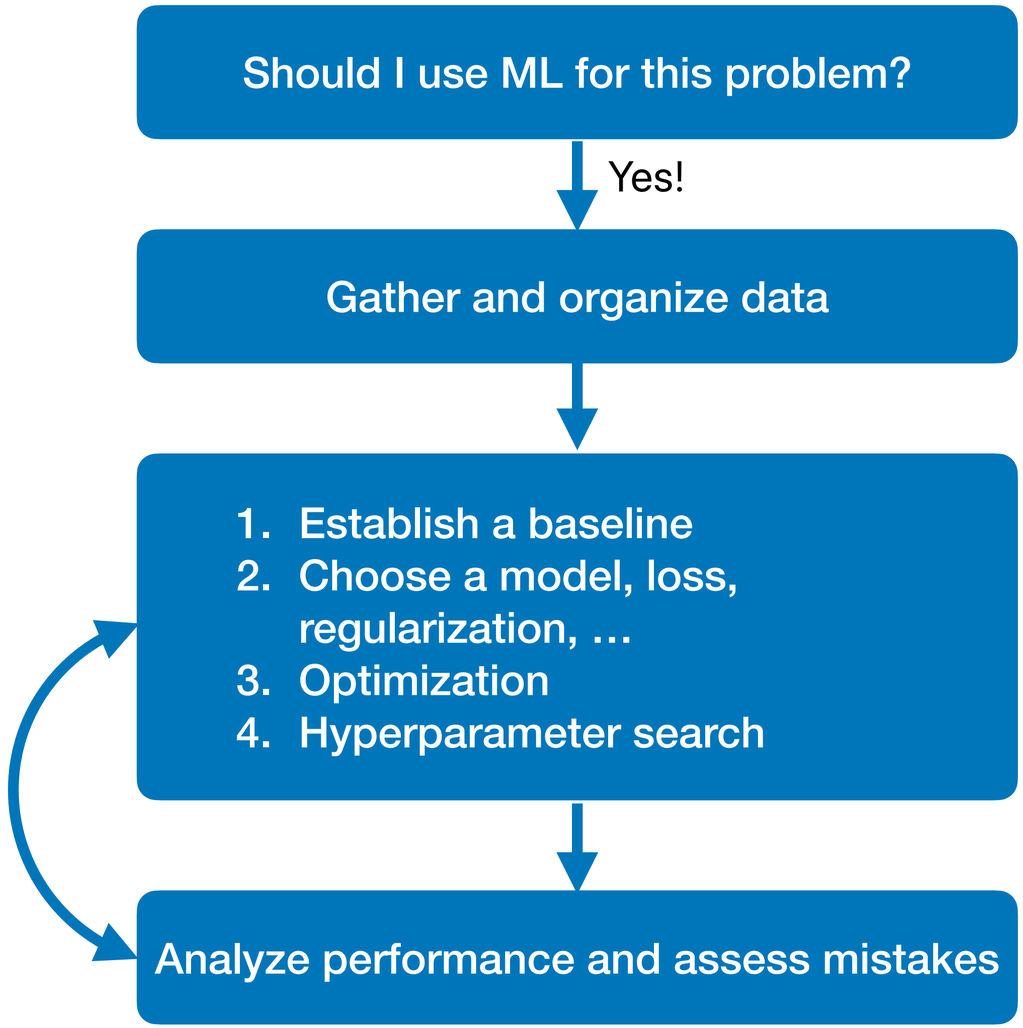
Challenges for machine learning in healthcare

- Data silos
  - **Why:** Countries have regulations (such as HIPAA) that require patient data to be kept private
  - What do we need:
    - New ideas in federated learning for institutions not comfortable with data-sharing
    - Automated methods for de-identification
- Deploying ML software in the clinic
  - **Why:** Machine learning models can stop working after a period of time
  - What do we need:
    - New techniques for lifelong learning
    - Ways to handle domain shift/covariate shift

Break

10 minutes

A typical ML Workflow



- Is there a pattern in the data?
- Can I solve it analytically?
- Can I gather proper data for it?

- Collect data
- Processing and cleaning
- Visualizing and understanding

Supervised learning

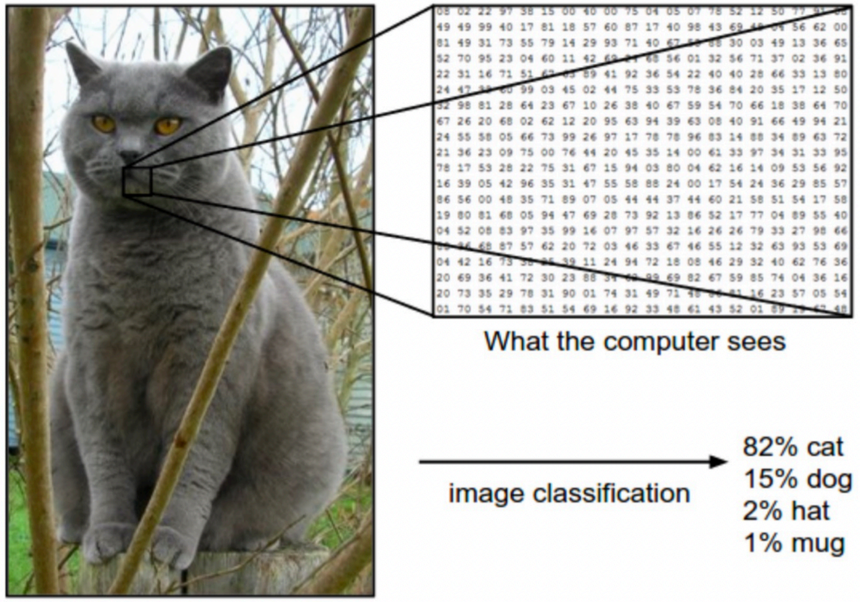
The next few lectures, we will focus on Supervised learning method, which means we will be given a training set consisting of **inputs** and **labels**.

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

Supervised learning

Input vectors

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



Supervised learning
training data

- Mathematically, our **training set** consists of a collection of pairs of an input vector $x \in \mathbf{R}^d$ 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, ...,C}
  - These days, t is can be a highly structured object
- Denote the training set:
  $\{(x^{(1)}, t^{(1)}), (x^{(2)}, t^{(2)}), \ldots, (x^{(N)}, t^{(N)})\}$

$x^{(1)}$ $t^{(1)} = \text{Muffin}$

$x^{(2)}$ $t^{(2)} = \text{Chihuahua}$

$x^{(3)}$ $t^{(3)} = \text{Muffin}$

$x^{(4)}$ $t^{(4)} = \text{Chihuahua}$

Nearest Neighbour

The very first supervised learning algorithm

- Suppose we are given a novel input vector x we'd like to classify.
- The idea: find the nearest input vector to 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:

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

2. Output $y = t^*$

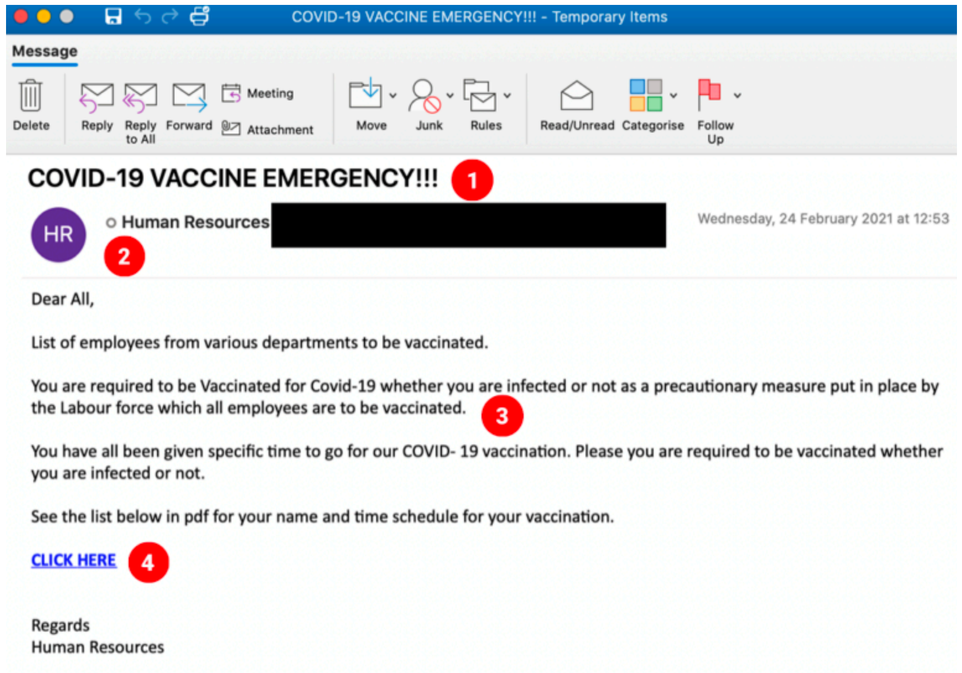
Note: We don't need to compute the square root. Why?

Nearest Neighbour

Example: Email spam detection

- Input variable: Variables describing the email. E.g. Number of capital letters in the message, number of symbols and punctuations, existence of an external link, ...
- Target variable: Spam or not

Test sample:

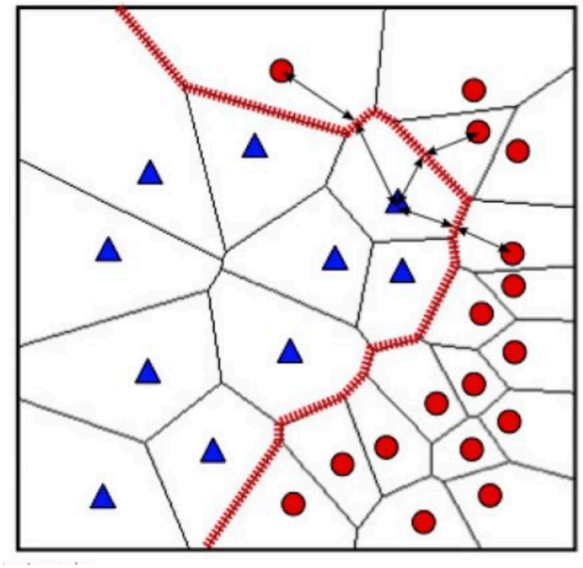


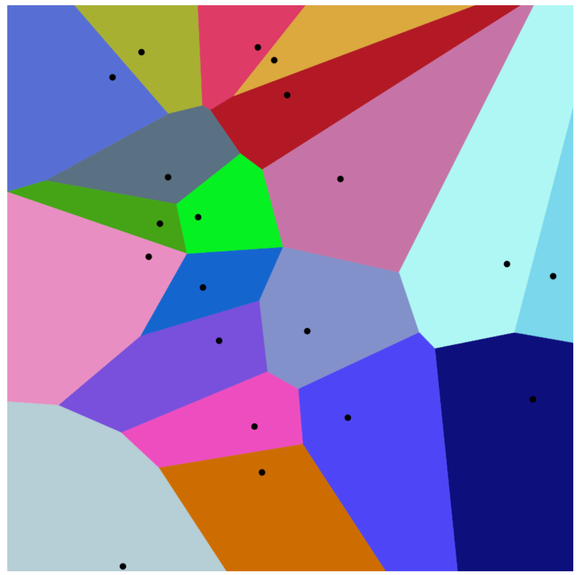
Courtesy of https://scholarlyoa.com/

Nearest Neighbour

Decision boundary

- Decision boundary: The boundary between regions of inputs space assigned to different categories.



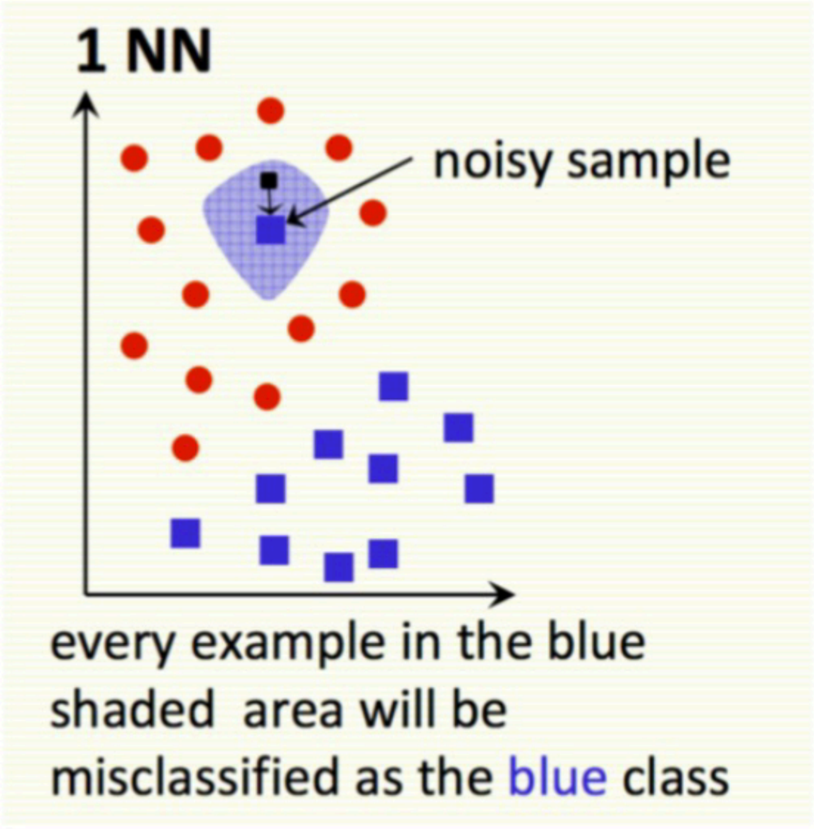


- We can visualize the behaviour in the classification setting using a **Voronoi diagram**

Nearest Neighbour

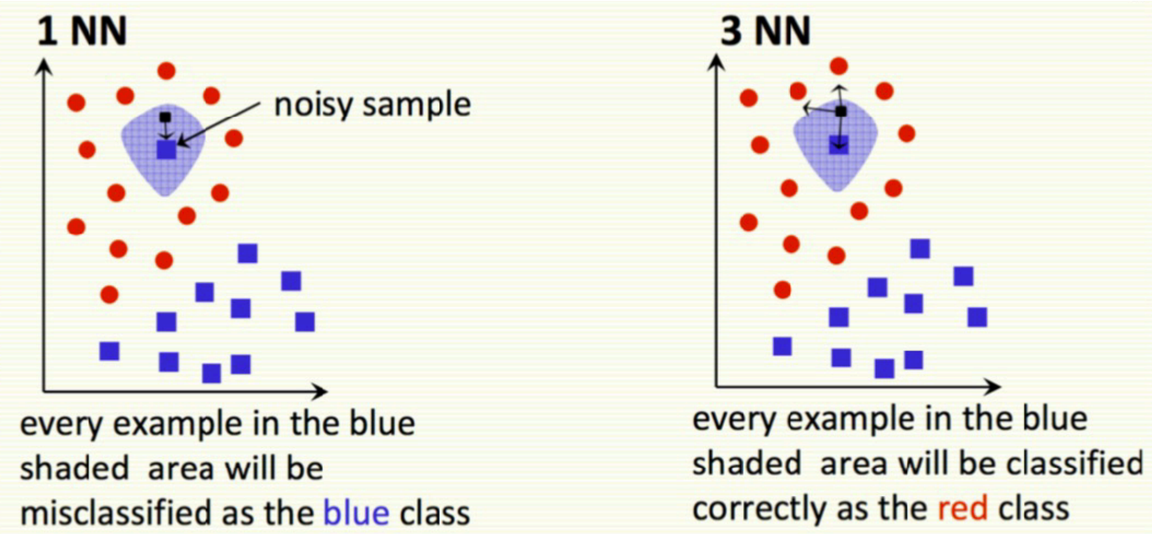
Pitfalls

- Nearest neighbours is sensitive to noise or mis-labeled data. Solution?



K-Nearest Neighbour (KNN)

- Smooth decision boundary by having k nearest neighbour vote



K-Nearest Neighbour (KNN)

- Smooth decision boundary by having k nearest neighbour 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

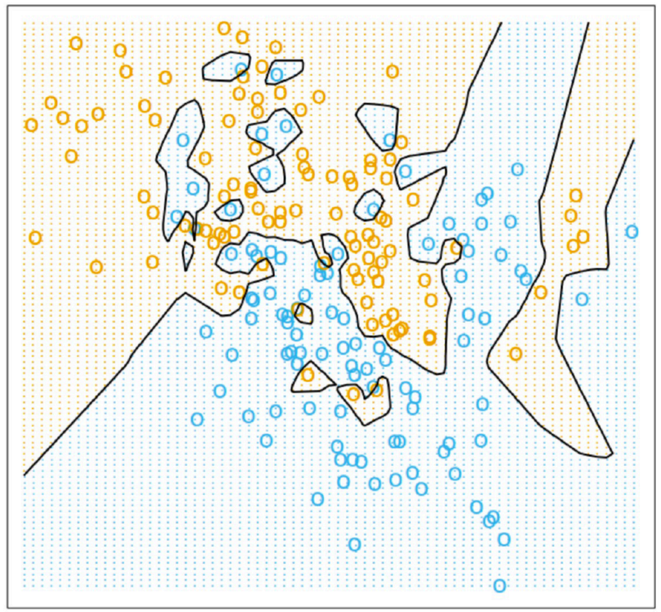
$$y = \underset{t^{(z)}}{\operatorname{argmax}} \sum_{i=1}^{k} \mathbb{I}\{t^{(z)} = t^{(i)}\}$$

Identity function, also shown as $\delta(t^{(z)}, t^{(i)})$

K-Nearest Neighbour (KNN)

Decision boundaries

K = 1



K = 15

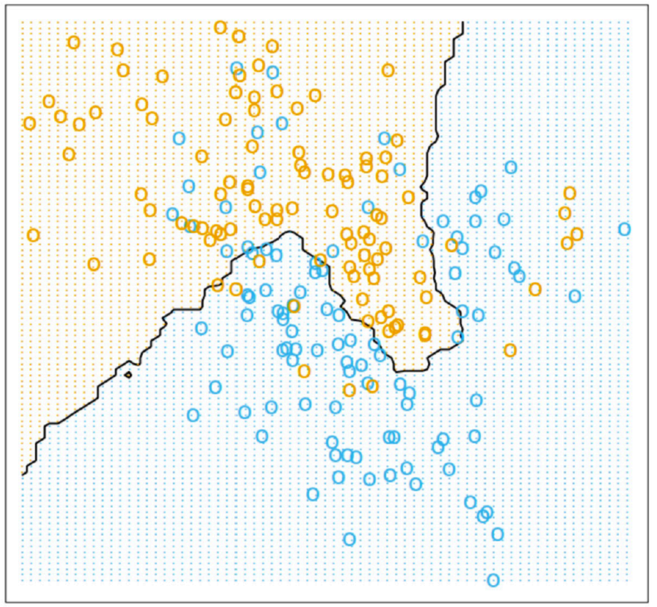


Image credit: The elements of statistical learning

K-Nearest Neighbour (KNN)

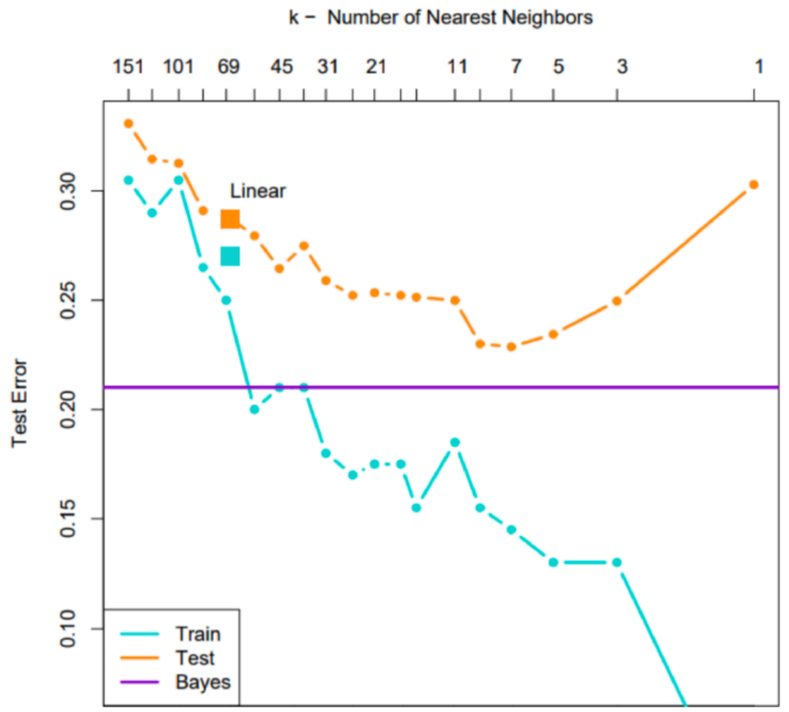
How to choose K?

- 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.
  - Rule of thumb: Choose $k = n^{\frac{2}{2+d}}$.
  - We explain an easier way to choose k using data.

K-Nearest Neighbour (KNN)

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 data) using a test set

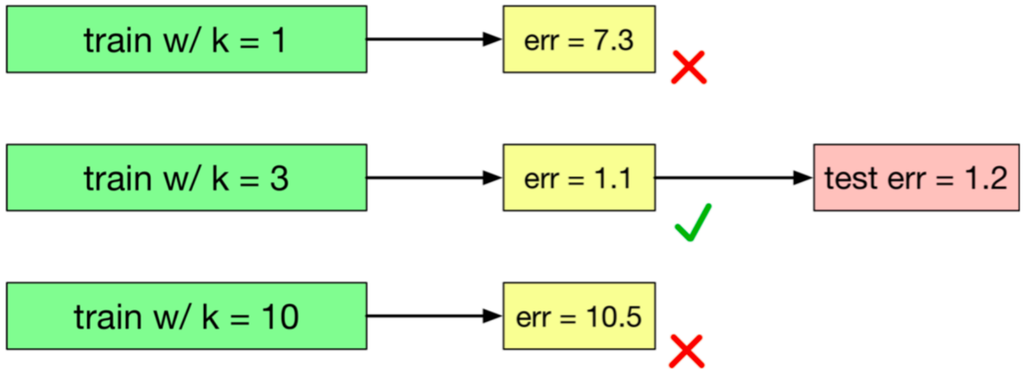


K-Nearest Neighbour (KNN)

How to choose K?

- k is an example of a **hyper-parameter**, something we can't fit as part of the learning algorithm.

- We can tune hyper parameters using a **validation set**.

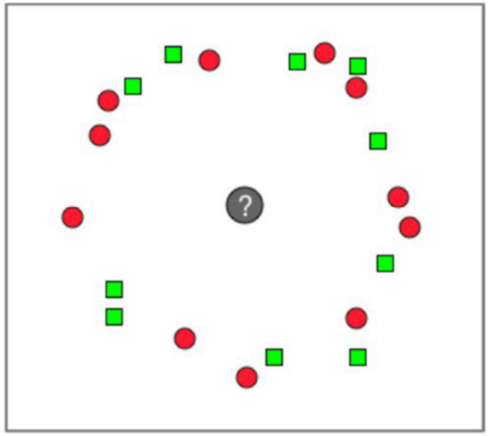


- The test set is used at the very end, to use the generalization performance of the final configuration.

KNN Pitfalls

Curse of dimensionality

- In high dimensions, most points are approximately the same distance. Why?



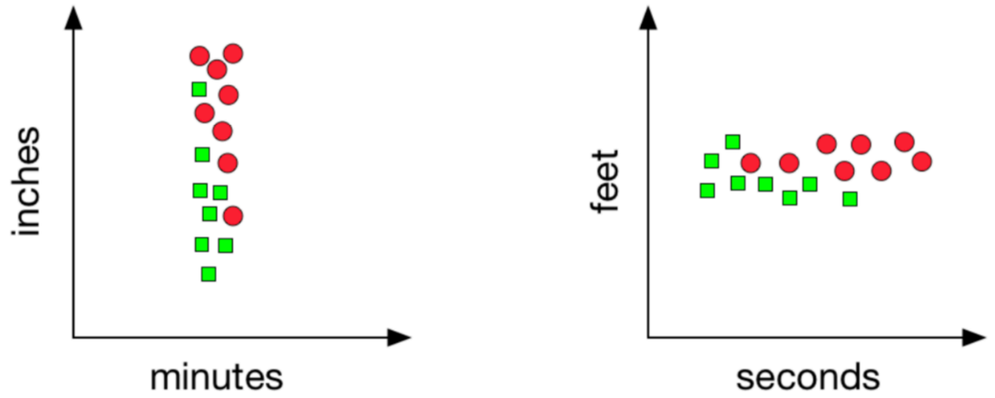
- With increasing dimensions, the amount of data required to represent the space effectively also increases. This means the amount of data required to produce reliable results can become prohibitively large.

Possible solutions: feature selection, dimensionality reduction, and data preprocessing

KNN Pitfalls

Normalization

- Nearest neighbour can be sensitive to a range of different features. Often the units are arbitrary. (Can you think of a healthcare example?)



- Simple fix: Normalize each dimension to be mean zero and anti 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.

KNN Pitfalls

Computation cost

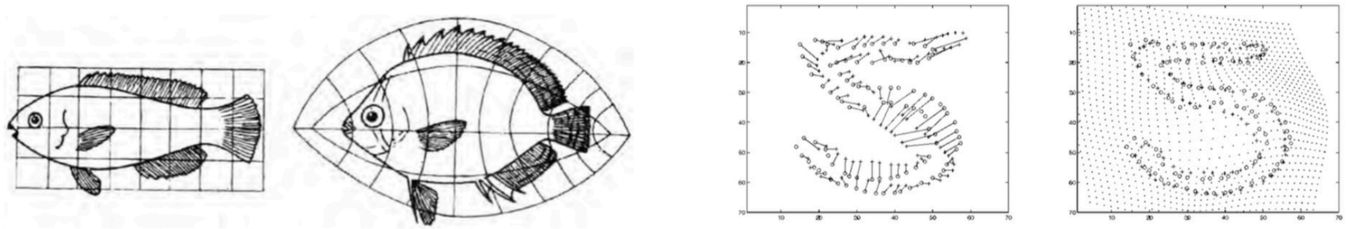
- Number of computations at training time: 0
- Number of computations at test time, per query (naive algorithm)
  - Calculate D-dimensional Euclidean distances with N data points: $\mathcal{O}(ND)$
  - Sort the distances: $\mathcal{O}(N \log N)$
- This must be done for each query, which is very expensive 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 neighbours with high dimensions and/or large datasets.

KNN 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.]

KNN in healthcare

AMIA Annual Symposium Proceedings Archive

Proc AMIA Symp. 2000 : 759–763.

PMCID: PMC2243774
PMID: 11079986

Application of K-nearest neighbors algorithm on breast cancer diagnosis problem.

M. Sarkar and T. Y. Leong

ELSEVIER

Procedia Technology
Volume 10, 2013, Pages 85-94

Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm ☆

M. Akhil jabbar [a], B.L. Deekshatulu [b], Priti Chandra [c]

ELSEVIER

Expert Systems with Applications
Volume 40, Issue 1, January 2013, Pages 263-271

An efficient diagnosis system for detection of Parkinson's disease using fuzzy *k*-nearest neighbor approach

Hui-Ling Chen [a], Chang-Cheng Huang [a], Xin-Gang Yu [b], Xin Xu [c], Xin Sun [d], Gang Wang [d], Su-Jing Wang [d]

K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus

Publisher: IEEE Cite This PDF

Madhuri Panwar ; Amit Acharyya ; Rishad A. Shafik ; Dwaipayan Biswas All Authors

An Extended K Nearest Neighbors-Based Classifier for Epilepsy Diagnosis

Publisher: IEEE Cite This PDF

Junying Na ; Zhiping Wang ; Siqi Lv ; Zhaohui Xu All Authors

Conclusion

- KNN is a simple and intuitive algorithm.
- It does all its work at test time, i.e. needs no training.
- We talked about using KNN for classification, but we can use KNN for regression too (will talk about it in future lectures)
- KNN has a number of limitations including computational cost, sensitivity to scale and distance measure, and can suffer from the curse of dimensionality.
- With smart solutions, we can overcome some of the shortcomings of KNNs and turn them into powerful tools.
- Next lecture:
  - Decision trees
  - Introduction to Python