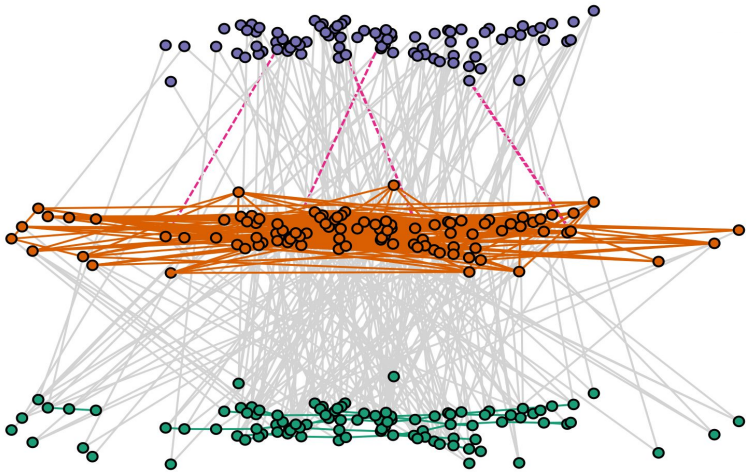
EPAH6054: Health Data Science



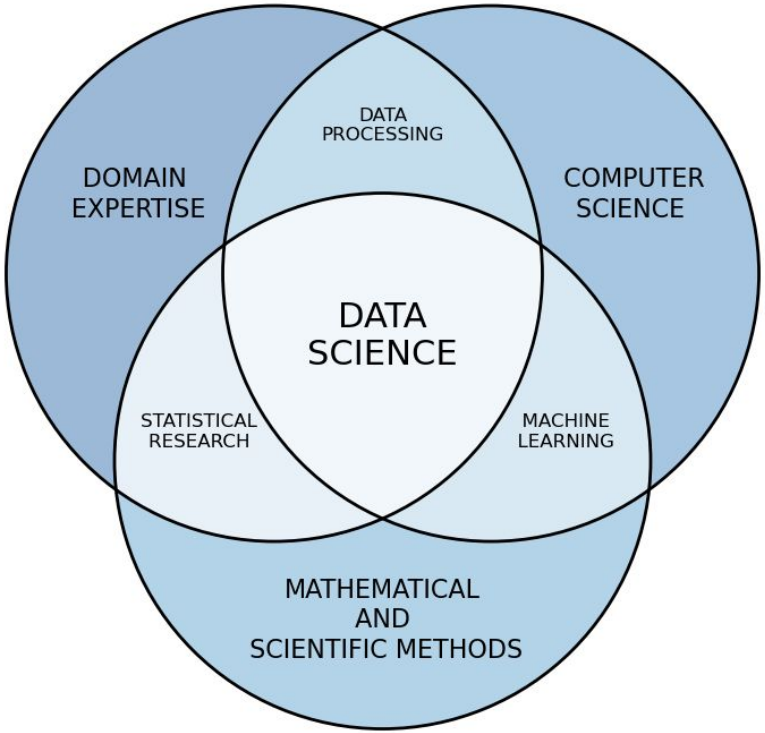
Finlay Maguire
Dalhousie University
Shared Hospital Laboratory
Public Health Alliance for Genomic Epidemiology

Overview

- What is health data science?
- How does it differ from traditional analyses of secondary data?
- What are the main types of Machine Learning?
  - Unsupervised Learning
  - Supervised Learning
- What is Deep Learning?
- Counter-intuitive nature of data and models in high dimensions

What is ~~health~~ data science?

Data Science: *data-intensive interdisciplinary approaches for exploration and prediction with large noisy datasets*

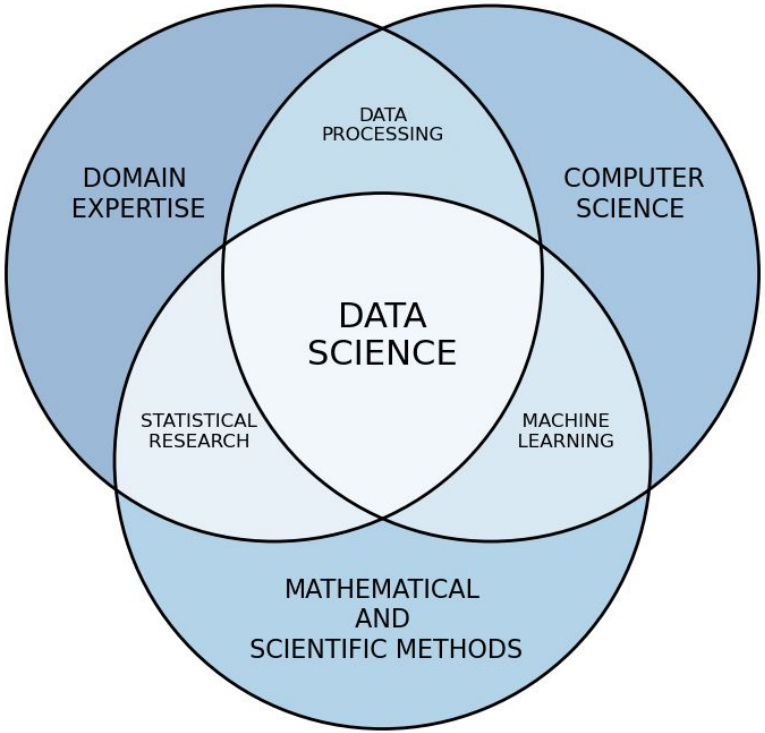


https://www.data.cymru/data-science

Data Science: *data-intensive interdisciplinary approaches for exploration and prediction with large noisy datasets*

Overlapping terms to varying degrees:

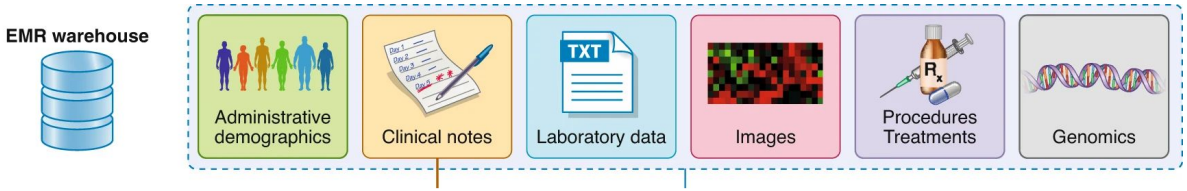
- Data Analytics
- Data Engineering
- Data Mining
- {Health,Bio,Medical}Informatics
- Database Analytics
- Business Intelligence
- Pattern Recognition
- Knowledge Discovery
- Predictive Analytics
- Quantitative Research
- Science
- **Epidemiology**
- **Statistics**



https://www.data.cymru/data-science

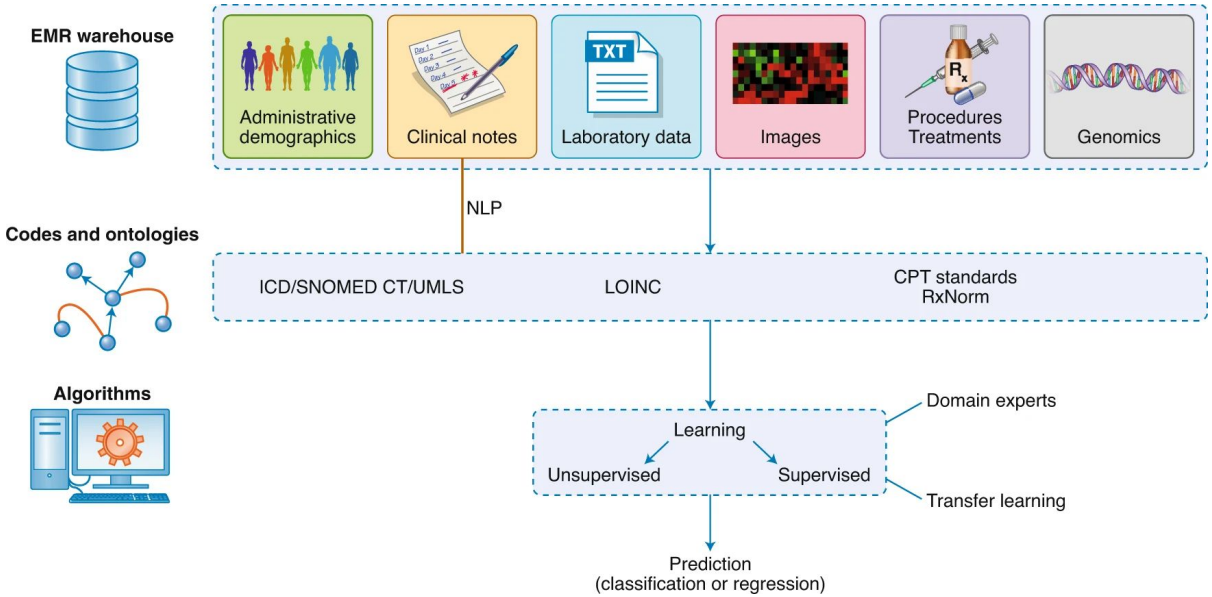
OK, what is **Health** Data Science?

Data Science applied to Health Data



Why “health data” instead of “medical data”: health encompasses medical ***(contentious)***

Data Science applied to Health Data

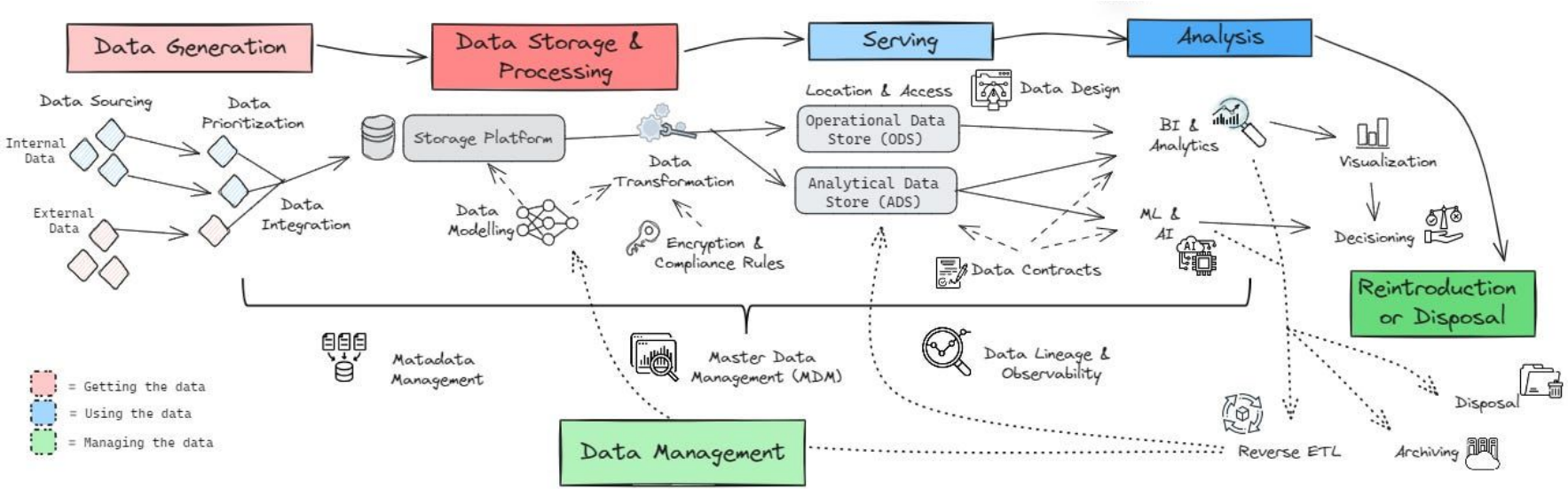


https://www.nature.com/articles/s41588-020-0698-y/figures/2

Why “health data” instead of “medical data”: health encompasses medical ***(contentious)***

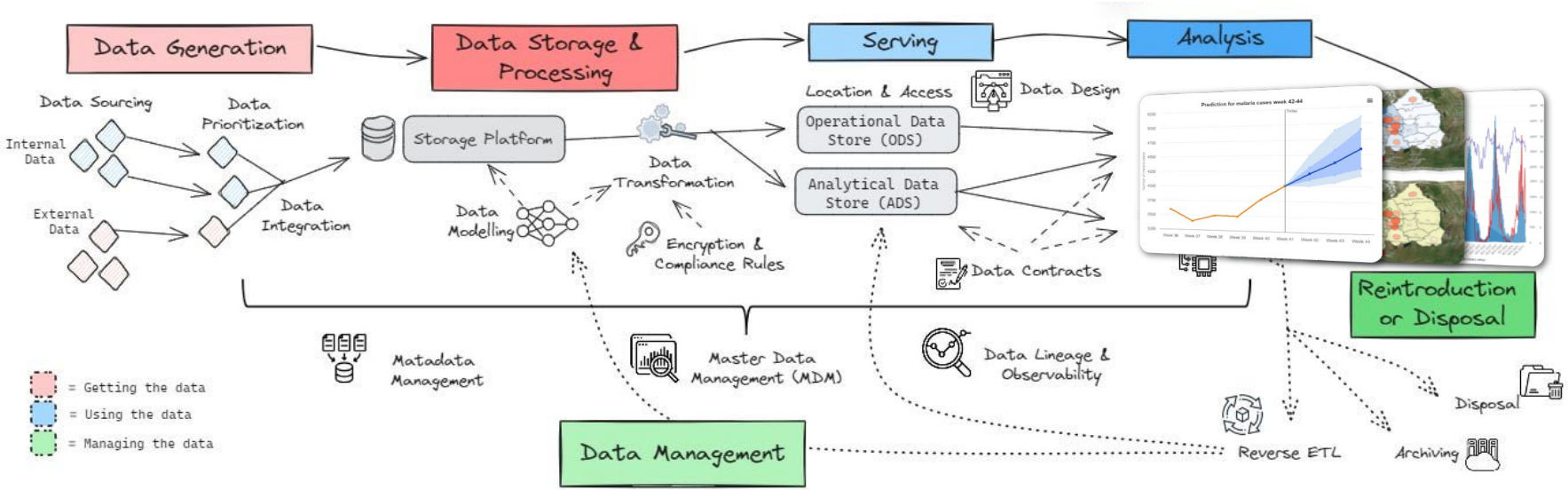
How does (health) data science (generally) differ from traditional epidemiology using secondary data?

Data science integrates within the wider data ecosystem



https://thedataecosystem.substack.com/p/issue-10-the-data-lifecycle
https://xkcd.com/2054/

Data science integrates within the wider data ecosystem



https://thedataecosystem.substack.com/p/issue-10-the-data-lifecycle
https://xkcd.com/2054/

Important as datafication pervades medicine (and more)

Medical Notes:
- 1-50MB

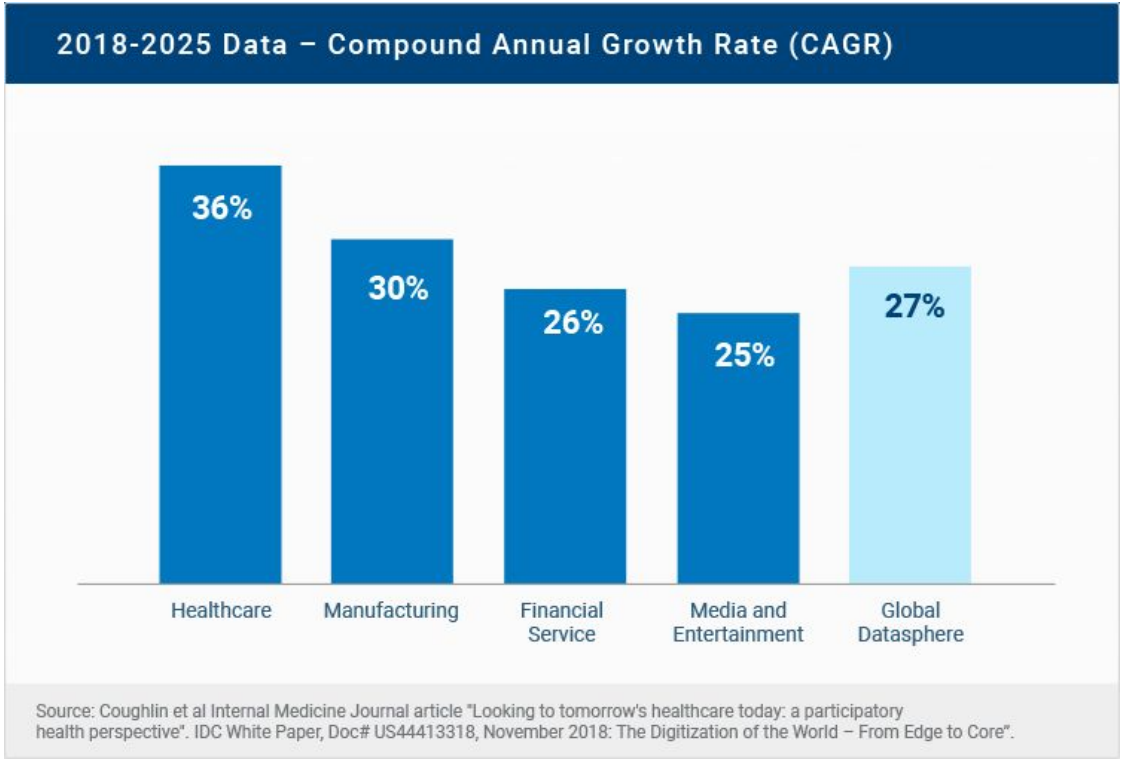
Laboratory Values:
- 10-2000MB

Physiological Sensors:
- 0.1-200GB

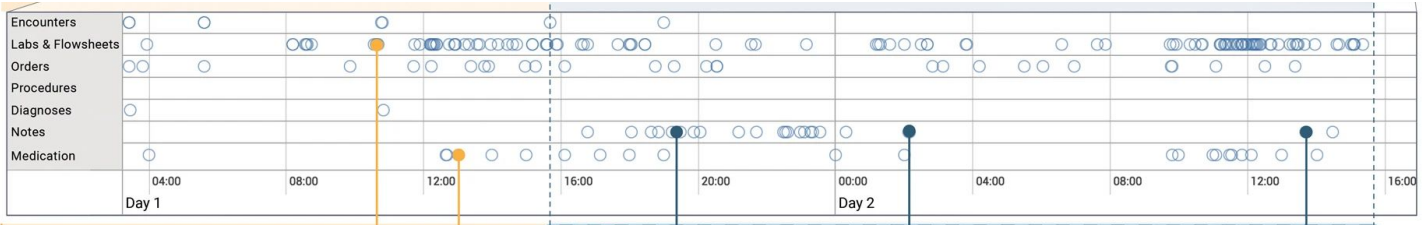
Genomics:
- 25-250GB

Medical Imaging:
- 0.1-10TB

Hospital:
- 0.5-50PB per day



Data Science handles unstructured and multi-modal data



−11:42 hours

Pegfilgrastim

-2:42 hours

Medication

Vancomycin, Metronidazole

+3:33 hours

Physician Note

"... PMH **of metastatic breast cancer, R lung malignant** effusion, and **R lung empyema** who presents with increased drainage from **R lung pleurx** tract ... "

+7:38 hours

Radiology Report - CT CHEST ABDOMEN PELVIS

"... FINDINGS : CHEST LUNGS AND PLEURA: Redemonstration of a moderate **left pleural effusion**. **interval** removal of a right chest tube within a loculated **right pleural effusion** which contains foci of air. [..]. IMPRESSION: 1. Interval progression of disease in the chest and abdomen including **increased** mediastinal **lymphadenopathy, pleural/parenchymal** disease within the right lung, probable new hepatic metastases and subcutaneous nodule within the thorax [..]"

+22:47 hours

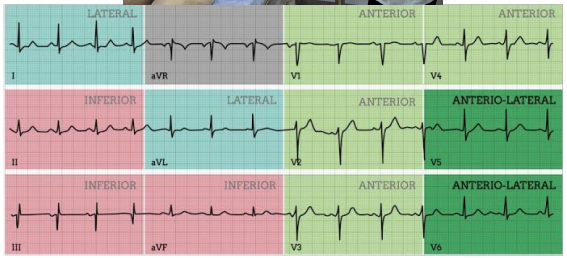
Pulmonary Consult Note

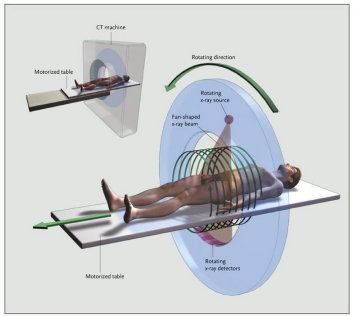
"... has a **complicated pleural space** that requires IR guidance. CT scan showing **increased loculted effusion** on R compared to date ..."

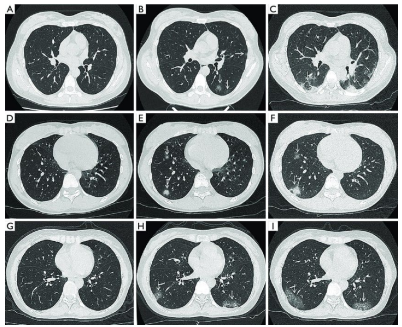
−3:23 hours

Nursing Flowsheet

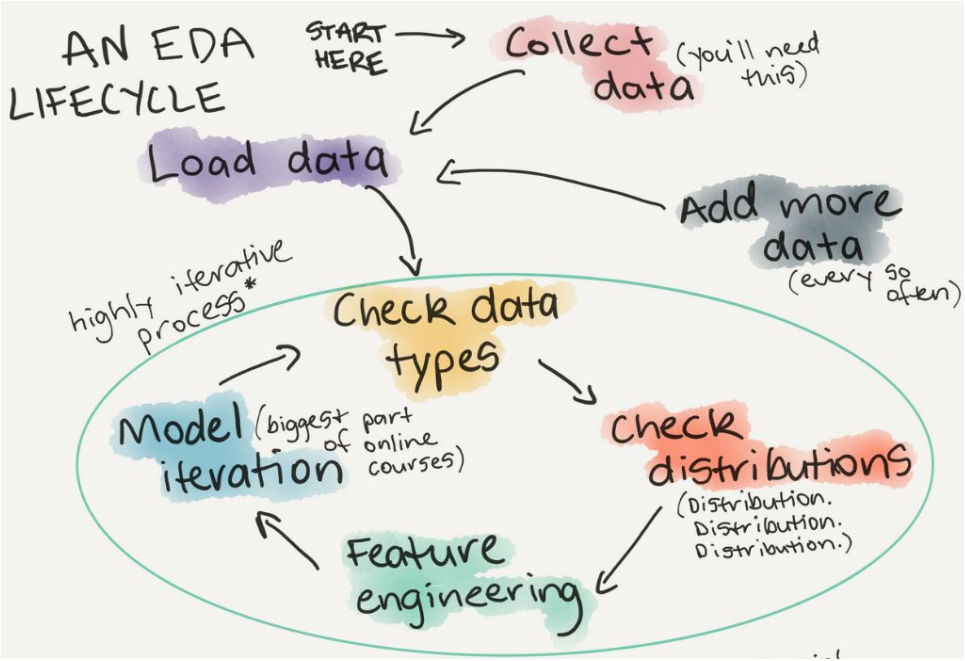
NUR RS BRADEN SCALE SCORE : 22

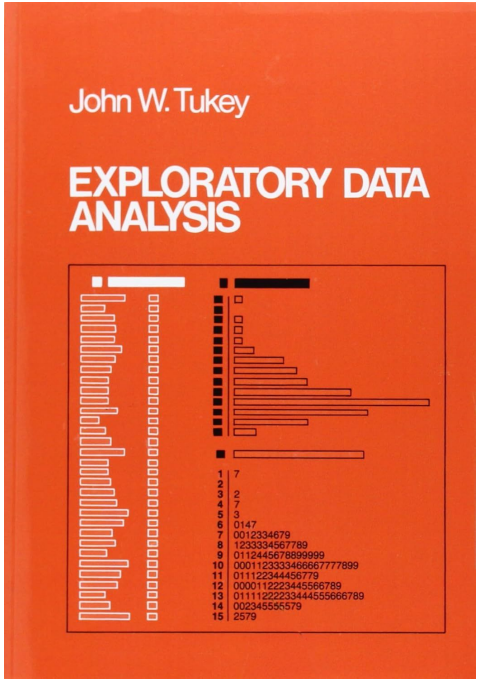






Data Science centers Exploratory Data Analysis



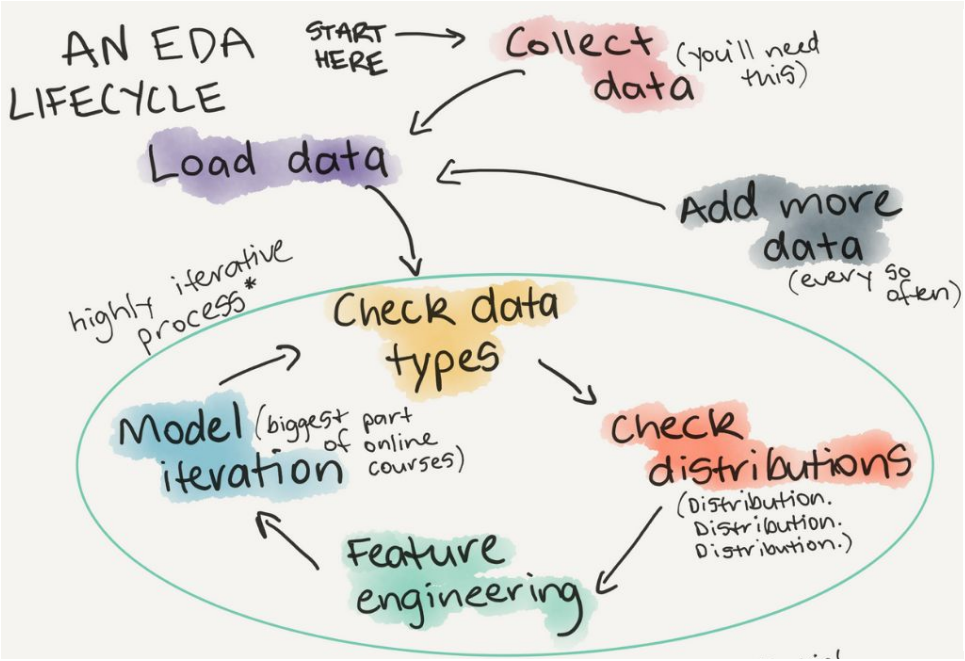


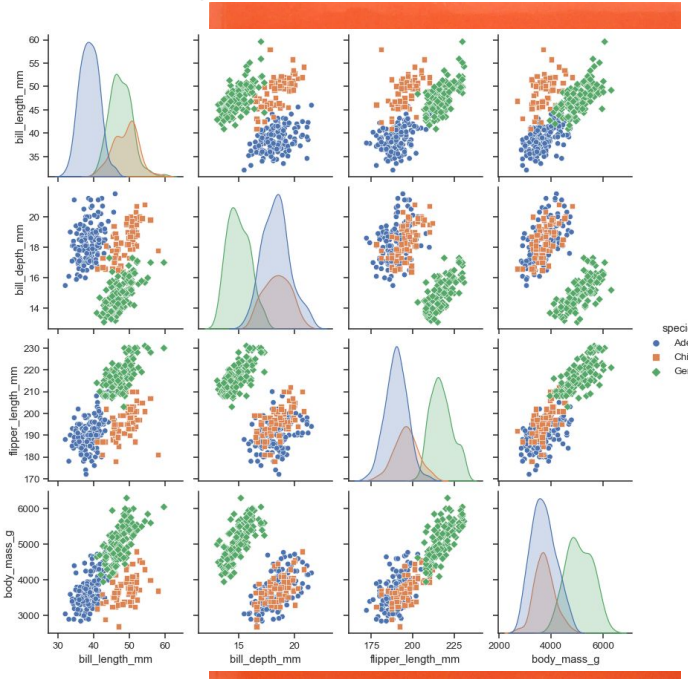
Ingest →Univariate→ Bi/Multivariate→Missing →Outliers → Hypothesise → Modify Features →Repeat

Plots!

https://www.mrdbourke.com/a-gentle-introduction-to-exploratory-data-analysis/

Data Science centers Exploratory Data Analysis





Ingest →Univariate→ Bi/Multivariate→Missing →Outliers → Hypothesise → Modify Features →Repeat

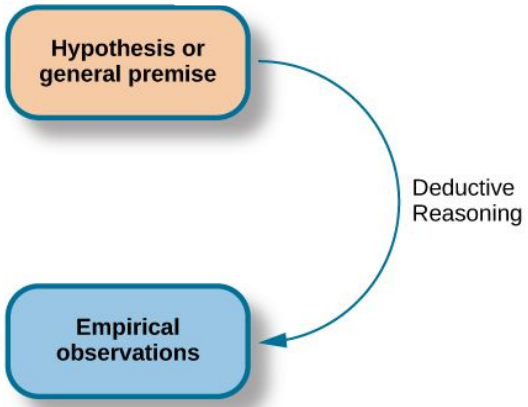
Plots!

https://www.mrdbourke.com/a-gentle-introduction-to-exploratory-data-analysis/

Explicitly incorporates inductive approaches

Deductive:

- “Condition X causes Y”
- Collect data
- Perform (typically) frequentist statistical tests
- Reject or confirm null hypothesis



https://opened.cuny.edu/courseware/lesson/14/student/?task=3

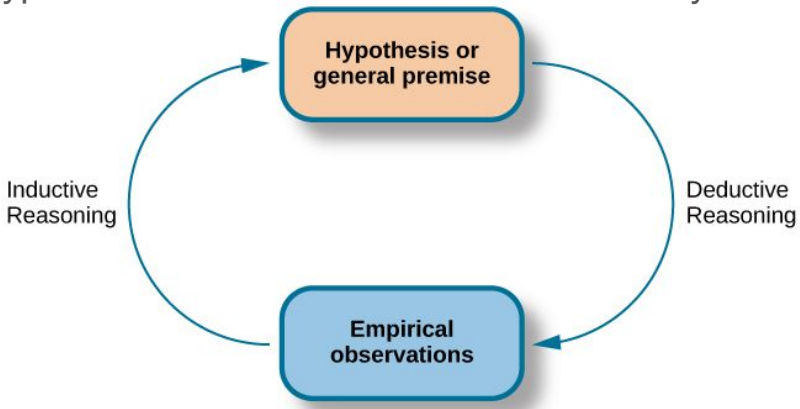
Explicitly incorporates inductive approaches

Deductive:

- “Condition X causes Y”
- Collect data
- Perform (typically) frequentist statistical tests
- Reject or confirm null hypothesis

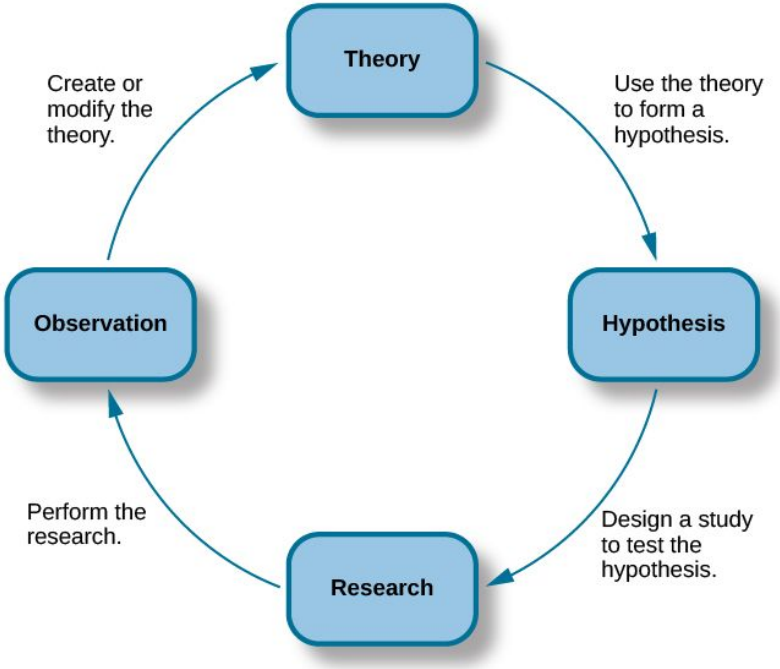
Inductive:

- Collect data
- Identify patterns in the data
- Observe X and Y seem connected somehow
- Quantify strength of association



https://opened.cuny.edu/courseware/lesson/14/student/?task=3

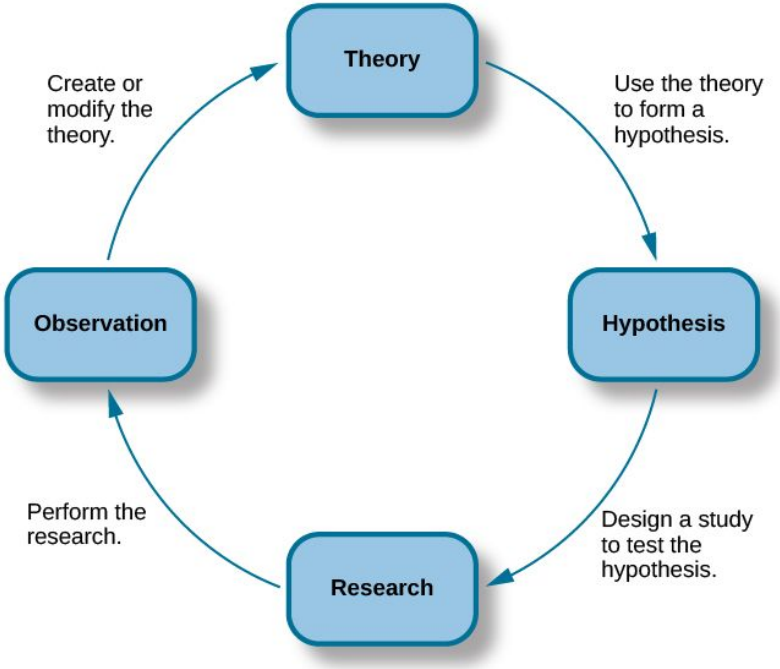
Inductive-Deductive loop aligns with knowledge cycle



https://opened.cuny.edu/courseware/lesson/14/student/?task=3

Inductive-Deductive loop aligns with knowledge cycle

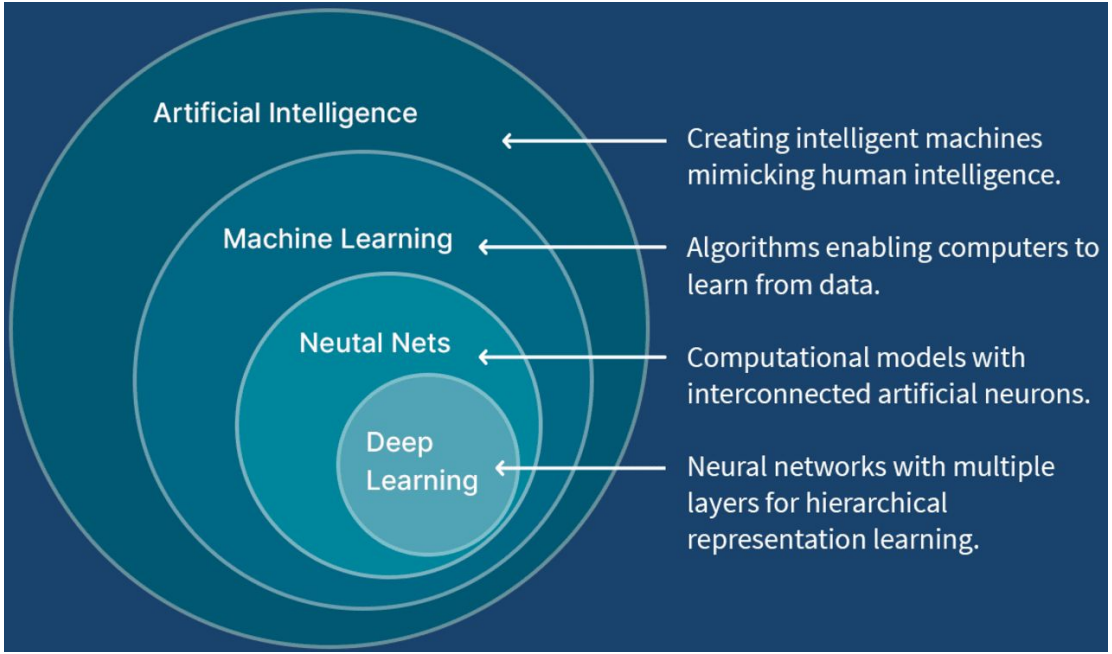
Best work: closes the loop



https://opened.cuny.edu/courseware/lesson/14/student/?task=3

Most obvious difference: Data Science focuses on Machine Learning

Machine Learning is a subset of Artificial Intelligence



https://www.qlik.com/us/augmented-analytics/machine-learning-vs-ai

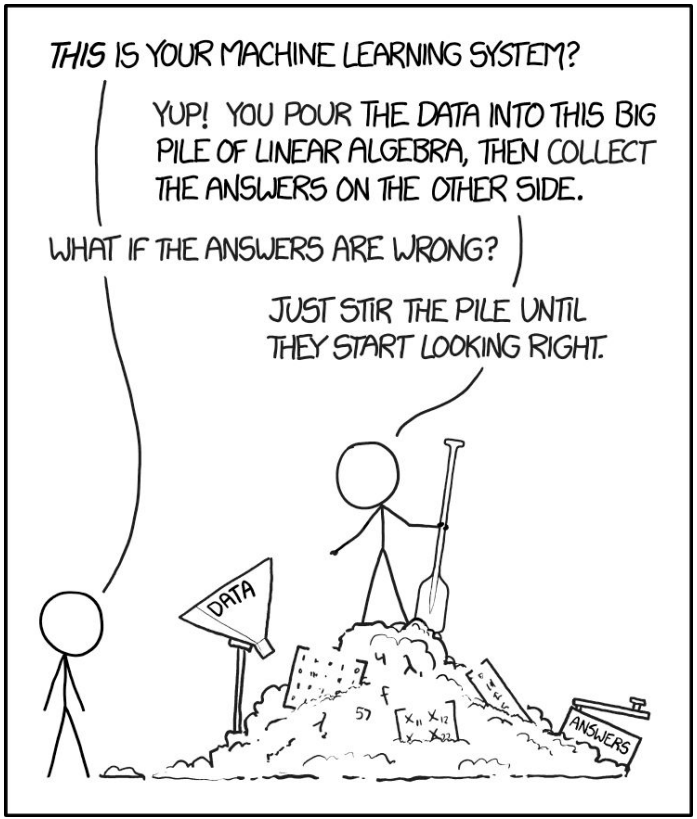
ML involves finding and using patterns in data

- “Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed”

- “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.”

- **T**ask *(play checkers)*
- **E**xperience *(data)*:
  - games played by the program (with itself)
- **P**erformance *(metric)*:
  - How often does it win

- *Training models which identify patterns in data*



https://xkcd.com/1838/

Is Machine Learning just a CS-flavoured rebrand of statistics?

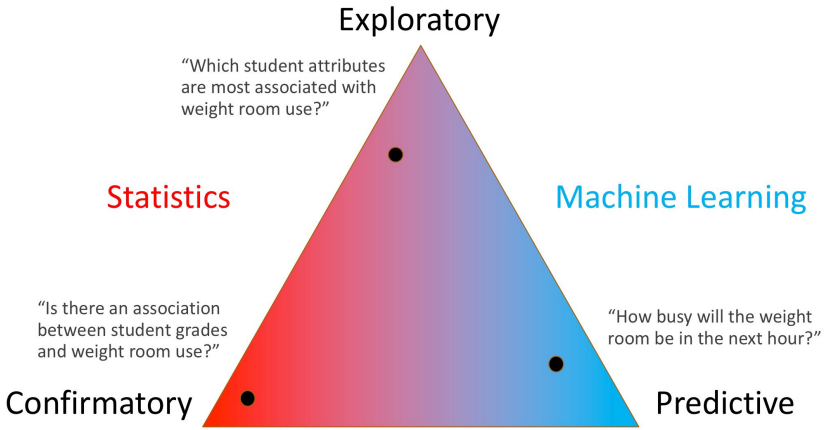
Large overlap but difference in people and focus

- Many shared methods

https://stats.stackexchange.com/questions/442128/machine-learning-vs-statistical-learning-vs-statistics

Large overlap but difference in people and priorities

- Many shared methods
- Difference in focus/priorities/culture

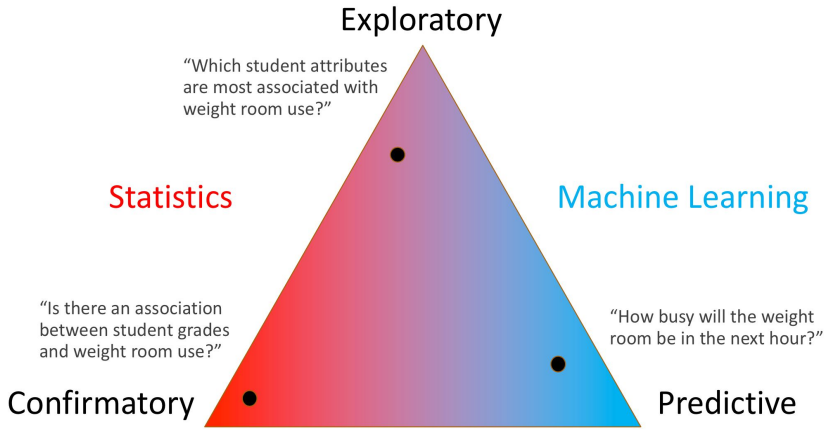


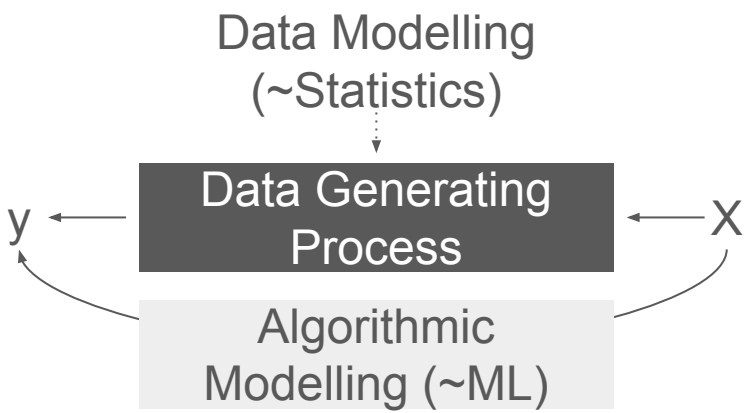
https://stats.stackexchange.com/questions/442128/machine-learning-vs-statistical-learning-vs-statistics

Large overlap but difference in people and priorities

- Many shared methods
- Difference in focus/priorities/culture

- Alternative framing:
  - Data Modelling
  - Algorithmic Modelling





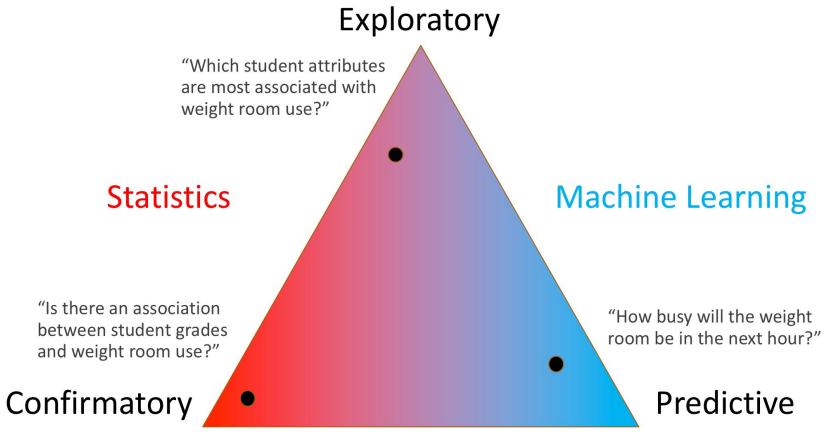
https://stats.stackexchange.com/questions/442128/machine-learning-vs-statistical-learning-vs-statistics

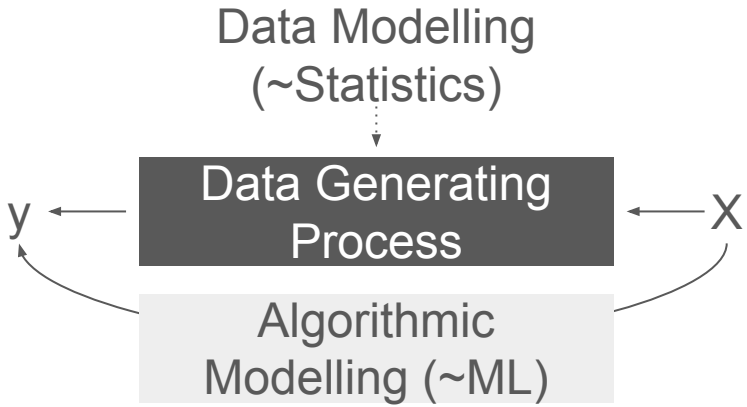
Large overlap but difference in people and priorities

- Many shared methods
- Difference in focus/priorities/culture

- Alternative framing:
  - Data Modelling
  - Algorithmic Modelling

- ML Pitfalls (can be):
  - Less rigorous/principled
  - Prone to reinventing the wheel



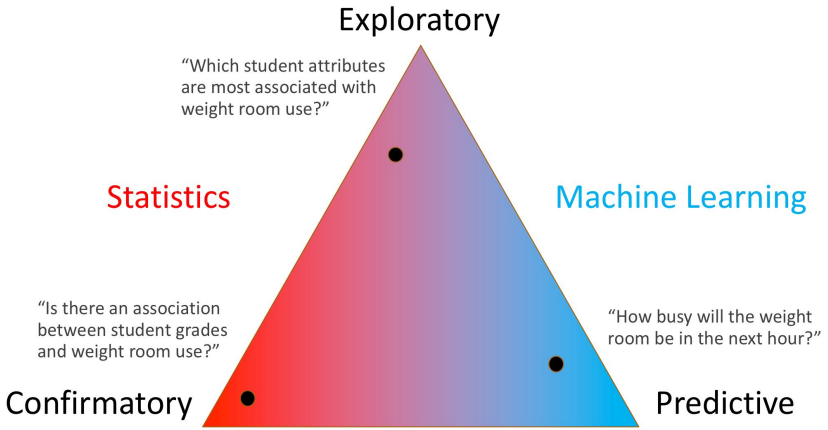


https://stats.stackexchange.com/questions/442128/machine-learning-vs-statistical-learning-vs-statistics

Large overlap but difference in people and priorities

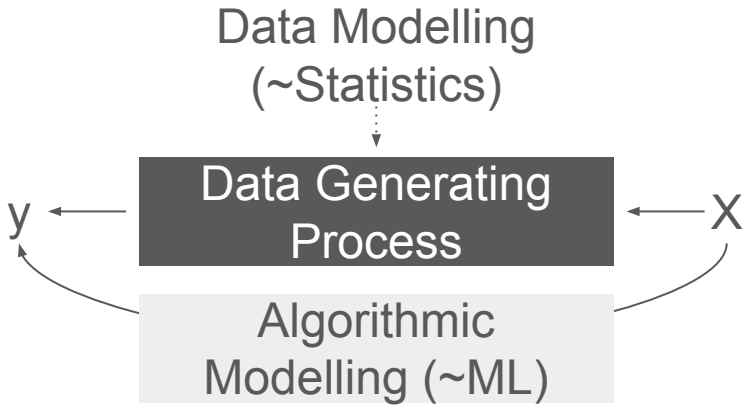
- Many shared methods
- Difference in focus/priorities/culture

- Alternative framing:
  - Data Modelling
  - Algorithmic Modelling



- ML Pitfalls (can be):
  - Less rigorous/principled
  - Prone to reinventing the wheel

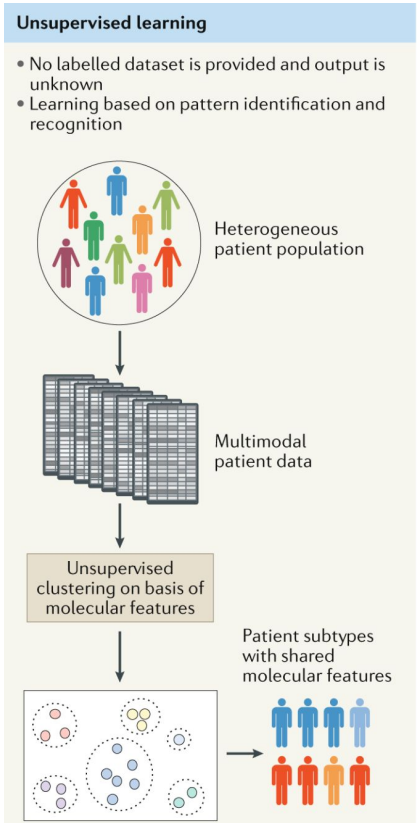
- ML Benefits (can be):
  - More flexible
  - Less prescriptive/intimidating



https://stats.stackexchange.com/questions/442128/machine-learning-vs-statistical-learning-vs-statistics

What are the main types of Machine Learning?

Types of Machine Learning

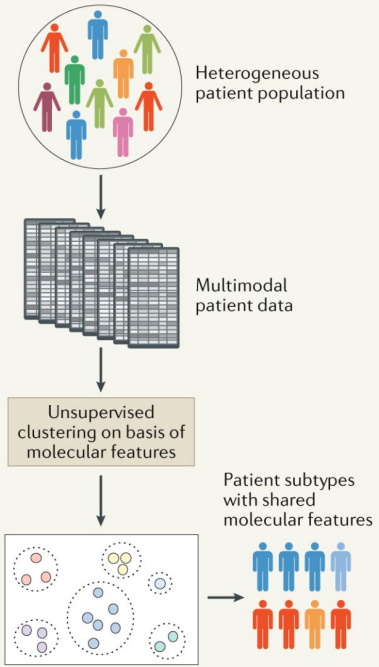


Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M.B. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol **16**, 440–456 (2020). https://doi.org/10.1038/s41582-020-0377-8

Types of Machine Learning

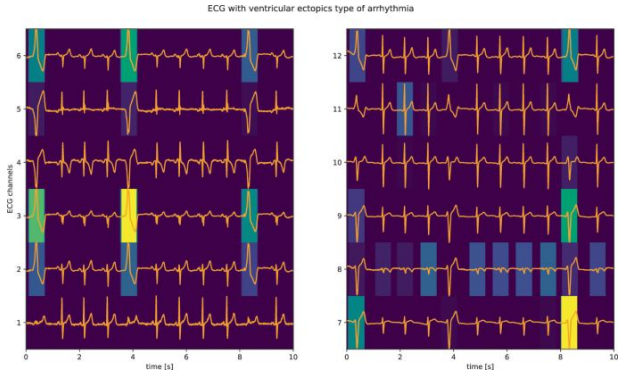
Unsupervised learning

- No labelled dataset is provided and output is unknown
- Learning based on pattern identification and recognition

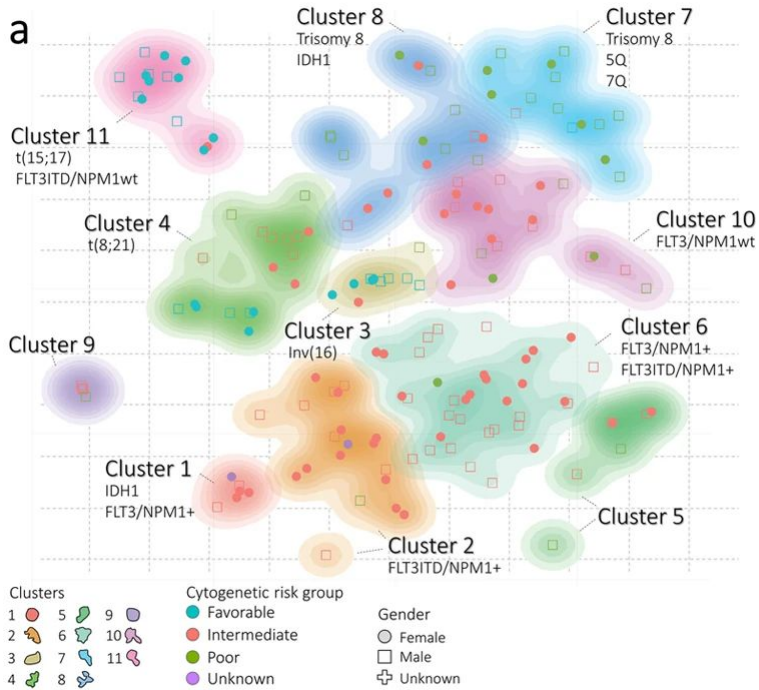


Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M.B. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol **16**, 440–456 (2020). https://doi.org/10.1038/s41582-020-0377-8

Anomaly Detection - ECG

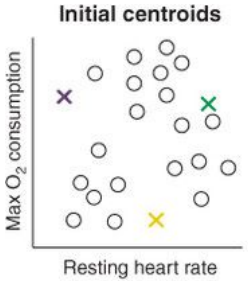


Clustering - Acute Myeloid Leukemia Subtypes



Taskesen, E., Huisman, S., Mahfouz, A. et al. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. Sci Rep **6**, 24949 (2016). https://doi.org/10.1038/srep24949

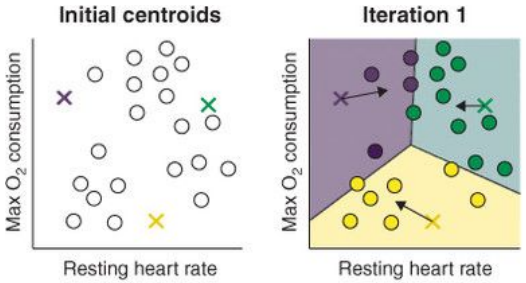
Unsupervised Learning: Finding Structure within Datasets



K-means Clustering:

1. Initialise cluster centroids (randomly)

Unsupervised Learning: Finding Structure within Datasets



K-means Clustering:

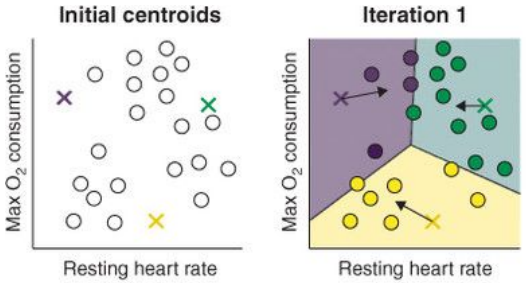
1. Initialise cluster centroids (randomly)

2. Assign observations to nearest centroid

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

Inferred label for obs i, whereas supervised learning has given label y_i

Unsupervised Learning: Finding Structure within Datasets



K-means Clustering:

1. Initialise cluster centroids (randomly)

2. Assign observations to nearest centroid

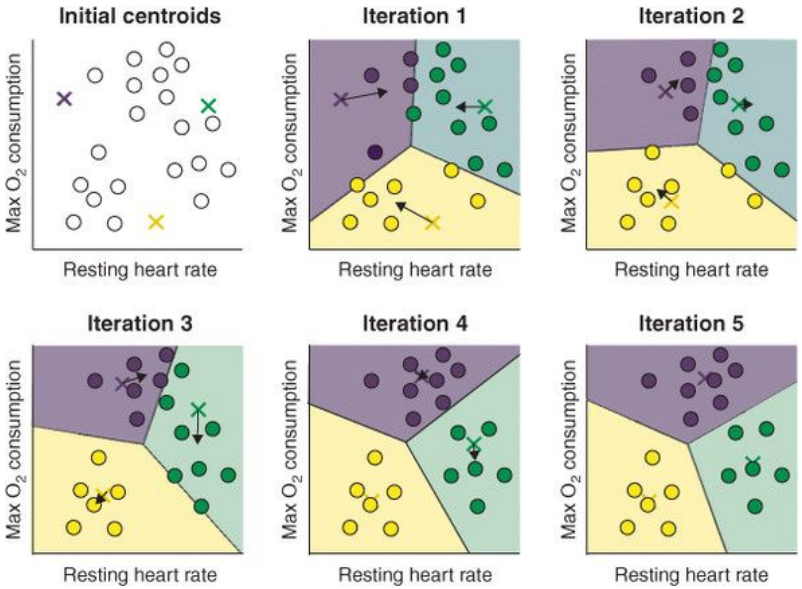
$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

Inferred label for obs i, whereas supervised learning has given label y_i

3. Move centroids to mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=j} \mathbf{x}_i$$

Unsupervised Learning: Finding Structure within Datasets



K-means Clustering:

1. Initialise cluster centroids (randomly)

Repeat until convergence:

2. Assign observations to nearest centroid

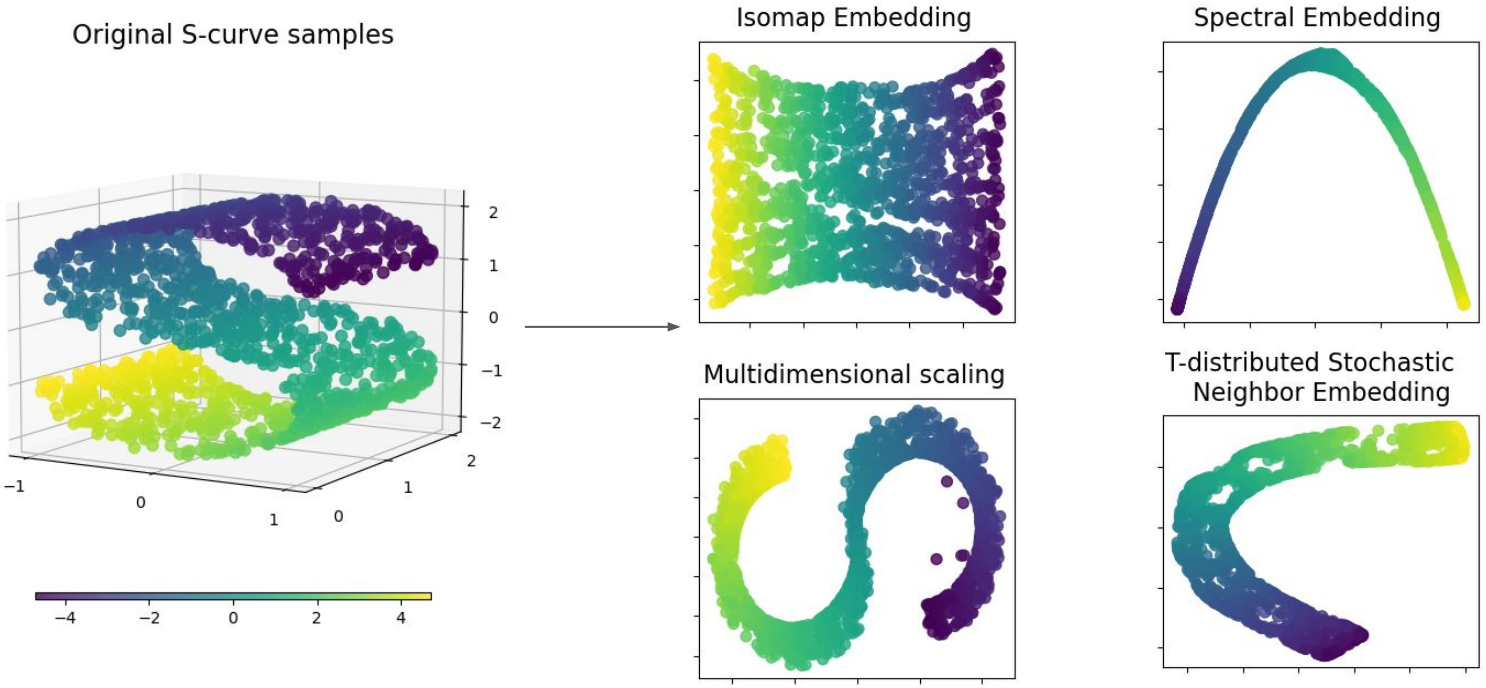
$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

Inferred label for obs i, whereas supervised learning has given label y_i

3. Move centroids to mean of assigned observations

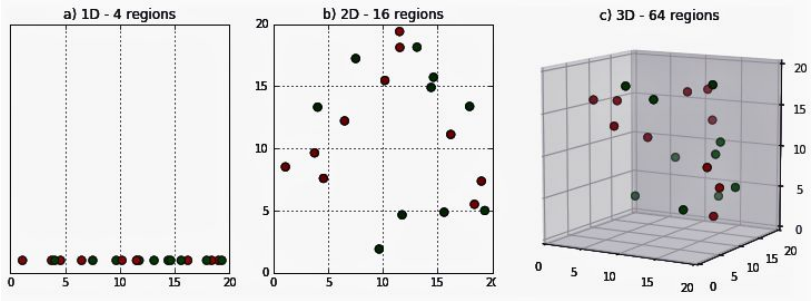
$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=j} \mathbf{x}_i$$

Unsupervised Learning: Finding Lower Dimension Projections



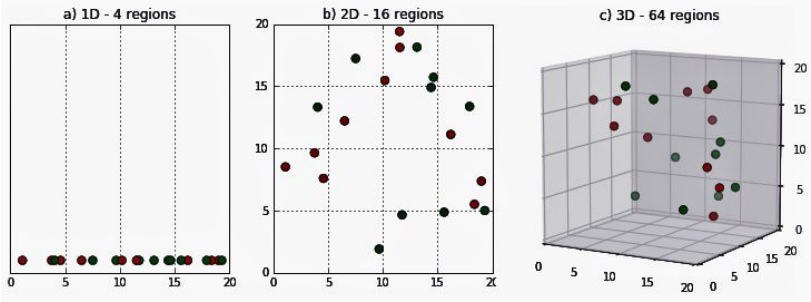
Why is this hard? High dimensions are counter-intuitive!

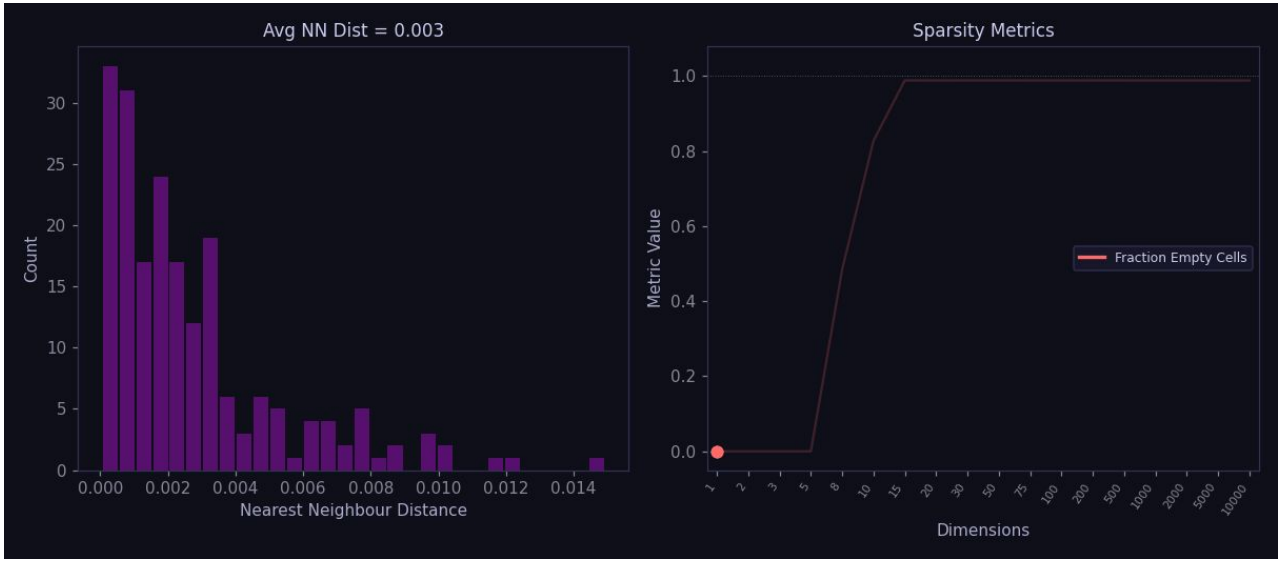
Data becomes increasingly sparse in high dimensions



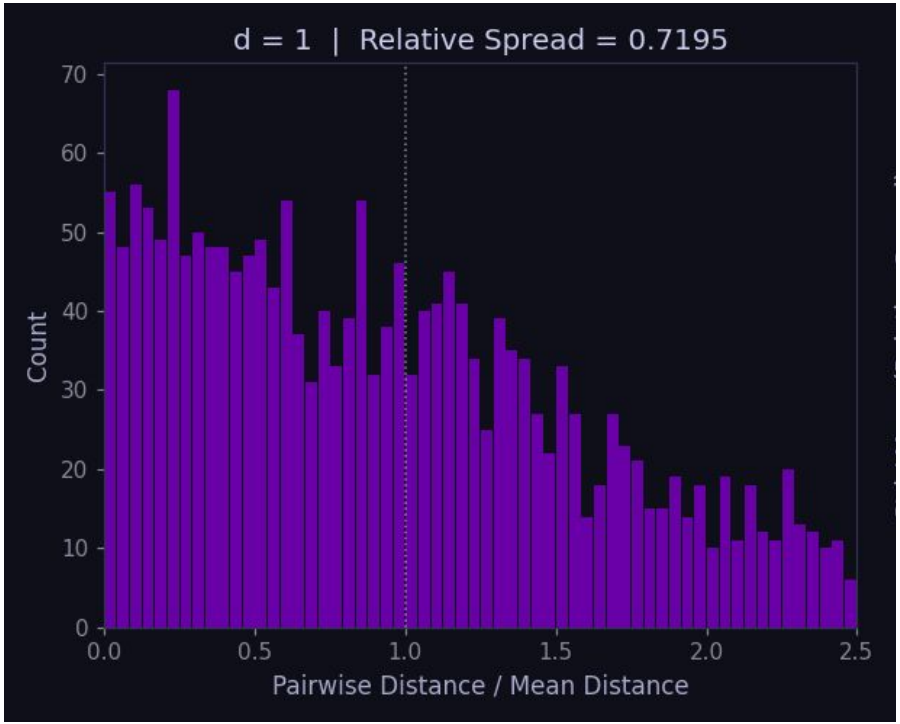
https://medium.com/analytics-vidhya/the-curse-of-dimensionality-and-its-cure-f9891ab72e5c

Data becomes increasingly sparse in high dimensions

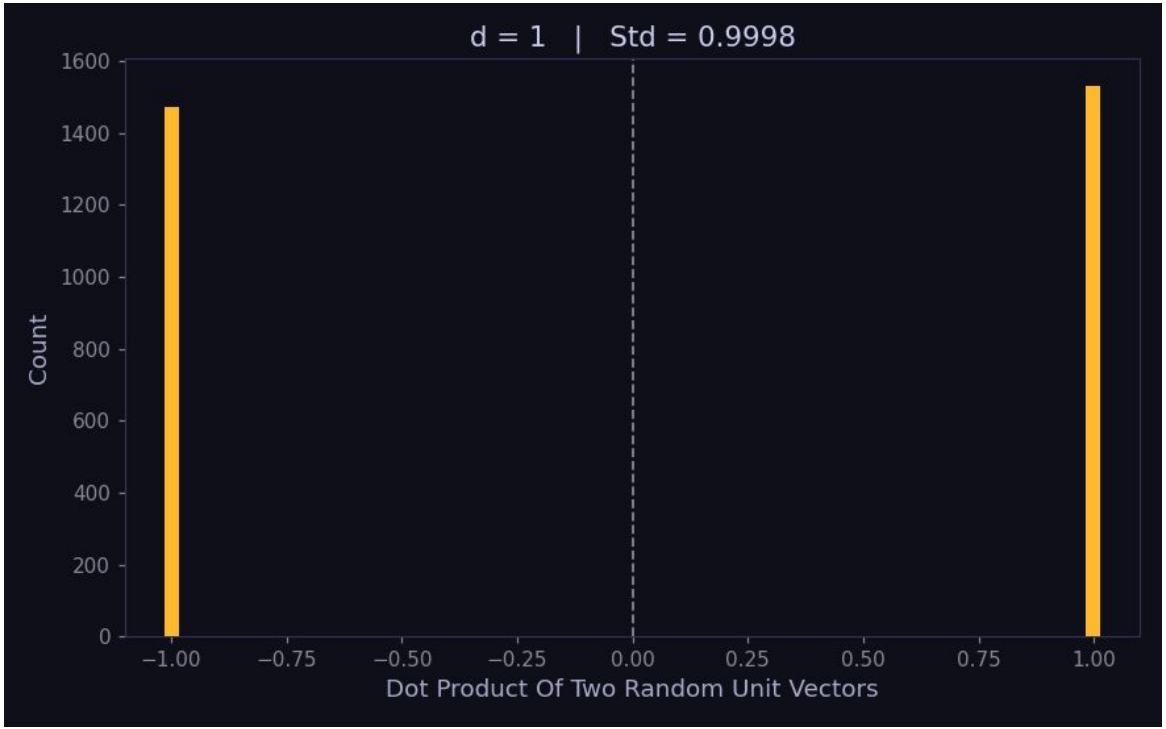


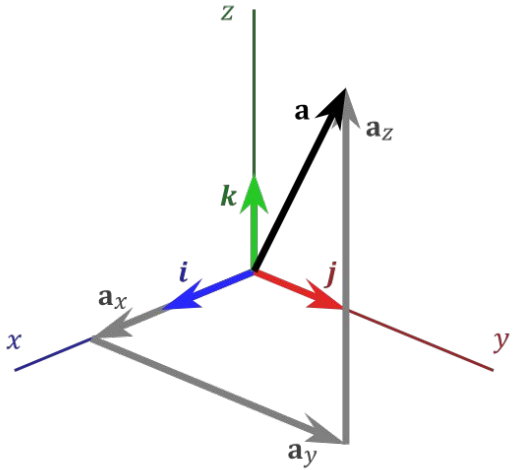


All points start to look equally far apart

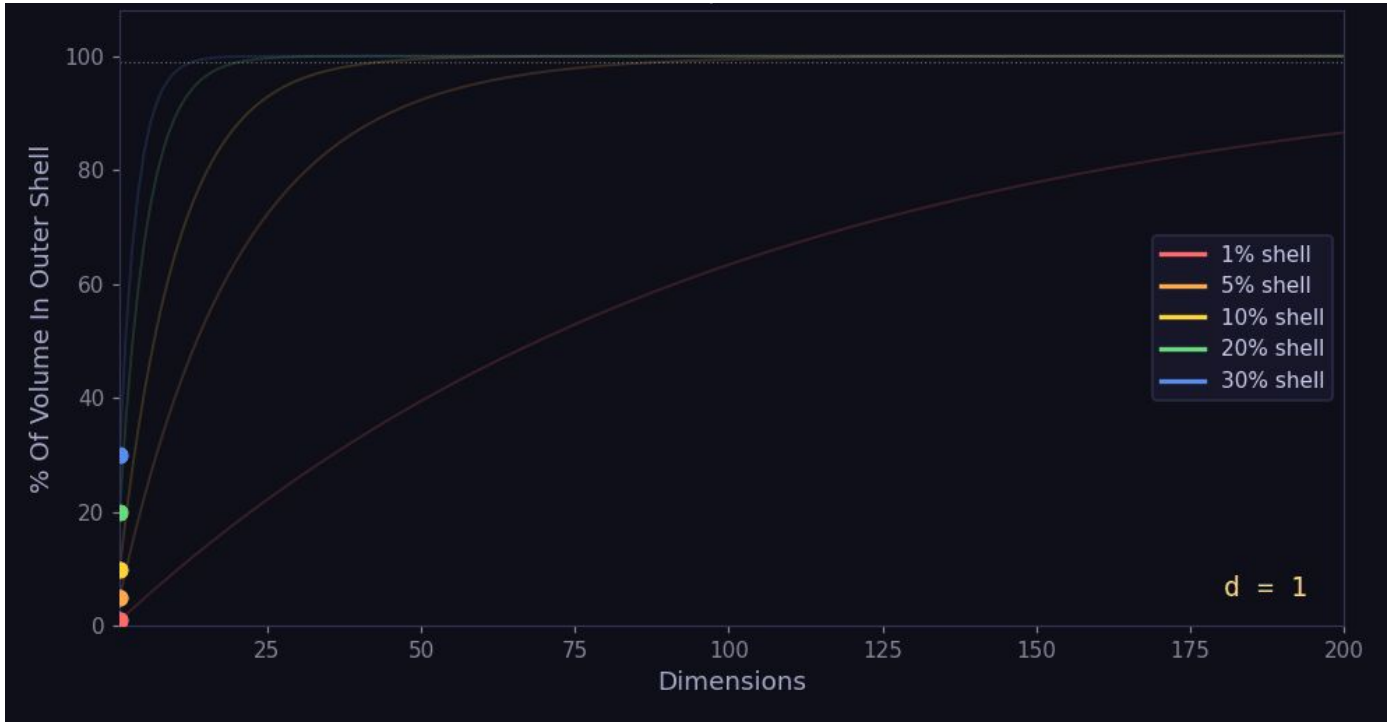


Everything becomes (near) orthogonal

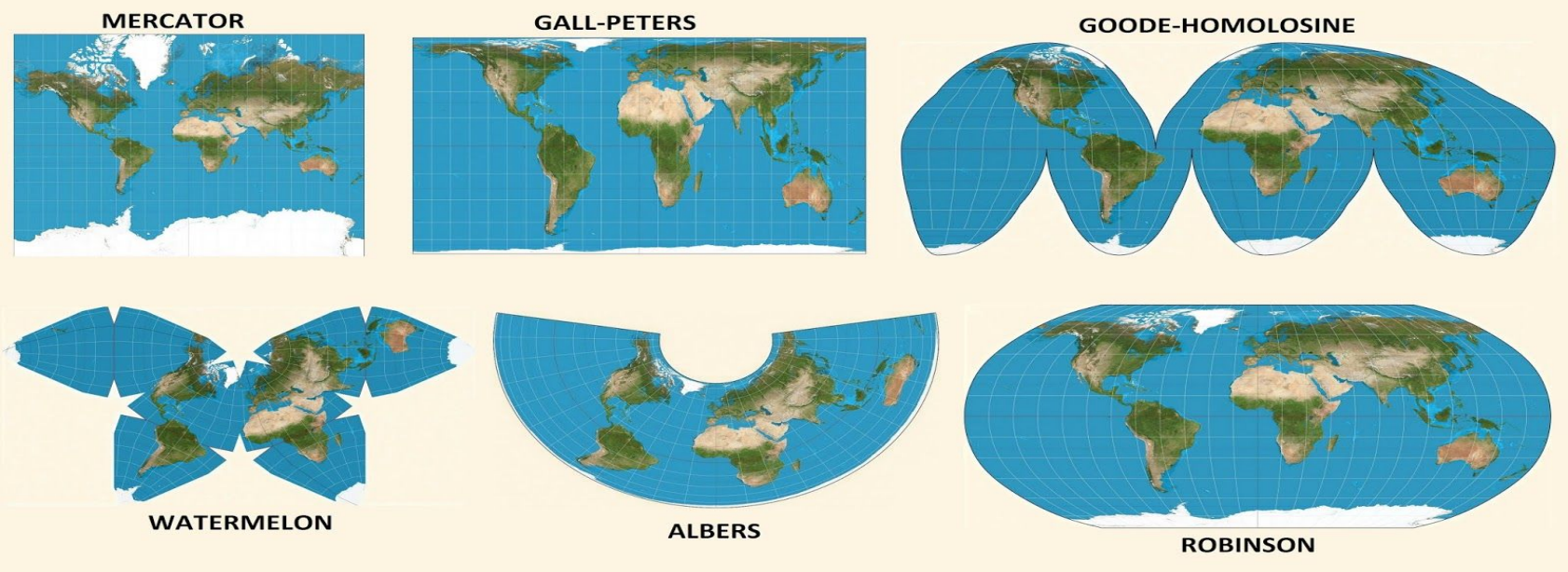




Weird distributions: hypersphere volume concentrates on surface



No lower dimension representation will ever be perfect

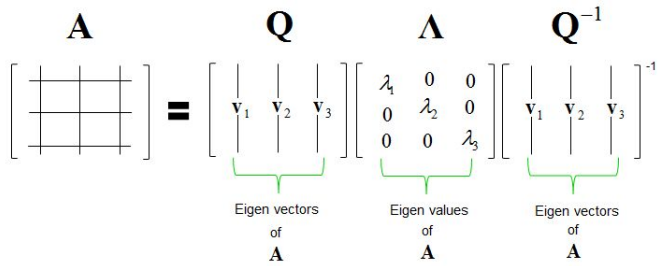


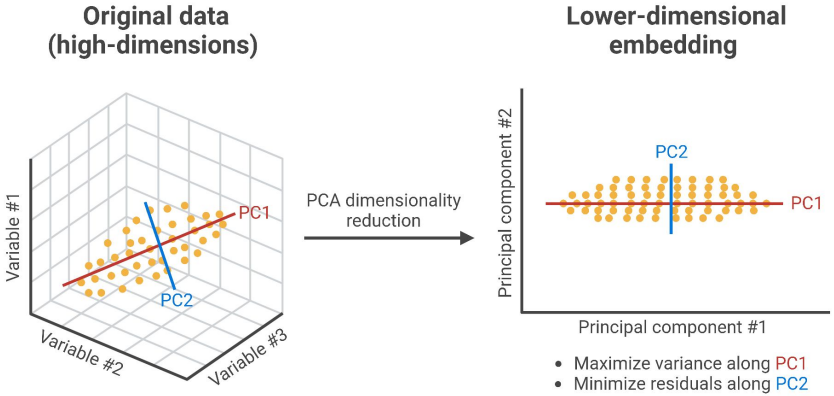
So, how can we do it?

Principal Component Analysis - Simplest Method

Reorient the data in the direction of maximal variance

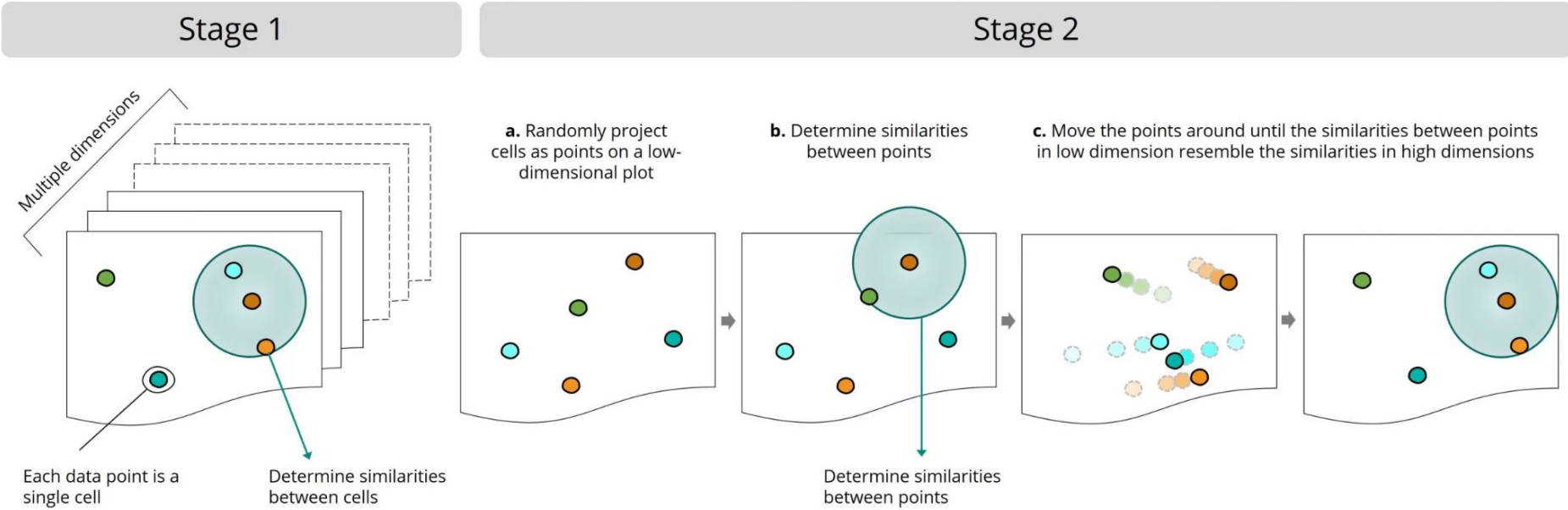
1. Center the data
2. Calculate the covariance matrix
3. Perform eigendecomposition
4. Sort and select n principal components
5. Project the data onto the reduced space





https://www.biorender.com/template/principal-component-analysis-pca-transformation

t-SNE (stochastic neighbour embedding) and UMAP



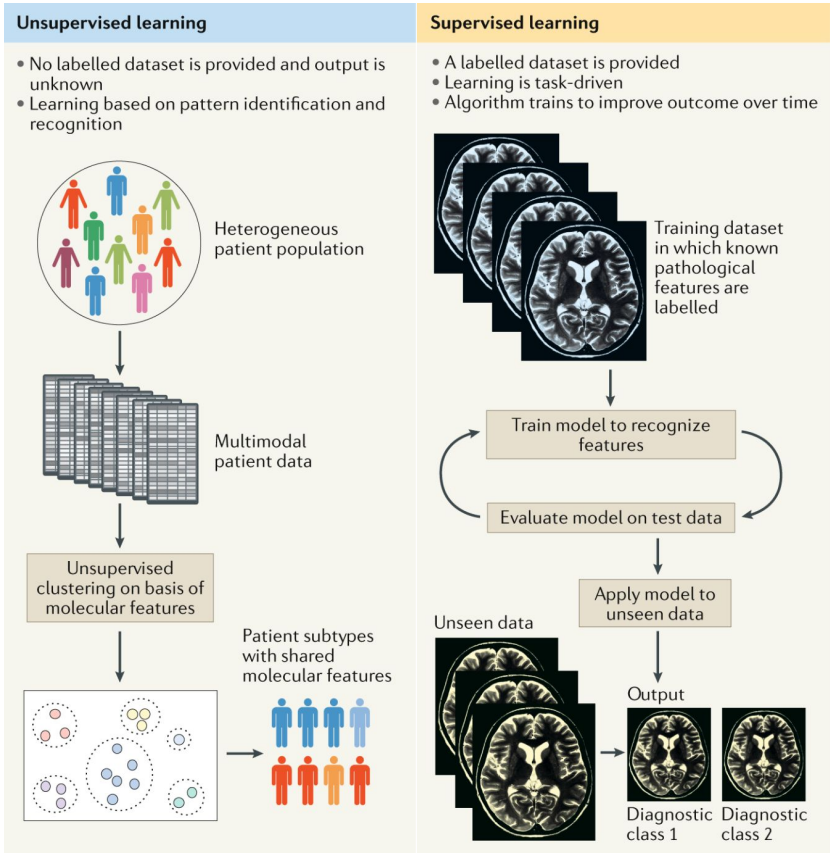
- Pairwise probability distribution in all dimensions
- Pairwise probability distribution in few dimensions
- Stochastic minimisation of KL divergence between distributions

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." - von Neumann

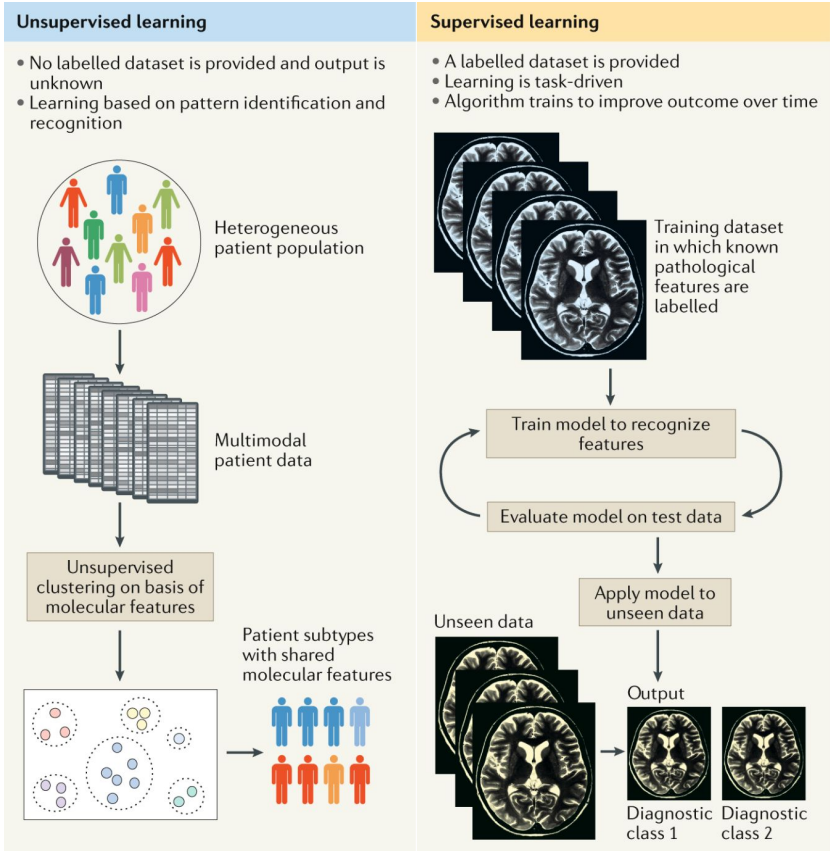
https://www.scdiscoveries.com/blog/knowledge/what-is-t-sne-plot/

What other types of ML are used?

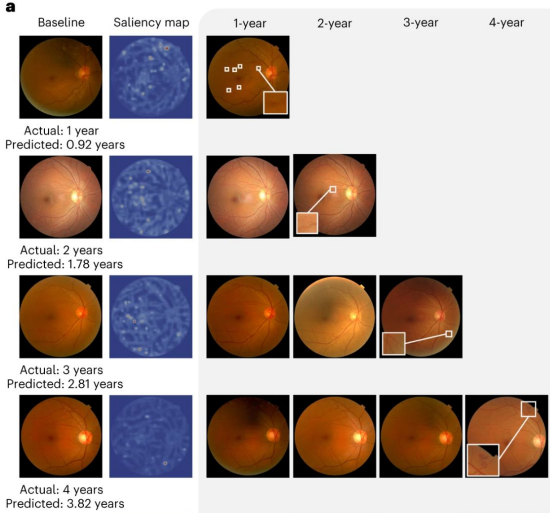
Types of Machine Learning - Supervised Learning



Types of Machine Learning - Supervised Learning

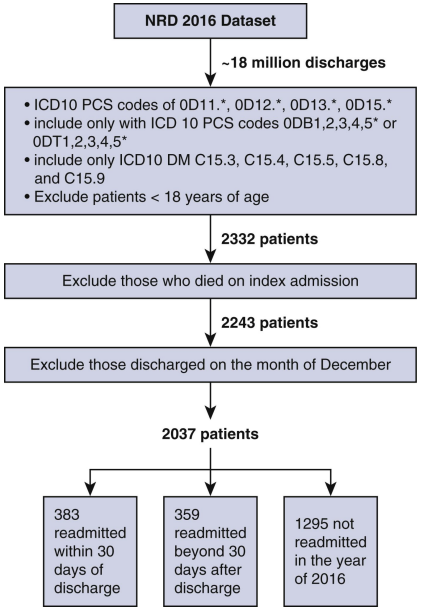


Prediction of Diabetic Retinopathy Progression



*Dai, L., Sheng, B., Chen, T. et al. A deep learning system for predicting time to progression of diabetic retinopathy. Nat Med **30**, 584–594 (2024). https://doi.org/10.1038/s41591-023-02702-z*

Prediction of 30-Day Readmission



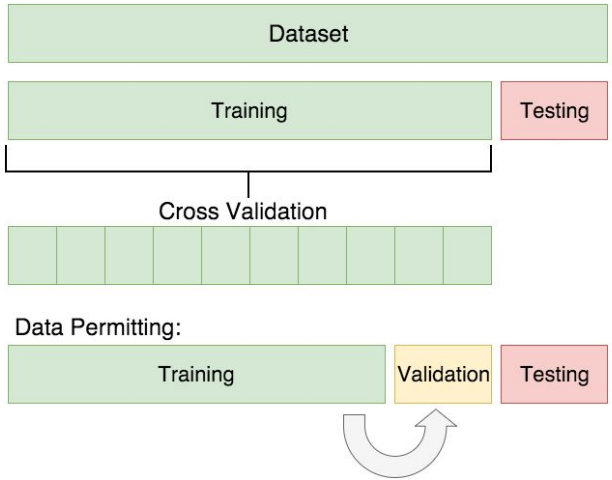
Bolourani, Siavash, et al. "Using machine learning to predict early readmission following esophagectomy." The Journal of Thoracic and Cardiovascular Surgery 161.6 (2021): 1926-1939.

Supervised Learning: Classification and Regression

Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P002	41	F	22.1	118	75	4.9	5.2	4.7	No	No	Healthy	4.1%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P005	58	M	29.9	145	91	6.9	6.5	5.8	No	No	Diabetic	14.7%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	38.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%

Supervised Learning: Classification and Regression

Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P002	41	F	22.1	118	75	4.9	5.2	4.7	No	No	Healthy	4.1%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P005	58	M	29.9	145	91	6.9	6.5	5.8	No	No	Diabetic	14.7%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	38.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%



What does it mean to fit a supervised model?

Decision boundaries and loss functions

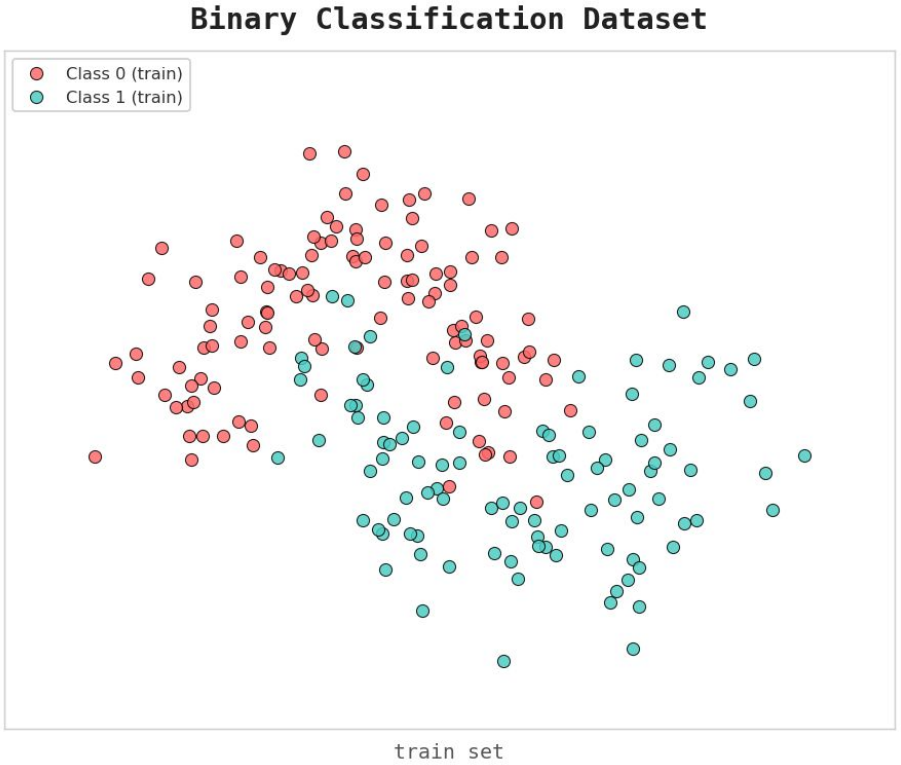
Find β so our function maps X to y

$$\hat{y} = f_\beta(x)$$

In other words: find β that defines boundary between labels

Minimising some loss function like binary cross-entropy (log-loss)

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$



Decision boundaries and loss functions

Find β so our function maps X to y

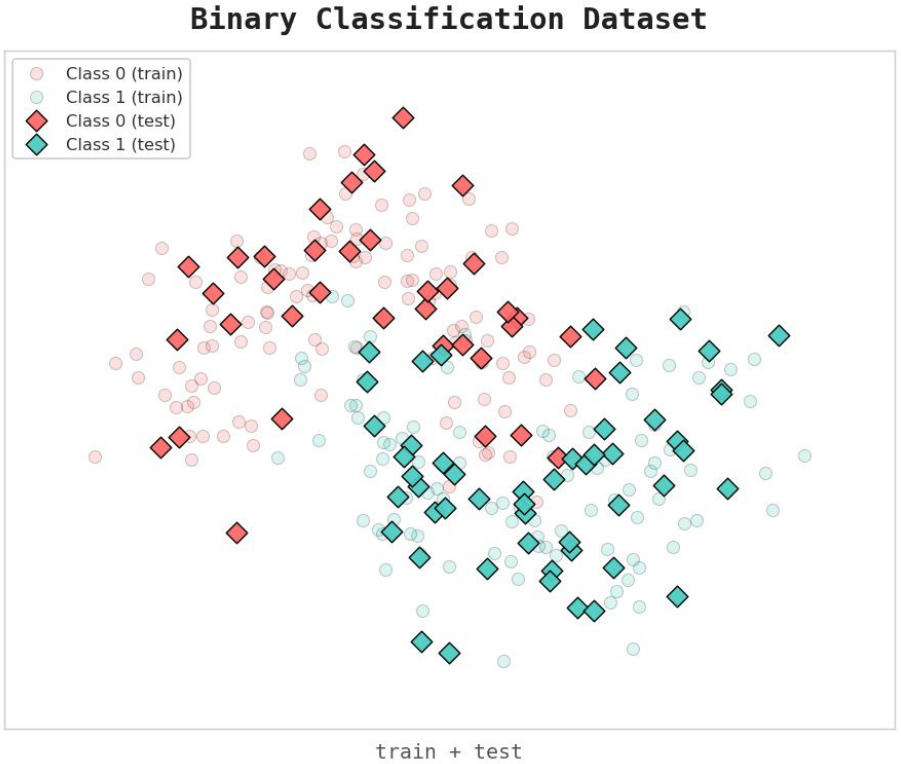
$$\hat{y} = f_\beta(x)$$

In other words: find β that defines boundary between labels

Minimising some loss function like binary cross-entropy (log-loss)

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

Ideally in a way that generalises to new data!



Decision boundaries and loss functions

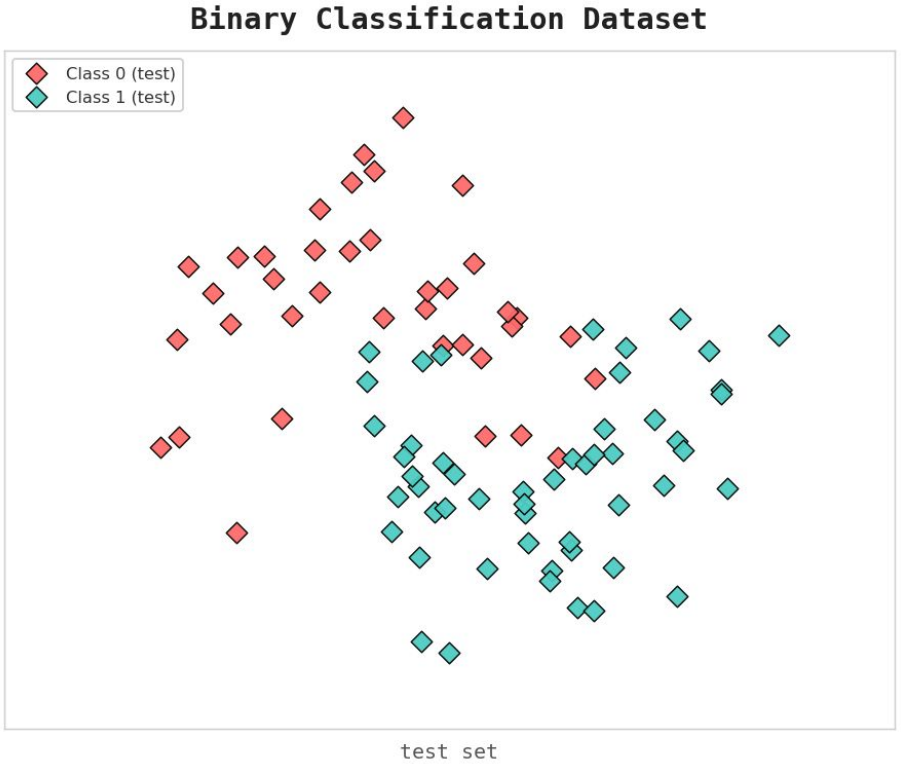
Find β so our function maps X to y

$$\hat{y} = f_\beta(x)$$

In other words: find β that defines boundary between labels

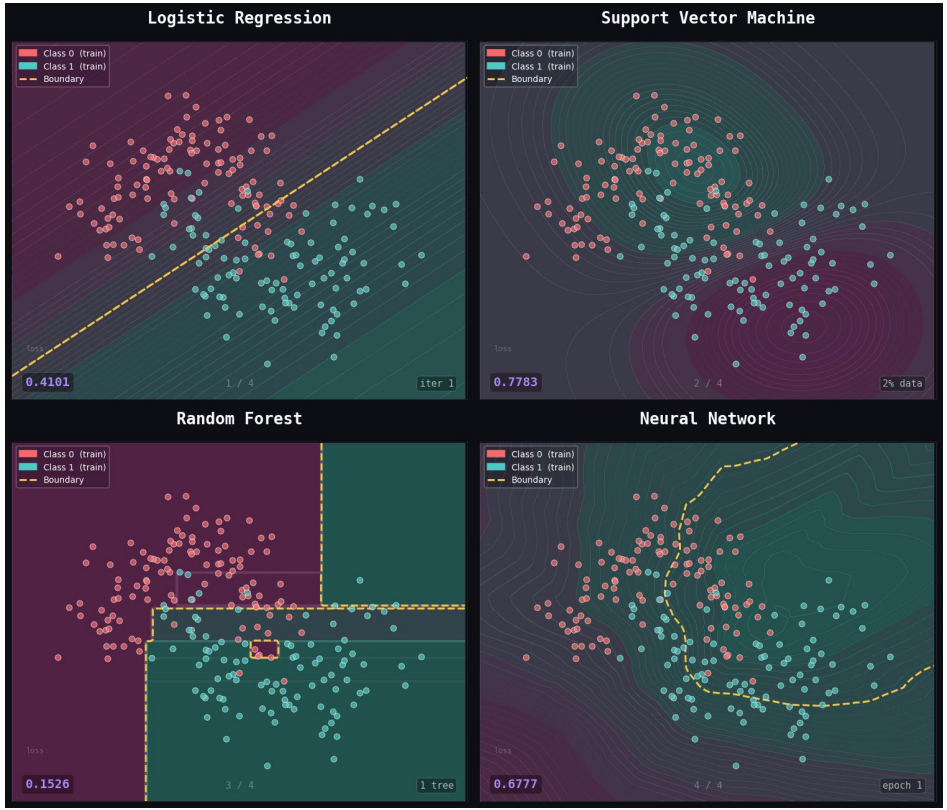
Minimising some loss function like binary cross-entropy (log-loss)

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

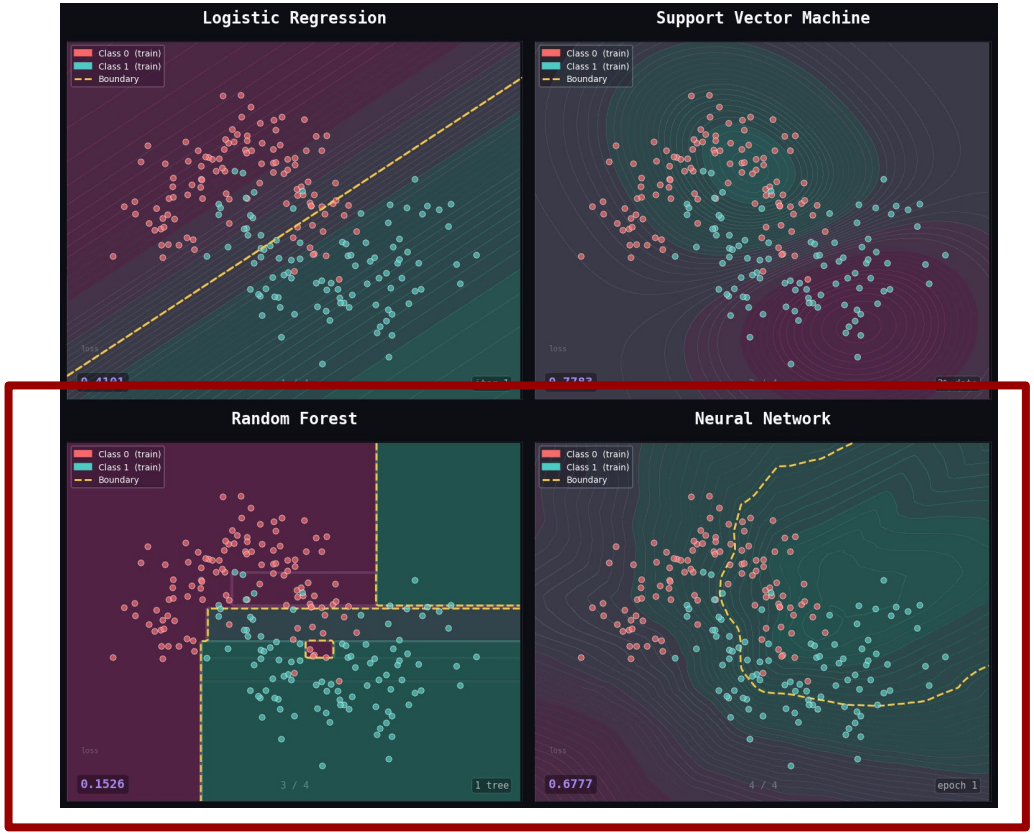


Ideally in a way that generalises to new data!

Decision boundaries and loss functions

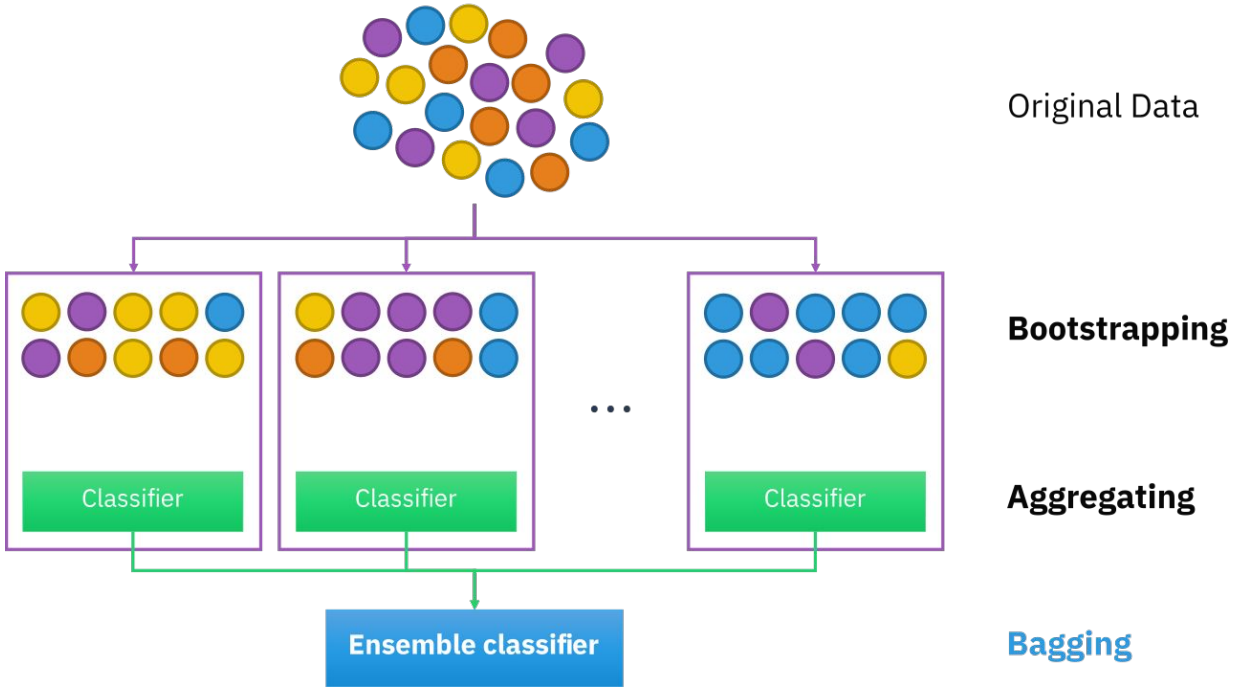


Decision boundaries and loss functions



Can’t get a well-performing model? Just combine lots of weak ones!

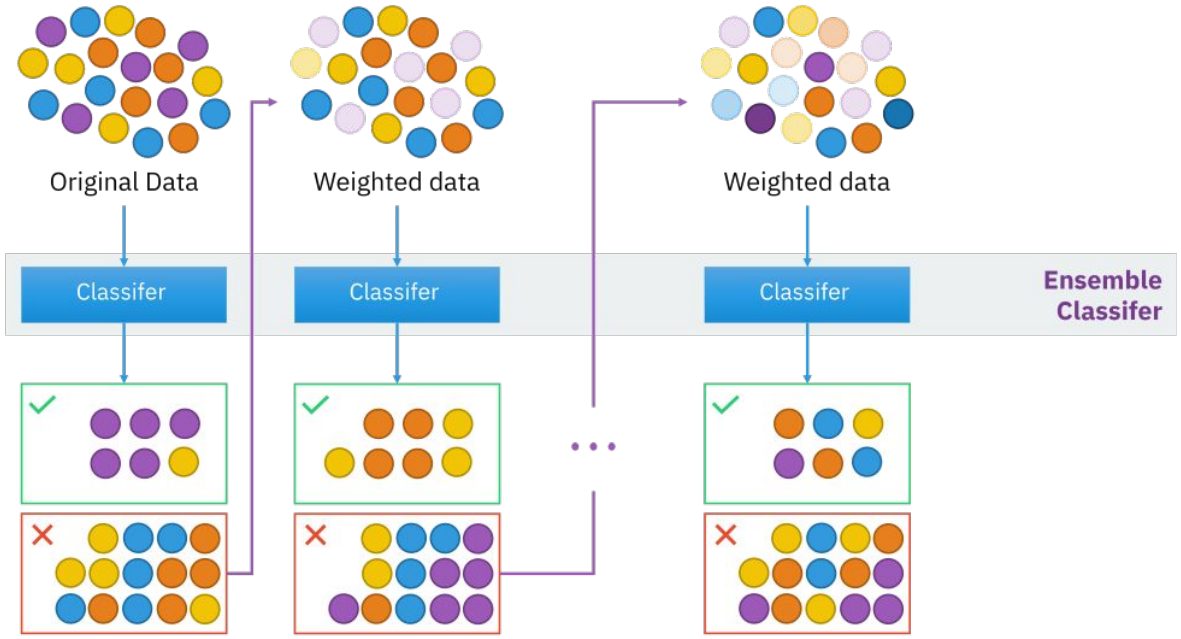
Bagging: combine models trained on random data subsets



Random Forest

https://commons.wikimedia.org/wiki/File:Ensemble_Bagging.svg

Boosting: train model on what the last model got wrong

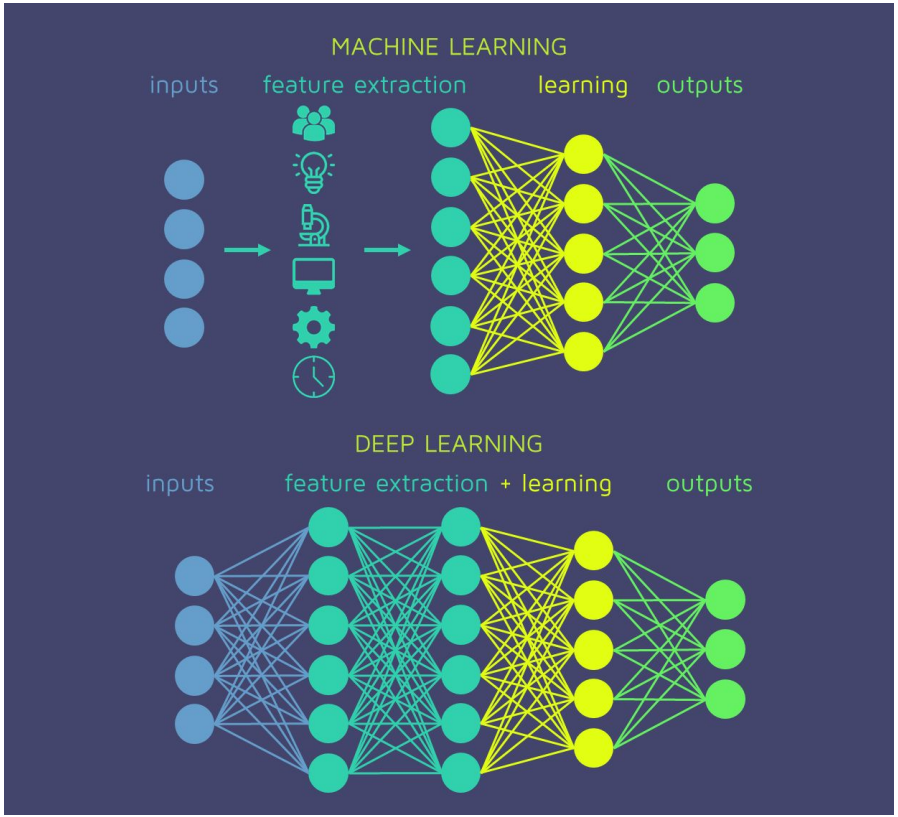


AdaBoost, Gradient Boosting, XGBoost

https://commons.wikimedia.org/wiki/File:Ensemble_Boosting.svg

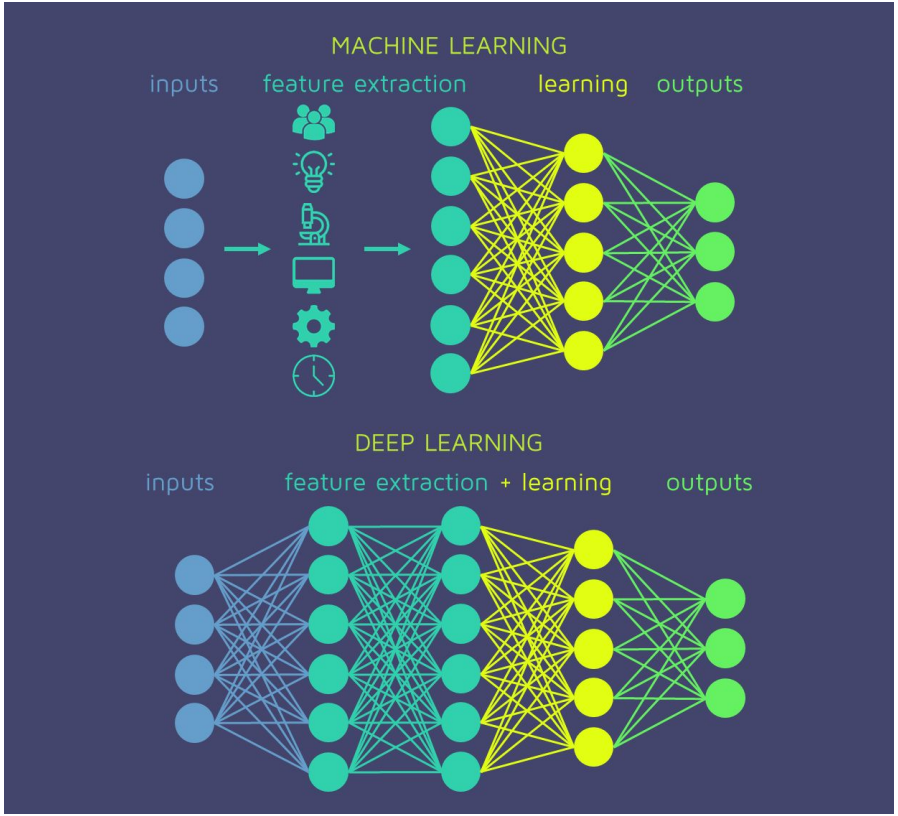
What is “Deep Learning”?

Deep Learning also learns a data representation for task

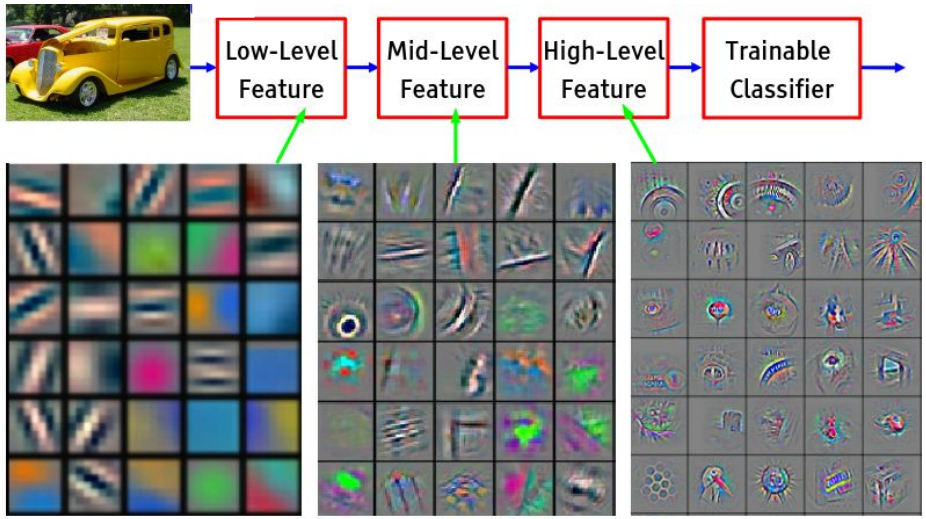


https://aithority.com/machine-learning/21-key-differences-of-deep-learning-vs-machine-learning/

Deep Learning also learns a data representation for task



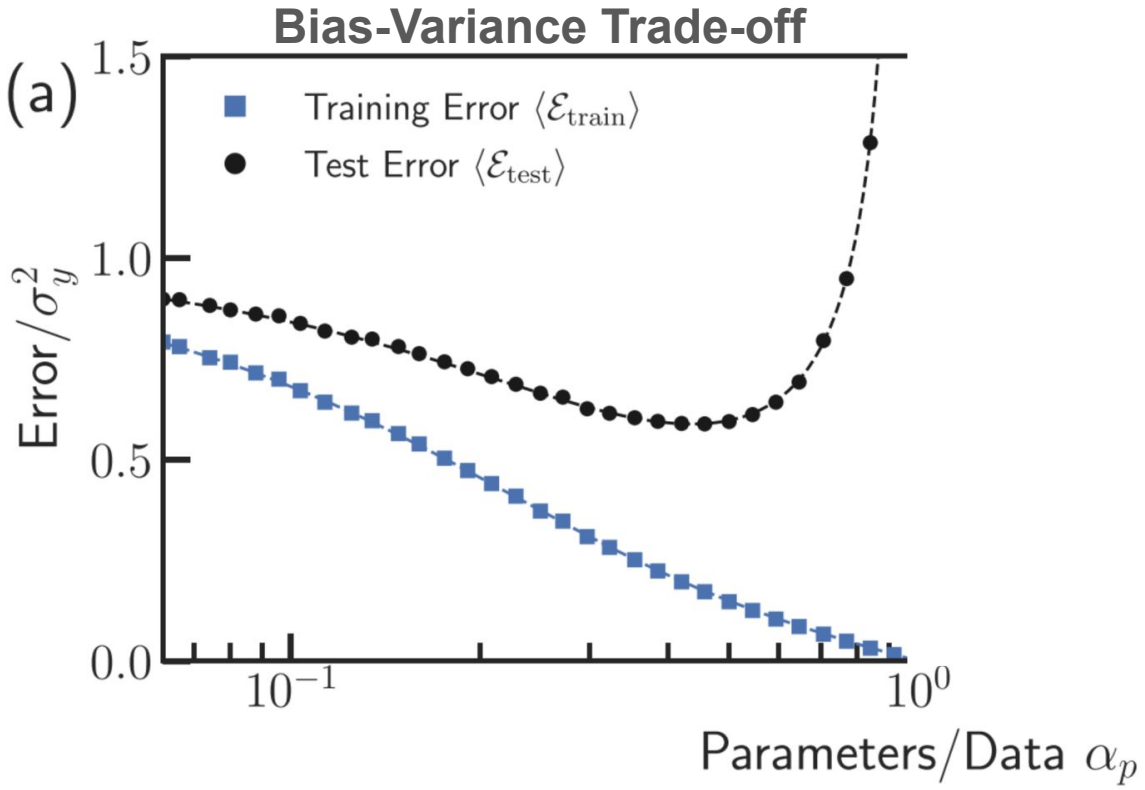
https://aithority.com/machine-learning/21-key-differences-of-deep-learning-vs-machine-learning/



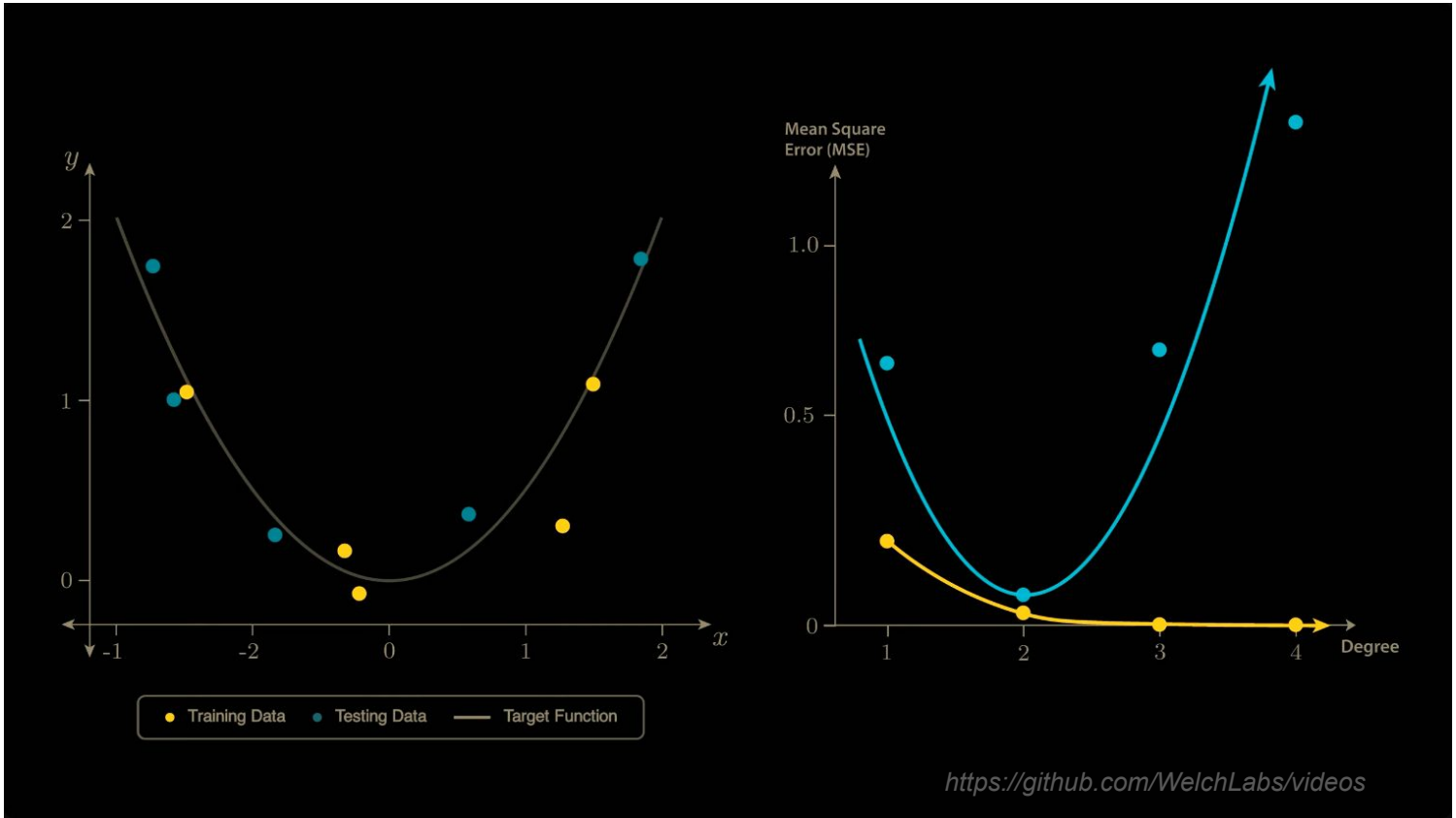
https://cs.nyu.edu/~yann/talks/lecun-ranzato-icml2013.pdf

Won’t massively complex models just lead to overfitting?

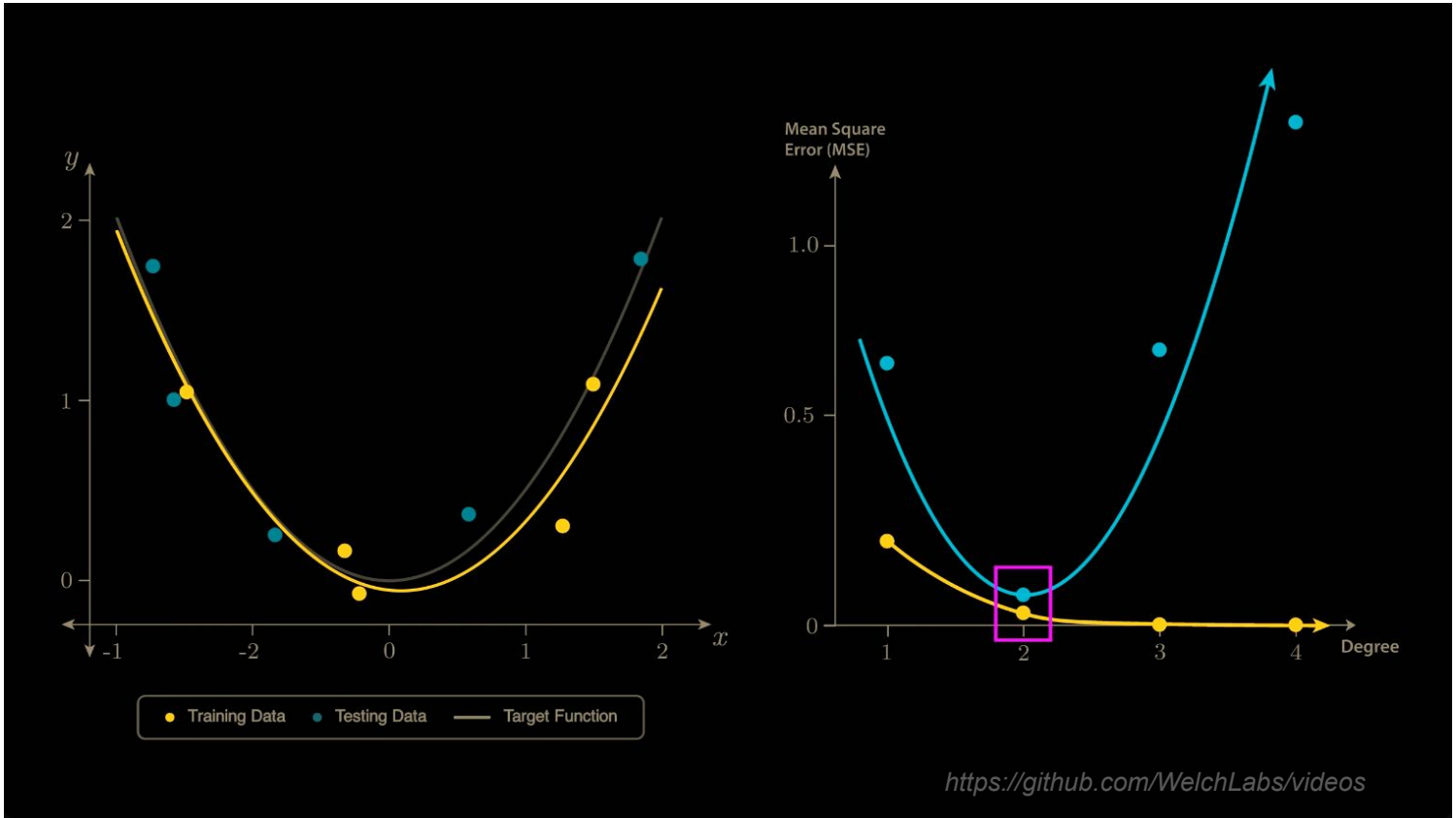
High dimensional models are counter-intuitive



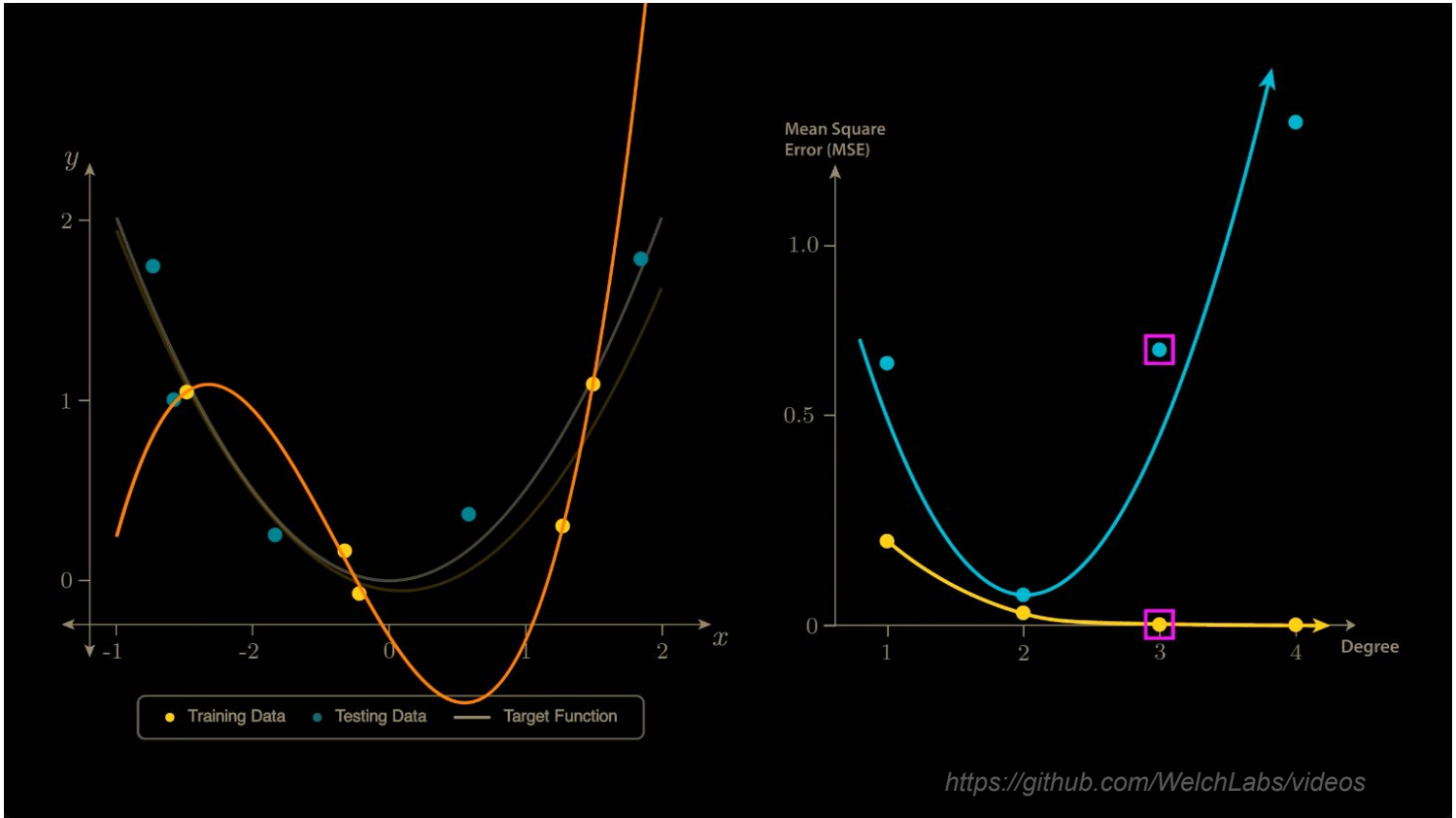
Bias-variance trade-off as model “complexity” increases



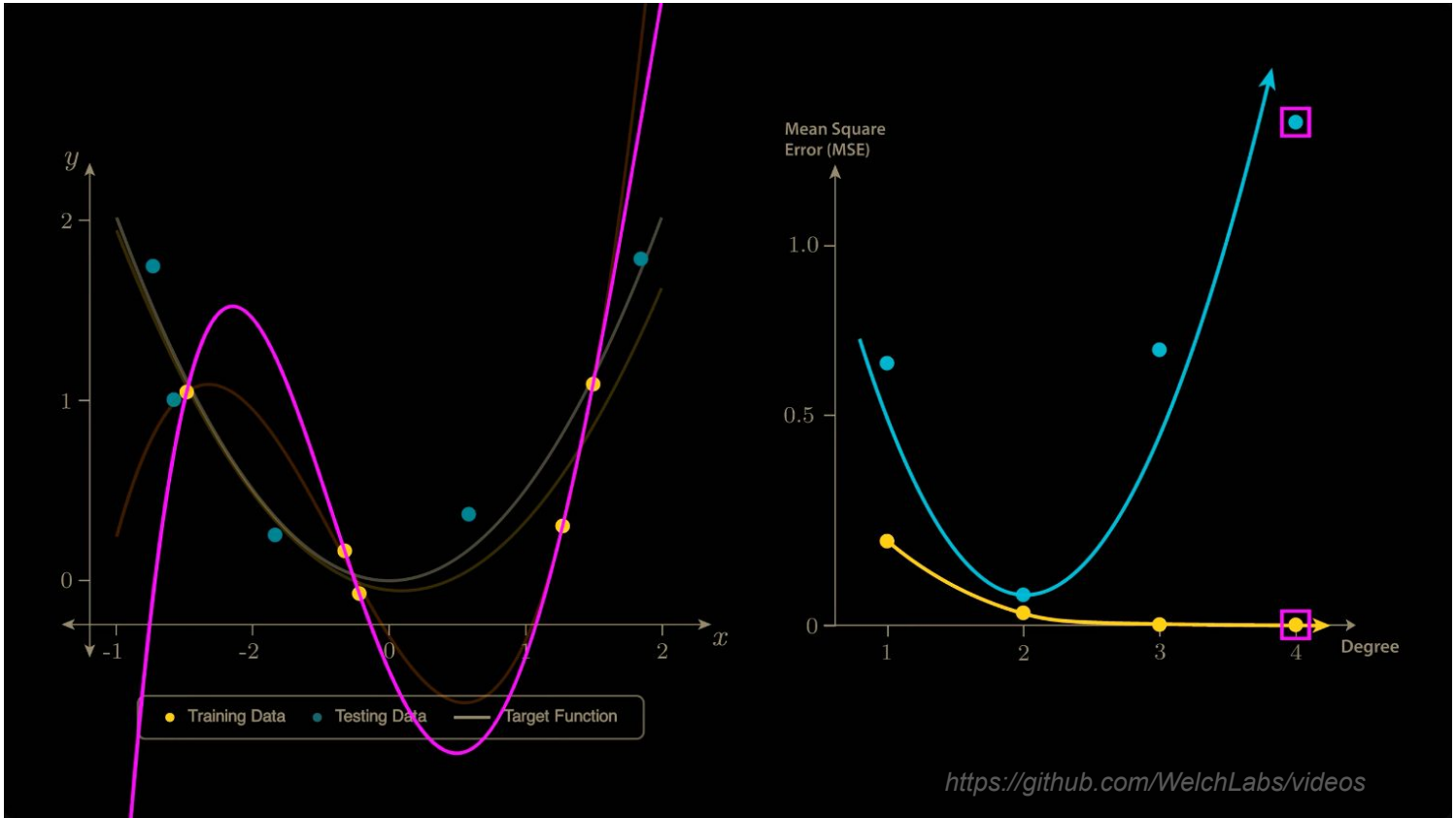
Bias-variance trade-off as model “complexity” increases



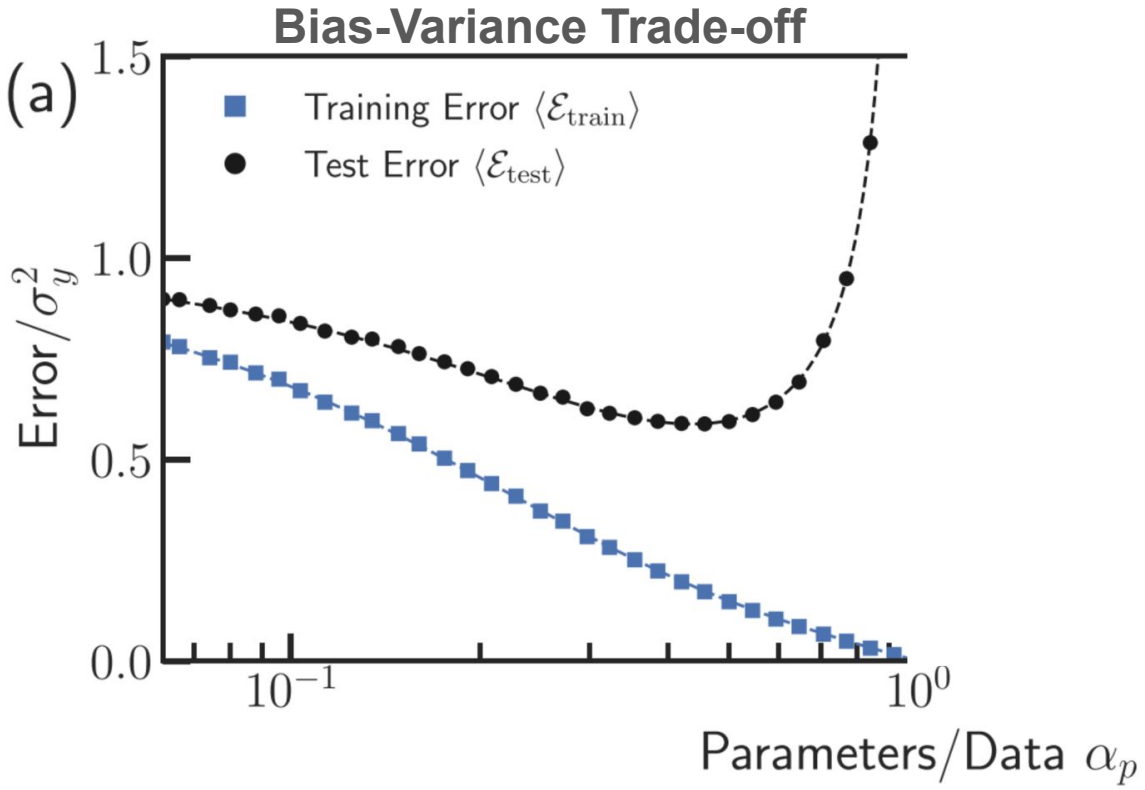
Bias-variance trade-off as model “complexity” increases



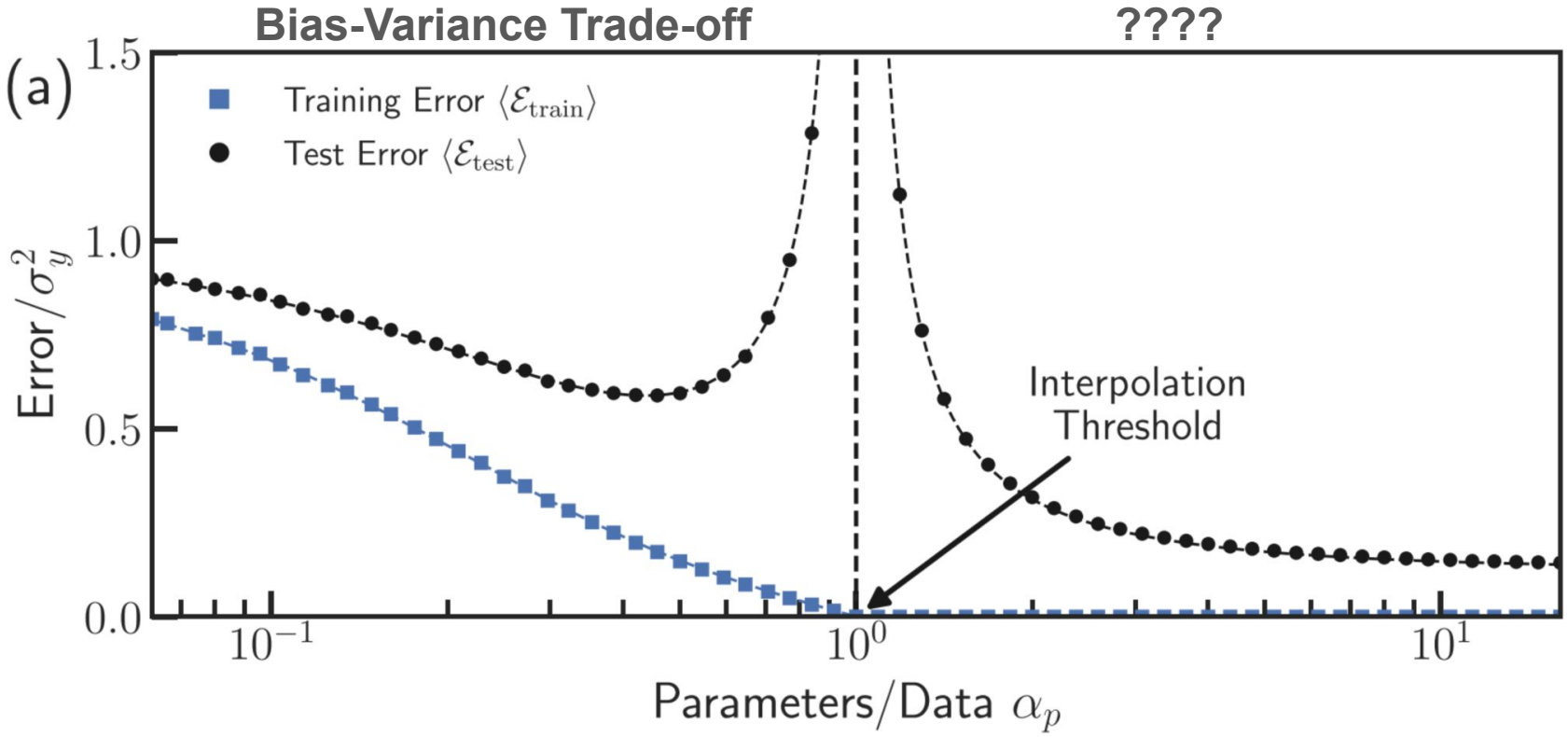
Bias-variance trade-off as model “complexity” increases



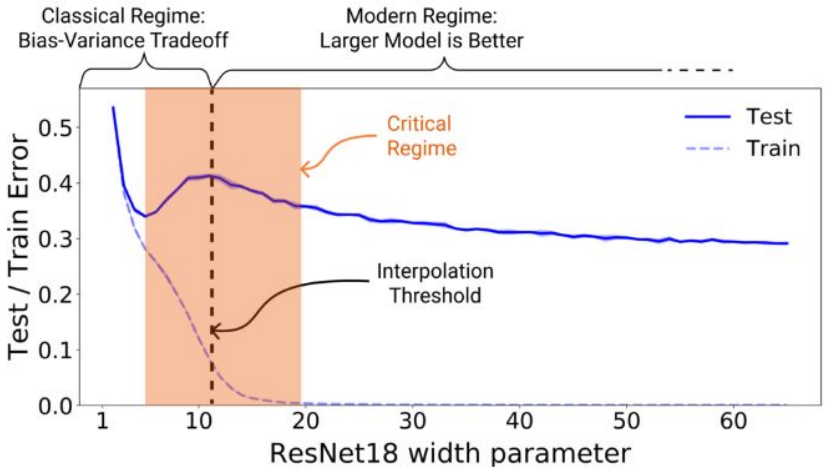
High dimensional models are counter-intuitive

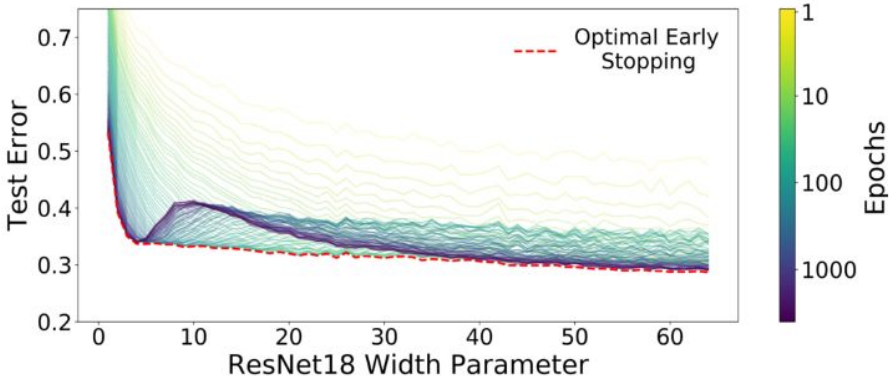


“Double Descent” beyond interpolation threshold



Empirically observed double descent with training length in large neural networks

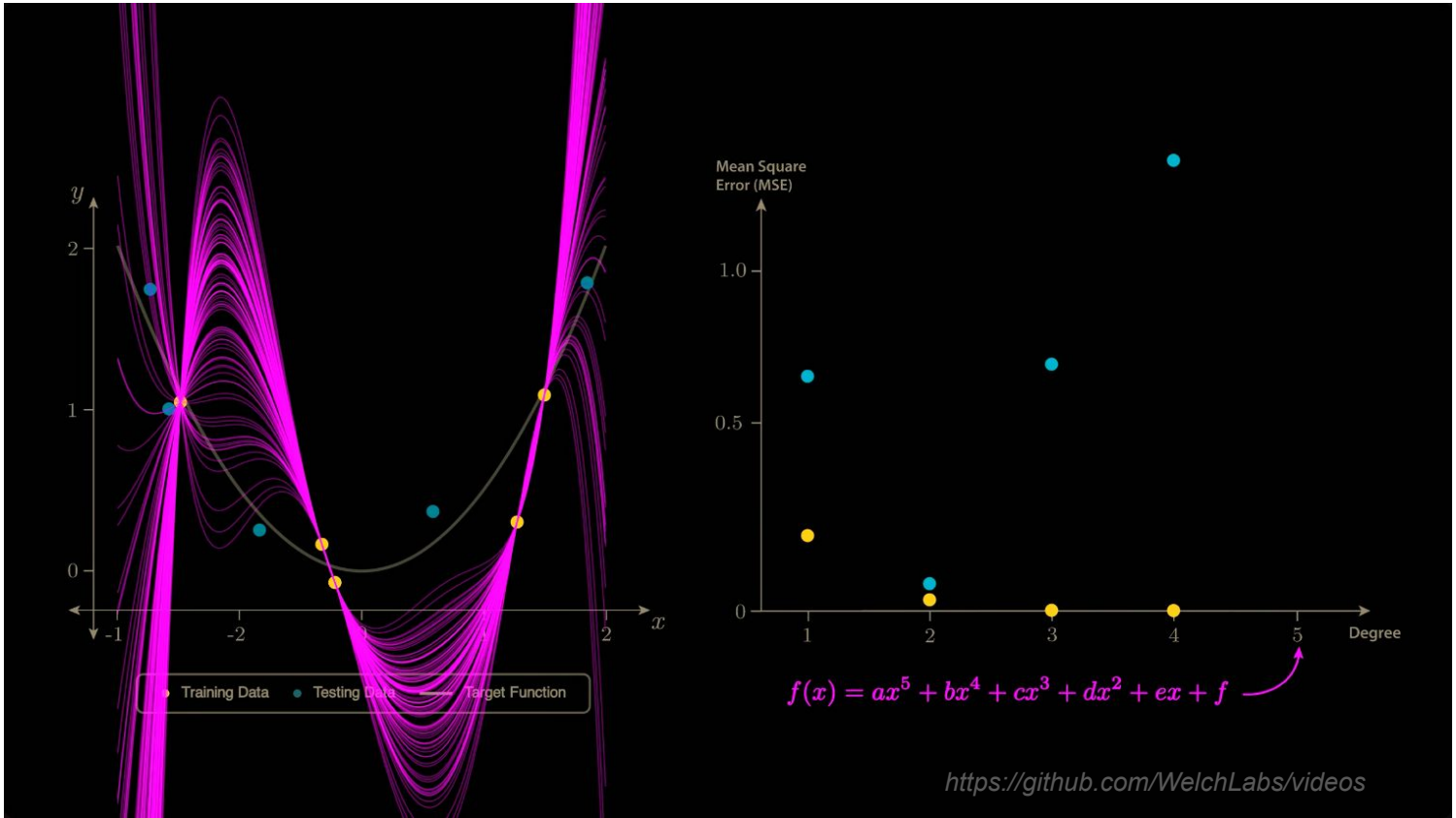




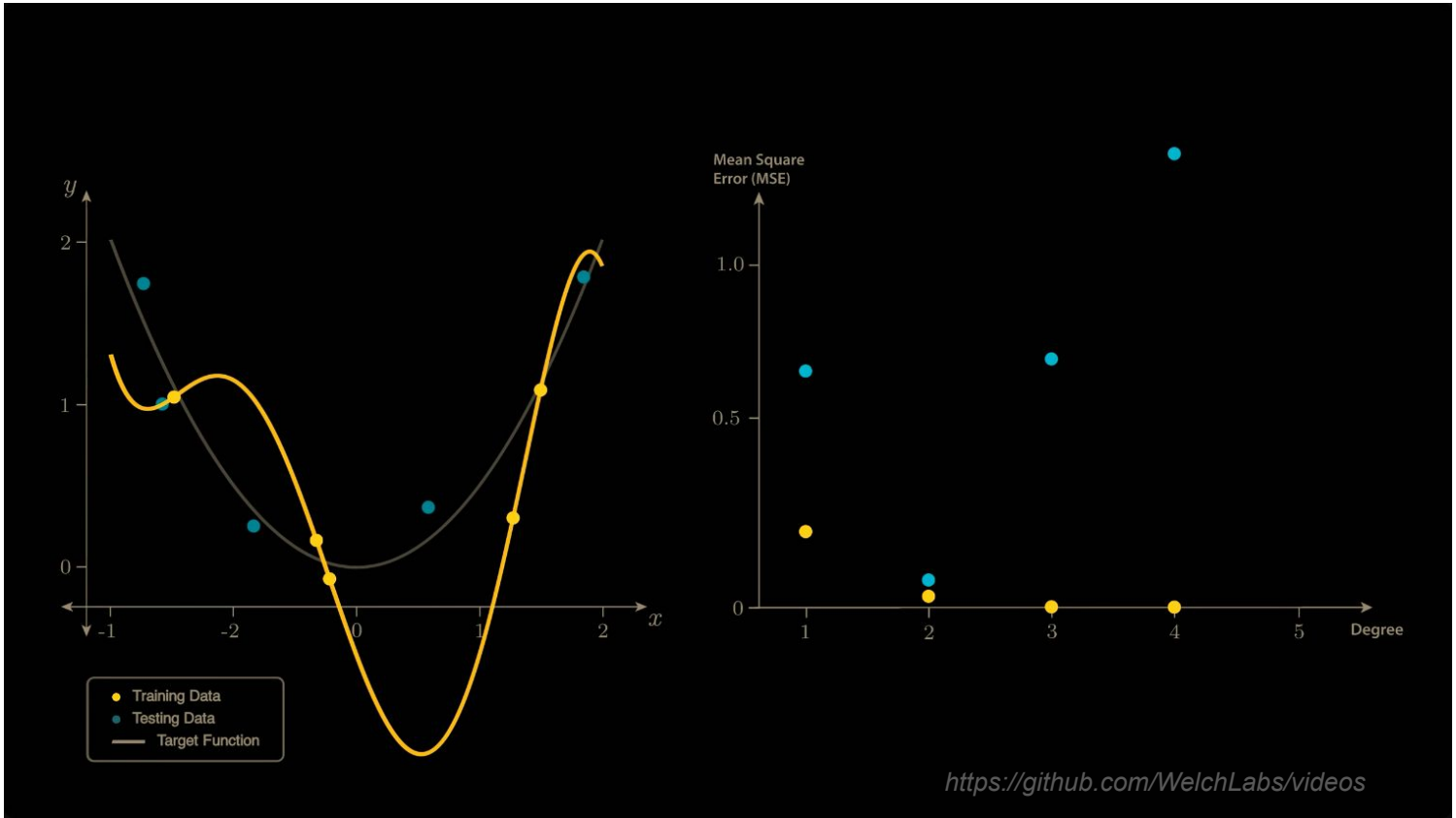
Nakkiran, Preetum, et al. "Deep double descent: Where bigger models and more data hurt." Journal of Statistical Mechanics: Theory and Experiment 2021.12 (2021): 124003.

Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

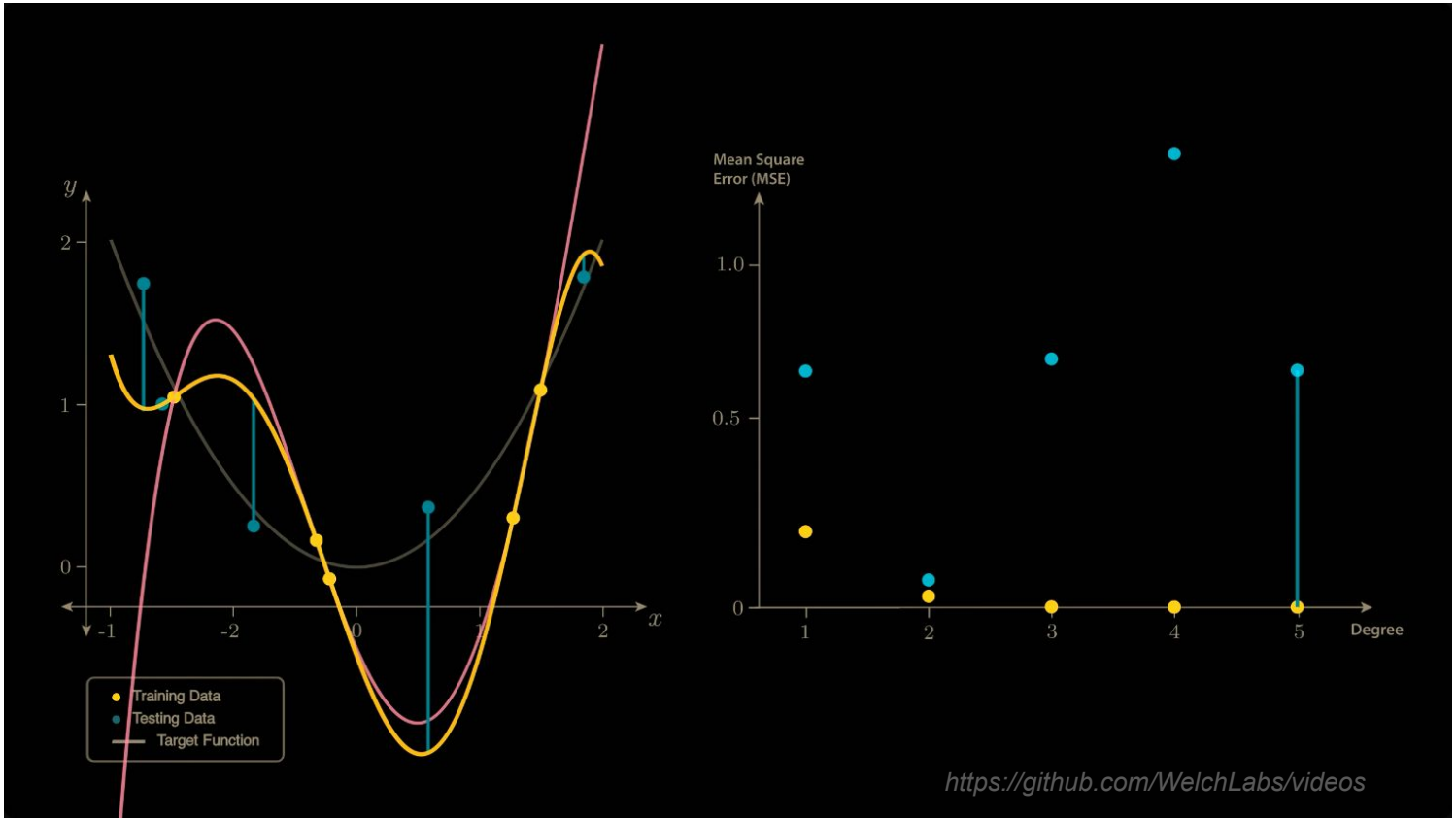
Regularised polynomial regression can show double descent



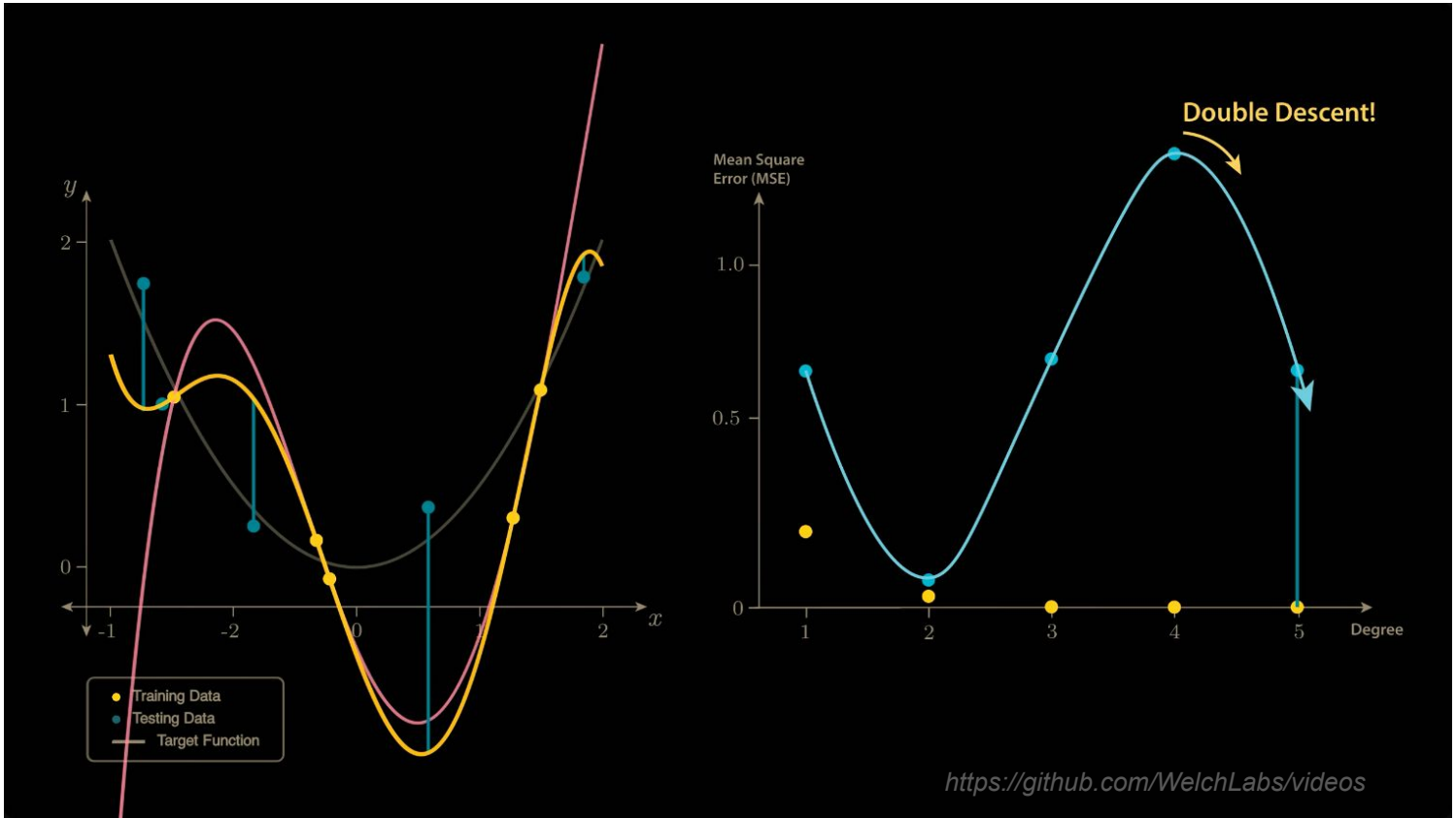
Regularised polynomial regression can show double descent



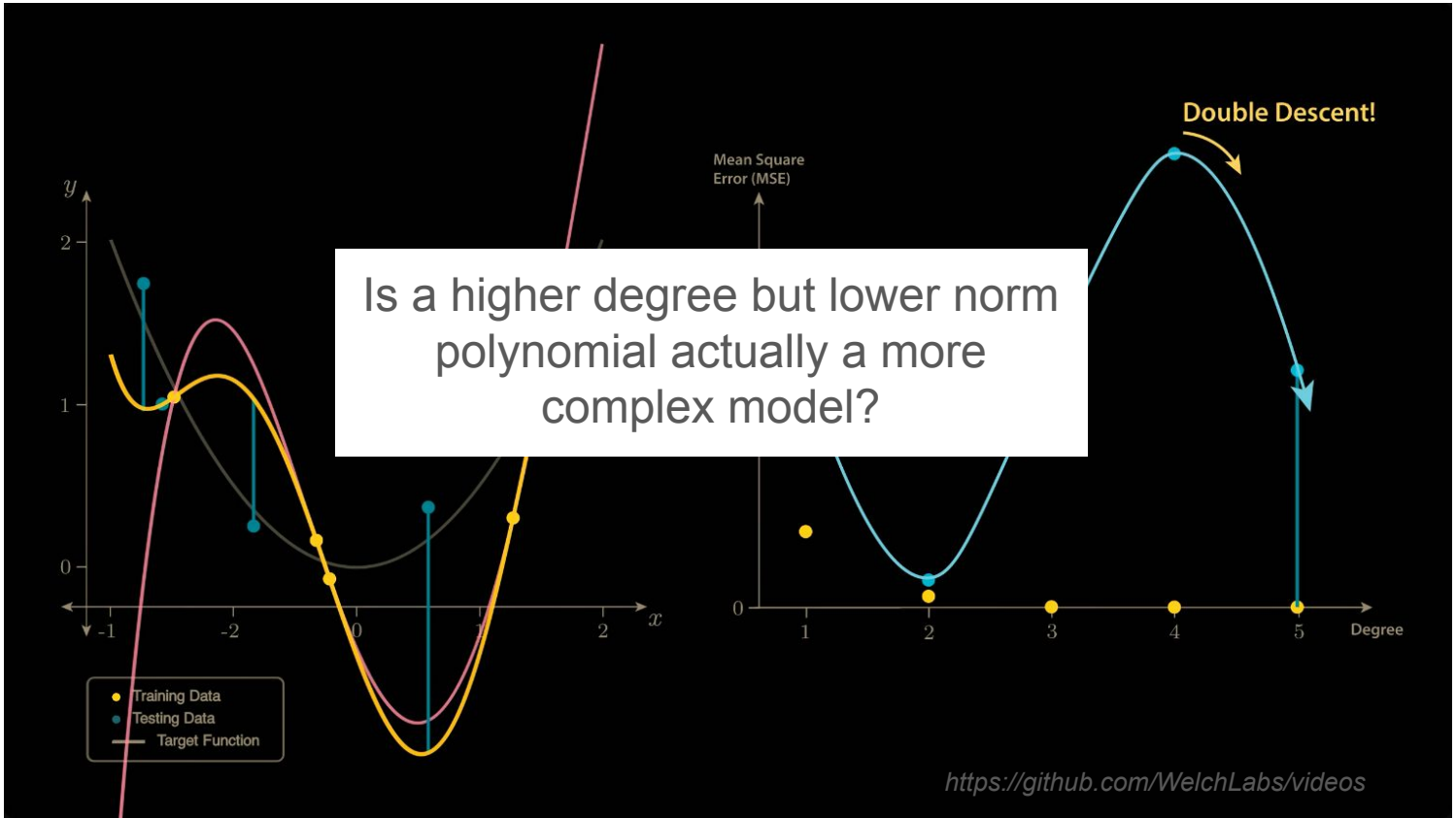
Regularised polynomial regression can show double descent



Regularised polynomial regression can show double descent



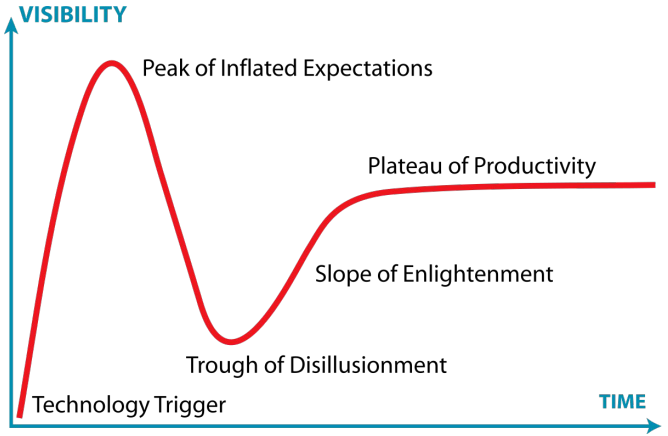
Regularised polynomial regression can show double descent



So, does health data science solve all my problems?

(Some) Challenges of Health Data Science (and AI/ML…)

- Lots of hype
- Lots of grifters
- Data quality issues
- Contextual/Metadata quality issues
- Regulatory challenges
- Influence of US health system
- Ethical pitfalls
- Treatment to the mean
- Knowledge Translation and Operations: **Hard**

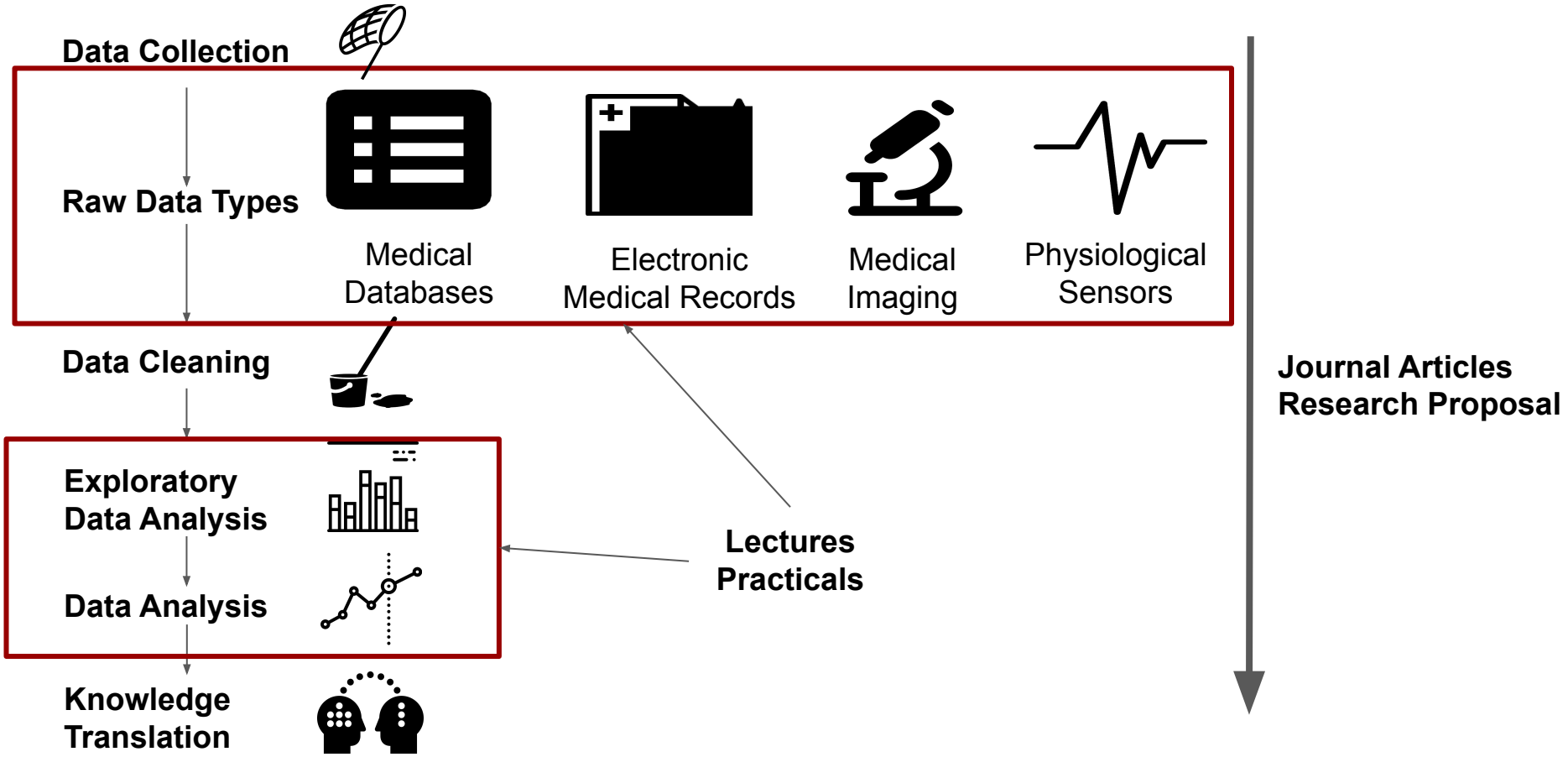


https://www.r-bloggers.com/2019/08/new-course-learn-advanced-data-cleaning-in-r/

https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner_Hype_Cycle.svg
https://commons.wikimedia.org/wiki/File:Clark_Stanley%27s_Snake_Oil_Liniment.png

Where can I learn more?

EPAH6410: Applied Research in Health Data Science



Summary

- What is health data science and how does it differ from traditional analyses of secondary data?
  - Integration with data ecosystem
  - Unstructured & multi-modal data
  - Centers exploratory data analyses
  - Incorporates Inductive methodologies
  - AI ⊃ Machine Learning ⊃ Deep Learning
- What are the main types of Machine Learning?
  - Unsupervised Learning
    - Anomaly Detection
    - Clustering (K-means)
    - Dimensionality Reduction (UMAP/t-SNE)
    - Counter-Intuitive properties of high dimensional data
  - Supervised Learning
    - Test-Train Split vital and Cross-Validation useful
    - Different models can fit different decision boundaries (Logistic, SVM, Random Forests, Neural Networks)
    - Combining Weak Predictors can be effective: Bagging (Random Forest) and Boosting (XGBoost)
    - Deep Learning of feature representations
    - Bias-Variance trade-off and the counterintuitive finding of double descent

Questions?

Terminology: Every Definition Is Wrong But Some Are Useful

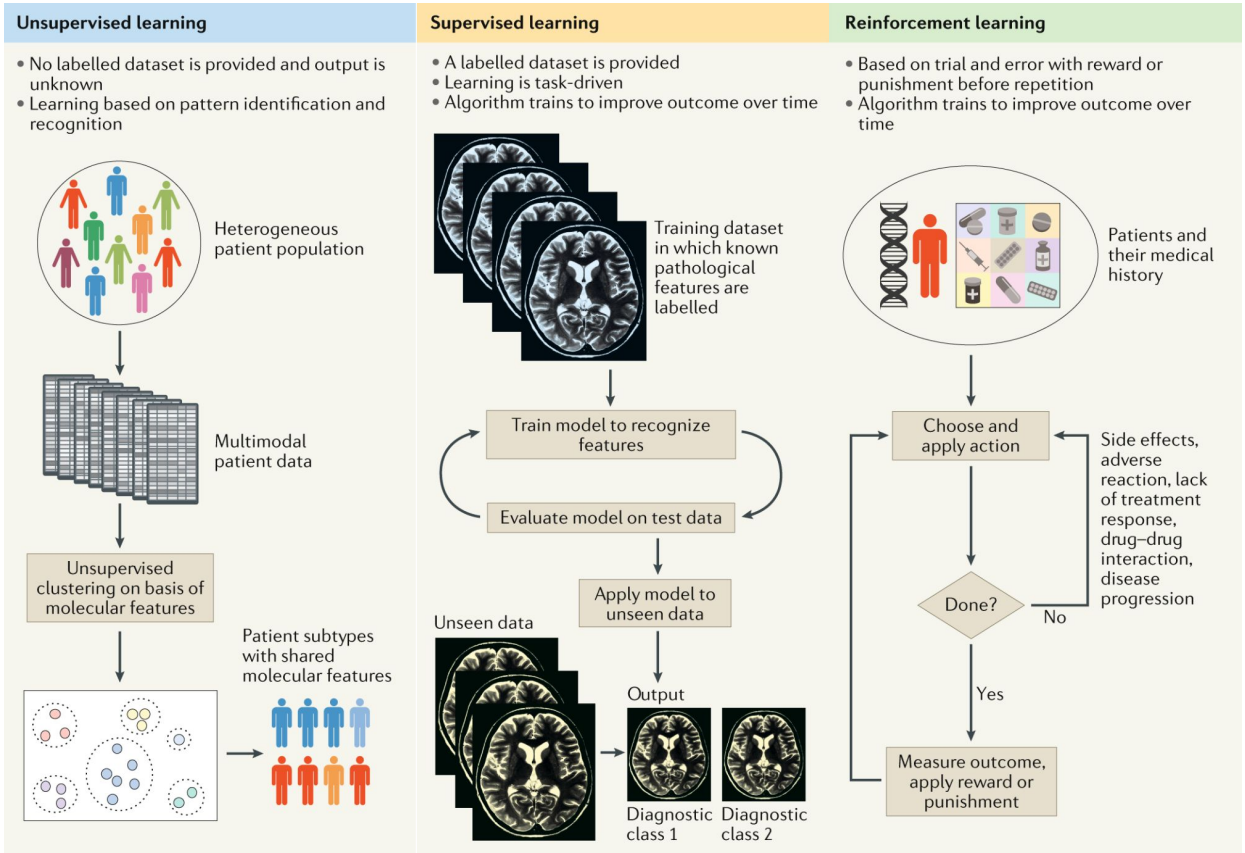
- Humans love taxonomies and ontologies.
- All systems oversimplify: incomplete, overlapping, and/or contradictory terms.
- Best system for you is the ones whose oversimplifications matter least to you.

- Are novelists and journalists both just writers? Depends who's asking.

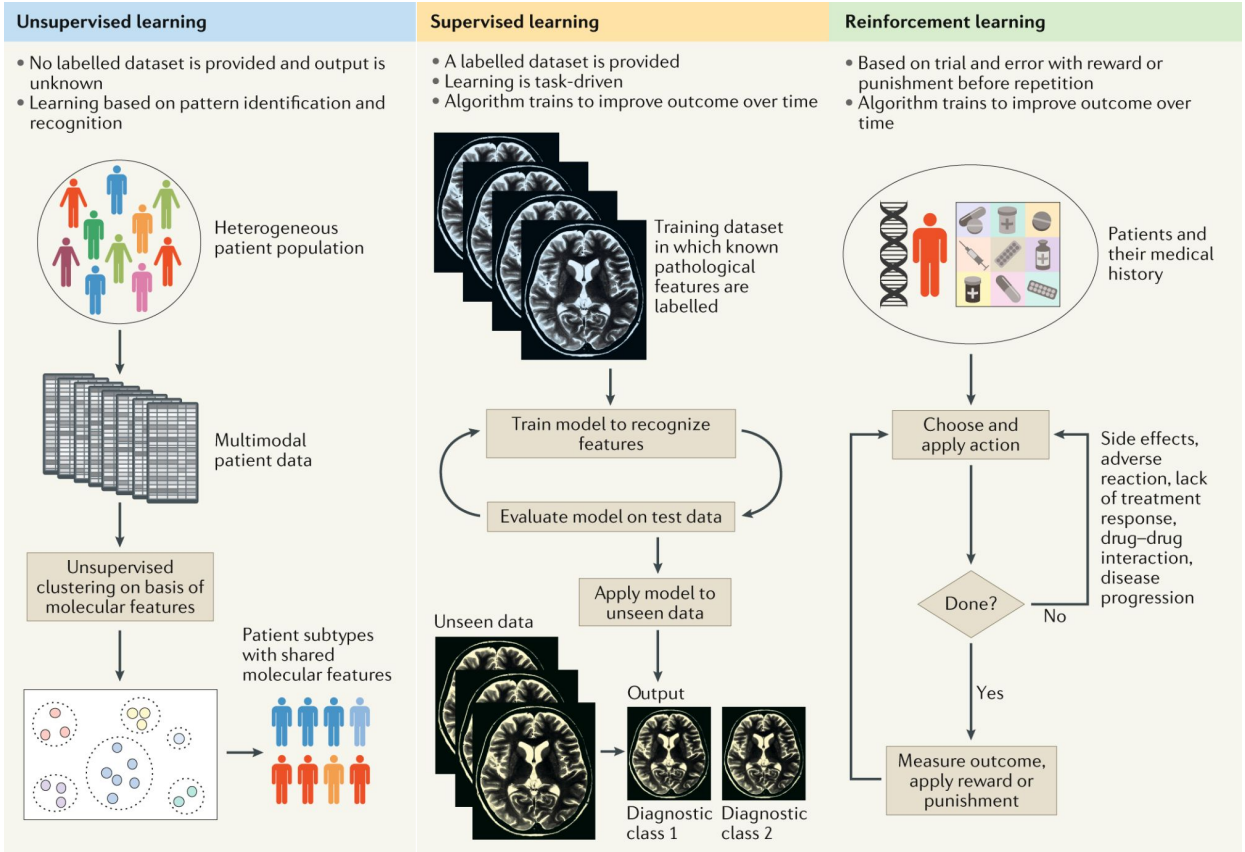
- **Generally**: communication more important than specific terminology.

u/a_wandering_chemist

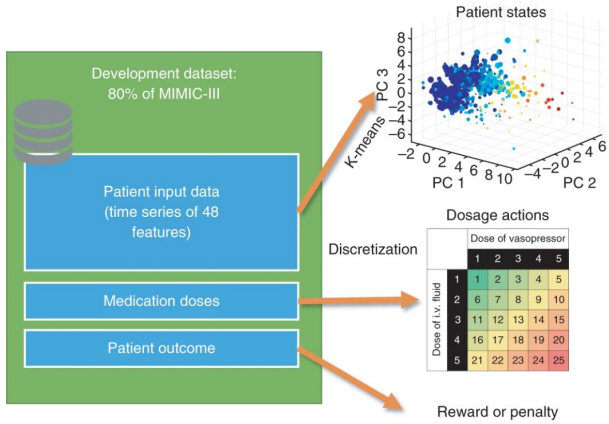
Types of Machine Learning



Types of Machine Learning

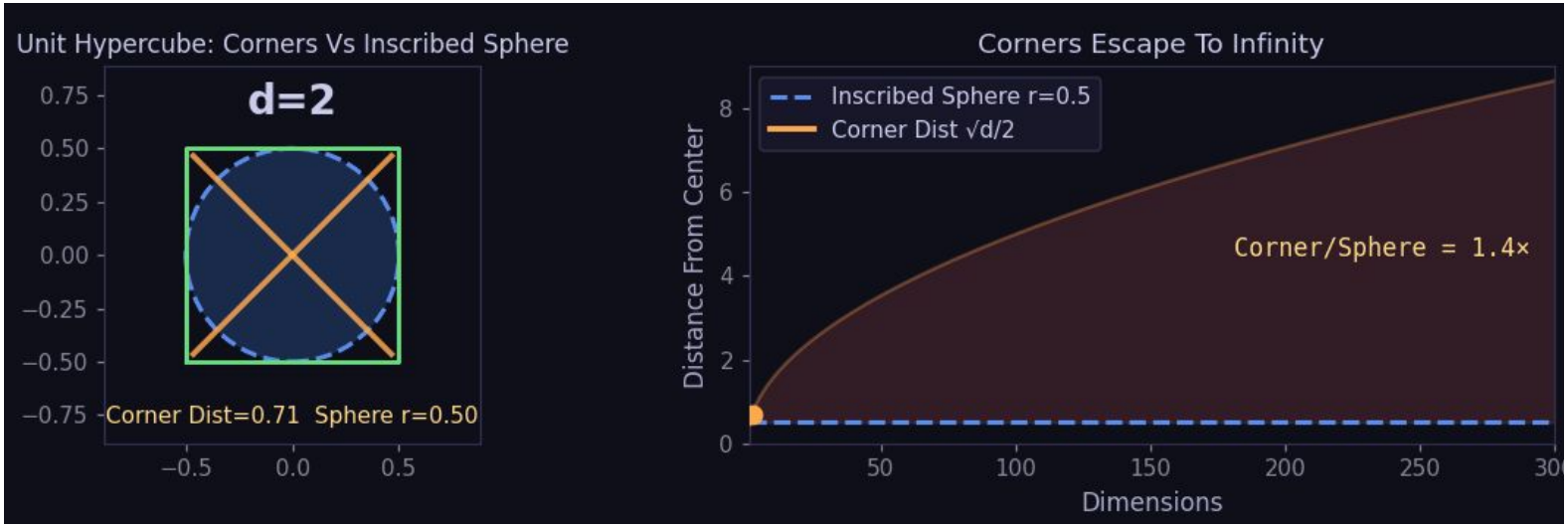


Sepsis Treatment Policy Optimisation



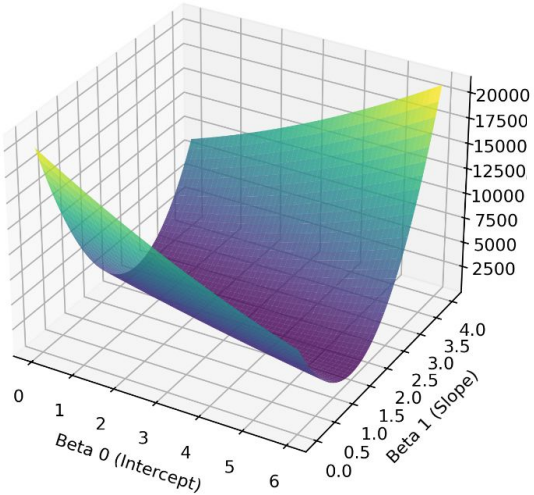
*Komorowski, M., Celi, L.A., Badawi, O. et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med **24**, 1716–1720 (2018). https://doi.org/10.1038/s41591-018-0213-5*

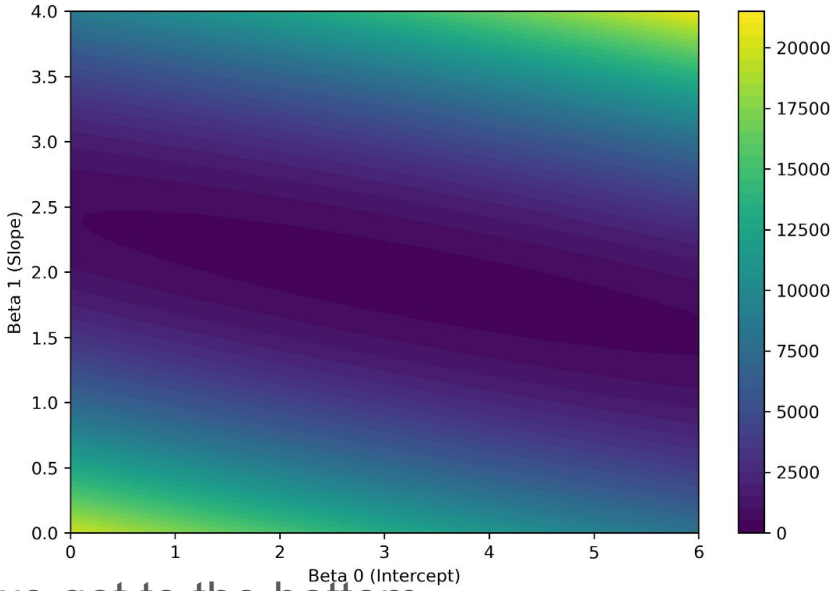
Weird distributions: Hypercubes get infinitely spiky



Relationship between loss and model parameters

$$\hat{y} = \frac{1}{1 + e^{-(\beta^\top \mathbf{x})}}$$





Minimise loss: go down this surface until we get to the bottom.

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

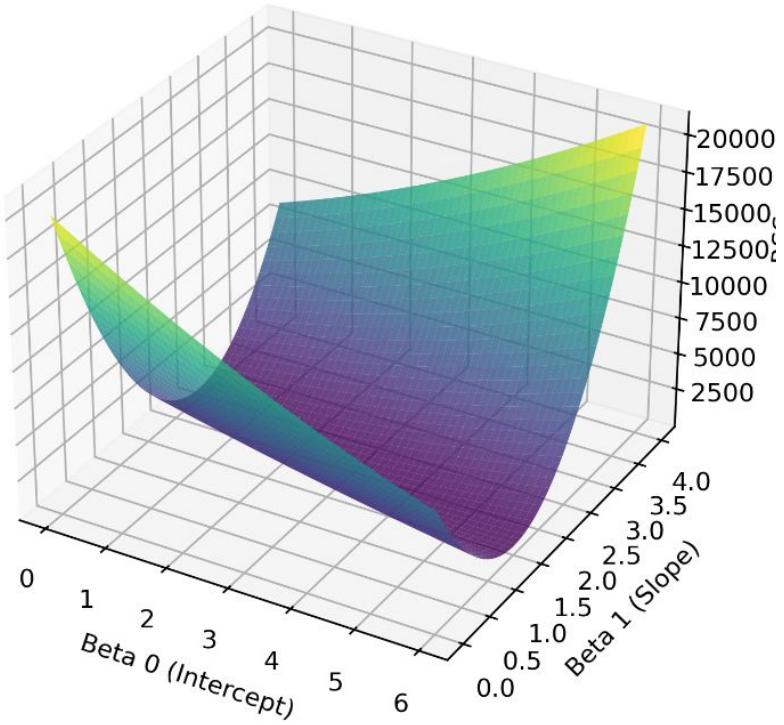
But how do we update model parameters to achieve this?

Gradient descent using partial derivatives of loss

Calculate partial derivative of:

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

with respect to each β parameter:



Gradient descent using partial derivatives of loss

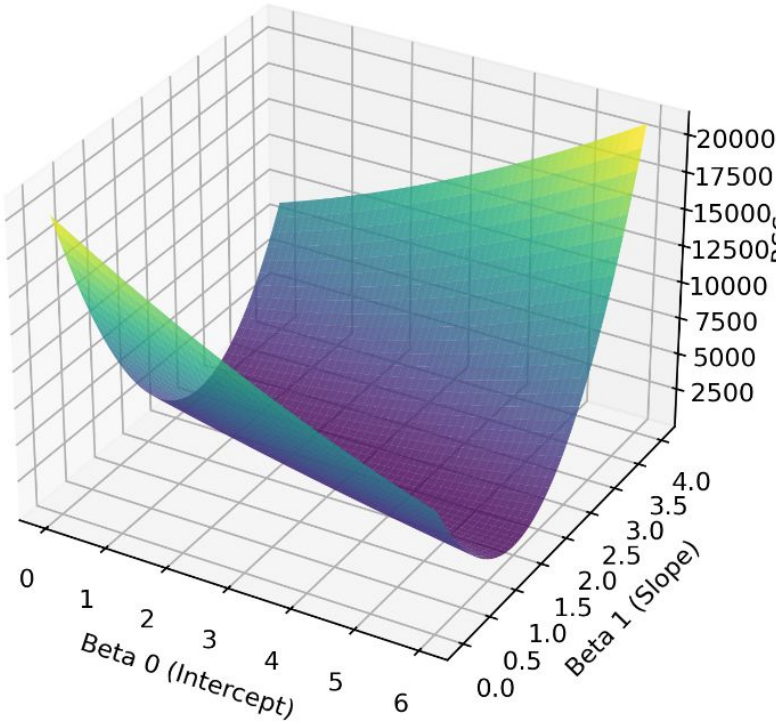
Calculate partial derivative of:

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

with respect to each β parameter:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)x_{i,0}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_1} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)x_{i,1}$$



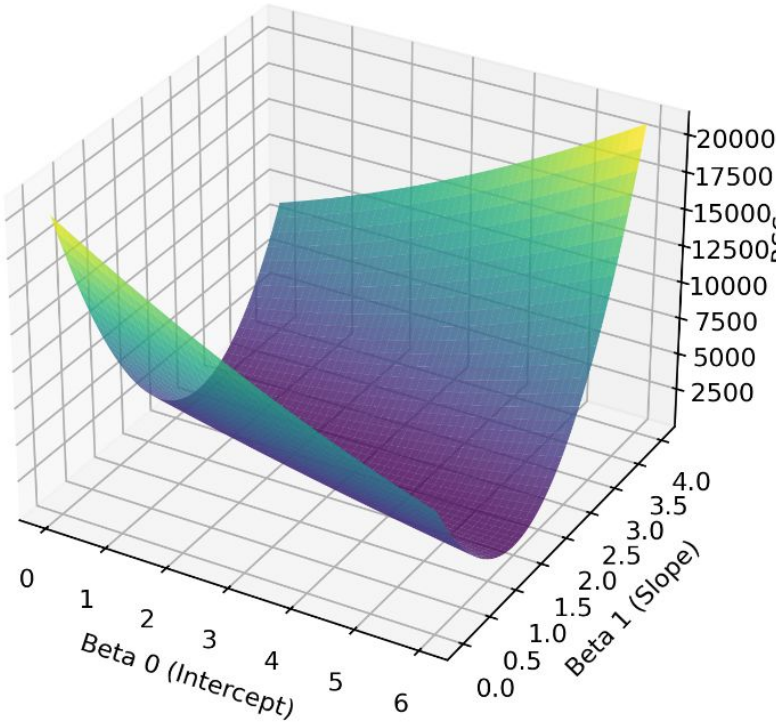
Gradient descent using partial derivatives of loss

Calculate partial derivative of:

$$\mathcal{L}(\beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

with respect to each β parameter:

$$\nabla_\beta \mathcal{L} = \frac{1}{N} X^\top (\hat{y} - y)$$



Learning Rate is an important parameter

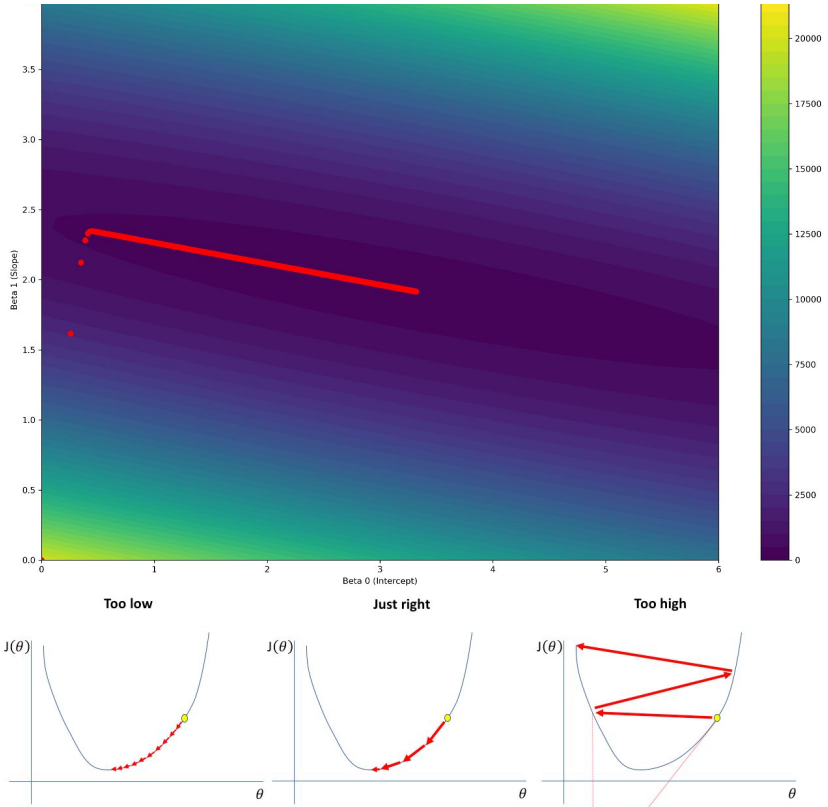
Learning rate is essentially relative “step” size when sliding down the gradient.

Learning rate too small:

- Convergence becomes extremely slow
- Stuck in local minima
- Can run out of iterations!

Learning rate too large:

- Overshoot optimal value and oscillate
- Failure to converge
- Float overflows



Autoencoder: find reduction by reconstructing original data

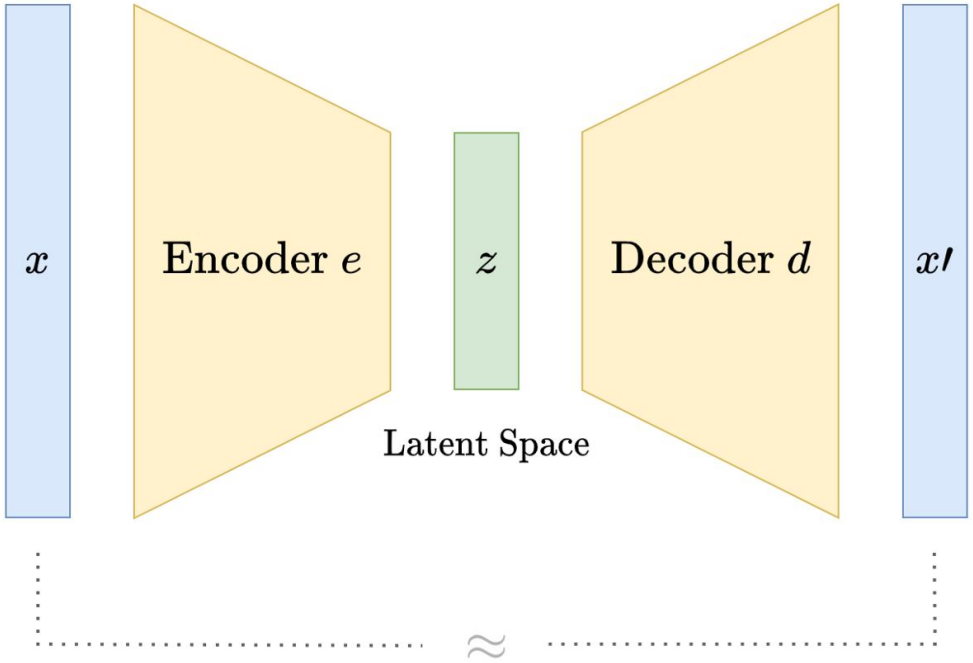
For each observation in dataset:

Find parameters:

e that compress **x** to dim(**z**)

d that recover **x** from **z**

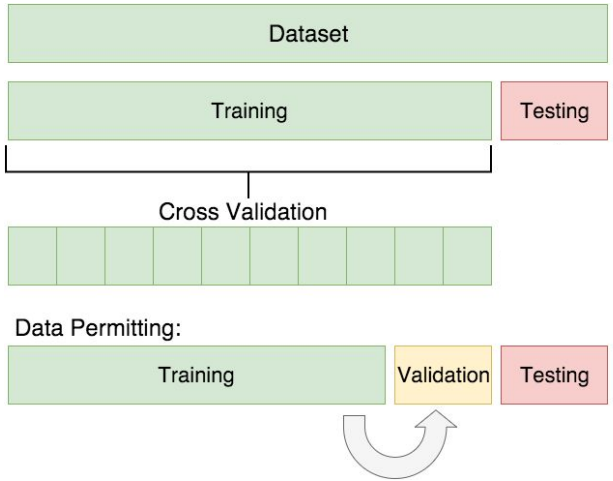
Loss: difference between input **x** and reconstructed output **x'**



"Self-supervised"

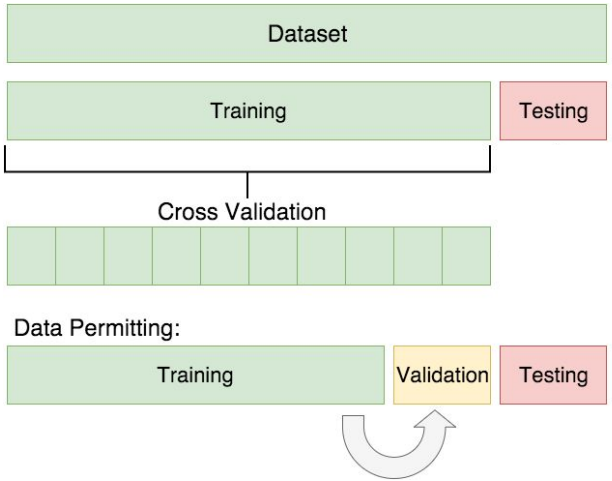
Supervised Learning: Classification and Regression

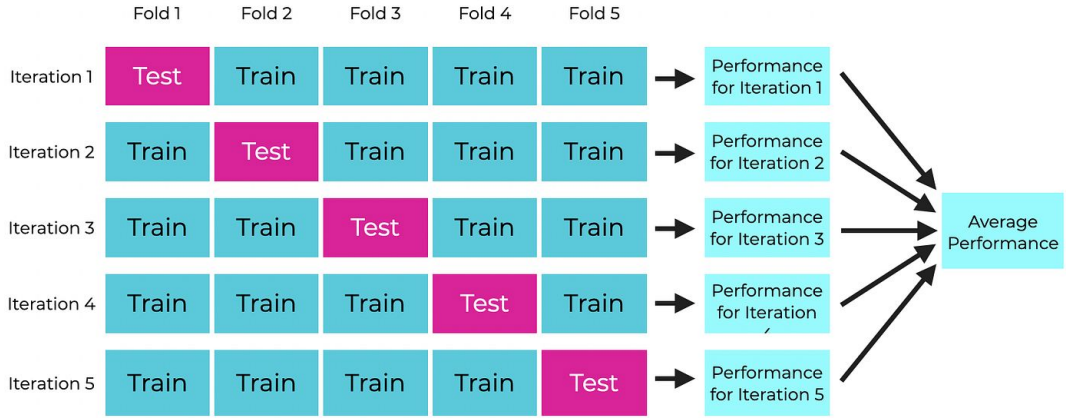
Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	38.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%



Supervised Learning: Classification and Regression

Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	38.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%





Supervised Learning: Classification and Regression

Patient ID	Age	Sex	BMI	Systolic BP (mmHg)	Diastolic BP (mmHg)	Fasting Glucose (mmol/L)	HbA1c (%)	Cholesterol (mmol/L)	Smoker	Family History Diabetes	Diagnosis (Classification)	10-yr CVD Risk (Regression)
P001	54	M	28.3	138	88	6.2	6.8	5.1	Yes	Yes	Diabetic	18.4%
P003	67	M	31.7	155	95	7.8	7.4	6.3	Yes	Yes	Diabetic	31.2%
P004	35	F	25.6	122	80	5.3	5.5	4.2	No	Yes	Healthy	5.8%
P006	72	F	27.4	160	97	8.4	8.1	6.9	Yes	Yes	Diabetic	38.5%
P007	29	M	23.8	115	73	4.7	5	4	No	No	Healthy	2.3%
P008	63	F	33.2	148	93	7.1	7	5.5	No	Yes	Diabetic	25.6%