Introduction to machine learning

Master SID

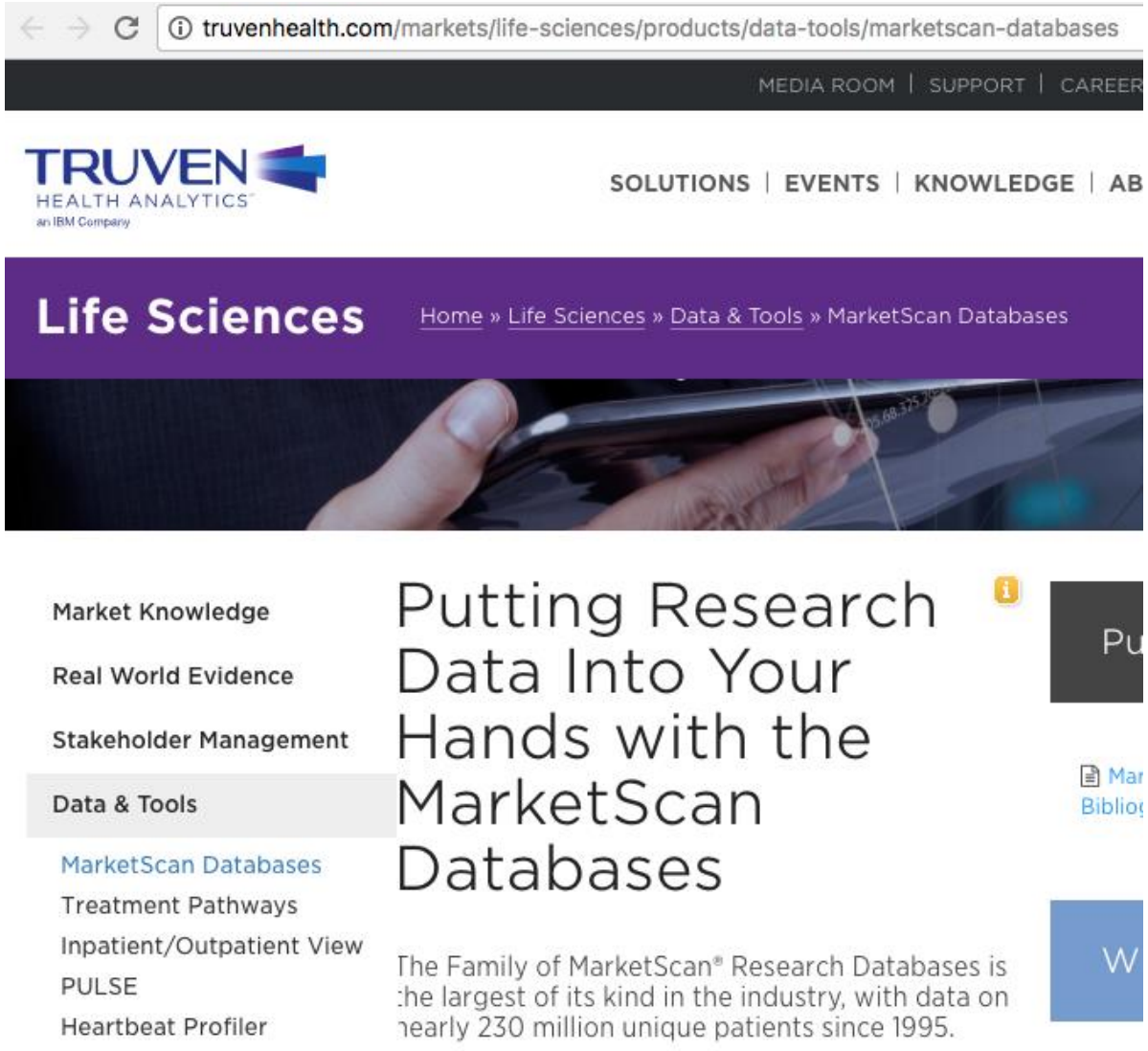
Machine Learning

Raquel Urena – raquel.urena@univ-amu.fr

SESSTIM, Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

Why Machine Learning now?

- Large datasets

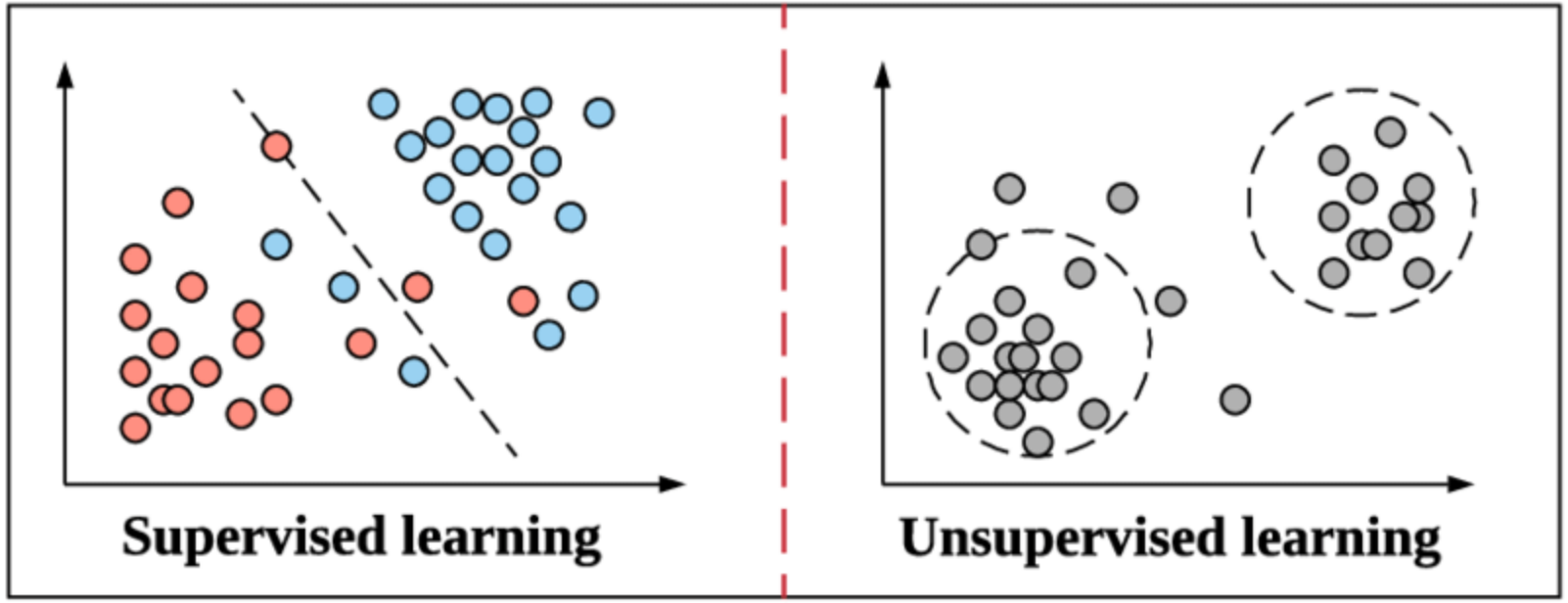


Why Machine Learning now?

- Diversity of digital health data

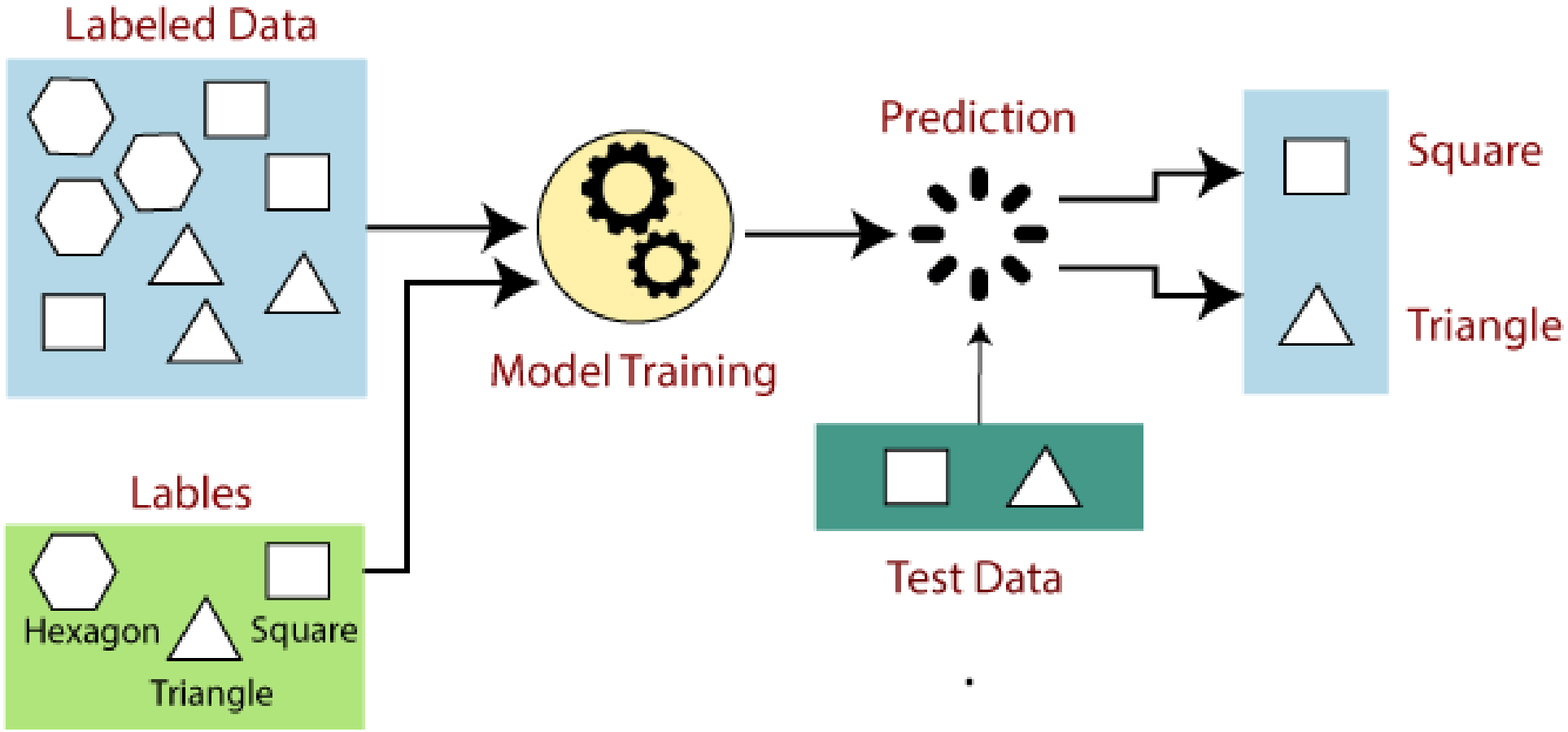
What is Machine Learning?

- "The field of study that gives computers the ability to learn without being explicitly programmed." Arthur Samuel
- "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
- Example: playing checkers.
  - E = the experience of playing many games of checkers
  - T = the task of playing checkers.
  - P = the probability that the program will win the next game.
- In general, any machine learning prob
  - Supervised learning a
  - Unsupervised learning.
  - Semi-supervised learning, which uses a mix



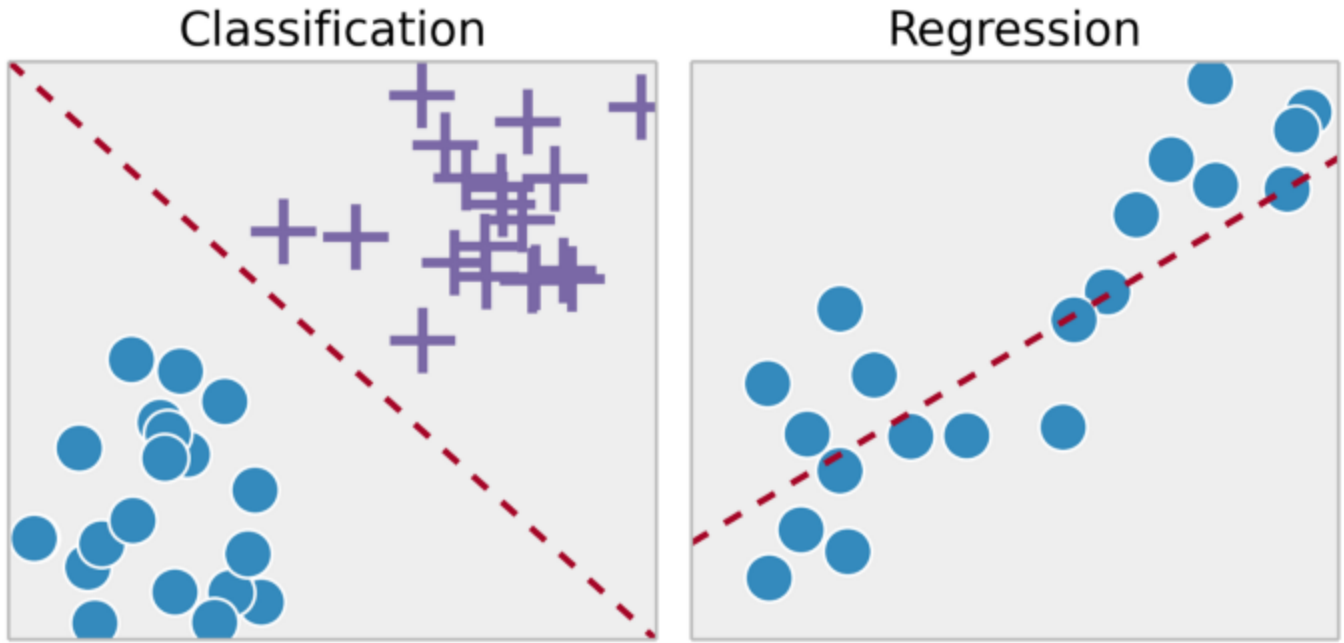
Supervised learning

- In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.



Supervised learning

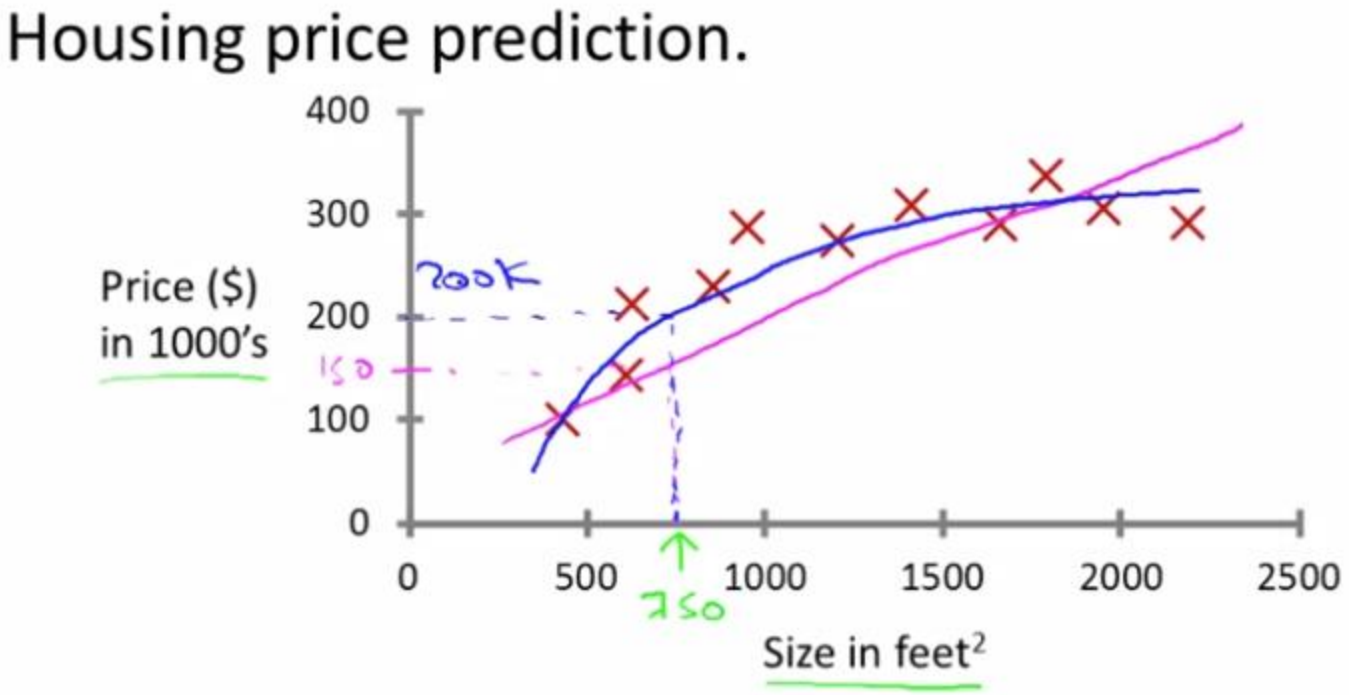
- Supervised learning problems are categorized into "regression" and "classification" problems:
  - In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function.
  - In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.



Examples of supervised learning

Example 1:

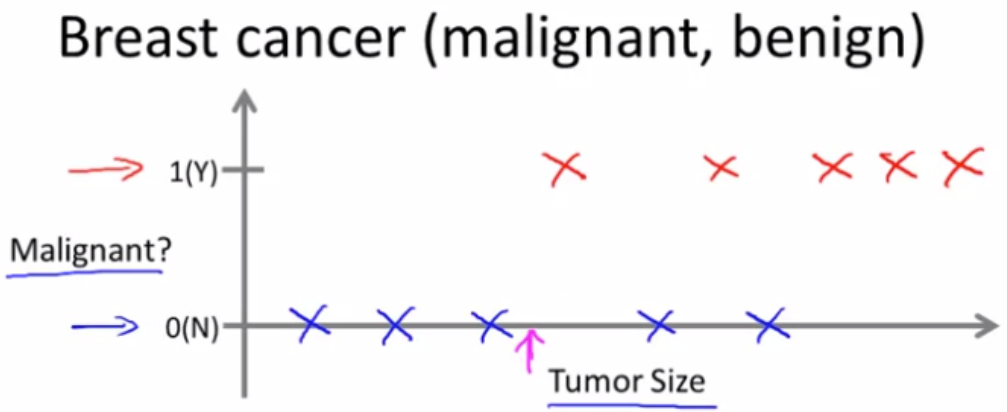
- Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

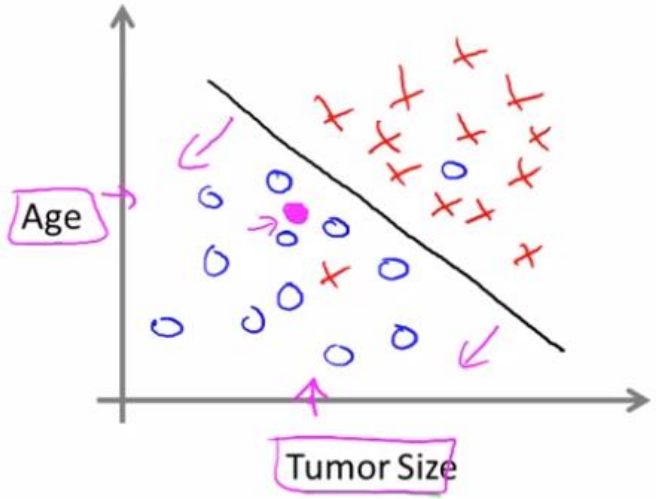


Examples of supervised learning

Example 2:

- (b) Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.





- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

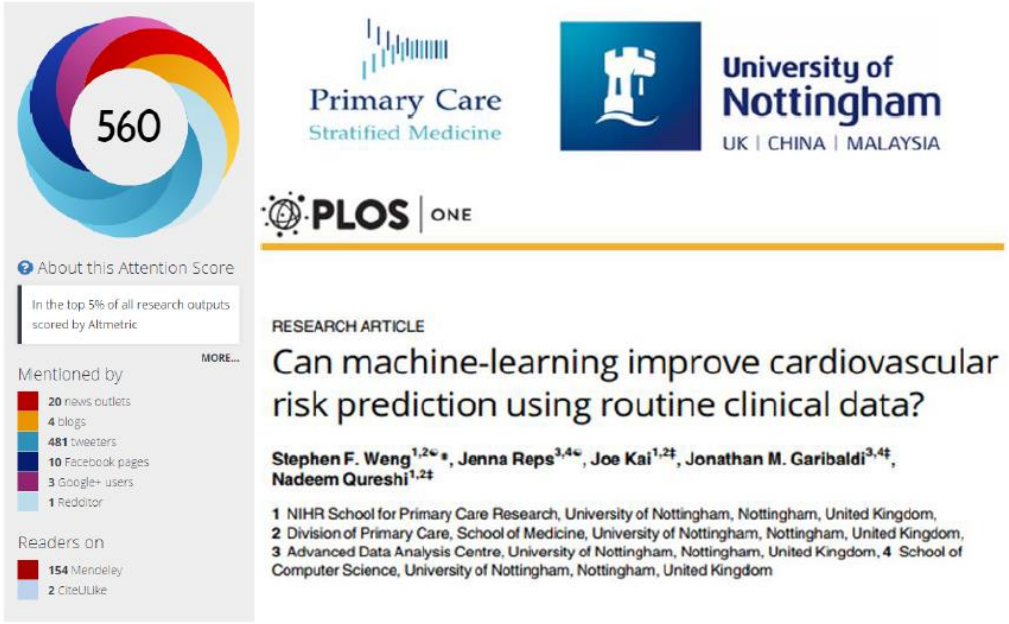
...

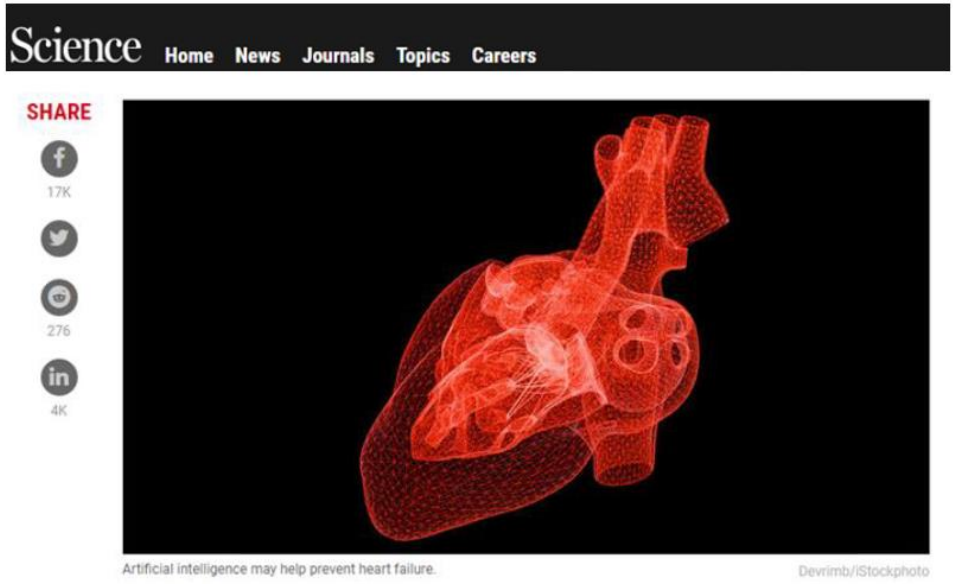
Using supervised learning to predict cardiovascular disease

- We want to predict whether someone will have a heart attack in the future.
- We have data on previous patients characteristics, including biometrics, clinical history, lab tests results, co-morbidities, drug prescriptions
- Importantly, your data requires "the truth", whether or not the patient did in fact have a heart attack.

Using supervised learning to predict cardiovascular disease

- 681 UK General Practices
- 383,592 patients free from CVD registered 1st of January 2005 followed up for years
- Two-fold cross validation (similar to other epidemiological studies): n = 295,267 "training set"; n = 82,989 "validation set"
- 30 separate included features including biometrics, clinical history, lifestyle, test results, prescribing
- Four types of models: logistic, random forest, gradient boosting machines, and neural networks



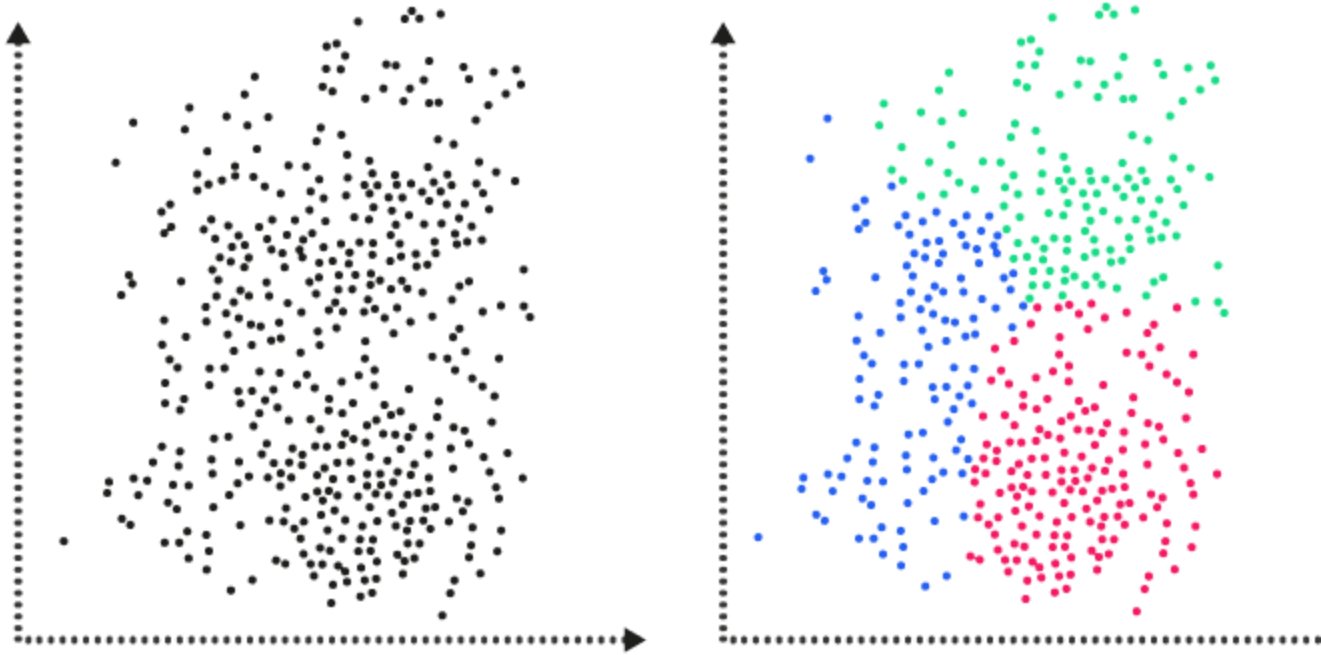


Most important supervised learning algorithms

- KNN
- Linear Regression
- Logistic Regression
- Support Vector Machines
- Decision Trees and Random Forest
- Neural networks

Unsupervised learning

- Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.
- We can derive this structure by clustering the data based on relationships among the variables in the data.
- With unsupervised learning there is no feedback based on the prediction results.



Examples of unsupervised learning

- Clustering: Take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on.

Improving phenotyping of heart failure patients to improve therapeutic stratifies

- 172 patients hospitalised with acute decompensation heart failure from the ESCAPE trial
- Performed cluster analysis (hierarchical clustering) to determine similar patient groups based on combined measures characteristics
- Researchers conducing analysis had no knowledge of clinical outcomes for patients
- 14 candidate variables, including demographics, biometrics, cardiac biomarkers

Characteristic	Cluster 1 (n = 75)	Cluster 2 (n = 33)	Cluster 3 (n = 29)	Cluster 4 (n = 25)	p-value†
Age, years	58 (46–67)	52 (44–59)	51 (42–57)	69 (59–79)	<0.001
Female, %	5	100	3	24	<0.001
Race					<0.001
White, %	87	52	0	76	
Minority, %	12	45	100	24	
Ischemic etiology, %	65	30	10	84	<0.001
LVEF, %	20 (15–23)	20 (15–25)	15 (13–18)	20 (19–25)	0.001
BMI, kg/m^2	29 (25–34)	26 (23–36)	28 (25–30)	24 (22–26)	0.013
Edema, %	72	56	79	60	0.145
Symptom score	40 (30–60)	44 (30–60)	35 (20–50)	50 (34–60)	0.295
MLHF score	78 (68–87)	76 (63–95)	83 (72–89)	74 (64–78)	0.212
Orthopnea, %	88	85	86	76	0.529
SBP, mmHg	100 (90–111)	109 (97–120)	110 (103–124)	100 (90–114)	0.005
DBP, mean	65 (60–70)	66 (56–70)	76 (68–85)	59 (55–70)	<0.001
Atrial fibrillation, %	44	15	7	24	<0.001
Angina pectoris, %	36	21	21	44	0.127
Prior CABG, %	32	15	7	64	<0.001
COPD, %	13	9	24	24	0.235
Depression, %	21	27	14	20	0.634
Diabetes, %	39	30	25	40	0.493
Hypertension, %	43	49	62	28	0.084
ICD, %	33	12	28	28	0.156
CVA, %	12	6	3.4	8	0.601
Peak VO_2, mL/kg/min	10.4 (8.0–11.9)	9.1 (7.3–10.6)	8.7 (7.6–9.3)	9.0 (7.6–10.4)	0.517
RAP, mmHg	13 (8–18)	11 (6–14)	17 (13–22)	14 (9–20)	0.005
PCWP, mmHg	27 (19–34)	22 (15–28)	32 (28–38)	23 (20–27)	<0.001
Cardiac index, L/min/m^2	1.9 (1.6–2.3)	2.0 (1.5–2.2)	1.6 (1.2–2.2)	1.8 (1.6–2.5)	0.120
Sodium, mEq/L	137 (134–139)	138 (136–139)	137 (136–139)	136 (134–138)	0.403
BUN, mg/dL	29 (20–41)	20 (12–26)	29 (23–41)	80 (47–98)	<0.001
Creatinine, mg/dL	1.4 (1.2–1.6)	0.9 (0.9–1.2)	1.4 (1.3–1.8)	2.5 (2.1–3.1)	<0.001
BNP, pg/mol	469 (174–963)	489 (183–860)	877 (89–1391)	1398 (518–4513)	0.001

AhmadT, DesaiN, WilsonF, SchulteP, DunningA, et al. (2016)Clinical Implications of Cluster Analysis-Based Classification of Acute Decompensated Heart Failure and Correlation with Bedside Hemodynamic Profiles. PLOS ONE 11(2):e0145881.https://doi.org/10.1371/journal.pone.0145881

Improving phenotyping of heart failure patients to improve therapeutic stratifies

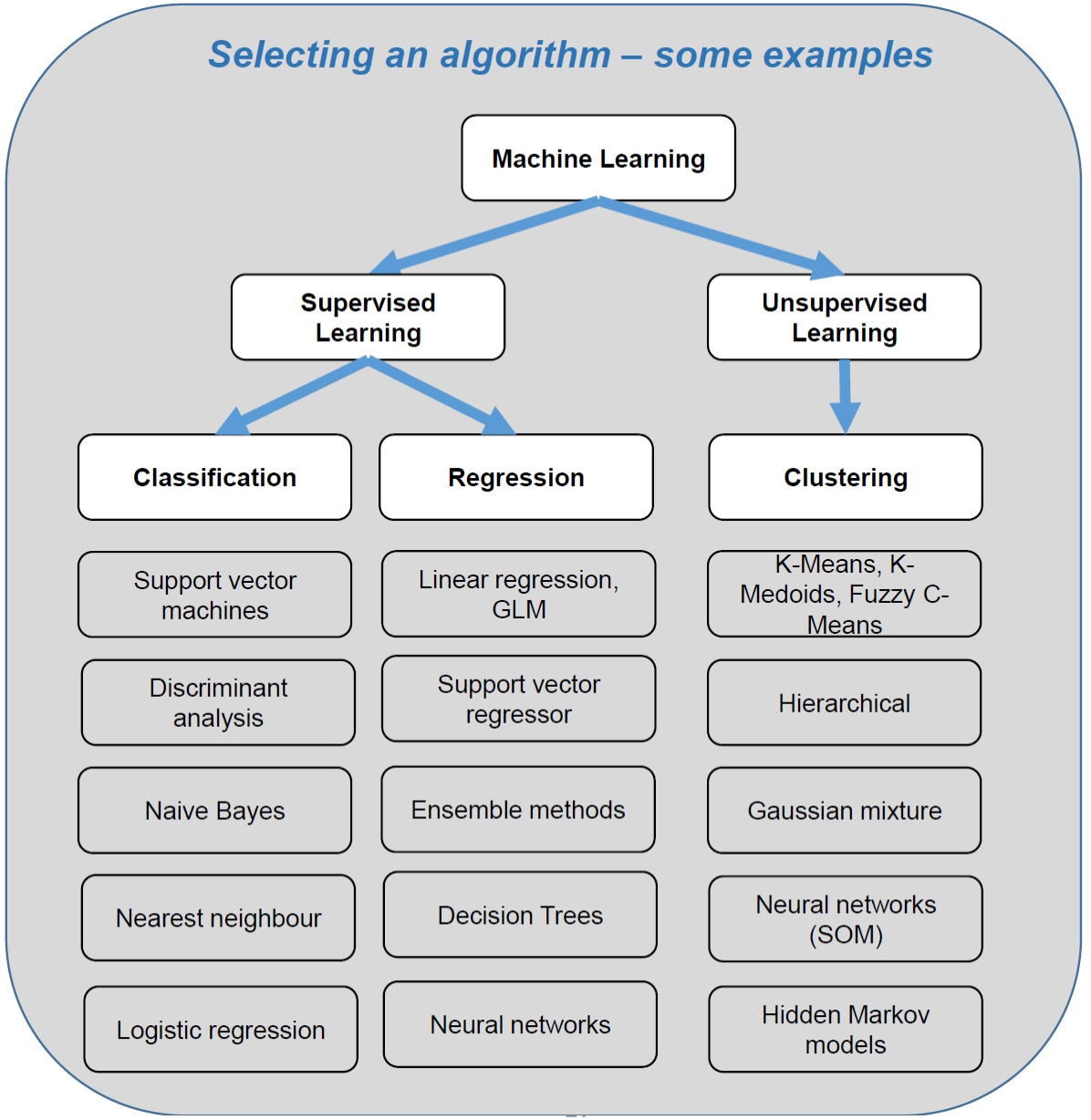
- Cluster 1: male Caucasians with ischemic cardiomyopathy, multiple comorbidities, lowest BNP levels
- Cluster 2: females with non-ischemic cardiomyopathy, few co-morbidities, most favourable hemodynamics, advanced disease
- Cluster 3: young African American males with non-ischemic cardiomyopathy, most adverse hemodynamics, advanced disease
- Cluster 4: older Caucasians with ischemic cardiomyopathy, concomitant renal insufficiency, highest BNP levels

Choosing the best ML algorithm

Choosing the right algorithm can seem overwhelming

Considerations:

- There is no best method or one size fits all.
- Trial and error.
- Size and type of data.
- The research question and purpose.
- How will the outputs be used?



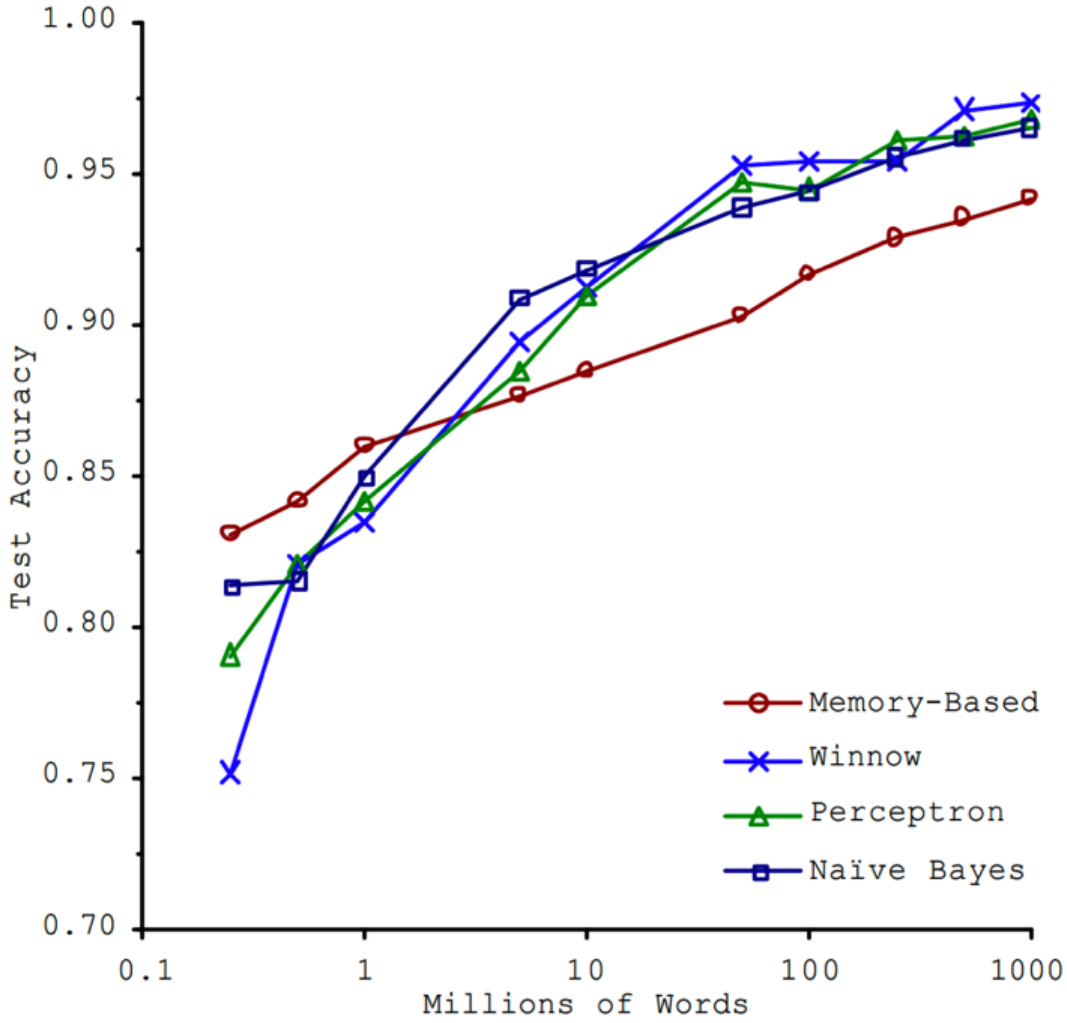
IA4Health simplified Workflow

Once a clear research question has been defined, it's time to look for the answer in the data by carrying out the following steps:

1. **ACCESS: format and load the data**
2. **PREPROCESS: data management, cleaning, coding, organising**
3. **DERIVE: features (variables) using the cleaned data**
5. **TRAINING: select algorithm, train models using derived features**
6. **ITERATE: different algorithms to find the best model**
7. **VALIDATE: trained model on separate dataset**
8. **INTERPRETATION: clinical verification and interpretation of outputs**
9. **DISSEMINATION: integrate into production system/publish in journals**

Main challenges of Machine Learning

- Issufficient Quantity of Trainning Data : Even for simple problems, you typically need thousand of examples.
  - The unreasanable effectiveness of data



Main challenges of Machine Learning

- Non representative trainning data: It is crucial to use a trainning set that is representative of the cases we want to generalize to.
- If the sample is too small you will have sample noise, (non representative data as a result of chance), but even very large samples can be non representative if the sampling method is flawed. This is called sample bias;

Main challenges of Machine Learning

- Overfitting

Let's start with an example, say one day you are walking down a street to buy something, a dog comes out of nowhere you offer him something to eat but instead of eating he starts barking and chasing you but somehow you are safe. After this particular incident, you might think all dogs are not worth treating nicely.

So this **overgeneralization** is what we humans do most of the time, and unfortunately machine learning model also does the same if not paid attention. In machine learning, we call this overfitting i.e model performs well on training data but fails to generalize well.

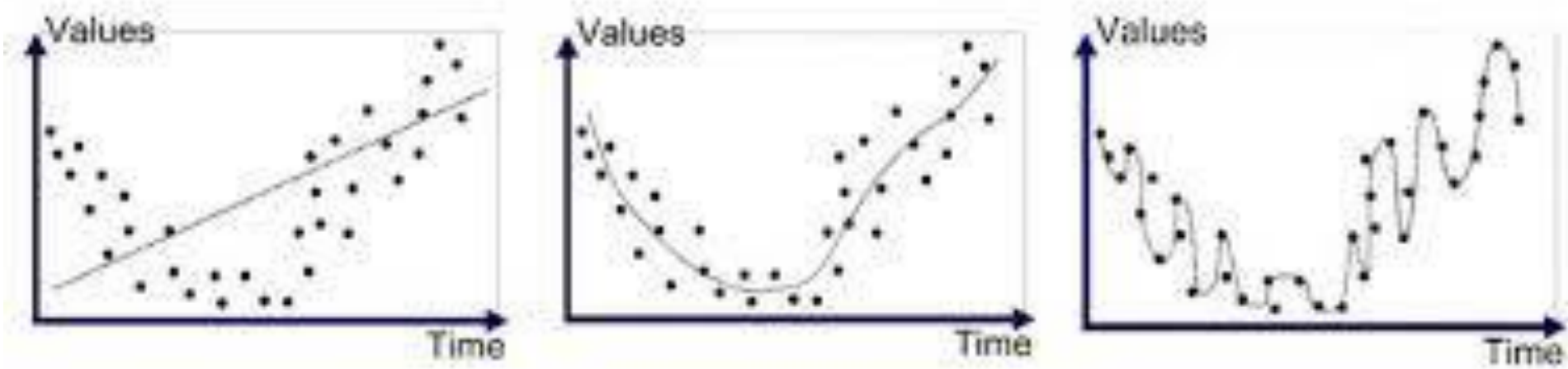
Overfitting happens when our model is too complex.

Things which we can do to overcome this problem:

- Simplify the model by selecting one with fewer parameters.
- By reducing the number of attributes in training data.
- Constraining the model.
- Gather more training data.
- Reduce the noise.

Main challenges of Machine Learning

- Underfitting is the opposite of overfitting. It happens when our model is too simple to learn something from the data. For E.G., you use a linear model on a set with multi-collinearity it will for sure underfit and the predictions are bound to be inaccurate on the training set too.
- Things which we can do to overcome this problem:
  - Select a more advanced model, one with more parameters.
  - Train on better and relevant features.
  - Reduce the constraints.



Key challenges for health care data

- Most challenges come from handling your data and finding the "right" model
- **Data comes in all shapes and sizes:** Real-world datasets are messy, incomplete, and come in a variety of formats
- **Pre-processing your data requires clinical knowledge and the right tools:** For example to select the correct features (variables) and codes to use in primary care datasets, you'll need clinical verification and knowledge of NHS coding and content expertise
- **Can your question be answered without ML:** many research questions don't actually require ML. For instance, accurate risk prediction models can be developed stepwise regression models.
- **Choosing the "right" model:** Highly flexible models tend to over-fit while simple models make too many assumptions. Trial and error is at the core of machine learning
- **Understand the limitations:** Not recommended for causal inferences, interpretation of results can be difficult

Key challenges for health care data

- Life or death decisions
  - Need **robust** algorithms
  - Checks and balances built into ML deployment (Also arises in other applications of AI such as autonomous driving)
  - Need **fair** and **accountable** algorithms
- Many questions are about unsupervised learning
  - Discovering disease subtypes, or answering question such as "characterize the types of people that are highly likely to be readmitted to the hospital"?
- Many of the questions we want to answer are *causal*
  - Naïve use of supervised machine learning is insufficient

Key challenges for health care data

- Often very little labeled data (e.g., for clinical NLP)
  - Motivates semi-supervised learning algorithms
- Sometimes small numbers of samples (e.g., a rare disease)
  - Learn as much as possible from other data (e.g. healthy patients)
  - Model the problem carefully
- Lots of missing data, varying time intervals, censored labels

Key challenges for health care data

- Difficulty of de-identifying data
  - Need for data sharing agreements and sensitivity.
- Difficulty of deploying ML
  - Commercial electronic health record software is difficult to modify.
  - Data is often in silos; everyone recognizes need for interoperability, but slow progress.
  - Careful testing and iteration is needed.