Machine Learning For Design

Lecture 2 - Fundamentals of Machine Learning

Alessandro Bozzon

11/02/2022

mlfd-io@tudelft.nl
www.ml4design.com

Admin

Week 1 Tasks

- 98 Students self-subscripted to a group
  - Still 20/25 students missing
  - Whatsapp chat for group composition
    - https://chat.whatsapp.com/DE36WPV7NjL8bL99eLNFmE
- 24 students presented themselves on Discourse
- 20 Questions
  - Thank you!

Previously, on ML4D….

Machine Learning

- *The field of study that gives computers the ability to learn **without being explicitly programmed***

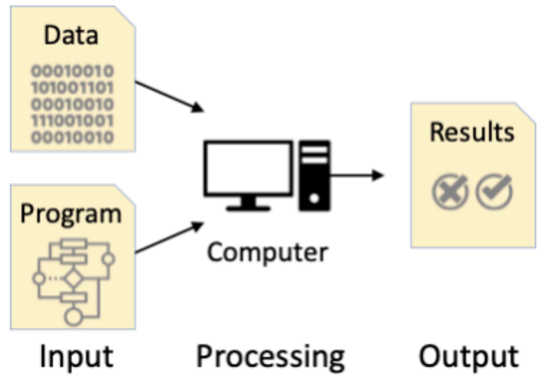
Arthur Samuel

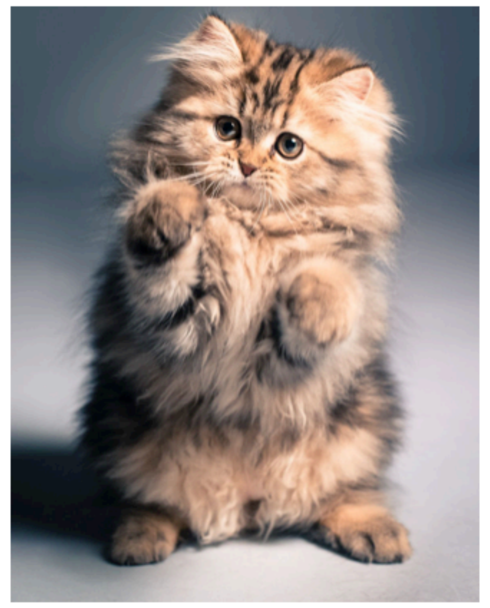
- Machine learning is the science (and art) of programming computers **so they can learn from data**

Is this a cat?

- **Traditional Programming**

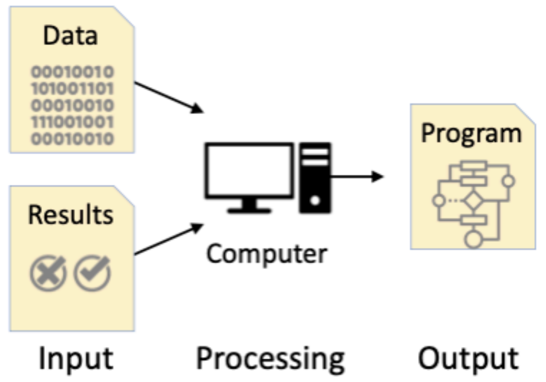
Rules to detect a cat:
1. It has whiskers
2. It is furry
3. It is small





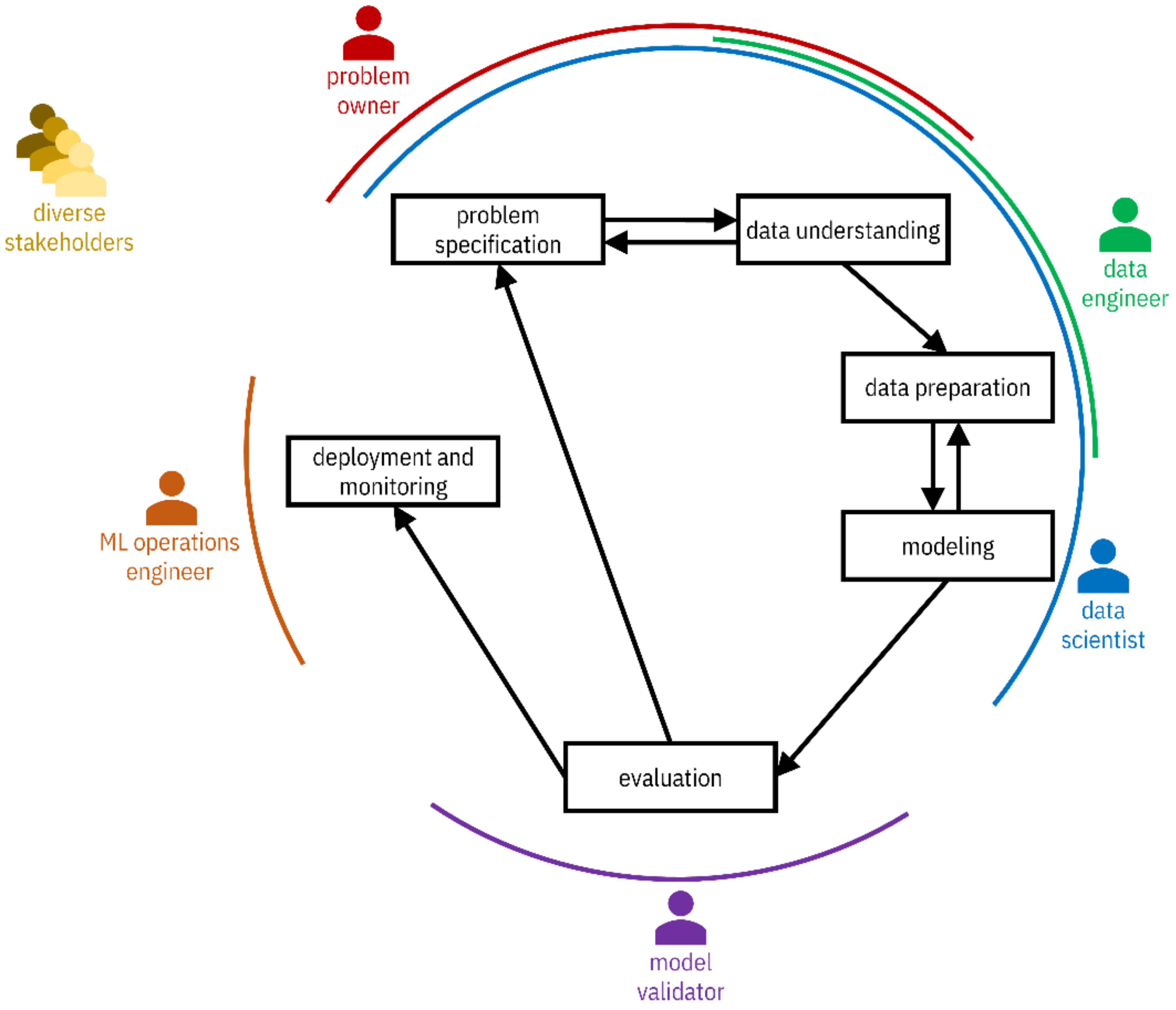
- **Machine Learning**

Let me guess how I can distinguish a cat :)



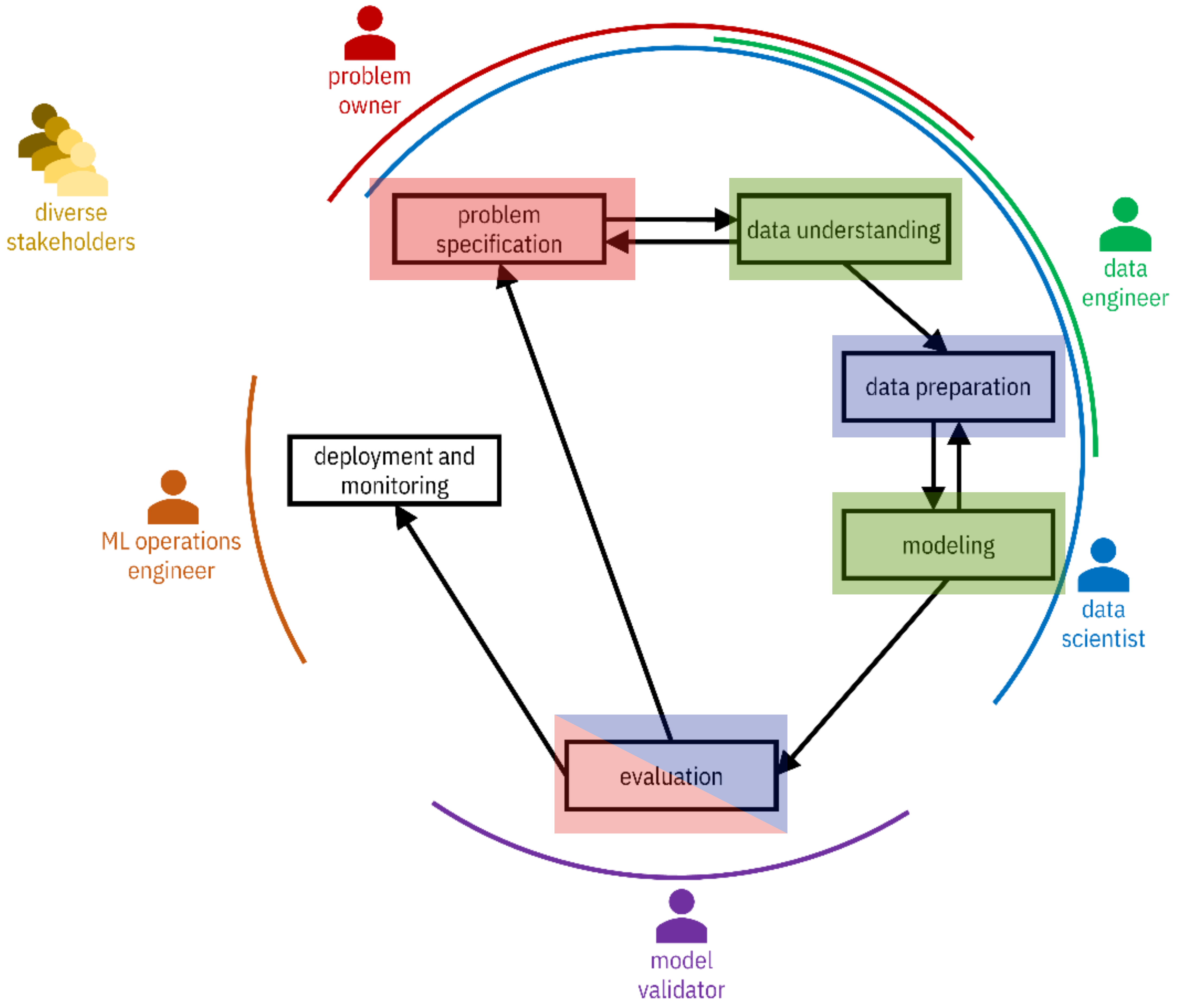
The Machine Learning Life-cycle

Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology



https://www.the-modeling-agency.com/crisp-dm.pdf

CRISP-DM in our course



Today and in all modules

In Module 4

In Module 3

Problem Specification

- What is the problem owner hoping to accomplish and why?
- Why am I (being asked to) solve it?
- Am I the right person to solve this problem?
- What are the repercussions of building this technology?
- Should this thing be built at all?
- What are the metrics of success?

Data Understanding

- Data need to be collected —> **Datasets**
- What data is available?
- What data should be available, but isn't?
- What population / system / process is your data representing?
- And what properties of such population / system / process are included (or excluded)?
- What biases (social, population, temporal) are present in your datasets?

Know your data!

Data Preparation

- Data integrations
  - Extracting, transforming, and loading (ETL) data from disparate relevant databases and other data sources
  - This step is most challenging when dealing with big data sources
- Data cleaning
  - Filling missing values
  - Transforming value types (e.g. binning)
  - Dropping features that should not be considered
- *Feature engineering*
  - Transform the data to derive new features

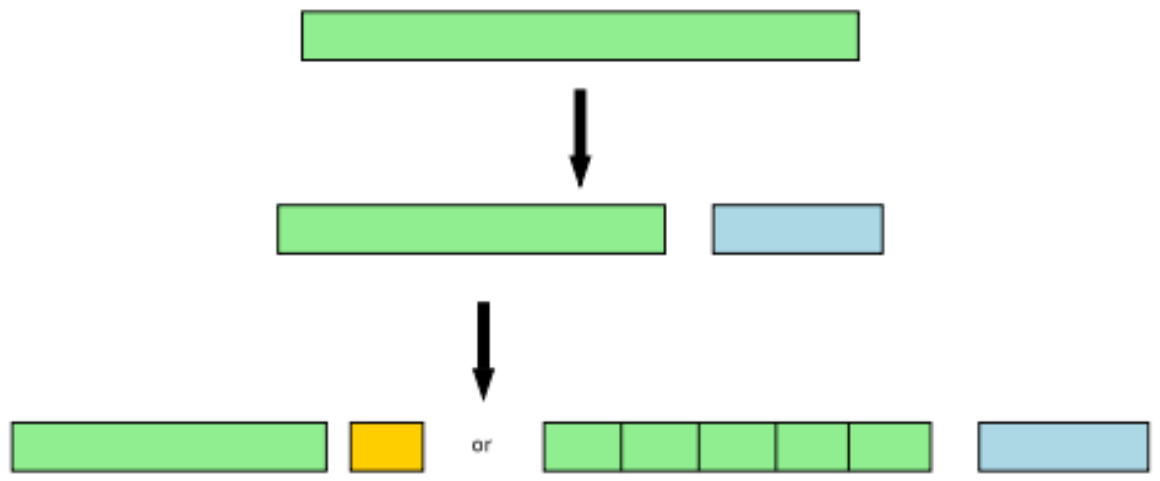
Modeling

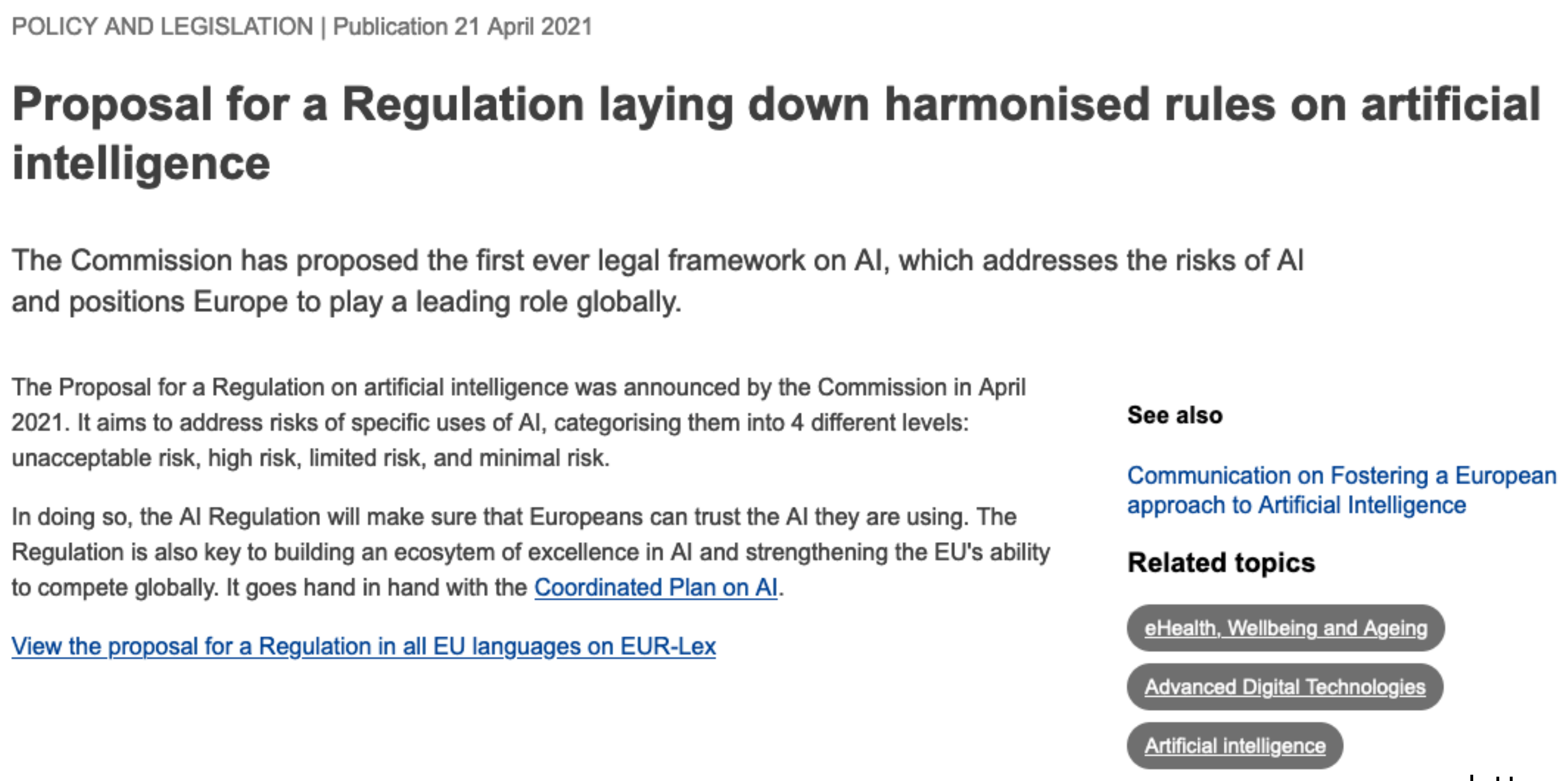
- Select a training algorithm
- Use it to find patterns in the training dataset
- Generalize from them to fit a statistical model
- Enhance the model to to satisfy additional objectives and constraints captured in the problem specification (e.g. increase reliability, mitigate biases, generate explanations)

- ***<u>No free-lunch theorem</u>***
  - There is no one best machine learning algorithm for all problems and datasets

Evaluation

- Testing and validation of the model
  - Also against the *problem specification requirements*
- Performed on data not used for training
  - *Held out* dataset
- Model auditing / risk management




POLICY AND LEGISLATION | Publication 21 April 2021

Proposal for a Regulation laying down harmonised rules on artificial intelligence

The Commission has proposed the first ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

The Proposal for a Regulation on artificial intelligence was announced by the Commission in April 2021. It aims to address risks of specific uses of AI, categorising them into 4 different levels: unacceptable risk, high risk, limited risk, and minimal risk.

In doing so, the AI Regulation will make sure that Europeans can trust the AI they are using. The Regulation is also key to building an ecosystem of excellence in AI and strengthening the EU's ability to compete globally. It goes hand in hand with the Coordinated Plan on AI.

View the proposal for a Regulation in all EU languages on EUR-Lex

See also

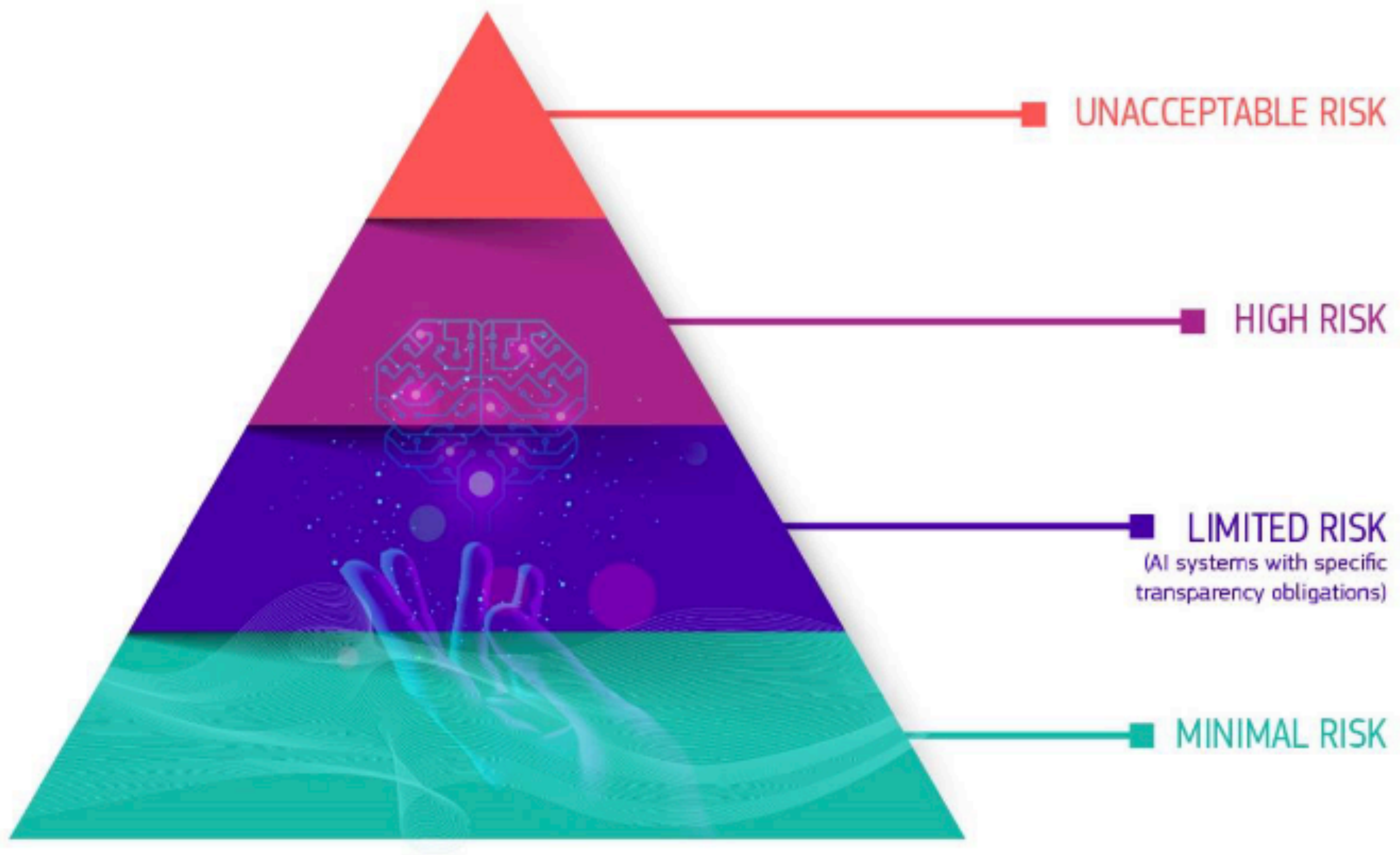
Communication on Fostering a European approach to Artificial Intelligence

Related topics

eHealth, Wellbeing and Ageing

Advanced Digital Technologies

Artificial intelligence



The Pyramid of Criticality for AI Systems

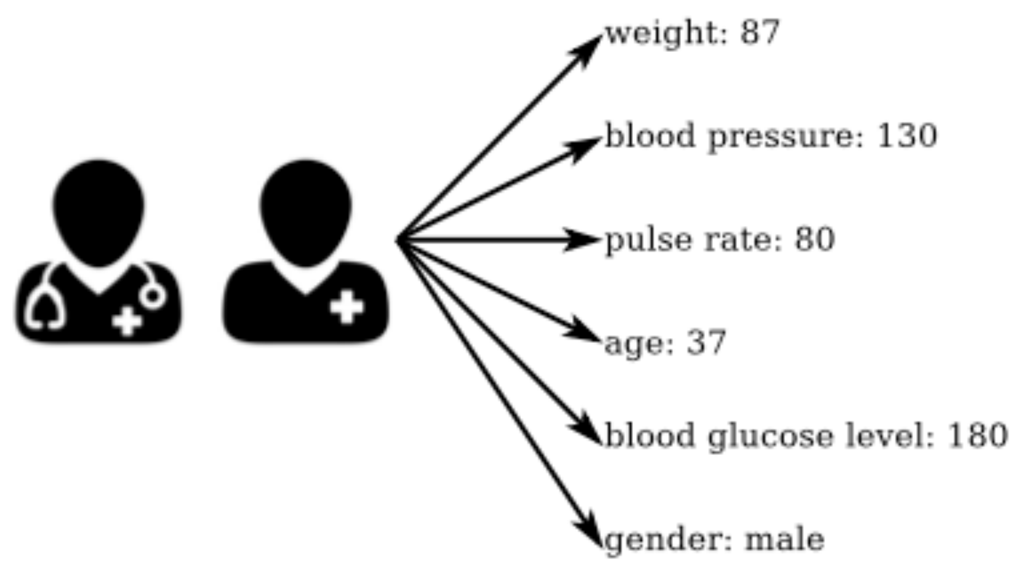
https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682

Deployment and monitoring

- What infrastructure will bring new data to the model?
- Will predictions be made in batch or one-by-one?
- How much latency is allowable?
- How will the user interact with the system?

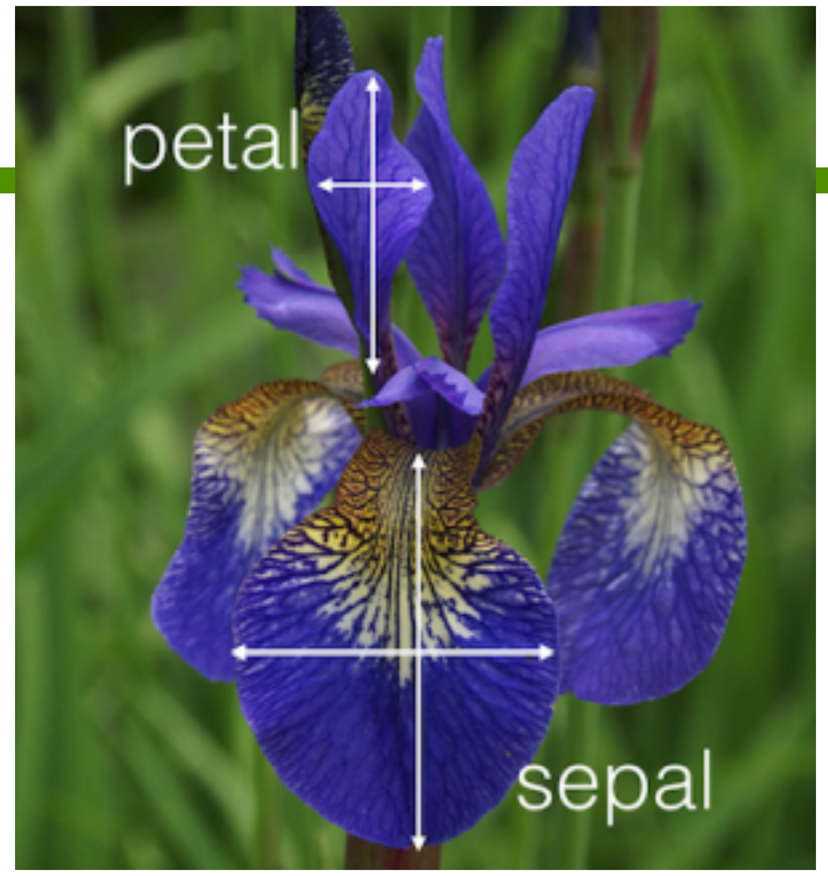
- Tools to monitor the model's performance
  - And ensure it is operating as expected

Data



The raw material

Data



Feature Feature Feature Feature Label

sepal_lenght	sepal_width	petal_lenght	petal_width	Class
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica

Record / Sample / Data Item

Label Value

Feature Value

Dataset Size

Dataset Dimensionality

https://archive.ics.uci.edu/ml/datasets/iris

Types of Feature / Label Values

Categorical

- Named data
- Can take numerical values, but no mathematical meaning

Nominal

- No order
- No direction

- **Marital status**
- **Gender**
- **Ethnicity**

Ordinal

- Order
- Direction

- **Letter grades (A, B, C, D)**
- **Socio-economic status (poor, rich)**
- **Ratings (dislike, neutral, like)**

Numerical

- Measurements
- Take numerical values
  - Discrete or continous

Interval

- Difference between measurements
- No true zero or fixed beginning

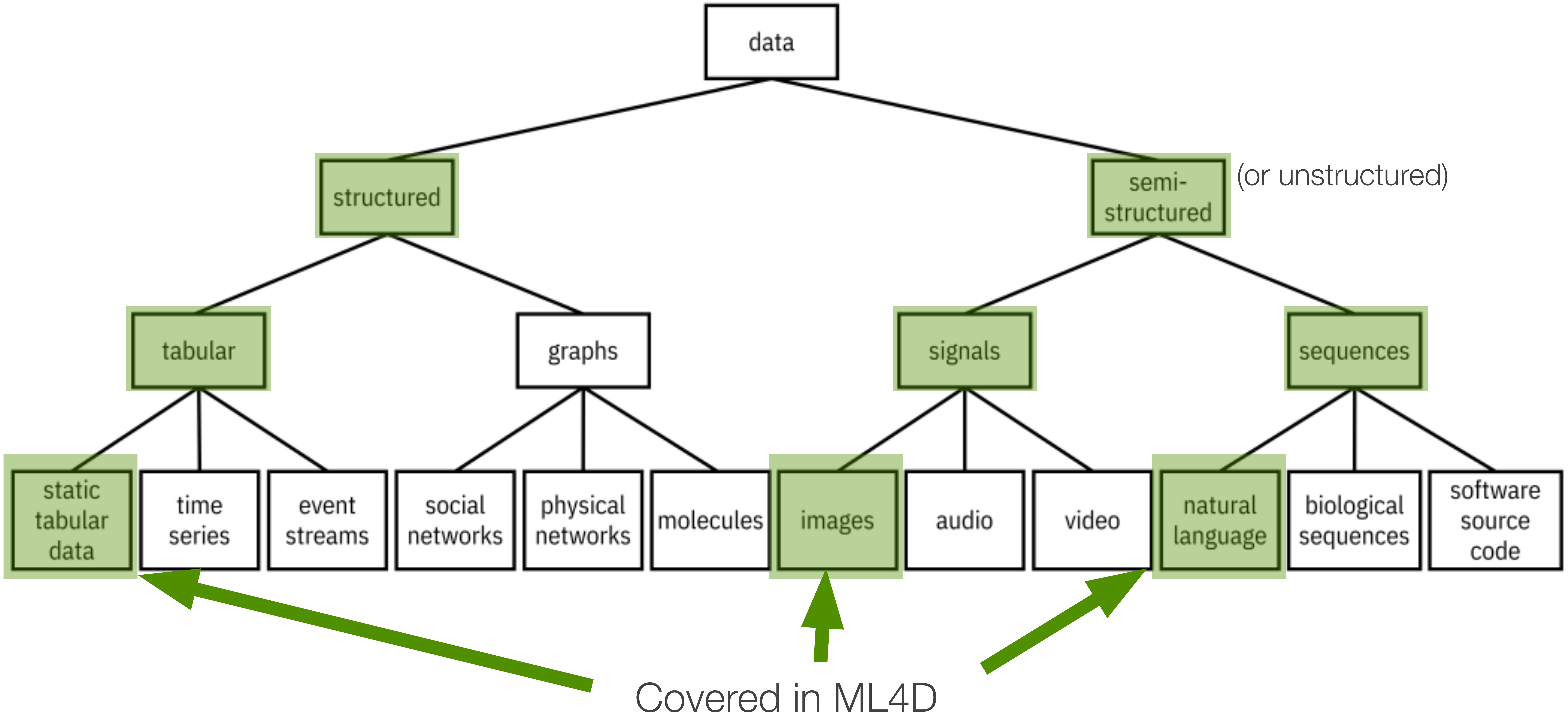
- **Temperature (C or F)**
- **IQ**
- **Time, Dates**

Ratio

- Difference between measurements
- True zero exists

- **Temperature (K)**
- **Age**
- **Height**
- **Weight**

Data Modalities



Key Dimensions

Modality	Quantity	Quality	Freshness	Cost
▪ Structured ▪ Semi-structured	▪ Number of records ▪ Number of features	▪ Errors ▪ Missing data ▪ Bias	▪ Rate of collections	▪ Of acquisition ▪ Licensing ▪ Cleaning and integration

Static Tabular Data

Feature Feature Feature Feature Label

sepal_lenght	sepal_width	petal_lenght	petal_width	Class
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica

Record / Sample / Data Item

Label Value

Feature Value

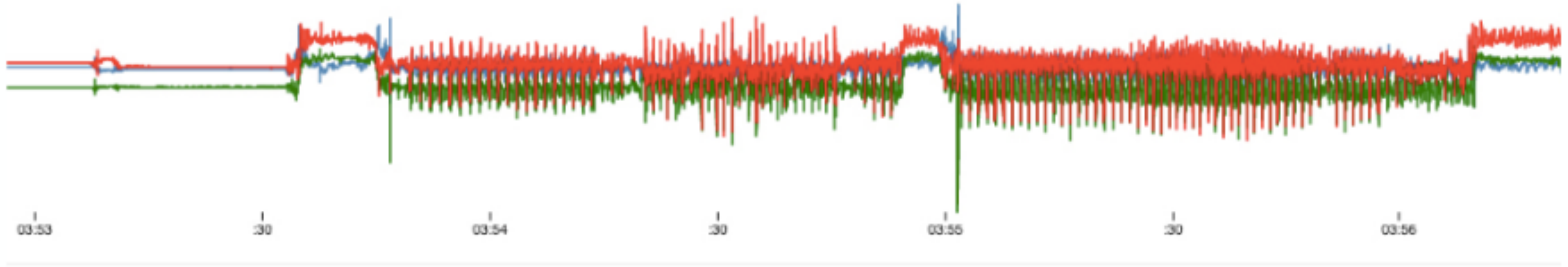
Dataset Size

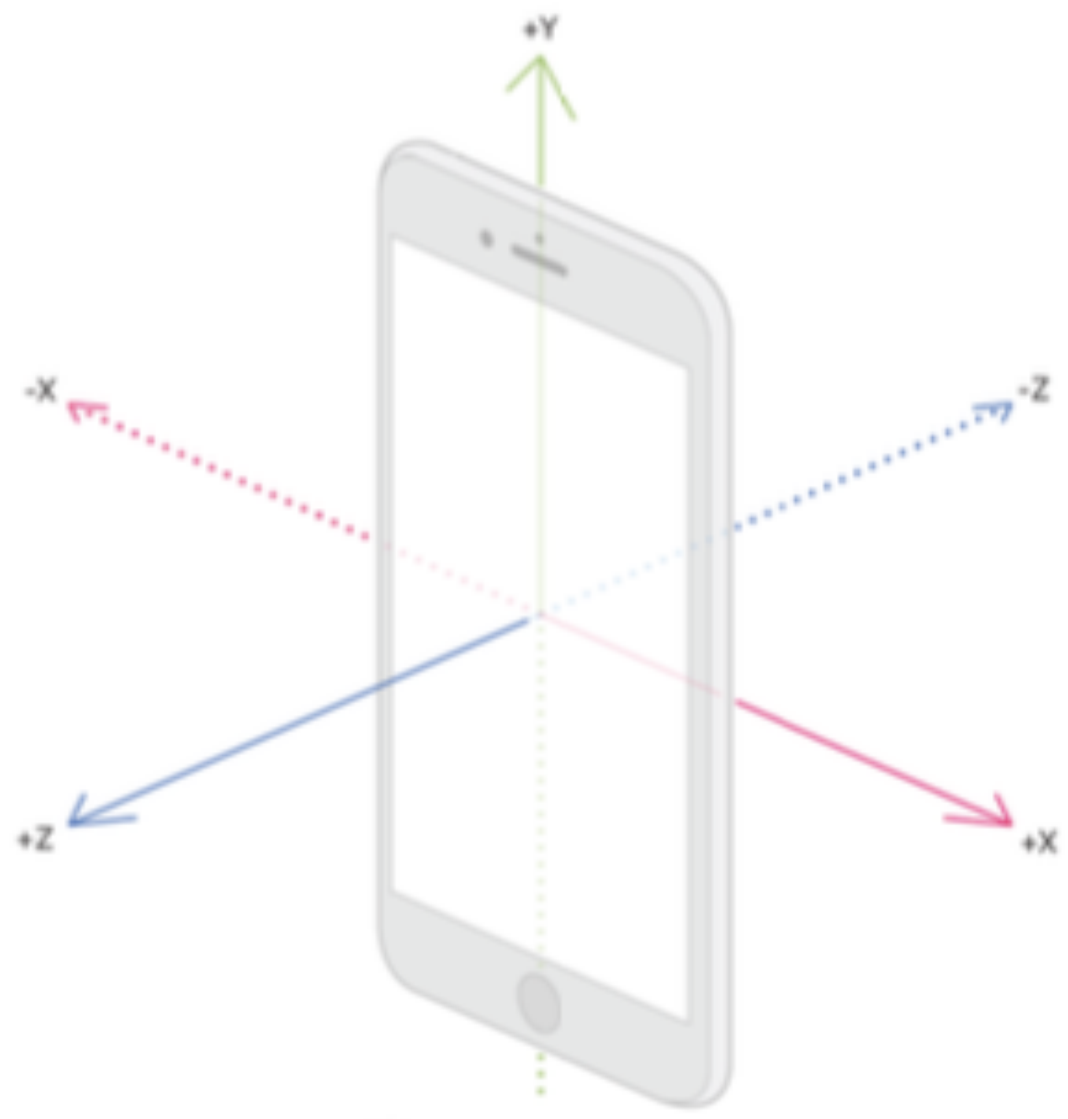
Dataset Dimensionality

Time Series

- Regularly captured tabular data
  - **Time feature**
- For instance
  - Sensor data, Stock market data
- Label is usually associated to set of records
  - e.g. a continues movement of the phone indicating an action

Accelerometer



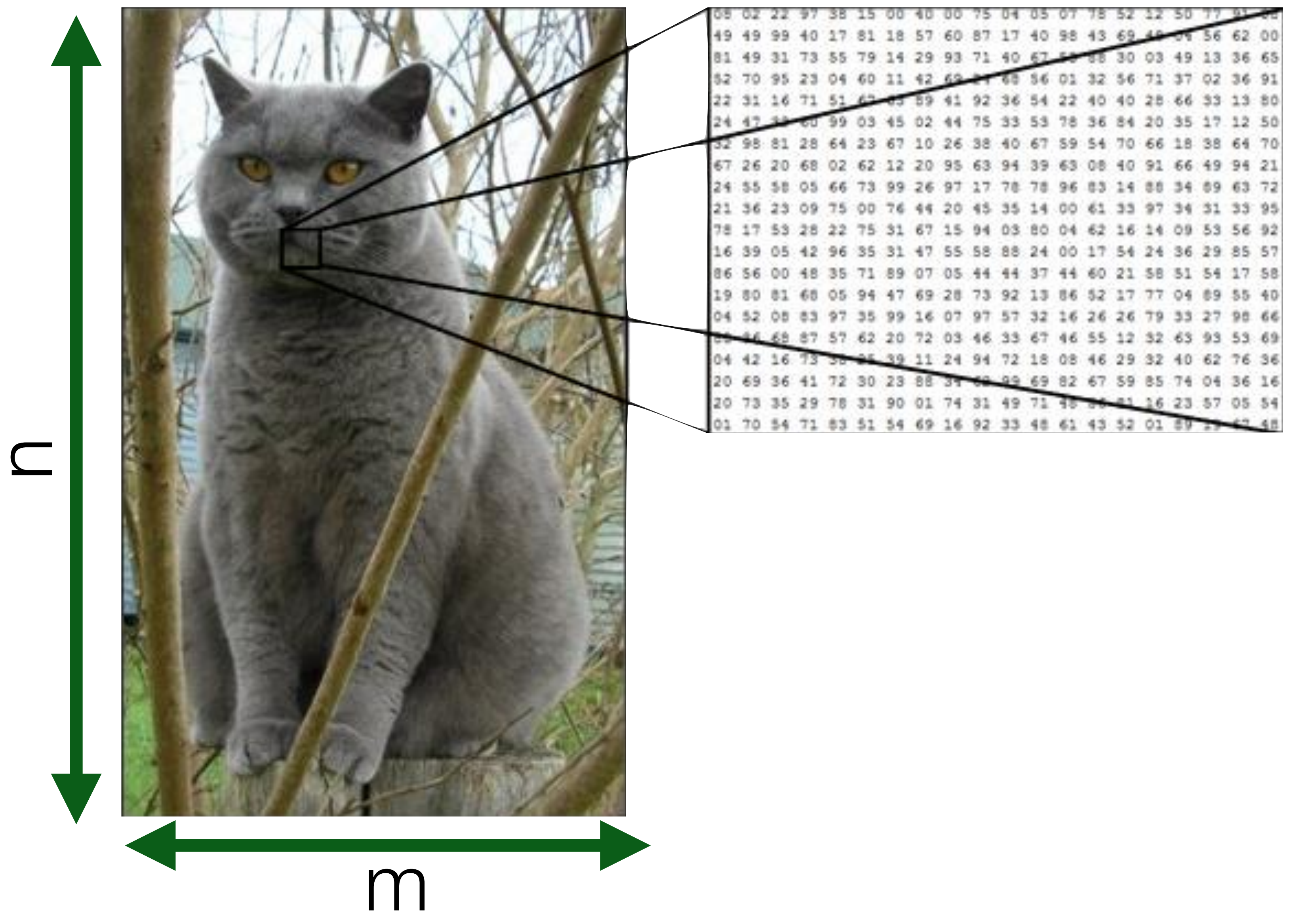


Time Feature

Timestamp	X	y	Z	Class
15060015925	2.04	3.72	8.12	Device Rotation
15060015943	1.96	4.73.68	7.56	
15060015980	1.63	3.56	6.53	
1506001610	1.06	3.76	5.81	

Images

- Visual content acquired through cameras, scanners, etc.
- Each pixel in an image is a feature
  - But spatially and geometrically organised
  - e.g. edges, corners
- Feature values are numerical values across channels
  - e.g. R,G,B
- Dimensionality —> n x m



Image

P(1,1)	P(2,1)	P(3,1)	...	P(n,m)	Class
255, 0, 0	255, 1, 1	255, 0, 0		R,G,B	Cat
255, 213, 0	255, 213, 1	255, 213, 4		R,G,B	Dog
					Cat
					Duck

More in Module 1

Textual documents

- Sequence of alphanumerical characters
  - Short: e.g. tweets
  - Long: e.g Web documents, nterview transcripts

- Features are (set of) words
  - Words are also syntactically and semantically organised
- Feature values are (set of) words occurences
- Dimensionality —> at least dictionary size

I wear this mask to sing lullabies to my children ..., 24 May 2015

By **Sir Chubs**

Verified Purchase (What is this?)

This review is from: Overhead Rubber Penguin Mask Happy Feet Animal Fancy Dress (Toy)

I wear this mask to sing lullabies to my children. They are terrified of the mask. Whenever they protest about their bed time, or ask for too many sweets, I whip on the mask, and they soon know who is the King Penguin.

Document

I	Wear	Mask	...	W(n)	Class
1	1	1		0	Spam
0	0	1		0	Not Spam
					Spam

More in Module 2

Data Sources

Purposefully Collected Data

- Surveys
- Census
- Scientific experiments
- Economic indicators
- Ad-hoc sensing infrastructure

- **Modality**: mostly structured
- **Quantity**: low
- **Quality**: high
- **Freshness**: low
- **Cost**: high

Administrative Data

- Call records
- Financial transaction data
- Travel data
- GPS data

- **Modality**: mostly structured
- **Quantity**: high
- **Quality**: high
- **Freshness**: high
- **Cost**: high

Social Data

- Web pages
- Social media
- Apps
- Search engines

- **Modality**: mostly semi-structured
- **Quantity**: high
- **Quality**: low
- **Freshness**: high
- **Cost**: low

Crowdsourcing

- Distributed sensing
- Implicit crowd work (e.g. captcha)
- Micro-work platforms (e.g. Amazon Mechanical Turk)

- **Modality**: all
- **Quantity**: mid-low
- **Quality**: mid
- **Freshness**: mid
- **Cost**: mid-low

Categories of Machine Learning

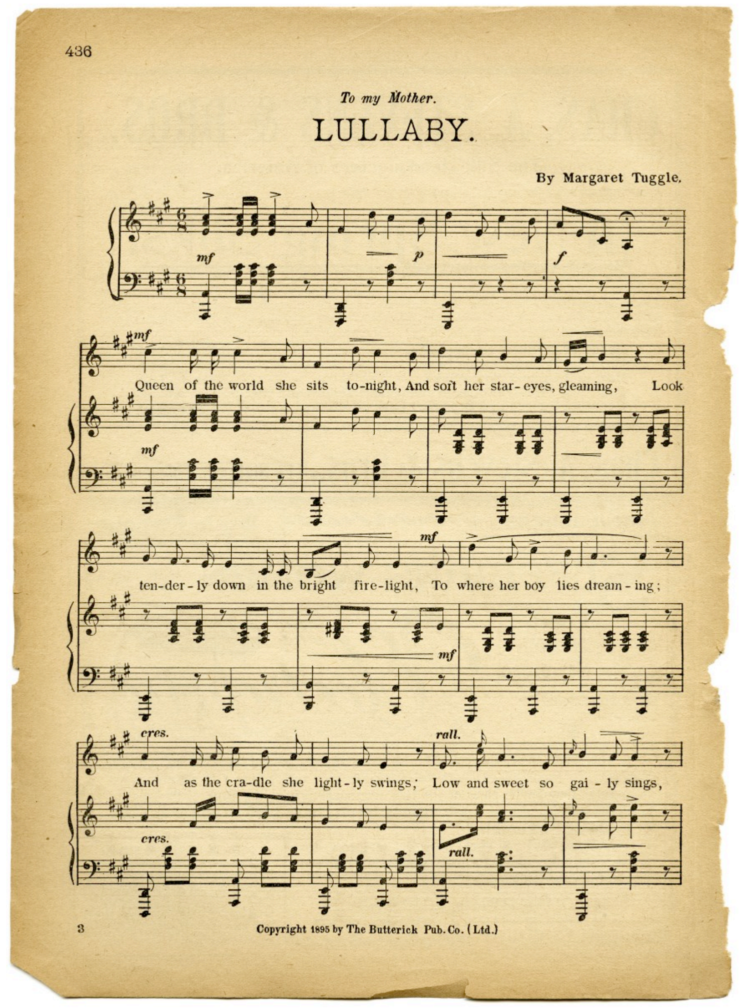
How do machines learn?

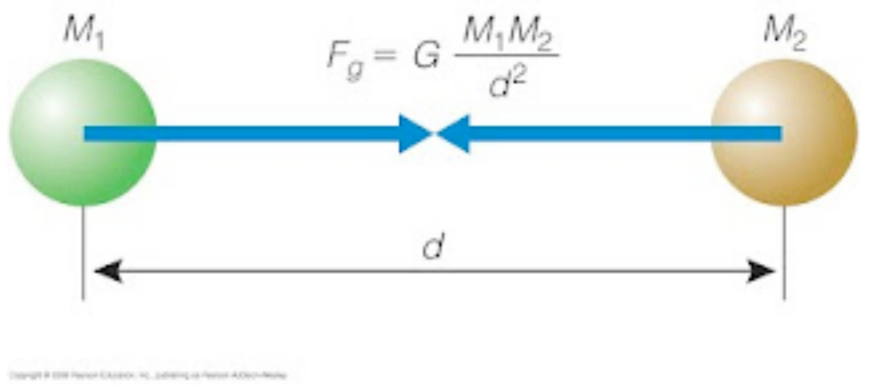
Training Data → Model Training → ML Model → Model Enhancement → Final ML Model

On Models

A physical, **mathematical**, logical, or conceptual representation of a system, entity, phenomenon, or process

- A simple(r) representation of reality, that helps us to understand how something **works**, or **will work**
  - Not a truthful representation of reality, just an useful one
- The goal of models is to make a particular part or feature of the world easier to **understand**, **define**, **quantify**, **visualize**, or **simulate**
- Examples of models
  - Architecture plans
  - Maps
  - Music Sheet
  - Mathematical laws of physics!
  - Machine Leaning (statistical) Models





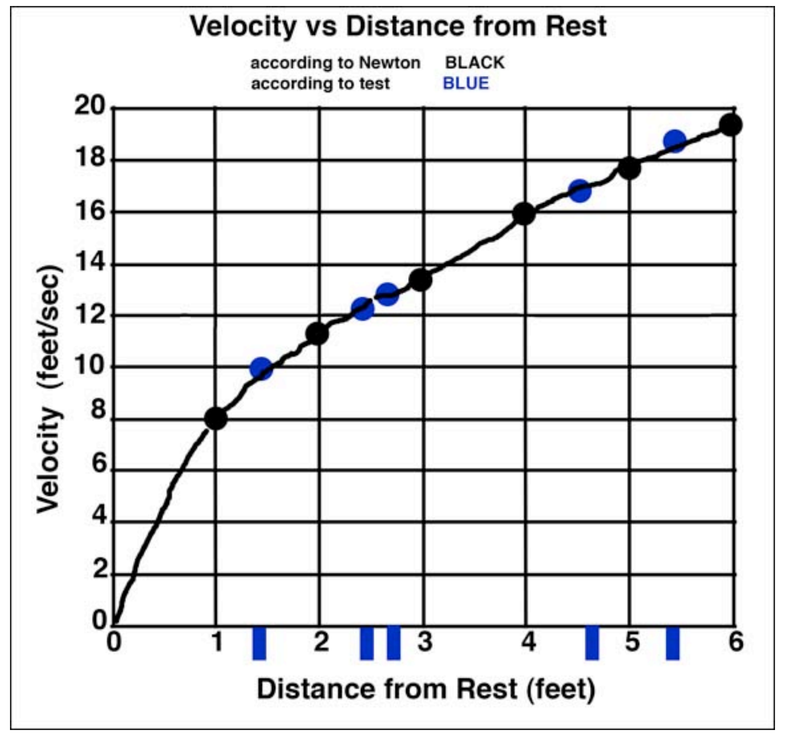
On Models / Scientific Models

- **GOAL**: **<u>explain reality</u>**
- Models are created to make predictions about the outcomes of future experiments
  - E.g. apples on the moon
- Models are tested against the outcome
- If data from new experiments don't agree, the model has to be modified / extended / refined
  - Falsifiability

- Scientific models should be small and simple
- They should generalize to phenomena observed in new ways

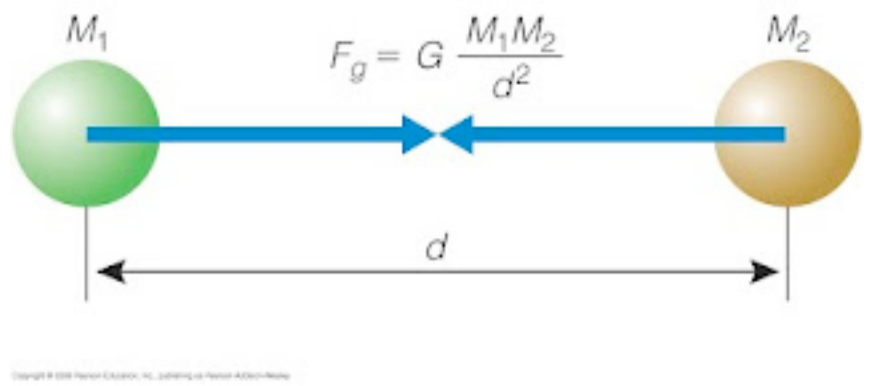
Piece of reality

Observation Procedure



https://people.rit.edu/andpph/text-newton-right-wrong.html

Dectect Regularities



Make predictions and test in experiments

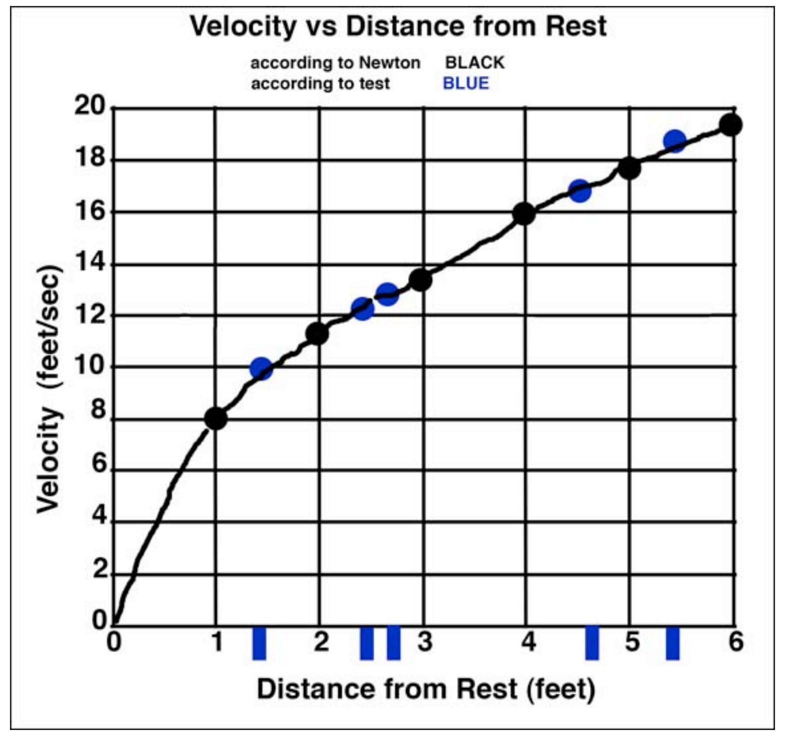
On Models / ML Models

- **GOAL**: **describe the data**
- ML models are designed to capture the variability in observational data, by exploiting regularities / symmetries / redundancies
- A good ML model doesn't need to explain reality, **just describe data**
- Therefore, they don’t need to be simple or transparent or intelligible. Just **accurate**
  - *Black box*

- ML models may be large and complex
- They should generalize to new data obtained in the same way as the training data
  - Same application context and data acqusition process

Piece of reality

Observation Procedure



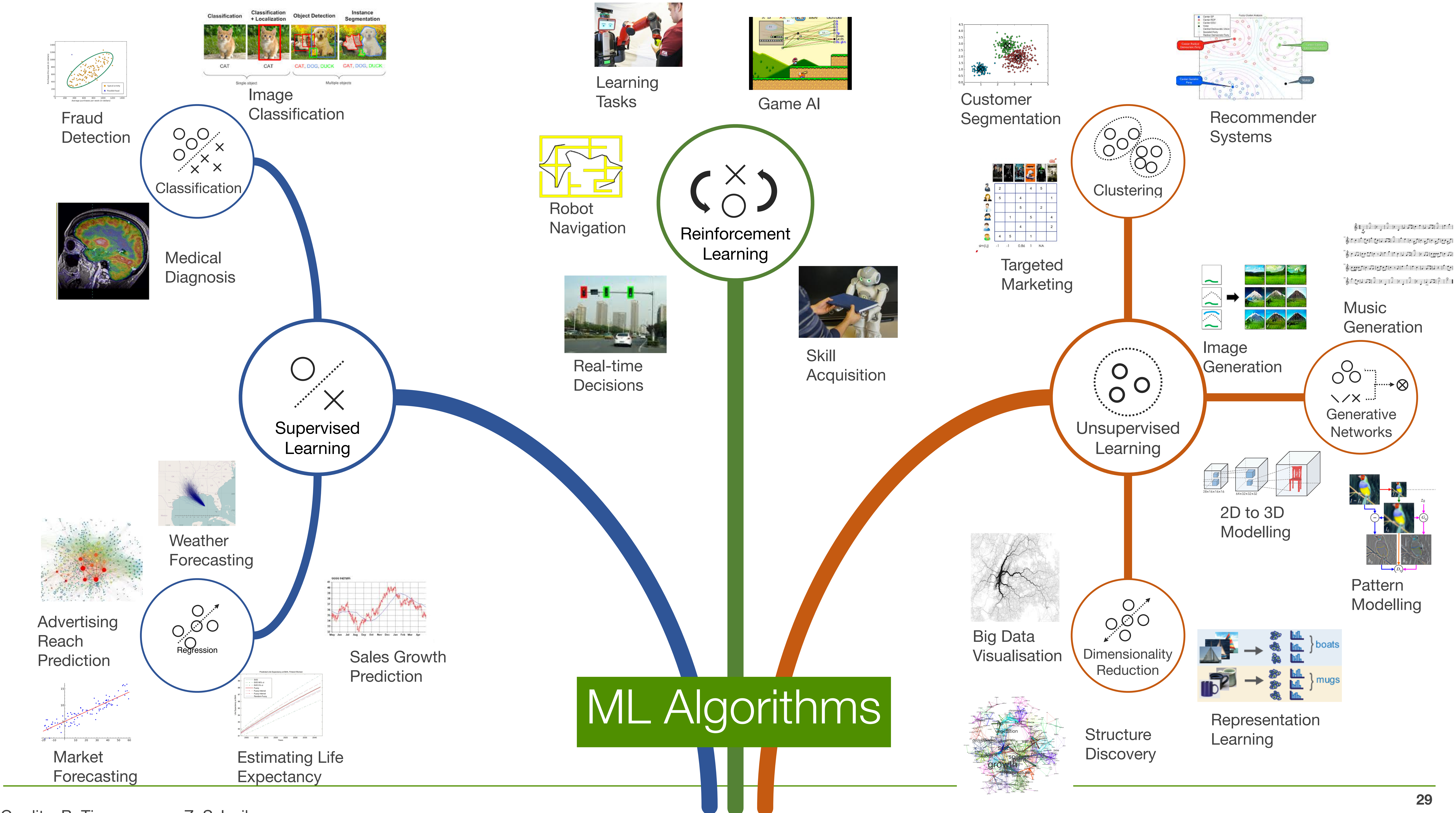
https://people.rit.edu/andpph/text-newton-right-wrong.html

Test models on more data

Dectect Regularities In training data

$V = f_{model}(D)$

Model



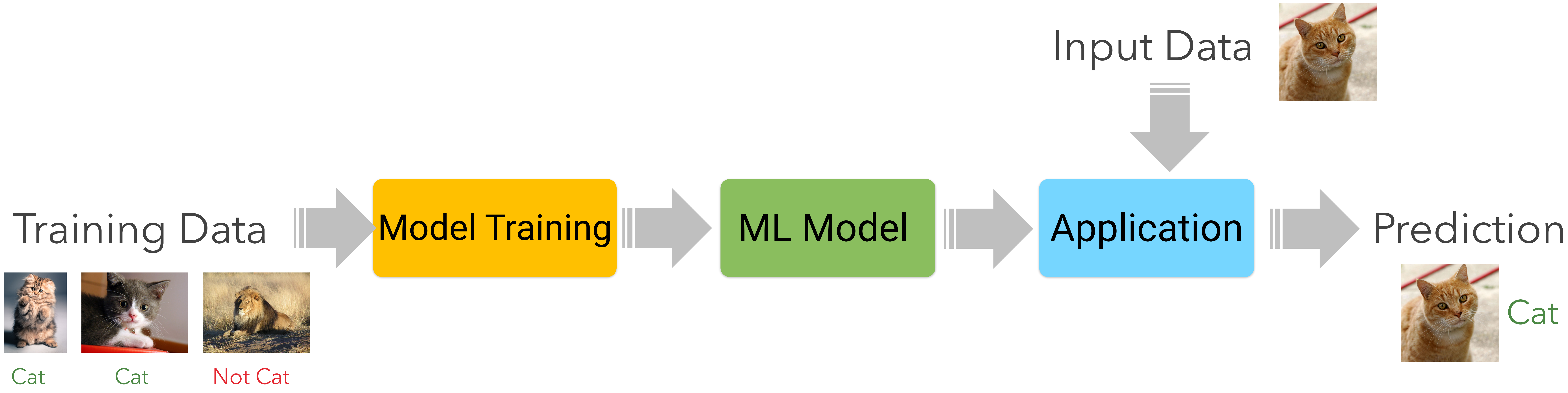
Credits: B. Timmermans, Z. Szlavik

Supervised Learning

Labelled data in input

The machine exploits labels to associate patterns to outputs

It learns how to make input-output **predictions**



Classification | **Regression** | **Ranking** | **Recommendation**

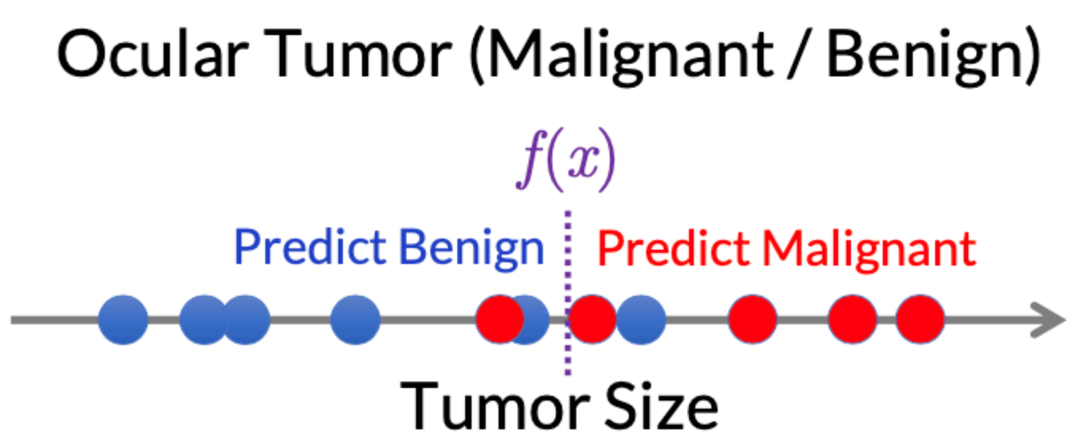
Classification / Regression

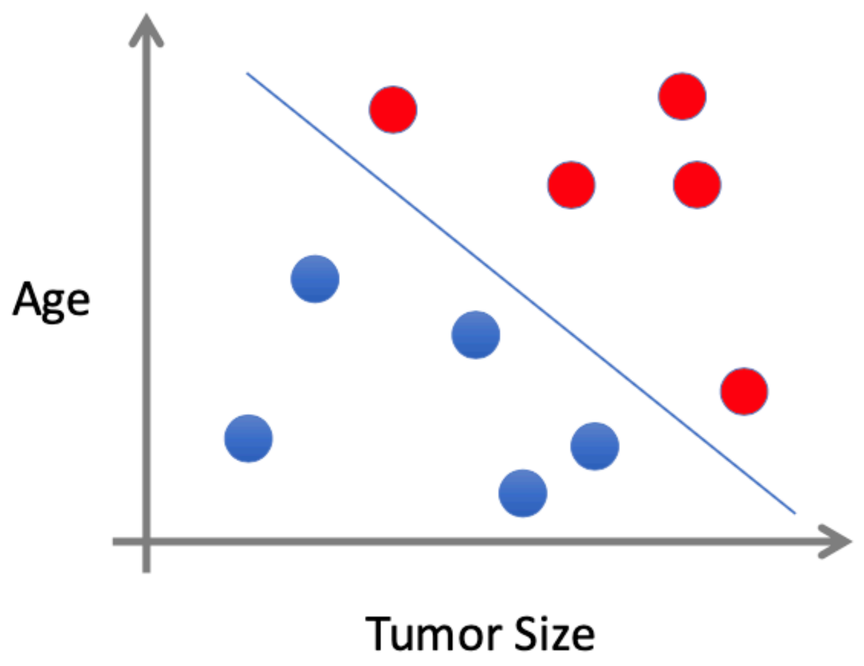
Classification

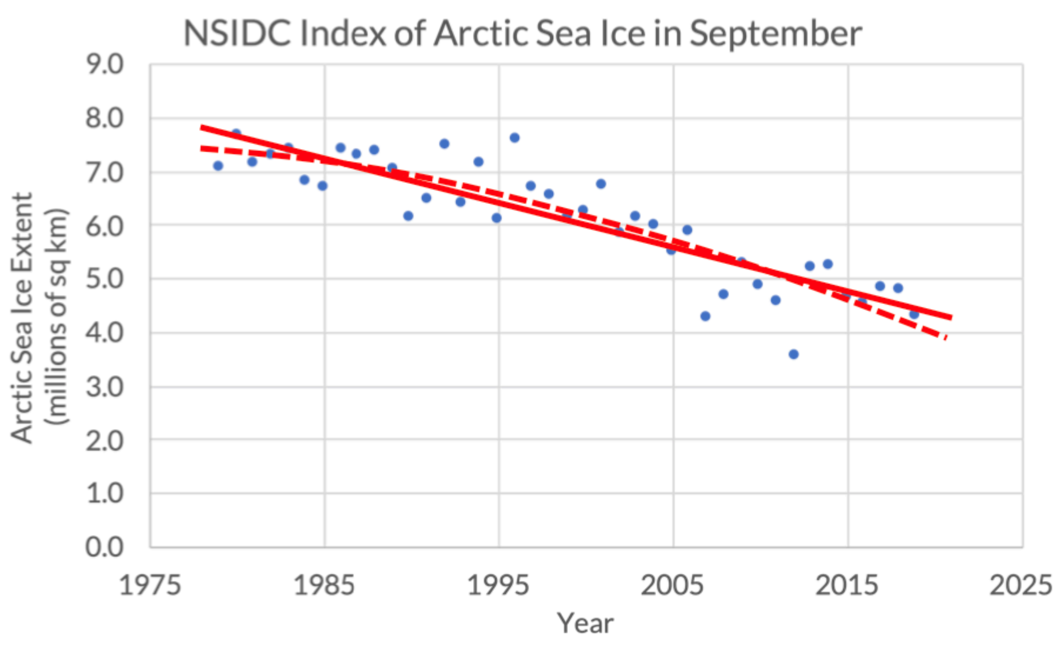
- Learn to output a **category** label
  - Binary (e.g. Spam / not Spam, Cat / not cat)
  - Multi-class (e.g. cat, dog, bird)

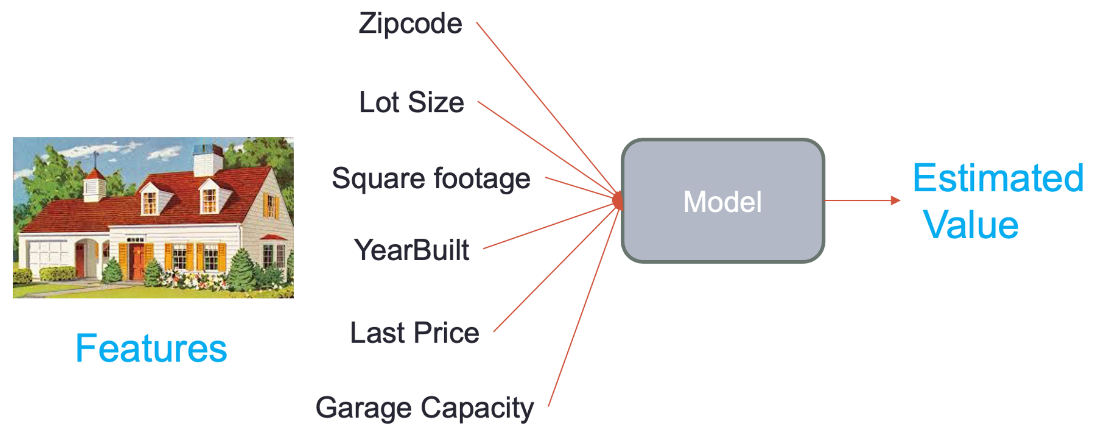
Regression

- Learn to guess one or more numbers
  - e.g. value of a share, number of stars in a review

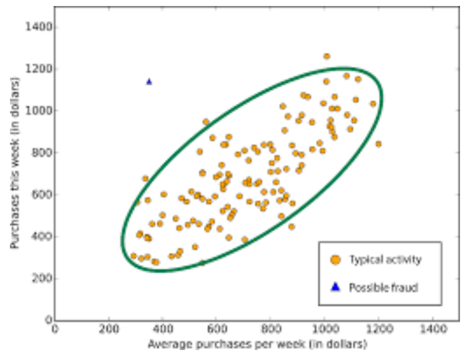




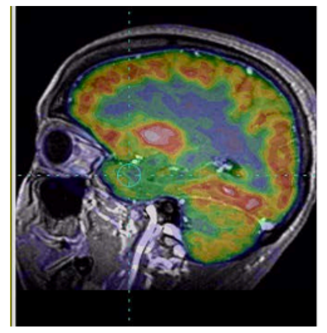




Estimating Home Prices



Fraud Detection



Medical Diagnosis

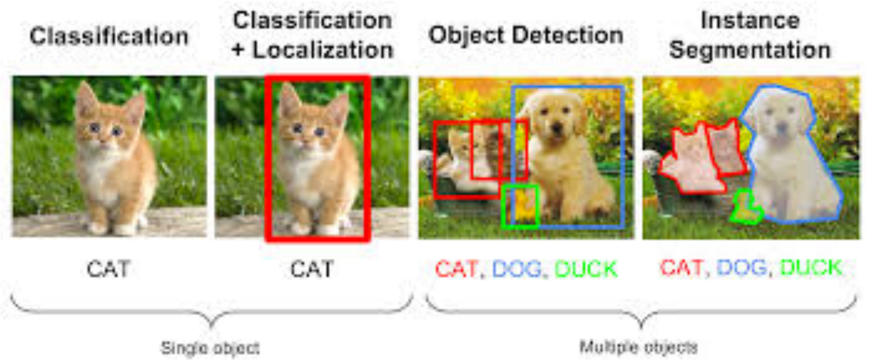
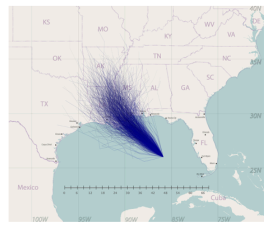
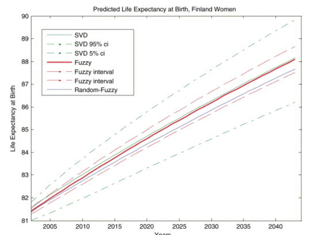


Image Classification

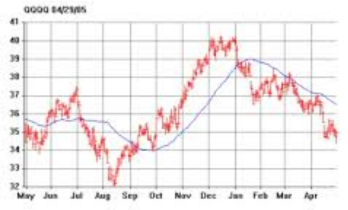
Advertising Reach Prediction



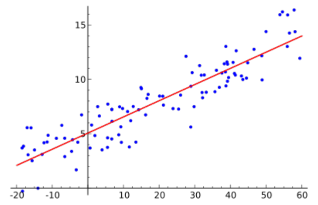
Weather Forecasting



Estimating Life Expectancy



Sales Growth Prediction

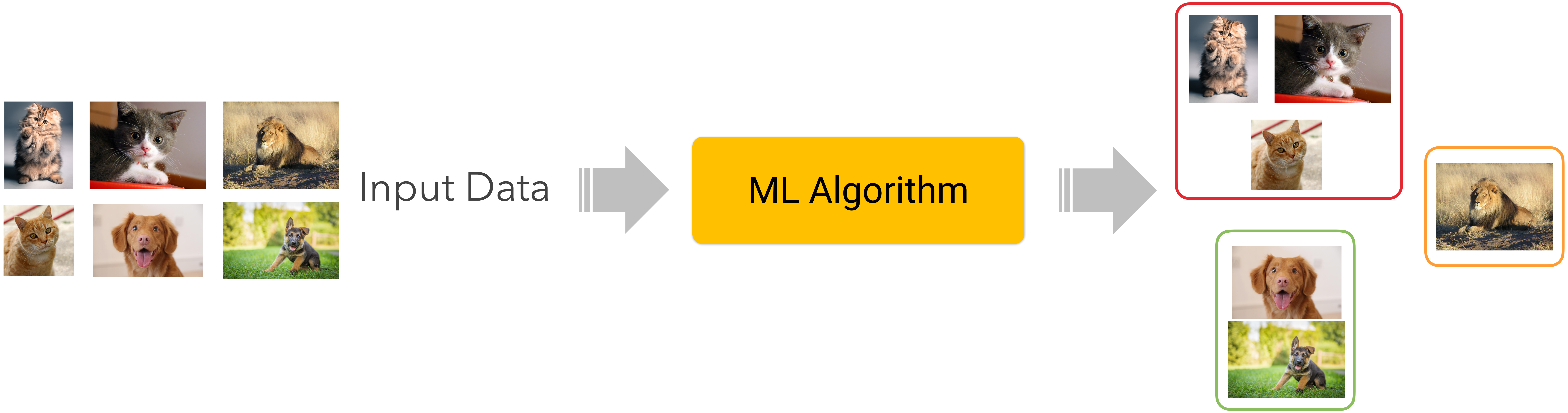


Market Forecasting

Unsupervised Learning

Unlabelled data in input

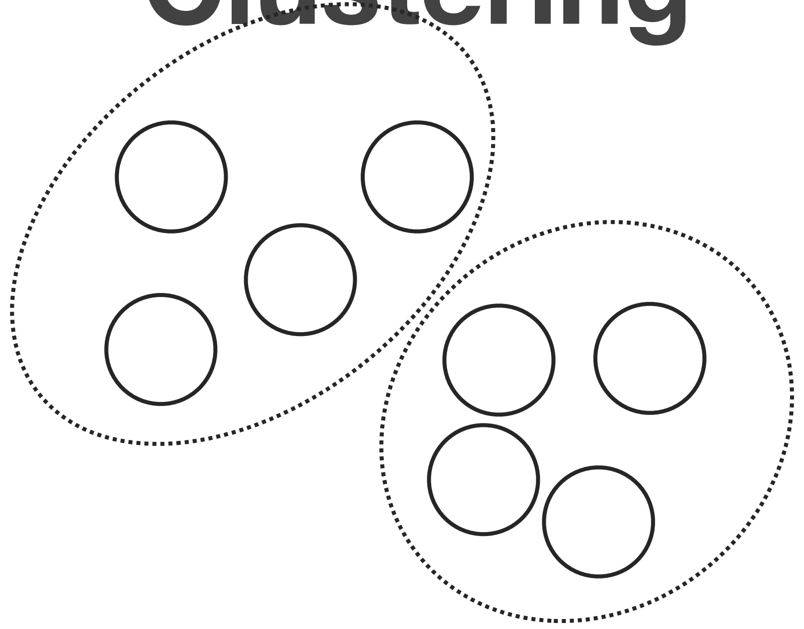
The machine learns structures (patterns) from the data without human guidance

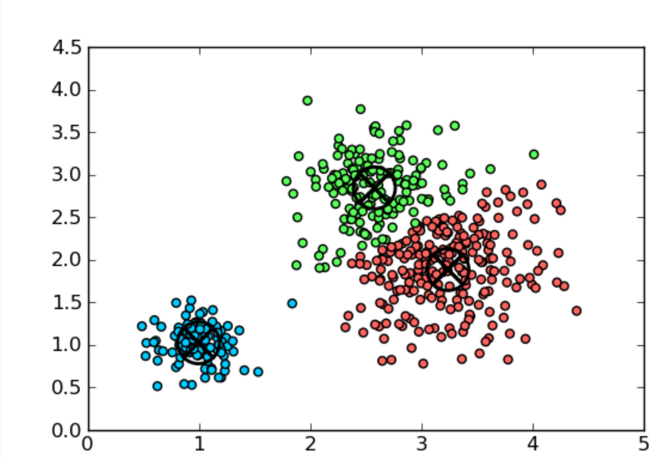


Clustering	Dimensionality Reduction	Anomaly Detection	Representation Learning

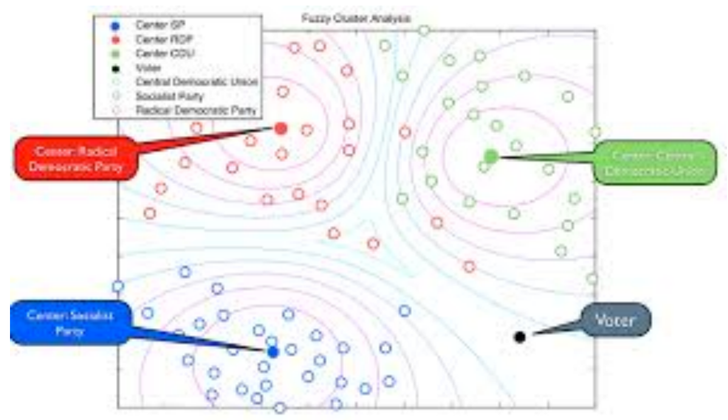
Example applications

Clustering





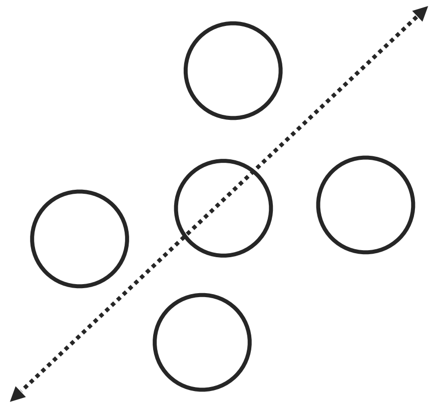
Customer Segmentation

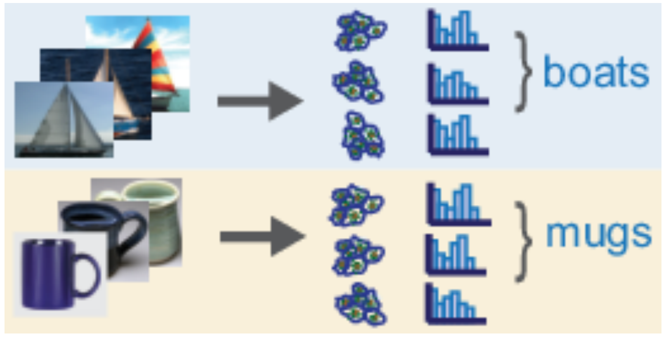


Targeted Marketing

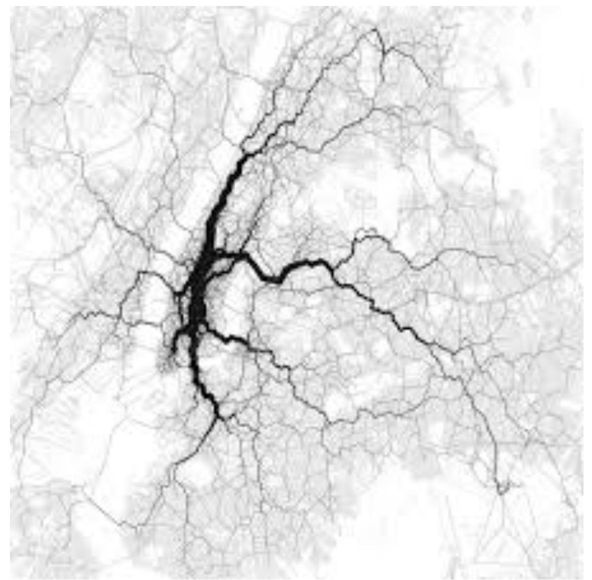
Recommender Systems

Dimensionality Reduction

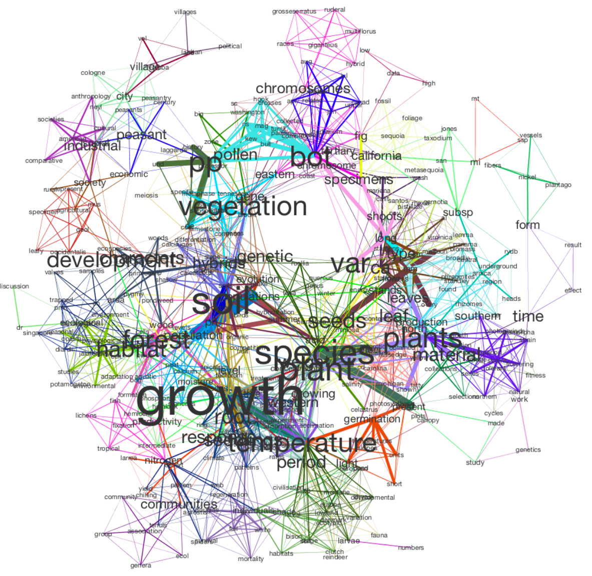




Foundational Models For Transfer Learning

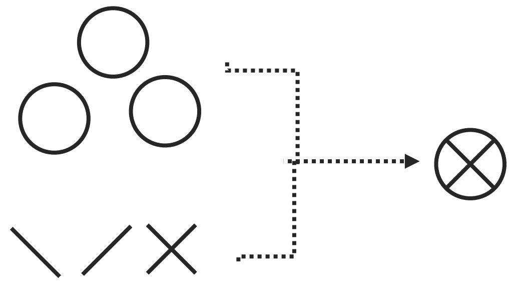


Big Data Visualisation



Structure Discovery

Generative Networks



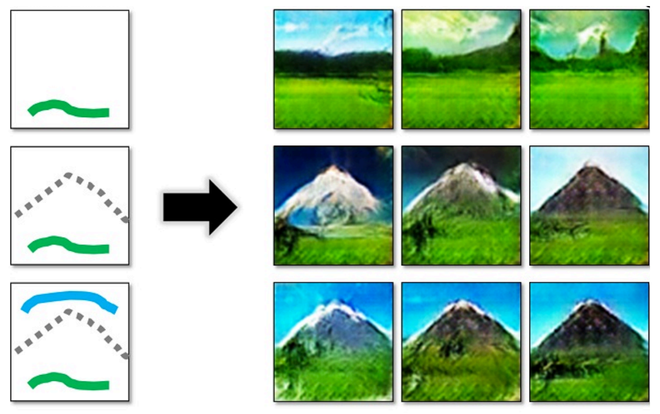
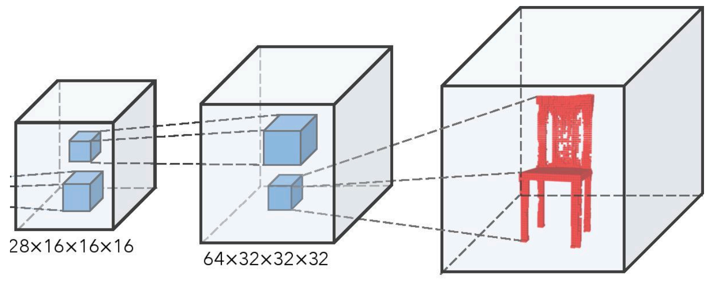
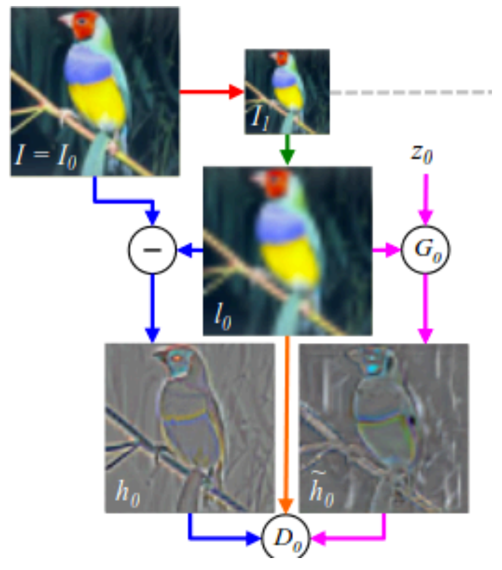


Image Generation

Music Generation



2D to 3D Modelling



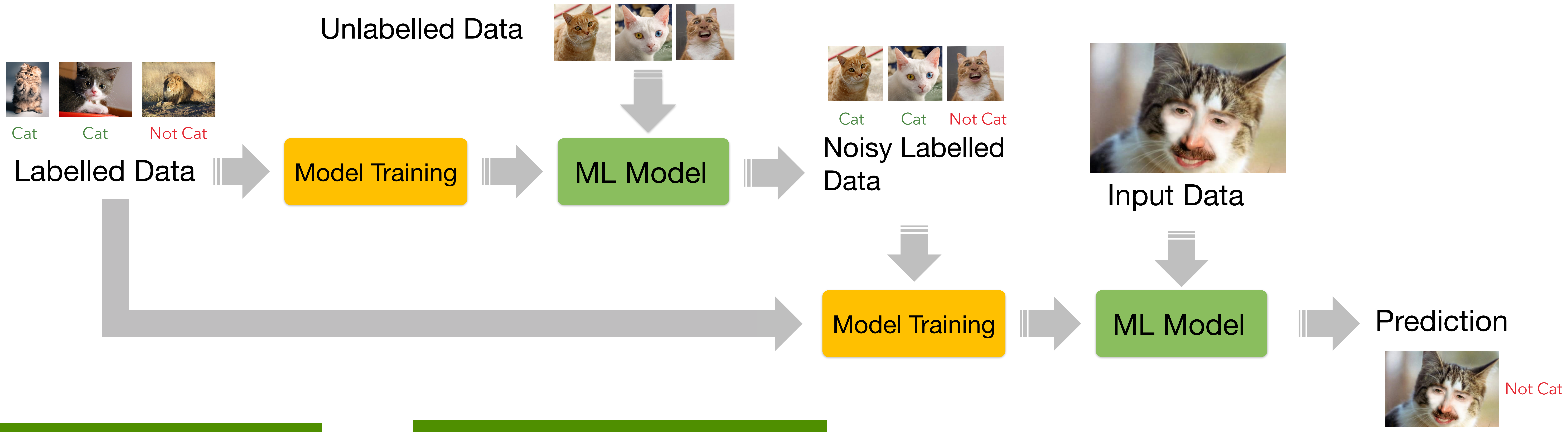
Pattern Modelling

Semi-Supervised Learning

Combination of **supervised** and **unsupervised** learning

Few **labelled** data in input are used to create **noisy labelled data**

With more labelled data, the machine learns how to make input-output **predictions**

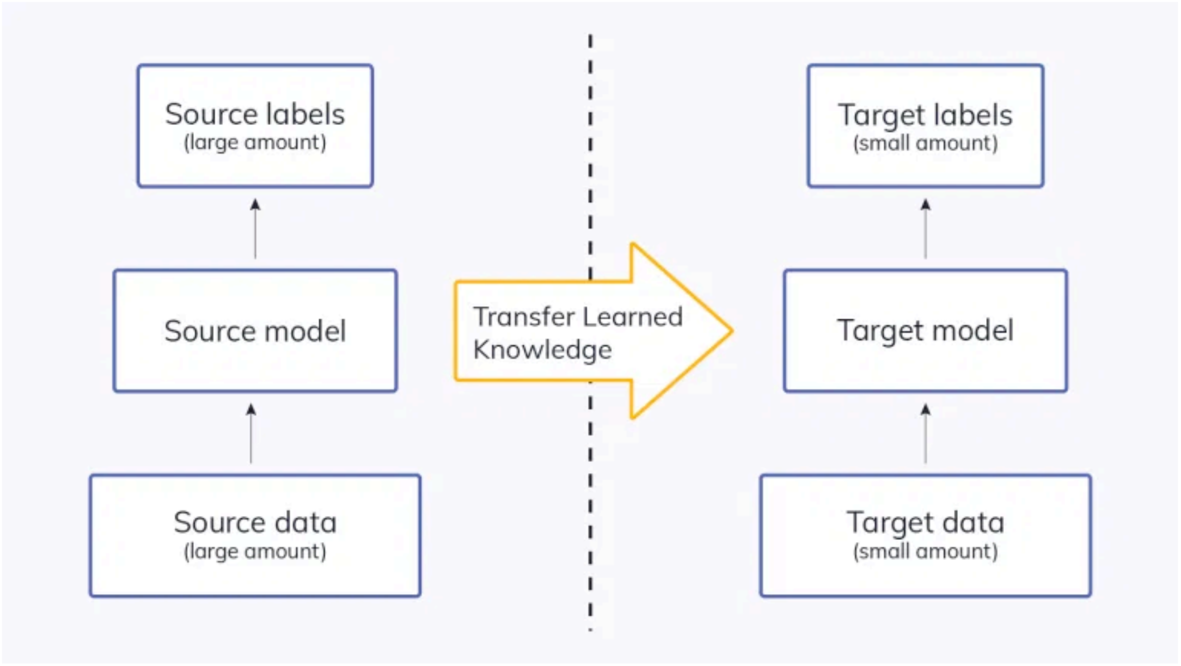


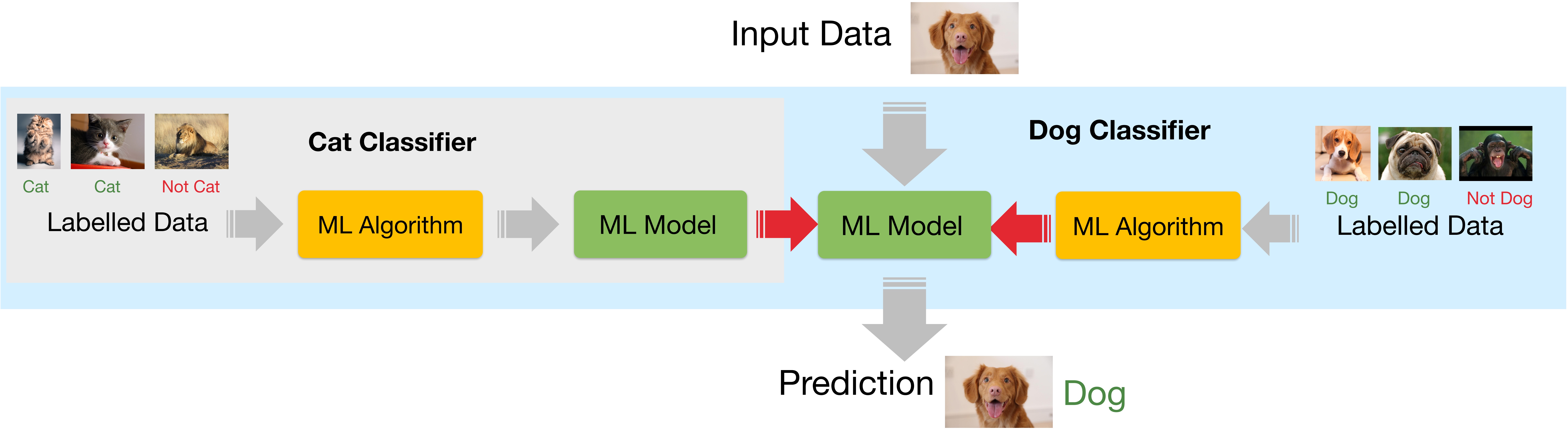
Supervised Learning + **Semi-supervised Clustering**

Transfer Learning

Reuse a model trained for one task is re-**purposed** (tuned) on a different but related task

Useful in tasks laking abundant data



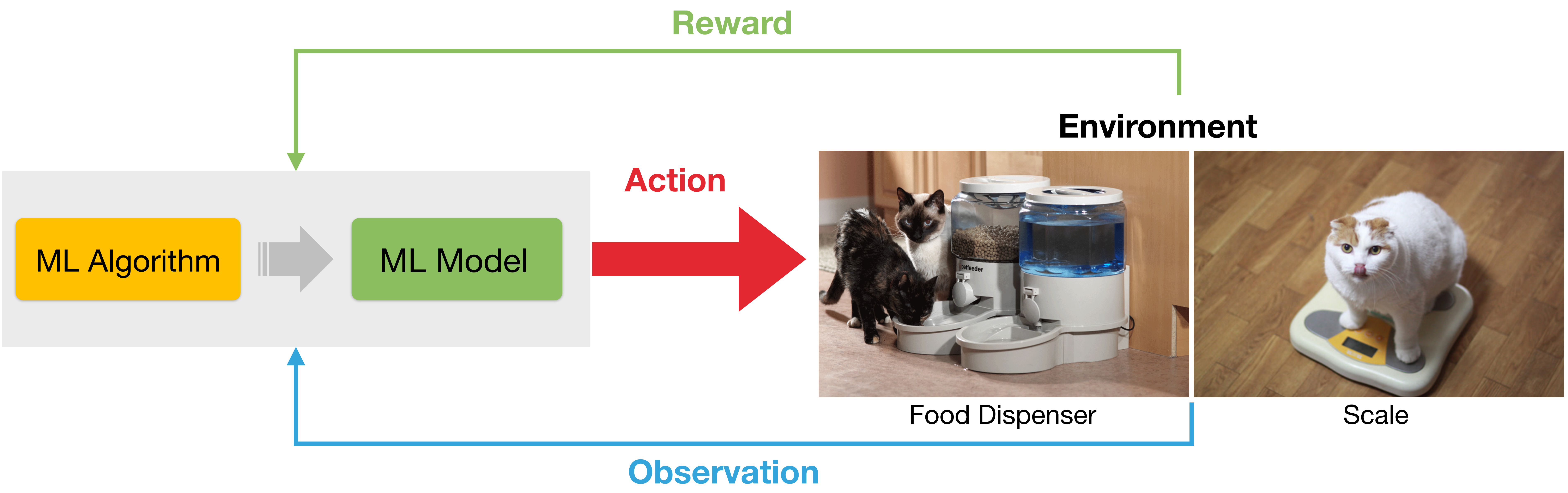


Reinforcement Learning

Data about the **environment** and **reward function** as input

The machine can perform **actions** influencing the environment

The machine learns behaviours that result in **greater reward**



Don’t forget domain expertise

- ML makes some tasks automatic, but we still need our brains
  - Defining the prediction task
  - Define the evaluation metrics
  - Designing features
  - Designing inclusions and exclusion criteria for the data
  - Annotating (hand-labeling) training (and testing) data
  - Select right model
  - Error analysis

More in Module 3 and 4

Machine Learning For Design

Lecture 2 - Fundamentals of Machine Learning

Alessandro Bozzon

11/02/2022

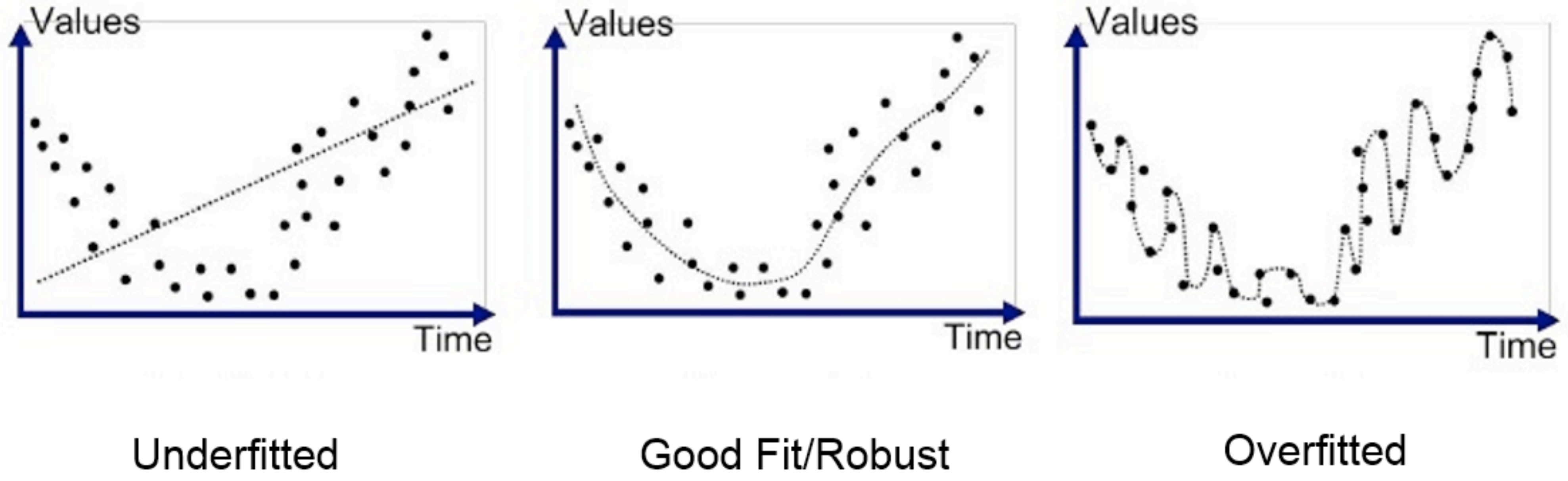
mlfd-io@tudelft.nl
www.ml4design.com

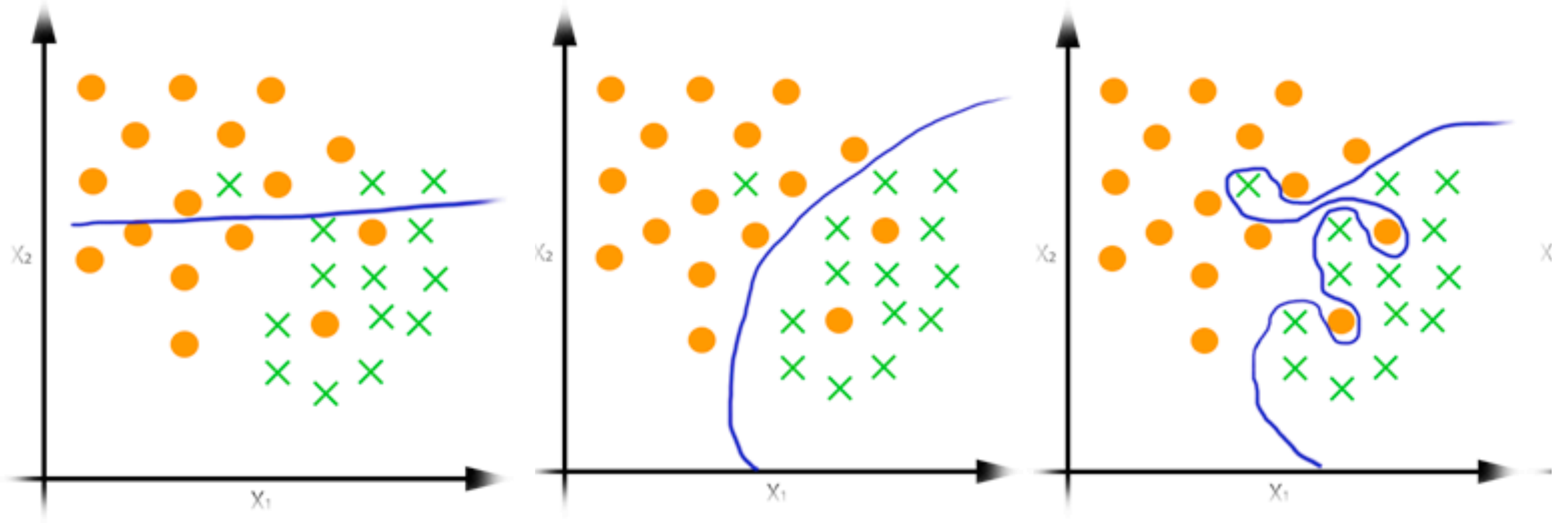
No free-lunch

- There is no one best machine learning algorithm for all problems and datasets

- Challenge: achieving good generalization and a small error rate
  - protect against **overfitting**
    - learning a model that too closely matches the idiosyncrasies of the training data
  - **underfitting**
    - learning a model that does not adequately capture the patterns in the training data

Regression





Classification

How to evaluate?

- Errors are almost inevitable!
  - How to measure errors?
- Select an evaluation procedure (a “metric”)
  - **Ok, but which one?**

Classification

- **Accuracy**
  - In Classification, the model with highest accuracy is not necessarily the best model
  - Some errors (e.g. False Negative) may be much more expensive than others
    - Usually due to imbalanced trained datasets

$$Accuracy = \frac{\#CorrectPredictions}{\#Predictions}$$

- **Confusions Matrix**
  - Describes the complete performance of the model

True Positive

Actual Class

Predicted Class		**Yes**	**No**
	Yes	50	10
	No	40	100

False Negative (**Type-1 Error**)

True Negative

False Positive (**Type-2Error**)

$$Accuracy = \frac{\#TruePositices + \#FalseNegatives}{\#AllPredictions}$$

Classification

- **Sensitivity** (True positive rate)
  - probability of a positive classification, conditioned on being in the correct class

$$Sensitivity = \frac{TruePositive}{FalseNegative + TruePositive}$$

- **Specificity** (False positive rate)
  - probability of a negative classification, conditioned on not being in the correct class

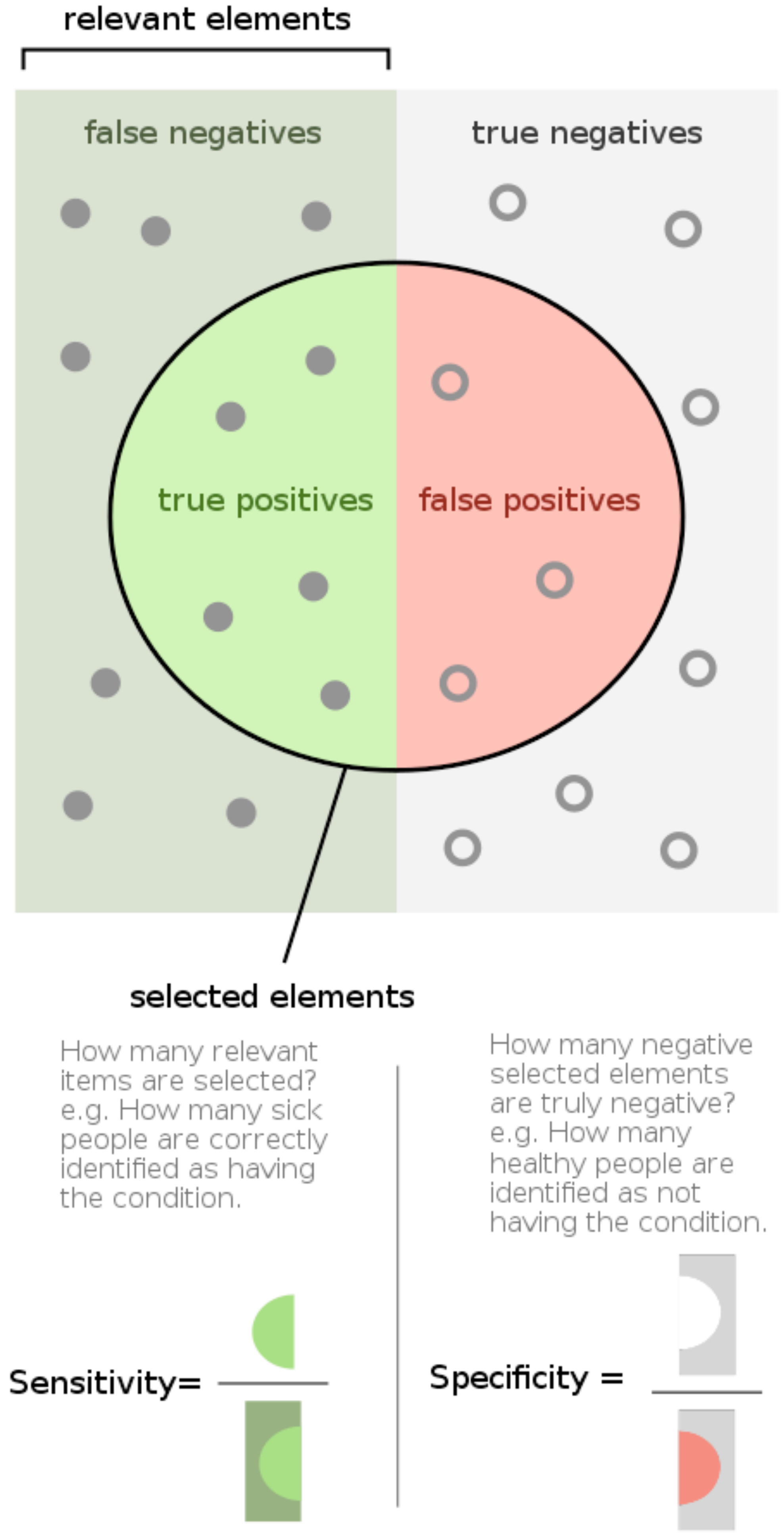
$$Specificity = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

- **F1-Score**
  - Harmonic mean between **precision** (how many instances correctly classified), and **recall** (how many relevant instance are correctly classified)

$$F_1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$



https://upload.wikimedia.org/wikipedia/commons/thumb/5/5a/Sensitivity_and_specificity_1.01.svg/512px-Sensitivity_and_specificity_1.01.svg.png

How to evaluate?

- Errors are almost inevitable!
  - How to measure errors?
- Select an evaluation procedure (a “metric”)
  - **Ok, but which one?**
- Compare to one or more baselines
  - trivial solution
  - rule-based solution
  - existing solution
- Apply your model to a held-out test set and evaluate
  - the test set must be different from the training set

More in Module 3 and 4