Applied Data Science

L1. Learning from Data. Overview of Machine Learning/ Data Science approaches.

Irina Mohorianu

Head of Bioinformatics/ Scientific Computing @CSCI

Email: iim22@cam.ac.uk

Structure of the module. Requirements.

Learning from data

L1 Learning from data. Overview of Machine Learning/ Data Science approaches

L2 Learning from data. Cross-validation. Data pre-processing

Data Science toolkit. Python Scientific Computing Ecosystem

L3 Data Science toolkit. Scikit learn. Visualisation

L4 Data Science toolkit. NumPy. Matplotlib

L5 Data Science toolkit. SciPy. Help and documentation

W4. Exercise sheet 1.
Cross-validation. Linear Regression

Supervised Learning

L6 Regression 1 [linear, logistic]

L7 Regression 2 [+ overfitting/ underfitting and regularisation approaches]

L8 Classification

L9 k nearest neighbours

L10 Support vector machines

L11 Kernelisation

W5/W6. Exercise sheet 2.
kNNs, SVMs

Structure of the module. Requirements.

Supervised Learning

L12 Tree based models

L13 Ensemble methods. Random Forests

L14 Ensemble methods. Bagging

L15 Ensemble methods. Boosting

AdaBoost; Stack methods will also be briefly discussed.

W7. Exercise sheet 3.
Tree based approaches

W8. Exercise sheet 4.
Dimensionality Reduction. Clustering

Unsupervised learning

L16 Dimensionality reduction. Linear methods. PCA

L17 Dimensionality reduction. Non-Linear methods. tSNE & UMAP

L18 Traditional clustering algorithms. Hierarchical clustering. K-means

L19 Modern clustering algorithms

Density based clustering. Graph based clustering

L20 Assessment of clustering outputs

Cluster-based methods and element centric methods

L21 Fuzzy clustering. C-means & consensus clustering

L22 Outlier detection

Structure of the module. Requirements.

Neural Networks

L23 Neural networks 1

L24 Neural networks 2

Non-examinable.

Assessments:

67% Assessed coursework (end of term).

focused on a large, curated input with hundreds of samples and measurements
comprising three parts – covering aspects of both supervised and unsupervised learning

33% Written exam (January 2024)

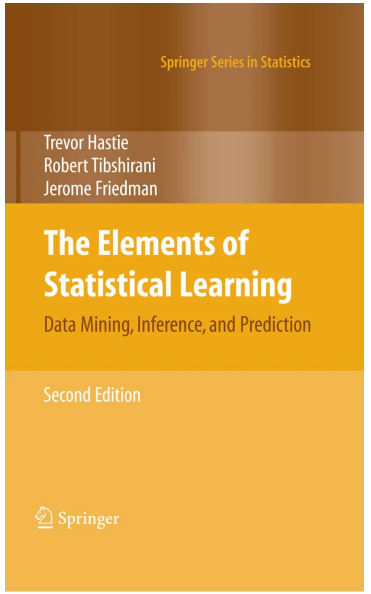
Useful references

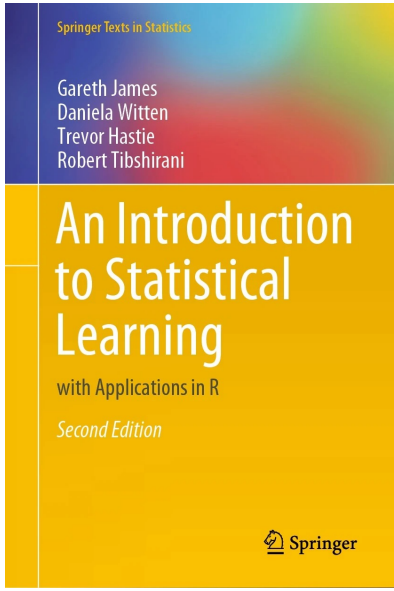
James, Witten, Hastie, Tibshirani Introduction to Statistical Learning
Python: https://www.statlearning.com/resources-python
R: https://www.statlearning.com/

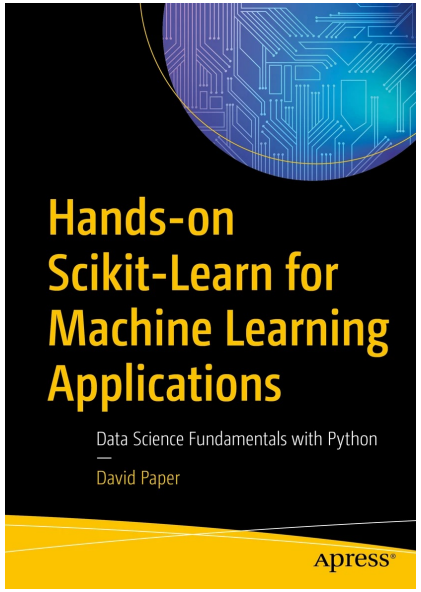
Hastie, Tibshirani, Friedman Elements of Statistical Learning
https://link.springer.com/book/10.1007/978-0-387-84858-7

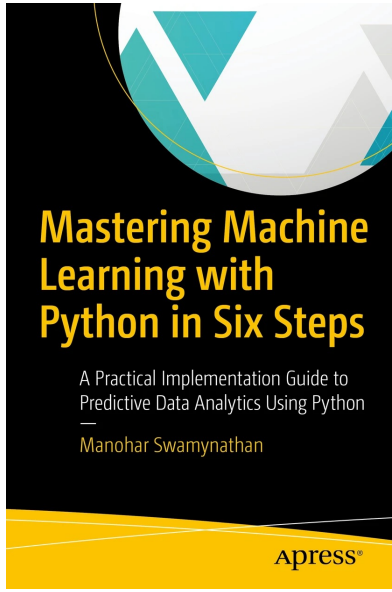
David Paper Hands-on Scikit-Learn for Machine Learning applications
Data Science Fundamentals with Python
https://link.springer.com/book/10.1007/978-1-4842-5373-1

M Swamynathan Mastering Machine Learning with Python in Six Steps
A practical implementation guide to predictive data analysis using Python.
https://link.springer.com/book/10.1007/978-1-4842-2866-1
[two editions are available]









This list is not exhaustive.

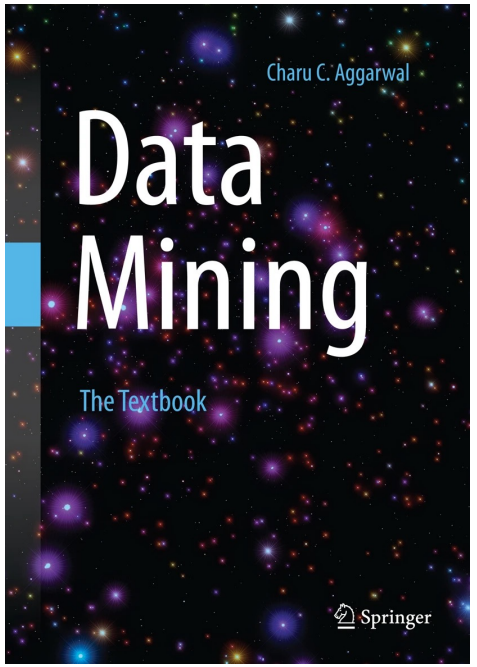
Useful references

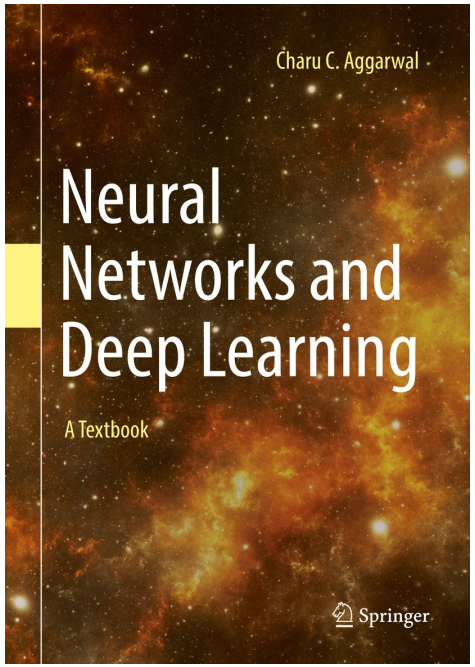
Aggarwal Neural Networks and Deep Learning
https://link.springer.com/book/10.1007/978-3-319-94463-0

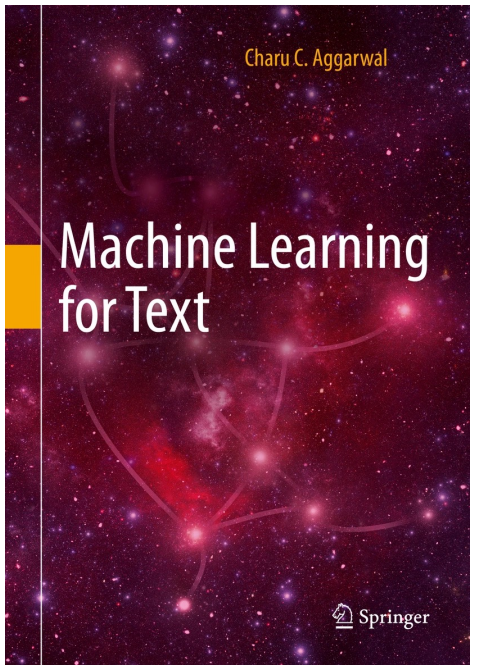
Aggarwal Data mining
https://link.springer.com/book/10.1007/978-3-319-14142-8

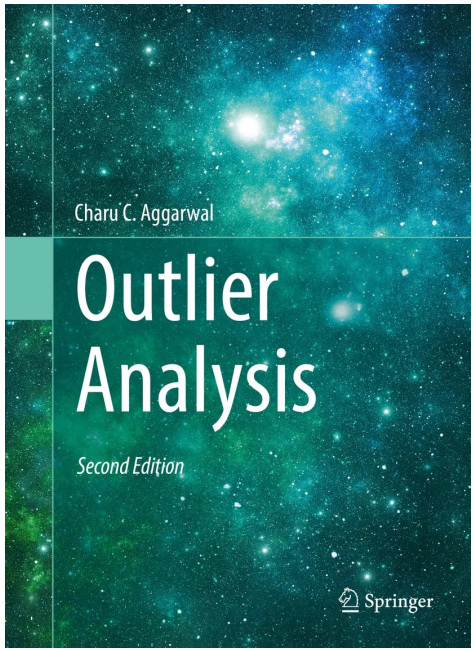
Aggarwal Machine Learning for Text
https://link.springer.com/book/10.1007/978-3-319-73531-3

Aggarwal Outlier Analysis
https://link.springer.com/book/10.1007/978-3-319-47578-3









This list is not exhaustive.

What is Applied Data Science/ Machine Learning?

What do you see?

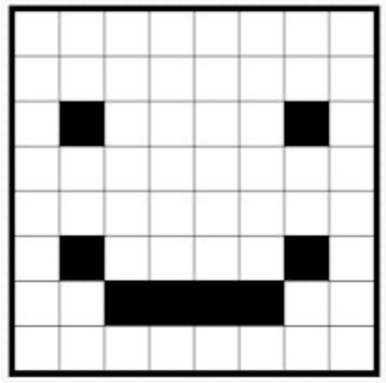
We need to classify the pictures with a dog and the ones with the muffin.

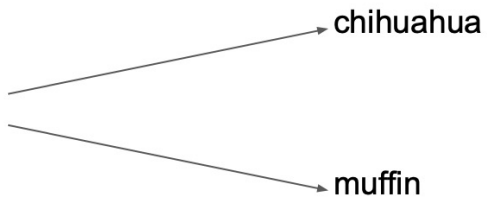
Or do we need to partition the images and manually annotate some of them?

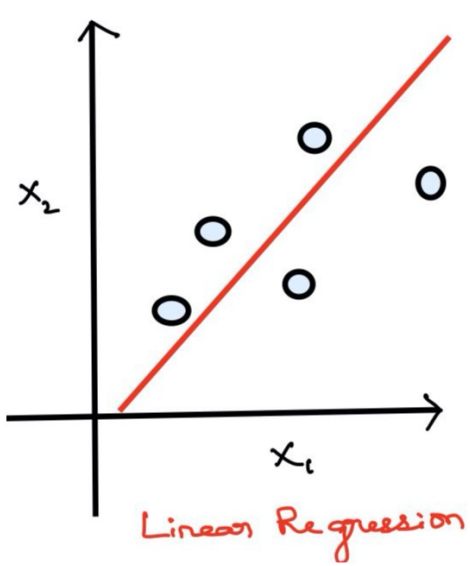
What is Applied Data Science/ Machine Learning?

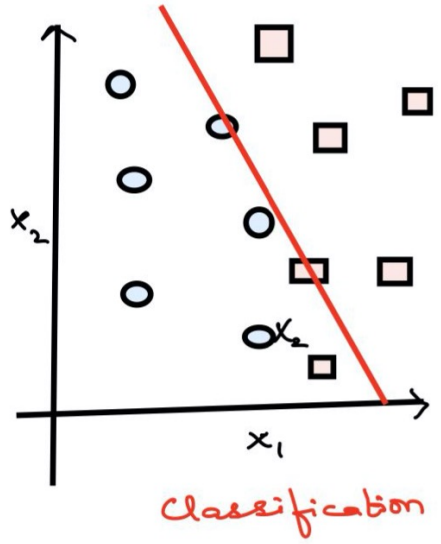
New York Times reported in 1958 that the invention was the beginning of a computer that would “be able to walk, talk, see, write, reproduce itself and be conscious of its existence”

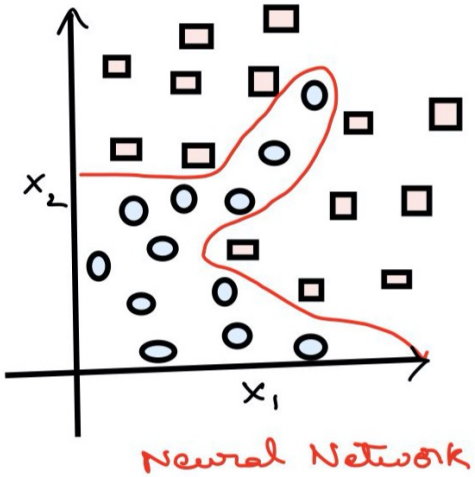
What is Applied Data Science/ Machine Learning?











Artificial Intelligence vs Machine Learning

Greek Mythology

Pygmalion of Cyprus sculpts a marble statue of Galateea.
He falls in love with the sculpture and it comes to life.

George Bernard Shaw's Pygmalion

Higgins teaches Eliza Doolittle to speak Queen's English

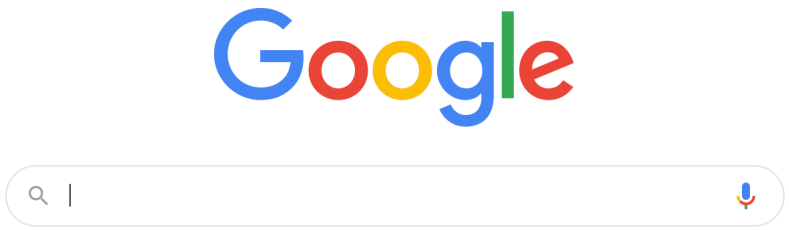
The rain in Spain stays mainly in the plain.

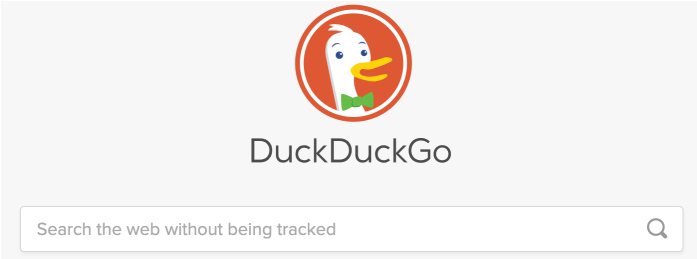
Eliza is the first AI program.
Weizenbaum, MIT

Artificial Intelligence vs Machine Learning

We can argue that today AI is ubiquitous.

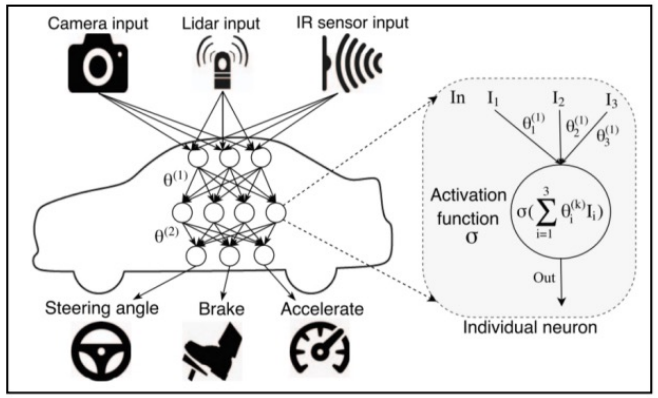
Automated searches





Speech recognition

Autonomous vehicles



Artificial Intelligence vs Machine Learning

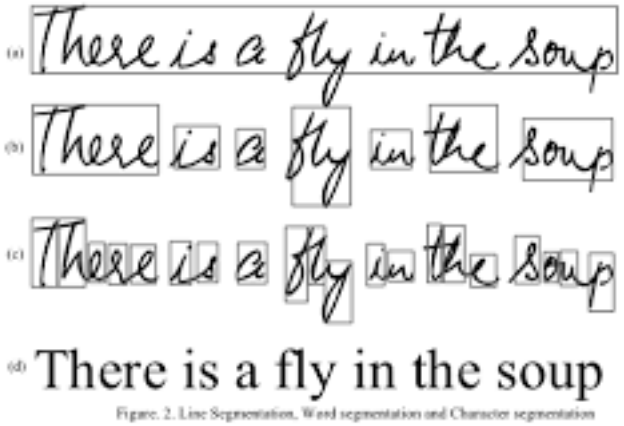


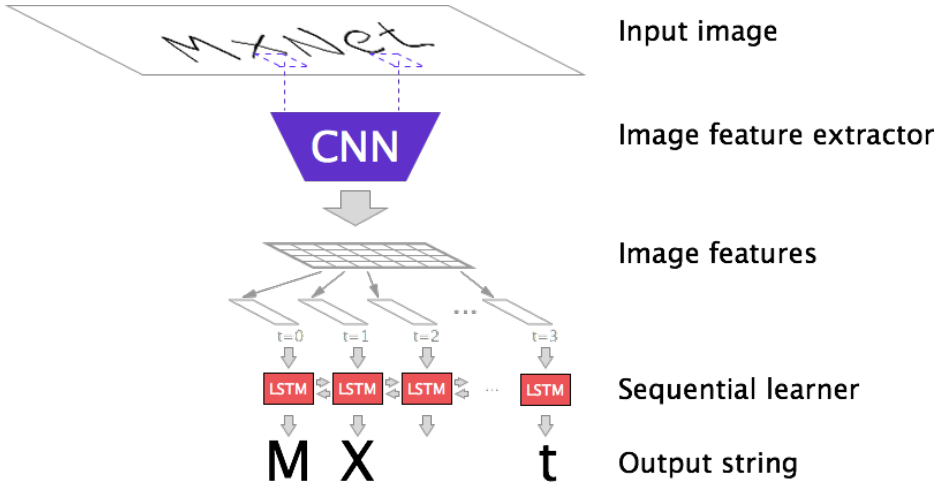
Figure. 2. Line Segmentation, Word segmentation and Character segmentation

Hand writing recognition

Rule-based approaches, numerous exceptions

Relies on training data

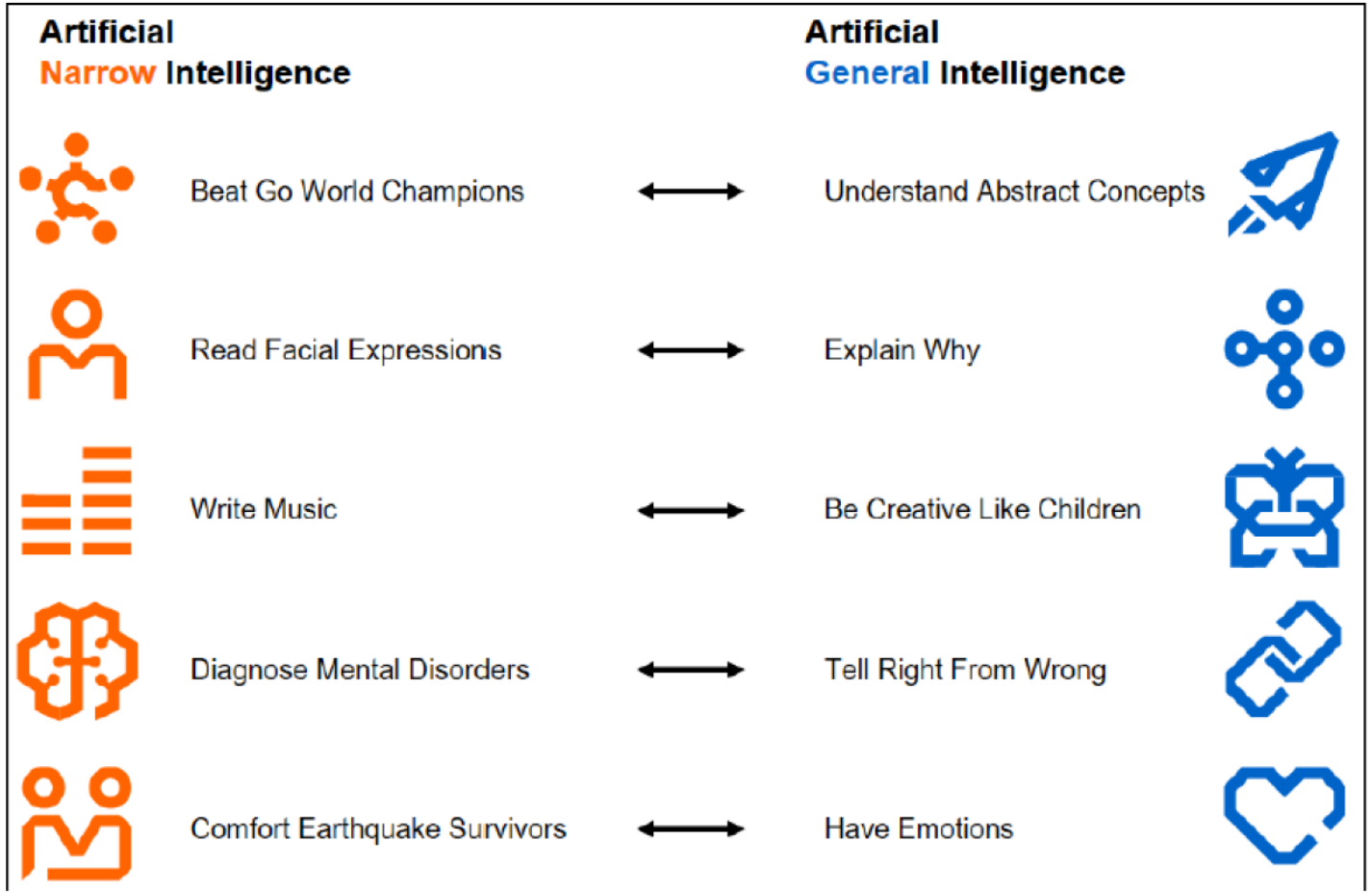
Cannot be done without Machine Learning



AI paradox

Hard problems for people are easy for AI

Easy problems for people are hard for AI



Where does the Machine Learning fit?

Tasks requiring intelligence:

Reasoning

puzzles, judgements

Planning

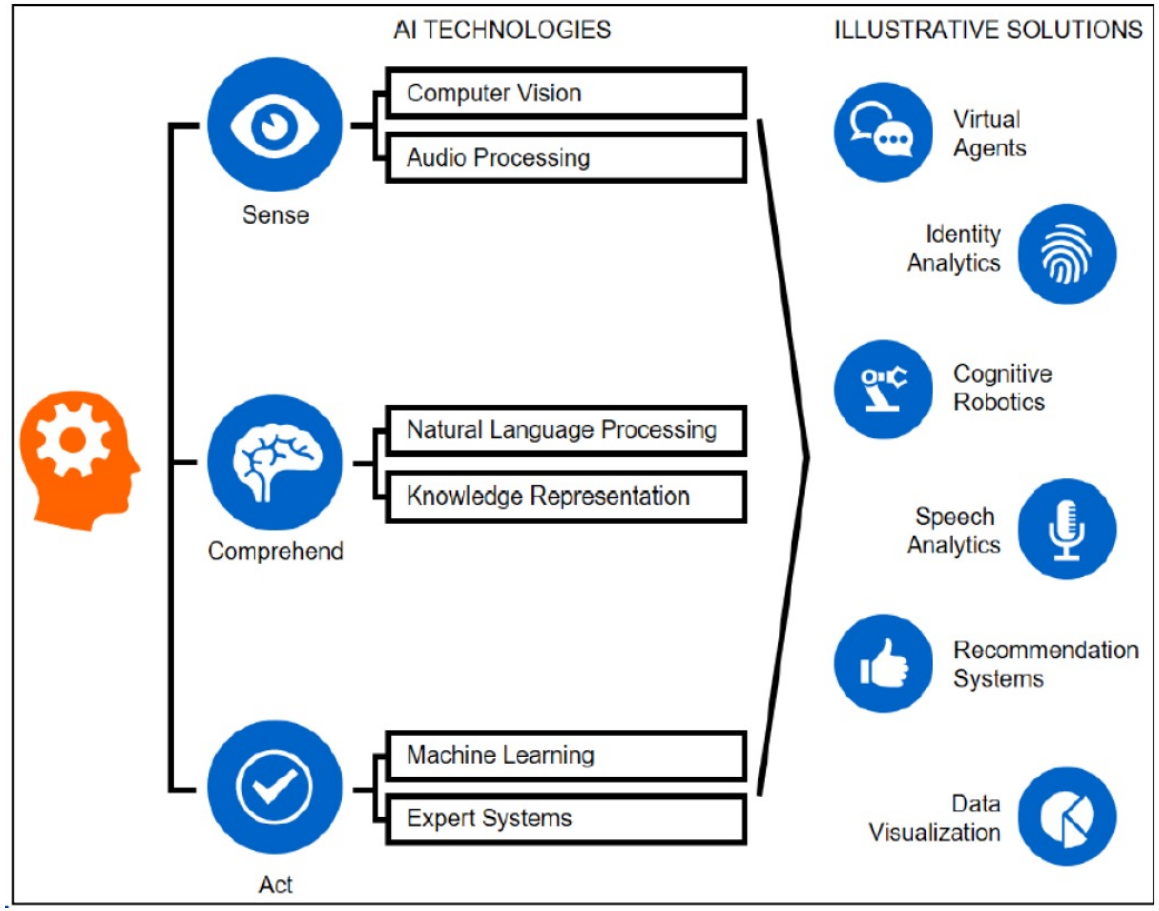
sequences of actions

Learning

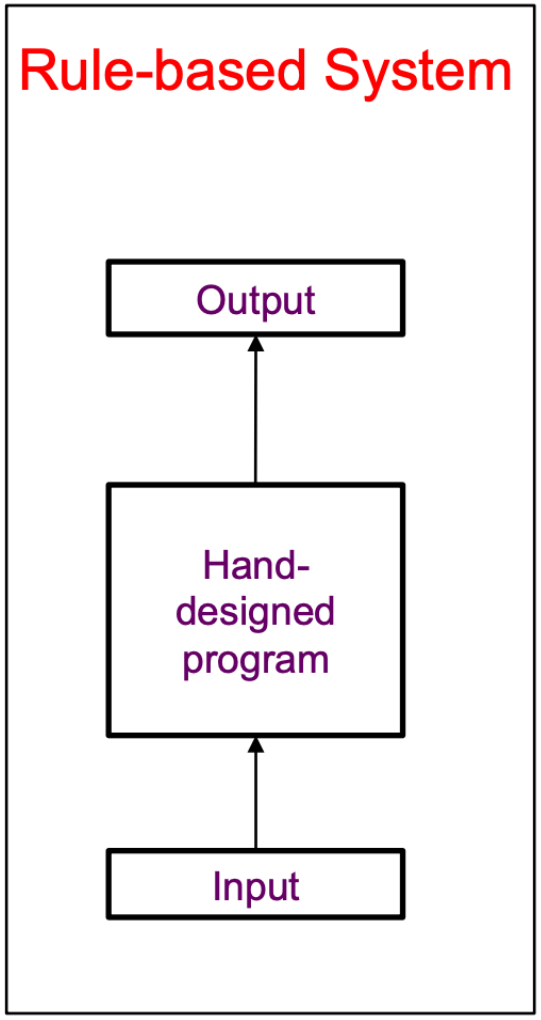
extracting information from data

Natural language

Integrating skills

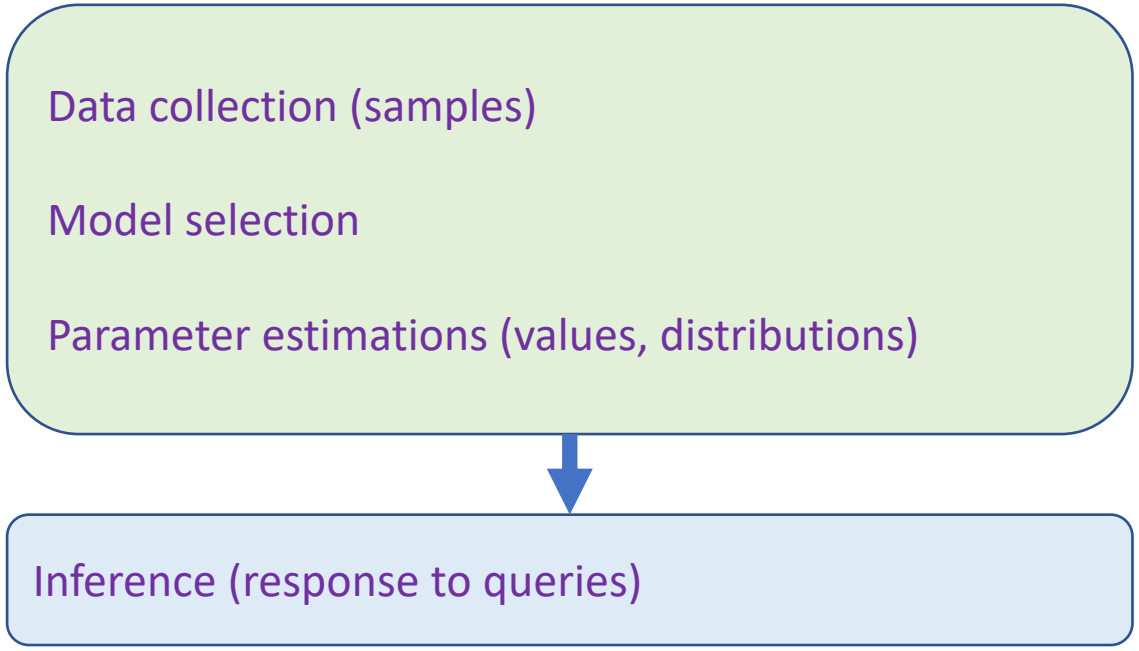


Knowledge based AI



Unwieldly process
Requires human experts (time and knowledge)
People struggle to formalise rules with enough complexity to capture variability

Machine Learning Approach



Training the model

Testing the model (decision inference)

What is Machine Learning?

Machine Learning (ML) is generating abstract hypotheses (models) based on data which can be later used predict results on new data.

Types of ML:
[1] Supervised
[2] Unsupervised
[3] Semi-supervised
[4] Reinforcement learning

What is Machine Learning?

Types of ML:
[1] Supervised
[2] Unsupervised
[3] Semi-supervised
[4] Reinforcement learning

Based on Type of Data

[1] supervised, unsupervised, semi-supervises
[2] reinforcement learning

Based on Type of Output

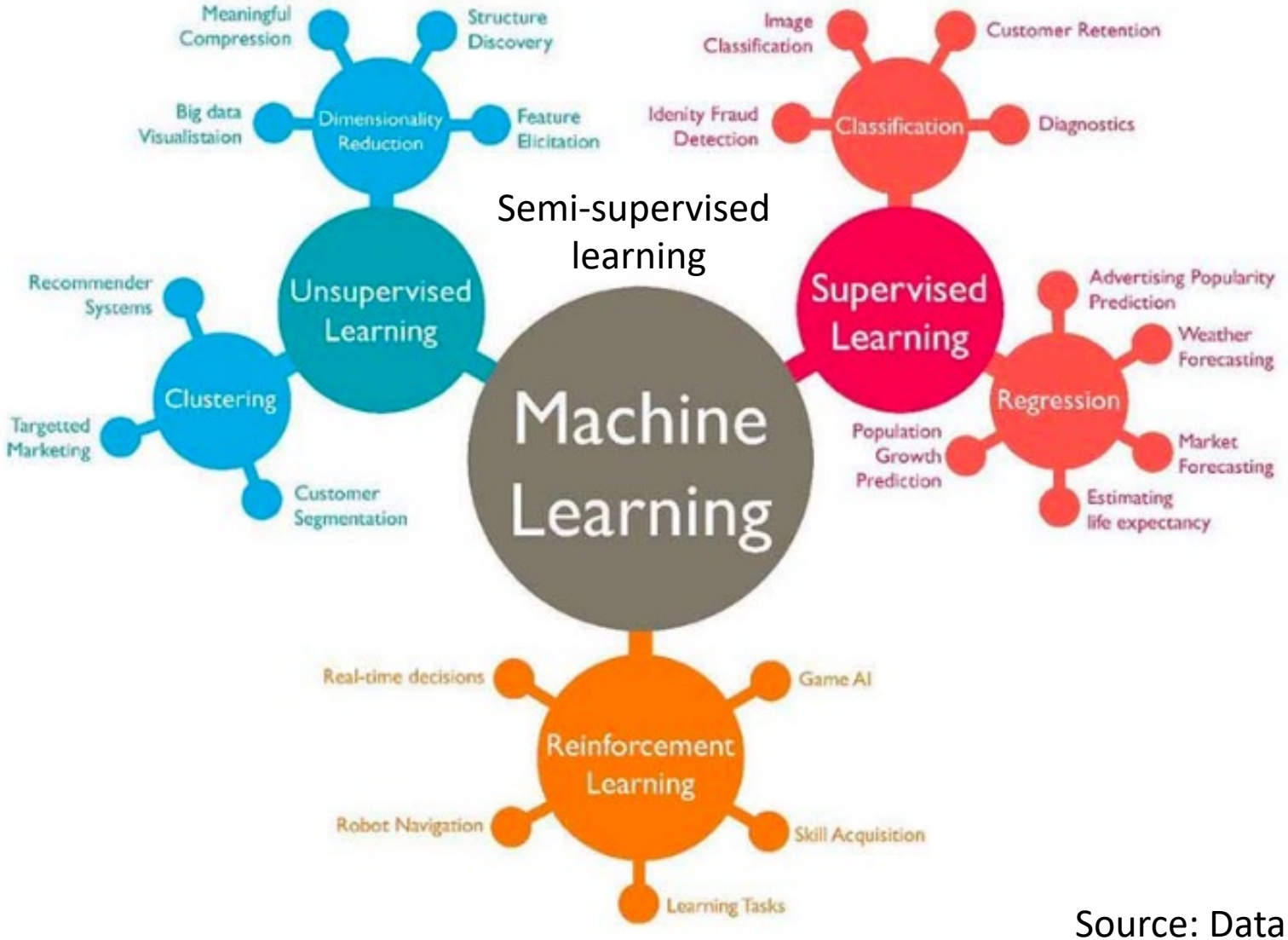
[supervised learning]

[1] regression
[2] classification

Based on Type of Model

[1] generative
[2] discriminative

Types of Machine Learning Algorithms

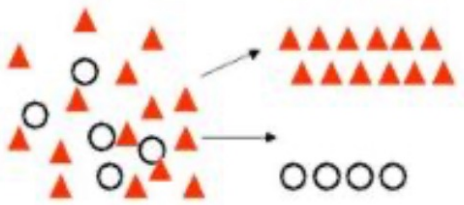


Source: Data Science Central

Common models used in Machine Learning

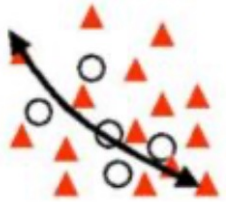
Techniques

Classification



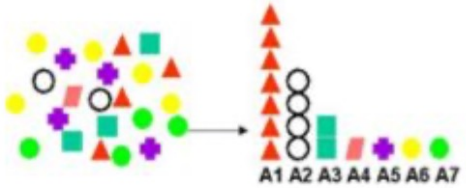
Separation of data point into classes.
Typical example: yes/no answers

Regression



Continuous numerical outcome.
Prediction of future values based on observed ones

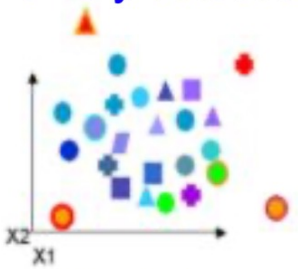
Attribute Ranking



Identification of important/ discriminative features based on their relationship with the target attribute.

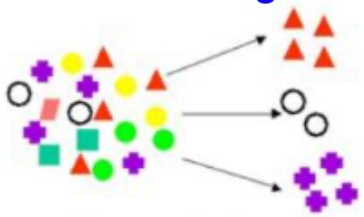
Common models used in Machine Learning

Anomaly detection



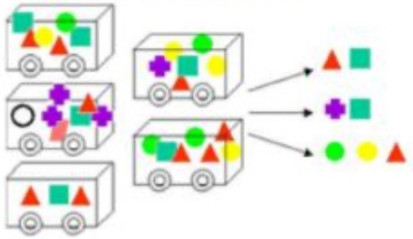
Identification of outliers

Clustering



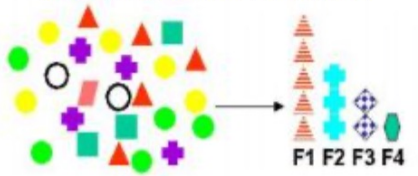
Finding groups/ structure in the data

Association



Finding rules associated with naturally co-occurring terms

Feature selection



Selection of best features to describe a target attribute

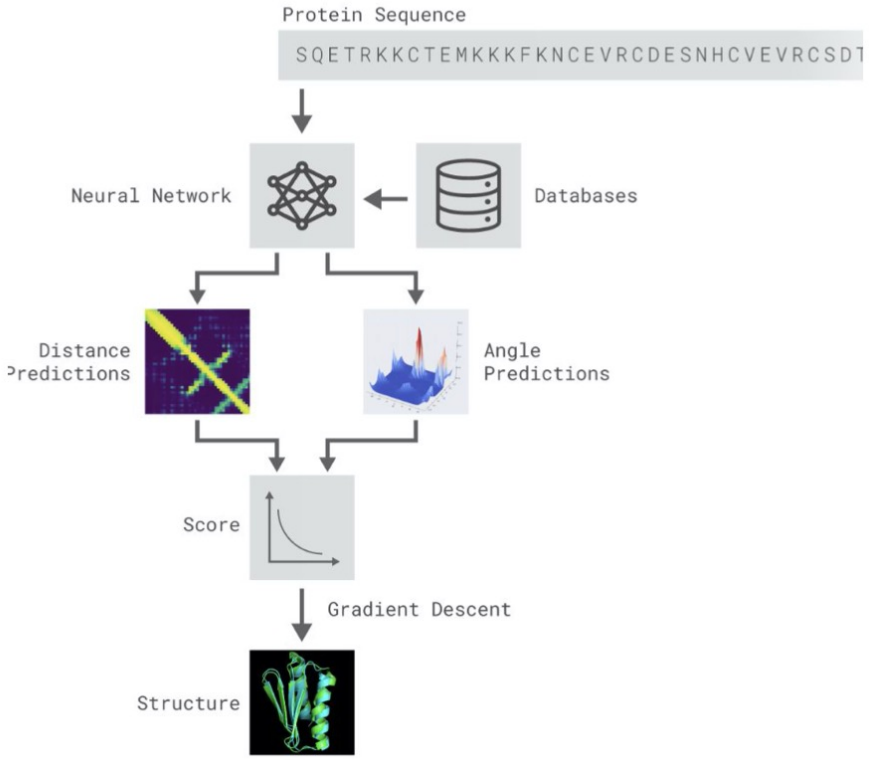
DS/ML examples. AlphaFold

AlphaFold: Using AI for scientific discovery

Today we're excited to share DeepMind's first significant milestone in demonstrating how artificial intelligence research can drive and accelerate new scientific discoveries. With a strongly interdisciplinary approach to our work, DeepMind has brought together experts from the fields of structural biology, physics, and machine learning to apply cutting-edge techniques to predict the 3D structure of a protein based solely on its genetic sequence.

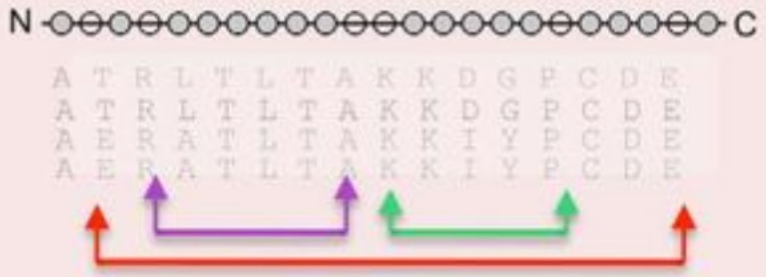
Our system, **AlphaFold**, which we have been working on for the past two years, builds on years of prior research in using vast genomic data to predict protein structure. The 3D models of proteins that AlphaFold generates are far more accurate than any that have come before—making significant progress on one of the core challenges in biology.

What is the protein folding problem?



DS/ML examples. AlphaFold

Align evolutionary diverged sequences



Calculate covariance matrix for each pair of sequence positions for all pairs of amino acids (A,B)

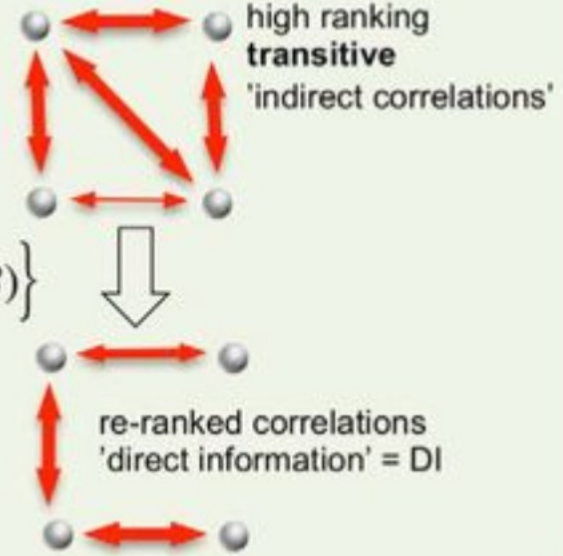
$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)P_j(B)$$

$$C_{ij}^{-1}(A,B) = -e_{ij}(A,B)_{i \neq j}$$

Identify maximally informative pair couplings using **statistical model** of entire protein to infer residue-residue co-evolution

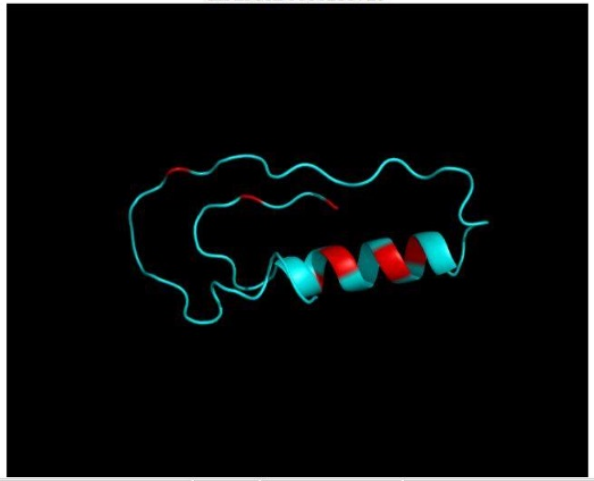
$$P_{ij}^{Dir}(A,B) = \frac{1}{Z}\exp\left\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}$$

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{Dir}(A,B)\ln\frac{P_{ij}^{Dir}(A,B)}{f_i(A)f_j(B)}$$



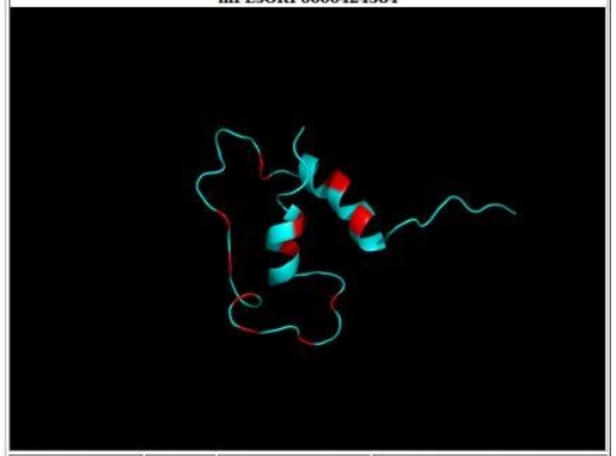
DS/ML examples. AlphaFold

mPLsORF0000239729



Mutation ID(s)	Change	Effect on sORF	sORF secondary structure
COSM3816594-AA=p.W105* COSM3816595-AA=p.W105*	G>A	W16>* (PDB W17)	H
COSM4056367-AA=p.Y91H COSM4056368-AA=p.Y91H	T>C	Y2>H (PDB Y3)	E
COSM5632701-AA=p.L109V COSM5632702-AA=p.L109V	C>G	L20>V (PDB L21)	H
COSM6324258-AA=p.A94V COSM6324259-AA=p.A94V	C>T	A5>V (PDB A6)	C
COSM6761144-AA=p.E103D COSM6761145-AA=p.E103D	G>T	E14>D (PDB E15)	H
COSM964154-AA=p.D111Y	G>T	D22>Y (PDB D23)	H
COSM964155-AA=p.R129H	G>A	R40>H (PDB R41)	C

mPLsORF0000424584



Mutation ID(s)	Change	Effect on sORF	sORF secondary structure
COSN1260962 COSN5770834	A>G	Y10>C (PDB Y10)	H
COSN18723188	C>T	P32>L (PDB P32)	C
COSN20080080	C>T	R13>C (PDB R13)	C
COSN20121983	C>T	I28>I (PDB I28)	C
COSN20648546 COSN21264986 COSN23079780	G>A	W50>* (PDB R50)	H
COSN20650206 COSN21677358 COSN22583798 COSN23084029	C>T	G42>G (PDB G42)	E
COSN20650207 COSN22687956 COSN23081188	T>C	W50>R (PDB R50)	H
COSN21264985 COSN22557586 COSN23085391	A>G	K49>K (PDB K49)	H
COSN21675097	T>C	Y43>Y (PDB Y43)	H
COSN23082255	G>A	K34>K (PDB K34)	H
COSN23087874	A>G	K26>E (PDB E26)	H
COSN24397730	G>A	R23>H (PDB R23)	H
COSN26666338	A>T	E67>V	H
COSN27001349 COSN8016762	G>A	T20>T (PDB T20)	C
COSN27001446	C>T	T20>M (PDB T20)	C
COSN27001936	G>T	R70>M	C
COSN27003013	G>A	G54>D (PDB G54)	C

DS/ML examples. Chemistry example

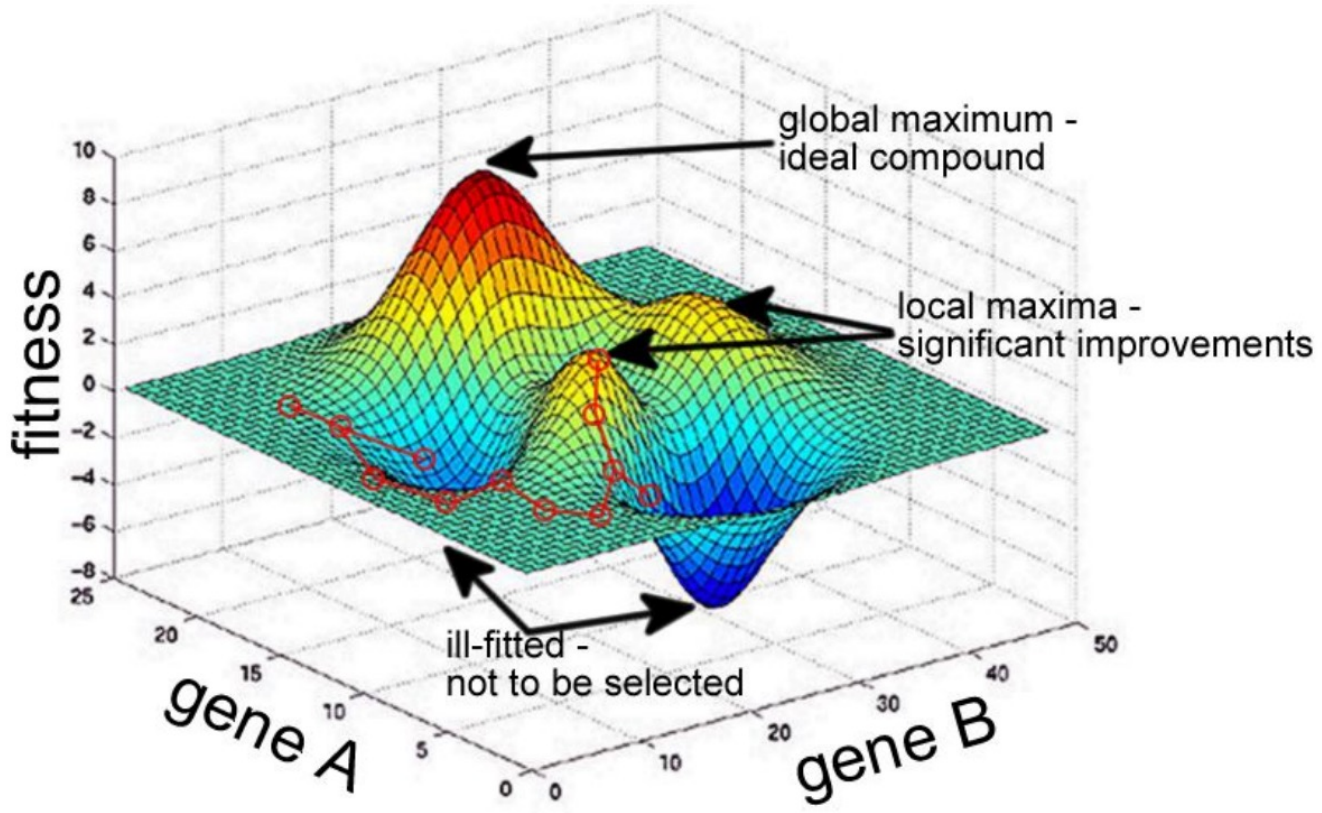


Figure 3 - A fitness landscape considering two hypothetical genes as, for example, block size and processing temperature of a polymer synthesis. Fitness could be hardness, for instance. The landscape contains local maxima, a global maximum, and a global minimum. In red, the path of a genetic algorithm along 11 iterations.

Source: Rodrigues et al, A survey of Big Data and Machine Learning for Chemistry

DS/ML examples. Chemistry example

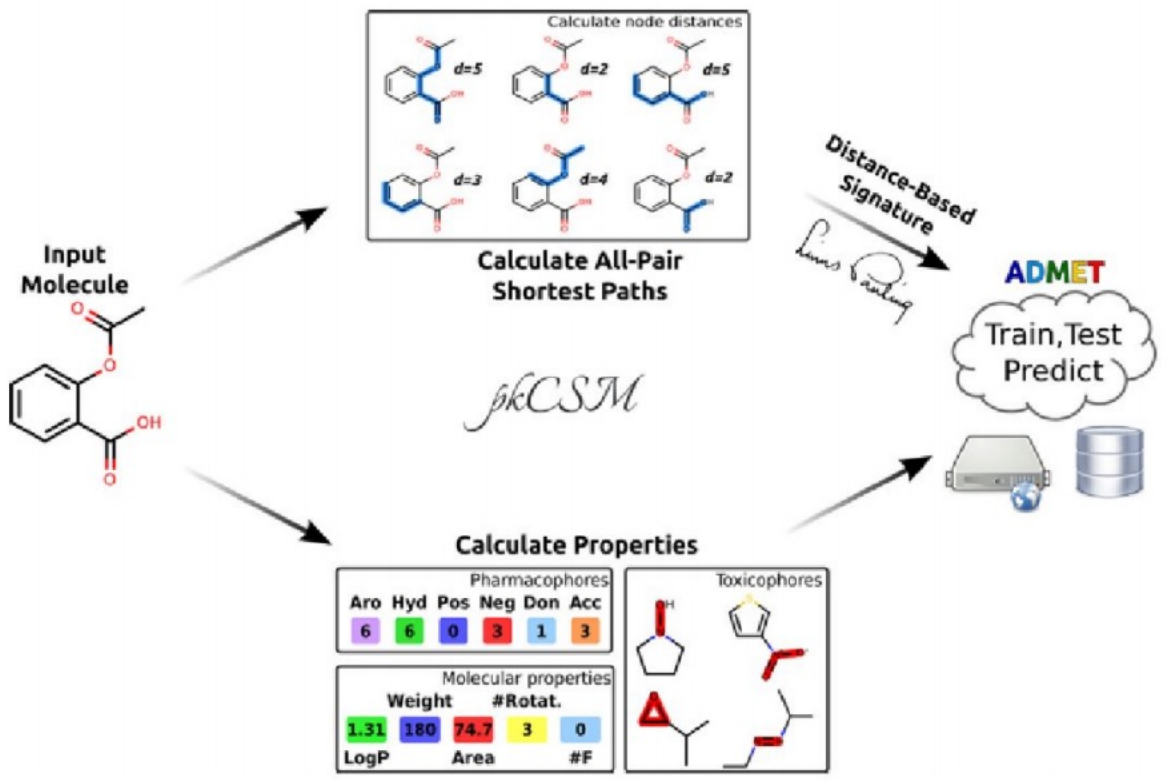


Figure 6 – The workflow of pkCSM is represented by the two main sources of information, namely the calculated molecular properties and shortest paths, for an input molecule. With these pieces of information, the ML system is trained to predict ADMET properties. Reproduced from [64].

Source: Rodrigues et al, A survey of Big Data and Machine Learning for Chemistry

DS/ML examples. Physics example

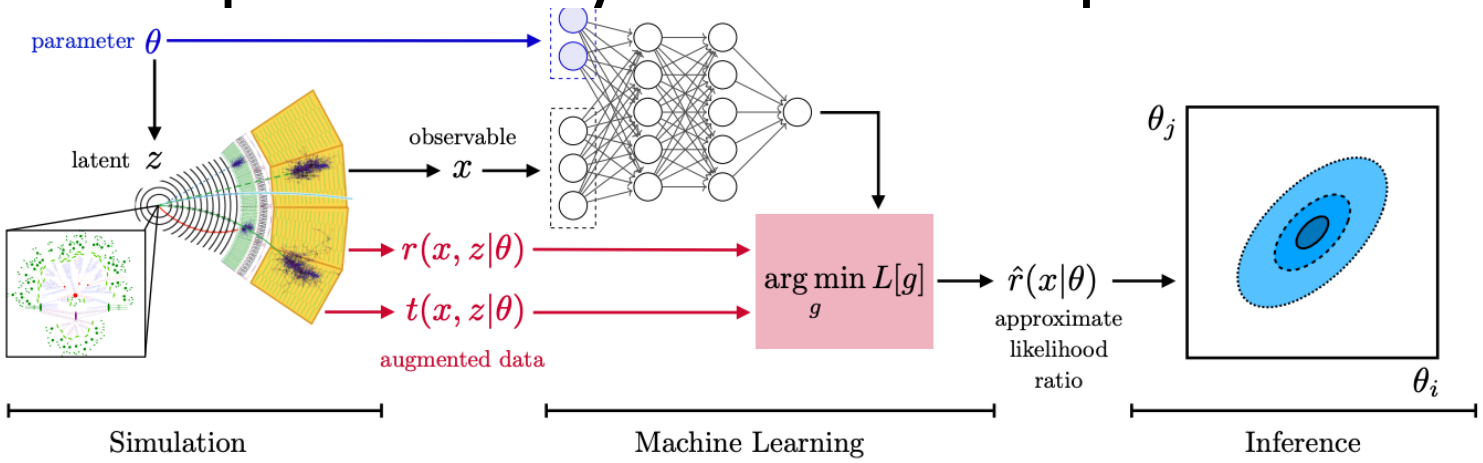


Figure 2 A schematic of machine learning based approaches to likelihood-free inference in which the simulation provides training data for a neural network that is subsequently used as a surrogate for the intractable likelihood during inference. Reproduced from (Brehmer *et al.*, 2018b).

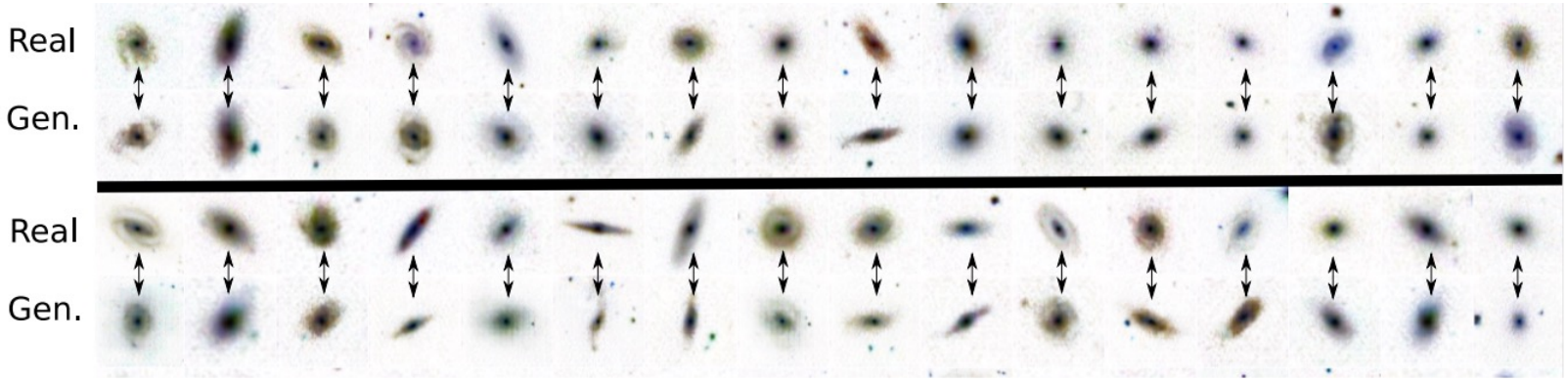
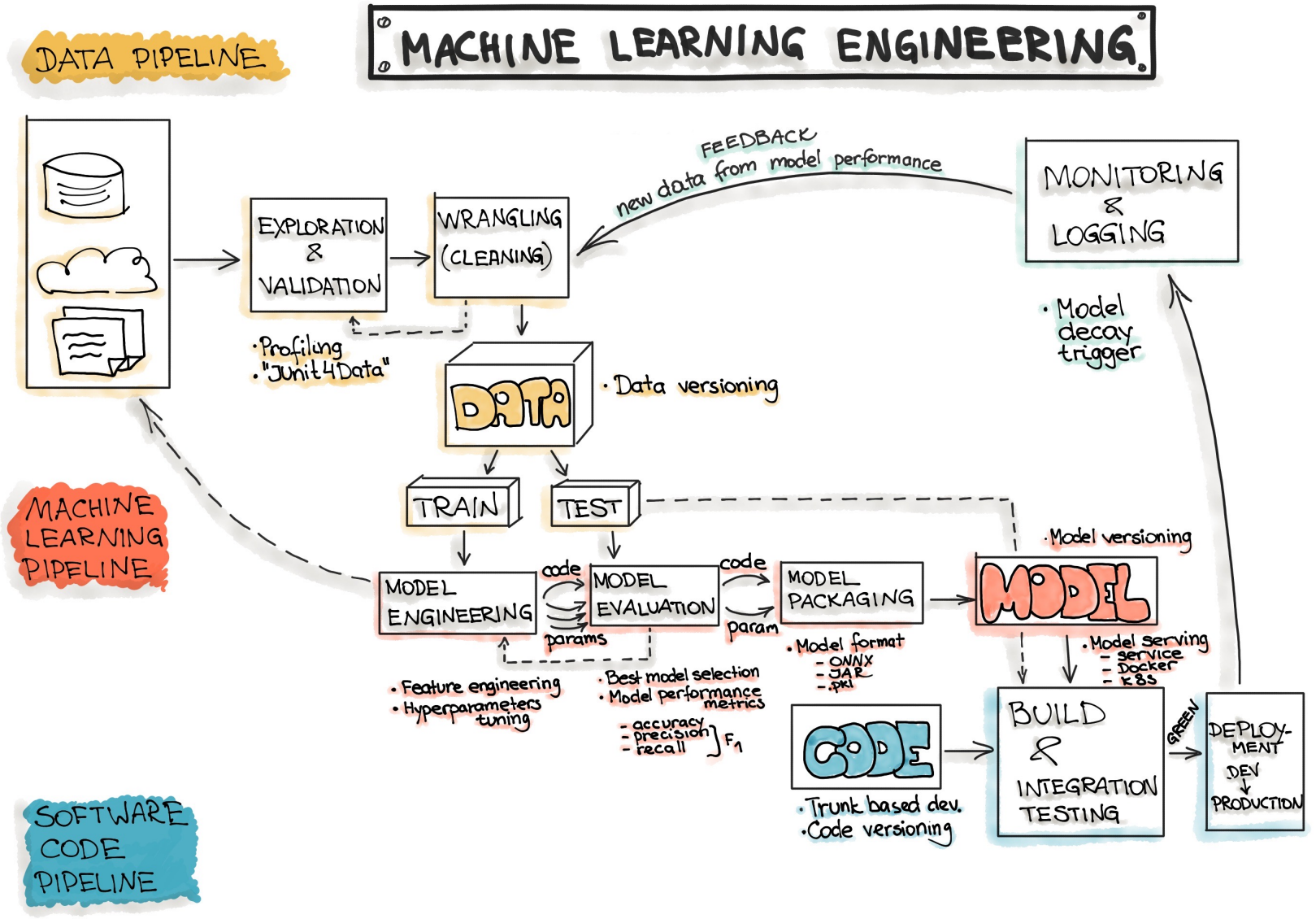


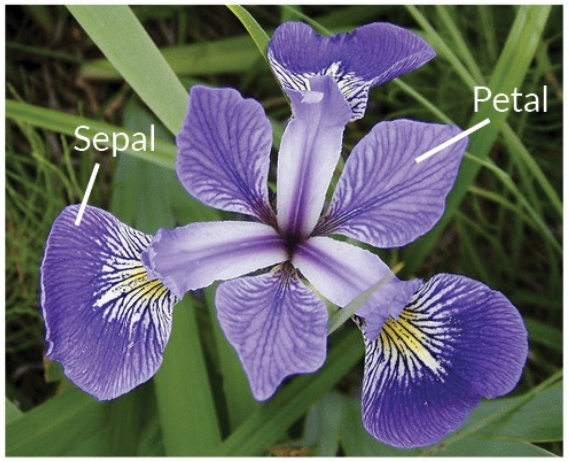
Figure 3 Samples from the GALAXY-ZOO dataset versus generated samples using conditional generative adversarial network. Each synthetic image is a 128×128 colored image (here inverted) produced by conditioning on a set of features $y \in [0, 1]^{37}$. The pair of observed and generated images in each column correspond to the same y value. Reproduced from (Ravanbakhsh *et al.*, 2016).

Figures from Machine Learning and the physical science, Carleo et al

ML workflow



Iris dataset



Iris Versicolor

Iris Setosa

Iris Virginica

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

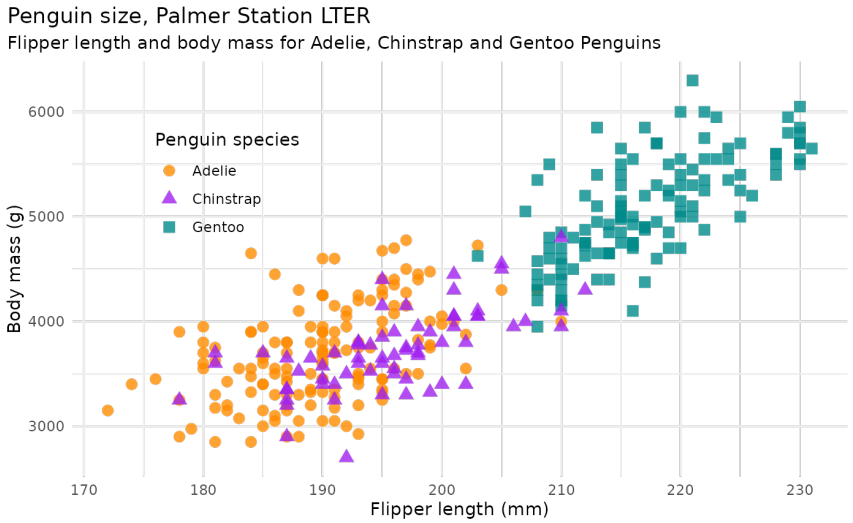
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

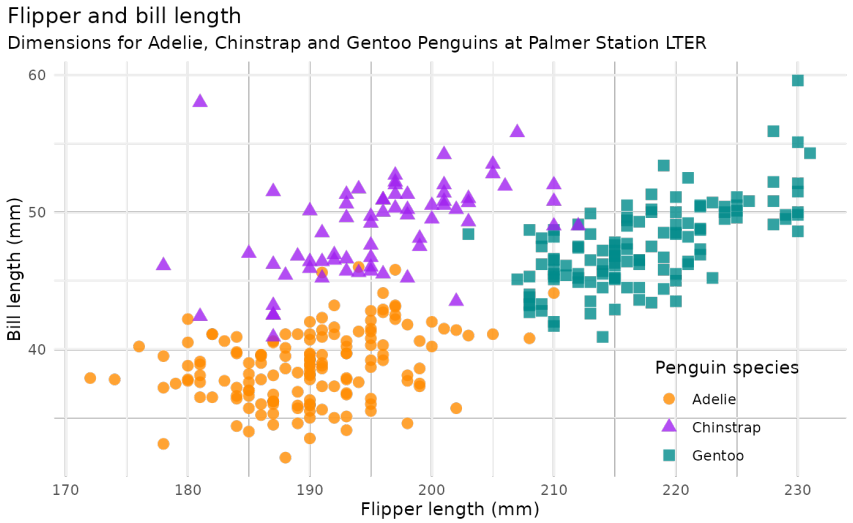
Data Set Information:

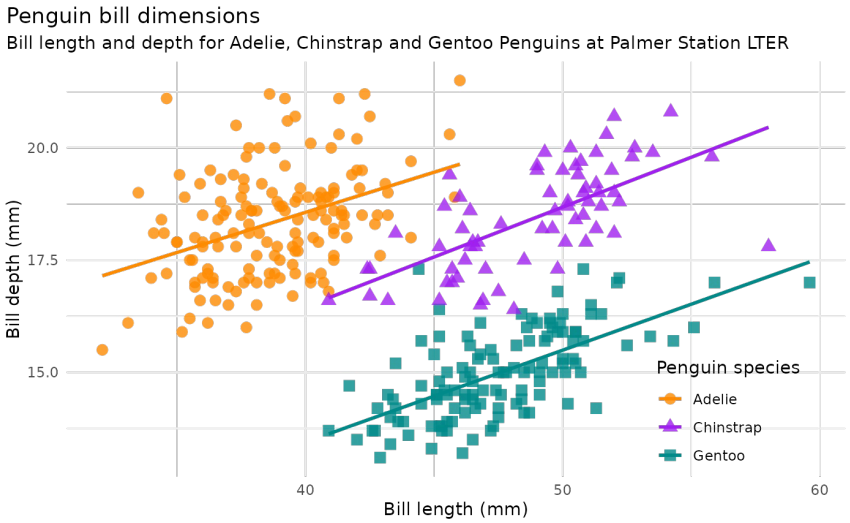
This is perhaps the best known database to be found in the pattern recognition literature.
Fisher's paper is a classic in the field and is referenced frequently to this day.
The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

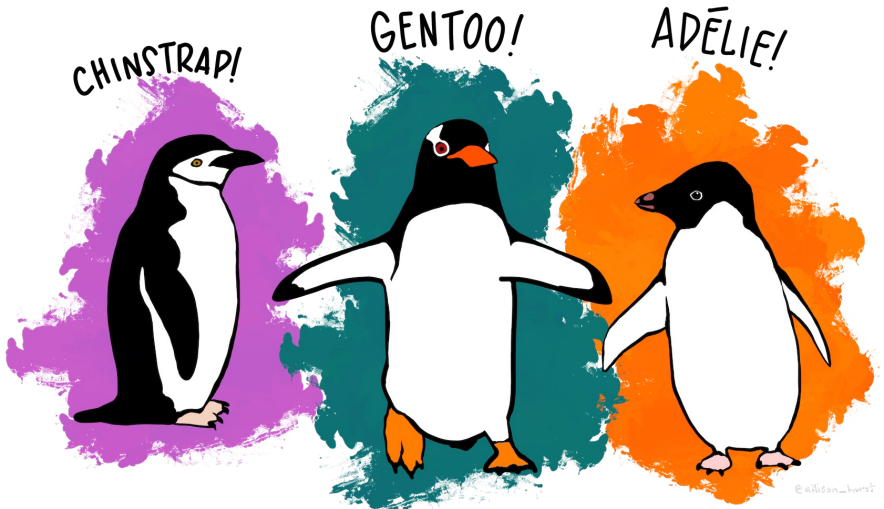
Predicted attribute: class of iris plant.

Palmer Penguin Dataset









We've got data. Find the signal!
Data (pre-processing)

Finding patterns in the data

i.e. the technical characteristics of the data are less important than the pattern (model)

- Input: some data
- Methods: statistical analysis, machine learning, programming
- Output: a robust, reproducible, generalizable model

Raw data

[1] **Accuracy** – how well does the data reflect a real object/ event
[2] **Completeness** – are all mandatory/ necessary features present?
[3] **Consistency** – repeated entries/ consistent labels?
[4] **Invariance in time** – is the data outdated or a mix of old & new?
[5] **Believability** – the presented values are trustworthy
[6] **Interpretability** – the data and conclusions can be interpreted

Parametric/
non-parametric

Can handle continuous/
categorical vars?

Data (pre-processing). Feature engineering Overview of concepts.

Input	Pre-processing	ML model
Raw data		

[1] **Accuracy**
[2] **Completeness**
[3] **Consistency**
[4] **Invariance in time**
[5] **Believability**
[6] **Interpretability**

[a] **Understanding the structure of the data**

type of features: continuous, categorical
ranges of features: [min, max], number of categories
missing information [labels, features]
discriminative power of features (redundancy)

[b] **Adjusting data without tampering with signal**

Expression ranges and One-hot encoding
Standardisation vs scaling
Near zero variance
Multi-collinearity
Dimensionality reduction

[c] **creating robust models – cross-validation. Bias/ variance**

Training/ Validation/ Test splitting
Cross-validation

Parametric/
non-parametric

Can handle continuous/
categorical vars?

Feature selection

Next Lecture …

Applied Data Science
L2. Learning from Data. Cross-validation. Data pre-processing