Adaptive Data Analysis

Machine learning in science and society

Christos Dimitrakakis

August 21, 2019

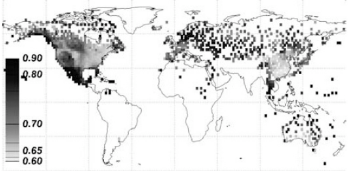
Introduction

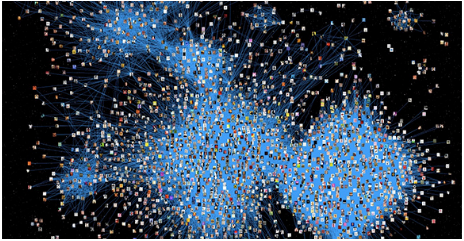
1 Introduction to machine learning
- Data analysis, learning and planning
- Experiment design
- Bayesian inference.
- Course overview

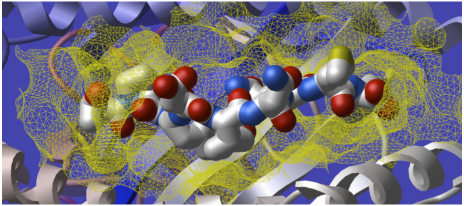
2 Nearest neighbours

3 Reproducibility

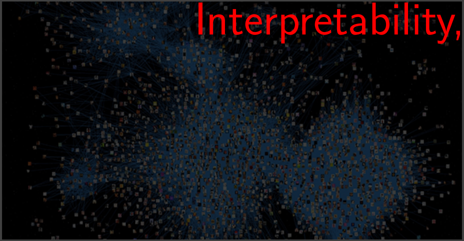
Scientific applications

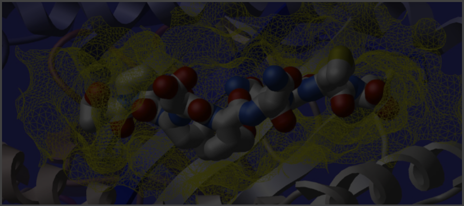






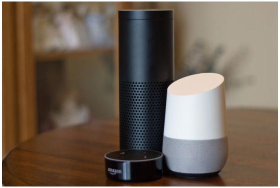
Scientific applications





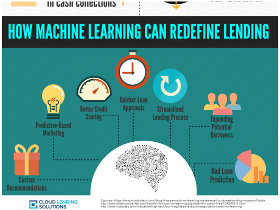
Interpretability, Reproducibility

Pervasive "intelligent" systems



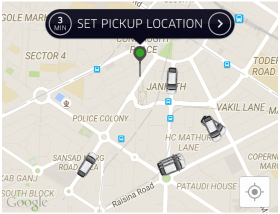
Home assistants

Web advertising



Lending

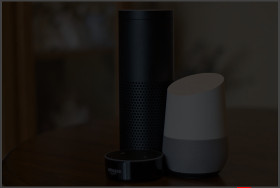
Autonomous vehicles



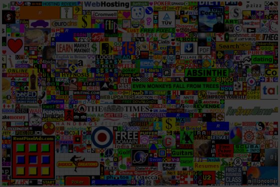
Ridesharing

Public policy

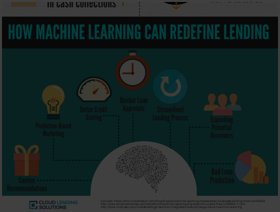
Pervasive "intelligent" systems



Home assistants



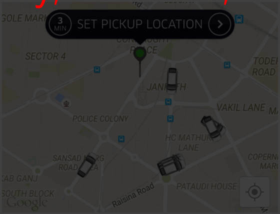
Web advertising



Lending

Privacy, Fairness, Safety

Autonomous vehicles

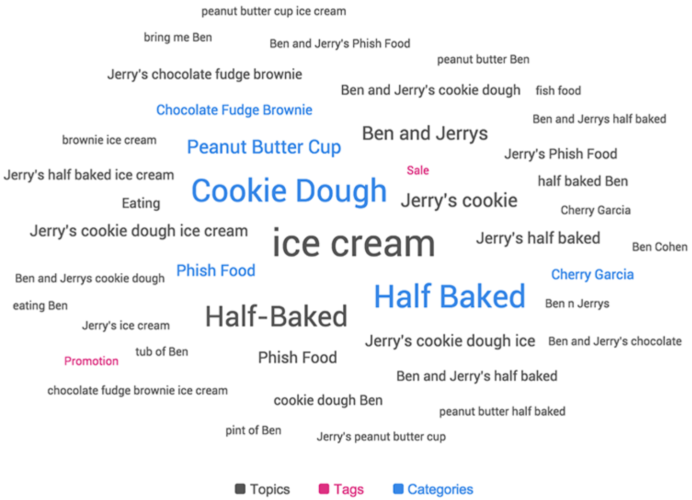


Ridesharing

Public policy

What can machine learning do?

Can machines learn from data?

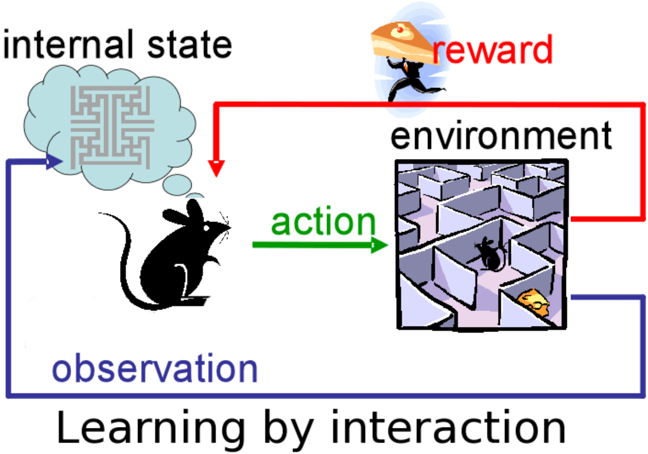


An unsupervised learning problem: topic modelling

Can machines learn from data?

A supervised learning problem: object recognition

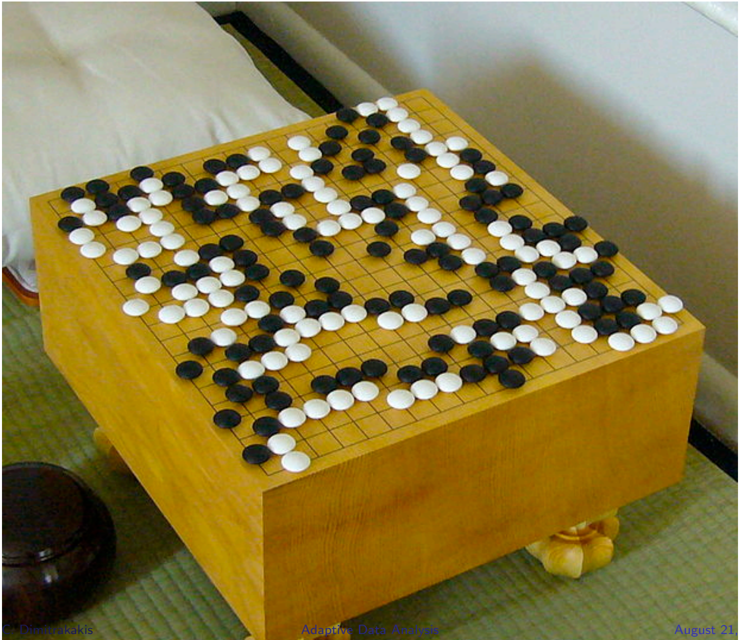
Can machines learn from their mistakes?



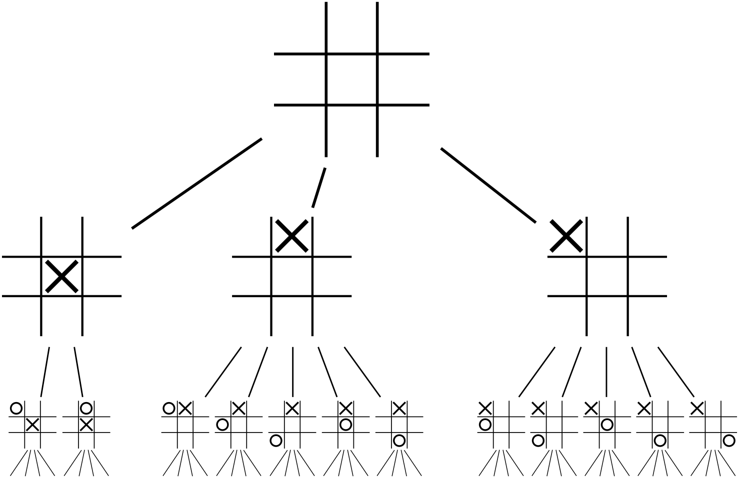
Reinforcement learning

Take actions $a_1, \ldots, a_t$, so as to maximise utility $U = \sum_{t=1}^{T} r_t$

Can machines make complex plans?



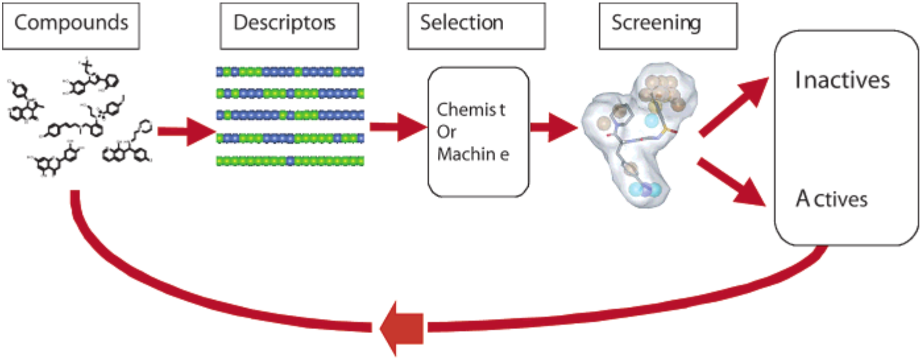
Machines can make complex plans!



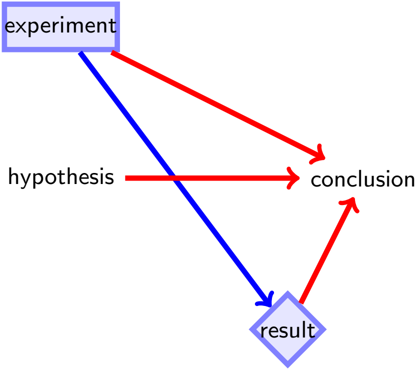
The scientific process as machine learning

Adam, the robot scientist

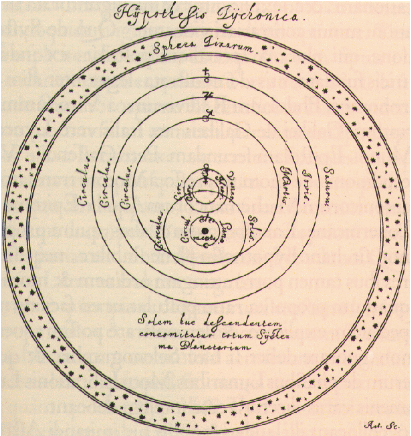
Drug discovery



Drawing conclusions from results



Tycho Brahe's minute eye measurements



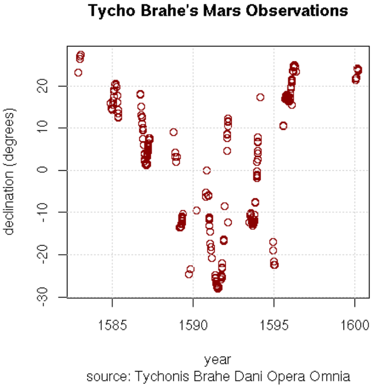
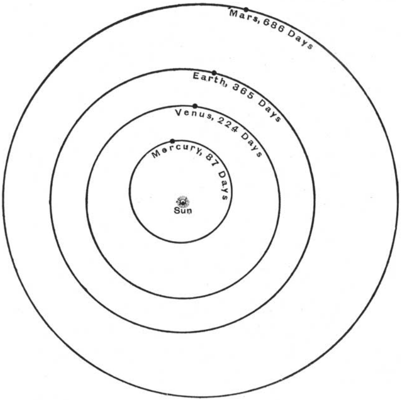
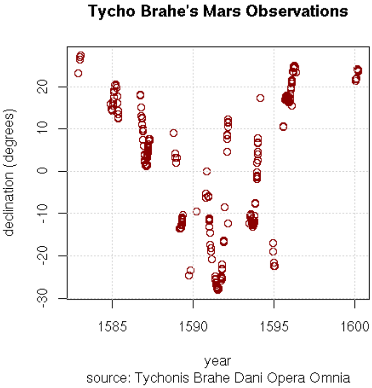


Figure: Tycho's measurements of the orbit of Mars and the conclusion about the actual orbits, under the assumption of an earth-centric universe with circular orbits.

- Hypothesis: Earth-centric, Circular orbits
- Conclusion: Specific circular orbits

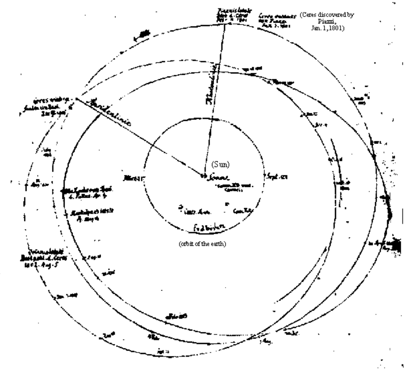
Johannes Kepler's alternative hypothesis



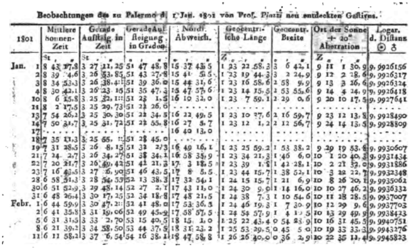


- Hypothesis: Circular or elliptic orbits
- Conclusion: Specific elliptic orbits

200 years later, Gauss formalised this statistically



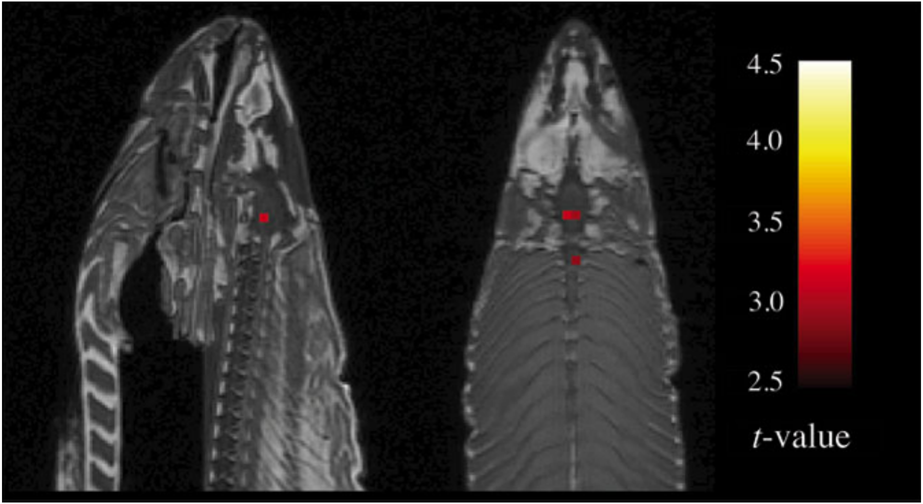
Sketch of the orbits of Ceres and Pallas (nachlaß Gauß, Handb. 4). Courtesy of Universitätsbibliothek Göttingen.



Beobachtungen des zu Palermo d. 1 Jan. 1801 von Prof. Piazzi neu entdeckten Gestirns.

1801		Mittlere Sonnen-Zeit	Gerade Aufsteig. in Zeit	Gerade Aufsteigung in Graden	Nördl. Abweich.	Geocentrische Länge	Geocentr. Breite	Ort der Sonne + 20" Aberration	Logar. d. Distanz ⊙ ♁
		St ' "	St ' "	° ' "	° ' "	Z ° ' "	° ' "	Z ° ' "	
Jan.	1	8 43 27,8	3 27 11, 25	51 47 48, 8	15 37 43, 5	1 23 22 58, 3	3 6 42, 1	9 11 1 30, 9	9, 9926156
	2	8 39 4,6	3 26 53, 85	51 43 27, 8	15 41 5, 5	1 23 19 44, 3	3 2 24, 9	9 12 2 28, 6	9, 9926317
	3	8 34 53,3	3 26 38, 4	51 39 36, 0	15 44 31, 6	1 23 16 58, 6	2 58 9, 9	9 13 3 26, 6	9, 9926324
	4	8 30 42,1	3 26 23, 16	51 35 47, 3	15 47 57, 6	1 23 14 15, 5	2 53 55, 6	9 14 4 24, 9	9, 9926418
	10	8 6 15,8	3 25 32, 1	51 23 1, 5	16 10 32, 0	1 23 7 59, 1	2 29 0, 6	9 20 10 17, 5	9, 9927641
	11	8 2 17,5	3 25 29, 73	51 22 26, 0	....	....	.....	....	.....
	13	7 54 26,2	3 25 30, 30	51 22 34, 5	16 22 49, 5	1 23 10 27, 6	2 16 59, 7	9 23 12 13, 8	9, 9928490
	14	7 50 31,7	3 25 31, 72	51 22 55, 8	16 27 5, 7	1 23 12 1, 2	2 12 56, 7	9 24 14 13, 5	9, 9928809
	17	......	....	....	16 40 13, 0	....	.....	....	.....
	18	7 35 11,3	3 25 55,	51 28 45, 0	....	....	.....	....	.....
	19	7 31 28,5	3 26 8, 15	51 32 2, 3	16 49 16, 1	1 23 25 59, 2	1 53 38, 2	9 29 19 53, 8	9, 9930607
	21	7 24 2,7	3 26 34, 27	51 38 34, 1	16 58 35, 9	1 23 34 21, 3	1 46 6, 0	10 1 20 40, 3	9, 9931434
	22	7 20 21,7	3 26 49, 42	51 42 21, 3	17 3 18, 5	1 23 39 1, 8	1 42 28, 1	10 2 21 32, 0	9, 9931886
	23	7 16 43,5	3 27 6, 90	51 46 43, 5	17 8 5, 5	1 23 44 15, 7	1 38 52, 1	10 3 22 22, 7	9, 9932348
	28	6 58 51,3	3 28 54, 55	52 13 38, 3	17 32 54, 1	1 24 15 15, 7	1 21 6, 9	10 8 26 20, 1	9, 9935062
	30	6 51 52,9	3 29 48, 14	52 27 2, 1	17 43 11, 0	1 24 30 9, 0	1 14 16, 0	10 10 27 46, 2	9, 9936332
	31	6 48 26,4	3 30 17, 25	52 34 18, 8	17 48 21, 5	1 24 38 7, 3	1 10 54, 6	10 11 28 28, 5	9, 9937007
Febr.	1	6 44 59,9	3 30 47, 2	52 41 48, 0	17 53 36, 3	1 24 46 19, 3	1 7 30, 9	10 12 29 9, 6	9, 9937703
	2	6 41 35,8	3 31 19, 06	52 49 45, 9	17 58 57, 5	1 24 54 57, 9	1 4 1, 5	10 13 29 49, 9	9, 9938423
	5	6 31 31,5	3 33 2, 70	53 15 40, 5	18 15 1, 0	1 25 22 43, 4	0 54 23, 9	10 16 31 45, 5	9, 9940751
	8	6 21 39,2	3 34 58, 50	53 44 37, 5	18 31 23, 2	1 25 53 29, 5	0 45 5, 0	10 19 33 33, 3	9, 9943276
	11	6 11 58,2	3 37 6, 54	54 16 38, 1	18 47 58, 8	1 26 26 40, 0	0 36 2, 9	10 22 35 11, 4	9, 9945823

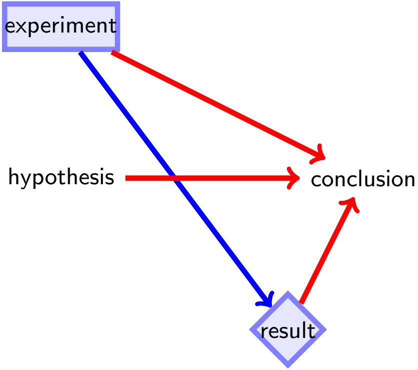
A warning: The dead salmon mirage



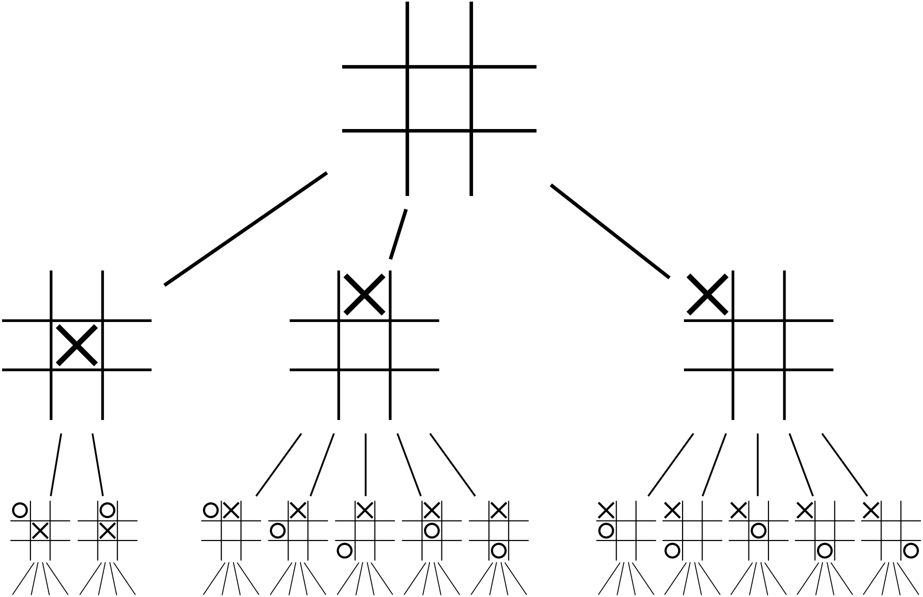
A simple simulation study

`src/reproducibility/mri_analysis.ipynb`

Planning future experiments



Planning experiments is like Tic-Tac-Toe



Eve, another robot scientist

Di

a malaria drug

Machine learning in practice

Avoiding pitfalls

- Choosing hypotheses.
- Correctly interpreting conclusions.
- Using a good testing methodology.

Machine learning in society

- Privacy
- Fairness
- Safety

Machine learning in practice

Avoiding pitfalls

- Choosing hypotheses.
- Correctly interpreting conclusions.
- Using a good testing methodology.

Machine learning in society

- Privacy — Credit risk.
- Fairness
- Safety

Machine learning in practice

Avoiding pitfalls

- Choosing hypotheses.
- Correctly interpreting conclusions.
- Using a good testing methodology.

Machine learning in society

- Privacy — Credit risk.
- Fairness — Job market.
- Safety

Machine learning in practice

Avoiding pitfalls

- Choosing hypotheses.
- Correctly interpreting conclusions.
- Using a good testing methodology.

Machine learning in society

- Privacy — Credit risk.
- Fairness — Job market.
- Safety — Medicine.

Course structure

Module structure

- Activity-based, hands-on.
- Mini-lectures with short exercises in each class.
- Technical tutorials and labs in alternate week.

Modules

Three mini-projects.

- Simple decision problems: Credit risk.
- Sequential problems: Medical diagnostics and treatment.

Technical topics

Machine learning problems

- Unsupervised learning. .
- Supervised learning.
- Reinforcement learning.

Algorithms and models

- Bayesian inference and graphical models.
- Stochastic optimisation and neural networks.
- Backwards induction and Markov decision processes.

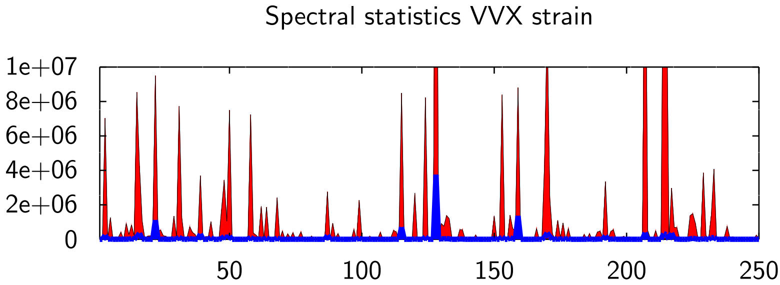
Further reading

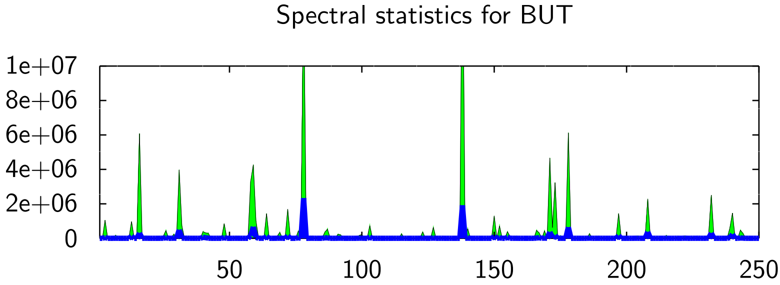
- Bennett et al. [2] describe how the usual uncorrected analysis of fMRI data leads to the conclusion that the dead salmon can reason about human images.
- Bennett et al. [1] discuss how to perform analyses of medical images in a principled way. They also introduce the use of simulations in order to test how well a particular method is going to perform.

Resources

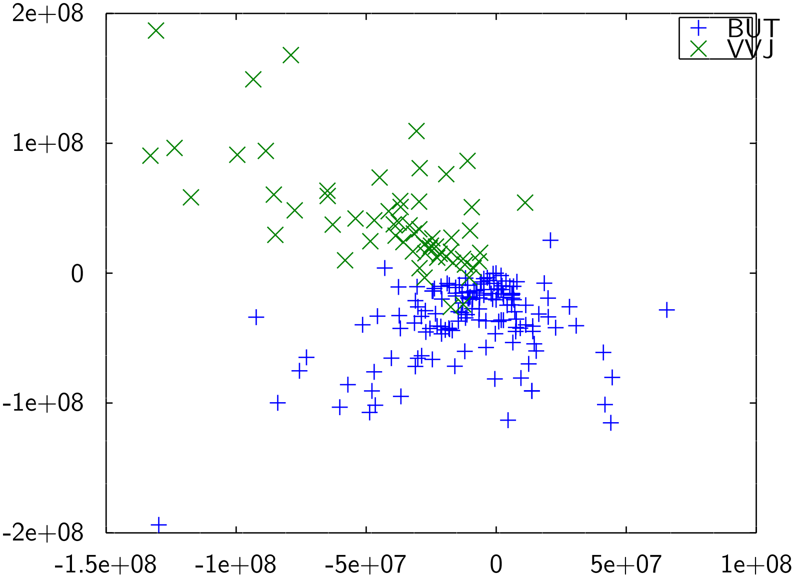
- Online QA platform: `https://piazza.com/class/jufgabrw4d57nh`
- Course code and notes: `https://github.com/olethrosdc/ml-society-science`
- Book `https://github.com/olethrosdc/ml-society-science/notes.pdf`

Discriminating between diseases

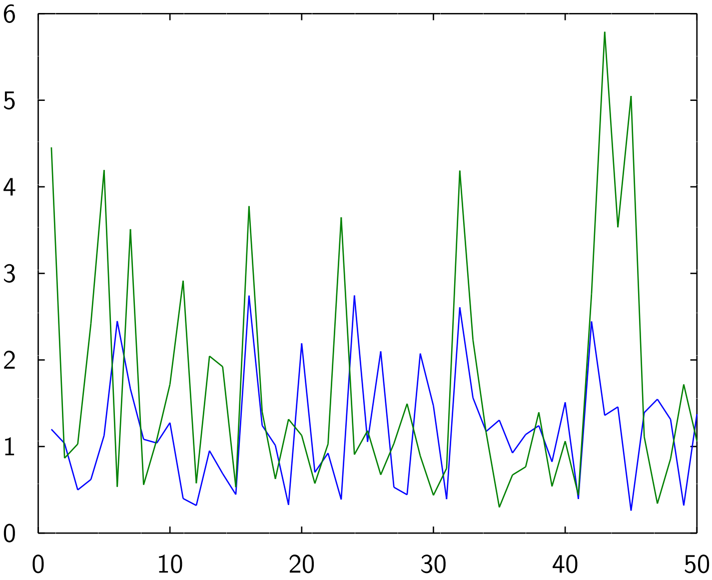




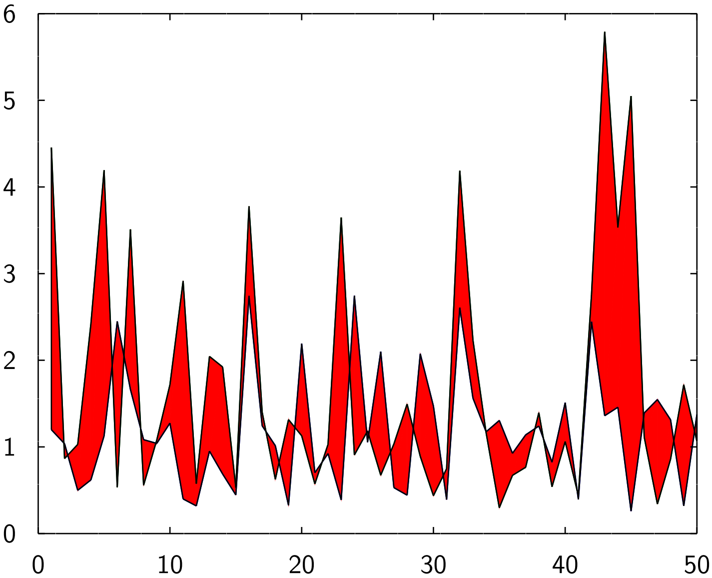
Nearest neighbour: the hidden secret of machine learning



Comparing spectral data



Comparing spectral data



The nearest neighbour algorithm

Algorithm 1 K-NN Classify

Input Data $D = \{(x_1, y_1), \ldots, (x_T, y_T)\}$, $k \geq 1$, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, new point $x \in \mathcal{X}$
$D = \texttt{Sort}(D, d)$ % Sort D so that $d(x, x_i) \leq d(x, x_{i+1})$.
$p_y = \sum_{i=1}^{k} \mathbb{I}\{y_i = y\} / k$ for $y \in \mathcal{Y}$.
Return $\boldsymbol{p} \triangleq (p_1, \ldots, p_k)$

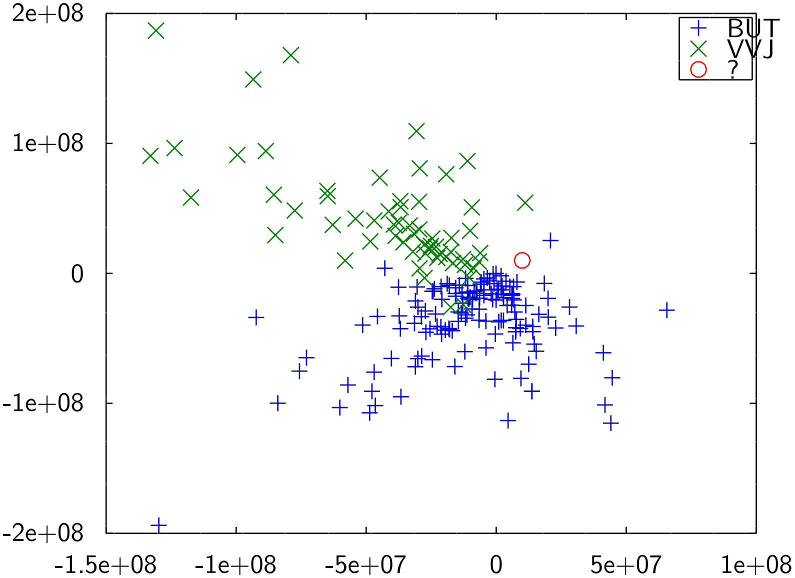
Algorithm parameters

- Neighbourhood $k \geq 1$.
- Distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

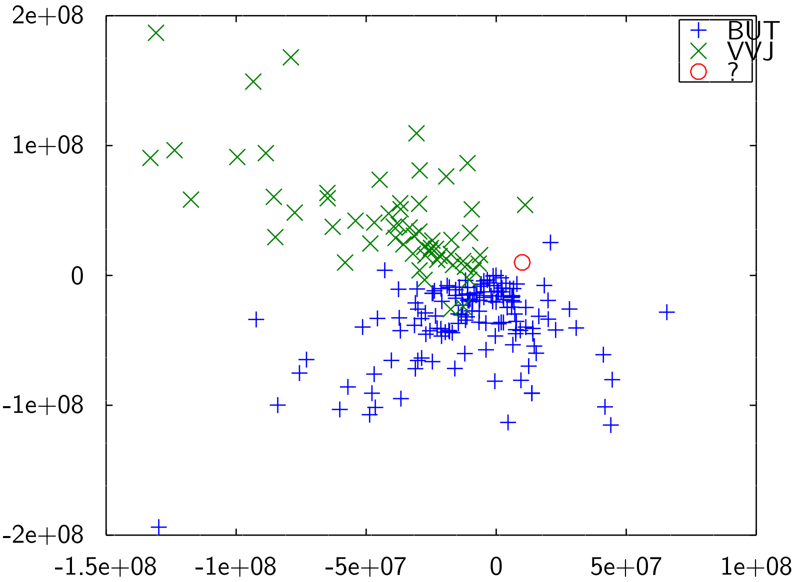
What does the algorithm output when $k = T$?

Figure: The nearest neighbours algorithm was introduced by Fix and Hodges Jr[3], who also proved consistency properties.

Nearest neighbour: What type is the new bacterium?



Nearest neighbour: What type is the new bacterium?



What if it a **completely different strain**?

Separating the model from the classification policy

- The $\textsc{k-NN}$ algorithm returns a model giving class probabilities for new data points.
- Deciding a class given the model

$$\pi(a \mid x) = \mathbb{I}\left\{p_a \geq p_y \forall y\right\}, \qquad \boldsymbol{p} = \textsc{k-NN}(D, k, d, x)$$

Hands on with Python console

- `src/decision-problems/knn-classify.py`
- `src/decision-problems/KNN.ipynb`

Discussion: Shortcomings of k-nearest neighbour

- Choice of k
- Choice of metric d.
- Representation of uncertainty.
- Scaling with large amounts of data.
- Meaning of label probabilities.

Learning outcomes

Understanding

- How kNN works
- The effect of hyperameters k, d for nearest neighbour.
- The use of kNN to classify new data.

Skills

- Use a standard kNN class in python
- Optimise kNN hyperparameters in an unbiased manner.
- Calculate probabilities of class labels using kNN.

Reflection

- When is kNN a good model?
- How can we deal with large amounts of data?
- How can we best represent uncertainty?

Computational reproducibility: Can the study be repeated?

Can we, from the available information and data, exactly reproduce the reported methods and results?

- `jupyter` notebooks
- `svn`, `git` or `mercurial` version control systems

Scientific reproducibility: Is the conclusion correct?

Can we, from the available information and a new set of data, reproduce the conclusions of the original study?

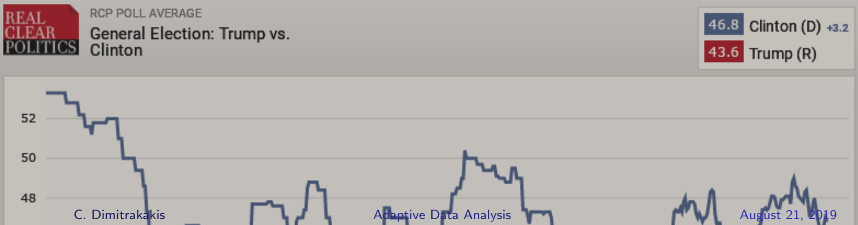
When publishing results about a new method, computational reproducibility is essential for scientific reproducibility.

RealClear Politics

Polls Election 2018 Video Writers More

Poll	Date	Sample	MoE	Clinton (D)	Trump (R)	Spread
Final Results	--	--	--	48.2	46.1	Clinton +2.1
RCP Average	11/1 - 11/7	--	--	46.8	43.6	Clinton +3.2
Bloomberg	11/4 - 11/6	799 LV	3.5	46	43	Clinton +3
IBD/TIPP Tracking	11/4 - 11/7	1107 LV	3.1	43	42	Clinton +1
Economist/YouGov	11/4 - 11/7	3669 LV	–	49	45	Clinton +4
LA Times/USC Tracking	11/1 - 11/7	2935 LV	4.5	44	47	Trump +3
ABC/Wash Post Tracking	11/3 - 11/6	2220 LV	2.5	49	46	Clinton +3
FOX News	11/3 - 11/6	1295 LV	2.5	48	44	Clinton +4
Monmouth	11/3 - 11/6	748 LV	3.6	50	44	Clinton +6
NBC News/Wall St. Jrnl	11/3 - 11/5	1282 LV	2.7	48	43	Clinton +5
CBS News	11/2 - 11/6	1426 LV	3.0	47	43	Clinton +4
Reuters/Ipsos	11/2 - 11/6	2196 LV	2.3	44	39	Clinton +5

All General Election: Trump vs. Clinton Polling Data



The principle of independent evaluation

Data used for estimation cannot be used for evaluation.

Figure: The decision process in classification.

Figure: The decision process in classification.

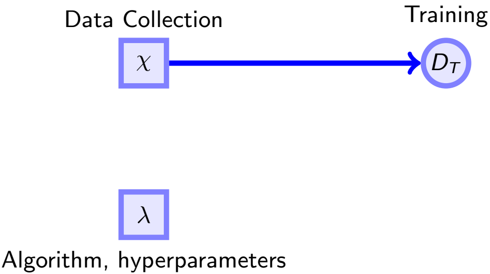


Figure: The decision process in classification.

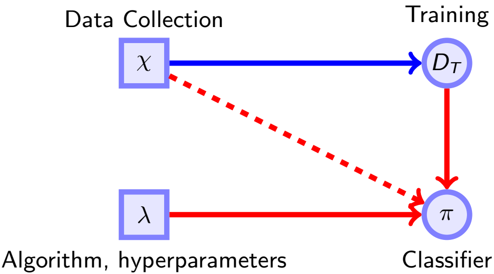


Figure: The decision process in classification.

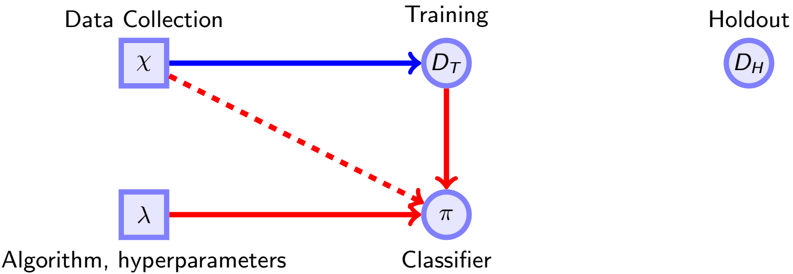


Figure: The decision process in classification.

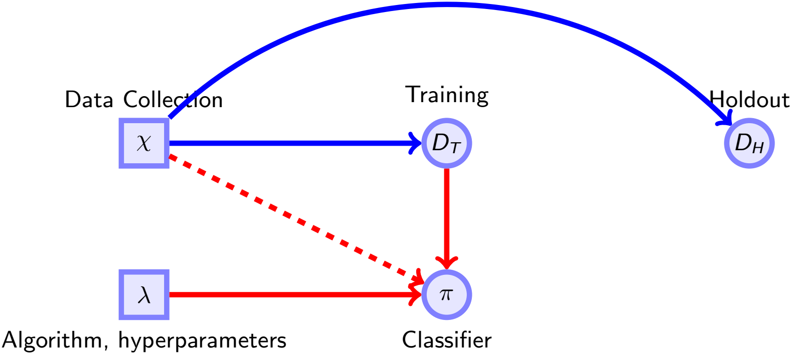


Figure: The decision process in classification.

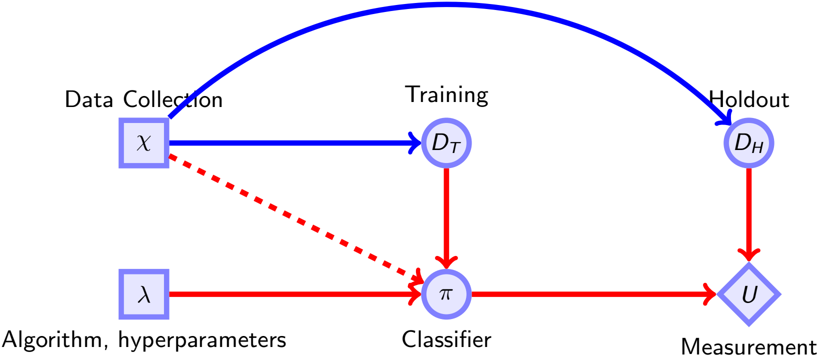


Figure: The decision process in classification.

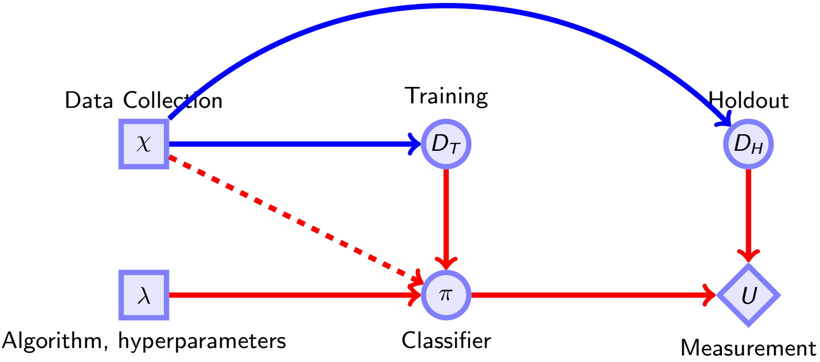


Figure: The decision process in classification.

Classification accuracy

$$\mathbb{E}_\chi[U(\pi)] = \sum_{x,y} \underbrace{\mathbb{P}_\chi(x, y)}_{\text{Data probability}} \overbrace{\pi(a = y \mid x)}^{\text{Decision probability}}$$

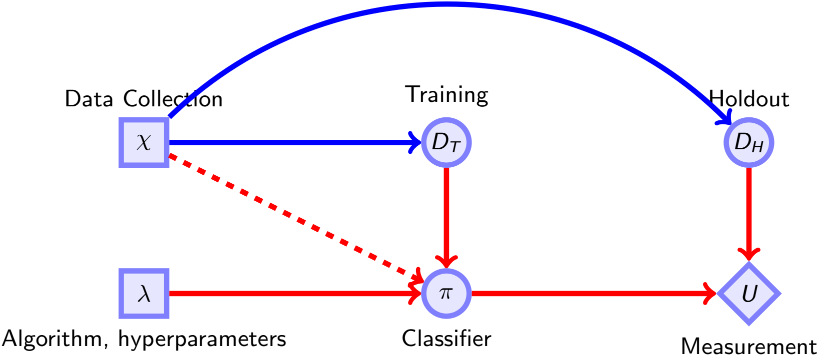


Figure: The decision process in classification.

Classification accuracy

$$\mathbb{E}_{D_H} U(\pi) = \sum_{(x,y)\in D_H} \pi(a = y \mid x)/|D_H|.$$

The human as an algorithm.

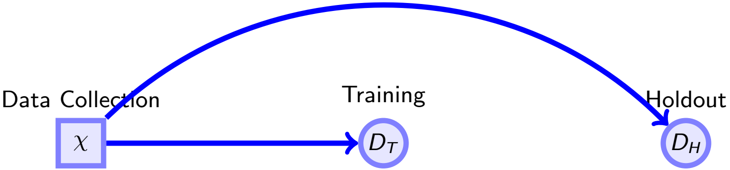


Figure: Selecting algorithms and hyperparameters through holdouts

The human as an algorithm.

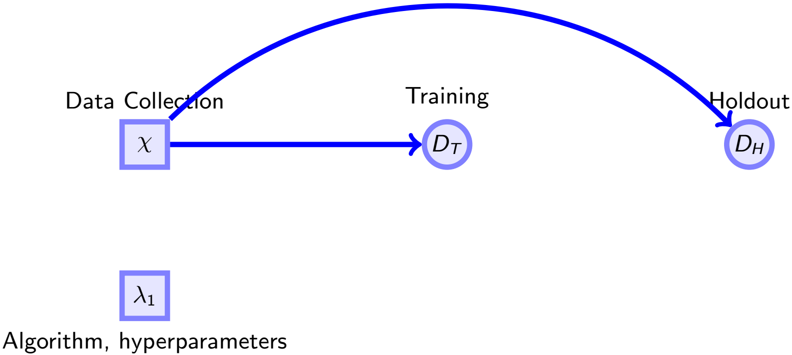


Figure: Selecting algorithms and hyperparameters through holdouts

The human as an algorithm.

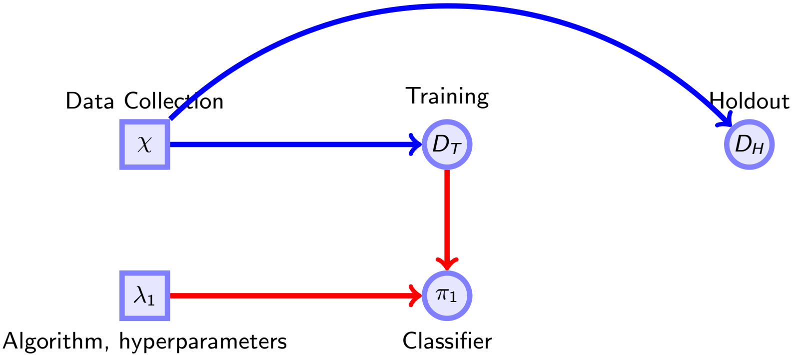


Figure: Selecting algorithms and hyperparameters through holdouts

The human as an algorithm.

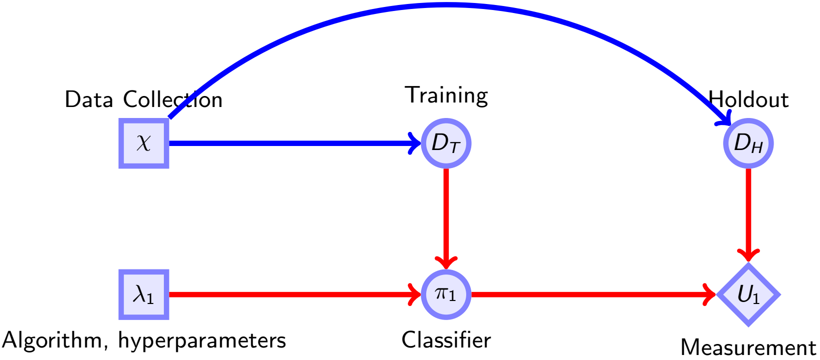


Figure: Selecting algorithms and hyperparameters through holdouts

The human as an algorithm.

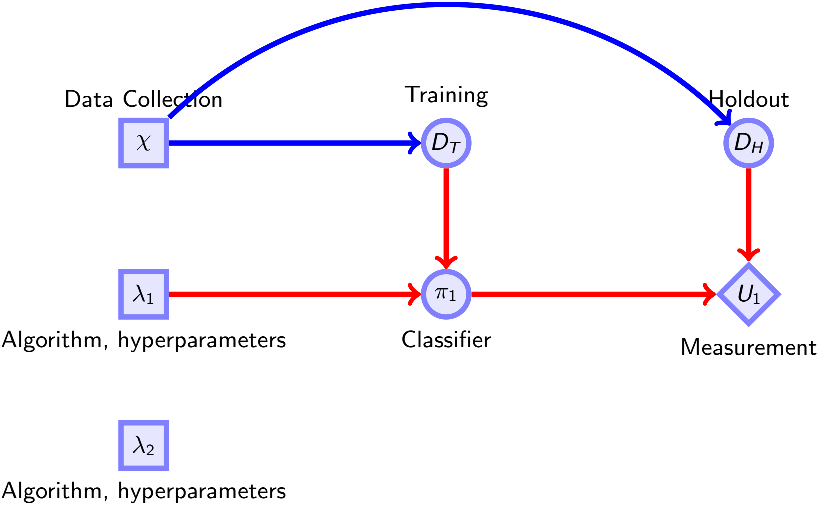


Figure: Selecting algorithms and hyperparameters through holdouts

The human as an algorithm.

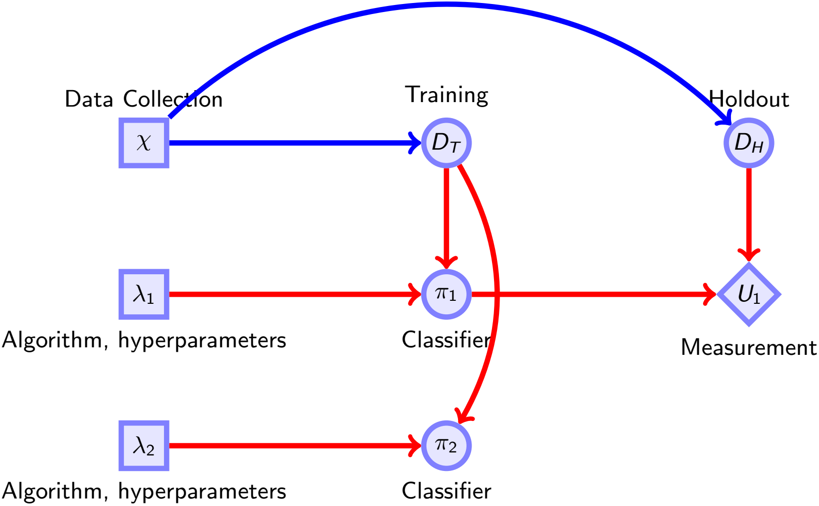


Figure: Selecting algorithms and hyperparameters through holdouts

The human as an algorithm.

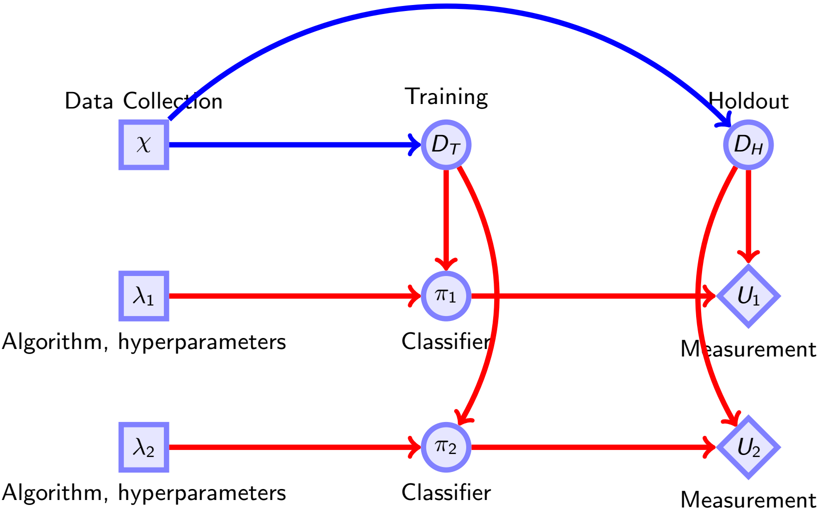


Figure: Selecting algorithms and hyperparameters through holdouts

Holdout sets

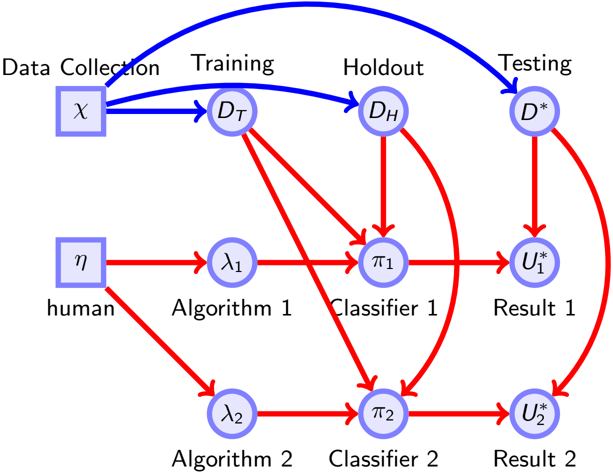
- Original data D, e.g. $D = (x_1, \ldots, x_T)$.
- Training data $D_T \subset D$, e.g. $D_T = x_1, \ldots, x_n$, $n < T$.
- Holdout data $D_H = D \setminus D_T$, used to measure the quality of the result.
- Algorithm λ with hyperparametrs ϕ.
- Get algorithm output $\pi = \lambda(D_T, \phi)$.
- Calculate quality of output $U(\pi, D_H)$

Holdout and test sets for unbiased algorithm comparison

Algorithm 2 Unbiased adaptive evaluation through data partitioning

Partition data into D_T, D_H, D^*.
for $\lambda \in \Lambda$ **do**
  for $\phi \in \Phi_\lambda$ **do**
    $\pi_{\phi,\lambda} = \lambda(D_T, \phi)$.
  end for
  Get π^*_λ maximising $U(\pi_{\phi,\lambda}, D_H)$.
  $u_\lambda = U(\pi^*_\lambda, D^*)$.
end for
$\lambda^* = \arg\max_\lambda u_\lambda$.

Final performance measurement



Independent data sets

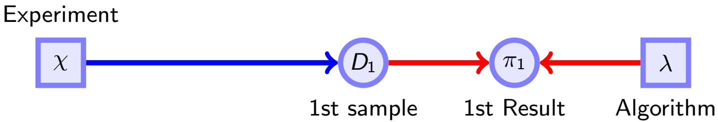


Figure: Multiple samples

Independent data sets

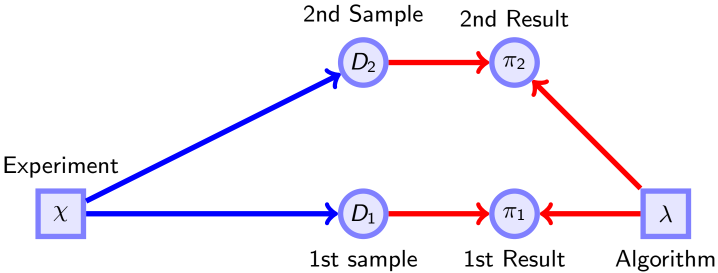


Figure: Multiple samples

Bootstrap samples

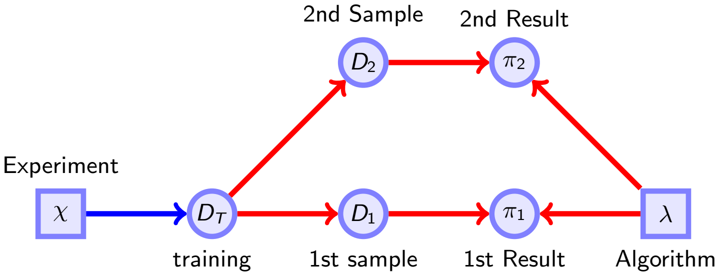


Figure: Bootstrap replicates of a single sample

Bootstrapping

Bootstrapping is a general technique that can be used to:

- Estimate the sensitivity of λ to the data x.
- Obtain a distribution of estimates π from λ and the data x.
- When estimating the performance of an algorithm on a small dataset D^*, use bootstrap samples of D^*.

Bootstrapping

1. **Input** Training data D, number of samples k.
2. **For** $i = 1, \ldots, k$
3. $\quad D^{(i)} = \mathrm{Bootstrap}(D)$
4. **return** $\left\{ D^{(i)} \;\middle|\; i = 1, \ldots, k \right\}$.

where $\mathrm{Bootstrap}(D)$ samples with replacement $|D|$ points from D_T.

Cross-validation

k-fold Cross-Validation

1. **Input** Training data D_T, number of folds k, algorithm λ, measurement function U
2. Create the partition $D^{(1)} \dots, D^{(k)}$ so that $\bigcup_{i=1}^{k} D^{(k)} = D$.
3. Define $D_T^{(i)} = D \setminus D^{(i)}$
4. $\quad \pi_i = \lambda(D_T^{(i)})$
5. **For** $i = 1, \dots, k$:
6. $\quad \pi_i = \lambda(D^{(i)})$
7. $\quad u_i = U(\pi_i)$
8. **return** $\{y_1, \dots, y_i\}$.

Simulation

Steps for a simulation pre-study

1. Define a data-generating process as close to the original dataset as possible.
2. Collect data according to your protocol.
3. Run the intended analysis.
4. See if the results are reasonable, or if you need more power.

Simulation

Simulation study

1. Create a simulation that allows you to collect data similar to the real one.
2. Collect data from the simulation and analyse it according to your protocol.
3. If the results are not as expected, alter the protocol or the simulation. In which cases do you get good results?
4. Finally, use the best-performing method as the protocol.

Independent replication

Replication study

1. Reinterpret the original hypothesis and experiment.
2. Collect data according to the original protocol, unless flawed.
3. Run the analysis again, unless flawed.
4. See if the conclusions are in agreement.

Learning outcomes

Understanding

- What is a hold-out set, cross-validation and bootstrapping.
- The idea of not reusing data input to an algorithm to evaluate it.
- The fact that algorithms can be implemented by both humans and machines.

Skills

- Use git and notebooks to document your work.
- Use hold-out sets or cross-validation to compare parameters/algorithms in Python.
- Use bootstrapping to get estimates of uncertainty in Python.

Reflection

- What is a good use case for cross-validation over hold-out sets?
- When is it a good idea to use bootstrapping?
- How can we use the above techniques to avoid the false discovery problem?
- Can these techniques fully replace independent replication?

[1] Craig M. Bennett, George L. Wolford, and Michael B. Miller. The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience*, 4 4: 417–22, 2009. URL https://pdfs.semanticscholar.org/19c3/d8b67564d0e287a43b1e7e0f496eb1e8a945.pdf.

[2] Craig M Bennett, Abigail A Baird, Michael B Miller, and George L Wolford. Journal of serendipitous and unexpected results. *Journal of Serendipitous and Unexpected Results (jsur. org)-Vol*, 1(1):1–5, 2012. URL https://teenspecies.github.io/pdfs/NeuralCorrelates.pdf.

[3] Evelyn Fix and Joseph L Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California Univ Berkeley, 1951.