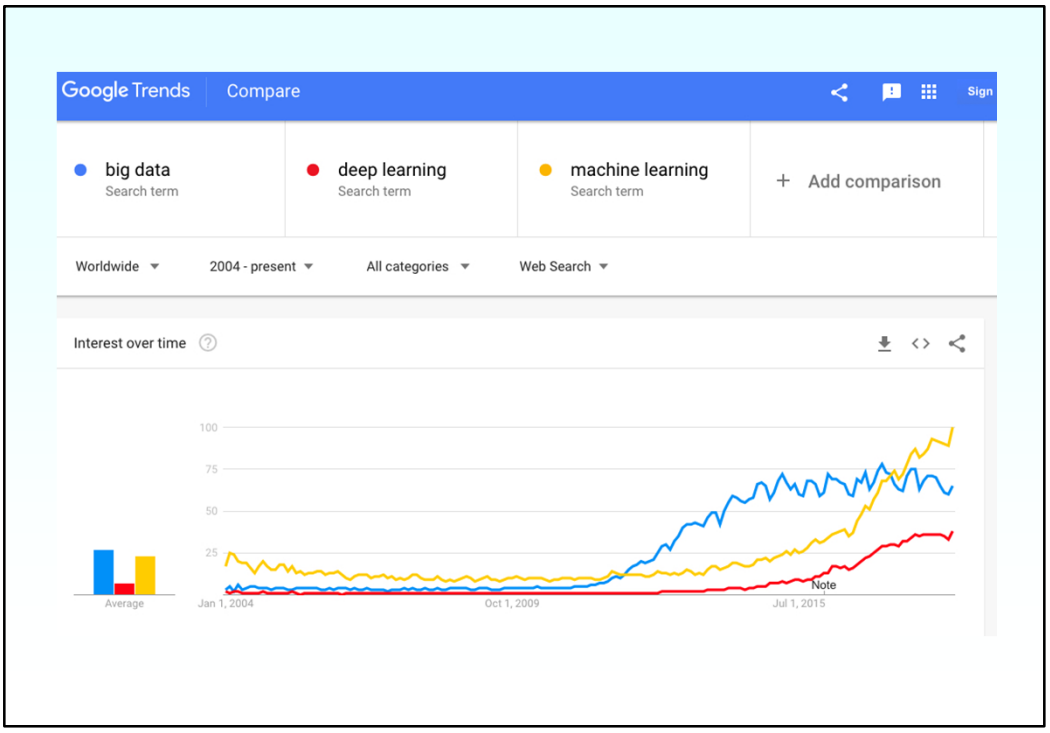
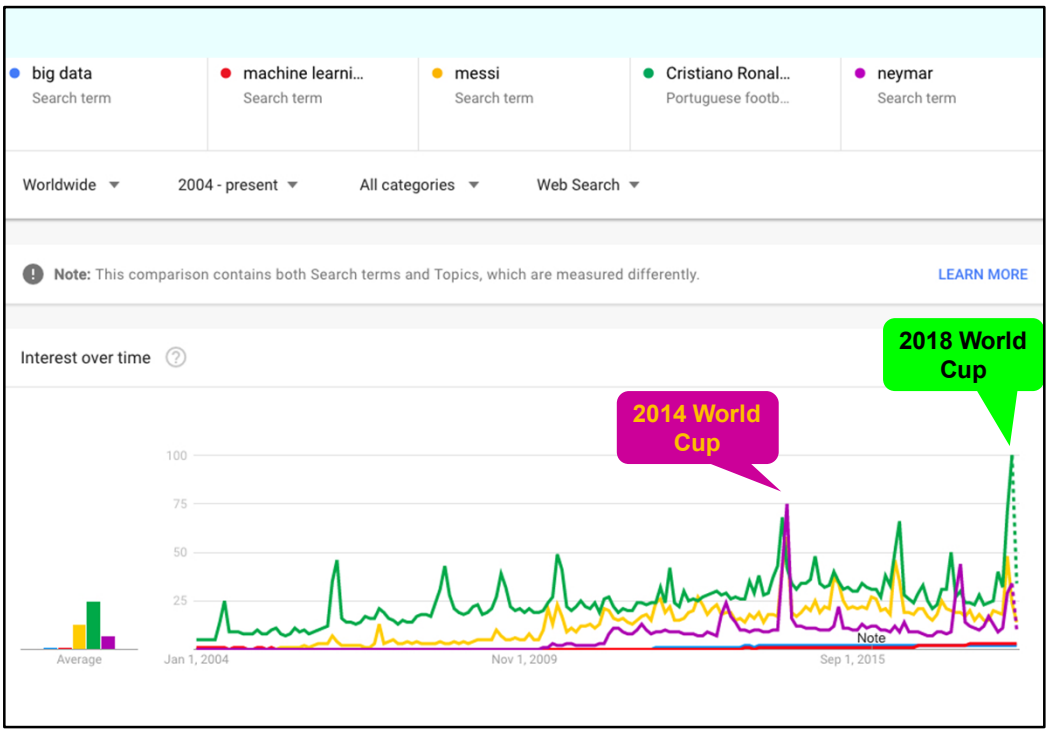

Complex Trait Genomic Analysis
Machine
Learning

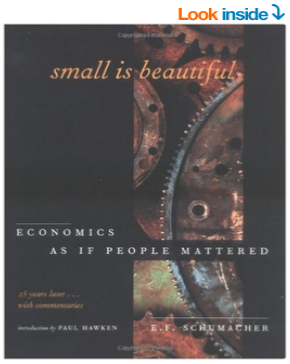

Big Data and Machine Learning
BIG
IS BEAUTIFUL


Google Trends
Compare
Sign
big data
Search term
deep learning
Search term
machine learning
Search term
Add comparison
Worldwide
2004 - present
All categories
Web Search
Interest over time
100
75
50
25
Note
Average
Jan 1, 2004
Oct 1, 2009
Jul 1, 2015


big data
Search term
machine learni...
Search term
messi
Search term
Cristiano Ronal...
Portuguese footb...
neymar
Search term
Worldwide
2004 - present
All categories
Web Search
Note: This comparison contains both Search terms and Topics, which are measured differently.
LEARN MORE
Interest over time
2018 World Cup
2014 World Cup
100
75
50
25
Note
Average
Jan 1, 2004
Nov 1, 2009
Sep 1, 2015

Big Data and Machine Learning

1. Big is Beautiful
2. Big Datasets in Genomics
3. Machine Learning vs. Statistics
4. Ensemble Methods
5. Deep Learning



Small Is Beautiful, 25th Anniversary Edition: Economics As If People Mattered: 25 Years Later . . . With Commentaries Paperback – June 15, 2000

by E. F. Schumacher (Author)

12 customer reviews

See all 15 formats and editions

Hardcover	**Paperback**	Mass Market Paperback
from $63.71	**from $3.81**	from $63.71
1 New from $63.71	23 Used from $3.81 10 New from $46.01	1 New from $63.71

Small is Beautiful is the perfect antidote to the economics of globalization. As relevant today as when it was first published, this is a landmark set of essays on humanistic economics. This 25th anniversary edition brings Schumacher's ideas into focus for the end-of-the-century by adding commentaries by contemporary thinkers who have been influenced by Schumacher. They analyze the impact of his philosophy on current political and economic thought. Small is Beautiful is the classic of common-sense economics upon which many recent trends in our society are founded. This is economics from the heart rather than from just the bottom line.

Best Books of 2015 So Far

Best Books of the Year So Far

Looking for something great to read? Browse our editors' picks for 2015's Best Books of the Year So Far in fiction, nonfiction, mysteries, children's books, and much more.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

SPECIALS

See all specials

2008

EDITORIAL

Community cleverness required

Researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching.

(3 September 2008)

PLOS | BIOLOGY

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

1 Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **2** Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3** Carl R. Woese Institute for Genomic Biology & Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **4** School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **5** Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **6** Carl R. Woese Institute for Genomic Biology, Department of Entomology, and Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

* mschatz@cshl.edu (MCS); sinhas@illinois.edu (SS); generobi@illinois.edu (GER)

Abstract

Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.

OPEN ACCESS

Citation: Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

Published: July 7, 2015

Table 1. Four domains of Big Data in 2025.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLOS Biology 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

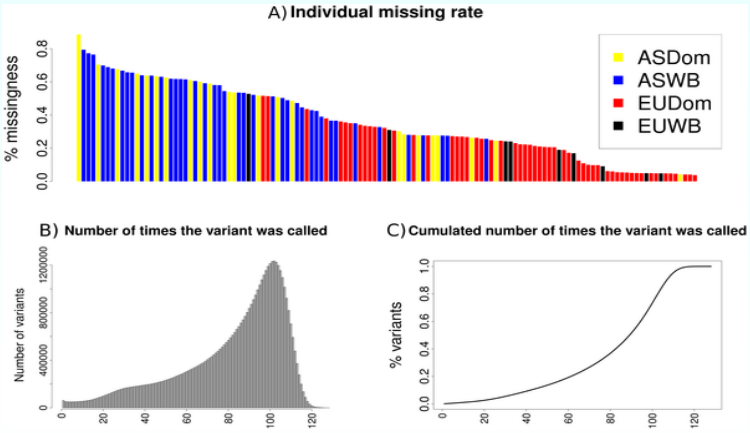
WHAT IS BIG?

- Animal breeding has always been 'big'.
- Is not because big in size only that are we going to improve.
- Big means not only large but mainly heterogeneous, maybe experiment-free collected.
- Data are non normal (e.g., expression levels, epigenetics), undefined statistical properties (e.g., reactome, ontology), unknown reliability (AnimalQTLdb, text mining).

HOW COME BIG?

- Obviously, because of internet.
- Today, social networks are the main provider of data, i.e., soft rather than hard sciences.
- In general data are collected experiment-free.
- Bias is potentially the most dangerous feature (eg, online surveys)
- Open access policy has been extremely beneficial for research.

Big data always come with (big) holes



128 complete genomes at coverage 4-20x.

Individual missing rate (a); number of times a variant was called (b) and cumulative number of times a variant was called (c).

Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M (2015) A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig. PLoS ONE 10(3): e0118867. doi:10.1371/journal.pone.0118867
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0118867

PLOS | ONE

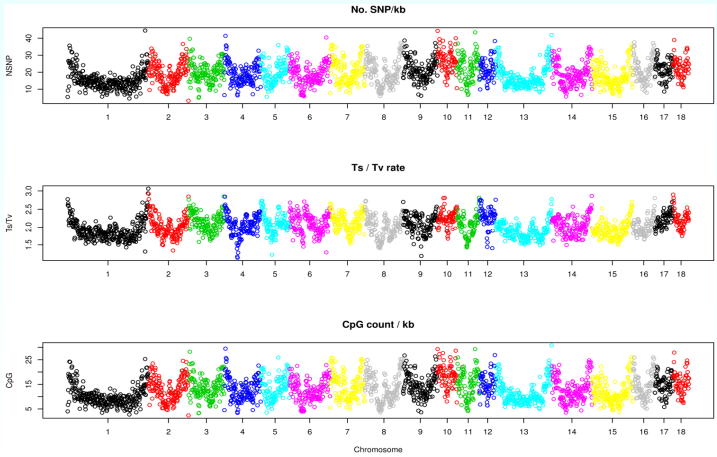
Sometimes, big is not what you need:
new omics data shrink even more the n/p ratio!

- Traditionally, sequence data have been in the form of small fragments over a moderate n size → a large number of site frequency spectrum based tests (Tajima's D, Fay-Wu…).
- New genome wide population sequence data are still in very small n compared to number SNPs.
- Good for inferring demographic history as different genome regions are interchangeable, but not for local selective tests, eg Fst estimates are highly biased for small n.
- Much theoretical work remains to be done here.

WHY BIG?

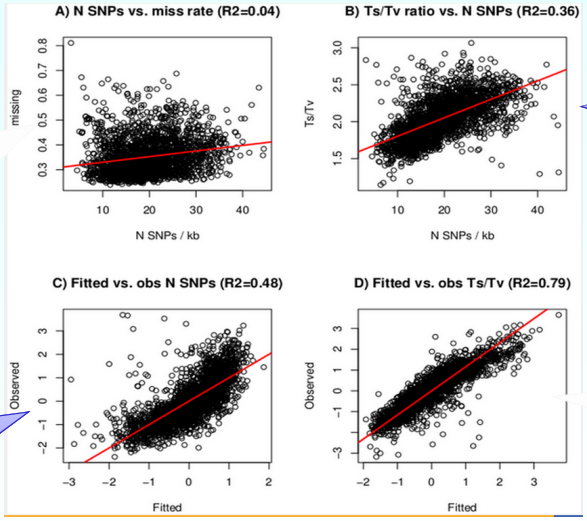
- To an extent, because there is no other option.
- Collecting and storing 'data' has become ridiculously inexpensive: this has created a sort of 'inverse experimental design' approach.
- Undeniable success of big data paradigm in other fields, such as marketing and opinion influence.

Sometimes big is great:
it allows you to get new insights



Bianco et al. 2015. 128 complete genomes at coverage 4-20x.

No correlation between missing data and no. SNPs, ergo no. SNPs is a good proxy for variability



Correlation between Ts/Tv ratio and no. SNPs exists but <<1

We can predict relatively well, but not perfectly, no. of SNPs

Instead, Ts/Tv can be adjusted quite well

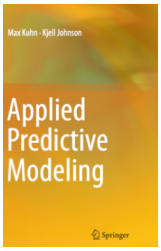
Variability in Ts/Tv rates is well explained by a differential composition in CpG in the genome and varying recombination rates, GC% and gene density unimportant. Our analyses also suggest that the correlation of number of SNPs and Ts/Tv ratio that we observe is likely an indirect consequence of both variables being affected by the same genome features, i.e., recombination rate and high mutability of CpG rich regions.

Big Data in Genomics

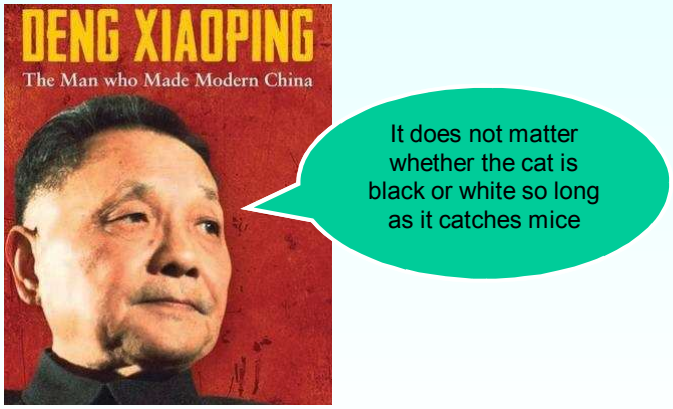
They are primarily in the Human area

- 2005: Wellcome trust case control consortium (WTCCC), a few common diseases, n ~ 19k
- 2015: Biobank UK: n = 500k , p ~ 800k, numerous phenotypes
- …

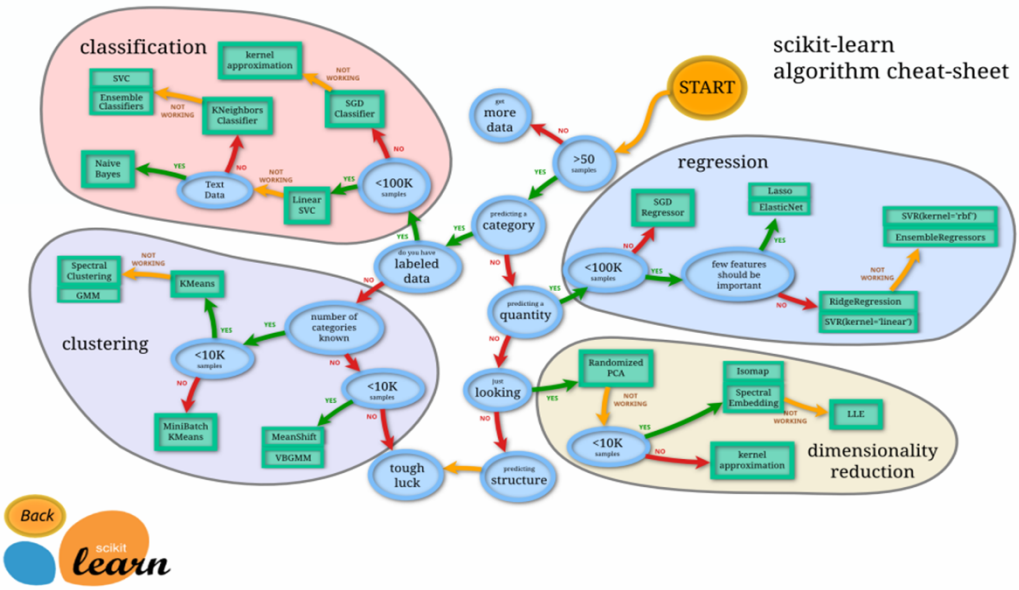
Machine Learning vs Statistics



Statistics	Machine Learning
➢ Focused on inference	➢ The target is prediction
➢ Based on Models	➢ Model free
➢ Theoretically founded	➢ Pragmatic
➢ Problem constrained	➢ Data heterogeneity is no problem
➢ 'Clear' interpretation	➢ Often cannot be interpreted
➢ General solutions	➢ Specific solutions

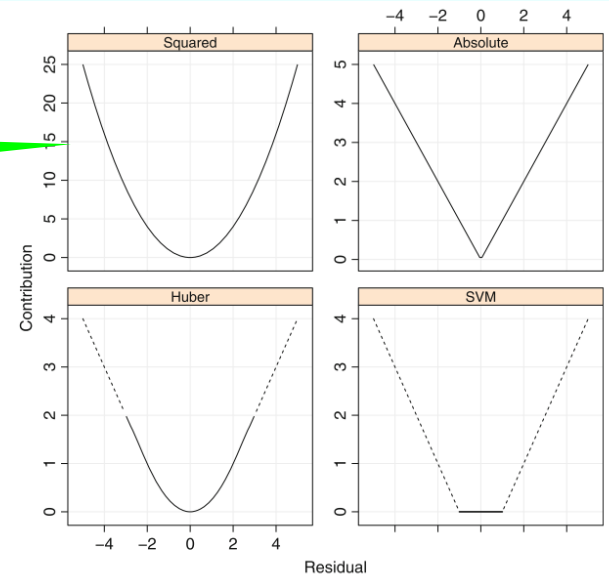


Machine Learning Visionary



Machine Learning: Support Vector Machines

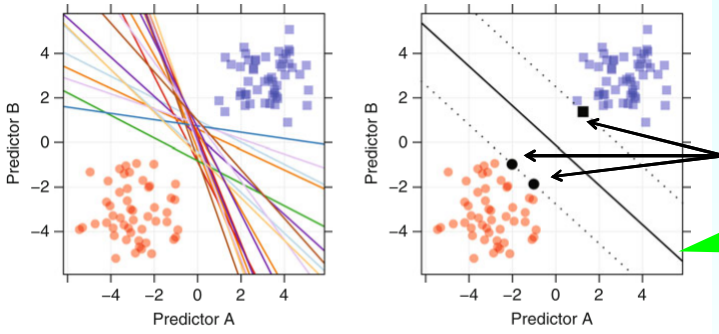
- Develop by Vapnik mid 60's, one of the most flexible and effective ML tools.
- Originally for classification, but regression also possible.
- One disadvantage of SSE in regression is that an outlier observation contributes enormously to the solution.
- In SVM, given a threshold T, residuals e < T do not contribute to error, whereas e > T contribute linearly.



Regression

Lasso

Fig. 7.6: The relationship between a model residual and its contribution to the regression line for several techniques. For the Huber approach, a threshold of 2 was used while for the support vector machine, a value of $\epsilon = 1$ was used. Note that the y-axis scales are different to make the figures easier to read



Support vectors

Maximum margin classifier

Fig. 13.9: *Left*: A data set with completely separable classes. An infinite number of linear class boundaries would produce zero errors. *Right*: The class boundary associated with the linear maximum margin classifier. The solid black points indicate the support vectors

- The criterion is to obtain the maximum margin classifier.
- If classes are not completely separable, a penalty is introduced that controls model complexity (soft margin case).

$$Min(\alpha) \sum_{j=1}^{n} \left[1 - y_j \left(\alpha_0 + \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) \right) \right] + \lambda \alpha' \mathbf{K} \alpha$$

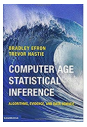
Variables for i-th observation, is a vector

parameters

observation

Kernel: allows non linear separation

penalty



The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

Machine Learning: Kernel methods

- SVM are part of a more general class of prediction methods, kernel methods.
- Very flexible and accurate.
- Allow complex separations
- BLUP is related to kernel methods, as the G matrix is a kernel, matrix of similarity.
- Allow combining different sources of information

Machine Learning: Tree / random forests

- Decision trees are classifiers / regressors that work by a set of hierarchical binary decisions.
- Easily visualized and, to an extent, easy to interpret.
- Parameters to be determined are predictor to split, tree depth, criteria to split and stop.
- Random forests are a special class of ensemble methods that combine several trees where the variables have been randomly sampled.

Some advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. Other techniques are usually specialised in analysing datasets that have only one type of variable. See algorithms for more information.
- Able to handle multi-output problems.
- Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

The disadvantages of decision trees include:

- Decision-tree learners can create over-complex trees that do not generalise the data well. This is called overfitting. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.
- There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.
- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

Machine Learning: Ensemble methods

Ensemble methods aims at combining the output of different methods which individual performance can be poor.

Two families of ensemble methods are usually distinguished:

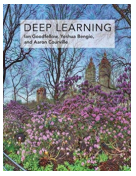
- In **averaging methods**, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.

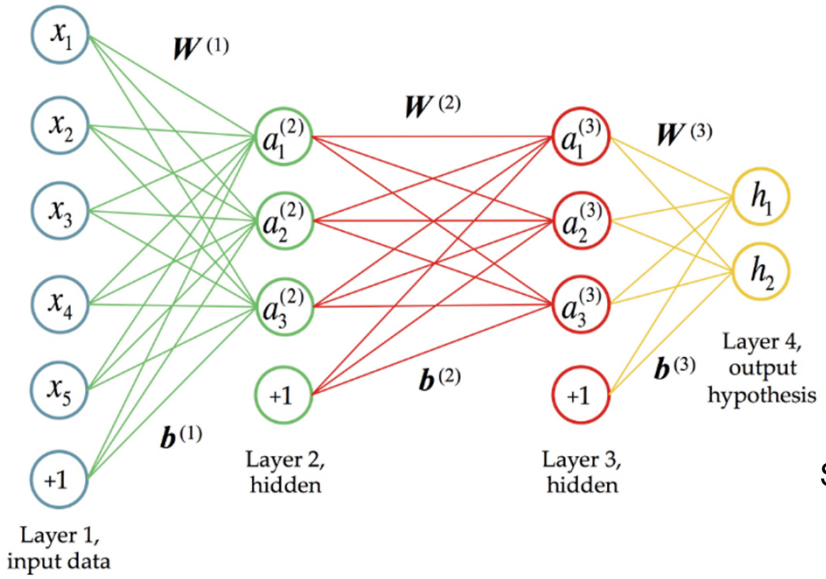
  Examples: Bagging methods, Forests of randomized trees, …
- By contrast, in **boosting methods**, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.

  Examples: AdaBoost, Gradient Tree Boosting, …

Machine Learning: Deep learning

- They are very much on fashion, despite the fact that the simplest versions (the perceptron or one layer neuron network) have been known for decades.
- Recent revival is due to new advances mainly in optimization and to success in many areas (image, speech, …)
- The main component is a 'neuron', which is a (non-)linear transformation of inputs (beginning with variable values).
- Modern predictors are made up of several layers of neurons.





Sheehan, Son, 2016

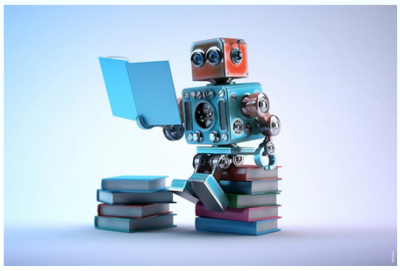
Fig 6. An example of a deep neural network with two hidden layers. The first layer is the input data (each dataset has 5 statistics), and the last layer predicts the 2 response variables. The last node in each input layer (+1) represents the bias term. Here the number of layers $L = 4$, and the number of nodes (computational units) in each layer is $u_1 = 5$, $u_2 = 3$, $u_3 = 3$, and $u_4 = 2$ (these counts exclude the biases).

- Multilayer Perceptrons (MLPs) depend on numerous parameters.
- Determining the optimum of these values is done by internal crossvalidation within the training data set.
- LOT of cooking and others' experience!

Some hyperparameters

Parameter	Issues
No. of layers	Risk of over/underfitting
No. of neurons / layer	Same
Activation function	Problem specific
Weight regularization	Constraint on weights

Will the future be unsupervised?



https://robohub.org/

ARTICLE

doi:10.1038/nature24270

Mastering the game of Go without human knowledge

David Silver[1]*, Julian Schrittwieser[1]*, Karen Simonyan[1]*, Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules. AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo's own move selections and also the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.

Nature Oct 2017

https://www.youtube.com/watch?v=5BrNt38OraE
https://www.youtube.com/watch?v=IbjF5VjniVE