

The Two Cultures

Aldo Solari

Data Mining



Outline

① The Two Cultures

② Data Science

③ Big Data



The two cultures

Breiman (2001) distinguished between

Statistical modeling

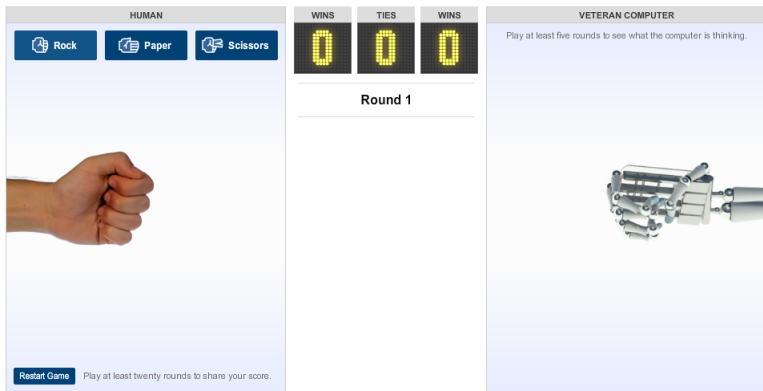
- emphasis on **probability models**
- the goal is **explanation** (by assessing the **uncertainty** of the estimates)

Machine learning

- emphasis on **algorithms**
- the goal is **prediction accuracy**



Rock-paper-scissors: you vs. the machine











<http://www.nytimes.com/interactive/science/rock-paper-scissors.html>



What the computer is thinking


Hide What the Computer is Thinking

Your Throw History




Your last throws were **PAPER, SCISSORS, SCISSORS, ROCK**

Continue



I am going to search over 200,000 rounds of Rock-Paper-Scissors data and find all the times when the human played **PAPER, SCISSORS, SCISSORS, ROCK** when I played **SCISSORS, PAPER, PAPER, SCISSORS**

Continue



Of all the times humans played **PAPER, SCISSORS, SCISSORS, ROCK** and I played **SCISSORS, PAPER, PAPER, SCISSORS**, they played **PAPER** as their next throw the most. I am going to assume that you will do the same this time.

Continue



Machine learning algorithm

Fact

A truly random game would result in a tie

Hypothesis

A human is not truly random

Machine strategy

Learn humans non-random patterns from the data

Human counter-attack

?



Same thing, different name?

Machine Learning

Statistics

target variable, output

response variable

Y

attribute, feature, input

predictor, explanatory variable

X

supervised learning

regression

model Y as a function of X

hypothesis

model, regression function

$$Y = f(X) + \epsilon$$

Source: Hothorn (2015) Big Data-Big Knowledge?



Same thing, different name?

Machine Learning

Statistics

instances, examples

samples, observations

$$(y_1, x_1), \dots, (y_n, x_n)$$

learning

estimation, fitting

$$\hat{f}(x) = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \text{loss}(y_i, f(x_i))$$

classification

prediction

$$\hat{y} = \hat{f}(x)$$

generalization error

risk

$$\mathbb{E}[\text{loss}(Y, f(X))]$$



So, what's the difference?

Machine Learning

Statistics

focus

prediction

significance

culture

algorithmic/optimization

modeling

methods

decision trees

k-nearest-neighbors

neural networks

support vector machines

adaboost

random forests

...

linear/logistic regression

discriminant analysis

mixed models

ridge/lasso regression

GAM

random forests

...

random forests?



Working together

Leo Breiman's fundamental contributions:

- Bagging
- Random Forests
- Boosting

Breiman's work helped to bridge the gap between statistical modeling and machine learning:

Statistical Learning = Statistical Modeling + Machine Learning



Statistical learning

Unsupervised learning

- the data consists of a set of variables X_1, \dots, X_p ; no variable has a special status
- the goal is clustering, dimensionality reduction, etc.

Supervised learning

- the data consists of both the response Y and the predictors X_1, \dots, X_p
- the problem is called supervised learning since the response 'supervises' the learning process
- the goal is (usually) prediction of the response



Supervised learning

Variables

- Response: Y
- Predictors: $X_{p \times 1} = (X_1, \dots, X_p)^T$

Problems

- Regression: Y is quantitative
- Classification: Y is binary or multi-class



The data: training and test

Training data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Test data

$$(x_1^*, ?), (x_2^*, ?), \dots, (x_m^*, ?)$$

Goal

Learn from the training data

$$(x_1, y_1), \dots, (x_n, y_n) \mapsto \hat{f}(x)$$

and predict the unseen y_1^*, \dots, y_m^* by

$$\hat{y}_1^* = \hat{f}(x_1^*), \dots, \hat{y}_m^* = \hat{f}(x_m^*)$$



Convergence

Data mining

finding patterns in data

Statistical learning

machine learning from a statistical point of view

...

converge to data science



Outline

① The Two Cultures

② Data Science

③ Big Data



What is data science?



“ A data scientist is a statistician who
lives in San Francisco.

Data Science is statistics on a Mac.

A data scientist is someone who is
better at statistics than any software
engineer and better at software
engineering than any statistician. ”

www.quora.com/Data-Science/What-is-the-difference-between-a-data-scientist-and-a-statistician



Data science

is an interdisciplinary field [...]

to extract knowledge or insights from data [...]

Wikipedia



Statistics \subset data science

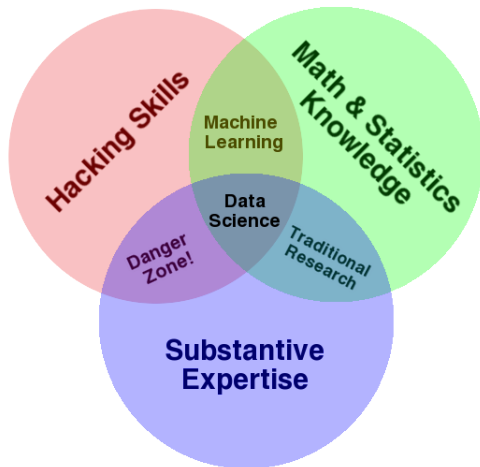
While statistics - as the science of learning from data - is necessary for turning data into knowledge and action, it's not the only critical component within data science

When statistics, database management, and distributed/parallel computing combine, we will see growth in cross-trained experts who are better equipped to solve complex challenges in today's massive data revolution

Jessica Utts, ASA president



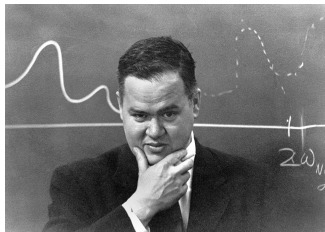
Interdisciplinary field



Source: Drew Conway, 2010



Data science = applied statistics



1915 - 2000

John W. Tukey (1962) *The future of data analysis* revisited by
Donoho (2015) *50 years of Data Science*



Key word

Most people hyping Data Science have focused on the first word:
Data

*The key word in "Data Science" is not Data, it is
Science*

Jeff Leek @ simply statistics

Data Science is only useful when the data are used to answer a
scientific question



Is Carl-Friedrich Gauß a data scientist?



1777 - 1855



Astronomy problem

Predict the position of the asteroid Ceres at 31 December 1801 on the basis of data provided by the italian astronomer Giuseppe Piazzi

Statistical problem

Response $\mathbf{y}_{n \times 1}$, design matrix $\mathbf{X}_{n \times p}$ and parameters $\boldsymbol{\beta}_{p \times 1}$

Find $\hat{\boldsymbol{\beta}}$ such that minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

Solution: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Computational problem

Solve (by hand!) the system of equations $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$

Solution: fast optimization algorithm (Gaussian elimination)



Outline

① The Two Cultures

② Data Science

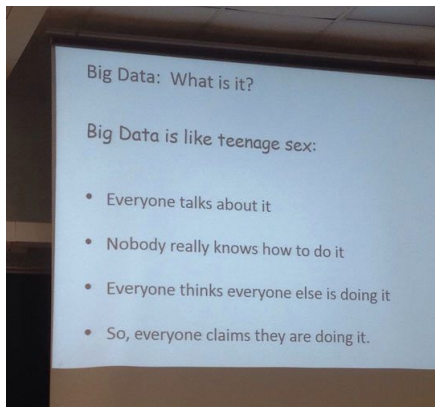
③ Big Data



Big data revolution



Big data?



Source: Dan Ariely, Facebook post on January 2013



Big data

is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them

Wikipedia

- If you have several gigabytes of data or several million observations, standard relational databases become unwieldy; databases to manage data of this size are generically known as “NoSQL” databases
- Tools for manipulating big data are given in the next table



Tools for Manipulating Big Data

<i>Google name</i>	<i>Analog</i>	<i>Description</i>
Google File System	Hadoop File System	This system supports files so large that they must be distributed across hundreds or even thousands of computers.
Bigtable	Cassandra	This is a table of data that lives in the Google File System. It too can stretch over many computers.
MapReduce	Hadoop	This is a system for accessing and manipulating data in large data structures such as Bigtables. MapReduce allows you to access the data in parallel, using hundreds or thousands of machines to extract the data you are interested in. The query is “mapped” to the machines and is then applied in parallel to different shards of the data. The partial calculations are then combined (“reduced”) to create the summary table you are interested in.
Sawzall	Pig	This is a language for creating MapReduce jobs.
Go	None	Go is flexible open-source, general-purpose computer language that makes it easier to do parallel data processing.
Dremel, BigQuery	Hive, Drill, Impala	This is a tool that allows data queries to be written in a simplified form of of Structured Query Language (SQL). With Dremel it is possible to run an SQL query on a petabyte of data (1,000 terabytes) in a few seconds.

Source: Varian (2013) Big Data: New Tricks for Econometrics



Data: tall and fat

The outcome of the big-data processing:

data set = $n \times p$ matrix

BIG n = tall

= computational problem

BIG p = fat

= curse of dimensionality



The end of theory?



Petabytes allow us to say: correlation is enough

Source: Anderson (2008) *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired Magazine 16.07



Experimental data

Scientists have

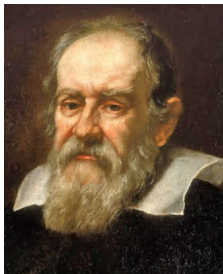
- a theory/hypothesis, and thus a model
- a planned experiment - and thus experimental data

Traditional statistical methods use data to

- estimate parameters in the model and assess their uncertainty
- provide means to falsify a theory/hypothesis and/or to formulate a better theory/hypothesis



Galileo Galilei



1565-1642



Inclined plane experiment (1604)



Question of interest

If a ball rolls down a ramp, what is the relationship between time and distance?



Aristotle vs Galileo

Aristotle theory

Constant velocity (zero acceleration): distance \propto time

Galileo theory

Increasing velocity (constant acceleration): distance \propto time²

Model and null hypothesis

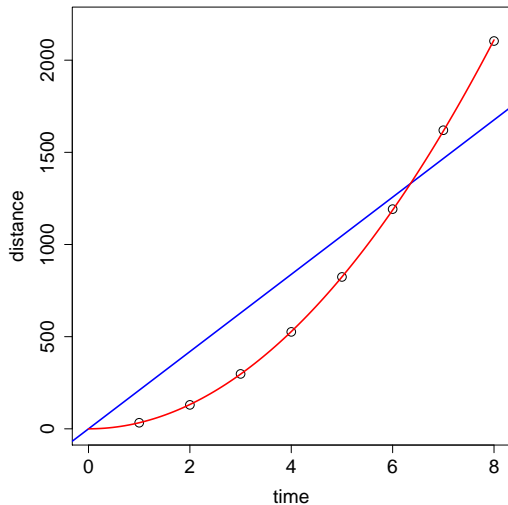
- $d = \beta_1 \cdot t + \beta_2 \cdot t^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$
- $H_0 : \beta_2 = 0$

Data

time	1	2	3	4	5	6	7	8
distance	33	130	298	526	824	1192	1620	2104



Model fit



blue line is Aristotle model, red curve is Galileo model



```
# data
time = 1:8
distance = c(33, 130, 298, 526, 824, 1192, 1620, 2104)
rolling = data.frame(distance, time)

# aristotle model
fit0 = lm(distance ~ 0 + time, rolling)
# galileo model
fit1 = lm(distance ~ 0 + time + I(time^2) , rolling)

anova(fit0, fit1)
# Null hypothesis rejected with p-value = 9.81e-13
```



What's different in big data?

Data scientists have

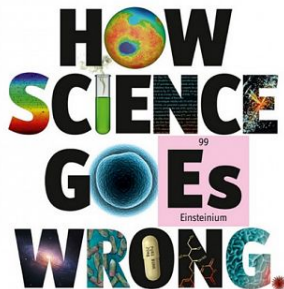
- large and complex data set from unplanned source
- data-driven theories/hypotheses

Modern statistical methods should take into account

- the exploratory approach that provides hypotheses/models
- risk of overfitting and p -hacking

If you torture the data long enough, Nature will always confess (R.H. Coase)





Ioannidis (2005) *Why Most Published Research Findings Are False.*
PLoS Med



p-hacking

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- ☐ Presidents
- ☐ Governors
- ☒ Senators
- ☐ Representatives

How do you want to measure economic performance?

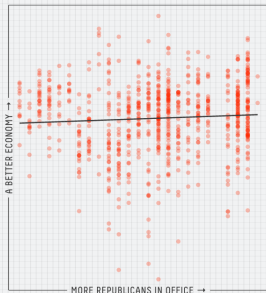
- ☐ Employment
- ☒ Inflation
- ☐ GDP
- ☒ Stock prices

Other options

- ☐ Factor in power
Weight more powerful positions more heavily
- ☐ Exclude recessions
Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



Result: Almost

Your **0.08** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Christie Aschwanden (2015) Science isn't broken

