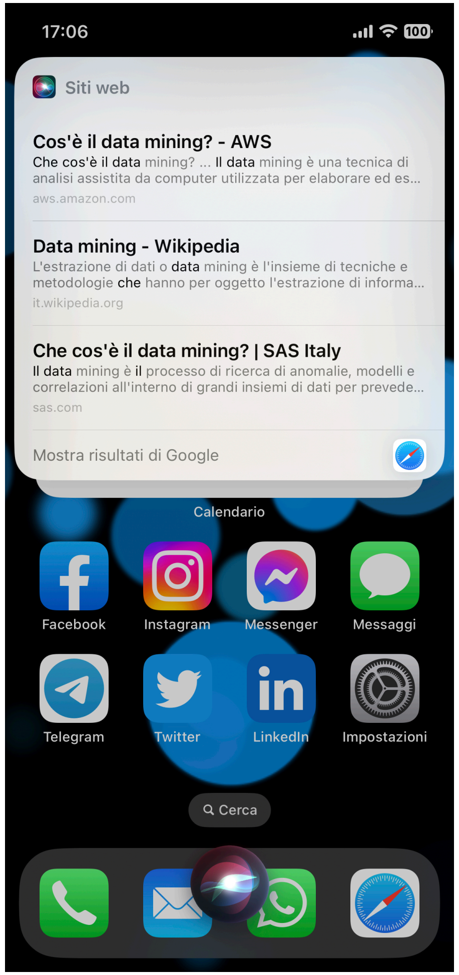
Introduction

Data Mining - CdL CLAMSES

Tommaso Rigon
Università degli Studi di Milano-Bicocca

Home page

Homepage



Siti web

Cos'è il data mining? - AWS
Che cos'è il data mining? ... Il data mining è una tecnica di analisi assistita da computer utilizzata per elaborare ed es...
aws.amazon.com

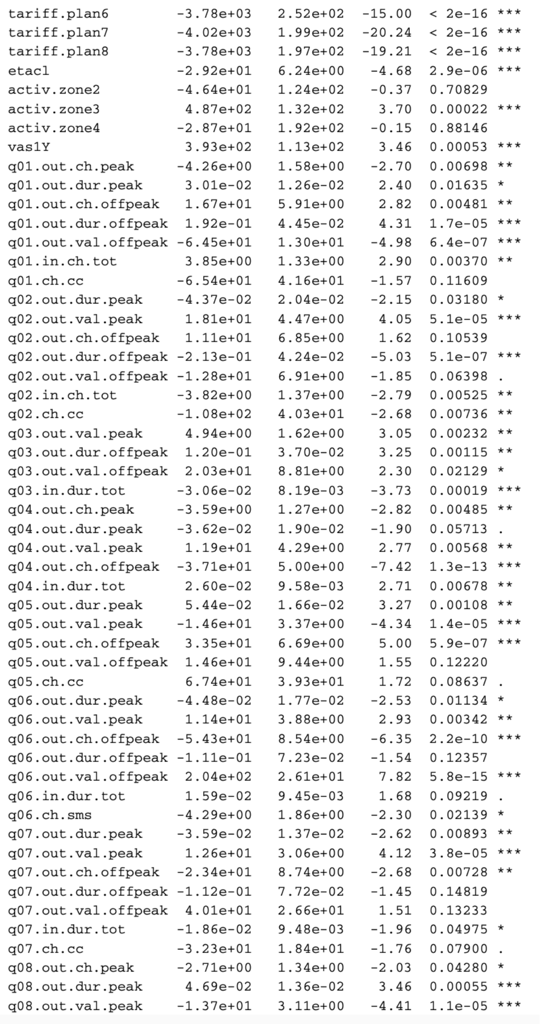
Data mining - Wikipedia
L'estrazione di dati o data mining è l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di informa...
it.wikipedia.org

Che cos'è il data mining? | SAS Italy
Il data mining è il processo di ricerca di anomalie, modelli e correlazioni all'interno di grandi insiemi di dati per prevede...
sas.com

Mostra risultati di Google

- Nowadays, **predictive algorithms** have become **mainstream** in the **popular culture** due to some spectacular successes:
  - iPhone's Siri;
  - Google translate;
  - recommendation systems (Netflix challenge);
  - business analytics (churn prediction, credit risk assessment);
  - and, more recently, chatGPT.
- And yet, there is a lot of **confusion** about the **history** and the **boundaries** of the field. For instance, what is "data mining"?
- And what are then the differences, **if any**, with statistics, machine learning, statistical learning, and data science?
- What applied problems cannot be solved with **classical statistical** tools? Why?
- Let us consider some real **case studies**...

Traffic prediction in telecommunications



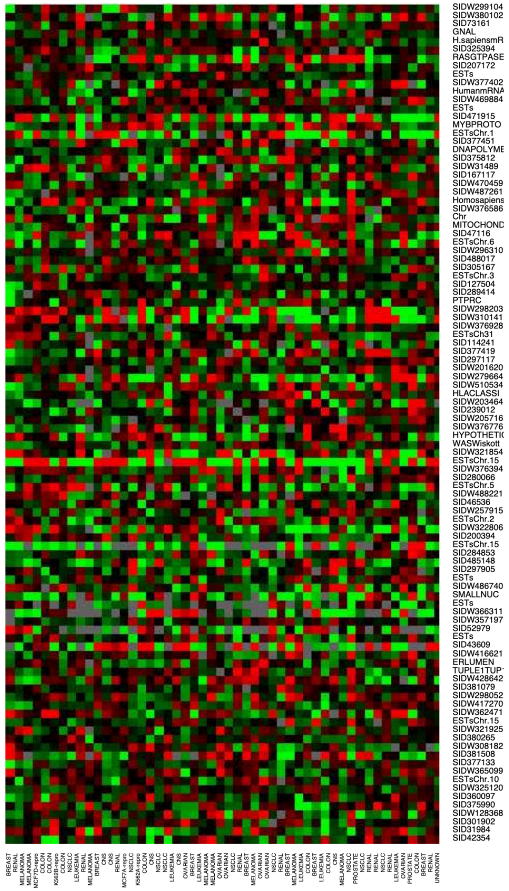
```
tariff.plan6        -3.78e+03   2.52e+02  -15.00  < 2e-16 ***
tariff.plan7        -4.02e+03   1.99e+02  -20.24  < 2e-16 ***
tariff.plan8        -3.78e+03   1.97e+02  -19.21  < 2e-16 ***
etacl               -2.92e+01   6.24e+00   -4.68  2.9e-06 ***
activ.zone2         -4.64e+01   1.24e+02   -0.37  0.70829
activ.zone3          4.87e+02   1.32e+02    3.70  0.00022 ***
activ.zone4         -2.87e+01   1.92e+02   -0.15  0.88146
vas1Y                3.93e+02   1.13e+02    3.46  0.00053 ***
q01.out.ch.peak     -4.26e+00   1.58e+00   -2.70  0.00698 **
q01.out.dur.peak     3.01e-02   1.26e-02    2.40  0.01635 *
q01.out.ch.offpeak   1.67e+01   5.91e+00    2.82  0.00481 **
q01.out.dur.offpeak  1.92e-01   4.45e-02    4.31  1.7e-05 ***
q01.out.val.offpeak -6.45e+01   1.30e+01   -4.98  6.4e-07 ***
q01.in.ch.tot        3.85e+00   1.33e+00    2.90  0.00370 **
q01.ch.cc           -6.54e+01   4.16e+01   -1.57  0.11609
q02.out.dur.peak    -4.37e-02   2.04e-02   -2.15  0.03180 *
q02.out.val.peak     1.81e+01   4.47e+00    4.05  5.1e-05 ***
q02.out.ch.offpeak   1.11e+01   6.85e+00    1.62  0.10539
q02.out.dur.offpeak -2.13e-01   4.24e-02   -5.03  5.1e-07 ***
q02.out.val.offpeak -1.28e+01   6.91e+00   -1.85  0.06398 .
q02.in.ch.tot       -3.82e+00   1.37e+00   -2.79  0.00525 **
q02.ch.cc           -1.08e+02   4.03e+01   -2.68  0.00736 **
q03.out.val.peak     4.94e+00   1.62e+00    3.05  0.00232 **
q03.out.dur.offpeak  1.20e-01   3.70e-02    3.25  0.00115 **
q03.out.val.offpeak  2.03e+01   8.81e+00    2.30  0.02129 *
q03.in.dur.tot      -3.06e-02   8.19e-03   -3.73  0.00019 ***
q04.out.ch.peak     -3.59e+00   1.27e+00   -2.82  0.00485 **
q04.out.dur.peak    -3.62e-02   1.90e-02   -1.90  0.05713 .
q04.out.val.peak     1.19e+01   4.29e+00    2.77  0.00568 **
q04.out.ch.offpeak  -3.71e+01   5.00e+00   -7.42  1.3e-13 ***
q04.in.dur.tot       2.60e-02   9.58e-03    2.71  0.00678 **
q05.out.dur.peak     5.44e-02   1.66e-02    3.27  0.00108 **
q05.out.val.peak    -1.46e+01   3.37e+00   -4.34  1.4e-05 ***
q05.out.ch.offpeak   3.35e+01   6.69e+00    5.00  5.9e-07 ***
q05.out.val.offpeak  1.46e+01   9.44e+00    1.55  0.12220
q05.ch.cc            6.74e+01   3.93e+01    1.72  0.08637 .
q06.out.dur.peak    -4.48e-02   1.77e-02   -2.53  0.01134 *
q06.out.val.peak     1.14e+01   3.88e+00    2.93  0.00342 **
q06.out.ch.offpeak  -5.43e+01   8.54e+00   -6.35  2.2e-10 ***
q06.out.dur.offpeak -1.11e-01   7.23e-02   -1.54  0.12357
q06.out.val.offpeak  2.04e+02   2.61e+01    7.82  5.8e-15 ***
q06.in.dur.tot       1.59e-02   9.45e-03    1.68  0.09219 .
q06.ch.sms          -4.29e+00   1.86e+00   -2.30  0.02139 *
q07.out.dur.peak    -3.59e-02   1.37e-02   -2.62  0.00893 **
q07.out.val.peak     1.26e+01   3.06e+00    4.12  3.8e-05 ***
q07.out.ch.offpeak  -2.34e+01   8.74e+00   -2.68  0.00728 **
q07.out.dur.offpeak -1.12e-01   7.72e-02   -1.45  0.14819
q07.out.val.offpeak  4.01e+01   2.66e+01    1.51  0.13233
q07.in.dur.tot      -1.86e-02   9.48e-03   -1.96  0.04975 *
q07.ch.cc           -3.23e+01   1.84e+01   -1.76  0.07900 .
q08.out.ch.peak     -2.71e+00   1.34e+00   -2.03  0.04280 *
q08.out.dur.peak     4.69e-02   1.36e-02    3.46  0.00055 ***
q08.out.val.peak    -1.37e+01   3.11e+00   -4.41  1.1e-05 ***
```

- The marketing section of a telecommunications company is interested in analyzing the **customer behavior**.
- Hence, the data science team would like to **predict**, for every single customer, the **telephone traffic**.
- Traffic is measured as the total **number of seconds** of **outgoing calls** made in a given month by each customer.
- Appropriate estimations of the **overall traffic** provide necessary elements for:
  - predicting the company's budget;
  - early identification of possible dissatisfaction;
  - finding issues in the primary services of the company;
  - spotting fraudulent situations.
- The dataset has $n = 30.619$ customers and $p = 99$ **covariates**, i.e., the customer activity in **previous months**.

Traffic prediction in telecommunications II

- The focus is on **prediction** and on **learning something useful** from the data, not much on hypothesis testing.
- These are **observational data**, which have been collected for other purposes, not for their analysis. Data "exists," there is **no sampling design**.
- Data are **dirty** and often stored in big data warehouse (DWH).
- The **dimension of the data** is **large** in both directions: large n and large p. Hence:
  - All **p-values** are **ultra-significant** and not very informative in this setting;
  - **Computations** are a crucial analysis aspect.
- The relationship between covariates and the response is complex, thus, it is hard to believe our models will be "true." They are **all wrong**!
- However, having a lot of data means we can **split** them, using the first half for estimation and the other half for testing.

Microarray cancer data

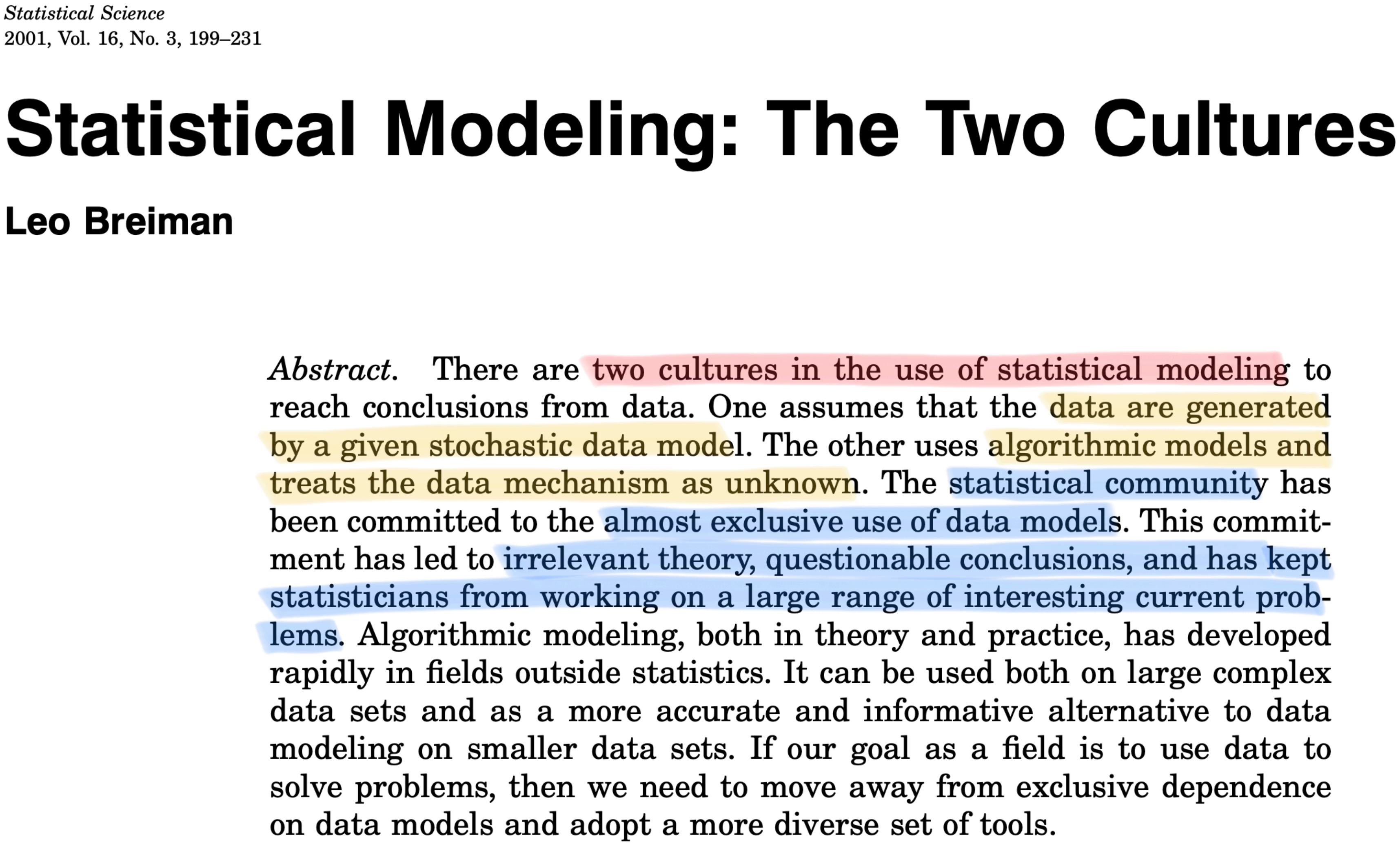


- **Expression matrix** of $p = 6830$ genes (rows) and $n = 64$ samples (columns), for the **human tumor data**.
- 100 randomly chosen rows are shown
- The picture is a **heatmap**, ranging from bright green (underexpressed) to bright red (overexpressed).
- **Missing values** are gray. The rows and columns are displayed in a randomly chosen order.
- Goal: **predict** cancer class based on expression values.
- The main **statistical difficulty** here is that $p > n$!
- Logistic regression and discriminant analysis wouldn't work; the estimates do not exist.
- Is it even possible to fit a model in this context?

The pitfalls of the old-fashioned way

- All the previous case studies cannot be solved using traditional tools; in fact:
  - there are **tons of variables**, sometimes even with $p > n$ and most of them are irrelevant. It is not clear how to select the most useful ones.
  - **p-values** are always **ultra-significant** and potentially meaningless.
  - there are **no true models** in these contexts. There is little hope that reality follows a linear specification.
- The objective is **predicting** a response variable in the most accurate way. Classical statistics has **broader goals** including, but not limited to, prediction.
- We need a **paradigm shift** to address the above issues.
- For instance, if reality is non-linear, what about going **nonparametric**? We could let the data speak without making any assumption about the relationship between y and $\boldsymbol{x}$.
- Moreover, if p-values and residual plots are no longer informative in this context, how do we **validate** our predictions?

A highly influential paper (Breiman, 2001)



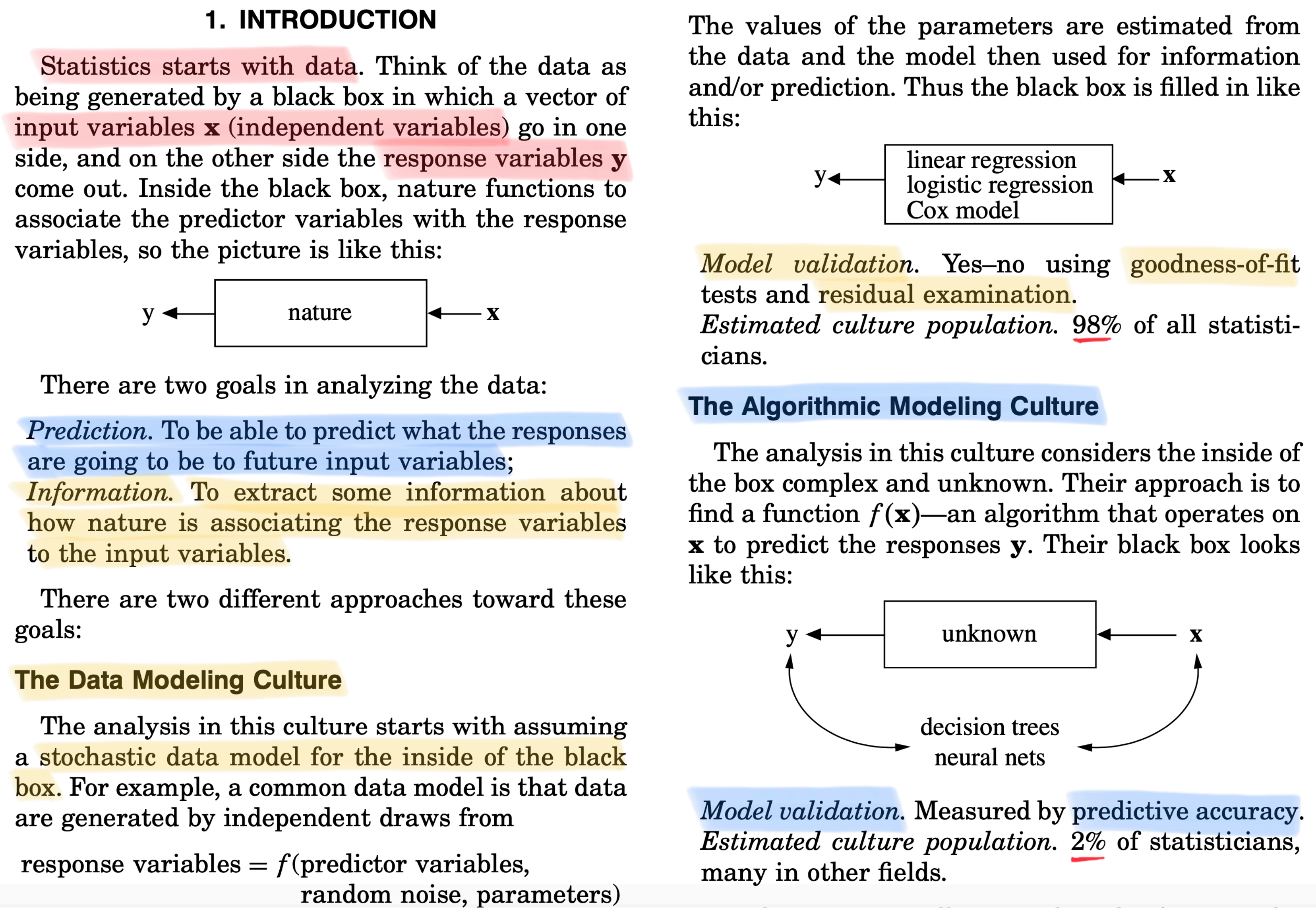
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Data models vs. algorithmic models



1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

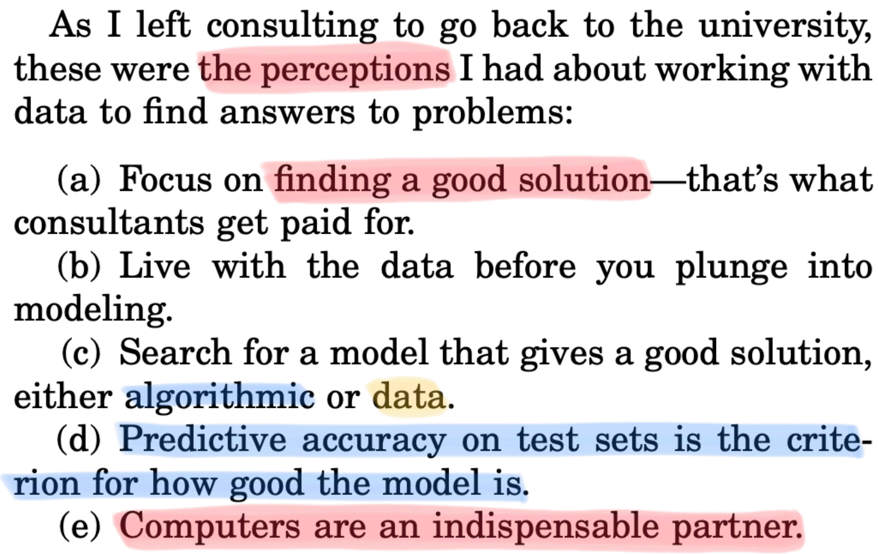
Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:

Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Focus on predictive accuracy & business solutions



As I left consulting to go back to the university, these were the perceptions I had about working with data to find answers to problems:

(a) Focus on finding a good solution—that's what consultants get paid for.

(b) Live with the data before you plunge into modeling.

(c) Search for a model that gives a good solution, either algorithmic or data.

(d) Predictive accuracy on test sets is the criterion for how good the model is.

(e) Computers are an indispensable partner.

Leo Breiman, 2003

- After the Ph.D., Breiman resigned and went into full-time **free-lance consulting**, and it worked as a consultant for thirteen years.
- Breiman joined the UC Berkeley **Statistics** Department in 1980.
- Leo Breiman died in 2005 at the age of 77. He **invented** many of the mainstream predictive tools: CART, bagging, random forests, stacking.

Statistical modeling: the two cultures

- It is **tempting** to fully embrace the **pure predictive viewpoint**, as Breiman did in his career, especially in light of the recent media attention and public interest.
- "*Statistical modeling: the two cultures*" has been a highly influential paper written by an outstanding statistician.
- In some cases, the paper may sound exaggerated and at times **confrontational**. These were different times.
- It was also a **discussion paper**!
- Two other giants of the discipline, **Sir David Cox** (died in 2022) and **Bradley Efron** were among the discussants and raised several critical points.
- It is **premature** to delve into those criticisms. We will get back to them at the end of the course once you have enough knowledge to understand them.

Prerequisites

- If you are in this class today, it means...
  - You already studied a lot of **real analysis**, **linear algebra** and **probability**;
  - You know how to **estimate the parameters** of a statistical model, to construct and interpret confidence intervals, p-values, etc. You know the **principles** of **inference**;
  - You know how to **explore data** using the **R** statistical software and other tools (SAS, python, etc.). You know **principal component analysis** and perhaps even factor models;
  - You know how to fit **linear models** and how to interpret the associated empirical findings. You are familiar with R^2s, likelihood ratio tests, **logistic regression**, and so on;
  - You may have attended a course named "data mining" before, and studied essential tools like linear discriminant analysis, k-nearest neighbors...
- These classical statistical tools are the **prerequisites** of **Data Mining M**. We will start from there.

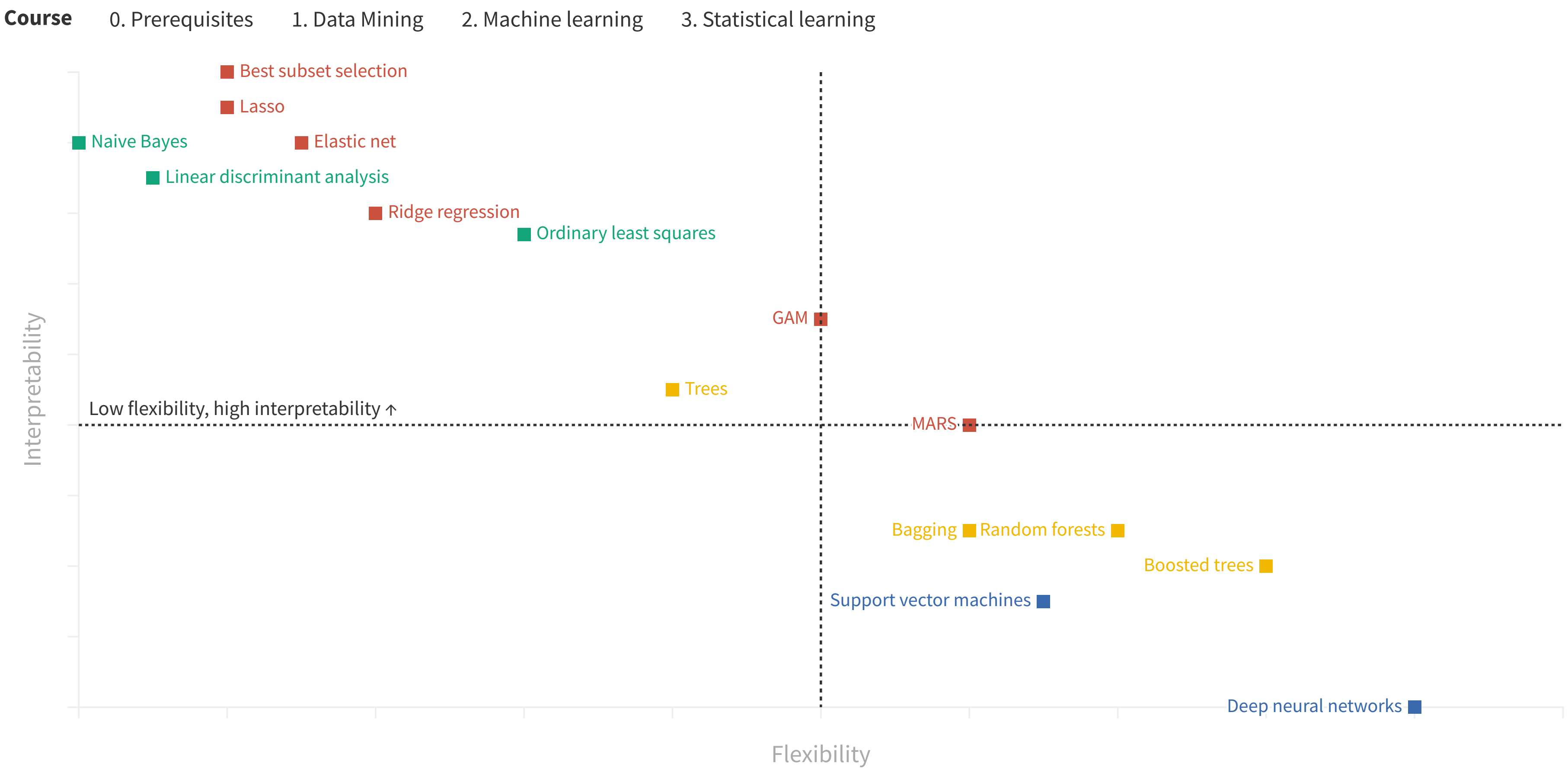
Overview of the topics

Unit	Description
A-B-C	Linear models. Data modeling, the old-fashioned way. Advanced computations.
Optimism, conflicts and trade-offs	Bias-variance trade-off. Training and test paradigm, cross-validation. Information criteria, optimism
Shrinkage and variable selection	Best subset selection, principal component regression. Ridge regression. Lasso and LARS. Elastic net.
Nonparametric estimation	Local linear regression. Regression and smoothing splines.
Additive models	Generalized additive models (GAM). Multivariate adaptive regression splines (MARS).

- Trees, bagging, random forests, boosting, neural networks, support vector machine are not in the program due to **time constraints**... but you will study them in other courses!

A tension between prediction and interpretability

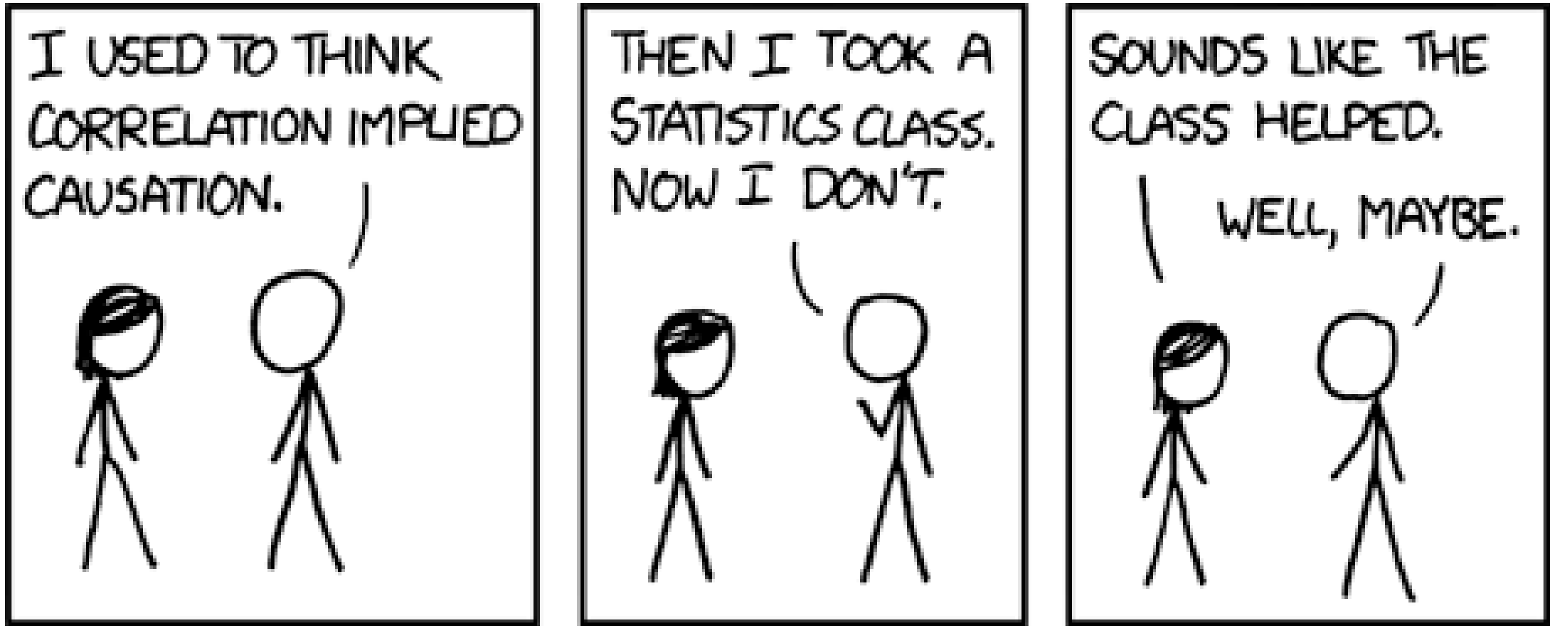
- Important **caveat**. Less flexible methods may have **more accurate predictions** in many case studies, on top of being more interpretable!



Home page

Made with Flourish • Create a scatter plot

Predictive interpretability $\neq$ causality


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

- Predictive **interpretability** means transparent understanding the **driving factors** of the predictions. An example is linear models with few (highly relevant) variables.
- For example, if I change the value of a set of covariates, what is the impact on predictions?
- This is **useful**, especially within the context of **ethical** AI and machine learning.
- However, the **predictive relevance** of a variable does **not** imply a **causal effect** on the response.
- Finding causal relationship requires **careful thinking**, a suitable **sampling design**, or both.

A definition of “data mining”

Azzalini & Scarpa (2011)

Data mining represents the work of processing, graphically or numerically, **large** amounts or continuous streams of **data**, with the aim of **extracting information** useful to those who possess them.

Hand et al. (2001)

Data mining is fundamentally an applied discipline [...] data mining requires an understanding of both **statistical** and **computational** issues. (p. xxviii)

[...]

The most fundamental difference between classical statistical applications and data mining is the **size** of the **data**. (p. 19)

Back to the 1800s

- At this point, it may sound natural to ask yourself: what is **statistics**?
- Statistics existed before data mining, machine learning, data science, and all these fancy new names.
- Statistical regression methods trace back to Gauss and Legendre in the early 1800s. Their goal was indeed **prediction**!

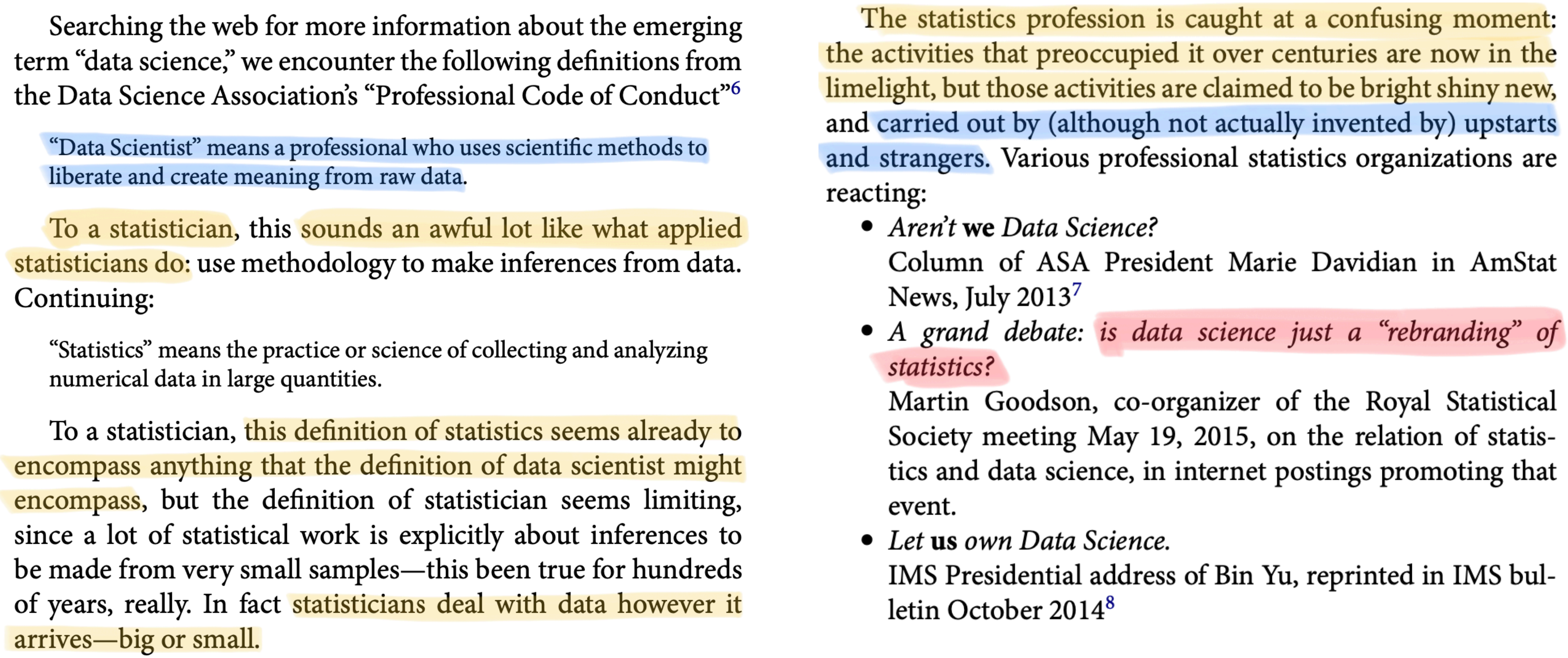
Davison (2003)

Statistics concerns what can be **learned** from **data**.

Hand (2011)

Statistics [...] is the technology of **extracting meaning** from **data**.

50 years of data science (Donoho, 2017)



Searching the web for more information about the emerging term "data science," we encounter the following definitions from the Data Science Association's "Professional Code of Conduct"[6]

> "Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.

To a statistician, this sounds an awful lot like what applied statisticians do: use methodology to make inferences from data. Continuing:

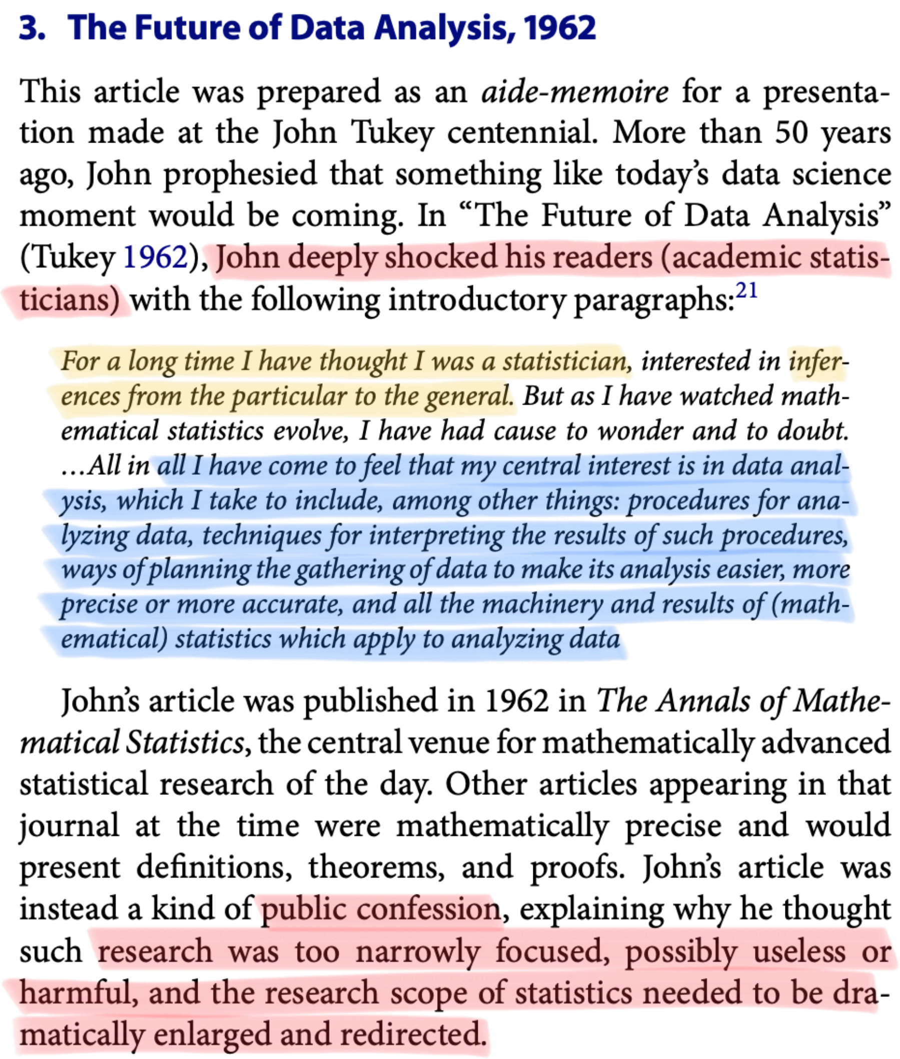
> "Statistics" means the practice or science of collecting and analyzing numerical data in large quantities.

To a statistician, this definition of statistics seems already to encompass anything that the definition of data scientist might encompass, but the definition of statistician seems limiting, since a lot of statistical work is explicitly about inferences to be made from very small samples—this been true for hundreds of years, really. In fact statisticians deal with data however it arrives—big or small.

The statistics profession is caught at a confusing moment: the activities that preoccupied it over centuries are now in the limelight, but those activities are claimed to be bright shiny new, and carried out by (although not actually invented by) upstarts and strangers. Various professional statistics organizations are reacting:

- *Aren't* **we** *Data Science?*
  Column of ASA President Marie Davidian in AmStat News, July 2013[7]
- *A grand debate: is data science just a "rebranding" of statistics?*
  Martin Goodson, co-organizer of the Royal Statistical Society meeting May 19, 2015, on the relation of statistics and data science, in internet postings promoting that event.
- *Let* **us** *own Data Science.*
  IMS Presidential address of Bin Yu, reprinted in IMS bulletin October 2014[8]

Statistics, an evolving science



3. The Future of Data Analysis, 1962

This article was prepared as an *aide-memoire* for a presentation made at the John Tukey centennial. More than 50 years ago, John prophesied that something like today's data science moment would be coming. In "The Future of Data Analysis" (Tukey 1962), John deeply shocked his readers (academic statisticians) with the following introductory paragraphs:[21]

> *For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. …All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data*

John's article was published in 1962 in *The Annals of Mathematical Statistics*, the central venue for mathematically advanced statistical research of the day. Other articles appearing in that journal at the time were mathematically precise and would present definitions, theorems, and proofs. John's article was instead a kind of public confession, explaining why he thought such research was too narrowly focused, possibly useless or harmful, and the research scope of statistics needed to be dramatically enlarged and redirected.

- Sure, old-fashioned statistics is often insufficient to address **modern challenges**.
- But statistics has profoundly changed over the years, **broadening** its **boundaries**.
- The road was paved by **Tukey** in the 60s, with further exhortations by **Breiman**.
- **Modern statistics** encompasses **also**:
  - Data gathering and representation;
  - Computational aspects;
  - Algorithmic and modeling culture to prediction.
- Feel free to call it "data science" if you like the bright, shiny new term.

A glossary

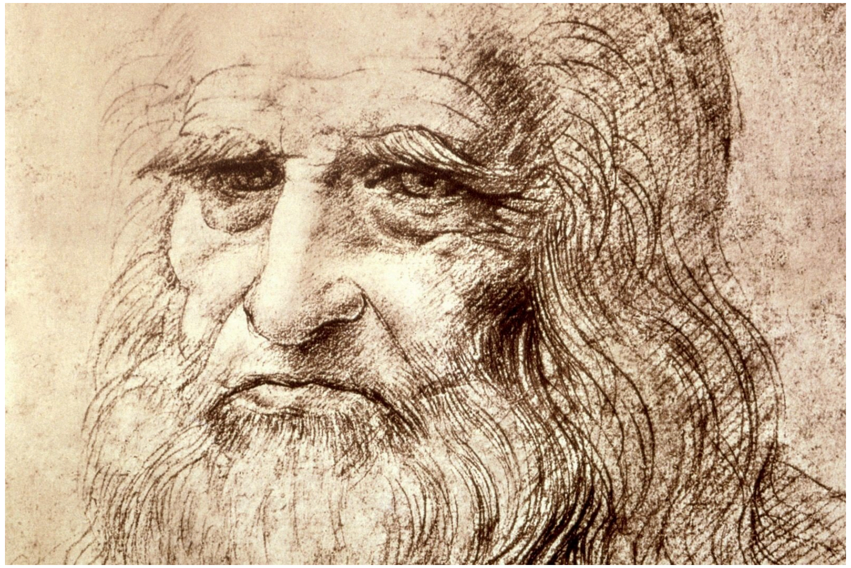
- While it might seem that **data science** and **data mining** have **strong roots** in **statistics**, it cannot be denied the existence of two distinct, albeit often overlapping, **communities**.
- For the lack of a better term, we will call these communities the **statisticians** and the **computer scientists**, as identified by their **background** and studies.

Statisticians	Computer Scientists
Parameters	Weights
Covariate	Feature
Observation	Instance
Response	Label
R	Python
Regression / Classification	Supervised learning
Density estimation, clustering	Unsupervised learning
Lasso / Ridge penalty	L^1 and L^2 penalty

Press the button?

- Several "automatic" tools have been developed over the years, tempting generations of analysts with **automatic pipelines**.
- Those who choose to "**press the button**":
  - do not know which method is used, they may only know the name of the method they are using;
  - are not aware of its **advantages** and **disadvantages**.
- More or less advanced **knowledge** of the methods is essential for:
  - choosing the most suitable method;
  - interpreting the results.
- Competence in **computational aspects** is helpful to evaluate better the output of the computer, e.g., in terms of its **reliability**.
- If you are not making the choices, somebody else is!

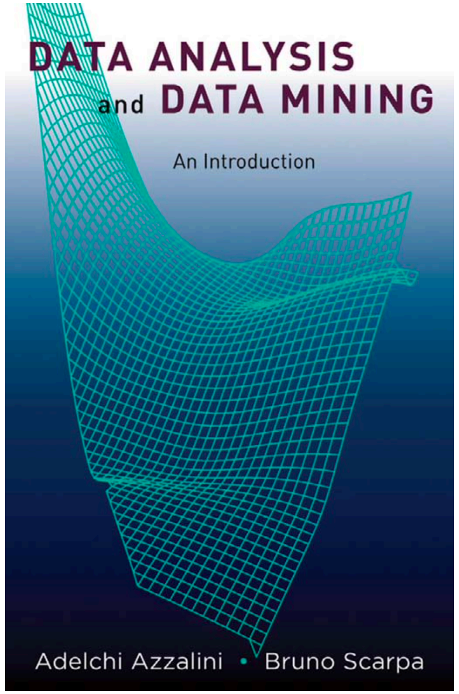
A matter of style



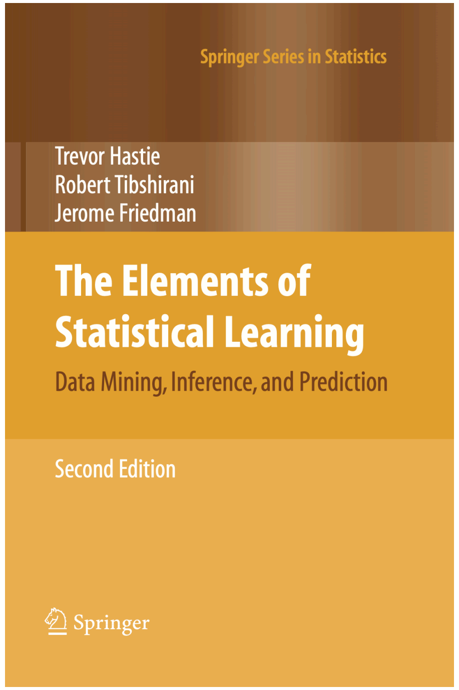
“*Quelli che s’innamoran di pratica sanza scienzia son come ’l nocchier ch’entra in navilio senza timone o bussola, che mai ha certezza dove si vada.*”

Leonardo da Vinci

Course material



A&S (2011)

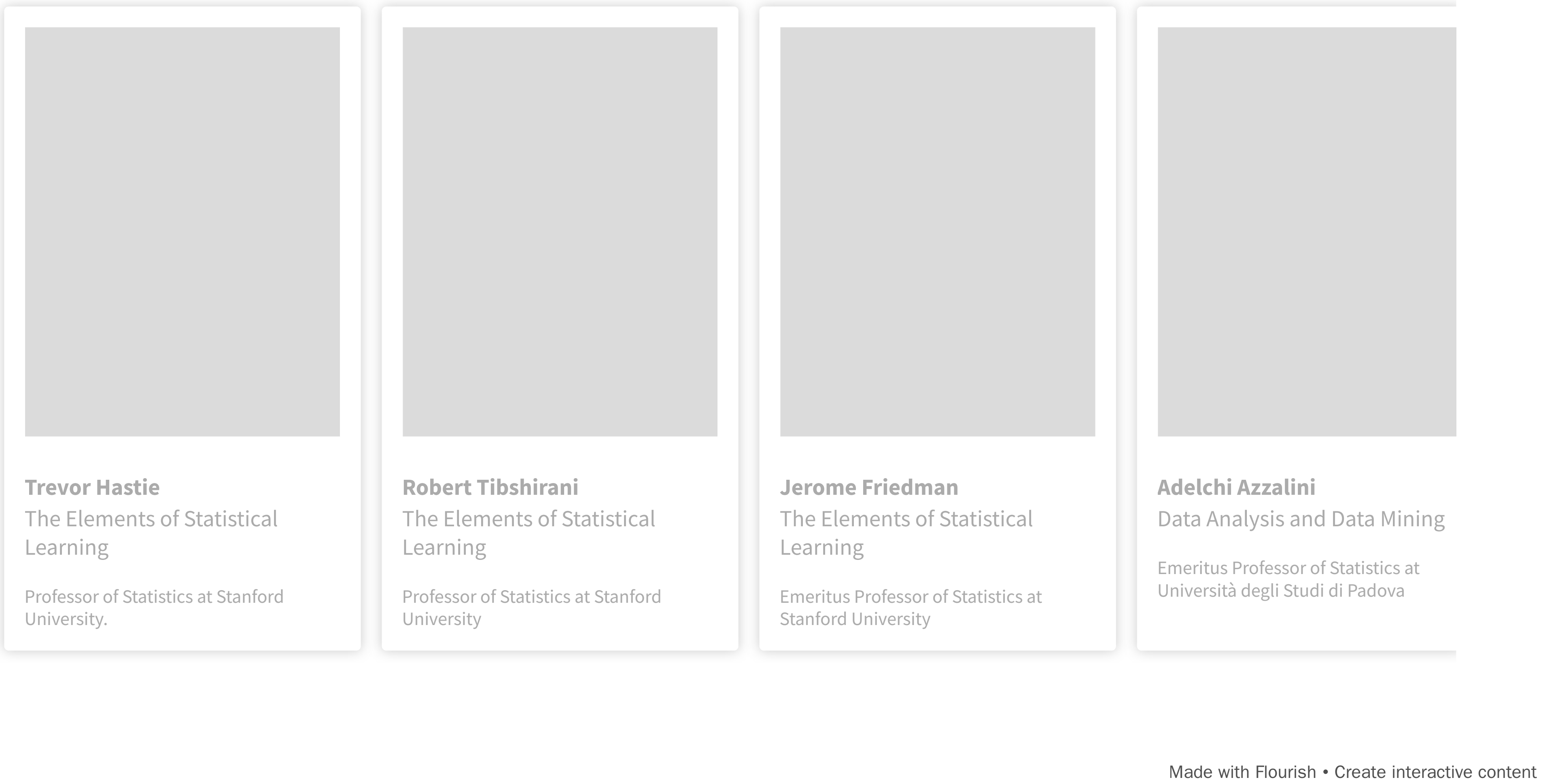

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

HTF (2009)

- Azzalini, A. and Scarpa, B. (2011), *Data Analysis and Data Mining*, Oxford University Press.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, Second Edition, Springer.

Home page

Data mining people



Trevor Hastie
The Elements of Statistical Learning

Professor of Statistics at Stanford University.

Robert Tibshirani
The Elements of Statistical Learning

Professor of Statistics at Stanford University

Jerome Friedman
The Elements of Statistical Learning

Emeritus Professor of Statistics at Stanford University

Adelchi Azzalini
Data Analysis and Data Mining

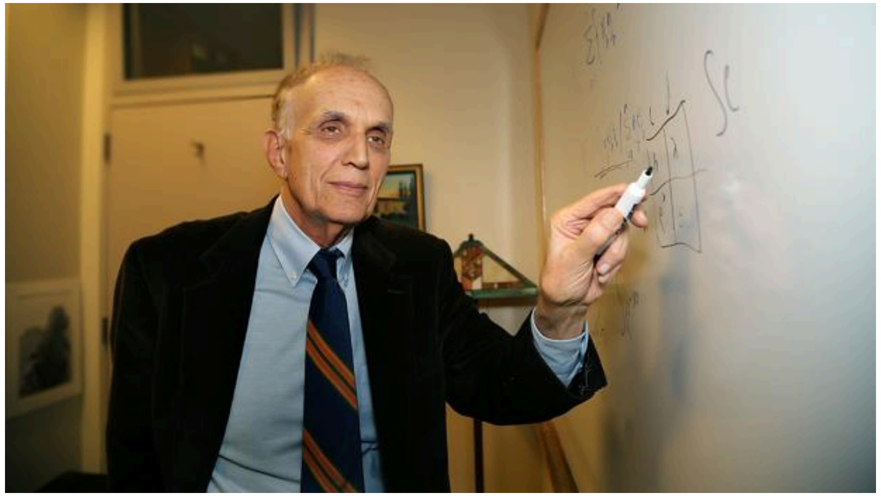
Emeritus Professor of Statistics at Università degli Studi di Padova

Made with Flourish • Create interactive content

Exam

- The exam is made of **two parts**:
- (20 / 30) **Written examination**: a pen-and-paper exam about the theoretical aspects of the course.
- (10 / 30) **Individual assignment**: a data challenge.
  - You will be given a prediction task, and you will need to submit your **predictions** and produce a **report** of maximum 4 pages;
  - You will make use of the Kaggle platform;
  - Further info will be provided in due time.
- Both parts are **mandatory**, and you need to submit the assignment **before** attempting the written part. The report expires after one year from the end of the course.
- The final grade is obtained as the **sum** of the above scores.

Epilogue



“*Those who ignore statistics are condemned to reinvent it.*”

Bradley Efron, Stanford University.

References

- **Main references**
  - Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, **16** (3), 199–215.
  - Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, **26**, 745-766
- **Specialized references**
  - Efron, B. (2020). Prediction, Estimation, and Attribution. *Journal of the American Statistical Association*, **115** (530), 636–55.
  - Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, **33**, 1–67.