# ENVS615

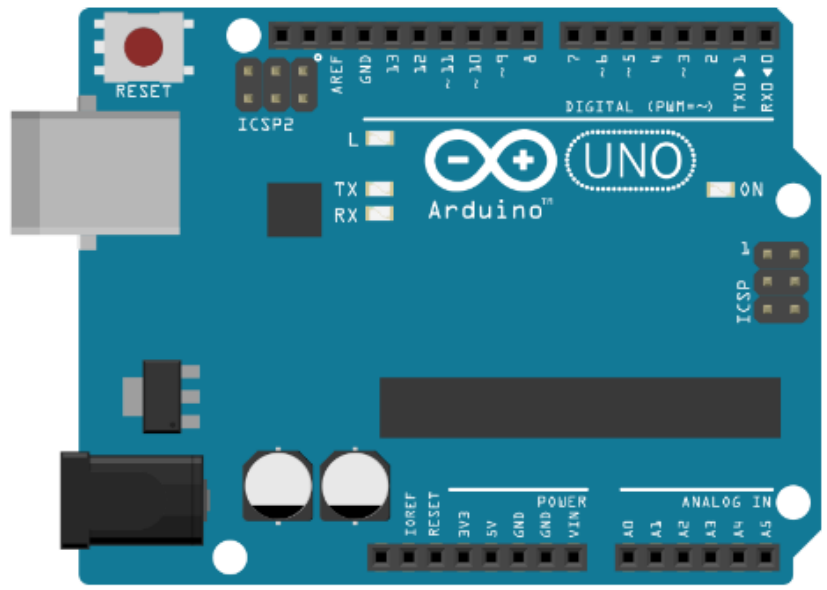
Dani Arribas-Bel

- *Why* Data Science?
- *What* is Data Science?
- This course

# *Why?*

# Data, data, data

OPEN DATA


RESET
ICSP2
AREF
GND
13
12
~11
~10
~9
8
7
~6
~5
4
~3
2
TX→1
RX←0
DIGITAL (PWM=~)
L
TX
RX
ON
UNO
Arduino™
ICSP
IOREF
RESET
3V3
5V
GND
GND
VIN
POWER
A0
A1
A2
A3
A4
A5
ANALOG IN


AM
12:46
LG

# *Accidental* data (Arribas-Bel, 2014)

- Availability is a *side effect*
- *More* data, *more* often, of *more* places
- Different (shape, form, nature, quality…)

# Lazer & Radford (2017)

- Digital life ("the *fraction of life [that is] intrinsically digitally mediated*")
- Digital traces ("*records that chronicle actions taken*")
- Digitalized life ("*nonintrinsically digital life [...] in digital form*")

# But…

Data, in itself, is not very useful… *insights* are

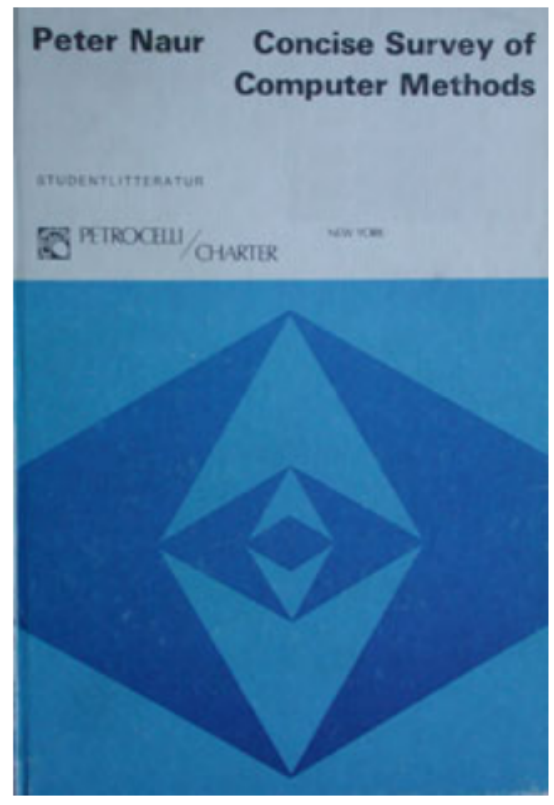
Data Science

# *What?*

# Peter Naur: Concise Survey of Computer Methods, 397 p.

Studentlitteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974

ISBN/Petrocelli 0-88405-314-8, 1975




## Part 1 - Basic Concepts, Tools and Methods

**1. Data and their Applications** (Reprinted as section 1.2 of Peter Naur: Computing, a Human Activity, 1992, ACM Press, ISBN 0-201-58069-1)

1.1. Data and What They Represent; 1.2. Data Processes and Models; 1.3. Data Recognition and Context; 1.4. Data Representations; 1.5. Numbers and Numerals; 1.6. Data Conversions; 1.7. Data Processing; 1.8. A Basic Principle of Data Science; 1.9. Limitations of Data Processing; 1.10. Summary of the Chapter; 1.11. Exercises.

# Tukey (1962)

*Four major influences act on data analysis today:*

1. *The formal theories of statistics*
2. *Accelerating developments in computers and display devices*
3. *The challenge, in many fields, of more and ever larger bodies of data*
4. *The emphasis on quantification in an ever wider variety of disciplines*

Harvard Business Review


DATA

# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 745–766
https://doi.org/10.1080/10618600.2017.1384734

Taylor & Francis
Taylor & Francis Group



# 50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

**ABSTRACT**

More than 50 years ago, John Tukey called for a reformation of academic statistics. In "The Future of Data Analysis," he pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or "data analysis." Ten to 20 years ago, John Chambers, Jeff Wu, Bill Cleveland, and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland and Wu even suggested the catchy name "data science" for this envisioned field. A recent and growing phenomenon has been the emergence of "data science" programs at major universities, including UC Berkeley, NYU, MIT, and most prominently, the University of Michigan, which in September 2015 announced a $100M "Data Science Initiative" that aims to hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; yet many academic statisticians perceive the new programs as "cultural appropriation." This article reviews some ingredients of the current "data science moment," including recent commentary about data science in the popular media, and about how/whether data science is really different from statistics. The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for "scaling up" to "big data." This chosen superset is motivated by commercial rather than intellectual developments. Choosing in this way is likely to miss out on the really important intellectual event of the next 50 years. Because all of science itself will soon become data that can be mined, the imminent revolution in data science is not about mere "scaling up," but instead the emergence of scientific studies of data analysis science-wide. In the future, we will be able to predict how a proposal to change data analysis workflows would impact the validity of data analysis across all of science, even predicting the impacts field-by-field. Drawing on work by Tukey, Cleveland, Chambers, and Breiman, I present a vision of data science based on the activities of people who are "learning from data," and I describe an academic field dedicated to improving that activity in an evidence-based manner. This new field is a better academic enlargement of statistics and machine learning than today's data science initiatives, while being able to accommodate the same short-term goals. *Based on a presentation at the Tukey Centennial Workshop, Princeton, NJ, September 18, 2015.*

**ARTICLE HISTORY**
Received August 2017
Revised August 2017

**KEYWORDS**
Cross-study analysis; Data analysis; Data science; Meta analysis; Predictive modeling; Quantitative programming environments; Statistics

# Greater Data Science (Donoho, 2017)

- `GDS1`: Data Gathering, Preparation, and Exploration
- `GDS2`: Data Representation and Transformation
- `GDS3`: Computing with Data
- `GDS4`: Data Visualisation and Presentation
- `GDS5`: Data Modeling
- `GDS6`: Science about Data Science

# This course

# Day 1

- Computational environment **[GDS3]**
- Data manipulation **[GDS1 + GDS2]**
  - Gathering data
  - Reading/writing data
  - "Wrangling" patterns (query, join, groupby)

# Day 2

- Exploring data visually **[GDS4]**
- Unsupervised learning: clustering **[GDS5]**

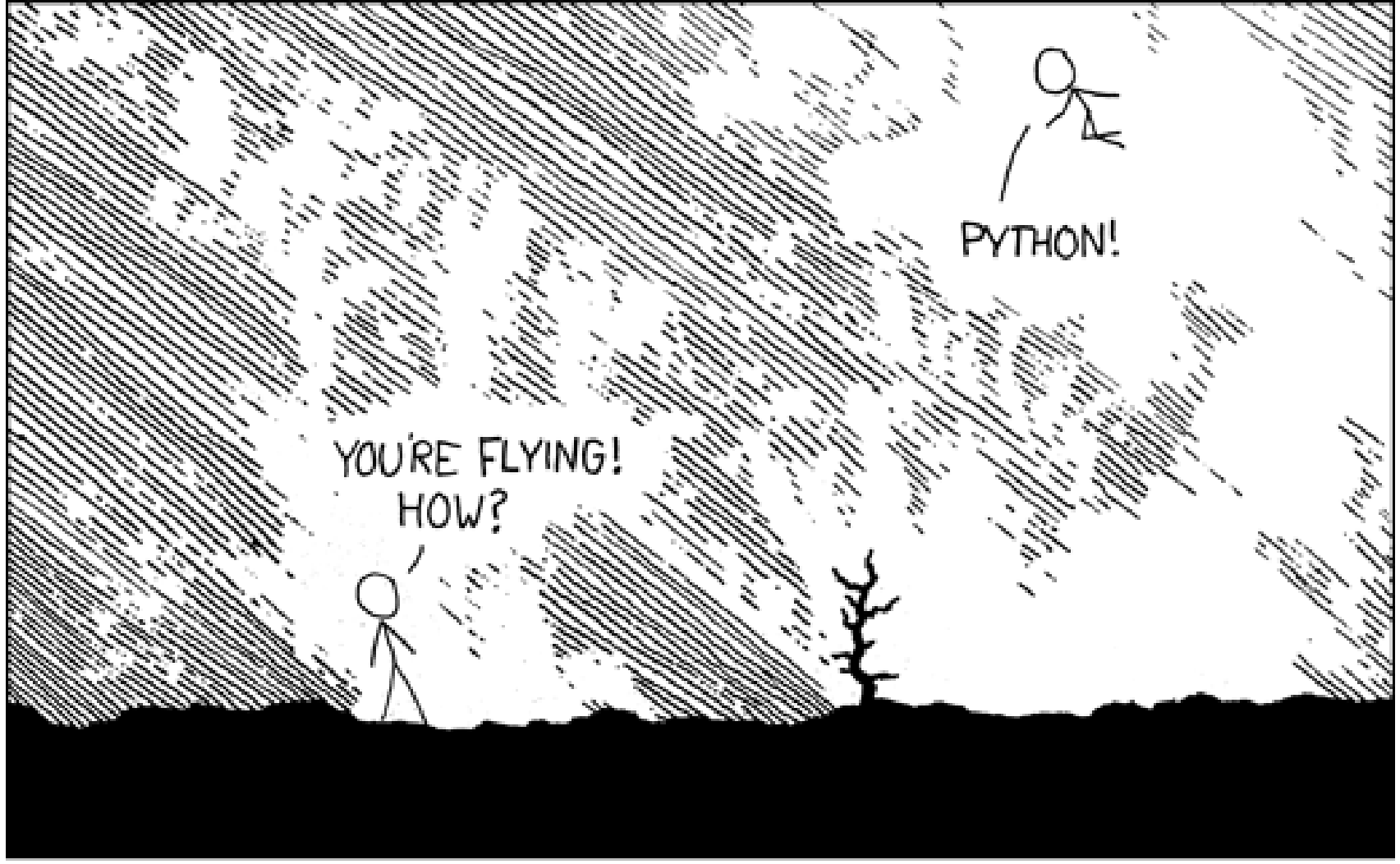
# Day 3

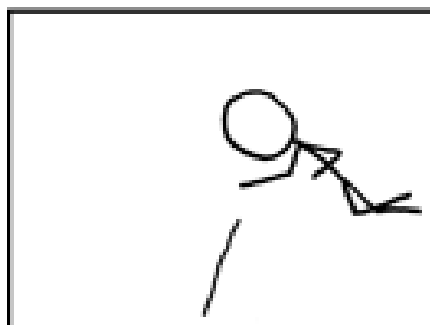
- Modeling: the "*two cultures*" **[GDS5]**
- Predictive checking **[GDS5]**
- Overfitting & Cross-validation **[GDS5]**
- Model predictive performance **[GDS5]**
- More advanced methods **[GDS5]**
- Discuss datasets for assignment

# Day 4

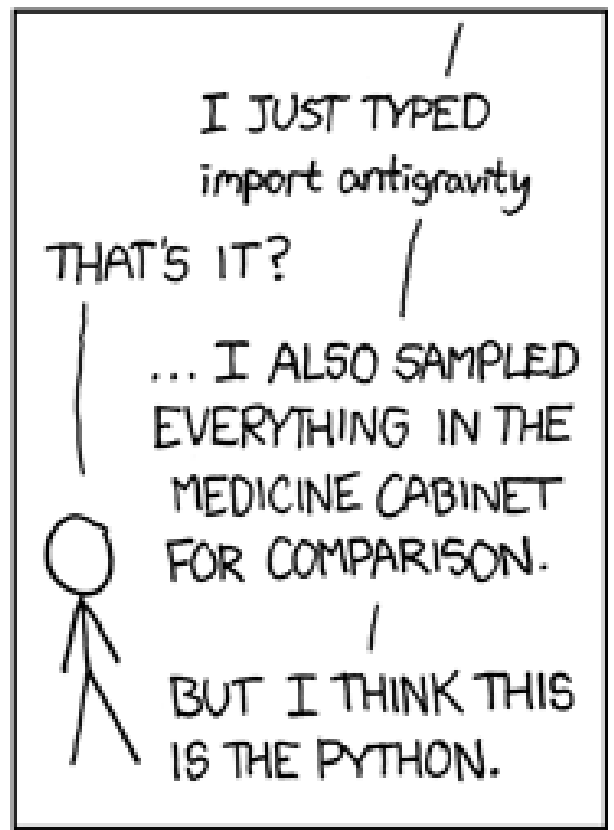
- Pick your favorite: Geo Vs API
- Data Studio
  - Use **your own** data
  - Explore the skills learnt over the course
  - 1-on-1
  - Work on assignment

# Python


PYTHON!
YOU'RE FLYING!
HOW?


I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"


I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?


I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

# Data

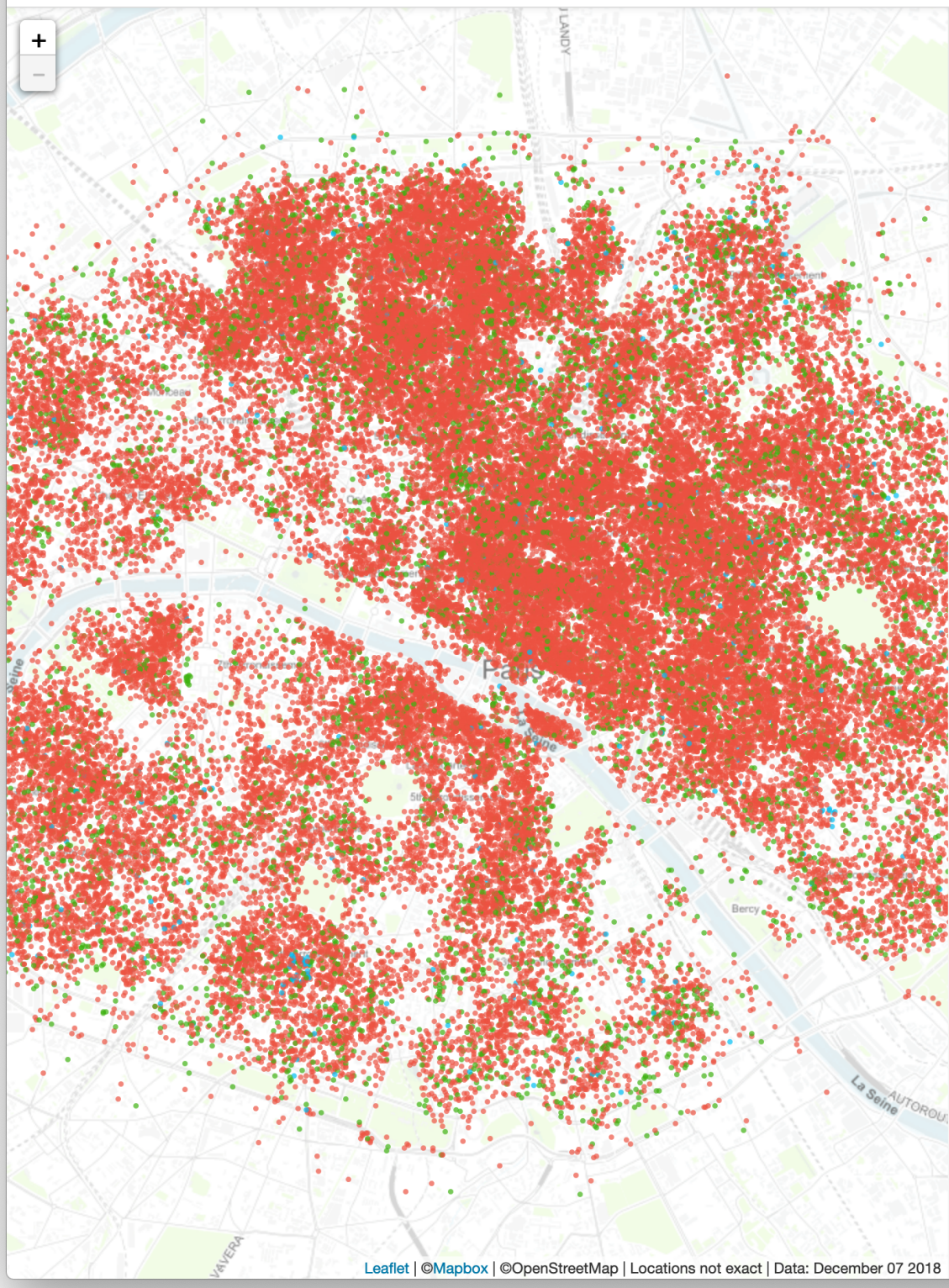
Inside Airbnb: Paris. Adding Dat

insideairbnb.com/paris/

# Inside Airbnb

Adding data to the debate

About Behind Get the Data Tweet




## Paris

Filter by: Paris

**##,###**
out of **59,881** listings (##.#%)

About Airbnb in Paris

### How is Airbnb really being used in and affecting your neighbourhoods?

## Room Type

☐ Only entire homes/apartments

Airbnb hosts can list entire homes/apartments, private or shared rooms.

Depending on the room type, availability, and activity, an airbnb listing could be more like a hotel, disruptive for neighbours, taking away housing, and illegal.

**##.#%**
entire homes/apartments

**€###**
price/night

**##,### (##.#%)**
entire home/apartments

**#,### (##.#%)**
private rooms

**#,### (##.#%)**
shared rooms

## Activity

☐ Only **recent** and **frequently** booked

Airbnb guests may leave a review after their stay, and these can be used as an indicator of airbnb activity (although not all guests leave a review, so the actual booking activity would be **much** higher).

The minimum stay, price and number of reviews have been used to estimate a **minimum income per month** for each listing.

How does the income from running a defacto hotel (an Airbnb room or apartment) compare to a long-term lease?

And more importantly, what is renting to a tourist rather than a resident doing to our neighbourhoods?

**###**
estimated nights/year

**##.#**
reviews/listing/month

**###,###**
reviews

**€###**
price/night

**##.#%**
estimated occupancy

**€###**
estimated income/month

# Assessment

# Assessment

## Computational essay

- Type: `Coursework`
- [Equivalent to 5,000 words] Up to five figures and three tables + code + comments + up to 2,000
- Chance to be reassessed
- Due on **March 9th at 2pm**
- Electronic submission only. Static HTML with NO interactive cells