

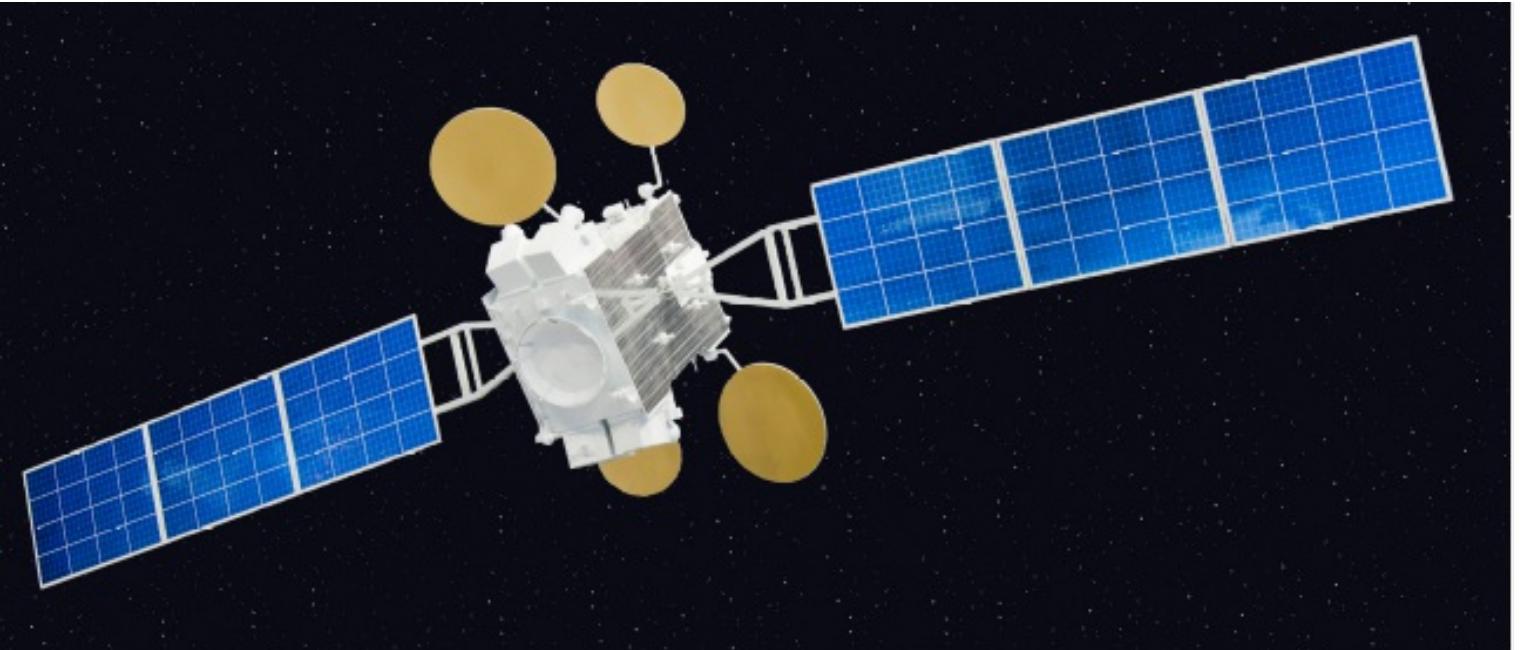
Geographic Data Science Studio

Dani Arribas-Bel

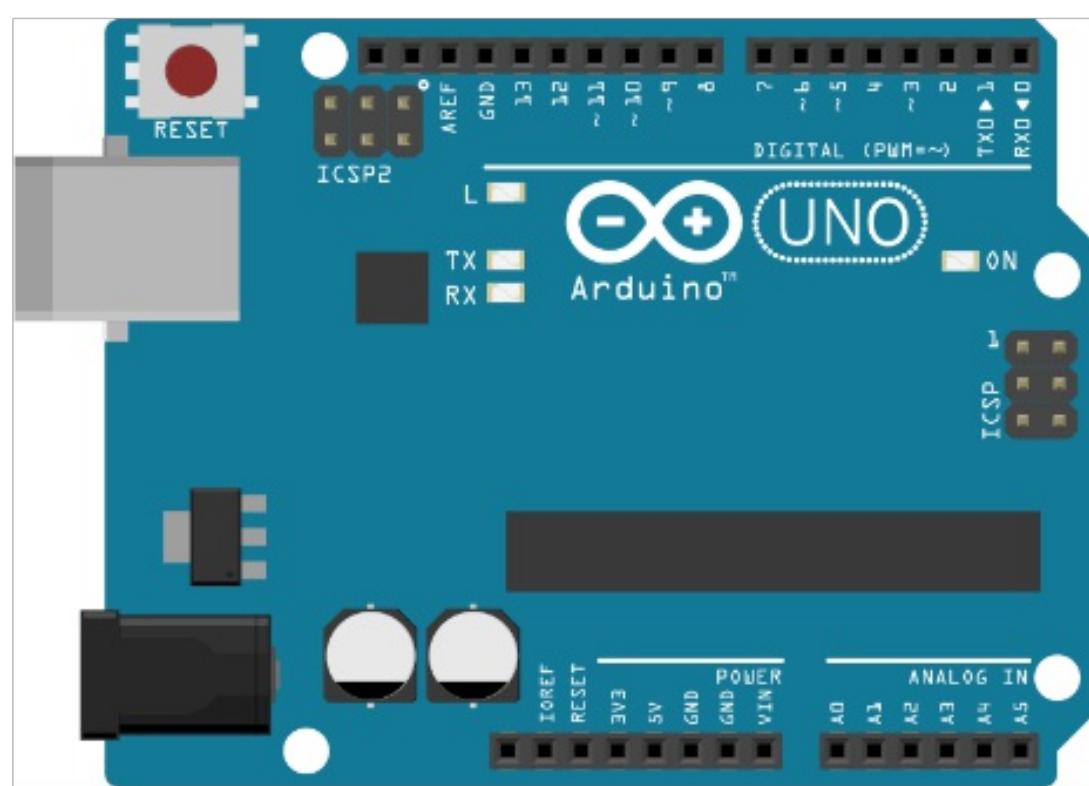
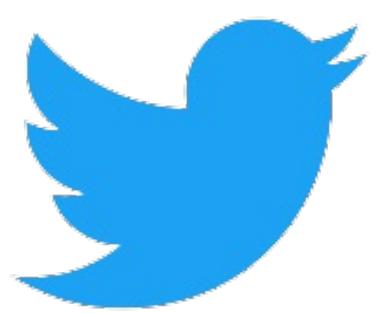
- *Why* Data Science?
- *What* is Data Science?
- This course

Why?

Data, data, data



OPEN DATA



946644442424443444545432243364355525743464453232123246423334344554434333246423212323544643475255534633
94466712564662651353263252502355254434422343223223465245754543464554643454575425643222322343223443224453
945323664356534454323554335442234322322346524575454346455464345457542564322232234322344322445324354522
8552334374115527424325423247154455664335156120466022245351141753664568355226353332642254245324354522
935677213315546434243465445565443345445642344353453134554434674224764344554313543534432465445433445655
2166554462431377236123115341373565852242445804224314141142334674672663335636324442562443545535323
12323433433546755325554335215546422333236433454454467644324413314423446764454454334632333224645512
7462346762676474432435377622452152223350425442226543425432261222434346263134424325325436441356325425
4232442463224242235466453222321322335435444322585343323323445544332332334358522344534533223123222
432234431445525212213366525545435633444332454333343335446521333312564453333433334542334443365345455
2443444235227421324431535435164661436123447334226316311431712446244243403521236753224547445244475544
454432313554566312124675333323554343334432477745446676311232233223211367664454777423443334345532333
943443124421263247554212211322432211335657413544542445665334322234335665442454453147565331122342231
5676524426344232366243345443436323305424337436647252473523243345355245455325233526765133233222334
3332332544644534256654146544344533310123314664312344331443345566554334413344921346641332101333544344
241234133223645344213433441232236762445224364445054345352273243736332413641246635332433421424421227
1146645513213535732343534632321556545564224313423642343123542344324532134324632431342246554565512323
7735443315566346268414535253341725766068135423765326411611333233666422258526163453253341322435534344
3225346422222575335641343466434565533663134323553675221146542355324564112257635532343136633556543466
274366367763214557553345645586215656567540123323333467513673732344646357532224324663332322457531533
246663257554423547654234642344674563146441422332243123232464554455464232321342233224144641365476443
1526322156433537552366314620164552252233121463626253227331463114464733110534576353534511102272564321
1243344243554332223333443345434433544432433233565367763355346533335643553367163565332334234453344344
7422135743135351317544385345762425333413661550305254766514855532312632241545565845534364121422531641
4265353335325423345454644442114553234344233543323323553345444444543355323323345332443432355411241
1523133237227215245242546554431570367431561733377617034585053512353011334584673334166343322133144122
4442253323535444365136554564421224544244654435621675313554244035530442455313576126534456442445422124
13334134133434142542775443122344123640225114656307736631373633254764433546335277621113873752113310516
6245353113334354345555543434354454442043123323654334564353323423324323353465433456323321340244454453
2414531123127565256341545757504634351331223556476033255182632548363563337236167723345232376652227232
95522534654643225623666524245346413444364434665542255454542443435534344245454552245566434463444314643
2553246544026522558643533756563115454652318133211524761442547661462003344533463221222357663124215414
2564556411343112463543444456742573123345543333344443345665324515445154235665493444433334554332137524
511231532266541316426164152372484027522556222544415637653623572461532586573566657462232231026322623
45543452452113225534433336664343234344335635544543333465343463122136434356433334544553653344323434
5313636322354013243264654142231203574657117563256244234336332453435525653223444656437457443034332422

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...

But...

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

But...

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

Data Science

What?



→



www.naur.com/Conc.Surv.html



...



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

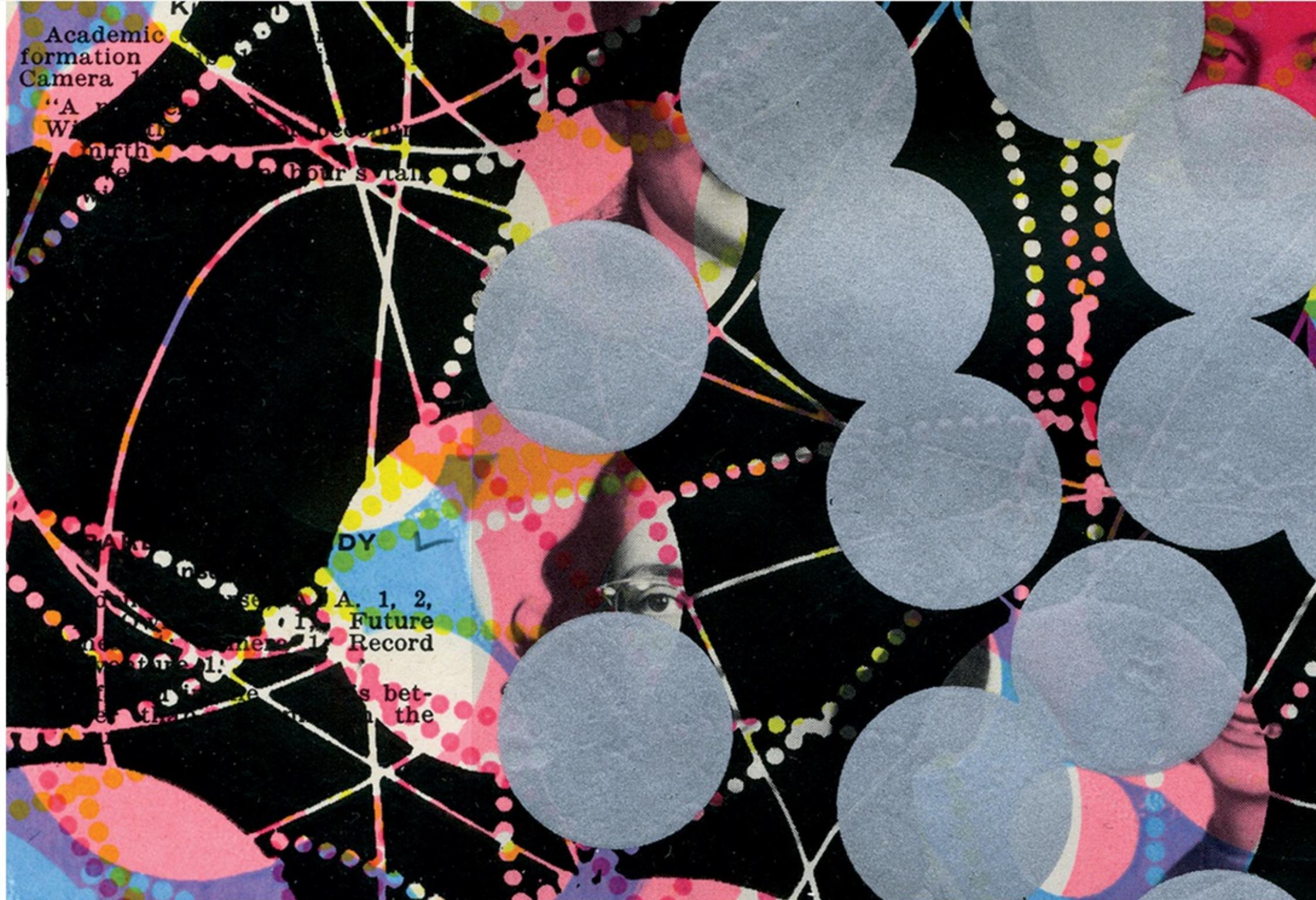
Four major influences act on data analysis today:

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*

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*



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

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 wrangling **[GDS1 + GDS2]**
 - Gathering data
 - Reading/writing data
 - Types of data
 - Wrangling patterns

Day 2

- Working with APIs [**GDS1**]
- Exploring data visually [**GDS4**]
- Unsupervised learning: clustering [**GDS5**]

Day 3

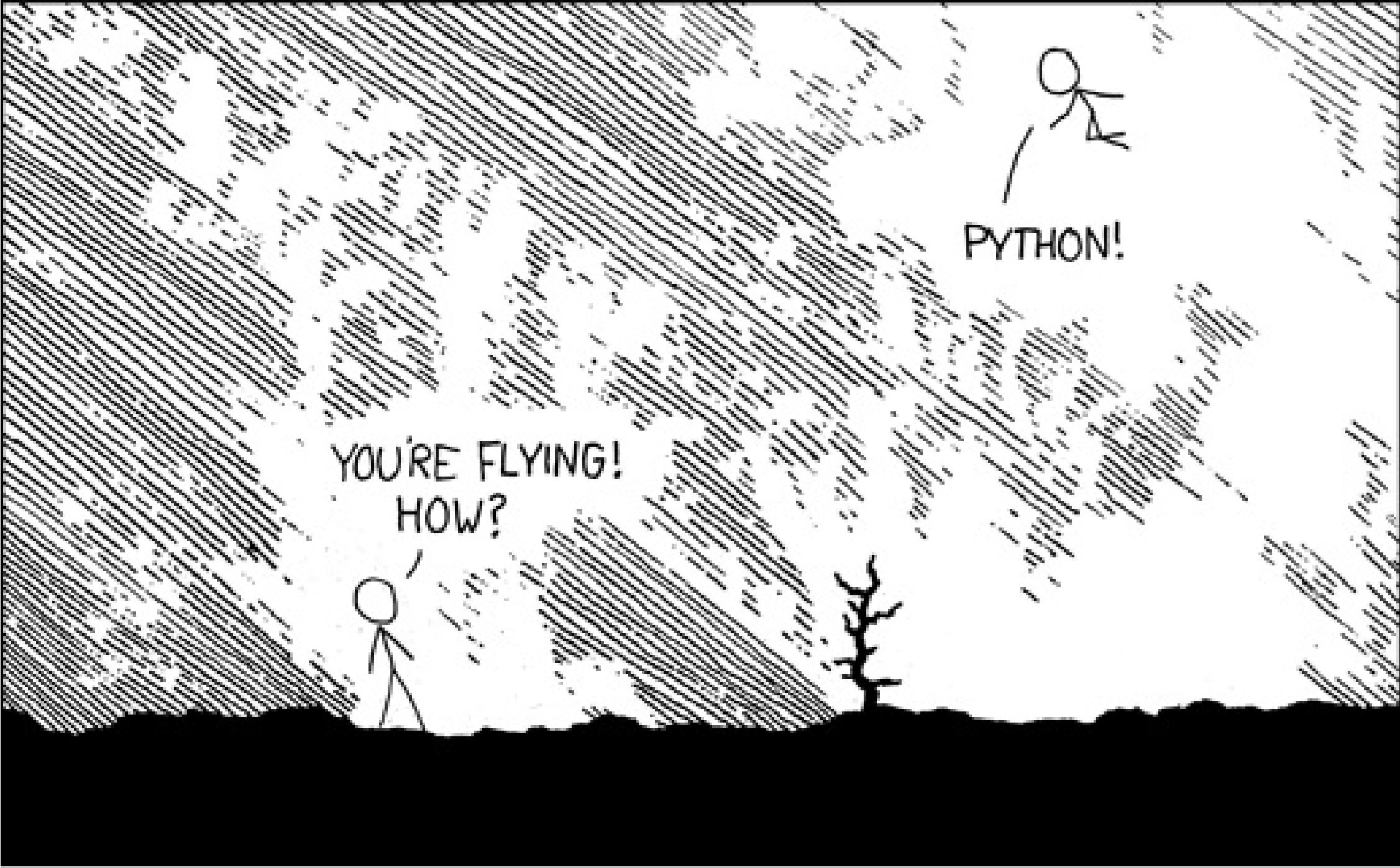
- Modeling: the "*two cultures*" [**GDS5**]
- Predictive checking [**GDS5**]
- Overfitting & Cross-validation [**GDS5**]
- Model predictive performance [**GDS5**]
- More advanced methods [**GDS5**]
- Discuss datasets for assignment

Day 4

Data Studio

- Use your own data
- Explore the skills learnt over the course
- 1-on-1
- Work on assignment

Python



I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
I
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!

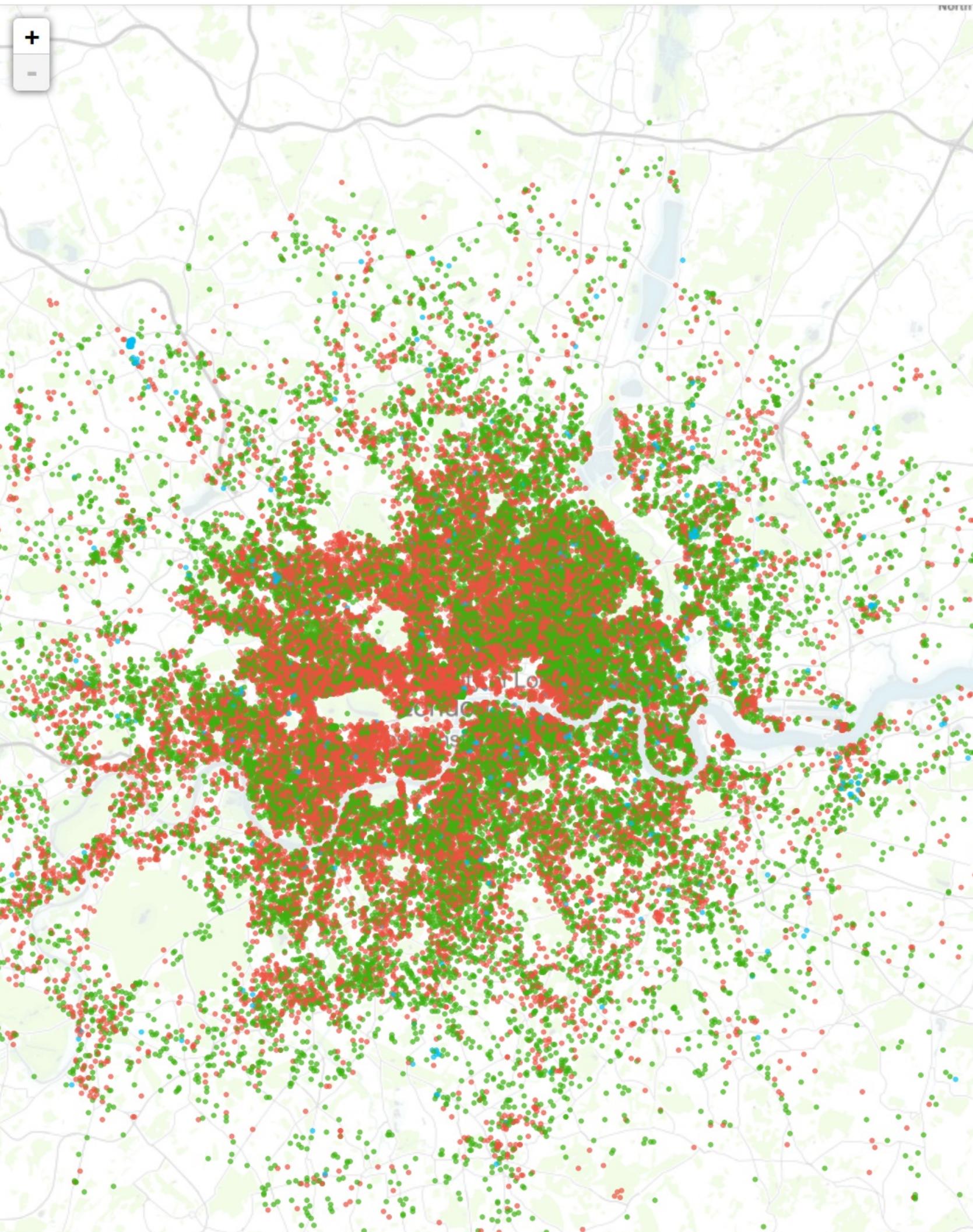
BUT HOW ARE
YOU FLYING?

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



London

Filter by:

London

49,348

out of 49,348 listings (100%)

[About Airbnb in London](#)

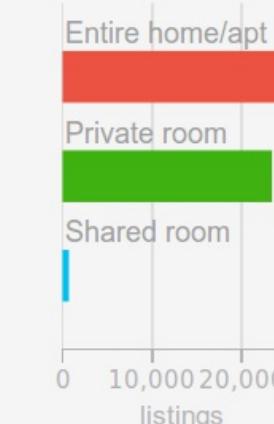
How is Airbnb really being used in and affecting your neighbourhoods?

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](#).



51.2%

entire homes/apartments

£98

price/night

25,285 (51.2%)

entire home/apartments

23,357 (47.3%)

private rooms

706 (1.4%)

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.

The minimum stay, price and number of reviews have been used to estimate the [occupancy rate](#), the number of [nights per year](#) and the [income per month](#) for each listing.

How does the income from Airbnb compare to a long-term lease?

Do the number of nights booked per year make it impossible for a listing to be used for residential housing?

And what is renting to a tourist full-time rather than a resident doing to our neighbourhoods and cities?

89

estimated nights/year

1

reviews/listing/month

564,297

reviews

£98

price/night

24.5%

estimated occupancy

£661

estimated income/month

Assessment

Assessment (pt. I)

Computational essay

- Find a dataset of your interest
[Discuss with me throughout Day 1-3, confirm on D3 - S3]
- Prepare it for analysis
- Explore the dataset visually, identifying the key patterns

Assessment (pt. II)

Computational essay

- Perform a **clustering** exercise & analyse results
- Fit a **regression** model and:
 - Interpret the coefficients
 - Evaluate its predictive performance both with and without cross-validation
 - Reflect on the differences between the two approaches and the reasons for the potential divergences
- Consider potential **predictive improvements** of the model through alternative techniques

Assessment (pt. III)

Computational essay

- HTML of Jupyter notebook
- Submit through Turnitin
- Fully document your code
- Max. 2,000 words (Code, comments)
- Deadline: May 15th