SDS Base Camp

History, Pandas Preview & Data Ethics

Samantha Breslin
Sept 8, 2021

KØBENHAVNS UNIVERSITET

Overview for Today

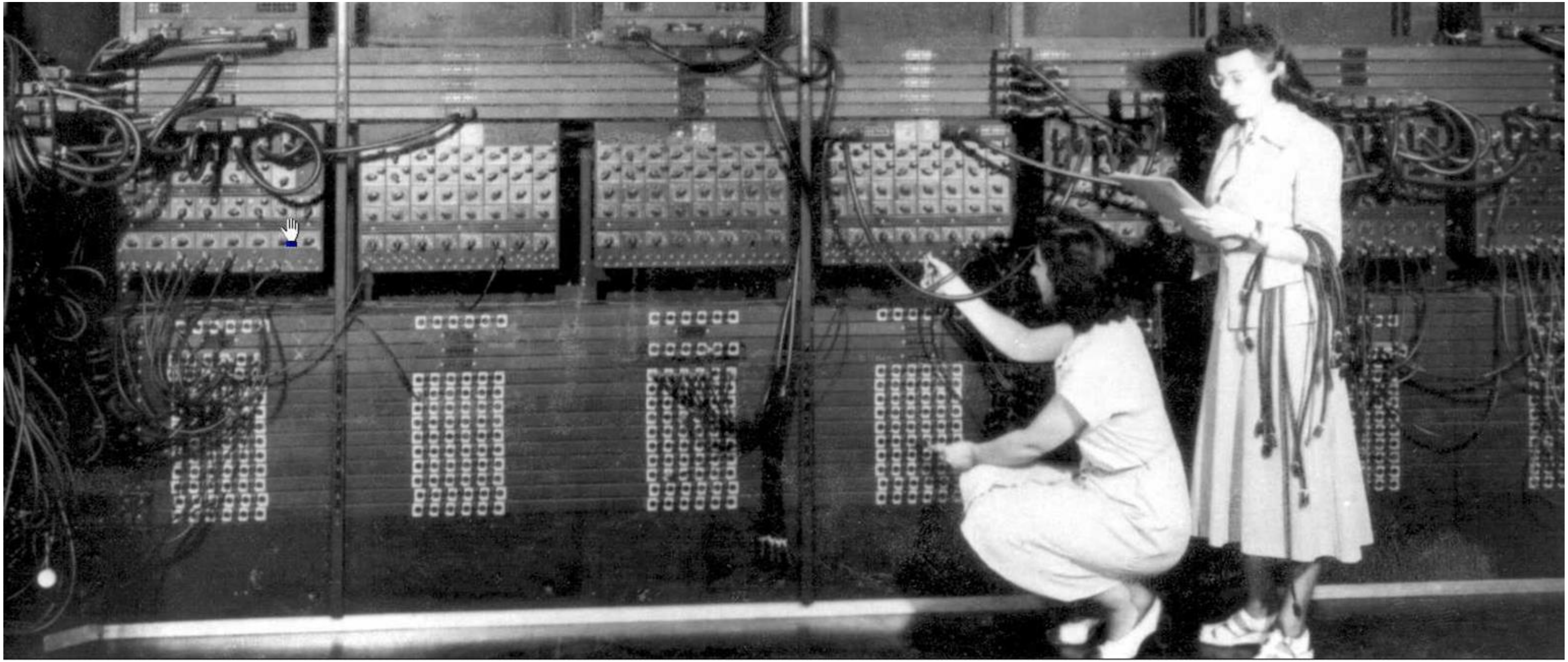
- A brief history of programming languages
- Introducing pandas and the SODAS DataFrame
- Introduction to ethics in Social Data Science

A *Very Brief* History of Programming Languages

Contextualizing programming

- For our methods
- For why programming languages work in certain ways
  - Products of historical and social circumstances
- For where you fit within networks of people (programmers, data scientists, etc.) and things (programming language communities, etc.)
- To help you learn programming in productive & reflexive ways

ENIAC - wiring a program



http://www.columbia.edu/cu/computinghistory/eniac.html

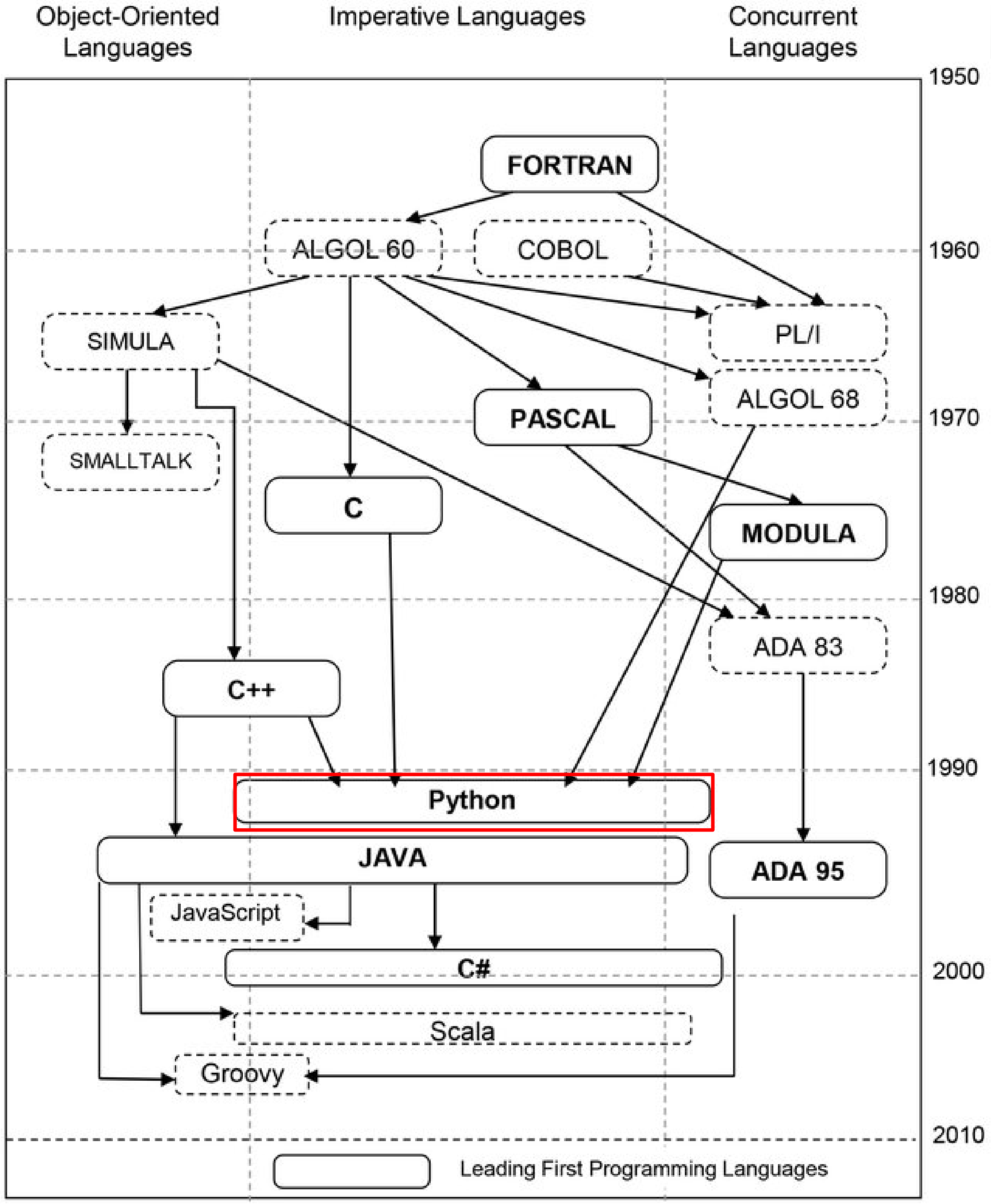
https://youtu.be/bGk9W65vXNA?t=101

What is a programming language?

- A computer operates using binary code (i.e. machine language)
  - Sequences of 1s and 0s that have (are given) specific meanings
- Programming language: a language designed to help humans program computers
  - To make programming easier & more readable for people
  - Must still be translatable into machine language

Python: Hello World!

- print(“Hello World!”)
  - (Relatively) easily readable
  - Based on prior languages, e.g. Fortran, C

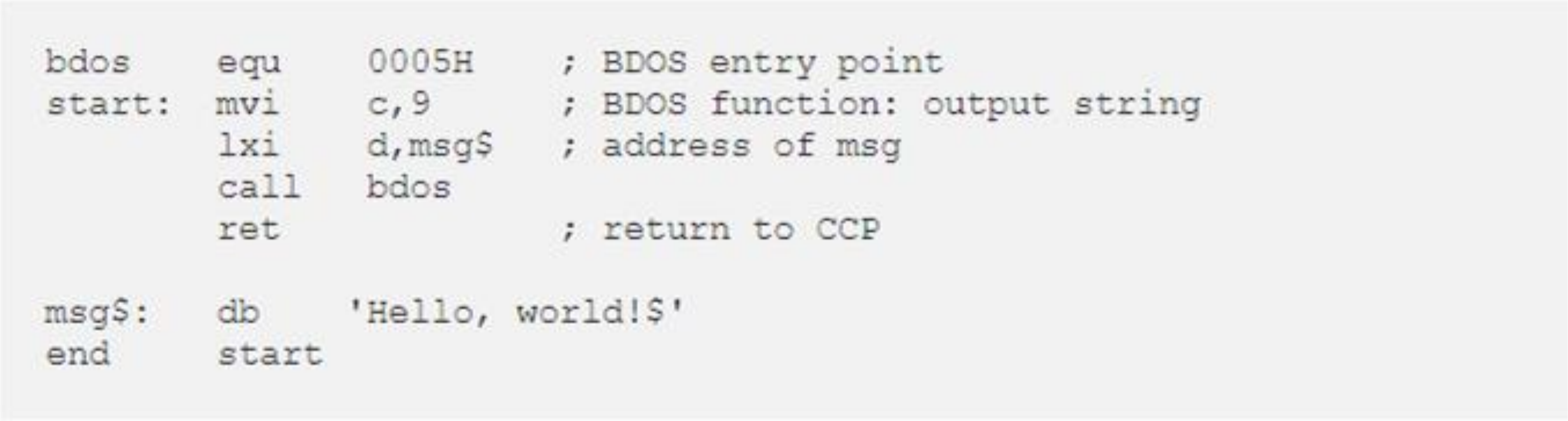


Hello World in binary (base 16)

```
00000000 7f 45 4c 46 01 01 01 00 00 00 00 00 00 00 00 00 |.ELF............|
00000010 02 00 03 00 01 00 00 00 80 80 04 08 34 00 00 00 |............4...|
00000020 c8 00 00 00 00 00 00 00 34 00 20 00 02 00 28 00 |........4. ...(.|
00000030 04 00 03 00 01 00 00 00 00 00 00 00 00 80 04 08 |................|
00000040 00 80 04 08 9d 00 00 00 9d 00 00 00 05 00 00 00 |................|
00000050 00 10 00 00 01 00 00 00 a0 00 00 00 a0 90 04 08 |................|
00000060 a0 90 04 08 0e 00 00 00 0e 00 00 00 06 00 00 00 |................|
00000070 00 10 00 00 00 00 00 00 00 00 00 00 00 00 00 00 |................|
00000080 ba 0e 00 00 00 b9 a0 90 04 08 bb 01 00 00 00 b8 |................|
00000090 04 00 00 00 cd 80 b8 01 00 00 00 cd 80 00 00 00 |................|
000000a0 48 65 6c 6c 6f 2c 20 77 6f 72 6c 64 21 0a 00 2e |Hello, world!...|
000000b0 73 68 73 74 72 74 61 62 00 2e 74 65 78 74 00 2e |shstrtab..text..|
000000c0 64 61 74 61 00 00 00 00 00 00 00 00 00 00 00 00 |data............|
000000d0 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 |................|  *
000000f0 0b 00 00 00 01 00 00 00 06 00 00 00 80 80 04 08 |................|
00000100 80 00 00 00 1d 00 00 00 00 00 00 00 00 00 00 00 |................|
00000110 10 00 00 00 00 00 00 00 11 00 00 00 01 00 00 00 |................|
00000120 03 00 00 00 a0 90 04 08 a0 00 00 00 0e 00 00 00 |................|
00000130 00 00 00 00 00 00 00 00 04 00 00 00 00 00 00 00 |................|
00000140 01 00 00 00 03 00 00 00 00 00 00 00 00 00 00 00 |................|
00000150 ae 00 00 00 17 00 00 00 00 00 00 00 00 00 00 00 |................|
00000160 01 00 00 00 00 00 00 00 |........|
```

Assembly

- Used starting in the 1940s



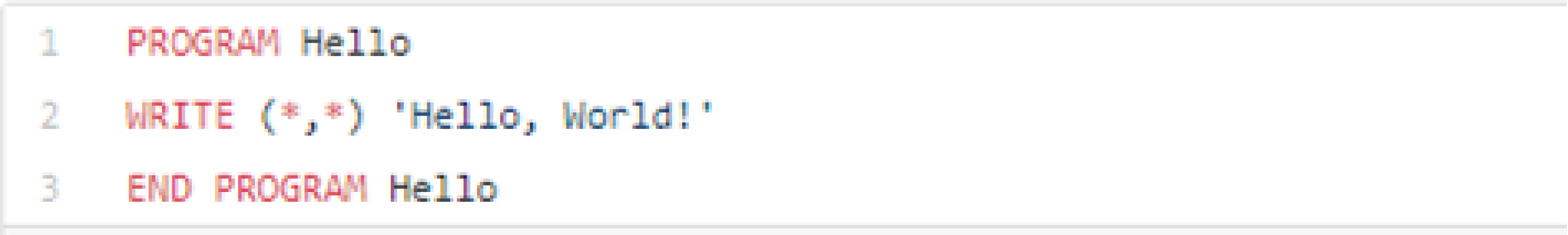
```
bdos    equ    0005H    ; BDOS entry point
start:  mvi    c,9      ; BDOS function: output string
        lxi    d,msg$   ; address of msg
        call   bdos
        ret             ; return to CCP

msg$:   db    'Hello, world!$'
end     start
```

https://medium.com/javarevisited/70-years-of-hello-world-with-50-programming-languages-2400de893a97

FORTRAN

- Language for mathematical and scientific computing
- Other functions (i.e. working with text data) more difficult



```
PROGRAM Hello
WRITE (*,*) 'Hello, World!'
END PROGRAM Hello
```

https://medium.com/javarevisited/70-years-of-hello-world-with-50-programming-languages-2400de893a97

Languages over time

- Programming languages as historical, material, social, and culturally constructs
- Influenced by:
  - Hardware
  - Companies
  - Goals for the language
  - Previous languages
  - Users

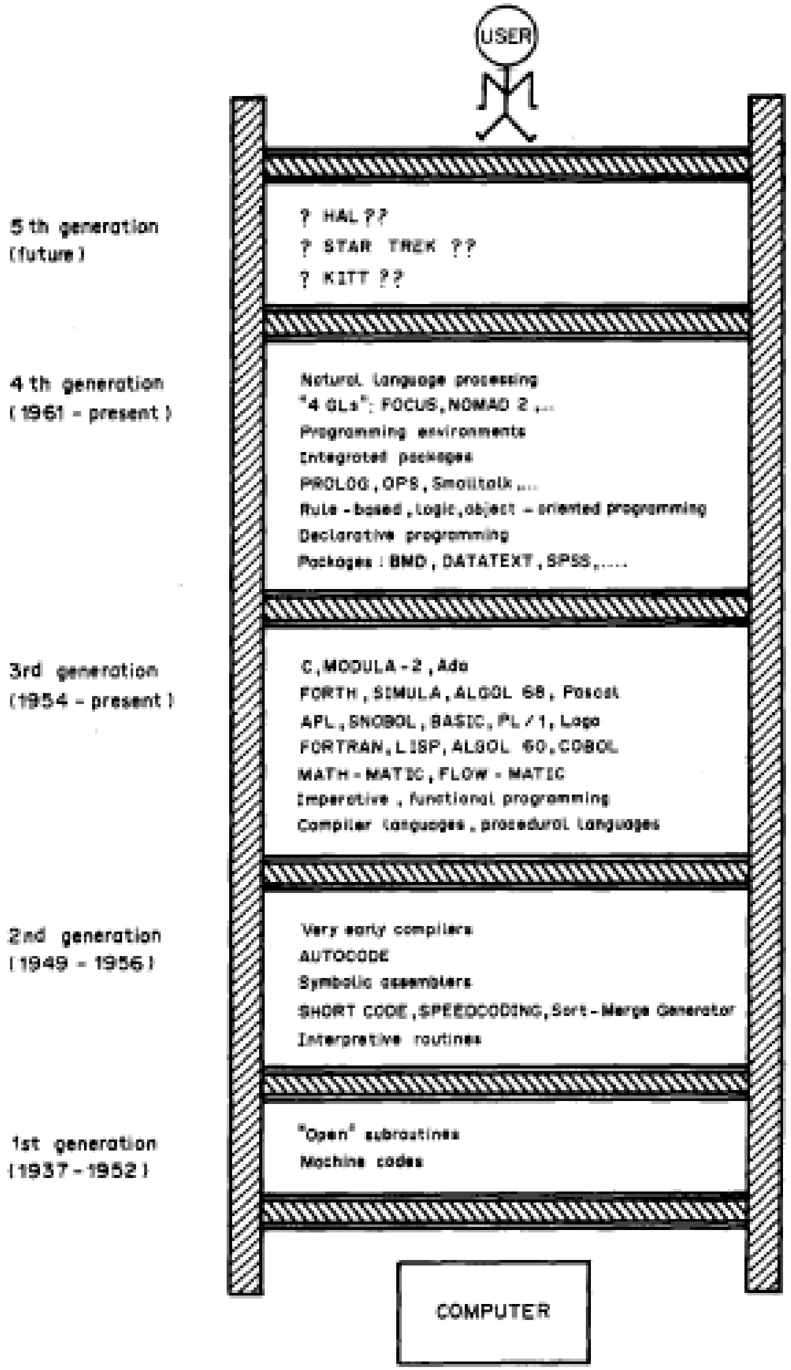

USER
5 th generation
(future)
? HAL ??
? STAR TREK ??
? KITT ??
4 th generation
(1961 - present)
Natural language processing
"4 GLs": FOCUS, NOMAD 2 ,...
Programming environments
Integrated packages
PROLOG, OPS, Smalltalk ...
Rule - based, logic, object - oriented programming
Declarative programming
Packages : BMD, DATATEXT, SPSS,....
3rd generation
(1954 - present)
C, MODULA - 2, Ada
FORTH, SIMULA, ALGOL 68, Pascal
APL, SNOBOL, BASIC, PL / 1, Logo
FORTRAN, LISP, ALGOL 60, COBOL
MATH - MATIC, FLOW - MATIC
Imperative , functional programming
Compiler languages , procedural languages
2nd generation
(1949 - 1956)
Very early compilers
AUTOCODE
Symbolic assemblers
SHORT CODE, SPEEDCODING, Sort - Merge Generator
Interpretive routines
1st generation
(1937 - 1952)
"Open" subroutines
Machine codes
COMPUTER

Fig. 1. The software ladder.

(Friedman 1992)

Chef (Programming Language)

```
Hello World Souffle.
This recipe prints the immortal words "Hello world!",
in a basically brute force way. It also makes a lot of
food for one person.

Ingredients.
72 g haricot beans
101 eggs
108 g lard
111 cups oil
32 zucchinis
119 ml water
114 g red salmon
100 g dijon mustard
33 potatoes

Method.

Put potatoes into the mixing bowl. Put dijon
mustard into the mixing bowl. Put lard into the
mixing bowl. Put red salmon into the mixing bowl.
Put oil into the mixing bowl. Put water into the
mixing bowl. Put zucchinis into the mixing bowl. Put
oil into the mixing bowl. Put lard into the mixing
bowl. Put lard into the mixing bowl. Put eggs into
the mixing bowl. Put haricot beans into the mixing
bowl. Liquefy contents of the mixing bowl. Pour
contents of the mixing bowl into the baking dish.

Serves 1.
```

https://dangermouse.net/esoteric/chef.html

Pandas

Based on Slides from Friedolin Merhout

What is Pandas?

- Module to facilitate work with tabular data
- Geared toward data manipulation and data analysis
- Introduces objects familiar to quantitative social scientist

Ash Warner
@AlsBoy

Don’t you love how literally almost every picture of pandas looks like chaos

7:57 PM · Oct 23, 2018 · Tweetbot for iOS

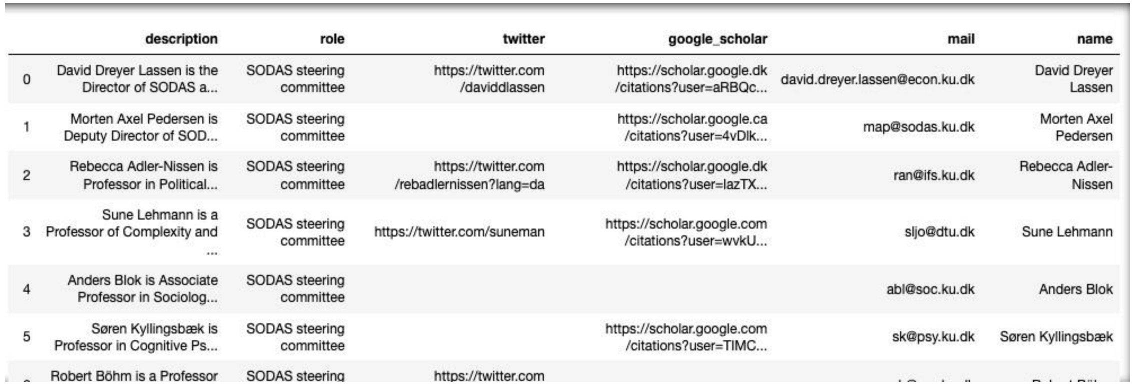
202 Retweets 15 Quote Tweets 497 Likes

Pandas History

- Started by Wes McKinney while working at AQR Capital Management in 2008
  - Created to conduct quantitative analyses on *financial data*
- Turned into Open Source project
  - McKinney is the “Benevolent Dictator for Life”

Why Pandas First?

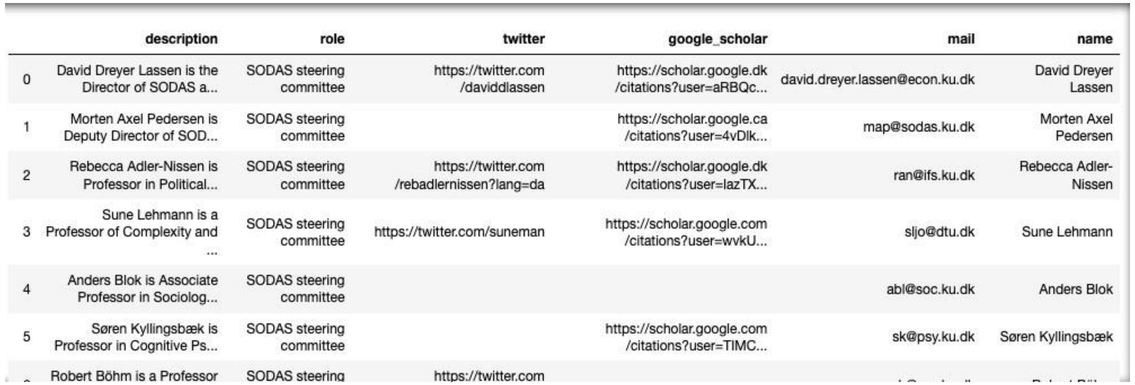
- Preview of where we are going
- Relatable format coming from statistical programming
- Motivating learning of basics



	description	role	twitter	google_scholar	mail	name
0	David Dreyer Lassen is the Director of SODAS a...	SODAS steering committee	https://twitter.com/daviddlassen	https://scholar.google.dk/citations?user=aRBQc...	david.dreyer.lassen@econ.ku.dk	David Dreyer Lassen
1	Morten Axel Pedersen is Deputy Director of SOD...	SODAS steering committee		https://scholar.google.ca/citations?user=4vDlk...	map@sodas.ku.dk	Morten Axel Pedersen
2	Rebecca Adler-Nissen is Professor in Political...	SODAS steering committee	https://twitter.com/rebadlernissen?lang=da	https://scholar.google.dk/citations?user=lazTX...	ran@ifs.ku.dk	Rebecca Adler-Nissen
3	Sune Lehmann is a Professor of Complexity and ...	SODAS steering committee	https://twitter.com/suneman	https://scholar.google.com/citations?user=wvkU...	sljo@dtu.dk	Sune Lehmann
4	Anders Blok is Associate Professor in Sociolog...	SODAS steering committee			abl@soc.ku.dk	Anders Blok
5	Søren Kyllingsbæk is Professor in Cognitive Ps...	SODAS steering committee		https://scholar.google.com/citations?user=TIMC...	sk@psy.ku.dk	Søren Kyllingsbæk
	Robert Böhm is a Professor	SODAS steering	https://twitter.com			

Getting Data I – Downloading

- Base Camp dataset
  - Sources: SODAS website, Twitter, Google Scholar
  - Combines downloading, scraping, and API techniques
- Content:
  - Information on SODAS affiliated individuals, including roles, names, publications, and social media accounts



	description	role	twitter	google_scholar	mail	name
0	David Dreyer Lassen is the Director of SODAS a...	SODAS steering committee	https://twitter.com /daviddlassen	https://scholar.google.dk /citations?user=aRBQc...	david.dreyer.lassen@econ.ku.dk	David Dreyer Lassen
1	Morten Axel Pedersen is Deputy Director of SOD...	SODAS steering committee		https://scholar.google.ca /citations?user=4vDlk...	map@sodas.ku.dk	Morten Axel Pedersen
2	Rebecca Adler-Nissen is Professor in Political...	SODAS steering committee	https://twitter.com /rebadlernissen?lang=da	https://scholar.google.dk /citations?user=lazTX...	ran@ifs.ku.dk	Rebecca Adler-Nissen
3	Sune Lehmann is a Professor of Complexity and ...	SODAS steering committee	https://twitter.com/suneman	https://scholar.google.com /citations?user=wvkU...	sljo@dtu.dk	Sune Lehmann
4	Anders Blok is Associate Professor in Sociolog...	SODAS steering committee			abl@soc.ku.dk	Anders Blok
5	Søren Kyllingsbæk is Professor in Cognitive Ps...	SODAS steering committee		https://scholar.google.com /citations?user=TIMC...	sk@psy.ku.dk	Søren Kyllingsbæk
	Robert Böhm is a Professor	SODAS steering	https://twitter.com			

Getting Data I – Downloading

- Open data
  - Social science research data:
    - ICPSR: https://www.icpsr.umich.edu/icpsrweb/ICPSR/
    - GESIS database: https://search.gesis.org
    - Dataverse, e.g. https://dataverse.harvard.edu/
  - Google Data Search: https://datasetsearch.research.google.com/
  - Data Is Plural: https://tinyurl.com/ydfgkq8u
  - Official statistics:
    - Eurostat: https://ec.europa.eu/eurostat
    - Statistics Denmark: https://www.dst.dk/en/Statistik/statistikbanken
  - Communal and regional data: https://www.opendata.dk/
  - Competition datasets: https://kaggle.com/datasets

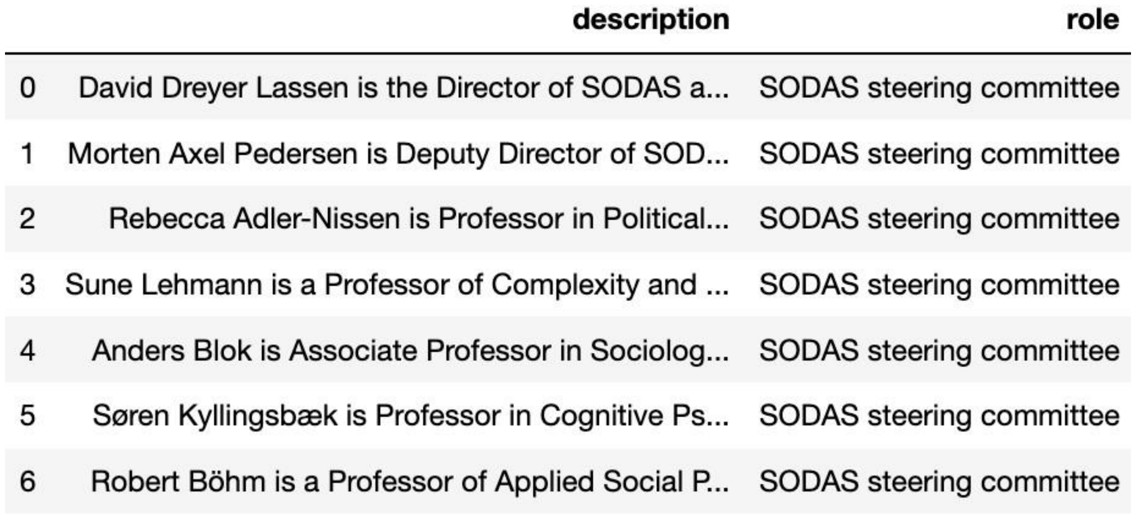
CSV

- "Comma Separated Values"
- Tabular data in text form

Data formatting

- Gist:
  - Varying nature and application of data leads to wide variation in representation and storage formats

- Three main types of formats
  - **Tabular (CSV, XLSX)**
    - Structure: rows and columns
    - Common terminology: rows/observations/cases, columns/vectors/variables



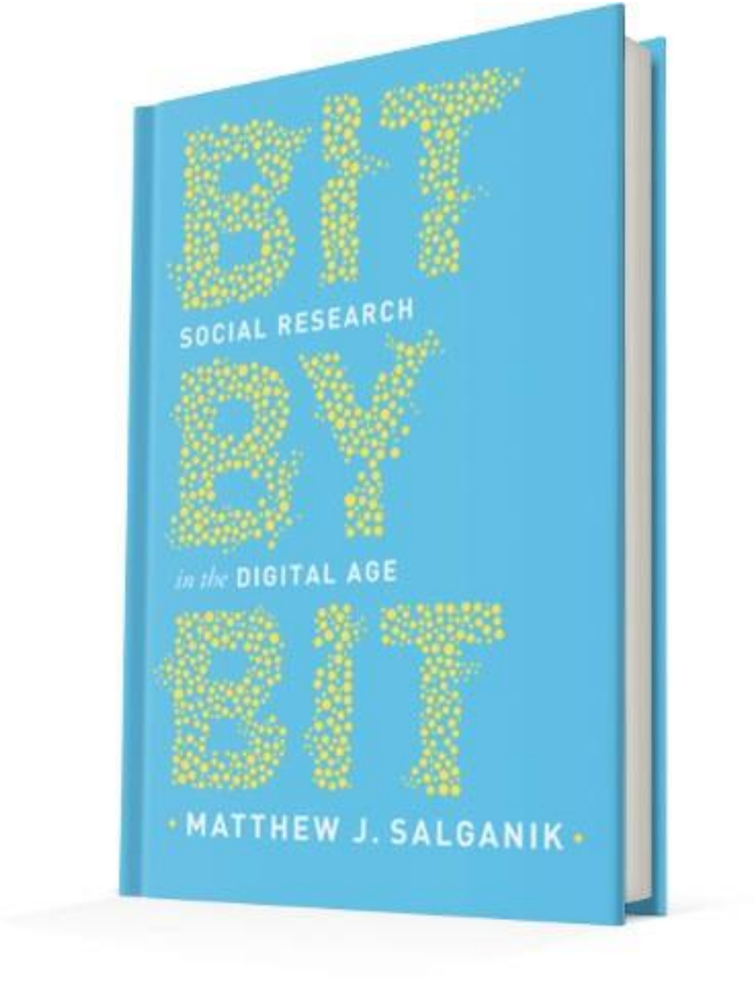
	description	role
0	David Dreyer Lassen is the Director of SODAS a...	SODAS steering committee
1	Morten Axel Pedersen is Deputy Director of SOD...	SODAS steering committee
2	Rebecca Adler-Nissen is Professor in Political...	SODAS steering committee
3	Sune Lehmann is a Professor of Complexity and ...	SODAS steering committee
4	Anders Blok is Associate Professor in Sociolog...	SODAS steering committee
5	Søren Kyllingsbæk is Professor in Cognitive Ps...	SODAS steering committee
6	Robert Böhm is a Professor of Applied Social P...	SODAS steering committee

Ethics

Based on Slides from Friedolin Merhout

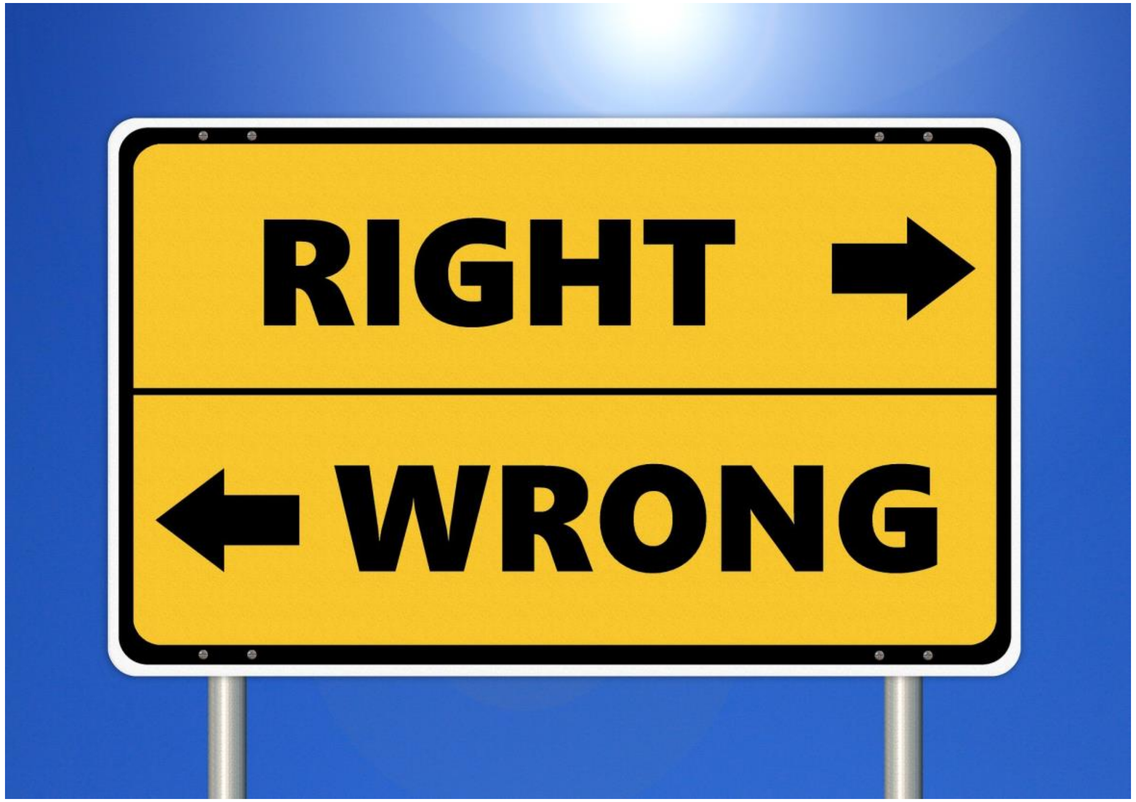
Ethics

- Further discussion in ESDS and Data Governance (Block 3)
- Fundamental part of doing Social Data Science
- See also Salganik lecture on Ethics at the Summer Institute in Computational Social Science: https://youtu.be/A- 5QaX5ZiK8



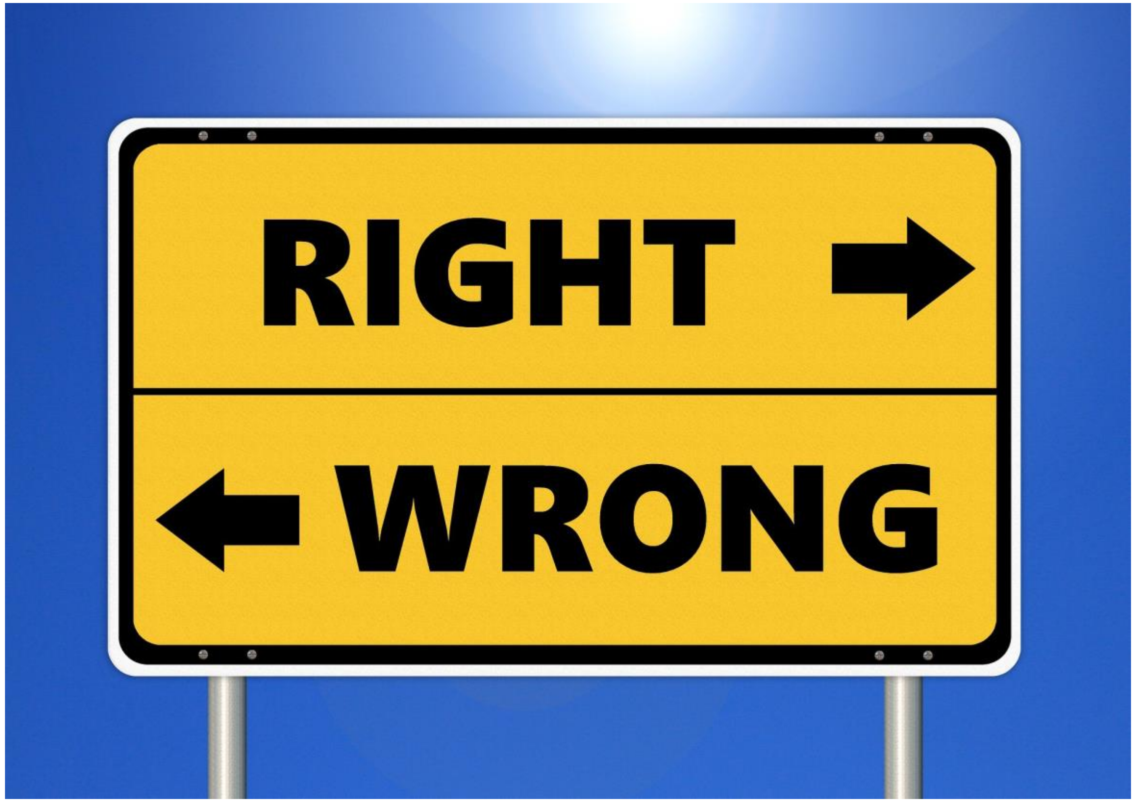
Ethics

- Three approaches
  - Rules-based approach
  - Ad hoc approach
  - Principles-based approach



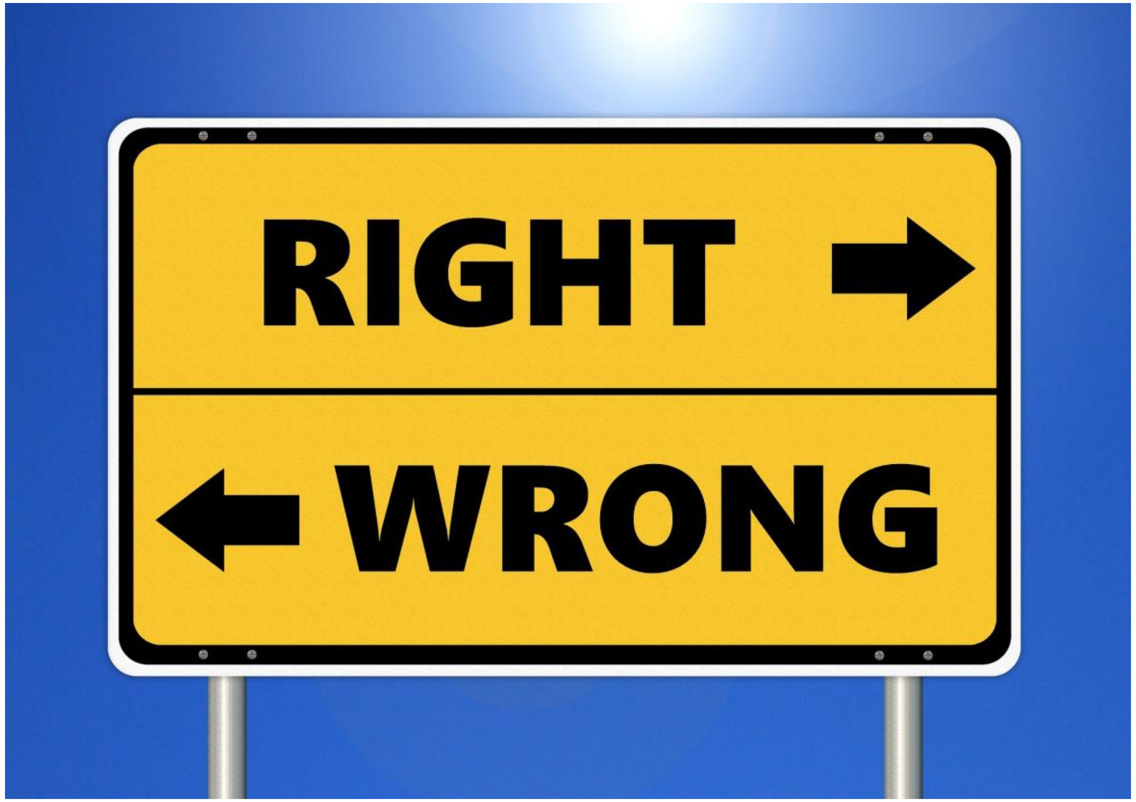
Ethics

- Prominent examples
  - Emotional contagion
  - Tastes, Ties, and Time
  - Encore



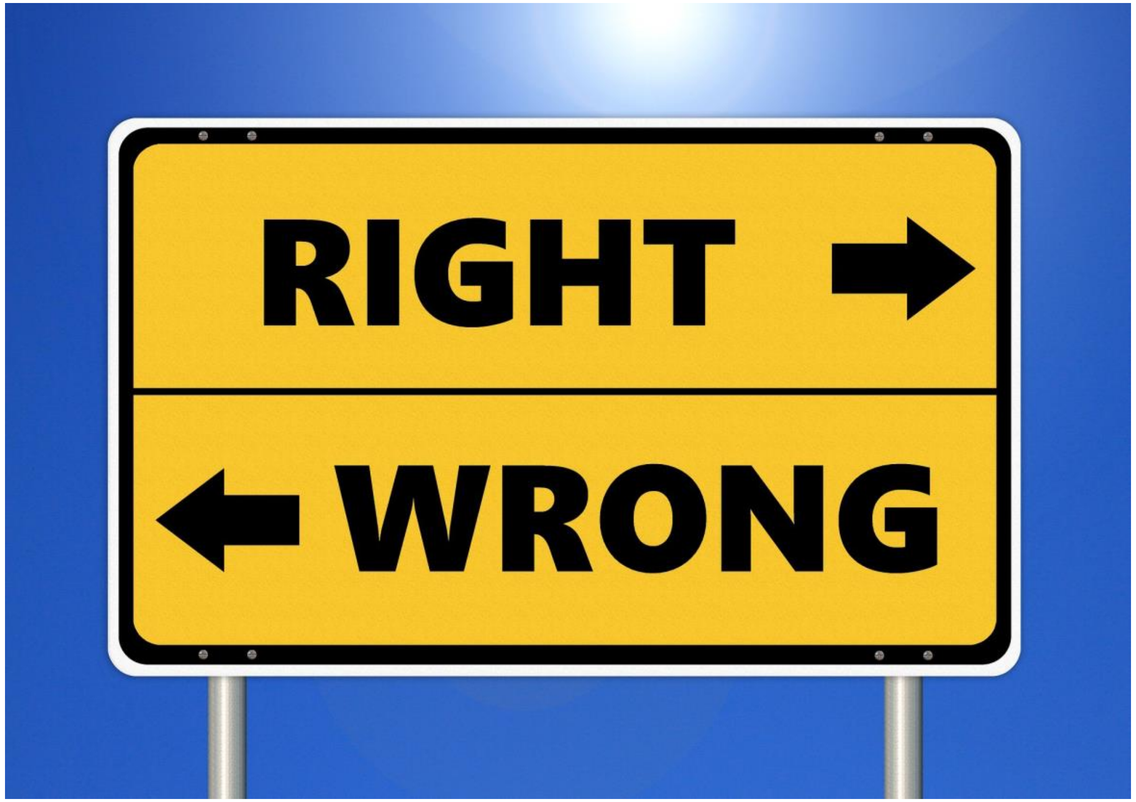
Ethics

- Ethics Exercise A
  - Consider the Base Camp dataset constructed from SODAS, Twitter, and Google Scholar data
  - Which example is this most related to and how?
    - Emotional Contagion
    - Taste, Ties, and Time
    - Encore
  - Do you foresee specific ethical concerns? If so, what are they?
  - Think-Pair-Share



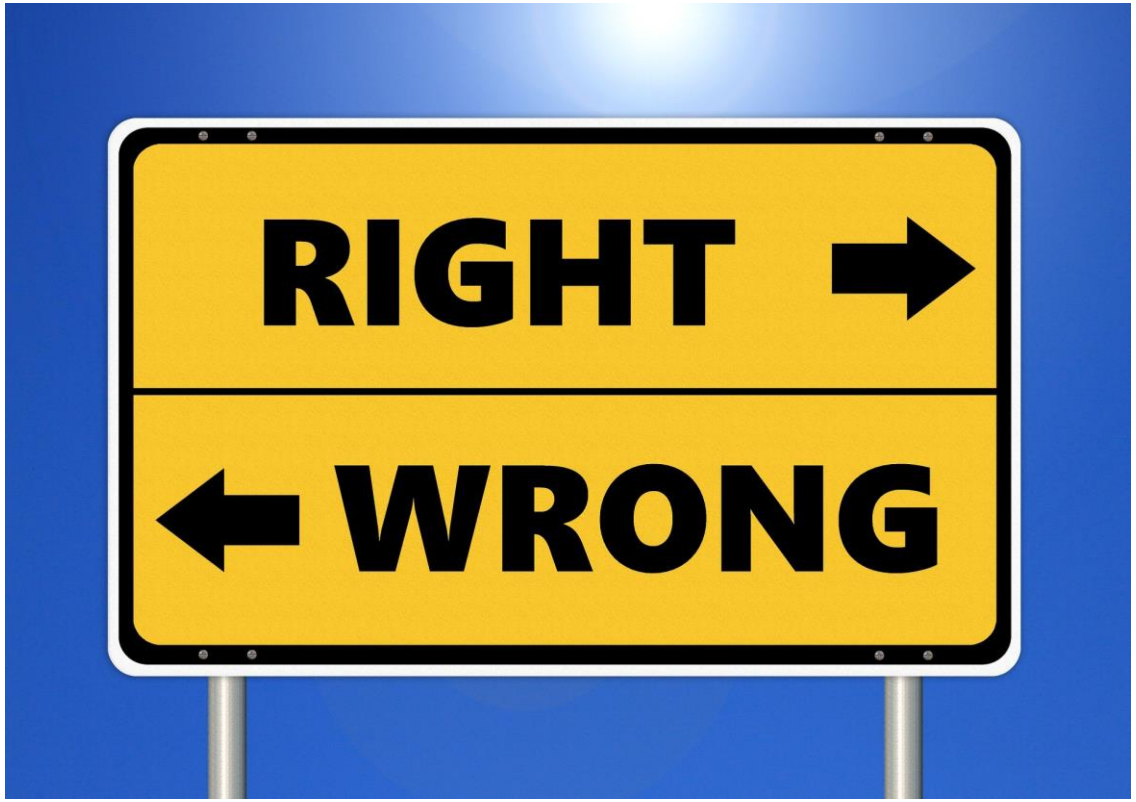
Ethics

- Principles
  - Respect for persons
  - Beneficence
  - Justice
  - Respect for Law and Public Interest



Ethics

- Ethics Exercise B
  - Returning to the Base Camp dataset constructed from SODAS, Twitter, and Google Scholar data
  - Consider the four principles. How would you weigh and, if appropriate, address each of them in this case?
    - Respect for persons
    - Benefice
    - Justice
    - Respect for Law and Public Interest
  - Think-pair-share



Exercise Preview

- Get to know the SODAS data
- Explore data sources and gain first familiarity with pandas
- More data ethics
- Daily reflections
- And more…

Groups Update

- A few shuffles in groups/classes (updated pdf on Absalon)
- If you have new group members, please welcome them!