Computational Social Science Methods and Tools

Aleksi Aaltonen
Warwick Business School

Innovation in the Digital Age seminar 6 February 2017
Warwick Business School

Quantitative Analysis of Culture?

What is
good, interesting, insightful,
about the study?

What is
bad, uninteresting, obvious,
about the study?

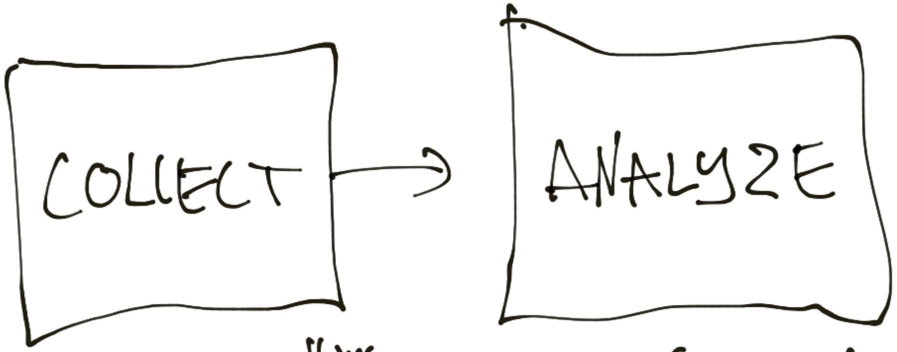
Michel, J.-B., Shen Y. K., Aiden, A. P, Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014): 176–182.

The 'microscope' for social sciences?

Computational methods do not form a distinct methodology.

It’s about **rigorously hacking** together software, new kinds of data sources, old and new methods.

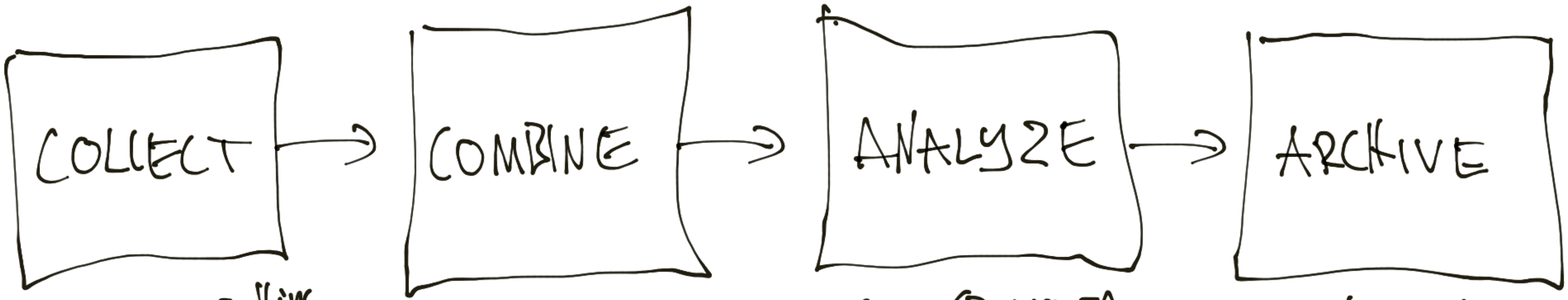
Before



- observe a setting
- make experiments
- run simulation

use QUANTITATIVE AND QUALITATIVE METHODS TO COME UP WITH FINDINGS.

In the future



- observe a setting
- make experiments
- run simulation

Aggregate, transform and filter data so that it can be analyzed.

use QUANTITATIVE AND QUALITATIVE METHODS TO COME UP WITH FINDINGS.

Archive the entire process.

Agenda, sort of

1. **Collecting** new kinds of data

2. **Do I need to learn to program?**

3. Some thoughts on the methods of data analysis

4. **PITFALLS** and a great opportunity!

Collecting new kinds of data

‘Naturally occurring’ data

1. Data is not produced for research purposes.

2. The operationalization of variables, the representativity and reliability of data require special care.

3. Excel cannot deal with 100M+ rows – requires learning to work with data that does not fit into application memory.

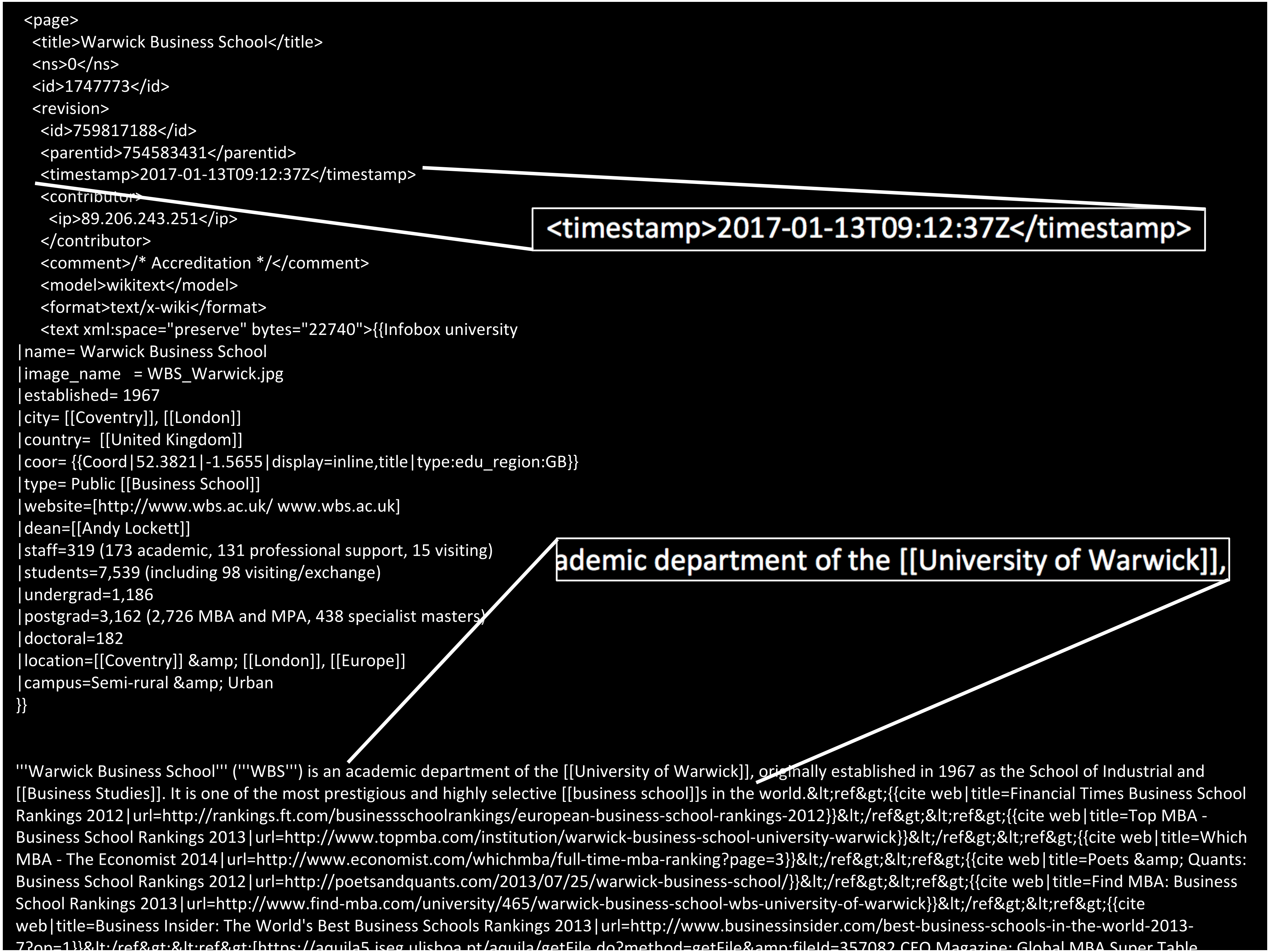
4. There are massive new opportunities to combine various datasets and sources.

‘Naturally occurring’ data

5. Newly available digital data is often more or less unstructured. E.g. images, narrative texts, etc.

6. Unstructured data does not have predefined data model (rows, columns, fields) that would suggest what do the individual data items stand for.

7. Most of the world’s digital information is believe to be stored as unstructured data.



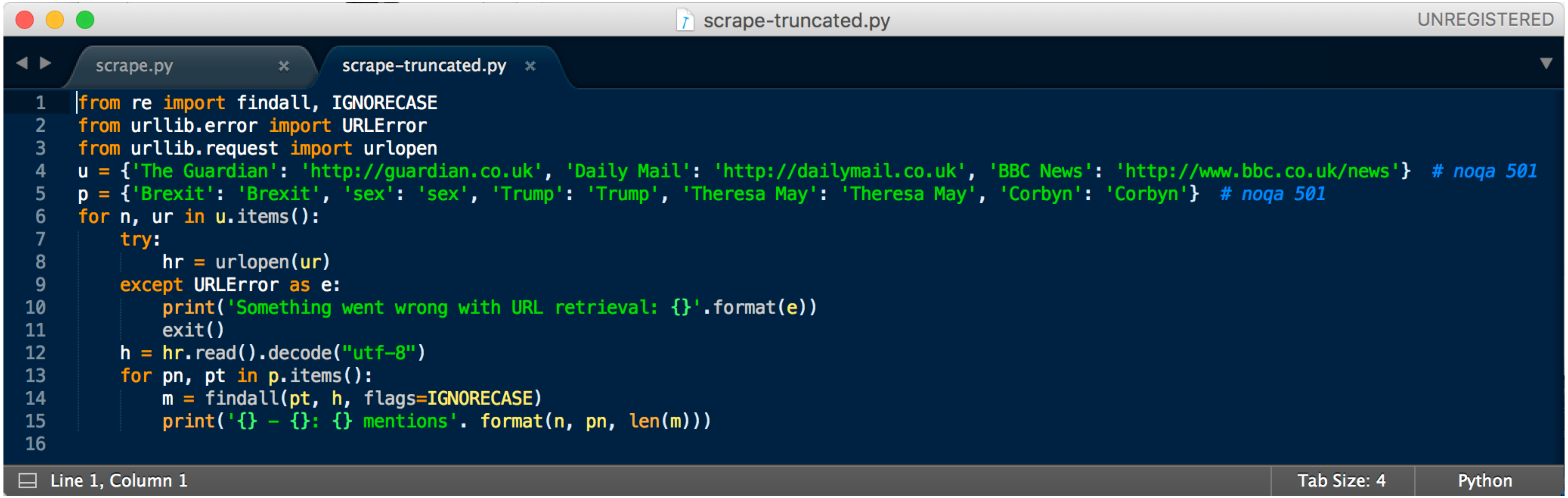
<timestamp>2017-01-13T09:12:37Z</timestamp>
ademic department of the [[University of Warwick]],

```
  <page>
    <title>Warwick Business School</title>
    <ns>0</ns>
    <id>1747773</id>
    <revision>
      <id>759817188</id>
      <parentid>754583431</parentid>
      <timestamp>2017-01-13T09:12:37Z</timestamp>
      <contributor>
        <ip>89.206.243.251</ip>
      </contributor>
      <comment>/* Accreditation */</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text xml:space="preserve" bytes="22740">{{Infobox university
|name= Warwick Business School
|image_name   = WBS_Warwick.jpg
|established= 1967
|city= [[Coventry]], [[London]]
|country=  [[United Kingdom]]
|coor= {{Coord|52.3821|-1.5655|display=inline,title|type:edu_region:GB}}
|type= Public [[Business School]]
|website=[http://www.wbs.ac.uk/ www.wbs.ac.uk]
|dean=[[Andy Lockett]]
|staff=319 (173 academic, 131 professional support, 15 visiting)
|students=7,539 (including 98 visiting/exchange)
|undergrad=1,186
|postgrad=3,162 (2,726 MBA and MPA, 438 specialist masters)
|doctoral=182
|location=[[Coventry]] &amp; [[London]], [[Europe]]
|campus=Semi-rural &amp; Urban
}}


'''Warwick Business School''' ('''WBS''') is an academic department of the [[University of Warwick]], originally established in 1967 as the School of Industrial and
[[Business Studies]]. It is one of the most prestigious and highly selective [[business school]]s in the world.&lt;ref&gt;{{cite web|title=Financial Times Business School
Rankings 2012|url=http://rankings.ft.com/businessschoolrankings/european-business-school-rankings-2012}}&lt;/ref&gt;&lt;ref&gt;{{cite web|title=Top MBA -
Business School Rankings 2013|url=http://www.topmba.com/institution/warwick-business-school-university-warwick}}&lt;/ref&gt;&lt;ref&gt;{{cite web|title=Which
MBA - The Economist 2014|url=http://www.economist.com/whichmba/full-time-mba-ranking?page=3}}&lt;/ref&gt;&lt;ref&gt;{{cite web|title=Poets &amp; Quants:
Business School Rankings 2012|url=http://poetsandquants.com/2013/07/25/warwick-business-school/}}&lt;/ref&gt;&lt;ref&gt;{{cite web|title=Find MBA: Business
School Rankings 2013|url=http://www.find-mba.com/university/465/warwick-business-school-wbs-university-of-warwick}}&lt;/ref&gt;&lt;ref&gt;{{cite
web|title=Business Insider: The World's Best Business Schools Rankings 2013|url=http://www.businessinsider.com/best-business-schools-in-the-world-2013-
7?op=1}}&lt;/ref&gt;&lt;ref&gt;[https://aquila5.iseg.ulisboa.pt/aquila/getFile.do?method=getFile&amp;fileId=357082 CEO Magazine: Global MBA Super Table
```

Computational data collection approaches

1. Get the data from somebody else
2. Write a scraper from scratch
3. Build on an existing software library
4. Use a web-based service or tool
5. Download freely available datasets

Write a scraper from scratch!



```python
from re import findall, IGNORECASE
from urllib.error import URLError
from urllib.request import urlopen
u = {'The Guardian': 'http://guardian.co.uk', 'Daily Mail': 'http://dailymail.co.uk', 'BBC News': 'http://www.bbc.co.uk/news'}  # noqa 501
p = {'Brexit': 'Brexit', 'sex': 'sex', 'Trump': 'Trump', 'Theresa May': 'Theresa May', 'Corbyn': 'Corbyn'}  # noqa 501
for n, ur in u.items():
    try:
        hr = urlopen(ur)
    except URLError as e:
        print('Something went wrong with URL retrieval: {}'.format(e))
        exit()
    h = hr.read().decode("utf-8")
    for pn, pt in p.items():
        m = findall(pt, h, flags=IGNORECASE)
        print('{} - {}: {} mentions'. format(n, pn, len(m)))

```

Simple example

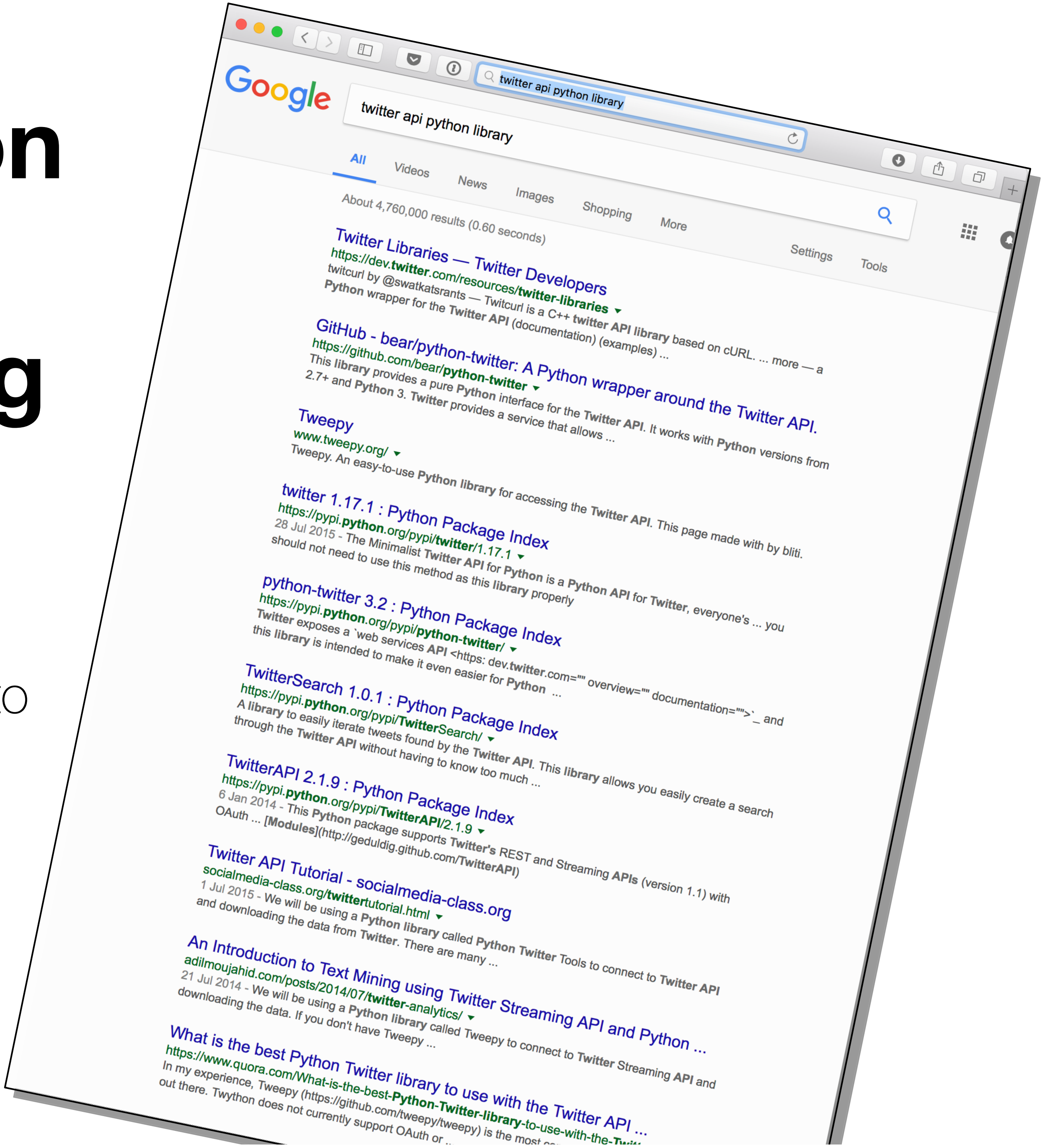
1. We want to asses the editorial values of **BBC News**, **the Guardian**, and **Daily Mail.**

2. Let’s assume that the front page of the website represents what the publication regards most important and appealing to its audience.

3. We observe occurrences of distinctive words that suggest certain emphasis in reporting: “Brexit”, “sex”, “Trump”, “Theresa May”, “Corbyn”

Build on an existing library

You still need to program a little…



twitter api python library

About 4,760,000 results (0.60 seconds)

Twitter Libraries — Twitter Developers
https://dev.twitter.com/resources/twitter-libraries
twitcurl by @swatkatsrants — Twitcurl is a C++ twitter API library based on cURL. ... more — a Python wrapper for the Twitter API (documentation) (examples) ...

GitHub - bear/python-twitter: A Python wrapper around the Twitter API.
https://github.com/bear/python-twitter
This library provides a pure Python interface for the Twitter API. It works with Python versions from 2.7+ and Python 3. Twitter provides a service that allows ...

Tweepy
www.tweepy.org/
Tweepy. An easy-to-use Python library for accessing the Twitter API. This page made with by bliti.

twitter 1.17.1 : Python Package Index
https://pypi.python.org/pypi/twitter/1.17.1
28 Jul 2015 - The Minimalist Twitter API for Python is a Python API for Twitter, everyone's ... you should not need to use this method as this library properly

python-twitter 3.2 : Python Package Index
https://pypi.python.org/pypi/python-twitter/
Twitter exposes a `web services API <https: dev.twitter.com="" overview="" documentation="">`_ and this library is intended to make it even easier for Python ...

TwitterSearch 1.0.1 : Python Package Index
https://pypi.python.org/pypi/TwitterSearch/
A library to easily iterate tweets found by the Twitter API. This library allows you easily create a search through the Twitter API without having to know too much ...

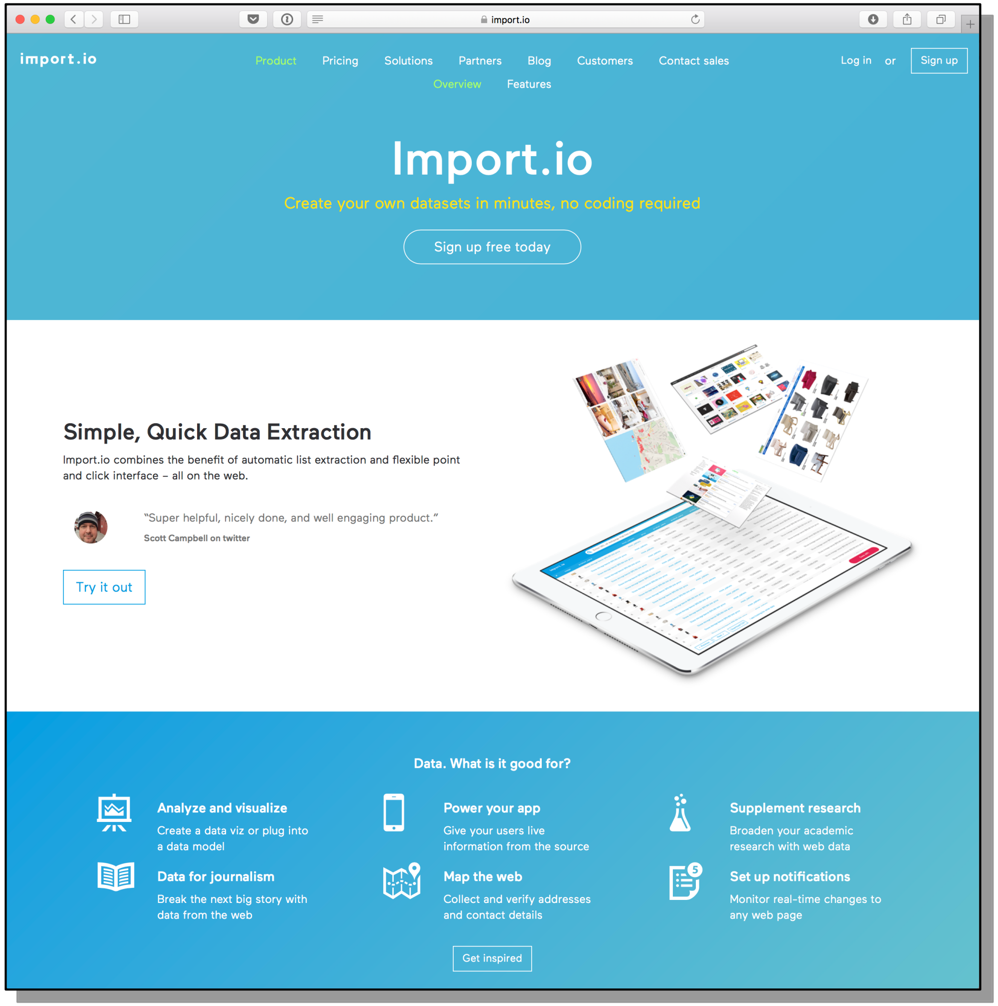
TwitterAPI 2.1.9 : Python Package Index
https://pypi.python.org/pypi/TwitterAPI/2.1.9
6 Jan 2014 - This Python package supports Twitter's REST and Streaming APIs (version 1.1) with OAuth ... [Modules](http://geduldig.github.com/TwitterAPI)

Twitter API Tutorial - socialmedia-class.org
socialmedia-class.org/twittertutorial.html
1 Jul 2015 - We will be using a Python library called Python Twitter Tools to connect to Twitter API and downloading the data from Twitter. There are many ...

An Introduction to Text Mining using Twitter Streaming API and Python ...
adilmoujahid.com/posts/2014/07/twitter-analytics/
21 Jul 2014 - We will be using a Python library called Tweepy to connect to Twitter Streaming API and downloading the data. If you don't have Tweepy ...

What is the best Python Twitter library to use with the Twitter API ...
https://www.quora.com/What-is-the-best-Python-Twitter-library-to-use-with-the-Twi
In my experience, Tweepy (https://github.com/tweepy/tweepy) is the most
out there. Twython does not currently support OAuth or ...

Using a service or a tool



Import.io

Create your own datasets in minutes, no coding required

Simple, Quick Data Extraction

Import.io combines the benefit of automatic list extraction and flexible point and click interface – all on the web.

"Super helpful, nicely done, and well engaging product."

Scott Campbell on twitter

Data. What is it good for?

Analyze and visualize
Create a data viz or plug into a data model

Power your app
Give your users live information from the source

Supplement research
Broaden your academic research with web data

Data for journalism
Break the next big story with data from the web

Map the web
Collect and verify addresses and contact details

Set up notifications
Monitor real-time changes to any web page

import.io
scrapinghub.com
grepsr.com
datahut.co
webrobots.io
webscraper.io
etc.

See also:
http://scraping.pro/choosing-web-scraping-service/
https://www.quora.com/What-are-the-best-web-crawling-services

Sometimes, you can just download the freely available dataset…

Storing the data

Text files are great for linear processing; they are easy to debug and can deal with huge amounts of data.

SQL (relational) databases (e.g. MySQL, Microsoft SQL server) are good for processing complex relational data stored as tables.

NoSQL databases (e.g. MongoDB, Apache Cassandra, Redis) do away with rigidities of relational databases to gain speed and flexibility of development.

Graph databases (e.g. Neo4j) support various operations on graph data.

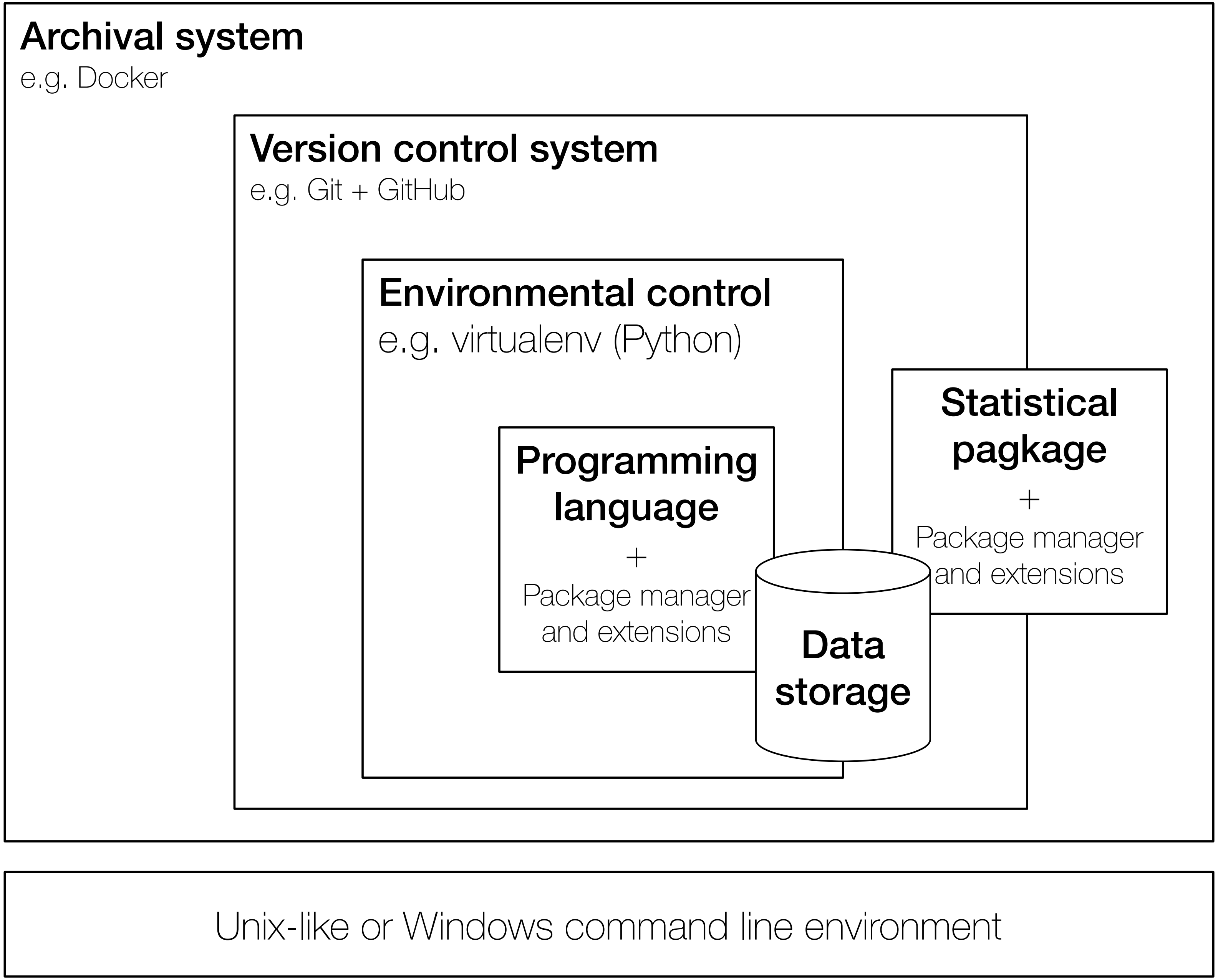
Do I need to learn to program?

YES. Taking the full control of the computational methods requires working directly with program code.

No. There are various tools with graphical user interface that may allow you to do what you need for your research.

Good news is that programming is not rocket science – they teach Python to schoolchildren so I am sure you are capable of learning it as well.

Bad news is that developing software (academic or not) is not just about knowing a programming language.


Archival system
e.g. Docker
Version control system
e.g. Git + GitHub
Environmental control
e.g. virtualenv (Python)
Programming language
+
Package manager and extensions
Data storage
Statistical pagkage
+
Package manager and extensions
Unix-like or Windows command line environment

Programming languages

Compiled languages (e.g. C and its variants) are fast to execute but difficult to learn and slow to develop.

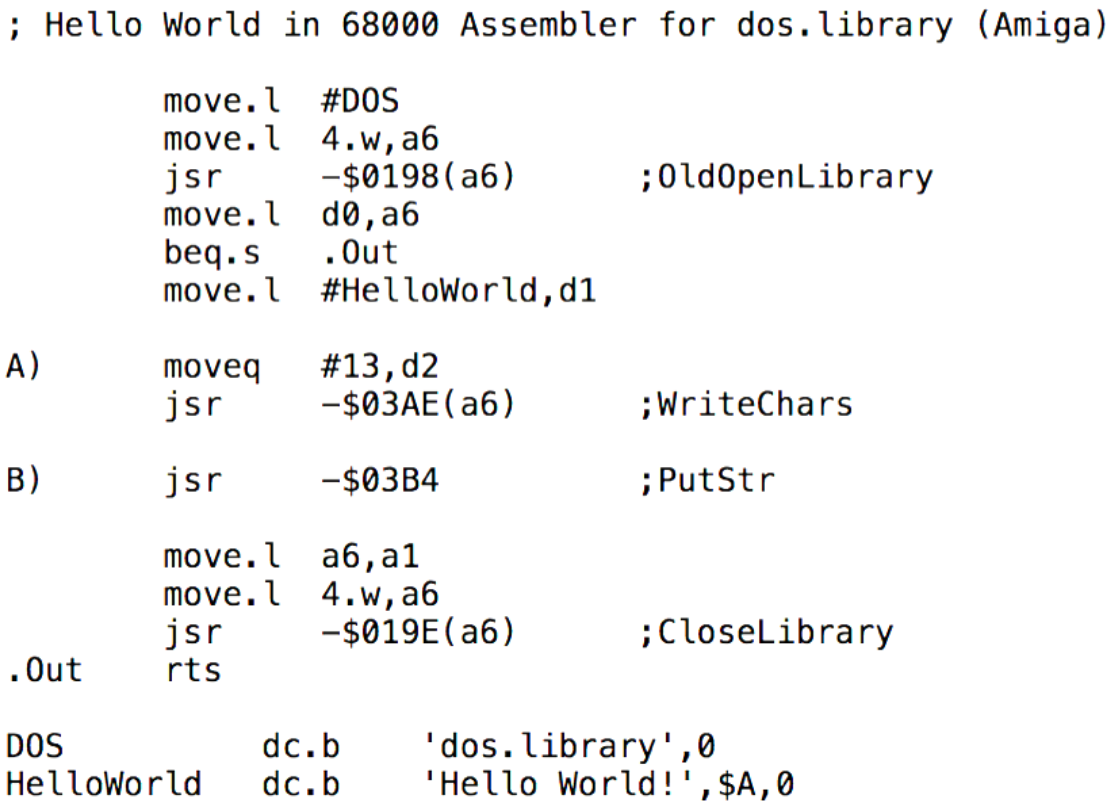
Interpreted languages (e.g. Python) slow to execute but easy to learn and fast to develop.

You may also need to learn a little bit of some **domain-specific languages** such as SQL (data storage), R (statistics), HTML/CSS/Javascript (web interfaces), and *definitely* Unix or Windows shell.

… but don't worry – it's much easier than learning Finnish!

Print “Hello World!”

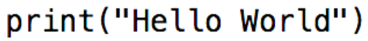
Amiga MC68000 assembler



```
; Hello World in 68000 Assembler for dos.library (Amiga)

        move.l  #DOS
        move.l  4.w,a6
        jsr     -$0198(a6)      ;OldOpenLibrary
        move.l  d0,a6
        beq.s   .Out
        move.l  #HelloWorld,d1

A)      moveq   #13,d2
        jsr     -$03AE(a6)      ;WriteChars

B)      jsr     -$03B4          ;PutStr

        move.l  a6,a1
        move.l  4.w,a6
        jsr     -$019E(a6)      ;CloseLibrary
.Out    rts

DOS          dc.b    'dos.library',0
HelloWorld   dc.b    'Hello World!',$A,0
```

Print “Hello World!”

Python 3



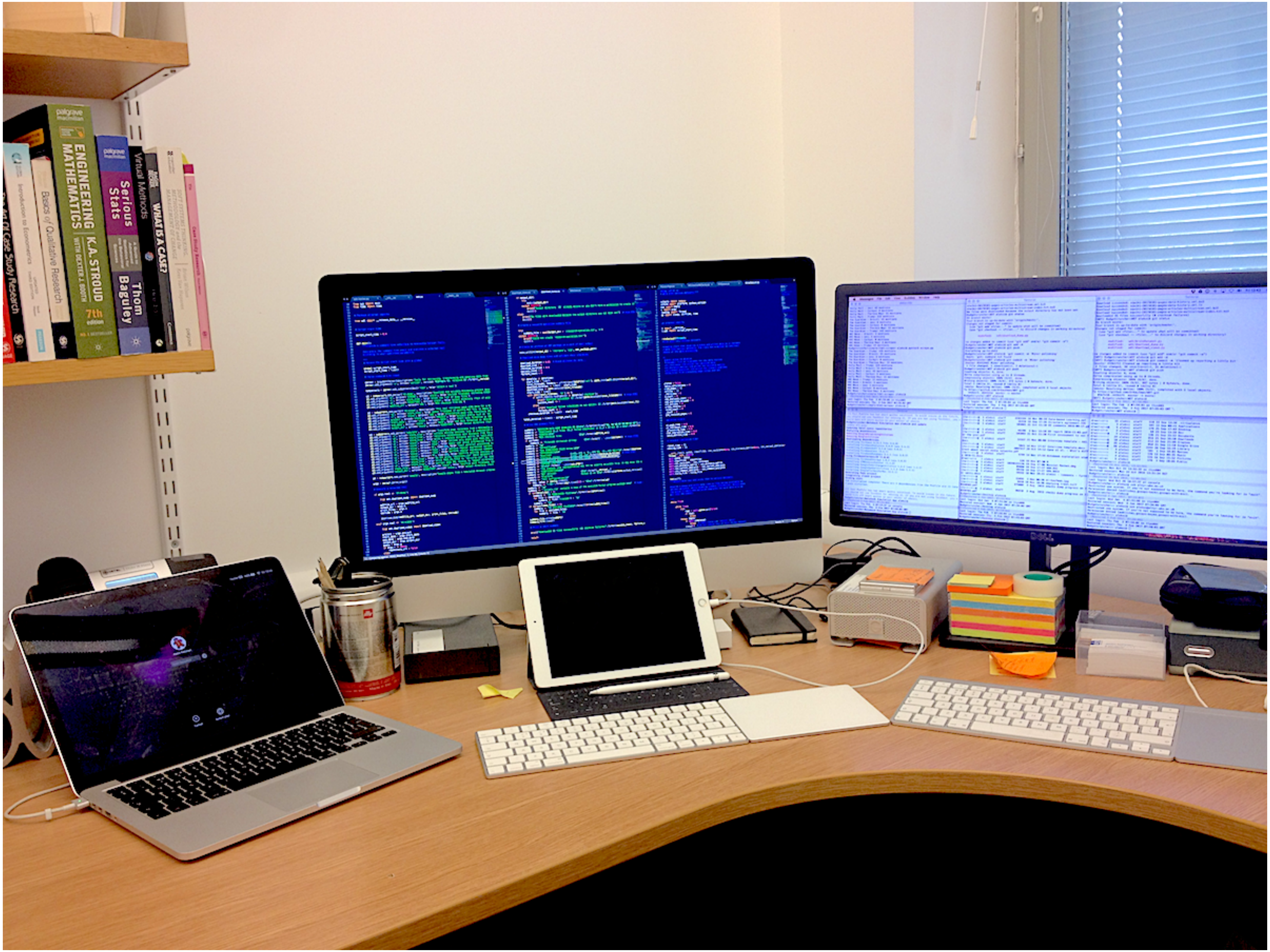
```
print("Hello World")
```

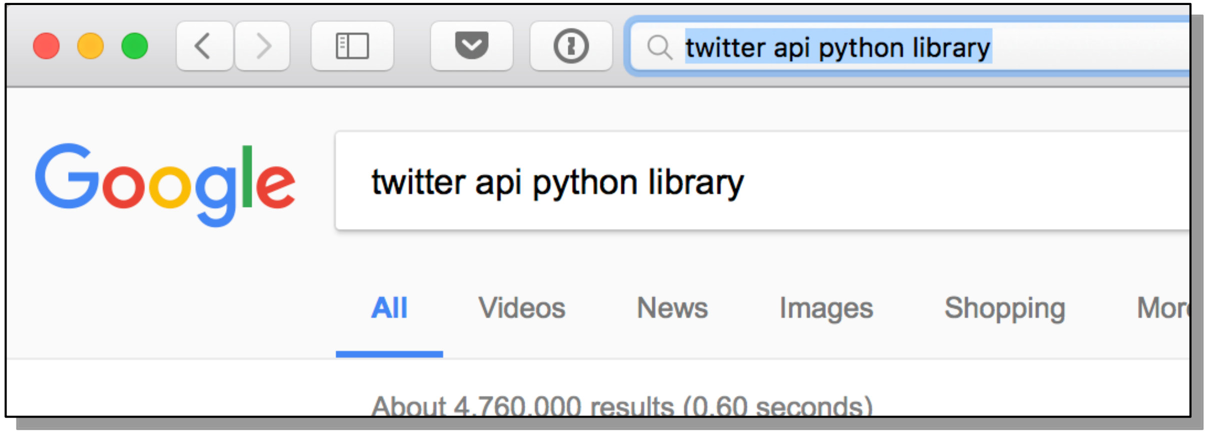
Lower your own barriers to do coding!

1. Develop practices that reveal your progress.

2. Create an environment in which you can leave and pick up your programming task any time.

3. Don’t plan too much in advance; instead, build something that produces some output and then iterate furiously toward your goal.


The Art of Case Study Research
Introduction to Econometrics
Basics of Qualitative Research
palgrave macmillan
ENGINEERING MATHEMATICS
K.A. STROUD
WITH DEXTER J. BOOTH
7th edition
Serious Stats
Thom Baguley
Virtual Methods
WHAT IS A CASE?
Case Study Research
DELL

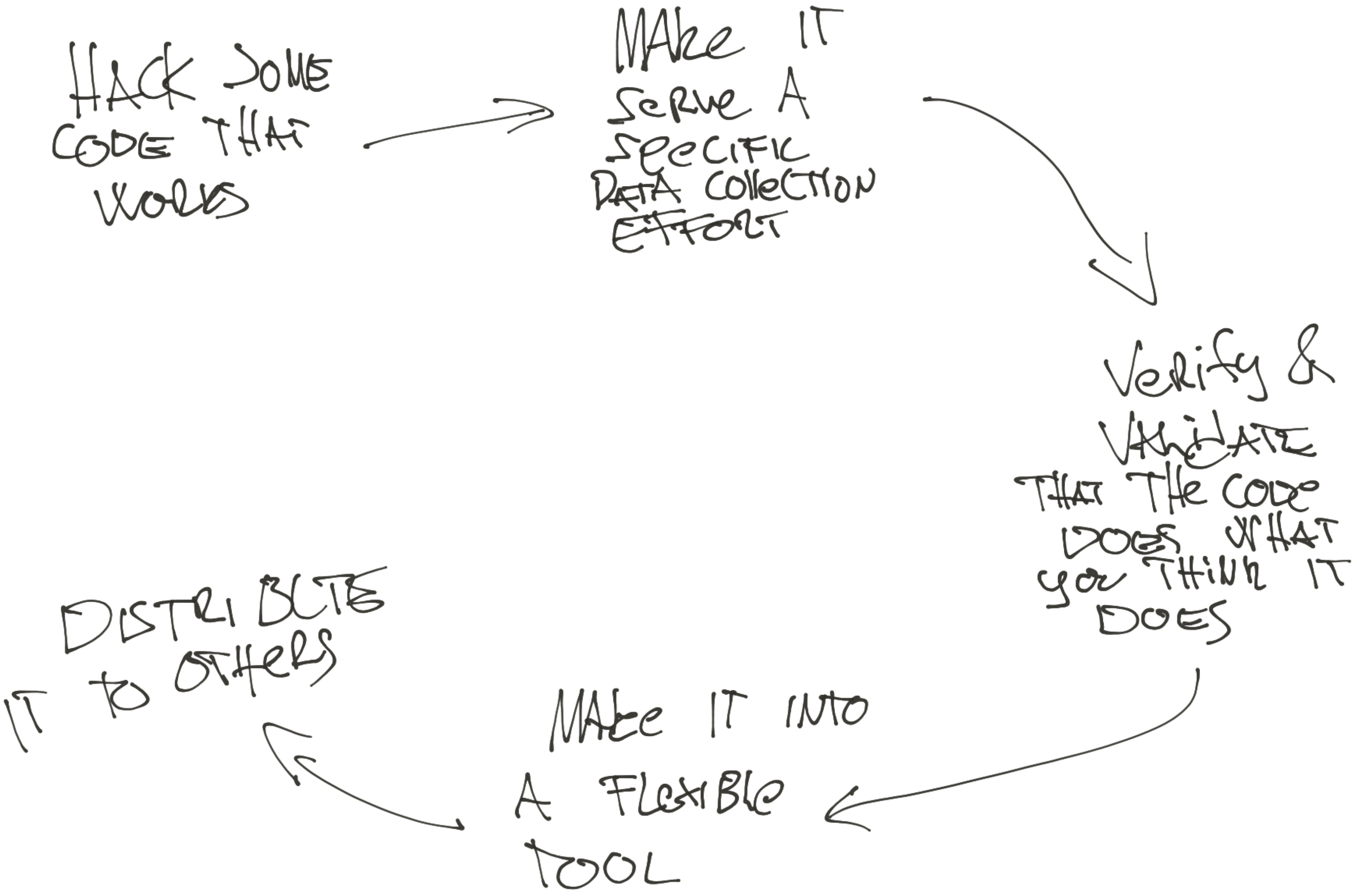


twitter api python library

twitter api python library

About 4,760,000 results (0.60 seconds)

Learn to Google the good stuff!


HACK SOME CODE THAT WORKS
MAKE IT SERVE A SPECIFIC DATA COLLECTION EFFORT
Verify & VALIDATE THAT THE CODE DOES WHAT YOU THINK IT DOES
MAKE IT INTO A FLEXIBLE TOOL
DISTRIBUTE IT TO OTHERS

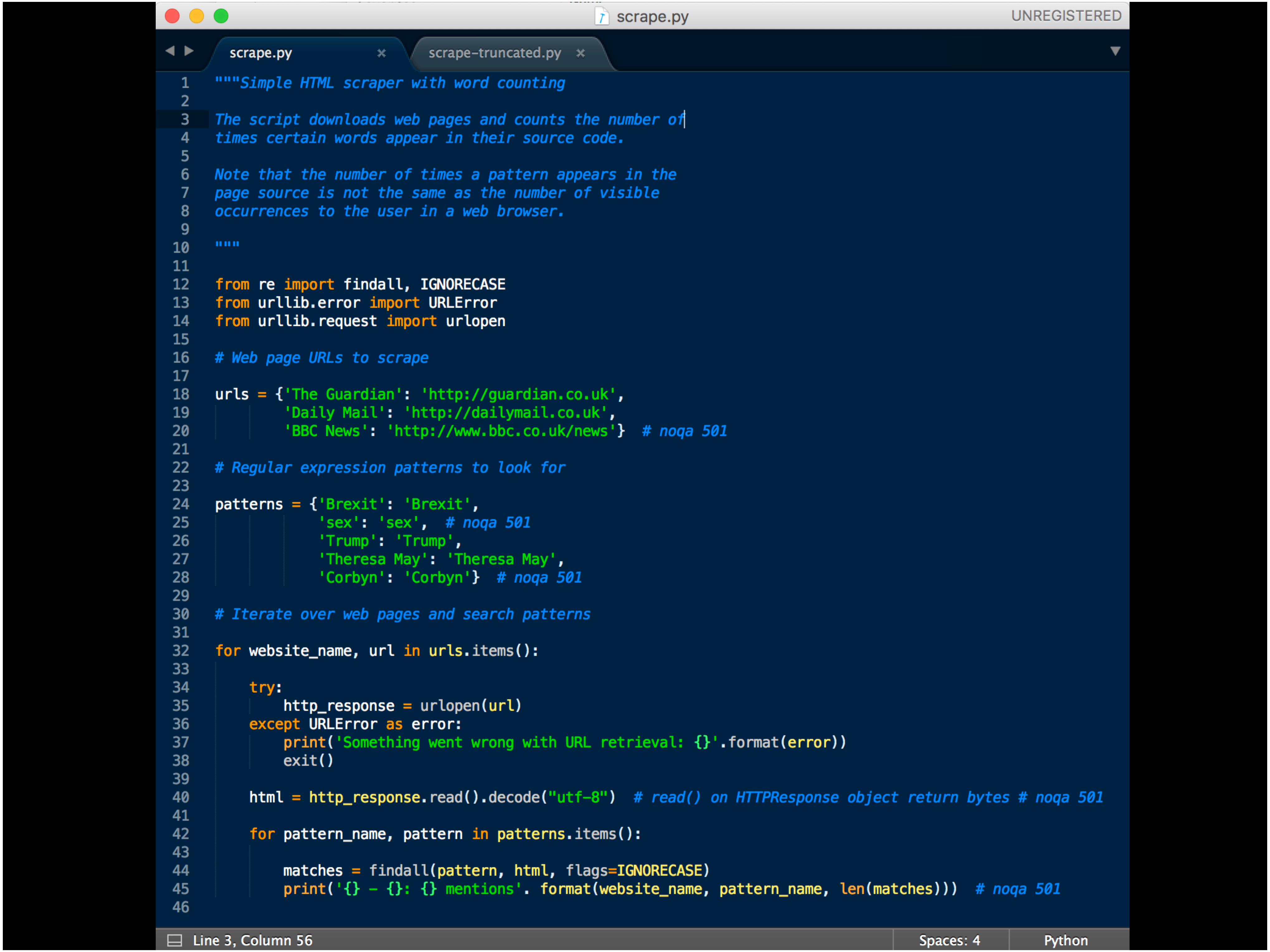
Documentation matters!

Document code **for yourself** to allow you to pick up after six months where you left it today.

Adapt (not adopt) general guidelines and practices to develop your own documentation style that is suitable for academic purposes.

CONSISTENCY is EVERYTHING!

Programming languages often have **style guides** available in the web, e.g. Google Python Style Guide: (https://google.github.io/styleguide/pyguide.html)



```python
"""Simple HTML scraper with word counting

The script downloads web pages and counts the number of
times certain words appear in their source code.

Note that the number of times a pattern appears in the
page source is not the same as the number of visible
occurrences to the user in a web browser.

"""

from re import findall, IGNORECASE
from urllib.error import URLError
from urllib.request import urlopen

# Web page URLs to scrape

urls = {'The Guardian': 'http://guardian.co.uk',
        'Daily Mail': 'http://dailymail.co.uk',
        'BBC News': 'http://www.bbc.co.uk/news'}  # noqa 501

# Regular expression patterns to look for

patterns = {'Brexit': 'Brexit',
            'sex': 'sex',  # noqa 501
            'Trump': 'Trump',
            'Theresa May': 'Theresa May',
            'Corbyn': 'Corbyn'}  # noqa 501

# Iterate over web pages and search patterns

for website_name, url in urls.items():

    try:
        http_response = urlopen(url)
    except URLError as error:
        print('Something went wrong with URL retrieval: {}'.format(error))
        exit()

    html = http_response.read().decode("utf-8")  # read() on HTTPResponse object return bytes # noqa 501

    for pattern_name, pattern in patterns.items():

        matches = findall(pattern, html, flags=IGNORECASE)
        print('{} - {}: {} mentions'. format(website_name, pattern_name, len(matches)))  # noqa 501

```

Don’t be afraid of version control!

Git + GitHub or Bitbucket

It’s great not just for rescuing screwed up code, but as a backup, collaboration and distribution tool.

Some thoughts on the methods of data analysis

1. New opportunities arise from innovative combinations of old and new datasets.

2. New methods emerge and need to be adopted: supervised and unsupervised learning (ML), sequence analysis (biology), predictive and real-time modeling, etc.

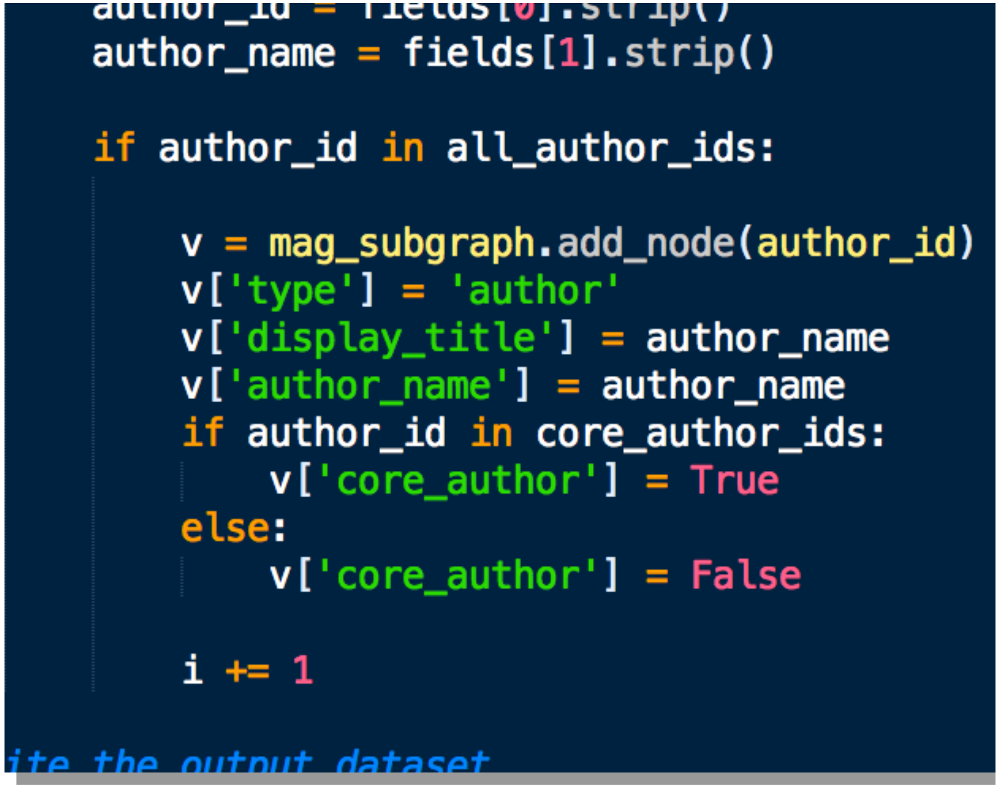
3. Nevertheless, old methods and methodological learnings still largely apply – computational research is not a license to do sloppy empirical research.

4. We still need to theorize and understand causality!

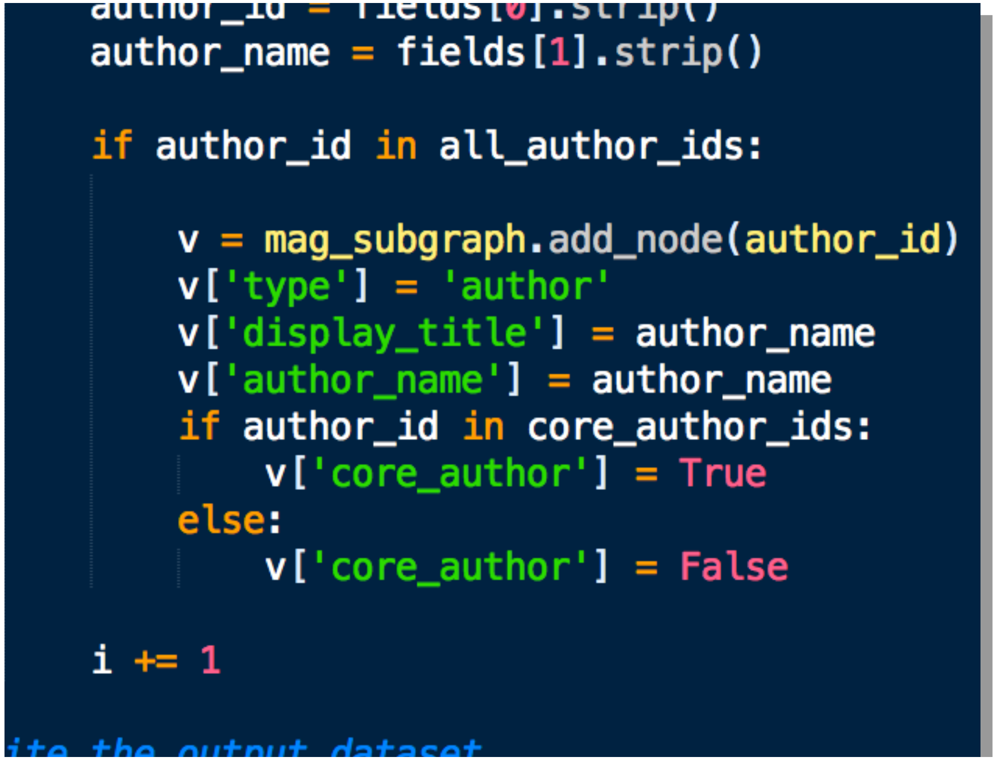
5. Statistical significance loses its role as a proxy for ‘importance’ if you can simply increase sample size to anything significant.

PITFALLS and a great opportunity!

Spot the difference!



```
    author_id = fields[0].strip()
    author_name = fields[1].strip()

    if author_id in all_author_ids:

        v = mag_subgraph.add_node(author_id)
        v['type'] = 'author'
        v['display_title'] = author_name
        v['author_name'] = author_name
        if author_id in core_author_ids:
            v['core_author'] = True
        else:
            v['core_author'] = False

        i += 1

ite the output dataset
```



```
    author_id = fields[0].strip()
    author_name = fields[1].strip()

    if author_id in all_author_ids:

        v = mag_subgraph.add_node(author_id)
        v['type'] = 'author'
        v['display_title'] = author_name
        v['author_name'] = author_name
        if author_id in core_author_ids:
            v['core_author'] = True
        else:
            v['core_author'] = False

    i += 1

ite the output dataset
```

Which one is correct?

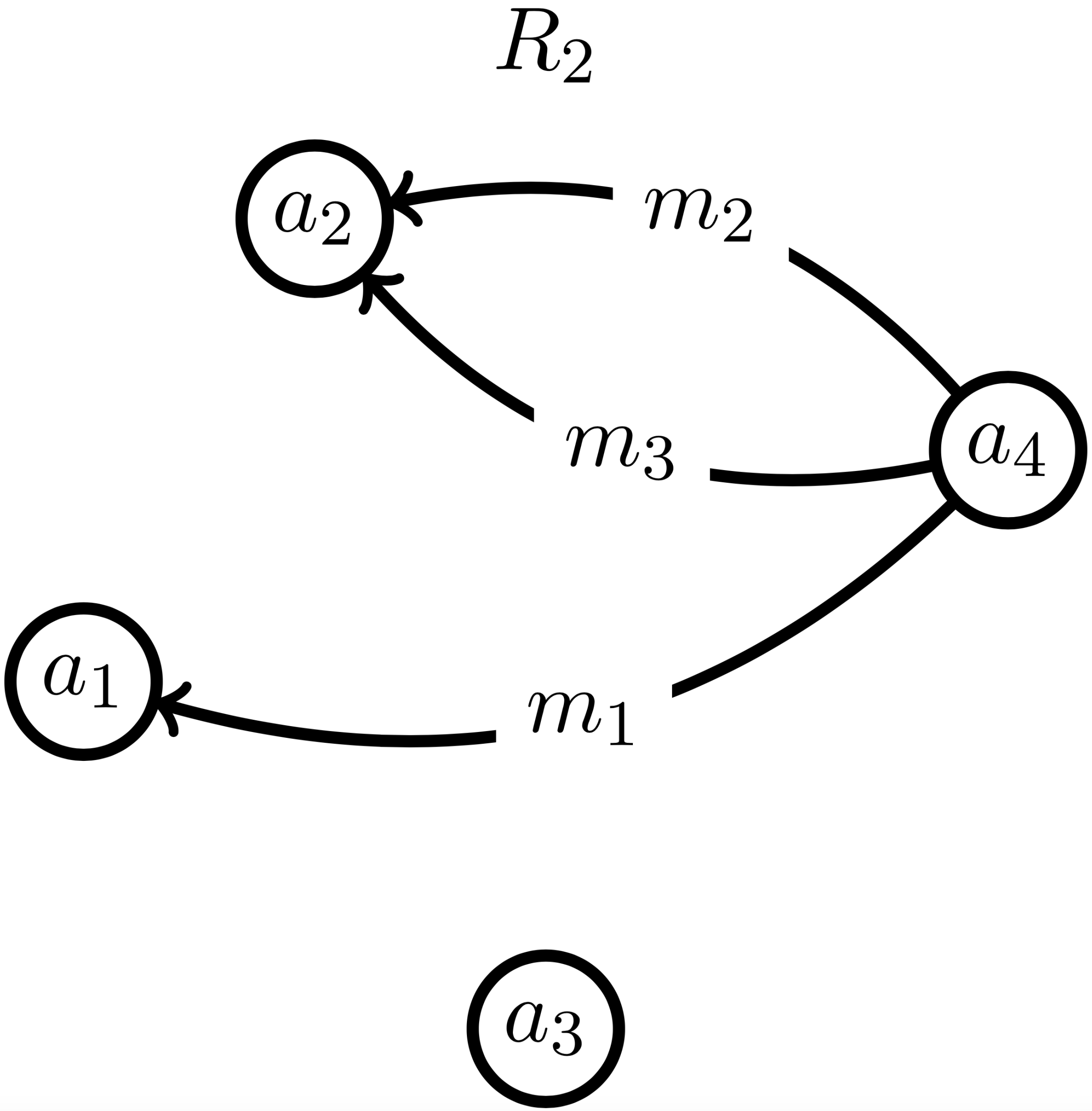
Research quality

Every step in the increasingly complex research process is an opportunity for something to go wrong.

When you have 100M observations, you cannot validate every row manually.

The principles of replicability, reliability, repeatability, etc. will become much more tangible, operational and important in the context of individual research project.

We may need to do ’academic’ software testing and develop new ways to describe research processes.


R_2
a_2
m_2
m_3
a_4
a_1
m_1
a_3

Ethics and good conduct

The fact that you can get your hands on the data does not mean that you can ignore research ethics.

Getting access to non-public big data can become more and more difficult as parties perceive its potential value/sensitivity.

Performance bottlenecks

A lot can be done on your laptop.

Keep development vs. execution time balance in mind.

Programming a specific analysis vs. developing a flexible toolset is a tricky balance.

High-performance facilities are relatively easy to access if needed.

Computational Tooling Group at WBS

Email Aleksi.Aaltonen@wbs.ac.uk

Bibliography

Burton, R. M., and Obel, B. 2011. Computational Modeling for What-Is, What-Might-Be, and What-Should-Be Studies — and Triangulation. Organization Science 22(5): 1195–1202.

Cioffi-Revilla, C., Gries, D., and Schneider, F. B. 2014. Introduction to Computational Social Science: Principles and Applications. Texts in Computer Science. London: Springer.

Conte, R., and Paolucci, M. 2014. On Agent-Based Modeling and Computational Social Science. Frontiers in Psychology, 5.

Evans, J., and Rzhetsky, A. 2010. Machine Science. Science, 329(5990): 399–400.

Gonçalves, B., and Perra, N. 2015. Social Phenomena from Data Analysis to Models. London: Springer.

Heiberger, R. H., and Riebling, J. R. 2016. Installing Computational Social Science: Facing the Challenges of New Information and Communication Technologies in Social Science. Methodological Innovations 9: 1–11.

Howison, J., Wiggins, A., and Crowston, K. 2011. Validity Issues in the Use of Social Network Analysis with Digital Trace Data. Journal of the Association for Information Systems 12(12): 767–97.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N. et al. Computational Social Science. Science, 323(5915): 721–723.

Mann, A. 2016. Computational Social Science. PNAS, 113(3): 468–470.

Miller, J. H., and Page, S. E. 2007. Computation as Theory. Princeton University Press.

Michel, J.-B., Shen Y. K., Aiden, A. P, Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 331(6014): 176–182.