TOOLS AND BEST PRACTICES IN QUANTITATIVE RESEARCH

Mercè Crosas, IQSS, Harvard University

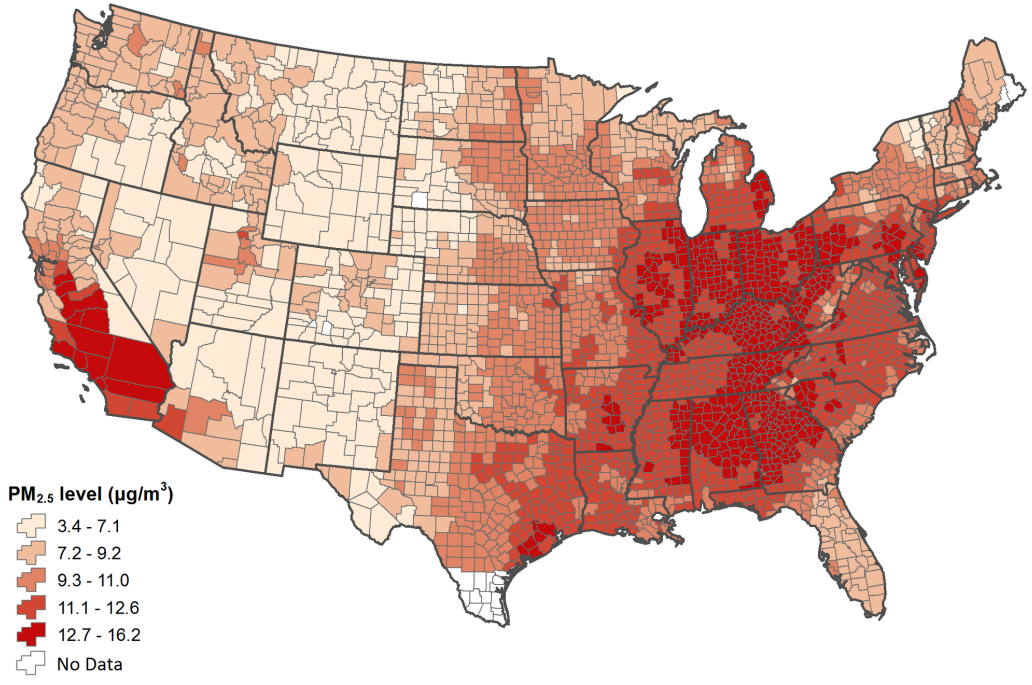
@mercecrosas

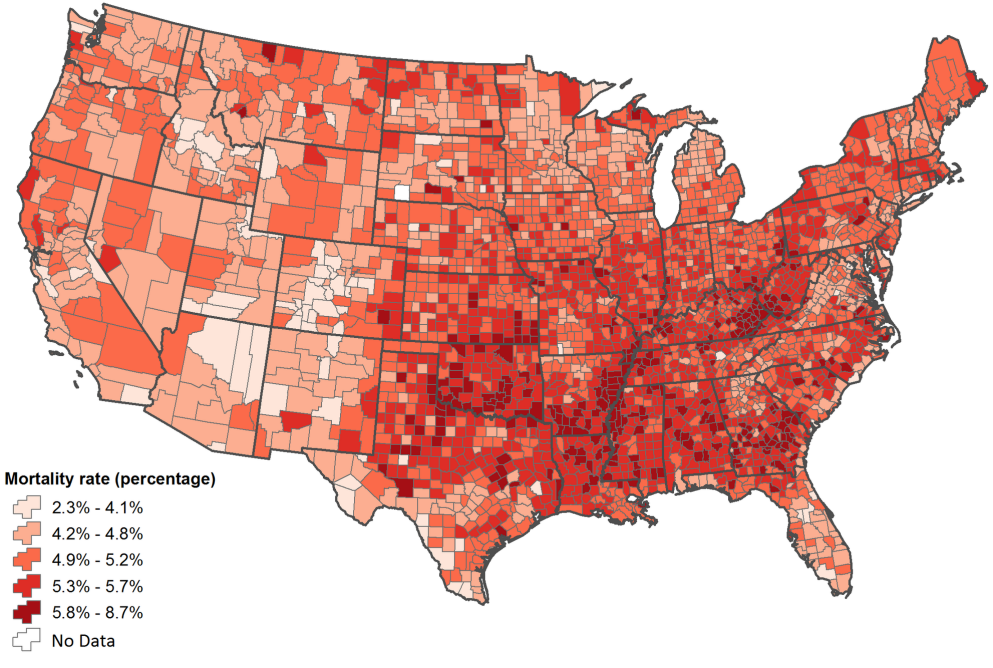
XXXVII JORNADAS DE ECONOMÍA DE LA SALUD, BARCELONA, 6-8 SEPTIEMBRE

SOME EXAMPLES

OF WHAT YOU CAN DO USING QUANTITATIVE DATA AND METHODS

- quantitative and geospatial data
- unstructured text as data





Imagine that you have data for all the deaths of all Medicare beneficiaries in the US 2000-2012 (~half a million person-years) and want to model **the effect of air pollution levels on death**, controlling for other factors that also affect death (such as smoking, BMI).

The NEW ENGLAND JOURNAL of MEDICINE

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D., Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

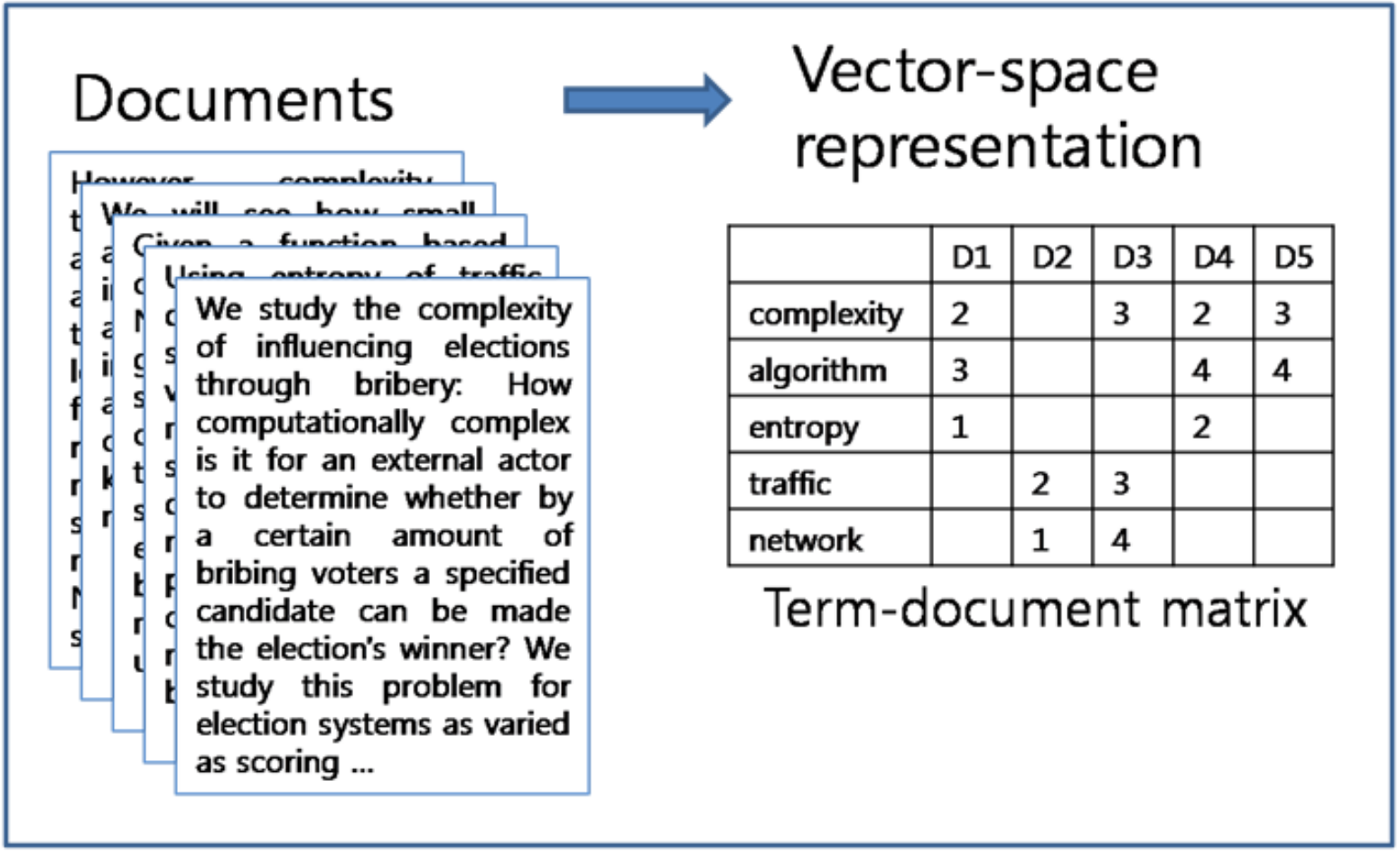
N Engl J Med 2017; 376:2513-2522 | June 29, 2017 | DOI: 10.1056/NEJMoa1702747

Concludes that levels of PM2.5 below the current standard are still harmful

WHAT IS USED TO COMPUTE?

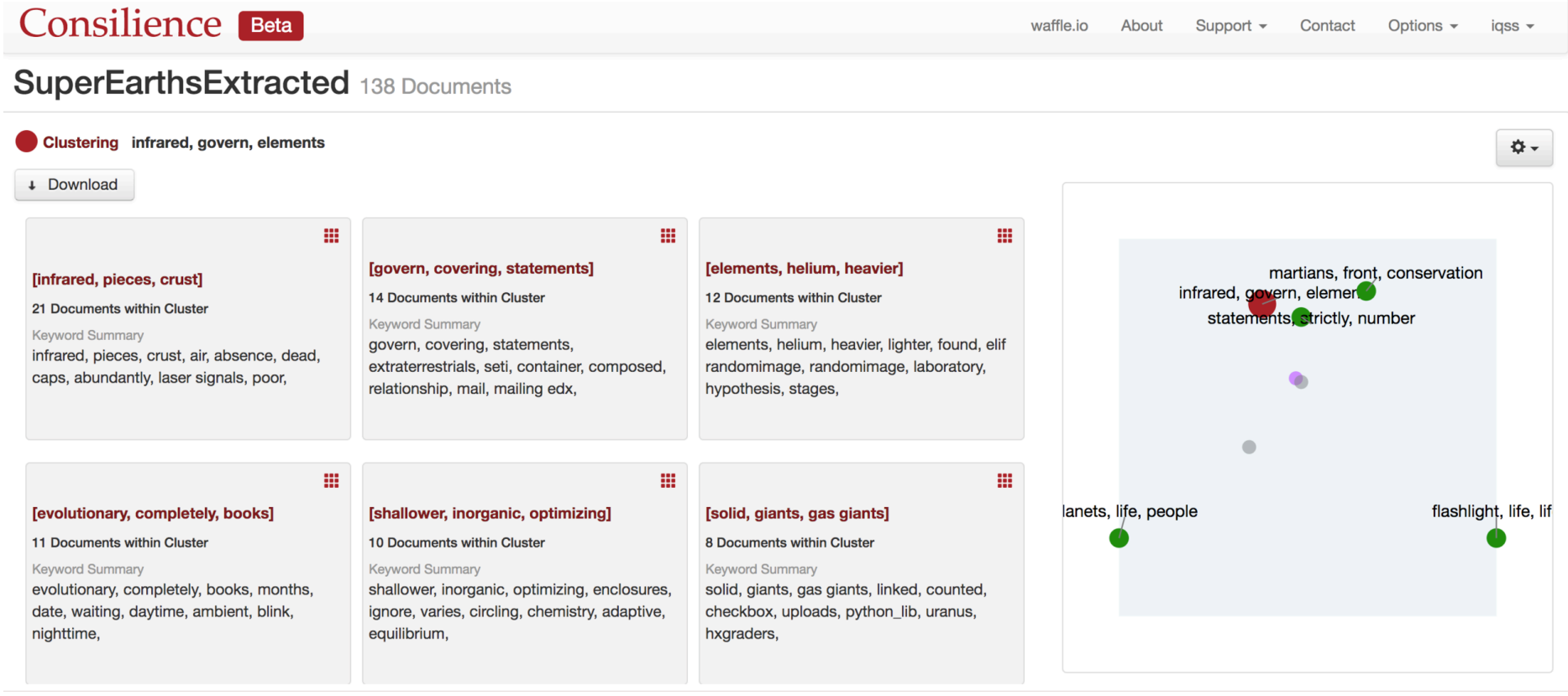
- Medicare data:
  - 5 TB
  - Privacy requirements
- Air pollution grids
  - 50 TB
- Statistical model from survival analysis (Cox proportional hazard, R package)
- WorldMap (GIS) for geospatial visualizations
- Massive computations performed on a secure cluster (1.3 years of combined runtime, ~700 CPU's, 24TB of memory, on IQSS Research Computing Environment)

FROM TEXT TO QUANTITATIVE DATA



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

source: https://www.linkedin.com/pulse/text-mining-its-applications-industry-subhajit-mukherjee



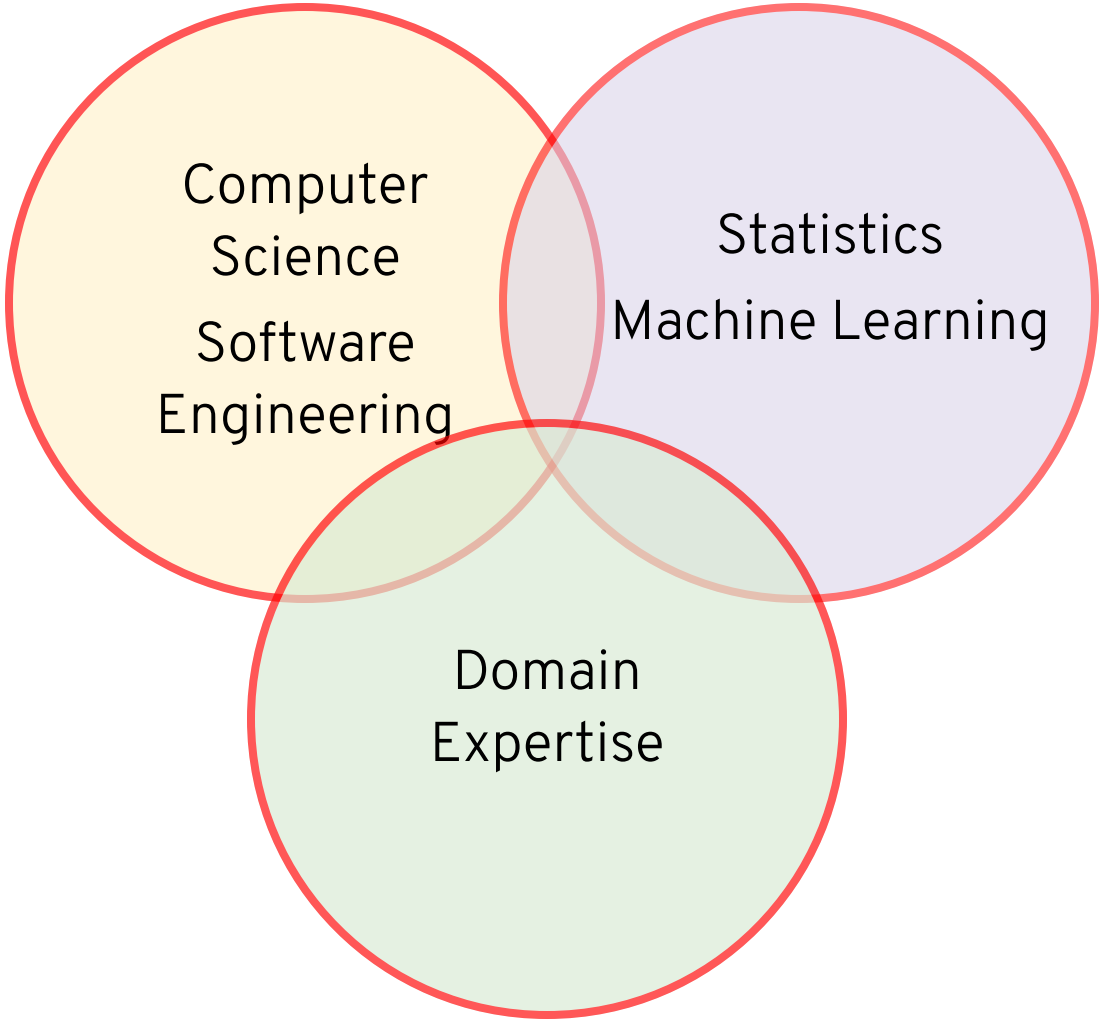
Consilience: a tool that enables you to quickly read, understand, categorize, and derive insights from large quantities of unstructured text.

WHAT IS USED TO COMPUTE?

- On Microsoft Azure cloud
- Using Java, Scala, Python, R
- MangoDB (document database)
- Apache Spark (large-scale data processing engine)
- Elasticsearch (search and analytics engine)
- Machine Learning Library (MLlib) for Spark
- 6 worker VMs with 112 GB memory each (total 672 GB), 16 cores each (total 64 cores)

SKILLS, LANGUAGES, AND TOOLS

USING QUANTITATIVE METHODS IN TODAY'S DATA-INTENSIVE RESEARCH



Program comparison

Program	Statistics	Visualization	Machine learning	Ease of use	Power/flexibility	Fun
Stata	Good	Serviceable	Limited	Very easy	Low	Some
SPSS	OK	Serviceable	Limited	Easy	Low	None
SAS	Good	Not great	Good	Moderate	Moderate	None
Matlab	Good	Good	Good	Moderate	Good	Some
R	Excellent	Excellent	Good	Moderate	Excellent	Yes
Python	Good	Good	Excellent	Moderate	Excellent	Yes
Julia	OK	Excellent	Good	Hard	Excellent	Yes

Language specific editors and IDEs

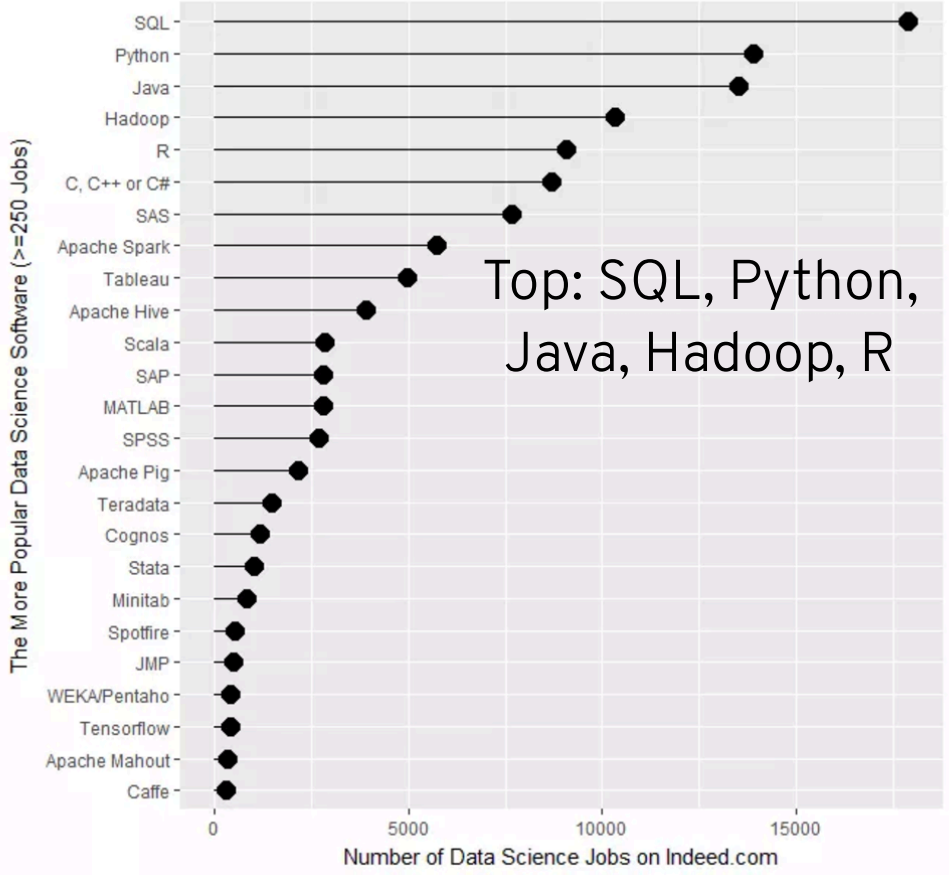
Editor	Features	Ease of use	Language
RStudio	Excellent	Easy	R
Spyder	Excellent	Easy	Python
Stata do file editor	OK	Easy	Stata
SPSS syntax editor	OK	Easy	SPSS

...

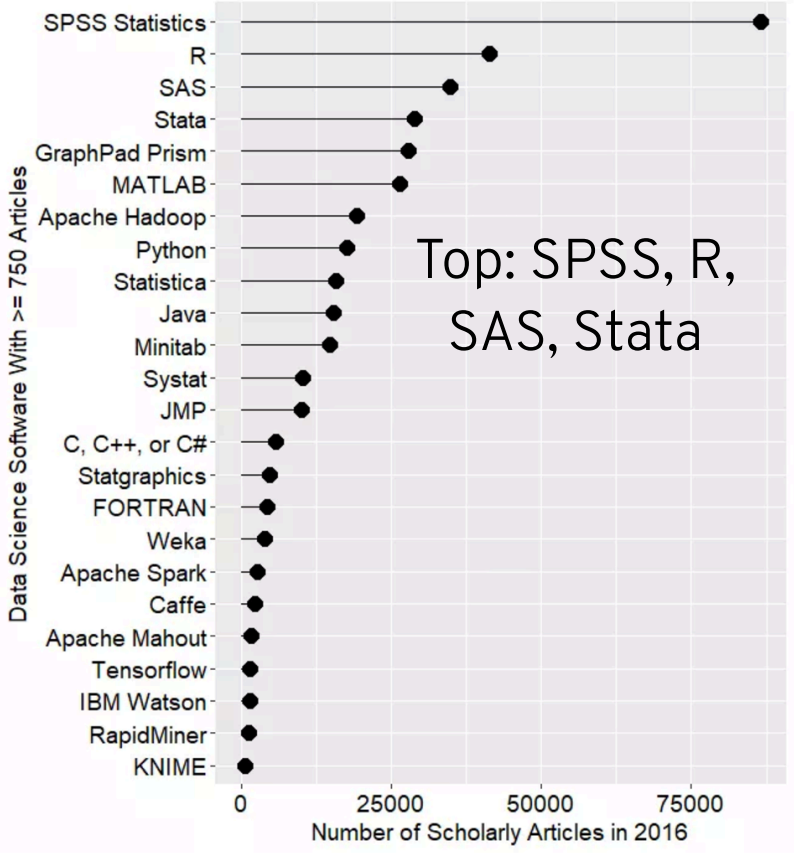
Ista Zhan's Data Science Tools Workshop IQSS:
https://rawgit.com/IQSS/workshops/master/DataScienceTools/DataScienceTools.html

POPULARITY OF TOOLS/LANGUAGES

BASED ON JOB POSTINGS

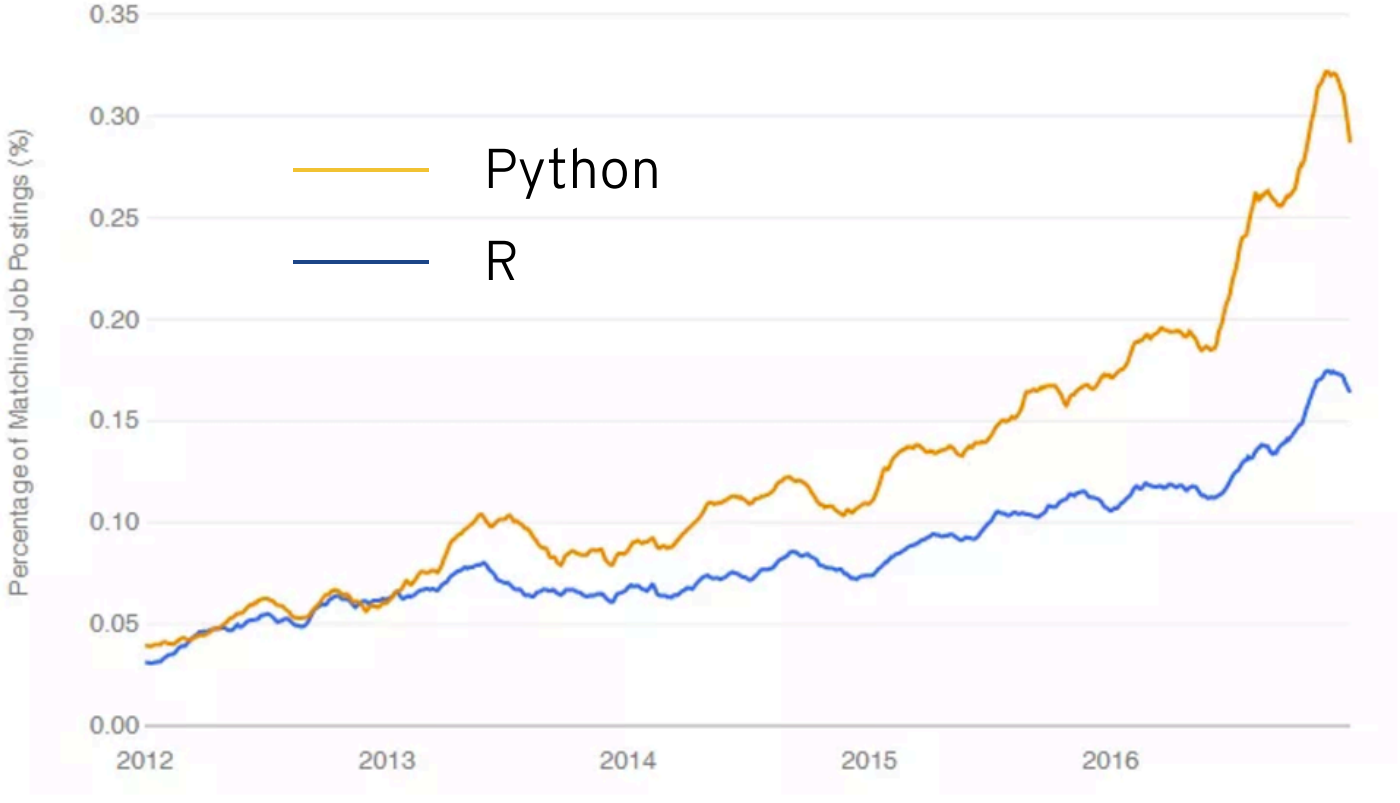


BASED ON SCHOLARLY ARTICLES



Robert Muenchen: http://r4stats.com/articles/popularity

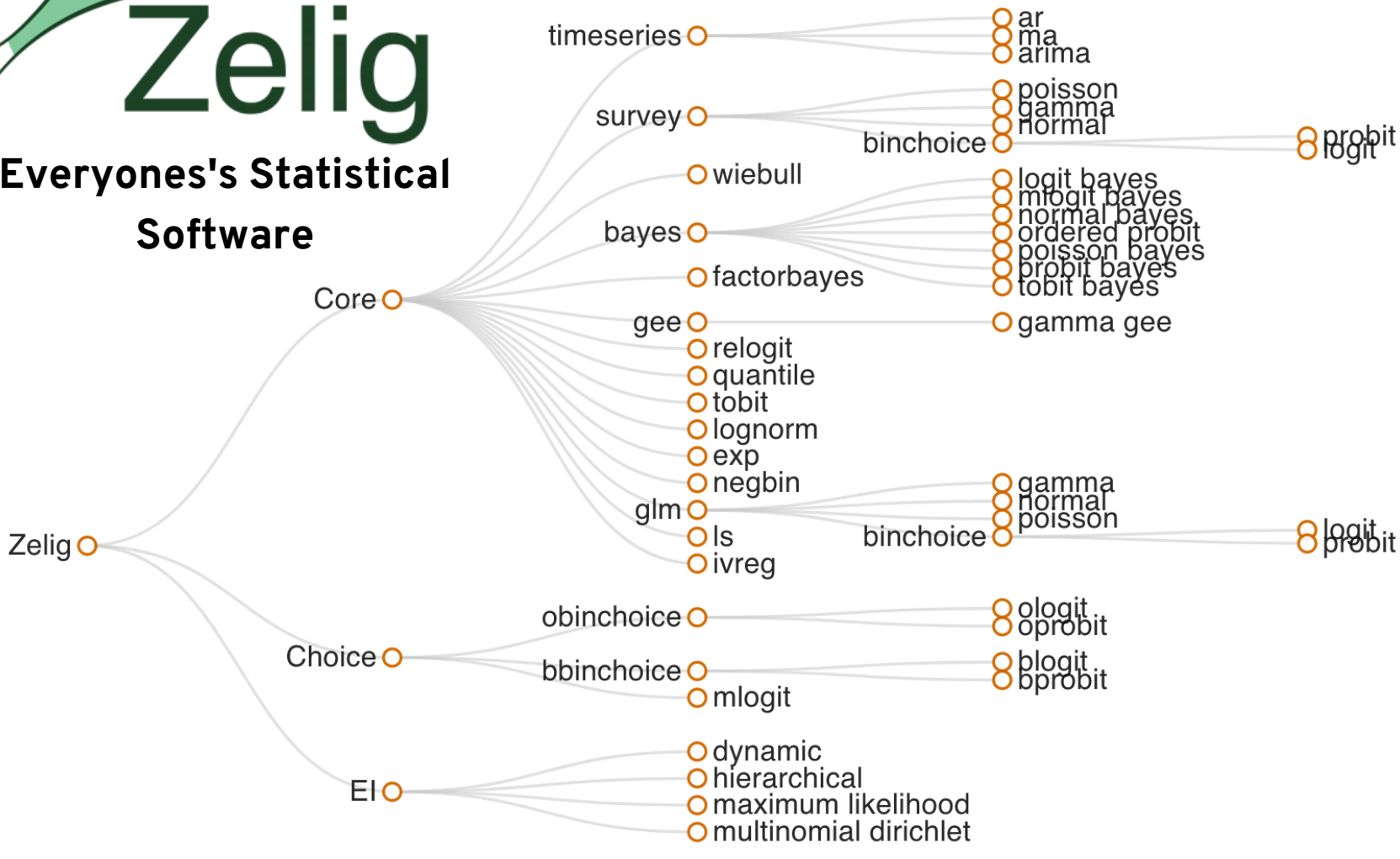
R VS PYTHON BASED ON JOB POSTINGS



Popularity of SPSS, Stata, SAS is decreasing compared to Python, R, and Julia.

Zelig

Everyones's Statistical Software



R Packages developed at Harvard's Institute for Quantitative Social Science (IQSS)

zeligproject.org

BEST PRACTICES

"Convert the raw results of any specific statistical procedure into expressions that:

1) convey numerically precise measurements of the **quantities** of greatest substantive **interest**,
2) include reasonable measurements of uncertainties about those estimates,
3) require little specialized knowledge to understand"

Gary King, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." American Journal of Political Science, 44, Pp. 341–355. Copy at http://j.mp/2n65duA

Science AAAS

Home | News | Journals | Topics | Careers

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE

POLICY FORUM | REPRODUCIBILITY

Enhancing reproducibility for computational methods

Victoria Stodden[1], Marcia McNutt[2], David H. Bailey[3], Ewa Deelman[4], Yolanda Gil[4], Brooks Hanson[5], Michael...

+ See all authors and affiliations

Science 09 Dec 2016:
Vol. 354, Issue 6317, pp. 1240-1241
DOI: 10.1126/science.aah6168

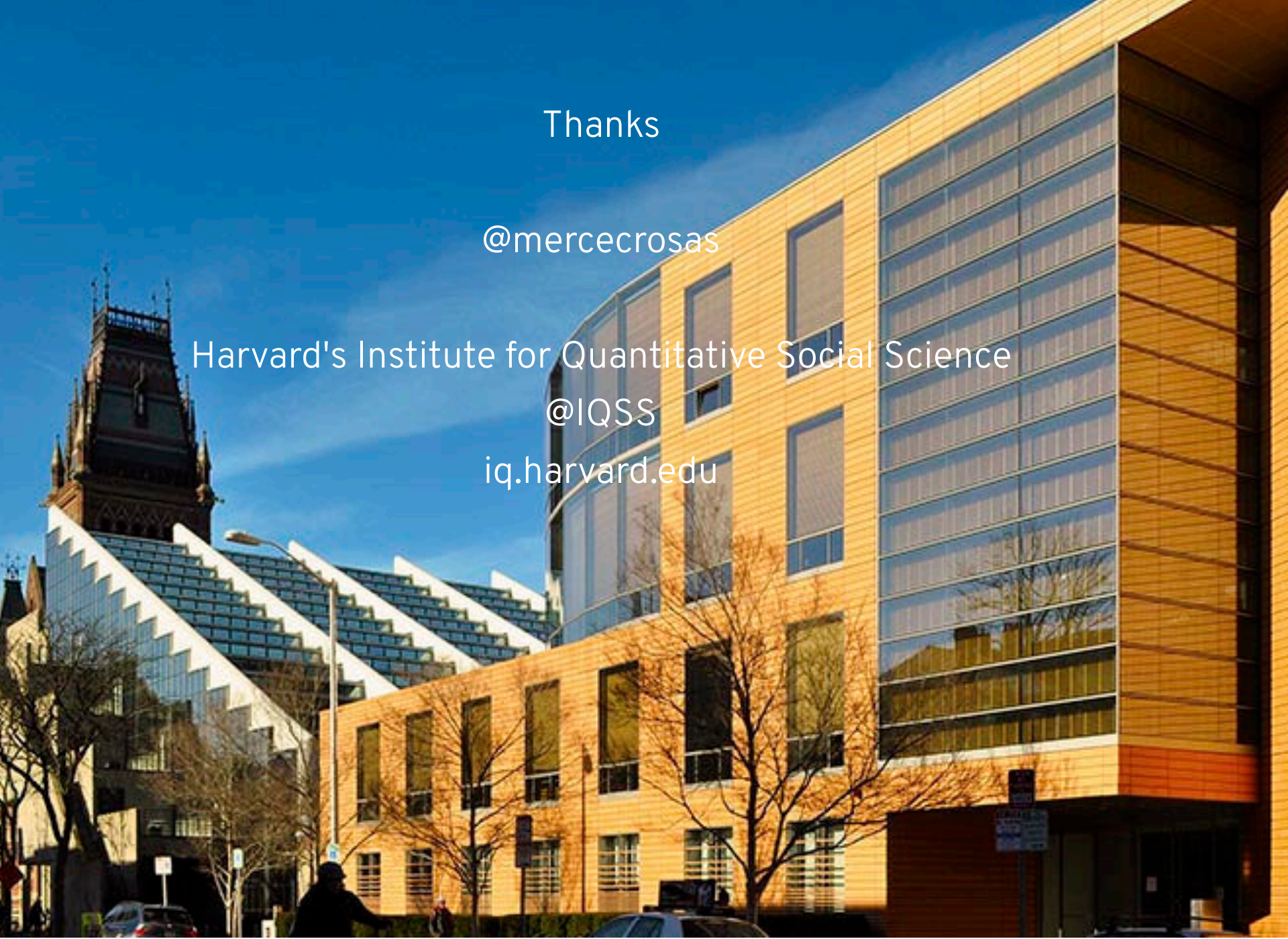
BEST PRACTICES FOR REPRODUCIBILITY

- Share data and code in open trusted repositories
- Use persistent links from publication to data and code
- Citation to data and code should be a standard
- Document data, code, workflows, and computational environment
- Use open license for your code and data

WHEN WRITING METHODS, YOUR CODE SHOULD BE:

1. Informatively documented
2. Open source
3. Comprehensible and automatically tested
4. Developed using version control
5. Stored in a public repository
6. Clearly citable

Christopher Gandrud, from IQSS Software Best Practices workshop


Thanks
@mercecrosas
Harvard's Institute for Quantitative Social Science
@IQSS
iq.harvard.edu