Contemporary challenges in digital social science methodologies

Eetu Mäkelä

This presentation:
http://j.mp/meth4dss-td

Anna Haverinen?

Kansatiede

Technology

- Hopefully more than just word frequencies.
- statistics
- digital methods in learning
- They are overestimating the fame of digihum
- future studies
- Overview of current research projects in DH would be nice
- computational methods for poetry
- Python <3
- Translation technology
- The use of computational methods for humanistic research and sciences
- game theory
- Theory of Digitalisation
- British Mass-Observation data as DH project material
- the history of concepts (Begriffsgeschichte)
- Text mining
- Corpus linguistics
- I don't even know what corpus means
- CAT tools
- simulation of history
- is DH science or just a method?
- UML models & history !?
- network analysis
- machine learning
- semantic vector space modelling
- Technological tools for traditionally non-technological fields
- Data visualization
- What is digital humanities?
- OPTIMIZATION
- Also research of the cultural effects of digital technologies
- Researching natural languages with the help of programming languages
- problems of communication across fields of science
- popularization of science
- Eye-movement research?
- natural language processing
- language technology
- Big Data
- Bringing humanities to this century
- Searching for new questions for humanities
- STUDYING E-BOOKS!
- Keep it simple and practical, please. :)
- interactive study of social systems
- R
- saving time by being efficient
- Not only teaching digital methods to humanists but also vice versa
- the future of humanities
- yes please
- study of digital societies
- Keep it interesting to those who already know something too!
- DH & film studies
- remember that one bill gates quote about giving lazy men hard jobs
- = digital humanities
- DH & semiotics?
- Updating methodologies in humanities
- intersection of humanities and technology/ CS
- Digitalization of OLD resources such as 1000 yo manuscripts
- that old book presence hype though
- we want it all
- My future career

Li et al., 2014: What a Nasty day: Exploring Mood-Weather Relationship from Twitter

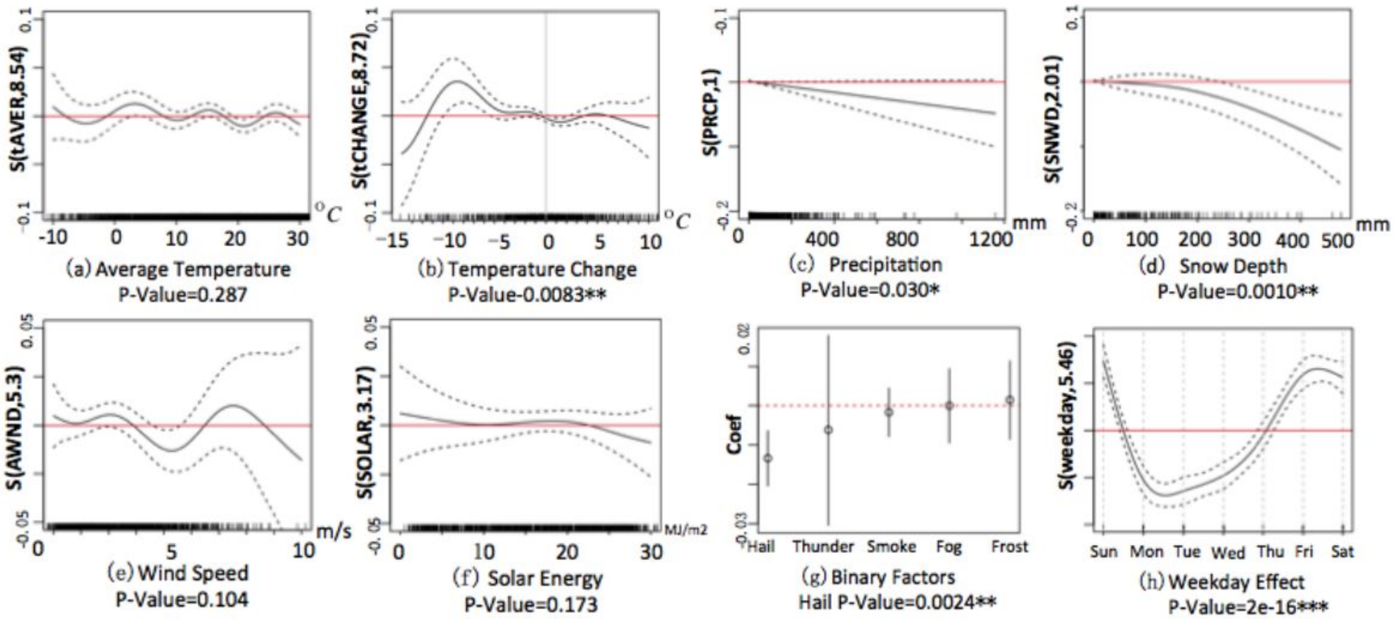


Figure 3: Positive/Negative mode analysis regarding multiple meteorological factors. Red solid line corresponds to 0 line. Black dotted lines correspond to boundary of confidence interval. Black solid line corresponds to regression curve. y-axis corresponds to smooth regression value from GAM model. Positive value of smooth regression means positive contribution to up-mood state while negative value means the opposite. Label for y-axis corresponds to S(meteorological factor, degree of freedom)

Nature **457**, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; Corrected 19 February 2009

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

PNAS, 2014
10.1073/pnas.1320040111

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Significance

We show, via a massive ($N = 689{,}003$) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

PNAS is publishing an Editorial Expression of Concern regarding the following article: "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111). This paper represents an important and emerging area of social science research that needs to be approached with sensitivity and with vigilance regarding personal privacy issues.

Questions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University.

FINAL FINAL

POLICYFORUM

Science, March 2014

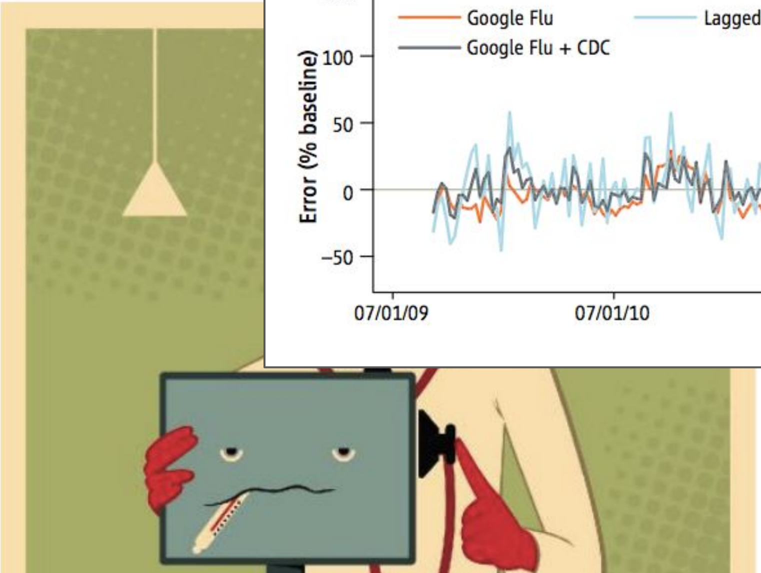
BIG DATA

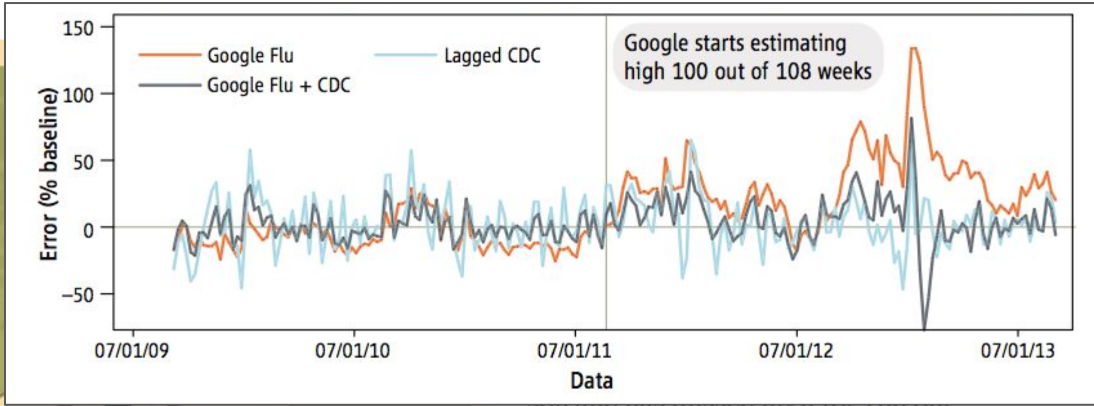
The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held

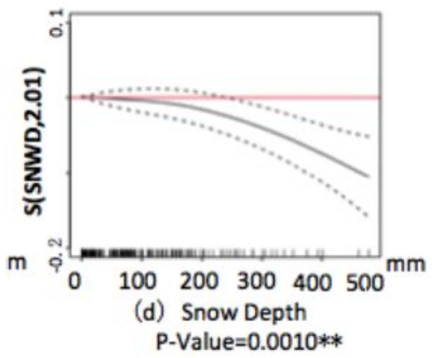


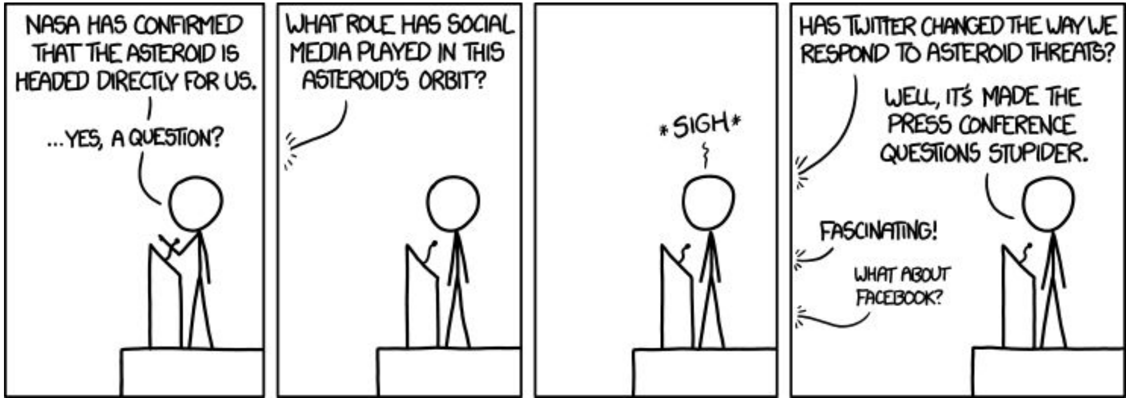


of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal auto-

Cihon & Yasseri, 2016: A Biased Review of Biases in Twitter Studies on Political Collective Action

This literature offers insight into particular social phenomena on Twitter, but often **fails to** use standardized methods that **permit interpretation beyond individual studies**. Moreover, the literature **fails to ground methodologies** and results in social or political theory, **divorcing empirical research from the theory needed to interpret it**. Rather, investigations focus primarily on methodological innovations for social media analyses, but these too often fail to sufficiently demonstrate the validity of such methodologies.

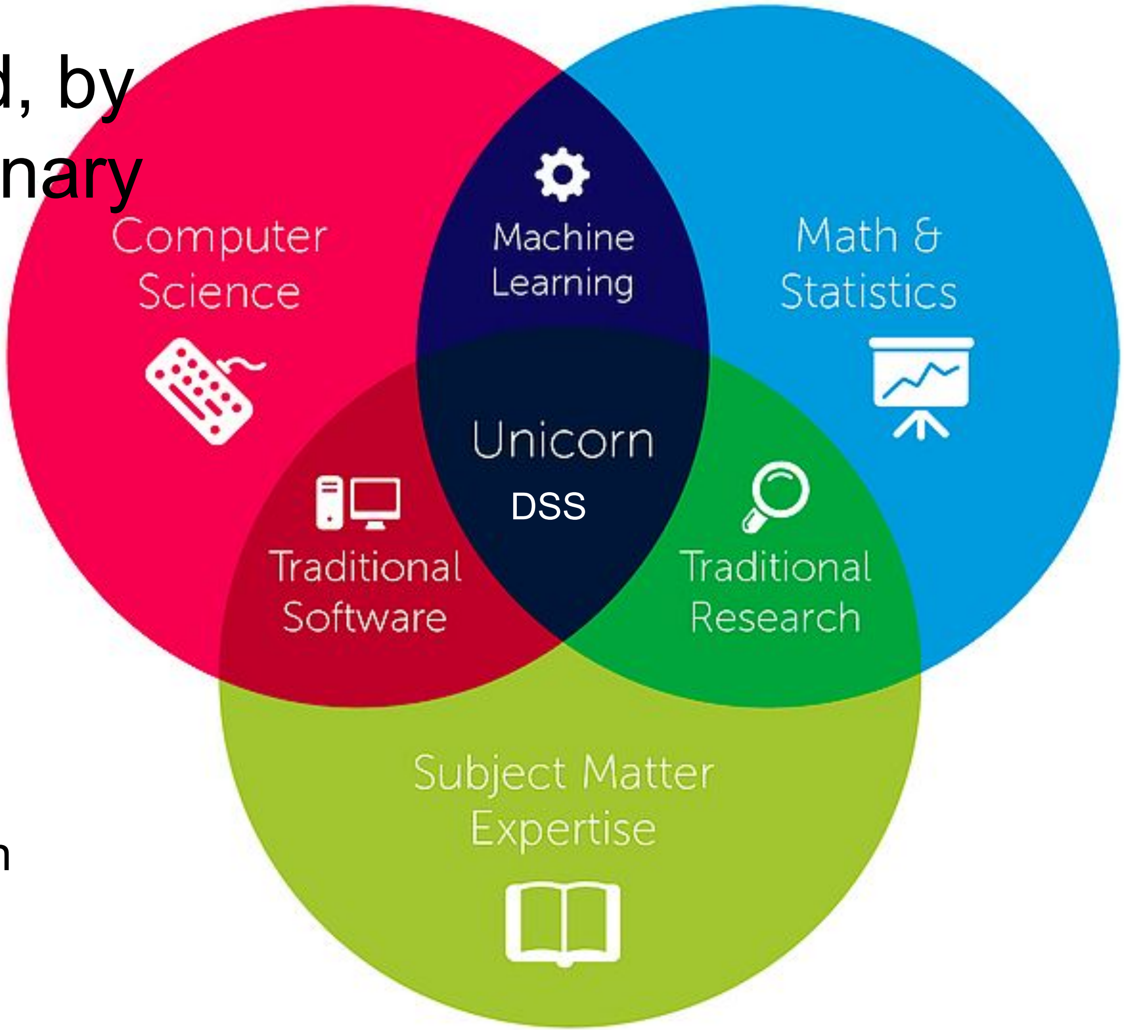




Why does this happen?

DSS is complex, hard, by necessity interdisciplinary

- Data is big, complex and inaccessible
- CS needed to access, process and explore it
- Knowledge of statistics needed to make reliable conclusions
- Social science subject expertise needed to ground results, provide interpretation and ensure depth



DSS is being done without social scientists!

A final challenge for computational social science is that, in spite of many thousands of papers published on topics related to social networks, financial crises, crowdsourcing, influence and adoption, group formation, and so on, **relatively few are published in traditional social science journals or even attempt to engage seriously with social scientific literature**. The result is that much of **computational social science has effectively evolved in isolation** from the rest of social science, largely ignoring much of what social scientists have to say about the same topics, and largely being ignored by them in return.

Duncan J. Watts (Microsoft Research): Computational Social Science: Exciting Progress and Future Directions. The Bridge on **Frontiers of Engineering**, December 20, 2013, Volume 43, Issue 4

Eur. Phys. J. Special Topics **214**, 325–346 (2012)

DOI: 10.1140/epjst/e2012-01697-8

THE EUROPEAN
PHYSICAL JOURNAL
SPECIAL TOPICS

Regular Article

Manifesto of computational social science

R. Conte[1,a], N. Gilbert[2], G. Bonelli[1], C. Cioffi-Revilla[3], G. Deffuant[4], J. Kertesz[5], V. Loreto[6], S. Moat[7], J.-P. Nadal[8], A. Sanchez[9], A. Nowak[10], A. Flache[11], M. San Miguel[12], and D. Helbing[13]

[1] ISTC-CNR, Italy
[2] CRESS, University of Surrey, UK
[3] Center for Social Complexity, George Mason University, USA
[4] National Research Institute of Science and Technology for Environment and Agriculture (IRSTEA), France
[5] Institute of Physics, Budapest University of Technology and Economics, Hungary
[6] Sapienza University of Rome, Italy
[7] University College London, UK
[8] CNRS, France
[9] GISC, Carlos III University of Madrid, Spain
[10] Center for Complex Systems, University of Warsaw, Poland
[11] ICS, University of Groningen, The Netherlands
[12] IFISC (CSIC-UIB), Campus Universitat Illes Balears, 07071 Palma de Mallorca, Spain
[13] ETH Zurich, Switzerland

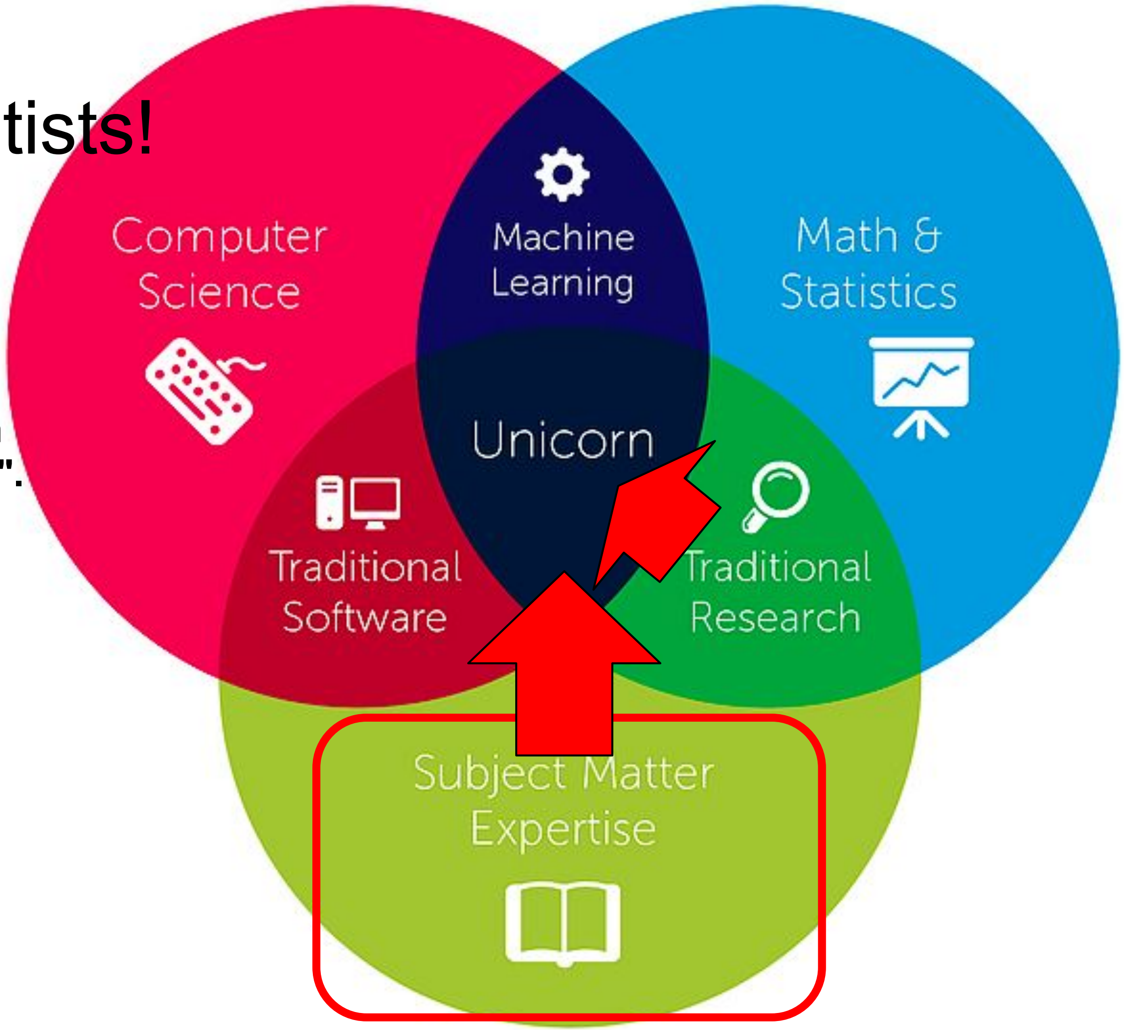
AT THE CROSSROADS: LESSONS AND CHALLENGES IN COMPUTATIONAL SOCIAL SCIENCE

EDITED BY: Javier Borge-Holthoefer, Yamir Moreno and Taha Yasseri

PUBLISHED IN: Frontiers in Physics

Niche for social scientists!

"I have the solution, but it works only in the case of spherical cows in a vacuum".



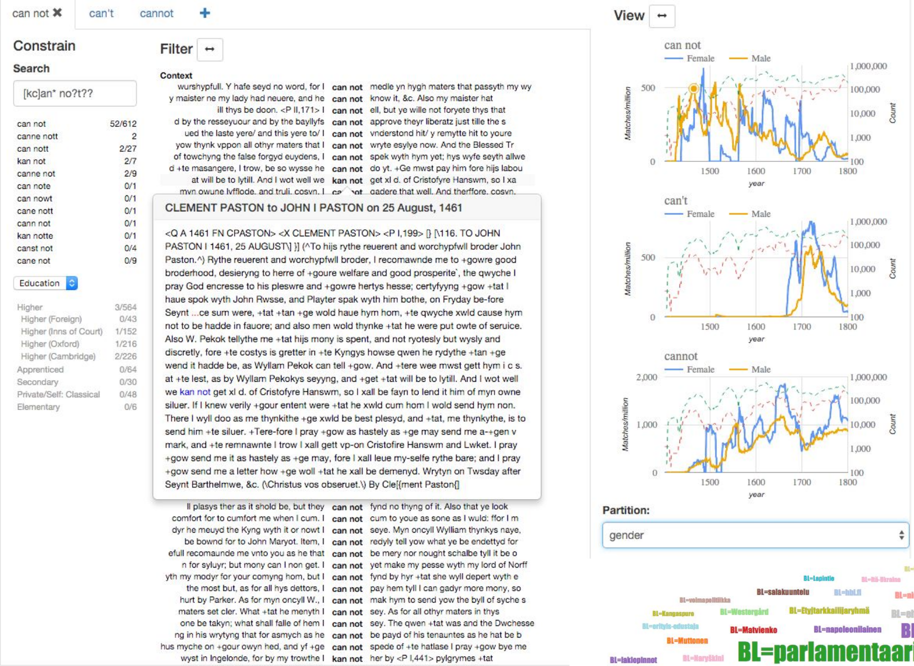
And they know they need you!

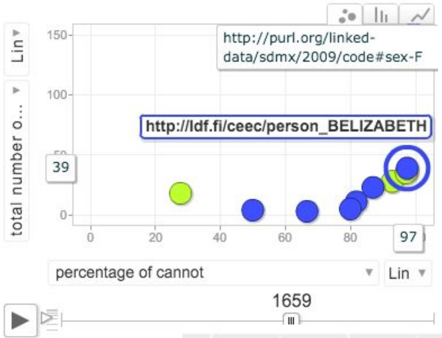
Olemme **fyysikkotaustaisia** Aalto-yliopiston tutkijoita tekemässä hakemusta MATINE:lle koskien aatteiden ja ideologioiden muodostumista ja kehittymistä agenttipohjaisissa simulaatioissa, ja **etsimme hakemukseen halukkaita yhteistyökumppaneita sosiaalitieteiden puolelta**. Lähestymme tutkimusaihettamme sen oletuksen kautta, että ihmisten pääasiallisena viettinä on maksimoida oma "paremmuutensa" sosiaalisessa ympäristössään. Tämä viitekehys on lähellä Adlerin yksilöpsykologian koulukunnan perusajatuksia, ja siinä ideologioita voidaan kuvata tapoina laittaa asiat ja ihmiset arvojärjestyksiin.

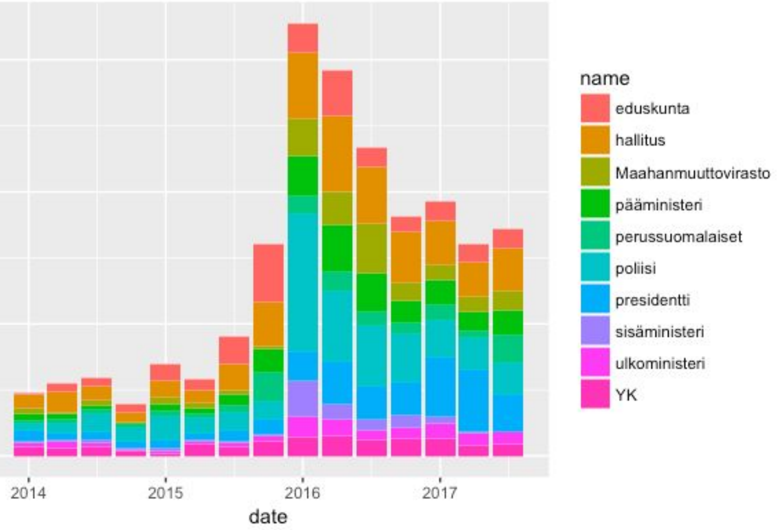
Hakemus on jätettävä viimeistään 14.6.2017, joten toivomme yhteistyötarjouksia mahdollisimman pian, ja pahoittelemme tiukasta aikataulusta mahdollisesti aiheutuvaa vaivaa.
Yhteystiedot: Prof. Kimmo Kaski, kimmo.kaski'at'aalto.fi, FT Jan Snellman, jan.snellman'at'aalto.fi

What to learn?

1. Knowledge of easy to use end-user data processing and exploration tools
    - Easy to use for their intended purpose, but limited
2. Knowledge of the fundamentals concepts of programming
    - Frees you to process your data more efficiently
    - Allows you to more freely apply analyses etc based on ready libraries and tutorials on the Internet
3. High-level understanding of what types of things can be accomplished with advanced CS methods
    - To be able to communicate in **collaborative projects**

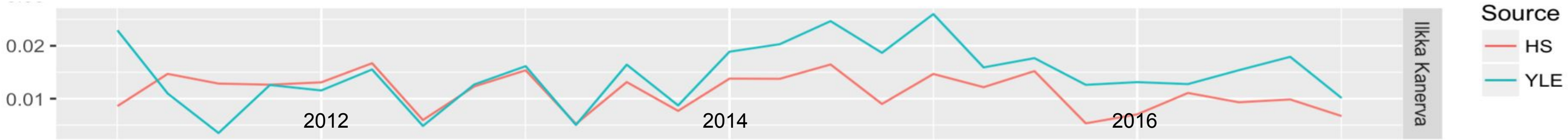






For computer scientists, DSS offers:

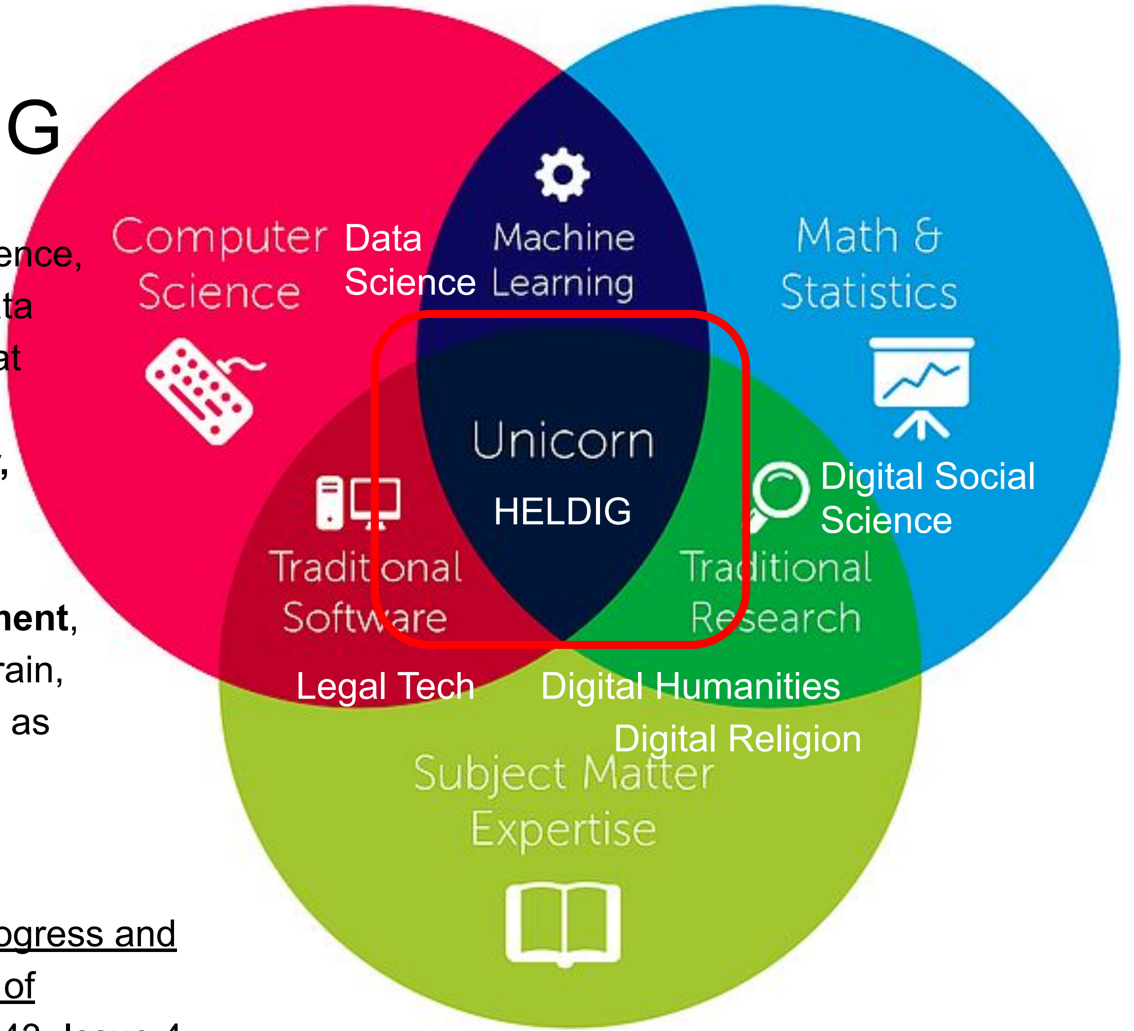
- complex, meaningful challenges
- both in terms of data as well as use cases



Longer term = HELDIG

Deep and significant progress in social science, in other words, will require not only new data and methods but also **new institutions** that are designed from the ground up to foster **long-term, large-scale, multidisciplinary, multimethod, problem-oriented social science research**. To succeed, such an institution will **require substantial investment**, on a par with existing institutes for mind, brain, and behavior, genomics, or cancer, as well as the active cooperation of industry and government partners.

Duncan J. Watts (Microsoft Research): Computational Social Science: Exciting Progress and Future Directions. The Bridge on Frontiers of Engineering. December 20, 2013. Volume 43, Issue 4



What kinds of data do you want us to use for examples?

Cookbooks

New sources of data (eg. Facebook feeds)

The dead and the dying

linguistic data

Tools for transcribing speech

Public documents

Music

Text documents

Skeletons

Musical instruments

Maps

archaeological grave databases etc?

METH4DH - what are you interested in?

Translating tools

Tools for analyzing the use & meaning of concepts

What should the methods make possible?

Analyze video games

Quantitative methods through coding and digitalization

Quantitative methods as a supporting toool for historical research

Distant reading for history studies

Methods of investigating musics

How to apply these methods to social sciences

What kind of tools and methods there are?

how to visualize data

sorting through audio

Methods for archaeology

how to manage resources

For a beginner - give an overview of what is possible already (I feel I cannot imagine it yet)

search through a large amount of articles

will definitely do!

Understand pitfalls of dh tools

Ability to analyze larger amounts of data reduces the risk that something goes unnoticed

Digital tools may increase the efficiency of research

Thanks! Will make me feel less lost ;)

important part of the course

this is what we're after, yes..

METH4DH background questionnaire

Pertinent background information

If you want to tell us more deeply about your study subject or interests, as they relate to the course

Your answer

Why are you taking this course?

Your answer

What would you especially like to learn during this course / where would you like us to focus on?

Your answer

SUBMIT

Never submit passwords through Google Forms.

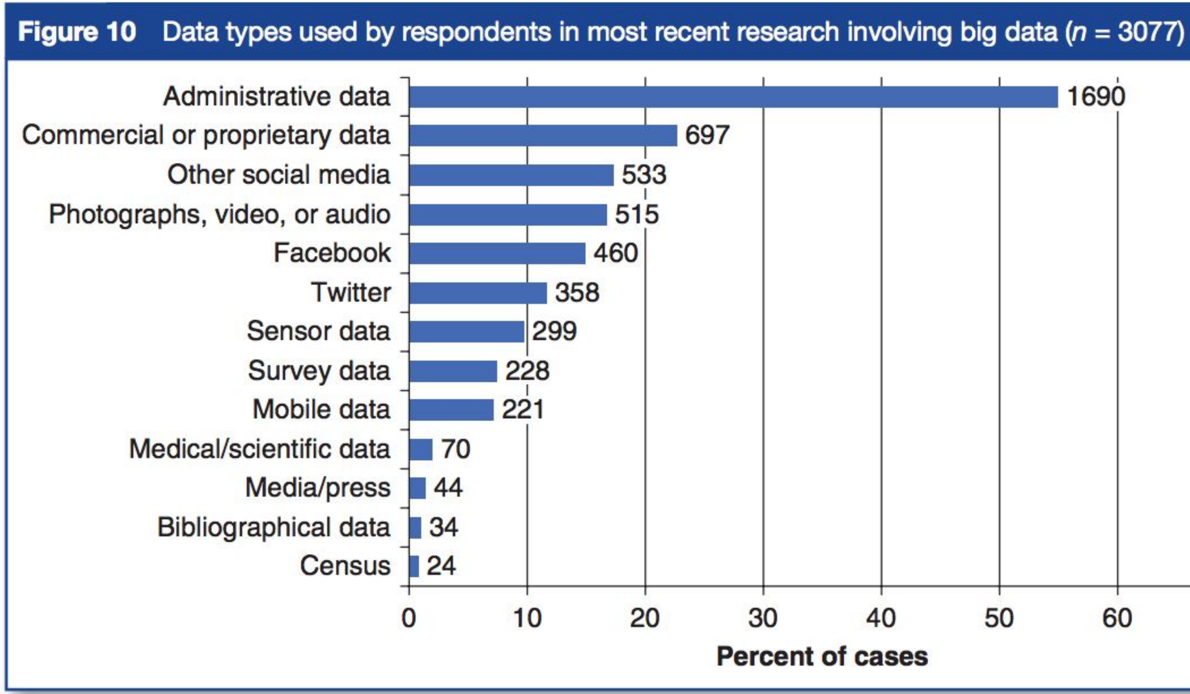
This presentation:
http://j.mp/meth4dss-td

Unused slides follow →

Challenge 1 - access to data

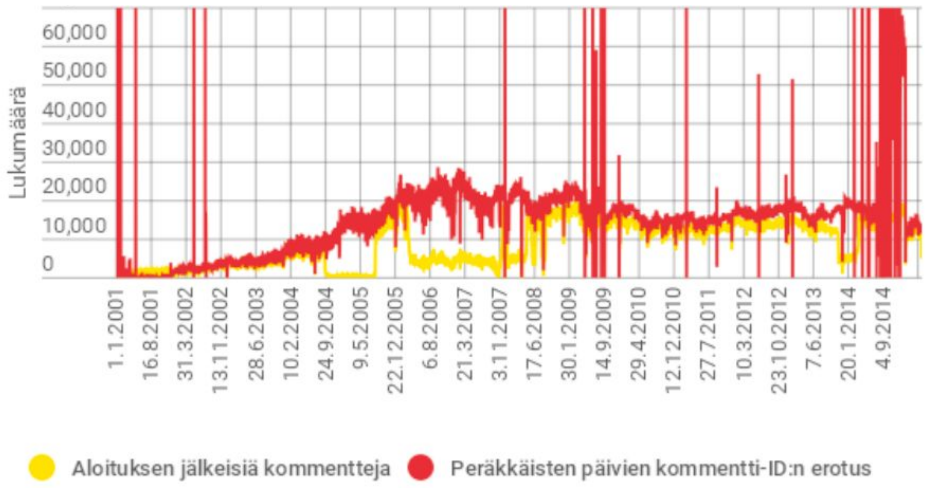
Metzler et al, 2016: Who is Doing Computational Social Science?, SAGE white paper, September 2016

- One of the biggest problems cited by researchers doing big data research was **getting access to commercial or proprietary data**, suggesting that more needs to be done to unlock data sets for social science research.



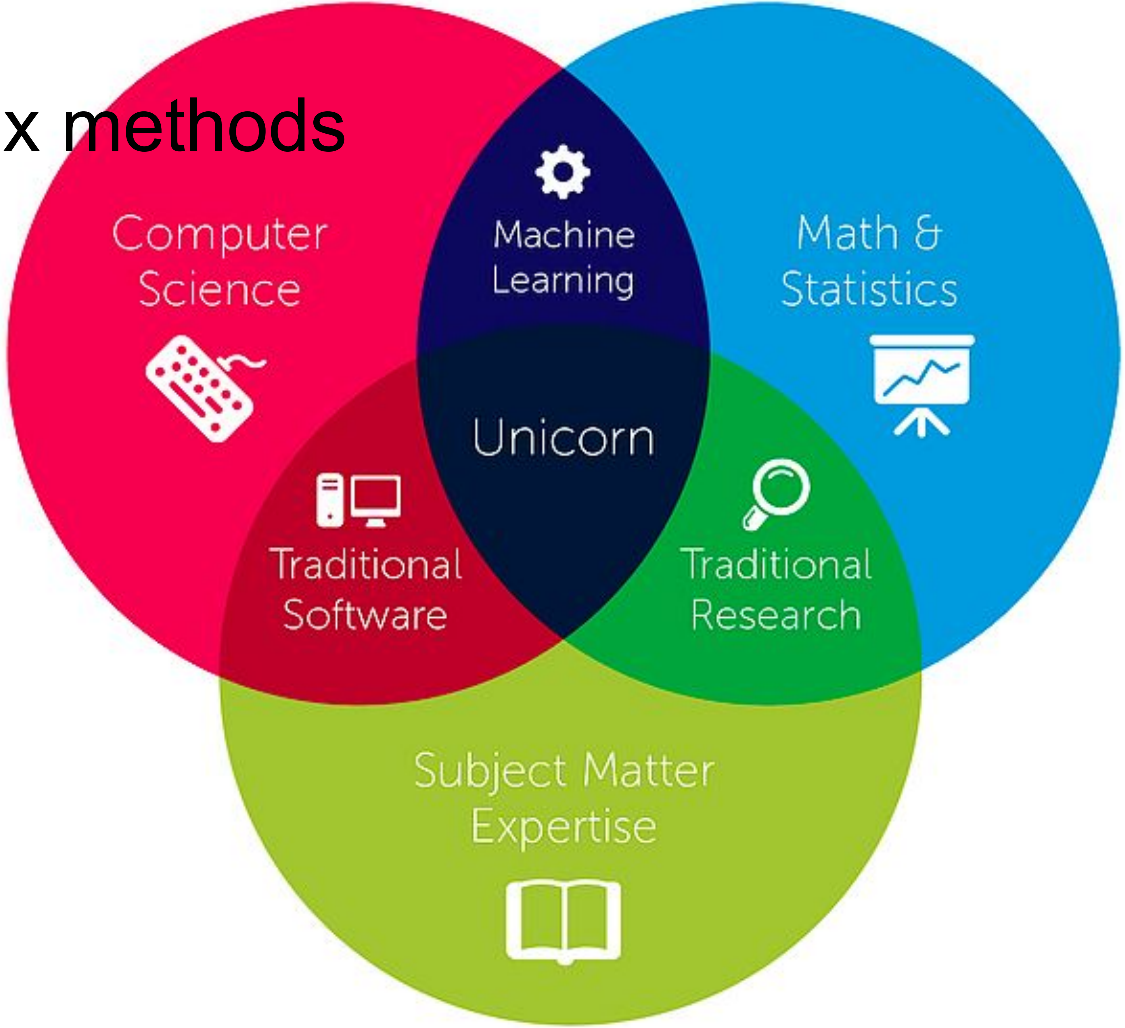
Challenge 2 - complexity of data

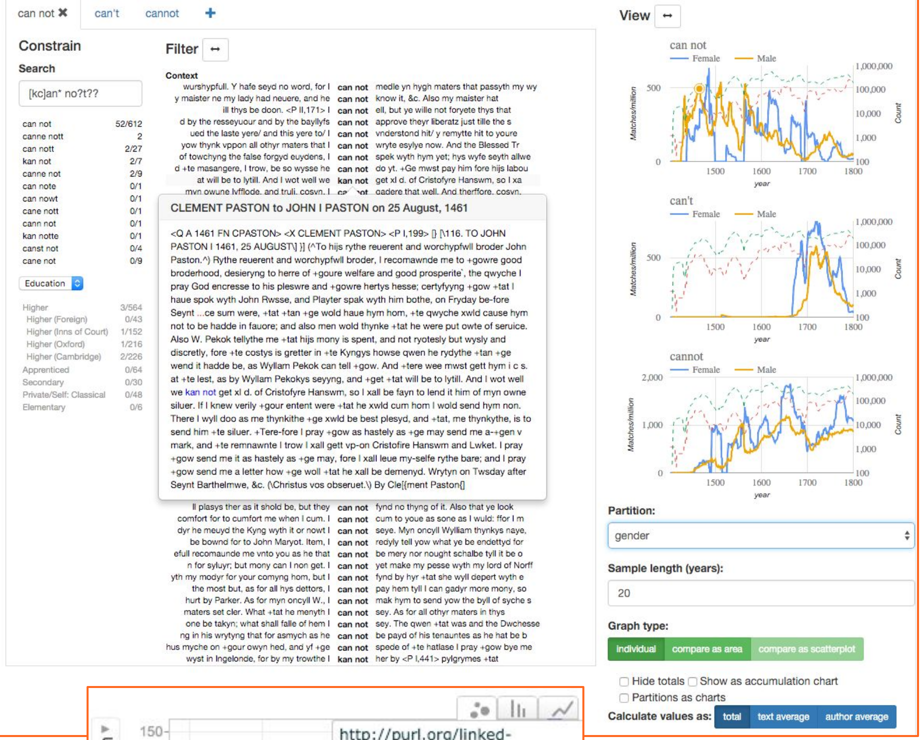
- In the social sciences, the new sources of data … derive overwhelmingly from mixed sources (e.g., social media, unstructured text, digital sensors, financial and administrative transactions) **not designed to produce valid and reliable data** for social scientific analysis (Lazer, Kennedy, King, & Vespignani, 2014), resulting in the **challenge of harmonizing and extracting meaningful features**
- …, **social scientific “big data” are notable less for absolute size per se than for the complexity** that renders conventional methods inadequate (Doorn, 2014).

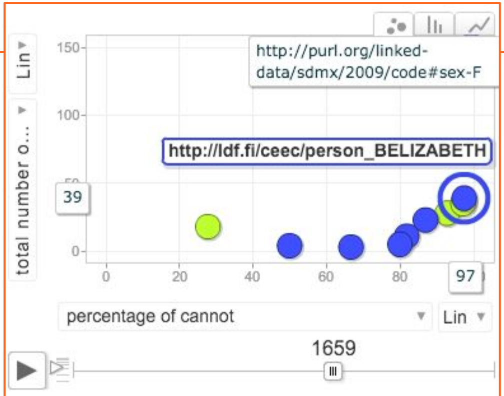


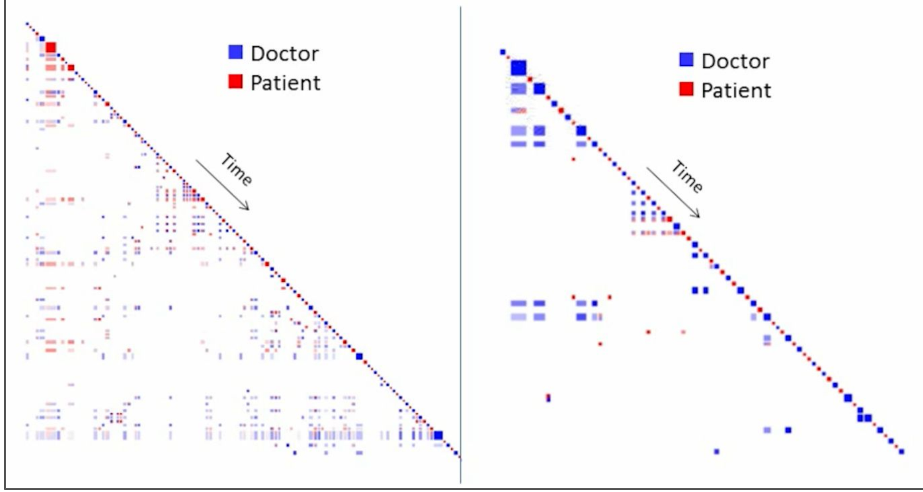
Challenge 3 - complex methods

- Our survey respondents listed finding **collaborators with the right skills** and the **amount of time required to learn** a new field as the biggest barriers to entry.
- A characteristic of researchers doing big data research is that they are more likely to **collaborate with other academics** (79 percent of big data researchers in our survey). Considering that **a large number of social science papers are single authored** (about 40 percent, according to Thomson Reuters (King, 2013), this information is significant.

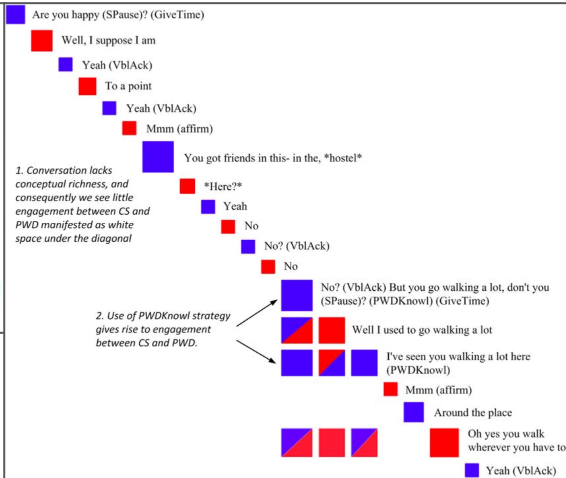



can not
can't
cannot
Constrain
Search
[kc]an* no?t??
can not 52/612
canne nott 2
can nott 2/27
kan not 2/7
canne not 2/9
can note 0/1
can nowt 0/1
cane nott 0/1
cann not 0/1
kan notte 0/1
canst not 0/4
cane not 0/9
Education
Higher 3/564
Higher (Foreign) 0/43
Higher (Inns of Court) 1/152
Higher (Oxford) 1/216
Higher (Cambridge) 2/226
Apprenticed 0/64
Secondary 0/30
Private/Self: Classical 0/48
Elementary 0/6
Filter
Context
CLEMENT PASTON to JOHN I PASTON on 25 August, 1461
View
can not
can't
cannot
Female
Male
Matches/million
Count
year
Partition:
gender
Sample length (years):
20
Graph type:
individual
compare as area
compare as scatterplot
Hide totals
Show as accumulation chart
Partitions as charts
Calculate values as:
total
text average
author average


http://purl.org/linked-data/sdmx/2009/code#sex-F
http://ldf.fi/ceec/person_BELIZABETH
total number o...
Lin
39
97
percentage of cannot
1659


Doctor
Patient
Time
Doctor
Patient
Time


ideas
people
animals
doing
Utterance Pair: 2, 12
Channel: Peter
Similarity: 0.284
Similar Concepts: doing, life, whole, people

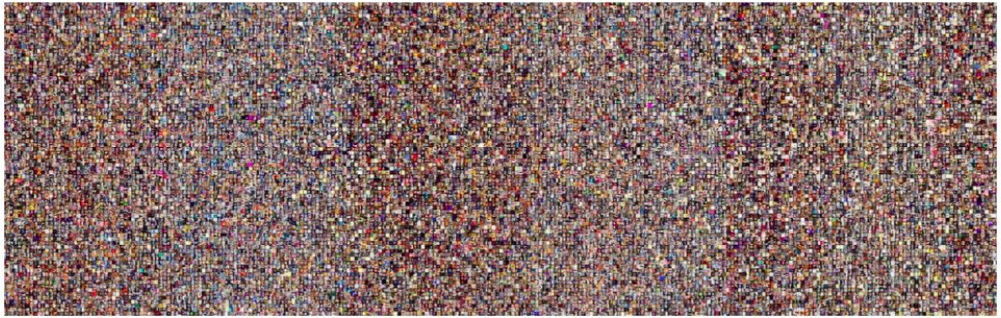

Are you happy (SPause)? (GiveTime)
Well, I suppose I am
Yeah (VblAck)
To a point
Yeah (VblAck)
Mmm (affirm)
You got friends in this- in the, *hostel*
Here?
Yeah
No
No? (VblAck)
No
No? (VblAck) But you go walking a lot, don't you (SPause)? (PWDKnowl) (GiveTime)
Well I used to go walking a lot
I've seen you walking a lot here (PWDKnowl)
Mmm (affirm)
Around the place
Oh yes you walk wherever you have to
Yeah (VblAck)
1. Conversation lacks conceptual richness, and consequently we see little engagement between CS and PWD manifested as white space under the diagonal
2. Use of PWDKnowl strategy gives rise to engagement between CS and PWD.

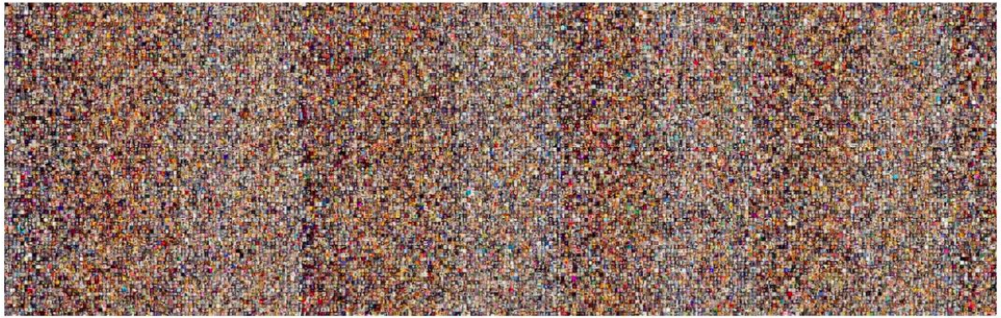

INTERNATIONAL CONFERENCE ON COMPUTATIONAL SOCIAL SCIENCE
June 8- 11, 2015
Finlandia Hall, Helsinki, Finland
IC^2S^2

International Conference on Computational Social Science Luminaries

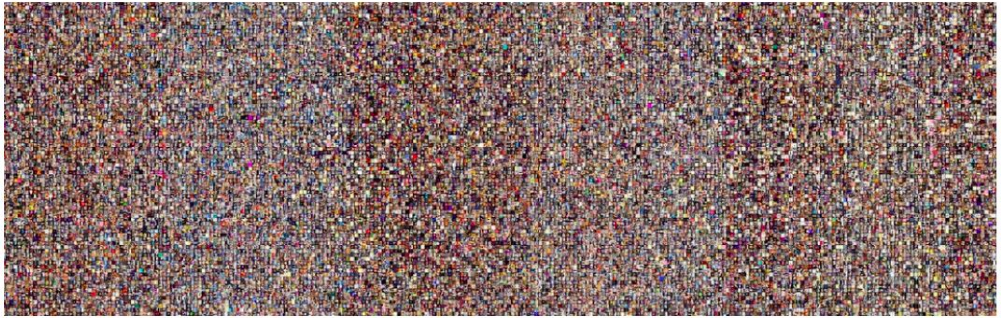
- Santo Fortunato is **Professor of Complex Systems** at the Department of Biomedical Engineering and Computational Science
- Lada Adamic is a **computational social scientist** at Facebook and previously an **associate professor at the School of Information and the Center for the Study of Complex Systems**
- Albert-László Barabási directs the **Center for Complex Network Research**, and holds appointments in the **Departments of Physics and College of Computer and Information Science**
- Nicholas Christakis MD, PhD, MPH, is a **social scientist and physician**
- Alessandro Vespignani is the Sternberg Distinguished **Professor of Physics, Computer Science and Health Sciences**
- Dirk Helbing is **Professor of Sociology**, in particular of Modeling and Simulation, at the Department of Humanities, Social and Political Sciences and member of the Computer Science Department at ETH Zurich. He earned a **PhD in physics**…

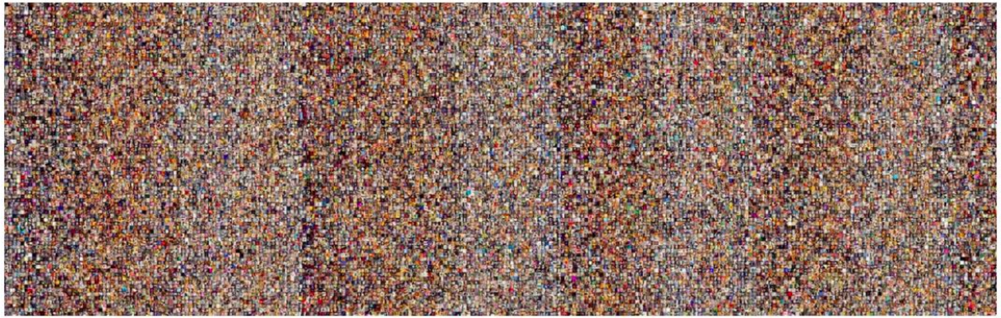
Indaco & Manovich, 2016: Urban Social Media Inequality: Definition, Measurements, and Application





Indaco & Manovich, 2016: Urban Social Media Inequality: Definition, Measurements, and Application





- Social media inequality of visitors' images in Manhattan (Gini = 0.669) is larger than income inequality of most unequal country in the world (Seychelles where Gini = 0.658).
- On the other hand, social media shared by locals has a Gini coefficient similar to countries that rank between 25 and 30 in the list of countries by income inequality. These are countries like Costa Rica (0.486), Mexico (0.481) and Ecuador (0.466). (The World Bank, 2015).

Since Instagram did not support downloading large volumes of historical data, we had to download data and images continuously during the period we wanted to cover. A single iMac computer running 24/7 continuously was used for downloading this data.

Solutions to data issues

- Be at Facebook
- Do local stuff
- Make the peculiarity of the data an asset, a part of the research
- Be opportunistic

Research process

1. Have data
2. Magic (?)
3. Something interesting shows up
4. Profit!

Research process

1. Have data
2. Magic (?)
3. Something interesting shows up
4. Profit!

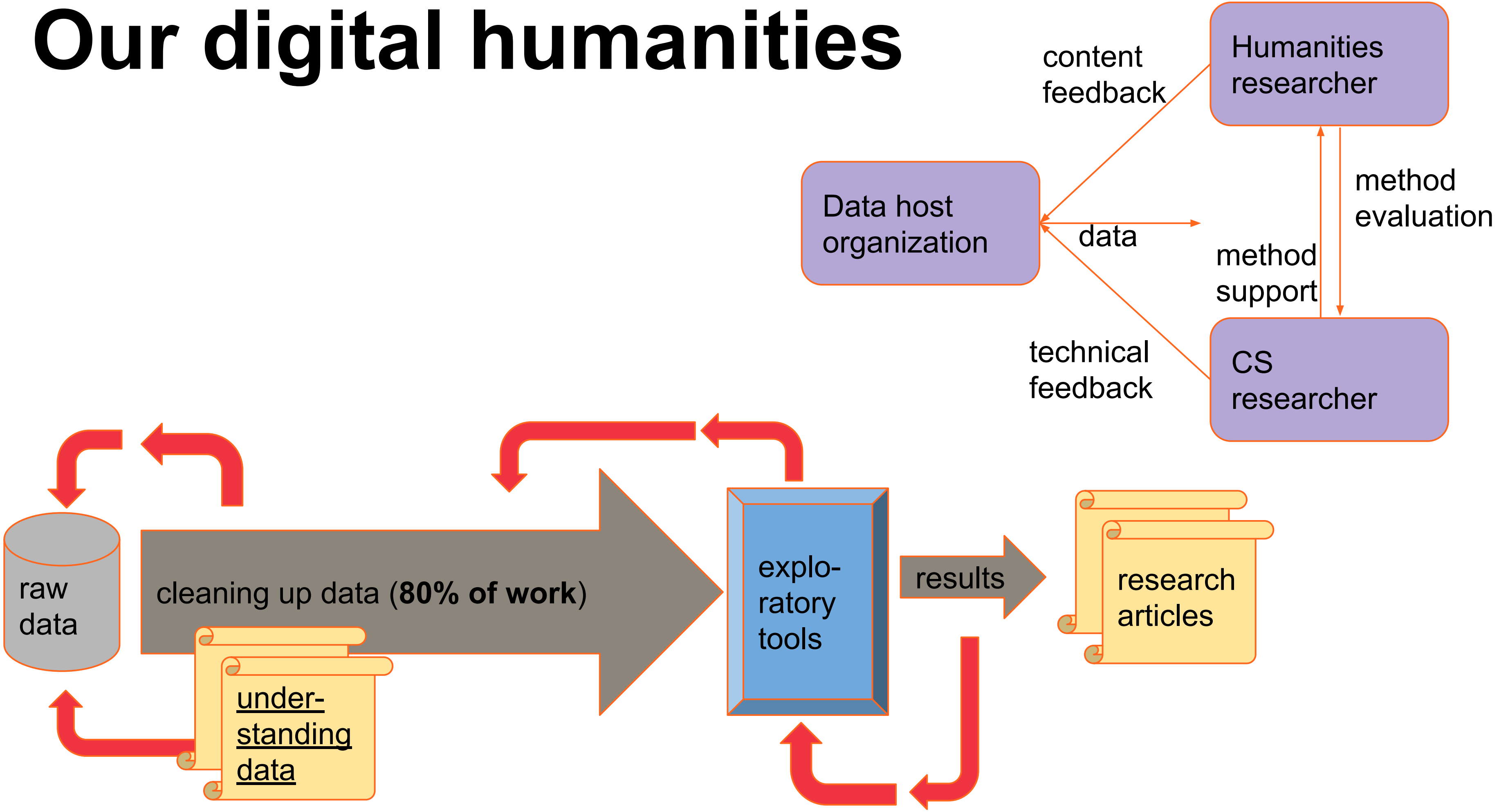
“Any sufficiently advanced technology is indistinguishable from magic.”

- Arthur C. Clarke

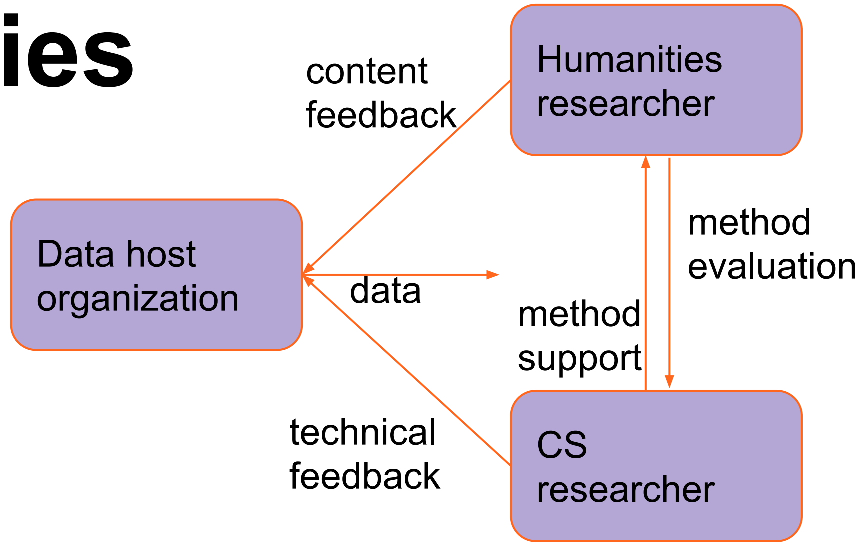
Research process

1. Have data
2. Magic (?)
   a. Hedge magic (spreadsheets, Excel graphs)
   b. Common ritual magic (statistics: correlation, ANOVA, PCA)
      - Relatively simple, commonly understood formulae you could mostly go through with pen and paper if you wanted to
   c. Higher ritual magic (SVM, LSA, LDA, SnE)
      - More complex, harder to follow formulae, impossible to work through manually
      - Well-grounded black box oracles (e.g. you feed a machine learning algorithm stuff, it processes it based on complex but well-defined rules, out comes results)
   d. Black magic (Deep learning)
      - True black box oracles (you feed a neural network both an input and a desired output, it derives mostly unintelligible black box rules that link the two)
   e. Flashy magic (proper visualizations)
3. Something interesting shows up

Our digital humanities



Our digital humanities



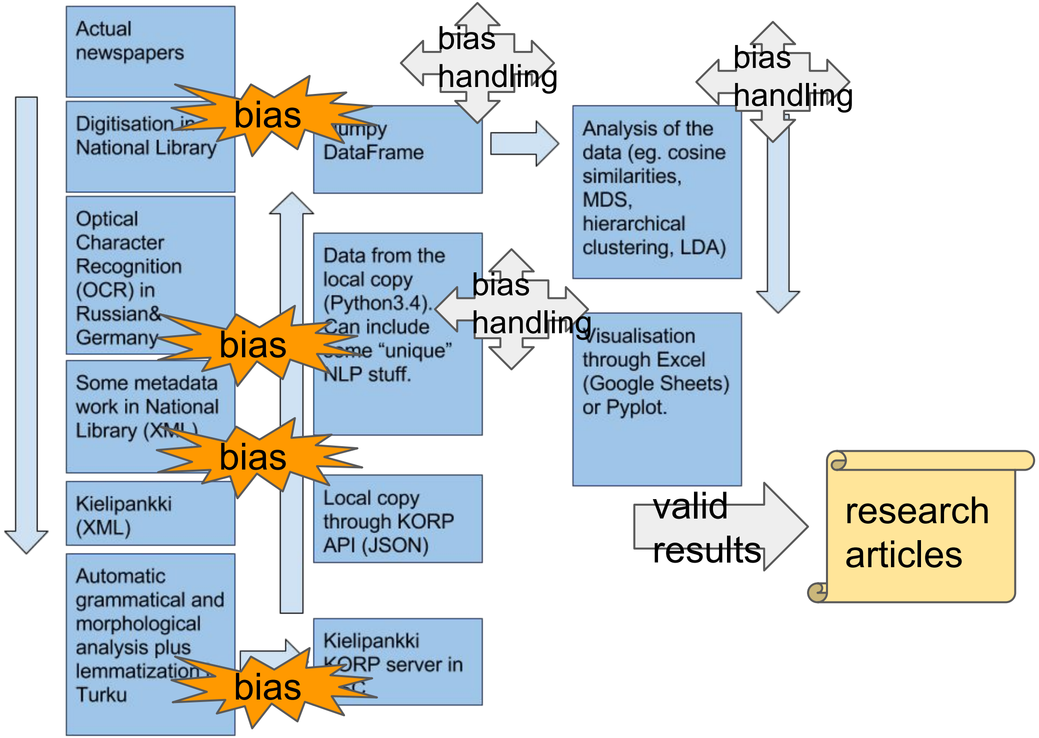
At its best, such close collaboration offers **benefits for everyone involved**

- scholars in the humanities are able to tackle **questions too labour-intensive for manual study**
- computer scientists encounter **new and challenging use cases** for the tools and algorithms they develop
- data providers **gain insight** into their own data

Don’t get carried away by fancy methods!

1. Your dataset must be applicable to the methods you choose. Complex methods often make presuppositions about the data they apply to - if you don’t understand these deeply, you’ll end up with invalid results
2. In typical DH research, 90% of your time will go to gathering and understanding the data and transforming it into a form you can use - using complex methods, another 90% of your time may go to altering them to fit your data, and it’ll run out
3. Complex methods are often unnecessary for DH work. On the contrary, often simpler methods are actually better.

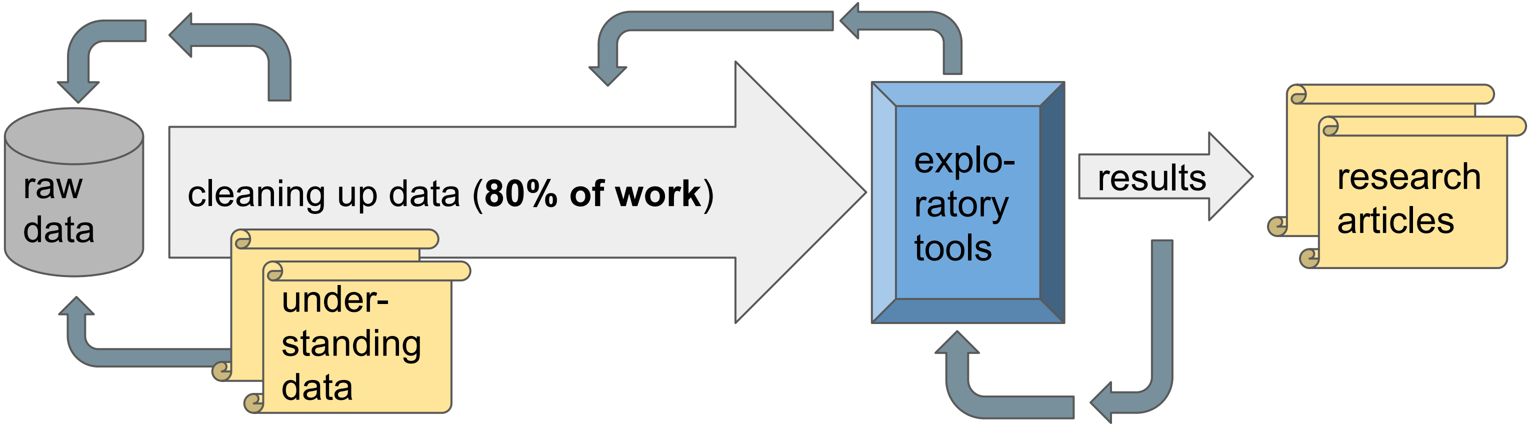
KLK Newspaper Pipeline: from archives to a hypothetical researcher



Our digital humanities

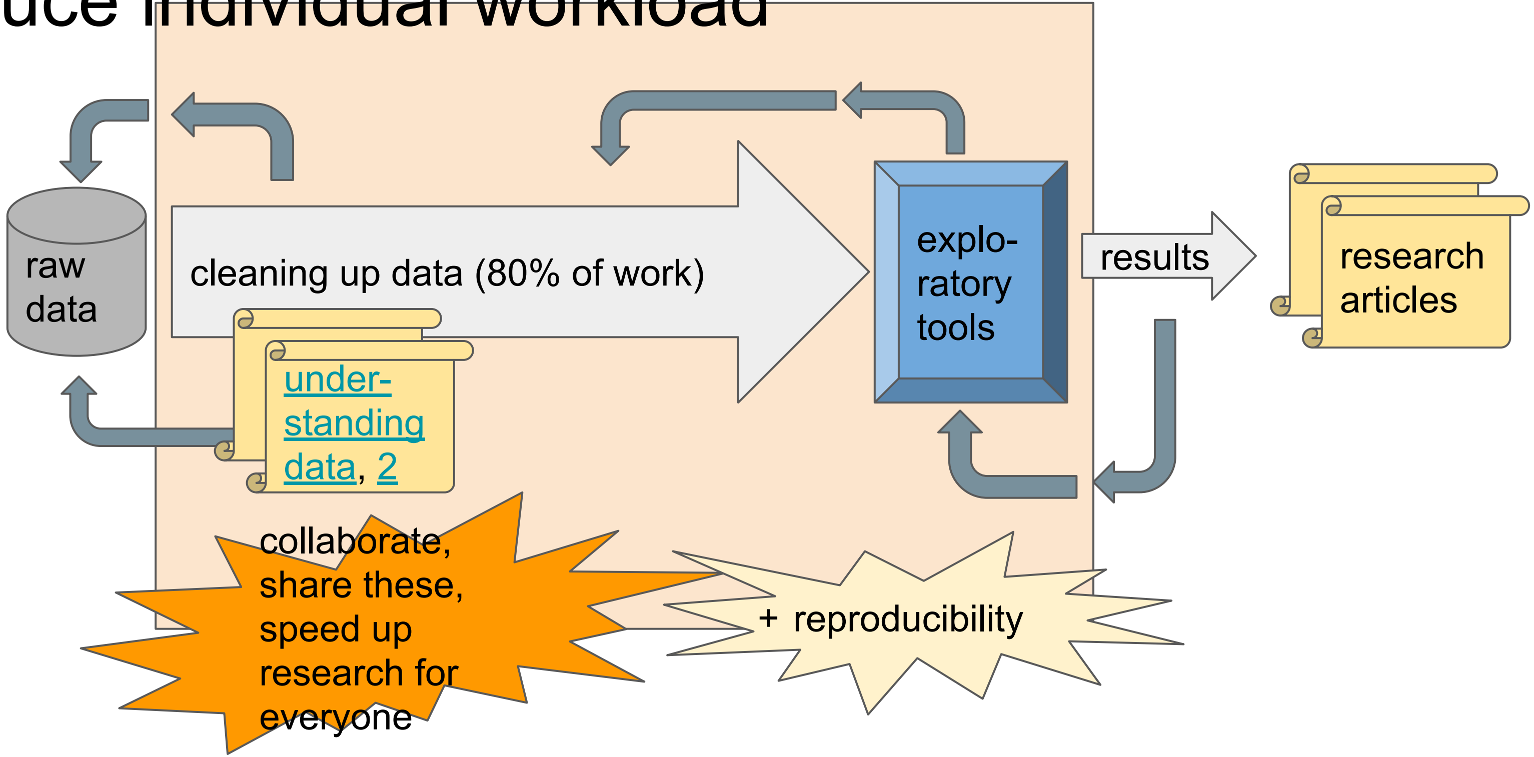
- Scholars in the humanities and computer sciences collaborating, **applying novel computer science** to solve **humanities research questions**

Digital humanities research process



80% of your time for data cleanup, another
80% for algorithms, …

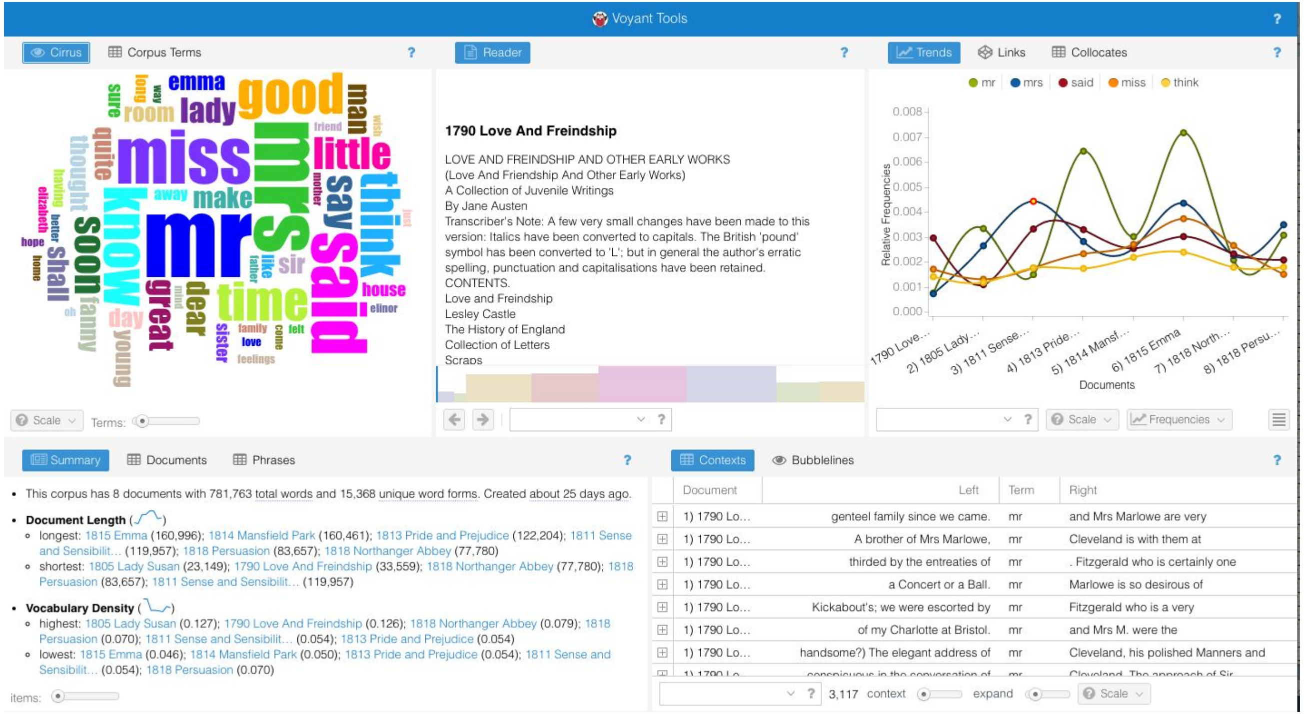
Leverage collaboration, open science workflows to reduce individual workload



Workflow/Tools

1. Data access
2. Possible preprocessing: R, Python, tm (for texts), OpenRefine, …
3. Zero or more of:
   - Statistics: R, stats, pandas, …
   - Topic modeling: Mallet, topicmodels, LDAvis, gensim, … (for texts)
   - Dimensionality reduction/clustering: stats, lsa, BayesLCA, pvclust, Weka, … (also for texts)
   - Social network analysis: igraph, sna, statnet, sonia, Gephi, …
   - Simulation: NetLogo, …
   - Neural networks: som, TensorFlow™, … (also for texts)
   - Association rule learning: arules, Weka, …
   - Anomaly detection: AnomalyDetection, …
4. Structured visualization: Tableau, Palladio, RAW, nodegoat, matplotlib, ggplot2, iPlots, plot.ly, Leaflet, Gephi, CartoDB, … or text visualization: Voyant Tools, Textexture, Wordsift, …

Voyant tools



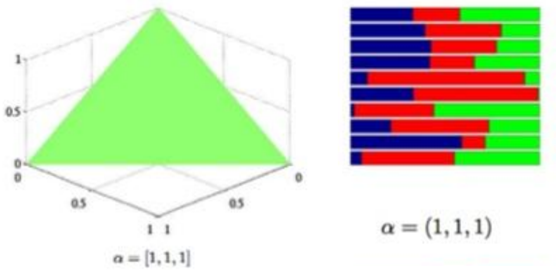
Types of data

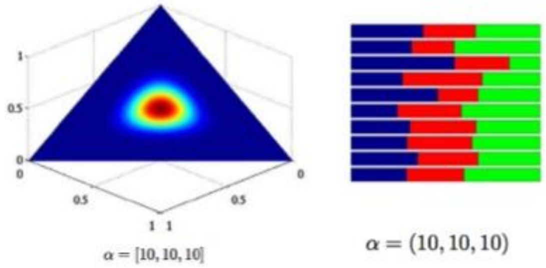
- Structured (databases) vs unstructured (text, image, video, audio)
- Clean vs messy
- **Biased? <- incomplete, messy, badly sampled**

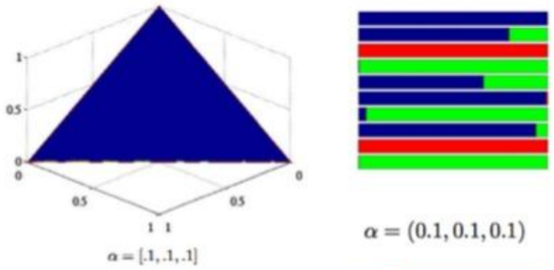
Topic Modeling: LDA - Assumptions

- A document collection contains N topics
- A single document can consist of multiple topics (e.g. 30% war and 70% cooking)
- The N topics are in essence probability distributions over words (e.g. there is a 1,5% chance that a random word from a ‘war’ topic is ‘attack’, while only a 0,00001% chance in a ‘cooking’ topic)
- There are two distributions that give the prior probabilities of:
    a. the probability of topic mixes in documents (e.g. how likely is it that a single document talks about all the topics vs. only a few) , and
    b. the probability mix of words in a topic (e.g. do individual topics mainly contain many words or just a few)

Topic Modeling: LDA - Role of (symmetric) priors







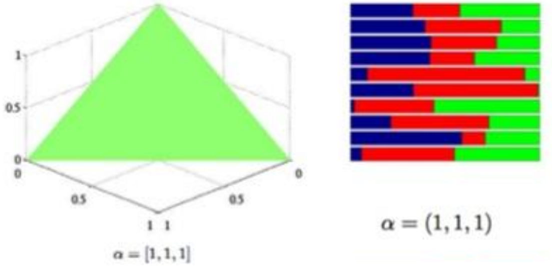
Topic Modeling: LDA - How it works

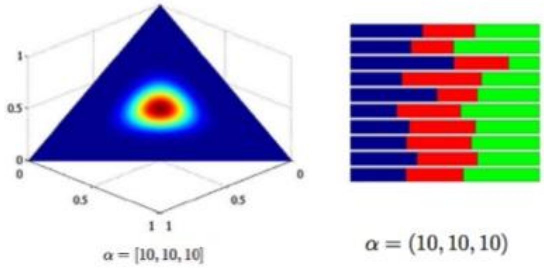
- Take all words and documents and randomly assign them to topics (based on the prior distributions)
- Calculate the combined probability of this combination producing the documents we have
- Update the topic assignments as well as the prior distributions so the probability increases
- Repeat many many times until we’re happy

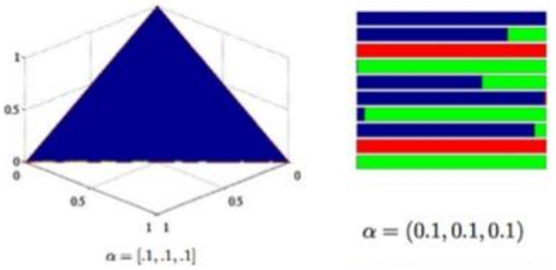
LDA in Practice

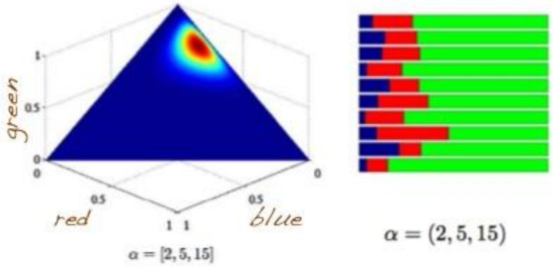
```
corpus <-
VCorpus(DirSource("/srv/data/varieng/ceec-subcorpora/scot-17
00-1719/"))
corpus <- tm_map(corpus,content_transformer(tolower))
corpus <- tm_map(corpus,removeNumbers)
corpus <- tm_map(corpus,removePunctuation)
corpus <- tm_map(corpus,removeWords,stopwords("SMART"))
corpus <- tm_map(corpus,stripWhitespace)
numtopics <- 20
lda <- LDA(DocumentTermMatrix(corpus), numtopics)
```

Topic Modeling: LDA - Role of priors









Topic Modeling: LDA - Effect of priors

- Traditional LDA supposed uniform priors
- Turns out non-uniform priors make sense for how topics appear in documents, but not for how words appear in topics

  → as-LDA, which also turns out to need less pre-filtering of e.g. stopwords, numbers, because these can be sequestered into a common topic without constraining how other topics appear

Figure 6 Primary discipline of respondents who have been involved in big data research (n = 9195)

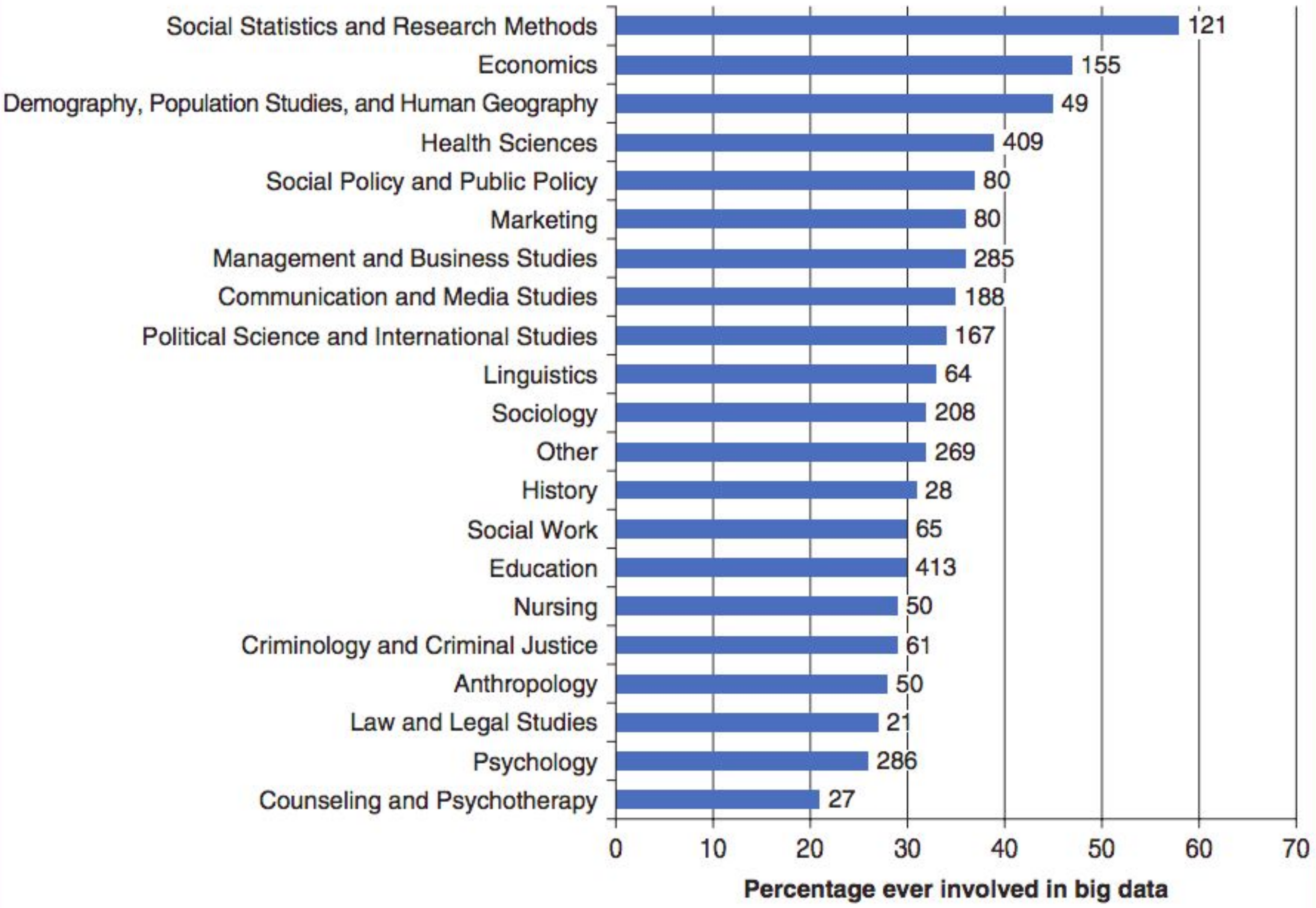


Figure 15 Challenges facing big data researchers ($n = 2273$)

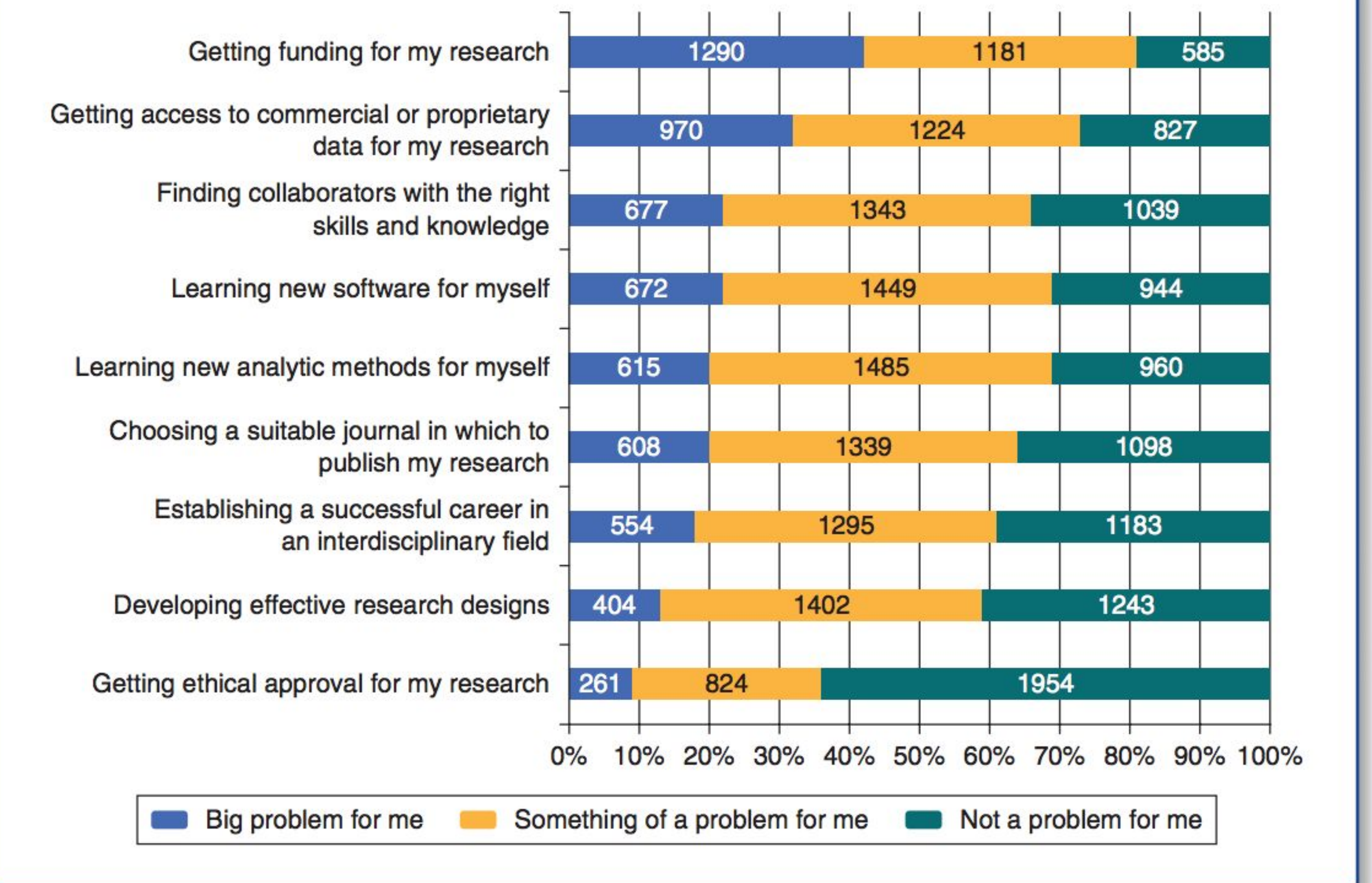
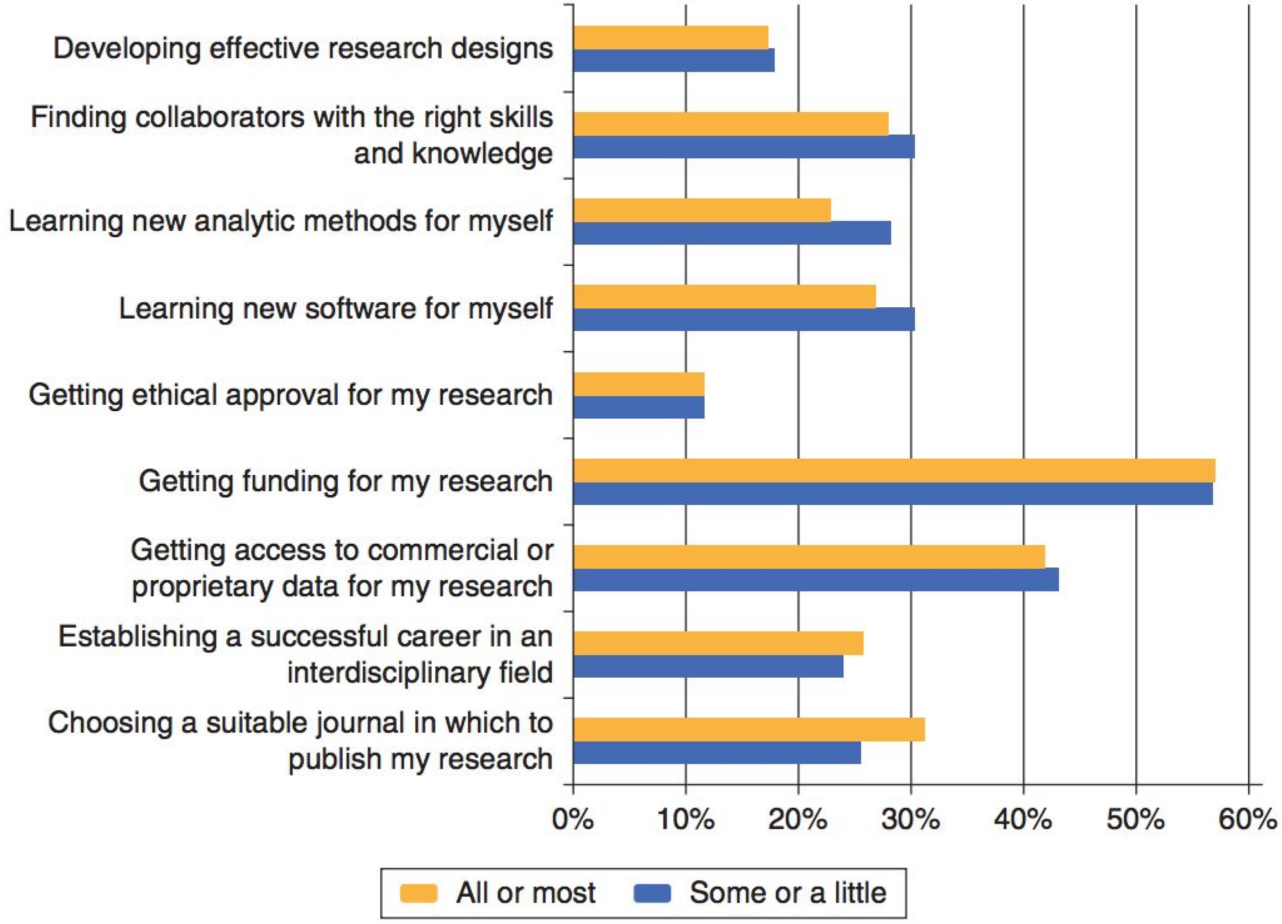
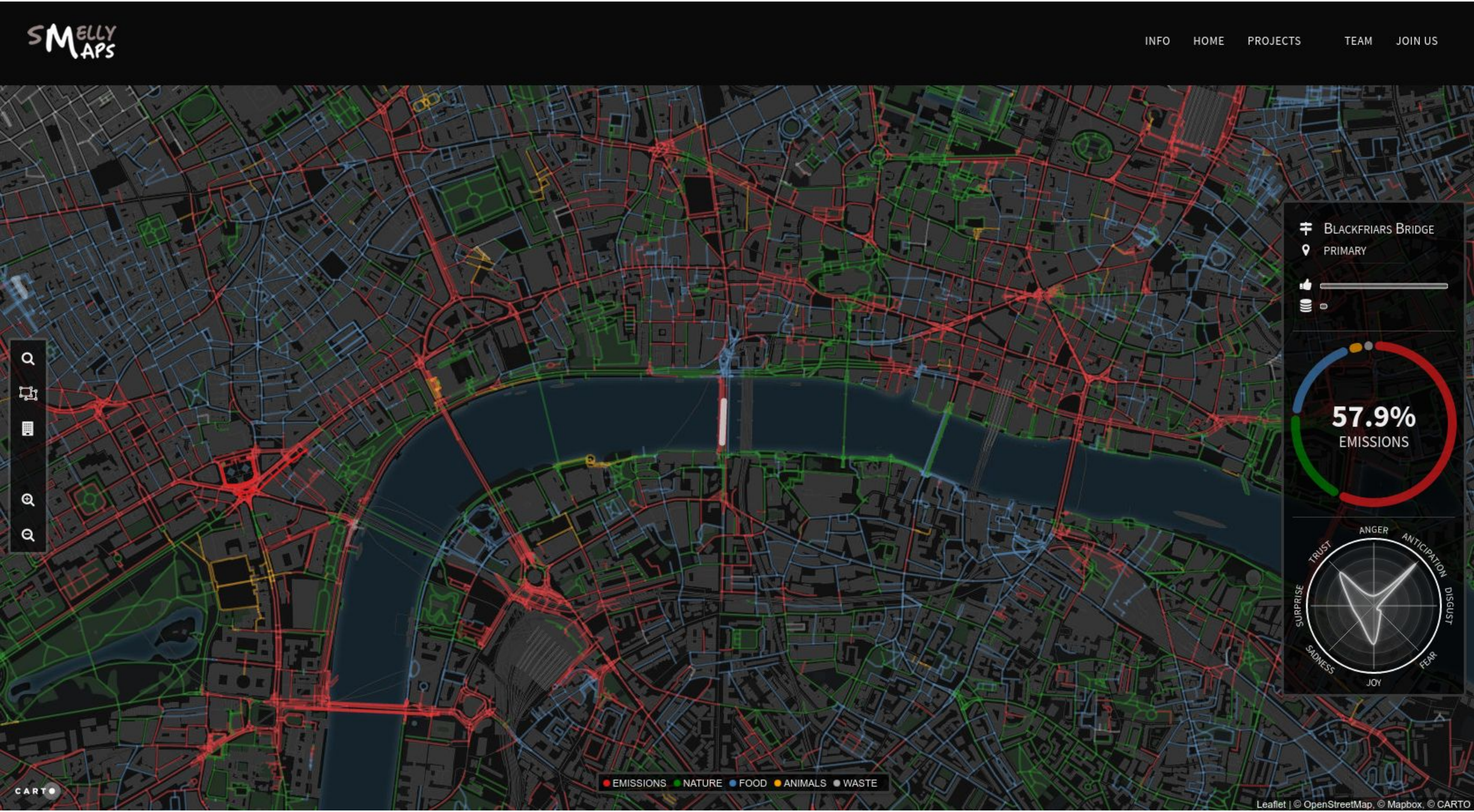


Figure 19 Problems encountered by amount of research using big data (n = 2266)




SMELLY MAPS
INFO
HOME
PROJECTS
TEAM
JOIN US
BLACKFRIARS BRIDGE
PRIMARY
57.9%
EMISSIONS
ANGER
ANTICIPATION
DISGUST
FEAR
JOY
SADNESS
SURPRISE
TRUST
EMISSIONS
NATURE
FOOD
ANIMALS
WASTE
CARTO
Leaflet | © OpenStreetMap, © Mapbox, © CARTO