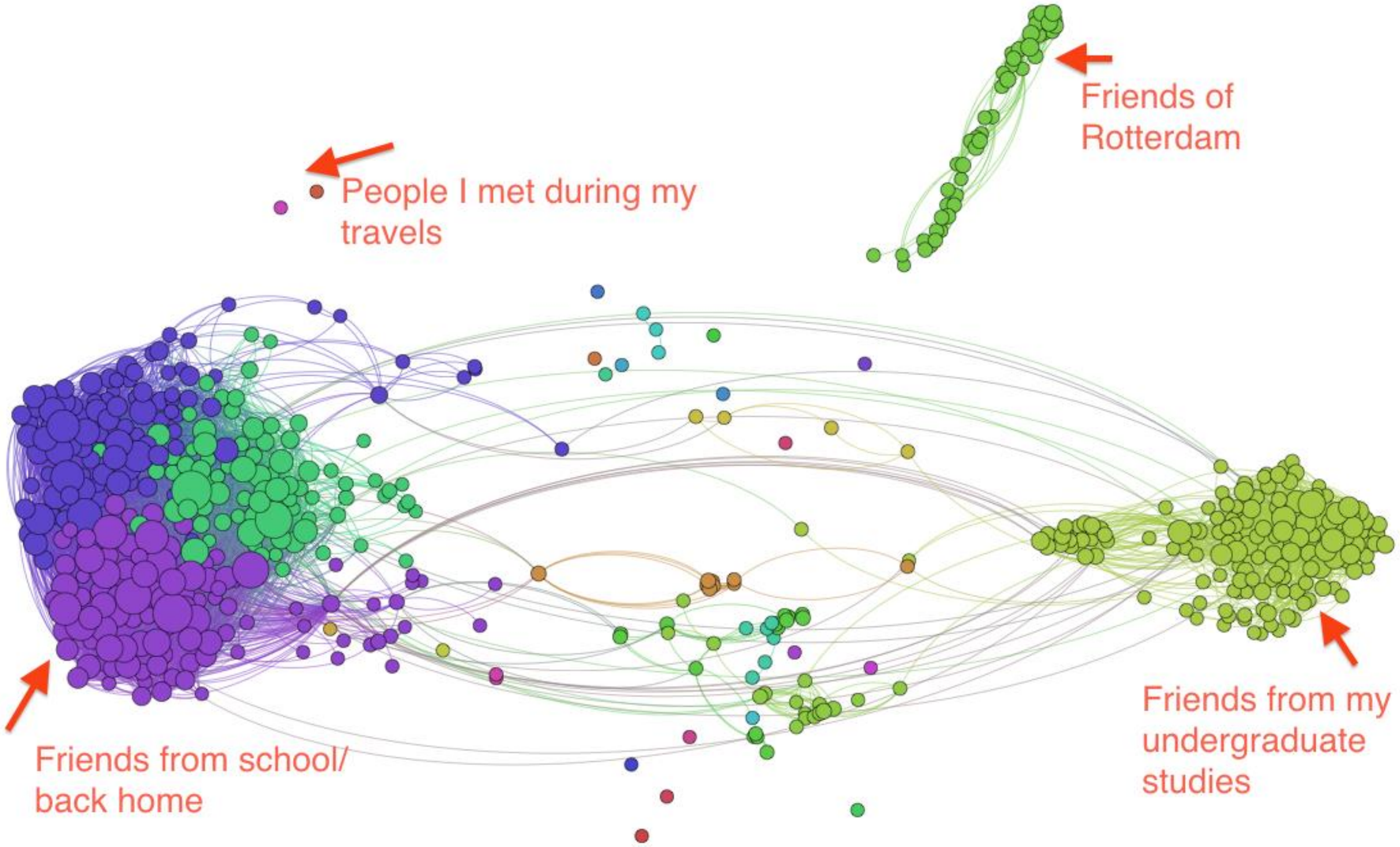
BIG DATA ANALYSIS & VISUALISATION

EGS course
in collaboration with Erasmus Studio



by 412753ibeur

today

- intro & meet and greet
- lecture on big data & DRM
- course outline

break

- examples of visualisation (tools)
- exercise 1: Wordle
- exercise 2: Facebook data
- group project explanation

meet and greet

name,
discipline(s),
PhD research,
aims for this course,
subjects of interest etc..

course outline

general outline per session

- intro in week's theme + short update
- a lecture by me (or Franciska)
- explanation of a tool
- exercises and/or project work
- q and a about literature/ lecture

themes per week

week 1: big data introduction
week 3: network analysis/ text mining*
week 2: network analysis & visualisations
week 4: sentiment analysis/text mining*
& final presentations!

* up for discussion with participants & type of projects

deliverables/ work load

group project:

presentation in week 4

project report

enhanced publication

(http://en.wikipedia.org/wiki/Enhanced_publication)

individual:

readings/ literature

engage with research tools

big data

BEST PRACTICE / **APPS FOR AMSTERDAM**

Met Best Practice presenteren we de beste en meest opvallende e-cultuur in Nederland.

Apps For Amsterdam

Gemeenten hebben enorm veel data over de samenleving tot hun beschikking; van criminaliteitscijfers tot vuilnisophaalroutes. Apps for Amsterdam daagt makers uit om deze overheidsdata op een creatieve manier te gebruiken en toegankelijk te maken.

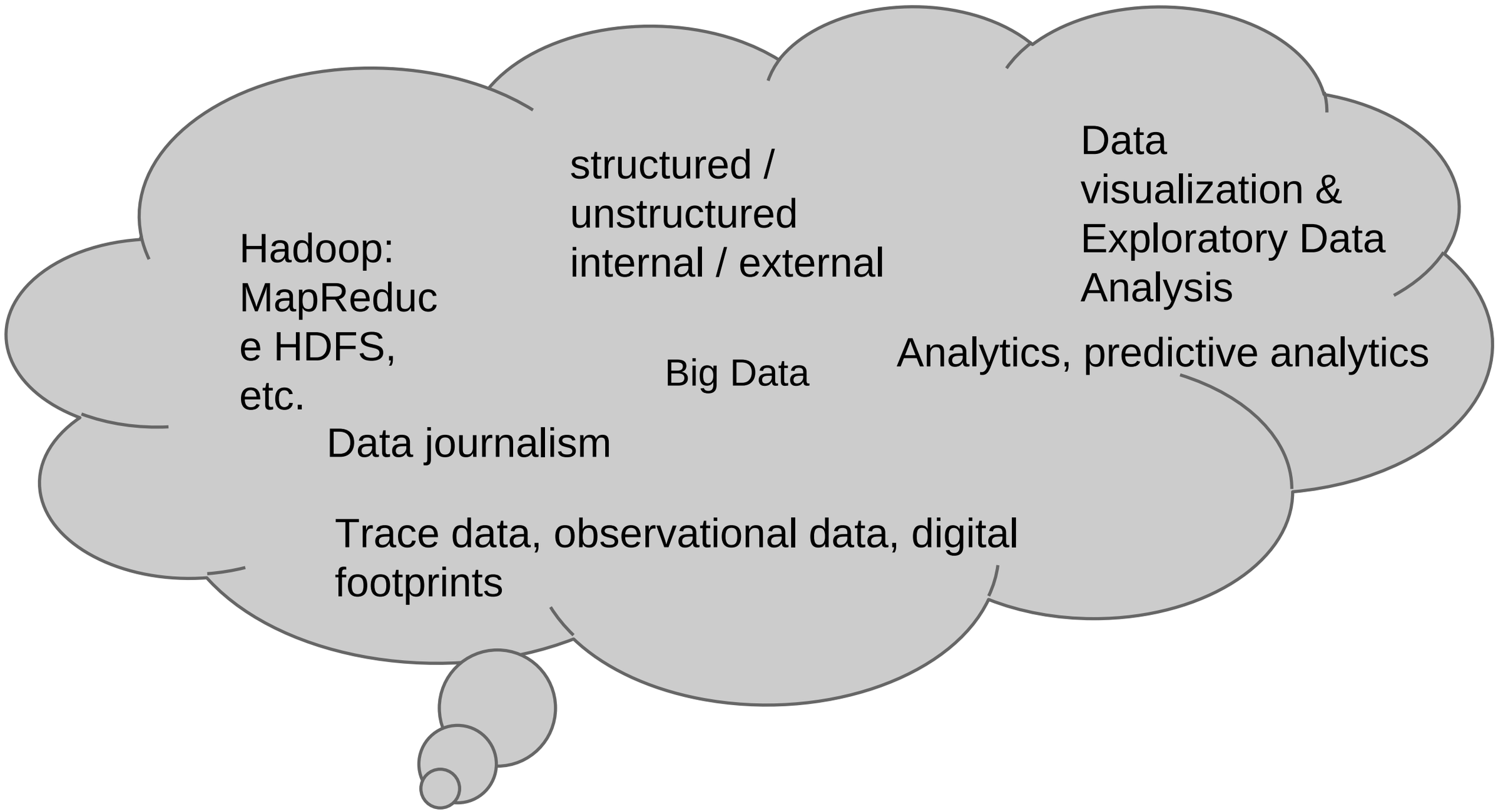
Moore's Law of Big Data:

"The Amount of Nonsense Packed Into the Term "BIG DATA" Doubles Approximately Every Two Years."

-Mike Pluta, 2014-08-10

What’s new?

“We have reason to fear that the multitude of books which grows every day in a prodigious fashion will make the following centuries fall into a state as barbarous as that of the centuries that followed the fall of the Roman Empire. Unless we try to prevent this danger by separating those books which we must throw out or leave in oblivion from those which one should save and within the latter between what is useful and what is not.” (Baillet, 1685, Quoted in Blair, 2003, p. 11)


Hadoop: MapReduce HDFS, etc.
structured / unstructured
internal / external
Data visualization & Exploratory Data Analysis
Big Data
Analytics, predictive analytics
Data journalism
Trace data, observational data, digital footprints

big data and science

big data for social science

“The power of ‘big data,’ ” Leetaru said, “means that a single person working purely on his personal time was able to extract [and tag] the images of 600 million pages of books spanning 500 years.” http://wapo.st/1AaXfLo (see also http://movies.benschmidt.org/)

“From this social science-based perspective, *big data are big because their* analytical potential *is qualitatively different than those researchers have been able to collect and assess in the past*, not due to their technical properties.” Menchen-Trevino, 2013

in short…

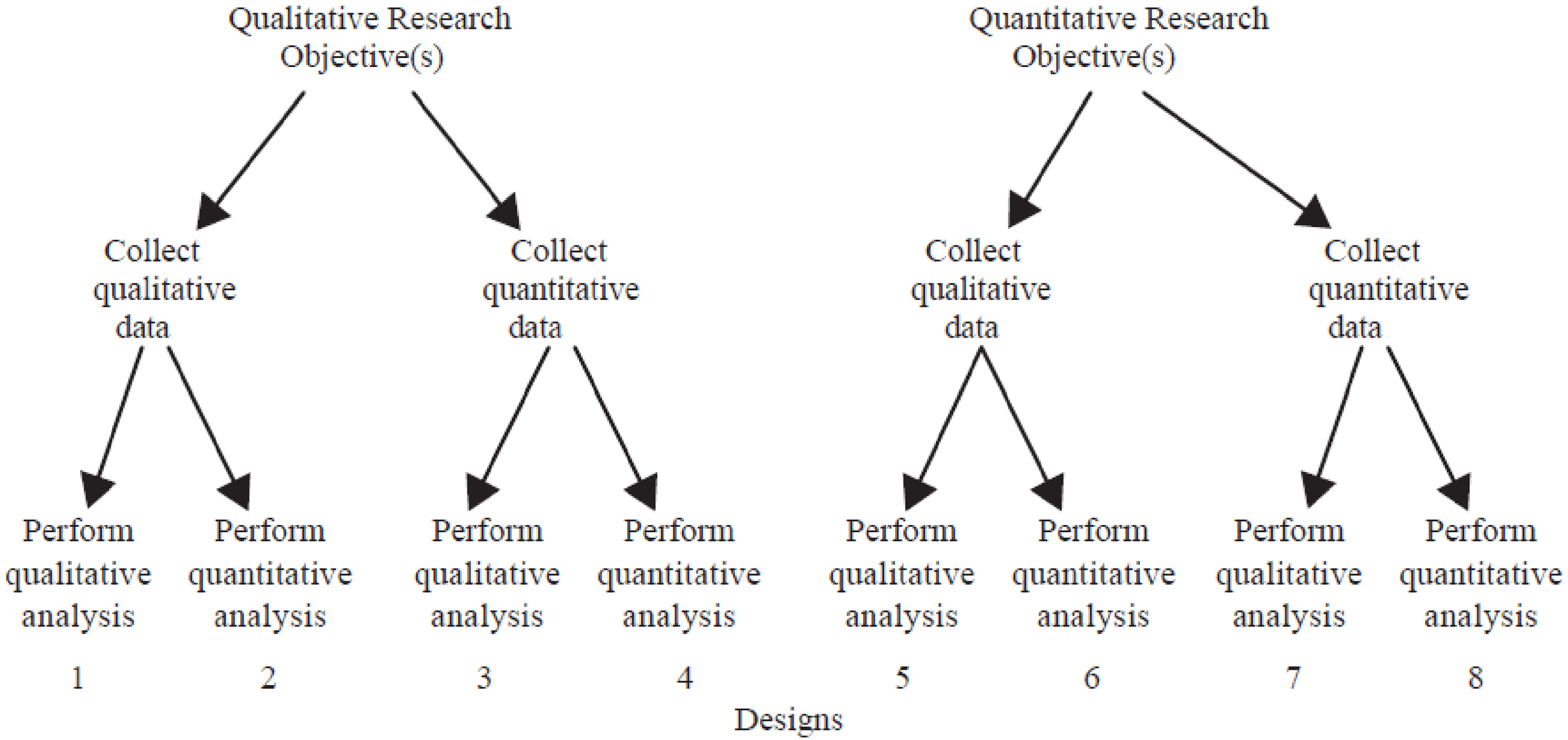
- what big data means depends on who you're asking..

- social science 'big data' is small (compared to computer science)

- in social science: analytical potential > 1 individual researcher

- grounded theory/ data-driven hypothesis

Methods overview



Note. Designs 1 and 8 on the outer edges are the monomethod designs. The mixed-model designs are Designs 2, 3, 4, 5, 6, and 7.[10]

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. Journal of mixed methods research, 1(2), 112-133.

Empirical Social Science Methods:

- Quantitative
  - Survey
  - Experiment
  - Content analysis
  - Network analysis
- Qualitative
  - Ethnography
  - Interview
  - Discourse analysis

For the purpose of this course:

- Mixed = Quant/Qual
- Multiple = Quant/Quant or Qual/Qual

Computational turn in science

- although we have been making records before, the computer has **accelerated** the generation of data

- the computer was introduced in **almost all types of work** in Western societies. Also in science. This is called the ‘computational turn’

- as a result of the **computational turn** (every type of science is being done via a computer), there is a tendency for the social sciences to become more **quantitative**, trying to compete with the natural sciences

Difference between statistical research and computational turn

> “traditional **distinction** between quantitative/statistical methods and qualitative/ ethnographic methods **with the addition of the new computational methods** that offer specific characteristics in terms of quantity and the nature of the data […]
>
> The computational approach is different from the statistical one because **data are not organized in a matrix of variables** and cases. Data are instead organized in a structure that recalls more a relational database than a spreadsheet […]
>
> This is the reason why **computational approaches do not necessarily need the use of statistics**, even if univariate or bivariate data representations are useful to visualize some results”

hypothesis driven

‘normal’ data
social science

‘normal’ data
natural science

qualitative

quantitative

big data
social science

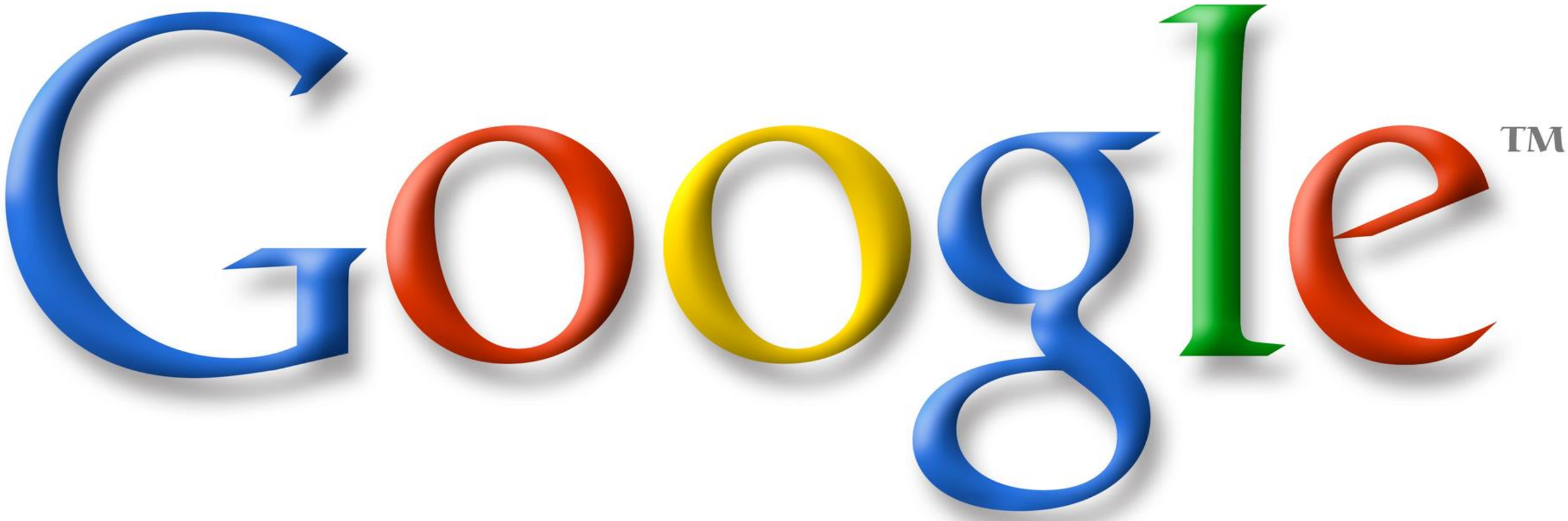
big data
natural science

data driven

from static to mobile computing…

to immersion and ubiquitousness

Has Google made social science obsolete?



What would Google do?

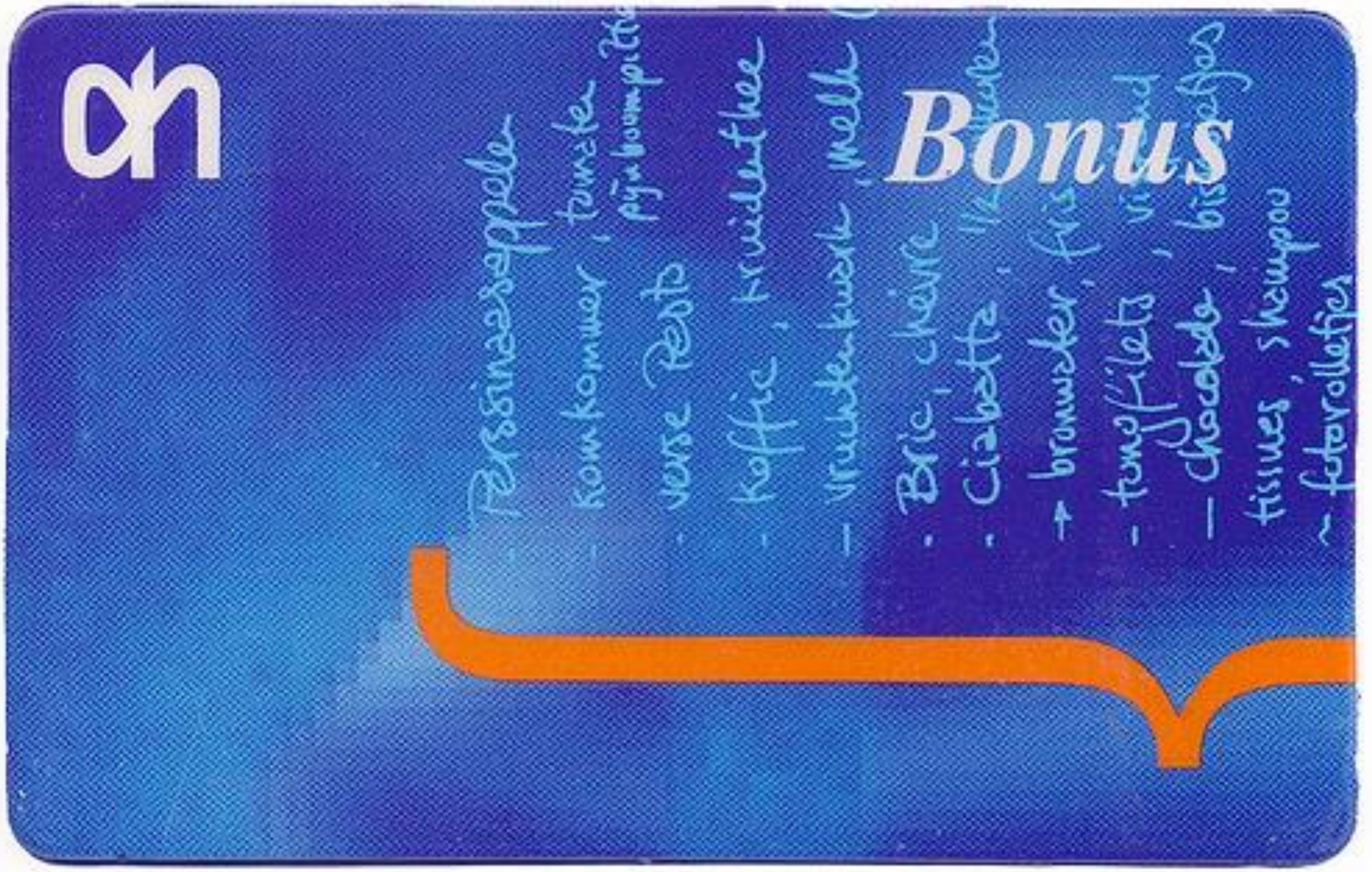
- **what would Google do?** became the question for social science and big data

- how to **diagnose** social and cultural conditions by means of the Internet

- the Internet becomes a **source**, not the object of study

- however, Google **cannot solve** each and every question (data needs context)

the Bonus card is one of many instances where data is stored



understanding digital humanities (Berry):

- objects have turned into **code**

- overabundance of sources and books and articles: all kinds of **cultural objects**

- however the **format** is discrete coding (in the end)

- from computation as **support** to computation as **a subject**: the computational turn (Berry, 2011)

some methodological questions

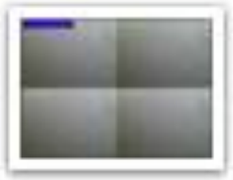
- **taxonomy** (or: knowledge-ordering systems; the problem of categories)

- **ethics** (when is data 'open' and free to use as a scientist?)

- cutting an **filtering data**: and time stamping it; when it what relevant to capture, scrape, or crawl?

- **onto-theology** and **ontologies**: what and how to study not only content of digital media, but also what they are.

0642.jpg



0675.jpg

0680.jpg

1250.jpg

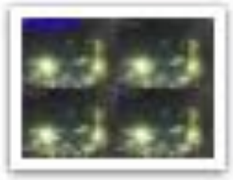
1275.jpg

1297.jpg

1400.jpg

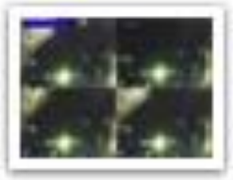
1423.jpg

1429.jpg



1433.jpg

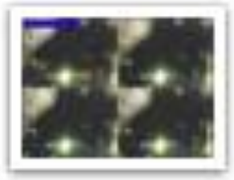
1461.jpg



1464.jpg

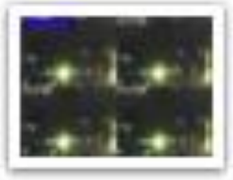
1479.jpg

1492.jpg



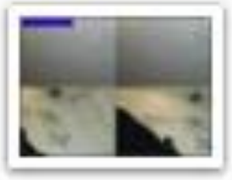
1499.jpg

1507.jpg



1516.jpg

1853.jpg



1882.jpg

1923.jpg

q & a

so far..?

visualisations

big big data

big data
natural science

qualitative

quantitative

big data
social science

small big data

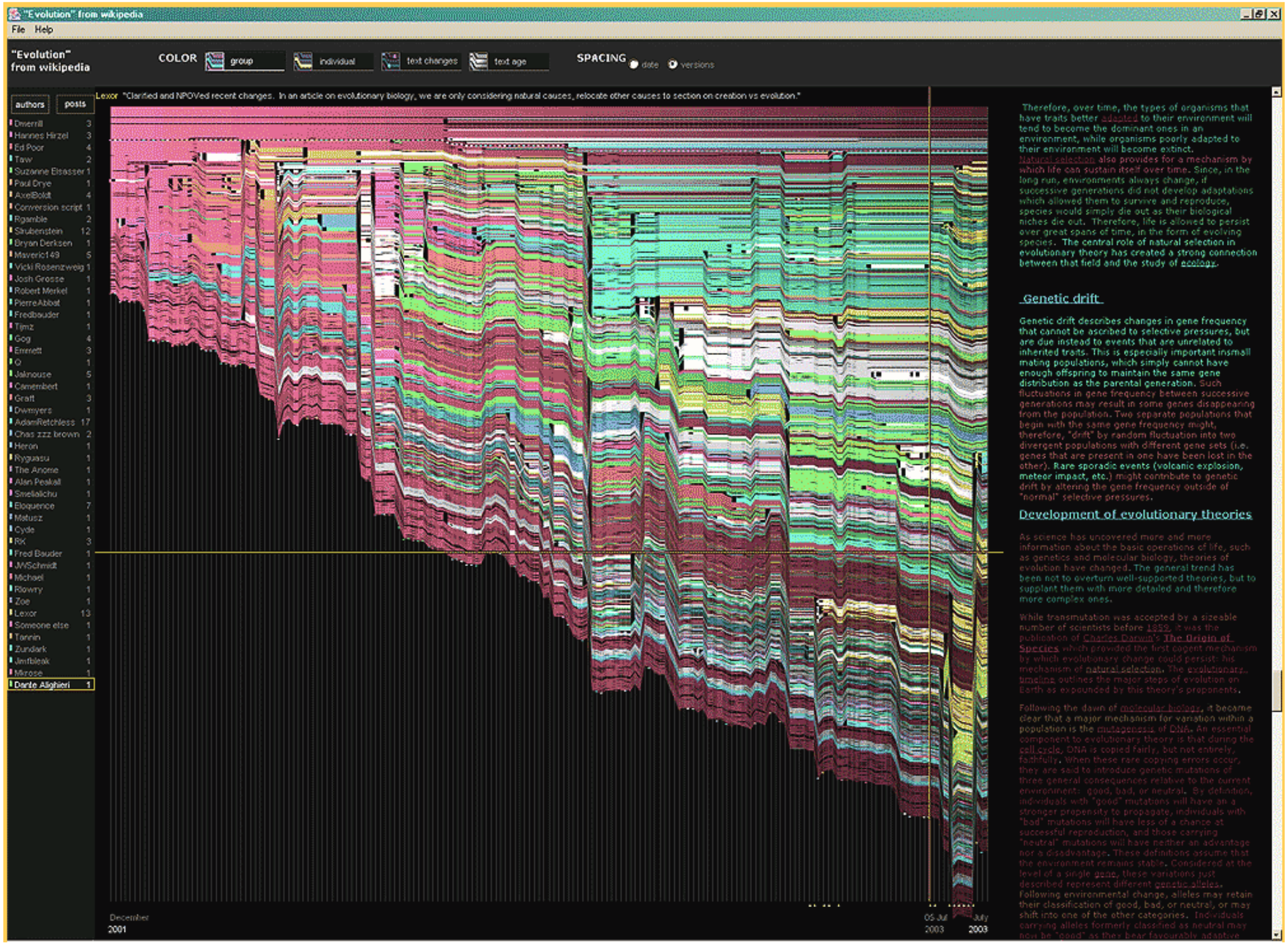
big data visualisations

- data > big data, but how to make it **accessible**?

- from static visualisations to **interactive** ones

- **aesthetics** and readability

- **medium-specific** information design



https://www.research.ibm.com/visual/projects/history_flow/gallery.htm

Data-Driven Documents

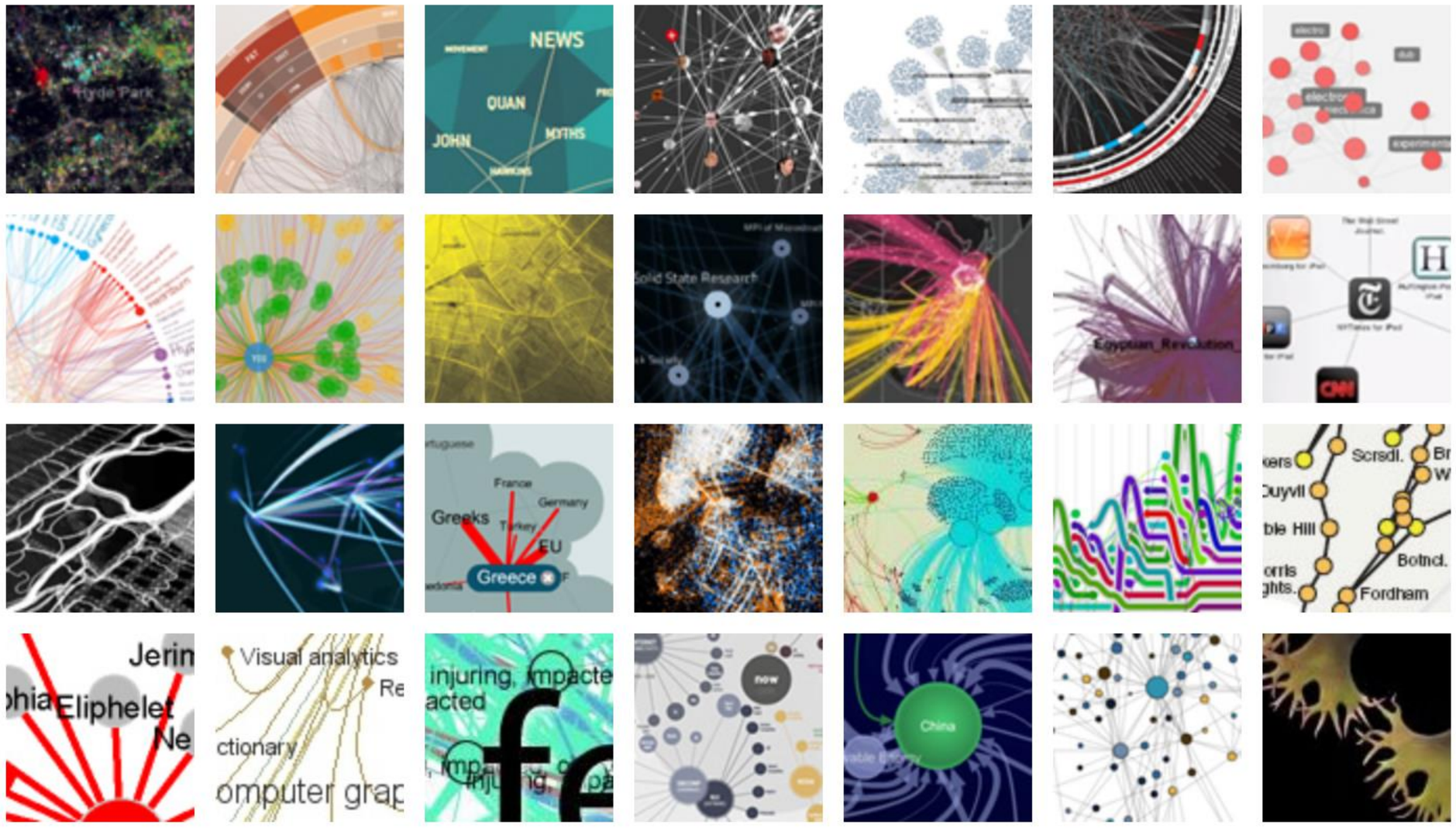
d3

visual complexity

Search the VC database:

GO

Latest Projects:



[visual complexity]

[social networks]

2. Shares of human, bot and software-assisted human edits in the top twenty most active Wikipedia entries in the English-language versi

Processing 2

Cover

Download

Exhibition

Reference
Libraries
Tools
Environment

Tutorials
Examples
Books

Overview
People
Foundation

Shop

» Forum
» GitHub
» Issues
» Wiki
» FAQ
» Twitter

» Download Processing
» Play With Examples
» Browse Tutorials

Processing is a programming language, development environment, and online community. Since 2001, Processing has promoted software literacy within the visual arts and visual literacy within technology. Initially created to serve as a software sketchbook and to teach computer programming fundamentals within a visual context, Processing evolved into a development tool for professionals. Today, there are tens of thousands of students, artists, designers, researchers, and hobbyists who use Processing for learning, prototyping, and production.

» Free to download and open source

» Interactive programs with 2D, 3D or PDF output

» OpenGL integration for accelerated 3D

» For GNU/Linux, Mac OS X, and Windows

» Over 100 libraries extend the core software

» Well documented, with many books available

» Exhibition

Filament Sculptures
by Lia

Fall in Love - Phantogram
by Timothy Saccenti and Joshua Davis

Keyflies
by Miles Peyton

processing

VISUAL ANALYTICS FOR EVERYONE

Tableau can help anyone find real answers in their data.

SEE TABLEAU IN ACTION

FAST A
FOR E

Tableau

BUSIN
INTELI

Tableau

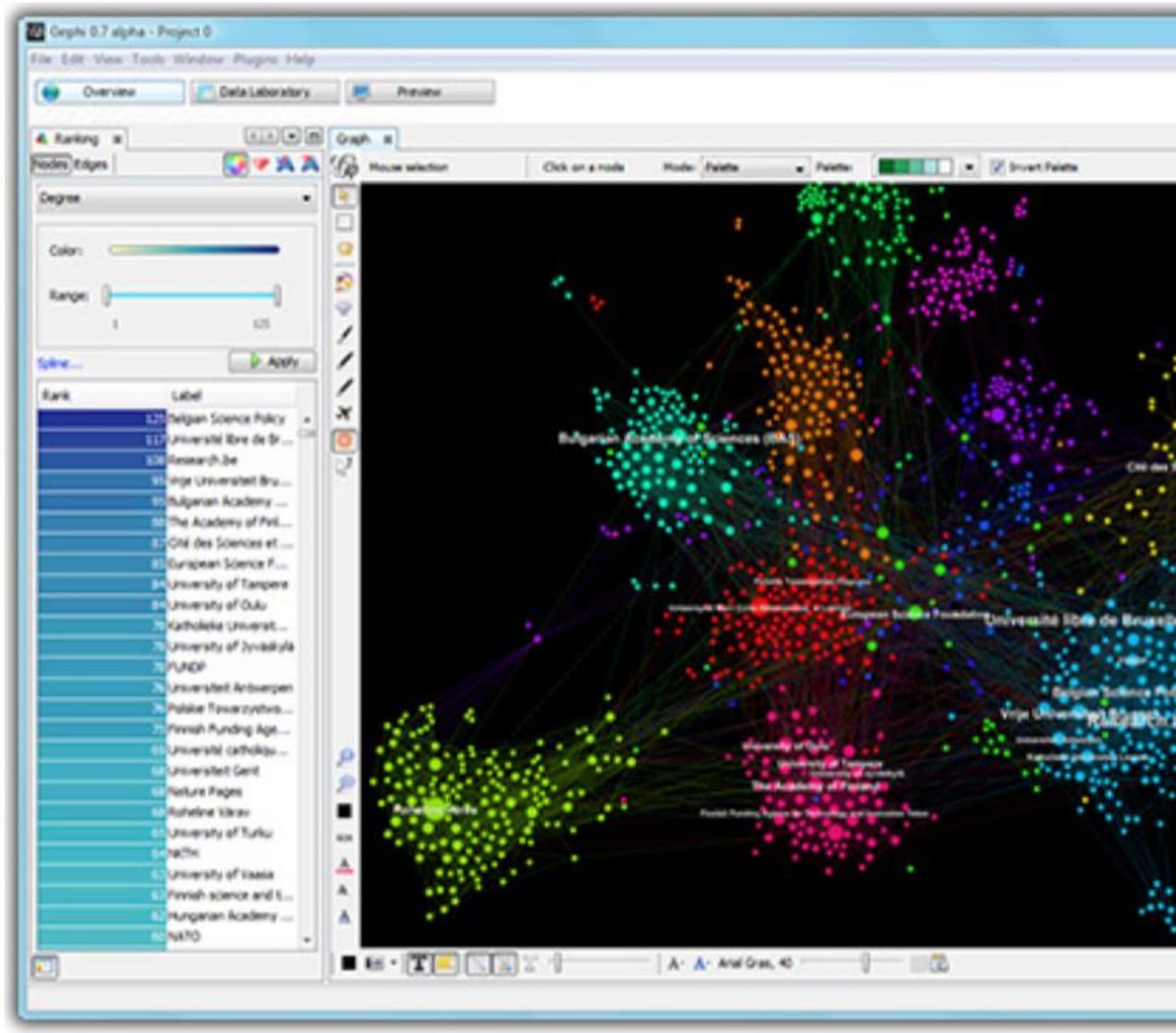
ANALY
THE C

The Open Graph Viz Platform

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.

Runs on Windows, Linux and Mac OS X. Gephi is open-source and free.

Learn More on Gephi Platform »



Download FREE
Gephi 0.8.2-beta

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos

Support us! We are non-profit. Help us to innovate and empower the community by donating c

Donate

APPLICATIONS

PAPERS

Explore

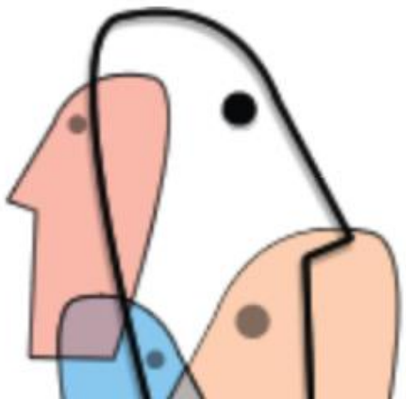
- Visualizations
- Data sets
- Comments
- Topic centers

Participate

- Create a visualization
- Upload a data set
- Create a topic center
- Register

Learn more

- Quick start
- Visualization types
- About Many Eyes
- Privacy
- Blog



Try our featured visualizations

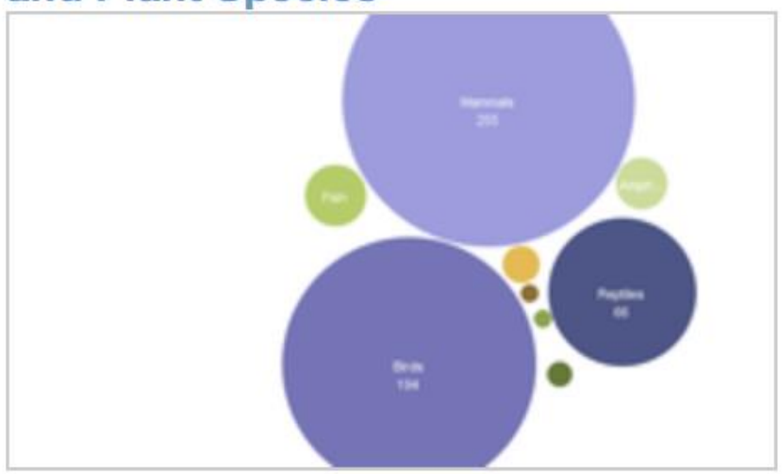
Film Comedy Titles

Based on IMDB
by shngl

Haiti

Haiti DFID Aid Data
by ChrisGarde

Threatened and Endangered Wildlife and Plant Species



2011 - Endangered species
by LJean

KFC franchises worldwide (excl US and China)

2013
by brenda233

GAPMINDER for a fact-based world view

Blog | FAQ

HOME | GAPMINDER WORLD | DATA | VIDEOS | DOWNLOADS | FOR TEACHERS | LABS | IGNORANCE

Explore the world

Gapminder World shows the World's most important trends

- Wealth & Health of Nations
- CO_2 emissions since 1820
- Africa is not a country!
- Is child mortality falling?
- Where is HIV decreasing?

Load Gapminder World

"Without d
control, th
increas from 1
next months.

Like 53,

Subscribe

e-mail

Latest videos

Demographic Party Trick 1 – Hans Rosling & Bill Gates

Gapminder Labs

Gapminder Agriculture
700 indicators from the Food and Agriculture Organization (FAO).

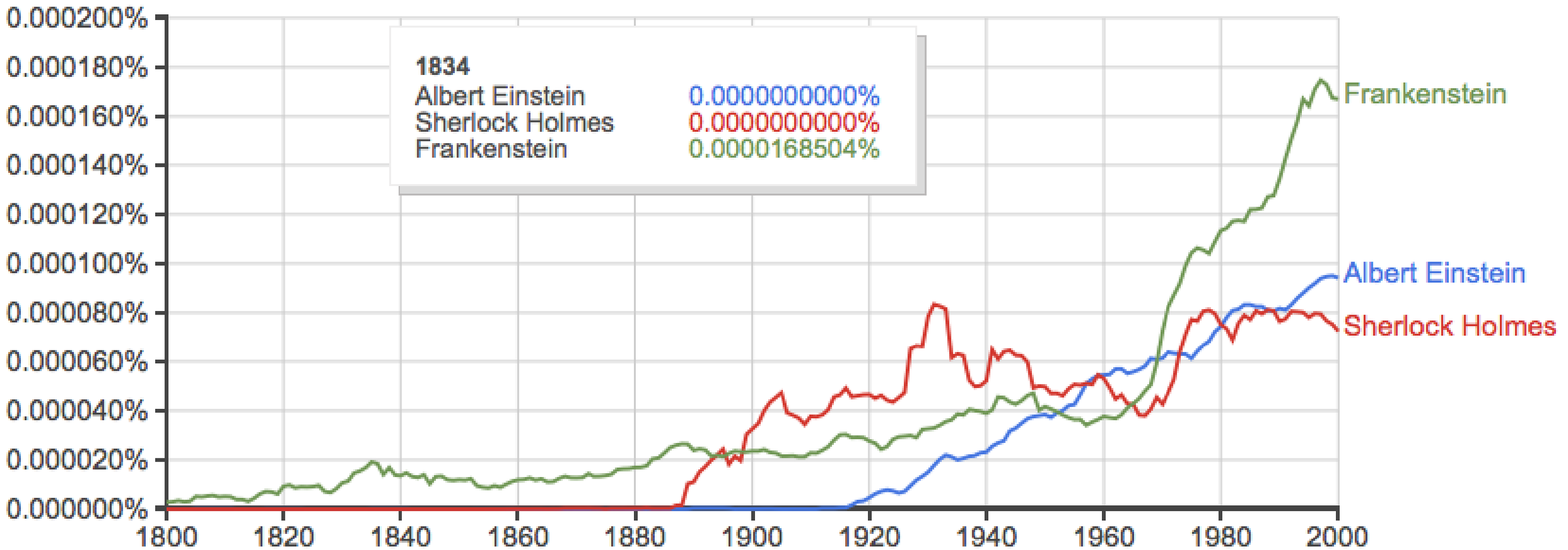
News

- Ignorance Survey in Ger
- Swedish and Norwegian

Google books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3 . Search lots of books



(click on line/label for focus)

Exploring Datasets

Google Trends
Gap Minder (video)
Google Data Explorer: Unemployment Europe
Many Eyes: BrBa
Sentiment 140 (Requires Twitter sign in)
A typical friend network
realtime twitter maps
N-gram viewer: Big Data and analytics based on term frequency

break

exercise 1:
making a Wordle

Text mining and word clouds

1. get a dataset of strings and/or a piece of text

2. get the main words out of the text (excl. ‘the’ etc..) via e.g., a tag cloud-generator

3. transform tag cloud into a Wordle

4. ponder about what you have actually made

DMI Tools

Media Analysis: Media Monitoring | Mapping | Clouding | Comparative Media Analysis
Data Treatment: Data Collection | Data Analysis | Information Visualization
Natively Digital: The Link | The URL | The Tag | The Domain | The PageRank | The Robots.txt
Device Centric: Google | Google Images | Google News | Google Blog Search | Yahoo | YouTube | Flickr | Del.icio.us | Wikipedia | Alexa | IssueCrawler | Twitter | Facebook | Amazon
Spherical: Web Sphere | News Sphere | Blogosphere | Tag Sphere | Video Sphere | Image Sphere

Actor Profiler

Launch tool **Instructions & Scenarios of Use**

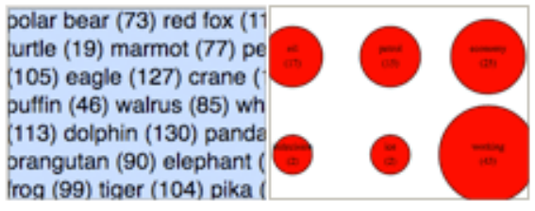
The Actor Profiler works in tandem with the Issue Crawler. It calculates the top ten actors on an Issue Crawler map, and profiles each of them. The profile consists of each actor's inlinks a...

Amazon Book Explorer

Launch tool **Instructions & Scenarios of Use**

Provides different analytics for Amazon.com's book search

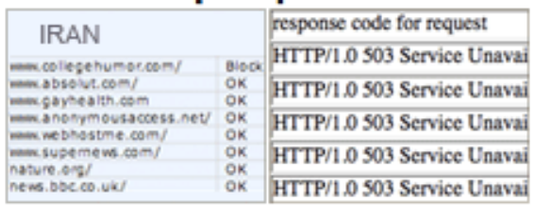
Bubble Lines



Launch tool **Instructions & Scenarios of Use**

Input tags and values to produce relatively sized bubbles. Output is an svg.

Censorship Explorer



Launch tool **Instructions & Scenarios of Use**

Check whether a URL is censored in a particular country by using proxies located around the world.

https://wiki.digitalmethods.net/Dmi/ToolDatabase

(step 2: to create a tag cloud from raw text)

Tag Cloud Generator

Input

Enter a text in the box below. All words will be counted, the output is a tagcloud.

Options

Minimum number of characters 3

Minimum frequency 3

Stopword treatment (•) keep ( ) remove

Ordering of tags ( ) by frequency ( ) by term name (•) unordered

Additional input for the SVG and PDF rendering

Describe the method used in analysis (max 55 characters)

Which query produces the input? (max 55 characters)

Give the cloud a title (max 45 characters)

Tag cloud

Tag Cloud Generator, an Introduction

Takes raw text, counts the words and returns an ordered, unordered or alphabetically ordered tagcloud.

Enter text. All words will be counted, the output is a tagcloud.

Tag Cloud Generator, Sample Output (Toggle)

https://tools.digitalmethods.net/beta/tagcloud/

(step 3: to convert a tag cloud into a Wordle)

Tag Cloud To Wordle

Input

Input tag cloud in the format tag1 (13) tag2 (2) ...

remove frequency from displayed tagcloud

to Wordle format

Tag Cloud To Wordle, an Introduction

This tool allows one to transform a normal tag cloud into a fancy Wordle one.

Insert a tag cloud in the following format: war (4) peace (2) United Nations (17)
Copy paste the output in wordle.net/advanced, click 'Go'
Choose the font, layout, and color to your liking in Wordle.

https://tools.digitalmethods.net/beta/tagcloudToWordle/

exercise 2: your own Facebook data

http://www.wolframalpha.com/input/?i=fac ebook+report

Netvizz and Gephi

- install Netvizz as a plugin for Facebook
- download the Netvizz file (all three networks)
- open the file in Gephi (pre installed?)
- follow the steps as given here: http://gephi.github.io/users/ (my Facebook network)

  or here: http://www.reddit.com/r/DRMatEUR/comments/2h50i2/visualizations_of_my_fb_network_an_easy_guide_how/

Some specific assignments

1. create your own Facebook **friend** network
2. create a network of **work-relations**
3. try to find the **most popular post(s)** in your network

exercise 3: getting Twitter data (for next week)

steps:

- installing Python & PIP
- installing twitter library for Python
- installing custom program (see link)
- enter query and get tweets!
- load tweets into Tableau software

Twitter API

follow the steps as explained here:

https://github.com/erickaakcire/ornithologist

- You can only get tweets from the past 6-9 days
- Results options: Recent / Popular / Mixed
- Language: English
- Just 100 Tweets

Twitter Data: Tweets

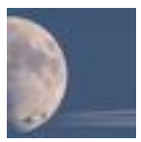
user name	HSWorldwide
time created	Thu Sep 04 08:01:35 2014
retweets	23
favorites	0
tweet source	Twitter Web Client
language	en
withheld from	None
tweet	RT @djbrennanheart: You can now listen to #WERHardstyle on all @KLM flights! http://t.co/jil5aHPPid
to user	
retweet of	djbrennanheart

all user mentions
djbrennanheart, KLM
hashtags
#WERHardstyle
links
http://t.co/jil5aHPPid
data gathered: 2014-09-04 08:04:04.304042 UTC

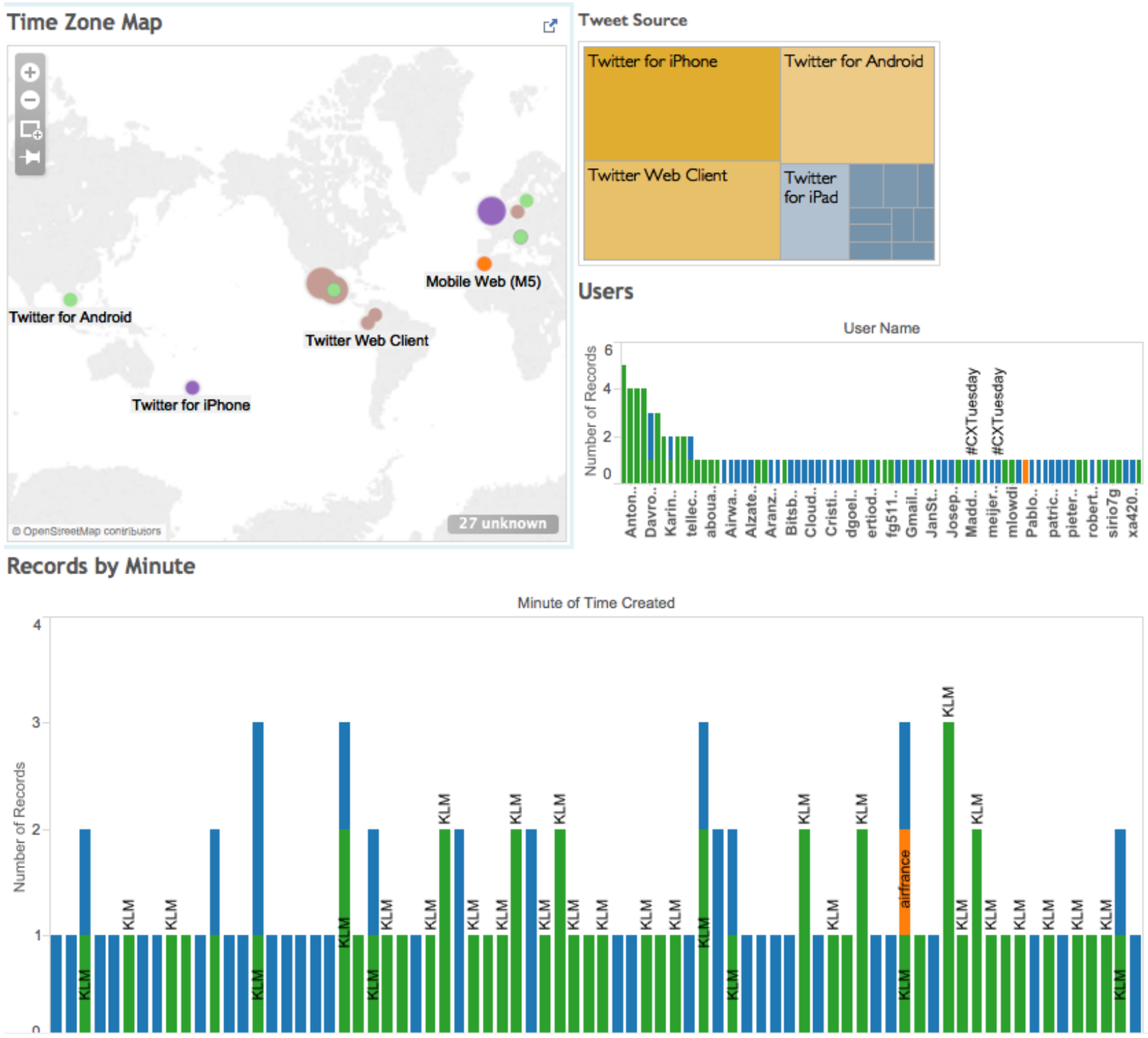
Twitter Data: Users

user id 402914779
screen name DrTrx
location Brabant, The Netherlands.
time zone None
utc offset None
profile image URL https://pbs.twimg.com/profile_images/2706211644/54bd2a2eeea206e939ea685dceeb1681_normal.jpeg

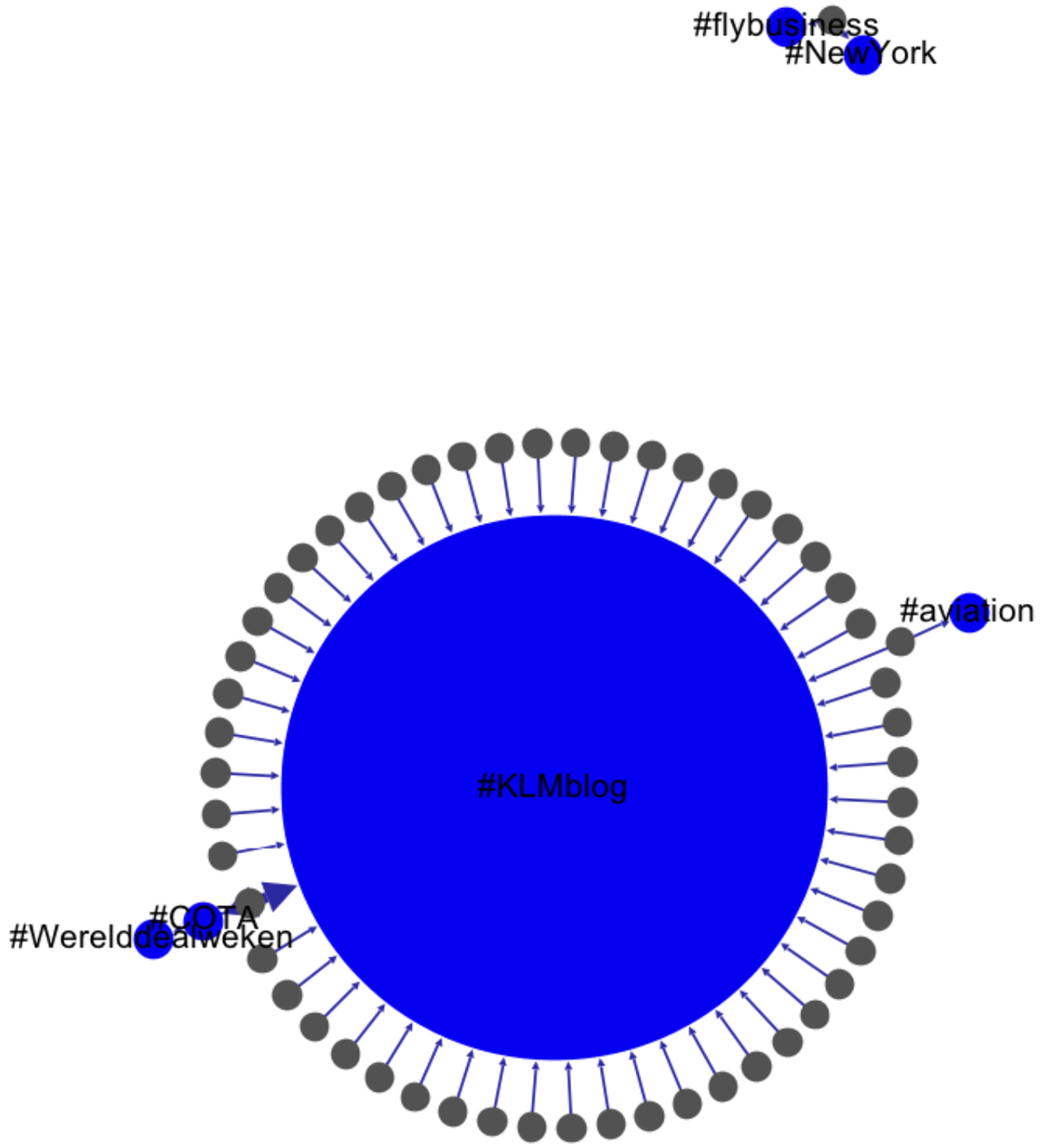
description Passionate, Photographing(Canon), Nursing, Aviation-loving Individual with open mind and opinion. Dutchman



Tableau


Time Zone Map
Twitter for Android
Twitter Web Client
Mobile Web (M5)
Twitter for iPhone
© OpenStreetMap contributors
27 unknown
Tweet Source
Twitter for iPhone
Twitter for Android
Twitter Web Client
Twitter for iPad
Users
User Name
Number of Records
#CXTuesday
#CXTuesday
Records by Minute
Minute of Time Created
Number of Records
KLM
airfrance

Gephi


#flybusiness
#NewYork
#KLMblog
#aviation
#COTA
#Werelddeelweken

the group project

- form groups of 2
- perform a short research project that can be presented in week 4
- try to follow a protocol for data research (get data, clean it up, analyse, generate visualisation, conclude and make presentable)

examples of projects and enhanced publicat

- http://minordigitalculture.wordpress.com/2014/01/24/projects-created-by-students-in-the-course-digital-visual-culture/

- (hier nog link naar voorbeelden MA course vorig jaar)

recap and to-do for next week

- hand in research plan (short)

- visit our website!
http://www.eur.nl/erasmusstudio/EGScourse2014/
(also for the slides and links to data- and software sources)

- play with software tools!