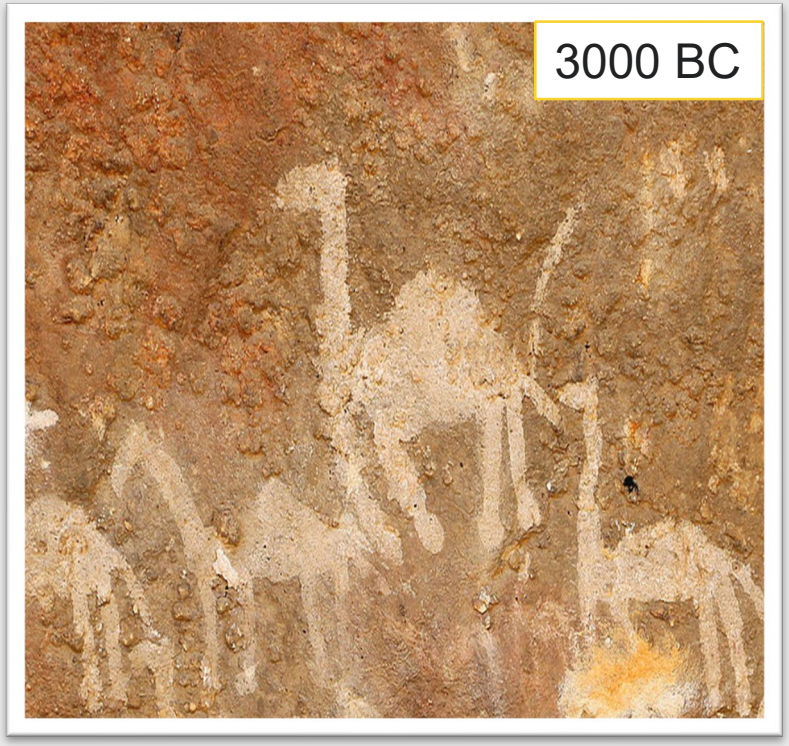
Visualize my Corpus

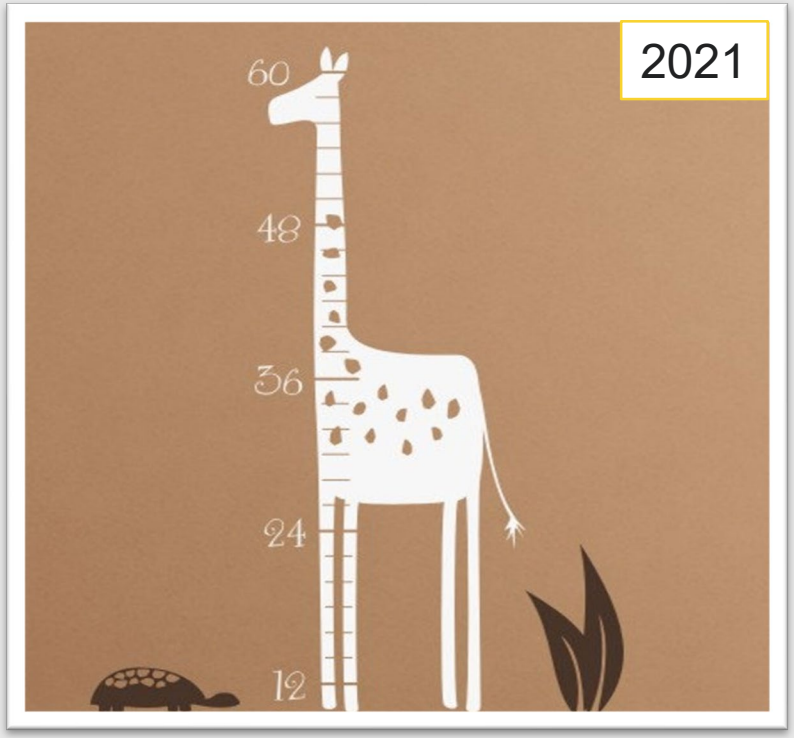
Dr Mahmoud El-Haj

Lecturer

SCC, Lancaster University

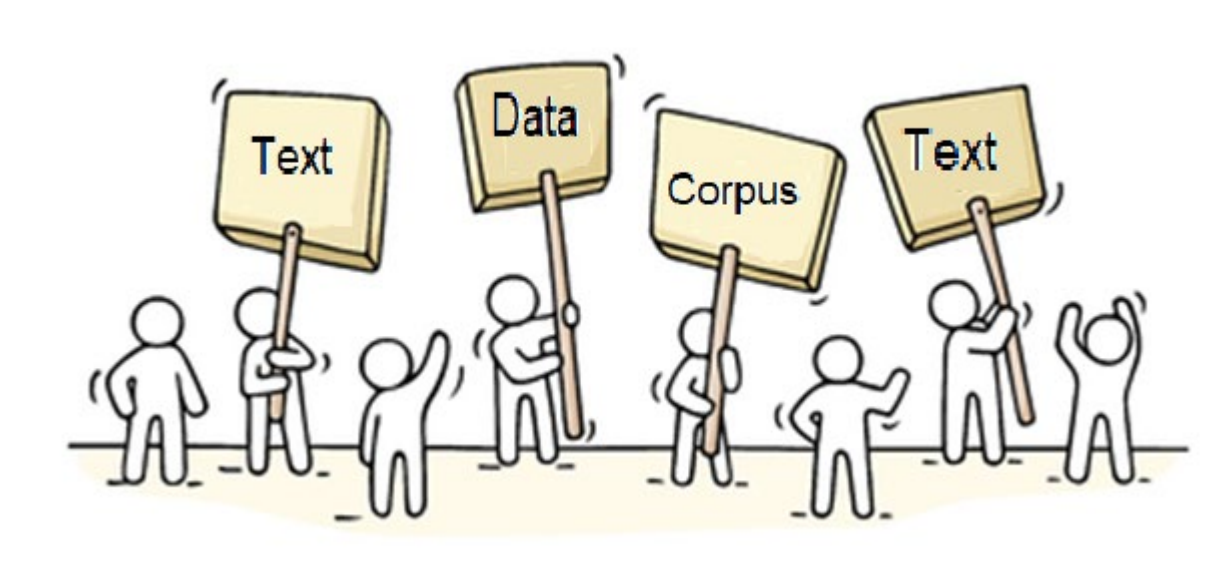
What is visualisation?





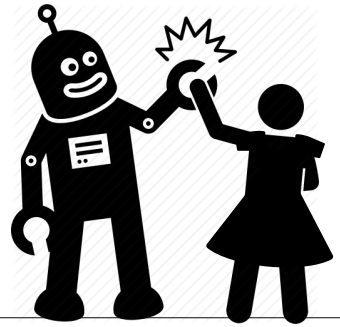
communicate a message

What to visualise?



WHY VISUALISE?

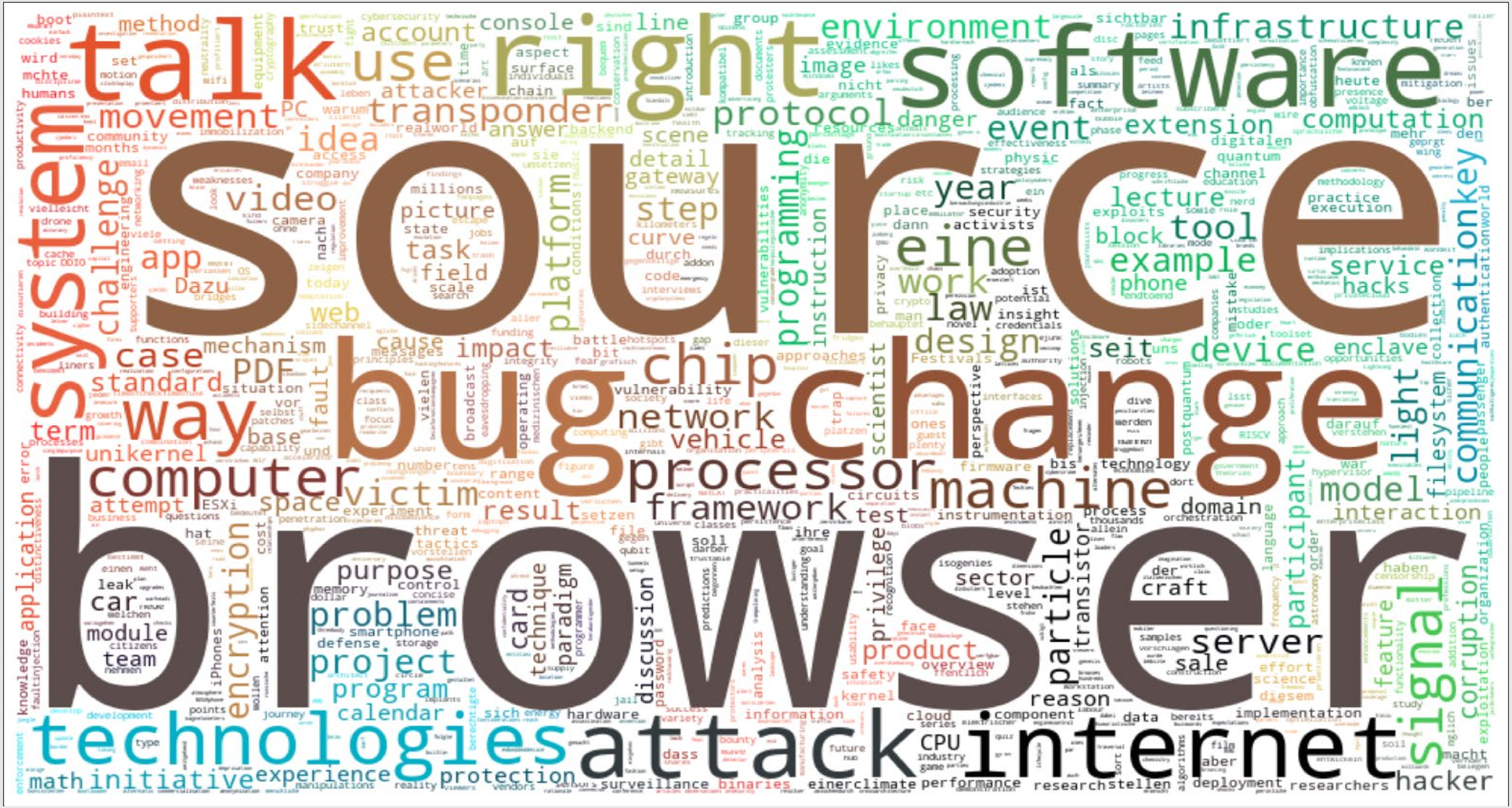
Machine: I can read it!

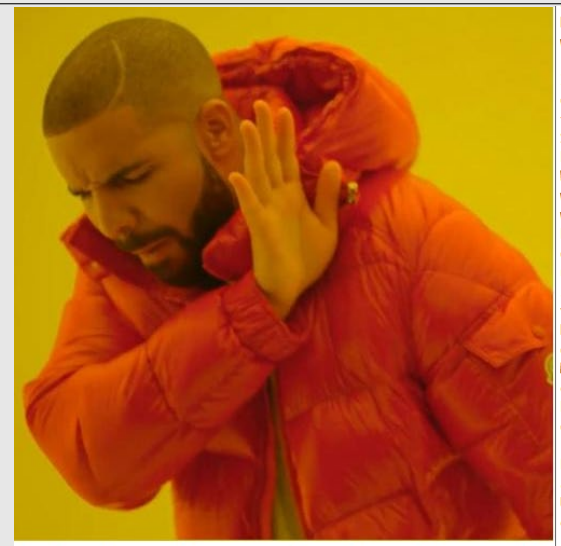


Human: I can read it too!

Isn’t data enough as it is?

ckelte sich ueber Computer virenz. Zwei Redner die mit Ralph Burger (Autor des empty"Bi
en war die niederlaendische Gruppe hackz-TIC die eine Maschine (ca. 1500$) vorfuehrte, die K
es CCC) arrangiert wurde um die Unsecurityz des Plastikgeldes zu demonstrieren. Waehrend
eichen CCC attackze auf die BTX/HASPA Konten war) sagte, das dies unmoeglich waere, beka
opie seine Karte. Allerdings wurde bei dem Versuch Geld aus einem Automaten zu beko
Workshop ueber 'Phreaking'. Versuche und Methoden wie man 'so weit, mit so vielen Phr
beschrieben. Genauso wie Tricks mit der 130-Nummer der Post(sowie Verbindungen zu den
kutiert wurden, um preiswertes Telefonie nkonferenzen und Voice Mailbox
l von einem US phreak zu hoeren, dass d MHz, rote Boxen um den Muen
Bericht war der Kommunikation zwis aten gewidmet, wo bestim
nd. Verwendung von Mailboxenv Infrastruktur fuer Veroeff
e Sitzung beschaeftigte sich mit der zuhelfen, indem ein zivile
en Hilfe von Computerhaendlern, d modems spendeten, und
e das Projekt mit der Wiedervereinig heren Diskussionen des CCC
udent (Tommi) seine Diplomarbeit: vo logie. Workshop Work
p Workshop Workshop Workshop Hier vaten Netze IN, SubNet, Maus
und wie man sich anschliessen kann Recht interessant, um einen kleinen Einblick in den Hint
nd Irc wurden den Anwesenden negebracht. Besonders hervorzuheben ist der Beitrag von
en der News rueberbrachte u auch eine Diskussion ueber die Zensur von Newsgruppen ir
fundiert war, die Meinung jedoch nicht zusammenzubringen waren. Empty Workshopem
y"Datenschutz"empty" wurde hauptsaechlich ueber die Arbeit und die Aufgaben des Da
wieder die implizite nsecurityz der heutigen Betriebssysteme Unix und DOS (PC) angefue
h gegen die Ueberschrift empty"Techno-terrorz"empty" gewandt hatte und erstmal eine Be
e"empty" sein? Die diversen zu diskutierenden Begriffe? Eingeschraenkt auf das Thema "empt
Pilotprojekt empty"DAWIN"empty" vorgestellt Funknetze: C-Netz, D-Netz, Cityruf etc. - Wie
e abgehört werden? Die von engagierten MailBoxen geschaffenen unabhängigen Bürger
edens- und Menschenrechtsgruppen ermöglichen. Seit vielen Monaten erhalten wir über die
slawien. Auf dem Congress wird eine Direktverbindung zu verschiedenen MailBoxen dort he
as sind Computervirenz; - Wie arebietne virenz? - Gibt es Schutz vor virenz? - Umgang mit
“hackz-Jäger” Protokolle: IP, TCP, UDP, ICMP Routing, Adressen Funktionsweisen: DNS, Eth
nputer Club treffen sich zum jährlichen Austausch organisatorischer Informationen. Insbeson

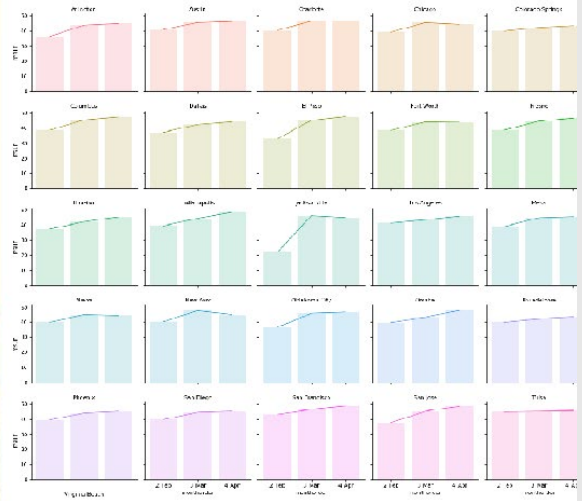

source
browser
bug
change
software
right
talk
system
attack
internet
technologies
processor
machine
computer
chip
eine
way
transponder
protocol
infrastructure
environment
event
extension
computation
movement
use
account
method
console
idea
video
app
challenge
platform
programming
instruction
vulnerabilities
scientist
privilege
particle
transistor
signal
communication
key
authentication
world
filesystem
light
participant
corruption
feature
model
device
enclave
example
service
hacks
tool
lecture
block
phone
year
law
work
design
network
vehicle
framework
victim
result
space
attempt
case
standard
PDF
term
unikernel
step
gateway
curve
detail
scene
picture
task
field
web
impact
mechanism
encryption
purpose
problem
project
program
calendar
initiative
experience
protection
math
module
team
car
leak
card
technique
paradigm
discussion
sector
craft
server
sale
reason
product
kernel
hacker
interaction
domain
test
CPU
hardware
surveillance
binaries
climate
performance
research
stellen
deployment
researchers
implementation
component
cloud
data
application
knowledge
seit
tool



Data in plain text

DDR zuhelfen, indem ein ziviles Computernetz (DDRNET) eingefuehrt
werden sollte. Trotz der grossen Hilfe von Computerhaendlern, die
spontan PC's, Software und modems spendeten, und trotz lokaler
Interessen, Kosten und Organisationsproblemen musste das Projekt mi
der Wiedervereinigung gestoppt Werden. Frueheren Diskussionen des C
folgend, ueber soziologische Aspekte des hackzens, beschreibt ein
Student (Tommi) seine Diplomarbeit: Die Verwandheit von Computer un
Psychologie. Workshop Workshop Workshop Diskussion Workshop
Workshop Workshop Workshop Workshop Workshop Workshop Hier
wurden die privaten Netze IN, SubNet, Maus und andere beschrieben,
was fuer Moeglichkeiten die Netze haben und wie man sich anschliess
kann. Recht interessant, um einen kleinen Einblick in den Hintergru
der Netzwerke zu bekommen. Mail, News, FTP, Remote Login
und Irc wurden den Anwesenden nahegebracht. Besonders hervorzuheben
ist der Beitrag von Princess ueber die News, der engagiert und gut
verstaendlich das Wesen der News rueberbrachte und auch eine
Diskussion ueber die Zensur von Newsgruppen in Gang setzte, die tro
der Zielgruppe des Vortrags (Netz-Laien) recht fundiert war, die
Meinungen jedoch nicht zusammenzubringen waren.
empty"Workshopempty" Cornflakes-Pfeifen (BlueBoxing)" In der
Podiumsdiskussion empty"Datenschutz"empty" wurde hauptsaechlich ueb
die Arbeit und die Aufgaben des Datenschutzbeauftragten geredet sow
speziell von einer Seite immer wieder die implizite Unsecurityz der
heutigen Betriebssysteme Unix und DOS (PC) angefuehrt." Es fing dam
an, dass Wau zur Vorgeschichte sagte, dass er sich gegen die
Ueberschrift empty"Techno-terrorz"empty" gewandt hatte und erstmal
eine Begriffsdefinition wuenschte. Das sollte also die
"empty"Definitionsfrage"empty" sein? Die diversen zu diskutierenden
Begriffe? Eingeschraenkt auf das Thema "empty"Technoterrorz"empty"?"
Hier wurde ausfuehrlich das Muensteraner Pilotprojekt
empty"DAWIN"empty" vorgestellt

Data in a plot

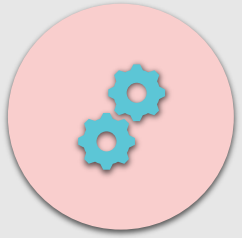


What can we do with text!

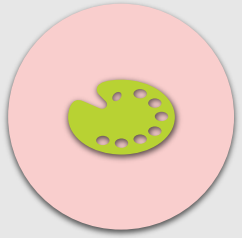
TOKENIZATION

ANNOTATION

TOPIC MODELLING



MACHINE LEARNING



VISUALISATION

Or as in the language of memes

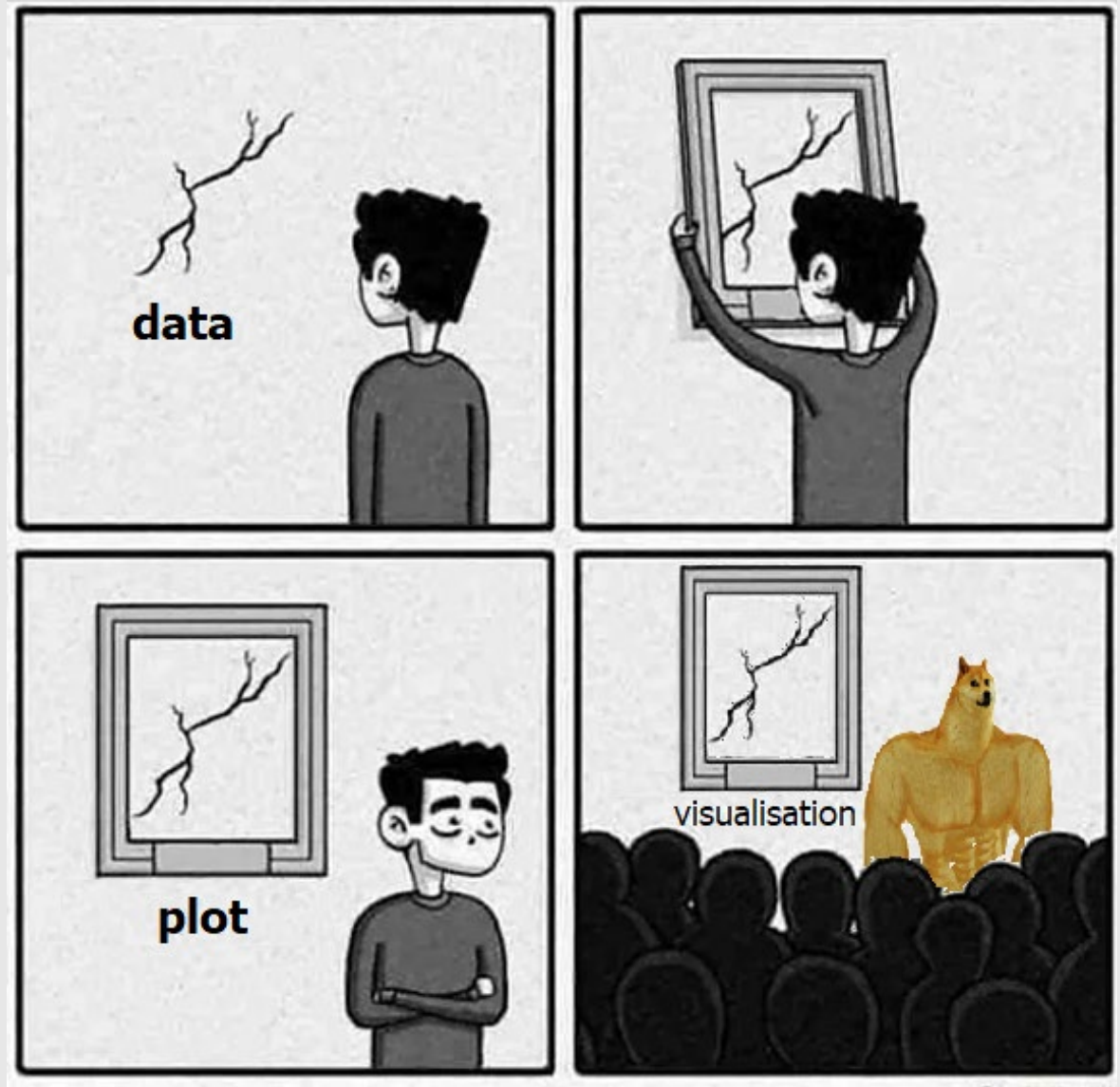
tokenization

annotation

topic modelling

machine learning

VISUALISATION


data
plot
visualisation

Plain text sample

<s>Visualise my corpus.</s>

<s>Text visualisation is awesome!</s>

<s>Visually representing the content of a text document is important.</s>

<s>Today is March 18th and Mahmoud is showing us how to visualising text at Lancaster University, well online!</s>

Added tags (annotations)

<s>Visualise [VERB] my [DET] corpus [NOUN].</s>

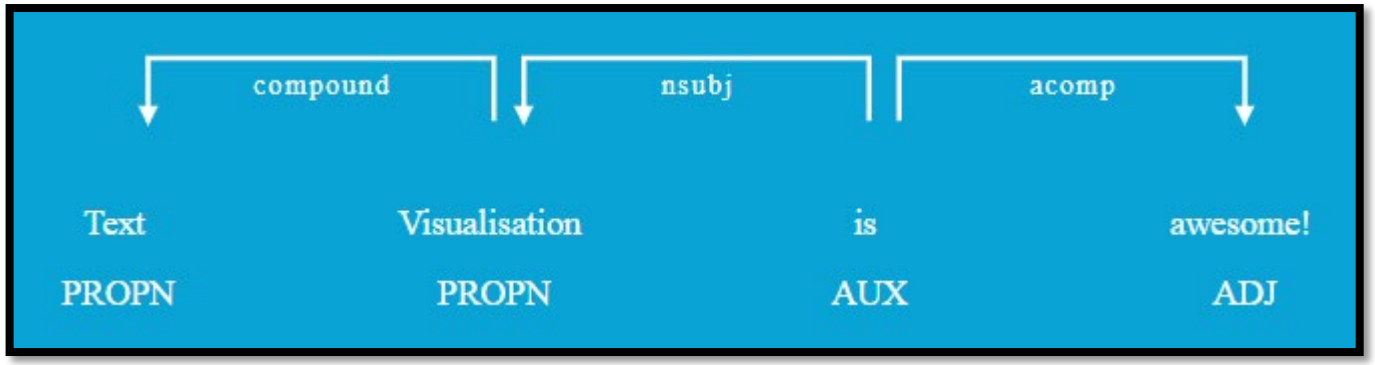
<s>Text [PROPN] visualisation [PROPN] is [AUX] awesome! [ADJ]</s>

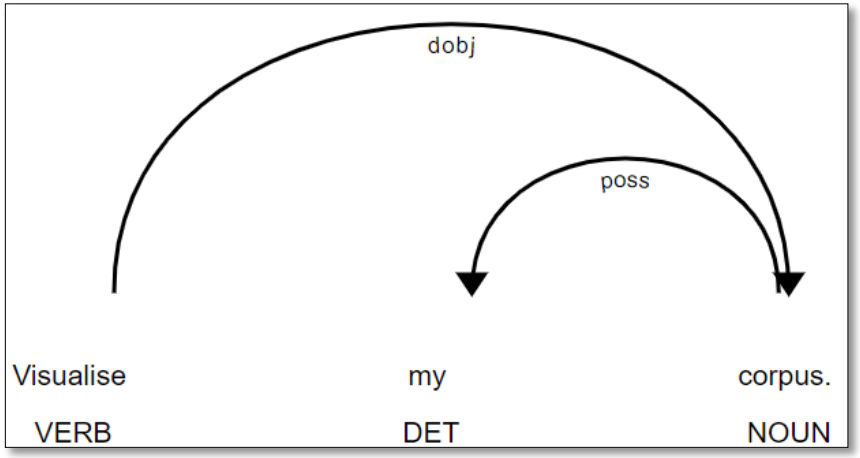
<s>Visually [ADV] representing [VERB] the [DET] content [NOUN] of [ADP] a [DET] text [NOUN] document [NOUN] is [AUX] important [ADJ].</s>

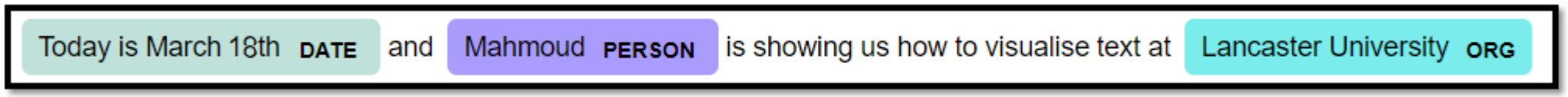
POST

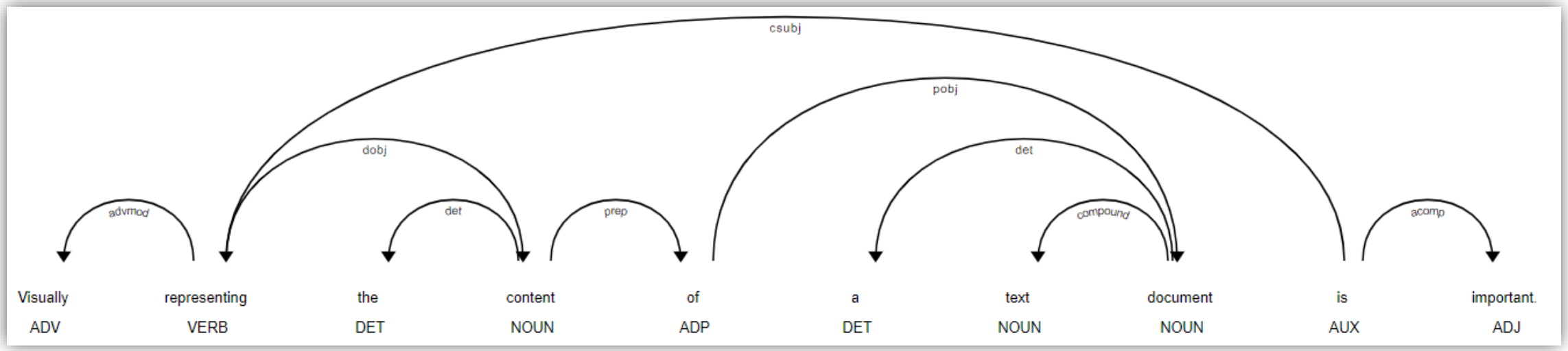
<s>{Today is March 18th}[DATE] and {Mahmoud}[PERSON] is showing us how to visualising text at {Lancaster University}[ORG].</s>

NER

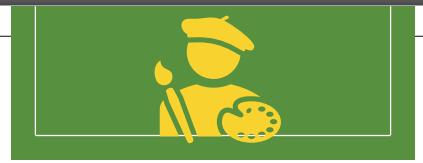








https://universaldependencies.org/docs/en/dep/



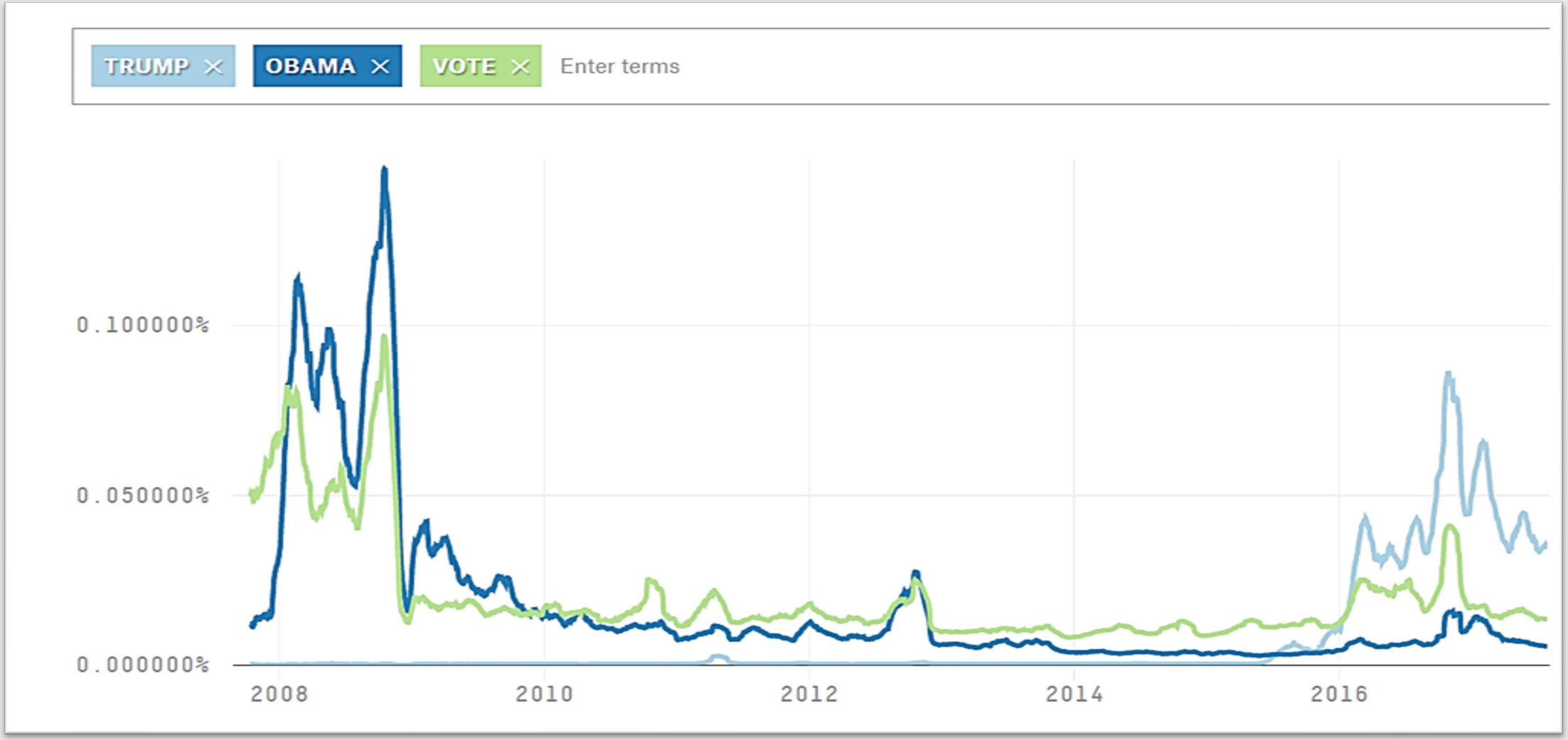
What can visualisation tell us?

Things we don't directly see!

Reddit.com

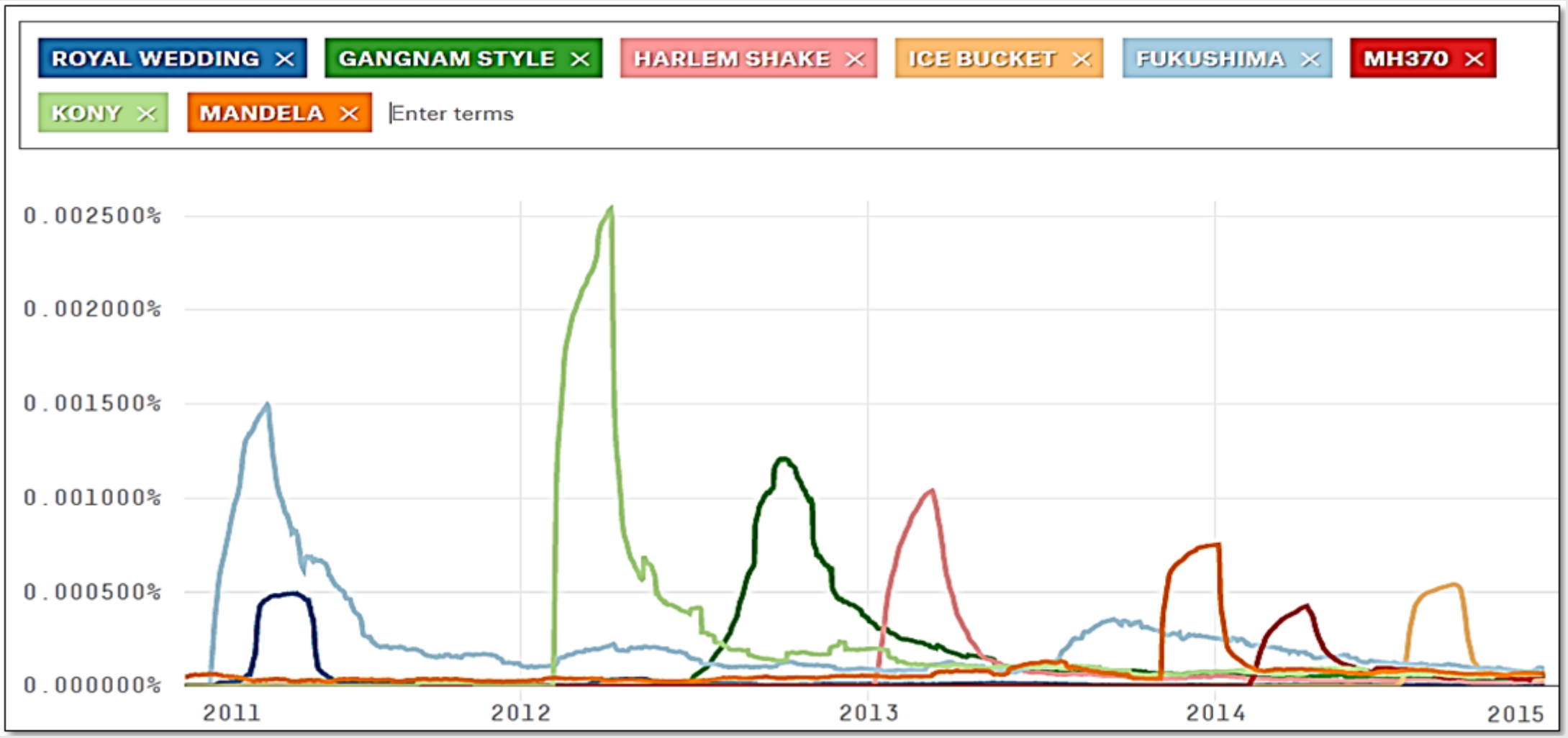
a social news aggregation, web content rating, and discussion website

TRUMP vs OBAMA vs VOTE



https://projects.fivethirtyeight.com/reddit-ngram/?keyword=trump.obama.vote&start=20071015&end=20170731&smoothing=22

How the Internet Talks!

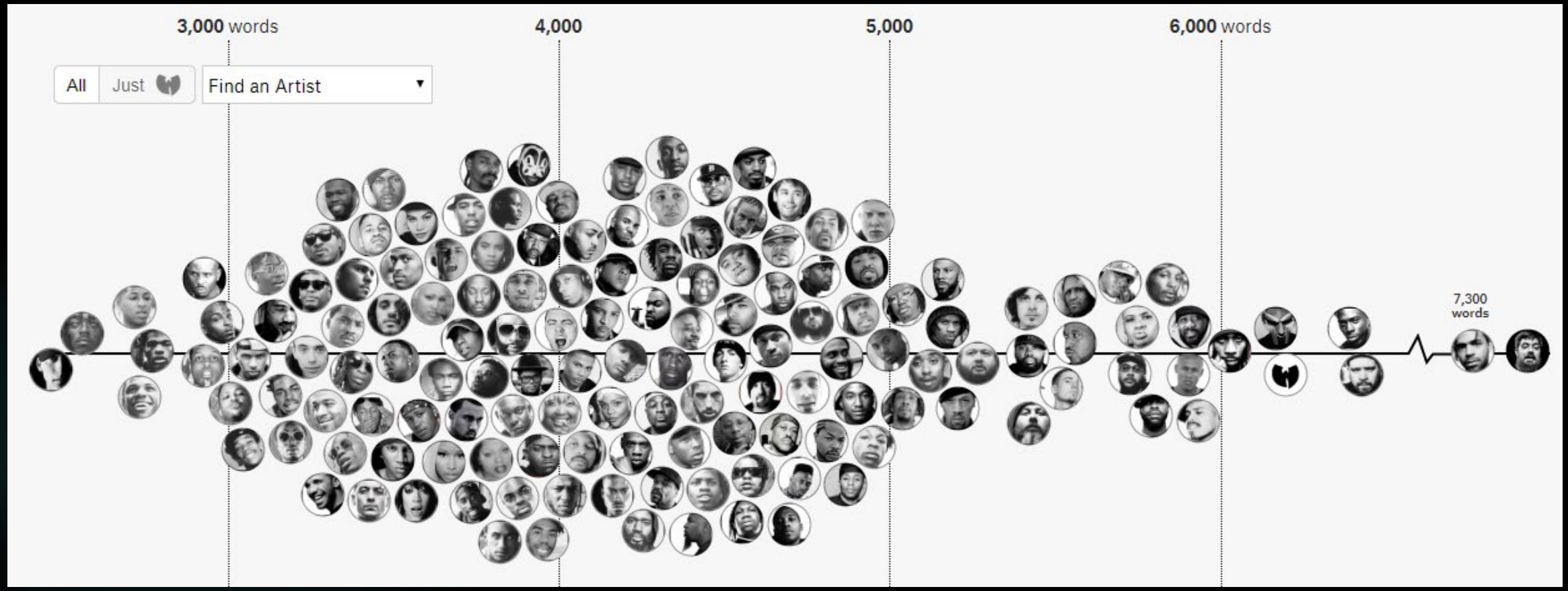


Major events/trends between 2011 - 2015

https://projects.fivethirtyeight.com/reddit-ngram

The Largest Vocabulary In Hip Hop

Rappers, ranked by the number of unique words used in their lyrics



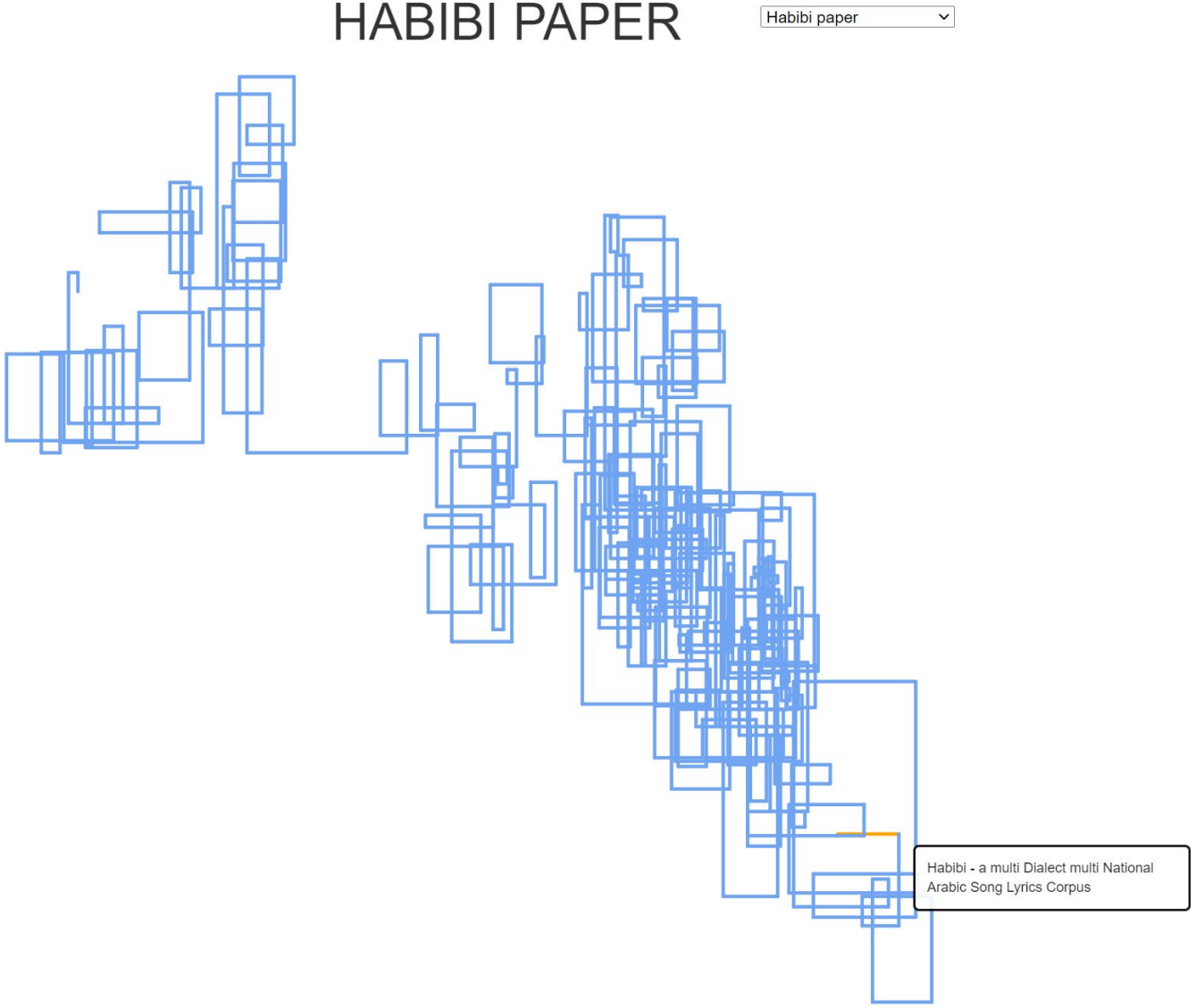
In May 2014, a study by Matt Daniels found that Aesop Rock's vocabulary in his music surpassed 85 other major hip-hop and rap artists, as well as Shakespeare's works and Herman Melville's Moby Dick, being named the largest vocabulary in Hip Hop [43][44] Wikipedia

Matt Daniels: https://pudding.cool/projects/vocabulary/

Stanford PhD Theses

https://nlp.stanford.edu/projects/dissertations/browser.html

Stanford's PhD dissertation abstracts from 1993-2008



Drawing one of my papers*

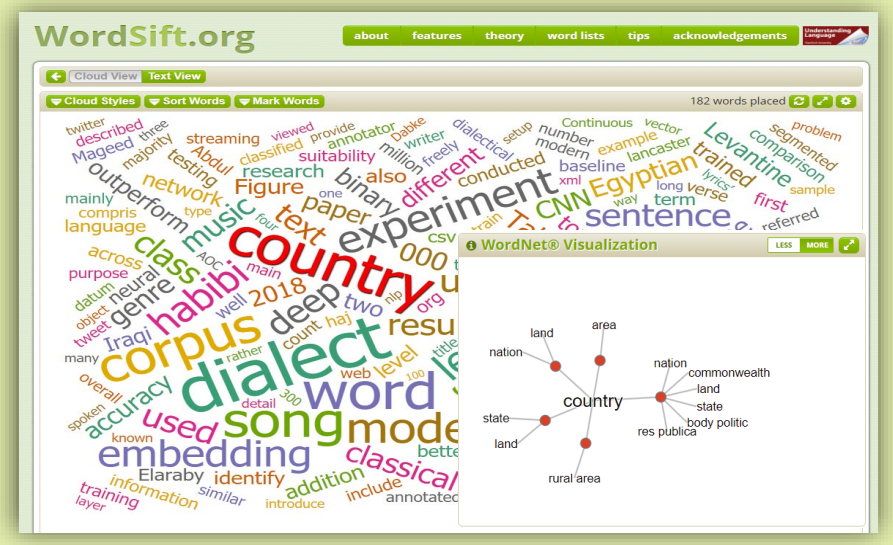
Try it out:

https://www.lancaster.ac.uk/staff/elhaj/draw/index.html#habibi

*https://www.lancaster.ac.uk/staff/elhaj/docs/habibi.pdf

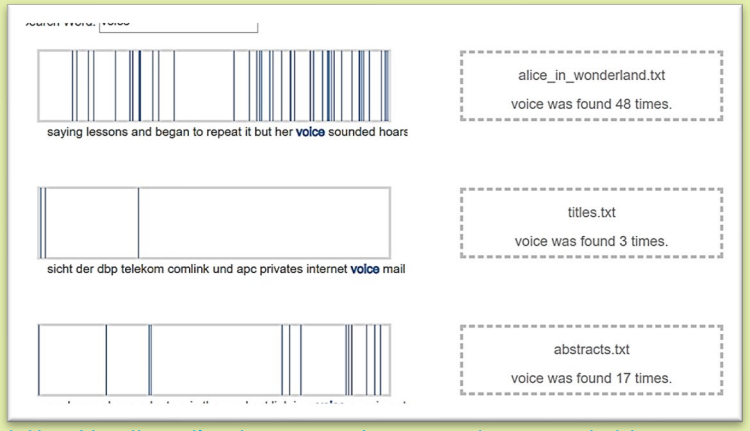
GitHub Tyler Rinker Blog

Stefanie Posavec
Understand more about text through Art!
Unlock hidden data!
https://youtu.be/y1wkGMLEktQ

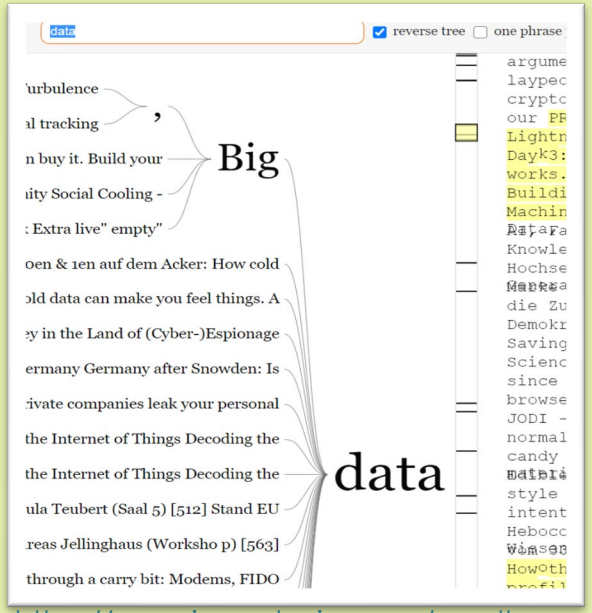


https://wordsift.org

http://wordwanderer.org/



http://vallandingham.me/concordance_plot/



https://www.jasondavies.com/wordtree

More

Text Visualisation Browser

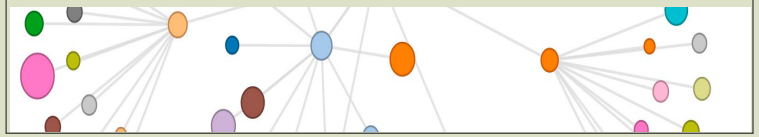
https://textvis.lnu.se/

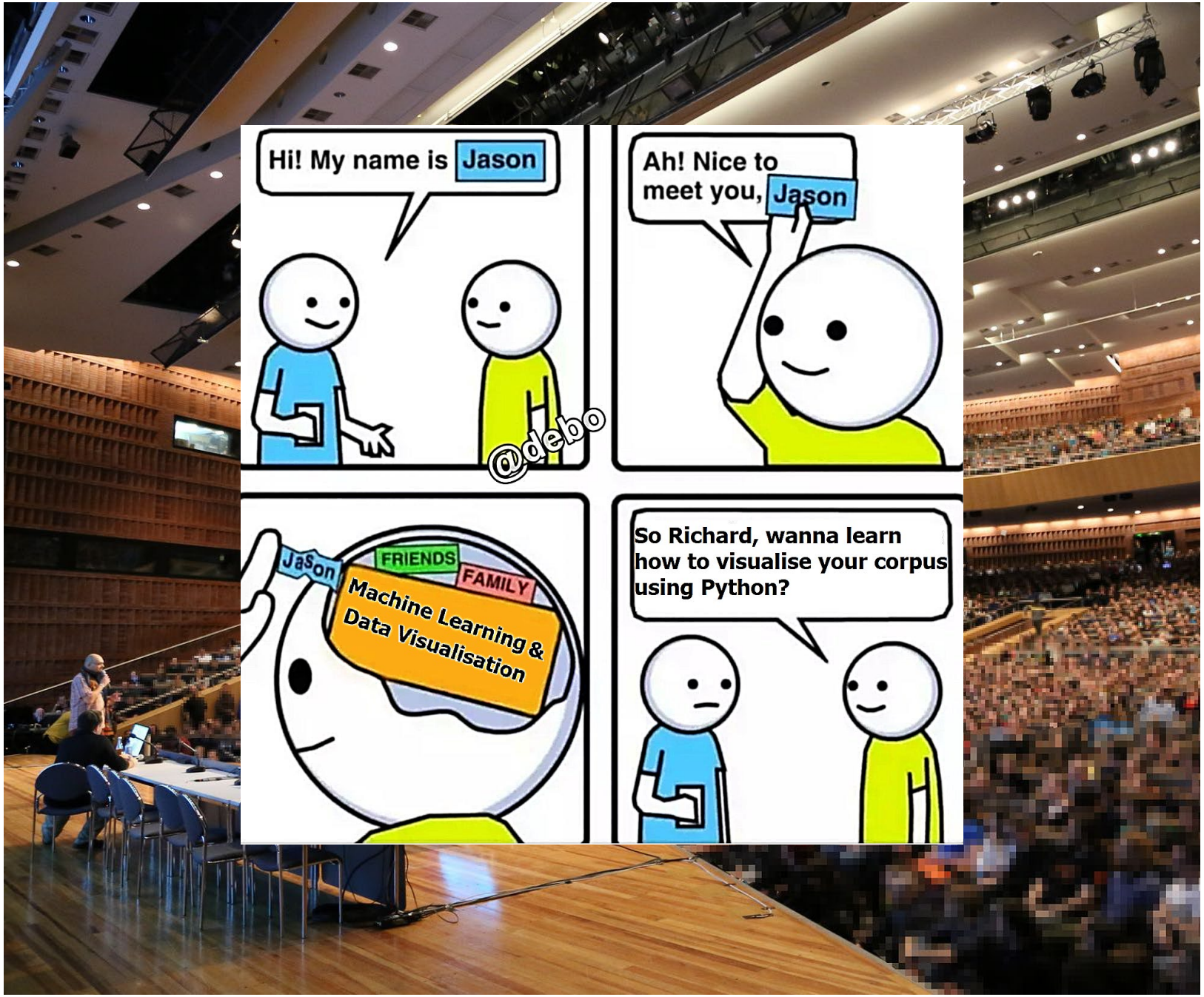
Data Driven Documents

https://d3js.org/

Jim Vallandingham

http://vallandingham.me





What are we visualising today?

Chaos Communication Congress talks (https://www.ccc.de/en/)

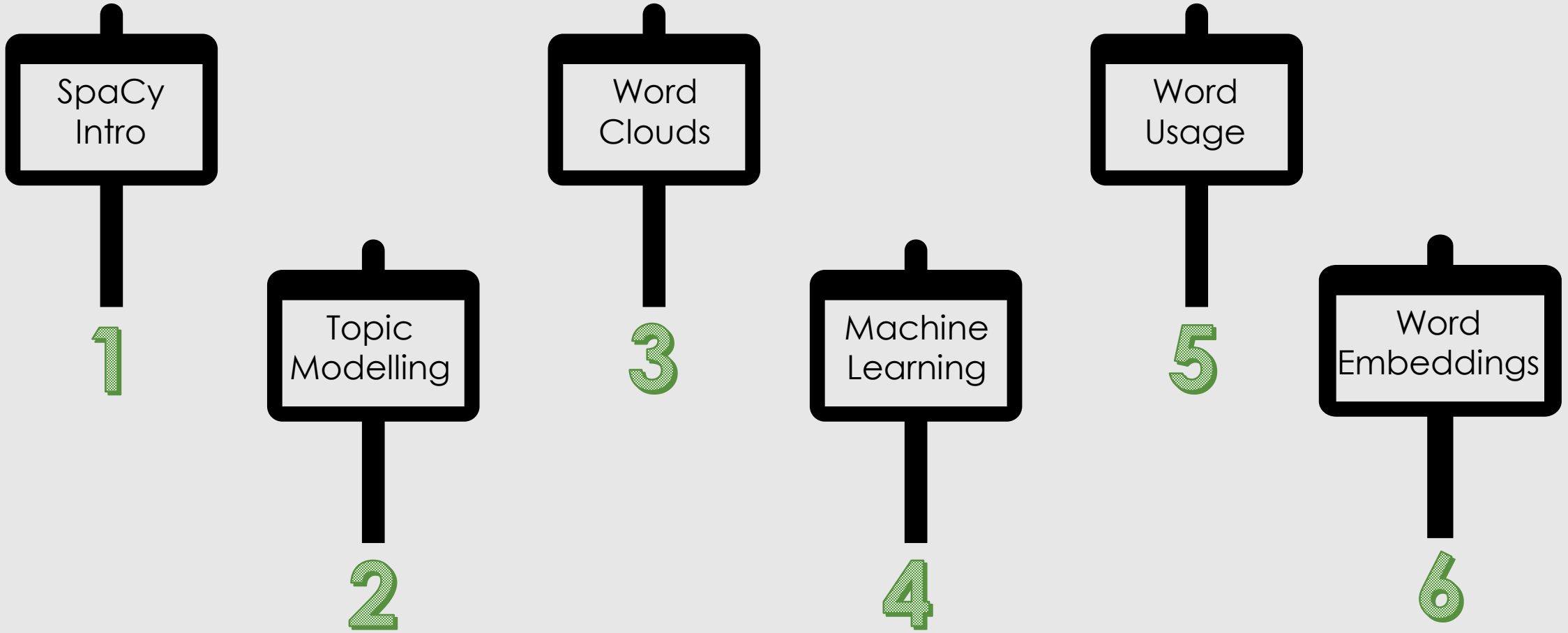
Tutorial Material:

https://github.com/drelhaj/NLP_ML_Visualization_Tutorial

Conference Data:

https://gitlab.com/maxigas/cccongresstalks/

Visualisation Roadmap



https://github.com/drelhaj/NLP_ML_Visualization_Tutorial