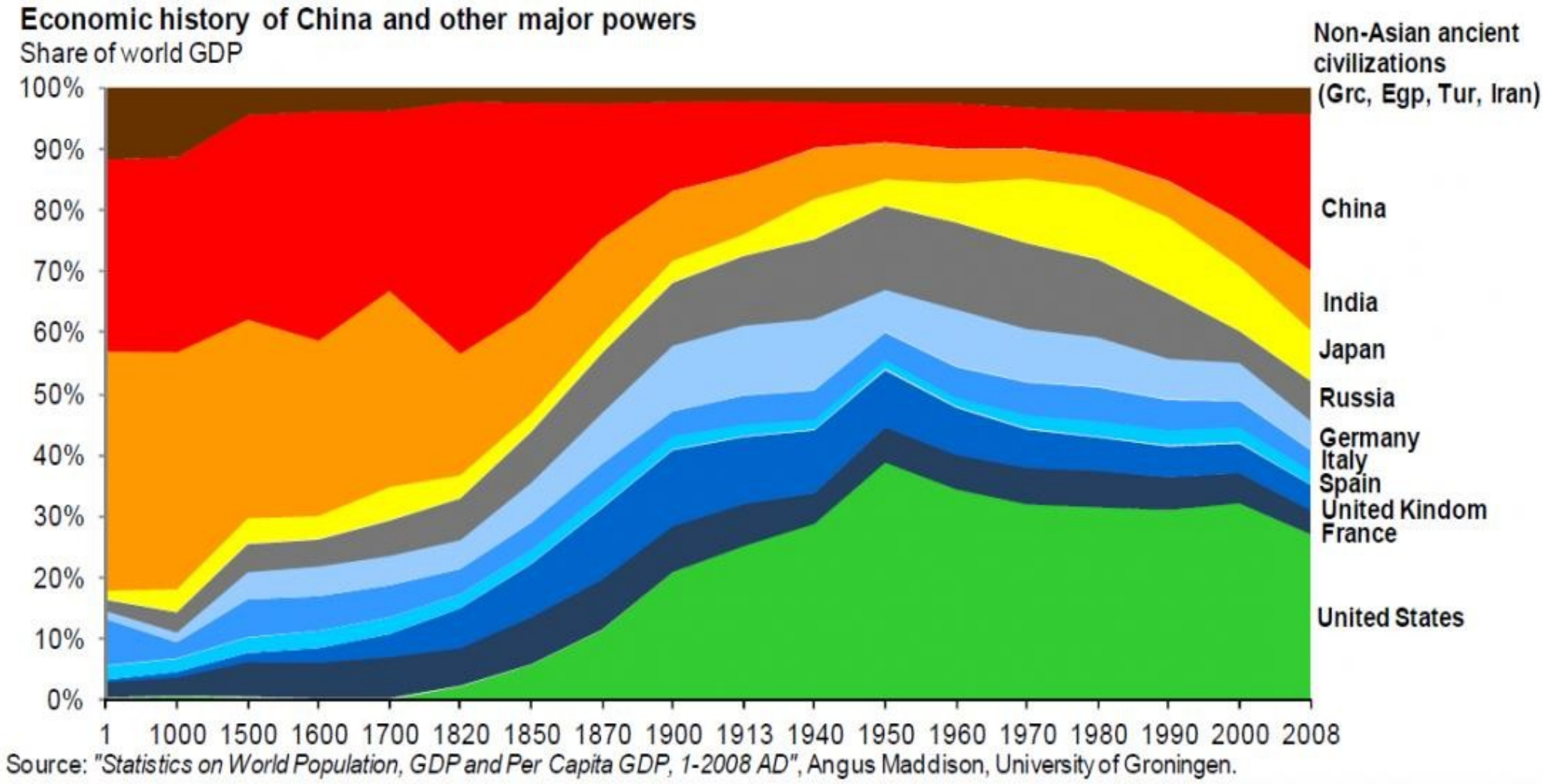
Data Visualisation
Lecture 6 – Visualising Text & Documents

Dr. Cathy Ennis

Learning Outcomes Week 6

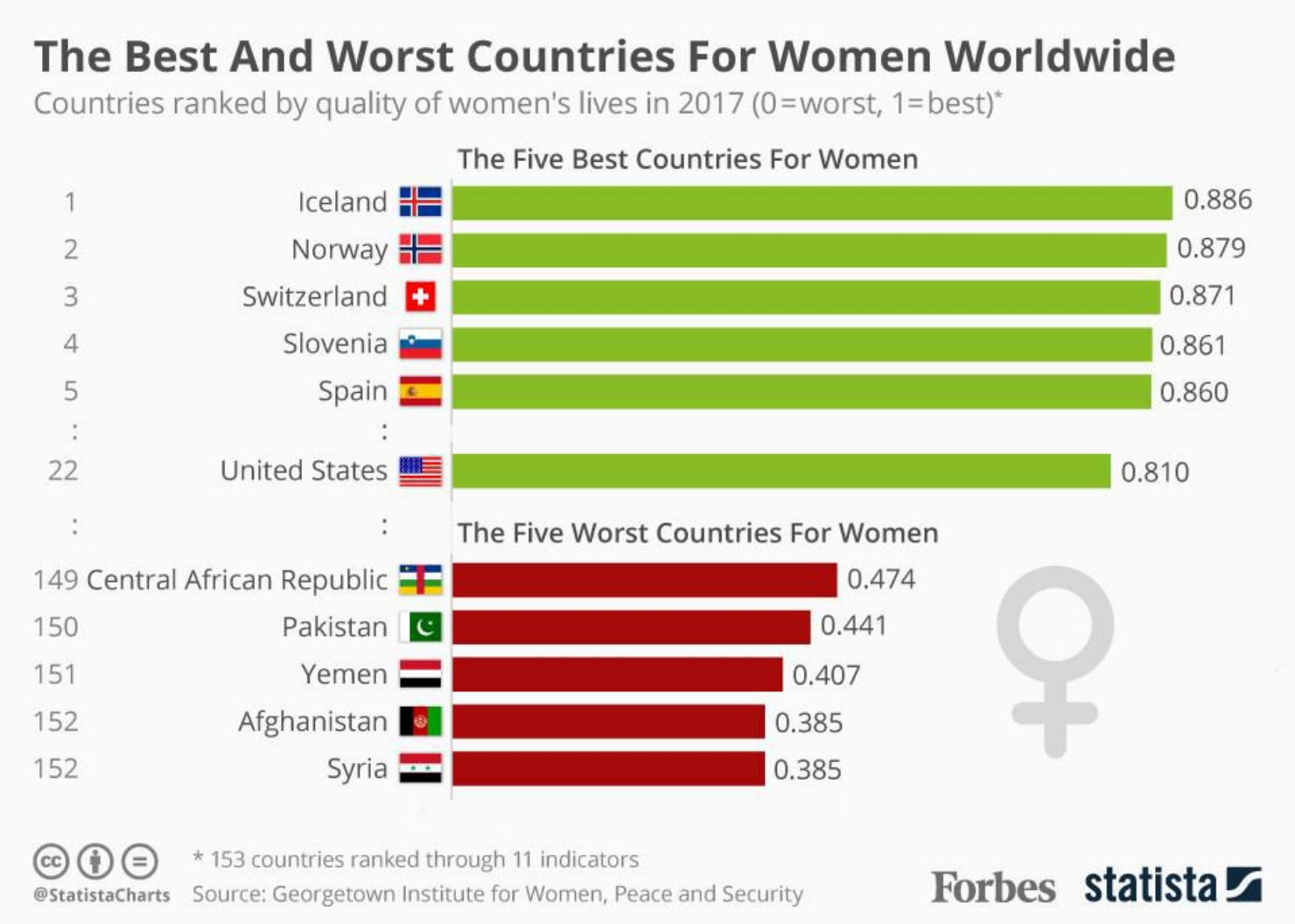
- Design effective Visualisations based on principles from perceptual psychology, cognitive science, graphic design and visual art
- Create and deploy successful data visualisations using leading software tools
- Demonstrate an understanding how visualisation is used in date journalism to communicate complex ideas and stories
- Demonstrate understanding how visualisation is used in story telling

Visualisation of the Week



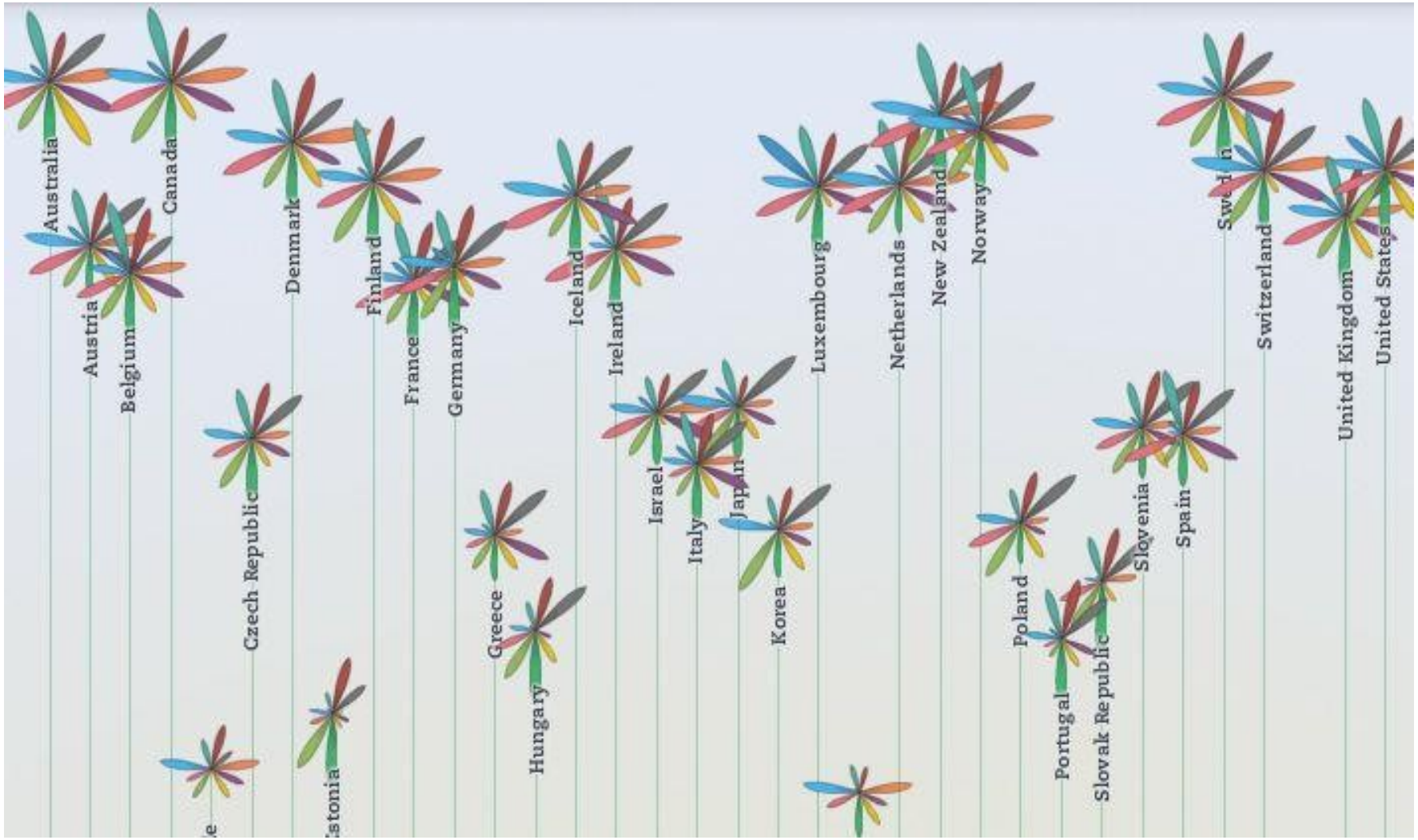
Visualisation of the Week

Data available: https://giwps.georgetown.edu/the-index/



https://www.forbes.com/sites/niallmccarthy/2017/11/08/the-best-and-worst-countries-for-women-worldwide-infographic/

(Un)Visualisation of the Week



Overview

In this lecture we will cover:

- Some Wrangling
  - Bag of words representation , n-grams
  - Word embeddings
- Single document visualisations
  - Tag & word clouds, Word trees
  - Semantics
  - Arc diagrams
- Corpus visualisations
- Lab + Project Work

Why Visualise Text?

- Understanding
  - get the “gist” of a document
- Grouping
  - cluster for overview or classification
- Comparison
  - compare document collections, or inspect evolution of collection over time
- Correlation
  - compare patterns in text to those in other data, e.g., correlate with social network

Text as Data

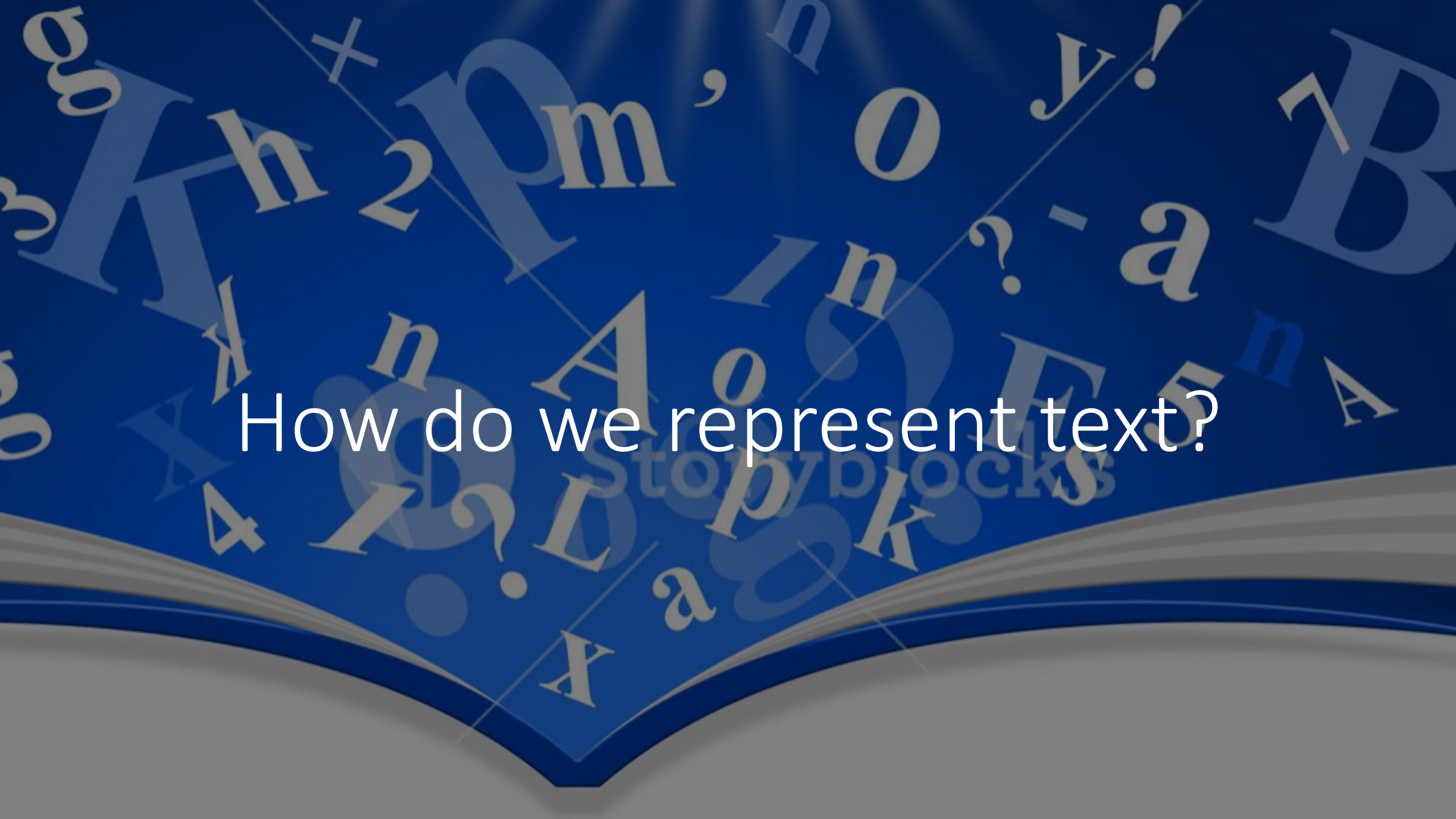
Documents

- Articles, books and novels
- E-mails, web pages, blogs
- Tags, comments
- Computer programs, logs

Collections of Documents

- Messages (e-mail, blogs, tags, comments)
- Social networks (personal profiles)
- Academic collaborations (publications)

DOCUMENT REPRESENTATIONS


How do we represent text?

Representing Text

- No numerical values

- Techniques to convert text to numerical values (feature extraction):
  - Bag of Words
  - Word N-grams
  - Character n-grams
  - Part of Speech
  - Vocabulary richness
  - Punctuation
  - Word length And others……

Bag Of Words

- Documents in text visualisation are often represented using a feature-vector model
  - Also called a bag of words

- In this approach, we look at the histogram of the words within the text, i.e. considering each word count as a feature

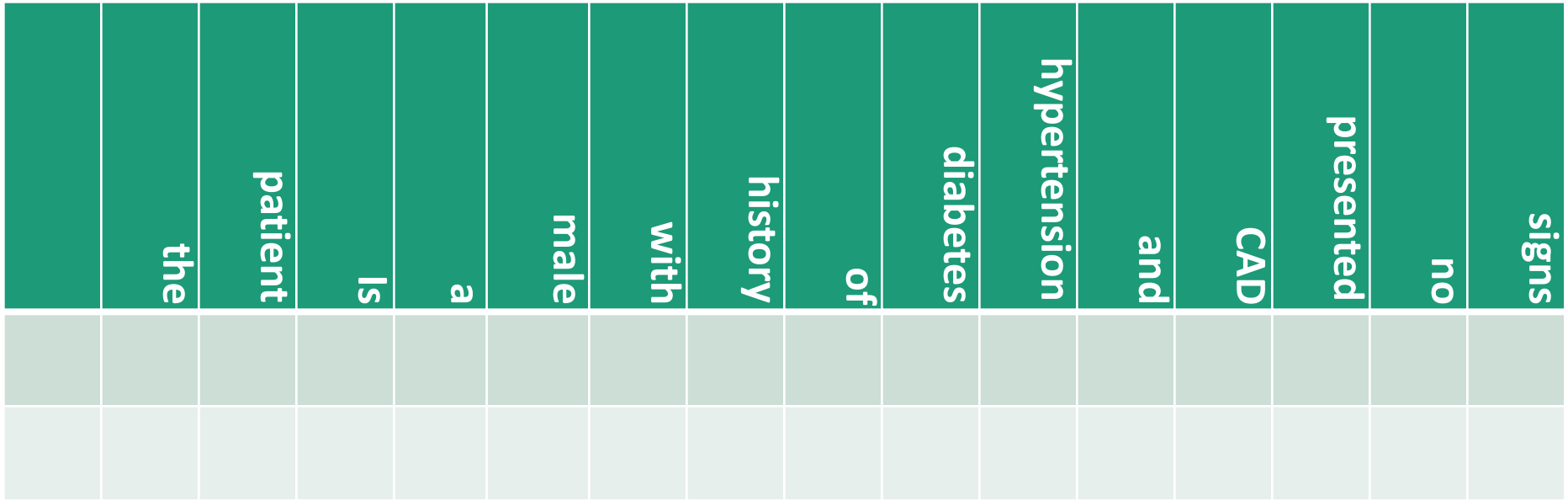
Bag of Words

Patient Record 1

The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.



Bag of Words - Binary

Patient Record 1

The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.

	the	patient	Is	a	male	with	history	of	diabetes	hypertension	and	CAD	presented	no	signs
PR1	1	1	1	1	1	1	1	1	1	1	1	1			
PR2															

Bag of Words - Binary

Patient Record 1

The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.

	the	patient	Is	a	male	with	history	of	diabetes	hypertension	and	CAD	presented	no	signs
PR1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
PR2	1	1	0	0	0	1	0	1	0	0	0	1	1	1	1

Bag of Words - Numeric

Patient Record 1

The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.

	the	patient	Is	a	male	with	history	of	diabetes	hypertension	and	CAD	presented	no	signs
PR1	1	1	1	2	1	1	1	1	1	1	1	1			
PR2	1	1				1		1				1	1	1	1

Bag of Words - Normalized

- The problem with counting term frequencies is that the frequently used terms become dominant in the document and begin to represent the document

- Normalized values are used instead showing the relative frequency

Bag of Words – TF-IDF

- Term Frequency - Inverse document frequency
- Statistical measure used to evaluate how important a word is to a document in a collection or corpus
- The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus

Limitations of Frequency Statistics

- Typically focus on unigrams (single terms)
- Often favours frequent (TF) or rare (IDF) terms
  - Not clear that these provide best description
- A “bag of words” ignores additional information
  - Grammar / part-of-speech
  - Position within document
  - Recognizable entities

N-grams

- An n-gram is a contiguous sequence of n items from a given sample of text or speech

- A unigram is one word, a bigram is a sequence of two words, a trigram is a sequence of three words etc
  - It is common to use more than one type of n-gram

- The items inside an n-gram may not have any relation between them apart from the fact that they appear next to each other

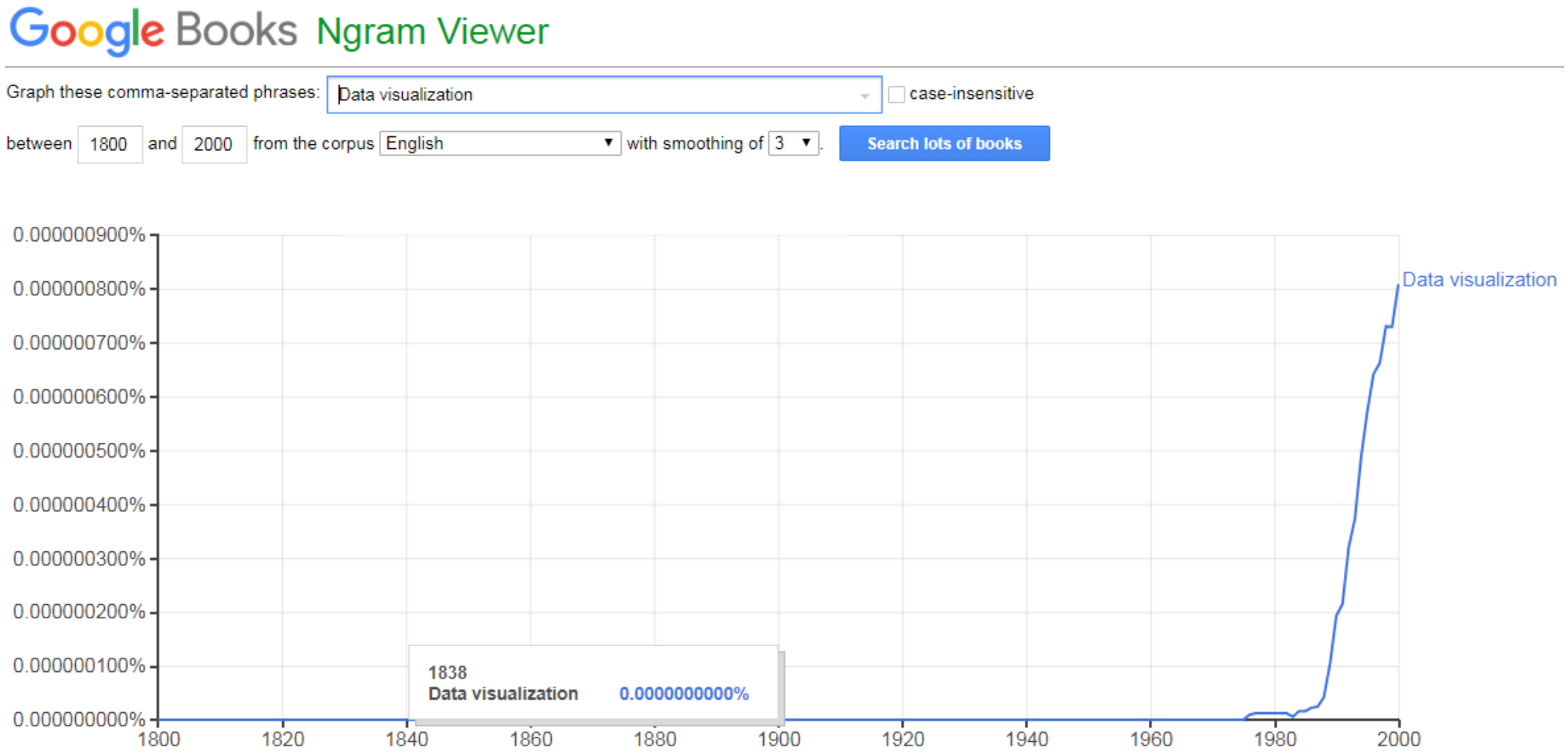
N-grams

The office building was demolished yesterday.

Contains 5 bigrams:

1. the office
2. office building
3. building was
4. was demolished
5. demolished yesterday

N-grams



https://books.google.com/ngrams

Too Many Features

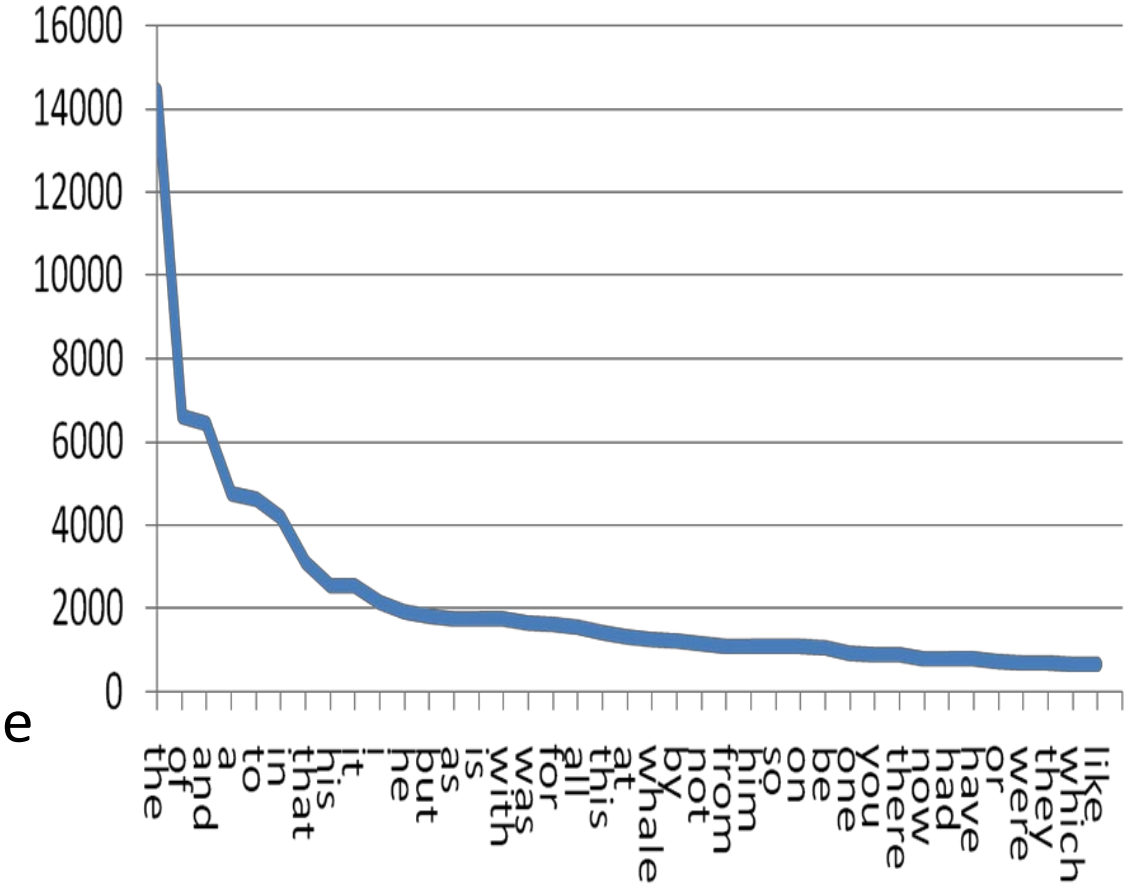
- One of the unique characteristics of textual data is the very large number of features involved – typically in the range of tens of thousands of features

- It is common practice to reduce the number of features as much as possible before attempting to build a useful model or visualisation

Zipf's Law

- Zipf's Law captures the distribution of the frequency of occurrence of words in natural language texts

- Zipf's law suggests two things:
  - High frequency words occur in too many documents to be useful in prediction
  - Low frequency words are too rare to be of value



http://www.wordcount.org/

Stop-Word Removal & Document Frequency

- We utilise two common feature reduction techniques:
  - Stop-word removal to remove the high frequency words – achieved using a list of known stop words or tf-idf
  - Document frequency to remove the low frequency words – remove terms that occur at most n times (1 <= n <= 3)

- These will typically reduce a feature set massively

always
am
amid
amidst
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are
aren't
i'm
immediate
in
inasmuch
inc
inc.
indeed
indicate
indicated
indicates
inner
inside
insofar
instead
into
inward
is
isn't
it
it'd
it'll
its
it's
somebody
someday
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
take
taken
taking
both
brief
but
by
c
came
can
cannot
cant
can't
caption
cause
causes
certain
certainly
changes
clearly
c'mon
co
co.
com
come
comes
mainly
make
makes
many
may
maybe
mayn't
me
mean
meantime
meanwhile
merely
might
mightn't
mine
minus
miss
more
moreover
most
mostly
mr
mrs
they're
they've
thing
things
think
third
thirty
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
till
to
together
too
took
toward

eg
eight
eighty
either
else
elsewhere
end
ending
enough
entirely
especially
et
etc
even
ever
evermore
every
everybody
everyone
everything
everywhere
ex
exactly
notwithstanding
novel
now
nowhere
o
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
one's
only
onto
opposite
or
other
via
viz
vs
w
want
wants
was
wasn't
way
we
we'd
welcome
well
we'll
went
were
we're
weren't
we've
what
whatever
what'll
what's
few
fewer
fifth
first
five
followed
following
follows
for
forever
former
formerly
forth
forward
found
four
from
further
furthermore
g
get
gets
getting
ourselves
out
outside
over
overall
own
p
particular
particularly
past
per
perhaps
placed
please
plus
possible
presumably
probably
provided
provides
q
que
quite
whereas
whereby
wherein
where's
whereupon
wherever
whether
which
whichever
while
whilst
whither
who
who'd
whoever
whole
who'll
whom
whomever
who's
whose
why
will

For Example: Stop Word Removal

Stop Word Removal

Patient Record 1

The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.

	the	patient	Is	a	male	with	history	of	diabetes	hypertension	and	CAD	presented	no	signs
PR1	1	1	1	2	1	1	1	1	1	1	1	1			
PR2	1	1				1		1				1	1	1	1

Stop Word Removal

Patient Record 1

The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.

	the	patient	Is	a	male	with	history	of	diabetes	hypertension	and	CAD	presented	no	signs
PR1	1	1	1	2	1	1	1	1	1	1	1	1			
PR2	1	1				1		1				1	1	1	1

Stop Word Removal

Patient Record 1

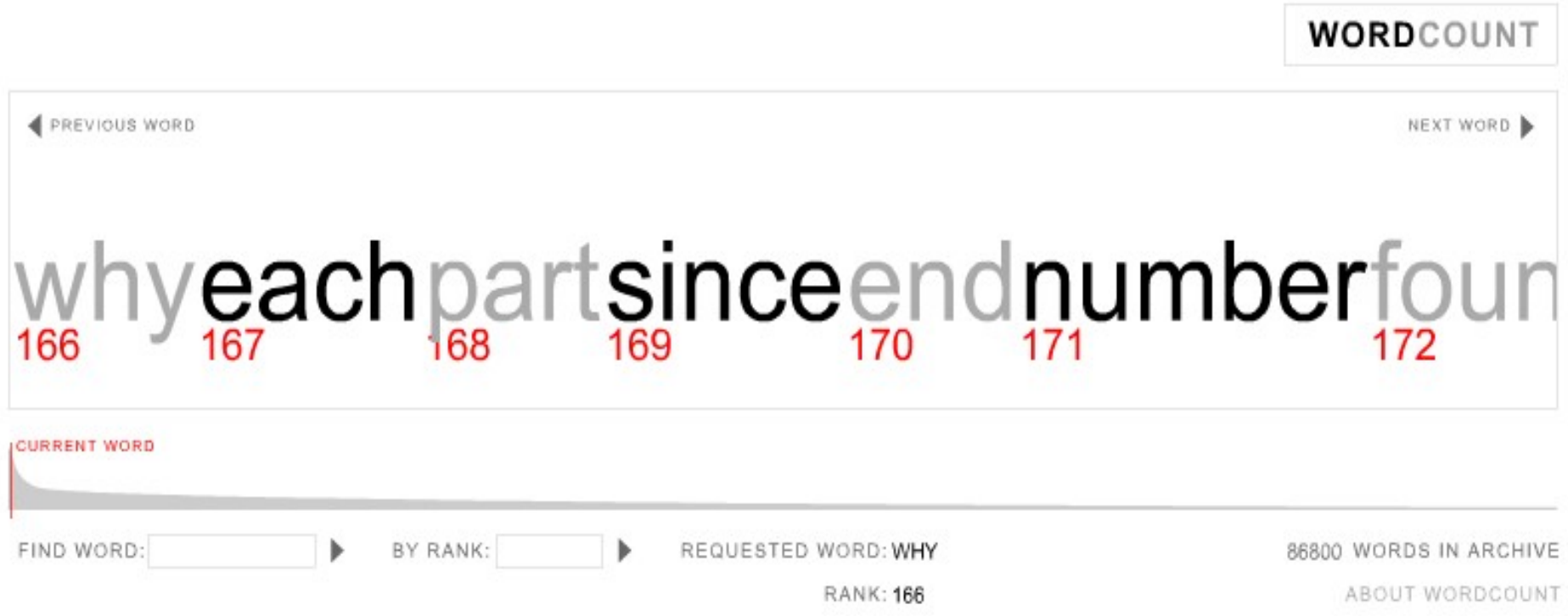
The patient is a male with a history of diabetes, hypertension and CAD.

Patient Record 2

The patient presented with no signs of CAD.

	patient	male	history	diabetes	hypertension	CAD	presented	signs
PR1	1	1	1	1	1	1		
PR2	1					1	1	1

Wordcount.org



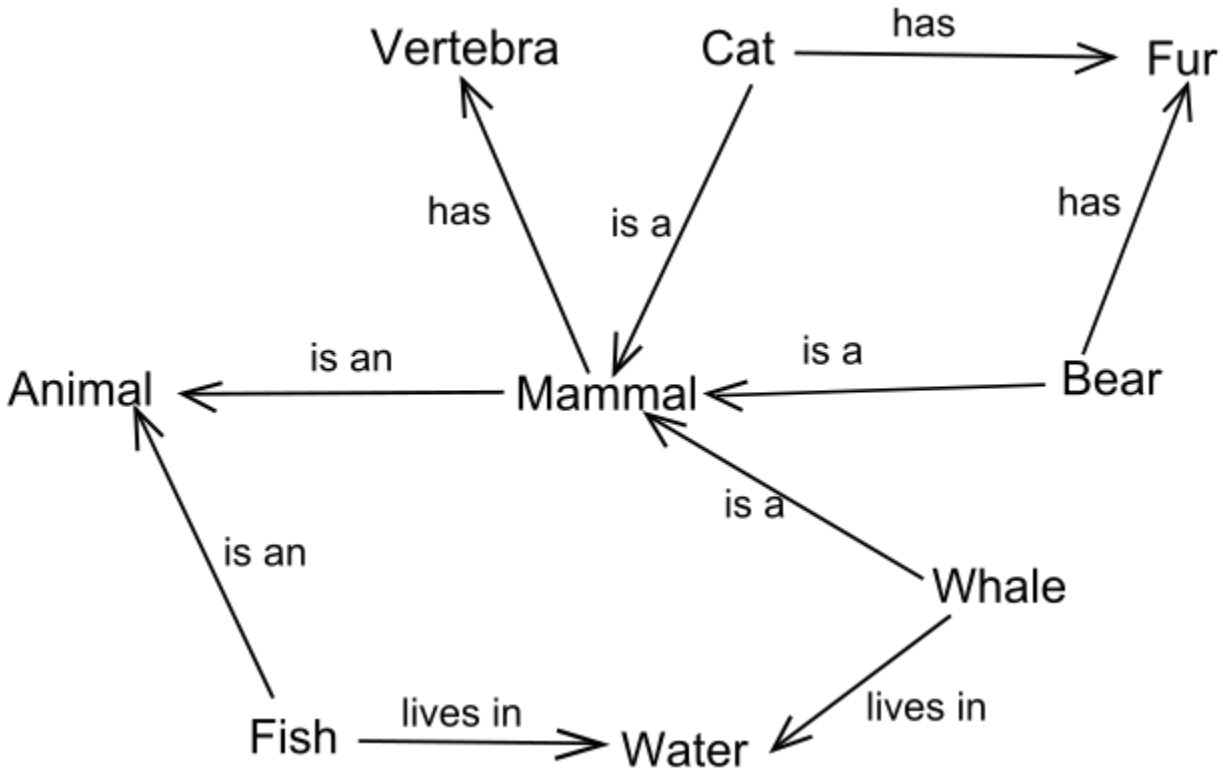
- Wordcount data currently comes from the British National Corpus®, a 100 million word collection of samples

Capturing Meaning

- Semantic networks
- Word embeddings

Semantic Networks

- Attempt to illustrate the semantics – or meaning – of text based on the relationship of words



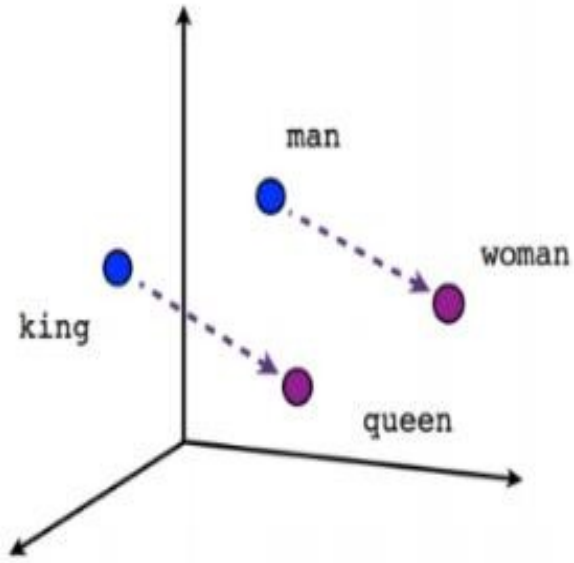
Word Embeddings

- Word embeddings are unsupervised techniques used to map words or phrases from a text to a corresponding vector of real numbers

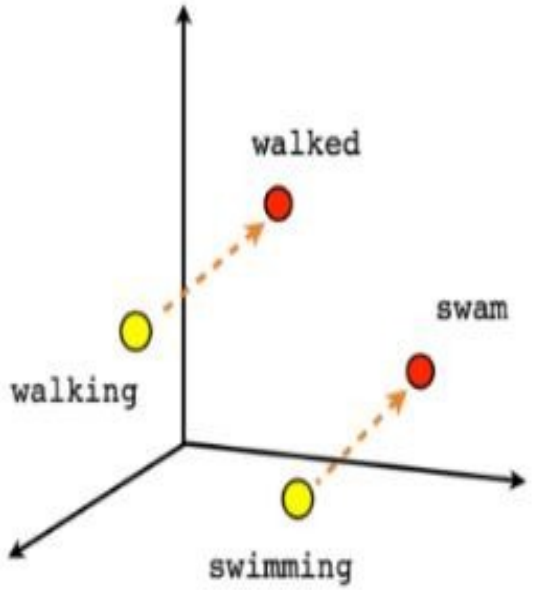
- This representation involves building a low dimensional continuous vector space from a high dimensional space (one dimension per word)

- The obtained vector space preserves the contextual similarity of words, therefore words that appear regularly together in text will also appear together in the vector space

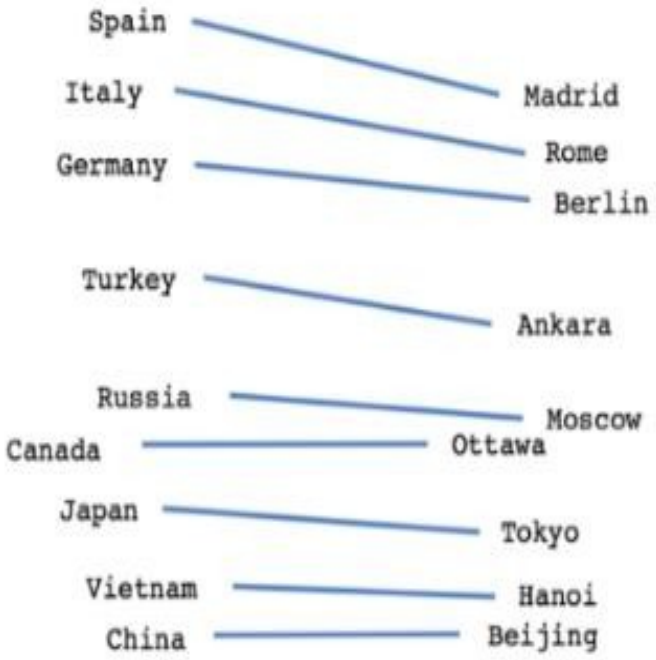
Word Embeddings



Male-Female



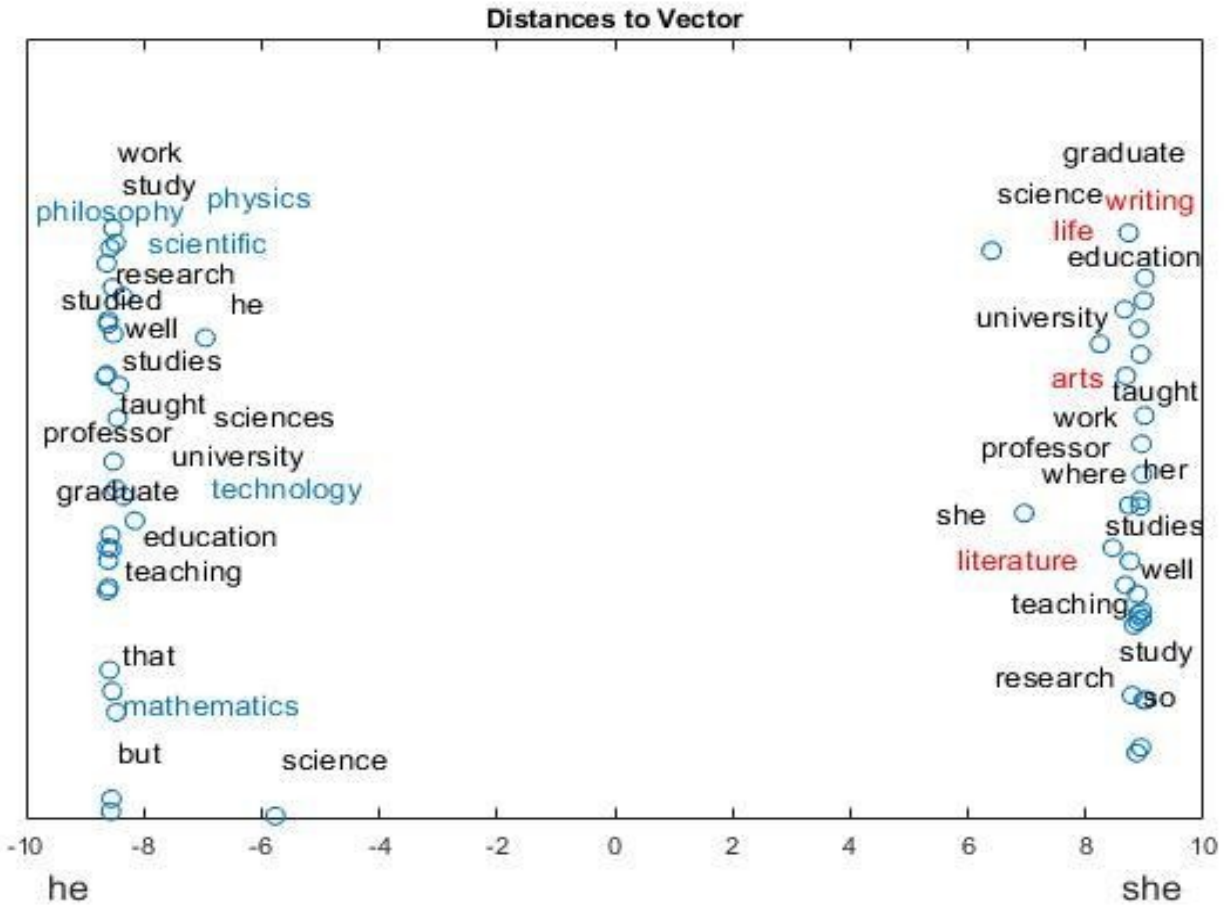
Verb tense



Country-Capital

https://www.tensorflow.org/tutorials/representation/word2vec

Word Embeddings



Llorens, Marisa (2018) "Text Analytics Techniques in the Digital World: Word Embeddings and Bias,"
Irish Communication Review: Vol. 16: Iss. 1, Article 6.

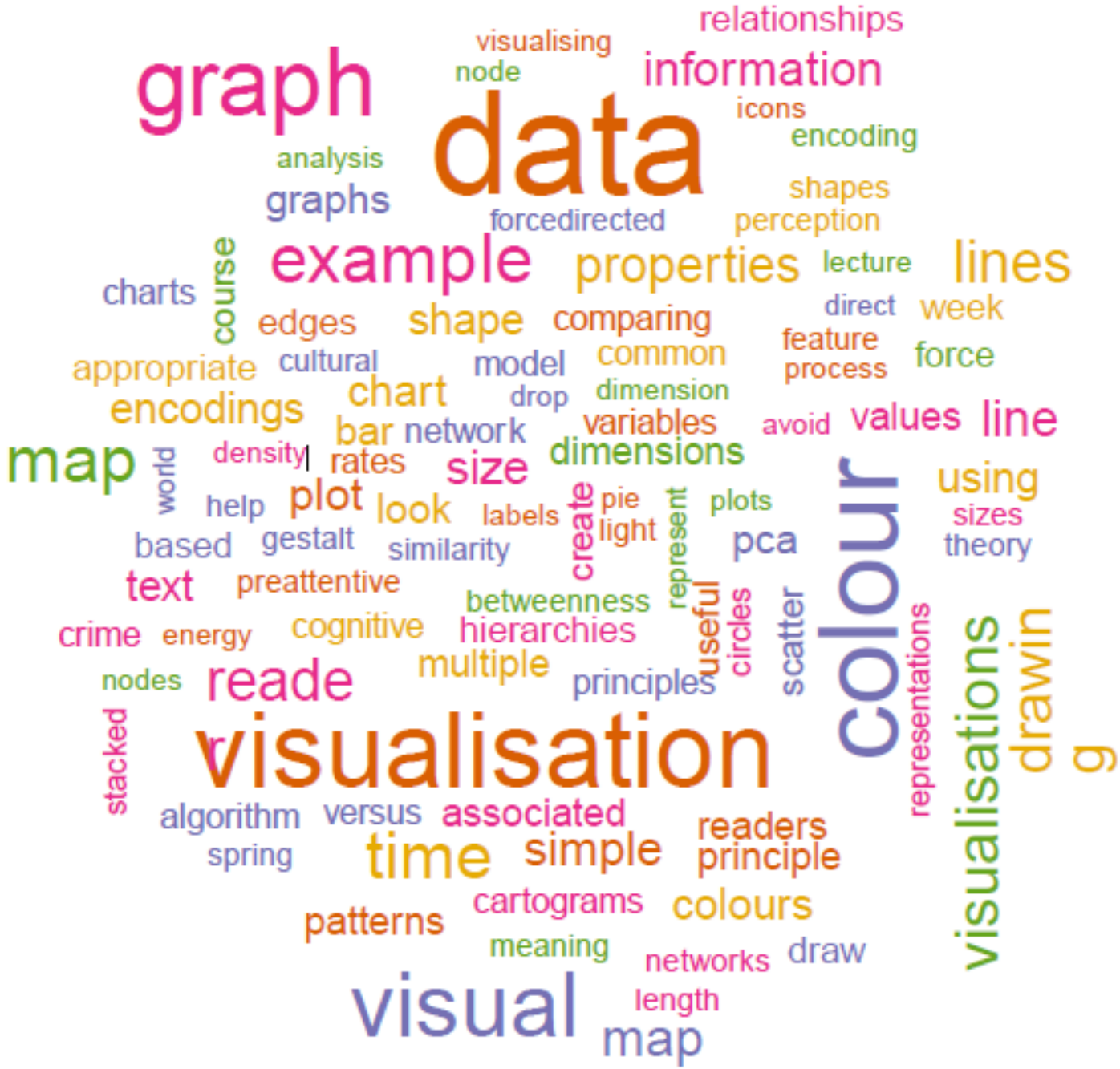
SINGLE DOCUMENT VISUALISATIONS

Single Document Visualisations

- In single document visualisations we try to quickly give readers a sense of the contents of a document without requiring that they read the actual text

- There is lots of debate about whether or not this is a sensible thing to do!

Word Clouds


relationships
visualising
graph
node
information
data
icons
encoding
analysis
shapes
graphs
forcedirected
perception
example
properties
lecture
lines
course
charts
direct
week
edges
shape
comparing
feature
process
force
appropriate
cultural
model
common
chart
drop
dimension
encodings
bar
network
variables
avoid
values
line
map
density
rates
size
dimensions
world
colour
using
help
plot
look
labels
create
pie
plots
sizes
light
represent
based
gestalt
similarity
pca
theory
text
preattentive
betweenness
useful
circles
scatter
representations
crime
energy
cognitive
hierarchies
visualisations
drawin
multiple
nodes
reade
principles
stacked
visualisation
g
algorithm
versus
associated
readers
spring
time
simple
principle
cartograms
colours
patterns
meaning
networks
draw
visual
length
map

Word Clouds

- Word clouds are probably the most common single document visualisation tool

- The important things to consider in developing word clouds:
  - What words will be displayed?
  - How will words be scaled?
  - How will the sizes be normalised?
  - What ordering of words is used?
  - How do we measure overlaps/collisions?


VISUALISATION
DATA
VISUAL
COLOUR
READER
GRAPH
EXAMPLE
MAP
TIME
INFORMATION
SIMPLE
USE
DRAWING
CHART
USED
BAR
VISUALISATIONS
PROPERTIES
MULTIPLE
USING
SMALL
DIMENSIONS
MAPS
ASSOCIATED
NETWORKS
PRE-ATTENTIVE
PCA
RELATIONSHIPS
CARTOGRAMS
SHAPE
DIFFERENT
BASED
SIZE
FORCE
ONE
PRINCIPLES
ALGORITHM
SCATTER
ENCODING
CRIME
VALUES
STACKED
COMMON
CREATE
RATES
VISUALISING
PROCESS
POSSIBLE
LENGTH
IMPORTANT
DIRECT
FRY'S
SCALE
SIMILARITY
LECTURE
THEORY
CHANGE
AVOID
NUMBER
GRAPHS
COMPARING
SIZES
OFTEN
MANY
WORK
ICONS
FONTS
WELL
SPRING
CASE
MAKE
EXPLORATION
PRINCIPLE
USEFUL
CHARTS
ANALYSIS
COLOURS
VARIABLE
RANDOM
MEANING
AREA
FEATURE
PROJECTION
APPROPRIATE
REPRESENTATIONS
LINE
FORCE-DIRECTED
WAY
NETWORK
THINGS
GESTALT
TYPES
SHAPES
TWO
PIE
TREE
POINTS
ENCODINGS
MULTI
HIERARCHIES
DRAW
TEXT
PLOTS
BC
GOOD
RED
SET
DROP
CIRCLES
DESIGN
MAY
HELP
WEEK
PLOT
LOOK
KEY
BEN
LINES
ALSO
COURSE
PERCEPTION
GIVEN
VERSUS
CORRELATION
EDGES
LABELS
LIGHT
BETWEENNESS
DENSITY
NODES
NODE
ENERGY
SINGLE
WORLD
DIMENSION
NEED
COGNITIVE
READERS
STEPS
ÄÖT
REPRESENT

Wordle Algorithm

Wordle answers these questions as follows:

- What words will be displayed?
  - Wordle removes stopwords which otherwise clutter the display
  - It contains lists of stopwords in 26 languages and by default uses whichever list has the most words in the input text
- How will words be scaled?
  - Font size is proportional to word frequency

Beautiful Visualisation (Chapter 3), Jonathan Feinberg, Eds Steele & Illinsky, 2010

Wordle Algorithm

Wordle answers these questions as follows:

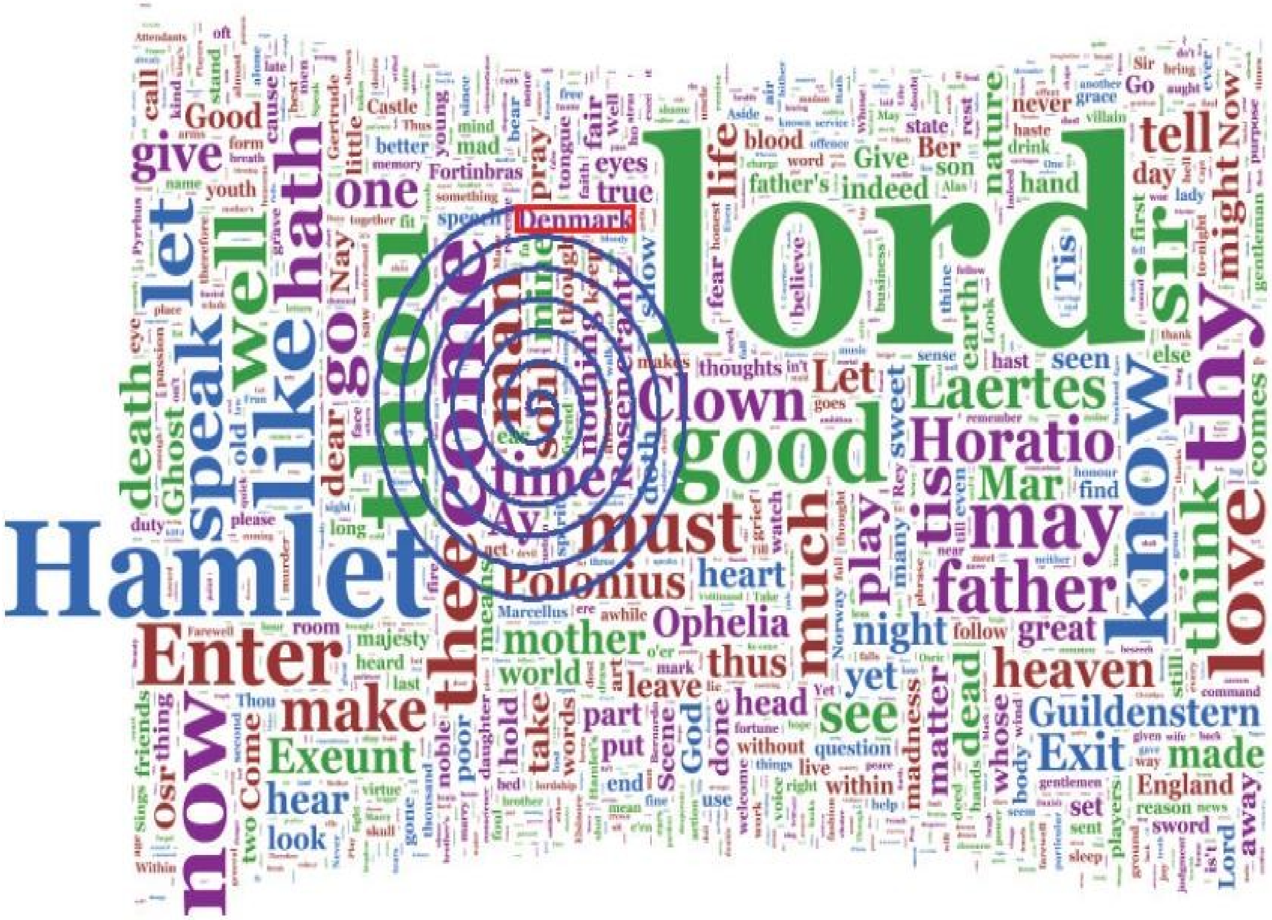
- How will the sizes be normalised?
  - The overall playing pitch size is determined and available fonts are scaled so the biggest word will fit
  - All word frequencies are then normalised to give a range of smallest to biggest font
- What ordering of words is used?
  - Wordle allows a number of different placement strategies - clustering is probably the most interesting
  - Results are approximate though

Beautiful Visualisation (Chapter 3), Jonathan Feinberg, Eds Steele & Illinsky, 2010

Wordle Algorithm

- For each word w in sorted words:
- placementStrategy.place(w)

- While w intersects any previously placed words:
- Move w a little bit along a spiral path



Beautiful Visualisation (Chapter 3), Jonathan Feinberg, Eds Steele & Illinsky, 2010

Wordle Algorithm

- How do we measure overlaps/collisions
  - Wordle is quite sophisticated and actually measures the internal shapes of words, other approaches use a simple bounding box

Beautiful Visualisation (Chapter 3), Jonathan Feinberg, Eds Steele & Illinsky, 2010

Word Cloud Challenges

- But Wordle present some challenges

1. Semantics – syntactic
2. Context
3. Comparison
4. What is the relationship between the words?

Word Clouds

Word cloud of Tea Party feelings about Obama

Largest words are

like

policy

http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/

Word Clouds

Word cloud of Tea Party feelings about Obama

Largest words are

like policy

- Stop word removal effect
- I really don’t like Obama’s policy on gun control
- I like nothing about …
- He is not like …

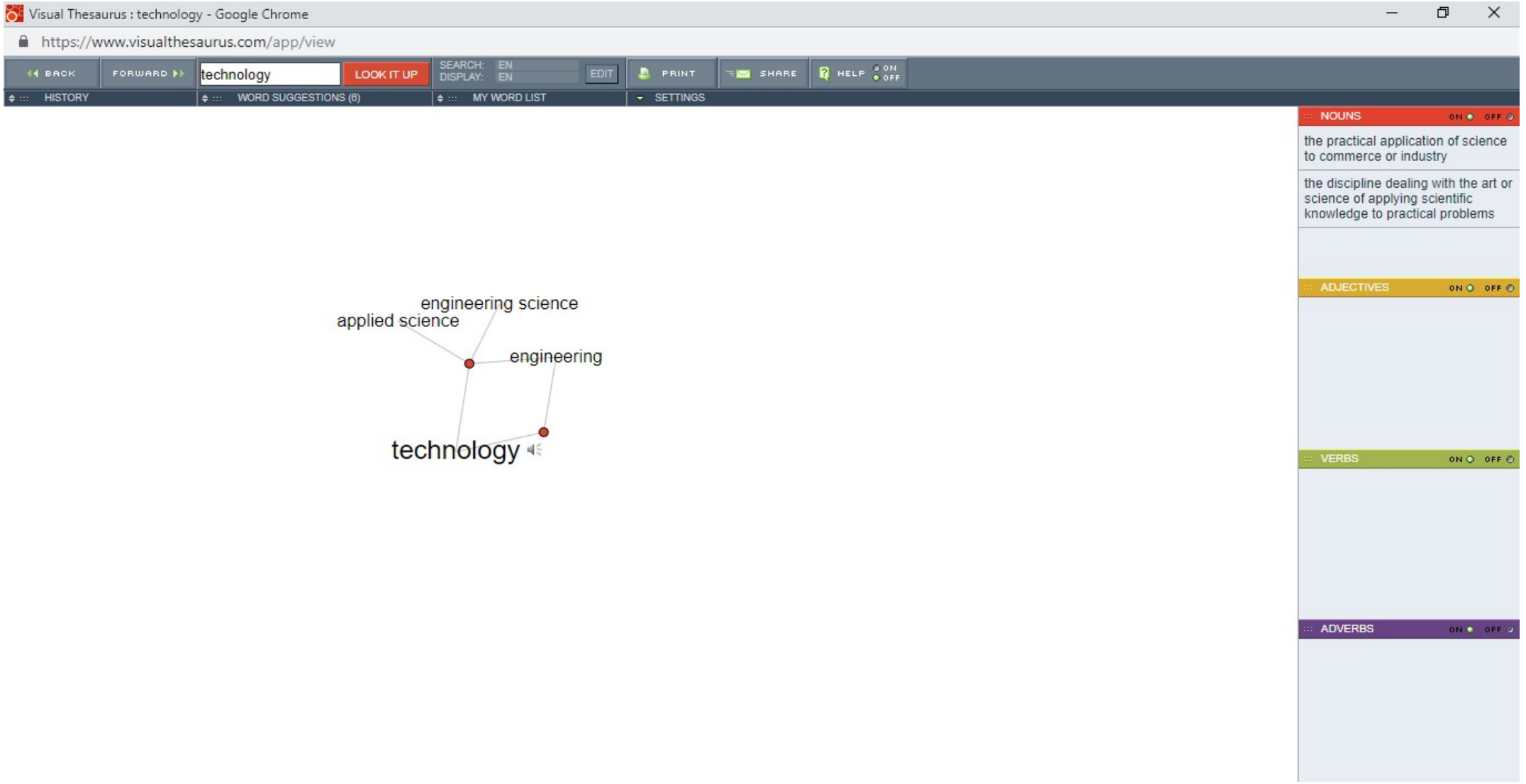
http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/

Semantic Networks

- Semantic Networks address some of the challenges of a wordle

- Semantic Networks attempt to illustrate the semantics – or meaning – of text based on the relationship of words in a traditional thesaurus (Wordnet)

Visual Thesaurus



Word Trees

- A word tree places a tree structure onto the words that follow a particular search term, and uses that structure to arrange those words spatially

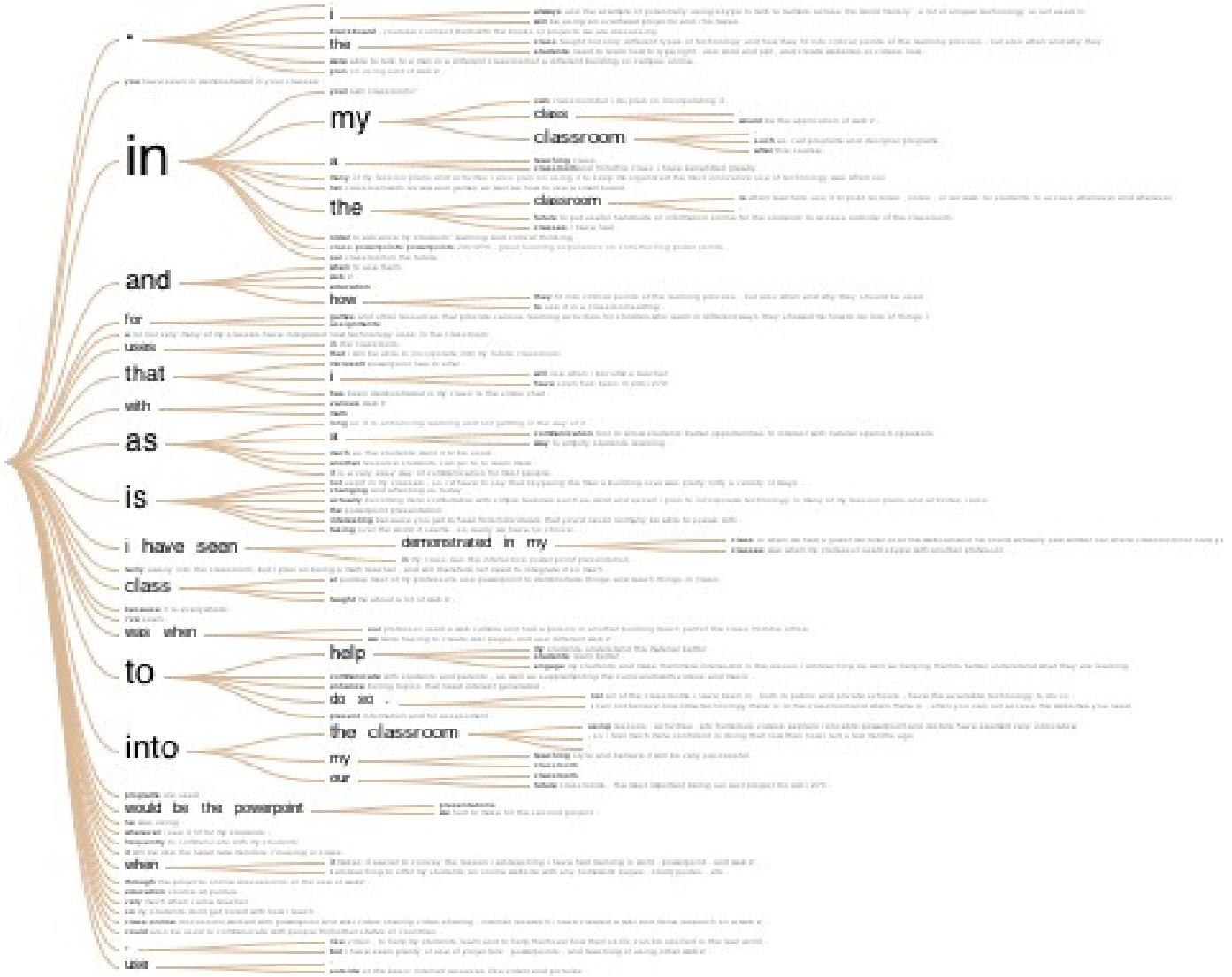
- Interactions are key in allowing the viewer to explore relationships

The Word Tree, an Interactive Visual Concordance, Martin Wattenberg and Fernanda B. Viégas, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 14, NO. 6, NOVEMBER/DECEMBER 2008
ManyEyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)

Word Trees

105
hits

technology



Word Trees

- Implementing word trees is surprisingly simple
- Words are scaled according to frequency - similarly to word clouds
- The data structure used is a **suffix tree**, which has been common in computer science string-processing for decades

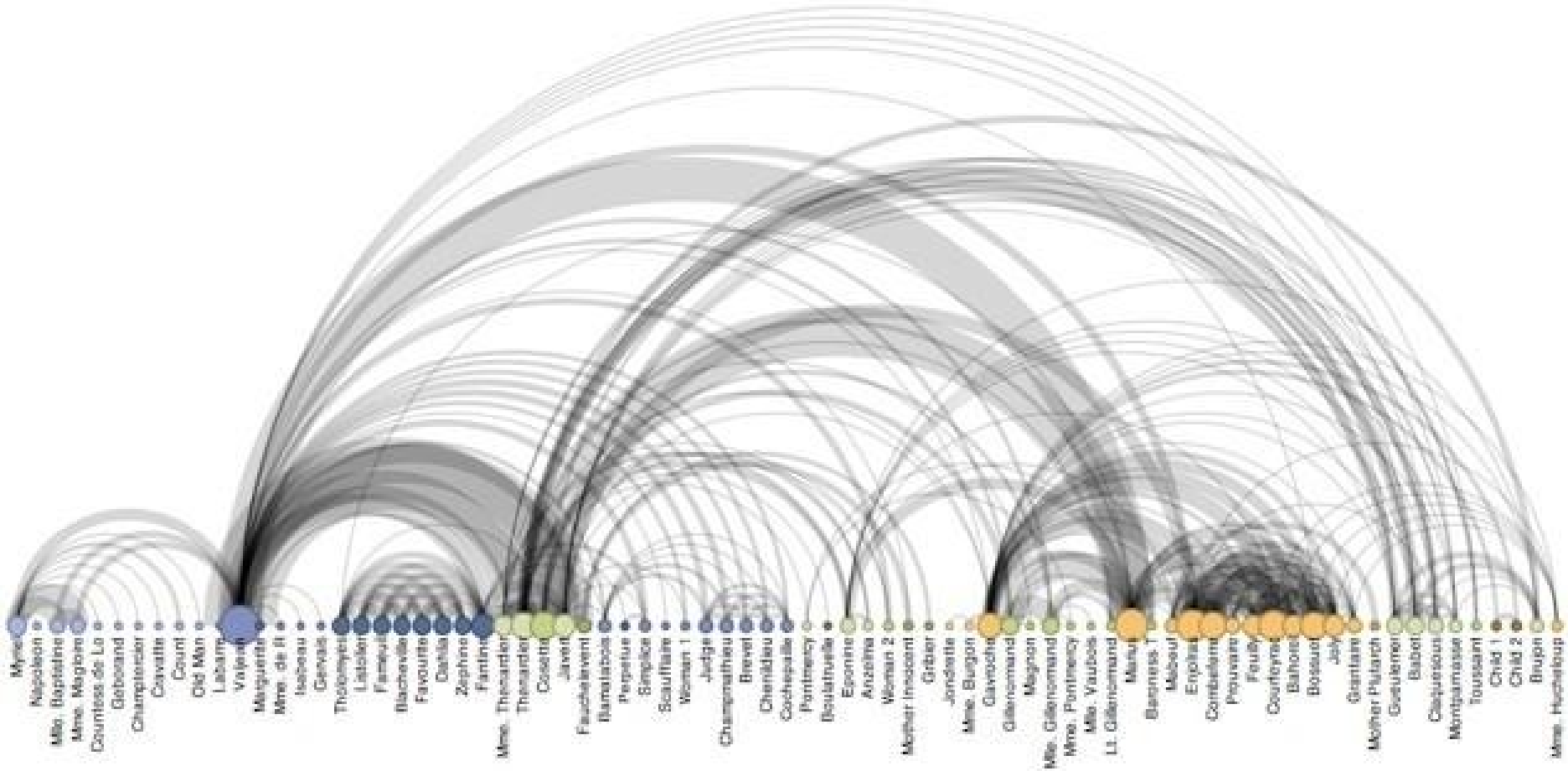
ManyEyes (http://www-958.ibm.com/software/data/cognos/manyeyes/) via "Interactive Data Visualisation", M. Ward, G Grinstein & D Keim, AK Peters, 2010; Daniel A. Keim, Christian Panse, and Stephen C. North. "Medial-Axis-Based Cartograms." IEEE Comput. Graph. Appl. 25:3 (2005), 60–68.

ENTITY RELATIONSHIPS

Arc Diagrams

- Arc diagrams are an interesting network layout approach that can be adapted to text
- Lay out all entities on a single line and draw arcs between co-occurring entities
- Strengths can be shown through colour intensity
- Groups can be shown with bands

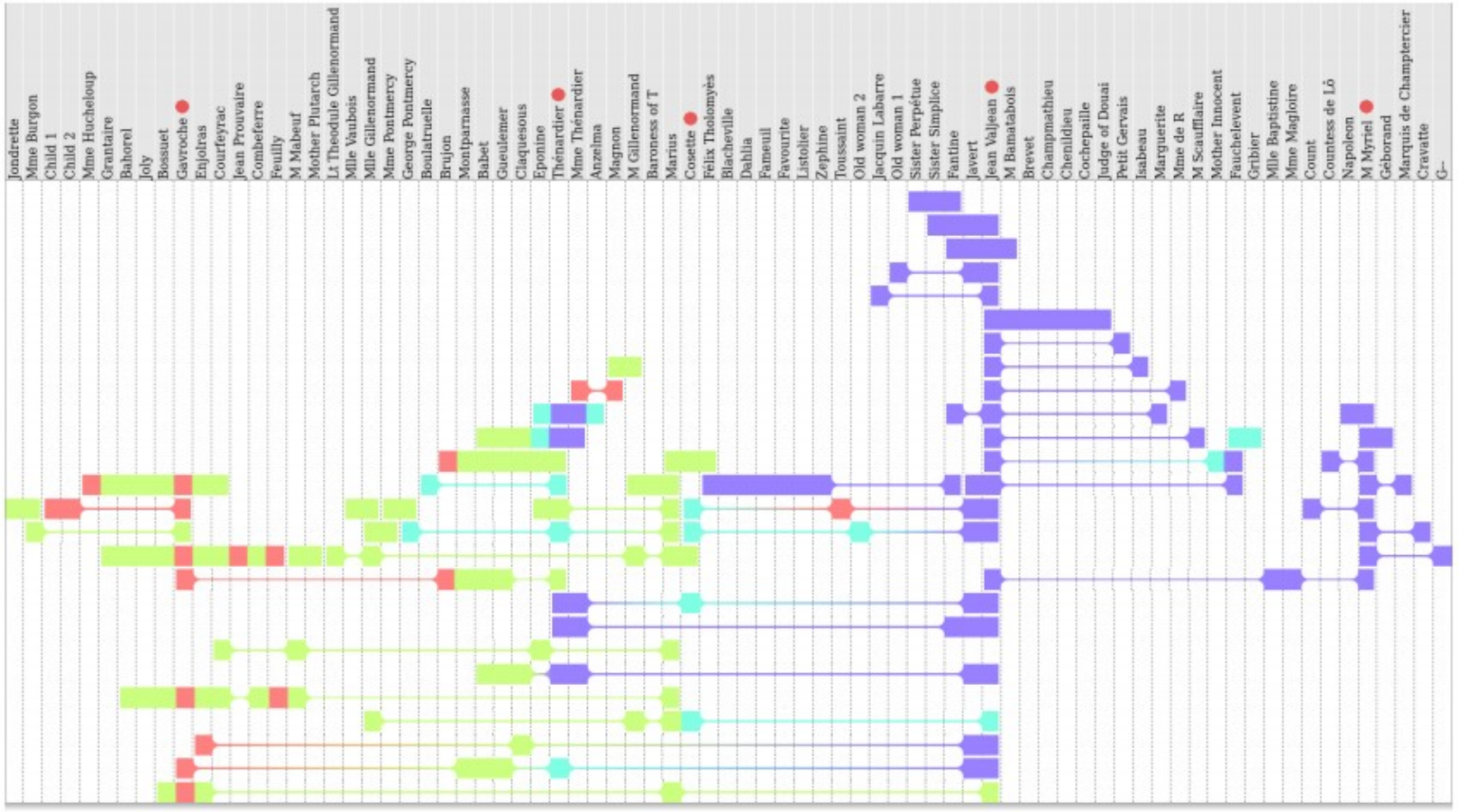
Arc Diagrams



Co-occurrence of characters in Les Miserables

Source: Knuth, D. E. 1993. The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley

Rainbow Boxes



Co-occurrence of characters in Les Miserables

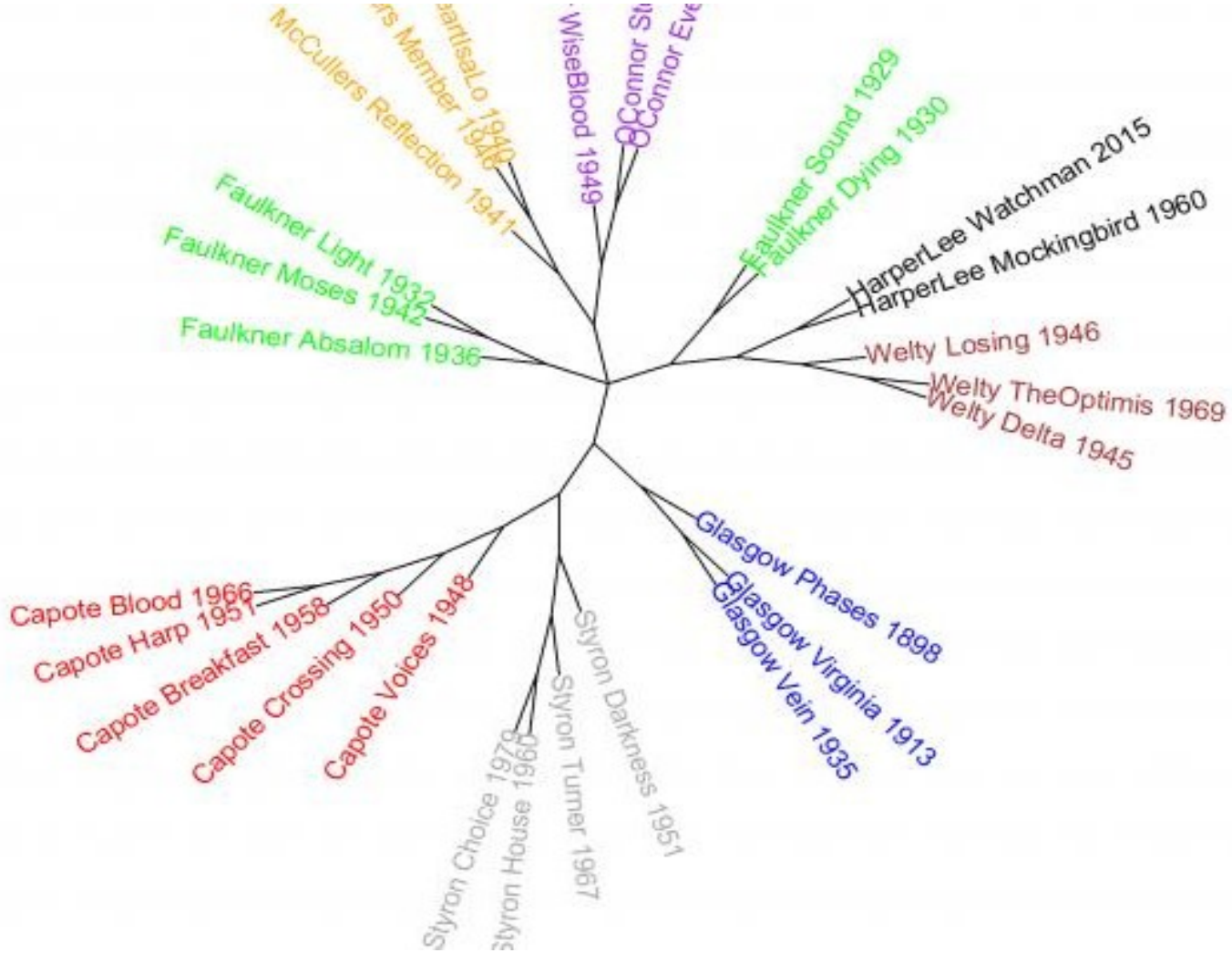
Jean-Baptiste Lamy. Visualizing symmetric square matrices with rainbow boxes: methods and application to character co-occurrence matrices in literary texts. International Conference Information Visualisation (iV), Jul 2018, Salerno, Italy.

MULTIPLE DOCUMENT VISUALISATIONS

Multiple Document Visualisations

- Visualising document collections is not particularly well standardised
- We can split approaches into two groups:
  - Visualising connections between documents
  - Visualising change in a document
- Wide open for Exploration!!

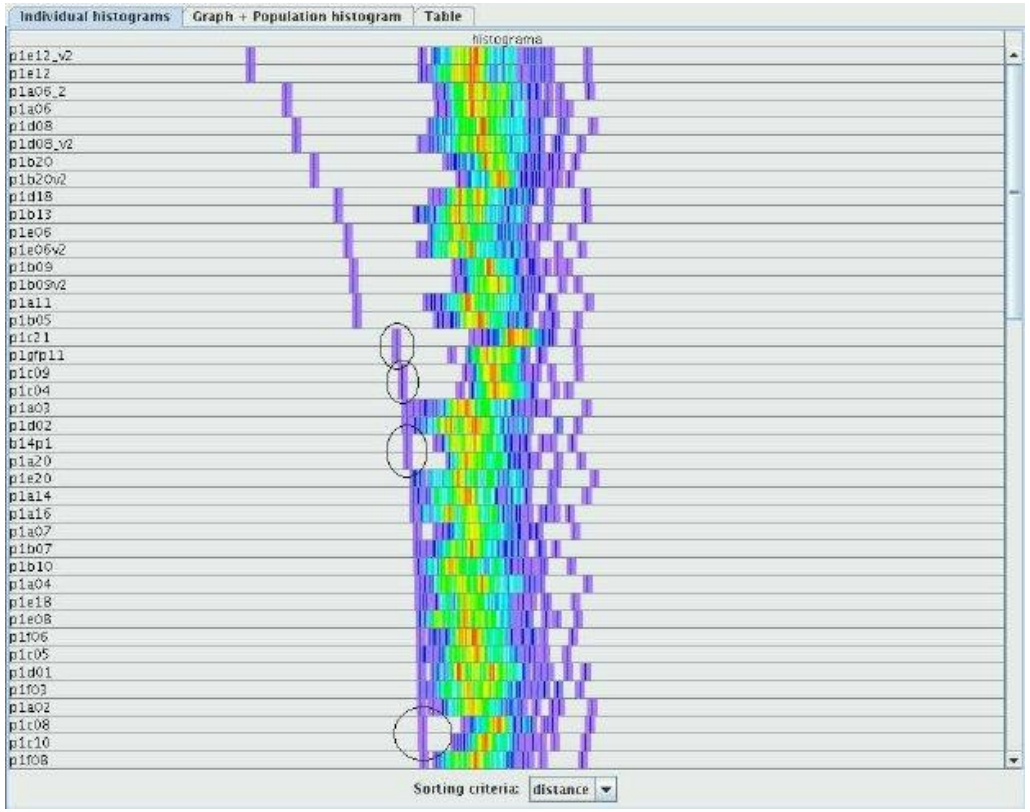
Connections - Dendrograms



https://cran.r-project.org/web/packages/stylo/index.html

Connections - Plagiarism Detection

- Visualising distances between documents



Freire, Manuel & Cebr, Manuel. (2018). Design of the AC Academic Plagiarism Detection System.

Visualising Connections Between Documents

- Collections of documents can be seen as a network
- All network visualization approaches can also be applied to collections of documents
- The real challenge is finding the most appropriate document representation for the task
- This, however, is a NLP challenge not a visualisation one

Visualising Changing Documents

- Visualising changes within documents typically involves monitoring changes in, for example:
  - Topic
  - Style
  - Author

and visualizing these using trend visualisation techniques

Changing - Style change detection

- A simple, yet challenging question to answer for style change detection is as follows:
  - Given a document, is it written by a single author or by multiple authors?

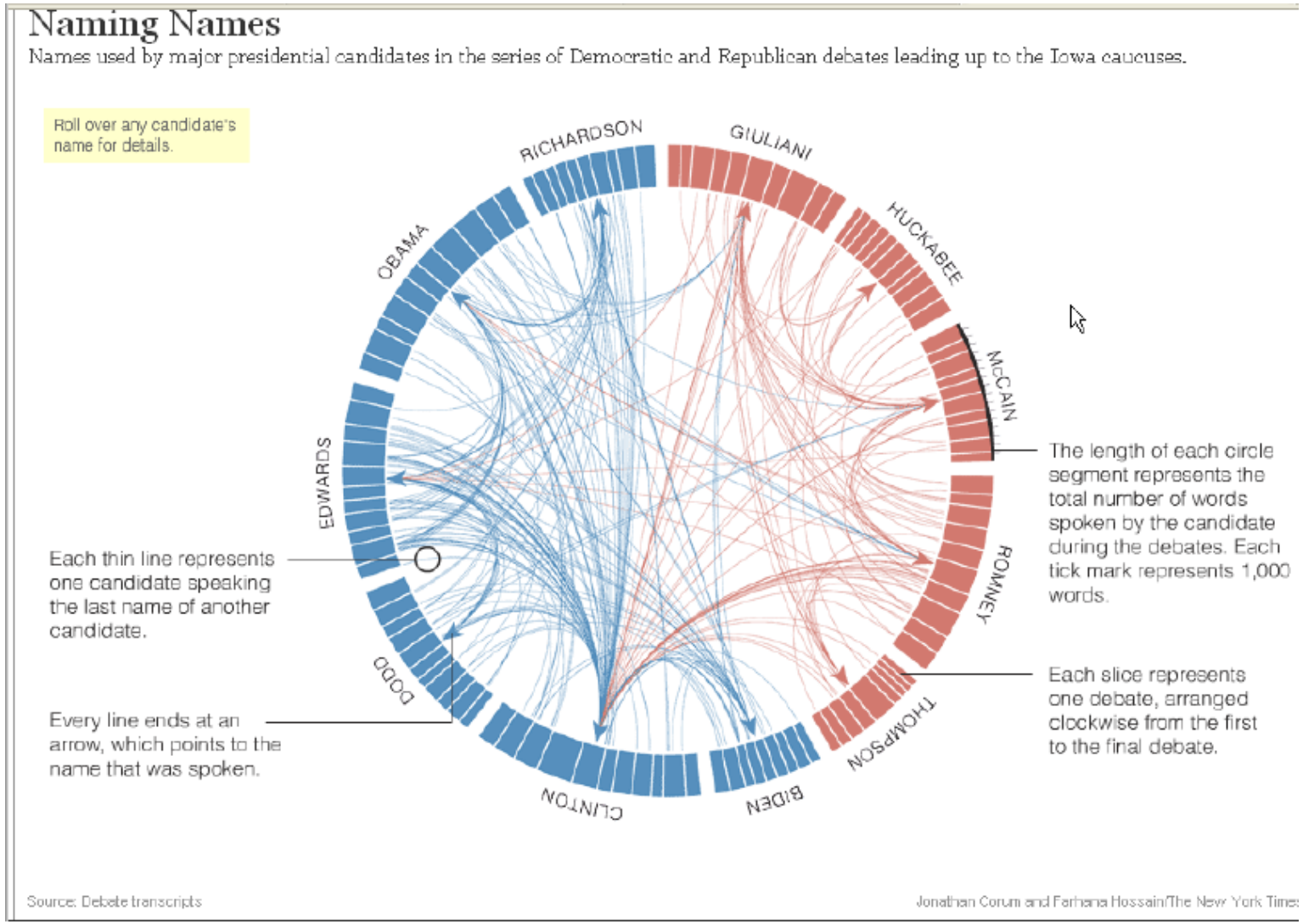
- To be able to provide an answer, the document has to be intrinsically analysed, i.e., changes of authorship have to be determined by capturing changes of writing styles

Visualising Conversations

- Many dimensions to consider:
  - Who (senders, receivers)
  - What (the content of communication)
  - When (temporal patterns)

- Interesting cross-products:
  - What x When -> Topic “Zeitgeist”
  - Who x Who -> Social network
  - Who x Who x What x When -> Information flow

Visualising Conversations



CONCLUSIONS

Conclusions

- Visualizing text seems like a sensible idea, but can be quite tricky.
- The main problem working with texts is document representation and feature extraction.
- Visualisations of texts include wordles, frequency xy scatter/histograms, word trees, arc diagrams, dendrograms amongst others.

Conclusions

- The field of text analytics and natural language processing is a very interesting open research area with plenty of applications
  - Forensic linguistics
  - Plagiarism
  - Genre/topic/Author studies
  - Style changes (authorship, health of author, dating)

- Visualisation of text can offer an extra insight into text analytics

This Week's Lab

- Exploratory Analysis of data
- Visualisation as a tool for exploring Bias and Fairness in Machine Learning

Assignment 1 30%

Specification

You have been hired as a visualisation designer to design an effective dashboard providing insights into a dataset. As part of the visualisation process you will first explore the data and produce a dashboard useful for exploration, then you will set your editorial thinking and produce a dashboard with at least 3 insights from the data.

Marking scheme

1. Select a Dataset – 2%
2. Decide on an audience (user story) – 3%
3. Using Tableau Public, design a Dashboard that allows the exploration of the data - 8%
4. Using Tableau Public, design a Dashboard that shows at least three insights from the data - 12%
5. Show evidence of previous iterations or alternatives - 5%

Assignment 1 30%

Sample sources of data

- https://toolbox.google.com/datasetsearch
- https://archive.ics.uci.edu/ml/index.php
- https://data.gov.ie/
- https://public.tableau.com/en-us/s/resources
- Make Over Monday challenges

Thanks To

- Marisa Llorens-Salvador, John McAuley, Colman McMahon and Brian Mac Namee for an earlier version of these lecture notes