Processing and exploring data

Björn Ross

Tod Van Gunten

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

Challenge of The Course

- Every Week:
  Understanding complex social phenomena using big data
- This Week:
  Where do we get ‘big’ social data?
  What does it look like?
  How can we visualise it?

Overview

- Data collection
- Data sources
- File formats and data types
- Descriptive statistics
- Visualisations

Basic Research Framework

- Research Question & Hypothesis
  - RQ: What makes persuasion effective?
  - Hypothesis: Evidence makes persuasion effective
- Data Collection
  - Download Posts & Replies from Change My View Subreddit
- Sample Population
  - Redditors in Change My View as proxy for “people engaging in persuasion online”
- Methods & Analysis
  - Identify “Evidence” in replies (e.g., search for “http” links)
- Measure & Report Outcome
  - Persuasion rate of replies with evidence vs. without evidence

Data Collection

- Asking
  - Surveys
  - Interviews
- Observing
  - Experiments
  - Big Data Analysis

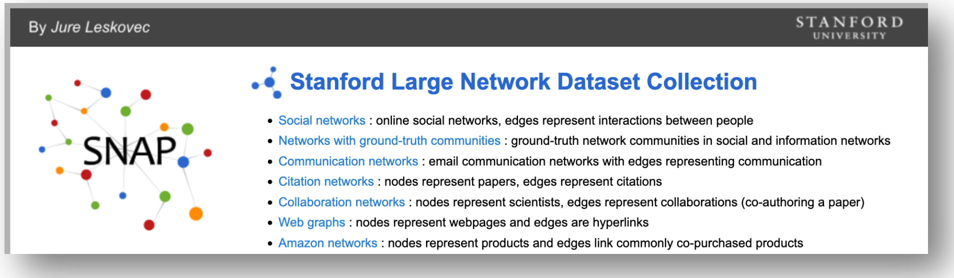
Data sources

Where do we get ‘big’ social data?

- APIs
- Scraping
- Existing datasets
  - Github, OSF, Kaggle, government open data portals

Data sources: Existing data sets

- Most data sets available online were collected **ad hoc** for a specific purpose – useful for reproducing research, not so much for planning new research
- Some datasets can be used to answer a wider set of RQs
  - Historical example: MyPersonality data https://sites.google.com/michalkosinski.com/mypersonality
  - Example: Stanford Large Network Dataset Collection https://snap.stanford.edu/data/


By *Jure Leskovec*

STANFORD UNIVERSITY

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products

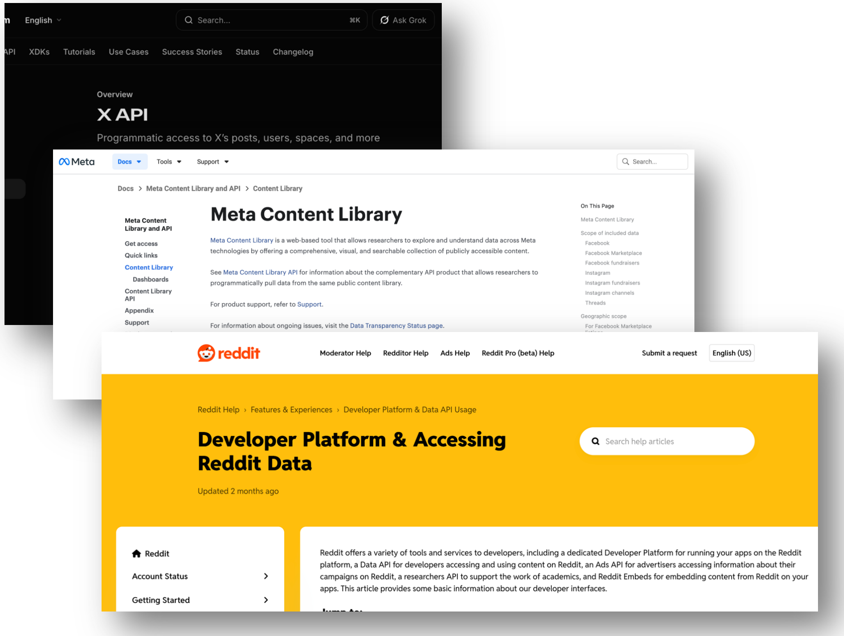
Data sources: APIs

API = application programming interface

Often the first choice for collecting “new” data

APIs differ in

- Pricing (free / paid)
- Rate limiting (e.g. number of requests per minute)
- Scope of data available (comprehensive / limited)
- Exclusivity (access for everyone / selected applicants only)
- Documentation and support
- Reliability
- …



Overview

X API

Programmatic access to X’s posts, users, spaces, and more

Meta Content Library

Meta Content Library is a web-based tool that allows researchers to explore and understand data across Meta technologies by offering a comprehensive, visual, and searchable collection of publicly accessible content.

See Meta Content Library API for information about the complementary API product that allows researchers to programmatically pull data from the same public content library.

For product support, refer to Support.

For information about ongoing issues, visit the Data Transparency Status page.

Developer Platform & Accessing Reddit Data

Updated 2 months ago

Reddit offers a variety of tools and services to developers, including a dedicated Developer Platform for running your apps on the Reddit platform, a Data API for developers accessing and using content on Reddit, an Ads API for advertisers accessing information about their campaigns on Reddit, a researchers API to support the work of academics, and Reddit Embeds for embedding content from Reddit on your apps. This article provides some basic information about our developer interfaces.

Data sources: Scraping

Often a "last resort" for obtaining data, but issues with

- Reliability
- Technical barriers
- Ethical considerations; 'politeness' (e.g. robots.txt)
- Data quality and consistency
- Possibly legal concerns (terms of service, privacy regulations e.g. GDPR, copyright, ..)

-> **Do not scrape without** explicit **permission** from the website owner!

File formats

- CSV
- JSON
- XML
- HTML

CSV (Comma-Separated Values)

```
id,neighbourhood,avg_rent
1,Old Town,1200
2,New Town,1350
3,Leith,900
4,Stockbridge,1100
5,Morningside,1300
6,Bruntsfield,1250
7,Marchmont,1150
```

- Tabular data: Columns and rows, like an Excel spreadsheet
- No hierarchical structure, or nested data

JSON (JavaScript Object Notation)

```
[
    {
        "id": 1,
        "neighbourhood": "Old Town",
        "averageRent": 1200
    },
    {
        "id": 2,
        "neighbourhood": "New Town",
        "averageRent": 1350
    },
    {
        "id": 3,
        "neighbourhood": "Leith",
        "averageRent": 900
    }
]
```

- Key-value pairs
- Objects {} and arrays []
- Supports hierarchical structures
- Commonly used in data storage and exchange in web programming (e.g. between web server and client)

XML (eXtensible Markup Language)

```xml
<Rents>
    <Neighbourhood id="1">
        <Name>Old Town</Name>
        <AverageRent>1200</AverageRent>
    </Neighbourhood>
    <Neighbourhood id="2">
        <Name>New Town</Name>
        <AverageRent>1350</AverageRent>
    </Neighbourhood>
    <Neighbourhood id="3">
        <Name>Leith</Name>
        <AverageRent>900</AverageRent>
    </Neighborhood>
</Rents>
```

- Tags and elements (root, parent elements, child elements)
- Hierarchical (tree structure)
- Commonly used in data storage and exchange

HTML (Hypertext Markup Language)

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Average Rent in Edinburgh</title>
    <style>
        table {
            width: 50%;
            border-collapse: collapse;
            margin: 20px auto;
        }
        th, td {
            border: 1px solid #ddd;
            padding: 8px;
            text-align: left;
        }
        th {
            background-color: #f2f2f2;
        }
    </style>
</head>
<body>
    <h1>Average Rent in Edinburgh per Neighbourhood</h1>
    <table>
        <thead>
            <tr>
                <th>ID</th>
                <th>Neighbourhood</th>
                <th>Average Rent (£)</th>
            </tr>
        </thead>
        <tbody>
            <tr>
                <td>1</td>
                <td>Old Town</td>
                <td>1200</td>
            </tr>
            <tr>
                <td>2</td>
                <td>New Town</td>
                <td>1350</td>
            </tr>
            <tr>
                <td>3</td>
                <td>Leith</td>
                <td>900</td>
            </tr>
        </tbody>
    </table>
</body>
</html>
```

Tree structure (like XML) but

- Designed for creating web pages
- Focuses on presentation of content
- Predefined tags
- Can include CSS and JavaScript
- Not designed as a format for data storage / exchange

Units of analysis

- Users
- Individual messages (e.g. Instagram posts, TikTok videos, ..)
- Interactions (e.g. friendship ties, retweets, replies, ..)
- Groups (e.g. subreddits, Facebook groups, ..)
- Geographical areas (neighbourhoods, countries, ..)

…

Choose your unit of analysis wisely!

Data cleaning

- Handling missing data (removal, imputation..)
- Handling duplicates
- Standardisation; transforming variables
- Error correction; handling outliers
- Validation

Data types

- Numerical data
  - Of a person: Age, income
  - Of a tweet: Number of retweets, ..
- Text data
  - Of a person: Name, Occupation, ..
  - Of a tweet: Text
- Network/relational data
  - Family relationships
  - Friendships
  - Follower / followee relationships
  - …

Descriptive statistics for numerical variables

$X = (1, 2, 7, 18, 23, 456, 1234)$

- Means

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i \qquad \bar{x} = \frac{1+2+7+18+23+456+1234}{2} \sim 248.71$$

- Medians

$$M = P_{50} = 23$$

1, 2, 7, 18, 23, 456, 1234
~50% of data fall below this number

- Percentiles

$$P_{15} = 2$$

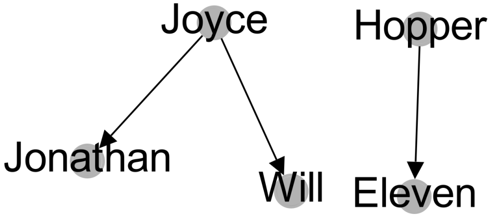
1, 2, 7, 18, 23, 456, 1234

~15% of data fall below this number

Representing network data as tabular data

Edge list

```
parent,child
Joyce,Will
Joyce,Jonathan
Hopper,Eleven
```



Adjacency matrix

```
  ,JC,WI,JN,HO,EL
JC, 0, 1, 1, 0, 0
WI, 0, 0, 0, 0, 0
JN, 0, 0, 0, 0, 0
HO, 0, 0, 0, 0, 1
EL, 0, 0, 0, 0, 0
```

- Each node is a row and a column
- “1” indicates a directed edge from row node to column node

Incidence matrix

```
JC,-1,-1, 0
WI, 1, 0, 0
JN, 0, 1, 0
HO, 0, 0,-1
EL, 0, 0, 1
```

- Each node is a row
- Each edge is a column
- “-1” for outgoing edges, “1” for incoming edges

Turning text data into numerical data

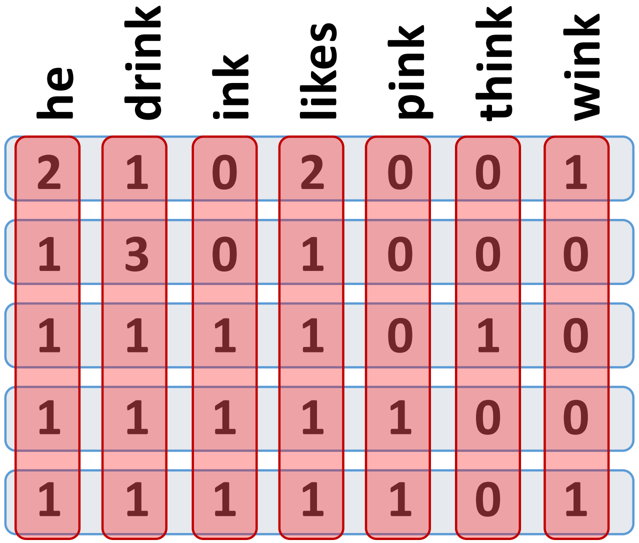
Five **documents**…

… represented as fixed-length document vectors

Turning text data into numerical data

Term vectors represent **words** as fixed-length vectors:

- Sparse (most values are 0)
- Incidentally capture semantics (similar vectors are terms that appear together)



he	drink	ink	likes	pink	think	wink	
2	1	0	2	0	0	1	← **D1:** He likes to wink, he likes to drink
1	3	0	1	0	0	0	← **D2:** He likes to drink, and drink, and drink
1	1	1	1	0	1	0	← **D3:** The thing he likes to drink is ink
1	1	1	1	1	0	0	← **D4:** The ink he likes to drink is pink
1	1	1	1	1	0	1	← **D5:** He likes to wink, and drink pink ink

Word embeddings

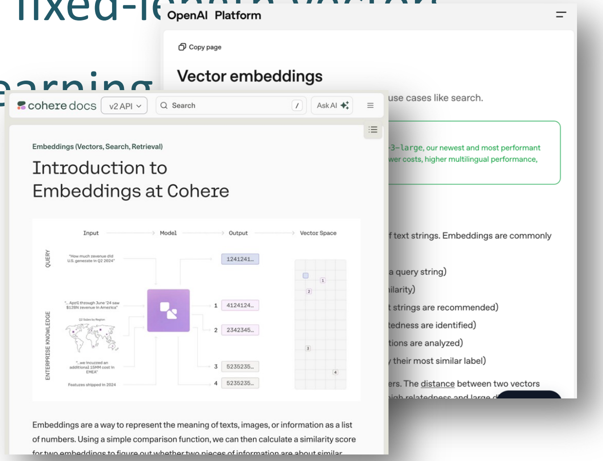
More complex representation of **words** as vectors

- Dense (all entries are non-zero)
- Capture semantics even better (similar words have similar vectors)

he	drink	ink
0.123	0.521	0.313
0.451	0.987	0.812
0.938	0.141	0.411
…	…	…

(many dimensions e.g. 300)

- Modern embeddings can represent entire **sentences**, **paragraphs** or **documents** as fixed-length vectors
- Obtained through machine learning collection of training data
- Pre-trained embeddings available online:


OpenAI Platform

Copy page

Vector embeddings

use cases like search.

-3-large, our newest and most performant
wer costs, higher multilingual performance,

f text strings. Embeddings are commonly

a query string)
ilarity)
t strings are recommended)
tedness are identified)
tions are analyzed)
their most similar label)

ers. The distance between two vectors

cohere docs v2 API Search / Ask AI

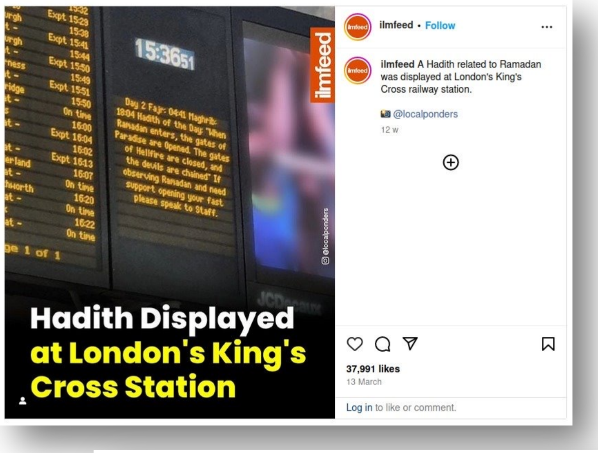
Embeddings (Vectors, Search, Retrieval)

Introduction to Embeddings at Cohere

Embeddings are a way to represent the meaning of texts, images, or information as a list of numbers. Using a simple comparison function, we can then calculate a similarity score

Case study: Religious discussions online

- Hadiths are the **recorded actions and words** of the prophet **Muhammad**
- One of the main sources of religious knowledge in Islam
- Widely used in daily life, also by politicians
- We wanted to know: **Who is using hadith on social media, how and why?**


ilmfeed
15:3651
Day 2 Fajr: 04:41 Maghrib: 18:04 Hadith of the Day: "When Ramadan enters, the gates of Paradise are Opened. The gates of Hellfire are closed, and the devils are chained" If observing Ramadan and need support opening your fast please speak to Staff.
@localponders
Hadith Displayed at London's King's Cross Station
ilmfeed • Follow
Log in to like or comment.

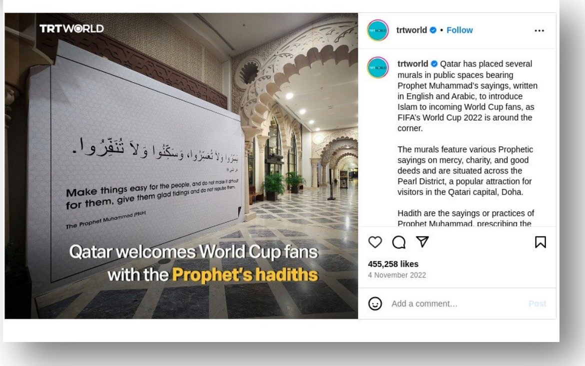
ilmfeed A Hadith related to Ramadan was displayed at London's King's Cross railway station.

📷 @localponders

12 w

37,991 likes

13 March


TRTWORLD
يَسِّرُوا وَلَا تُعَسِّرُوا، وَسَكِّنُوا وَلَا تُنَفِّرُوا.
Make things easy for the people, and do not make it difficult for them, give them glad tidings and do not repulse them.
The Prophet Muhammad (PBUH)
Qatar welcomes World Cup fans with the Prophet's hadiths
trtworld • Follow
Add a comment...
Post

trtworld Qatar has placed several murals in public spaces bearing Prophet Muhammad's sayings, written in English and Arabic, to introduce Islam to incoming World Cup fans, as FIFA's World Cup 2022 is around the corner.

The murals feature various Prophetic sayings on mercy, charity, and good deeds and are situated across the Pearl District, a popular attraction for visitors in the Qatari capital, Doha.

Hadith are the sayings or practices of Prophet Muhammad, prescribing the

455,258 likes

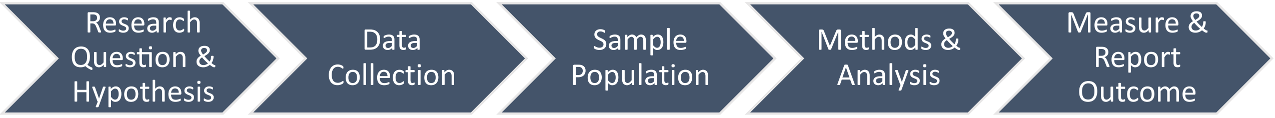
4 November 2022

Mahmoud Fawzi, Walid Magdy, and Björn Ross. 2025. "The Prophet said so!": On Exploring Hadith Presence on Arabic Social Media. In *Proceedings of ACM on Human-Computer Interaction (CSCW '25)*. ACM, New York, NY, USA, Article CSCW192, 23 pages. https://doi.org/10.1145/3711090

Religious discussions online: Research framework

- Research Questions
  - Which hadiths are most frequently shared by users of Arabic social media?
  - What topics are they about? Are they authentic or fabricated? When do they share them?
- Data Collection
  - Existing tweet dataset from archive.org (originally obtained from Twitter API)
  - Hadiths with topic categories scraped with permission from www.sonnaonline.com
  - Authenticity data from existing dataset (LK Corpus + MAHADDAT)
- Sample Population
  - Tweets in Arabic that contain specific phrase
- Methods & Analysis
  - Match tweets to hadiths (Jaccard similarity)
  - Calculate seasonality (Gini coefficient)
- Measure & Report Outcome
  - Compare topics in hadiths shared on social media with topics in all hadiths
  - Report most seasonal hadiths, authenticity distribution

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science


Research Question & Hypothesis

Data Collection

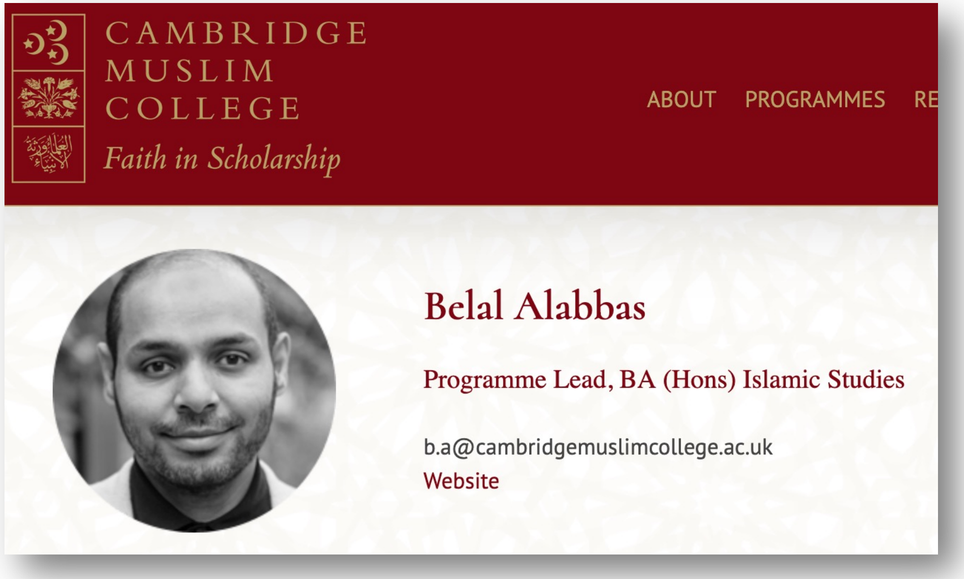
Sample Population

Methods & Analysis

Measure & Report Outcome

THE UNIVERSITY *of* EDINBURGH
informatics

Religious discussions online: Data collection



CAMBRIDGE MUSLIM COLLEGE

Faith in Scholarship

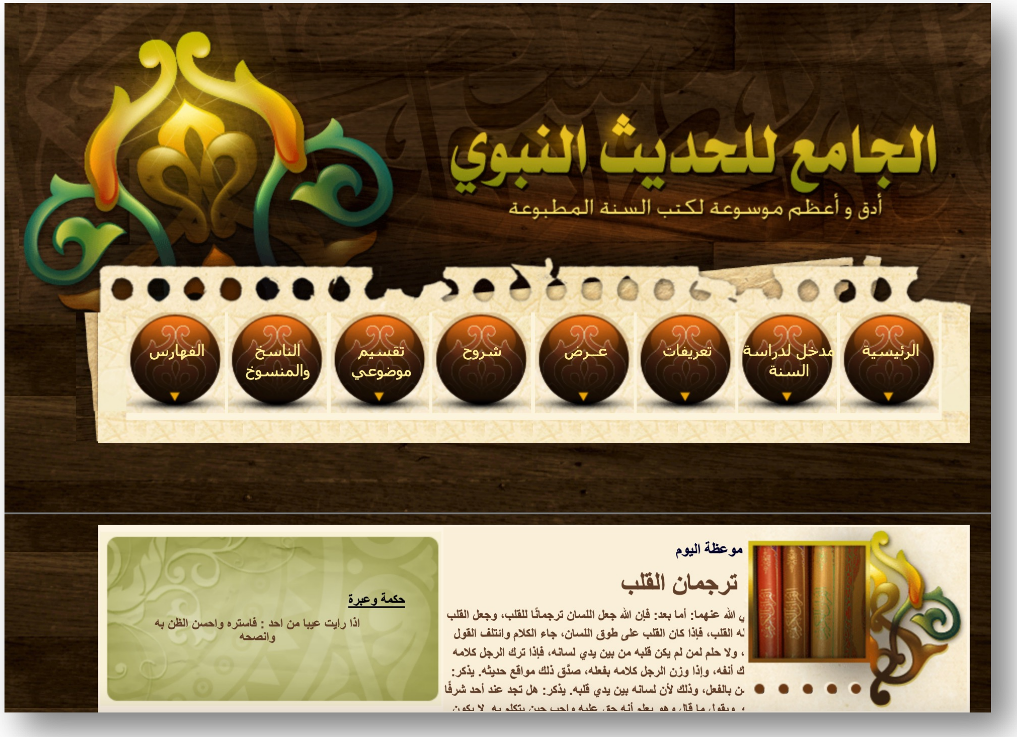
ABOUT PROGRAMMES RE

Belal Alabbas

Programme Lead, BA (Hons) Islamic Studies

b.a@cambridgemuslimcollege.ac.uk

Website



الجامع للحديث النبوي

أدق و أعظم موسوعة لكتب السنة المطبوعة

موعظة اليوم

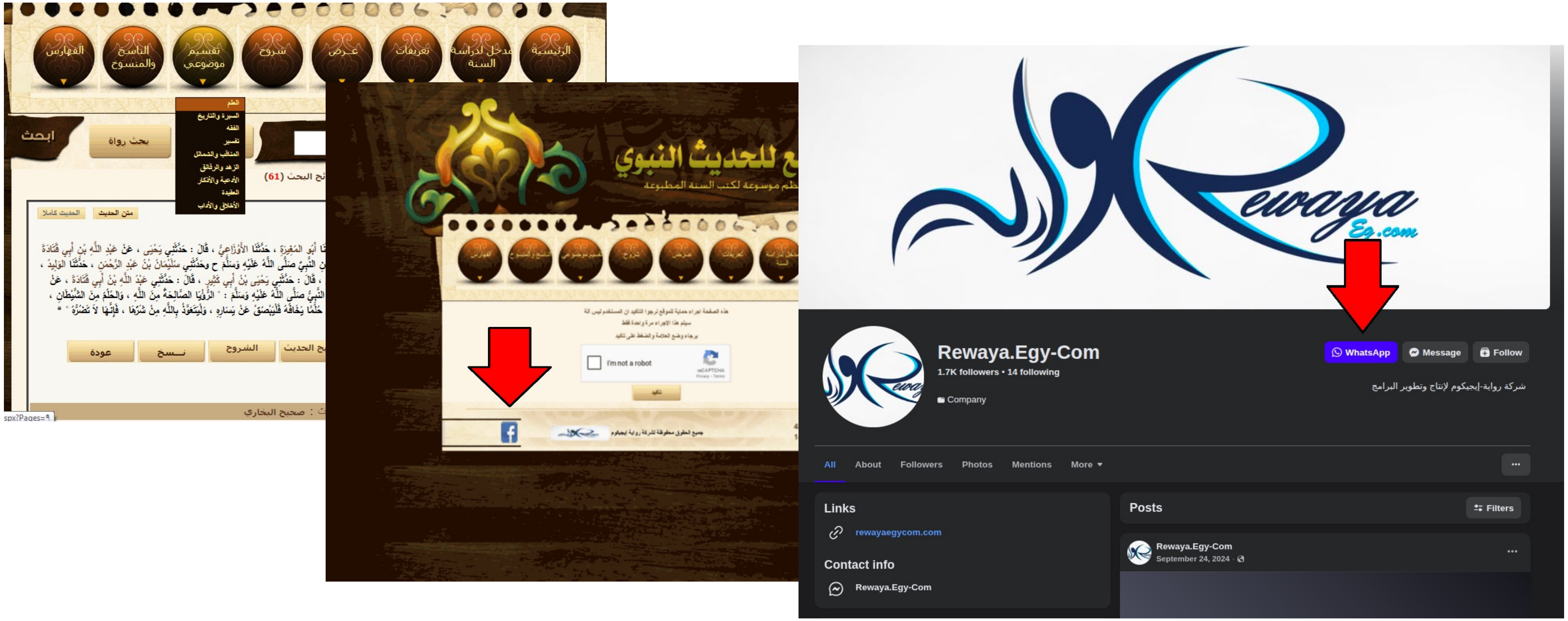
ترجمان القلب

ي الله عنهما: أما بعد: فإن الله جعل اللسان ترجمانًا للقلب، وجعل القلب
له القلب، فإذا كان القلب على طوق اللسان، جاء الكلام وائتلف القول
، ولا حلم لمن لم يكن قلبه من بين يدي لسانه، فإذا ترك الرجل كلامه
ك أنفه، وإذا وزن الرجل كلامه بفعله، صدَّق ذلك مواقع حديثه. يذكر:
ن بالفعل، وذلك لأن لسانه بين يدي قلبه. يذكر: هل تجد عند أحد شرفًا
ه ويقول ما قال وهو يعلم أنه حق عليه واجب حين يتكلم به. لا يكون

حكمة وعبرة

اذا رايت عيبا من احد : فاستره واحسن الظن به
وانصحه

Religious discussions online: Data collection



نتائج البحث (61)

نَا أَبُو الْمُغِيرَةِ ، حَدَّثَنَا الْأَوْزَاعِيُّ ، قَالَ : حَدَّثَنِي يَحْيَى ، عَنْ عَبْدِ اللَّهِ بْنِ أَبِي قَتَادَةَ
نِ النَّبِيِّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ ح وحَدَّثَنِي سُلَيْمَانُ بْنُ عَبْدِ الرَّحْمَنِ ، حَدَّثَنَا الْوَلِيدُ ،
، قَالَ : حَدَّثَنِي يَحْيَى بْنُ أَبِي كَثِيرٍ ، قَالَ : حَدَّثَنِي عَبْدُ اللَّهِ بْنُ أَبِي قَتَادَةَ ، عَنْ
النَّبِيُّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ : " الرُّؤْيَا الصَّالِحَةُ مِنَ اللَّهِ ، وَالْحُلْمُ مِنَ الشَّيْطَانِ ،
حُلْمًا يَخَافُهُ فَلْيَبْصُقْ عَنْ يَسَارِهِ ، وَلْيَتَعَوَّذْ بِاللَّهِ مِنْ شَرِّهَا ، فَإِنَّهَا لَا تَضُرُّهُ " *

: صحيح البخاري

spx?Pages=٩

Rewaya.Egy-Com

1.7K followers • 14 following

Company

شركة رواية-إيجيكوم لإنتاج وتطوير البرامج

Links

rewayaegycom.com

Contact info

Rewaya.Egy-Com

Posts

Rewaya.Egy-Com
September 24, 2024

Data visualisation

- Data visualization is often the first step in a project
  - What kind of **variation** is present in the data?
  - What are key **comparisons** between groups?
  - What kinds of **problems** can we anticipate (e.g. missing data, data not what we expected, selection effects)

Image: thenewstack.io

Why visualize?

- The table shows the trade balance between England and Denmark/Norway in the 1700s
- In what year did the trade balance begin to favour England (exports higher than imports)?

Source: https://data.europa.eu/apps/data-visualisation-guide/data-visualisation-is-accessibility

Year	Imports	Exports
1700	71.1	32.8
1705	74.5	40.9
1710	82.6	59
1715	87.2	77.9
1720	96.8	75.2
1725	102.6	71.3
1730	96.4	64.7
1735	93.7	60.5
1740	92.9	65.1
1745	92.5	74.3
1750	90.1	77.4
1755	79.9	82.8
1760	76.6	117.5
1765	79.6	151.8
1770	83.8	163.8
1775	90.4	175.7
1780	92.7	185.4

Why visualize?

- Patterns in data are often much easier to see graphically than numerically.
- Famous early data visualization by Scottish Engineer William Playfair

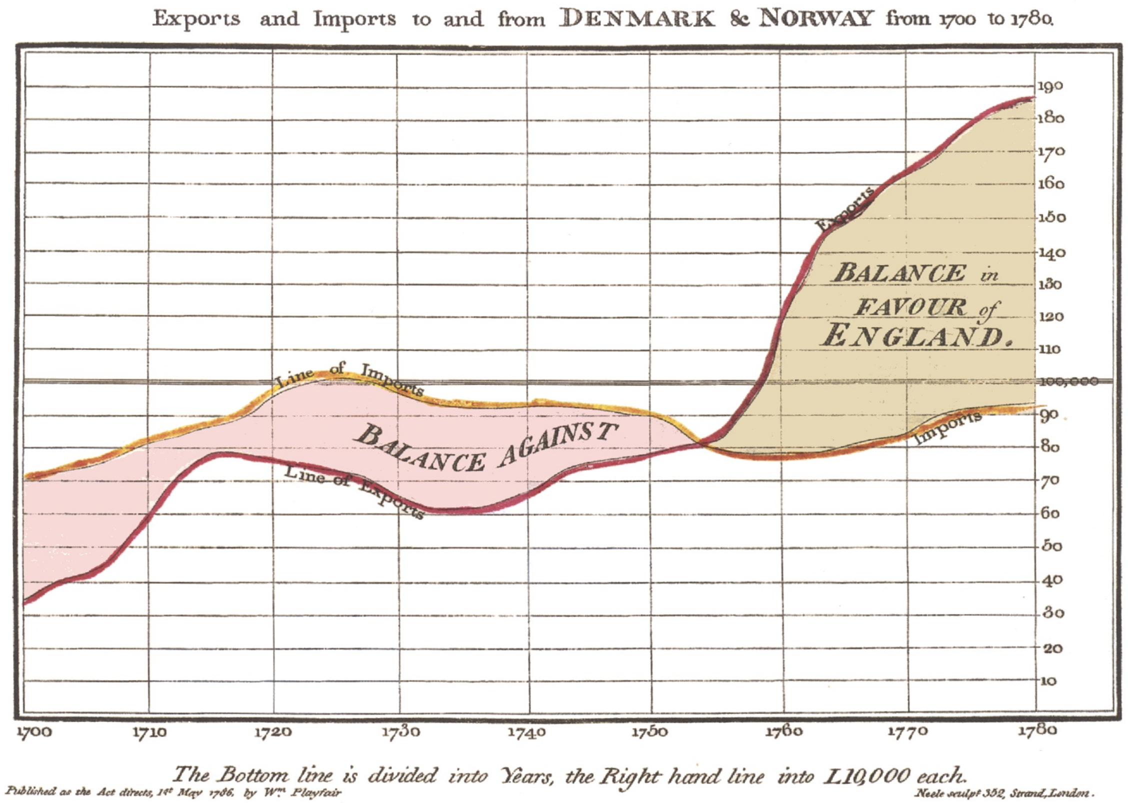
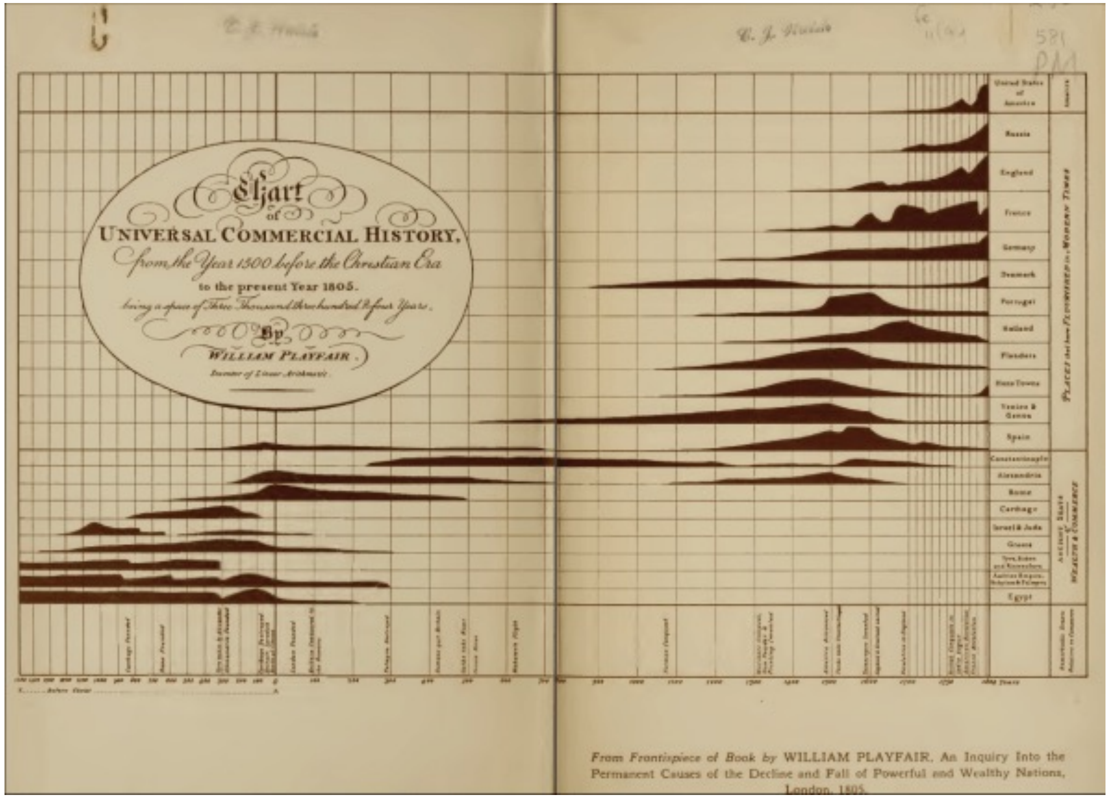


Image source: https://commons.wikimedia.org/wiki/File:Playfair_TimeSeries-2.png

- Playfair’s visualisations exemplify many aspects of modern data visualization best practice.
- Here, the idea of **small multiples**.

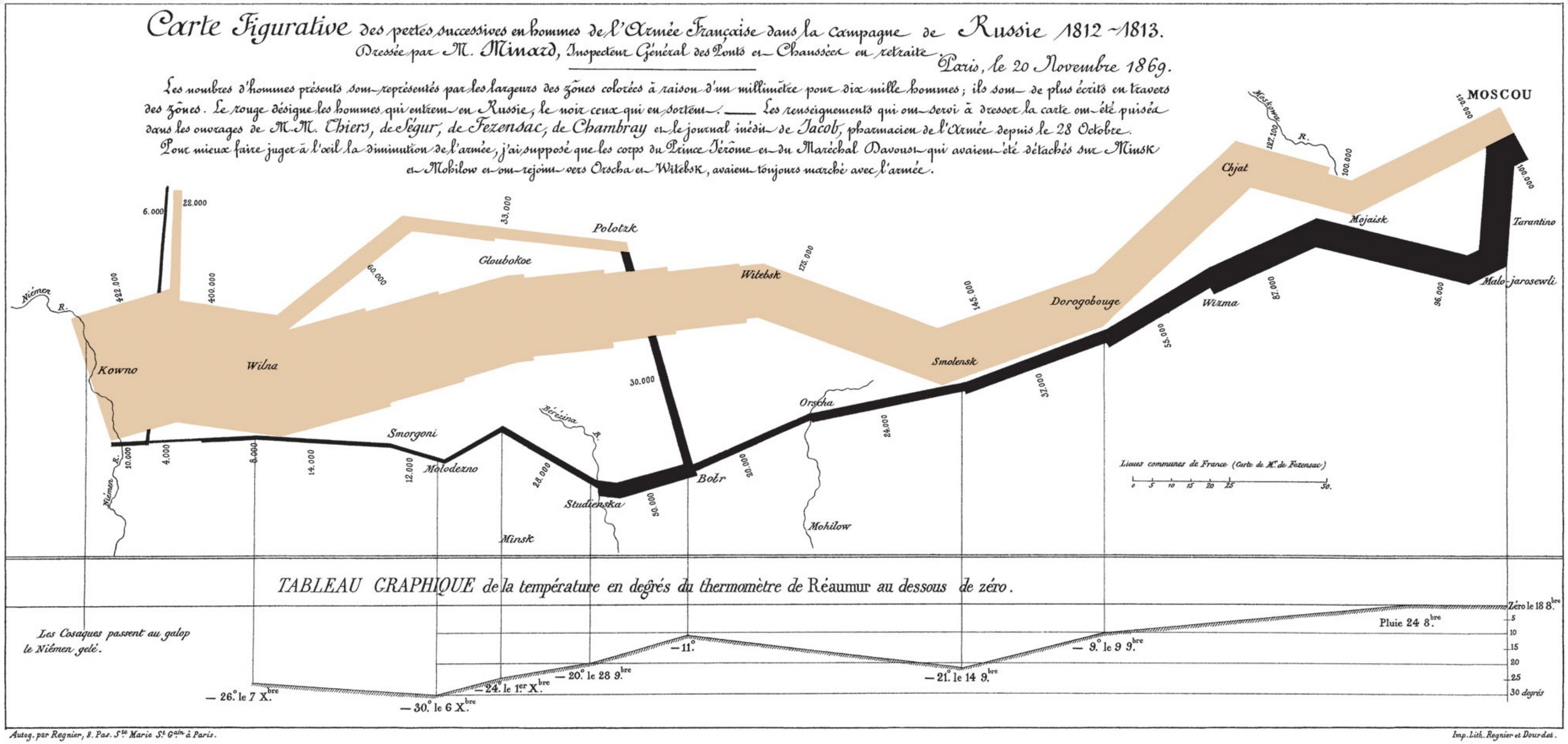


From Frontispiece of Book by WILLIAM PLAYFAIR. An Inquiry Into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations, London, 1805.

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science

THE UNIVERSITY of EDINBURGH
informatics

Napoleon’s march



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

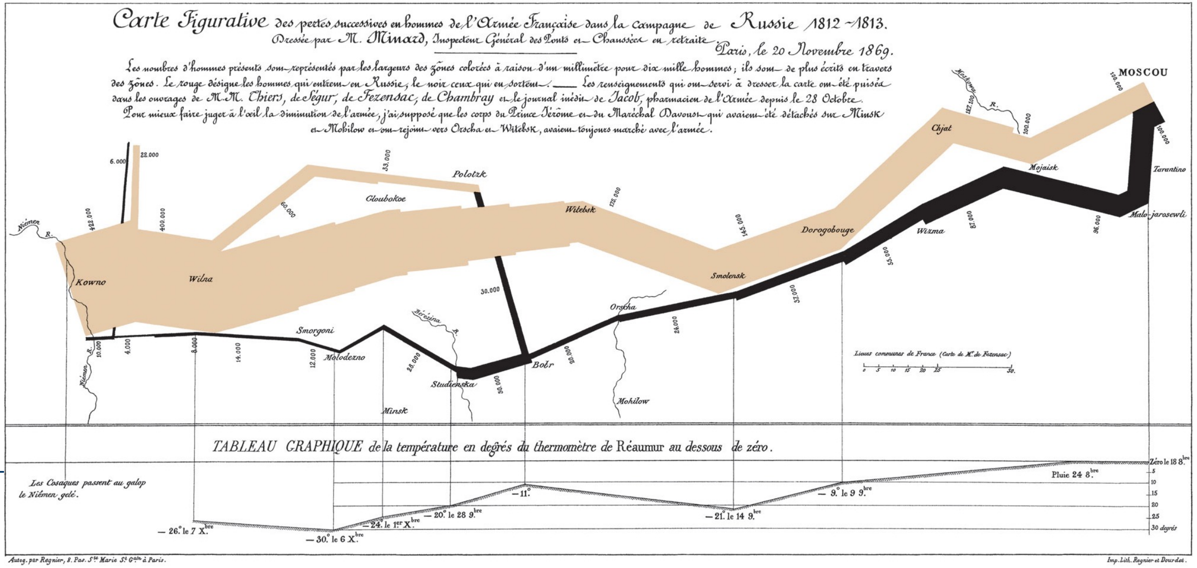
Imp. Lith. Regnier et Dourdet.

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science

THE UNIVERSITY of EDINBURGH
informatics

‘The best statistical graphic ever drawn’

- Pierre Minard’s depiction of Napoleon’s retreat shows several variables:
  - Size of the army (width)
  - Location of the army (position)
  - Direction of movement (colour)
  - Temperature (secondary line plot)



Carte Figurative des pertes successives en hommes de l’Armée Française dans la campagne de Russie 1812 – 1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

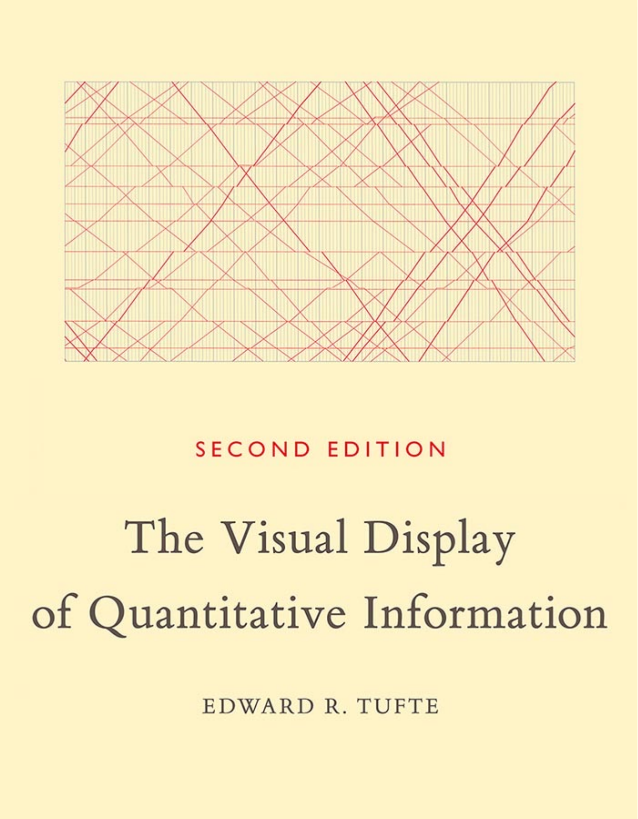
Les nombres d’hommes présents sont représentés par les largeurs des zônes colorées à raison d’un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l’Armée depuis le 28 Octobre.
Pour mieux faire juger à l’oeil la diminution de l’armée, j’ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l’armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

The science of data visualisation

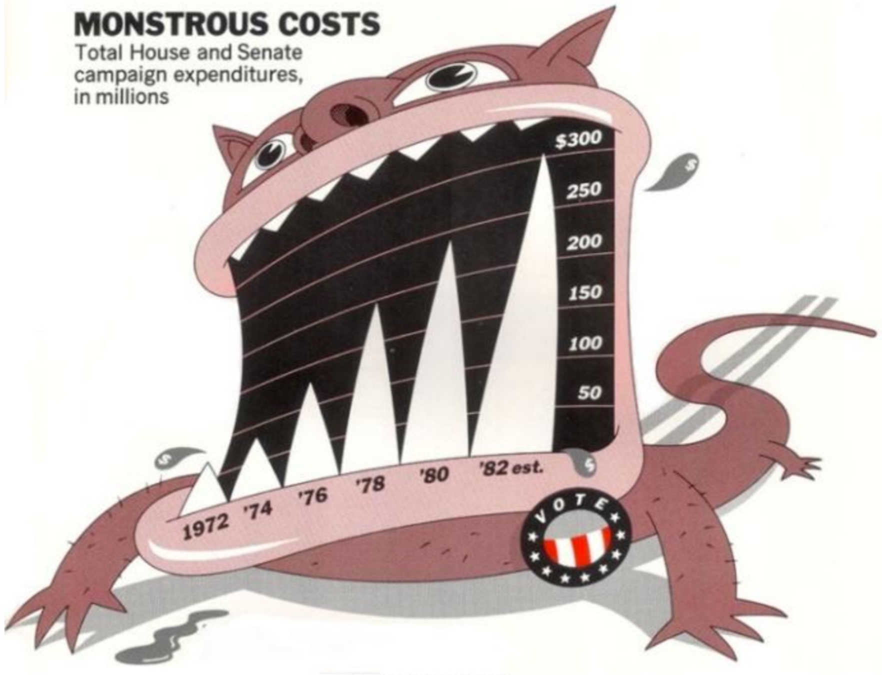
- Data visualisation can look pretty, but it is not just aesthetic
- There are principles of clear, effective, honest visualization
- Data-ink ratio:
  - Show the data
  - Minimise not-data
  - Minimise redundancy



Failure to minimize not-data

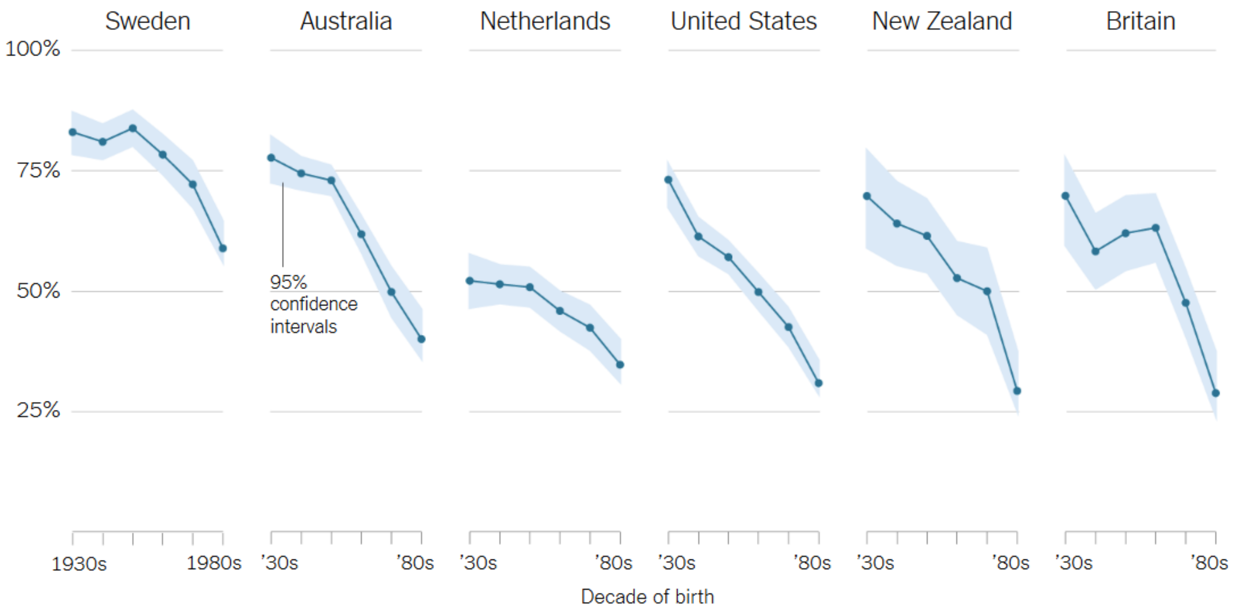
- Tufte argues against gimmicky graphics with a low data-ink ratio ('chartjunk')
- How else could we represent these data?

Example: Healy, Data Visualisation



What does this graph show?

Percentage of people who say it is "essential" to live in a democracy



Source: Yascha Mounk and Roberto Stefan Foa, "The Signs of Democratic Deconsolidation," Journal of Democracy | By The New York Times

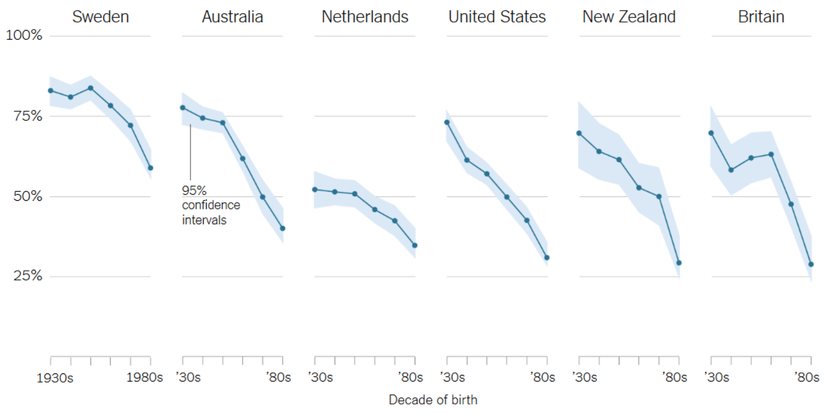
This and the following draw on Healy, Data Visualization

Always read the fine print

- The figure shows the proportion of people who answered “10” to the question: “How important is it to you to live in a country that is governed democratically”
- The horizontal (x) axis shows year of birth, not year of survey
  - Differences between age cohorts, not change over time

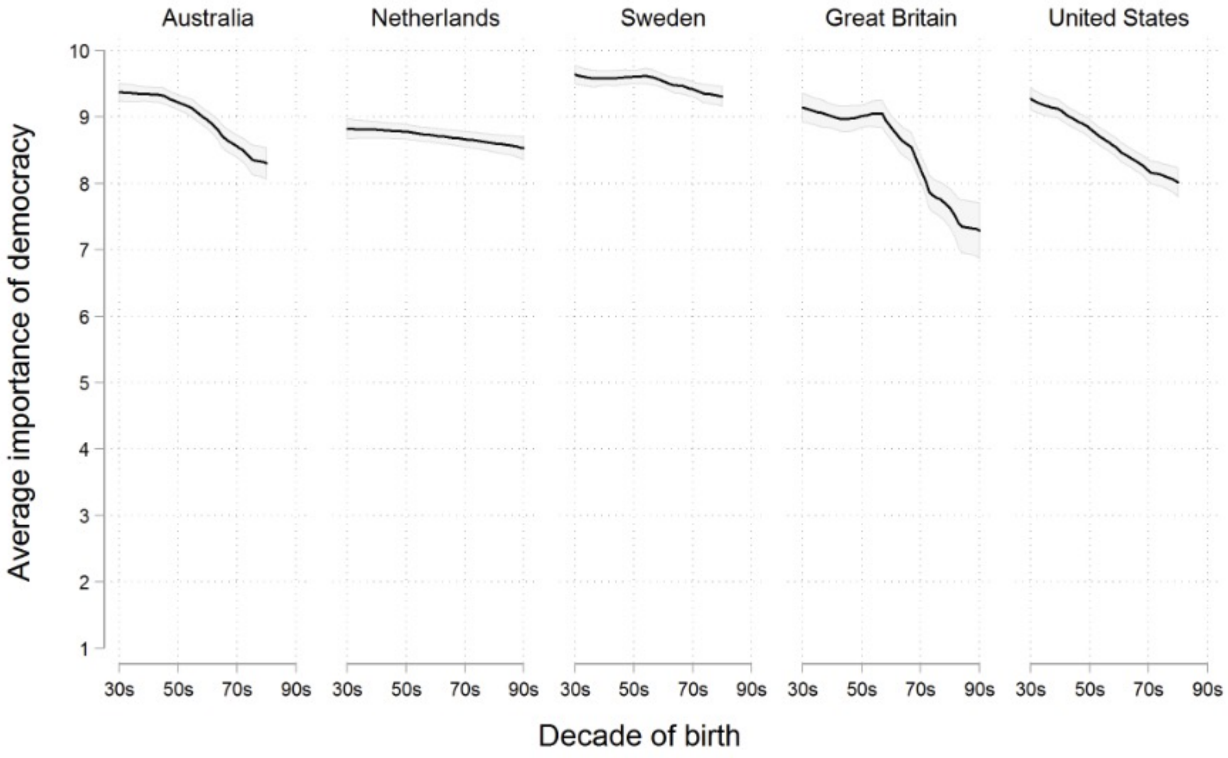
Example from Healey, Data Visualisation

Percentage of people who say it is “essential” to live in a democracy



Source: Yascha Mounk and Roberto Stefan Foa, “The Signs of Democratic Deconsolidation,” Journal of Democracy | By The New York Times

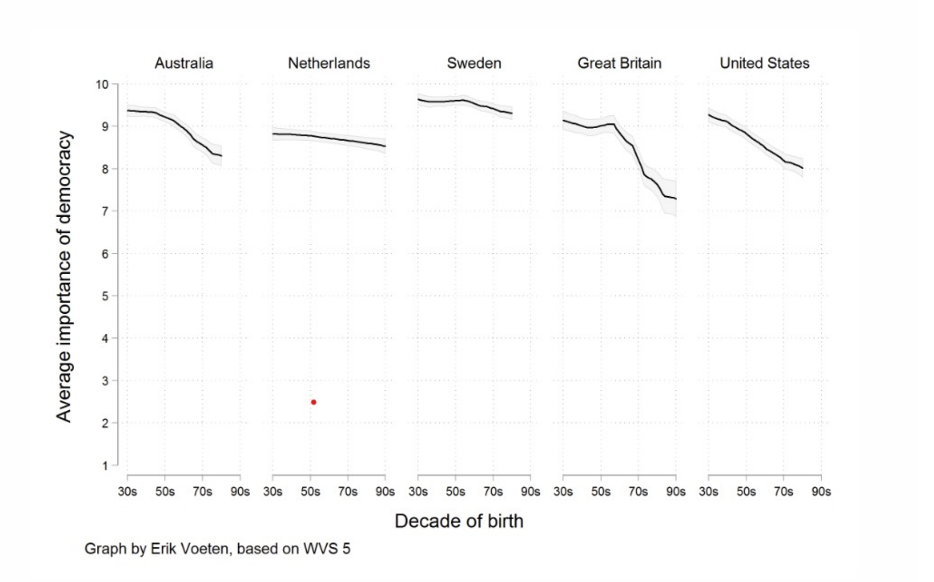
How NOT to lie with graphics



Graph by Erik Voeten, based on WVS 5

Visualisation choices matter

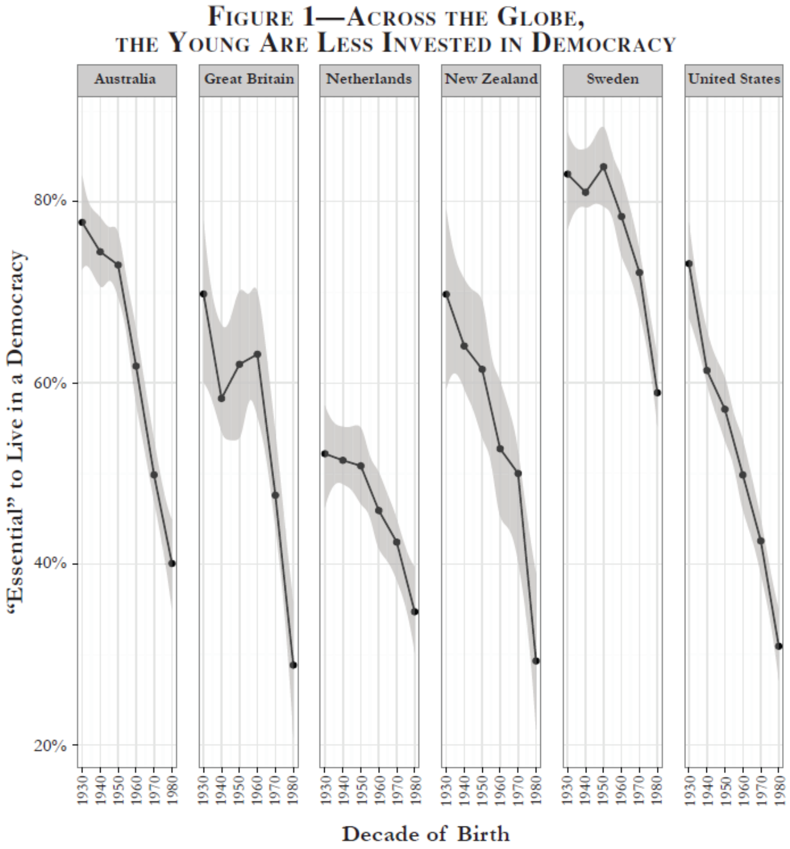
- The alternative plot shows the same data, but plots the average value of the 10-point scale (rather than the proportion of 10s)
- Suggests decline of importance of democracy, but not as much.
- Respondents still overwhelmingly say it is important to live in a democracy.



Graph by Erik Voeten, based on WVS 5

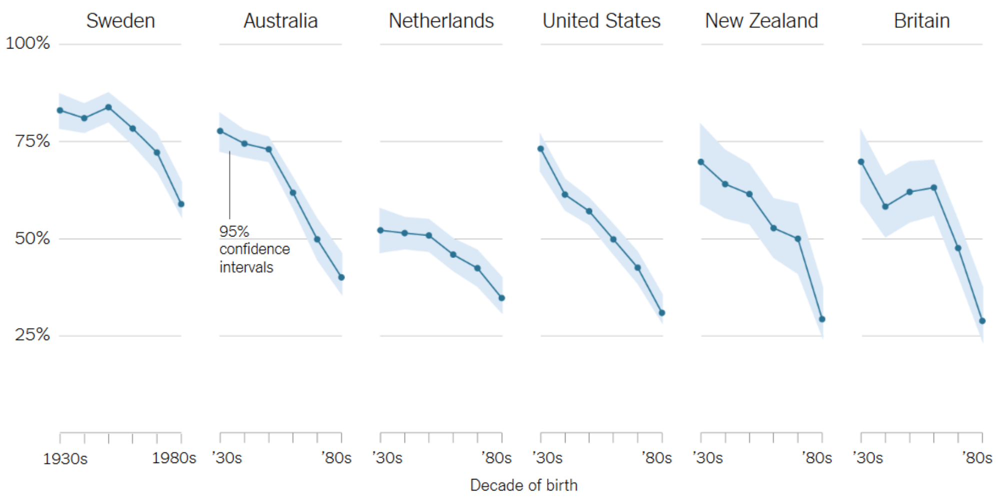
Coordinate space matters

- The figure in the originally published article had another problem: a compressed **aspect ratio** accentuated the impression of decline
- Too much vertical space, not enough horizontal space



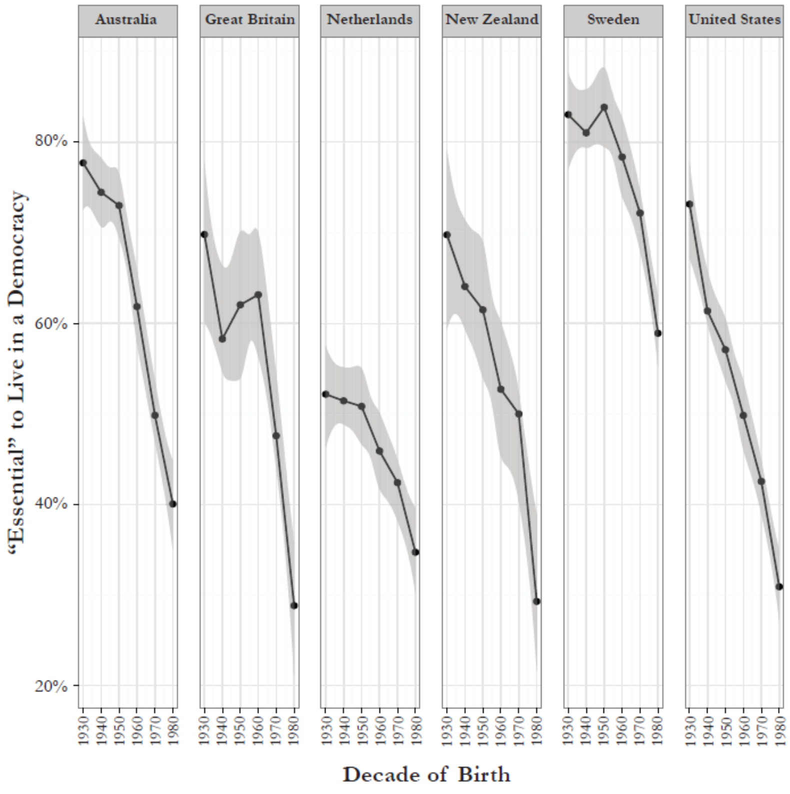
Two versions of the same plot

Percentage of people who say it is “essential” to live in a democracy



Source: Yascha Mounk and Roberto Stefan Foa, “The Signs of Democratic Deconsolidation,” Journal of Democracy | By The New York Times

Figure 1—Across the Globe, the Young Are Less Invested in Democracy

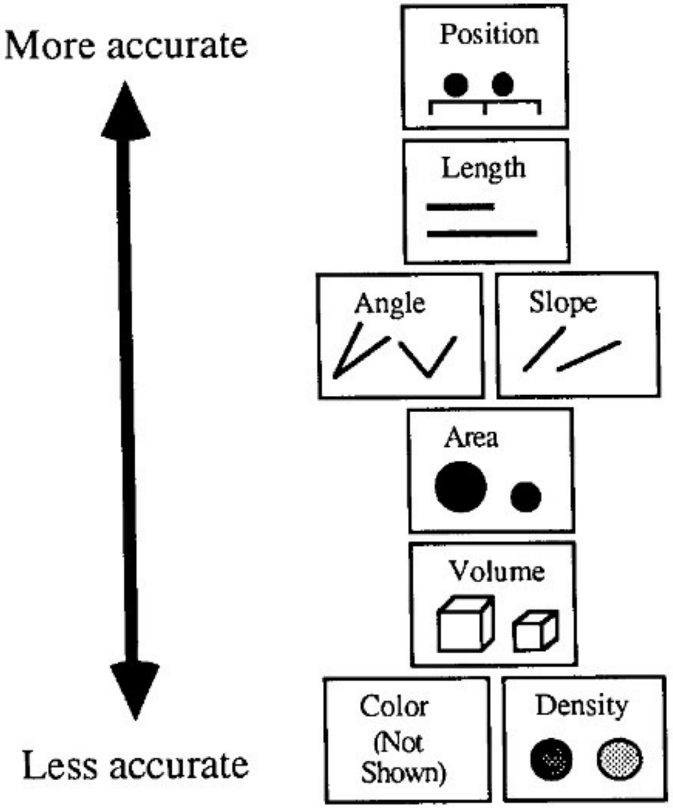


Perceptual accuracy

Most to least accurate:

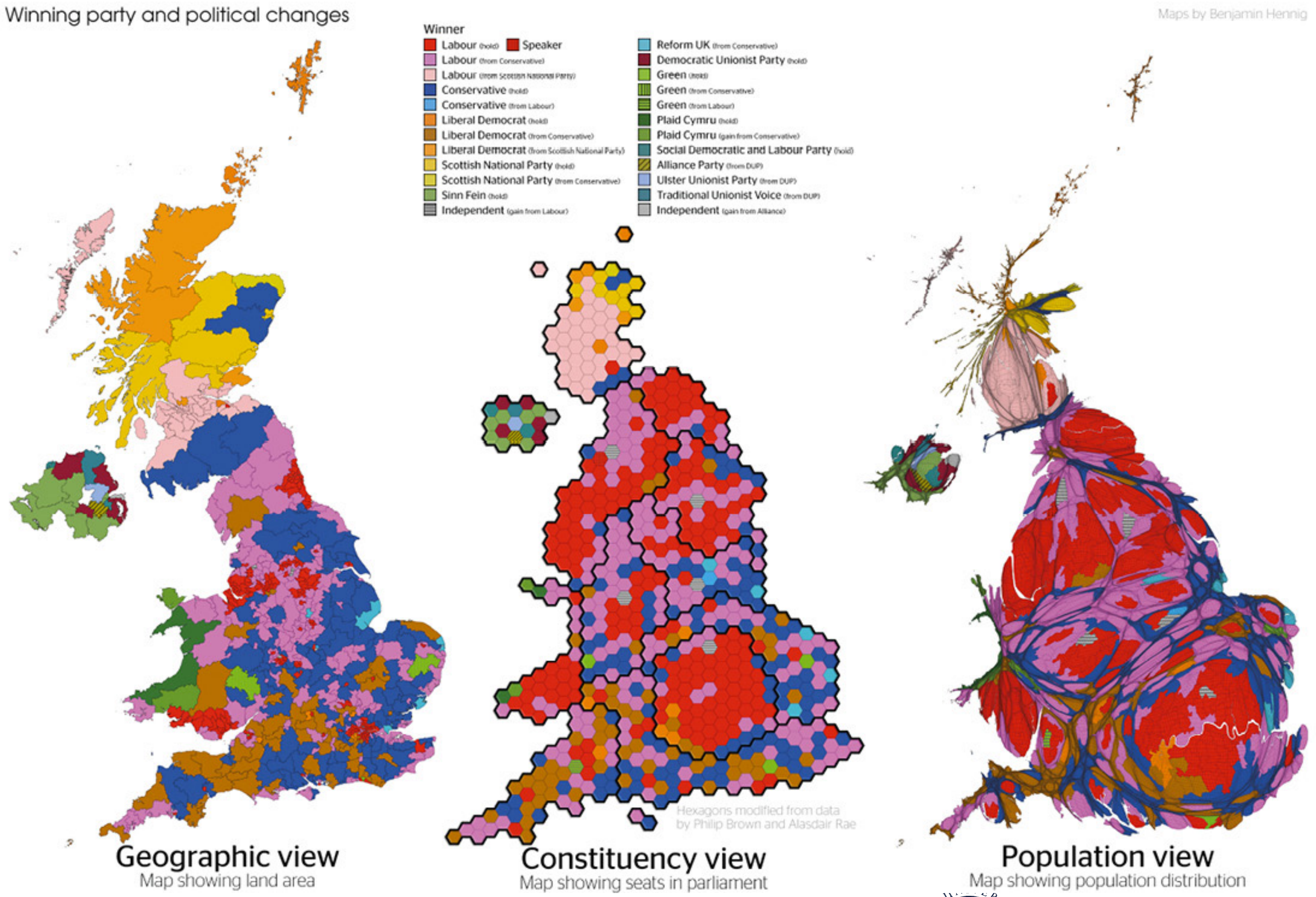
- Position on a common scale
- Length
- Direction
- Angle and slope
- Area
- Volume
- Density, color

Source: Mackinlay (1986); ucdavisdatalab.github.io



Using colour

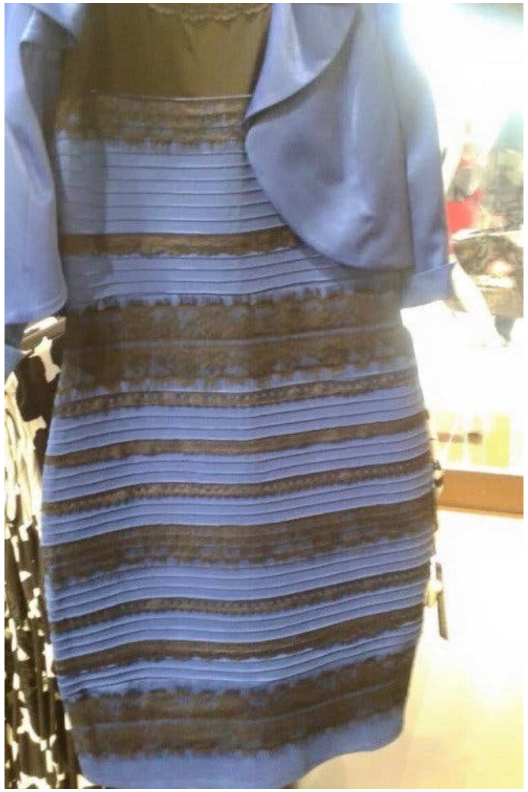
- Colour can convey key information, but is also hard to perceive clearly.


Winning party and political changes
Maps by Benjamin Hennig
Winner
Labour (hold)
Speaker
Labour (from Conservative)
Labour (from Scottish National Party)
Conservative (hold)
Conservative (from Labour)
Liberal Democrat (hold)
Liberal Democrat (from Conservative)
Liberal Democrat (from Scottish National Party)
Scottish National Party (hold)
Scottish National Party (from Conservative)
Sinn Fein (hold)
Independent (gain from Labour)
Reform UK (from Conservative)
Democratic Unionist Party (hold)
Green (hold)
Green (from Conservative)
Green (from Labour)
Plaid Cymru (hold)
Plaid Cymru (gain from Conservative)
Social Democratic and Labour Party (hold)
Alliance Party (from DUP)
Ulster Unionist Party (from DUP)
Traditional Unionist Voice (from DUP)
Independent (gain from Alliance)
Hexagons modified from data by Philip Brown and Alasdair Rae
Geographic view
Map showing land area
Constituency view
Map showing seats in parliament
Population view
Map showing population distribution

The problem with colour

- But not everyone sees colour in the same way
- Do you need colour in your plot? If not, don't use it.

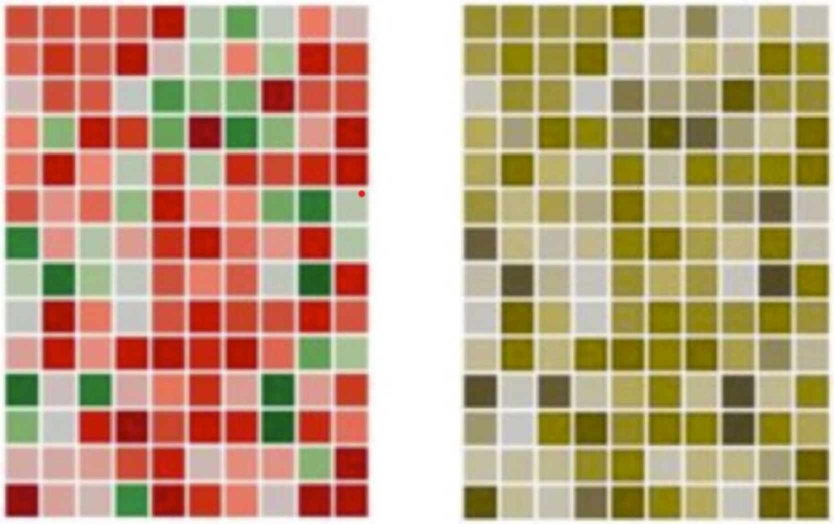
White and gold or blue and black? A 2015 meme →



Colour blindness

- About 8% of men and 0.5% of women have some form of "colour-blindness"
- Most common form: red-green colour blindness
- Avoid red-green contrasts

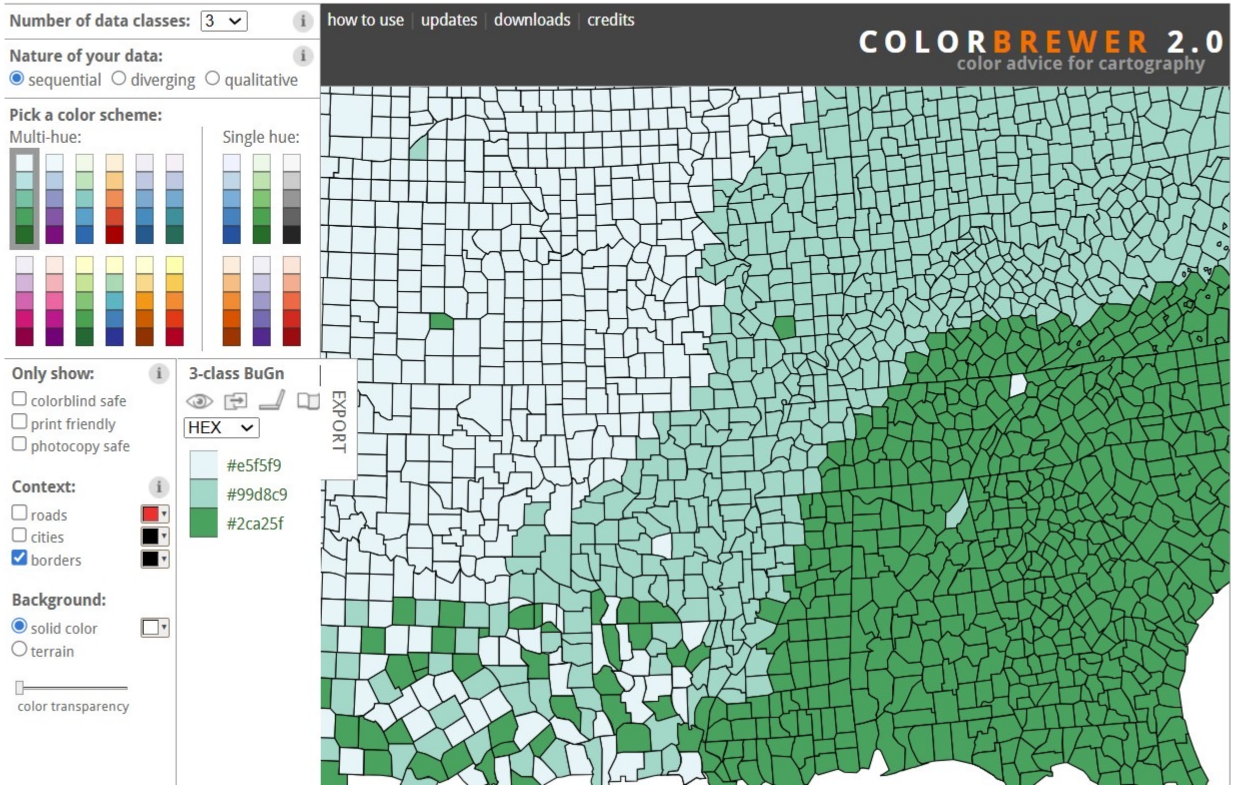
Source: https://www.tableau.com/en-gb/blog/examining-data-viz-rules-dont-use-red-green-together



Choose and appropriate palette

- When using color, pick an appropriate palette:
  - Sequential
  - Diverging
  - Qualitative

See colorbrewer2.org



Number of data classes: 3

Nature of your data:
● sequential ○ diverging ○ qualitative

Pick a color scheme:
Multi-hue: Single hue:

Only show:
□ colorblind safe
□ print friendly
□ photocopy safe

Context:
□ roads
□ cities
☑ borders

Background:
● solid color
○ terrain

color transparency

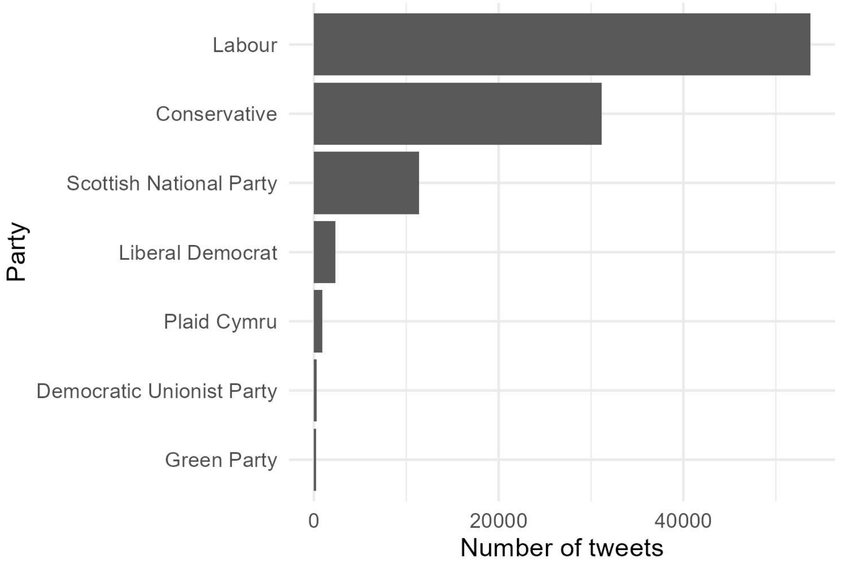
3-class BuGn

HEX

▢	#e5f5f9
▢	#99d8c9
🟩	#2ca25f

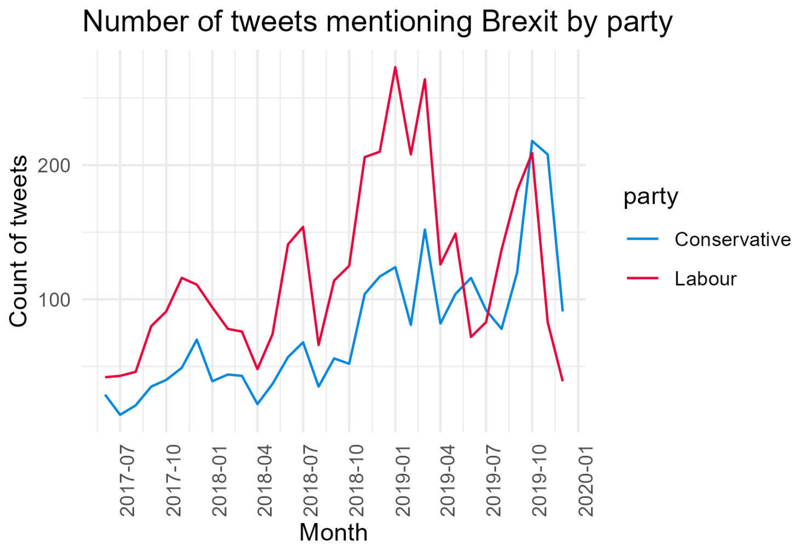
Common plot types: bar charts

- Tried and true choice – if it works, use it
- Good for comparison between groups
- Works with more than one grouping factor (grouped bar chart)



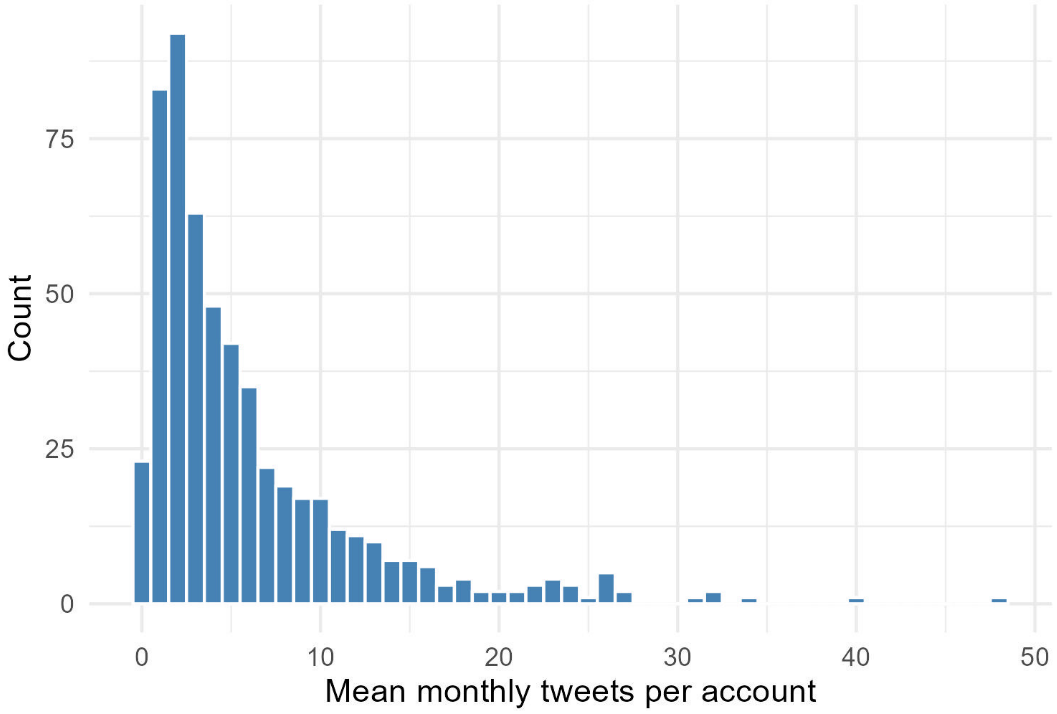
Common plot types: line plots

- Good for time trends
- Can add group information using colour or other elements



Common plot types: Distributions

- Visualising the distribution of key variables of interest can inform analysis
- A histogram shows the number of observations falling bins (small increments of the data)
- Alternatives: density plot, violin plot

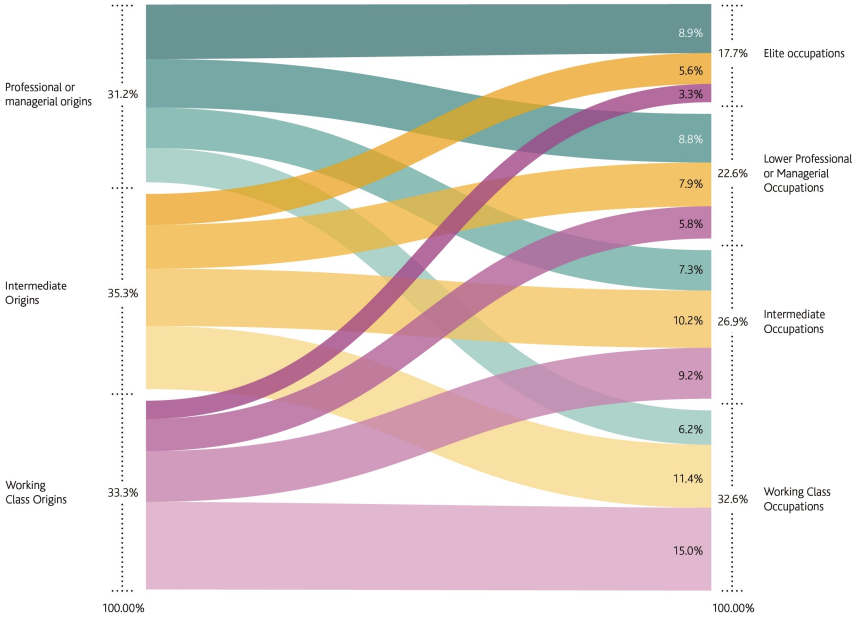


Small multiples

- Tufte: “At the heart of quantitative reasoning is a single question: Compared to what?”
- Small multiple: a series of similar plots showing the difference (in trend, distribution, etc) by relevant group variables.

Visualising flows

- A Sankey plot is a good way to visualize flows (changes from one state to another)
- This plot: visualising social mobility (change in class from parent’s to children’s generation)

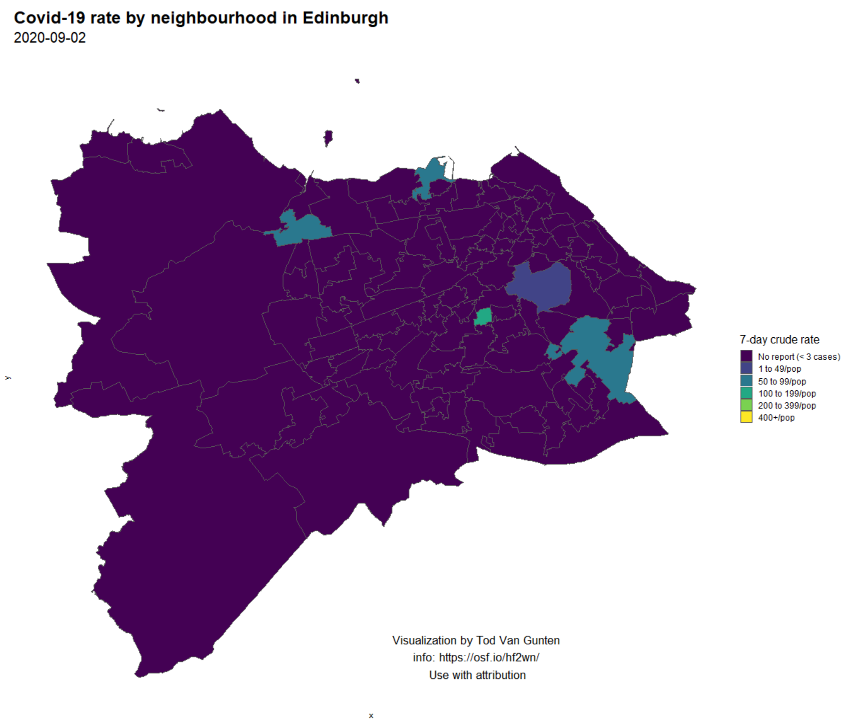


Animated plots

- Animated plots can make it easier to spot trends

Animated plots

- Cloropleth maps: represent spatially-distributed data using colour and familiar geographic outlines.
- Animated version can

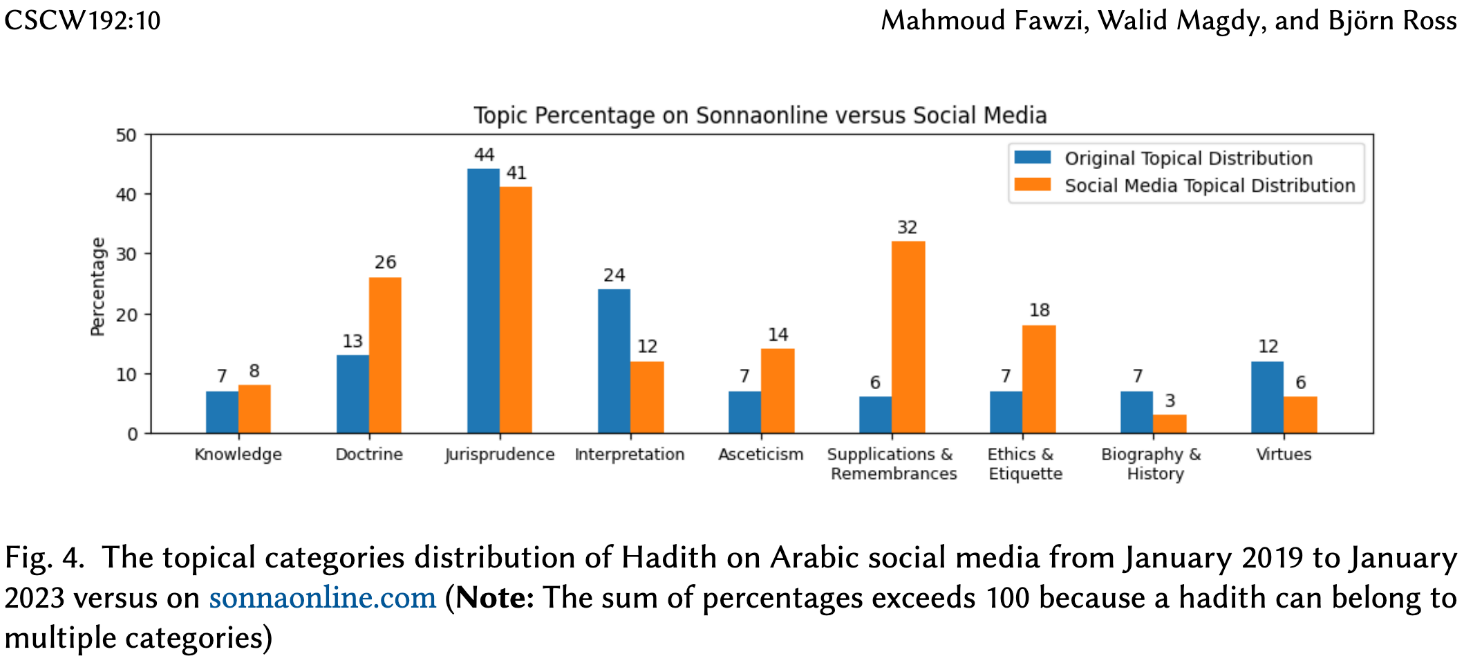


Covid-19 rate by neighbourhood in Edinburgh

2020-09-02

Visualization by Tod Van Gunten
info: https://osf.io/hf2wn/
Use with attribution

Religious discussions online: visualisations



CSCW192:10 Mahmoud Fawzi, Walid Magdy, and Björn Ross

Fig. 4. The topical categories distribution of Hadith on Arabic social media from January 2019 to January 2023 versus on sonnaonline.com (**Note:** The sum of percentages exceeds 100 because a hadith can belong to multiple categories)

Religious discussions online: visualisations

Exploring Hadith Presence on Arabic Social Media CSCW192:11

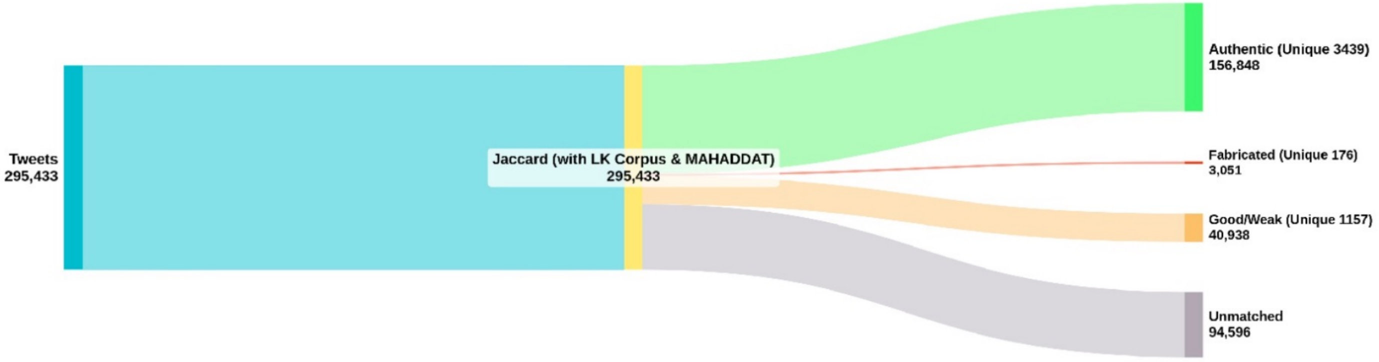
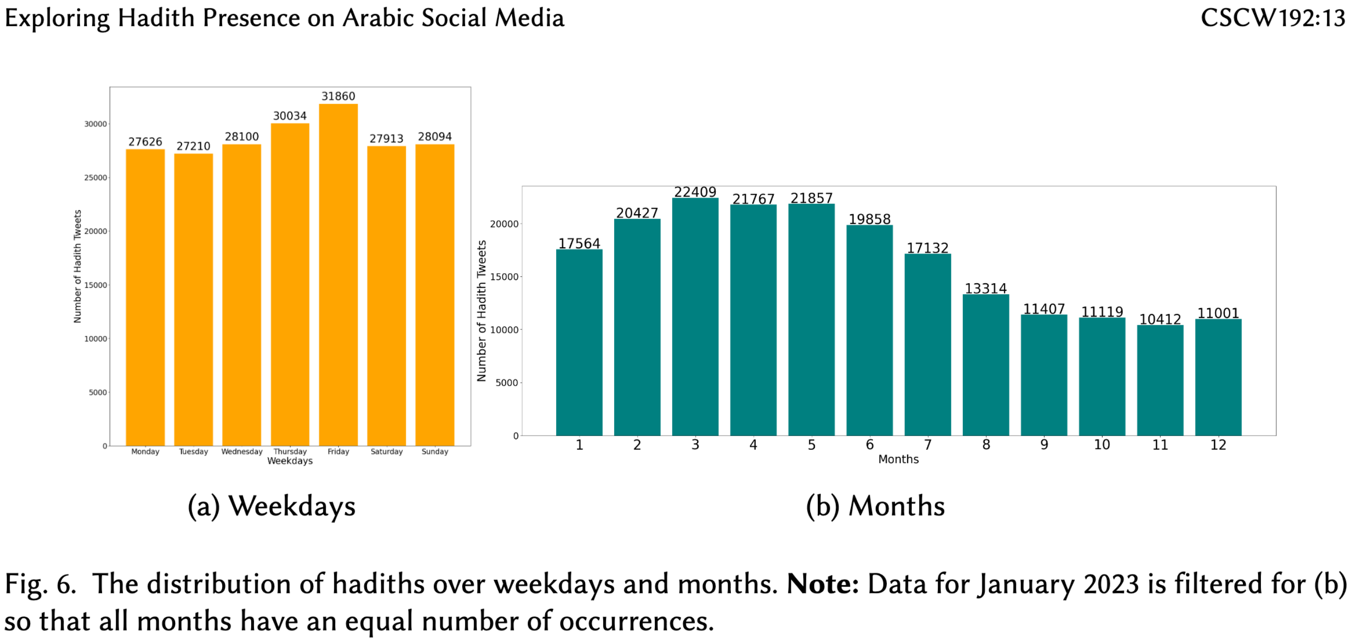


Fig. 5. The authenticity distribution of hadiths on Arabic social media from January 2019 to January 2023

Religious discussions online: visualisations



Exploring Hadith Presence on Arabic Social Media CSCW192:13

(a) Weekdays

(b) Months

Fig. 6. The distribution of hadiths over weekdays and months. **Note:** Data for January 2023 is filtered for (b) so that all months have an equal number of occurrences.

Questions?

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science