

Text Analysis

Introduction to Data Visualization

The Graduate Center at CUNY | Summer 2018

June 12, 2018

Five Minute Reflection

Moretti presents us with a problem: “a canon of two hundred novels sounds very large for nineteenth-century Britain, but is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows—and close reading won’t help here, a novel a day every day of the year would take a century”—Moretti, 2003, pg 1.

Computers cannot “read” the way people do. Given this, do you think that reading these texts computationally (distant reading) is possible? Why or why not? Do you think there is anything to be gained in this approach?

Texts as Data

- Corpus/Corpora
- Collection of documents
 - Tweets
 - Books
 - Zines
 - Text messages
 - Newspaper Articles
 - Supreme Court proceedings

Working with text

- Given a text, what can you quantify?

Working with Corpora

MetaData

- Date (Time)
- Location
- Creators (author, publisher, etc.)
- Platform/Format
- Sales/Retweet/Distribution
- Length
- Keywords
- Title
- Revision History

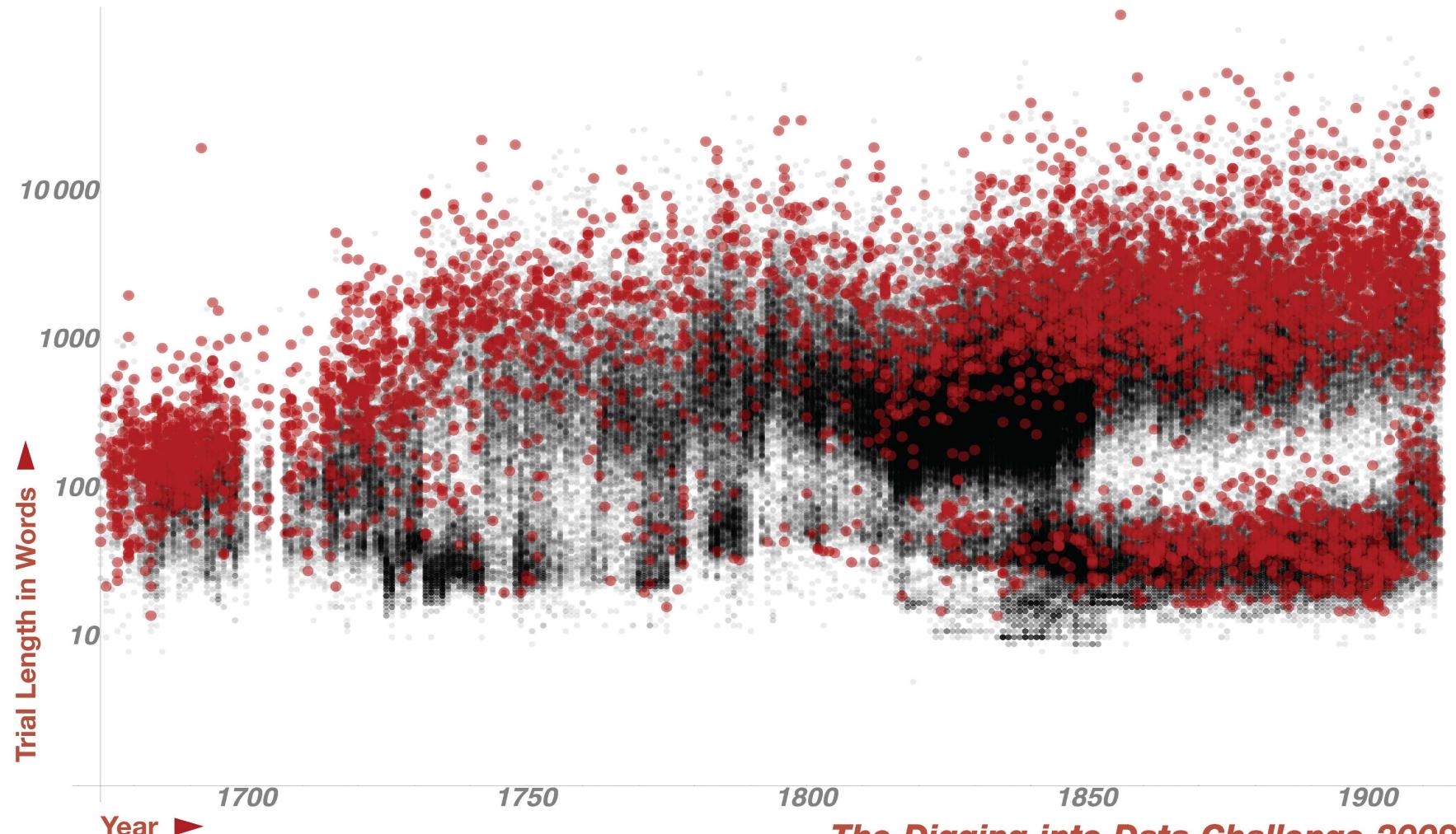
Content

- Word frequencies
- N-grams
- Word offset
- Topic Modelling
- Sentiment Analysis
- Predictions

Data Mining With Criminal Intent

Cyril Briquet • Dan Cohen • Frederick Gibbs • Tim Hitchcock • Jamie McLaughlin • Geoffrey Rockwell
Joerg Sander • Robert Shoemaker • John Simpson • Stéfan Sinclair • Sean Takats • William J. Turkel

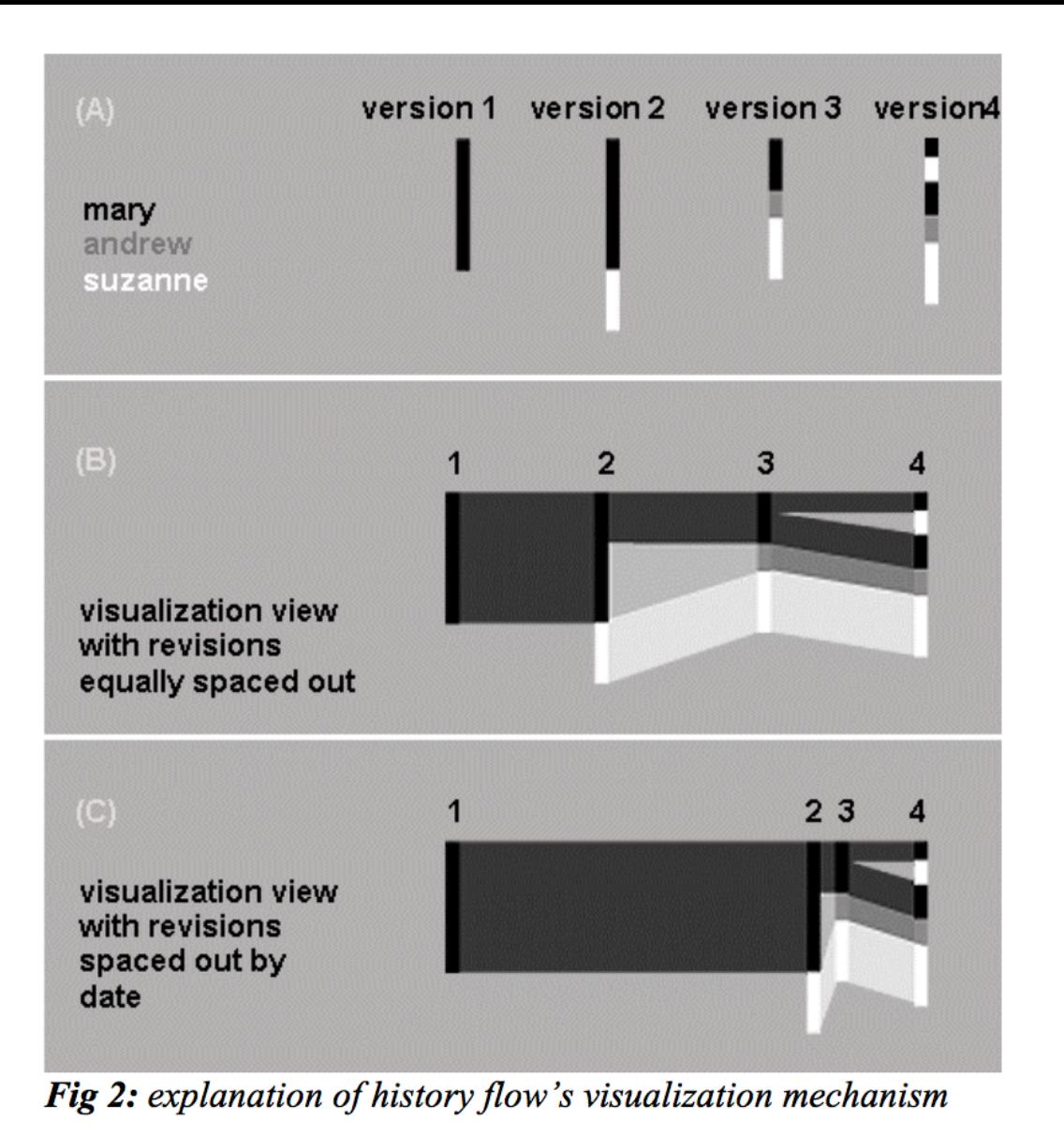
Metadata



Length in words of trials involving 'killing' (red) versus all other trials (black) from the Old Bailey Proceedings, plotted with a logarithmic scale for the x axis. 197,745 Trials.

The Digging into Data Challenge 2009

NEH • JISC • SSHRC



Visualizing Edits

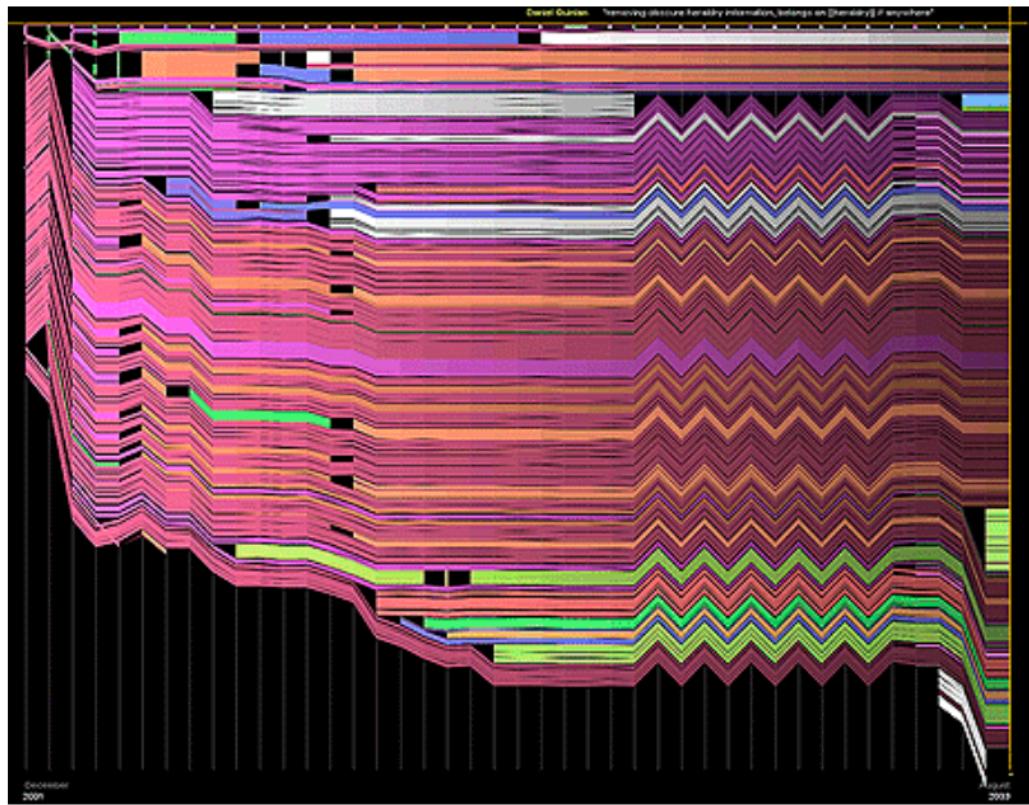


Fig 6: “Chocolate” page spaced out by number of versions; we can see the zigzag pattern of an edit war.

Visualizing Text

Metadata

- Darwin's Origin of the Species
Edits <https://fathom.info/traces/>

Content

- Presidential Debates (n-grams)
https://public.tableau.com/views/Election2016-SpeechAnalysis/Newspaper?:embed=y&:display_count=yes&:toolbar=no&:showVizHome=no
- <https://books.google.com/ngrams/>

Metadata

- Can usually be formatted for visualization with minimal cleaning
- Categorical or Numerical
- Tells a story about the text



Working with Text

Some Cleaning

- Word Counts/Frequencies
- Offsets
- N-grams

A Lot of Cleaning

- Comparing two or more texts
- Lexical Density/Diversity
- N-grams
- Topic Modelling
- Sentiment Analysis
- Predictions

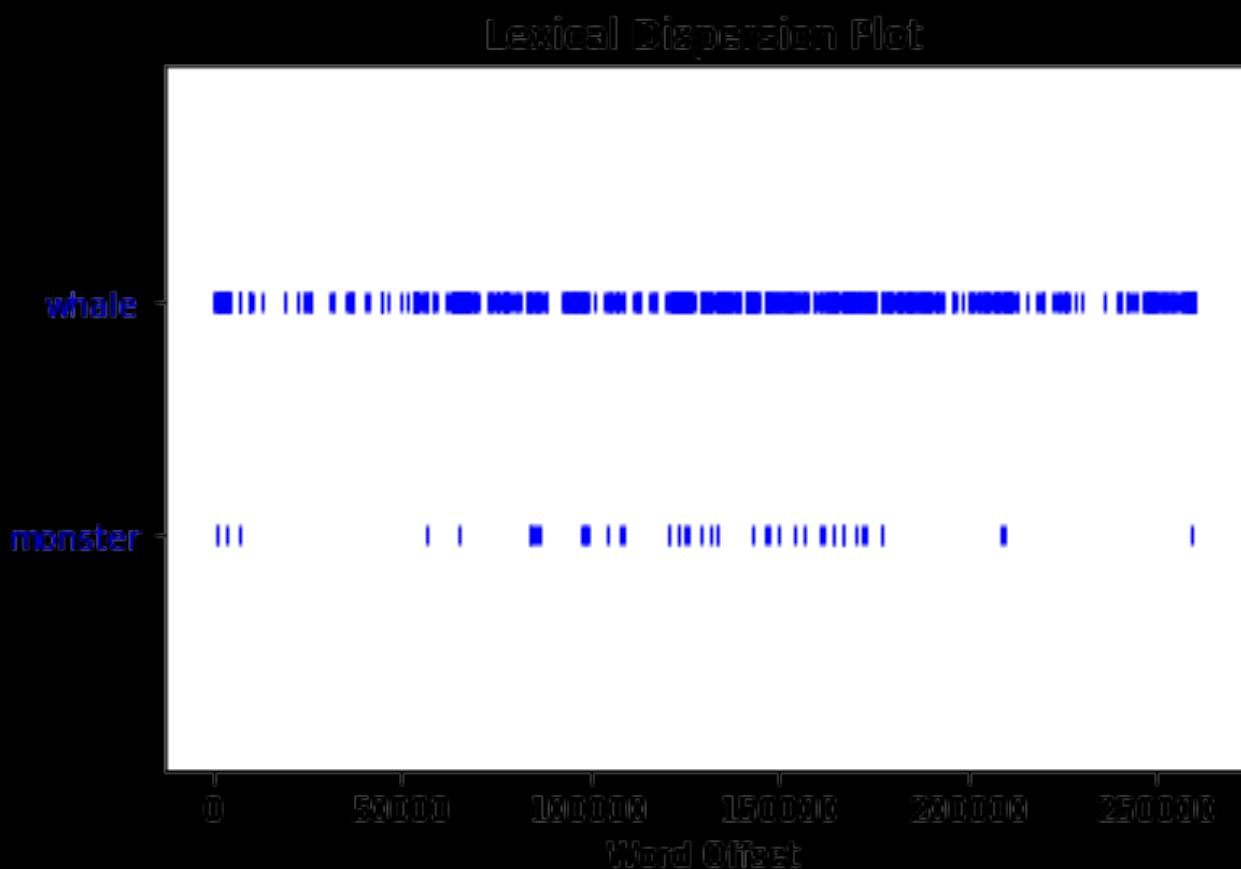
Cleaning Text

DON'T DO THIS WITH TABLEAU

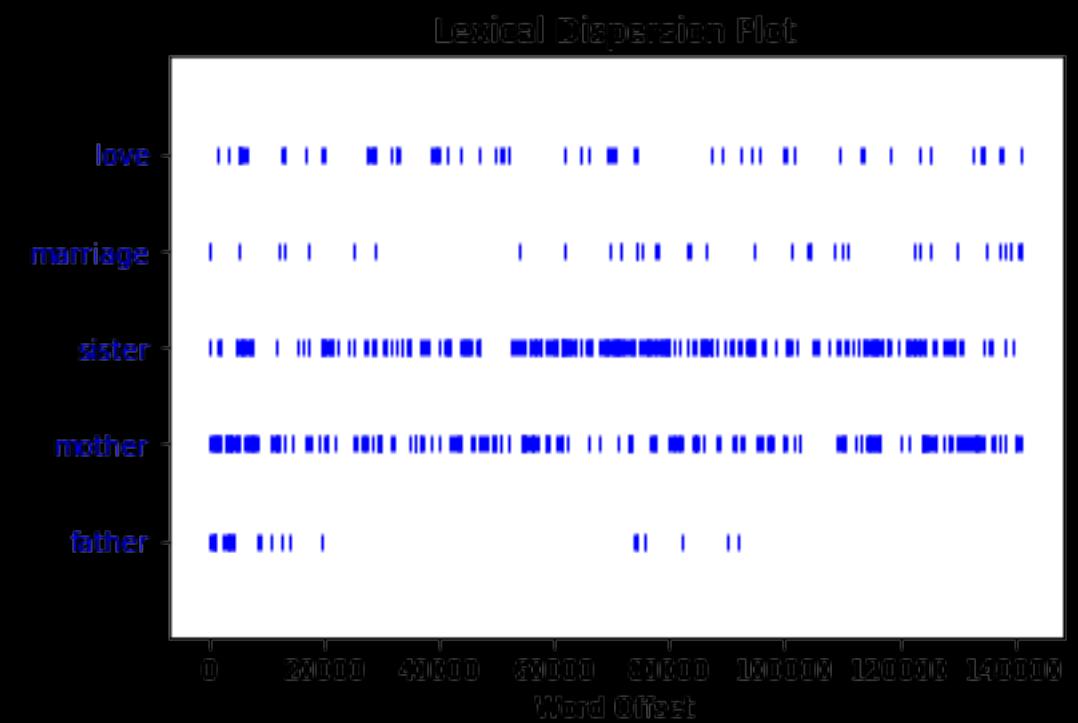
- Tokenize (split into words)
- Normalize (remove punctuation and capitalization)
- Remove Stopwords (determiners, prepositions, auxiliaries, etc.)
- Stem or Lemmatize (transform “run”, “runs”, and “running” into “run”)

Frequency Distributions

Moby Dick (Herman Melville)



Sense and Sensibility (Jane Austen)



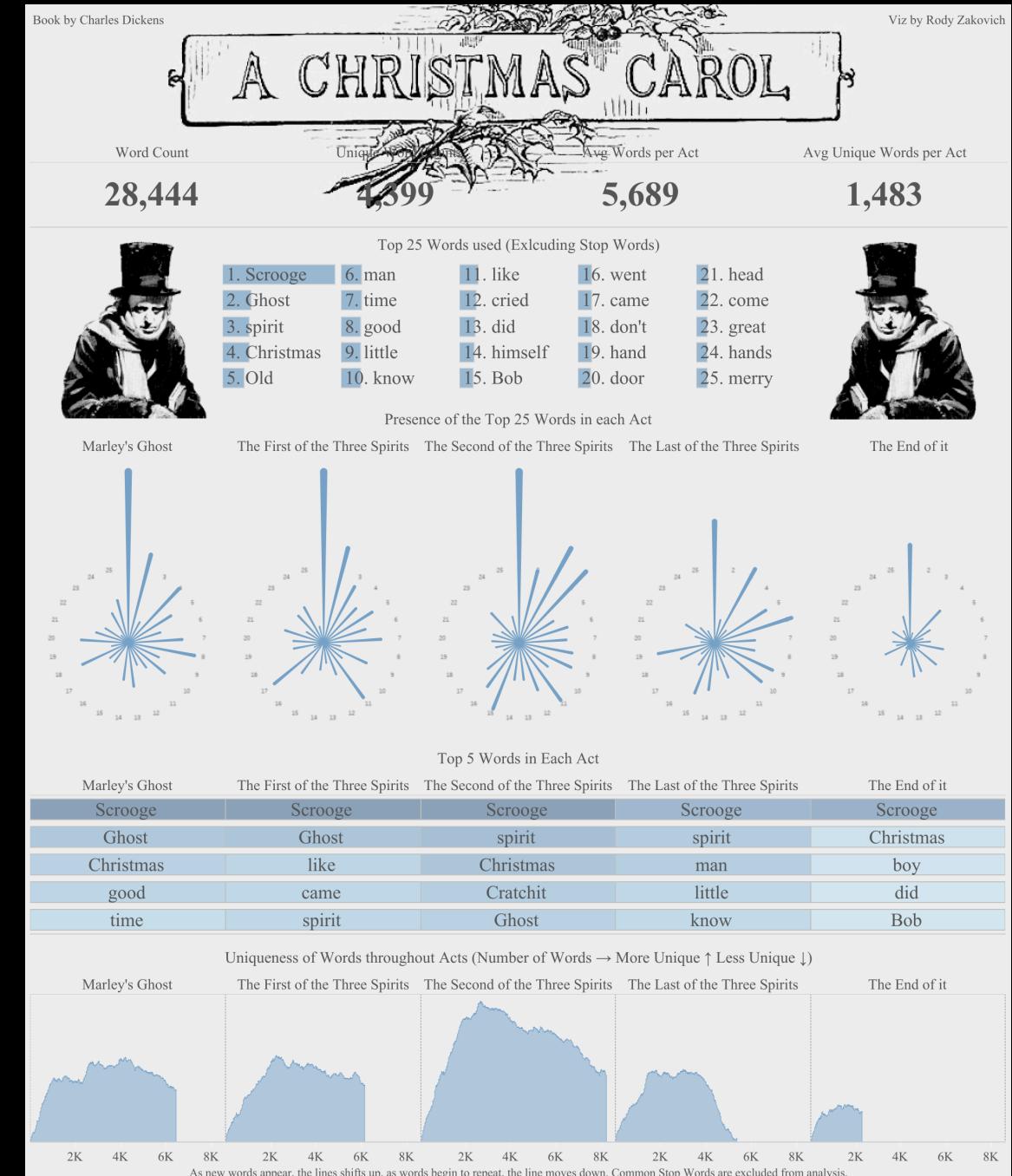
Cleaning Text

Hunger Games (Dirty Text)

1 iby 1s 2 3 4 5 6 7 7:00 7:30 8 8:30 8th 9 10 11 12 12. 12s 13 13's 13s 14 15 15:00 16 17 18 18:00 18:30 19 19.I 20 21 22 22:00
22:30 23 24 25 25.Mutations 26 27 39 75 100 150-pound 220-pound 250 307 451 2212 3901 3902 3903 3908 `` -- - a -Leeg -takes -there -to -two
; ; ! ? . . ' ' 'Because 'Bread 'Bring 'd 'I 'l 'm 'Next, 'propaganda 're 's 'seize 'Star 'the 'The Thirteen 'to 've 'We 'Yeah () A a Aa
aback abandon abandoned abandoning abandonment abandons abate abdomen abduct Abernathy aberration abhorrent abide abilities ability ablaze able ablebodied
abnormally aboard abounds about ABOUT About above Above aboveground abrasive abreast abrupt abruptly absconded absence absences absent absently
absolute absolutely Absolutely absorb absorbs absorbed Absorbent absorbing absorbs absorption abstain abstract absurdity abundance abuse abused
abuses abusing abyss accelerated accent accents accept acceptable accepted accepting accepts access accessible accessories accessory accident accidental
accidentally accidents accommodate accompanied accompanies accompany accompanying accomplice accomplish accomplished according account accountable
accounted accounts accumulated accuracy accurate accurately accusation accusations accused accusingly accustomed acerbic ache aches achieve achieved
achievement aching acid acid-damaged acidic acker acknowledge acknowledged acknowledging acknowledgment ACKNOWLEDGMENTS acloth acorns acquaintances
acquainted acquire acquired acquisition acrid Across across act acted action Action actions.Numerous activate activated activates active actively
activities activity actor acts actual actually Actually acutely adapt adapting Add add added additive addicts adding additional Additional Additions additions
address addressed addresses addressing adds adequately adhere adjacent adjoining adjourned adjourns adjust adjusted Adjusting adjusting adjusts administered
Admiration admiration admire admired admirers admiring admission admit admits admitted admitting adopt adopted Adoration adores adoring adorned adorns
adrenaline Adrenalin Adrienne drift adult adults advance advanced advantage advantages adventures adversaries advice advisable advises advisor advisors
aerodynamics affair affect affectations affected affecting affection affectionate affects affixed affixes afford afforded afield afloat afoot afraid Afraid After
after aftereffects aftermath afternoon afternoons afterthought afterward Again again Against against age aged agenda agent ages aggravating
aggravation aggression aggressive aghast aging Agitated agitation ago agonies agonized agonizing Agonizing agony agree agreed Agree agreeing agreement
agrees Agriculture agriculture Ah ahead Ahs aids ail Ains aim aimed aiming aimlessly Air air Airborne aircraft aired airing airlift airpower airs
airspace airstrip Airtime airtime airwaves airway aisles ajar alarm alarmed alarming alcohol alcove alcoves alert alerted alertness alerts alibi alien alight
alignment alike alive Alive all All all-consuming all-powerful all-too-familiar allied allegations allegiance alleviate alley allies alliance alliances allied allies Allies
allotted allow allowance allowed allowing allows allude ally Alma almost Almost aloft alone Alone Along along alongside Aloof aloud already Already
alreadyFull Also also alterations alteration altered alternate alternately alternates alternating alternative alternatives although Although altogether always
Always am Am amaze amazed amazes amazing Amazingly amber ambiguous ambush America amiable amiss among Among amount ample amplifies amputate
amused amusement amuses amusing An an analyze anatomy ancestors anchor ancient And and andburned andtend Andrea anecdote anesthetize anger Anger
angers angle angles angrily angry anguish angular animal animals animate animated ankle ankle-deep ankles ankles Annie Annie's annihilation anniversary
announce Announced announced announcement announcer announcers announces announcing annoy annoyance annoyed annoying annoys annual annulling anoint
anointed anonymously another Another another's anotherproblem another's answer answered answering answers anthem anthems anti-Capitol anti-climactic
anti-infection antiaircraft anticipate anticipated anticipates anticipating Anticipating anticipation antis antisepsis antlers Ants ants anxiety anxious
anxiously Any any anybody anymore Anyone anyone anyone Anything anything anytime anyway Anyway anywhere aorta space apart Apart apartment
apartments apologetically apologize apologized apology apothecaries apothecary Appalachia appalled apparent Apparently apparently appeal appealing appear
appearance appearances appeared appearing appears appetite appetites appetizing applauding applause apple apple-sized apples applesauce applications
applied applies apply appointed appointment appraise appreciated appreciates appreciation appreciative apprehend apprehended apprehensively approach
approached approaches approaching appropriate appropriated approval approve approved approving approximate apron aqua arbitrarily arbitrary arc arched archery
arching architectural arduous Are are Area area areas arena Arenas arenas arent argue argued argues arguing argument arguments arise arises arm armband
armed armful arms arming armload armor armored armor arms army aroma arose around Around around Arrange arrange arranged arrangement
arrangements arranges arranging array arrest Arrest arrested arrival Arrival arrivals arrive arrived arrives arrogance Arrogant arrogant arrow arrowheads ...

The Hunger Games (Clean Text)

Cleaned Text Analysis



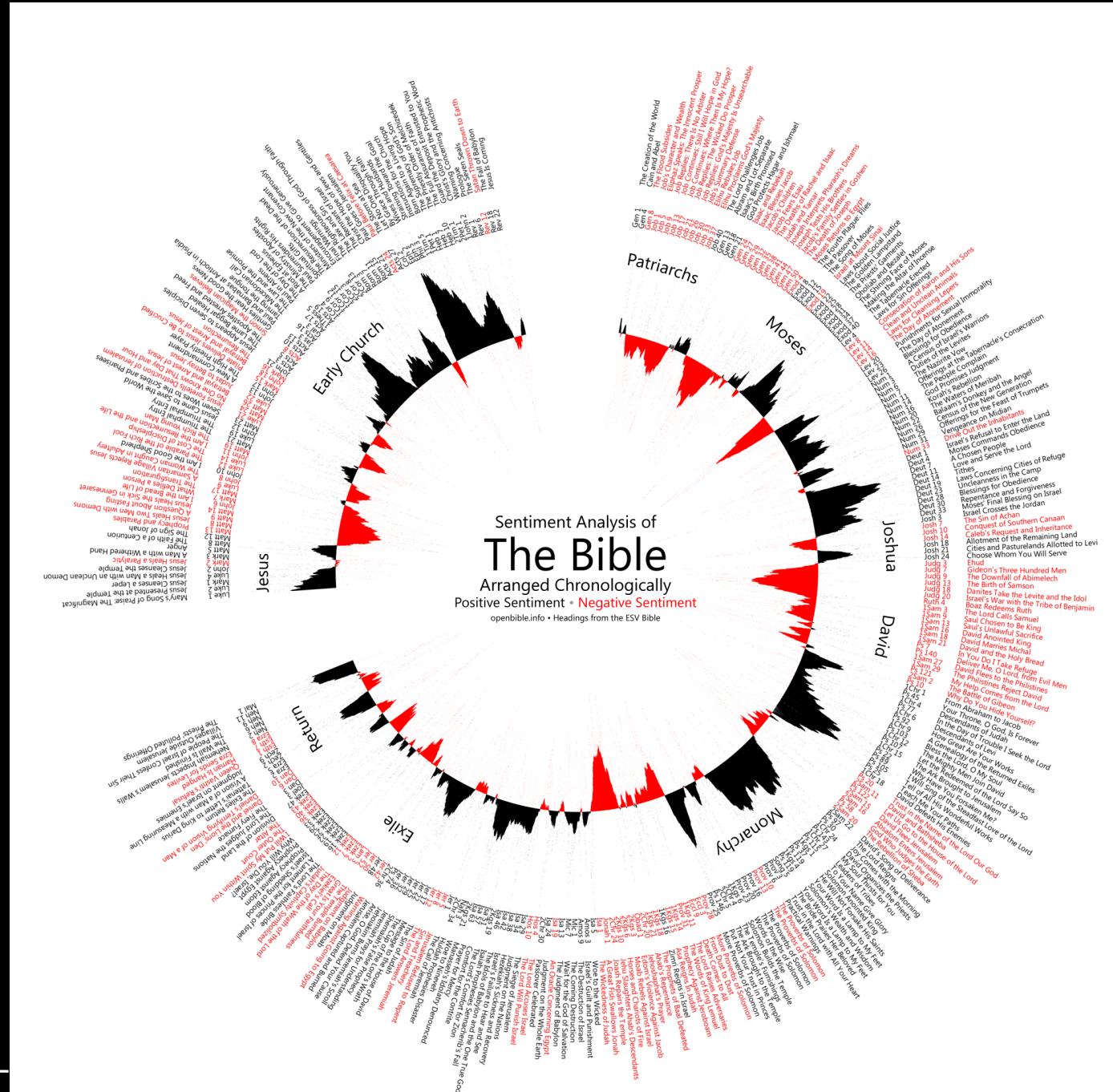
<https://public.tableau.com/profile/rody.zakovich#/vizhome/AChristmasCarolTextAnalysis/AChristmasCarolTextAnalysis>

Texts as visualizations: Sentiment Analysis

Sentiment analysis is a Natural Language Processing technique using Machine Learning

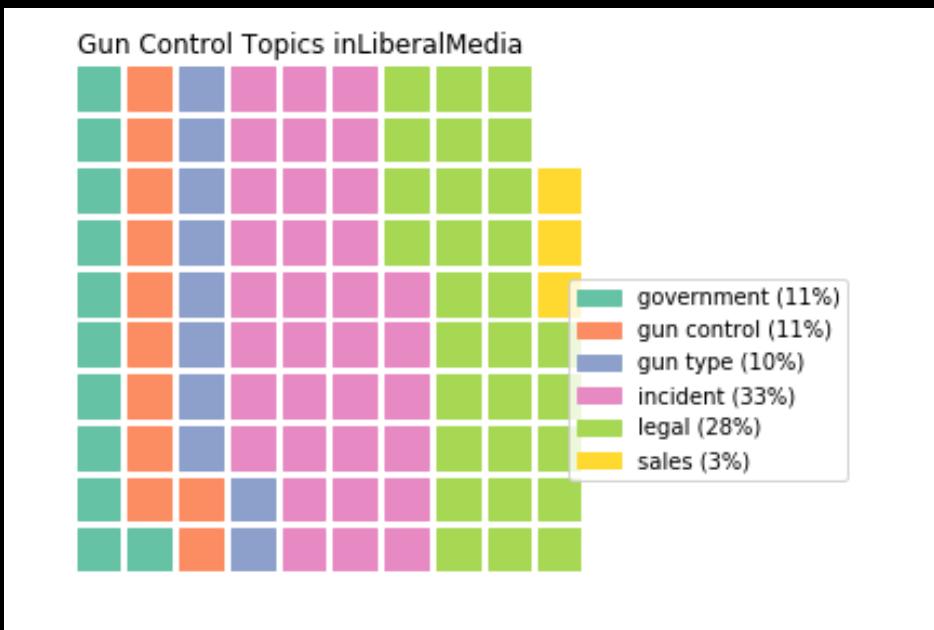
It is a Classification Task (labels each instance) to tag positive or negative usually based on a previously tagged corpus (but not always)

[https://www.biblegateway.com/blog/2011/10/the-ups-and-downs-of-the-bible-a-sentiment-analysis-](https://www.biblegateway.com/blog/2011/10/the-ups-and-downs-of-the-bible-a-sentiment-analysis/)



Texts as visualizations: Topic Modelling

Topic Modelling is a Natural Language Processing technique using Machine Learning



topicID	description
0	gun control violence shooting owner law mass america crime policy
1	police officer department incident city county video killed say man
2	ar 15 rifle automatic semi used round style shooter store
3	school student high shooting florida parkland medium dead control left
4	trump president white house wednesday policy ha support lawmaker also
5	one ha say like people think know would could use
6	state law carry enforcement public texas federal owner official local
7	get news whether often let say already told want life
8	video street 000 country want people show ha world shooting
9	firearm act gun federal home number according crime death handgun
10	shot killed shooting two man time fired one fire year
11	company store sale rifle ha year make group say end
12	check background sale system purchase gun record federal buy criminal
13	com second amendment twitter news follow bullet armed american host
14	nra national association member rifle group right republican political amendment
15	bill republican house legislation congress lawmaker measure would support president
16	new york city year old ha week part time street
17	said told official county member authority statement two attack home
18	read 2016 report attack weapon news control government member official
19	court right federal case amendment arm second law family handgun
20	wa year old told man report life went family one
21	weapon assault ban magazine rifle used automatic military style mass

Distant Reading

- Understanding literature through data
- Scientific Method
 - Hypothesis
 - Experiment
 - Analysis
- Topic Modelling / Sentiment Analysis
- Named Entity Recognition/ Relation Detection

Moretti

FIGURE 1. The Hamlet network

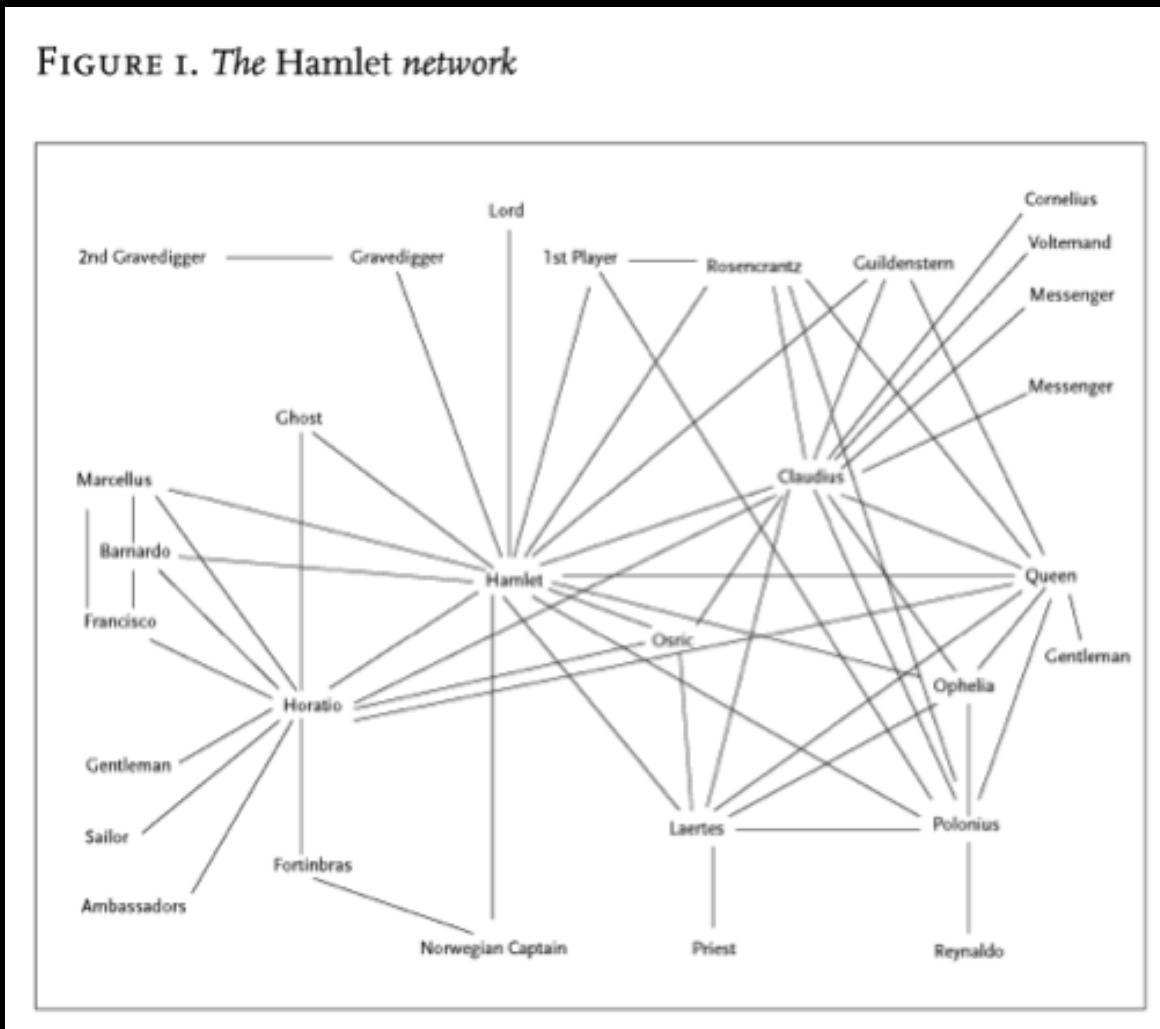
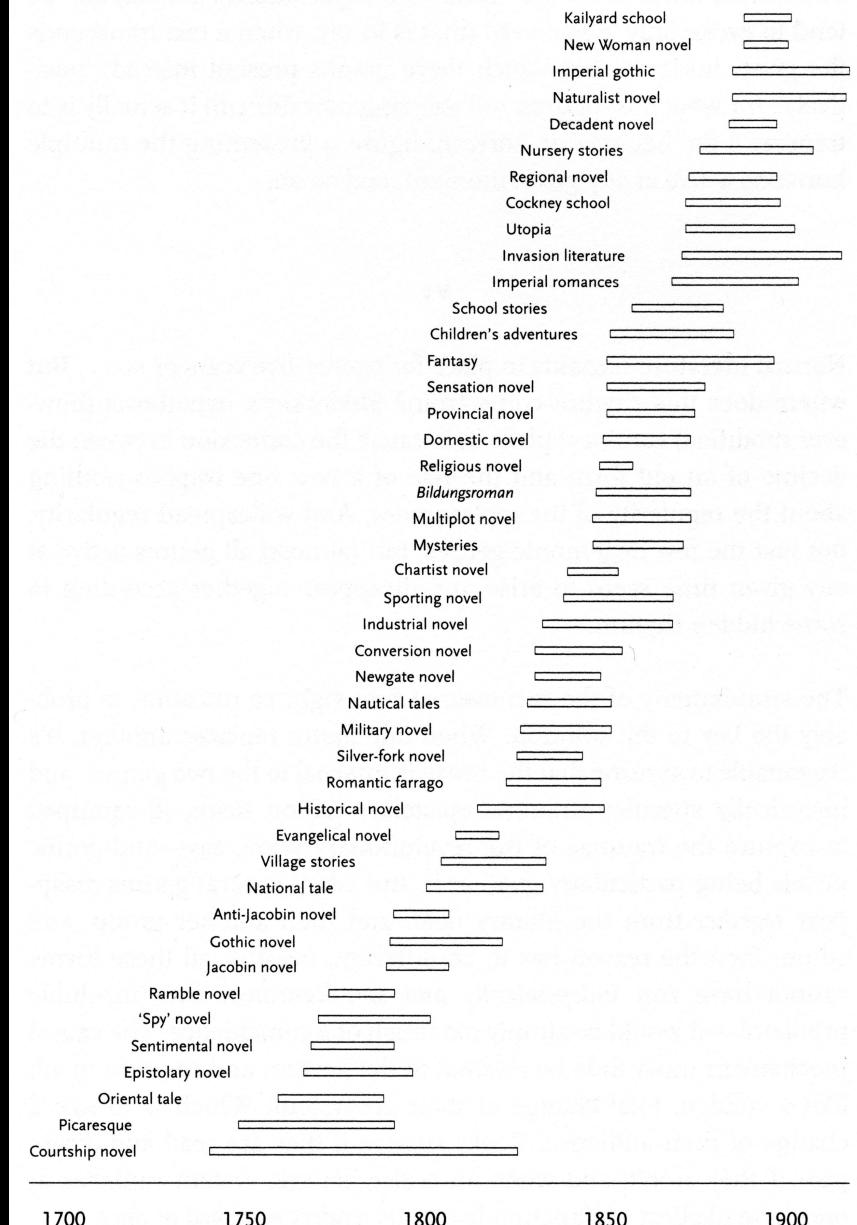


FIGURE 9: *British novelistic genres, 1740–1900*



For sources, see 'A Note on the Taxonomy of the Forms', page 31.

Maps

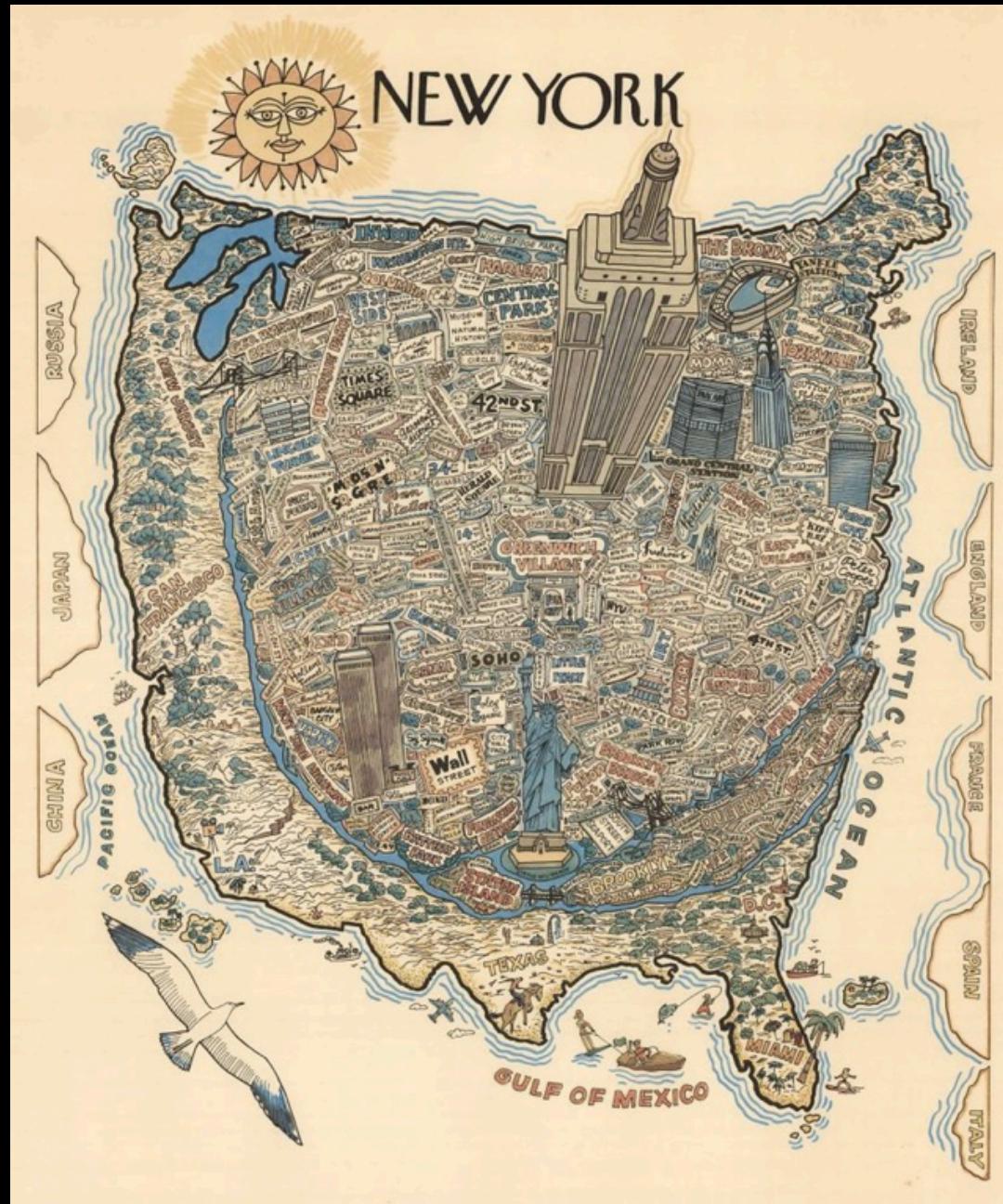
Introduction to Data Visualization

The Graduate Center at CUNY | Summer 2018

June 13, 2018

Five Minute Reflection

- What makes a map different from a visualization?



Maps As...

- Referents
 - Event is happening in this place
 - Proxy for spatial categories
- Containers
 - Map boundaries are essential to the visualization
- Spatial Analysis
 - New findings by combining spatial datasets

Maps to represent...

- People
 - Places
 - Things
 - Events
-
- ... All against a backdrop of space/location

BUT...

Critical Cartography

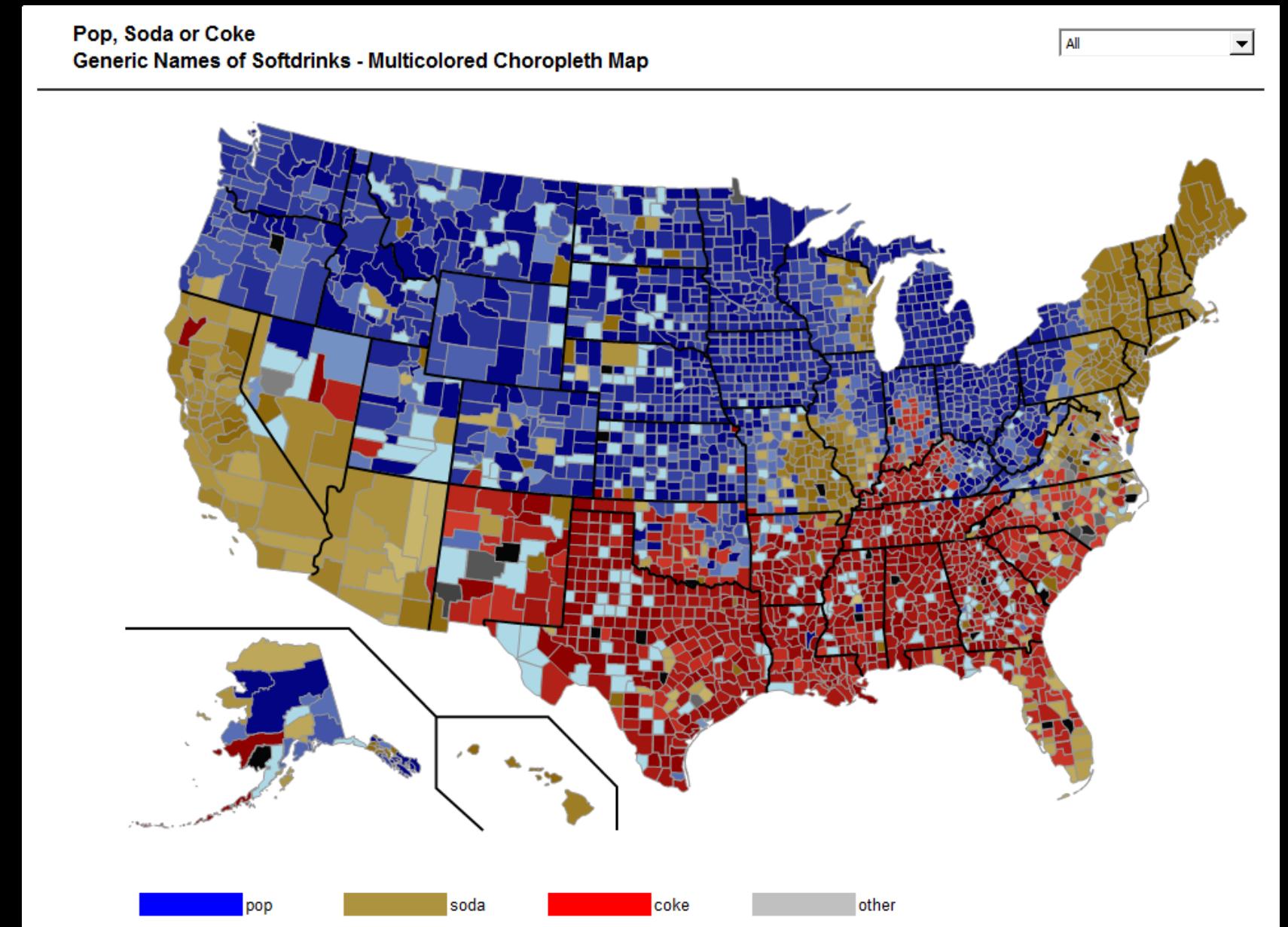
- Gods Eye View
- Supremacy of “science”
- The map (or any visualization) is not always the “truth”...
- Incompatible stories aren’t always wrong

Referents

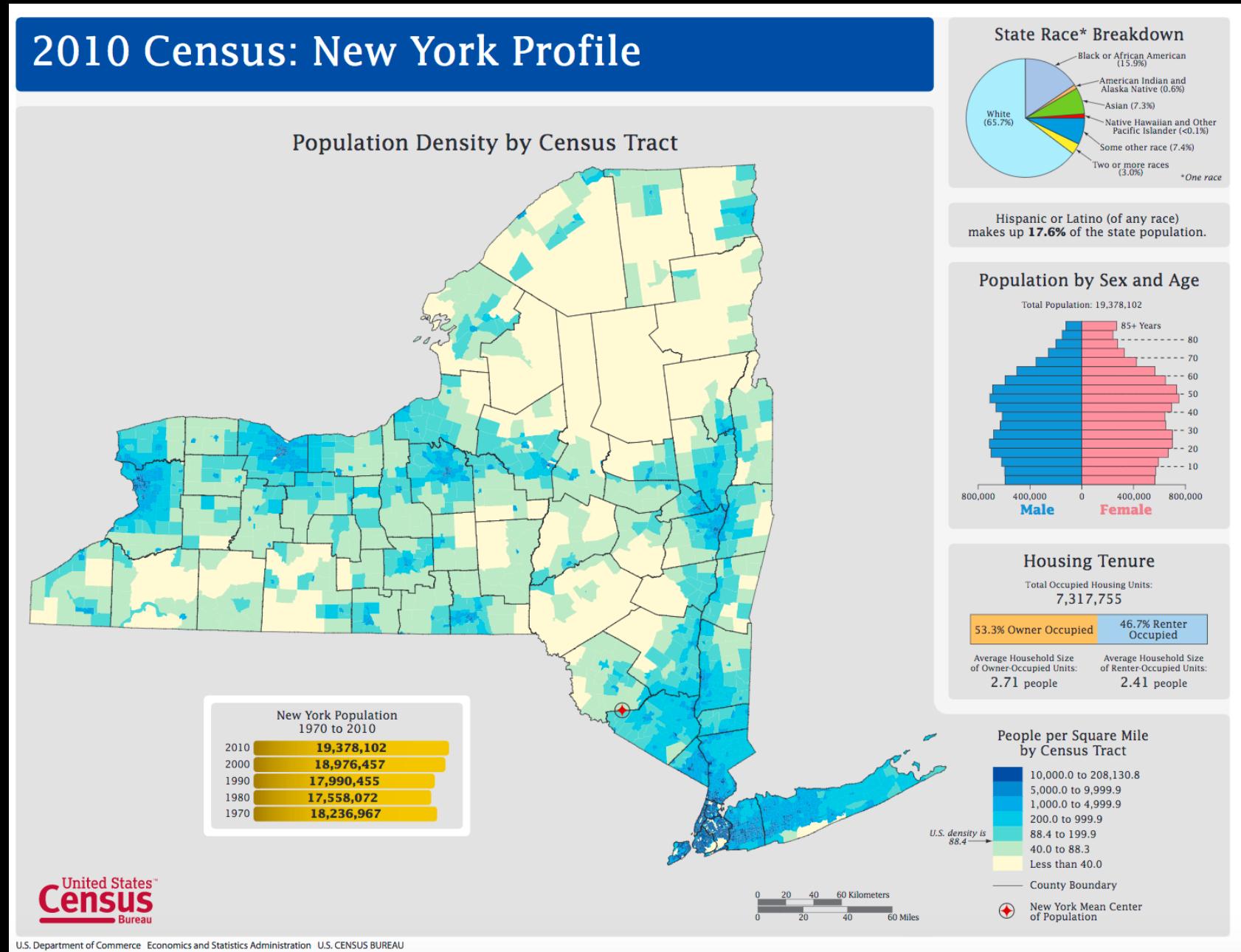
- Stacked Bar Chart
 - www.lucify.com/the-flow-towards-europe/
- Graduated Symbols
 - www.therefugeeproject.org/#/2016
- Wind Map
 - hint.fm/wind



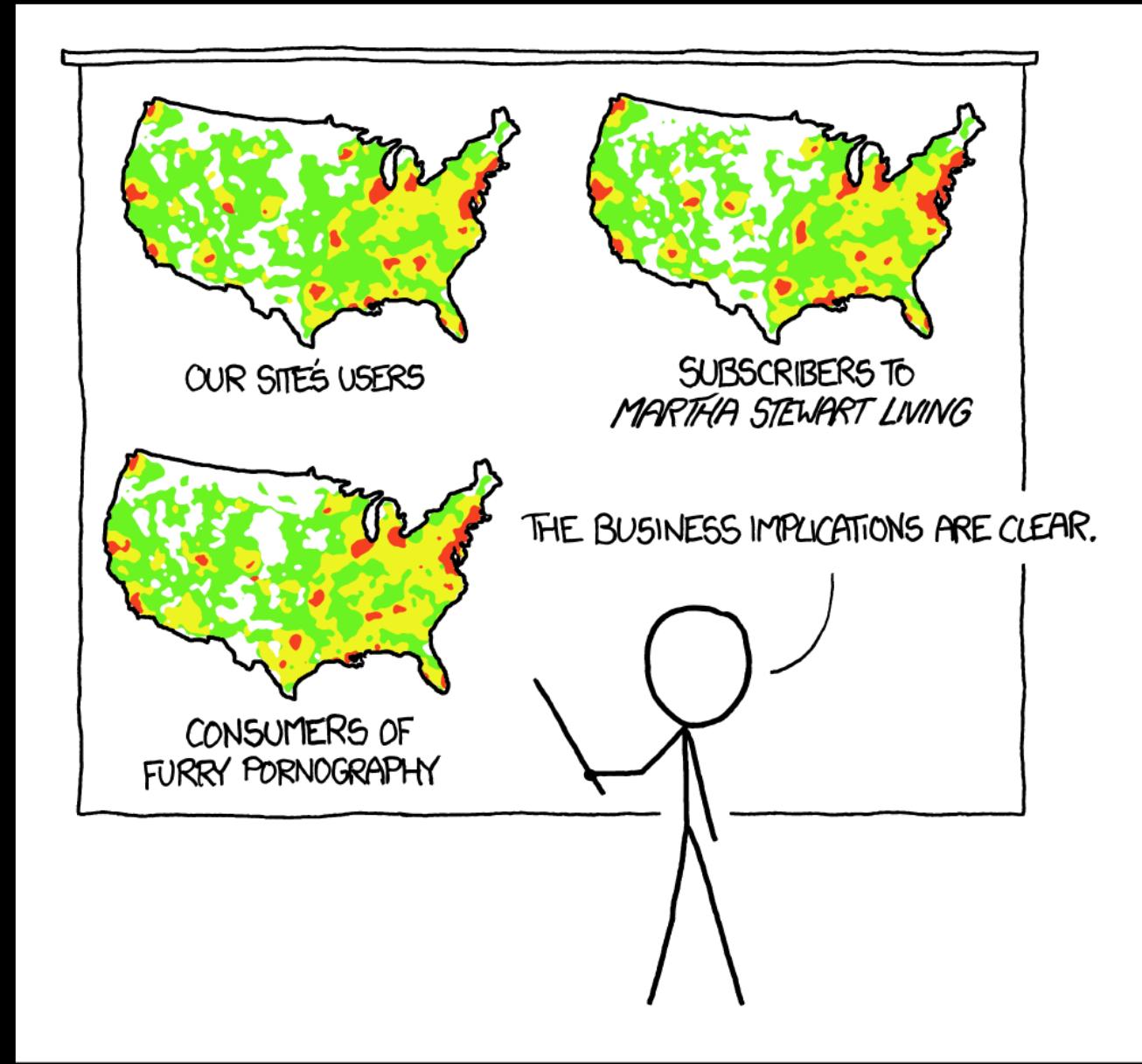
Containers: Categorical Chloropleth



Containers: Numerical Chloropleth



Traps in Normalization & Causation vs. Correlation



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

Chloropleth & Small Multiples

- <https://www.nytimes.com/interactive/2017/08/07/upshot/music-fandom-maps.html?mcubz=3>

Spatial Analysis

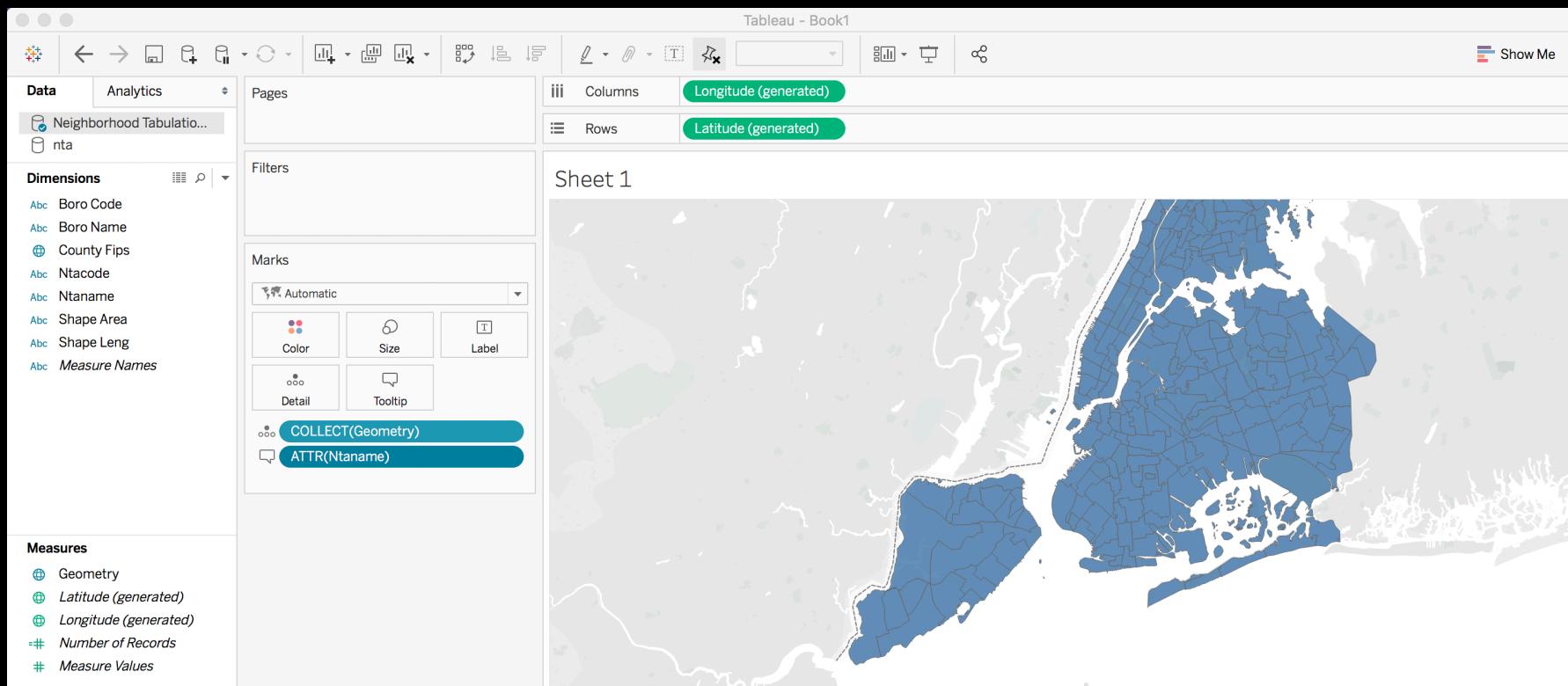
www.subwaylanguages.
michelleajohnson.com

<http://tubecreature.com/#metric=tongues&year=2015&layers=TTTTTF&zoom=13&lon=-0.1500&lat=51.5200>



Adding Shape Files

- <https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq>
- Geojson or KML are easiest (can use ESRI or MapInfo)



Advanced Visualization Techniques & Considerations

Introduction to Data Visualization

The Graduate Center at CUNY | Summer 2018

June 14, 2018

Five minute reflection

- Mapping Activity - Solnit

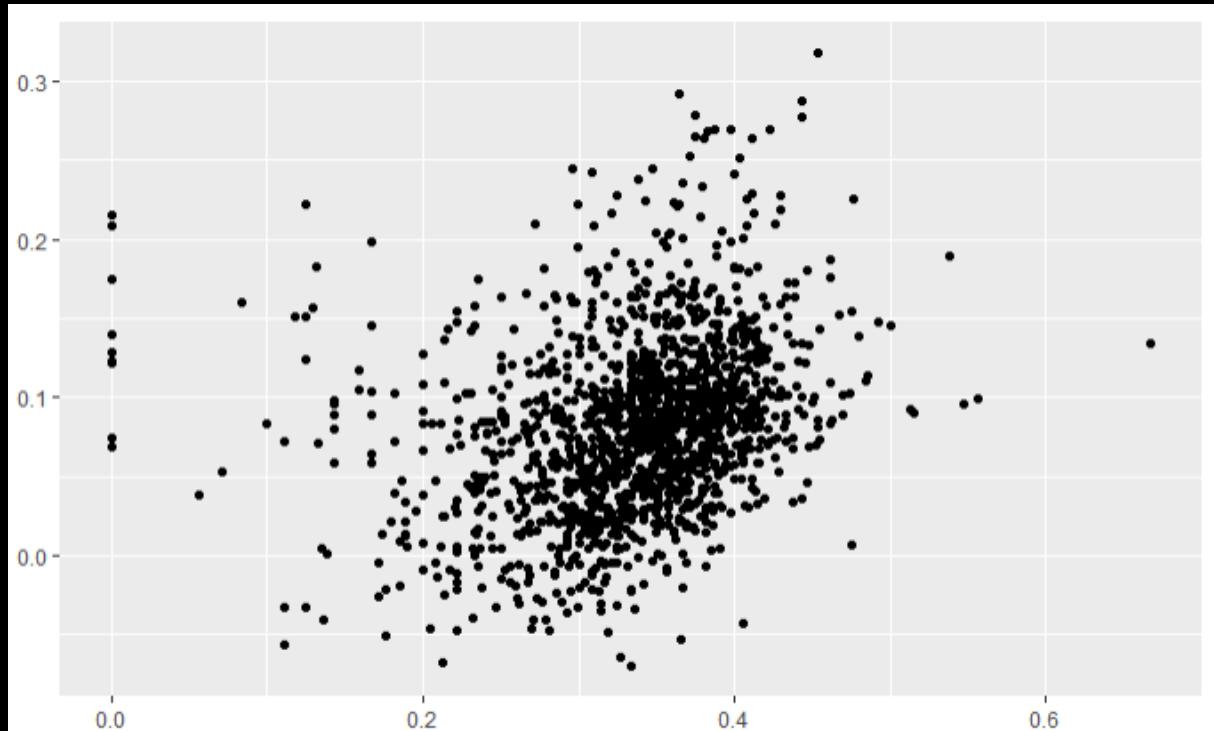
Tufte's Principles

- Labels – if more is needed, write an explanation
- Standardize and normalize units (including money over time)
- Variable dimensions should reflect data dimensions
- Context is essential

Tableau Manages these:

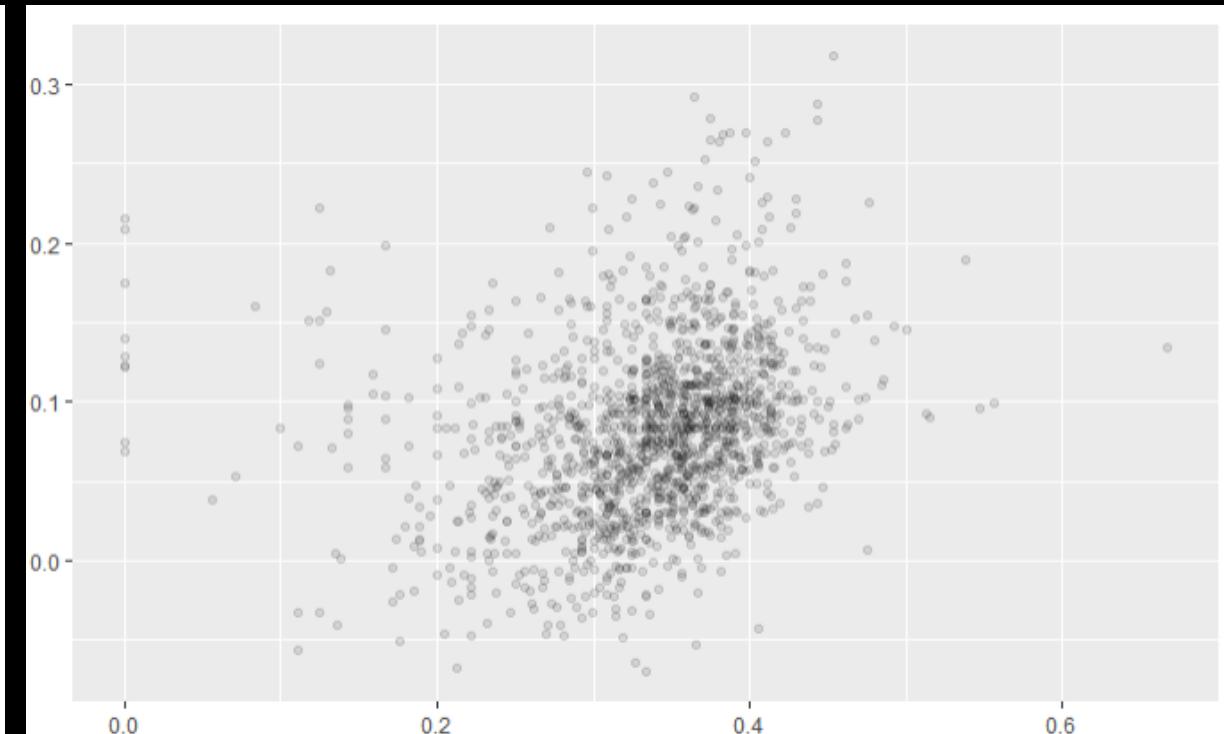
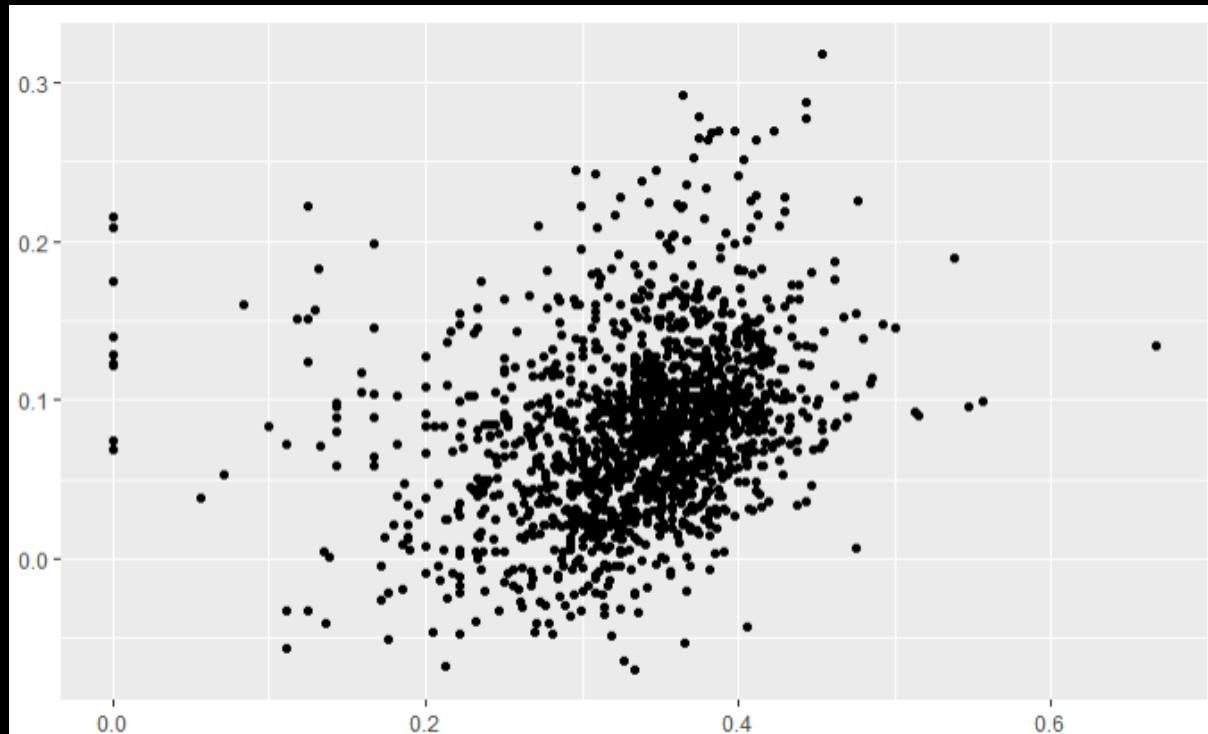
- Keep the design of the plot consistent
- Numerical proportions should be accurately represented by visual

OVERPLOTTING: Point Density

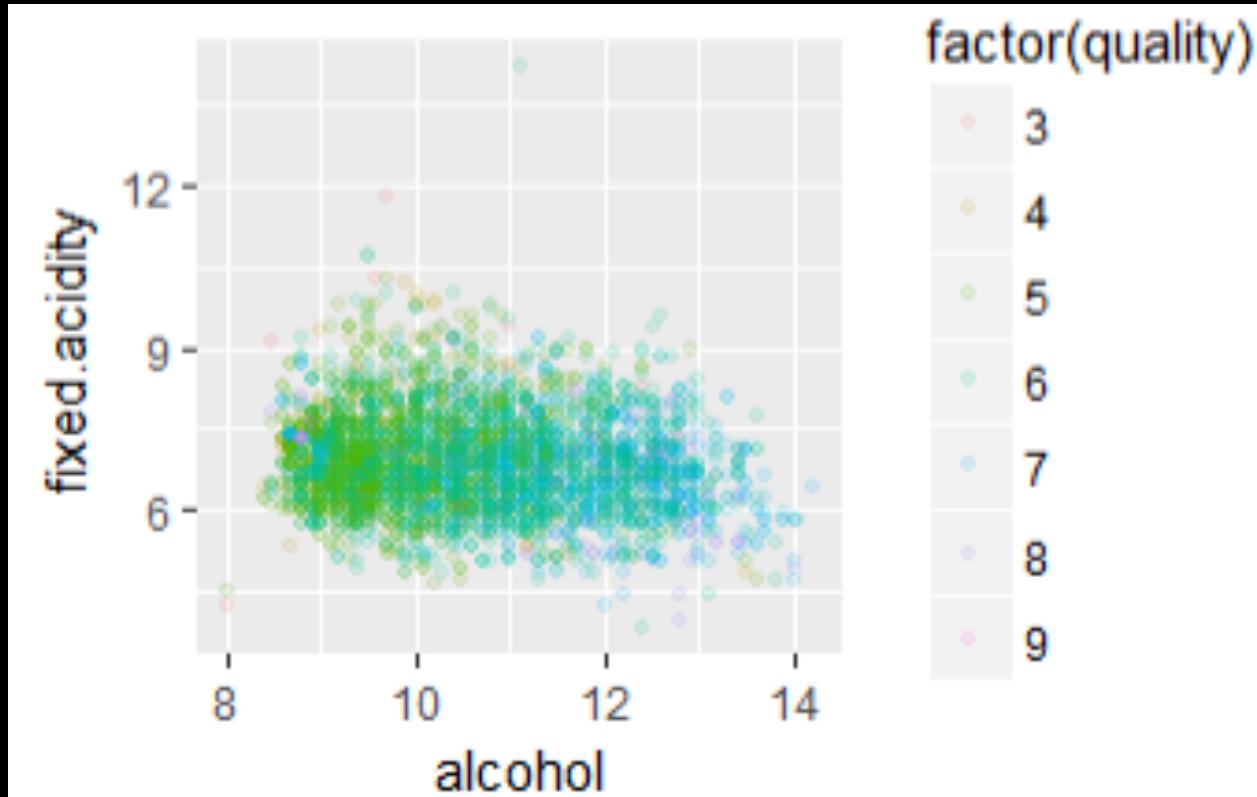


OVERPLOTTING: A SOLUTION

Transparency

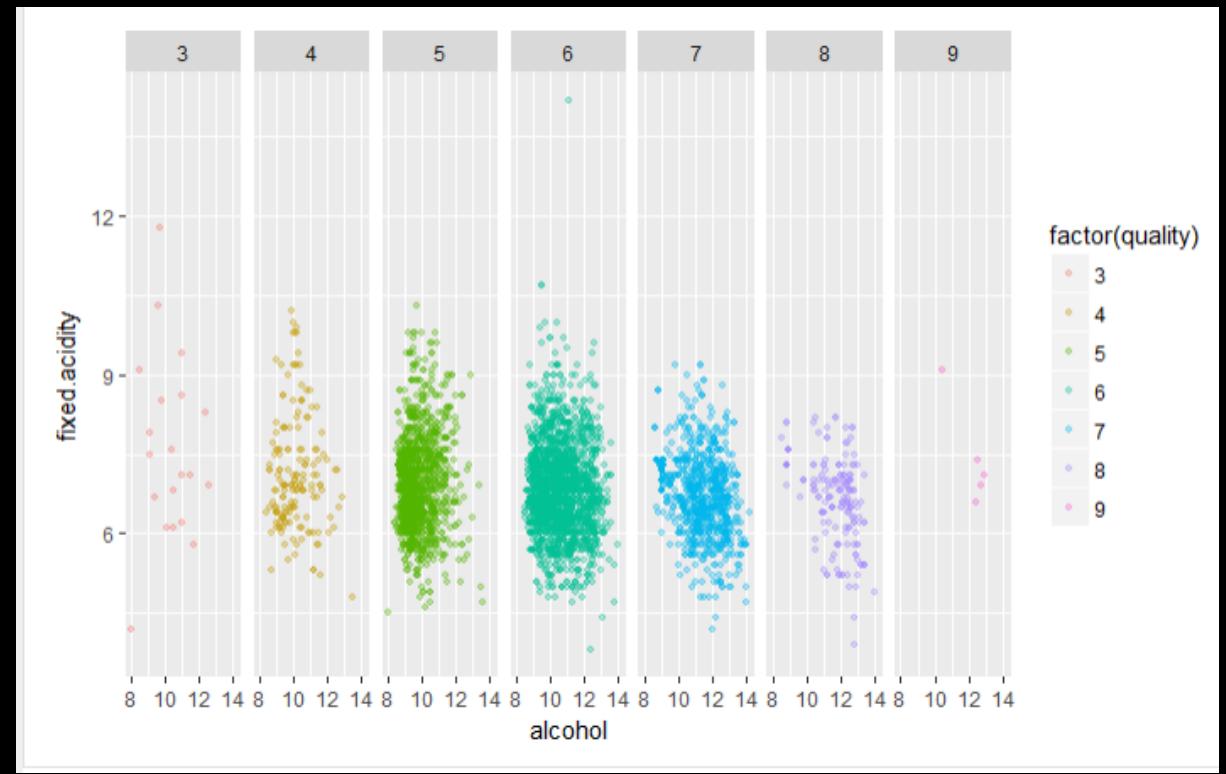
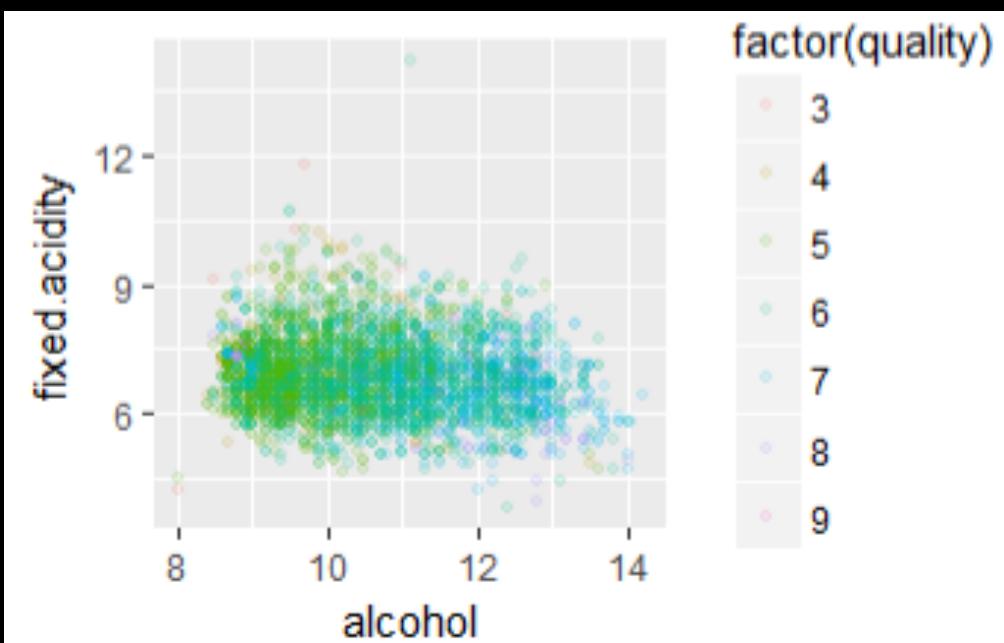


OVERPLOTTING: 3 or more Variables

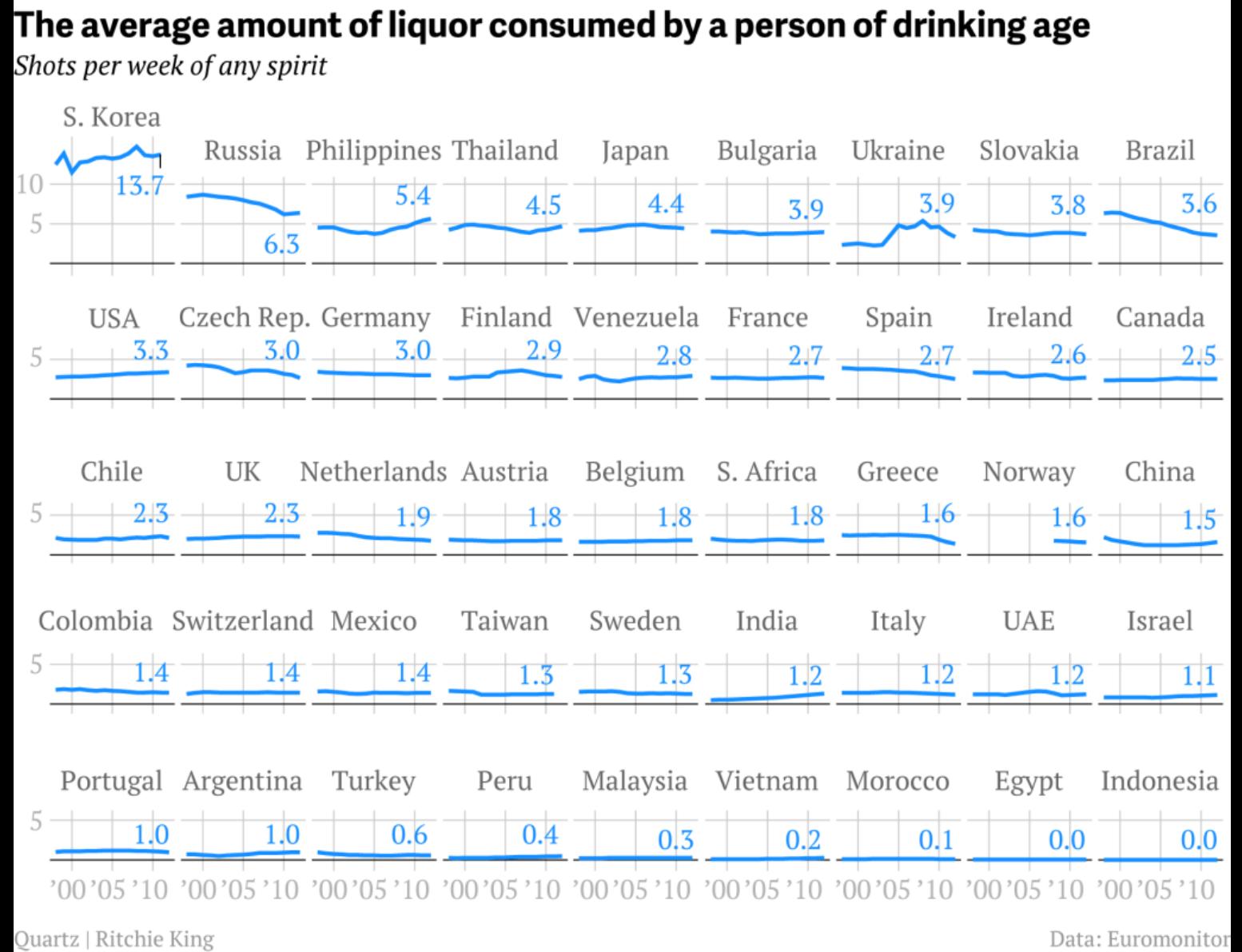


OVERPLOTTING: A SOLUTION

Small Multiples



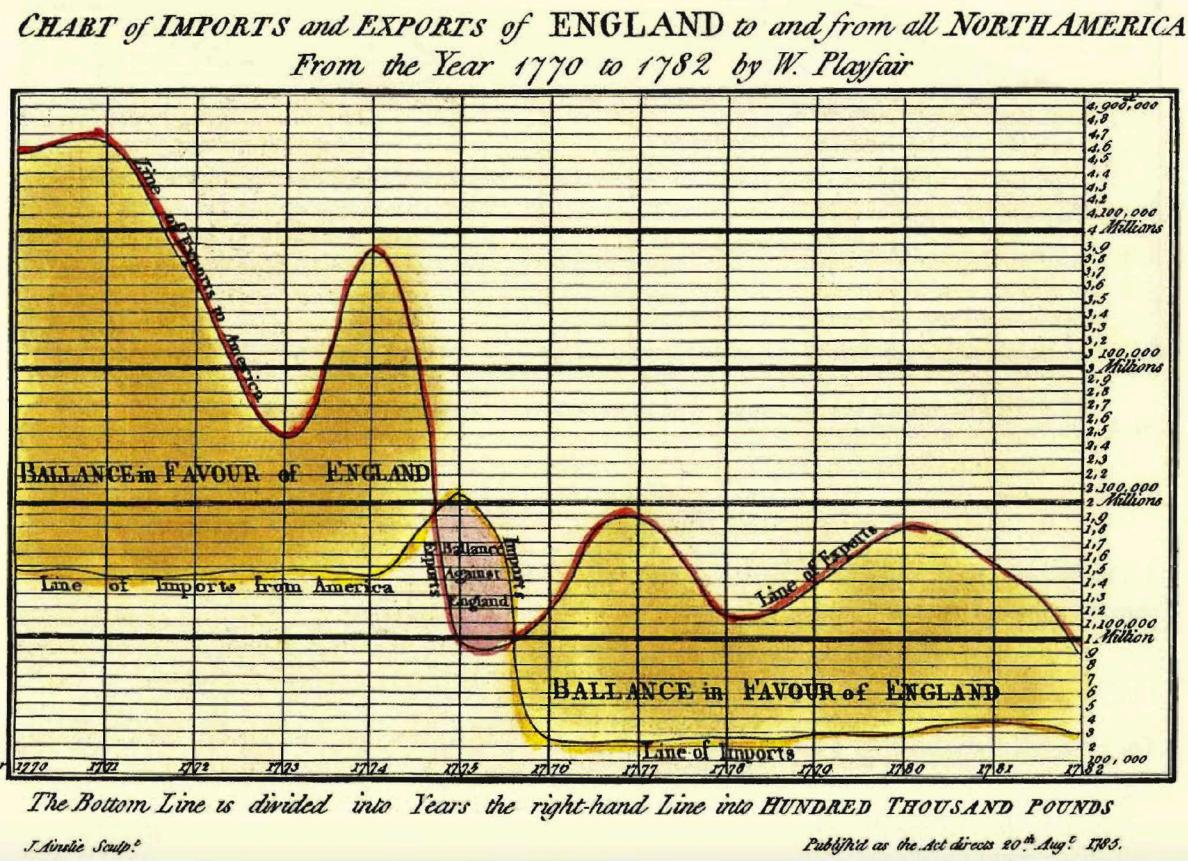
Small Multiples



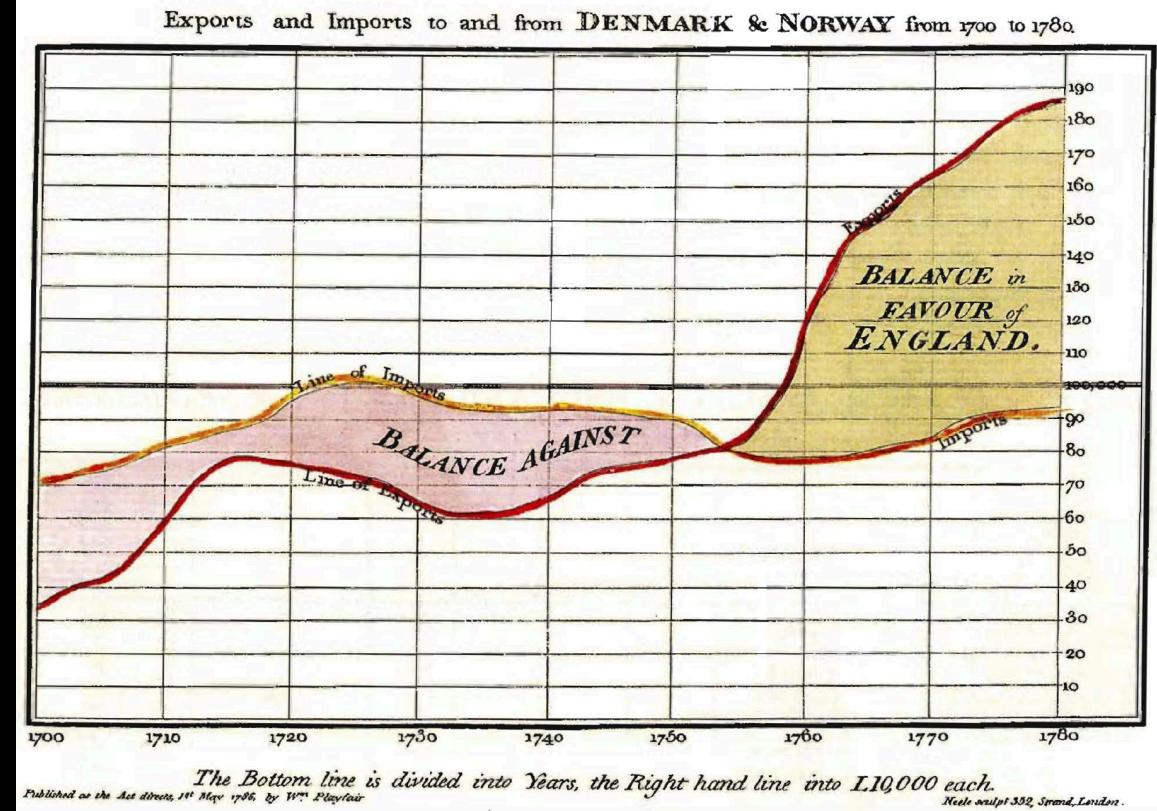
Small Multiples with Tableau

- <https://public.tableau.com/profile/beatriz.woods#/vizhome/PresidentialElectionResults-smallmultiples/Smallmultiples>
- <https://public.tableau.com/s/blog/2016/07/finding-small-stories-ncaa-football-data>

Making Polished Graphics



Playfair 1785



Playfair 1876

Polish: Simplification

- Maximize the data : ink ratio
 - Minimize grids & tick marks
 - Simplify axes
 - Remove all non-essential lines
 - Minimize redundancy without purpose
 - Labels are data - but use sparingly

Polish: ChartJunk

- No 3D.
- Patterns should be used VERY carefully if at all
- If gridlines are necessary, make them light
- Single boxes or no boxes
- Plots on plots cannot be read

Polish: Redundancy

- If the information can be recovered elsewhere, remove it
 - Titles
 - Legends
 - Labels

But don't go crazy