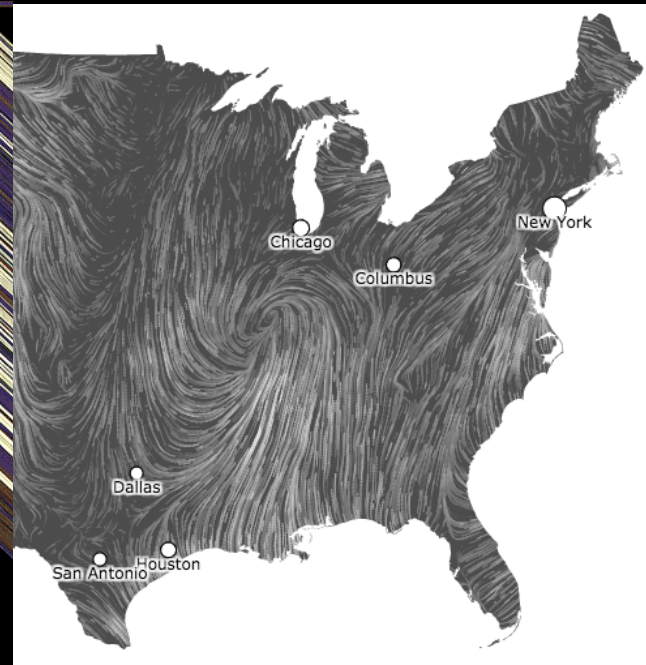
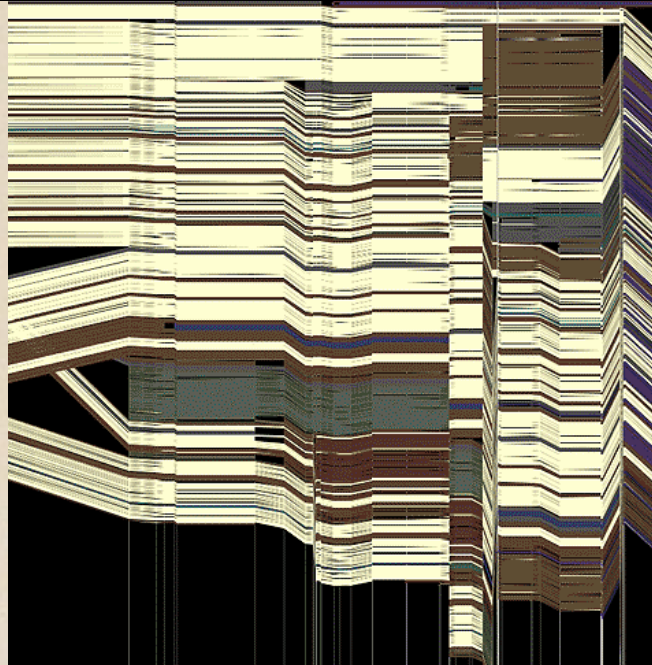
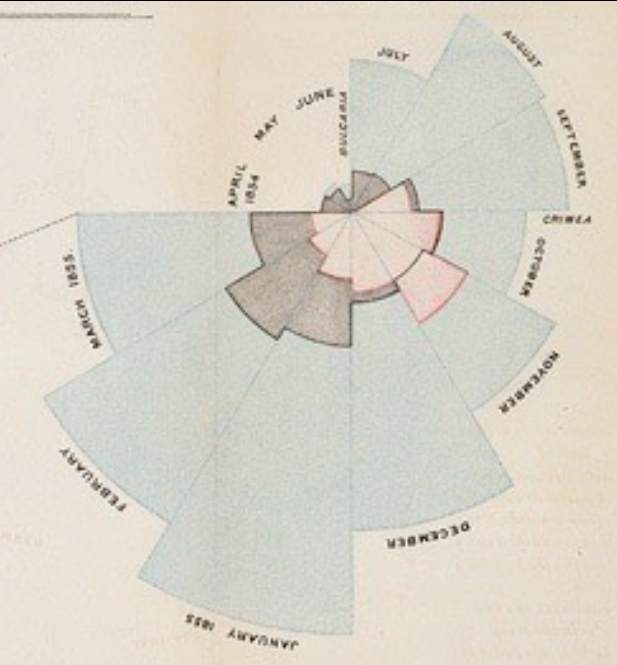


CSE 442 - Data Visualization

Exploratory Data Analysis

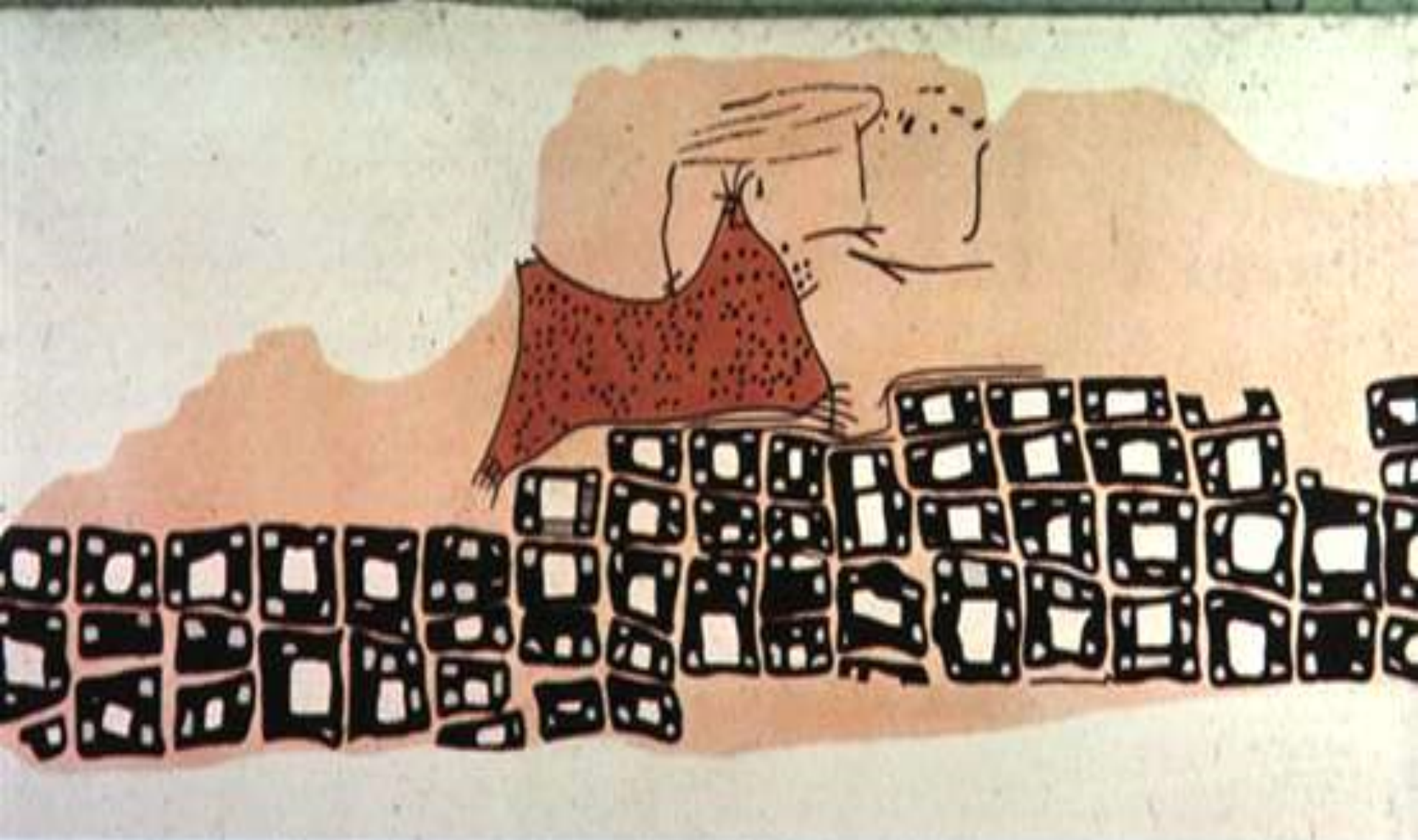


Jeffrey Heer University of Washington

What was the **first**
data visualization?

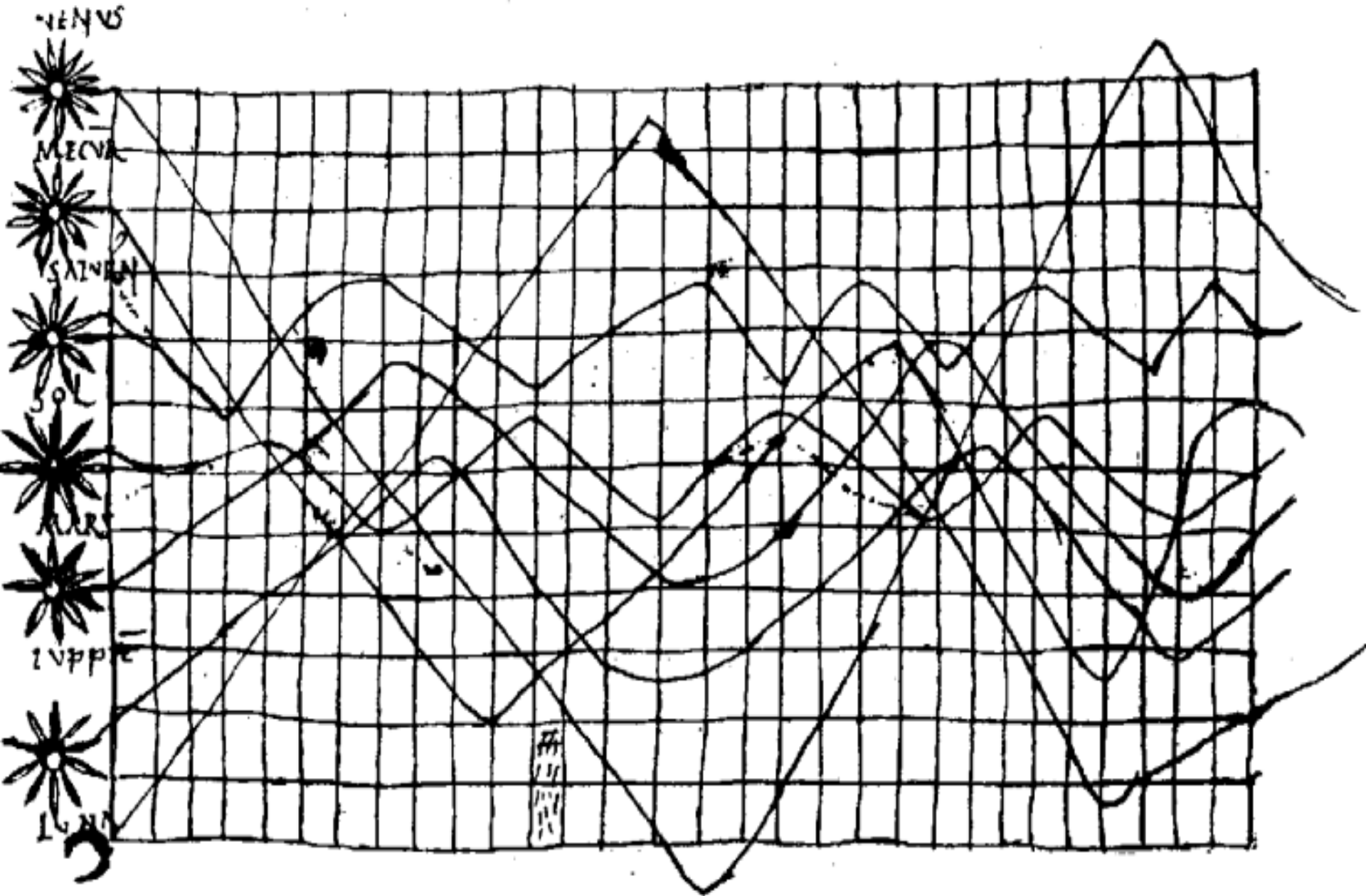
0 BC





~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey

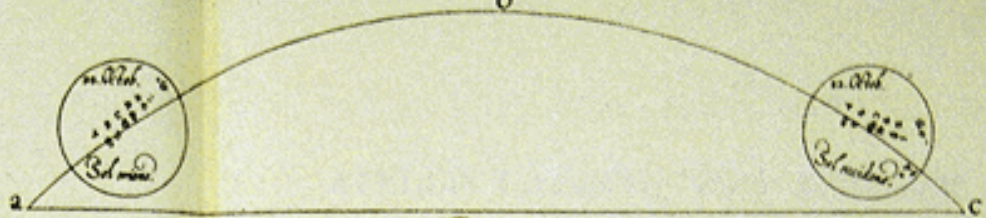
0 BC



~950 AD Position of Sun, Moon and Planets



MACVLAE IN SOLE APPARENTES, OBSERVATAE
 anno 1611. ad latitudinem grad. 48. min. 40.



a c, horizon. a b c, arcus solis diurnus. Sol oriens ex parte a, maculas exhibet quas vides, occidens vero c, easdem ratione primj motus, nonnihil inuertit. Et hanc matutinam vespertinamq; mutationem, omnes maculae quotidie subeunt. Quod semel exhibuisse et mouisse, sufficiat.



Macula M, est
 haec tenus usque
 maxima, nulliq;
 prima magnitudinis
 sideri fixo cedit.

Macula I fuit
 valde conspicua,
 propter notabi-
 lem pra reliquis
 magnitudinem.

Figura qua
 habet sinuatum
 signum X, est
 Omittere.

Alm. New. Ang. m. 1611.

Com. p. 1611. (1611)

TOLEDO.

GRADOS DE LA LONGITUD.



G. Janssonius.

G. Mercator.

J. Schonerus.

P. Lansbergius.

T. Blaez.

L. Regiomontanus.

Oronius.

C. Clavius.

C. Ptolomaeus.

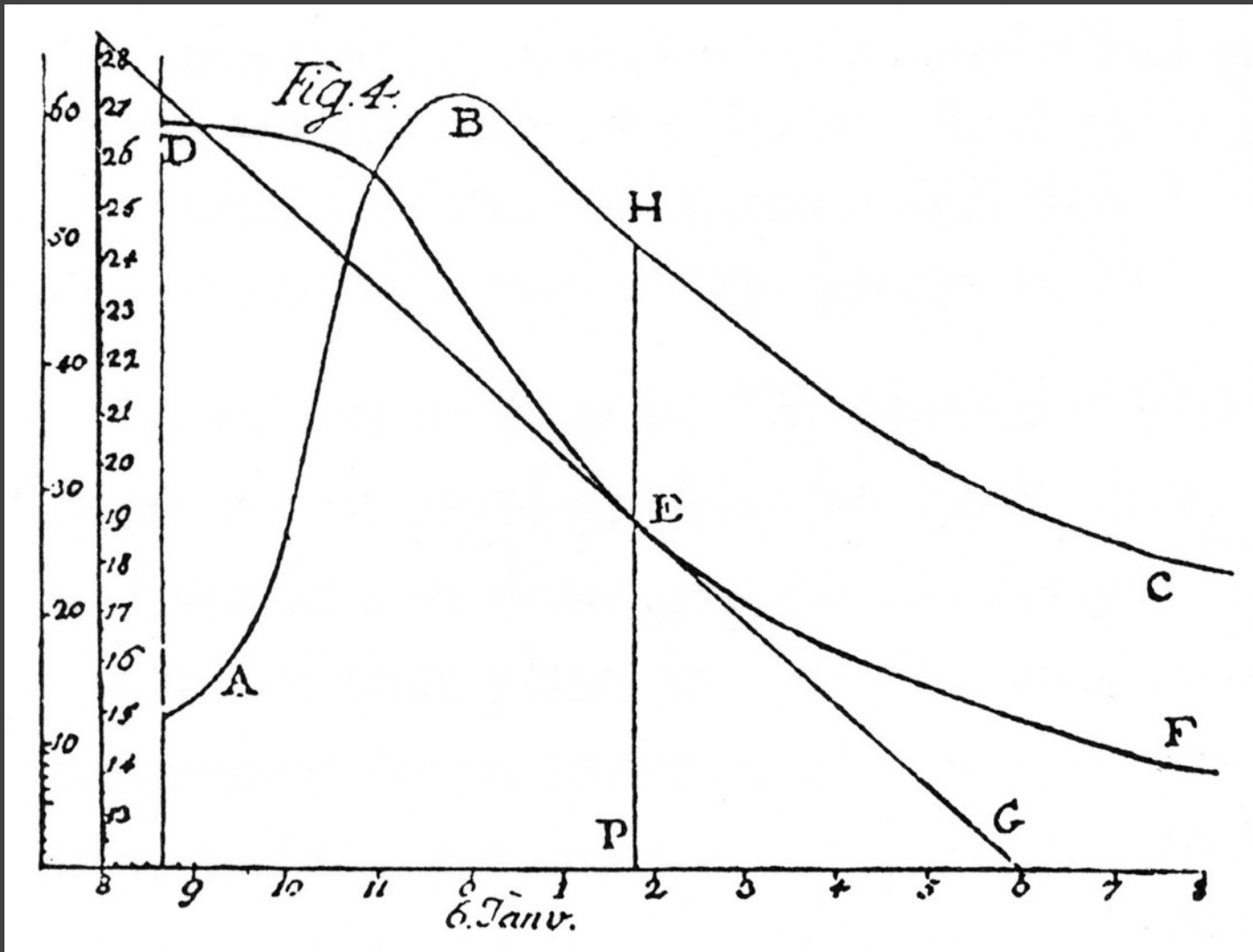
A. Argelius.

A. Maginus.

D. Origanus.

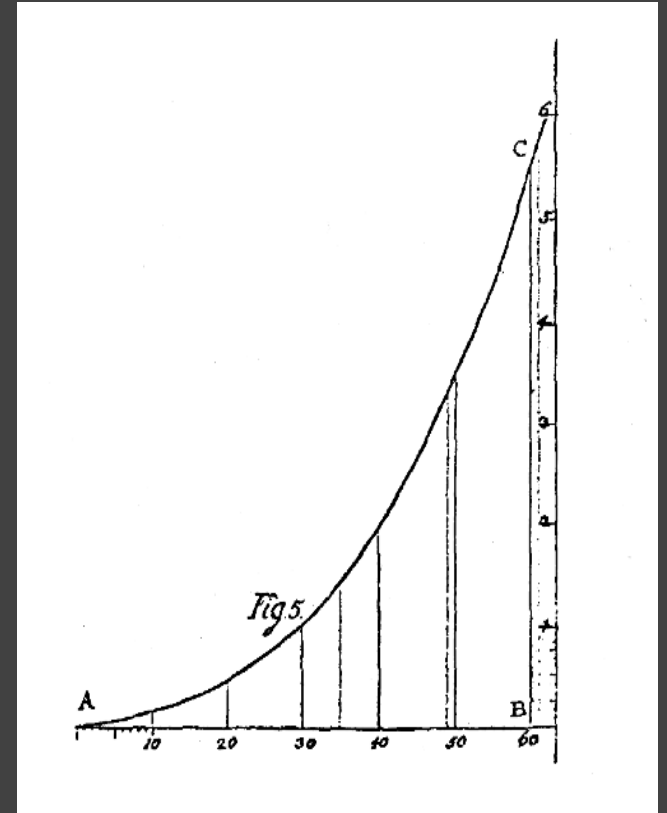
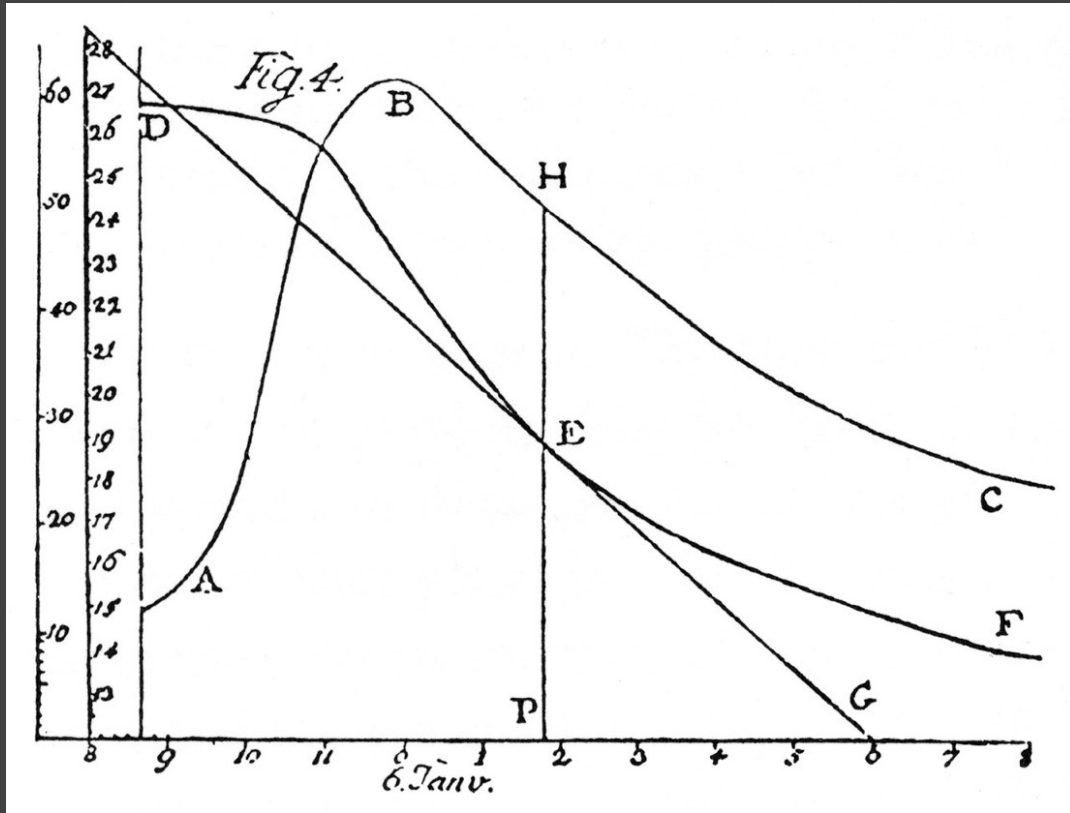
ROMA

Longitudinal distance between Toledo and Rome, van Langren 1644



The Rate of Water Evaporation, Lambert 1765





The Rate of Water Evaporation, Lambert 1765

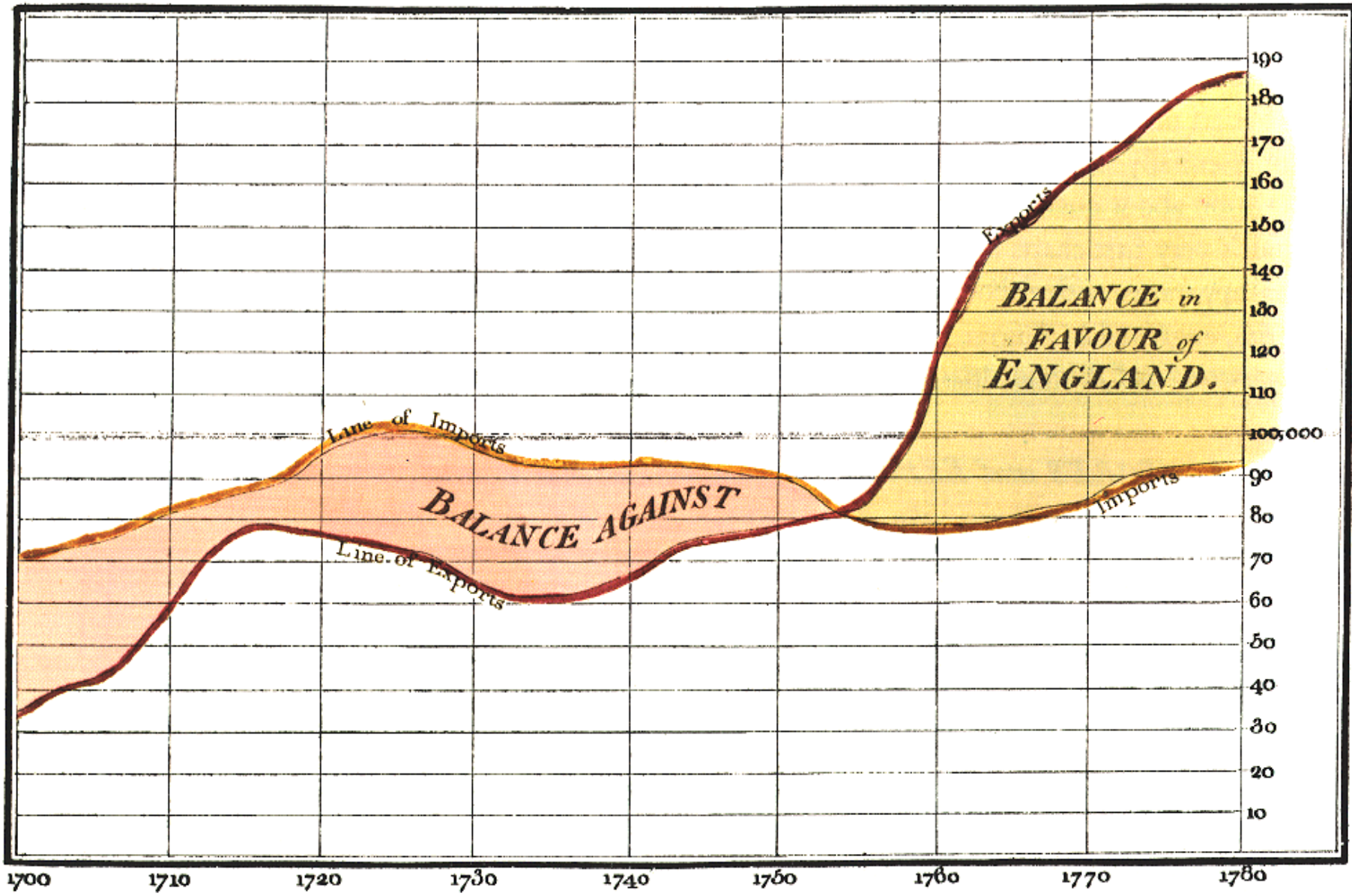


The “**Golden Age**” of Data Visualization

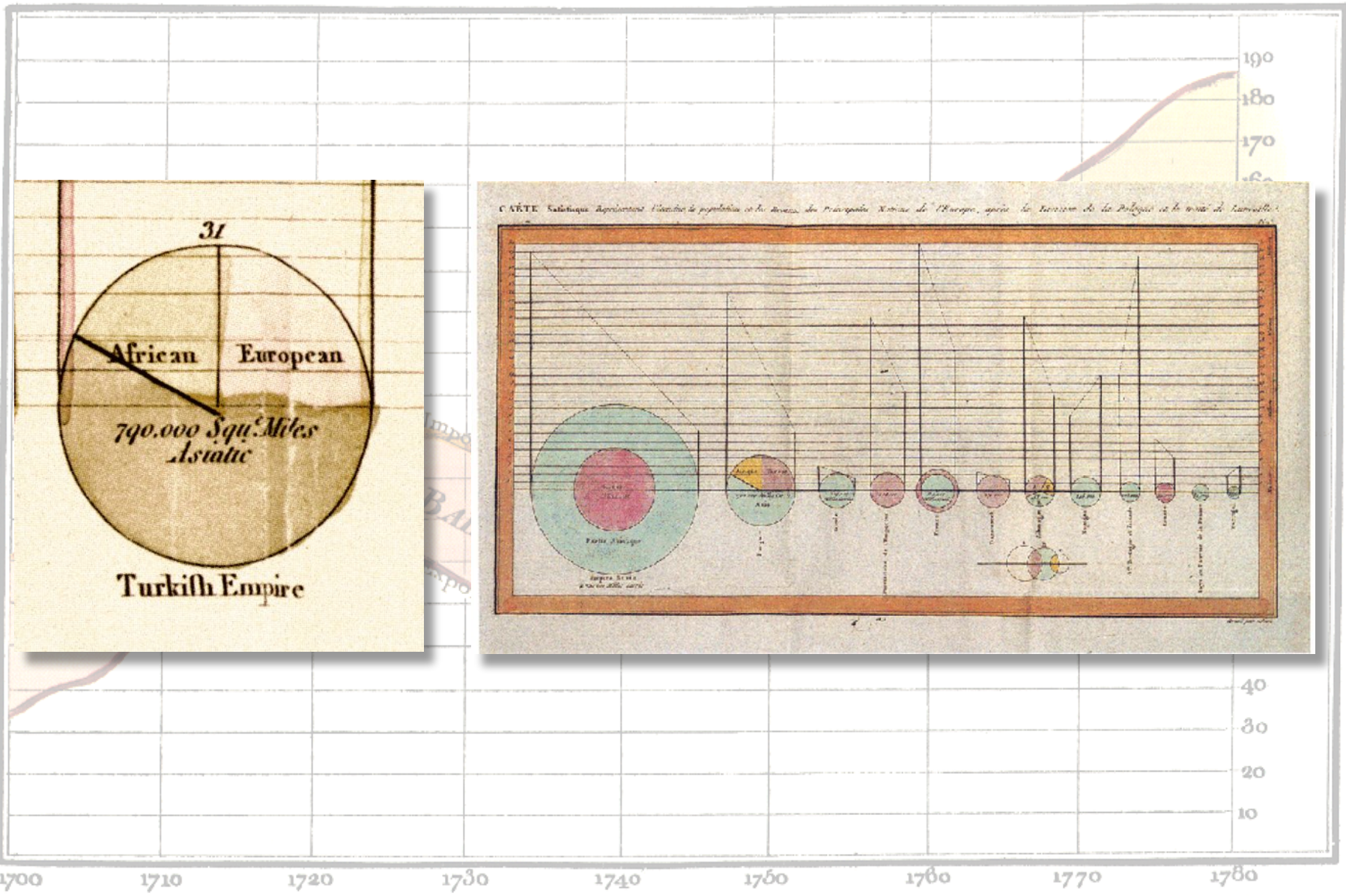
1786 1900

A horizontal white line at the bottom of the slide serves as a timeline. A small vertical tick mark is on the left. A red rectangular segment is positioned on the right side of the line, corresponding to the years 1786 and 1900.

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.

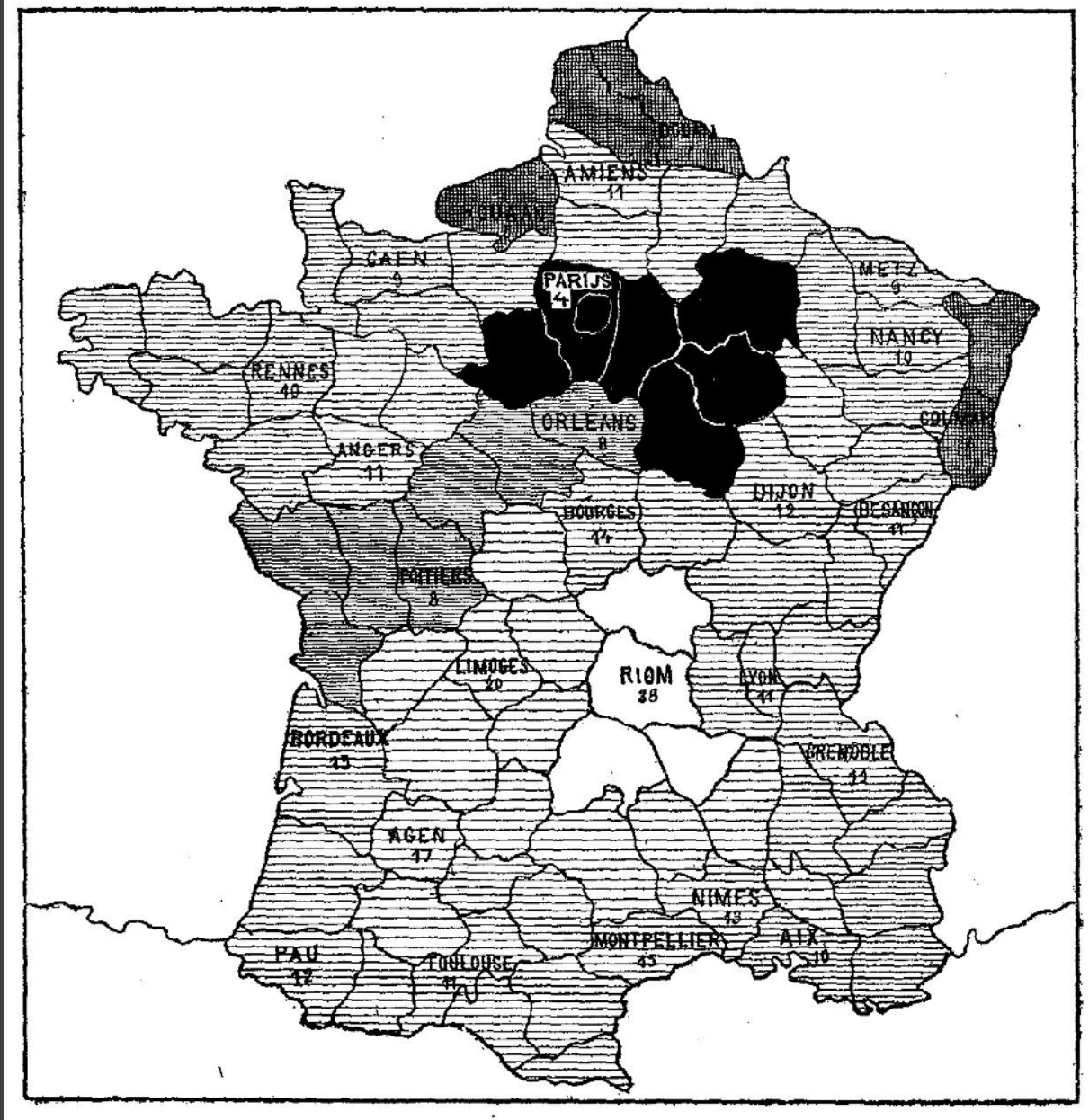


Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



Statistical Breviary, William Playfair 1801



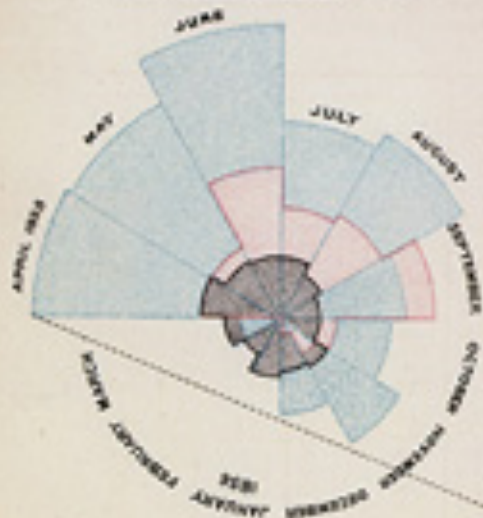


1786 1826(?) Illiteracy in France, Pierre Charles Dupin

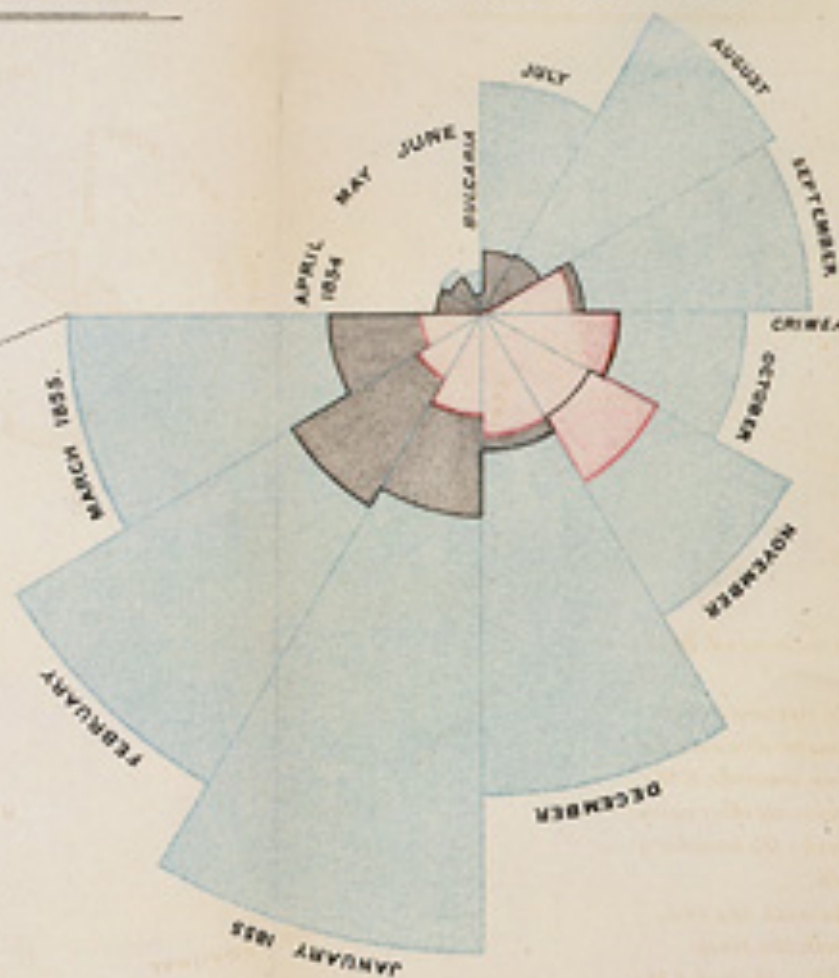


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856



1.
APRIL 1854 TO MARCH 1855

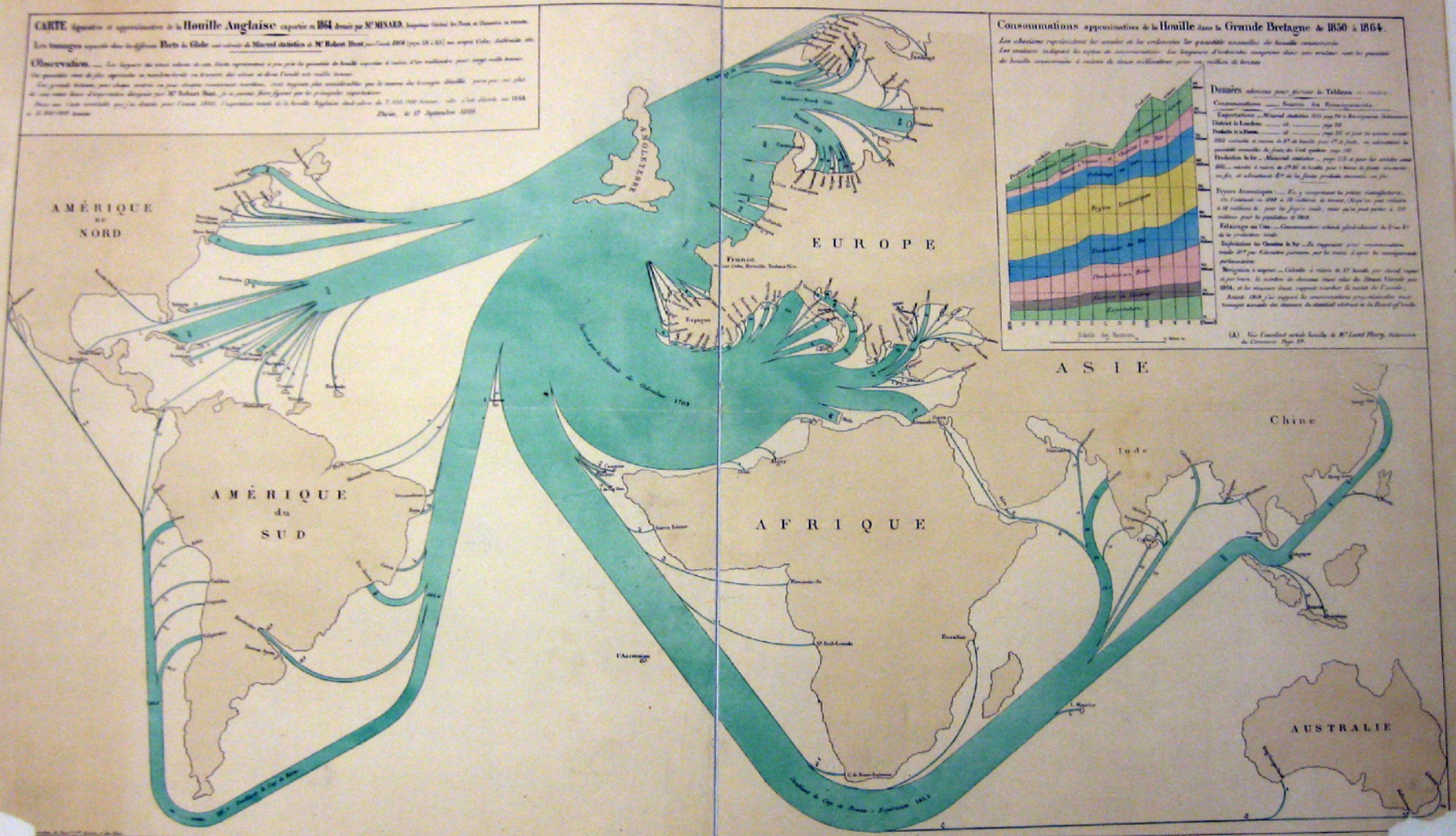
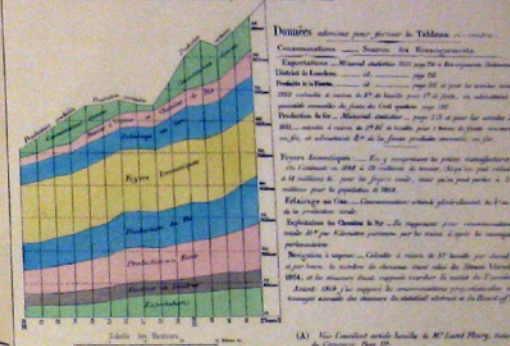


“to affect thro’ the Eyes
what we fail to convey to
the public through their
word-proof ears”



CARTE figurative et approximative de la **houille Anglaise** exportée en 1864 dessinée par M. MINARD, Ingénieur civil des Ponts et Chaussées à Paris.
 Les tracés expriment dans les lignes l'Écart et le volume de Marchés établis à St Robert (Brest) par l'année 1864 (page 18 & 19) au moyen d'un Code, détaillé en
Observation — Les lignes de même couleur de cette Carte représentent à peu près la quantité de houille exportée et celle de son transport pour chaque année.
 Une grande ligne de plus épaisseur ou marquée de traits en travers des autres se dit l'année elle-même.
 Les grands tracés, pour chaque année, ou pour chaque mouvement particulier, sont toujours plus ou moins épais, pour ou en plus
 de son volume dans l'année désignée par M. Robert (Brest) ou le premier jour figuré par les principales importations.
 Dessiné par l'auteur en 1864 pour l'année 1864. L'exportation totale de la houille Anglaise était alors de 7 100 000 tonnes. Elle l'est de 10 000 en 1864
 et de 10 000 000 tonnes.
 Paris, le 27 Septembre 1864.

Consommations approximatives de la Houille dans la Grande Bretagne & 1850 & 1864.
 Les chiffres représentent les années et les couleurs les quantités annuelles de houille consommées.
 Les couleurs indiquent les types de consommation. Les longueurs d'indicateurs expriment dans une échelle non les quantités
 de houille consommées à raison de deux millions pour un million de tonnes.



1786

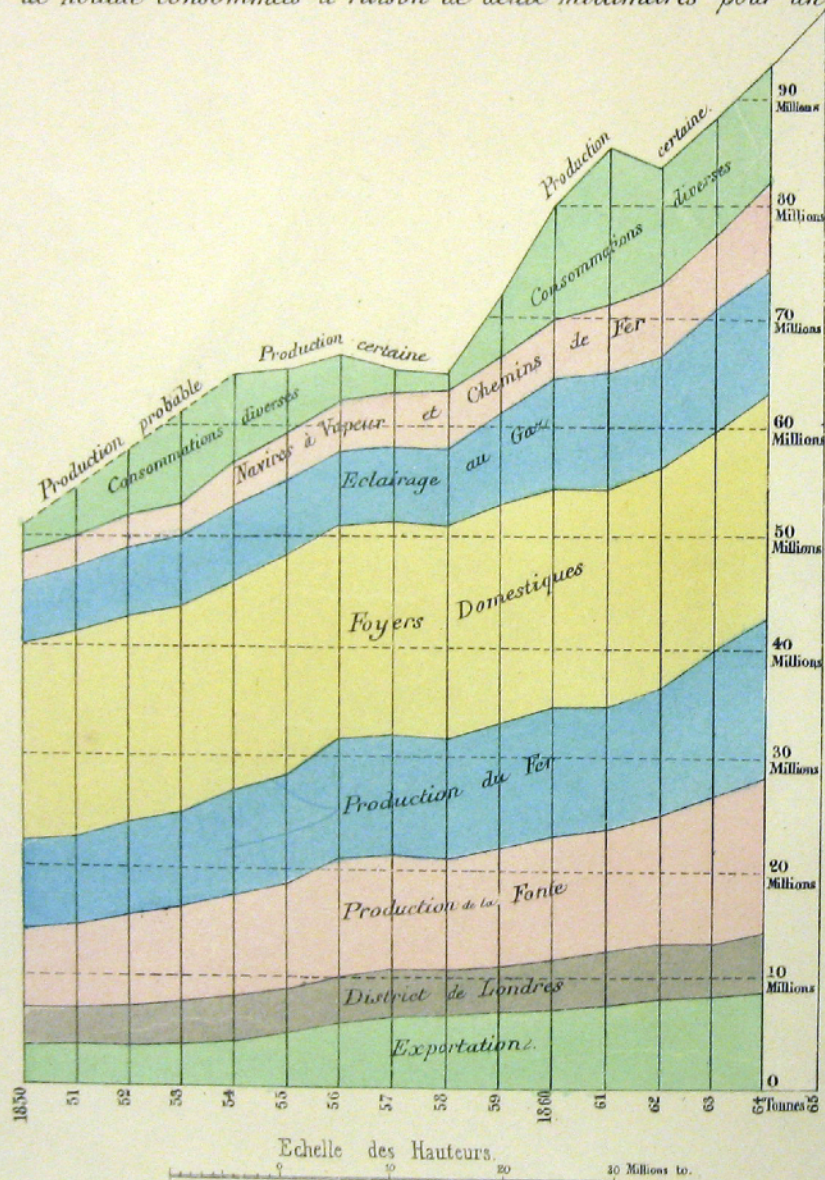
1864 British Coal Exports, Charles Minard



Consommations approximatives de la Houille dans la Grande Bretagne de 1850 à 1864.

Les abscisses représentent les années et les ordonnées les quantités annuelles de houille consommée.

Les couleurs indiquent les espèces de consommations. Les longueurs d'ordonnées comprises dans une couleur sont les quantités de houille consommées à raison de deux millimètres pour un million de tonnes.



Données admises pour former le Tableau ci-contre.

Consommations. — Sources des Renseignements.

Exportations. — *Mineral statistics 1865 page 214 et Renseignements Parlementaires.*

District de Londres. — *id.* — page 213

Produits de la Fonte. — *id.* — page 215 et pour les années avant 1855 calculée à raison de 3^{to} de houille pour 1^{to} de fonte, en admettant les quantités annuelles de fonte du Coal question page 192.

Production du fer — *Mineral statistics* — page 215 et pour les années avant 1855 — calculée à raison de 3^{to} 35 de houille pour 1 tonne de fonte convertie en fer, et admettant $\frac{2}{10}$ ^{es} de la fonte produite convertis en fer.

Foyers domestiques. — En y comprenant les petites manufactures.

On l'estimait en 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.

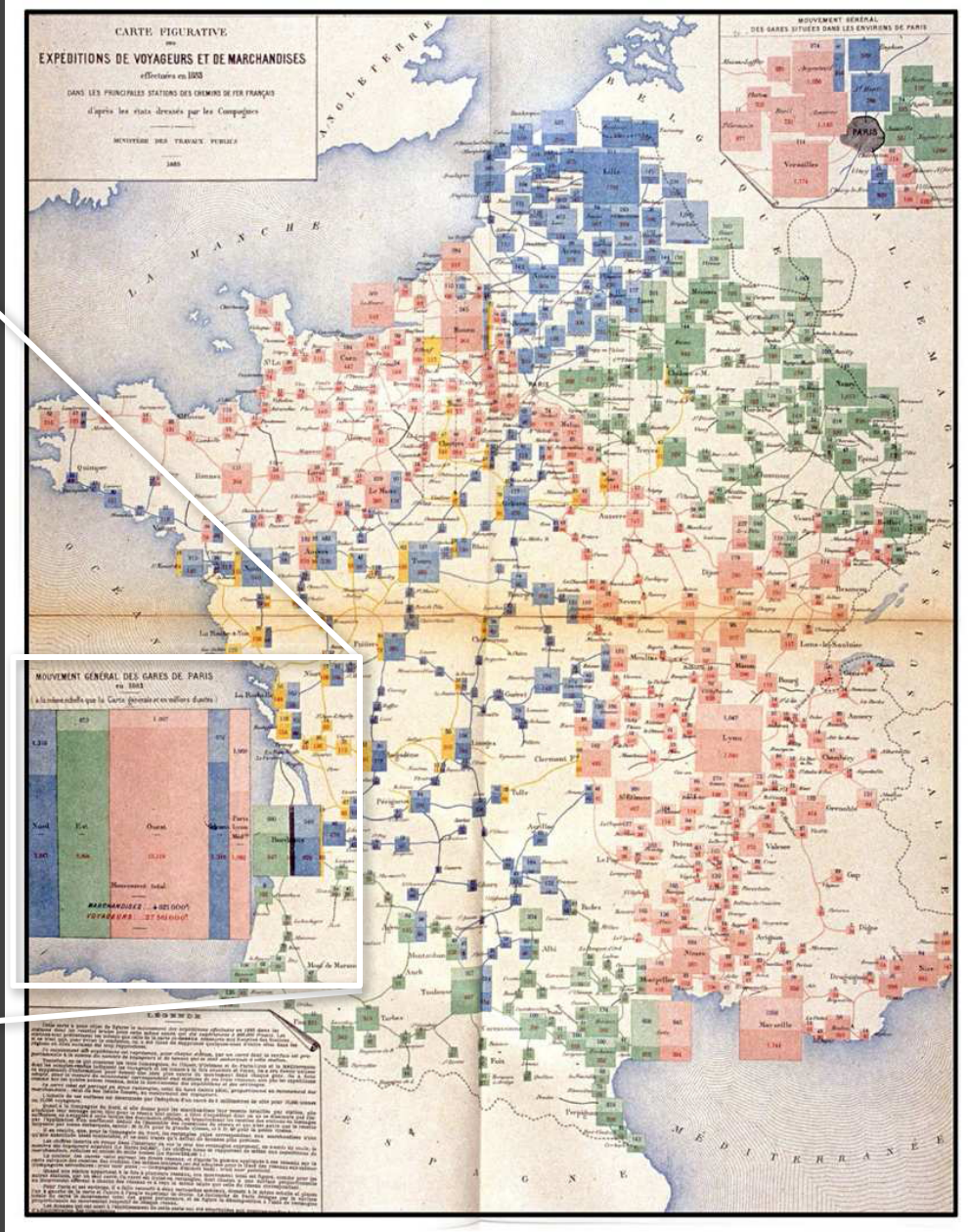
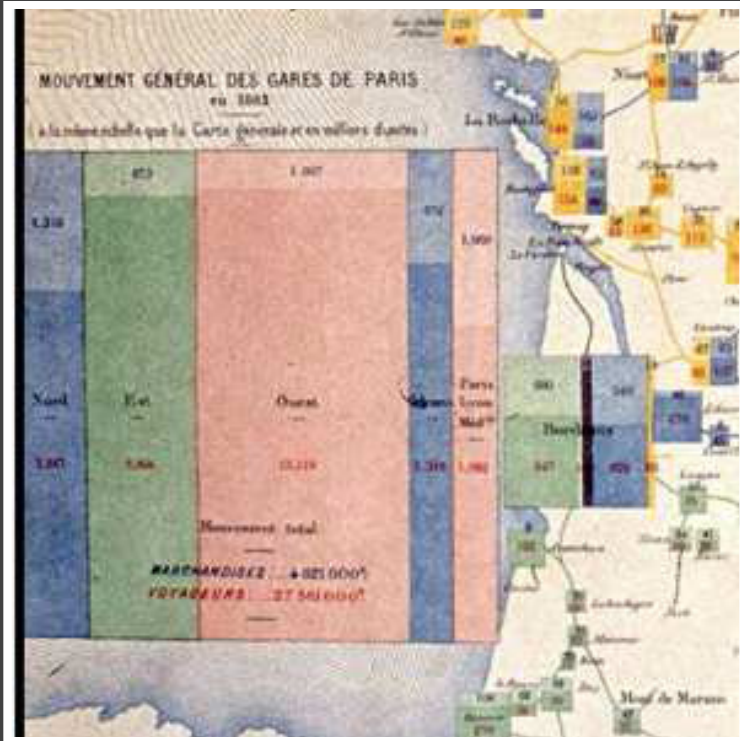
Eclairage au Gaz. — Consommation estimée généralement du $\frac{1}{3}$ ^o au $\frac{1}{8}$ ^o de la production totale.

Exploitation des Chemins de Fer. — En supposant pour consommation totale 10^{es} par Kilomètre parcouru par les trains d'après les renseignements parlementaires.

Navigation à vapeur. — Calculée à raison de 5^{es} houille par cheval vapeur et par heure, le nombre de chevaux étant celui du Steam Vessels pour 1864, et les steamers étant supposés marcher la moitié de l'année;

Avant 1864 j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers du statistical abstract et du Board of trade.

(A) Voir l'excellent article houille de M.^r Lamé Fleury, Dictionnaire du Commerce Page III.



1786

1884 Rail Passengers and Freight from Paris

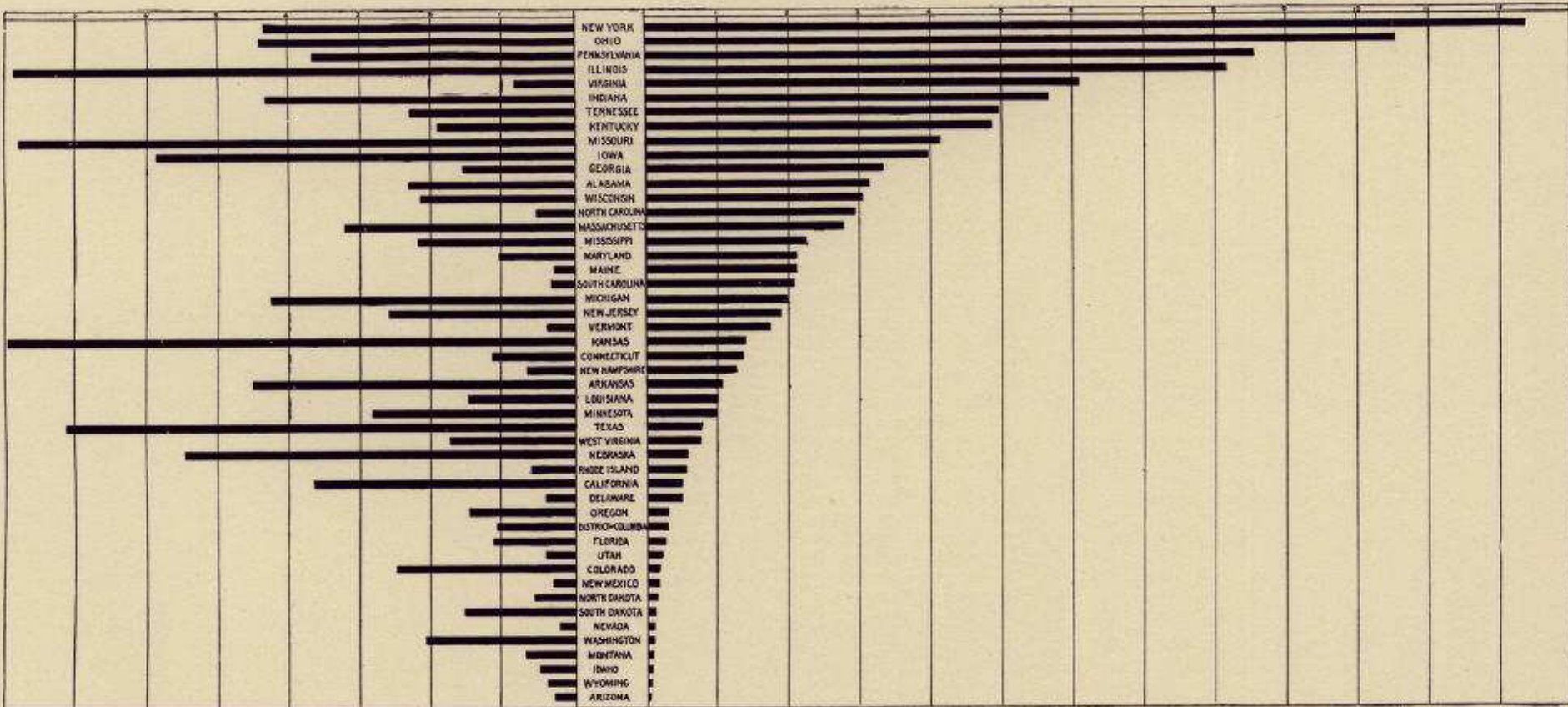


66. INTERSTATE MIGRATION—NUMBER OF NATIVE IMMIGRANTS AND NATIVE EMIGRANTS, BY STATES AND TERRITORIES: 1890.

Native immigrants.

[Hundreds of thousands.]

Native emigrants.



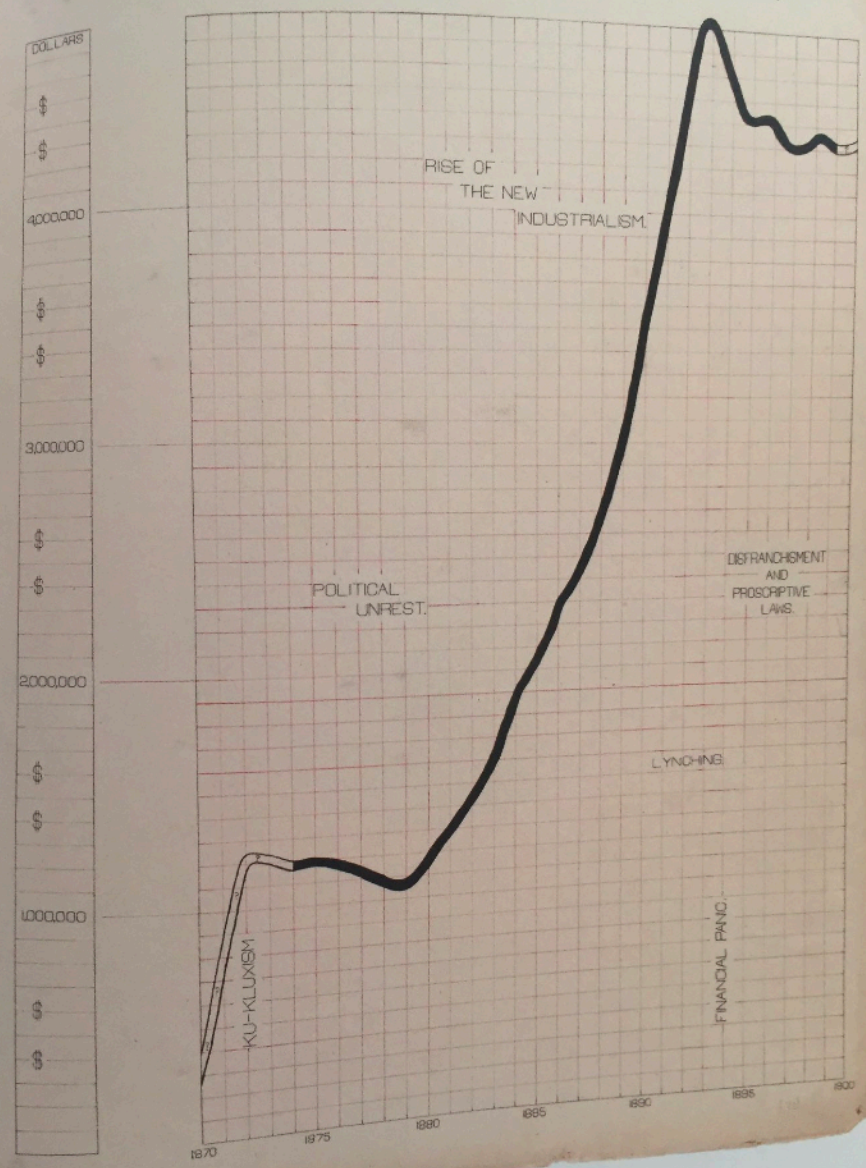
Negro business men in the United States.

Nègres Américains dans les affaires.

Done by Atlanta University.



VALUATION OF TOWN AND CITY PROPERTY OWNED BY GEORGIA NEGROES.



1786

1900 Visualizing Black America, W. E. B. DuBois et al.



The Rise of Statistics

1786



1900



1950

Rise of **formal statistical methods** in the physical and social sciences

Little innovation in graphical methods

A period of **application and popularization**

Graphical methods enter textbooks, curricula, and **mainstream use**

1786

1900

1950





LIFE

1786

Data Analysis & Statistics, Tukey 1962





Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.



The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.

LIFE



While some of the influences of statistical theory on data analysis have been helpful, others have not.

LIFE



Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality** and **flexibility** appropriate to the **exploratory character of exposure** can be fitted into any of the structures of formal statistics so far proposed.



Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human mind.**

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention.**

LIFE

Set A

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set B

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set C

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set D

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Summary Statistics

$$u_X = 9.0 \quad \sigma_X = 3.317$$

$$u_Y = 7.5 \quad \sigma_Y = 2.03$$

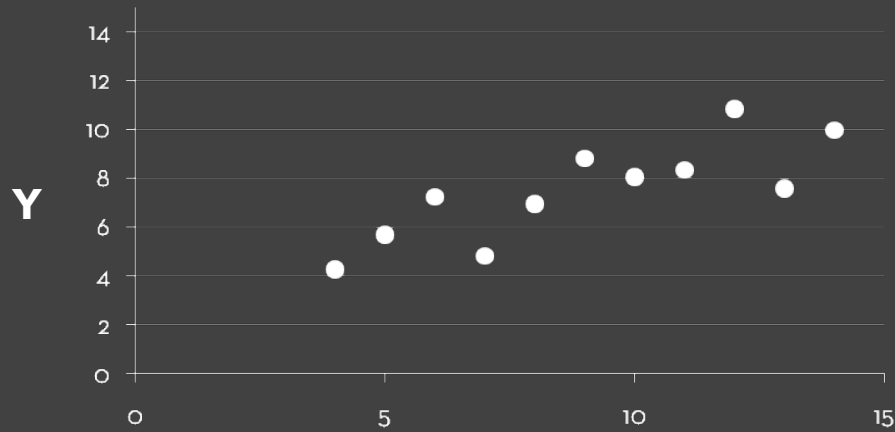
Linear Regression

$$Y = 3 + 0.5 X$$

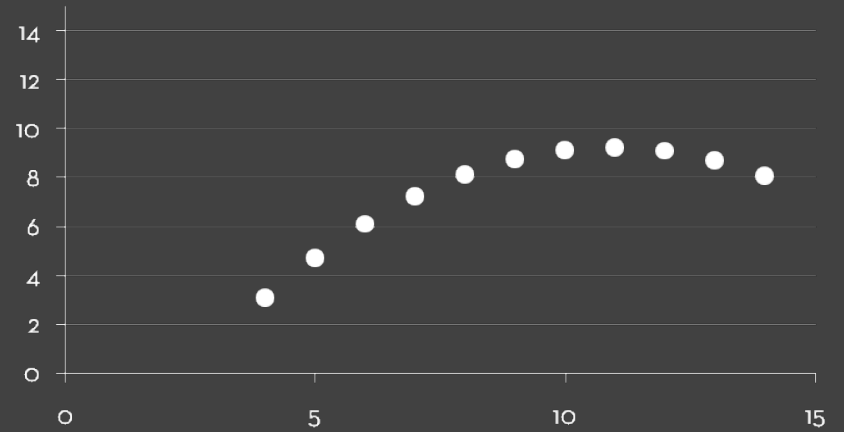
$$R^2 = 0.67$$

[Anscombe 1973]

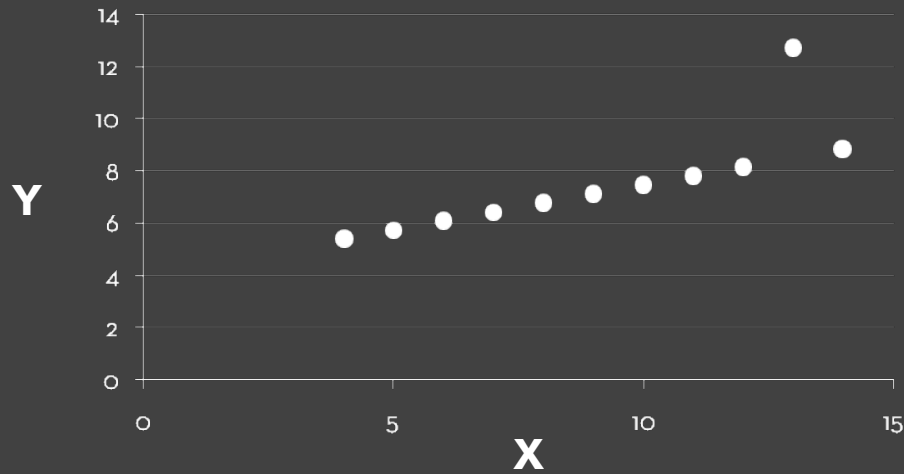
Set A



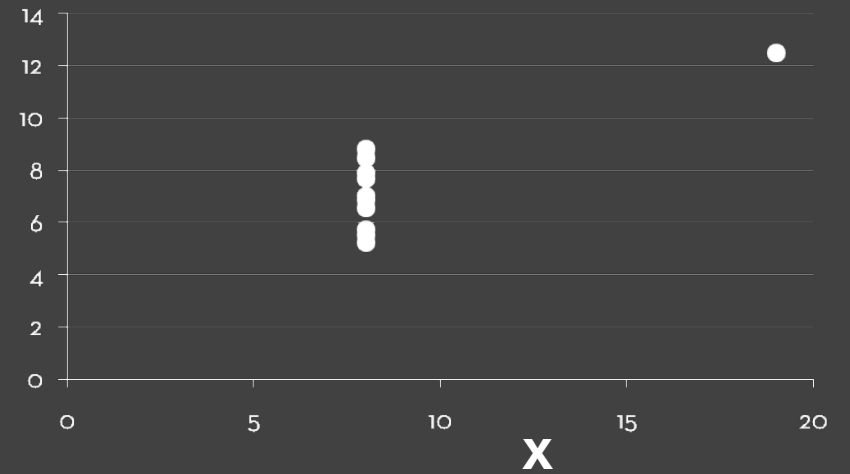
Set B



Set C



Set D



Topics

Exploratory Data Analysis

Data Wrangling

Exploratory Analysis Examples

Tableau / Polaris

Data Wrangling

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all.

Anonymous Data Scientist

[Kandel et al. '12]





**Big Data
Borat**

@BigDataBorat



Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.



Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

DataWrangler

The screenshot displays the DataWrangler interface. On the left, there are two panels: 'Suggestions' and 'Script'. The 'Suggestions' panel lists four actions: 'Delete rows 8,10', 'Delete empty rows', 'Delete rows where Property_crime_rate is null', and 'Delete rows where Year is null'. The 'Script' panel has an 'Export' button and two expandable items: 'Split data repeatedly on newline into rows' and 'Split data repeatedly on \','. On the right, a data table is shown with 408 rows. The table has two columns: '# Year' and '# Property_crime_rate'. The data is grouped into sections: 'Reported crime in Alabama' (rows 1-7), 'Reported crime in Alaska' (rows 9-14), and an empty row (row 8). The table is currently displaying rows 1 through 14.

#	Year	#	Property_crime_rate
1	Reported crime in Alabama		
2			
3	2004	4	4029.3
4	2005	5	3900
5	2006	6	3937
6	2007	7	3974.9
7	2008	8	4081.9
8			
9	Reported crime in Alaska		
10			
11	2004	12	3370.9
12	2005	13	3615
13	2006	14	3582
14	2007		3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel et al. *CHI'11*

Data Wrangling

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration.

Approaches include:

Manual manipulation in spreadsheets

Code: arquero (JS), dplyr (R), pandas (Python)

Tableau Prep

Open Refine

Tidy Data [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In “tidy” data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.

Our pivoted table variant was not “tidy”!

(This is a variant of normalized forms in DB theory)

Data Quality

“The first sign that a visualization is good is that it shows you a problem in your data...

...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

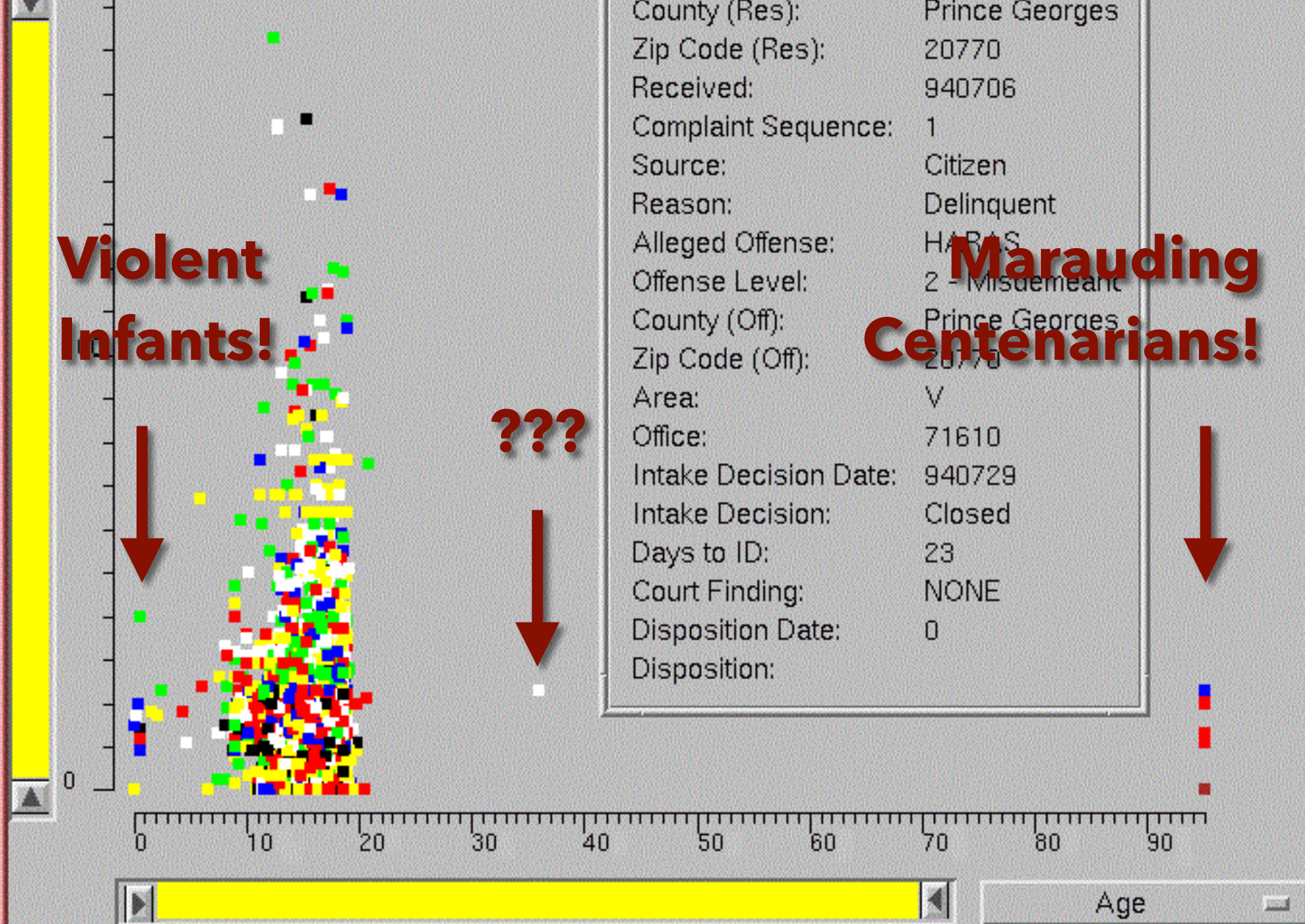
Martin Wattenberg

County (Res):	Prince Georges
Zip Code (Res):	20770
Received:	940706
Complaint Sequence:	1
Source:	Citizen
Reason:	Delinquent
Alleged Offense:	HARAS
Offense Level:	2 - Misdemeanor
County (Off):	Prince Georges
Zip Code (Off):	20770
Area:	V
Office:	71610
Intake Decision Date:	940729
Intake Decision:	Closed
Days to ID:	23
Court Finding:	NONE
Disposition Date:	0
Disposition:	

**Violent
Infants!**

**Marauding
Centenarians!**

???



Query Result: 4792 out of 4792 (100%)

Graph Viewer

Roll-up by:

All

Visualization:

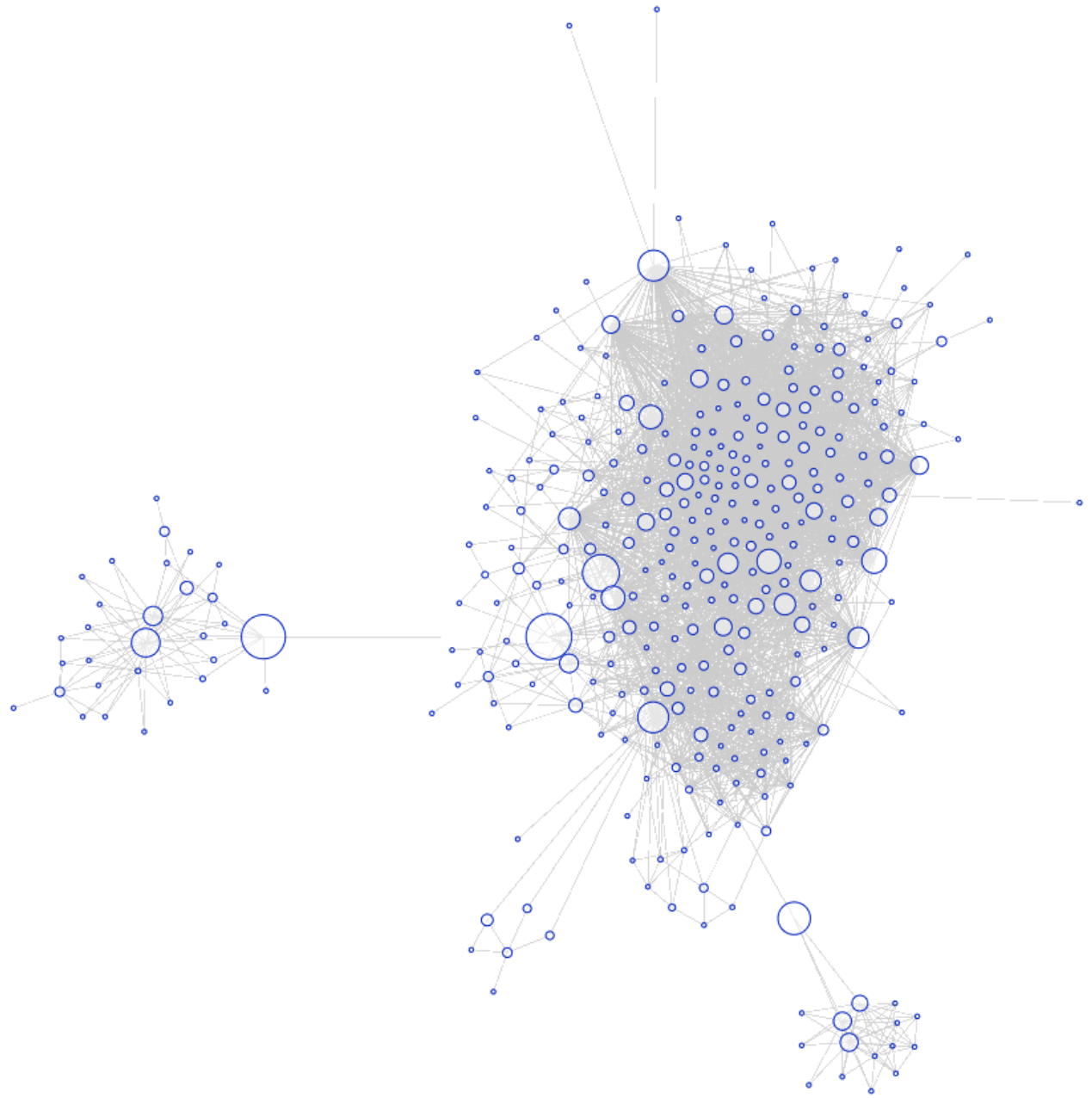
Node-Link

Sort by:

None

Edge centrality filters:

Two horizontal sliders for edge centrality filters, both currently at the minimum position.



- Images
- Animate

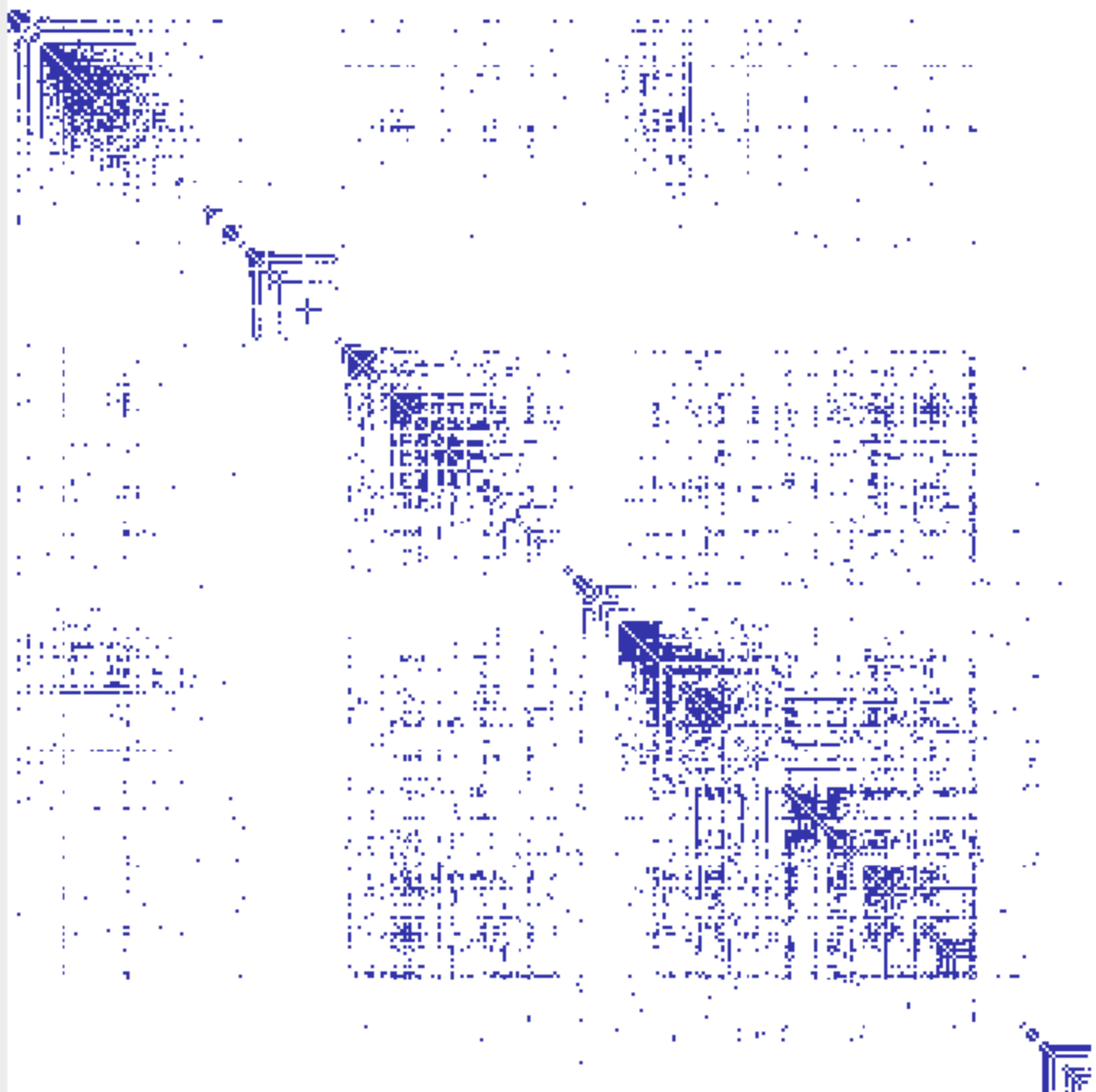
Graph Viewer

Roll-up by:

Visualization:

Sort by:

Edge centrality filters:



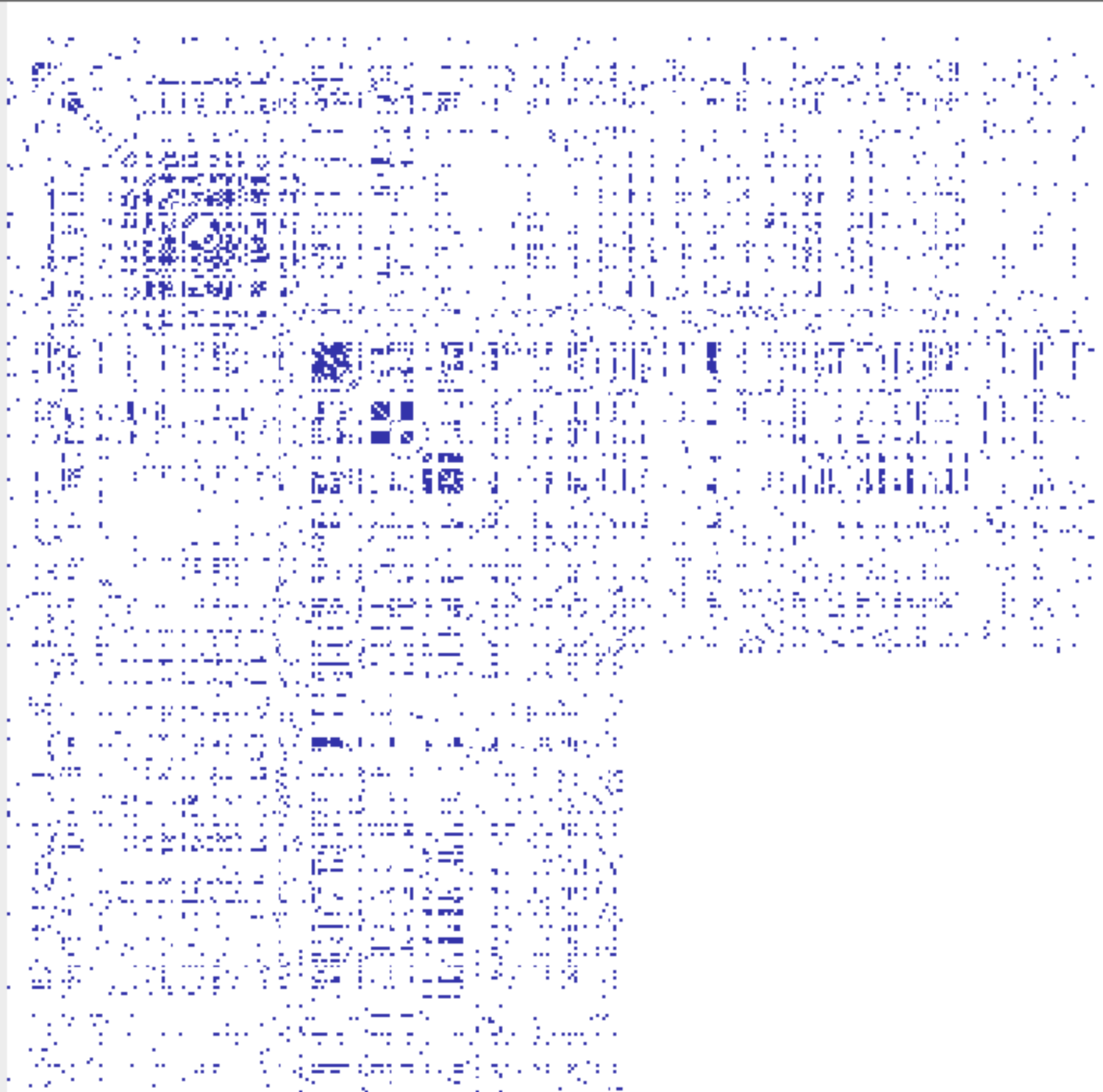
Graph Viewer

Roll-up by:

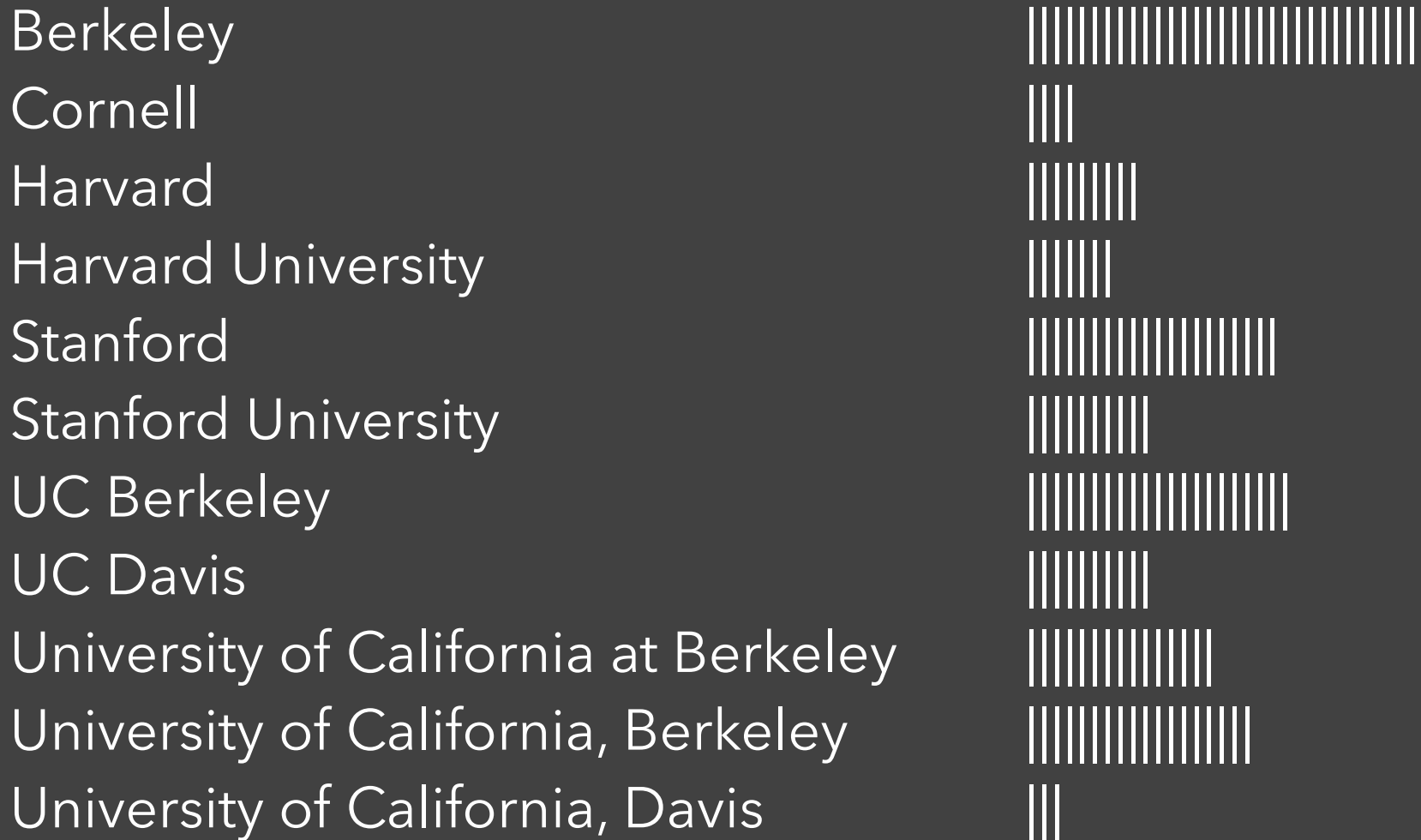
Visualization:

Sort by:

Edge centrality filters:



Visualize Friends by School?



Data Quality Hurdles

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

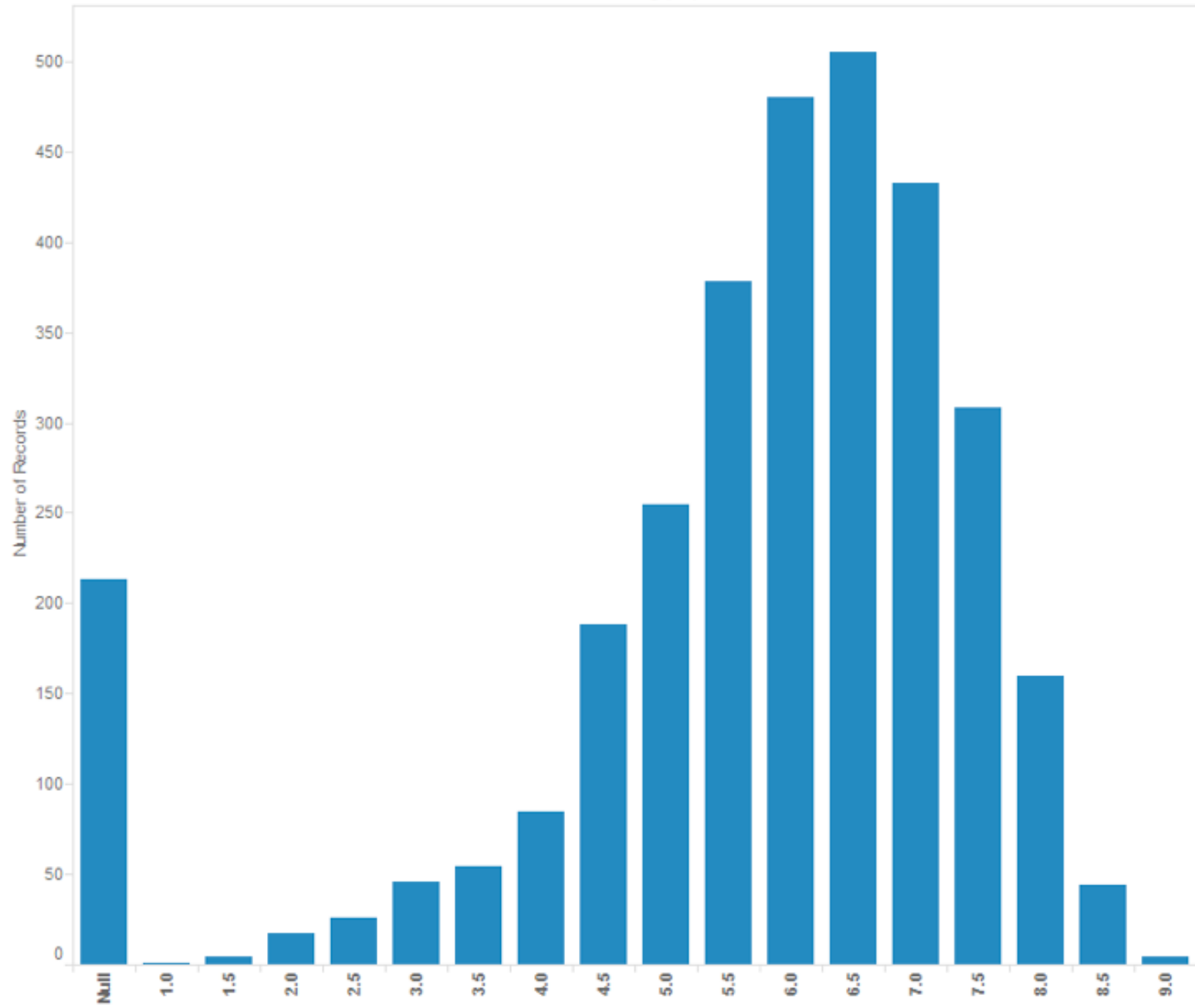
LESSON: Anticipate problems with your data.
Many research problems around these issues!

Analysis Example: Motion Pictures Data

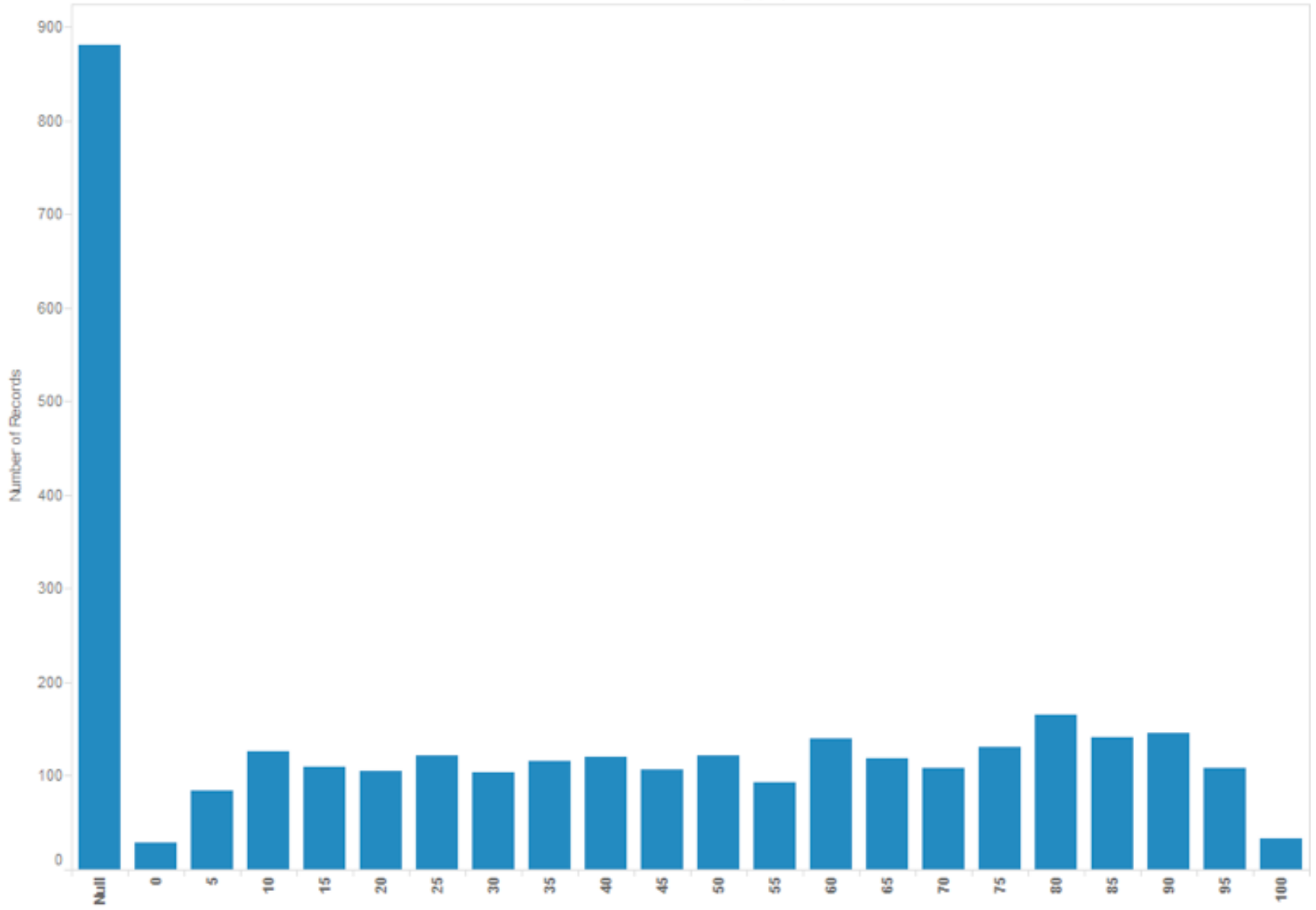
Motion Pictures Data

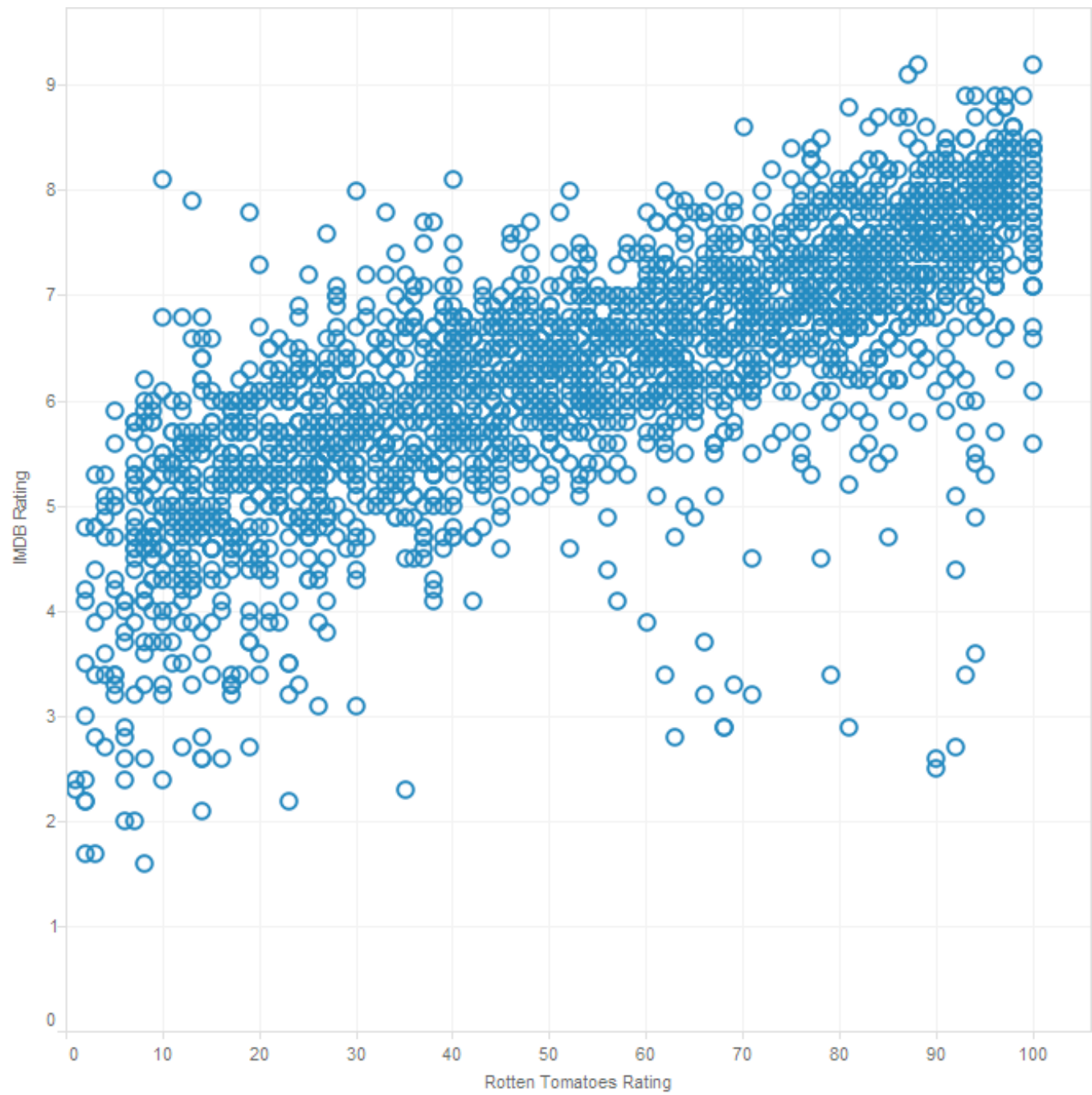
Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)

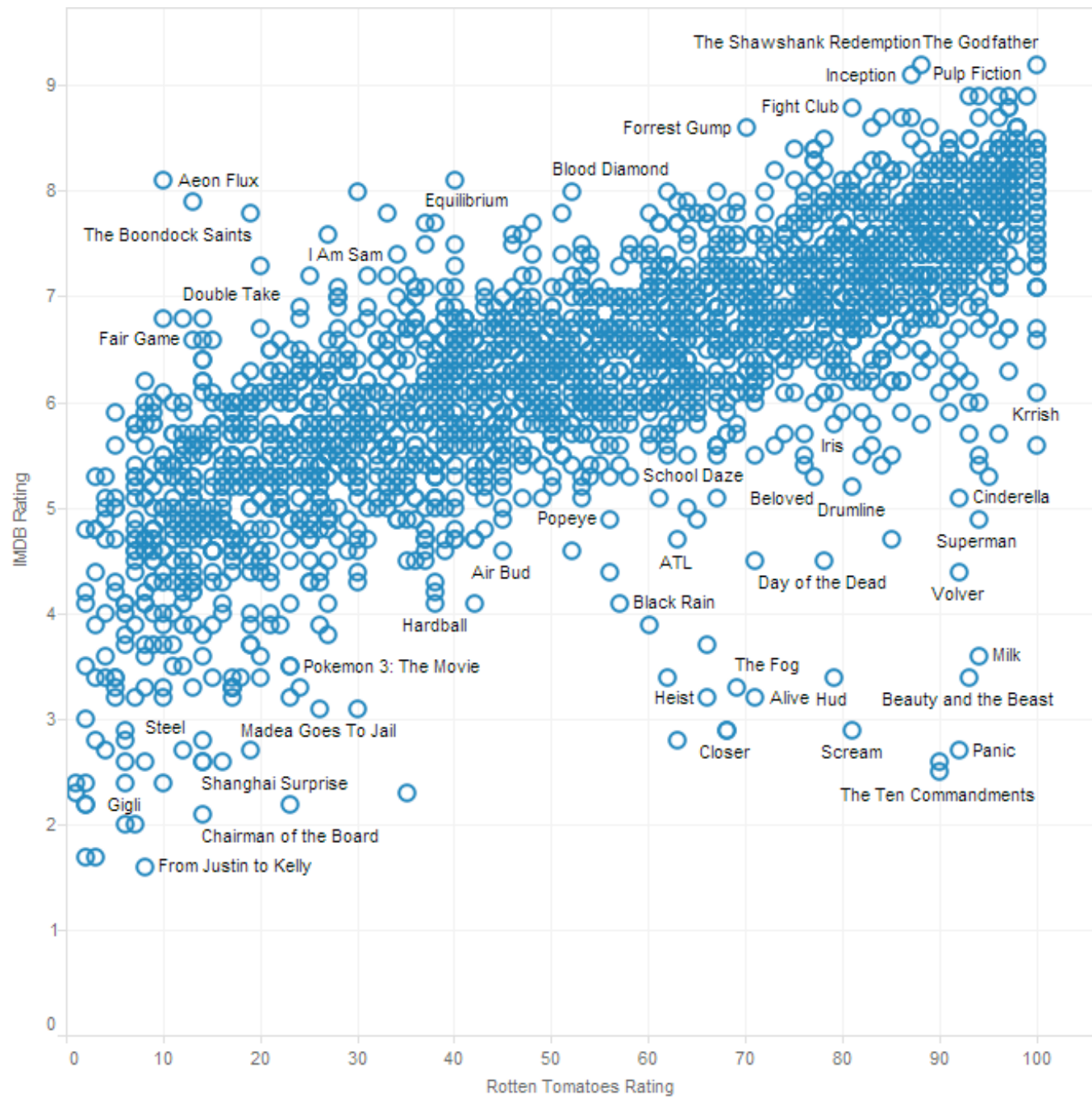
IMDB Rating (bin)

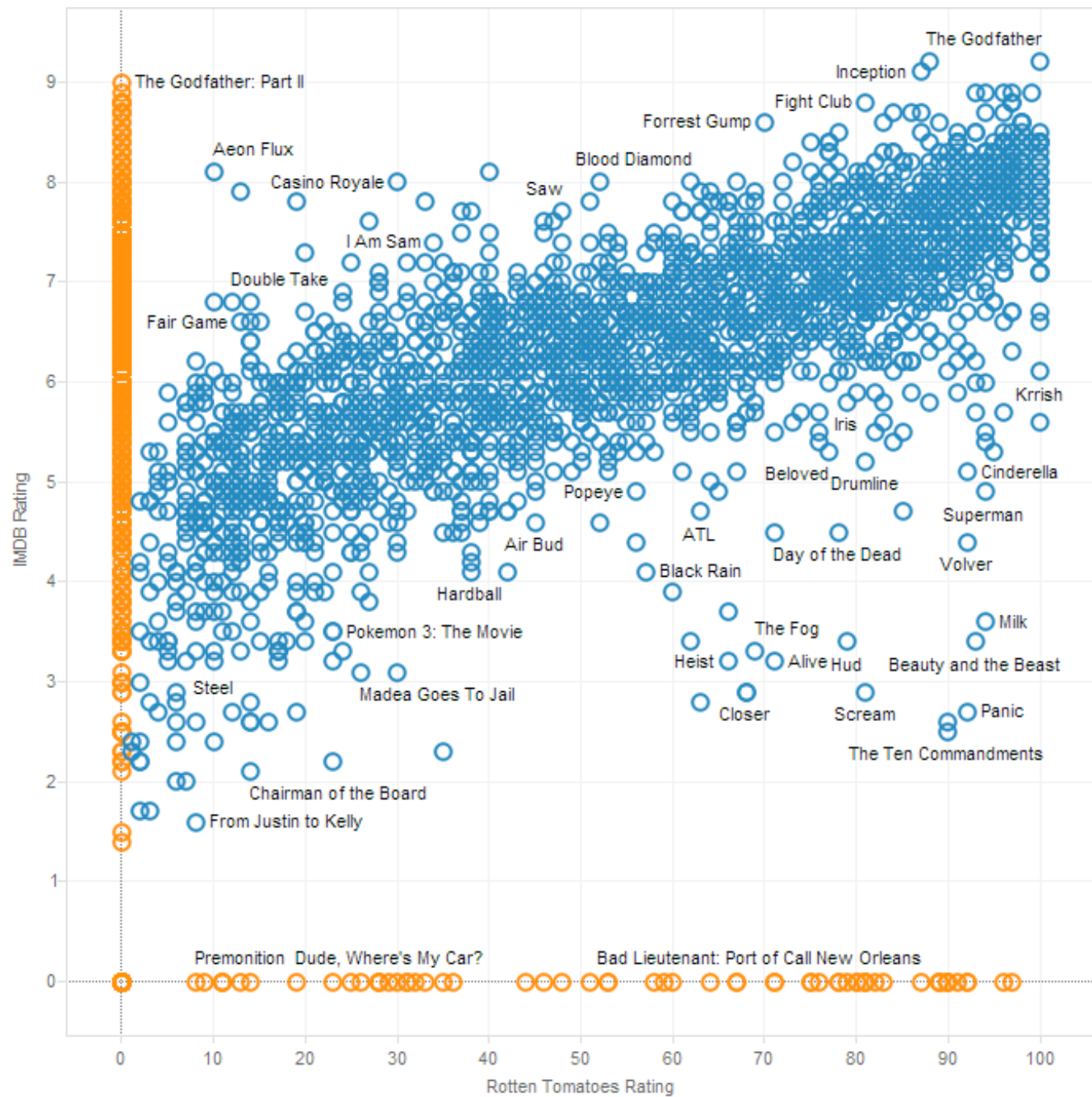


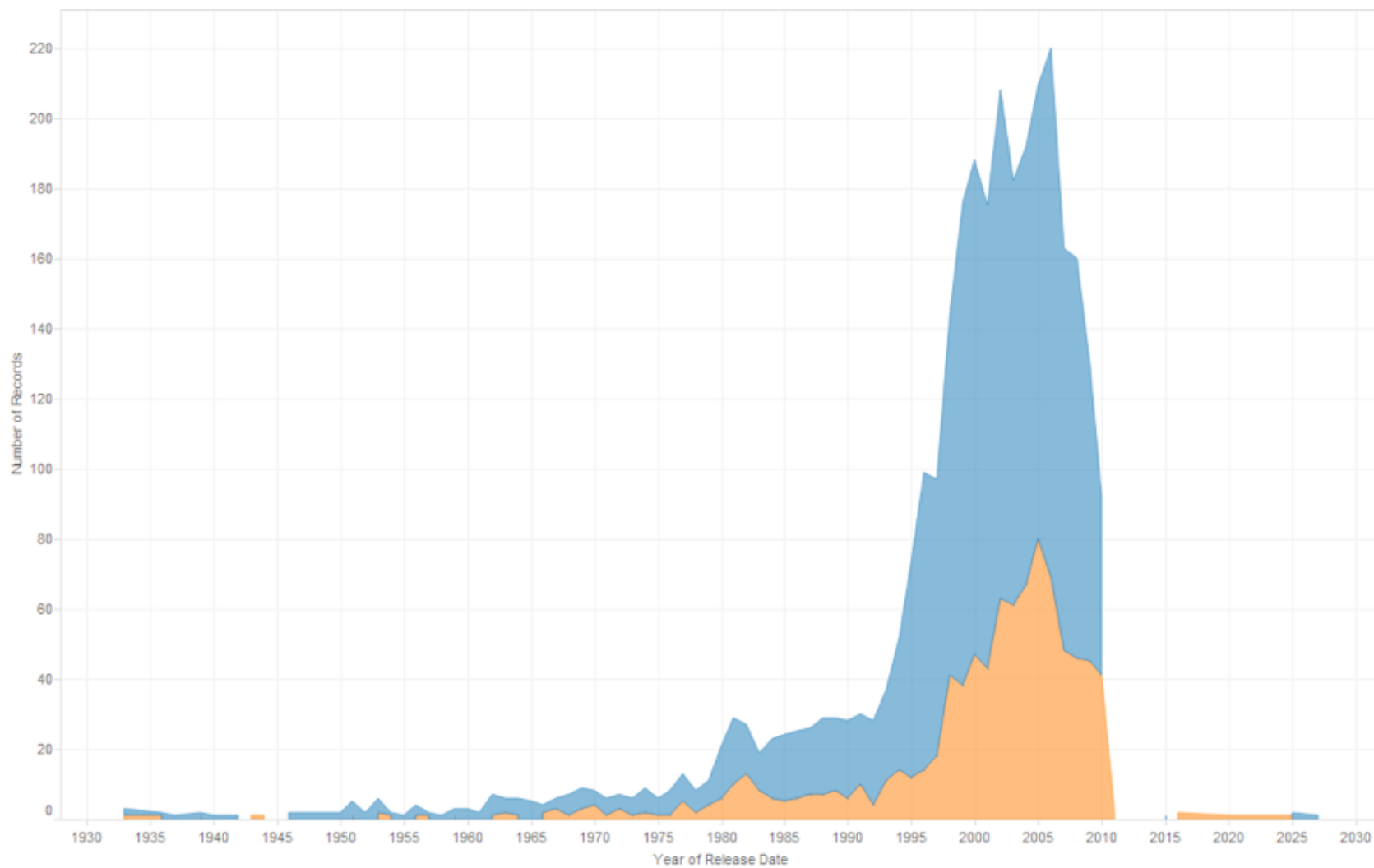
Rotten Tomatoes Rating (bin)











Lesson: Exercise Skepticism

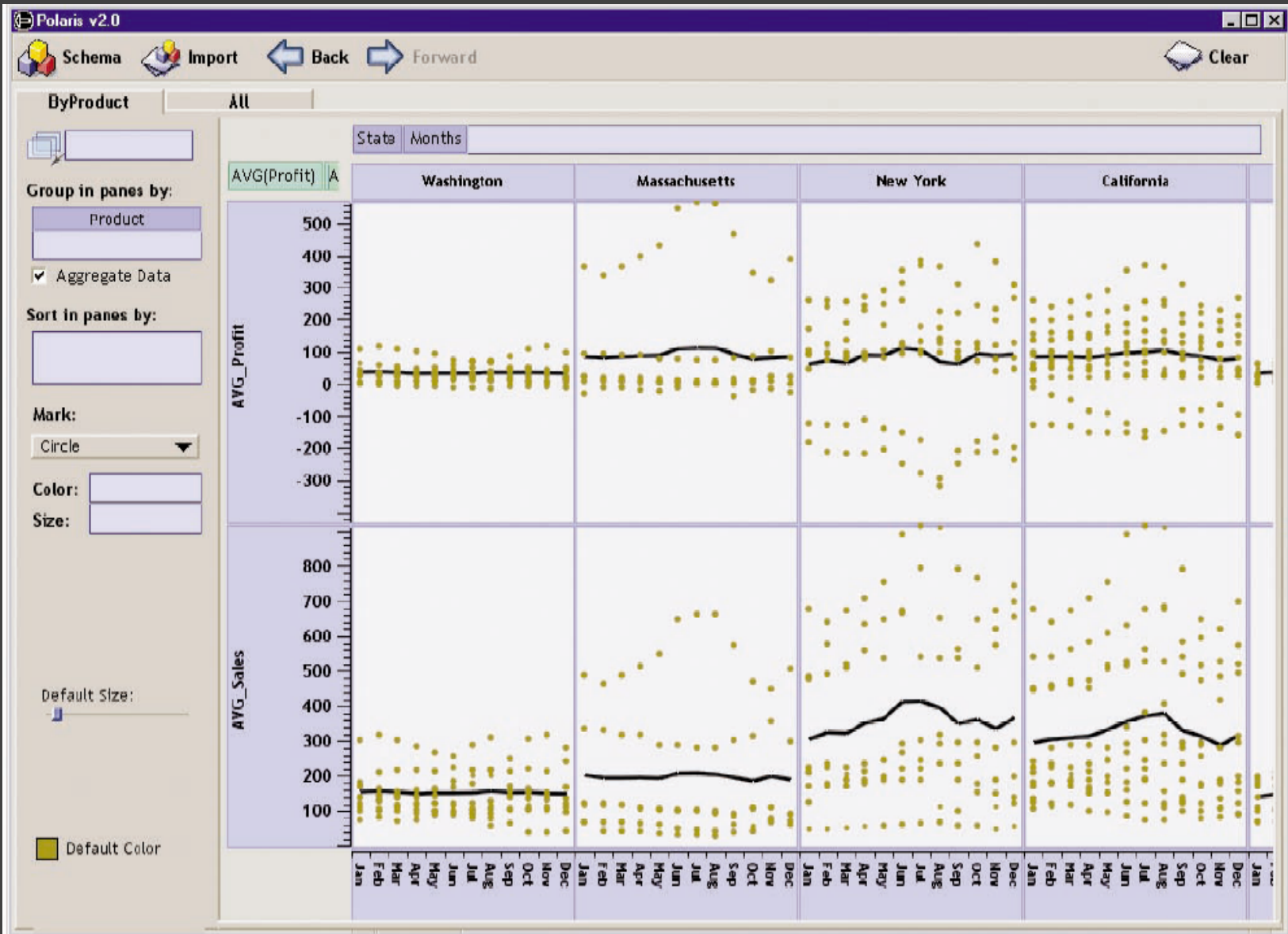
Check **data quality** and your **assumptions**.

Start with **univariate summaries**, then start to consider **relationships among variables**.

Avoid premature fixation!

Tableau / Polaris

Polaris [Stolte et al.]



Tableau

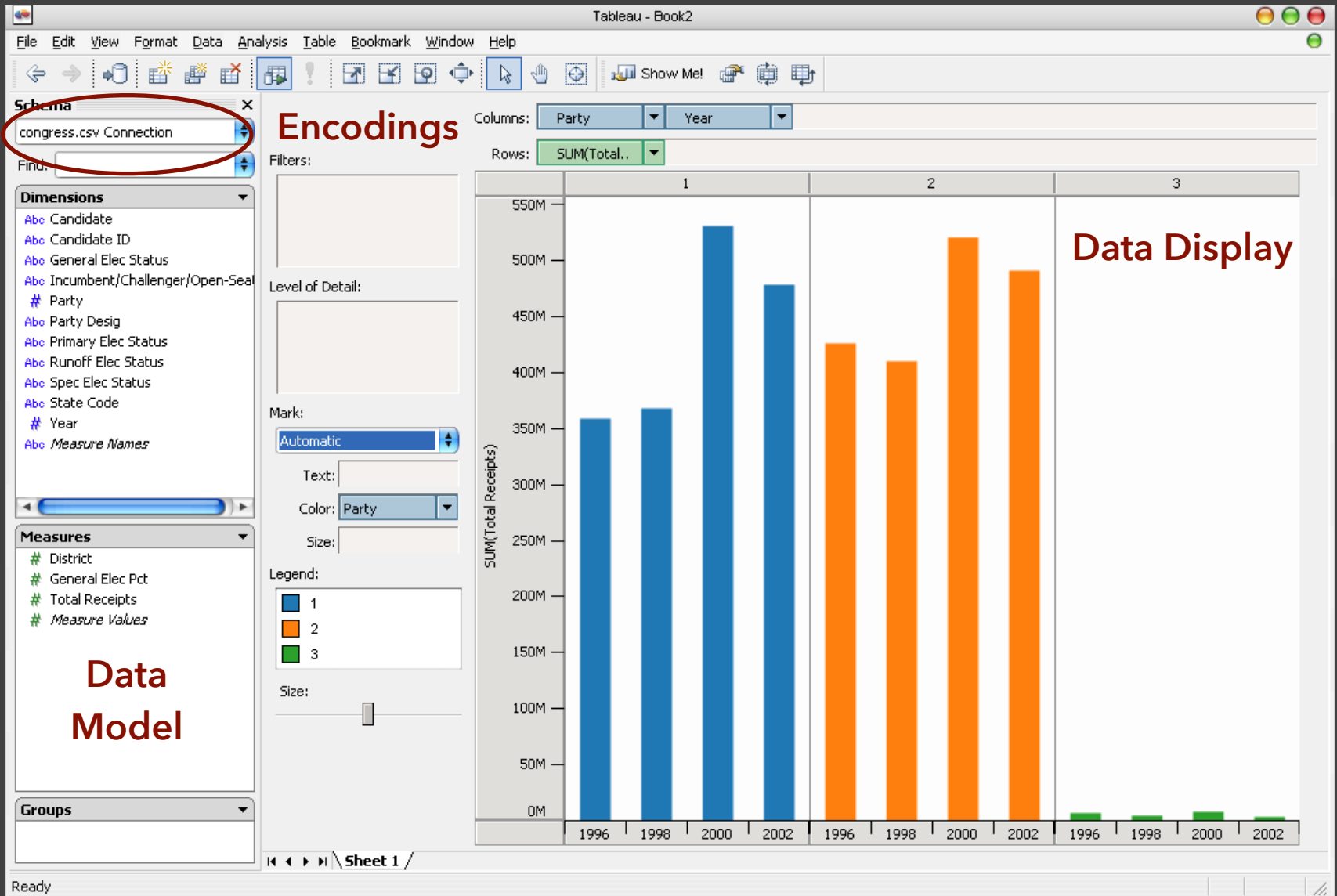


Tableau / Polaris Approach

Insight: can simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest encodings upon request

Tableau Demo

The dataset:

Federal Elections Commission Receipts

Every Congressional Candidate from 1996 to 2002

4 Election Cycles

9216 Candidacies

Dataset Schema

Year (Qi)

Candidate Code (N)

Candidate Name (N)

Incumbent / Challenger / Open-Seat (N)

Party Code (N) [1=Dem,2=Rep,3=Other]

Party Name (N)

Total Receipts (Qr)

State (N)

District (N)

This is a subset of the larger data set available from the FEC.

Hypotheses?

What might we learn from this data?

Hypotheses?

What might we learn from this data?

Correlation between receipts and winners?

Do receipts increase over time?

Which states spend the most?

Which party spends the most?

Margin of victory vs. amount spent?

Amount spent between competitors?

Tableau Demo

EDA Summary

Exploratory analysis combines graphical methods, data transformations, and statistics.

Use questions to uncover more questions.

Formal methods may be used to confirm, sometimes on held-out or unseen data.

Visualization can further aid assessment of fitted statistical models.

More to come in the *Uncertainty* lecture!

Dimensionality Reduction

Dimensionality Reduction (DR)

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

In contrast, multidimensional scaling (MDS) attempts to preserve pairwise distances.

Reduction Techniques

LINEAR - PRESERVE GLOBAL STRUCTURE

Principal Components Analysis (PCA)

Linear transformation of basis vectors, ordered by amount of data variance they explain.

NON-LINEAR - PRESERVE LOCAL TOPOLOGY

t-Dist. Stochastic Neighbor Embedding (t-SNE)

Probabilistically model distance, optimize positions.

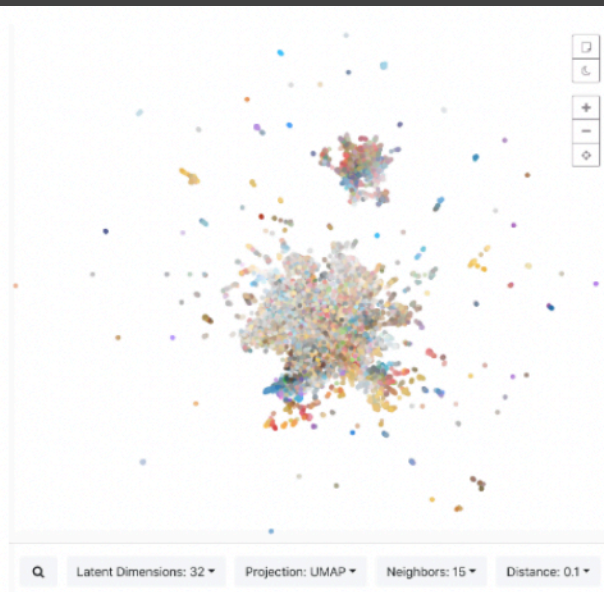
Uniform Manifold Approx. & Projection (UMAP)

Identify local manifolds, then stitch them together.

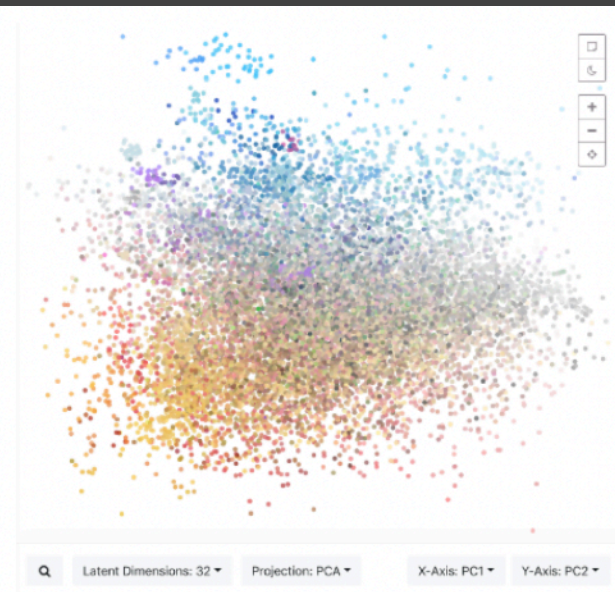
Mapping Emoji Images



t-SNE

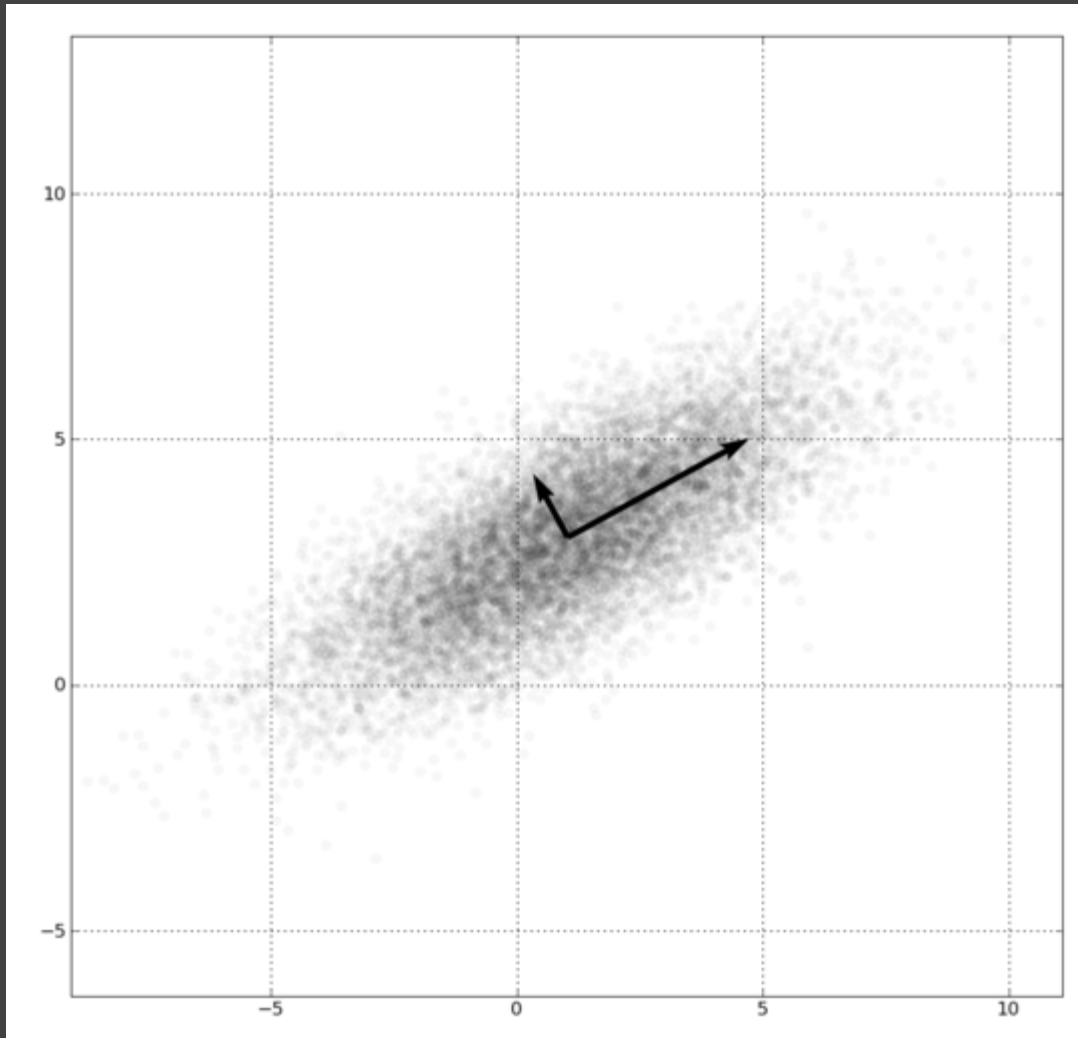


UMAP



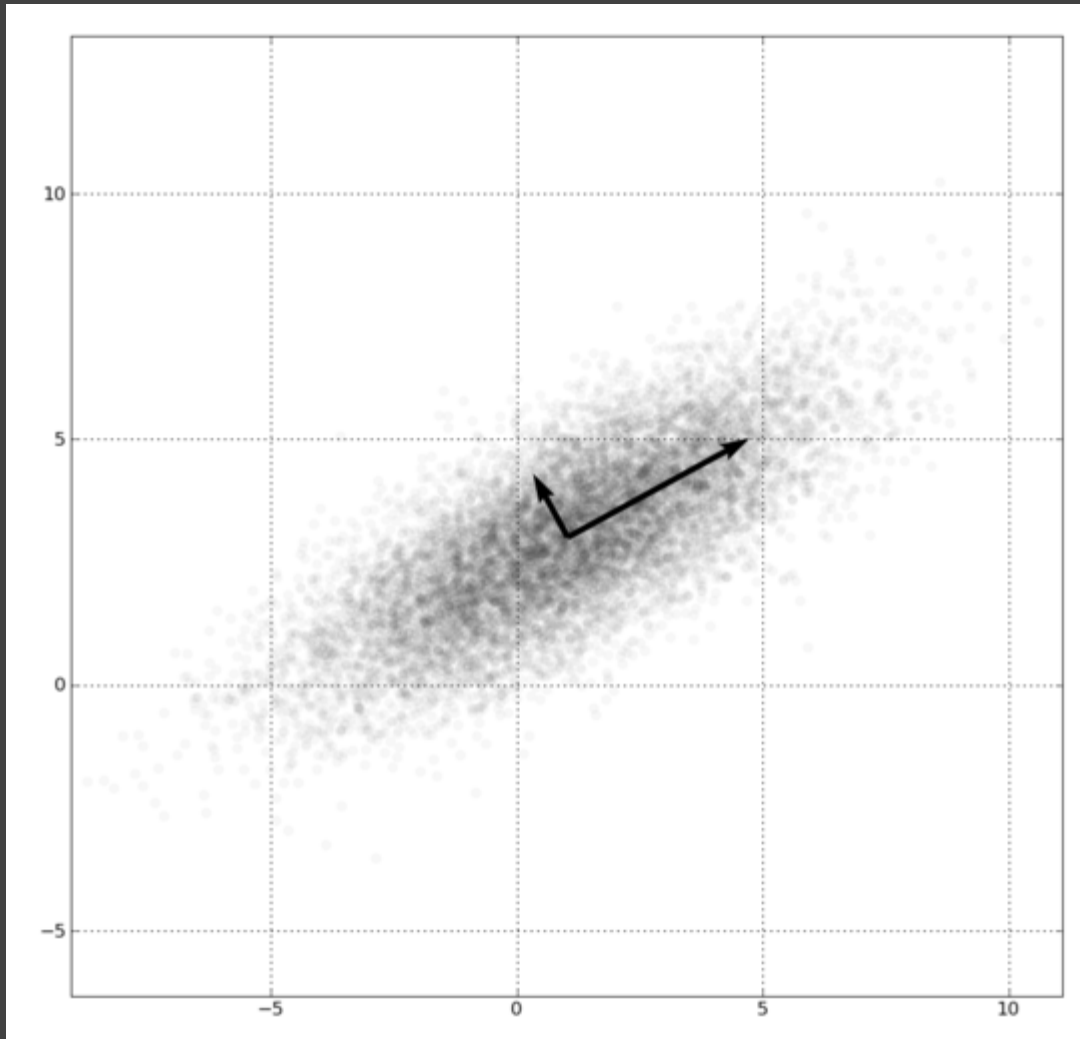
PCA

Principal Components Analysis



1. Mean-center the data.
2. Find \perp basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

Principal Components Analysis

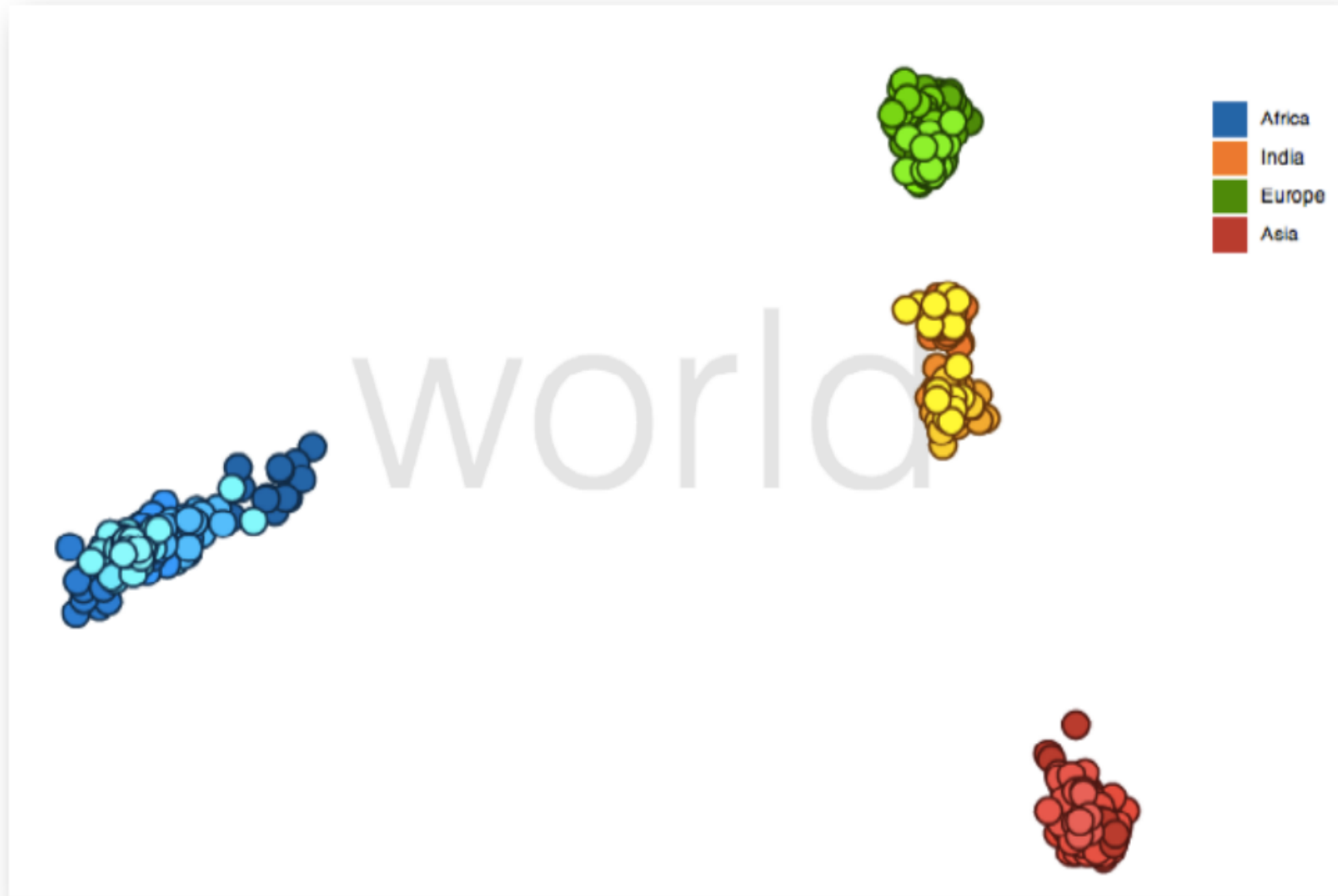


Linear transform:
scale and rotate
original space.

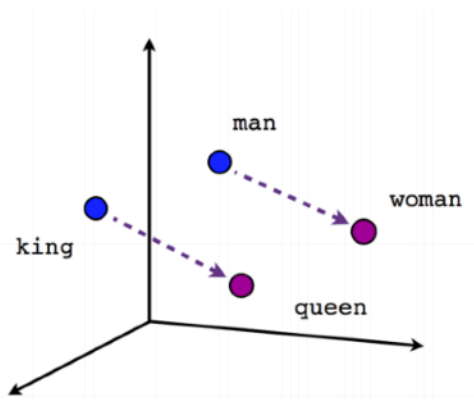
Lines (vectors)
project to lines.

Preserves global
distances.

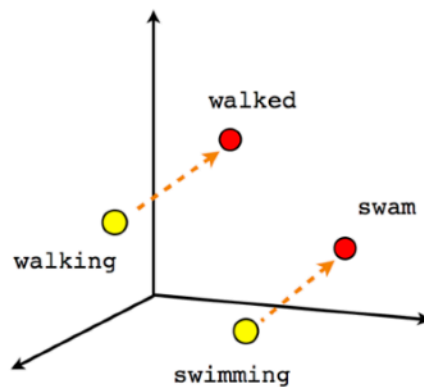
PCA of Genomes [Demiralp et al. '13]



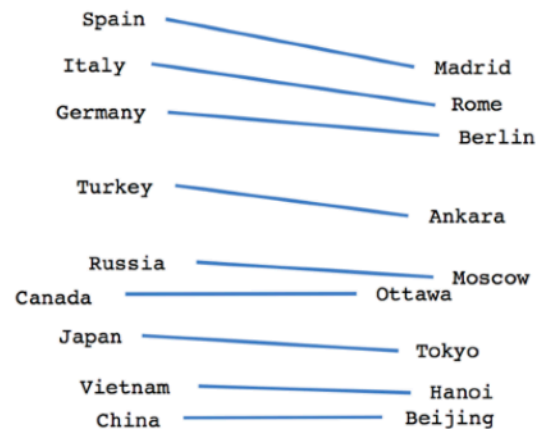
Word Embeddings (word2vec, GloVe)



Male-Female



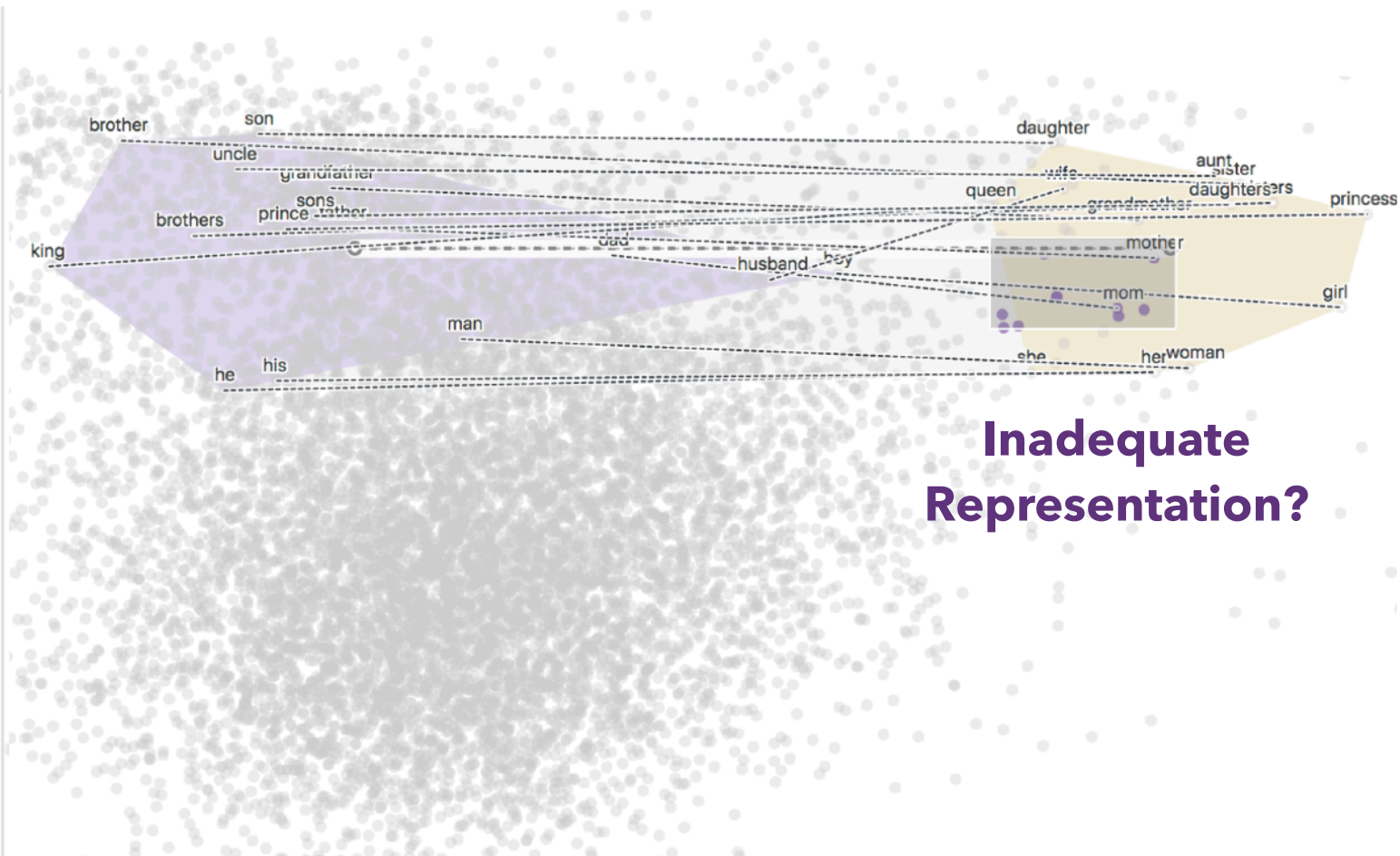
Verb tense



Country-Capital

Mapping Latent Spaces [Liu 2019]

Brushed	
mother	+
ms.	+
wedding	+
pink	Bias? +
mom	+
nurse	+
bedroom	+
ladies	+
householder	+
butterfly	+



Non-Linear Techniques

Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

t-SNE - probabilistic interpretation of distance

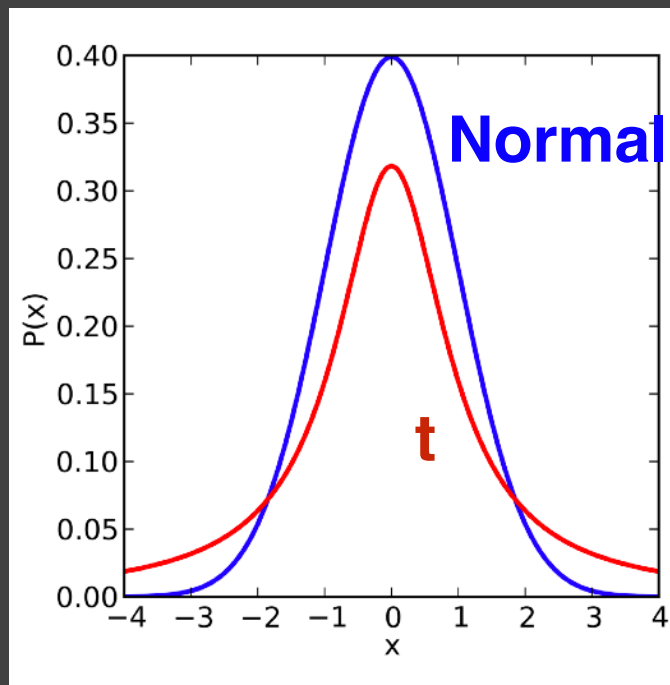
UMAP - tries to balance local/global trade-off

t-SNE [Maaten & Hinton 2008]

1. Model probability \mathbf{P} of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

t-SNE [Maaten & Hinton 2008]

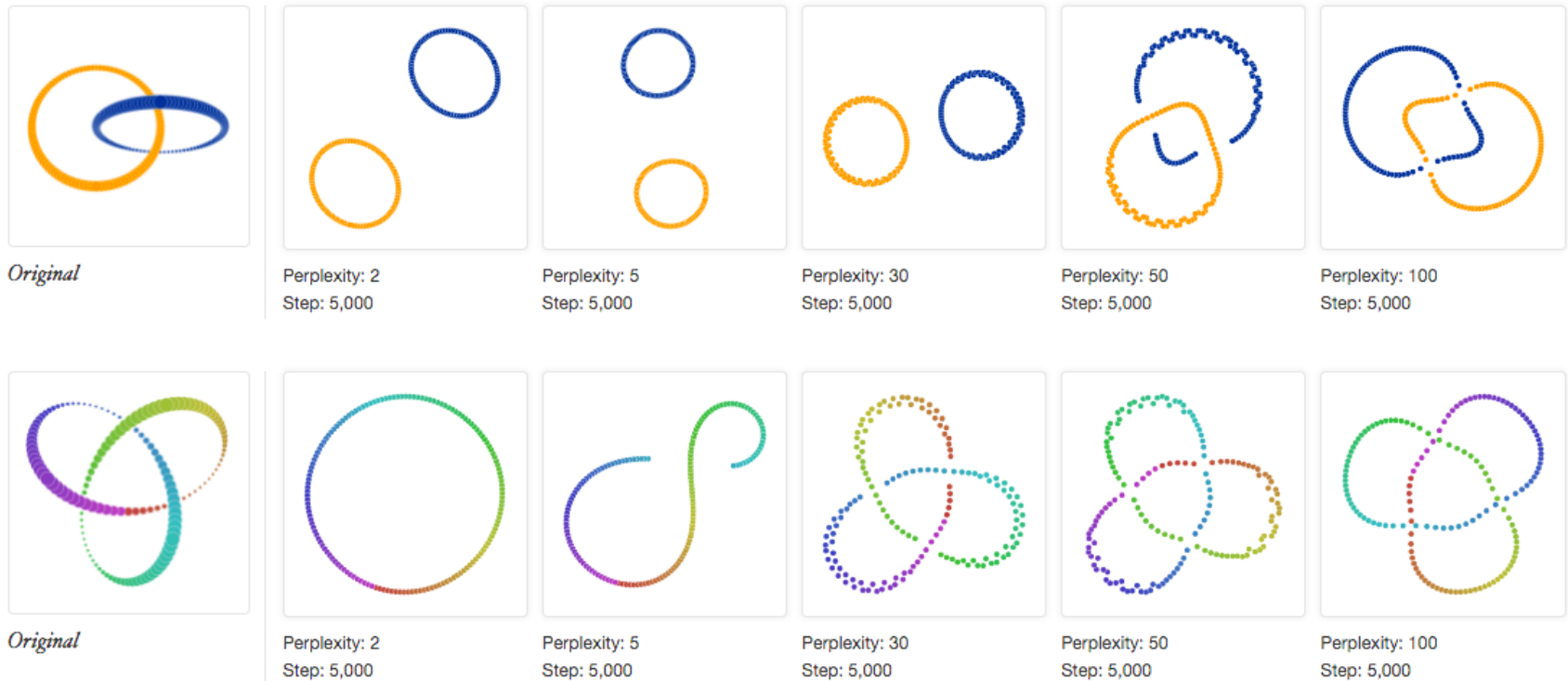
2. Define a similar probability Q in the low-dimensional (2D or 3D) embedding space, using a Student's t distribution (hence the "t-" in "t-SNE"!). The t -distribution is heavy-tailed, allowing distant points to be even further apart.



t-SNE [Maaten & Hinton 2008]

1. Model probability \mathbf{P} of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
2. Define a similar probability \mathbf{Q} in the low-dimensional (2D or 3D) embedding space, using a Student’s t distribution (*hence the “t-” in “t-SNE”!*). The t -distribution is heavy-tailed, allowing distant points to be even further apart.
3. Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the \mathbf{P} and \mathbf{Q} distributions: $KL(P \parallel Q)$

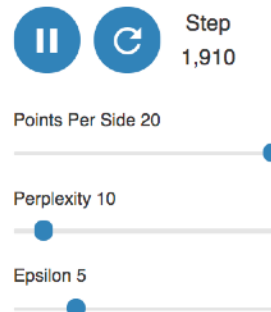
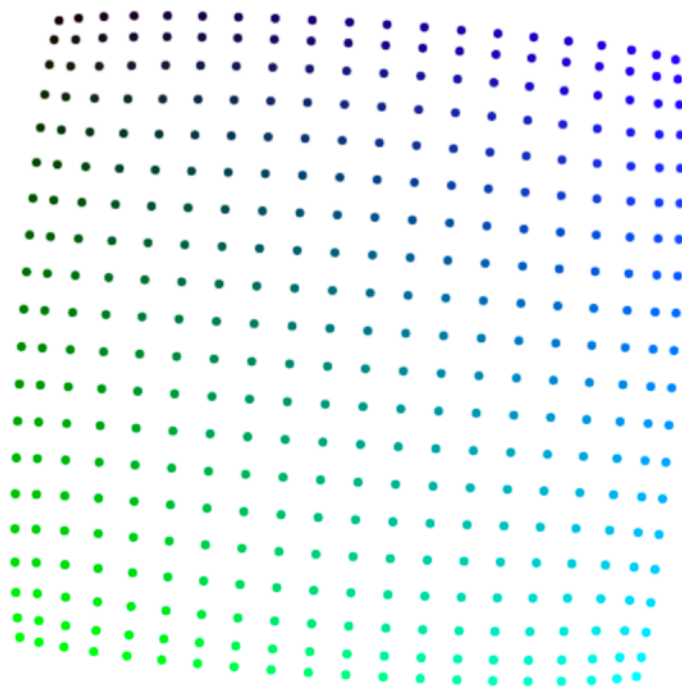
Visualizing t-SNE [Wattenberg et al. '16]



Results can be highly sensitive to the algorithm parameters!
Are you seeing real structures, or algorithmic hallucinations?

How to Use t-SNE Effectively

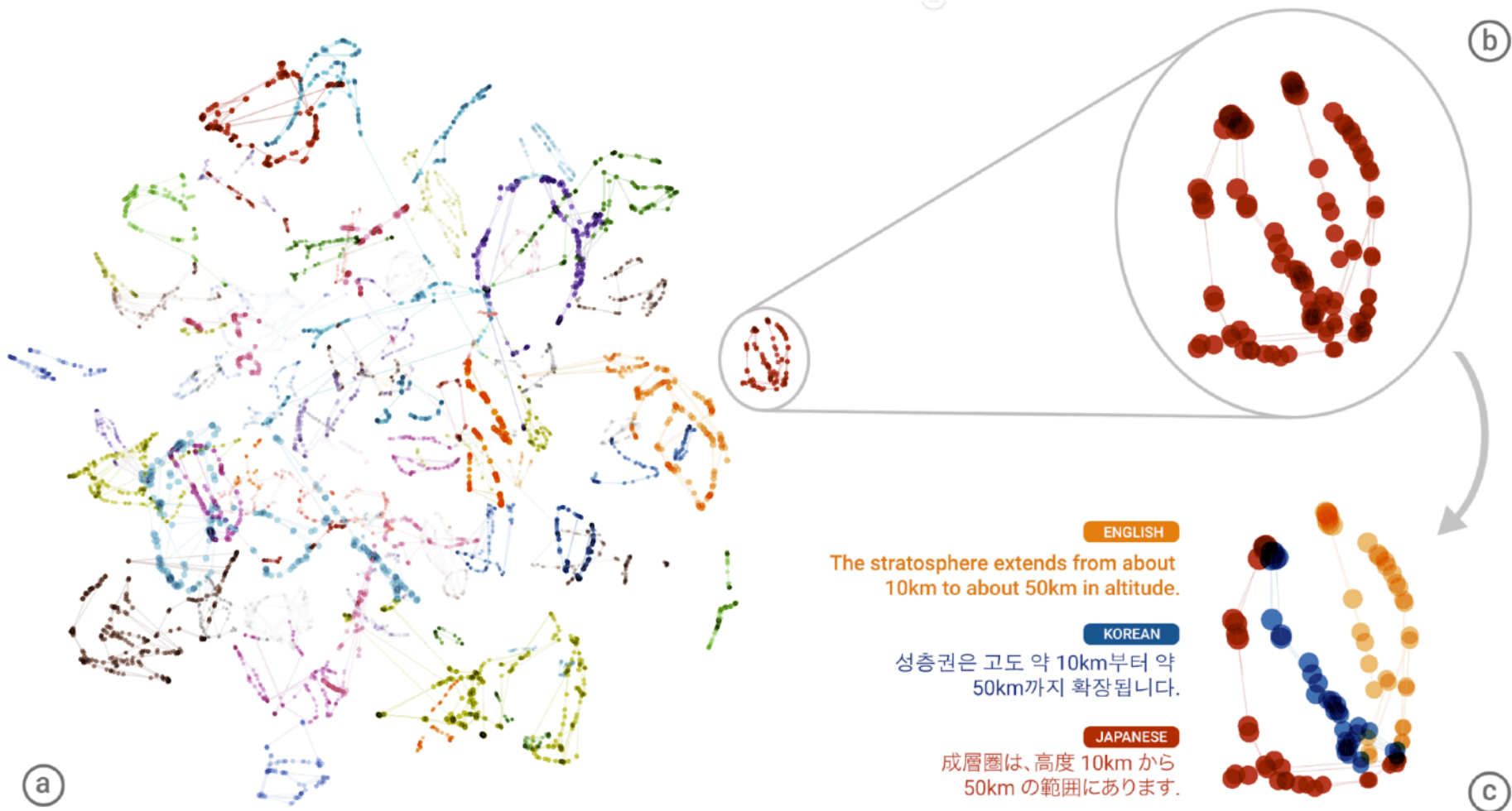
Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.



A square grid with equal spacing between points. Try convergence at different sizes.

distill.pub

MT Embedding [Johnson et al. 2018]



t-SNE projection of latent space of language translation model.

UMAP [McInnes et al. 2018]

Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

“Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance.” - McInnes et al. 2018

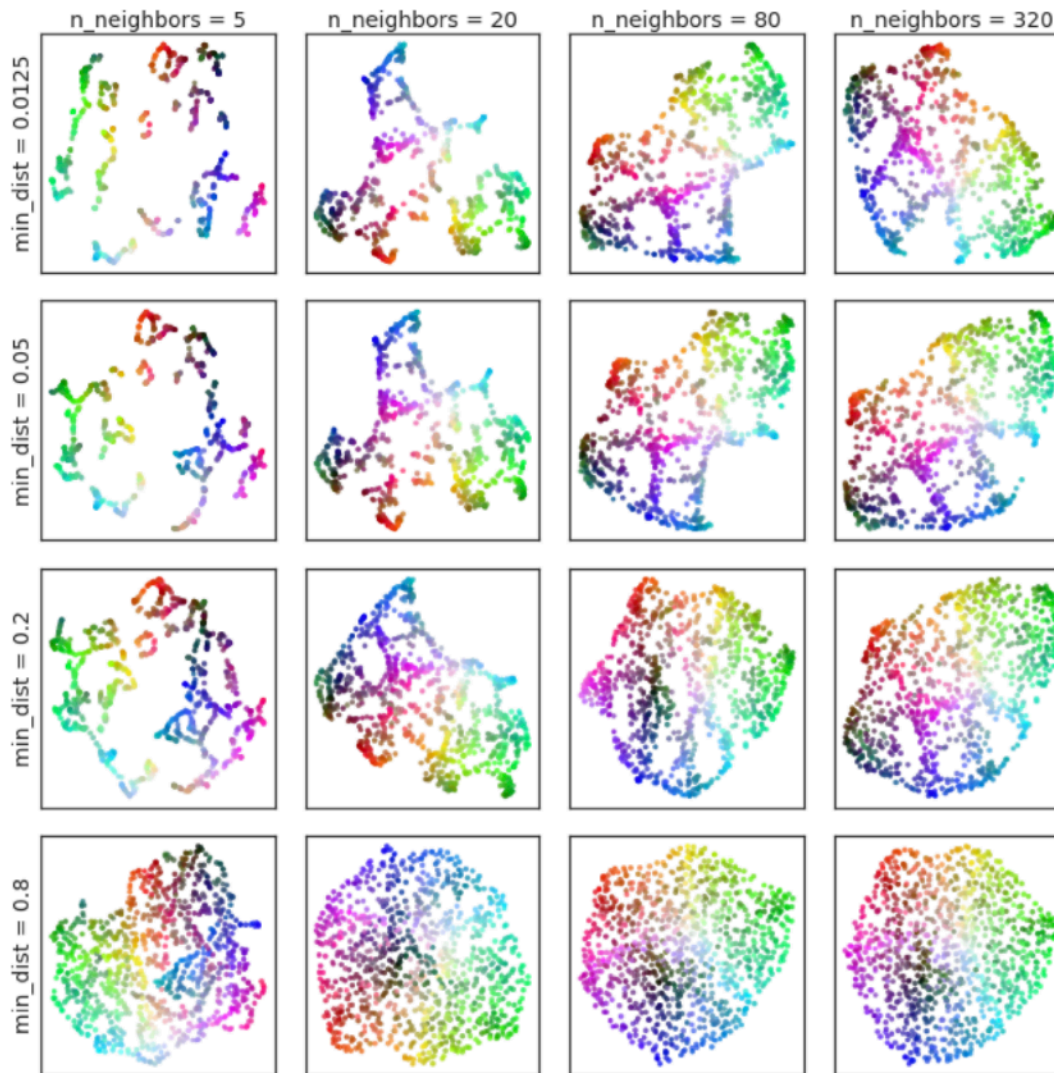


Figure 1: Variation of UMAP hyperparameters n and min_dist result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of n spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.

User Activity in Interactive Articles

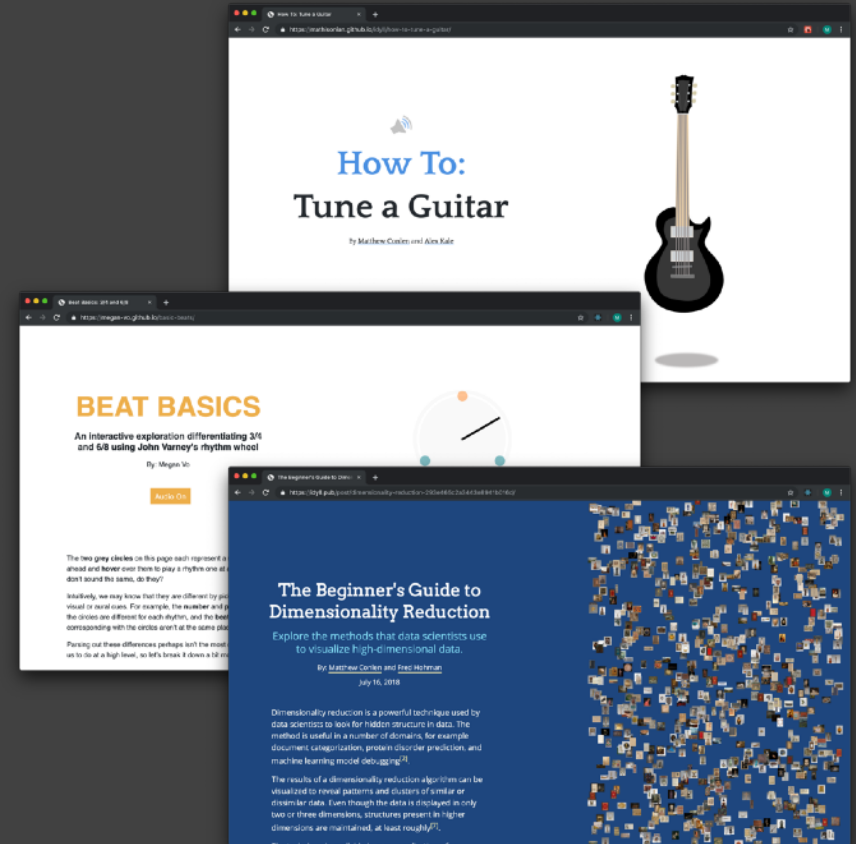
Represent reader sessions as a feature vector with:

- time spent in each section
- count of variable changes

Provide an overview of usage patterns of interactive features.

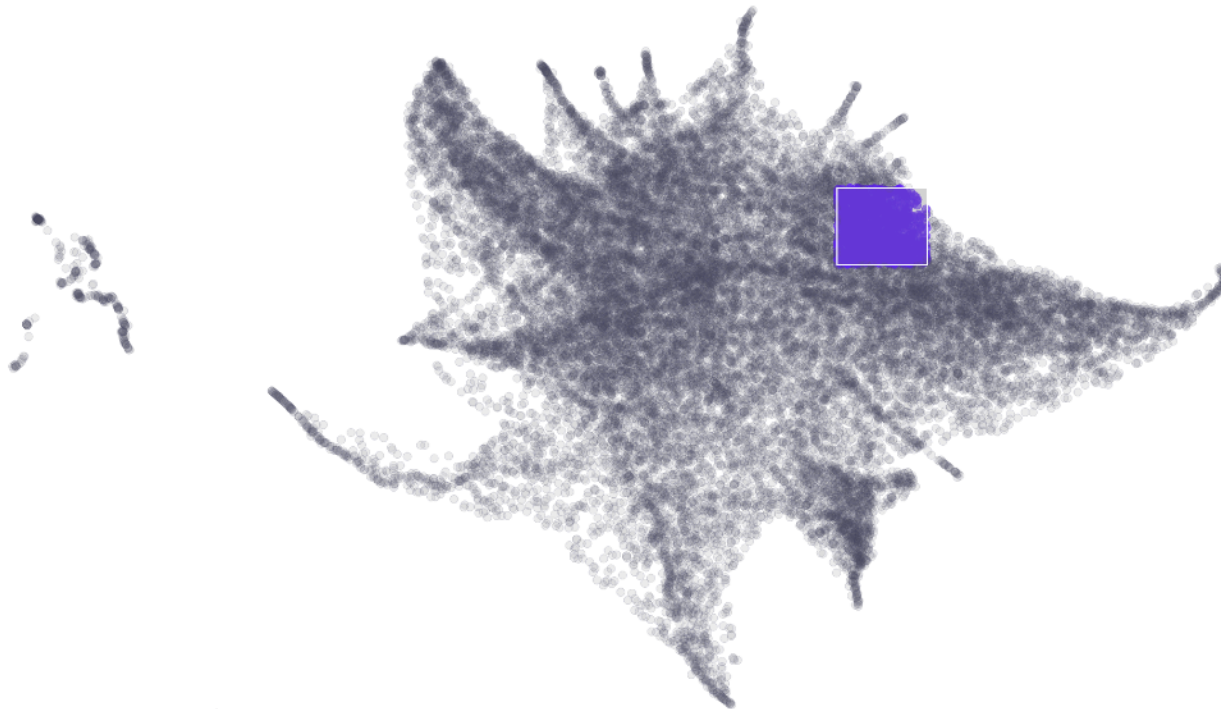
Identify variations in usage.

[Conlen '19]

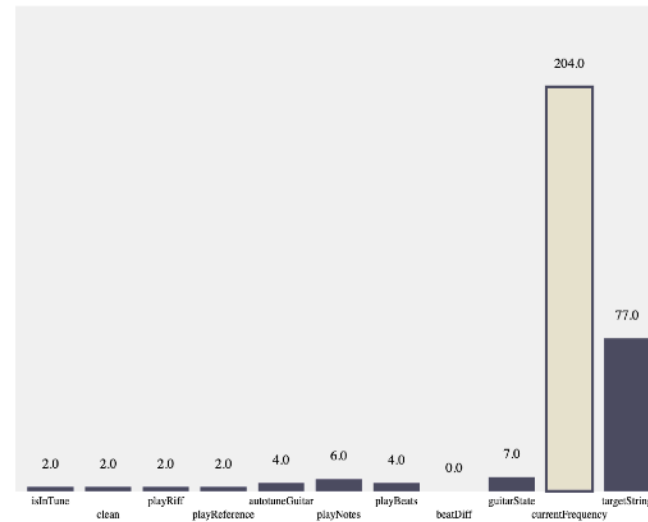
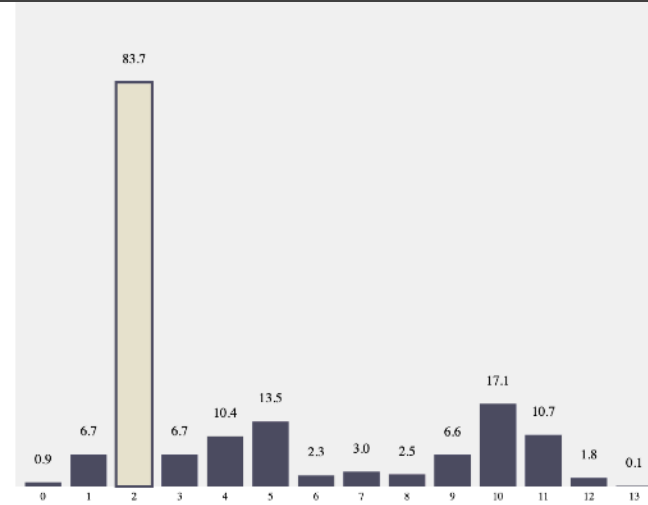


User Activity in Interactive Articles

Showing 1233 users.



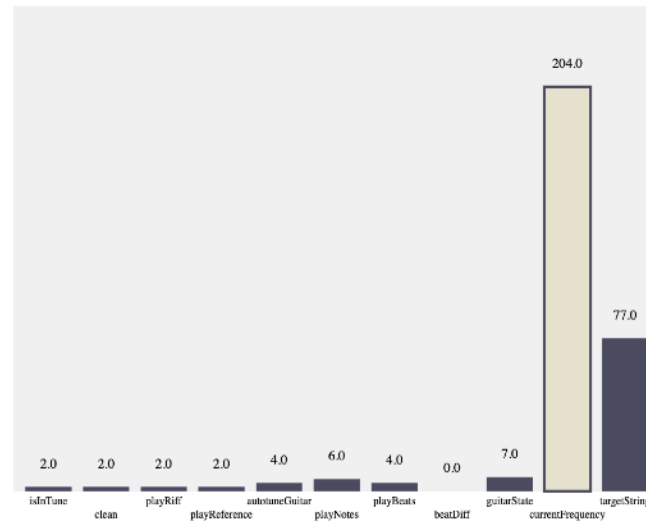
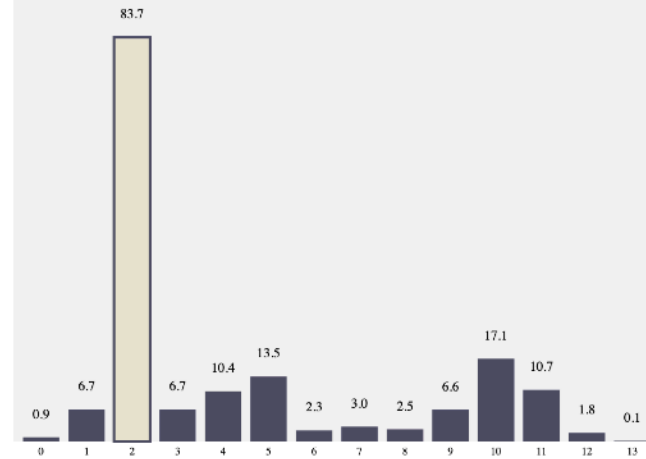
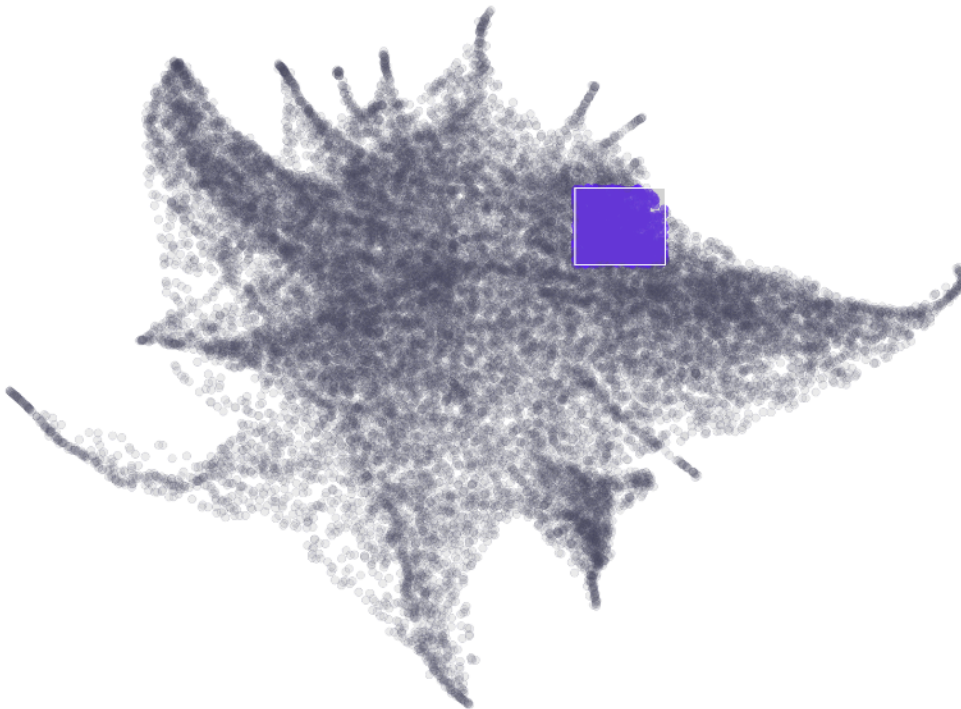
Each point represents a readers session, projected via UMAP.



User Activity in Interactive Articles

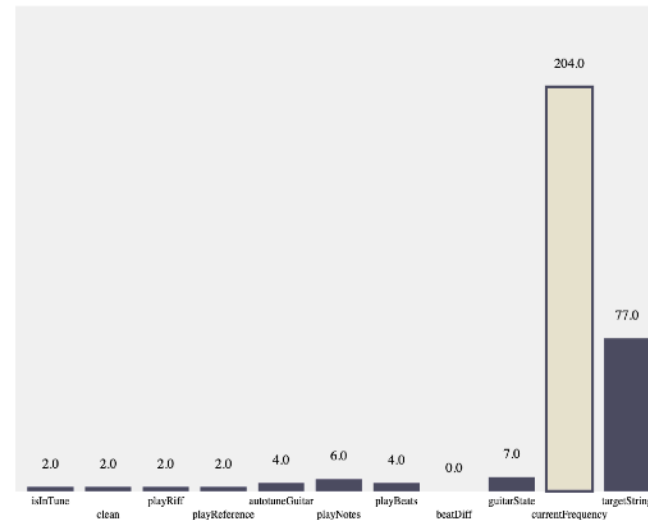
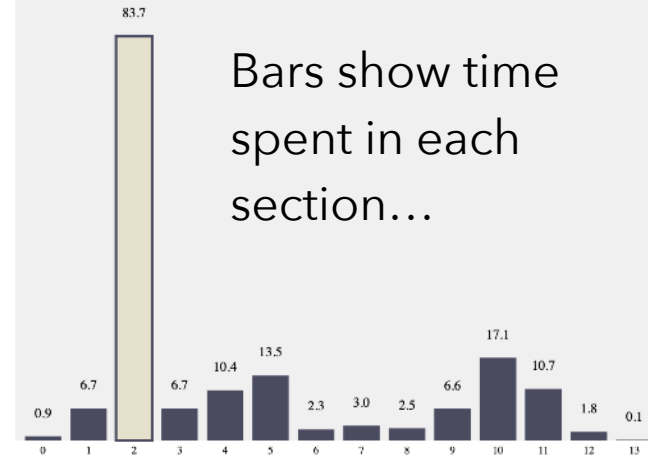
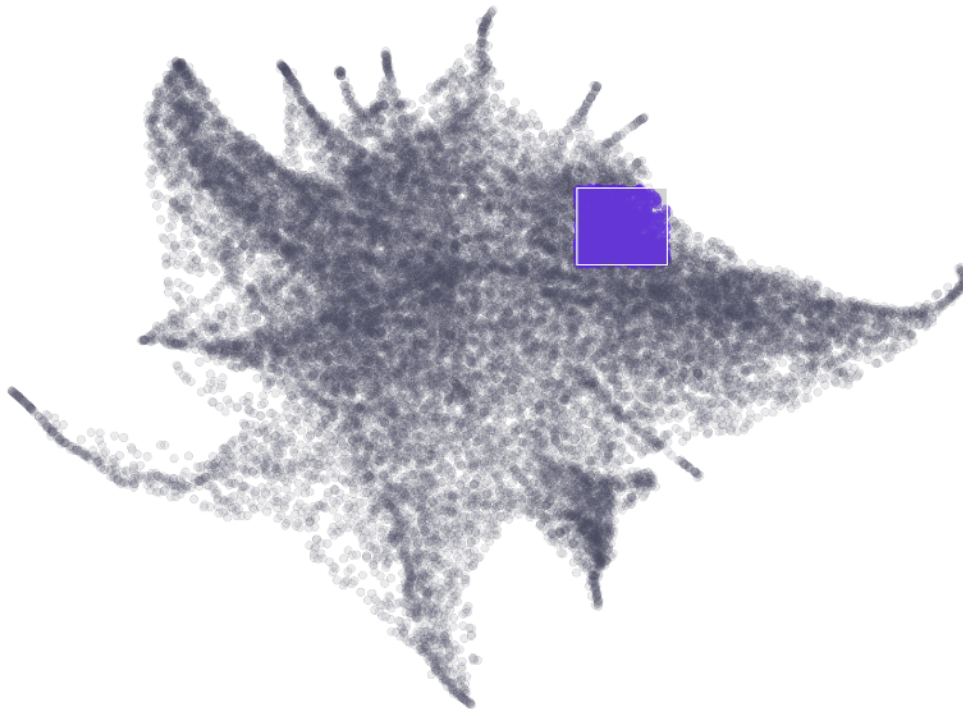
Showing 1233 users.

Brushing to select a subset of sessions.



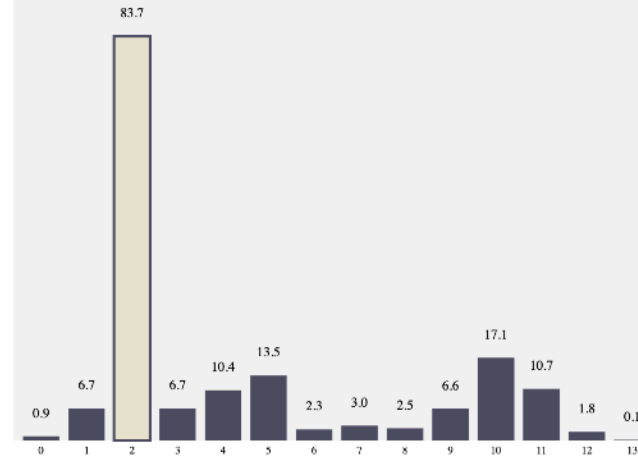
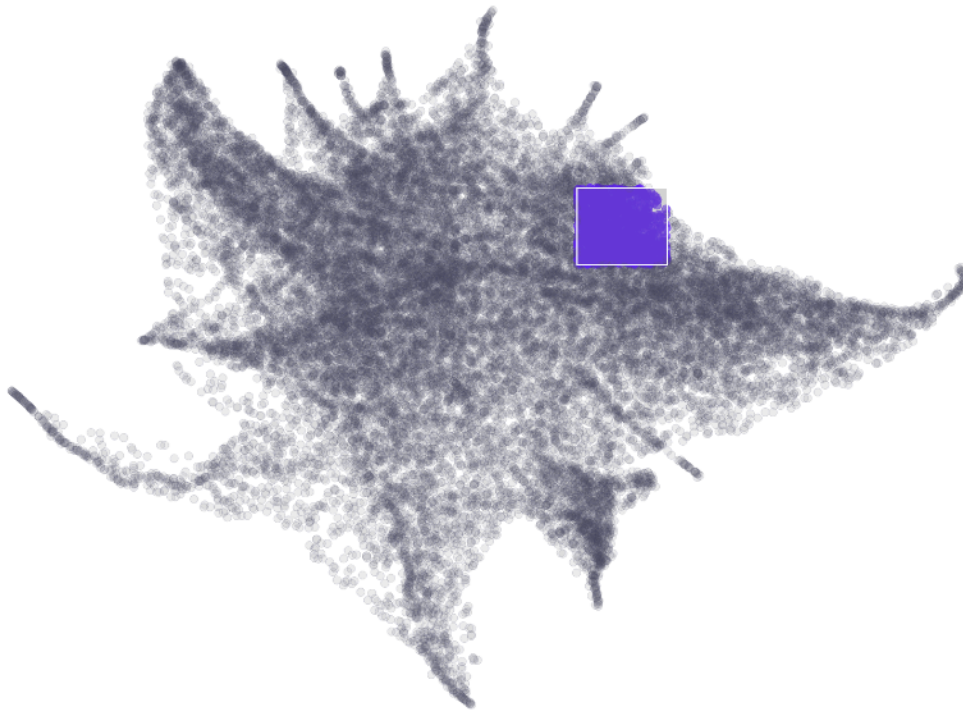
User Activity in Interactive Articles

Showing 1233 users.

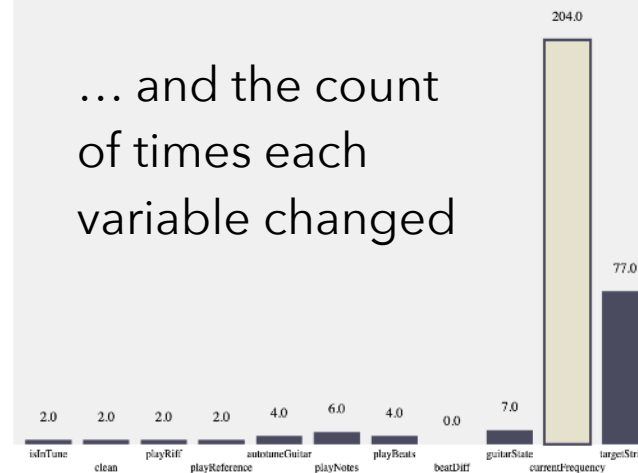


User Activity in Interactive Articles

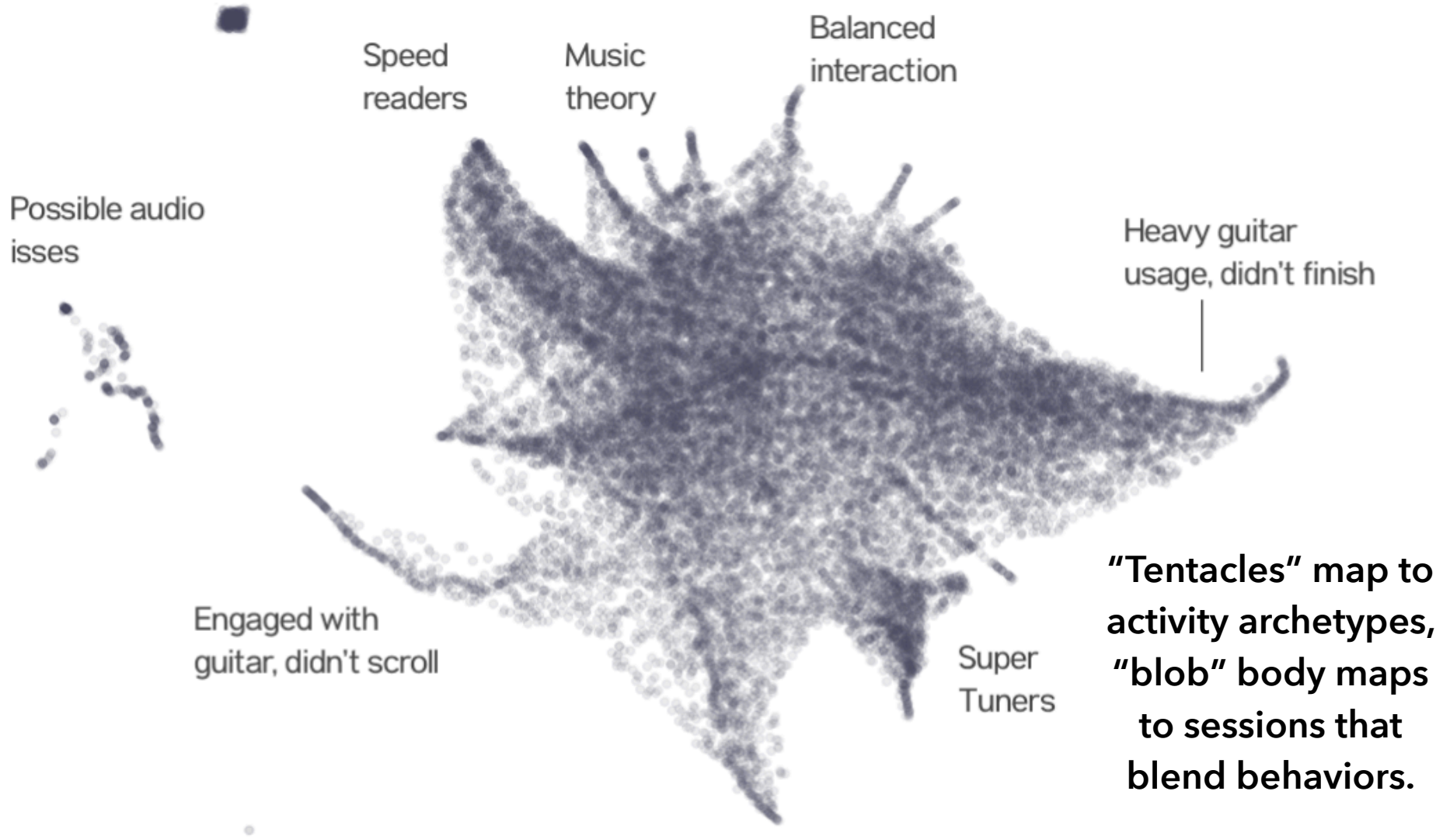
Showing 1233 users.



... and the count of times each variable changed

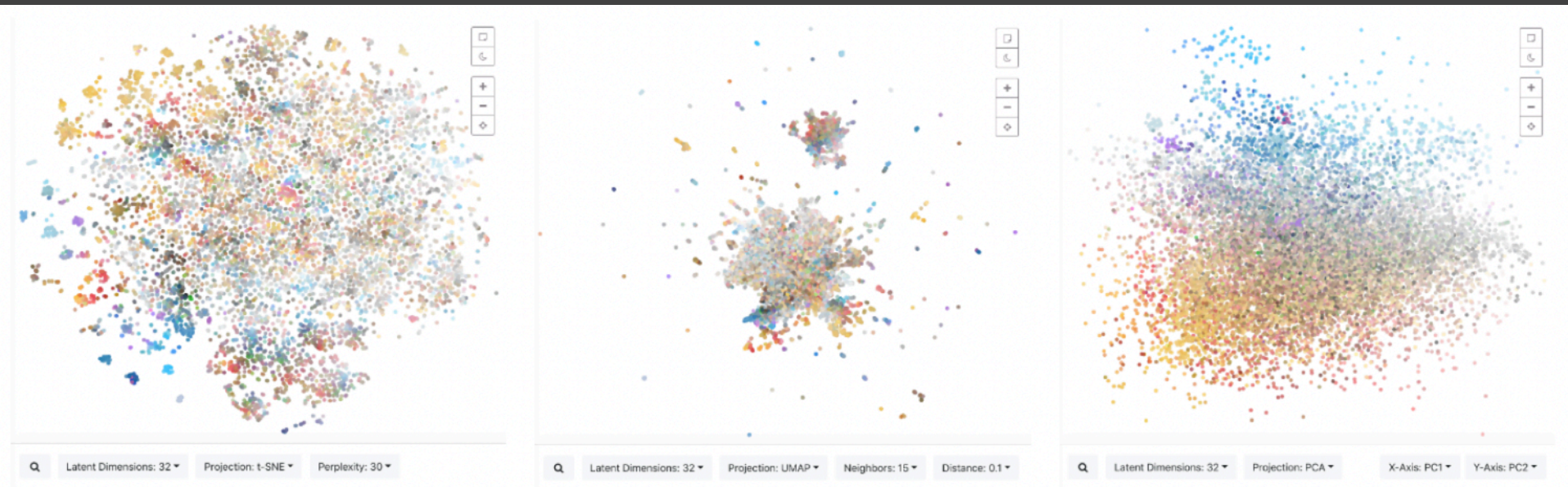


Reader Behavior [Conlen et al. 2019]



UMAP projection of reader activity for an interactive article.

Mapping Emoji Images



t-SNE

UMAP

PCA

Each has strengths and weaknesses – and they can be used in tandem!

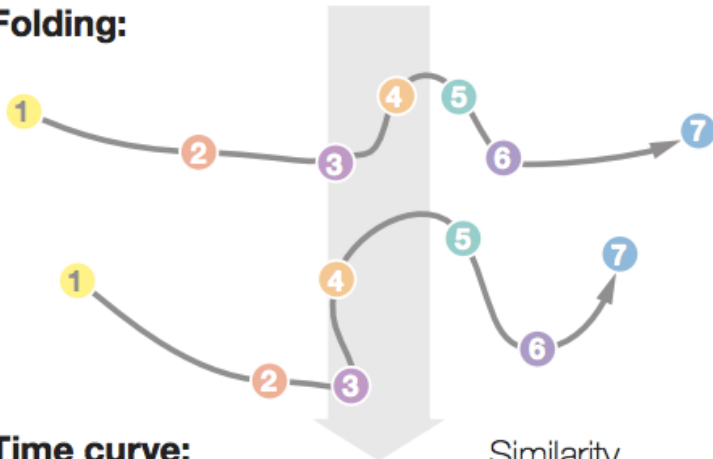
Time Curves [Bach et al. '16]

Timeline:



Circles are data cases with a time stamp.
Similar colors indicate similar data cases.

Folding:

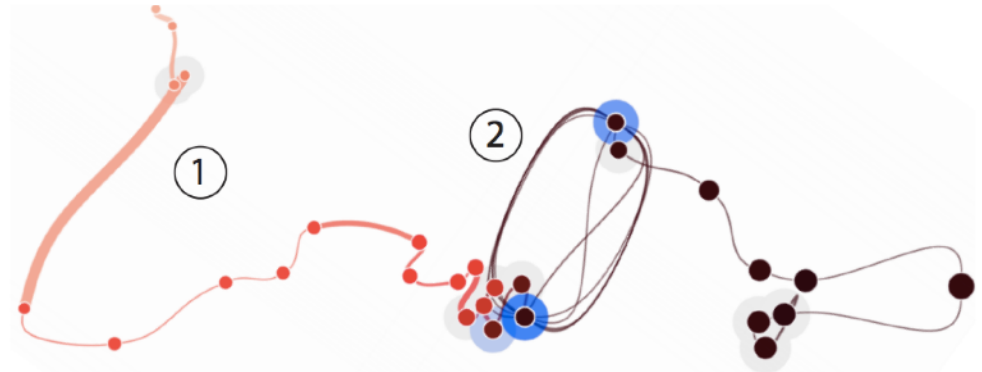


Time curve:

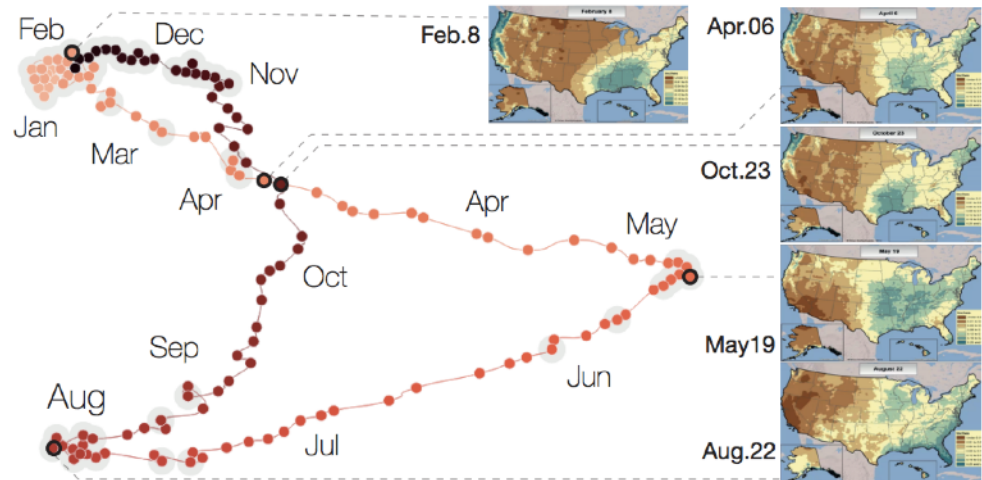


The temporal ordering of data cases is preserved.
Spatial proximity now indicates similarity.

(a) Folding time



Wikipedia "Chocolate" Article

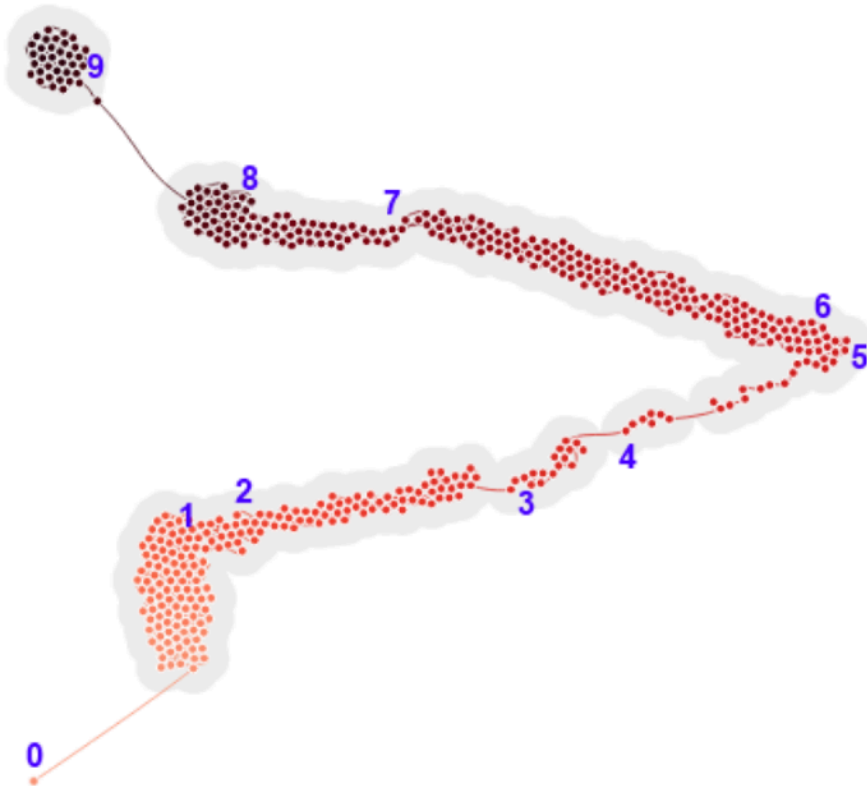


U.S. Precipitation over 1 Year

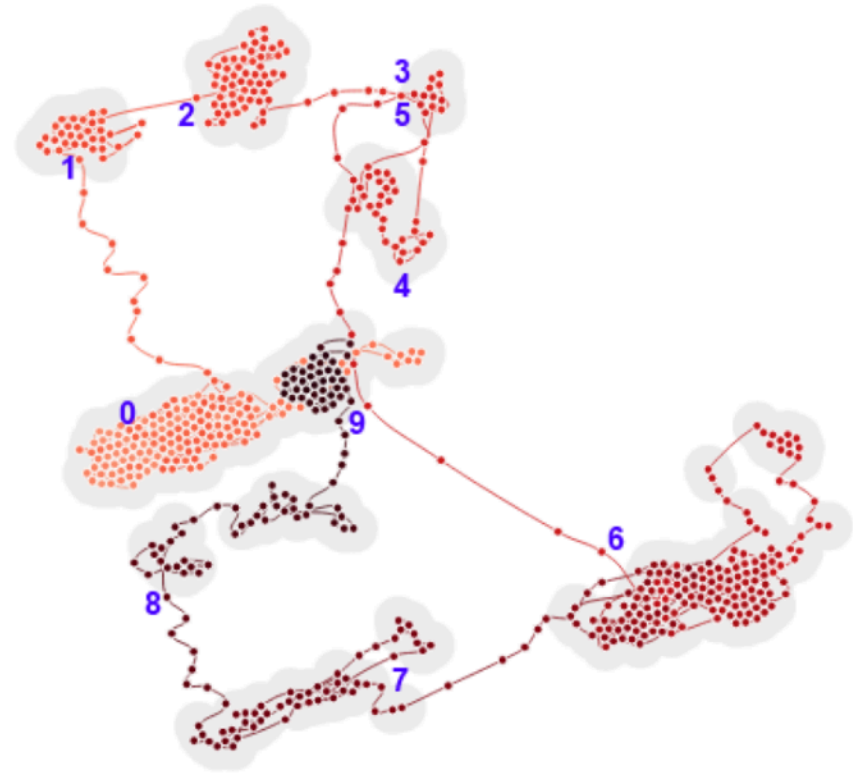
Rover Telemetry

[Guy '16]

How to track high-dimensional state?



Using Raw Multi-D Data



Using Pearson Correlation Matrix

Dimensionality Reduction Issues

Reproducible?

Projections are *data-dependent*. Fitting a new projection with different data can give rise to different results.

Reusable?

PCA and UMAP provide reusable projection functions that can map new points from high-D to low-D. t-SNE (and others, like MDS) do not provide this.

Interpretable?

DR plots are hard to interpret! Try multiple methods and hyperparameter settings. Inspect via interaction!