“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

Data Science

“Data Science” is thinking with data

How to categorize data

How to computationally explore data

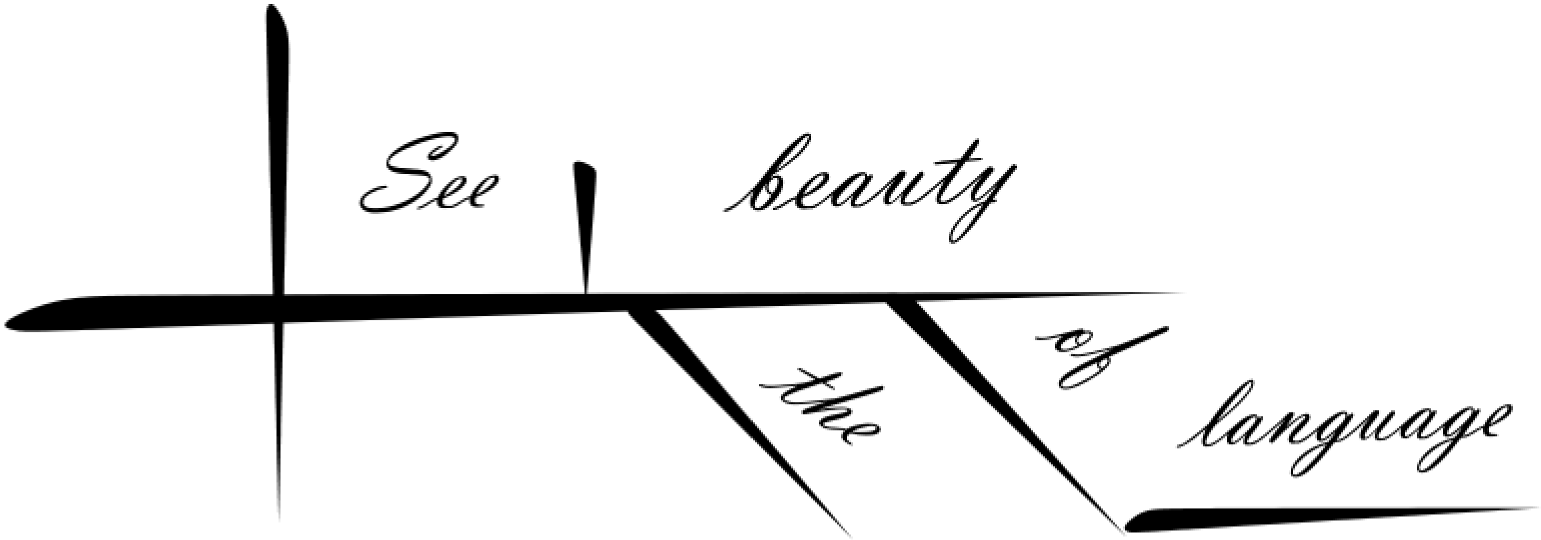
How to visually explore data

Please ask questions if you have them!!!

How to categorize data

How to computationally explore data

How to visually explore data



What are the different properties of the data

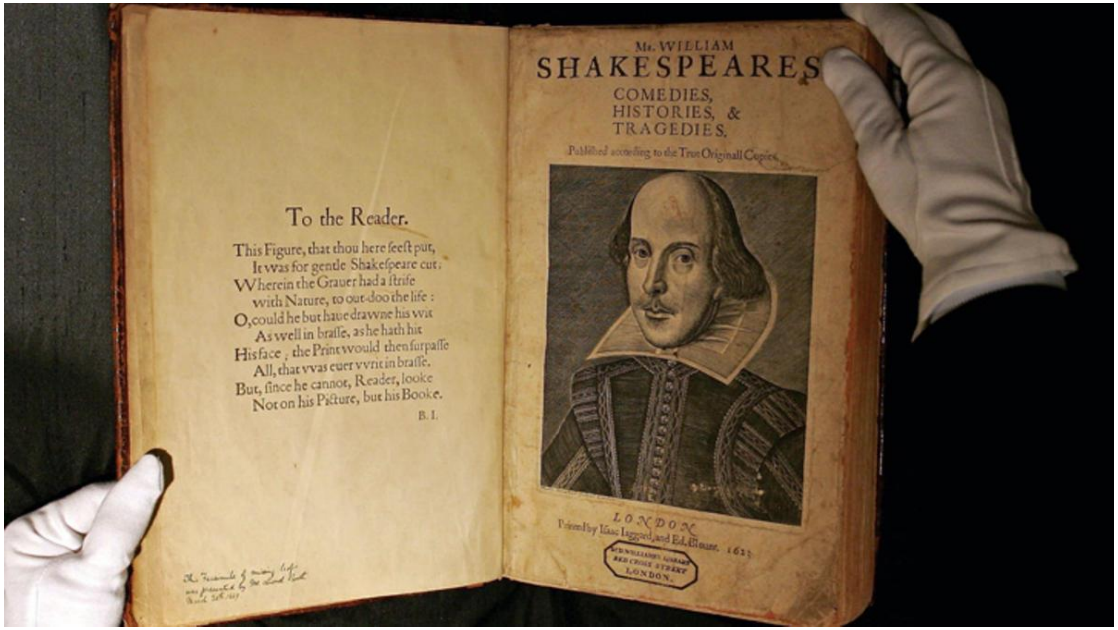
Data falls into two categories:

Quantitative:
Numeric measures

Qualitative:
Descriptions, categories, and observations

Data about this book:

Quantitative:
142 pages
20,000 words
1,700 nouns



To the Reader.

This Figure, that thou here ſeeſt put,
It was for gentle Shakeſpeare cut;
Wherein the Grauer had a ſtrife
with Nature, to out-doo the life:
O, could he but haue drawne his wit
As well in braſſe, as he hath hit
His face; the Print would then ſurpaſſe
All, that vvas euer vvrit in braſſe.
But, ſince he cannot, Reader, looke
Not on his Picture, but his Booke.

B. I.

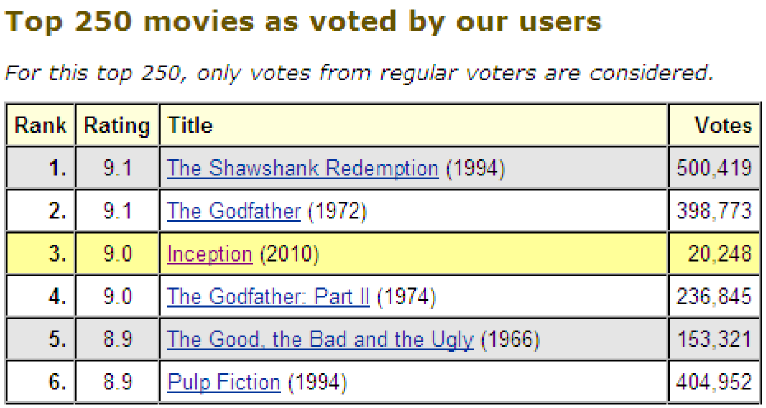
Mr. WILLIAM
SHAKESPEARES
COMEDIES,
HISTORIES, &
TRAGEDIES.
Publiſhed according to the True Originall Copies.

LONDON
Printed by Iſaac Iaggard, and Ed. Blount. 1623.

DR WILLIAMS'S LIBRARY
RED CROSS STREET
LONDON.

Qualitative:
Old
By Shakespeare
Published in London

Data can also be:



Top 250 movies as voted by our users

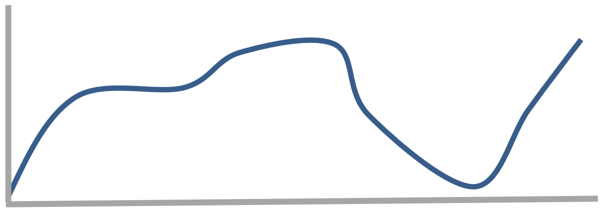
For this top 250, only votes from regular voters are considered.

Rank	Rating	Title	Votes
1.	9.1	The Shawshank Redemption (1994)	500,419
2.	9.1	The Godfather (1972)	398,773
3.	9.0	Inception (2010)	20,248
4.	9.0	The Godfather: Part II (1974)	236,845
5.	8.9	The Good, the Bad and the Ugly (1966)	153,321
6.	8.9	Pulp Fiction (1994)	404,952

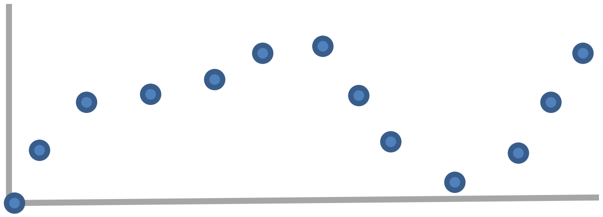
Ordinal

Pigs
Cows
Dogs

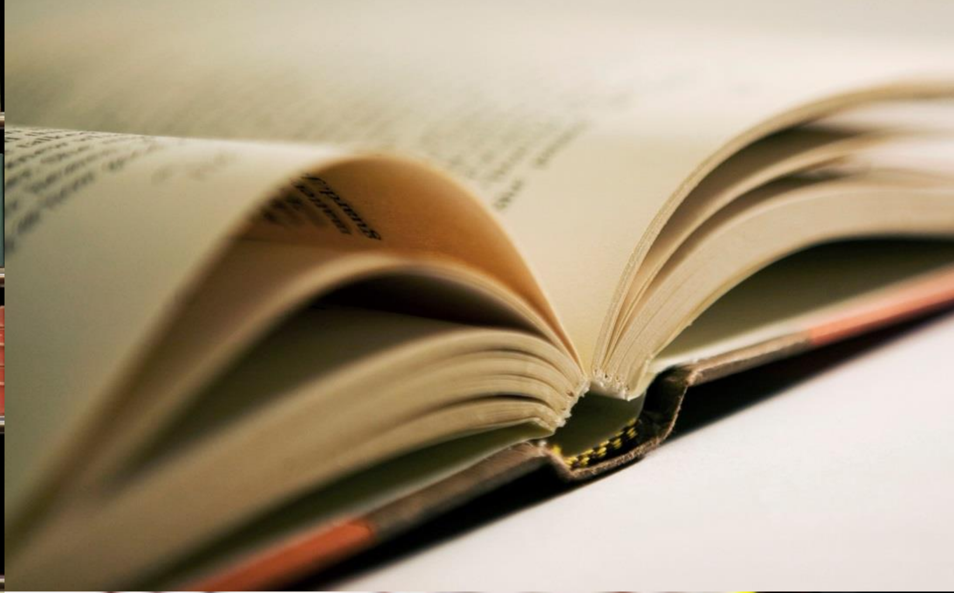
Categorical/Nominal

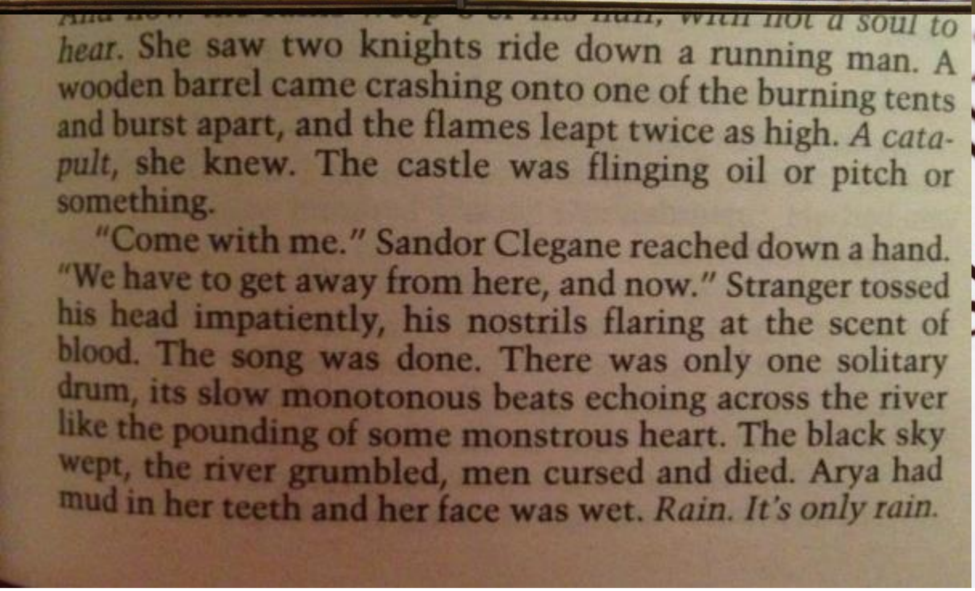


Continuous



Discrete





with not a soul to *hear*. She saw two knights ride down a running man. A wooden barrel came crashing onto one of the burning tents and burst apart, and the flames leapt twice as high. *A catapult*, she knew. The castle was flinging oil or pitch or something.

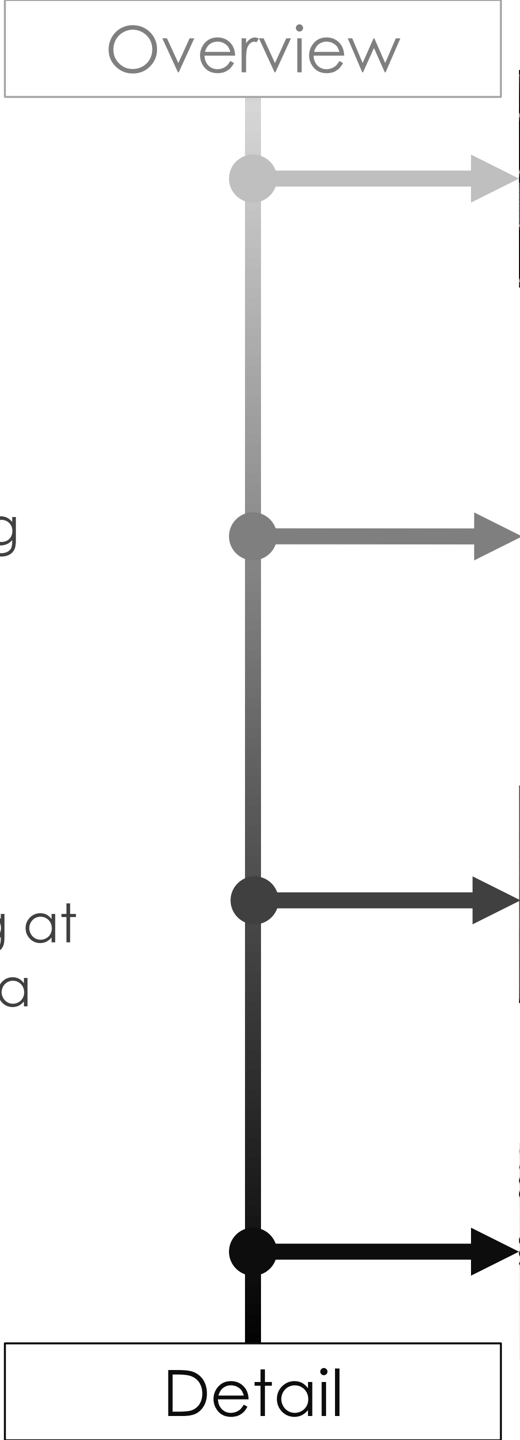
"Come with me." Sandor Clegane reached down a hand. "We have to get away from here, and now." Stranger tossed his head impatiently, his nostrils flaring at the scent of blood. The song was done. There was only one solitary drum, its slow monotonous beats echoing across the river like the pounding of some monstrous heart. The black sky wept, the river grumbled, men cursed and died. Arya had mud in her teeth and her face was wet. *Rain. It's only rain.*

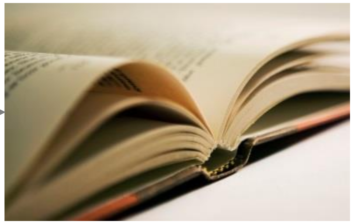
What are the different levels of detail we can look at?

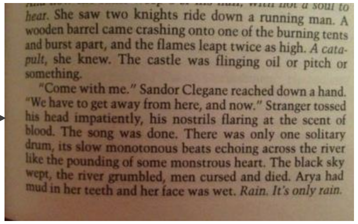
Scales

Overview:
High-level patterns looking across all the data

Detail:
Low-level patterns looking at specific pieces of the data







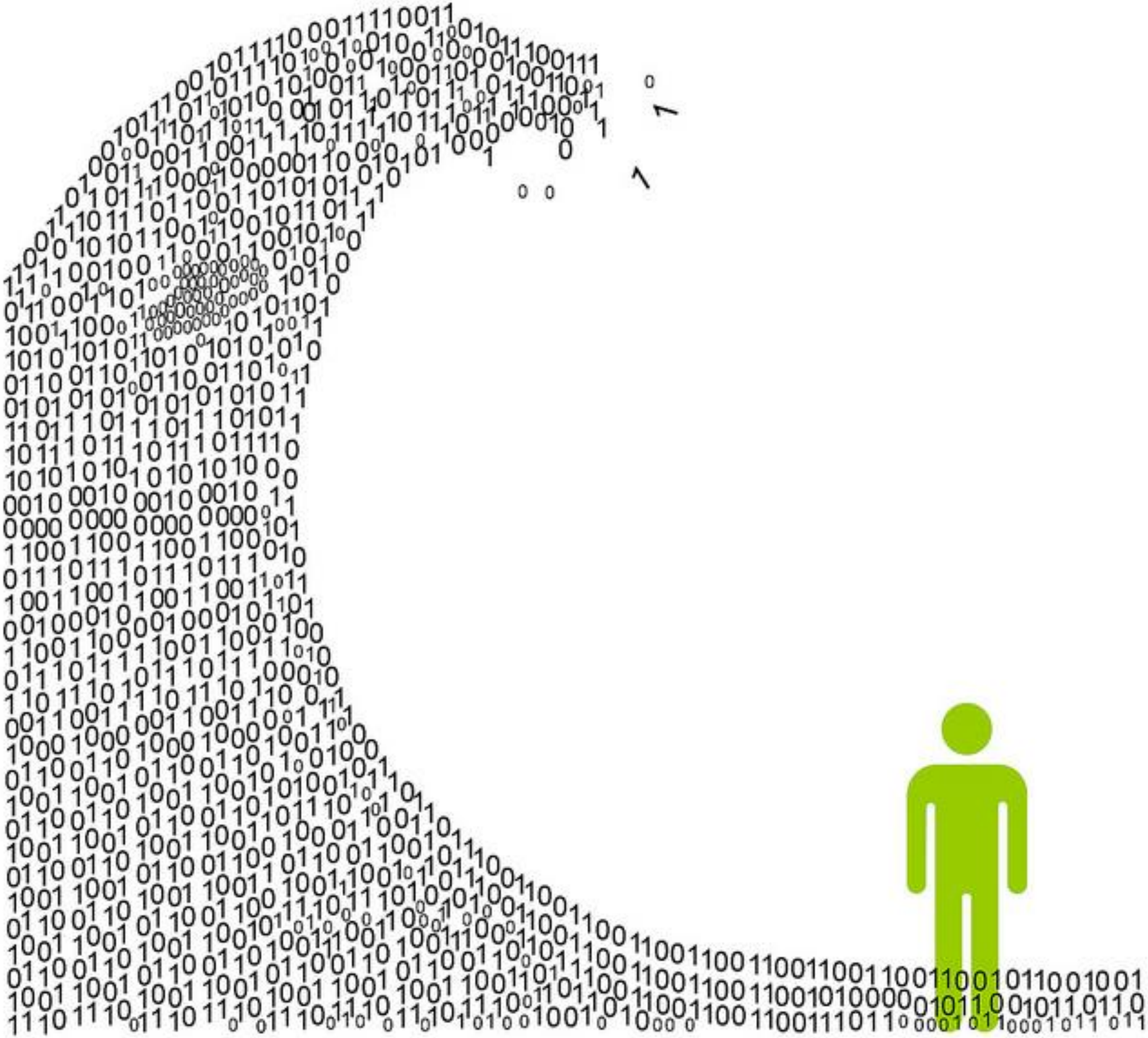
with not a soul to *hear*. She saw two knights ride down a running man. A wooden barrel came crashing onto one of the burning tents and burst apart, and the flames leapt twice as high. *A catapult*, she knew. The castle was flinging oil or pitch or something.

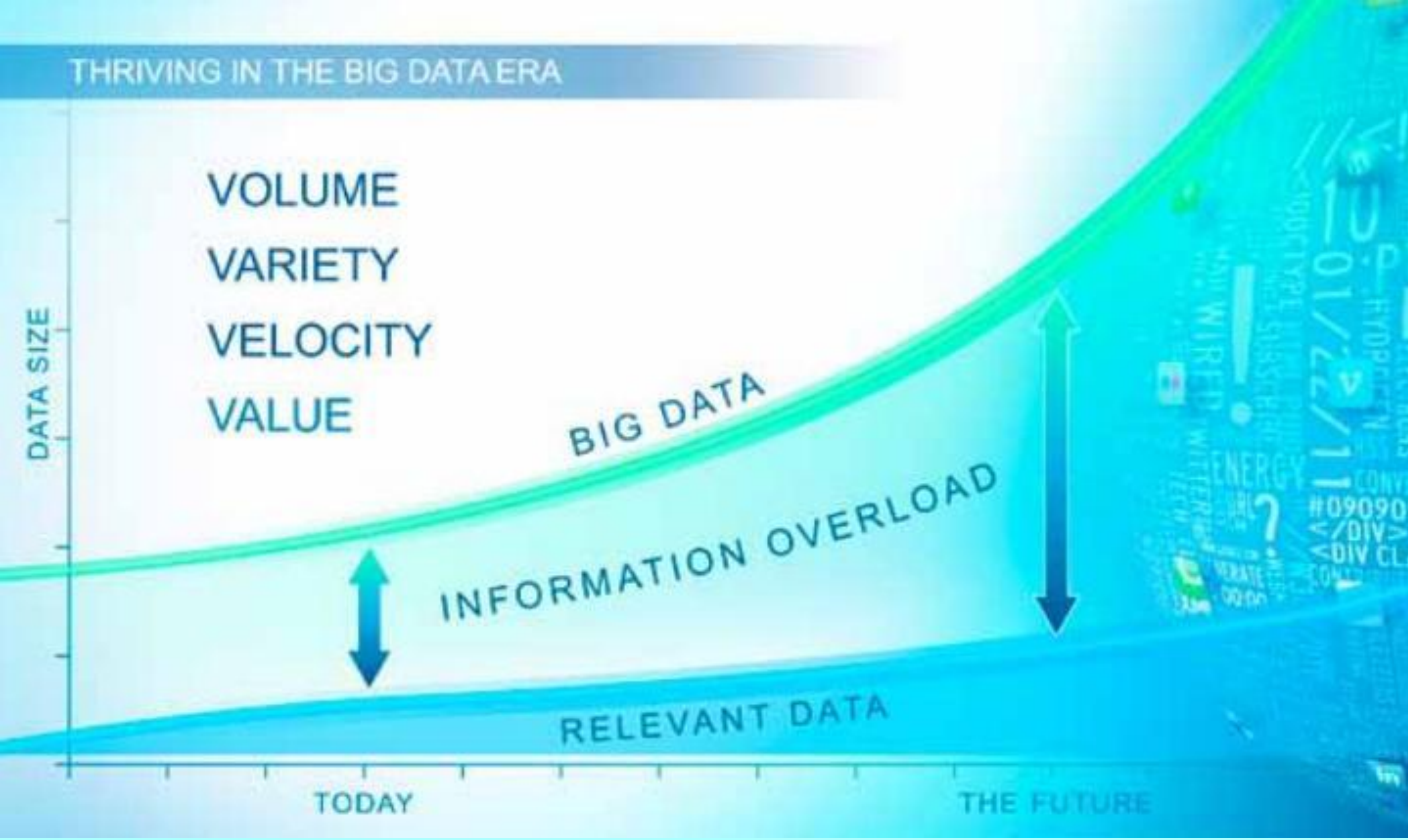
"Come with me." Sandor Clegane reached down a hand. "We have to get away from here, and now." Stranger tossed his head impatiently, his nostrils flaring at the scent of blood. The song was done. There was only one solitary drum, its slow monotonous beats echoing across the river like the pounding of some monstrous heart. The black sky wept, the river grumbled, men cursed and died. Arya had mud in her teeth and her face was wet. *Rain. It's only rain.*

How to categorize data

How to computationally explore data

How to visually explore data

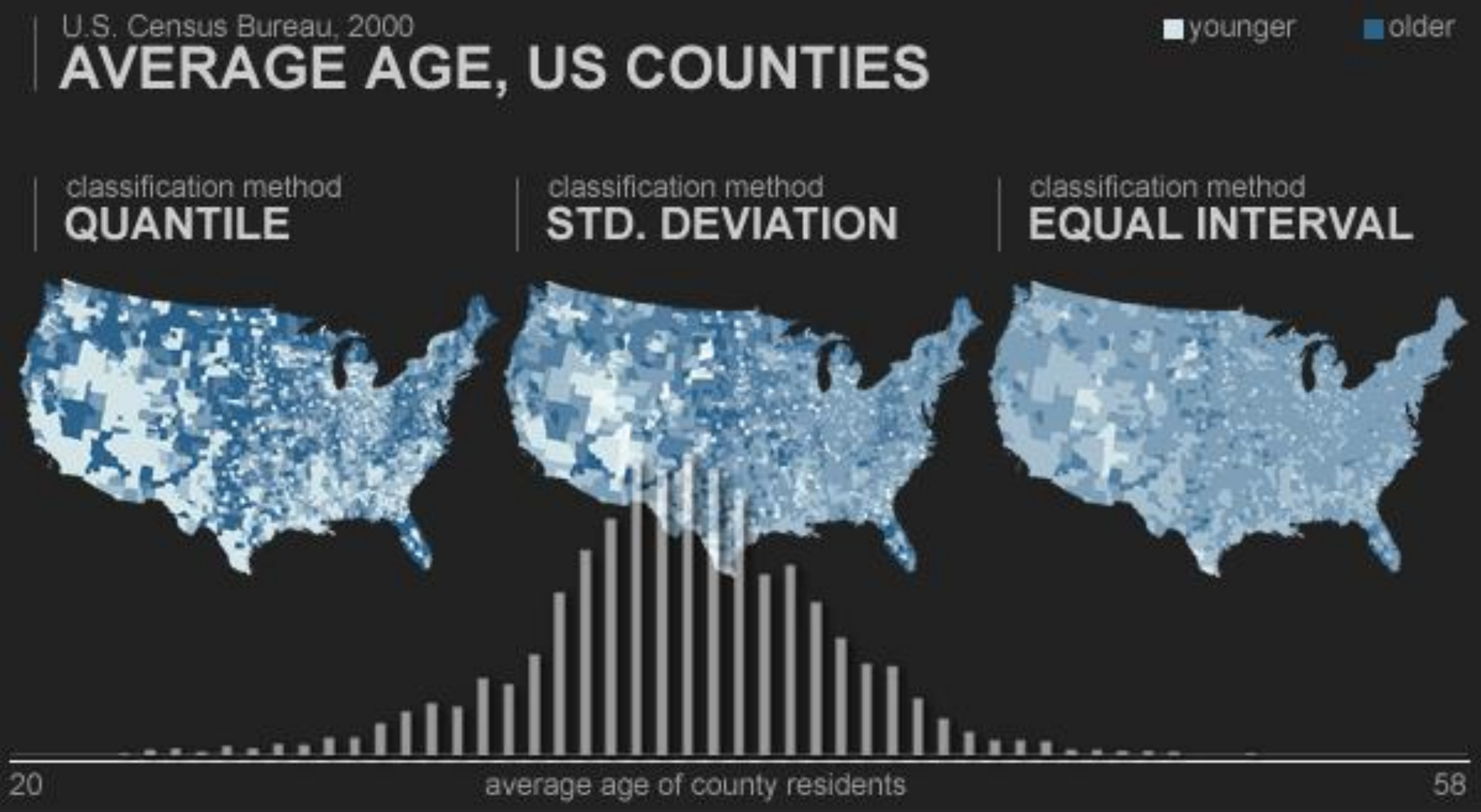




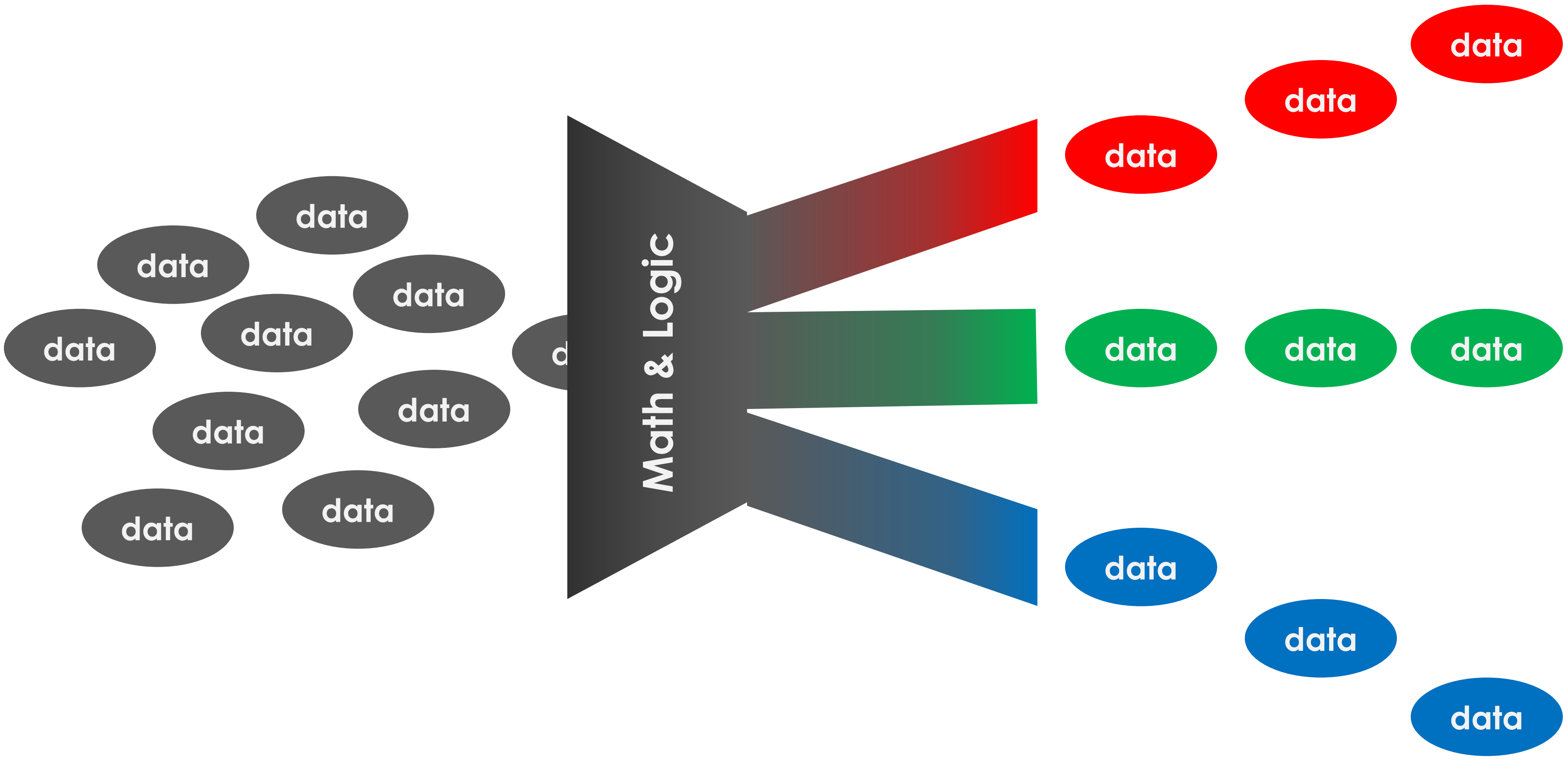
Reduce the dataset using mathematics and logic

All models are wrong, but some are useful.

--George E. P. Box



Use statistics to group the data into manageable units



Algorithmically categorize dataset based on properties of the data

Topic Models:

Identify words that categorize groups of texts in a corpus

Clustering:

Identify groups of datapoints with similar properties

Bayes Nets:

Compute how likely it is that a text belongs to different groups based on its properties

Explainers:

Determine how similar different texts are to an example text

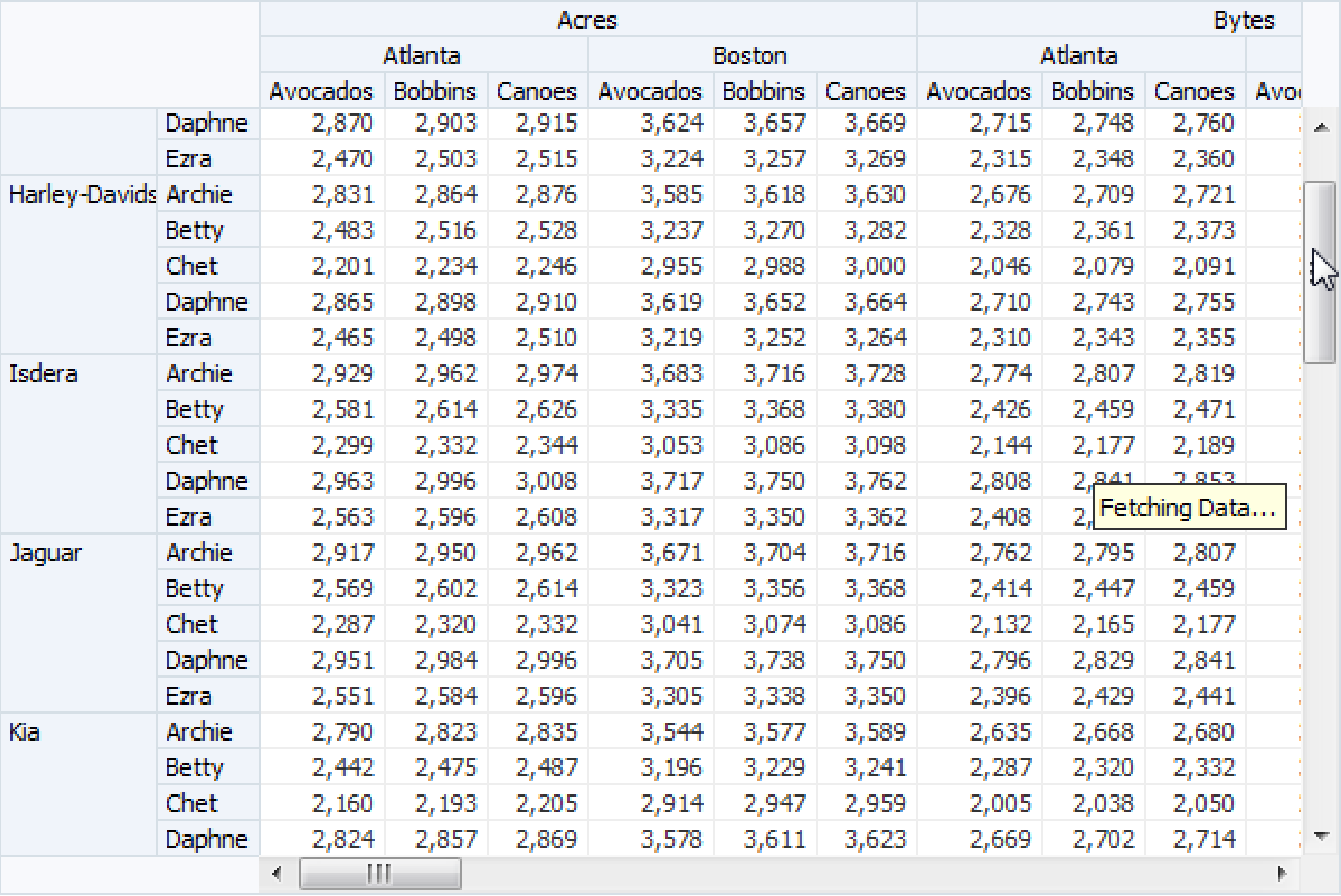
How to categorize data

How to computationally explore data

How to visually explore data

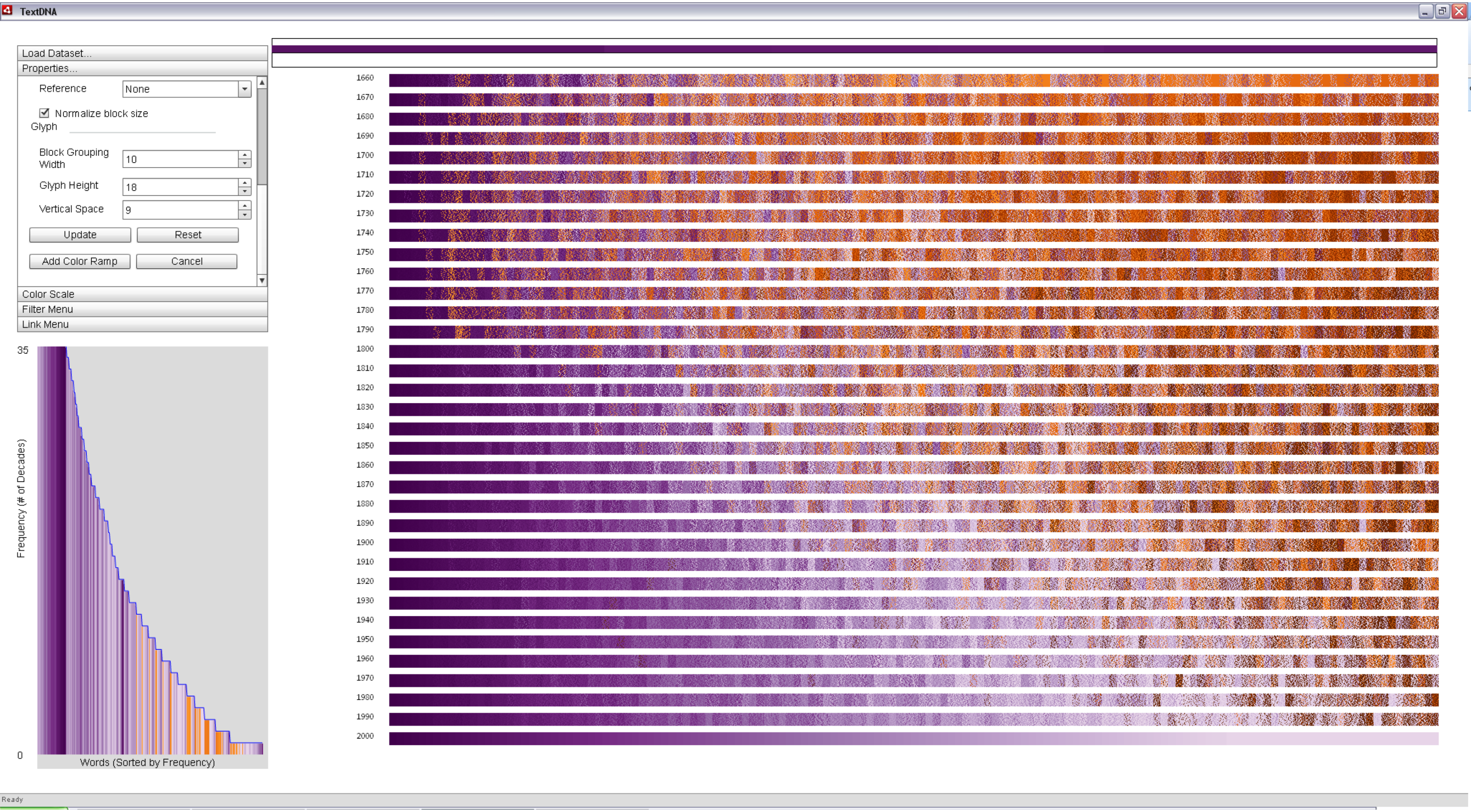
You need statistics to describe data, but then visualization to see it in context.

-- *Andy Kirk*



		Acres						Bytes			
		Atlanta			Boston			Atlanta			
		Avocados	Bobbins	Canoes	Avocados	Bobbins	Canoes	Avocados	Bobbins	Canoes	Avo
	Daphne	2,870	2,903	2,915	3,624	3,657	3,669	2,715	2,748	2,760	
	Ezra	2,470	2,503	2,515	3,224	3,257	3,269	2,315	2,348	2,360	
Harley-Davids	Archie	2,831	2,864	2,876	3,585	3,618	3,630	2,676	2,709	2,721	
	Betty	2,483	2,516	2,528	3,237	3,270	3,282	2,328	2,361	2,373	
	Chet	2,201	2,234	2,246	2,955	2,988	3,000	2,046	2,079	2,091	
	Daphne	2,865	2,898	2,910	3,619	3,652	3,664	2,710	2,743	2,755	
	Ezra	2,465	2,498	2,510	3,219	3,252	3,264	2,310	2,343	2,355	
Isdera	Archie	2,929	2,962	2,974	3,683	3,716	3,728	2,774	2,807	2,819	
	Betty	2,581	2,614	2,626	3,335	3,368	3,380	2,426	2,459	2,471	
	Chet	2,299	2,332	2,344	3,053	3,086	3,098	2,144	2,177	2,189	
	Daphne	2,963	2,996	3,008	3,717	3,750	3,762	2,808	2,841	2,853	
	Ezra	2,563	2,596	2,608	3,317	3,350	3,362	2,408	2,	[illegible]	
Jaguar	Archie	2,917	2,950	2,962	3,671	3,704	3,716	2,762	2,795	2,807	
	Betty	2,569	2,602	2,614	3,323	3,356	3,368	2,414	2,447	2,459	
	Chet	2,287	2,320	2,332	3,041	3,074	3,086	2,132	2,165	2,177	
	Daphne	2,951	2,984	2,996	3,705	3,738	3,750	2,796	2,829	2,841	
	Ezra	2,551	2,584	2,596	3,305	3,338	3,350	2,396	2,429	2,441	
Kia	Archie	2,790	2,823	2,835	3,544	3,577	3,589	2,635	2,668	2,680	
	Betty	2,442	2,475	2,487	3,196	3,229	3,241	2,287	2,320	2,332	
	Chet	2,160	2,193	2,205	2,914	2,947	2,959	2,005	2,038	2,050	
	Daphne	2,824	2,857	2,869	3,578	3,611	3,623	2,669	2,702	2,714	

Fetching Data...



Reference: None

☑ Normalize block size

Block Grouping Width: 10

Glyph Height: 18

Vertical Space: 9

Visualizations let us explore and communicate large amounts of data visually

1) Visually encode the data

2) Arrange the encoded data to highlight patterns of interest

3) Design complementary methods for looking at the data that can answer complex analysis questions

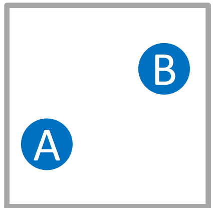
4) Design ways for interacting with the encoded data that support your analysis

1) Visually encode the data

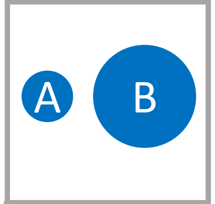
2) Arrange the encoded data to highlight patterns of interest

3) Design complementary methods for looking at the data that can answer complex analysis questions

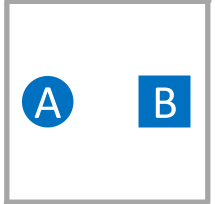
4) Design ways for interacting with the encoded data that support your analysis



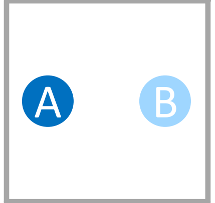
Position



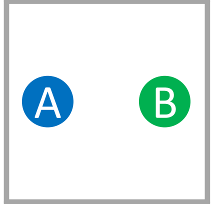
Size



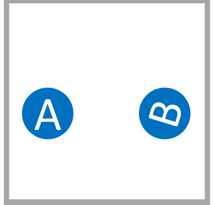
Shape



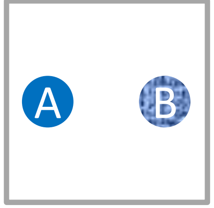
Value/Lightness



Color



Orientation

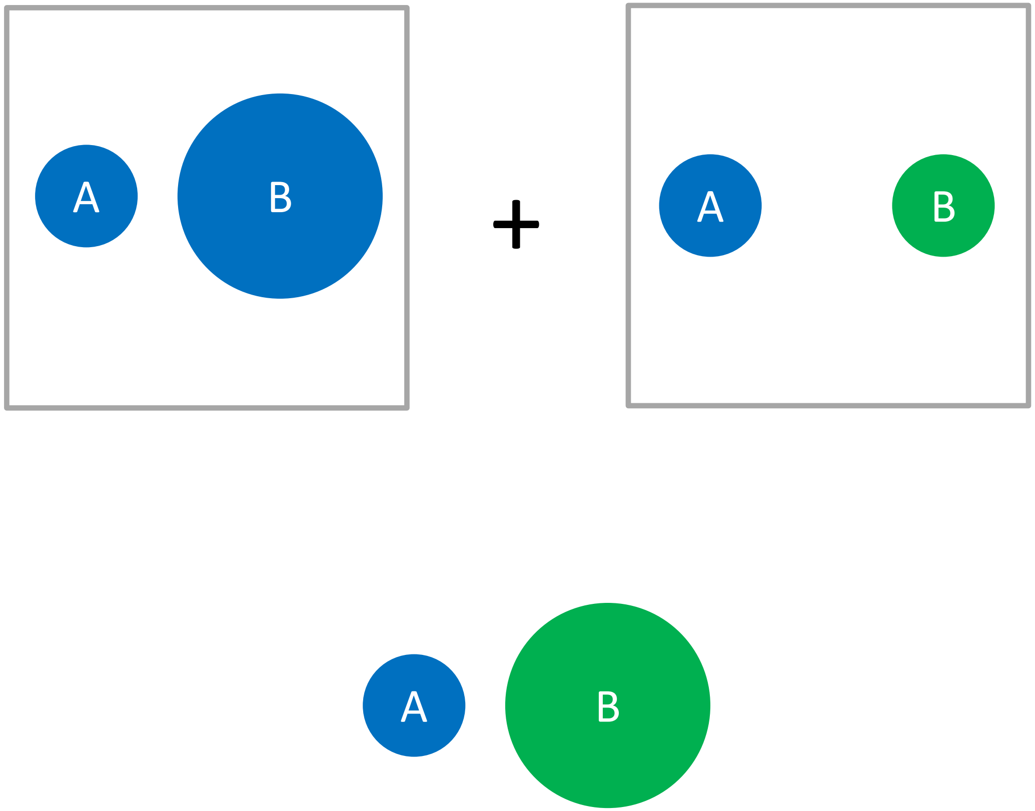


Texture

Visual Encodings:

Ways to map data values to visual marks

Different visual encodings highlight different properties in the data



Encodings can be combined to communicate multiple properties of the data

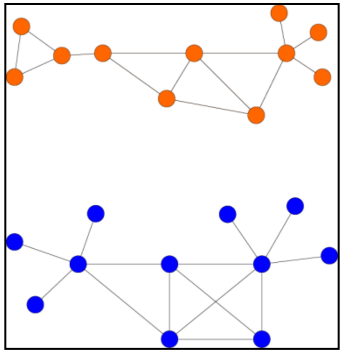
1) Visually encode the data

2) Arrange the encoded data to highlight patterns of interest

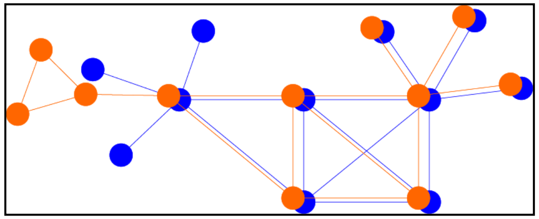
3) Design complementary methods for looking at the data that can answer complex analysis questions

4) Design ways for interacting with the encoded data that support your analysis

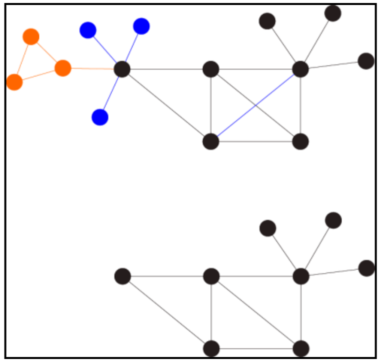
Once data is encoded, we can highlight relationships in the data by:



Juxtapositioning encoded data side-by-side



Superpositioning encoded data in the same space

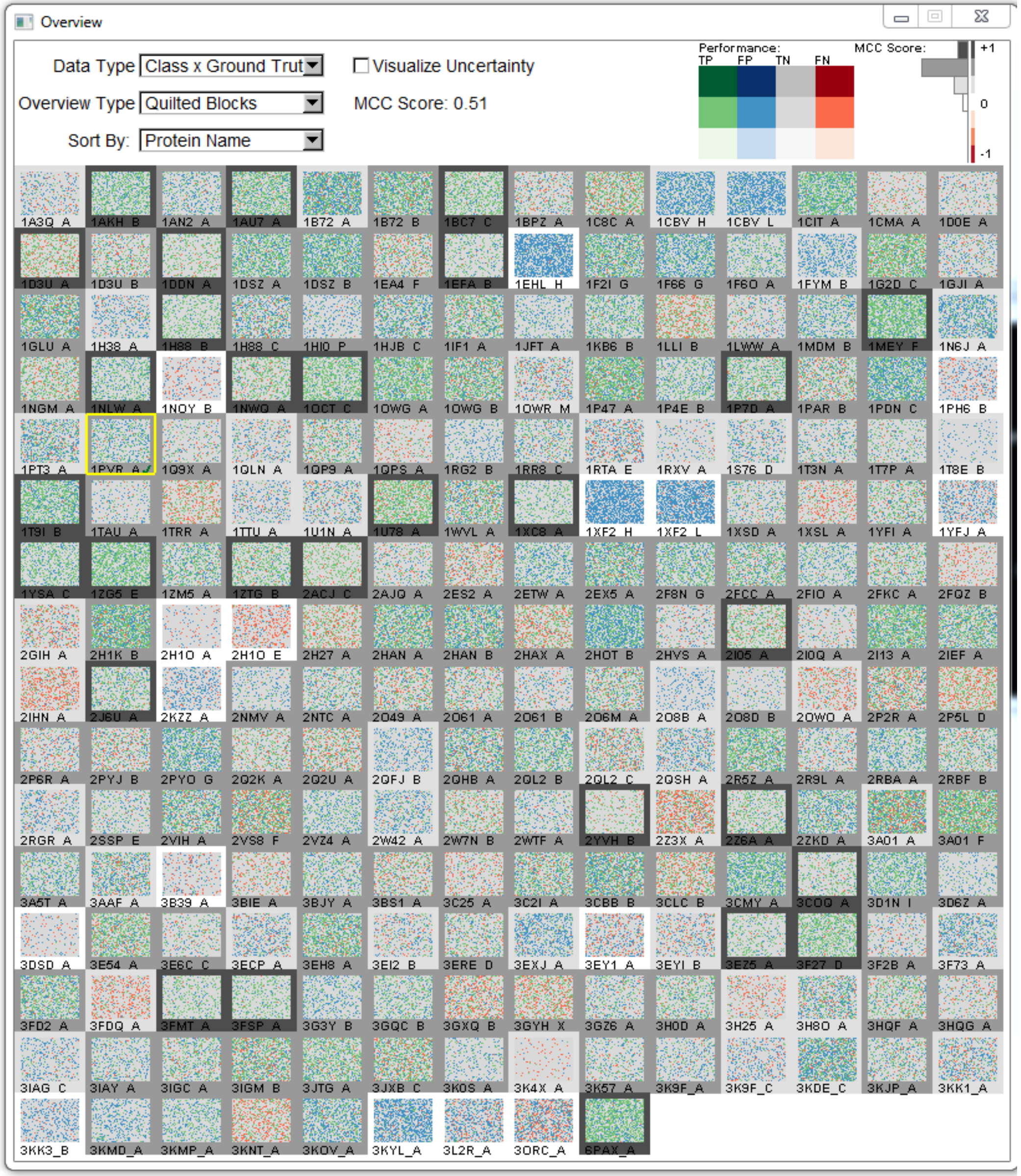


Explicitly encoding relationships of interest between datapoints

Small Multiples:

Juxtapose large numbers of small visualizations to communicate high-level patterns

Can either subdivide the data or properties of the data



Data Type Class x Ground Trut

Overview Type Quilted Blocks

Sort By: Protein Name

☐ Visualize Uncertainty

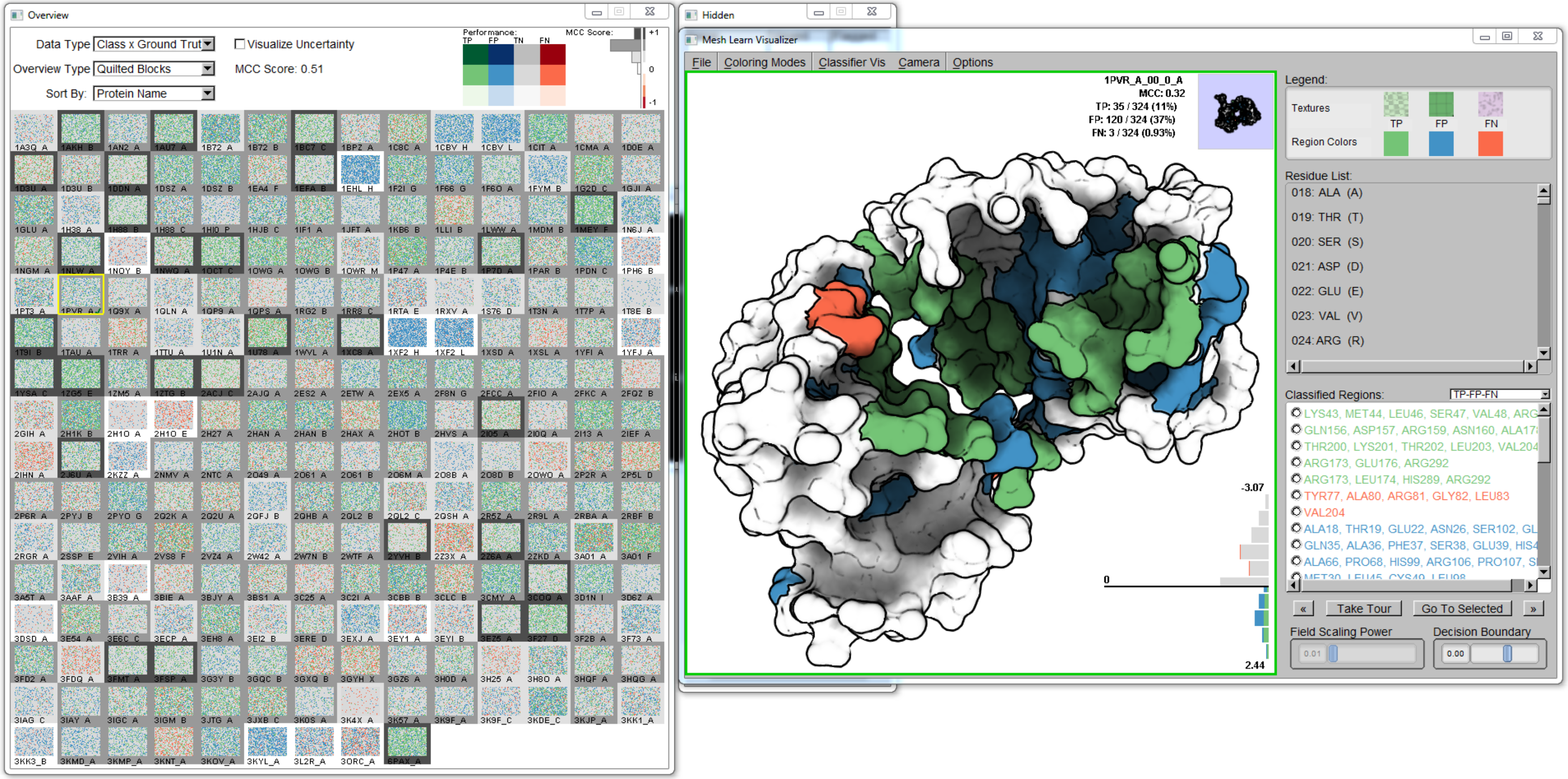
MCC Score: 0.51

1) Visually encode the data

2) Arrange the encoded data to highlight patterns of interest

3) Design complementary methods for looking at the data that can answer complex analysis questions

4) Design ways for interacting with the encoded data that support your analysis



Overview

Data Type: Class x Ground Trut

☐ Visualize Uncertainty

Overview Type: Quilted Blocks

MCC Score: 0.51

Sort By: Protein Name

Hidden

Mesh Learn Visualizer

1PVR_A_00_0_A
MCC: 0.32
TP: 35 / 324 (11%)
FP: 120 / 324 (37%)
FN: 3 / 324 (0.93%)

Residue List:
018: ALA (A)
019: THR (T)
020: SER (S)
021: ASP (D)
022: GLU (E)
023: VAL (V)
024: ARG (R)

Classified Regions: TP-FP-FN

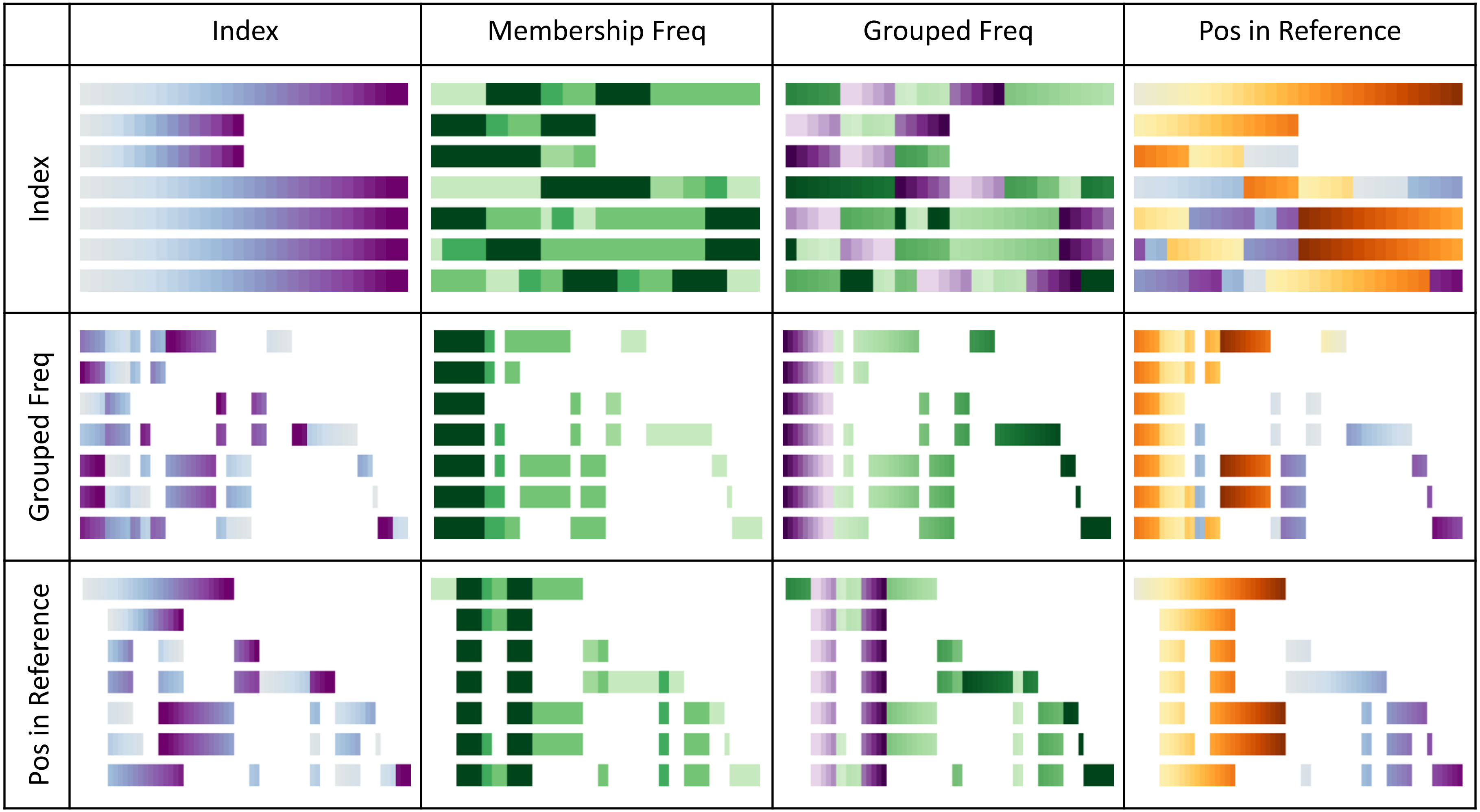
○ LYS43, MET44, LEU46, SER47, VAL48, ARG
○ GLN156, ASP157, ARG159, ASN160, ALA17
○ THR200, LYS201, THR202, LEU203, VAL204
○ ARG173, GLU176, ARG292
○ ARG173, LEU174, HIS289, ARG292
○ TYR77, ALA80, ARG81, GLY82, LEU83
○ VAL204
○ ALA18, THR19, GLU22, ASN26, SER102, GL
○ GLN35, ALA36, PHE37, SER38, GLU39, HIS4
○ ALA66, PRO68, HIS99, ARG106, PRO107, S
○ MET30, LEU45, CYS49, LEU98

Field Scaling Power: 0.01

Decision Boundary: 0.00

Coordinated Views:

Create multiple visualizations that work together to support complex analysis



Dynamic Remapping:

Allow the user to change what data maps to which visual channels to highlight different patterns

1) Visually encode the data

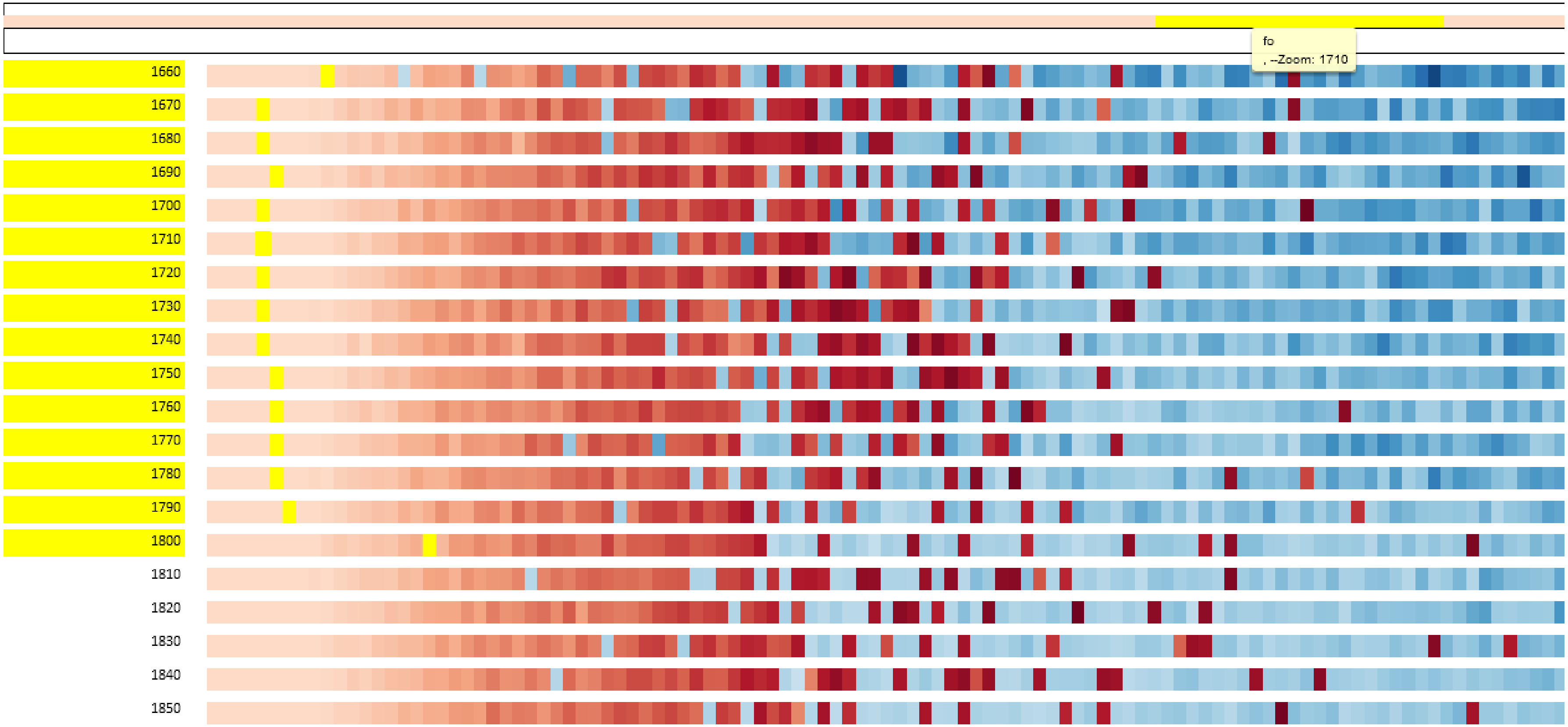
2) Arrange the encoded data to highlight patterns of interest

3) Design complementary methods for looking at the data that can answer complex analysis questions

4) Design ways for interacting with the encoded data that support your analysis

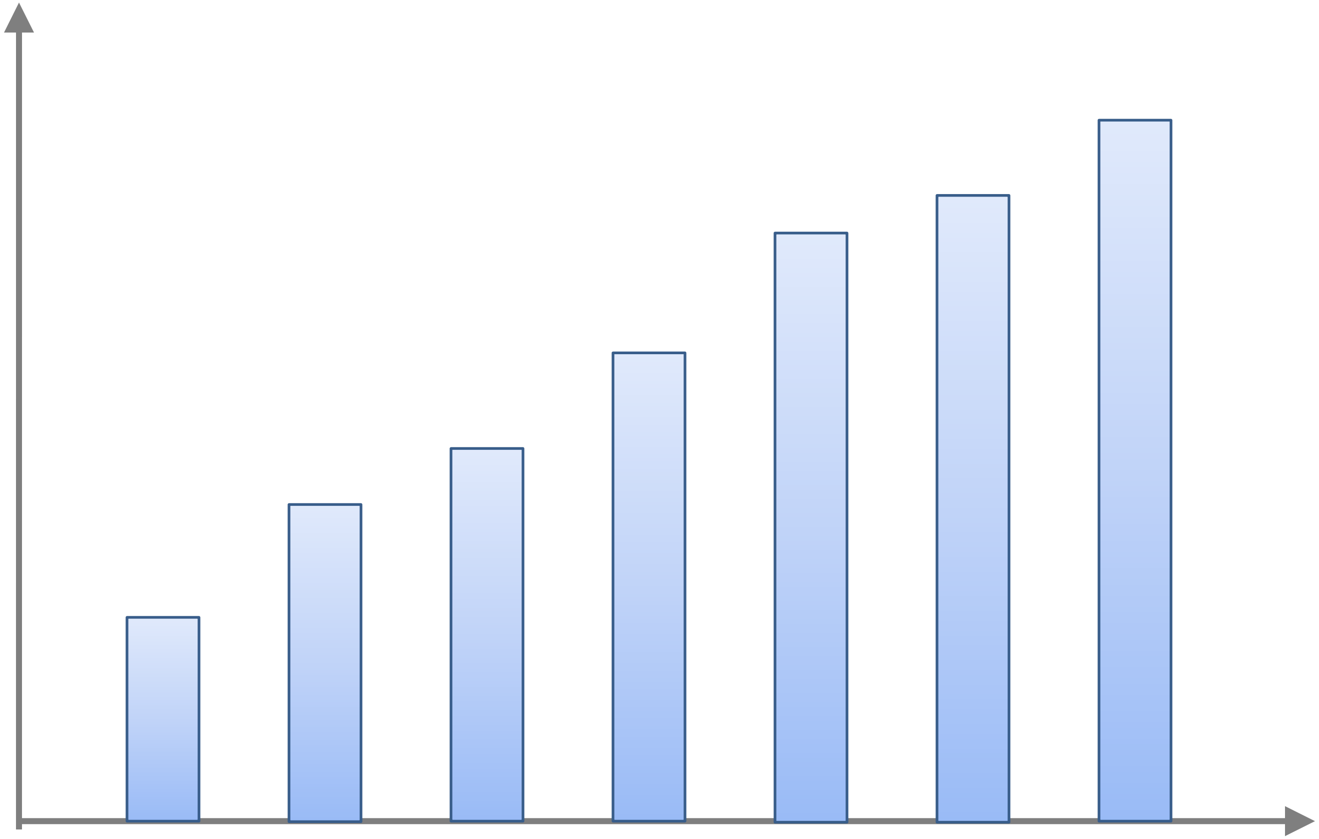
Always connect back to the person:
how can we make insights meaningful?

Interaction



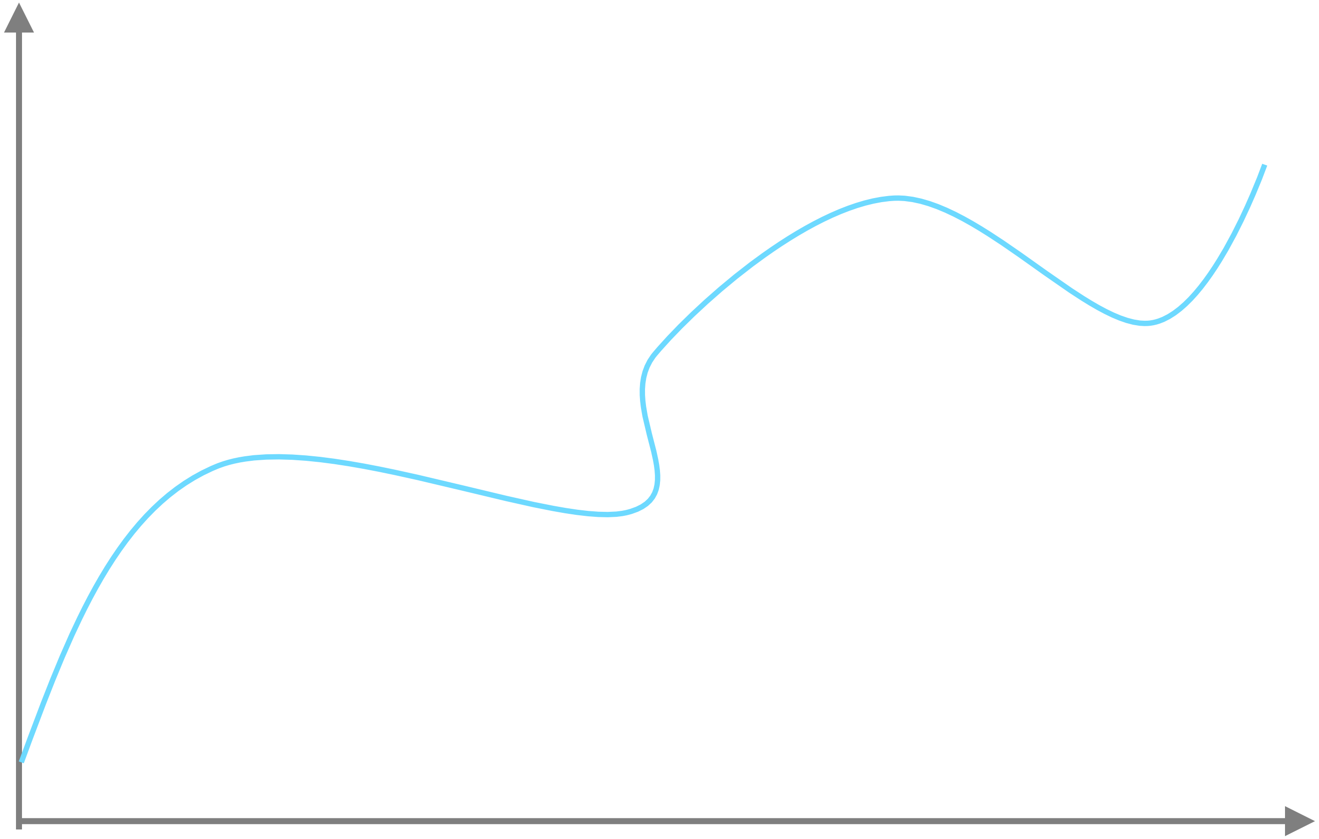
Some techniques for visualizing data…

Bar Charts



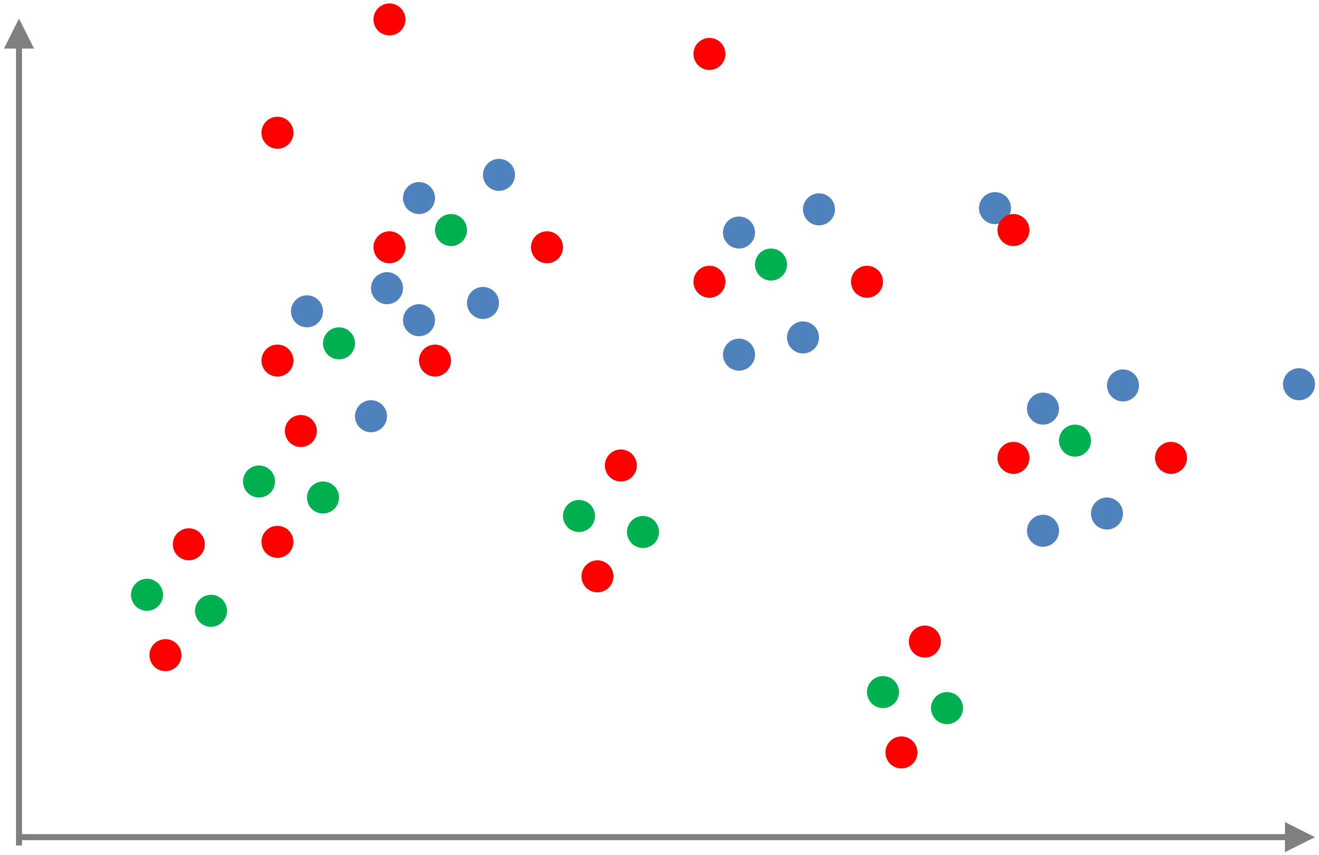
Compare values

Line Graphs



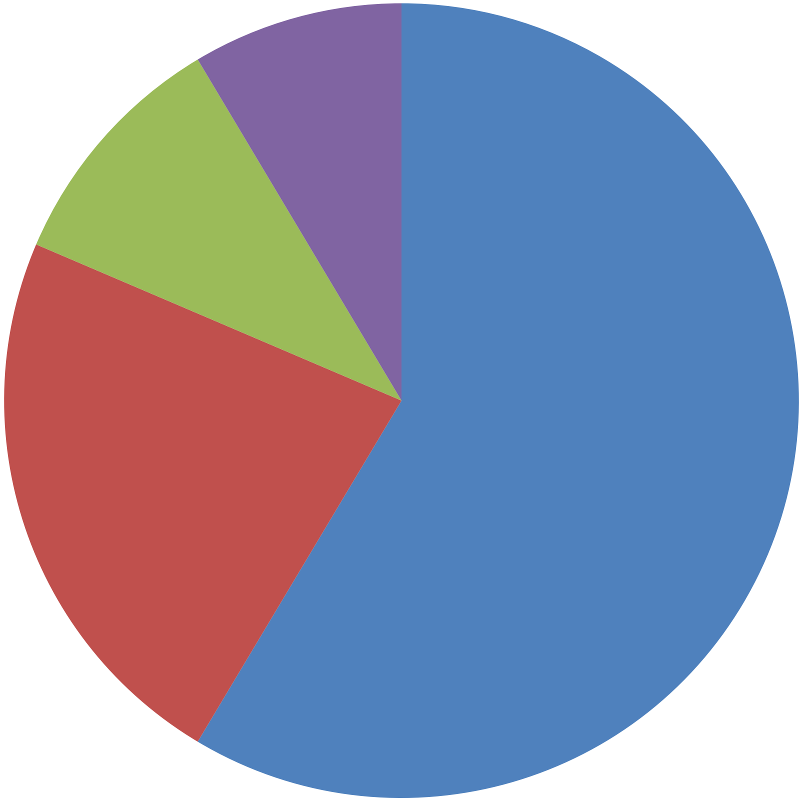
Identify trends

Scatterplots



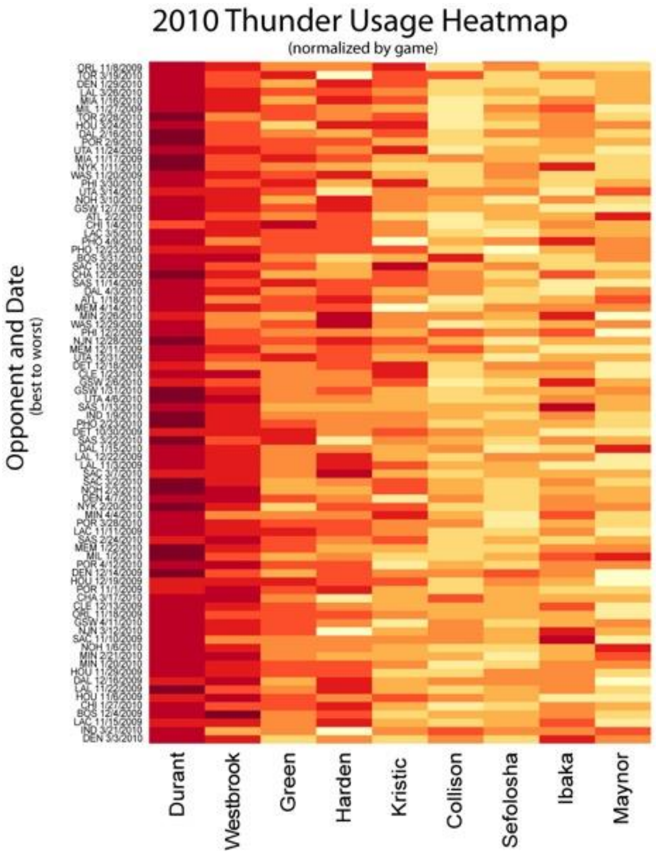
Identify clusters

Pie Charts



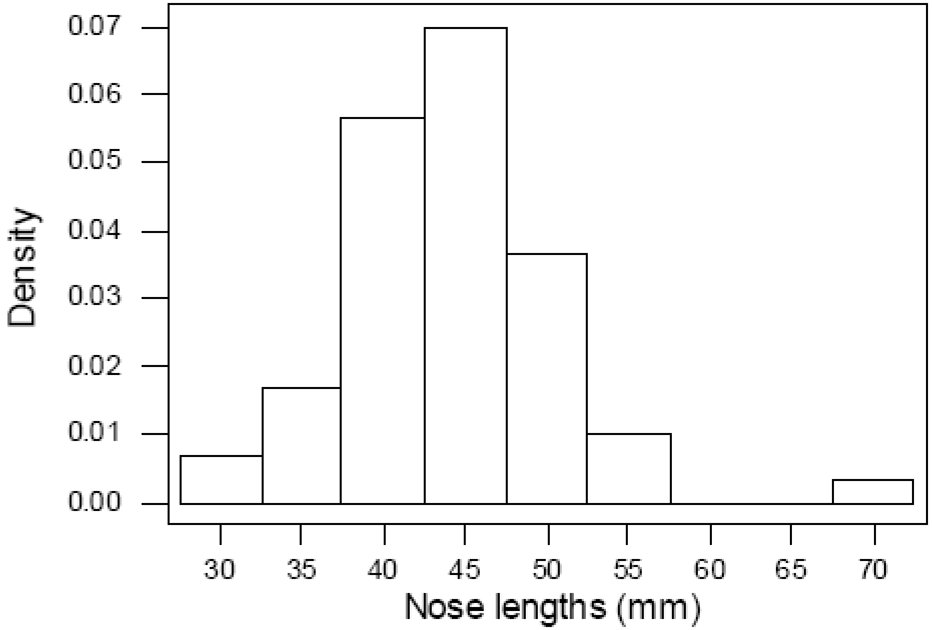
Communicate proportions of a whole

Heatmaps



Color to convey values compactly

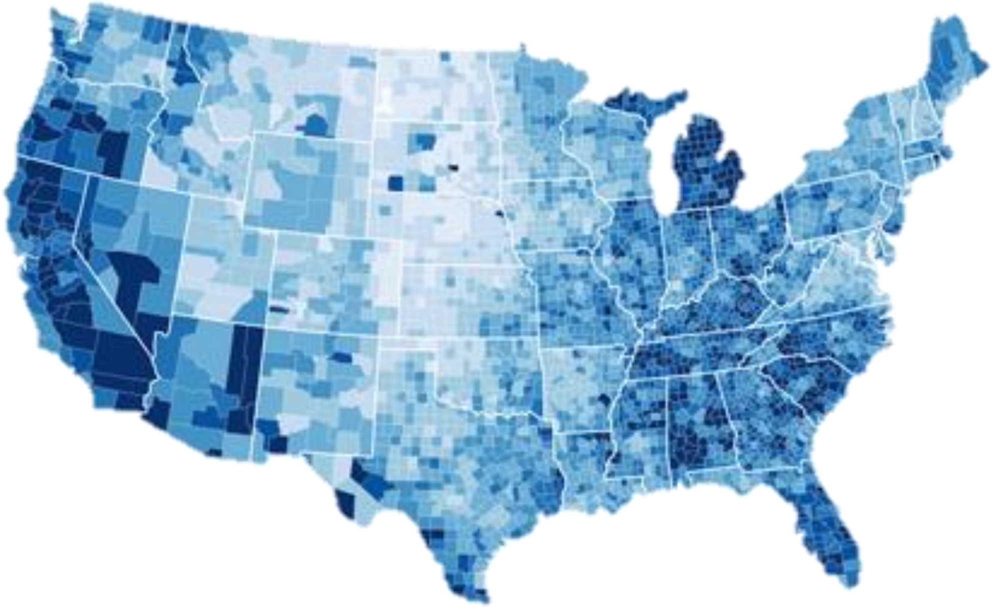
Histogram

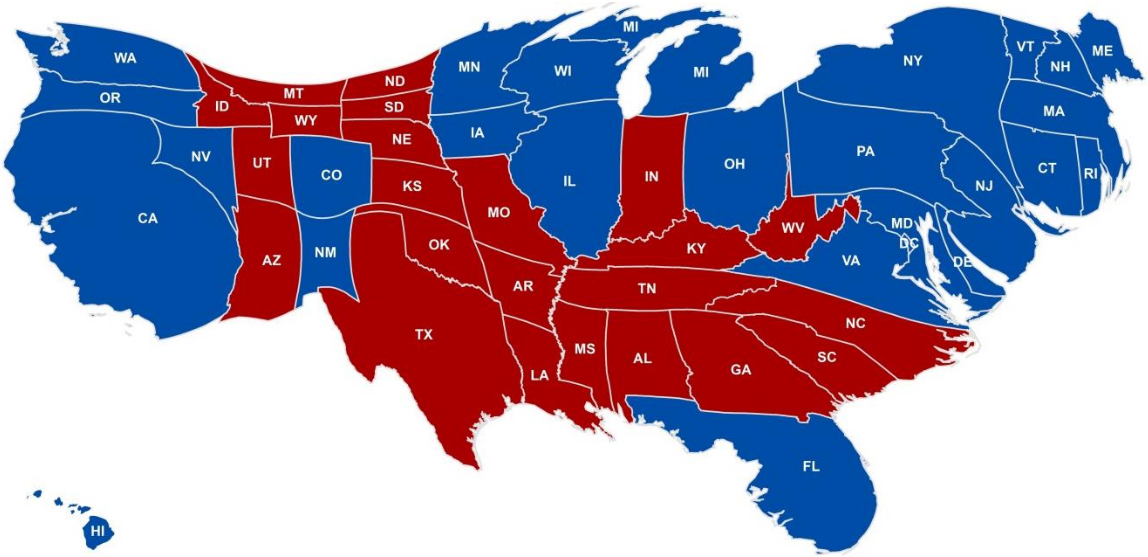


Distribution over different properties

Choropleths:

Color to convey values

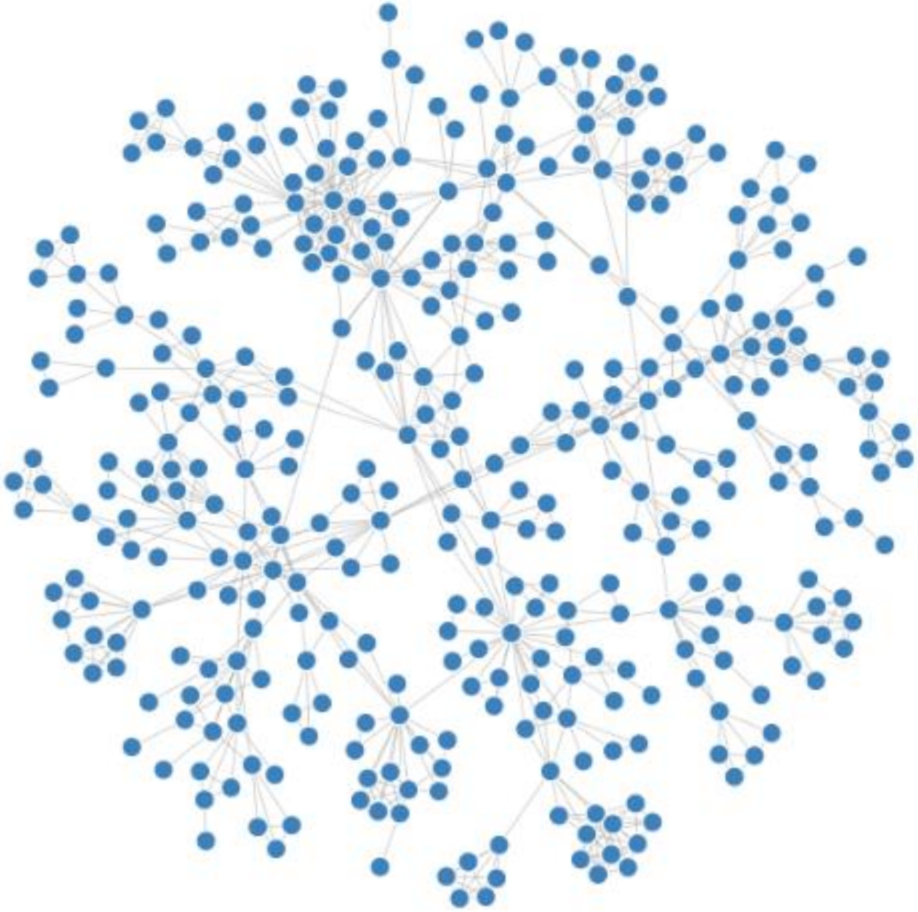




Cartograms:

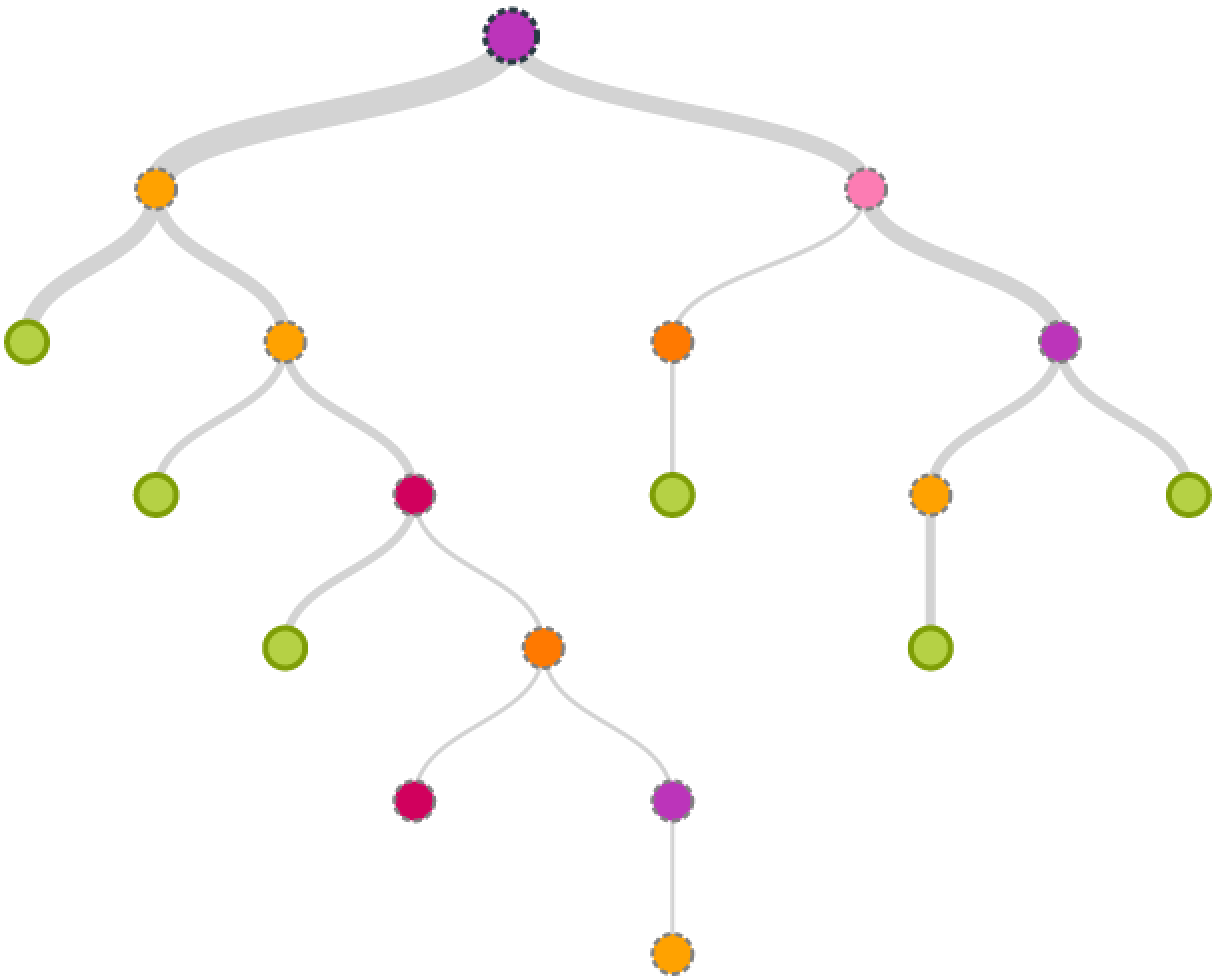
Size to convey values

Networks/Node-Link Diagrams



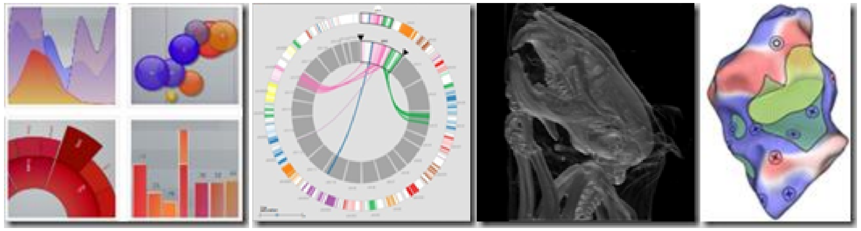
Connect related objects

Trees & Hierarchies



Communicate hierarchical relationships

Learn More about Visualization



CS638/838: Visualization
Prof. Michael Gleicher
11:00-12:15 Tu/Th

Visualization Reading Group
2pm every other Thursday

Visualization
at UW-Madison

1) Break into groups—mix “techies” and “humanists”

2) Pick one dataset from your group to talk about

3) Sketch how you might approach analyzing this data

4) Rinse and repeat

5) Group critique

What are the different properties of the data?

What are the interesting relationships between these properties and why?

What are common or informative labels that can describe different aspects of the data?

What, if any, questions do you want to explore in the data?

What levels of detail are interesting?

What would be some interesting ways to “look” at this data?

What patterns (or lack thereof) would you hope to find in this data and what would they mean?