

Data Visualization: What Can you See In Your Data?



Prof. S. George Djorgovski

Astronomy and Center for Data-Driven Discovery, Caltech

Lecture 2

XXX Canary Islands Winter School

November 2018

Caltech

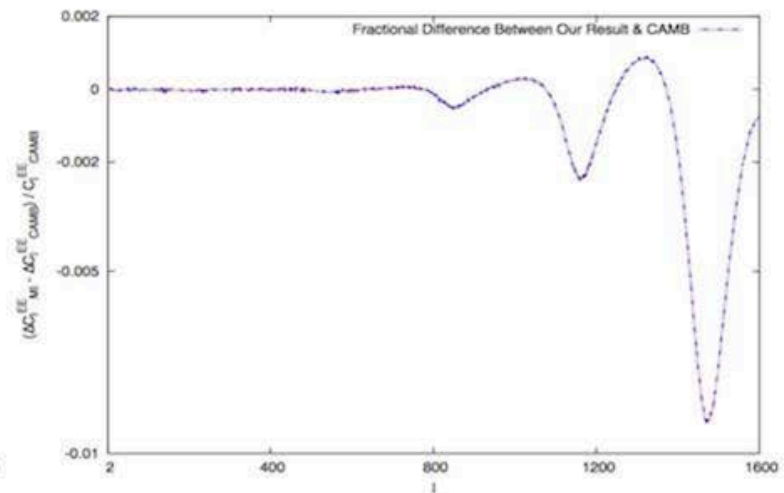
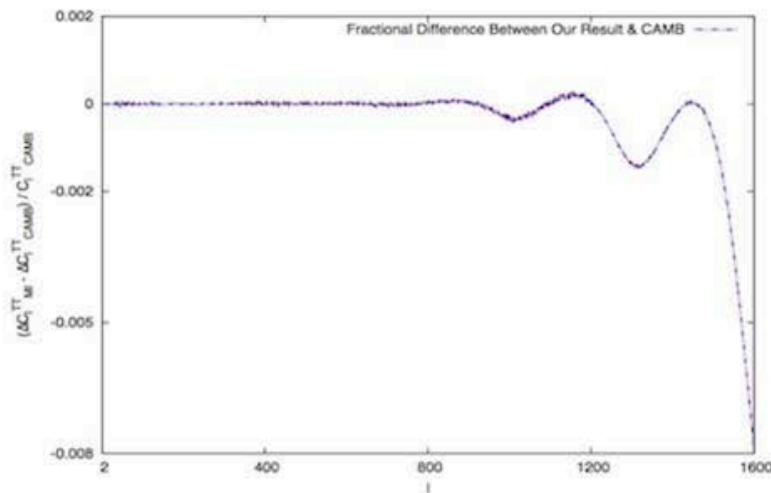
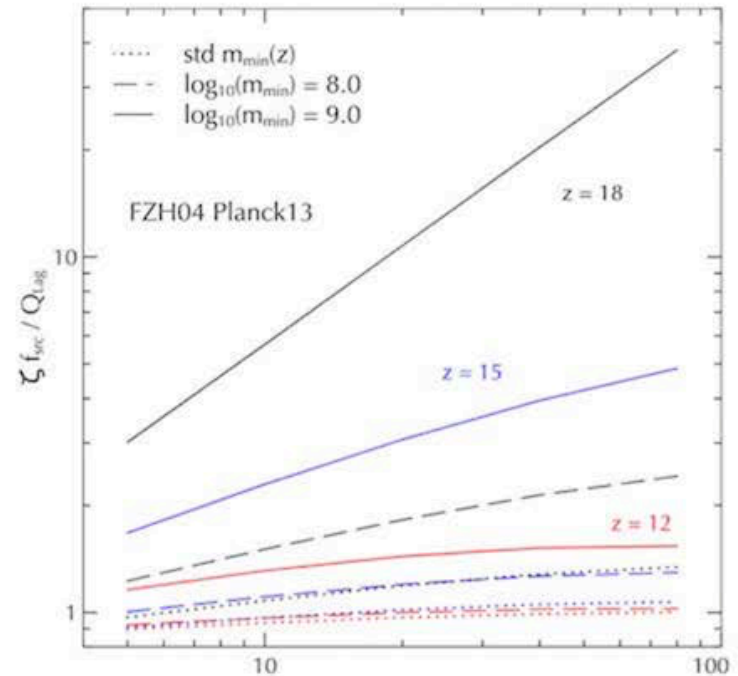


CENTER FOR DATA-DRIVEN DISCOVERY

Never Do This!

A figure made for a print may not look good on the screen:
Paper ~ 5000 by ~ 6500 pixels
Powerpoint usually 768 by 1024

Figure axes and labels must be legible: use a large font



We consume information, and information consumes us

What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Herb Simon

Scientific American, 1995

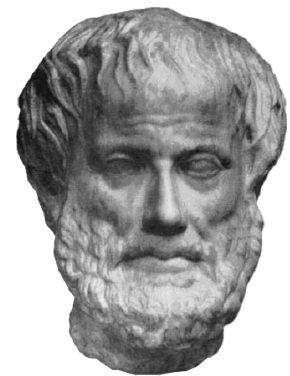
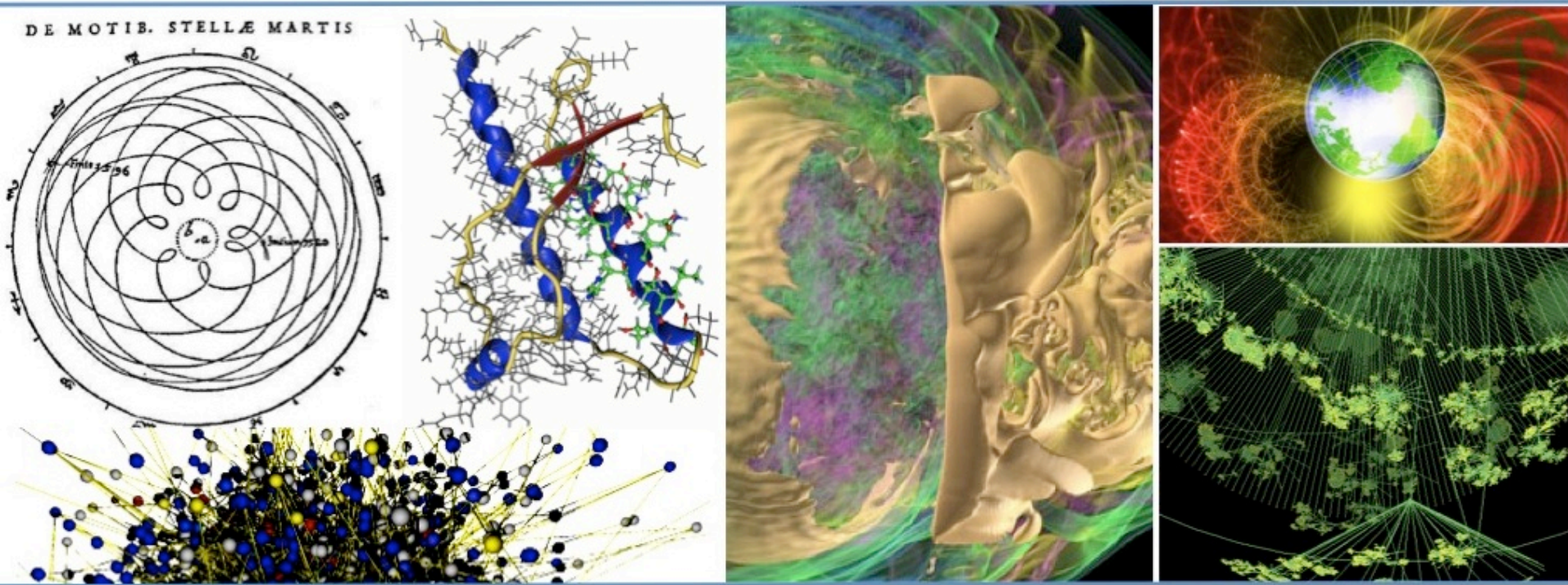
What are the computers for?

Increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems.

Douglas Engelbart

Augmenting Human Intellect:
A Conceptual Framework

Effective visualization is the bridge between quantitative information and human intuition



Man cannot understand without images

Aristotle, *De Memoria et Reminiscentia*

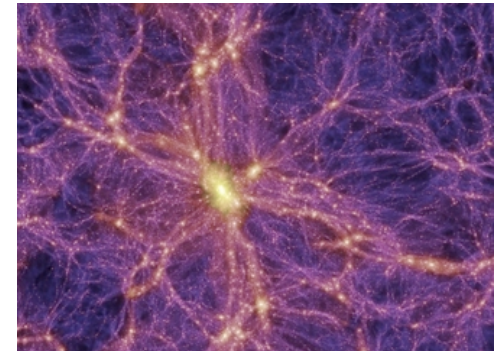
You can observe a lot just by watching

Yogi Berra, an American philosopher



A Key Challenge: Visualizing Complexity

- Hyperdimensional structures (clusters, correlations, etc.) are likely present in many complex data sets, whose dimensionality is commonly in the range of $D \sim 10^2 - 10^4$, and will surely grow
- It is not only the matter of ***data understanding***, but also of choosing the appropriate data mining algorithms, and interpreting the results
- We are biologically limited to perceiving 3 - 12(?) dimensions



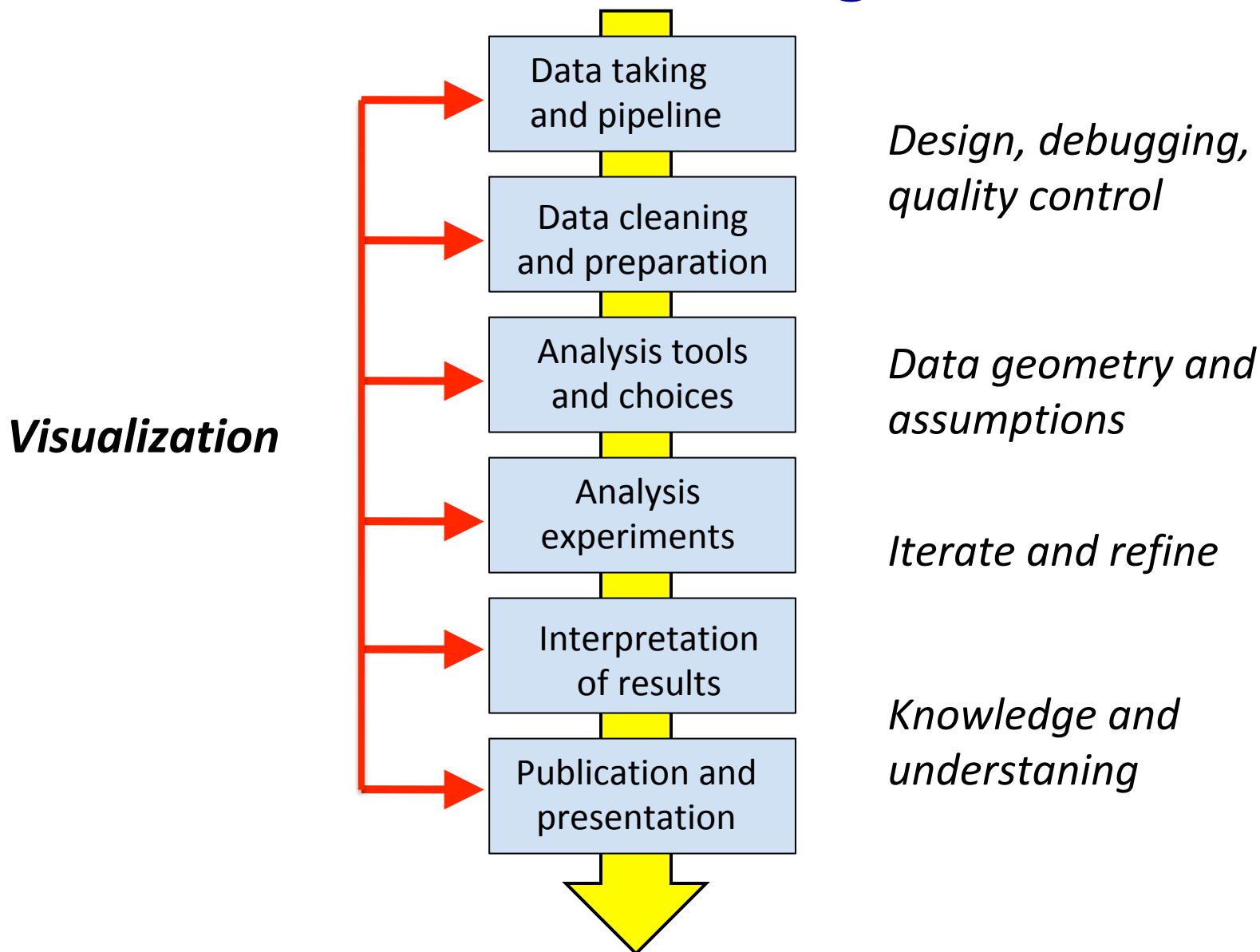
What good are the data if we cannot effectively extract knowledge from them?

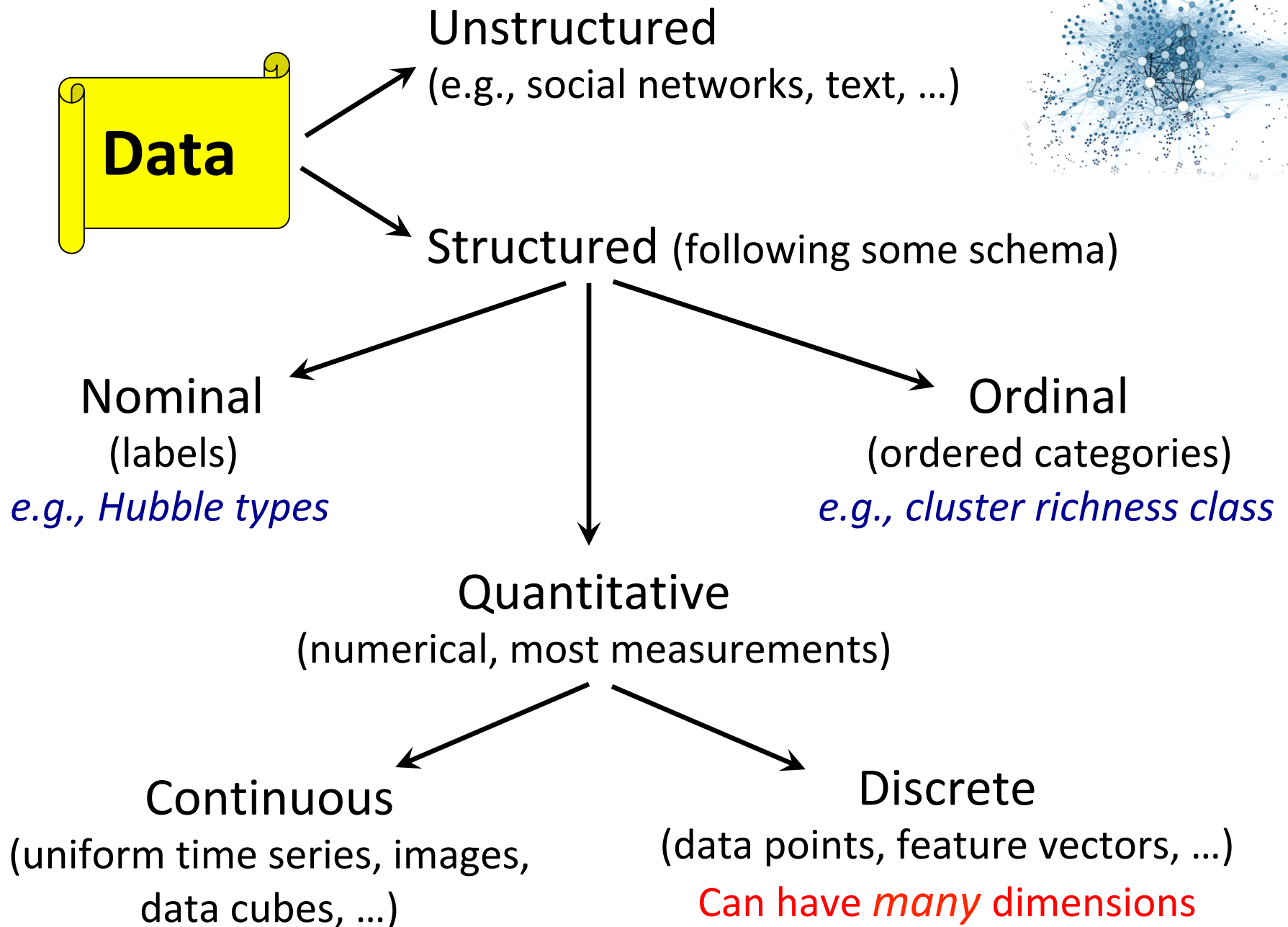
“A man has got to know his limitations”

Dirty Harry, another American philosopher

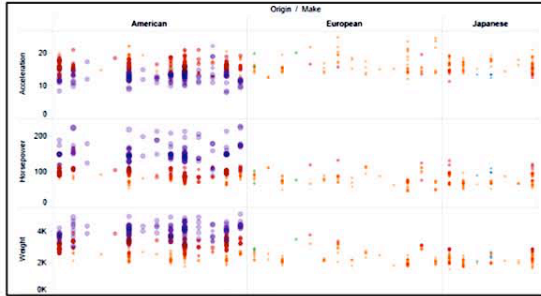
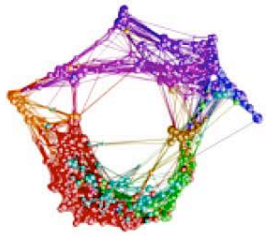


Visualization is an Essential Component of the Entire Data-to-Knowledge Process

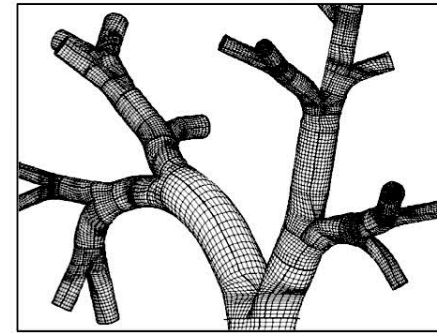




Geometric Structure of Data

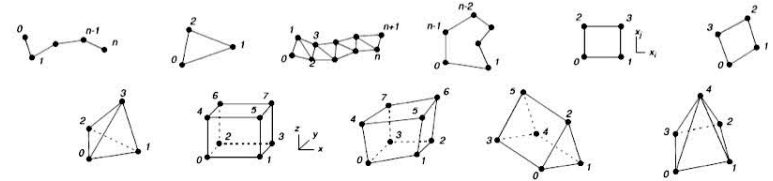


UNSTRUCTURED
ABSTRACT DATA
multi-dimensional data RECORDS



STRUCTURED
2D/3D DATA
scalar/vector/tensor + time

MPG	Cylinders	Horsepower	Weight	Acceleration	Year	Origin
8.50	4.2	8.2	4.4	40.2	50.4	1500.5
8.00	4.0	8.0	4.0	25.0	4.4	1500.5
15.00	8.0	165.0	3693.0	11.5	70.0	1.0
18.00	8.0	150.0	3433.0	12.0	70.0	1.0
17.00	8.0	140.0	3449.0	10.5	70.0	1.0
15.00	8.0	165.0	3693.0	11.5	70.0	1.0
18.00	8.0	150.0	3436.0	11.0	70.0	1.0
16.00	8.0	150.0	3433.0	12.0	70.0	1.0
17.00	8.0	140.0	3449.0	10.5	70.0	1.0
15.00	8.0	165.0	3693.0	11.5	70.0	1.0
18.00	8.0	150.0	3436.0	11.0	70.0	1.0
16.00	8.0	150.0	3433.0	12.0	70.0	1.0
17.00	8.0	140.0	3449.0	10.5	70.0	1.0
16.00	8.0	150.0	3433.0	12.0	70.0	1.0

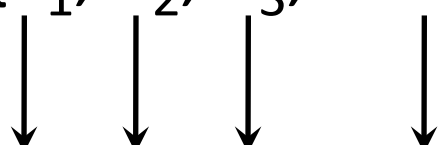


The two kinds can
 be interchangeable

From Data Space to Visualization Space

If data consists of feature vectors with N independent measurements, they form an N -dimensional data space

Each of the data dimensions is mapped to one “axis” of the visualization space:

$$\text{Data} = \{x_1, x_2, x_3, \dots, x_N\}$$


Visualization space: { XYZ positions, point sizes, shapes, RGB α or HSV colors, textures, glyphs, point orientations, animations, ... }

The choice of this mapping is *critical*

Quantitative perception (visual or other)

Many senses are organized around the “just noticeable difference”

Ratio is more important than magnitude

Most continuous variation in stimuli is perceived in discrete steps

Two important visualization principles:

Principle of consistency

Properties of the image (visual encoding) should match the properties of the data

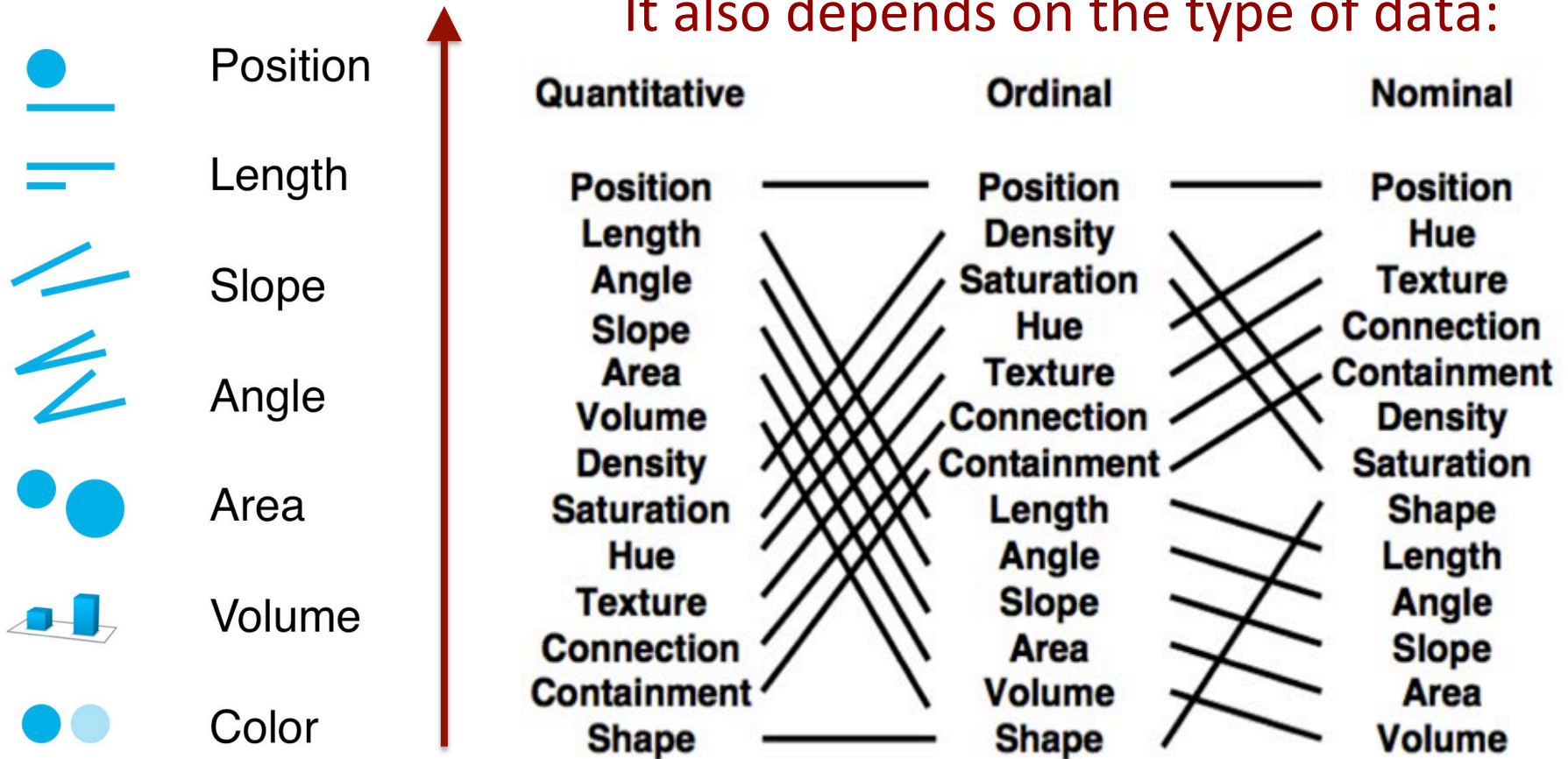
Principle of importance ordering

Encode the most important variables in the most effective way

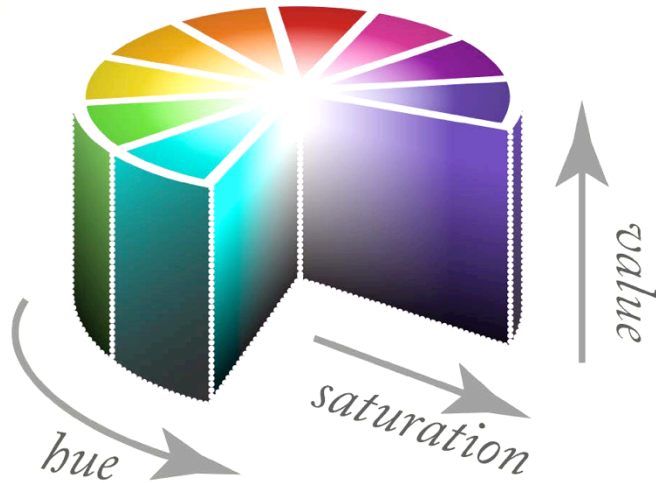
Map the most important variables to the visual “axis” that corresponds to the most accurate perception:

Increasing accuracy

It also depends on the type of data:



How Many Dimensions for Color?

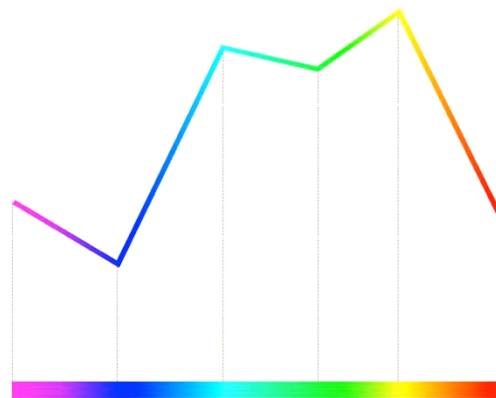
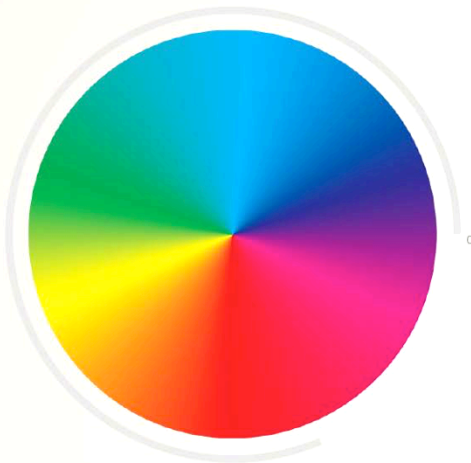


3? RGB or HSV

2? R/G, G/B

Actually, effectively only **1**

Perceptions of luminosity are different:



e.g., at a given value, yellow looks brighter than blue

(from S. Lombeyda)

How Many Shades of Gray Can you Distinguish?



How Many Shades of Gray Can you Distinguish?



Value easily encodes ordinal variables



Value encodes continuous variables (less well)

How Many Shades of Gray Can you Distinguish?



Value easily encodes ordinal variables



Value encodes continuous variables (less well)

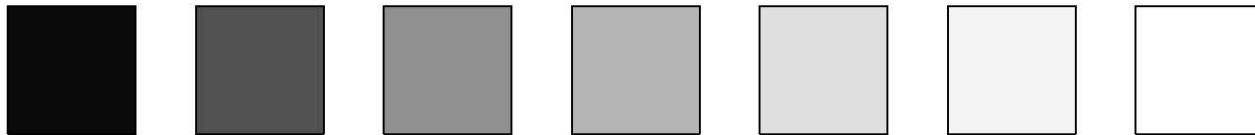
How Many Colors?



How Many Shades of Gray Can you Distinguish?



Value easily encodes ordinal variables

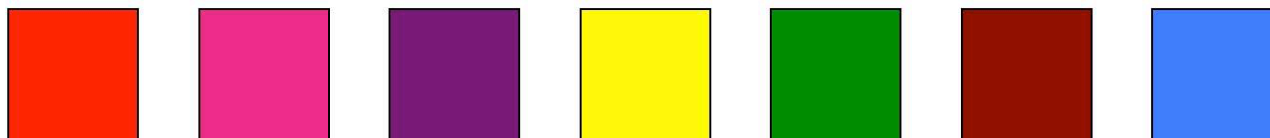


Value encodes continuous variables (less well)

How Many Colors?



Hue encodes nominal variables

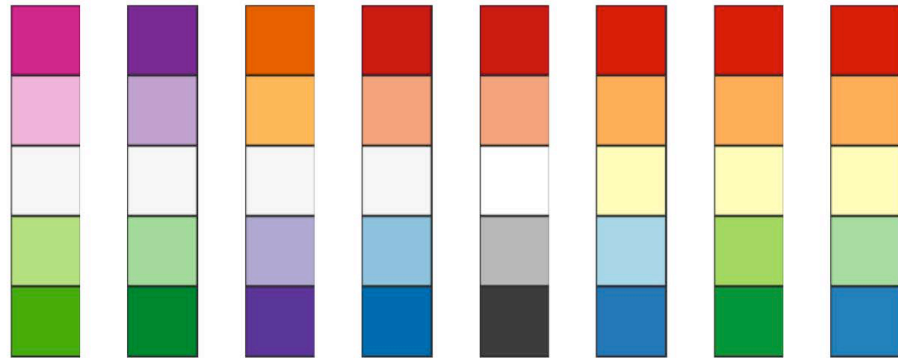


(from S. Davidoff)

Choosing the color palette

Discrete rather than continuous

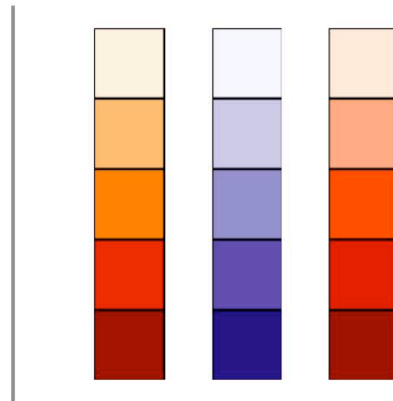
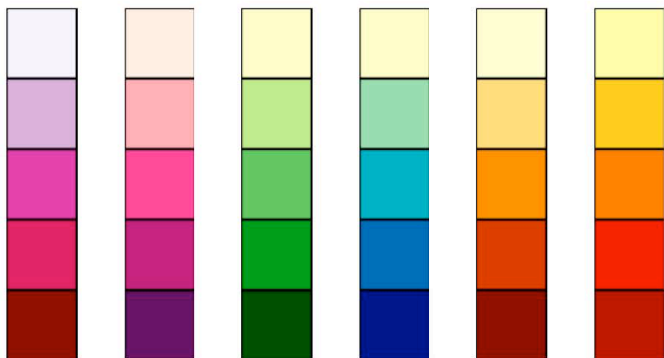
Diverging color



Data maps to meaningful mid-point

Color midpoint neutral, saturation at endpoints

Sequential color



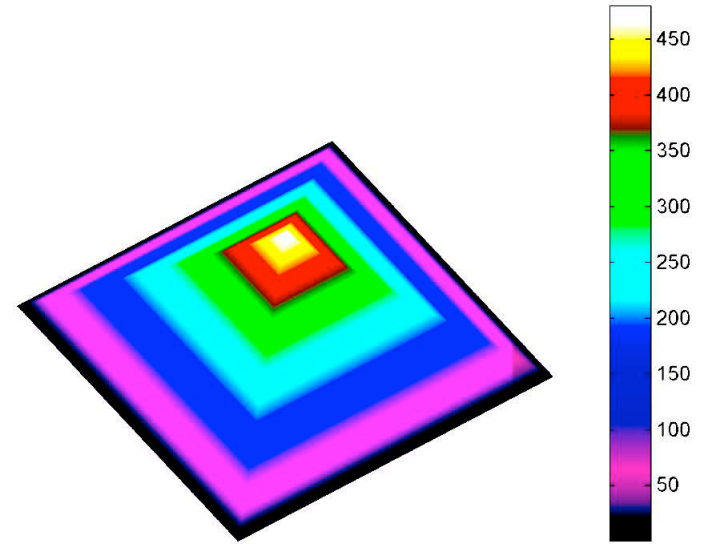
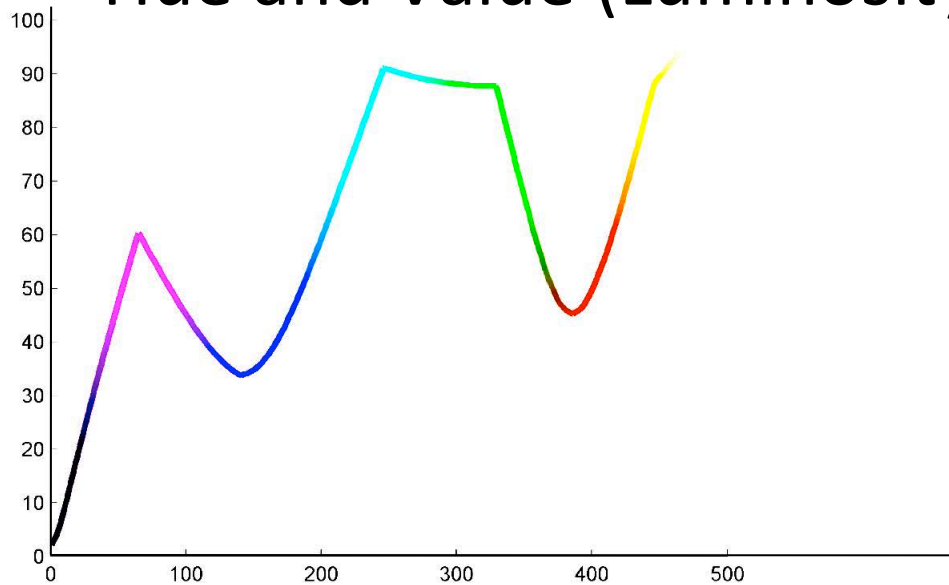
It depends on your purpose

Vary luminance and saturation

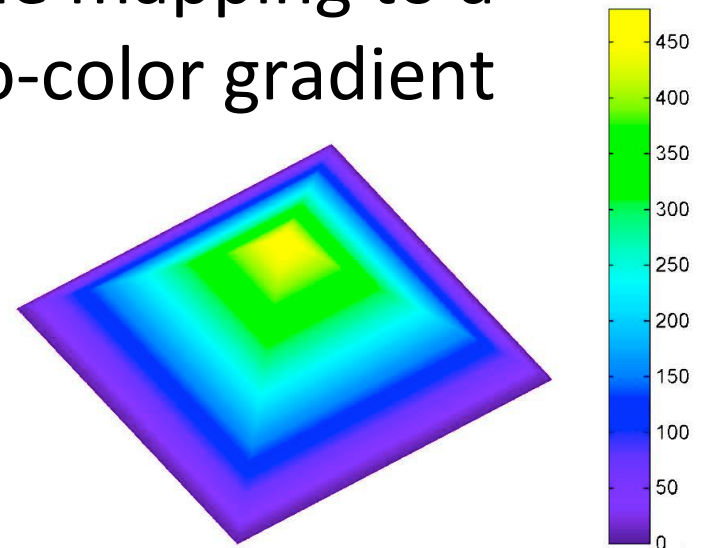
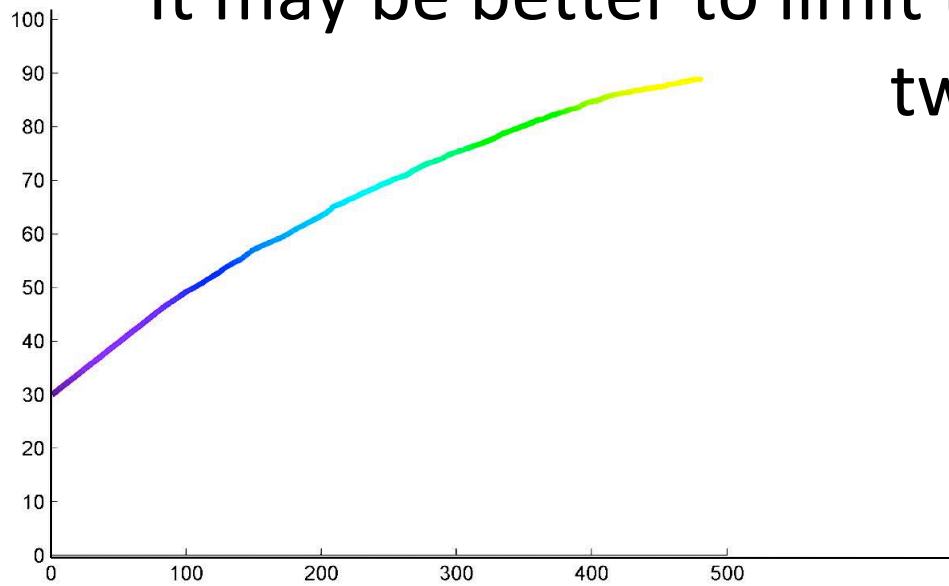
Map higher values to darker colors

(from S. Davidoff)

Hue and Value (Luminosity) can be confusing



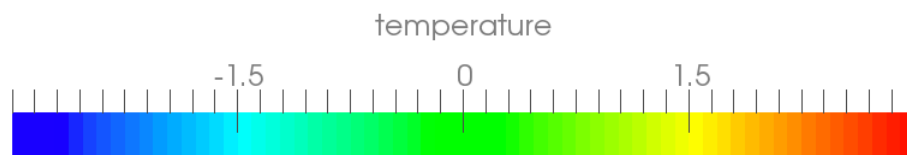
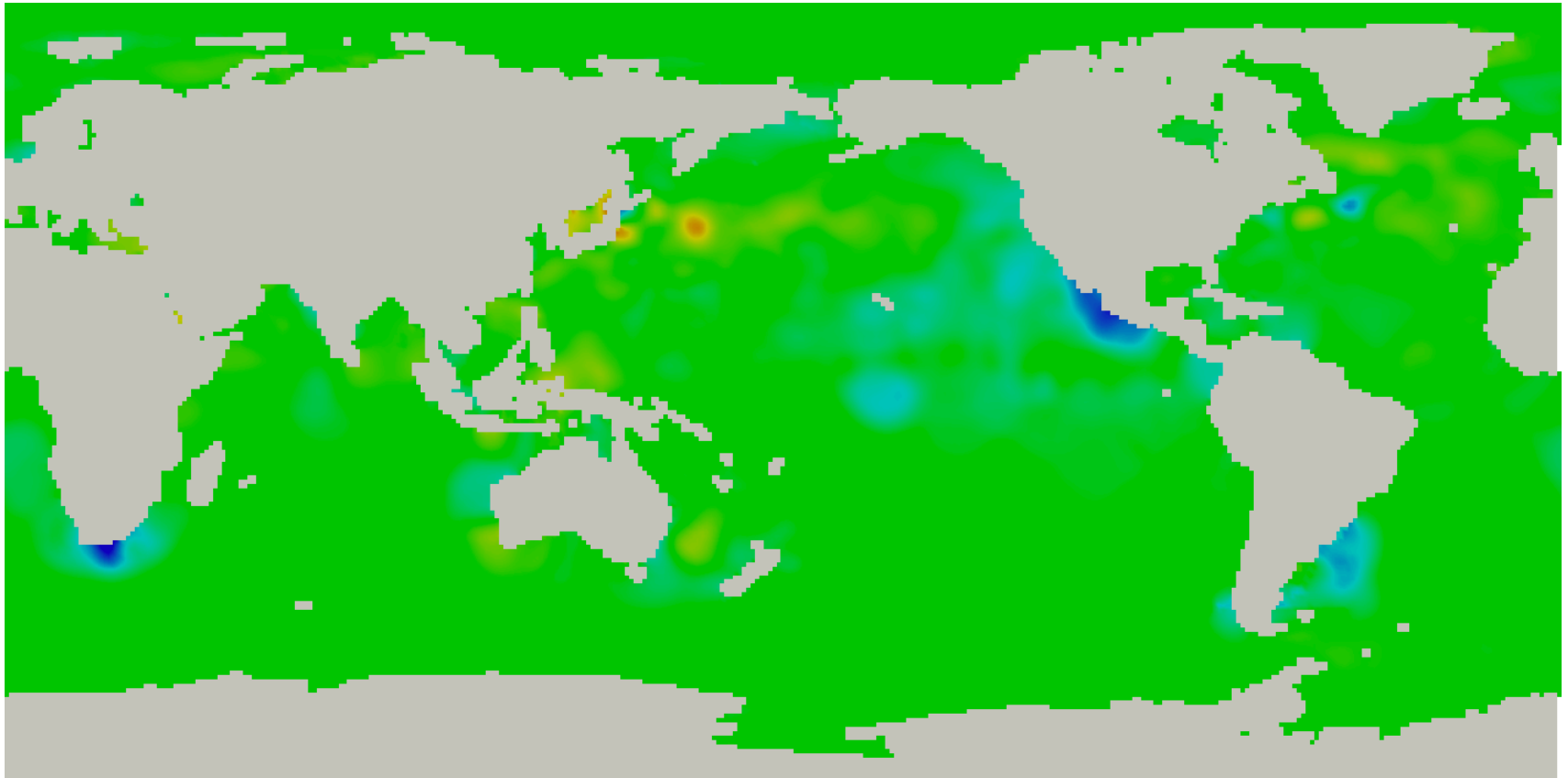
It may be better to limit the mapping to a two-color gradient



(from S. Lombeyda)

Color Map Choice Emphasizes Different Things

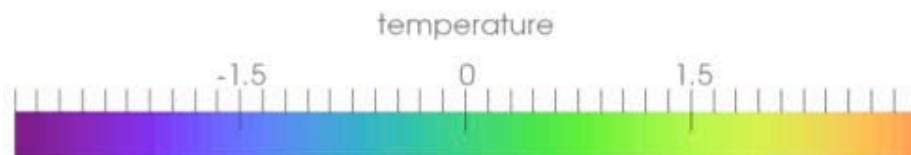
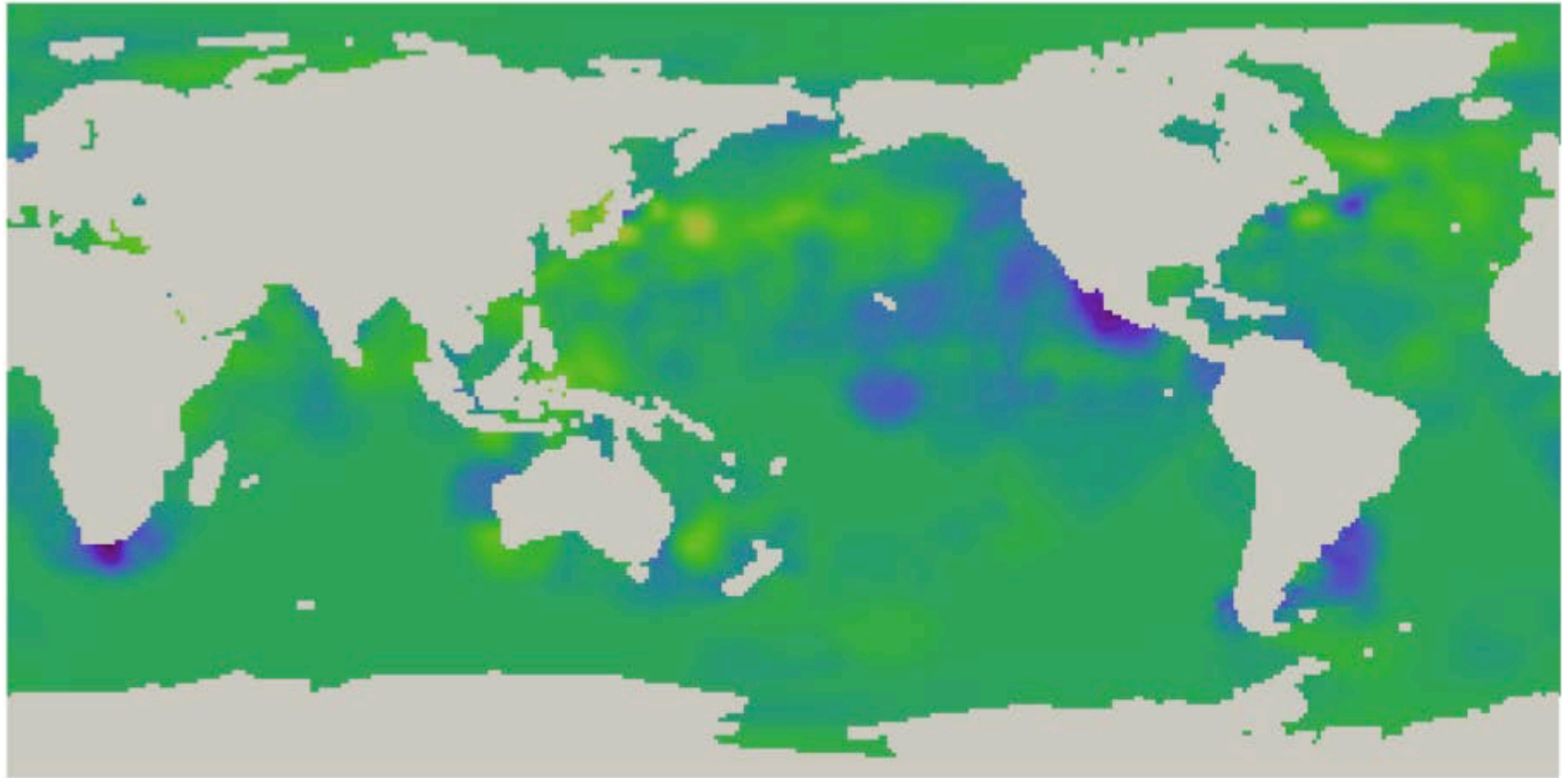
NOAA satellite data on the annual average ocean temperature at a 100m depth



Standard rainbow: dominated by the most common pixel values

Color Map Choice Emphasizes Different Things

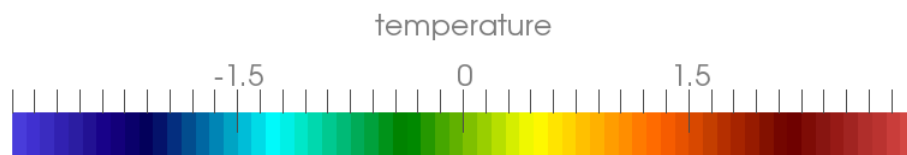
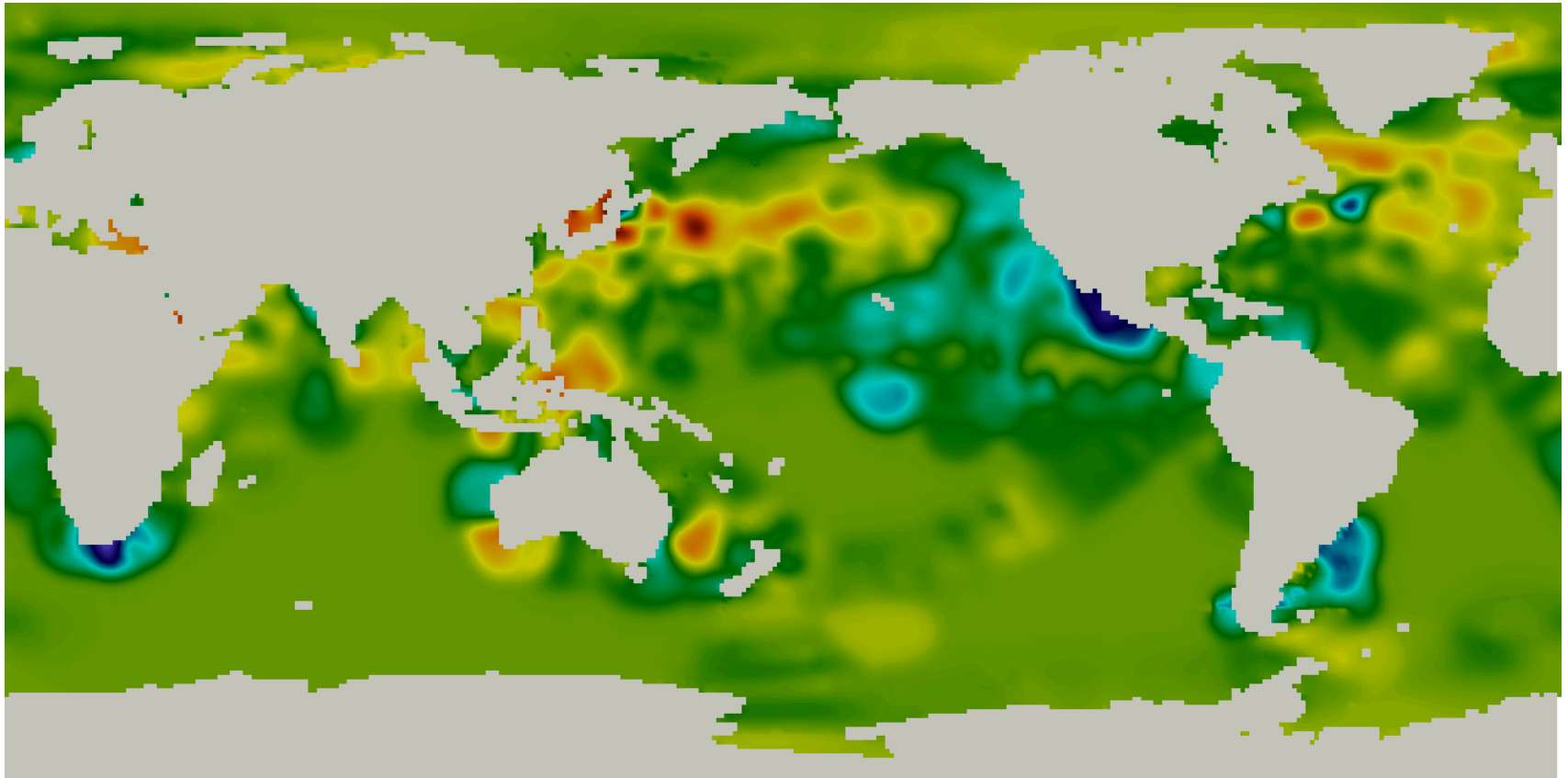
NOAA satellite data on the annual average ocean temperature at a 100m depth



Squeeze the middle to emphasize the tails of the distribution

Color Map Choice Emphasizes Different Things

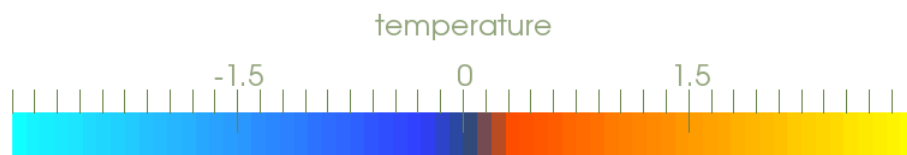
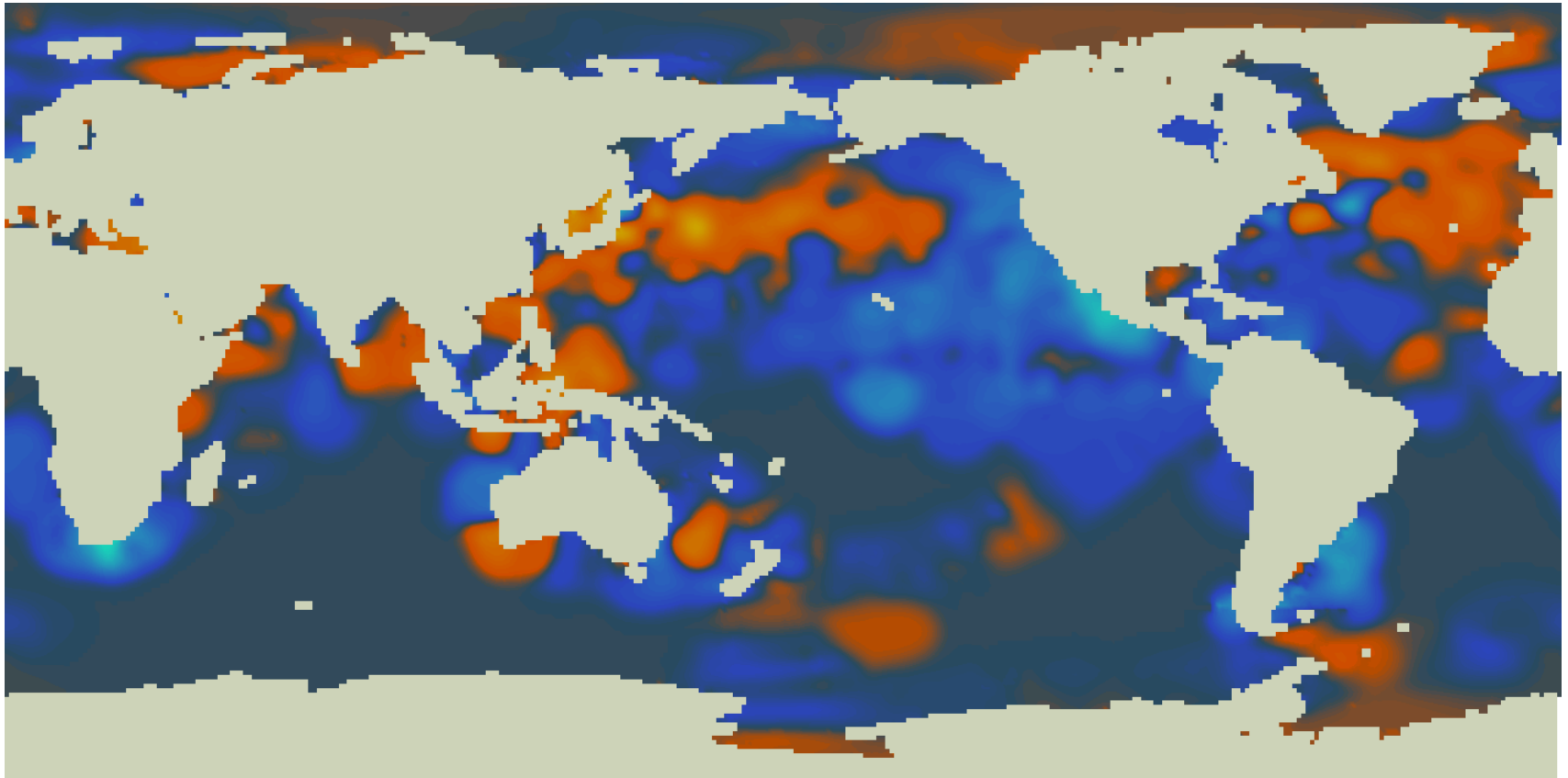
NOAA satellite data on the annual average ocean temperature at a 100m depth



Expand the ends to see most of the variations

Color Map Choice Emphasizes Different Things

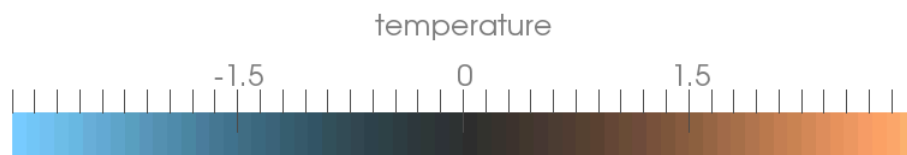
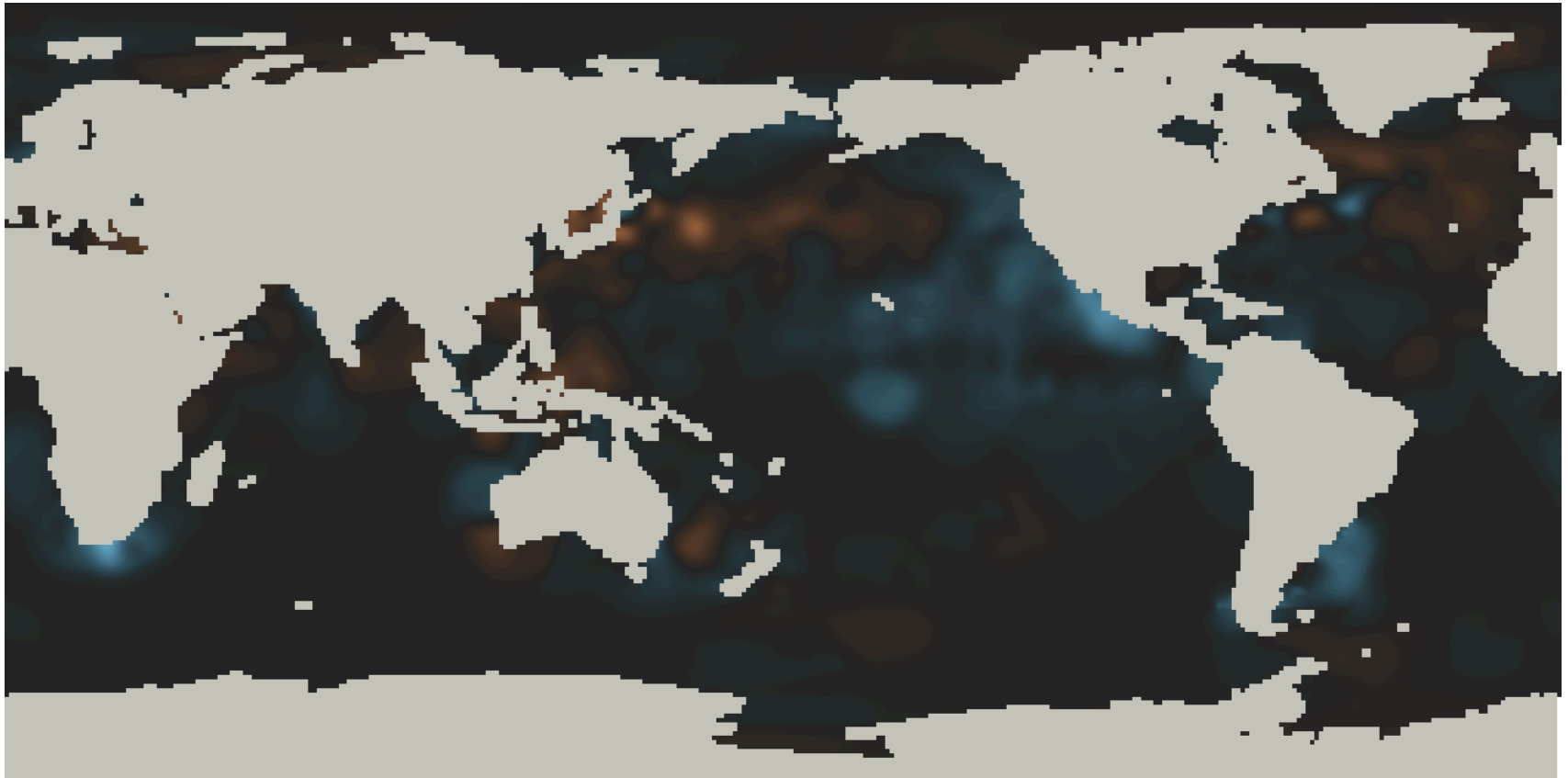
NOAA satellite data on the annual average ocean temperature at a 100m depth



Squeeze the middle dramatically, to really emphasize the tails

Color Map Choice Emphasizes Different Things

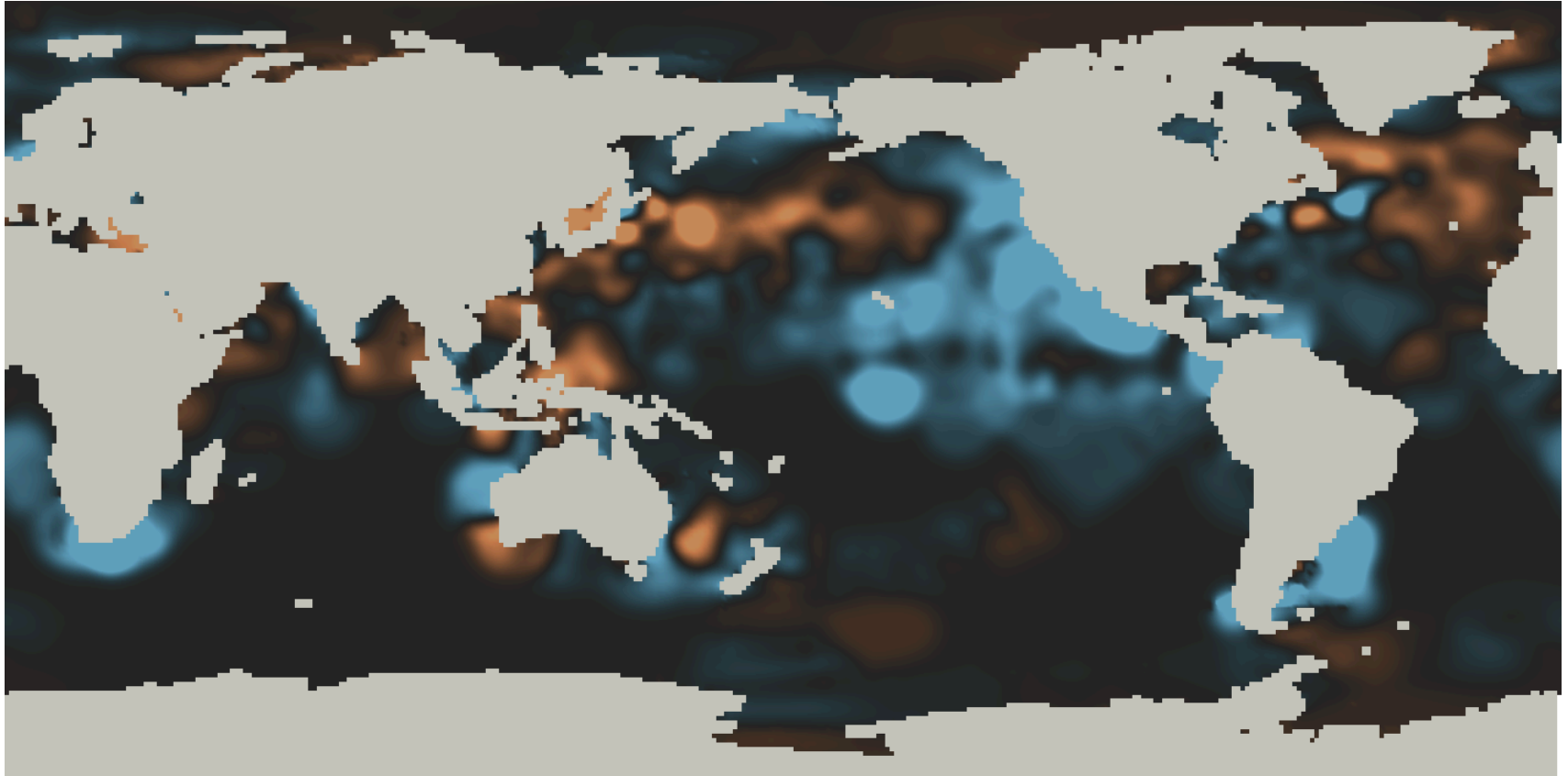
NOAA satellite data on the annual average ocean temperature at a 100m depth



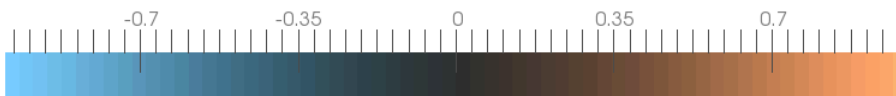
Black out the middle, color the ends of the distribution

Color Map Choice Emphasizes Different Things

NOAA satellite data on the annual average ocean temperature at a 100m depth



temperature



Really emphasize the extreme values

Guidelines for Color in Data Visualization

Use only a few (6 is ideal, 9 is max)

Colors should be distinctive and named

Strive for color harmony

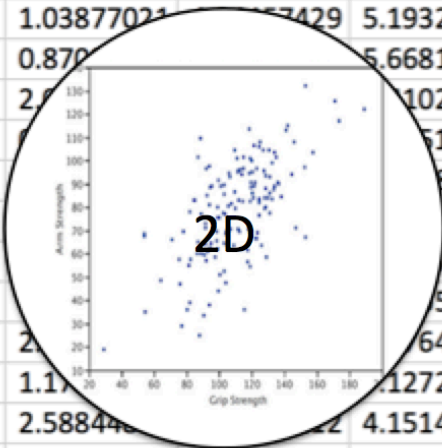
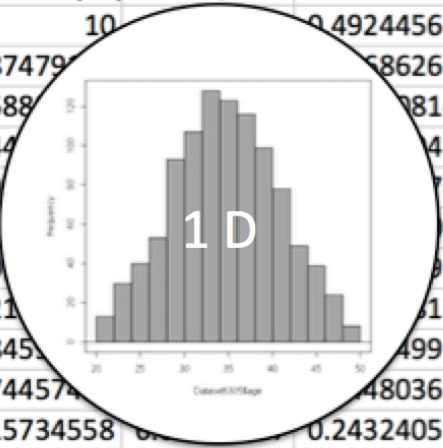
Be aware of cultural conventions

Beware bad interactions

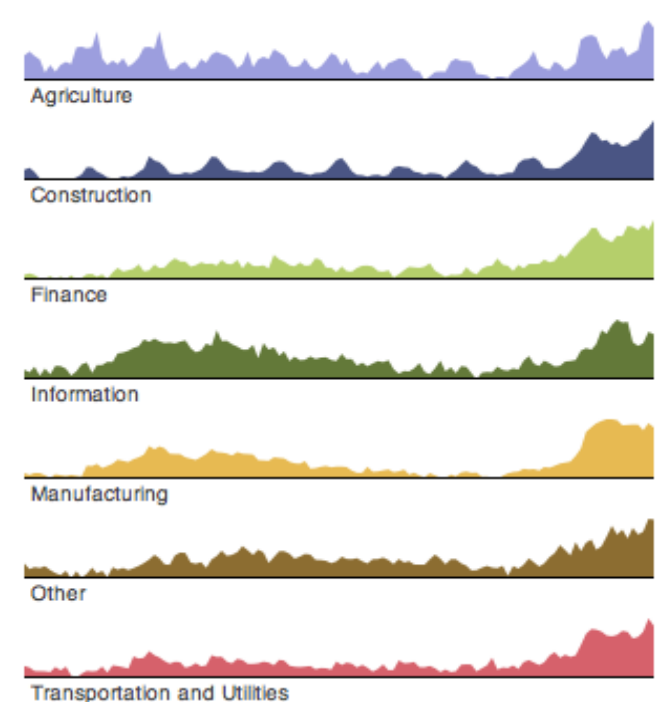
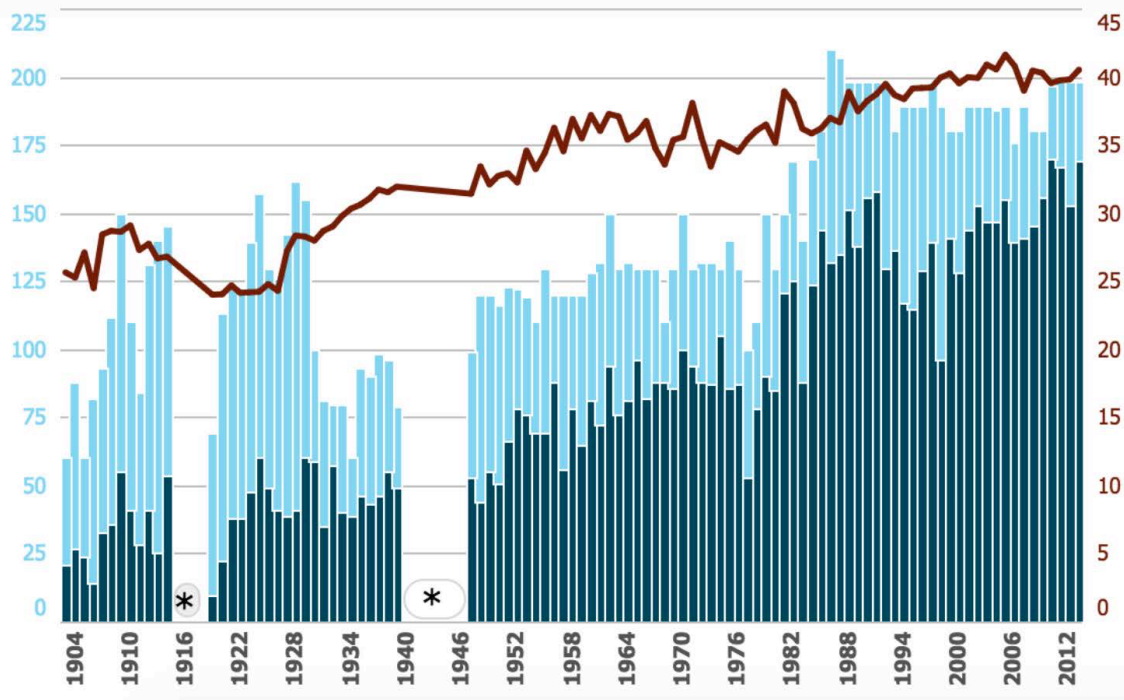
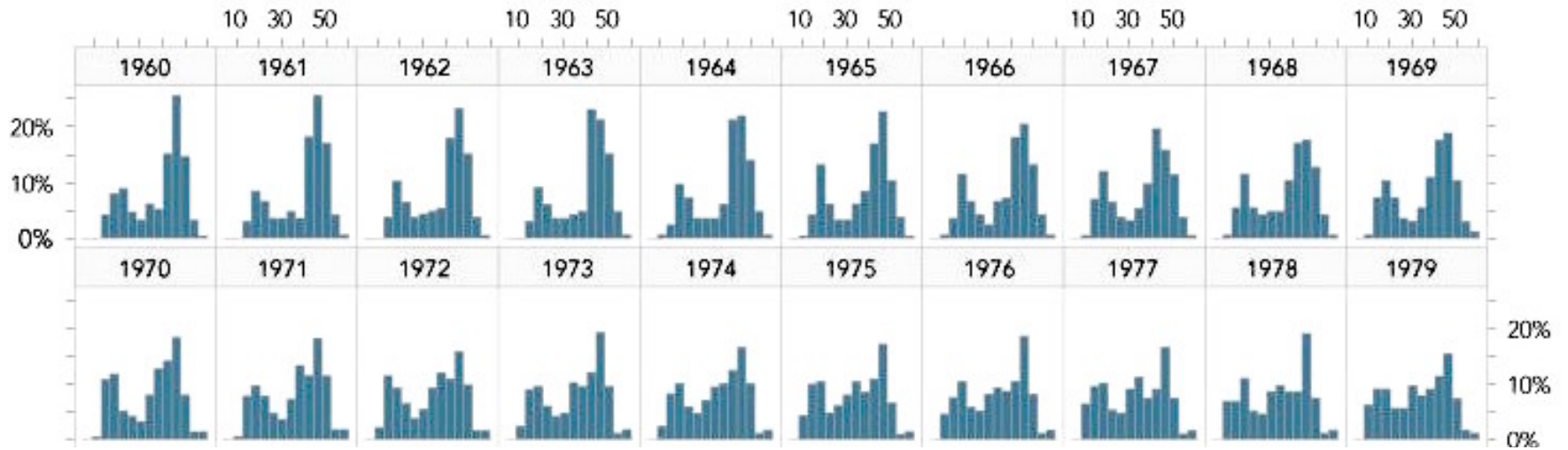
Get it right in black and white

Traditional data visualization fails to reveal the complex patterns - the hidden knowledge - that may be present in the data

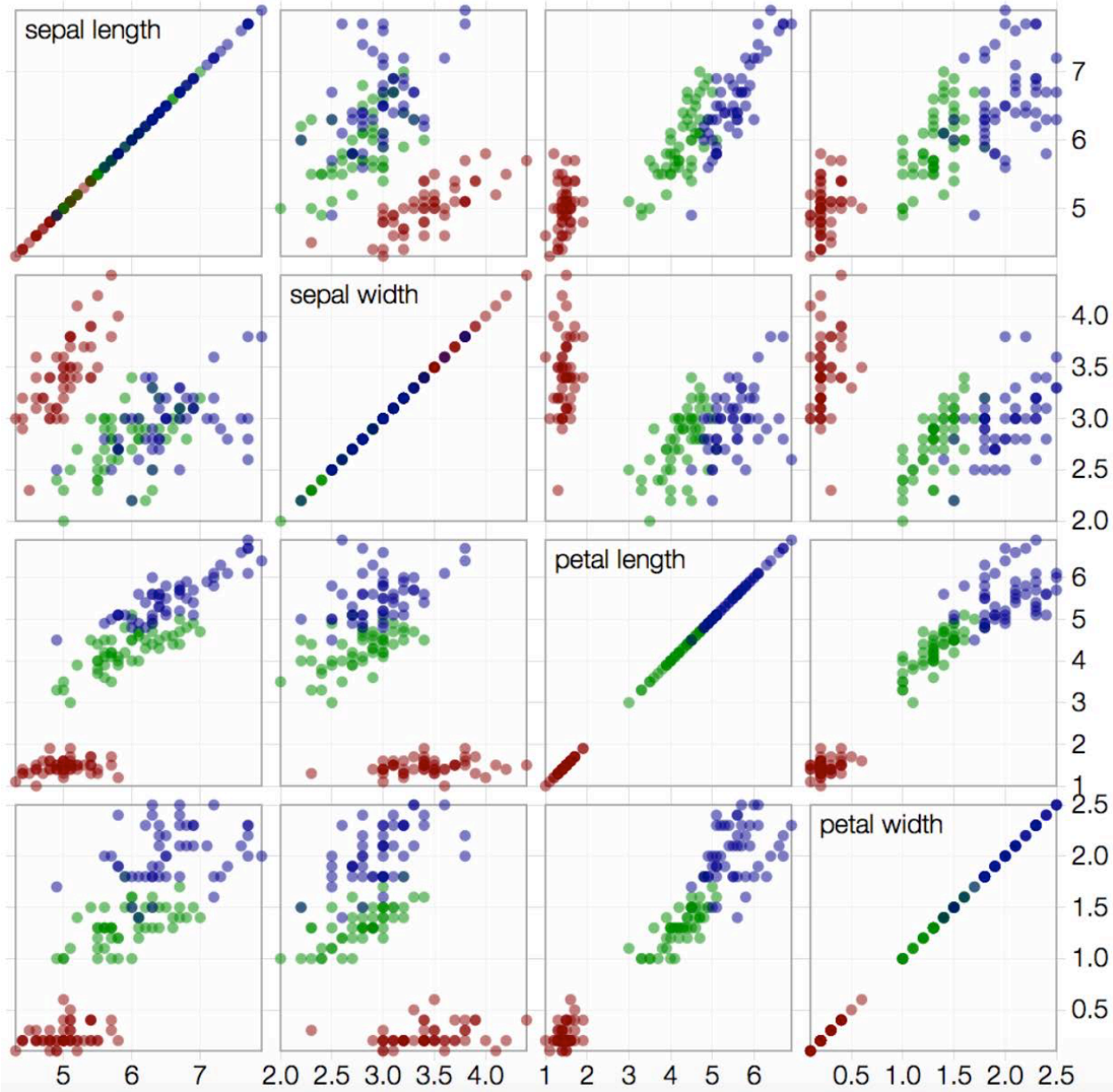
ID-Ticker	X-ret03/13-0	Y-volatility-1y	Z-ret03/15-0	size-LOG-AU	alpha_1	trans/norm	beta_2/Y	Y-derivative	LOG_Volume	Tau/risk
FCG	2.12501413	10	0.49244565	2.12501413	17	1	10	0.49244565		
EWZS	0.15372443	4.87479	0.34686263	0.15372443	4.87479132	2.78193643	4.87479132	0.34686263	2.781936	
BRF	0.24188991	4.588	0.42750814	0.24188991	4.58889816	3.28261136	4.58889816	0.42750814	3.282611	
YMLP	2.83938058	6.44	0.43089047	2.83938058	6.44198664	0.18161738	6.44198664	0.43089047	0.181617	
REMX	1.19475528	3.9	0.36430776	1.19475528	3.99207012	2.52178181	3.99207012	0.36430776	2.521781	
KOL	1.76104894	4.2	0.3934906	1.76104894	4.25500835	2.00638115	4.25500835	0.3934906	2.006381	
URA	1.31796089	4.0	0.46457595	1.31796089	4.01502504	3.16357835	4.01502504	0.46457595	3.163578	
PSCE	2.77608229	8.21	0.35876813	2.77608229	8.21577629	0.946	8.21577629	0.35876813	1.35599	
GREK	2.13179609	8.845	0.51094991	2.13179609	8.84599332	2.49723893	8.84599332	0.51094991	2.497238	
COPX	1.62880072	6.74457	0.34480365	1.62880072	6.74457429	3.58326175	6.74457429	0.34480365	3.583261	
SILJ	0.31423081	8.15734558	0.24324053	0.31423081	8.15734558	6.68793717	8.15734558	0.24324053	6.687937	
GDXJ	0	8.13230384	7.9224447	0.66265945	0	8.13230384	7.9224447	0.66265945	7.922444	
SLVP	0.9325195	6.34599332	5.40802553	0.33450945	0.9325195	6.34599332	5.40802553	0.33450945	5.408025	
GXG	1.03877021	4.24457429	5.19327525	0.42990263	1.03877021	4.24457429	5.19327525	0.42990263	5.193275	
SIL	0.87035153	7.14732888	5.66818015	0.49087651	0.87035153	7.14732888	5.66818015	0.49087651	5.668180	
XES	2.08093139	7.21410685	3.410234	0.49447508	2.08093139	7.21410685	3.410234	0.49447508	3.410234	
RSXJ	0.67706567	6.3516996	0.37451031	0.37451031	6.3516996	4.33222037	6.3516996	0.37451031	6.35169	
EWZ	1.8819939	6.03505843	3.87286784	0.67265086	1.8819939	6.03505843	3.87286784	0.67265086	3.872867	
XOP	3.25760145	6.99290484	2.40888453	0.66048546	3.25760145	6.99290484	2.40888453	0.66048546	2.408884	
XME	2.4968916	5.7909015	3.65075469	0.55648664	2.4968916	5.7909015	3.65075469	0.55648664	3.650754	
PXJ	2.3951622	5.92445743	3.86550497	0.38525767	2.3951622	5.92445743	3.86550497	0.38525767	3.865504	
PICK	2.84955352	4.83931553	3.27647564	0.44358701	2.84955352	4.83931553	3.27647564	0.44358701	3.276475	
GDX	1.17893071	7.89858097	7.12725488	0.75688682	1.17893071	7.89858097	7.12725488	0.75688682	7.127254	
ENY	2.58844806	5.35267112	4.15142962	0.35967998	2.58844806	5.35267112	4.15142962	0.35967998	4.151429	
RING	1.01616367	7.55843072	7.63529267	0.4383076	1.01616367	7.55843072	7.63529267	0.4383076	7.635292	



Multiple 1-D \neq Multi-D



Multiple 2-D \neq Multi-D

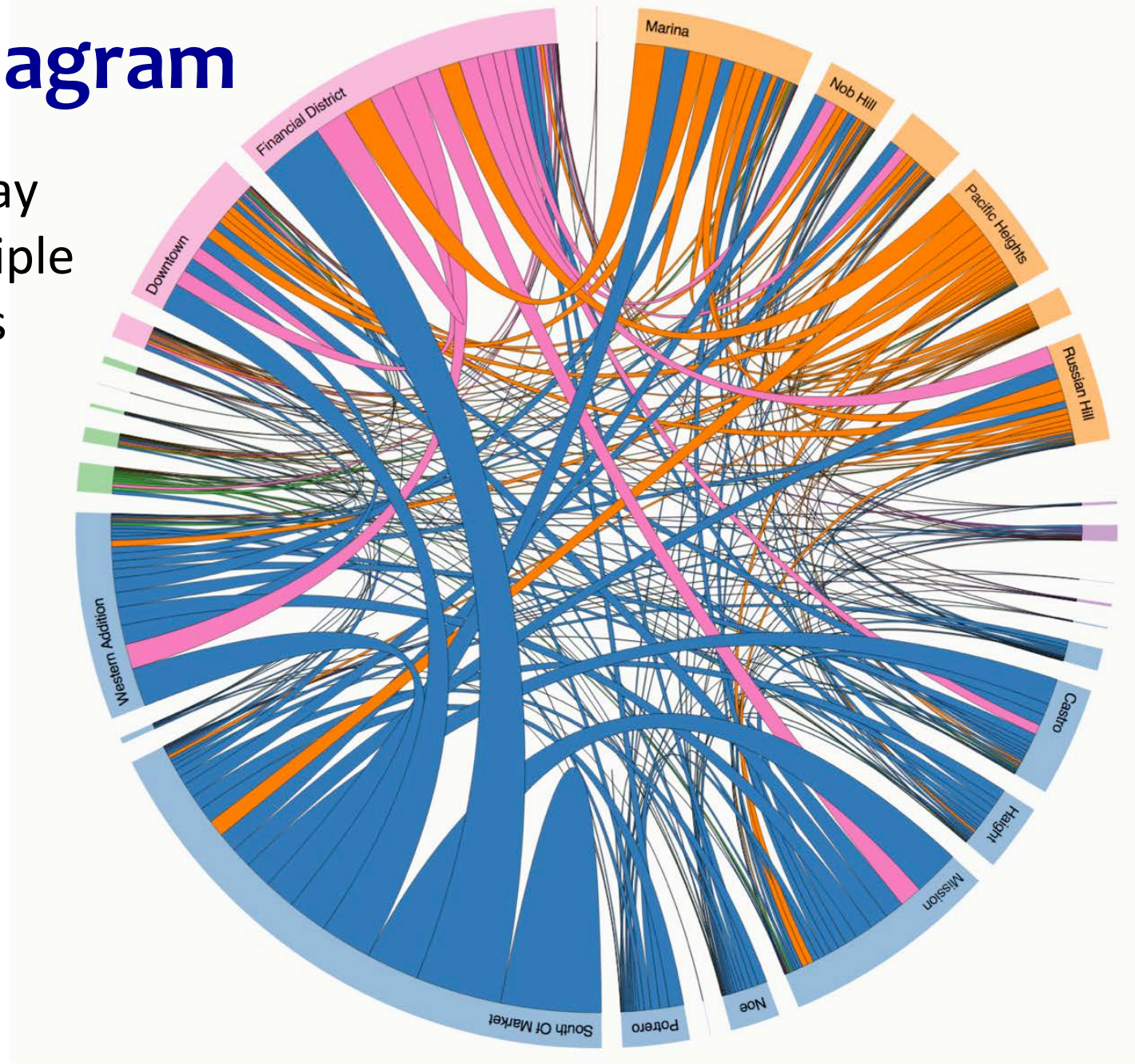


A grid plot of pairwise XY projections

Structures (e.g., clusters) that are present in 3-D (or higher) may not project well on any 2-D plane

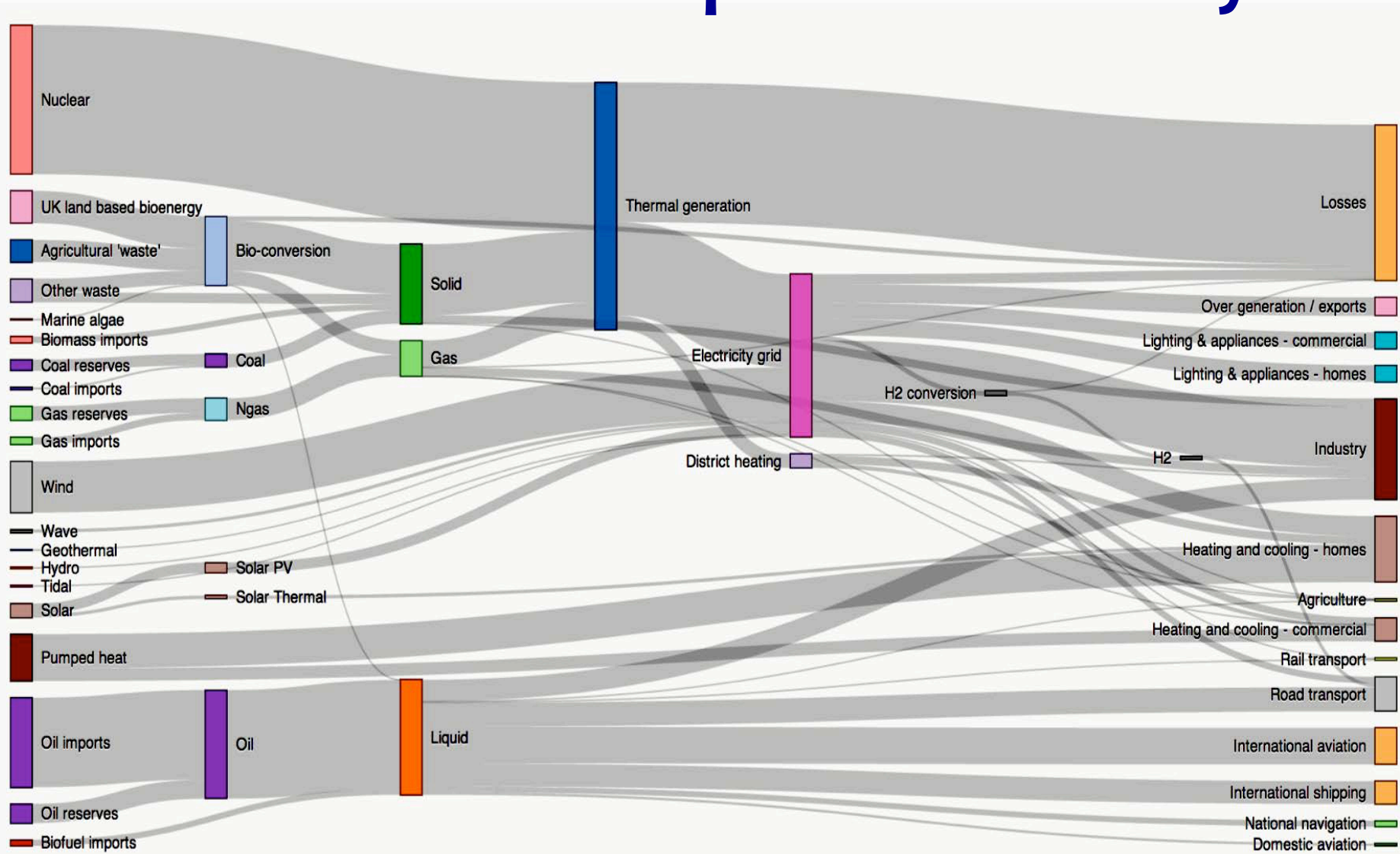
Chord Diagram

A compact way
to show multiple
connectivities

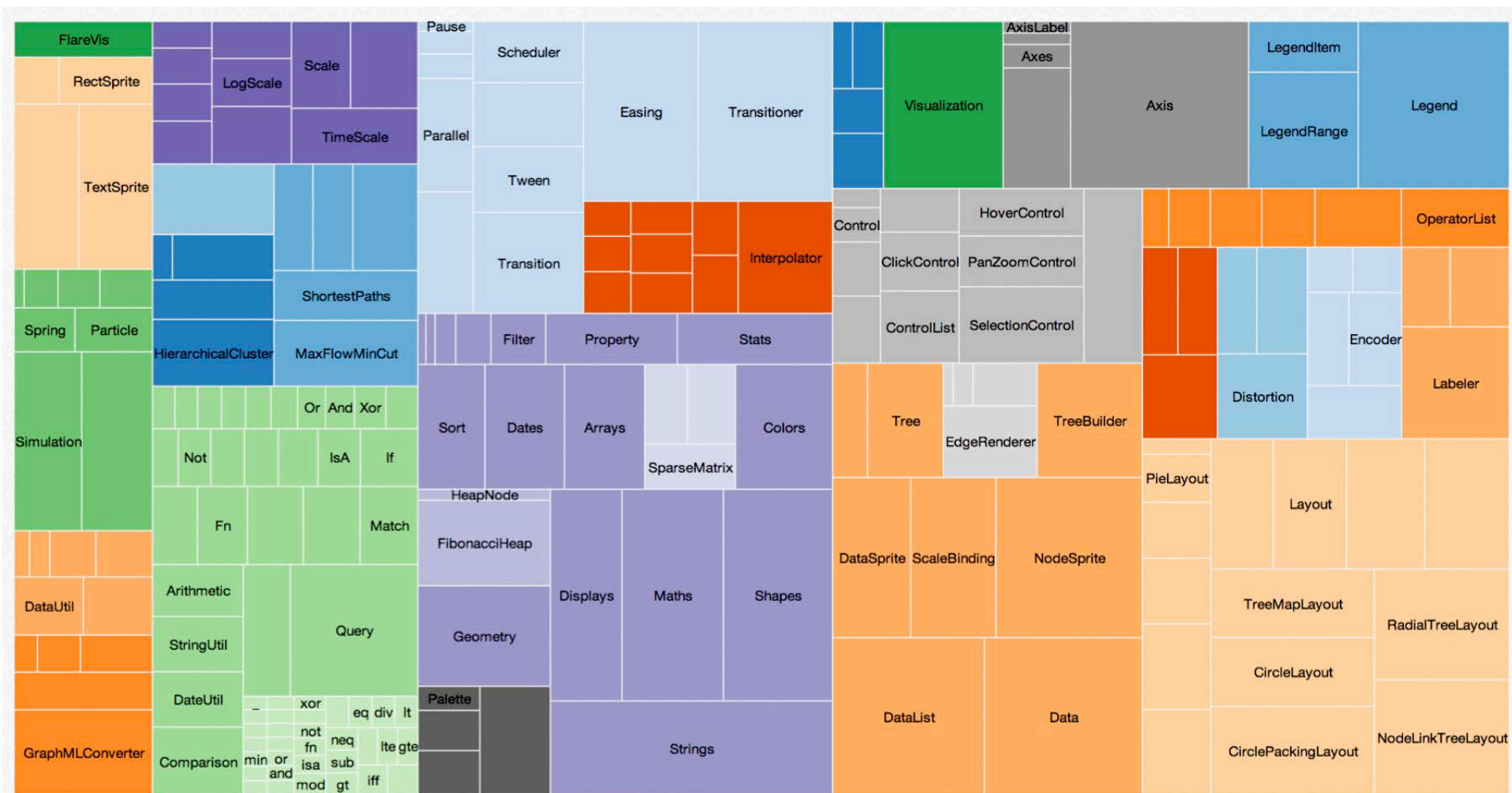


Chord diagram: Uber rides by San Francisco neighborhood

Sankey Diagram: another way to visualize a multiple connectivity



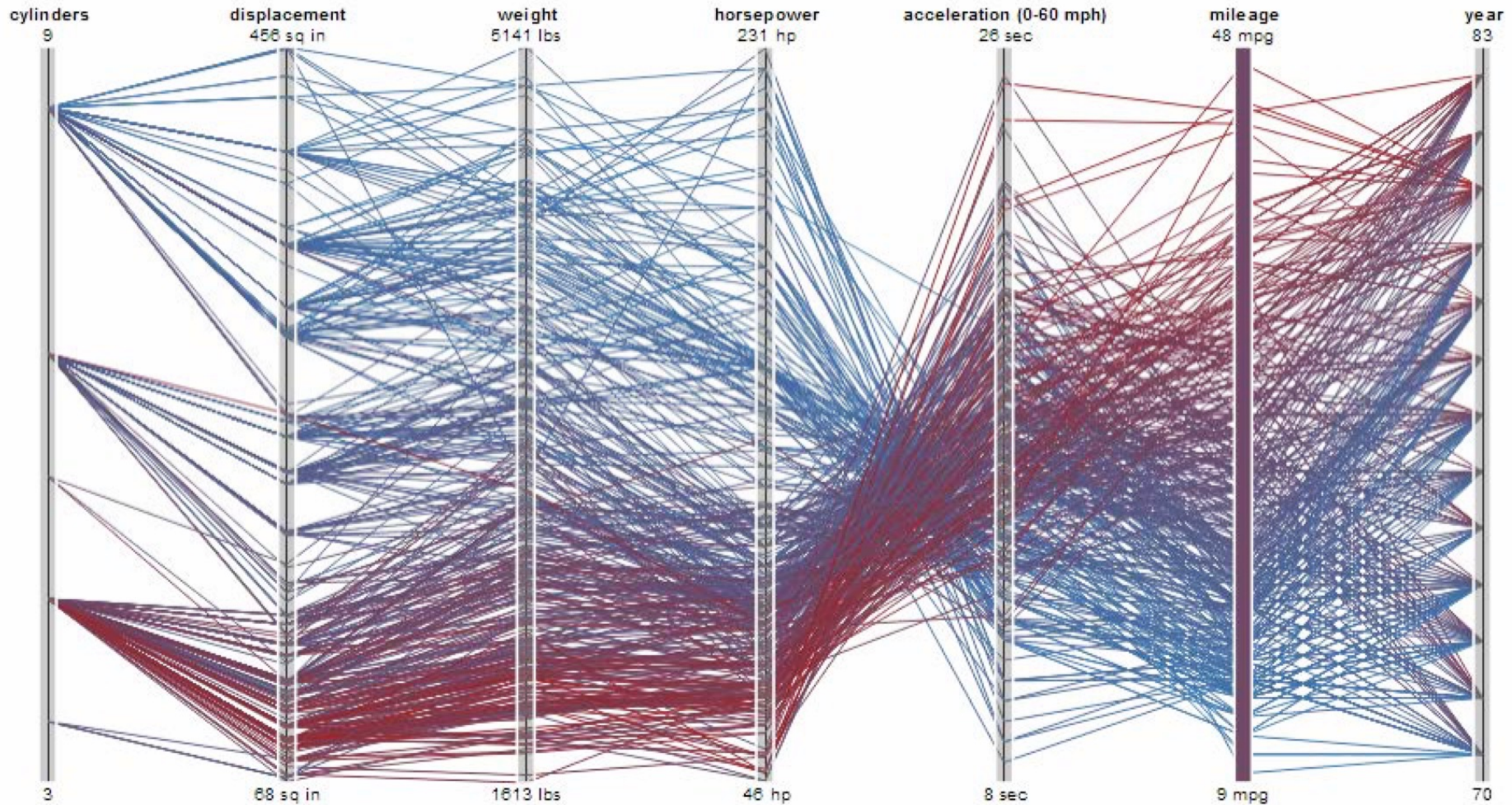
Hierarchical Visualization: Treemap



Area = 1 quantitative dimension

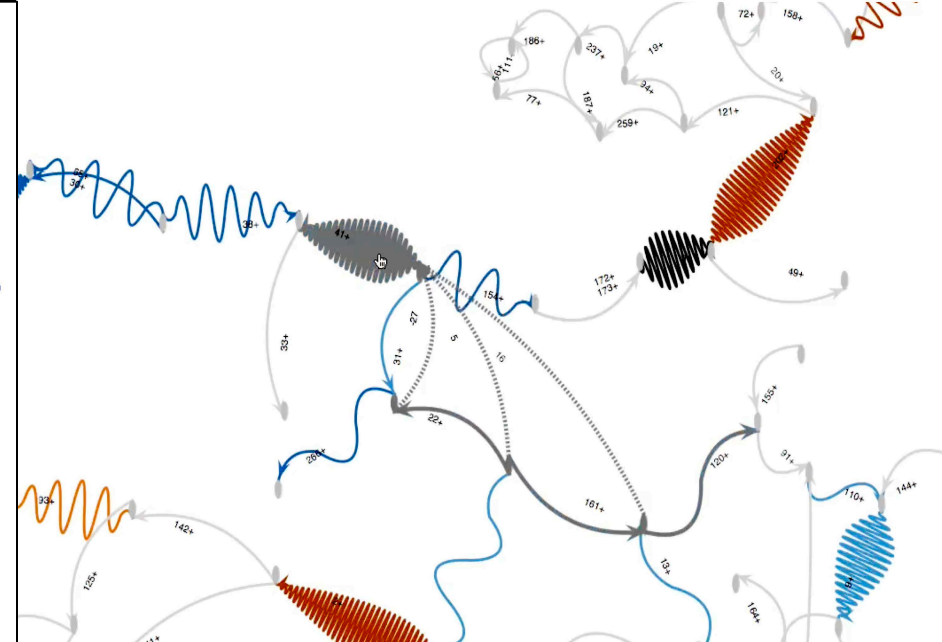
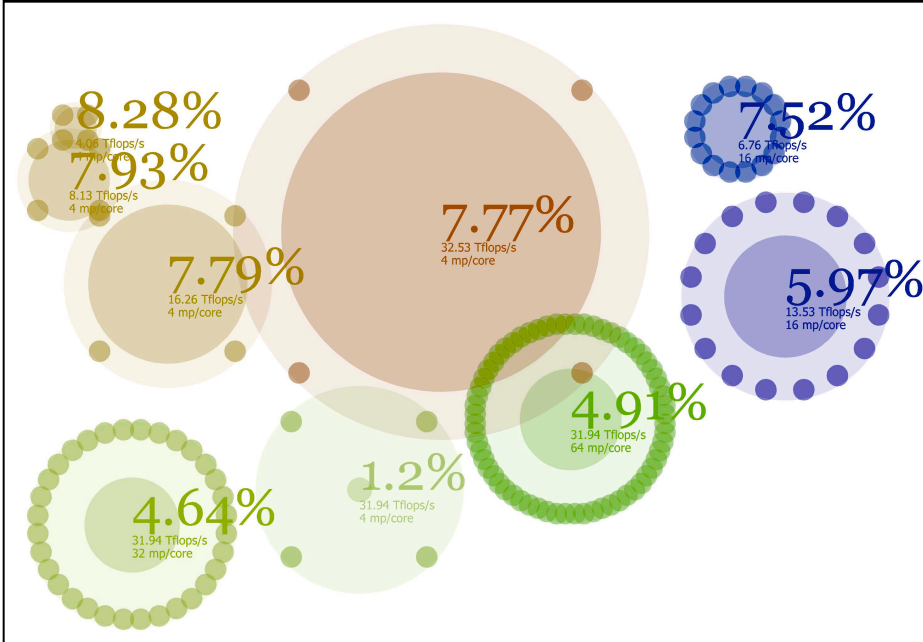
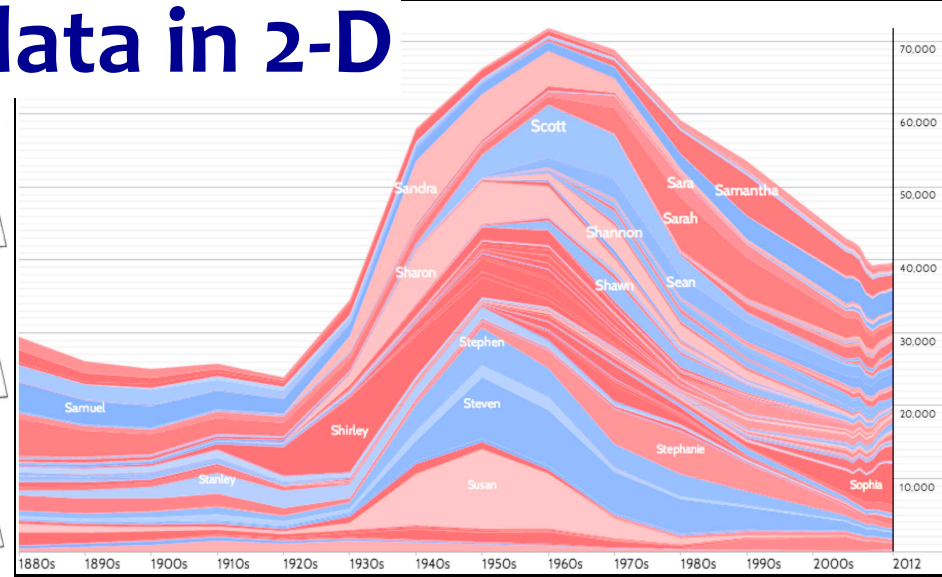
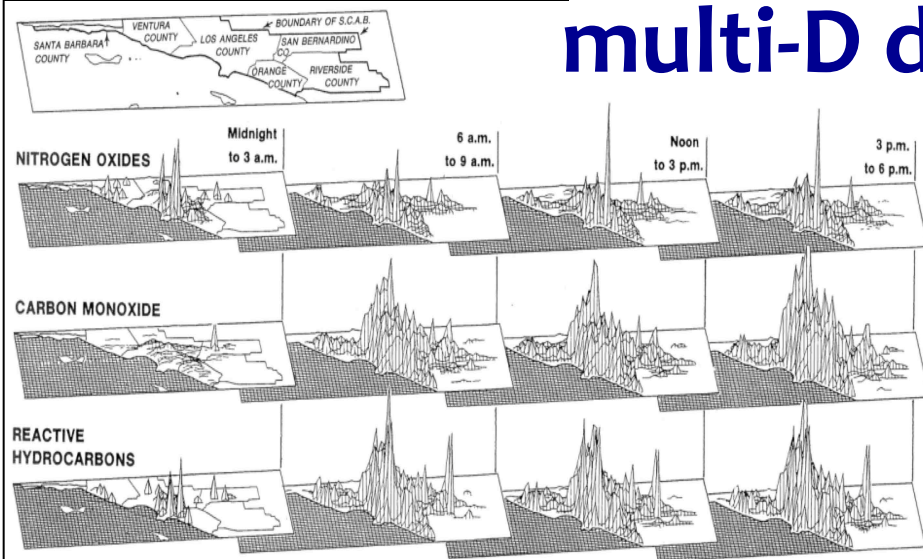
Color = 1 nominal (or ordinal?) dimension

Parallel Coordinates: visualize clustering in multiple dimensions



Note: the order of the axes is arbitrary!

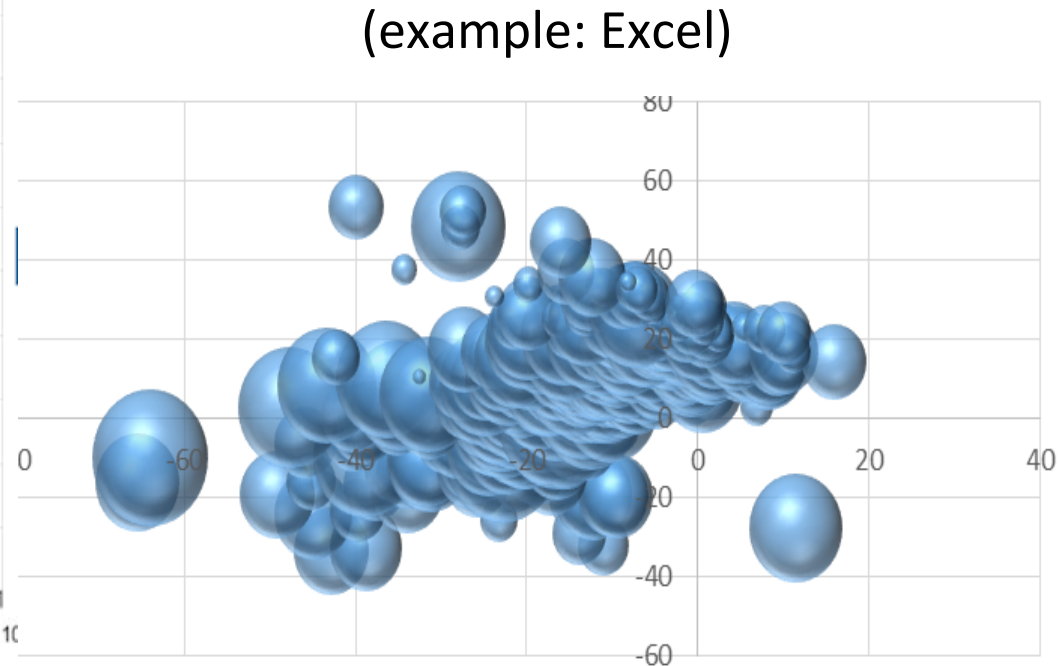
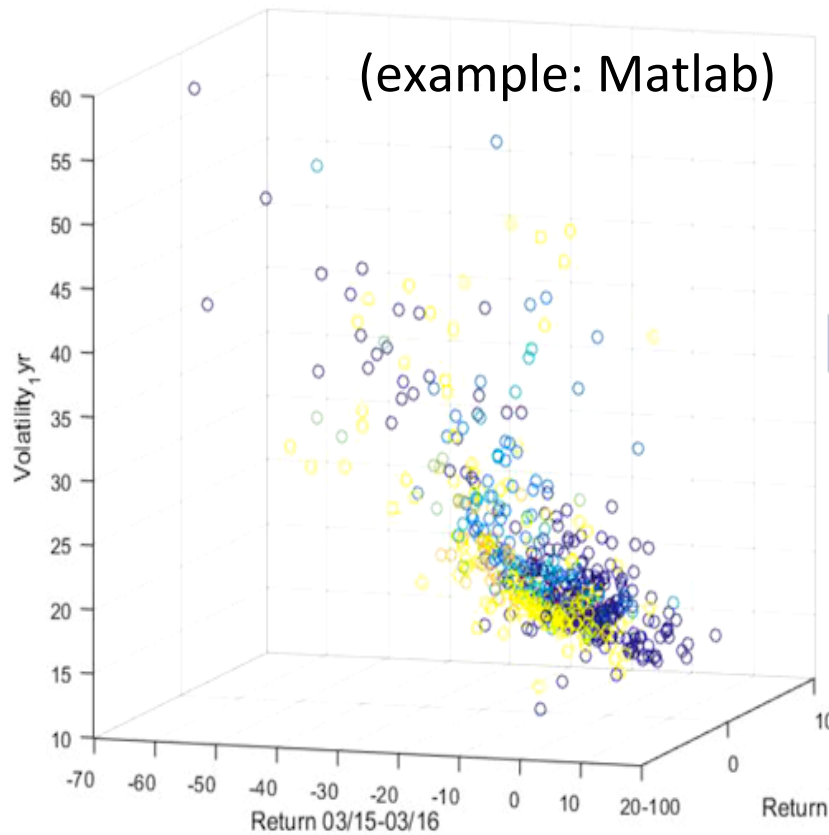
There are many other ways to try to visualize multi-D data in 2-D



(examples from S. Davidoff and S. Lombeyda)

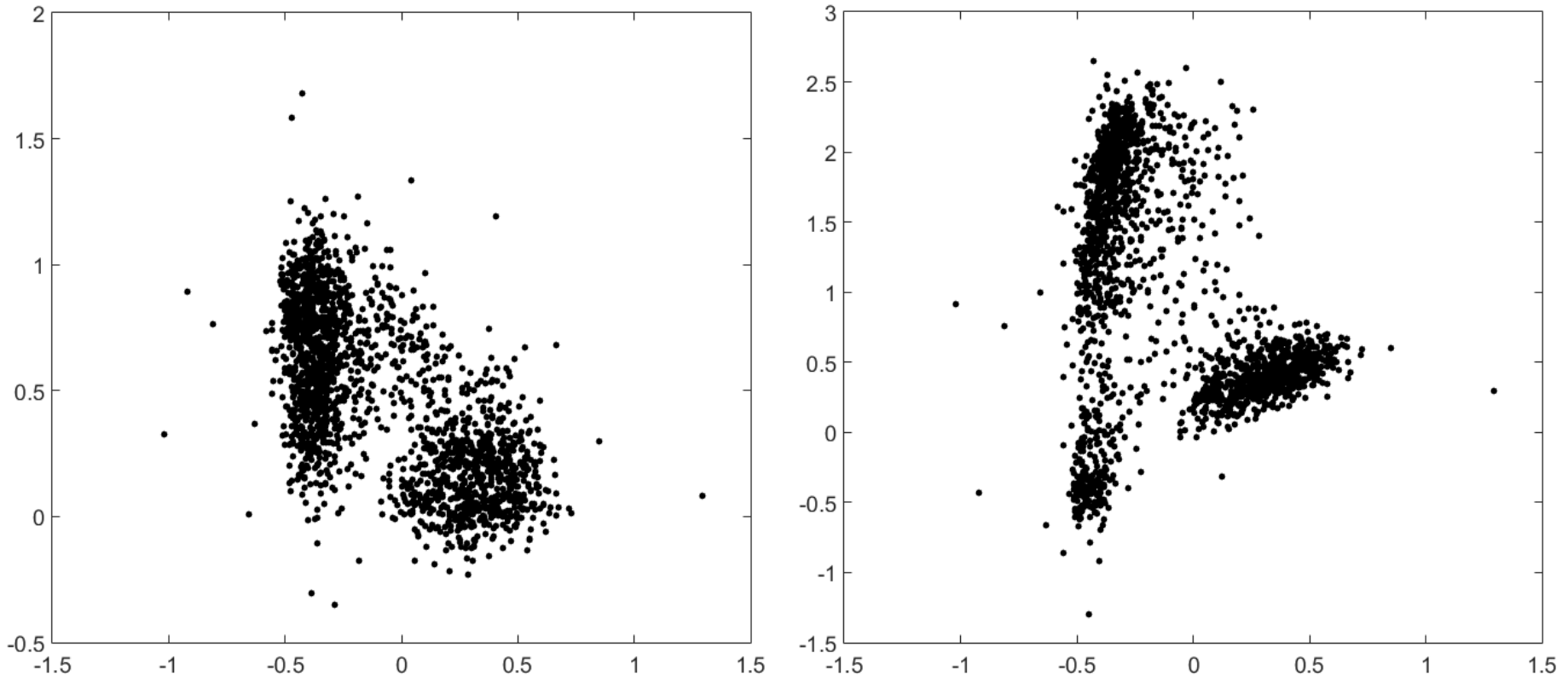
3D Data Visualization is Not Simple

Traditional 3D suffers from many problems (navigation, selection, manipulation, anchoring, perspective, occlusion, and inability to transition to 2D and back) that can hide the structure present in the data



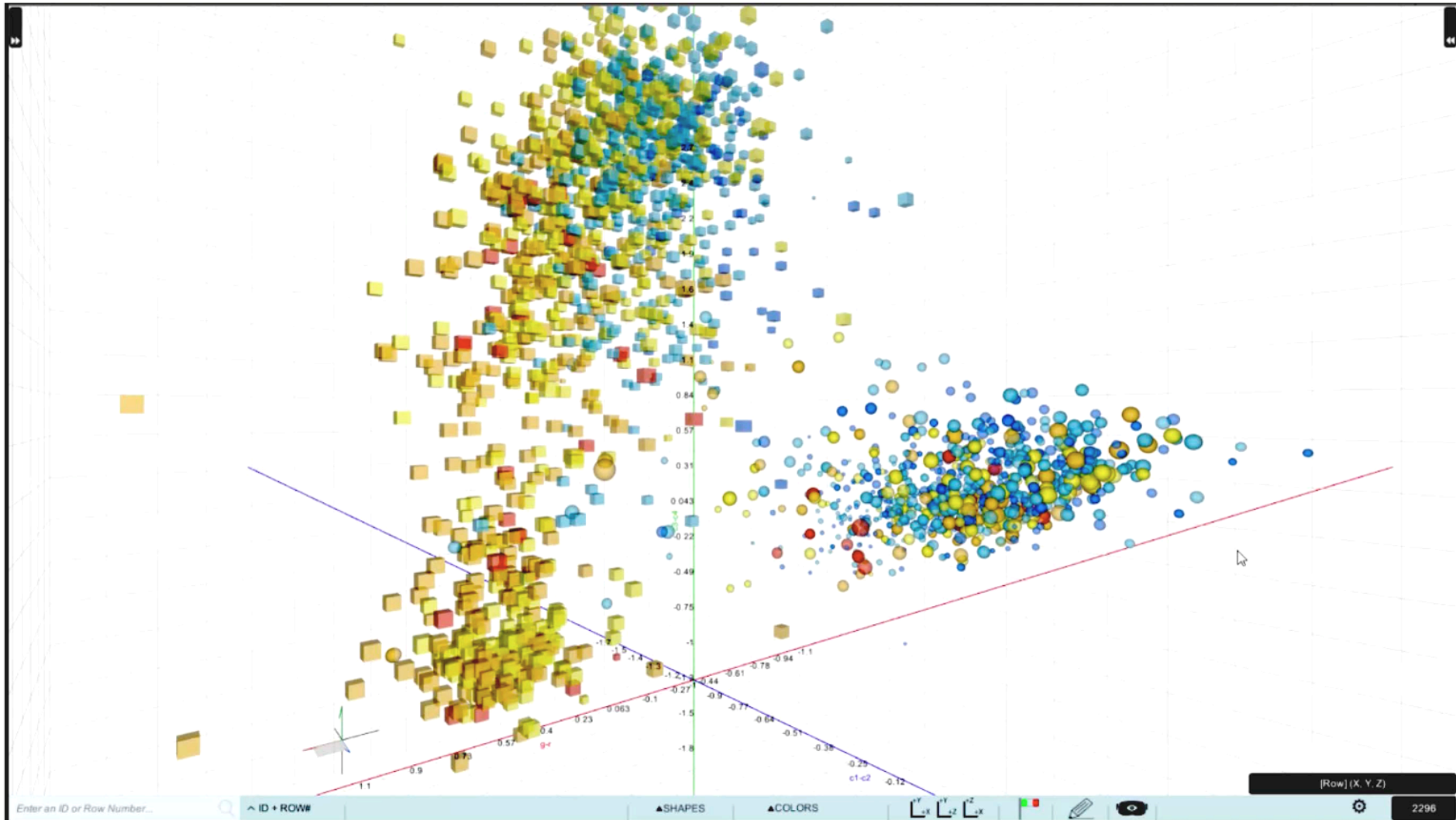
Traditional Data Visualization

An example from astronomy: a subset of data on quasar properties, from a 6-dimensional data space

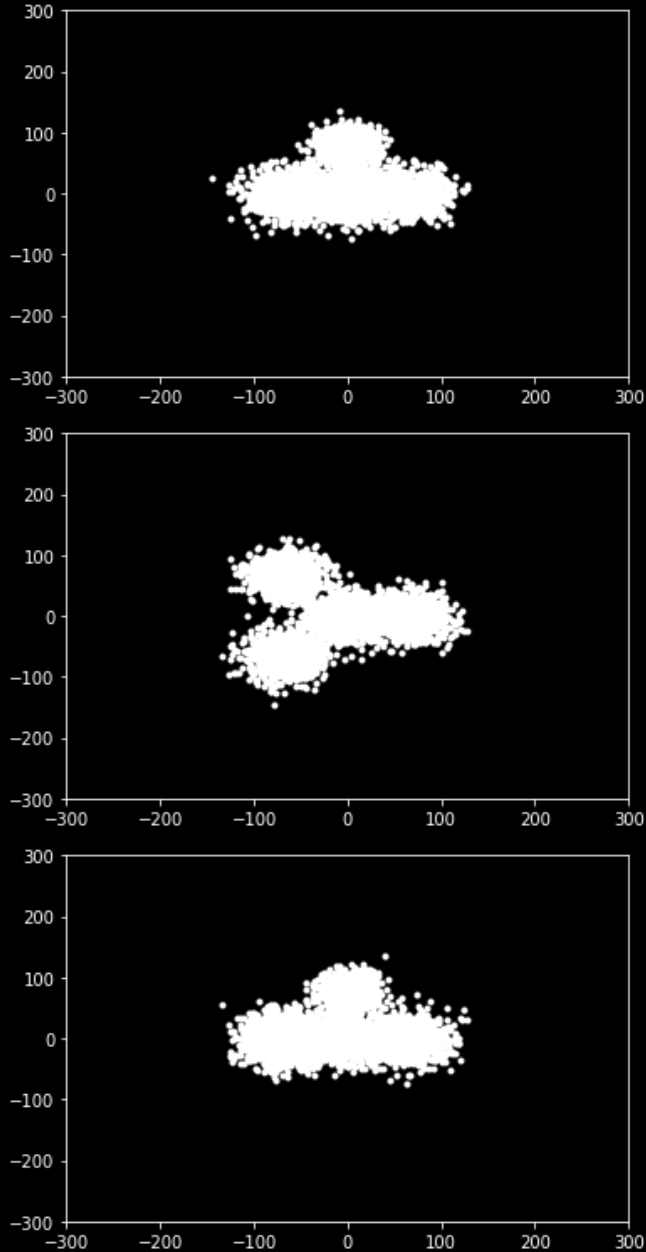


These are 2 out of the 15 possible 2-D plots, but even then relationships involving >2 variables are lost

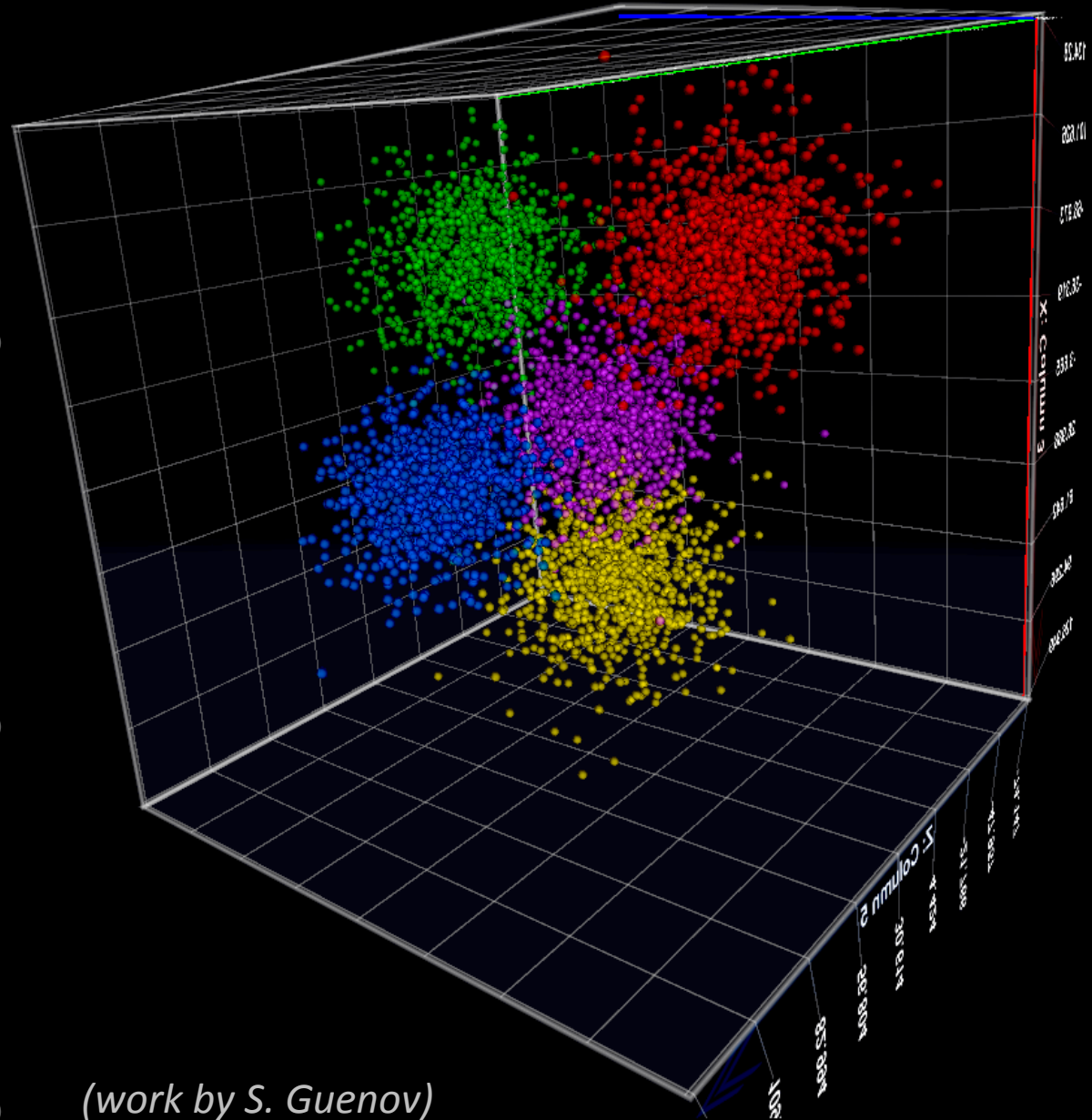
Diving Into the 6-Dimensional Data Space in Virtual Reality



XYZ Projections



3D Visualization + Clustering Analysis



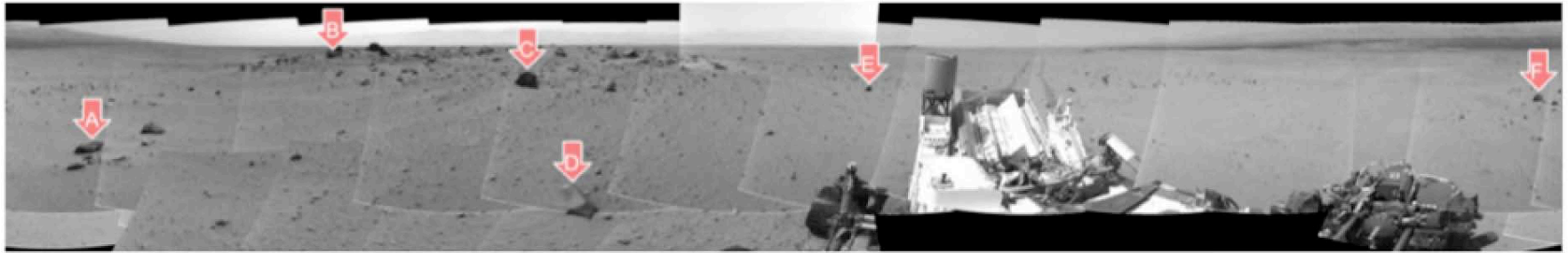
(work by S. Guenov)

Exploring the Virtual Mars at JPL

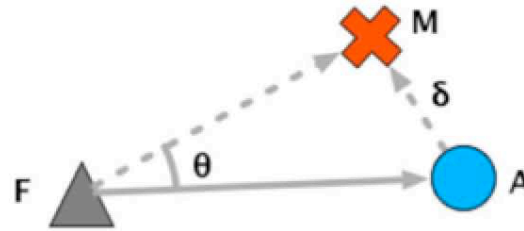
S. Davidoff, J. Norris, et al.



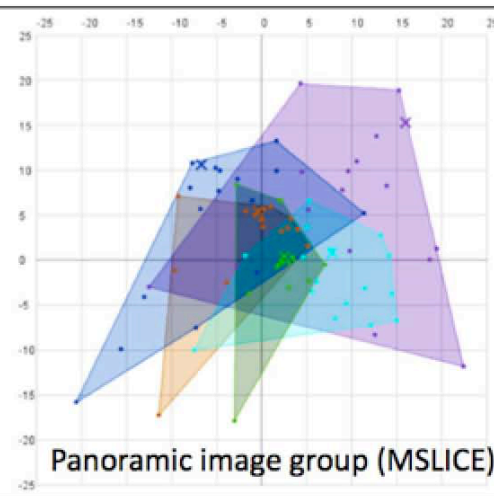
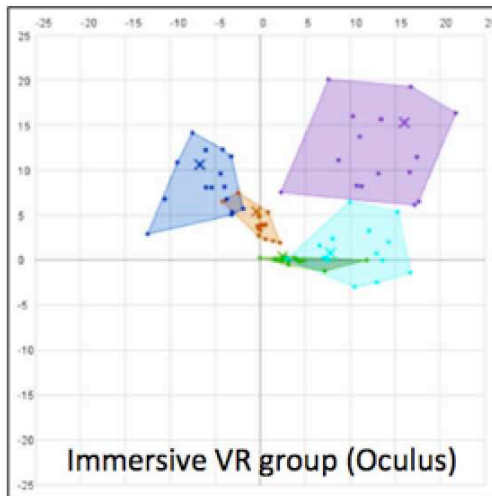
Navigating on Mars using VR



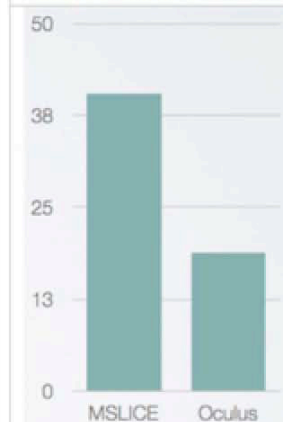
S. Davidoff,
J. Norris,
et al.



Users in VR are **4 times better** in estimating the relative positions, distances, etc.

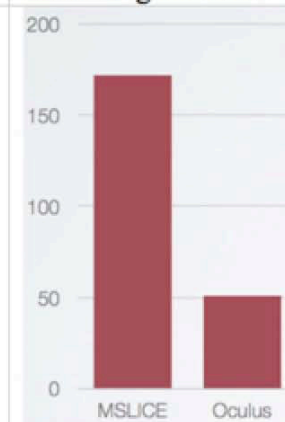


Mean Distance Error



$t(16) = 2.91, p < 0.02$

Mean Angular Error



$t(16) = 3.27, p = 0.01$

Why Virtual Reality?

- VR/AR is the ***next computing platform***, following on the mainframe, desktop, and mobile
- VR ***solves the problems*** that traditionally plagued 3-D visualization: occlusion, perspective, navigation, etc.
- Immersion gives a ***qualitatively different perception*** of the patterns present in the data
 - The key concepts are *proprioception* (sense of the relative position) and *kinesthesia* (movement sense)
- VR is a natural platform for ***a collaborative visual exploration*** and collaboration
- ***Leverages*** a multi- $\$Z$ investment by the games industry



VR/AR as the New Computing/ Information/Communication Platform



From Mainframes...

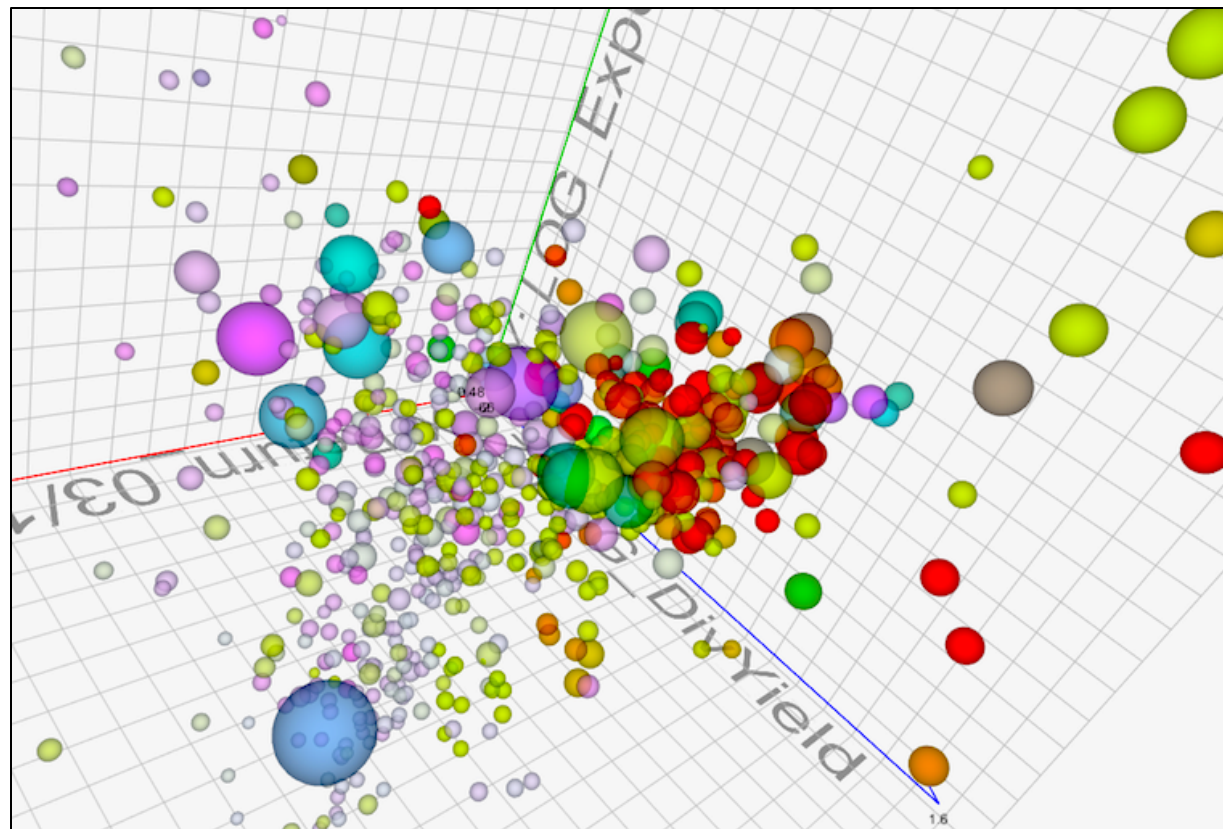
... to VR/AR Headsets

Increasing computing power, usability, fidelity,
information content and rate, immediacy

We don't use computers only to compute – we use
them to access information and to communicate

Quantitative Improvements

- Preliminary tests using both artificial and real data indicate that *dramatic improvements* are possible in time-to-insight, as compared to the traditional data analysis/visualization tools (e.g., Excel): from days/months to minutes/hours in some cases
- Some results are simply impossible to achieve using traditional tools, due to the projection effects
- Work still in progress



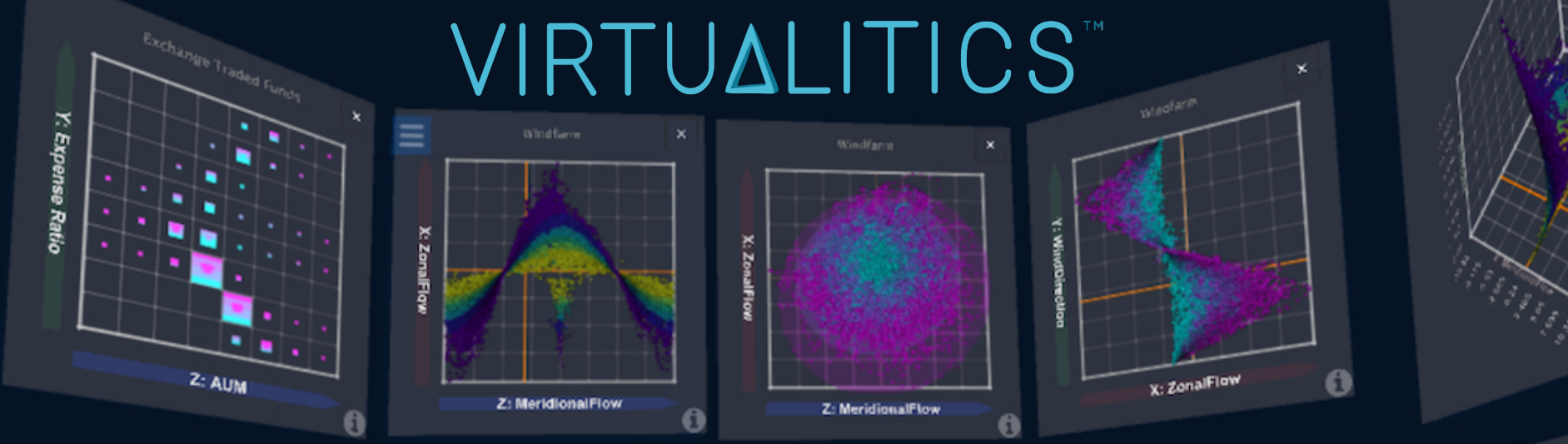
Keck Institute for Space Studies Symposium on Virtual and Augmented Reality for Space Science and Exploration

Caltech, Jan. 30, 2018

Videos: www.kiss.caltech.edu/symposia/space_science



VIRTUALITICS™



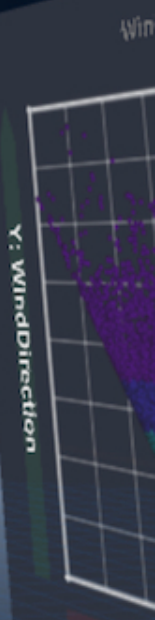
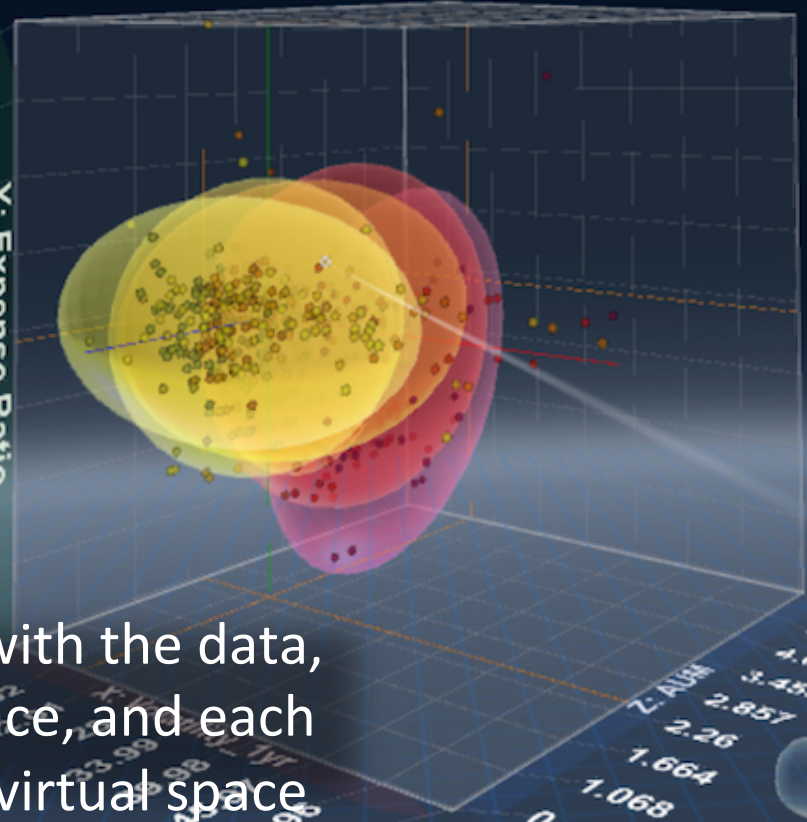
Ciro



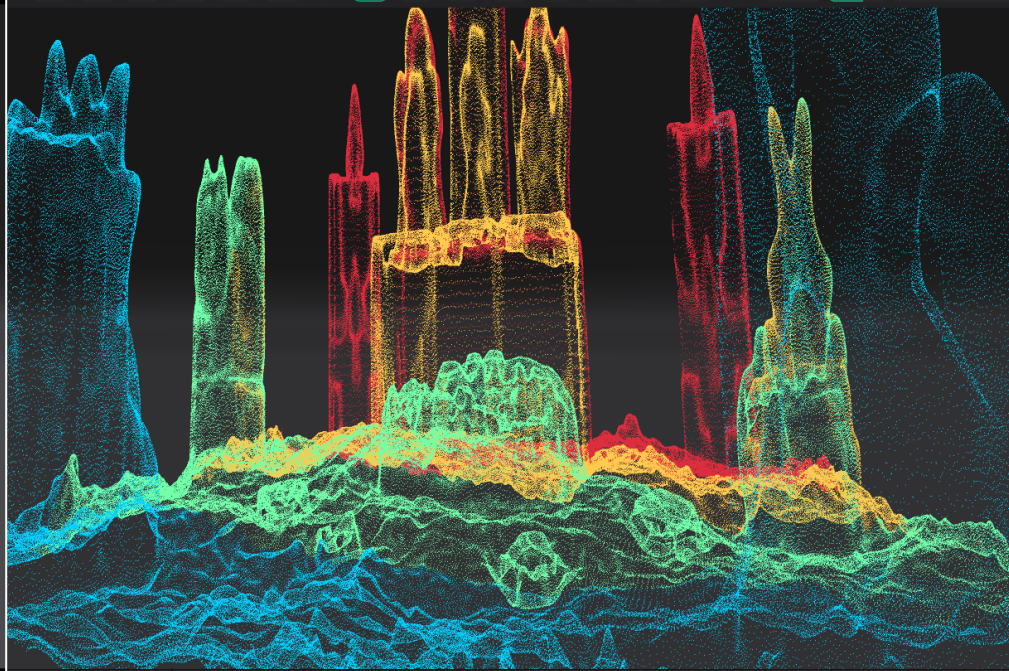
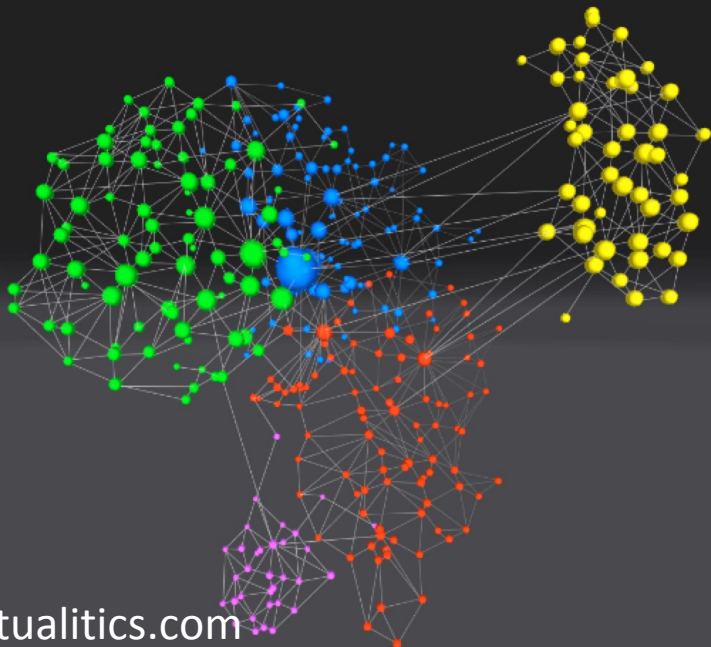
Michael



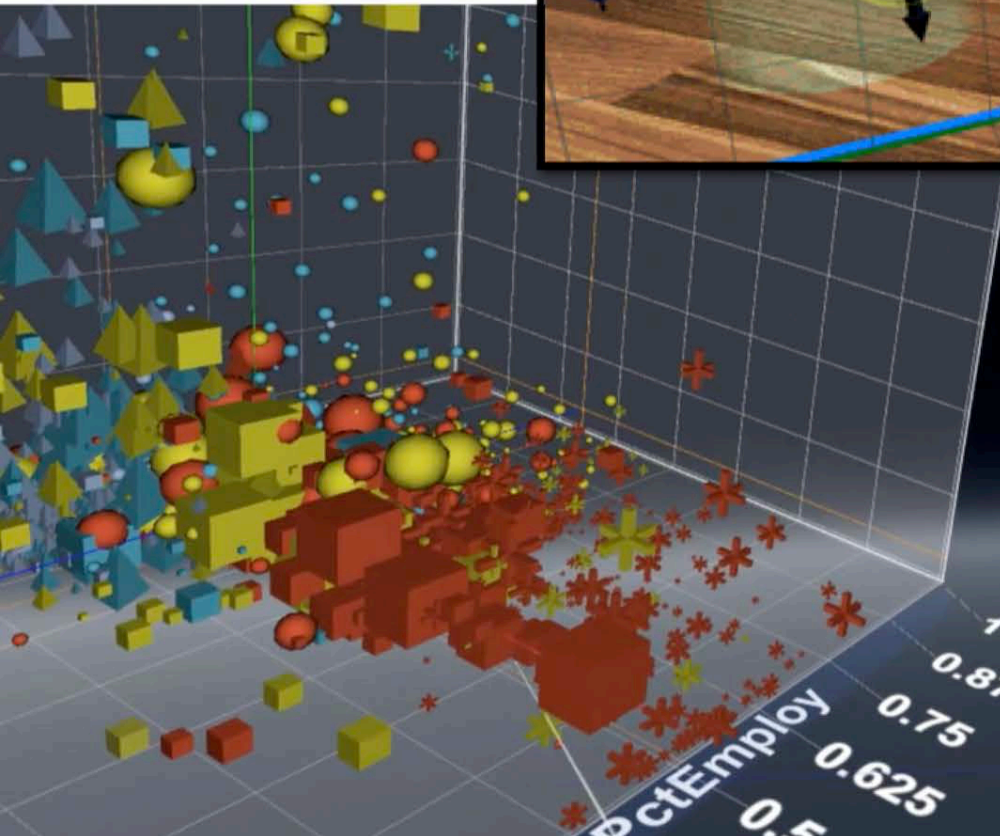
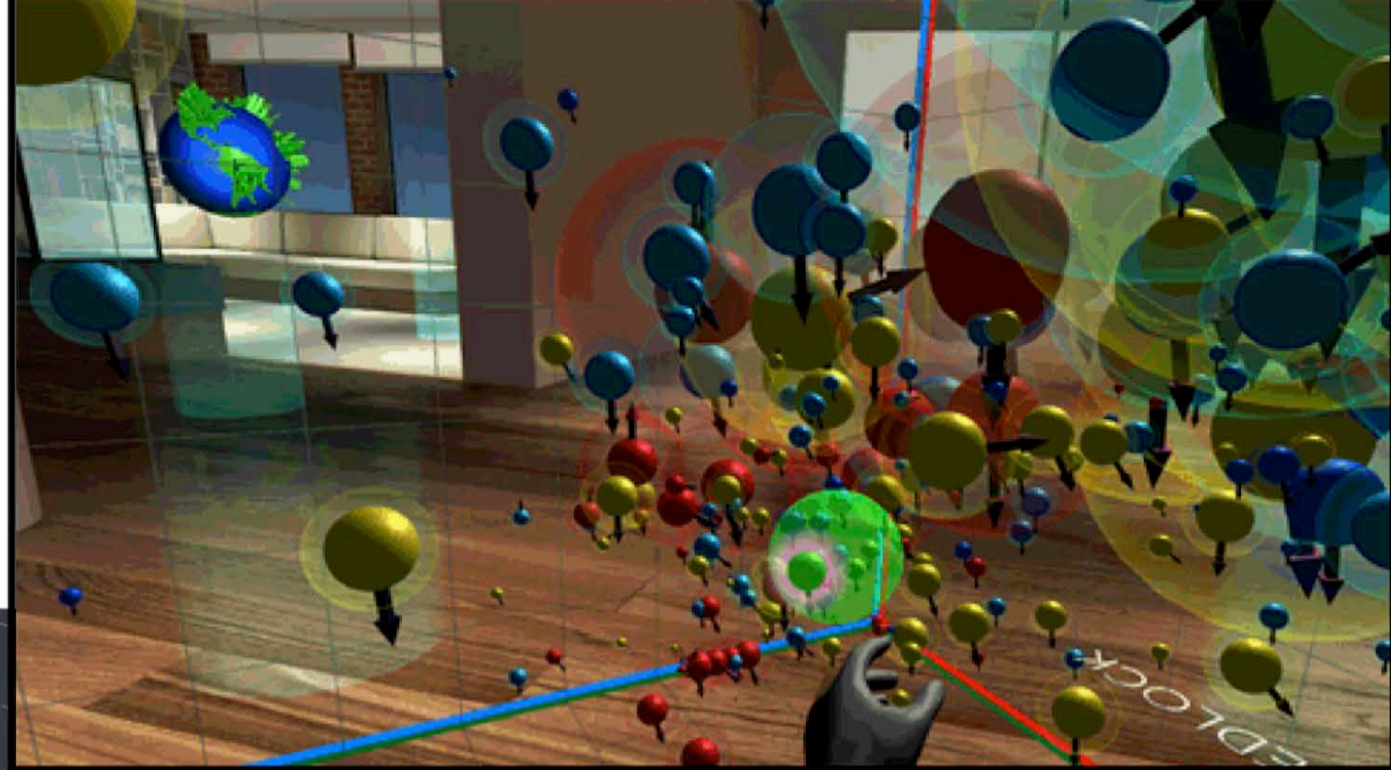
Users interacting with the data, machine intelligence, and each other in a shared virtual space



Different types of data visualized in VR



Interacting with data in VR



Legend & Insights

Mapping

Legend Insights

Smart Mapping

Cluster

Outlier Detection

Datasets History

Tools

Color

Shape

Show Insights

Color	Shape	Count
[0, 0.2]	497/499	
[0.2, 0.33]	496/499	
[0.33, 0.44]	494/499	
[0.52, 1]	489/497	

Insights

Tip: Transform axes by HST

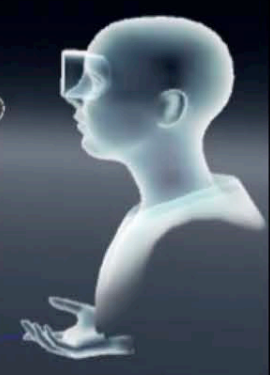
Press A to Select

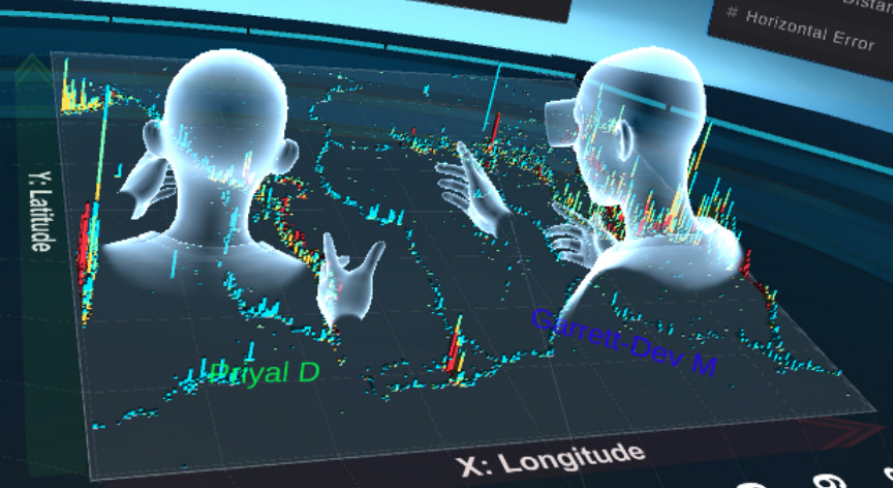
When [0, 0.2] is Below Median, and [0.2, 0.33] is Below Median, then: Out of 383 records, 60% are [0, 0.2], vs. 25% overall

Press A to Select

When [0.33, 0.44] is Above Median, and [0.52, 1] is Below Median, then: Out of 492 records, 52% are [0.52, 1], vs. 25% overall

Press A to Select





Collaborative Data Visualization in VR

VR+ML platform API smooth interaction with a Python notebook (and other popular data analytics platforms)

The image displays a dual-screen setup. The left screen shows a Jupyter notebook with the following content:

```
Run Smart Mapping and Export Results
```

```
In [4]: # Smart Mapping with target feature of the Price-to-Sales ratio to determine most
```

```
features = list(df_orig)
features.remove('Company Name')
features.remove('Exchange:Ticker')
features.remove('PS')
features.remove('rand')
features.remove('link')
```

```
vip.smart_mapping(df_orig, "PS", features, apply=False)
vip.normalize("softmax", apply=True)
```

Smart Mapping completed successfully with target Feature PS.

Ranked list of Features:

- 1: Market Debt to capital ratio
- 2: Bottom up levered beta
- 3: Marginal Tax Rate
- 4: Broad_Group
- 5: Effective Tax Rate
- 6: Return on Capital (ROC or ROIC)
- 7: Cost of equity in US\$
- 8: Historical growth in Revenues - Last 3 years
- 9: Sub_Group
- 10: Historical growth in Revenues - Last 5 years

Correlated Groups:

- Group 1 -
 - 1: Bottom up levered beta
 - 2: Cost of equity in US\$
- Group 2 -
 - 1: ERP for Country
 - 2: ERP for Country
- Group 3 -
 - 1: Total Default Spread for cost of debt (Company + Country)
 - 2: Pre-tax cost of debt in US\$
 - 3: After-tax cost of debt in US\$
- Group 4 -

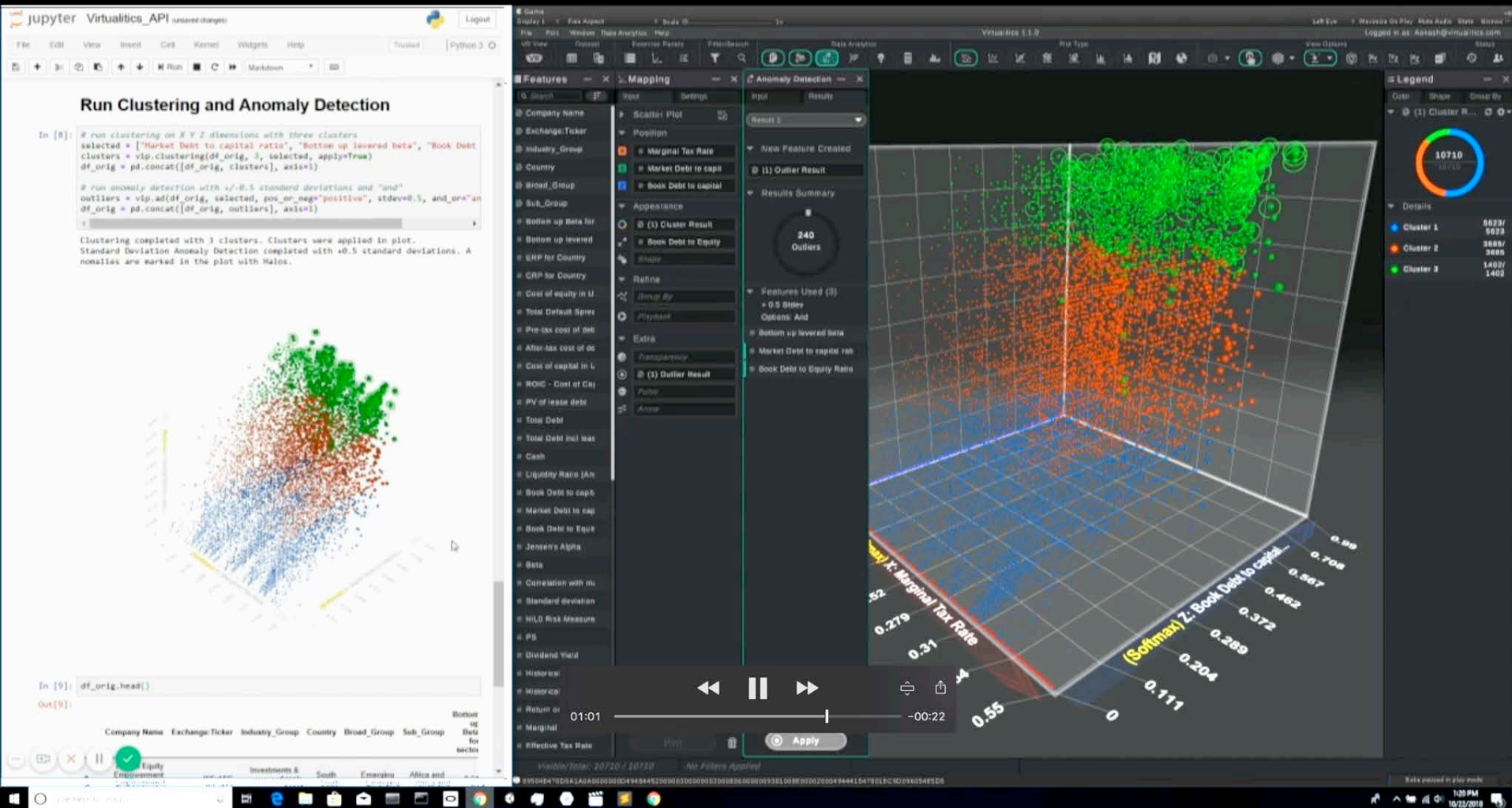
```
Run Clustering and Anomaly Detection
```

```
In [ ]: # run clustering on X Y Z dimensions with three clusters
selected = ["Market Debt to capital ratio", "Bottom up levered beta", "Book Debt to capital ratio"]
clusters = vip.clustering(df_orig, 3, selected, apply=True)
df_orig = pd.concat([df_orig, clusters], axis=1)
```

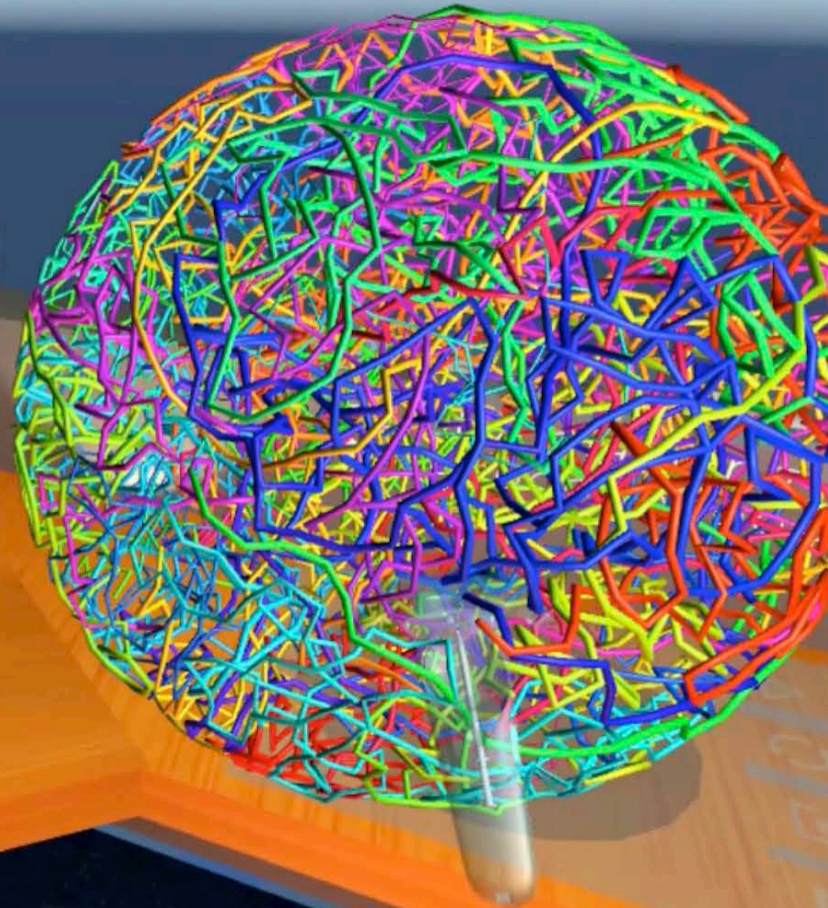
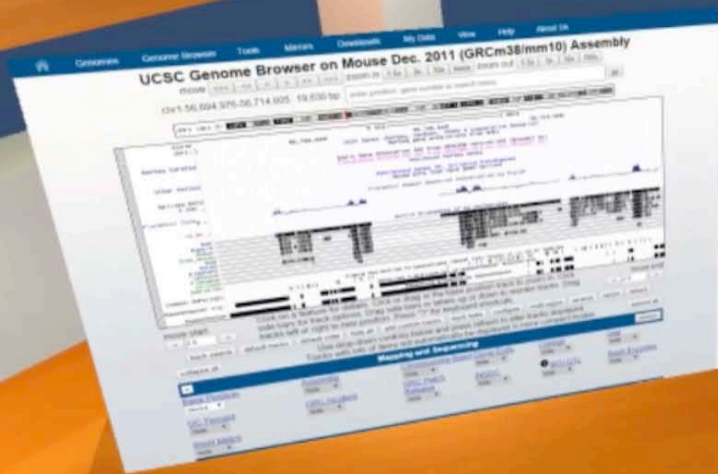
```
# run anomaly detection with +/-0.5 standard deviations and "and"
outliers = vip.ad(df_orig, selected, pos_or_neg="positive", stdev=0.5, and_or="and")
df_orig = pd.concat([df_orig, outliers], axis=1)
```

The right screen shows a VR interface with a 3D scatter plot of data points. The plot is titled "Smart Mapping" and shows a dense cluster of points in a 3D space. The axes are labeled with feature names: "Market Debt to capital ratio" (X-axis), "Marginal Tax Rate" (Y-axis), and "Softmax" (Z-axis). A legend on the right side of the plot shows a color scale for "PS" ranging from 0.05 to 3.84. The plot is viewed through a VR headset, with a virtual hand visible on the right side. The interface includes a "Features" panel on the left, a "Mapping" panel in the center, and a "Legend" panel on the right. The bottom of the screen shows a video player interface with a play button and a progress bar.

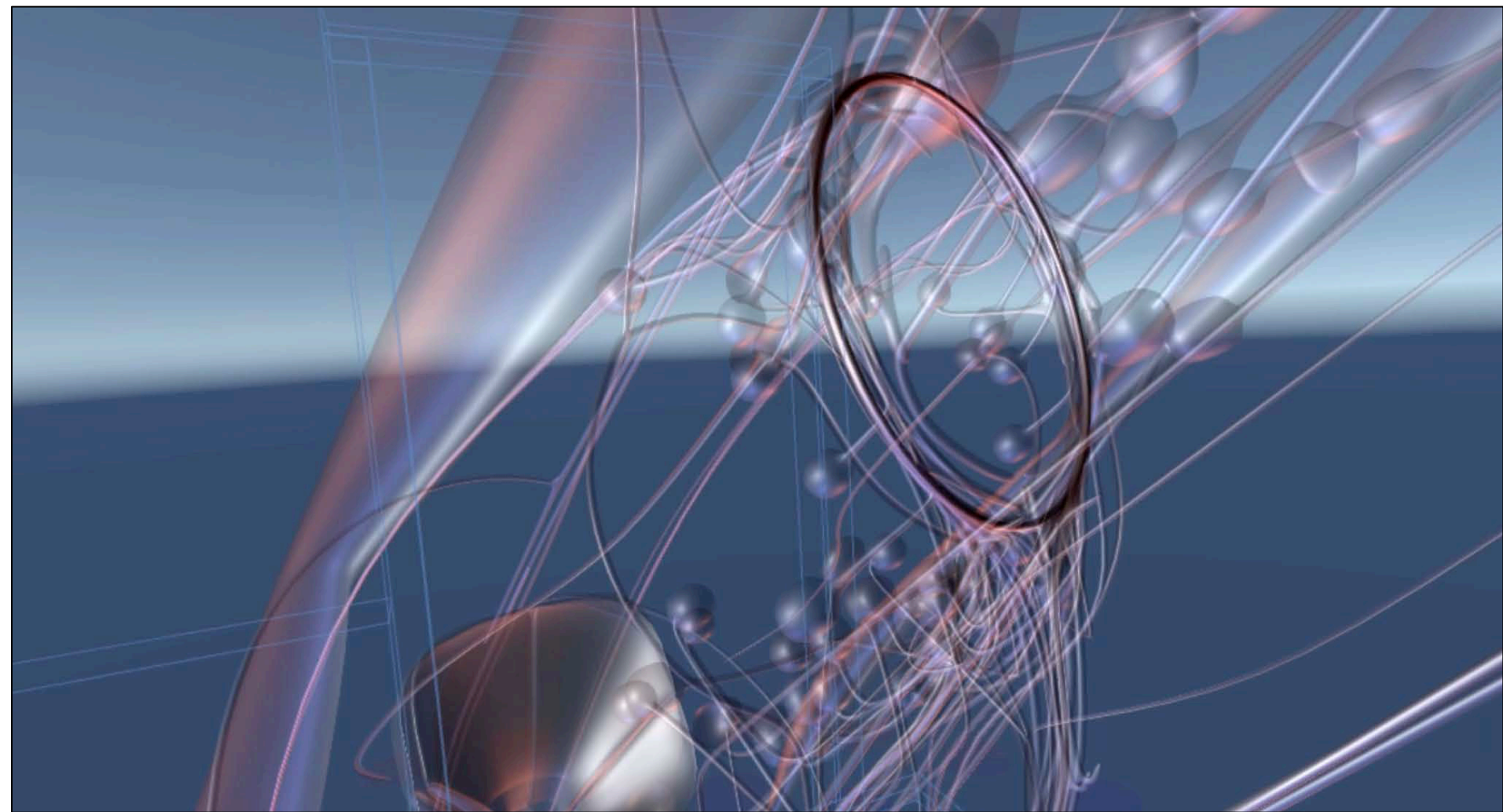
Export the results back to a Python notebook (publication quality 3D plots)



3D Mapping of the DNA: Santiago Lombeyda, CD3, and Mitch Guttman, Biology, Caltech



Visualising *C. Elegans*: Santiago Lombeyda, CD3, and Paul Sternberg, Biology, Caltech

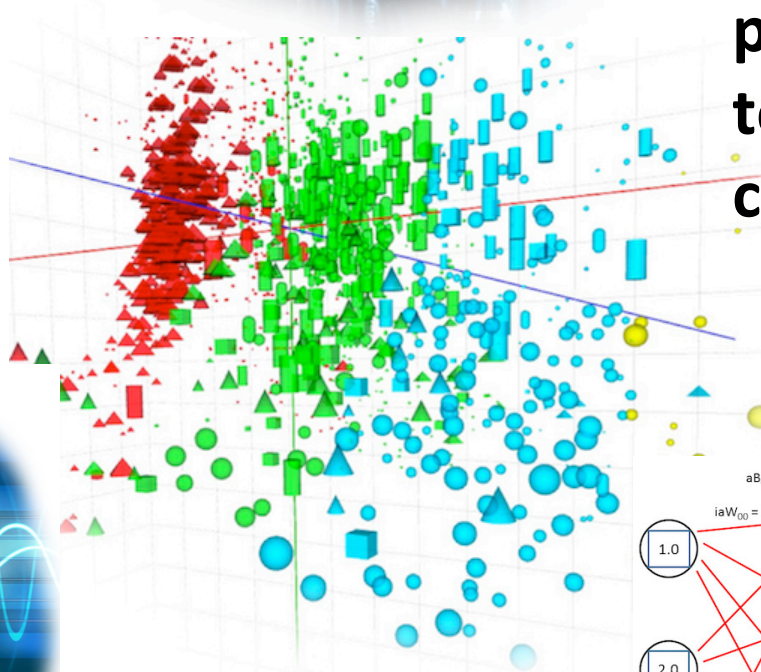


Virtual Teaching Labs

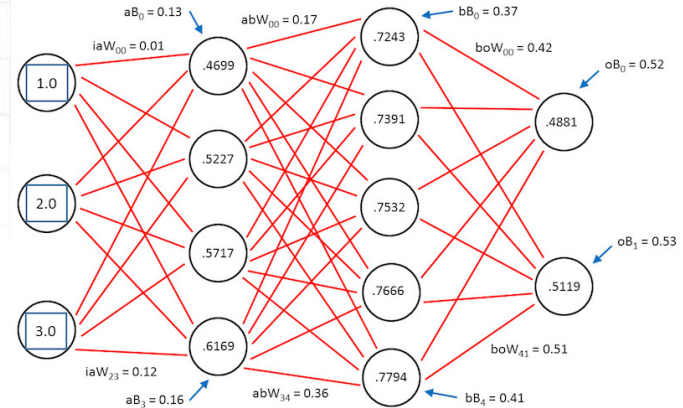
- VR enables a better comprehension and recall of the complex patterns and simulated phenomena
- You can do things in VR that are impossible (or too dangerous) in real life: from building or repairing nuclear reactors to entering the cells and the molecules
- This solves one of the major challenges of on-line education



Visualization is at the nexus of **Data**, **Human pattern recognition and understanding**, and **Machine Intelligence**

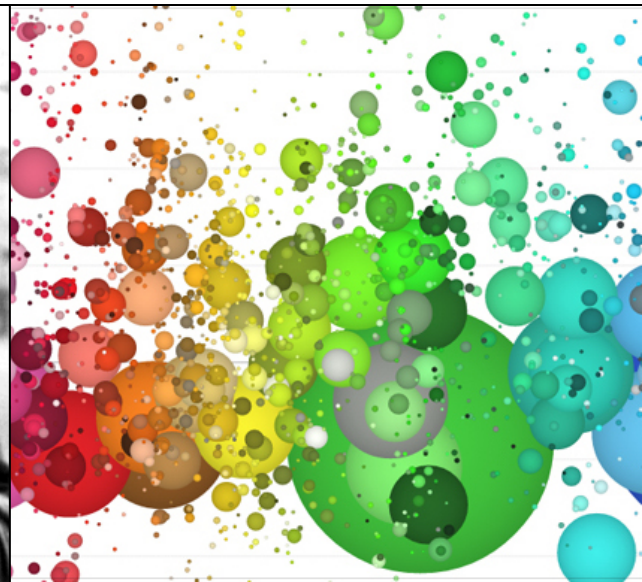
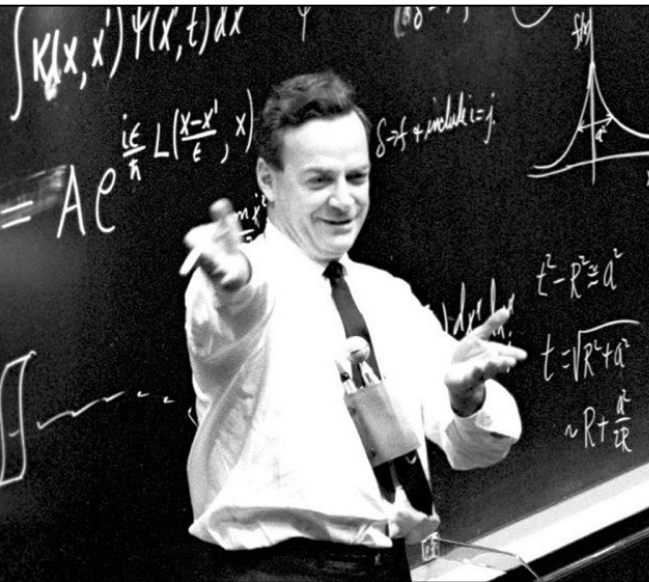


VR/AR is the most powerful platform to date that connects them



Ideas, Discoveries, and Learning Occur at the Interfaces

- Between the human minds, theoretical constructs, and data/information
- Between different fields or domains
- Improving technologies for communication and information access facilitate these interactions



Technology changes how we communicate and convey information



Increasing immediacy, increasing fidelity

What comes after the Web and the Internet?



Looking Ahead



- **VR will become more pervasive and better, driven largely by the entertainment industries, but other domains will follow**
 - ✧ This will be a natural technology evolution for the digital natives (the future workforce)
- **AI will increasingly permeate all aspects of the modern society, science included**
 - ✧ It will be essential for a rapid knowledge discovery

VR is the natural interaction environment for the humans and information technology

AI + VR = Cognition Technology

Summary

- Effective data visualization is an essential component of data exploration and discovery
 - Especially when coupled with machine learning
- Most off-the-shelf data visualization tools are fairly limited, and/or poorly designed
- Learn how to design your data visualizations well
- Visualization of high-dimensionality data spaces may be ***the key bottleneck of data-driven discovery***
 - The challenge is not data size, it is ***data complexity***
- ***Virtual Reality*** is a powerful, intuitive new platform for multi-dimensional, collaborative data exploration and visual analytics
 - It is not a game any more...*

Information Dashboard Design

STEPHEN FEW

Visualizing Data

BEN FRY

Visual Explanations

EDWARD TUFTE

Envisioning Information

EDWARD TUFTE

The Visual Display of Quantitative Information

EDWARD TUFTE

Visual Strategies: A Practical Guide to Graphics for Scientists and Engineers

FELICE FRANKEL + ANGELA DEPACE

Information Visualization: Perception for Design

COLIN WARE

Visual Thinking for for Design

COLIN WARE

Interactive Visualization—Insights into Inquiry

BILL FERSTER

