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

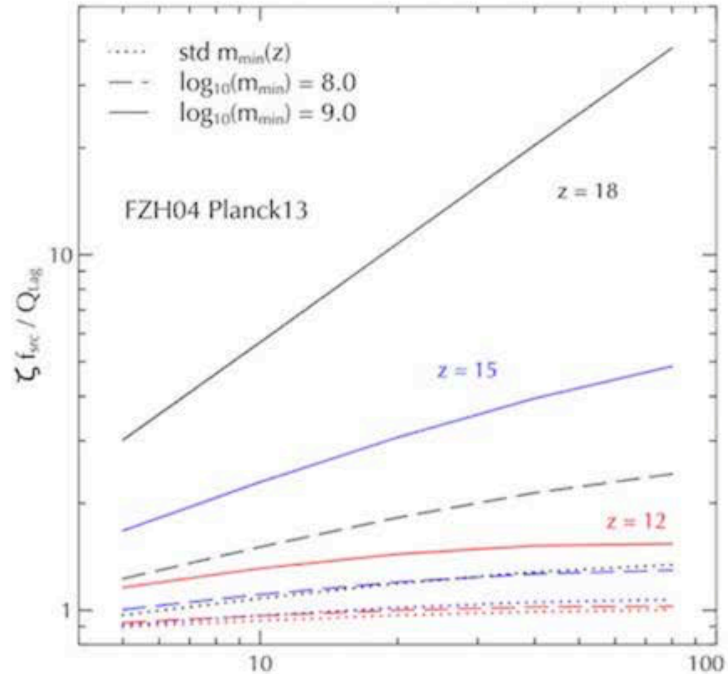
Never Do This!

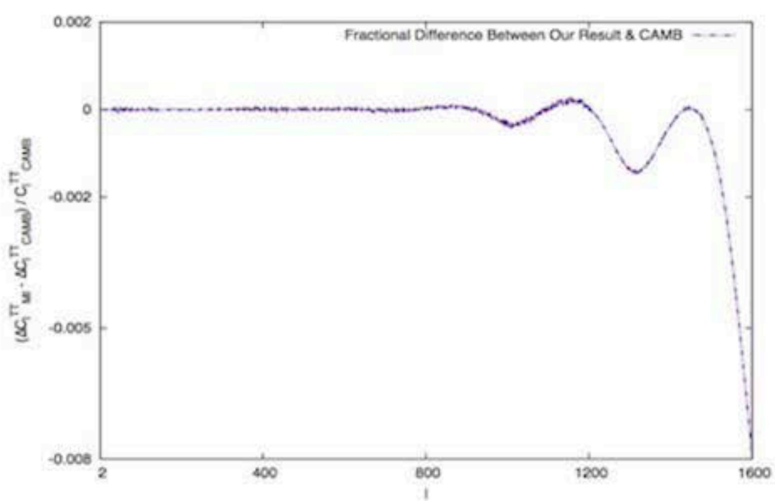
A figure made for a print may not look good on the screen:

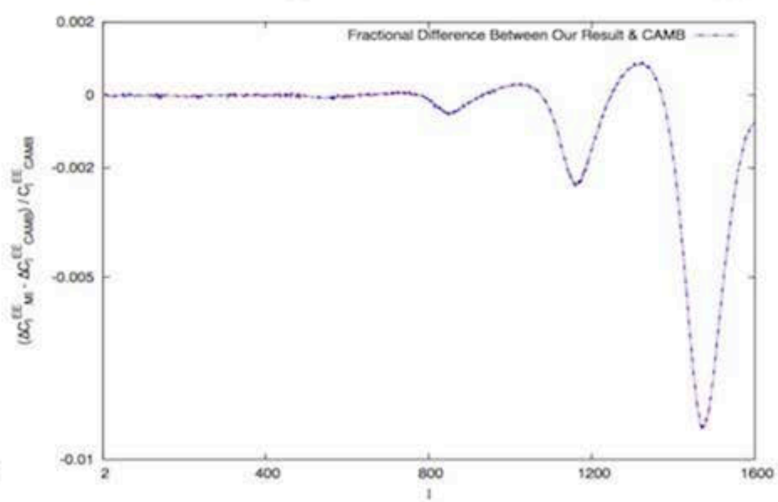
Paper ~ 5000 by ~ 6500 pixels

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?

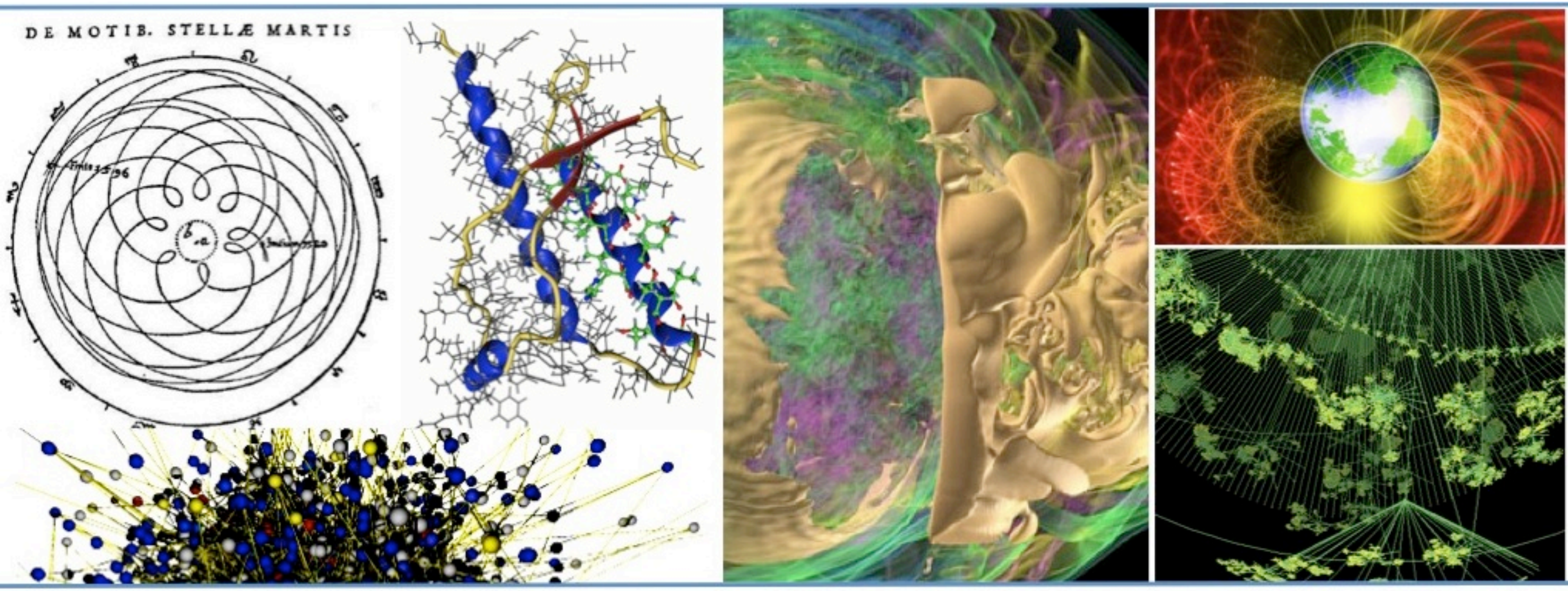
Increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems.

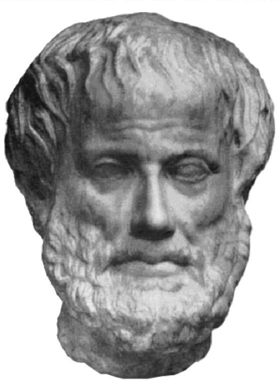
Douglas Engelbart

Augmenting Human Intellect:

A Conceptual Framework

Effective visualization is the bridge between quantitative information and human intuition



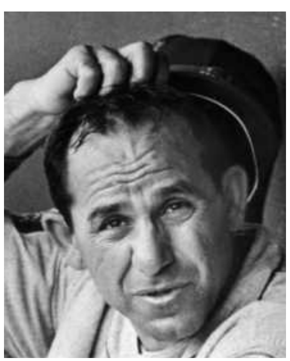


Man cannot understand without images

Aristotle, *De Memoria et Reminiscentia*

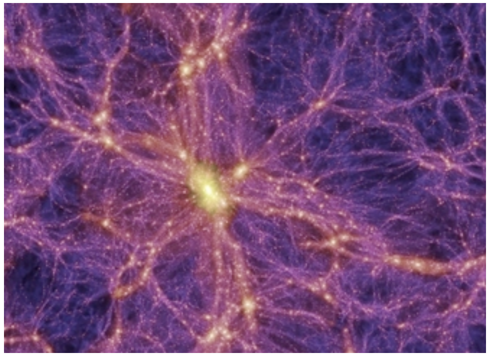
You can observe a lot just by watching

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



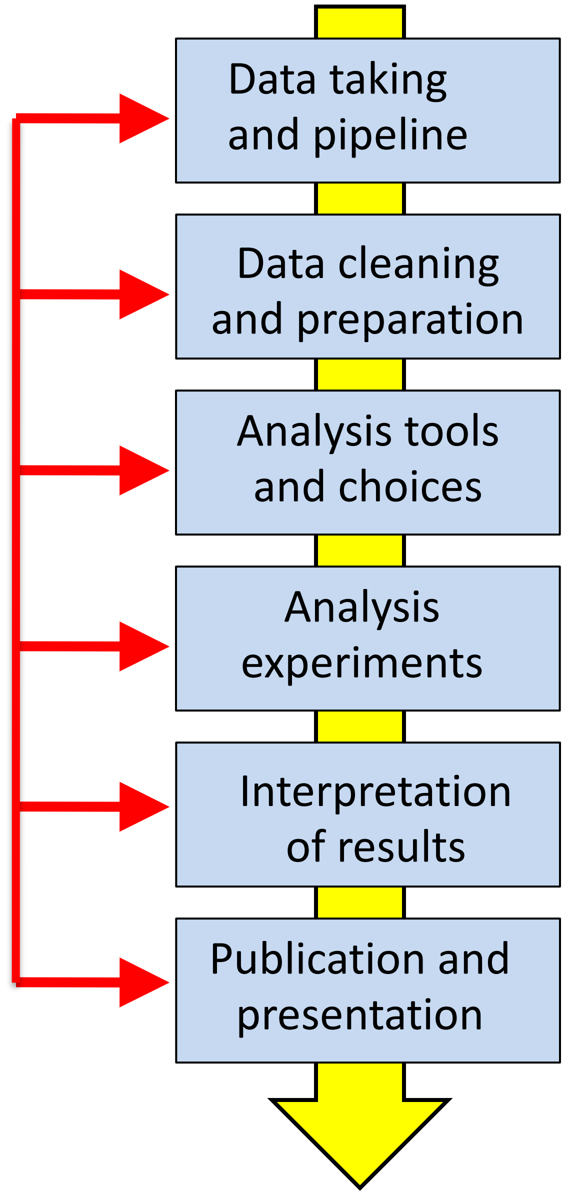
What good are the data if we cannot effectively extract knowledge from them?

"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

Visualization



Design, debugging, quality control

Data geometry and assumptions

Iterate and refine

Knowledge and understaning

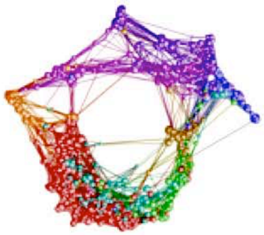
Data

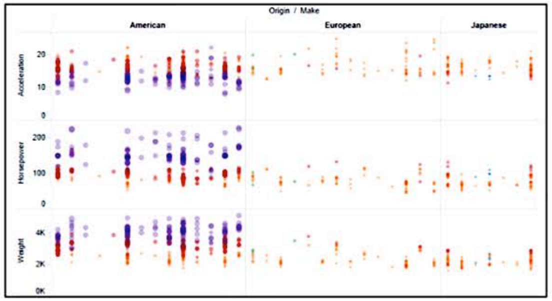
Unstructured
(e.g., social networks, text, …)

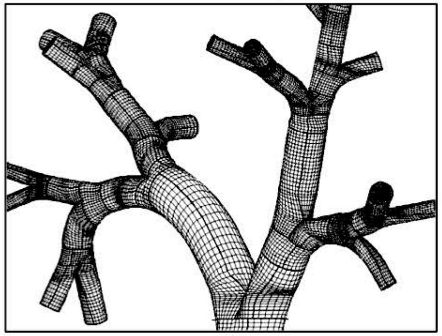
Structured (following some schema)

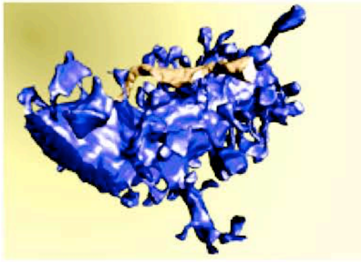
- Nominal
  (labels)
  e.g., Hubble types
- Ordinal
  (ordered categories)
  e.g., cluster richness class
- Quantitative
  (numerical, most measurements)
  - Continuous
    (uniform time series, images, data cubes, …)
  - Discrete
    (data points, feature vectors, …)
    Can have *many* dimensions

Geometric Structure of Data









UNSTRUCTURED

ABSTRACT DATA

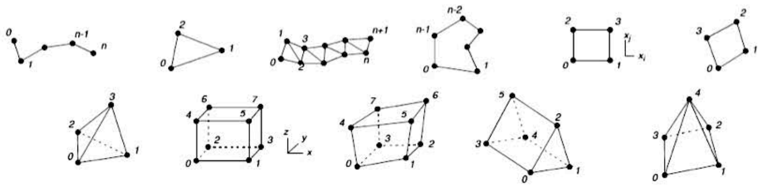
multi-dimensional data RECORDS

STRUCTURED

2D/3D DATA

scalar/vector/tensor + time

MPG Cylinders Horsepower Weight Acceleration Year Origin
8. 50.4 2.8 8.2 4 40. 250. 4 1500. 5500. 4 5. 30. 4 69.5 82.5 4 .8 3.2 3
18.000000 8.000000 130.000000 3504.000000 12.000000 70.000000 1.000000
15.000000 8.000000 165.000000 3693.000000 11.500000 70.000000 1.000000
18.000000 8.000000 150.000000 3436.000000 11.000000 70.000000 1.000000
16.000000 8.000000 150.000000 3433.000000 12.000000 70.000000 1.000000
17.000000 8.000000 140.000000 3449.000000 10.500000 70.000000 1.000000
15.000000 8.000000 165.000000 3693.000000 11.500000 70.000000 1.000000
18.000000 8.000000 150.000000 3436.000000 11.000000 70.000000 1.000000
16.000000 8.000000 150.000000 3433.000000 12.000000 70.000000 1.000000
17.000000 8.000000 140.000000 3449.000000 10.500000 70.000000 1.000000
16.000000 8.000000 150.000000 3433.000000 12.000000 70.000000 1.000000
... ...



The two kinds can be interchangeable

Slide: S. Lombeyda

From Data Space to Visualization Space

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

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

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

$$\downarrow \quad \downarrow \quad \downarrow \quad \quad \downarrow$$

Visualization space: { XYZ positions, point sizes, shapes, RGBα 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:

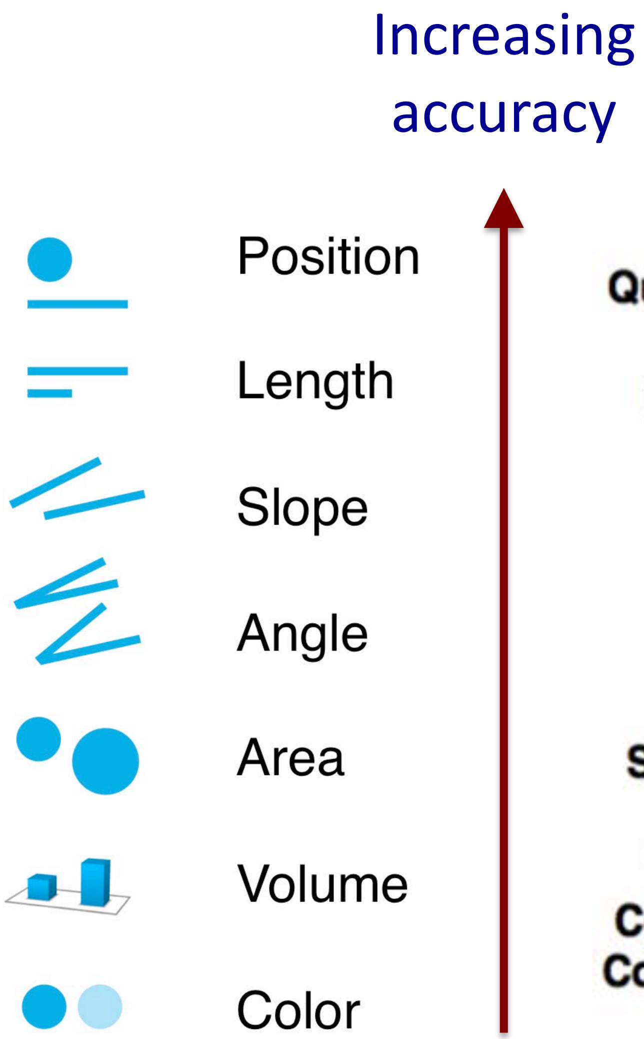
Principle of consistency

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

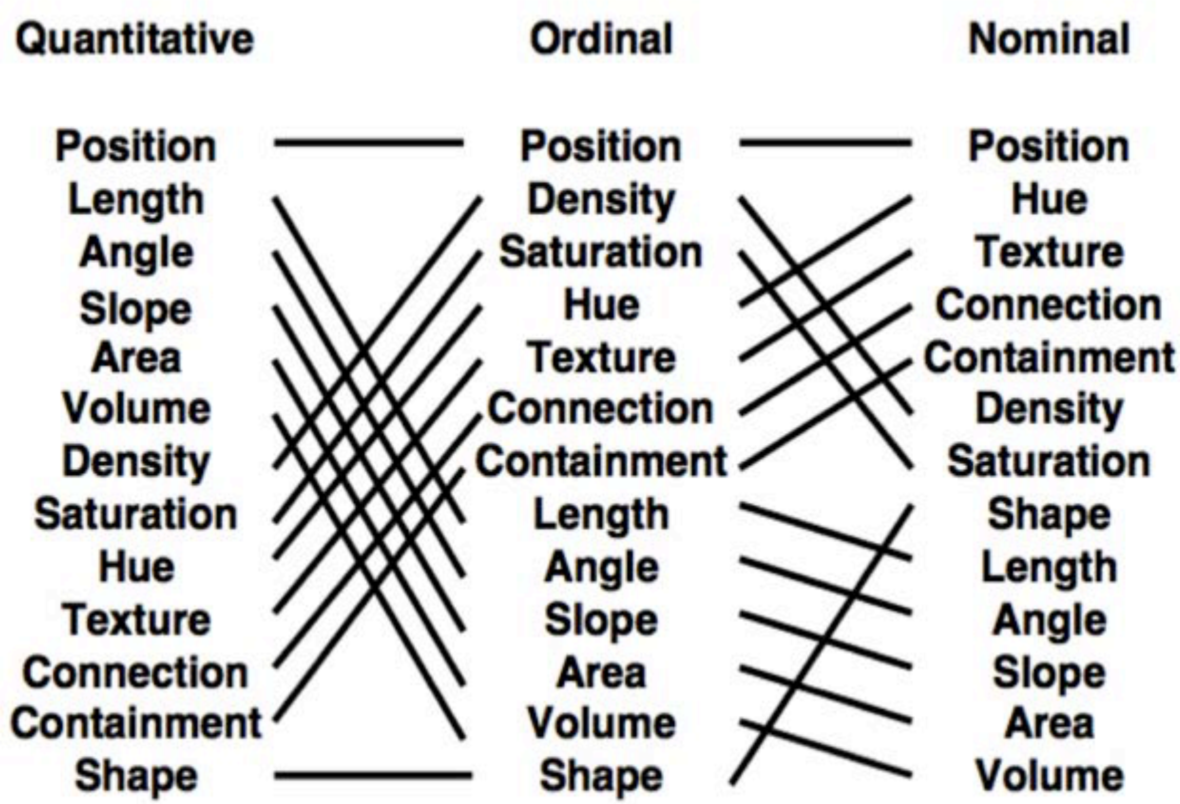
Principle of importance ordering

Encode the most important variables in the most effective way

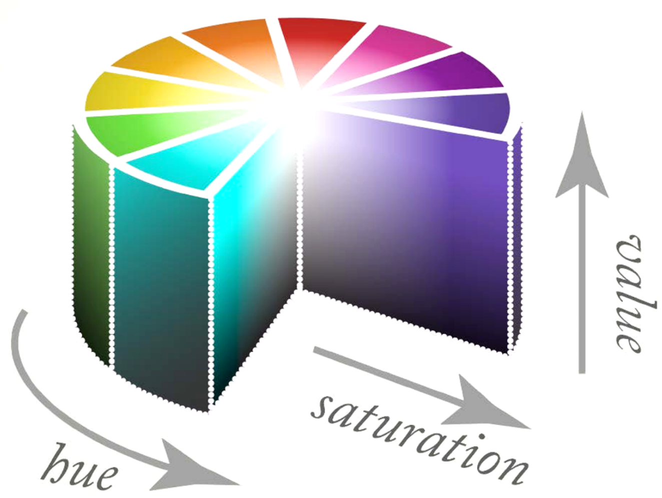
Map the most important variables to the visual "axis" that corresponds to the most accurate perception:



It also depends on the type of data:



How Many Dimensions for Color?

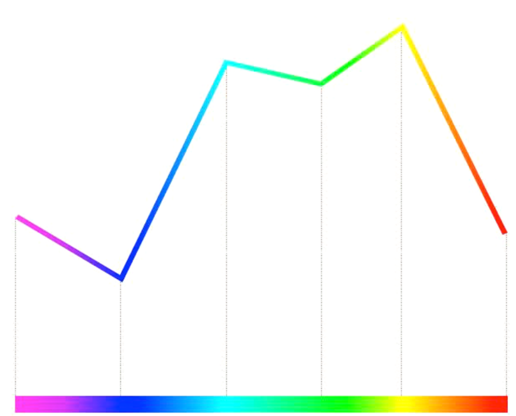


3? RGB or HSV

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

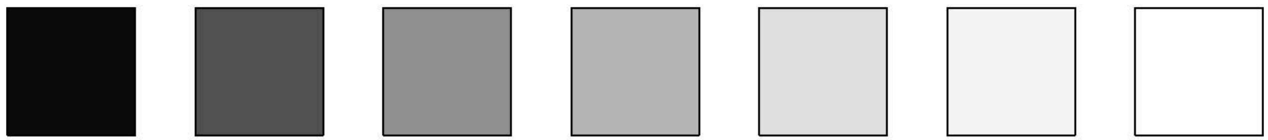
(from S. Lombeyda)

How Many Shades of Gray Can you Distinguish?

(from S. Davidoff)

How Many Shades of Gray Can you Distinguish?

Value easily encodes ordinal variables

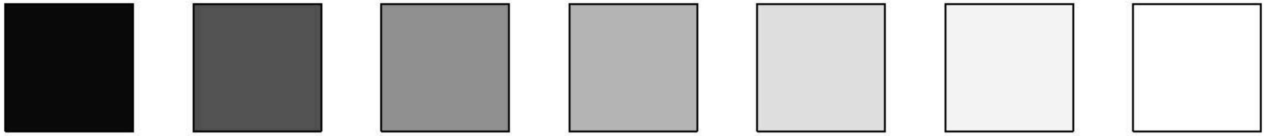


Value encodes continuous variables (less well)

(from S. Davidoff)

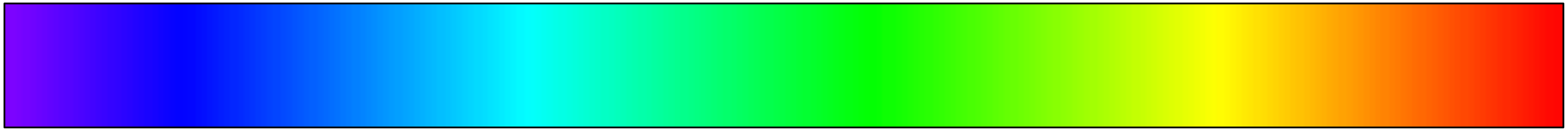
How Many Shades of Gray Can you Distinguish?

Value easily encodes ordinal variables



Value encodes continuous variables (less well)

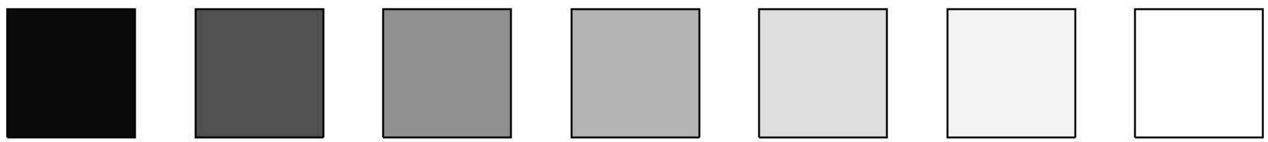
How Many Colors?



(from S. Davidoff)

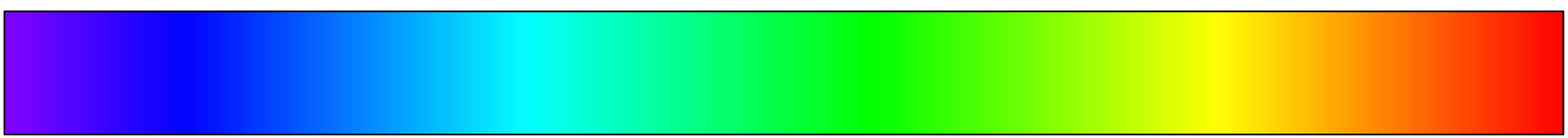
How Many Shades of Gray Can you Distinguish?

Value easily encodes ordinal variables

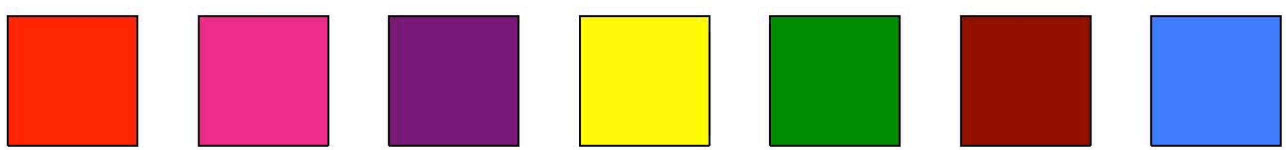


Value encodes continuous variables (less well)

How Many Colors?



Hue encodes nominal variables

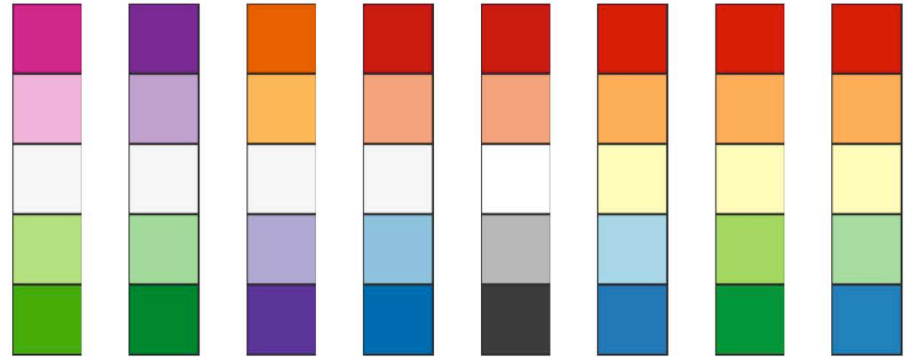


(from S. Davidoff)

Choosing the color palette

Discrete rather than continuous

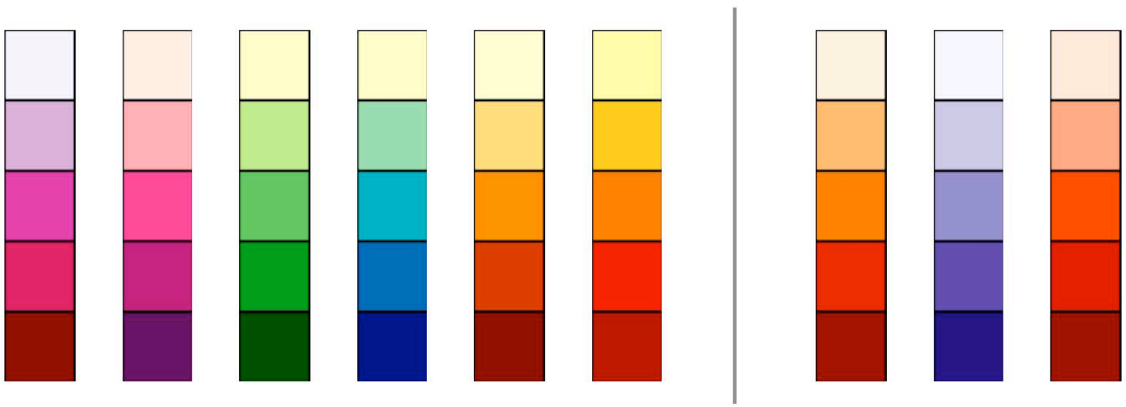
Diverging color



Data maps to meaningful mid-point

Color midpoint neutral, saturation at endpoints

Sequential color



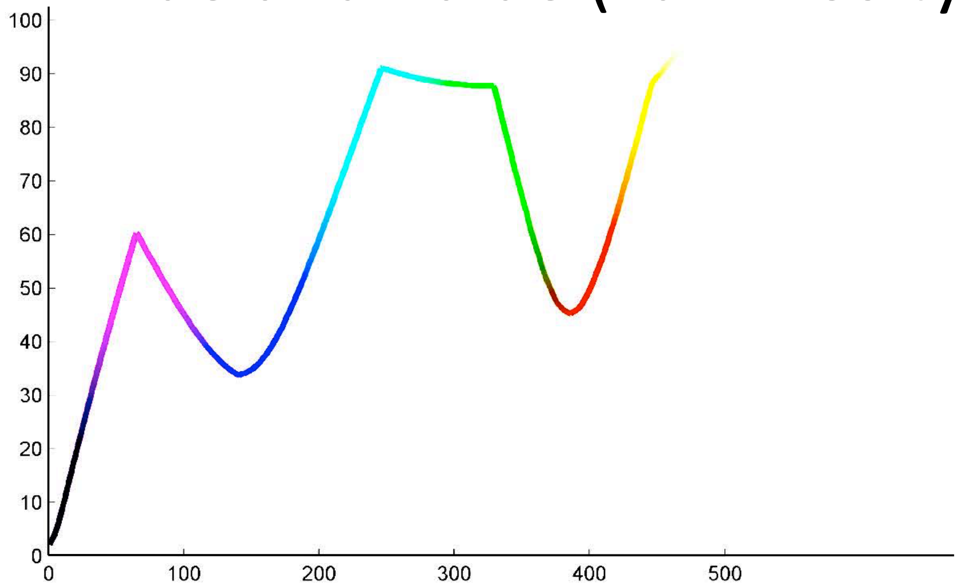
It depends on your purpose

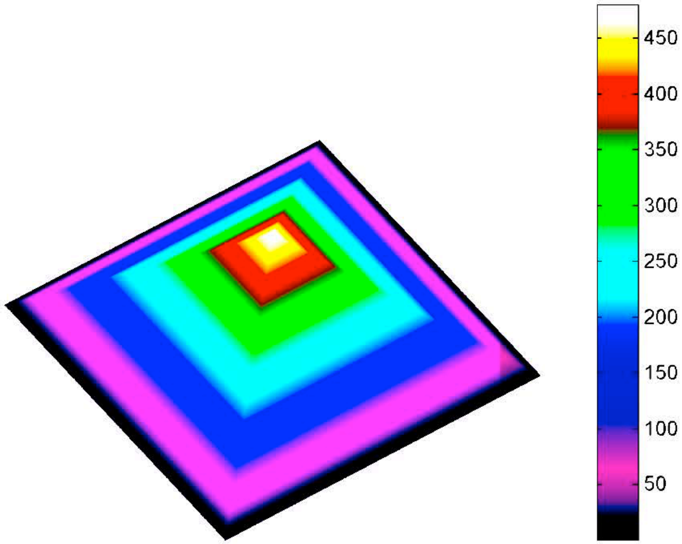
Vary luminance and saturation

Map higher values to darker colors

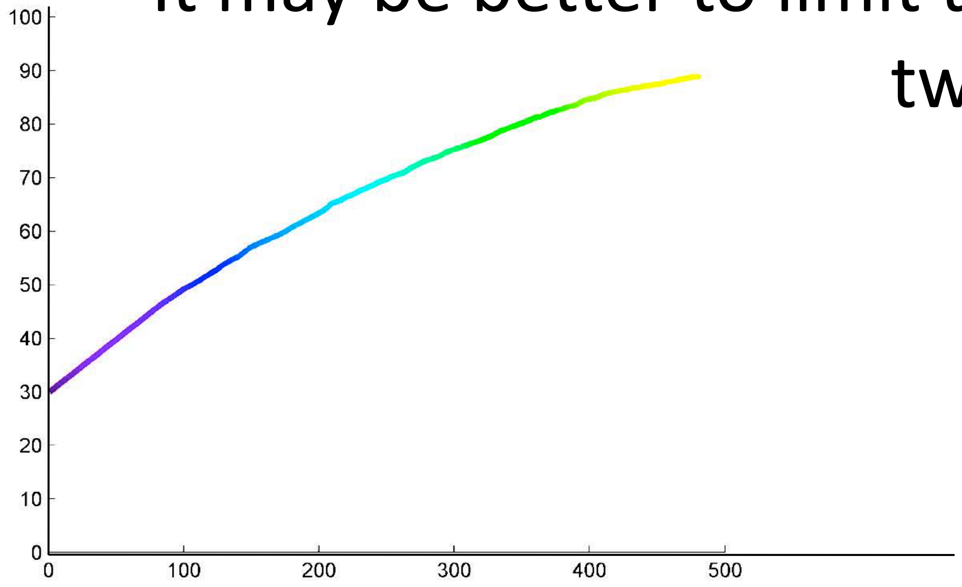
(from S. Davidoff)

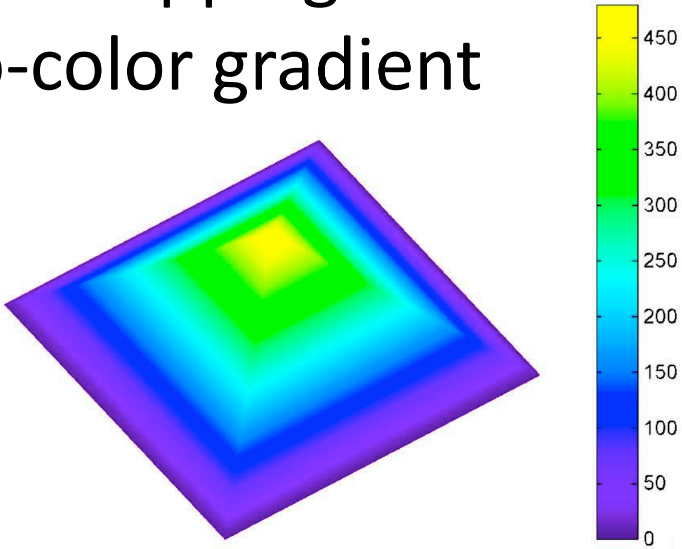
Hue and Value (Luminosity) can be confusing





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

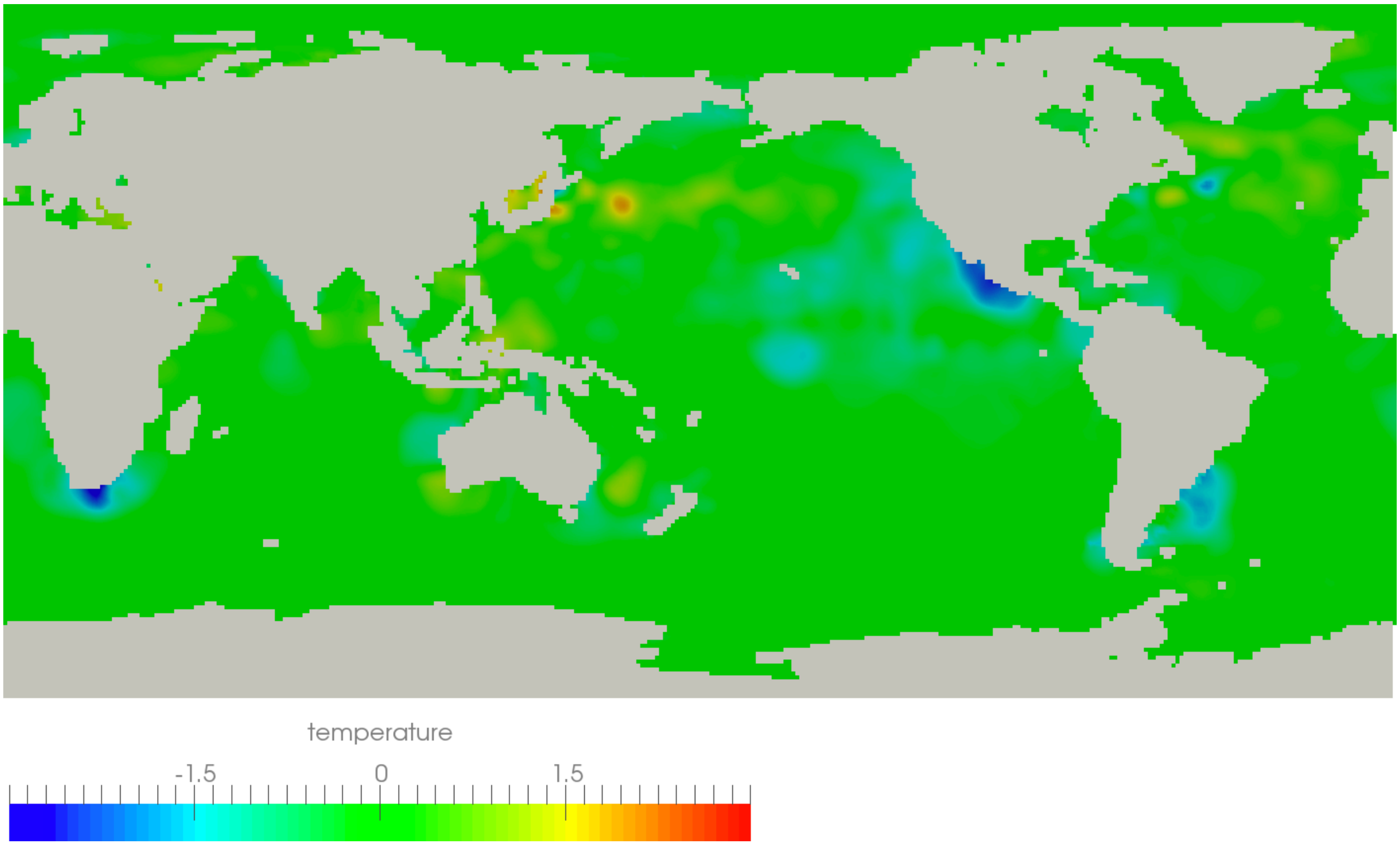




(from S. Lombeyda)

Color Map Choice Emphasizes Different Things

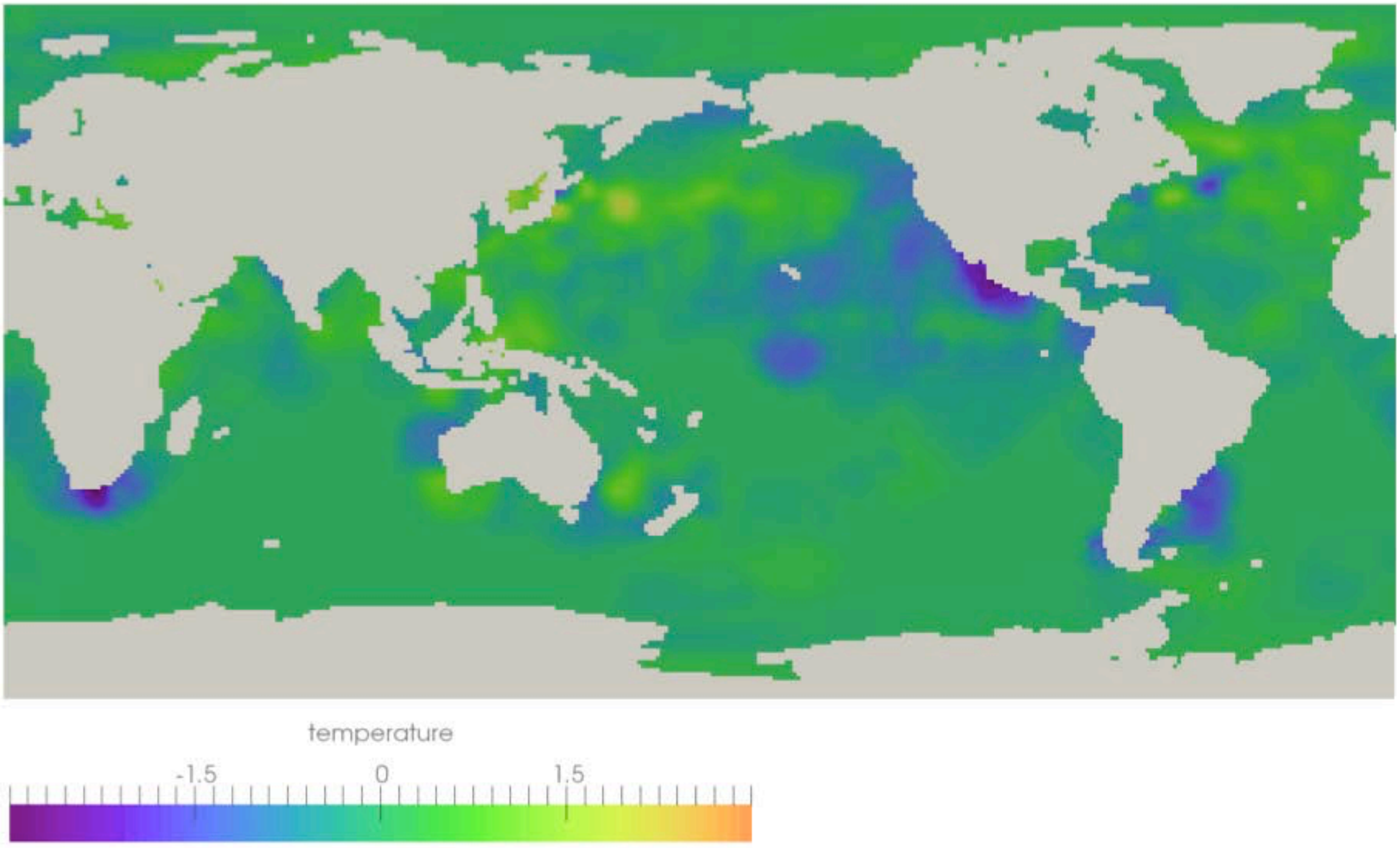
NOAA satellite data on the annual average ocean temperature at a 100m depth



Standard rainbow: dominated by the most common pixel values

Color Map Choice Emphasizes Different Things

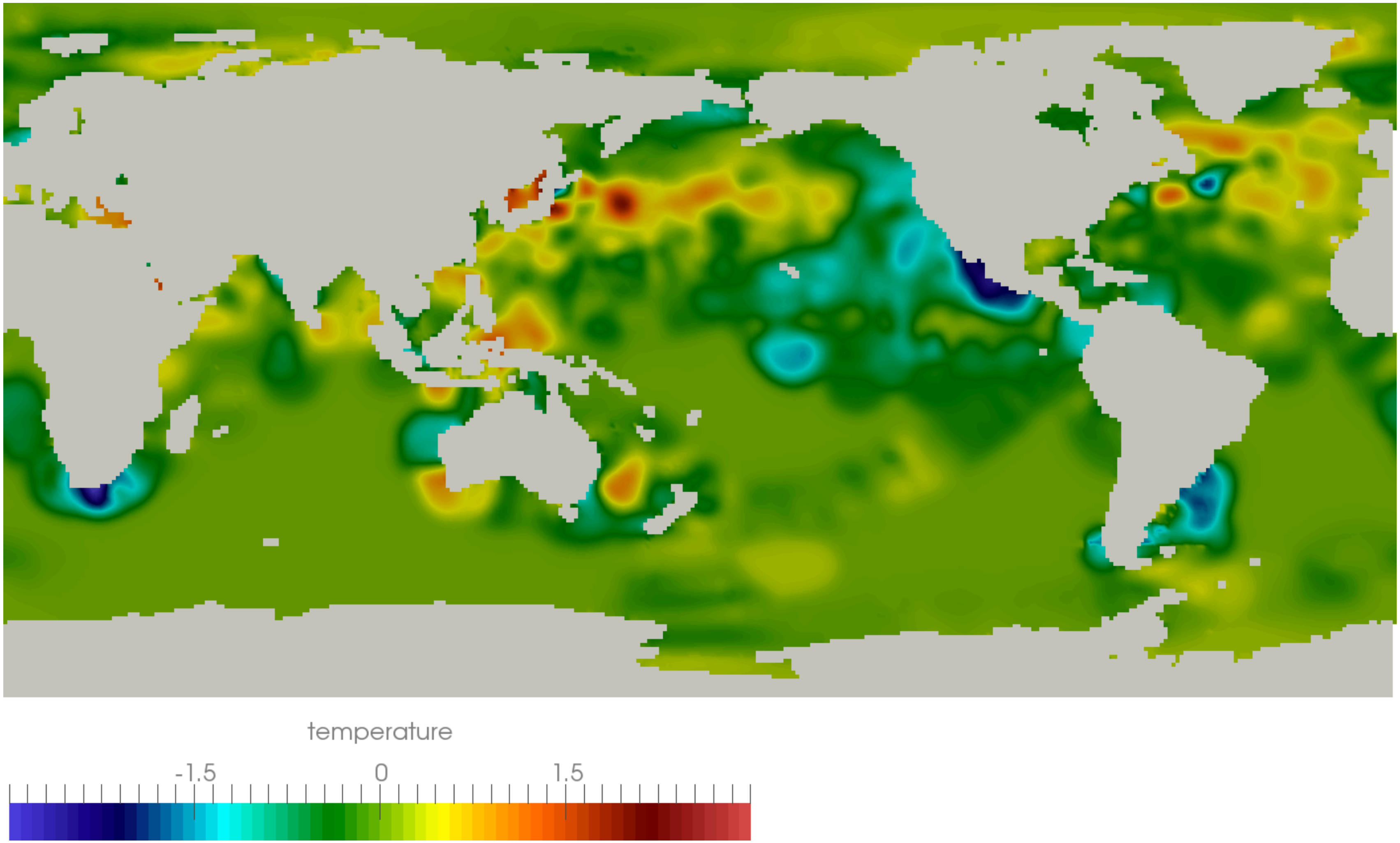
NOAA satellite data on the annual average ocean temperature at a 100m depth



Squeeze the middle to emphasize the tails of the distribution

Color Map Choice Emphasizes Different Things

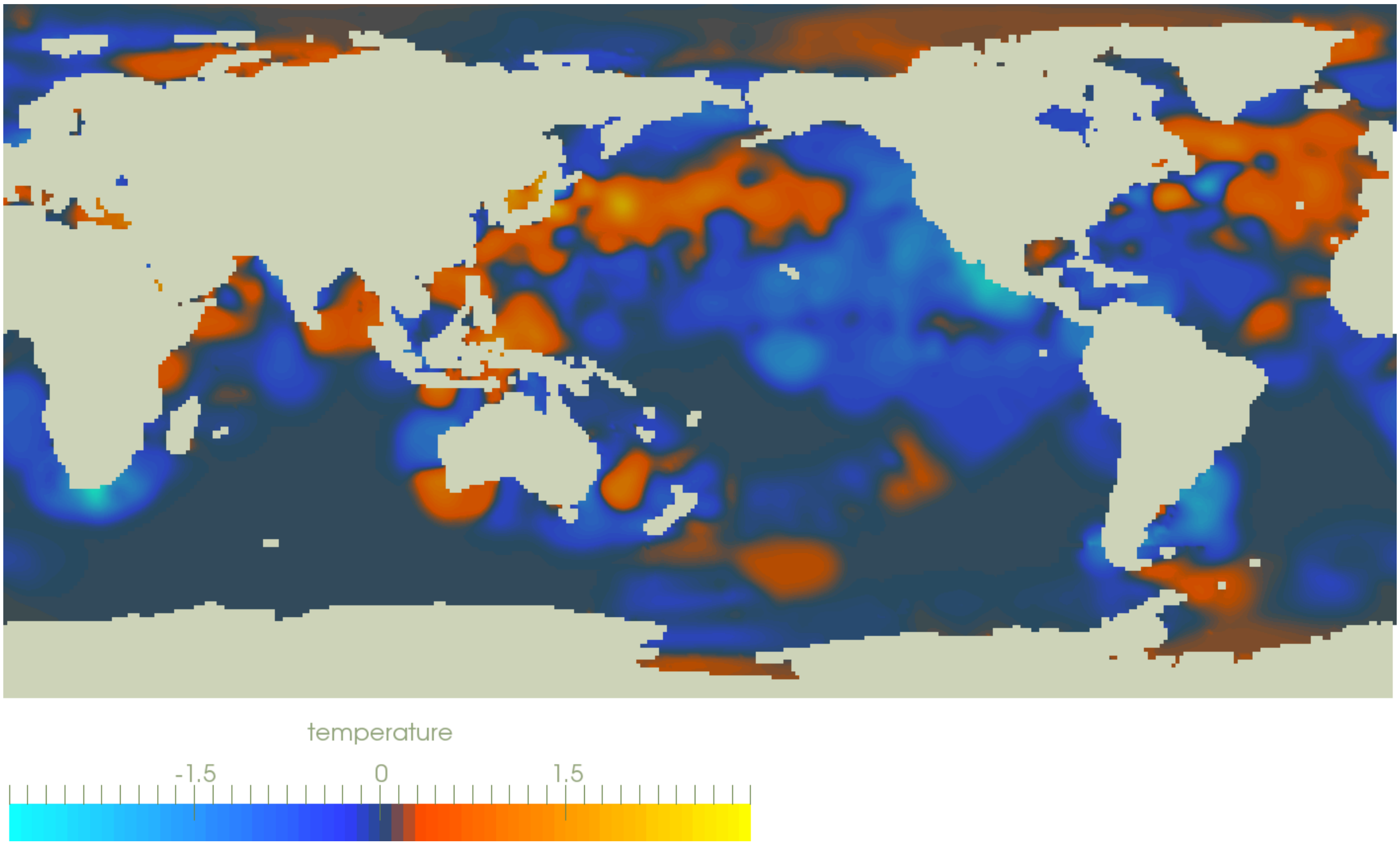
NOAA satellite data on the annual average ocean temperature at a 100m depth



Expand the ends to see most of the variations

Color Map Choice Emphasizes Different Things

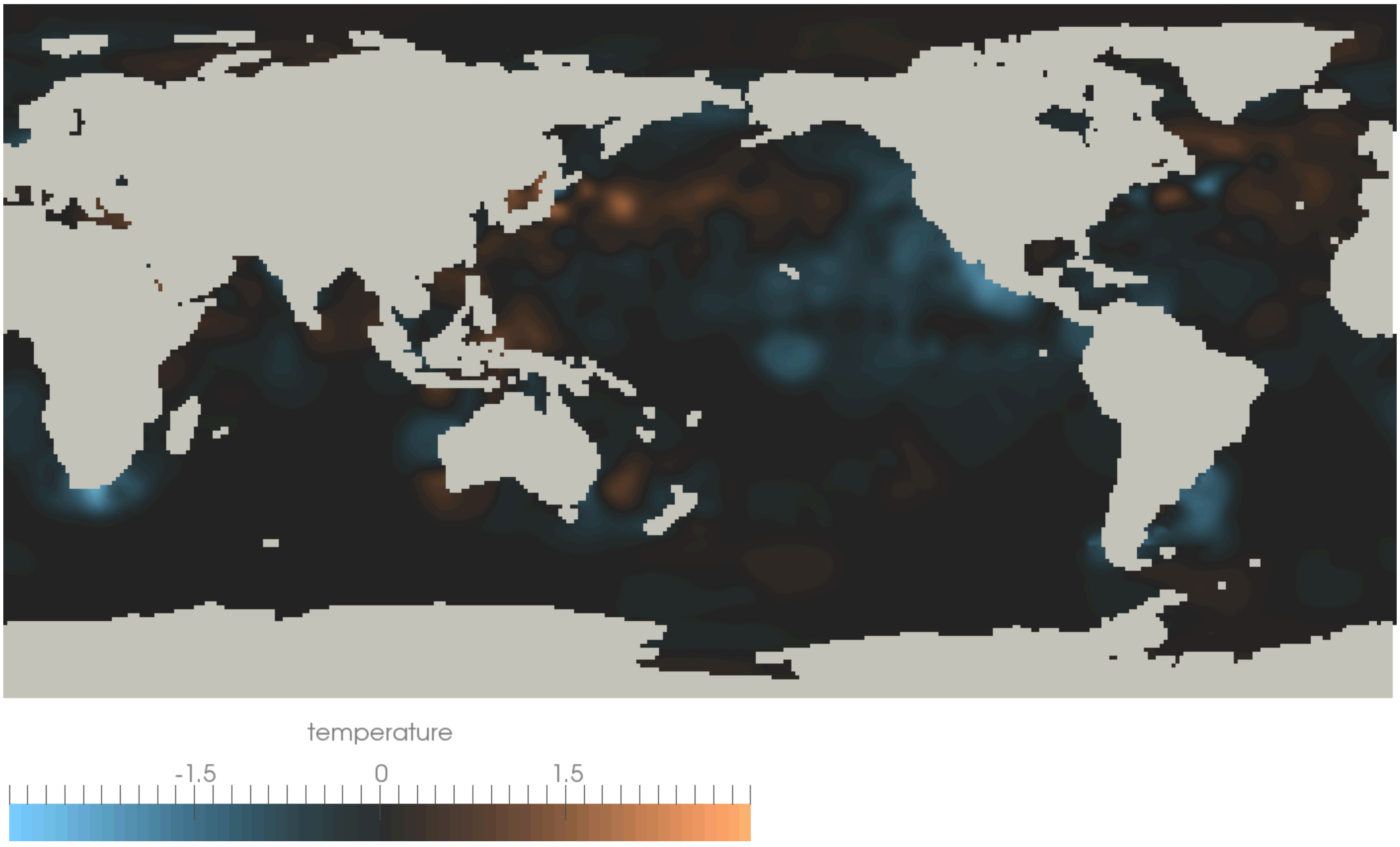
NOAA satellite data on the annual average ocean temperature at a 100m depth



Squeeze the middle dramatically, to really emphasize the tails

Color Map Choice Emphasizes Different Things

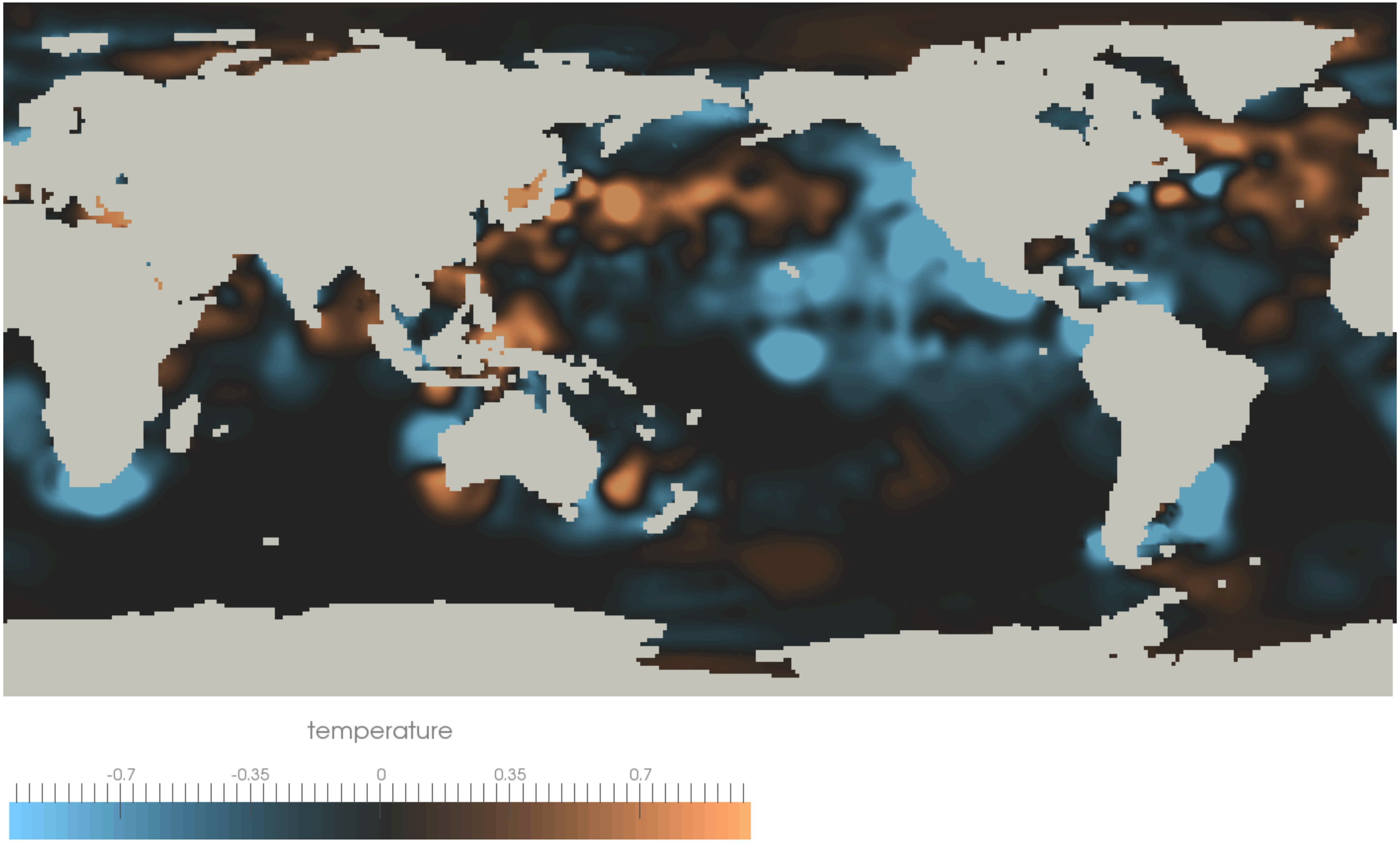
NOAA satellite data on the annual average ocean temperature at a 100m depth



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



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

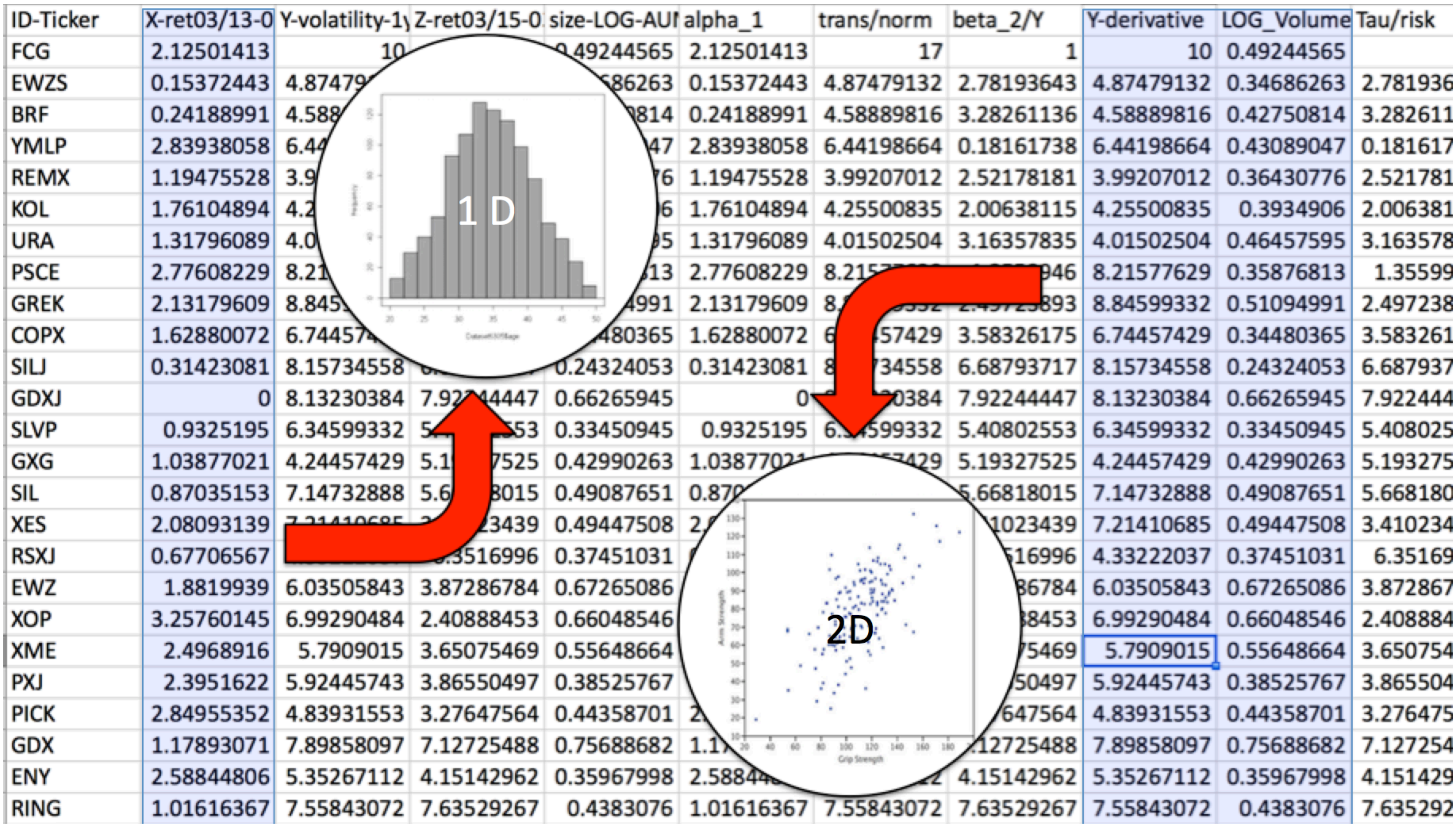
Strive for color harmony

Be aware of cultural conventions

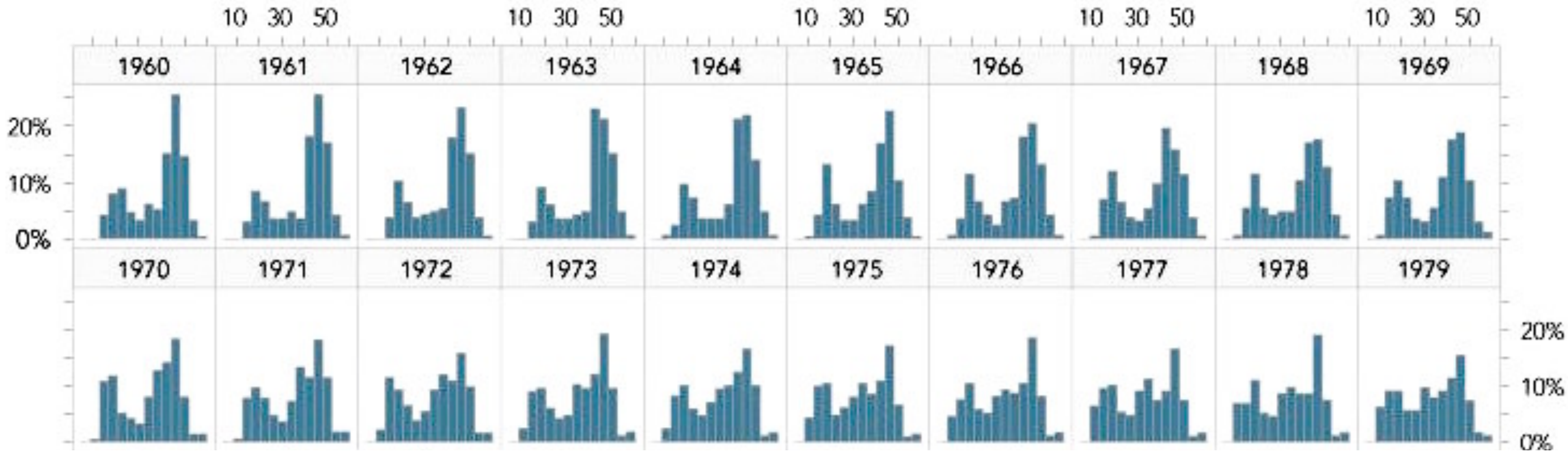
Beware bad interactions

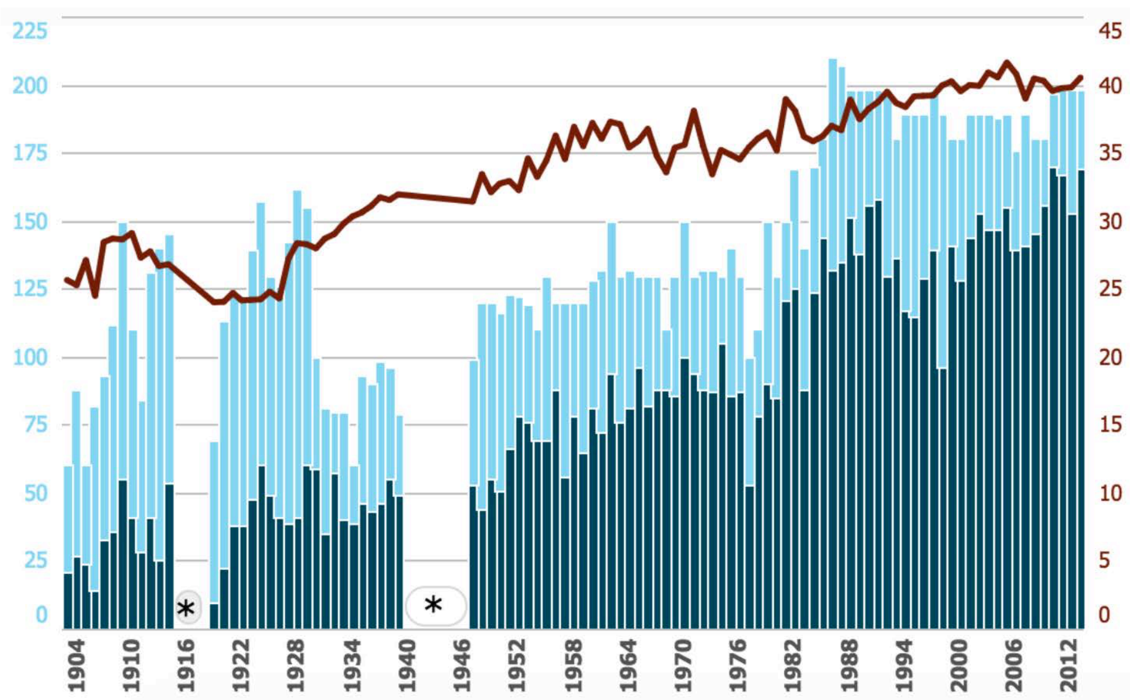
Get it right in black and white

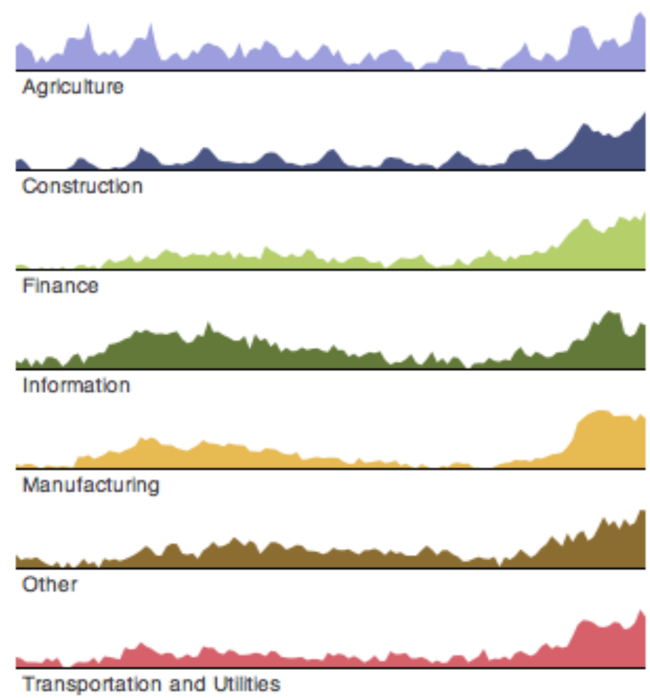
Traditional data visualization fails to reveal the complex patterns - the hidden knowledge - that may be present in the data



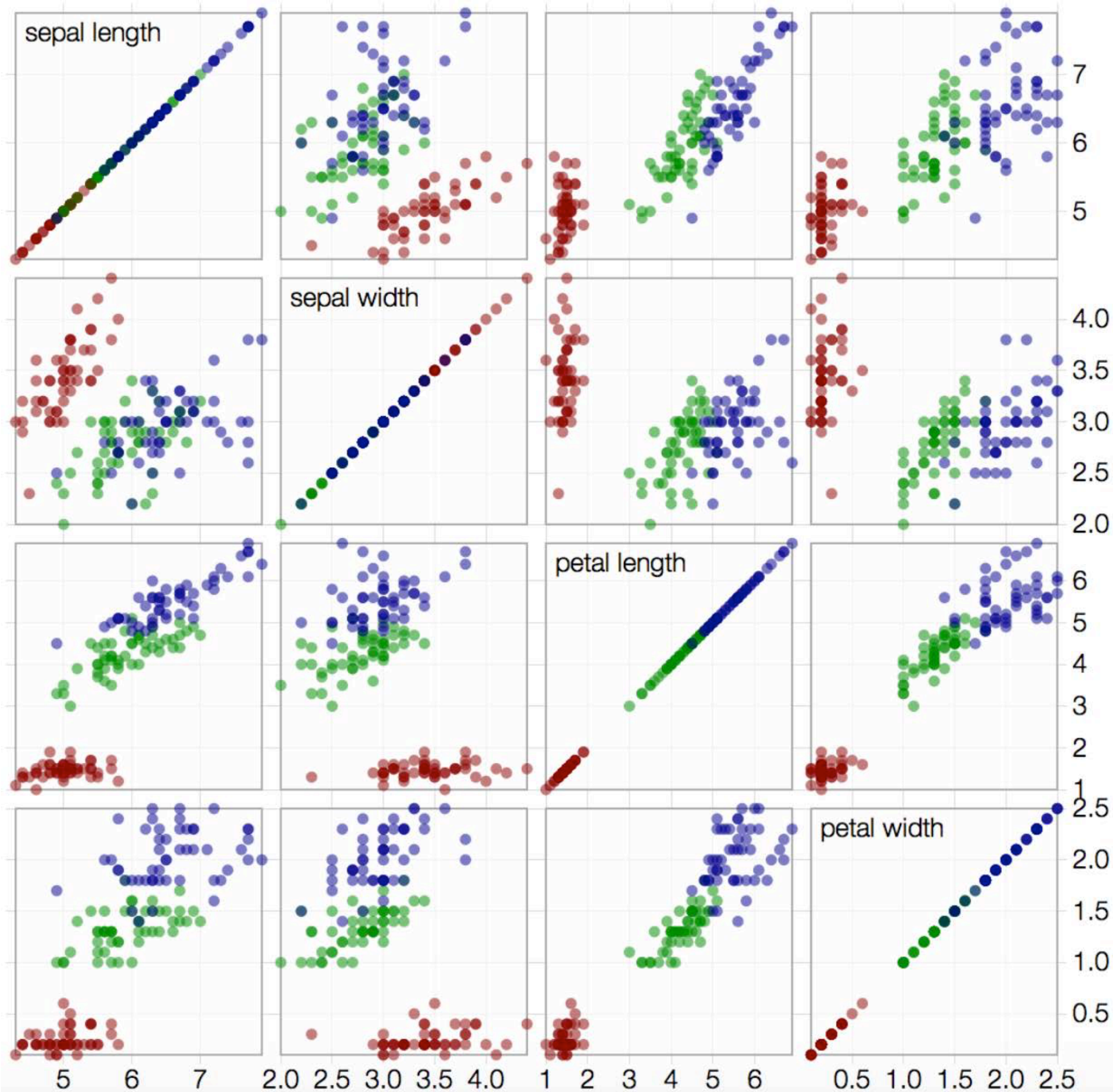
Multiple 1-D ≠ Multi-D







Multiple 2-D ≠ Multi-D

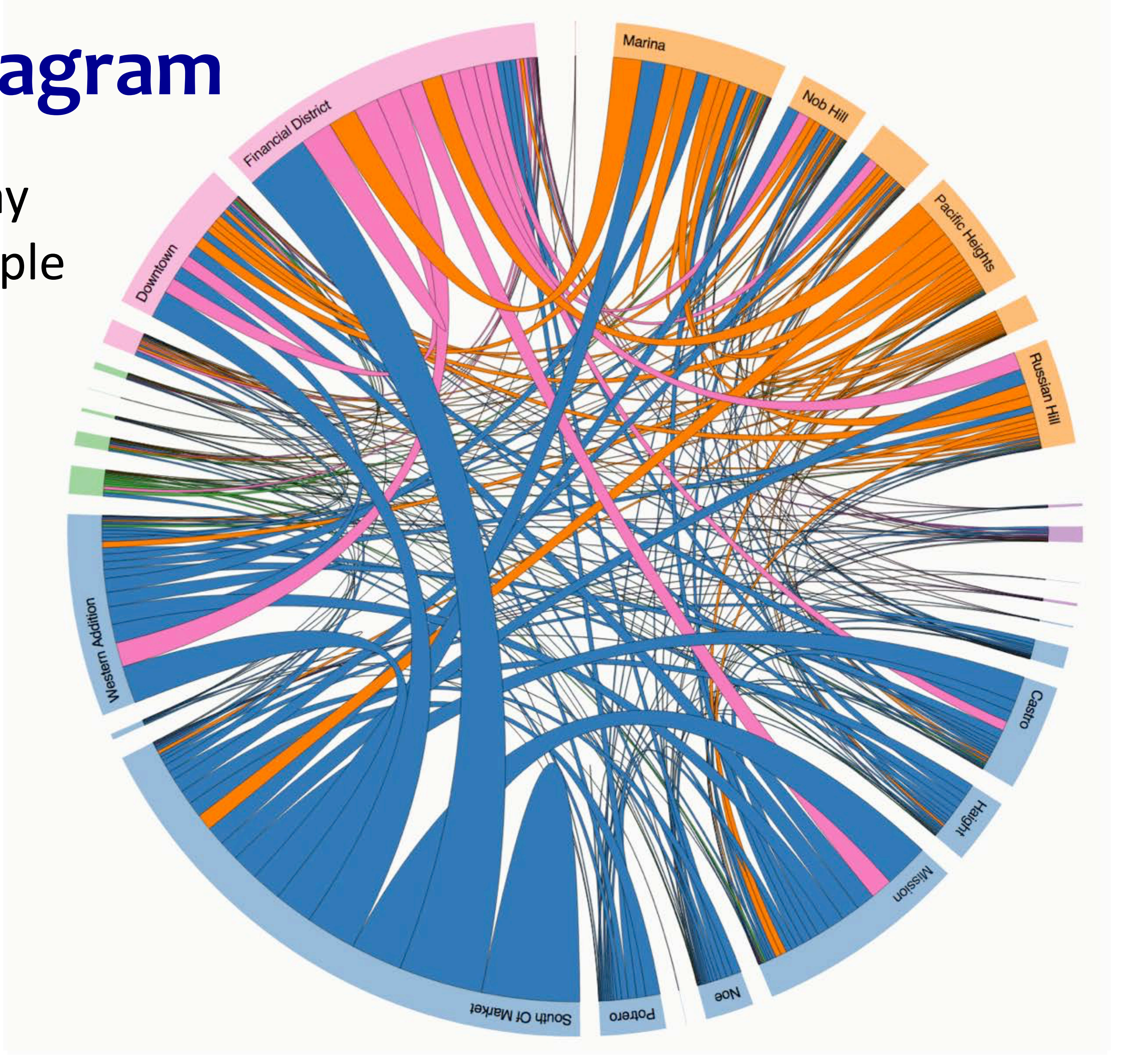


A grid plot of pairwise XY projections

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

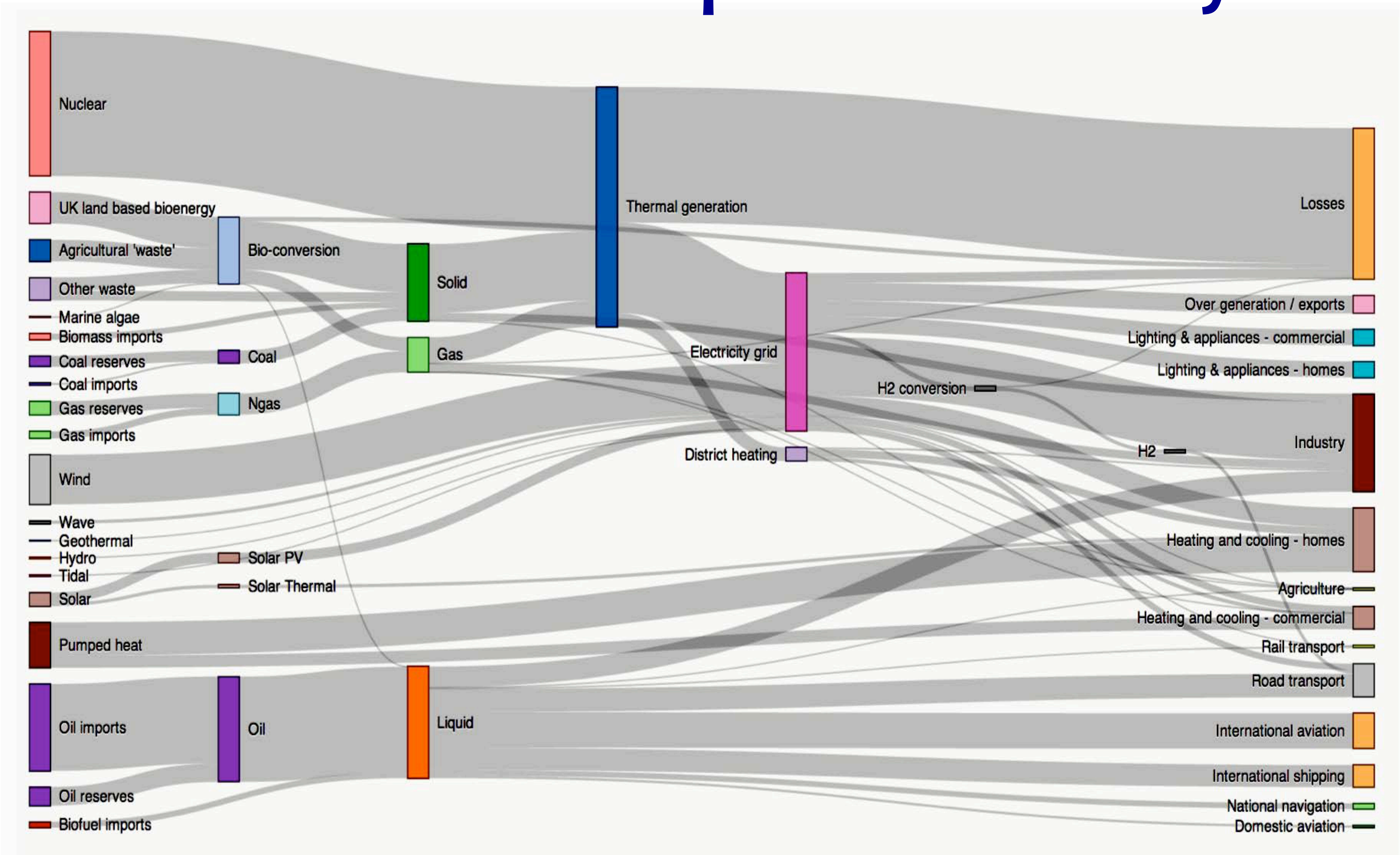
Chord Diagram

A compact way to show multiple connectivities

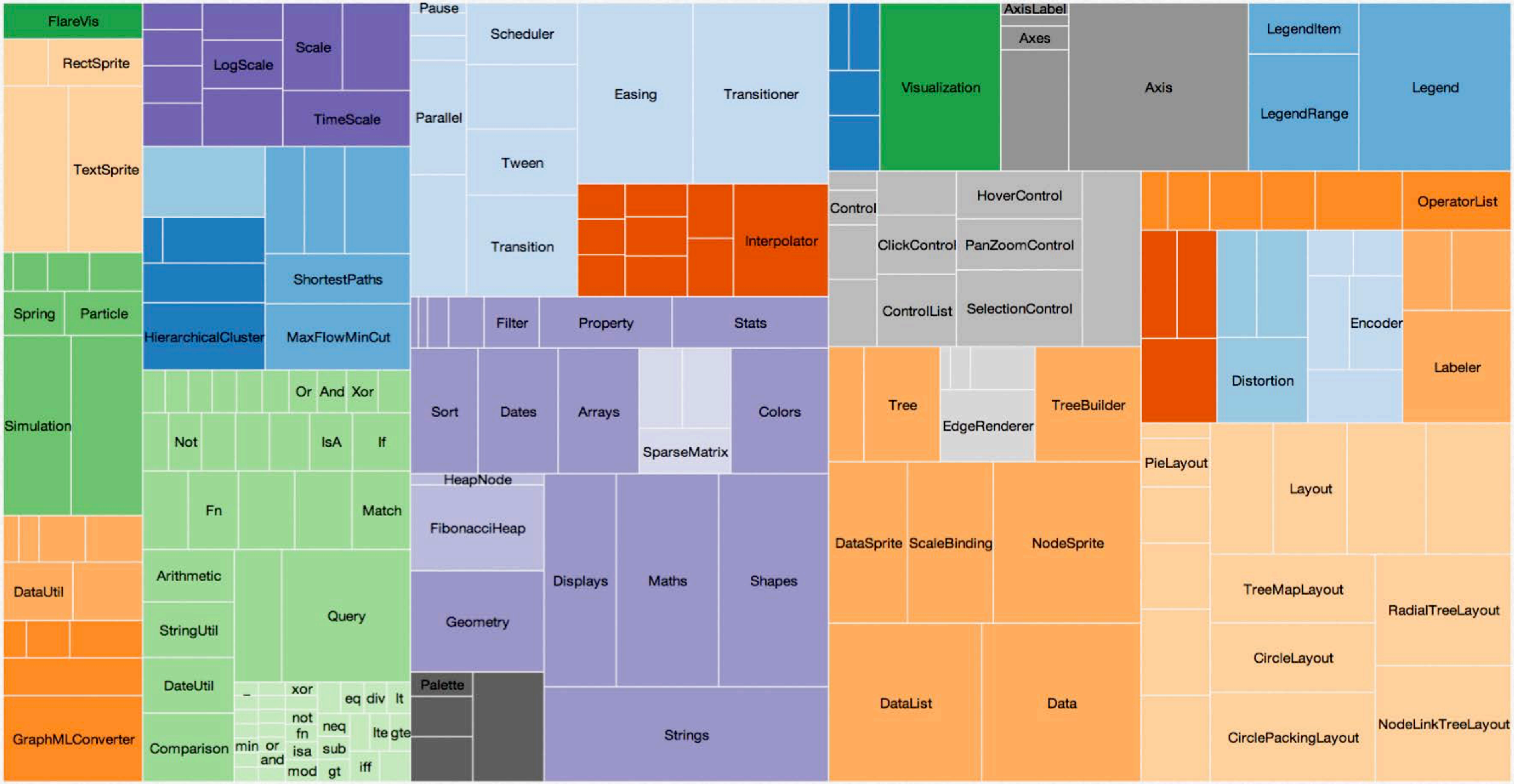


Chord diagram: Uber rides by San Francisco neighborhood

Sankey Diagram: another way to visualize a multiple connectivity



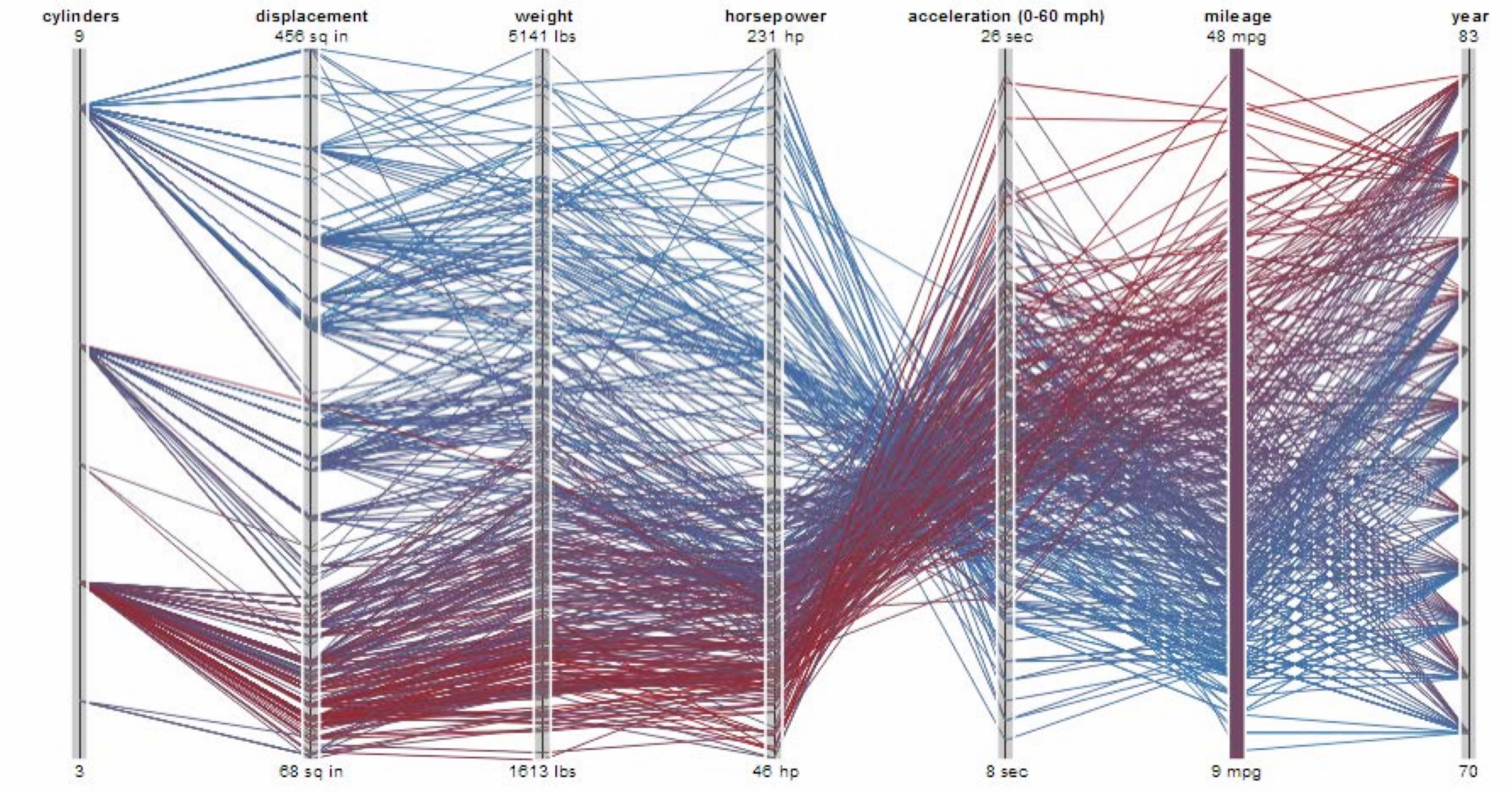
Hierarchical Visualization: Treemap



Area = 1 quantitative dimension

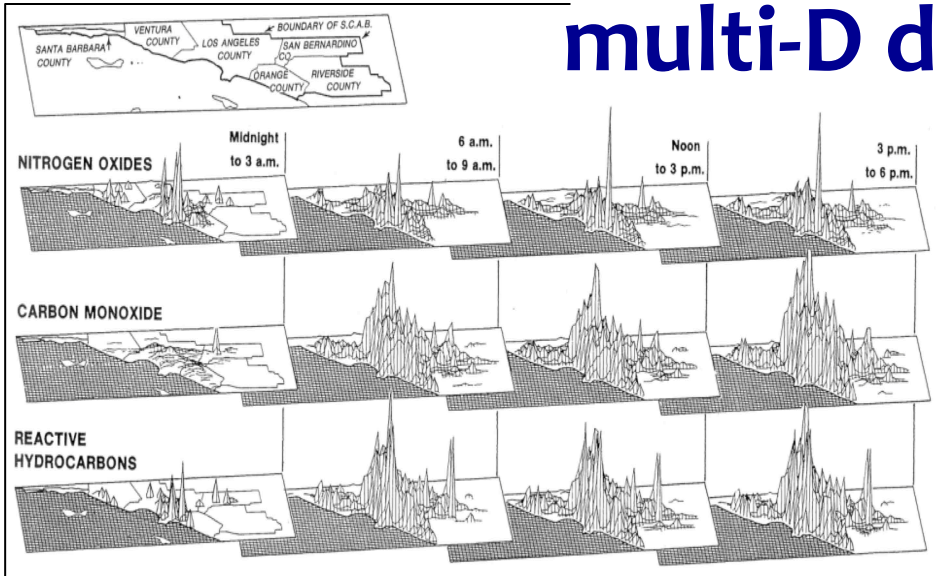
Color = 1 nominal (or ordinal?) dimension

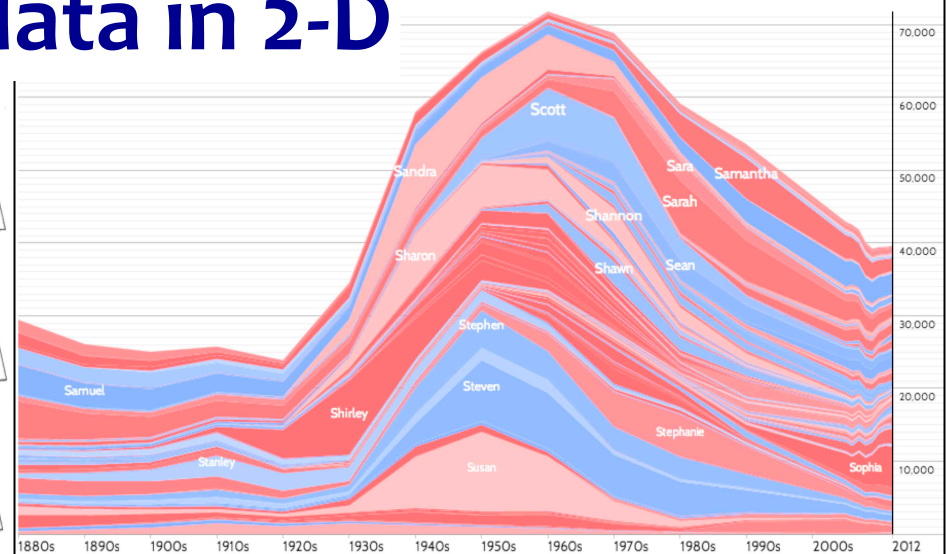
Parallel Coordinates:
visualize clustering in multiple dimensions

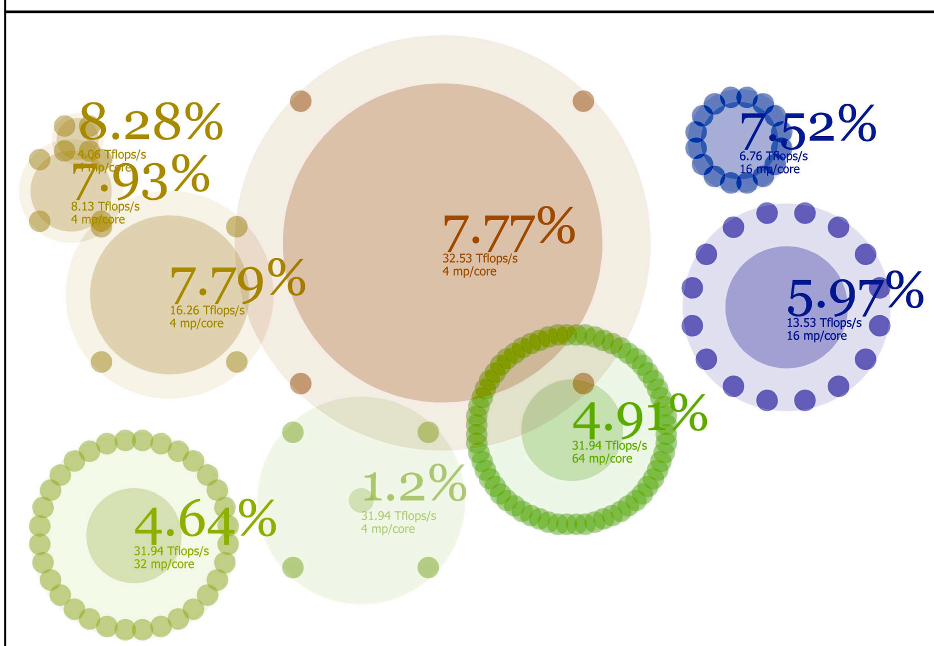


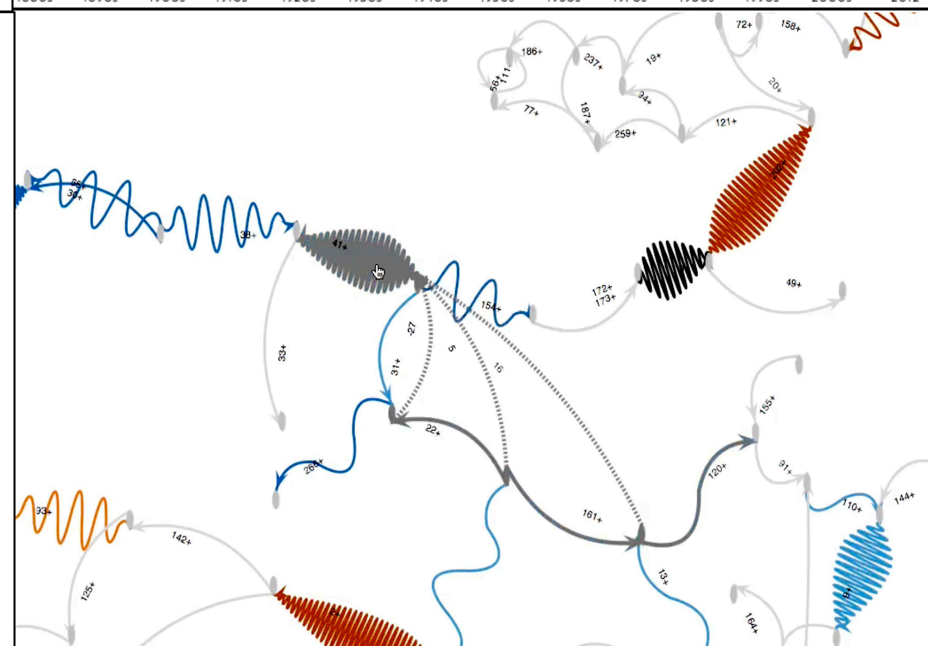
Note: the order of the axes is arbitrary!

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







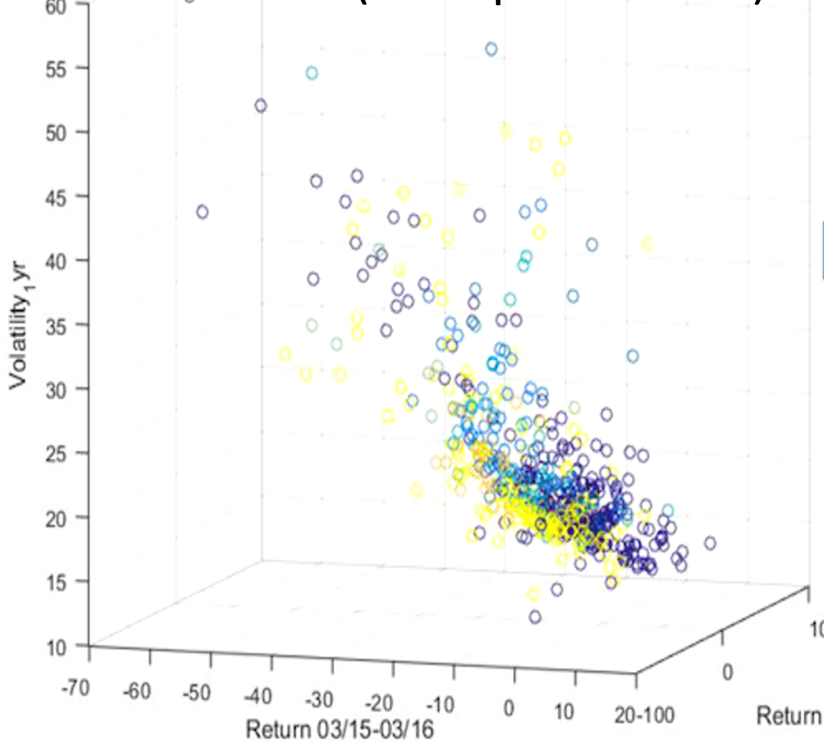


(examples from S. Davidoff and S. Lombeyda)

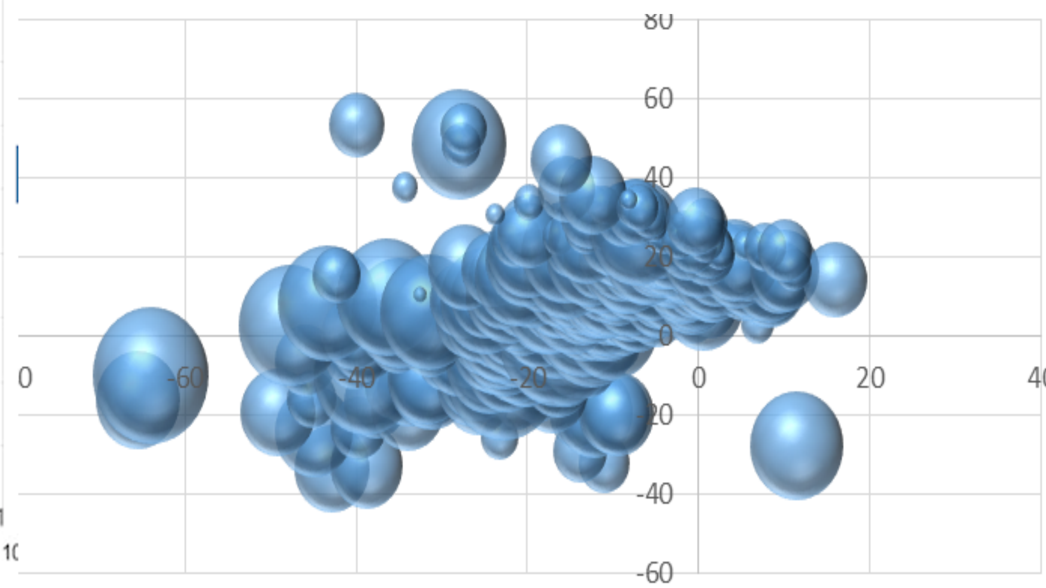
3D Data Visualization is Not Simple

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

(example: Matlab)

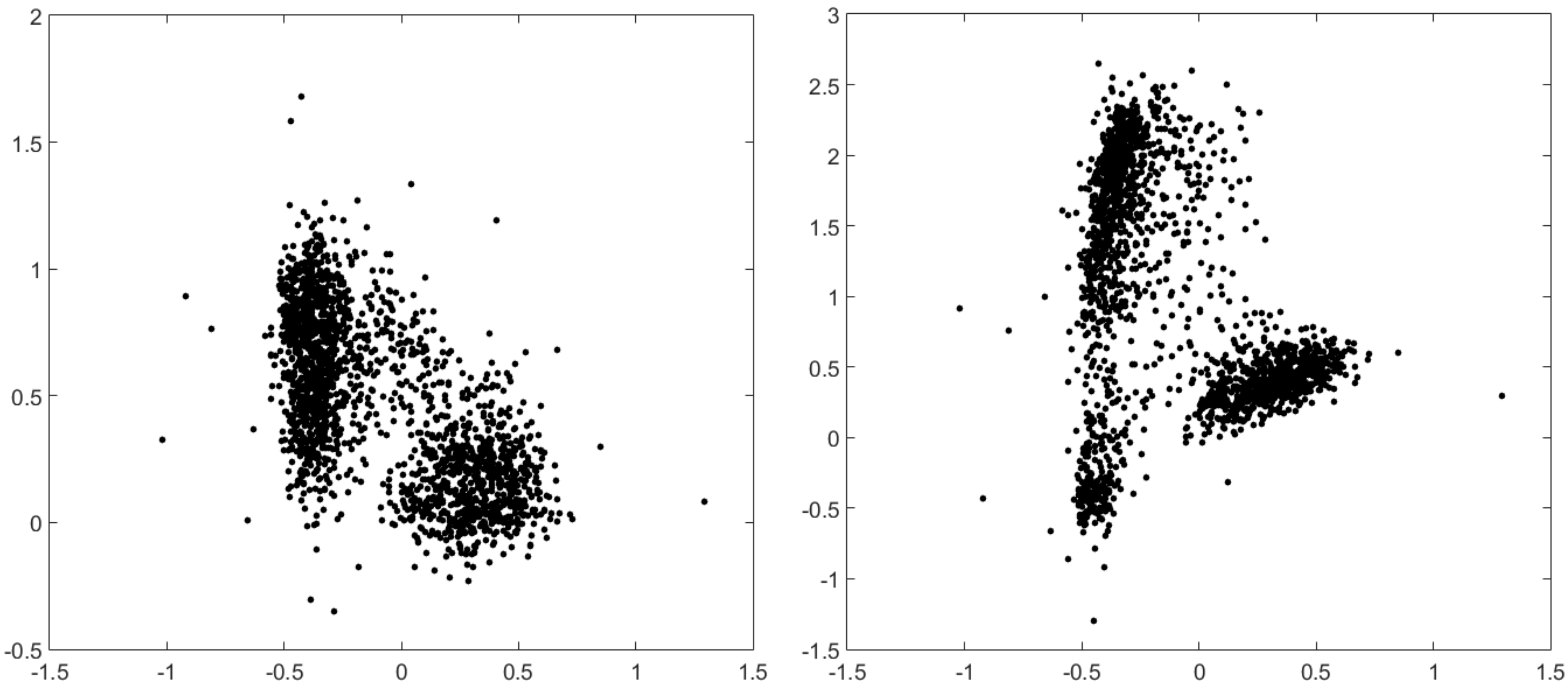


(example: Excel)



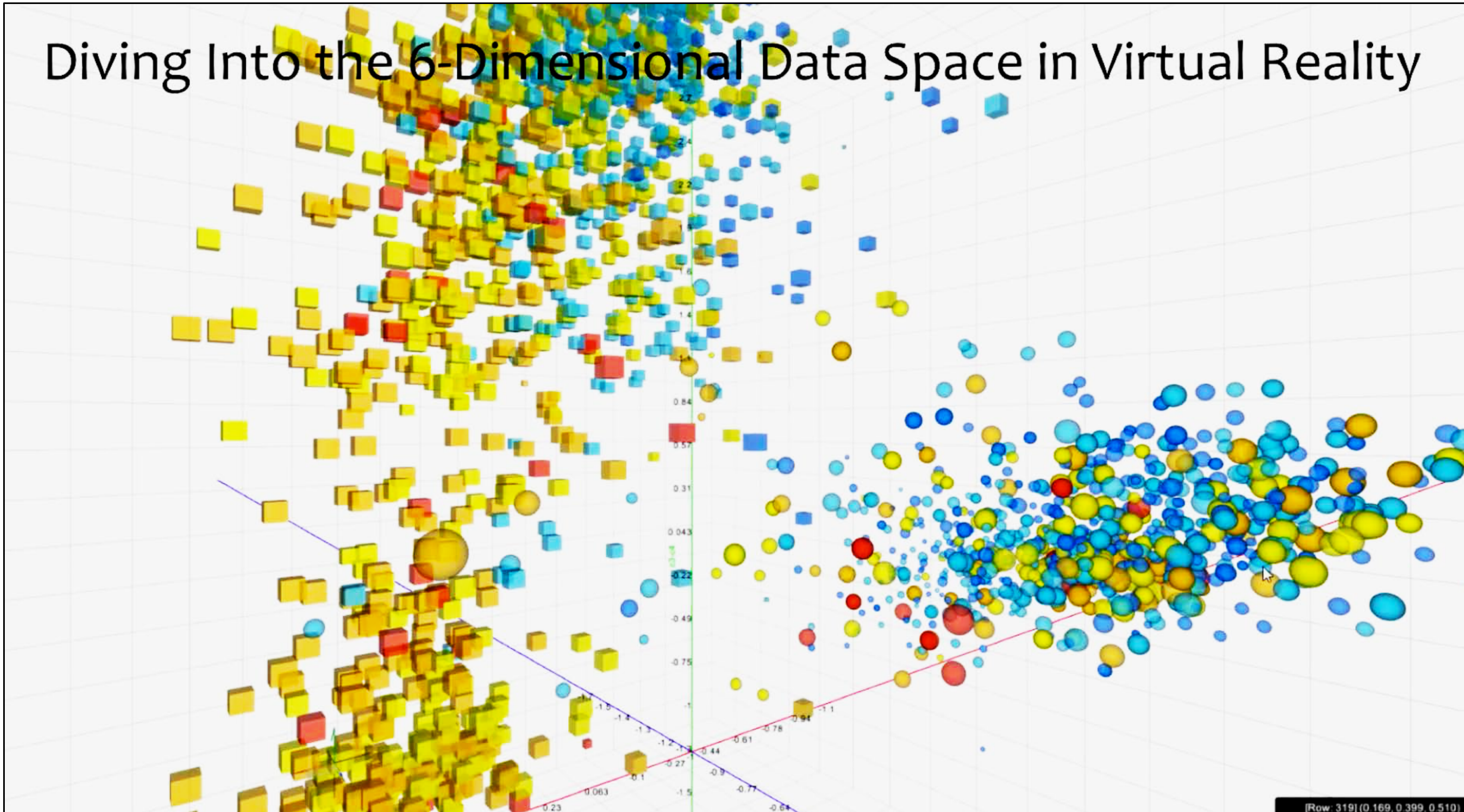
Traditional Data Visualization

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

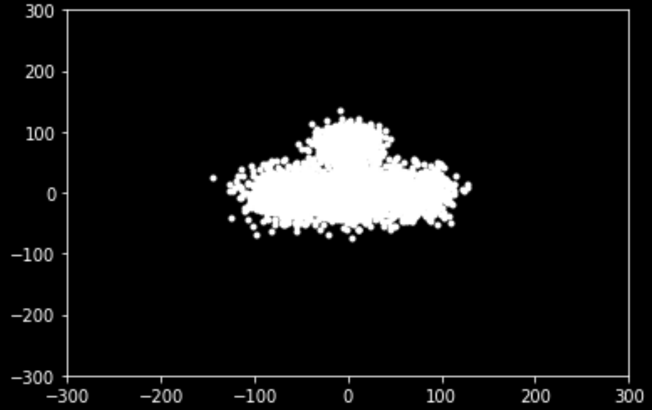


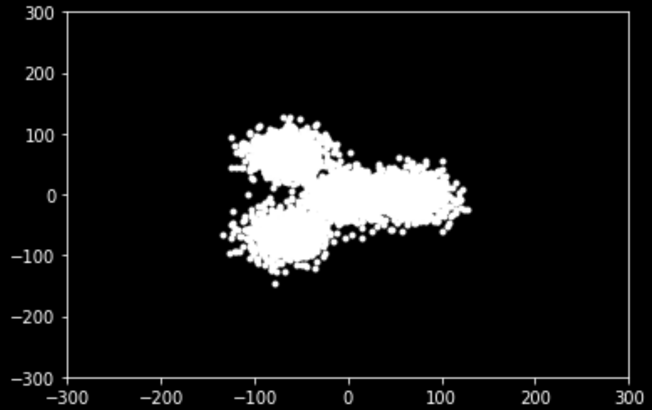
These are 2 out of the 15 possible 2-D plots, but even then relationships involving >2 variables are lost

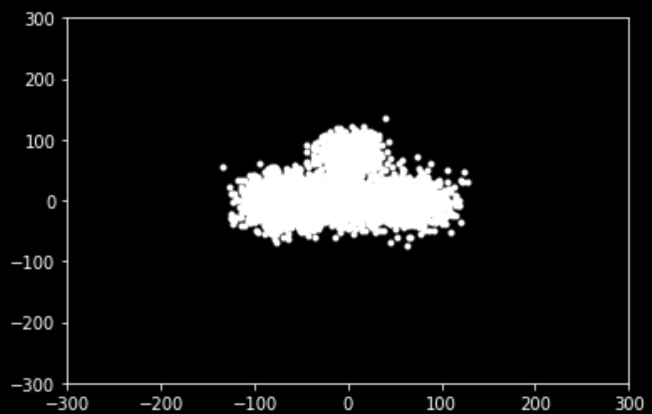
Diving Into the 6-Dimensional Data Space in Virtual Reality



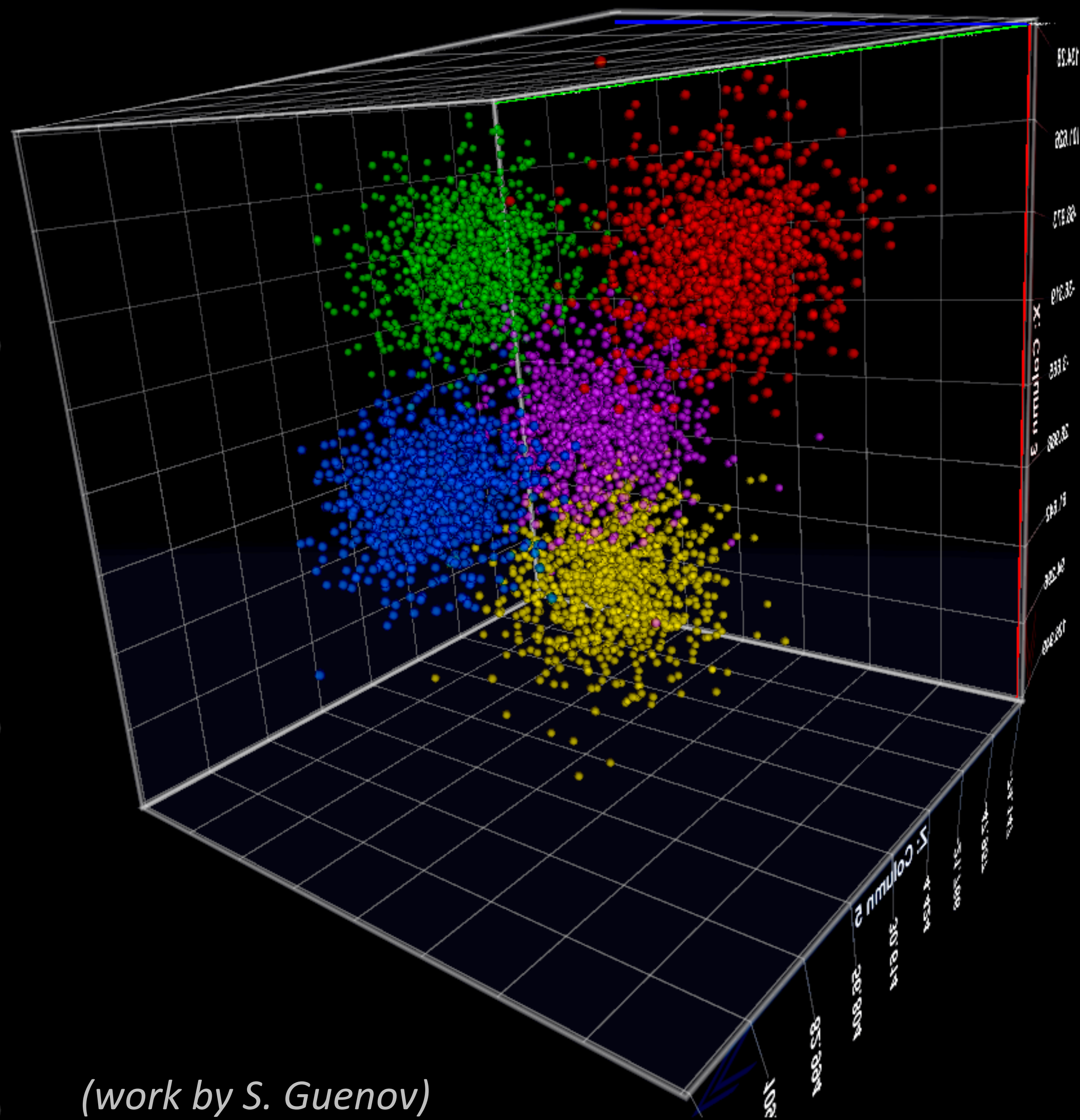
XYZ Projections







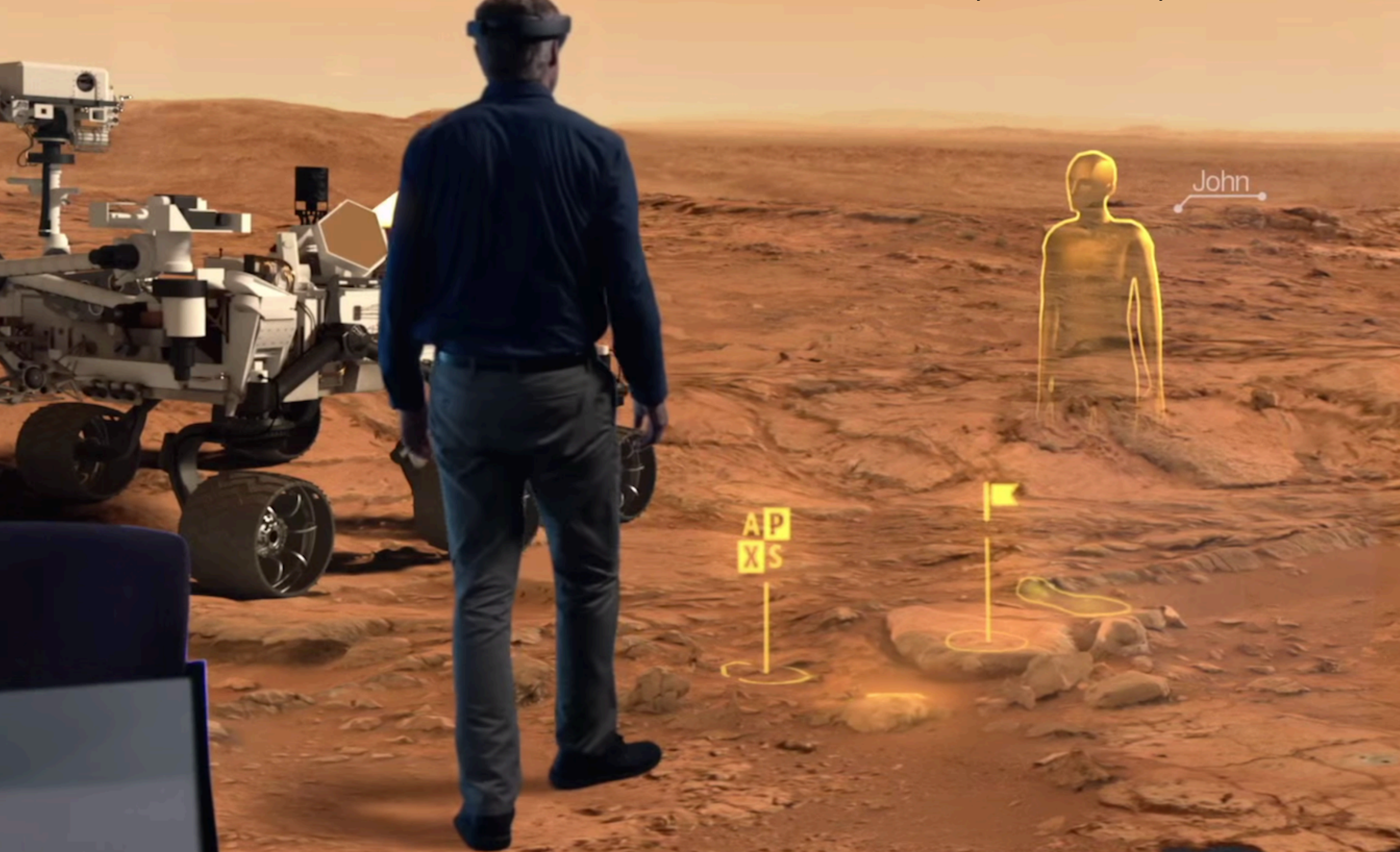
3D Visualization + Clustering Analysis



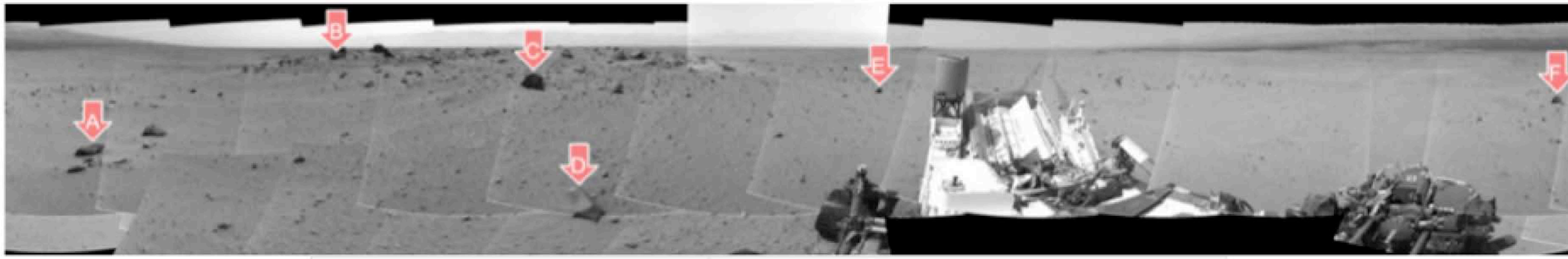
(work by S. Guenov)

Exploring the Virtual Mars at JPL

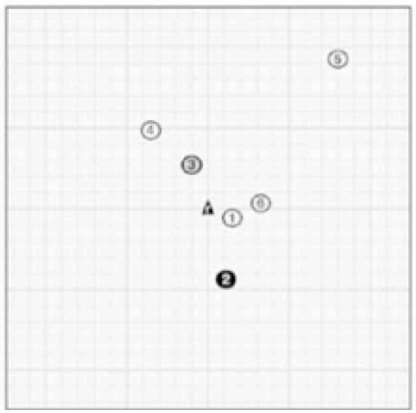
S. Davidoff, J. Norris, et al.

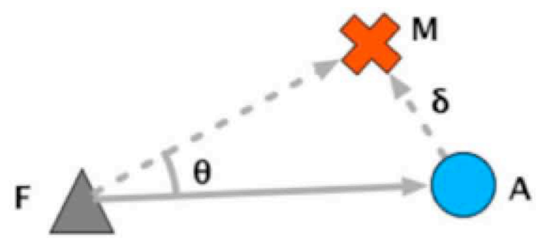


Navigating on Mars using VR

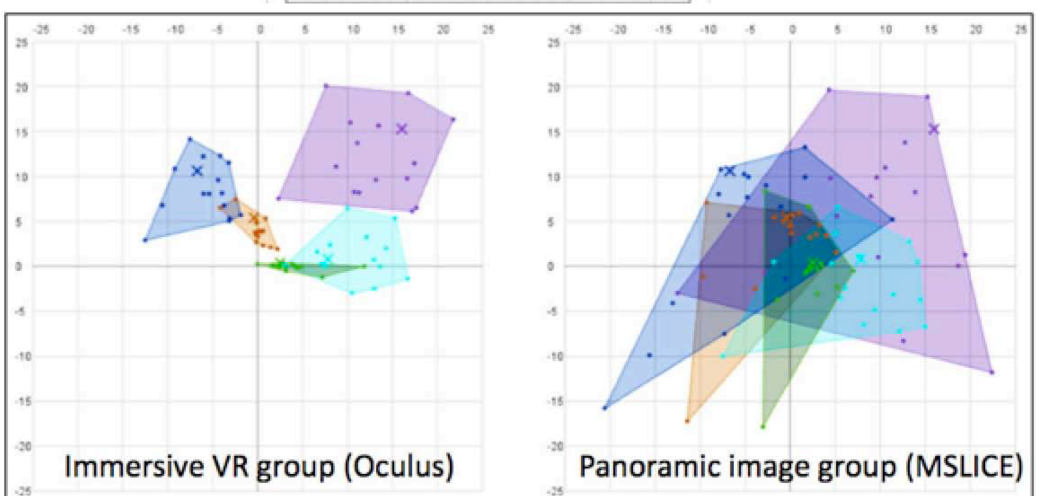


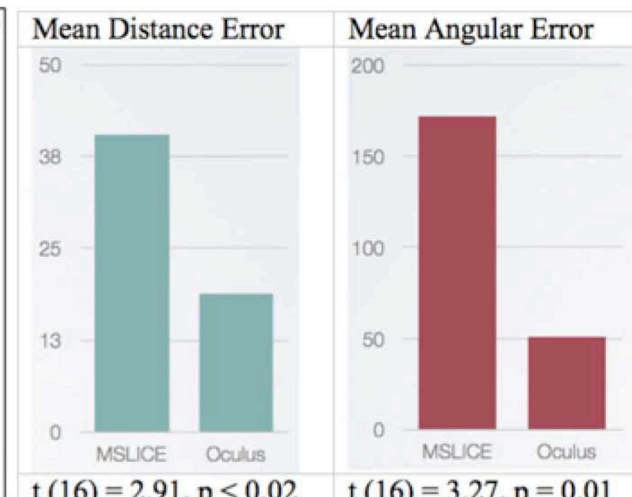
S. Davidoff,
J. Norris,
et al.





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





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



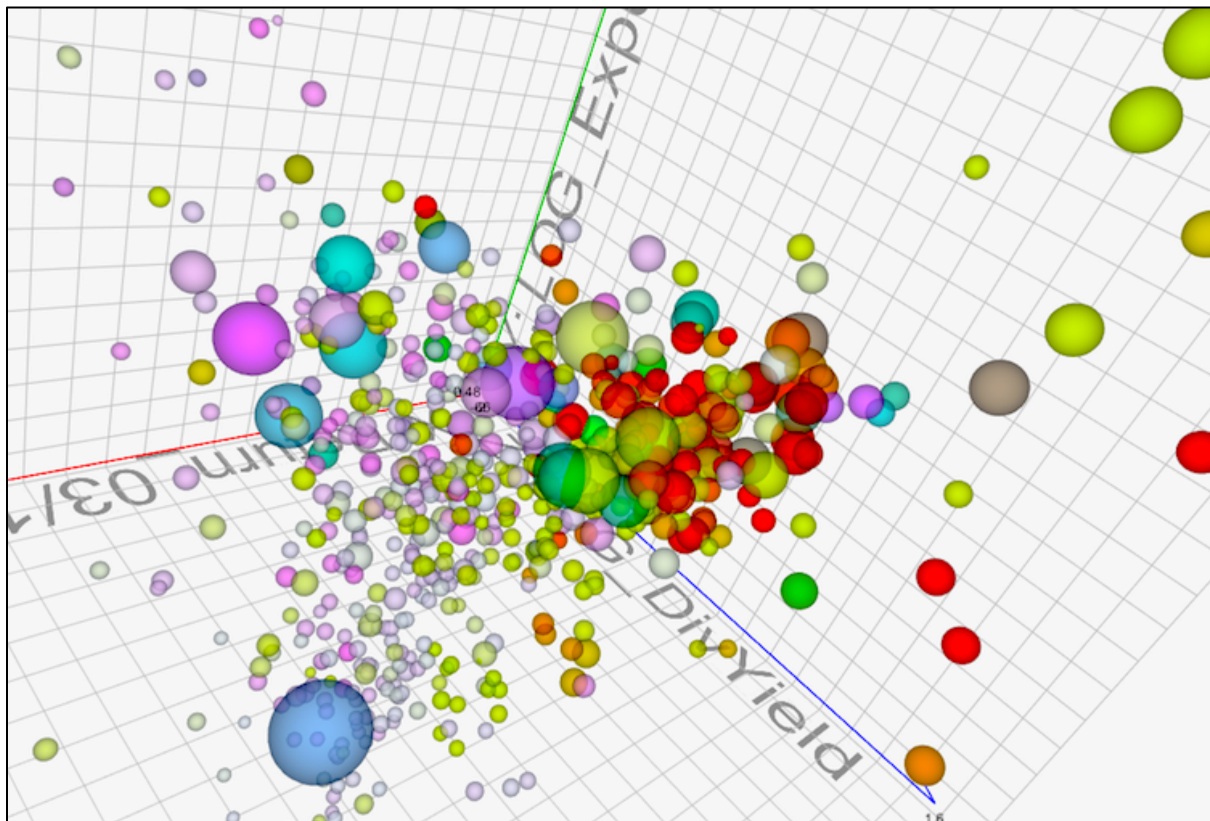

From Mainframes… … to VR/AR Headsets

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

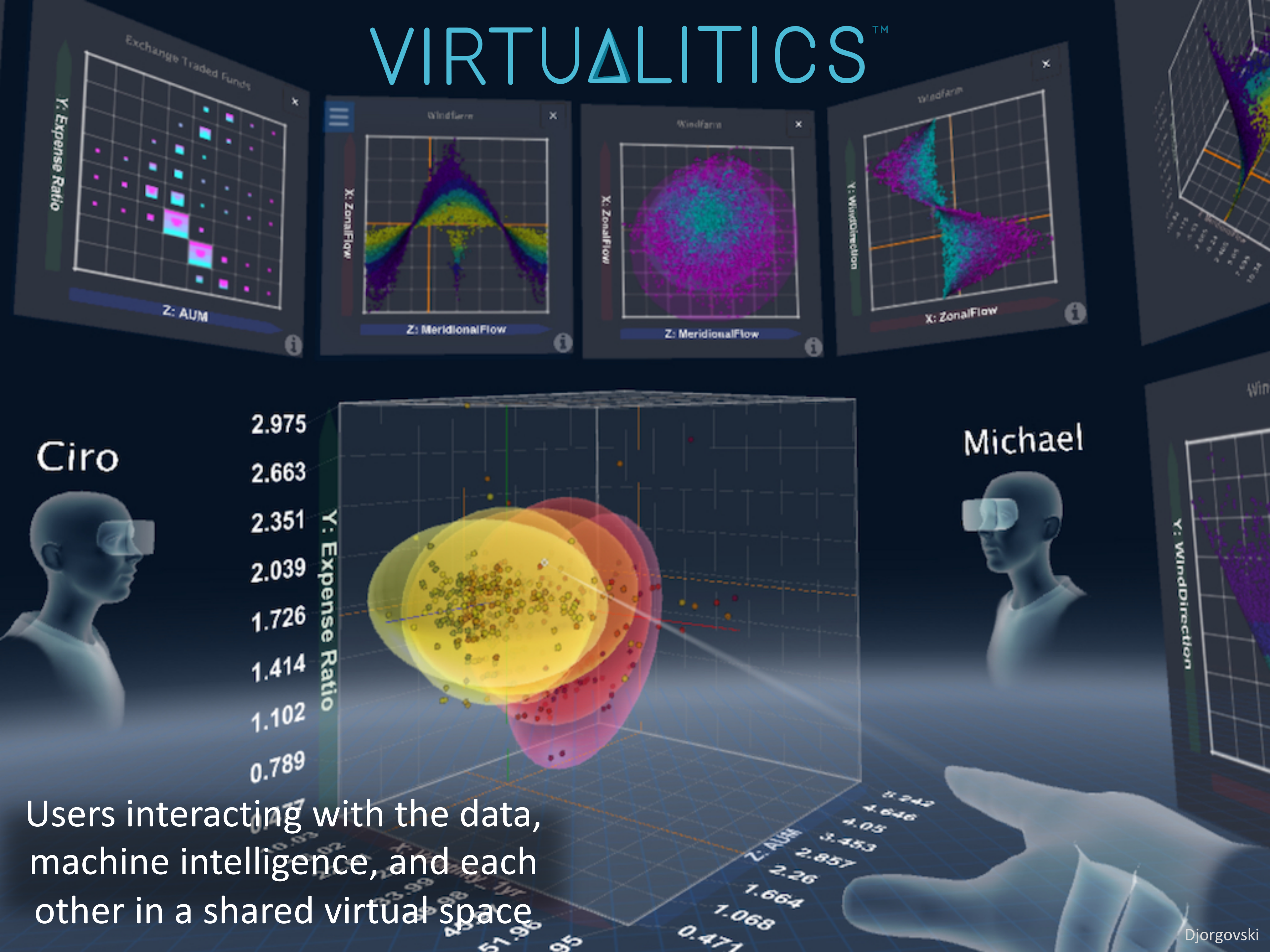
- Preliminary tests using both artificial and real data indicate that *dramatic improvements* are possible in time-to-insight, as compared to the traditional data analysis/visualization tools (e.g., Excel): from days/months to minutes/hours in some cases
- Some results are simply impossible to achieve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

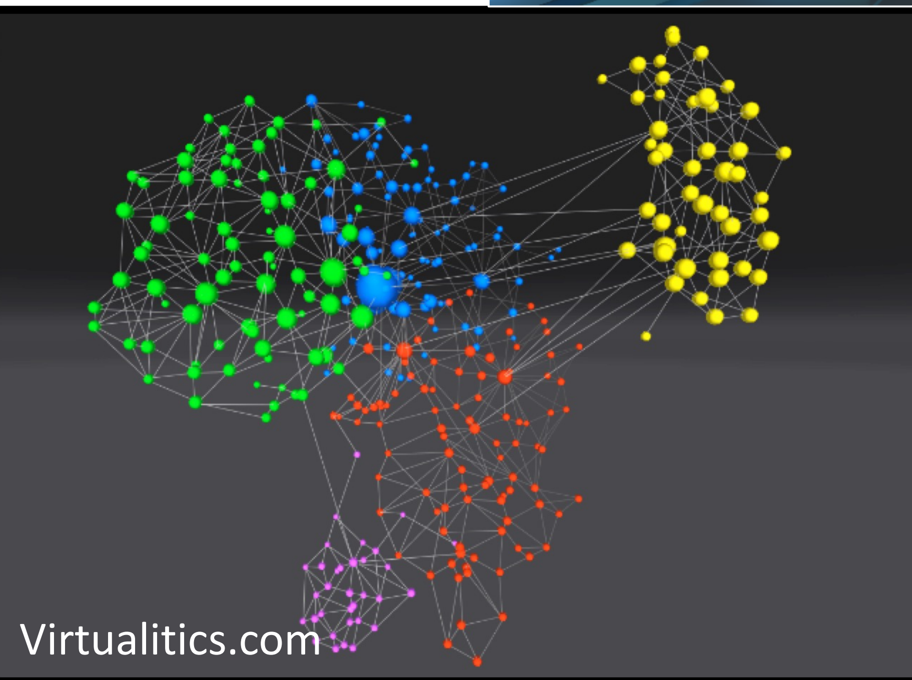
Videos: www.kiss.caltech.edu/symposia/space_science

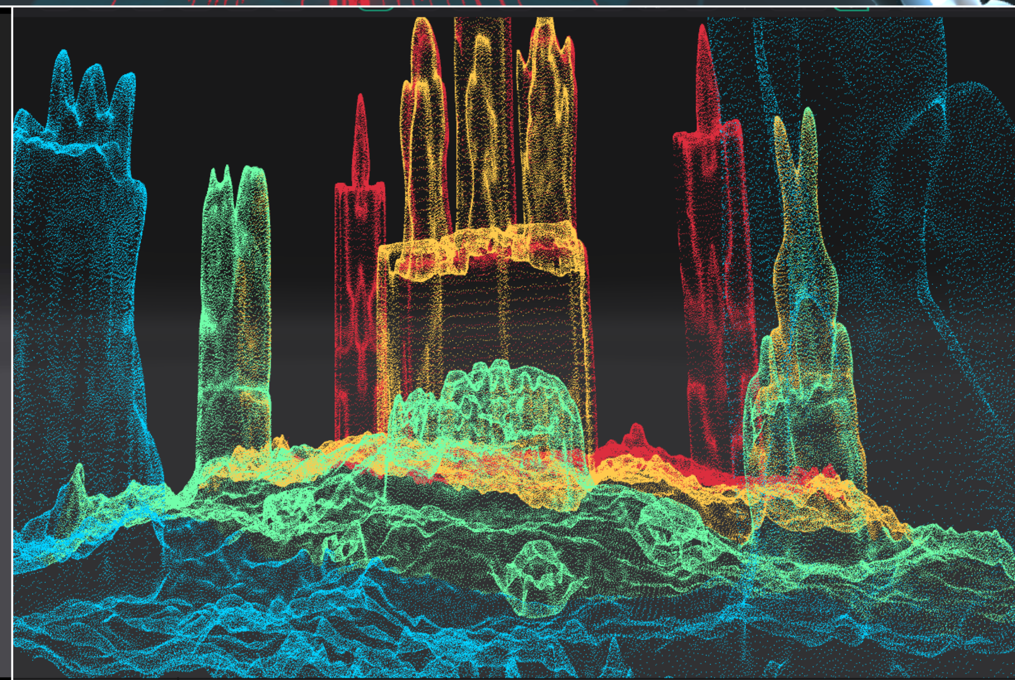


Users interacting with the data, machine intelligence, and each other in a shared virtual space

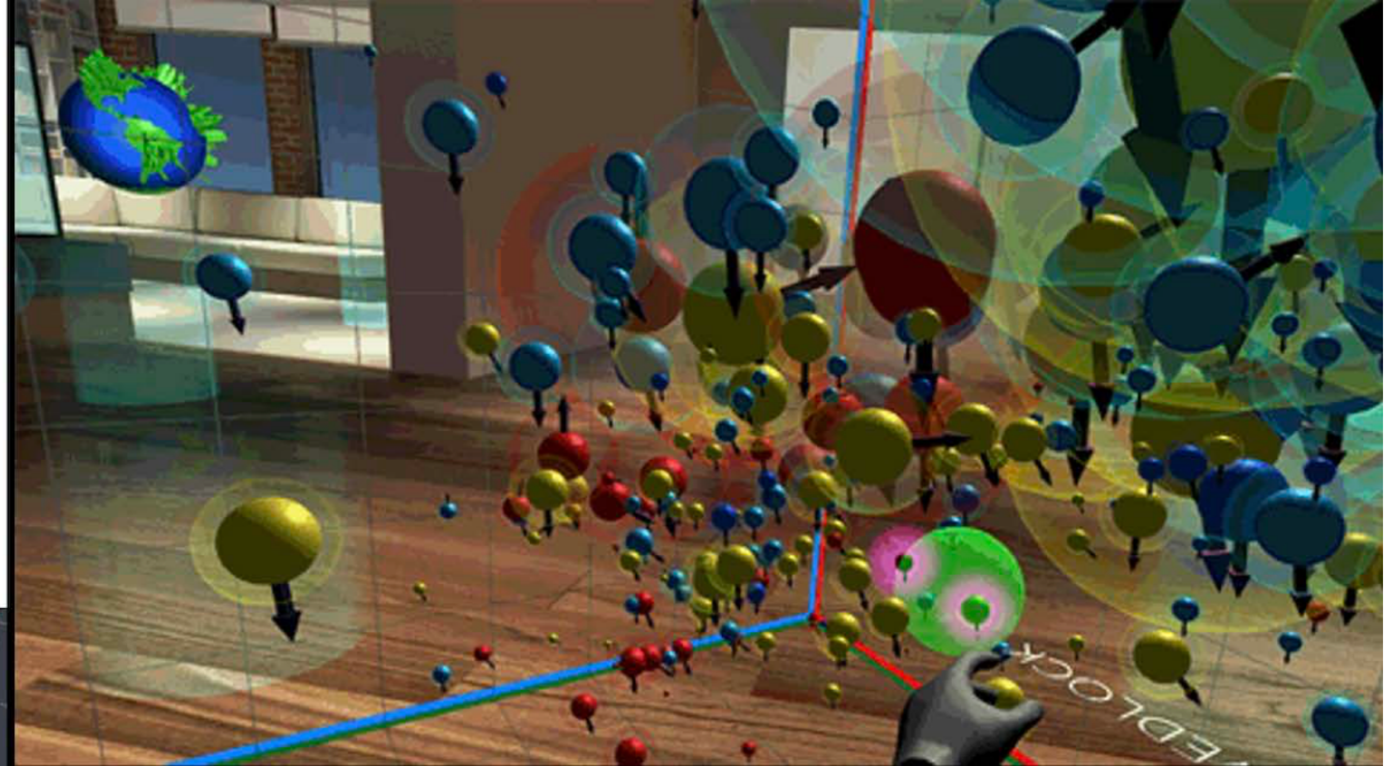
Djorgovski

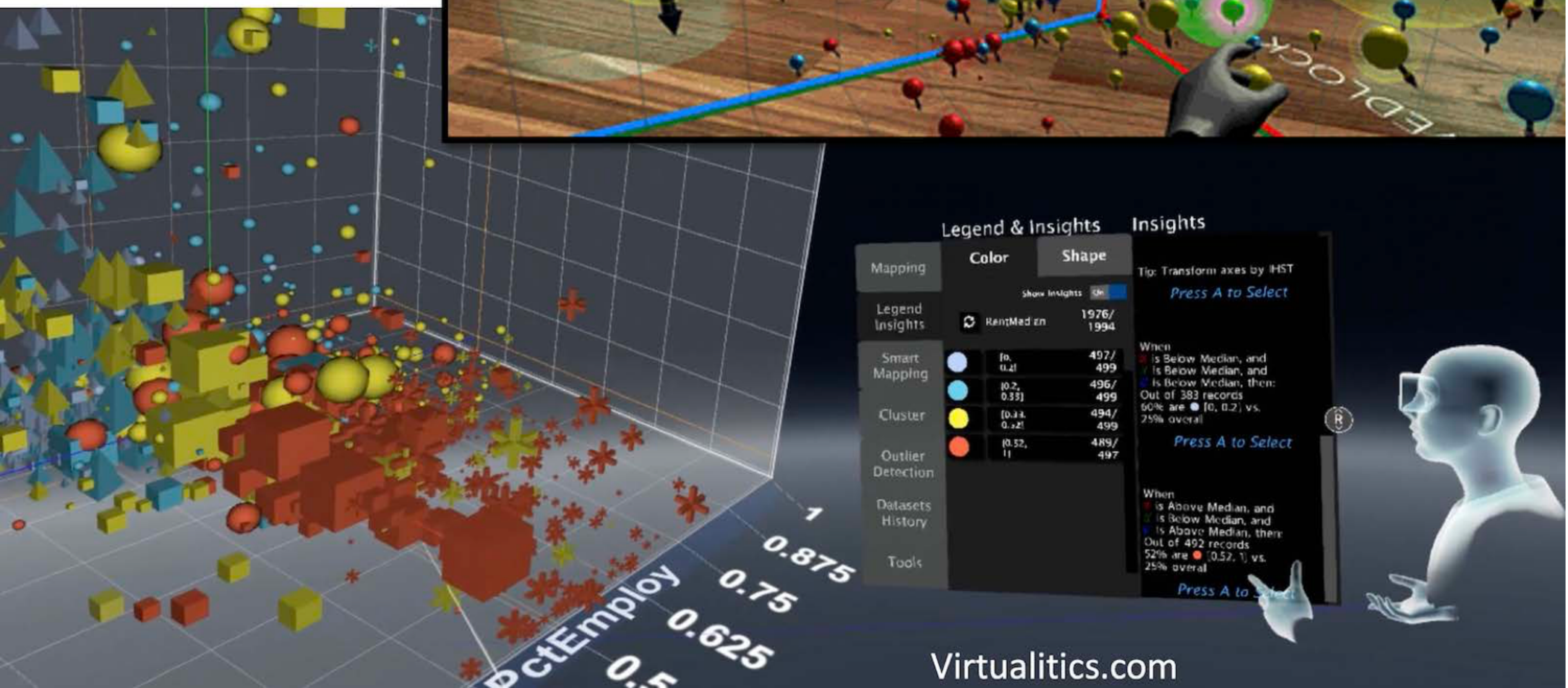
Different types of data visualized in VR





Interacting with data in VR

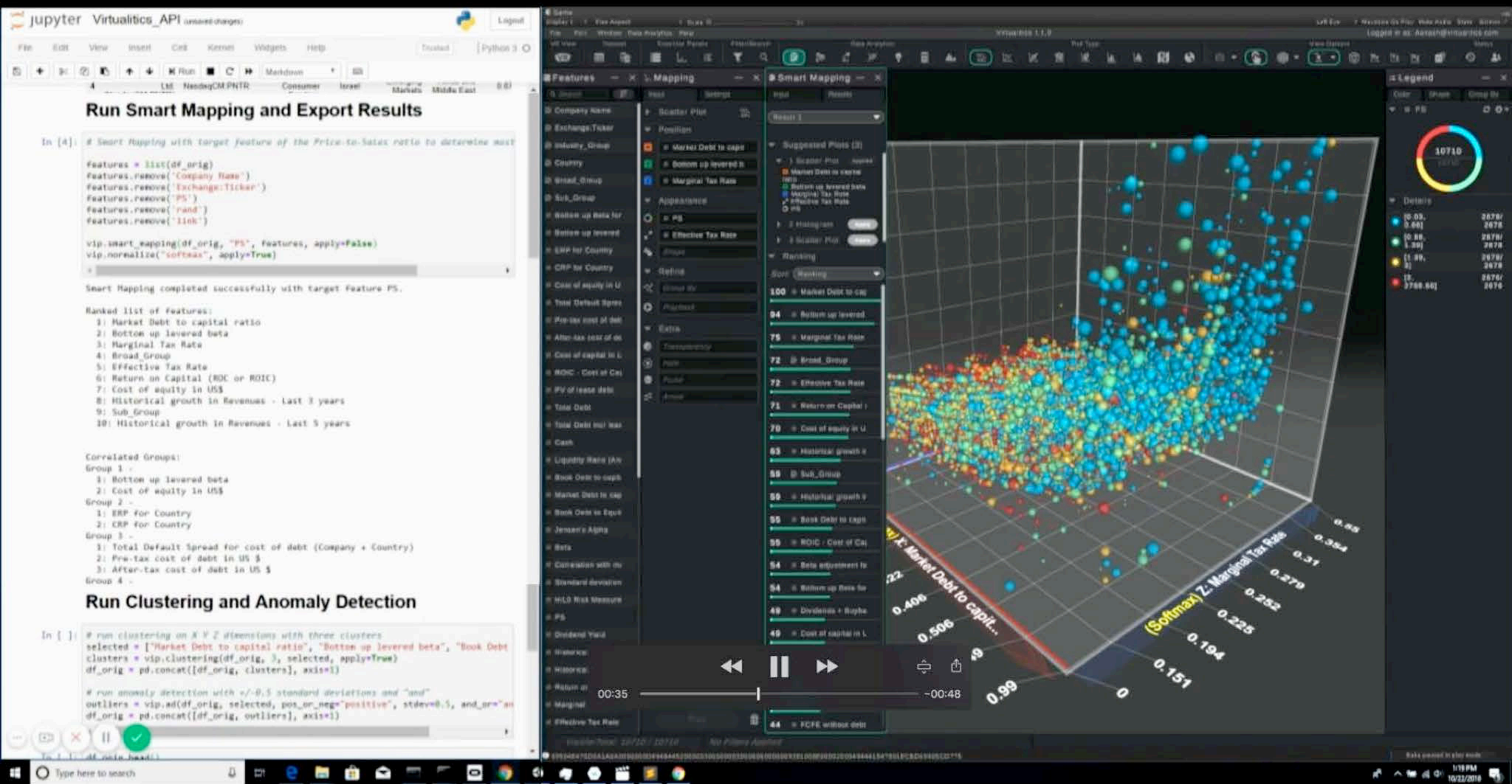




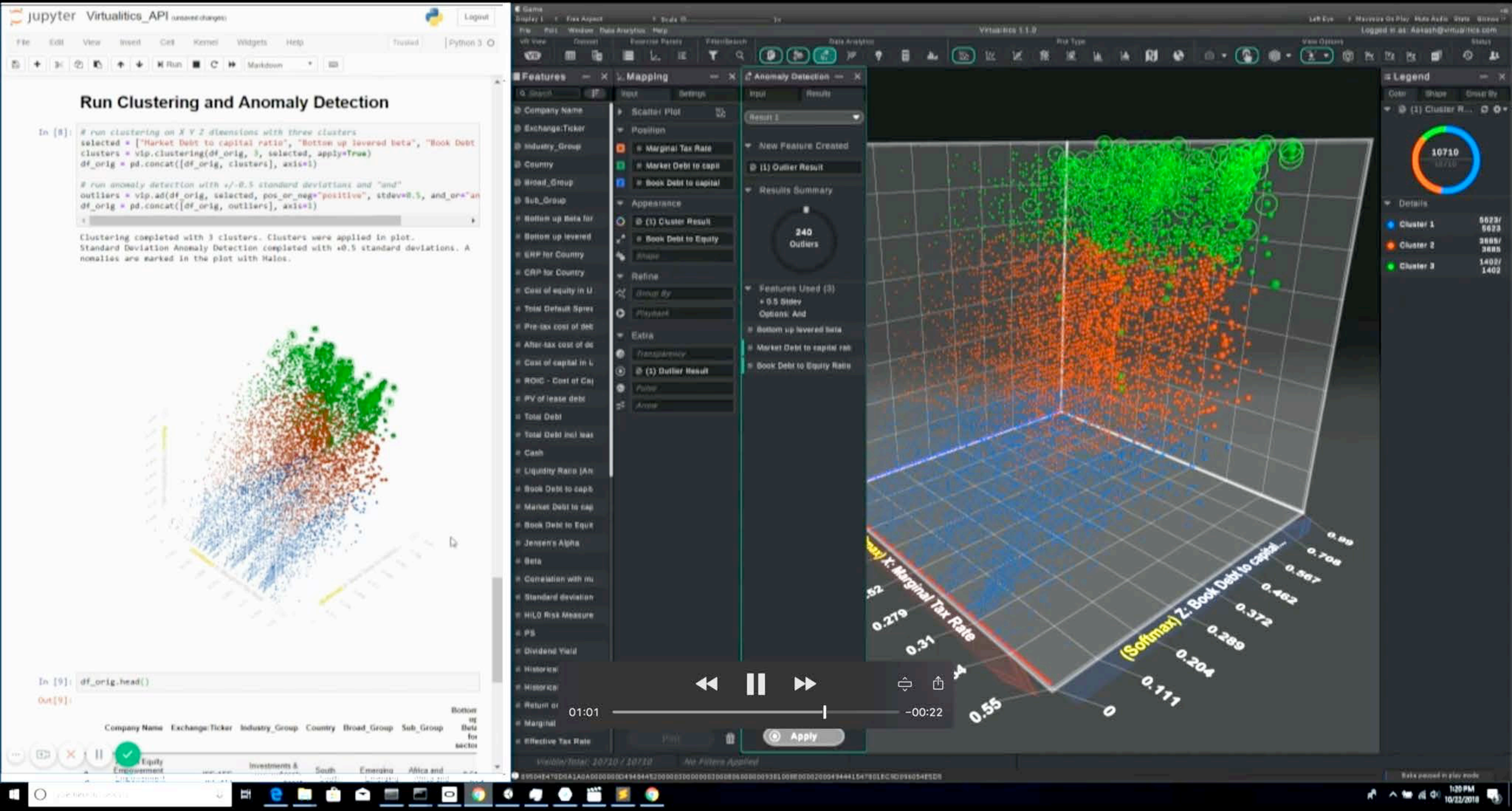

Horizontal Error
Y: Latitude
X: Longitude
Garrett-Dev M
X: Latitude
Paused
Latitude
Longitude
Depth
Depth
X: Latitude
Z: Depth
Collaborative Data
Visualization in VR

VR+ML platform API smooth interaction with a Python notebook

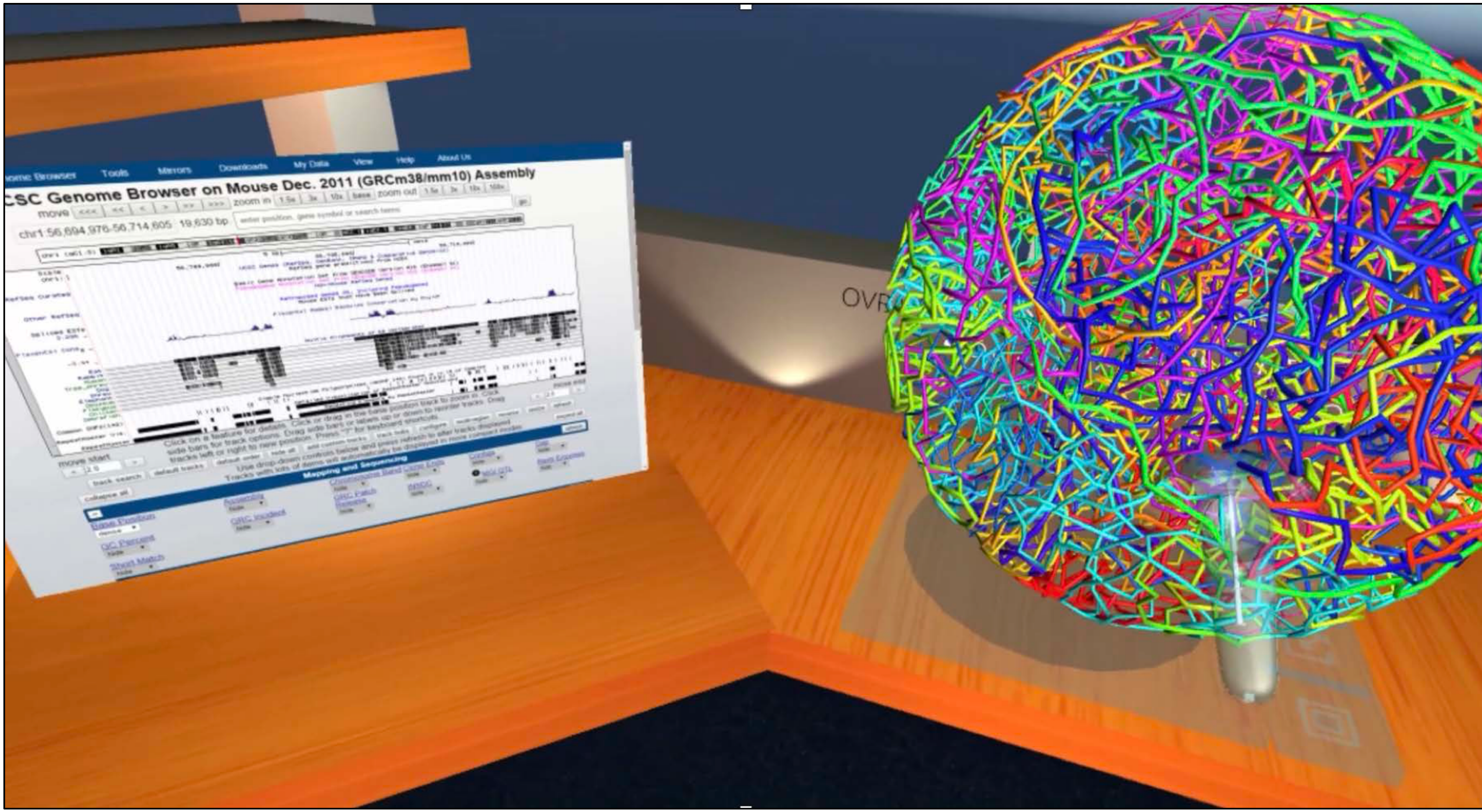
(and other popular data analytics platforms)



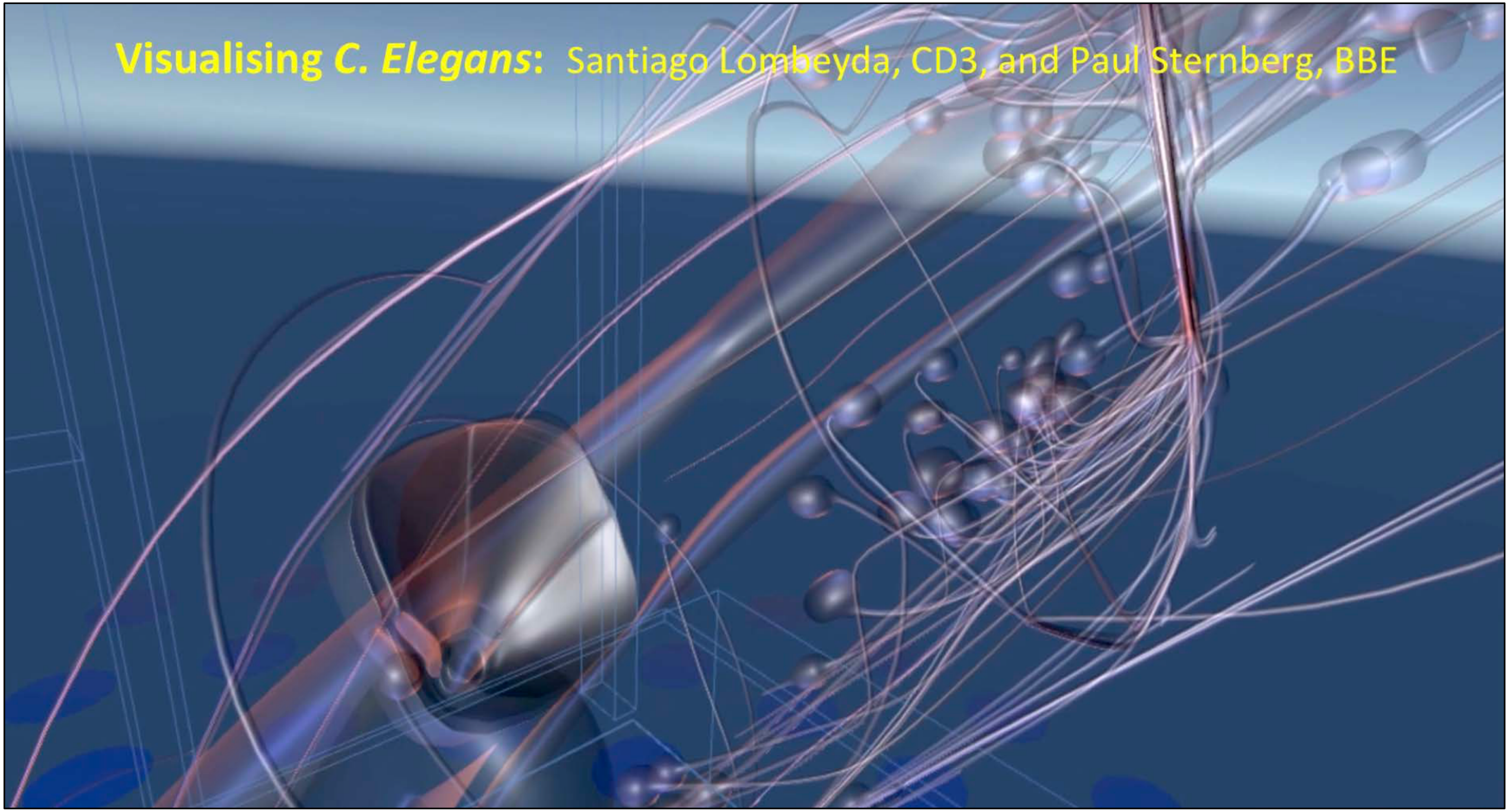
Export the results back to a Python notebook (publication quality 3D plots)



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

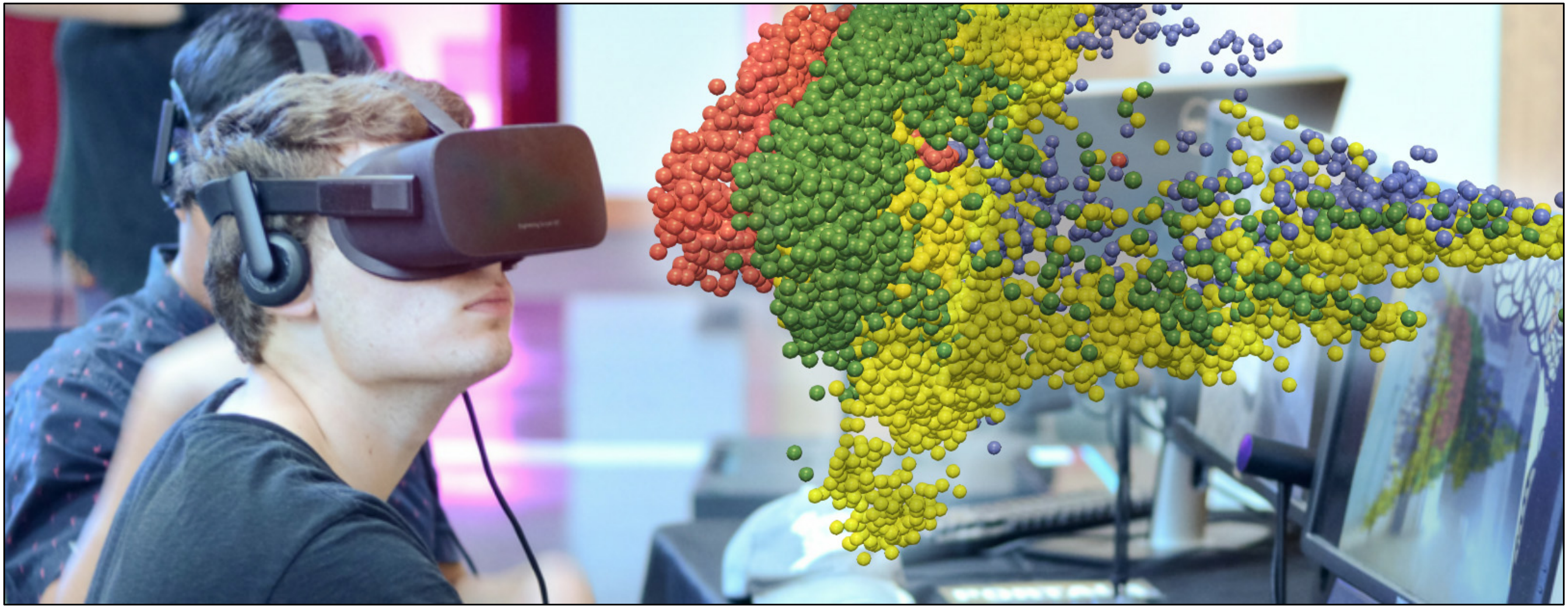


Visualising C. *Elegans*: Santiago Lombeyda, CD3, and Paul Sternberg, Biology, Caltech



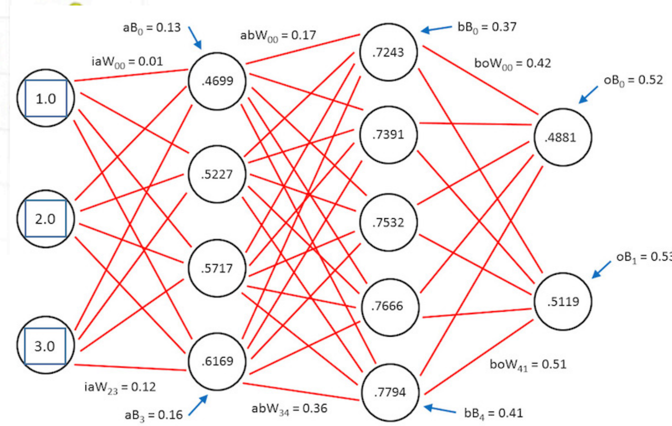
Virtual Teaching Labs

- 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



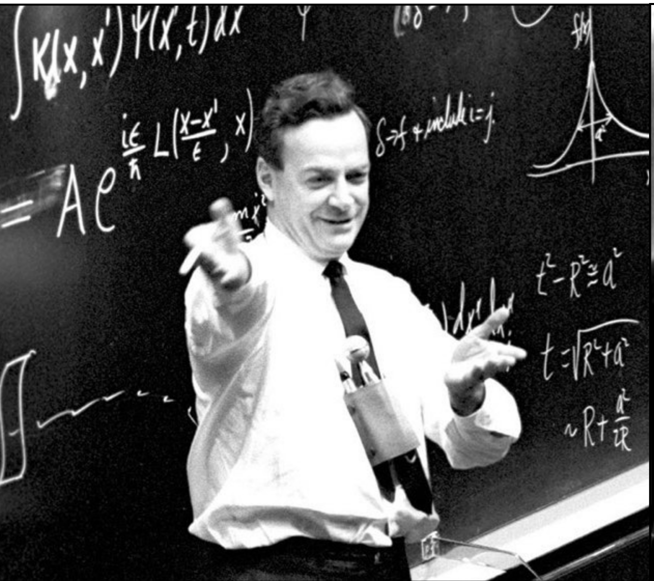
Visualization is at the nexus of Data, Human pattern recognition and understanding, and Machine Intelligence

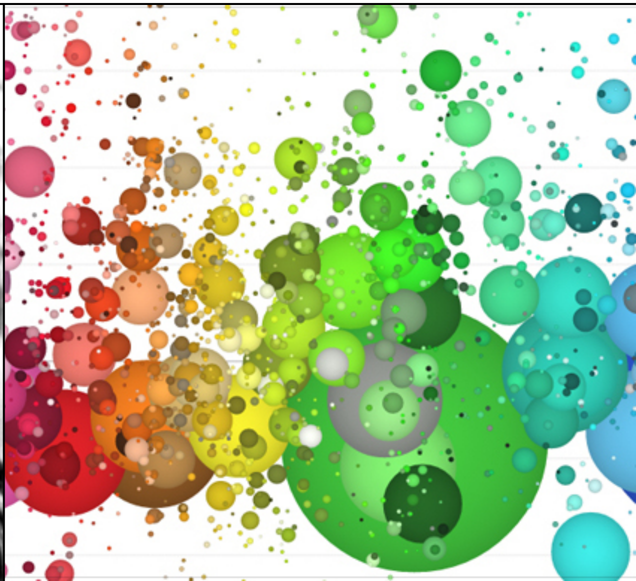
VR/AR is the most powerful platform to date that connects them



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

What comes after the Web and the Internet?

Djorgovski

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…

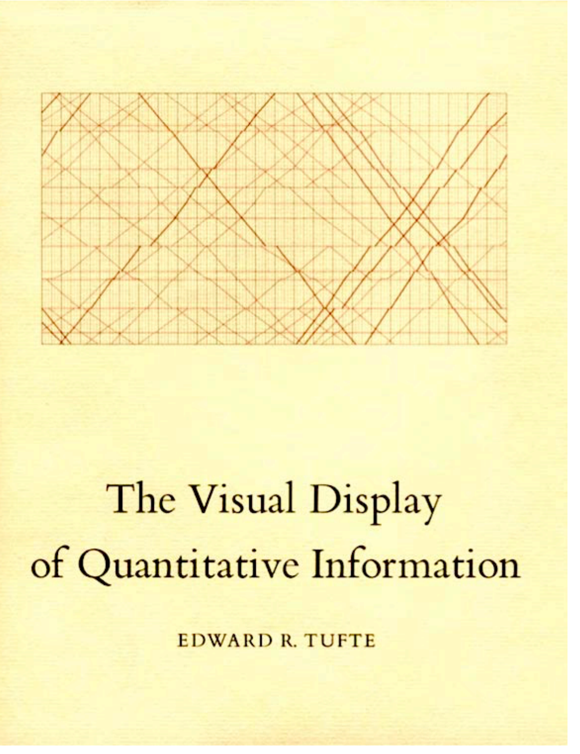
Information Dashboard Design
STEPHEN FEW

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