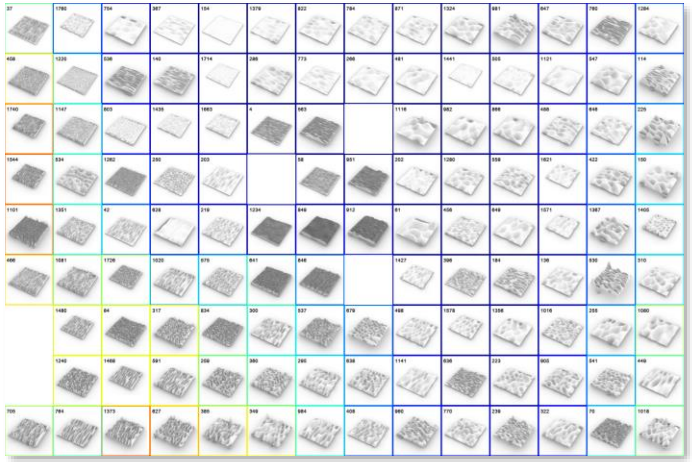
Data Processing and Visualization 2

Lecture overview: Data Processing and Visualization 2

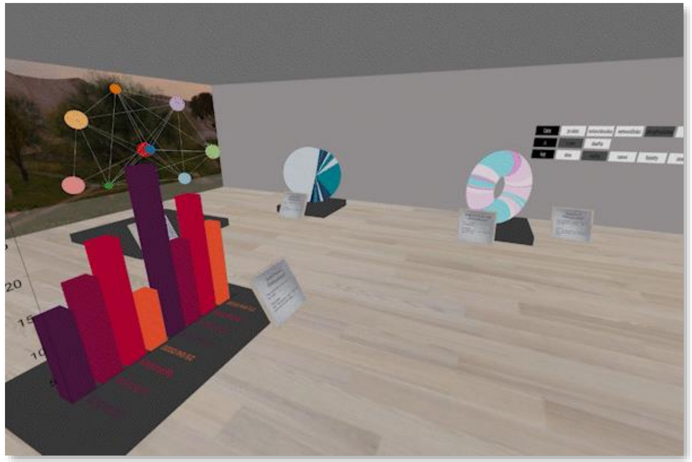
"Understanding and exploring high dimensional data"

Dimensionality reduction



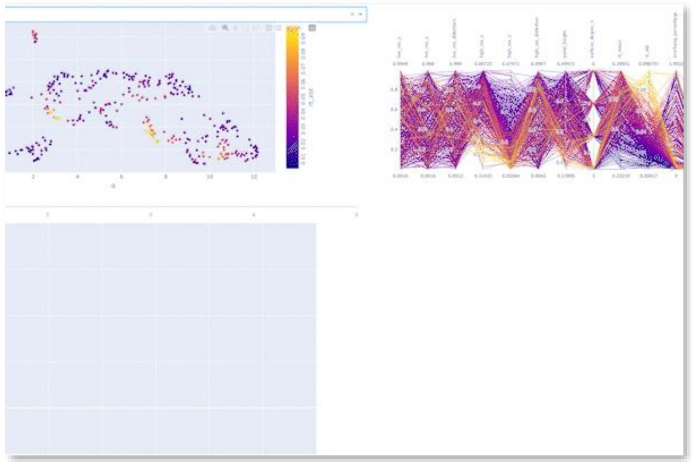
PCA, SOM, UMAP

Immersion



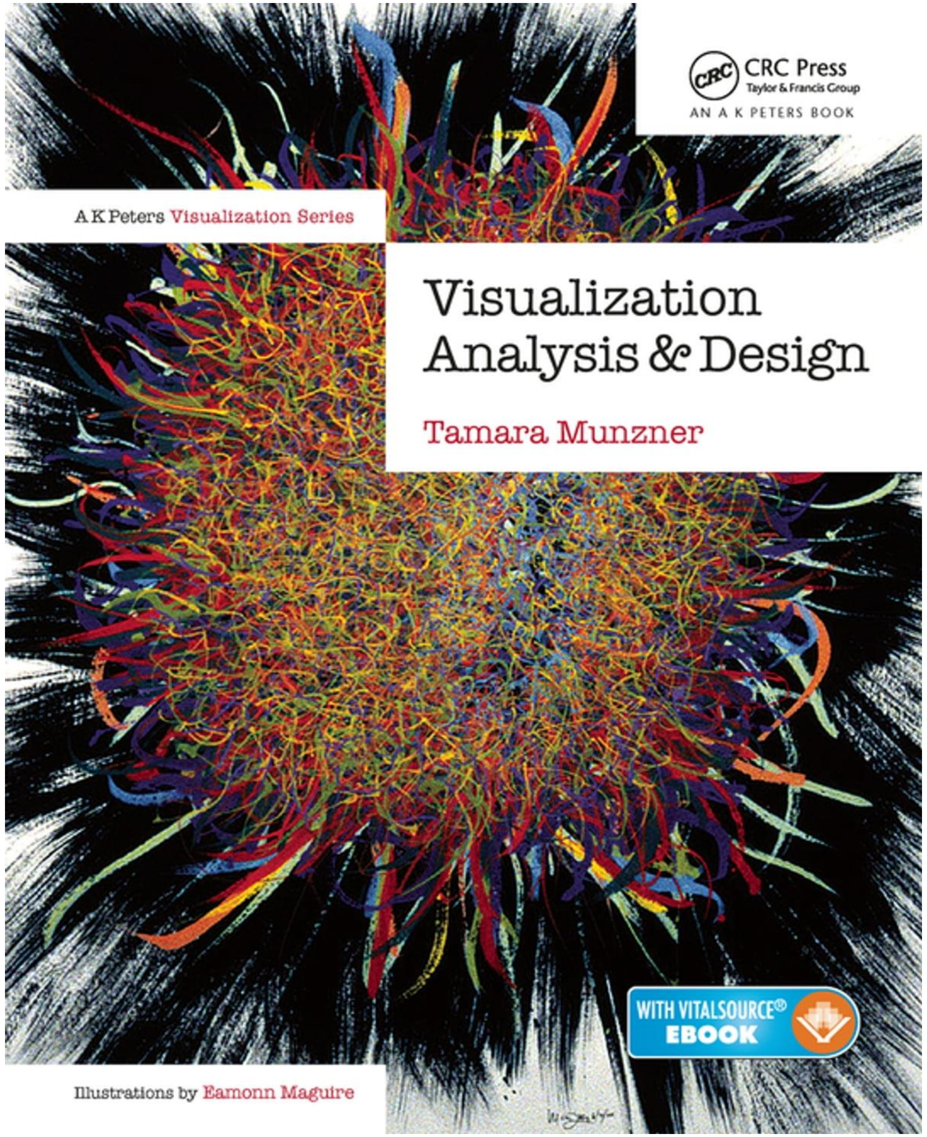
Data representation in 3D

Interaction



Interactive data visualization and annotation

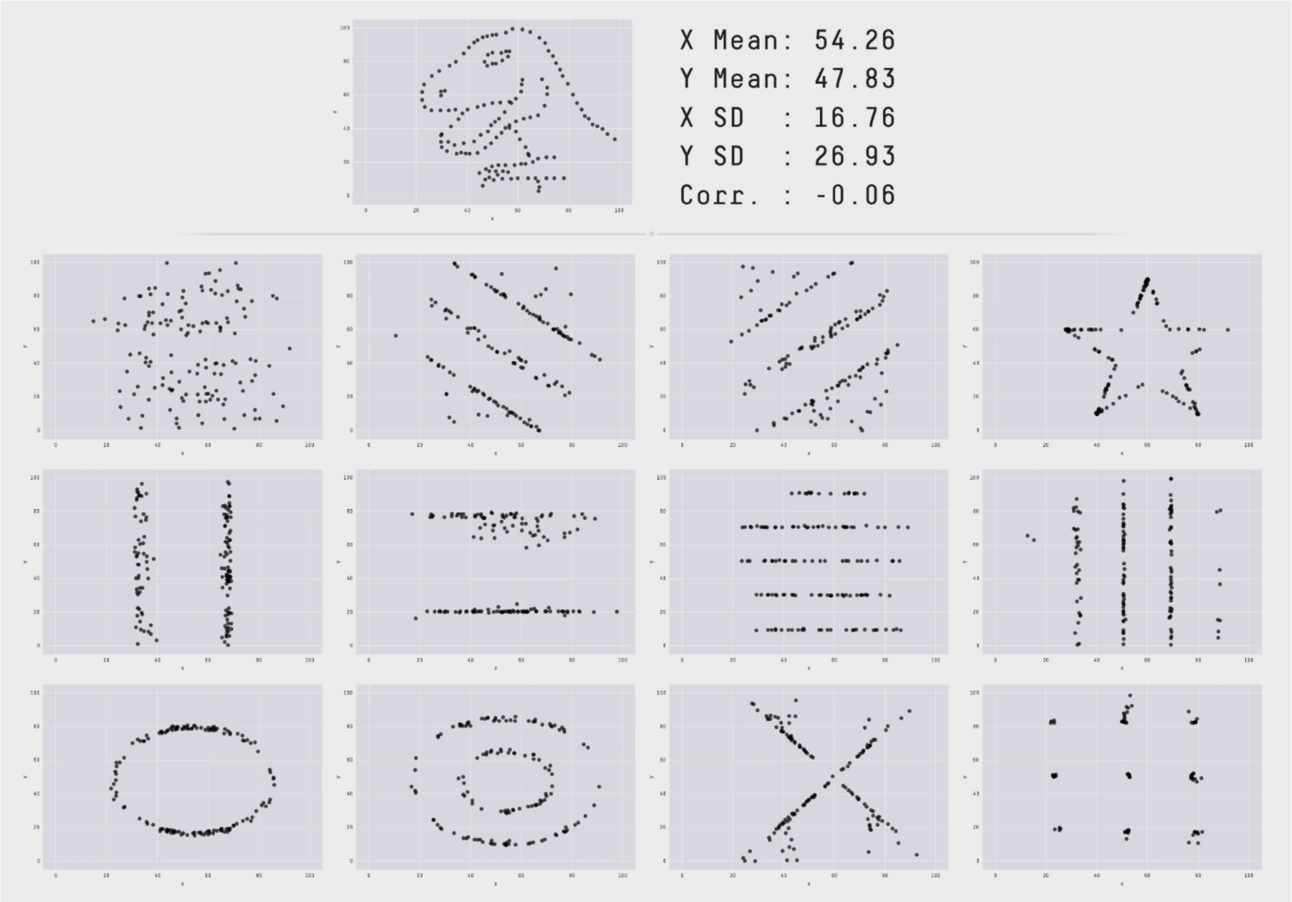
Why?

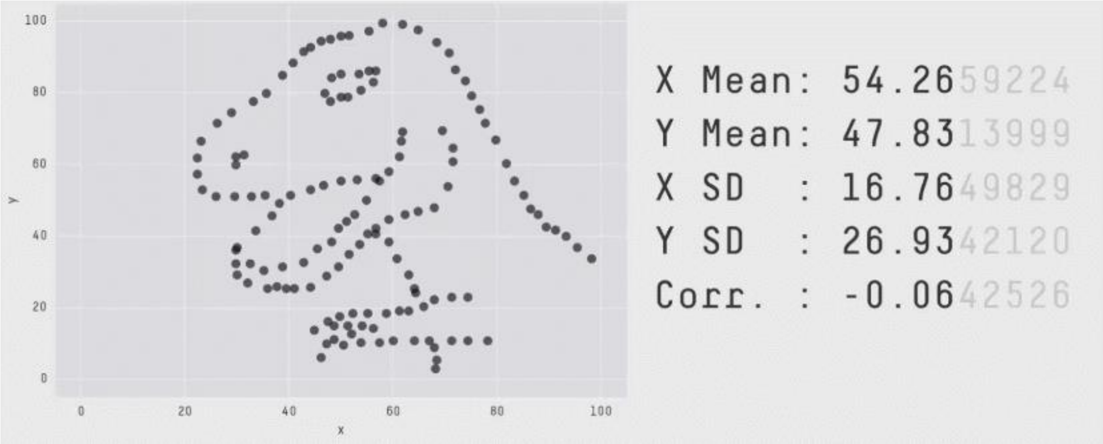

CRC Press
Taylor & Francis Group
AN A K PETERS BOOK
A K Peters Visualization Series
Visualization
Analysis & Design
Tamara Munzner
WITH VITALSOURCE®
EBOOK
Illustrations by Eamonn Maguire

Why have a human in the decision-making loop?

- Questions unknown in the beginning
- Analysis problems ill specified
- Visualisation if goal is to augment human capabilities

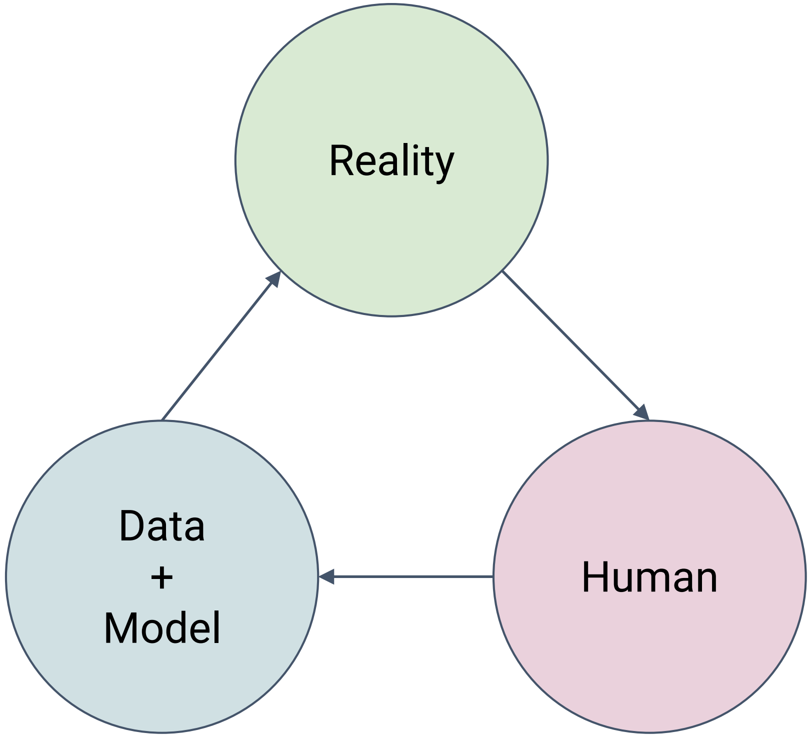
The Datasaurus Dozen





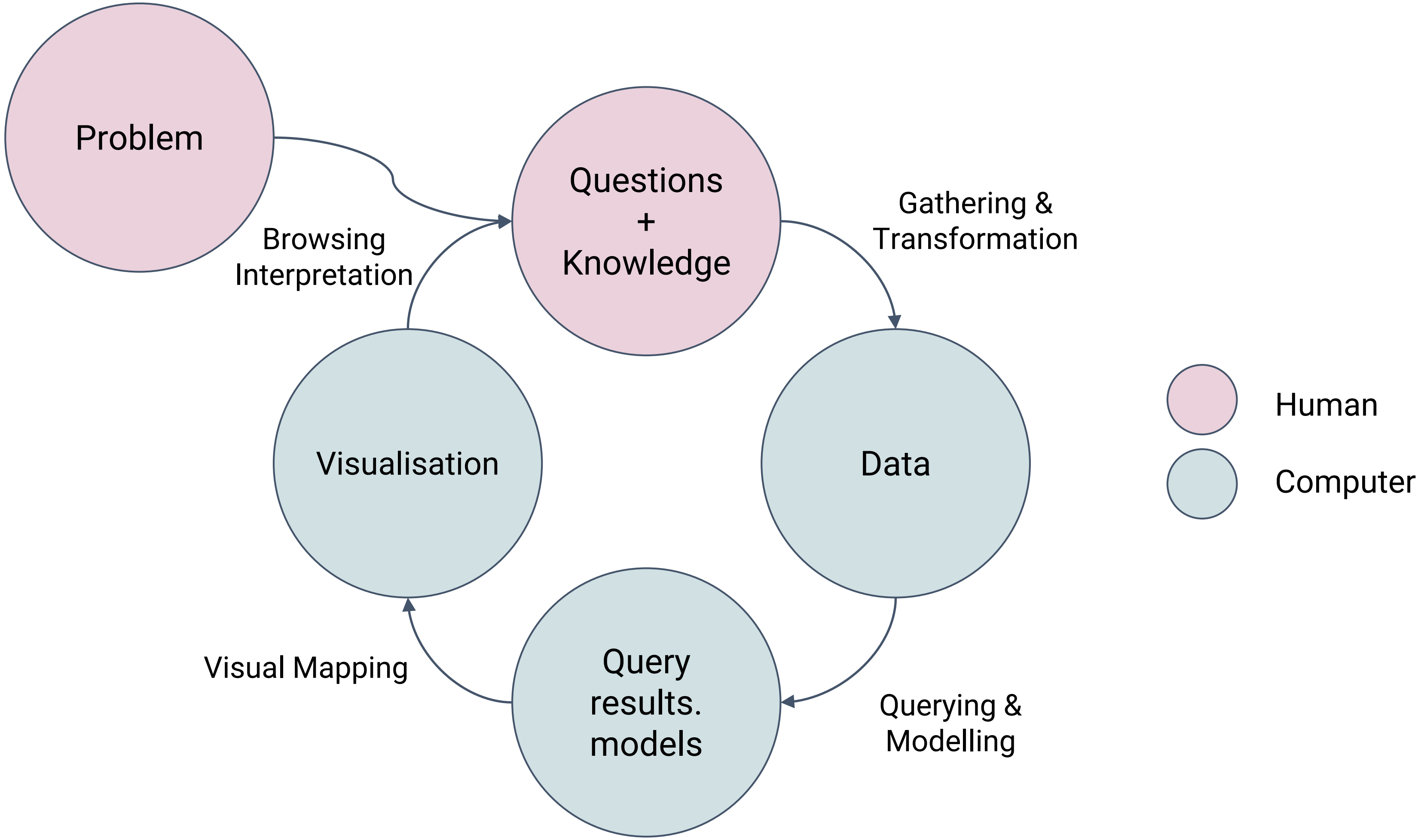
https://www.autodesk.com/research/publications/same-stats-different-graphs

Understanding something



How does interactive data analysis work?

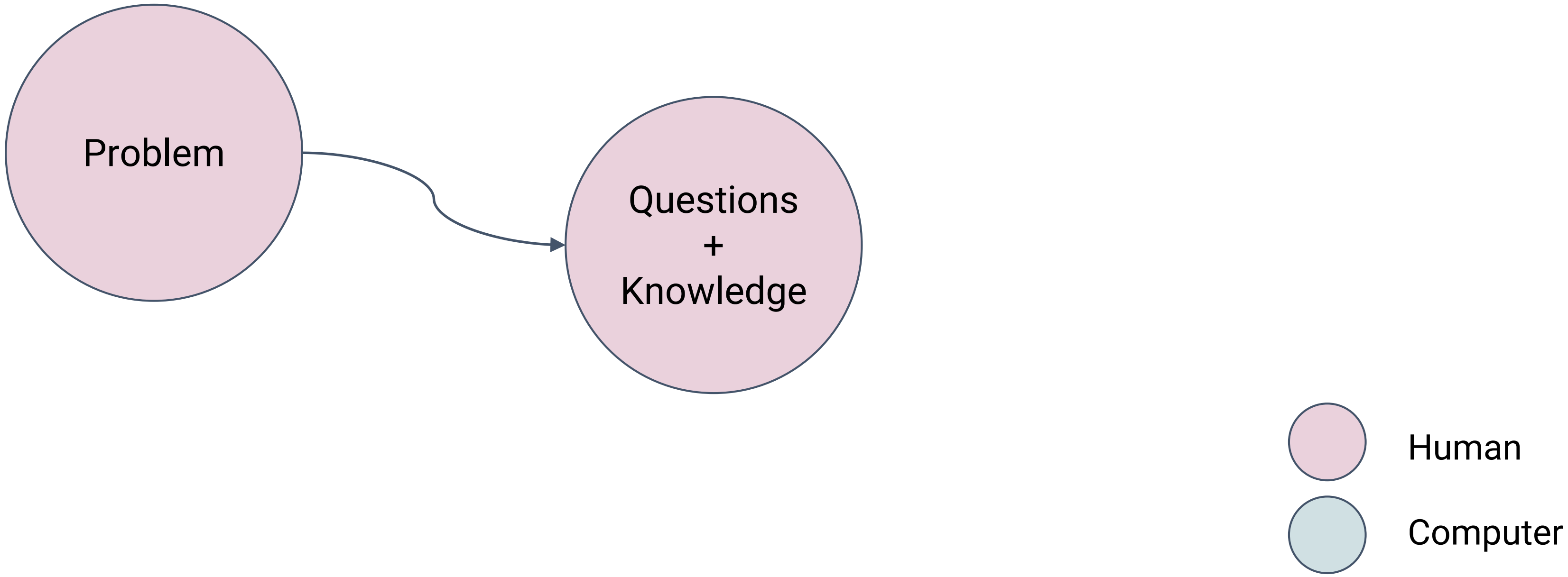
How Does Interactive Data Analysis Work?



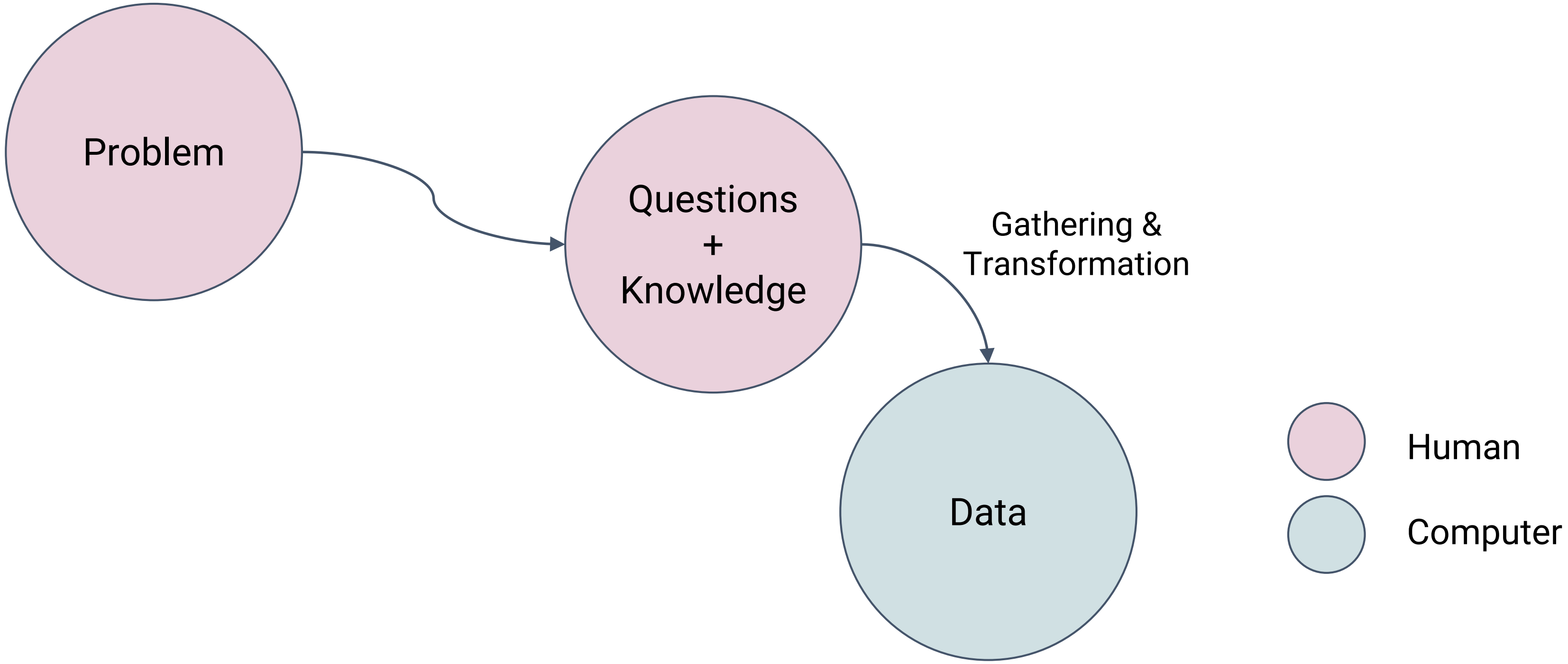
cf. https://medium.com/@FILWD/from-data-visualization-to-interactive-data-analysis-e24ae3751bf3

How Does Interactive Data Analysis Work?

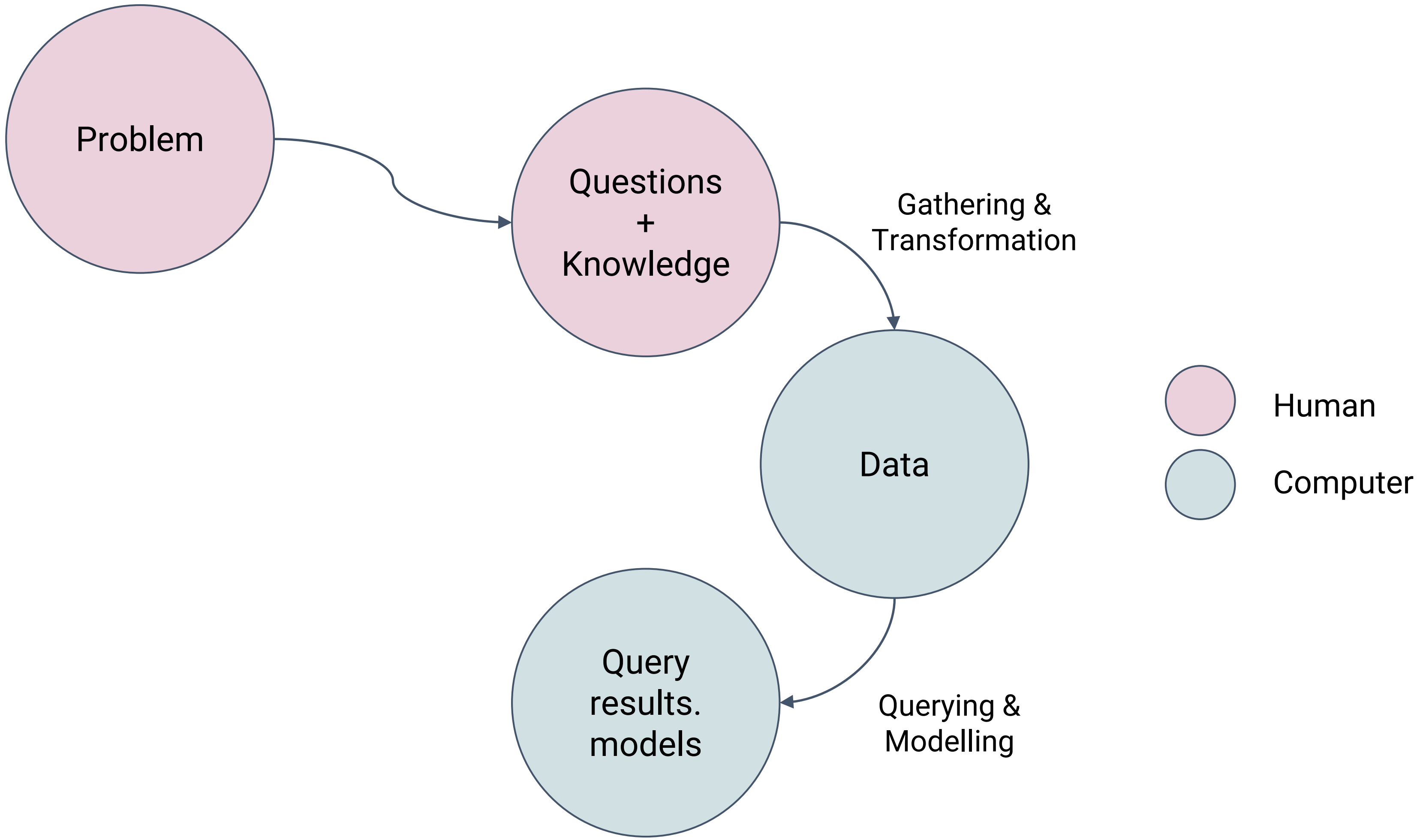
How Does Interactive Data Analysis Work?



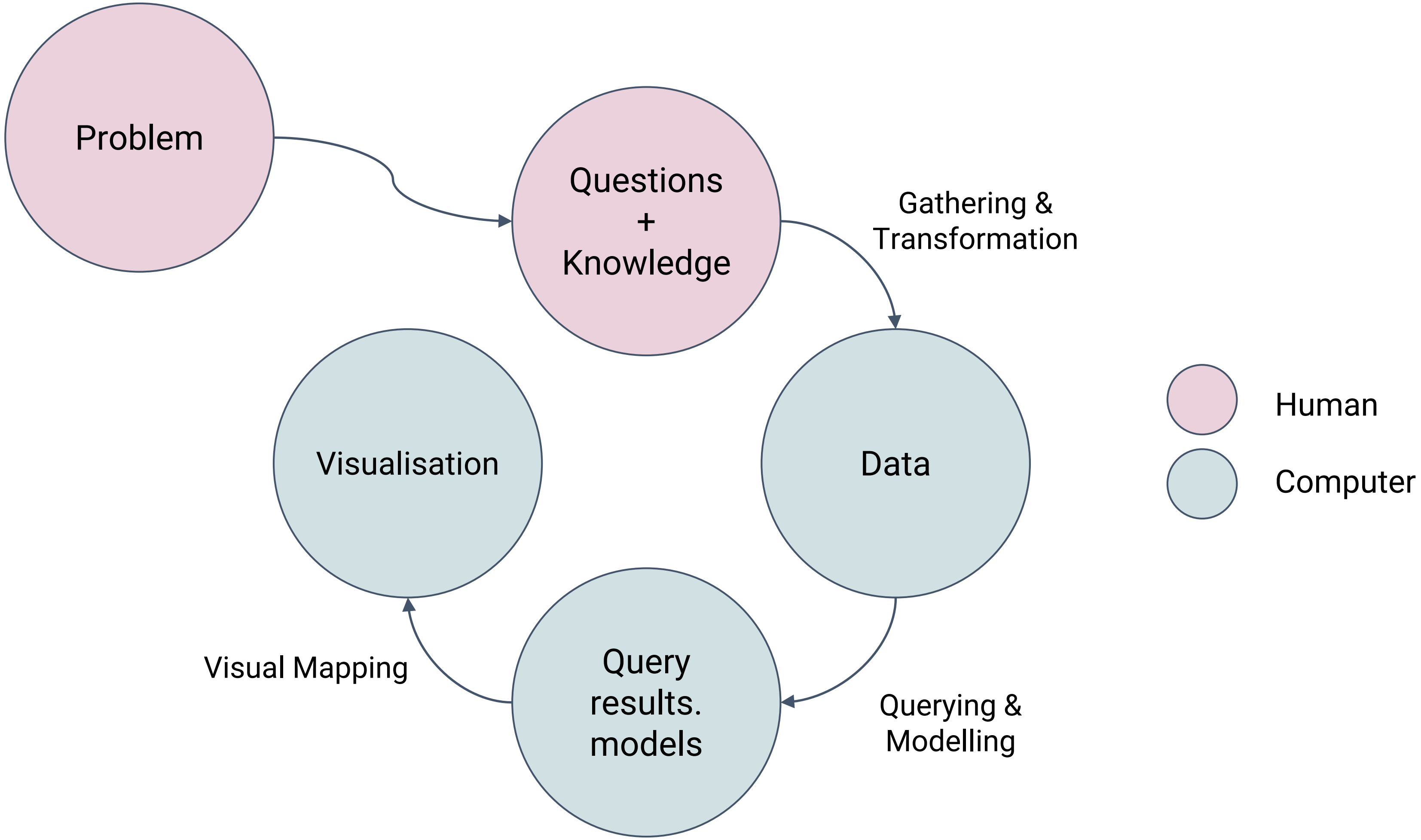
How Does Interactive Data Analysis Work?



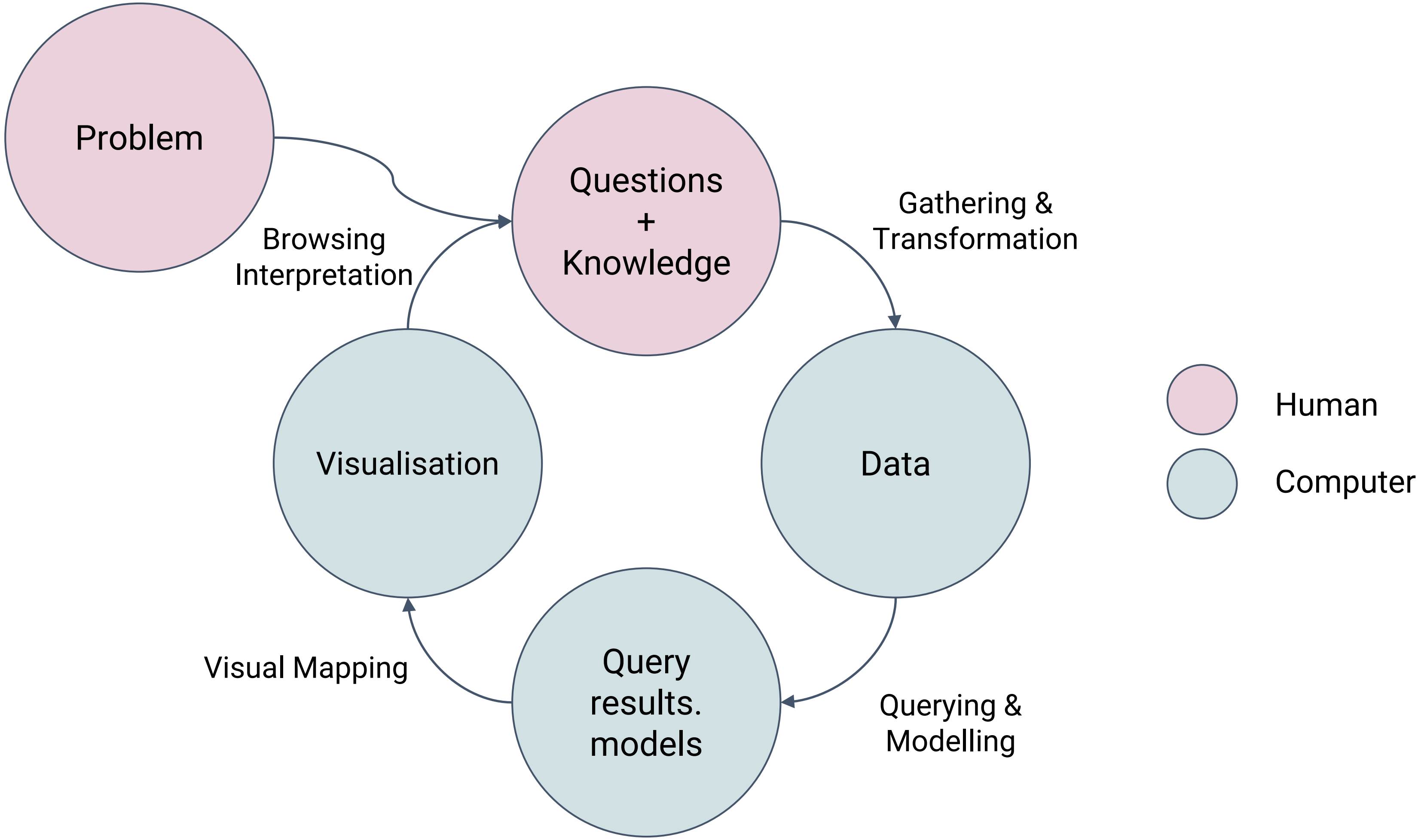
How Does Interactive Data Analysis Work?



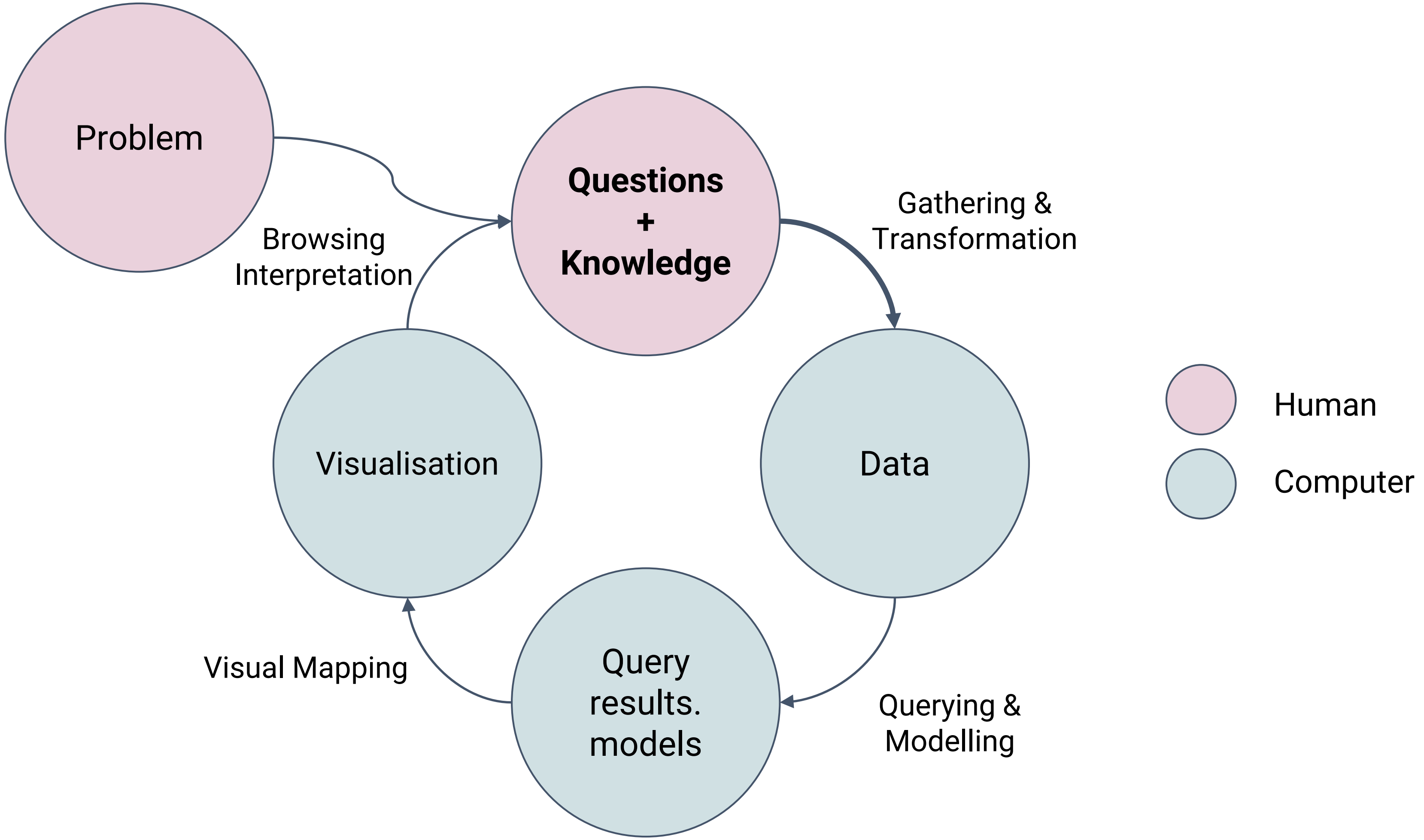
How Does Interactive Data Analysis Work?



How Does Interactive Data Analysis Work?



How Does Interactive Data Analysis Work?

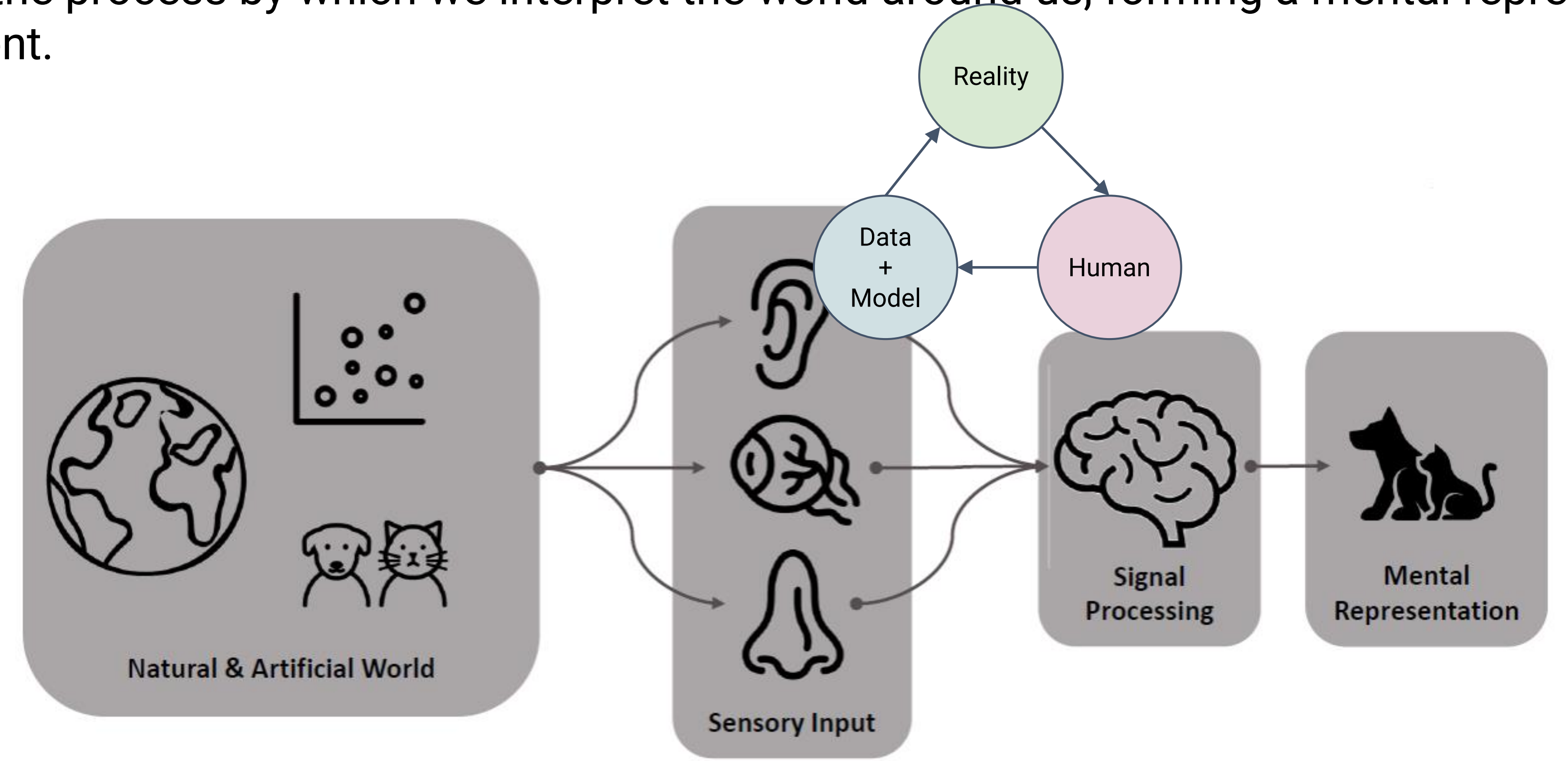


Summary

1. **The process is not sequential and is highly iterative.** While these steps are presented as a sequence, the real process is not. People jump from one step to another as more of the problem, requirements and limitations are understood. It's also highly iterative. You typically come up with an initial question, do the work to generate an answer and as you go through this process generate new questions and needs and start over again.
2. **Some activities are exclusively human.** Many steps are exclusively human: defining problems, generating questions, interpreting the results and generating inferences and new questions. It's all about human activities, not computational activities. Which leads us to ask: how much do we know about how humans think with data? How can we expand our knowledge to improve this process?
3. **Visualization is just a small portion of the process.** For data visualizers, this is a crucial observation. As much as we love the visualization step, we have to recognize that when visualization is used for data analysis, it represents just a small portion of a much more varied set of activities. This is not to say that visualization is not important or challenging, but it's crucial to realize what is the big picture. The effectiveness of the whole process depends on all of the steps above, not just visual representation.

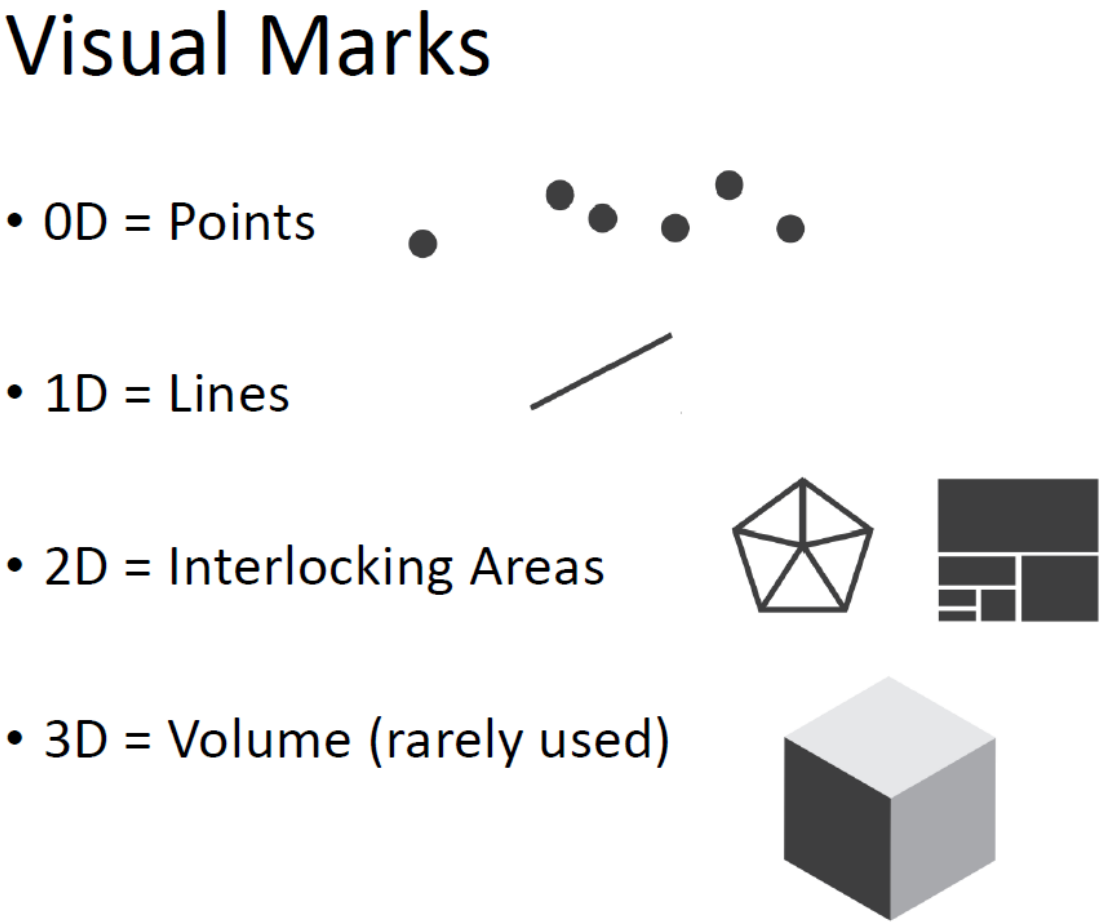
What is perception?

Perception is the process by which we interpret the world around us, forming a mental representation of the environment.



Visualization Building Blocks

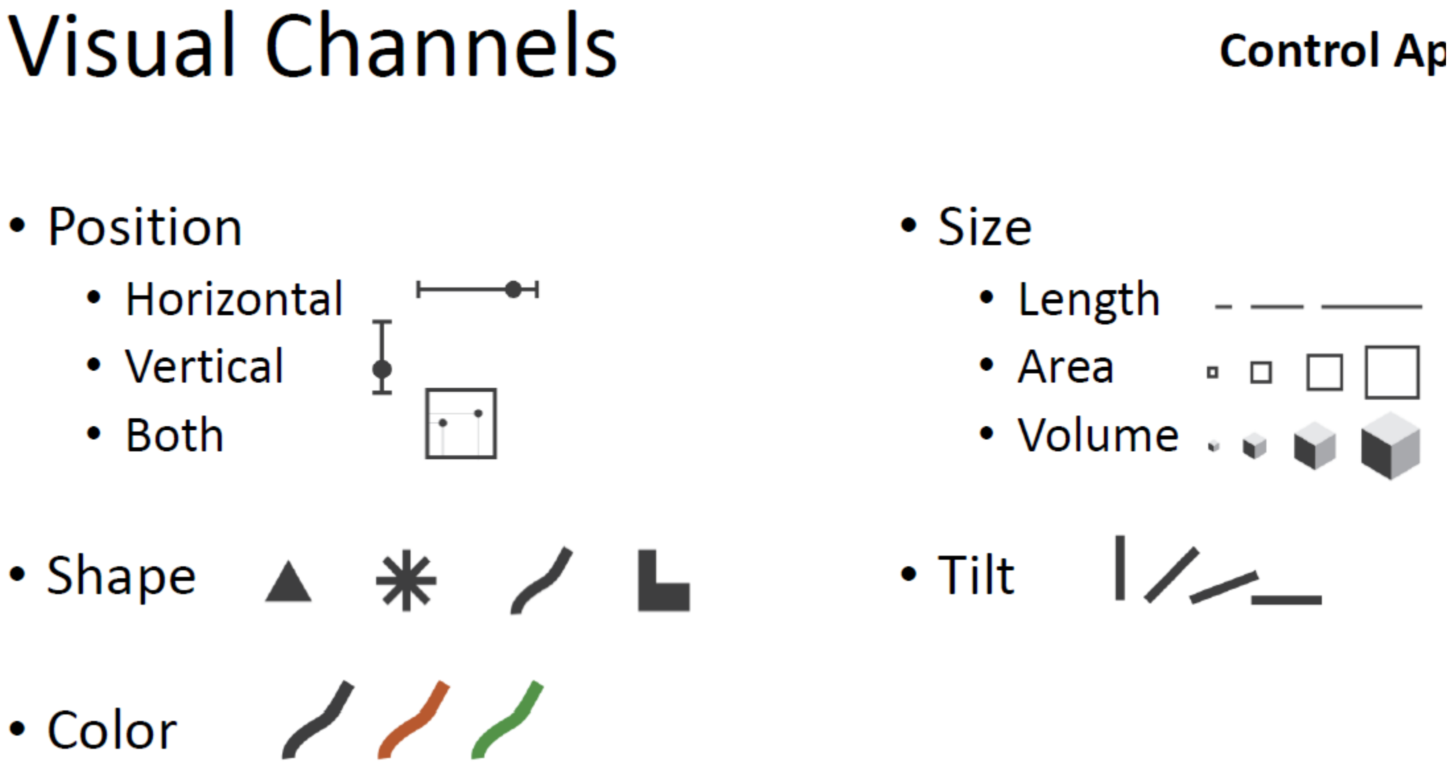
- Visual Marks and Channels
  - Gestalt Grouping
  - Color
  - Design Rules of Thumb



Munzner, Tamara. Visualization analysis and design. CRC press, 2014

Visualization Building Blocks

- Visual Marks and Channels
  - Gestalt Grouping
  - Color
  - Design Rules of Thumb

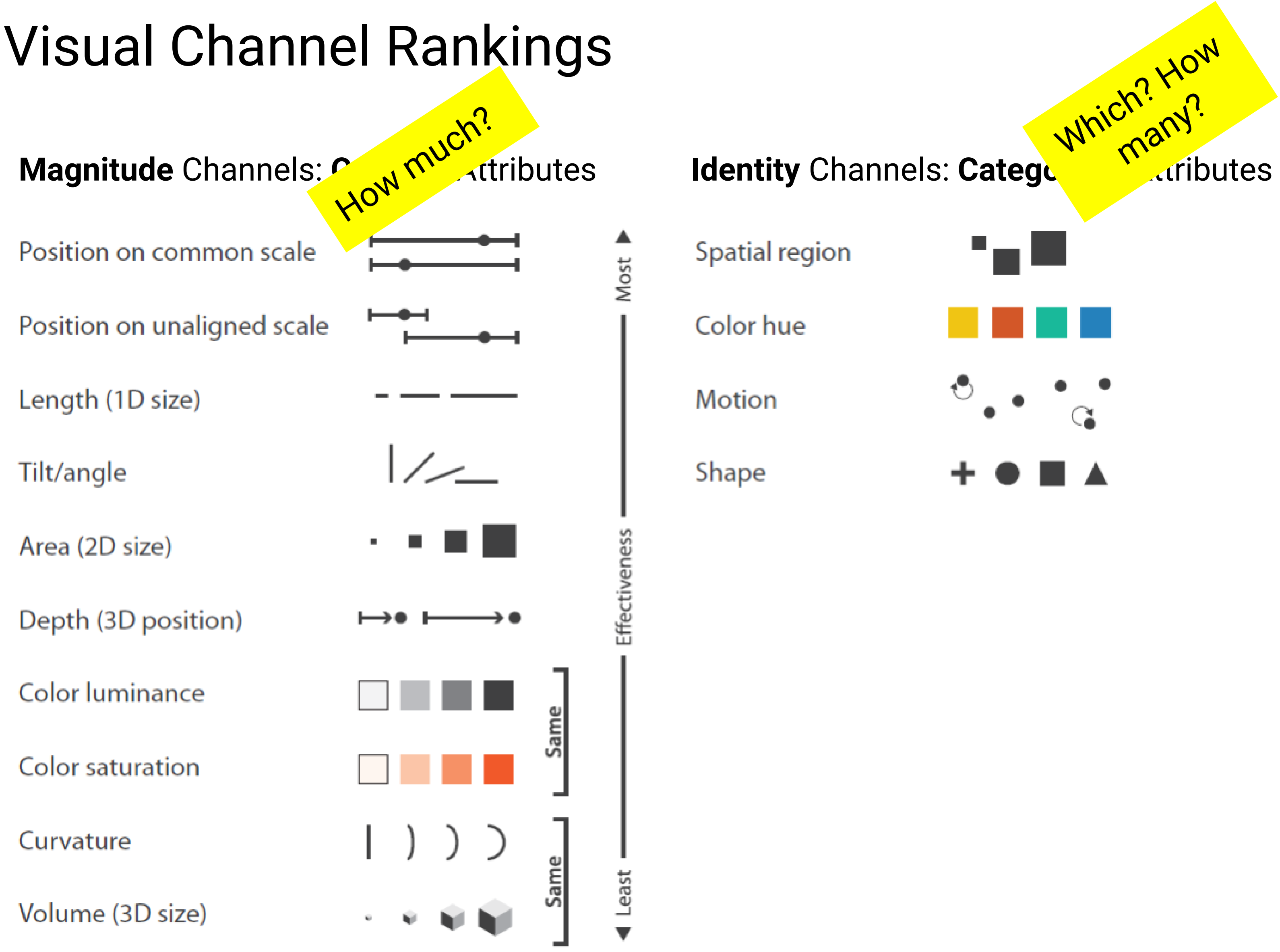


Munzner, Tamara. Visualization analysis and design. CRC press, 2014

Channel Effectiveness

- **Accuracy**
  - How precisely can we tell the difference between encoded items?
- **Discriminability**
  - How many unique steps can we perceive?
- **Separability**
  - Is our ability to use this channel affected by another one?
- **Popout**
  - Can things jump out using this channel?

Visual Channel Rankings



Munzner, Tamara. Visualization analysis and design. CRC press, 2014

Stevens' Power Law

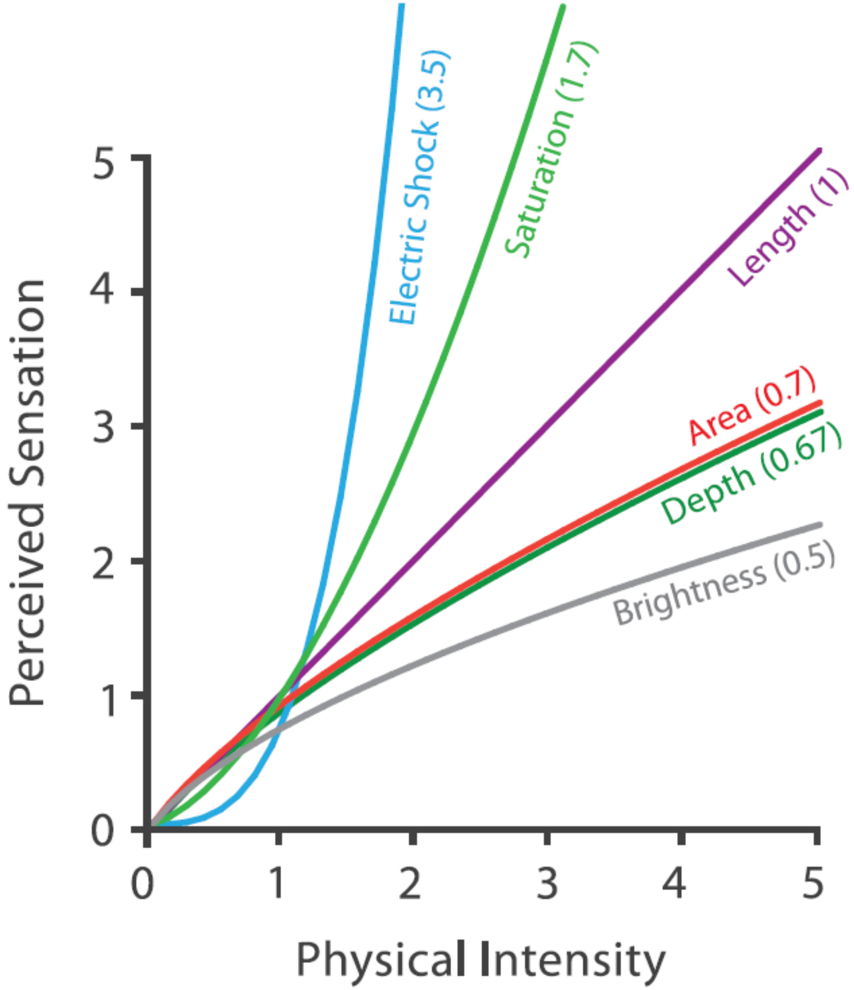
Describes a relationship between

- a physical stimulus/sensation (S) &
- its perceived, physical intensity (I).

$$S = I^N$$

$N \longrightarrow$ Determined Empirically

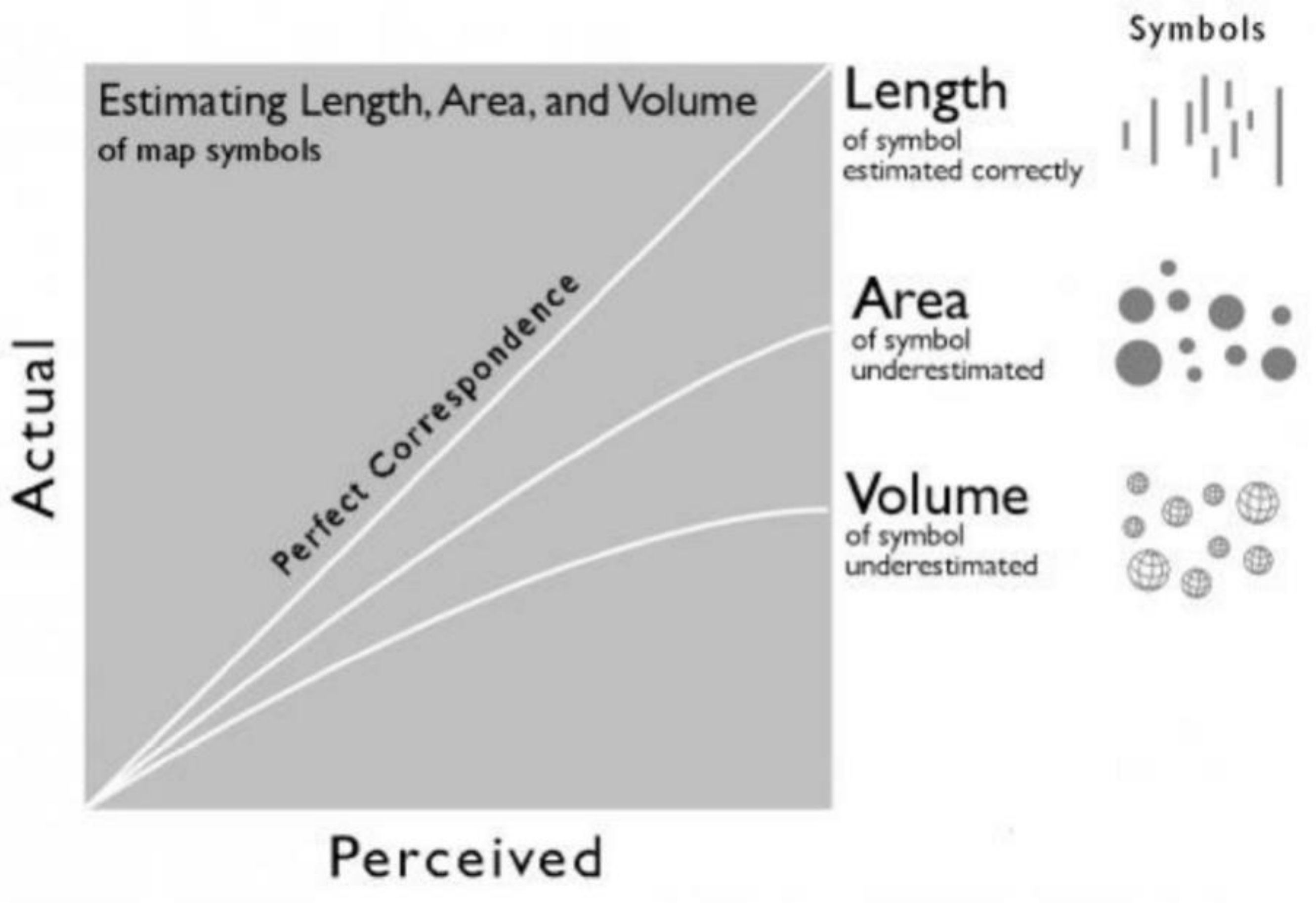
- $N<1$ = **under**estimation
- $N>1$ = **over**estimation



Munzner, Tamara. Visualization analysis and design. CRC press, 2014

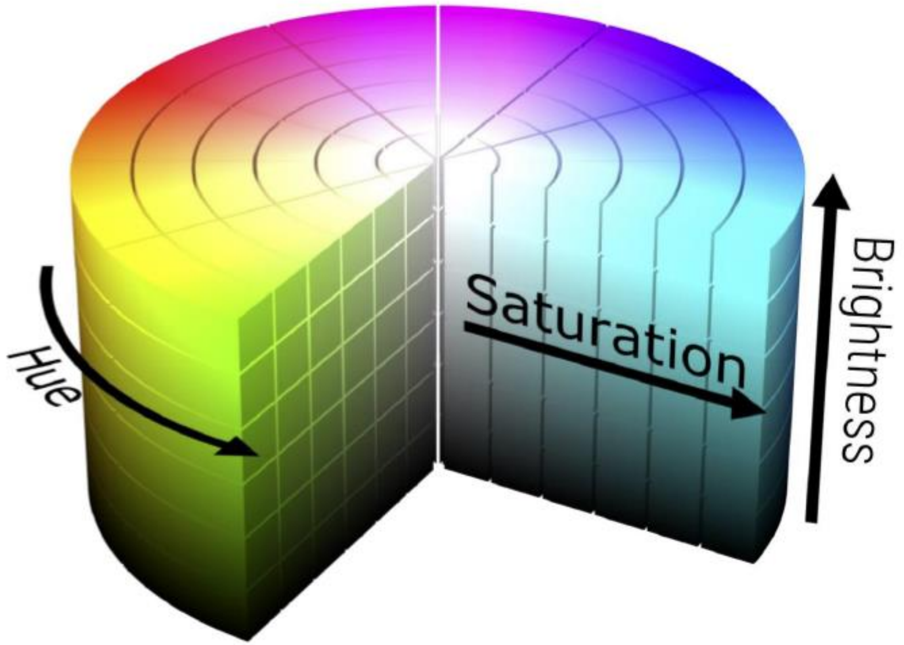
Stevens' Power Law

- People tend to correctly estimate lengths.
- They tend to underestimate areas and volumes
- This tendency gets worse as area or volume grow.



Munzner, Tamara. Visualization analysis and design. CRC press, 2014

Color Spaces



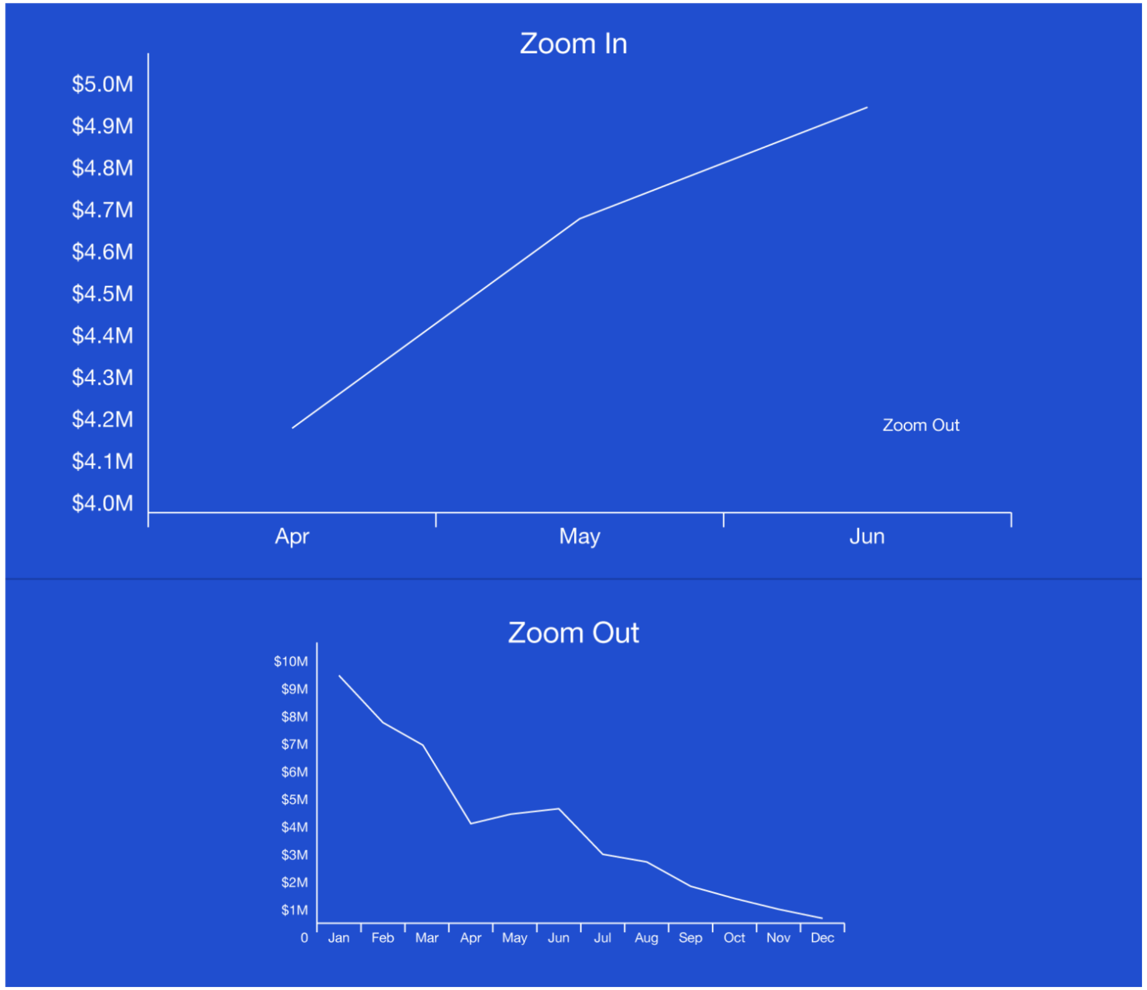
Designing Visualizations

Visualizations help:

- To describe some structure, patterns or anomaly in the data.
- To explore and analyze large datasets.
- To make effective use of the information overflow.
- To communicate information to people.

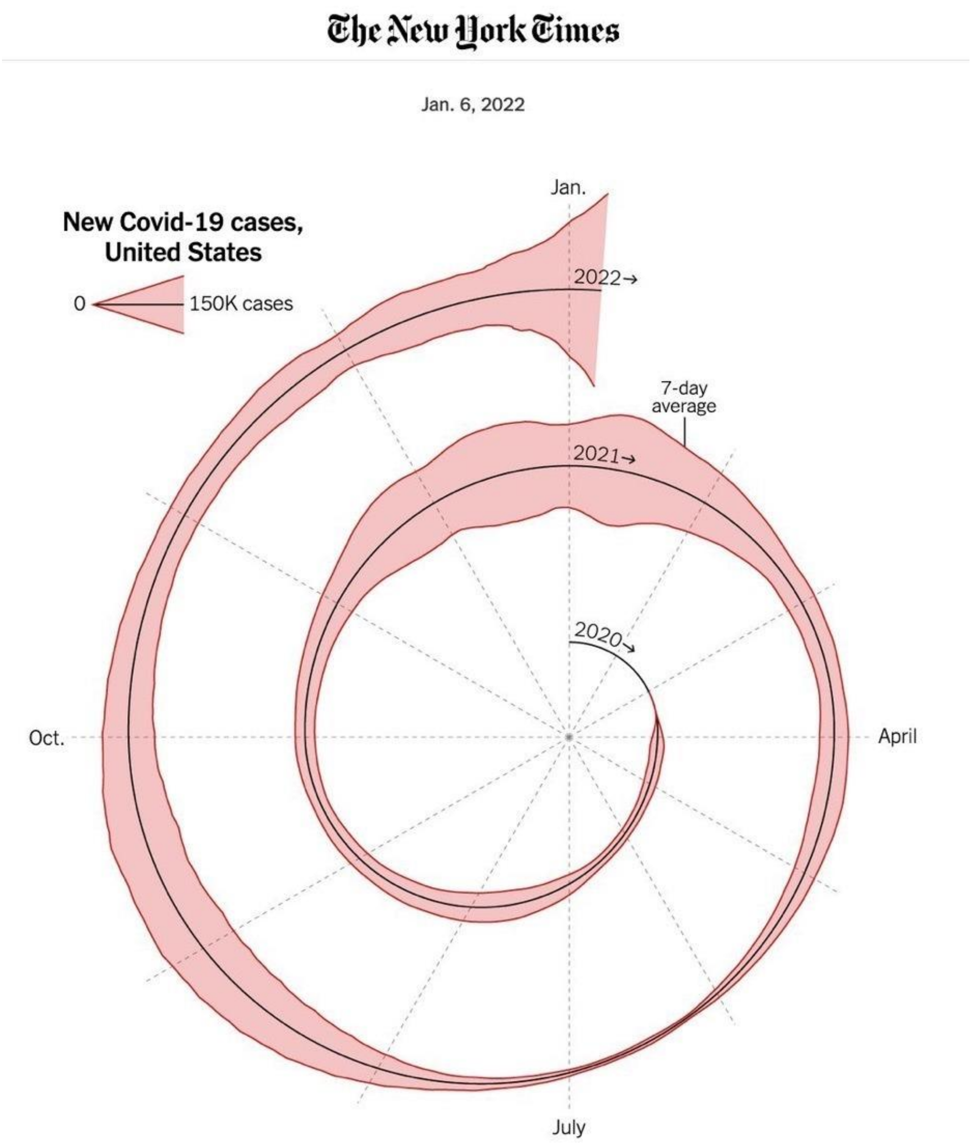
However, visualization can distort the "truth"!

Visualisations that distort the truth



https://www.toptal.com/designers/ux/data-visualization-mistakes

Visualisations that make no sense



https://viz.wtf/

Designing Visualizations

Expressiveness

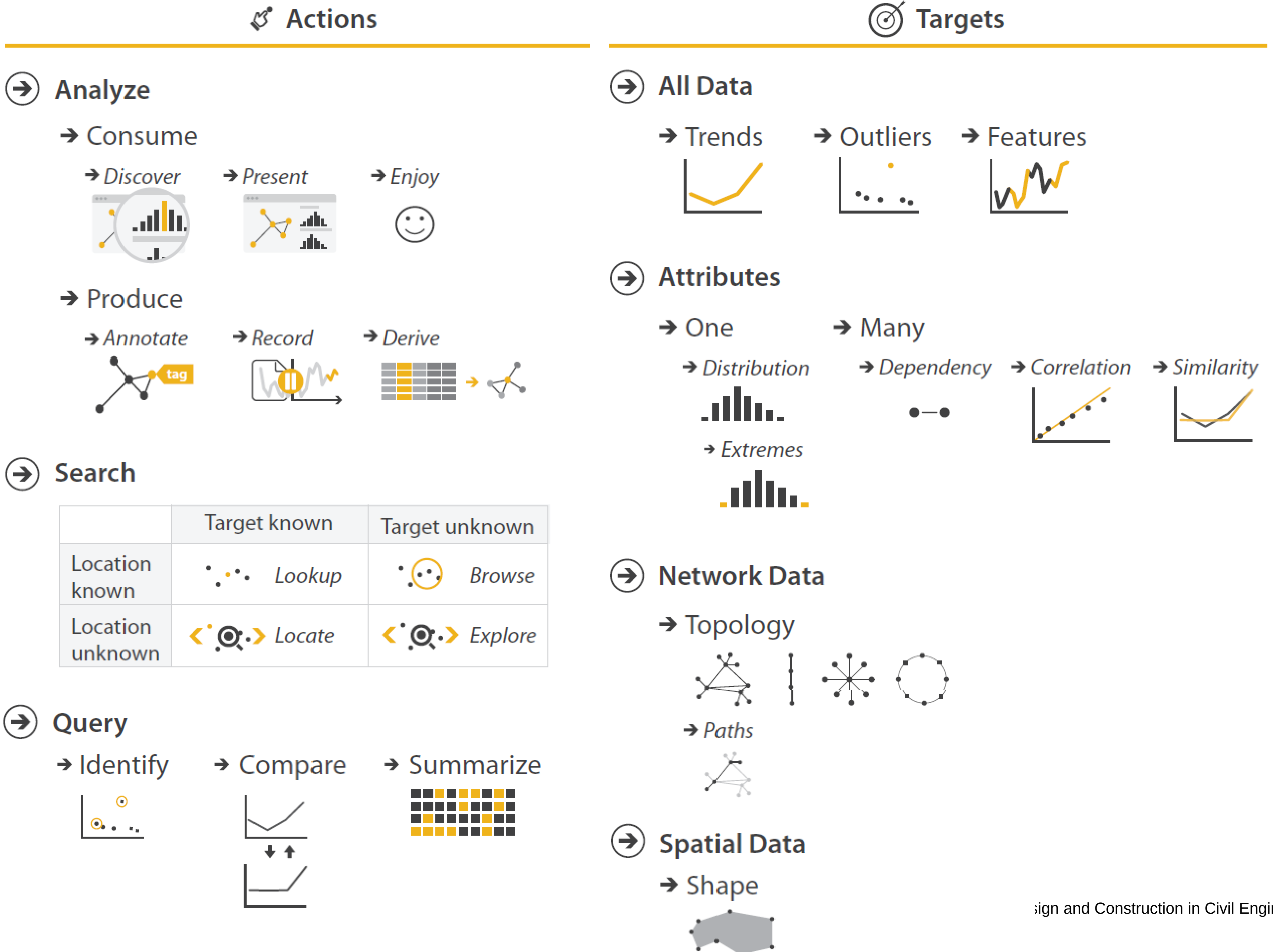
Visualization presents all the information and only the information.

Effectiveness

Visualization is effective when it can be interpreted accurately and quickly,

and when it can be rendered in a cost-effective manner.

Lie Factor = (Size of effect in graphic)/(Size of effect in data)



Actions
Analyze
Consume
Discover
Present
Enjoy
Produce
Annotate
tag
Record
Derive
Search
Target known
Target unknown
Location known
Lookup
Browse
Location unknown
Locate
Explore
Query
Identify
Compare
Summarize
Targets
All Data
Trends
Outliers
Features
Attributes
One
Distribution
Extremes
Many
Dependency
Correlation
Similarity
Network Data
Topology
Paths
Spatial Data
Shape

What is high-dimensional data?

What is high-Dimensional Data?

Multi- and high-dimensional data is typically provided in a table like format:

- rows = data records/objects
- columns = their dimensions, attributes, features, or descriptors

	d1	d2	d3	d4	d5	d6	d7
r1							
r2							
r3							
r4							
r5							
r6							
r7							

Dimensions can be:

- Nominal
- Ordinal
- Numeric (Interval/Ratio)
- Complex Objects
  - Vectors
  - References
  - Geo-Spatial Objects
  - …

Challenges

- Datasets with a mix of nominal, ordinal, and numeric dimensions make it **difficult to compute relations** between objects (e.g., similarity).

- The number of dimensions highly **influence the interpretability** of similarity measures.

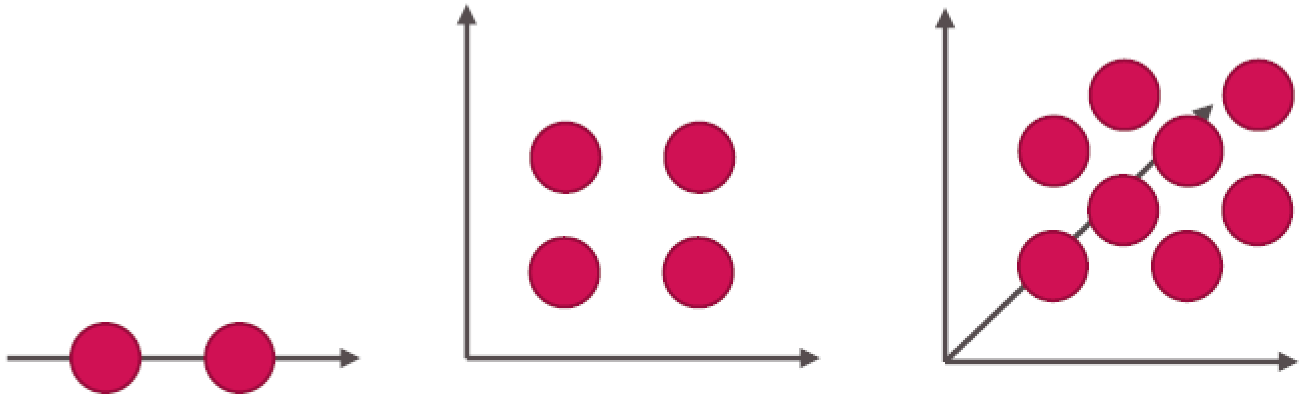
- Non-relevant, redundant, and conflicting dimensions may **hide interesting patterns** in a sea of noise.

- A large number of dimensions increase the **possible visual mappings** and the arrangement of dimensions.

Ther Curse of Dimensionality

More Dimensions → Data Sparsity

To preserve data intervals, we need exponentially more points!



N^D $\quad 2^1 = 2 \quad 2^2 = 4 \quad 2^3 = 8$

Impact on Distance/Similarity Calculations: If your distance/similarity function assumes all dimensions are equally relevant, adding more (irrelevant) dimensions will decrease distances between data points.

https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e

Dimensionality reduction

What is Dimensionality Reduction?

Most Dimensionality Reduction applications are used for:

- **Data Compression**
- **Noise Reduction**
- **Data Classification**
- **Data Visualization**

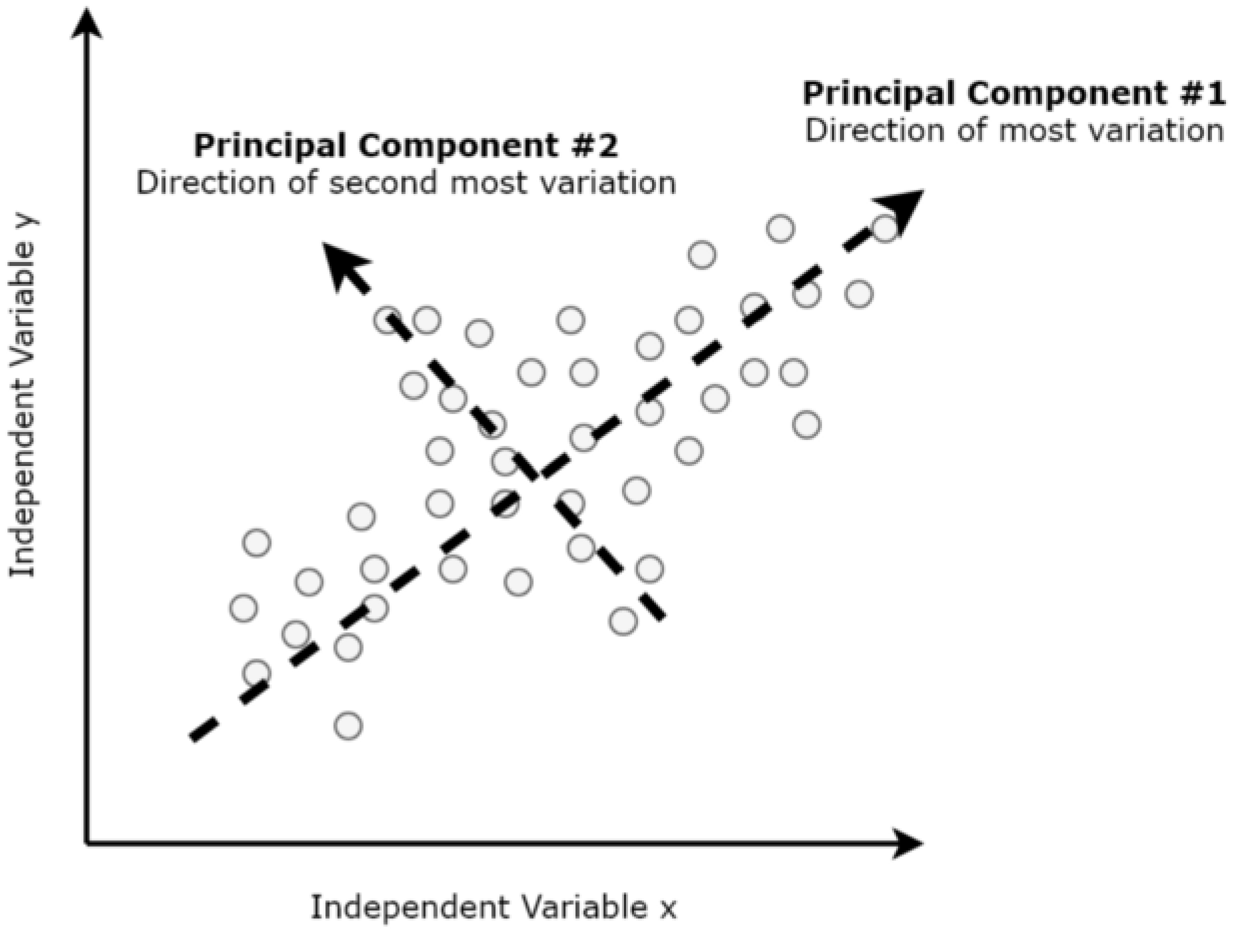
What is Dimensionality Reduction?

The two main approaches to reducing dimensionality: **Linear Projection** and **Non-Linear Manifold Learning**

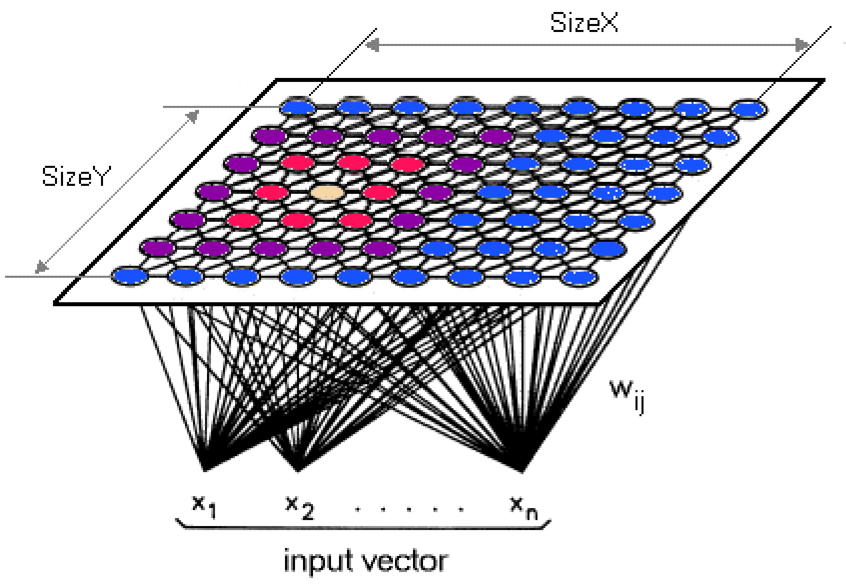
Projection: This technique deals with projecting every data point which is in high dimension, onto a subspace suitable lower-dimensional space in a way which approximately preserves the distances between the points.

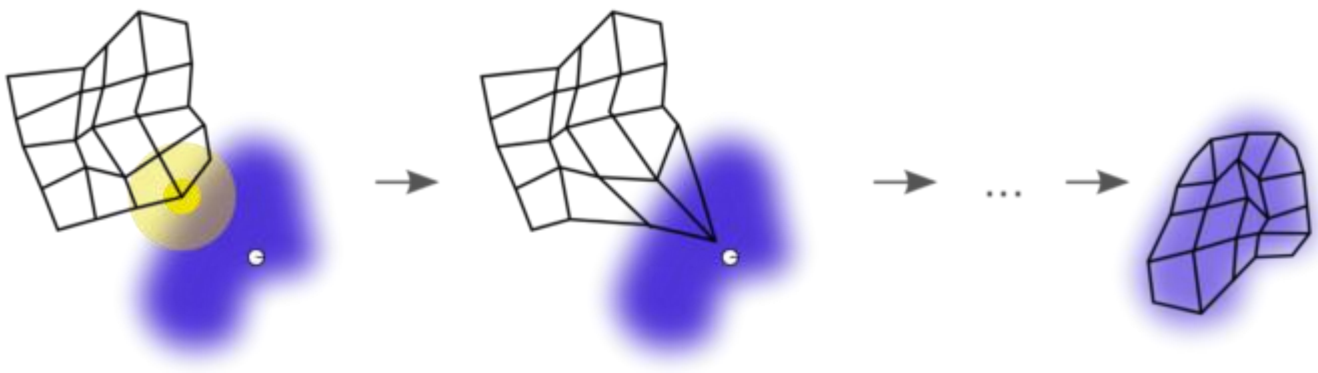
Manifold Learning: Many dimensionality reductions algorithm work by modelling the manifold on which the training instance lie; this is called *Manifold learning*. It relies on the manifold hypothesis or assumption, which *holds that most real-world high-dimensional datasets lie close to a much lower-dimensional manifold,* this assumption in most of the cases is based on observation or experience rather than theory or pure logic.

PCA


Principal Component #1
Direction of most variation
Principal Component #2
Direction of second most variation
Independent Variable y
Independent Variable x

SOM



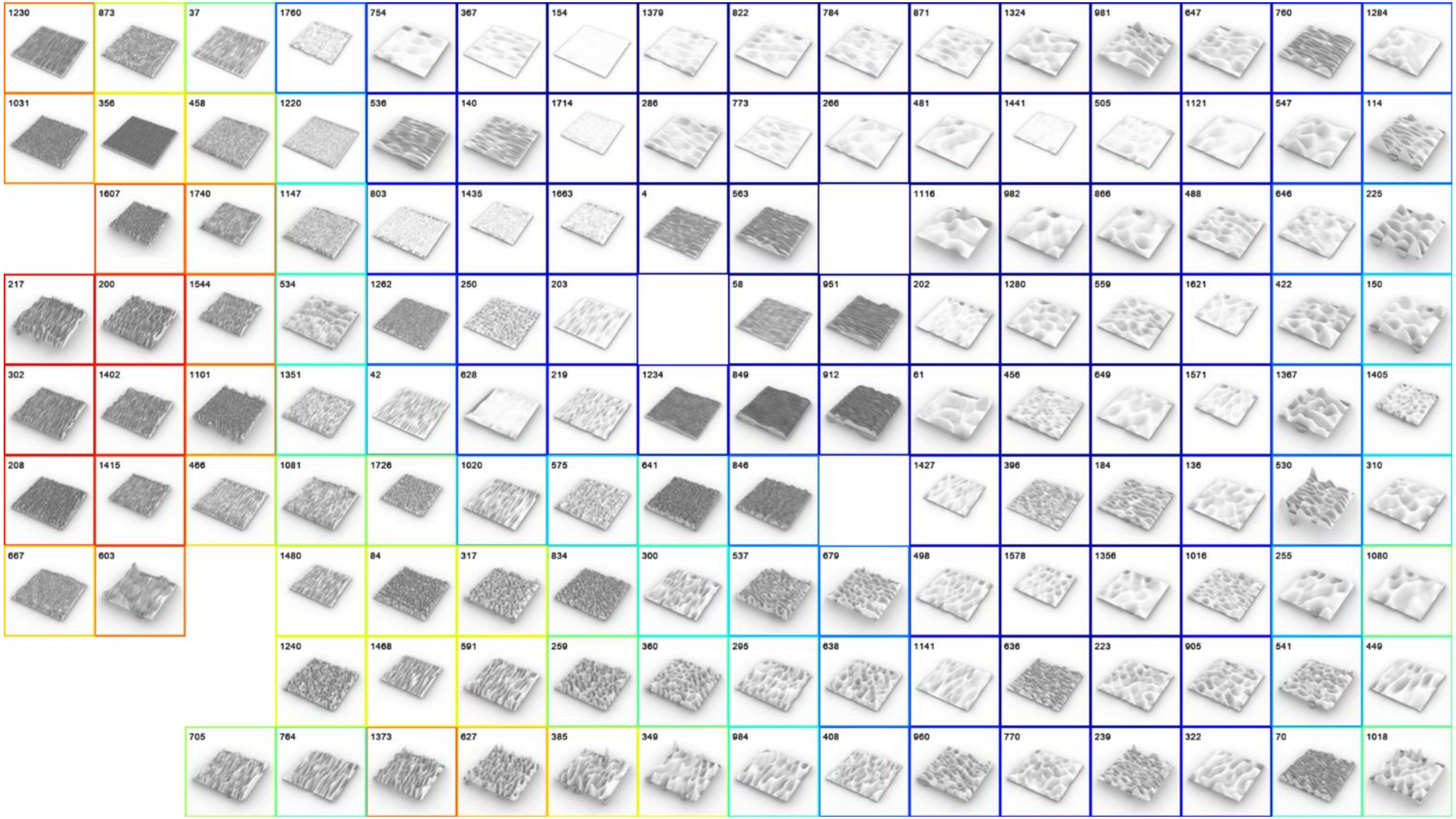


$$\Delta w_{ji} = \eta(t)\ .\ T_{j,I(\mathbf{x})}(t)\ .\ (x_i - w_{ji})$$

$$T_{j,I(\mathbf{x})} = \exp(-S_{j,I(\mathbf{x})}^2 / 2\sigma^2)$$

$$d_j(\mathbf{x}) = \sum_{i=1}^{D} (x_i - w_{ji})^2$$

https://medium.com/analytics-vidhya/introduction-to-unsupervised-deep-learning-with-self-organizing-map-with-its-architecture-and-7ac67b0c5977



AAAD Acoustics

Interactive Visualization

Interactive Visualization

What is Interactive Data Visualization?

Interactive data visualization refers to the use of modern data analysis software that enables users to directly manipulate and explore graphical representations of data. Data visualization uses visual aids to help analysts efficiently and effectively understand the significance of data. Interactive data visualization software improves upon this concept by incorporating interaction tools that facilitate the modification of the parameters of a data visualization, enabling the user to see more detail, create new insights, generate compelling questions, and capture the full value of the data.

Considerations for Interactive Visualization

- **Visual Patterns =! Data Patterns**
  - Be Careful of Visualization Artifacts!
  - Communicate Data Transformation Steps.
  - The Lack of Patterns in the Visualization does not imply a Lack of Patterns in the Data.

- **Correlations =! Causality**
  - Be Careful of Rush Interpretations of Visual/Data Patterns.
  - Clearly Communicate Uncertainties.

Data visualisation in VR => Data representation & Immersive Analytics

Opportunities

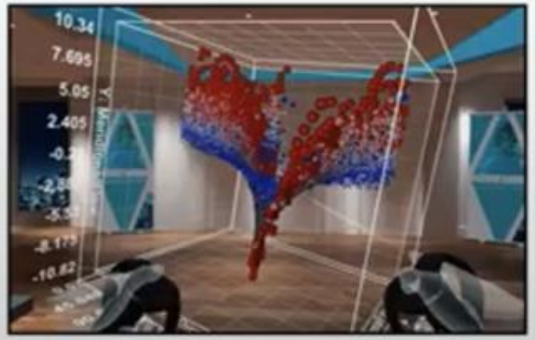
- Visual representation of 3D data
- Natural interfaces for 3D data exploration
- Multi-sensory feedback for immersive analytics

Challenges

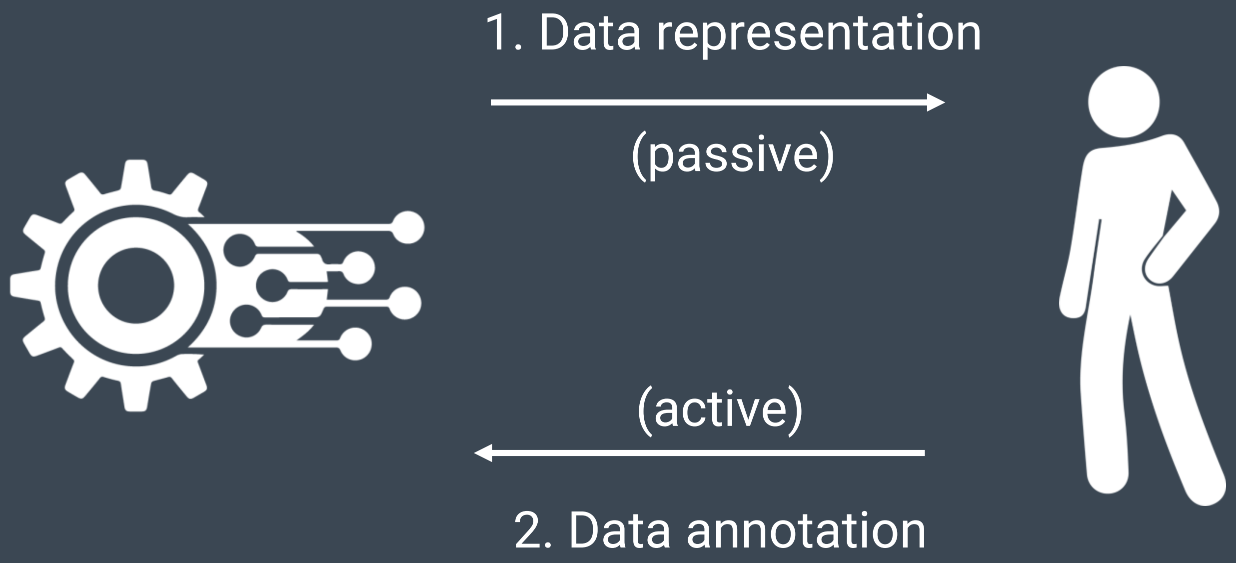
- Increase in complexity of data
- Occlusion and distortion problems
- Accurate utilization of space

2D Interfaces

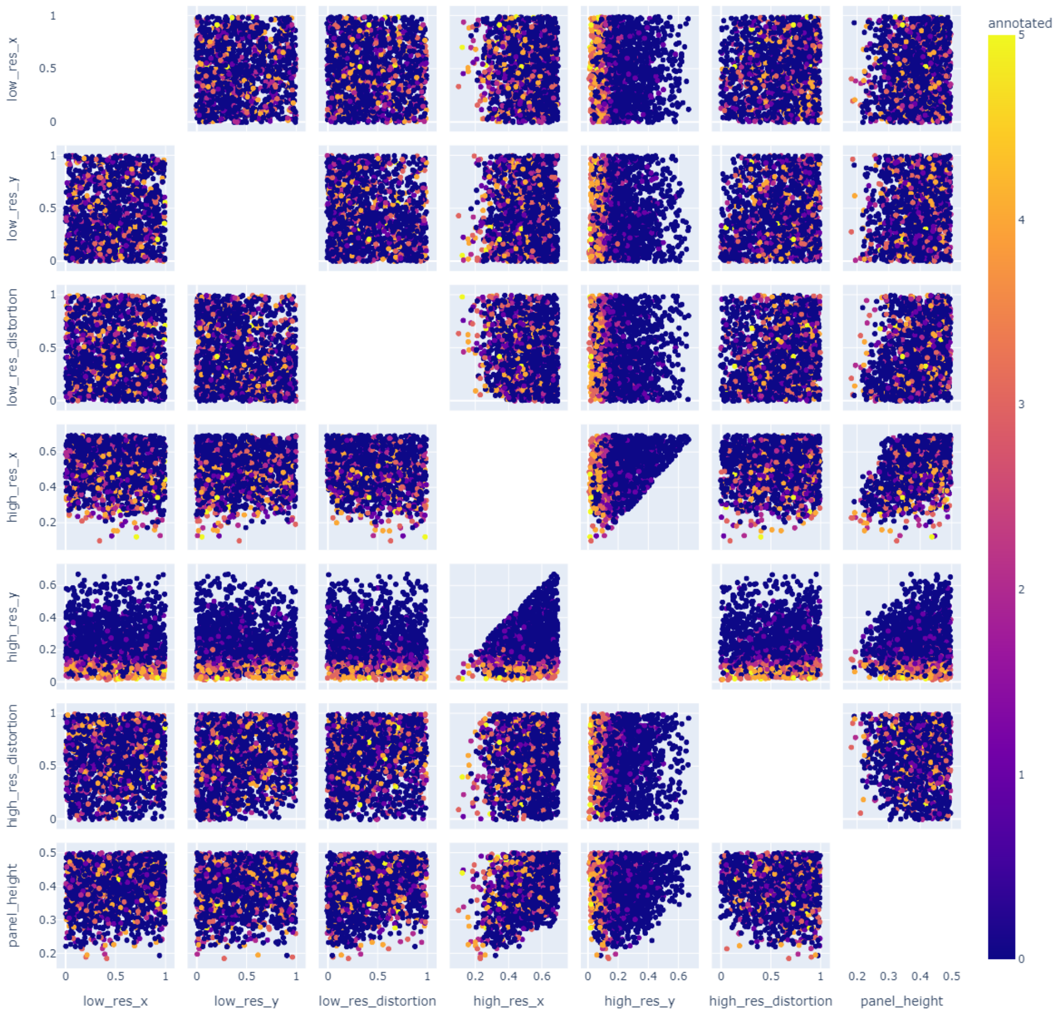
3D Interfaces

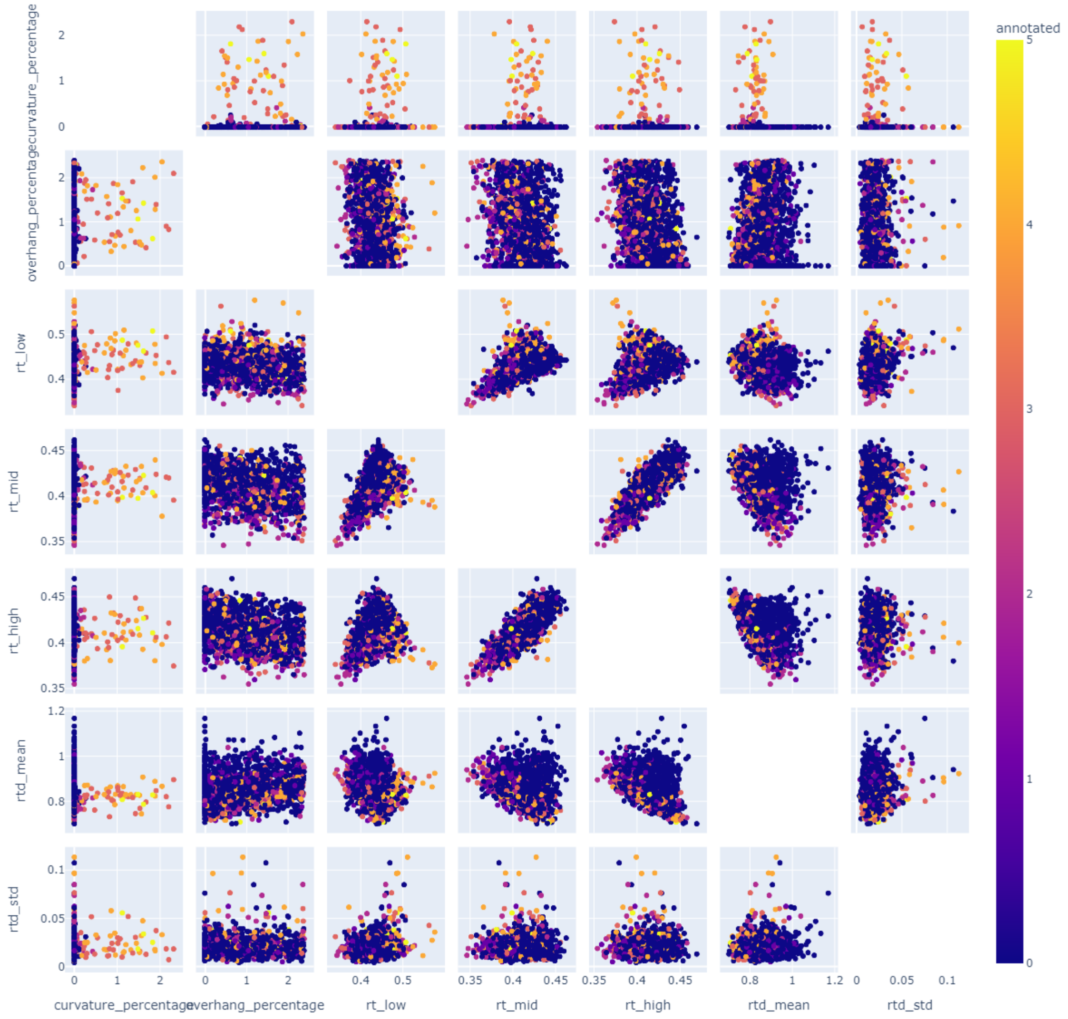


https://www.youtube.com/watch?v=cr7U1kRw-HE


1. Data representation
(passive)
(active)
2. Data annotation

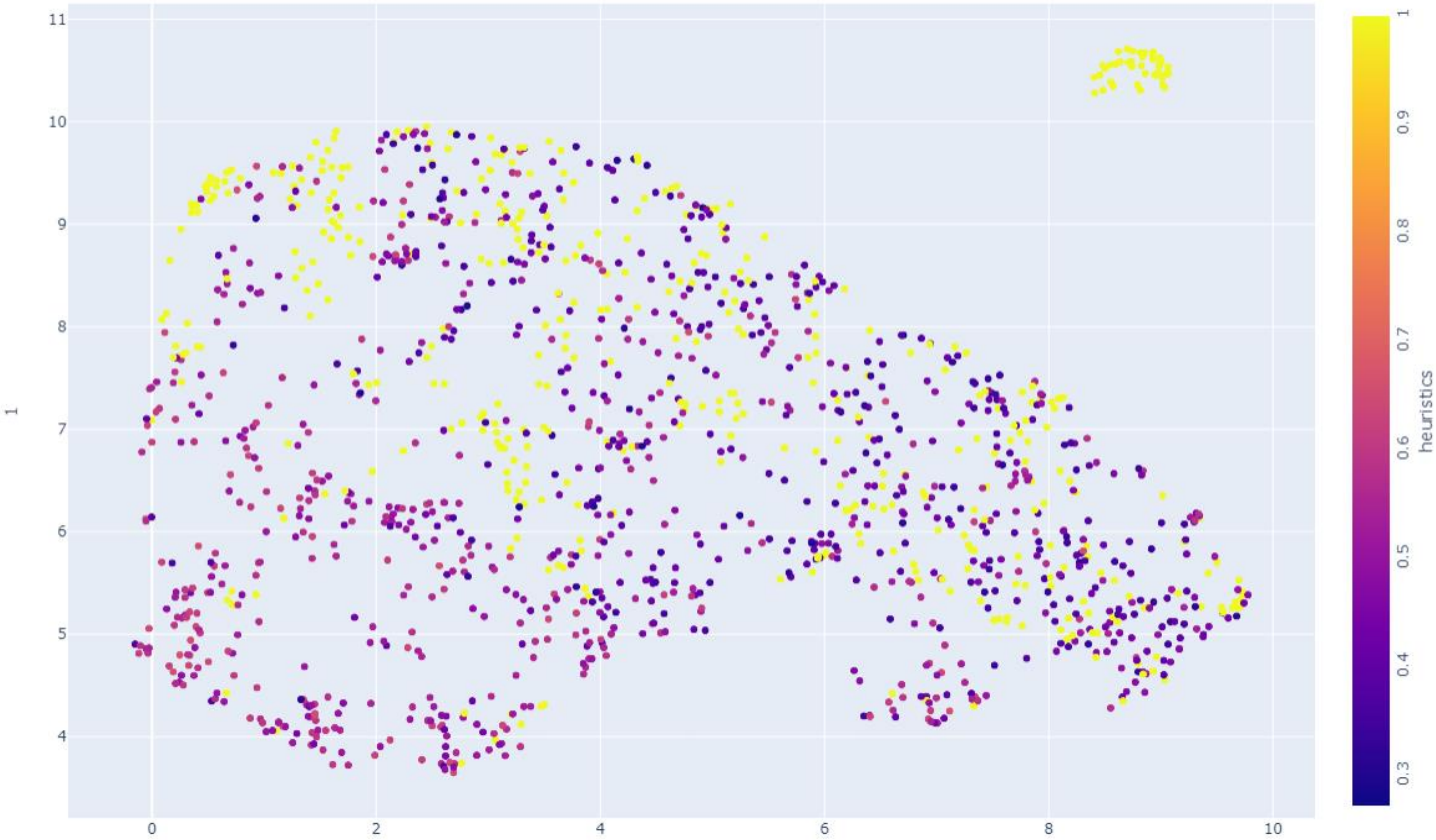
Acoustic Panel System, Gramazio Kohler Research, Immersive Design Lab, ETH Zurich, 2020-2021



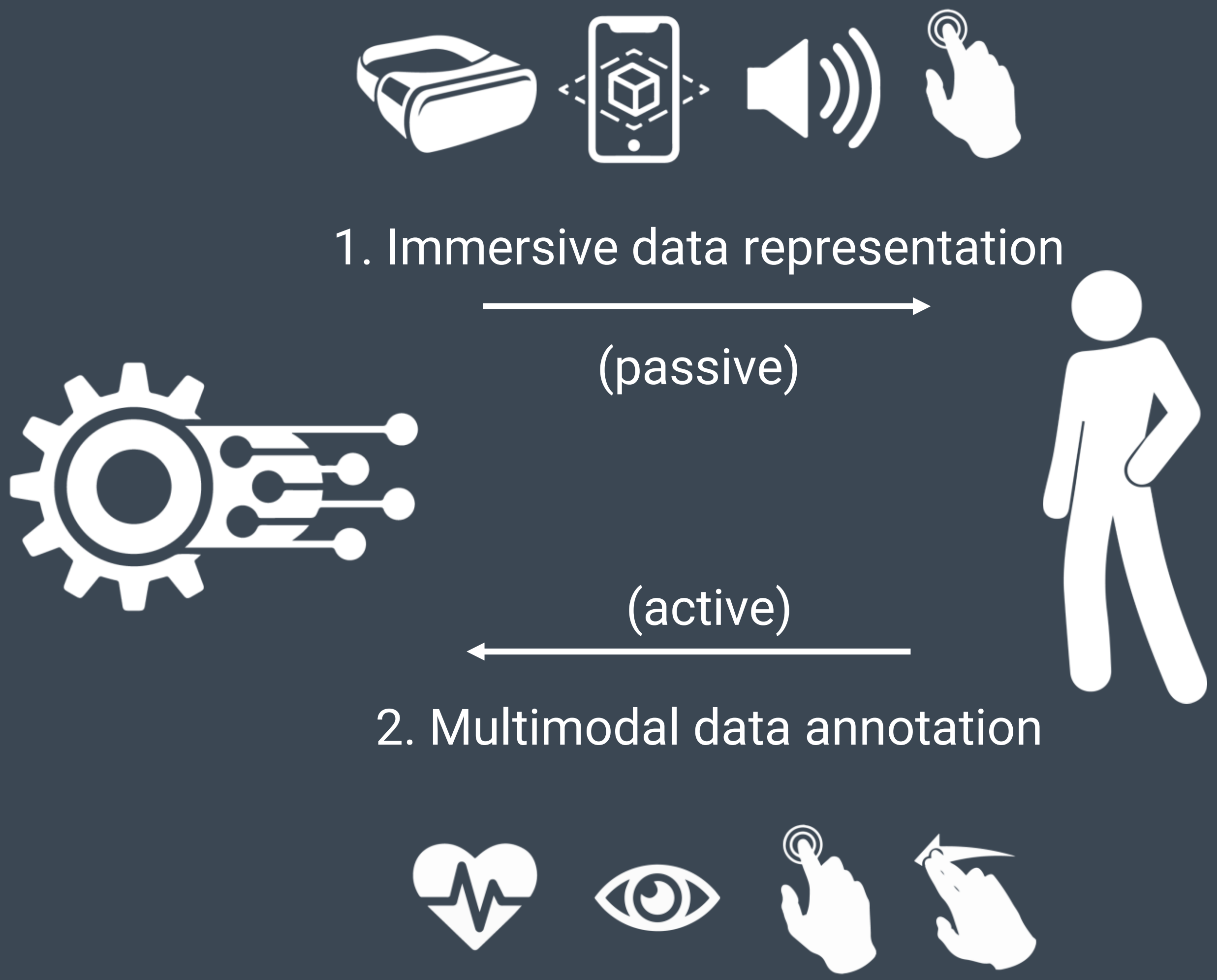


=> we like panels with low `high_res_y` and panels with high `curvature_percentage`

GRAMAZIO KOHLER RESEARCH ETH zürich SDSC



11
10
9
8
7
6
5
4
1
0
2
4
6
8
10
heuristics
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3


1. Immersive data representation
(passive)
(active)
2. Multimodal data annotation

Other resources and examples

- https://www.bloomberg.com/graphics/2015-whats-warming-the-world/?leadSource=uverify%20wall
- https://www.heavy.ai/demos
- https://www.youtube.com/watch?v=cr7U1kRw-HE
- https://medium.com/dataman-in-ai/plot-with-plotly-114ac106e25f
- https://dash.plotly.com/interactive-graphing
- https://www.tableau.com/learn/articles/interactive-map-and-data-visualization-examples
- http://manpopex.us/
- https://www.analyticsvidhya.com/blog/2021/10/interactive-plots-in-python-with-plotly-a-complete-guide/


FLOW

https://babiaxr.gitlab.io