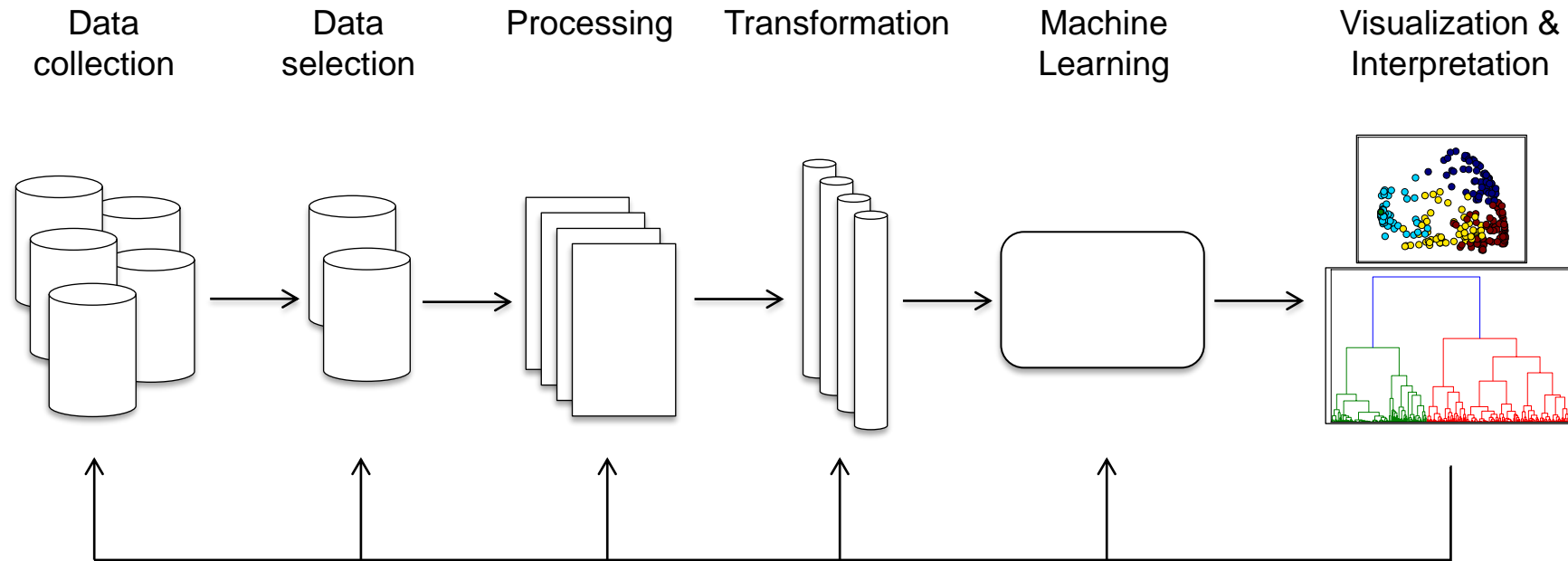


Lecture overview

- Student Project Pitches
- Data Processing and Visualization Lecture
 - Knowledge discovery process
 - Data selection
 - Data processing
 - Data visualization
- Data Processing and Visualization Exercise

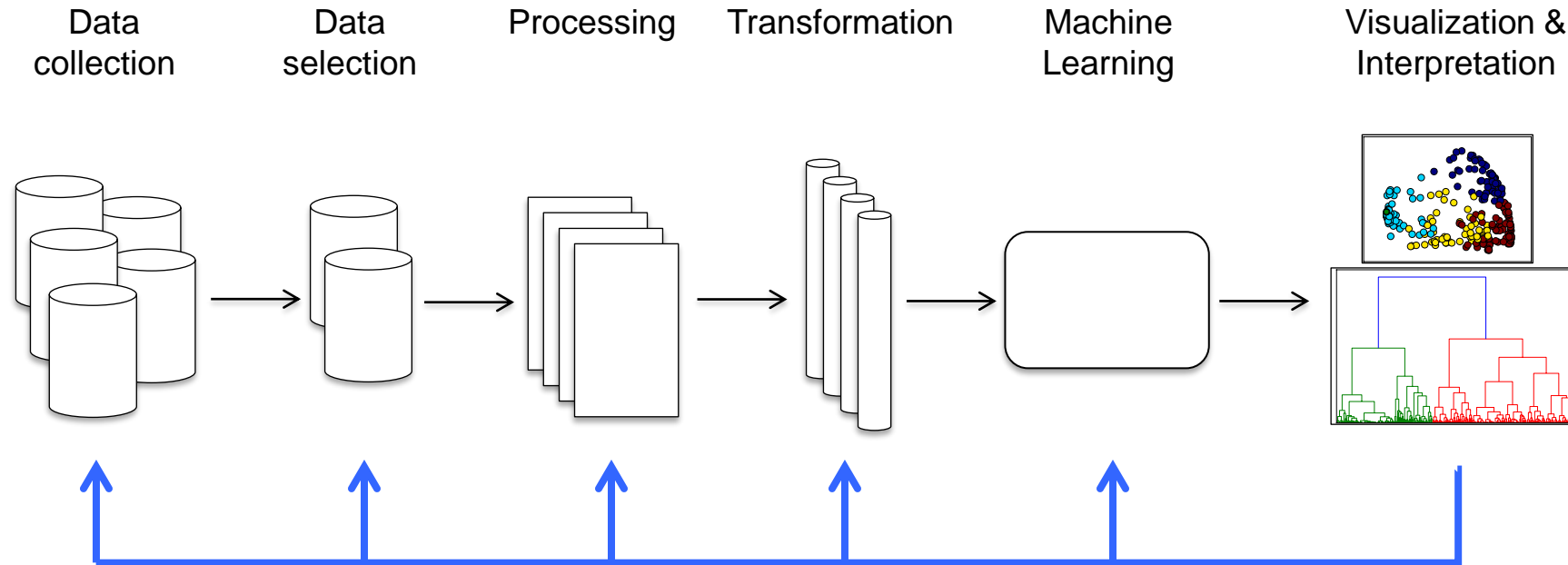
Knowledge discovery process



Typical Knowledge Discovery Diagram (KDD)

Knowledge discovery process

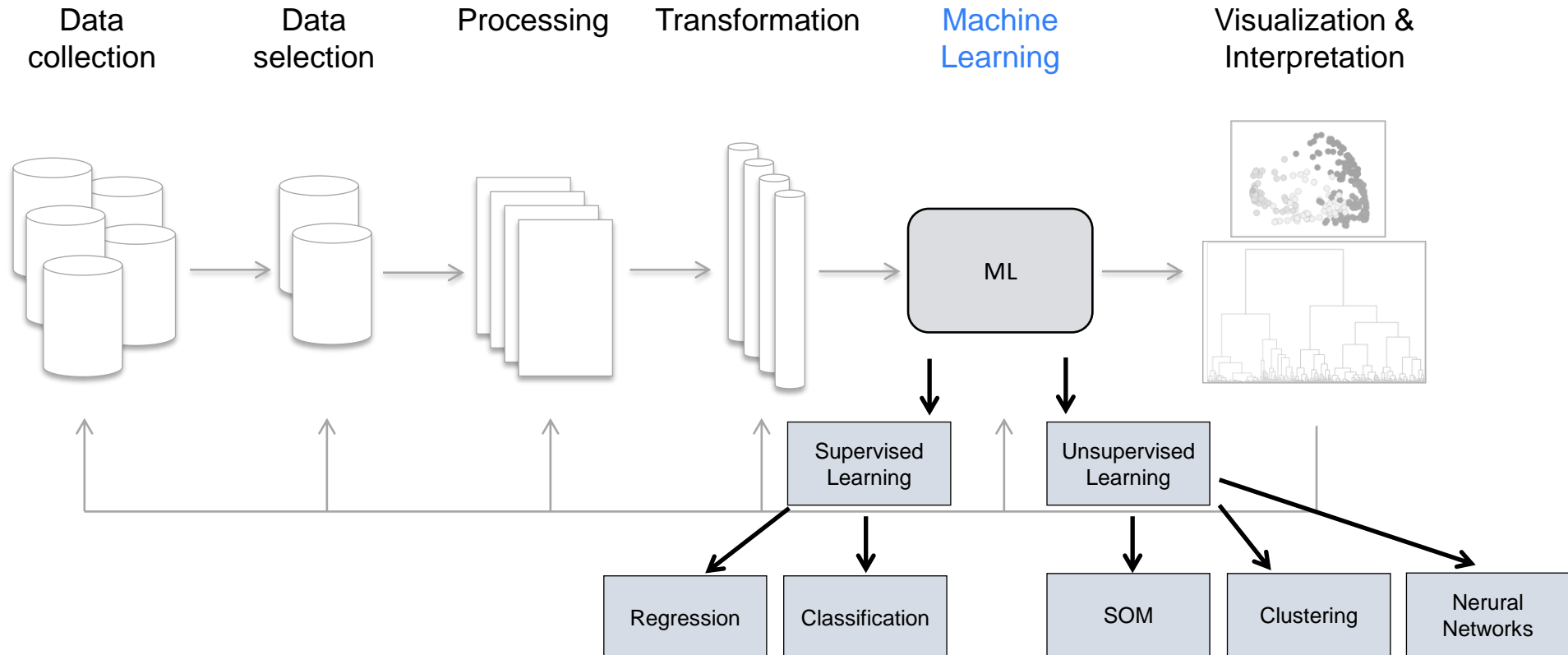
It is an exploratory and iterative process



Typical Knowledge Discovery Diagram (KDD)

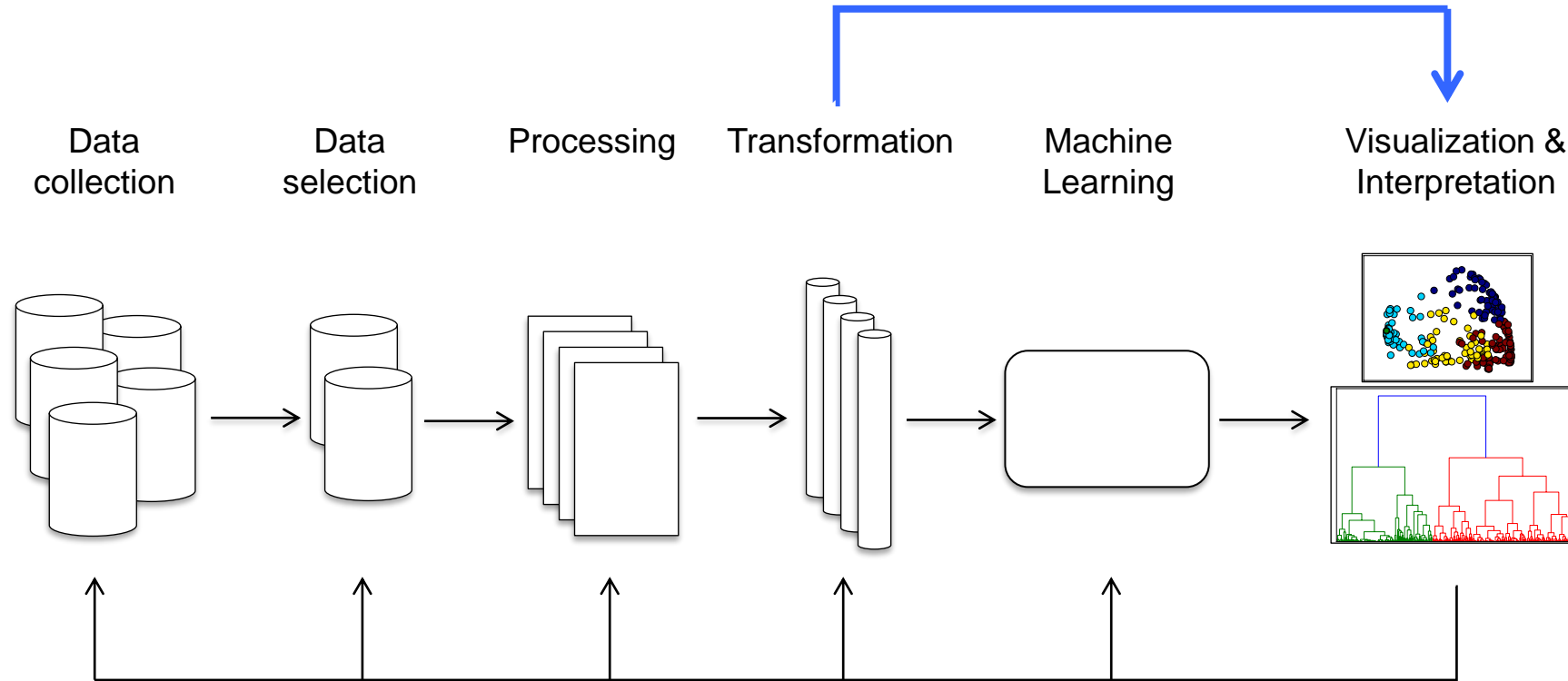
Knowledge discovery process

Where does machine learning fit into the process?



Knowledge discovery process

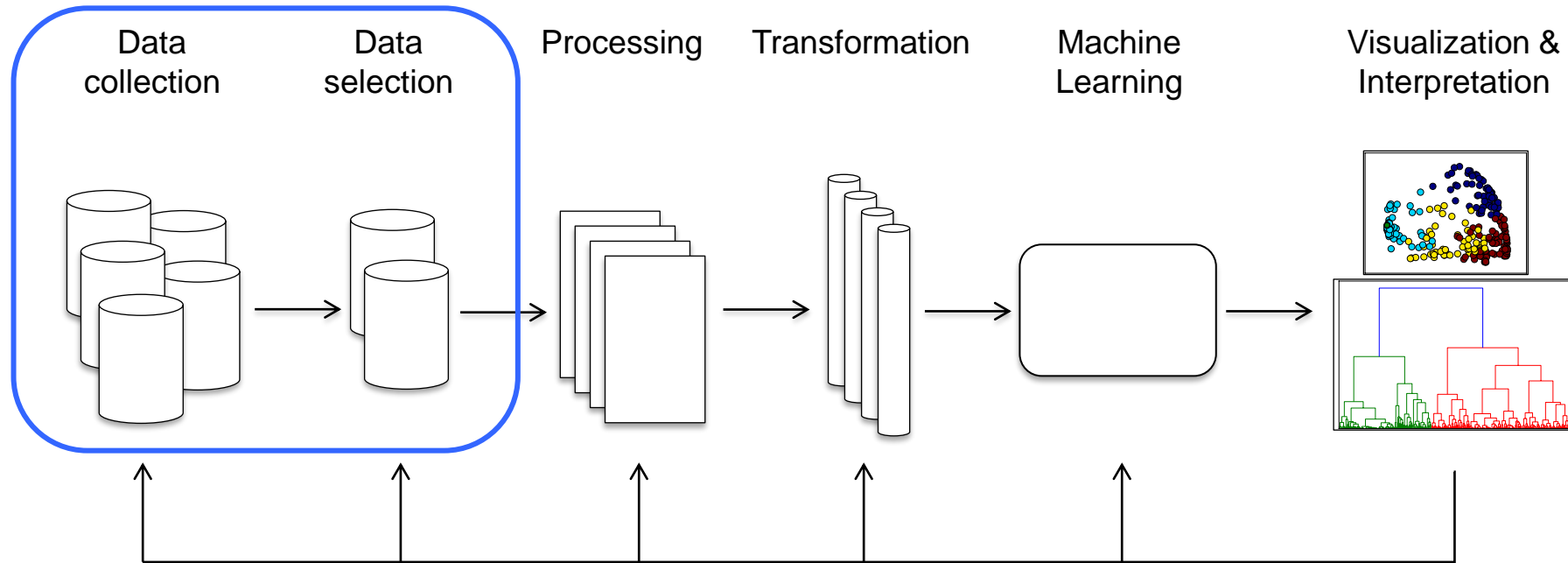
Data mining does not always include machine learning, for example time-series analysis and geo-referenced data visualization



Typical Knowledge Discovery Diagram (KDD)

Knowledge discovery process

Leverage expertise through domain
specific data source(s)



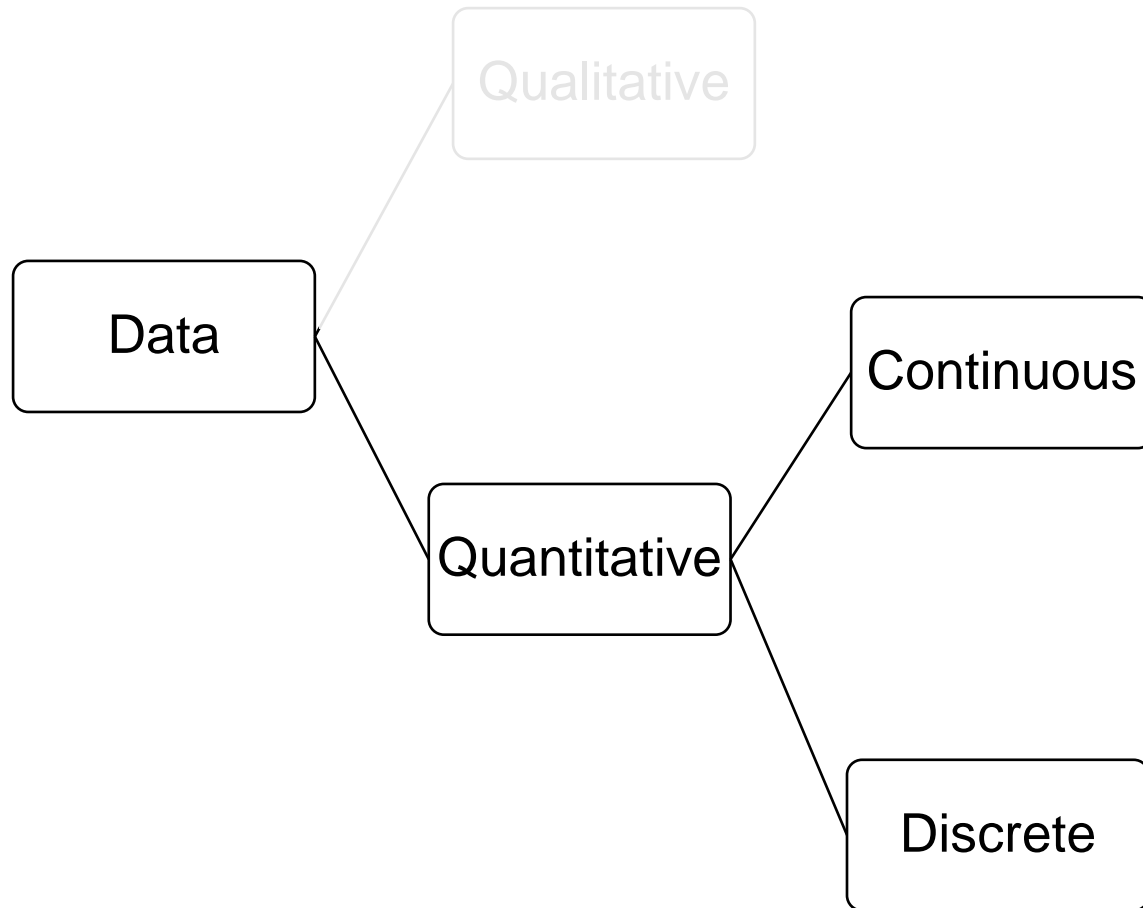
Typical Knowledge Discovery Diagram (KDD)

Data Structures

What is data in machine learning context?

- ML can process data in all imaginable ways like
 - Pictures, videos, Excel spreadsheets, SQL databases, ...
- When using machine learning, you will need to understand the basic data types to build your ML pipeline
 - Numerical data
 - Categorical data
 - Time series data
 - Text data

Numerical Data



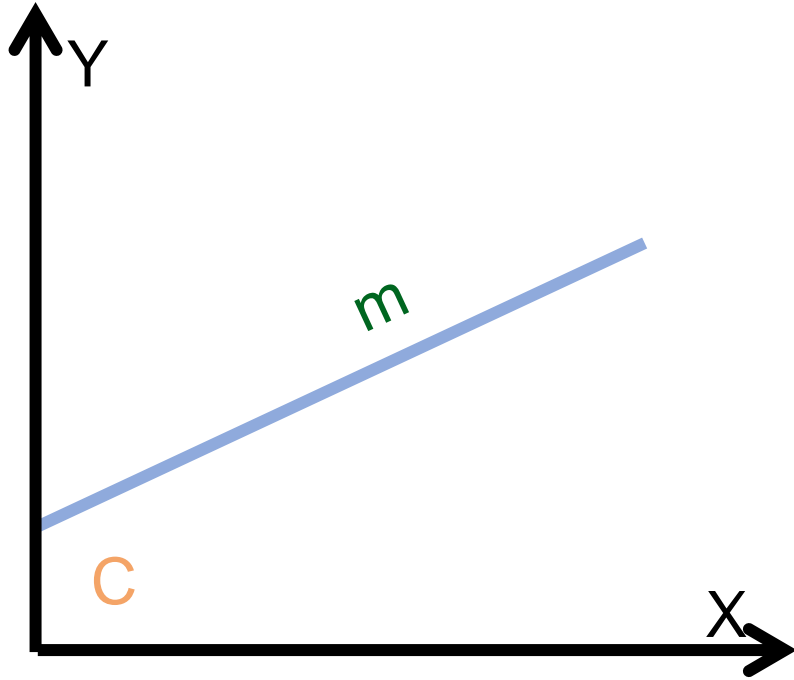
Continuous

- Always numeric
- Can be any number, positive or negative
- Something that can be measured, e.g., temperature

Discrete

- Ordinal variable – Survey ratings (0 – 5)
- Binary variables – (0 /1)
- Something that can be counted, e.g., number of students

Numerical Data

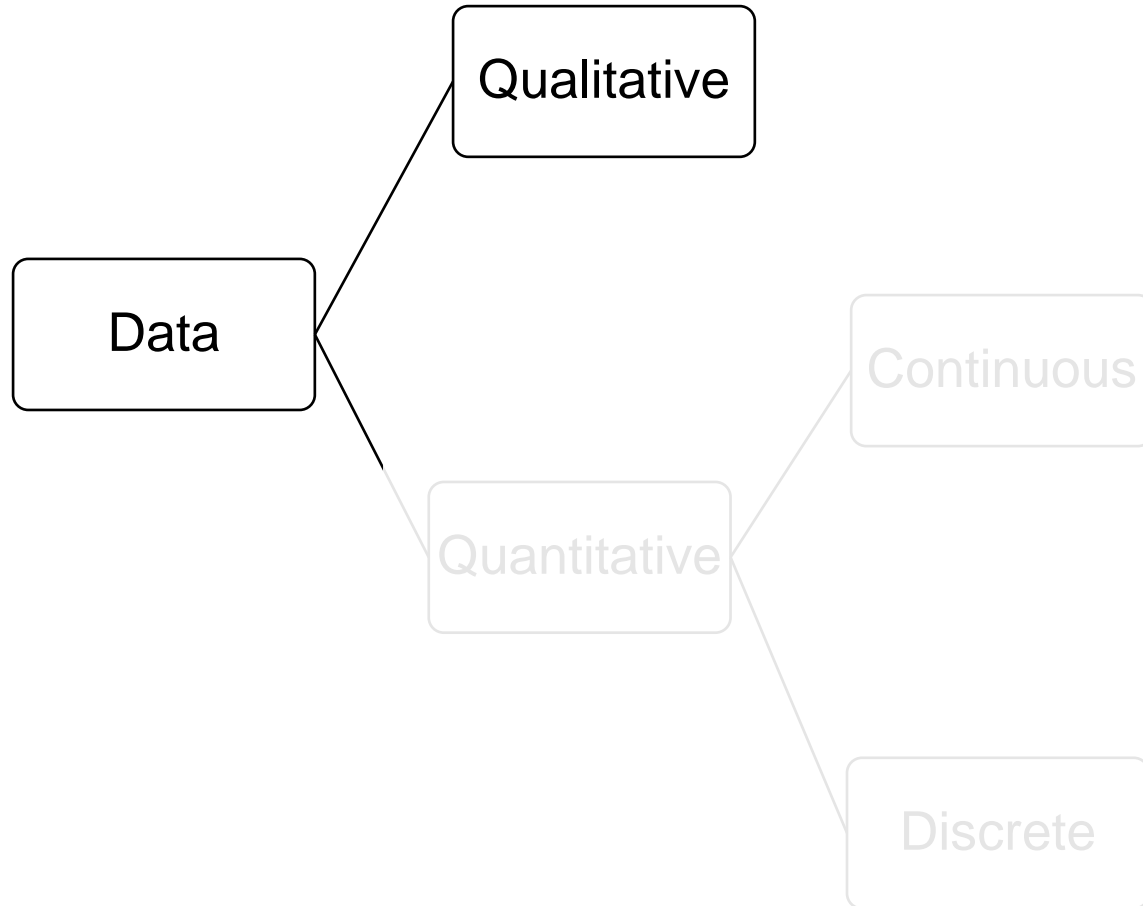


Independent / Dependent

e.g., $Y = mX + C$

Value of Y depended on m , X , and C

Categorical Data



Qualitative data

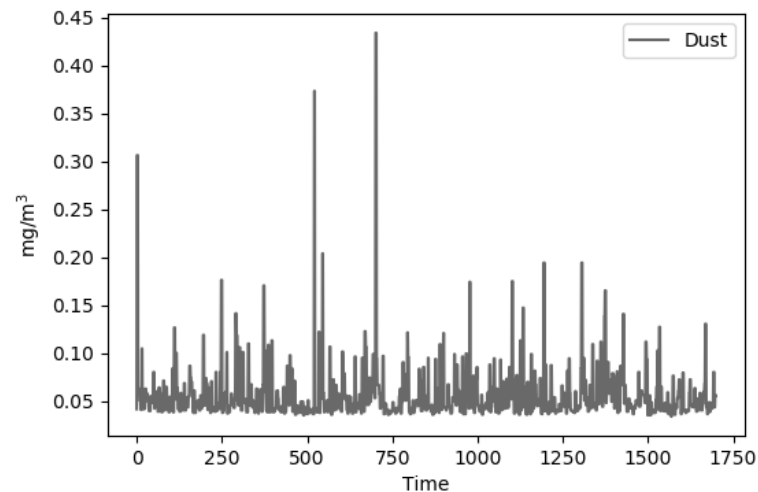
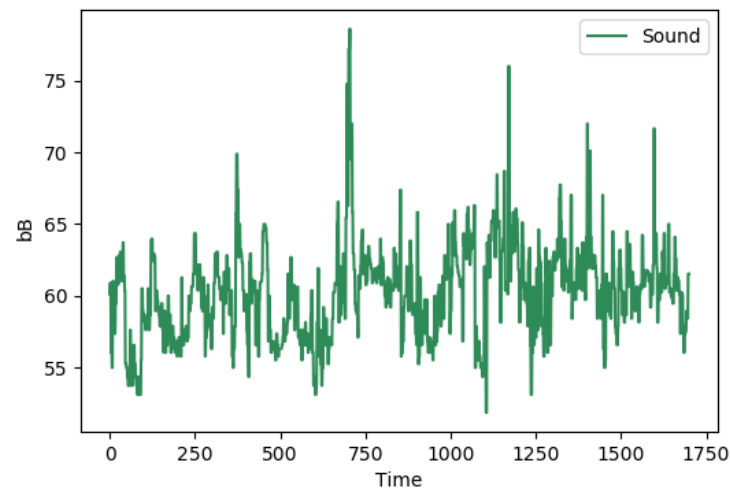
- Subjective ratings - excellent, good, fair, poor
- Meta data – gender (male, female)
- Categorical data may derive from observations made of qualitative data
- Observations of quantitative data grouped within given intervals



https://image.freepik.com/free-vector/temperature-measurement-from-cold-hot_53562-2741.jpg

Time Series Data

- Time series data is a sequence of numbers collected at regular intervals over some period of time
- Is a sequence taken at successive equally spaced points in time, thus it is a sequence of discrete-time data
- Can be applied to real-valued, continuous data, discrete numeric data, or discrete symbolic data

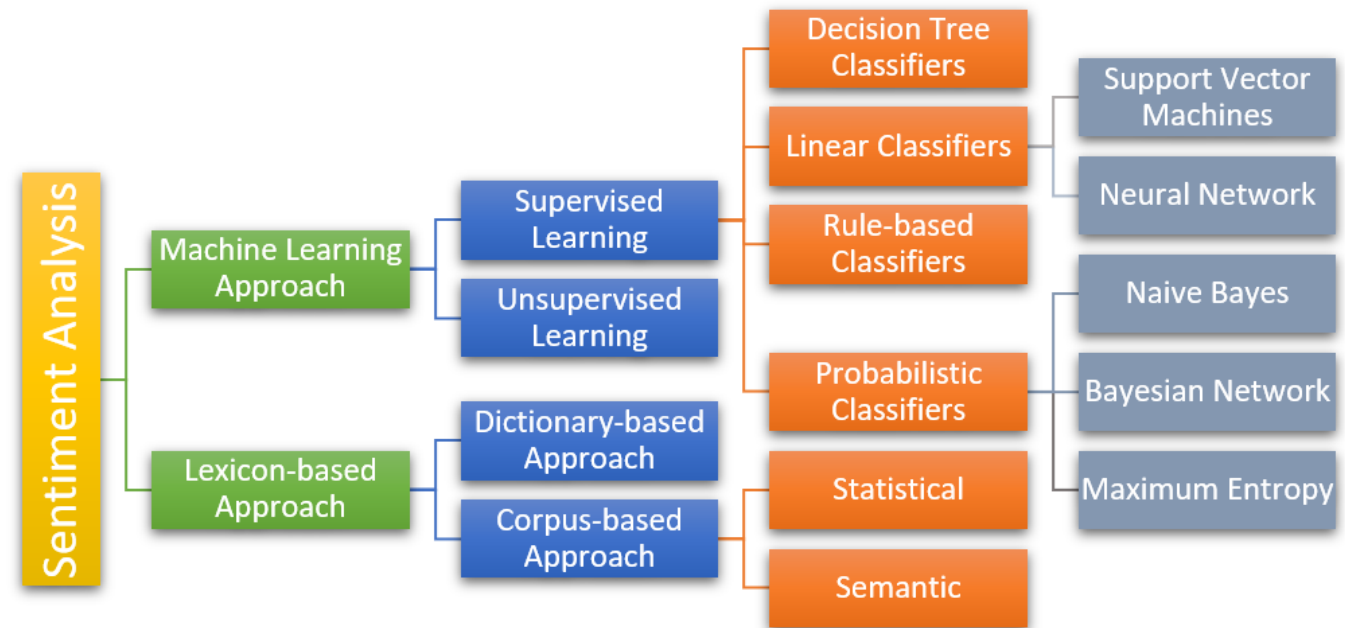


Ojha VK, Griego D, Kuliga S, Bielik M, Buš P, Schaeben C, Treyer L, Standfest M, Schneider S, König R, Donath D, Schmitt G (2018) Machine learning approaches to understand the influence of urban environments on human's physiological response, *Information Sciences*, Elsevier ([pdf](https://archive.arch.ethz.ch/esum/data.html)). <https://archive.arch.ethz.ch/esum/data.html>

Text Data

Text Mining:

- Process of deriving high-quality information from text
- Automation of extracting information of unknown text: websites, books, emails...
- Text analysis processes are
 - Sentiment analysis
 - Information retrieval
 - Dimensionality reduction



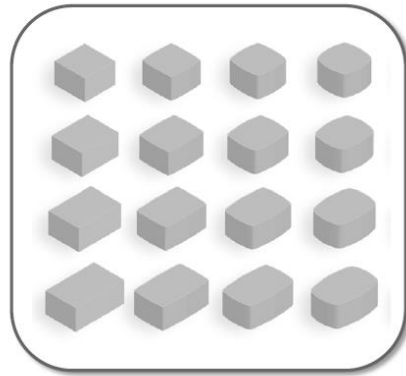
<https://devopedia.org/images/article/105/8215.1532752754.png>

Acquiring data for machine learning?

Limitations of collected data in AEC



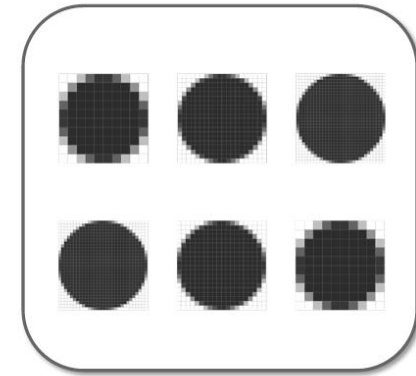
NOT ENOUGH



NOT DIVERSE



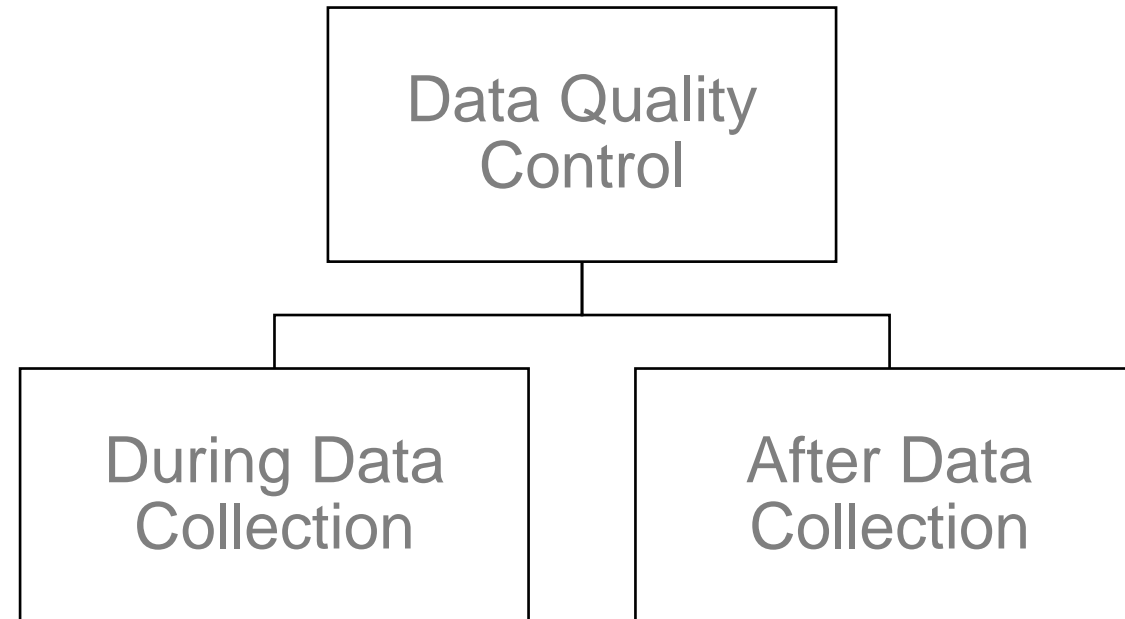
INCOMPLETE



INCOSISTENT
QUALITY & FORMAT

Data Quality Control

Improving data quality



Data Collection







Improving data quality

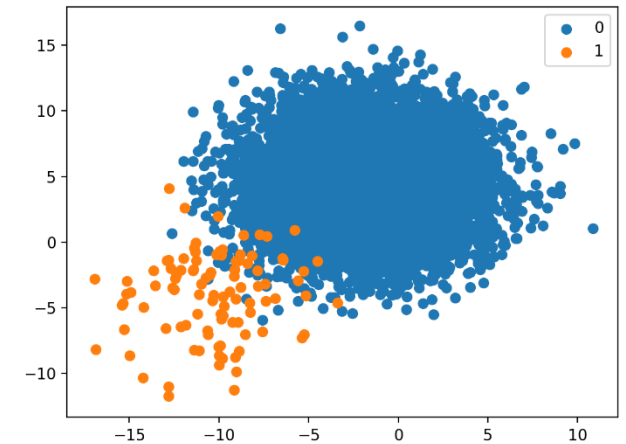
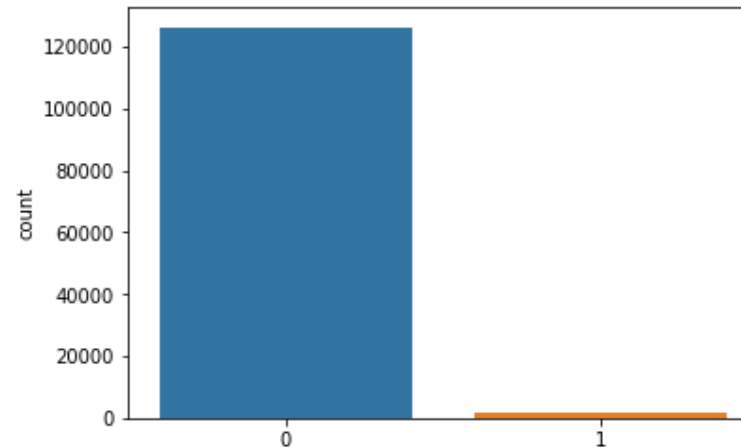


Reference: Mónica Bobrowski, Martina Marré, Daniel Yankelevich, Measuring Data Quality, Report no.: 99-002, Pabellón 1 - Planta Baja - Ciudad Universitaria

Properties of a good dataset

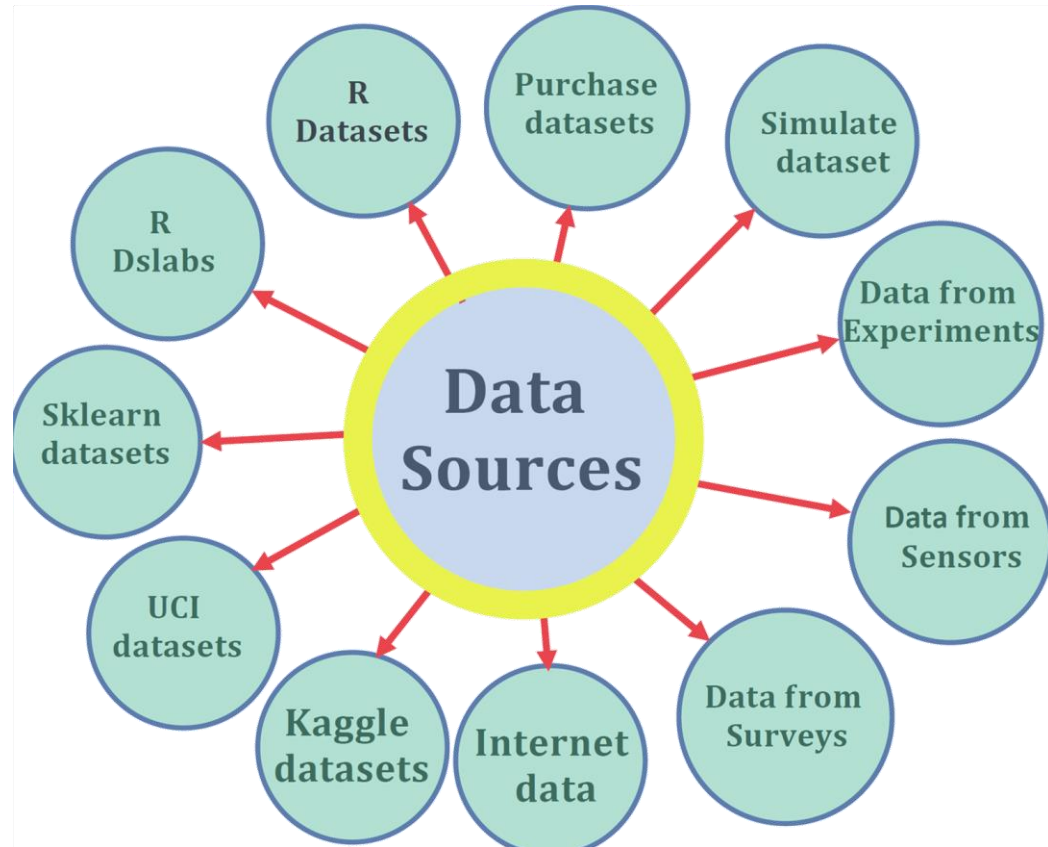
Koch et al. (2019), ABC: A Big CAD Model Dataset For Geometric Deep Learning

	LARGE SIZE
	GROUND TRUTH LABELS
	PARAMETRIC REPRESENTATION
	EXPANDABLE
	VARIATION
	BALANCED



<https://3qepr26caki16dnhd19sv6by6v-wpengine.netdna-ssl.com/wp-content/uploads/2019/10/Scatter-Plot-of-Binary-Classification-Dataset-With-1-to-100-Class-Distribution.png>

Data Sources



https://miro.medium.com/max/3978/1*yPcYNnAVxcRWSDQlgKDUxg.png



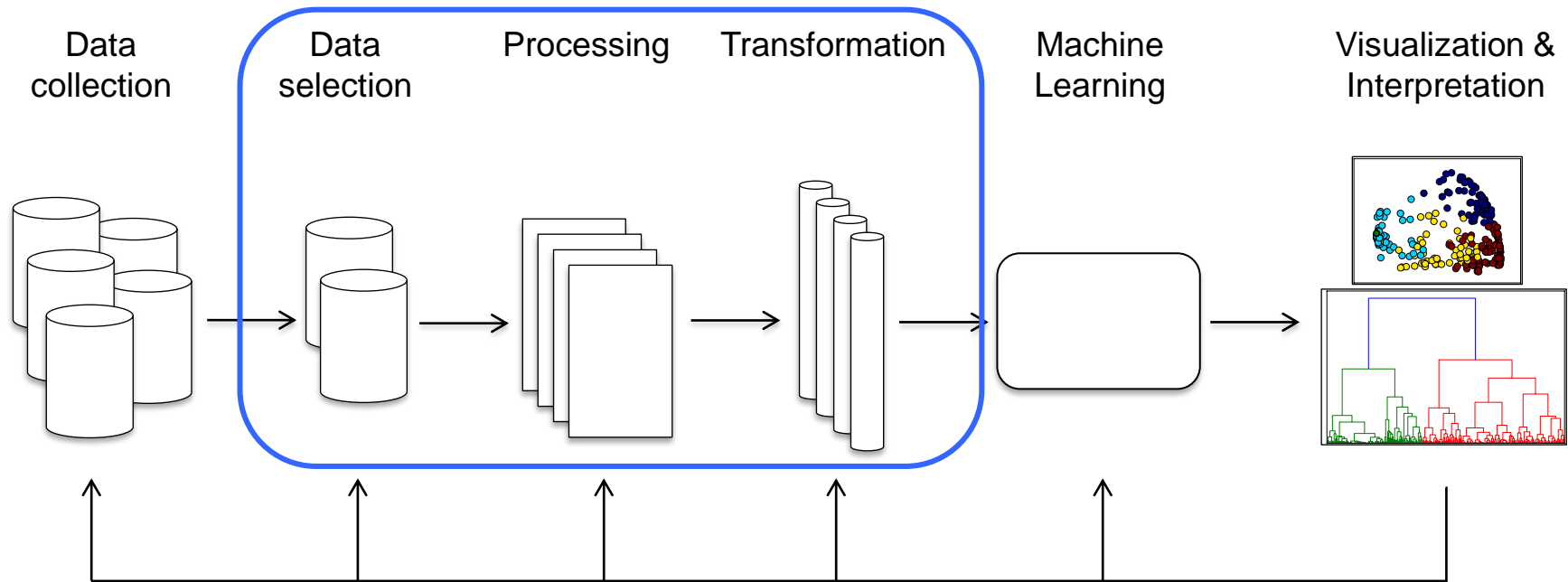
Floor plans from the CVC-FP dataset <http://dag.cvc.uab.es/resources/floorplans/>



Fusion 360 Gallery Dataset - <https://github.com/AutodeskAILab/Fusion360GalleryDataset>

Knowledge discovery process

Is the data usable? The not-so fun, but essential part of the process.



Typical Knowledge Discovery Diagram (KDD)

The not-so fun, but essential part of the process

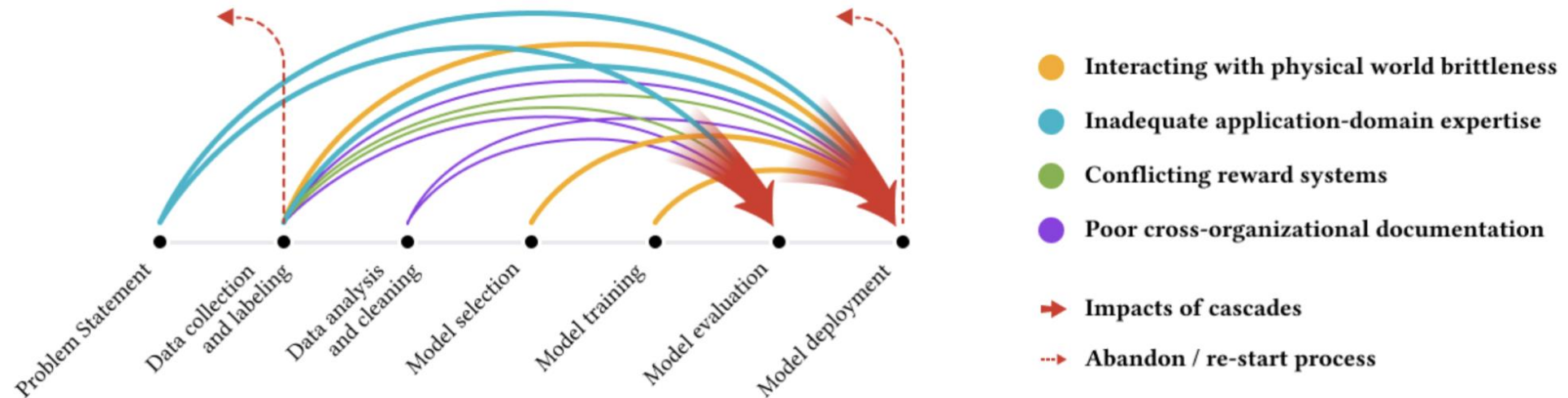
“Everyone wants to do the model work, not the data work”: **Data Cascades in High-Stakes AI**

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora

Aroyo

[nithyasamba,kapania,hhighfill,dakrong,pkp,loraa]@google.com

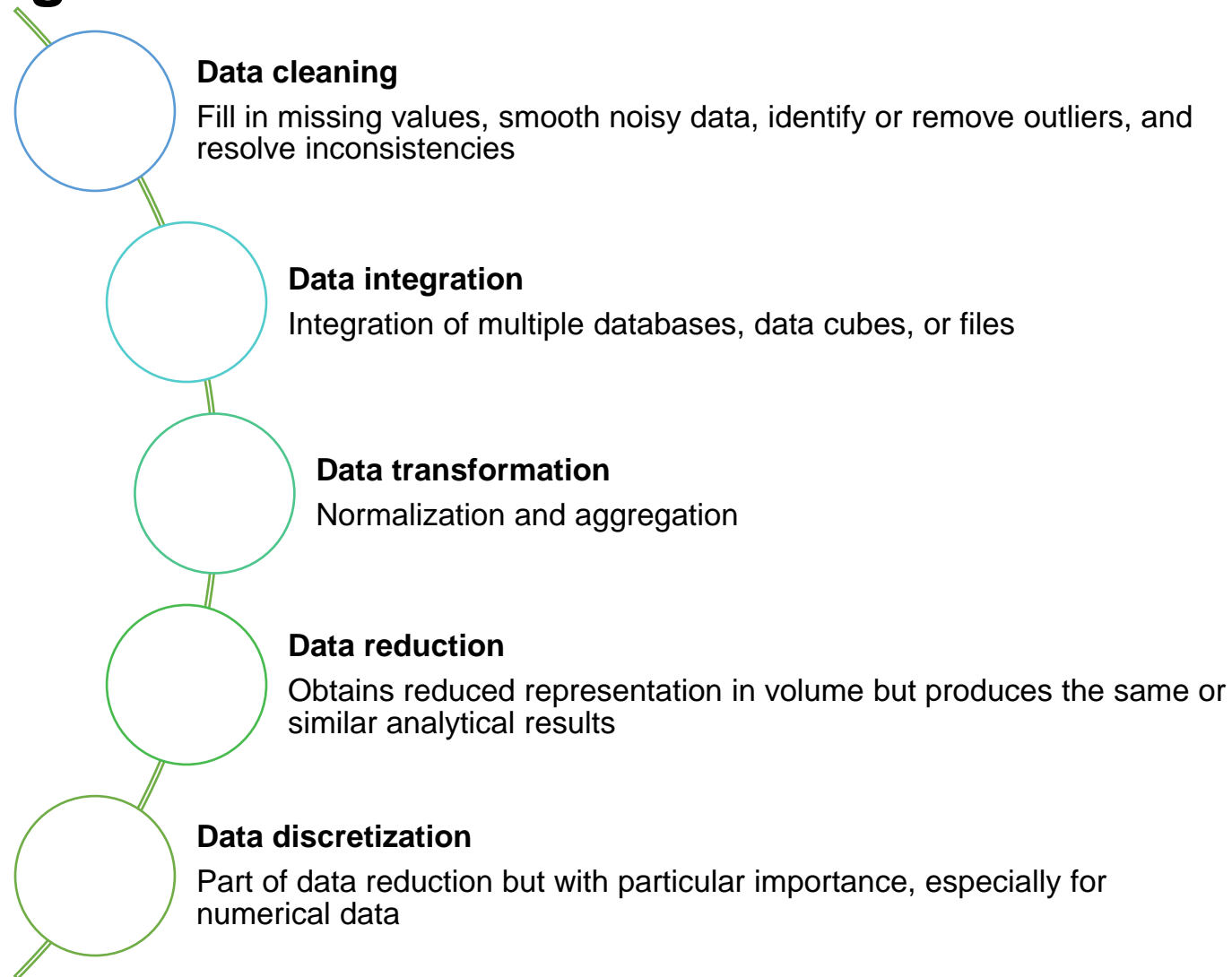
Google Research
Mountain View, CA



<https://ai.googleblog.com/2021/06/data-cascades-in-machine-learning.html>

Data Processing

Improving Data Quality



Reference: <http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf>

Data Processing

Data Cleaning

Samples	Features (Variables)				Output feature (dependent variable)
	Input features (Independent Variables)				
	Input V1	Input V2	Input V3	Input V4	
Sample 1	2.3	0.25	Good	1.5	y_1
Sample 2	4.5	43598.21	Good	1.8	y_2
Sample 3	4.7	0.33	Excellent	1.9	y_3
Sample 4	?	0.22	Good	3.9	y_4
Sample 5	?	0.19	Average	1.2	y_5
:	6.7	0.88	Good	1.8	:
Sample N	5.5	0.36	Bad	1.6	y_N



Missing Value



Outlier



Noise (required smoothing)

y_i

Output variable can be continuous/discrete

Data Processing

Data Cleaning

Considered physiological signals

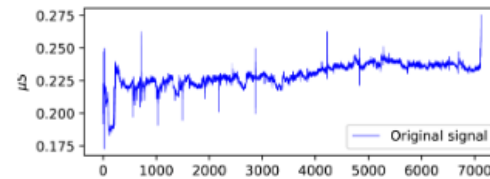


Fig Type 1

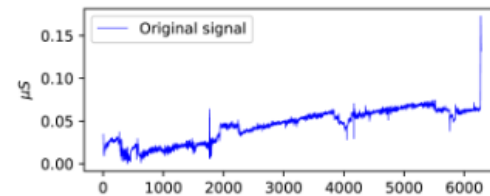


Fig Type 2

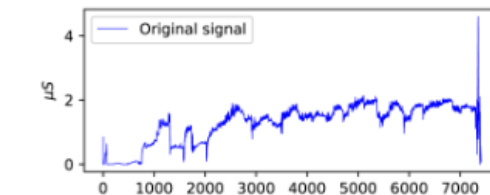


Fig Type 3

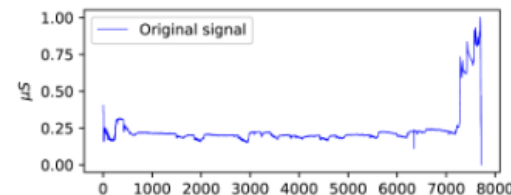


Fig Type 4

Discarded physiological signals

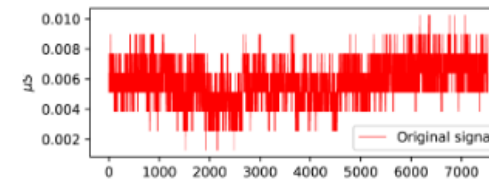


Fig Type 1: step function like signal

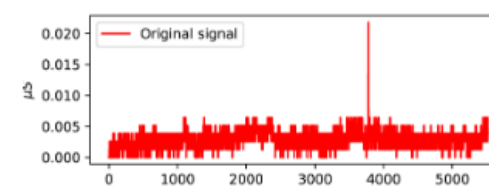


Fig Type 2: step function with major sensor loss

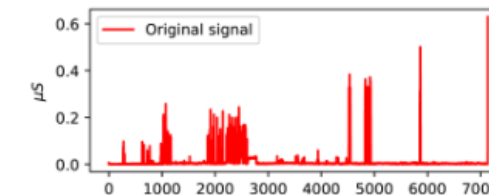


Fig Type 3: major sensor loss

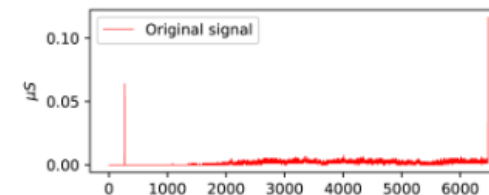
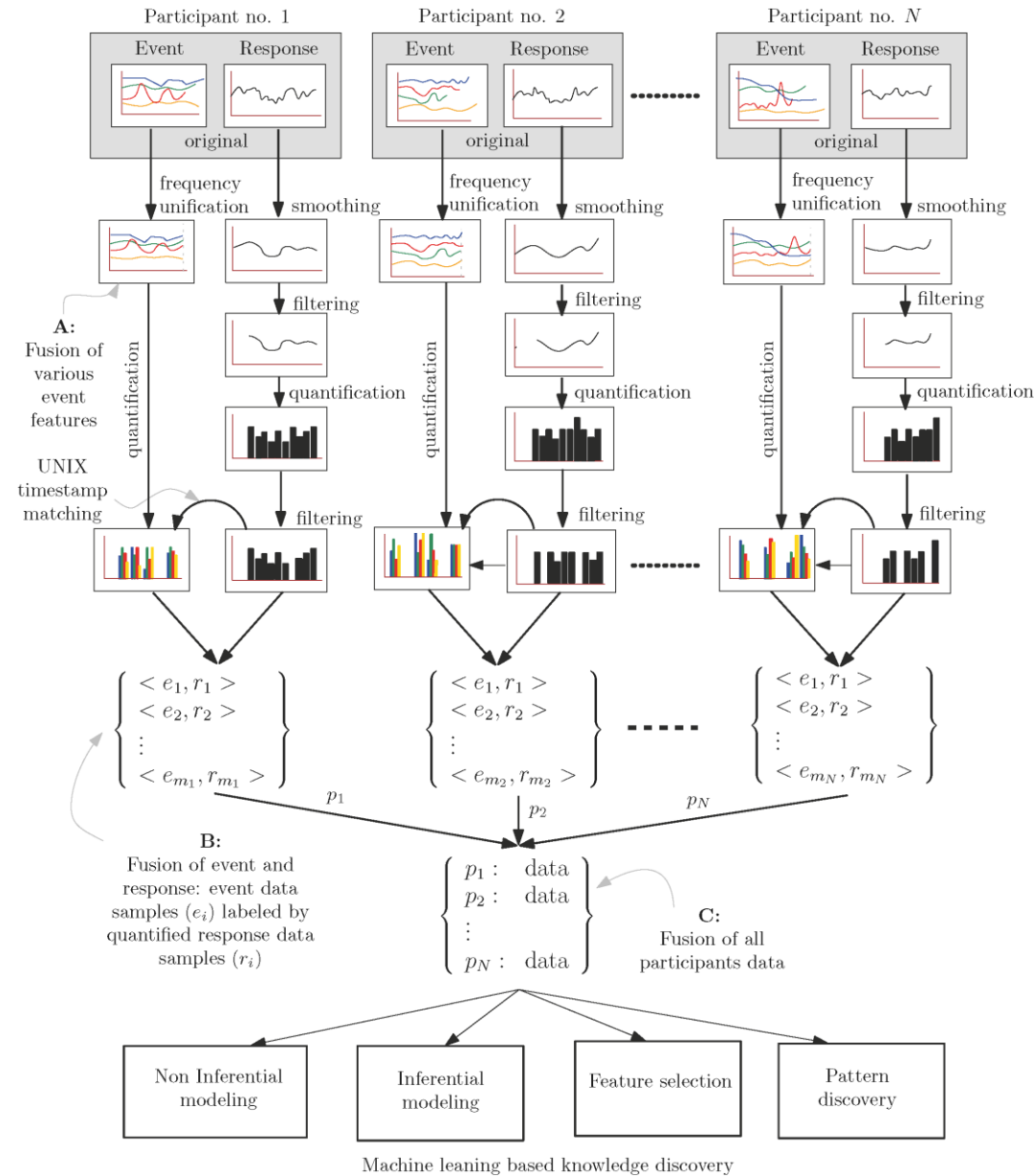


Fig Type 4: insignificant observations

Ojha VK, Griego D, Kuliga S, Bielik M, Buš P, Schaeben C, Treyer L, Standfest M, Schneider S, König R, Donath D, Schmitt G (2018) Machine learning approaches to understand the influence of urban environments on human's physiological response, *Information Sciences*, Elsevier ([pdf](https://archive.arch.ethz.ch/esum/data.html)) <https://archive.arch.ethz.ch/esum/data.html>

Data Processing

Data Integration

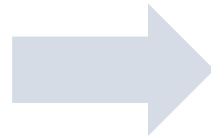


Ojha VK, Griego D, Kuliga S, Bielik M, Buš P, Schaeben C, Treyer L, Standfest M, Schneider S, König R, Donath D, Schmitt G (2018) Machine learning approaches to understand the influence of urban environments on human's physiological response, *Information Sciences*, Elsevier
[pdf\) https://archive.arch.ethz.ch/esum/data.html](https://archive.arch.ethz.ch/esum/data.html)

Data Processing

Data Transformation

Samples	Features (Variables)		
	Input features (Independent Variables)		Output feature (dependent variable)
	Input V1	Input V2	Output V1
Sample 1	2.3	0.25	y_1
Sample 2	4.5	0.39	y_2
Sample 3	4.7	0.33	y_3
Sample 4	2.99	0.22	y_4
Sample 5	3.18	0.19	y_5
:	6.7	0.36	:
Sample N	5.5	0.88	y_N



Samples	Features (Variables)		
	Input features (Independent Variables)		Output feature (dependent variable)
	Input V1	Input V2	Output V1
Sample 1	0.00	0.09	y_1
Sample 2	0.50	0.29	y_2
Sample 3	0.55	0.20	y_3
Sample 4	0.16	0.04	y_4
Sample 5	0.20	0.00	y_5
:	1.00	0.25	:
Sample N	0.73	1.00	y_N

	Max value
	Min value
	Transformed Variable


Data Processing

Data Reduction

Features (Variables)				
Samples	Input features			Output feature
	Input V1	Input V2	Input V4	Output V1
Sample 1	2.3	0.25	1.5	y_1
Sample 2	4.5	0.39	1.8	y_2
Sample 3	4.7	0.33	1.9	y_3
Sample 4	2.99	0.22	1.6	y_4
Sample 5	3.18	0.19	1.2	y_5
:	6.7	0.88	1.8	:
Sample N	5.5	0.36	1.6	y_N

Feature Extraction

Feature Selection

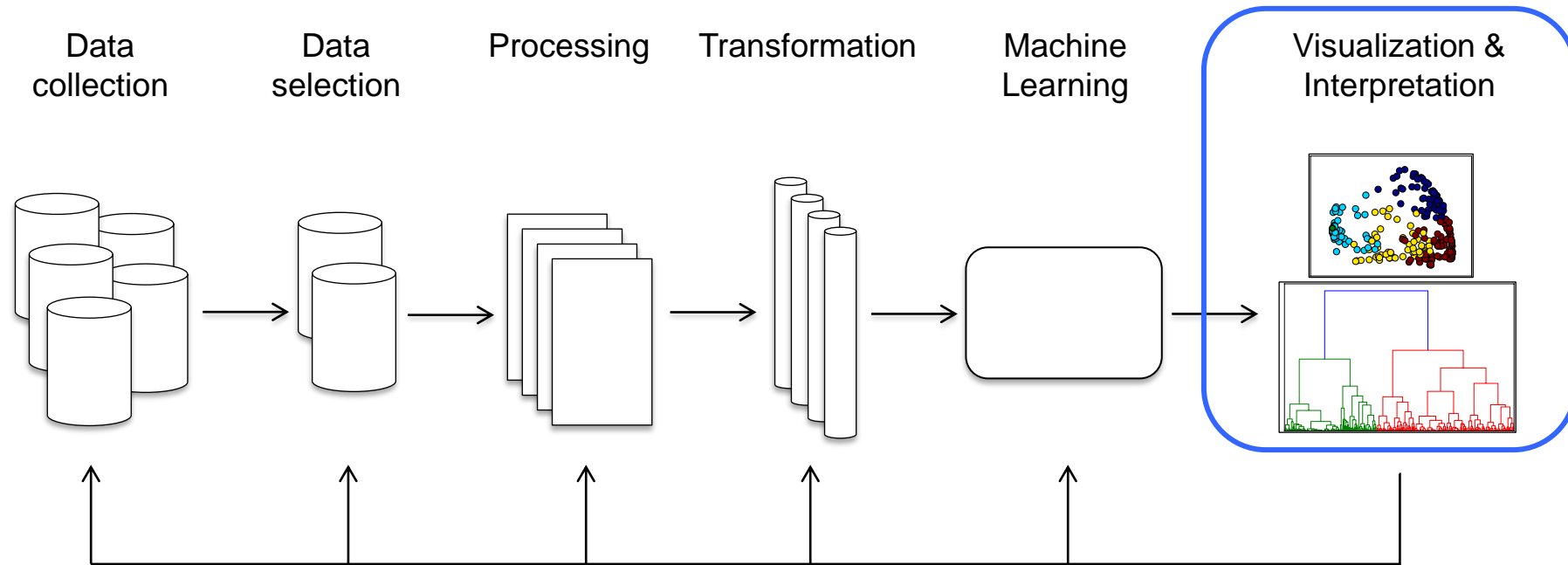
 Extracted Features

Samples	Features (Variables)			
	Input features			Output feature
	PCA 1	PCA 2		Output V1
Sample 1	-1.97	0.06		y_1
Sample 2	0.25	0.15		y_2
Sample 3	0.45	0.22		y_3
Sample 4	-1.28	0.09		y_4
Sample 5	-1.14	-0.33		y_5
:	2.48	-0.04		:
Sample N	1.21	-0.15		y_N

Samples	Features (Variables)			
	Input features			Output feature
	Input V1		Input V4	Output V1
Sample 1	2.3		1.5	y_1
Sample 2	4.5		1.8	y_2
Sample 3	4.7		1.9	y_3
Sample 4	2.99		1.6	y_4
Sample 5	3.18		1.2	y_5
:	6.7		1.8	:
Sample N	5.5		1.6	y_N

Knowledge discovery process

The visualizations tell the final story.
What do we want to know?



Typical Knowledge Discovery Diagram (KDD)

Data Visualization

Good data visualization helps to:

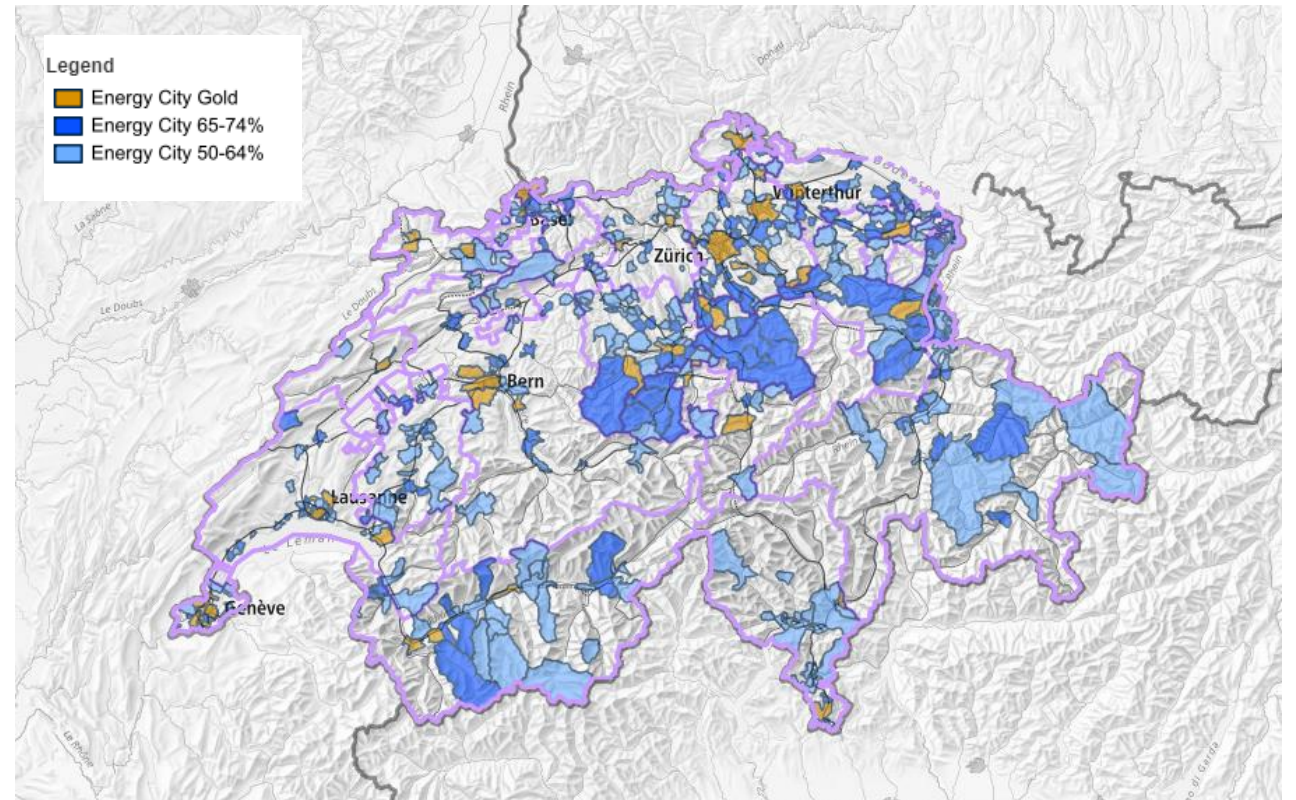
- make information easy to read and retain
- identify trends and patterns
- prove theories and answer questions
- control the focus and capture the audience's attention
- improves the impact of your message



Definitions- What is information visualization?

Information visualization represents data or **information that is already somewhat understood**

Visual representations of abstract data to **reinforce human cognition**

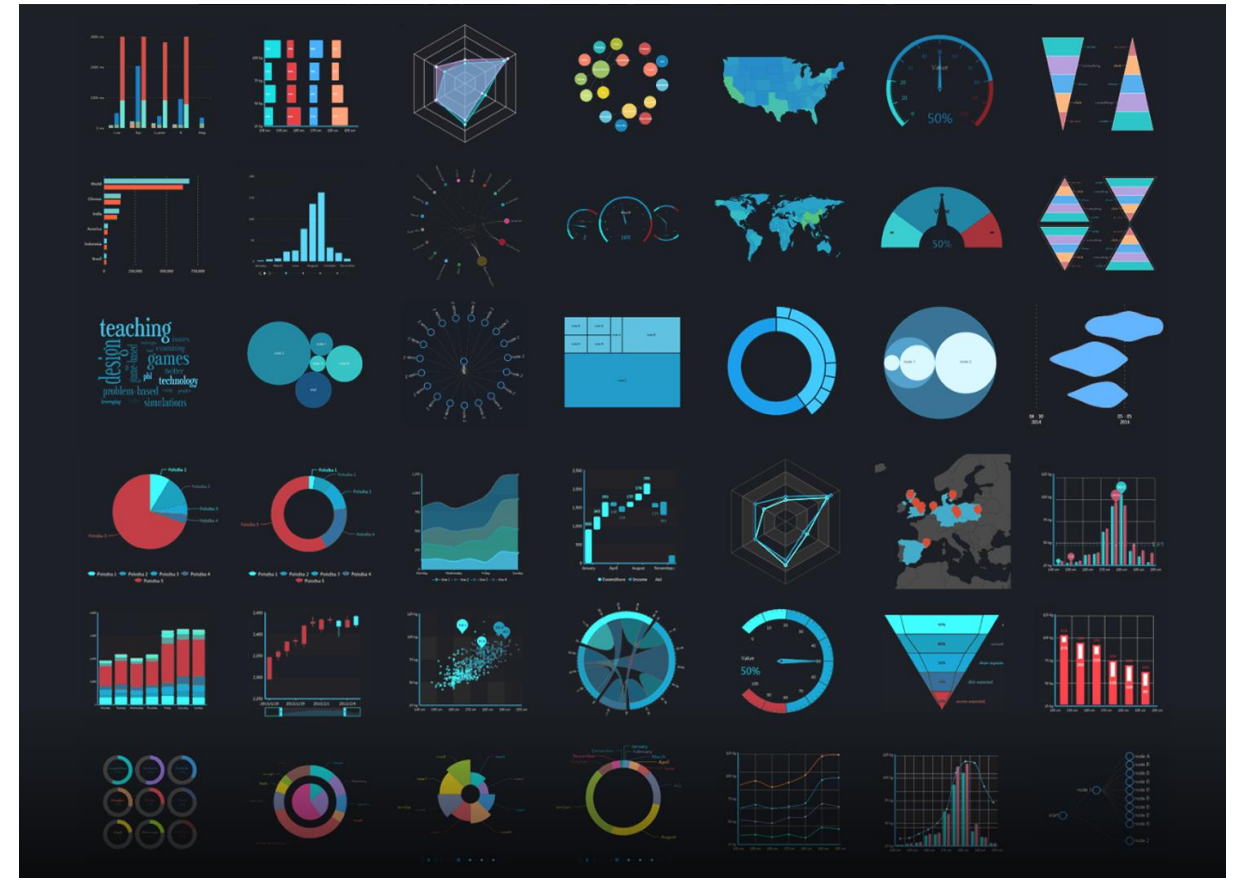


Map of the Energy Cities in Switzerland
<https://s.geo.admin.ch/8c096ea987>

Definitions- What is data visualization?

Data visualization helps researchers find patterns and relationships from data by presenting information in a **clear, efficient and meaningful way**

It involves an **abstracted representation** of raw data, in form of graphs, charts and drawings, in order to **enhance comprehension** and **direct the focus**



Source: <https://www.theguardian.com/guardian-masterclasses/2015/aug/07/data-visualisation-a-one-day-workshop-tobias-sturt-adam-frost-digital-course>

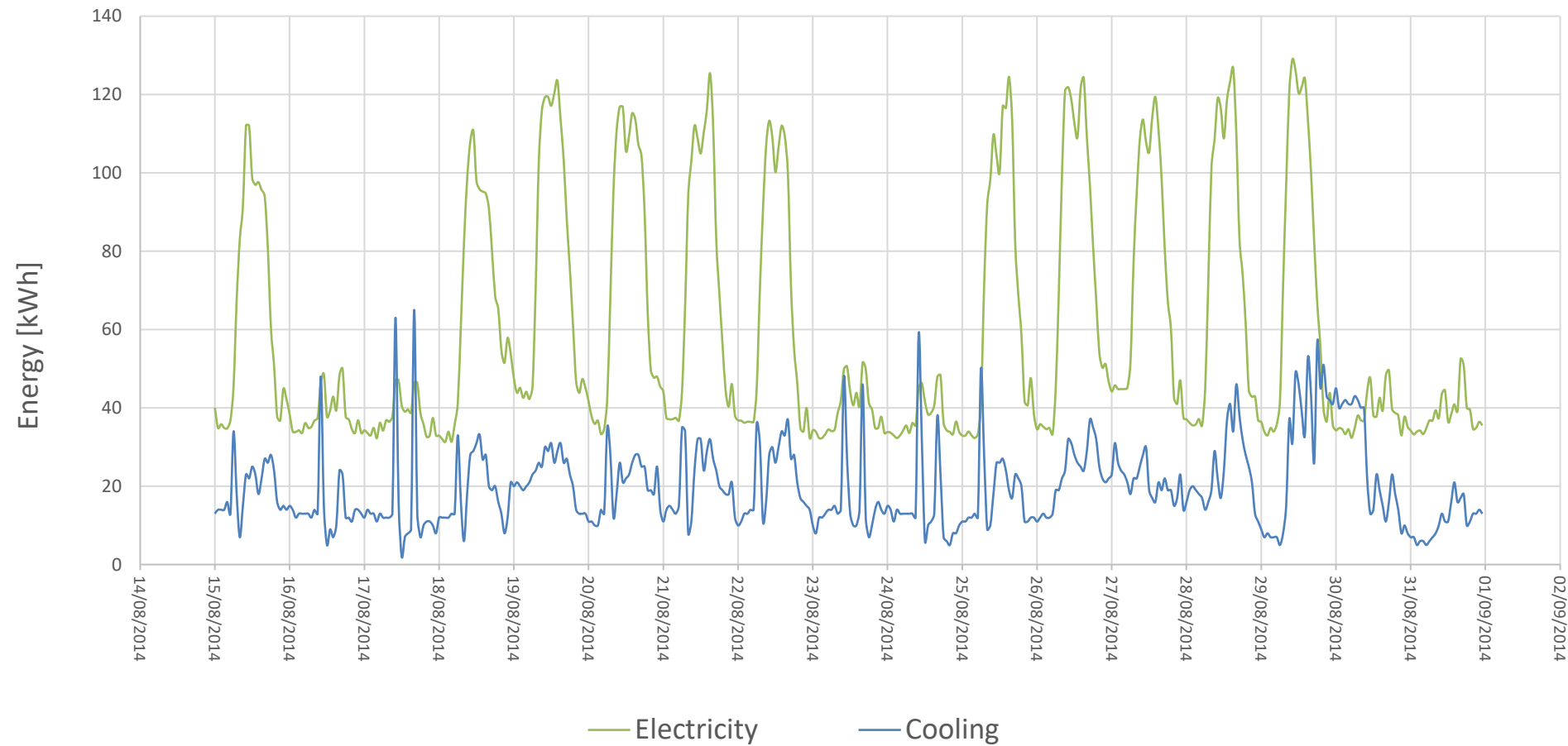
What's the point?

Electricity and cooling demand of the ETL building, 15-21 August 2014

Elect.	Cooling	Elect.	Cooling	Elect.	Cooling	Elect.	Cooling	Elect.	Cooling	Elect.	Cooling	Elect.	Cooling
15.08.2014		16.08.2014		17.08.2014		18.08.2014		19.08.2014		20.08.2014		21.08.2014	
40.03	13.00	38.41	15.00	34.28	12.00	32.94	12.00	47.63	20.00	41.66	11.00	44.16	11.00
34.88	14.00	33.97	14.00	33.59	14.00	32.03	12.00	43.88	21.00	37.69	11.00	37.47	14.00
35.84	14.00	33.88	12.00	32.94	13.00	31.34	12.00	45.13	20.00	35.91	10.00	37.09	15.00
34.91	14.00	34.28	13.00	34.91	13.00	33.94	12.00	42.56	19.00	36.75	10.00	37.13	14.00
34.81	16.00	33.59	13.00	32.25	11.00	31.34	13.00	44.13	20.00	33.28	14.00	37.47	13.00
36.84	13.00	36.13	13.00	36.22	13.00	35.84	13.00	42.25	21.00	34.59	13.00	36.78	15.00
45.47	34.00	34.91	13.00	34.19	12.00	41.25	33.00	45.44	23.00	42.50	35.00	44.13	35.00
67.47	18.00	35.19	12.00	36.84	12.00	61.16	16.00	70.06	24.00	68.56	28.00	65.97	34.00
82.91	7.00	36.78	14.00	36.47	12.00	82.22	6.00	102.44	26.00	95.63	12.00	94.09	8.00
91.47	15.00	37.47	13.00	37.75	13.00	98.22	18.00	115.84	25.00	110.72	18.00	103.00	11.00
112.06	23.00	46.06	48.00	47.03	63.00	107.84	28.00	119.31	30.00	116.84	26.00	112.03	24.00
112.00	22.00	48.66	15.00	47.09	16.00	110.75	29.00	119.38	29.00	116.78	21.00	108.81	32.00
98.56	25.00	37.75	5.00	40.63	2.00	97.91	31.00	117.13	31.00	105.56	22.00	104.94	32.00
96.94	23.00	39.34	9.00	39.03	7.00	95.82	33.20	120.31	26.00	109.44	23.00	110.41	24.00
97.63	18.00	42.91	7.00	39.63	8.00	95.21	26.80	123.50	29.00	115.19	26.00	116.16	29.00
95.63	22.00	39.38	10.00	38.78	9.00	94.72	28.00	113.66	31.00	113.66	28.00	125.41	32.00
94.09	27.00	48.28	24.00	46.41	65.00	90.91	20.00	103.34	26.00	107.19	28.00	112.03	27.00
80.97	26.00	49.97	23.00	46.38	13.00	79.97	19.00	88.00	27.00	104.31	25.00	82.84	24.00
60.47	28.00	37.75	12.00	38.72	7.00	68.53	20.00	75.47	23.00	89.63	25.00	69.16	20.00
50.91	24.00	37.09	12.00	35.84	10.00	65.25	16.00	61.16	20.00	64.00	19.00	56.34	19.00
37.78	16.00	34.56	11.00	32.63	11.00	54.75	13.00	46.72	14.00	49.56	19.00	44.47	18.00
36.75	14.00	33.59	14.00	32.97	11.00	51.50	8.00	43.84	13.00	47.72	18.00	40.31	18.00
44.84	15.00	36.84	14.00	37.44	10.00	57.88	12.00	47.38	13.00	47.97	25.00	46.09	21.00
42.19	14.00	33.56	13.00	32.97	8.00	53.84	21.00	45.09	13.00	45.44	14.00	38.03	12.00

What's the point?

Electricity and cooling demand of the ETL building, 15-21 August 2014



The point is ...

**The human brain can capture an image
in as little as 13 milliseconds!**

Can you say the same about text or numbers?

“while the images are seen for only 13 milliseconds before the next image appears, part of the brain continues to process those images for longer than that”

Source: <http://news.mit.edu/2014/in-the-blink-of-an-eye-0116>

Selecting the right visualization

- Define a **focus sentence** that summarizes what you want to show
- Decide how many **variables** you want to show in a single chart
- Decide how many **items** you want to display for every variable
- Decide if values are **spread** or **grouped**
- Identify **data types** and choose **representation styles**
- Choose the appropriate chart type
- Check if the chart fulfils the requirements of the focus sentence.

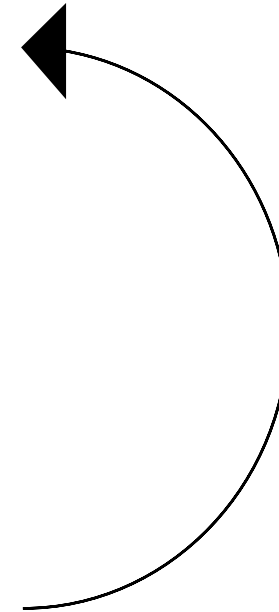
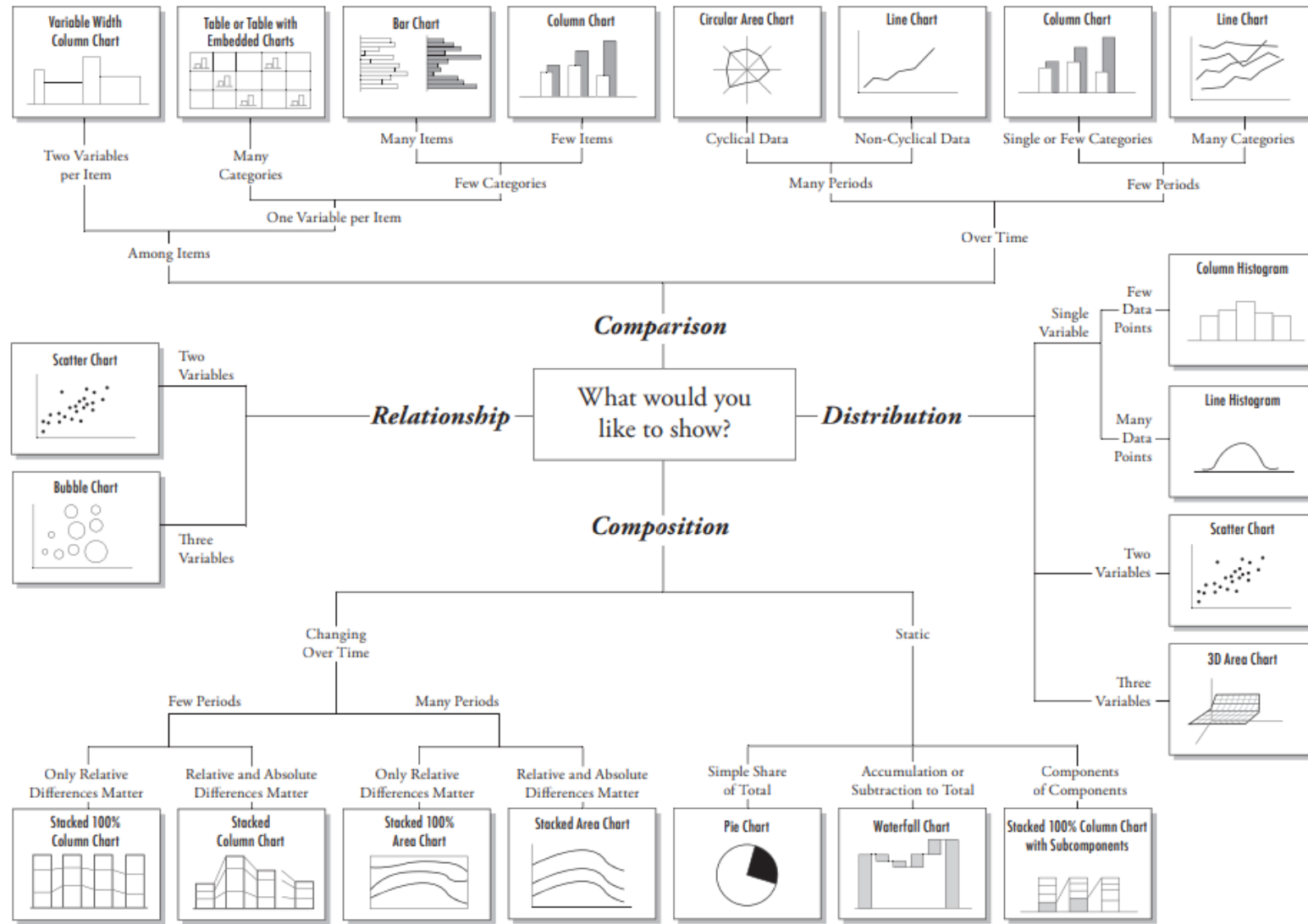
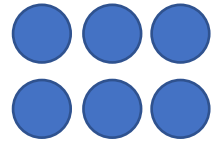


Chart Suggestions—A Thought-Starter



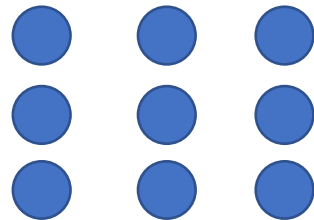
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

Data types



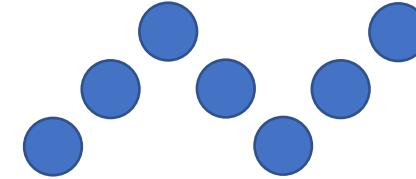
Quantitative

Data that can be counted or measured; has numerical values



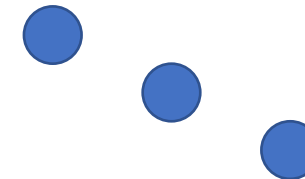
Qualitative / Categorical

Data that can be sorted in groups or categories



Continuous

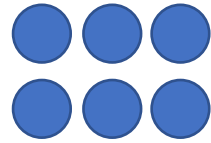
Data that can take any value within a certain range, even if data points are missing



Discrete

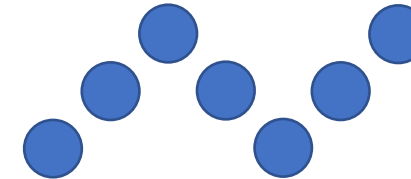
Data with a finite number of possible values; countable

Data types



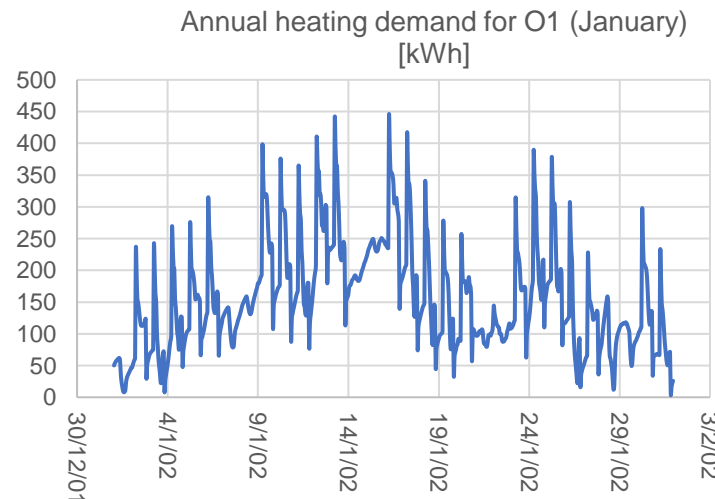
Quantitative

Data that can be counted or measured; has numerical values

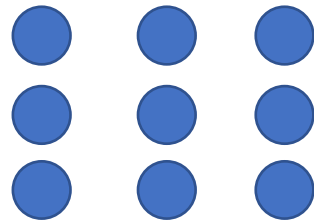
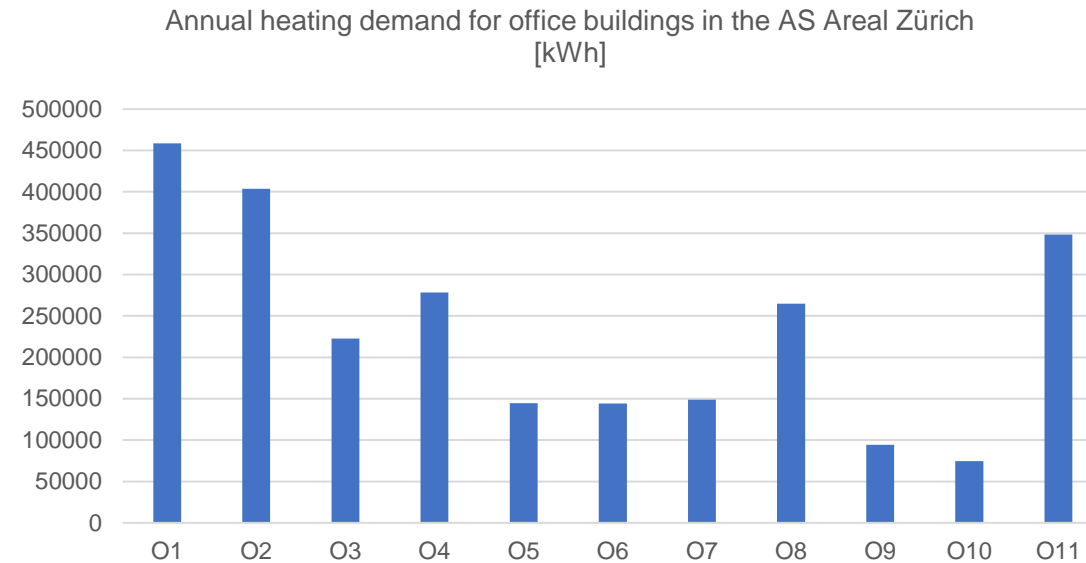


Continuous

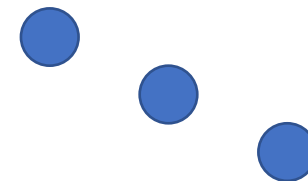
Data that can take any value within a certain range, even if data points are missing (interpolation)



Data types

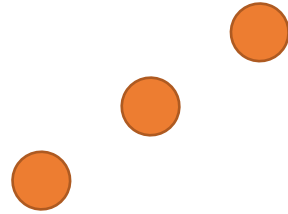


Qualitative / Categorical
Data that can be sorted in
groups or categories



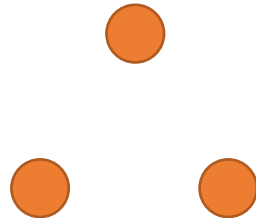
Discrete
Data with a finite number of possible
values; directly countable

Representation types



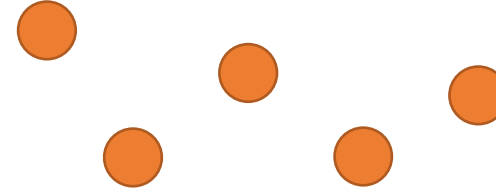
Comparison

Between two or multiple items, with emphasis on the difference or ranking



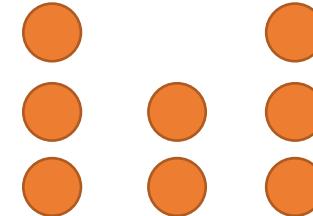
Relationship

Between two or more parameters of a series of items



Distribution

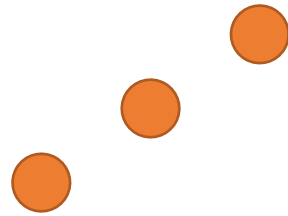
Of one or more parameters over a series of categories, without emphasis on difference or ranking



Composition

Shows subsets of data as part of the “whole” for a series of items

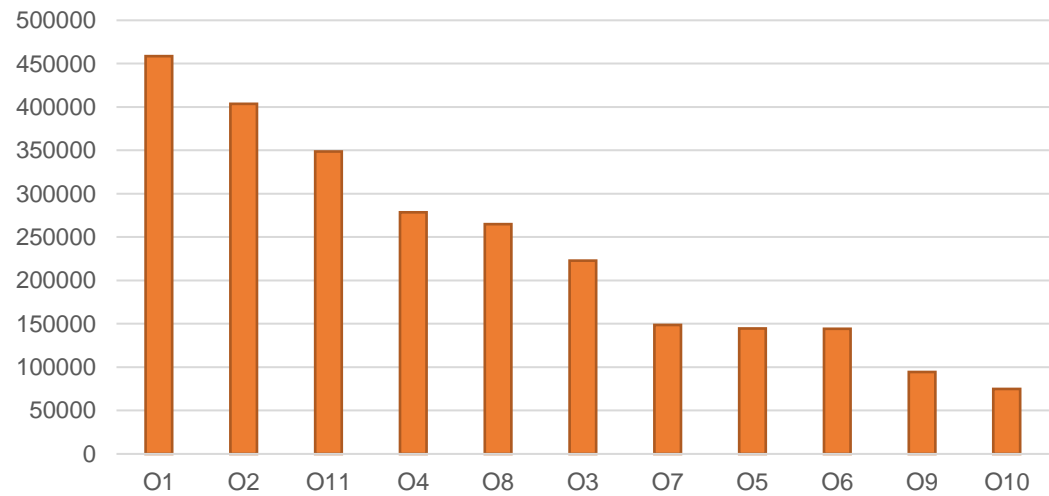
Representation types



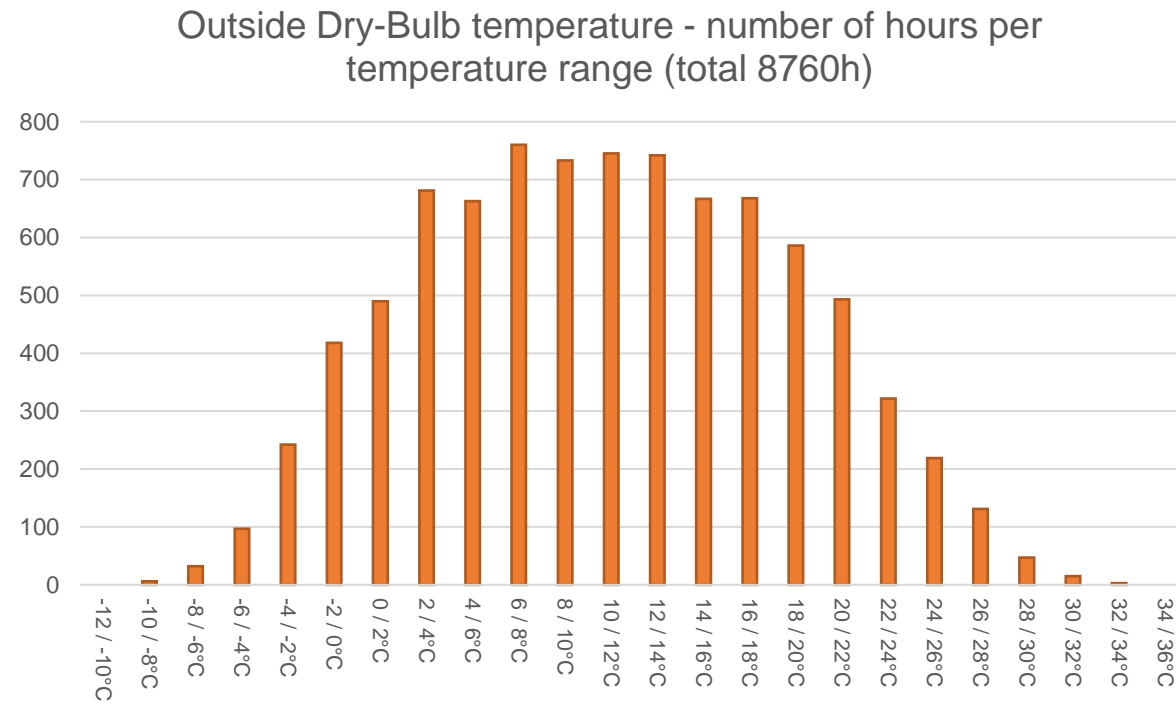
Comparison

Between two or multiple items, with emphasis on the difference or ranking

Annual heating demand for office buildings in the AS Areal Zürich
[kWh]



Representation types

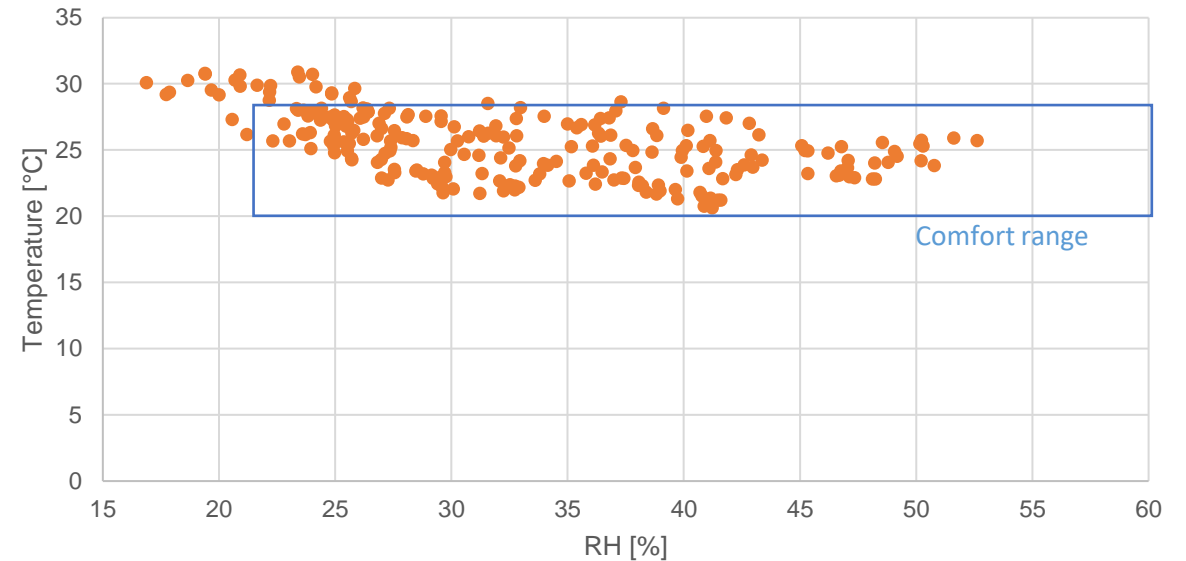


Distribution

No emphasis on difference or ranking.
Order of horizontal categories has priority

Representation types

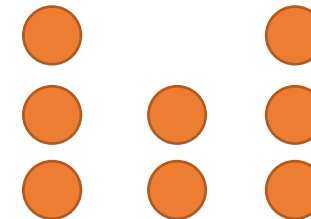
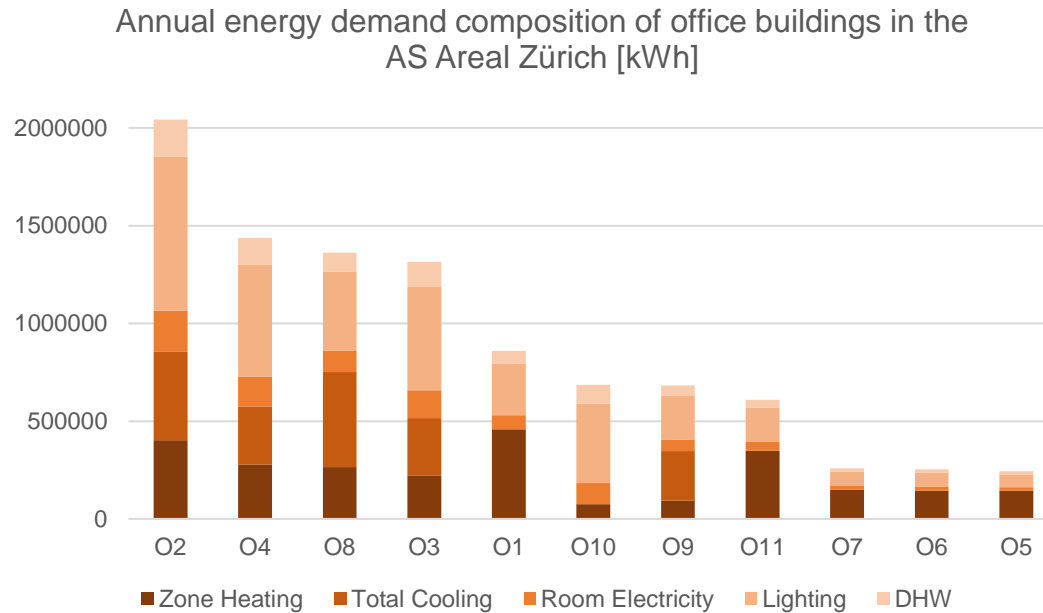
Hourly Temperature vs. RH plot (1-10 June)



Relationship

Between two or more parameters of a series of items

Representation types



Composition

Shows subsets of data as part of the “whole” for a series of items

Select the right visualization – iterative process

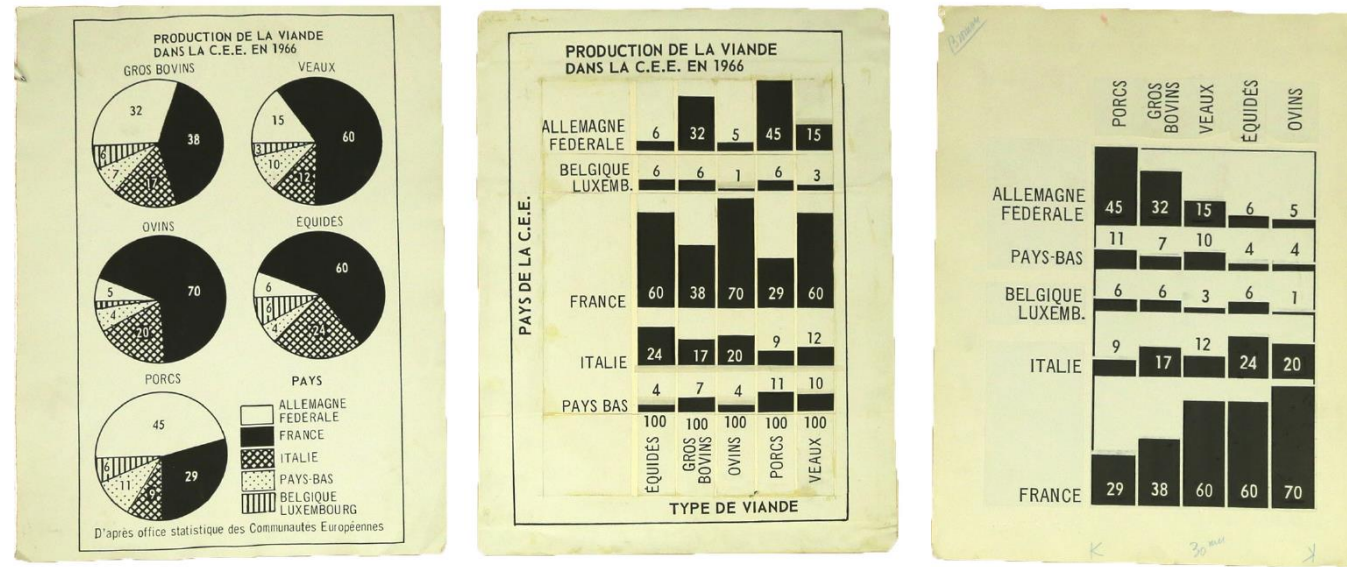
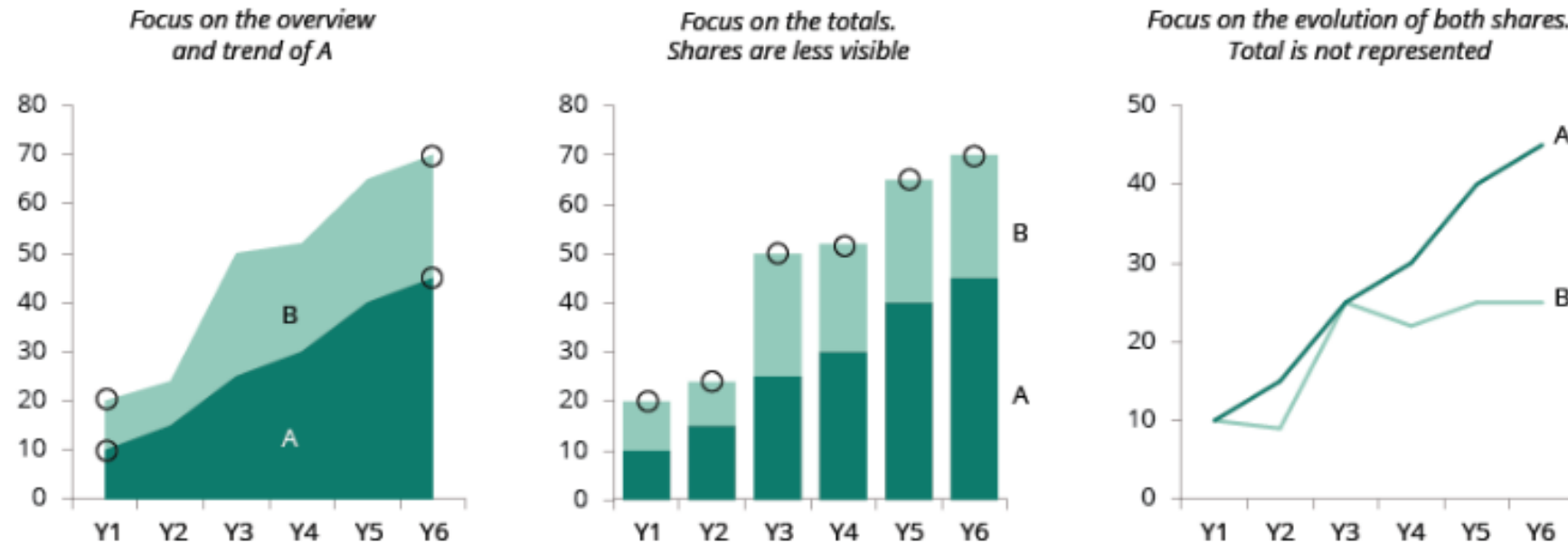


Figure 2. *Left:* Pie charts showing the contribution of different countries in the production of different types of meat, which Bertin qualified as “useless”. *Middle:* With a matrix visualization, high-level patterns become immediately visible. *Right:* Since countries and meats do not have a natural order, many other matrices can be produced, including this one, which is more effective. Thus, being able to try different orderings was essential. Drafts for the book *La Graphique* (Bertin, 1977) Courtesy of EHESS/AN ref. 20010291/36. All rights reserved.

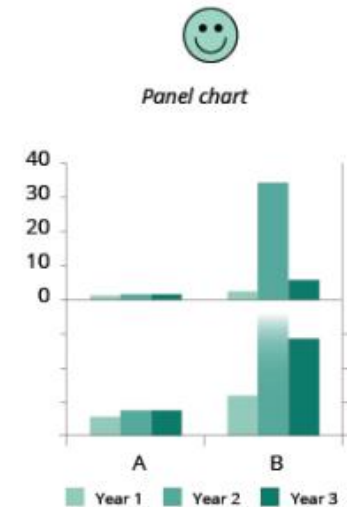
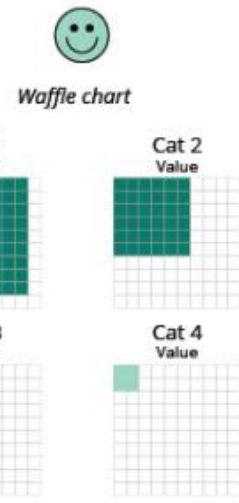
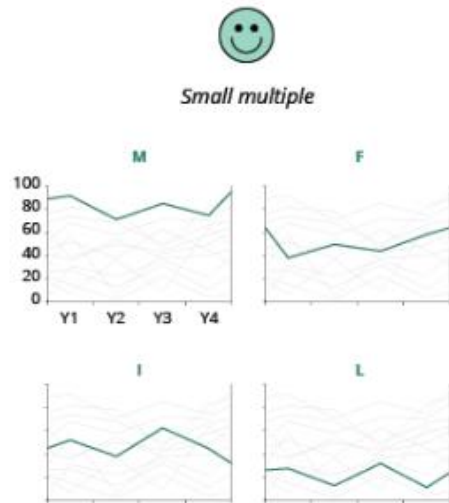
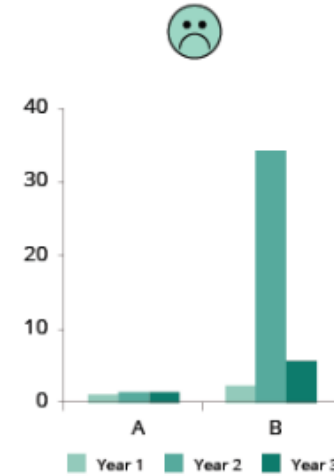
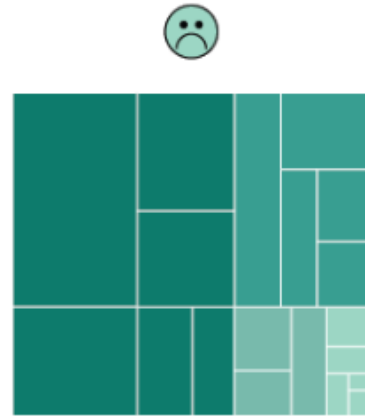
Source: Charles Perin, “Jacques Bertin’s Legacy in Information Visualization and the Reorderable Matrix”

Select the right visualization



Source: [European Environment Agency-Chart dos and don'ts](#)

Select the right visualization



Source: [European Environment Agency-Chart dos and don'ts](#)

Chart design

- Keep the design simple and elegant
- Limit the number of colours you use (use tones)
- Do not use the same colour palette for different categories
- Keep the colour scheme coherent in all your charts
- Consider a colour-blind colour scheme
- Use a legend only if necessary



Color resources to use in python

- Seaborn: https://seaborn.pydata.org/tutorial/color_palettes.html
- Matplotlib: https://matplotlib.org/stable/gallery/color/named_colors.html#sphx-glr-gallery-color-named-colors-py

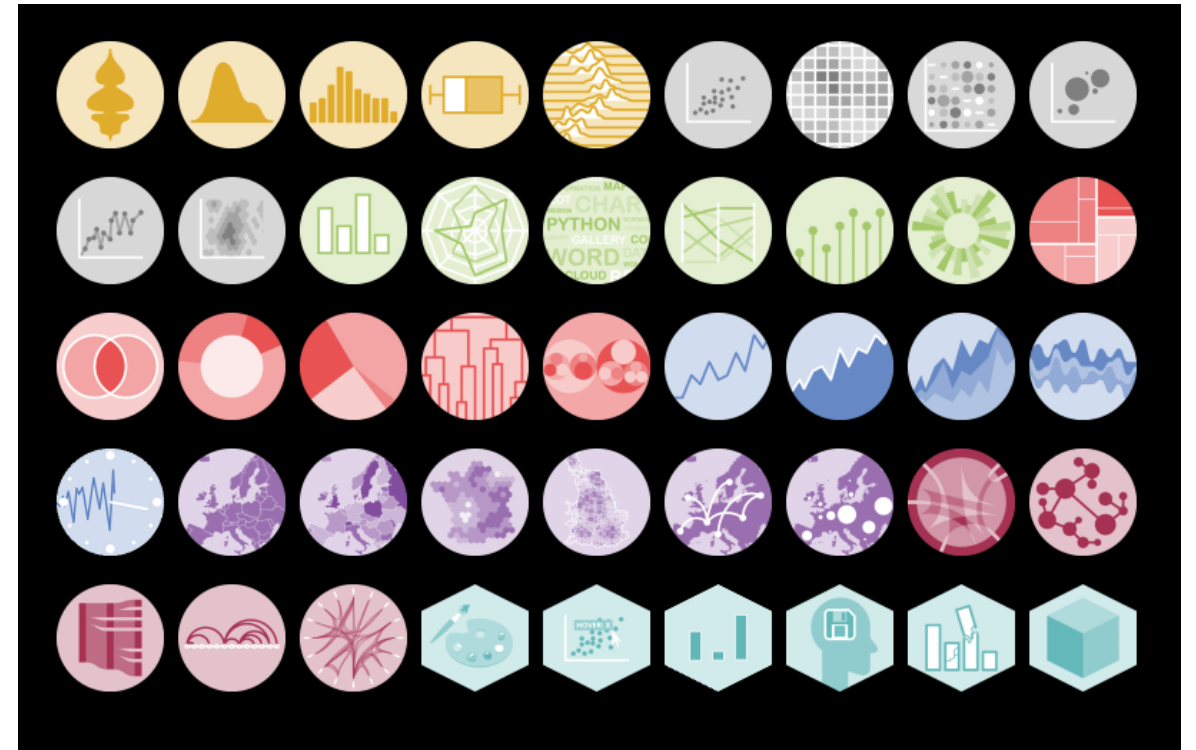
Visualization Resources

For inspiration



<https://datavizcatalogue.com/index.html>

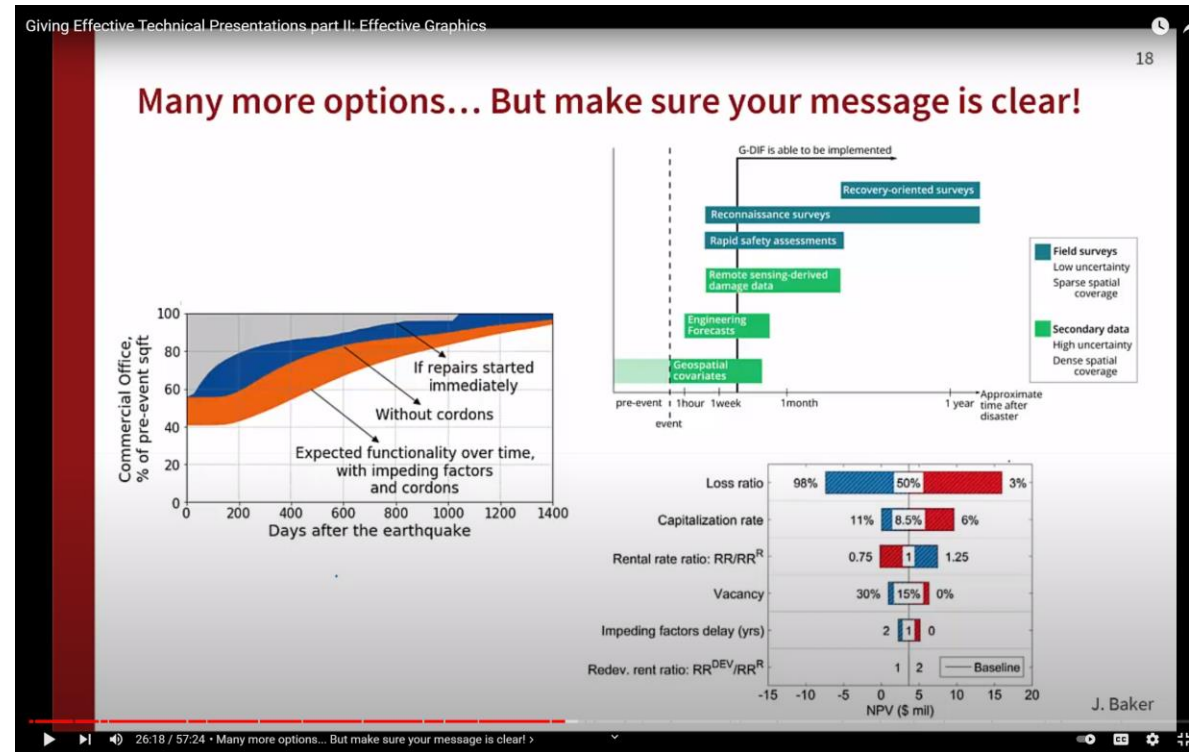
For implementation



<https://www.python-graph-gallery.com/>

Further Resources

Effective Graphics

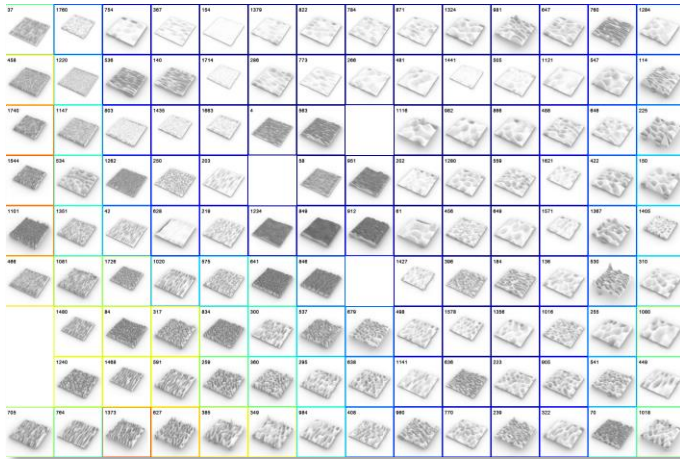


<https://youtu.be/Z0BCD6f9b4I>

Sneak Peak: Data Processing and Visualization 2

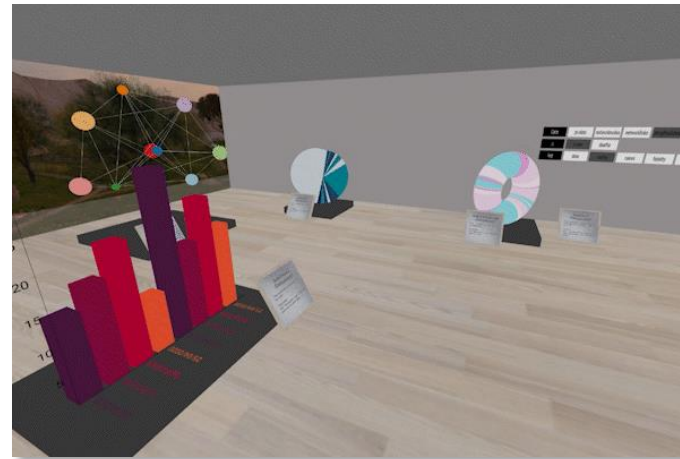
“Understanding and exploring high dimensional data”

Dimensionality reduction



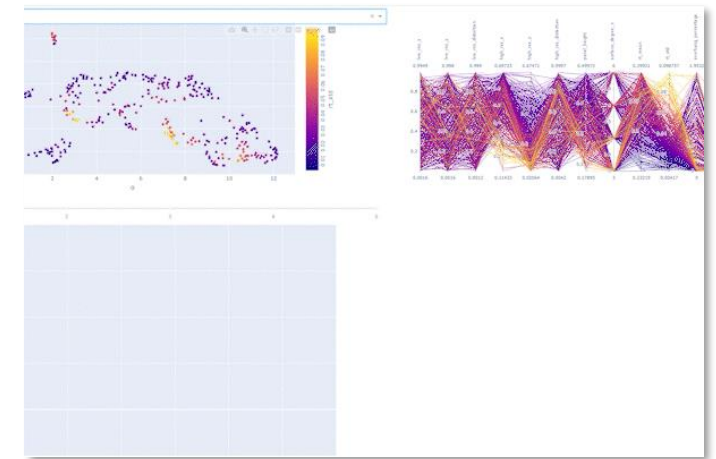
PCA, TSNE, UMAP, SOM

Immersion



Data representation: using more than visual sense, BabiaXR

Interaction



Interactive data visualization and annotation