3050571 Practical Clin Data Sci

Session 4: Data exploration and visualization

February 8, 2024

Sira Sriswasdi, PhD

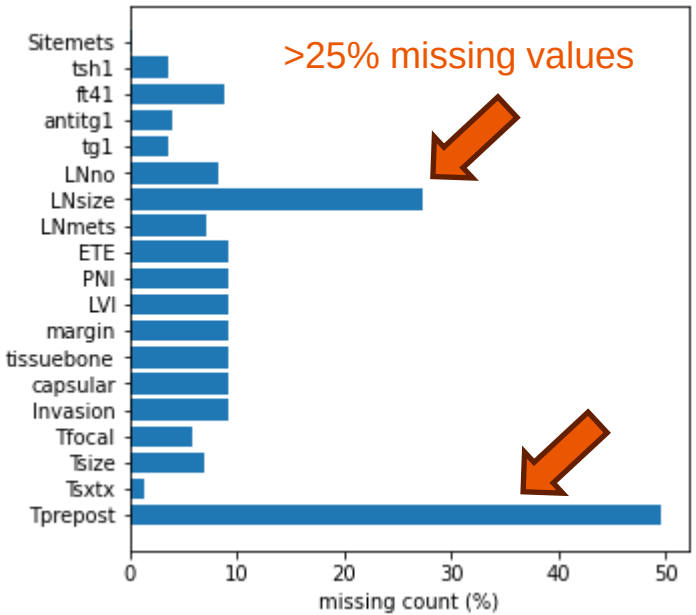
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

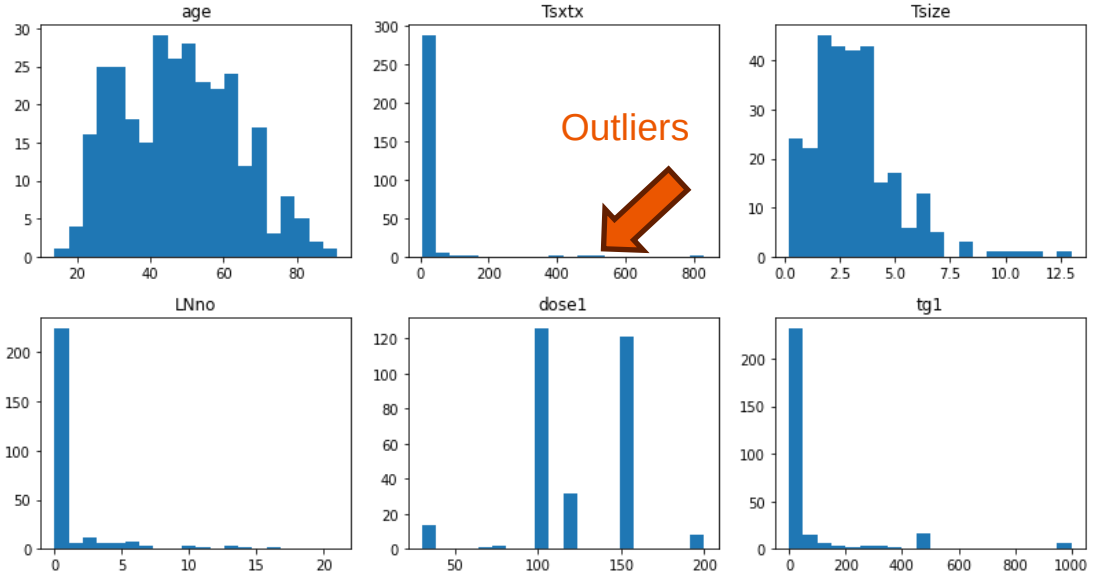
Exploratory data analysis (EDA)

General thought process

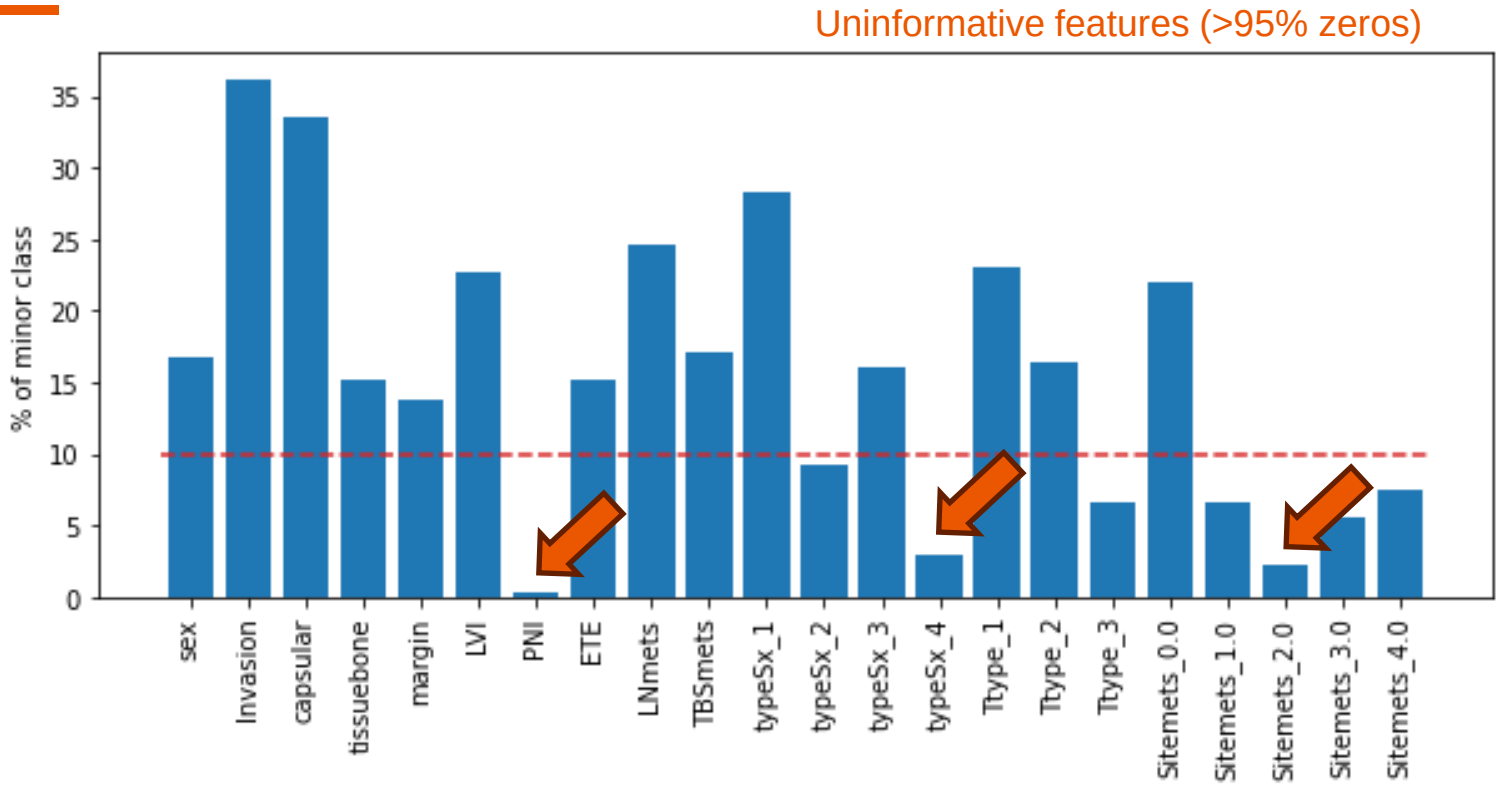
- Quality check: Identify outliers and missing values
- Data distribution: Catalog the amount of usable data
- Sanity test: Evaluate some trivial hypotheses
- Hypothesis development
  - Knowledge driven
  - Data-driven
- Hypothesis testing: How do I support / disprove my hypothesis?

Identify bad features



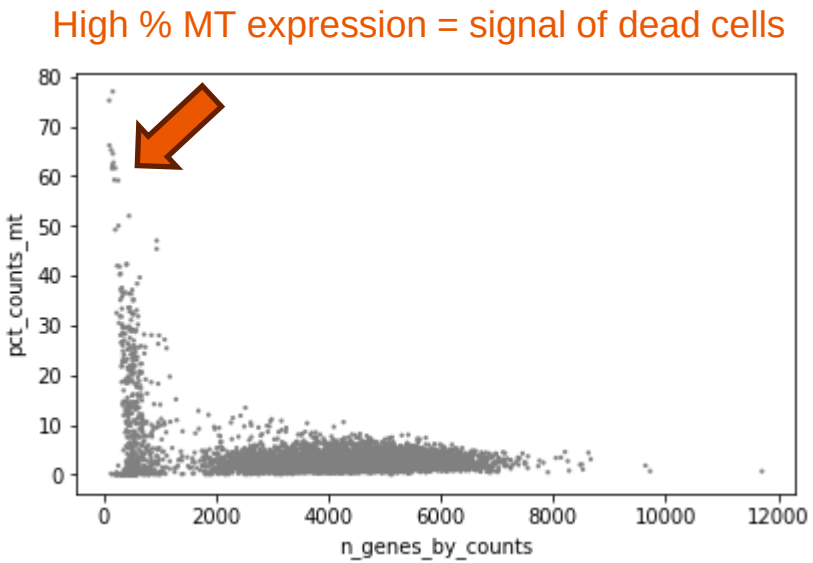


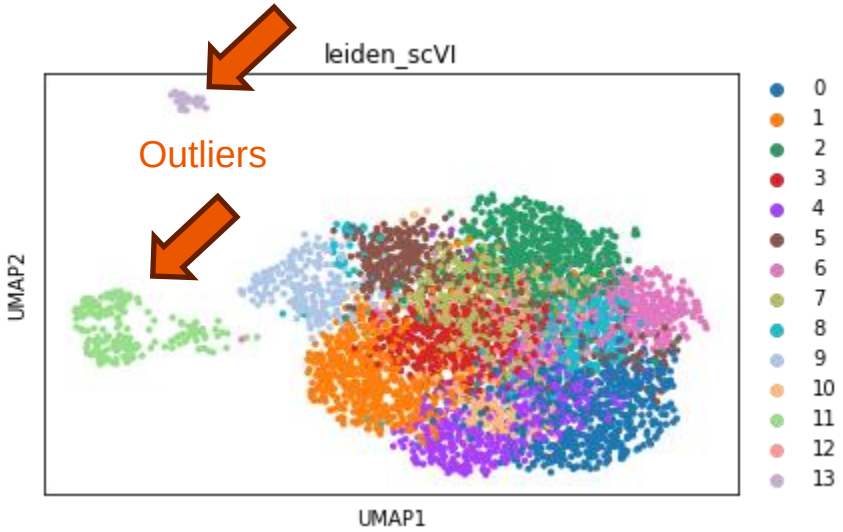
Identify bad features



Identify outliers and data clusters

High % MT expression = signal of dead cells





Catalog usable data

- Sample:
  - How many samples are there from each group?
  - Is there any subpopulation structure (cluster)?

- Features:
  - How many features are usable?
  - Are they expected to be informative?

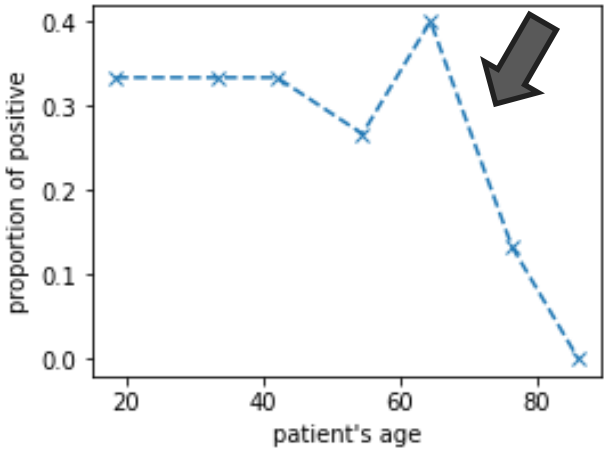
- What questions can or cannot be investigated with this dataset?

Sanity test

- What should be true?
  - Obvious associations between features
  - Prior knowledge
  - Expected sample subpopulation

- What should not be true?
  - Pattern that goes against prior knowledge
  - Could still happen:
    - Bias in data collection
    - Small sample size
    - Exception to the

Lower rate of drug allergy in patients 65 years and older!!



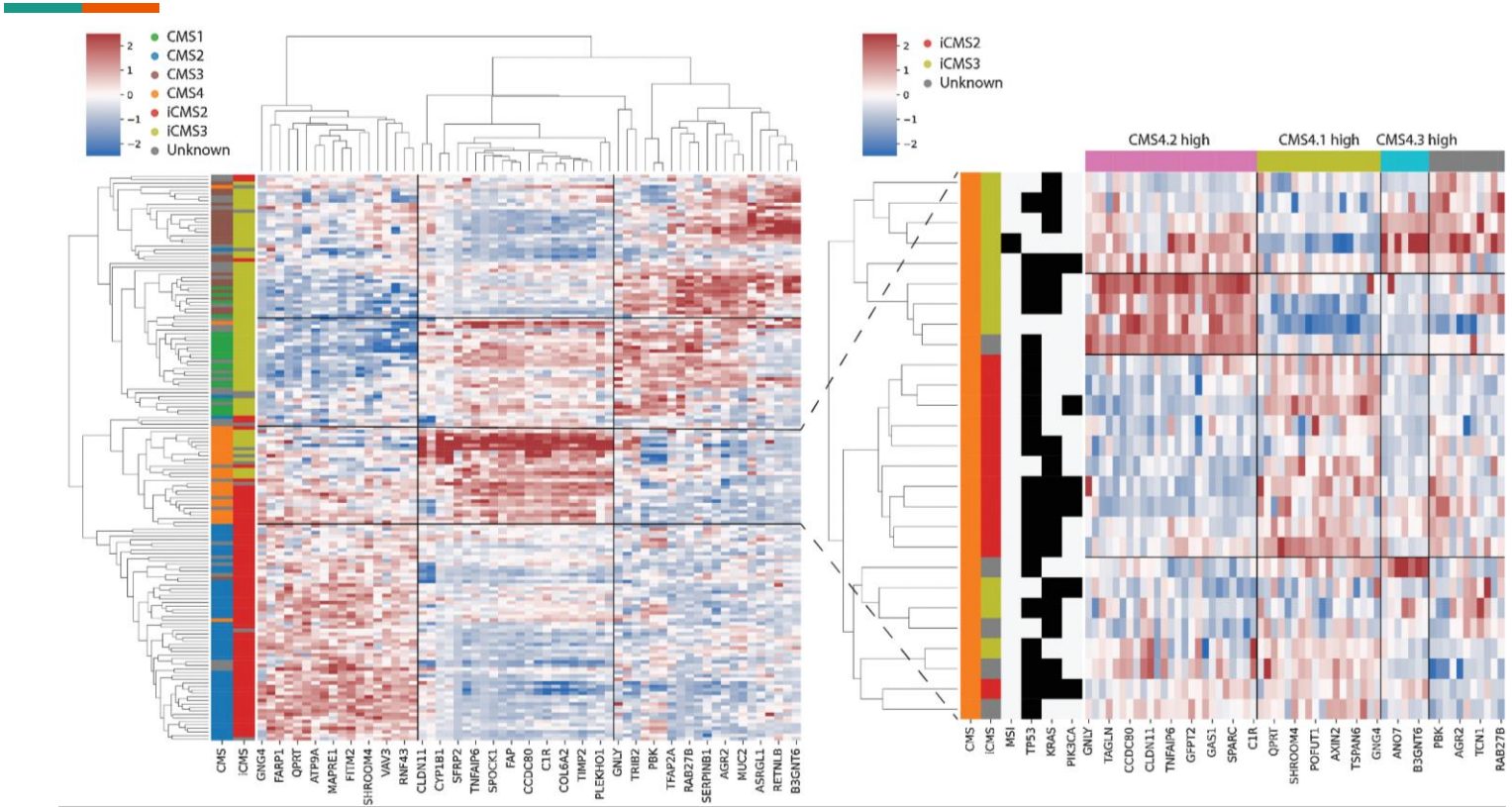
Hypothesis development & testing

- Knowledge driven

- Data driven
  - Correlation between all pairs of features
  - Unsupervised learning
    - Dimensionality reduction for visualization
    - Clustering for sample subpopulation discovery

- Your weapons = statistical test + visualization

Clustermap: Bi-clustering of samples & features

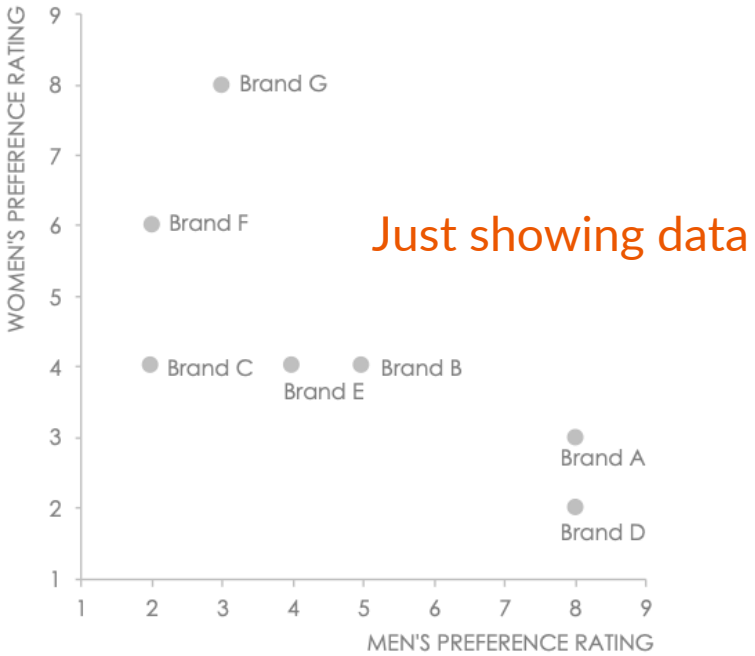


Visualization

Key rules for data visualization

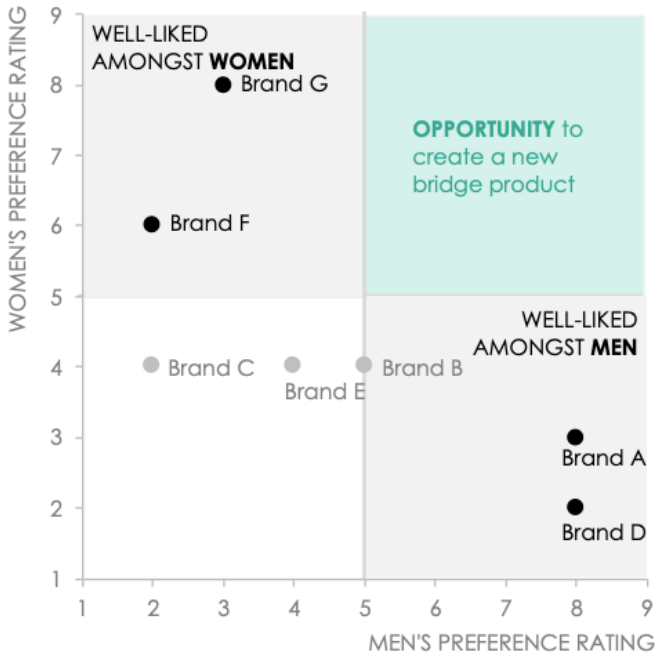
- Tell a story: Each graph has a purpose
- Reduce noise: Remove unnecessary graph elements
- Focus attention:
  - Each graph convey only one or few messages
  - Guide the reader's eyes
- Provide context: Readers should be able to understand the graph even without your explanation

Tell a story



Just showing data

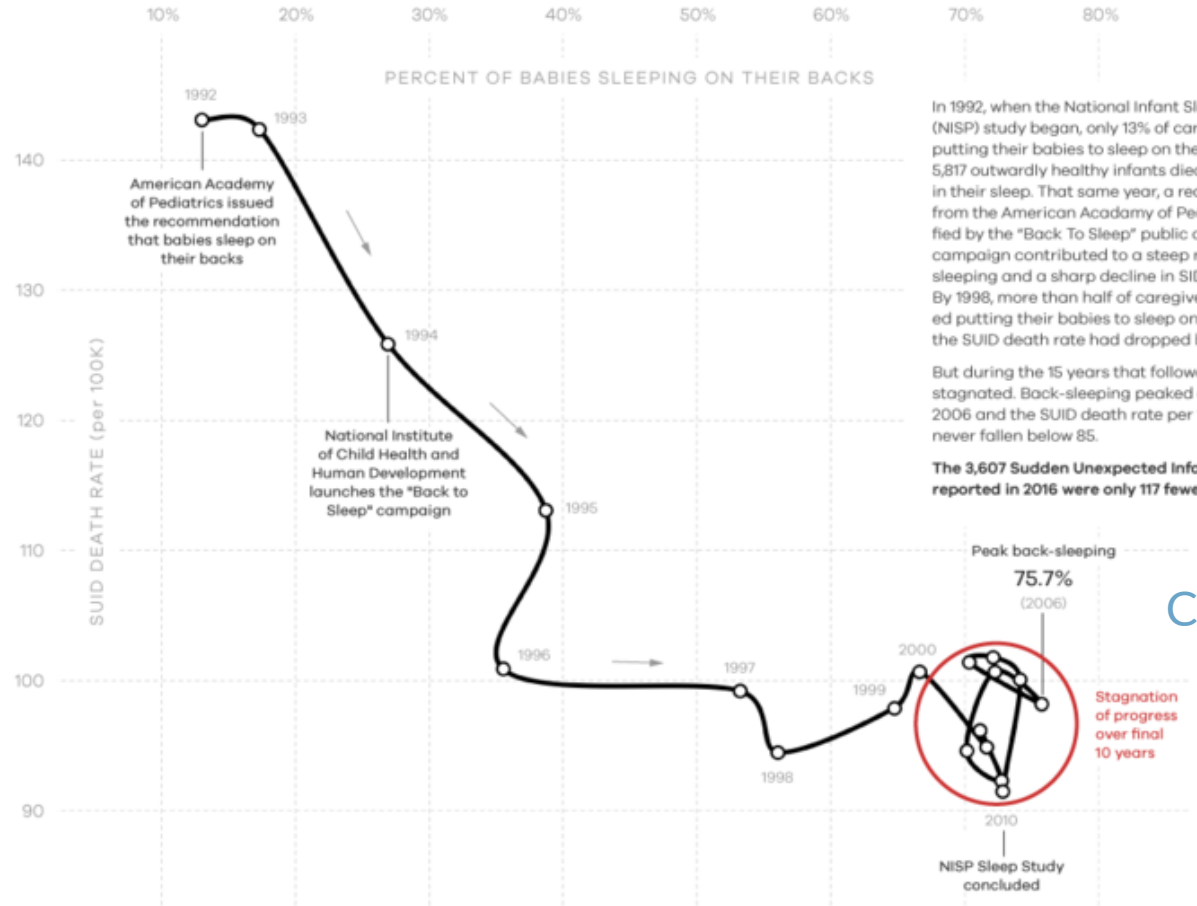
Clear message



https://www.storytellingwithdata.com/blog/2020/5/27/what-is-a-scatterplot

Back To Sleep and the War on SIDS

SUID Death Rate Relative to Sleep Position, 1992–2010



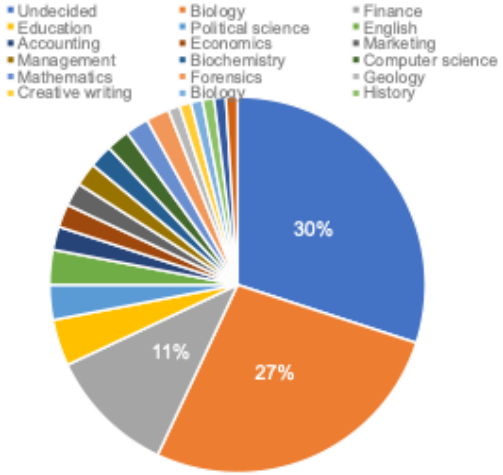
In 1992, when the National Infant Sleep Position (NISP) study began, only 13% of caregivers reported putting their babies to sleep on their backs, and 5,817 outwardly healthy infants died unexpectedly in their sleep. That same year, a recommendation from the American Academy of Pediatrics, amplified by the "Back To Sleep" public awareness campaign contributed to a steep rise in back sleeping and a sharp decline in SIDS/SUID deaths. By 1998, more than half of caregivers polled reported putting their babies to sleep on their backs and the SUID death rate had dropped by 31%.

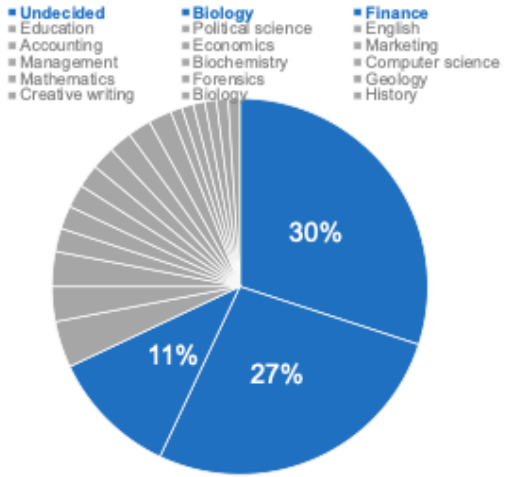
But during the 15 years that followed, progress stagnated. Back-sleeping peaked at about 75% in 2006 and the SUID death rate per 100,000 has never fallen below 85.

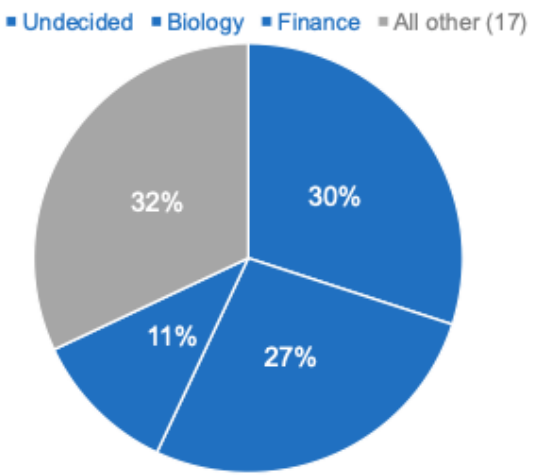
The 3,607 Sudden Unexpected Infant Deaths reported in 2016 were only 117 fewer than in 1998.

Clear message

Remove noise



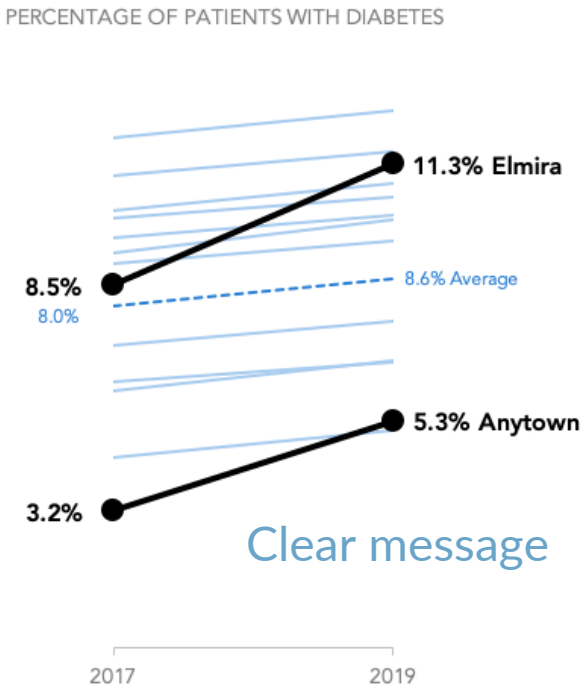




https://www.storytellingwithdata.com/blog/2020/5/14/what-is-a-pie-chart

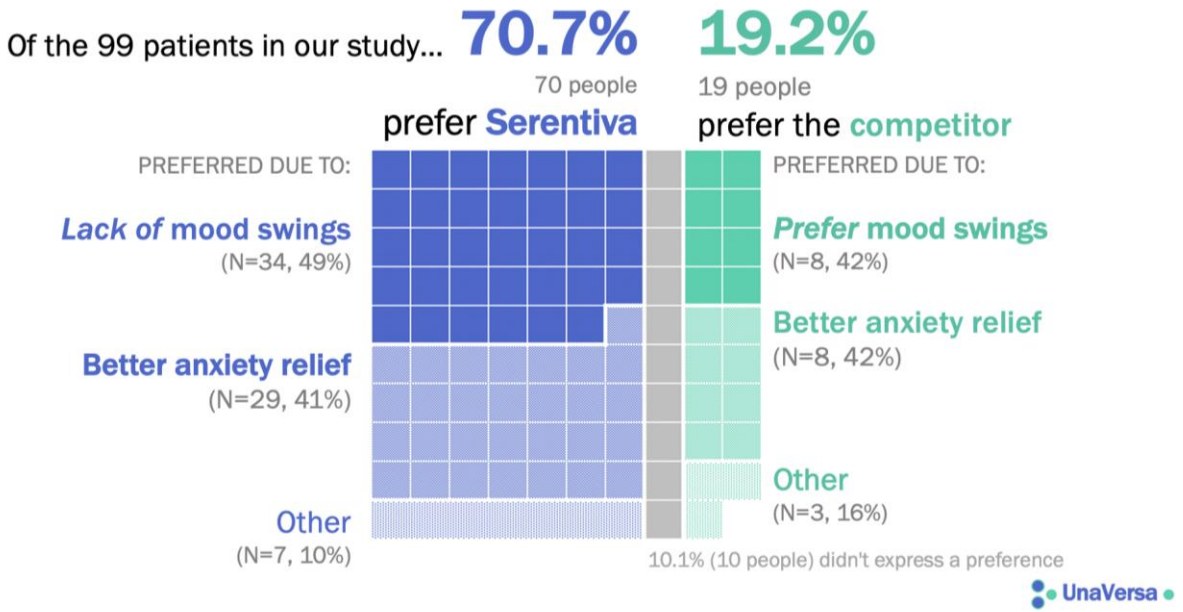
- Showing complete data is good during initial investigation
- But not necessary when telling a story

Focus attention



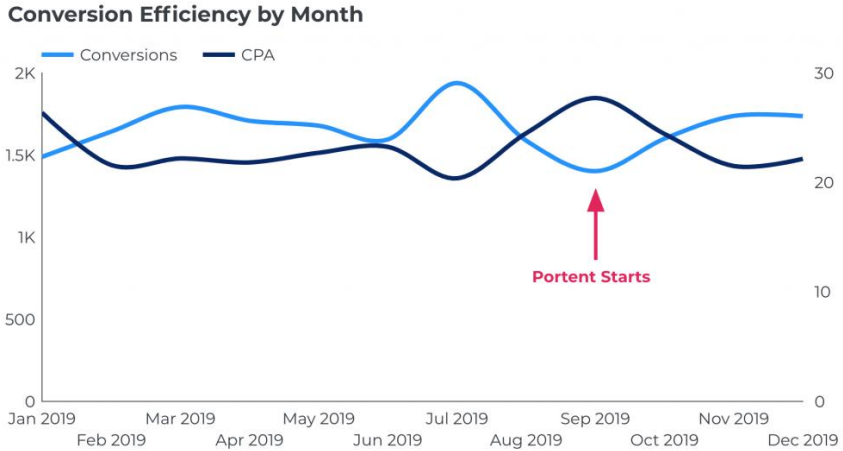
Clear message

Too many messages



https://www.storytellingwithdata.com/blog/2020/7/27/what-is-a-slopegraph

Provide context



	Campaign	Country	Device	Impressions	Clicks	CTR ▾
1.	Vests	Mexico	Desktop	2,225,384	12,560	0.56%
2.	Boots	United States	Mobile	2,419,478	13,526	0.56%
3.	Boots	Canada	Tablet	2,320,431	12,804	0.55%
4.	Boots	Mexico	Desktop	2,592,762	14,007	0.54%
5.	Gloves	United States	Mobile	2,471,944	13,320	0.54%
6.	Gloves	Mexico	Desktop	2,499,738	13,420	0.54%
7.	Vests	United States	Mobile	2,467,340	12,915	0.52%
8.	Pants	United States	Mobile	2,523,501	13,179	0.52%
9.	Pants	Canada	Tablet	2,521,257	12,851	0.51%
10.	Shirts	Canada	Tablet	2,672,690	13,465	0.50%

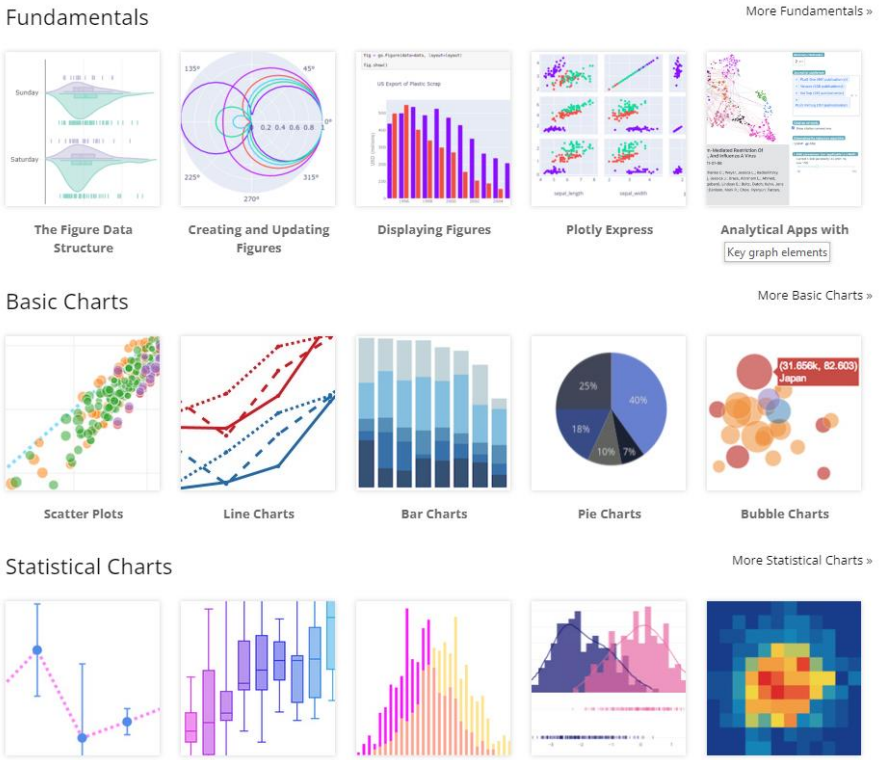
1 - 21 / 21

BENCHMARK CTR: 0.45%

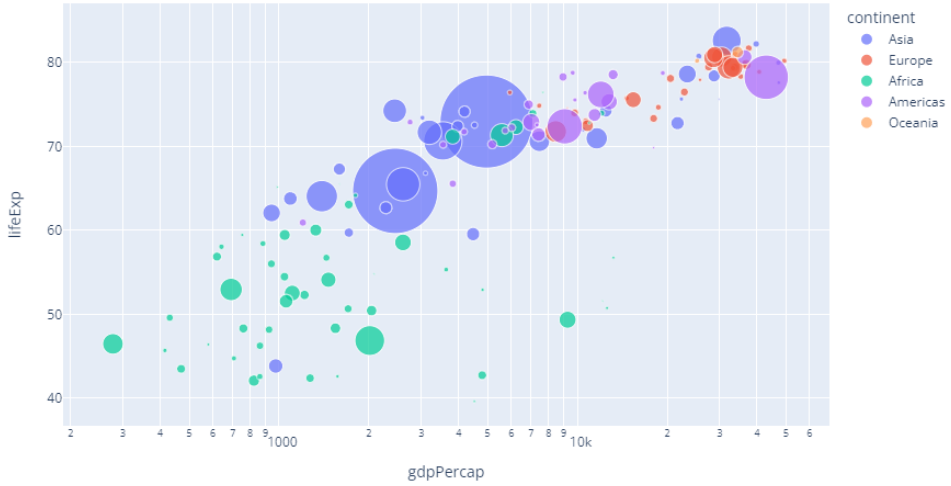
https://gfchart.com/2022/08/adding-context-to-your-in-content-charts-and-graphs/

Graph elements

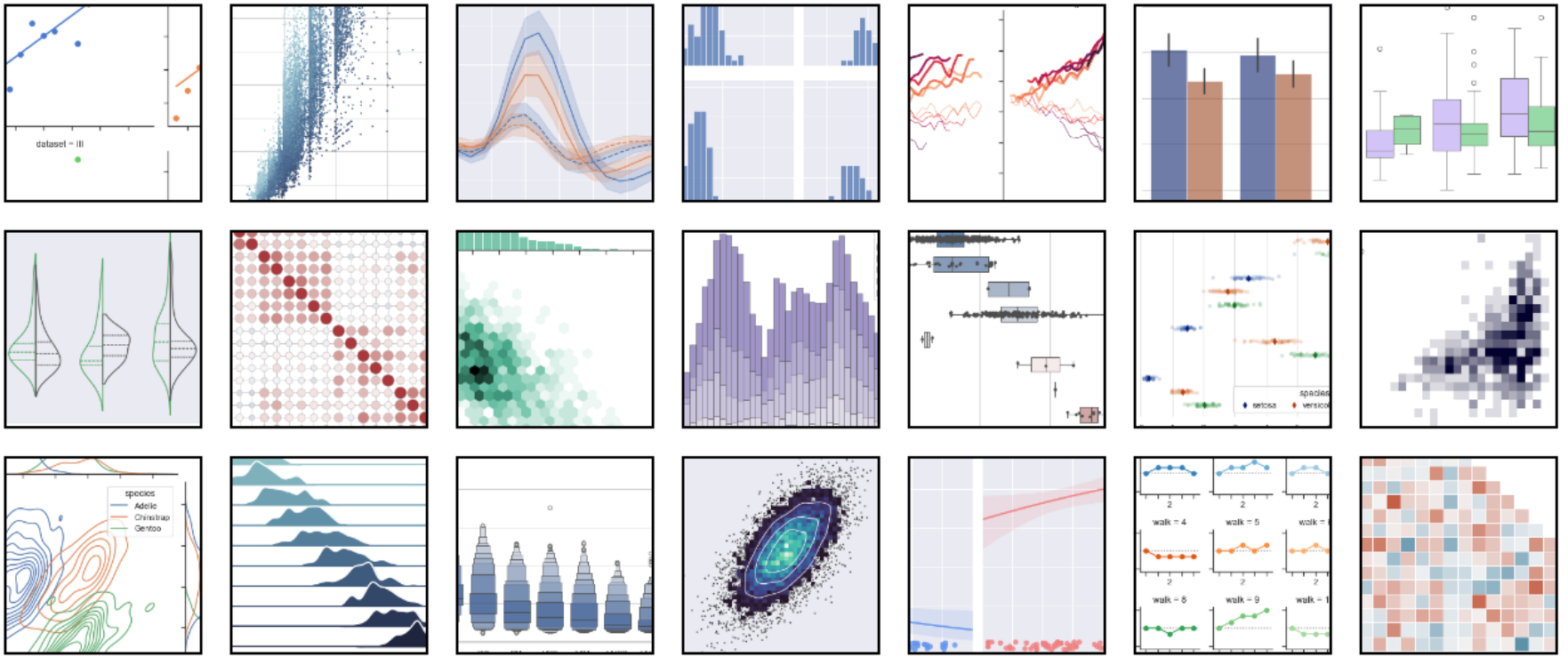
Plotly (https://plotly.com/python/)



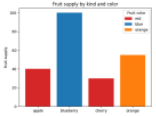

```
import plotly.express as px
df = px.data.gapminder()

fig = px.scatter(df.query("year==2007"), x="gdpPercap", y="lifeExp",
                 size="pop", color="continent",
                 hover_name="country", log_x=True, size_max=60)
fig.show()
```



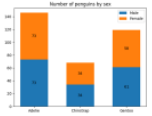
Seaborn (https://seaborn.pydata.org)



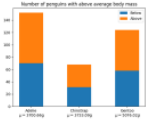
Matplotlib (https://matplotlib.org/stable/)



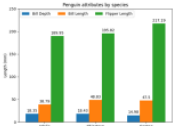
Bar color demo



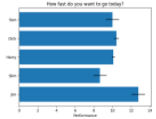
Bar Label Demo



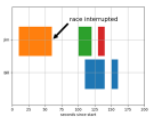
Stacked bar chart



Grouped bar chart with labels

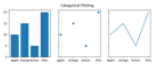


Horizontal bar chart

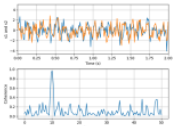


Broken Barh

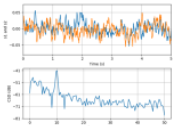
CapStyle



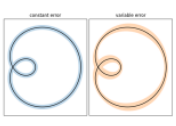
Plotting categorical variables



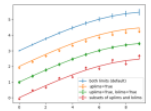
Plotting the coherence of two signals



Cross spectral density (CSD)



Curve with error band



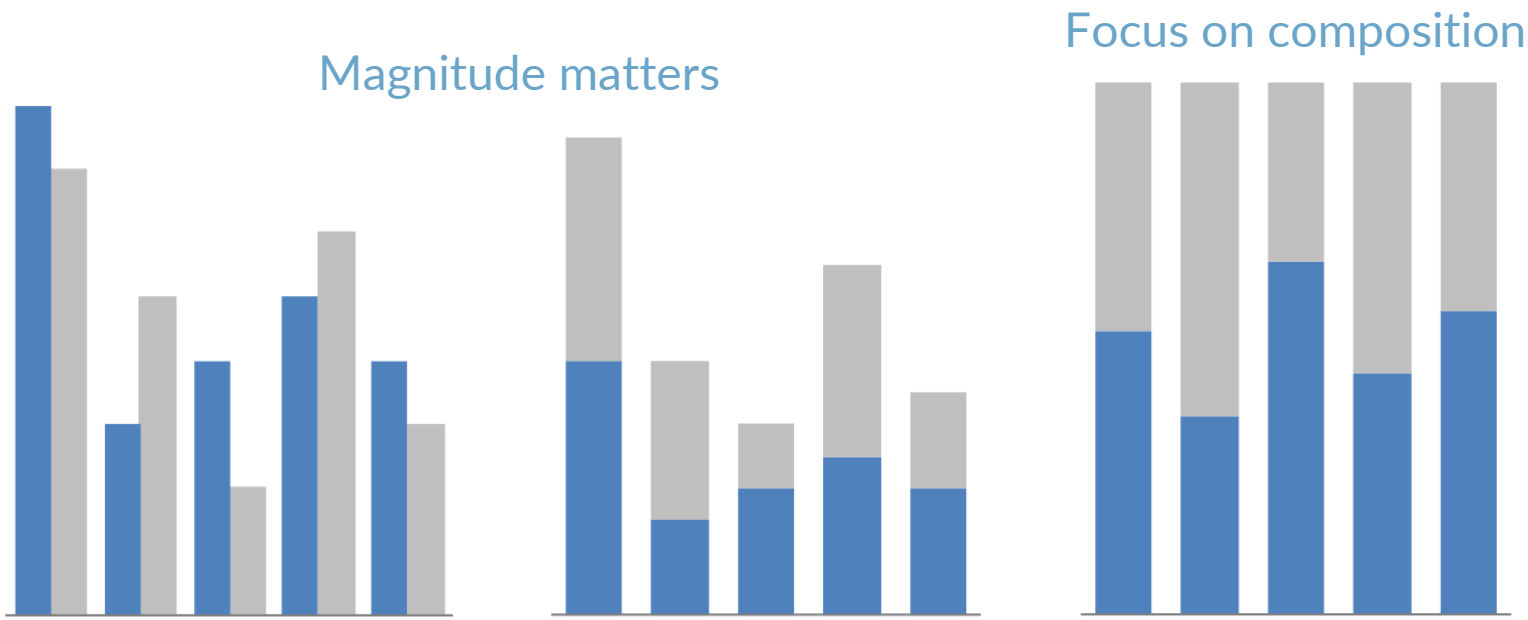
Errorbar limit selection

```python
import matplotlib.pyplot as plt
import numpy as np

# data from https://allisonhorst.github.io/palmerpenguins/

species = (
    "Adelie\n $\\mu=$3700.66g",
    "Chinstrap\n $\\mu=$3733.09g",
    "Gentoo\n $\\mu=5076.02g$",
)
weight_counts = {
    "Below": np.array([70, 31, 58]),
    "Above": np.array([82, 37, 66]),
}
width = 0.5

fig, ax = plt.subplots()
bottom = np.zeros(3)

for boolean, weight_count in weight_counts.items():
    p = ax.bar(species, weight_count, width, label=boolean, bottom=bottom)
    bottom += weight_count

ax.set_title("Number of penguins with above average body mass")
ax.legend(loc="upper right")

plt.show()
```

Key graph elements

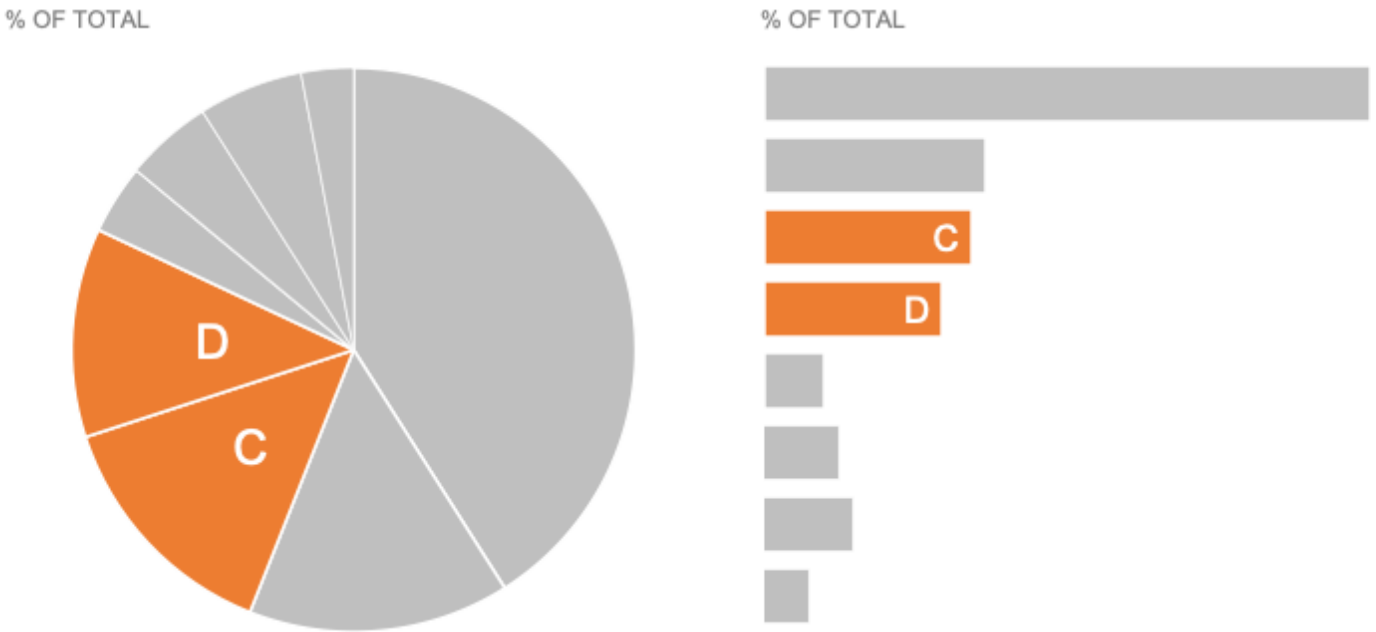
- Graph style
- Axis
- Label / Annotation
- Color / Highlight
- Marker / shape
- Zoom / Scale
- Reference / Normalization
- Multiplot / Inset

Graph style



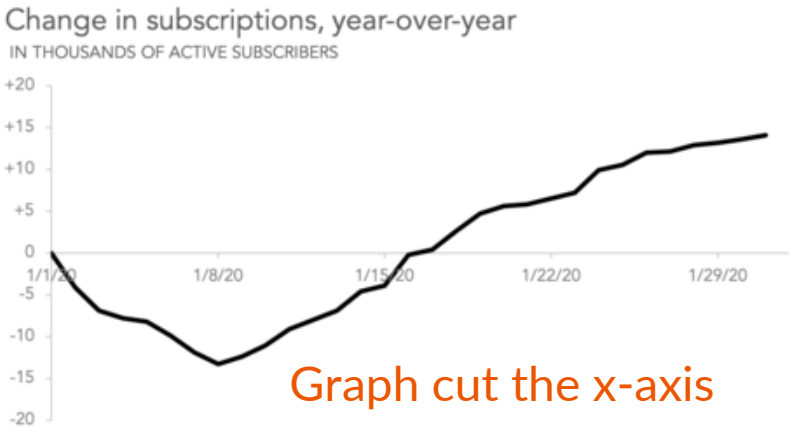
https://www.storytellingwithdata.com/blog/2020/2/19/what-is-a-bar-chart

Graph style



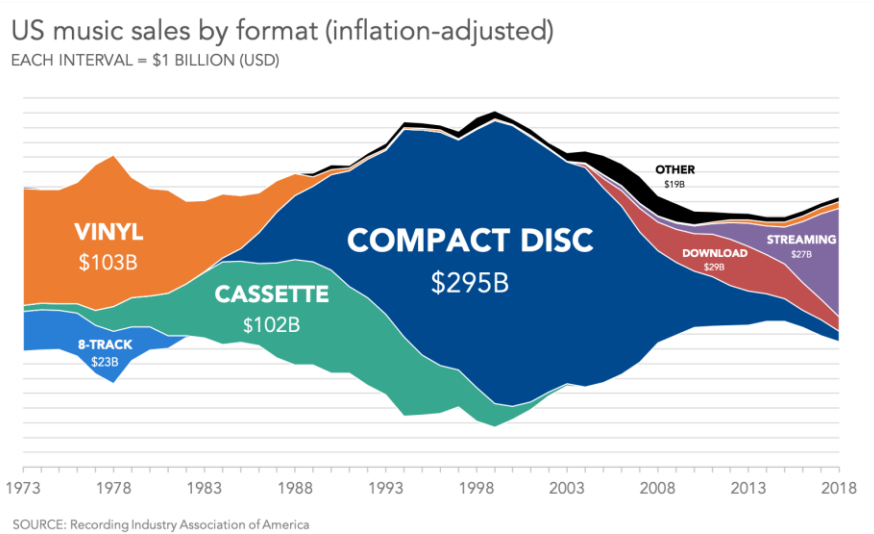
https://www.storytellingwithdata.com/blog/2020/2/19/what-is-a-pie-chart

Axis



Graph cut the x-axis

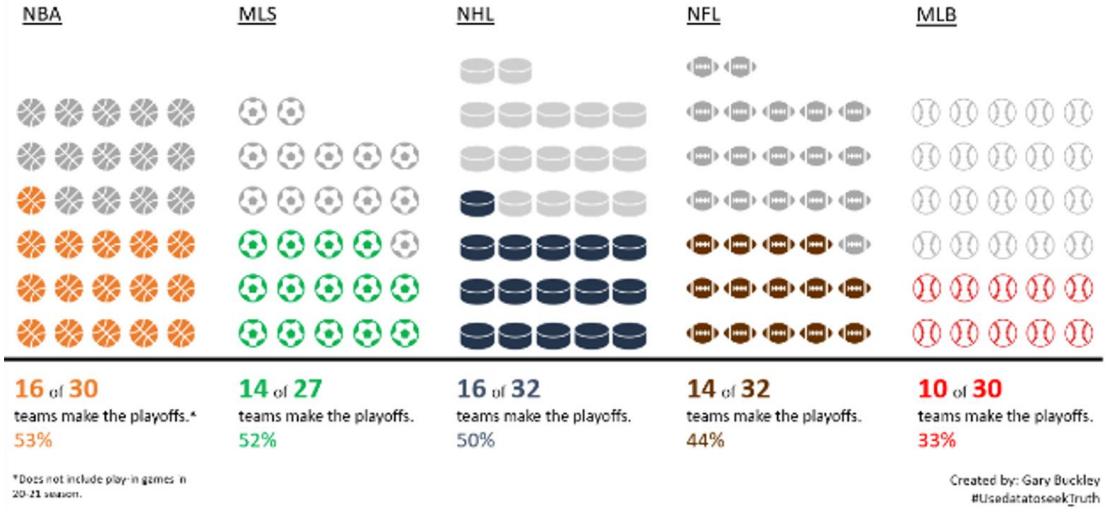
No unit on the y-axis



https://www.storytellingwithdata.com/blog/2020/3/24/what-is-a-line-graph

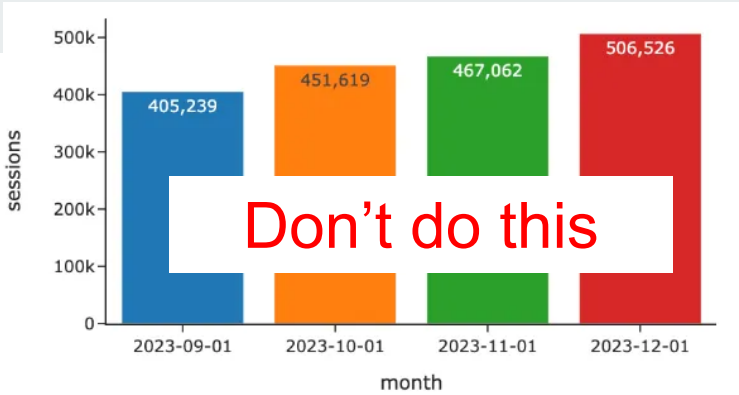
https://www.storytellingwithdata.com/blog/2020/4/9/what-is-an-area-graph

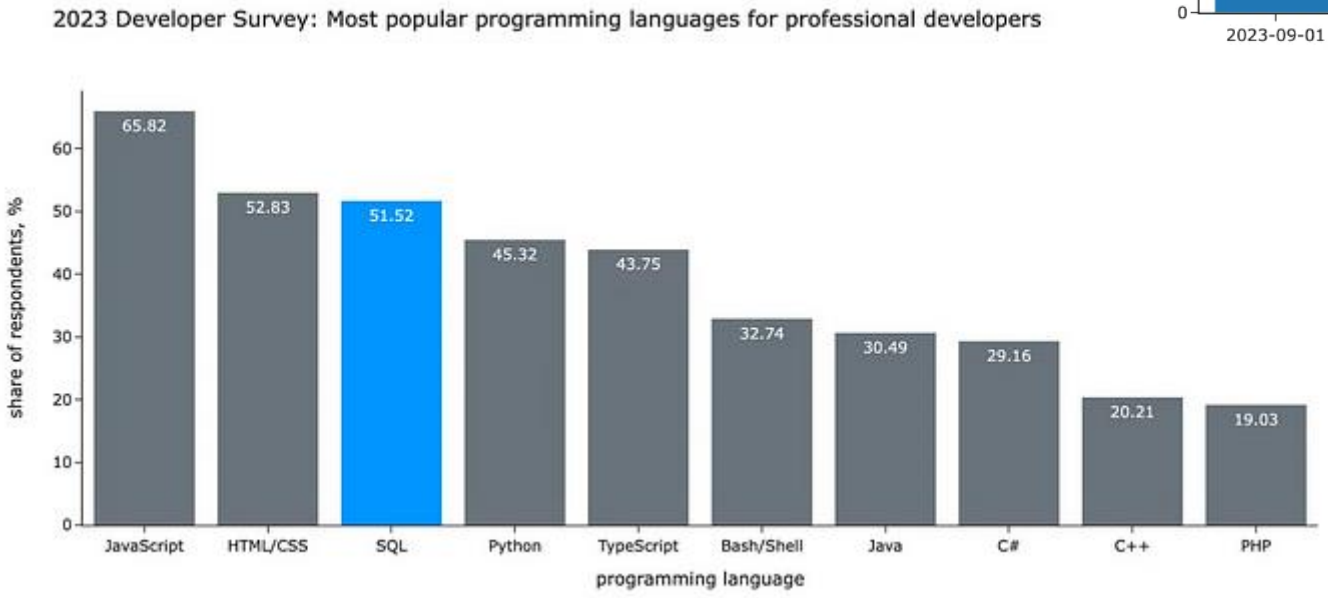
Label / Annotation

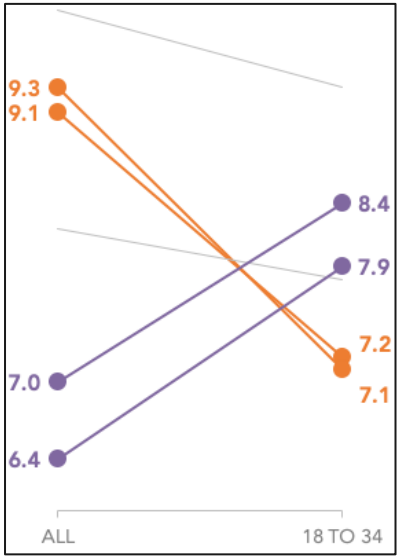


https://www.storytellingwithdata.com/blog/what-is-a-unit-chart

Color / Highlight

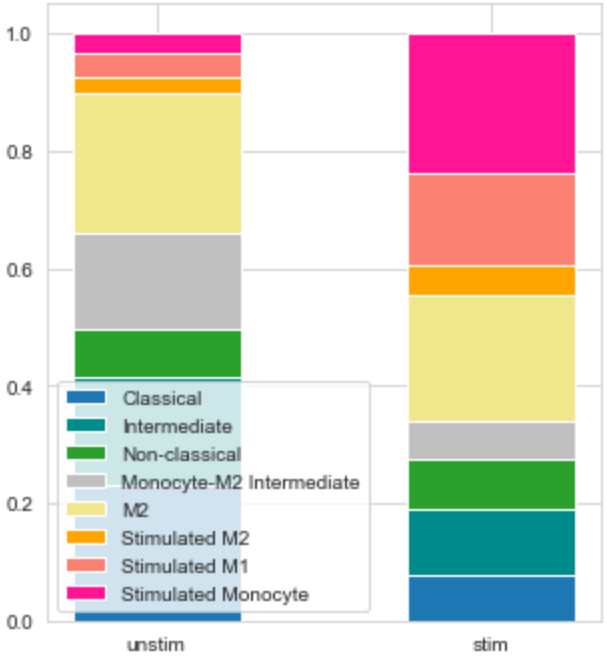


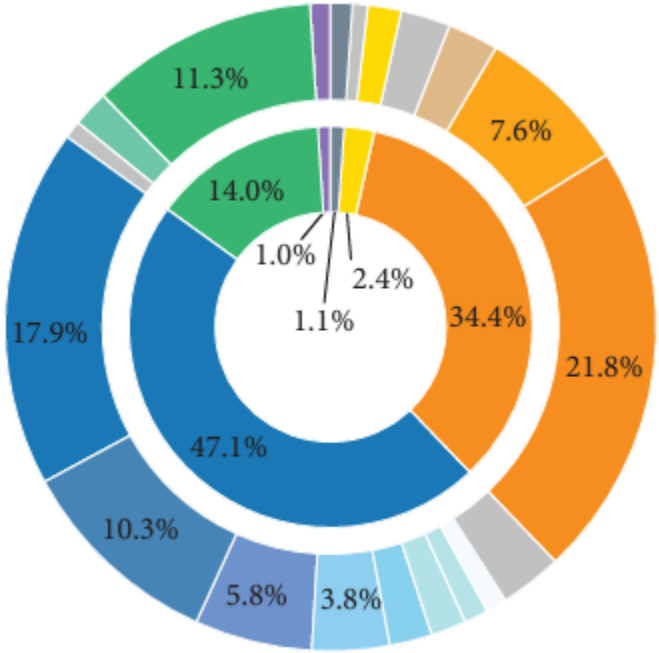




https://towardsdatascience.com/data-visualisation-101-playbook-for-attention-grabbing-visuals-c5b9faa7a950

Color / Highlight



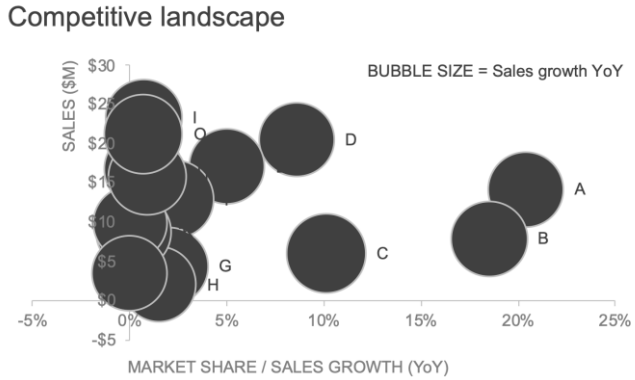


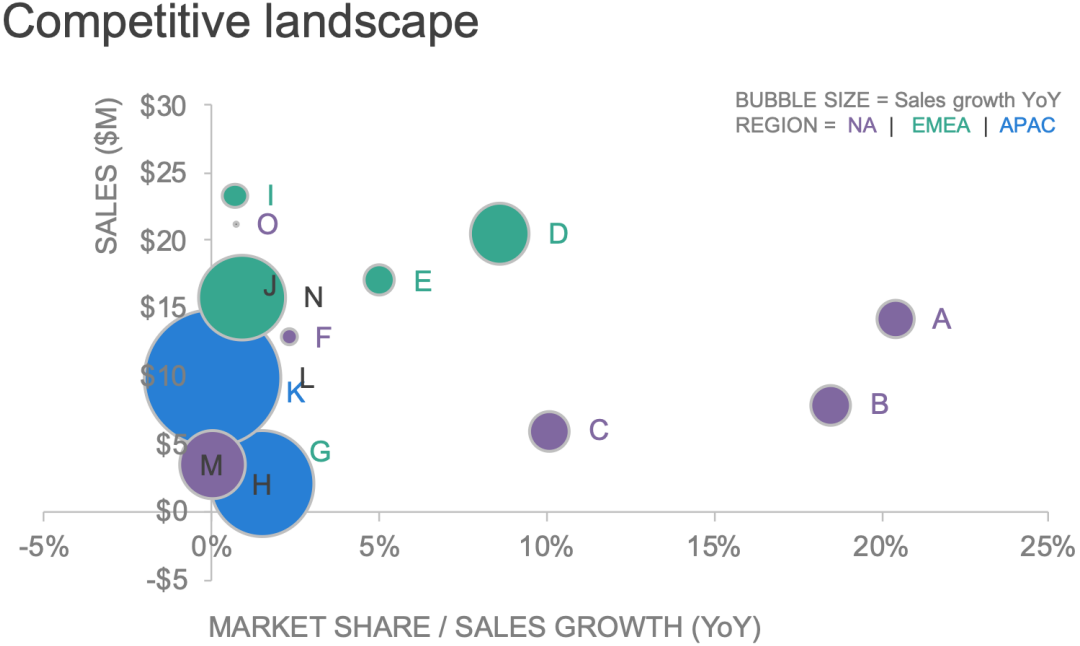
Phylum (inner circle)
- Fusobacteria
- Proteobacteria
- Firmicutes
- Bacteroidetes
- Actinobacteria
- Other phyla

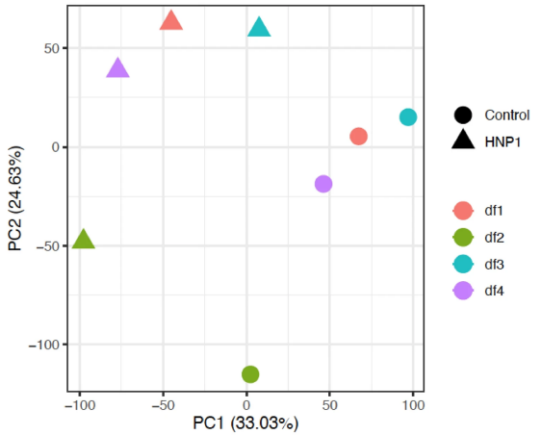
Family (outer circle)
- Fusobacteria: Fusobacteriaceae
- Fusobacteria: Other
- Proteobacteria: Enterobacteriaceae
- Proteobacteria: Sutterellaceae
- Proteobacteria: Other
- Firmicutes: Lachnospiraceae
- Firmicutes: Ruminococcaceae
- Firmicutes: Veillonellaceae
- Firmicutes: Acidaminococcaceae
- Firmicutes: Oscillospiraceae
- Firmicutes: Peptostreptococcaceae
- Firmicutes: Eubacterium coprostanoligenes
- Firmicutes: Selenomonadaceae
- Firmicutes: Other

Marker / Shape

Bubble sizes are too similar





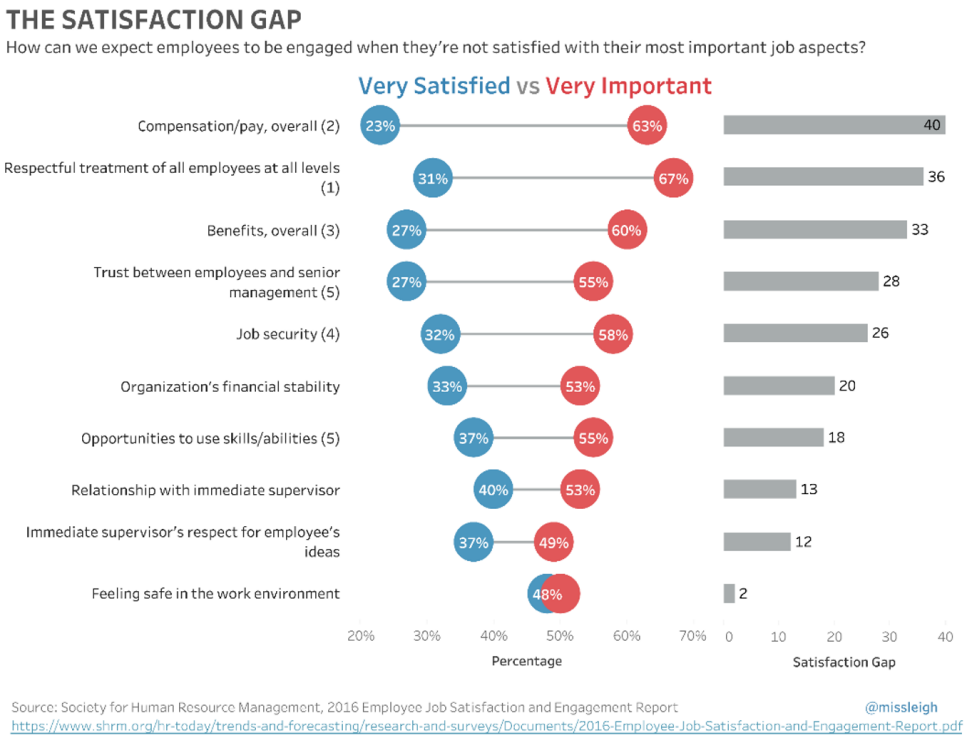


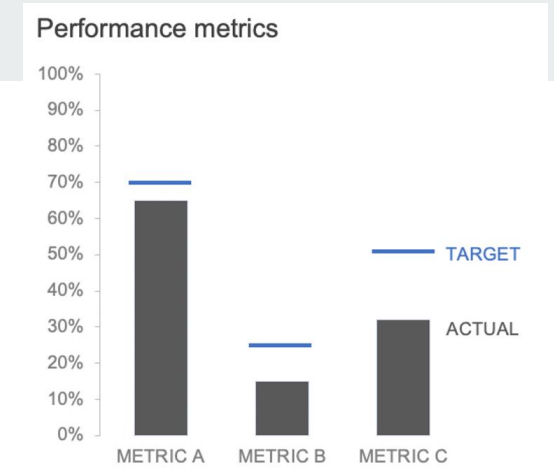
https://www.storytellingwithdata.com/blog/2021/5/11/what-is-a-bubble-chart

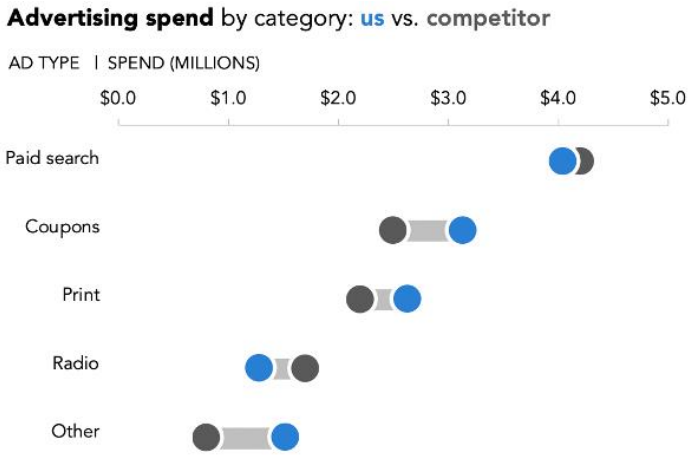
Zoom / Scale

https://www.storytellingwithdata.com/blog/2020/3/24/what-is-a-line-graph

Reference / Normalization

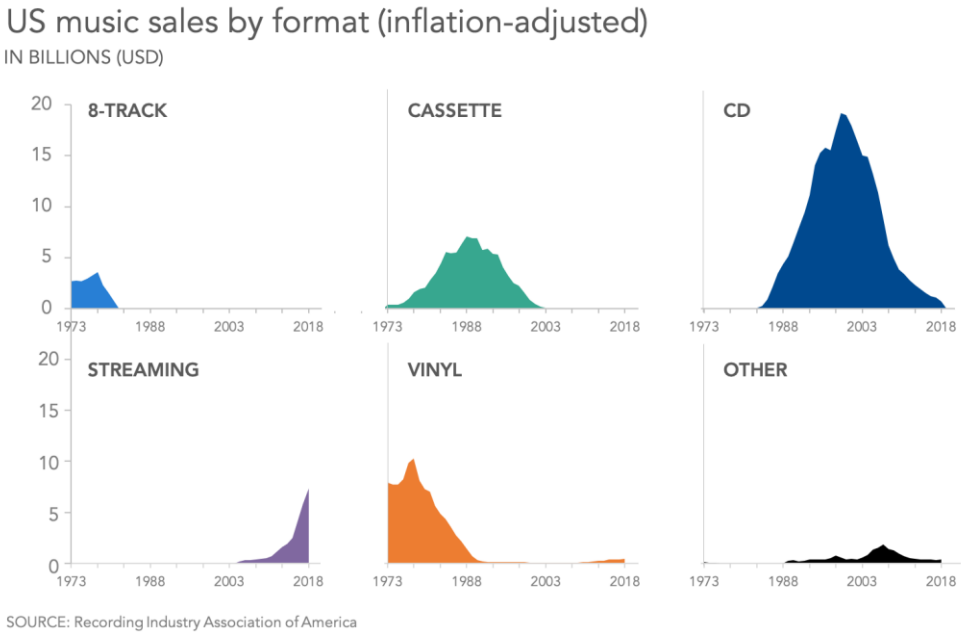


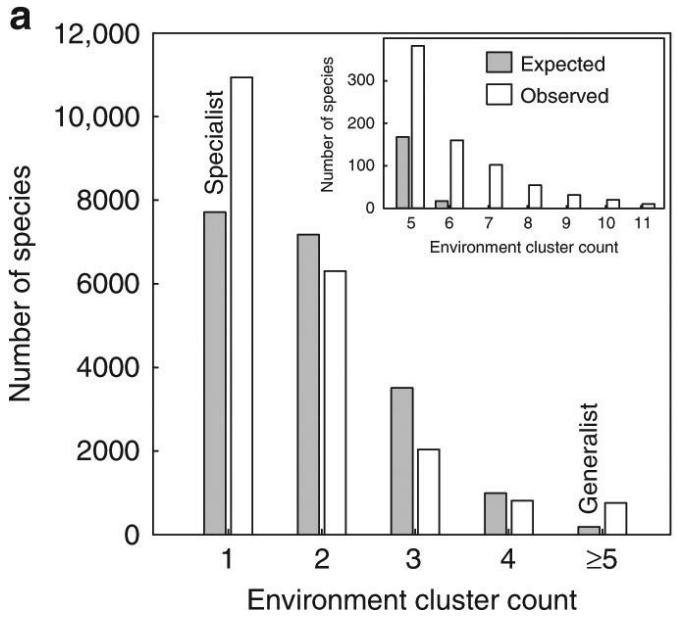




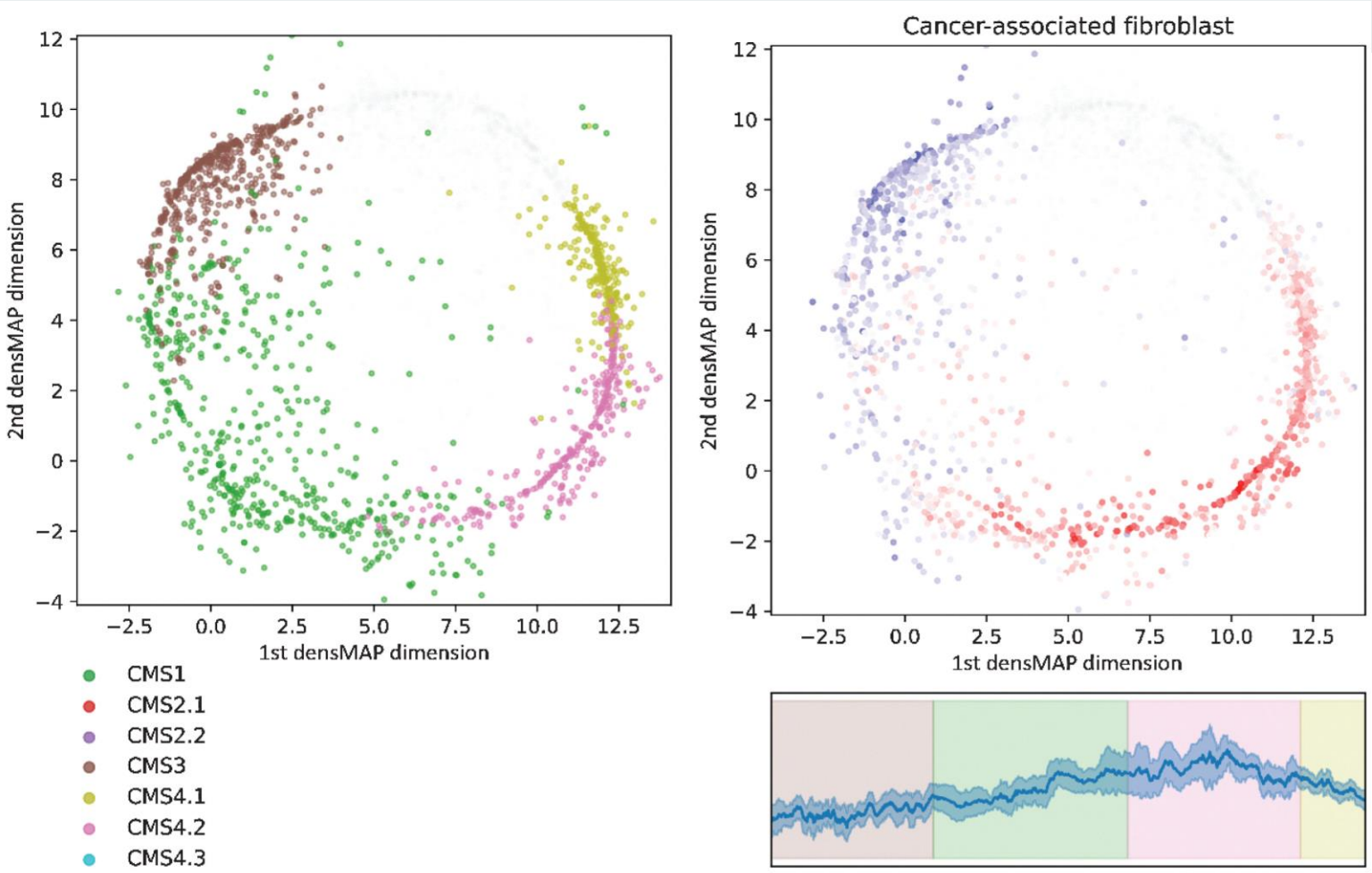
https://www.storytellingwithdata.com/blog/2020/12/9/what-is-a-dot-plot

Multiplot / Inset



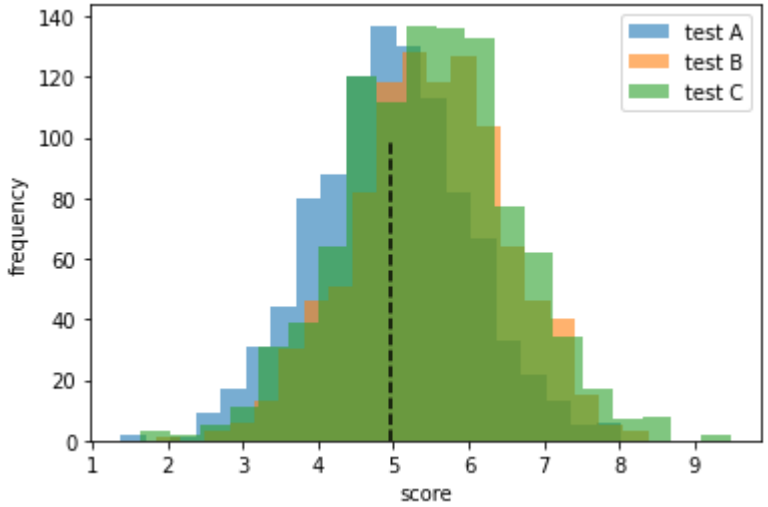


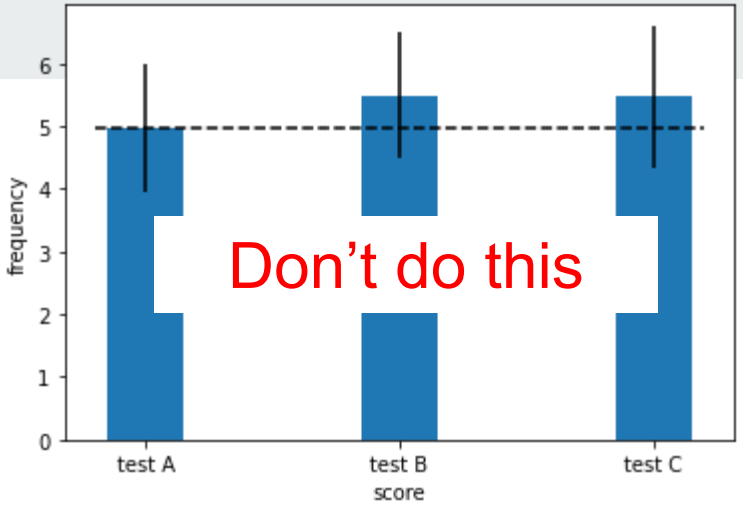
https://www.storytellingwithdata.com/blog/2020/4/9/what-is-an-area-graph

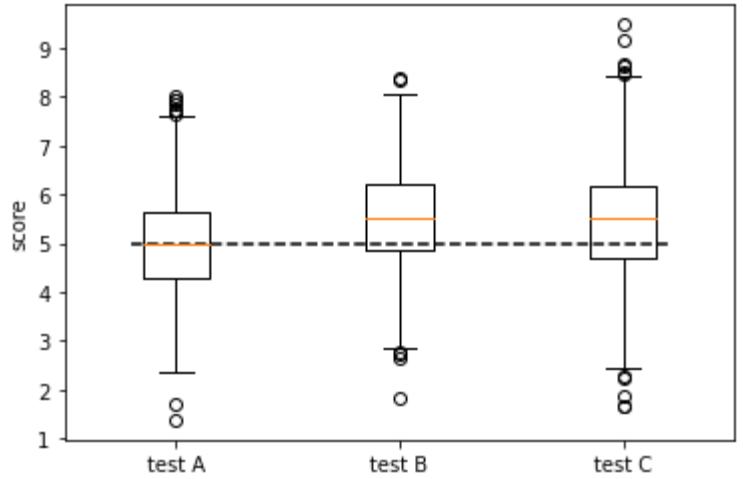

2nd densMAP dimension
1st densMAP dimension
CMS1
CMS2.1
CMS2.2
CMS3
CMS4.1
CMS4.2
CMS4.3
Cancer-associated fibroblast
2nd densMAP dimension
1st densMAP dimension

Some more examples

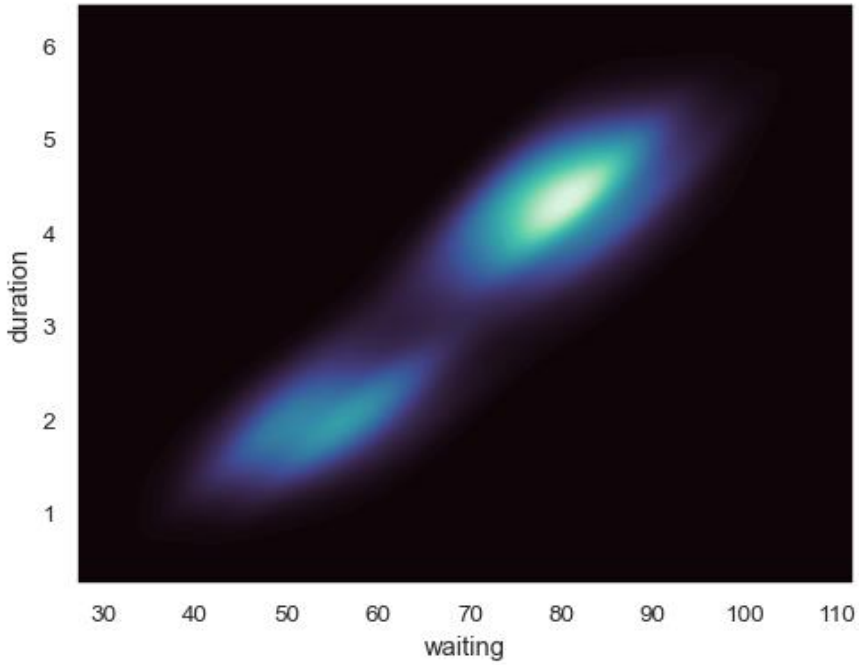
Distribution

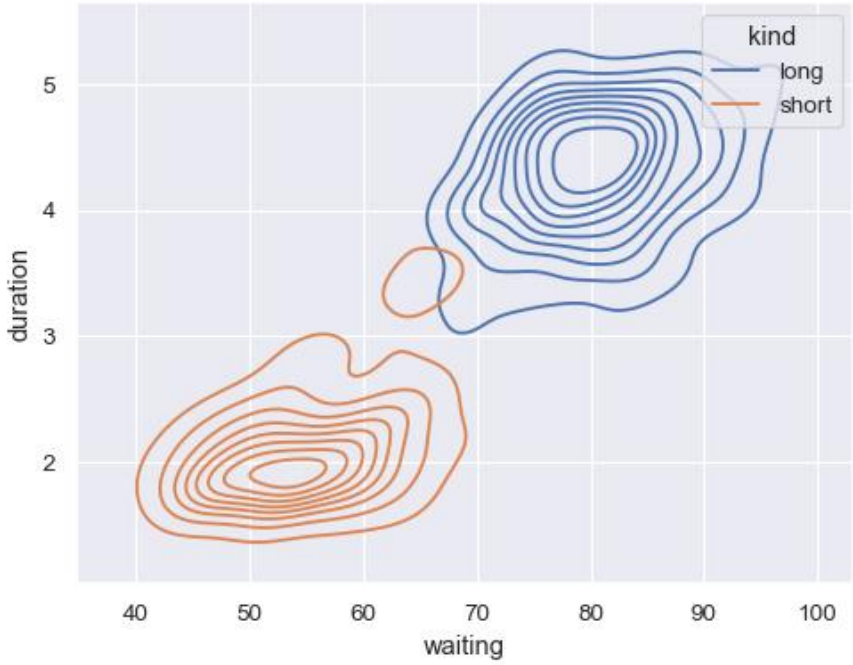






Kernel Density Estimate (KDE) plot

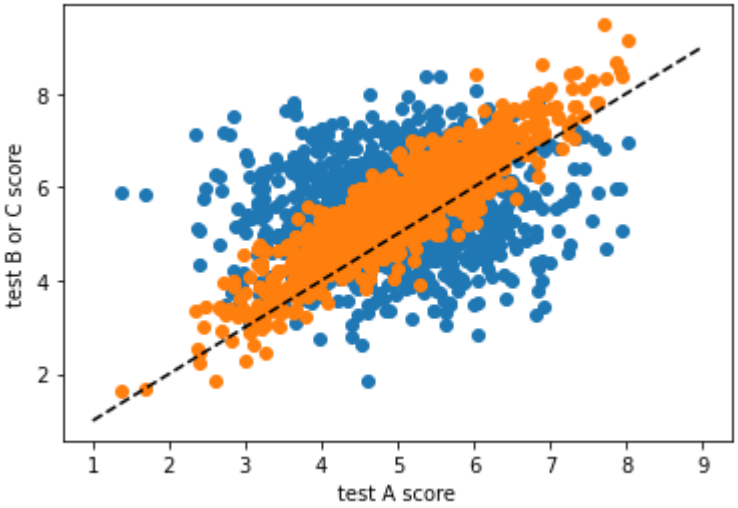


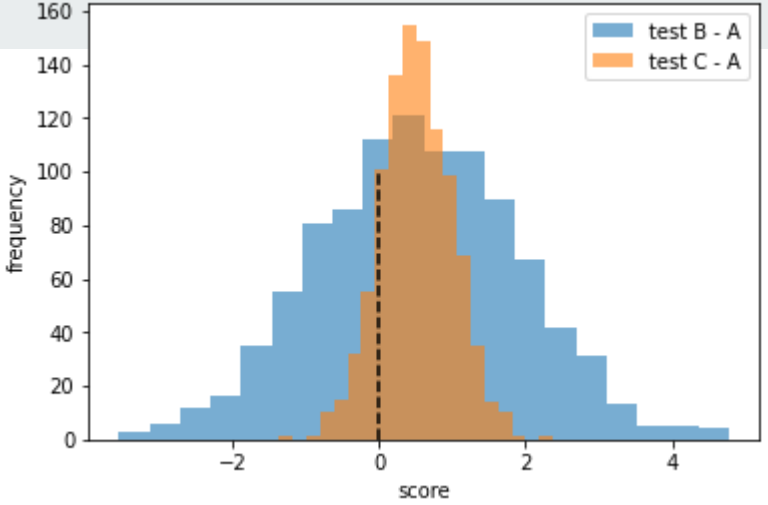


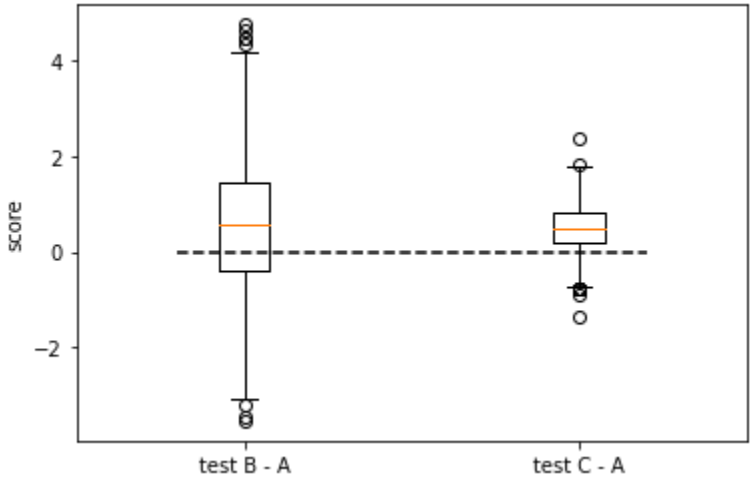
https://seaborn.pydata.org/generated/seaborn.kdeplot.html

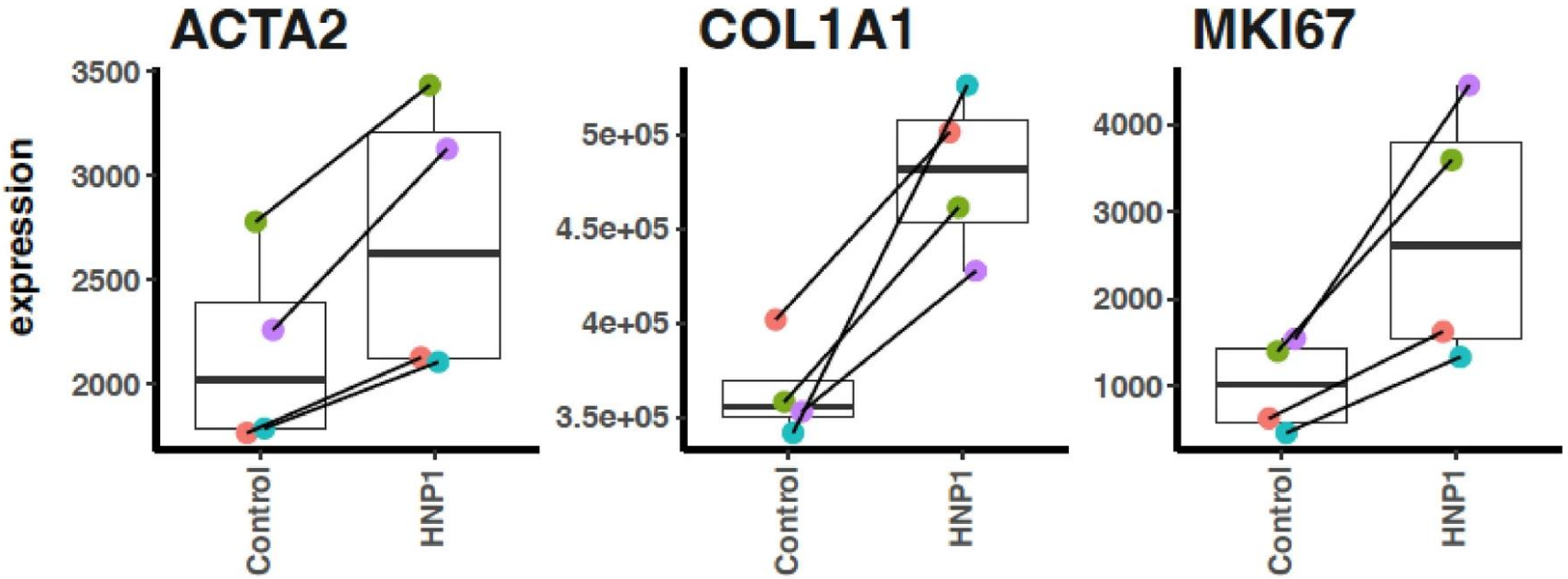
Paired vs unpaired data

Show paired differences









- Box plot of raw values is not appropriate for paired data
- Line plot and color coding for the four patients were added

Dot plot

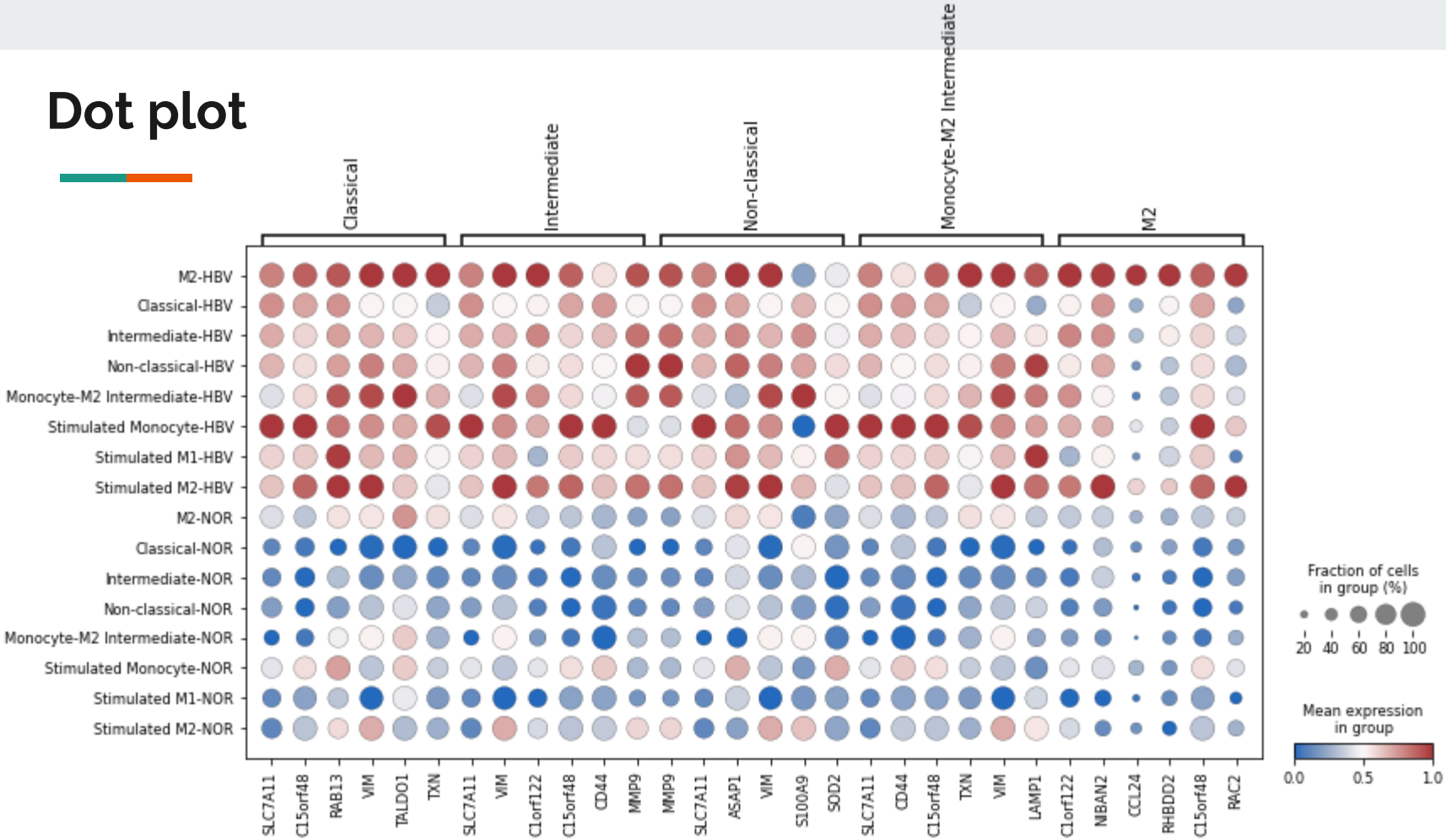
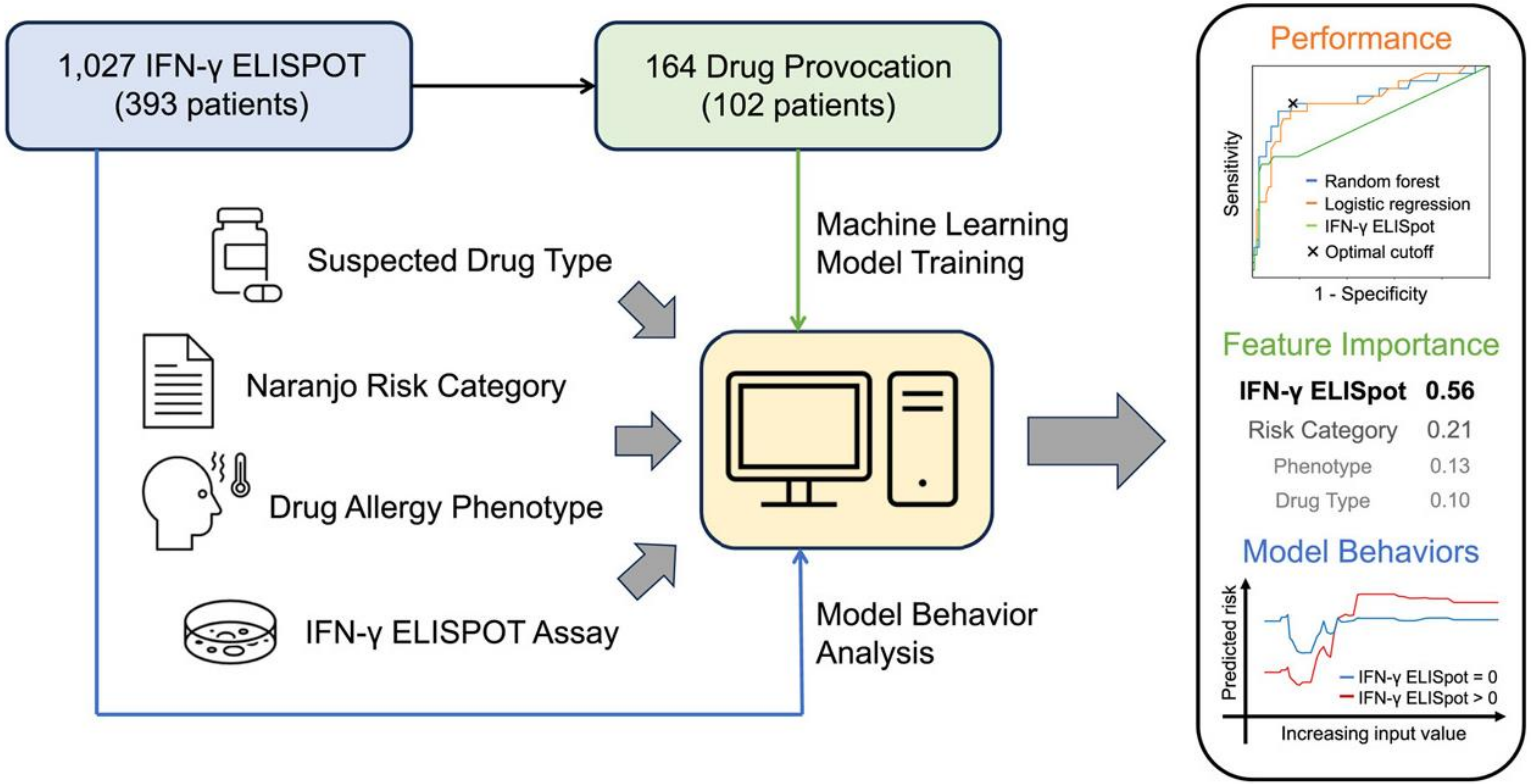
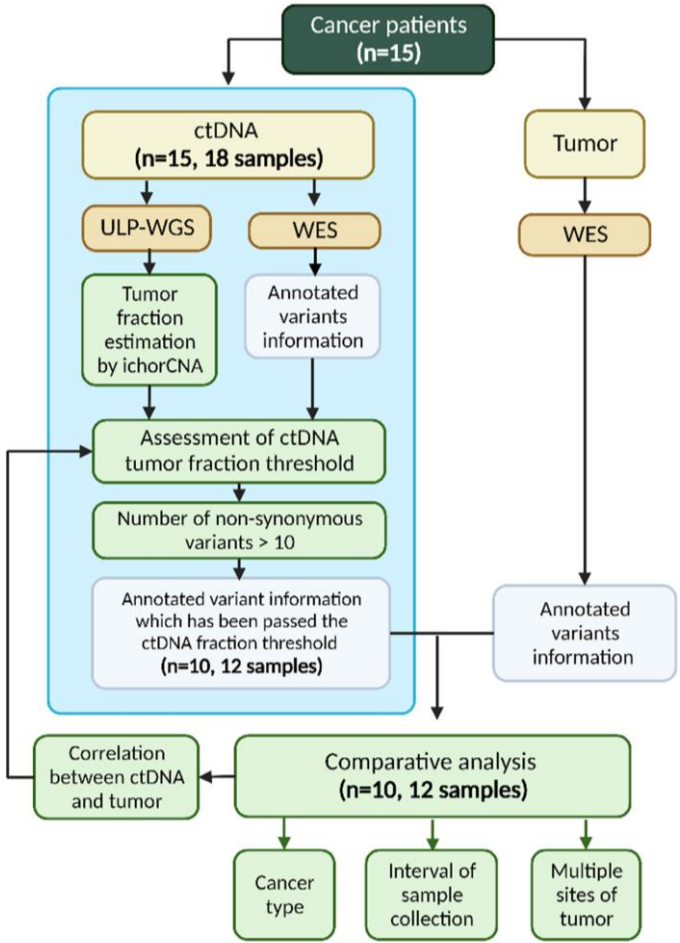

Classical
Intermediate
Non-classical
Monocyte-M2 Intermediate
M2
M2-HBV
Classical-HBV
Intermediate-HBV
Non-classical-HBV
Monocyte-M2 Intermediate-HBV
Stimulated Monocyte-HBV
Stimulated M1-HBV
Stimulated M2-HBV
M2-NOR
Classical-NOR
Intermediate-NOR
Non-classical-NOR
Monocyte-M2 Intermediate-NOR
Stimulated Monocyte-NOR
Stimulated M1-NOR
Stimulated M2-NOR
SLC7A11
C15orf48
RAB13
VIM
TALDO1
TXN
SLC7A11
VIM
C1orf122
C15orf48
CD44
MMP9
MMP9
SLC7A11
ASAP1
VIM
S100A9
SOD2
SLC7A11
CD44
C15orf48
TXN
VIM
LAMP1
C1orf122
NIBAN2
CCL24
RHBDD2
C15orf48
RAC2
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0.0 0.5 1.0

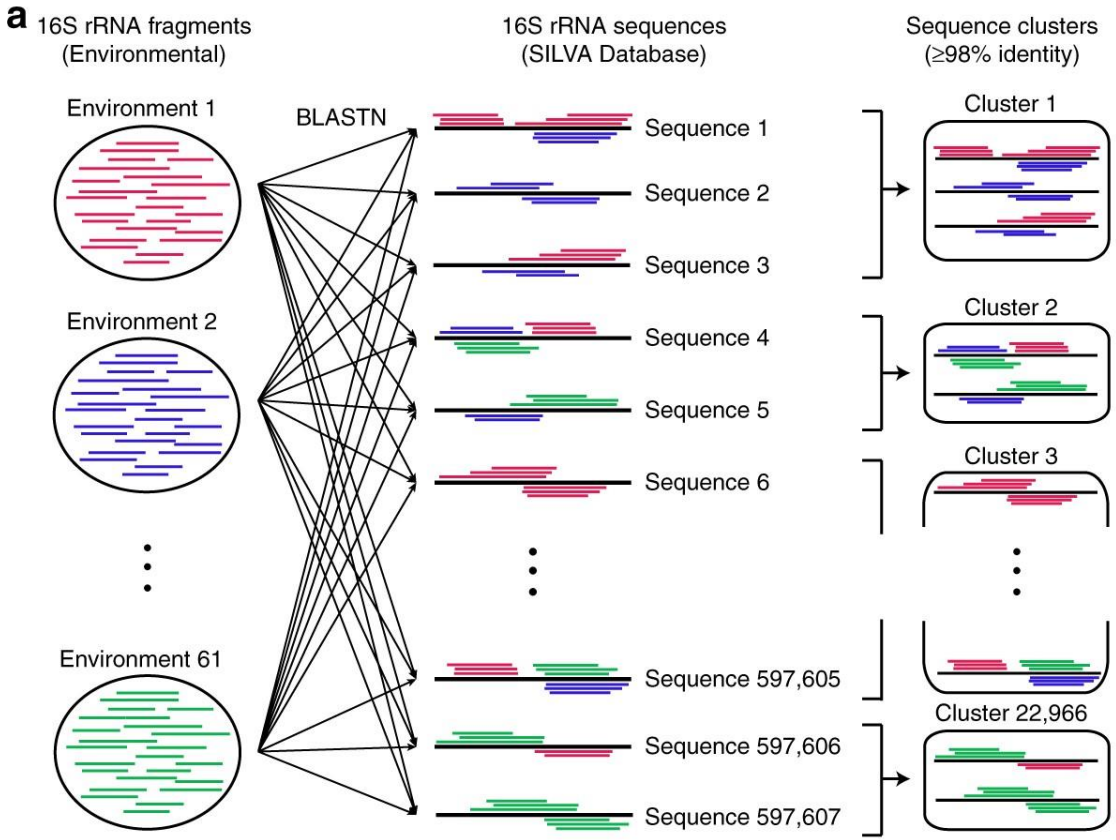
Diagram / Conceptual figure

Mix of real data, graph, and cartoon





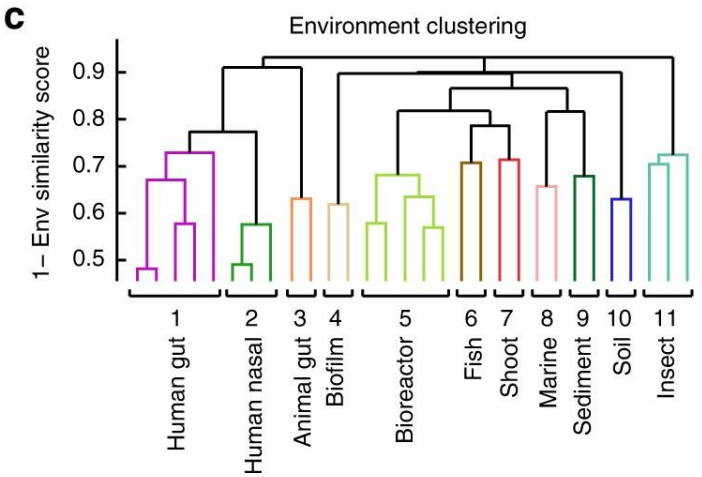
- Top-to-bottom hierarchy
- Color coding
- Grouping of related steps



b

Environment profiles of aligned 16S rRNA fragments

Sequence Cluster	Env 1	Env 2	• • •	Env 61
Cluster 1	23	12	• • •	
Cluster 2		33	• • •	21
⋮	⋮	⋮		⋮
Cluster 22,966			• • •	13



Any questions?

See you on February 9th