

---

# 3050571 Practical Clin Data Sci

## Session 5: 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

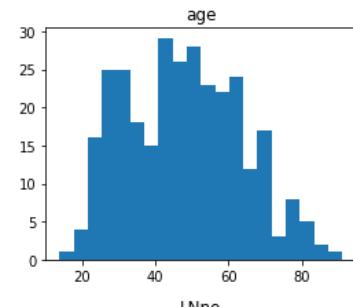
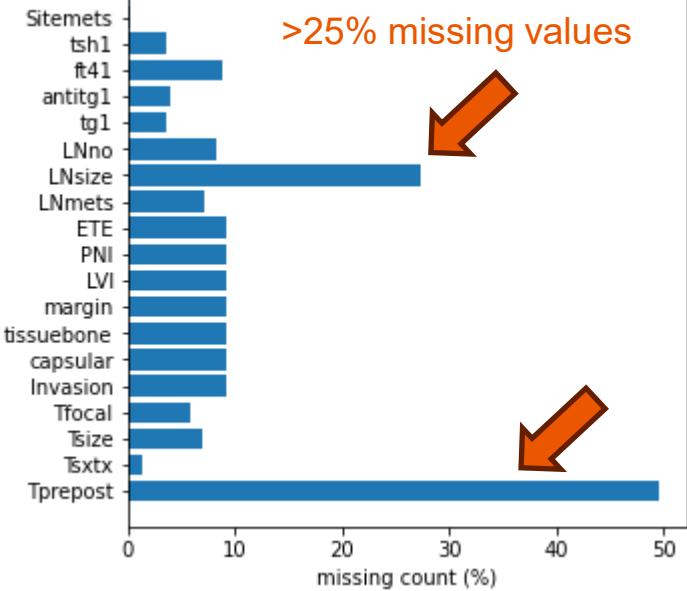
---

- **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

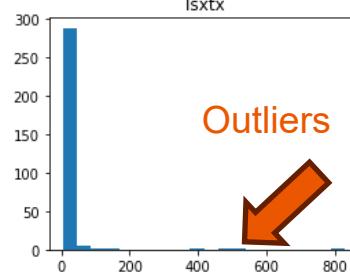


>25% missing values

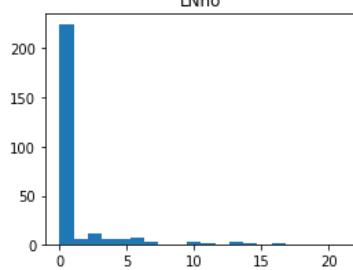
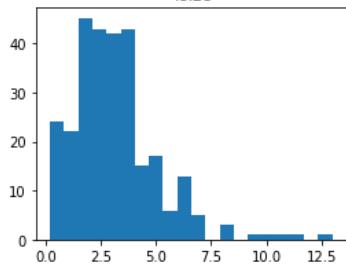


Tsxtx

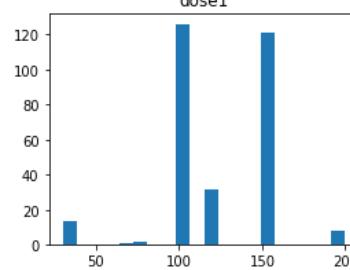
Outliers



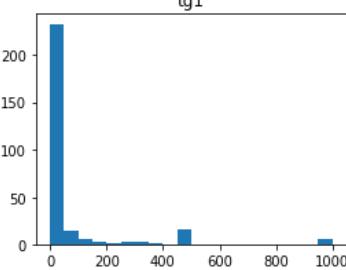
Tsize



LNno

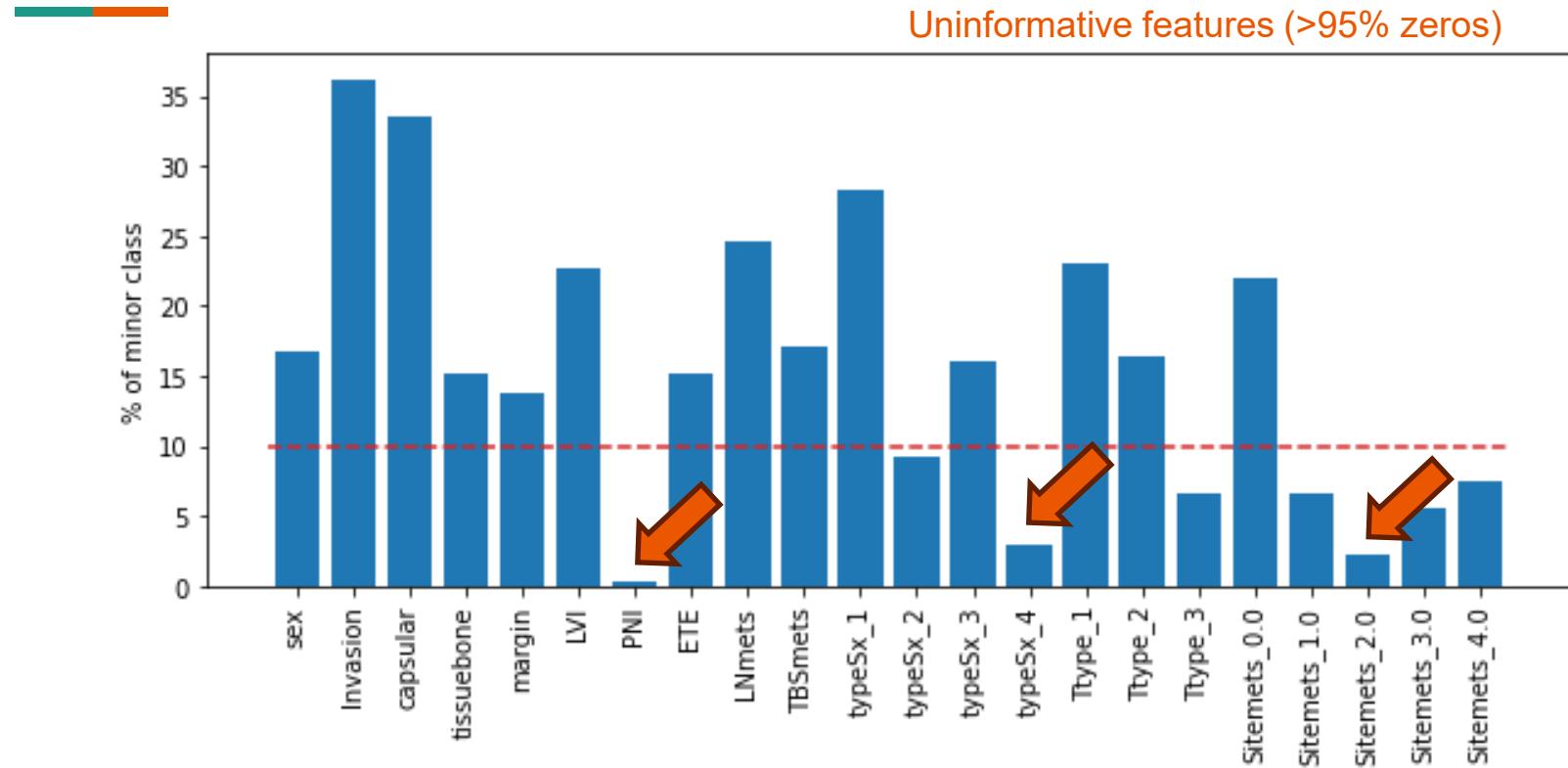


dose1



tg1

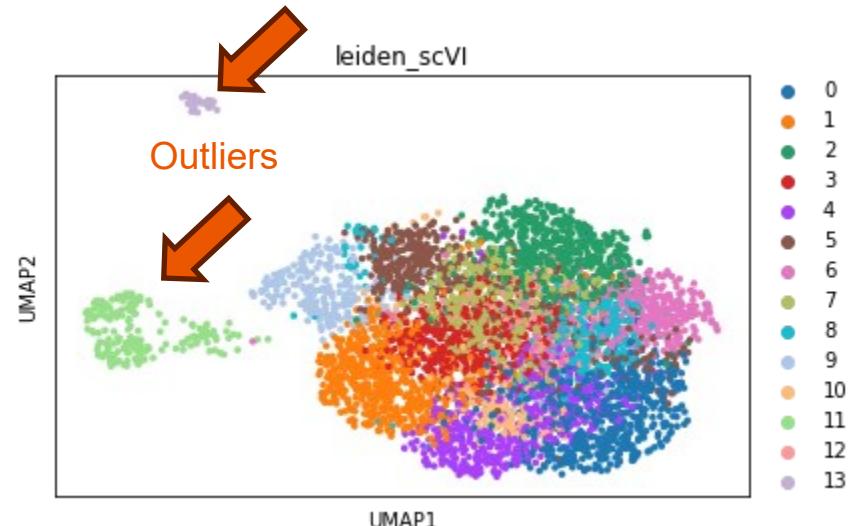
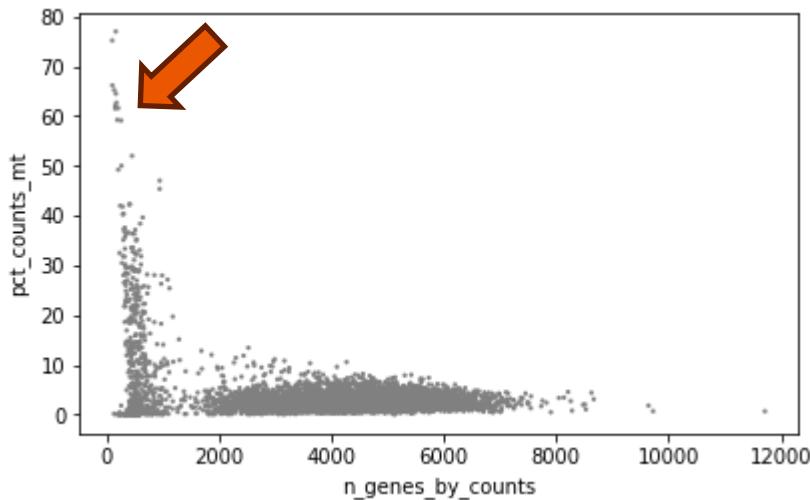
# Identify bad features



# Identify outliers and data clusters



High % MT expression = signal of dead cells



# Catalog usable data

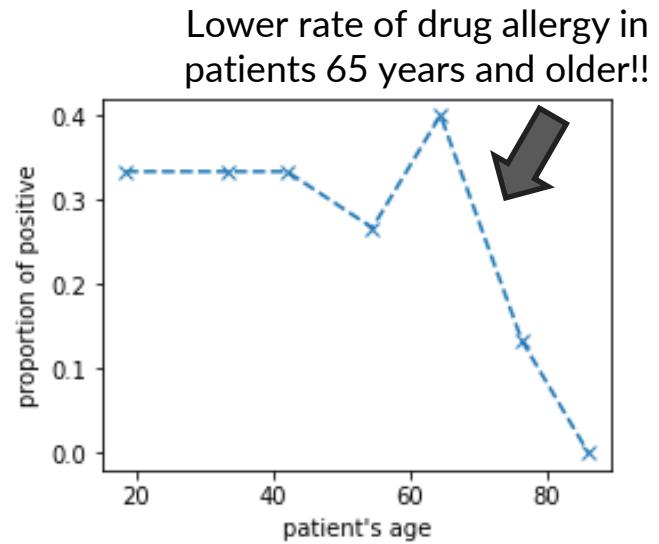
---

- **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

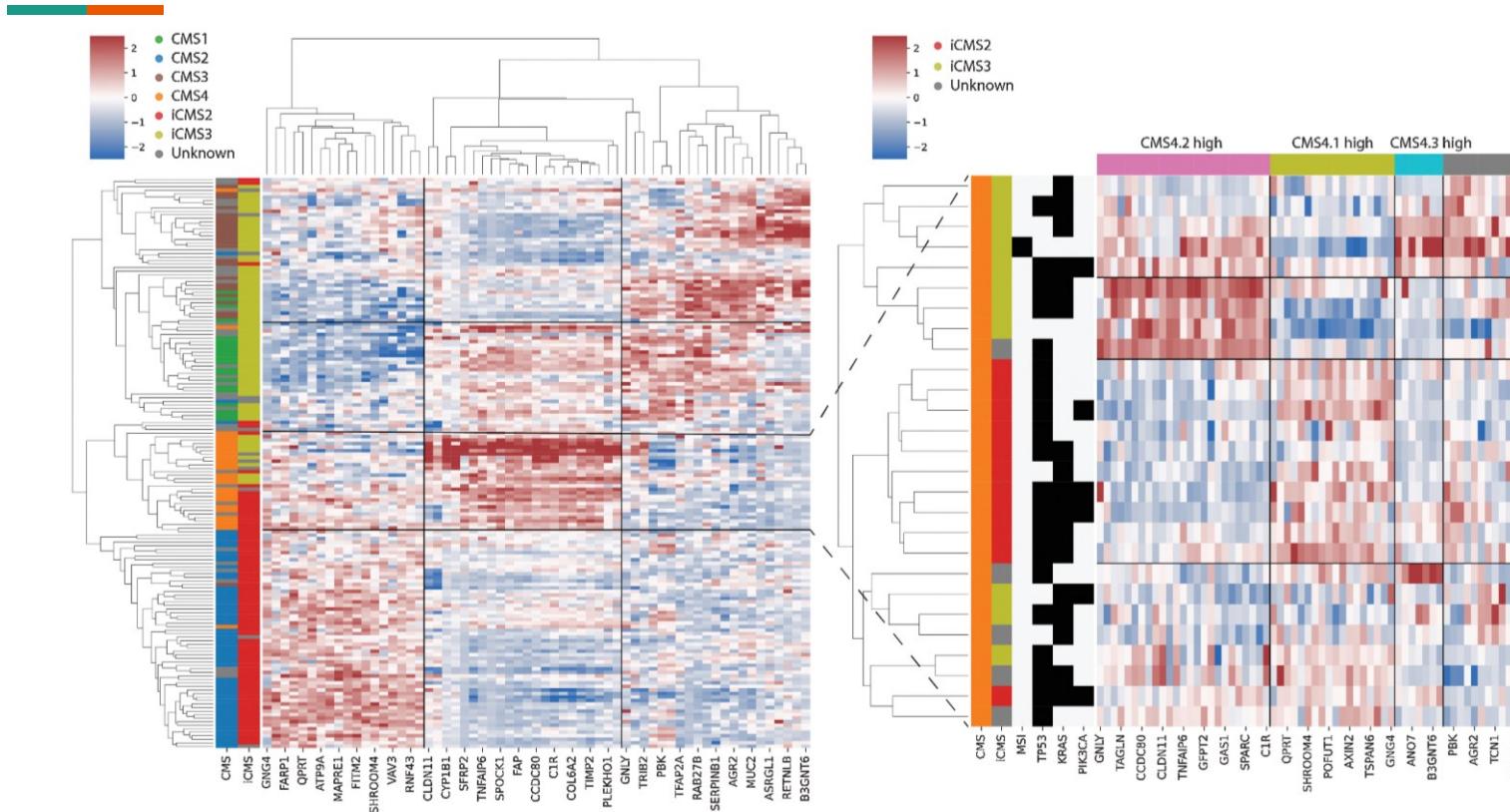


# Hypothesis development & testing

---

- 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





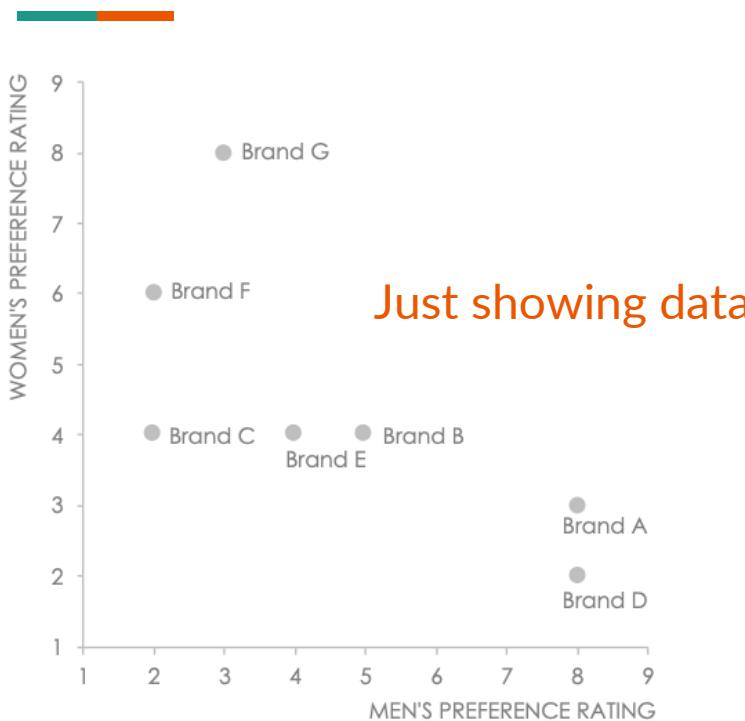
# Visualization

# Key rules for data visualization

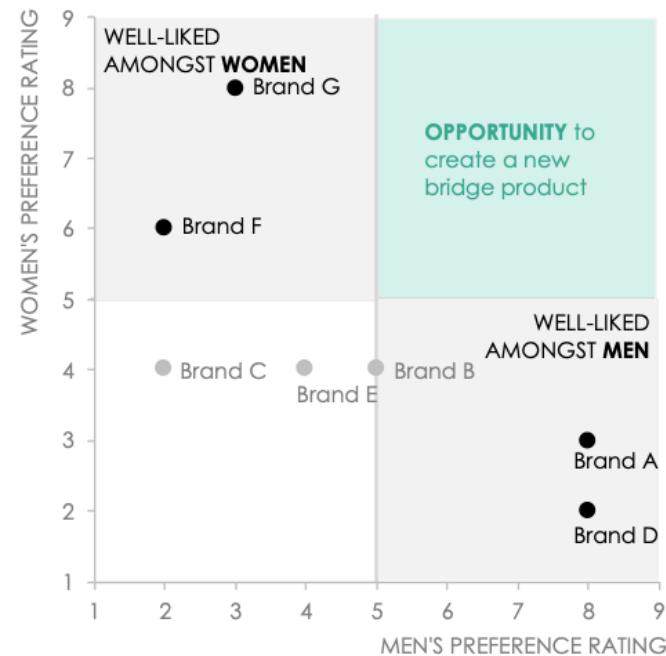
---

- 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

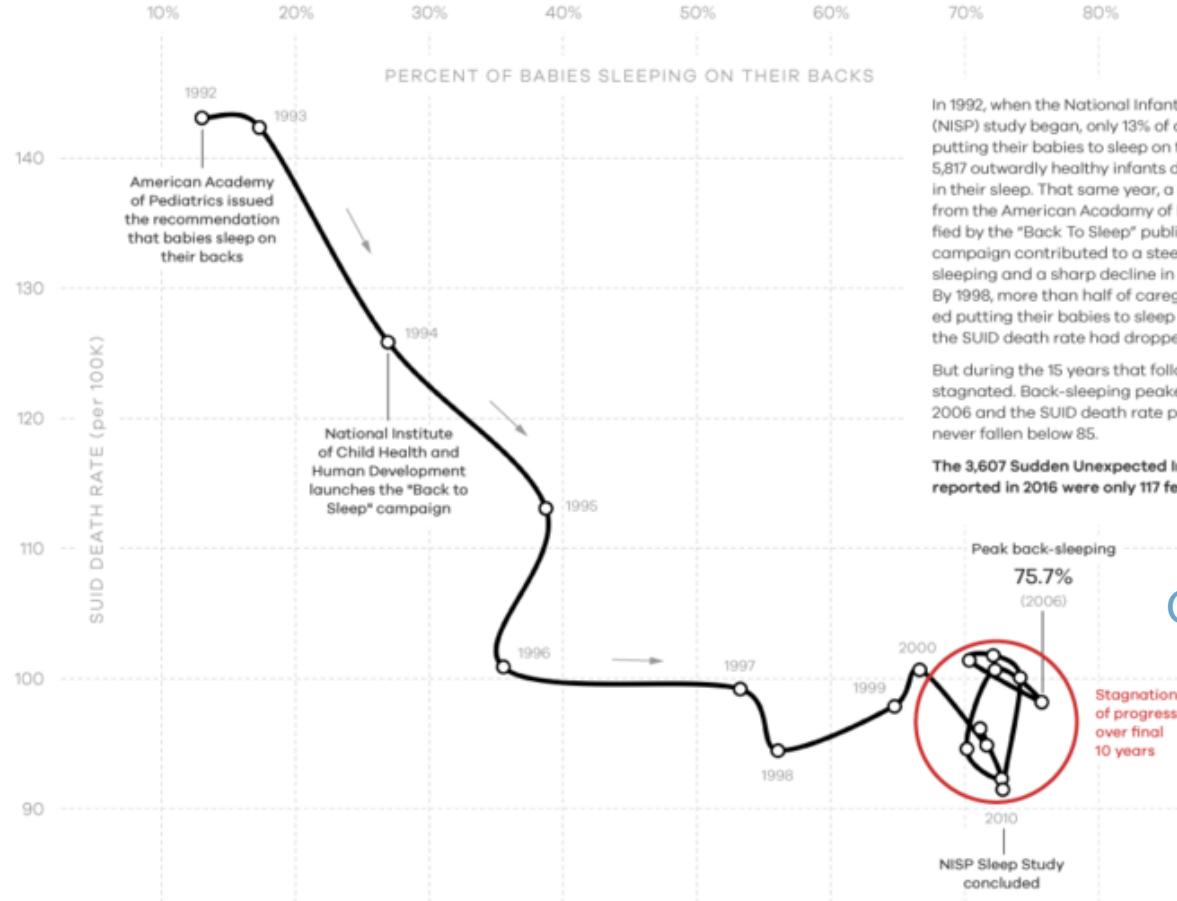


Clear message



# 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.

Peak back-sleeping  
75.7%  
(2006)

Stagnation of progress over final 10 years

NISI Sleep Study concluded

Clear message

# Remove noise

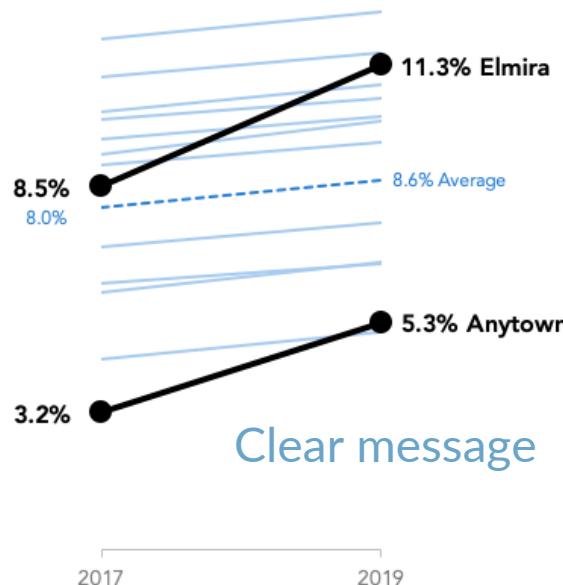


<https://www.storytellingwithdata.com/blog/2020/5/14/what-is-a-pie-chart>

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

# Focus attention

PERCENTAGE OF PATIENTS WITH DIABETES



## Too many messages

Of the 99 patients in our study...

**70.7%**

70 people prefer **Serentiva**

**19.2%**

19 people prefer the **competitor**

PREFERRED DUE TO:

**Lack of mood swings**  
(N=34, 49%)

**Better anxiety relief**  
(N=29, 41%)

**Other**  
(N=7, 10%)

PREFERRED DUE TO:

**Prefer mood swings**  
(N=8, 42%)

**Better anxiety relief**  
(N=8, 42%)

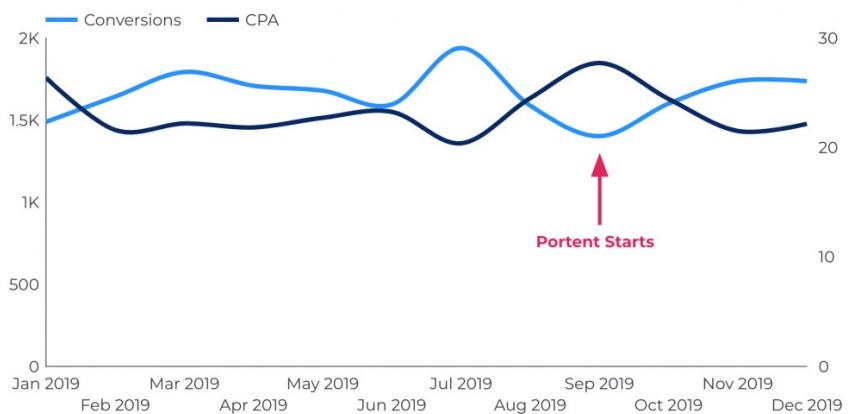
**Other**  
(N=3, 16%)

10.1% (10 people) didn't express a preference



# Provide context

Conversion Efficiency by Month



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% ←

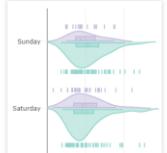
<https://gfchart.com/2022/08/adding-context-to-your-in-content-charts-and-graphs/>



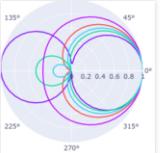
# Graph elements

# Plotly (<https://plotly.com/python/>)

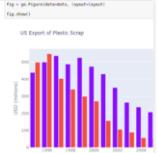
## Fundamentals



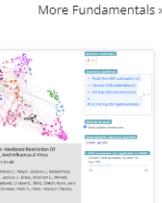
The Figure Data Structure



Creating and Updating Figures



Displaying Figures



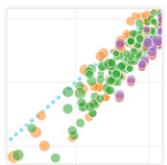
Plotly Express  
Analytical Apps with  
Key graph elements

More Fundamentals »

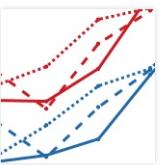
```
import plotly.express as px  
df = px.data.gapminder()
```

```
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()
```

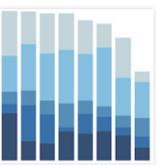
## Basic Charts



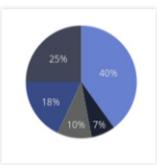
Scatter Plots



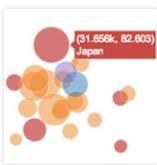
Line Charts



Bar Charts

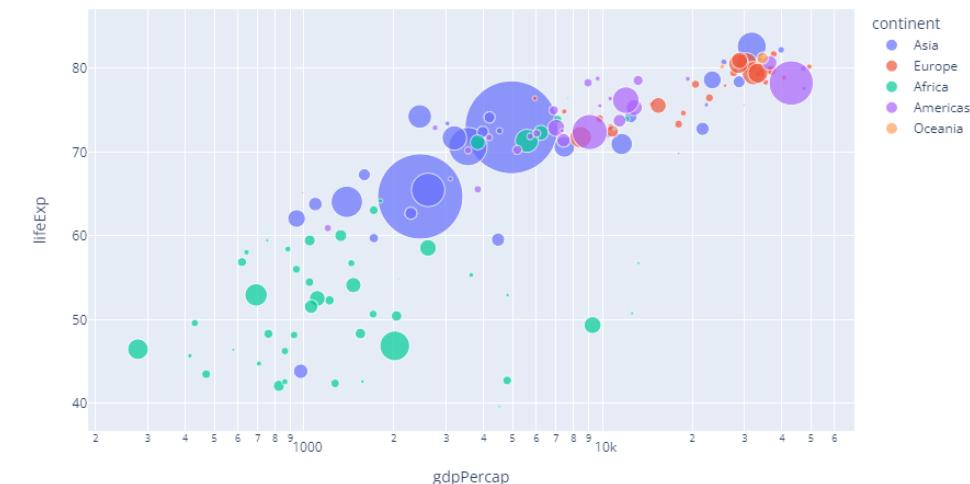


Pie Charts

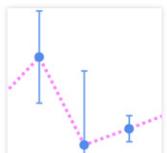


Bubble Charts

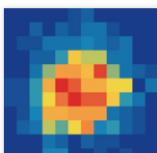
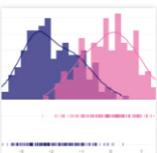
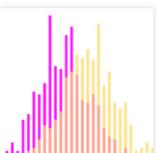
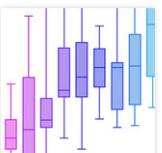
More Basic Charts »



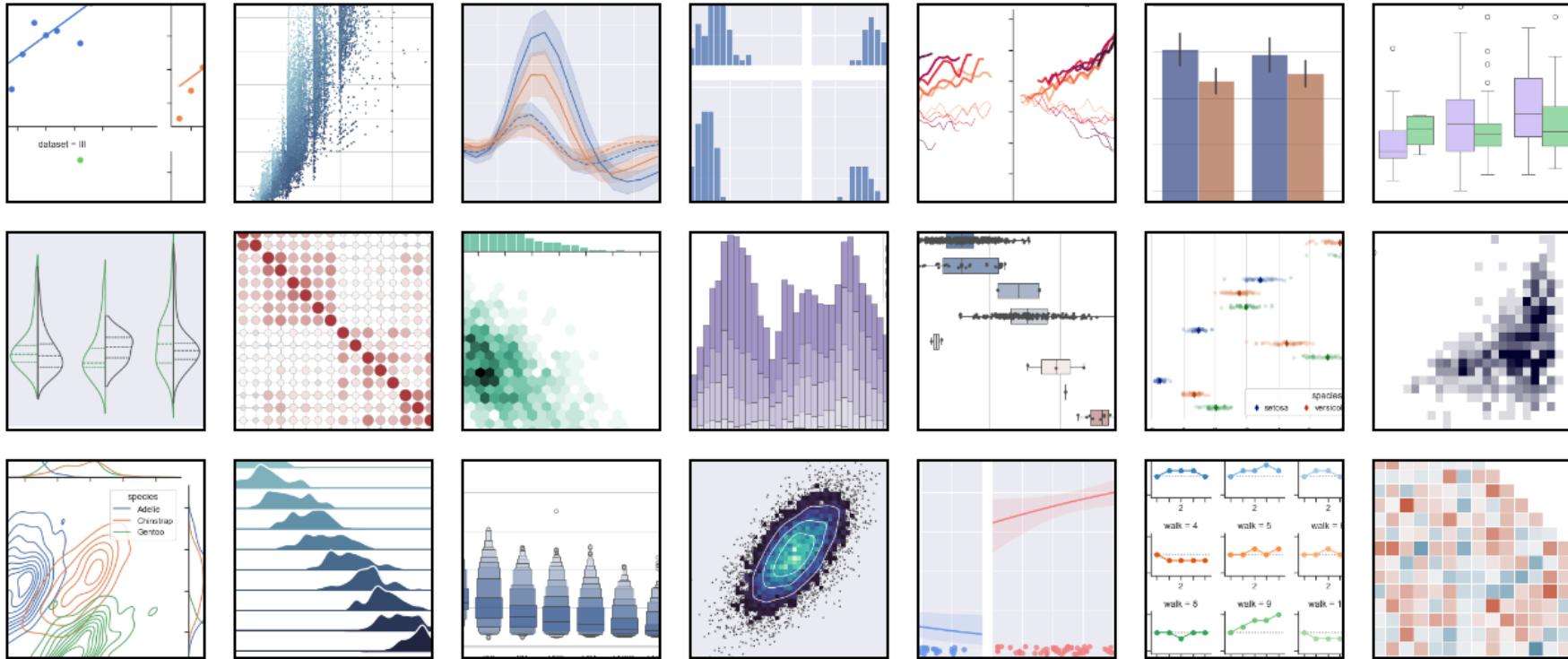
## Statistical Charts



More Statistical Charts »



# Seaborn (<https://seaborn.pydata.org>)



# Matplotlib (<https://matplotlib.org/stable/>)



```
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

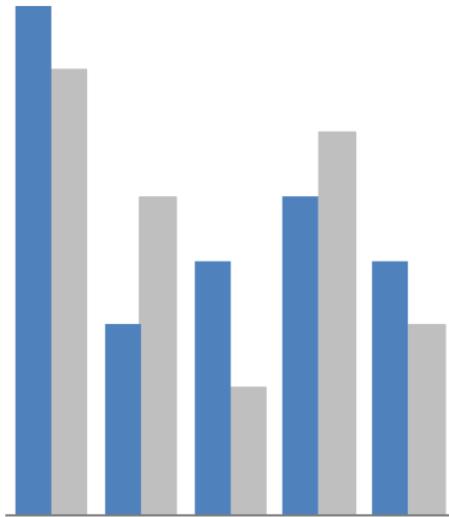
---

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

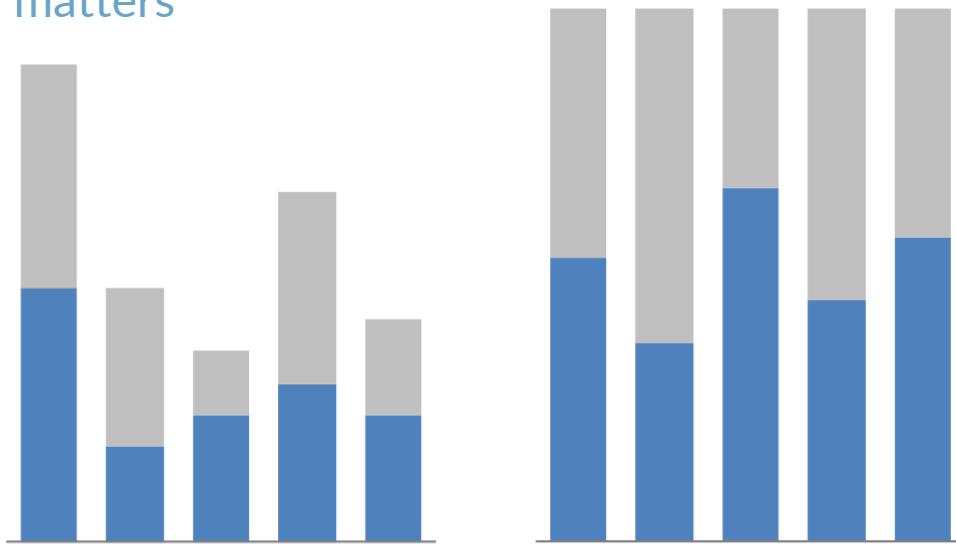
# Graph style



Magnitude matters



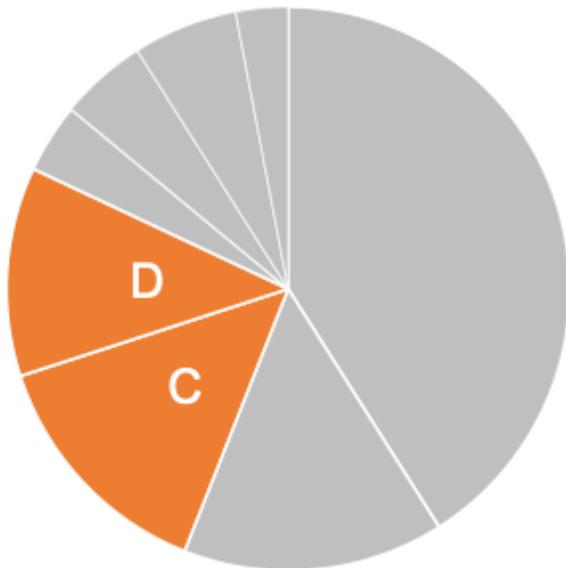
Focus on composition



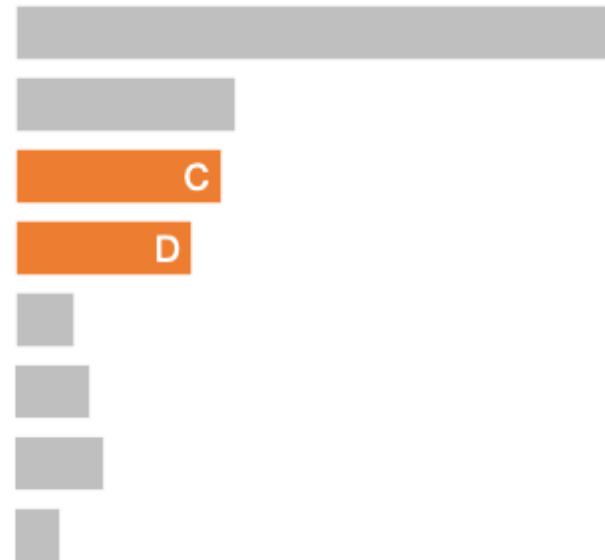
# Graph style



% OF TOTAL

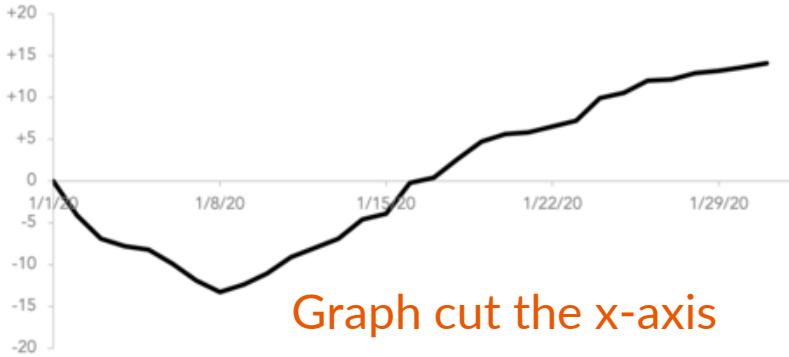


% OF TOTAL



# Axis

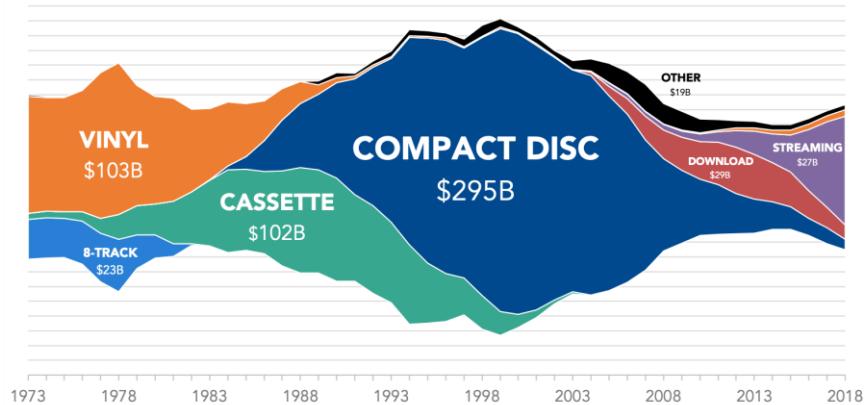
Change in subscriptions, year-over-year  
IN THOUSANDS OF ACTIVE SUBSCRIBERS



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

## No unit on the y-axis

US music sales by format (inflation-adjusted)  
EACH INTERVAL = \$1 BILLION (USD)

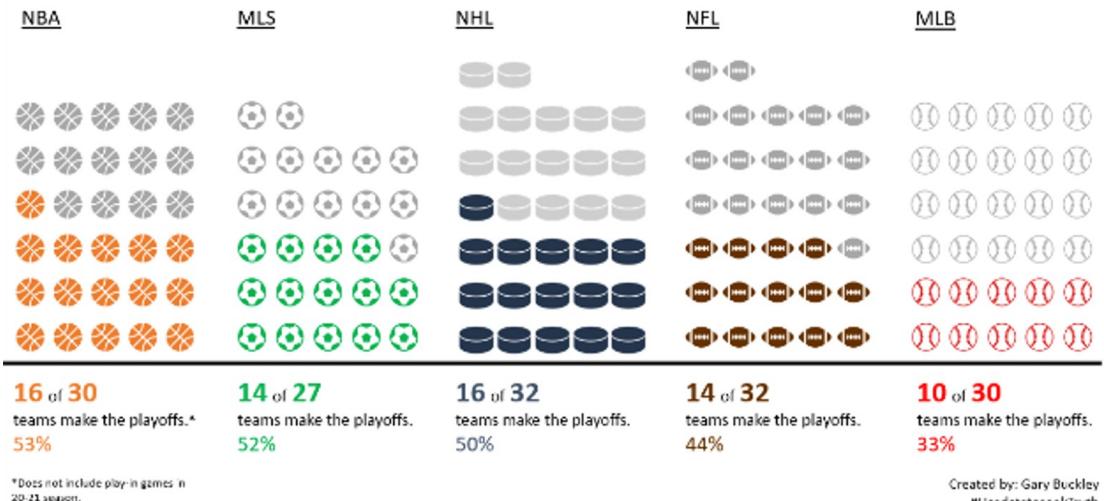
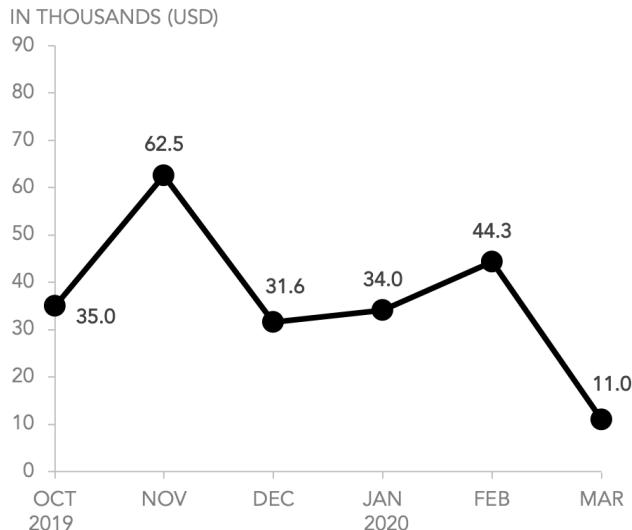


SOURCE: Recording Industry Association of America

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

# Label / Annotation

## 6-Month sales report and forecast



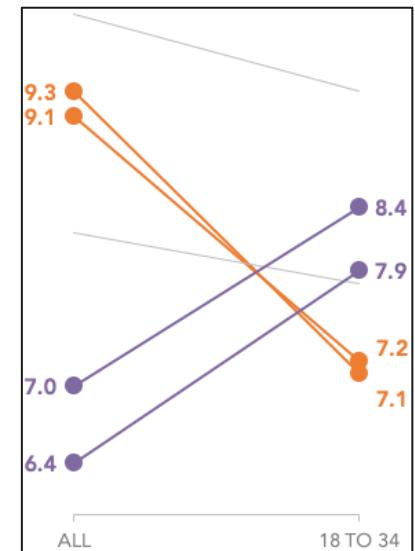
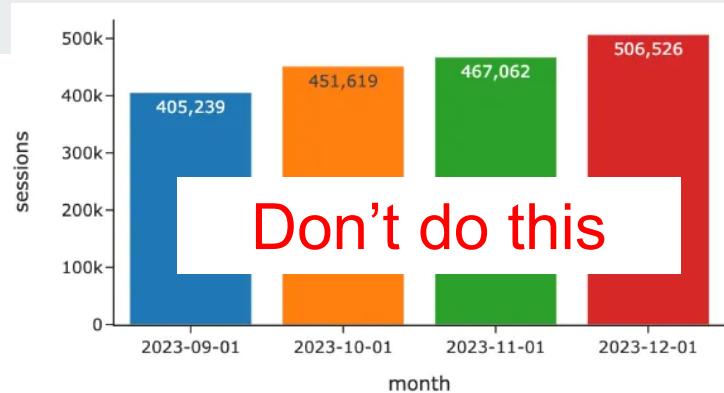
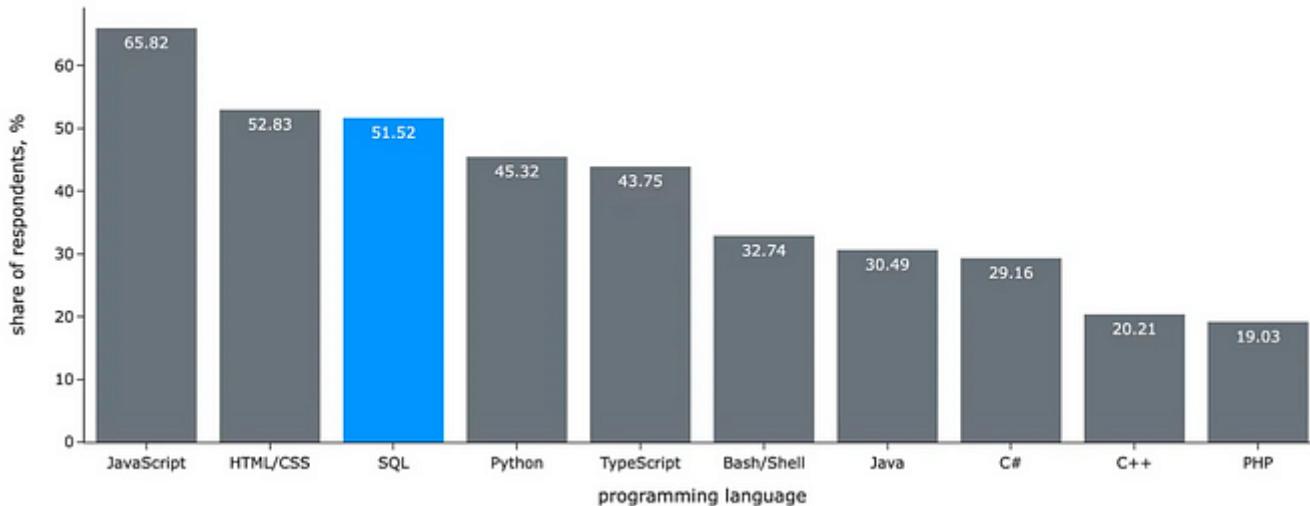
<https://www.storytellingwithdata.com/blog/what-is-a-unit-chart>

Created by: Gary Buckley  
#UsadatatoseekTruth

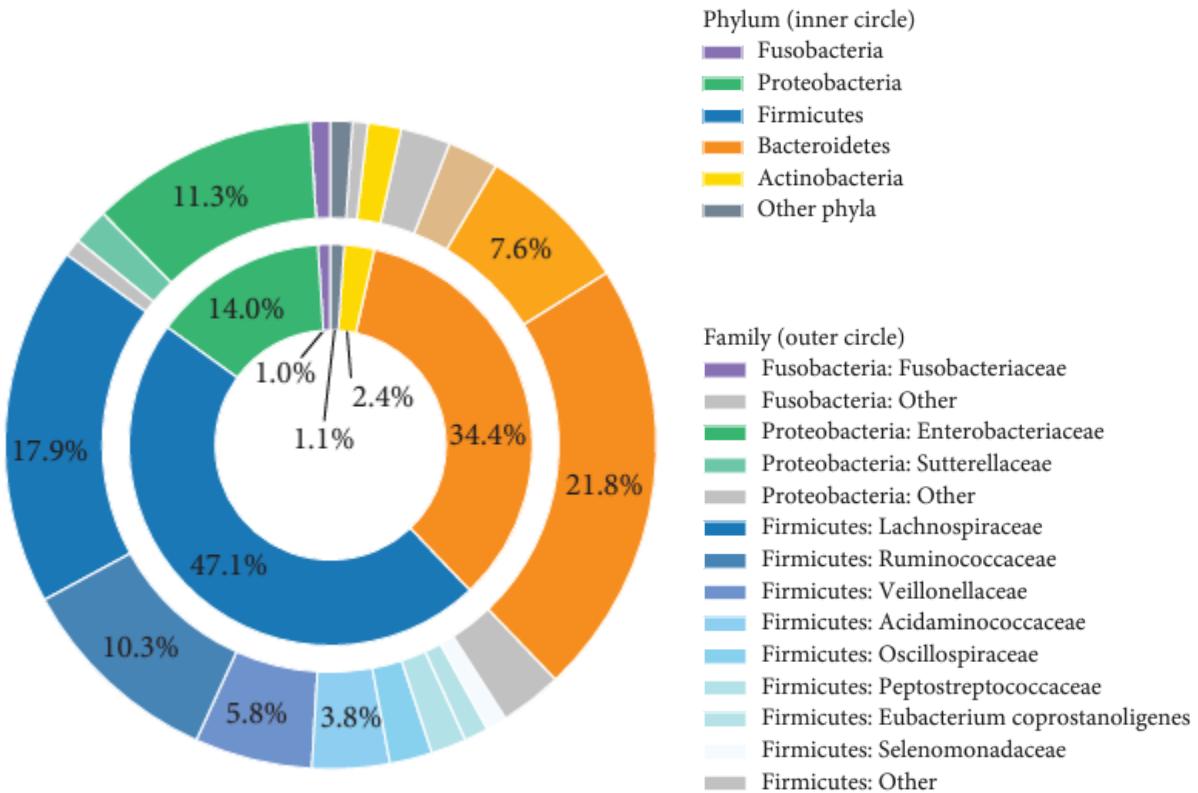
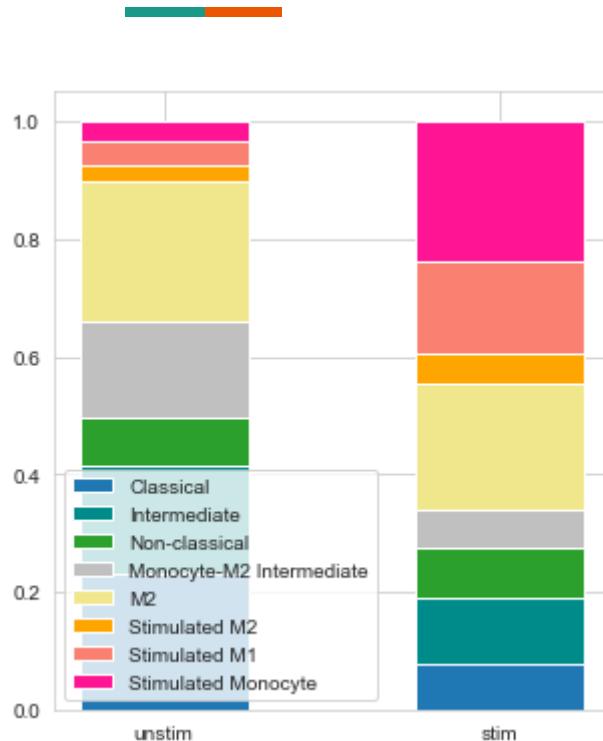
# Color / Highlight



2023 Developer Survey: Most popular programming languages for professional developers



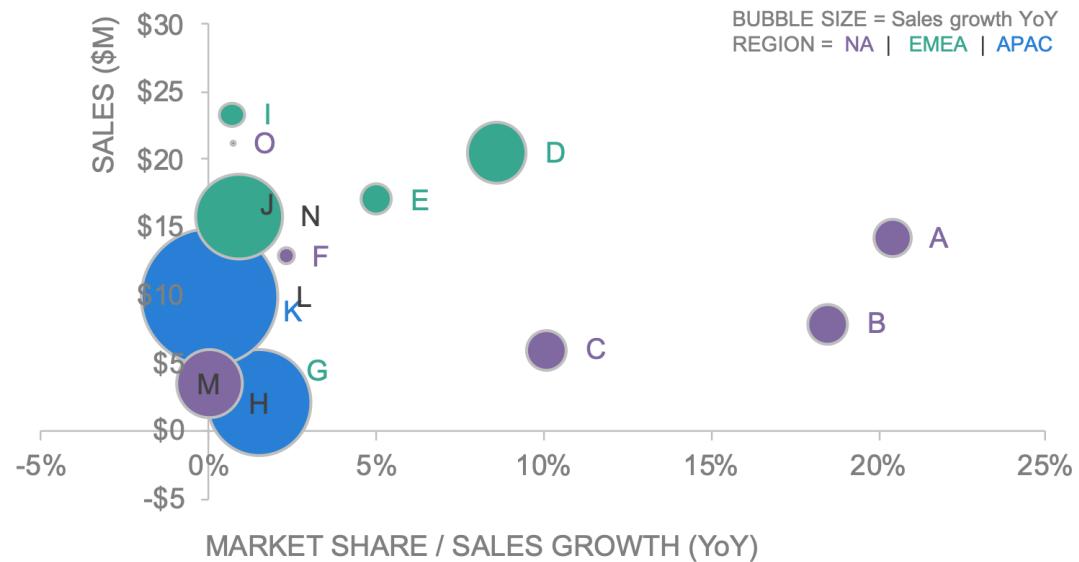
# Color / Highlight



# Marker / Shape



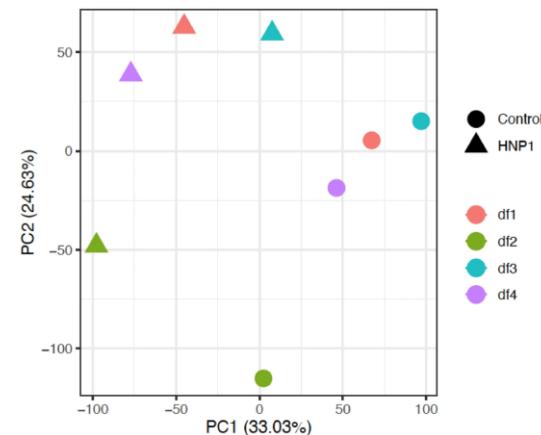
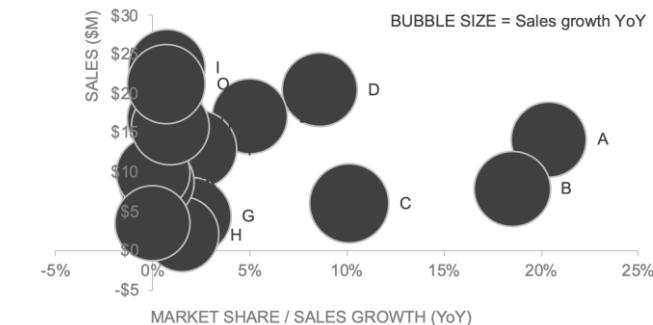
## Competitive landscape



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

## Bubble sizes are too similar

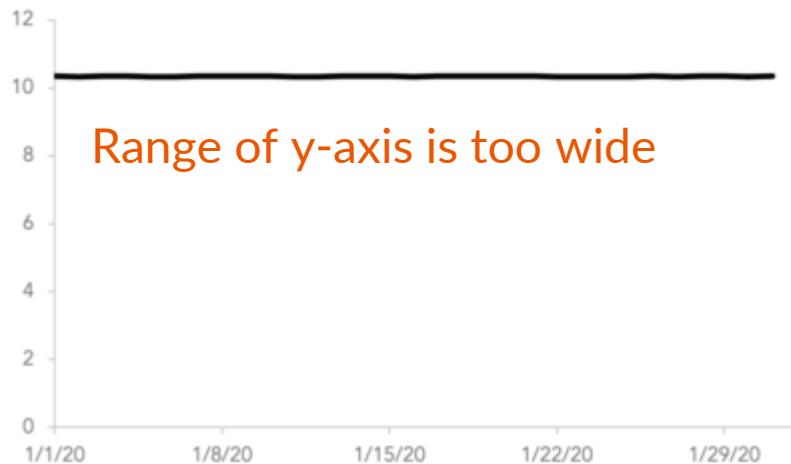
### Competitive landscape



# Zoom / Scale

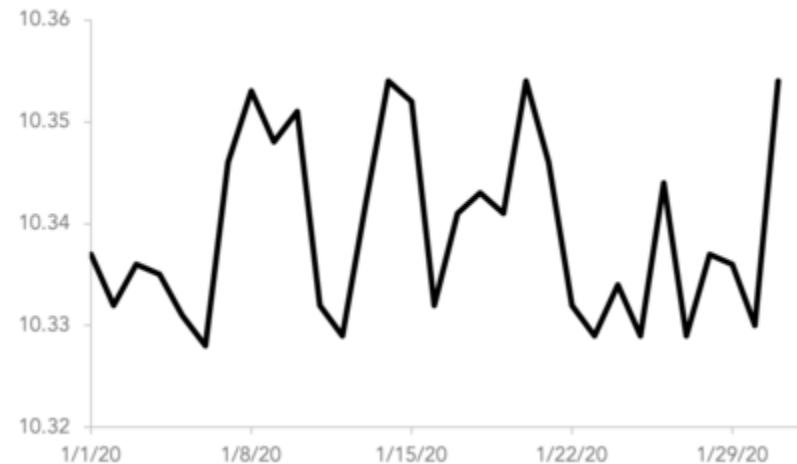


ABC Corporation January daily valuation  
IN BILLIONS (USD)



Range of y-axis is too wide

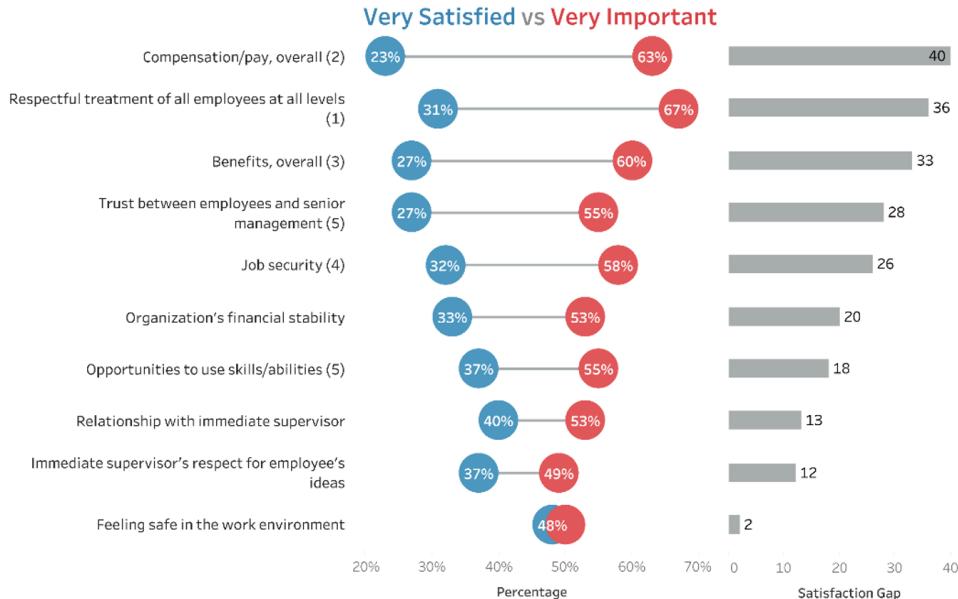
ABC Corporation January daily valuation  
IN BILLIONS (USD)



# Reference / Normalization

## THE SATISFACTION GAP

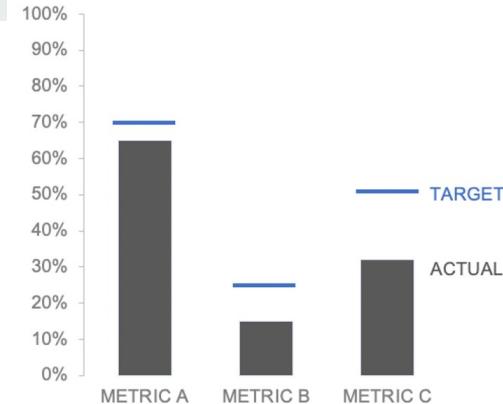
How can we expect employees to be engaged when they're not satisfied with their most important job aspects?



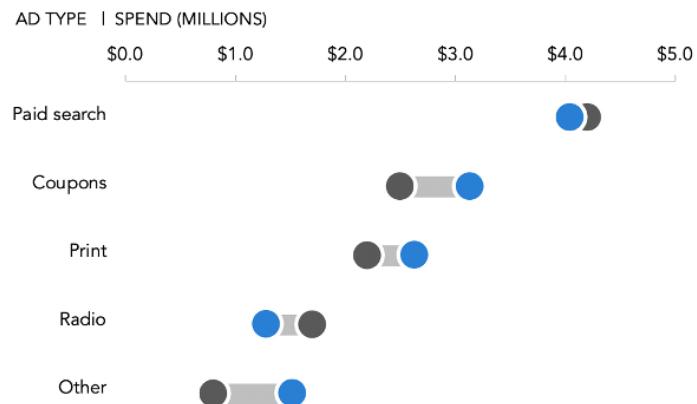
Source: Society for Human Resource Management, 2016 Employee Job Satisfaction and Engagement Report

<https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/2016-Employee-Job-Satisfaction-and-Engagement-Report.pdf>

## Performance metrics



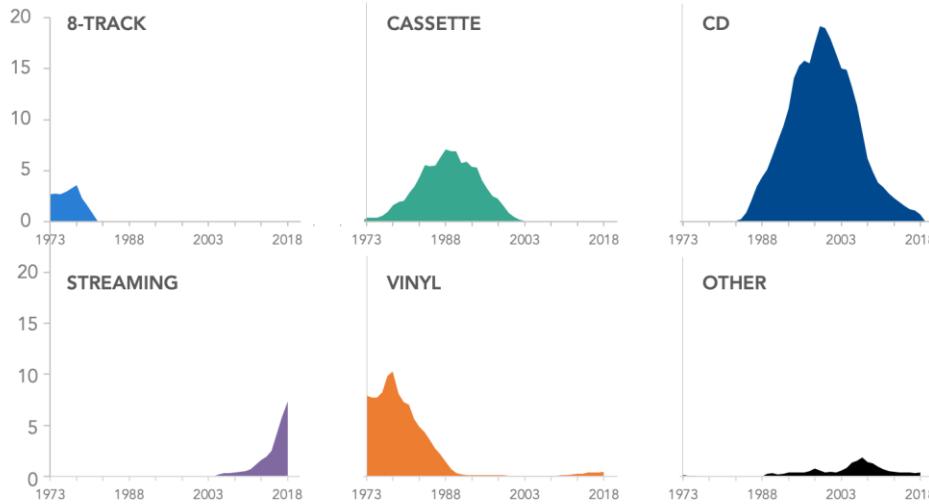
## Advertising spend by category: us vs. competitor



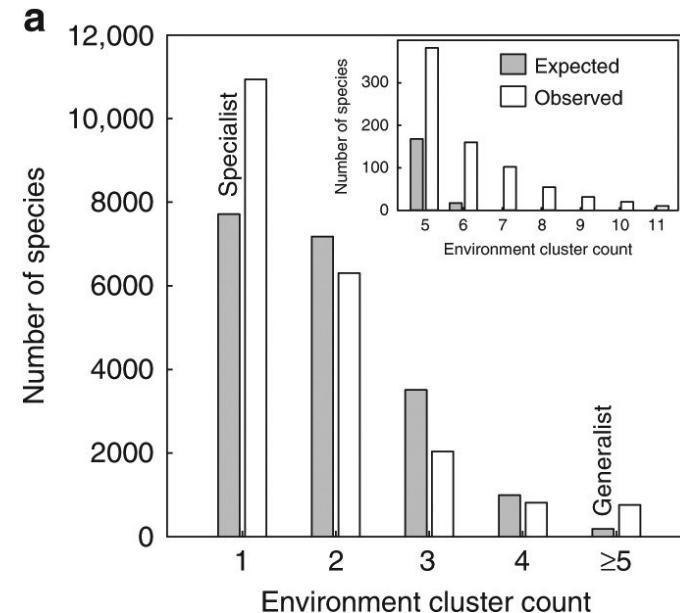
# Multiplot / Inset

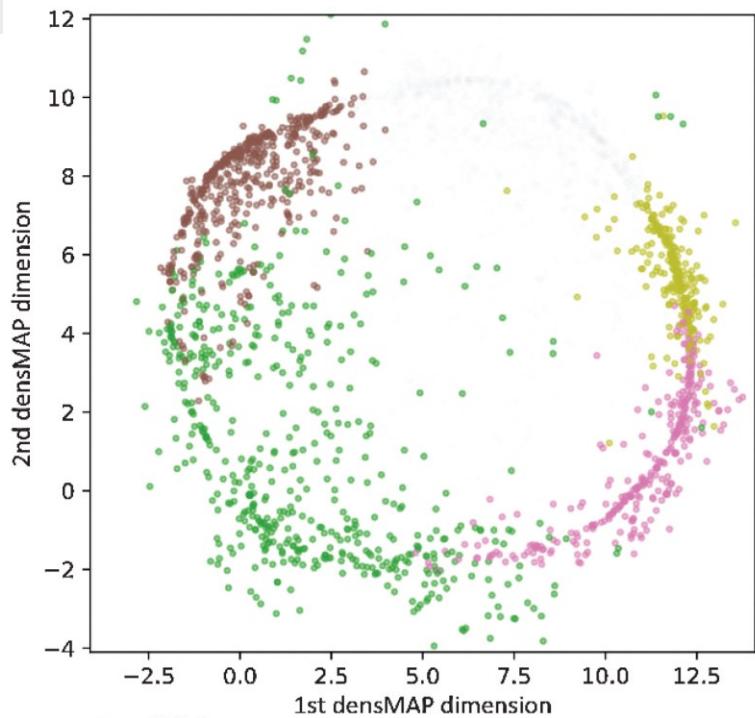


US music sales by format (inflation-adjusted)  
IN BILLIONS (USD)

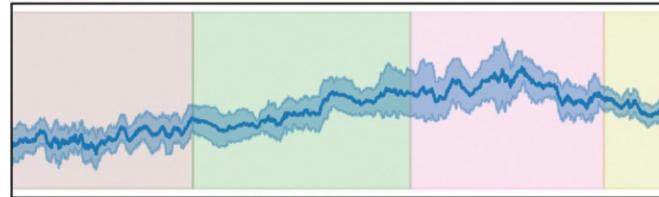
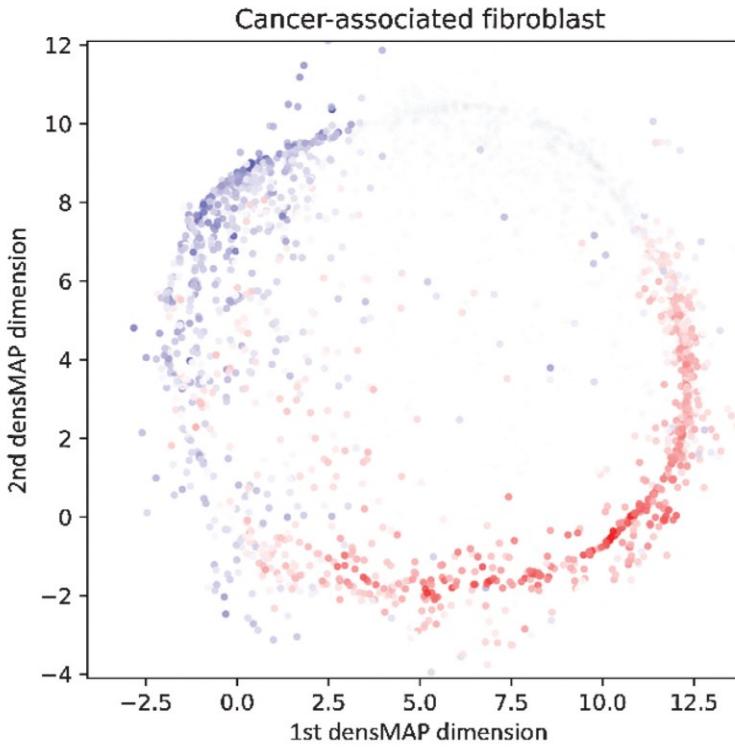


SOURCE: Recording Industry Association of America





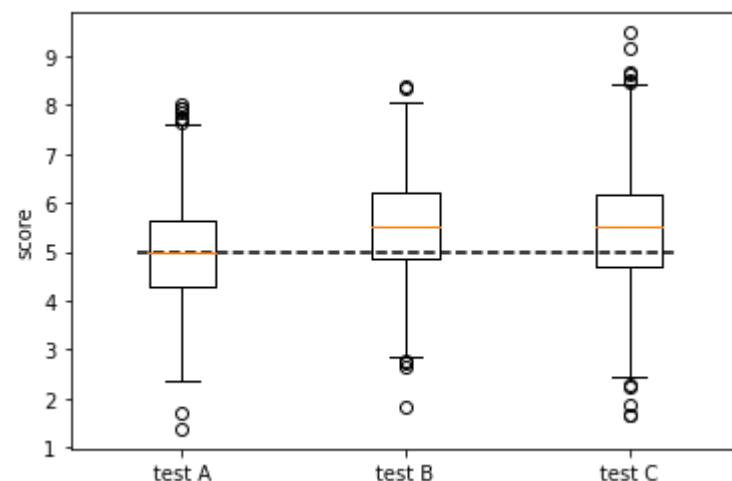
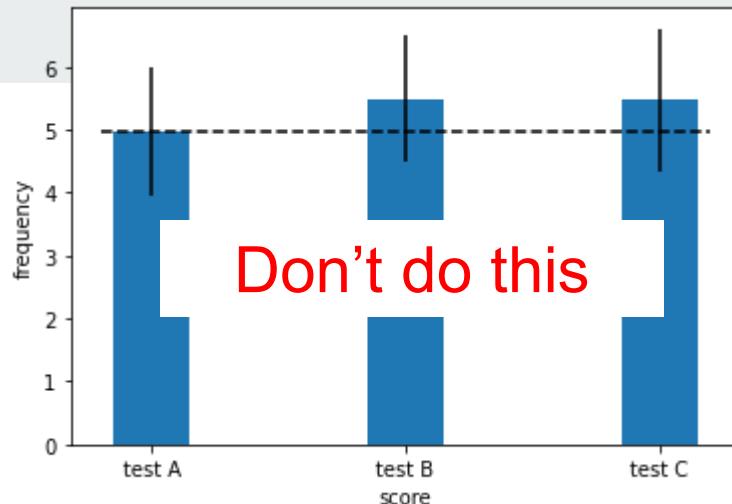
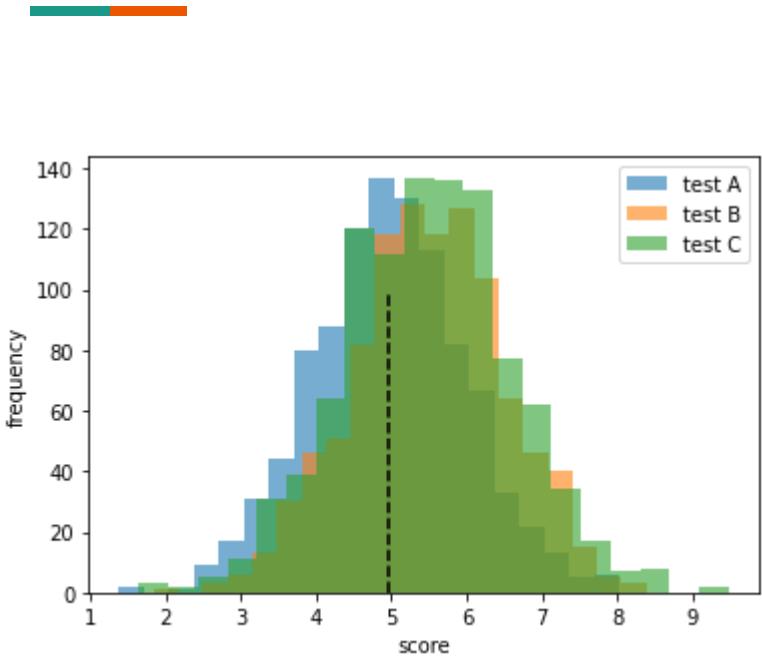
- CMS1
- CMS2.1
- CMS2.2
- CMS3
- CMS4.1
- CMS4.2
- CMS4.3





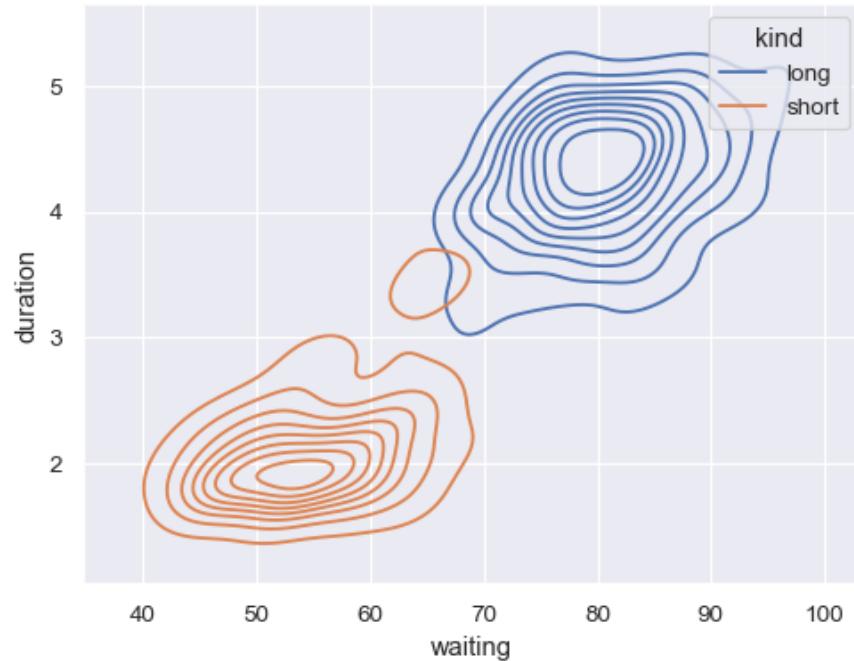
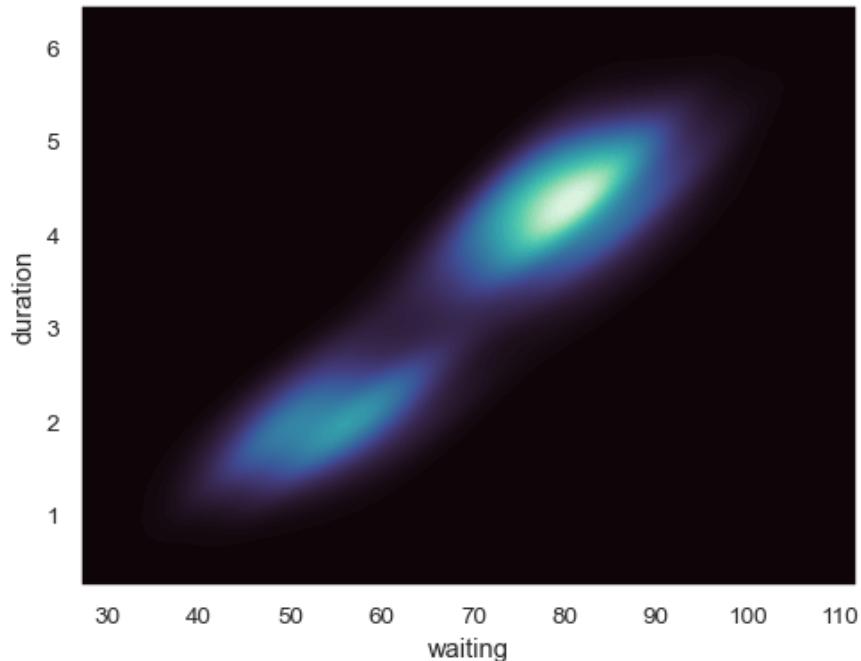
# Some more examples

# Distribution



# Kernel Density Estimate (KDE) plot

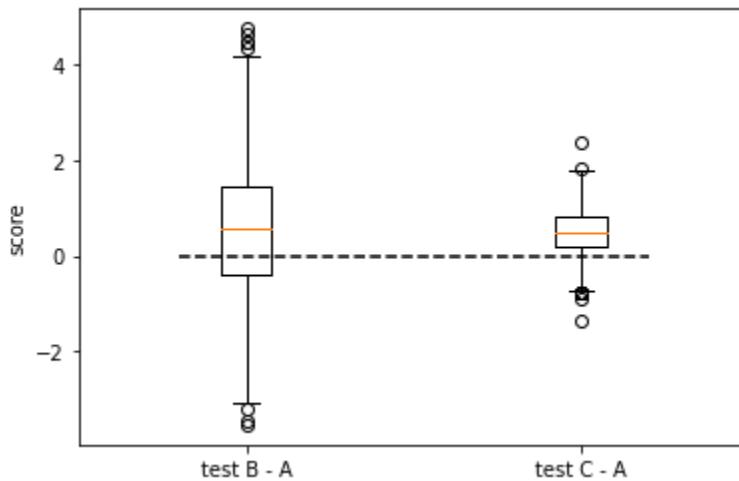
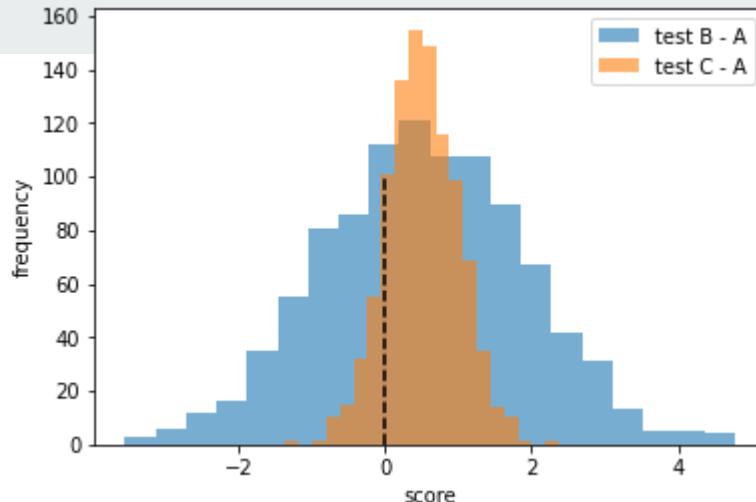
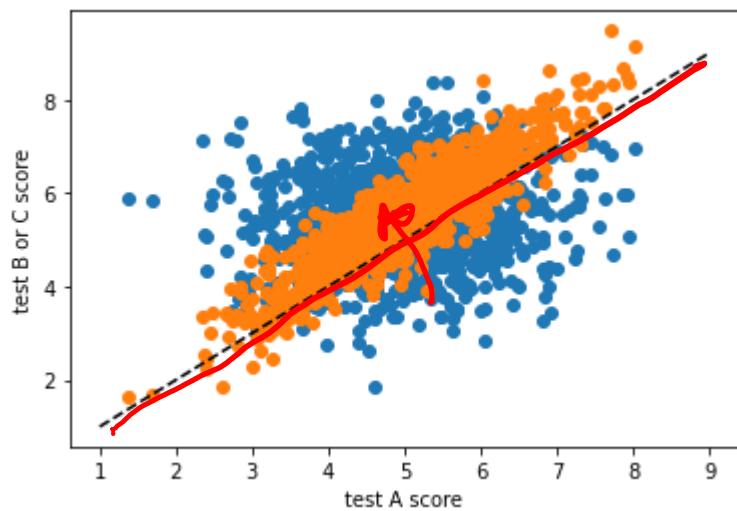
duration  
waiting

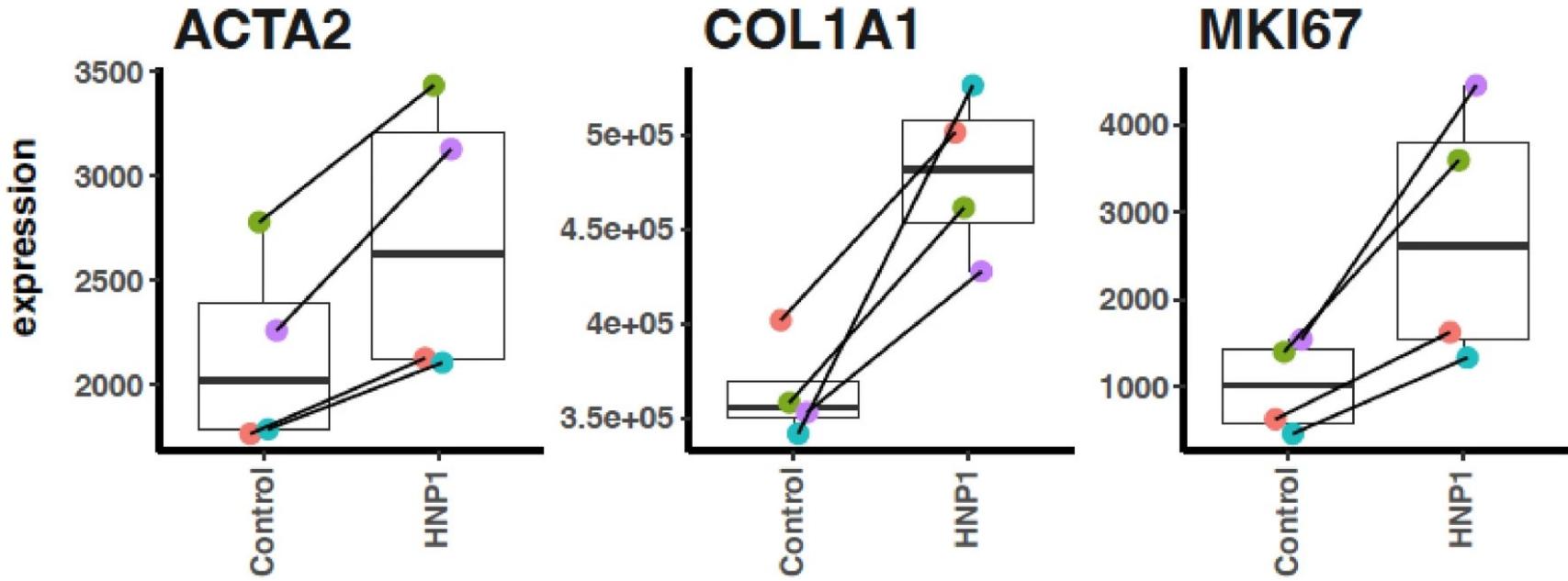


# Paired vs unpaired data



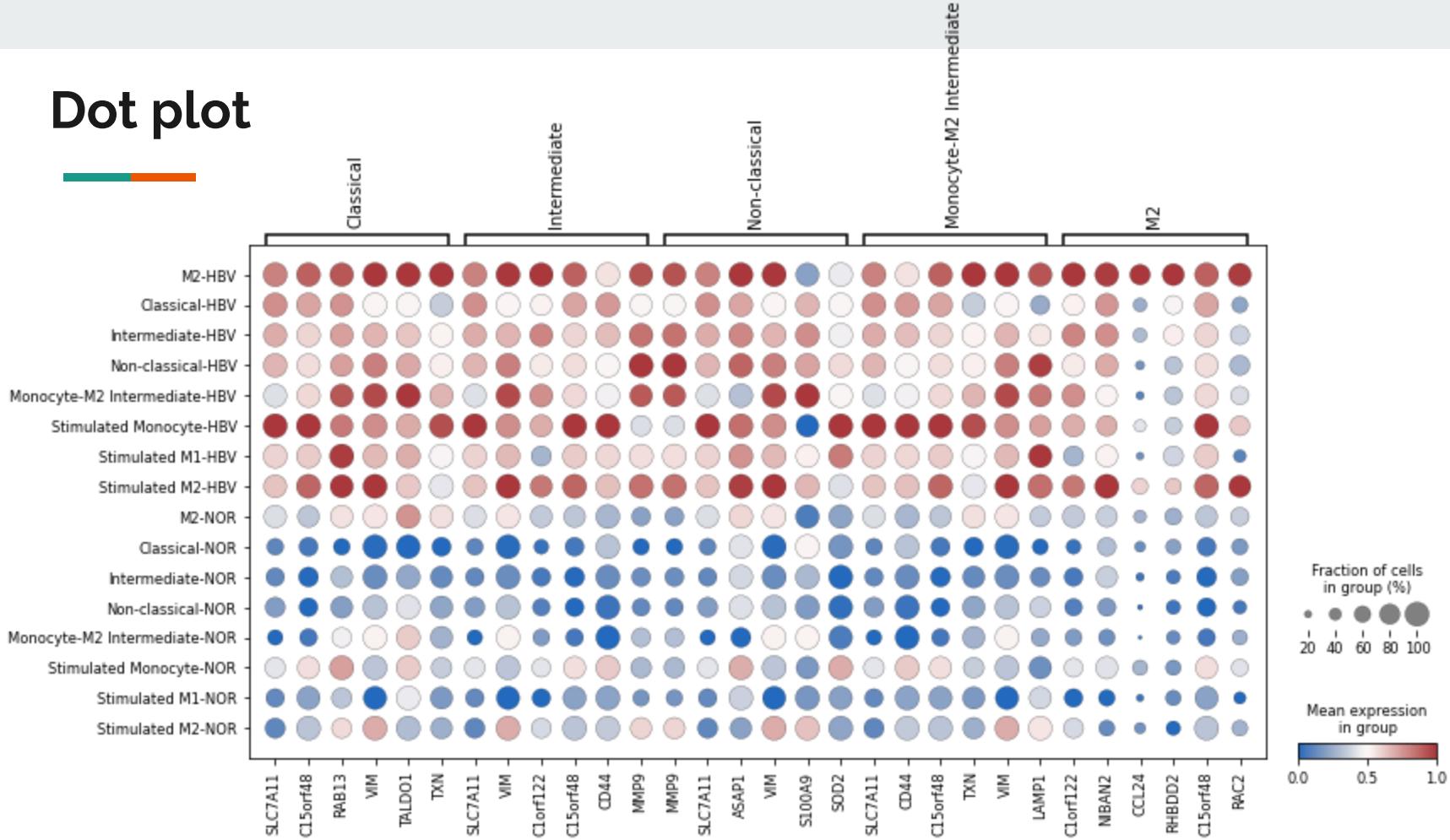
Show paired differences





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

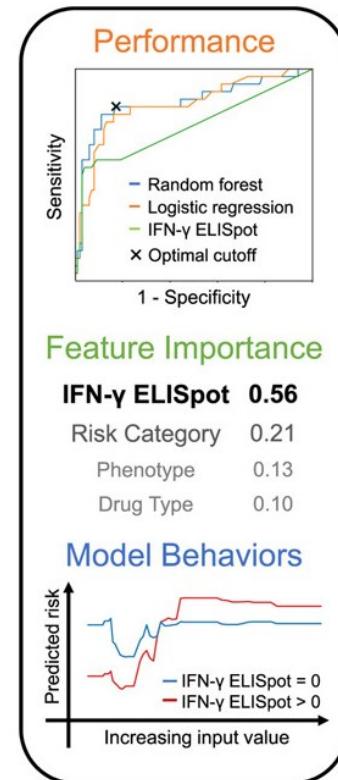
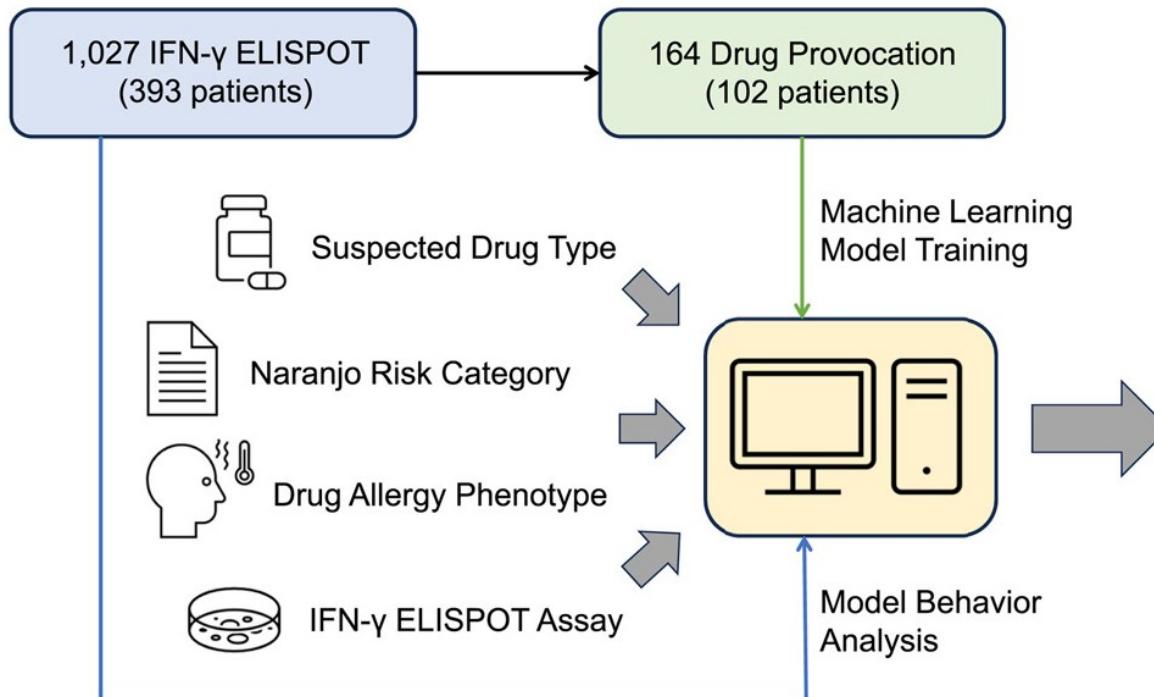
# Dot plot

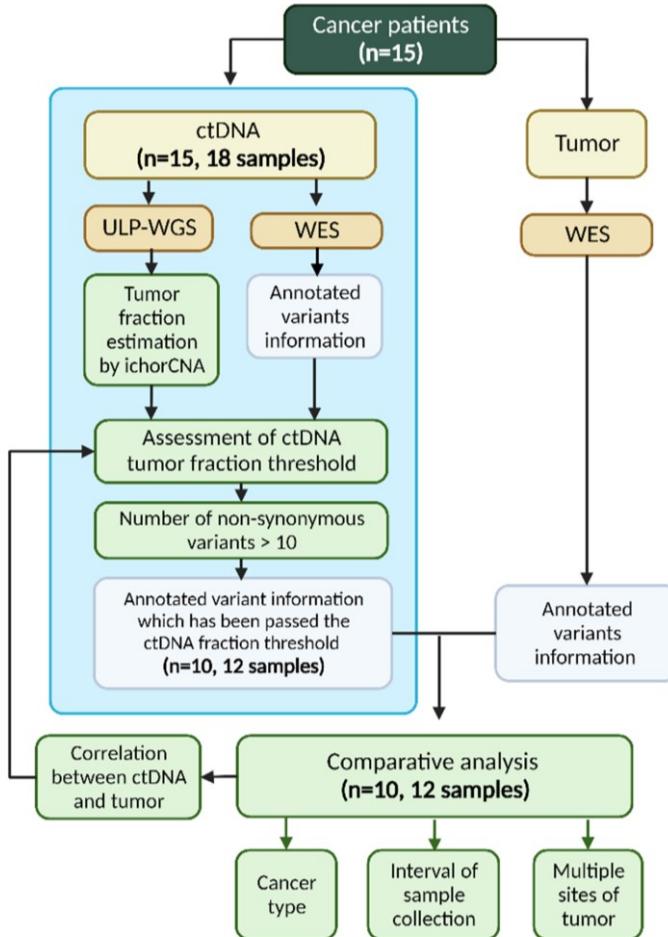




# Diagram / Conceptual figure

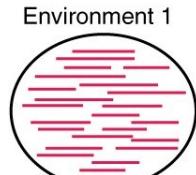
# Mix of real data, graph, and cartoon



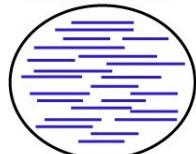


- Top-to-bottom hierarchy
- Color coding
- Grouping of related steps

**a** 16S rRNA fragments  
(Environmental)

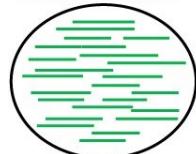


Environment 2

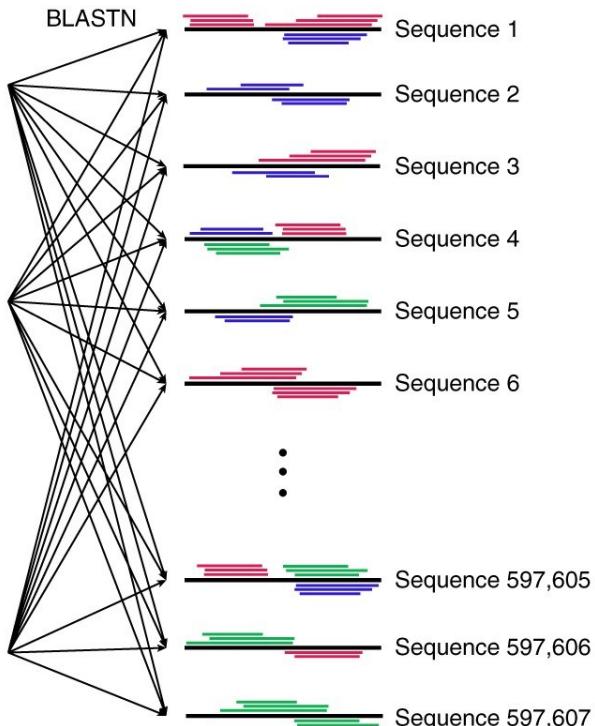


⋮

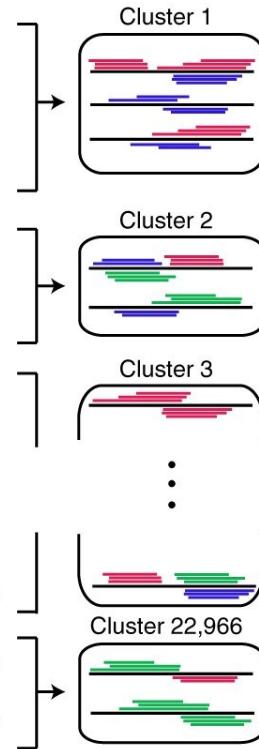
Environment 61



16S rRNA sequences  
(SILVA Database)



Sequence clusters  
( $\geq 98\%$  identity)



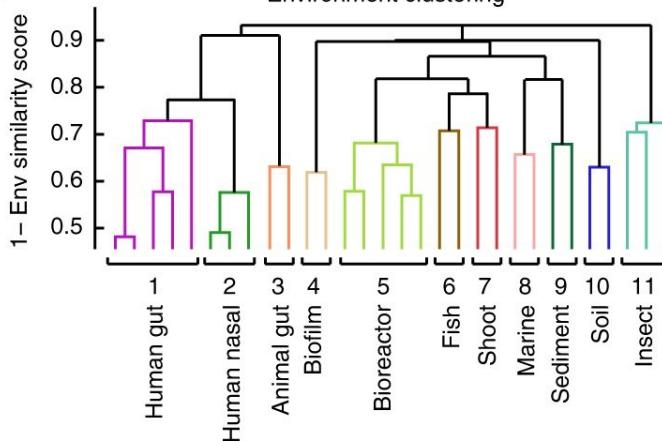
**b**

Environment profiles of aligned 16S rRNA fragments

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

**c**

Environment clustering



# Any questions?

---

See you on February 9<sup>th</sup>