GHGA THE GERMAN HUMAN GENOME-PHENOME ARCHIVE

A Beginner's Guide to Data Visualisation

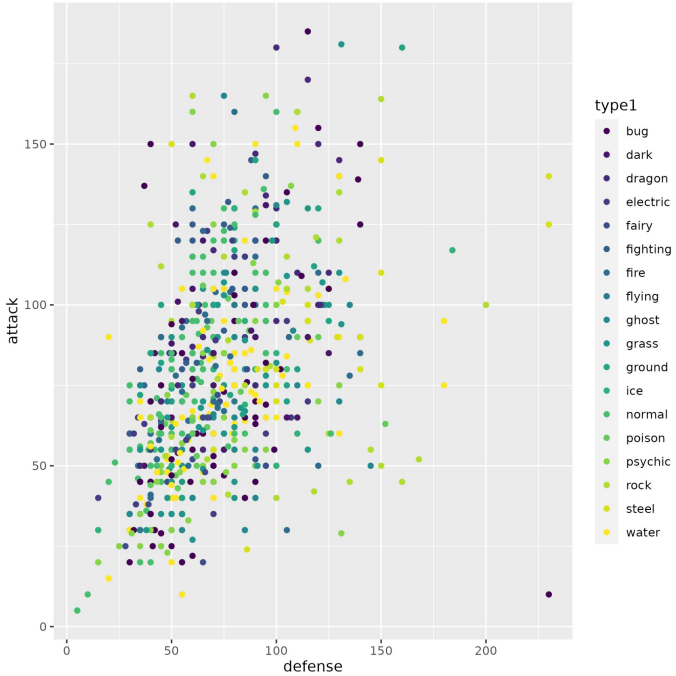
Florian Heyl

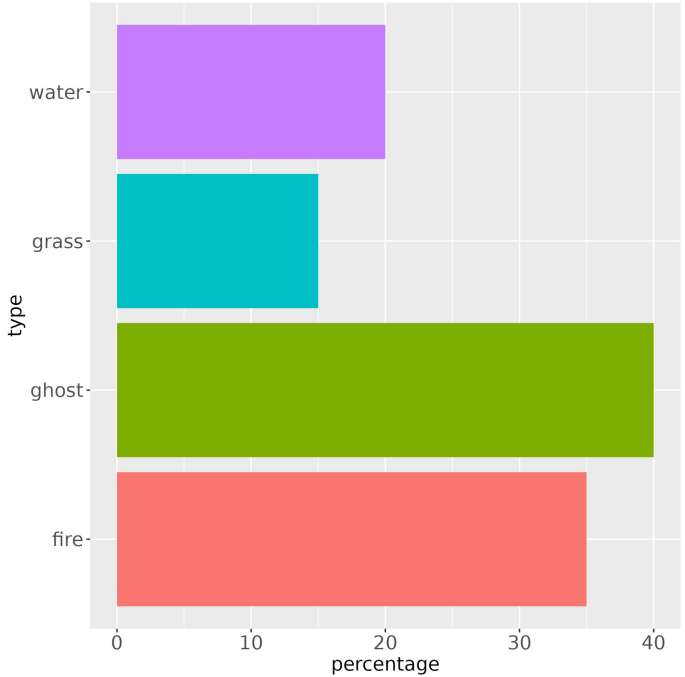
In cooperation with

nfdi

Data visualization is a form of communication

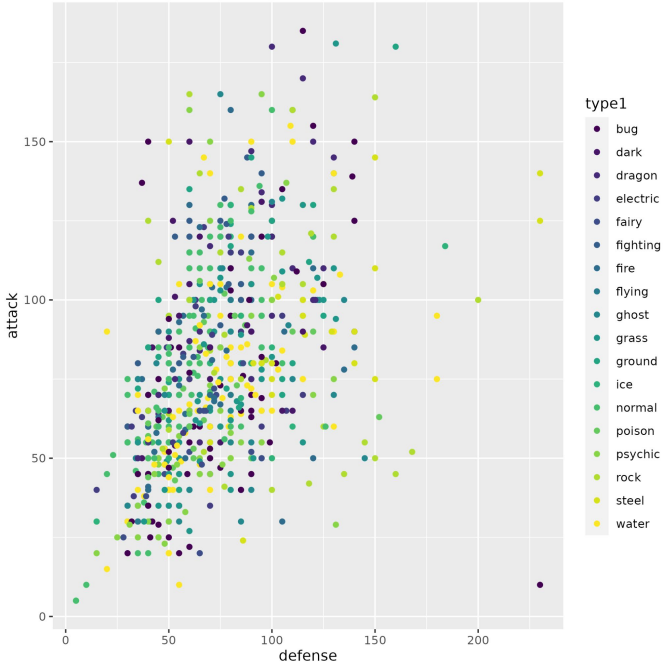
Inform ⟷ Misinform

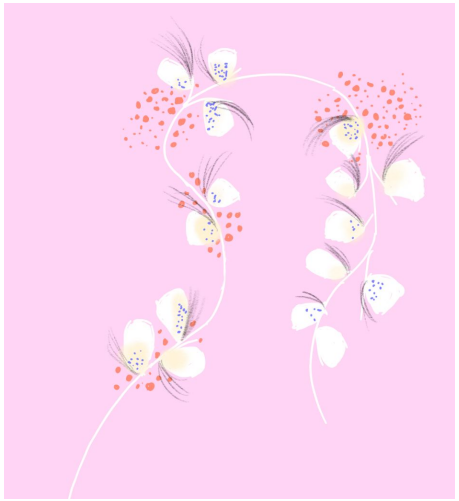




Data visualization is a form of communication

Objective ⟷ Subjective

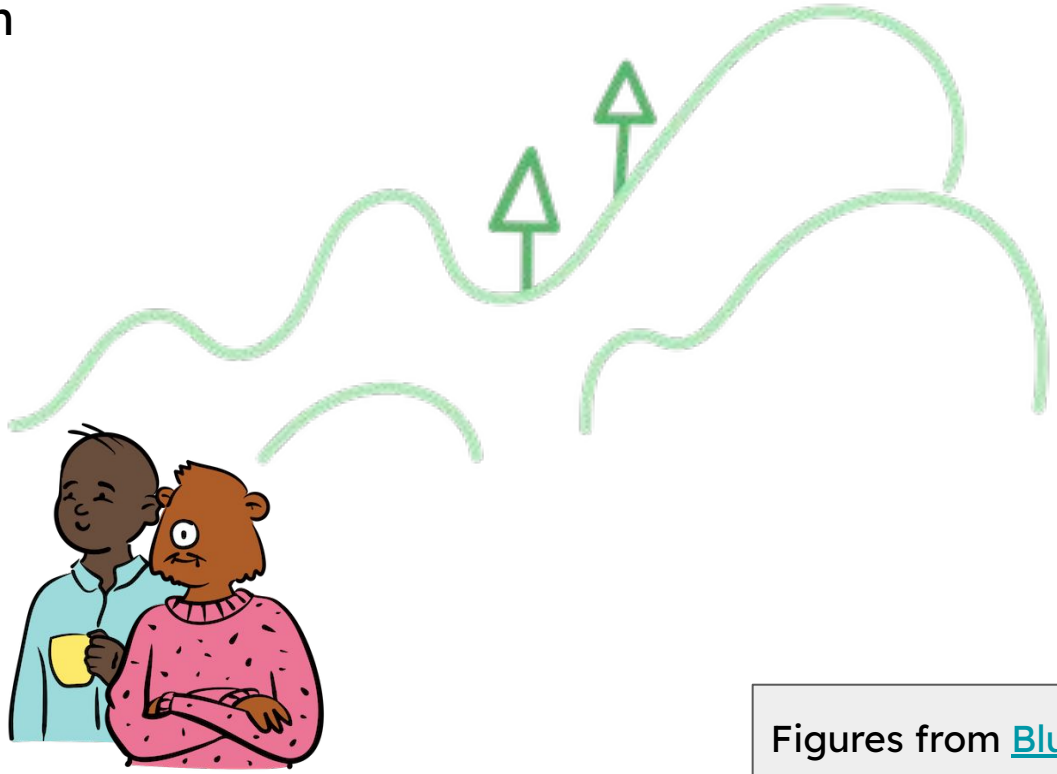




[Giorgia Lupi and Kaki King — Bruises: The Data We Don't See]

Data visualization is a form of communication

- Data + Visualization form
- Audience
- Location / Context



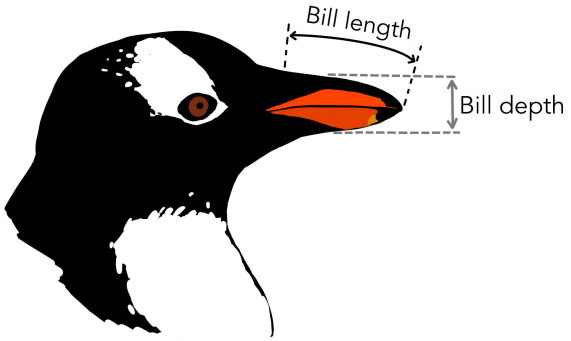
Figures from Blush

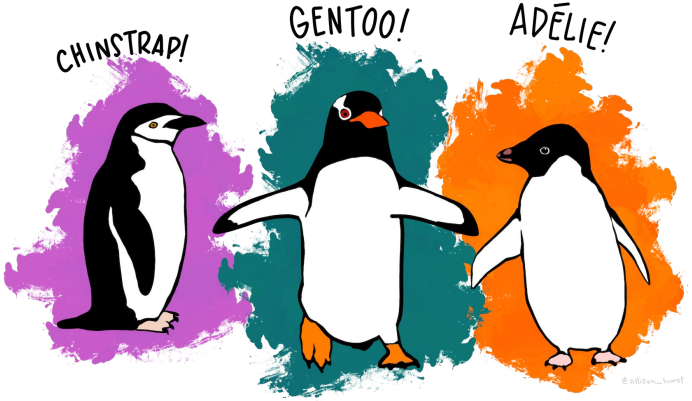
DATA

https://en.wikipedia.org/wiki/Penguin#/media/File:Penguins_collage.png

Types of Data

1. Numerical
    a. Expression data (RNA bulk)
    b. Methylation data (WGBS)
    c. Variant data (WGES)
2. Categorical
    a. Cell composition data
    b. Image recognition
    c. Transcription factor prediction

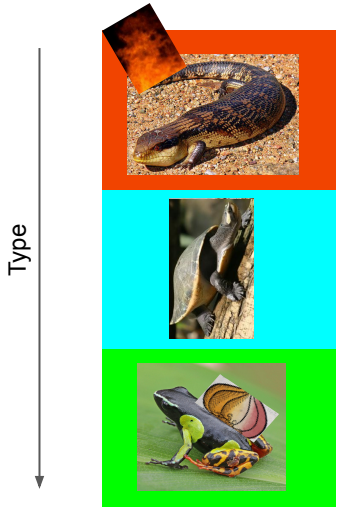




[Allison Horst, Alison Hill, Kristen Gorman - palmerpenguins (example dataset)]

Types of Data

1. Numerical
    a. Expression data (RNA bulk)
    b. Methylation data (WGBS)
    c. Variant data (WGES)
2. Categorical
    a. Cell composition data
    b. Image recognition
    c. Transcription factor prediction



https://en.wikipedia.org/wiki/Frog#/media/File:Variegated_golden_frog_(Mantella_baroni)_Ranomafana.jpg
https://en.wikipedia.org/wiki/Onion#/media/File:An_onion,_woodcut,_1547_Wellcome_L0029214.jpg
https://en.wikipedia.org/wiki/Turtle#/media/File:Turtle_diversity.jpg
https://en.wikipedia.org/wiki/Lizard#/media/File:Lizard_Collage.jpg
https://en.wikipedia.org/wiki/Lizard#/media/File:Lizard_Collage.jpg

Important data characteristics

1. Data integrity
    a. Accuracy
    b. Completeness
    c. Consistency
    d. Validity
2. Quality
3. Security
4. Collect more data if needed

Good statistical practice / experimental design - W. Huber
Modern statistics for modern biology - S. Holmes & W. Huber

Messy Data

Data that is not correctly formatted

Correspondence | Open access | Published: 23 June 2004

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett & John N Weinstein

BMC Bioinformatics **5**, Article number: 80 (2004) | Cite this article

NEWS | 13 August 2021 | Correction 25 August 2021

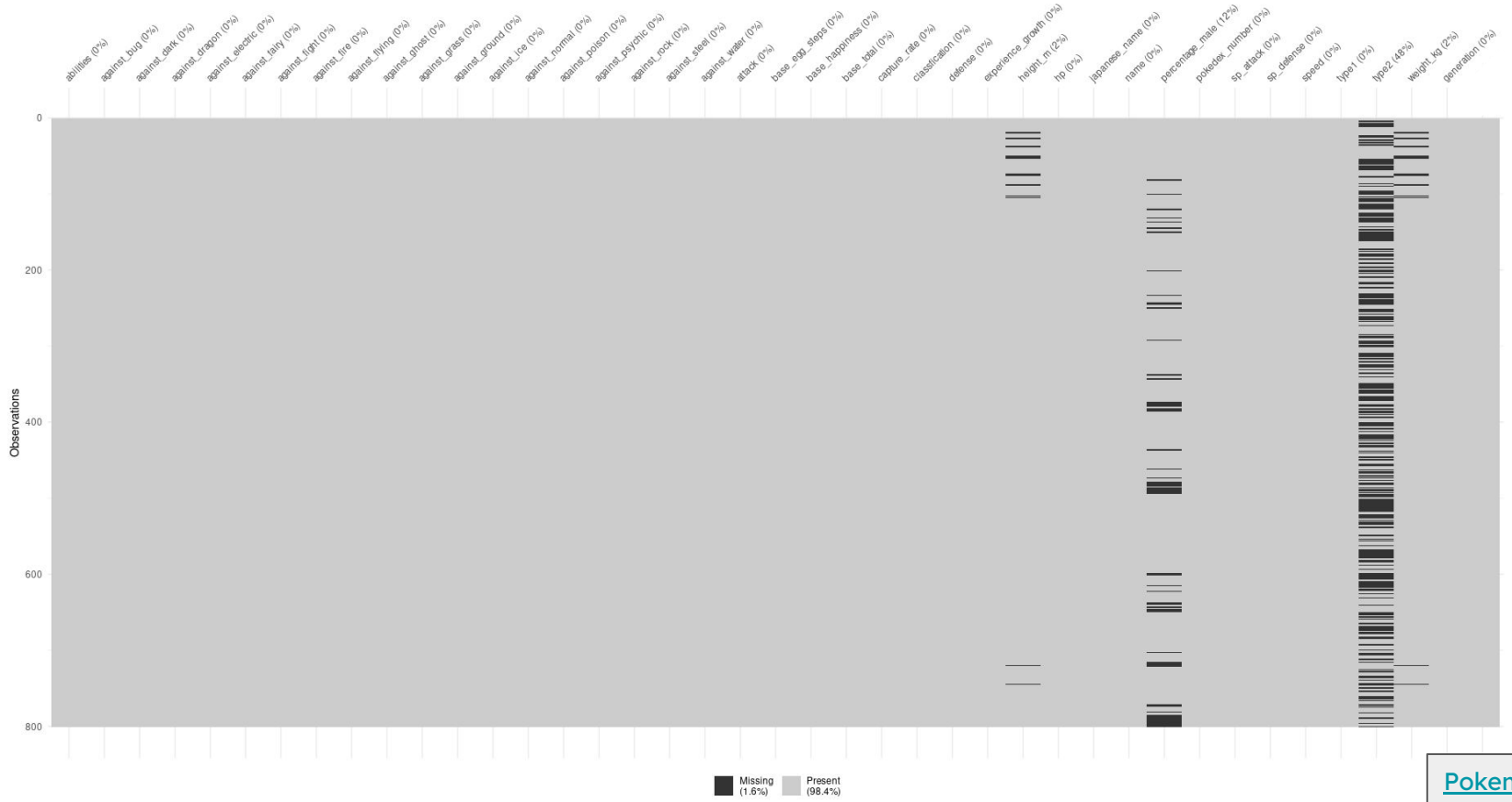
Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

By Dyani Lewis

FAIR Data / Research Data Management
https://www.ghga.de/resources/training/

Missing Data



Pokemon Data

Types of Analysis

1. Descriptive (describing your data)
2. Exploratory (relationship of different descriptive analysis)
3. Inferential (testing a hypothesis)
4. Causal (finding the cause of an observation)
5. Prediction (predict an outcome based on your data)

Good statistical practice / experimental design - W. Huber
Modern statistics for modern biology - S. Holmes & W. Huber

PLOTS – Critic: “It is fun, but I could barely see it.”

75%

TOMATOMETER

20 Reviews

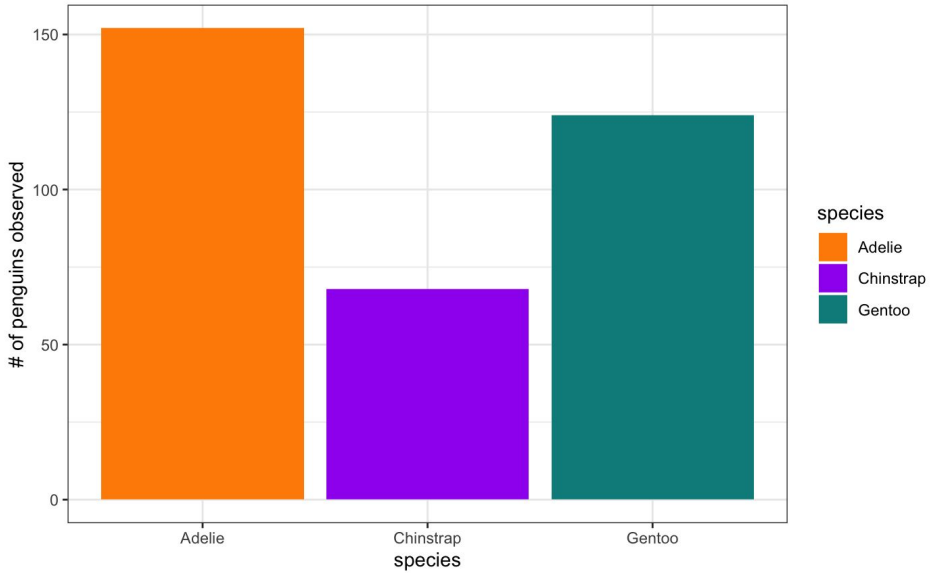
33%

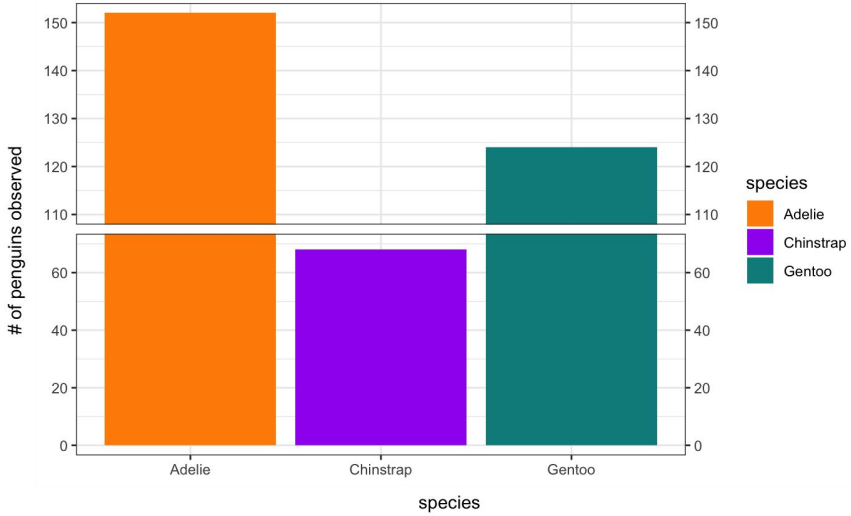
AUDIENCE SCORE

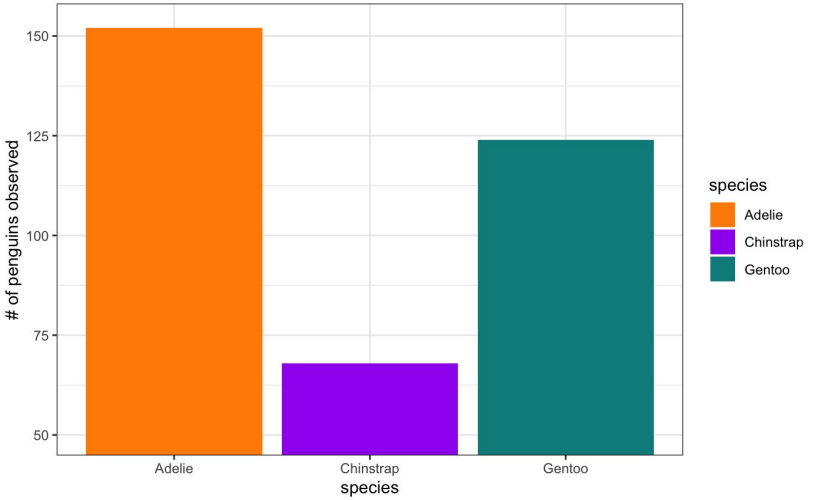
5000+ Ratings

https://en.wikipedia.org/wiki/Tomato#/media/File:Tomatoes_plain_and_sliced.jpg
https://en.wikipedia.org/wiki/Popcorn#/media/File:Cinema_popcorn_bucket.jpg

Axes


150
100
50
0
# of penguins observed
Adelie
Chinstrap
Gentoo
species
species
Adelie
Chinstrap
Gentoo

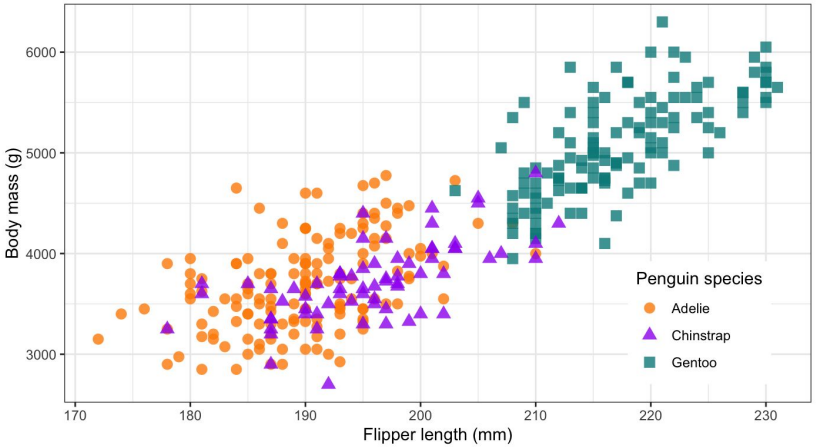

150
140
130
120
110
60
40
20
0
# of penguins observed
Adelie
Chinstrap
Gentoo
species
species
Adelie
Chinstrap
Gentoo


150
125
100
75
50
# of penguins observed
Adelie
Chinstrap
Gentoo
species
species
Adelie
Chinstrap
Gentoo

Scaling

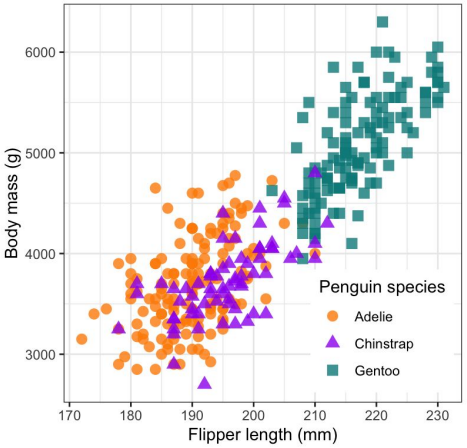
Flipper length and body mass

Dimensions for Adelie, Chinstrap and Gentoo Penguins at Palmer Station LTER



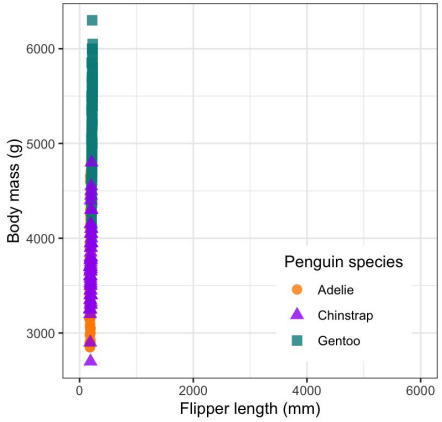
Flipper length and body mass

Dimensions for Adelie, Chinstrap and Gentoo Penguins at Palmer Station LTER



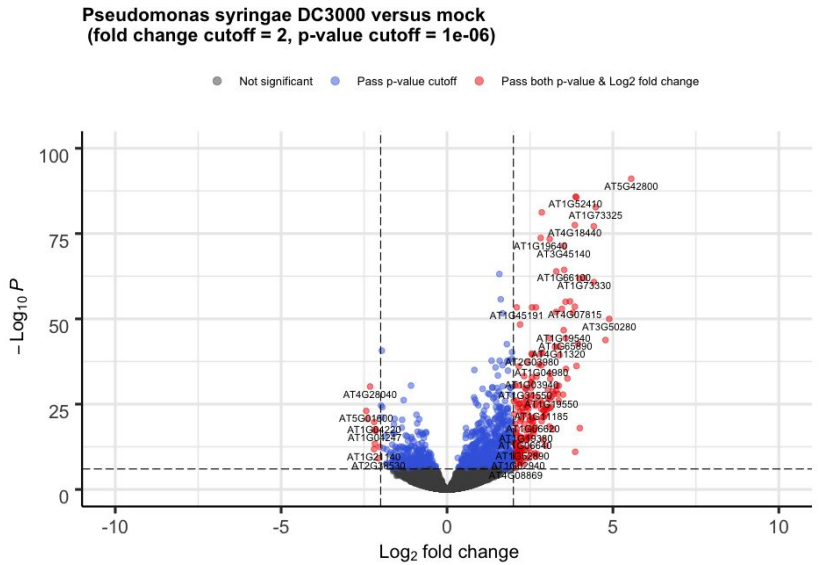
Flipper length and body mass

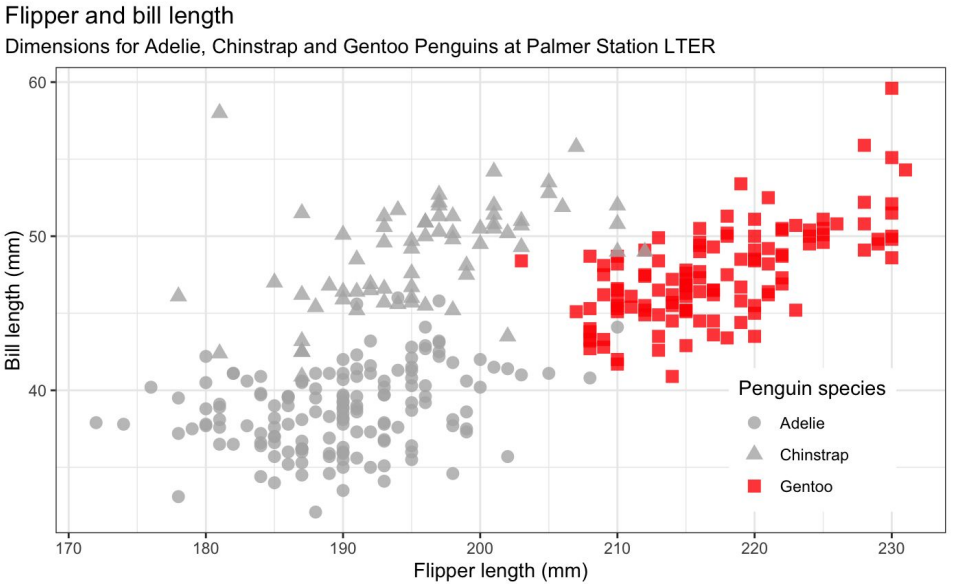
Dimensions for Adelie, Chinstrap and Gentoo Penguins at Palmer Station LTER



Colors

Associations: grey = unimportant, red = very important

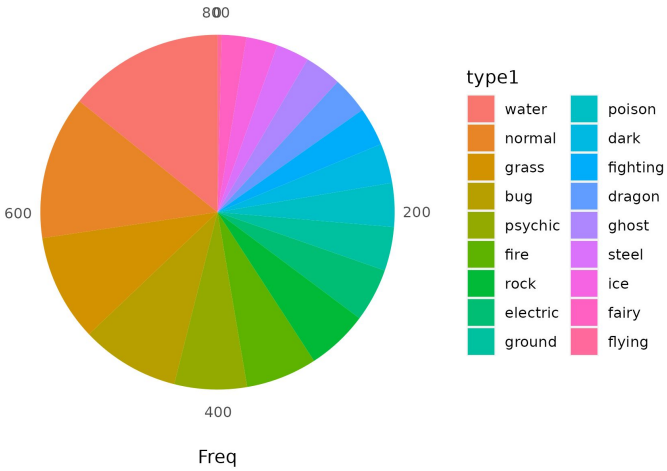


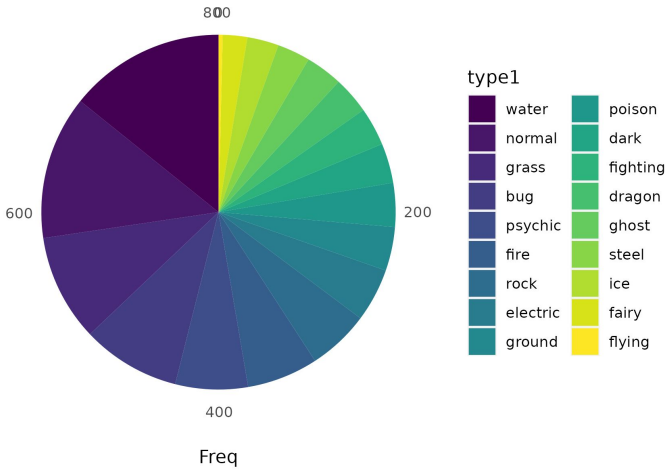


https://scienceparkstudygroup.github.io/rna-seq-lesson/06-differential-analysis/index.html#3-volcano-plot

Colors

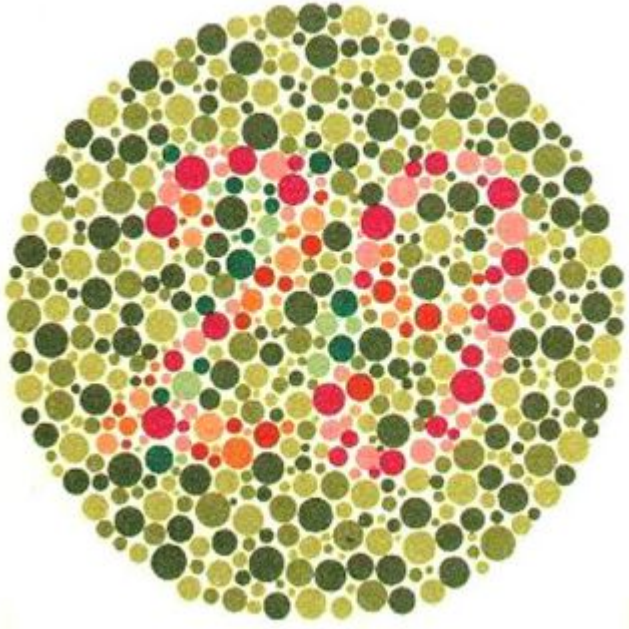
- Mind sequential color schemes for categorical data
- Is it portrayable with grey scale
- Have a consistent color schema
- (Print vs online colors might be different)

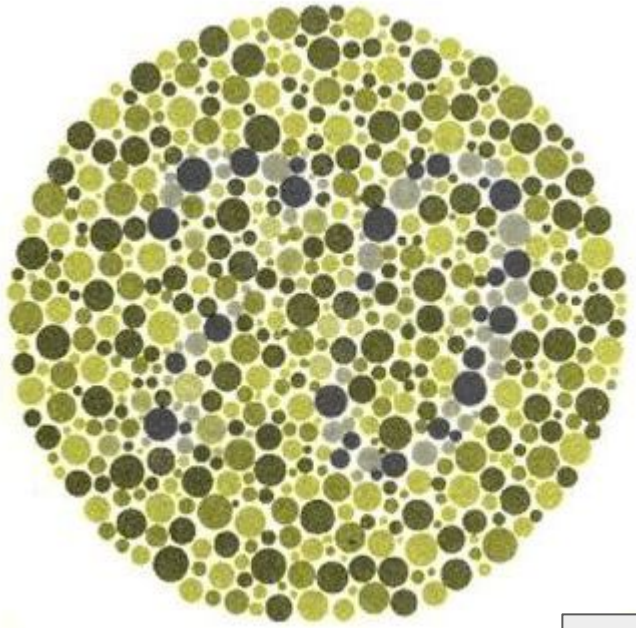




Color impairment simulation tools

Normal

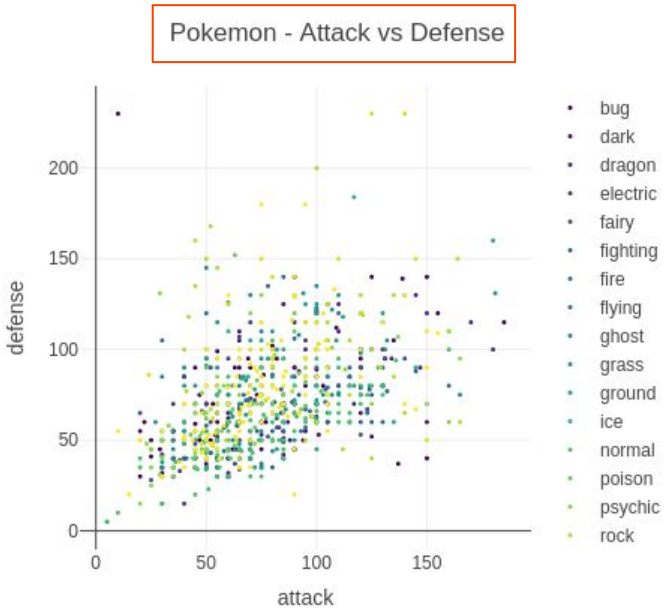


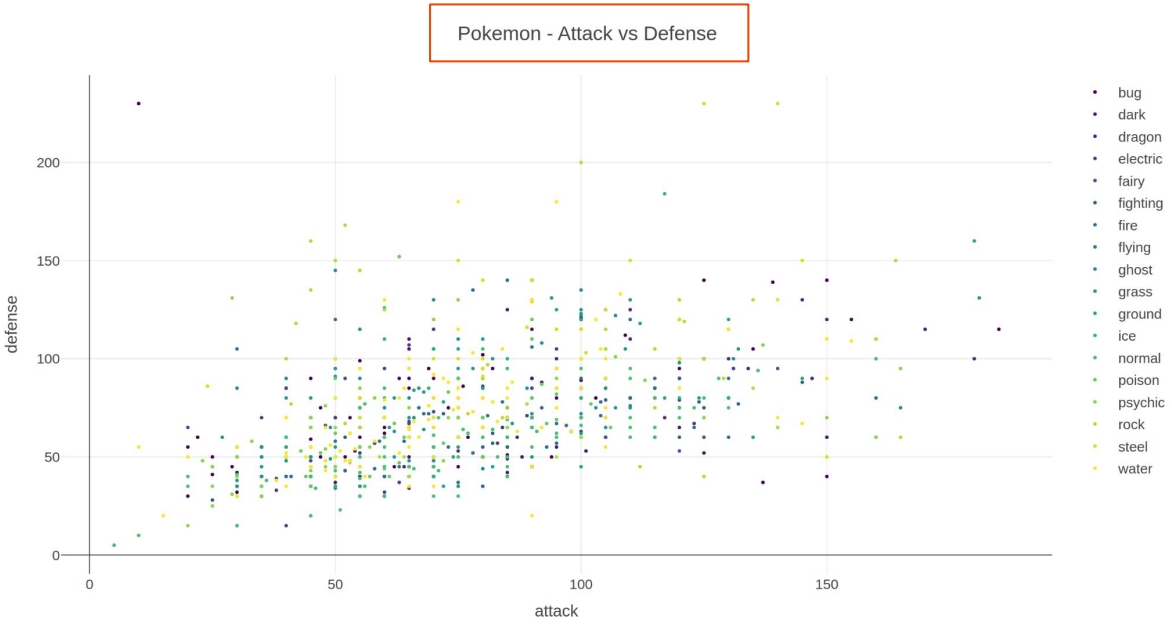


- Daltonize (pyhton)
- Colorblindr (R)

Resolution and Title

Is it needed?



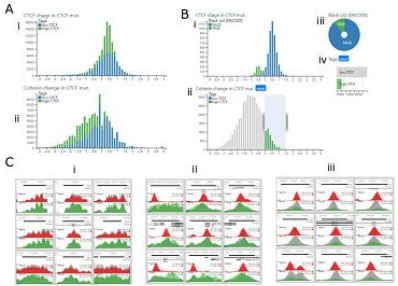


Interactivity

[Multi Locus View]

FEATURED PROJECTS

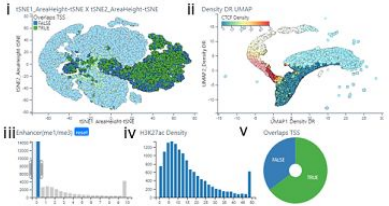
The structural basis for cohesin-CTF-anchored loops



This project takes data from *Li et al* and looks at CTCF and cohesin ChIP-seq experiments. As well as confirming the findings of the paper, it also offers new insights into the data. There are two tutorial videos showing how the project was created, Uploading Data and Basic Analysis and Creating Images and Tagging Data

VIEW

Looking at ChIP-seq signals in enhancers and prometers

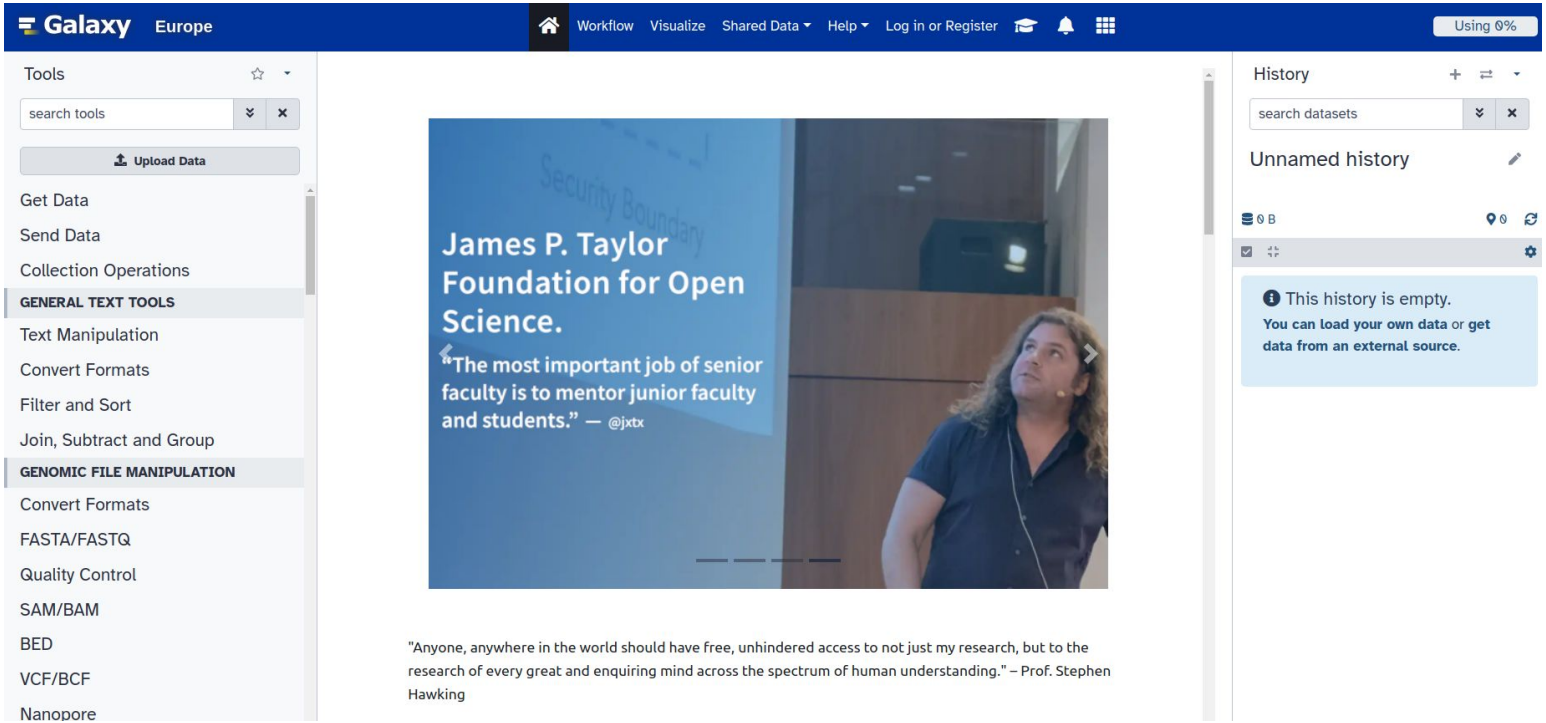


This project takes ChIP-seq data from *Kowalczyk et al* and clusters the data based on various CHiP-seq signals using the dimension reduction algorithms tSNE and UMAP. Along with distance from TSSs, this allows the exploration of promoters and enhancers. It forms the basis of the third tutorial Clustering Data and Further Analysis

VIEW

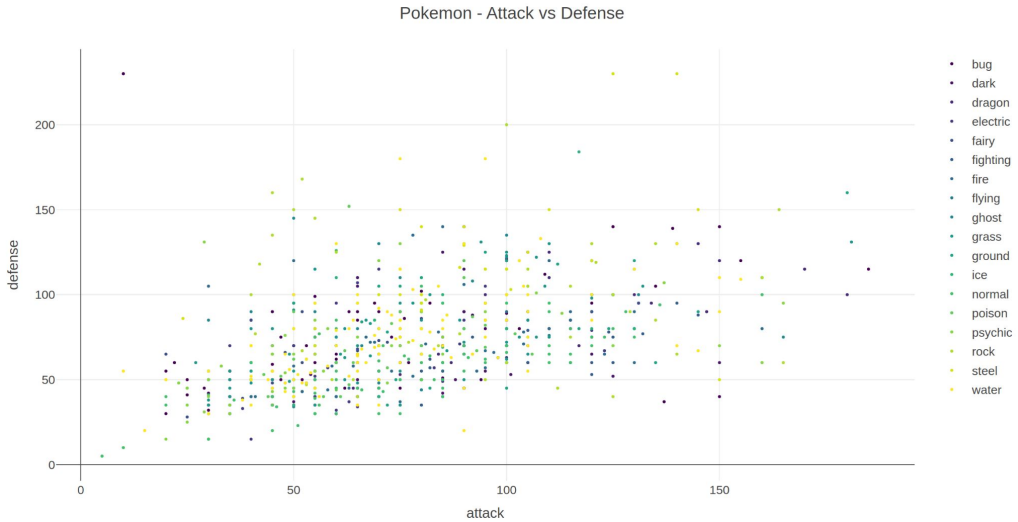
Interactivity

Galaxy



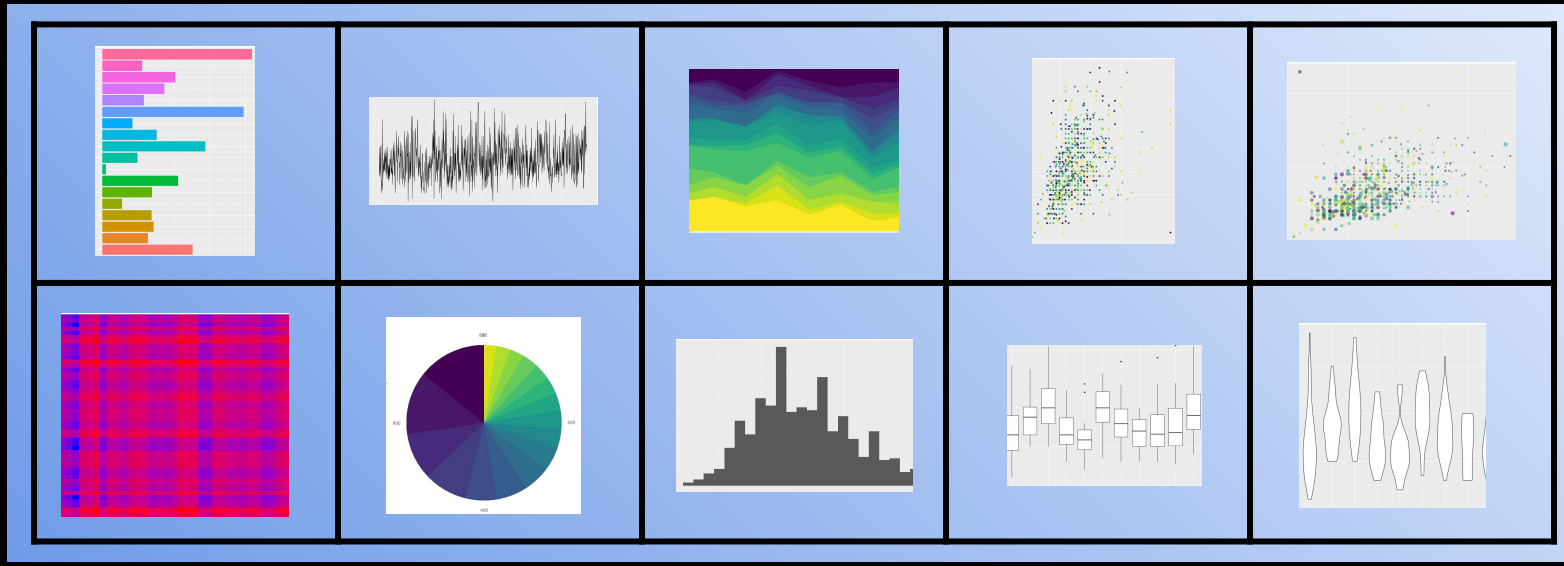
Interactivity

Plotly

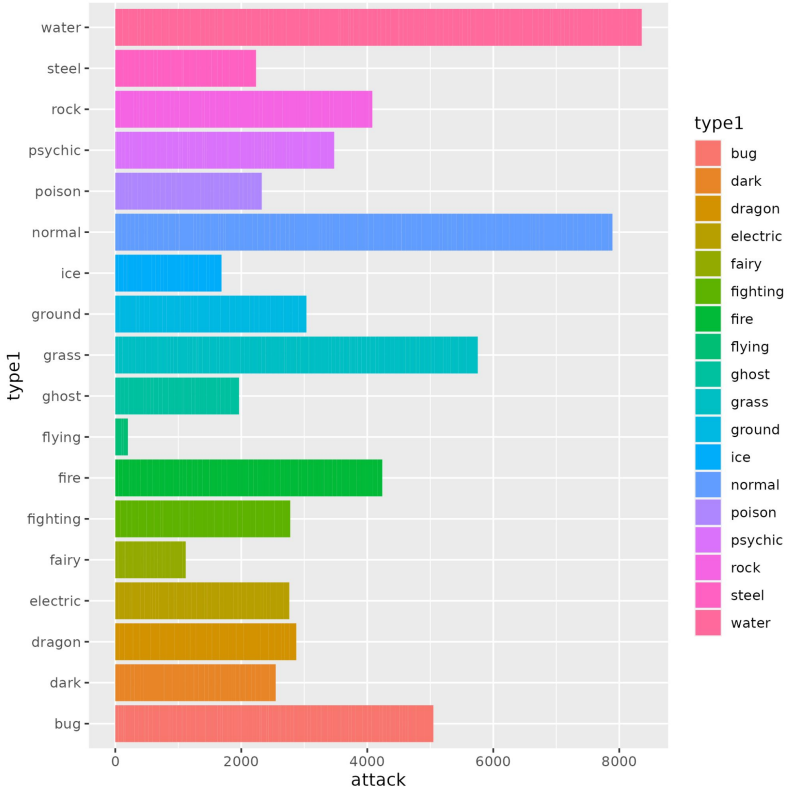


CHARTS – Choose Your Character

Solo Battle



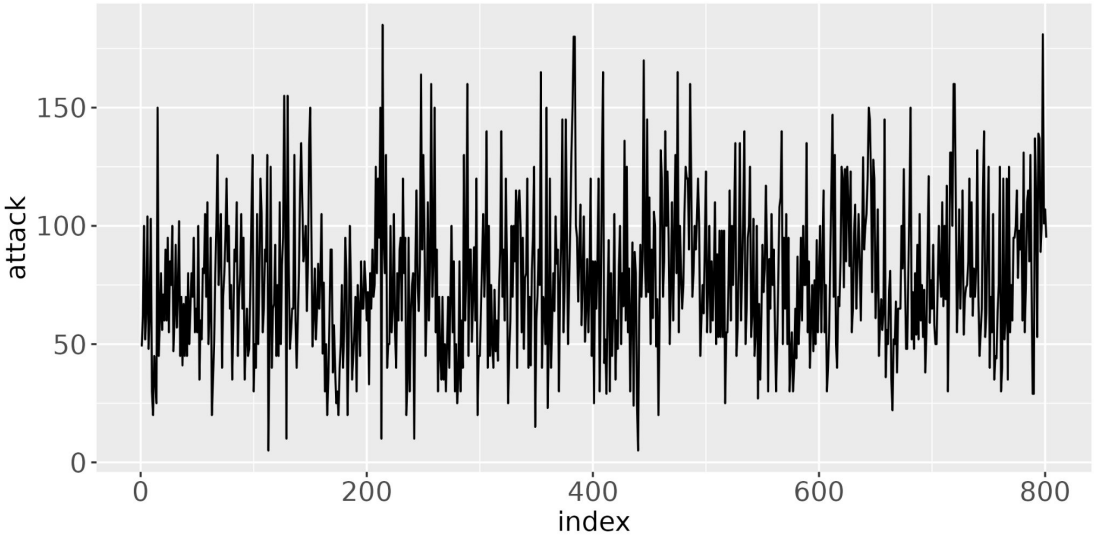
When to use **Bar Chart**?



- Univariate
- Order alphabetically
- Labels horizontal

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

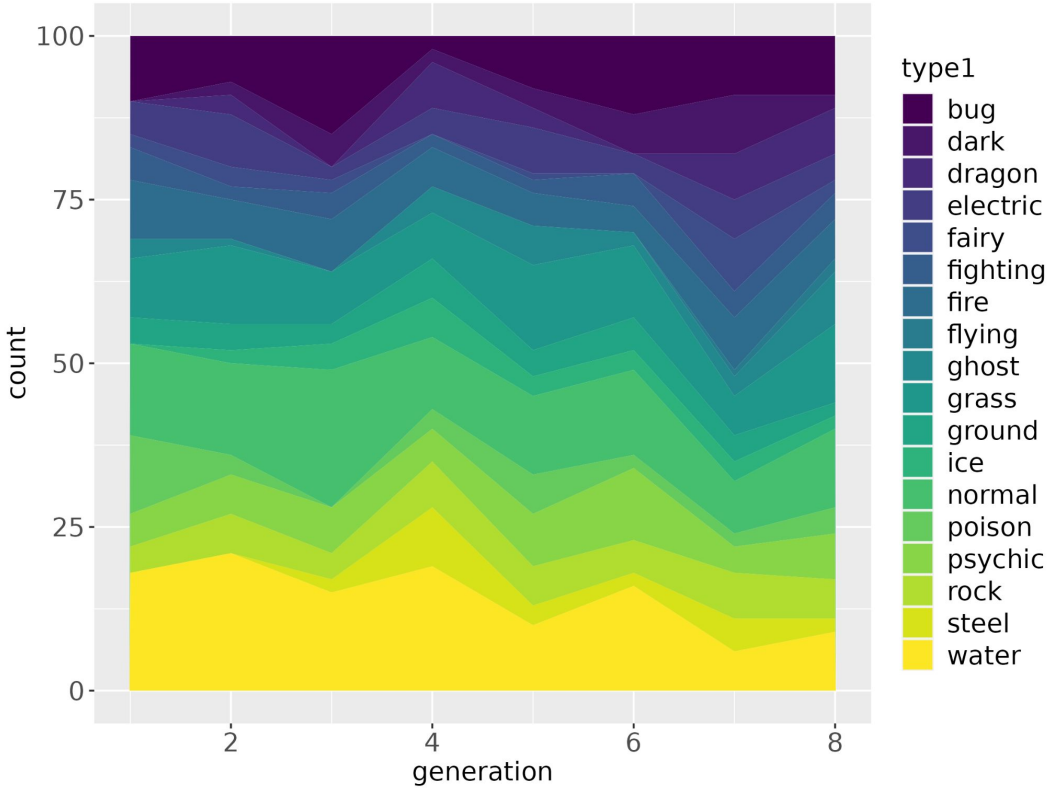
When to use **Line Chart**?



- Non-cyclical data
- Show trends
- Show uncertainties

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

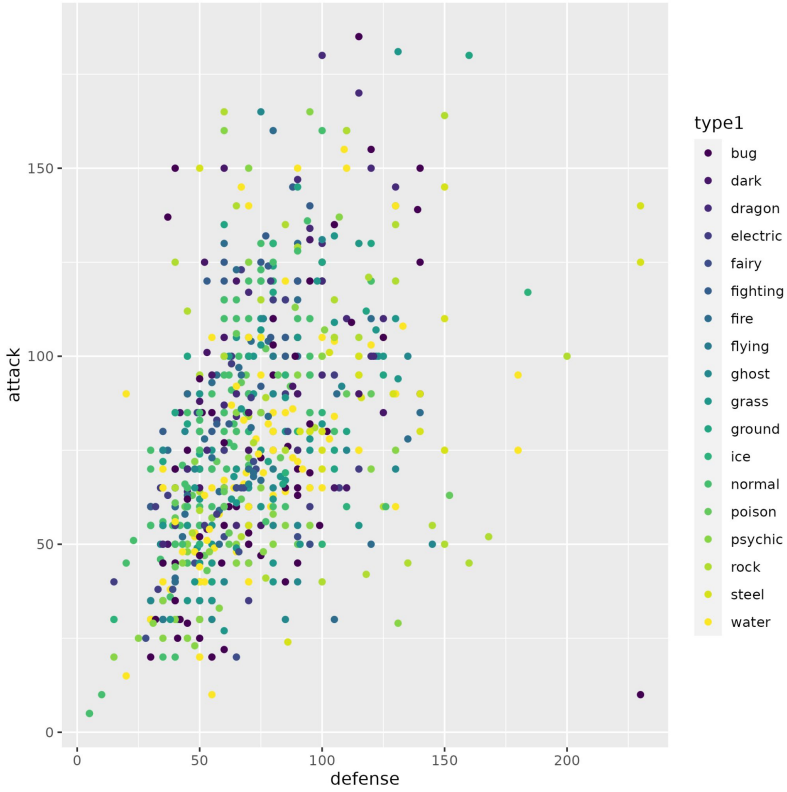
When to use **Area Graph**?



- Relative difference matters
- Time series
- You might need to add annotation / explanation

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

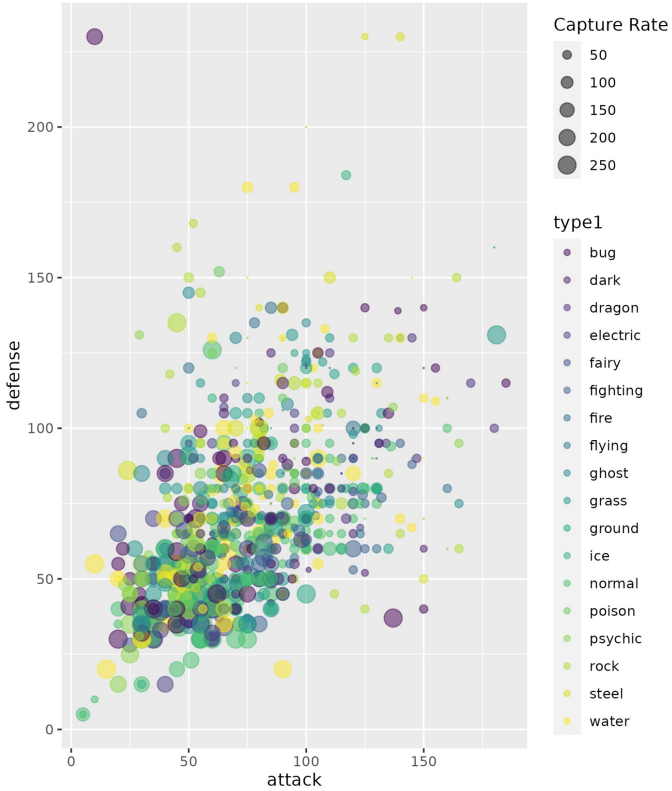
When to use **Scatter Plot**?



- Relationships between two variables
- Highlight areas of interest with colors or marker shapes

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

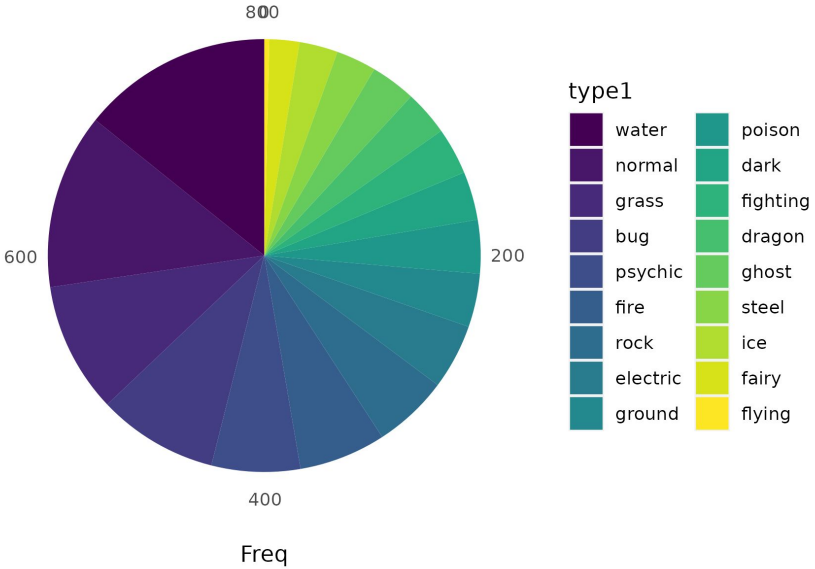
When to use **Bubble Chart**?



- Relationship between three or more variables
- Additional labels might be required

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

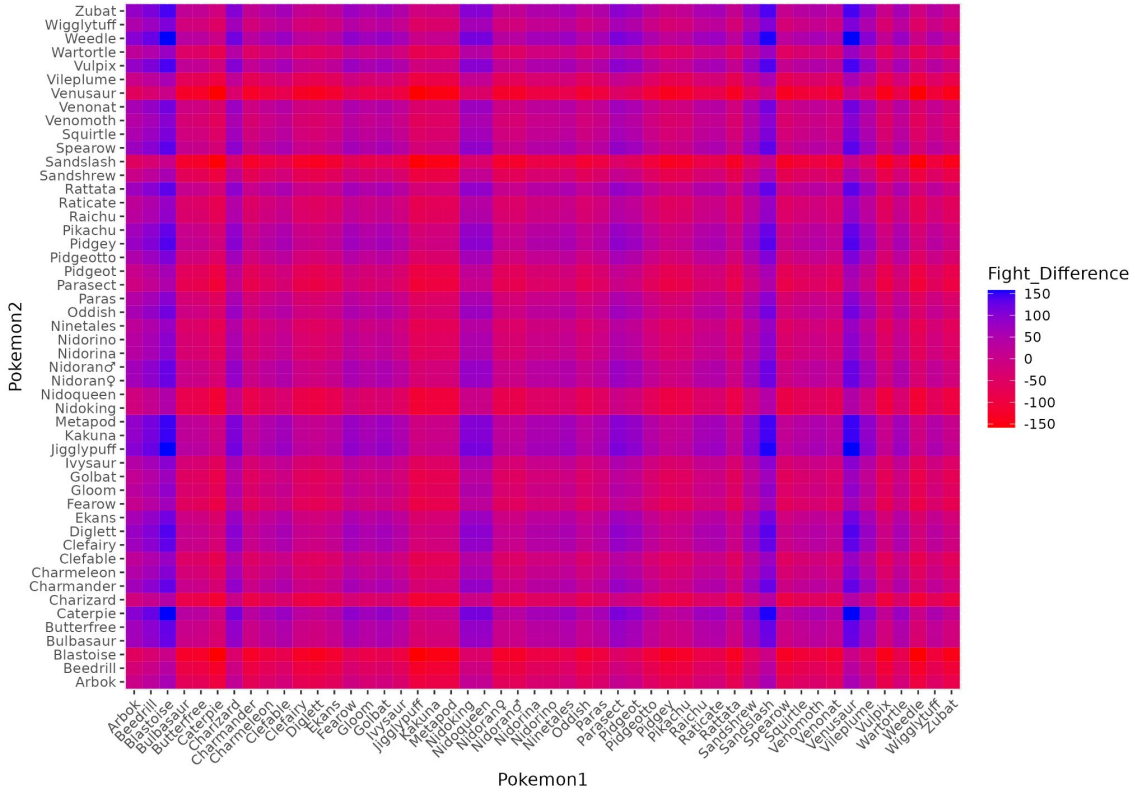
When to use **Pie Chart**?



- Subpopulation comparisons
- Check sum(subpopulations) = total population
- Highlight important subpopulations
- Order subpopulation based on size

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

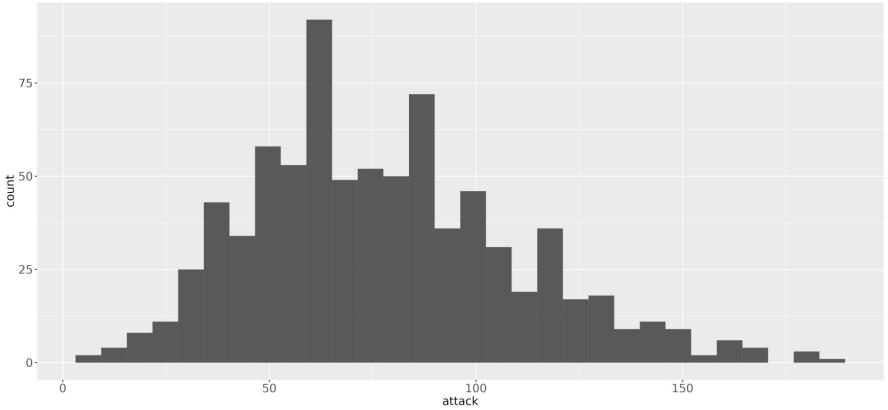
When to use **Heat Map**?

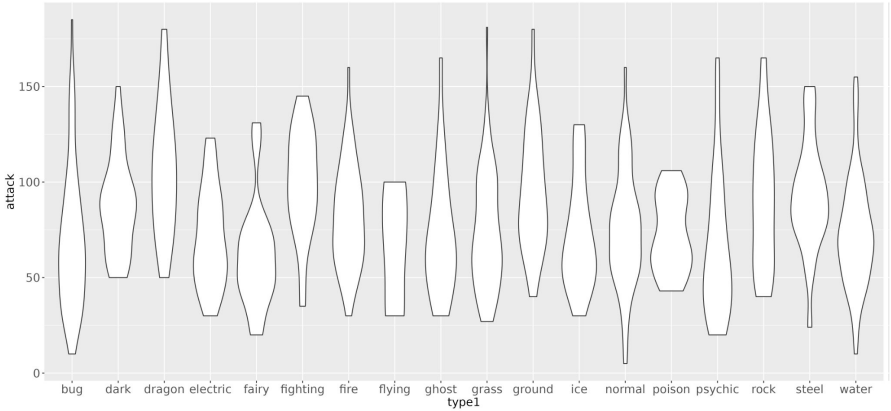


- Spatial or grid based data
- To show density and groups
- Choose an intuitive color gradient (e.g., hot to cold)

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

When to use **Histogram / Box Plot / Violin Plot**?





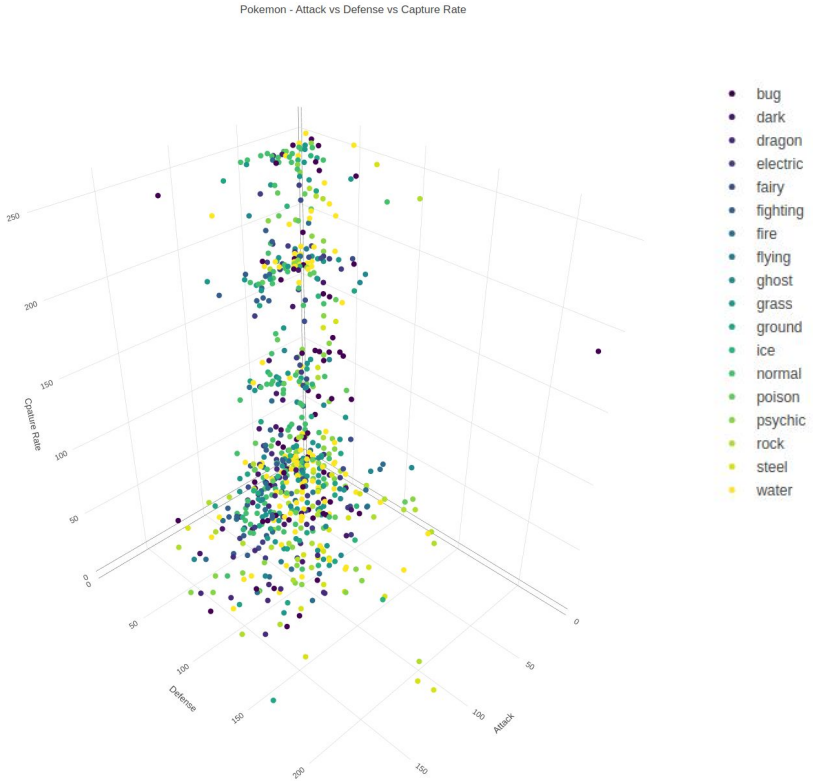
- Investigate distribution of data
- Shape of total population (data) = histogram
- Comparison of subpopulations:
  - Variance and outliers = box plot
  - Shape of multimodal distribution = violin

- Jānis Gulbis
- Kai Tomboc
- Sara A. Metwalli

PRINCIPLES (Audience and Context)

Blank – out of principle

Everything, everywhere, all at once



https://en.wikipedia.org/wiki/Bagel#/media/File:Bagel_with_sesame_3.jpg

[Jacob Johnson - 3 Common Data Visualization Mistakes to Avoid (2023, codecademy)]

Other forms of Inclusion

- Make it clear and concise
- Use dyslexic friendly fonts and styles (e.g., extra space around text)
- Have an alternative to consume your visualization (e.g., table)

- Sarah L. Fossheim - An intro to designing accessible data visualizations
- Kim Marriott et al. (2021, Interactions)
- Tuija Marin - Why should your organization care about accessibility of your data visualizations?

Cognitive Load

Intrinsic Load: Inherent complexity of the data

quantitative	← **Measurement** →	qualitative
certain	← **Knowability** →	uncertain
precise	← **Specificity** →	ambiguous
concrete	← **Relatability** →	abstract

Eva Sibinga & Erin Waldron - Cognitive Load as a Guide: 12 Spectrums to Improve Your Data Visualizations (2021, Nightingale)

Cognitive Load

Germane Load: Familiarity of the audience to new Information

intentional ←	**Connection**	→ coincidental
slow ←	**Pace**	→ fast
expert ←	**Knowledge**	→ novice
confident ←	**Confidence**	→ anxious

Eva Sibinga & Erin Waldron - Cognitive Load as a Guide: 12 Spectrums to Improve Your Data Visualizations (2021, Nightingale)

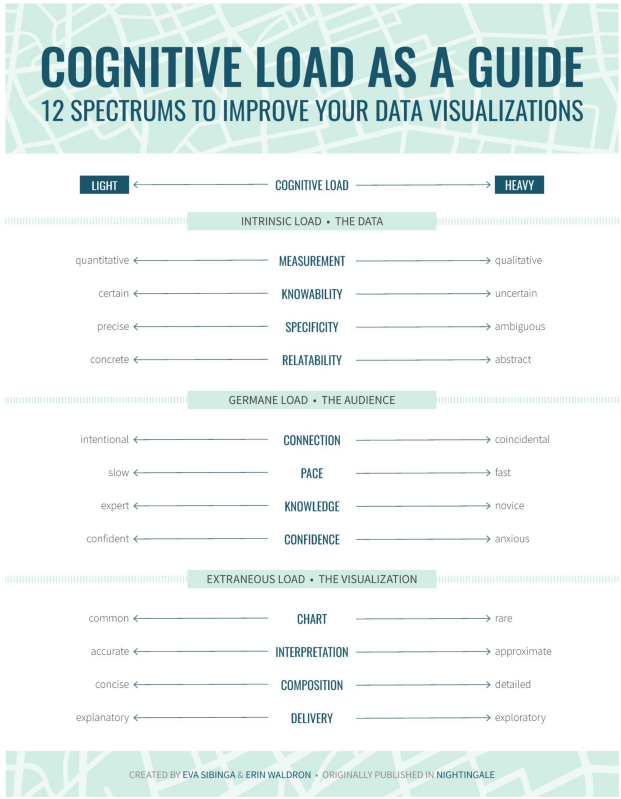
Cognitive Load

Extraneous Load: Form of Visualization

common	**Chart Type**	rare
accurate	**Interpretation**	approximate
concise	**Composition**	detailed
explanatory	**Delivery**	exploratory

Eva Sibinga & Erin Waldron - Cognitive Load as a Guide: 12 Spectrums to Improve Your Data Visualizations (2021, Nightingale)

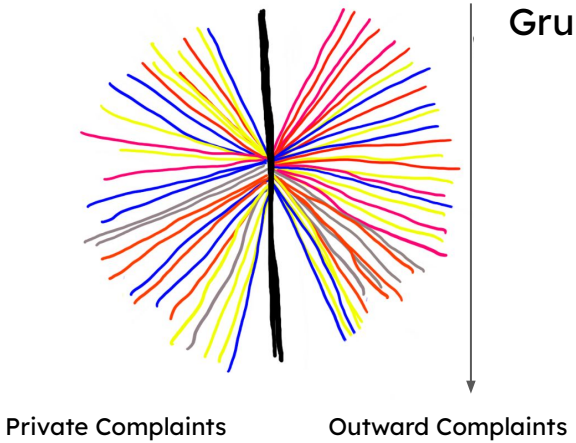
Cognitive Load



[Eva Sibinga & Erin Waldron - Cognitive Load as a Guide: 12 Spectrums to Improve Your Data Visualizations (2021, Nightingale)]

Data Humanism

“Instead of using data just to become more efficient, we argue we can use data to become more human and to connect with ourselves and others at a deeper level.” - Giorgia Lupi



Dear Data:
Week of Complaints and general Grumpiness

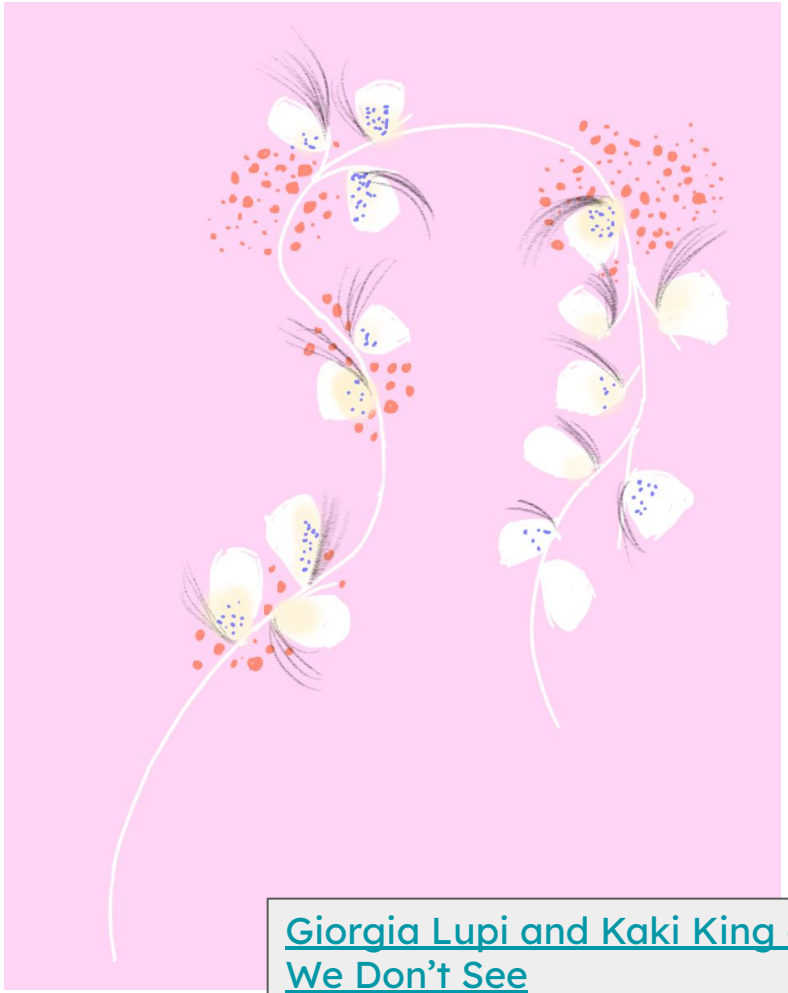
[Giorgia Lupi and Stefanie Posavec - Dear Data]

Data Humanism

“Can a data visualization evoke empathy and activate us at an emotional level, and not only at a cognitive one?” - Giorgia Lupi

“Can looking at a data visualization make you feel part of a story of someone’s life?” - Giorgia Lupi

“But sometimes we completely forget these softer, more intimate, even fuzzy type of data.” - Giorgia Lupi

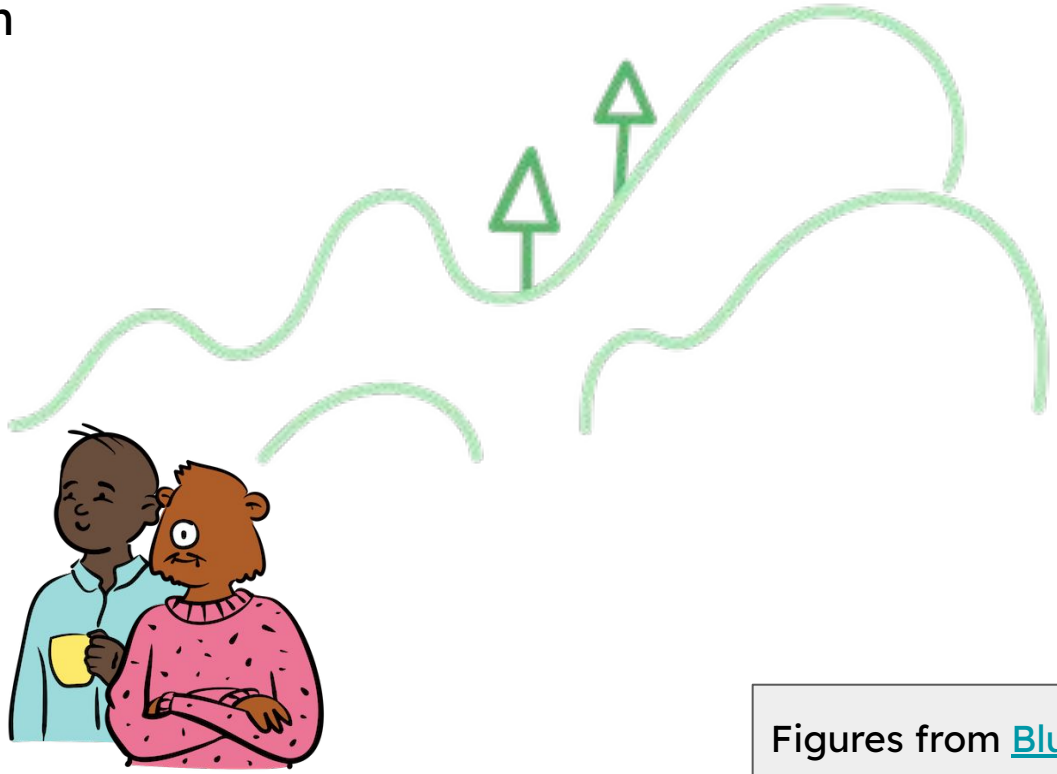


Idiopathic Thrombocytopenic Purpura (ITP)

[Giorgia Lupi and Kaki King — Bruises: The Data We Don’t See]

Data visualization is a form of communication

- Data + Visualization form
- Audience
- Location / Context



Figures from Blush