Lecture #4: Exploratory data analysis (EDA)

Slides by Artem Dembitskiy

Previously on

- Python for atomistic modeling
  - ASE's Atoms and Pymatgen's Structure
  - Neighbor list
  - Voronoi partitioning
- Data in materials informatics
  - Computational data
  - The Materials project API

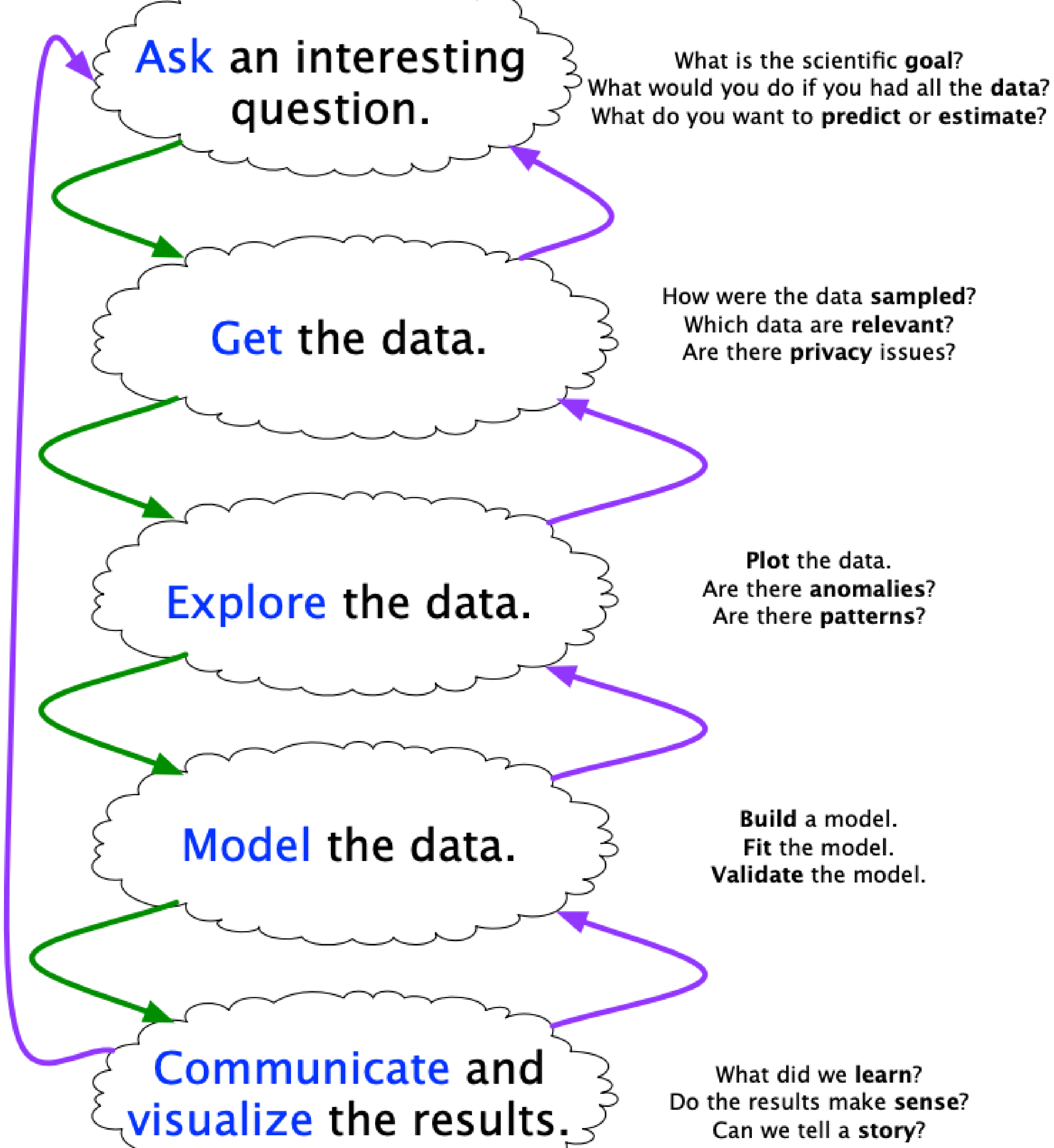
Goals/Agenda

- Explain why visualizing data is important when analyzing data
- Provide tips on how to use visualization to explore data

Attribution

- Parts of these slides are adopted from the excellent lecture on exploratory data analsysis from the course CS 109A: Introduction to Data Science by Pavlos Protopapas & Kevin Rader shared under MIT licence
  - https://harvard-iacs.github.io/2018-CS109A/lectures/lecture-3/presentation/lecture3.pdf
- Consider the following materials your reading homework

The data science workflow



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

From CS 109a: Data Science, Effective Exploratory Data Analysis and Visualization by Pavlos Protopapas & Kevin Rader slide #2

Descriptive statistics

"...is a summary statistic that quantitatively describes or summarizes features from a collection of information"

https://en.wikipedia.org/wiki/Descriptive_statistics

Sample size

Number of observations in a dataset (study)

```
len(data)
```

Mean

```
np.mean(data)
```

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median

```
np.median(data)
```

"The median of a set of numbers is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution."

$$1, 3, 3, \mathbf{6}, 7, 8, 9$$

$$\text{Median} = \underline{\underline{6}}$$

$$1, 2, 3, \mathbf{4}, \mathbf{5}, 6, 8, 9$$

$$\begin{aligned}\text{Median} &= (4 + 5) \div 2 \\ &= \underline{\underline{4.5}}\end{aligned}$$

Standard deviation

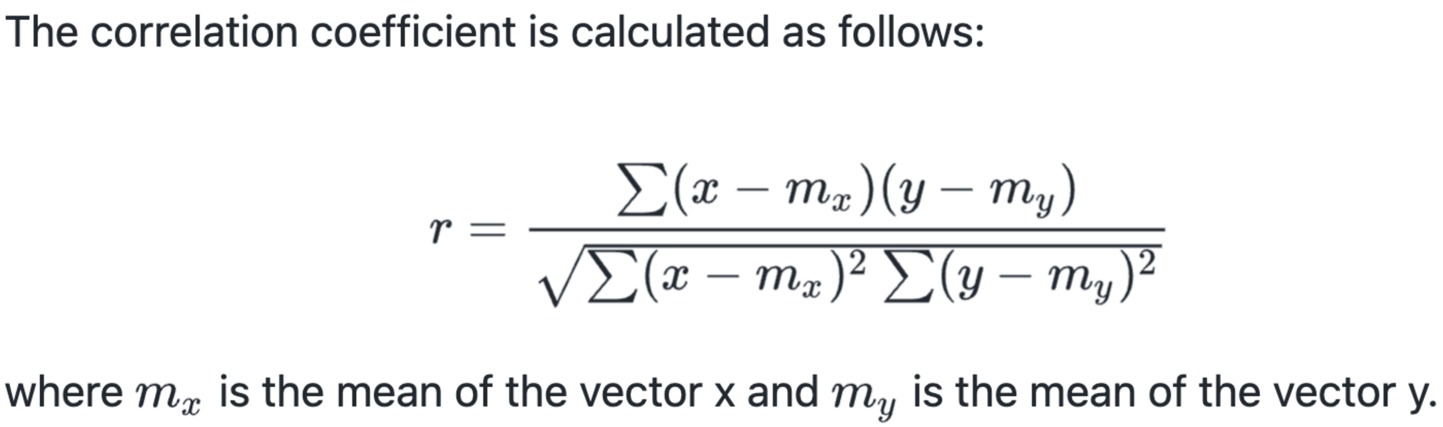
"...is a measure of the amount of variation of the values of a variable about its mean."

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

Correlation coefficient

The Pearson correlation coefficient measures the linear relationship between two datasets. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation."

scipy docs



The correlation coefficient is calculated as follows:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

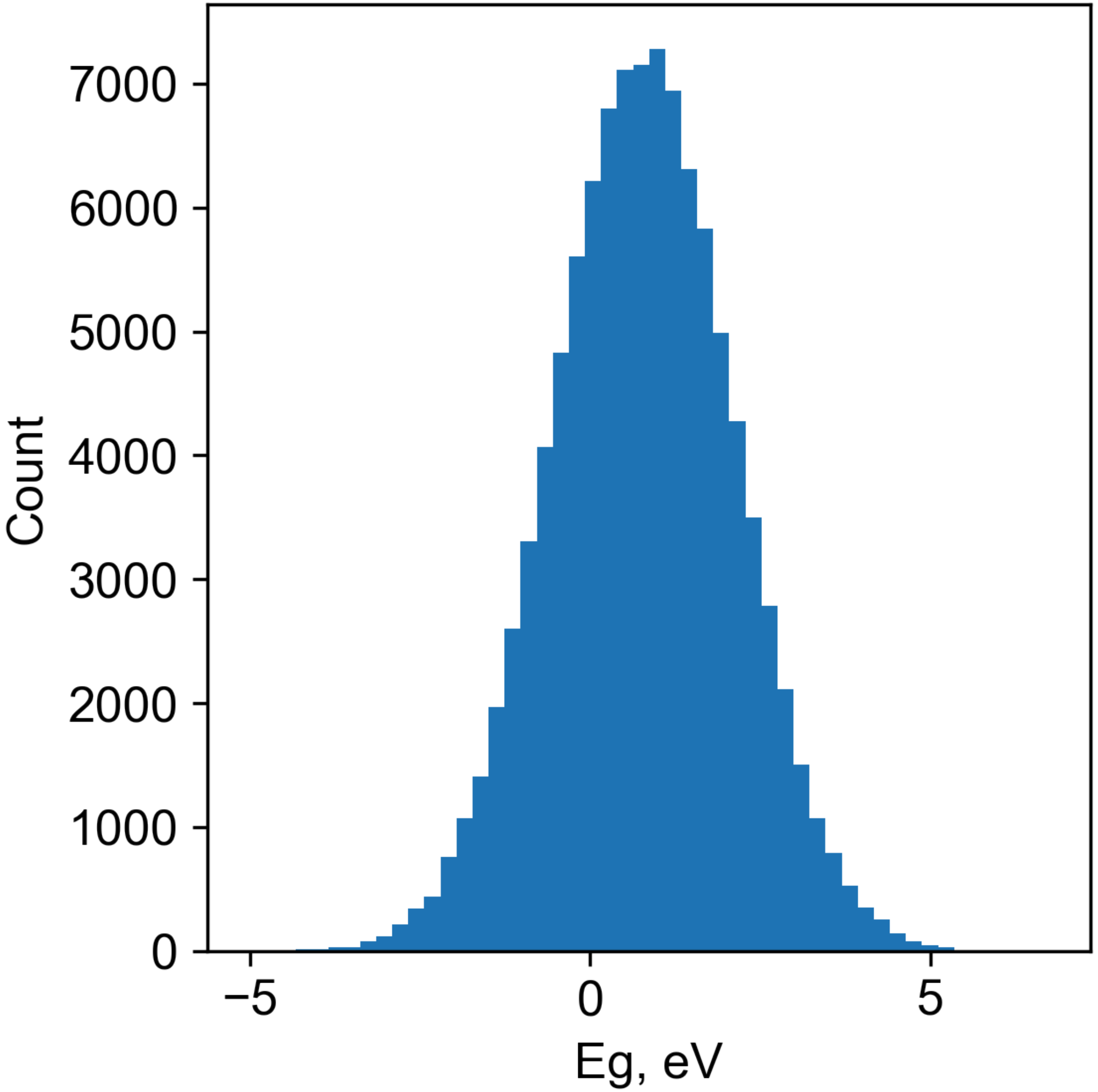
where m_x is the mean of the vector x and m_y is the mean of the vector y.

Descriptive statistics of band gap (Eg) distribution in the Materials Project

- Sample size
    - 103,217
- Mean of Eg
    - 0.79
- Standard deviation of Eg:
    - 1.37

Is it what you expected?

- What's wrong with this distribution?



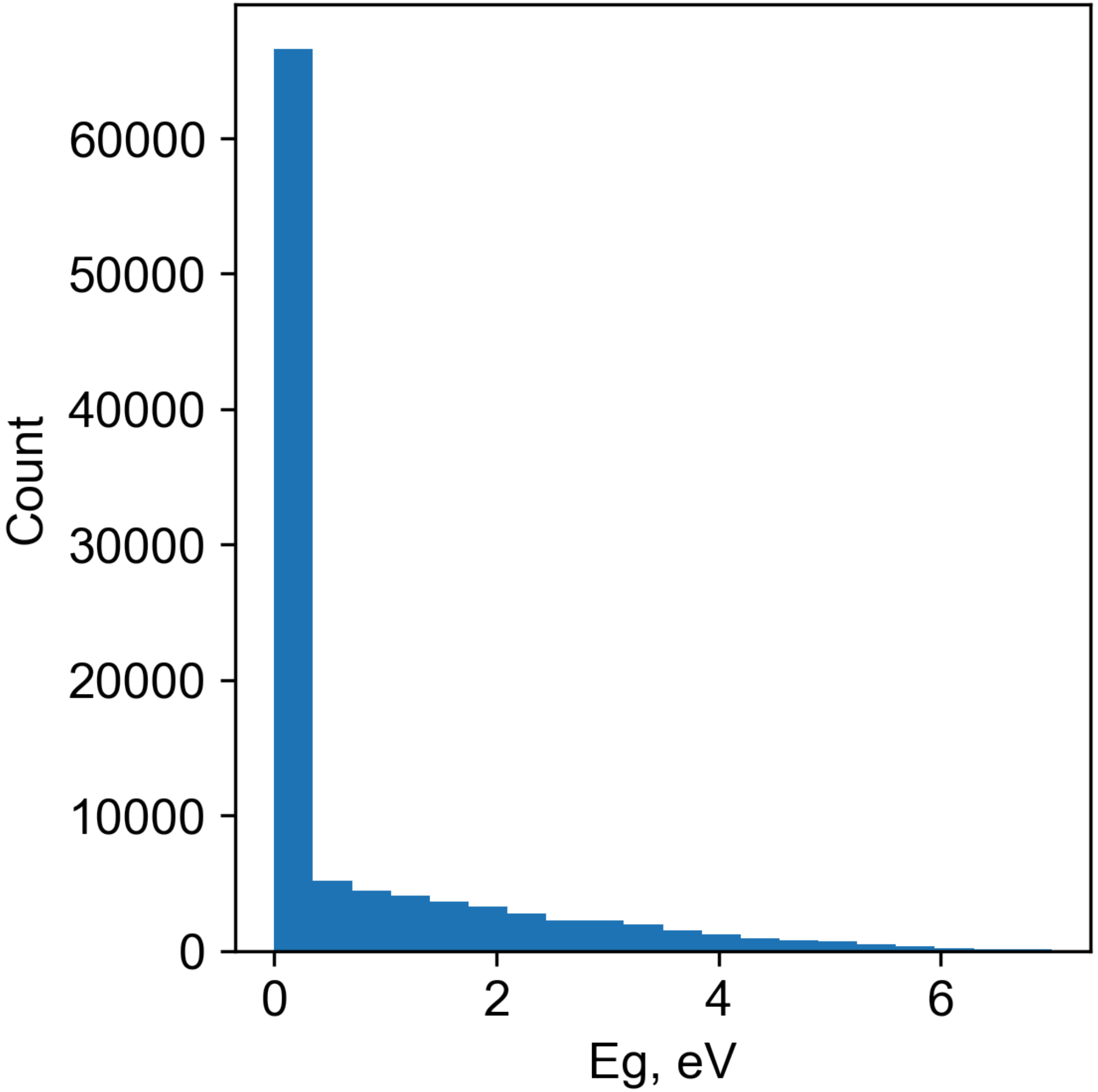
Any ideas?

- Sample size
    - 103,217
- Median of Eg:
    - 0.0 <--- ???

This is the real distribution

- Metals have a zero Eg

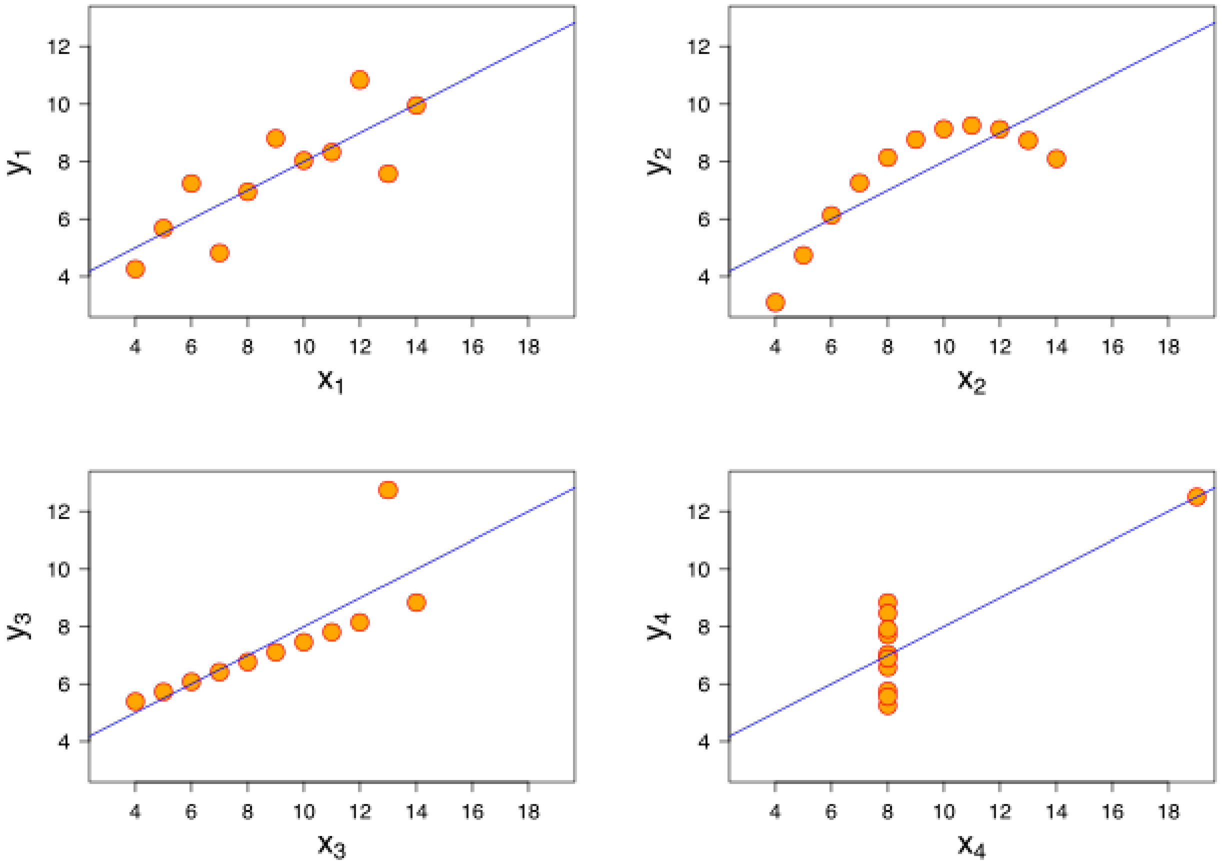
> Median(Eg) = 0.0 says that metals represent at least half of the sample



Why is visual inspection of data important?

- Same descriptive statistics
- Very different distributions

https://en.wikipedia.org/wiki/Anscombe's_quartet



Visulaization goals

Communicate (Explanatory)

- Present data and ideas
- Explain and inform
- Provide evidence and support
- Influence and persuade

Analyze (Exploratory)

- Explore the data
- Assess a situation
- Determine how to proceed
- Decide what to do

From CS 109a: Data Science, Effective Exploratory Data Analysis and Visualization by Pavlos Protopapas & Kevin Rader slide #23

Communicate

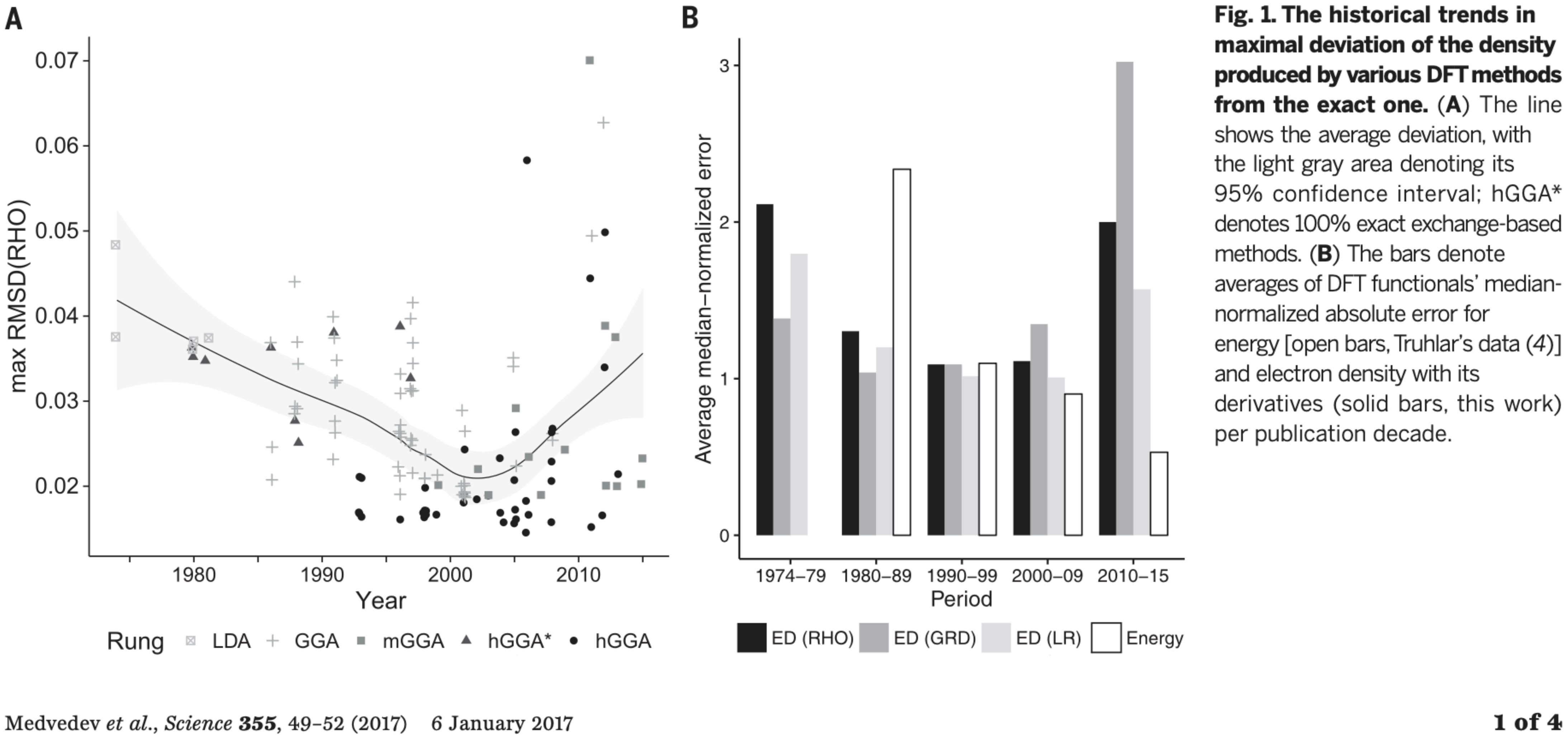


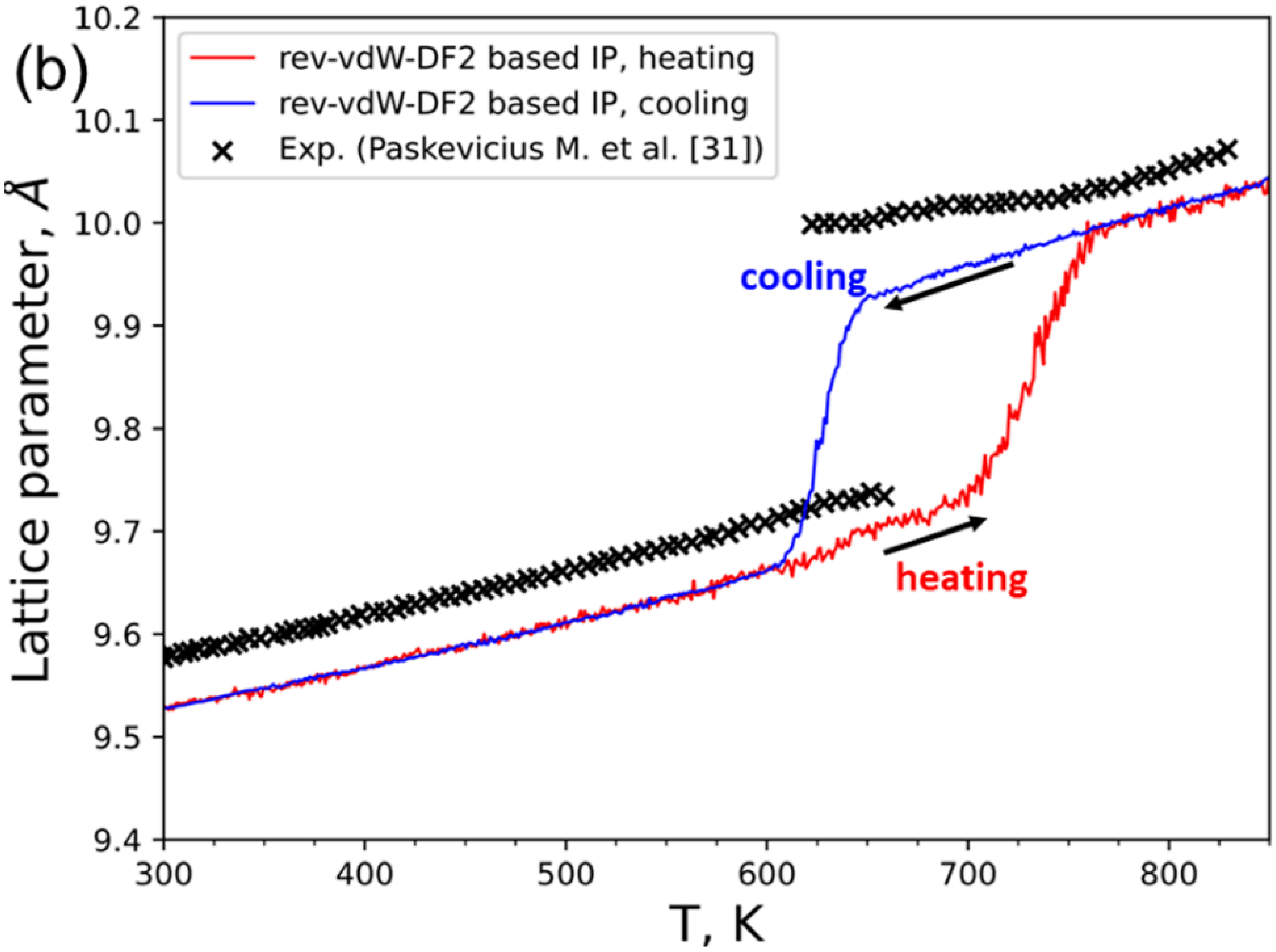
Fig. 1. The historical trends in maximal deviation of the density produced by various DFT methods from the exact one. (**A**) The line shows the average deviation, with the light gray area denoting its 95% confidence interval; hGGA* denotes 100% exact exchange-based methods. (**B**) The bars denote averages of DFT functionals' median-normalized absolute error for energy [open bars, Truhlar's data (*4*)] and electron density with its derivatives (solid bars, this work) per publication decade.

Medvedev *et al.*, *Science* **355**, 49–52 (2017)   6 January 2017

1 of 4

Density functional theory is straying from the path toward the exact functional

Explore



Order–Disorder Phase Transition and Ionic Conductivity in a Li2B12H12 Solid Electrolyte

Exploratory data analysis pipeline

- Build data
- Clean data
- Explore global features
- Explore group features

Build (read) data in a structured format

- Pandas DataFrame
- One row per variable

```
df = pd.read_csv('eg_data.csv')
```

Clean the data

- outliers
- NaNs (missing values)
- constant rows
- duplicates

```
df.dropna()
```

- plus visual support: histogram, box plot

Study the global summary statistics

```
df.describe()
```

```
df.aggregate(
                {
                "column_name": ["min", "max", "median", "skew"]
                }
)
```

- plus visual support: histogram, scatter plot, bar plot

Study the summary statistics of the subgroups

```
df[["bandgap, chemsys"]].groupby("chemsys").mean()
```

- plus visual support: histogram, scatter plot, bar plot

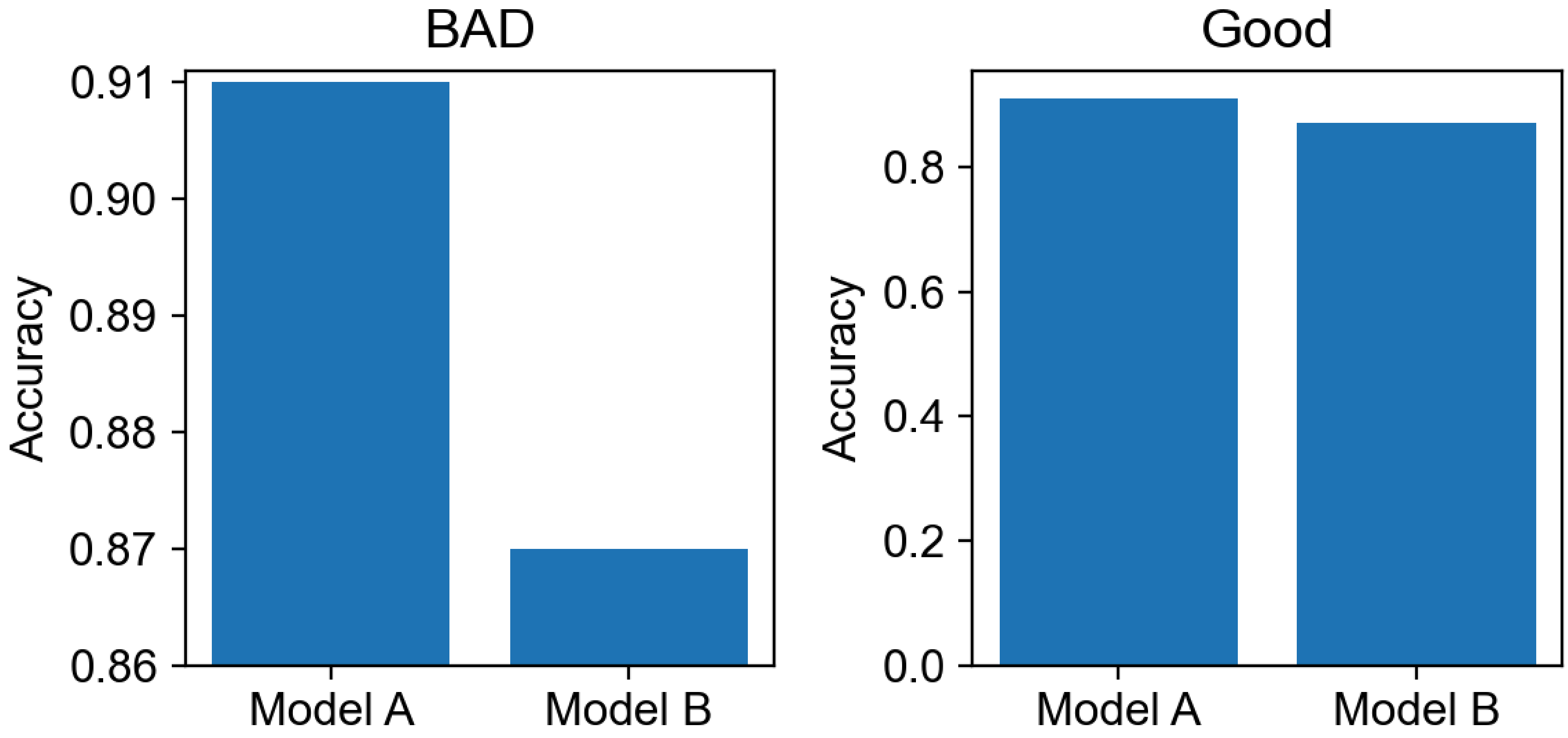
Some principles for effective EDA

Avoid misleading graphs

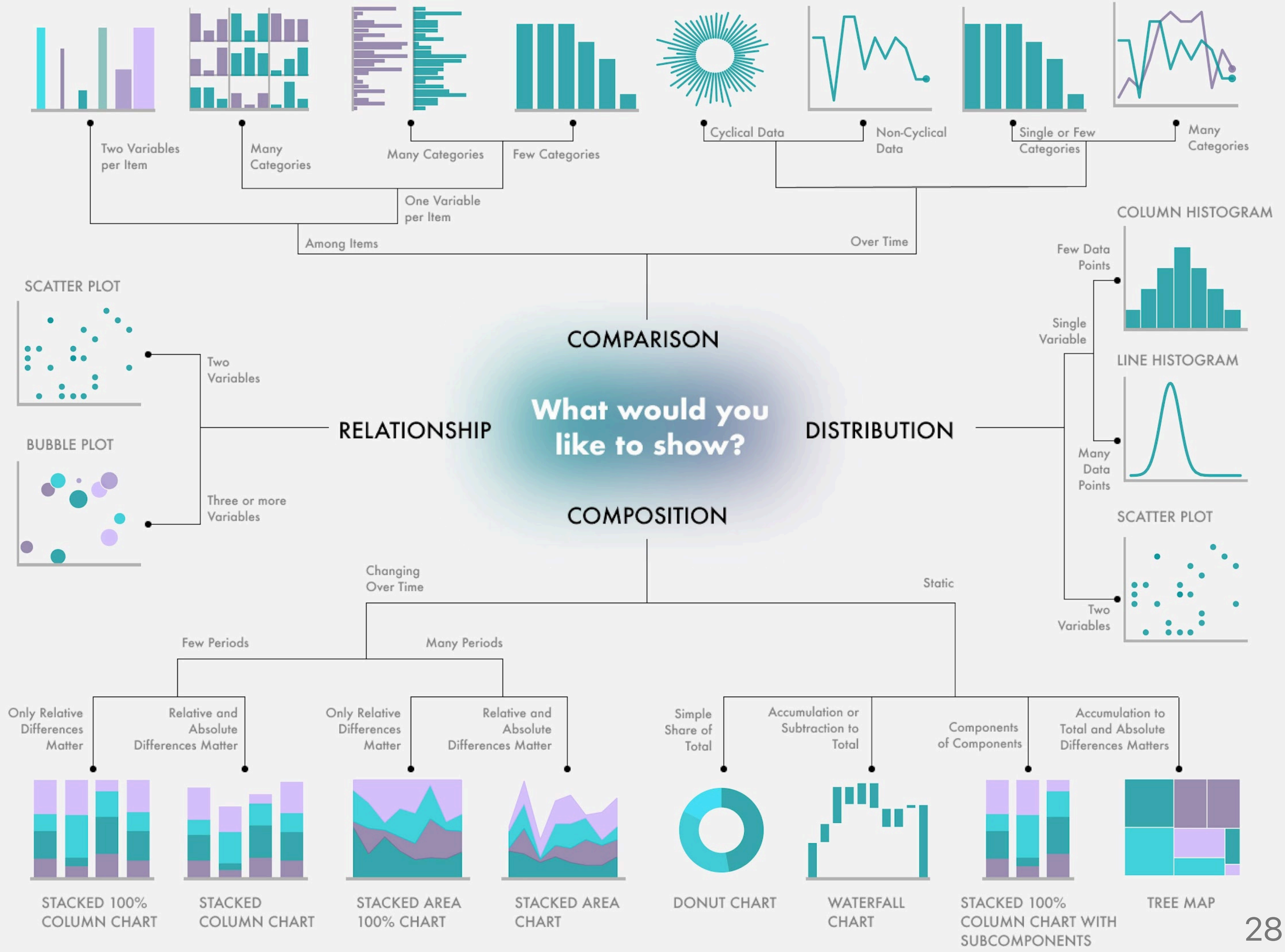
- Do not distort scales
- Do not truncate graph when comparing the data
  - or indicate the truncation
- Avoid 3D charts
- Do not change y(or x)-axis maximum
- Aspect ratio determines the perception of steepness in slope
  - be proportional

Have a look at this page: https://en.wikipedia.org/wiki/Misleading_graph

Lie factor


BAD
Accuracy
0.91
0.90
0.89
0.88
0.87
0.86
Model A
Model B
Good
Accuracy
0.8
0.6
0.4
0.2
0.0
Model A
Model B

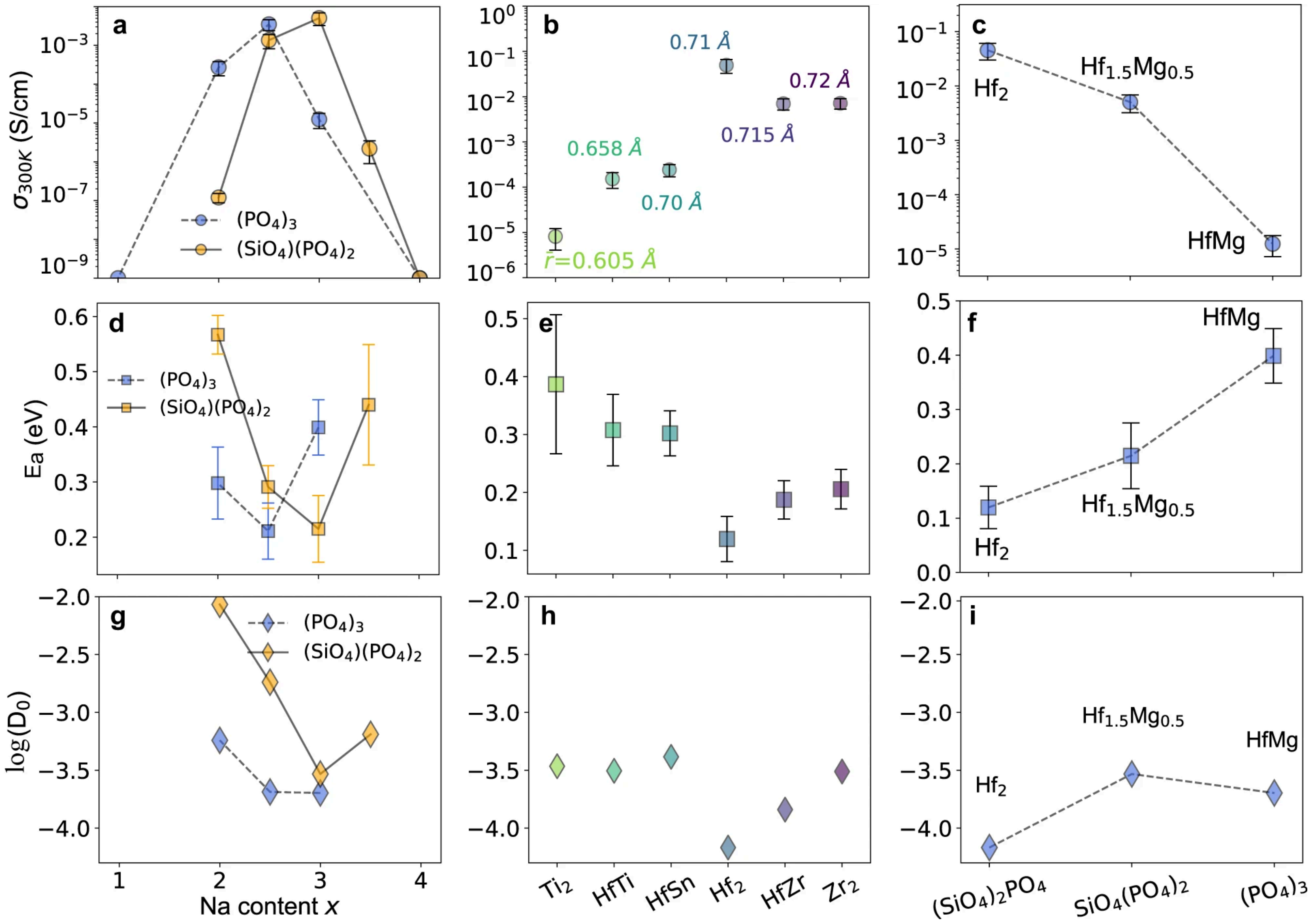
Use the right display



Correlations

- scatter plot, correlation matrix

Is it a good graph? Why?

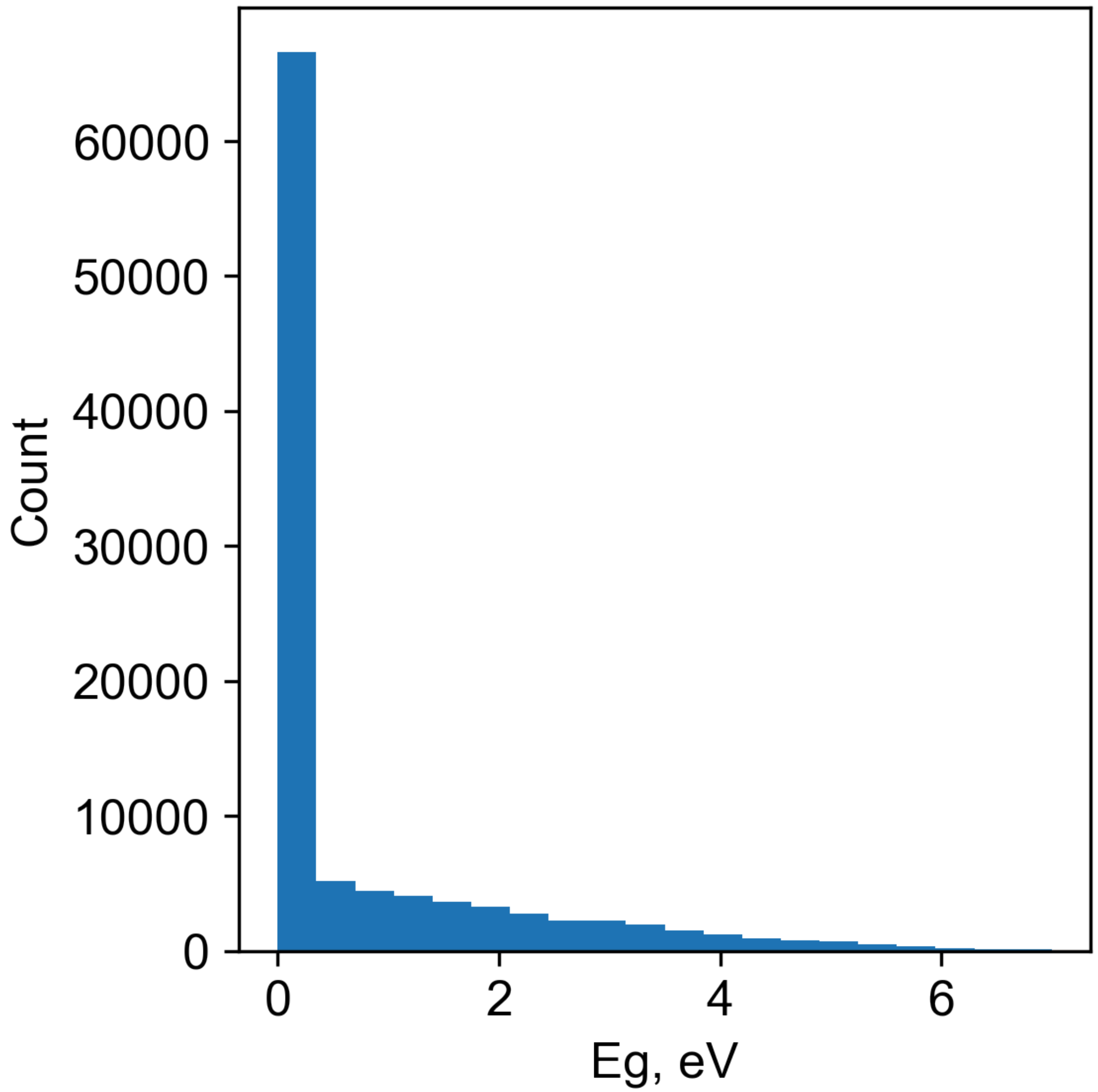


Design principles for NASICON super-ionic conductors

Distribution

- histogram, density plot

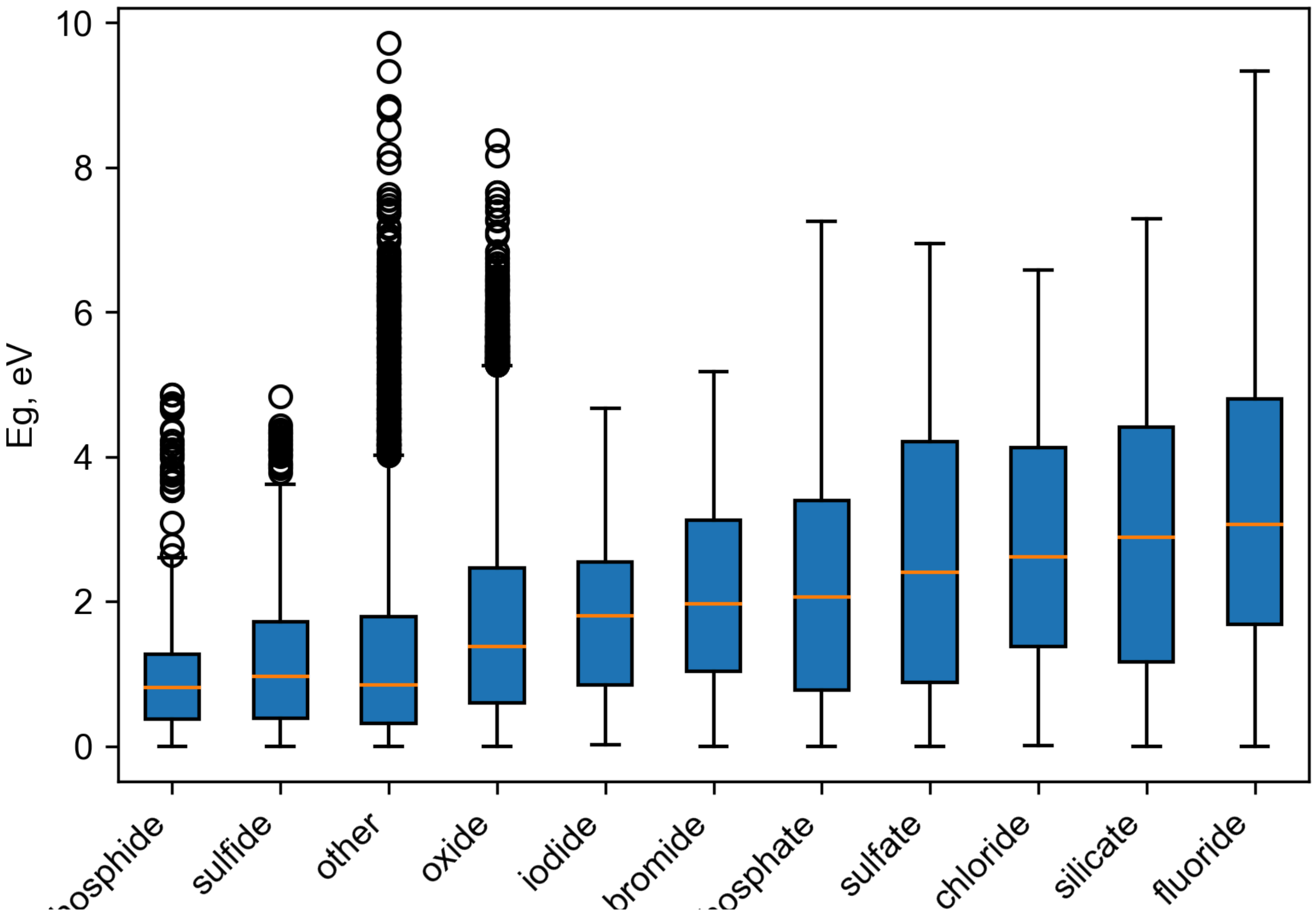
Is it a good graph? Why?



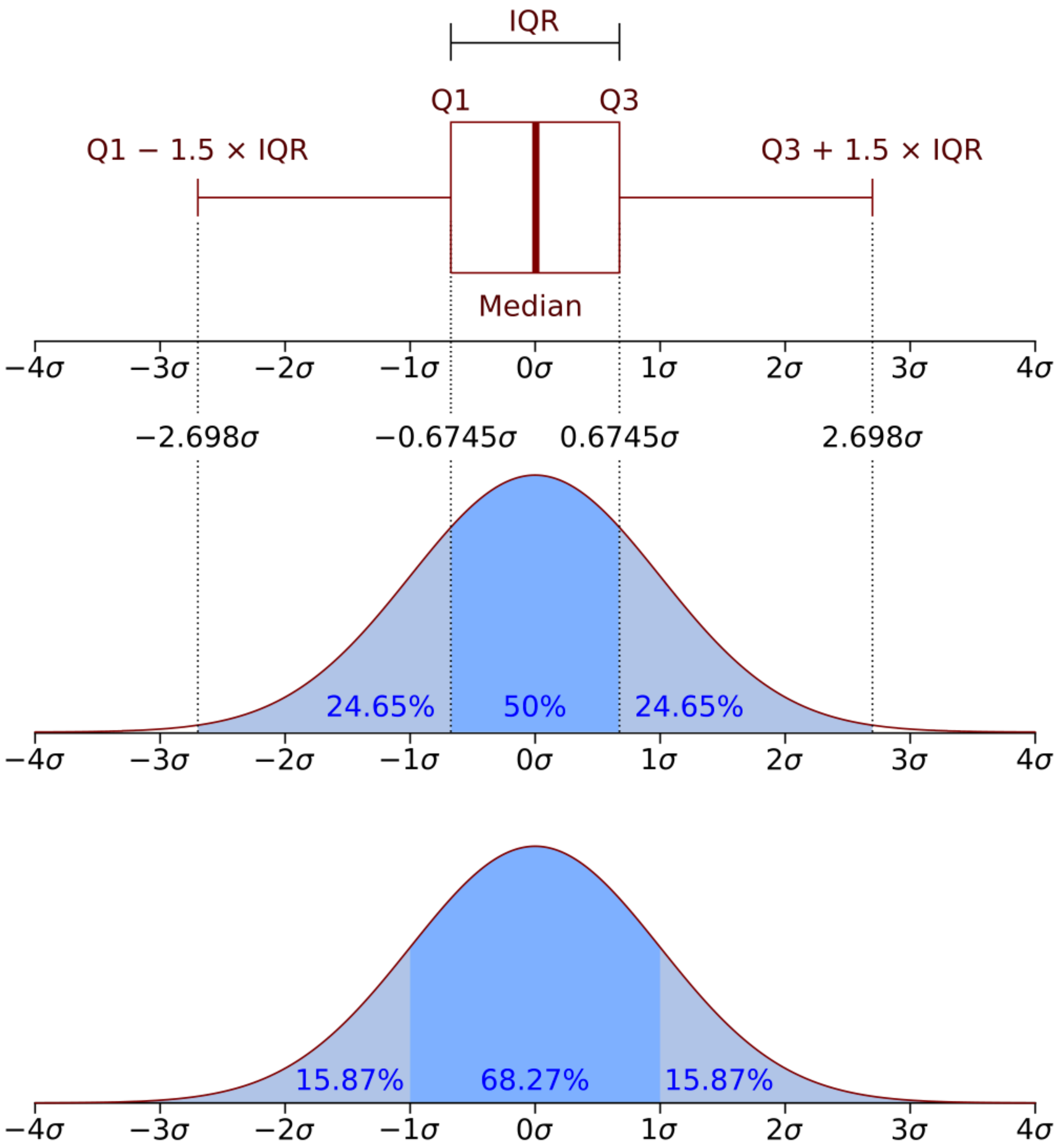
Comparison

- bar plot, box plot

Is it a good graph? Why?

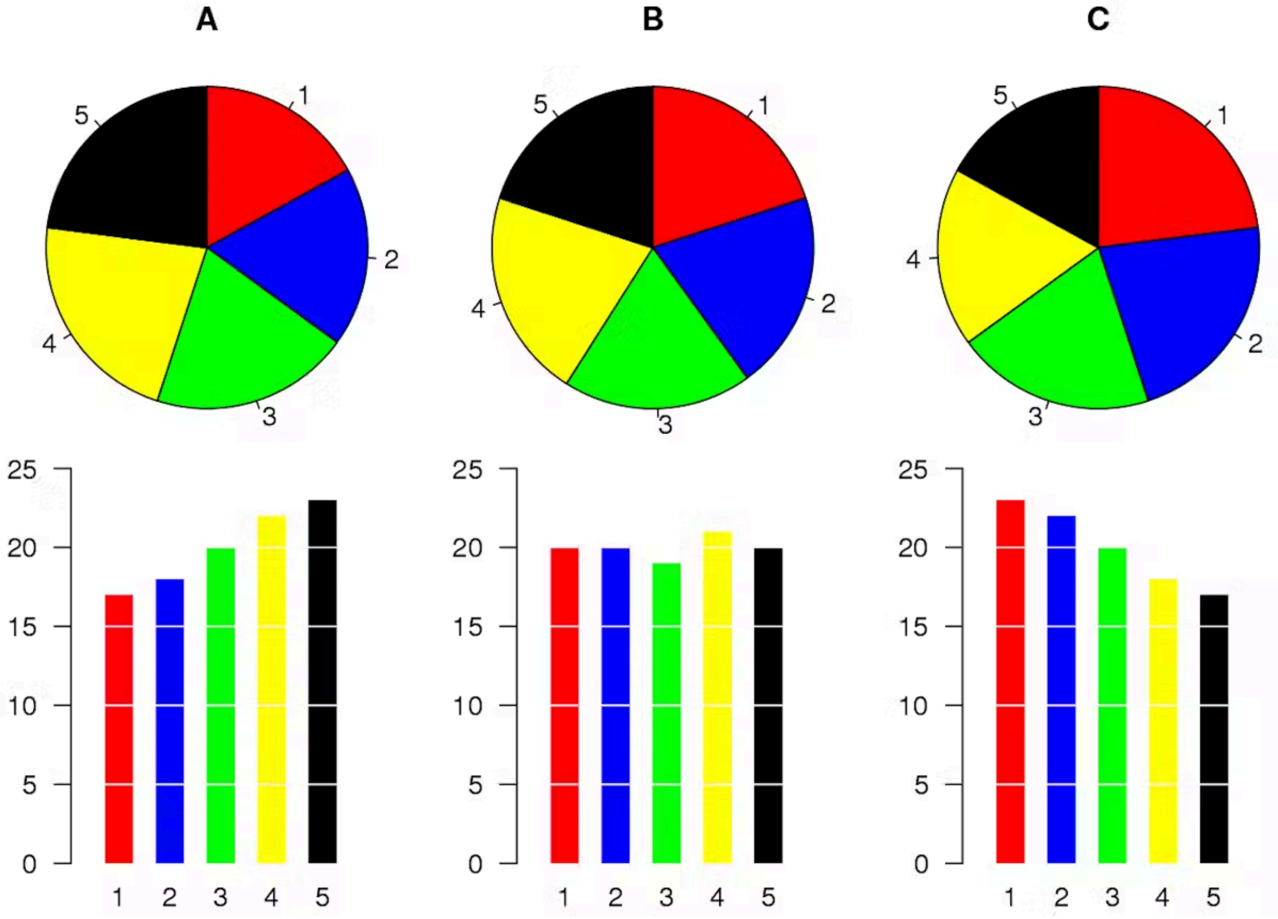


Box plot


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
15.87%
68.27%
15.87%
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ

Don't use pie charts

Barplots are easier to compare

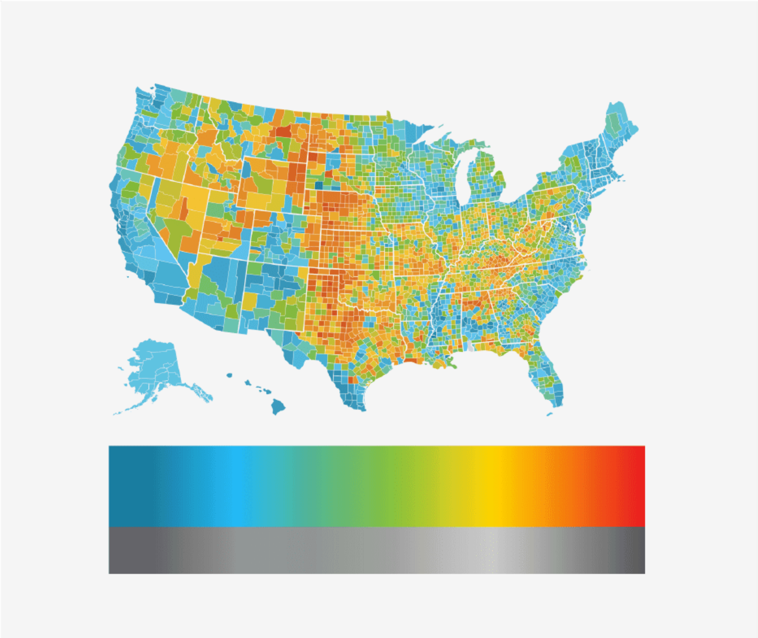


Here's why you should (almost) never use a pie chart for your data

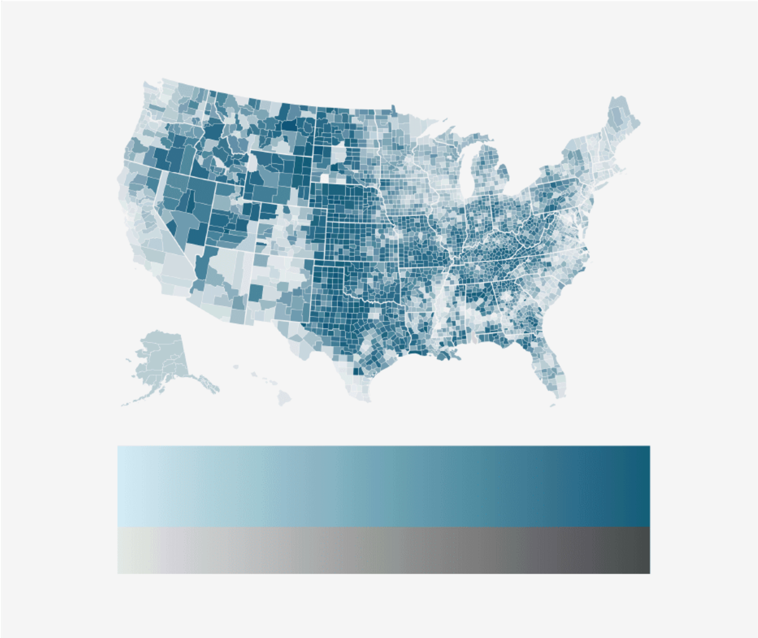
Use color

Have a look at this page:

https://blog.datawrapper.de/colors/



NOT IDEAL

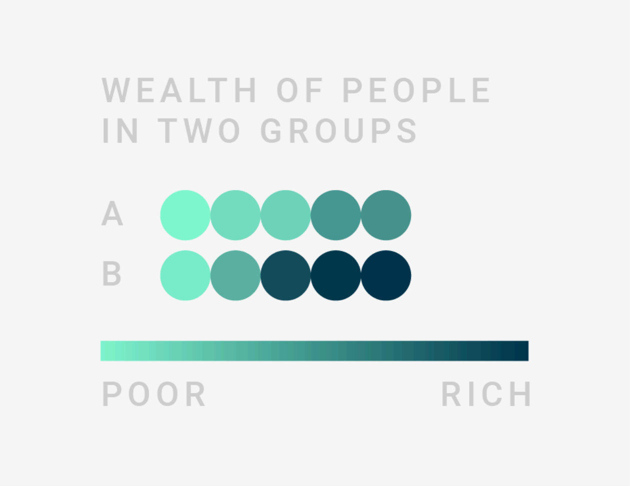


BETTER

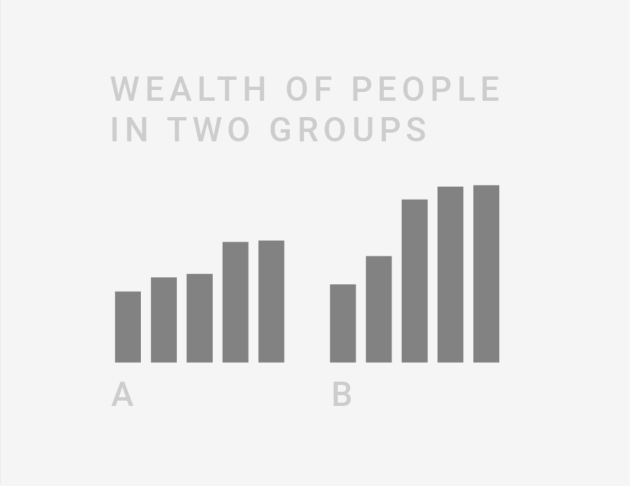
But consider a better alternative if possible

- the simpler the better

Have a look at this page:
https://blog.datawrapper.de/colors/

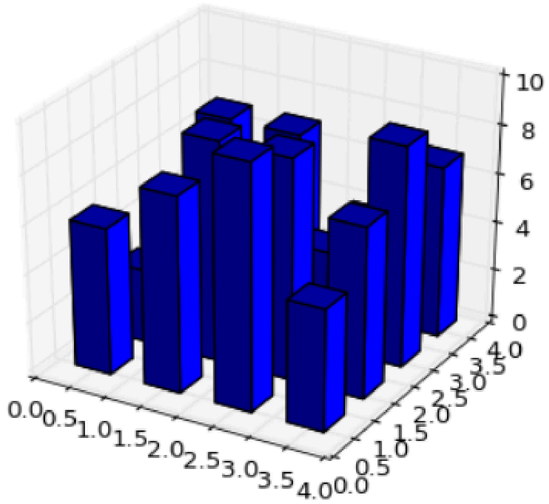


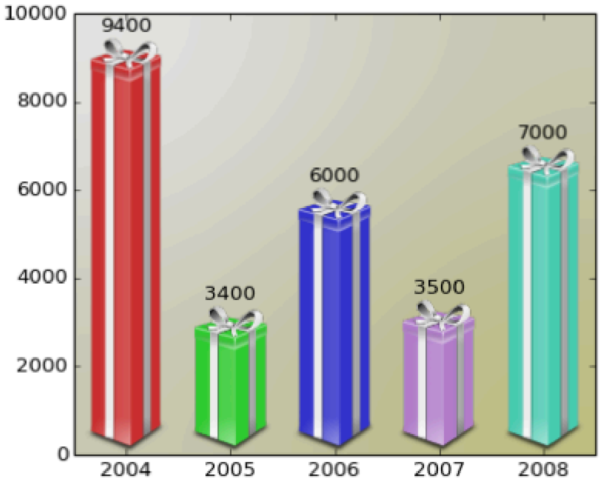
NOT IDEAL



BETTER

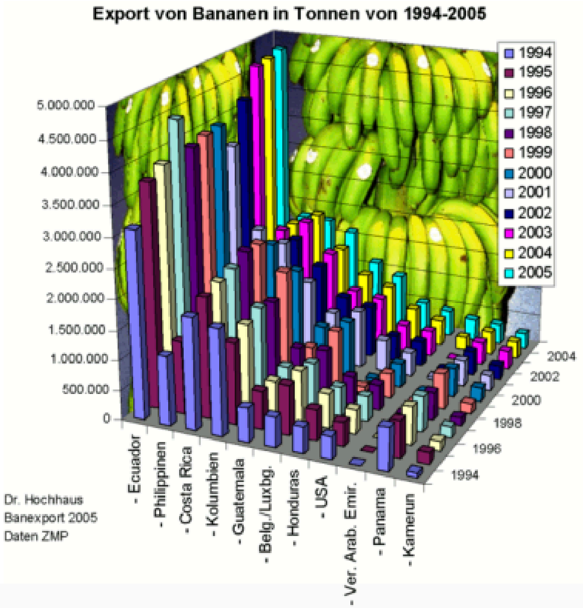
Don’t!

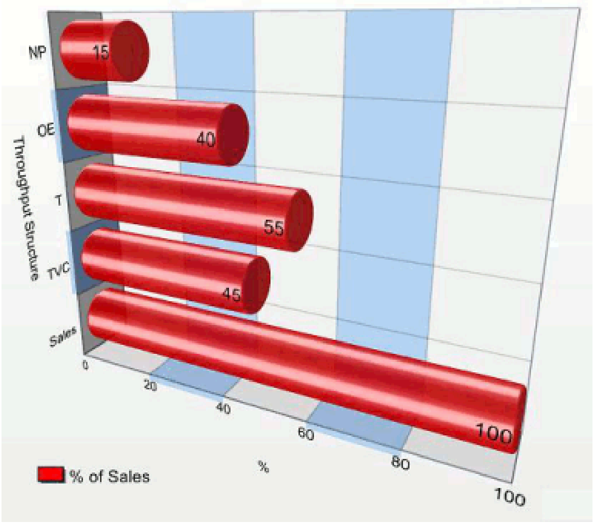




matplotlib gallery

My favorite





Excel Charts Blog

From CS 109a: Data Science, Effective Exploratory Data Analysis and Visualization by Pavlos Protopapas & Kevin Rader slide #55

Take home message

- Visualizing data helps you
  - Present data and ideas
  - Analyze results
  - Define future steps
- The data is more important than the design
  - Represent the data in a right way
  - Avoid misleading graphs

Resources:

https://harvard-iacs.github.io/2018-CS109A/lectures/lecture-3/presentation/lecture3.pdf

https://en.wikipedia.org/wiki/Misleading_graph

https://en.wikipedia.org/wiki/Anscombe's_quartet

https://blog.datawrapper.de/colors/

Thank you for your attention!