m.sc. data science

data visualisation lab

a.y. 2024/2025

there is a magic in graphs. the curve informs the mind, awakens the imagination, convinces. (*h.d. hubbard, 1939*)

Monica Moroni

who

Monica Moroni

m.sc. maths 2015

ph.d. comp neuroscience 2016-2019

postdoc comp neuroscience 2019-2022

data scientist 2022-today

mmoroni@fbk.eu

Shahryar Noei

m.sc. biomedical engineering 2017

data scientist 2017-2018

PhD comp neuroscience 2018-2022

data scientist 2023-now

snoei@fbk.eu

logistics

4 March - 27 May

Lecture in presence

TH:Tuesdays 8:30-10:30
Room B108

EX:Wednesdays 10:30-12:30
Room A203
w/ dr. ***Shahryar Noei***

contacts

mmoroni@fbk.eu

snoei@fbk.eu

office hours:
Available on appointment
(send email)
thesis opportunities

course material:
`Moodle`

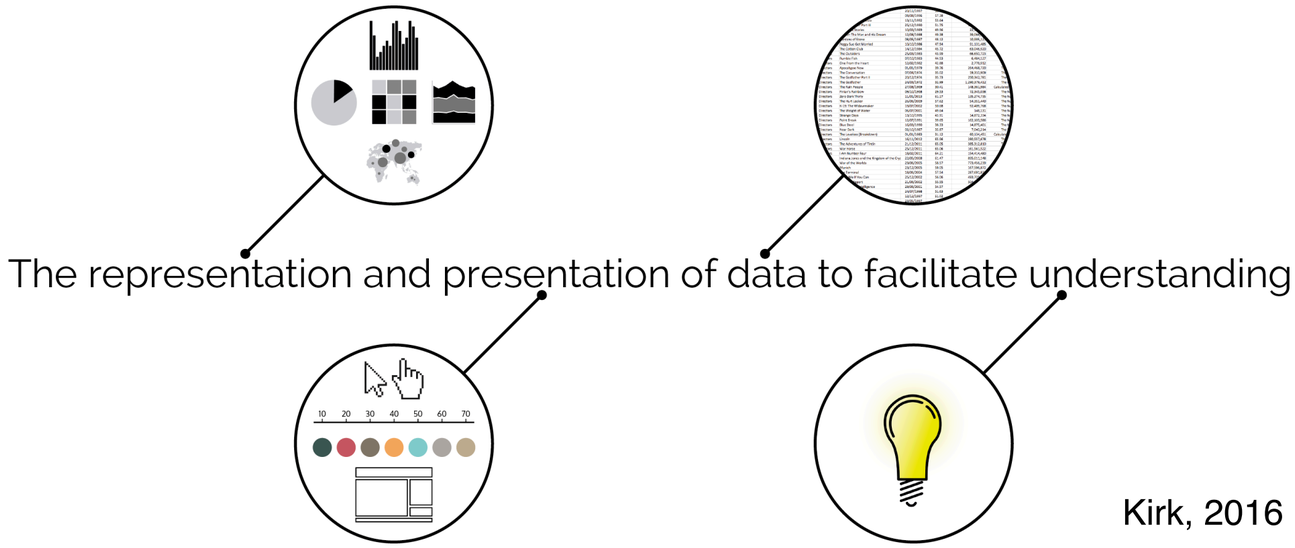
exam:

- written theoretical questions + practical R/Python exercises
- no mid-term tests
- no projects

what

data visualisation elements

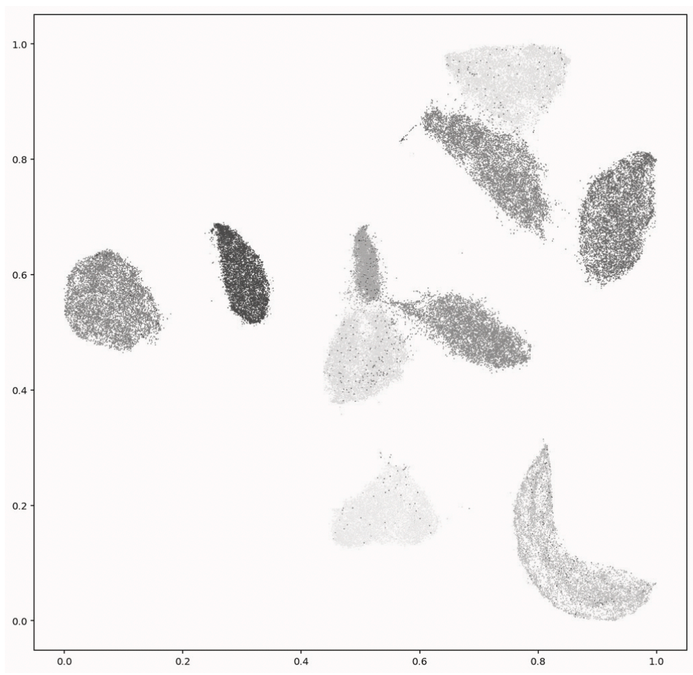
the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics



Kirk, 2016

dimensionality reduction algorithms

n-dim data



why

… off the beaten path

information requires…

salience

the relevance of information for an actor's decision choices, or for the choices that affect a given stakeholder.

credibility

whether an actor perceives information as meeting standards of scientific plausibility and technical adequacy.

legitimacy

whether an actor perceives the process in a system as unbiased and meeting standards of political and procedural fairness.

… but scientists and policy makers give them different interpretations

why

… off the beaten path

Big Data, Vol. 4, No. 2 | Editorial

Visualization for Data Science: Adding Credibility, Legitimacy, and Saliency

Ross Maciejewski and Douglas C. Montgomery

Published Online: 17 Jun 2016 | https://doi.org/10.1089/big.2016.29007.vis

visualisation provides

of visualization becomes vital in data science. What visualization provides are methods that enable end users and decision makers to interact with underlying statistical algorithms, apply the algorithms at the appropriate scales, deal with noisiness inherent in real-world data, and then apply their domain knowledge to reason about the outcomes of the analysis. In terms of credibil-

why

… off the beaten path

Big Data, Vol. 4, No. 2 | Editorial

Visualization for Data Science: Adding Credibility, Legitimacy, and Saliency

Ross Maciejewski and Douglas C. Montgomery

Published Online: 17 Jun 2016 | https://doi.org/10.1089/big.2016.29007.vis

salience

enable analysts to choose meaningful features based on their own understanding of the questions and problems being addressed

credibility

communicate underlying distributions of the data variables, highlight potential errors, and provide interactions for the analyst to correct the errors.

legitimacy

enhanced through the user engagement that visualisation encourages

why

back down to earth: the anscombe quartet

x	y
10	8,04
8	6,95
13	7,58
9	8,81
11	8,33
14	9,96
6	7,24
4	4,26
12	10,84
7	4,82
5	5,68

number of observations	11
mean of x	9.0
mean of y	7.5
correlation between x and y	0.816
regression line	y=0.5x+3
variance of x	11
variance of y	4.12

why

back down to earth: the anscombe quartet

x	y
10	9,14
8	8,14
13	8,74
9	8,77
11	9,26
14	8,10
6	6,13
4	3,10
12	9,13
7	7,26
5	4,74

number of observations	11
mean of x	9.0
mean of y	7.5
correlation between x and y	0.816
regression line	y=0.5x+3
variance of x	11
variance of y	4.12

why

back down to earth: the anscombe quartet

x	y
10	7,46
8	6,77
13	12,74
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73

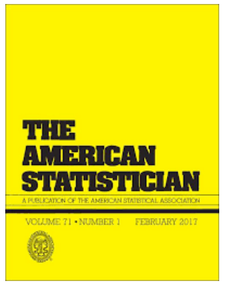
number of observations	11
mean of x	9.0
mean of y	7.5
correlation between x and y	0.816
regression line	y=0.5x+3
variance of x	11
variance of y	4.12

why

back down to earth: the anscombe quartet

x	y
8	6,58
8	5,76
8	7,71
8	8,84
8	8,47
8	7,04
8	5,25
19	12,50
8	5,56
8	7,91
8	6,89

number of observations	11
mean of x	9.0
mean of y	7.5
correlation between x and y	0.816
regression line	y=0.5x+3
variance of x	11
variance of y	4.12

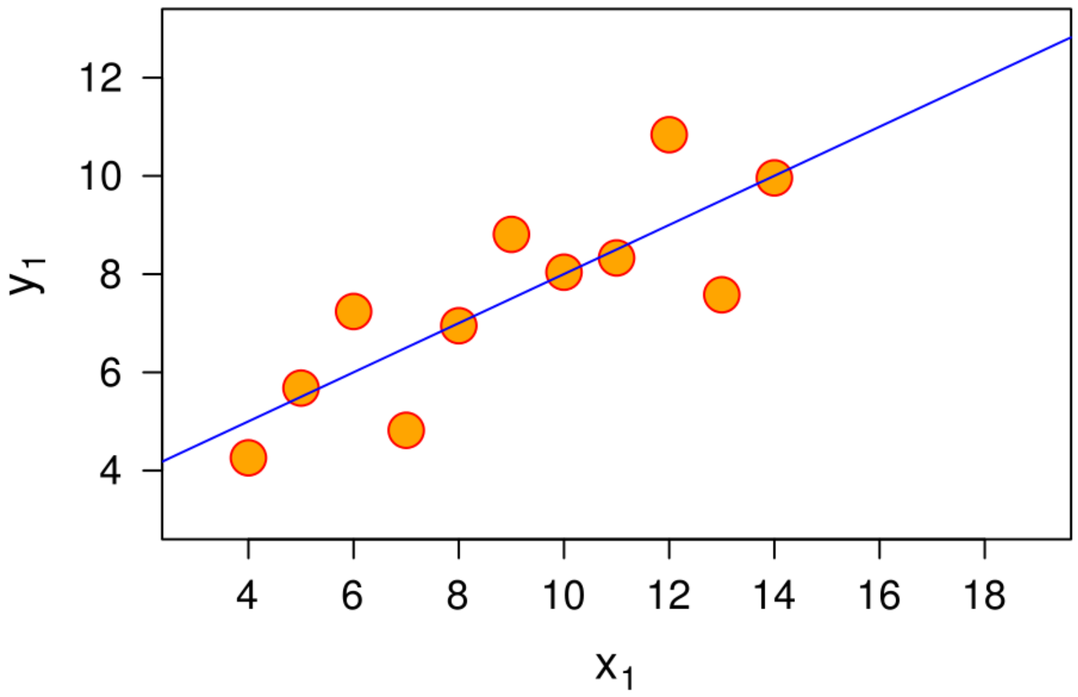


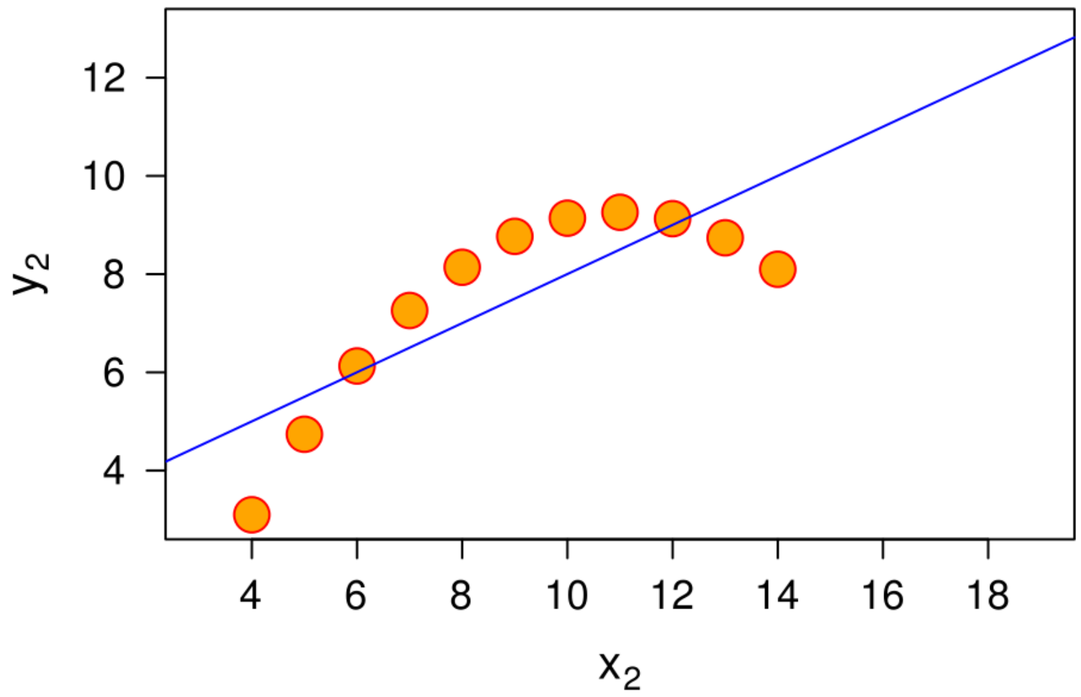
Graphs in Statistical Analysis*

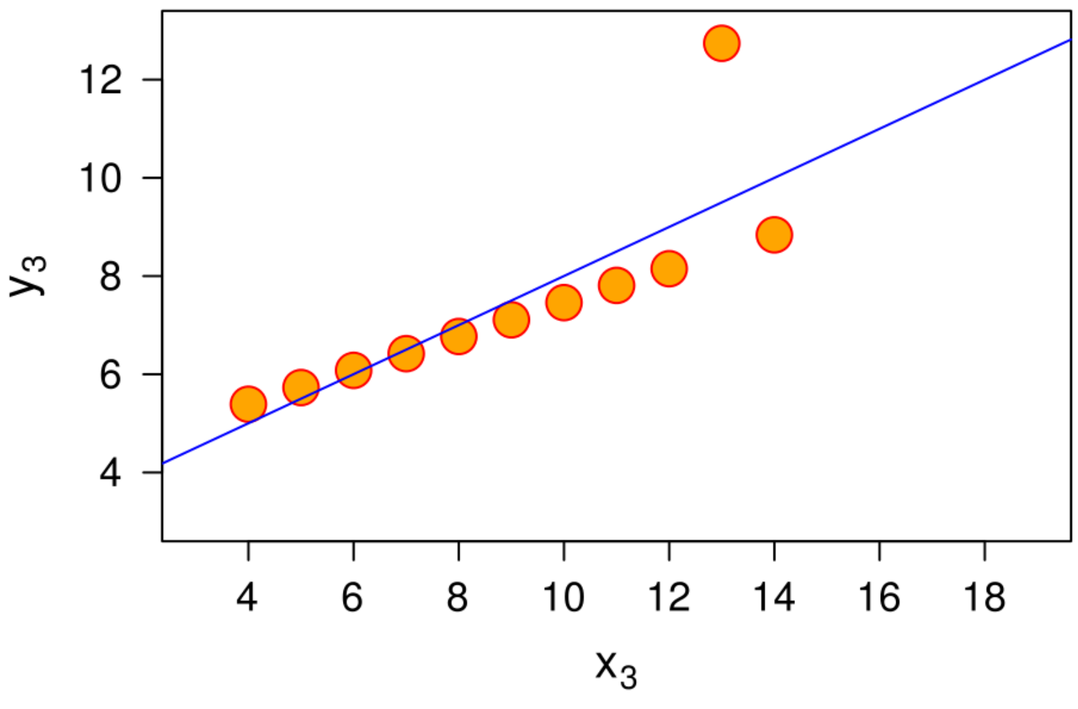
1973

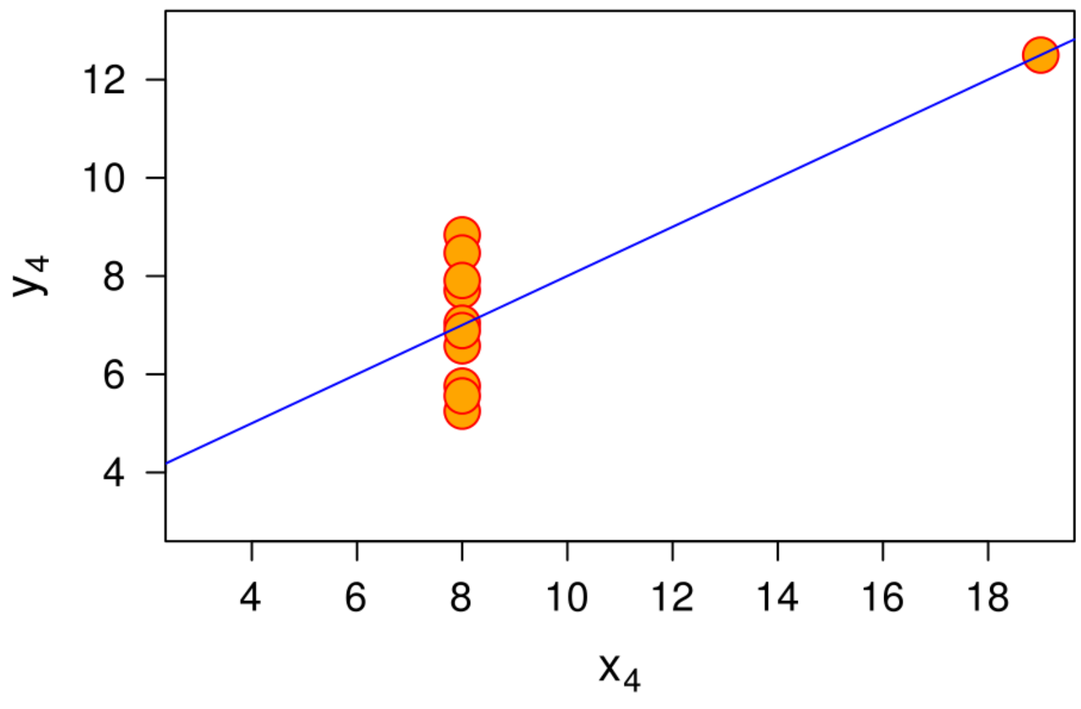
why

back down to earth: the anscombe quartet



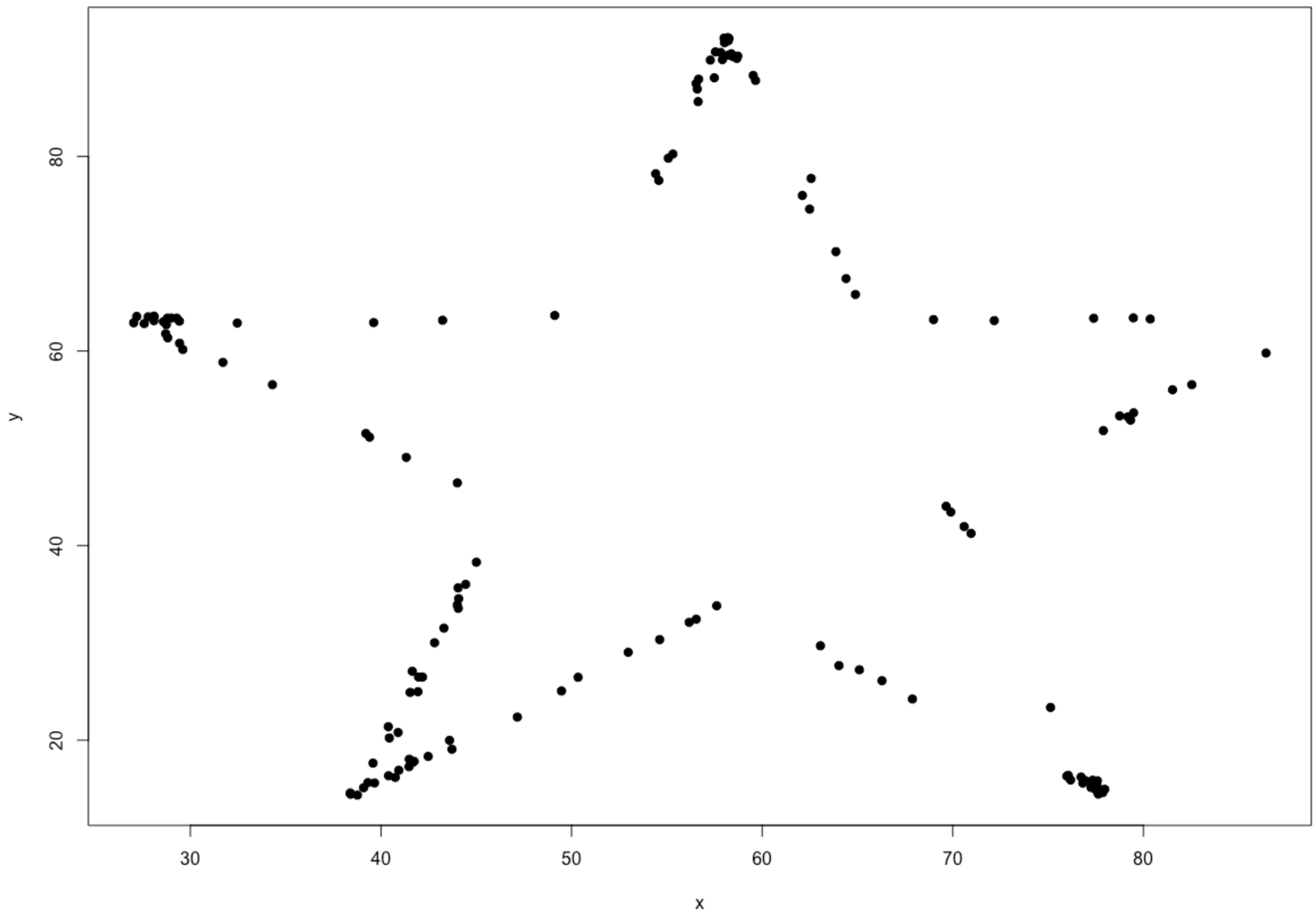






why

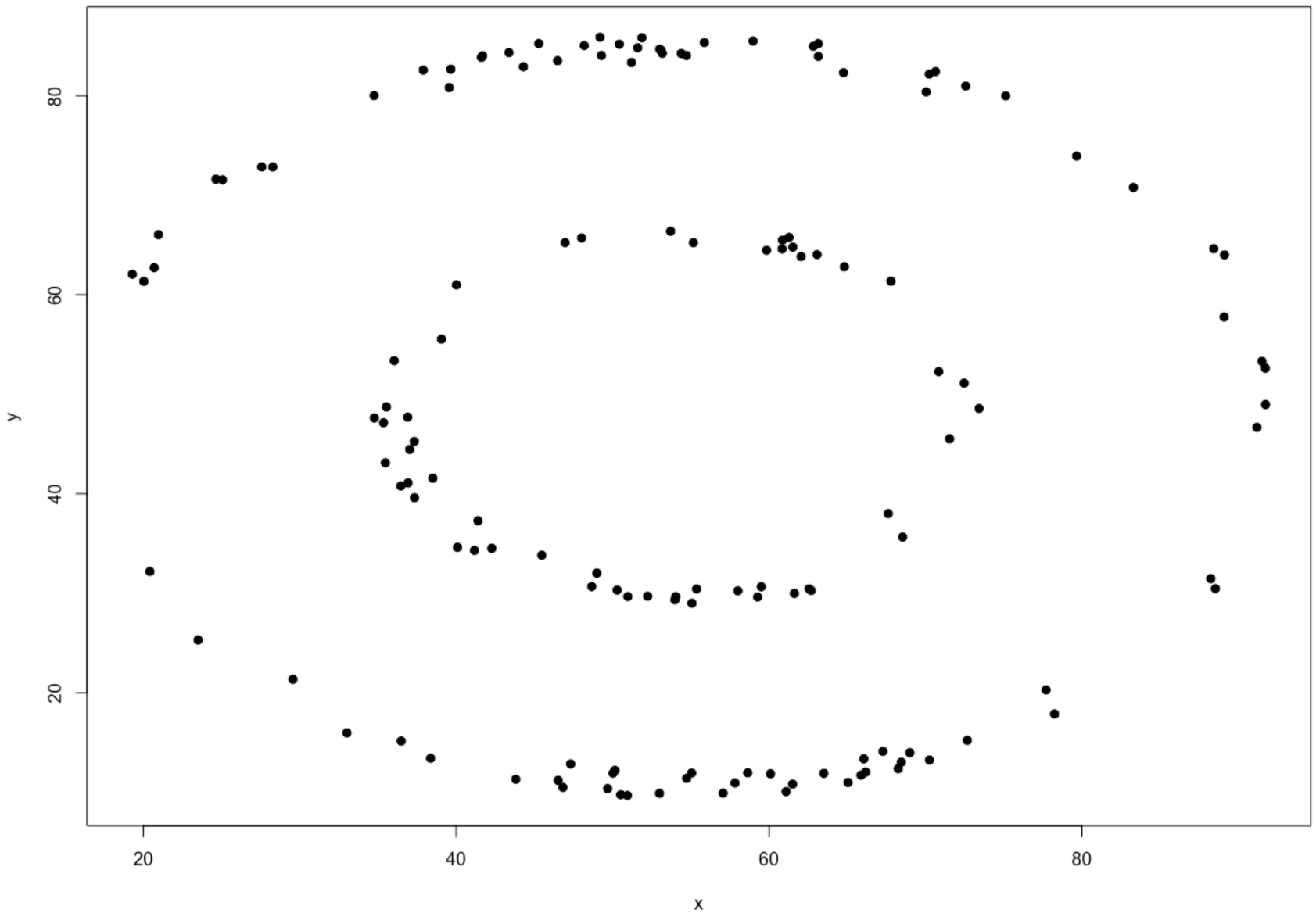
even harder: alberto cairo's datasaurus



w/ same mean/variance/correlation you can have…

why

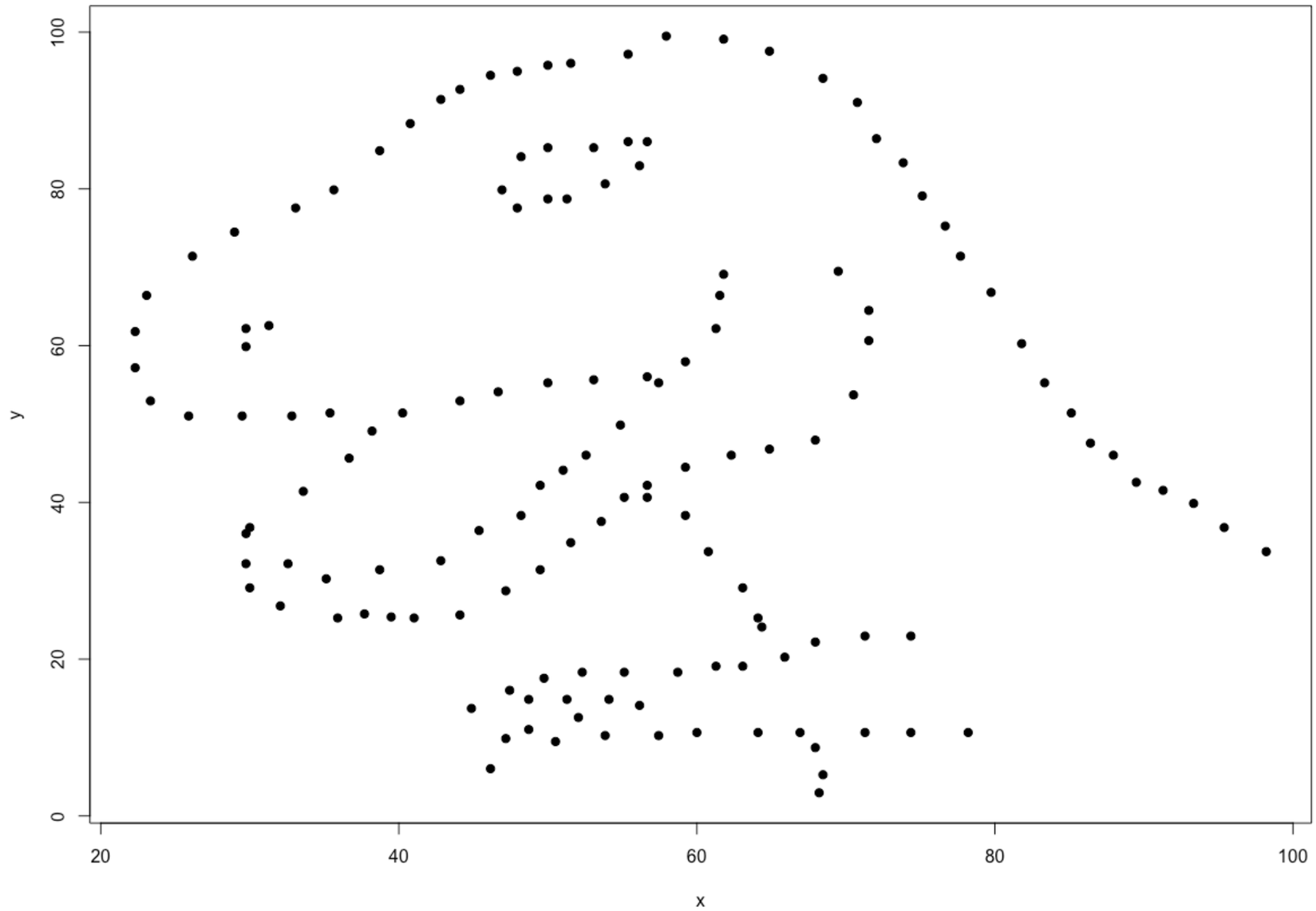
even harder: alberto cairo's datasaurus



and even more…

why

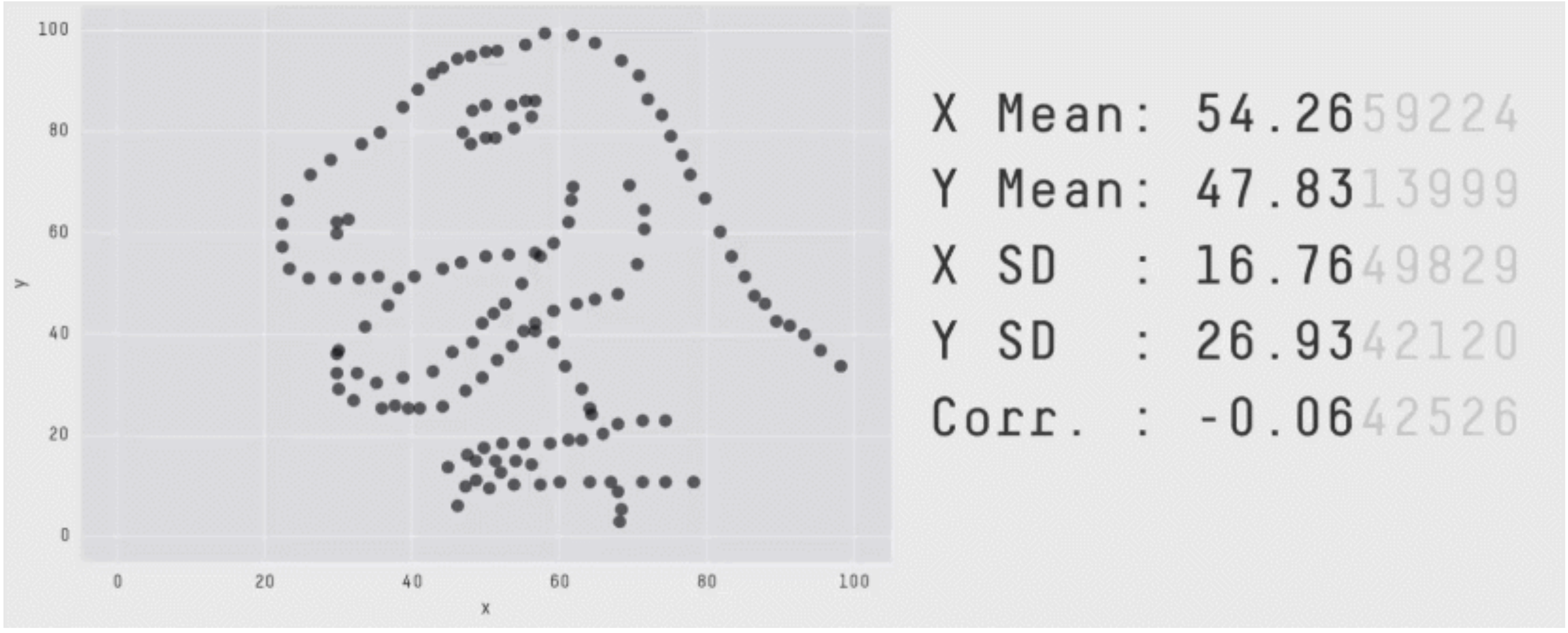
even harder: alberto cairo's datasaurus



actually, a wealth of nice shapes…

why

even harder: alberto cairo's datasaurus



never trust summary statistics alone;
always visualize your data
[a. cairo, 2017]

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

ACM SIGCHI Conference on Human Factors in Computing Systems

Honourable Mention

DOWNLOAD PDF

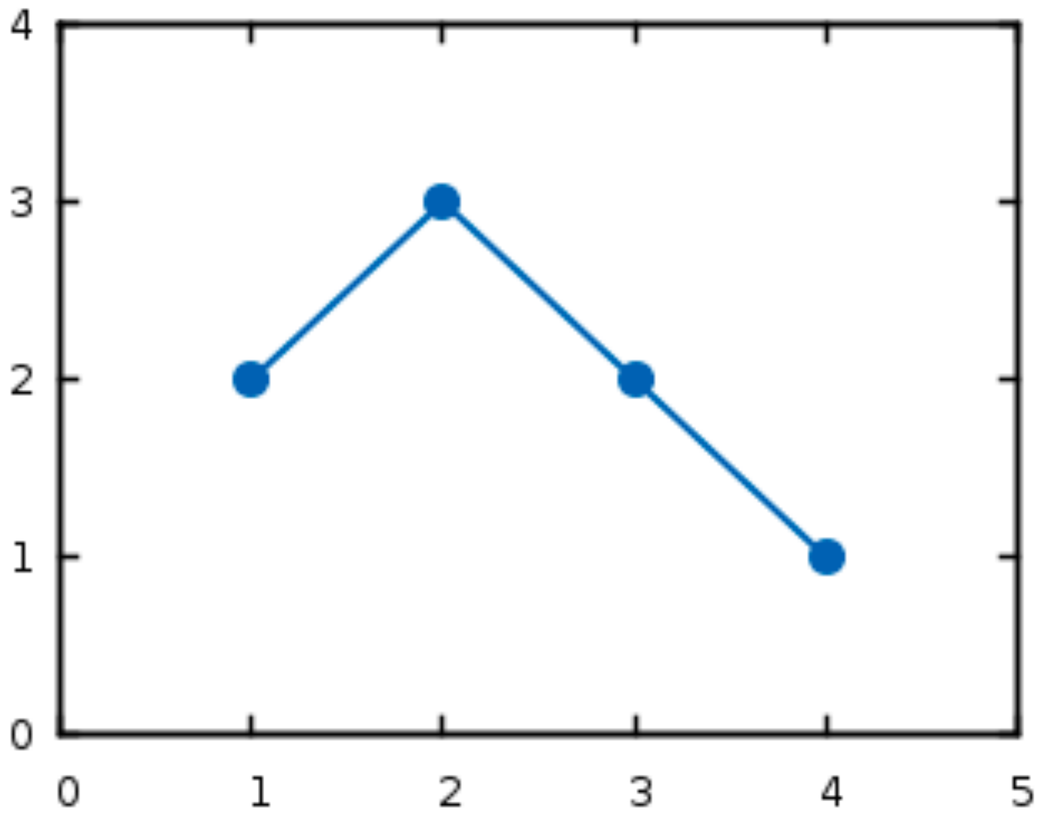
AUTHORS
Justin Matejka, George Fitzmaurice

2017

why

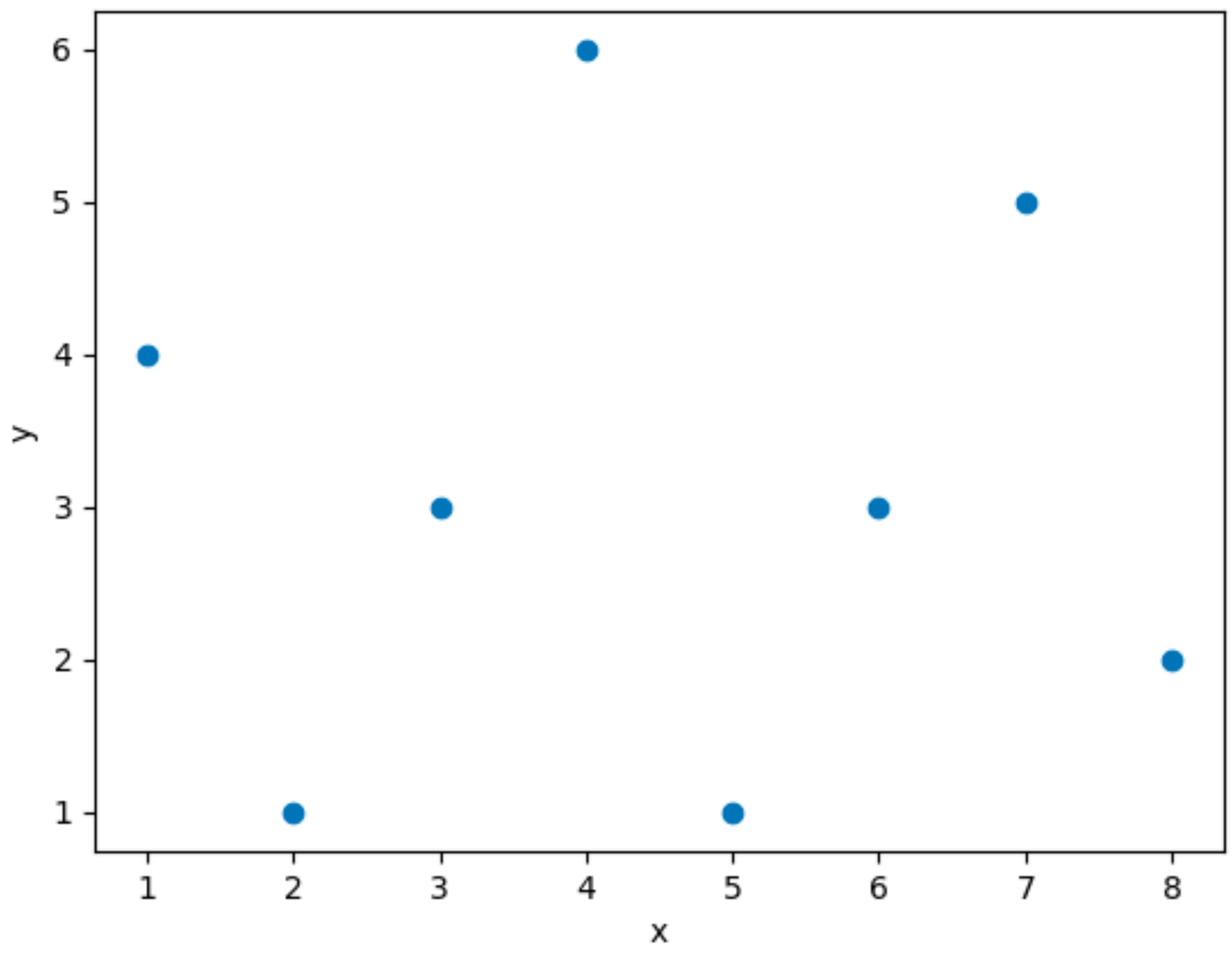
the goal

understanding and creating graphical representation of data



lineplot

scatterplot

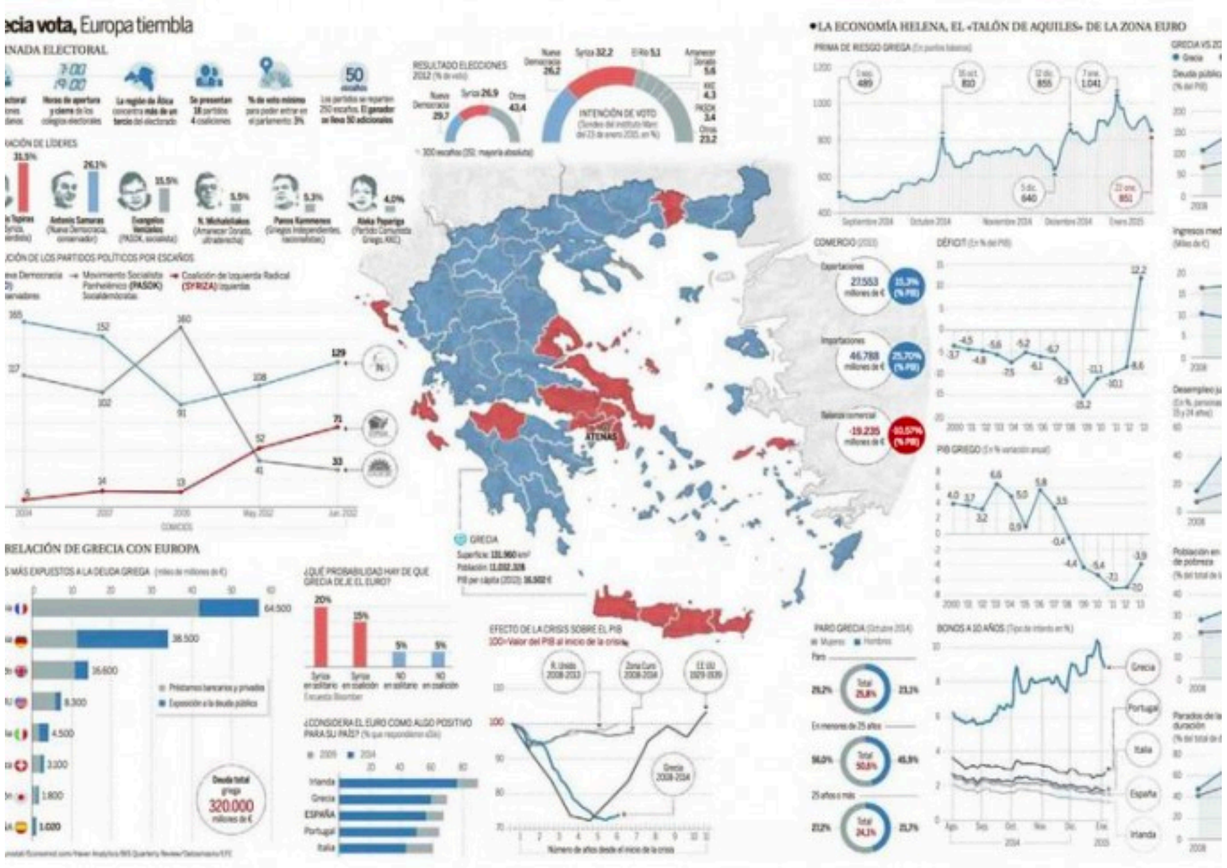


… from very simple plots …

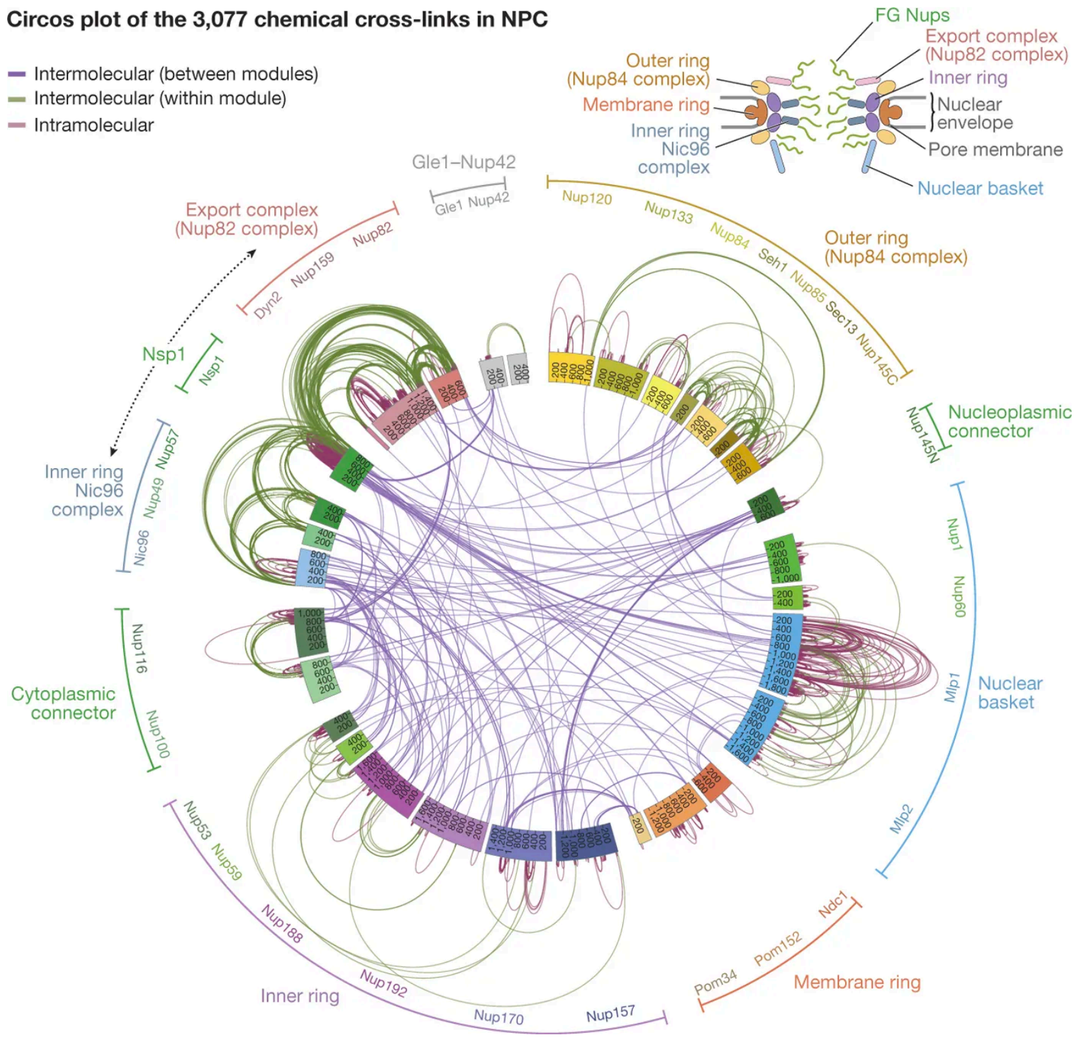
why

the goal

understanding and creating graphical representation of data



data journalism



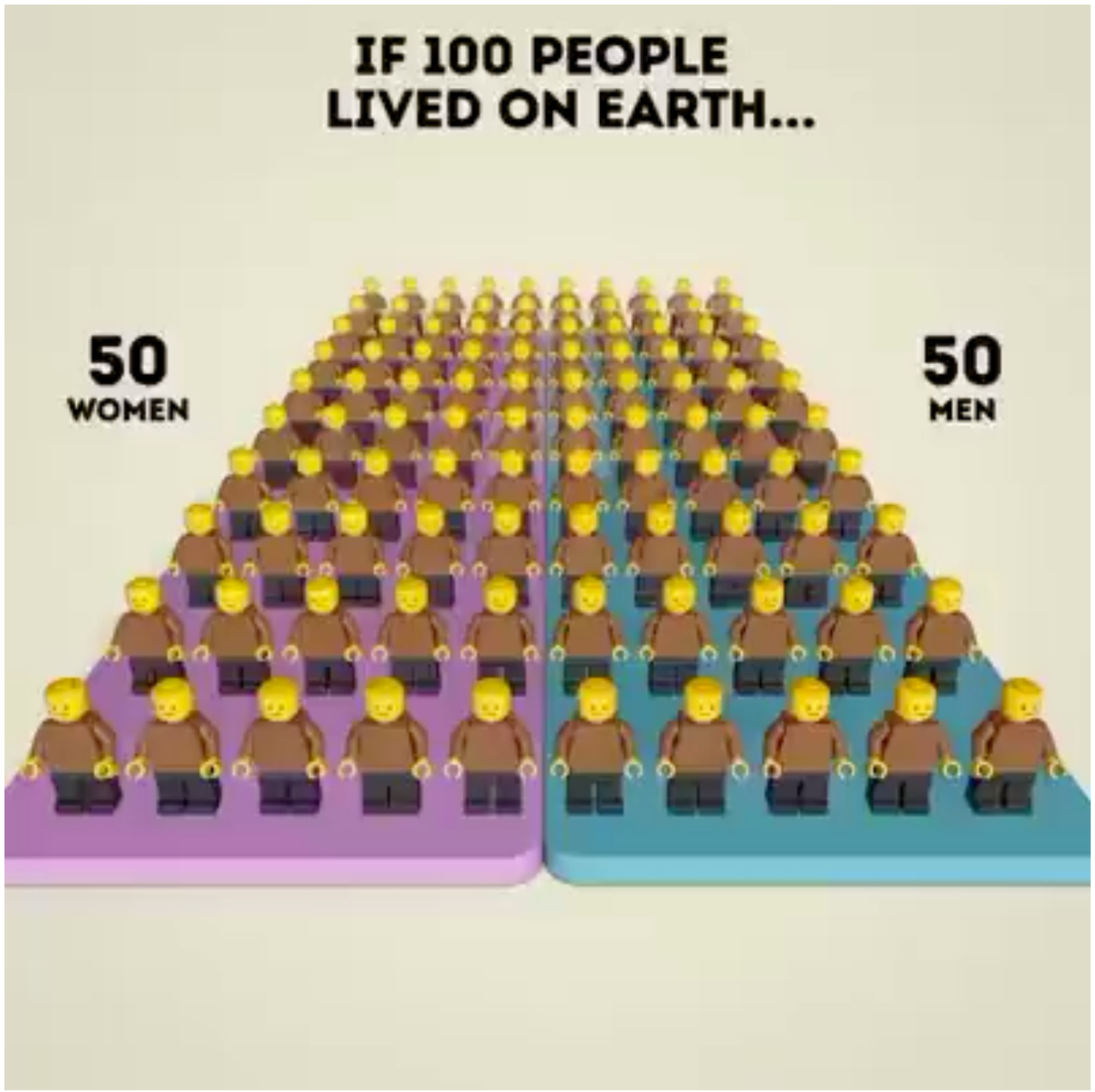
circos

… to complex panels …

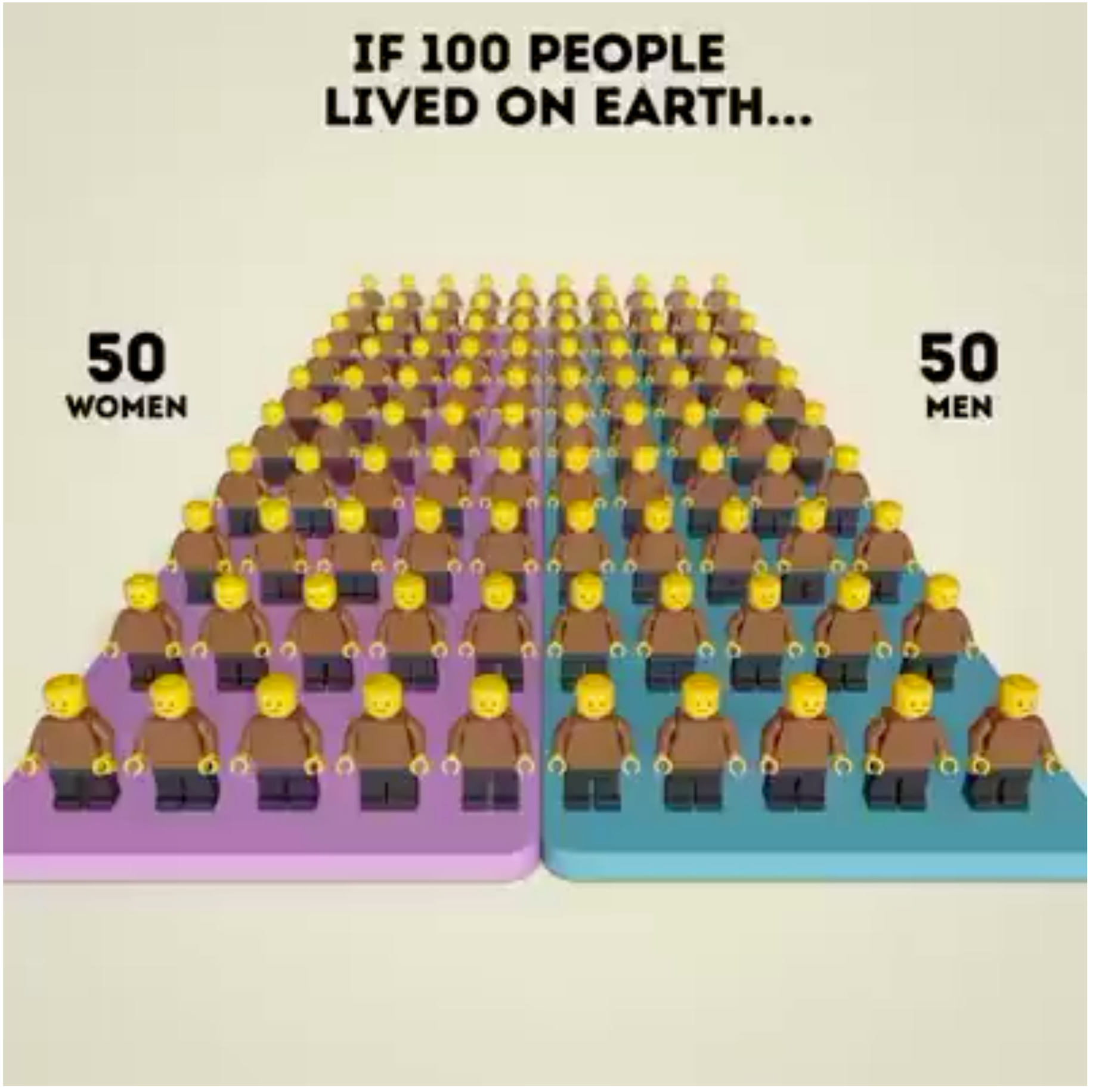
why

the goal

understanding and creating graphical representation of data



… telling stories …


IF 100 PEOPLE
LIVED ON EARTH...
50
WOMEN
50
MEN

why

the goal

understanding and creating graphical representation of data

- 50% are men, 50% are women
- 60% asian people, 15% africans, 14% americans, 11% europeans
- 26% 0-14 years, 16% 15-24 years, 41% 25-54 years, 9% 55-64 years, 8% 65+ years
- 33% Christianity, 21% Islam, 16% non-religious, 14% Hinduism, 10% other religions, 6% Buddhism
- 12% Chinese, 6% Spanish, 5% English, 4% Hindi, 3% Arabic, the rest 6500 languages
- 86% can read and write, 14% can’t
- …

… telling stories …

why

the goal

understanding and creating graphical representation of data

… raising issues …

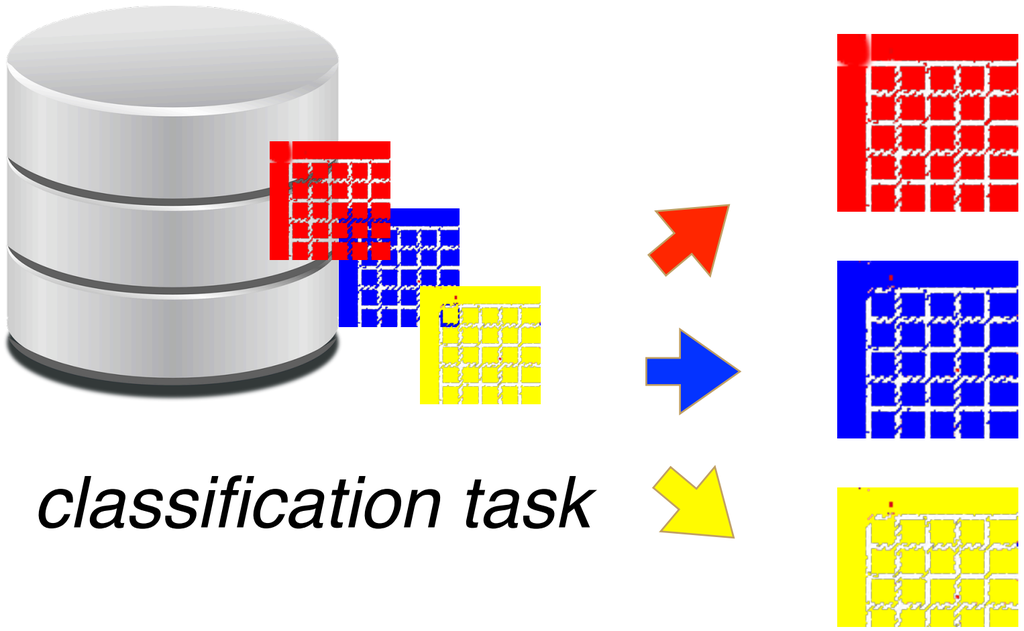
why

the goal

“a picture is worth a thousand words”

… if you know how to read it!

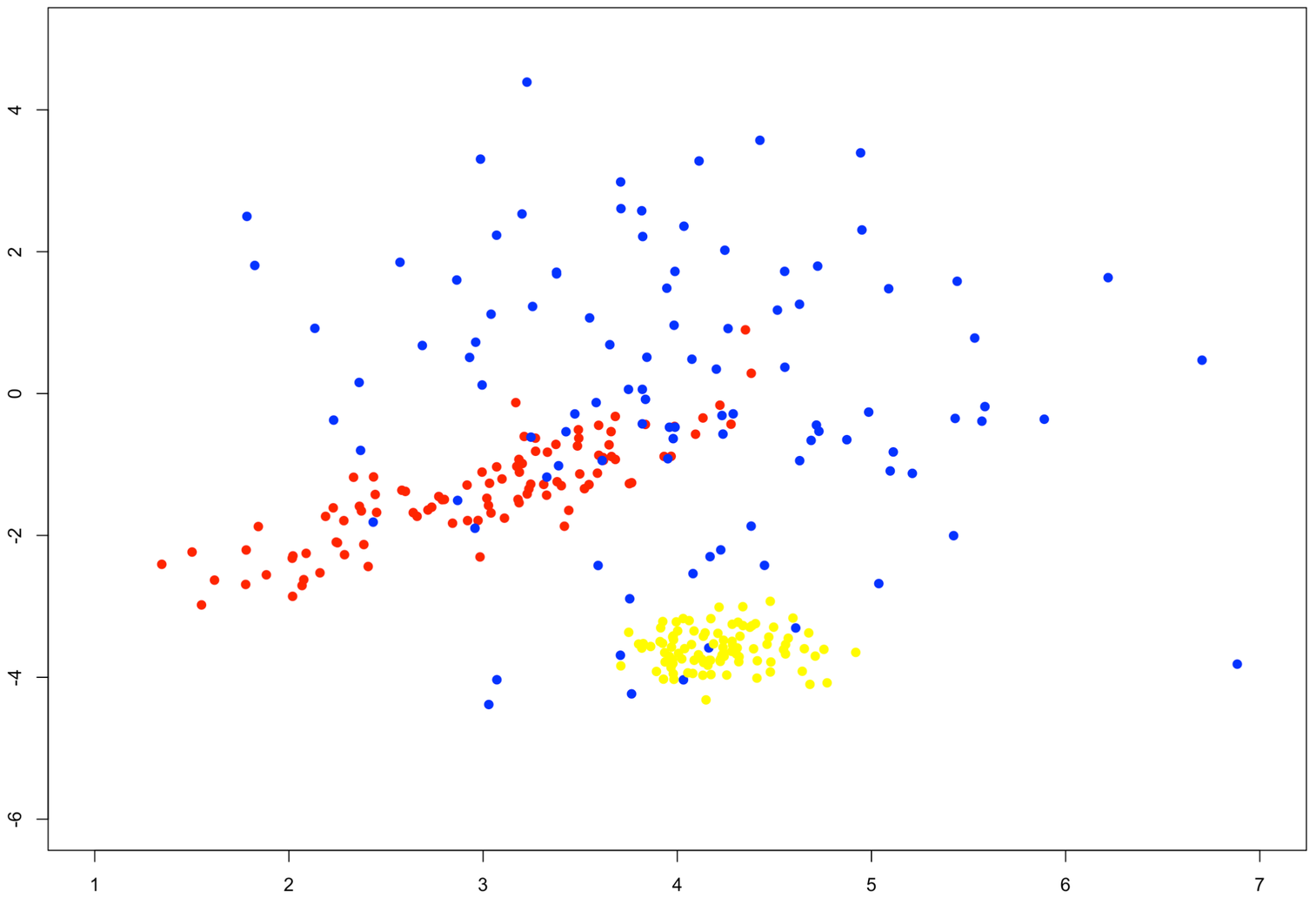
[a. cairo, 2019]



classification task

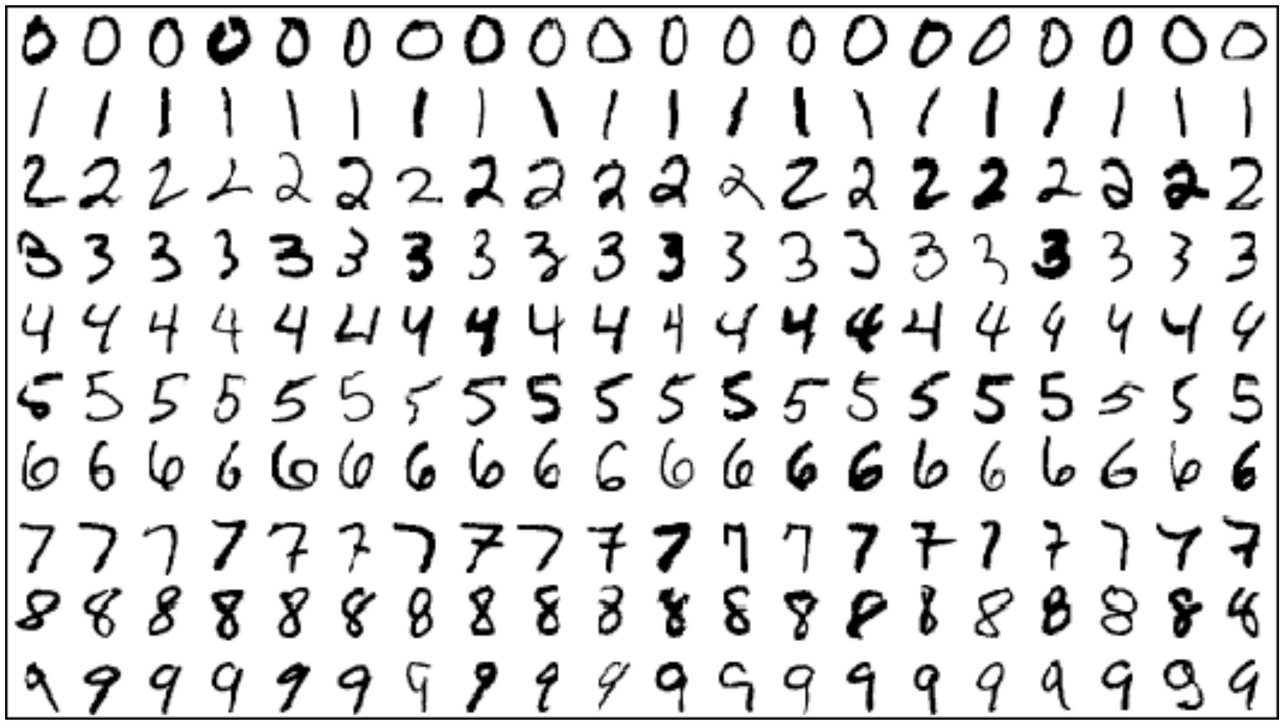
why

easy or hard?



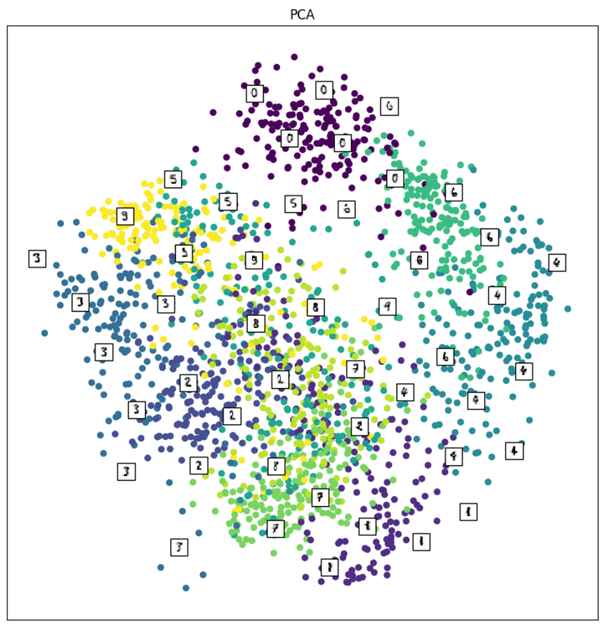
why

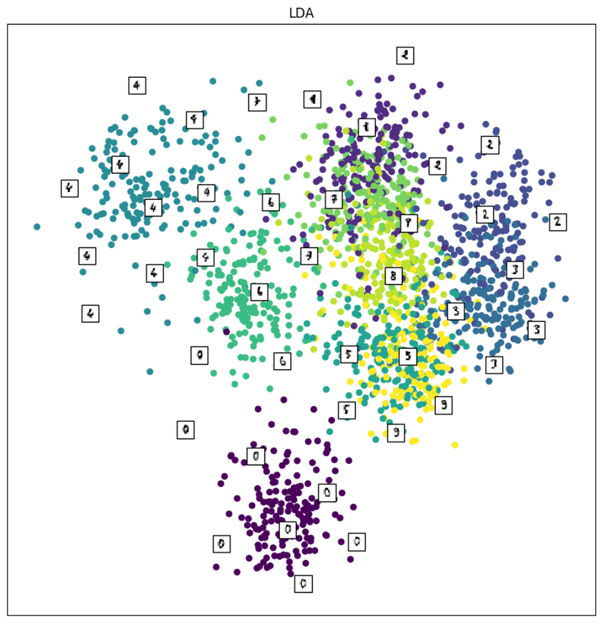
a wealth of choices…

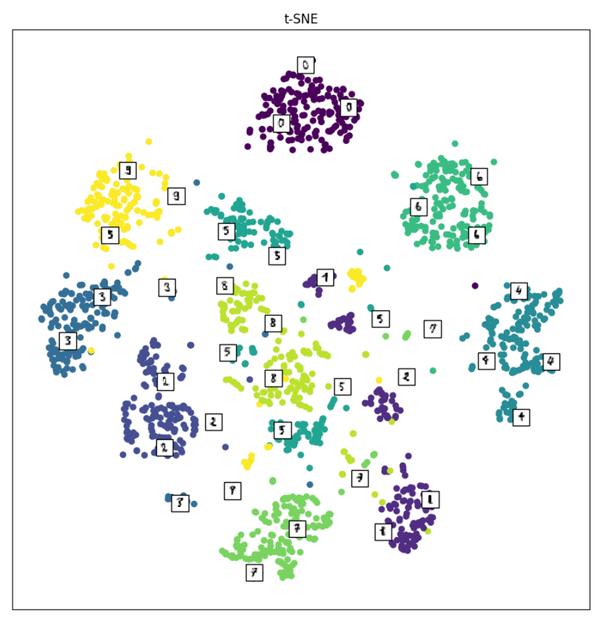


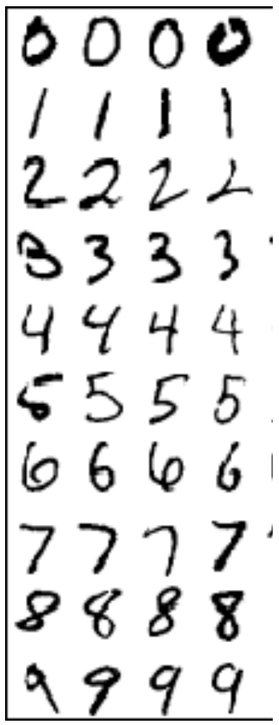
mnist dataset
[LeCun et al., 2013]

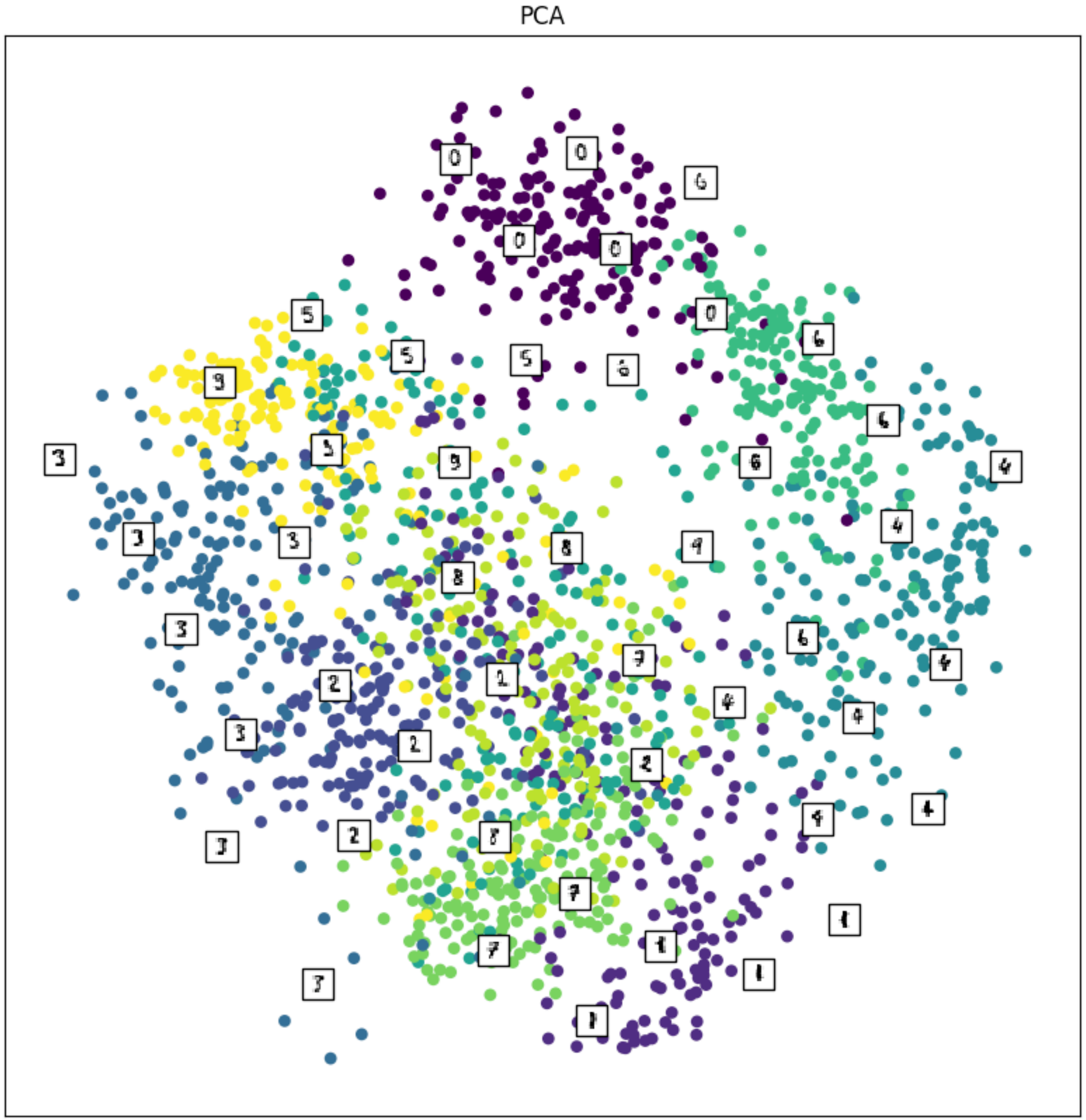
60K+10K images





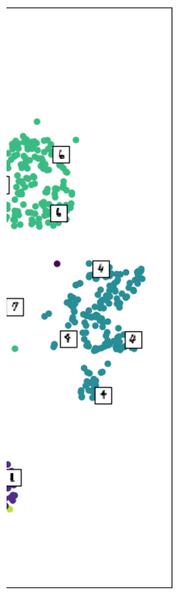






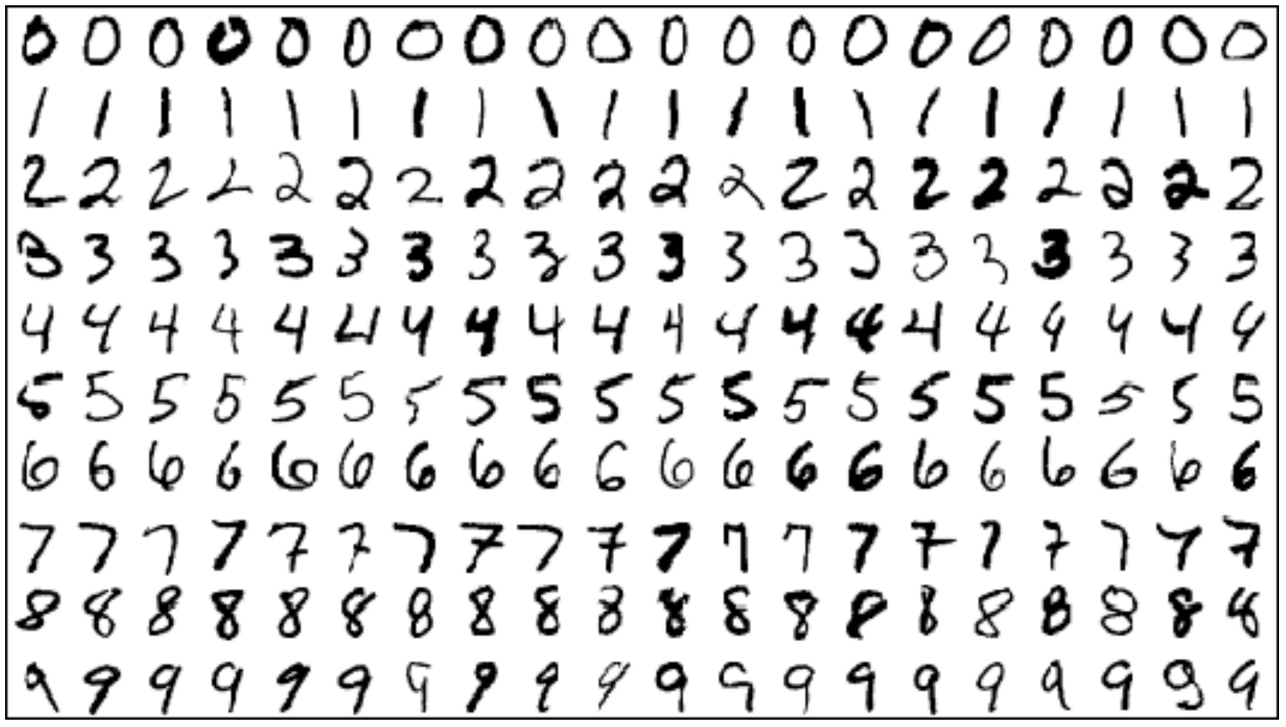
why

ces…



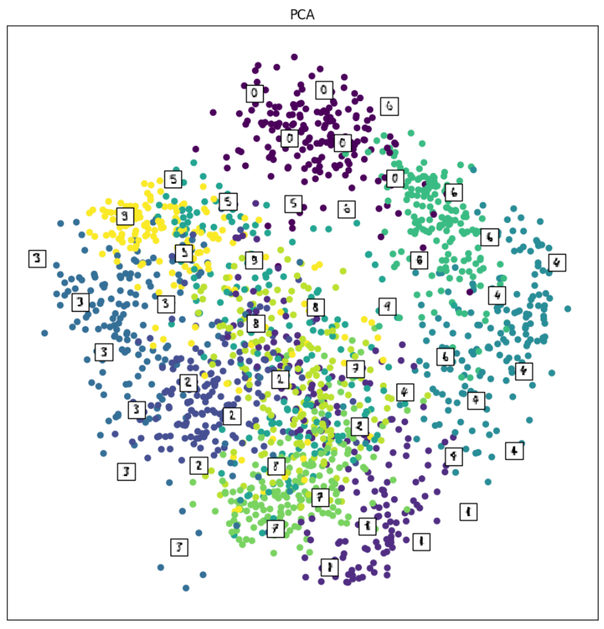
why

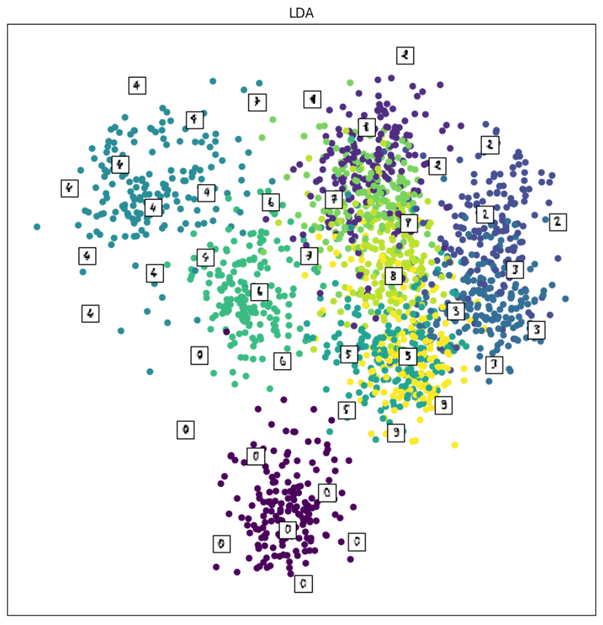
a wealth of choices…

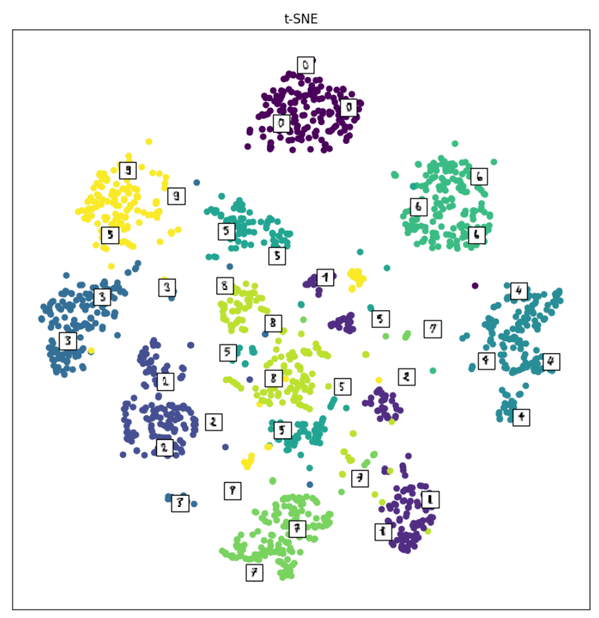


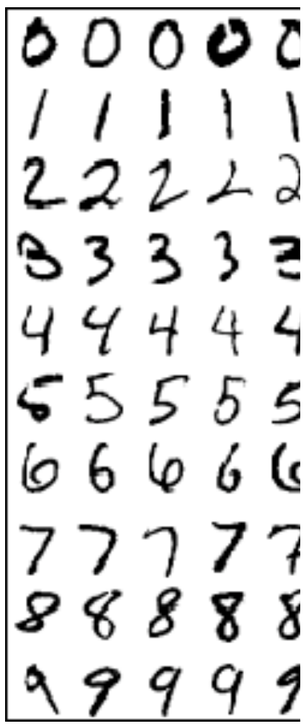
mnist dataset
[LeCun et al., 2013]

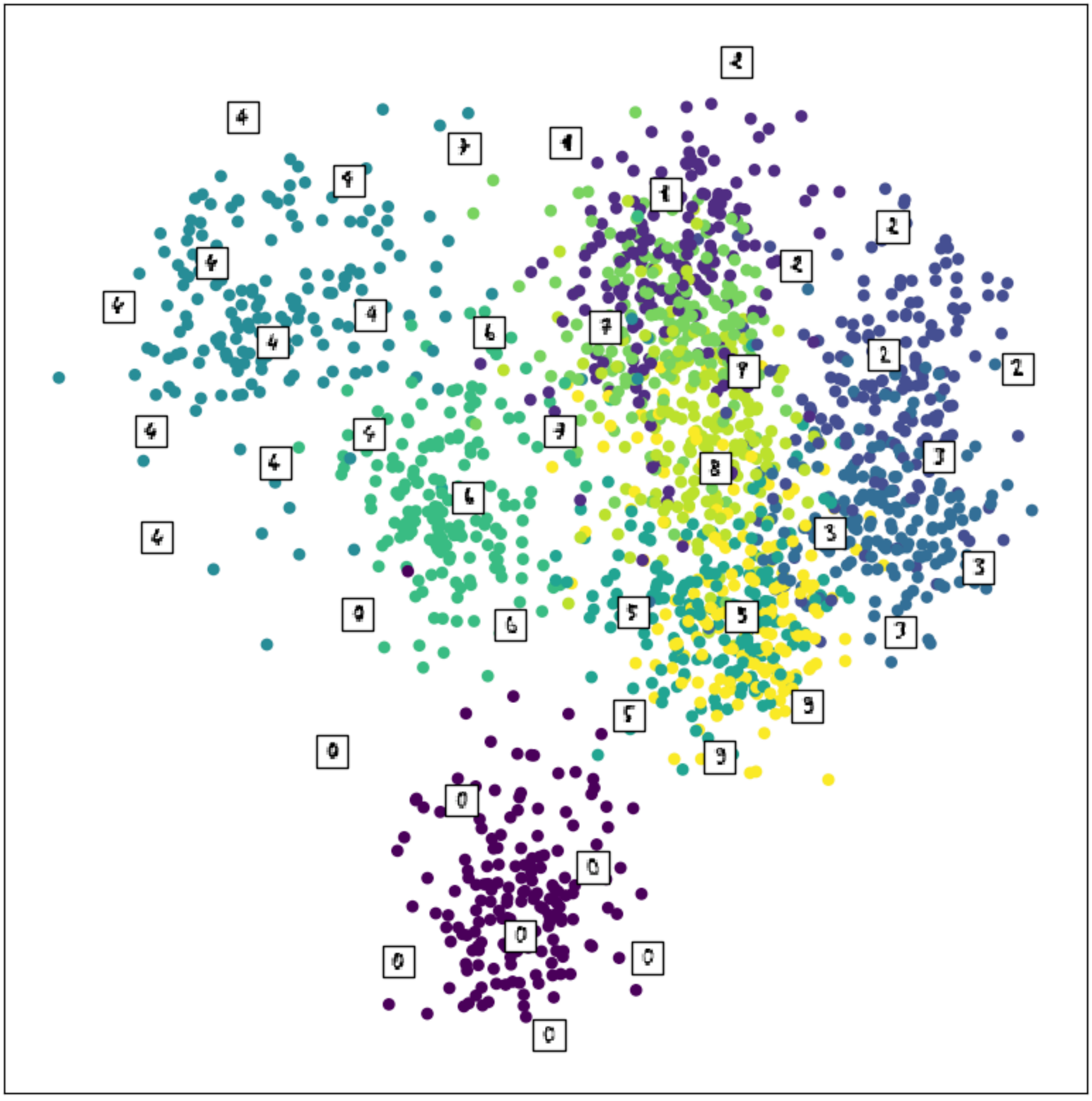
60K+10K images

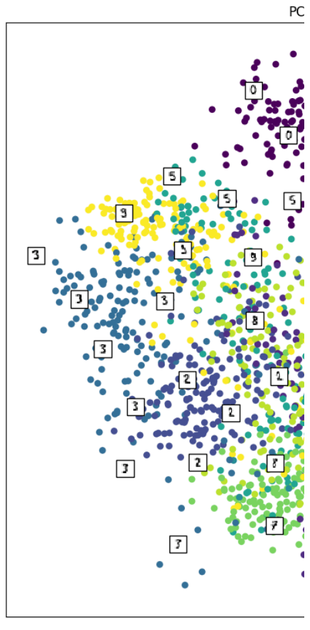


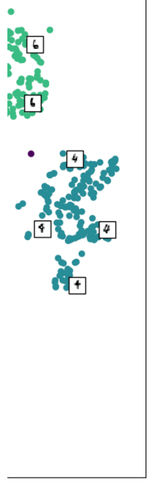










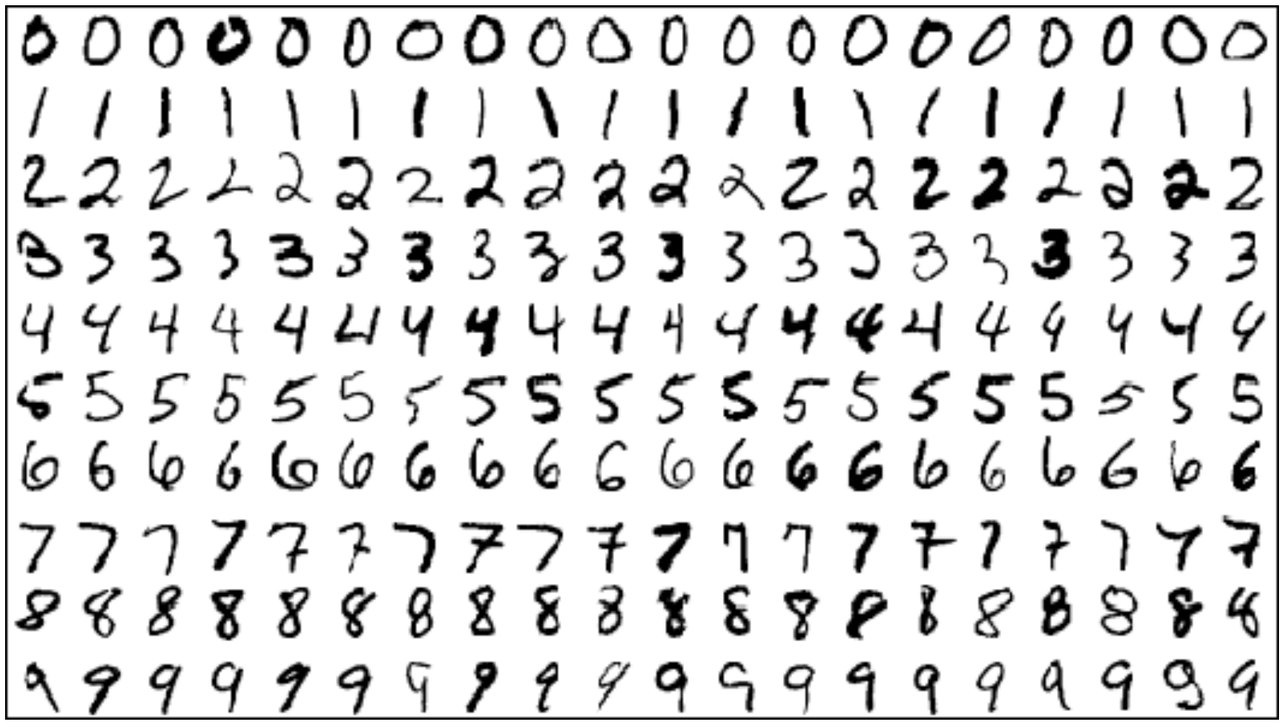


why

es…

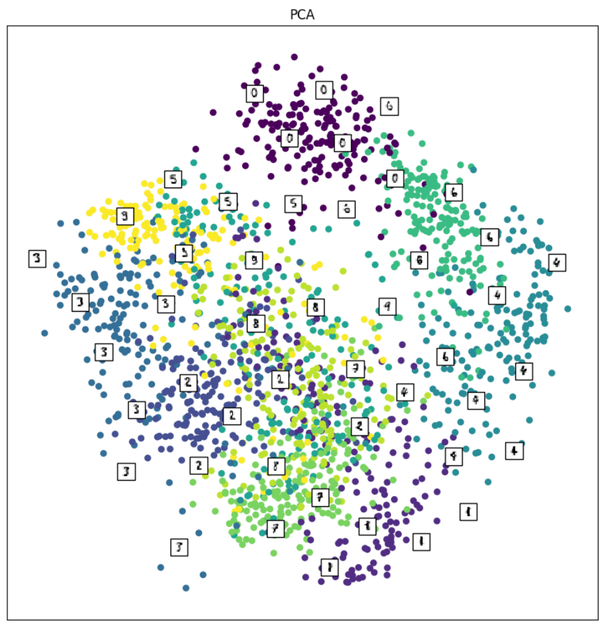
why

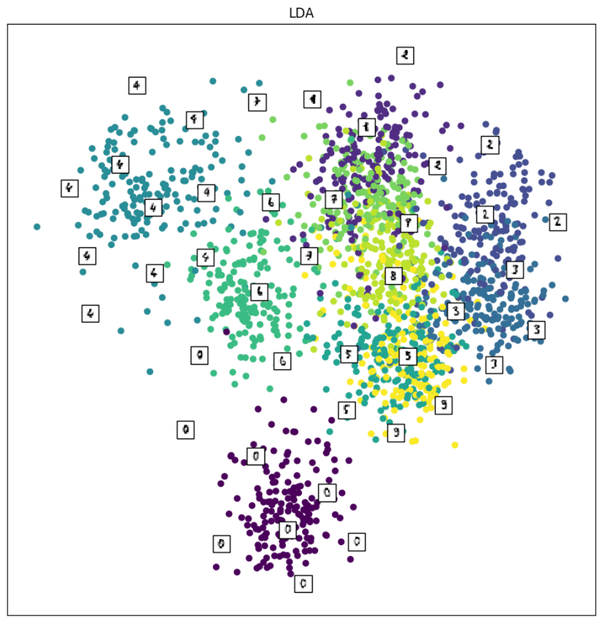
a wealth of choices…

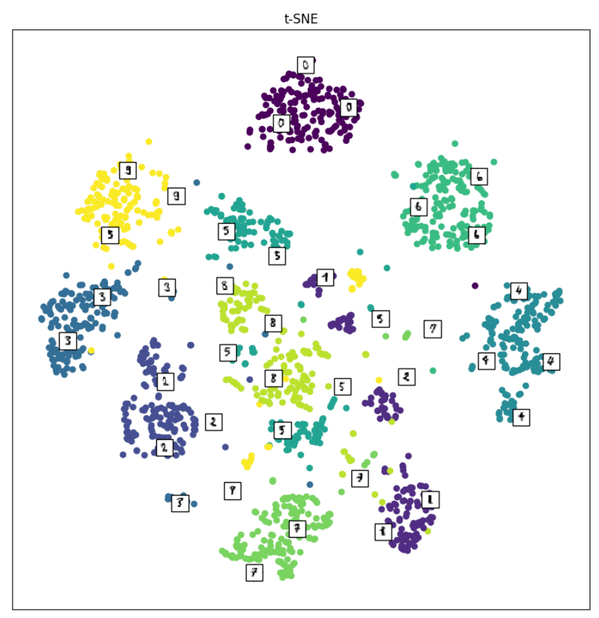


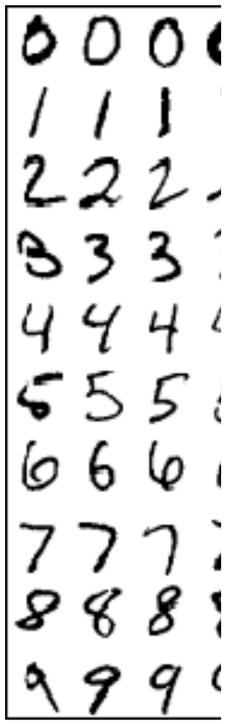
mnist dataset
[LeCun et al., 2013]

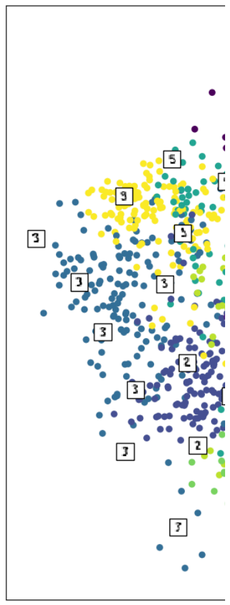
60K+10K images

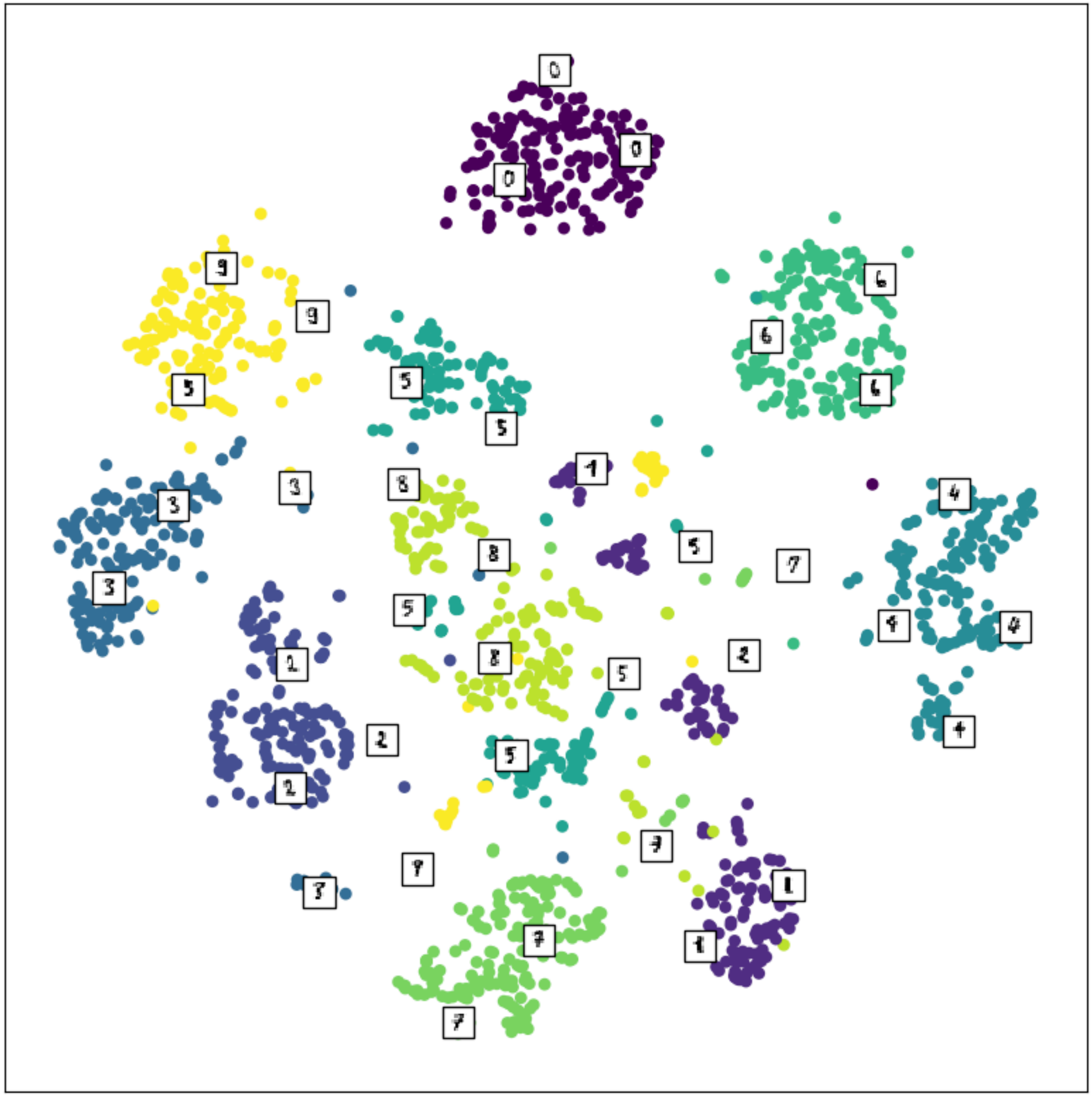










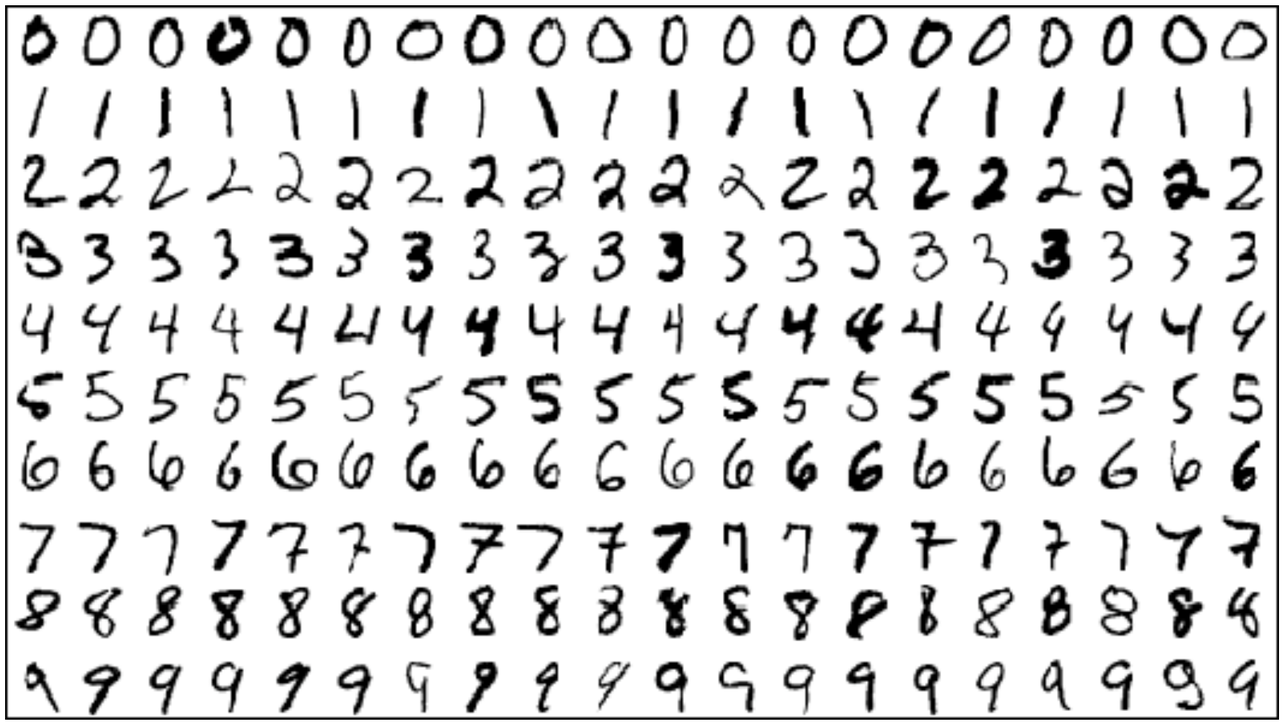


why

ices…

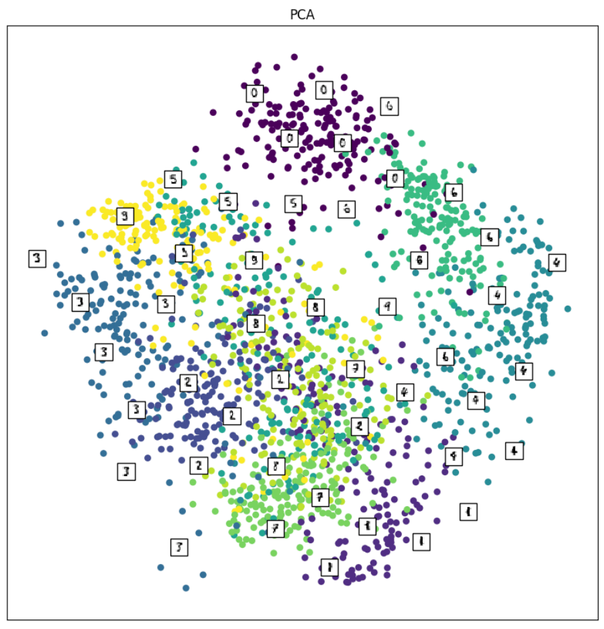
why

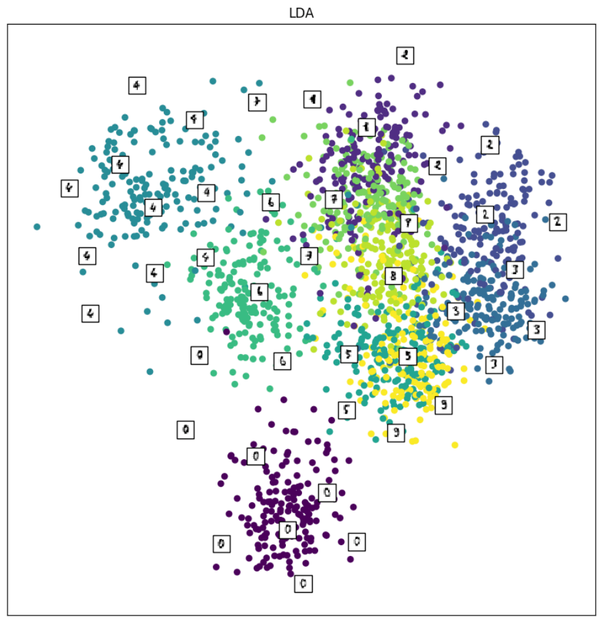
a wealth of choices…

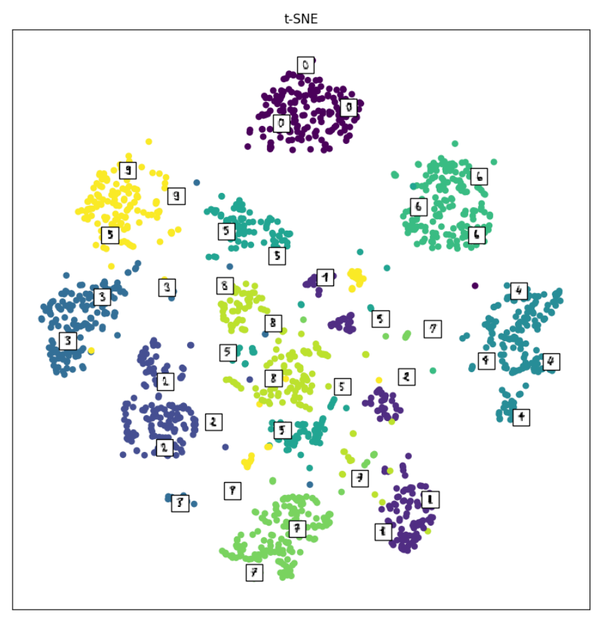


mnist dataset
[LeCun et al., 2013]

60K+10K images

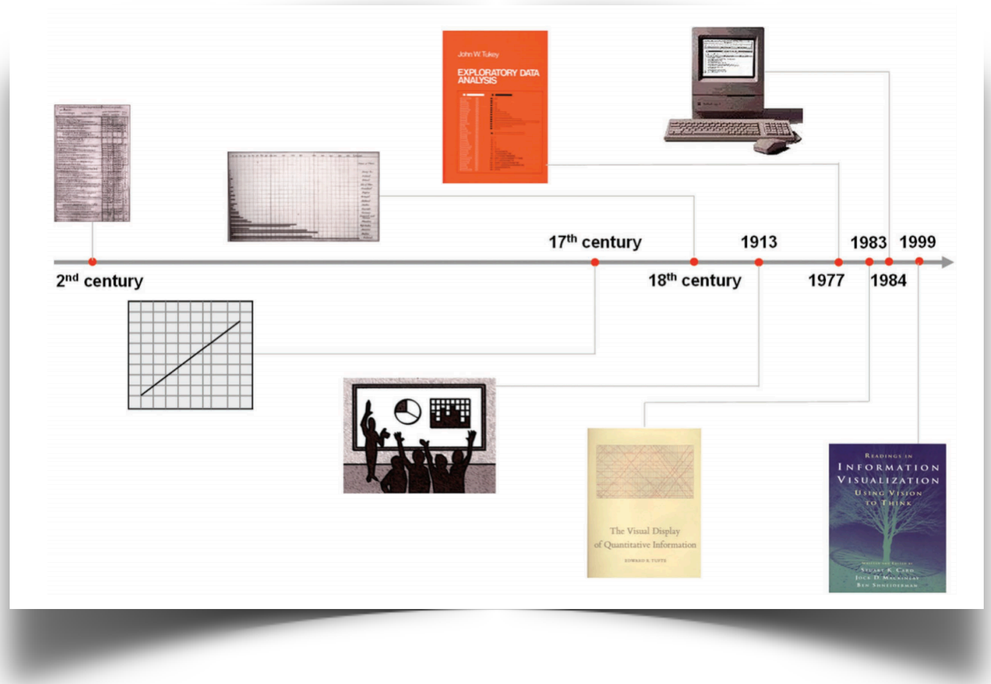


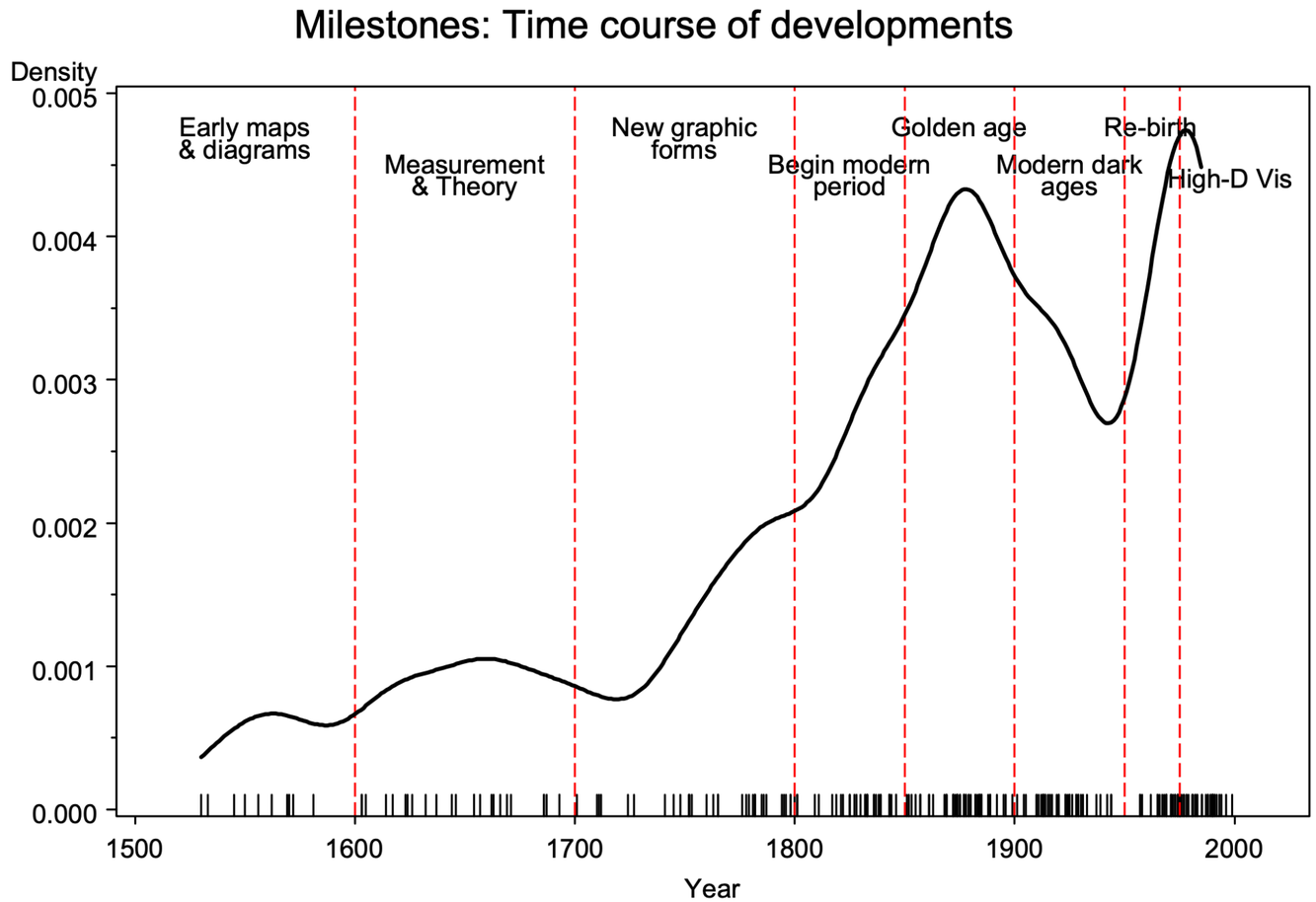




outline

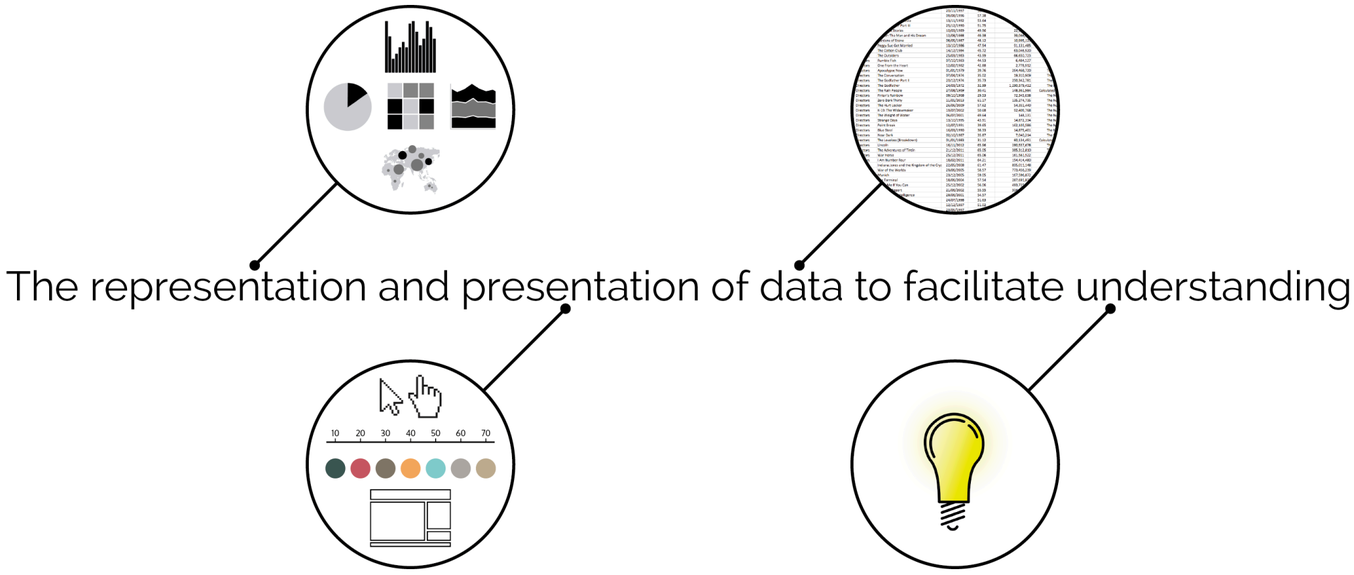
history



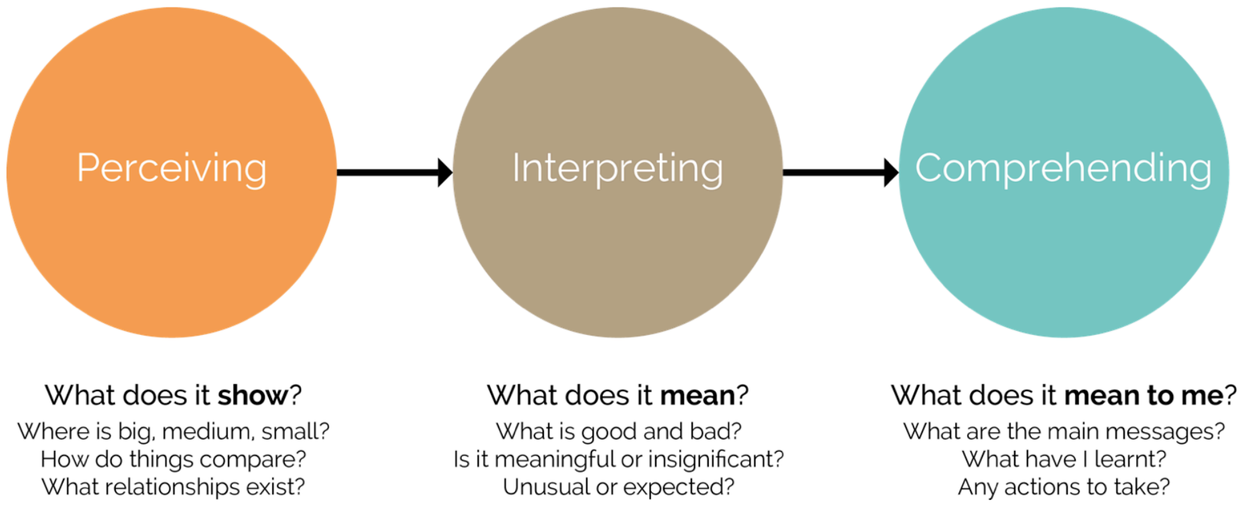


outline

foundations



The representation and presentation of data to facilitate understanding



What does it **show**?
Where is big, medium, small?
How do things compare?
What relationships exist?

What does it **mean**?
What is good and bad?
Is it meaningful or insignificant?
Unusual or expected?

What does it **mean to me**?
What are the main messages?
What have I learnt?
Any actions to take?

Principle 1

Good data visualisation is **TRUSTWORTHY**

Principle 2

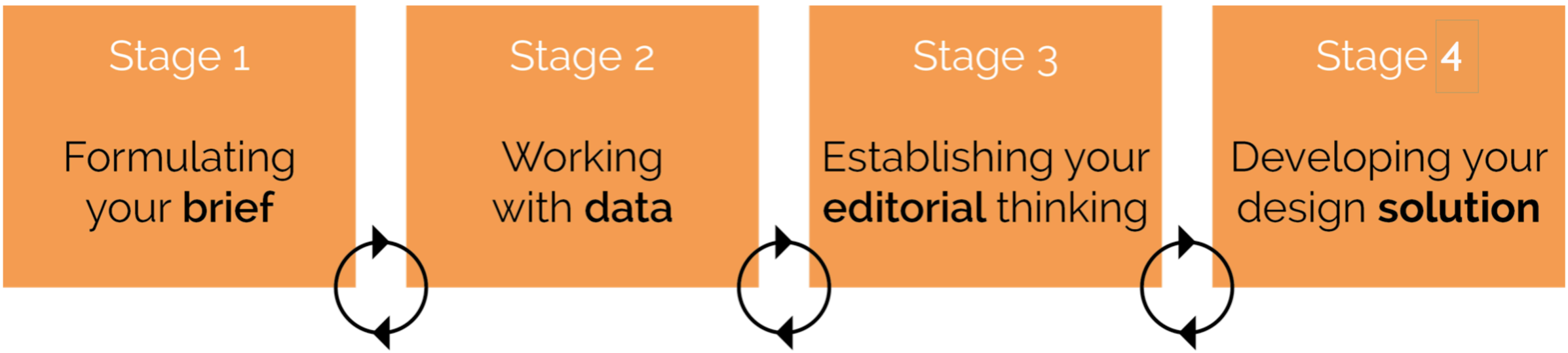
Good data visualisation is **ACCESSIBLE**

Principle 3

Good data visualisation is **ELEGANT**

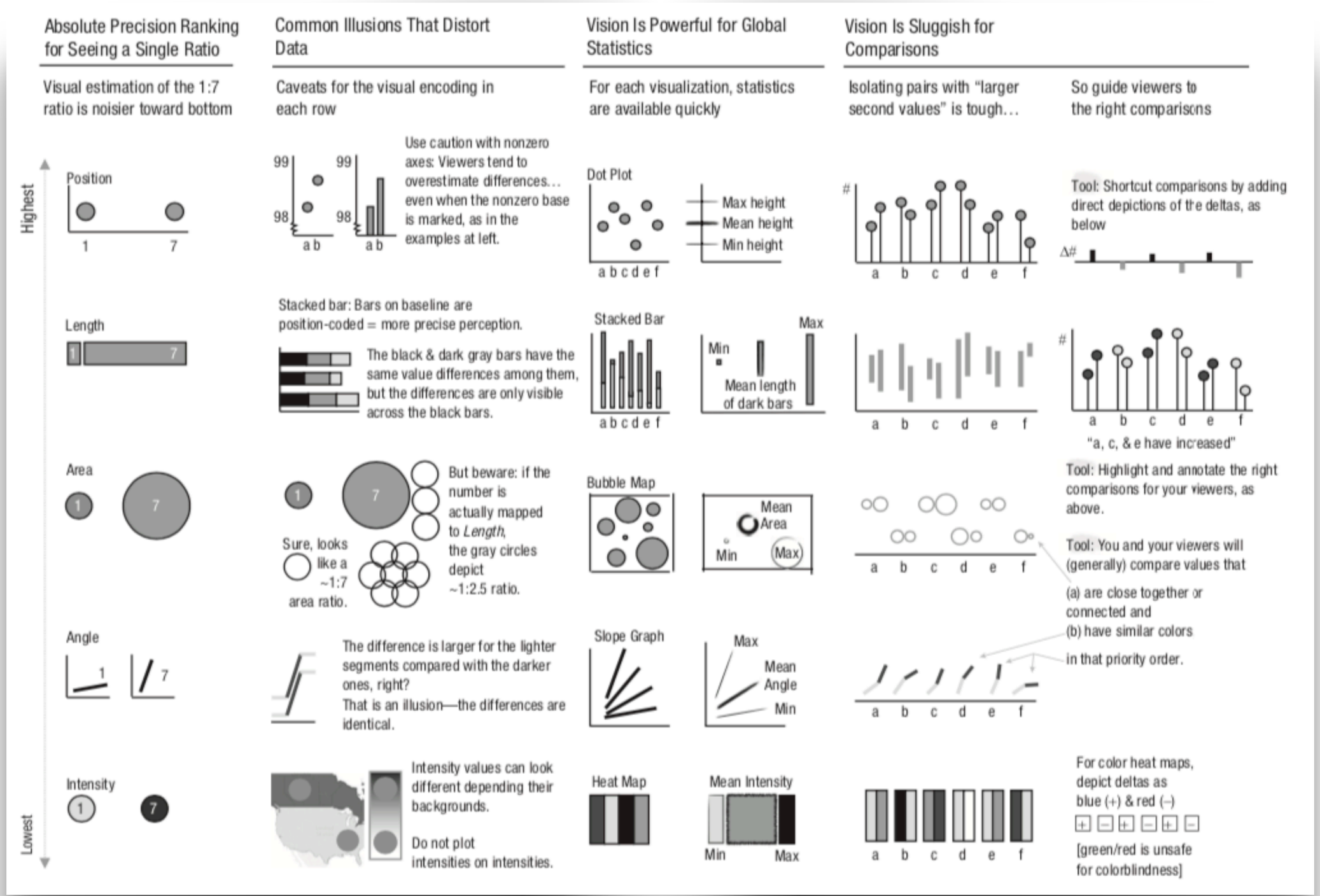
outline

workflow



outline

representation



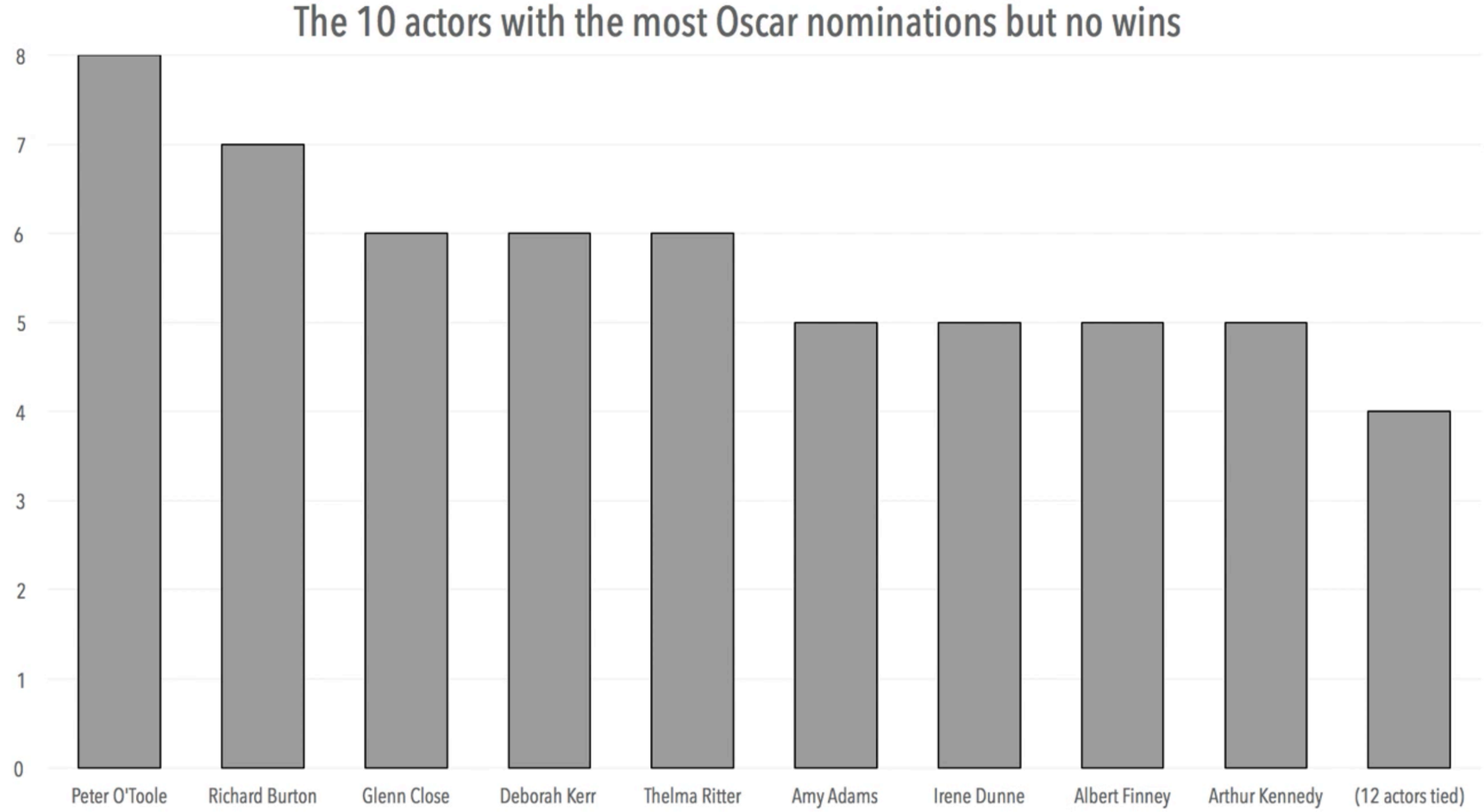
outline

zoo

chrts

bar chart

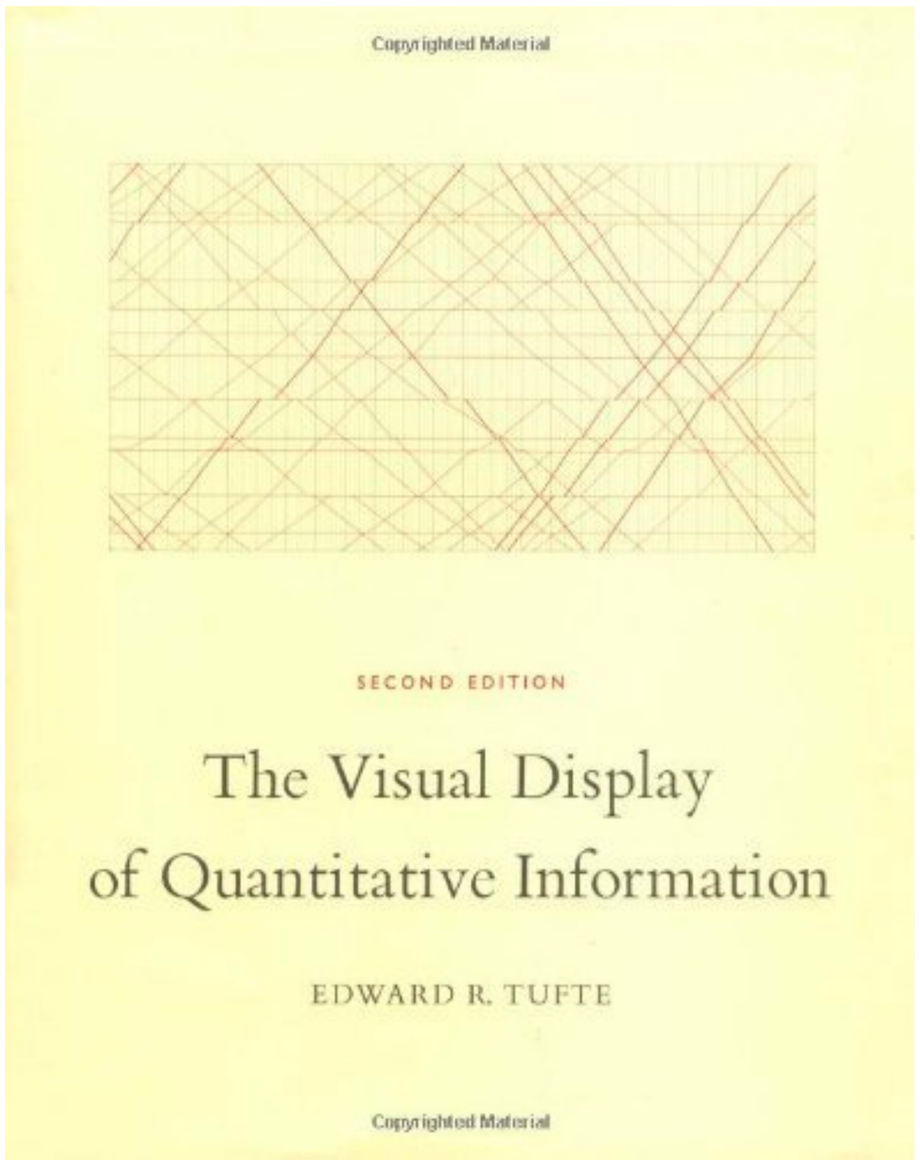
column chart — histogram



Source: https://en.wikipedia.org/wiki/List_of_actors_with_two_or_more_Academy_Award_nominations_in_acting_categories (as at March 2016)

outline

good practice

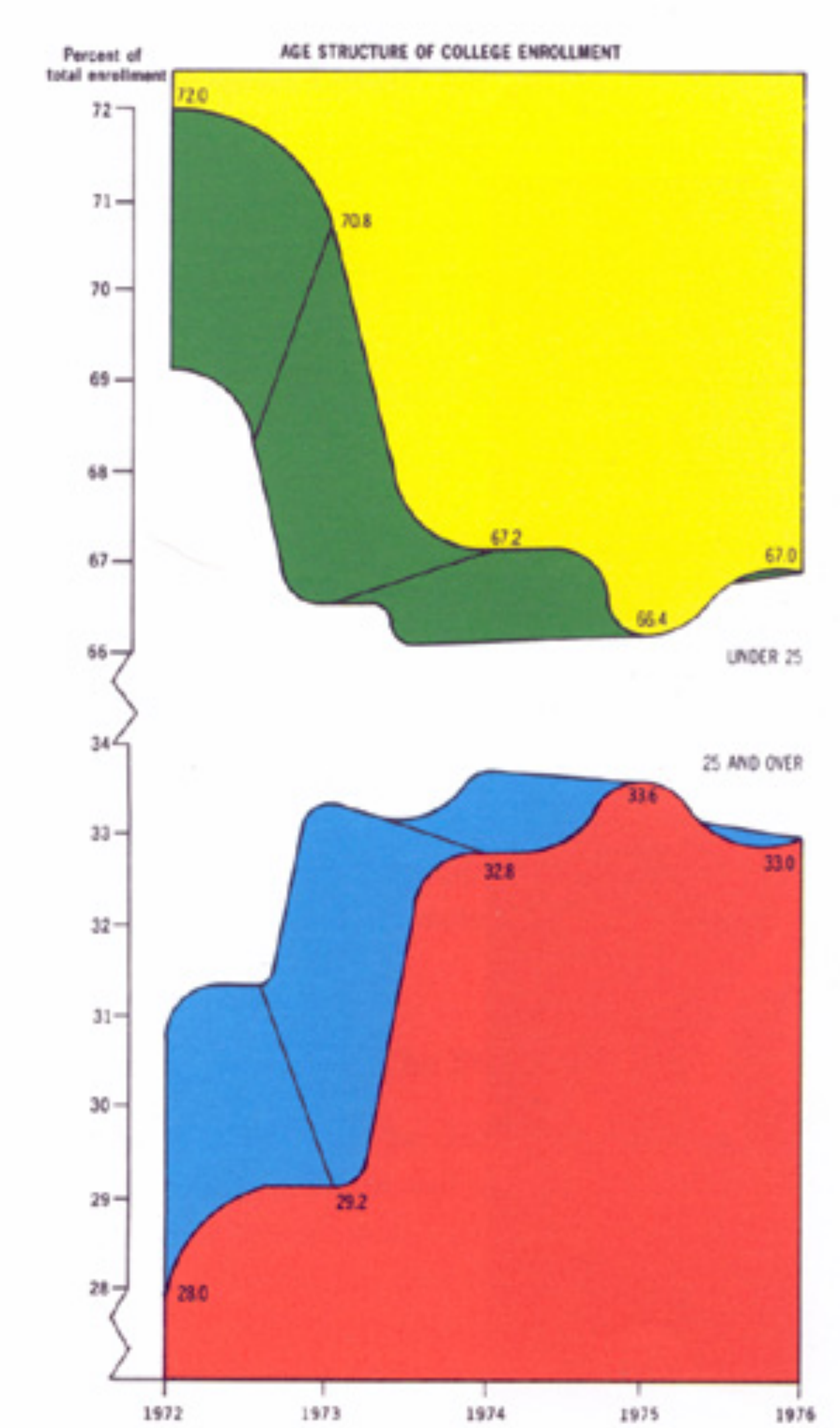


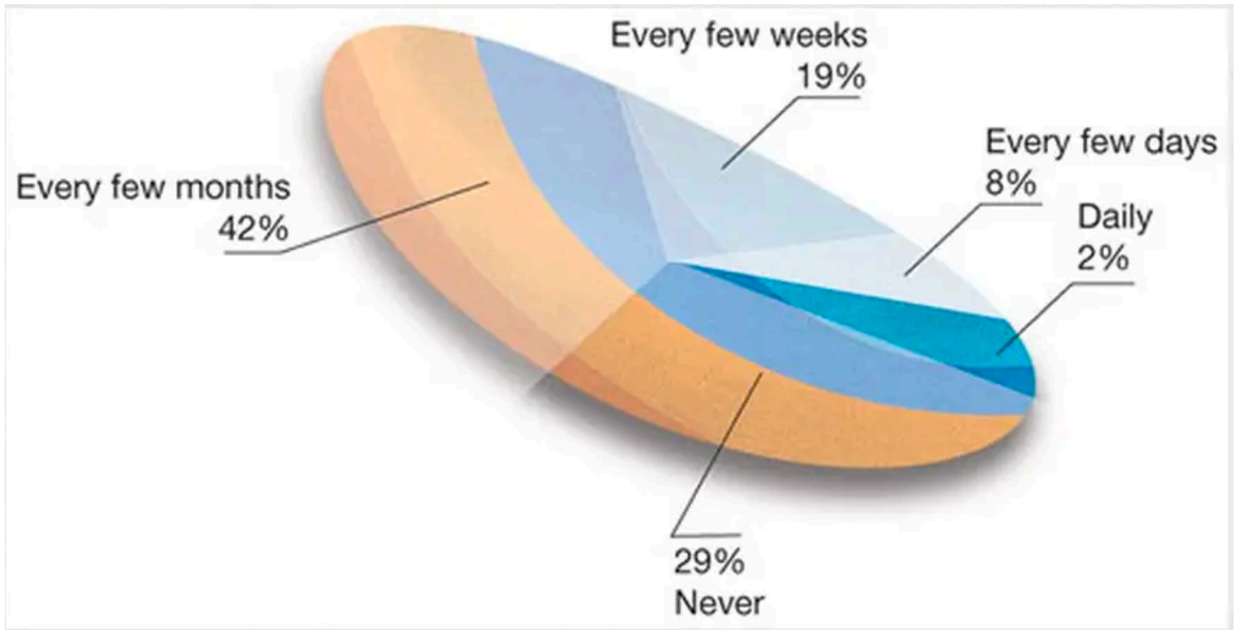
Tufte's Principles of Graphical Excellence:

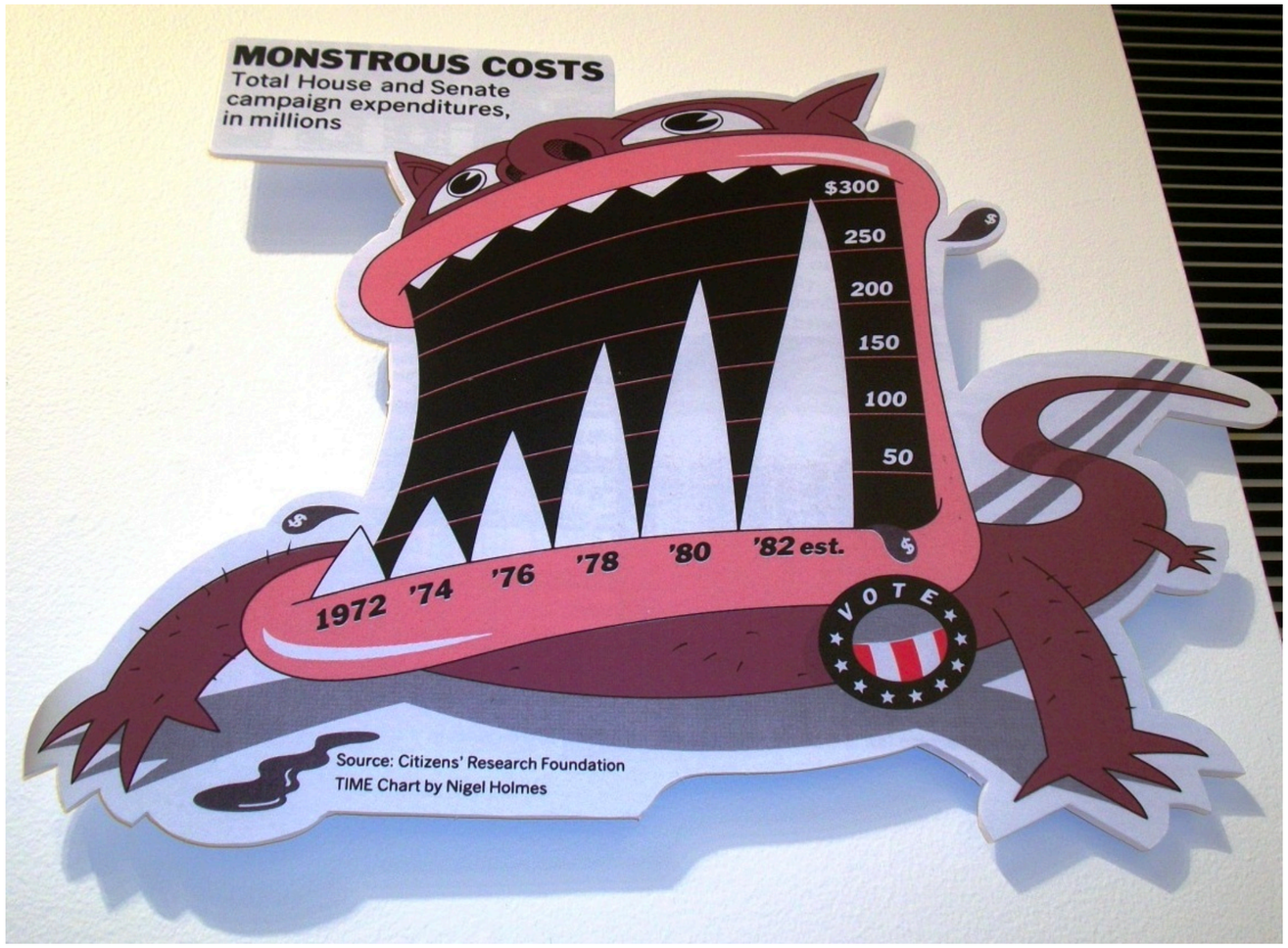
1. *Show the data*
2. *Induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else*
3. *Avoid distorting what the data have to say*
4. *Present many numbers in a small space*
5. *Make large data sets coherent*
6. *Encourage the eye to compare different pieces of data*
7. *Reveal the data at several levels of detail, from a broad overview to the fine structure*
8. *Serve a reasonably clear purpose: description, exploration, tabulation or decoration*
9. *Be closely integrated with the statistical and verbal descriptions of a data set*

outline

horror gallery

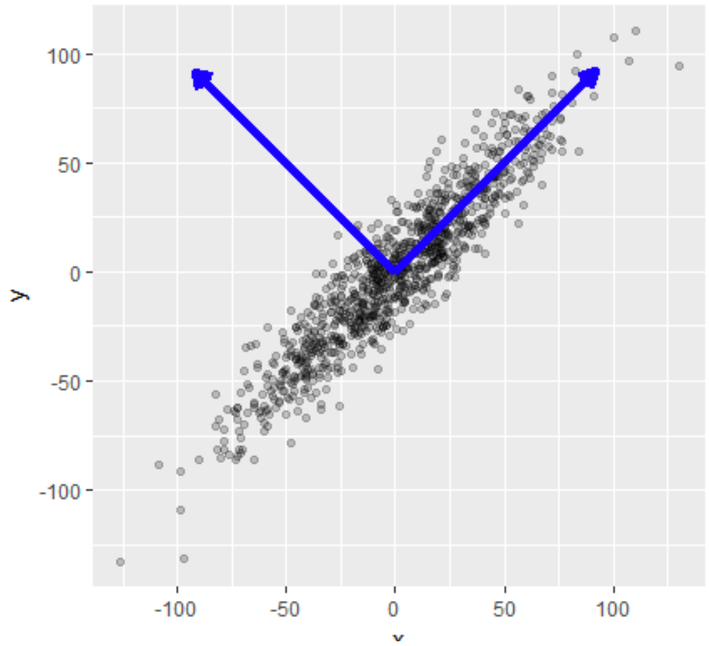


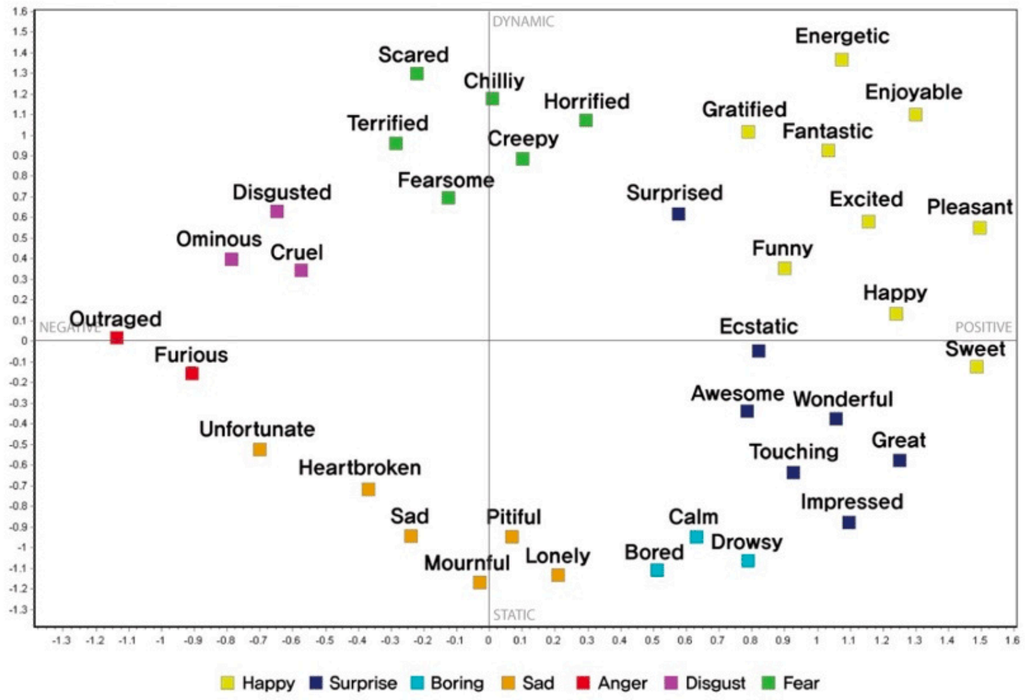


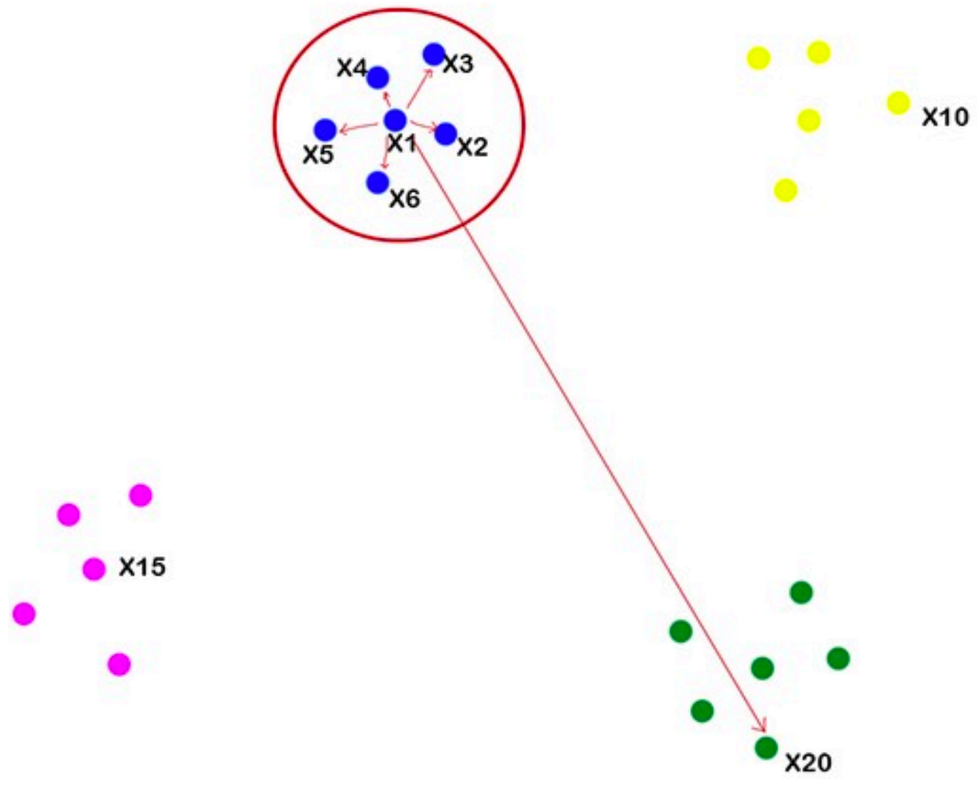


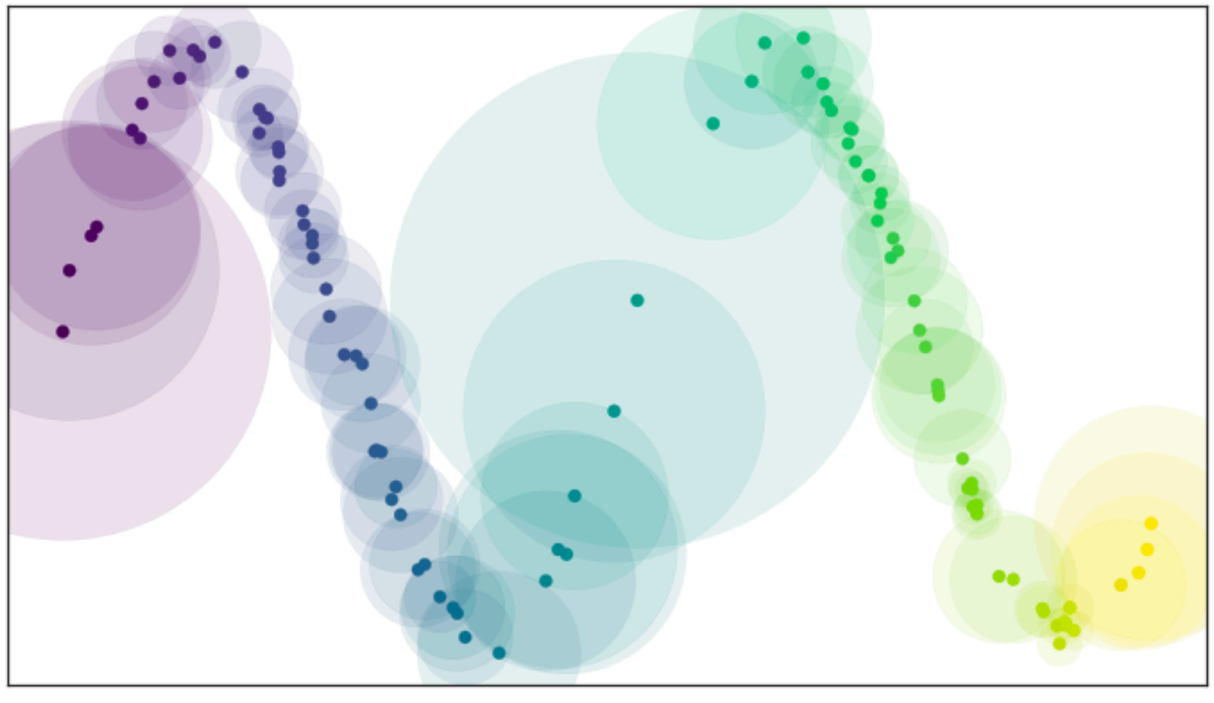
outline

a case study: dimensionality reduction for data viz

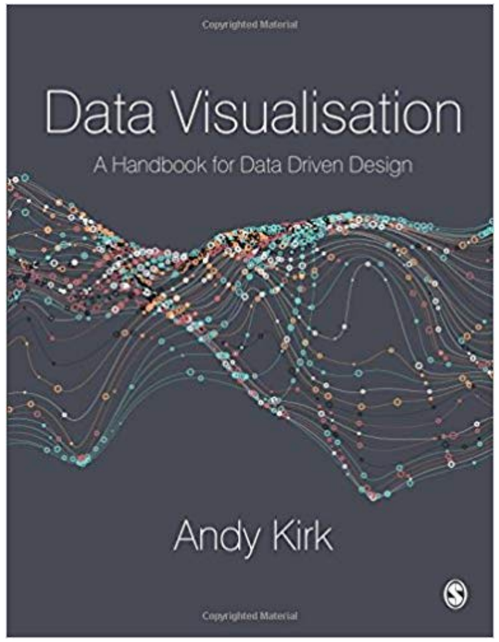


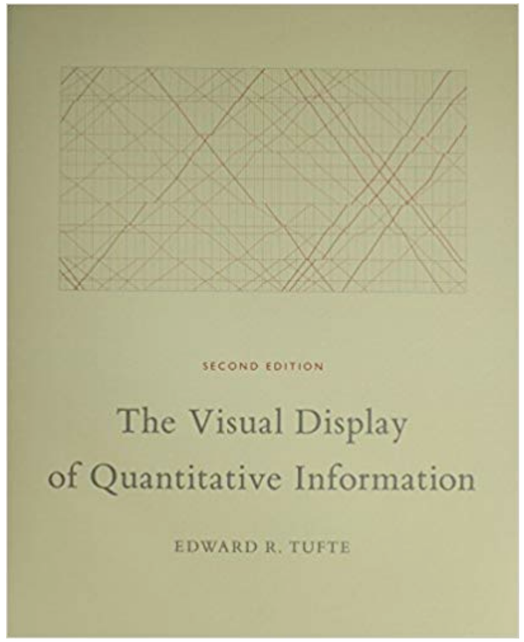


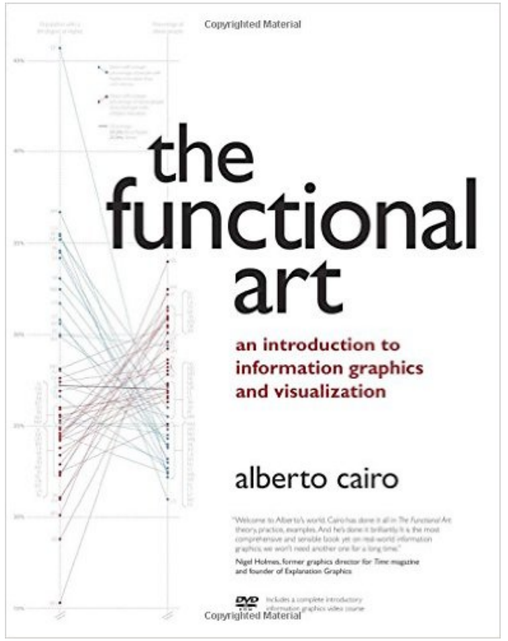




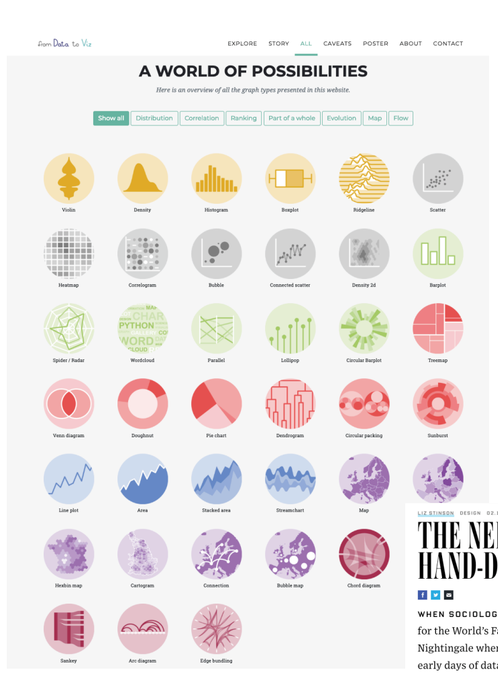
how/1

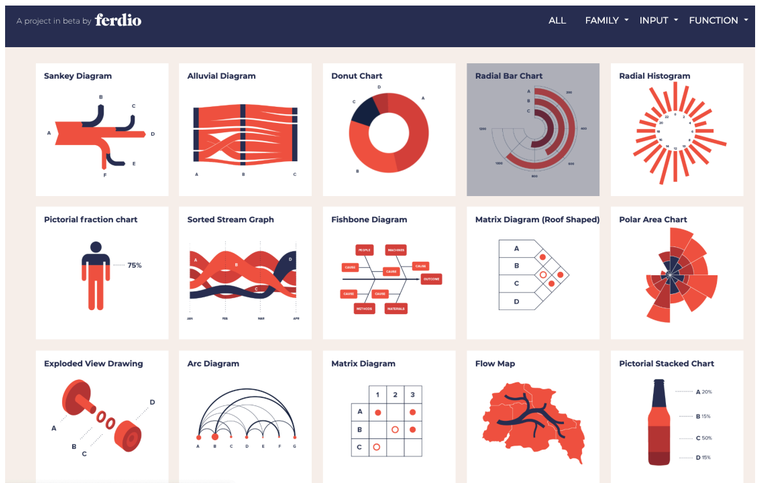


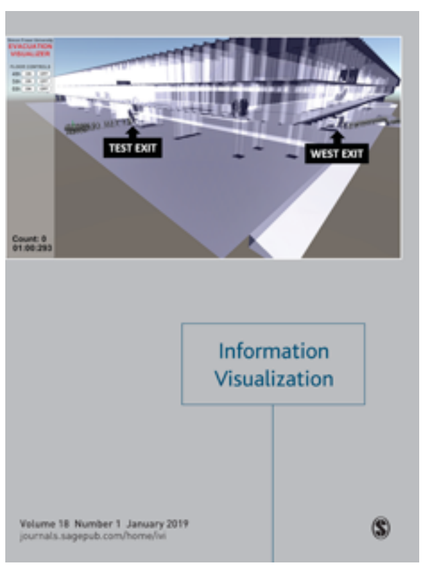


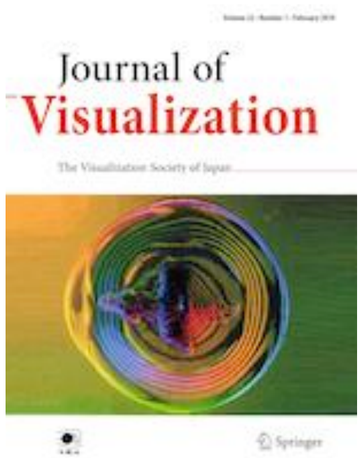


+









LIZ STINSON DESIGN 02.15.17 12:44 PM

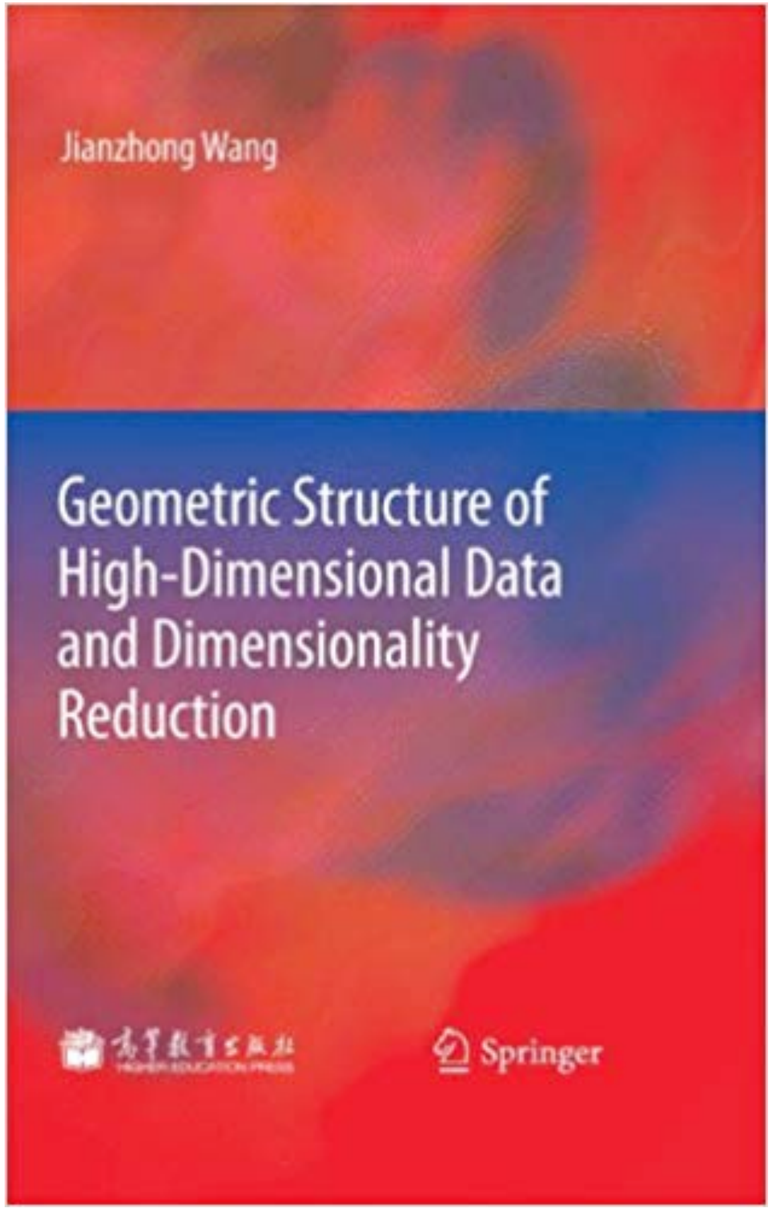
THE NERDY CHARM OF ARTISANAL, HAND-DRAWN INFOGRAPHICS

WHEN SOCIOLOGIST W. E. B. DU BOIS crafted his brilliant and colorful data visualizations for the World's Fair in 1900, he didn't have the help of a computer. Neither did Florence Nightingale when she visualized the causes of death in the Crimean War in the 1850s. In the early days of data visualization, people made infographics by hand because they had to. Today, that's not the case. Designers can use software, styli, and tablets to craft glossy data visualizations—and plenty of them do. But many still prefer simple tools—and use them to fantastic effect.

Giorgia Lupi and Stefanie Posavec

In a time where slick, computer-made visualizations are standard practice, information designers Giorgia Lupi and Stefanie Posavec prove analog illustrations can be just as compelling. A couple years ago, the pair started a project called *Dear Data*, for which the designers spent every week tracking a bit of personal data—moments of indecision or the number of times they laughed—and translating it into infographics. Those infographics went on postcards, which they mailed each other once a week. Lupi has described the postcards as small data, which hints not only at their quasi-qualitative nature, but their intimate aesthetic. The statistics aren't perfect, and neither are the drawings, but their imperfections lend the visualizations a more friendly, personal, and evocative feel.

STEFANI POSAVEC



JOSS
The Journal of Open Source Software

UMAP: Uniform Manifold Approximation and Projection

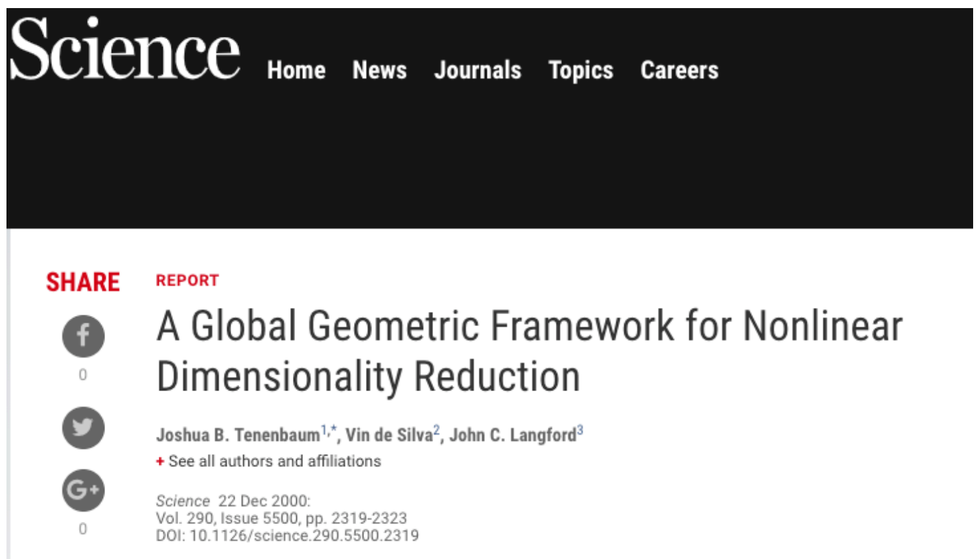
Leland McInnes[1], **John Healy**[1], **Nathaniel Saul**[2], **and Lukas Großberger**[3, 4]

Journal of Machine Learning Research 1 (2008) 1-48 Submitted 4/00; Published 10/00

Visualizing Data using t-SNE

Laurens van der Maaten L.VANDERMAATEN@MICC.UNIMAAS.NL
MICC-IKAT
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada



how/3

python™

SciPy

pandas

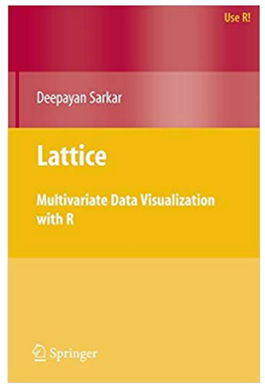
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

NumPy

matplotlib

Bokeh

ggplot2

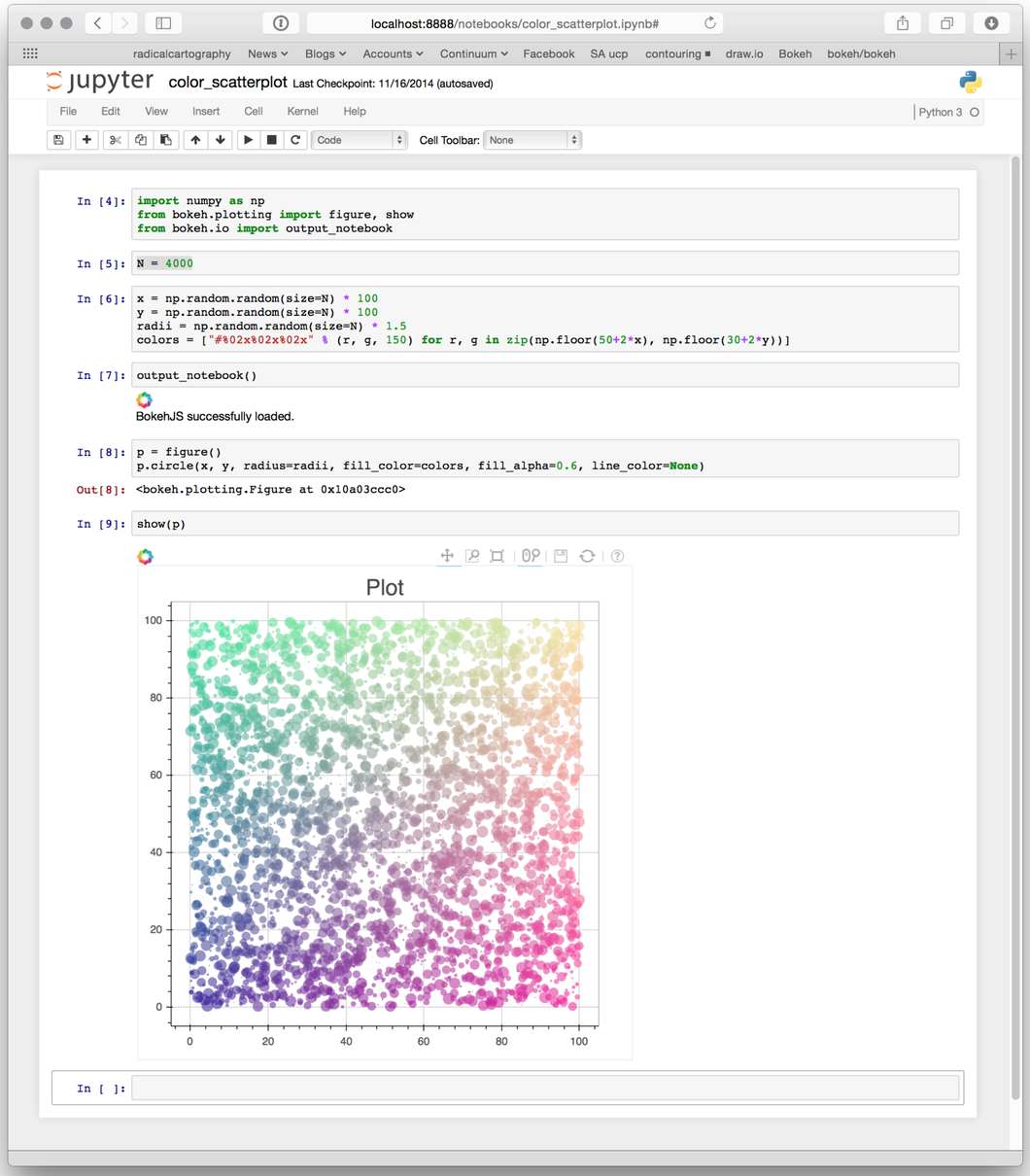


PyViz

jupyter

RStudio®

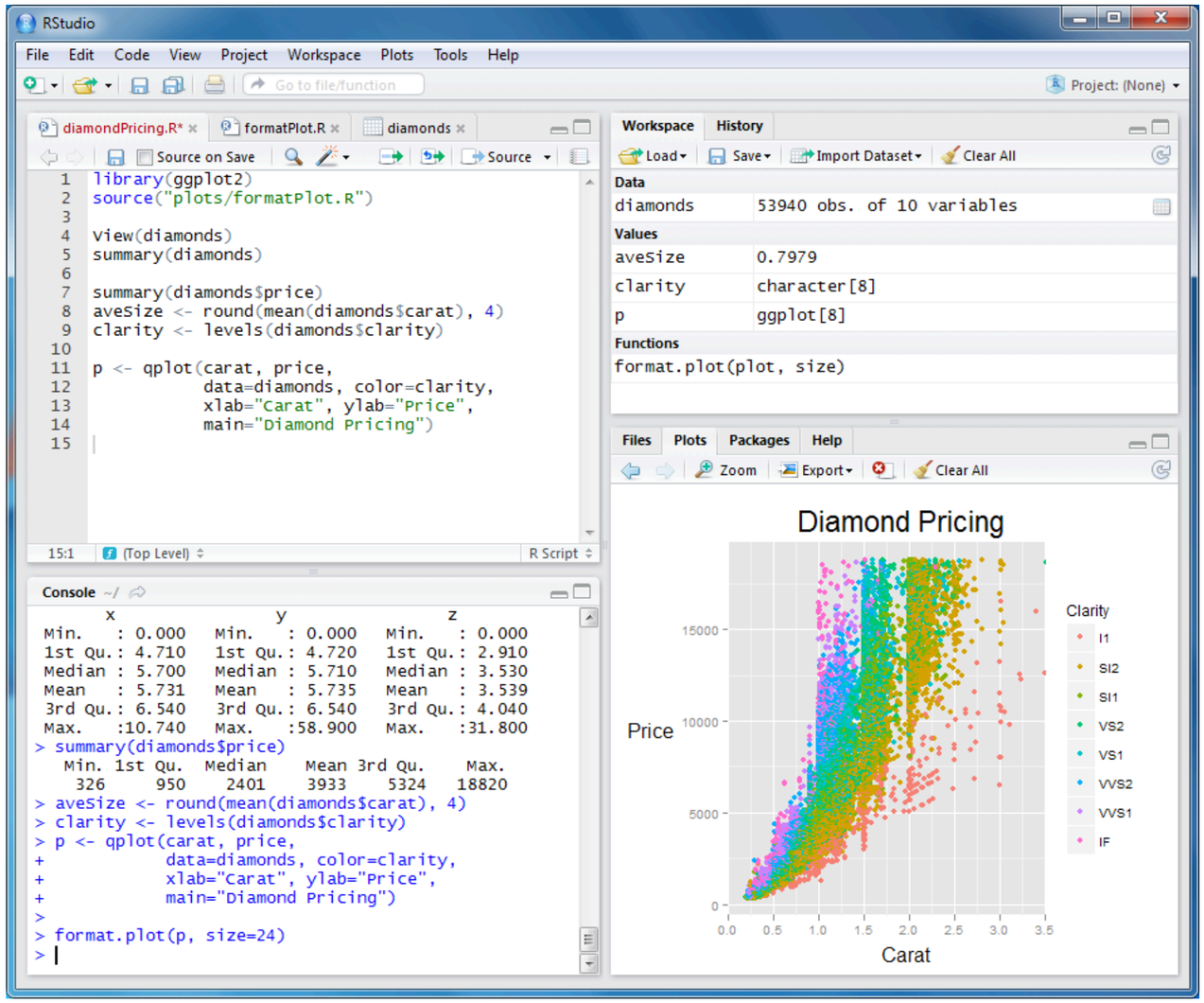
how/3



jupyter

… but I leave it to Shahryar to show you the details

R Studio



I THINK WE SHOULD GIVE IT ANOTHER SHOT.

WE SHOULD BREAK UP, AND I CAN PROVE IT.

HUH.

MAYBE YOU'RE RIGHT.

I KNEW DATA WOULD CONVINCE YOU.

NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.