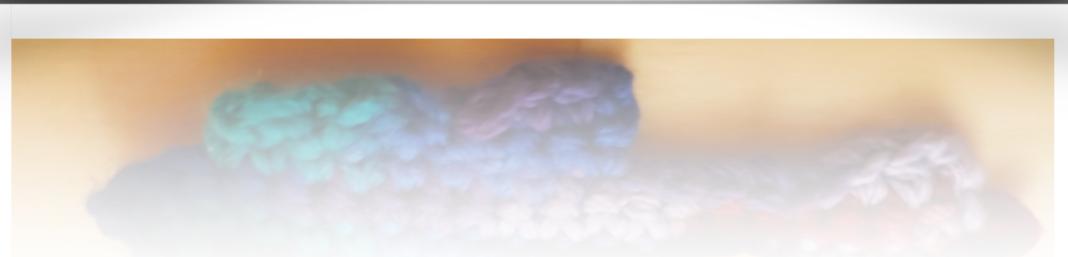




m.s.c. data science
data visualisation lab

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

4 March - 27 May

Lecture in presence

logistics

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

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

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

contacts



mmoroni@fbk.eu



snoei@fbk.eu

course material:
Moodle

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

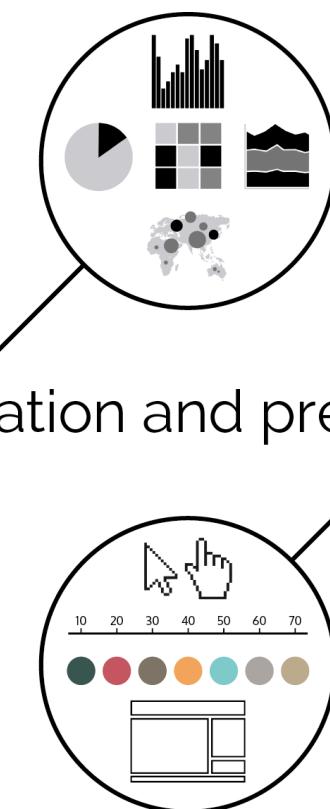
what

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



The representation and presentation of data to facilitate understanding

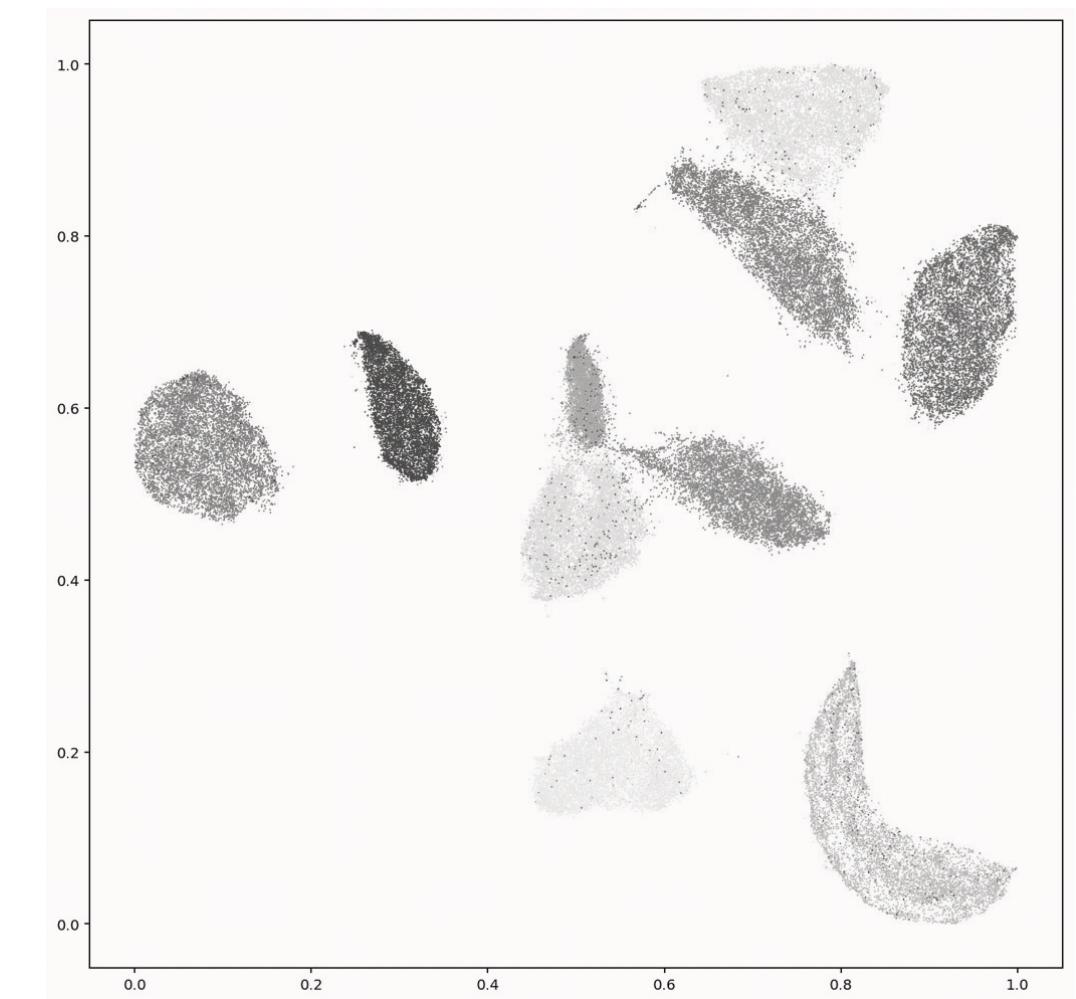
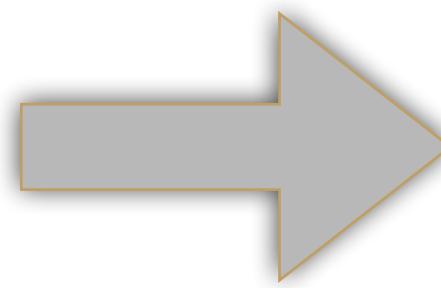


Kirk, 2016

dimensionality reduction algorithms



n-dim data



why

... off the beaten path

DASH DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD DASH.HARVARD.EDU

HARVARD LIBRARY Office for Scholarly Communication

DASH Home » Harvard Kennedy School » HKS Faculty Scholarship » View Item

Search Search DASH

Salience, Credibility, Legitimacy and Boundaries: Linking Research, Assessment and Decision Making

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

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	number of observations	11
10	8,04	mean of x	9.0
8	6,95	mean of y	7.5
13	7,58	correlation between x and y	0.816
9	8,81	regression line	$y=0.5x+3$
11	8,33	variance of x	11
14	9,96	variance of y	4.12
6	7,24		
4	4,26		
12	10,84		
7	4,82		
5	5,68		

why

back down to earth: the anscombe quartet

x	y	number of observations	11
10	9,14	mean of x	9.0
8	8,14	mean of y	7.5
13	8,74	correlation between x and y	0.816
9	8,77	regression line	$y=0.5x+3$
11	9,26	variance of x	11
14	8,10	variance of y	4.12
6	6,13		
4	3,10		
12	9,13		
7	7,26		
5	4,74		

why

back down to earth: the anscombe quartet

x	y	number of observations	11
10	7,46	mean of x	9.0
8	6,77	mean of y	7.5
13	12,74	correlation between x and y	0.816
9	7,11	regression line	$y=0.5x+3$
11	7,81	variance of x	11
14	8,84	variance of y	4.12
6	6,08		
4	5,39		
12	8,15		
7	6,42		
5	5,73		

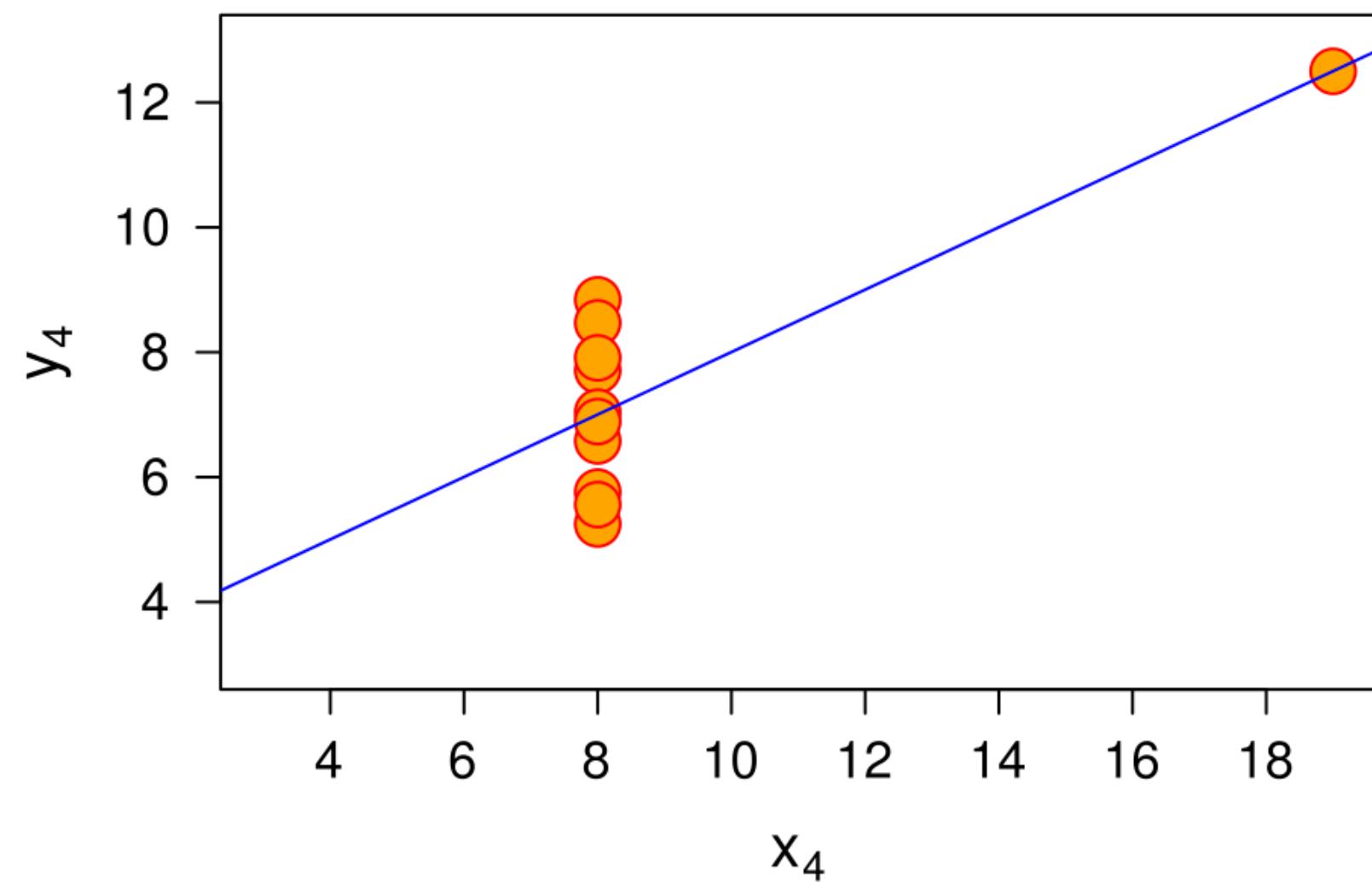
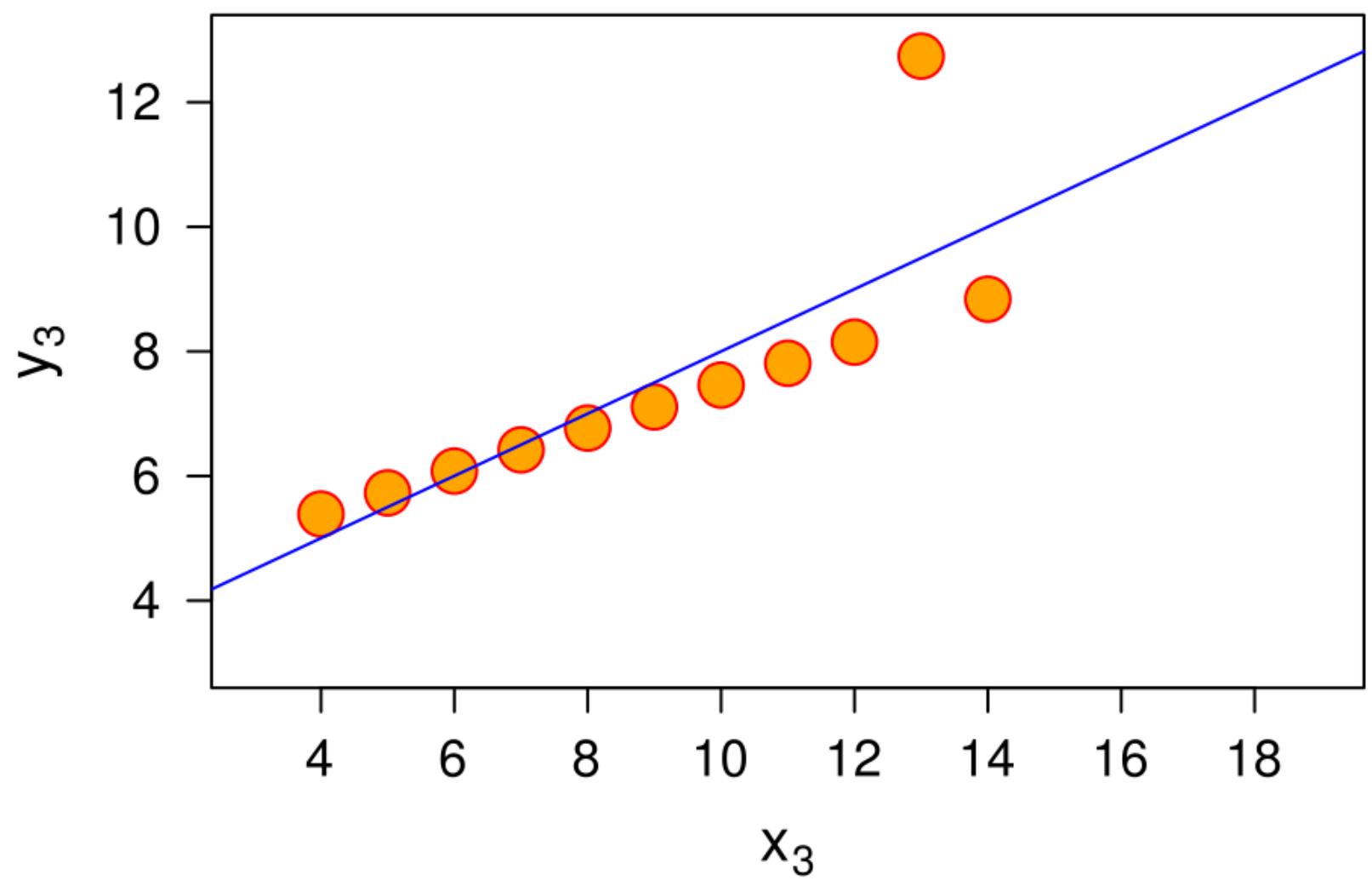
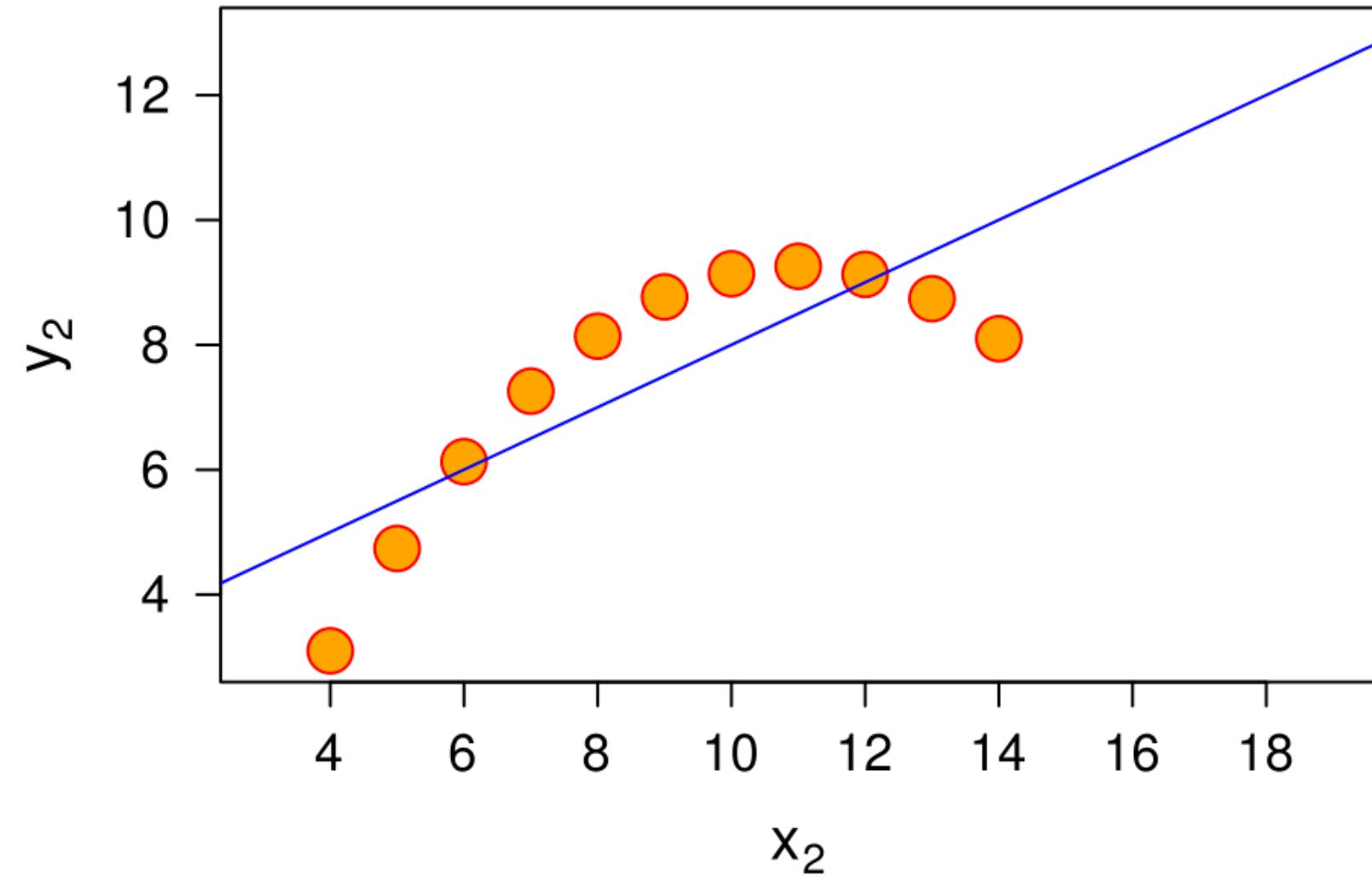
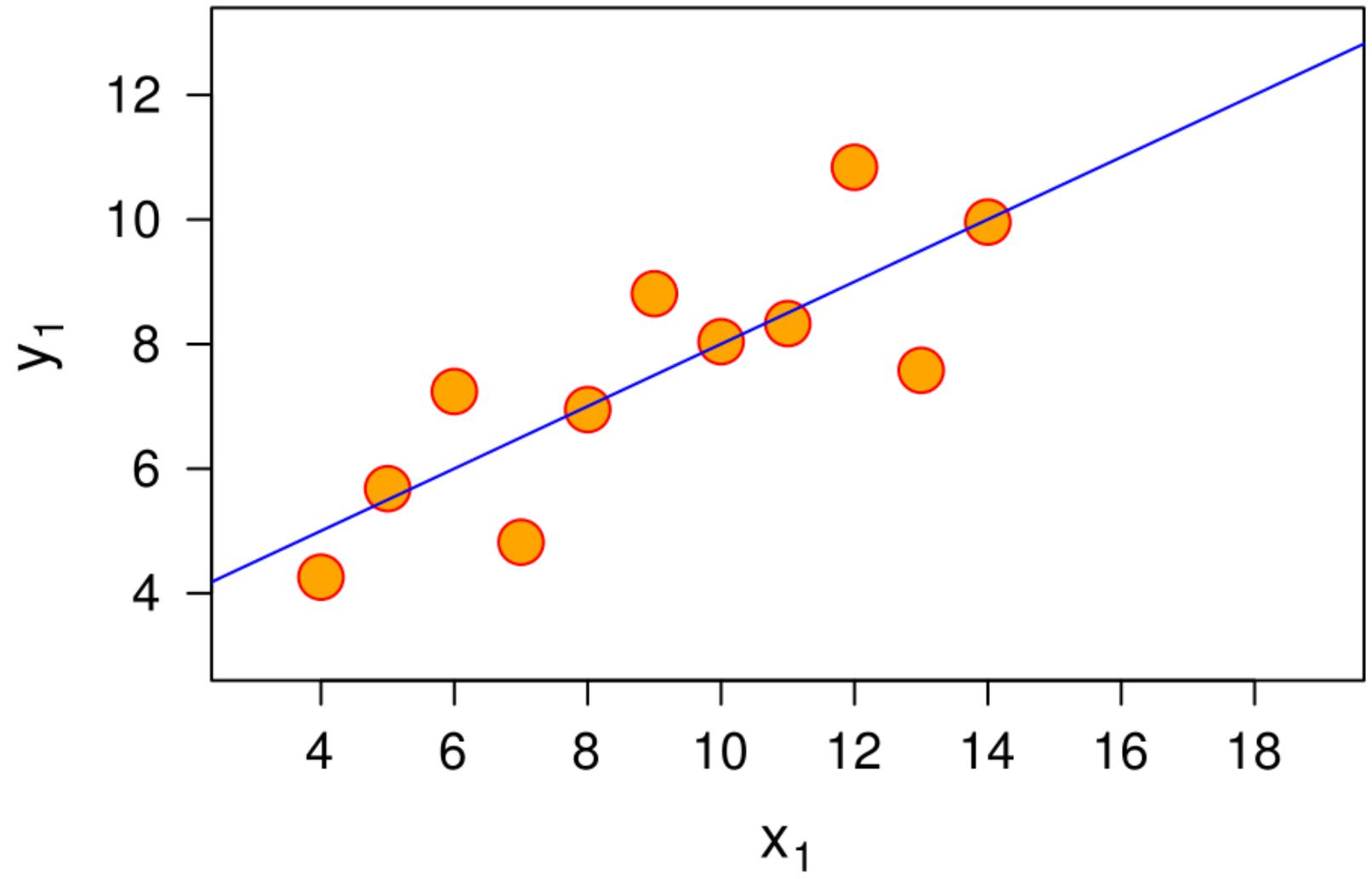
why

back down to earth: the anscombe quartet

x	y	number of observations	11
8	6,58	mean of x	9.0
8	5,76	mean of y	7.5
8	7,71	correlation between x and y	0.816
8	8,84	regression line	$y=0.5x+3$
8	8,47		
8	7,04		
8	5,25		
19	12,50	variance of x	11
8	5,56	variance of y	4.12
8	7,91		
8	6,89		

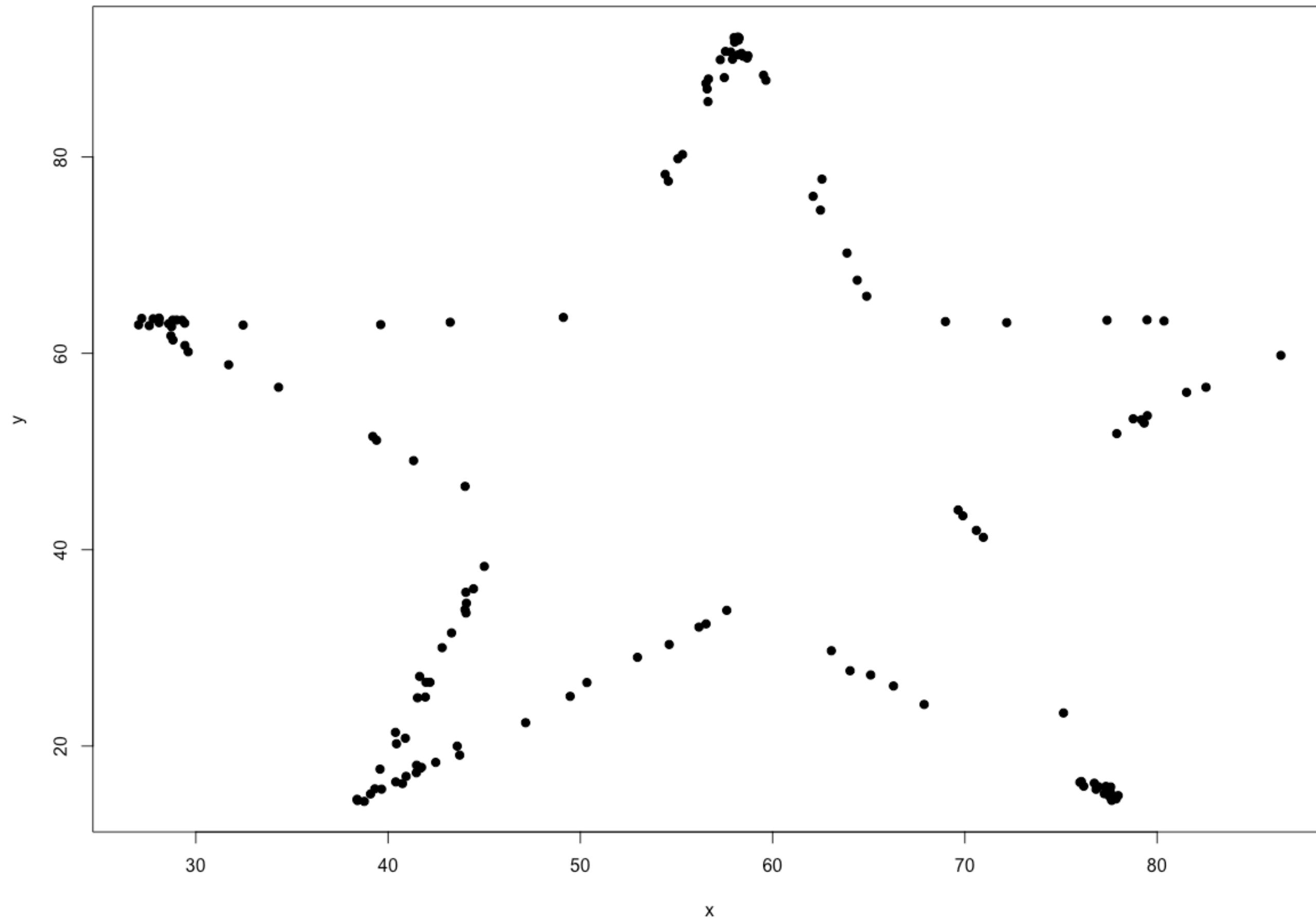
1973

back down to earth: the anscombe quartet



why

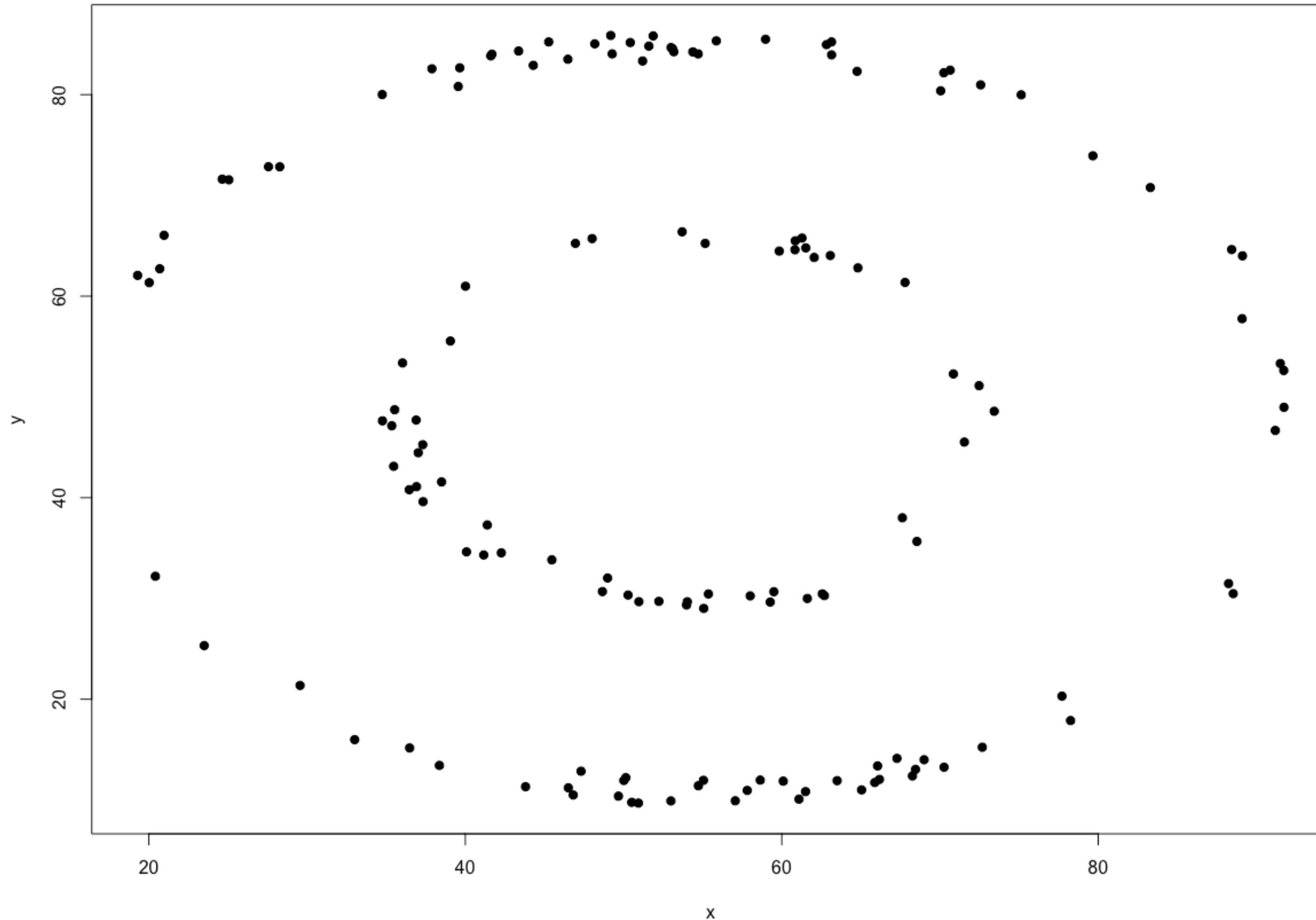
even harder: alberto cairo's datasaurus



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

why

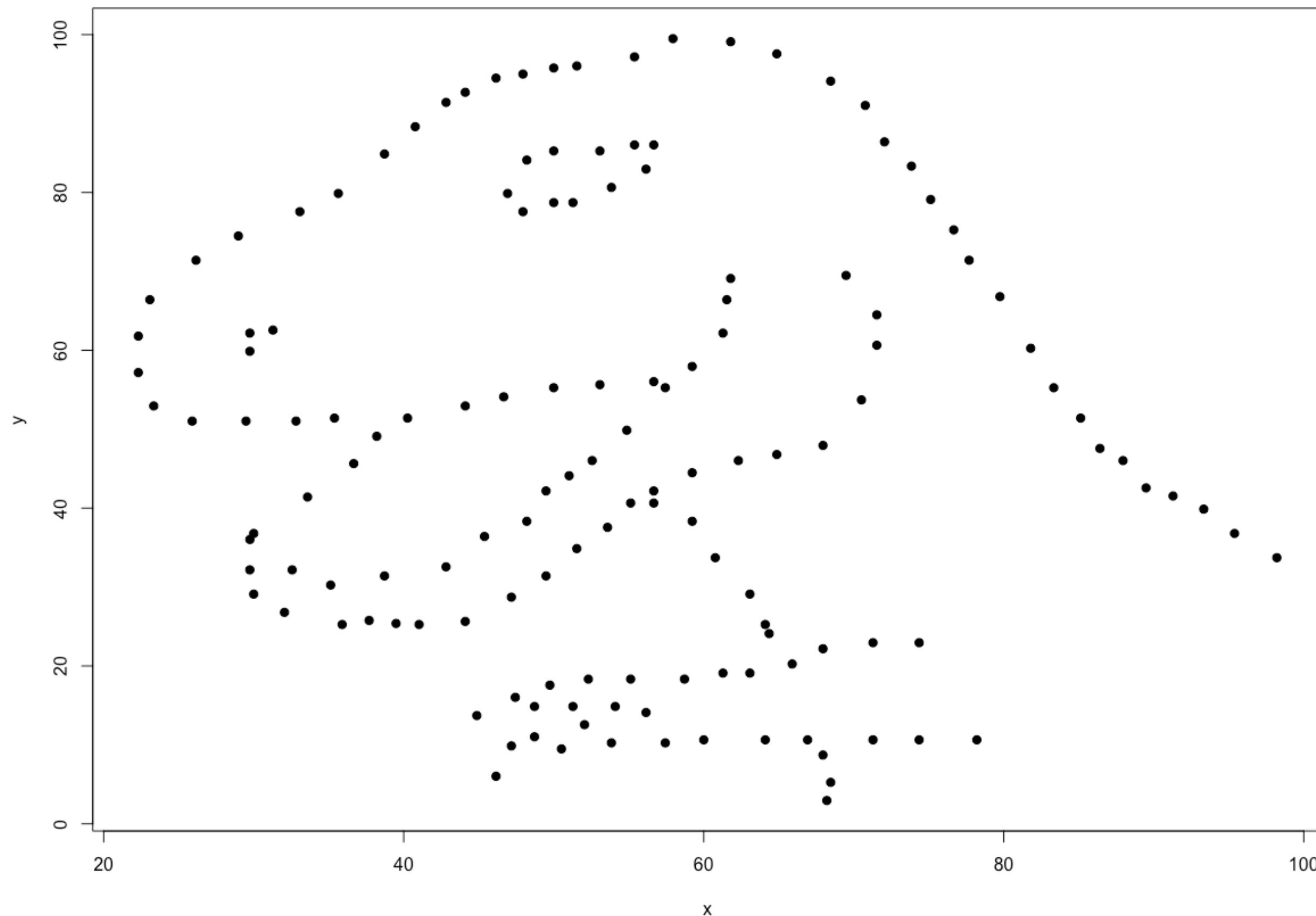
even harder: alberto cairo's datasaurus



and even more...

why

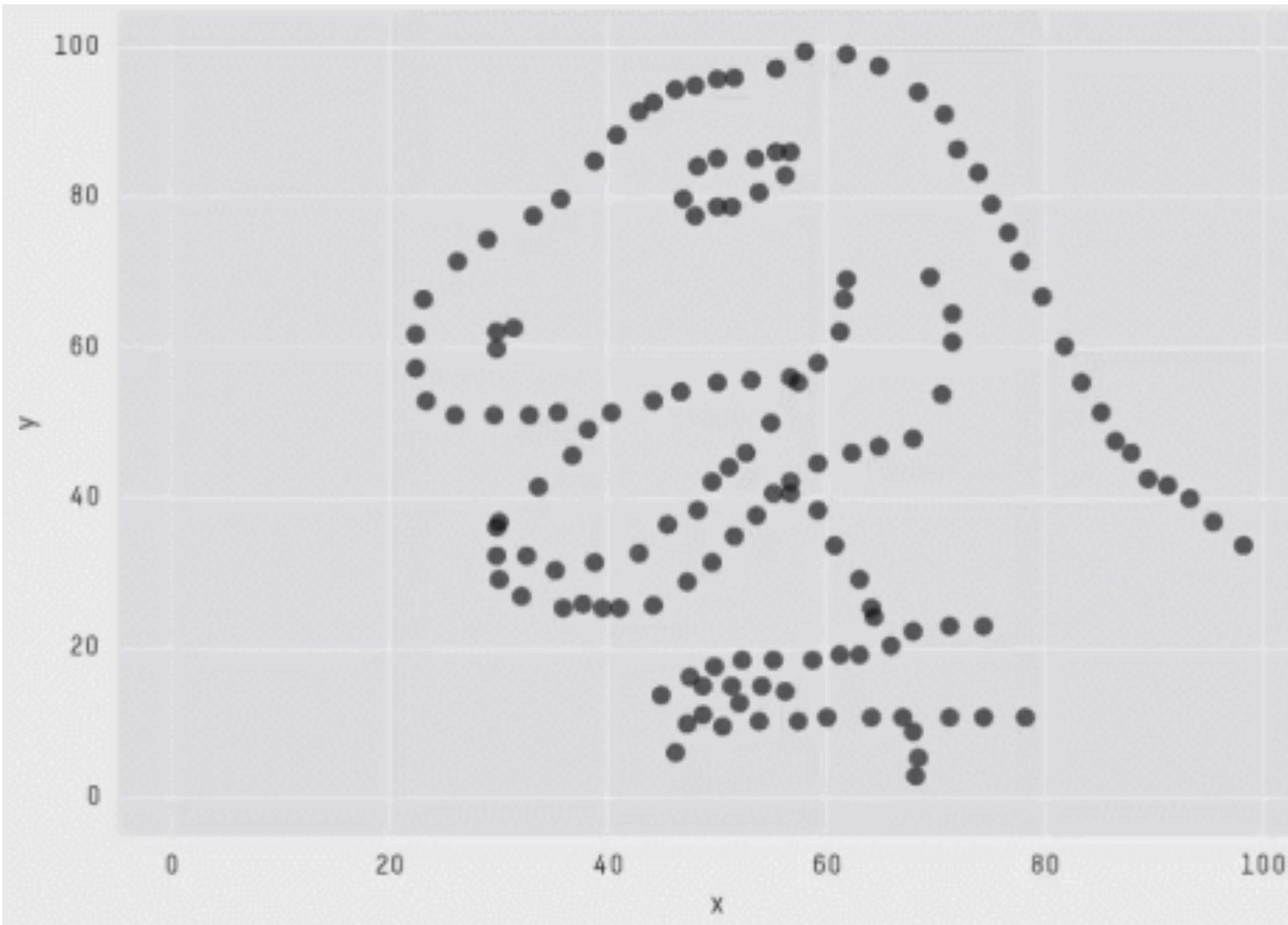
even harder: alberto cairo's datasaurus



actually, a wealth of nice shapes...

why

even harder: alberto cairo's datasaurus



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

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

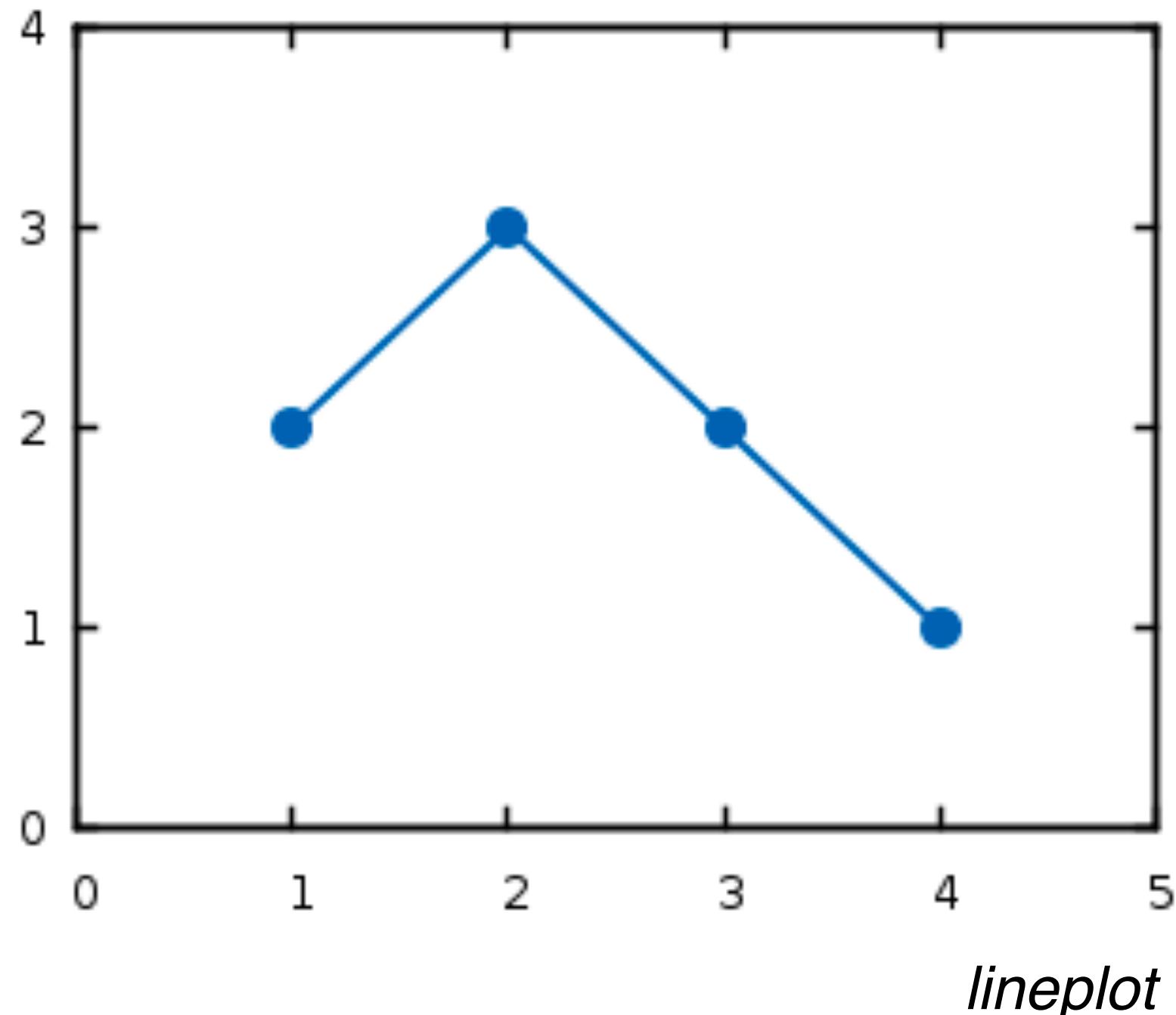
ACM SIGCHI Conference on Human Factors in Computing Systems



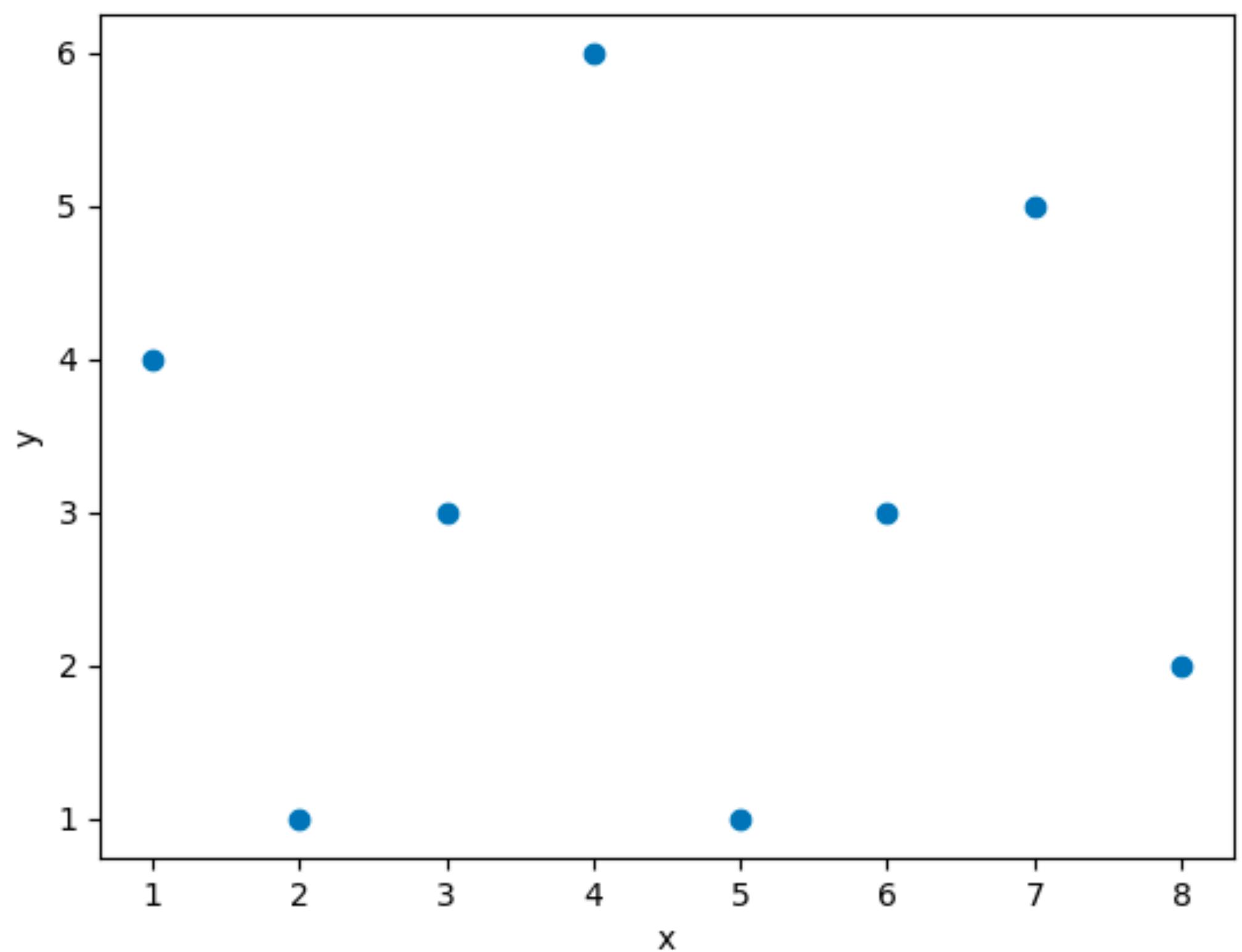
DOWNLOAD PDF

why
the goal

understanding and creating graphical representation of data



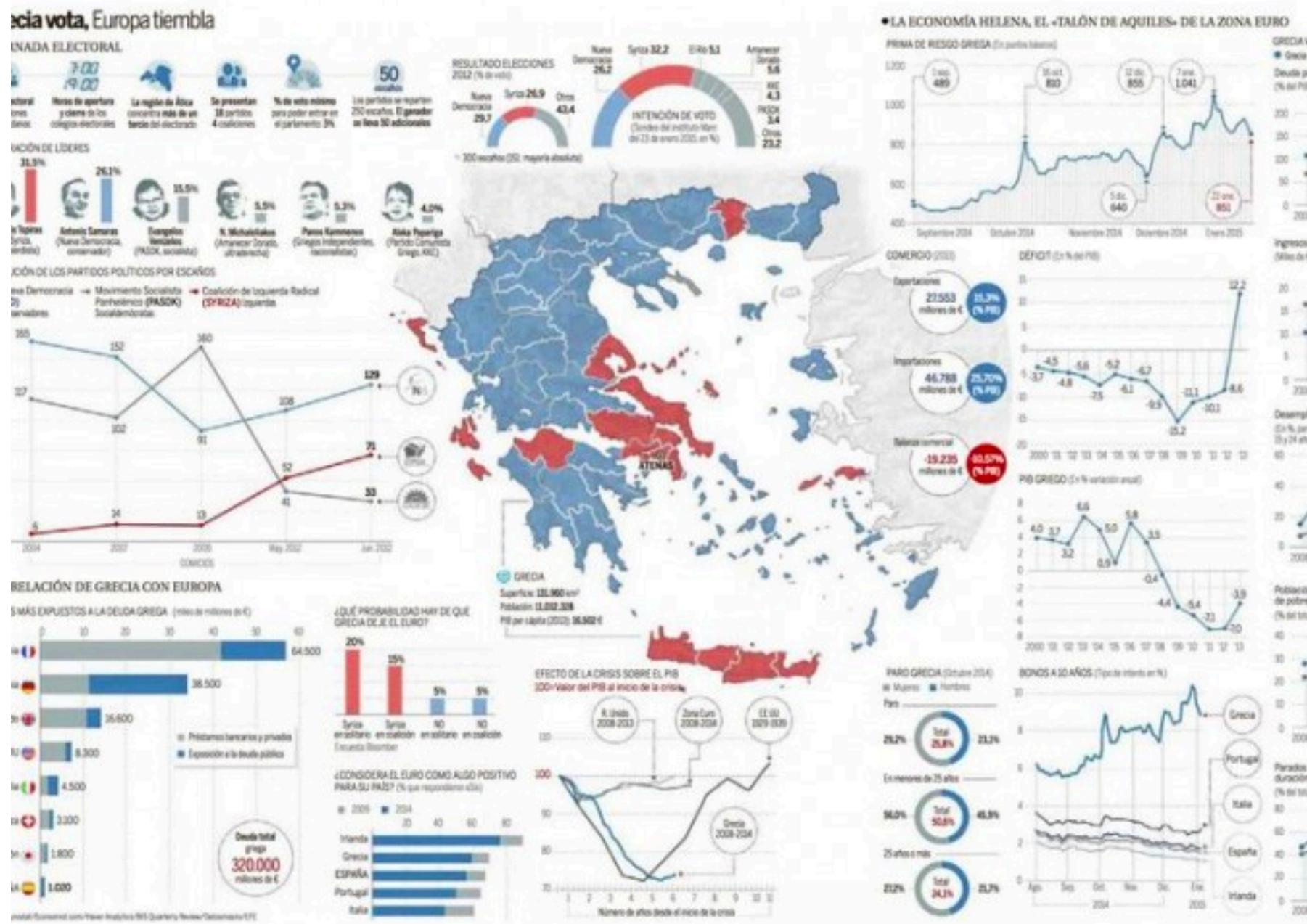
... from very simple plots ...



why

the goal

understanding and creating graphical representation of data

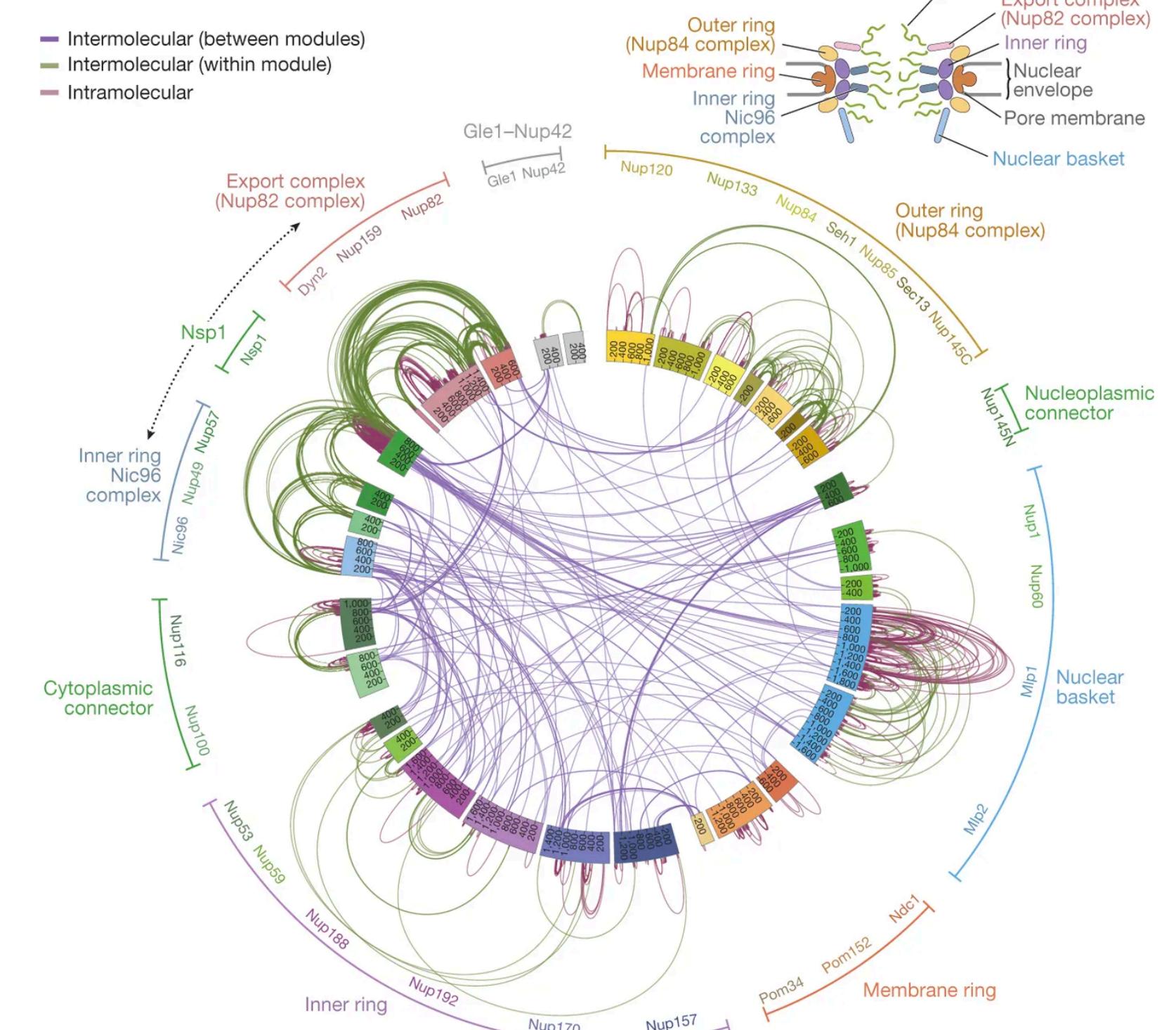


data journalism

... to complex panels ...

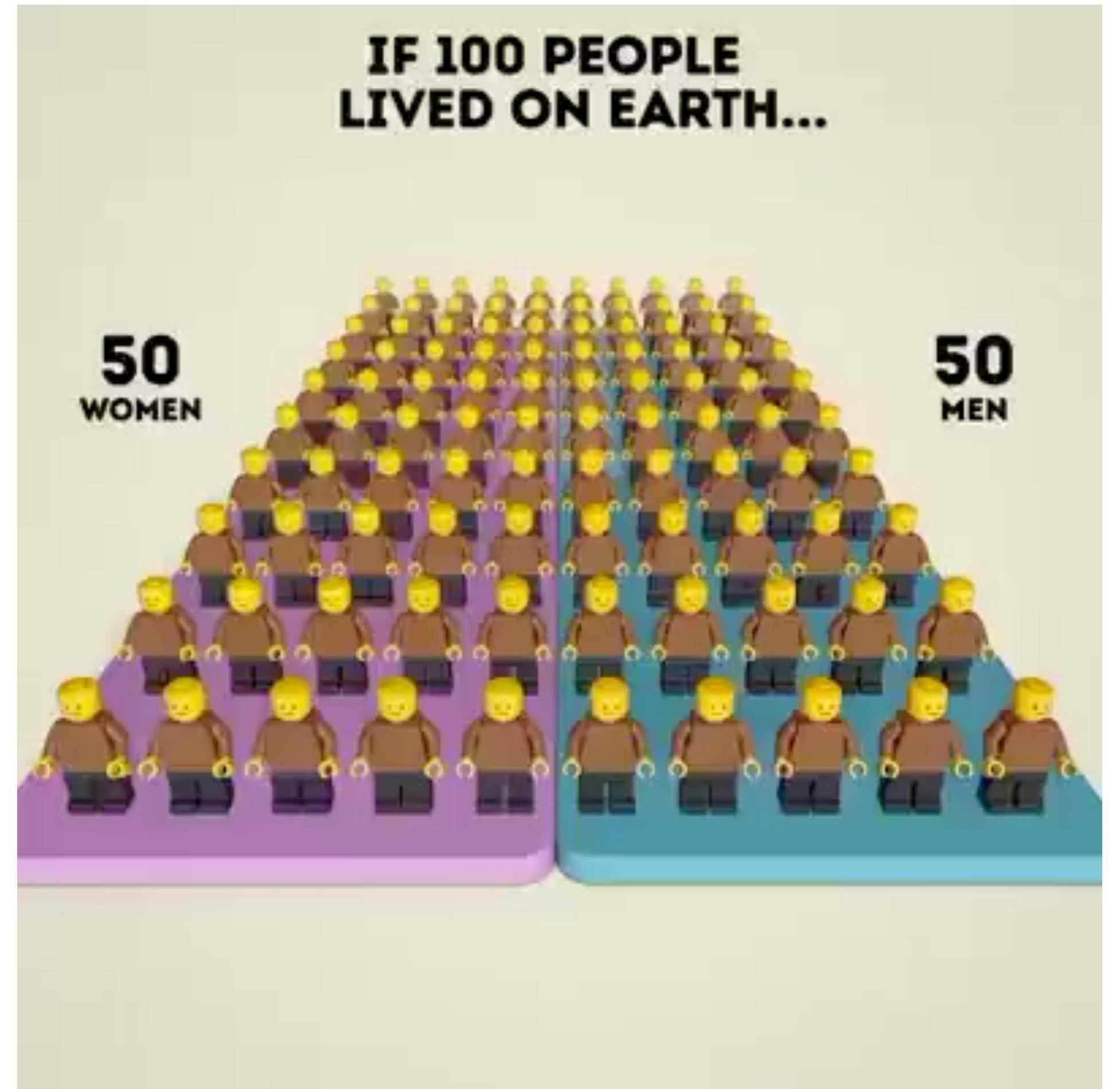
Circos plot of the 3,077 chemical cross-links in NPC

- Intermolecular (between modules)
- Intermolecular (within module)
- Intramolecular



why
the goal

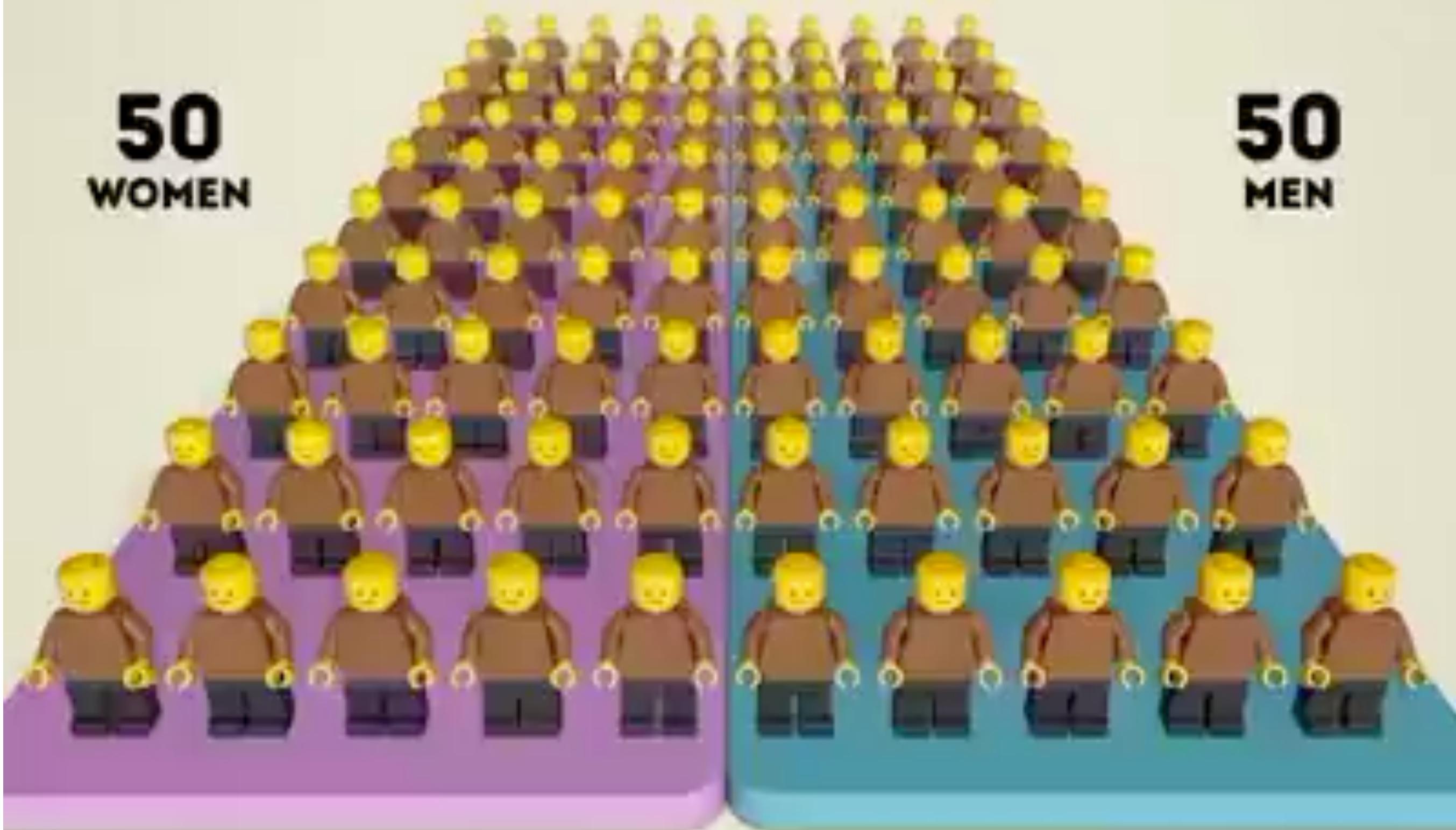
understanding and creating graphical representation of data



**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

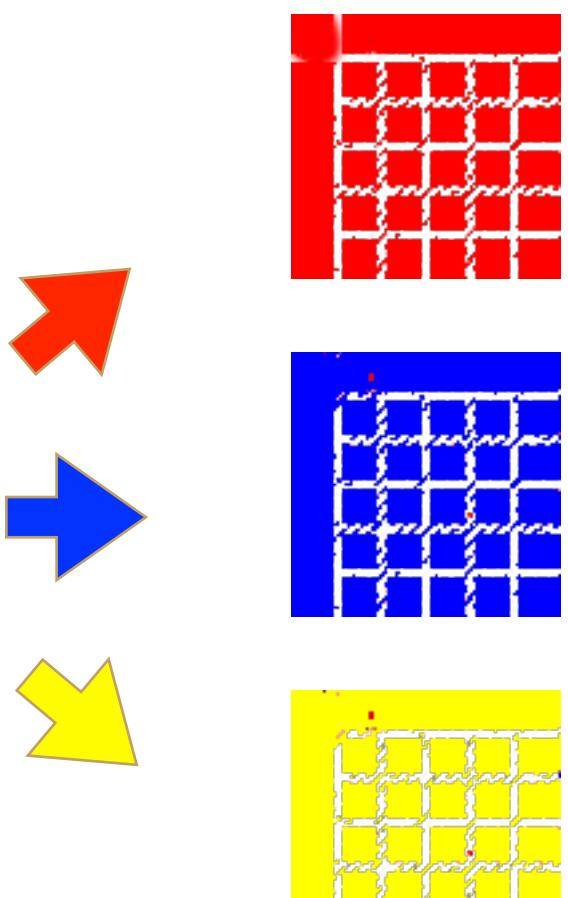
“a picture is worth a thousand words”

... if you know how to read it!

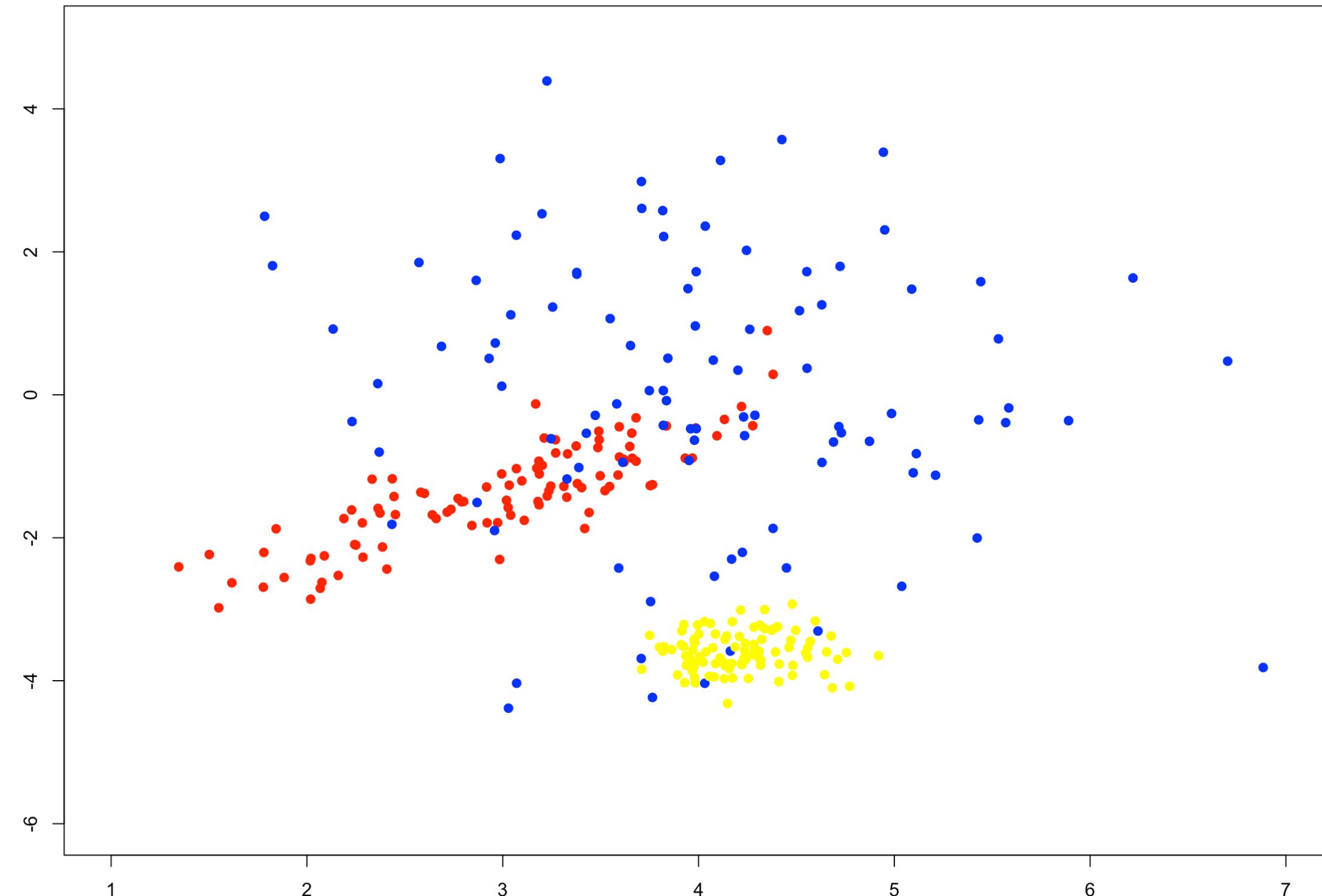
[a. cairo, 2019]



classification task



why
easy or hard?



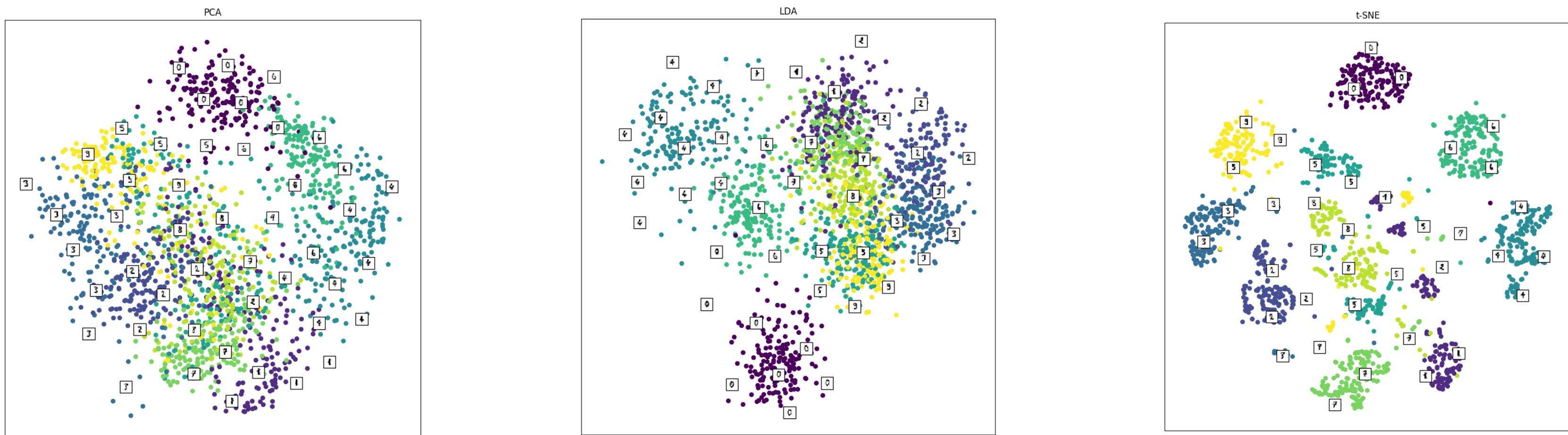


why

a wealth of choices...

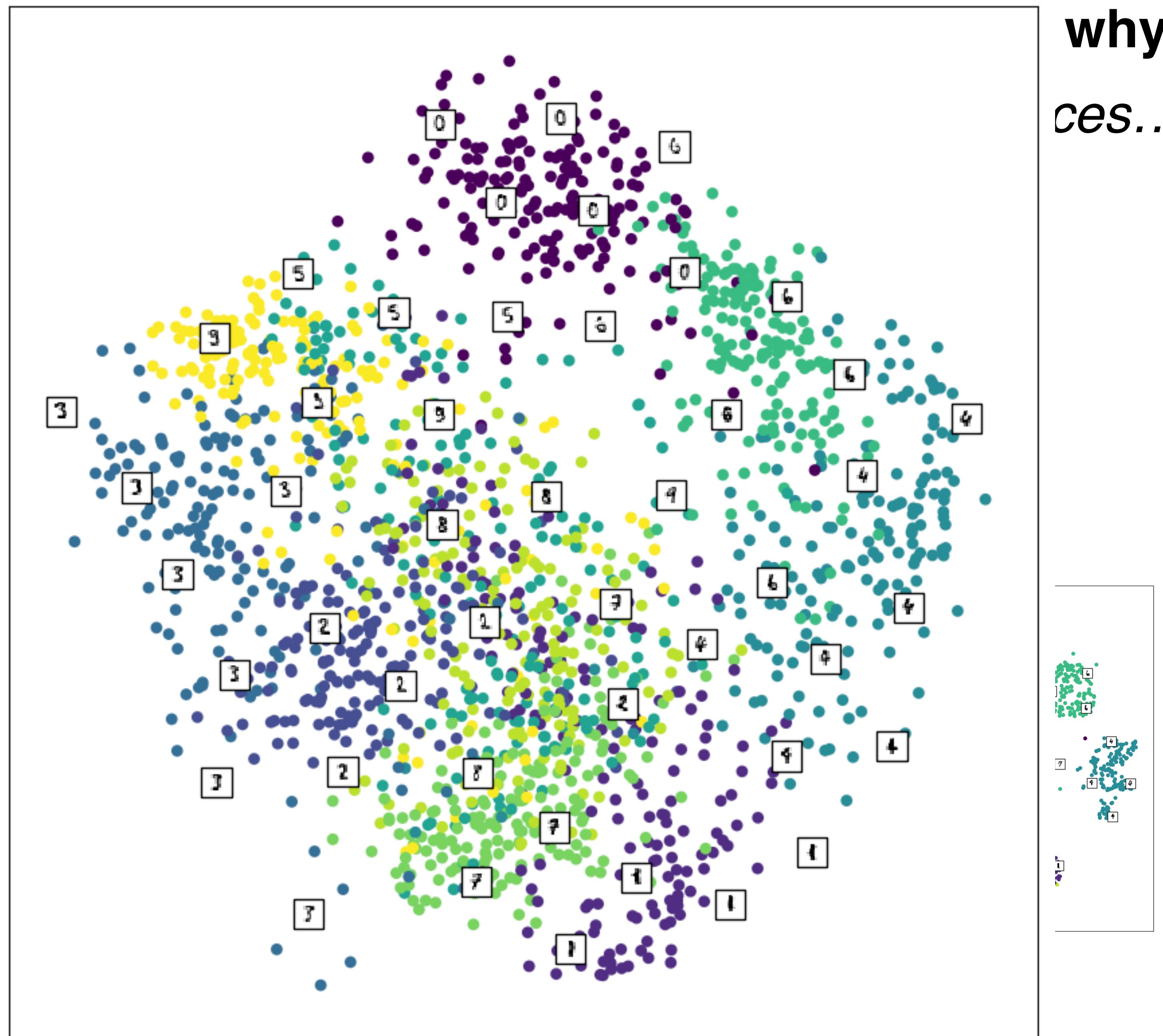
mnist dataset
[LeCun et al., 2013]

60K+10K images



PCA

0	0	0	0
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9

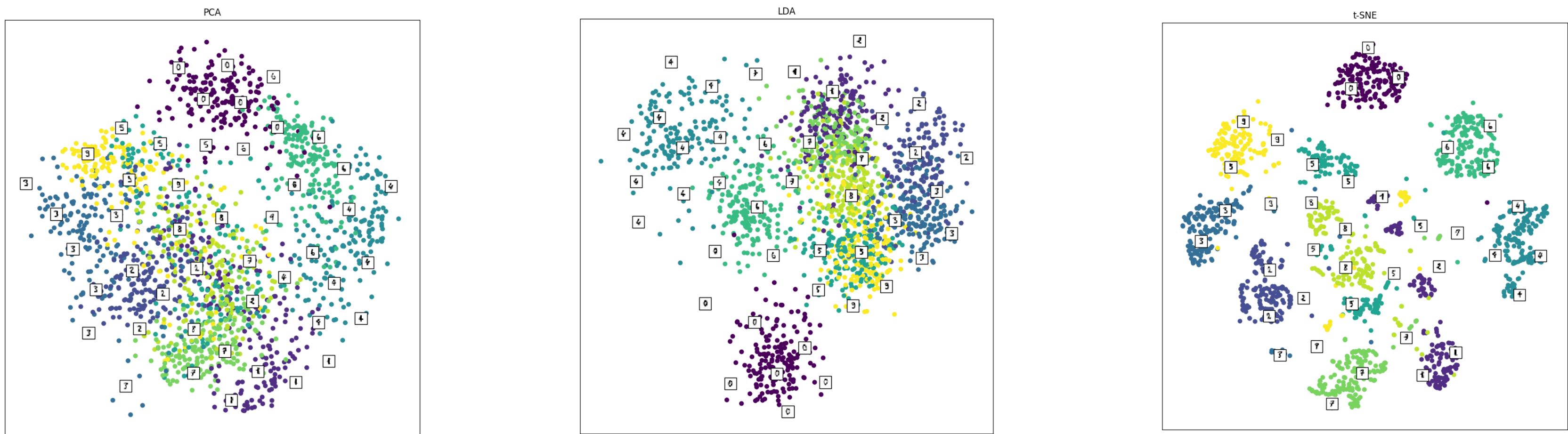


why
a wealth of choices...

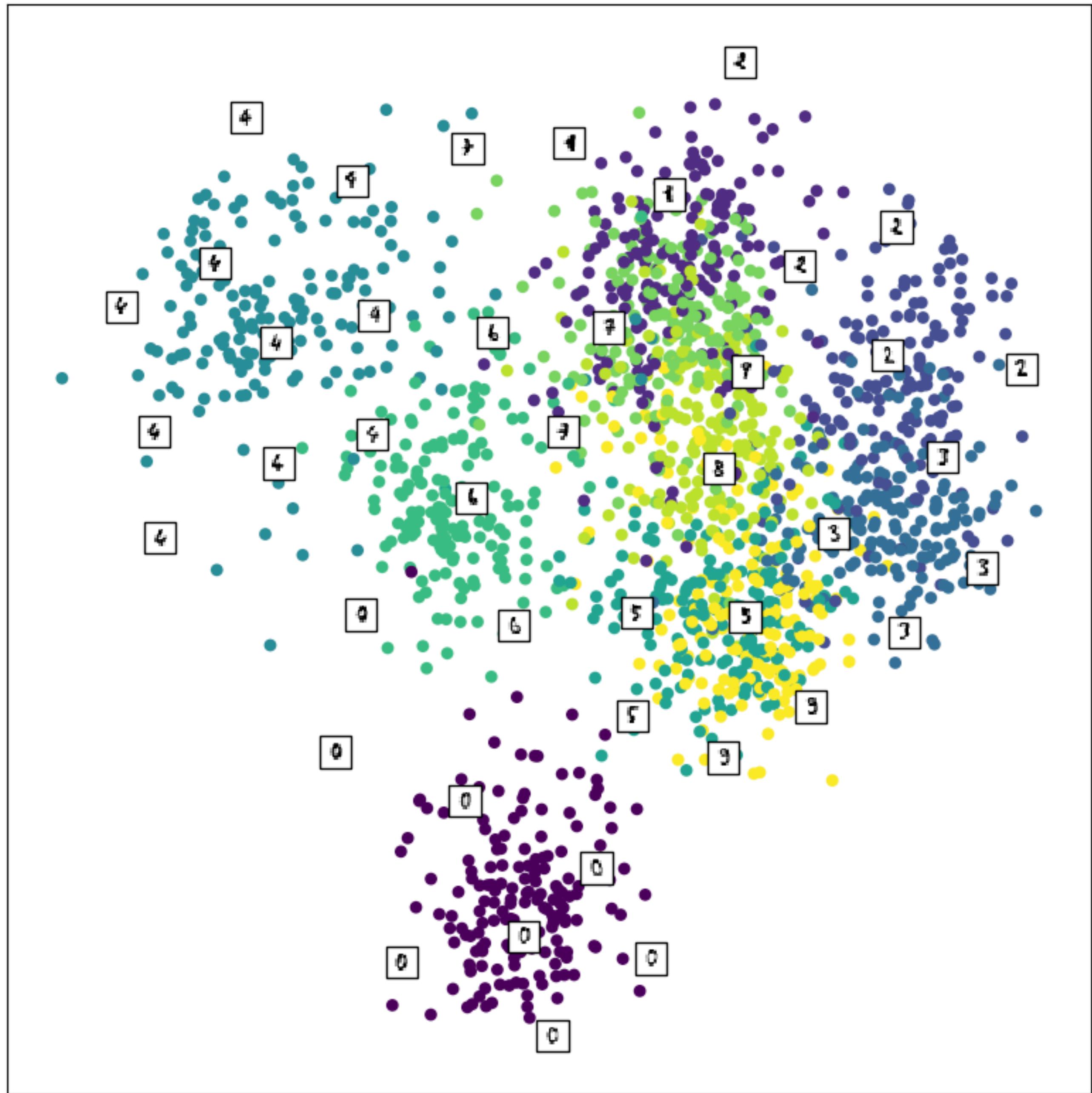
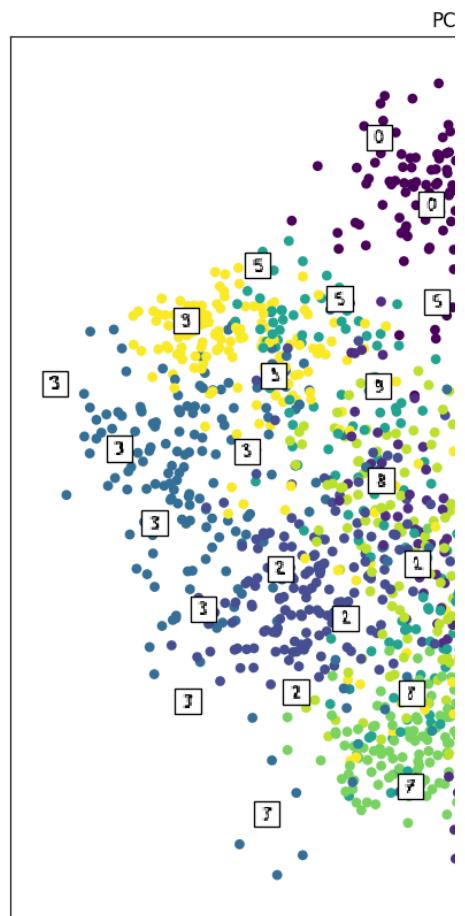


mnist dataset
[LeCun et al., 2013]

60K+10K images



0 0 0 0 0
1 1 1 1 1
2 2 2 2 2
3 3 3 3 3
4 4 4 4 4
5 5 5 5 5
6 6 6 6 6
7 7 7 7 7
8 8 8 8 8
9 9 9 9 9



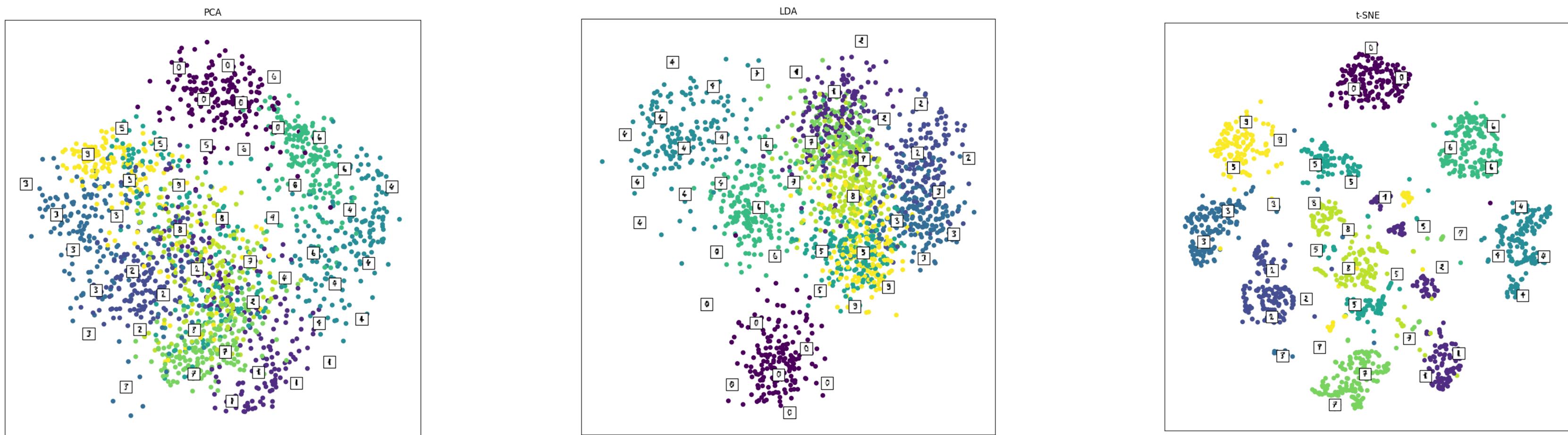
why
yes...

why
a wealth of choices...



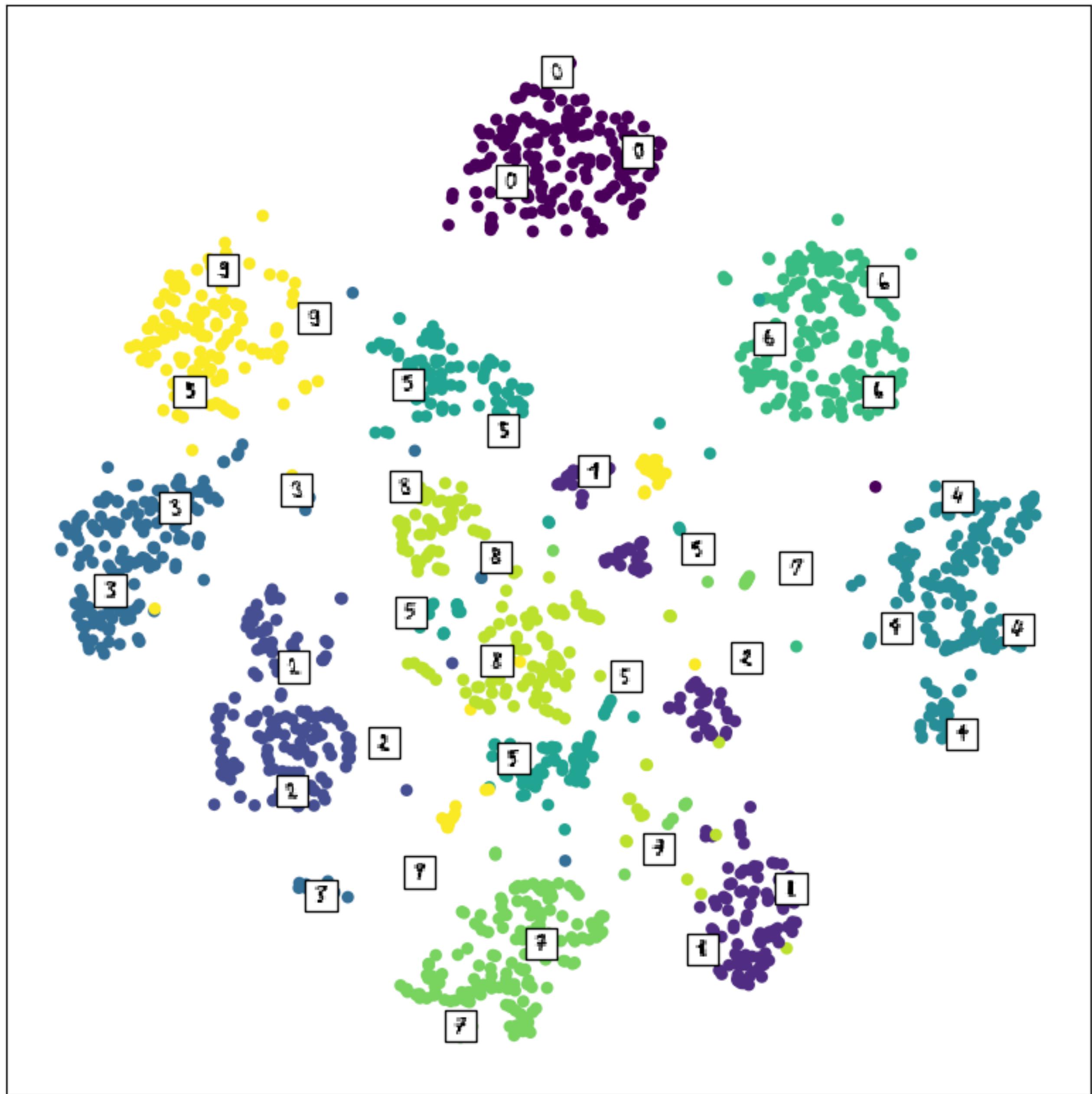
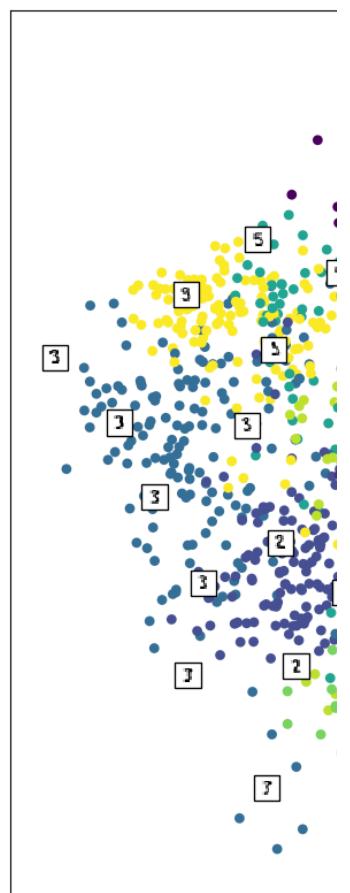
mnist dataset
[LeCun et al., 2013]

60K+10K images



why
dices...

0 0 0
1 1 1
2 2 2
3 3 3
4 4 4
5 5 5
6 6 6
7 7 7
8 8 8
9 9 9

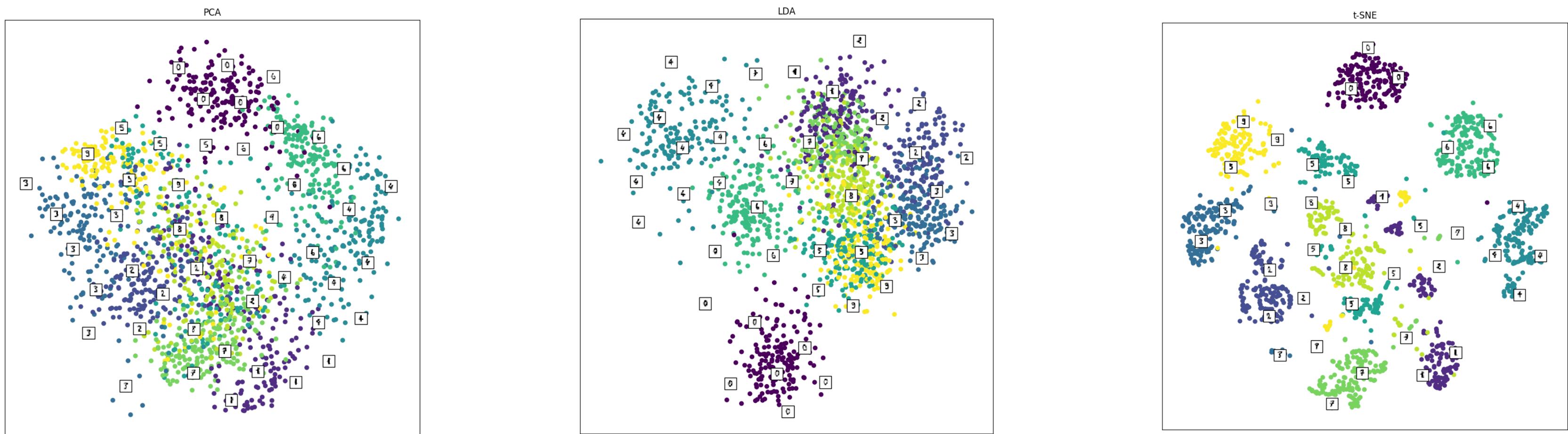


why
a wealth of choices...



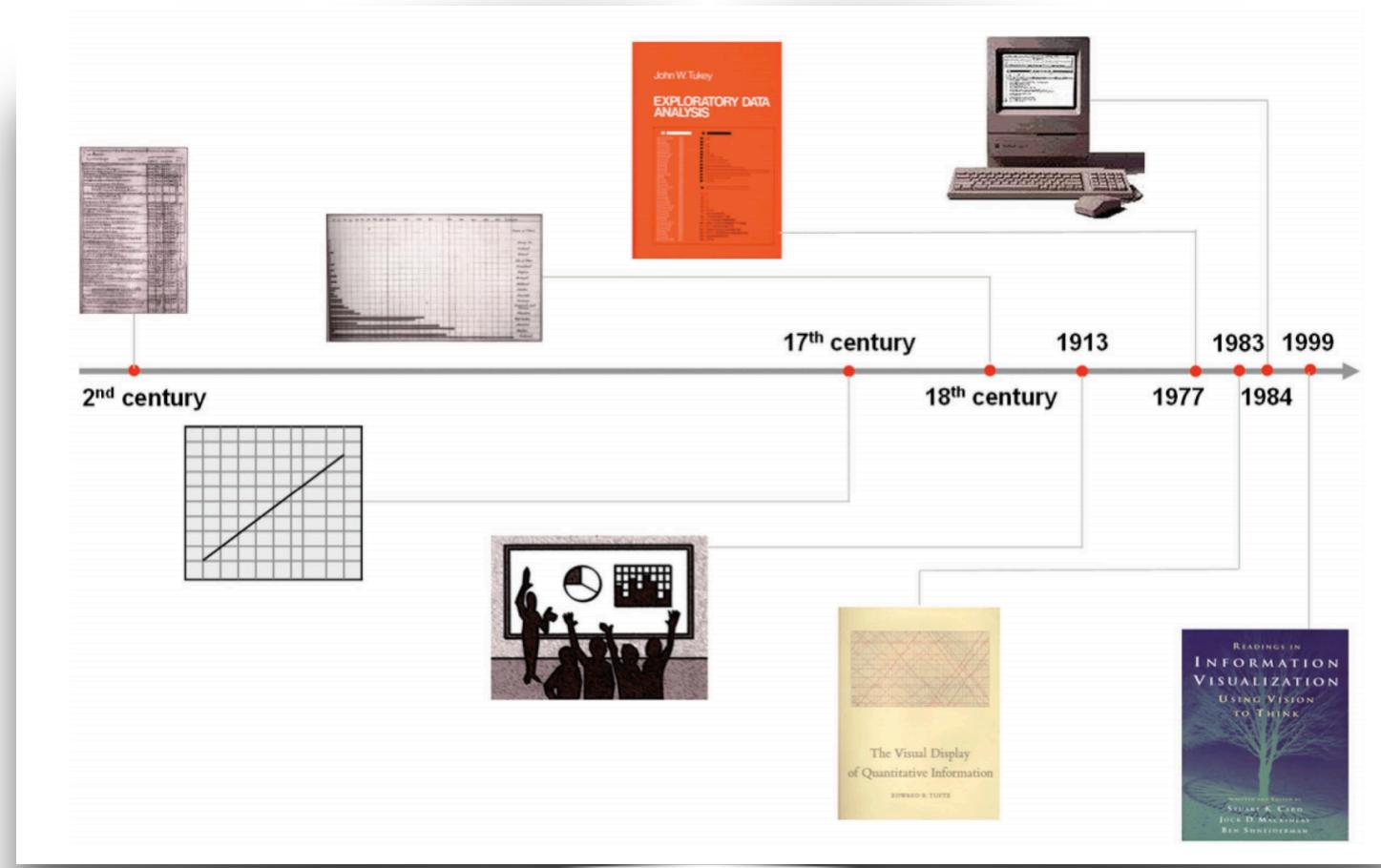
mnist dataset
[LeCun et al., 2013]

60K+10K images

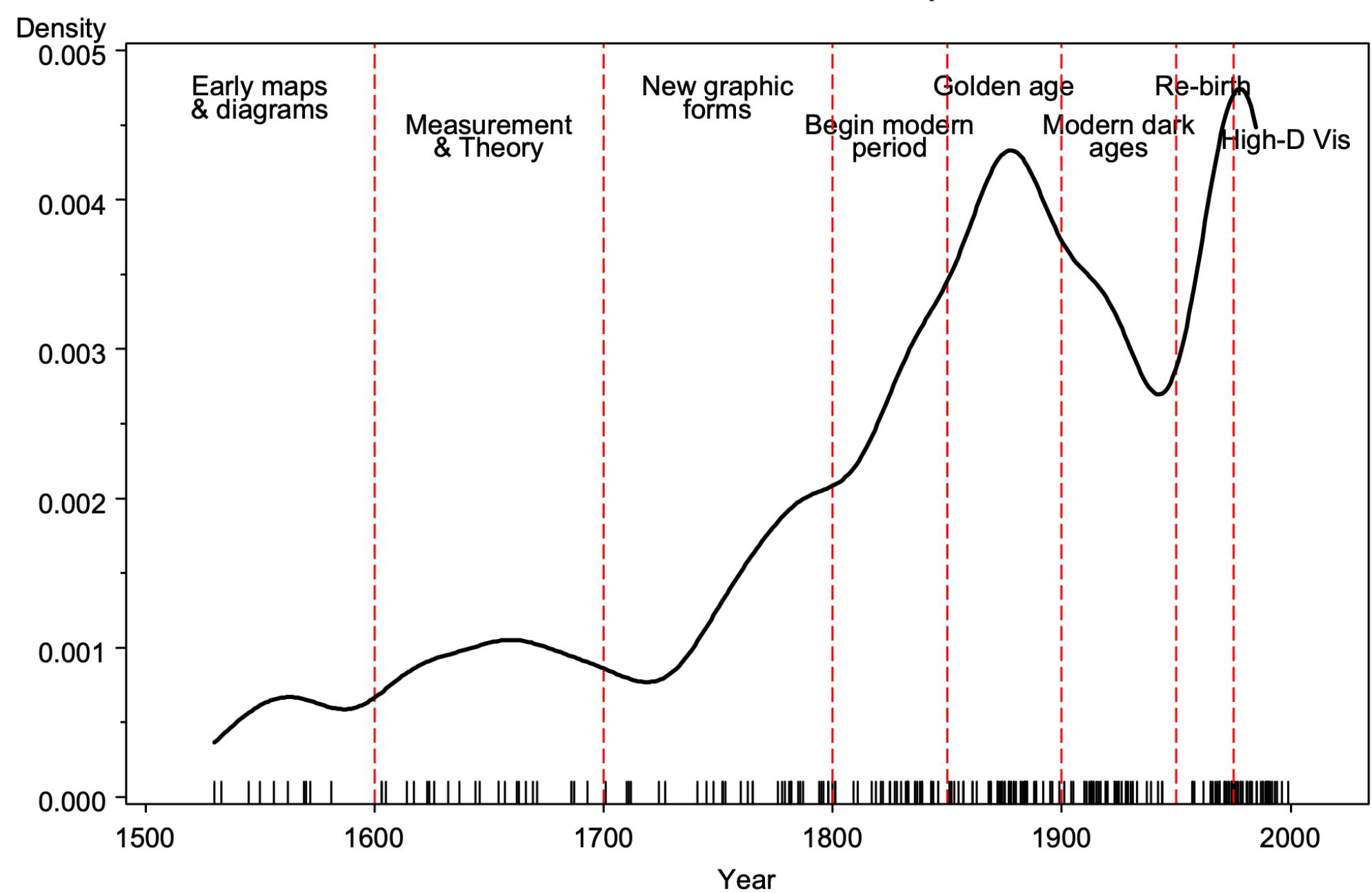


outline

history

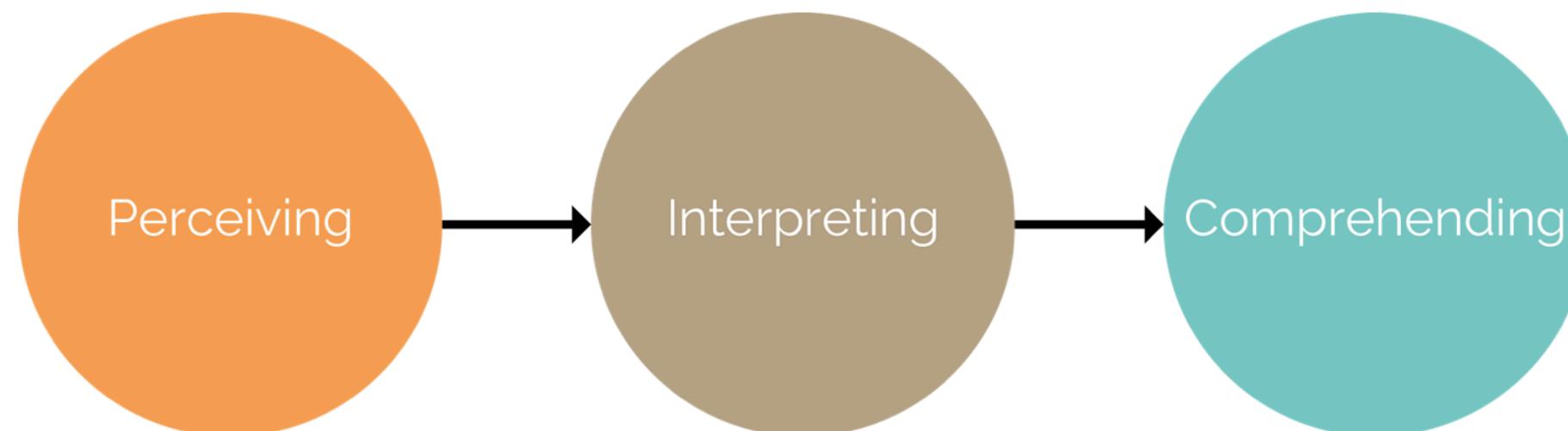


Milestones: Time course of developments



outline

foundations



What does it show?

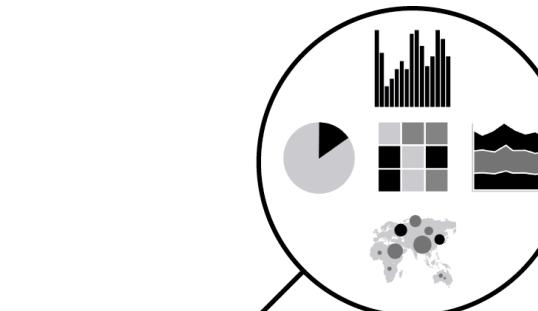
Where is big, medium, small?
How do things compare?
What relationships exist?

What does it mean?

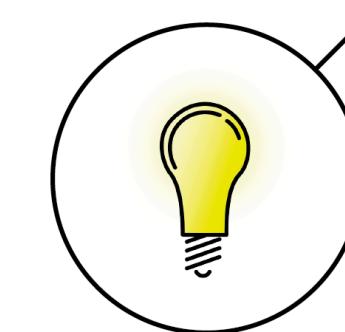
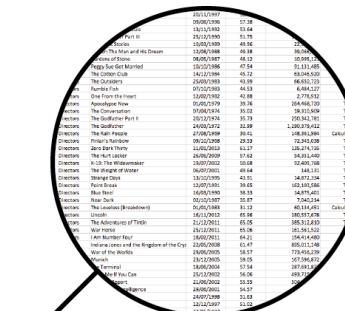
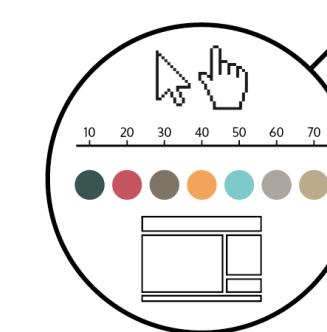
What is good and bad?
Is it meaningful or insignificant?
Unusual or expected?

What does it mean to me?

What are the main messages?
What have I learnt?
Any actions to take?



The representation and presentation of data to facilitate understanding



Principle 1

Good data visualisation
is **TRUSTWORTHY**

Principle 2

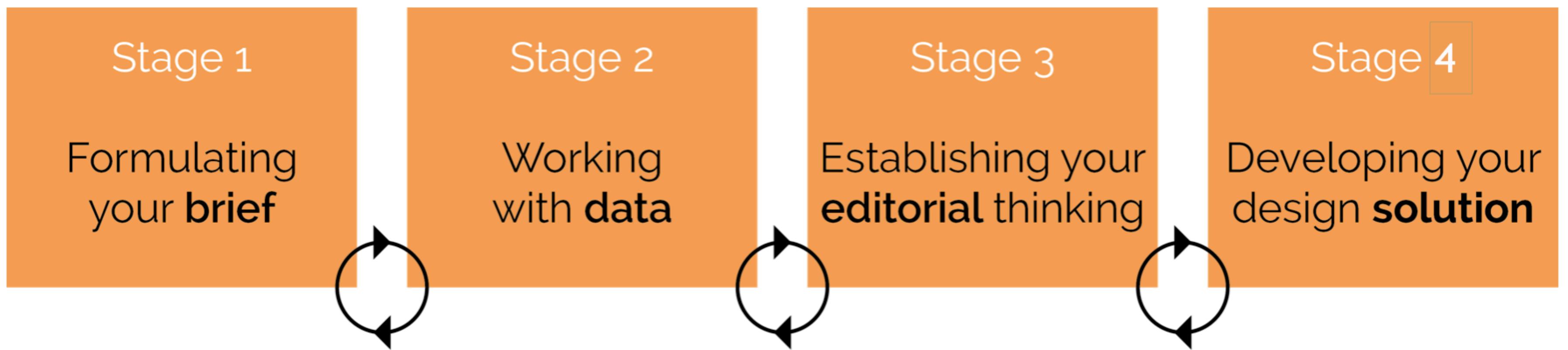
Good data visualisation
is **ACCESSIBLE**

Principle 3

Good data visualisation
is **ELEGANT**

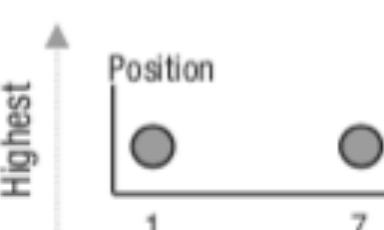
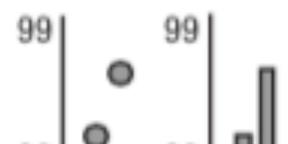
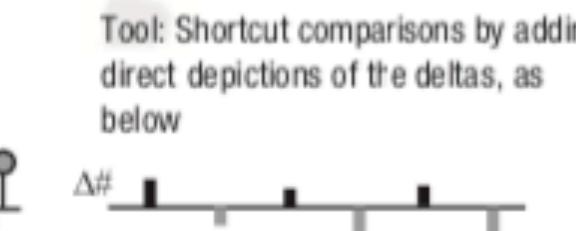
outline

workflow



outline

representation

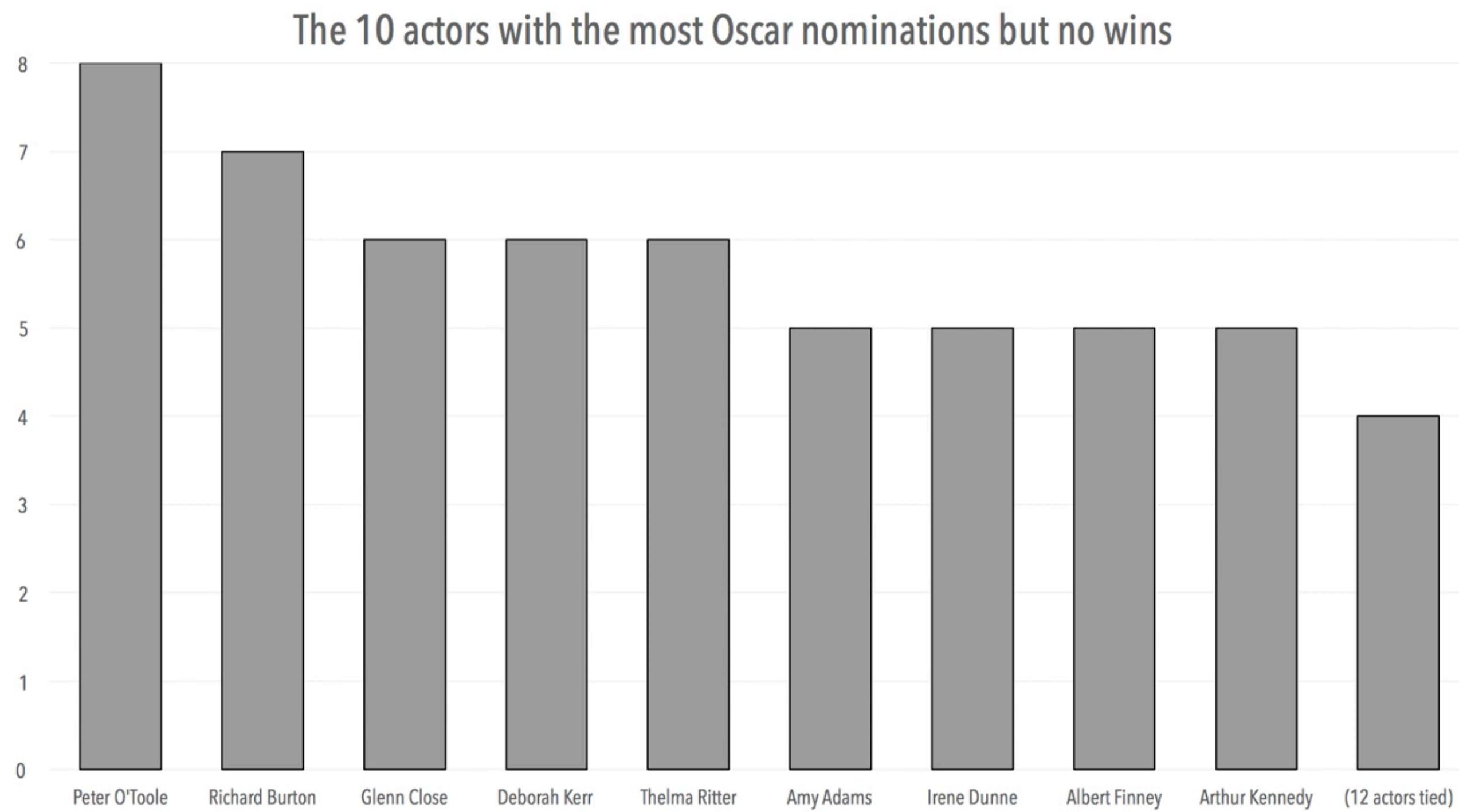
Absolute Precision Ranking for Seeing a Single Ratio	Common Illusions That Distort Data	Vision Is Powerful for Global Statistics	Vision Is Sluggish for Comparisons
Visual estimation of the 1:7 ratio is noisier toward bottom	Caveats for the visual encoding in each row	For each visualization, statistics are available quickly	Isolating pairs with "larger second values" is tough...
 Length: A horizontal bar chart where the length of the bar from position 1 to 7 is approximately 7 times longer than the bar from 1 to 2.	<p>Use caution with nonzero axes: Viewers tend to overestimate differences... even when the nonzero base is marked, as in the examples at left.</p>  Area: A bubble map showing two bubbles at areas 1 and 7. The y-axis is labeled 'Highest' at the top and 'Lowest' at the bottom.	Dot Plot: Shows data points with vertical error bars indicating range. Statistics: Max height, Mean height, Min height.	So guide viewers to the right comparisons
Angle: An angle graph showing two segments at angles 1 and 7. The y-axis is labeled 'Highest' at the top and 'Lowest' at the bottom.	<p>The difference is larger for the lighter segments compared with the darker ones, right?</p> <p>That is an illusion—the differences are identical.</p> 	Stacked Bar: Shows stacked bars for categories a-f. Statistics: Min, Max, Mean length of dark bars.	 Slope Graph: Shows lines with different slopes for categories a-f. Statistics: Max, Mean Angle, Min.
Intensity: A color bar showing intensity values from 1 to 7. The y-axis is labeled 'Highest' at the top and 'Lowest' at the bottom.	<p>Intensity values can look different depending their backgrounds.</p> <p>Do not plot intensities on intensities.</p>	 Mean Intensity: A color bar showing mean intensity levels from Min to Max.	 "a, c, & e have increased"

zoo

charts

bar chart

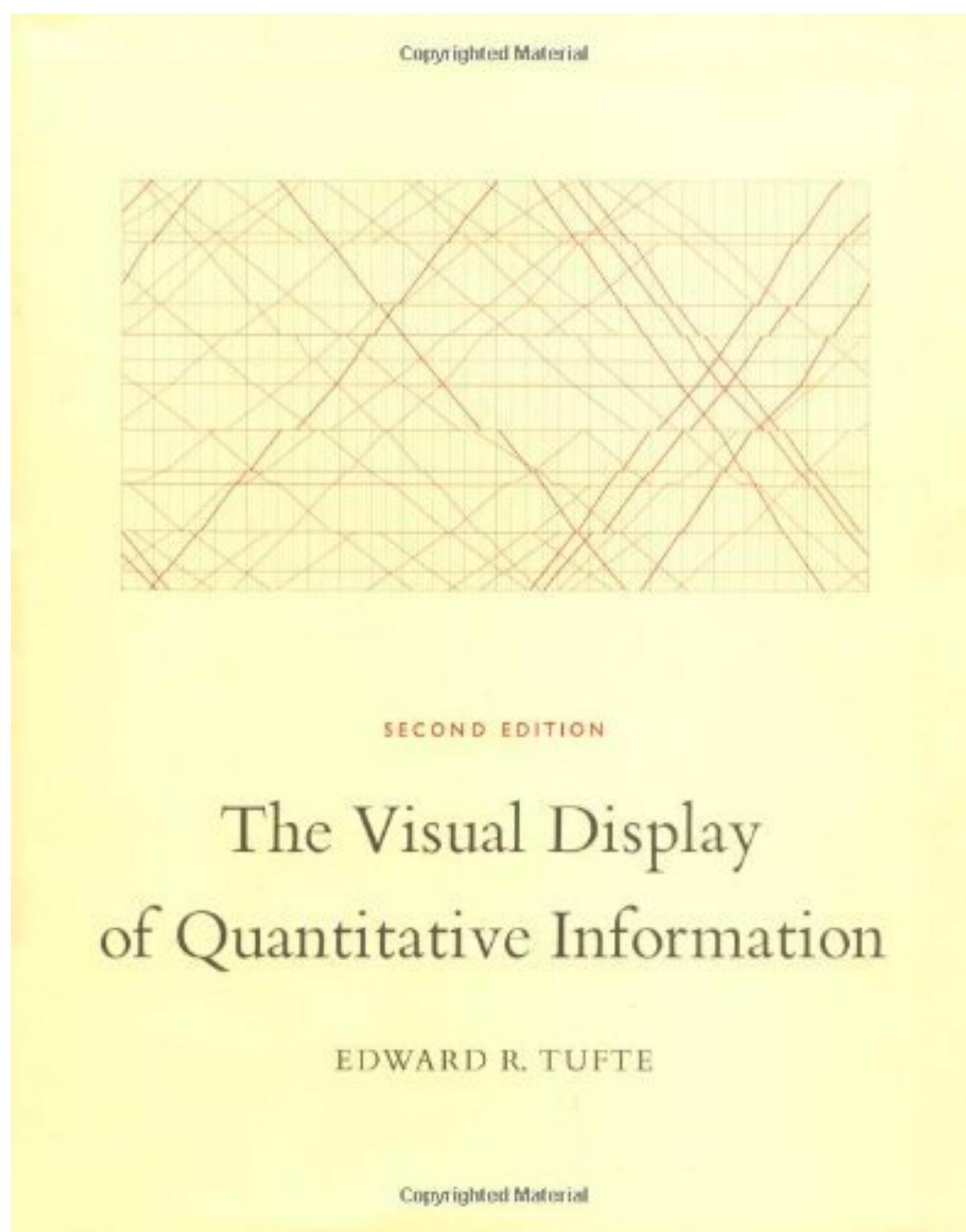
column chart — histogram



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

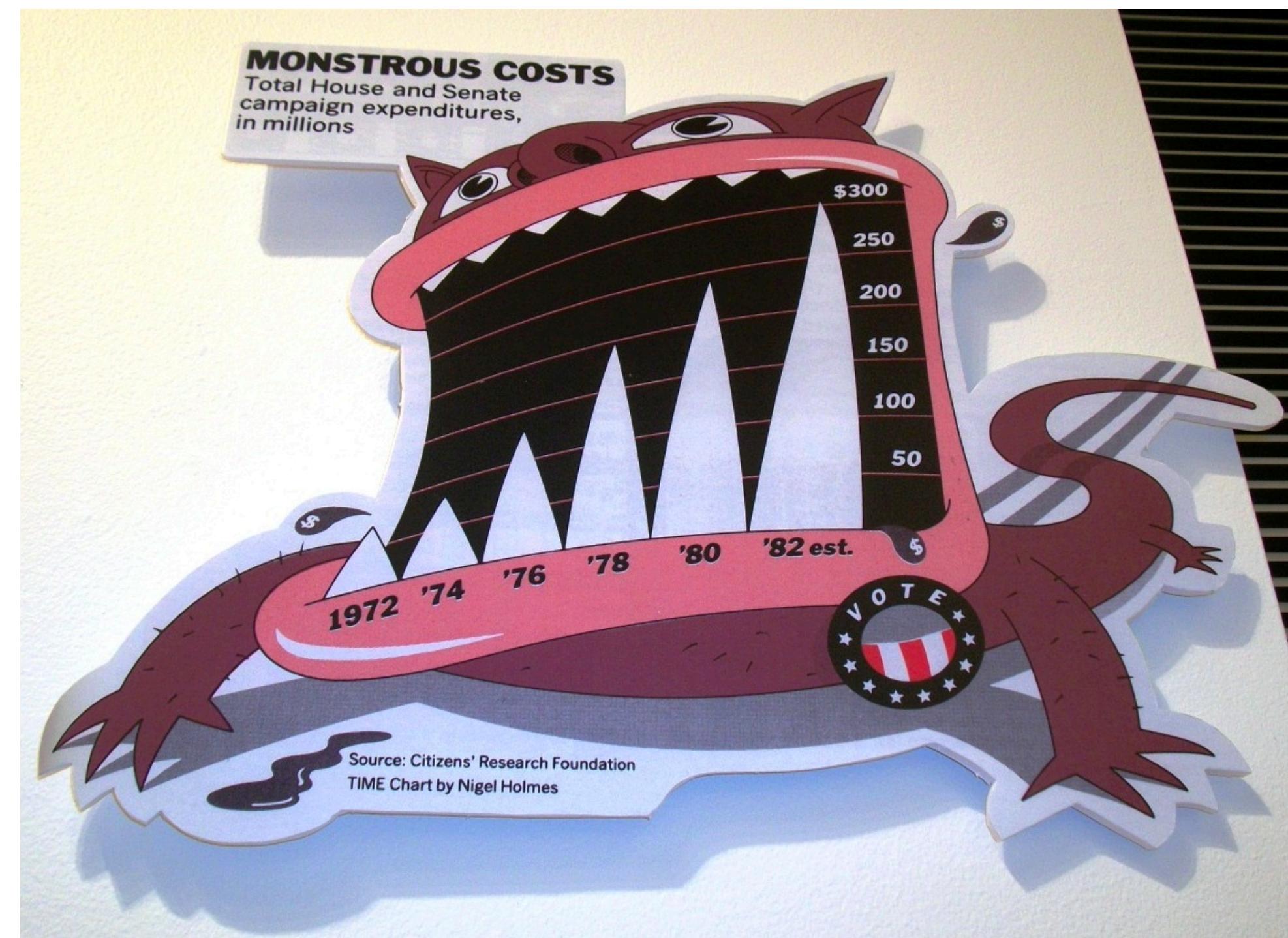
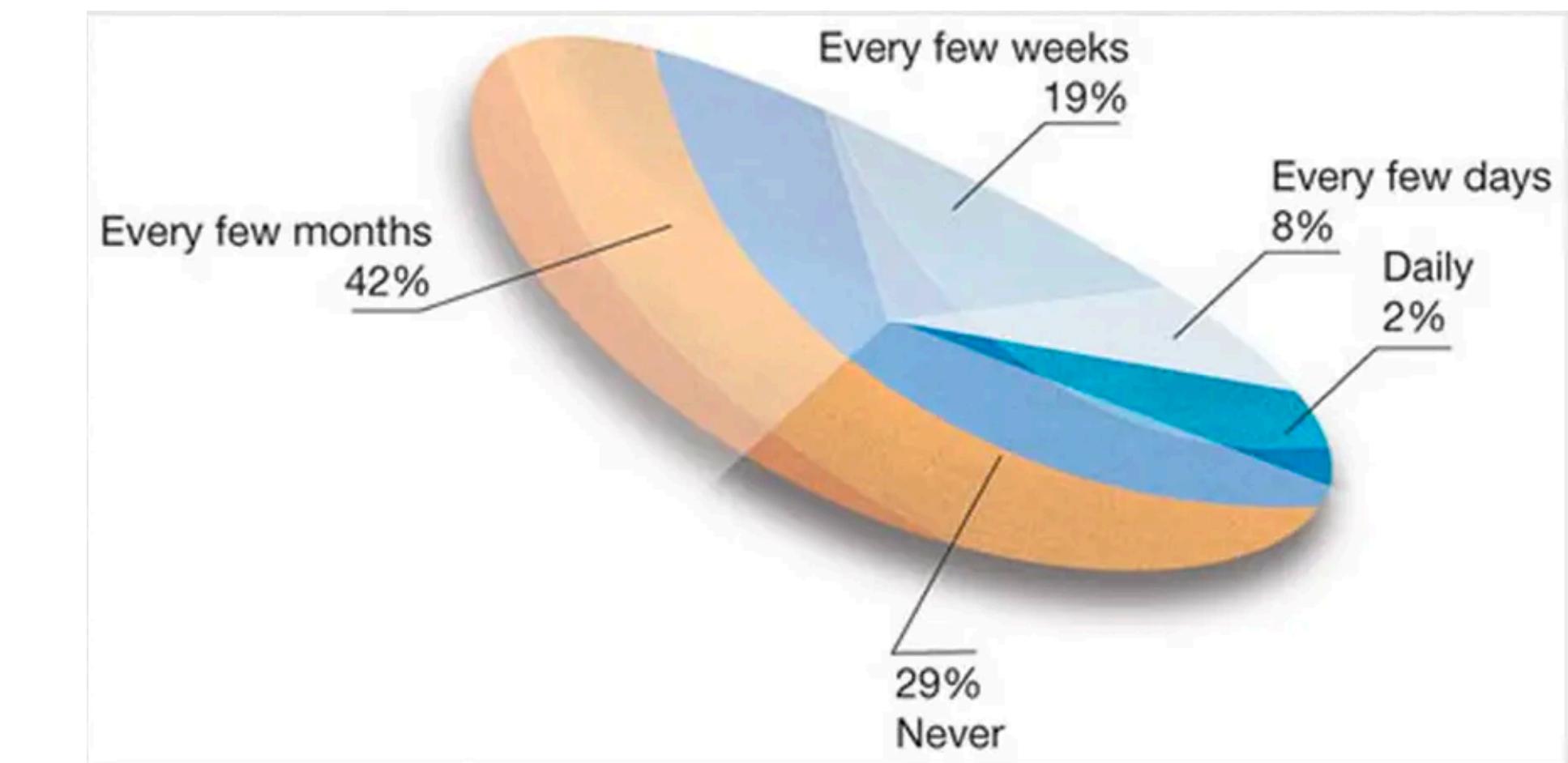
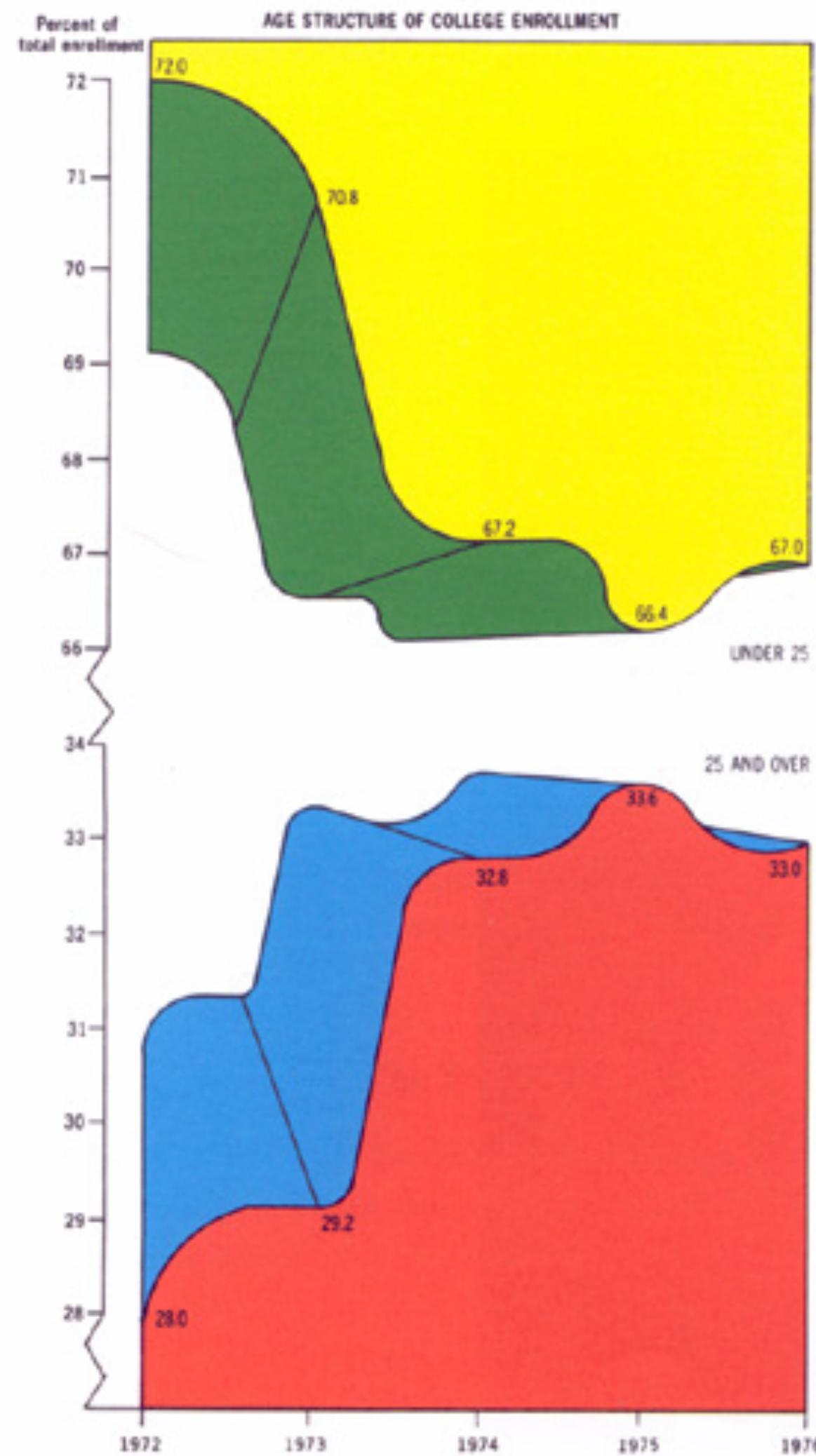


Tufte's Principles of Graphical Excellence:

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*

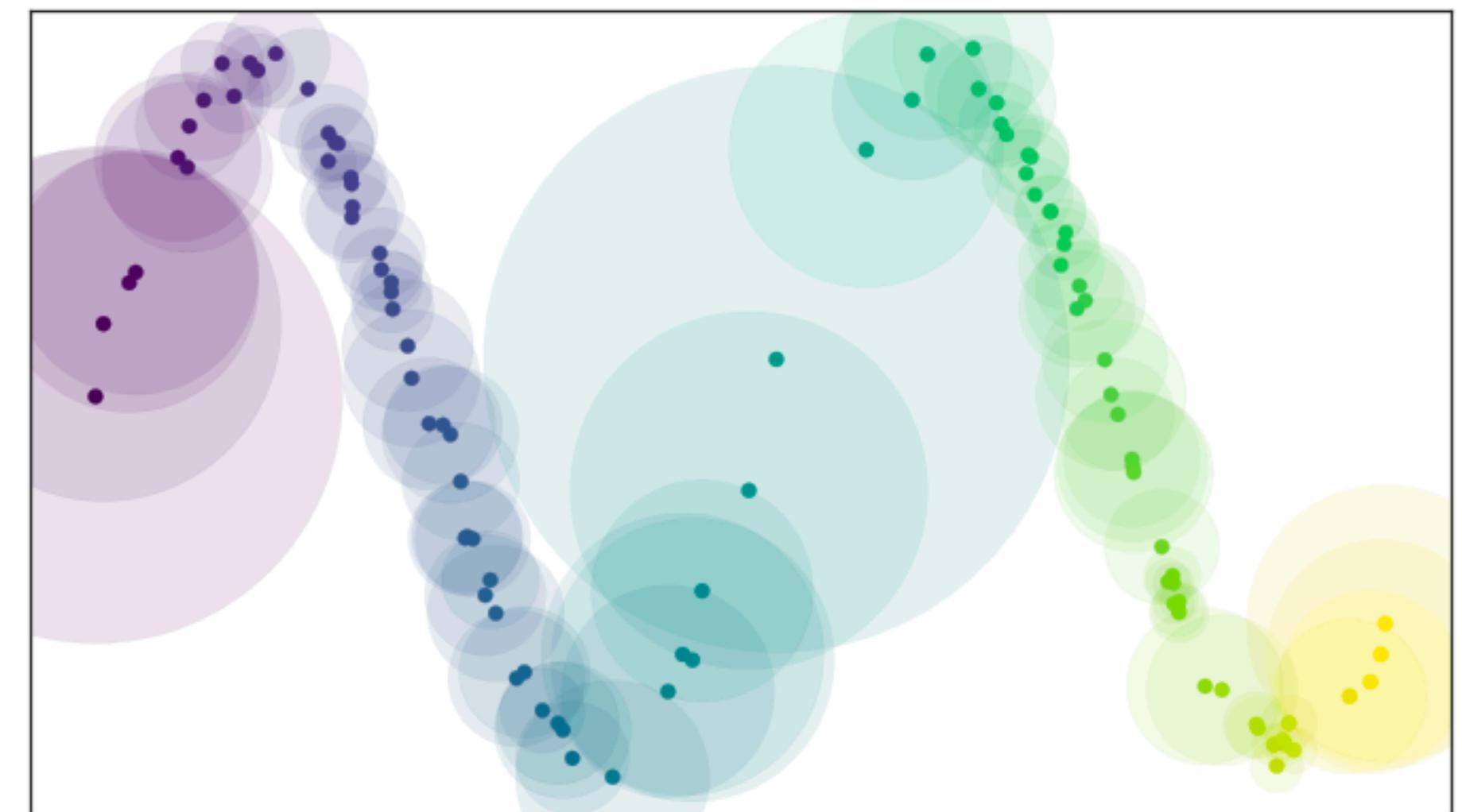
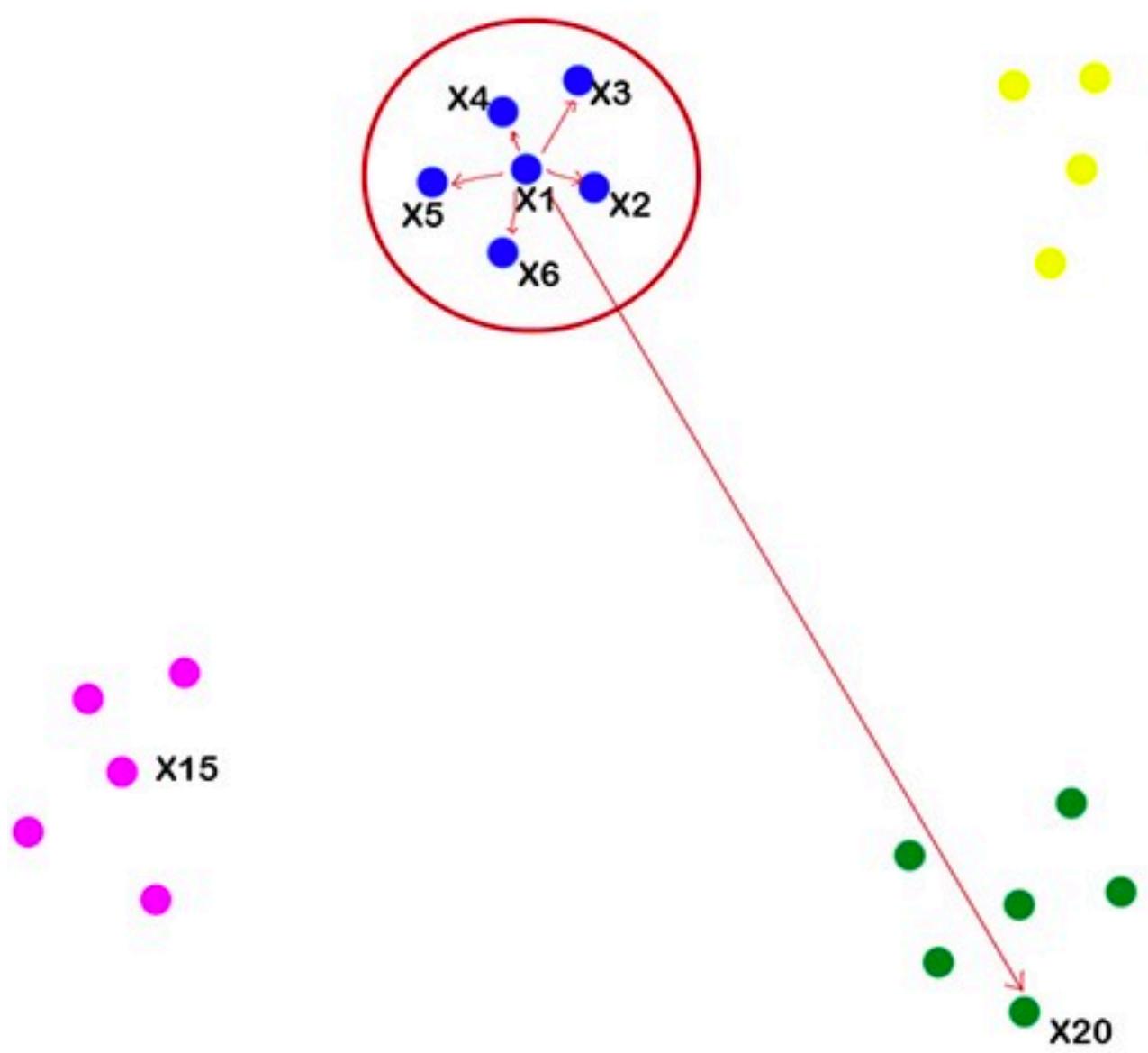
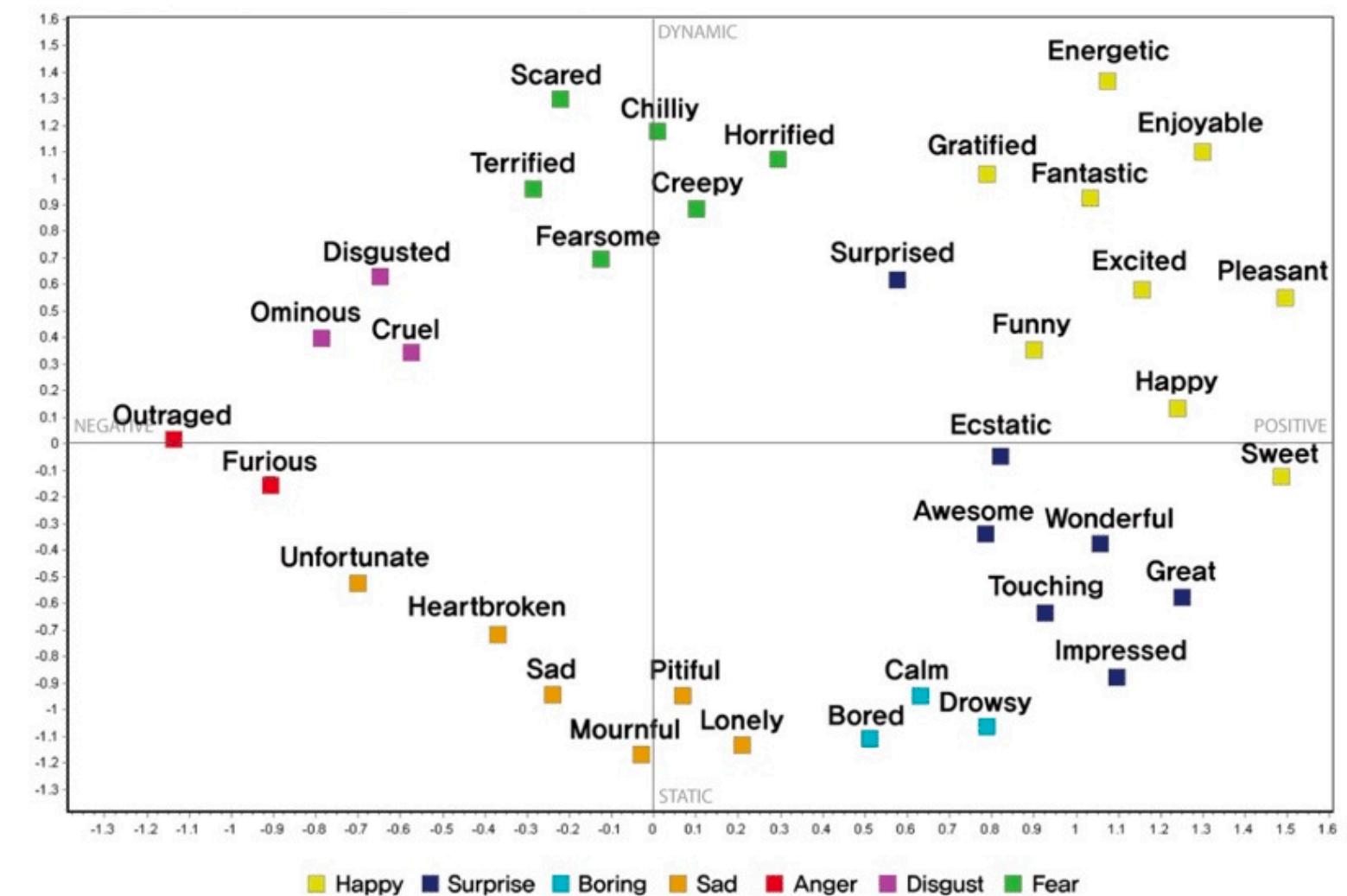
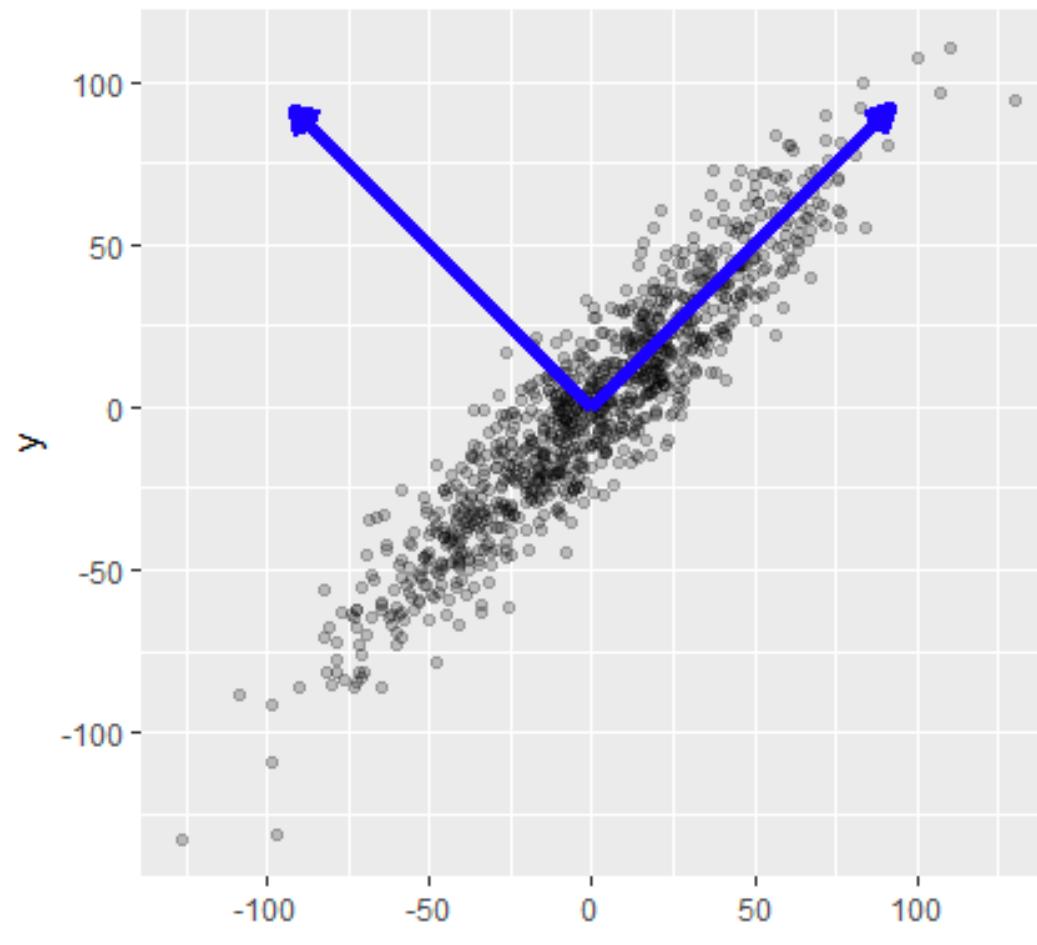
outline

horror gallery

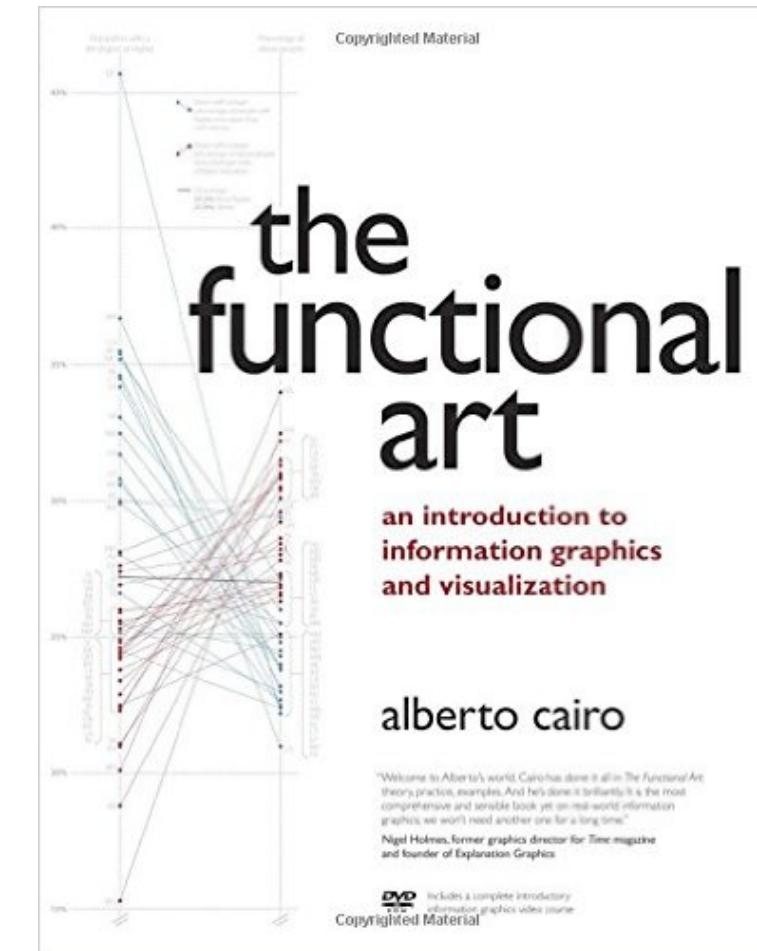
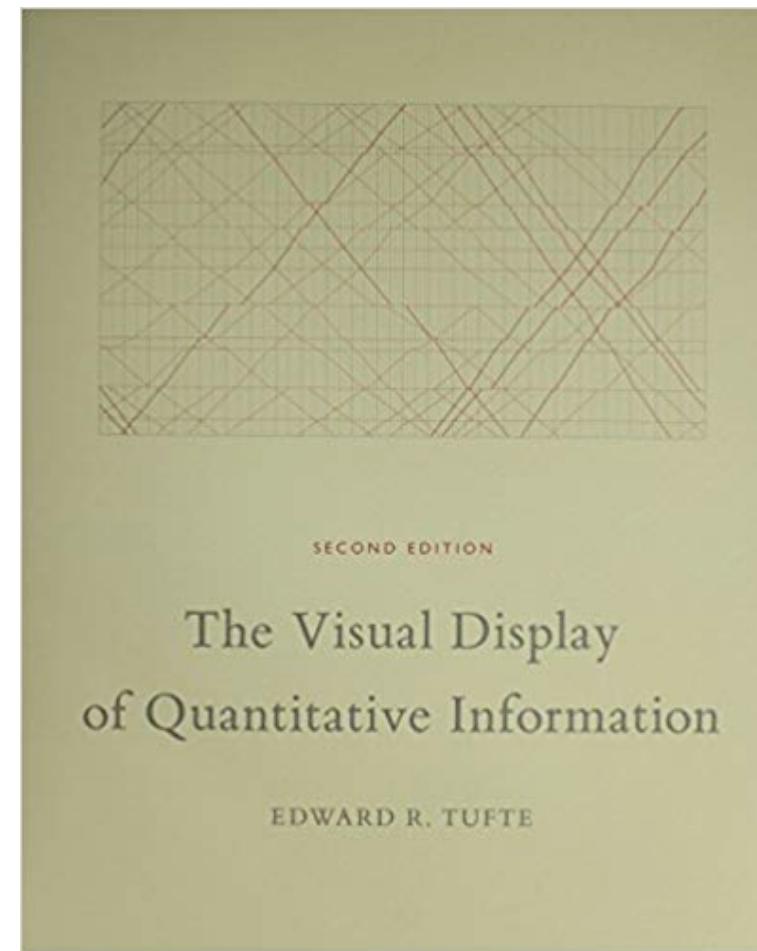
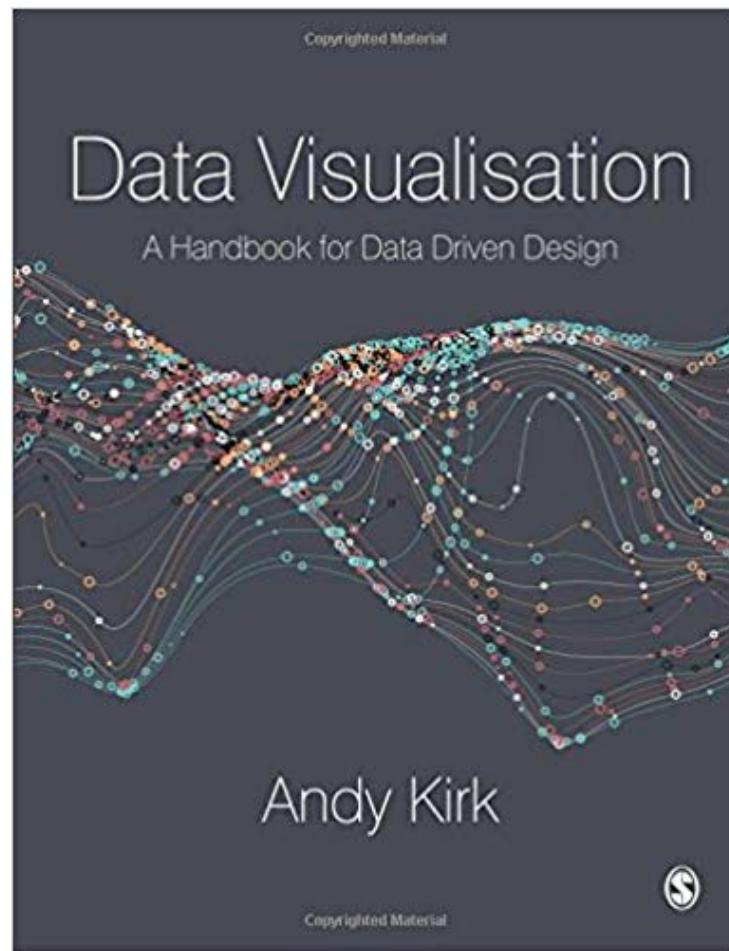


outline

a case study: dimensionality reduction for data viz



how/1



From Data to Viz

EXPLORE STORY ALL CAVEATS POSTER ABOUT CONTACT

A WORLD OF POSSIBILITIES

Here is an overview of all the graph types presented in this website.

Show all Distribution Correlation Ranking Part of a whole Evolution Map Flow

Venn diagram Density Histogram Boxplot Ridgeplot Scatter

Heatmap Correlogram Bubble Connected scatter Density 2d Barplot

Spider / Radar Wordcloud Parallel Lollipop Circular Barplot Treemap

Ven diagram Donut chart Pie chart Dendrogram Circular packing Sunburst

Line plot Area Stacked area Streamchart Map

Haben map Cartogram Connection Bubble map Chord diagram

Sectary Art diagram Edge bonding

LIZ STIBORI DESIGN - 02.15.17 - 12:44 PM

THE NERDY CHARM OF ARTISANAL, HAND-DRAWN INFOGRAPHICS

WEN 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, styl, and tablets to craft glossy data visualizations—and plenty of them do. But many still prefer simple tools—and use them to fantastic effect.

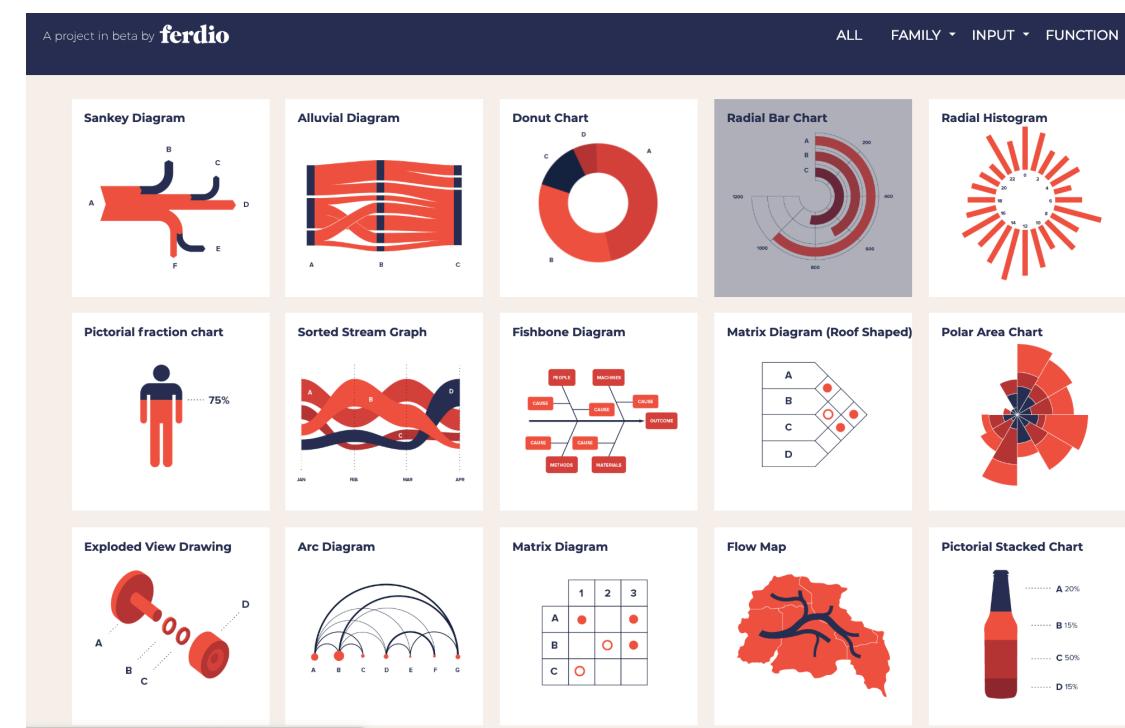
Giorgia Lupi and Stefanie Posavec

In a time where slick, computer-made visualizations dominate, hand-drawn infographics prove analog illustrations can be just as compelling. In 2015, data visualization designers Giorgia Lupi and Stefanie Posavec started a project called *Dear Data*, for which the designers spent every week tracking a bit of personal data—moments of joy, stress, and sadness, for example—they laughed—and translating it into infographics. Those infographics went on postcards, which they mailed to their friends and family. The postcards described the postcards as “giant data,” which holds not only at their quasi-quantitative nature, but their imperfections. They’re not perfect, and neither are the drawings, but their imperfections lend the visualizations a more friendly, personal, and evocative feel.

EVERYONE SECRETLY STORES DATA (EVEN IF THEY DON'T ADMIT TO IT)

Giorgia Lupi and Stefanie Posavec

STEFANI POSAVEC



IEEE computer society

The Community for Technology Leaders

Libraries & Institutions About Resources Subscribe

CSDL Home > IEEE Transactions on Visualization & Computer Graphics > 2018 vol. 24 > Issue No. 08 - Aug.

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS



Information Visualization

Volume 18 Number 1 January 2019 journals.sagepub.com/home/iv

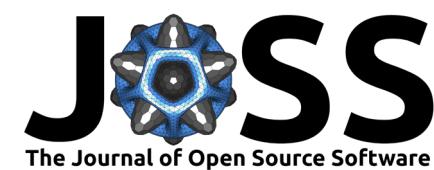
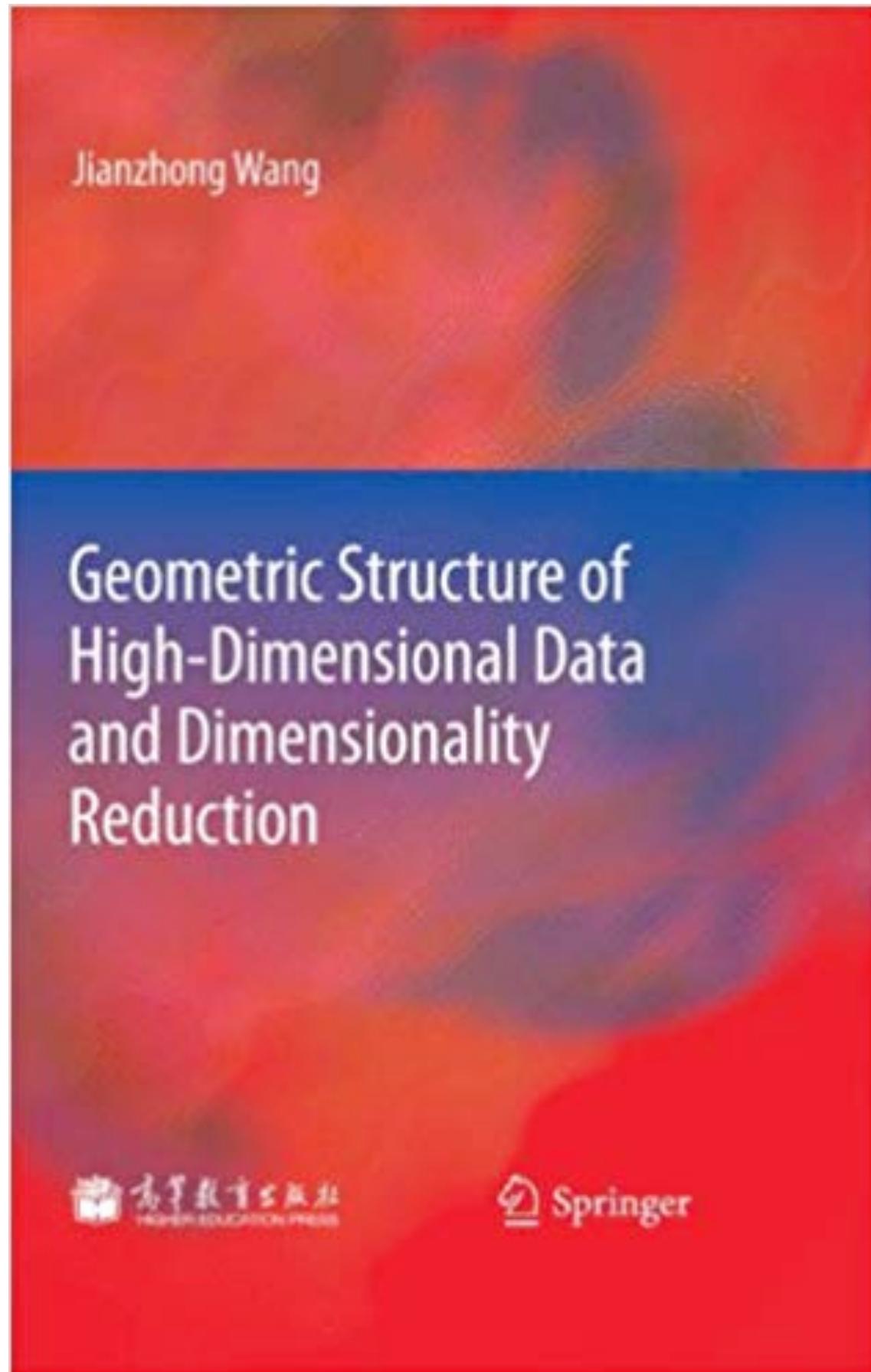


Springer

Journal of
Visualization

The Visualization Society of Japan

how/2



UMAP: Uniform Manifold Approximation and Projection

Leland McInnes¹, John Healy¹, Nathaniel Saul², 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

MICC-IKAT

Maastricht University

P.O. Box 616, 6200 MD Maastricht, The Netherlands

L.VANDERMAATEN@MICC.UNIMAAS.NL

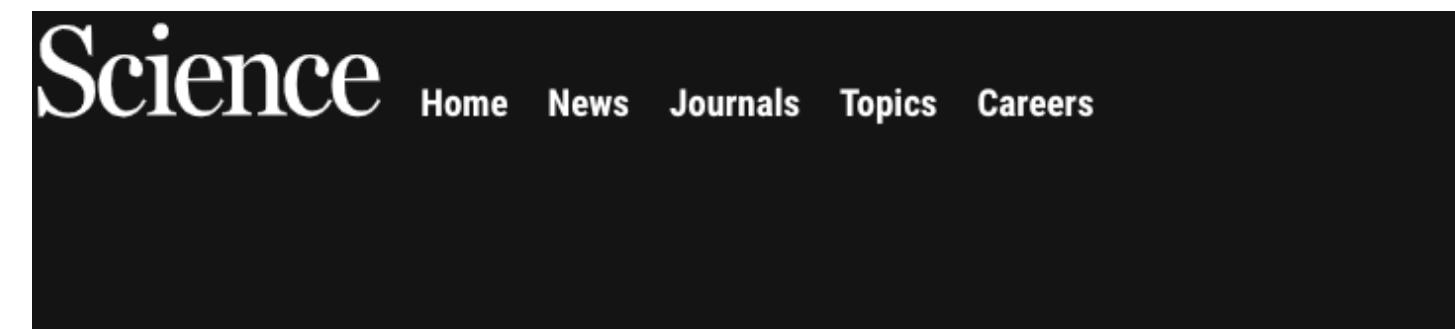
Geoffrey Hinton

Department of Computer Science

University of Toronto

6 King's College Road, M5S 3G4 Toronto, ON, Canada

HINTON@CS.TORONTO.EDU



SHARE REPORT



0



0



0

A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum^{1,*}, Vin de Silva², John C. Langford³

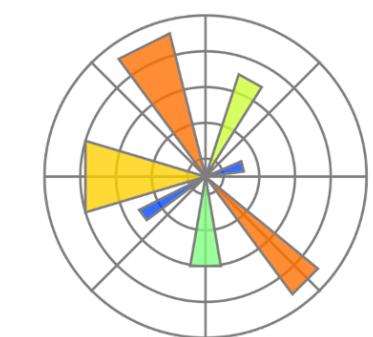
* See all authors and affiliations

Science 22 Dec 2000:
Vol. 290, Issue 5500, pp. 2319-2323
DOI: 10.1126/science.290.5500.2319

how/3



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



matplotlib



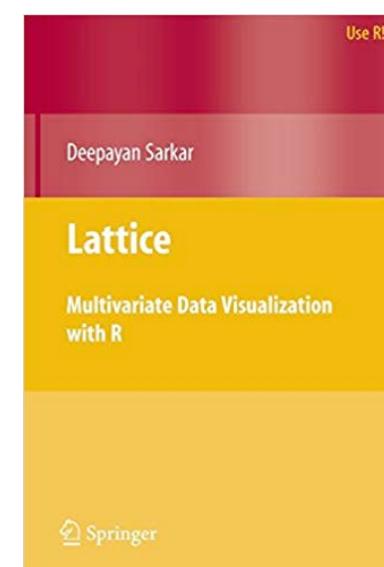
Bokeh



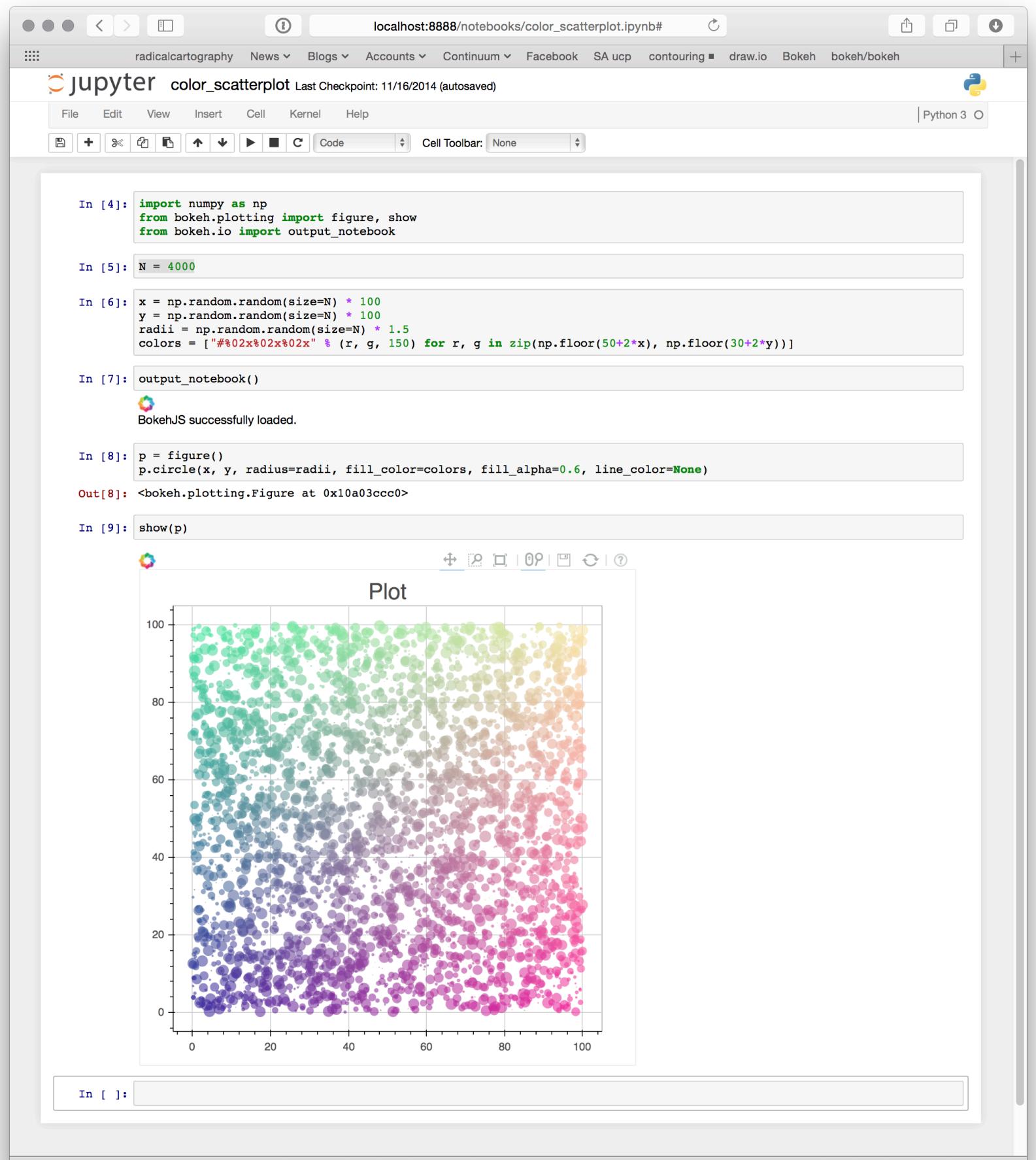
PyViz



ggplot2



how/3



A screenshot of a Jupyter Notebook interface. The top bar shows the URL "localhost:8888/notebooks/color_scatterplot.ipynb#". The menu bar includes File, Edit, View, Insert, Cell, Kernel, Help, and Cell Toolbar (None). A Python 3 logo is in the top right. The code cell In [4] contains imports for numpy, bokeh.plotting, and bokeh.io. In [5] defines N = 4000. In [6] generates random data for x, y, radii, and colors. In [7] calls output_notebook(). In [8] creates a Bokeh Figure p with circles. In [9] shows(p) displays the plot. The plot shows a grid of colored circles from purple to yellow, with axes ranging from 0 to 100.



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

