CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

An introduction to visualization grammars for exploratory data analysis in the Humanities: a modern perspective

Alejandro Benito-Santos (abenito@usal.es)
Margarita Salas Postdoc – Spanish Ministry of Universities
Universidad de Salamanca (Spain)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

CLSINFRA

HANDS-ON

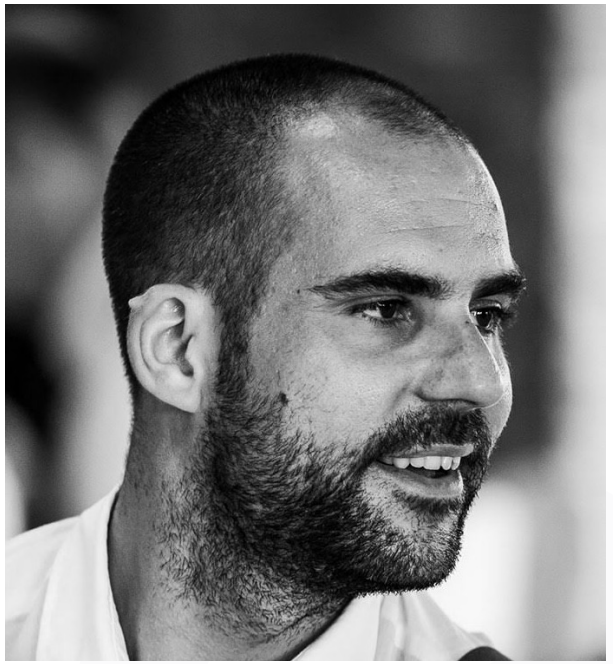
https://bit.ly/clsinfra_vis

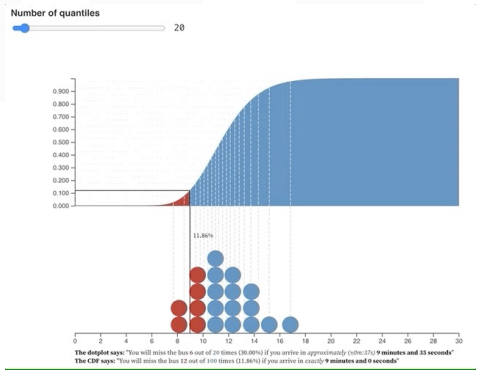
Outline

- Definitions:
  - visualization
  - exploratory data analysis
  - visual analytics
  - text visual analytics
- Why visualization is important
- Why you should always visualize your data
  - Visualization is not just an output, but a tool to support reasoning with data.
- How visualization and visual analysis grammars help us design effective visualizations
- Why interaction is important in data analysis
- How interactive visualizations are designed
- Hands-on session using an example typical of DH practice
  - Text-based visualization
  - Explore the output of an algorithm

About me

- Ph.D. Computer Engineering (2020)
- Text processing and representation, recommender systems, linguistics, NLP, digital humanities
- Currently in 1st year (out of 2) of a ”Margarita Salas” postdoc position funded by the Spanish Ministry of Universities
- More at https://alexbs.me





What is visualization?

CLSINFRA

- “Transformation of the symbolic into the geometric” [McCormick et al. 1987]
- “... finding the artificial memory that best supports our natural means of perception.” [Bertin 1967]
- “The use of computer-generated, interactive, visual representations of data to amplify cognition.” [Card, Mackinlay, & Shneiderman 1999]
- "Creating a visualization means encoding abstract data into visual channels with the goal of amplifying the user's cognitive capabilities.” [Moritz, 2022]

Why create visualizations?

CLSINFRA

- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support graphical calculation
- Find patterns
- Present an argument or tell a story
- Inspire

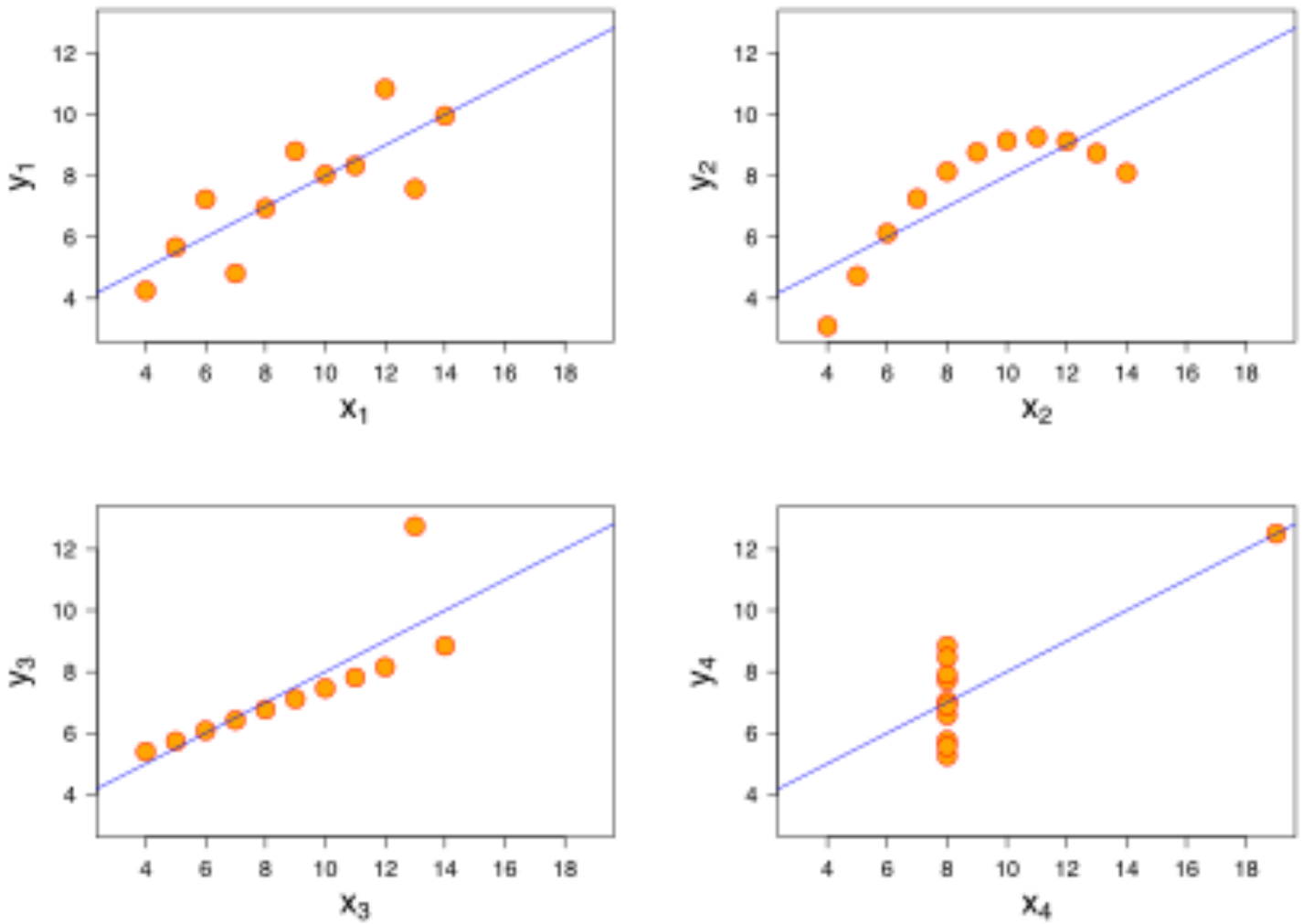
Why Data Visualization?

90% of information transmitted to the human brain is visual

The human brain processes visuals 60K times faster than text

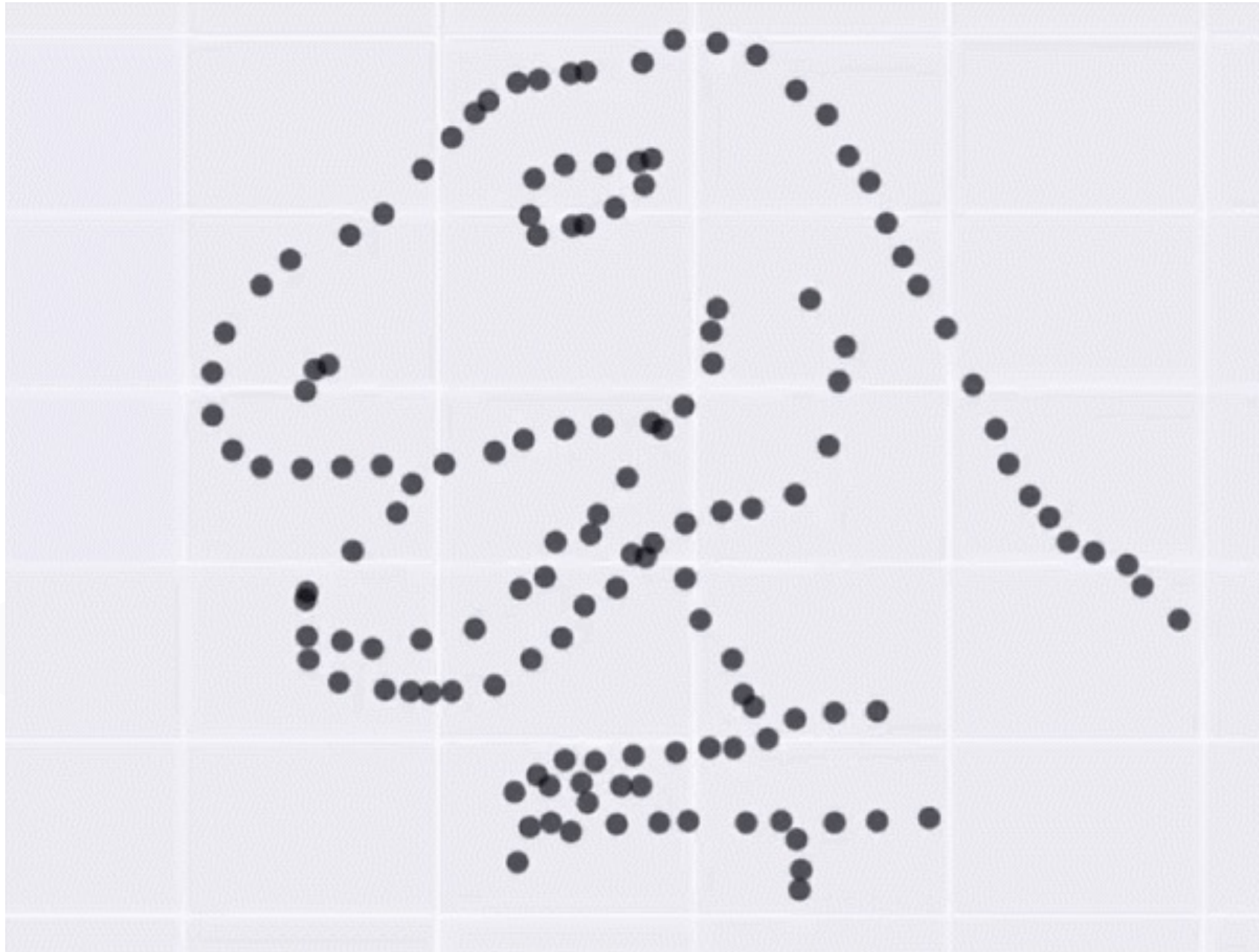
65% of humans are visual learners

The human brain can process an observed visual in 13 - 80 ms



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

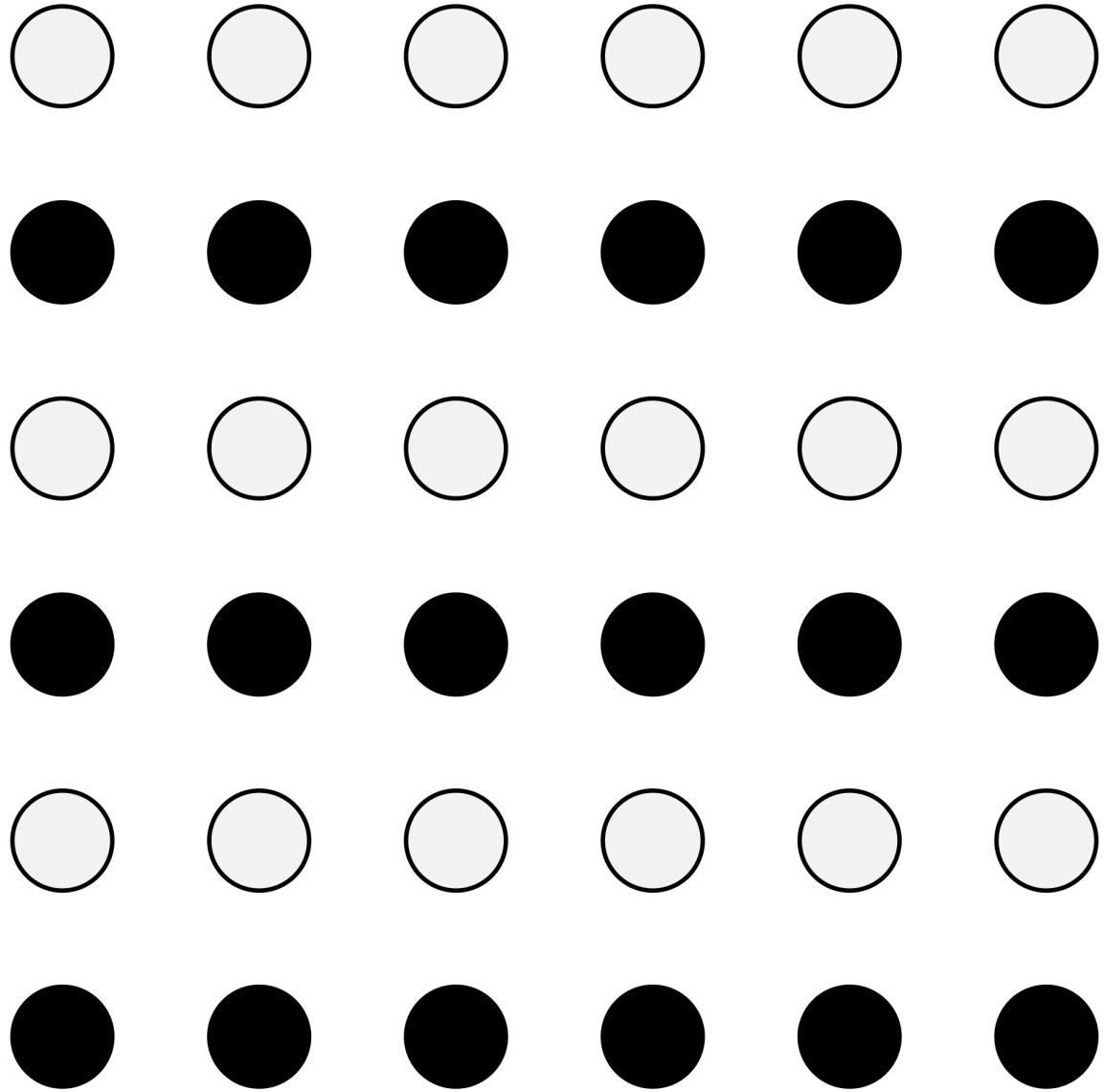
Anscombe's Quartet

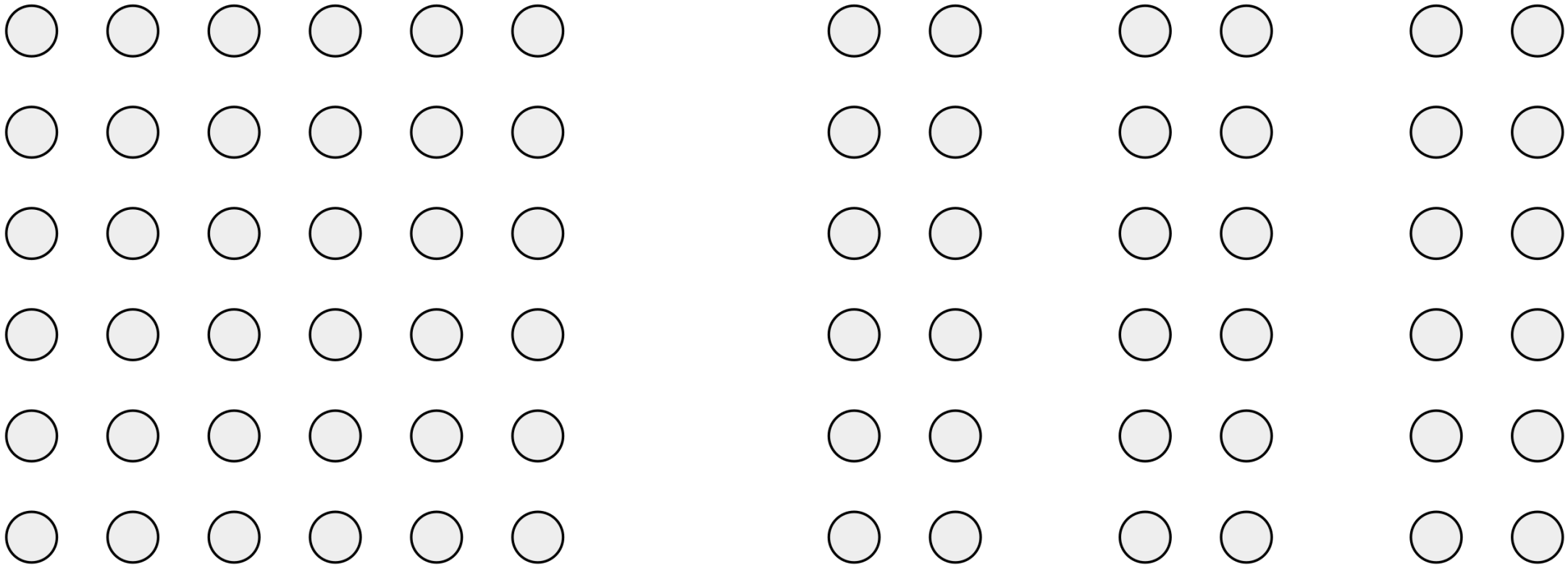


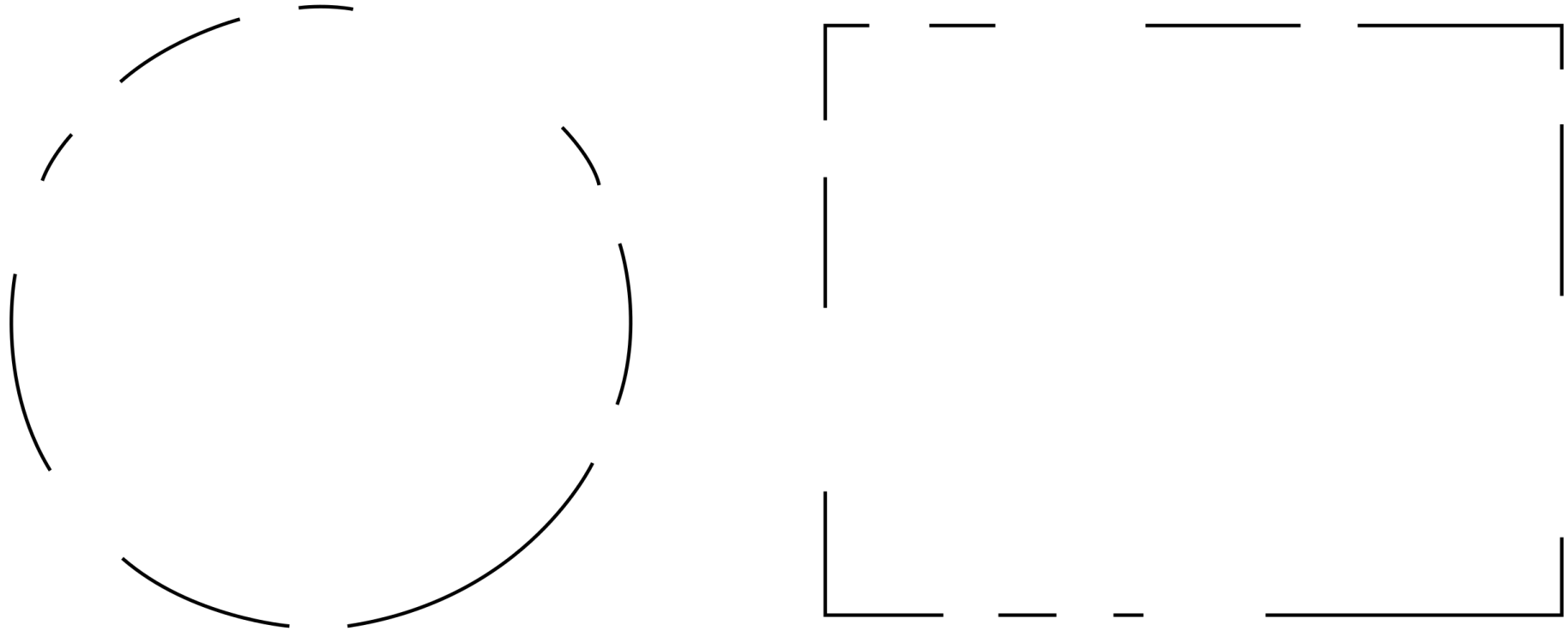
CLSINFRA

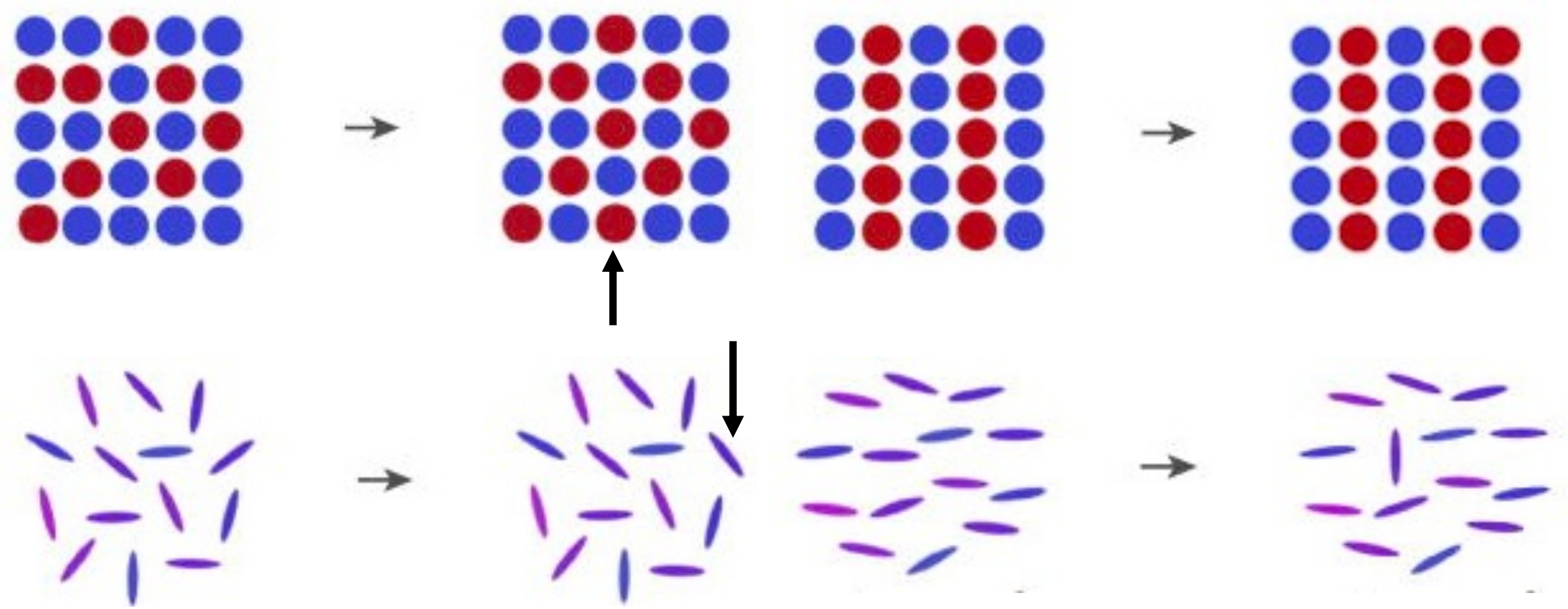
X Mean	:	54.26
Y Mean	:	47.83
X SD	:	16.76
Y SD	:	26.93
Corr.	:	-0.06

Alberto Cairo – The Datasaurus Dozen









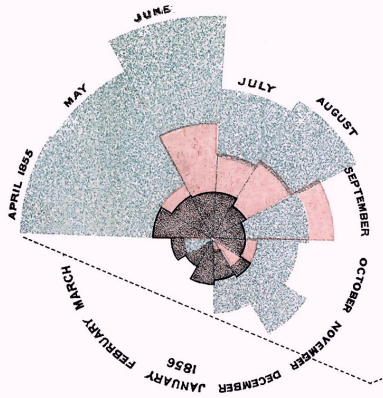
M. A. Cohen, D. C. Dennett, and N. Kanwisher, ‘What is the Bandwidth of Perceptual Experience?’, *Trends Cogn Sci*, vol. 20, no. 5, pp. 324–335, May 2016, doi: 10.1016/j.tics.2016.03.006.

Motivation for effective data visualization

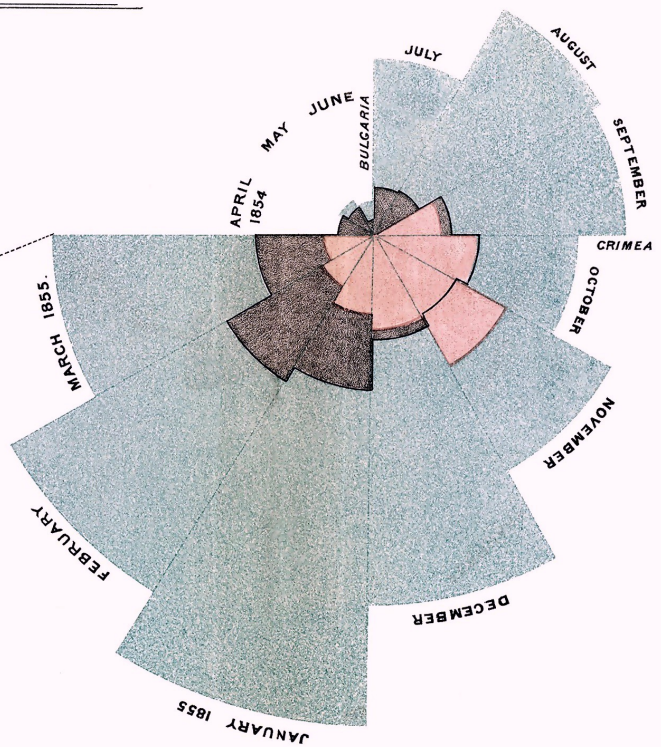
1. Effective Data Visualization is an art as well as a science
2. Focus should be on abstracting out unnecessary data, noise and clutter
3. Leverage concepts from the Grammar of Graphics to depict the right information using clean and concise visuals
4. “A Picture is worth a thousand words”
5. “The greatest value of a picture is when it forces us to notice what we never expected to see.” — John Tukey

DIAGRAM of the CAUSES of MORTALITY
IN THE ARMY IN THE EAST.

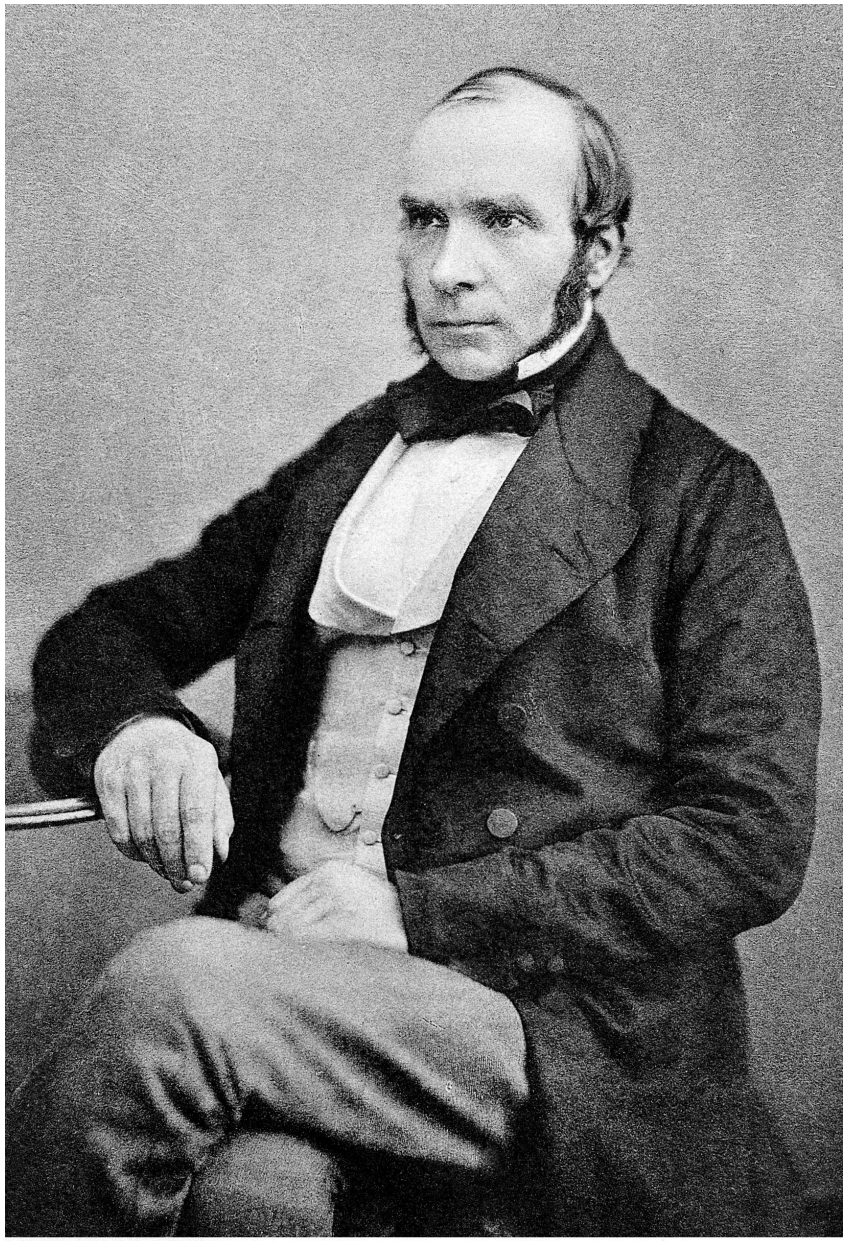
2.
APRIL 1855 to MARCH 1856.



1.
APRIL 1854 to MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.




OXFORD
GT. MARLBOROUGH ST.
REGENT
CONDUIT STREET
KING STREET
BROAD STREET
WARDOUR STREET
DEAN STREET
GOLDEN SQUARE
SAVILLE ROW
BREWER STREET

Exploratory Data Analysis

CLSINFRA

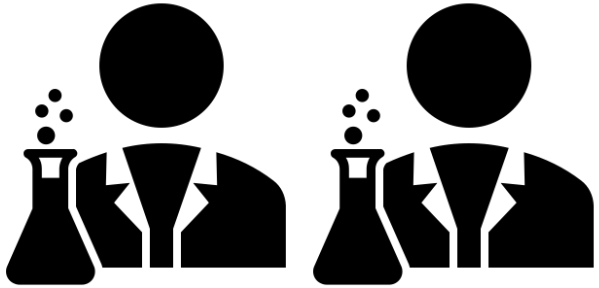
- Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data to
  - discover patterns,
  - spot anomalies,
  - test hypotheses, and
  - check assumptions
  - with the help of summary statistics **and graphical representations.**

- It is a good practice to understand the data first and try to gather as many insights from it.

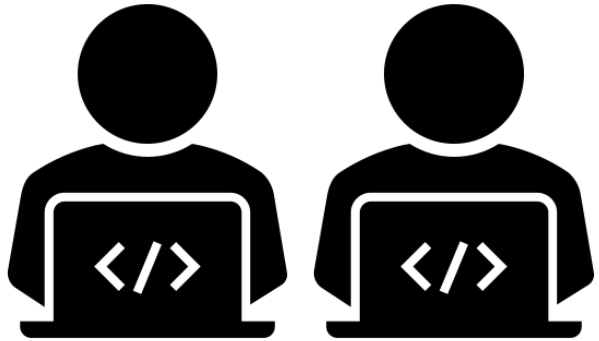
Problem-Driven Visualization Research

- Data
- Driving problems
- Tasks

- Algorithms
- Analysis Techniques
- Visualizations

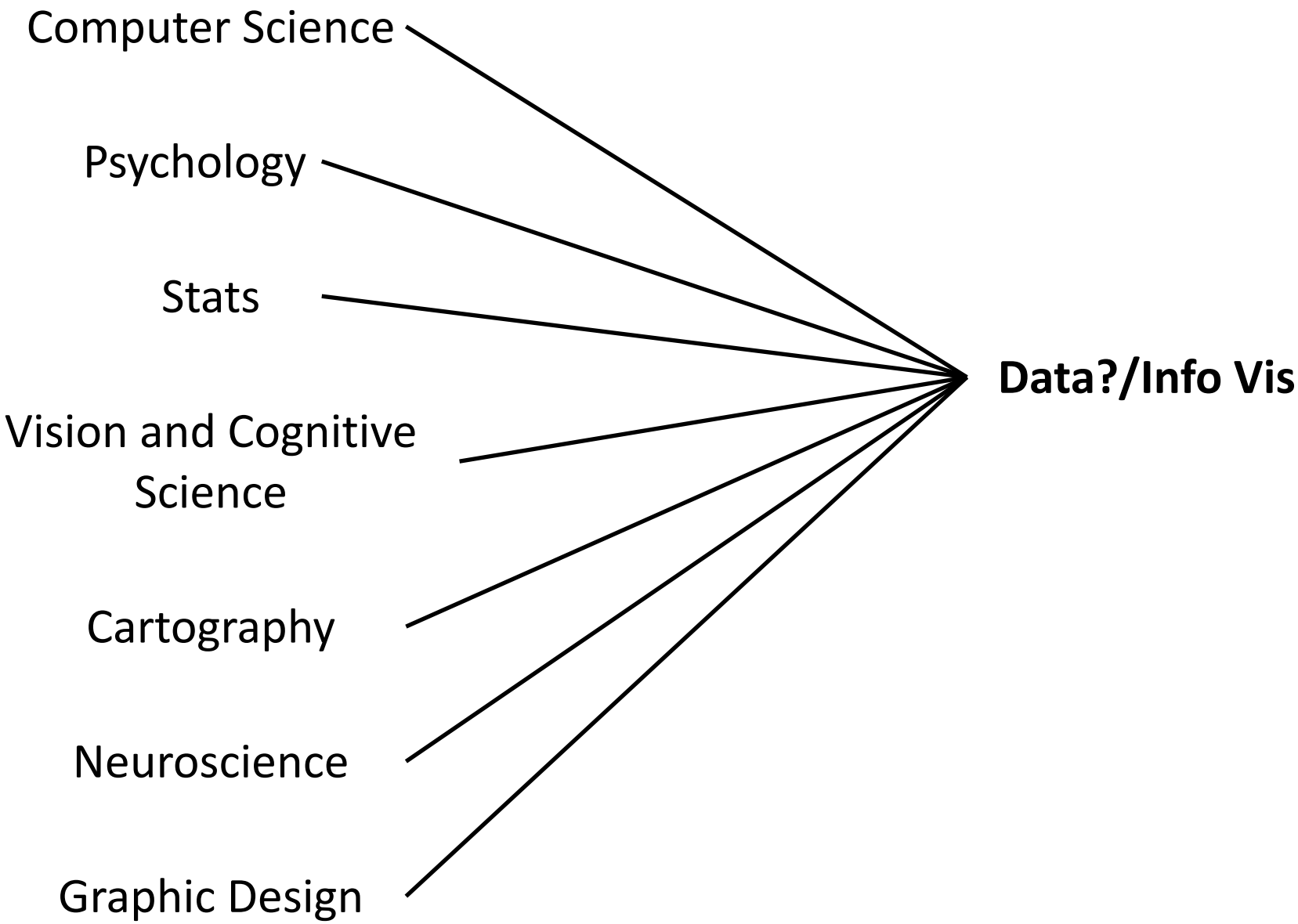


Domain Experts

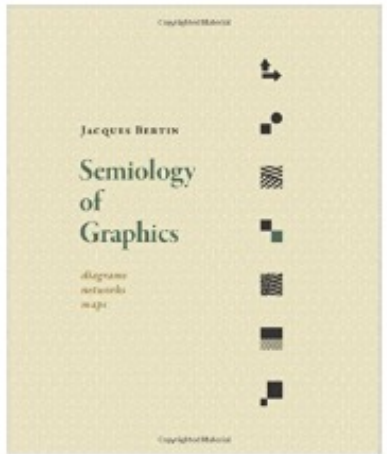


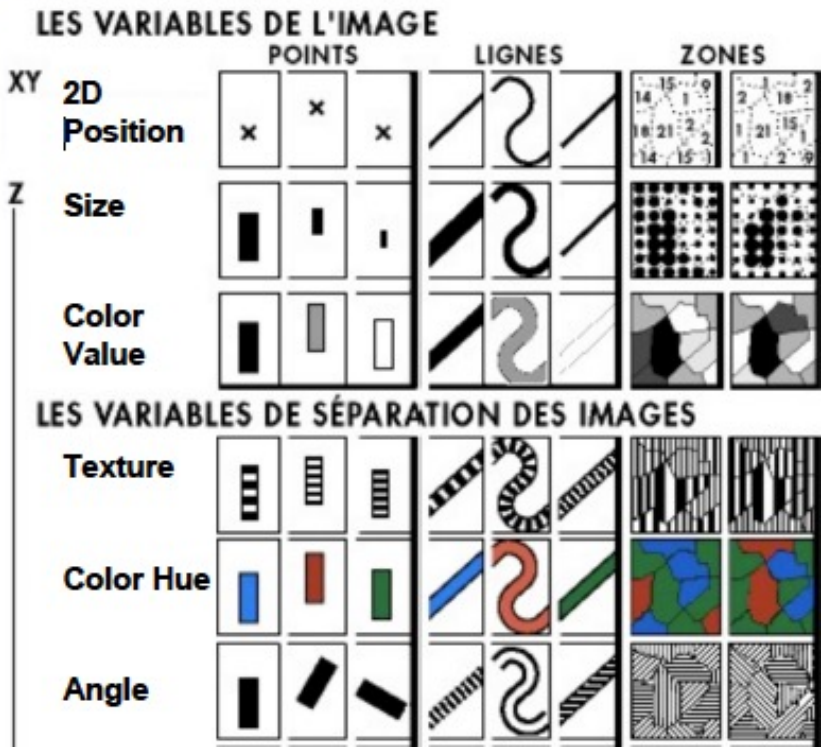
VIS Experts

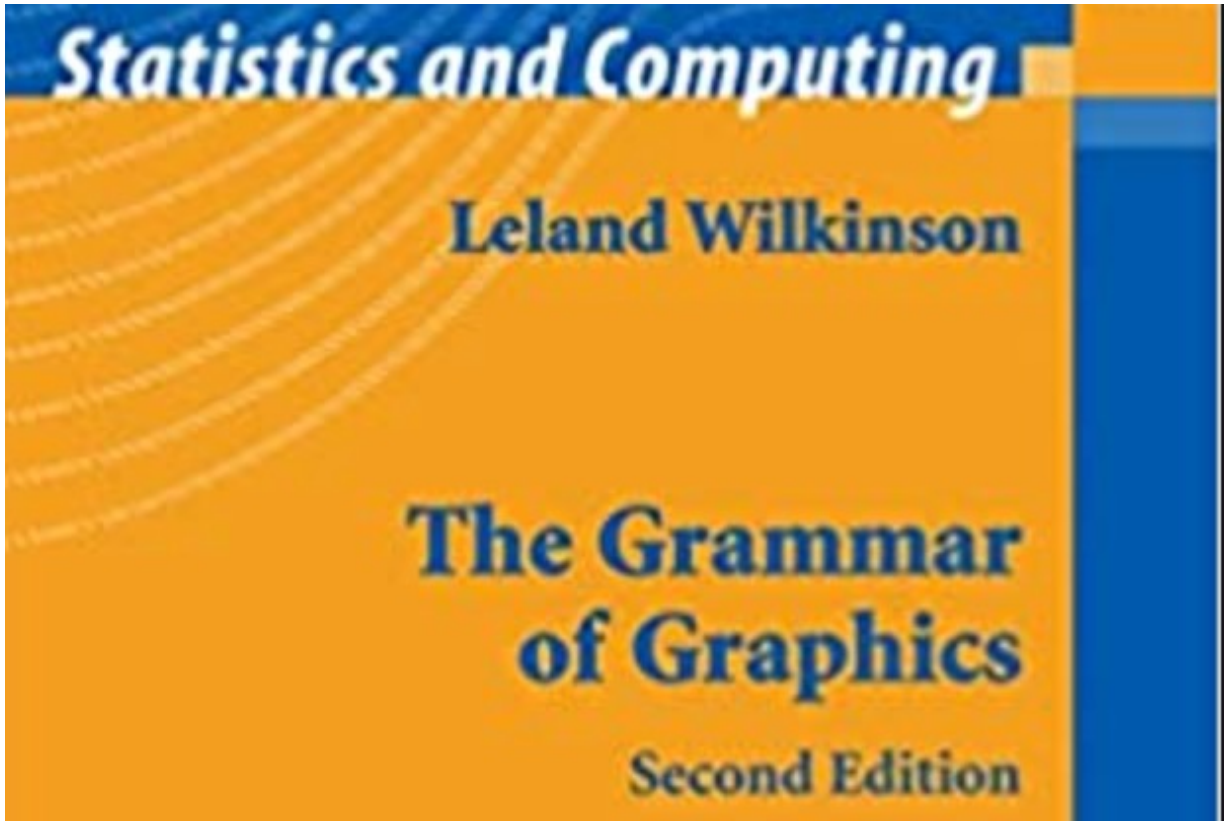
Information Visualization: A multidisciplinary approach



Bertin's Semiology of Graphics (1967)



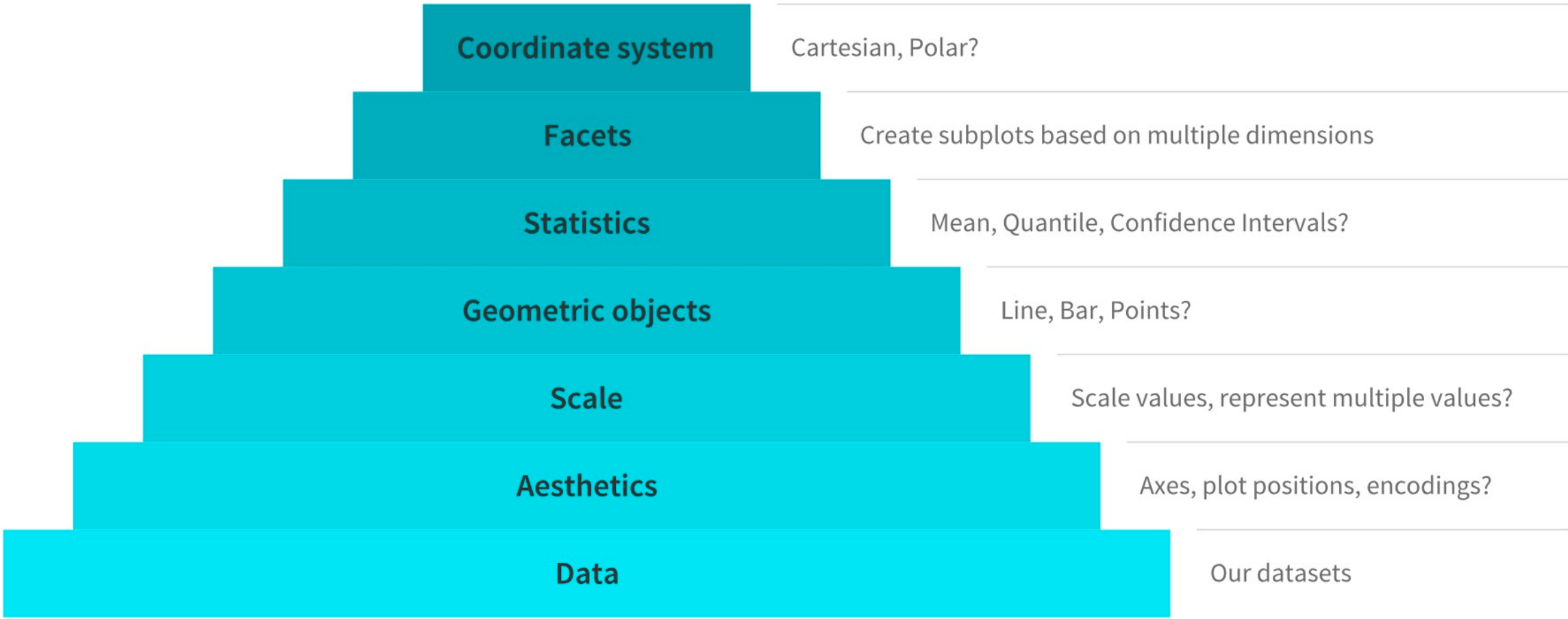




Effective Visualization with Grammar of Graphics

- Grammar is defined as a set of structural rules which helps define and establish the components of a language
- The whole system and structure of a language usually consists of syntax and semantics.
- A grammar of graphics is a framework that enables us to concisely describe the components of any graphic
- Instead of random trials and errors, follow a layered approach by using defined components to build a visualization

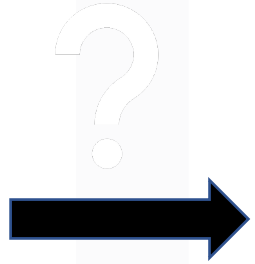
Major Components of the Grammar of Graphics

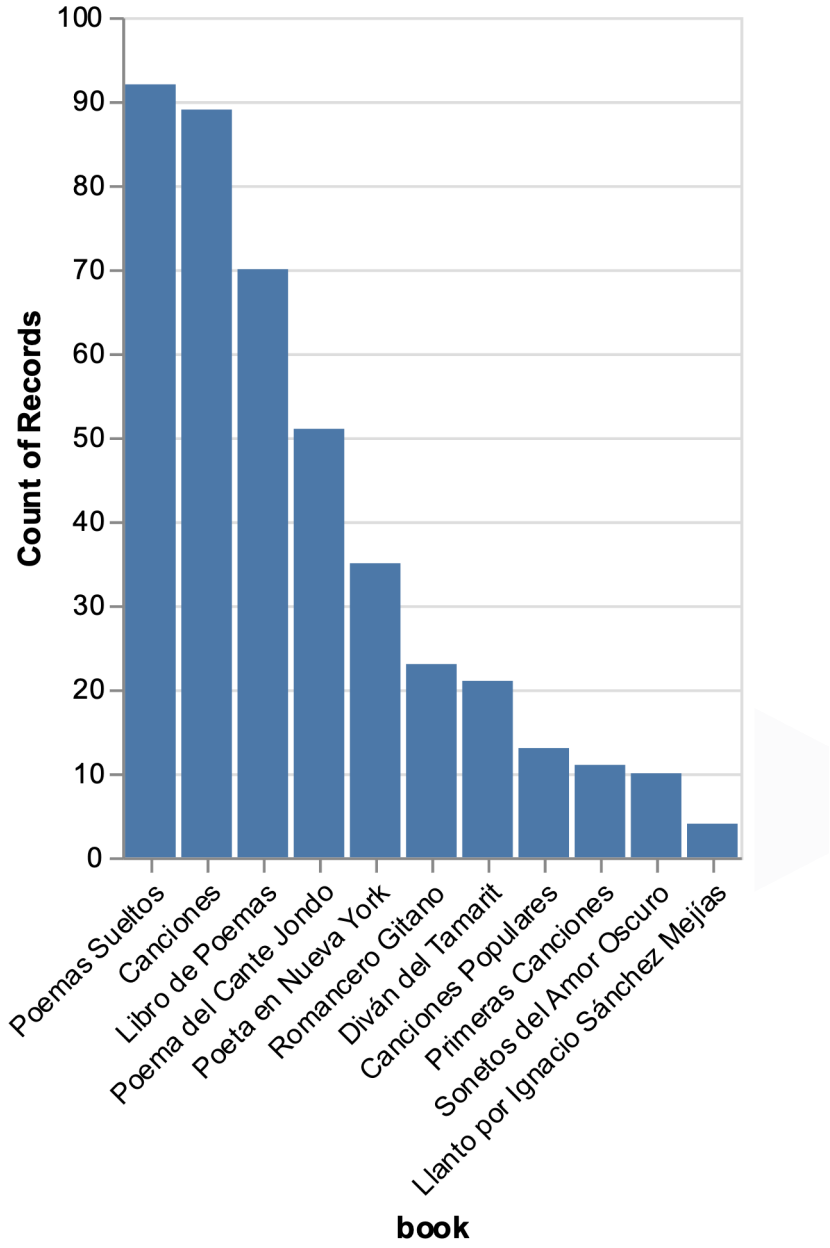


Everything starts with a table

	book	body	subtitle	title
0	Canciones	Cantan las siete\ndoncellas.\n\n(Sobre el ciel...	Canción de las siete doncellas (Teoría del arc...	Teorías
1	Canciones	Hinojo, serpiente y junco.\nAroma, rastro y pe...	Nocturno esquemático.	Teorías
2	Canciones	Sábado.\nPuerta de jardín.\n\nDomingo.\nDía gr...	Canción del colegial.	Teorías
3	Canciones	El canto quiere ser luz.\nEn lo oscuro el cant...	El canto quiere ser luz	Teorías
4	Canciones	Los días de fiesta\nvan sobre ruedas.\nEl tío-...	Tío vivo.	Teorías
...	...	...	...	...
414	Sonetos del Amor Oscuro	¡Ay voz secreta del amor oscuro!\n¡ay balido s...	NaN	Ay voz secreta del amor oscuro
415	Sonetos del Amor Oscuro	Tengo miedo a perder la maravilla\nde tus ojos...	NaN	Soneto de la dulce queja
416	Sonetos del Amor Oscuro	Noche arriba los dos con luna llena,\nyo me pu...	NaN	Noche del amor insomne
417	Sonetos del Amor Oscuro	¿Te gustó la ciudad que gota a gota\nlabró el ...	NaN	El poeta pregunta a su amor por la ciudad enca...
418	Sonetos del Amor Oscuro	Tú nunca entenderás lo que te quiero\nporque d...	NaN	El amor duerme en el pecho del poeta

419 rows × 4 columns





CLSINFRA

HOW DO WE CREATE VISUALIZATIONS?

Popular Data Visualization Tools & Frameworks

D3.js

Tableau

MS Excel

FusionCharts

Highcharts

Leaflet

Datawrapper

Plotly

Kibana

Python Data Visualization Frameworks

Matplotlib

Plotnine (ggplot)

Pandas

Seaborn

Bokeh

Pygal

Plotly

R Data Visualization Frameworks

ggplot2

lattice

ggiraph

taucharts

Plotly

Chart Typologies

Excel, Many Eyes, Google Charts

Charting Tools

Visual Analysis Grammars

VizQL, ggplot2, ***Vega-Lite, Vega-Altair***

Visualization Grammars

Protovis, D3.js, ***Vega***

Declarative Languages

Component Architectures

Prefuse, Flare, Improvise, VTK

Graphics APIs

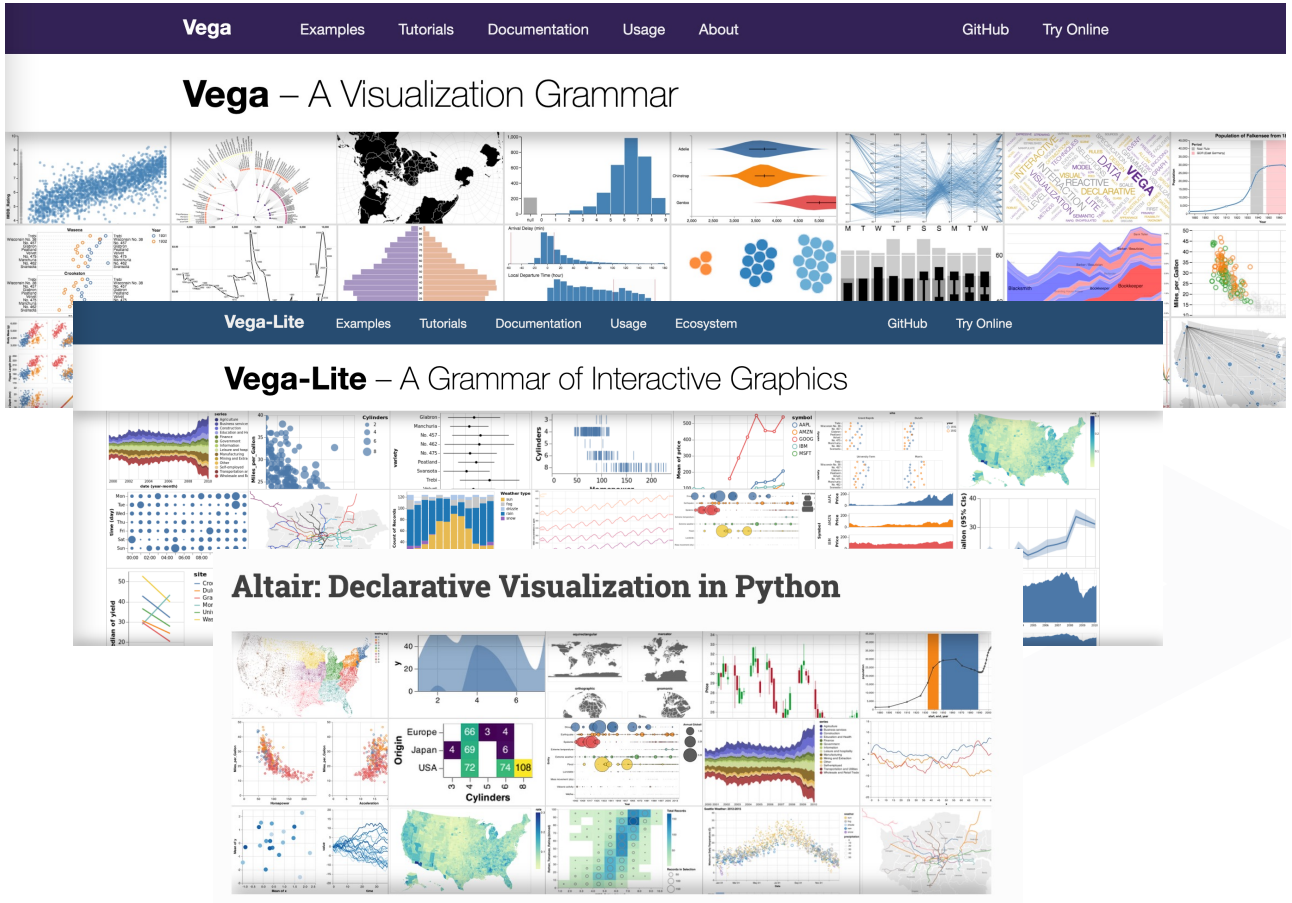
Processing, OpenGL, Java2D

Programming Toolkits

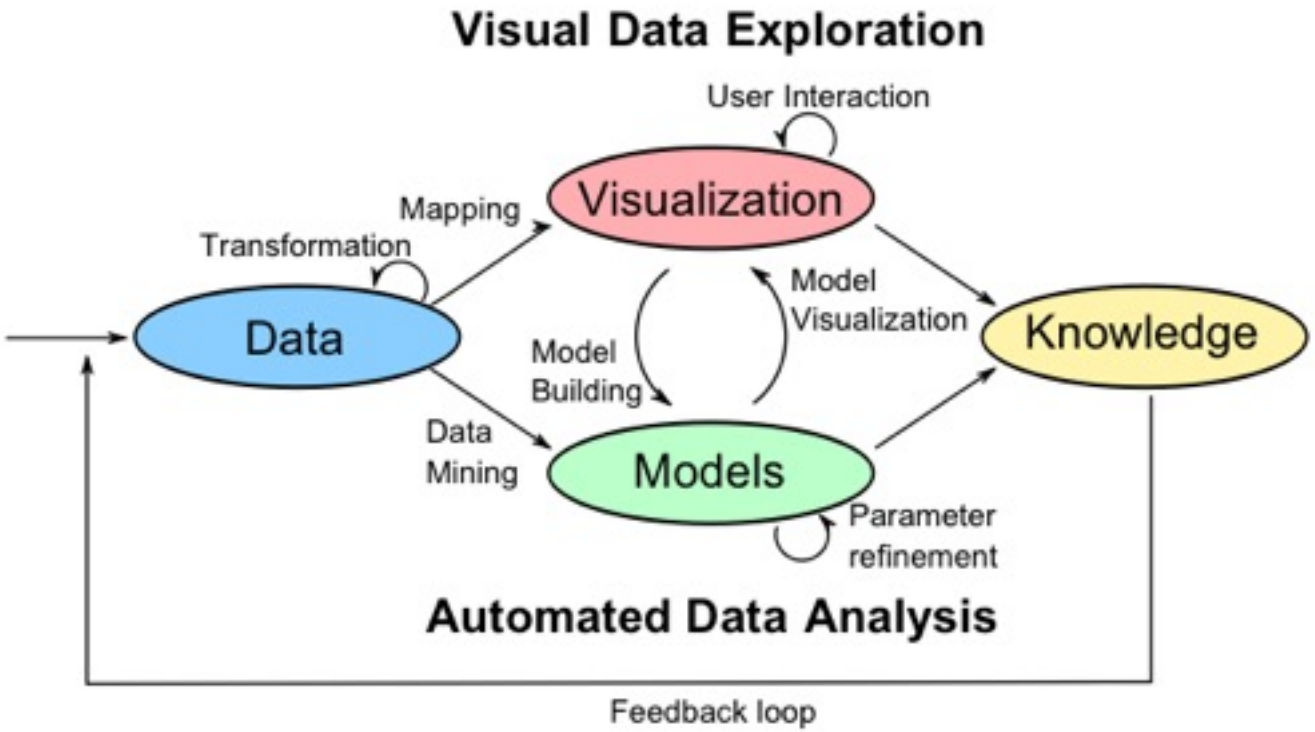
Why vega-altair

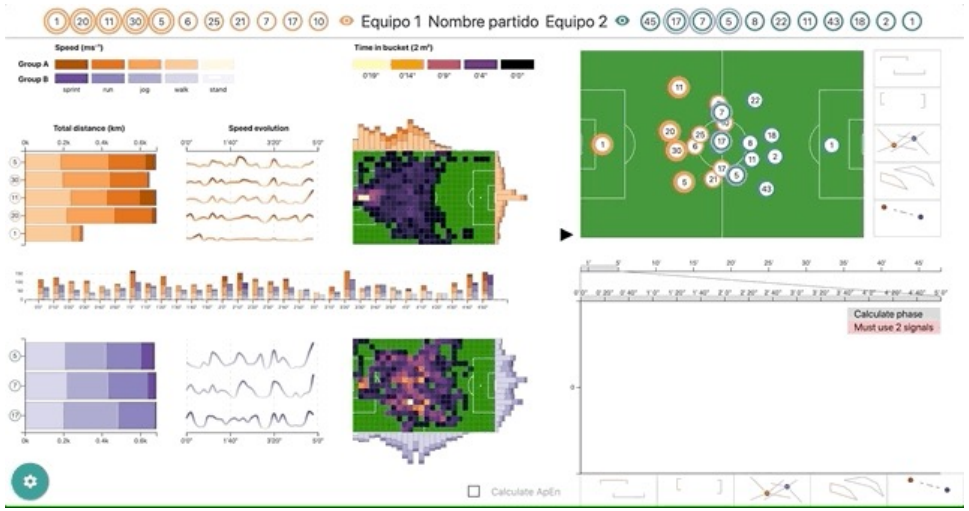
CLSINFRA

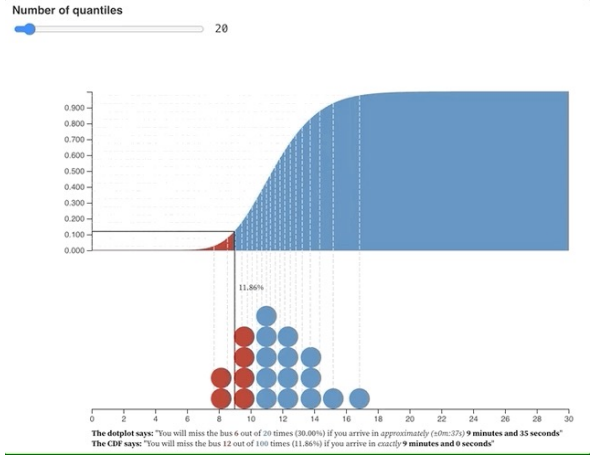
- Declarative syntax: Concise, readable code
- Vega-Lite foundation: Portable JSON specifications
- Easy interactivity: Tooltips, brushing, linking
- Consistent API: Intuitive, versatile usage
- Automatic data transformations: Simplified data handling
- Jupyter Notebook integration: Interactive exploration
- Customization options: Aesthetic control & flexibility
- Active community & development: Ongoing improvements



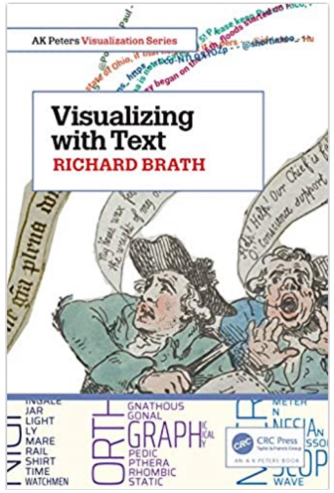
Beyond EDA: Visual Analysis



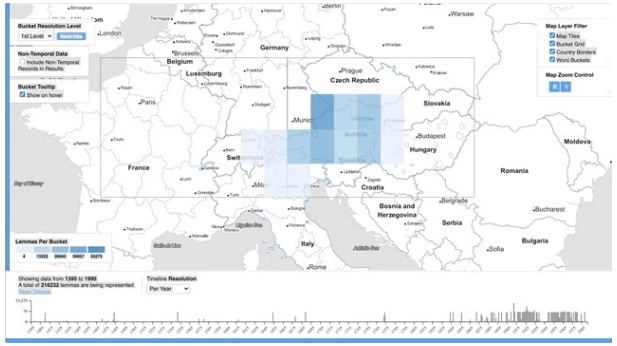


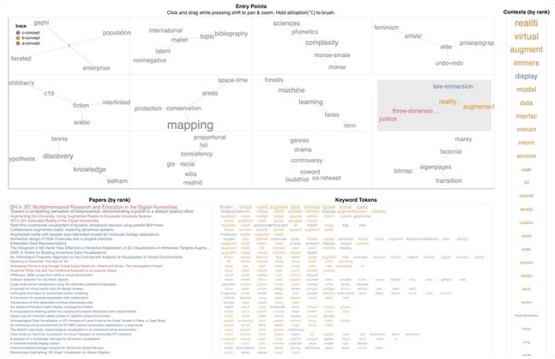


Visual Text Analytics



Link





- Novel specialization of visual analytics that focuses on (semi)structured and unstructured textual data.
- Employs NLP, visualization, and text mining techniques to enhance the comprehension of large bodies of text.
- Highly related to Digital Humanities and cultural heritage!

Context: Automatic scansion of Spanish poems

(1) *Cuando el alba me despierta*
*Cuan-do-el-**al**-ba-me-des-**pier**-ta*
− − + − − − +− 8
(Miguel de Unamuno)

```
!pip install rantanplan spacy-stanza
import rantanplan
!python -m spacy download es_core_news_md
!python -m spacy_affixes download es
from rantanplan.core import get_scansion
```

```
[{'phonological_groups': [{'is_stressed': True, 'syllable': 'Can'},
   {'is_stressed': False, 'is_word_end': True, 'syllable': 'tan'},
   {'is_stressed': False, 'is_word_end': True, 'syllable': 'las'},
   {'is_stressed': True, 'syllable': 'sie'},
   {'is_stressed': False, 'is_word_end': True, 'syllable': 'te'}],
  'rhythm': {'length': 5, 'stress': '+--+-', 'type': 'pattern'},
  'tokens': [{'pos': 'VERB',
    'stress_position': -2,
    'word': [{'is_stressed': True, 'syllable': 'Can'},
     {'is_stressed': False, 'is_word_end': True, 'syllable': 'tan'}]},
   {'pos': 'DET',
    'stress_position': 0,
    'word': [{'is_stressed': False, 'is_word_end': True, 'syllable':
'las'}]},
   {'pos': 'NUM',
    'stress_position': -2,
    'word': [{'is_stressed': True, 'syllable': 'sie'},
     {'is_stressed': False, 'is_word_end': True, 'syllable': 'te'}]}]},
 {'phonological_groups': [{'is_stressed': False, 'syllable': 'don'},
   {'is_stressed': True, 'syllable': 'ce'},
   {'is_stressed': False, 'is_word_end': True, 'syllable': 'llas'}],
  'rhythm': {'length': 3, 'stress': '-+-', 'type': 'pattern'},
  'tokens': [{'pos': 'NOUN',
```

References

- A. Satyanarayan, D. Moritz, K. Wongsuphasawat, y J. Heer, «Vega-Lite: A Grammar of Interactive Graphics», *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, n.º 1, pp. 341-350, ene. 2017, doi: 10.1109/TVCG.2016.2599030.
- J. VanderPlas et al., «Altair: Interactive Statistical Visualizations for Python», The Journal of Open Source Software, vol. 3, p. 1057, dic. 2018, doi: 10.21105/joss.01057.
- J. Chuang, C. D. Manning, y J. Heer, «Termite: Visualization Techniques for Assessing Textual Topic Models», en Proceedings of the International Working Conference on Advanced Visual Interfaces, New York, NY, USA, 2012, pp. 74–77, doi: 10.1145/2254556.2254572.
- J. de la Rosa, Á. Pérez, L. Hernández, S. Ros, y E. González-Blanco, «Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry», Procesamiento del Lenguaje Natural, vol. 65, n.º 0, Art. n.º 0, sep. 2020.
- Jake Vanderplas, *Exploratory Data Visualization with Vega, Vega-Lite, and Altair - PyCon 2018*. 2018. https://youtu.be/ms29ZPUKxbU
- *Wongsuphasawat, Moritz, and Satyanarayan*, *Vega Lite: A Grammar of Interactive Graphics - OpenVisConf* 2017. https://youtu.be/9uaHRWj04D4
- https://altair-viz.github.io/
- https://vega.github.io/vega-lite/tutorials/getting_started.html
- https://github.com/uwdata/visualization-curriculum
- https://courses.cs.washington.edu/courses/cse512/19sp/
- https://vis4dh.dbvis.de/
- https://courses.cs.washington.edu/courses/cse442/18au/lectures/CSE442-ValueOfVisualization.pdf