DSCI 554

COURSE OVERVIEW, INTRODUCTION TO DATA VISUALIZATION

Dr. Luciano Nocera

USC Viterbi
School of Engineering
Integrated Media Systems Center

OUTLINE

- Course information
- Data visualization
- Uses and examples
- Design considerations
- Tools and software
- Sample quiz questions

COURSE OBJECTIVE

- Learn to design plots, infographics and dashboards
- Learn techniques to create effective visual displays
- Learn to build visualizations in notebooks and web pages

WEEKLY QUIZZES 20%

On slides, MCQ & code, worst quiz does not count!

WEEKLY HOMEWORK ASSIGNMENTS 30%

2-4 hours to complete in one week, in GiHub

PROJECTS 30%

Data analysis (notebook), data presentation (web site)

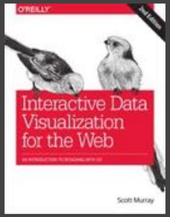
FINAL EXAM 20%

Summative, MCQ & code, Friday, December 9, 2-4 p.m.

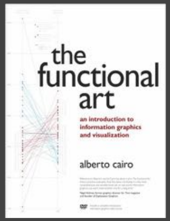
Class materials in blackboard and communications in Slack

READINGS

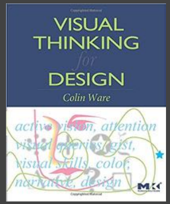
REQUIRED



Murray S. Interactive Data Visualization for the Web, 2nd Edition. 2nd ed. O'Reilly Media, Inc; 2017.

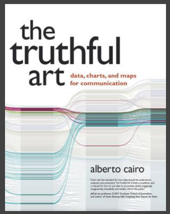


Alberto Cairo. The Functional Art: An Introduction to Information Graphics and Visualization. First. New Riders; 2012.

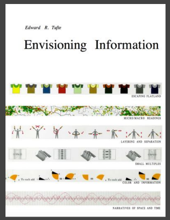


Colin Ware. Visual Thinking: For Design. 1st ed. Morgan Kaufmann Publishers Inc; 2008.

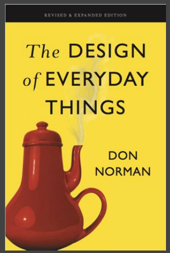
OPTIONAL



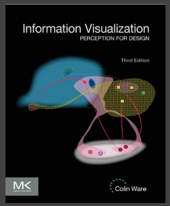
Cairo A. The Truthful Art. Pearson Education; 2016. *



Tufte ER. Envisioning Information . Graphics Press; 1990.



Norman DA. The Design of Everyday Things . 1st Basic paperback ed. Basic Books; 2002.



Ware C. Information Visualization Perception for Design . 3rd ed. Elsevier/MK; 2013. *

Most books are available online through USC Libraries

OUTLINE

- Course information
- Data visualization
- Uses and examples
- Design considerations
- Tools and software
- Sample quiz questions

DATA UNITS

The byte is the unit of data:
1 byte = 8 bits or possible values

Multiples of bytes					
Decimal			**Binary**		
Value	**Metric**		**Value**	**IEC**	
1000	kB	kilobyte	1024	KiB	kibibyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte
1000^6	EB	exabyte	1024^6	EiB	exbibyte
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte
1000^8	YB	yottabyte	1024^8	YiB	yobibyte

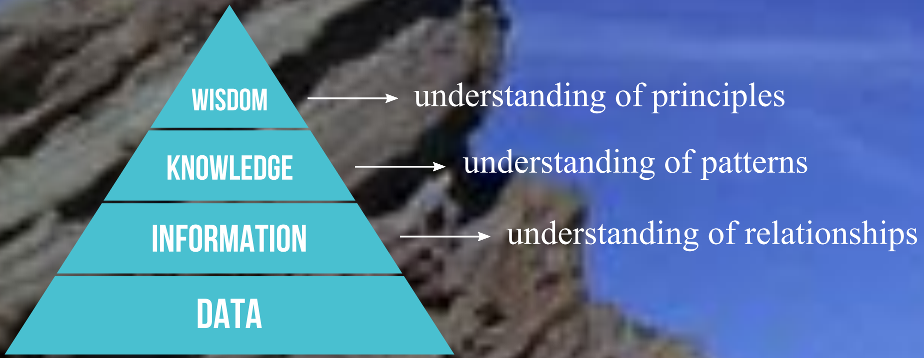
💡 Convert between metric values using powers of kilobytes (kB)

Ex:

$1PB = 1000^5 \text{ bytes} = 1$

Prefix	Symbol	Value
Tera	T	4
Peta	P	5
Exa	E	6
Zetta	Z	7
Yotta	Y	8

DIKW MODEL [ACKOFF 1989]



Data, Information, Knowledge and Wisdom (DIKW) pyramid

Wisdom To go from USC Park Campus to the Griffith Observatory at this time takes 45 min with traffic.

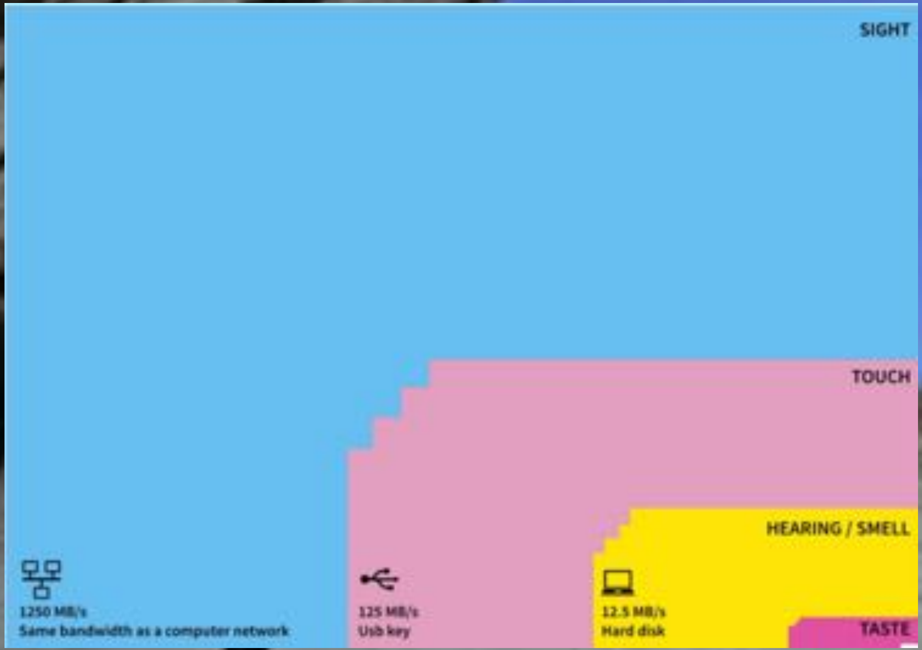
Knowledge It is best to visit the Griffith Observatory weekdays before 4 p.m., because it is less crowded.

Information [The Griffith Observatory is open Tuesday to Friday] [during 12:00 noon - 10:00 p.m.], [admission to building and grounds is always FREE].

Data The [Griffith Observatory] is [open] [Tuesday] to [Friday] during [12:00] [noon] [-] [10:00] [p.m.], [admission] to [building] and [grounds] is [always] [FREE].

Examples of Data, Information, Knowledge and Wisdom

DATA VISUALIZATION



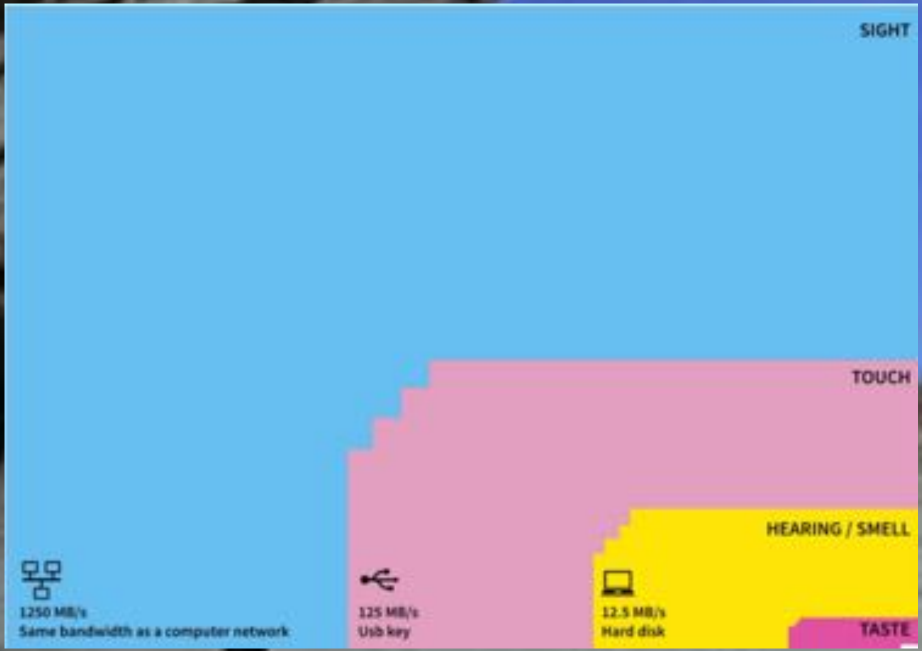
Nørretranders bandwidth of senses Graphic by David McCandless

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science.

Information visualization is the study of (interactive) visual representations of abstract data to reinforce human cognition.

Wikipedia definitions

DATA VISUALIZATION



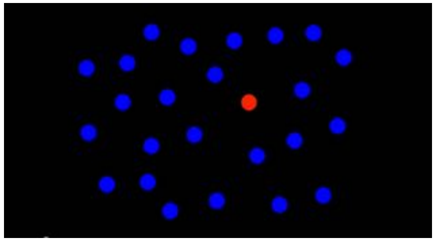
Nørretranders bandwidth of senses Graphic by David McCandless

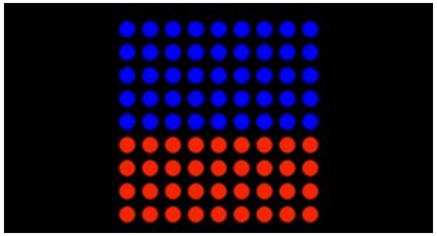
Visual information processing:

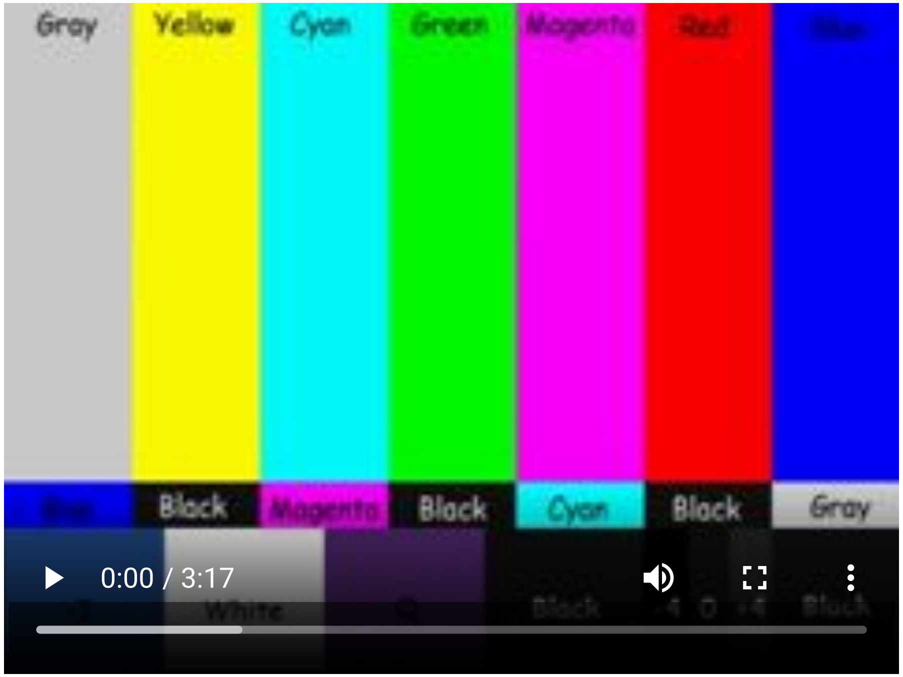
- Only aware of 0.7% of what we experience:
    - High-res limited to central 3° of visual field
    - Finite cognitive capabilities
- Some features are better/faster to see, e.g., preattentive features
- Some features/symbols are not seen/understood by everyone, e.g., universal vs. individual features/symbols

PREATTENTIVE FEATURES

- Typically seen in less than 1/10s
- Does not require eye movements
- Does not require focused attention
- Color and boundary can be detected preattentively







Christopher G. Healey - Preattentive features and tasks

UNIVERSAL VS. INDIVIDUAL CAPABILITIES

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

Universal examples:

- Some color combinations are differentiated by all
- Some symbols are understood across cultures 🙂

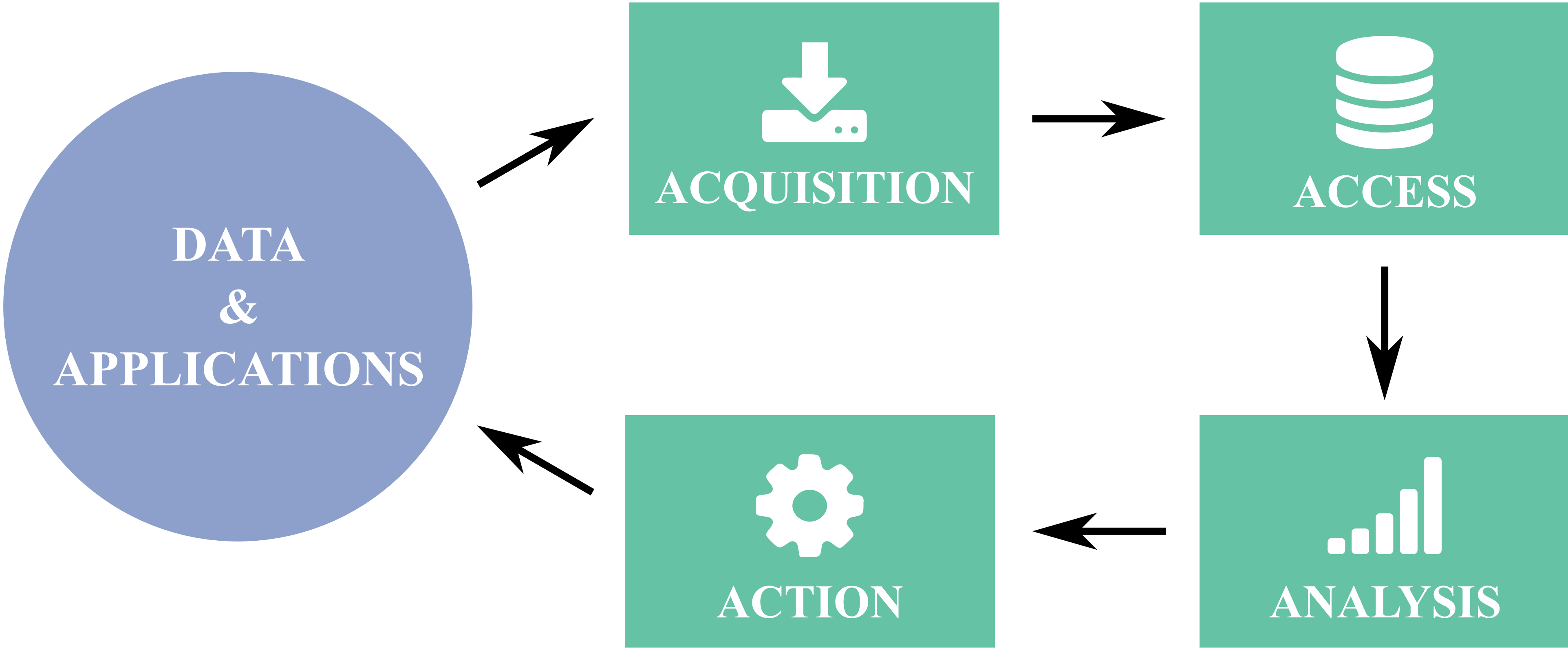
Individual examples:

- We interpret lighting differently
- Not everyone can differentiate certain colors
- Not everyone understands certain symbols ♶
- Not everyone can read small text!

OUTLINE

- Course information
- Data visualization
- Uses and examples
- Design considerations
- Tools and software
- Sample quiz questions

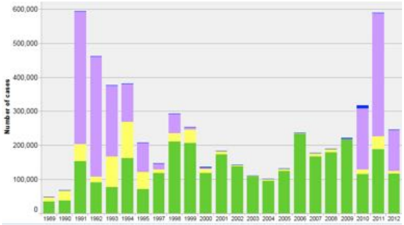
DATA VISUALIZATION IN DATA SCIENCE

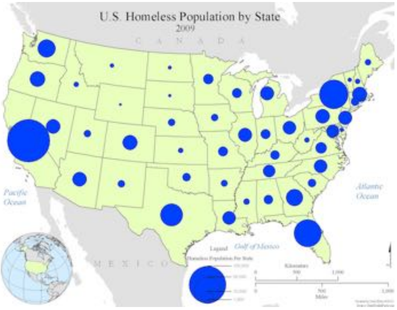


Data science process flowchart

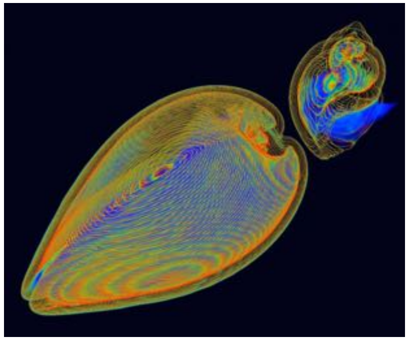
INFOVIS VS. SCIVIS

In information visualization (infovis) the representation is chosen, in scientific visualization (scivis) the representation is given





Examples Infovis



Examples Scivis

	Representation
Infovis	chosen
Scivis	given

AFFORDANCES AND SIGNIFIERS IN INTERACTION DESIGN*

Affordances are properties of objects which show users the actions they can take. Affordances define what actions are possible.

Signifiers are physical signs, for example a word or a sound, that has a meaning. Signifiers specify how people discover those possibilities: signifiers are signs, perceptible signals of what can be done.

* The Design of Everyday Things, Don Norman

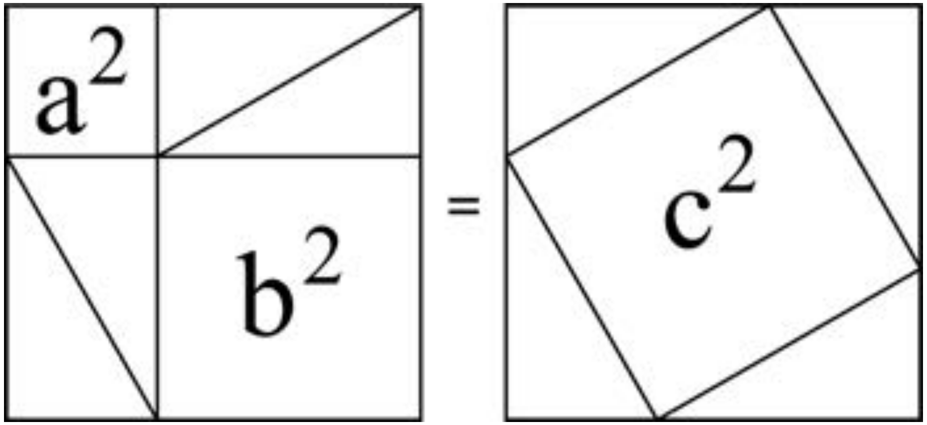
VISUALIZATION USES

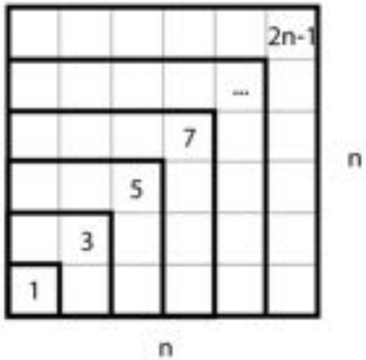
Scope	Goal	Examples
Communicate Information	◦ Inform ◦ Communicate ◦ Explain	◦ Presentations ◦ Hand-outs ◦ Instructions ◦ Infographics ◦ Signage

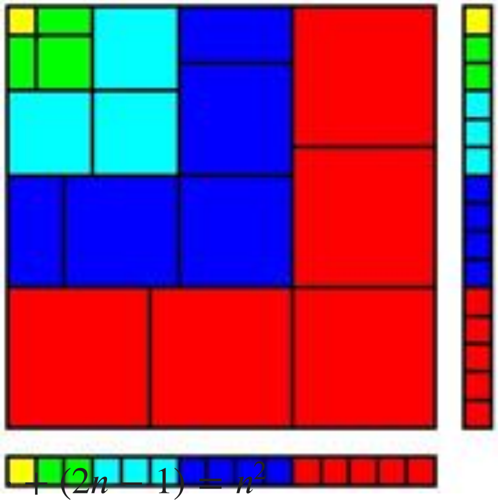
VISUALIZATION USES

Scope	Goal	Examples
Communicate Information	◦ Inform ◦ Communicate ◦ Explain	◦ Presentations ◦ Hand-outs ◦ Instructions ◦ Infographics ◦ Signage
Analyze & Model Data	◦ Explore ◦ Analyze ◦ Discover ◦ Decide	◦ Spreadsheets ◦ Dashboards ◦ Notebooks ◦ Interactive graphics

CAN REPLACE COMPLEX CALCULATIONS







$a^2 + b^2 = c^2$

$1 + 3 + 5 + \cdots + (2n-1) = n^2$

$\sum_{k=1}^{n} k^3 = \left(\sum_{k=1}^{n} k\right.$

CAN REVEAL COMPLEX PATTERNS, TRENDS AND OUTLIERS

193

189

297

311

247

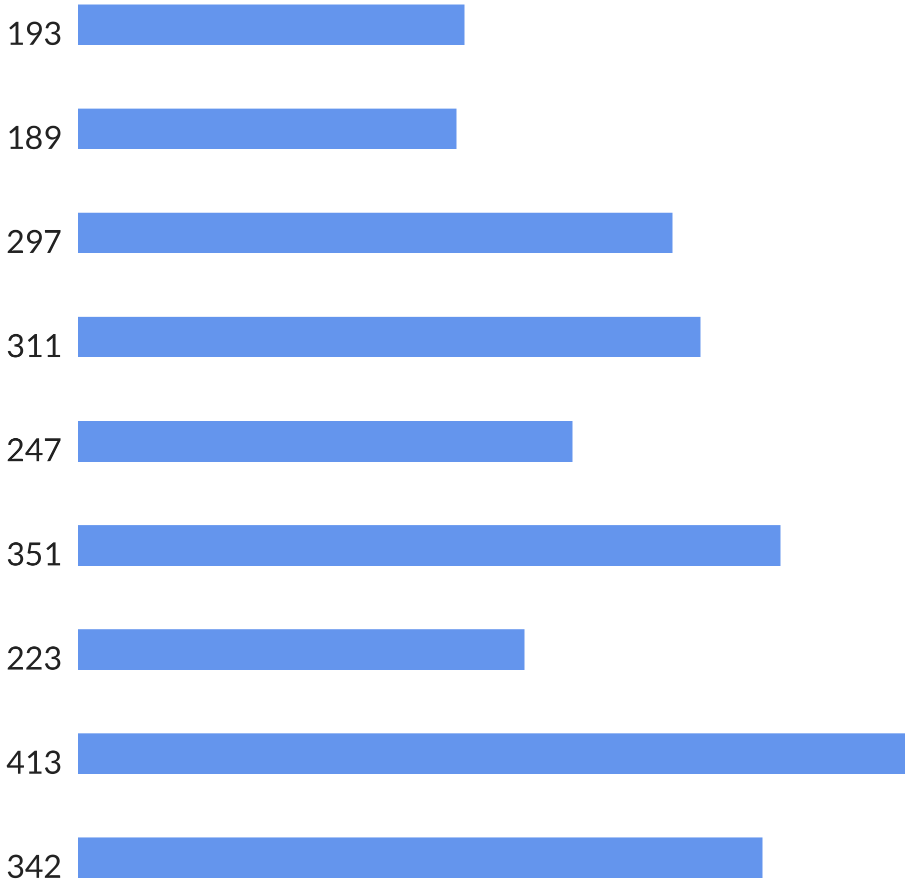
351

223

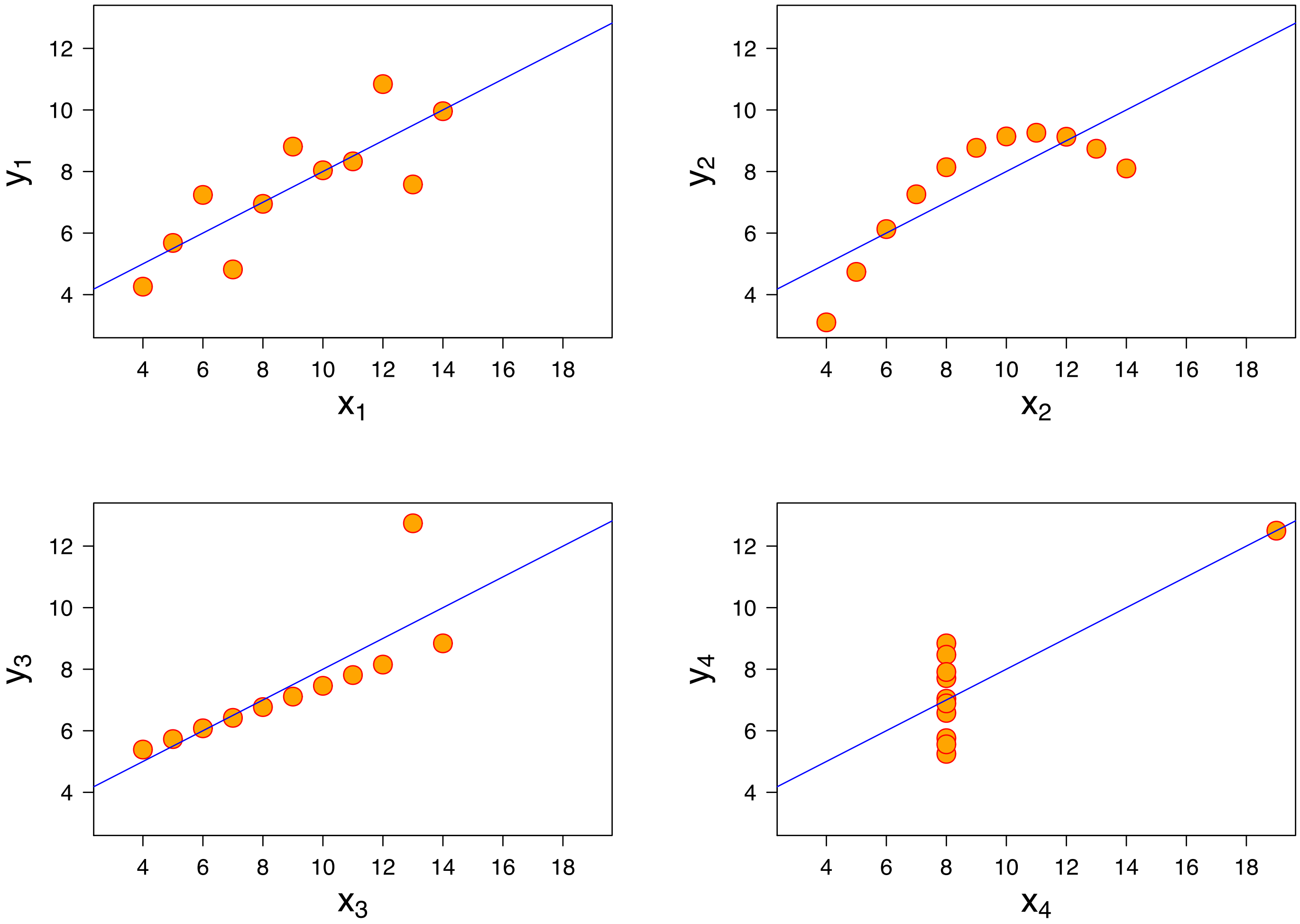
413

342

CAN REVEAL COMPLEX PATTERNS, TRENDS AND OUTLIERS

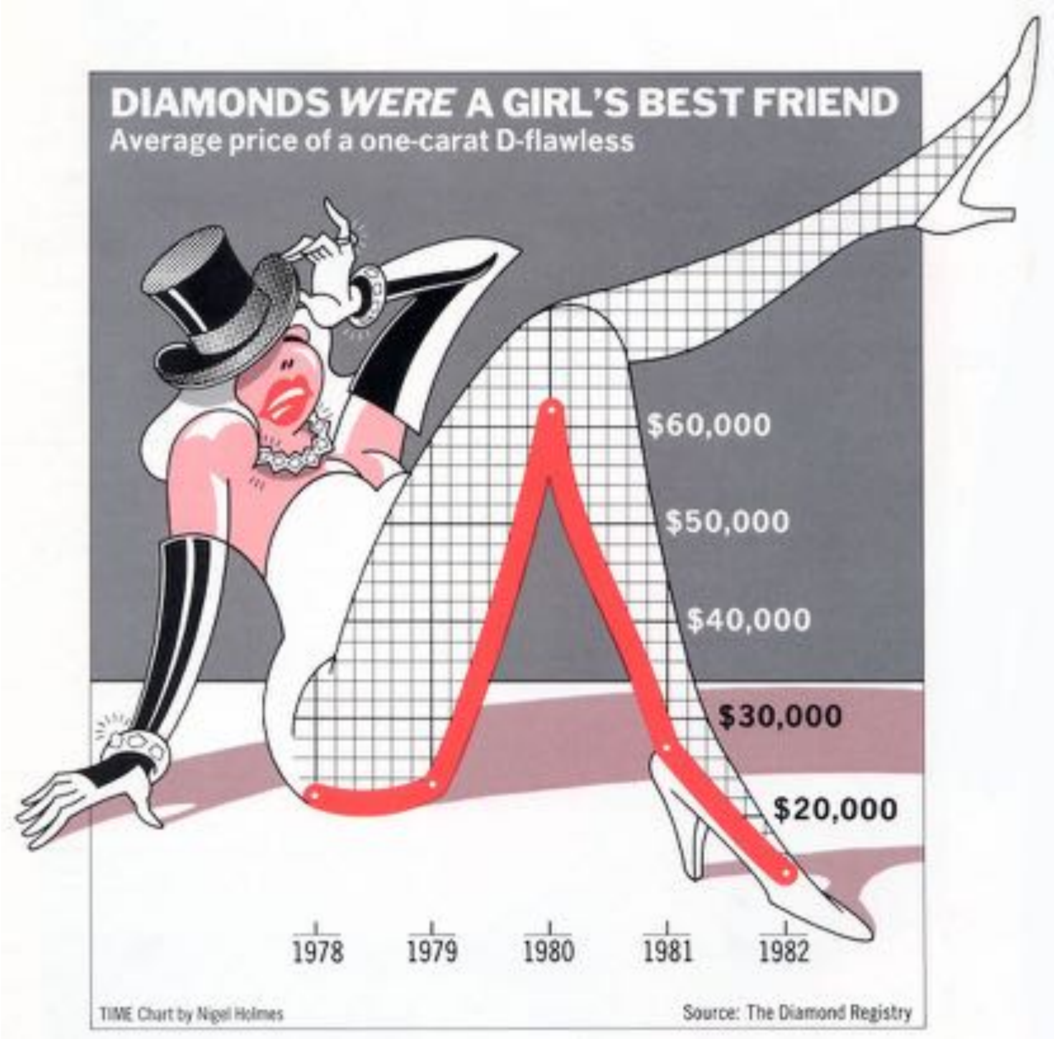


CAN REVEAL FEATURES NOT OTHERWISE APPARENT



Anscombe's quartet (1973): importance of graphing data before analysis

CAN SUPPORT MEMORY AND COMPREHENSION

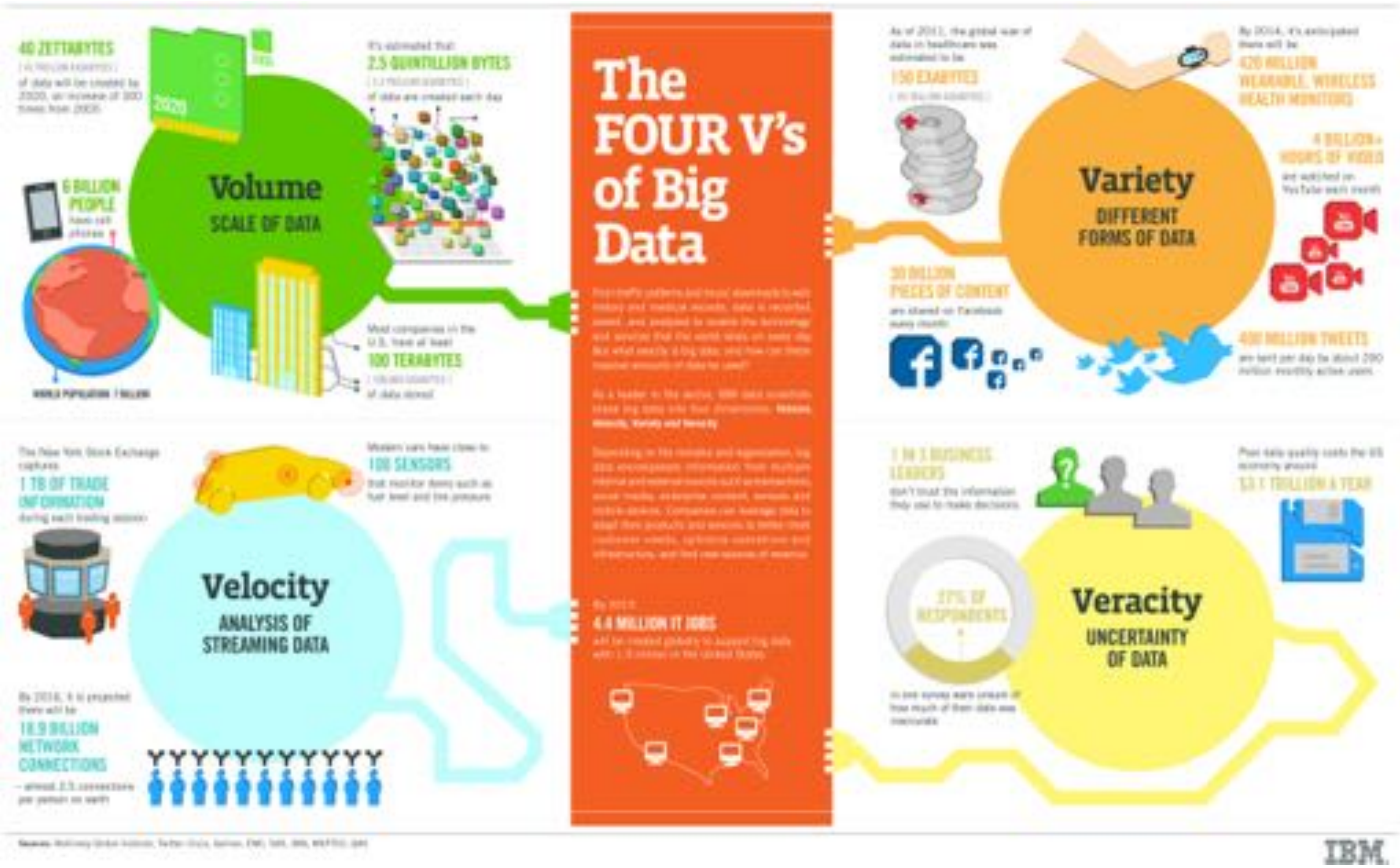


CAN TELL A STORY

Hans Rosling's 200 Countries, 200 Years

🔗 https://youtu.be/jbkSRLYSojo

CAN INFORM AND ENGAGE MORE DIVERSE AUDIENCES

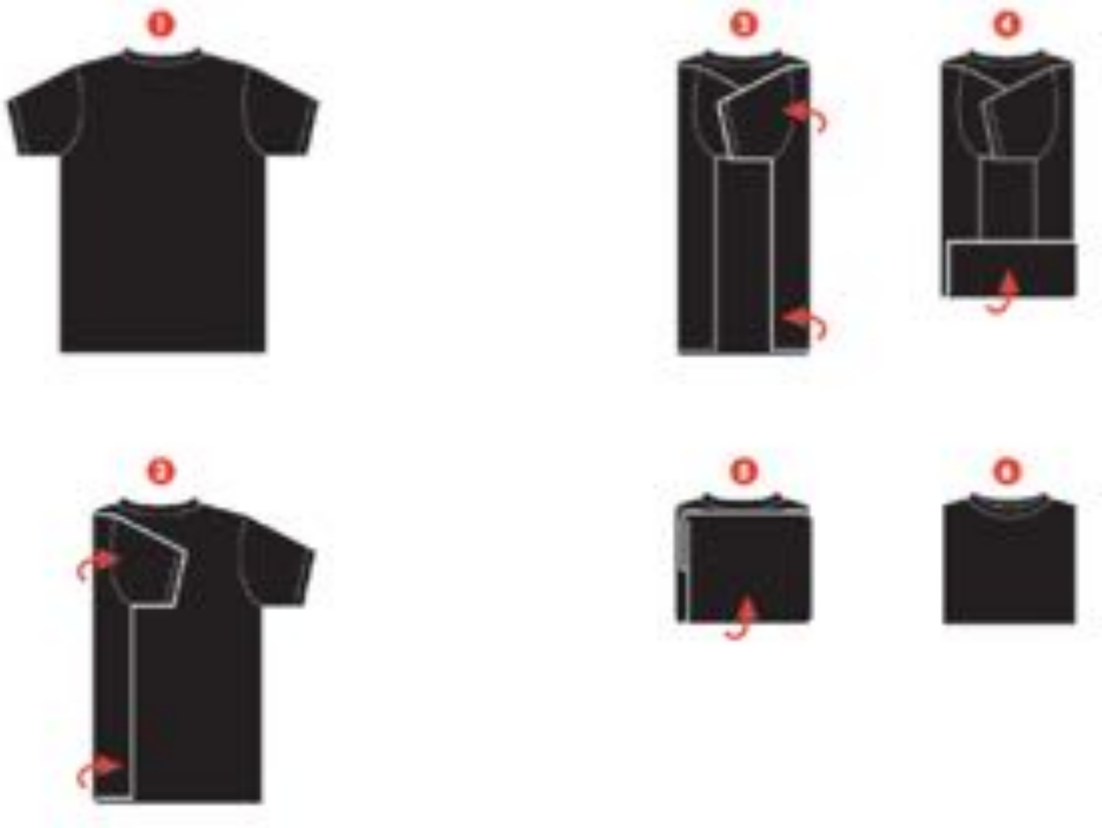


IBM Big Data & Analytics Hub - Infographics & Animations

OUTLINE

- Course information
- Data visualization
- Uses and examples
- Design considerations
- Tools and software
- Sample quiz questions

INFORMATION GRAPHICS* ARE DEVICES WHOSE AIM IS TO HELP AN AUDIENCE COMPLETE CERTAIN TASKS



Wordless Diagrams (2005) by Nigel Holmes.
* Information graphic or infographic

VISUALIZATIONS ARE MEANS TO REACH GOALS

NYT Buy rent calculator

HOW DESIGNERS ENCODE VISUAL INFORMATION FOR USERS

DESIGNER ENCODES		USER DECODES
Information → Visual encoding	→	Visual Decoding → Understanding

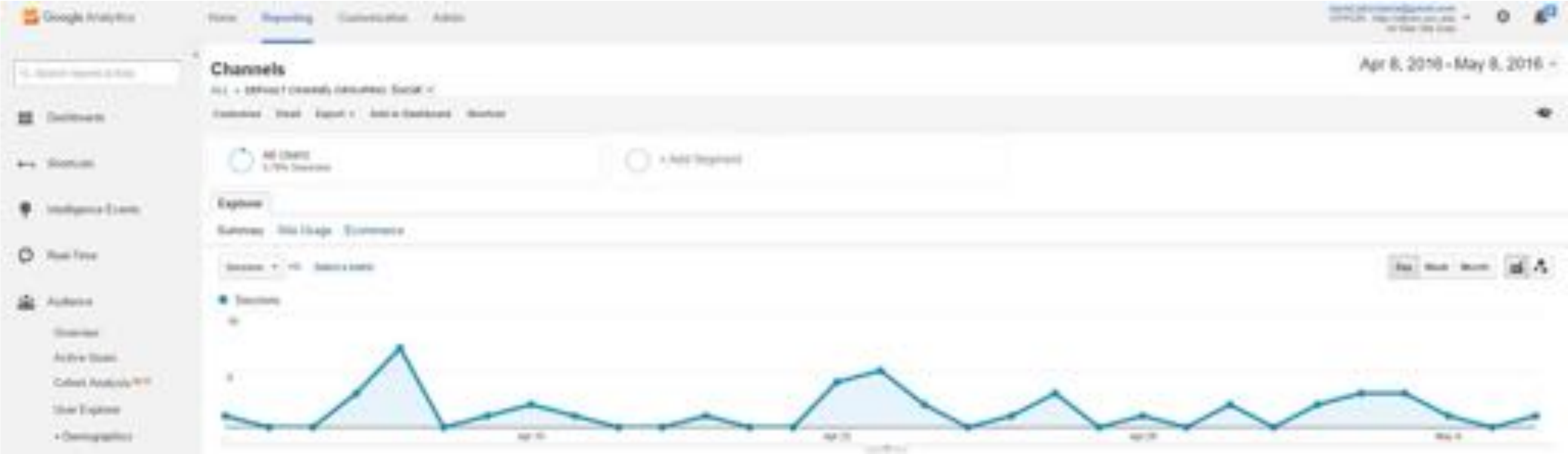
HOW DESIGNERS ENCODE VISUAL INFORMATION FOR USERS

DESIGNER ENCODES		USER DECODES
Information → Visual encoding	→	Visual Decoding → Understanding

INFORMATION DESIGNERS USE

DATA RELATED	USER RELATED
Form adapted to the nature of the information	◦ User familiarity with form ◦ User knowledge of topic ◦ User abilities ◦ Display type and size ◦ Context where the form is used

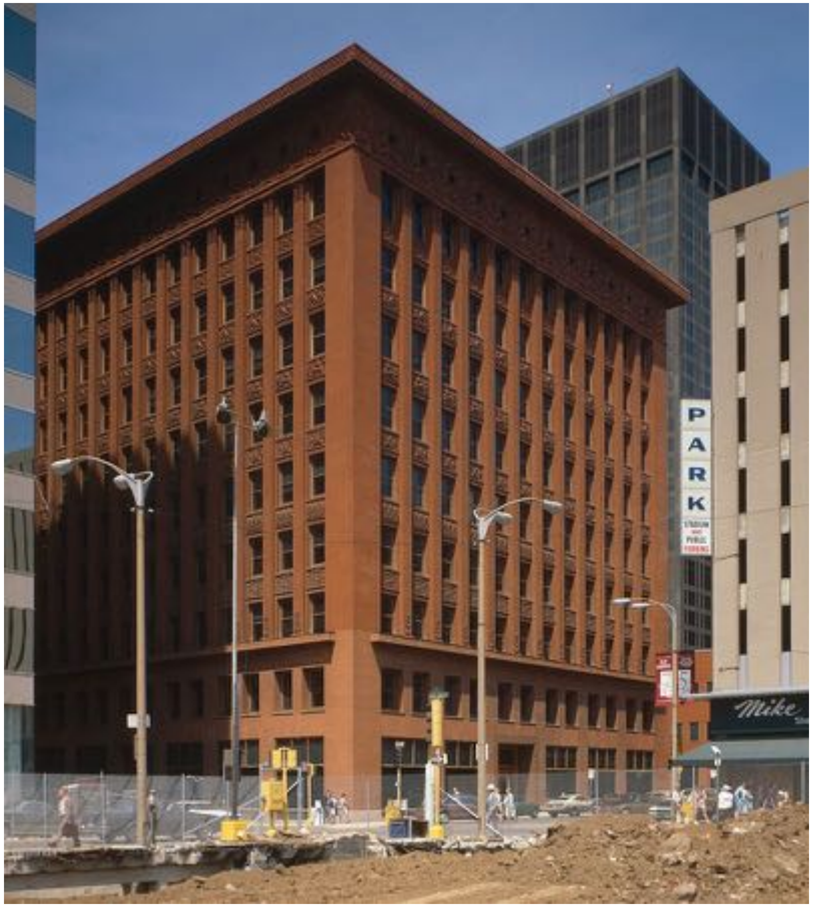
THE FORM SHOULD BE CONSTRAINED BY THE GOALS OF THE VISUALIZATION



Google Analytics dashboard

Form adapted to the nature of the information

FORM FOLLOWS FUNCTION



Sullivan, Louis H. (1896). "The Tall Office Building Artistically Considered". Lippincott's Magazine (March 1896).

20th-century modernist architecture and industrial design principle

The shape of an object should primarily relate to its intended function or purpose

Form adapted to the nature of the information

OUTLINE

- Course information
- Data visualization
- Uses and examples
- Design considerations
- Tools and software
- Sample quiz questions

VISUALIZATION TOOLS

Chart Typologies

Excel, Google Sheets, Matplotlib, Seaborn

Ease-of-Use

Expressiveness

Adapted from [Heer 2014]

Satyanarayan, Arvind, and Jeffrey Heer. "Lyra: An interactive visualization design environment." In Computer Graphics Forum, 2014.

VISUALIZATION TOOLS

Chart Typologies

Excel, Google Sheets, Matplotlib, Seaborn

Visual Analysis Grammars

VizQL, Tableau, ggplot2, Altair

Ease-of-Use

Expressiveness

Adapted from [Heer 2014]

Satyanarayan, Arvind, and Jeffrey Heer. "Lyra: An interactive visualization design environment." In Computer Graphics Forum, 2014.

VISUALIZATION TOOLS

Ease-of-Use

Chart Typologies
Excel, Google Sheets, Matplotlib, Seaborn

Visual Analysis Grammars
VizQL, Tableau, ggplot2, Altair

Visualization Grammars
Protovis, D3, Vega, Vega-Lite

Expressiveness

Adapted from [Heer 2014]
Satyanarayan, Arvind, and Jeffrey Heer. "Lyra: An interactive visualization design environment." In Computer Graphics Forum, 2014.

VISUALIZATION TOOLS

Ease-of-Use

Chart Typologies
Excel, Google Sheets, Matplotlib, Seaborn

Visual Analysis Grammars
VizQL, Tableau, ggplot2, Altair

Visualization Grammars
Protovis, D3, Vega, Vega-Lite

Component Architectures
Prefuse, Flare, Improvise, VTK

Expressiveness

Adapted from [Heer 2014]
Satyanarayan, Arvind, and Jeffrey Heer. "Lyra: An interactive visualization design environment." In Computer Graphics Forum, 2014.

VISUALIZATION TOOLS

Ease-of-Use

Chart Typologies
Excel, Google Sheets, Matplotlib, Seaborn

Visual Analysis Grammars
VizQL, Tableau, ggplot2, Altair

Visualization Grammars
Protovis, D3, Vega, Vega-Lite

Component Architectures
Prefuse, Flare, Improvise, VTK

Graphics applications
P5.js, three.js, Mapbox, WebGL, VTK.js

Expressiveness

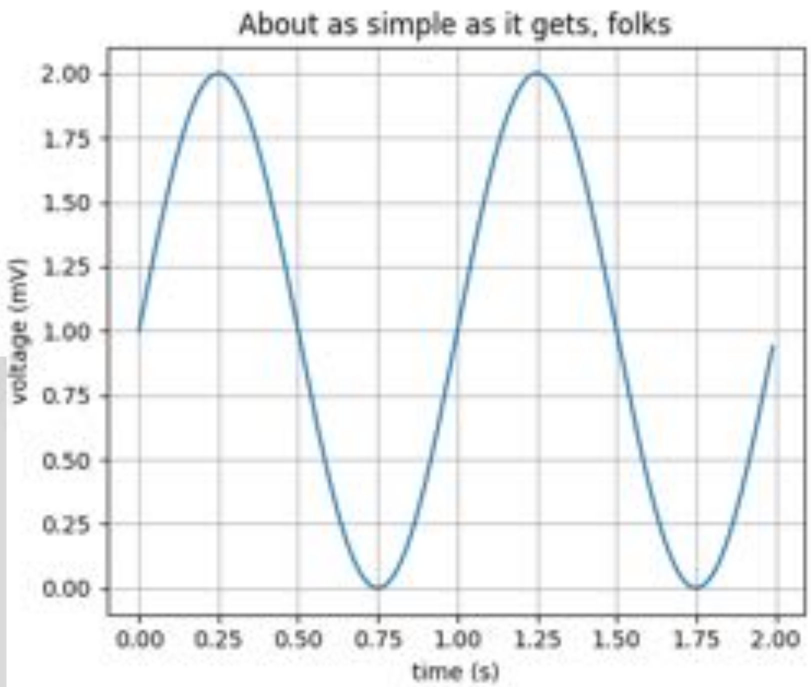
Adapted from [Heer 2014]

Satyanarayan, Arvind, and Jeffrey Heer. "Lyra: An interactive visualization design environment." In Computer Graphics Forum, 2014.

MATPLOTLIB

- https://matplotlib.org/
- Imperative (functional) programming
- Emulating the MATLAB® graphics commands

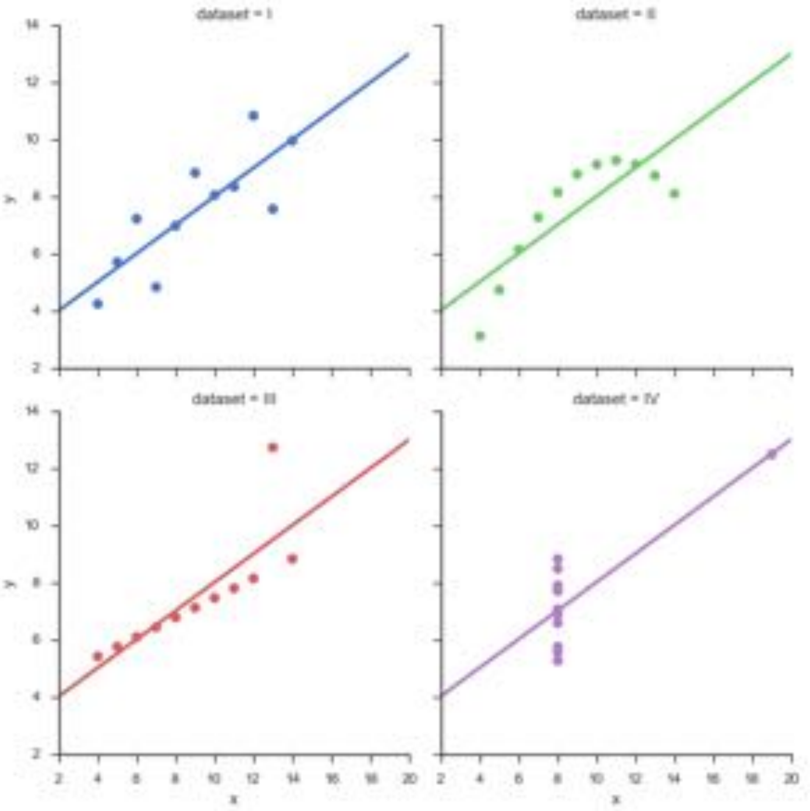
```
import matplotlib.pyplot as plt
import numpy as np

T = np.arange(0.0, 2.0, 0.01)
S = 1 + np.sin(2*np.pi*t)
plt.plot(T, S)

plt.xlabel('time (s)')
plt.ylabel('voltage (mV)')
plt.title('About as simple as it gets, folks')
plt.grid(True)
plt.savefig("test.png")
plt.show()
```



SEABORN

- http://seaborn.pydata.org
- Imperative (functional) programming
- Visualization library based on matplotlib
- High-level interface for statistical graphics
- Support for Pandas

```
import seaborn as sns
sns.set(style="ticks")

# Load the example dataset for Anscombe's quartet
df = sns.load_dataset("anscombe")

# Show the results of a linear regression within each dataset
sns.lmplot(x="x", y="y", col="dataset", hue="dataset", data=df,
           col_wrap=2, ci=None, palette="muted", size=4,
           scatter_kws={"s": 50, "alpha": 1})

sns.plt.show()
```



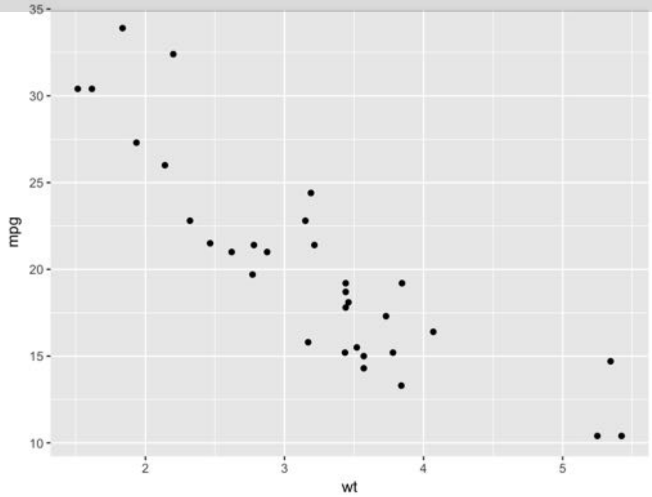
GGPLOT2

Grammar of graphics

- Defaults
  - Data
  - Mapping
- Mapping
  - Layer
  - Data
  - Mapping
  - Geom
  - Stat
  - Position
- Scale
- Coord
- Facet

```
                mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4      21.0 6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag  21.0 6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710     22.8 4 108.0  93 3.85 2.320 18.61  1  1    4    1
...
```

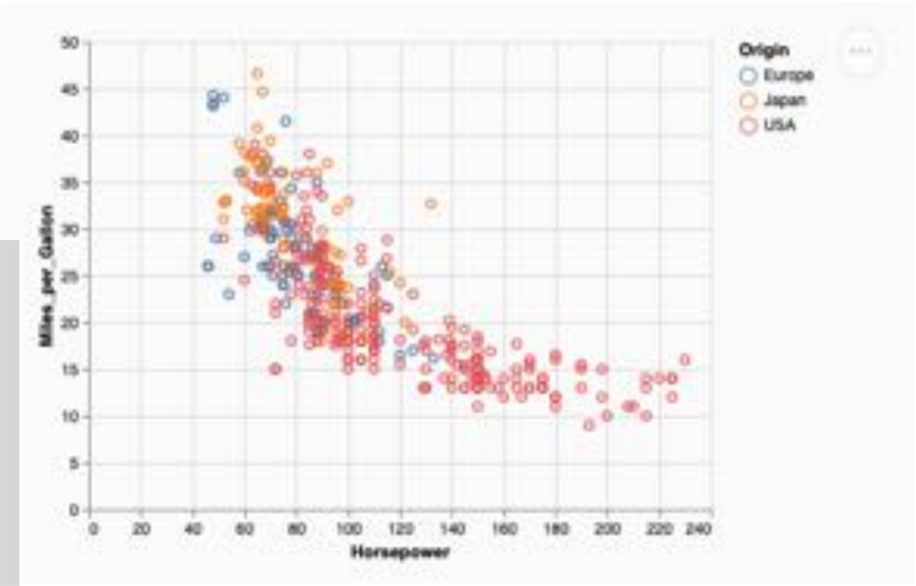
```
library(ggplot2)  #load ggplot2 library in R

#minimal plot: specify data, mapping and geometry:
#ggplot(data, mapping) + geom

ggplot(mtcars, aes(x = wt, y = mpg)) + geom_point()
```



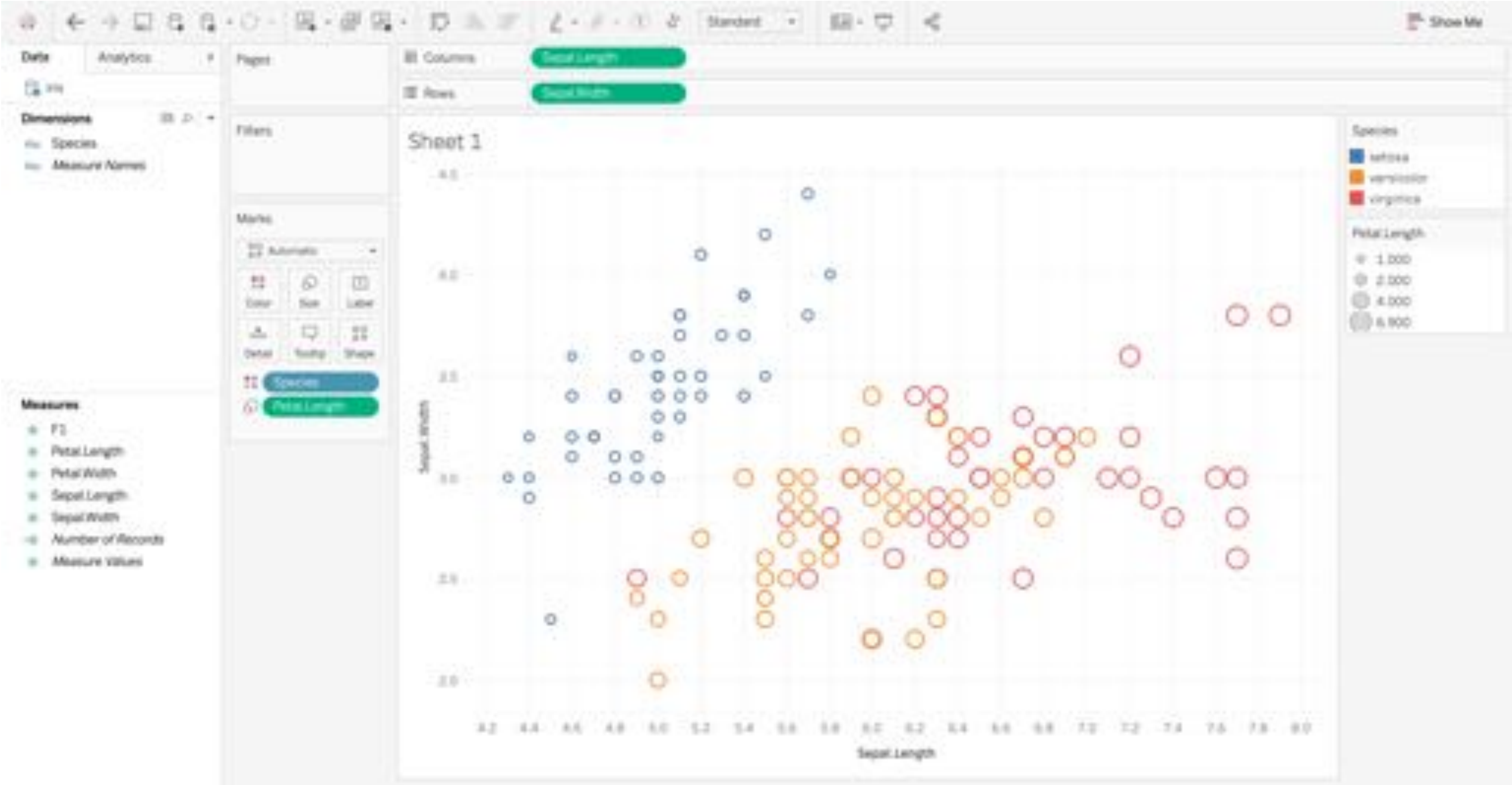
ALTAIR

- Altair page and repository
- Declarative statistical visualization library for Python, based on Vega and Vega-Lite
- Support for Pandas

```python
import altair as alt

# load a simple dataset as a pandas DataFrame
from vega_datasets import data
cars = data.cars()

alt.Chart(cars).mark_point().encode(
  x='Horsepower',
  y='Miles_per_Gallon',
  color='Origin',
).interactive()
```



TABLEAU



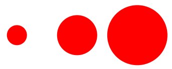
With data read from CSV:
Dimensions ↔ categorical visual variables
Measures ↔ numerical visual variables

D3.JS

What it is	○ Javascript client-side library ○ D3 stands for Data-Driven Documents ○ Uses HTML, SVG, and CSS ○ Primarily made to use SVG (not raster graphics, i.e., images)
What it does	○ Loads data in the browser memory ○ Create elements and bind data to elements within the document ○ Transform and customize elements ○ Transition elements in response to user input

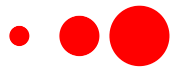
D3 (DATA DRIVEN DOCUMENTS)

[5, 10, 15]

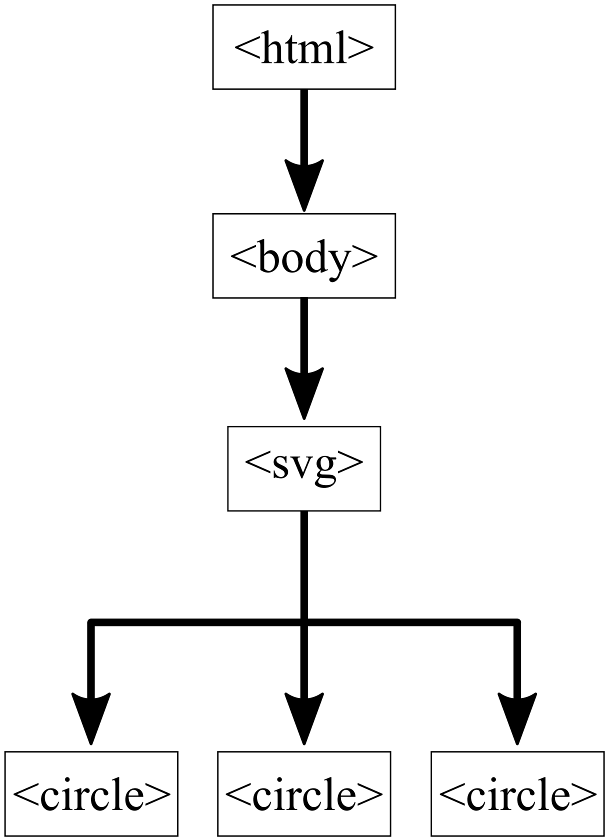


D3 (DATA DRIVEN DOCUMENTS)

[5, 10, 15]



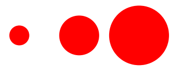
```
<svg>
  <circle r="5"  cx="30" cy="15" fill="red"><circle\>
  <circle r="10" cx="60" cy="15" fill="red"><circle\>
  <circle r="15" cx="90" cy="15" fill="red"><circle\>
</svg>
```



DOM: Document Object Model

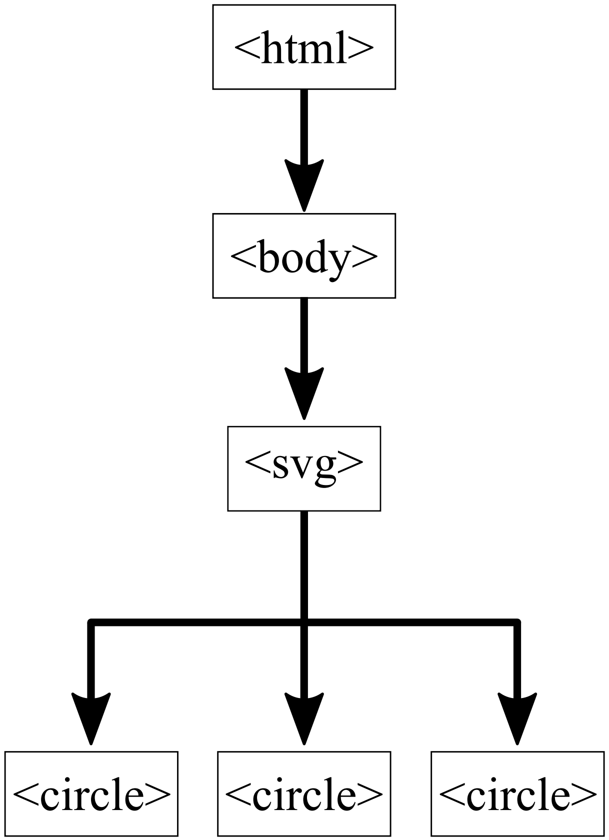
D3 (DATA DRIVEN DOCUMENTS)

[5, 10, 15]



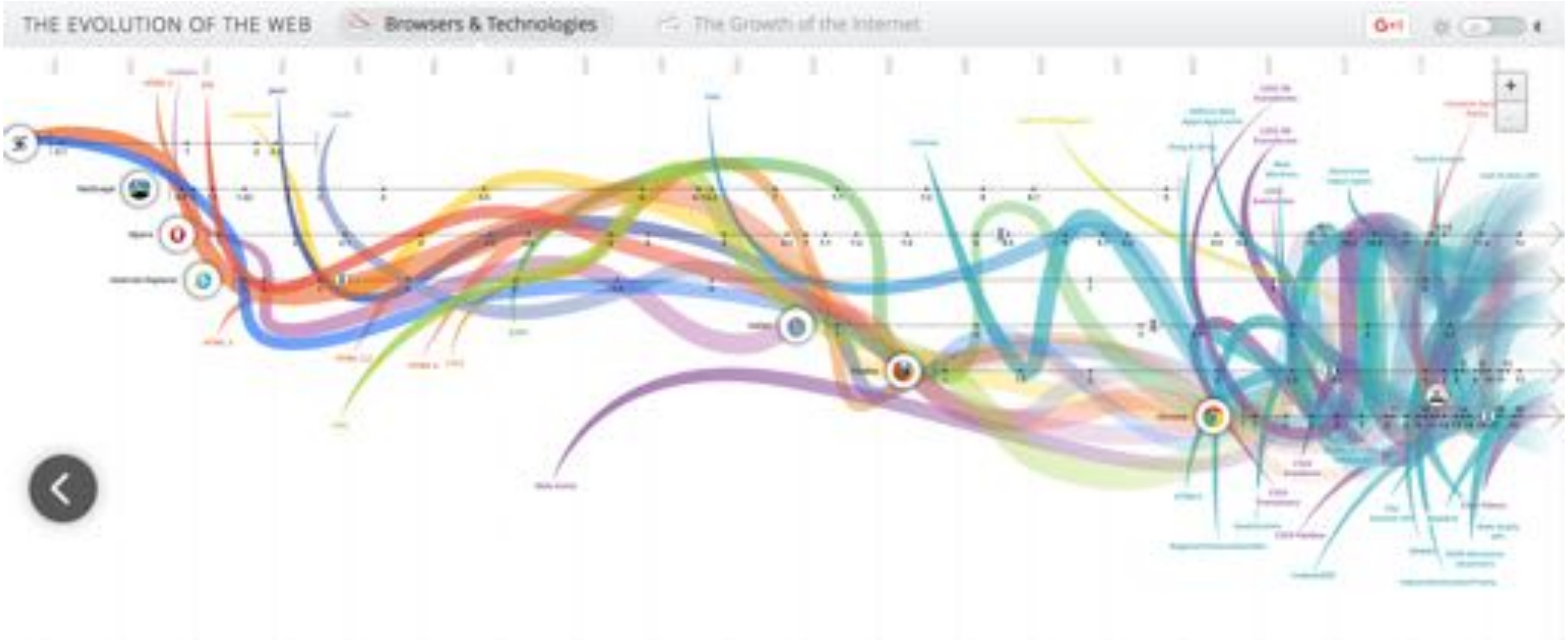
```
<svg>
  <circle r="5"  cx="30" cy="15" fill="red"><circle\>
  <circle r="10" cx="60" cy="15" fill="red"><circle\>
  <circle r="15" cx="90" cy="15" fill="red"><circle\>
</svg>
```

```
<html>
<body>
  <svg id="chart" height="30px"></svg>

  <script src="http://d3js.org/d3.min.js"></script>
  <script>
    var svg = d3.select('chart');
    var dataset = [5, 10, 15];

    svg.selectAll("circle")
      .data(dataset)
      .enter()
      .append("circle")
      .attr('cx', function(d, i) { return 30 * (i + 1); } )
      .attr('cy', '15')
      .attr('r', function(d) { return d; })
      .attr('fill', 'red');
  </script>
</body>
</html>
```



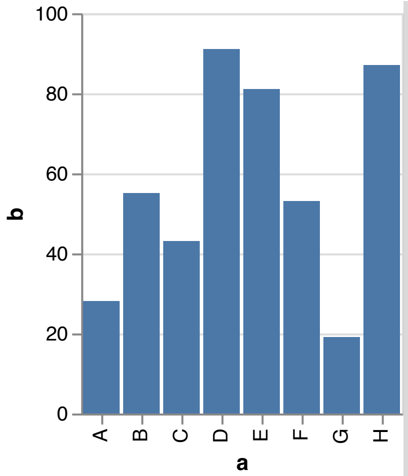
DOM: Document Object Model

INTERACTIVE VISUALIZATIONS WITH D3



http://www.evolutionoftheweb.com

VEGA



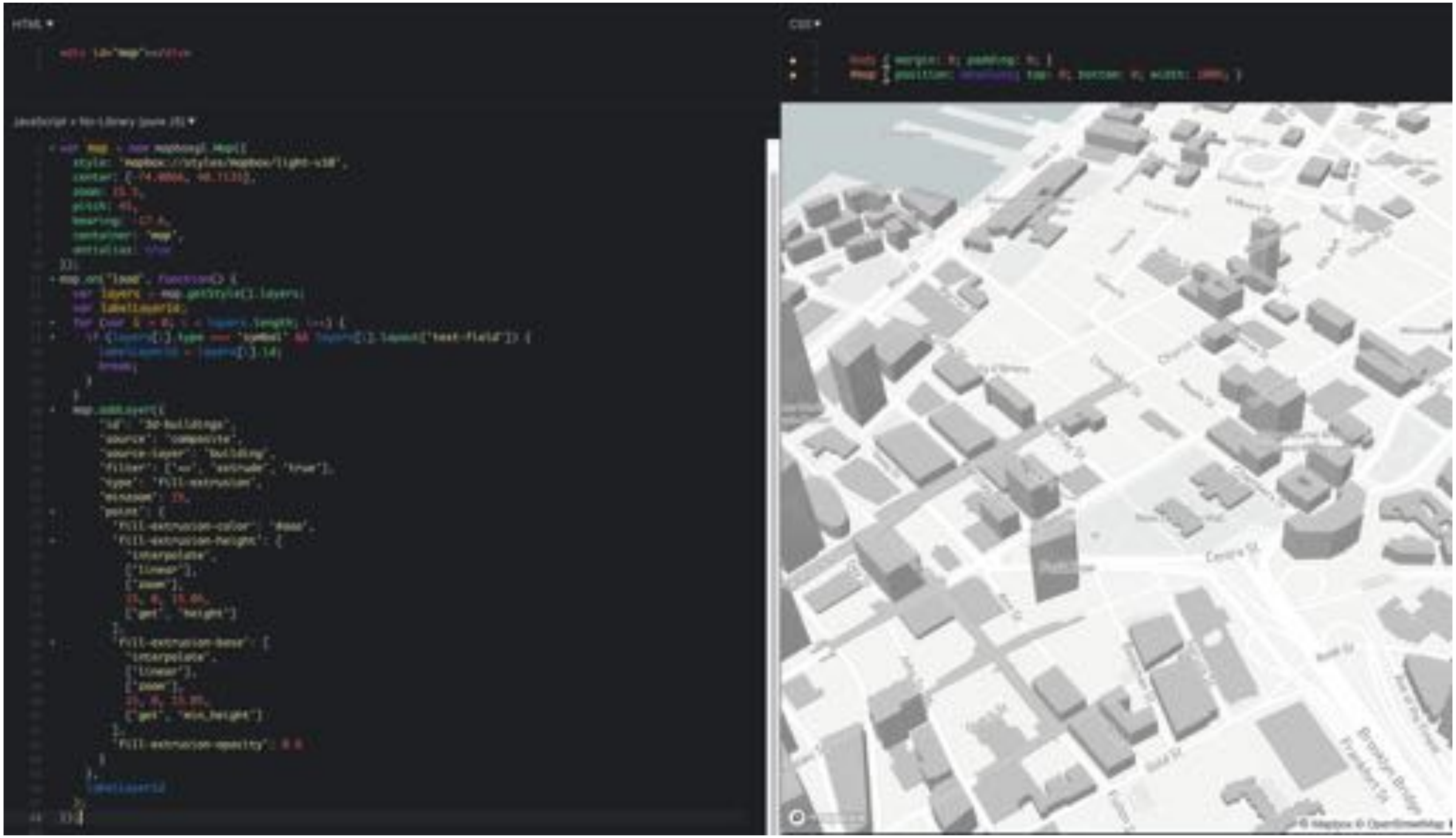

```
{
  "$schema": "https://vega.github.io/schema/vega-lite/v3.json",
  "description": "A simple bar chart with embedded data.",
  "data": {
    "values": [
    {"a": "A","b": 28},
    {"a": "B","b": 55},
    {"a": "C","b": 43},
    {"a": "D","b": 91},
    {"a": "E","b": 81},
    {"a": "F","b": 53},
    {"a": "G","b": 19},
    {"a": "H","b": 87}
    ]
  },
  "mark": "bar",
  "encoding": {
    "x": {"field": "a", "type": "ordinal"},
    "y": {"field": "b", "type": "quantitative"}
  }
}
```

Vega-Lite

```
{
  "$schema": "https://vega.github.io/schema/vega/v4.json",
  "width": 400,
  "height": 200,
  "padding": 5,

  "data": [
    {
      "name": "table",
      "values": [
        {"category": "A", "amount": 28},
        {"category": "B", "amount": 55},
        {"category": "C", "amount": 43},
        {"category": "D", "amount": 91},
        {"category": "E", "amount": 81},
        {"category": "F", "amount": 53},
        {"category": "G", "amount": 19},
        {"category": "H", "amount": 87}
      ]
    }
  ],

  "signals": [
    {
      "name": "tooltip",
      "value": {},
      "on": [
        {"events": "rect:mouseover", "update": "datum"},
        {"events": "rect:mouseout",  "update": "{}"}
      ]
    }
  ],

  "scales": [
    {
      "name": "xscale",
      "type": "band",
      "domain": {"data": "table", "field": "category"},
      "range": "width",
      "padding": 0.05,
      "round": true
    },
```

Vega

MAPBOX



See also Mapbox Display buildings in 3D example

OUTLINE

- Course information
- Data visualization
- Uses and examples
- Design considerations
- Tools and software
- Sample quiz questions

Astronomers expect to be processing 10 petabytes of data every hour from the Square Kilometer Array (SKA) telescope.

1. How many 1TB drives would be filled in a day?
2. How many days would it take to collect one exabyte?
3. How may zettabytes would be collected in a year?

1. 240000 drives

$$1\text{PB} = 10^{15}\text{bytes} = 10^3 \times 10^{12}\text{bytes} = 10^3\text{TB}$$

$$10\text{PB} \times 24\text{h} = 24x10 \times 10^3\text{TB} = 240000 \times 1\text{TB drives}$$

2. About 4 days

$$24\text{PB} \times x = 1\text{EB} \Rightarrow 24 \times 10^{16} \times x = 10^{18} \Rightarrow x = 100/24 \simeq$$

3. About 0.1 ZB/year

$$1ZB = 10^{21}\text{bytes}$$

$$365 \times 24 \times 10\text{PB} = 365 \times 24 \times 10^{16} \simeq 10^4 \times 10^{16} \simeq 10^{20} bytes \simeq 0.1\text{ZB/year}$$

Which information visualization use most relates to communicating information?

A. Explore
B. Analyze
C. Explain
D. Decide

Which information visualization use most relates to communicating information?

A. Explore
B. Analyze
C. Explain ←
D. Decide