Data Visualization

Spring Semester 20/21

Introduction

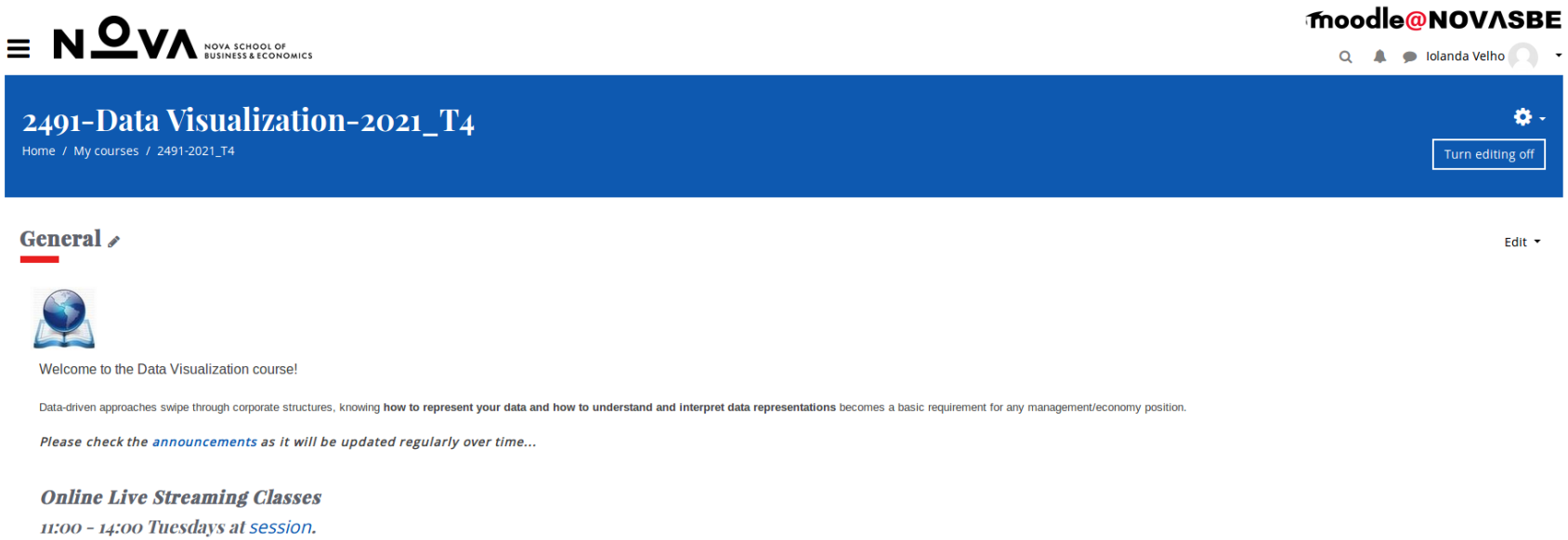
Iolanda Velho, Ph.D.
Nova School of Business and Economics

April 6th, 2021

NOVA SCHOOL OF BUSINESS & ECONOMICS

Moodle

- All course materials will be available on moodle (https://moodle.novasbe.pt/)
- Please enroll the course **2491_A-Data Visualization TA-2021_T4** with the password: **DV_2021_T4**



Course instructor

Iolanda Velho

iolanda.velho@novasbe.pt

Invited Assistant Professor, Nova SBE

M.S. in Biomedical Engineering at FCT/Universidade Nova de Lisboa

Ph.D. in Computational Engineering at Instituto Superior Técnico

Teaching Assistant

Alessandro Gambetti

gambetti.alessandro@novasbe.pt

Data Scientist & Researcher at Nova SBE

M.S. in Finance at Nova SBE

B.S. in Economics & Finance at Alma Mater Università di Bologna

Agenda

1. Welcome to the course
2. Syllabus
3. Setting up your environment

Syllabus

- **Course structure:**
  1. Introduction and setting up environments/installing libraries
  2. Why data visualization
  3. Story telling (because you need it for your project)
  4. Univariate data
  5. Multivariate data – categorical vs numerical
  6. Multivariate data – numerical variables and dimensionality reduction techniques
  7. Time series
  8. Graph representations
  9. Interactive visualizations
  10. Dashboards

- **Class hour:**
  - Tuesdays 11:00 – 14:00

Office hour:

- Monday / Wednesday / Thursday at 15:00 – 16:00
- Sessions with Alessandro Gambetti

Common Questions

- What if I have questions regarding course, assignment, or anything
  - Feel free to contact me via email
  - Project questions should be sent at least 24 hours before the deadline to ensure a timely response. Please CC the TA so that we can all stay coordinated, and include the course code 2491 in the subject line of your emails

- What is the late project policy?
  - The project need to be submitted by the due date and will lose 20% of points for each late day.
  - Example: if the project is late for 3 days, student may get 40% of points at maximum. If the assignment is late for 6 days, assignment will be returned without evaluation.

- Can I collaborate with others?
  - You are encouraged to discuss general approaches and clarification questions with your fellow students.
  - However, **DO NOT COPY** from another student's project

Assessment

Section	Description / Dates	Proportion
3 In-class quizzes	12 minutes 10 questions 20 Apr / 4 May / 11 May	20%
1 Project	1st Delivery: 18 April (proposal) 2nd Delivery: 16 May (final)	30%
Final exam	27 May	50%

Project Information

Objectives

Provide hands on experience using the tools and libraries covered during the course.

Provide an opportunity for students to develop and practice storytelling.

Evaluation

1) Project proposal - 20% of the final project grade
2) Adequate use of different representations - 40% of the final project grade
3) Proper storytelling - 30%
4) Non-obvious findings/advanced representations - 10%

Project Proposal

- **A project proposal must answer three questions:**
  1. what is the business problem you will be addressing,
  2. what data will you use and why is it relevant for business question,
  3. overview of the main analysis you will do, and how do they make the bridge between the data and the business questions.

- And it should make clear what you plan to deliver in the second delivery.

- **Important:** You do not need to follow the proposal blindly, i.e., if during the project you get interesting results in a different direction that you had planned, **it is more than ok to change your plan, if you justify your choices**.

Remaining Criteria

- **Adequate use of different representations**

You will choose a dataset and a story to tell on top of the dataset. I expect to find 2 plots per student, i.e., in a group of 2 students, you should have 4 plots (more is fine, less than 4 is not fine)

- **Adequate use of storytelling**

Tell a story that connects the different visualizations, explaining the insights you gathered from the data

- **Originality**

Do something different – pick your own dataset, get data from other sources, go for a more sophisticated representation. (Include your ideas in the proposal to ensure that it will get you the 10%)

Ideas for projects and datasets (1)

Car prices

Dataset - car specifications, e.g., power, size, brand, and price

Story Idea - You have a client that is designing a new car. You want to show him that there are:

- a few features that he could twick that would allow him to sell the car for a higher price.
- You can also show him how the brand value has been evolving, and what is your expectation for the near future in terms of price evolution

link to source: https://www.kaggle.com/goyalshalini93/car-data

Restaurants

Dataset - Zomato dataset, with restaurants price range, reviews, location, etc

Story Idea - You are working for a famous chef that wishes to open a new restaurant somewhere in Europe. You want to show him the differences in type of restaurants and prices across the continent to help him pick a location. You can also collect extra data, e.g., the size of the city to guess the actual market, or the average income to estimate costs

link to source: https://www.kaggle.com/damiensbeneschi/krakow-ta-restaurans-data-raw

Ideas for projects and datasets (2)

Bitcoin

Dataset - cryto currency daily values

Story Idea - You are wondering whether or not to invest on bitcoin. Check the evolution of bitcoin value and compare it with other crypto currencies

link to source: https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

Covid-19

Dataset: cases evolution across regions

Story Idea - You were tasked with creating a dashboard for showcasing the evolution over time and region of Covid-19. You can also add the overall sentiment in each place using a geolocated tweet dataset

link to source: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv
https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset

Ideas for projects and datasets (3)

Insurance

Dataset - Per policy: state/zip code, deductible, premium, age, income, fraud

Story Idea - You are working with a network of promoters and wish to generate new leads for them, i.e., good products to sell to a specific demography and location

link to source: https://www.kaggle.com/roshansharma/insurance-claim

Bank customer base

Dataset: Dataset mapping customer characteristics, e.g., income, age, balance, to default

Story Idea - Your client, one of the largest retail banks in Europe, asked you to characterize the customer basis, searching for segments more prone to default

link to source: https://www.kaggle.com/krishna0201/banckingmarket

Ideas for projects and datasets (4)

Cinema

Dataset - 4 tables with list of movies, actors, release years, duration, genre and ratings (NOTE: you will need to merge tables!)

Story Idea - You decided to try your hand as a producer. Before deciding the genre and who to cast, you want to take a closer look into what has been done and what are the current trends

link to source: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

Stock market

Dataset: Daily stock exchange prices for the main stock markets in US

Story Idea - You want to invest in a cool market, e.g., new technologies. But you need some money to start your journey. Check trends, find correlations between stocks, and prepare a pitch to get funding

link to source: https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs


jupyter
Environment
Setup

Anaconda, what it is?

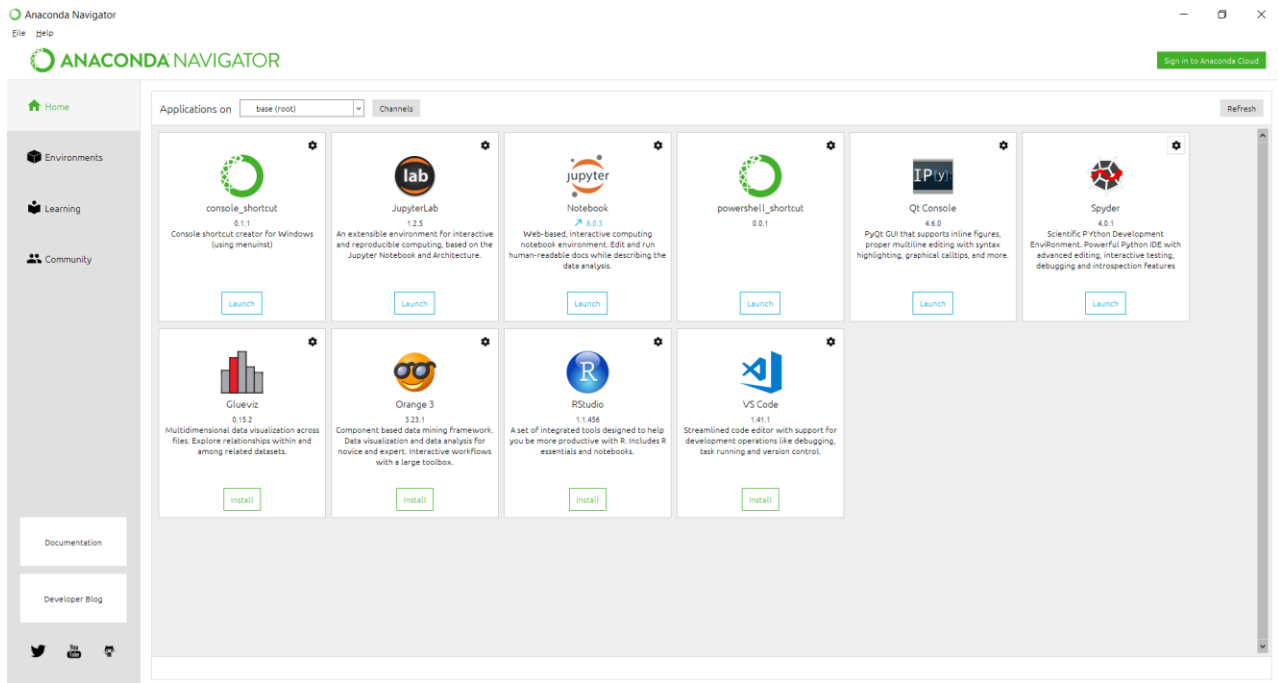
1. Open source **tool**
2. Open-source **distribution** of the **Python** and R programming languages for *scientific computing*

Anaconda, what about its features?

- Anaconda aims to simplify *package* management and *deployment*

- Python's package versions are managed by the package management system *conda*

- When download it also has some user interfaces aim for both **Python** and **R** *interpreters* (we will be using **Jupyter Notebook** , but Anaconda distribution also install **Spyder** and allows for **RStudio** download).



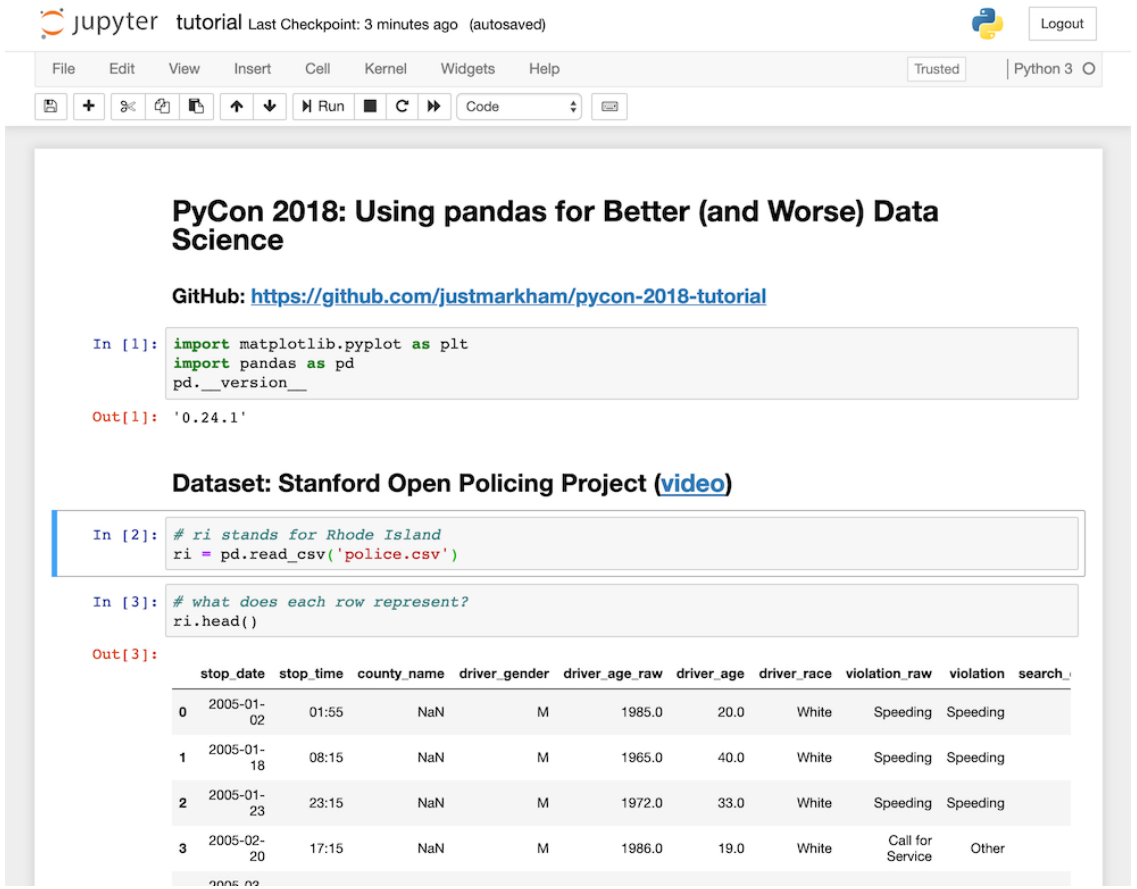
Open jupyter notebook

Jupyter

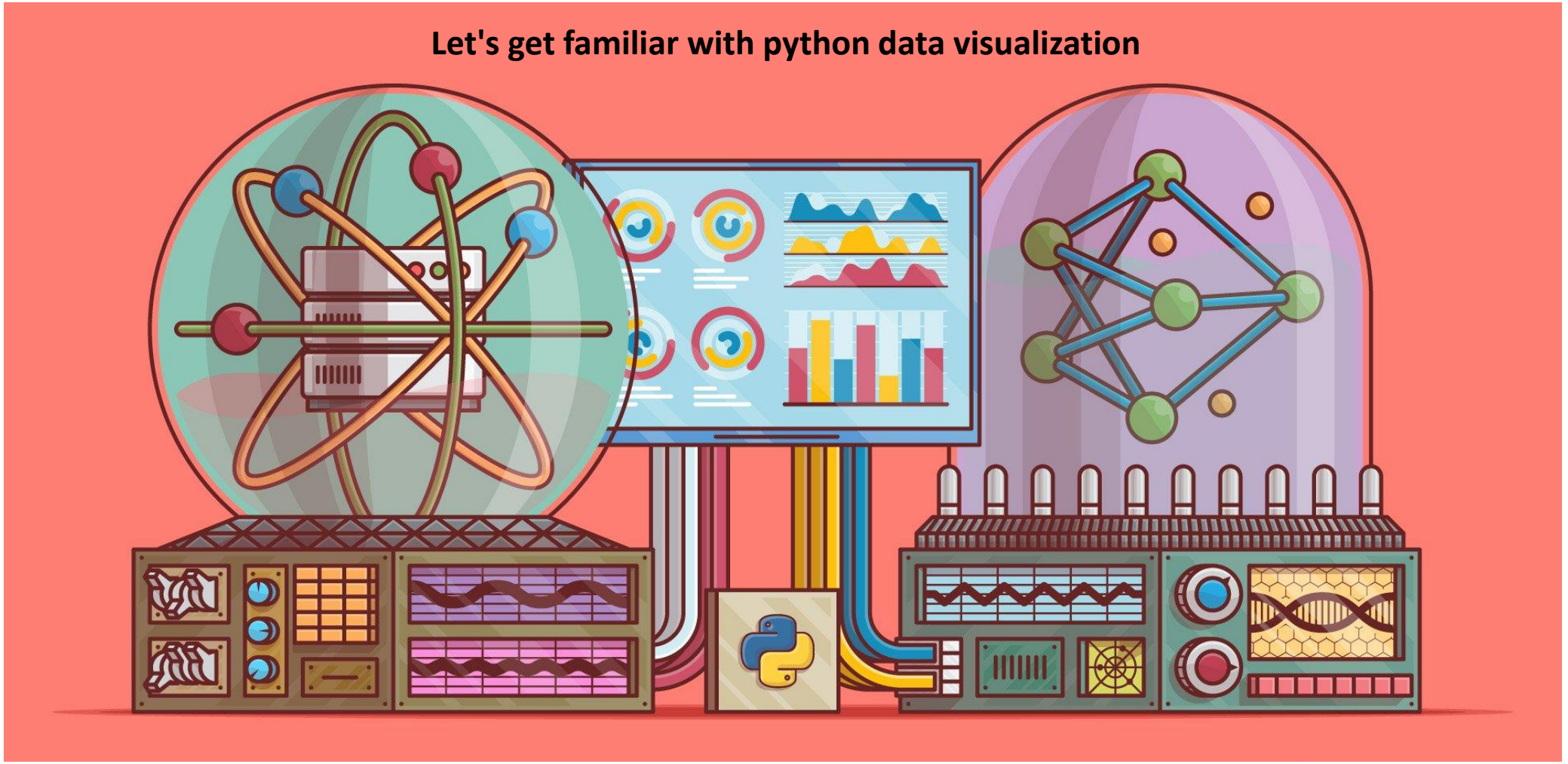
- Open source **web-application** that allows you to create and share documents that contain live code, equations, visualizations and narrative text

Jupyter use cases:

- Data cleaning and transformation
- Numerical simulation
- Statistical modeling
- Data visualization
- Machine learning
- And much more



Next



Agenda

1. Motivation
2. Theory of visualization
3. Science of perception

Motivation

What is the color of #thedress?

White and gold? Black and blue?

Motivation

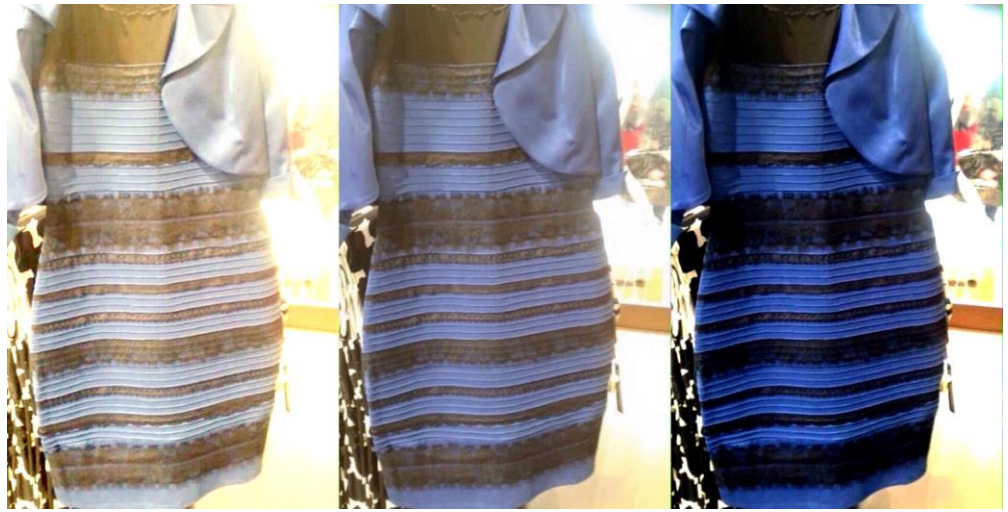
What is the color of #thedress?

Kim Kardashian West
@KimKardashian

What color is that dress? I see white & gold. Kanye sees black & blue, who is color blind?

4:41 AM - Feb 27, 2015

36.3K 26.5K people are talking about this



Justin Bieber
@justinbieber

And for everyone asking I see blue and black

4:21 AM - Feb 27, 2015

151K 160K people are talking about this

Mariah Carey
@MariahCarey

Of course it's white and GOLD dahhlings!! #TheDress

6:29 PM - Feb 27, 2015

2,336 2,551 people are talking about this

Lady Gaga
@ladygaga

@thecolorated its periwinkle and sand 💣

3:51 AM - Feb 27, 2015

9,568 7,484 people are talking about this

Taylor Swift
@taylorswift13

Follow

I don't understand this odd dress debate
and I feel like it's a trick somehow.
I'm confused and scared.
PS it's OBVIOUSLY BLUE AND BLACK

RETWEETS 113,207 FAVORITES 156,359

6:14 PM - 26 Feb 2015

© Twitter

Motivation

- A research survey showed that overall, 57% of interviewed subjects described the dress as blue/black (B/K); 30% as white/gold (W/G); 11% as blue/brown (B/B); and 2% as something else
- Differences between the two groups arose mainly from **differences in lightness**, rather than **chromaticity of the colors** they adjusted to match the dress
- Ambiguity arises in the case of this particular image because the distribution of **colors within the dress closely matches the distribution of natural daylights**
- **Asymmetries in color perception from natural lightning** has striking effects on the appearance of images

Science behind #thedress

Lafer-Sousa et al. (2015), Striking individual differences in color perception uncovered by 'the dress' photograph, Current Biology, 25:13, 545-546

Data Visualization for Human Perception

- Data visualization is the graphical display of abstract information for two purposes: **sense-making (also called data analytics)** and **communication**

- Visualization can work as a cognitive tool, and become a powerful strategy in the construction of knowledge using **human visual perception and cognitive capacities**

- **Data analytics**: map data onto visual elements in a way that will help humans to effectively perceive and reason about the structure in the data (that were previously hidden, either because of data complexity or data volume)

- **Communications**: when data is presented in a abstract form which our brains interpret through the use of verbal processing, becomes visible and understandable when communicated visually

Card et al. (1999) *Information Visualization: Using Vision to Think,* Morgan Kaufmann

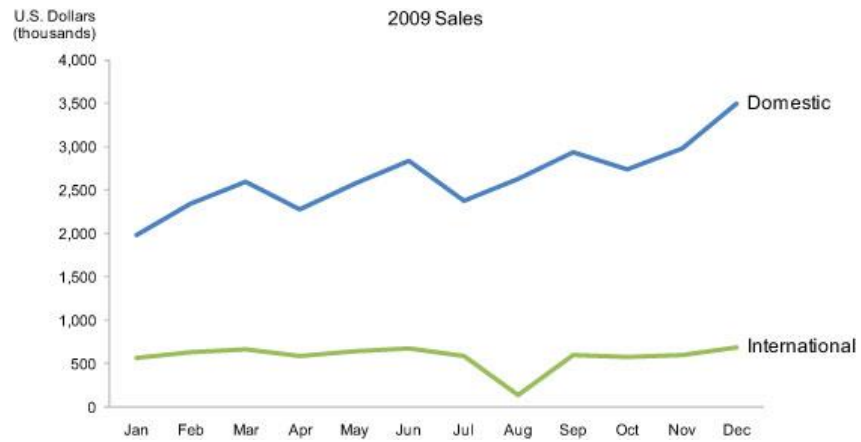
Illustrations for Quantitative Information

- This table does two things extremely well: it expresses these sales values precisely and it provides an efficient means to look up values for a particular region and month

2009 Sales (thousands of U.S. $)

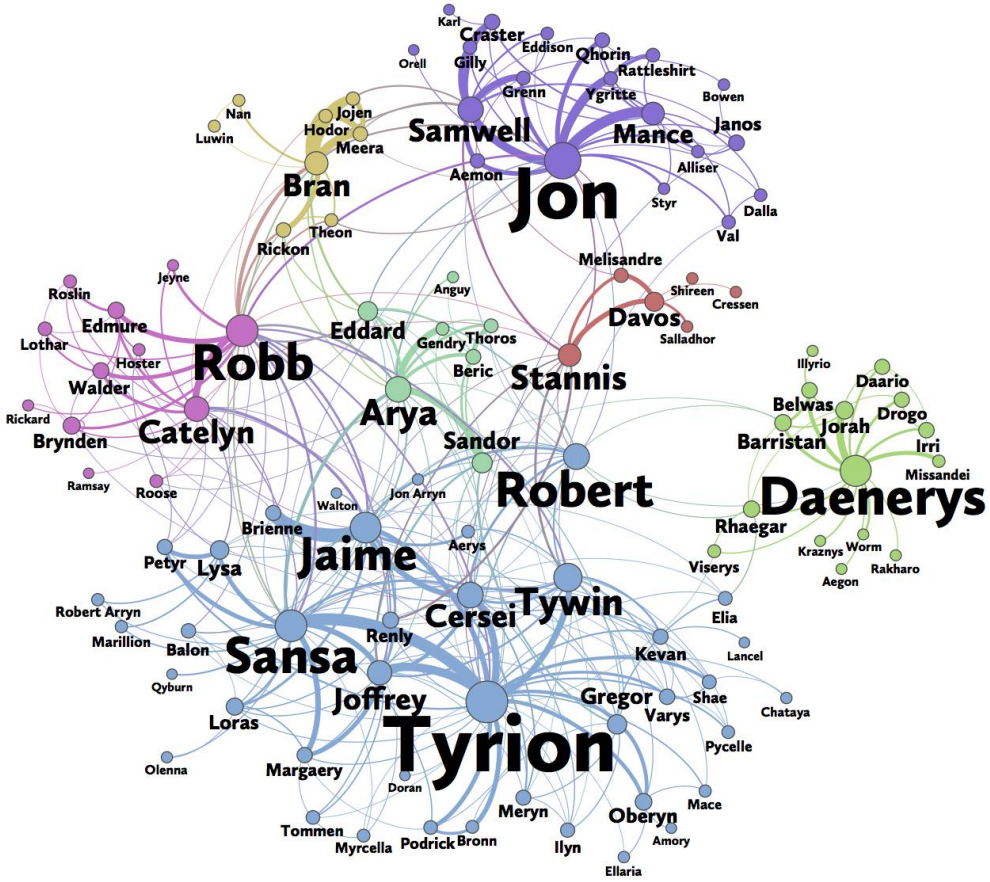
Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493	31,783
International	574	636	673	593	644	679	593	139	599	583	602	690	7,005
Total	2,557	2,979	3,266	2,876	3,218	3,517	2,975	2,773	3,537	3,322	3,585	4,183	38,788

- But if we're looking for patterns, trends, or exceptions among these values, if we want a quick sense of the story contained in these numbers, or we need to compare whole sets of numbers rather than just two at a time, this table fails.



Illustrations for Qualitative Information

- Each character in the Game of Thrones can be represented as a vertex
- A link between characters means that they interacted, spoke of one another, or that another character spoke of them together (a link between two characters whenever their names appeared within 15 words of one another)
- Qualitative information can also be transformed to represent the complex semantics of the data



Game of Thrones Characters Network

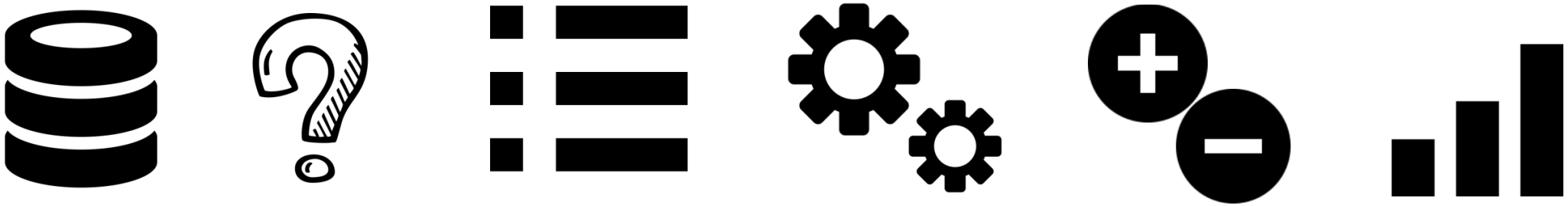
Goals for Visualization

- Exploratory Analysis:
  - Starting point: we intend to discover new knowledge from the input data
  - Process: explore the obtained visual representation and look for signs that could suggest indications of particular tendencies and relations
  - Results: visualization of data that can form the basis of a hypothesis

- Confirmatory Analysis:
  - Starting point: we already have hypothesis and objective about the data
  - Process: goal-oriented visual examination of the hypotheses
  - Results: determine evidence for the acceptance or the rejection of the pre-formulated hypothesis

- Presentation:
  - Starting point: facts to be presented are fixed a priori in graphical display
  - Process: choice of appropriate presentation techniques
  - Results: high-quality visualization of the facts to present the relationship, structure, behavior and other intrinsic characteristics of the data involved

Steps of Data Visualization



Data Type **Questions** **Requirements** **Preprocessing** **Pro/Cons** **Design**

Find out what kind of data you have

Which questions shall be answered by the data?

Derive requirements from these tasks

Preprocess data (missing data, outliers, transformations, …)

Consider the pros and cons of different visualization techniques (effectiveness, familiarity …)

Design a visualization system

Data Visualization Pipeline

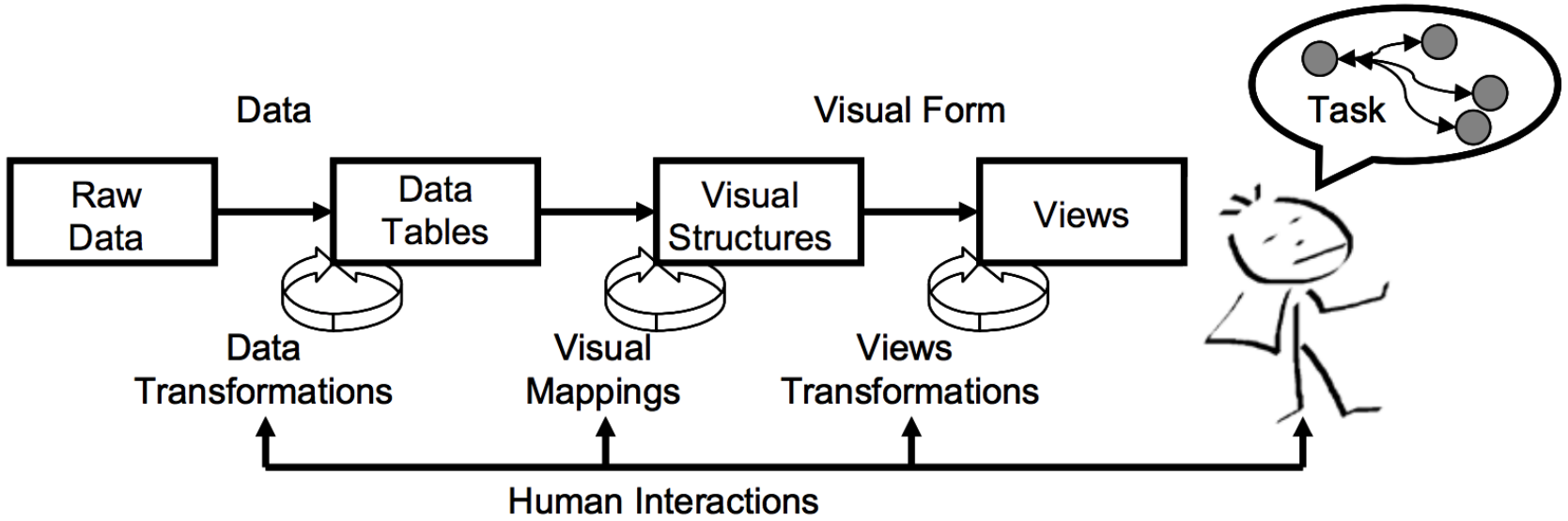


Figure 1 - Reference model for visualization (Card 1999).

In the first phase **raw data is input after the organization of some formatting operations** and/or normalizations, according to a logical and structured representation, for example, in the form of tables. In the mapping phase, the **association between data and graphic representations** takes place: a visual structure that supports the data previously formatted to be used in the rendering step, finally, in this last step**, the visual representation of the data is drawn**.

Data Models / Conceptual Models

- Data models are formal descriptions using mathematical operations
  - Example: integers with + and x operators

- Conceptual models are mental constructions that include semantics to support reasoning
  - Example: stock prices of the last year

- Data vs. conceptual models
  - 1D float numbers vs. temperatures
  - 3D vector of float numbers vs. spatial locations

Data Types

Nominal

- No quantitative relationship between categories
- Classification without ordering

➢ No intrinsic order

e.g. gender, nationality

Ordinal

- Attributes can be rank-ordered
- Distances between values do not have any meaning

➢ Ordered in a sequence

e.g. education, health conditions

Numerical

- Attributes can be rank-ordered
- Distances between values have a meaning
- Mathematical operations are possible

e.g. age, height/weight

C. Ware. 2012. Information Visualization: Perception for Design. Chapter 1.

Relational Data Models

- Data is represented as a table (or relation)
- Each row (or tuple) represents a record
  - Each record is a fixed-length tuple
- Each column (or field) represents a variable
  - Each field has a name and a data type
- A table's schema is the set of names and types
- A database is a collection of tables (relations)
- The data is in structured form and stored in different formats
  - comma separated values files
  - json files
  - database tables

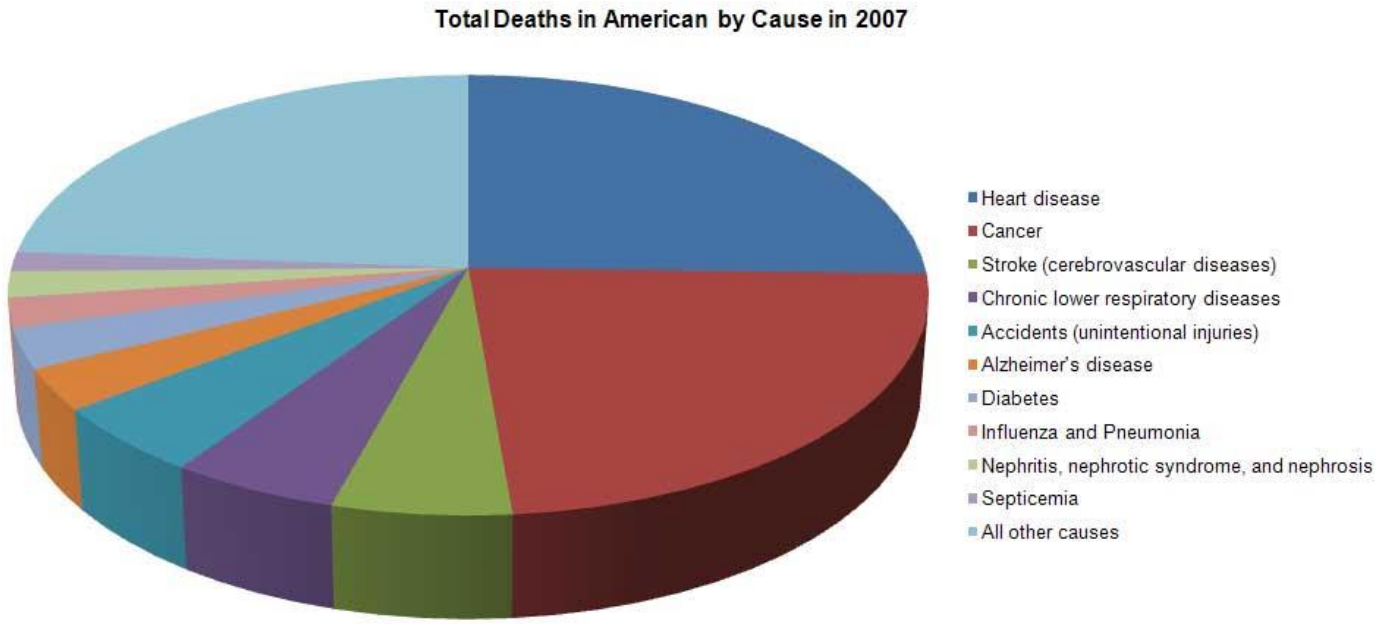
Example: Primary causes of death in America

- Data model
  - Cause of death
  - Integer numbers
- Conceptual model
  - Number of death per cause
- How to translate this data model into image model?
  - Clearly indicates how the values relate to one another
  - Represents the quantities accurately.
  - Makes it easy to compare the quantities.
  - Makes it easy to see the ranked order of values, such as from the leading cause of death to the least.
  - Makes obvious how people should use the information - what they should use it to accomplish - and encourages them to do this.

Causes of Death	Deaths per Year
Heart disease	616,067
Cancer	562,875
Stroke (cerebrovascular diseases)	135,952
Chronic lower respiratory diseases	127,924
Accidents (unintentional injuries)	123,706
Alzheimer's disease	74,632
Diabetes	71,382
Influenza and Pneumonia	52,717
Nephritis, nephrotic syndrome, and nephrosis	46,448
Septicemia	34,828
All other causes	577,181
Total	2,423,712

Example: Primary causes of death in America

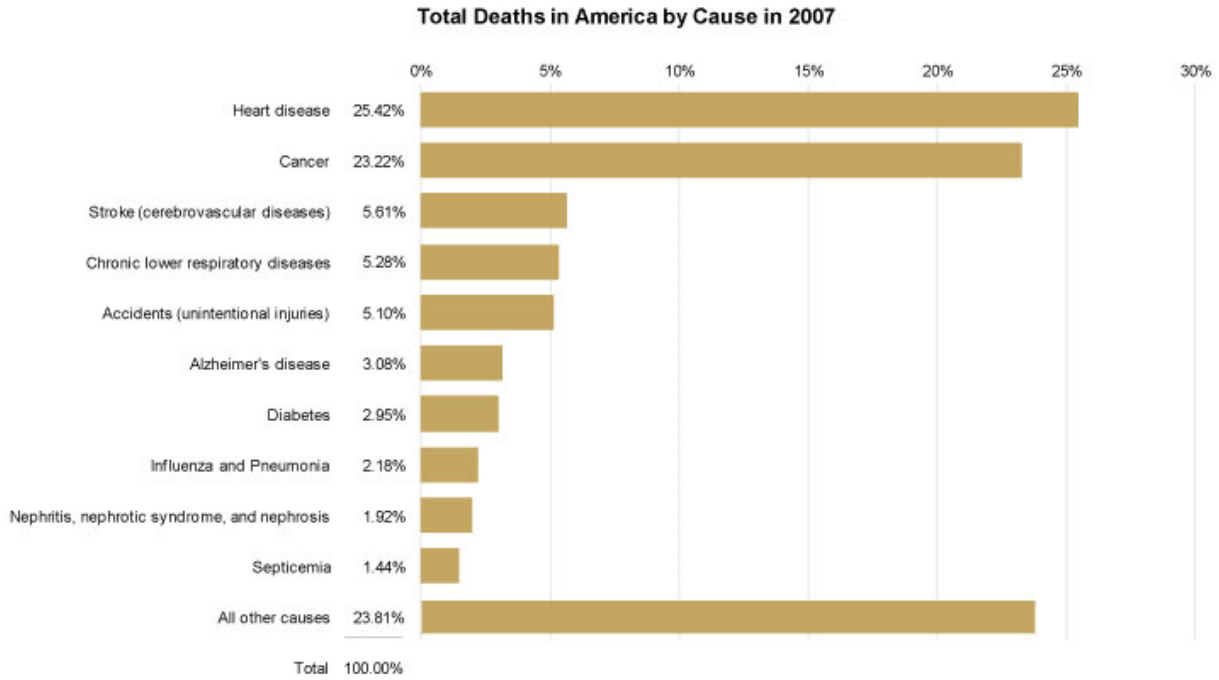
Pie Chart



- Pie chart is a circular graphic which is divided into slices to illustrate numerical proportion
  - *Clearly indicates the nature of the relationship?*
  - *Represents the quantities accurately?*
  - *Makes it easy to compare the quantities?*
  - *Makes it easy to see the ranked order of values?*
  - *Makes obvious how people should use the information?*

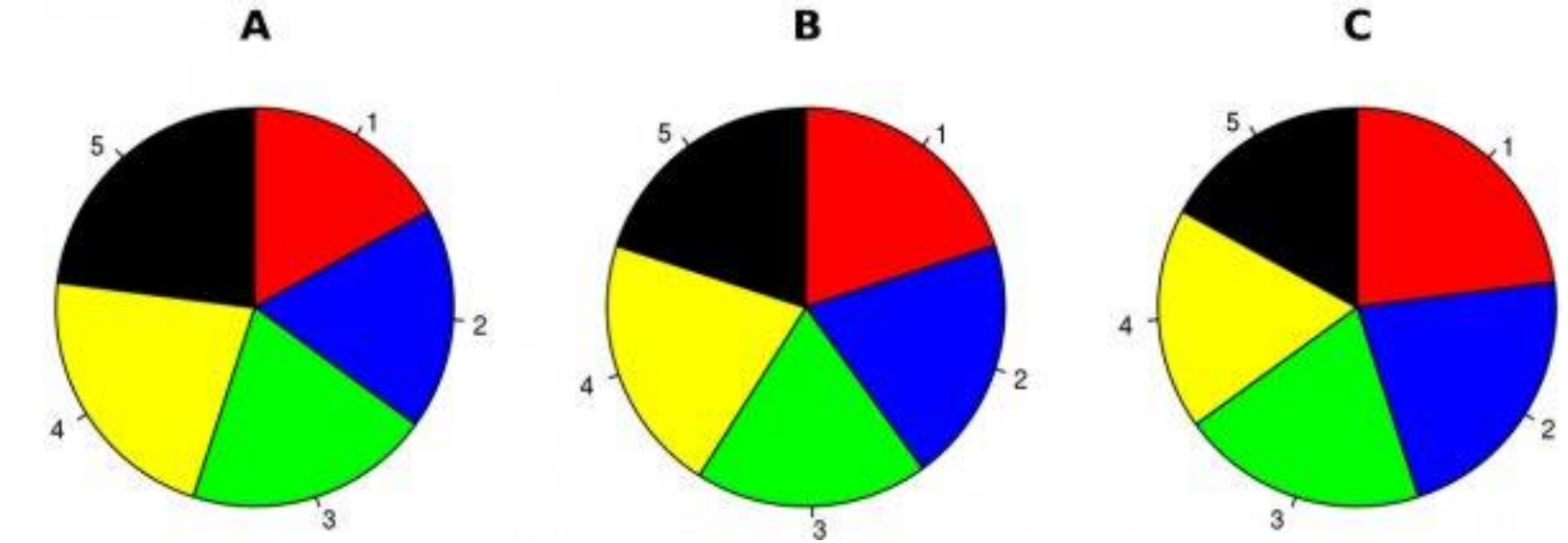
Example: Primary causes of death in America

Bar Plot



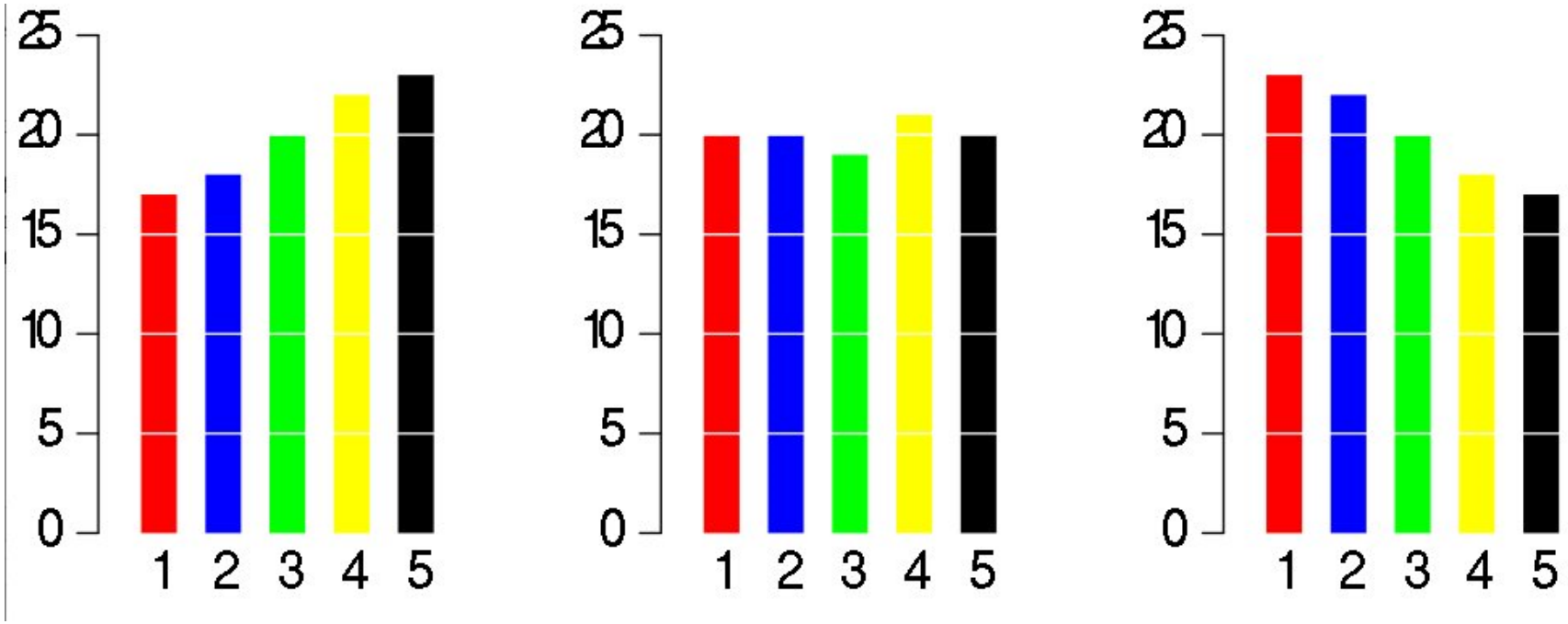
- Barplot shows the relationship between a **numerical variable** and a **categorical variable**.
  - *Clearly indicates the nature of the relationship?*
  - *Represents the quantities accurately?*
  - *Makes it easy to compare the quantities?*
  - *Makes it easy to see the ranked order of values?*
  - *Makes obvious how people should use the information?*

(Why) Pie charts are bad?



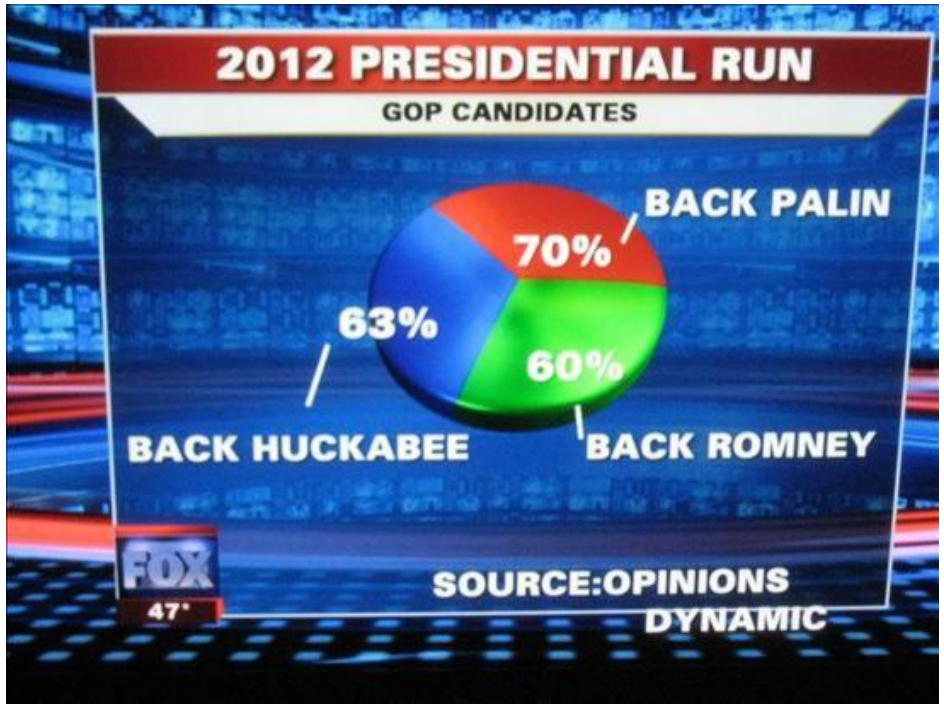
What can you conclude from the pie charts above?

(Why) Pie charts are bad?



What can you conclude from the bar plots above?

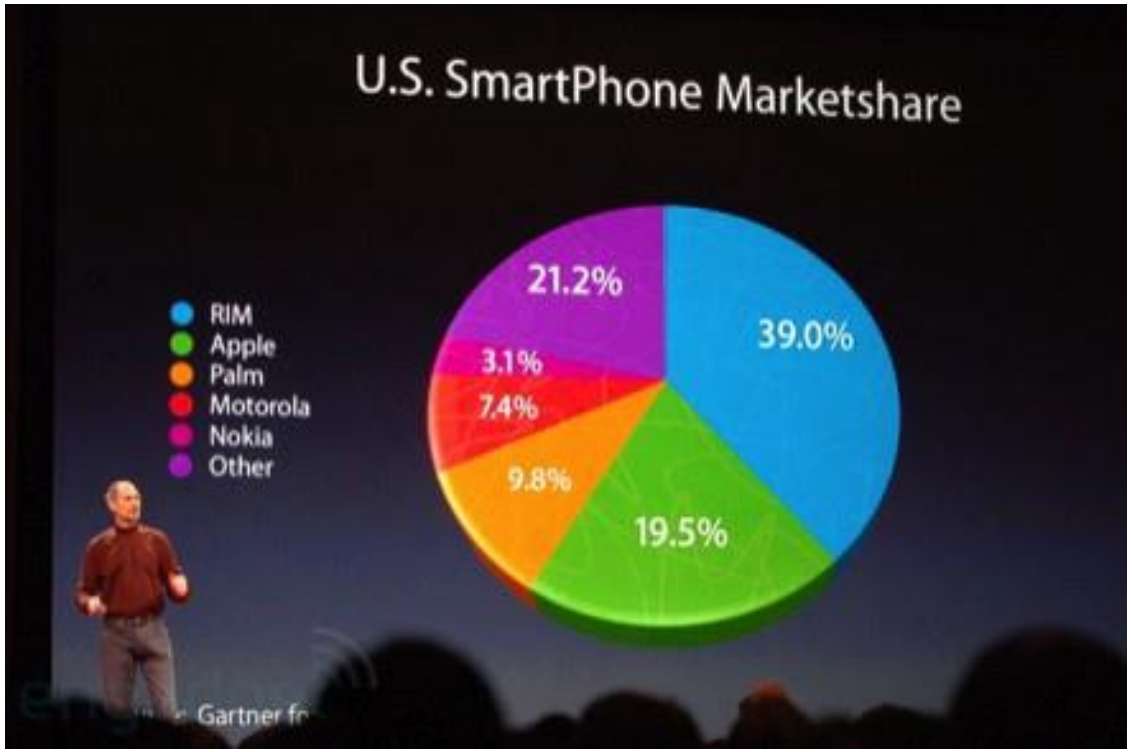
(Why) Pie charts are bad?



Sometimes it is even worse when you present the wrong proportions in the pie chart…

(Why) Pie charts are bad?

- Human brains have difficulty in **comparing the size of angles** and **reading accurate values without scale**
- Problems are exacerbated when making 3D pie charts (see right)
- Pie charts are not a recommended visualization, because it may **incur the wrong cognitive understandings** from human perception
- "**Pie charts are bad**" - Edward Tufte
- except:
  - only use them for a percentage breakdown
  - each slice represents a certain percentage out of 100%
  - order the slices in size to make it easier to read
  - never use a pie chart if it has more than 5 slices
  - never-ever make it 3D



Steve Jobs at MacWorld 2008

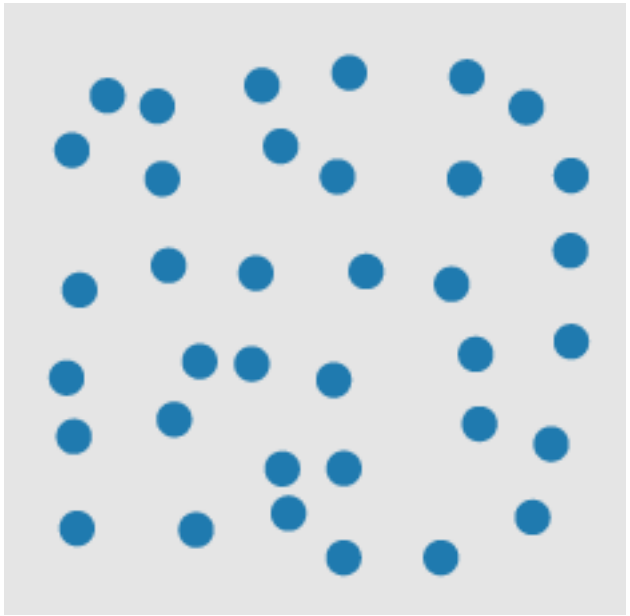
Science of Perception

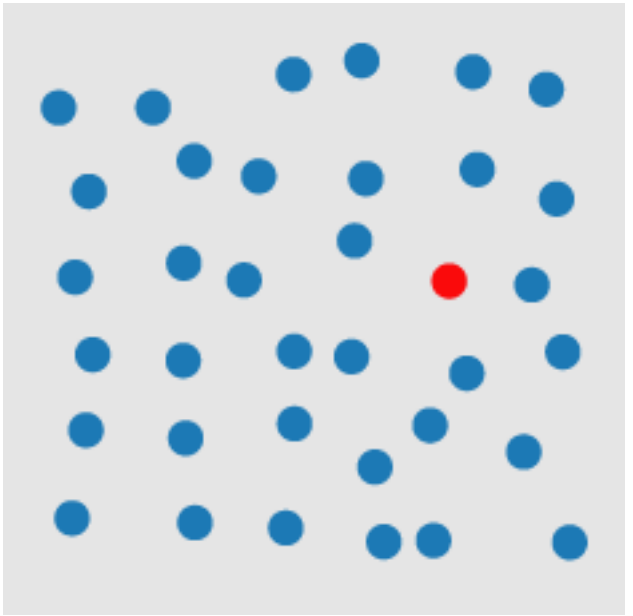
- Data visualization is effective because it **shifts the balance between perception and cognition** to take full advantage of the brain's abilities
- **How visual perception works**
  1. Light is reflected off an object that you're looking at and moves through the eyes
  2. It filters down into the retina in the brain via its photosensitive cells (rods and cones), where it is encoded as electrical signals
  3. Your brain now detects basic attributes, also known as **pre-attentive attributes**
  4. The brain performs more analysis and **encodes information within your memory** (iconic memory, working memory and long-term memory)
- **Pre-attentive visual processing**
  - Certain low level visual aspects are recognized before *conscious* awareness
- **Gestalt Laws**
  - The *tendency* to **perceive** elements as belonging to a group, based on certain visual properties
- **Visual Variables**
  - The different visual aspects that can be used to **encode** information

Pre-attentive Processing

- Perception of visual features managed by the **low-level visual system**
- Extremely fast: **< 200 msec** (the eyes take > 200 msec to initiate movement) => processed in parallel
- Pre-attentive attributes include:
  - Color
  - Length
  - Width
  - Orientation
  - Shape
  - Size
  - Enclosure
  - Hue
  - Intensity/Shade
  - Position

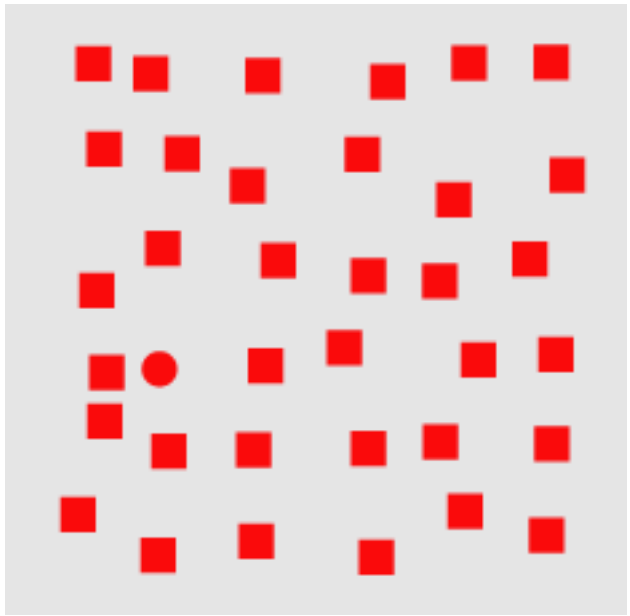
Pre-attentive Processing (Color)

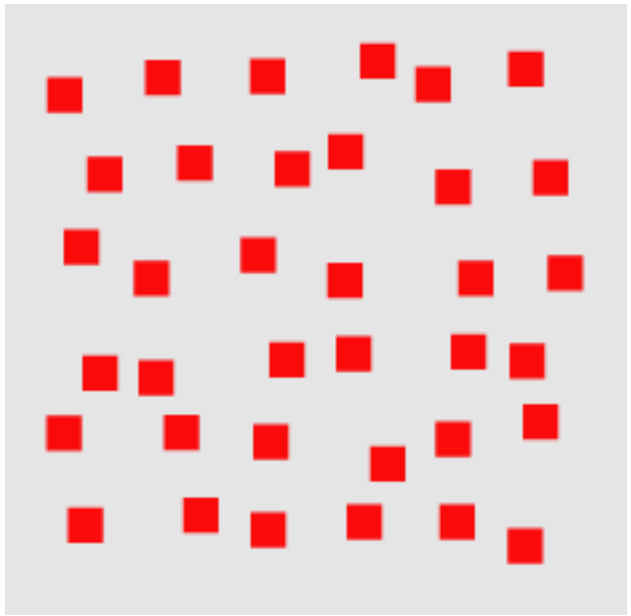




How fast can you spot the target red circle?

Pre-attentive Processing (Shape)

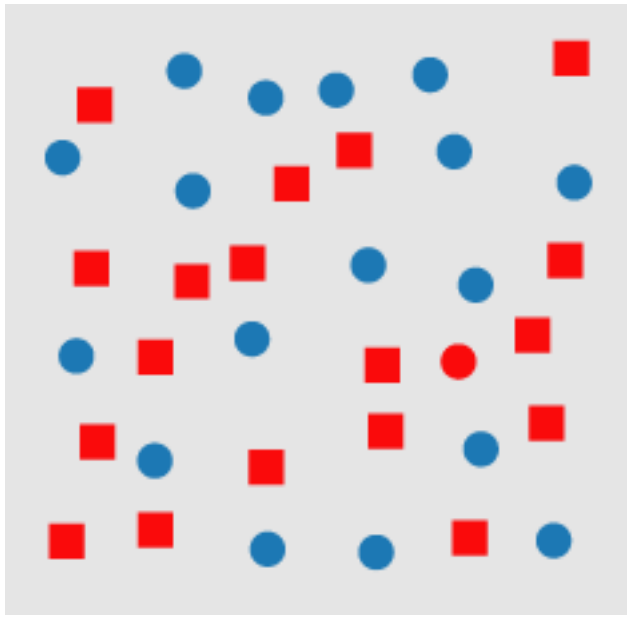


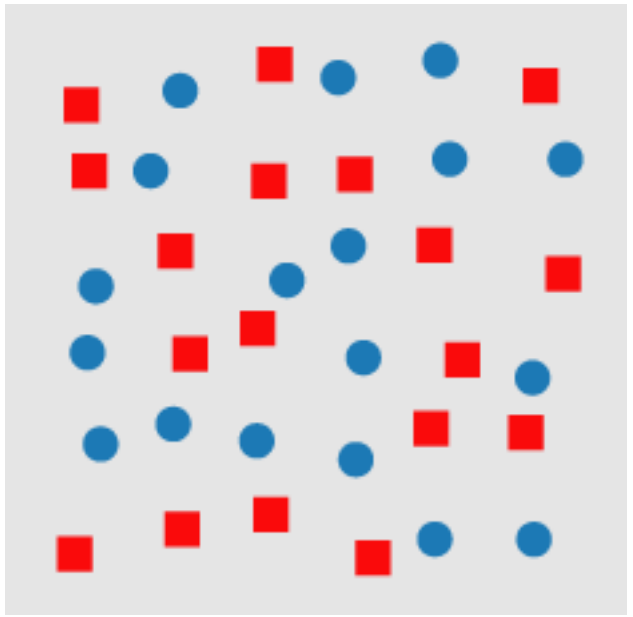


How fast can you spot the target red circle?

Serial Pre-attentive Processing (Shape)

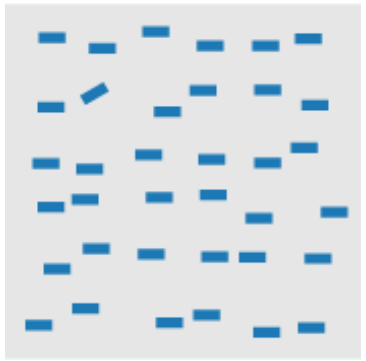
Serial Pre-attentive Processing (shape)



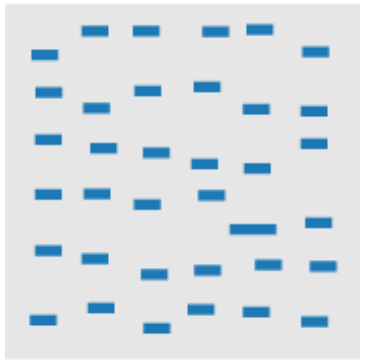


How fast can you spot the target red circle?

Many more Pre-attentive Processing

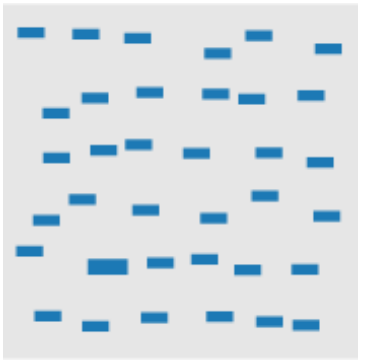


line (blob)
orientation

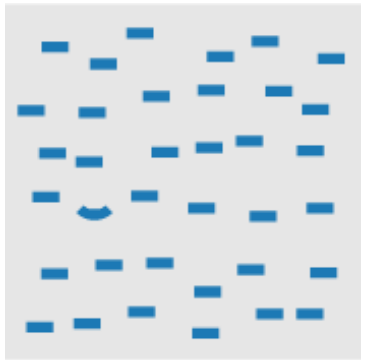


length, width

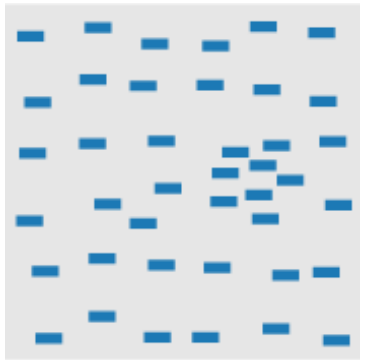
closure



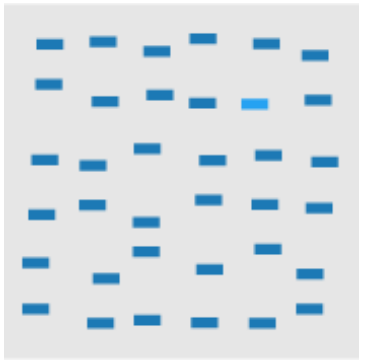
size



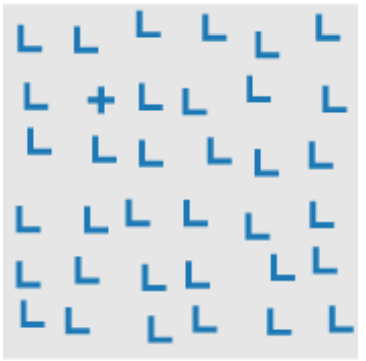
curvature



density



intensity



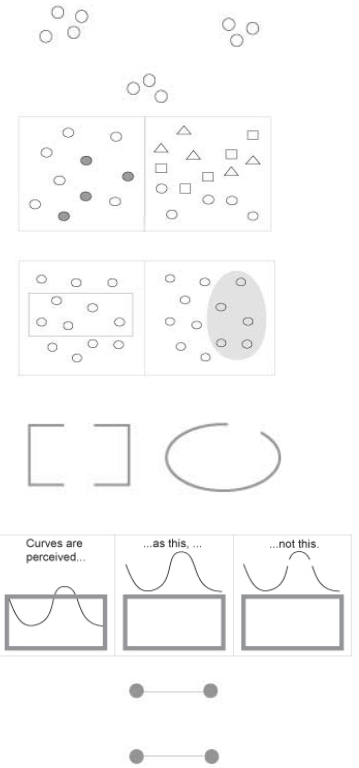
intersection

https://learnforeverlearn.com/preattentive/

Gestalt Laws

- Perceptual laws about how humans **group visual objects together** to form visual entities

Property	Description
Proximity	Objects that are close together are perceived as a group.
Similarity	Objects that share similar attributes (e.g., color or shape) are perceived as a group.
Enclosure	Objects that appear to have a boundary around them (e.g., formed by a line or area of common color) are perceived as a group.
Closure	Open structures are perceived as closed, complete, and regular whenever there is a way that they can be reasonably interpreted as such.
Continuity	Objects that are aligned together or appear to be a continuation of one another are perceived as a group.
Connection	Objects that are connected (e.g., by a line) are perceived as a group.



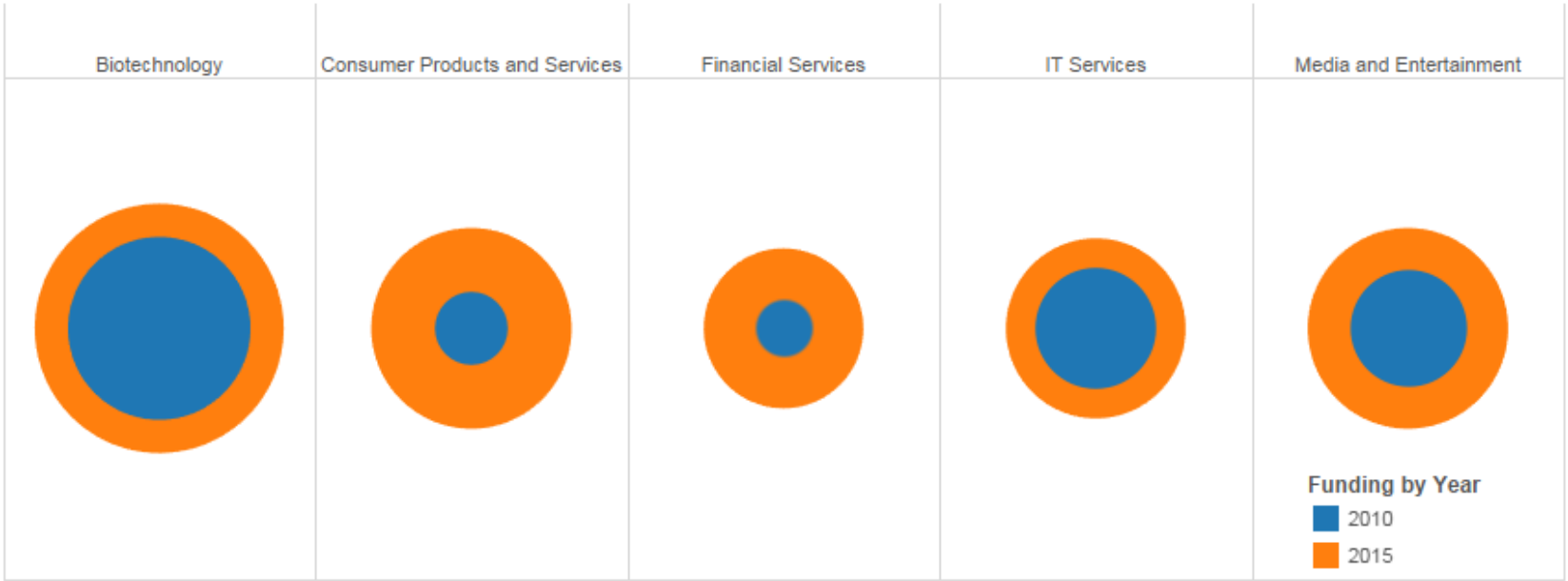
Gestalt Laws – Figure and Ground

- **figure**: object like, perceived to be in the foreground
- **ground:** whatever lies behind the figure
- The figure is distinguished from the background by Gestalt laws
- The object is only perceived as a figure after being separated from the background

face or vase?

Choosing the right visual variables

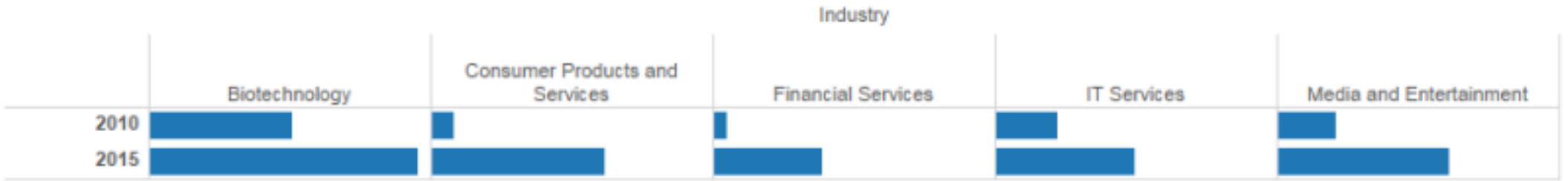
How much venture capital funding was received by various industries in 2010 versus 2015?



- Which industry received the greatest amount of venture capital funding in 2015?
- Which industry received the second most amount of venture capital funding in 2015?
- What percent of funding did the *Biotechnology* industry receive in 2010 compared to 2015?
- What percent of funding did the *Media and Entertainment* industry receive in 2010 compared to 2015?

Choosing the right visual variables

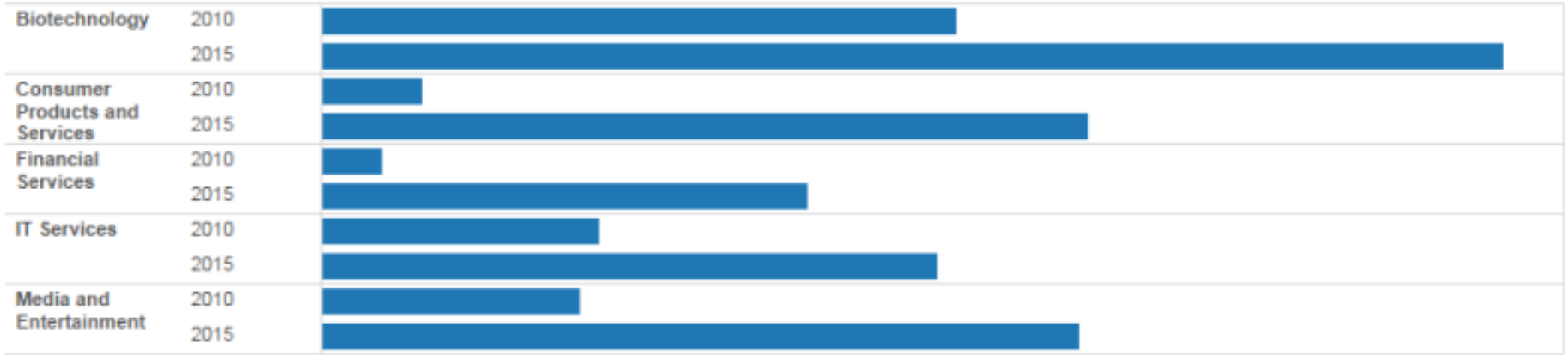
How much venture capital funding was received by various industries in 2010 versus 2015?



- Which industry received the greatest amount of venture capital funding in 2015?
- Which industry received the second most amount of venture capital funding in 2015?
- What percent of funding did the *Biotechnology* industry receive in 2010 compared to 2015?
- What percent of funding did the *Media and Entertainment* industry receive in 2010 compared to 2015?

Choosing the right visual variables

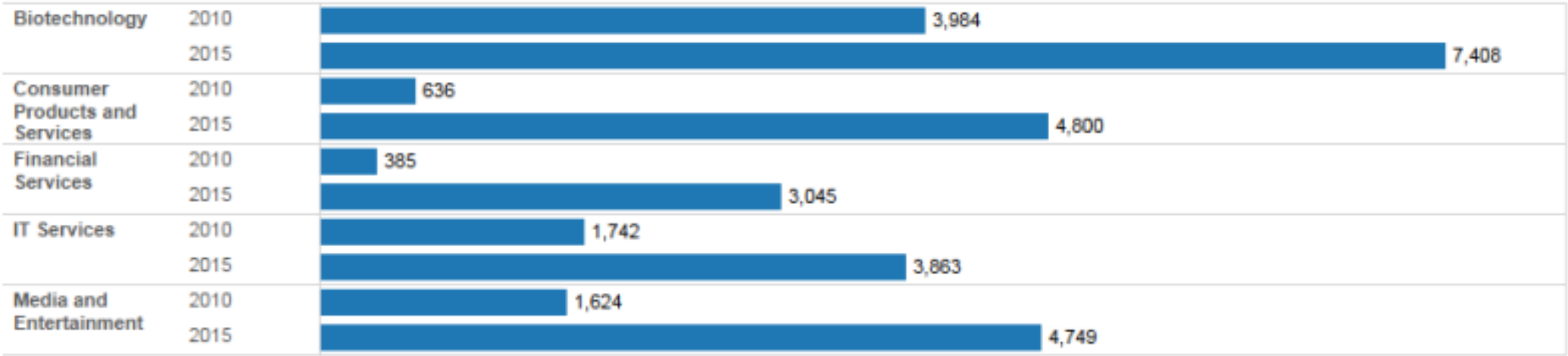
How much venture capital funding was received by various industries in 2010 versus 2015?



- Which industry received the greatest amount of venture capital funding in 2015?
- Which industry received the second most amount of venture capital funding in 2015?
- What percent of funding did the *Biotechnology* industry receive in 2010 compared to 2015?
- What percent of funding did the *Media and Entertainment* industry receive in 2010 compared to 2015?

Choosing the right visual variables

How much venture capital funding was received by various industries in 2010 versus 2015?



- Which industry received the greatest amount of venture capital funding in 2015?
- Which industry received the second most amount of venture capital funding in 2015?
- What percent of funding did the *Biotechnology* industry receive in 2010 compared to 2015?
- What percent of funding did the *Media and Entertainment* industry receive in 2010 compared to 2015?

Priorities table in relation to perceivable visual attributes (Mackinlay 1988)

Perception	Quantitative Data	Ordinal Data	Nominal Data
More	Position	Position	Position
	Length	Density	Color hue
	Angle	Color Saturation	Texture
	Slope	Color hue	Connections
	Area	Texture	Containment
	Volume	Connections	Density
	Density	Containment	Color Saturation
	Color Saturation	Length	Shape
	Color hue	Angle	Length
	Texture	Slope	Angle
	Connections	Area	Slope
	Containment	Volume	Area
Less	Shape	Shape	Volume

In grey are the attributes that are not considered relevant to the type of data in question

Summary

- Human brain relies on the eyes to supply an image of the outside world, but there are many processing steps between the **translation of light energy into electrical impulse**s that happens in the eye and **the neural activity** that corresponds to a conscious perception of the outside world

- Our perception corresponds to the brain's best guess of **what is going on in the outside world**, not necessarily to **the way things actually are**.

- Many parts of the brain contribute to any given perception of forms, motions, colors, shapes, also known as **preattentive attributes**

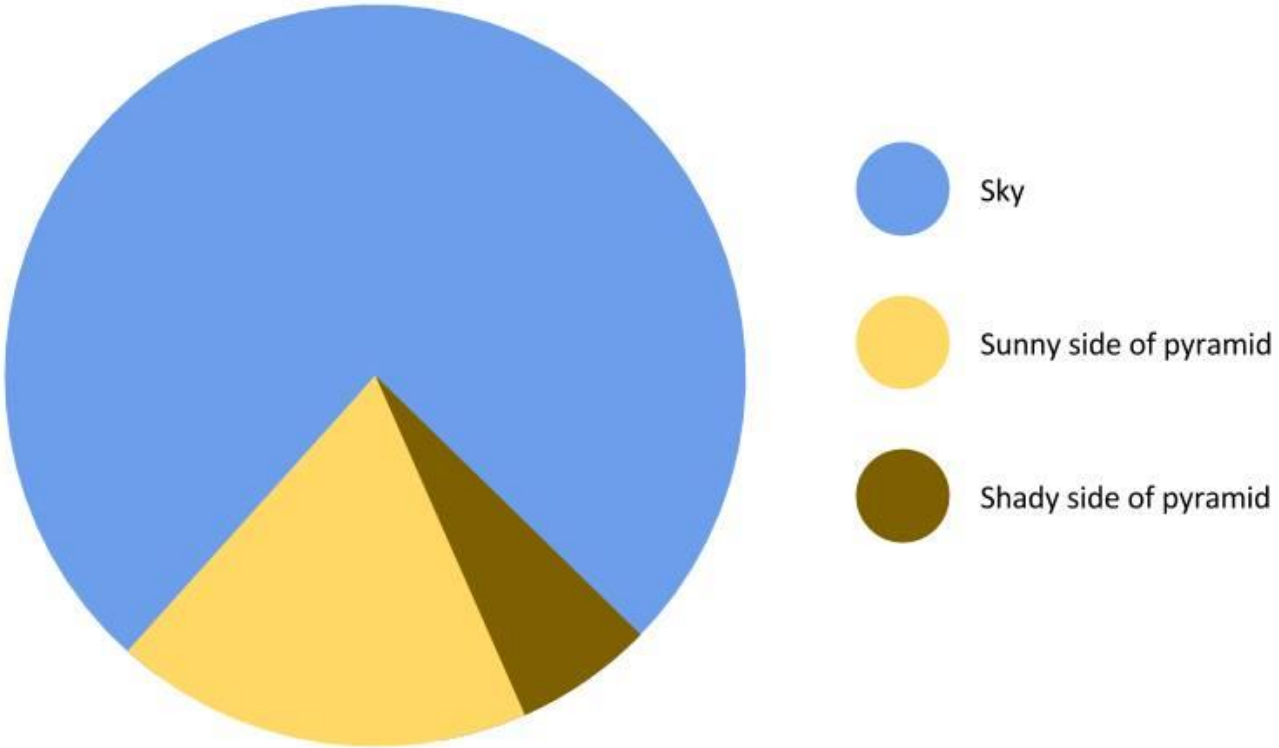
- The brain then performs analysis and **encodes information** within the memory

Principles of visualization design

- Choosing effective visual encodings requires knowledge of visual perception
- Formal specification
  - Data model: relational data; Nominal, Ordinal, Numerical types
  - Image model: visual encoding channels
  - Encodings map data to visual variables

- Visual features/attributes
  - Individual attributes often pre-attentive
  - Multiple attributes may be separable or integral

- Gestalt principles provide high-level guidelines
- Tell the truth and nothing but the truth (don’t lie, and don’t lie by omission)
- Use encodings that people decode better (where better = faster and/or more accurate)



Questions?

THE END

See you all next class