Universidade de Aveiro
Departamento de Electrónica,
Telecomunicações e Informática

Evaluation in Visualization

Beatriz Sousa Santos, Universidade de Aveiro, 2025

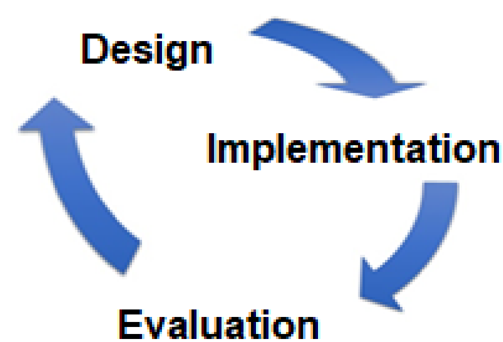
How can we produce a Visualization/ Visual data exploration app?

- There are **principles** (derived form human perception and cognition) and many **visualization techniques** (we study some…)

- To obtain **efficacy** a **Human-centered approach** is fundamental, involving:
  - a correct definition of **goal** and **user tasks**

    (the **questions**!)
  - apply adequate methods and **evaluate**

in **several iterations** until the goals are satisfied



How can we evaluate?

- **Many methods** can be used to evaluate a Visualization application (some specifically developed, others adapted)

- Evaluation methods from other disciplines may and have been adapted and used to evaluate Visualization applications, as methods from:
  - Human- Computer Interaction
  - Image Processing
  - S/W Engineering

- Applications to visually explore data are interactive and should be **usable**

- **Usability** is, according to ISO 9241-11:

"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"

- How to measure it??

- We can use methods used in Human-Computer Interaction

Usability evaluation Methods

- Methods used in **Human-Computer Interaction** to evaluate **usability** may be used to evaluate visualizations and visual data exploration applications and some have been adapted

- **Analytical** (without users)
  - Heuristic Evaluation
  - Cognitive Walkthrough
  - Model based methods
  - Review methods

  ...

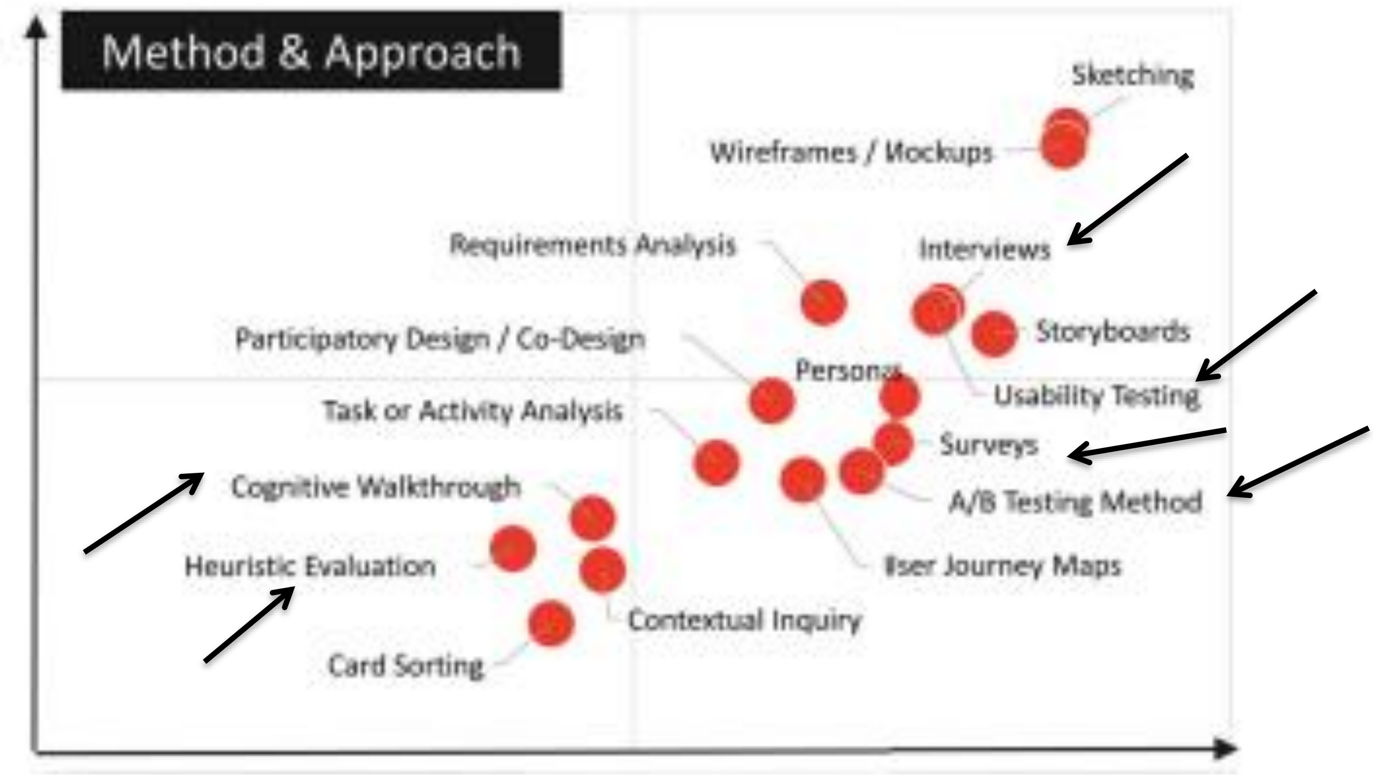
We will focus on some

- **Empirical** (involving users)
  - Observation
  - Query
  - Controlled Experiments

  ...

usability tests

Methods used by twenty data visualization practitioners:
(including evaluation methods)



P. Parsons, "Understanding Data Visualization Design Practice," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, n.1, 2022
https://ieeexplore.ieee.org/document/9555646

Heuristic Evaluation (Nielsen and Molich 1990)

- A “**discount usability engineering method**” for quick, cheap, and easy evaluation of a UI design
- Most popular usability inspection method; yet is **subjective**
- It is a **systematic inspection** of a design for usability
- Meant to find the usability problems in the design so that they can be attended to as part of an iterative design process
- Involves a small set of analysts judging the UI against a list of usability principles ("**heuristics**")

How to perform HE

- Should be performed by **several evaluators** (one person will never be able to find all the problems)

- Evaluators should work independently:
  - First get a general idea of the UI
  - Then perform a detailed inspection using a set of heuristics
  - Listing usability problems (heuristics not followed and severity degree)

- Findings of all evaluators should be **integrated in the same report**

- The report should **help the development teem to prioritize problem fixing**

Heuristic Evaluations: How to Conduct - NN/G

- Nielsen proposed **10 general usability heuristics**, yet **there are other sets** (e.g., visualization, web, mobile applications for seniors or children…)

- More details on how to conduct a heuristic evaluation at: [Heuristic Evaluations: How to Conduct - NN/G]()

- And how to rate severity of the usability problems found: [Severity Ratings for Usability Problems: Article by Jakob Nielsen - NN/G]()

- **The list of problems and severity rates should help the development team to prorityse problem fixing**

List of recognized usability principles (“the heuristics”)

1-Visibility of system status

2-Match between system and the real world

3-User control and freedom

4-Consistency and standards

5-Error prevention

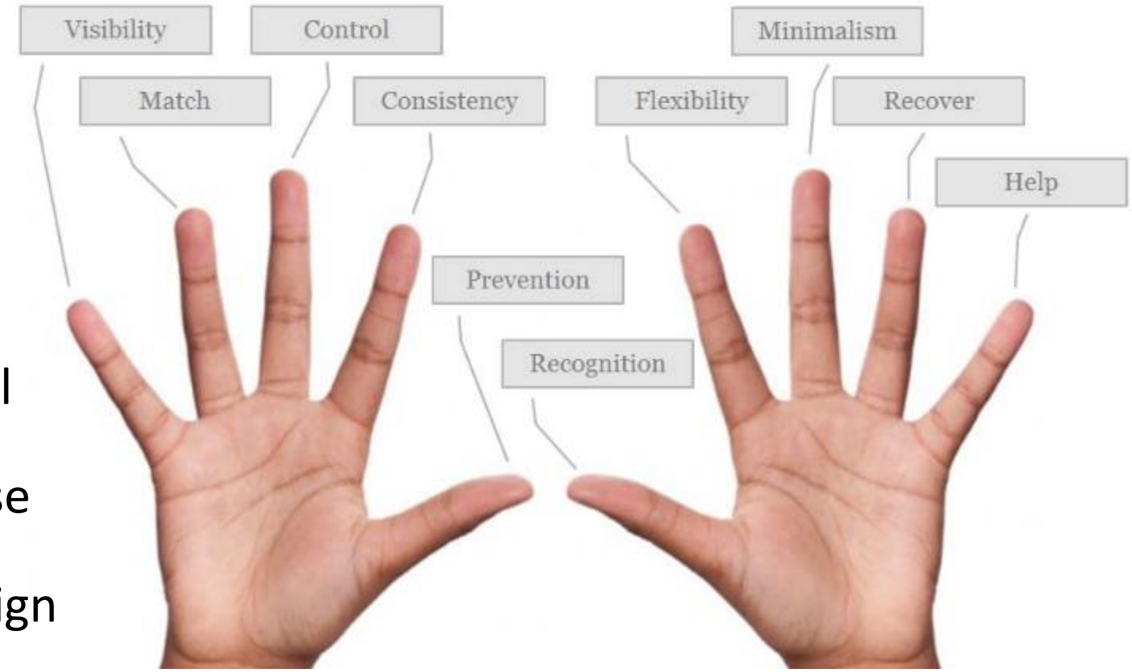
6-Recognition rather than recall

7-Flexibility and efficiency of use

8-Aesthetic and minimalist design

9-Help users recognize, diagnose, and recover from errors

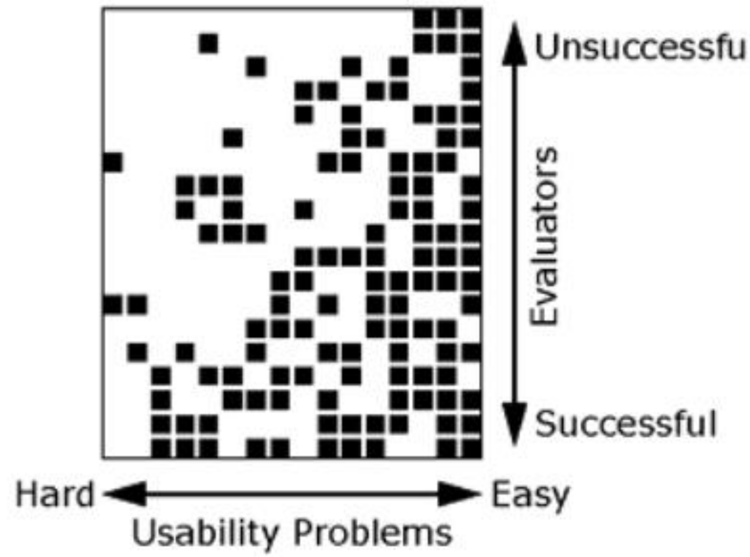
10-Help and documentation



[10 Usability Heuristics for User Interface Design - NN/G]

Number of problems found by several evaluators
Example:

- Heuristic evaluation of a banking system:
  - 19 evaluators
  - 16 usability problems



black square - problem found

white square – not found

- Conclusion : **in general 3 to 5 evaluators seems reasonable**

- More evaluators find more problems but cost more …

Specific Heuristics for Visualization

- Zuk's Perceptual and Cognitive heuristics (Zuk *et al*., 2006)
- Forsell's. heuristic set for evaluation in InfoVis (Forsell and Johanson, 2010)
- Shneiderman's "Visual Information-Seeking Mantra"
- Freitas's *et al*. Ergonomic Criteria for Hierarchical Information Visualization Techniques (Freitas et al., 2009)
- Amar and Stasko's Knowledge and task-based framework
- ...

Zuk and Carpendale's (2006) heuristics

1-Ensure visual variable has sufficient length

2-Don't expect reading order from color

3-Color perception varies with size of colored item

4-Local contrast affects color & gray perception

5-Consider people with color blindness

6-Preattentive benefits increase with field of view

7-Quantitative assessment requires position or size variation

8-Preserve data to graphics dimensionality

9-Put the most data in the least space

10-Remove the extraneous (ink)

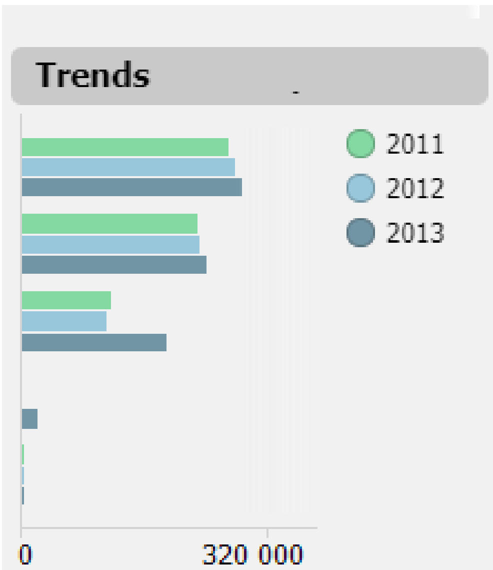
11-Consider Gestalt Laws

12-Provide multiple levels of detail

13-Integrate text whenever relevant

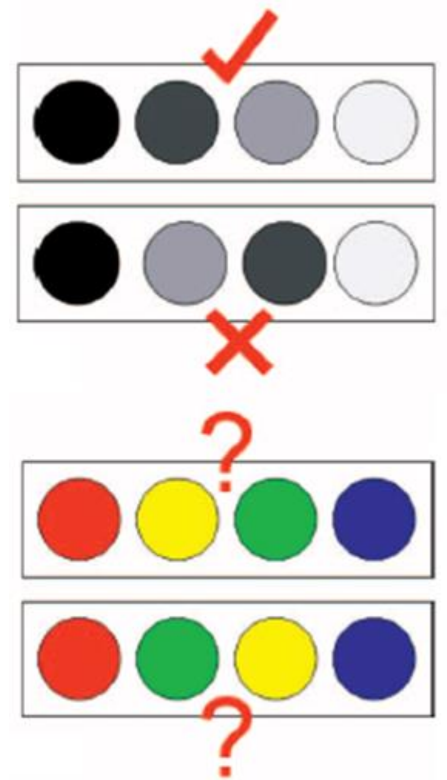
Explaining some of the Specific Heuristics for Visualization

Visual variables must have sufficient length

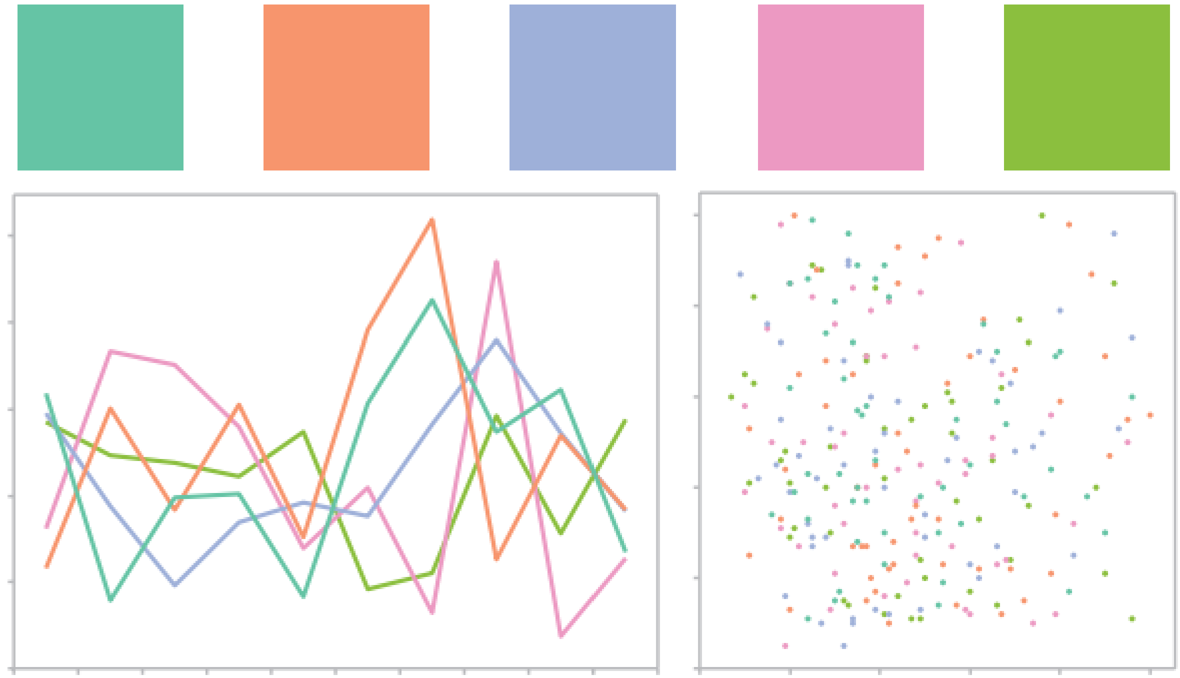


Too small

Do not expect to easily perceive order from color



Color perception varies with size of colored item



A set of colors with different hues but equal luminance (L* = 72).
In the squares, these colors are distinct and visually balanced (www.colorbrewer.org).
They are harder to distinguish in smaller items

Stone, M., "In color perception, size matters", IEEE *Computer Graphics & Applications*. 32, 2, 2012, pp. 8-13
https://ieeexplore.ieee.org/document/6155162

Consider people with color blindness

The most common form of color blindness is deuteranopia ("daltonism")

There are color blindness simulators

[Coblis — Color Blindness Simulator – Colblindor]()

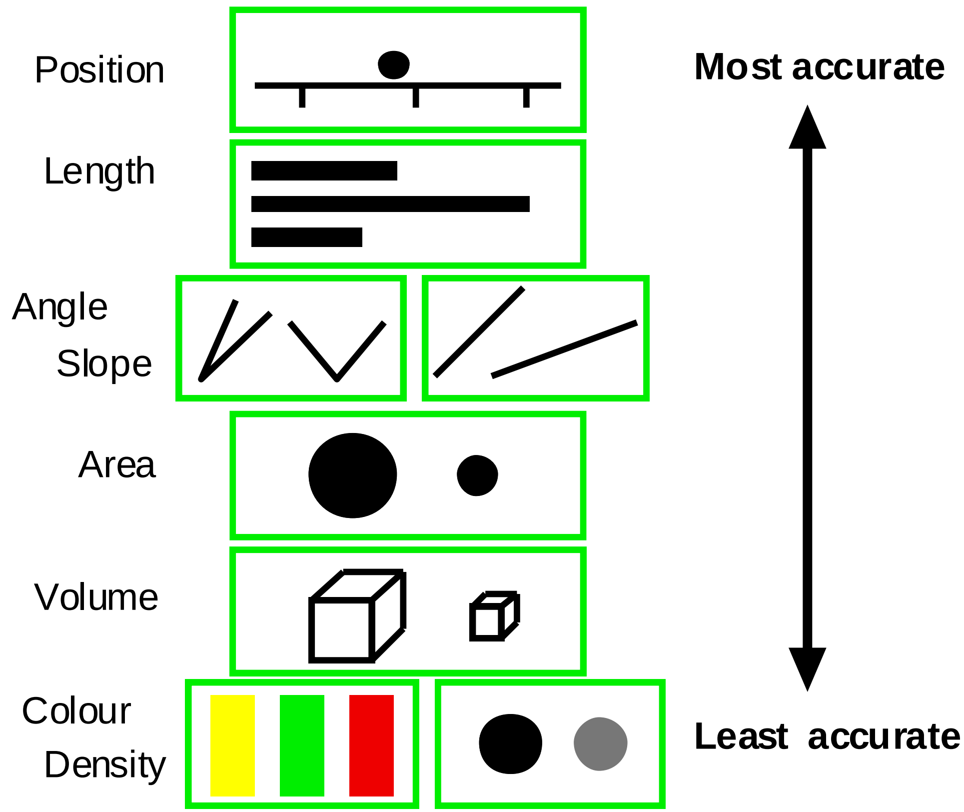
Normal vision

Deuteranopia

Tritanopia

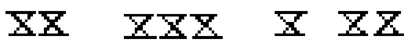
[Home - Colour Blind Awareness]()

Quantitative assessment requires position or size variation



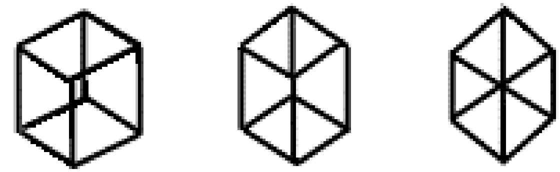
The relative difficulty of assessing quantitative value as a function of encoding mechanism, as established by Cleveland and McGill (Spence, 2007)

Gestalt Principles



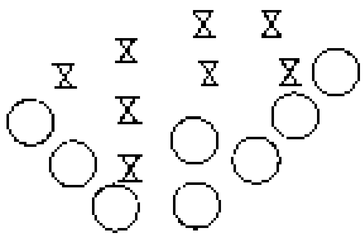
Proximity

Near stimuli are perceived as a group



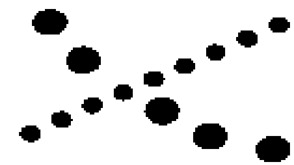
Simplicity

Ambiguous stimuli tend to be resolved
Using the simplest explanation



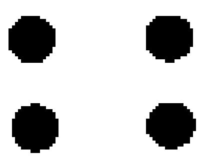
Similarity

Similar stimuli tend to be grouped
(may override proximity)



Good continuation

Stimuli tend to be grouped as to minimize variations or discontinuities



Closure

Stimuli tend to be grouped
in complete figures

Symmetry

Regions delimited by symmetric tend
to be perceived as coherent figures

[What are the Gestalt Principles? | IxDF]()

Forsell´s et al. (2010) heuristics

Uses heuristics from other sets:
B- Freitas et al.
C - Nielsen
D- Zuck and Carpendale
E- Bastien & Scapin

B5. Information coding. Perception of information is directly dependent on the mapping of data elements to visual objects. This should be enhanced by using realistic characteristics/techniques or the use of additional symbols.

E7. Minimal actions. Concerns workload with respect to the number of actions necessary to accomplish a goal or a task.

E11: Flexibility. Flexibility is reflected in the number of possible ways of achieving a given goal. It refers to the means available to customization in order to take into account working strategies, habits and task requirements.

B7: Orientation and help. Functions like support to control levels of details, redo/undo of actions and representing additional information.

B3: Spatial organization. Concerns users' orientation in the information space, the distribution of elements in the layout, precision and legibility, efficiency in space usage and distortion of visual elements.

E16: Consistency. Refers to the way design choices are maintained in similar contexts, and are different when applied to different contexts.

C6: Recognition rather than recall. The user should not have to memorize a lot of information to carry out tasks.

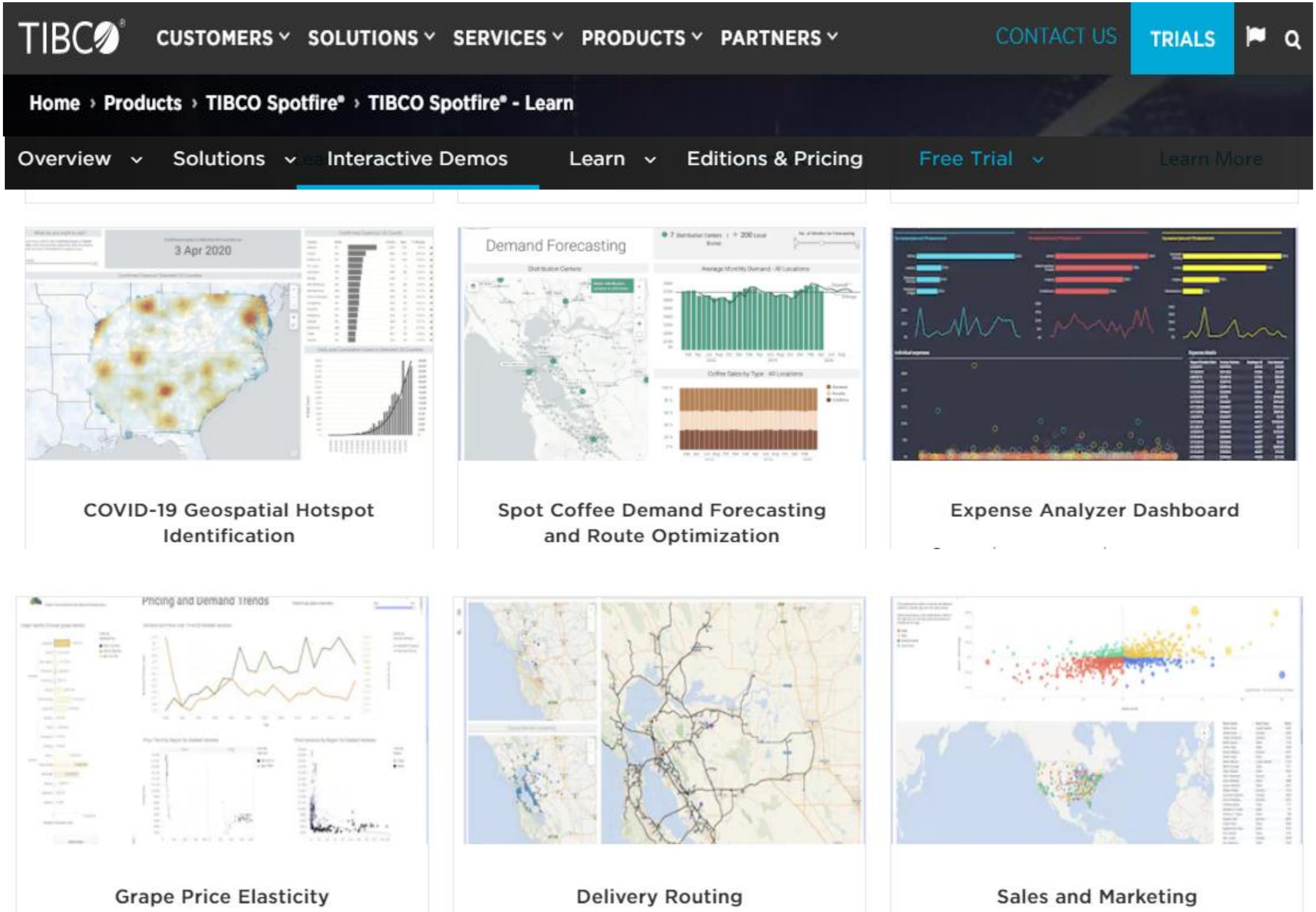
E1: Prompting. Refers to all means that help to know all alternatives when several actions are possible depending on the contexts

D10: Remove the extraneous. Concerns whether any extra information can be a distraction and take the eye away from seeing the data or making comparisons.

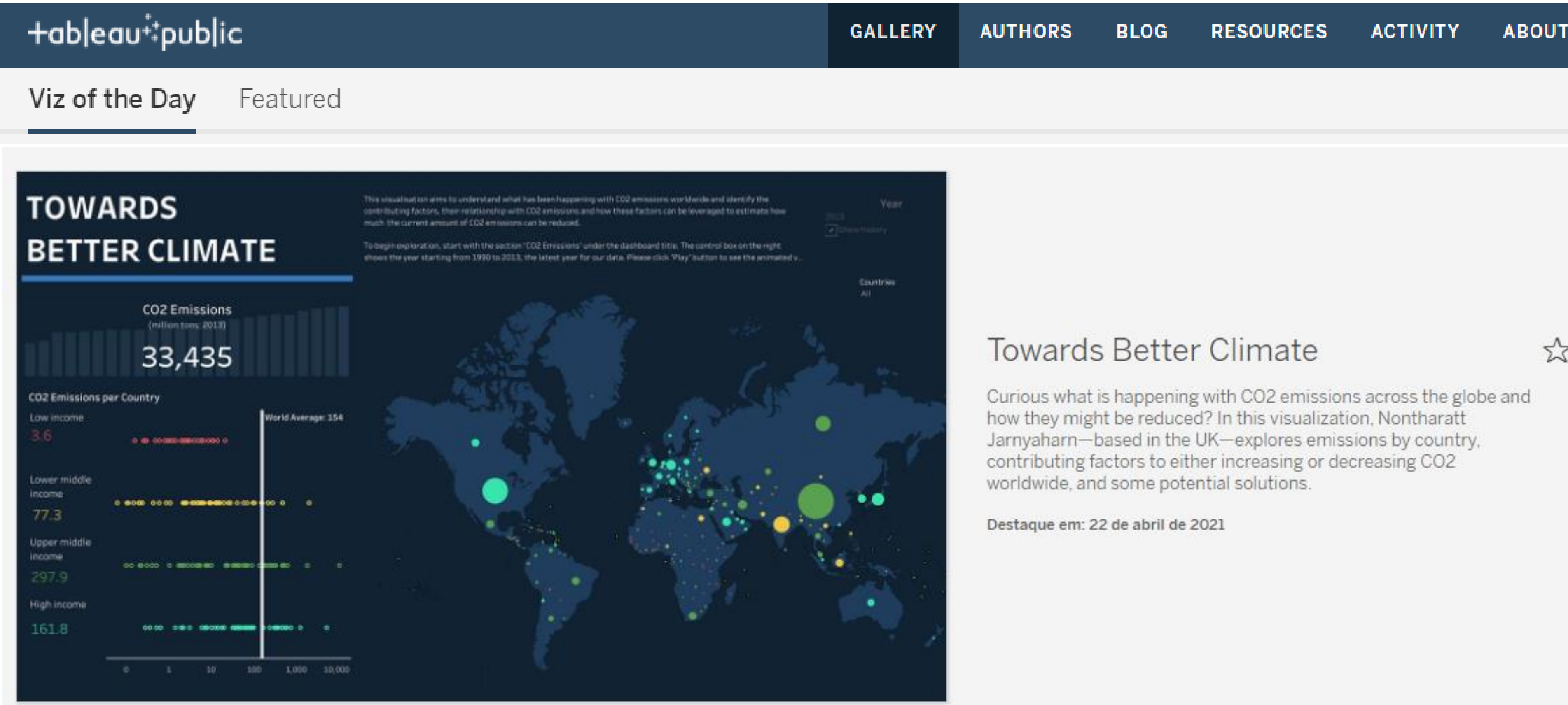
B9: Data set reduction. Concerns provided features for reducing a data set, their efficiency and ease of use

Practice using heuristic evaluation:
(groups of three students)

- Select a Visualization application from:
  - Spotfire demo gallery
  - Tableau Public
  - etc.
- Select a set of heuristics (e.g. Nielsen’s or Zuch’s)
- Explore and identify interesting characteristics or problems
- Perform an heuristic evaluation
- Prepare a short presentation with the most interesting results

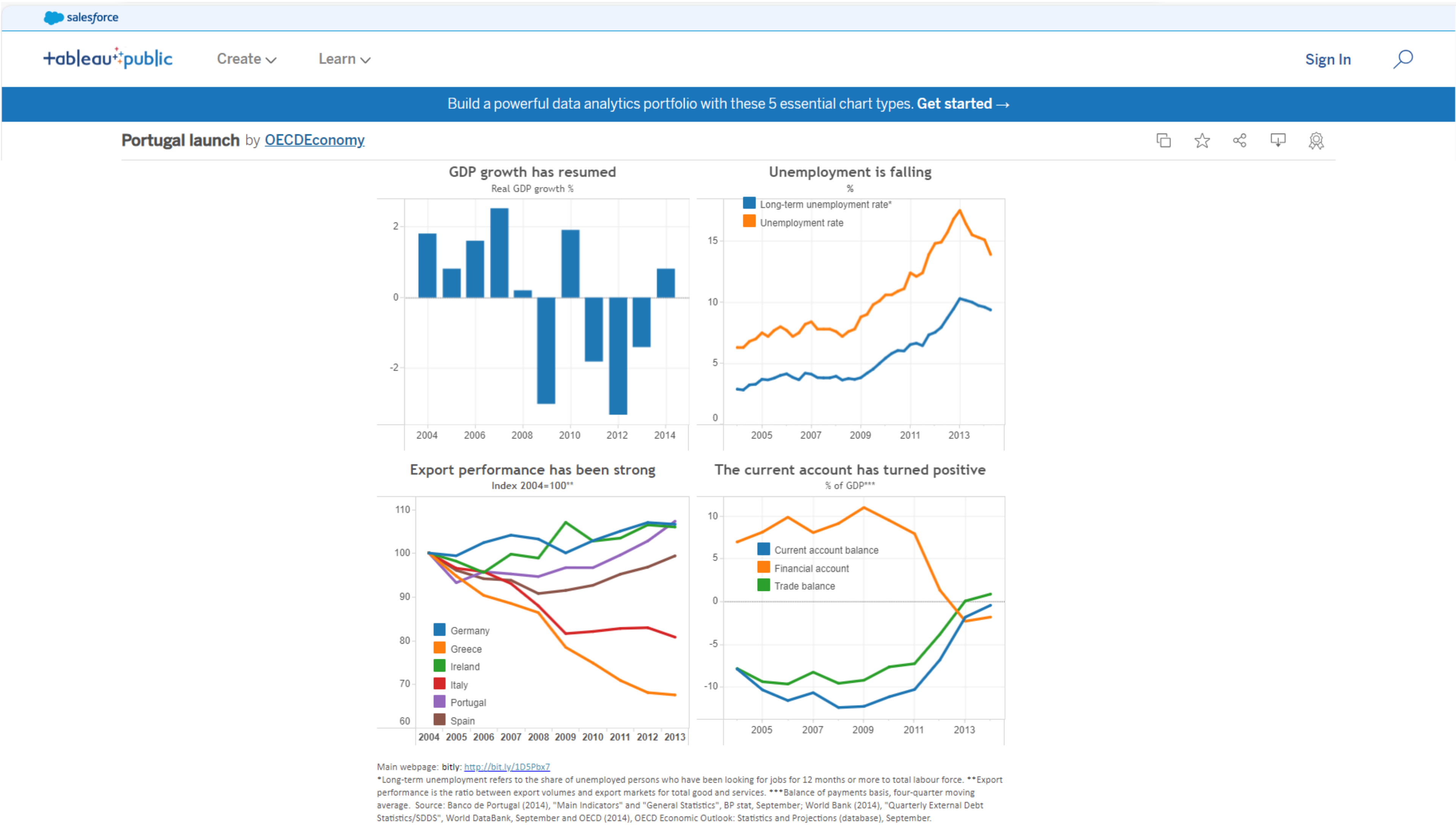


[Spotfire interactive demos]



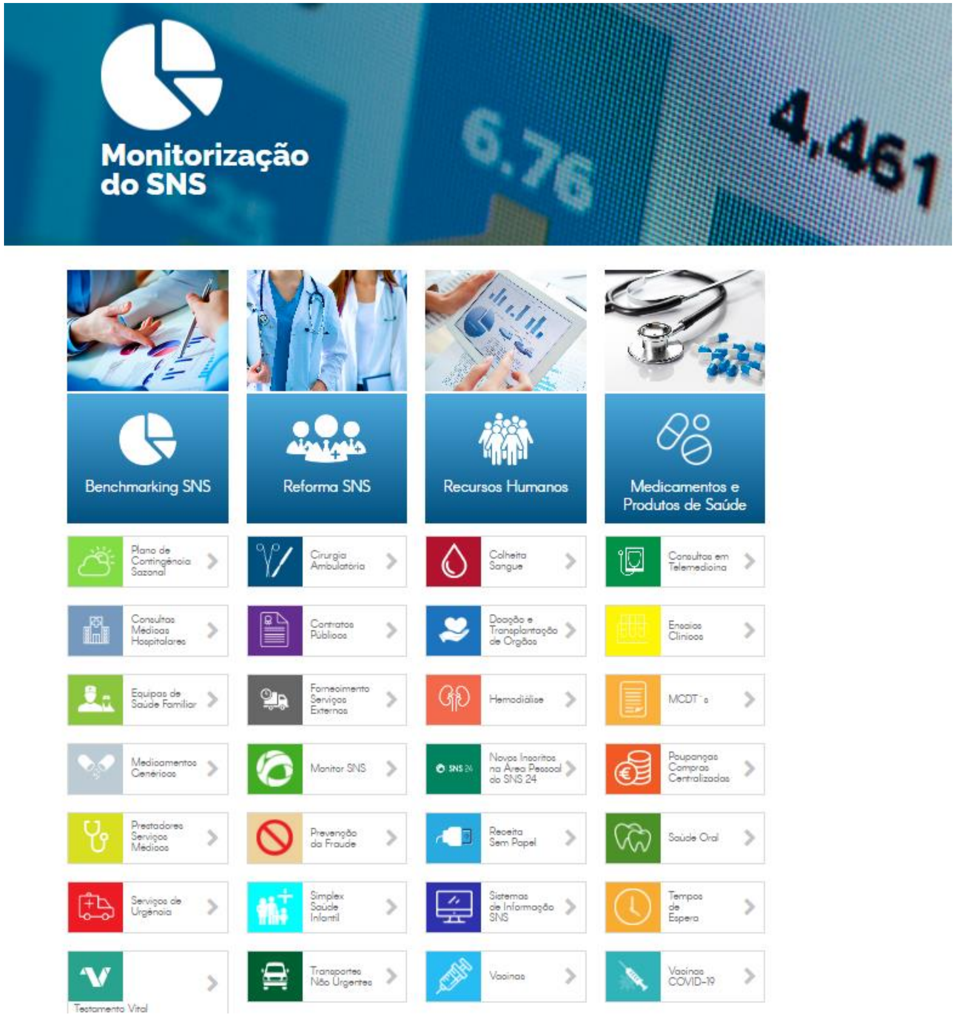
[Discover | Tableau Public]()

Example: Portugal Launch



[Portugal launch | Tableau Public]()

A Portuguese public example:
monitoring the National Health Service (SNS)



[Monitorização do SNS – SNS]

Usability tests

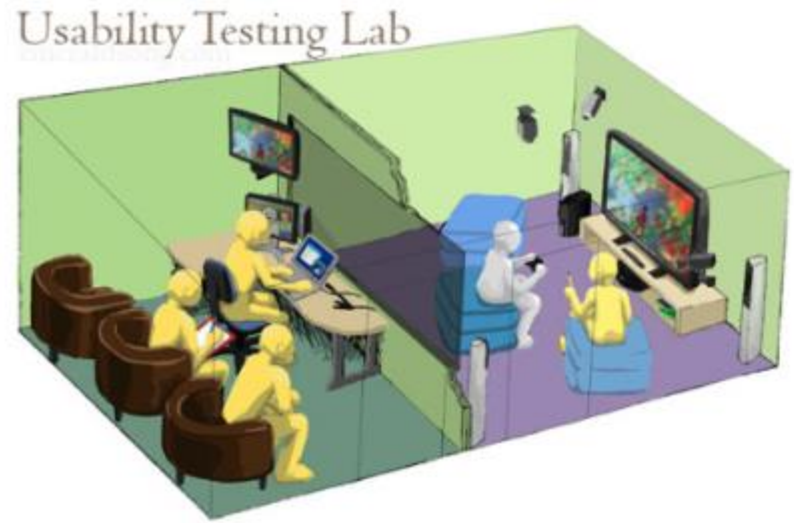
- “Engineering approach”

- Involve observation and query

- Main aspects:
  - Participants
  - Tasks
  - Test facilities and systems
  - Experimental design
  - Usability measures
  - Data analysis

- May have a complex logistics

Observation



Has many variants from very simple
to very complex and expensive:

- **Direct**: observer takes notes
- **Indirect**: through audio/ video – more complex and time consuming
- **Think Aloud**: users are asked to explain what they are doing
- **Logging**: users activity is logged by the system
- Combinations of the previous, etc.

Query

- Two main variants:
  - Questionnaire
    (reach more people; less flexible)
  - Interview
- **Should always be carefully prepared and tested**
- Collected data should be carefully analyzed

[Useful Survey Questions for User Feedback Surveys | IxDF]

[How to Conduct User Interviews | IxDF]

Well-known usability questionnaires

- System Usability Scale (**SUS**)

- Questionnaire for User Interface Satisfaction (**QUIS**)

- SUS provides a “quick and dirty”, reliable tool for measuring the usability
- It includes 10 questions with five response options

- QUIS is designed to assess a user's subjective satisfaction with the UI

- It is designed to be configured according to the needs of each UI analysis by including only the sections that are of interest to the user

- Both questionnaires should be completed following use of the UI in question

System Usability Scale (SUS)

- Provides a “quick and dirty”, reliable tool for measuring the usability
- It includes 10 questions with five response options
- It allows to evaluate a wide variety of products and services

  (H/W, S/W, mobile devices, websites and applications)
- Has become an industry standard, with references in over 1300 publications

Benefits of using a SUS

- Is a very easy scale to administer to participants
- Can be used on small sample sizes with reliable results
- Is valid – it can differentiate between usable and unusable systems

Usability – Digital.gov

SUS Questions

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.
- I would imagine that most people would learn to use this system very quickly.
- I found the system very cumbersome to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.

[Usability – Digital.gov]

Scoring SUS

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5

Note that:

odd numbered questions represent positive aspects

and

even numbered questions negative aspects…

Scoring SUS

- SUS provides a value in **[0-100]**
- To obtain the value:

  - Add the scores of all questions:

    - odd numbered questions - subtracting 1 from the score
    - even numbered questions - subtracting their value from 5

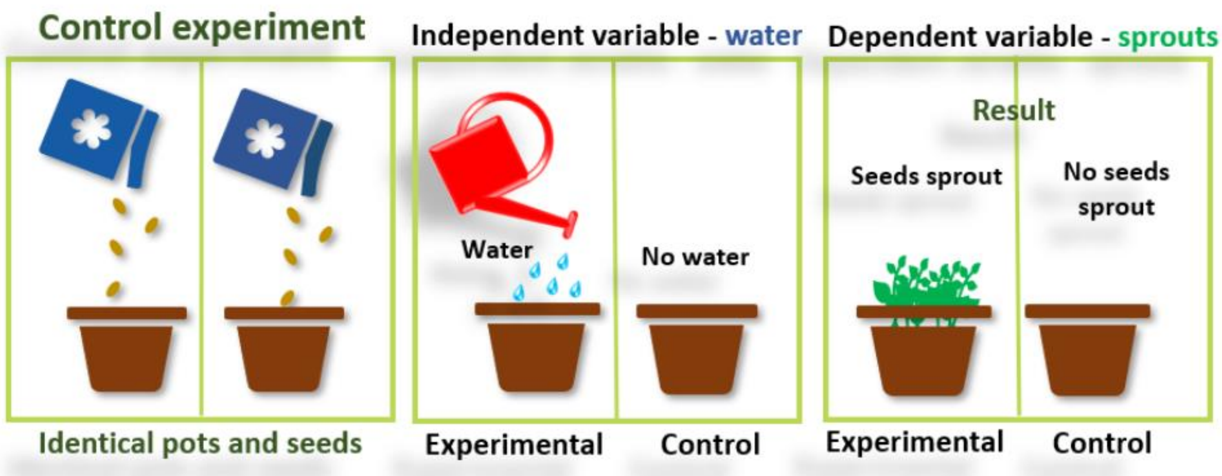
  - Multiply the sum by 2.5.

SUS > **68 would be considered above average**

Controlled experiments

- The “**work horse of experimental science”** ...
- Important issues to consider:
  - **Hypothesis**
  - **Variables** (input or independent; output or dependent)
  - **Secondary variables**
  - **Experimental design**
  - **Procedure**
  - **Statistics**

[Biology sample]



Controlled experiments with humans

Some times referred to as A/B testing [A/B testing - Wikipedia]

- Define an **hypothesis**
- Define input (independent), output (dependent) and secondary **variables**
- Define **conditions**/treatments (according to the input variables)
- Define **experimental design** (within-groups/subjects; between groups/subjects)
- Select the **participants**
- Define the procedure and prepare all the documentation:
  - informed consent
  - list of tasks and perceived difficulty
  - final questionnaire
  - list of tasks for the observer to take notes

  To the user

  To the observer
- Run a pilot test
- Take care of the logistics ... and after the experiment analyze data

Participants

Important issues in usability tests and controlled experiments:

- The total number of participants to be tested
  (a valid statistical analysis implies a sufficient number of subjects)

- Segmentation of user groups tested, if more than one

- Key characteristics and capabilities of user group
  (user profile: age, gender, computing experience, product experience, etc.)

- How to select participants

- Differences between the participant sample and the user population
  (e.g. actual users might have training whereas test subjects were untrained)

Tasks

- The task scenarios for testing (or experiments)
- Why these tasks were selected

  (e.g. the most frequent tasks, the most troublesome tasks)
- The source of these tasks

  (e.g. observation of users using similar products, product specifications)
- Any task data given to the participants
- Completion or performance criteria established for each task

  (e.g. n. of clicks < N, time limit)

Test Facilities and equipment

- The setting and type of space in which the evaluation will be done

  (e.g. usability lab, cubicle office, meeting room, home office, home family room, manufacturing floor, etc.)

- Any relevant features or circumstances that can affect the results

  (e.g. video and audio recording equipment, one-way mirrors, or automatic data collection equipment)

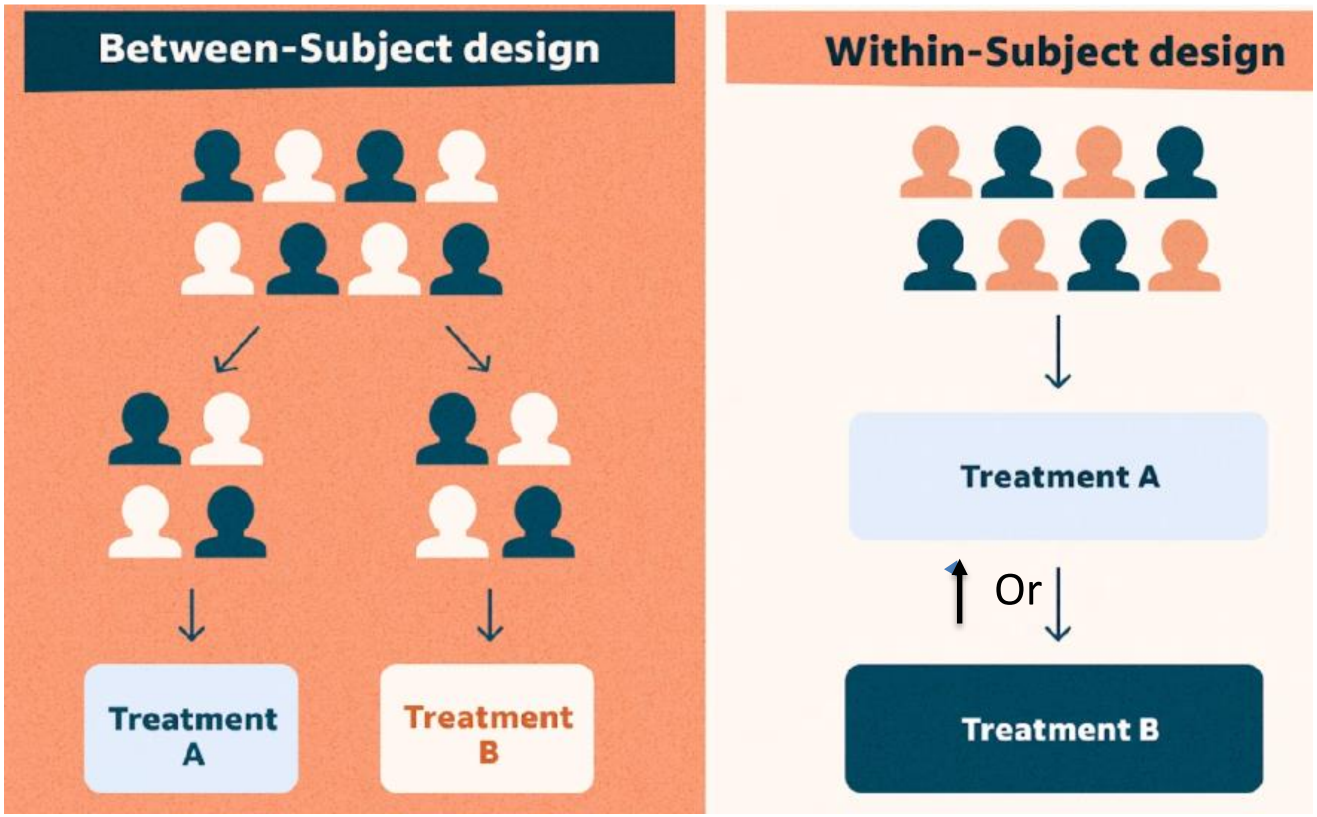
- Participant’s Computing Environment

  (e.g. computer configuration, including model, OS version, required libraries or settings, browser name and version; relevant plug-in, etc.)

- Display and input devices characteristics

- Any questionnaires to be used

Experimental design: **between *vs* within subjects/groups**



More participants
Less learning effects

Less participants
Possibly more learning/fatigue effects
(treatments order need to be randomized)

In some situations it is not possible to use a within subjects design

Procedure

- Procedure/ protocol: the logical design of the test/experiment
- Participant general instructions and task instructions
- The independent variables and control variables
- The usability measures to be used:
  - a) for effectiveness (completeness rate, errors, assists…)
  - b) for efficiency (times)
  - c) for satisfaction

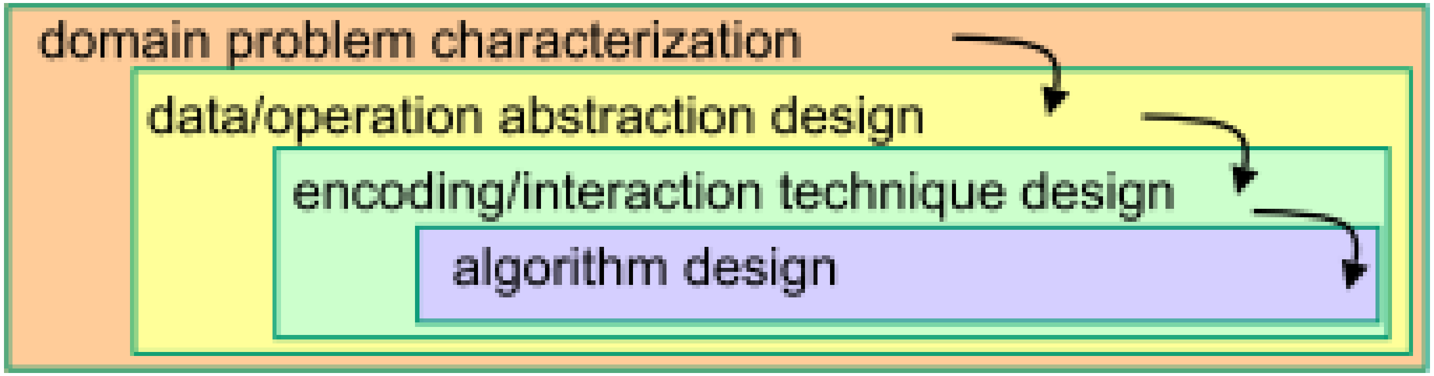
We know issues and methods, but how to use them?

A nested model for visualization design and validation:

- This model can be used :
  - to **analyze existing systems** or papers,

  or
  - to **guide the design** process

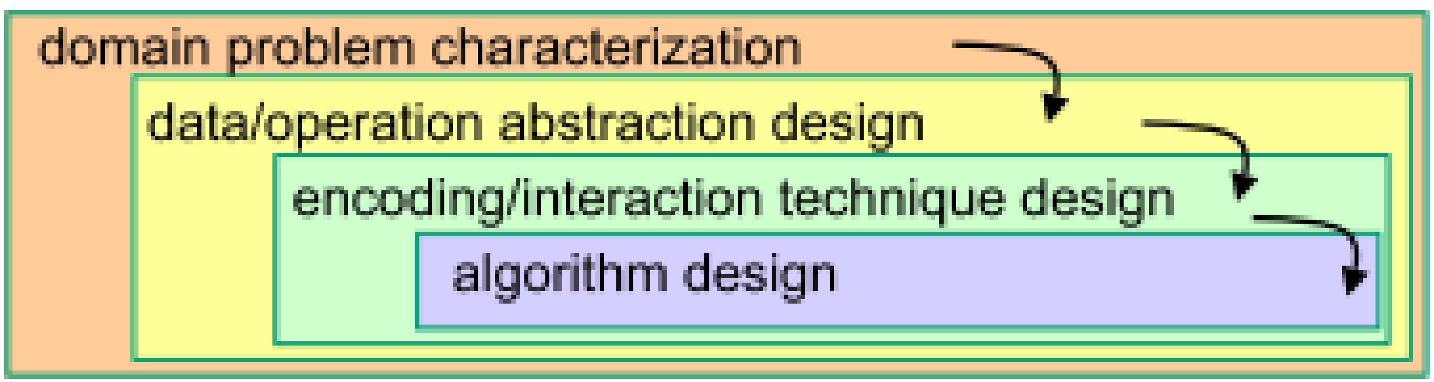
- Provides explicit guidance on what **evaluation methodology** is appropriate and identifies **threats to validity at each level**



Munzner, T.. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*. 15, 6, 2009, pp. 921–8
[Visualization Analysis & Design - Visualization Analysis and Design [Book]]

Four levels for visualization design and validation

- **characterize** the tasks and data in the vocabulary of the **problem domain**,
- **abstract into operations and data types**,
- **design visual encoding and interaction techniques**,
- **create algorithms** to execute these techniques with **efficiency** and **efficacy**



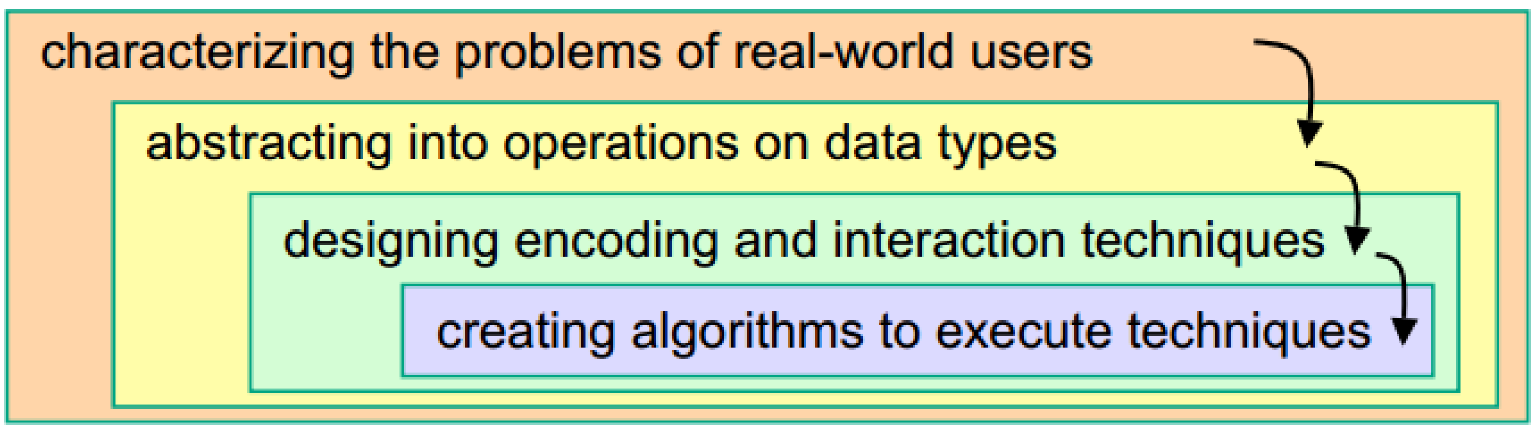
3 design levels

In each of the four levels it is necessary to :

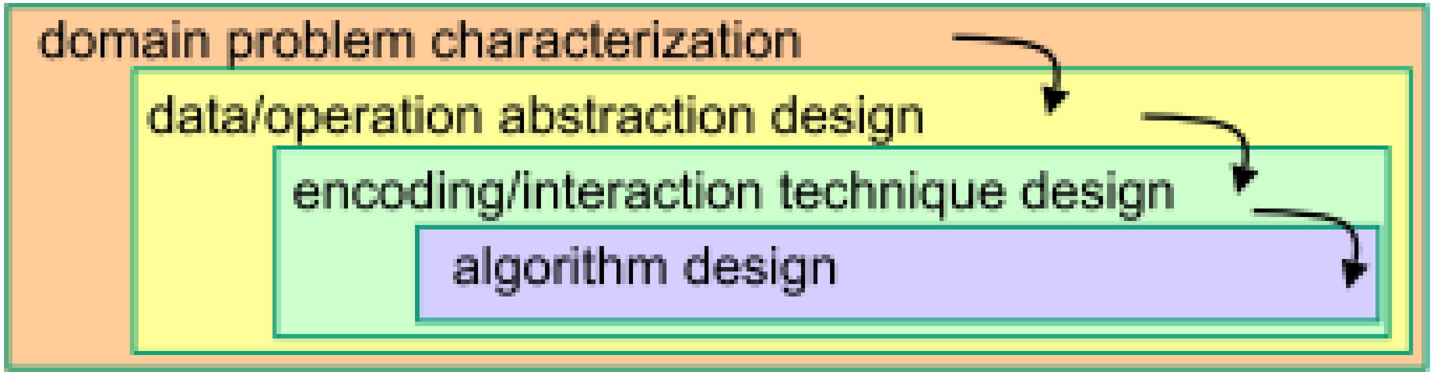
- 1- learn about the tasks and data of target users in some particular target **domain**,
- 2- map problems and data from the vocabulary of the specific domain into a more **abstract description** in the vocabulary of computer science (e.g. filter, retrieve value, sort, find extrema, etc.)
- 3- design the **visual encoding**, presentation and interaction,
- 4- create an **algorithm** to carry out the visual encoding and interaction designs automatically. The issues of algorithm design are not unique to visualization .

This split into levels is motivated by shared **threats to validity** at each one:

- Wrong problem: they don't do that;
- Wrong abstraction: you're showing them the wrong thing;
- Wrong encoding/interaction: the way you're showing the thing doesn't work;
- Wrong algorithm: your code is not adequate (e.g. too slow).



- Output from a level above is input to the level below
- An **upstream error inevitably cascades to all downstream levels**:
- E.g. a poor choice in the abstraction stage will not create a visualization system that solves the intended problem, even with perfect visual encoding and algorithm design .



- Evaluating Visualizations is **challenging**

- It will become more challenging as Visualization evolves to be more interactive, collaborative, mobile, immersive, multi-sensorial, …

- It is **fundamental** to:
  - evaluate solutions to specific cases
  - develop new visualization methods / systems
  - establish guidelines

  - i.e**. to make Visualization more useful, more usable, and more used**

Evaluation in Visualization Bibliography

- Zuk, T., L. Schlesier, Neumann, P., Hancock, M. and Carpendale, S. Heuristics for Information Visualization Evaluation," in *First Workshop on Beyond Time and Errors Novel Evaluation Methods for Visualization BELIV'06*, 2006, pp. 1–6.

- Also in Moodle: **Readings for exam preparation**

  Evaluation in Visualization