

# INF03 Expérience Utilisateur

---

## 7. Evaluation and tests – 3 Controlled experiments

Aurélien Tabard

# Evaluation and tests

---

Today:

- ▶ Experimental design
- ▶ Evaluation 2.0, scaling up

Last weeks:

- ▶ Introduction
- ▶ Approaches to evaluation
- ▶ Analytical methods
- ▶ Empirical methods

# Experimental design

---

- ▶ Introduction and examples
- ▶ The elements of experiments
- ▶ Experiment definition
- ▶ Conducting an experiment
- ▶ Gathering and cleaning data
- ▶ Analyzing data
  - ▶ Exploration
  - ▶ Statistics

# Experimental design

---

- ▶ **Introduction and examples**
- ▶ The elements of experiments
- ▶ Experiment definition
- ▶ Conducting an experiment
- ▶ Gathering and cleaning data
- ▶ Analyzing data
  - ▶ Exploration
  - ▶ Statistics

# Controlled experiments

---

## A scientific approach

- ▶ Answering specific questions with data
  - ▶ Performance
  - ▶ Learning
  - ▶ Satisfaction
- ▶ Providing basic knowledge generalizable across contexts.
- ▶ Demonstrate causality between different factors
  - ▶ correlation: show that a change in A occurs with a change in B
  - ▶ order: show that A takes place before B
  - ▶ no hidden cause: show that there is no C with  $C \rightarrow A$  and  $C \rightarrow B$

# Example 1: what is the best input device?

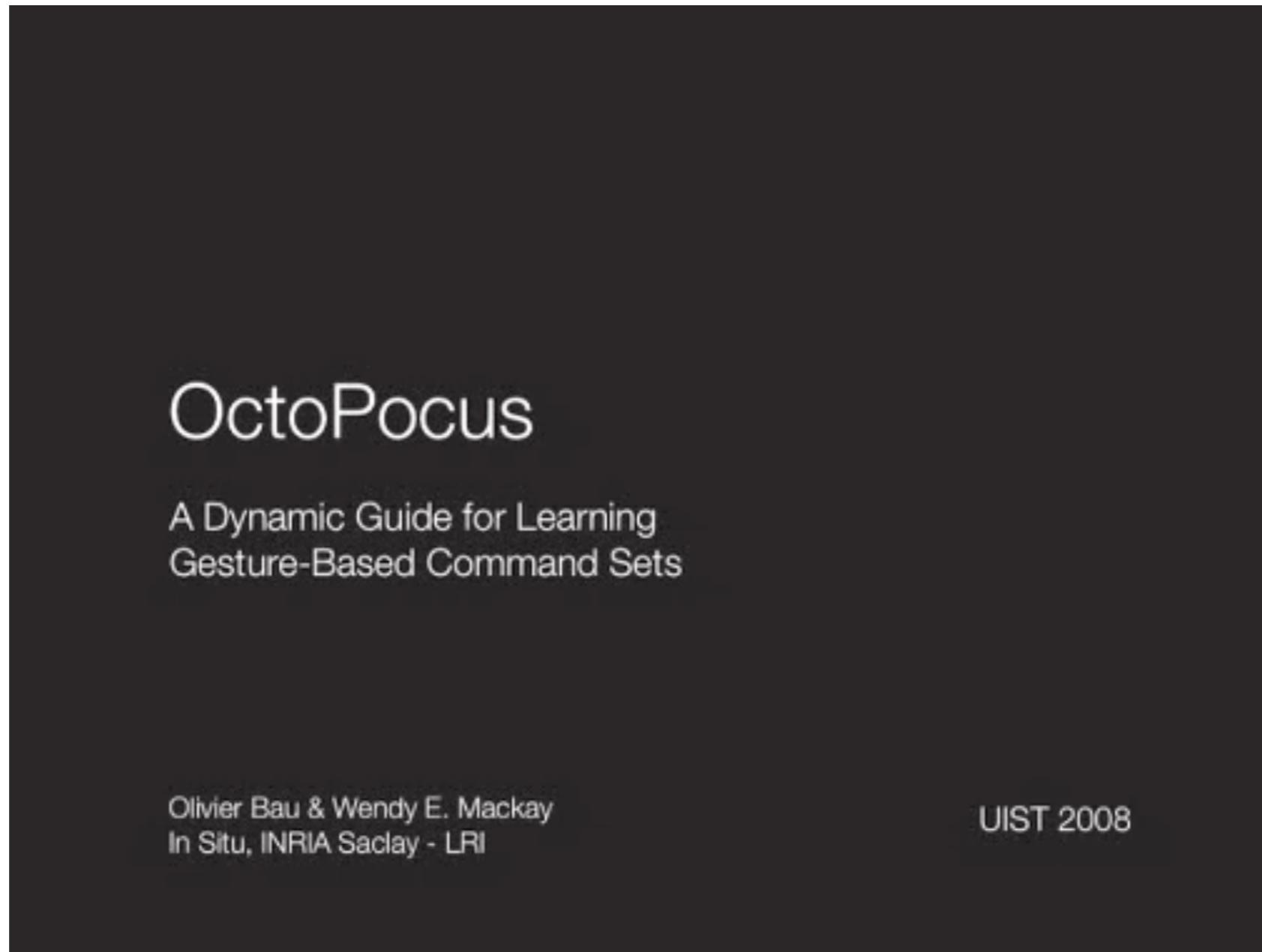
---

From two weeks ago:

Device	Study	IP (bits/s)
Hand	Fitts (1954)	10,6
Mouse	Card, English, & Burr (1978)	10,4
Joystick	Card, English, & Burr (1978)	5
Trackball	Epps (1986)	2,9
Touchpad	Epps (1986)	1,6
Eyetracker	Ware & Mikaelian (1987)	13,7

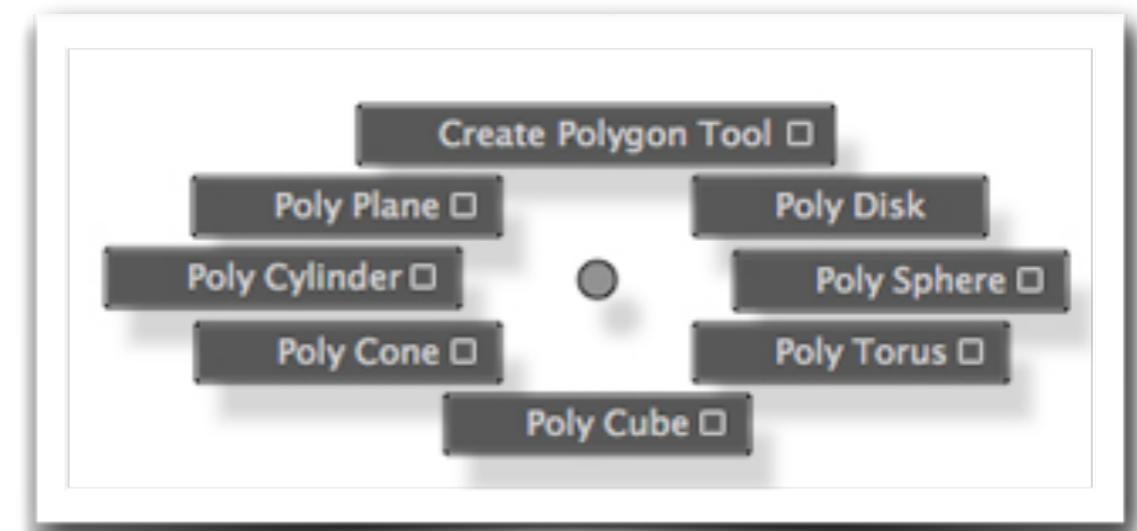
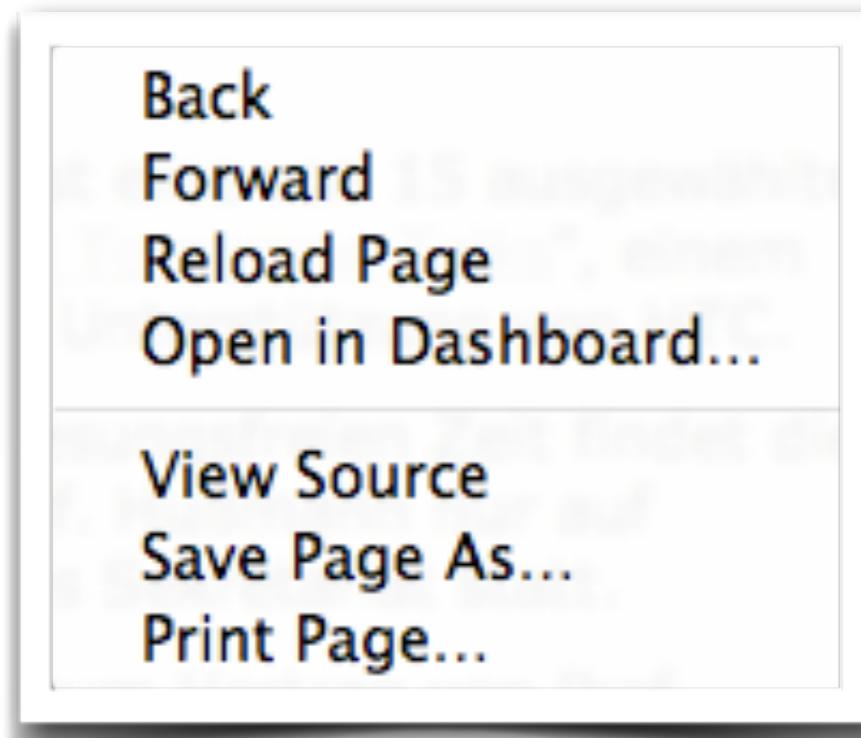
# Example 2: a better way do learn gestures?

---



# Exemple 3: which menu design is best?

---



# Experimental design

---

- ▶ Introduction and examples
- ▶ **The elements of experiments**
- ▶ Experiment definition
- ▶ Conducting an experiment
- ▶ Gathering and cleaning data
- ▶ Analyzing data
  - ▶ Exploration
  - ▶ Statistics

# Method (1)

---

1. Define what you are looking for: write down an hypothesis
  - ▶ Circular menus reduce search time
2. Design the experiment pick variables and fixed parameters
  - ▶ Define the menu structure (1 level or 2 level? how many items?)
3. Conduct a pilot to test the experiment
  - ▶ Fine tune data collection adjust the number of repetitions

# Method (2)

---

4. Recruit participants
  - ▶ Students spending more than 2h per day in front of their computer
5. Conduct experiment and gather data
6. Analyze and clean up data, to accept or reject the original hypothesis measure the effect (between conditions)
  - ▶ Average search time: 2.26 (Circular menu), 2.64 (Usual menu)
  - ▶ Difference is significant :  $p < .05$

# The elements of experiments

---

- ▶ Factors (or independent variables)
  - ▶ Variables we manipulate in each condition
  - ▶ The quantity of menu items, the quantity of sub-menus
- ▶ Levels (a.k.a. possible values for independent variables)
  - ▶ Menu with 8 items or a menu with 12 items
- ▶ Measures (or response): dependent variable(s)
  - ▶ The measured outcome of the experiment
  - ▶ Selection time of a menu item
- ▶ Replication
  - ▶ number of subjects assigned to each level

# Independent variables (factors)

---

The conditions of the experiment are set by independent variables

- ▶ The number of items in a list, text size, font, color

The number of different values used is the level

- ▶ The number of experimental conditions is the product of the levels
- ▶ E.g., font can be times or arial (2 levels), background can be blue, green, or white (3 levels). This results in 6 experimental conditions (times on blue, times, on green, ..., arial on white)

# Dependent variables (measures)

---

The dependent variables are the values to be measured:

- ▶ Objective values: e.g. time to complete a task, number of errors, etc.
- ▶ Subjective values: ease of use, preferred option, etc.
- ▶ They should only be dependent on changes of the independent variables.

# Exercise

---

Identify independent variables (factors) and dependent variables (measures) in each scenario.

Give possible levels for independent variables.

- ▶ A study to learn whether people who have followed a digital security training use more secure password.
- ▶ A study to learn whether a joystick or a mouse is more efficient to select static targets or moving targets
- ▶ A study to learn whether teams using video Hangout/Skype are more productive than those using text-only chat.

# Experimental design

---

- ▶ Introduction and examples
- ▶ The elements of experiments
- ▶ **Experiment definition**
- ▶ Conducting an experiment
- ▶ Gathering and cleaning data
- ▶ Analyzing data
  - ▶ Exploration
  - ▶ Statistics

# Participants / subjects

---

Within-subjects design:

- ▶ Same participant exposed to all test conditions
  - + Fewer participants needed (10 - 20)
  - Learning effects

Between-subjects design:

- ▶ Independent groups of participants for each test condition (control and treatment)
  - + No learning effect
  - Require more users

# Randomization and control

---

The order of presentation can have an impact measures:

- ▶ learning
- ▶ fatigue
- ▶ contrast (1st treatment leads to behavior/answers in 2nd treatment)

Solution

- ▶ Rest between treatment
- ▶ Counter-balancing, but can become complicated
- ▶ Latin square

# Latin square (Carré Latin)

via <http://hci.rwth-aachen.de/~chat/StatLecture/prerequisite.pdf>

- ▶ Each condition appears at each ordinal position
- ▶ Each condition precedes and follows each condition one time
- ▶ Example: six treatments: A, B, C, D, E, F

I	A	B	F	C	E	D
2	B	C	A	D	F	E
3	C	D	B	E	A	F
4	D	E	C	F	B	A
5	E	F	D	A	C	B
6	F	A	E	B	D	C

# Exercise

---

What kind of experimental design for:

- ▶ A study to learn whether people who have followed a digital security training use more secure password.
- ▶ A study to learn whether a joystick or a mouse is more efficient to select static targets or moving targets
- ▶ A study to learn whether teams using video Hangout/Skype are more productive than those using text-only chat.

# Hypotheses

---

- ▶ Prediction of the result of an experiment
- ▶ Stating how a change in the independent variables will affect the measured dependent variables

# Usual approach to hypotheses

---

- ▶ Stating a working hypothesis  $H_1$
- ▶ Stating a null hypothesis  $H_0$ 
  - ▶ intuition (naive) : if  $H_0$  is false then  $H_1$  must be true
- ▶ Carrying out the experiment and using statistical measures to disprove the null-hypothesis
- ▶ When a statistical test shows a significant difference it is probable that the effect is not random

# Exercise

---

Which  $H_0$  et  $H_1$  ?

- ▶ A study to learn whether people who have followed a digital security training use more secure password.
- ▶ A study to learn whether a joystick or a mouse is more efficient to select static targets or moving targets
- ▶ A study to learn whether teams using video Hangout/Skype are more productive than those using text-only chat.

# Validity

---

## Internal validity

- ▶ Manipulation of independent variable is cause of change in dependent variable
- ▶ Requires removing effects of confounding factors
- ▶ Requires choosing a large enough sample size, so the result couldn't have happened by chance alone.

## External validity

- ▶ Results generalize to real world situations
- ▶ Requires that the experiment be replicable
- ▶ No study “has” external validity by itself!

# Practice

---

Compare typing speed of  
a laptop keyboard vs. a smartphone keyboard.

- ▶ Which independent variables (factors)?
- ▶ Which dependent variables (measures)?
- ▶ inter- or intra- participants?
- ▶ Which hypotheses?

# Experimental design

---

- ▶ Introduction and examples
- ▶ The elements of experiments
- ▶ Experiment definition
- ▶ **Conducting an experiment**
- ▶ Gathering and cleaning data
- ▶ Analyzing data
  - ▶ Exploration
  - ▶ Statistics

# Collecting data

---

Records of what participants did

Log data

- ▶ Data of dependent variables (measures)
- ▶ time, errors, etc.

Table structure:

userid	group	condition	executontime	error

# Practice

---

- ▶ Which protocol?
- ▶ Which data to collect?

Go check

- ▶ <http://10fastfingers.com/>

# Experimental design

---

- ▶ Introduction and examples
- ▶ The elements of experiments
- ▶ Experiment definition
- ▶ Conducting an experiment
- ▶ **Gathering and cleaning data**
- ▶ Analyzing data
  - ▶ Exploration
  - ▶ Statistics

# Gathering data

---

Gather data from all groups in a pad  
in csv format (comma separated value).

Copy them on your plot.ly account

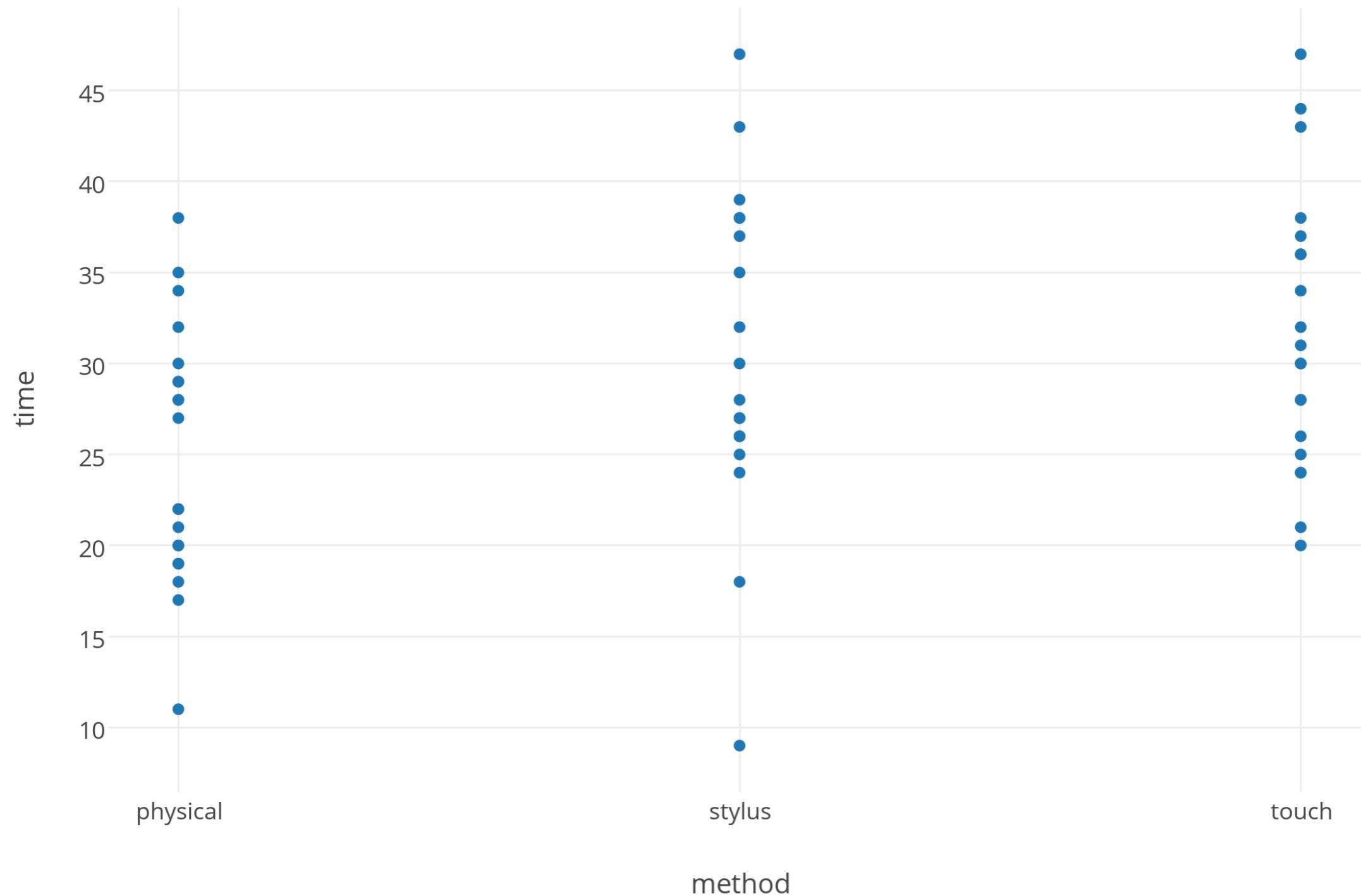
# Experimental design

---

- ▶ Introduction and examples
- ▶ The elements of experiments
- ▶ Experiment definition
- ▶ Conducting an experiment
- ▶ Gathering and cleaning data
- ▶ **Analyzing data**
  - ▶ Exploration
  - ▶ Statistics

# Who is fastest?

<https://plot.ly/~aurelient/41/>



# Who is fastest?

---

It depends of:

- ▶ the median differences
- ▶ the data distribution (standard deviation)
- ▶ the sample size
- ▶ whether averages are significantly different

First: the exploratory part

- ▶ look at the data with basic plots and statistics to get an idea.

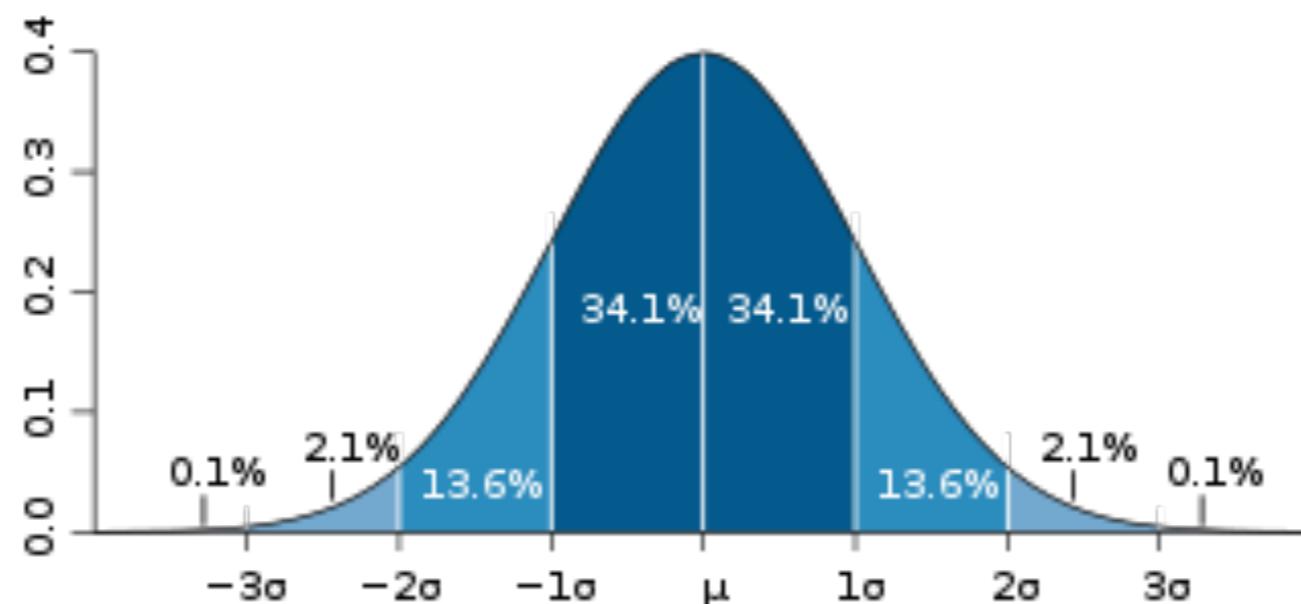
# (Student's) t-test

---

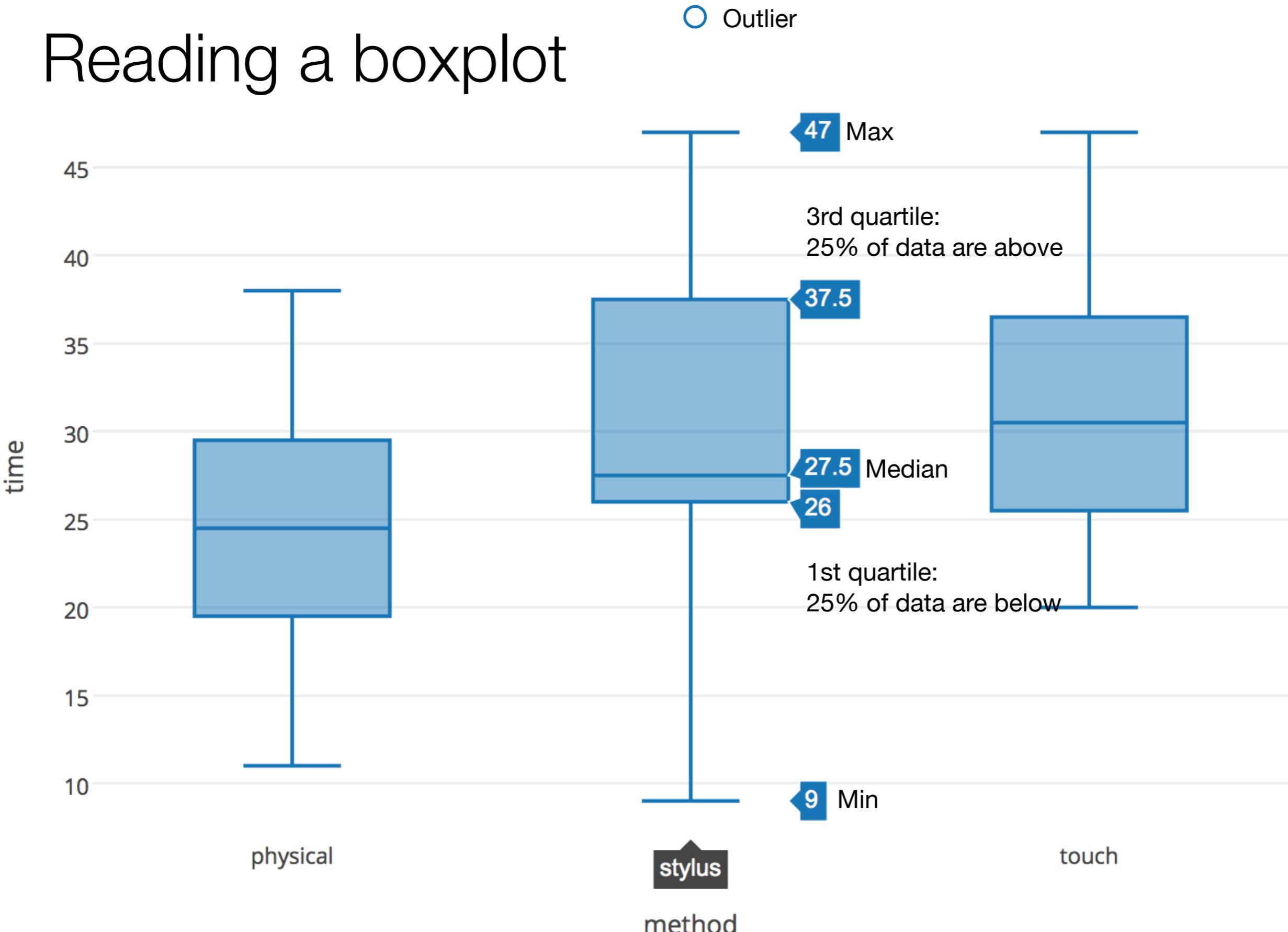
Looks at the relationship between two data sets

Designed for

- ▶ small sample (= few measurements)
- ▶ unknown (mean and) standard deviation
- ▶ but has to be **normally distributed**



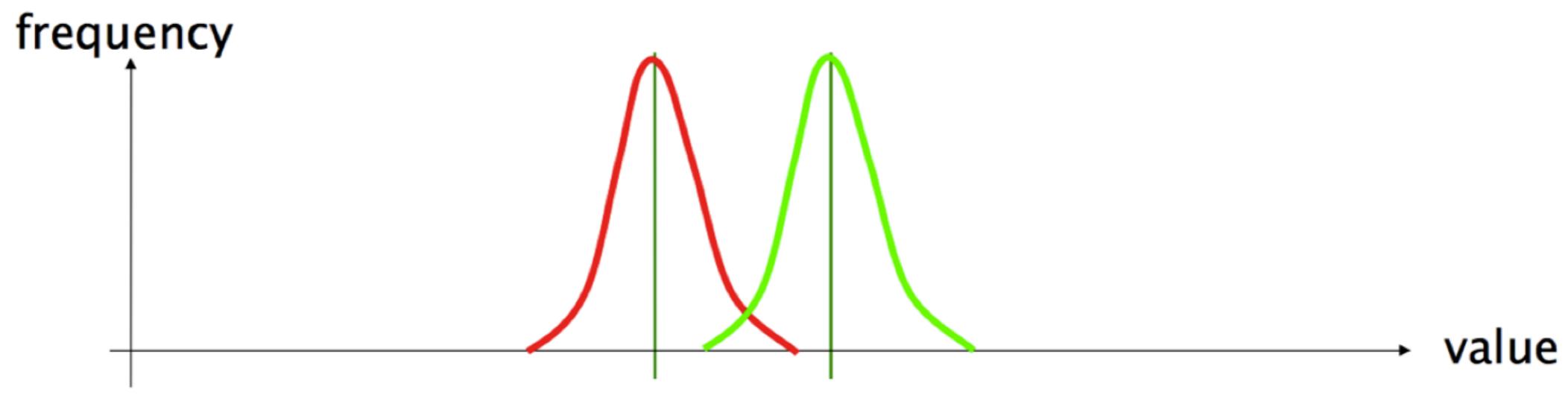
# Reading a boxplot



# Comparing values

via [http://www.medien\\_ifi.lmu.de/lehre/ws1213/mmi2/uebung/slides10.pdf](http://www.medien_ifi.lmu.de/lehre/ws1213/mmi2/uebung/slides10.pdf)

Is there a significant difference between two measures?



# t-test

---

Gives  $p$ :

- ▶ **the probability that two population have the same mean**
- ▶ Not probability the result is due to chance...

In UX :

- ▶  $p < 0.05$  (= 5% probability) is a convention (or 0.01)
- ▶ a smaller  $p$  (e.g. 0.00001) doesn't make the result more significatif.
- ▶ **a significant result is different from an important result**

# DO NOT

---

If  $p > 0.05$  say:

- ▶ “our tests showed that there was no difference”
- ▶ significant difference -> impact
- ▶ no significant difference -> nothing

You cannot show that there is no difference!

# Going further

<https://www.coursera.org/learn/designexperiments>

The screenshot shows the Coursera platform interface. On the left, a sidebar for 'UCSanDiego' lists navigation options: Home, Course Content, Assignments, Discussions, Classmates, and Course Info. The main content area displays the course title 'Designing, Running, and Analyzing Experiments' by 'University of California, San Diego'. It is part of an 8-course series, the 'Interaction Design Specialization'. The 'About this Course' section describes the course content, mentioning user experience testing, experiment design, and analysis using R. Subtitles are available in English, and the course requires 9 weeks, 2-3 hours per week. A sidebar indicates the learner is enrolled in session February 22 – May 1, with an upcoming session from March 21 - May 30. Financial aid information and course ratings are also present.

UCSanDiego

Coursera Catalog Search catalog Institutions AT

Home Course Content Assignments Discussions Classmates Course Info

## Designing, Running, and Analyzing Experiments

University of California, San Diego

Part of a 8-course series, the [Interaction Design Specialization](#)

### About this Course

You will never know whether you have an effective user experience until you have tested it with users. In this course, you'll learn how to design experiments, how to run experiments, and how to analyze data from these experiments in order to evaluate and validate user experiences. You will work through real-world examples of experiments from the fields of IxD and HCI, understanding issues in experiment design and analysis. You will analyze multiple data sets using recipes given to you in the R statistical programming language -- no prior programming experience is assumed or required. By the end of the course, you will be able to knowledgeably design, run, and analyze your own experiments for putting empirical and statistical weight behind your designs.

Subtitles available in English  
9 weeks, 2-3 hours/week

You're currently enrolled in this session:  
February 22 – May 1

Upcoming session:  
March 21 - May 30

Switch sessions

Following session begins April 4

Financial Aid is available for learners who cannot afford the fee. [Learn more and apply.](#)

Course Ratings Help Center

# Evaluation methods

---

Last weeks:

- ▶ Introduction
- ▶ Approaches to evaluation
- ▶ Analytical methods
- ▶ Empirical methods

Today:

- ▶ Experimental design
- ▶ **Evaluation 2.0, scaling up**

# Scaling user testing

---

Large Web audiences

Large audiences on mobile platforms

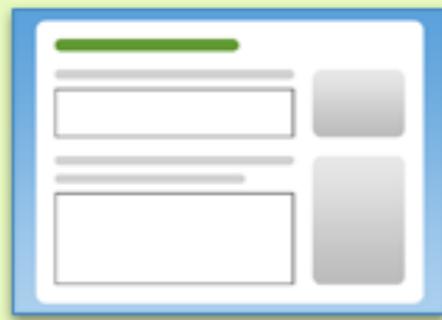
Ease of distribution of updates

Ease of logging

# Remote usability studies

## How It Works

### 1. Design Your Test



Choose one of our **professionally designed task templates** and then customize it for your site in seconds.

### 2. We Notify our User Panel



Within seconds, **representative users** start recording themselves using your site.

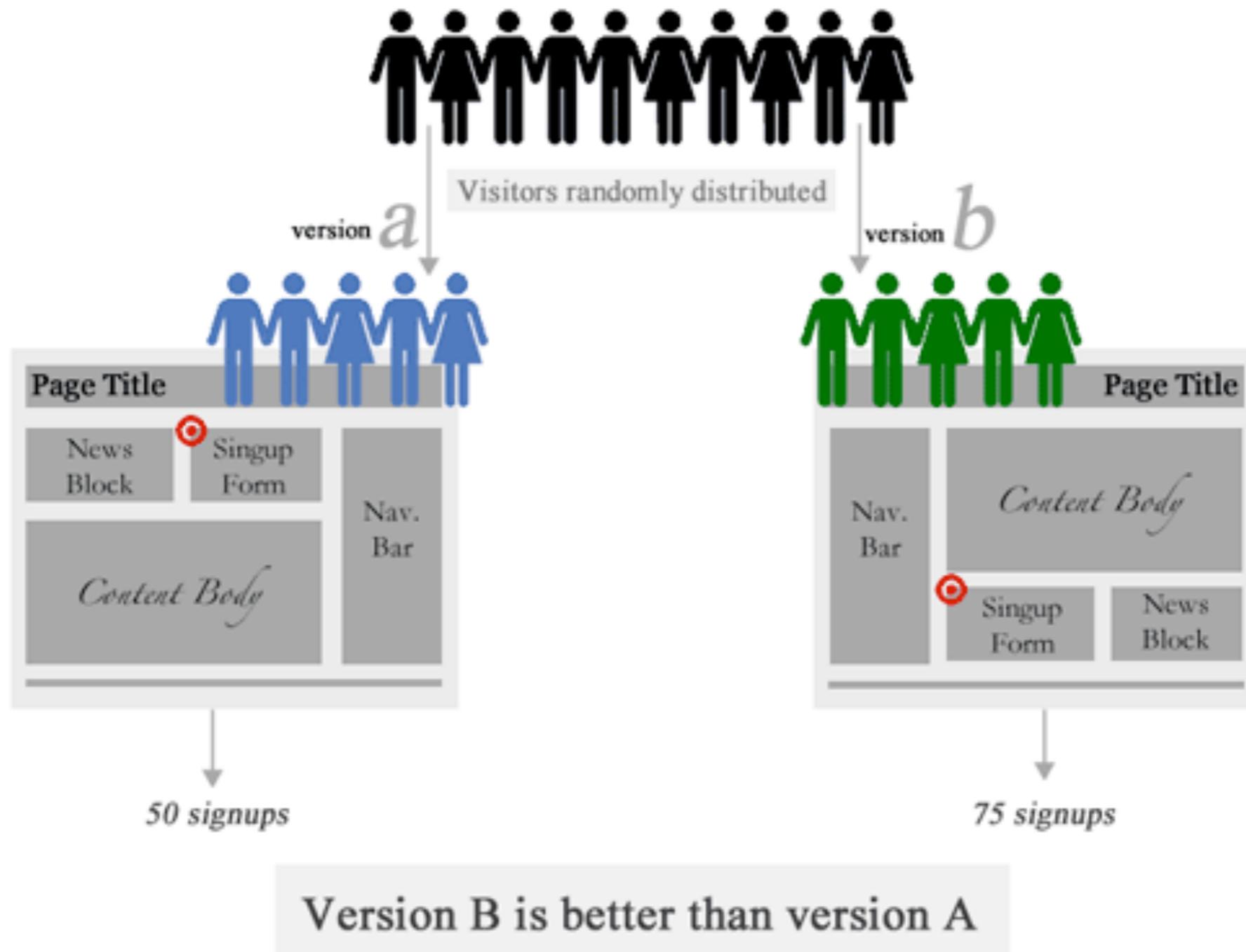
### 3. Get Feedback in an Hour



Receive a **video** and **written responses** from users.

E.g. [Usertesting.com](#)

# A/B testing



# A/B testing

---

see <http://optimizely.com/>

## Test the best version of a page

- ▶ a better form design
- ▶ a better conversion rate (for a newsletter)

## Limits :

- ▶ You need significant traffic
- ▶ Does not replace user studies!
- ▶ Does not provide explanations
- ▶ Arbitrary changes can be disturbing.
- ▶ Complex when there is tailored and social content, e.g., Facebook.
- ▶ Often used for incremental changes, complex for full redesigns.

# Controlled distribution of Beta versions



# Evaluation methods

---

Last weeks:

- ▶ Introduction
- ▶ Approaches to evaluation
- ▶ Analytical methods
- ▶ Empirical methods

Today:

- ▶ Experimental design
- ▶ Evaluation 2.0, scaling up
- ▶ **Homework**

# Devoir

---

Maintenant :

- ▶ Se mettre d'accord sur les deux modalités d'entrée

Pour la semaine prochaine (28 mars) :

- ▶ Description de l'expérience :  
variables, hypothèses, déroulé.
- ▶ Faire passer le test de frappe à 2 personnes (chacun)
- ▶ 2 conditions en alternance, 3 répétitions  
-> 3x2x1 minute d'XP, max 30 minutes par participant.

# Devoir

---

Pour le 18 avril (seul) :

- ▶ Analyse descriptive sur un échantillon de 12 personnes (*J'assigne les participants aux groupes le 03 avril*) :
  - Boxplot avec description des données, médiane, quartiles, distribution des données
  - présence (ou non) d'outliers
  - analyse des erreurs
- ▶ Conclusion visuelle
- ▶ Bonus: T-test et conclusion statistique