RESEARCH METHODOLOGY

Rémi Eyraud

Laboratoire Hubert Curien
Université Jean Monnet de Saint-Etienne
Faculté des Sciences et Techniques

Slides shamelessly based on the ones of Emilie Morvant (with her authorization)
`https://perso.univ-st-etienne.fr/er101405/RM_1.pdf`

Master 1 MLDM & Master 2 DSC - Semester 1

Who am I?

Dr. **Rémi Eyraud**

- Maître de Conférences (*junior professor*)

  at University Jean Monnet, St-Etienne, France
  in Laboratoire Hubert Curien
  in the Data Intelligence research Team

- Research fields:
  - Machine Learning
  - Grammatical Inference
  - Explainibility

- To reach me

  Mail : `remi.eyraud@univ-st-etienne.fr`

Objective

To answer to this question

What is research?

How to reach this objective?

- Presentation of some basic element about research

↪ do not hesitate to ask questions about research

Research is a permanent exchange between researchers

- You will be evaluated through an exam

Preliminary remarks

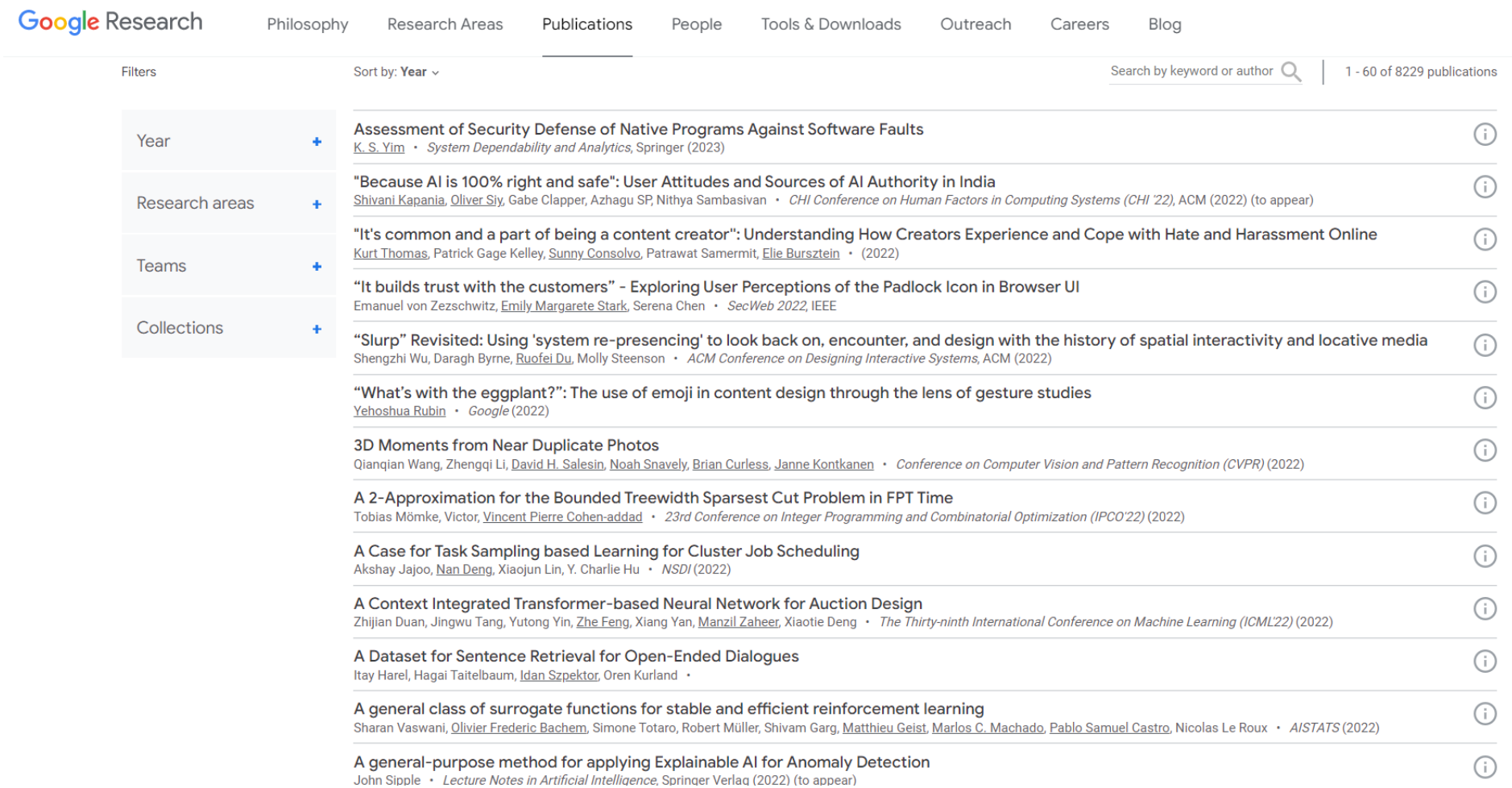
- This course is about research
- Biases
  - I am french
    and I work in a French University/Laboratory
  - I am a researcher
    and a teacher *(Maître de Conférences)*
  - I am a computer scientist in machine learning

Why it matters to everyone

- Computer science is a (relatively) young science
- New discoveries every day!
- Engineering works often rely on recent discovery (Machine Learning and Data Science are good examples but not the only one!)
- Even if you are not planning to work in research, knowing what is going on there will be helpfull (for your work AND your personal life)
- Some of you might be interested by doing by doing a PhD

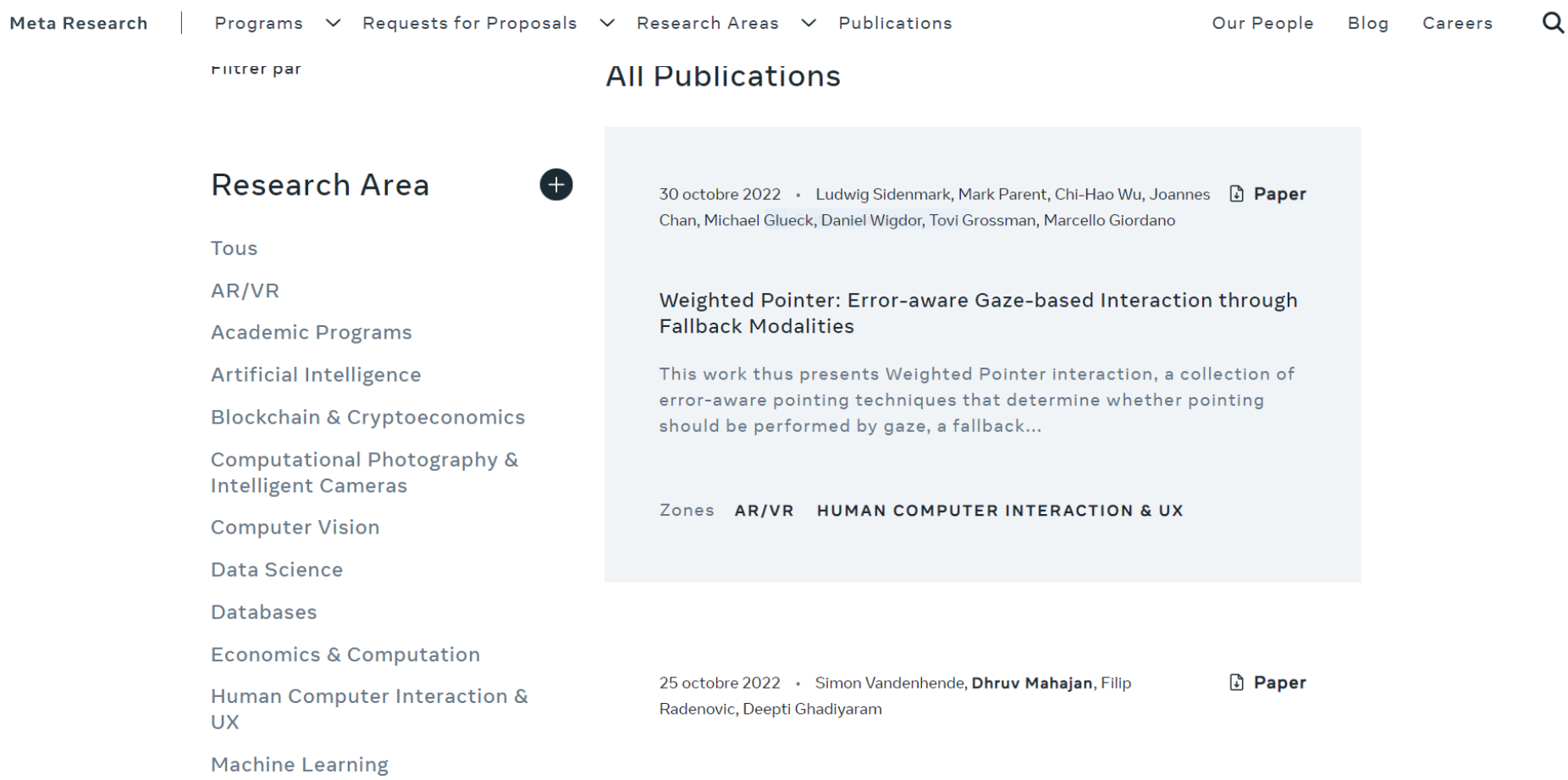
Why it matters to everyone

Most interesting engineering positions are in firms that enforce research



Why it matters to everyone

Most interesting engineering positions are in firms that enforce research



Why it matters to everyone

Most interesting engineering positions are in firms that enforce research

CRITEO AI Lab

About Us | Publications | Blog Posts | Datasets | Events | Programs | Careers

Criteo AI Lab > Publications

Publications

Publications 2018 – today (automatically extracted from HAL, using Haltools)

Publications before 2017

2017

- A Comparative Study of Counterfactual Estimators, T.Nedelec, N. Le Roux and V. Perchet, What If, What Next Workshop, NIPS 2017
- Attribution Modeling Increases Efficiency of Bidding in Display Advertising, E. Diemert, J. Meynet, P. Galland & D. Lefortier, AdKDD & TargetAd Workshop, KDD 2017 **(Best Paper Award Finalist)**
- Cost-sensitive Learning for Utility Optimization in Online Advertising Auctions, F. Vasile, D. Lefortier, & O. Chapelle, AdKDD & TargetAd Workshop, KDD 2017
- Contextual Sequence Modeling for Recommendation with Recurrent Neural Networks, E. Smirnova & F. Vasile, Deep Learning Workshop, RecSys 2017
- Specializing Joint Representations for the task of Product Recommendation, T. Nedelec, E. Smirnova & F. Vasile, Deep

Latest Tweets

Do you like challenges? Every day at 3:30pm, come and test your knowledge of Machine Learning/AI

Why it matters to everyone

Most interesting engineering positions are in firms that enforce research

Engineering
10.03.2022

Automatic Parameter Tuning for Big Data Pipelines

Big data frameworks generally constitute a pipeline, each having a different role. This makes tuning big data pipelines an important yet difficult task given the size of the search space. We propose to use a deep reinforcement learning algorithm to tune a fraud detection big data pipeline.

read more

Data science
20.10.2021

Multimodal Classifier For Space Target Recognition

We propose a multi-modal framework to tackle the SPARK Challenge by classifying satellites using RGB and depth images. Our framework is mainly based on Auto-Encoders to embed the two modalities in a common latent space in order to exploit redundant and complementary information between the two types of data.

read more

Data science
27.09.2021

AMI-Class: Towards a Fully Automated Multi-view Image Classifier

In this paper, we propose an automated framework for multi-view image classification tasks. The proposed

Data governance
27.07.2021

Policy-based Automated Compliance Checking

Under the GDPR requirements and privacy-by-design guidelines, access control for personal data should not be limited to a simple role-based scenario. For the

A true story

- Who: a DSC student
- Context: an internship at a construction firm
- Goal: build a software for the planning (employees, materials, machines, etc.)
- First idea: invent an algorithm $\rightarrow$ failed...
- Second idea: Adapt an existing code $\rightarrow$ failed
- Third idea: Find research work presenting an algorithm for the task $\rightarrow$ failed
- Fourth idea: find a good research work presenting an algorithm for the task $\rightarrow$ success!

What is research?

Definition

The systematic investigation into a topic and study of materials and sources in order to establish facts and reach new conclusions.

(from Concise Oxford English Dictionary)

How to become a researcher?

How to become a researcher ?

In academia

1. Obtain a Master's degree
2. Obtain a Doctorate/PhD
   - In France, in Computer Science: 3 years
3. Work as a "Postdoc" for few years
4. Become an Assistant Professor
   - In France: Maître de Conférences *(with teachings)*, or Chargé de Recherche *(potentially without teaching)*
5. Become a Tenured Professor
   - In France: Professeur, or Directeur de recherche

Outside academia

- Leave the process anywhere after Step 2

Case Study: me

- 2001: **Bachelor's degree in Math & Computer Science (CS)**
  University of Saint-Etienne
- 2003: **Master's degree in CS**
  Ecole des Mines de Saint-Etienne, France
- 2003-2006: **Ph.D. in CS - Machine Learning (ML)**
  University of Saint-Etienne, lab. EURISE
- 2006-2007: **Postdoc in CS - ML**
  University of Amsterdam, The Netherlands
- 2007-2011: **Maître de Conférences in CS - ML**
  Aix-Marseille University - Lab. d'Informatique Fondamentale
- 2011-2015: **Invited Researcher**
  University of Maryland, University of Delaware
- 2016-2020: **Maître de Conférences in CS - ML**
  Aix-Marseille University - Lab. d'Informatique et Systèmes
- Since 2021: **Maître de Conférences HDR in CS - ML**
  IUT of Saint-Etienne - Lab Hubert Curien

What's noteworthy?

- I have stayed in Computer Science field during my career

↪ Many researchers switch fields during their careers
It is really not a big deal

- I worked in different places

↪ Actually, experience abroad is very important
to show open-mindedness and stir up ideas

- I did only few months of Postdoc

↪ I was lucky (and did what we can call a "good" PhD)

A difference between France and other countries

Other countries

- You join as Assistant Professor with a **limited** time contract
- After few years, there is an evaluation
- If you pass, you get promoted to Professor with **permanent** contract (=Tenured)

In France

- You join as Maître de Conférences with a **permanent** contract
- After few years, similarly as a Ph.D. Thesis, you write an Habilitation Thesis (Habilitation à Diriger des Recherches, HDR)
- If you pass, you can apply to a Professeur position

How to have the best chances?

- Study at a good university
- Do your PhD in a good institution/laboratory
  - Apply to several PhD programs, but not too many
    You must be able to **tailor your application**
  - Select programs that fit your interest
    but don't be narrow-minded regarding topics
  - Start early
    up to one year between application and start of the program!
- Do your PostDoc in a good institution/laboratory
  - Similar than for a PhD
  - Change institution, ideally go abroad
  - Change topic: avoid to be stuck in a specific topic...

How to do research?

Steps in Research Process

source http://visual.ly/8-essential-steps-research-process

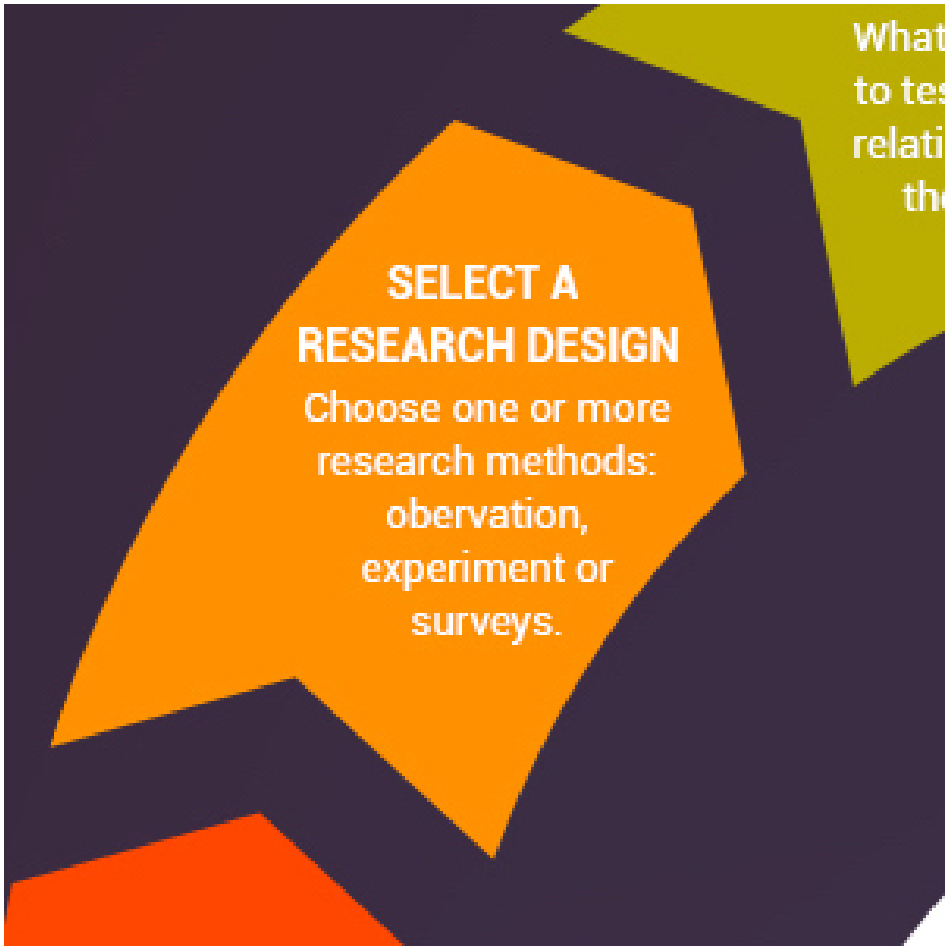
Steps in Research Process

source http://visual.ly/8-essential-steps-research-process

Steps in Research Process

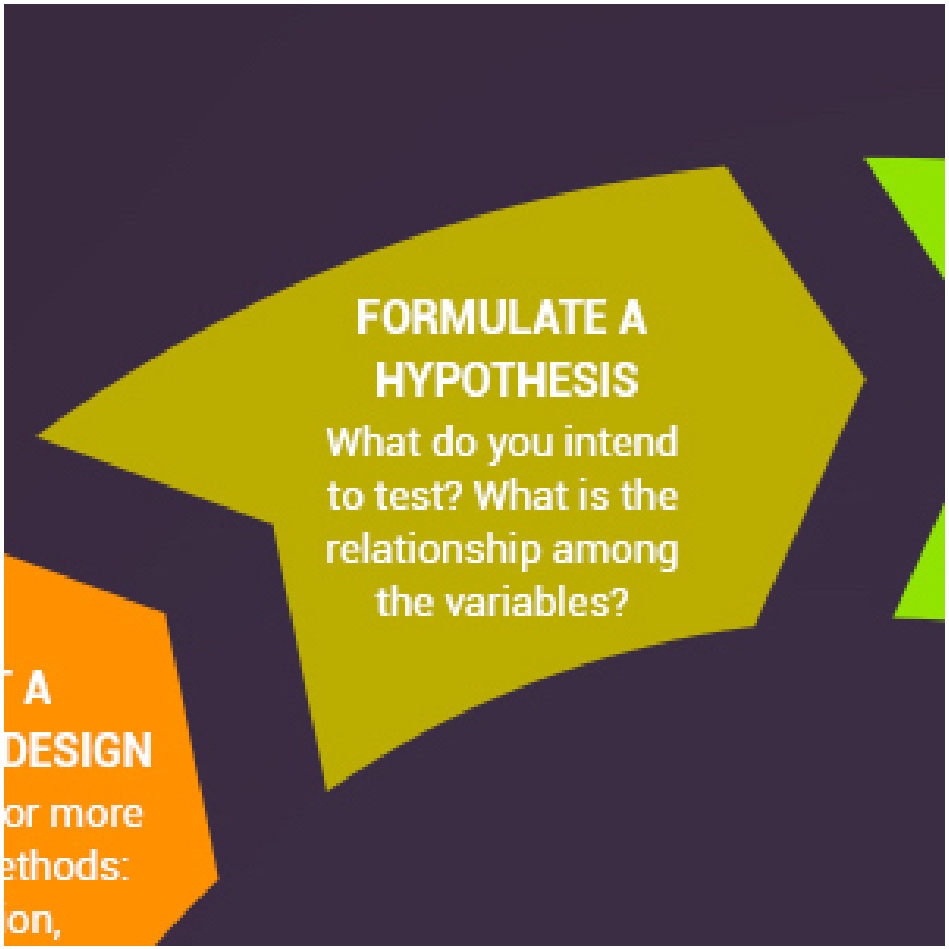
source http://visual.ly/8-essential-steps-research-process

Steps in Research Process



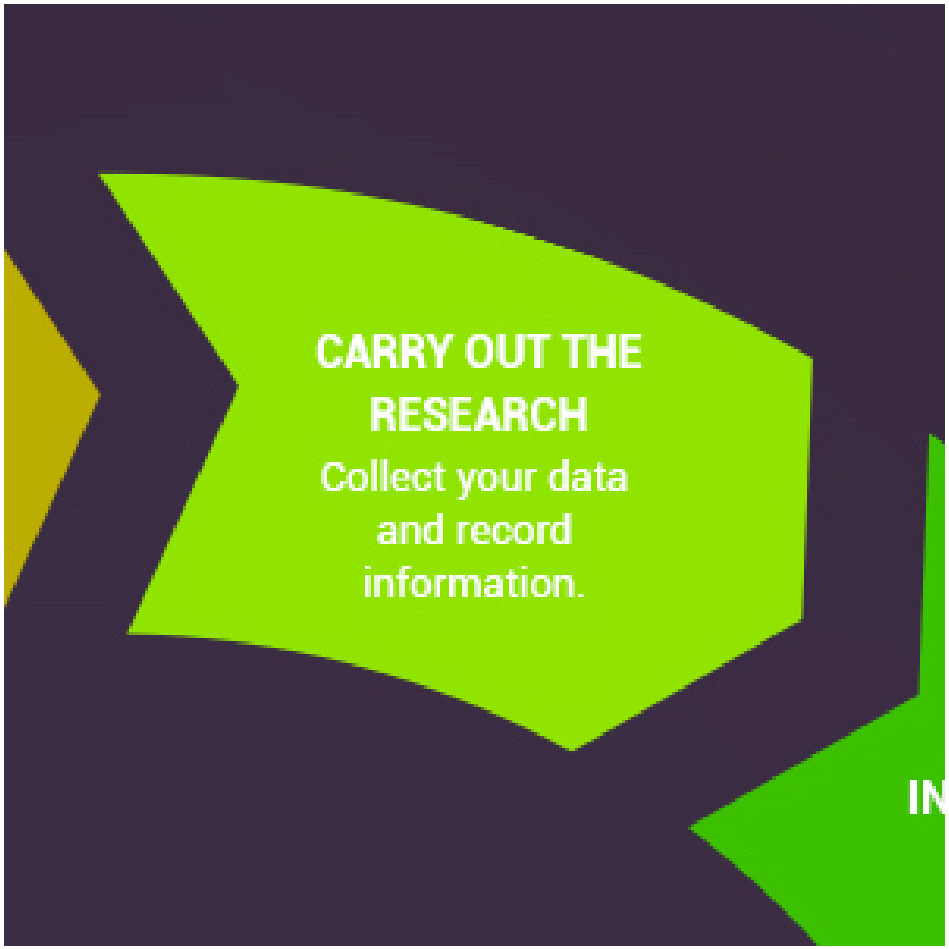
source http://visual.ly/8-essential-steps-research-process

Steps in Research Process



source http://visual.ly/8-essential-steps-research-process

Steps in Research Process



source http://visual.ly/8-essential-steps-research-process

Steps in Research Process

source http://visual.ly/8-essential-steps-research-process

Steps in Research Process

source http://visual.ly/8-essential-steps-research-process

Steps in Research Process

source http://visual.ly/8-essential-steps-research-process

Some Rules on Experiments in CS

Computer science is (in part) an experimental science
with a strong particularity: it is (relatively) easy to experiment

- Two different families of experiments

1. **Discovery experiments**

↪ the goal is to discover, to investigate,
to test hypotheses

2. **Validation experiments**

↪ the goal is to prove that the theory/idea/algorithm/setting is correct

Some Rules on Experiments in CS

WARNING 1

It is easy to cheat and present false results!!

OF COURSE **FORBIDDEN!!**

When you do research you have to be honest on your results (related to ethical issues)

Some Rules on Experiments in CS

WARNING 2

It is easy to produce wrong results
because of a bad experimental setting!!

When you do experiments, you have

- to set correctly and formally what you want to do
- to check if it corresponds to your theoretical setting
- to validate your results
- to check carefully your code
- to say if you are using an already done implementation
- to test on different datasets

Your experiments AND results have to be reproducible
by anyone

Some Rules on Experiments in CS

WARNING 3

Make your experiments reproducible

You have to give to the community

- the details of your setting
- the source code of the experiments/algorithms
- the dataset you used
- everything needed to obtain the same results than you

Some Rules on Experiments in CS

Solution

Nowadays a lot of conferences ask for making the source code of your experiments available

Well that's cool but...

To do research you need funding !

A part of the job is to find "money"

- Essentially done by writing proposals to answer to call for National/International Projects in collaboration with others Laboratories or Companies

How to evaluate research?

Through publications
to "good" conferences and journals

- It is easy to publish in "bad" conferences/journals
- It is hard to publish in "good" conferences/journals

Be careful when you check the publications of someone (*e.g.*, via google scholar or dblp)

How to evaluate if a conference/journal is good?

For Machine Learning and Data Mining conferences

- you can check the core rank website
  `http://portal.core.edu.au/conf-ranks/`
  (A, A, B, C, not ranked)*
- you can check the acceptance rate of the conference

For journals

- You can check their "IF" (Impact Factors) or their "SRJ" (Scientific Journal Ranking)

It is not necessarily the best solution,
but it can help you to evaluate publications...

A note on ArXiv.org

What is ArXiv? (wikipedia definition)

ArXiv is an open-access repository of electronic preprints and postprints (known as e-prints) approved for posting after moderation, but **not peer review**.

We can deposit papers that have not been published/evaluated

This means that

- no one checked if the results are correct
- you have to check if the paper has been published (and where)
- if it has been published, do not cite the arxiv version
- if it has not been published, avoid to base your work on it, and thus avoid to cite it as a reference paper

but then why do we put our papers on arxiv before publication?

It is a way to protect your idea/work

How to publish and share your work?

- Go for the best possible journal/conference
- Be sure it is a place where people will understand your work
- Communicate your work in a correct form (writing & orally)
- If it is a good work, repeat the message
  - But do not auto-plagiarize
  - Go to workshops, give talks, make further experiments...

To hear more about the process of scientific articles (7 minutes video):
`https://www.youtube.com/watch?v=K4qY8WvgZOw`

Some particularities of this job

- Meet interesting people
- Discuss interesting questions
- Have the feeling that you are a step ahead
- Share ideas
- *also "free" travel - in a non-COVID-19 situation...*

What's next?

In the following I will give you some insights on

- **How to write a paper/a scientific report**
- **How to do a "good" presentation/talk**

But first, some basics on an powerful tool : LaTeX

Some basics on LaTeX

Introduction

- TEX is a very powerful computer **typesetting** program
  - ↪ created by Donald E. Knuth

- LATEX is a large set of macros (shortcuts) for TEX
  - ↪ written by Leslie Lamport

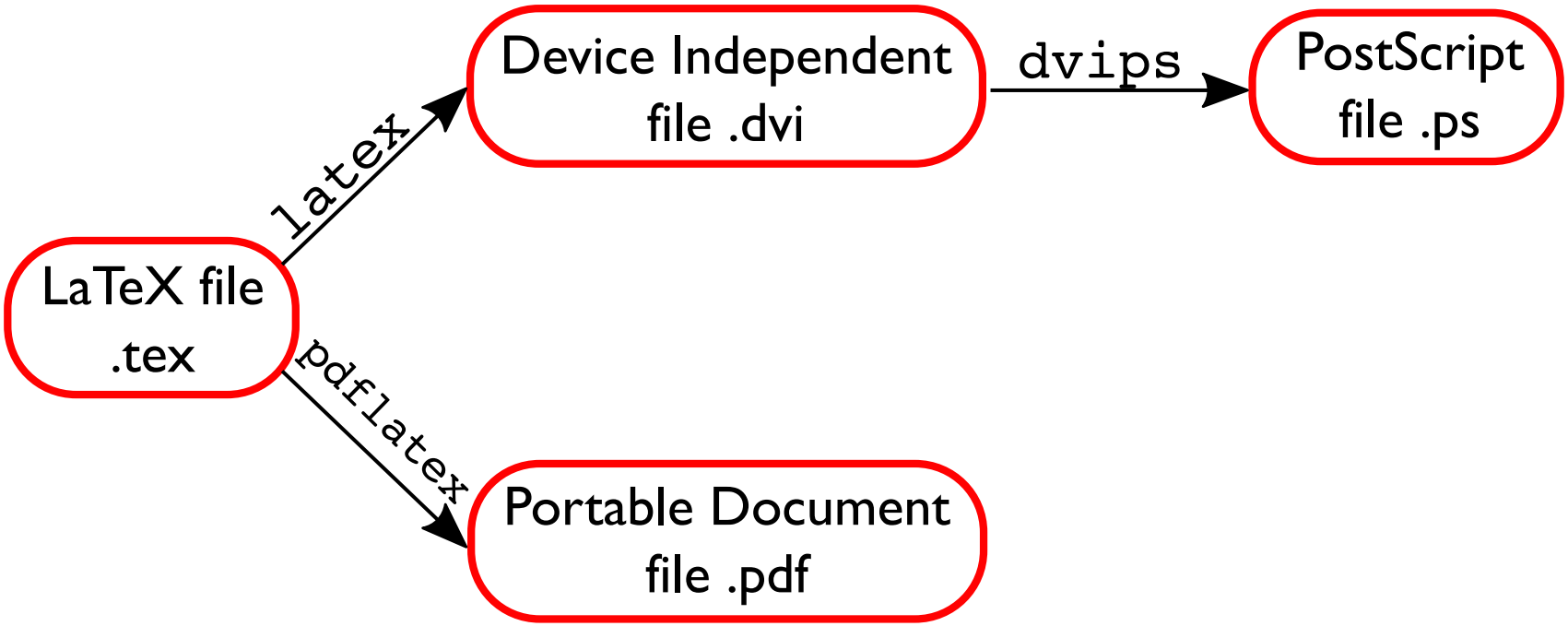
Basic principle

- Same principle than `html`
- Plain text that contains the text you want to typeset
- You indicate the way you want typesetting done by means of certain special codes, most of which begin with a backslash \

Advantages

- Professionally crafted layouts are available
  - make the document look as if "printed"
- Mathematics are supported in a convenient way
- Only need to learn easy-to-understand commands that specify the logical structure of the document
  - ↪ You never need to play with the layout of the document
- Complex structures (ex. footnotes, references, table of contents and bibliographies) can be generated very easily
- TEX is highly portable and free

Compilation of a LaTeX file



The LaTeX language

- LaTeX commands always start with a backslash \
- Required command arguments are between curly brackets { }
- Optional command arguments are between brackets [ ]
- Comments start with % (and finish at the end of the line)
- LaTeX takes care of the spacing between words and paragraphs
- The commands `\begin{ }` and `\end{ }` create environments

Characters

- Delimiters
  - the space '␣' is the delimiter between two words
  - the line break is the delimiter between two paragraphs
  - several delimiters are equivalent to one delimiter
- Special characters, *i.e.* "means something for LaTeX"
  - $ - delimiter for the math mode
  - \[\] - delimiters for the math mode
  - \ - start of a command
  - % - start of a comment
  - # - argument of a command
  - { } - group of characters
  - ^ - superscript in math mode
  - _ - subscript in math mode
  - ~ - non-breaking space
- Normal characters: the others

Document skeleton

```
\documentclass{article} %type or class of the document

%preamble: general format guidelines

\title{Title of the document} %optionnal
\author{Emilie Morvant} %optionnal
\date{\today}  %optionnal

\begin{document}
\maketitle %to automatically create the title
blablablabla

\end{document}
```

Commands

- A lot of commands are available
- A lot of commands are available **for the math mode**

$\Rightarrow$ The commands automatically format the text

Example

```
I am writing in \textbf{bold}
```

I am writing in **bold**

You need a special symbol? $\rightarrow$ *detexify.kirelabs.org/classify.html*

Exemple: a basic .tex file

```
\documentclass[12pt,a4paper]{article}
\usepackage[english]{babel}

\title{Manchuria Kung Fu}
\author{Emilie Morvant}

\begin{document}
\maketitle

\section{Introduction}
\textbf{Manchuria Kung Fu} is a traditional chinese martial art (kung fu).
This fighting art was secretly developped within the Chinese imperial guard
during the Ming Dynasty originating from Manchuria (1644-1911). More recently,
it spread throughout thanks to Grand Master R. Shekhar. Since his death in
2014, Master Derosi\'ere devotes himself to transmitting this art.

Manchuria Kung Fu is based on $12$ styles derived mainly from animal behavior, and
35 weapons. The origins of this Kung Fu makes it very comprehensive: it includes
not only specificities from the north of China with a lot of legs and jumps
techniques, but also a lot of techniques from the south.

\end{document}
```

Preamble: Document Classes

`\documentclass[<options>]{<class>}`

Classes

`article`: "normal" document

`beamer`: these slides

and a lot of different classes...

Options

`11pt, 12pt`: fontsize

`a4paper`: size of the paper

`twocolumn`: 2 columns

and a lot of different options

Preamble: Document Classes

`\usepackage[<options>]{<package>}`

- `graphicx`: for using images
- `geometry`: for modifying the margins of the document
- `fancyhdr`: for header and footer
- `amsmath`, `amsthm`, `amssymb`: very useful for math
- `algorithm2e`, `algorithmx`: for algorithms

and a lot of other packages

Preamble: Document Classes

`\pagestyle{<style>}`: for all the pages from this one

`\thispagestyle{<style>}`: only for the current page

- `plain`: page number in the middle of the footer
- `headings`: title and page number in the header
- `empty`: an empty page

Organisation of the document

Hierarchical organisation

```
\section{Section title}
  \subsection{Subsection title}
    \subsubsection{Subsubsection title}
      \paragraph{Paragraph title}
        \subparagraph{Subparagraph title}
```

The numbering is automatically done!!!!

Fonts, Sizes (non-exhaustive list)

`\textrm{roman}`	roman
`\textbf{bold}`	**bold**
`\texttt{typewriter}`	`typewriter`
`\textit{italic}`	*italic*
`\emph{text to emphase}`	*to emphase*
`\textsc{small capitals}`	SMALL CAPITAL
`{\tiny tiny}`	tiny
`{\scriptsize very small}`	very small
`{\footnotesize quite small}`	quite small
`{\small small}`	small
`{\large large}`	large
`{\Large larger}`	larger
`{\LARGE even larger}`	even larger
`{\huge huge}`	huge
`{\Huge enormous}`	enormous

Environment

```
\begin{<environment>}
.
.
\end{<environment>}
```

Few useful environments

center, flushleft, flushright, itemize, enumerate, table, tabular, figure, equation, align, theorem, minipage ...

Examples

```
\begin{itemize}
\item first item
  \begin{enumerate}
  \item first item \footnote{one footnote}
  \item second item
  \end{enumerate}
\item second item
\end{itemize}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{Images/image.pdf}
\caption{\label{fig:image} This is the caption
                        of Figure \ref{fig:image}.}
\end{figure}
```

Examples

- first item
  1. first item[1]
  2. second item
- second item

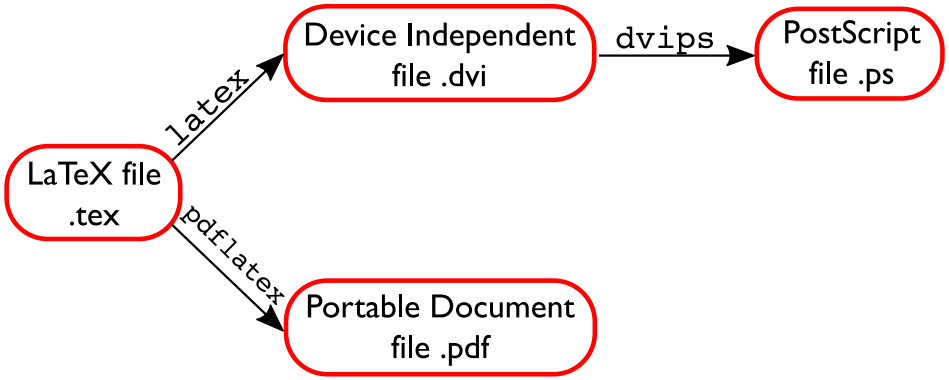


Figure 1: This is the caption of Figure 1.

[1] one footnote

Examples

```
\begin{table}
\begin{tabular}{|r||c|l|}
\hline
hello & world & !!!\\
\hline
My & name is & R\'emi\\
\hline
\end{tabular}
\caption{Caption of this table.}
\end{table}

\begin{align}
\sum_{i=1}^n i &=  1 + 2 + 3 + \cdots + n \label{eq:line1}\\
& = \frac{n\cdot(n+1)}{2} \label{eq:line2}
\end{align}
Equation \eqref{eq:line1} \textit{equals} Equation \eqref{eq:line2}
```

Examples

hello	world	!!!
My	name is	Rémi

Table: Caption of this table.

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + \cdots + n \tag{1}$$

$$= \frac{n \cdot (n+1)}{2} \tag{2}$$

Equation (1) *equals* Equation (2)

The bibliography file

The bibliography file (`biblio.bib`) contains the list of references

```
@inproceedings{ICML16,
  TITLE = {{A New PAC-Bayesian Perspective on Domain Adaptation}},
  AUTHOR = {Germain, P. and Habrard, A. and Laviolette,
  F. and Morvant, E.},
  BOOKTITLE = {{ICML 2016}},
  YEAR = {2016},
}
@article{Neurocomp17,
  TITLE = {{Risk Upper Bounds for General Ensemble Methods with an
           application to Multiclass Classification}},
  AUTHOR = {Laviolette, F. and Morvant, E. and  Ralaivola, L. and Roy, J.-F.},
  JOURNAL = {{Neurocomputing}},
  PUBLISHER = {{Elsevier}},
  VOLUME = {219},
  PAGES = {15-25},
  YEAR = {2017},
}
```

To Cite a Reference

`\cite{ICML16}`	Germain et al. (2016)
`\cite*{ICML16}`	Germain, Habrard, Laviolette, and Morvant (2016)
`\citealt{ICML16}`	Germain et al. 2016
`\citealt*{ICML16}`	Germain, Habrard, Laviolette, and Morvant 2016
`\citet{ICML16}`	Germain et al. (2016)
`\citet*{ICML16}`	Germain, Habrard, Laviolette, and Morvant (2016)
`\citep{ICML16}`	(Germain et al., 2016)
`\citep*{ICML16}`	(Germain, Habrard, Laviolette, and Morvant, 2016)
`\citep[p99]{ICML16}`	(Germain et al., 2016, p99)
`\citep[eg][]{ICML16}`	(eg Germain et al., 2016)
`\citep[eg][p99]{ICML16}`	(eg Germain et al., 2016, p99)
`\citeauthor{ICML16}`	Germain et al.
`\citeauthor*{ICML16}`	Germain, Habrard, Laviolette, and Morvant
`\citeyear{ICML16}`	2016

Example

```
\documentclass[12pt]{article}
\usepackage{natbib} %or another biblio package

\begin{document}

\section{Introduction}
You refer to \cite{ICML16} or \cite{Neurocomp17}, or you cite a
specific page: see \citet[p.2]{ICML16}, or you cite more than one
paper: \cite{ICML16,Neurocomp17}, or you make a parenthetical
reference to one or more articles by omitting the parentheses
around the year (\citealt{ICML16}).

\bibliographystyle{te}
\bibliography{research}

\end{document}
```

I Introduction

You refer to Germain et al. (2016) or Laviolette et al. (2017), or you cite a specific page: see Germain et al. (2016, p.2), or you cite more than one paper: Germain et al. (2016); Laviolette et al. (2017), or you make a parenthetical reference to one or more articles by omitting the parentheses around the year (Germain et al. 2016).

References

Germain, Pascal, Amaury Habrard, François Laviolette, and Emilie Morvant (2016), "A New PAC-Bayesian Perspective on Domain Adaptation." In *ICML 2016*.

Laviolette, François, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy (2017), "Risk Upper Bounds for General Ensemble Methods with an application to Multiclass Classification." *Neurocomputing*, 219, 15–25.

Typesetting citations

There are 3 major citation styles

- Harvard style: (Smith et al., 1996)
- Code in brackets: numeric [16] or generated [AMS+96] (to avoid)
- Superscripts: [16] (to avoid)

Citations should not be used as a noun

✗ [16] conducted a study.

✓ Smith et al. [16] conducted a study.

✗ (Smith et al., 1996) conducted a study.

✓ Smith et al. (1996) conducted a study.

Citations are also typically listed after a statement

✓ Better result is possible with hash methods [16, 30].

✗ Better result is possible with hash methods Smith et al. (1996).

✓ Better result is possible with hash methods (Smith et al., 1996)

Compilation of a LaTeX File with References

To compile the file `file.tex` with the bibliography file `biblio.bib`

```
> latex file ; bibtex file ; latex file ; dvips file
```

Or

```
> pdflatex file ; bibtex file ; pdflatex file
```

Beamer-to create slides

For slides: Beamer!

To learn on your own

How to write a paper

inspired from the tutorial "The data scientist's guide for writing paper"

by Nikolaj Tatti

Why Should We Care about Writing?

Writing is personal and unique

Reading is personal and depends on the background

Some find paper easy to read, some find paper hard to read

So why should we care ?

If you do not care about your document,
then you do not care about the reader!

The more you know about writing,
The less time it takes you to write

Objective of This Part

This course will give you

- Guidelines that you should follow by default
- These guidelines are illustrated on a research paper
- But are similar for a lot of documents

(such as the report of your internship)

Advice

You can break these guidelines, if needed

Be aware of the consequences of your choices

Good rules for writing process

START EARLY: Waiting for the deadline is a cardinal sin

- Do not wait the last moment to write
- Deadline writing introduces errors
- Deadline writing results in bad presentation and organisation of your ideas
- Good paper is a result of several revisions
- Start to write immediately once you have something

Calvin & Hobbes, Bill Watterson

Good rules for writing process

Make things easy for you

- Use LaTeX *(if you can)*, and LaTeX features
  - especially macros that you can modify easily if needed, with the command `\newcommand{}[]{}`

Examples

```
\newcommand{\R}{\mathbb{R}}                                   $\R$ → ℝ
\newcommand{\red}[1]{\textcolor{red}{\sc #1}    \red{Hello!} → HELLO!
```

- Use version control such as **git** *(if you can)*
  - especially with multiple authors
  - especially if you split each section in its own file
- Use real-time collaboration editors such as **overleaf**
- Learn these tools before the deadline ;-)

Good rules for writing process

Learn how to typeset technical content

- Figures
- Tables
- Pseudo-code
- Equations

Revise your text: It can always be improved

- Write a section, leave it and try to forget it, then read it as a reader
- If you get stuck while reading, rewrite it
- Do not fall in love with your text

Good rules for writing process

You have to deal with rejections

Reviewers are rarely 100% wrong or 100% right

- Do not cry out in despair
- Do not say that it is not fair
  - ↪ Try to find the reason why this paper got killed
  - ↪ If the reviewer misundertood, try to find the reason for it

- Decide if you should address the problem

Reviewer's Point-of-View

Reviewers typically look for the followings things:

- What is the problem that the author are tackling/solving?
- How are they tackling it?
- Is the solution logical and correct?
  Is there a better way to solve the problem?
- How is the approach related to the existing work?
- How much the contribution is significant?
- Is the problem relevant?
- Are the experiments reproducible and correct?
- Is it well-written?
- Are there any mistakes?
- Does the paper follow the imposed format?
- Is the paper relevant for the audience?

General Structure of a paper

What is a well-written, mature paper

- The goal of a paper is to **explain the work to the reader**

↪ It is **not** a chronological report on what you have done

↪ It should have a clear, logical story

- A paper is **not** an ad: Be honest, and neutral

↪ It should be **an objective take** on the topic

↪ Discuss the benefits and the limitations of your approach

↪ Be critical of your work

- Think about the reader when you are writing

General Structure of a ML/DM paper

Abstract: Very short summary

1. Introduction

   general presentation of the context, the existing work, the contribution
2. Problem Definition/Framework: definitions, notations, setting...
3. Contribution (theoretical and/or algorithmic) section(s)
4. Experimental section(s)
5. Conclusion and Perspectives
6. References
7. Appendices

This structure is specific to ML and DM papers

NB: Sometimes there is a section dedicated to the related works

Introduction

The introduction should explain **in plain text**

- Which problem are you solving (describe the existing work)
- Why this is an interesting problem
- An overview on how you solve it
- Why your approach is logical
- How is it new regarding the existing work
- What are your main (theoretical/empirical) results
- The organisation of the paper

After reading the introduction, the reader should know the paper on a high level, and what is the "**take home message**"

Problem Definition/Framework

The problem/framework definition section should present

- At least the minimal notations/definitions needed to describe and understand
  - the setting
  - the input/output data
  - the problem that you are tackling
- The problem definition (as soon as possible), stated formally

NB: Definitions related to your solutions should be in the theory section

Contribution(s) (theory and/or experiments)

Theory sections should contain

- Theoretical analysis of the problem *(if any)*
- Approach(es) to solve the problem *(if any)*
- Theoretical analysis of the algorithms *(if any)*

Experimental sections should contain

- An accurate and precise description of your setup
  ↪ **experiments have to be reproducible by the reader**
- An objective discussion about the obtained results

Conclusion and Perspectives

The conclusion should **summarize**

- Your contribution

  (you assume that the reader have read your paper)
- The advantages and disadvantages of your approaches

Perspectives should present the possible extensions

- On what you want to work
- To illustrate the usefulness of your contribution

Important Remarks on the Related Work Section

- A related works section can be present
  after the introduction or before the conclusion
- Sometimes the whole related work is within the introduction
- Some insist to have the related work after the introduction
- However it is easier to compare to your work,
  once you have given the formal definition

The Golden Rule

Treat the related works
the way you want your work to be treated!

Related Work

To compare your work to related works you have to

- Explain the connections to the prior work
- Do not dismiss related works if its setting is a variant of your setting
  - see if it can be still used
  - if not, then explain why
- Explain when your method is better
- Explain when your method is worse
- Compare theoretically, if you can
- Compare experimentally, it if makes sense

Do Not Hesitate to Give Examples

Goal of an example

To make sure that the reader understood the statement

The examples should be

- Complex enough to clarify the definition
- Simple enough to be understandable

How to Structure the Paper

- Paper must have a logical structure
- Explain the structure explicitly to the reader
- At the beginning of a section (or just before),
  answer why this section is needed
- Before or just after a definition/proposition/theorem
  explain why you need it and how you plan to use it
- Section and subsection titles should be specific

The reader should know where he is and where he is going

How to Structure the Paper-Common ways

There are many ways of structuring your work

- By chain
- By specificity
- By example
- By complexity

How to choose one way?

- The appropriate way depends on the nature of the work
- They are not mutually exclusive

How to Structure the Paper-By Chain

By Chain

The order is mandated by the topic

1. Problem definition
2. Previous approaches
3. Your approach
4. Demonstration that you do better than the baselines

How to Structure the Paper-By Specificity

1. High-level description
2. Descriptions of lower layers

An example of a structure by specificity

Description of the main algorithm

Then, description of the subroutines

How to Structure the Paper-By Example

By Example

To make your approach more concrete

through to a simple example

1. Apply your approach to a small and familiar problem

2. Introduce your approach

How to Structure the Paper-By complexity

By Complexity

To ease the learning curve,

especially in the case of complex notations

1. Introduce a simple/special case of your framework

2. Extend it to a more complex/general case

Style

The Main Problem in Writing

The goal of a paper is to explain your work but...

- It is **easy** to explain **simple** concepts in **simple** terms
- It is **easy** to explain **hard** concepts in **hard** terms
- It is **hard** to explain **hard** concepts in **simple** terms

$\rightsquigarrow$ **This is the main difficulty among novice writers**

Grammar is typically not the bottleneck

It is hard to write, and depends on the context

Only comes with experience !!!

General Tone

Guidelines (that you can break for a good reason)

- One idea per sentence
- Short sentences with simple structure
- Short paragraph
- One topic per section
- Use short words
- Avoid buzzwords (e.g. think outside the box), cliches, slang
- Be specific, not vague or abstract
- Avoid excess in length or style
- Omit unnecessary material

Economic Writing

- The length of the paper should reflect its content
- You often have a limited number of pages

> Every sentence should be necessary

Example:

✗ The volume of information has been rapidly increasing in the past few decades. While computer technology has played a significant role in encouraging the information growth, that latter has also had a great impact on the evolution of computer technology in processing data throughout the years. Historically, many different kinds of databases have been developed to handle information, including the early hierarchical and network models, the relational model, as well as the latest object-oriented and deductive databases. However, no matter how must these databases have improved, they still have their deficiencies. Much information is in textual format. This unstructured style of data, in contrast to the old structured record format data, cannot be managed properly by the traditional database models. Furthermore, since so much information is available, storage and indexing are not the only problems. We need to ensure that relevant information can be obtained upon querying the database.

Economic Writing

- The length of the paper should reflect its content
- You often have a limited number of pages

Every sentence should be necessary

Example:

✓ The volume of information has been rapidly increasing in the past few decades. While computer technology has played a significant role in encouraging the information growth, that latter has also had a great impact on the evolution of computer technology in processing data throughout the years. Historically, many different kinds of databases have been developed to handle information, including the early hierarchical and network models, the relational model, as well as the latest object-oriented and deductive databases. However, no matter how must these databases have improved, they still have their deficiencies. Much information is in textual format. This unstructured style of data, in contrast to the old structured record format data, cannot be managed properly by the traditional database models. Furthermore, since so much information is available, storage and indexing are not the only problems. We need to ensure that relevant information can be obtained upon querying the database.

Economic Writing

Warning

Do not cut the text too much: Clarity is the priority
and economic writing is means to that!

Example:

✗ Bit-stream interpretation requires external description of stored structures. Stored descriptions are encoded, not external.

✓ Interpretation of bit-streams requires external information such as description of stored structures. Such descriptions are themselves data, and if stored with the bit-stream become part of it, so that further external information is not required.

Voice

Prefer **active** voice over passive voice
And avoid more subtle indirect sentences

- It is easier to write and read
- Useful to distinguish your contribution from related work
- Use passive voice, if necessary

Examples:

✗ The following theorem can now be proved.

✓ We can now prove the following theorem.

✗ Local packet transmission was performed to test error rates.

✓ Error rates were tested by local packet transmission.

Voice

Prefer **present** tense over past tense

- You can use past tense if needed
- You can mix the two tenses if needed

Examples:

✗ The algorithm had asymptotic cost $O(n)$.

✓ The algorithm has asymptotic cost $O(n)$.

✗ The ideas are tested by experiment

✓ The ideas were tested by experiment

✓ Although, the suggests that the Klein algorithm has asymptotic cost $O(n^2)$, in our experiments the trend observed was $O(n)$.

Paragraph Structure

Paragraph should discuss one topic at a time

- The first sentence outlines the argument
- The following sentences elaborate on the topic

Warning: Lists

- Avoid bullet point lists
- Bullet point lists serve well if used correctly
- They tend to highlight the material
- Use it only for important material that needs enumeration

Sentence Structure

Sentences should have simple structure

- 1-2 lines long
- Do not say too much at once
- Avoid information in nested sentences

Example:

✗ In the first stage, which is the backtracking tokenizer with a two-element retry buffer, errors, including illegal adjacencies as well as unregognized tokens, are stored on an error stack for collation into a complete report.

✓ The first stage is the backtracking tokenizer with a two-element retry buffer. In this stage possible errors include illegal adjacencies as well as unregonized tokens; when detected, errors are stored on a stack for collation into a complete report.

Sentence Structure

Sentences should have simple structure

Beware of splitted "if" expressions

✗ If the machine is lightly loaded, then response time is acceptable whenever the data is on local disks.

✓ If the machine is lightly loaded and data is on local disks, then response time is acceptable.

✓ Response time is accetable when the machine is lightly loaded and data is on local disks.

Beware of misplaced modifiers

✗ We collated the responses from the users, which were usually short, into the following table.

✓ The users' responses, most of which were short, were collated into the following table.

Sentence Structure

Sentences should have simple structure

Avoid double negatives

✗ There do not seem to be any reasons not to adopt the new method.

⇒ Here, the impression is that of a condemnation:
we do not like the method but we are not sure why...

- The goal was opposite

✓ The new method is at least as good as the old, and should be adopted.

Sometimes "double" negatives are ok

✓ The two outcomes are not inconsistent

- Sounds like a double negative but it is different than
"The two outcomes are consistent"

Sentence Structure

Organize sentences so that can be parsed/understood

without too much backtracking

✗ Classifying handles can involve opening the files they represent

Without the context of the sentence, **classifying handles** is either **handles for classifying** or **classification of handles**

✓ Classification of handles can involve opening the files they represent

Typically, replacing "-ing" with "-ation of" is a good idea

✗ In this context, developing tools is not an option

✓ In the context, development of tools is not an option

✗ The final line in the table shows that removing features with low amplitude can dramatically reduce costs

✓ The final line in the table show that removal of features with low amplitude can dramatically reduce costs

Parallel Structures

Complementary concepts should explained as parallels,

or they are difficult to relate

✗ In SIMD, the same instructions are applied simultaneously to multiple data sets, whereas in MIMD different data sets are processed with different instructions.

✓ In SIMD, multiple data sets are processed simultaneously by the same instructions are, whereas in MIMD multiple data sets are processed simultaneously by different instructions.

Parallels can be based on antonyms

✗ Access is fast, but at the expense of slow update.

✓ Access is fast, but the update is slow.

Parallel Structures

Lack of parallel structure can result in ambiguity

✗ The performance gains are the result of tuning the low-level code used for data access and improved interface design.

✓ The performance gains are the result of tuning the low-level code used for data access and of improved interface design.

✓ The performance gains are the result of improved interface design and of tuning the low-level code used for data access.

Parallel structures should be used in lists

✗ For real-time response there should be sufficient memory, parallel disk arrays should be used, and fast processors.

✓ Real-time response requires sufficient memory, parallel disk arrays, and fast processors.

Choice of words

Use

- Short words-**Examples**:

firstly	→	first	component	→	part
secondly	→	second	utilize	→	use
			initiate	→	begin

- Direct words
- Non-ambigous words
- Specific words
- Familiar words
- Exact long word rather than an approximate short word

✗ The analysis derives the information about software.

✓ The analysis estimates the resource costs of software.

Choice of words

Avoid

- Slang
- Contraction
  - ✗ can't, don't, it's,...
  - ✓ cannot, do not, it is, ...
- vague terms

↪ it is better to repeat a word than use a vague synonym

- abstract terms

Qualifiers

Do not pile qualifiers on top of another!

Use at most one qualifiers (*e.g.*, might, may, perhaps, likely, could)

✗ It is perhaps possible the algorithm might fail on unusual input.

✓ The algorithm might fail on unusual input.

✗ We are planning to consider possible options to extend our work.

✓ We are considering how to extend our work.

NB: Avoid double negative that acts sometimes as a qualifier

✗ Merten's algorithm is not dissimilar to ours.

Qualifiers

Avoid meaningless qualifiers

✗ There is very little advantage to the networked approach.

✓ There is little advantage to the networked approach.

✗ The standard method is simply too slow.

✓ The standard method is too slow.

Padding

Avoid, if you can, phrases:

- in general
- of course (this can be even insulting)
- it is frequently the case $\rightarrow$ often
- a number of $\rightarrow$ several
- a large number of $\rightarrow$ many
- adjectives can be used as padding

✗ A well-known method such as the venerable quicksort is a potential practical alternative in instances of this kind.

✓ A method such as quicksort is a potential alternative.

Important Remark

Phrases note that and the fact that are not padding, they introduce something that the reader should deduce themselves

Foreign expressions and abbreviations

Avoid foreign expressions (latin expressions included)

↪ not everybody knows them

✗ Some writers feel that use of foreign words is de rigueur because it shows savoir-vivre.

✗ mutatis mutandis, prima facie, circa, mea culpa, vice verca, e.g., i.e.

Avoid abbreviations (unless you have introduced them in the text)

✗ w.r.t.

✓ with respect to

Caption of figures or tables

Captions should be informative and exact

- Tables and figures should be **self-contained**
- Explain the major elements in the table / figure
- Explain the non-trivial elements in the table / figure
- If needed, explain how to read the table
- If needed, explain which direction is better (*e.g*, low values are better)

How to deal with mathematics

Mathematical expressions

Mathematical expressions allow you to explain

exactly, clearly and formally

You can always say it in text, but often it is better to use math

✗ An inverted list for a given term is a sequence of pairs, where the first element in each pair is a document identifier and the second is the frequency of the term in the document to which the identifier corresponds.

✓ An inverted list for a term t is a sequence of pairs of the form (d, f), where each d is a document identifier and f is the frequency of t in d.

✓ Given an integer $n \geqslant 3$, there are no integers $x > 0$, $y > 0$, $z > 0$ such that

$$x^n + y^n = z^n.$$

Mathematical expressions

Be clear and avoid "math as jargon"

✗ Let $\langle S \rangle = \{\sum_{i=1}^{n} \alpha_i x_i | \alpha_i \in F, 1 \leqslant i \leqslant n\}$. For $x = \sum_{i=1}^{n} \alpha_i x_i$ and $y = \sum_{i=1}^{n} \beta_i x_i$, so that $x \in \langle S \rangle$ and $y \in \langle S \rangle$, we have $\alpha x + \beta y = \alpha(\sum_{i=1}^{n} \alpha_i x_i) + \beta(\sum_{i=1}^{n} \beta_i x_i) = \sum_{i=1}^{n} (\alpha\alpha_i + \beta\beta_i) x_i \in \langle S \rangle$.

✓ Let $\langle S \rangle$ be a vector space defined by

$$\langle S \rangle = \left\{ \sum_{i=1}^{n} \alpha_i x_i | \alpha_i \in F \right\}.$$

We now show that $\langle S \rangle$ is closed under addition. Consider any two vectors $x \in \langle S \rangle$ and $y \in \langle S \rangle$. Then $x = \sum_{i=1}^{n} \alpha_i x_i$ and $y = \sum_{i=1}^{n} \beta_i x_i$. For any constants $\alpha \in F$, $\beta \in F$, we have

$$\begin{aligned} \alpha x + \beta y &= \alpha \left(\sum_{i=1}^{n} \alpha_i x_i \right) + \beta \left(\sum_{i=1}^{n} \beta_i x_i \right) \\ &= \sum_{i=1}^{n} \left(\alpha\alpha_i + \beta\beta_i \right) x_i, \end{aligned}$$

so that $\alpha x + \beta y \in \langle S \rangle$.

Inline Vs Display

Mathematical formulas can be written

- inline : `$f(x) = x^2 + ax + b$` gives $f(x) = x^2 + ax + b$
- in display : `$$f(x) = x^2 + ax + b$$` gives

$$f(x) = x^2 + ax + b$$

Inline and display math are grammatically equivalent

⇒ **They have to be treated as a part of a sentence!**

✗ This allows us to define $f(x)$ as:

$$f(x) = \sum_i x_i \log x_i + C$$

where C is a normalization constant.

✓ This allows us to define $f(x)$ as

$$f(x) = \sum_i x_i \log x_i + C,$$

where C is a normalization constant.

Displays

You can also write text in a display

✓ This allows us to define $f(x)$ as

$$f(x) = \sum_i x_i \log x_i + C\,, \qquad \text{where } C = \sum_j g(x_j).$$

- Text should be wrapped in `\text{}`, `\textrm{}`, or `\mbox{}`
- Use decent space between two equations,

  e.g., with `\quad` or `\qquad`

Displays and Multi-Lines Equations

In LaTeX to deal with multi-lines equations you can use:

- The line separator is `\\`
- `\begin{align} \end{align}` → Use `&` to align each lines
- `begin{eqnarray} \end{eqnarray}` → To use like a tabular `lcr`

✗

$$\alpha x + \beta y = \alpha \left(\sum_{i=1}^{n} \alpha_i x_i \right) + \beta \left(\sum_{i=1}^{n} \beta_i x_i \right) \tag{3}$$

$$= \sum_{i=1}^{n} \left(\alpha \alpha_i + \beta \beta_i \right) x_i, \tag{4}$$

✓

$$\alpha x + \beta y = \alpha \left(\sum_{i=1}^{n} \alpha_i x_i \right) + \beta \left(\sum_{i=1}^{n} \beta_i x_i \right) \tag{5}$$

$$= \sum_{i=1}^{n} \left(\alpha \alpha_i + \beta \beta_i \right) x_i, \tag{6}$$

NB: if you add a $*$ in the name of the environment, the lines will not be numbered

Writing Formulas

Break down equations by using helper functions

✗ $f(x) = e^{-\frac{b}{a}x\sqrt{1-\frac{a^2}{x^2}}}$

✓ $f(x) = e^{2g(x)}, \quad \text{where } g(x) = -\frac{b}{a}x\sqrt{1-\frac{a^2}{x^2}}$

Introduce variables before using them

✓ Let

$$g(x) = -\frac{b}{a}x\sqrt{1-\frac{a^2}{x^2}},$$

and

$$f(x) = e^{2g(x)}.$$

Writing Formulas

Symbols in different formulas must be separated by words

✗ Consider S_q, $q < p$.

✓ Consider S_q, where $q < p$

Do **not** start a sentence with a symbol

✗ $x^n - a$ has n distinc zeroes.

✓ The polynomial $x^n - a$ has n distinc zeroes.

Use 'that' when it helps to parse a sentence

✗ Assume A is a group.

✓ Assume that A is a group.

Typically, assume and presume should follow with that. However,

✗ We have that $x = y$.

✓ We have $x = y$

Notations Design

What is a good notation?

- It has to allow to express yourself clearly
- It has to be easy to read

Modify your notations, it they does not work!!

$\Rightarrow$ Easy to do with the `\newcommand{}[]{}` command

Notations Design

Main Principle

- Objects of main focus should be well-presented in the notation
- Secondary objects should be in the background, or omitted
- When defining, state explicitly all the dependencies
- If needed, you can drop the trivial dependencies

✗ Given a clustering $\mathcal{C}$ of data D, we define the score $q_{\mathcal{C}}(D)$.

✓ Given a clustering $\mathcal{C}$ of data D, we define the score $q(\mathcal{C}, D)$.

↪ **If D is clear from the context**, we can write $q(\mathcal{C})$.

Notations Design

Avoid

- Parameters as subindices/superindices

- Long names

↪ Long names can be useful but can clutter the text

  - If you use an object often, use a short name

Subscripts

- Subscripts need to be justified
- Unnecessary indices is a waste of resources
- Use indices only if you are going to use the indices

✗ Let $(x_i, x_j) \in X^2$.

✓ Let $(x, y) \in X^2$.

✓ Let $(x_i, x_j) \in X^2$, where $i < j$.

- Prefer i and j than i_1 and i_2

✗ Let $(x_{i_1}, x_{i_2}) \in X^2$, where $i_1 < i_2$.

- Avoid subsubscripts and/or supersubscripts

✗ $x_{i_j}^{k_a^n}$

- You can use subindices/superindices to create instances

✓ $q_{greedy}(x)$, and $q_{opt}(x)$ are instances of the quality score $q(x)$.

Accents

- It is better to avoid accent, i and j is better than i and i'
- Do not combine accents

✗ $\bar{\hat{a}}$, $D'^{j'}_{i'}$.

- Avoid multiple accents

✗ a'', a'^4.

They work

- if you are low on symbols
- if you want to tie some symbols together

✓ $a^* = \max A$, and $a \in A$

✓ Let a' be alternative solution.

Alphabet

Mathematical alphabet consists of

- small letters: a, b, c ... `$a$, $b$, $c$`
- capital letters: A, B, C ... `$A$, $B$, $C$`
- bold small letters: $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ ... `$\bf a$, $\mathbf{b}$, $\bf c$`
- bold capital letters: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ ... `$\bf A$, $\bf B$, $\bf C$`
- caligraphic letters: $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, ... `$\cal A$, $\mathcal{B}$, $\cal C$`
- small greek letters: α, β, γ ... `$\alpha$, $\beta$, $\gamma$`
- capital greek letters: Γ, Δ, Θ ... `$\Gamma$, $\Delta$, $\Theta$`

Avoid

- exotic fonts: $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$, $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ ... `$\mathfrak{A}$ \mathscr{A}$`
- exotic greek letters: prefer α to ξ
- Σ (reserved for a sum), and Π (reserved for a product)

Typecasting

By convention, you should typecast your symbols with alphabet. Typically,

- a, b, c are elements
- A, B, C are sets
- $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ are complex obects, such as distributions
- $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ are vectors
- $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are matrices

By convention, you should typecast the nature of your symbols. Typically

- i and j are indices
- k, l, m, n are integers
- u, v, w are vertors

These conventions can be violated, if needed

LaTeX Typography

- Long variable names should be wrapped in `\mathit{}`

✗ $offer = surface \times force$

✓ $\mathit{offer} = \mathit{surface} \times \mathit{force}$

- Ellipsis obeys the vertical alignment of the operator

✗ $x_1, x_2, \cdots, x_n$ $\cdots$ are produced by `\cdots`

✓ $x_1, x_2, \ldots, x_n$ $\ldots$ are produced by `\ldots` or `\dots`

✓ $x_1 + x_2 + \cdots + x_n$

- $*$ is not a multiplication

✗ $a * b$

✓ ab or $a \times b$ or $a \cdot b$ (order of preference) `\times` for $\times$, `\cdot` for $\cdot$

NB: Be careful $\cdot$ often refers to the dot product between vectors

Theorems and Proofs

Theorems, propositions, Lemmas, Corollaries

Theorem and Proposition

Major statment

`\begin{theorem} \end{theorem} \begin{proposition} \end{proposition}`

Lemma

Intermediate result used to prove a theorem/proposition

`\begin{lemma} ... \end{lemma}`

Corollary

A proposition that follows from one already proved

`\begin{corollary} ... \end{corollary}`

The statements should be exact, ideally self-contained

The proof does not have to follow immediately the statement

↪ You can state a theorem as a goal, and slowly prepare for the proof with lemmas and additional notations

What is a Good Proof

A good proof is

- correct
- a logical sequence of steps
- not vague
- not an argument
- a recipe for the reader
- not just a sequence of formulas
- a guide for the reader towards the argument

- Explain the overall structure of the proof
- Explain the intermediate structure of the proof

Divide and Conquer Proofs

Long proofs typically consist of several large steps
$\Longrightarrow$ Explicit these steps (by enumerating them)

✓ *Proof.*...We prove the result in several steps.
Step *i*...
Step *ii*...

✓ *Proof.*...We prove the result in several steps.
Step *i* (construction of the graph)...
Step *ii* (reduction step)...

In LaTeX use the environment `\begin{proof} .... \end{proof}`

Breaking Proofs into Lemmas

Long proofs typically consist of several large steps
$\Longrightarrow$ An independent step can be written as a lemma

- If a lemma is used only once, then you can place it between the statement of the main proposition and the main proof

✓ **Statement of the main proposition**
To prove the the result, we need several technical lemmas
Lemma 1 with the proof
Lemma 2 with the proof
Proof of main proposition

- If some step is used by several proofs,
  then make that step into a lemma

- If your main proof is in the appendix,
  then the main lemmas can also go there

Reverse Order of the Proof

A proof can consist of two steps:
To prove that X has property P, you need to prove that X has property P', and that P' implies P

- X has property P' is hard to prove
- P' implies P is easy to prove

$\Rightarrow$ Logically, prove the hard step and then the easy step

$\Rightarrow$ However, it is often friendlier to reverse the order

✓ To prove that X has property P, we will prove that X has property P'. To see why this proves the theorem.
We will now prove that X has property P'. **(that can be a lemma)**

Conclusion

Take home messages

- Take writing seriously
- Take time and start early
- Make things easy for yourself, without forgetting the reader
- Write, forget, read, revise, ..., and do it again

- Your paper has to be an objective guide explaining your work
- Be precise, do not waste ink
- Try to write a simpler and clear paper (this is the most difficult part)
- Use math to make things simpler
- Have empathy for the reader
- Be logical, and explain your logic
- Follow the guidelines given by the venue

How to Do a Good Presentation/Talk

...or at least what to avoid

Why Should We Care about Presenting?

Talking is personal and unique

Understanding is personal
and depends on the background
Some find talks and slides easy to follow,
some find talks and slides hard to follow

So why should we care ?

If you do not care about your presentation,
then you do not care about the audience!

The more you know about preparing a talk,
the less time it takes you to prepare

Objective of This Part

This course will give you

- Guidelines that you can follow by default
- What you HAVE to avoid
- These guidelines are similar for a lot of presentations (such as for your internship)

Advice

You can break these guidelines, if needed

Be aware of the consequences of your choices

Good rules for preparing a talk

START EARLY: Waiting for the deadline is a cardinal sin

- Do not wait the last moment to prepare your presentation
- Preparing your presentation just before the deadline
  - introduces errors
  - results in bad presentation of your ideas
- Very good talk is often a result of many revisions & practices

General advice

The audience comes to listen, NOT to read

Unlike paper, the information is transmitted in one go, with no possibility of going back

Slides aim to facilitate the understanding of the audience

$\Rightarrow$ The slides should reinforce your speech, not to repeat it

$\Rightarrow$ A presentation should be

- linear and well-organized
- simple*: simple words, short sentences (avoid multiline sentence)*

Did you know this ?

You normally lose 90% of the audience within the first 5min of your talk

Get information

- Scientific background/level of the audience
- Approximate number of people in the audience
- Instructions given by the organizers

IMPORTANT Remark

Even if you already presented your work elsewhere,

ALWAYS adapt your talk to the venue

How to prepare your talk ?

Define the "storyline"
Identify the take home message

Define the main message of your presentation
(in few words, use simple concepts)

How to prepare your talk ?

Chose a simple style/template for your slides

keep the same template for all your slides

- Do not mix too much colors
  - ↪ up to 3 high contrasted colors per slide
  - ↪ Think about color blind people (5-10% of the population)
  - ↪ color blind rules:

    `www.tableau.com/about/blog/2016/4/examining-data-viz-rules-dont-use-red-green-together-53463`
- Choose a simple background (basically, white or black)
- Prefer to put in **bold** your text to highlight key points
- Choose an easily readable font
  - ↪ *e.g.*, avoid Times New Roman
- Use a large typeface (this is clearly too small)
- DO NOT WRITE A WHOLE SENTENCE IN CAPITAL LETTERS
- Do not use too many animations

How to prepare your talk ?

Organize your presentation

Classical organization:

1. First slide : Title of the presentation, name and affiliation of the speaker, the name of the authors if you present the work of someone else, logos, the place and the date of the presentation
2. Introduction
3. Context - Background
4. Contribution(s) and Result(s)
5. Conclusion and Perspectives
6. Thank you slide
7. References

Organization of your talk-First slide

Use this slide to breathe, to get the audience's attention,
and to restate the purpose of your presentation.
Reading the title is usually useless

RESEARCH METHODOLOGY

Emilie Morvant, Ievgen Redko

Laboratoire Hubert Curien
Université Jean Monnet de Saint-Etienne
Faculté des Sciences et Techniques

FACULTÉ DES SCIENCES ET TECHNIQUES
SAINT-ÉTIENNE

Master 1 MLDM & Master 2 DSC - Semester 1

Organization of your talk-Introduction

Objectives

- Introduce the context and the problem tackled
- Introduce the solutions/contributions
- Introduce the take home message

As for a paper, the audience should know where you are going

- Put figures or images to illustrate what you are saying
- Take common examples
- Do not use complex mathematical formula
- Explain with simple words *(depending on the audience)*

Highlight the take home message

Organization of your talk-Table of contents

A common question

Do we need to put a "table of contents" slide?

- YES **IF** you plan to spend time on it

↪ to "do something else" than reading the table of contents.

- NO **OTHERWISE**

↪ for example if you put the slide just to read the contents
or if you do NOT plan to talk on these slides

Additional advice

- For a short talk (<15min) avoid to put a table of contents

↪ You will loose time on these slides

- Instead of a table of contents, you can put a simple title slide to indicate that you are moving on to another part
- If you put a table of contents slide, place it AFTER the introduction (and not directly after the first slide)

One section per subject

- **At the end of each section**
  - Summarize in few words the main ideas of the section
  - You can highlight a take home message
- **If you present a scientific paper**
  - Do not necessarily follow the organization of the paper
  - Keep in mind that a talk is linear

Organization of your talk-Context

How to present the context, the background ?

- You cannot present all the bibliography during a talk

↪ present "just" what you need to explain the main concepts

↪ put at the end of the talk a slide with at least the main references on which you built your work

- Highlight the problem tackled/the bottlenecks
- Speech:
  - Make links to your contribution
  - Point out what will be important

  In a way: make spoilers of what you will present next

Organization of your talk-Contributions

How to present a theoretical contribution?

- If needed do not hesitate to repeat the messages
- **BE CAREFUL with maths/algorithms** (a talk is linear)

↪ the audience cannot come back to understand the specific notations in your slides

> Contrary to a paper, formalizing difficult result with maths will not make clear your message

- **Solution:**
  - Simplify the notations: use words, ideas, and/or colors
  - Explain with simple concept the difficult theoretical result
  - Then, and only then, can you put the "complete" wording.
    Feel free to draw the parallel with "your simplification".

Organization of your talk-Contributions

How to present empirical results ?

- If needed do not hesitate to repeat the messages

- **In few words** present the key points of the setting

↪ Contrary to a paper, you do not need to present the entire setting

- **BE CAREFUL with tables of results**, contrary to a paper the audience cannot spend time on reading the tables

Solution: Prefer graphics, with lines, bars, etc.

- Highlight the conclusion of the results

IMPORTANT

When you present a figure (or a table), explain how to read it!

Organization of your talk-Conclusion-Perspectives

Conclusion — in few words

- Summarize what you have presented
- Highlight again the take home message(s)

Perspectives — in few words

- Explain what are the most exciting perspectives
- Describe open-questions

One or two slides at most for a 15-20 min presentation

Organization of your talk-Thank you slide

Thank you for your attention.

you can put the references here

Organization of your talk-The hidden slides...

After a presentation there is a serie of questions

- When you prepare your talk, think about the questions
- You can prepare slides to help you to answer

  e.g., details on a setting, additional experiments, proofs of theoretical results, a part you didn't present, etc
- Take care to understand the questions
  - Listen until the end the question
  - Reformulate it to be sure to understand

    (also to be sure the audience understands the question)
  - If you do not know the answer, say: **I don't know**

    Do NOT make up an answer that will not make sense...

VERY IMPORTANT MESSAGES

YOU WILL NOT BE ABLE TO PRESENT EVERYTHING

YOU HAVE TO MAKE CHOICE ON WHAT YOU WILL PRESENT

How many slides?

As baseline, you can count 1 slide per minute

(1 to 2 minutes per slide)

Examples of Bad Slides

Any resemblance to existing or past presentations

is not accidental...

The worst...To avoid without a doubt

Definition

Also to avoid without a doubt

Setting

- $\mathbf{X}$
- $\mathbf{Y}$
- $f(\mathbf{x}) = \mathbf{y}$
- $min_{f \in F} R(f)$

Also to avoid without a doubt

Economic Writing

- The length of the paper should reflect its content
- You often have a limited number of pages

Every sentence should be necessary

Example:

✗ The volume of information has been rapidly increasing in the past few decades. While computer technology has played a significant role in encouraging the information growth, that latter has also had a great impact on the evolution of computer technology in processing data throughout the years. Historically, many different kinds of databases have been developed to handle information, including the early hierarchical and network models, the relational model, as well as the latest object-oriented and deductive databases. However, no matter how must these databases have improved, they still have their deficiencies. Much information is in textual format. This unstructured style of data, in contrast to the old structured record format data, cannot be managed properly by the traditional database models. Furthermore, since so much information is available, storage and indexing are not the only problems. We need to ensure that relevant information can be obtained upon querying the database.

Do not write too much things...

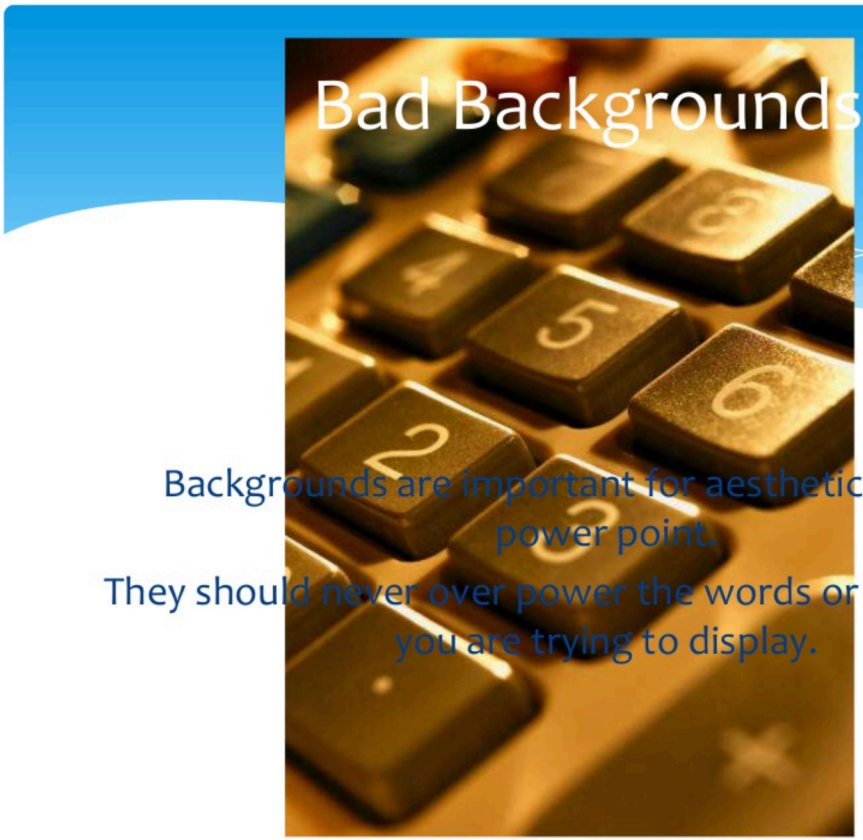
Quelques règles

- Il est important de bien veiller à ce que le texte des diapositives soit le plus simple possible afin d'être compris par touts les auditeurs, et précis pour éviter les éventuelles ambiguïtés d'interprétation. Il est rappelé ici que 75% des écrans sont du texte.
  - Il faut en conséquence choisir correctement son vocabulaire et toujours proposer le mot le mieux adapté à ce que l'on souhaite dire. La syntaxe peut évidemment être simplifiée afin de condenser le texte.
  - Il est également essentiel de soigner le texte des des titres afin de ne pas dérouter l'auditoire par des incohérences entre celui-ci et le contenu de la diapositive.
  - N'oubliez pas d'accentuer l'idée forte de votre diapositive, c'est ce qui doit être retenu en priorité. Le reste n'est qu'enrobage, enjolivure. Dans le pire des cas, si l'auditoire ne doit retenir qu'une chose de la diapositive c'est l'idée forte.
  - Le nombre de mots par ligne doit être le plus petit possible pour ne pas inciter les auditeurs à lire le texte plutôt qu'à écouter le présentateur ou la présentatrice.
    Pensez que tout le monde ne lit pas à la même vitesse. Que feront, à votre avis, ceux qui auront fini de lire les premiers ?... Et à quel moment le présentateur pourra-t'il reprendre la parole sans perturber la lecture ? Quand vous lisez (vos SMS par exemple...), vous n'écoutez plus le prof, et si vous l'écoutez, vous ne pourrez pas lire, et donc le texte ne servira à rien.
  - Également pour des raisons de lisibilité, il est préconisé de ne pas proposer trop de lignes de texte dans une diapositive, sinon vous serez amenés à diminuer la taille de la police et rendre illisible le texte de loin.
- Exercices :
  - Chronométrez la différence de temps de lecture entre cette diapositive (à surtout ne pas prendre comme modèle) et la suivante, où il est dit exactement la même chose, mais de façon bien plus concise. Et demandez-donc à votre chargé de TD de vous les projeter pour bien visualiser la différence !!!
  - Remarque : ce texte peut (doit) être dit lors de la projection du diaporama, mais surtout pas proposé à la lecture.

Montpellier La PréAO et Powerpoint Page 2

Ouch, it stings the eyes

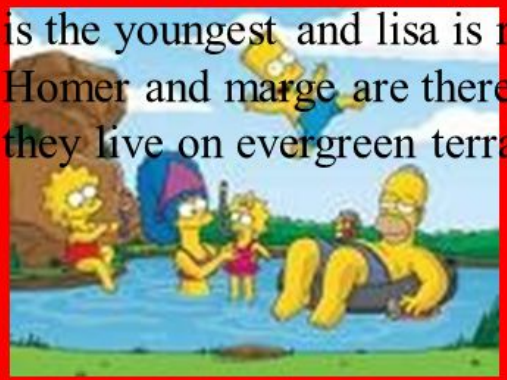
There is a problem with the background



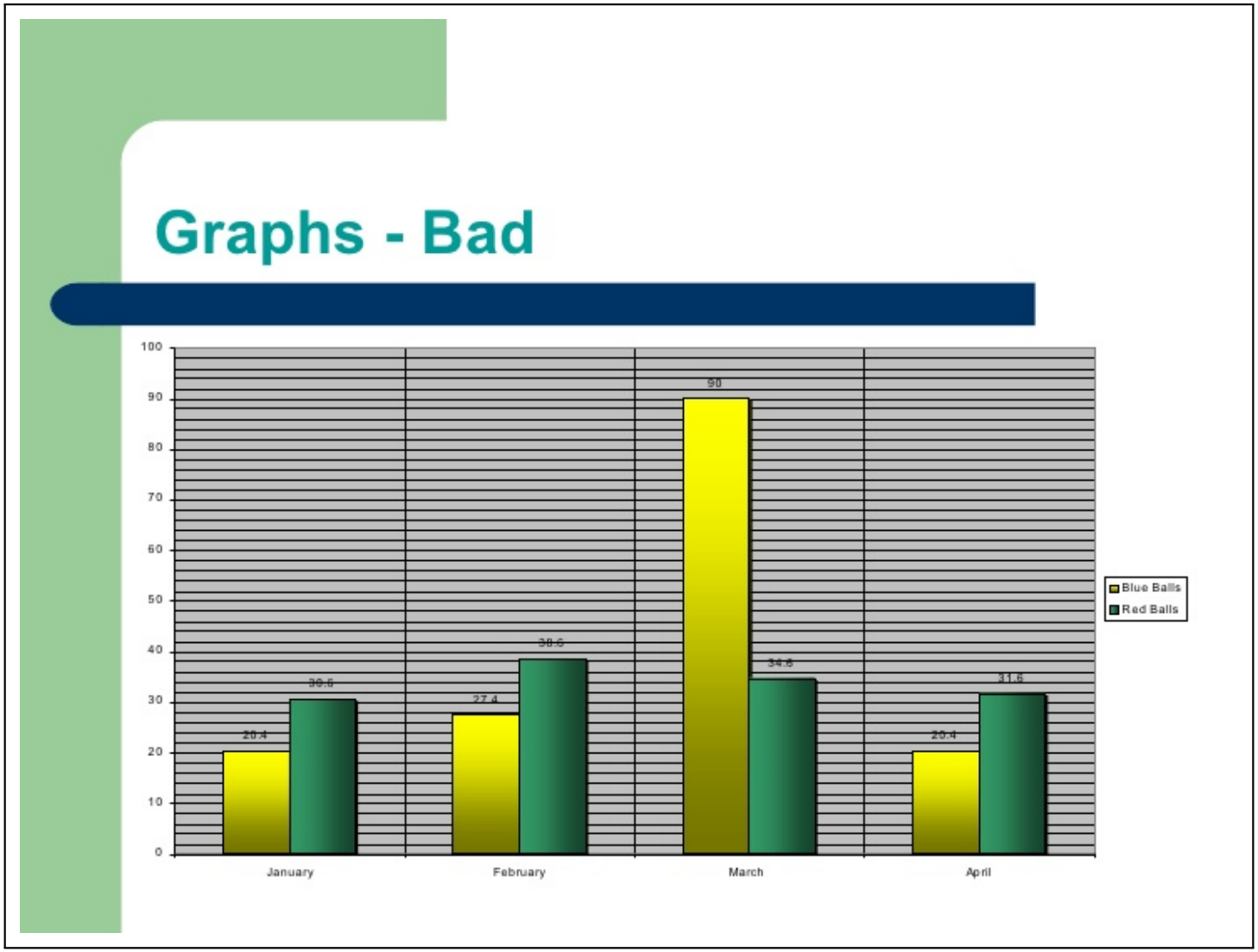
D'oh!

Example of a bad slide

This is a slide about the simpsons who are my favourite cartoon characters because the are funny and there are lots of characters in it and my favourite character is bart. The simpsons have 5 members in there family, bart, lisa, maggie, homer and marge, bart is the oldest out of all the children and maggie is the youngest and lisa is really brainy. Homer and marge are there parents and they live on evergreen terrace.



Sometimes a graph can be unreadable



Pseudo-code is not necessarily clear

Main Algorithm

Algorithm 3 Minimization of $F_{\mathcal{S}}^{e,d}(\mathcal{Q})$ by SGD

Given: learning sample $\mathcal{S}$
prior distribution $\mathcal{P}$ on $\mathcal{H}$
objective function $F_{\mathcal{S}}^{e,d}(\mathcal{Q})$ with $\alpha > 0$
Hyperparameters: learning rate η, batch size m'
number of iterations T

function MINIMIZE-$\mathcal{Q}$
  $\mathcal{Q} \to \mathcal{P}$
  for $t \leftarrow 1$ to T **do**
    Sample $\mathcal{S}' \subseteq \mathcal{S}$ with $|\mathcal{S}'| = m'$
    if $2e_{\mathcal{S}'}(\mathcal{Q}) + d_{\mathcal{S}'}(\mathcal{Q}) \geq 1$ **then**
      $\mathcal{Q} \leftarrow \mathcal{Q} - \eta \nabla_{\mathcal{Q}} \left[2e_{\mathcal{S}'}(\mathcal{Q}) + d_{\mathcal{S}'}(\mathcal{Q})\right]$
    else
      $(e, d) \leftarrow$MAXIMIZE-e-$d(e_{\mathcal{S}'}(\mathcal{Q}), d_{\mathcal{S}'}(\mathcal{Q}))$
      $\mathcal{Q} \leftarrow \mathcal{Q} - \eta \nabla_{\mathcal{Q}} F_{\mathcal{S}'}^{e,d}(\mathcal{Q})$
    end if
  end for
return $\mathcal{Q}$
end function

Bad slide

Empirical results

	Alg.3		Alg.3'		CB-Boost		MinCq	
	$r_{\mathcal{T}}^{\mathrm{MV}}$	Bnd	$r_{\mathcal{T}}^{\mathrm{MV}}$	Bnd	$r_{\mathcal{T}}^{\mathrm{MV}}$	Bnd	$r_{\mathcal{T}}^{\mathrm{MV}}$	Bnd
letter:OvsQ	.11	**.69**	.05	.90	.04	.87	**.02**	1.0
letter:DvsO	.05	**.51**	.03	.81	.04	.84	**.00**	1.0
letter:AvsB	.04	**.32**	.01	.69	.01	.67	**.00**	1.0
credit	.13	**.78**	.13	.79	.13	.89	**.12**	1.0
haberman	.24	**.99**	**.23**	1.0	.24	1.0	.24	1.0
heart	**.18**	**.93**	.23	.98	.21	.96	.19	1.0
glass	.07	**.81**	.06	1.0	.05	.86	**.04**	1.0
tictactoe	.29	**.97**	.08	1.0	.23	1.0	**.02**	1.0
usvotes	.04	**.59**	**.03**	.86	.04	.65	.04	1.0
wdbc	.04	**.61**	**.02**	.87	.05	.84	.03	1.0
adult	.19	.59	.18	.61	**.18**	**.56**	.24	1.0
mnist:1vs7	.03	**.19**	.03	.27	**.02**	.28	.03	1.0
mnist:4vs9	.12	**.48**	.07	.58	**.05**	.60	.07	1.0
mnist:5vs6	.06	**.37**	.04	.46	**.03**	.48	.04	1.0
fash:TOvsPU	.11	**.42**	.09	.47	**.06**	.45	.06	1.0

Bad without any context

The classical PAC-Bayesian Generalization Bound

General PAC-Bayesian Theorem

For any distribution $\mathcal{D}$ on $X \times Y$, any $\mathcal{H}$, any prior distribution π on $\mathcal{H}$, any $\delta \in (0,1]$, any convex function $D : [0,1] \times [0,1] \to \mathbb{R}$, with a probability at least $1-\delta$ over the random choice of the learning sample $S \sim (\mathcal{D})^m$, we have for all posterior distribution ρ on $\mathcal{H}$

$$D\left(\mathop{\mathbf{E}}_{h\sim\rho} R_S(h), \mathop{\mathbf{E}}_{h\sim\rho} R_{\mathcal{D}}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{h\sim\pi} e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right)\right]$$

With $D(a,b) = (a-b)^2$ (McAllester's form)

$$\mathop{\mathbf{E}}_{h\sim\rho} R_{\mathcal{D}}(h) \leq \mathop{\mathbf{E}}_{h\sim\rho} R_S(h) + \sqrt{\frac{1}{2m}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right)}$$

Emilie Morvant | PAC-Bayes & Domain Adaptation | 06/06/2019

This kind of slide is important, for a theoretical work, but if you do not present intuitively the results before, then it makes no sense...

VERY IMPORTANT MESSAGES-the return

- Prepare your speech, repeat, revise, change your slides
- Know your slides, don't give the feeling to discover them
- Speak to the audience, be dynamic
  - Do not talk "to your laptop"
  - Do not read notes on a paper
  - Do not use your slides as a teleprompter: do not write everything you want to say and do not read your slides
  - Look at the audience
  - Avoid to learn by heart all your speech
- Make choice on what you will present
  - You cannot present everything
- Do not prepare to much slides
  - On average, prepare 1 slide per minute

How to design good visualization

Visualization

Three most frequent mistakes in scientific plots:

- the fonts are too small
- axis ranges are wrong
- captions are not detailed enough

Your goal:

- learn how to make a simple plot well
- learn how to present complex plots

Tools for plotting

Prefer Vector graphics

that can be scaled up or down to any resolution with no aliasing

- Images defined in terms of points on a Cartesian plane
- The points are connected by lines and curves
- The points determine the direction of the vector path
- each path may have various properties

(color, shape, curve, thickness, and fill)

Tools for plotting

- Python (recommended for beginners)
- inkscape
- gnuplot
- R
- Pgfplots/tikz

Some basics on "how to read an image"

to help you to design slides/figures/posters

Disclamer

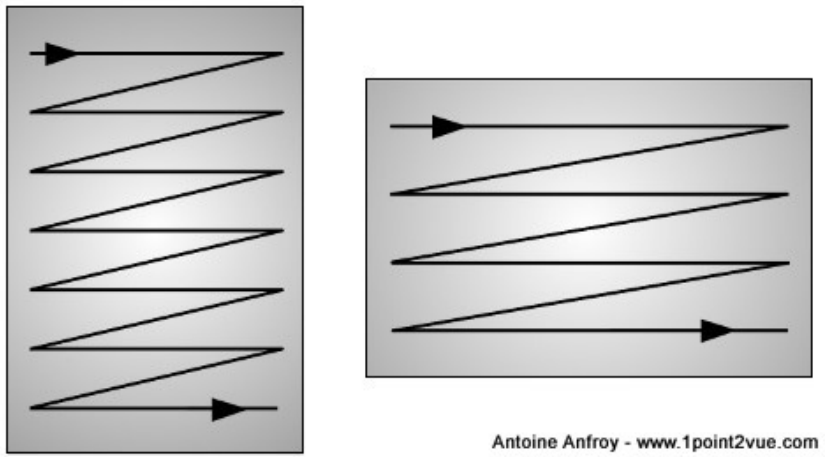
Everything I am going to tell you here is not directly related to the preparation of a talk, but can help you improve your visual communications (and therefore your slides, posters, etc)

Reading direction

Reading direction: "Z"

Our eyes are trained to read a text

$\Rightarrow$ The eyes follow the same principle than reading



Remark: Depending on your native language the reading direction can be different and can lead to a different understanding of an image

Rules of third

Rules of third

due to the reading the eyes focus on particular lines and point of interest in the image

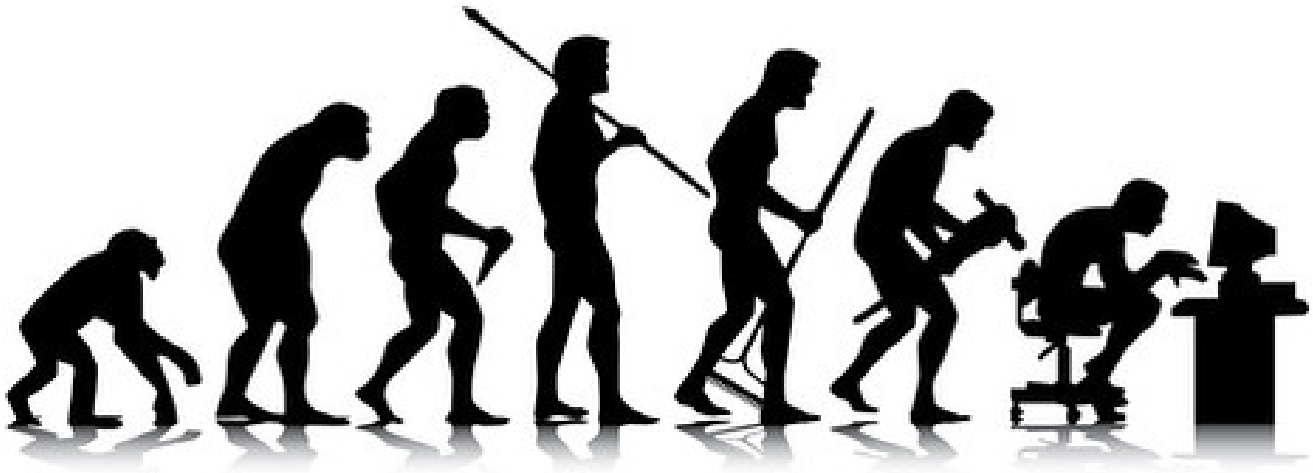
Image guidelines

Image guidelines

the eyes follow the explicit lines in the image (such as a road) and also implicit lines (such as the eyes of people in an image)

Temporality

The linear scanning of the image by the eyes adds a temporal notion



Lessons from The Visual display of quantitative information

by Edward Tufte

Lie factor

Lie factor (LF) is a measure of exaggeration of the interesting quantity in graphics compared to the data

$$LF = \frac{\text{relative change shown in graphics}}{\text{relative change in data}}$$

If the Lie factor is greater than 1.05 or less than 0.95, then graphics distort

- overstating ($LF > 1$)
- understating ($LF < 1$)

Overstating is the most common distortion.

Example: fuel economy standards

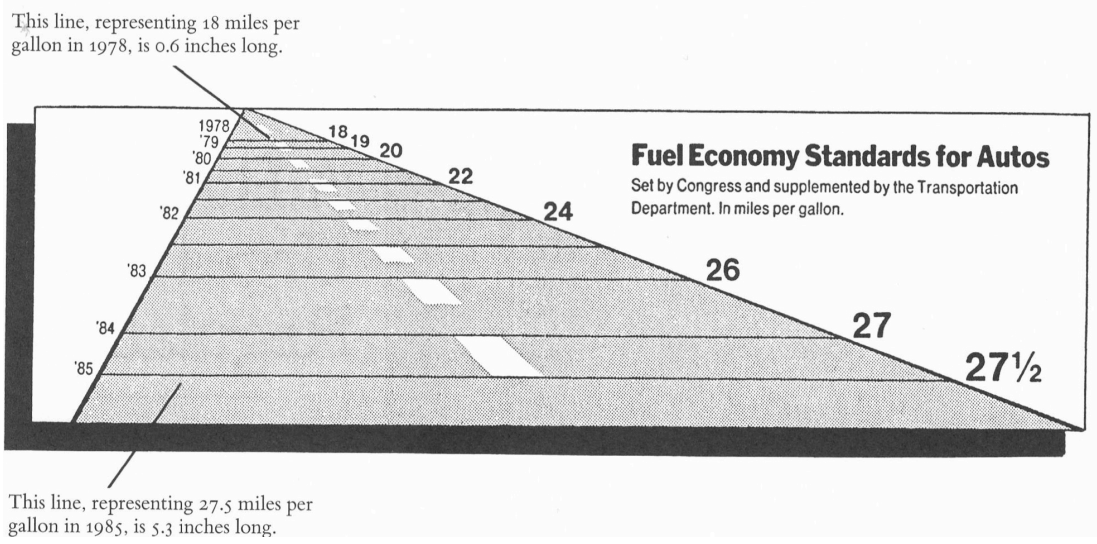
An extreme example in 1978:
series of fuel economy standards to be met by automobile manufacturers, in steps

miles per gallon

- from 18 mpg (in 1978)
- to 27.5 mpg (by 1985)

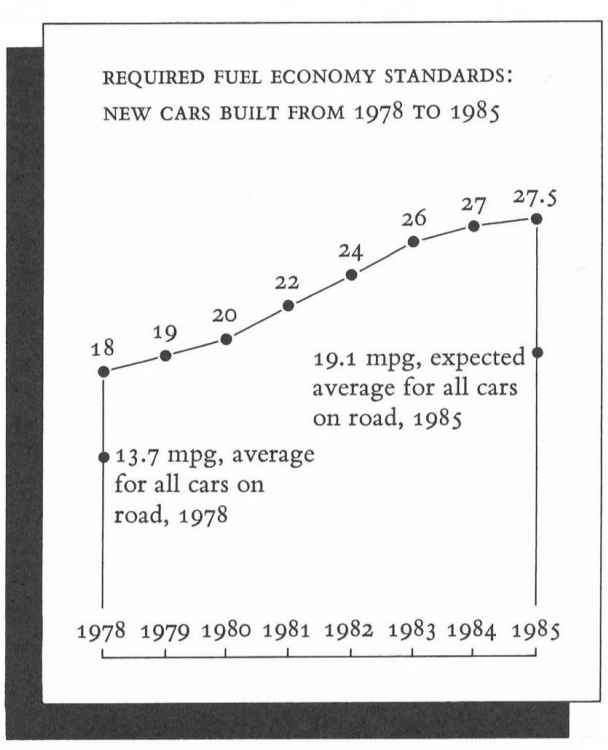
Lie factor $= 14.8$

- $\frac{(27.5-18)}{18} \times 100\% = 53\%$ increase in data
- 783% increase in length



The New York Times (1978)

Example: fuel economy standards (cont.)

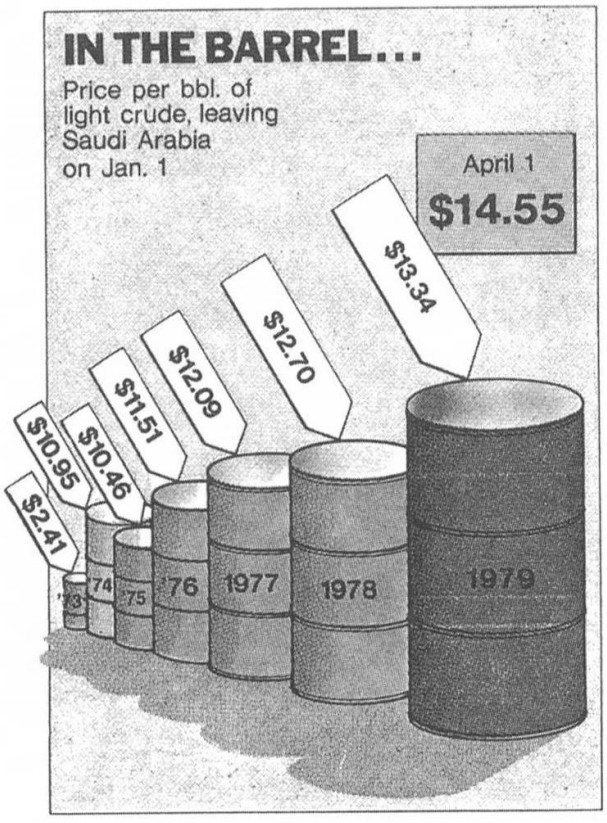


The non-lying version, with proper context

- new cars standards compared with average cars on the road
- plot reveals dynamics of fuel economy
  - slow startup
  - fast growth
  - final stabilization

better reflection of the reality

Visual area and numerical measure



- an increase of 454% is depicted as an increase of 4280%

The viewer gets mixed up by the fact that a barrel (3D) represented by area (2D) is used to show 1D data,

The Time magazine (1979)

Perception of surface

Many experiments on the visual perception of graphics have been conducted

- people look at lines of varying length, circles of different areas and then recording their assessments of the numerical quantities

E.g., the perceived area of a circle grows more slowly than the actual measured area,

$$\text{Perceived area} = (\text{Actual area})^{\alpha}, \quad \text{with} \quad \alpha = 0.8 \pm 0.3 \quad .$$

However, different persons see the same areas somewhat differently

- perceptions can be different depending on the culture
- perceptions change with experience
- perceptions are context-dependent

Data-ink ratio

$$\begin{aligned}
\text{data-ink} &= \text{ink (pixels) used to show data} \\
\text{redundant data-ink} &= \text{redundant ink used to show data} \\
\text{non data-ink} &= \text{remaining ink} \\
\text{data-ink ratio} &= \frac{\text{data-ink}}{\text{total ink}}
\end{aligned}$$

(Disclaimer: these are not formally defined concepts: it is not always clear what type of ink you have)

Data-ink ratio

Tufte:
you should always try to maximize the data-ink ratio, within reason

- every bit of ink on a graphic needs a reason
- nearly always that reason being that the ink presents new information

To increase the proportion of data-ink use two erasing principles

- erase non data-ink
- erase redundant data-ink

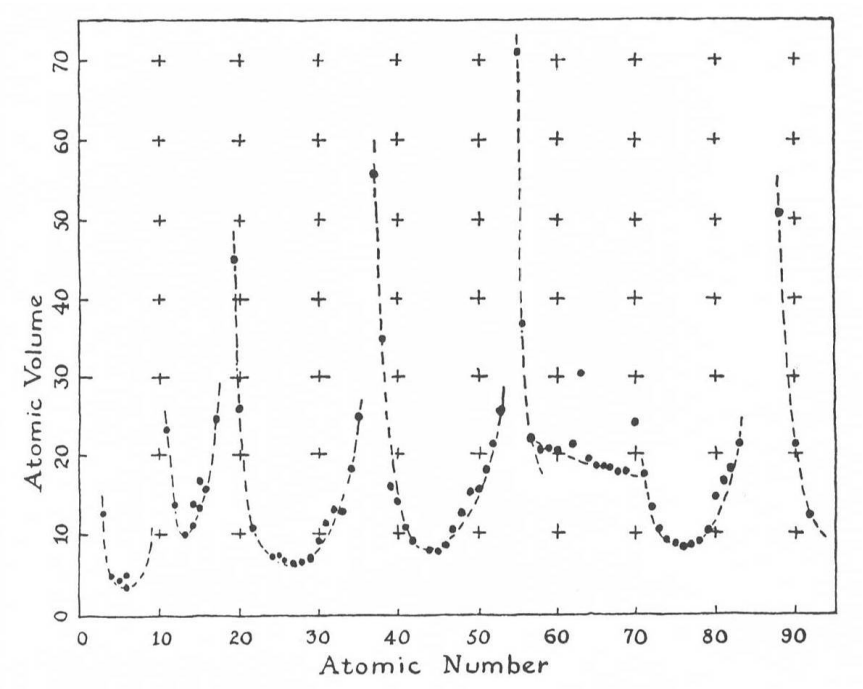
⇒ This will ease the reading and understanding !!!

Edit and redesign (cont.)

The data-ink ratio is about 0.6

- 76 data points and the reference curve are obscured by 63 grid marks

The grid and part of the frame can be erased to improve the data-ink ratio

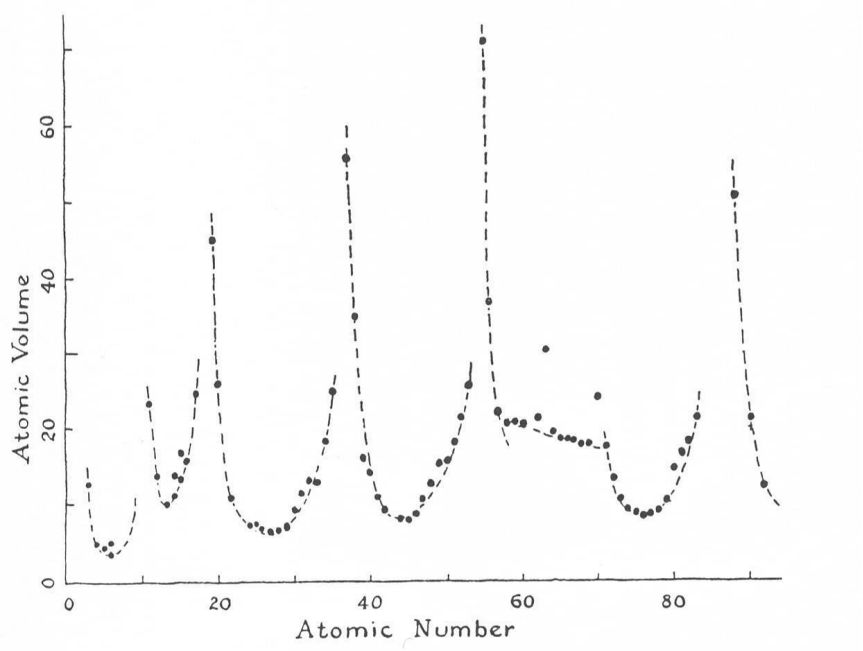


Linus Pauling, General Chemistry, p. 64, 1947

Edit and redesign (cont.)

Data-ink ratio improves to 0.9

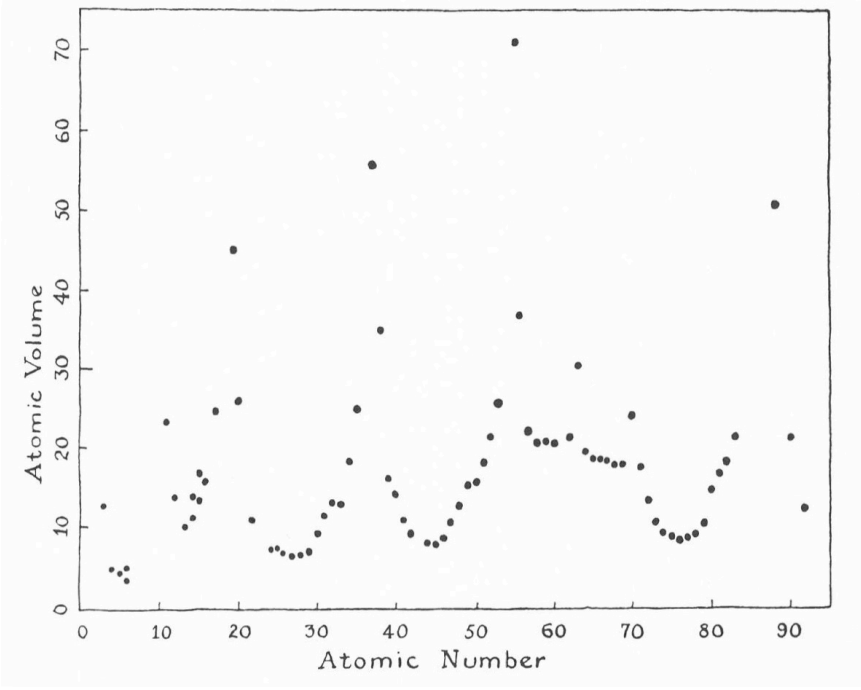
- only the frames line are uninformative
- erasing the grid marks highlights that several of the elements do not fit the smooth theoretical curve so well



The reference curve is essential in organizing the data, and shows the periodicity (the message) by creating a structure, and by giving ordering and hierarchy

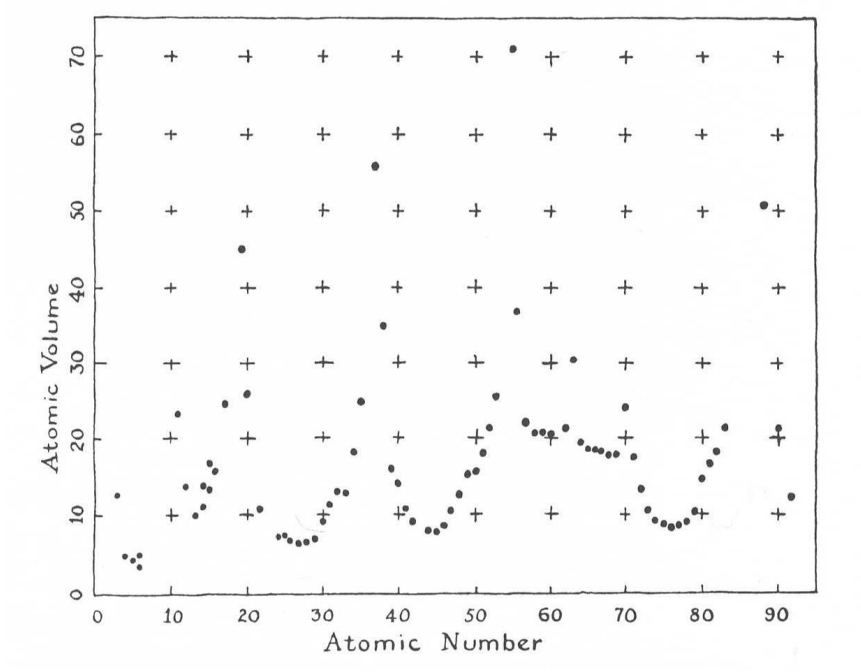
Edit and redesign (cont.)

BAD SOLUTIONS!



Without the curve we hardly detect the periodicity and the message

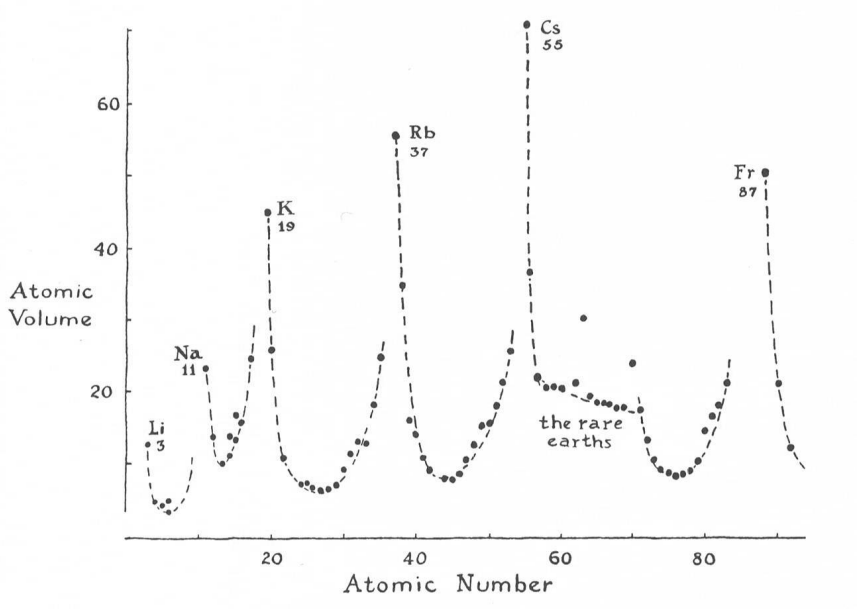
- the curve becomes necessary because the eye needs guidance



Restoring the grid totally fails to organize the data

- the grid marks are too powerful and induce visual vibration

Edit and redesign (cont.)



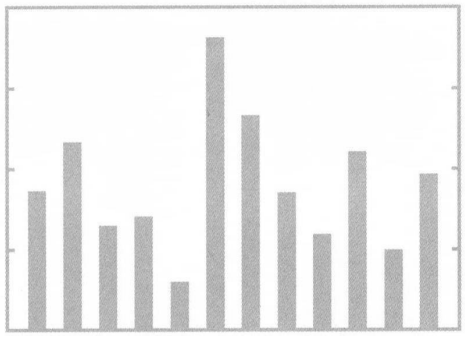
We can use the erased space

- labels for the initial elements of each period
- unusual rare-earths
- also, turned label and numbers on the vertical axis

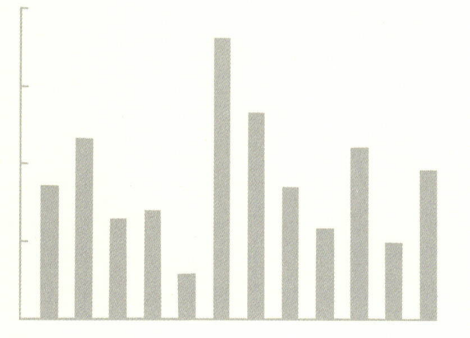
Take-home message:

Don't be happy with the initial version of your graphic!

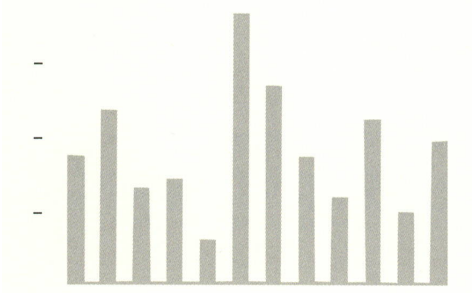
Redesign of the bar chart



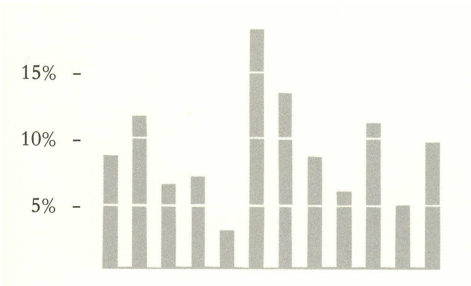
A standard model bar chart, with the design endorsed by the practices and the style sheets of many publications



erase the bounding box, for starters



and the vertical axes, keep the ticks



make a white grid, plus numerical labels

Chartjunk

Decoration of graphics requires a lot of ink and doesn't tell anything new.
So, why is it there, and what is the purpose of decoration?

- to make the graphic appear more scientific or precise
- to enliven the display
- to give the designer a chance to exercise artistic skills

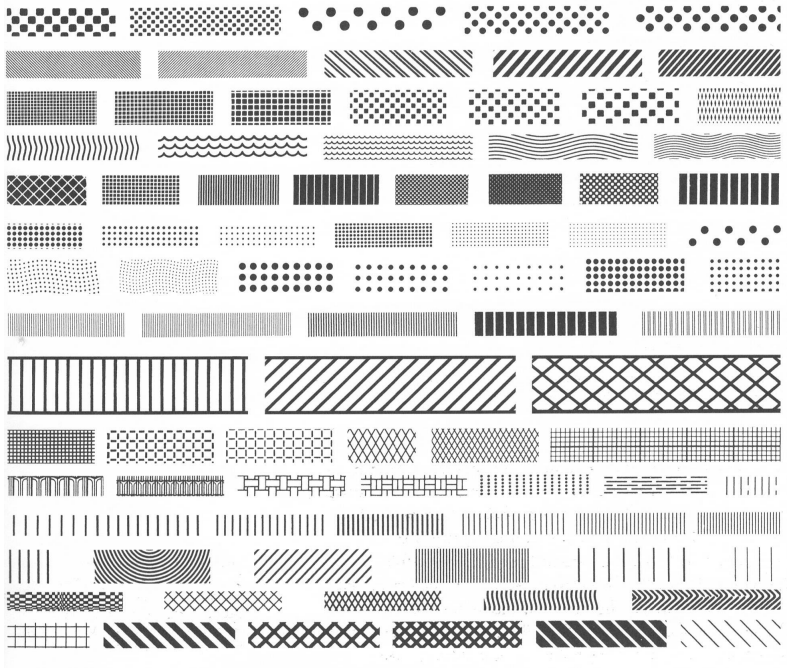
Chartjunk:

- visual elements not necessary to comprehend the information presented on the graph, or
- that distract the viewer from this information
- somewhat abstract concept (whereas non data-ink is very specific)

Unintentional optical art

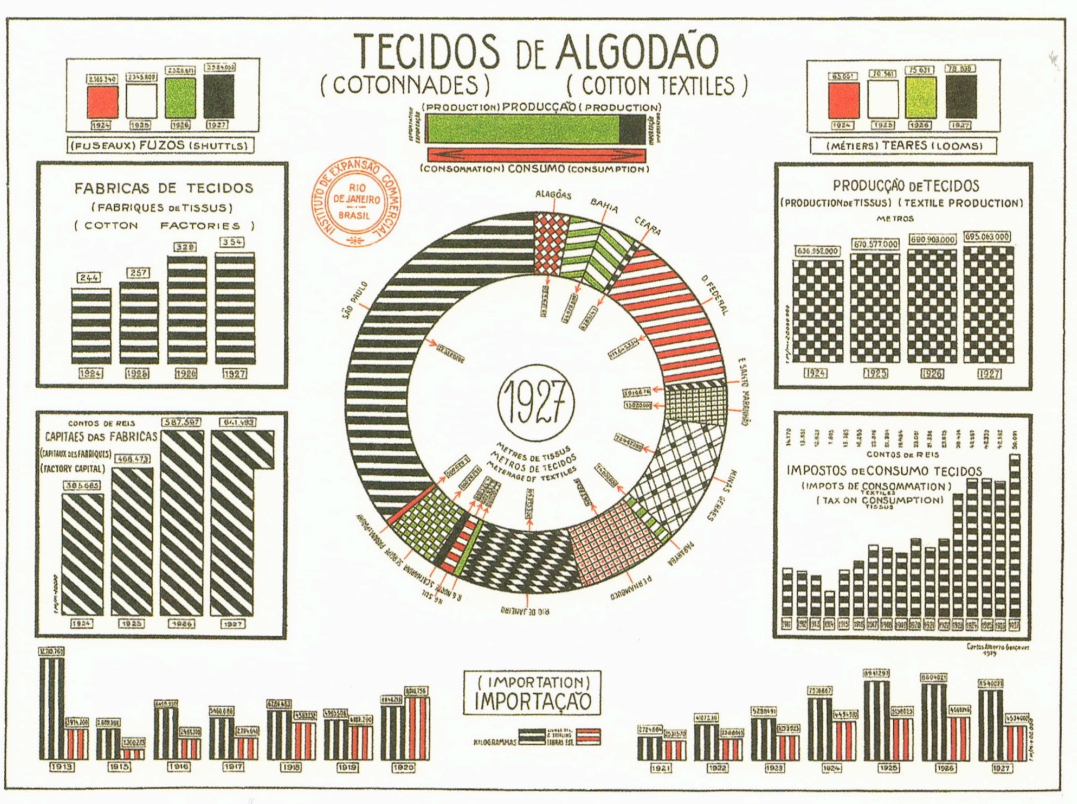
Happens in plots when areas are filled with patterns

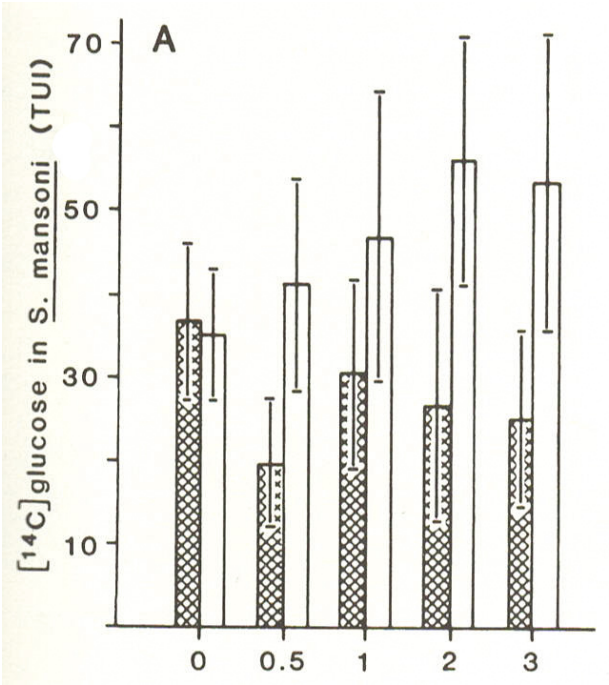
- draws attention away from the data
- clutters the plot
- induces shimmering



Unintentional optical art (cont.)

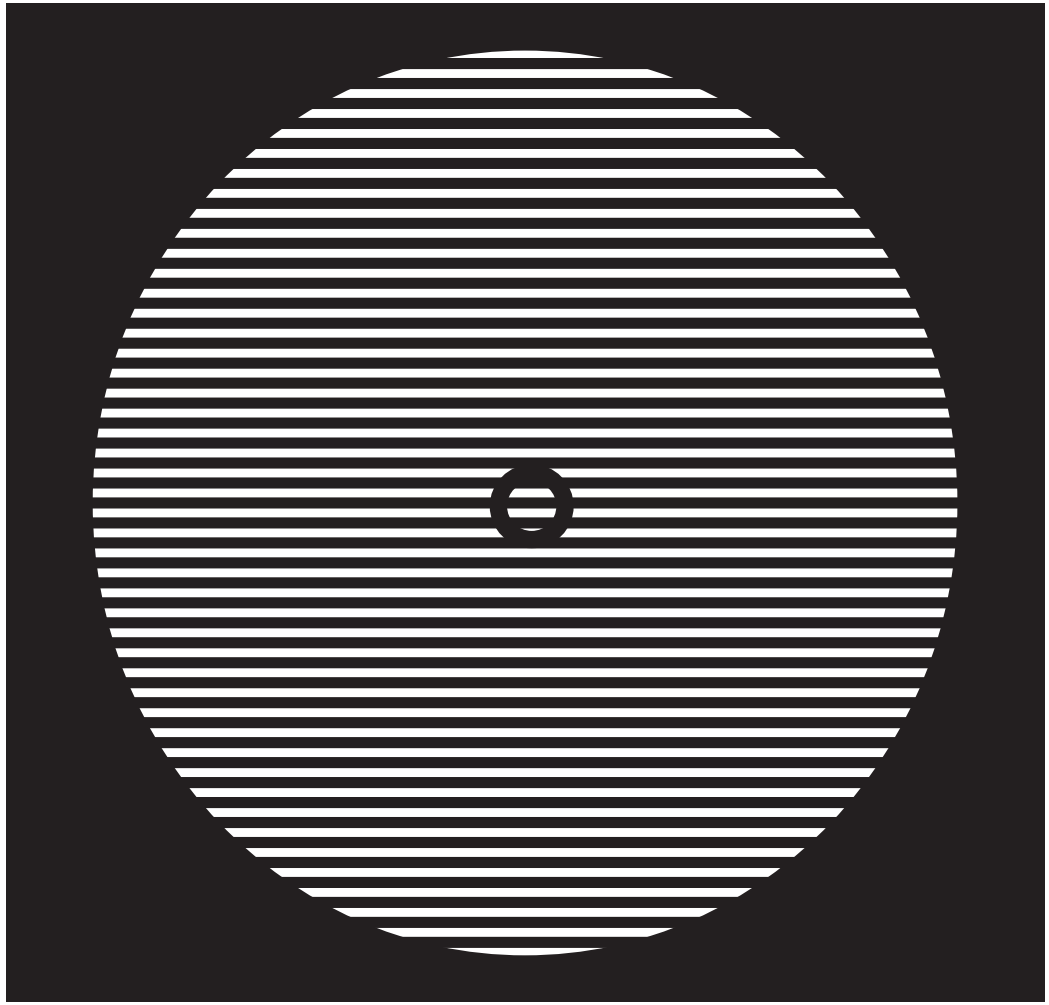
BAD EXAMPLES





bad data graphics

Patterns can cause seizures!



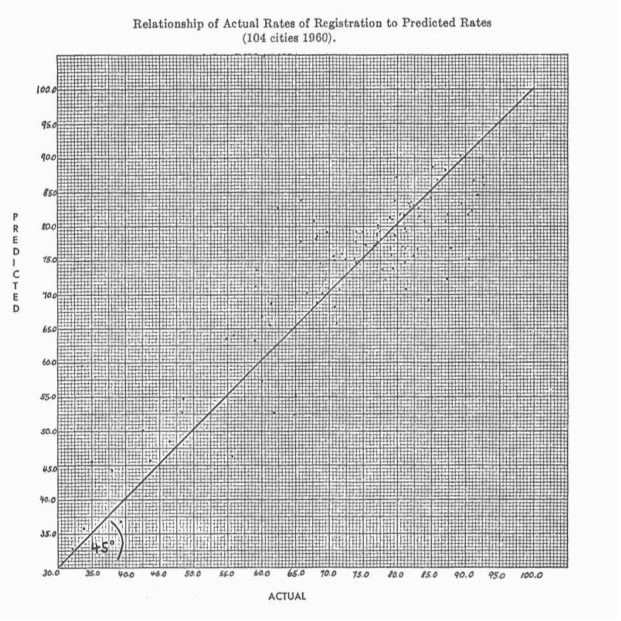
Grids

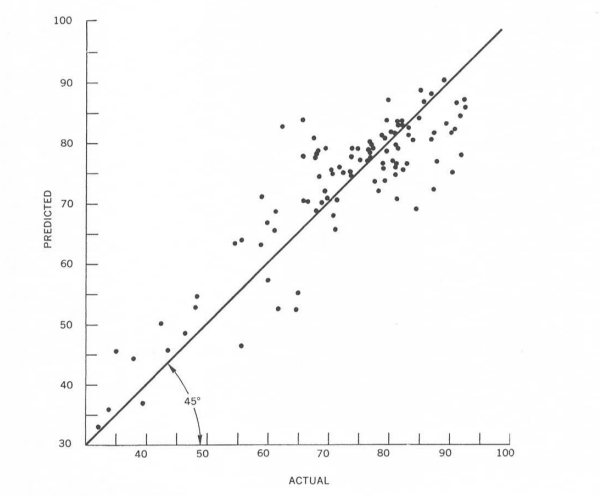
Grids are mostly for the initial plotting of data

- the grid should usually be muted or completely suppressed, so its presence is only implicit
- dark grid lines are chartjunk: they
  - do not carry any real information,
  - clutter up the graphics,
  - and generate perception not related to data information

Extremely active grid

Most of the ink devoted to matters other than data (1967) ⟶





Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

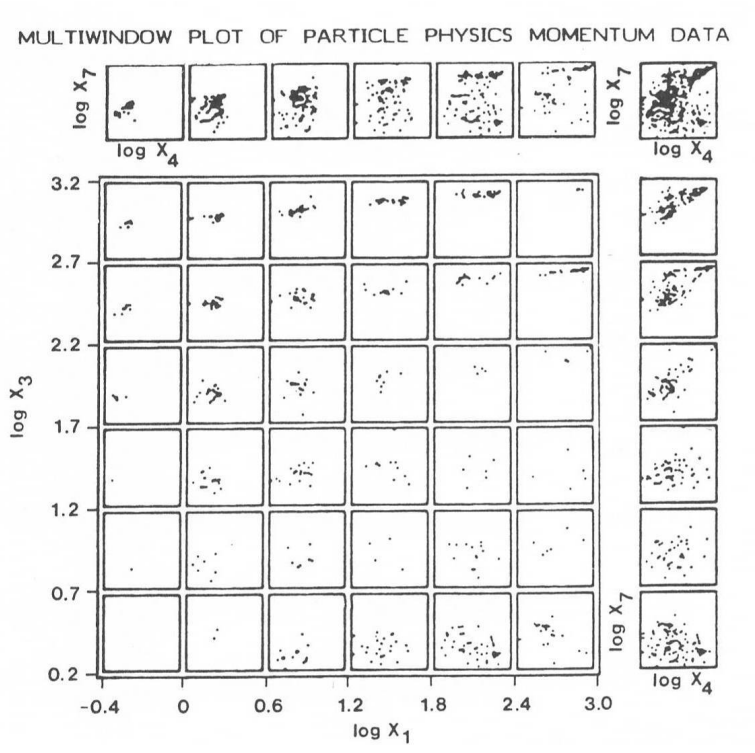
⟵ Improvements in a republished version of the same data (1970)

Double grid example

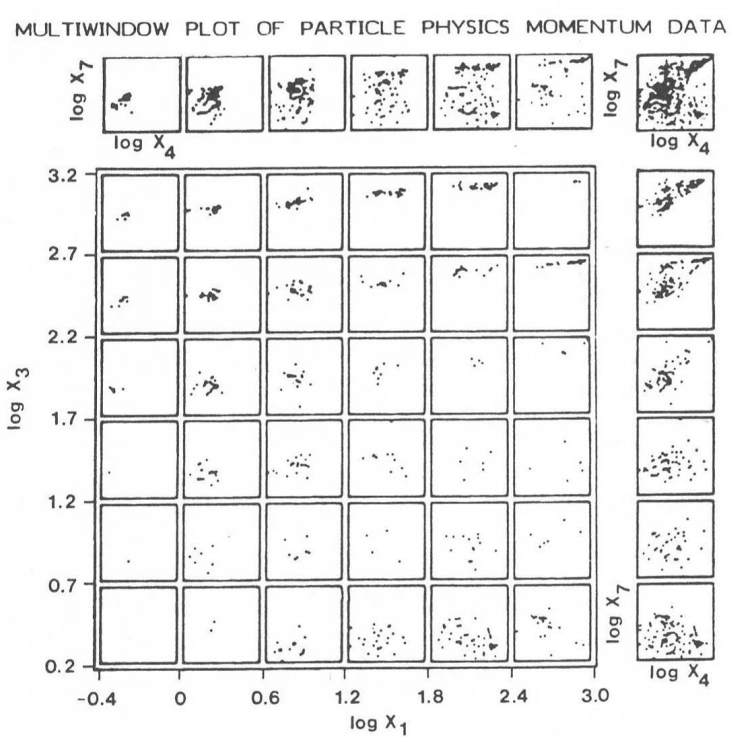
Double grid dominates the graphic and consumes 18% of the area

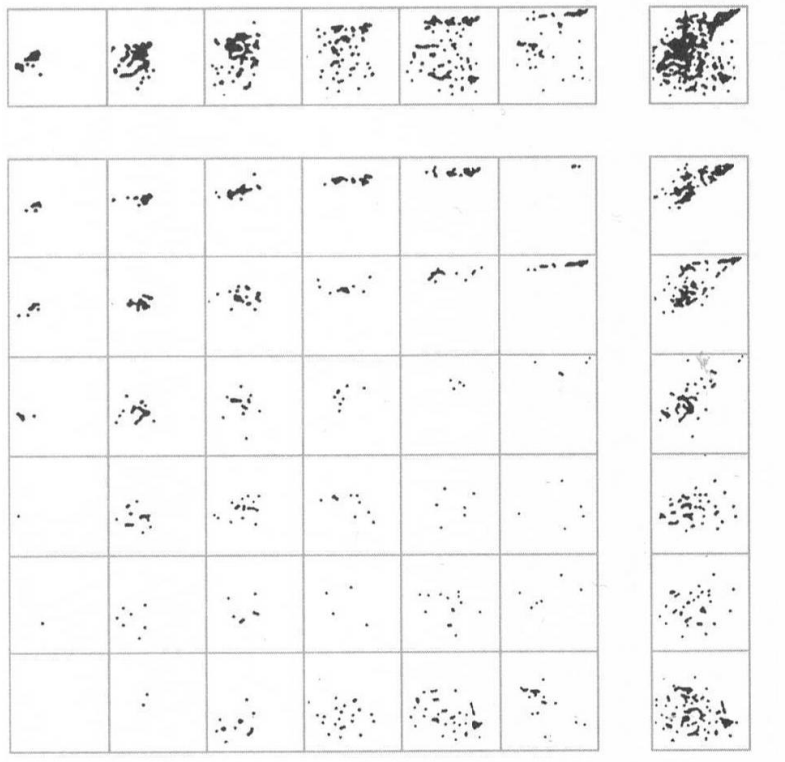
Optical white dots appear at the intersections of the grid lines

- Hermann grid illusion



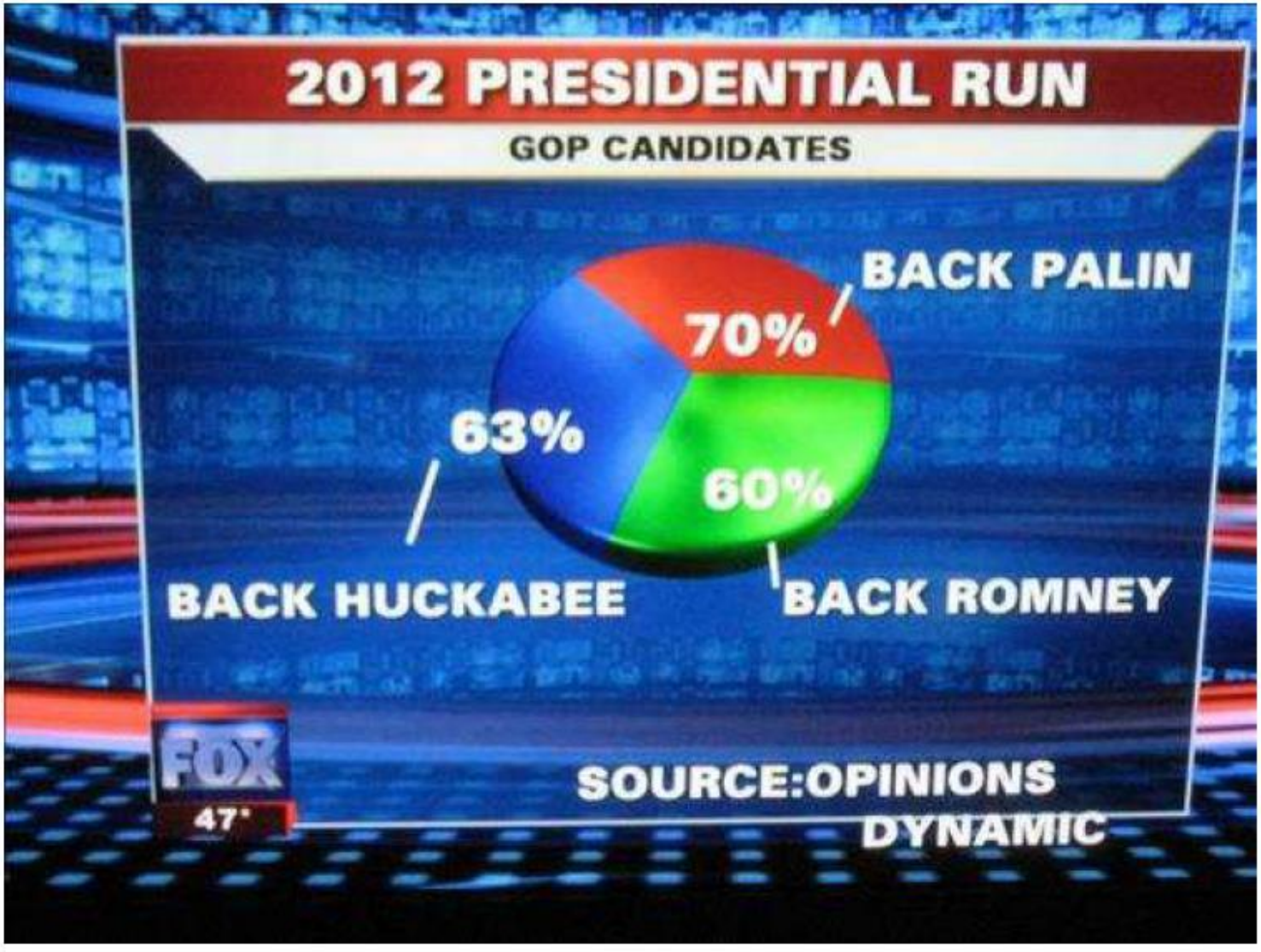
Double grid example, improved





Removing the double grid improves the graphic considerably

Really???


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS
DYNAMIC

Guidelines

a graphic can help to summarize a large table with a lot of numbers

(1) Choose properly format and design:

- table or figure or just plain text
- often combinations of these
- a simple table is often enough, don't force graphics
- choose the proper graphics format

(2) Often have a narrative quality, a story to tell about the data

(3) Use words, numbers and drawing together

- data graphics are paragraphs about data and should be treated as such

Guidelines

(4) Make your figure viewer-friendly:

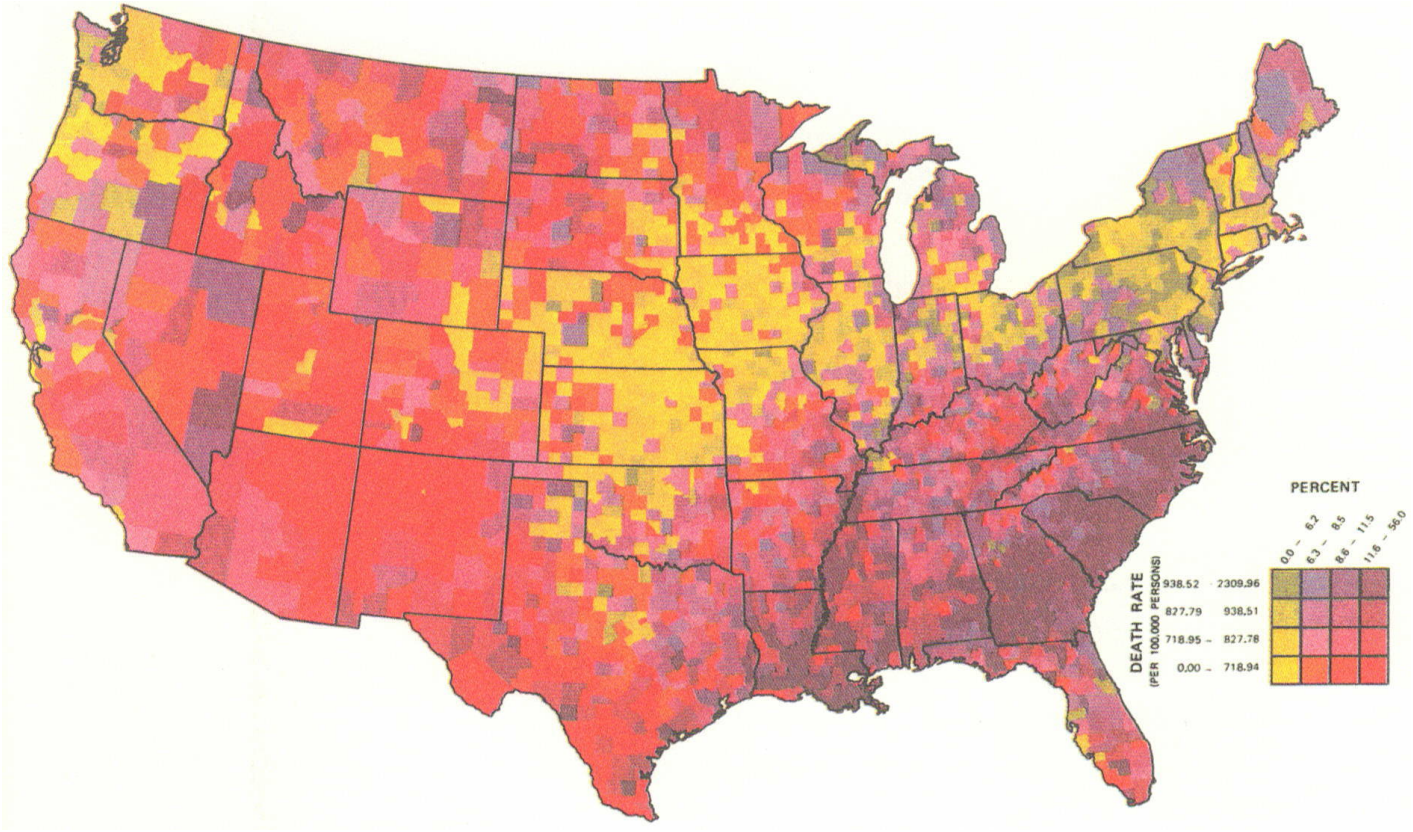
- explain what you are seeing (in caption)
- if needed, explain how you should read the plot
- spell out words, don't abbreviate
- if needed, add helping messages to the plot
- if possible, make words run from left to right: y-axis label.

(5) Colors:

- don't use patterns for filling areas
- if picture is going to be printed b&w, make sure it looks good
- don't use primary colors:
  - they look bad when printed b&w
  - problematic for color-blind
  - there are better color schemes
- if you use color for encoding, consider instead labelling surfaces

Guidelines

(6) Beware of creating puzzles, unnecessary complexity



'Now let's see, purple represent counties where there are both high levels of male cardiovascular disease mortality *and* 11.6–56.0 percent of the households have more than 1.01 persons per room. Uh!'

Guidelines

(7) Figures can be complex

- if your data and story requires it
- explain the figure in caption

(8) Figures can be very small

- the overall trend is still readable
- the fonts need to be the correct size
- controlling gaps becomes important

(9) Draw in a professional manner: make sure

- your font size is correct,
- axis ranges are proper,
- legend doesn't obstruct the lines,
- minimize clutter, rethink the format, if necessary, and
- avoid content-free decoration

Lessons from
Information Visualization: perception for design

by Colin Ware

Human vision is weird

Is the dress

- white and gold, or
- black and blue?

Semiotics

- study of symbols and how they convey meaning
- dominated mostly by philosophical arguments instead formal experiments

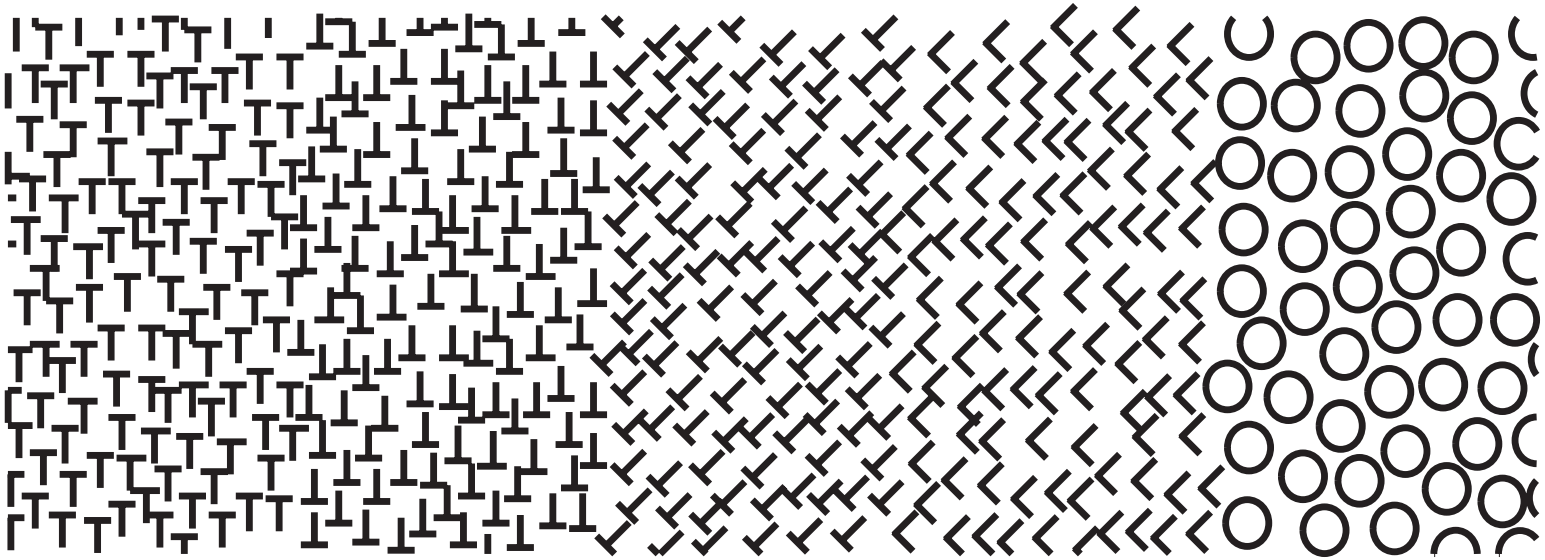
Arbitrary symbols

- depends on the culture
- needs to be learned
- easy to to forget

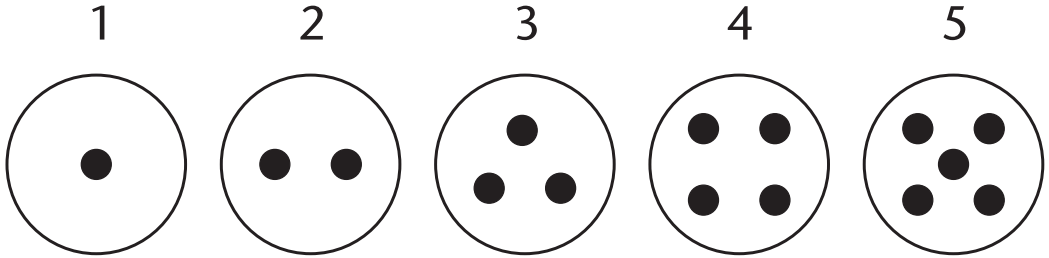
Sensory symbols

- works across cultures
- understanding without training
- resists to instruction bias

Semiotics



5 regions of texture: some are easier to separate from the other



arbitrary (arabic numbers) and sensory (count)

Gibson affordance theory

Gibson (1979) proposed that we perceive in order to operate on the environment.

- surfaces are perceived for walking
- handles are perceived for turning
- buttons are perceived for pressing

Very different than previous approaches: bottom-up approaches based on how light hits the eye

Problematic if taken literally

- the information that is shown can be very abstract
- we must learn that pressing buttons leads to action
- glosses over visual mechanisms

Anatomy of an eye

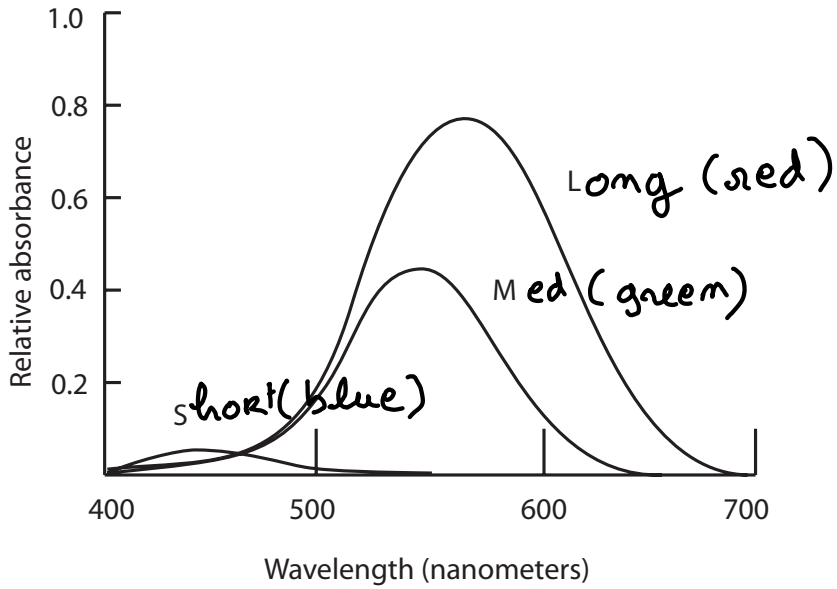
- In eye, the lens focuses an image to retina
- Retina has two types of receptor cells: cones and rods
- cones
  - responsible for normal color vision
  - 6–7 millions
  - concentrated around fovea
- rods
  - responsible for dim-light vision
  - no color information
  - roughly 90 millions
  - concentrated on the edges of the eye

Anatomy of an eye

- fovea = small area in retina, densely packed with cones. Sharpest vision is here.
- blind spot = area in retina without rods, connection to optic nerve
- Field of view (for one eye)
  - 60° up
  - 70°–75° down
  - 60° nasal
  - 100°–110° temporal

Trichromacy

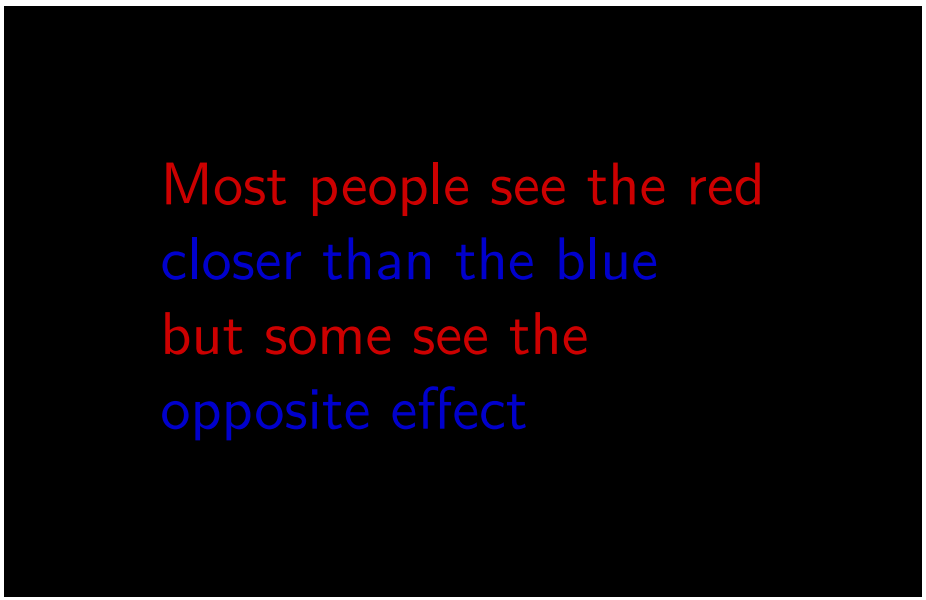
- Human eye have 3 distinct color receptors, cones
- (we also have rods, but they work only in dim conditions)
- Hence, we tend to express colors using 3 primary colors
- Chickens have 12 distinct color receptors



Cone sensitivity as a function of wavelength

Chromatic Aberration

- Human eye focus depends on the wavelength:
- 60% see red closer, 30% see blue closer, 10% no difference
- don't use pure blue with black background, esp. if red is present: mixing blue with red or green helps



Color blindness

Occurs in 10% of males, and 1% of females Most common deficiencies are

- protanopia = lack of long wavelength-sensitive cones
- deuteranopia = lack of medium wavelength-sensitive cones
- both result in inability to distinguish red from green

Opponent process theory

Theory by Ewald Hering: there are 6 primary colors and they are arranged pairs

- black-white
- red-green
- blue-yellow

The colors are 'opponents' to each other

Evidence supporting Opponent process theory

Naming

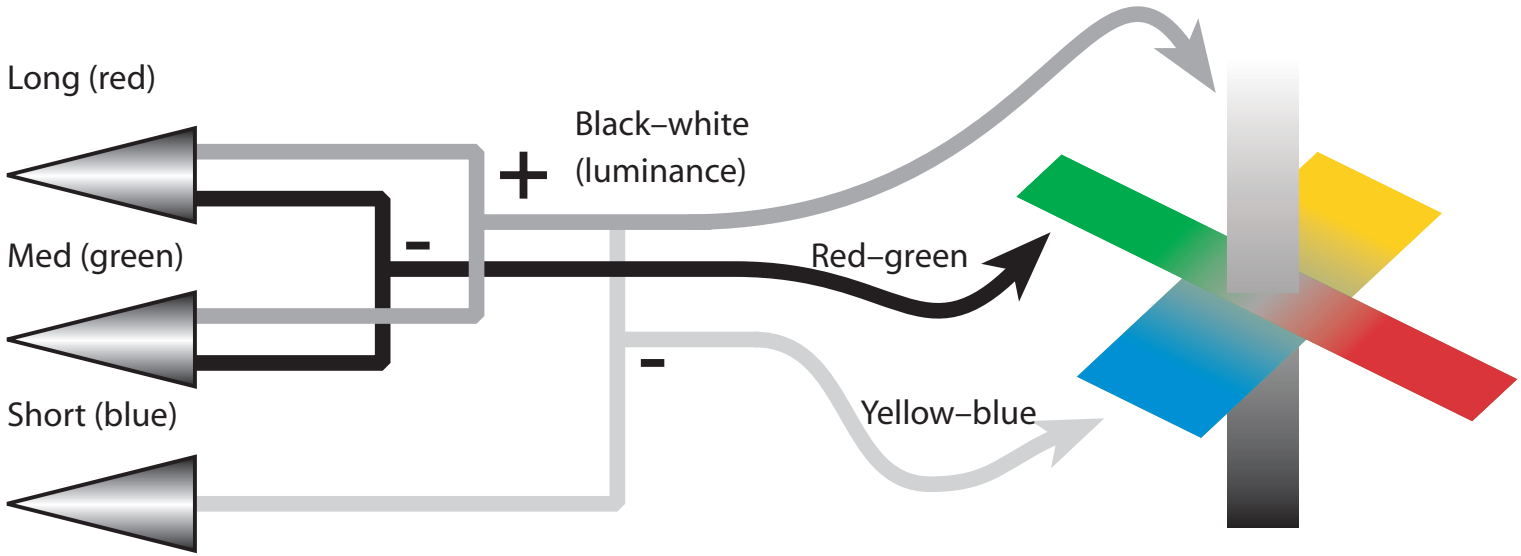
- it is OK to say 'yellowish green' or 'reddish blue'
- nobody says 'reddish green' or 'yellowish blue'

Cross-cultural naming

- A study of over 100 languages by Berlin and Kay 1969.
- all languages has words for black and white
- if language has a third color, then it's red
- then it's followed by yellow and green or by green and yellow
- then blue

Evidence supporting Opponent process theory

Inputs from cones are combined to channels on biological level



Color in visualization

- Always use vary luminance, not just color, between background and foreground
- Use only a few colors if they represent distinct codes: 6 is easy, 10 must be selected carefully
- Black or white borders around colorful objects help to separate from the background
- If you are color-coding large areas, use muted colors
- Small color objects should have high-saturation colors

Color in visualization

If color sequence is needed

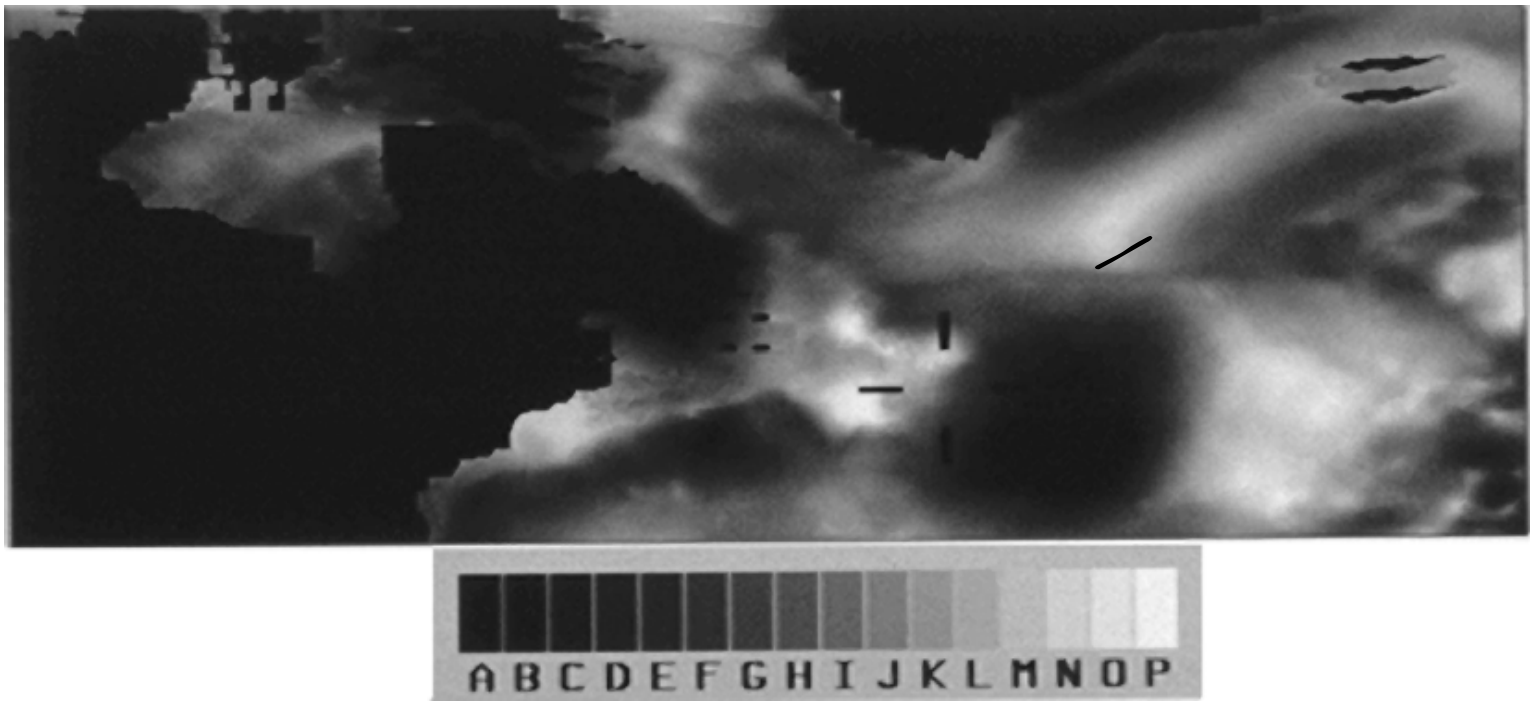
- use a sequence that varies monotonically on at least one of the opponent color channels: red-to-green, blue-to-yellow
- variation in multiple channels is often better: pale-yellow to dark-blue
- if 0 value is meaningful, use neutral color for 0, and saturate towards opposite colors to show negative and positive values: for example, red – gray – green

Gray scale as for coding data

Not a good idea, use colors instead

- local contrast induces errors
- the luminance channel is fundamental to perception
- it's a waste of its resources to use it for coding data

Local contrast illusion



Errors occur when reading the map

- local contrast between the legend labels
- local contrast between a map point and surrounding points

Acuity

Visual acuity is the measurement to see detail

- for example, separate the two lines

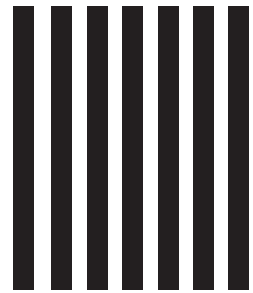
Expressed in visual angles

- degrees (360 full turn)
- arcminutes (1/60 of degree)
- arcseconds (1/60 of minute)

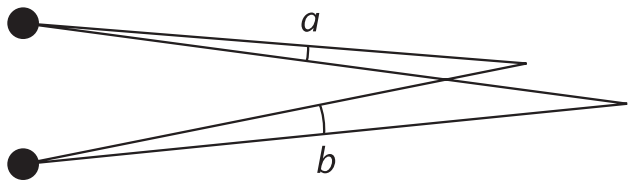
Normally, the best you get is 1 arcminute

- this corresponds to the density of cones in fovea
- but there are superacuities.

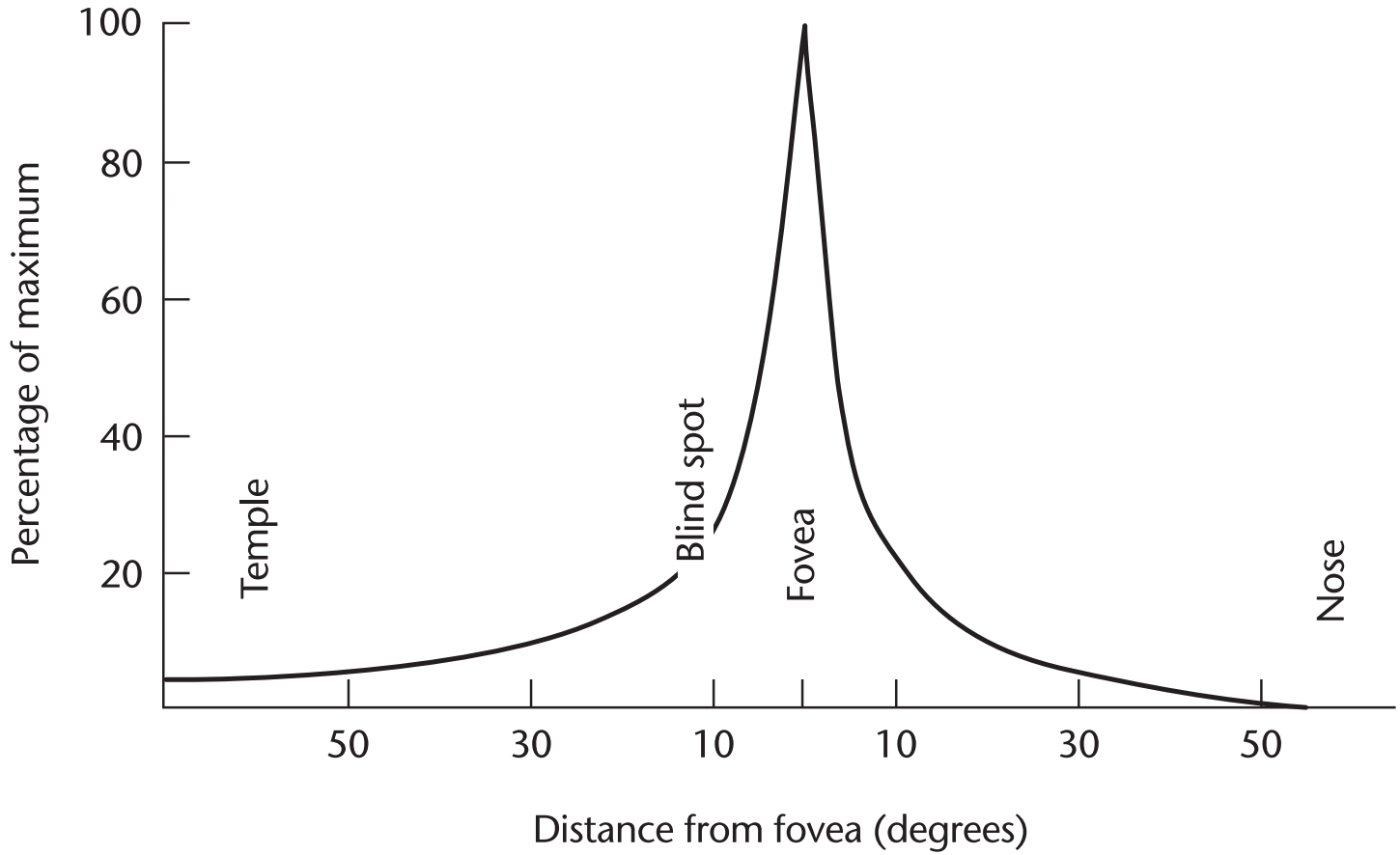
Acuities

Point acuity (1 minute of arc): The ability to resolve two distinct point targets.	
Grating acuity (1–2 minutes of arc): The ability to distinguish a pattern of bright and dark bars from a uniform gray patch.	
Letter acuity (5 minutes of arc): The ability to resolve letters. The Snellen eye chart is a standard way of measuring this ability. 20/20 vision means that a 5-minute letter target can be seen 90% of the time.	

Superacuities

Stereo acuity (10 seconds of arc): The ability to resolve objects in depth. The acuity is measured as the difference between two angles (*a* and *b*) for a just-detectable depth difference.	
Vernier acuity (10 seconds of arc): The ability to see if two line segments are collinear.	

Acuity: fovea vs. edge



Preattentive processing

How many 3s are there?

85689726984689762689764358922659865986554897689269898
02462996874026557627986789045679232769285460986772098
90834579802790759047098279085790847729087590827908754
98709856749068975786259845690243790472190790709811450
85689726984689762689764458922659865986554897689269898

To find them, you have to do a linear scan.

Preattentive processing

How many 3s are there?

85689726984689762689764358922659865986554897689269898
02462996874026557627986789045679232769285460986772098
90834579802790759047098279085790847729087590827908754
98709856749068975786259845690243790472190790709811450
85689726984689762689764458922659865986554897689269898

Now you only need to do a linear scan over the red ones.

Preattentive processing

- certain shapes or color pop out from their surroundings.
- mechanism underlying pop-out is called preattentive processing
- because it must occur before conscious attention.
- preattentive processing determines what visual objects gets attention

- used for designing symbols when displaying information
- certain symbols need more attention

Examples of preattentive features

Form

- line orientation
- line length
- line width
- collinearity
- size
- curvature
- spatial grouping
- blur
- added marks
- numerosity

Color

- hue
- intensity

Motion

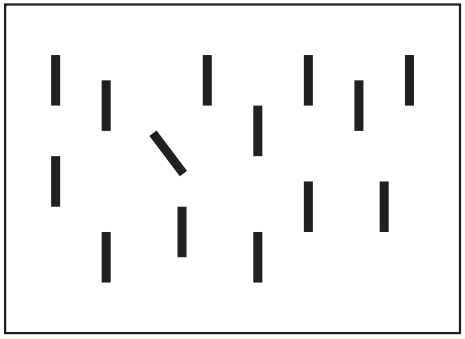
- flicker
- direction of motion

Spatial position

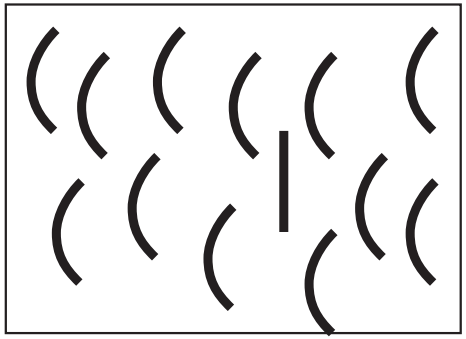
- 2D position
- Stereoscopic depth
- convex/concave shade

Examples of preattentive features

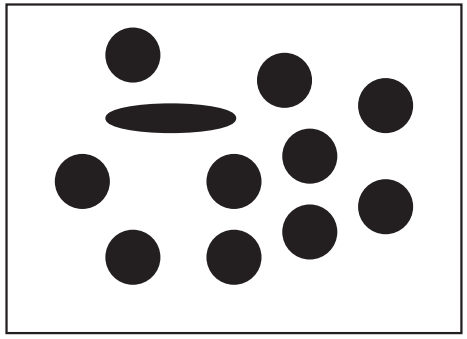
Orientation



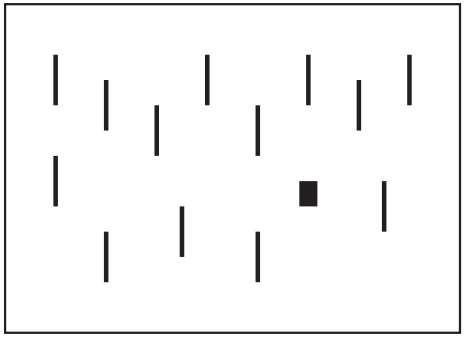
Curved/straight



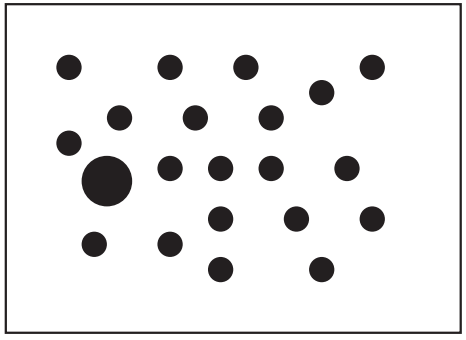
Shape



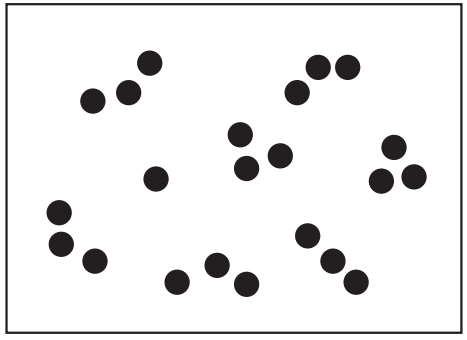
Shape



Size

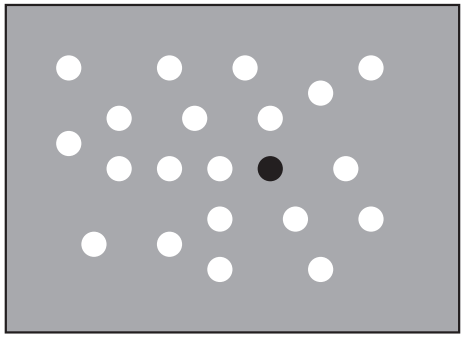


Number

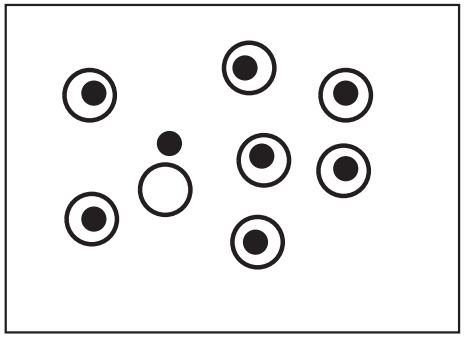


Examples of preattentive features

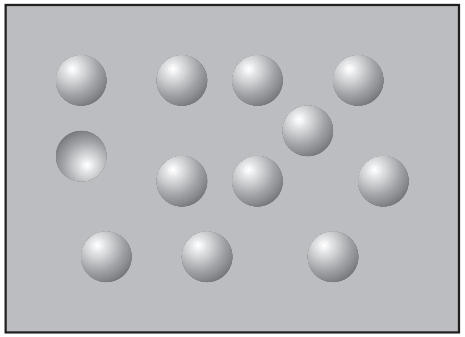
Gray/value



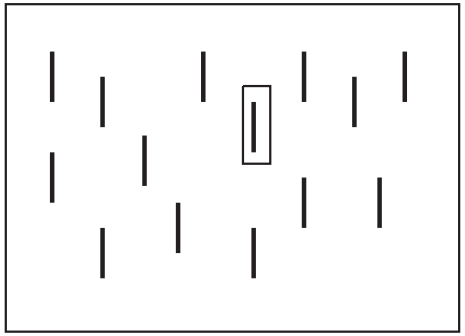
Enclosure



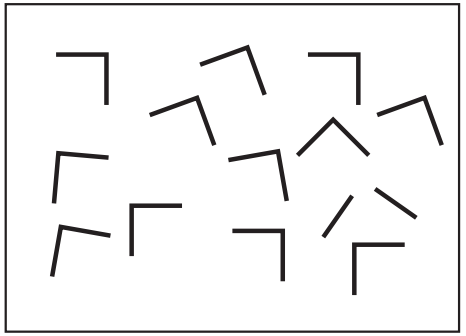
Convexity/concavity



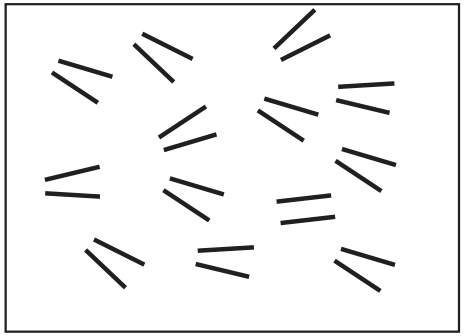
Addition



Juncture



Parallelism



The last two are not preattentive features

Preattentive processing

Are some features better than others?

Unfortunately, this depends heavily on surroundings.

Callaghan (1989) compared colors to orientation:
results depended on the

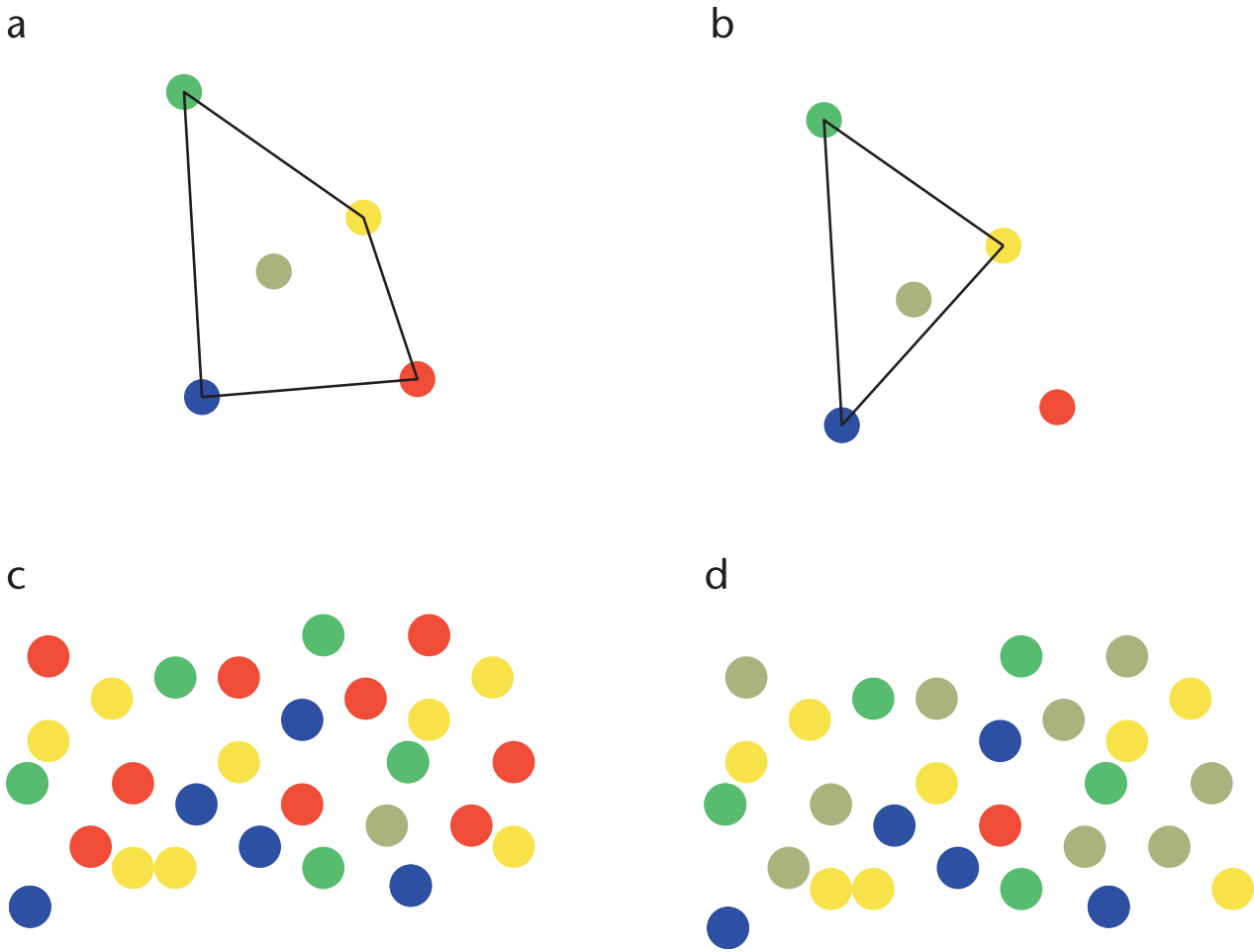
- the saturation
- size of the color patch
- difference from the surrounding colors
- length of the line
- difference degree
- contrast of the line pattern

Preattentive processing

Some generalizations are possible though

- adding marks to highlight something is better than taking them away
- we can see at glance that there are 1–4 objects in a group, more requires counting
- using color as a preattentive feature is well-established
- color should be outside the region defined by other neighboring colors (next slide)

Color as preattentive feature

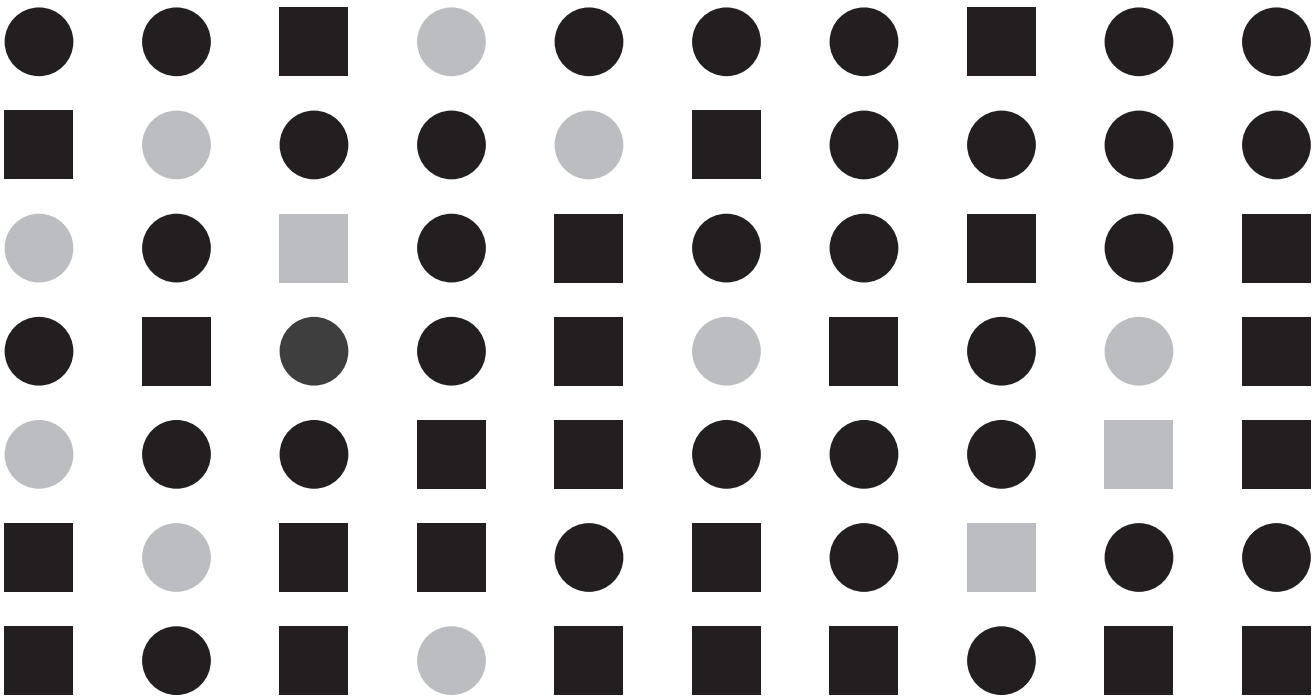


Grey is more difficult to locate in (c) than red in (d)

Conjunctive preattentive features

Is it possible to use preattentive features for complex queries?

Generally speaking, no



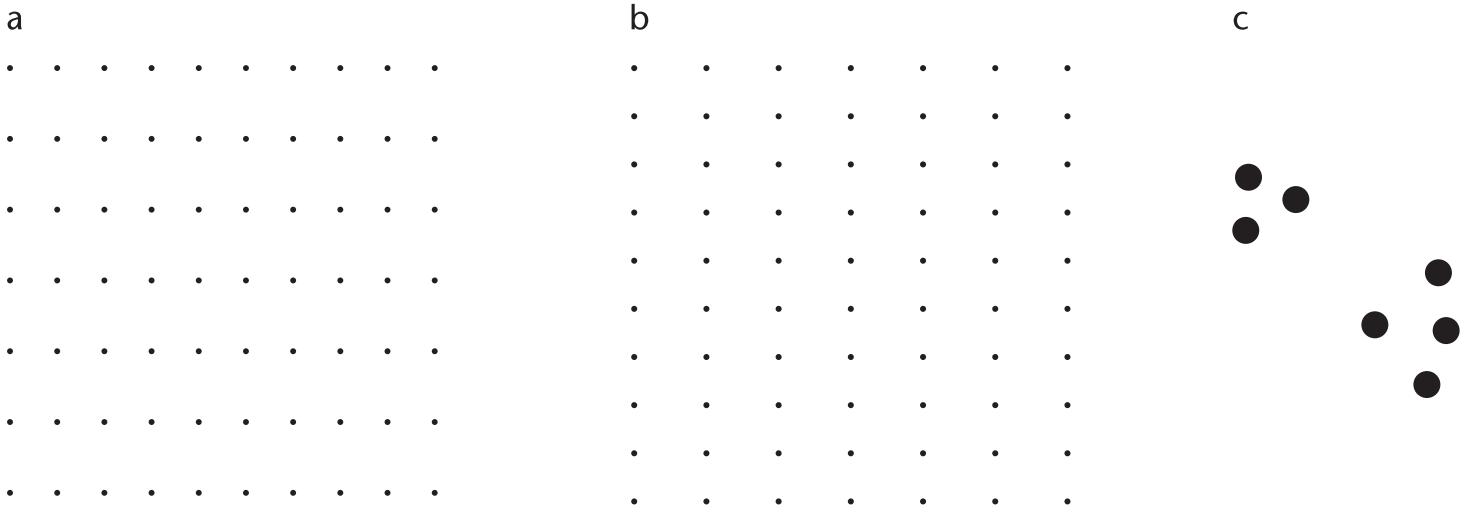
Finding grey squares is slow

Gestalt laws

- established by group of German psychologists in 1912
- rules how we see patterns in visual displays
- Gestalt = pattern in German

Gestalt law of proximity

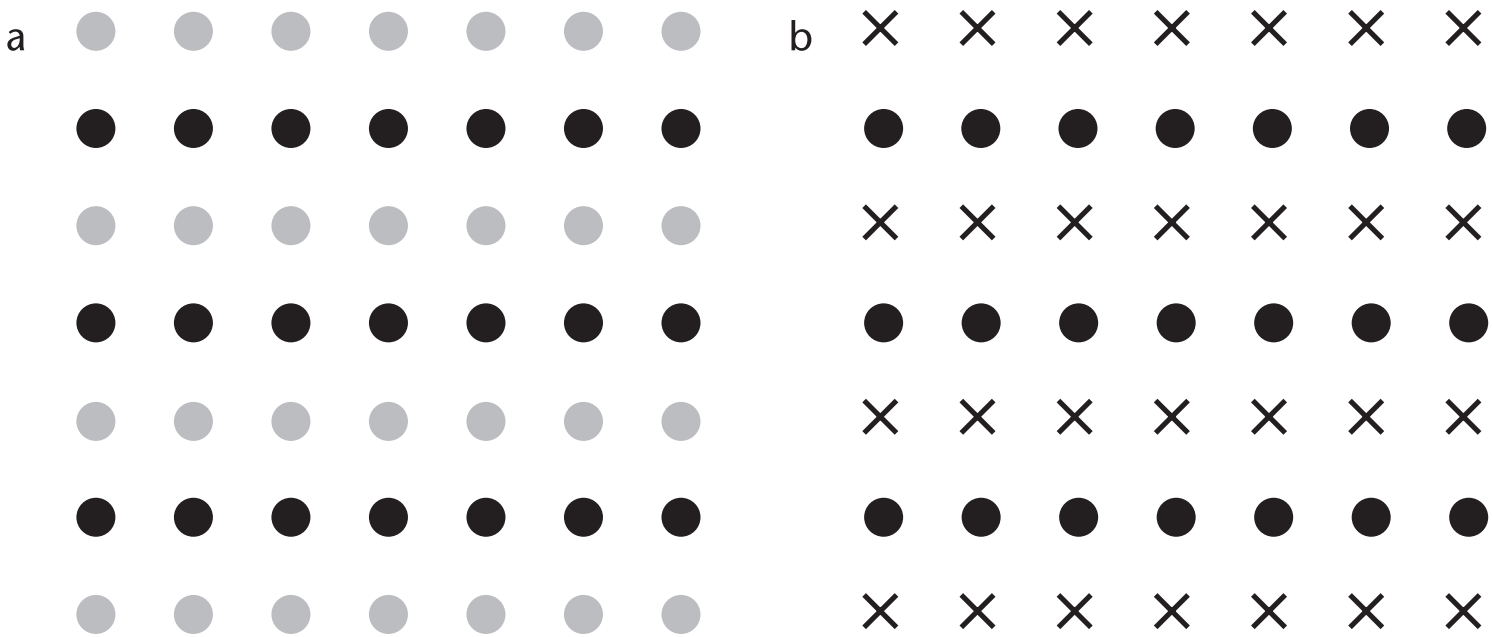
Things that are close are grouped together



- (left) rows are grouped
- (center) columns are grouped
- (right) two clusters

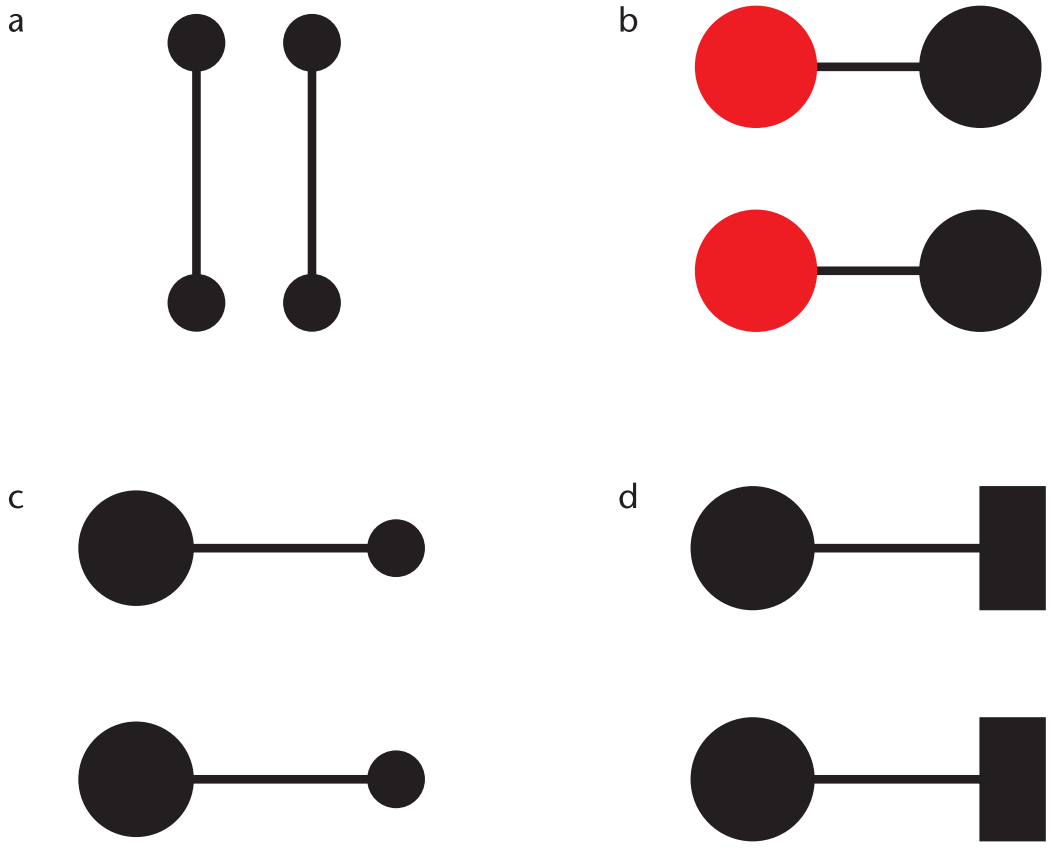
Gestalt law of similarity

Similar shapes are grouped together



Gestalt law of connectedness

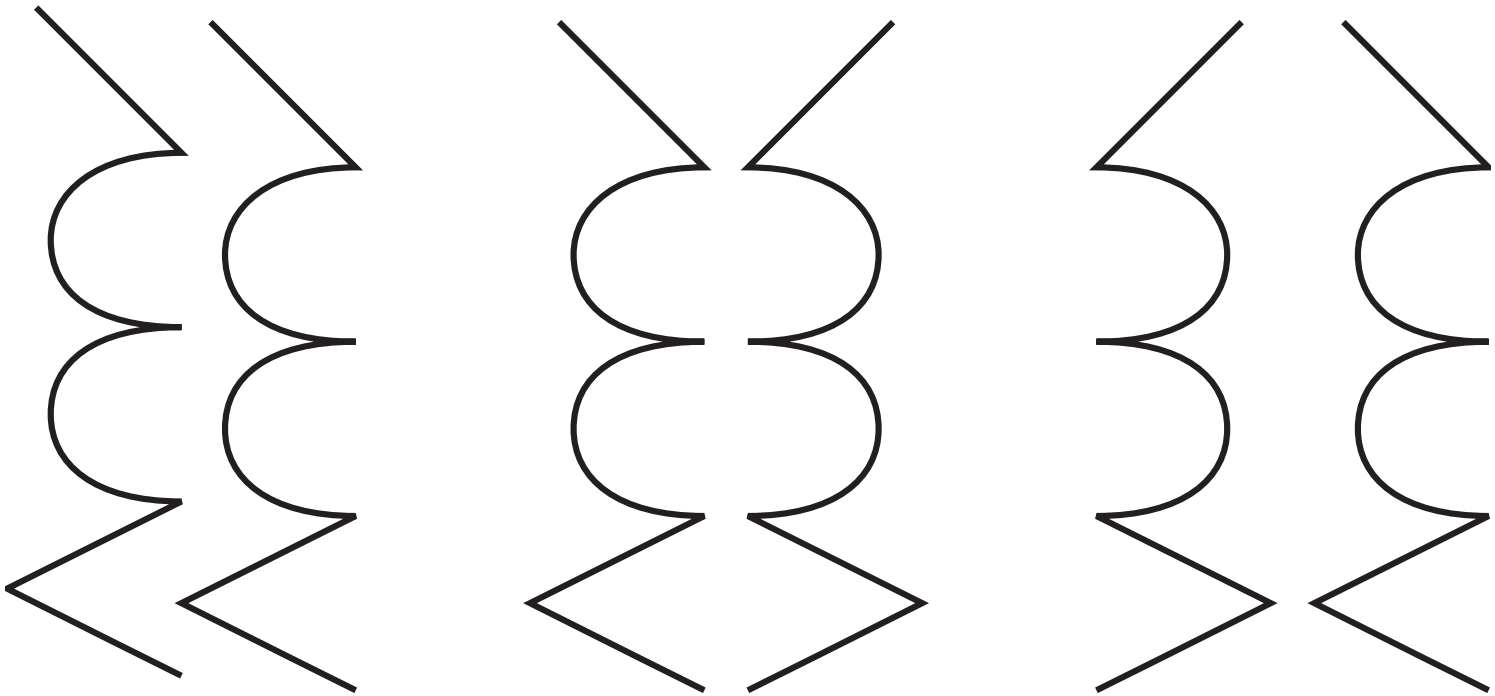
Connected shapes are grouped together



Connectedness is stronger than proximity or similarity

Gestalt law of symmetry

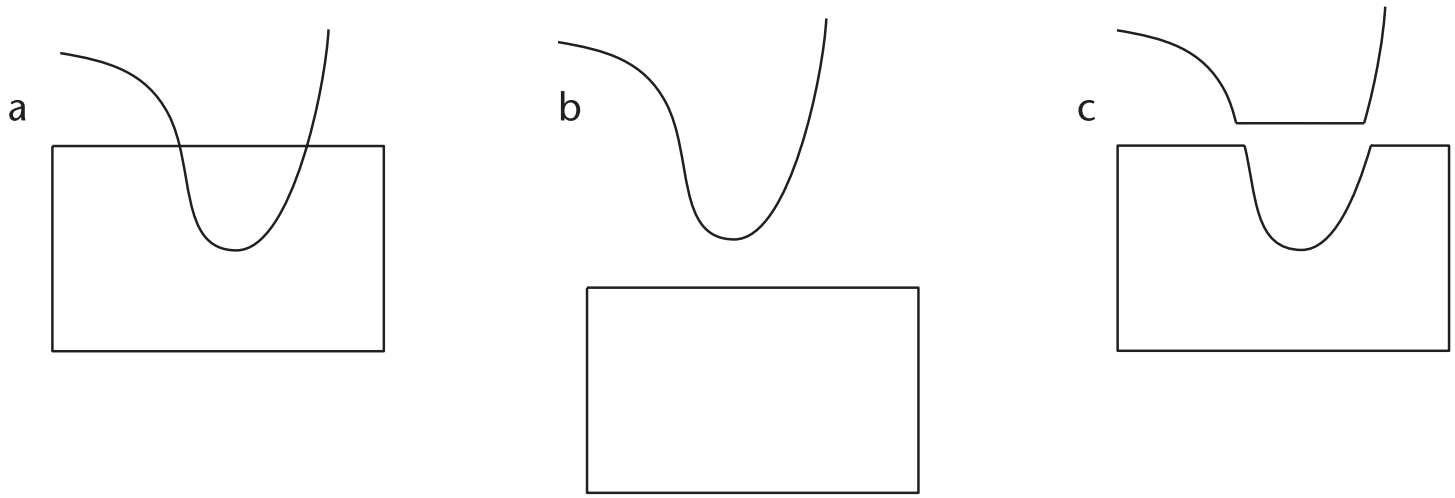
Symmetric shapes are interpret as a whole



left show two separate figures, center and right are whole

Gestalt law of continuity

We construct visual entities from smooth and continuous shapes

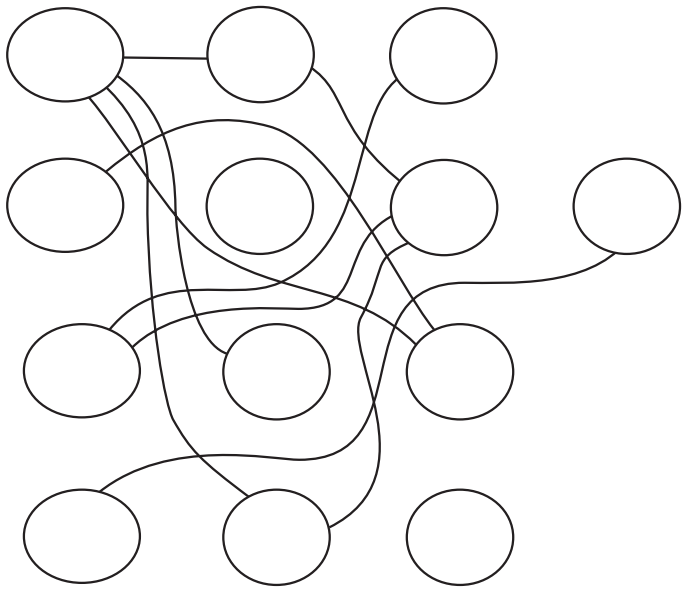


We interpret (a) as (b) rather than (c)

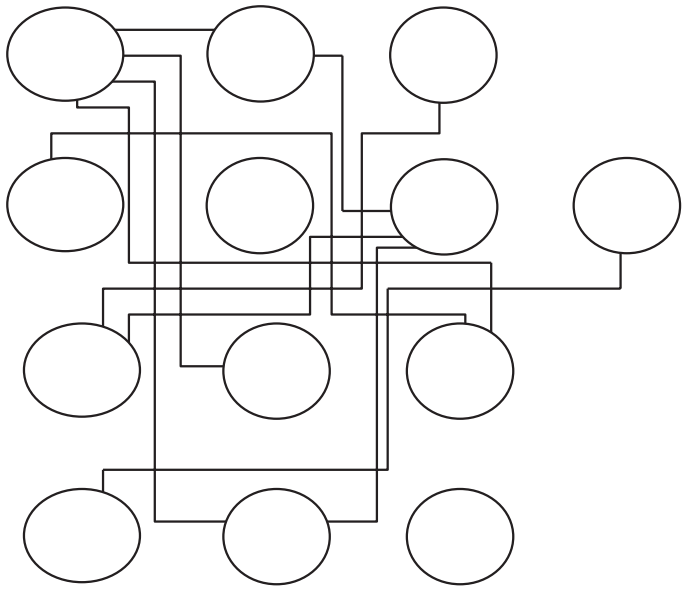
Gestalt law of continuity

Left is easier to read than right

a

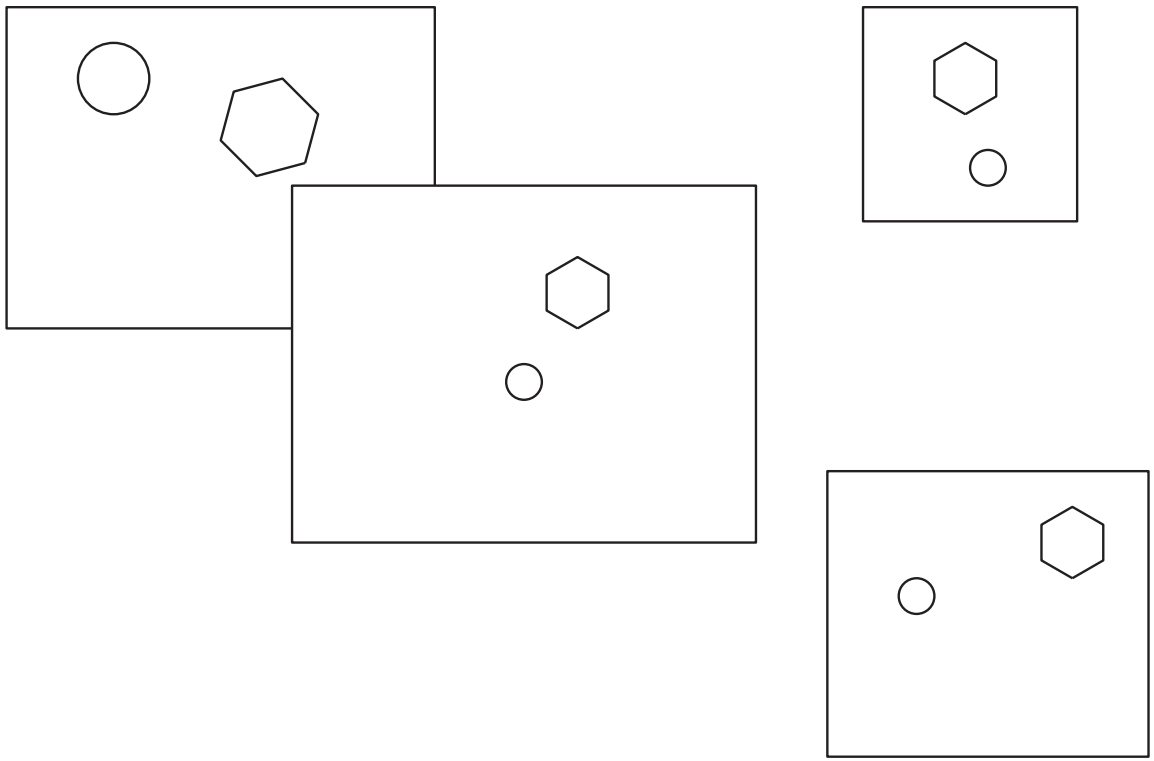


b



Gestalt law of closure

A closed contour tends to be seen as an object



closure is stronger organizing principle than proximity

Gestalt laws

Gestalt laws is a background for many visualization principles.

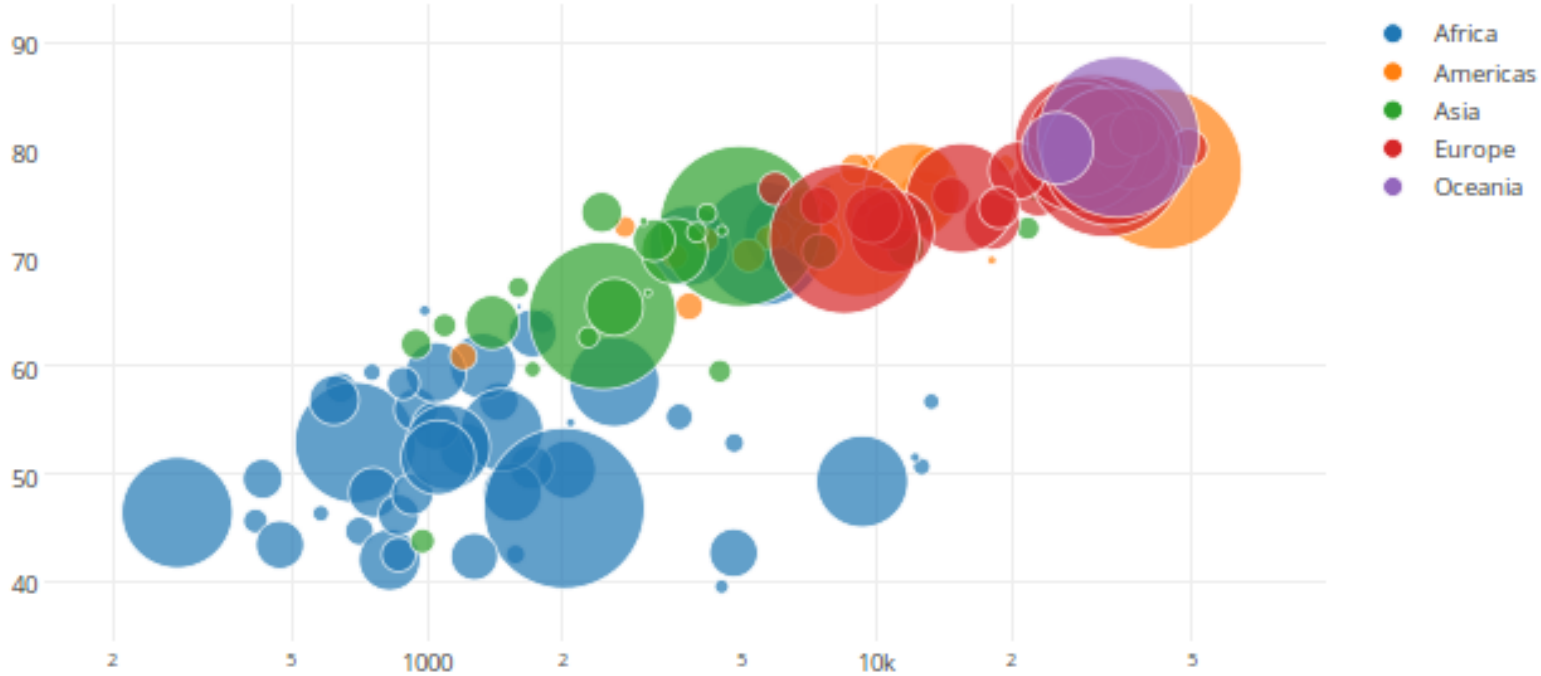
- similar elements should be placed closely
- similar elements should have similar shapes
- connectedness law leads to node-link diagrams
- connectedness law leads also to connecting dots in a plot
- continuity law leads to have smooth links
- closure allows to create groups

Nevertheless, they are commonly violated

- elements that are related are placed far away
- closure is used inconsistently

Glyphs

If the data has ≥ 2 dimensions, we can use glyphs to visualize it.



4 dimensions: coordinates, size, color

Popular visual variables

Spatial position of glyph (2–3 dimensions)

Color of glyph (3 dimensions, color opponent theory)

- luminance is needed for other graphical variables

Shape (2–3 dimensions, open problem)

- exact number of dimensions that can be processed fast is unknown
- some evidence that size and elongation are two primary ones

Popular visual variables

Orientation (1–3 dimensions)

- not independent of shape due to symmetry

Surface text (3 dimension: orientation, size, contrast)

- not independent of shape or symmetry
- uses one color dimension

Motion coding (2-3 dimensions, open problem)

Blink (1 dimension)

- high dependency on motion

Visual variables

- many of these channels are not independent
- if we are lucky, we can get 8 dimensions represented
- easy granularity for each dimension: 8 colors, 4 different orientations, 4 sizes
- ... about 2 bits per channel, on average
- leads to 2^{16} options
- however, conjunctions are not preattentive
- for preattentive processing $4 \times 8 = 32$ options

Lessons from Visualization analysis and design

by Tamara Munzner

Marks and channels

- mark is a basic graphical element in an image
- visual channel is a way to control the appearance of a mark

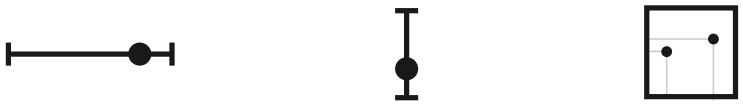
Examples of marks

Points **Lines** **Areas**

Examples of visual channels

Position

- Horizontal
- Vertical
- Both



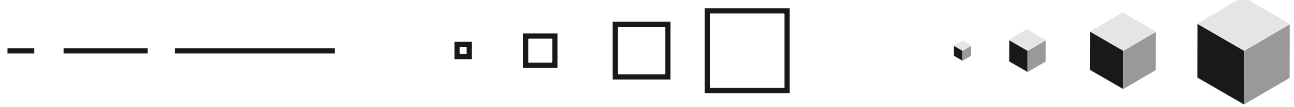
Color

Shape

Tilt

Size

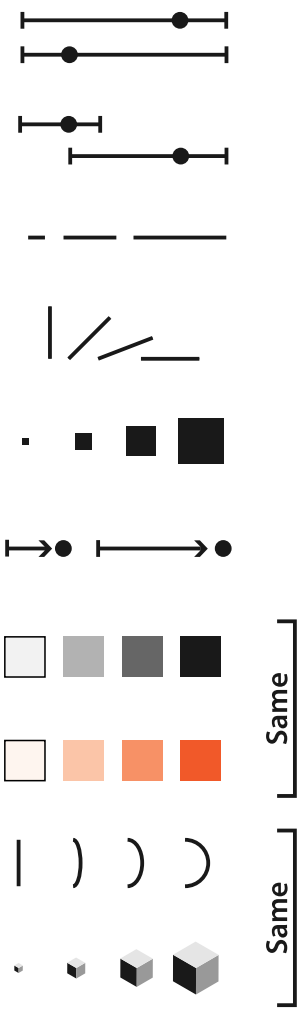
- Length
- Area
- Volume



Effectiveness of visual channels

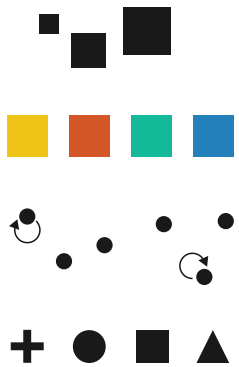
Magnitude Channels: Ordered Attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)



Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape



Eight rules of thumb

1. no unjustified 3D
2. no unjustified 2D
3. eyes beat memory
   - show related information simultaneously
4. function first, form next
   - visualize your data effectively, then make it pretty
5. get it right in black and white
   - literally, print your figure with b&w printer
   - use luminance
6. resolution over immersion
7. overview first, zoom and filter, details on demand
8. responsiveness is required

Framework for visualization proposed by Munzner

3 main layers

- What data is shown?
- Why is this task performed? (why is this plot shown)
  - is it to explore the data?
  - is it to present a result to somebody?
- How is the visualization constructed?
- we will focus on the how

Arrange

Express

- used for continuous data
- scatterplot, dot chart, line chart

Separate, order, and align

- used frequently if you have categorical attributes
- separate plane into regions
- order the regions
- align the regions
- bar chart, stacked bar chart, streamgraph

Spatial axis orientation

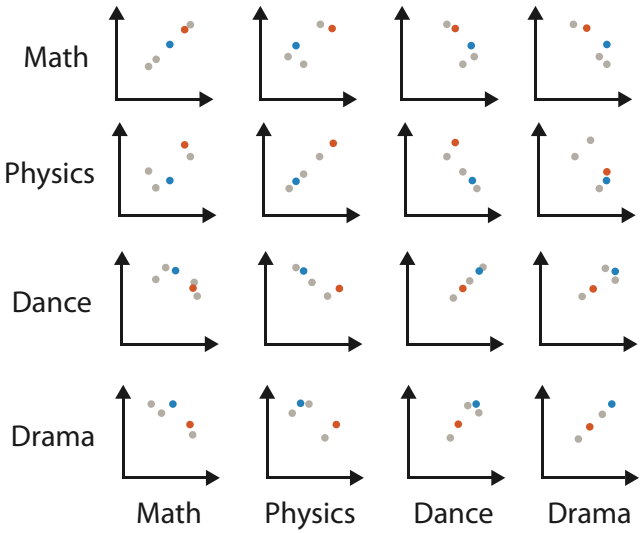
- Rectilinear layouts
- Parallel layouts
- Radial layouts

Parallel layouts

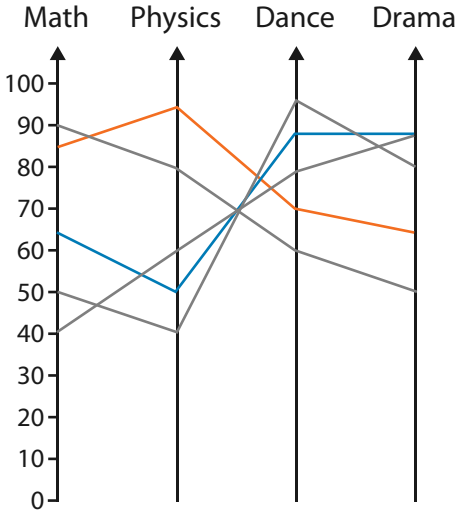
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrix



Parallel Coordinates

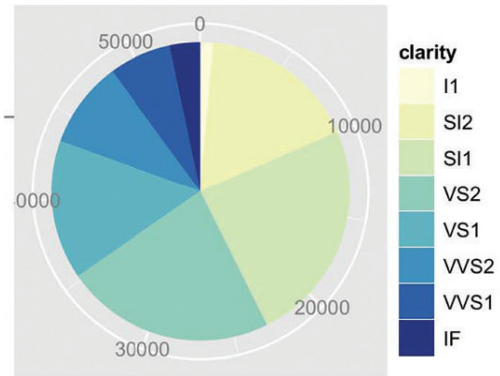


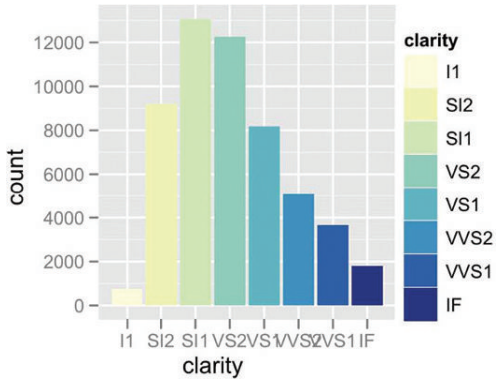
used for

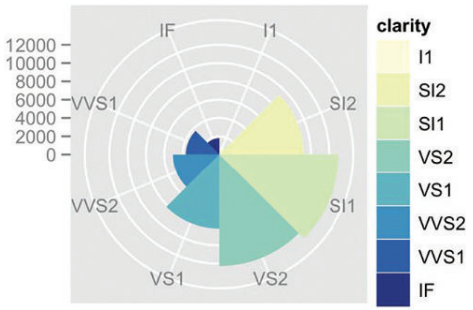
- overview over all the attributes
- range of the individual attributes
- outlier detection
- requires training time for the user

Radial layouts

- pie chart (avoid, bar chart is better)
- polar pie chart



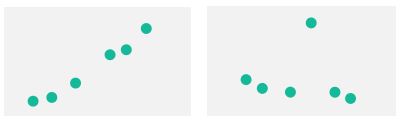




Facet: Juxtaposition

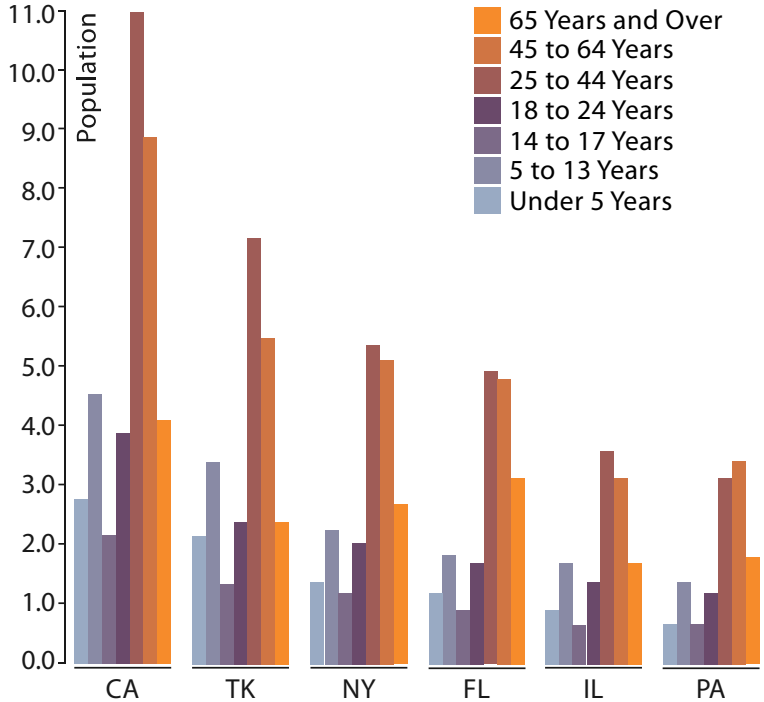
Multiple views

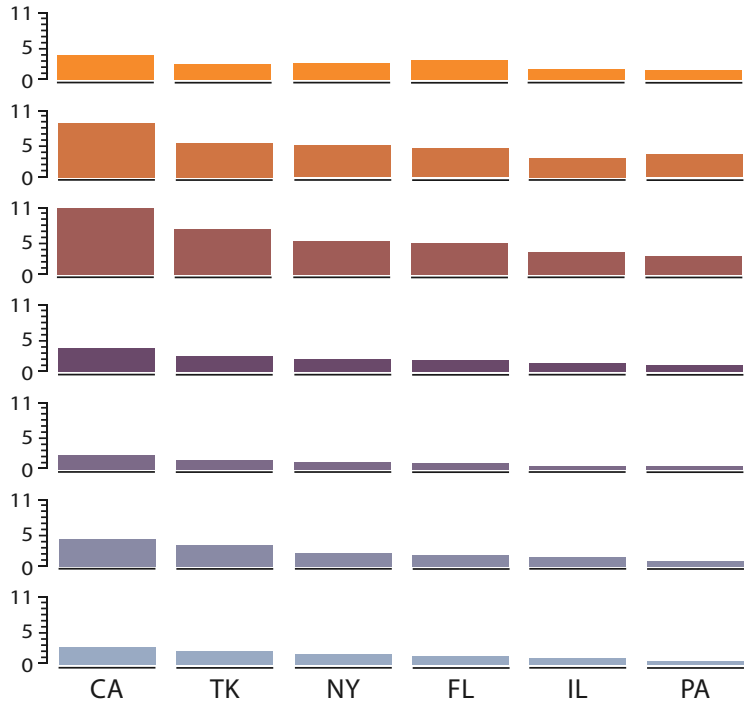
- is the encoding different or same?
- do views share data?

		Data		
		All	Subset	None
Encoding	Same	Redundant	 Overview/ Detail	 Small Multiples
	Different	 Multiform	 Multiform, Overview/ Detail	No Linkage

Facet: partition

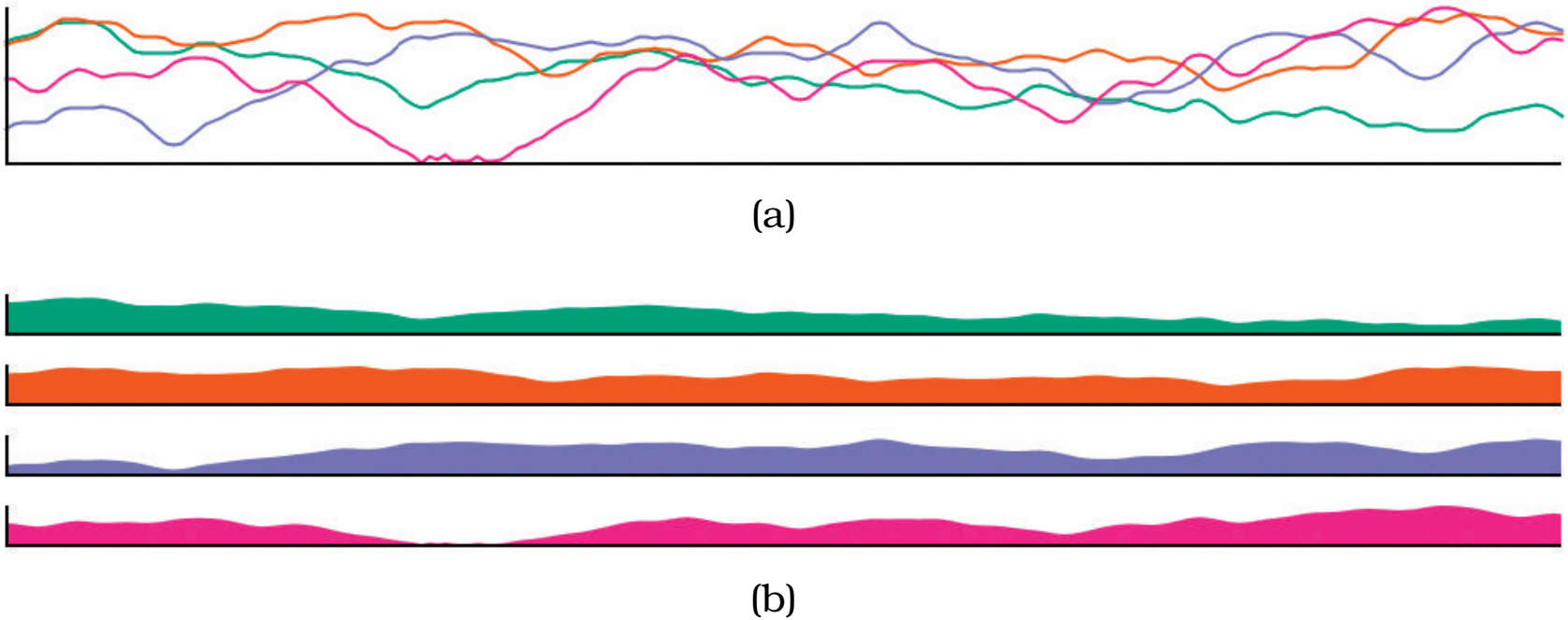
- partition is a design choice on how you separate data into groups
- these groups are assigned to regions (due to separate)
- trivial if you have only one key
- more choices, if you have multiple keys





Facet: superimposition

Superimposition vs. juxtaposition



(a)

(b)

- (local) find time series with the highest point at one time point
- (global) is series A at t_A higher than series B at t_B?
- Javed et al. (2010) showed that superimposition is better for local, juxtaposition is better for global.

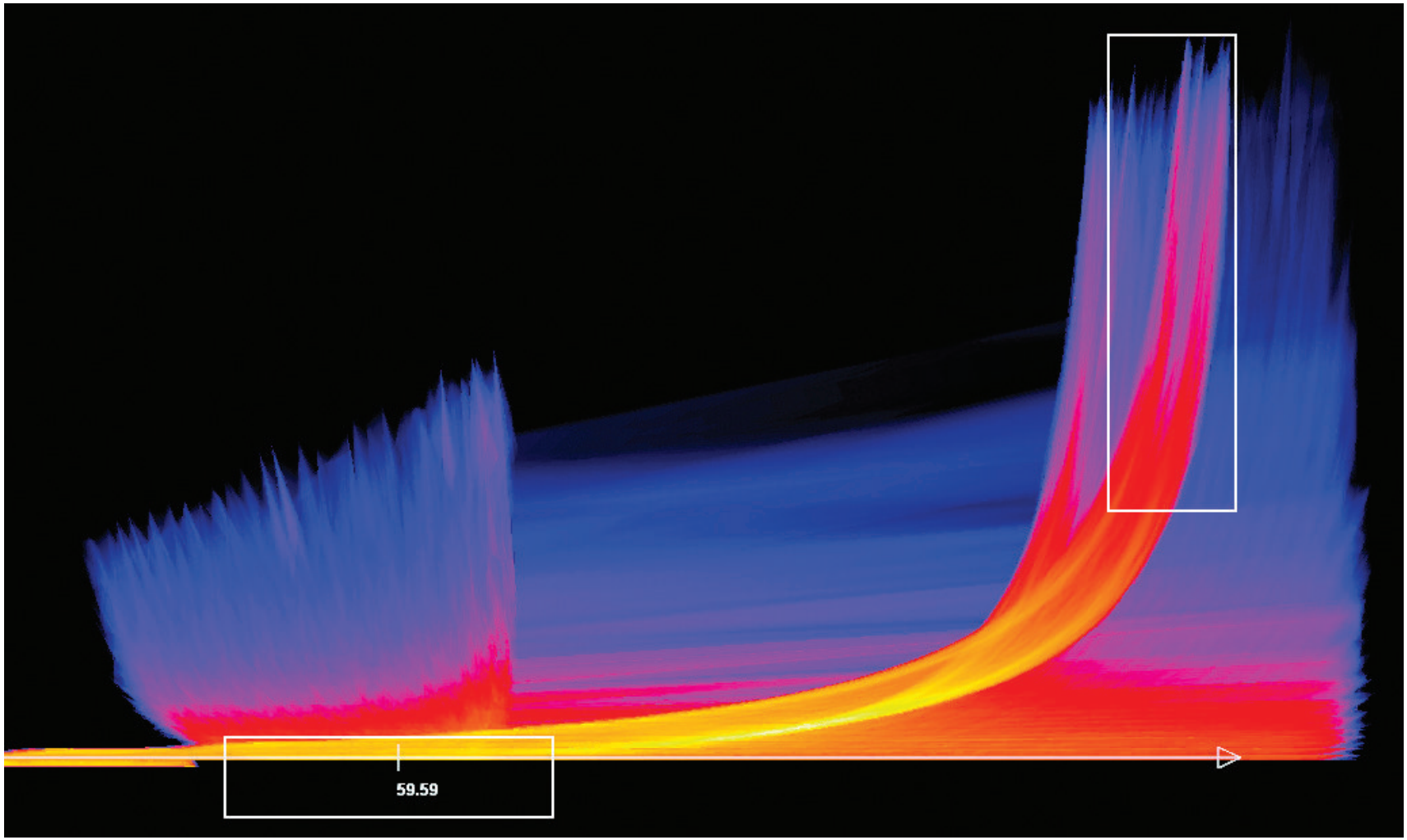
Filter: aggregate

Individual data points are combined, and new derived element represents the new group

- histograms
- continuous scatterplots
- boxplots
- dimensionality reduction techniques (PCA, MDS)

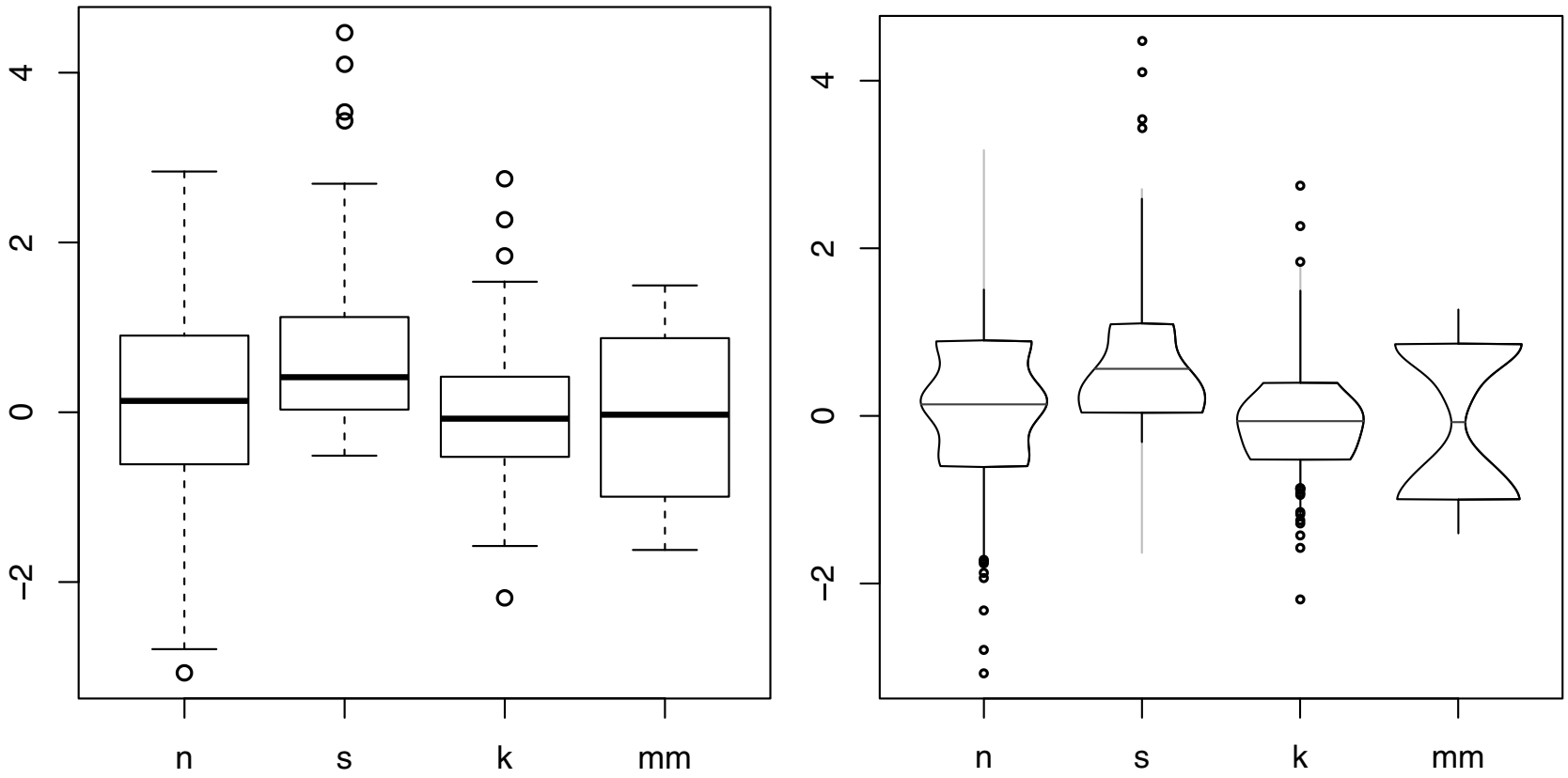
Continuous scatterplot

'heatmap' from data points



Boxplots

standard box plots vs. vase plots



Conclusions of Part II

- learn how to make simple plots well
- avoid: small fonts, bad ranges, vague captions
- plots should have a story
- the ink in the plot should be justified
- plots should correspond to the data and the story
- plots can be complex, if they are explained
- use luminance well, check if the plot works in b&w
- use colors well (opponent process theory, preattentive features)
- first function, then form
- design, forget, check, revise