Aalto University
School of Science
and Technology

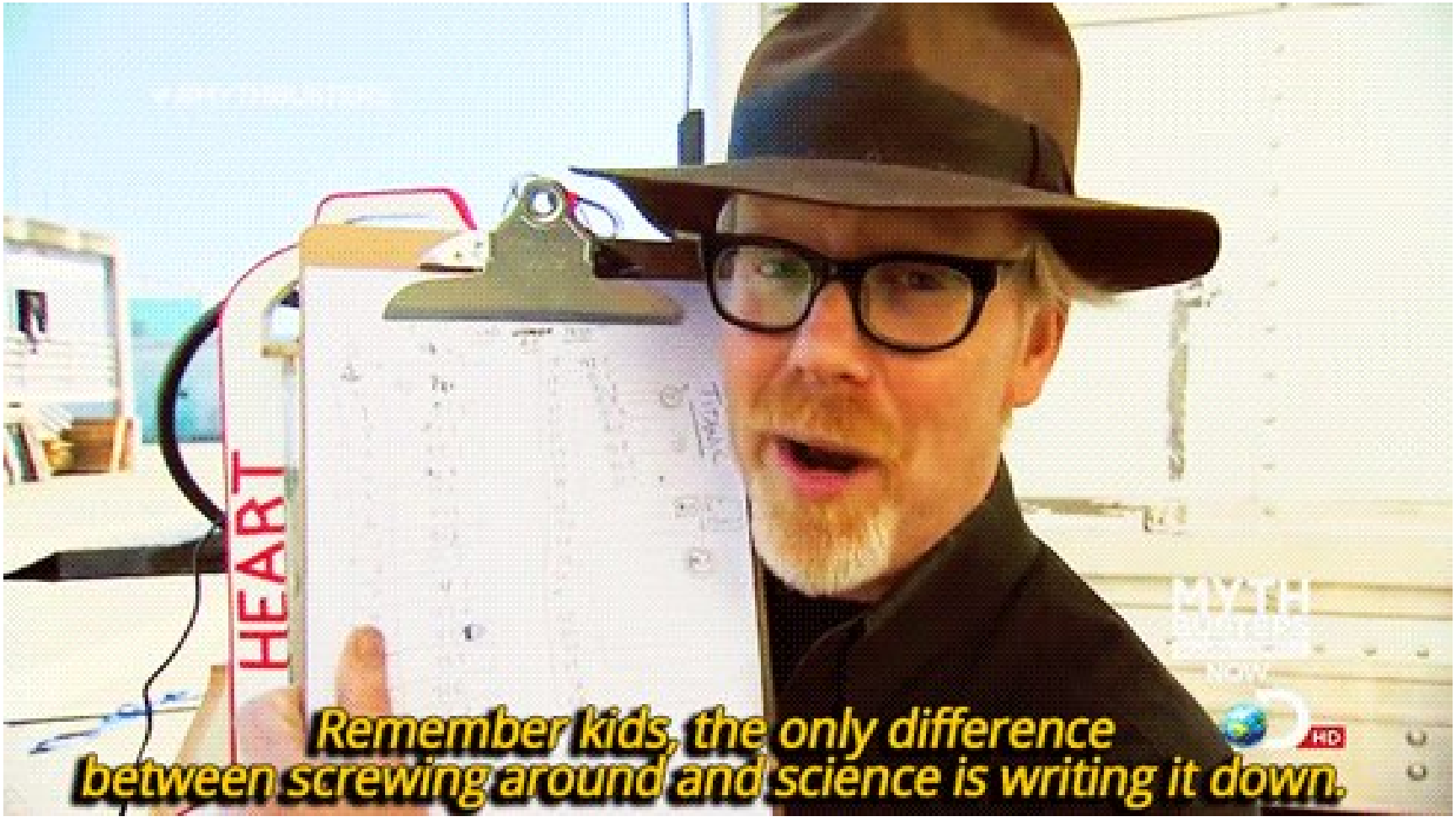
The data scientist's guide for writing papers

Nikolaj Tatti

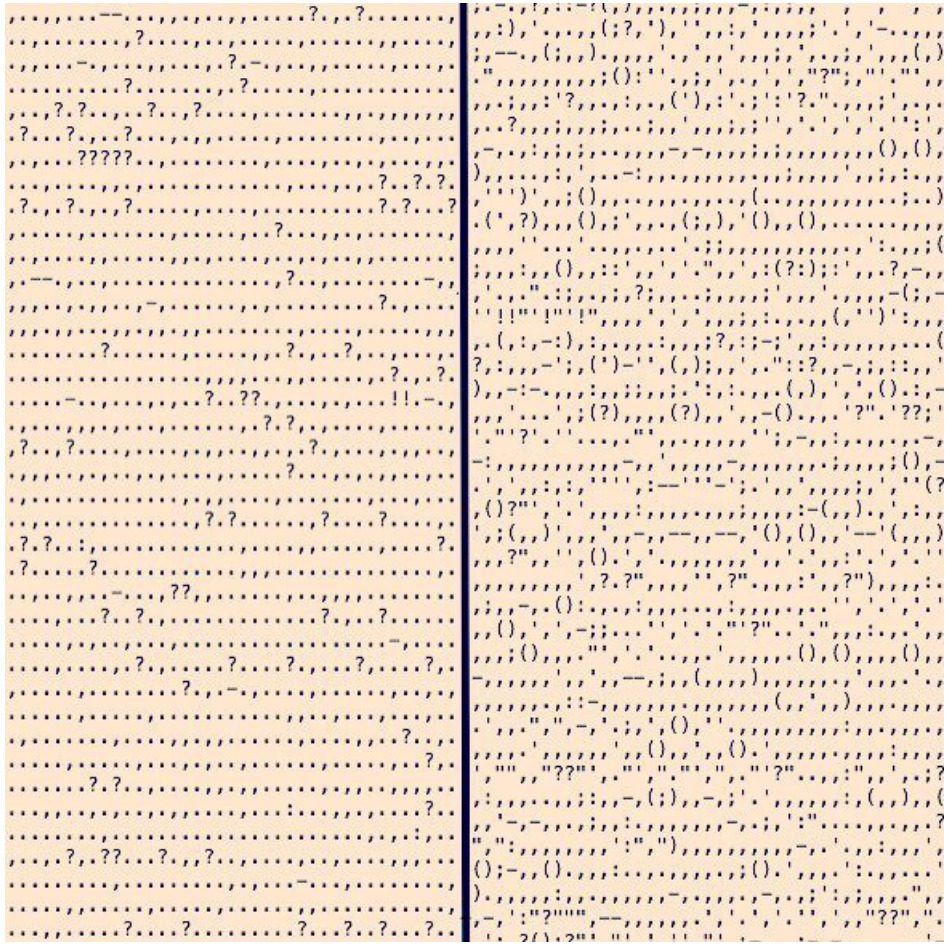
Department of Computer Science, Aalto University
nikolaj.tatti@aalto.fi

September 23, 2016

Why writing


HEART
Remember kids, the only difference
between screwing around and science is writing it down.

Writing is personal



Punctuation in Blood Meridian by Cormac McCarthy (left) and in Absalom, Absalom! by William Faulkner (right).

Why should we care about writing?

- Writing is personal and unique.
- Reading is also personal and depends heavily on the background: some find paper easy to read, some will strugle.
- So, can be there any rules, and why should we care?

- If you don't care about your paper, then you don't care about the reader.
- Why should the reader care about you?
- There are not that many strict rules.
- But there are guidelines that you should follow by default.
- You can break these guidelines, if there is a need.
- You should be aware of the consequences of your choices.

- The more you know about writing, the less time it takes you to write.
- More time to do research!

Structure of the tutorial

Part I: writing

1. introduction / literature
2. on the process of writing
3. structure
4. style
5. typesetting math
6. theorems and proofs

Part I: visualization

1. introduction
2. lessons from Tufte
3. lessons from Ware
4. lessons from Munzner

About this tutorial

Disclaimers

- There is little scientific content in this tutorial
- This tutorial is specifically about writing, not producing content
- You may have different opinions, and this is perfectly acceptable

Audience

- novice writers
- senior writers may find some parts trivial, and some parts may be very interesting

A better title for the tutorial: Stuff (about writing) I wanted to know earlier

What should I read?

- don't look for academic papers as a model example on how to write
- don't look latex manuals on how to write
  - latex manuals teach you how to typeset certain things
  - that doesn't that those thing should be used
- read 2–3 books on writing
- read 2–3 books on visualization
- read scientific papers with a critical eye on the style

Literature

Material on writing

- Writing for computer science by *Justin Zobel*
- The elements of style by *Strunk, White, and Kalman*
- Mathematical writing by *Donald Knuth* (online)

Material on visualization

- The visual display of quantitative information by *Edward Tufte*
- Information visualization: perception for design by *Colin Ware*
- Visualization analysis and design by *Tamara Munzner*

Literature

Useful latex manuals

- The not so short introduction to Latex by *Tobi Oetiker* [link]
- Math mode by *Herbert Voß* [link]
- Tikz / pgf for plotting diagrams [link]
- Pgfplots [link]

Books dedicated to proofs

- Proofs from THE BOOK by *Aigner and Ziegler*
- Mathematics made difficult by *Carl E. Linderholm*

Part I: Writing

On the process of writing

Good rules for writing process

Start early

- deadlining is a cardinal sin!
- writing is left to the last moment but this should not be the case
- deadline writing introduces errors and results in bad presentation of your ideas
- good paper is a result of several revisions, this takes time
- start writing immediately once you have something

Make things easier for yourself

- use latex (if you can)
- use version control (if you can)
  - especially with multiple authors
  - especially if you split each section in its own file
- use latex features
  - macros so that you can change your names later easily
- learn these tools while not deadlining!

Good rules for writing process

Learn how to typeset technical content

- learn how to typeset figures
- learn how to typeset tables
- learn how to typeset pseudo-code
- learn how to typeset equations
- don't do this while deadlining

Revise your text

- write a section, leave it and try to forget it, read it as a reader
- if you get stuck while reading, rewrite it
- don't fall in love with your text

Let it go

- don't overdo the finetuning

Writing order

The content can be divided into 4 groups

1. Content that requires technical typesetting:
   - definitions
   - theorems / proofs
   - pseudocode
   - plots, tables
2. Examples
3. Related work
4. Text
   - introduction / abstract
   - conclusions
   - text between technical content

Writing order

Write technical content first

- this is the contribution of your paper
- you may get stuck / spend a lot of time in writing technical content
- you may have to change presentation later

How to deal with rejections

Reviewers are rarely 100% wrong or 100% right

- don't cry over spilled milk
- instead, try to find the reason why this paper got killed
- if the reviewer misunderstood, try to find the reason for it
- decide should you address the problem, or not

General structure of the paper

What is a well-written, mature paper

- The goal of the paper is to explain the work to the reader.
- The paper is not an ad.
- The paper is not a (chronological) report on what you did.
- Instead, the paper should have a clear, logical story.
- The paper should be an objective take on the topic.
  - you should identify the benefits of your approach
  - and you should also discuss the limitations of your approach
- The writing is clear, to-the-point, and neutral.
- Think about the reader when you are writing.

Reviewer's point-of-view

Reviewers typically look for following things:

1. what is the problem that the authors are solving?
2. how are they solving it?
3. is the solution logical and correct?
   is there a better way to solve the problem?
4. how is the approach related to the existing work?
5. is the problem relevant?

Well-written paper should answer to these questions with ease.

Reviewer's point-of-view

- The main reason for paper being rejected is too thin technical content.
- Papers without content will (probably) get rejected, even if they are written well.
- Badly written papers with good content will also probably get rejected, if reviewers don't understand it.

Standard structure of data mining paper

1. Introduction
2. Preliminaries
3. Problem definition
4. Theory sections
5. Experimental section(s)
6. Conclusions

- Related work is somewhere between the introduction and the conclusions.
- This structure is specific to data mining papers.
- Papers in other fields have very different structure.
- Sometimes, it's even mandated by the venue.

Introduction

Introduction should explain

1. what are you solving in plain text,
2. why this is an interesting problem,
3. an overview on how do you solve it,
4. why your approach is logical,
5. what are your main (theoretical/experimental) results.

After reading the introduction, the reader should know the paper on a high level.

Preliminaries and problem definition

- Preliminary notation is the minimal notation needed to describe
  - the setting,
  - input data,
  - output data
  - the problem that you are solving.
- The problem definition should come as soon as possible.
- The problem definition should be stated formally.
- Definitions related to the solutions should be in the theory sections.

Theory / experimental sections

Theory sections should contain

- theoretical analysis of the problem, if any
- approach(es) to solve the problem
- theoretical analysis of the algorithms.

Experimental section(s) should contain

- an accurate description of your setup,
- an objective discussion about the obtained results.
- reader should, in theory, replicate your experiments.

Related work

- Related work section appears after the introduction and before conclusions.
- 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 definitions.
- The golden rule: treat the related work the way you want your work to be treated.

Related work can be roughly divided into two categories

- related problems
- related techniques

Related work

- Explain the connections to the prior work.
- Don't dismiss related work if its problem setting is a superficial 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, if it makes sense.

Examples

The purpose of example is to make sure that the reader understood the statement

- The examples should be complex enough so that the fine(r) point of the definition becomes clear.
- At the same time, they should be simple enough so that reading them is not a chore.
- Running example, a series of examples, sharing the same setup/data is typically a good idea.

Structure

- Paper must have logical structure.
- You should explain the structure explicitly to the reader.
- At the beginning of the section, answer why this section is needed.
- Before or immediately after a definition/proposition, explain why do you need it, and how are planning to use it.
- Before starting a logical argument, explain what is the goal of your argument.
- Section and subsection titles should be specific, instead of vague.

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

Common ways of structuring the work

There are many ways of structuring your work.

- The most appropriate one depends on the nature of the work.
- They are not mutually exclusive.

By chain:

- here the order is mandated by the topic
- problem definition
- previous approaches, if any
- your approach
- demonstration that you do better than the baselines

By specificity:

- first high-level description
- then descriptions of lower layers
- for example, main algorithm and then subroutines

Common ways of structuring the work

By example:

- apply your approach to a small, familiar problem
- then introduce your approach
- the example then makes your approach more concrete

By complexity:

- first introduce a simple, special case of your framework
- then extend it to a more complex case
- this may ease the learning curve, especially with complex notation

Style

The main problem in writing

The goal of the paper is to explain your work

- it's easy to explain simple concepts in simple terms
- it's easy to 'explain' difficult concepts in difficult terms
- it's difficult to explain difficult concepts in simple terms

- this is the main difficulty among novice writers
- very hard to teach as it depends heavily on the context
- only comes with experience
- to write clearly
  - writer should understand the core nature of the concept
  - should be able to emphasize with the reader
  - should have the experience to express himself
- grammar is typically not the bottleneck

General tone

- have one idea per sentence or paragraph, and one topic per section
- have a straightforward, logical organization
- avoid excess, in length or style
- omit unnecessary material
- keep paragraphs short
- use short sentences, with simple structure
- use short words
- avoid buzzwords cliches, and slang
- be specific, not vague or abstract
- break these rules if there is a good reason

Economy

- the length of the paper should reflect its content
- every sentence should be necessary

× 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.

Economy

- the length of the paper should reflect its content
- every sentence should be necessary

✓ 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.

Economy

- don't cut text too much
- clarity is the priority, and economic writing is means to that

× 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

× The following theorem can now be proved.

✓ We can now prove the following theorem.

- it is easier to write and read
- useful when distinguishing your contribution from related work
- use passive voice, if necessary

Also avoid (more subtle) indirect sentences

× Local packet transmission was performed to test error rates.

✓ Error rates were tested by local packet transmission.

- you can typically remove the following words: perform, utilize, achieved, carried out, conducted, done, occurred, effected

Tense

Present tense is meant for eternal truths

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

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

Past tense is for describing the work (and outcomes).

× the ideas are tested by experiment

✓ the ideas were tested by experiment

Occasionally, it is correct to mix tenses

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

Either past or present tense can be used for discussion of references.

Paragraph structure

- paragraph should discuss one topic at a time
- typically, the first sentence outlines the argument; the following sentences elaborate on the topic.
- context can be forgotten between paragraphs so be careful with cross-paragraph references

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

- usually they should be 1–2 lines long
- don't say too much at once

Avoid information in nested sentences.

× In the first stage, which is the backtracking tokenizer with a two-element retry buffer, errors, including illegal adjacencies as well as unrecognized 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 unrecognized tokens; when detected, errors are stored on a stack for collation into a complete report.

Sentence structure

Beware of fractured '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 acceptable 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.

Double negatives

Avoid double negatives

× There do not seem to be any reasons not to adopt the new approach.

- here, the impression is condemnation: we don't like the approach but we are not sure why.
- the goal was opposite

✓ The new approach 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.

Backtracking

Organize sentences so that they can be parsed without too much backtracking.

× Classifying handles can involve opening the files they represent.

- without the context of the rest 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 this 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 shows that removal of features with low amplitude can dramatically reduce costs.

Parallel structures

Complementary concepts should be 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, direct words rather than long, circumlocutionary words

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

- use an exact long word rather than an approximate short one

Use specific and familiar words, avoid vague terms that have different meanings for different readers

× The analysis derives the information about software.

✓ The analysis estimates the resource costs of software.

Choice of words

Be careful with abstract and vague terms:

- important, intelligent, method, paradigm, performance, semantic, difficult, hard, efficient.
- common reason to use vague terms is to avoid word repetition
- it's better to repeat the word than use vague synonym

Avoid slang and contractions

- use cannot instead of can't

Qualifiers

Don't pile qualifiers on top of another

- Within a sentence, use at most one of the qualifiers such as might, may, perhaps, possibly, likely, could.

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

✓ The algorithm might fail on unusual input.

× We are planning to consider possible options for extending our results.

✓ We are considering how to extend our results.

Double negative acts sometimes as a qualifier (avoid)

× Merten's algorithm is not dissimilar to ours.

Qualifiers

Consider deleting meaningless qualifiers:

- very, quite, totally, completely, truly, highly, usually, accordingly, certainly, necessarily, somewhat

× 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

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

Adjectives can be used as padding, and should be avoided

$\times$ 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.

Foreign expressions and abbreviations

Avoid foreign expressions

- not everybody knows them

× Some writers feel that use of foreign words is de rigueur because it lends the work certain je ne sais quoi and shows savois-vivre.

- scientists have irrational love for latin expressions
- it's also better to avoid them

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

Avoid abbreviations

- they save some space but they will slow the reader down

× w.r.t.

✓ with respect to

Wordy expressions

wordy	concise
completely random	random
completely unique	unique
completely optimized	optimized
adding together	adding
cancel out	cancel
conflated together	conflated
divided up	divided
cooperate together	cooperate
totally eliminated	eliminated
joined up	joined
merged together	merged
free up	free
separate into partitions	partition
reason why	reason
after the end of	after
during the course of	during

Wordy expressions

wordy	concise
in the region of	approximately
let us now consider	consider
currently . . . today	currently
give a description of	describe
of fast speed	fast
first of all	first
for the purpose of	for
in the view of the fact	given
of large size	large
the vast majority of	most
it is frequently the case that	often
at a fast rate	quickly
a number of	several
such as . . . etc.	such as . . .
in the majority of the cases	usually
whether or not	whether
it is a fact that	—

Typesetting citations

There are 3 major citation styles

- code in brackets: numeric [16] or generated [AMS+96].
- Harvard style: (Smith et al., 1996)
- superscripts: better performance might be possible with hash techniques[16].

Citation should not be used as a noun:

× [16] conducted a study.

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

- this is most obvious with superscript-style

Citation are also typically listed after a statement:

✓ Better performance is possible with hash techniques [16, 30].

Typesetting citations

Harvard style is tricky:

- (Smith et al., 1996) = [16] = [16], but
- Smith et al. (1996) = Smith et al. [16] = Smith et al.[16]

In the latter

- 'Smith et al.' is technically not part of the citation.
- (1996) is the shortened version of (Smith et al., 1996)
- it is shortened because Smith et al. is redundant

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

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

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

Typesetting citations

The citation style is typically constrained by a publisher.
Latex-tip:

- natbib-package is handy for spelling the names out.
- \citet $\rightarrow$ Smith et al. (1996) or Smith et al. [16]
- \citep $\rightarrow$ (Smith et al., 1996) or [16]
- by default uses Harvard style
- but can be configured for numeric style

Typesetting tables

- Tables should not contain vertical lines or double lines

Dataset	parameter	score	time
Dolphins	0.4	103	45s
Karate	0.2	42	2m4s
Lesmis	0.1	107	1m14s

Dataset	parameter	score	time
Dolphins	0.4	103	45s
Karate	0.2	42	2m4s
Lesmis	0.1	107	1m14s

- Booktabs package makes topline and bottomline thicker

Typesetting tables

- table headers can be arranged into a hierarchy
- use \cmidrule for intermediate lines
- in \cmidrule(lr){3-4}: (lr) will shorten the lines

			Algorithm 1		Algorithm 2	
	Dataset	parameter	score	time	score	time
✓	*Dolphins*	0.4	103	45s	92	5m
	Karate	0.2	42	2m4s	29	10m
	Lesmis	0.1	107	1m14s	31	3m10s

Captions

Captions should be informative and exact

- ideally, 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)

- it is possible to have more details in caption than in the main text, for example parameter values.

Mathematical notation

Mathematical expressions

- mathematical expressions are tools that allow you explain explain exactly and clearly
- you can always say it in text, but often it's 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.

- Math is not only meant for linear algebra!

History of mathematics

- Earliest number system, 3400 BC
- $+$ sign, around 1360
- $-$ sign, 1489
- $=$ sign, 1557
- $\lceil \cdot \rceil$, $\lfloor \cdot \rfloor$ signs, 1962

- it was common to write equations in text

× Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos & generaliter nullam in infinitum ultra quadratum potestatem in duos eiusdem nominis fas est dividere cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

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

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

Math as jargon

× Let $\langle S \rangle = \{\sum_{i=1}^{n} \alpha_i x_i \mid \alpha_i \in F, 1 \leq i \leq n\}$. For $x = \sum_{i=1}^{n} \alpha_i x_i$ and $y = \sum_{i=1}^{n} \beta_i x_i$, so that $x, y \in \langle S \rangle$, 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

$$\langle S \rangle = \{ \sum_{i=1}^{n} \alpha_i x_i \mid \alpha_i \in F \} \quad .$$

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

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

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

Inline vs. display

Mathematical formulas can be written either inline, $f(x) = x^2 + ax + b$, or in display,

$$f(x) = x^2 + ax + b \quad .$$

- displays and inline math are grammatically equivalent
- they should 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

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

- works well if you are writing single-column
- text should be wrapped in \text or \textrm
- it's a good idea to have a decent space (\quad) between two equations

Complex equations

- consider breaking down equations by using helper functions

✗

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

✓

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

- consider introducing variables before using them,

✓ Let

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

and

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

Writing formulas

Symbols in different formulas must be separated by words

× Consider S_q, $q < p$.

✓ Consider S_q, where $q < p$.

Don't start a sentence with a symbol.

× $x^n - a$ has n distinct zeroes.

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

Use 'that' when it helps to parse the 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 $x = y$.

Quantifiers

- avoid $\forall$, $\exists$, $\wedge$, $\vee$, unless you write about logic

× Here,

$$gcd(x, y) = \max\{z \mid \exists u, v : x = uz \wedge y = vz\}$$

✓ Here,

$$gcd(x, y) = \max\{z \mid x = uz, y = vz, \text{ for some } u, v\}$$

✓ Here, $gcd(x, y)$ is the largest integer that is a factor of both x and y.

Notation design

- design notation that allows you to express yourself clearly
- ... and that is read easily
- change it, if it doesn't work

Main principle:

- objects of main focus should be well-presented in the notation
- secondary objects should be in the background, or even omitted

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

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

- when defining, state explicitly all the dependencies.
- if needed, you can drop the trivial dependencies

✓ if D is clear from the context, we will write $q(\mathcal{C})$.

Notation design

- avoid parameters as subindices / superindices.
- However, you can use subindices to create instances

$$q_{\text{greedy}}(x), \quad \text{and} \quad q_{\text{opt}}(x)$$

  are instances of the quality score $q(x)$.
- long names vs. short names
- long names are useful but can clutter the text, if used often
- if you use object often, consider a short name

Subscripts

Subscripts need to be justified

× Let $x_i, x_j \in X$.

✓ Let $x, y \in X$.

- Unnecessary indices is a waste of resources: x doesn't carry any information.
- Use only if you are going to use the indices i and j

✓ Let $x_i, x_j \in X$, where $i < j$.

- Sometimes used to indicate connection between variables: i_1 and i_2 are two indices.
- Typically, i and j are better notation than i_1 and i_2.

Rewrite subsubscripts and supersubscripts

× $x_{i_j}^{k_a^n}$

Accents

Be vary of accents

- often it's better to avoid them, i and j is better than i and i'
- they work if you are low on symbols,
- or you want to tie some symbols together,

✓ $a^* = \max A$, and $a \in A$.

✓ Let a' be alternative solution.

- don't combine accents; avoid $\overline{\hat{a}}$ and $D'^{i'}_{j'}$.
- avoid multiple accents: a'' or a'^4.

Alphabet

Mathematical alphabet consists of

- small letters, a, b, c, ...
- capital letters, A, B, C, ...
- calligraphic letters, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, ...
- small greek letters, α, β, γ, ...
- capital greek letters, Γ, Δ, Θ, ...

Avoid

- exotic fonts $\mathfrak{A}, \mathfrak{B}, \mathfrak{C}$, $\mathscr{A}, \mathscr{B}, \mathscr{C}$
- exotic greek letters,
  α reads like 'alpha', whereas ξ reads like 'curly fry'
- Σ if you also use sum, Π if you also use product

Typecasting with alphabets

It's good to typecast your symbols with alphabet

- a, b, and c are items
- A, B, and C are sets
- $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are complex objects, for example, families of sets

You can also typecast the nature of your symbol

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

These conventions can be violated, if needed

Latex typography

Long variable names should be wrapped in \mathit

× $offer = surface \times force$

✓ $offer = surface \times force$

Ellipsis obeys the vertical alignment of the surrounding operator

× $x_1, x_2, \cdots, x_n$

✓ $x_1, x_2, \ldots, x_n$

✓ $x_1 + x_2 + \cdots + x_n$

$*$ is not a multiplication

× $a * b$

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

- Use \times, if you have long names, $surface \times force$

Theorems and proofs

Theorems, propositions and Lemmas

- Theorem and proposition is a major statement.
- Lemma is an intermediate result that is typically to use proof a theorem or proposition.
- The statements should be exact, ideally self-contained.

- The proof doesn't have to follow immediately.
- You can state the theorem as a goal, and slowly prepare for the actual proof with lemma and additional notation.

What is a good proof

A good proof

- is correct
- is not vague
- is not an argument
- is a recipe for the reader
- is not just a sequence of formulas
- guides the reader towards the statement

- The proof should be a logical sequence of steps.
- You should explain the overall structure of the proof.
- You should also explain the intermediate structure of the proof.

Divide & conquer proofs

- long proofs typically consist of several large steps
- to impose a structure, it's a good idea to explicitly enumerate them

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

- or name the steps:

✓ Step *i* (construction of the graph): . . .
Step *ii* (reduction step): . . .

Breaking proofs into lemmas

- in a long proof, a single, largely independent, step can be often written as a separate lemma
- this works well if you can also generalize and simplify the lemma

Example:

- as a part of the proof, we need to show that

$$\log\log\frac{1+\mu}{\mu} - \log\log\frac{1/n+\mu}{\mu} \leq \log\log\frac{1+1/n}{1/n} - \log\log\frac{2/n}{1/n},$$

  when $\mu \geq 1/n$.
- this is quite messy, but if we define $h(x,y) = \log\log\frac{x+y}{x}$, we can write the step as

$$h(\mu,1) - h(\mu,1/n) \leq h(1/n,1) - h(1/n,1/n) \quad .$$

Breaking proofs into lemmas

- We can prove a generalized version of the step in a separate lemma

Lemma

Let $h(x, y) = \log\log \frac{x+y}{x}$. *Then*

$$h(x_2, y_2) - h(x_2, y_1) \leq h(x_1, y_2) - h(x_1, y_1),$$

for $x_1 \leq x_2$ *and* $y_1 \leq y_2$.

- we have replaced μ and $1/n$ with general variables
- this states that to prove this step we only need the fact that $\mu \geq 1/n$ and $1 \geq 1/n$

Lemmas

- if 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 it's a good idea to make that step into a lemma.
- If your main proof is in the appendix, so the main lemmas can also go there.

Reverse order of the proof

- Sometimes, your proof consists 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
- Logically, prove the hard step and then the easy step.
- 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, [Step 2]

  We will now prove that X has property P'. [Step 1]

- Step 1 can also be in a separate lemma

Conclusions of Part I

- take writing seriously
- start early
- make things easy for yourself by learning the tools
- write, forget, read, revise

- your paper should be an objective guide explaining your work
- be precise, don't waste ink
- try to write a simple and clear as possible
  (this is the most difficult part)
- use math to make things simpler not to obfuscate
- have empathy for the reader
- be logical, and explain your logic

Part II: Visualization

Visualization

Your goal of visualization:

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

Three most frequent mistakes in scientific plots:

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

Tools for plotting

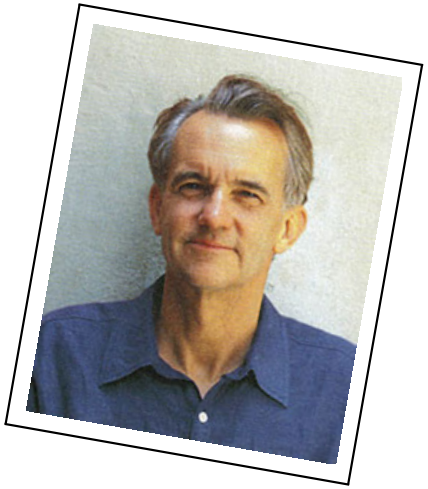
Learn how to make good plots with one of these:

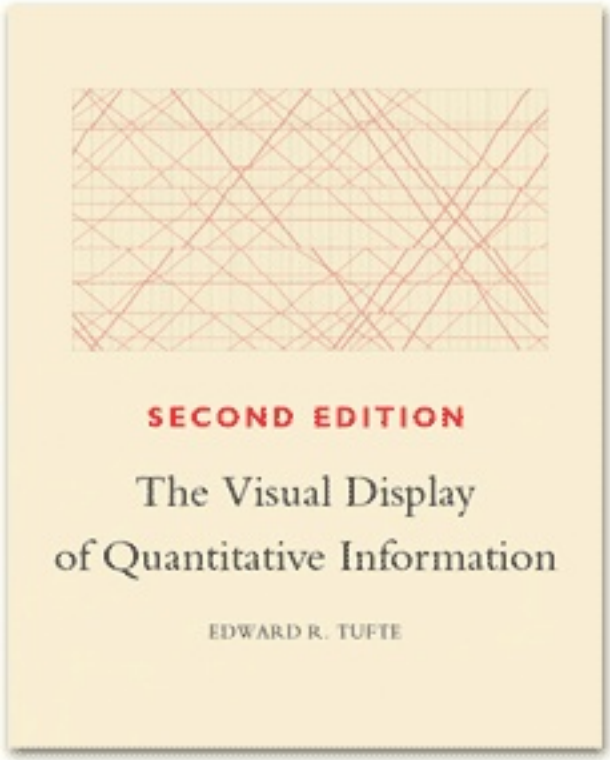
- Matlab
- R
- gnuplot
- Python
- Pgfplots/tikz (latex package)

Recommend Python for beginners, and pgfplots for control freaks.

Lessons from Tufte

The visual display of quantitative information





Tufte's archenemy are bad plots published for laymans

- the examples of bad plots are really extreme (and they are old)
- however similar mistakes can be also found in scientific plots, but they are more subtle

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.

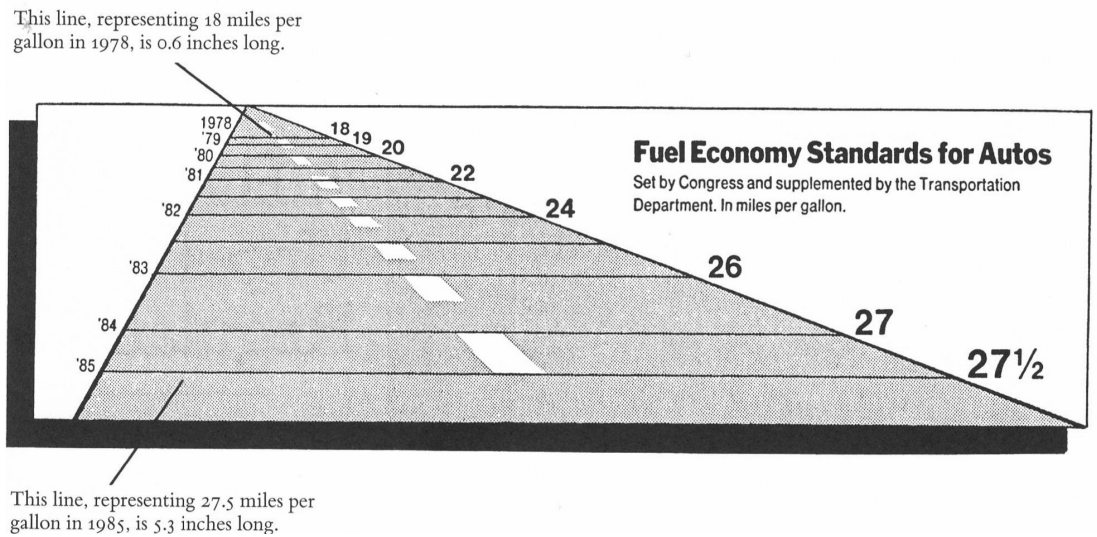
Example: fuel economy standards

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

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

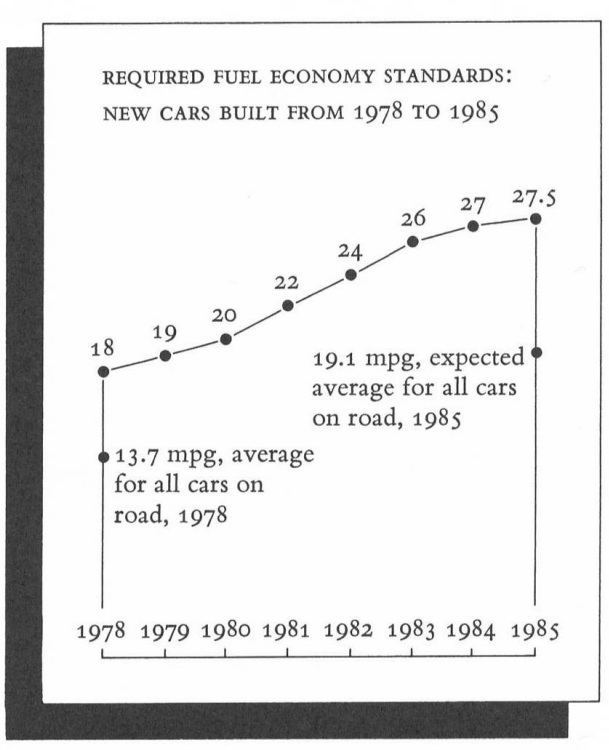
Lie factor $= 14.8$

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



The New York Times (1978)

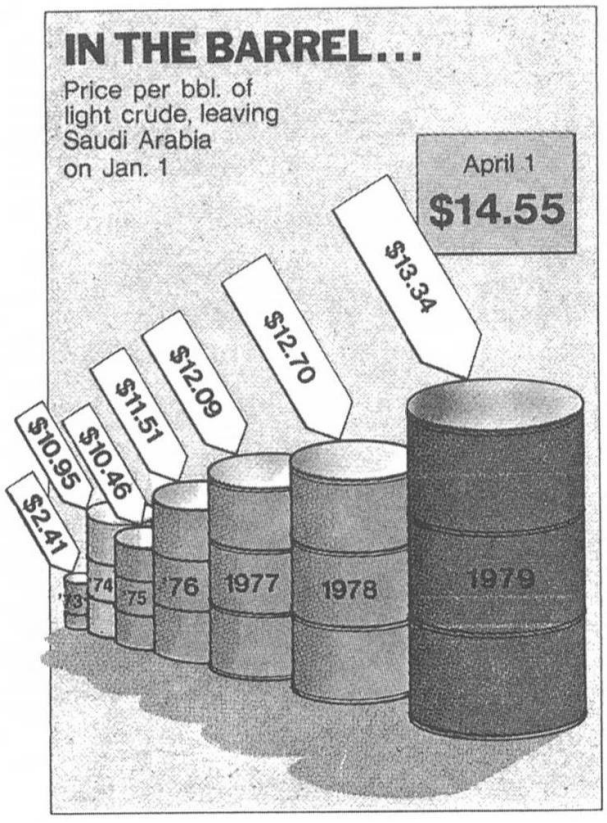
Example: fuel economy standards (cont.)



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

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 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

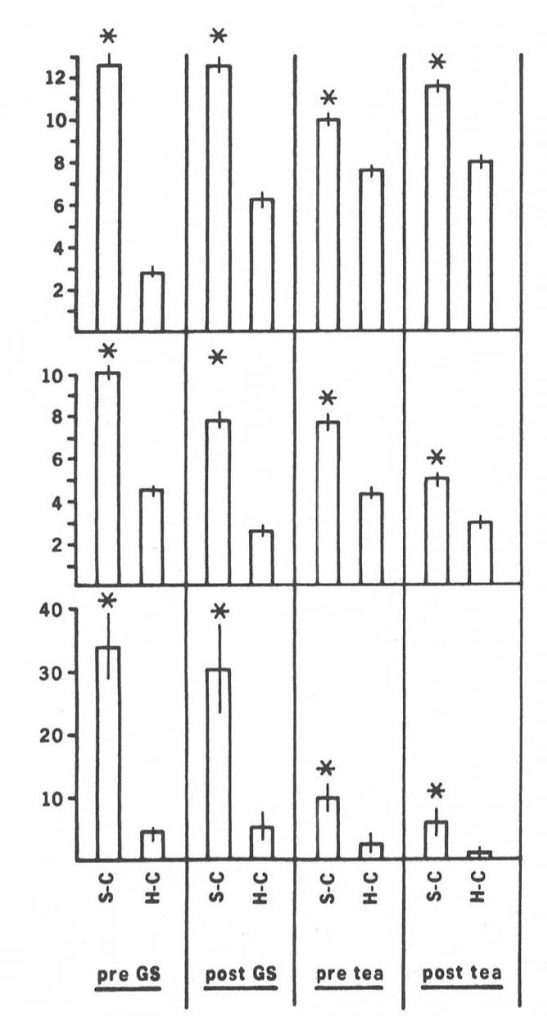
Edit and redesign

(A bad) example: This display tries to compare each long bar with the adjacent short bar, under various experimental conditions, to show that the long one is always longer (uh?)

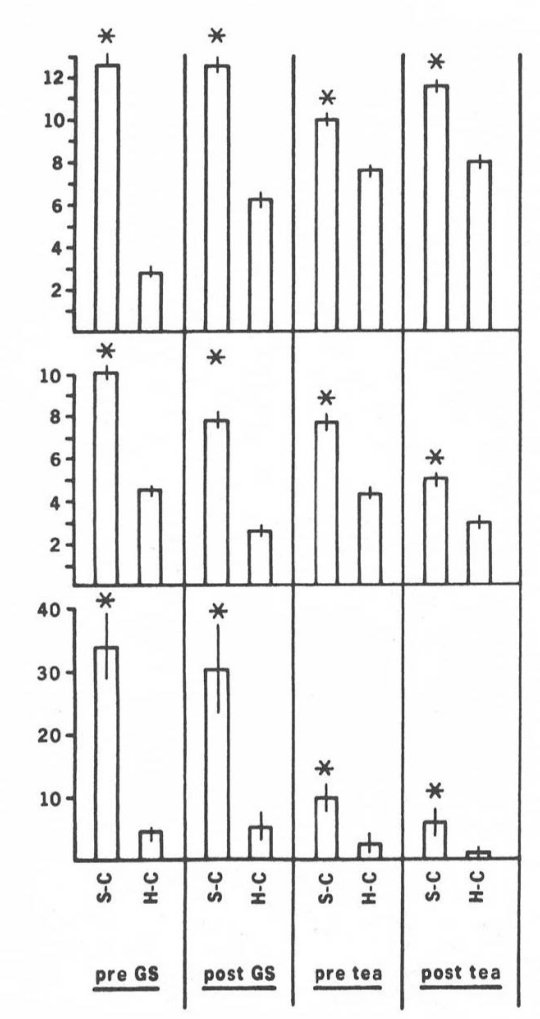
Improve figure with revising and editing

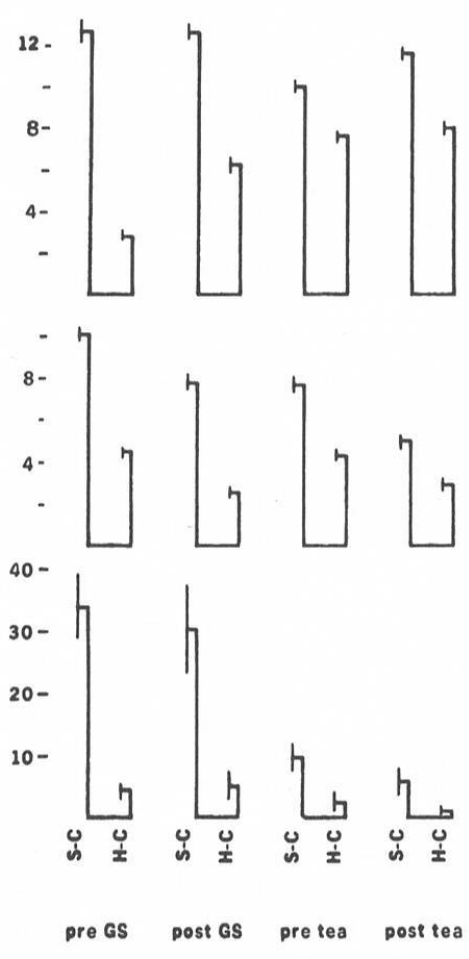
Maximize data-ink ratio by

- erasing non data-ink
- redundant data-ink.



Edit and redesign (cont.)





Pruning improves the graphic while retaining all the information in the original data

- erase bilateral symmetry
- asterisks are out because unnecessary
- the framing structure was erased, too

65% of the original was erased

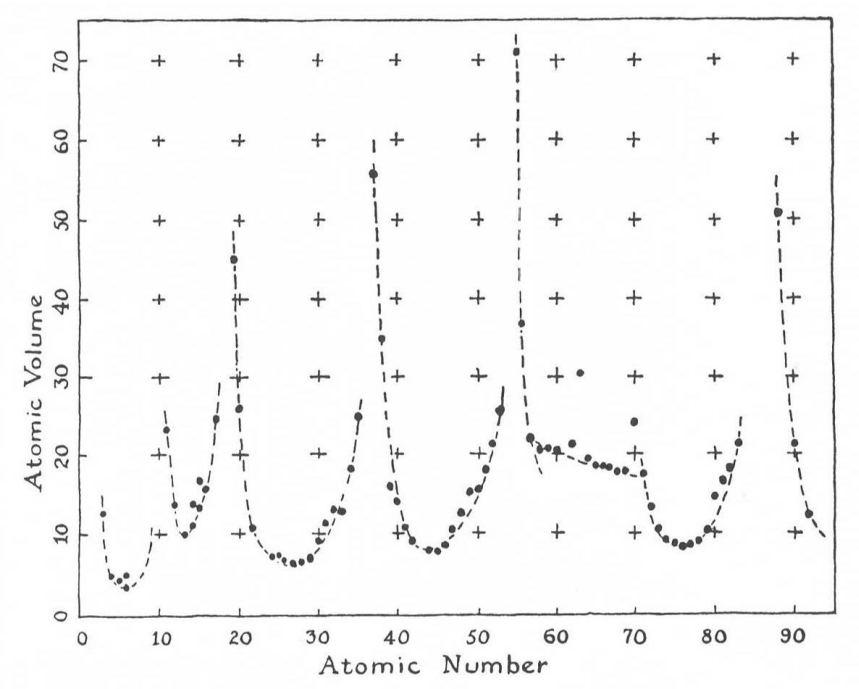
- a mix of non data-ink and redundant data-ink

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

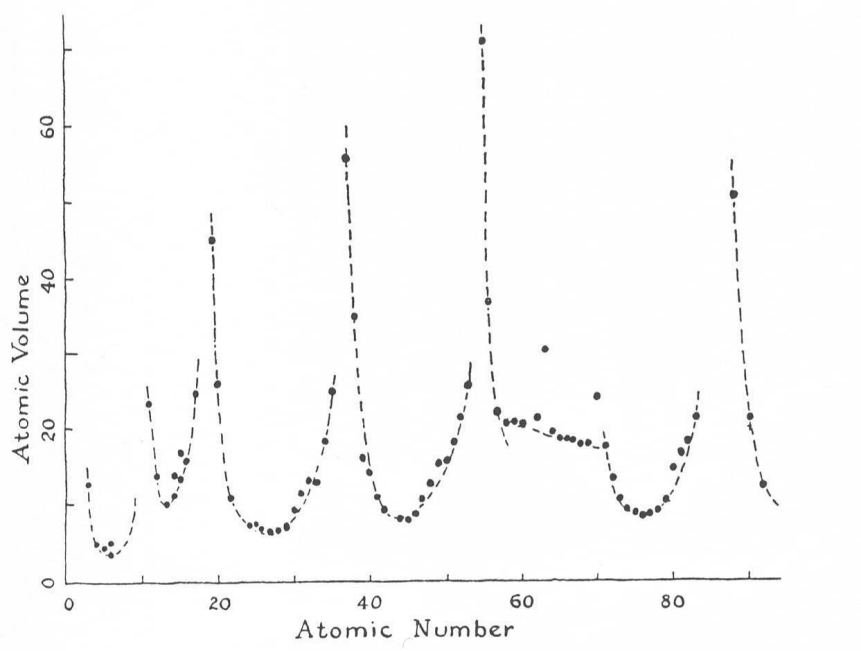


Linus Pauling, General Chemistry, p. 64, 1947

Edit and redesign (cont.)

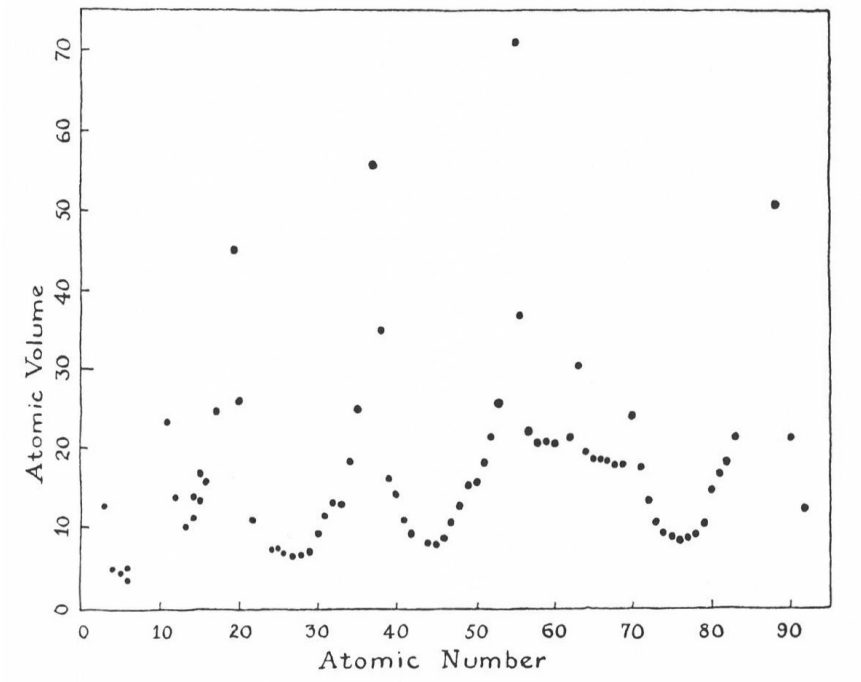
Data-ink ratio improves to 0.9

- 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



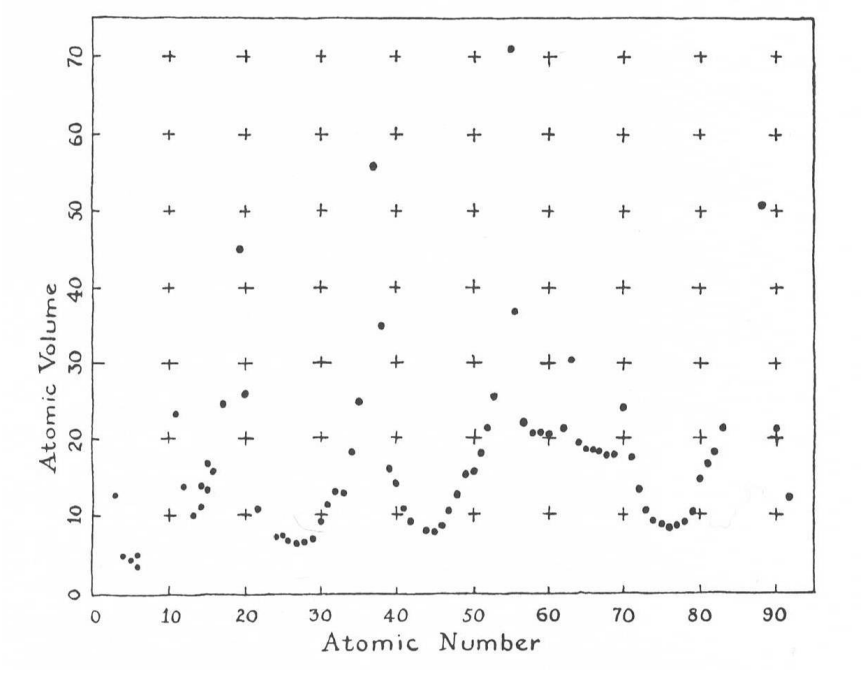
The reference curve is essential in organizing the data, and shows the periodicity (the message) by creating a structure, and by giving ordering and hierarchy

Edit and redesign (cont.)



Without the curve we hardly detect the periodicity

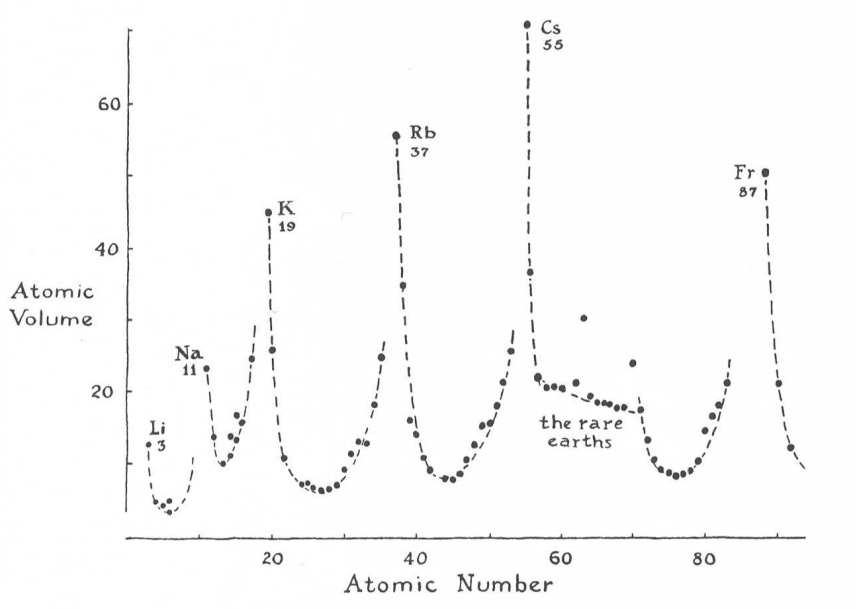
- the curve becomes necessary because the eye needs guidance



Restoring the grid totally fails to organize the data

- the grid marks are too powerful and induce visual vibration

Edit and redesign (cont.)



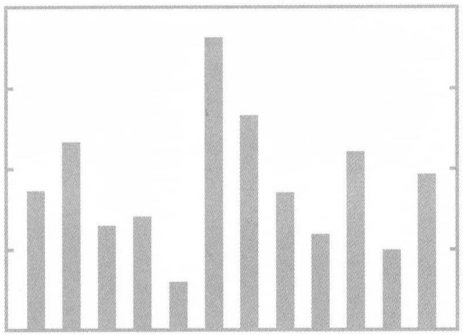
We can use the erased space

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

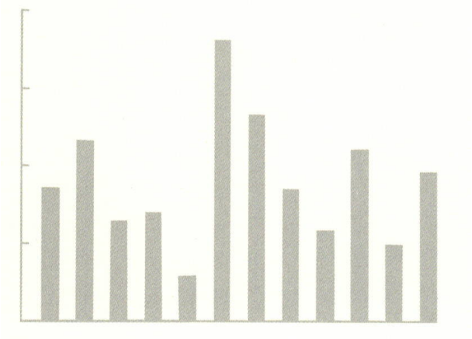
Take-home message:

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

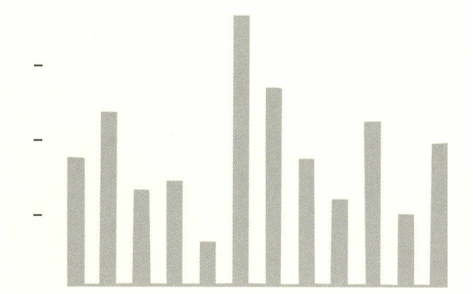
Redesign of the bar chart



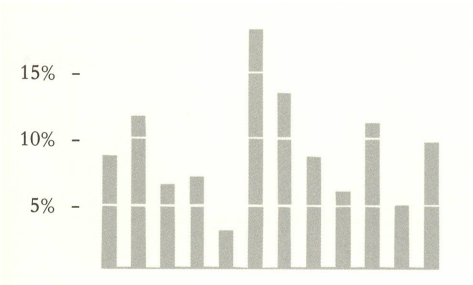
A standard model bar chart, with the design endorsed by the practices and the style sheets of many publications



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

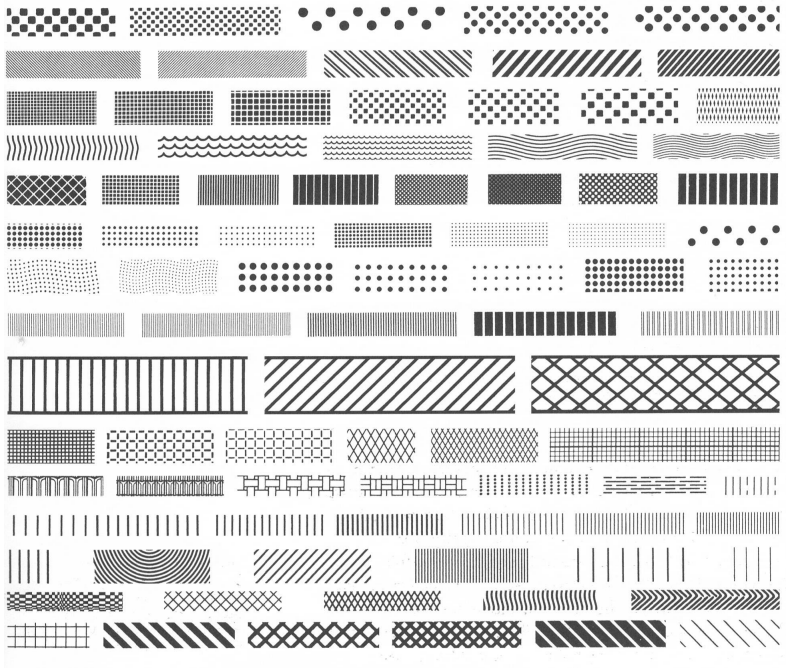
Chartjunk:

- 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)

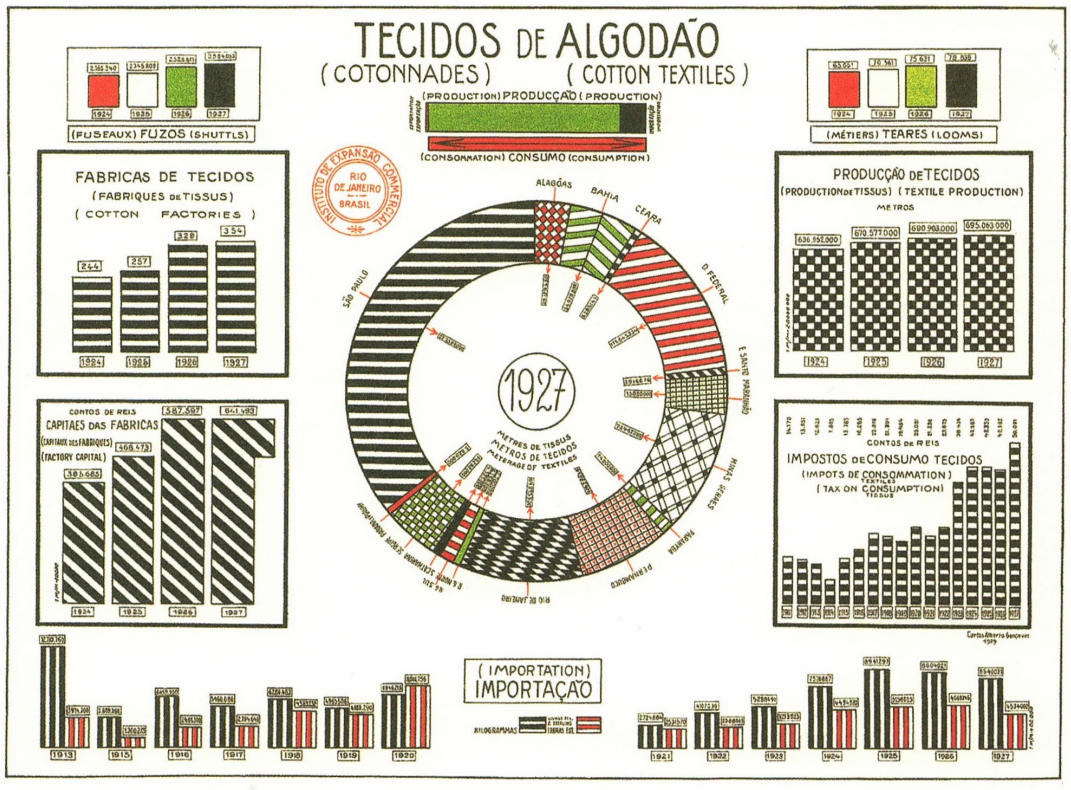
Unintentional optical art

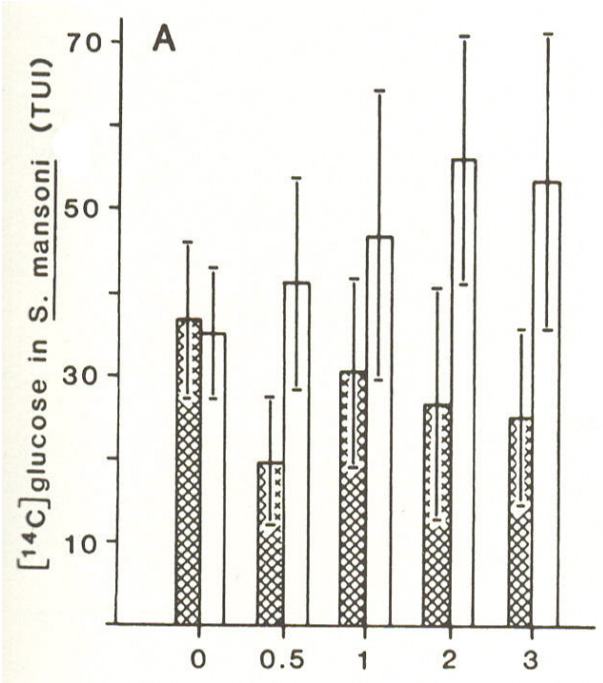
Happens in plots when areas are filled with patterns

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



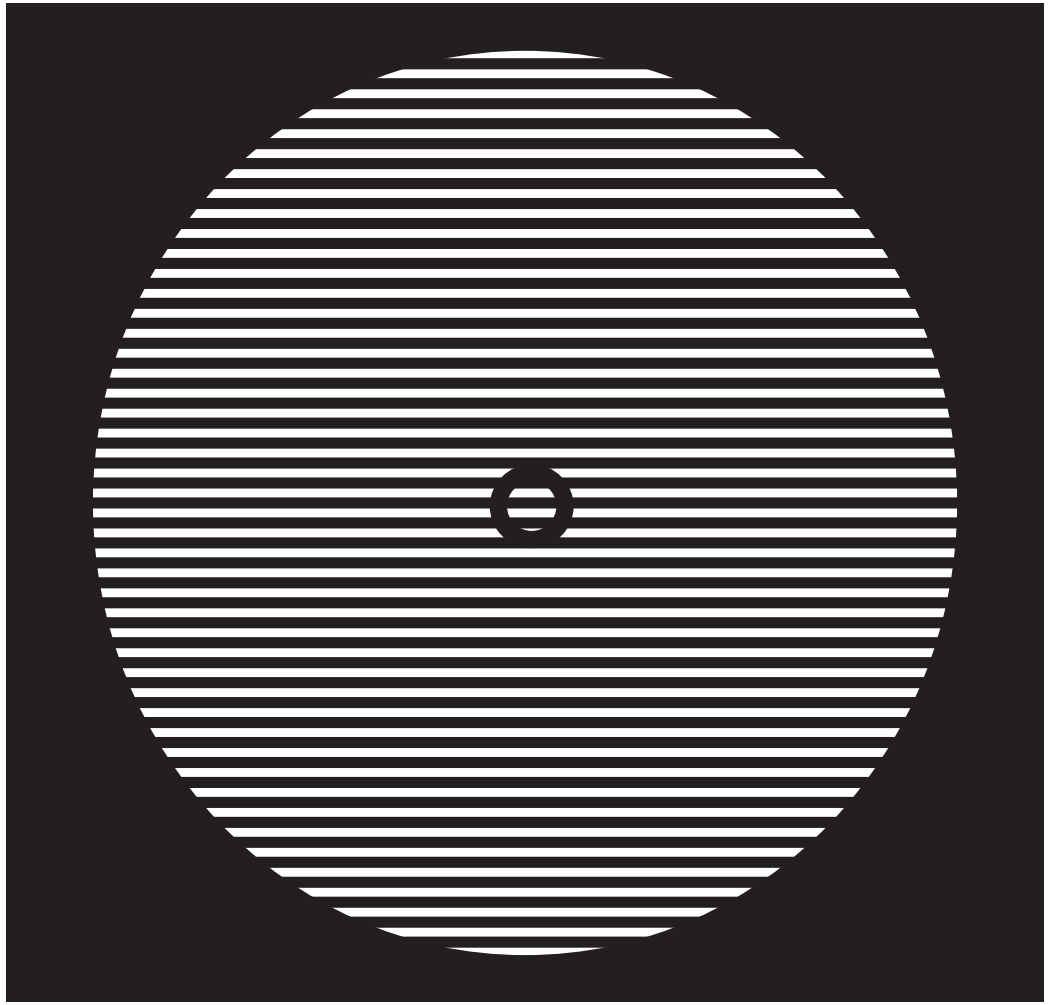
Unintentional optical art (cont.)





bad data graphics

Patterns can cause seizures!



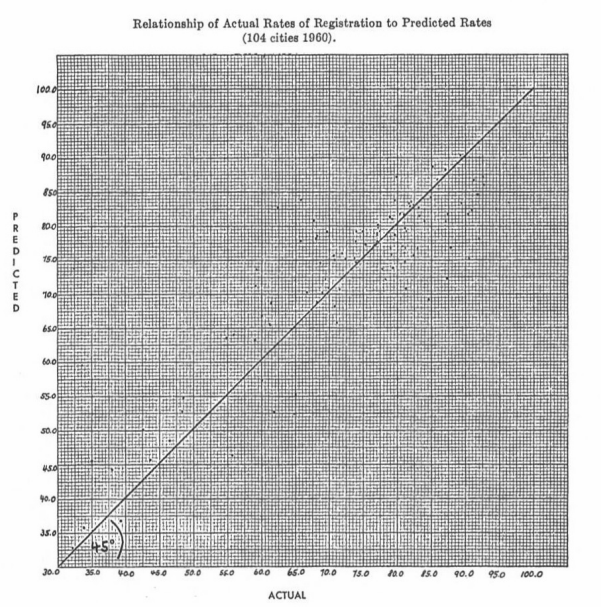
Grids

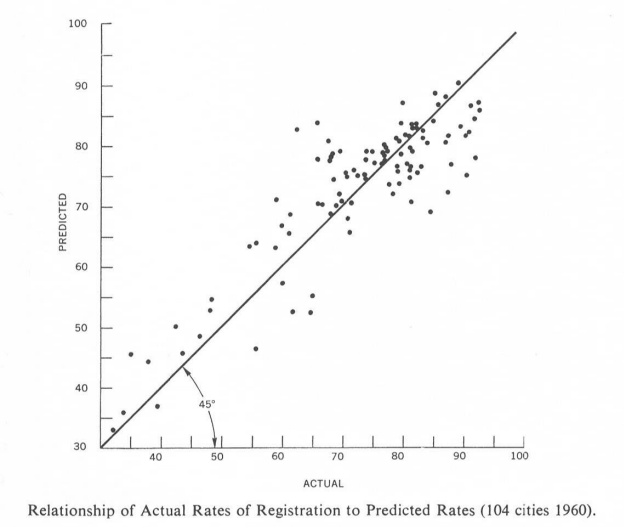
Grids are mostly for the initial plotting of data

- 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) $\rightarrowtail$





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

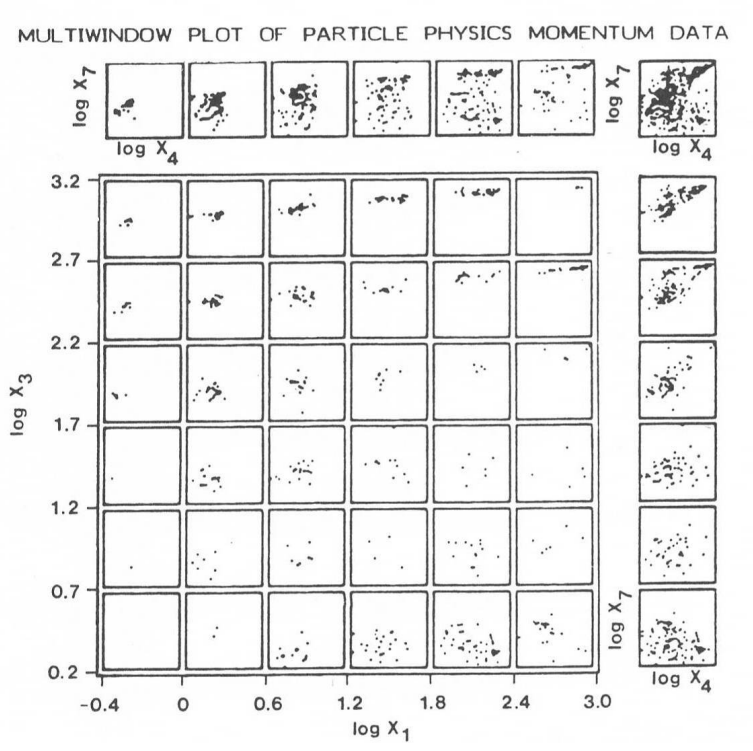
$\rightarrowtail$ Improvements in a republished version of the same data (1970)

Double grid example

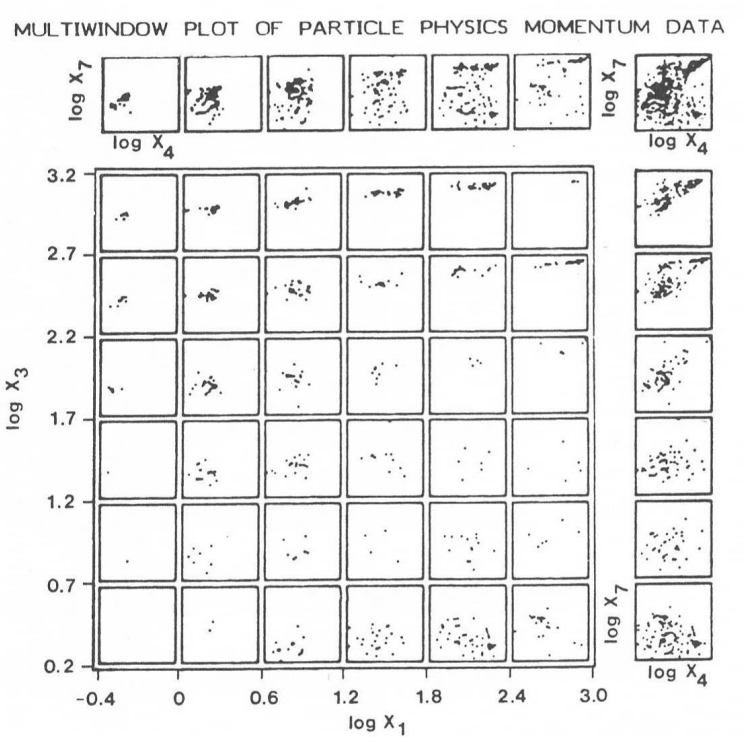
Double grid dominates the graphic and consumes 18% of the area

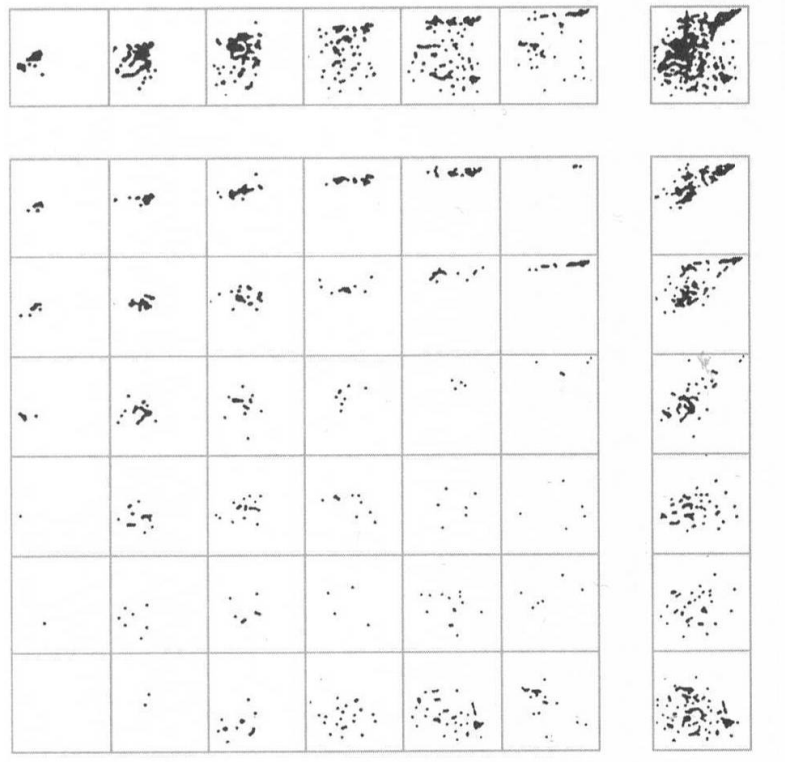
Optical white dots appear at the intersections of the grid lines

- Hermann grid illusion



Double grid example, improved





Removing the double grid improves the graphic considerably

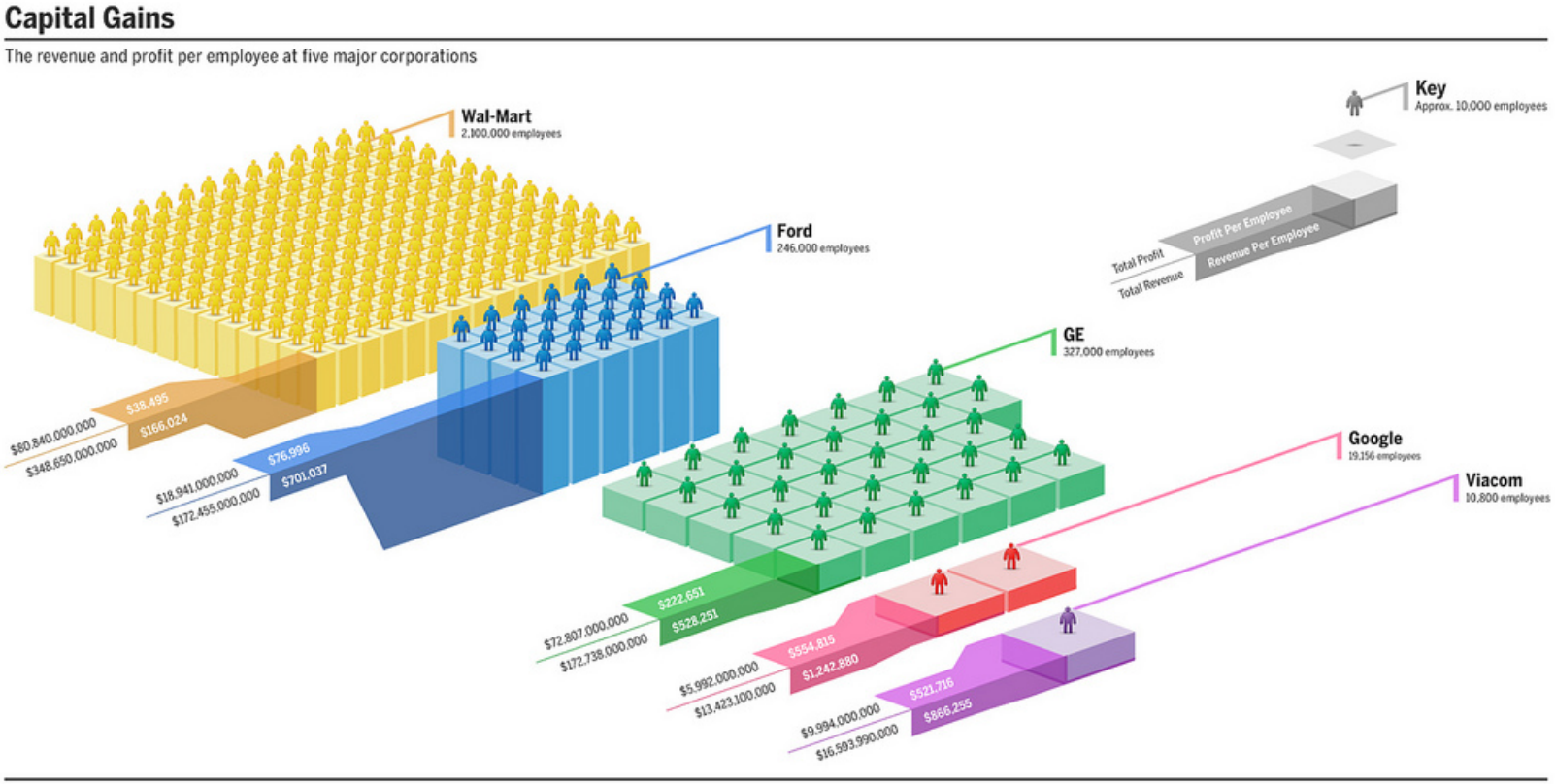
The self-promoting duck

Graphic is a duck if

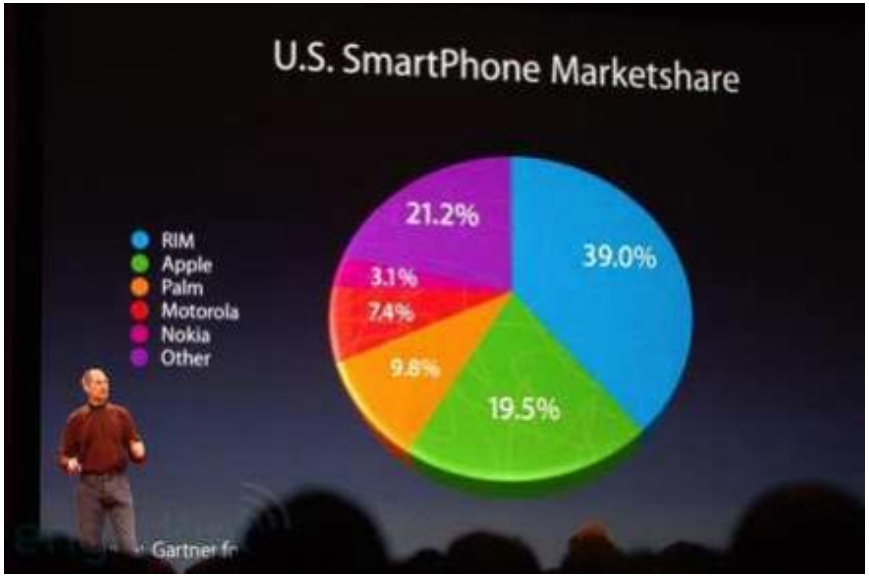
- it is taken over by decorative forms
- data measures and structures become design elements
- the overall design purveys style rather than quantitative information

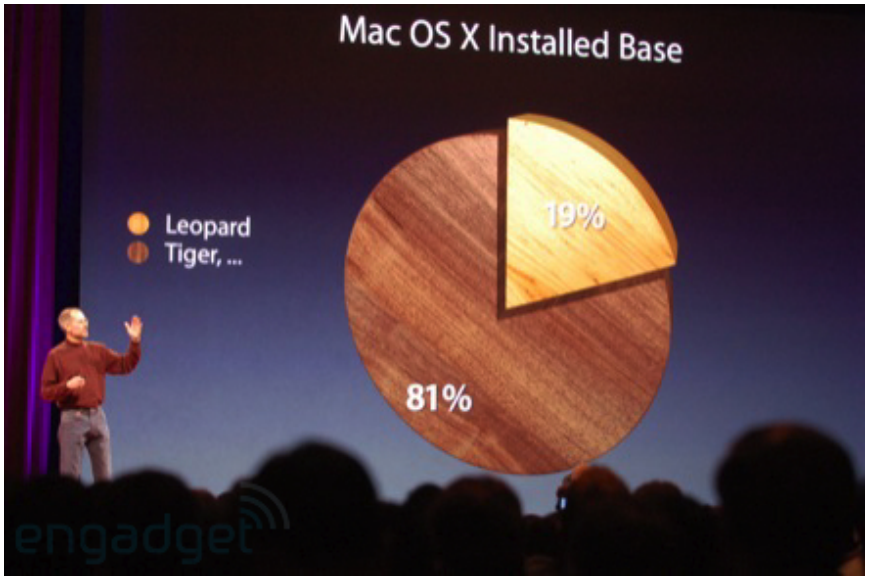
The good magazine

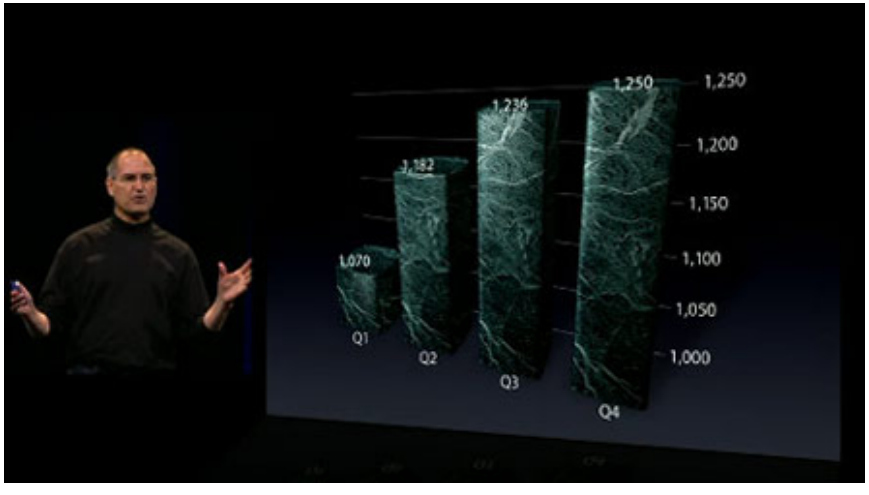
Really?

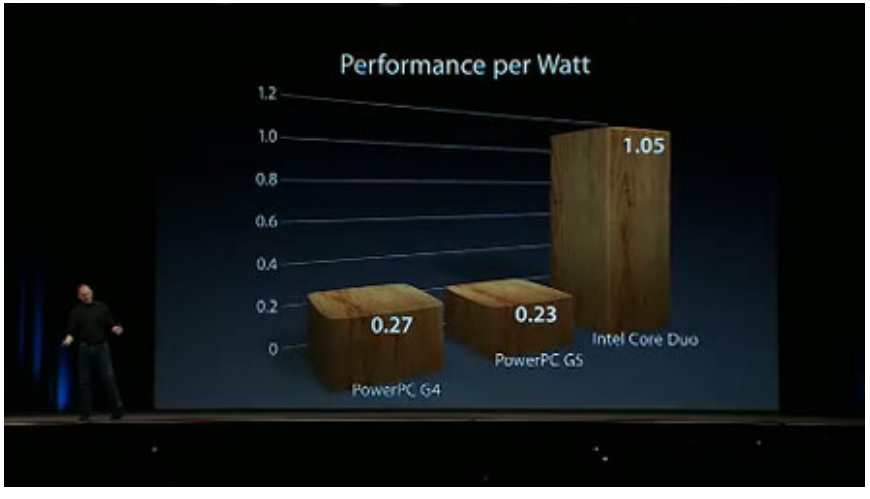

Capital Gains
The revenue and profit per employee at five major corporations
Wal-Mart
2,100,000 employees
$38,495
$166,024
$80,840,000,000
$348,650,000,000
Ford
246,000 employees
$76,996
$701,037
$18,941,000,000
$172,455,000,000
GE
327,000 employees
$222,651
$528,251
$72,807,000,000
$172,738,000,000
Google
19,156 employees
$554,815
$1,242,880
$5,992,000,000
$13,423,100,000
Viacom
10,800 employees
$521,716
$866,255
$9,994,000,000
$16,593,990,000
Key
Approx. 10,000 employees
Profit Per Employee
Revenue Per Employee
Total Profit
Total Revenue

Really??

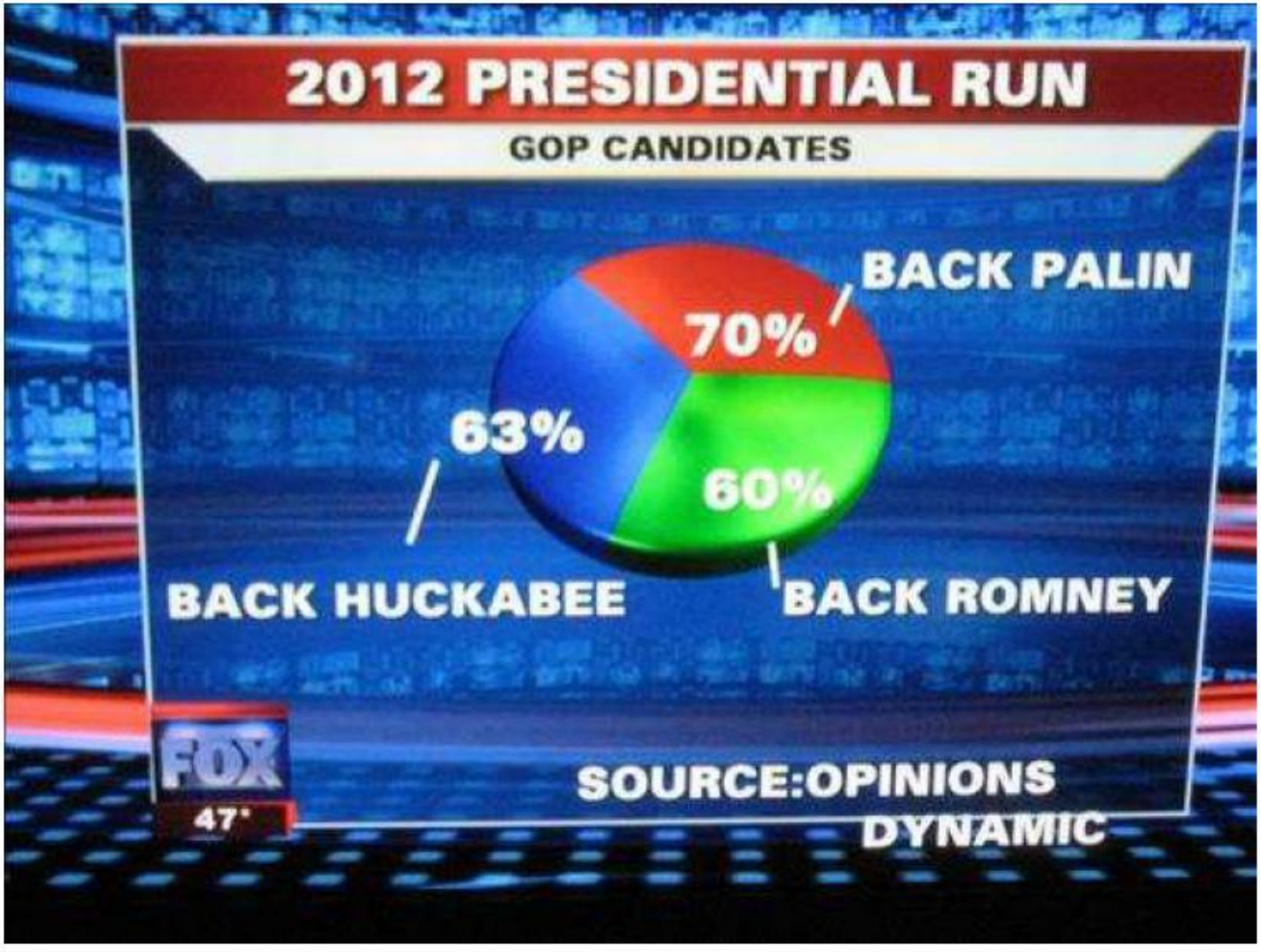

U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
39.0%
19.5%
9.8%
7.4%
3.1%
21.2%


Mac OS X Installed Base
Leopard
Tiger, ...
19%
81%
engadget


1,070
1,182
1,236
1,250
Q1
Q2
Q3
Q4
1,250
1,200
1,150
1,100
1,050
1,000


Performance per Watt
1.2
1.0
0.8
0.6
0.4
0.2
0
0.27
0.23
1.05
PowerPC G4
PowerPC G5
Intel Core Duo

Really???


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

Guidelines

(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:

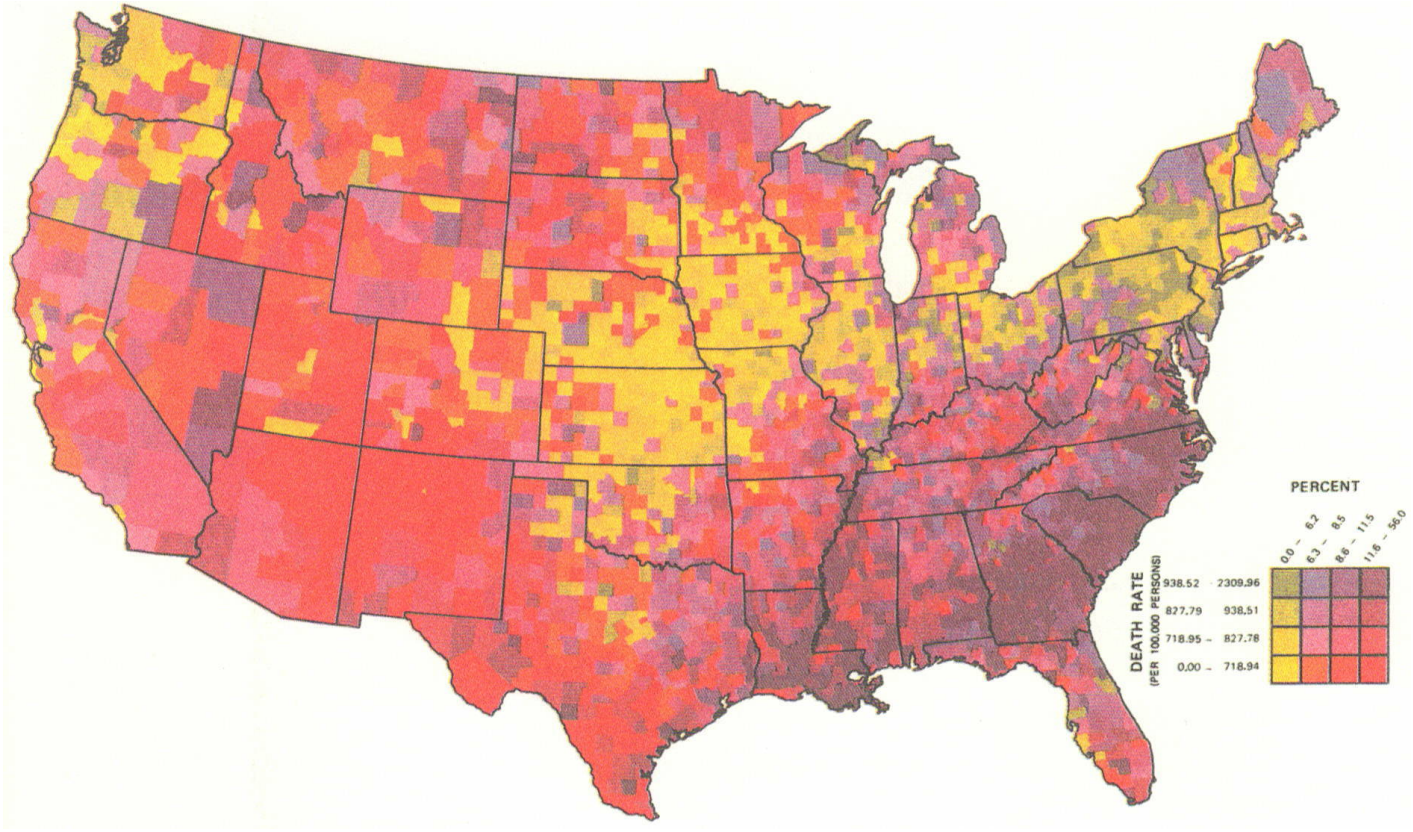
- explain what you are seeing (in caption)
- if needed, explain how you should read the plot
- spell out words, don't abbreviate
- if needed, add helping messages to the plot
- if possible, make words run from left to right: y-axis label.

(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

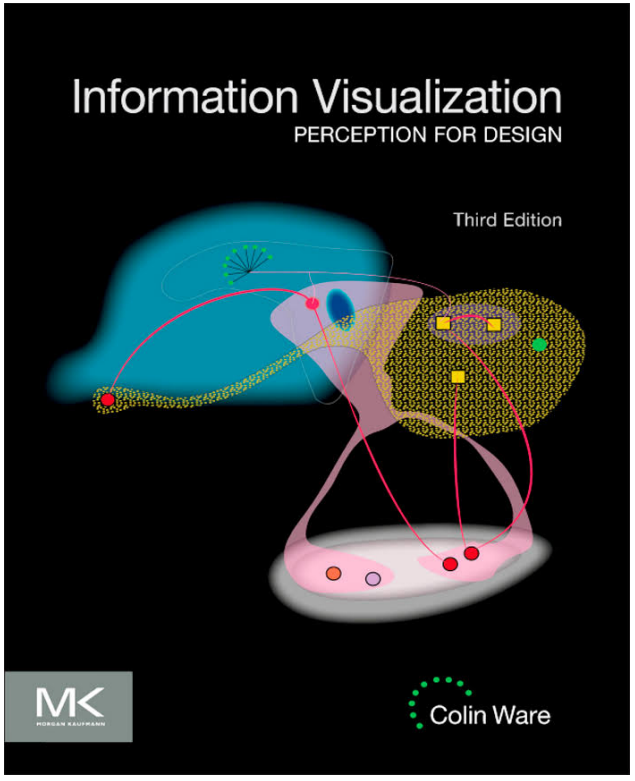
- the overall trend is still readable
- the fonts need to be the correct size
- controlling gaps becomes important

(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 Ware

Information visualization: perception for design



- study of human anatomy, neurology, and psychology
- its effect on visualization design

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

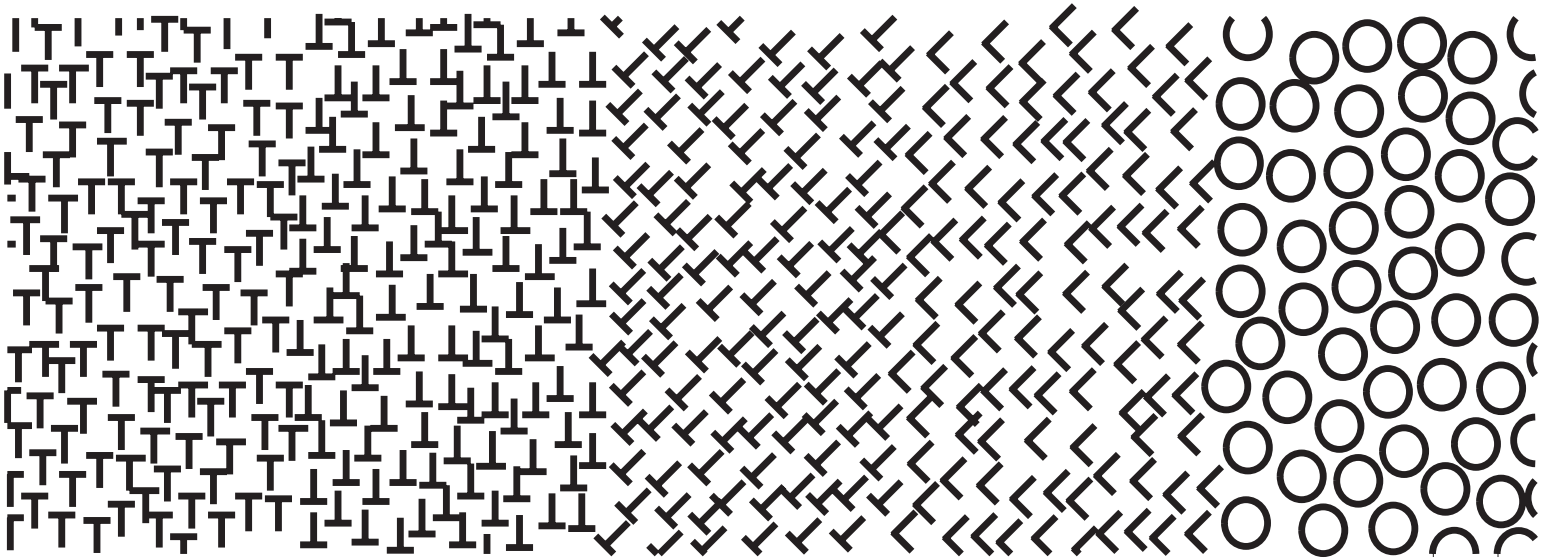
Arbitrary symbols

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

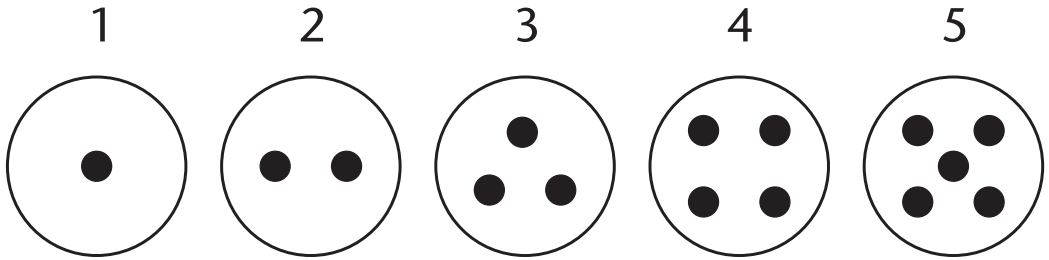
Sensory symbols

- 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

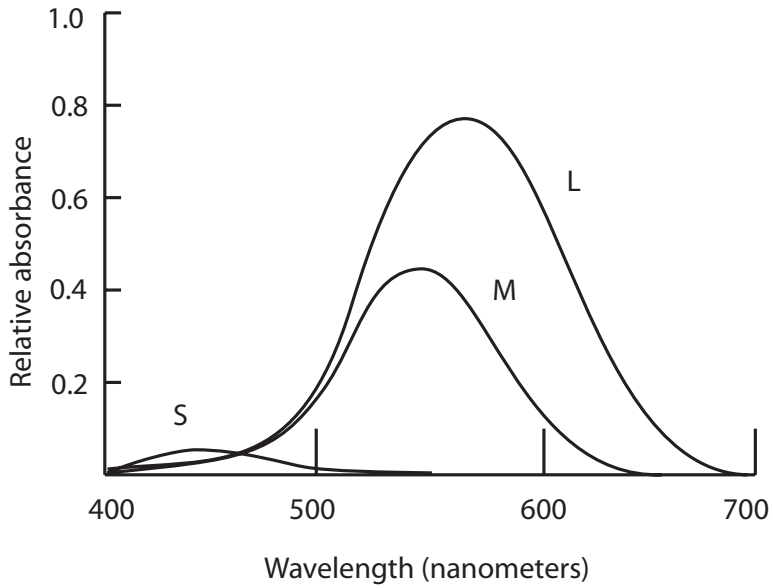
- In eye, the lens focuses an image to retina
- Retina has two types of receptor cells: cones and rods
- cones
  - responsible for normal color vision
  - 6–7 millions
  - concentrated around fovea
- rods
  - responsible for dim-light vision
  - no color information
  - roughly 90 millions
  - concentrated 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:
  - 60° up
  - 70°–75° down
  - 60° nasal
  - 100°–110° temporal

Trichromacy

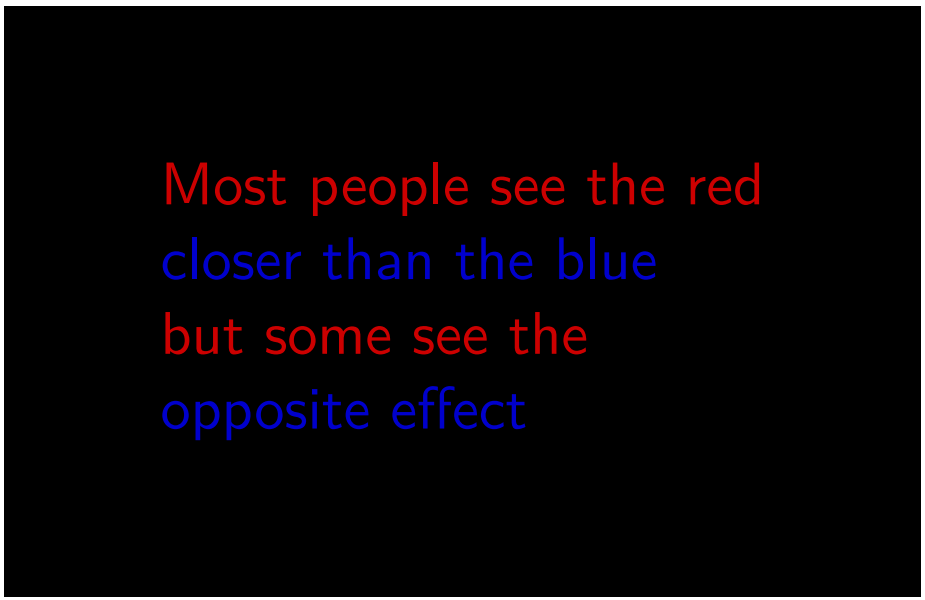
- Human eye have 3 distinct color receptors, cones
- (we also have rods, but they work only in dim conditions)
- Hence, we tend to express colors using 3 primary colors
- Chickens have 12 distinct color receptors



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

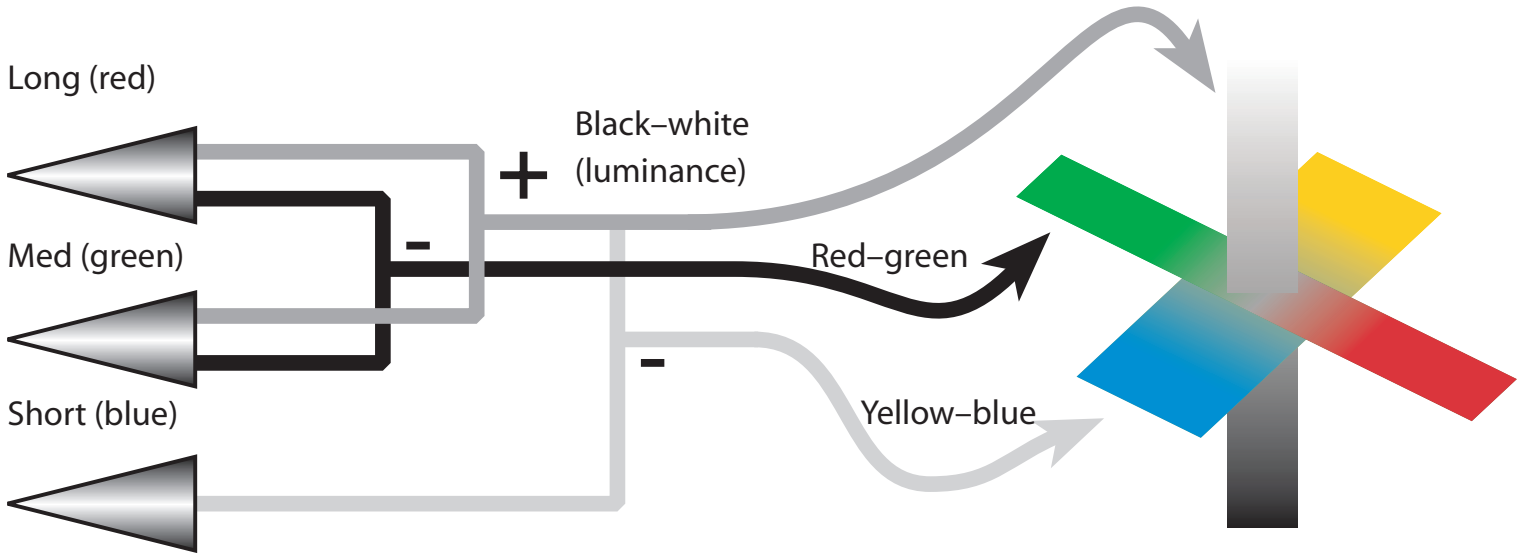
- it is OK to say 'yellowish green' or 'reddish blue'
- nobody says 'reddish green' or 'yellowish blue'

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

Ganglion cells

- inputs from rods and cones, after some processing are handled by ganglion cells.
- at fovea, a ganglion cell takes input from few rods/cones
- at the edge, a cell is responsible for thousands of receptors
- receptive field = area (in retina) used by ganglion cell

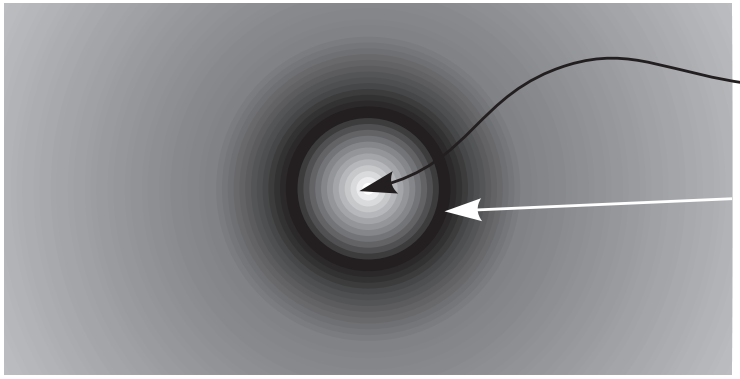
On-center ganglion cell:

- cones/rods excited at the center of the field $\rightarrow$ cell is excited
- cones/rods excited at the edge of the field $\rightarrow$ cell is inhibited
- modelled with a difference of Gaussians (DOG)

$$f(x) = \alpha_1 \exp\left(-\frac{x^2}{\omega_1^2}\right) - \alpha_2 \exp\left(-\frac{x^2}{\omega_2^2}\right),$$

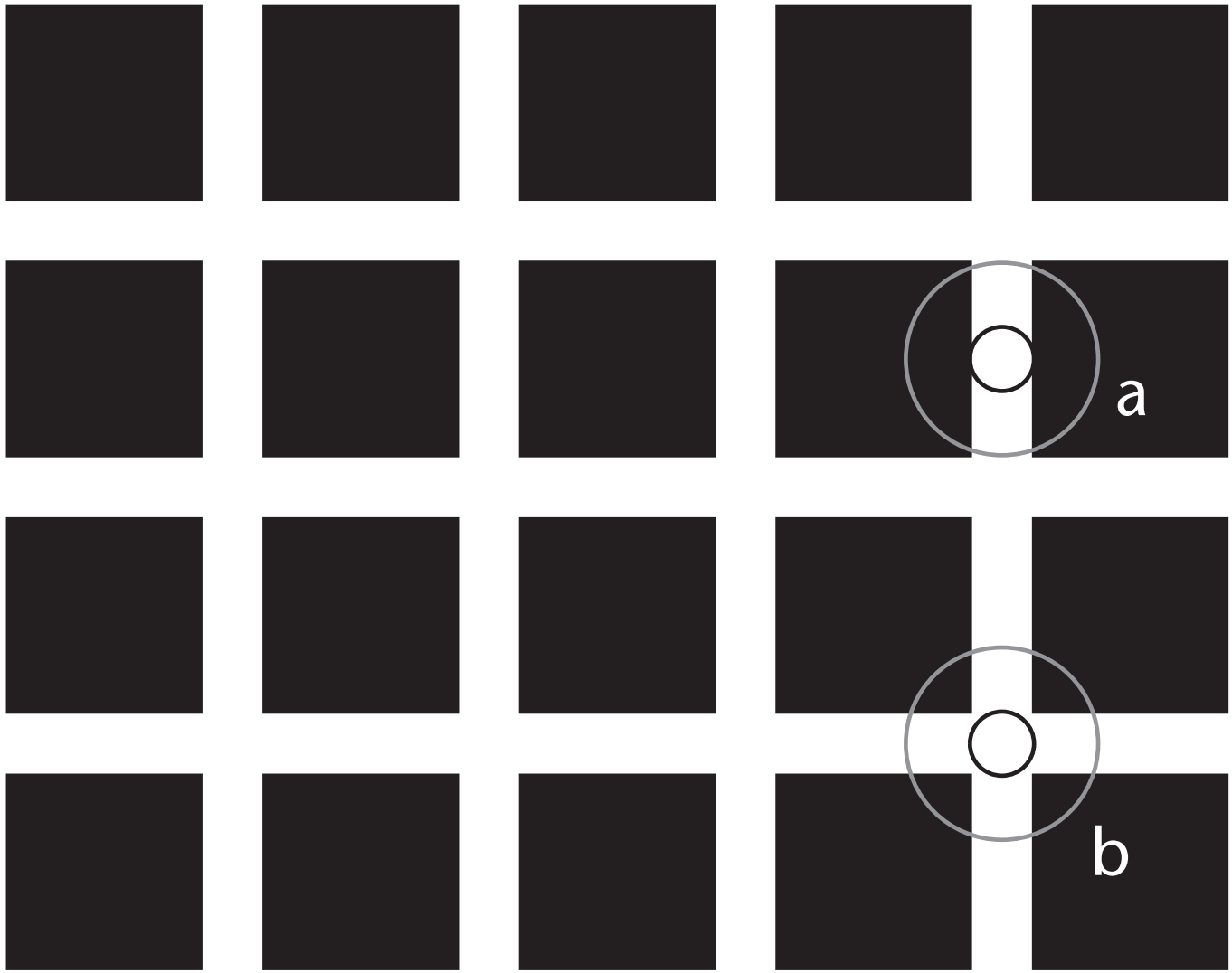
where x is a distance from the center of the receptive field.

DOG model

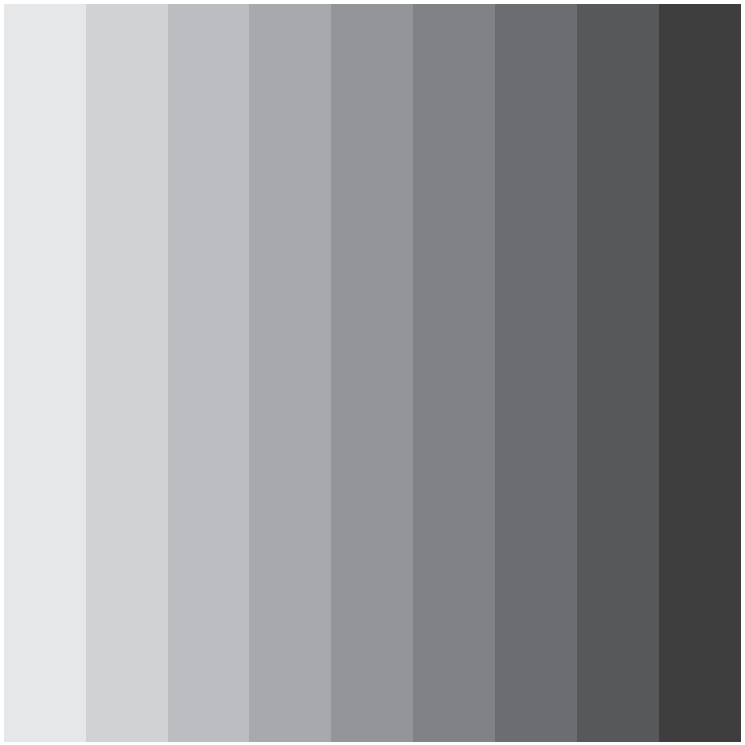


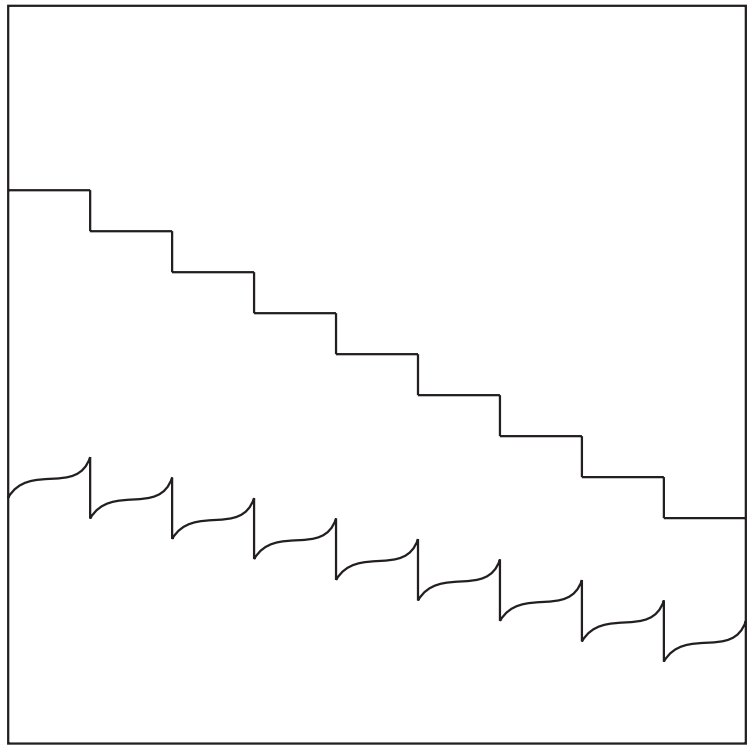
- essentially a hardwired edge detection
- explains many illusions

Hermann grid

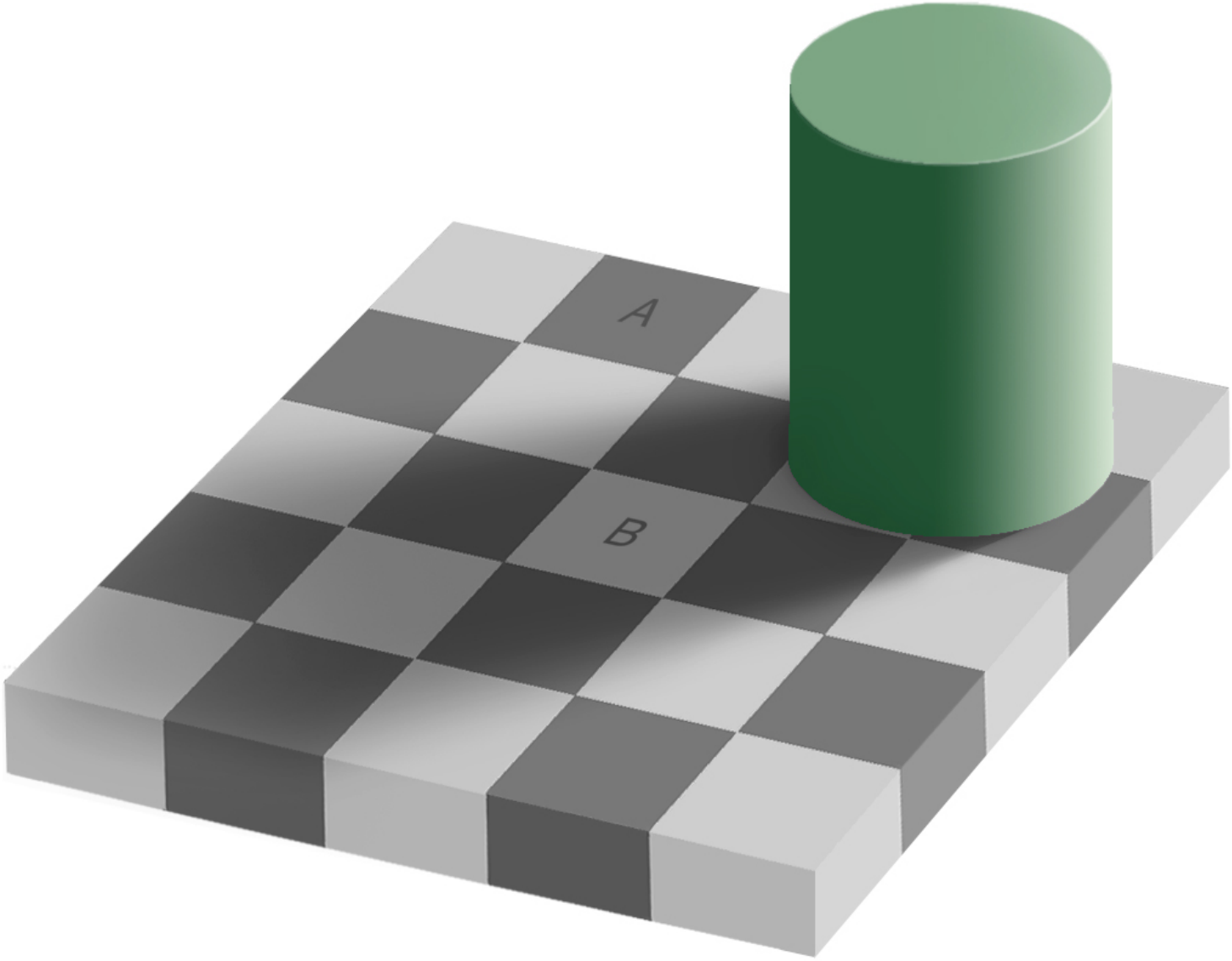


Chevreul illusion

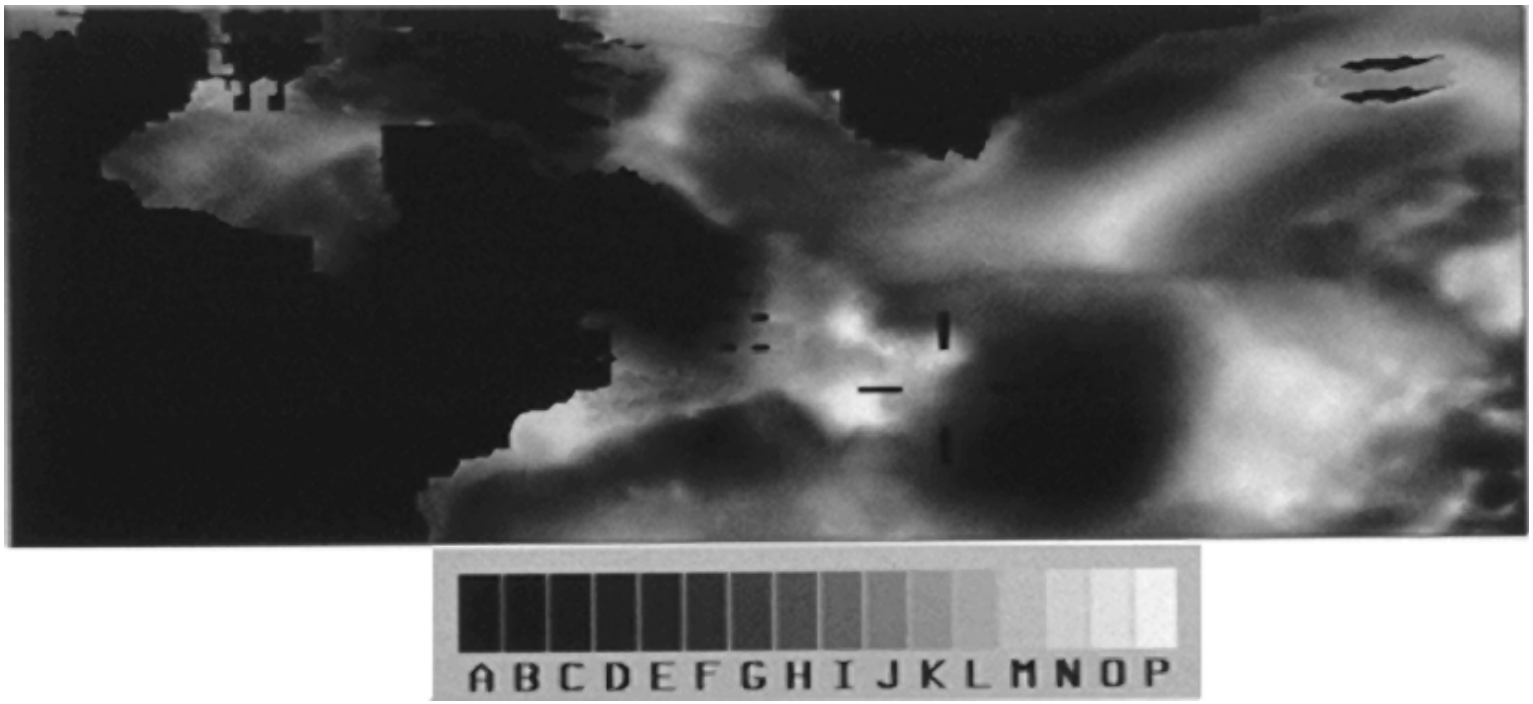




Local contrast illusion



Local contrast illusion



Errors occur when reading the map

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

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

Acuity

Visual acuity is the measurement to see detail

- for example, separate the two lines

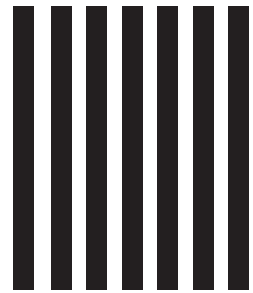
Expressed in visual angles

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

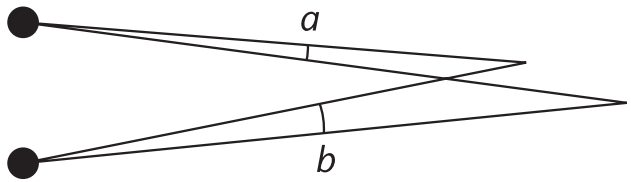
Normally, the best you get is 1 arcminute

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

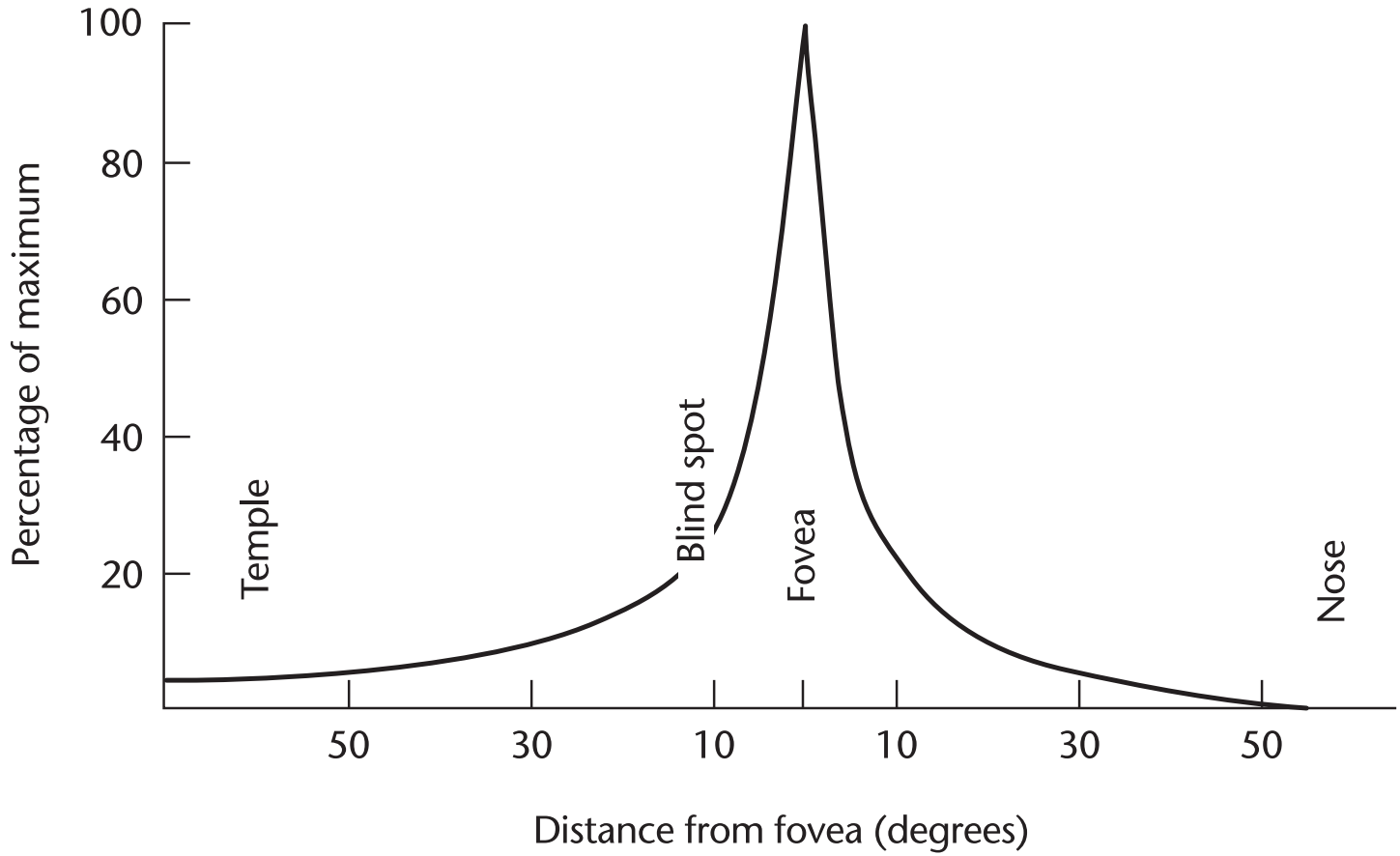
Acuities

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

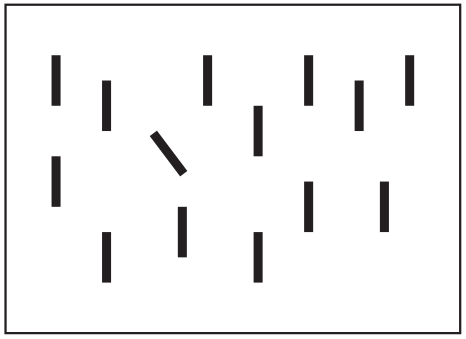
- flicker
- direction of motion

Spatial position

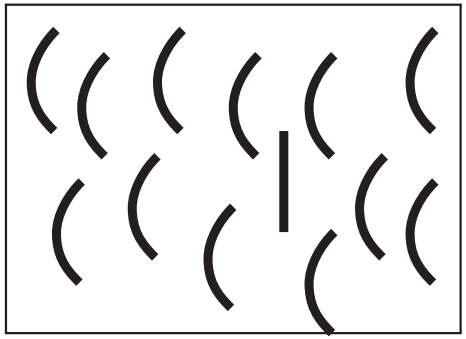
- 2D position
- Stereoscopic depth
- convex/concave shade

Examples of preattentive features

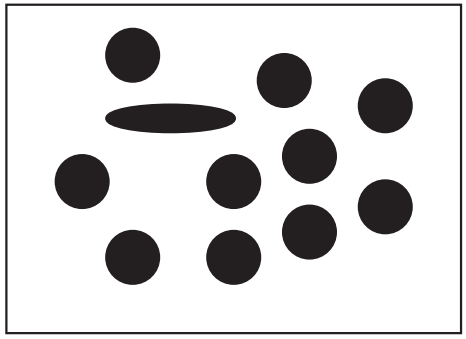
Orientation



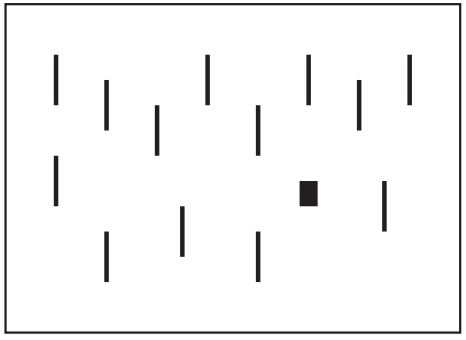
Curved/straight



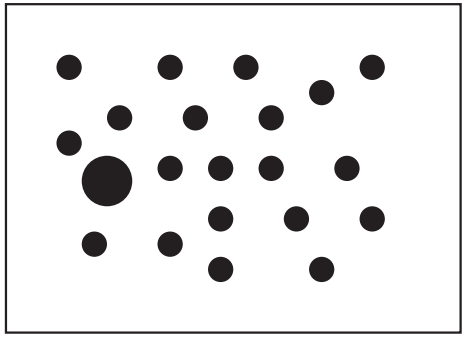
Shape



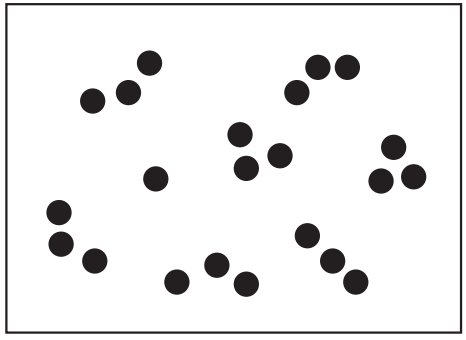
Shape



Size

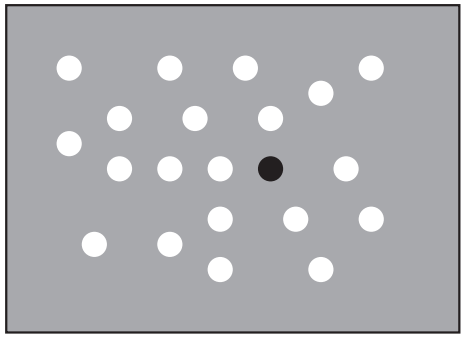


Number

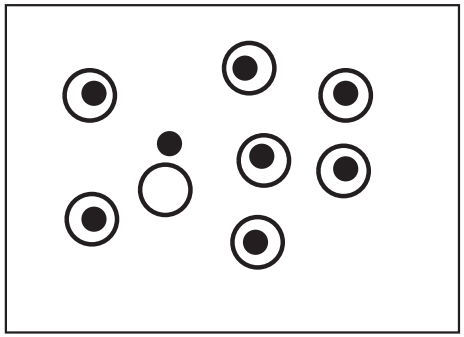


Examples of preattentive features

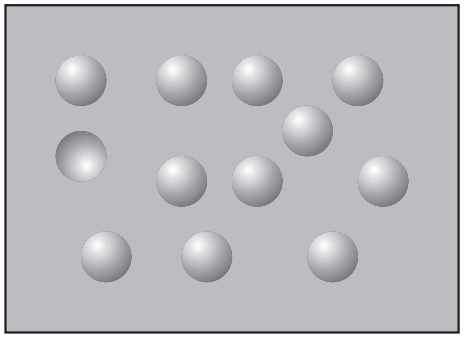
Gray/value



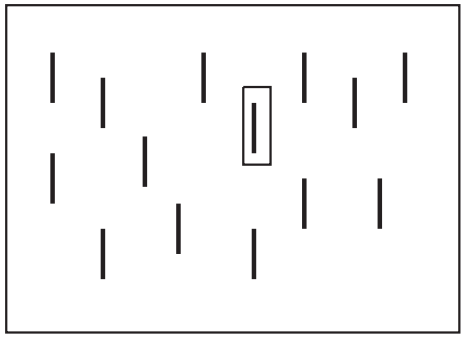
Enclosure



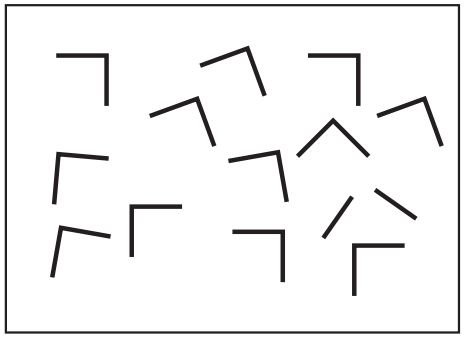
Convexity/concavity



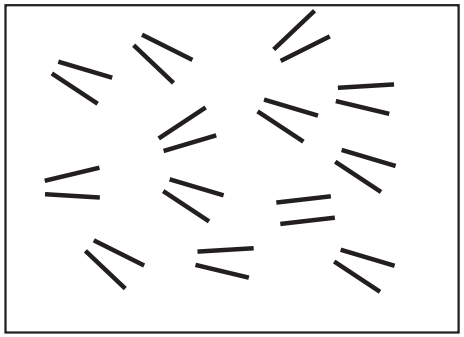
Addition



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

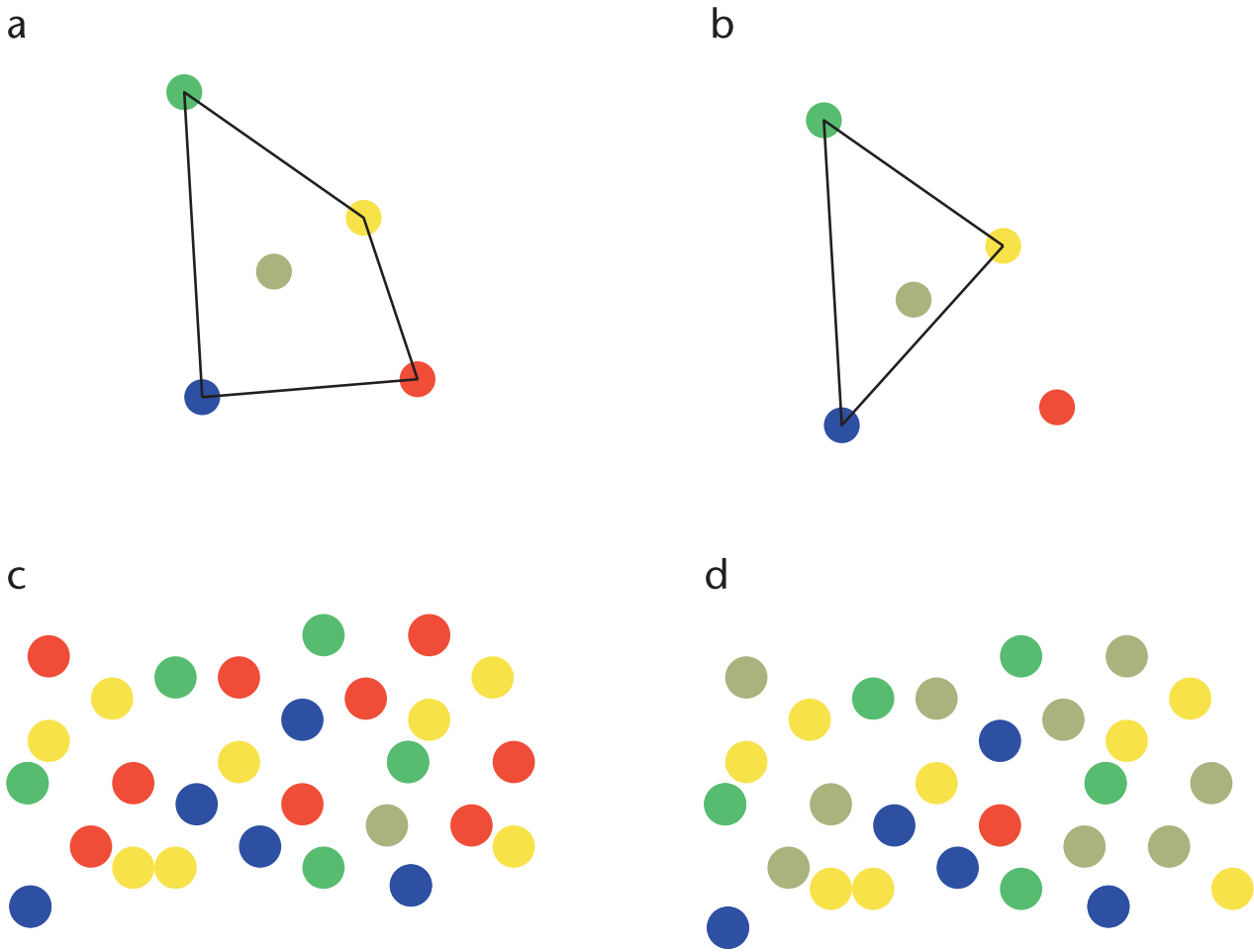
- the saturation
- size of the color patch
- difference from the surrounding colors
- length of the line
- difference degree
- contrast of the line pattern

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

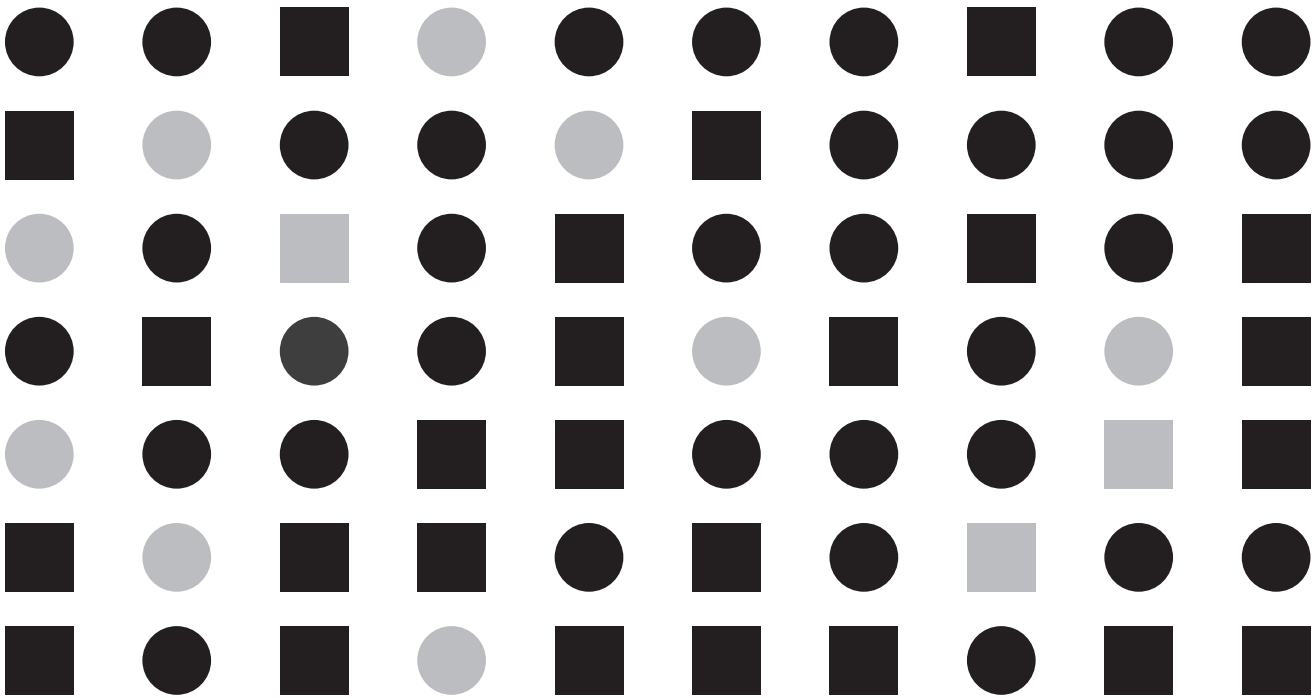


Grey is more difficult to locate in (c) than red in (d)

Conjunctive preattentive features

Is it possible to use preattentive features for complex queries?

Generally speaking, no

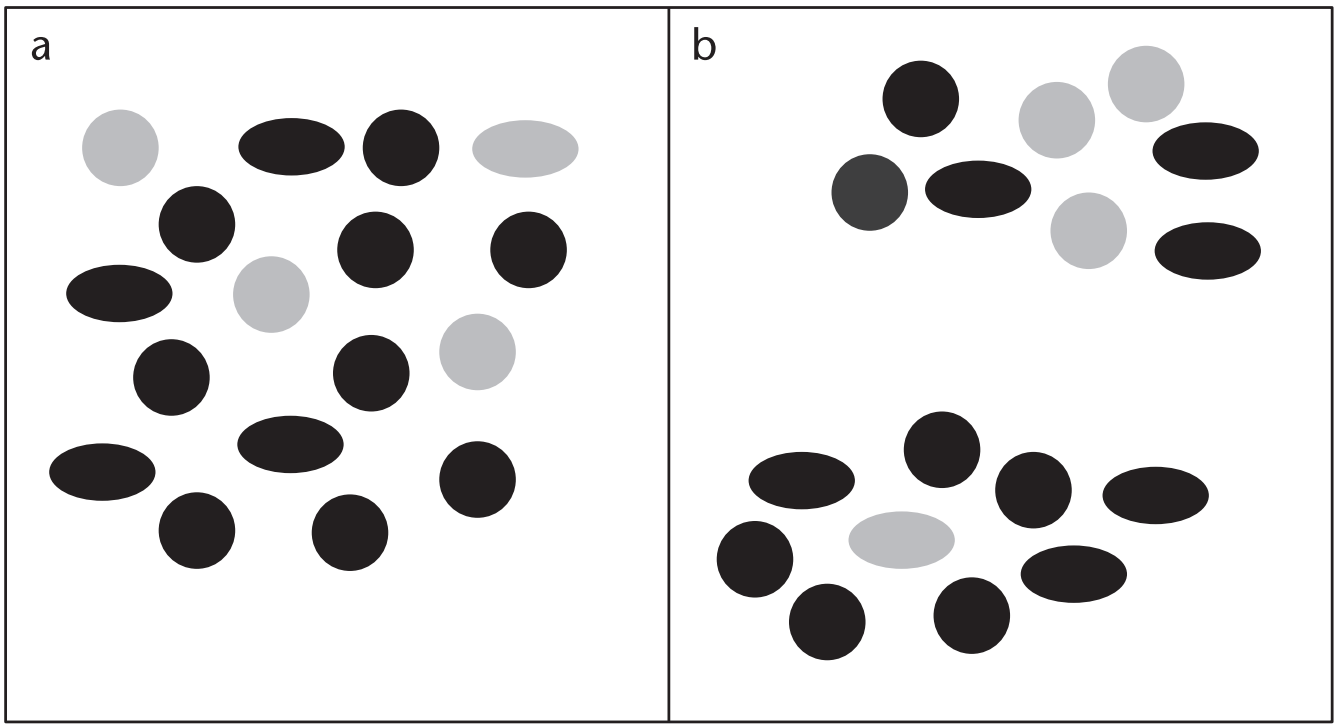


Finding grey squares is slow

Conjunctive preattentive features

In some special cases, it is possible to do complex queries:

- spatial grouping on the XY plane
- stereoscopic depth
- shade convexity/concavity and color
- motion

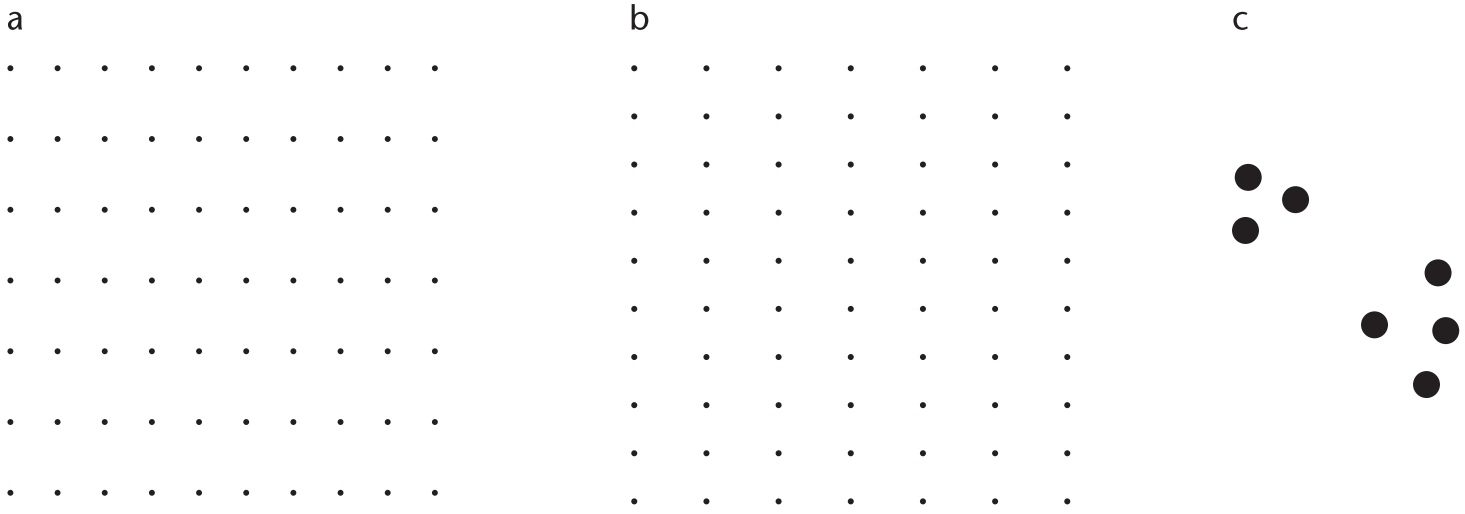


Gestalt laws

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

Gestalt law of proximity

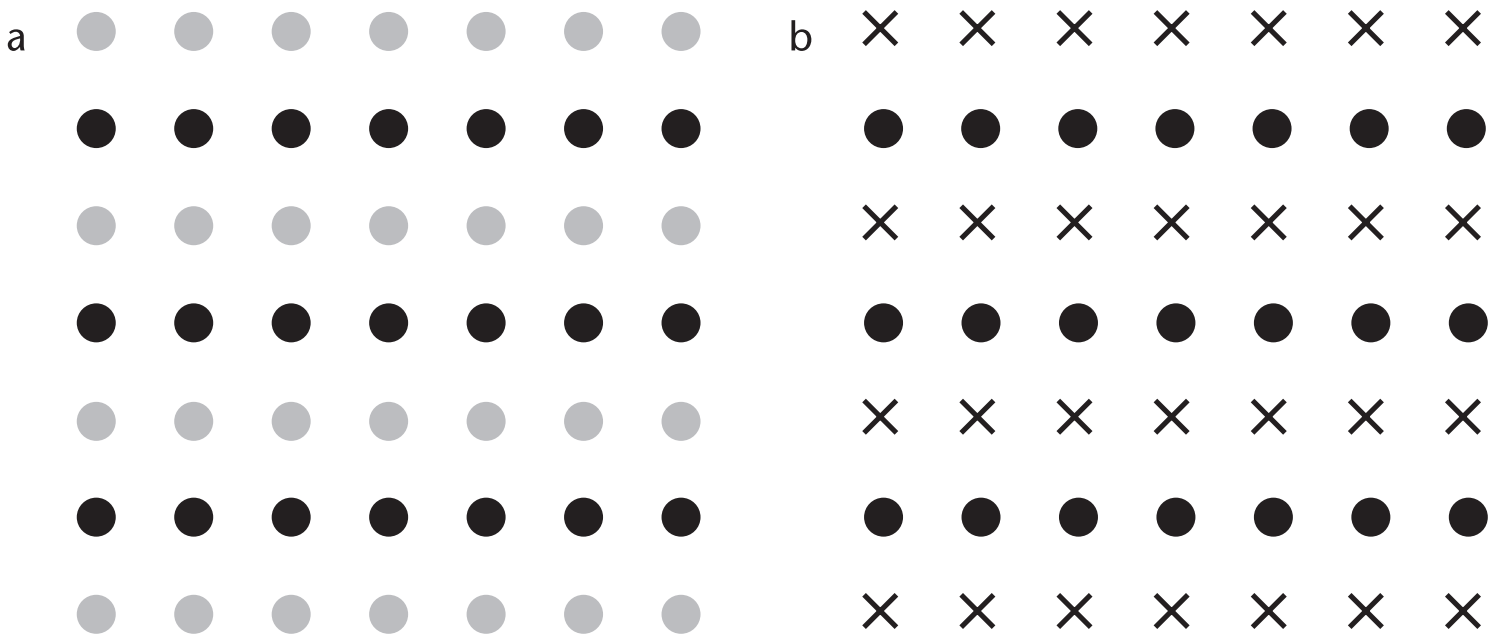
Things that are close are grouped together



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

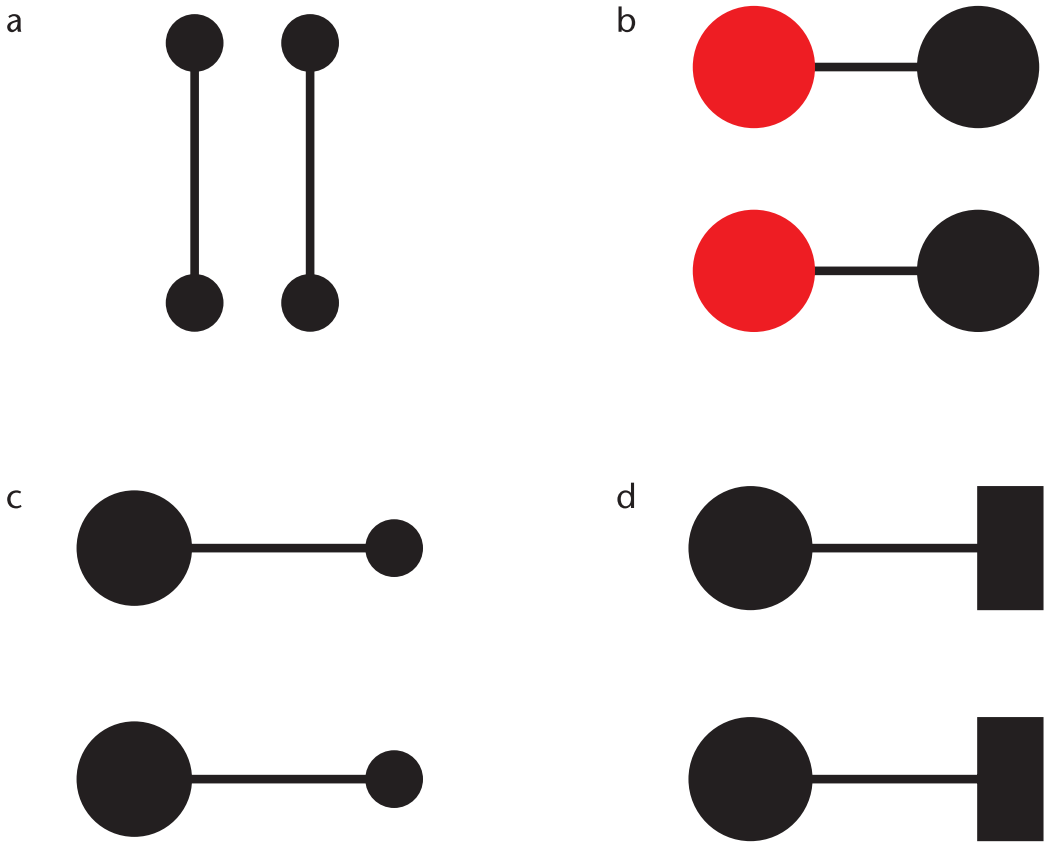
Gestalt law of similarity

Similar shapes are grouped together



Gestalt law of connectedness

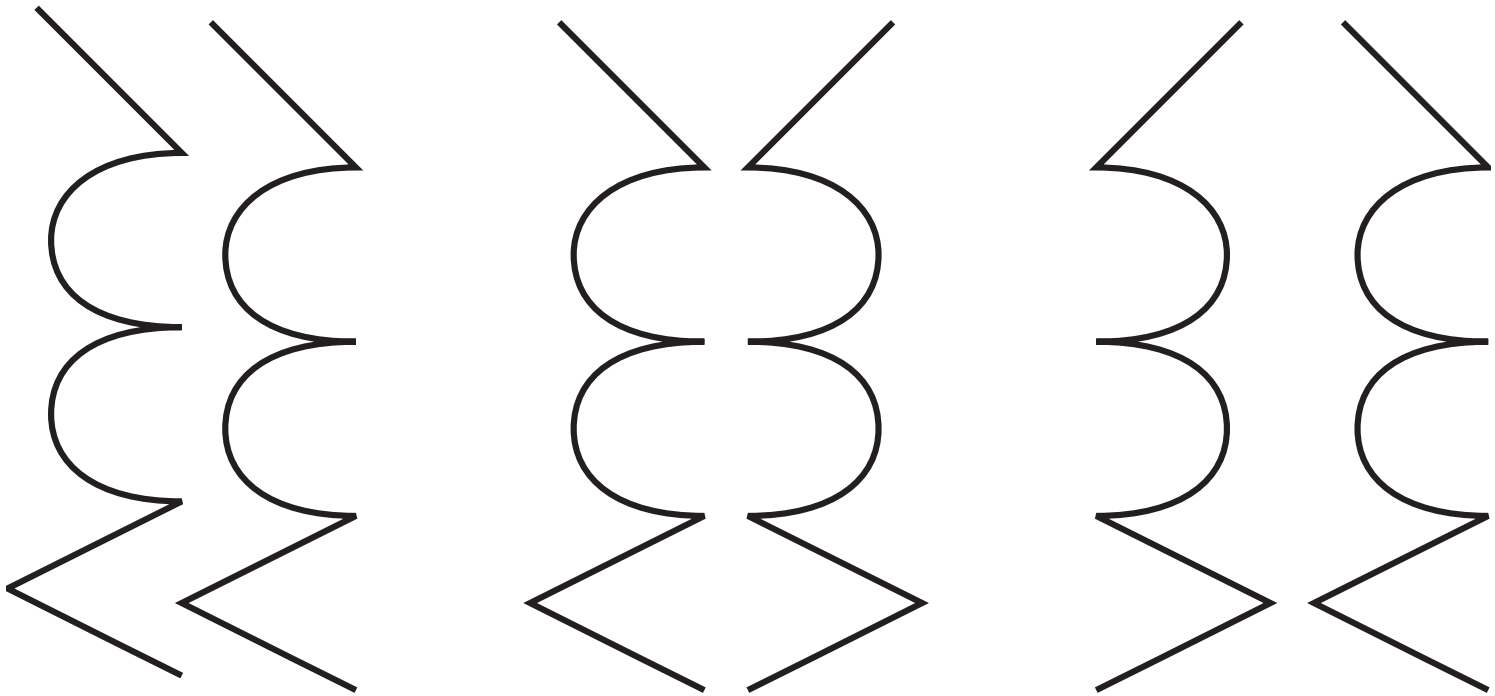
Connected shapes are grouped together



Connectedness is stronger than proximity or similarity

Gestalt law of symmetry

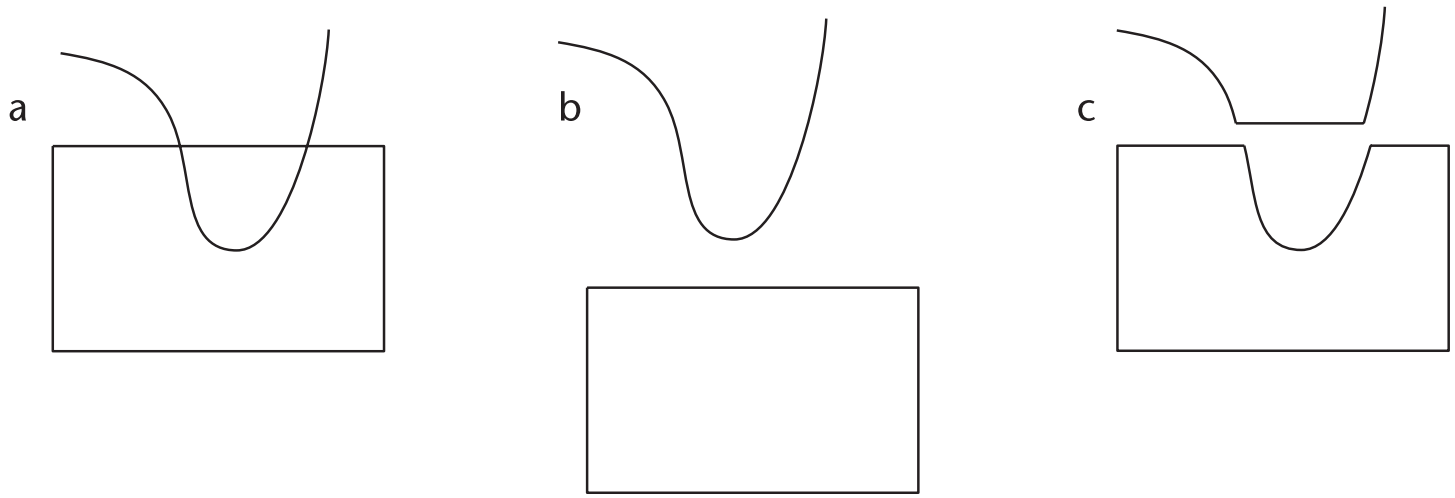
Symmetric shapes are interpret as a whole



left show two separate figures, center and right are whole

Gestalt law of continuity

We construct visual entities from smooth and continuous shapes

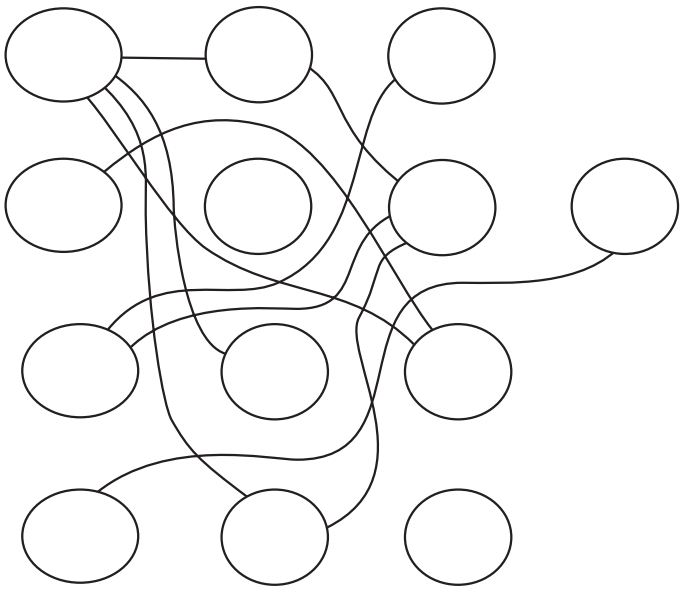


We interpret (a) as (b) rather than (c)

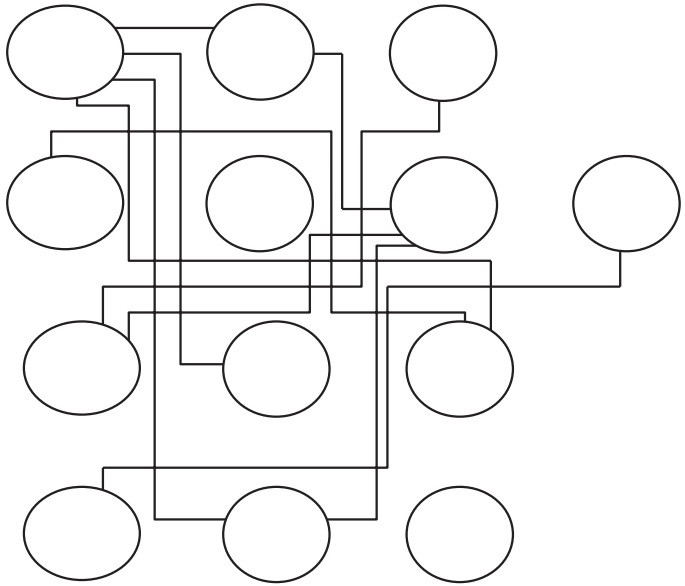
Gestalt law of continuity

Left is easier to read than right

a

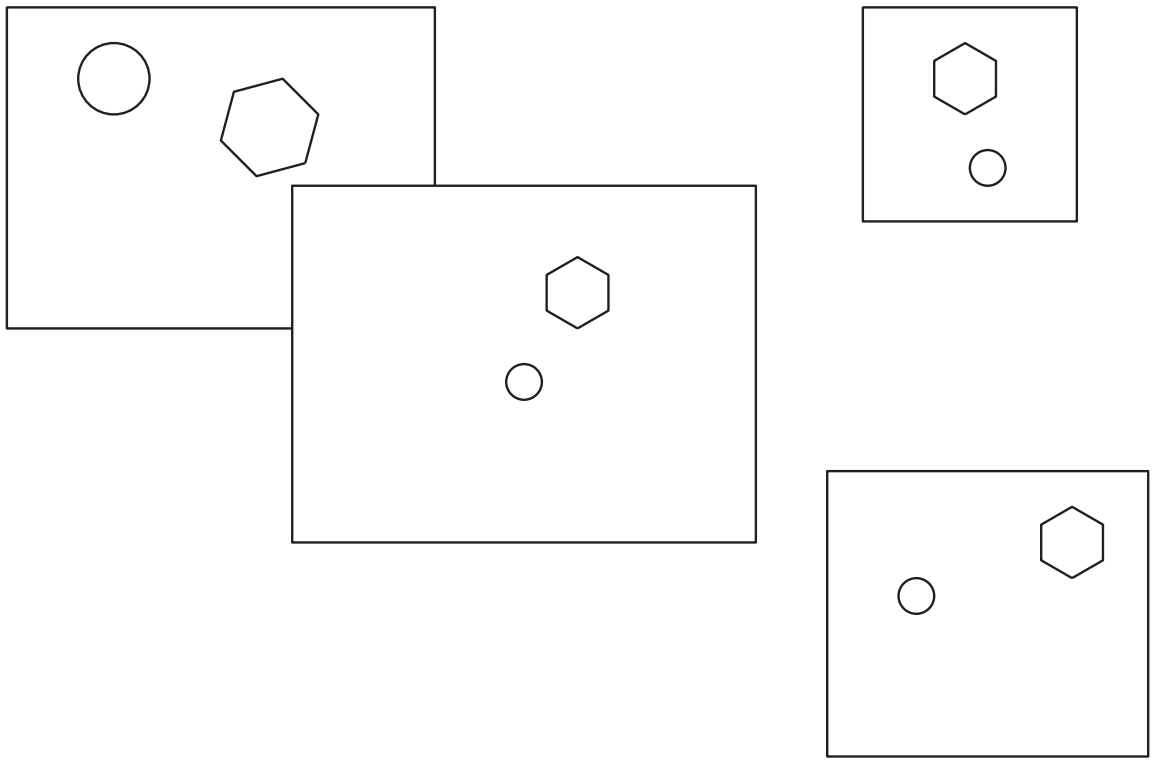


b



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.

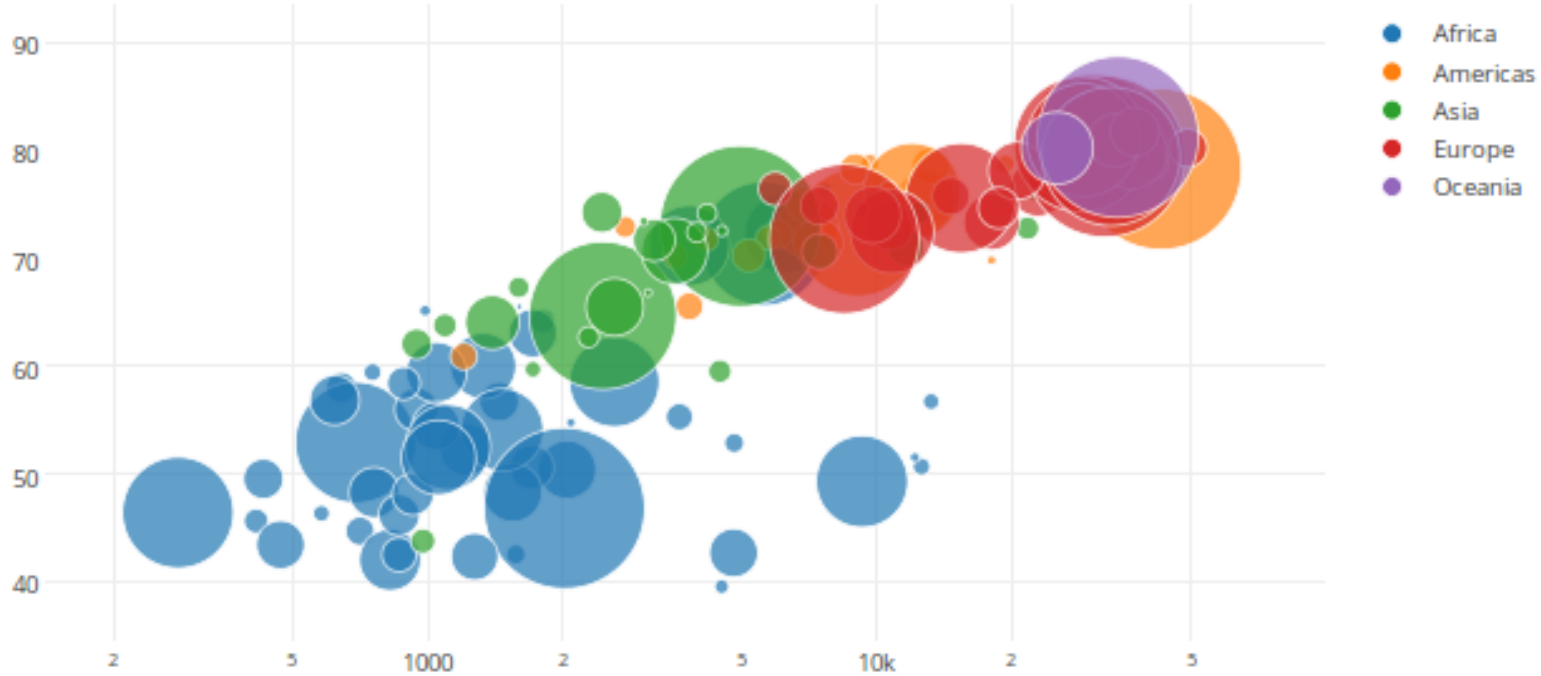
- similar elements should be placed closely
- similar elements should have similar shapes
- connectedness law leads to node-link diagrams
- connectedness law leads also to connecting dots in a plot
- continuity law leads to have smooth links
- closure allows to create groups

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.

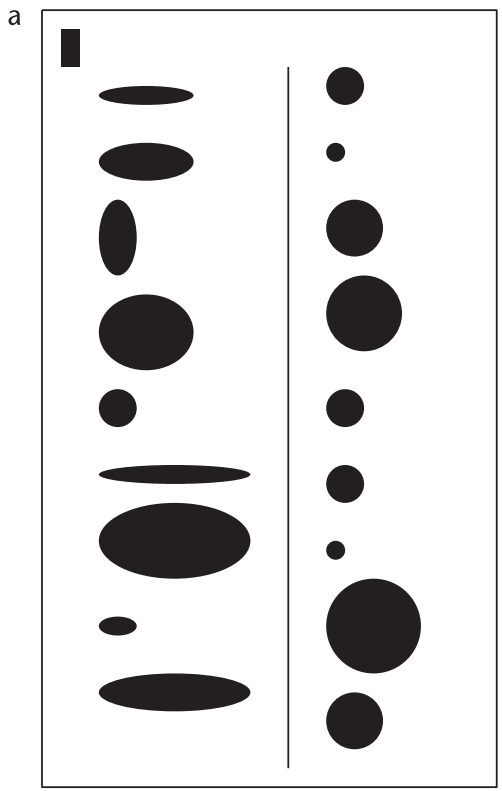


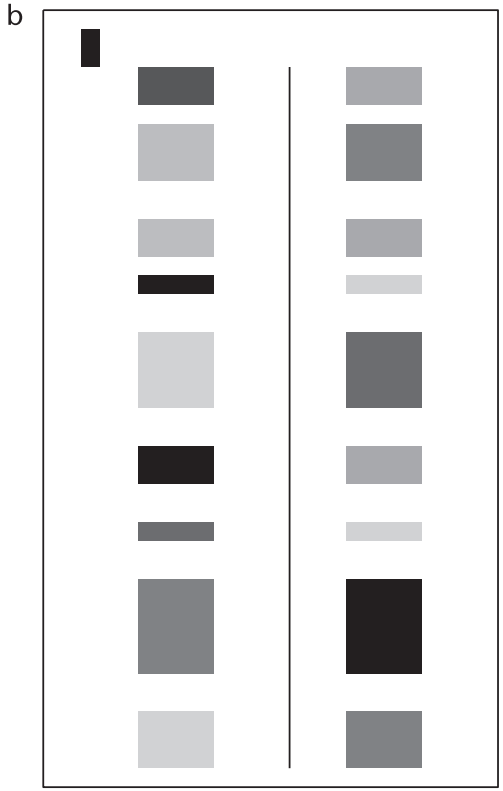
4 dimensions: coordinates, size, color

Integral vs. separable dimension pairs

Combining different visual dimensions is tricky:

- integral dimension pairs are interpreted as whole
- separate dimension pairs are interpreted separately

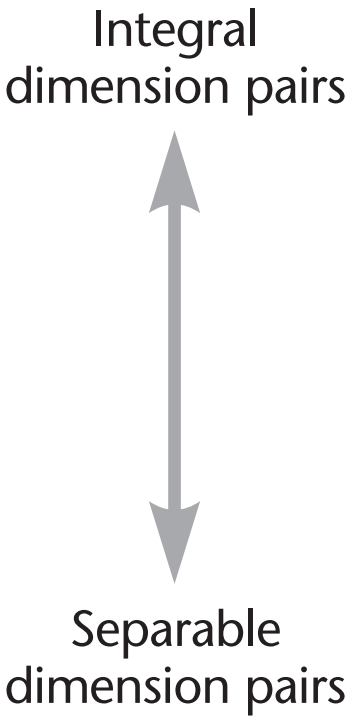




Find the same glyphs that have the same height as the top-left bar.

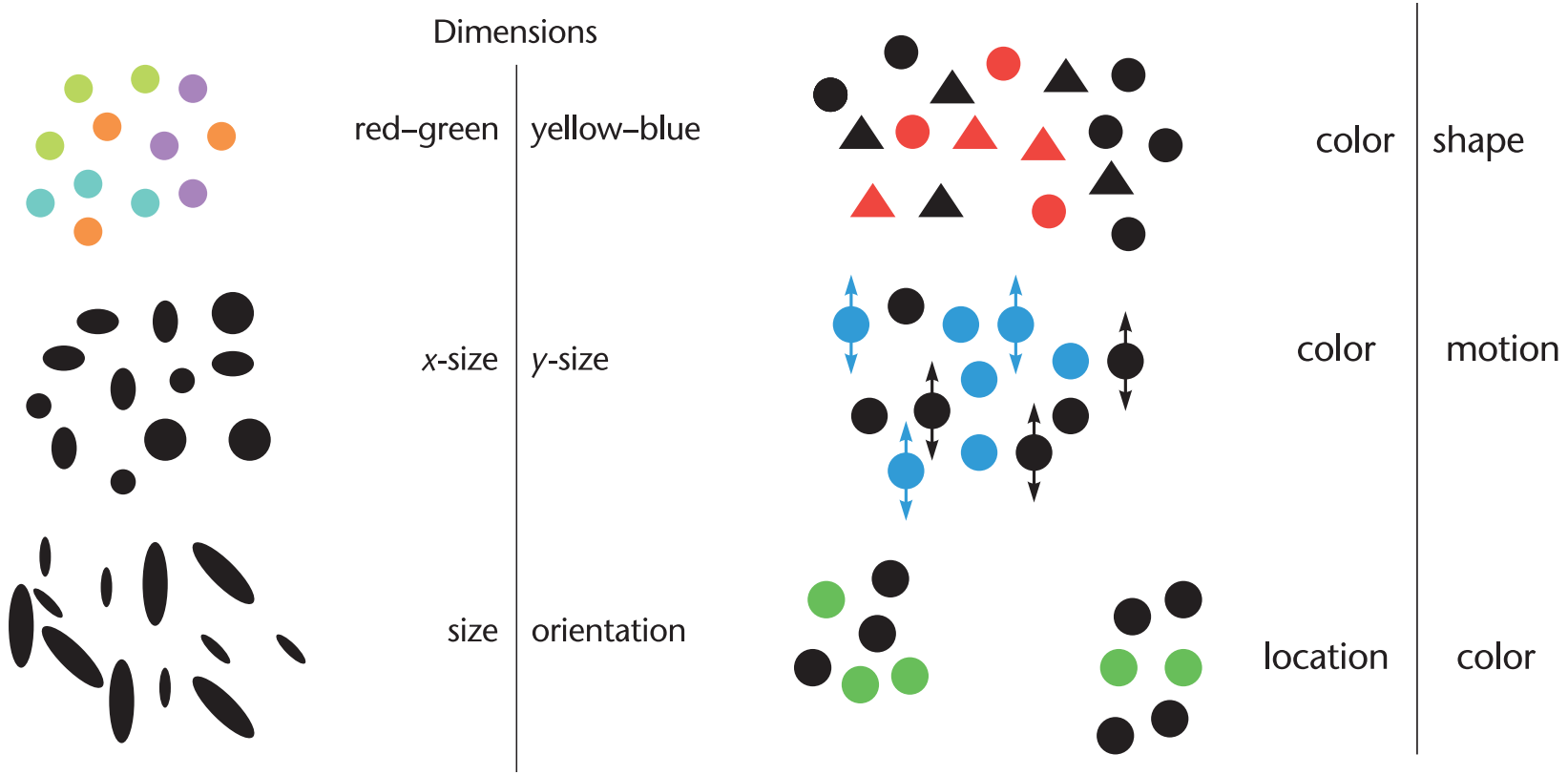
(a) integral pairs
(b) separate pairs

Integral vs. separable dimension pairs



red–green	yellow–blue
red–green	black–white
shape height	shape width
shape	size
color	size
direction of motion	shape
color	shape
color	direction of motion
x,y position	size, shape, or color

Integral vs. separable dimension pairs



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)

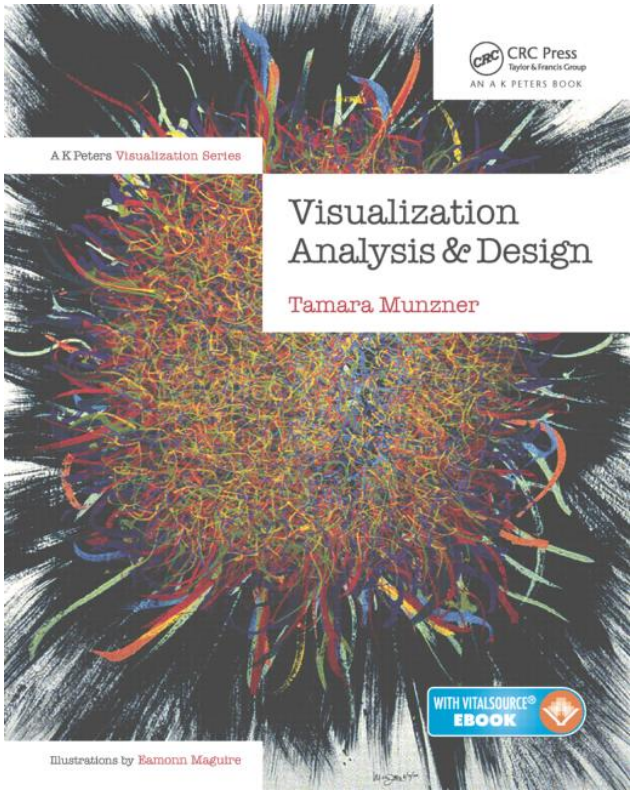
- high dependency on motion

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 Munzner

Visualization analysis and design



- modern take on the visualization in general
- we will go only through a subset of this framework

Marks and channels

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

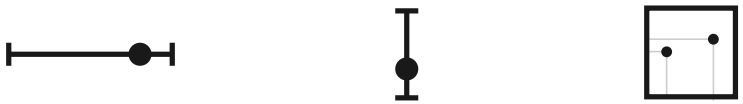
Examples of marks

Points **Lines** **Areas**

Examples of visual channels

Position

- Horizontal
- Vertical
- Both



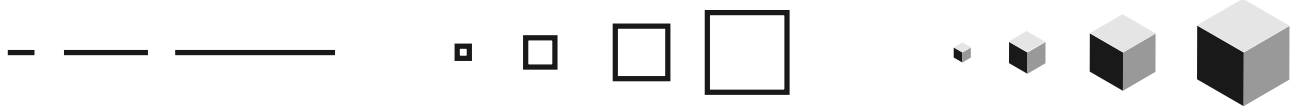
Color

Shape

Tilt

Size

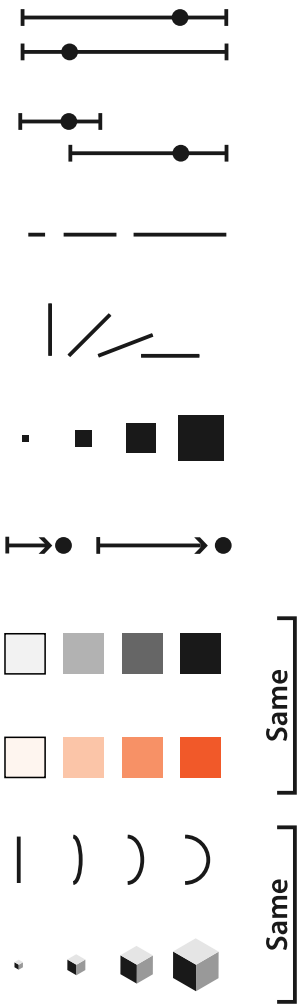
- Length
- Area
- Volume



Effectiveness of visual channels

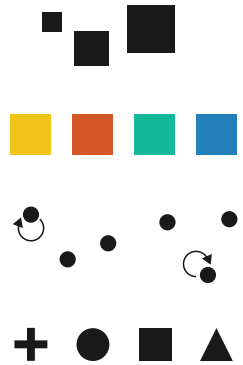
Magnitude Channels: Ordered Attributes

- 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)

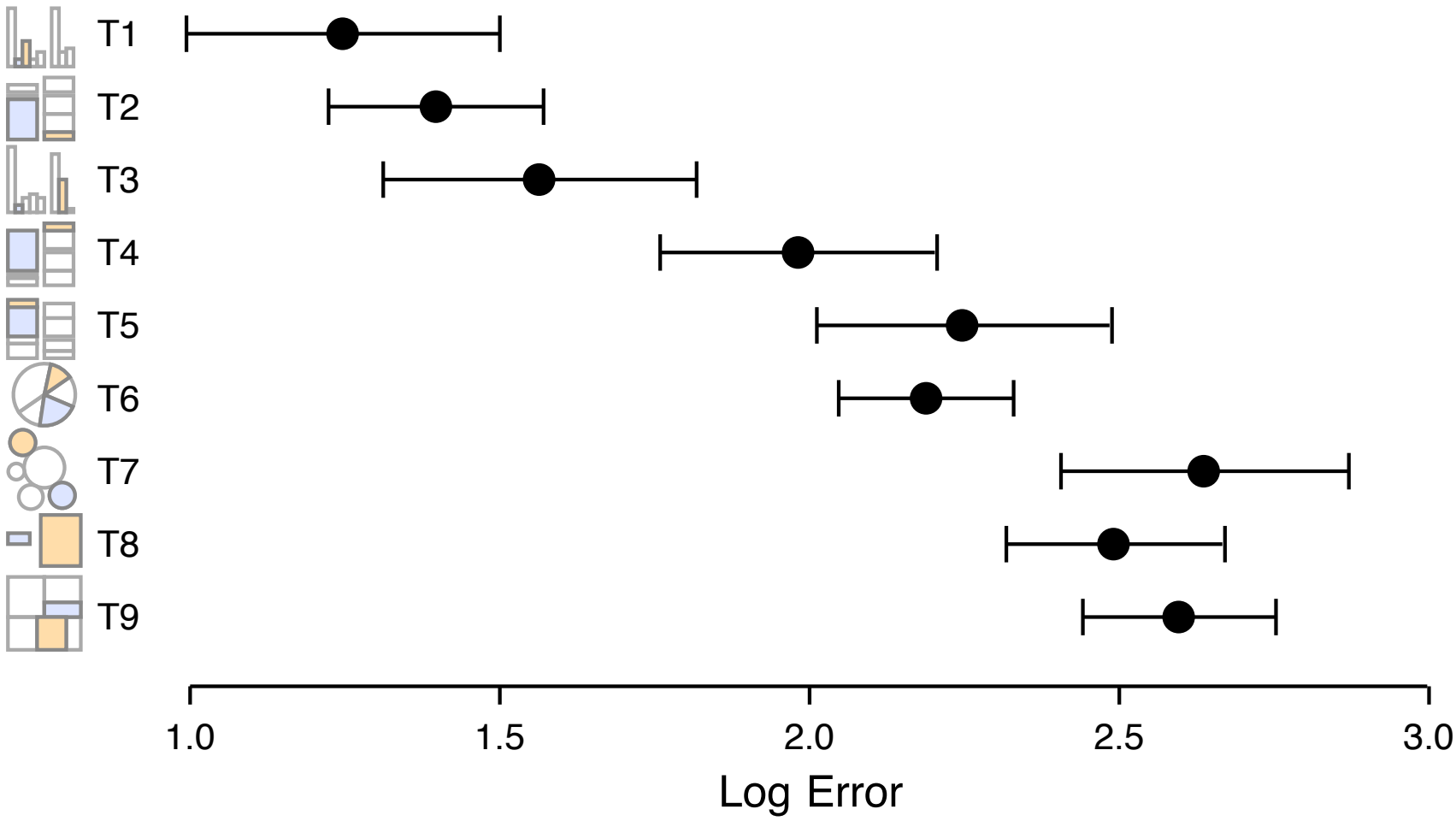


Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape



Effectiveness of visual channels



Heer and Bostock, 2010

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

How layer

Encode

- **Arrange**
  - Express
  - Separate
  - Order
  - Align
  - Use

- **Map**
  from categorical and ordered attributes
  - Color
    - *Hue*
    - *Saturation*
    - *Luminance*
  - Size, Angle, Curvature, ...
  - Shape
  - Motion
    Direction, Rate, Frequency, ...

Manipulate

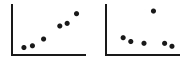
- **Change**

- **Select**

- **Navigate**

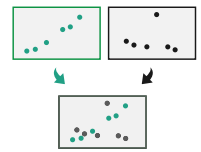
Facet

- **Juxtapose**



- **Partition**

- **Superimpose**

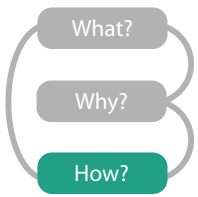


Reduce

- **Filter**

- **Aggregate**

- **Embed**



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

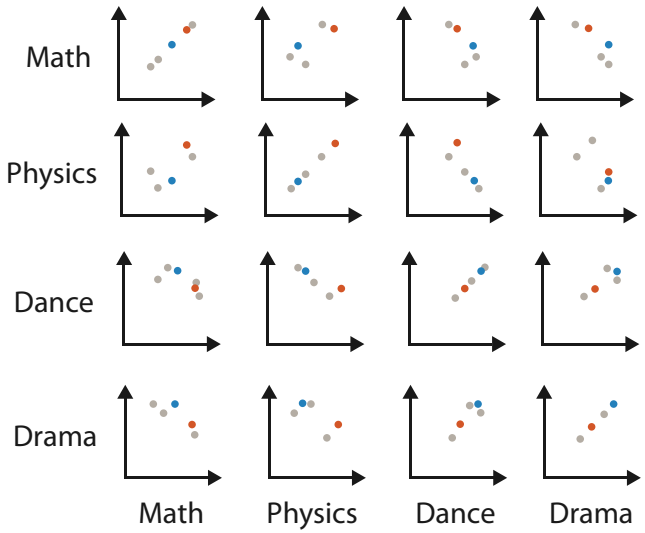
- Rectilinear layouts
- Parallel layouts
- Radial layouts

Parallel layouts

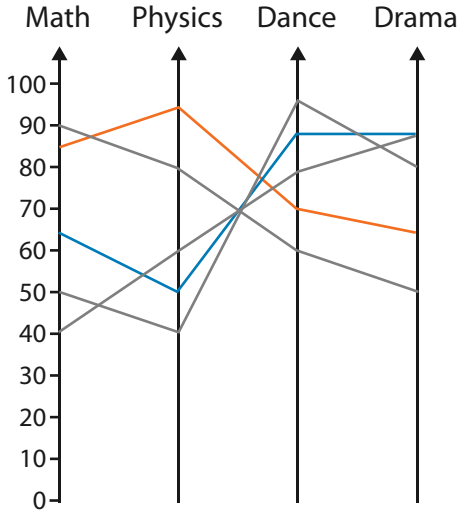
Table

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

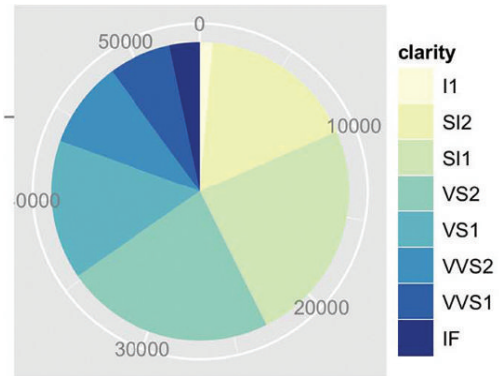


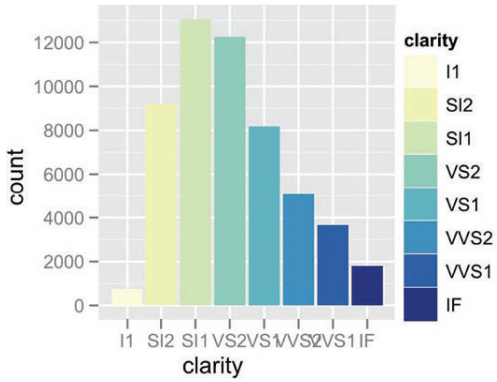
used for

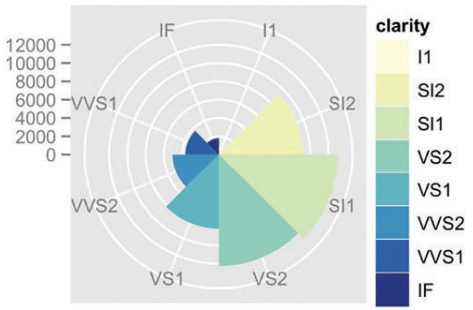
- overview over all the attributes
- range of the individual attributes
- outlier detection
- requires training time for the user

Radial layouts

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



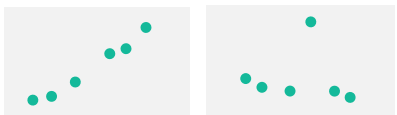




Facet: Juxtaposition

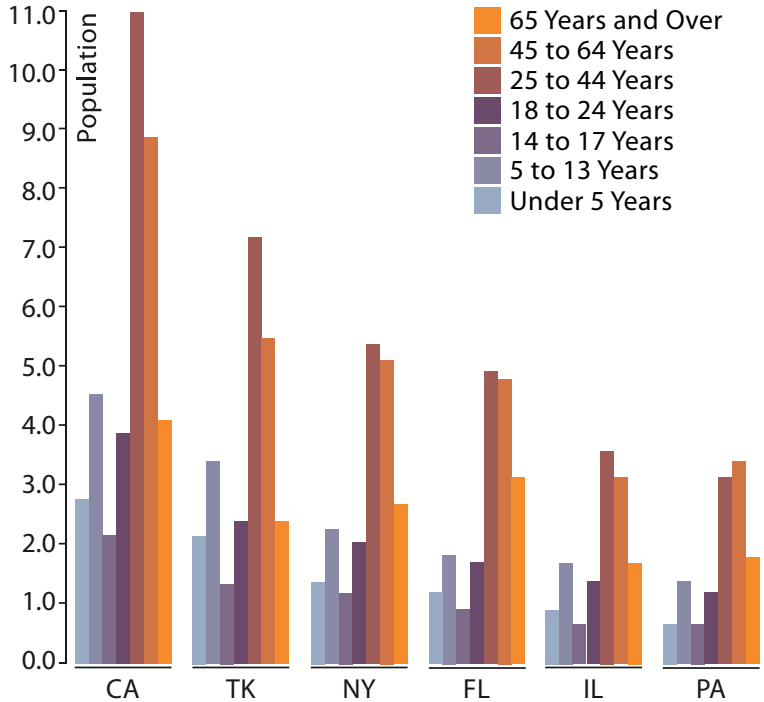
Multiple views

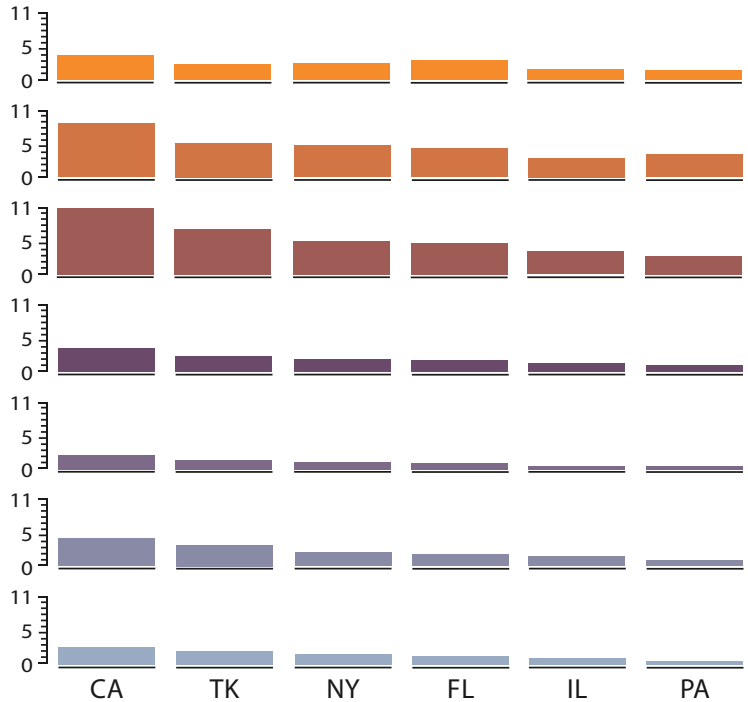
- is the encoding different or same?
- do views share data?

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

Facet: partition

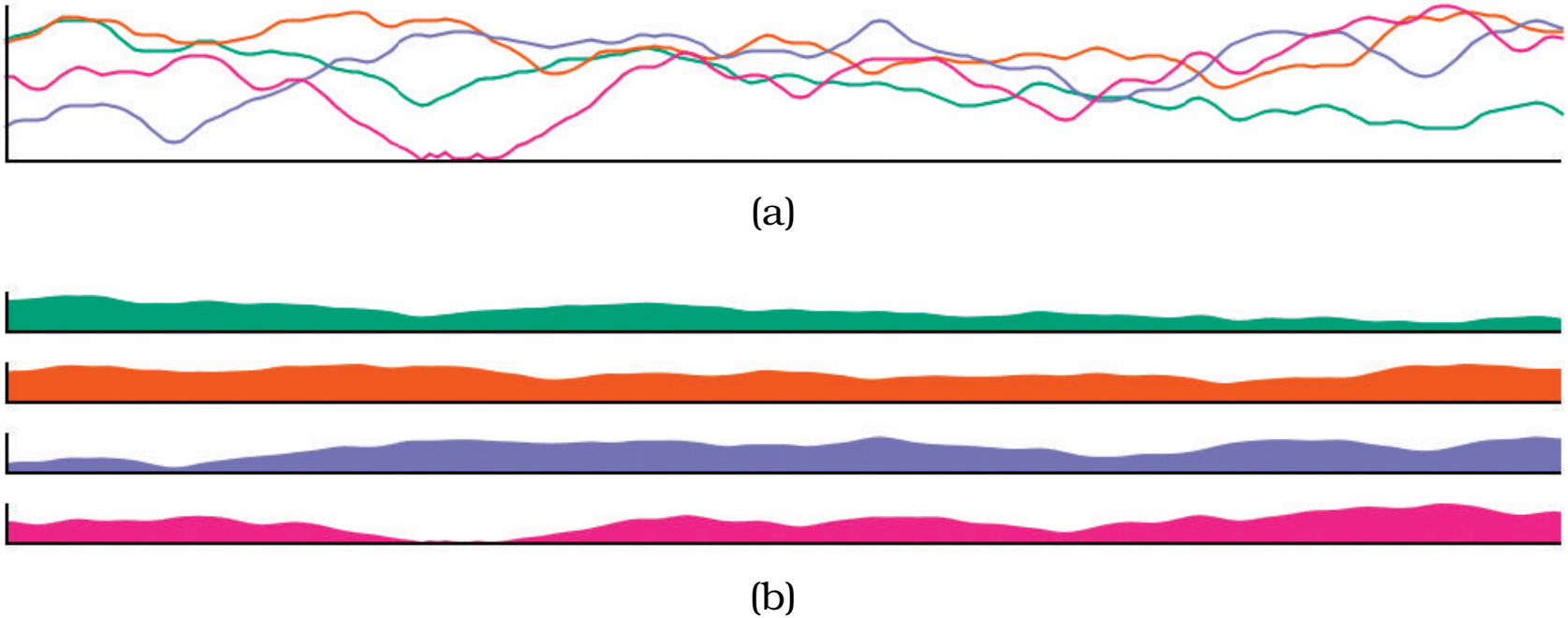
- partition is a design choice on how you separate data into groups
- these groups are assigned to regions (due to separate)
- trivial if you have only one key
- more choices, if you have multiple keys





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.

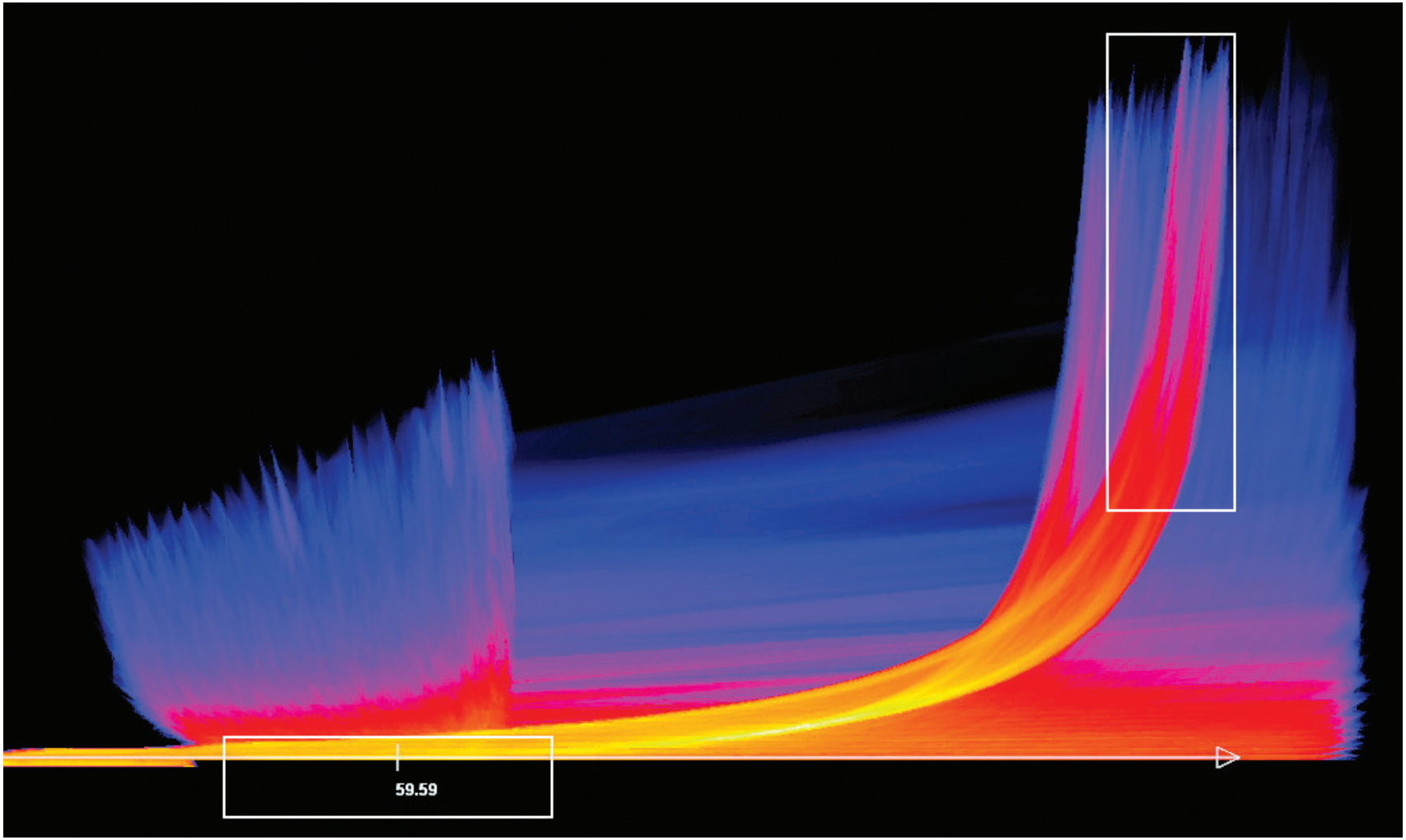
Filter: aggregate

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

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

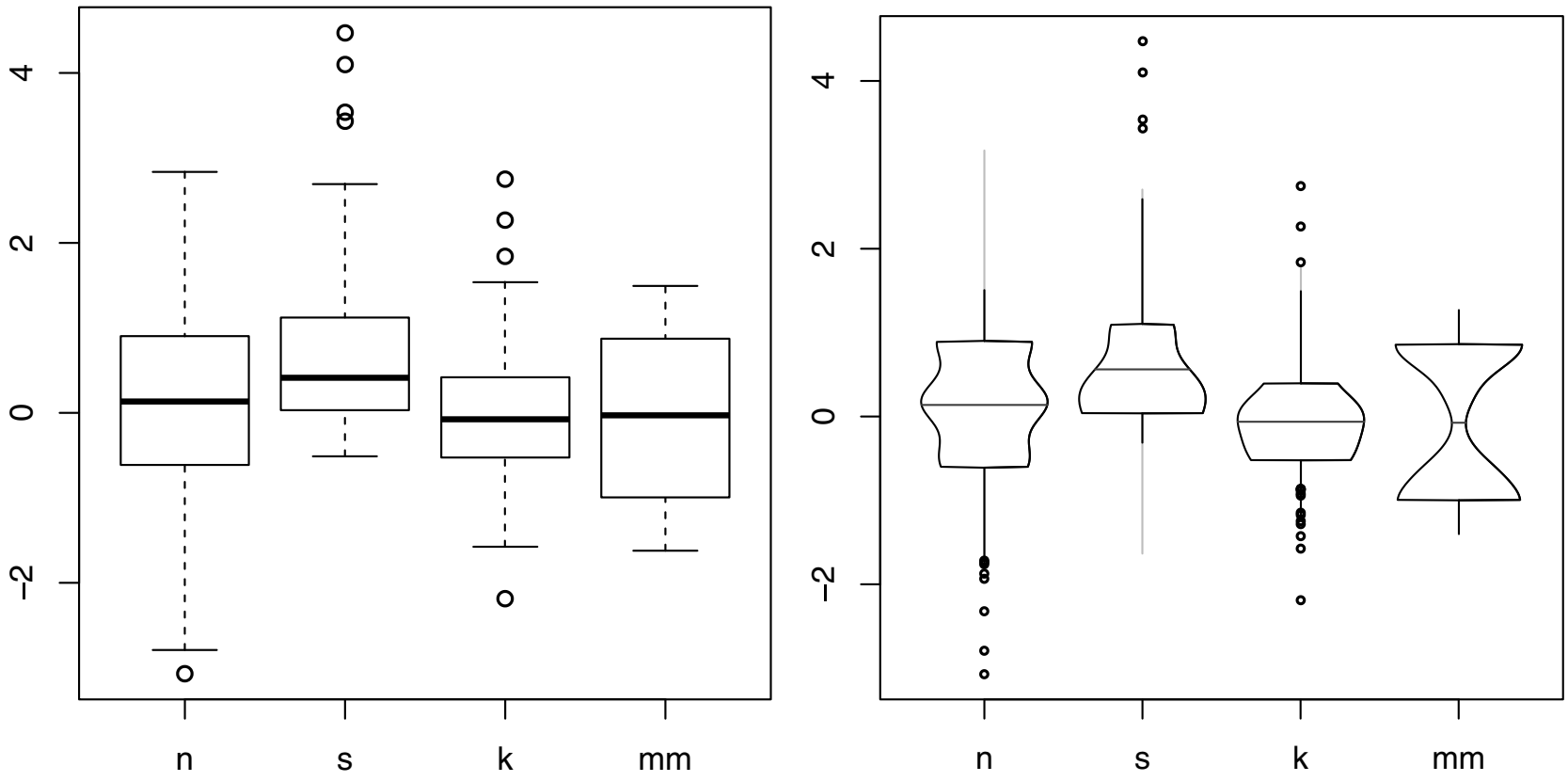
Continuous scatterplot

'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