Information visualization, data mining and machine learning

Fabrice Rossi

Télécom ParisTech

April 2011

Goals of this lecture

1. To give an introduction to Information Visualization (Infovis)
   - enhancement methods for classical displays
   - specialized displays
   - why you should leverage infovis in your everyday work
2. To outline links between Infovis and Machine Learning
   - why do they exist?
   - current solutions
   - open research problems
3. To give examples of successful joint researches:
   - Machine learning methods designed for visualization
   - Visualization of machine learning algorithm results

Outline

Information Visualization

> *The use of computer-supported interactive, visual representation of abstract data to amplify cognition*
>
> Card, Mackinlay & Shneiderman

Information Visualization

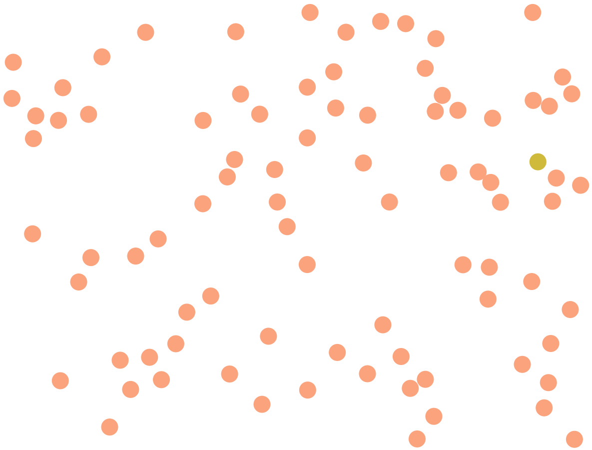
> *The use of computer-supported interactive, visual representation of abstract data to amplify cognition*
>
> Card, Mackinlay & Shneiderman

Human preattentive processing capabilities

- non conscious processing (no thinking involved)
- low level visual system
- extremely fast: 200 ms
- scalable (no browsing $\Rightarrow$ sublinear scaling)
- feature type must match data type (e.g., hue is suitable for categories, less for real values)

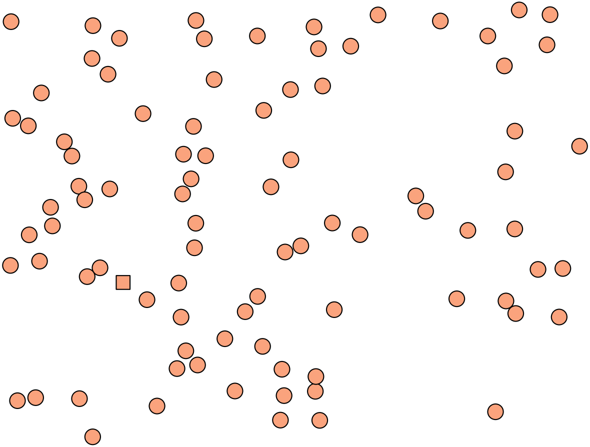
Preattentive features

Hue



Preattentive features

Shape



Information Visualization

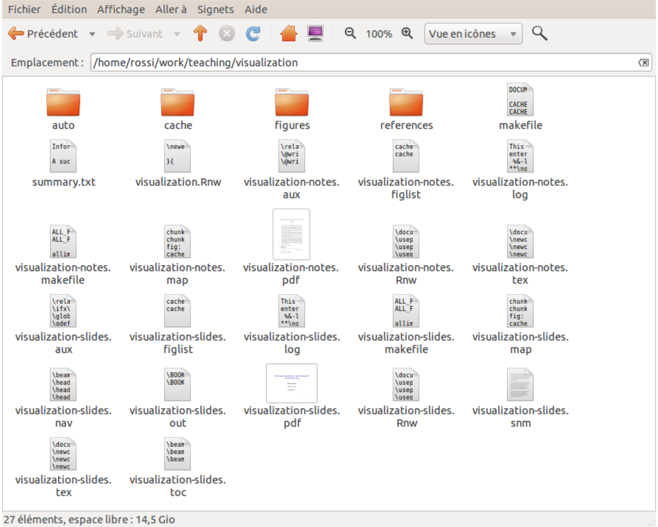
The use of computer-supported interactive, visual representation of abstract data to amplify cognition

Card, Mackinlay & Shneiderman

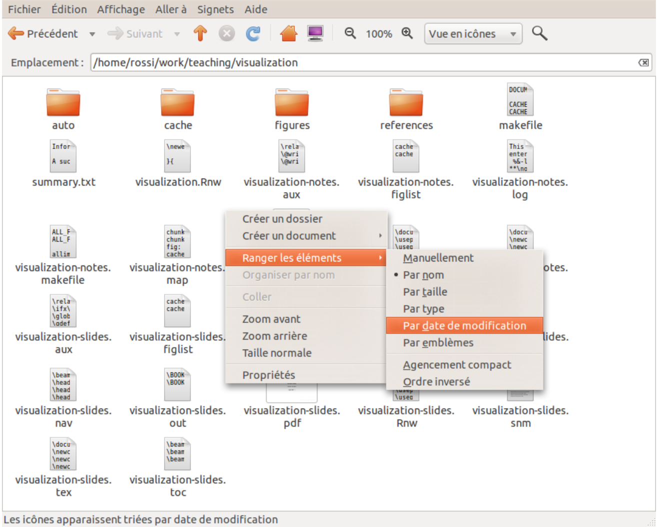
Tool metaphor (hammer, microscope, etc.)

- extending user possibilities:
  - more scalable processing (speed and/or volume)
  - details enhancement
  - multi-source fusion
  - etc.
- under user control

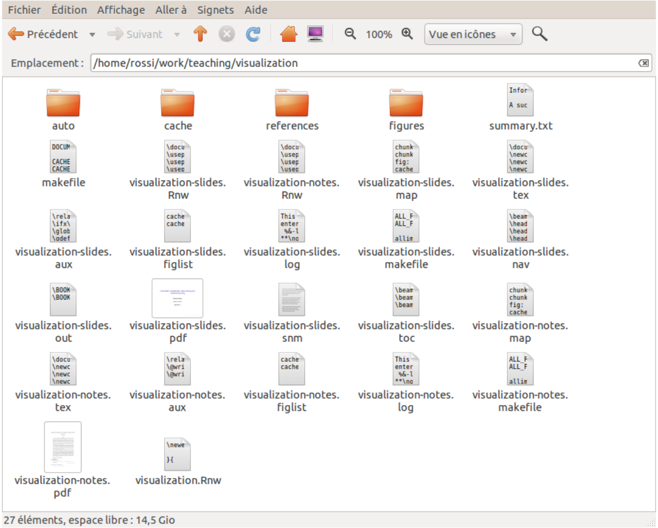
File explorer



File explorer



File explorer



Information Visualization

The use of computer-supported interactive, visual representation of abstract data to amplify cognition

Card, Mackinlay & Shneiderman

Overview first, zoom and filter, and then details-on-demand

Information Seeking Mantra, Shneiderman

Interactivity

- enables user control:
  - exploration (panning)
  - zooming
  - 3D world
- reduces clutter on the screen

Information Visualization

The use of computer-supported interactive, visual representation of abstract data to amplify cognition

Card, Mackinlay & Shneiderman

Abstract data

- digital data with no real world “visual” counterpart, e.g.:
  - sound
  - high dimensional vectors
- no “natural” visual representation of the data, e.g.:
  - requests received by a web server
  - file systems
  - source code

Infovis $\neq$ scientific visualization (Scivis)

Data visualization

Closely related to information visualization

- data $\simeq$ a data table (N objects described by P variables)
- no real consensus on a boundaries between infovis and datavis:
  - datavis $\subset$ infovis: data $\simeq$ preprocessed information (turned into numerical quantities)
  - datavis $\cap$ infovis $= \emptyset$: information (in infovis) is explicitly non numerical (is it?)
  - infovis $\subset$ datavis: because data include scientific data...

Also related to statistical graphics

Outline

What is infovis used for?

Some specific goals

- easier access (learning curve):
  - GUI in general
  - e.g., File system browsing
- productivity (doing the same things but faster):
  - IDE (on the fly documentation, multi-view, graphical programming, etc.)
  - on the fly search (Google suggest)
- organization:
  - tree paradigm (sorting)
  - metadata (image, music, etc.)
  - overview

Visual data mining (VDM)

A.k.a. *Visual Analytics* and *Visual Data Analysis*:

Interactive visual exploration of massive data sets

- cluster analysis
- outlier detection
- dependency assessment
- pattern detection (repetition, sub-structure, etc.)
- etc.

Interactive visualization of the results of data mining algorithms

- parameter tuning
- quality assessment
- mining on the results (e.g., meta-clustering)
- etc.

But...

Limited external impact

Many discoveries of infovis are not used (or even known) outside the infovis community

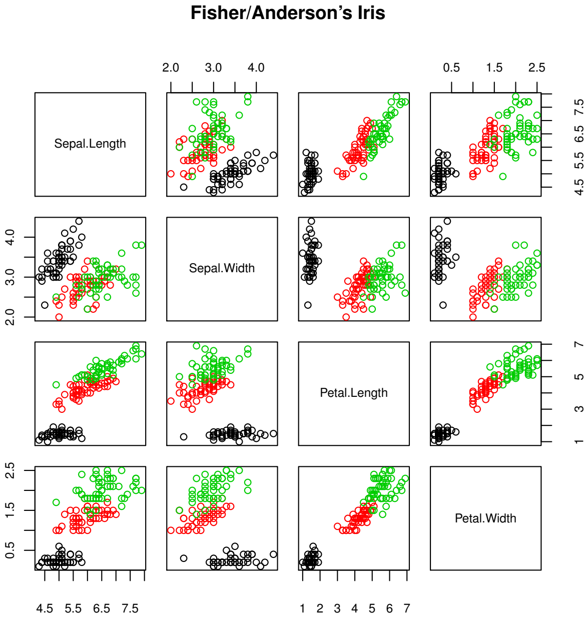
External "state-of-the-art"

- mostly static graphics
- with broken defaults (color-wise, shape-wise, etc.)

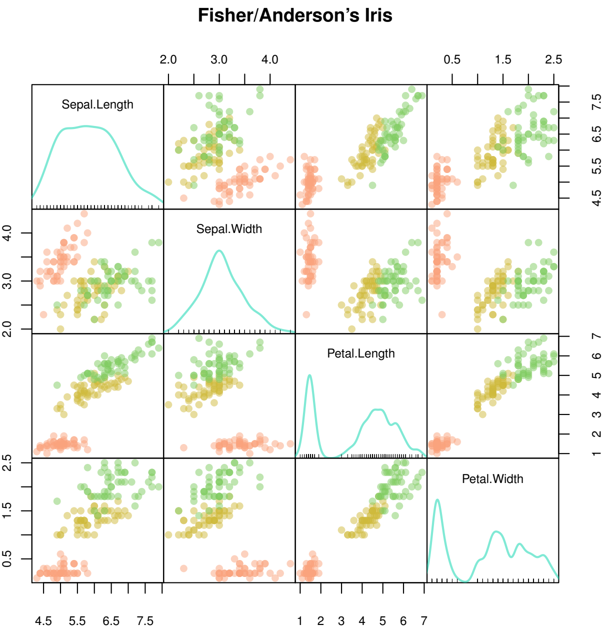
Serious state-of-the-art

- interactive graphics
- high quality static graphics
- sound defaults

For instance, in default R...



Rather than...

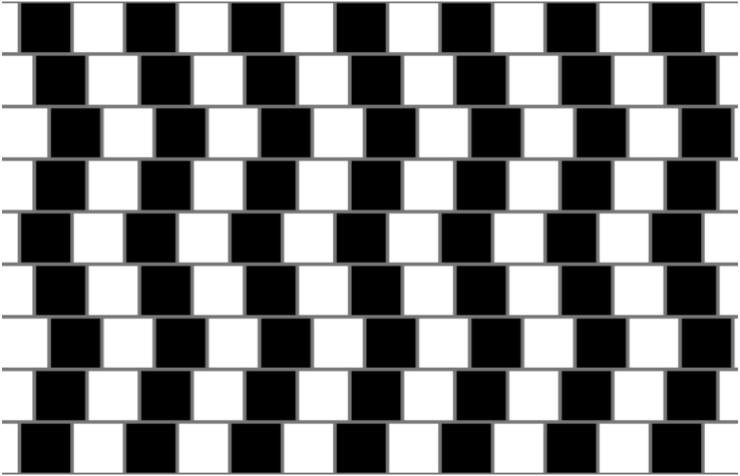

Fisher/Anderson's Iris
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0
3.0
4.0
0.5
1.5
2.5
7.5
6.5
5.5
4.5
4.0
3.0
2.0
7
6
5
4
3
2
1
2.5
1.5
0.5
4.5
5.5
6.5
7.5
1
2
3
4
5
6
7

Outline

Limitations of Infovis and VDM

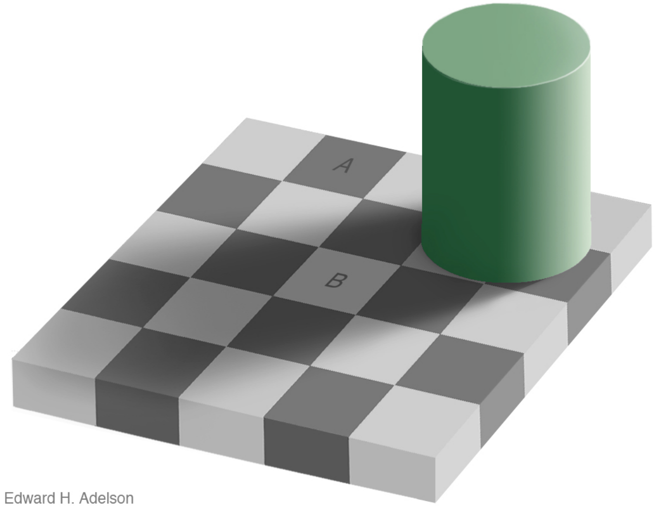
- Visual illusions
- Distortion, occlusion, etc.
- Visual semiotics
- Scalability
  - number of objects
  - number of descriptors
  - human scalability
  - computer scalability
- Art or Science?

Café wall illusion



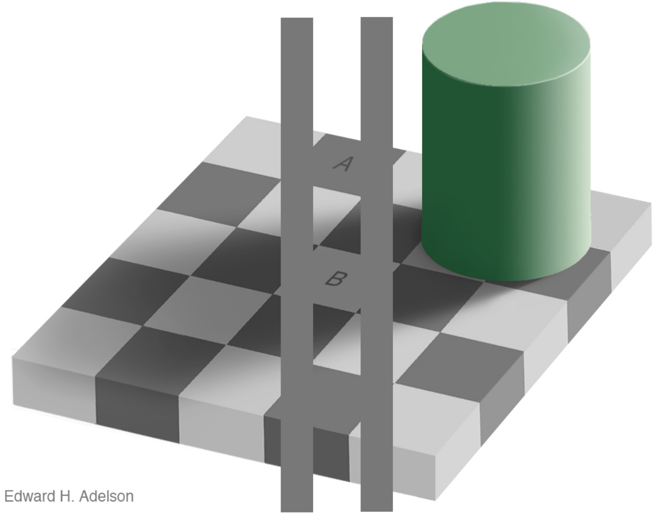
http://en.wikipedia.org/wiki/File:Caf%C3%A9_wall.svg

Grey levels



http://web.mit.edu/persci/people/adelson/index.html

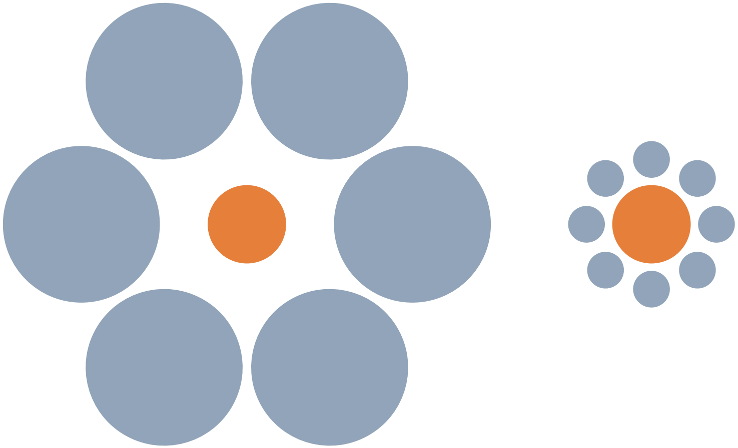
Grey levels



http://web.mit.edu/persci/people/adelson/index.html

Area and size

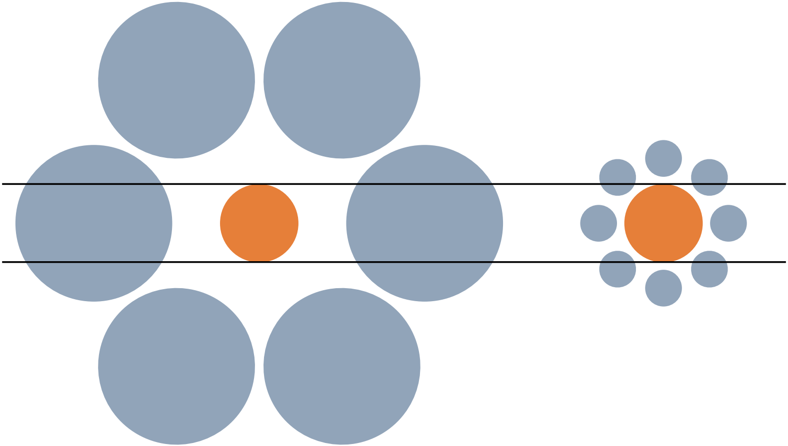
Ebbinghaus illusion



http://en.wikipedia.org/wiki/File:Mond-vergleich.svg

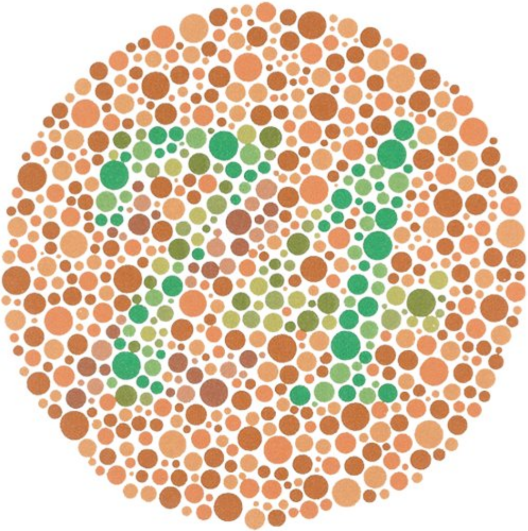
Area and size

Ebbinghaus illusion



http://en.wikipedia.org/wiki/File:Mond-vergleich.svg

Color blindness



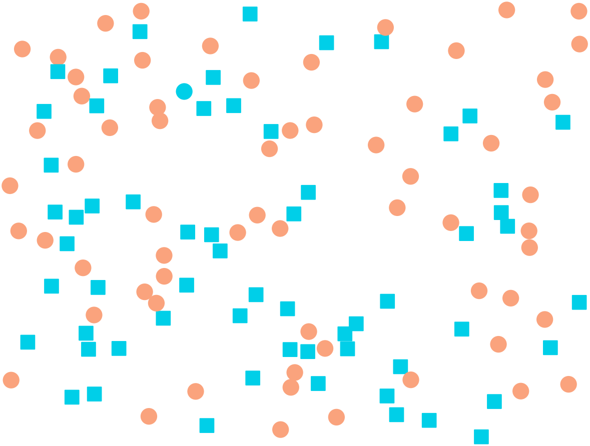
http://en.wikipedia.org/wiki/File:Ishihara_9.png

The scalability issue

- Vision is limited to 2 or 3 dimensions
- Position can be combined with other features:
  - color (intensity and hue)
  - shape (e.g., star icon)
  - texture
  - etc.
- But fast pre-attentive processing is limited to roughly 5 combined features (for some combination only!)
- Correlating distant things is difficult
- Computer screens have a “low” resolution (HD is 2 millions pixels)
- Complex HD interactive display requires dedicated graphic board and associated software (OpenGL and Direct3D, Shader languages)

Broken preattentive features

Hue + shape



How to scale?

Complementary solutions

- interactivity (zooming, distorting, details on demand, etc.)
- data transformation:
  - interaction between objects rather than objects themselves
  - similarity between objects
- data simplification:
  - reduction of the number of objects (summary, clustering, etc.)
  - reduction of the number of characteristics (selection, projection, etc.)
  - compact layout: one glyph per object or one pixel per measurement
- data ordering:
  - positioning related things closely on the screen
  - one to three dimensional ordering

Obviously linked to Machine Learning (clustering, projection, etc.)

Outline

A classical example

Displaying a one dimensional numerical variable

- dataset: N observations $(X_i)_{1 \leq i \leq N}$ described by P numerical variables
- objective: display (X_{ij}) for a fixed j and all i

A classical example

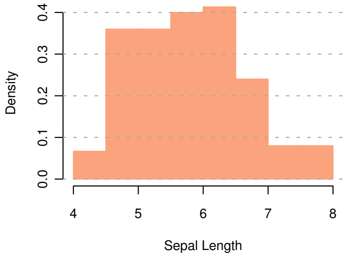
Displaying a one dimensional numerical variable

- dataset: N observations $(X_i)_{1 \leq i \leq N}$ described by P numerical variables
- objective: display (X_{ij}) for a fixed j and all i

Classical solution

A histogram:

- a rough density estimator
- enormous scalability



Histograms in detail

Visualization algorithm

1. choose a number of bins and their widths:
   - numerous possibilities: $\sqrt{N}$, $\log N$, etc.
   - no perfect one
2. for each bin $]a, b]$ compute the fraction of i such that $X_{ij} \in]a, b]$
3. display a bar per bin, with an area proportional to the fraction computed at step 2

Histograms in detail

Visualization algorithm

1. choose a number of bins and their widths:
   - numerous possibilities: $\sqrt{N}$, $\log N$, etc.
   - no perfect one
2. for each bin $]a, b]$ compute the fraction of i such that $X_{ij} \in]a, b]$
3. display a bar per bin, with an area proportional to the fraction computed at step 2

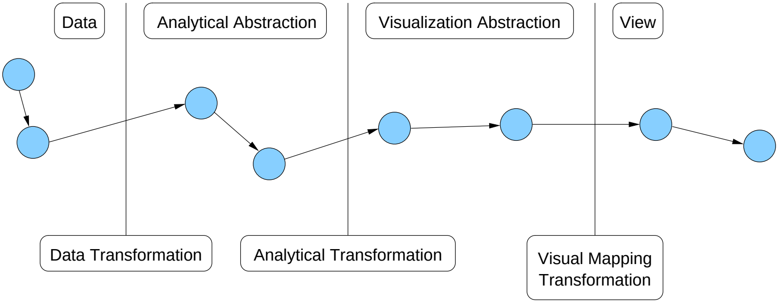
Remarks

- this is statistical learning: non parametric estimation of $p(X_{.j})$
- the original data is not displayed
- the drawing itself remains rather unspecified (at this point) but has crucial consequences: color, aspect ratio, labels, position, etc.

Chi & Riedl's Operator Model

A formal view of infovis [CR98]

Production of the *View* from the *Data*:



Arrows represent transformation operators.

Chi & Riedl's Operator Model

Applied to histograms

data	$(X_{ij})_{1\leq i\leq N}$ for a fixed j
analytical abstraction	K bins $(a_k, b_k, f_k)_{1\leq k\leq K}$, where f_k is the fraction of objects falling in bin k
visual abstraction	K adjacent bars $(w_k, h_k)_{1\leq k\leq K}$, where the width w_k is proportional to $b_k - a_k$ and the h_k is proportional to $\frac{f_k}{w_k}$
view	the histogram itself

Remarks

- the visual mapping step gathers most of the wizard's tricks
- machine learning may operate at numerous stages
- the model includes feedback and interactivity

Machine learning operators

Some examples

- **data transformation**: main field: density estimation, clustering, etc.
- **analytical transformation**: second field: dimensionality reduction techniques
- **visual mapping**: seldom used: optimal ordering and optimal coloring

Internal operators

- **data stage**: de-noising, clustering as a pre-processing
- **analytical stage**: some form of dimensionality reduction, feature selection, model building
- **visualization stage**: visual oriented transformation (e.g., second clustering)

Visual Mapping Transformation

View "Language"

Graphical primitives:

- coordinates (two or three)
- symbols (dots, ticks, *glyphs*)
- lines and areas (larger scale symbols)
- text
- color:
  - Hue and Saturation
  - Ligthness

Mapping

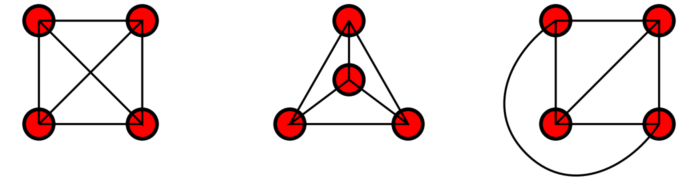
- expression of the visual abstraction in terms of the graphical language
- arrangement, color choice, axes, etc.

View design

The difficult part

- probably the most difficult part of infovis
- many heated debates...
- between art and science
- tricky evaluation:
  - experimental psychology
  - long and complex
  - task oriented

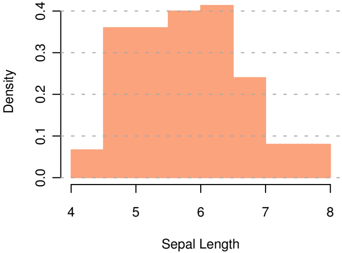
Graph visualization



Histogram mapping

The mapping

- **bar** a solid rectangle, no border
- **layout** horizontal data axis, vertical density axis
- **key** data description, value description, horizontal grid



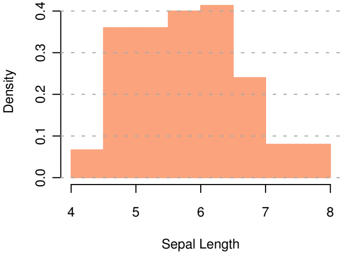
Histogram mapping

The mapping

bar a solid rectangle, no border

layout horizontal data axis, vertical density axis

key data description, value description, horizontal grid



Rationale?

- do we need color?
- don't we need bar borders?
- what are we trying to show?

Let's summon statistics...

Definition

A histogram is a (rough) density estimator. Given $X_1, \ldots, X_N$ distributed according to $P(X)$ is approximates $f(x)$ such that $P(X \in B) = \int_B f(x)dx$.

Let's summon statistics...

Definition

A histogram is a (rough) density estimator. Given $X_1, \ldots, X_N$ distributed according to $P(X)$ is approximates $f(x)$ such that $P(X \in B) = \int_B f(x)dx$.

"Natural" visualization

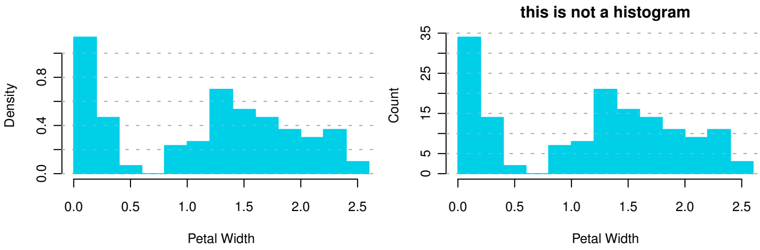
Then let's draw $f(x)$!

- natural axes: horizontal for the values, vertical for the likelihood
- colored and borderless bars emphasize the area (as opposed to the height): proper probabilistic interpretation

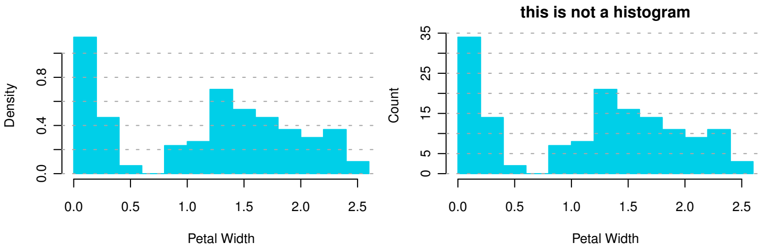
Use

- shape analysis: symmetric? unimodal? Gaussian like?
- quality control: outliers and other unexpected behaviors

Density or counting?



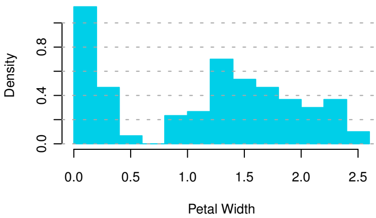
Density or counting?

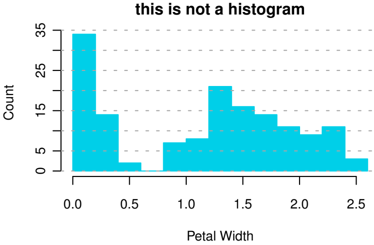


Counting as the main outcome

The right hand side "histogram" displays the number of objects who fall into each bin.

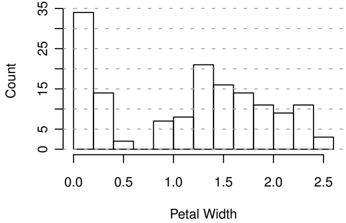
Density or counting?





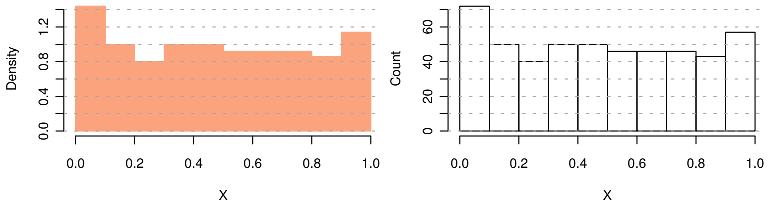
Counting as the main outcome

The right hand side “histogram” displays the number of objects who fall into each bin. Colorless bordered bars are more adapted to inference in this case.



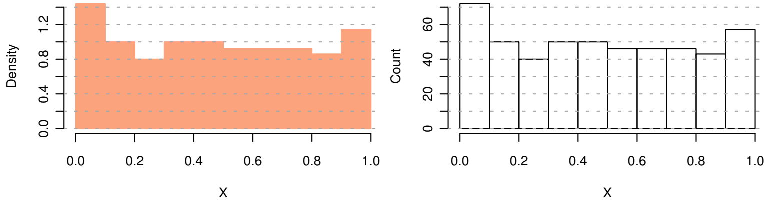
Unnatural visualization

So far, so good

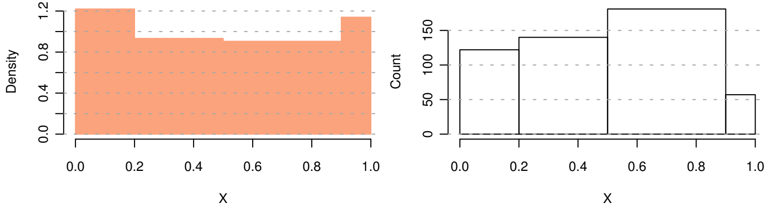


Unnatural visualization

So far, so good



Oops... (this is a uniform distribution)



The infovis process

Wrapping up

The process

1. extract information from the raw data
2. build an abstract visualization of this information
3. map it to a view using a sound graphical language

Major points

- a view should be adapted to some specific goal
- less is more: remove interferences
- self-contained views: axes, keys, labels are part of the mapping process
- faithful views: do what you claim to do; beware of “natural representations”

Outline

Information Visualization

- Definition
- Infovis applications
- Limitations of Infovis and VDM
- An introductory example: the histogram
- Another example: categorical data
- Interactivity

Machine learning and visualization

- One dimensional data
- Two dimensional data

Another classical example

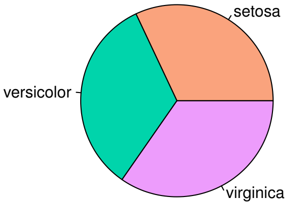
Displaying a one dimensional categorical variable

- dataset: N observations $(X_i)_{1\leq i\leq N}$ described by one categorical variable with Q possible values in $\mathcal{V} = \{V_1, \dots, V_Q\}$
- zero structure on $\mathcal{V}$ (in particular, no order)
- objective: display $(X_i)_{1\leq i\leq N}$

Classical solution

A pie chart

- a counting like approach
- enormous scalability with respect to N (none with respect to Q)



Another classical example

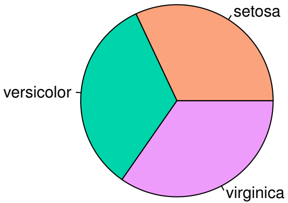
Displaying a one dimensional categorical variable

- dataset: N observations $(X_i)_{1 \leq i \leq N}$ described by one categorical variable with Q possible values in $\mathcal{V} = \{V_1, \ldots, V_Q\}$
- zero structure on $\mathcal{V}$ (in particular, no order)
- objective: display $(X_i)_{1 \leq i \leq N}$

Classical solution (but a very bad one)

A pie chart

- a counting like approach
- enormous scalability with respect to N (none with respect to Q)



Chi & Riedl's Operator Model

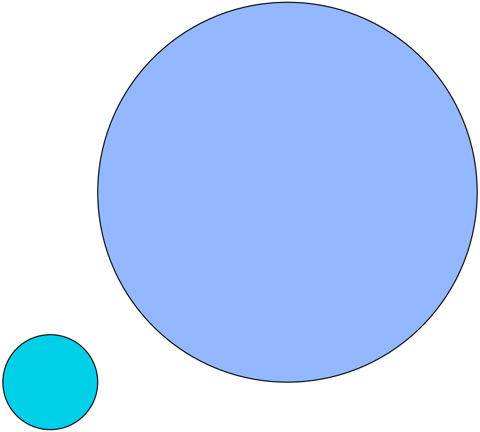
Applied to pie charts

data	$(X_i)_{1 \leq i \leq N}$
analytical abstraction	N_q the number of i such that $X_i = V_q$
visual abstraction	a sliced pie whose Q slices have an area proportional to N_q
view	the colored pie, with labeled slices ordered according to some rule

Representation principle

- the quantity of interest is encoded via an area (equivalently an angle)
- inference is based on area (angle) comparisons
- in histograms, inference is based on shape analysis (skewness, spread, etc.)
- in counting histograms, inference is based on bar lengths

Let's play...



Please write down your estimation of the ratio of the areas of those disks.

Let's play...

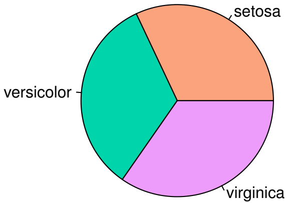
Please write down your estimation of the ratio of the lengths of those bars.

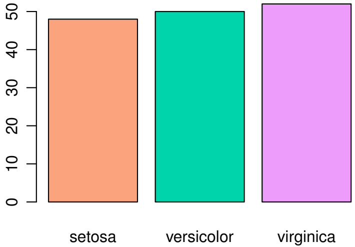
Bar chart

Another visual abstraction

Using the same counting data, replace the Q pie slices by Q bars with length/height proportional to N_q

And the views are



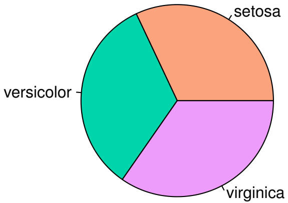


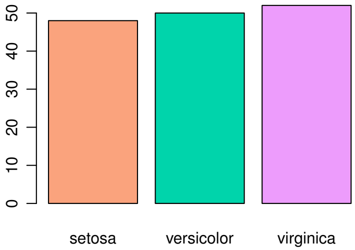
Bar chart

Another visual abstraction

Using the same counting data, replace the Q pie slices by Q bars with length/height proportional to N_q

And the views are





Did you noticed the unequal proportions in the first occurrence of this pie chart? Can you now?

Stevens' Power Law

Definition (Wikipedia)

Stevens' power law is a proposed relationship between the magnitude of a physical stimulus and its perceived intensity or strength, with the general form $\psi(I) = kI^{\alpha}$

Some examples in visualization

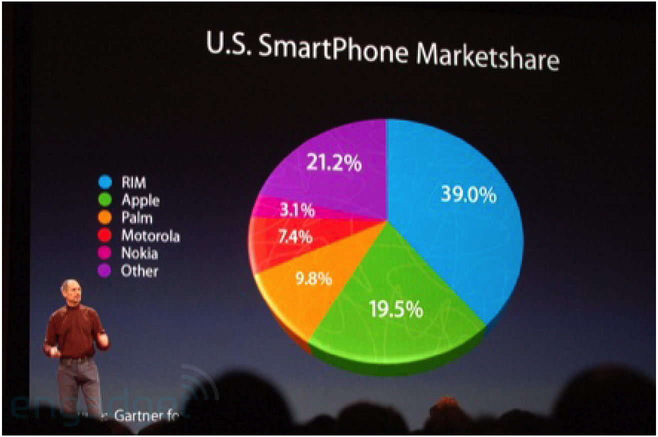
length: $\alpha \in [0.9; 1.1]$

area: $\alpha \in [0.6; 0.9]$

volume: $\alpha \in [0.5; 0.8]$

We tend to underestimate large areas and to overestimate small ones.

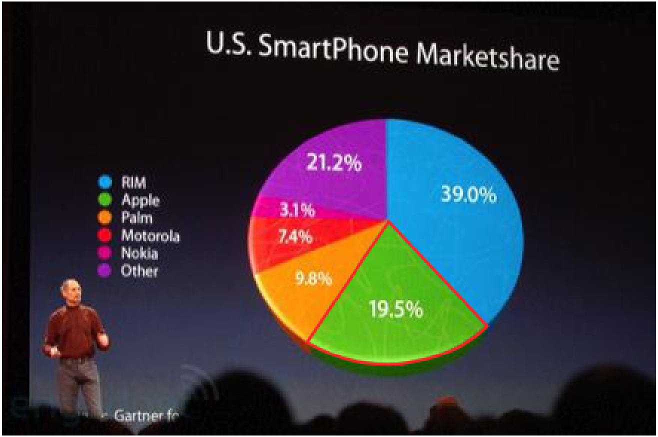
Cheating with the pie...



Steve Jobs' keynote at Macworld 2008, source:

`http://www.engadget.com/2008/01/15/live-from-macworld-2008-steve-jobs-keynote/`

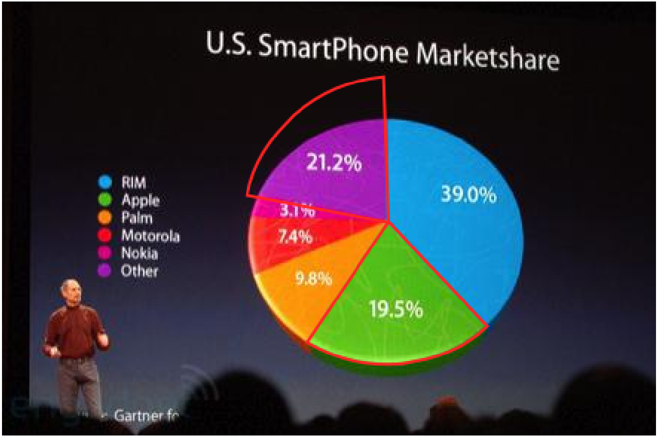
Cheating with the pie...



Steve Jobs' keynote at Macworld 2008, source:

http://www.engadget.com/2008/01/15/live-from-macworld-2008-steve-jobs-keynote/

Cheating with the pie...

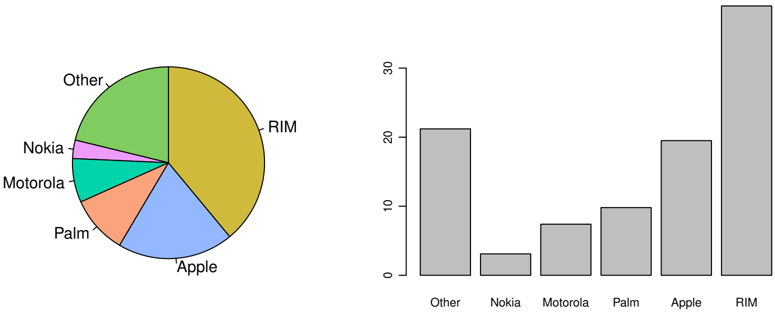


Steve Jobs' keynote at Macworld 2008, source:

`http://www.engadget.com/2008/01/15/live-from-macworld-2008-steve-jobs-keynote/`

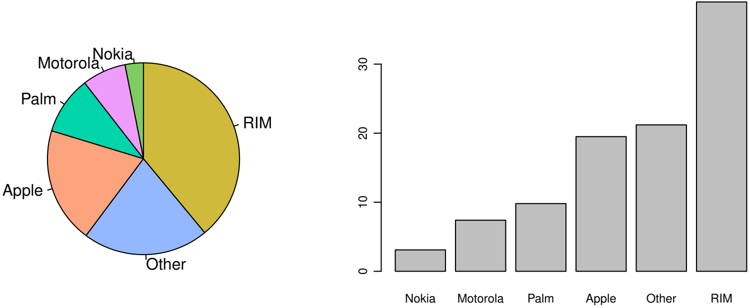
Back to the facts

With Jobs' ordering



Back to the facts

With a more natural ordering



The lie factor

Definition

The lie factor is the ratio between the (relative) size of a quantity on a graphic and its (relative) size in the data.

> *The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented.*

Tufte, 1991

Jobs' keynote lie factor

real effect	$19.5\%/21.2\% \simeq 0.92$
graphic effect	roughly 1.5
lie factor	roughly 1.6

Variations over the view

Remarks

- Pie charts and Bar charts differ mostly on the visual abstraction: area/angle versus length/height
- Cleveland and McGill have shown that lengths encode numerical values much more efficiently than areas and therefore that you should

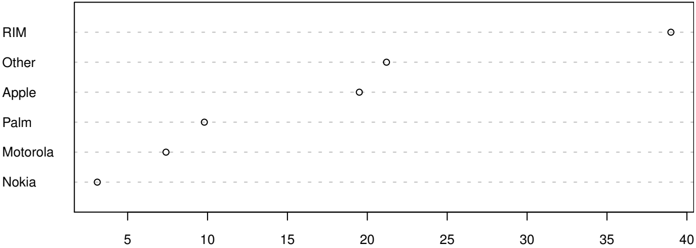
  Save the pies for dessert

  Stephen Few
- what about the view?
  - common challenges for both charts: ordering, labelling, coloring
  - but do we need to materialize the bars?

Cleveland's dot plot

Mostly a visual mapping variation

- use the same visual abstraction (bars of length proportional to counts)
- with a much simpler view



- much more scalable than the bar chart with respect to the number of modalities

Ranking visual features

Cleveland & McGill

Quality of numerical encoding

In decreasing order:

1. position on a common scale
2. position along identical scales but non-aligned
3. length
4. angle and slope
5. area
6. volume
7. color properties (brightness, etc.)

Visual elements

Summary

- the visual abstraction induces the main aspects of the view:
  - lengths will be mapped to positions or actual lengths
  - areas will be mapped to graphical primitives (disks, squares, etc.) with the requested areas
- visual features are not born equal:
  - positions and lengths are rather easily compared
  - areas and angles are much more misleading
  - colors (hue) are useful for differentiation not for coding quantities
- visual mapping can ruin the visual abstraction:
  - 3D is very dangerous: it does not preserve areas, angles, etc.
  - distraction, cluttering, etc. must be avoided

Outline

Interactivity

Core feature of information visualization

Infovis research emphasizes the practical relevance of interactivity

- exploration: panning, zooming, rotating (3D)
- details on demand, distortion: reduces clutter
- conditional analysis: helps testing "what if?" scenario
- user choice: parameter tuning, guided exploration

Nice bonus: some form of interactivity is mostly orthogonal to the rest of the infovis pipeline

Current status

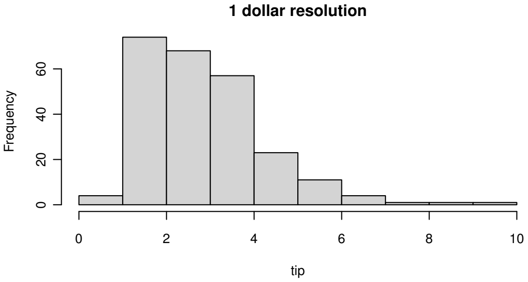
- rather limited diffusion except for basic tricks, such as 3D
- still computationally demanding; for some interactions, the hardware is there (even in smartphones!) but programming it is still hard
- major ideas (e.g., brushing and linking) are still not mainstream

A simple example

The tips dataset

- tips received during two and half months by a food server in the early 90s
- 7 variables including the tip received for the meal

Count histogram of the tips: bin number effect

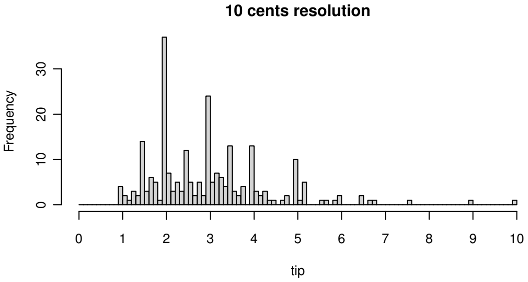


A simple example

The tips dataset

- tips received during two and half months by a food server in the early 90s
- 7 variables including the tip received for the meal

Count histogram of the tips: bin number effect



Parameter selection

General paradigm

- some parametric visualization method:
  - analytical abstraction (e.g. number of bins)
  - visual abstraction (e.g. histogram vs counting histogram)
  - view (e.g. aspect ratio)
- simple interactive way of modifying the parameter value (slider, keyboard, etc.)
- real time update of the view

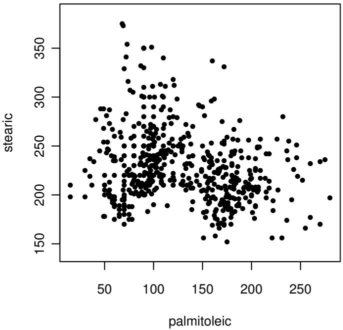
Strong requirements

- deterministic results (minimum surprise principle)
- real time performances
- animated transitions (recommended)

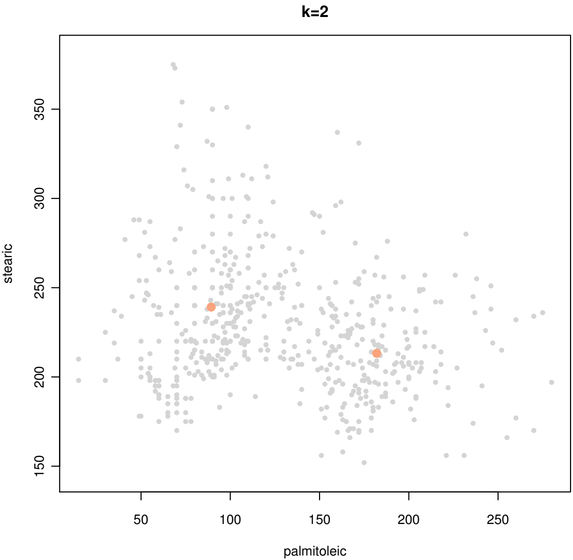
A naive example

K-means clustering

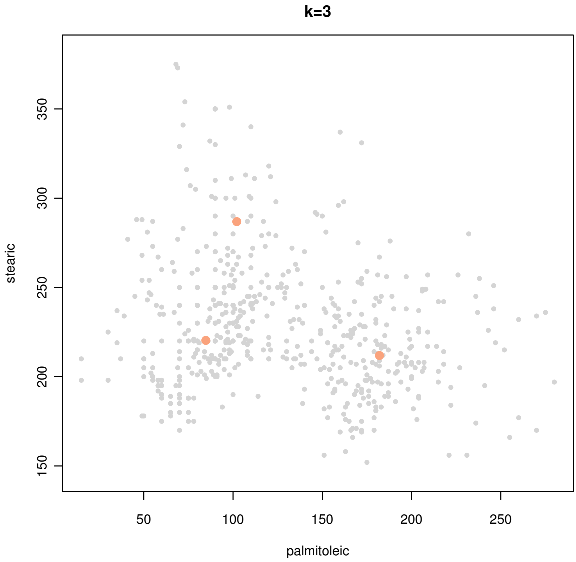
- a simple two dimensional dataset (for illustration)
- minimalist view: scatter plot with prototypes
- interaction: number of clusters



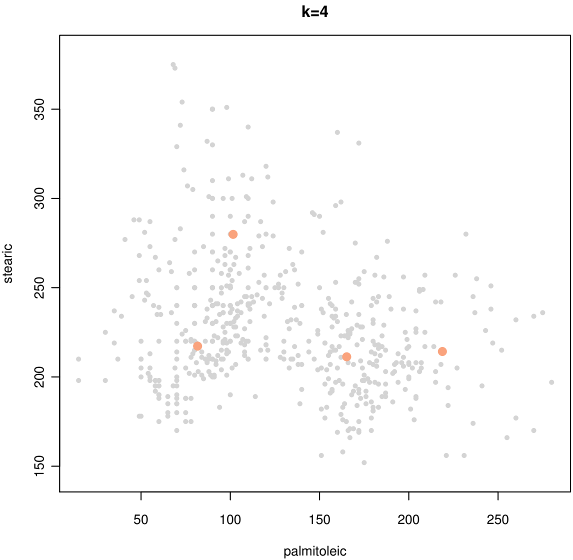
A naive example (of failure)



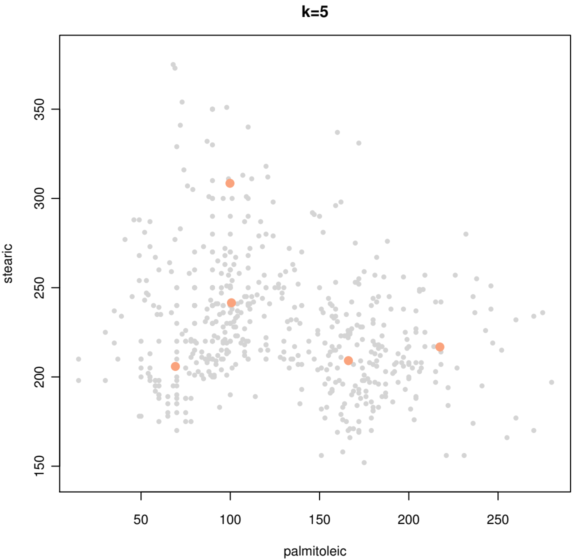
A naive example (of failure)



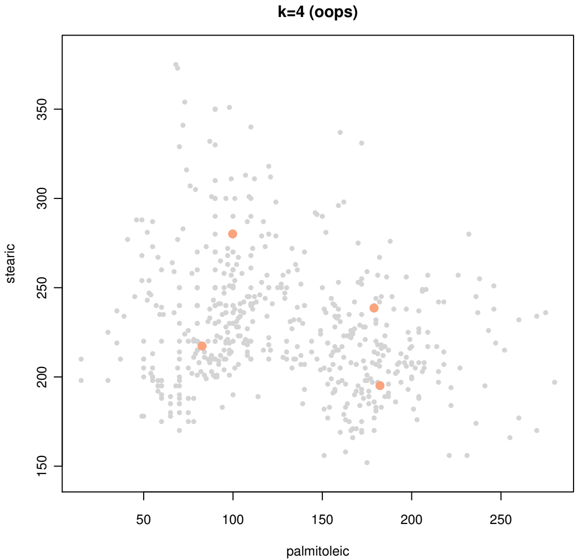
A naive example (of failure)



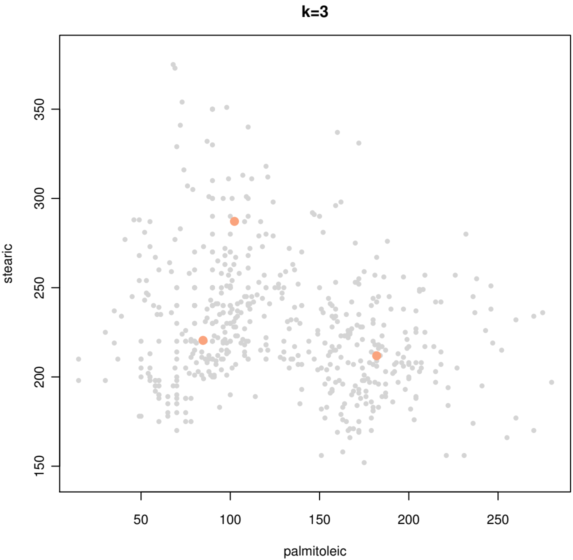
A naive example (of failure)



A naive example (of failure)



A naive example (of failure)



State of the art

Available

- generic view interaction (including 3D): zooming, panning, rotating, etc.
- specific view interaction: ordering, alpha blending, labelling, etc.
- visual abstraction interaction: switching from one representation to another
- data abstraction interaction: basic control of simple things (e.g., histogram bandwidth)

Todo...

- deeper interaction: complex data abstraction control (e.g., number of clusters, of variables, etc.)
- quality feedback
- hypotheses formulation and testing

Brushing and Linking

A.k.a. Conditioning

- simple but crucial technique
- represent whatever variables X and Y (numerical, categorical, high dimensional, etc.)
- show the effect on X of selecting part of the values of Y
- technically: compare e.g. $P(X|Y \in \mathcal{Y})$ and $P(X)$

Definition

Brushing: selecting a subset of the data items with an input device

Definition

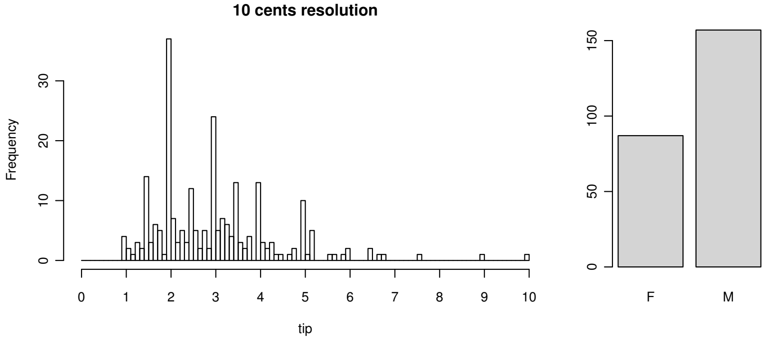
Linking: showing the effect of a brush on all representations of the data

Simple example

Back to the tips dataset

- tips received during two and half months by a food server in the early 90s
- effect of the gender of the tiper?

Count histogram of the tips

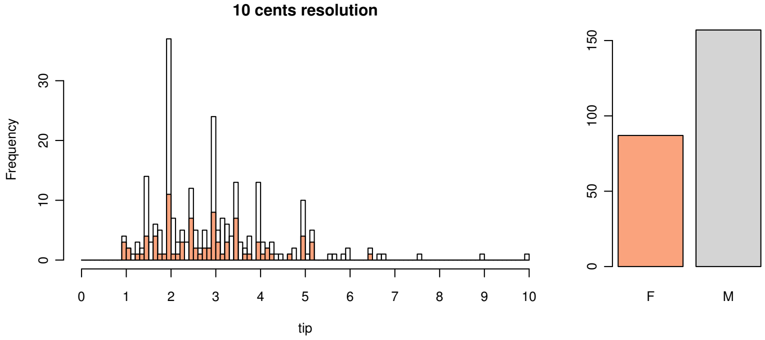


Simple example

Back to the tips dataset

- tips received during two and half months by a food server in the early 90s
- effect of the gender of the tiper?

Count histogram of the tips

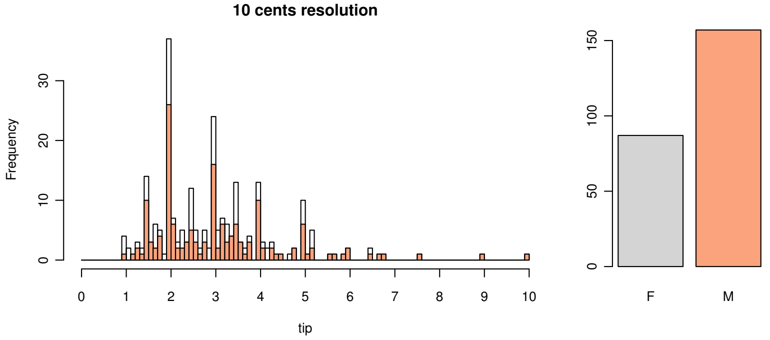


Simple example

Back to the tips dataset

- tips received during two and half months by a food server in the early 90s
- effect of the gender of the tiper?

Count histogram of the tips

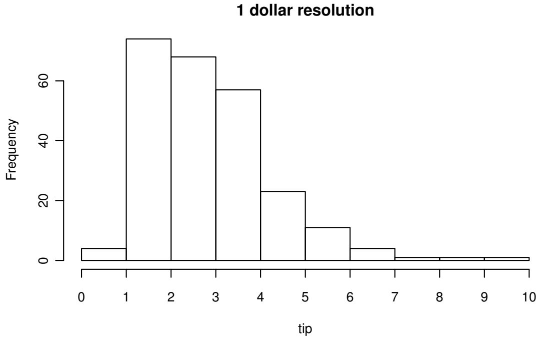


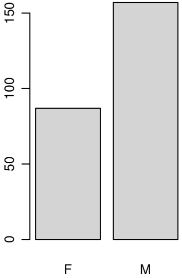
Simple example

Back to the tips dataset

- tips received during two and half months by a food server in the early 90s
- effect of the gender of the tiper?

Count histogram of the tips



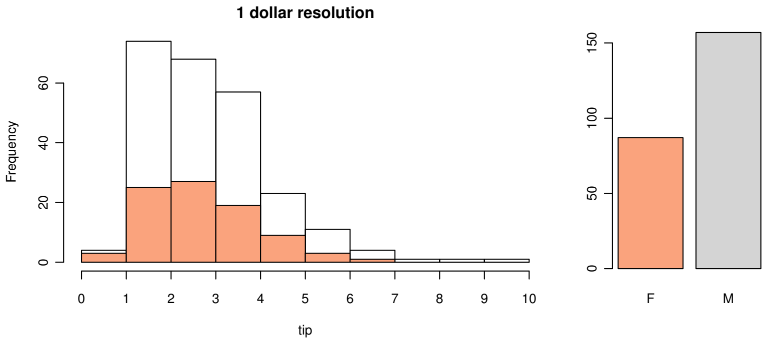


Simple example

Back to the tips dataset

- tips received during two and half months by a food server in the early 90s
- effect of the gender of the tiper?

Count histogram of the tips

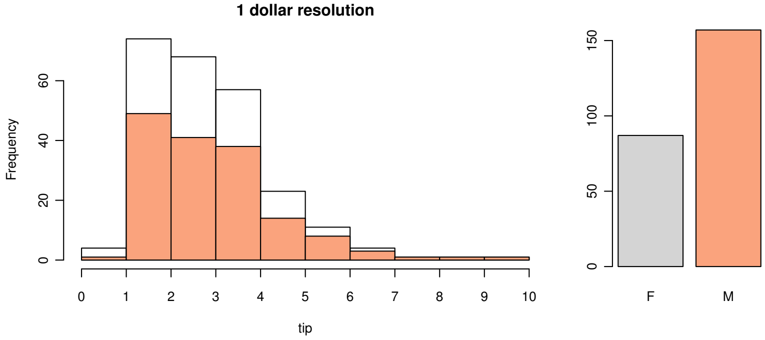


Simple example

Back to the tips dataset

- tips received during two and half months by a food server in the early 90s
- effect of the gender of the tiper?

Count histogram of the tips



Brushing and Linking

Implementation

is not that obvious...

- complex effect: goes all the way back to the data
- process:
  1. provide a selection technique
  2. map selection on the screen to selection on the original data
  3. apply the data transformation chain to this subset on all the views
  4. display the subset together with the full set on all the views
- numerous implied hypotheses:
  - deterministic methods
  - context aware methods (the full data context must be imposed to the subset)
  - meaningful superimposed displays
- highly goal dependent (again...)

Main points

- infovis methods consist of a succession transformation steps from the raw data to a view
- machine learning and statistics can play some role in most of the transformations
- visual mapping is crucial to readability and inference:
  - explicit encoding principles
  - visual features encoding quality ordering
- brushing and linking enable conditional analysis
- interactivity needs real-time, animation and determinism

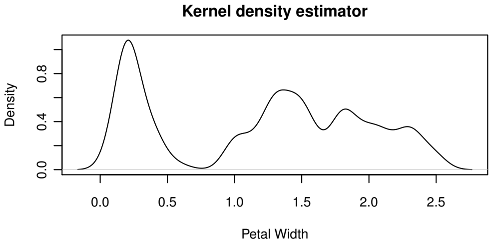
Outline

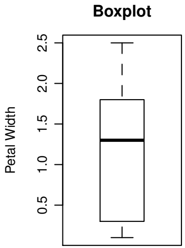
Histograms

Can we do better?

- "numerical" summaries: median, mean, standard deviation, etc.
- smooth density estimates

Associated visualizations





Analysis

Chi and Riedl's model

data	$(X_{ij})_{1 \leq i \leq N}$ for a fixed j
analytical abstraction	a kernel density estimate of $p(x)$ or some basic statistics
visual abstraction	a function $x \mapsto p(x)$ or a fancy box with whiskers (legs in French)
view	see previous slide

Variations

Can we do better?

- less distraction (simpler design)
- readability
- more context

Data-Ink ratio

Less is more (sort of)

Edward Tufte's version:

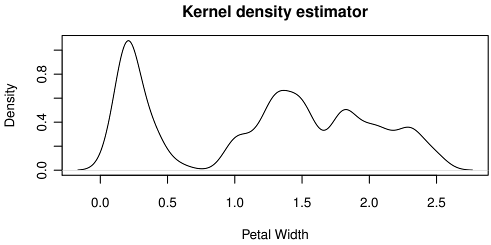
- ink is used to print data and non data
- one should maximize the ratio between the data ink and the total ink

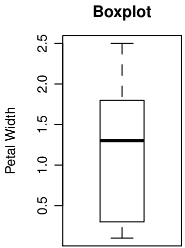
> *A large share of ink on a graphic should present data-information, the ink changing as the data change. Data-ink is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented.*

Tufte, 1983

Examples

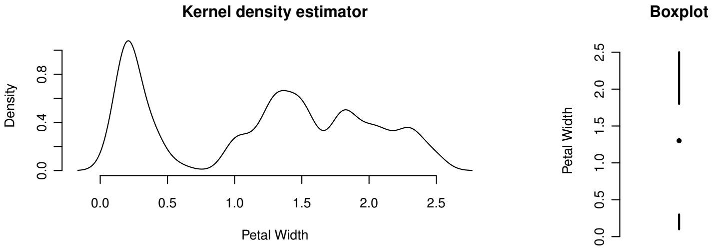
Original designs





Examples

Higher data-ink ratios



Readability

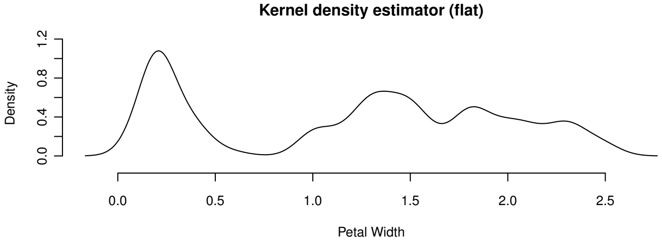
Aspect ratio

Drawing a function is simple, is it?

- the aspect ratio is the graph plays in fact a crucial role
- slopes are difficult to judge
- Cleveland, McGill and McGill have shown that slopes far away from 45° are difficult to estimate

To maximize readability, choose the aspect ratio according to this finding

Example



Readability

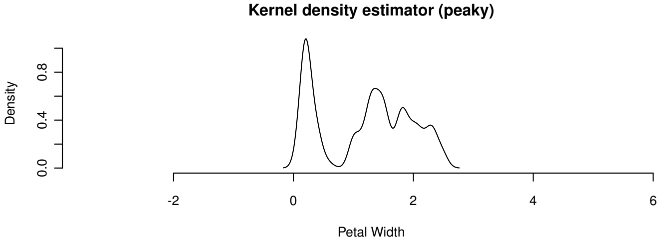
Aspect ratio

Drawing a function is simple, is it?

- the aspect ratio is the graph plays in fact a crucial role
- slopes are difficult to judge
- Cleveland, McGill and McGill have shown that slopes far away from 45° are difficult to estimate

To maximize readability, choose the aspect ratio according to this finding

Example



Readability

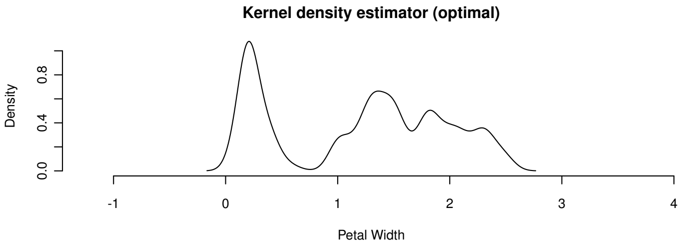
Aspect ratio

Drawing a function is simple, is it?

- the aspect ratio is the graph plays in fact a crucial role
- slopes are difficult to judge
- Cleveland, McGill and McGill have shown that slopes far away from 45° are difficult to estimate

To maximize readability, choose the aspect ratio according to this finding

Example



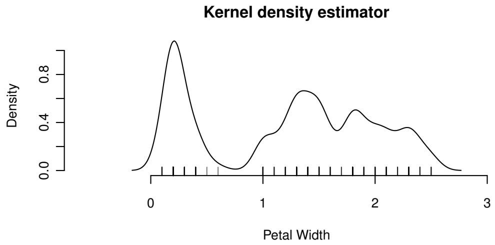
Context

Misleading aspects of density plots

- smoothing effect: generally assign weights to empty intervals
- smoothing effect again: generally extend the actual range of the observations

$\Rightarrow$ Context is needed

Adding a rug to the plot



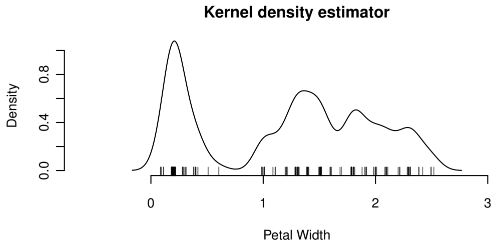
Context

Misleading aspects of density plots

- smoothing effect: generally assign weights to empty intervals
- smoothing effect again: generally extend the actual range of the observations

$\Rightarrow$ Context is needed

Adding a rug to the plot



Complex process

Chi and Riedl's model

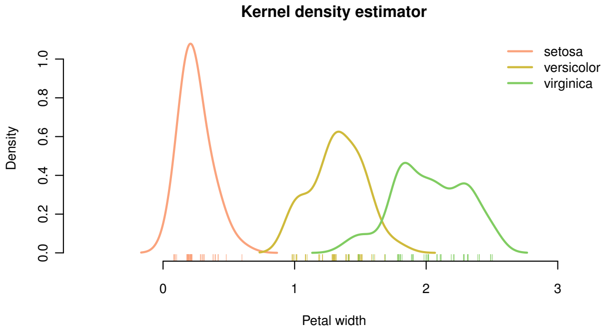
data	$(X_{ij})_{1\leq i\leq N}$ for a fixed j
analytical abstraction	a kernel density estimate of $p(x)$ and the data themselves
visual abstraction	a function $x \mapsto p(x)$
view	a function representation with optimal aspect ratio (computed via the median of the slopes of p), together with a rug obtained from $(X_{ij})_{1\leq i\leq N}$ (maybe with some added jitter)

Added value over histograms

- smooth estimator
- no border effects ($]a, b]$ or $[a, b[$...)
- less non data ink
- easier comparison between distribution

Conditional distribution

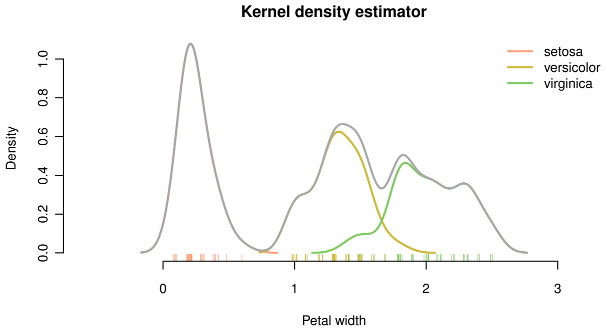
- estimation of $p(x|y)$ for a discrete y
- draw $p(x|y)p(y)$ to ease superposition with the global distribution



Conditioning
Comparing distributions

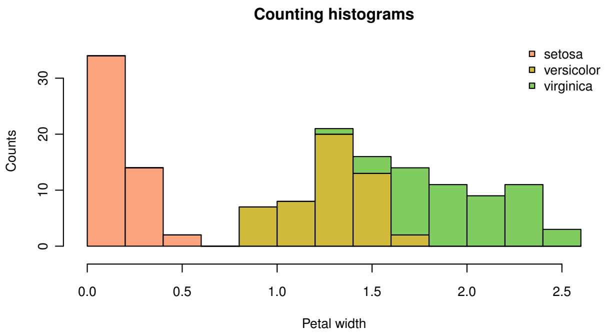
Conditional distribution

- estimation of $p(x|y)$ for a discrete y
- draw $p(x|y)p(y)$ to ease superposition with the global distribution



Conditional distribution

- estimation of $p(x|y)$ for a discrete y
- draw $p(x|y)p(y)$ to ease superposition with the global distribution



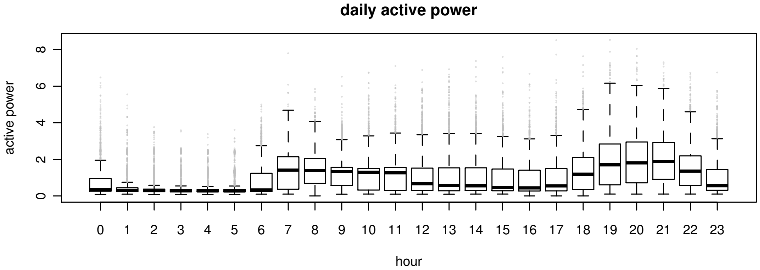
Scaling the comparison

Density estimator limitations

- only a small number of color codes can be distinguished quickly (less than 10)
- overlapping occurs also quickly

Boxplots are more scalable, up to ordering issues

An example



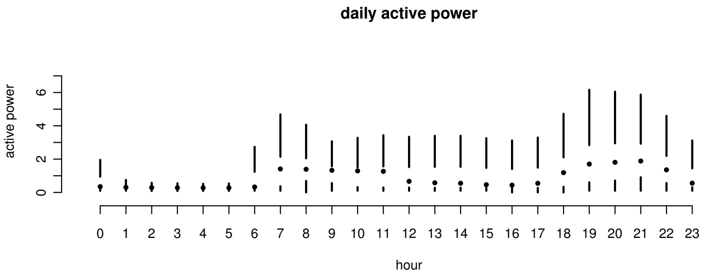
Scaling the comparison

Density estimator limitations

- only a small number of color codes can be distinguished quickly (less than 10)
- overlapping occurs also quickly

Boxplots are more scalable, up to ordering issues

An example



Interaction

Brushing and linking

- brushing induces conditioning (directly or indirectly)
- the conditional data should be superposed to or displayed close to the original data
- boxplots are easy to superpose and/or to display side by side
- density plots also: additionally, multiple brushing is possible (contrarily to histograms)

Other actions

- interactive bandwidth selection
- boxplots positioning (could be automated)

Outline

Moving to “higher” dimensions

Mixed case

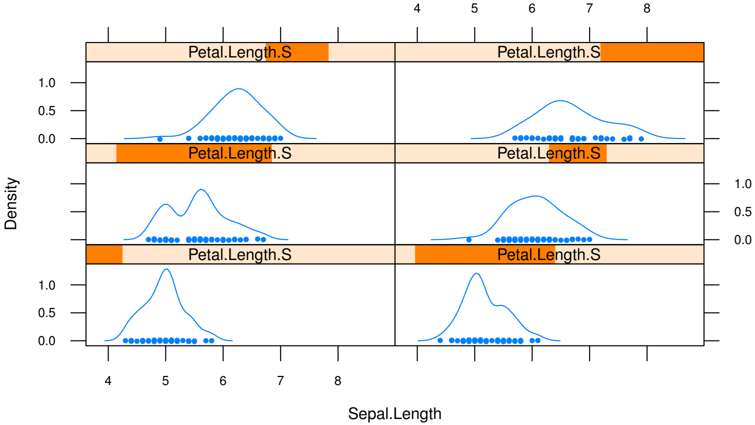
- one numerical variable, one categorical variable
- done by conditioning!
- density plots for less than 10 modalities, boxplots for more

Numerical case

- classical solution: scatter plot (to be discussed)
- less classical solution: discretize one variable
- *shingles*: overlapping intervals that span the discretized variable
- back to the mixed case

A trellis plot

Numerical case with shingles



Categorical variables

Stacked bar charts

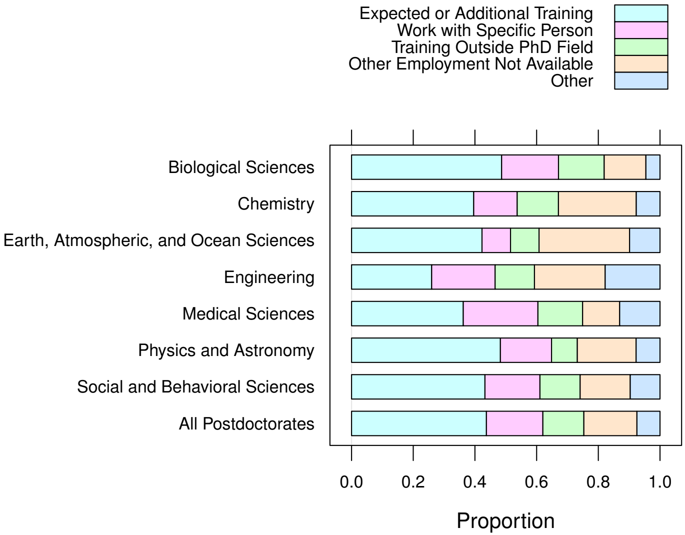
- conditional approach
- use spinograms: display $P(U|V)$ as a bar of length one with sub-bars of length proportional to the conditional probability of the modalities of U
- stack the spinograms

Mosaic plot

- recursive spinograms
- compute Q bars of width proportional to the probabilities of V
- split each bar in sub-bars with heights proportional to the conditional probabilities of U given V

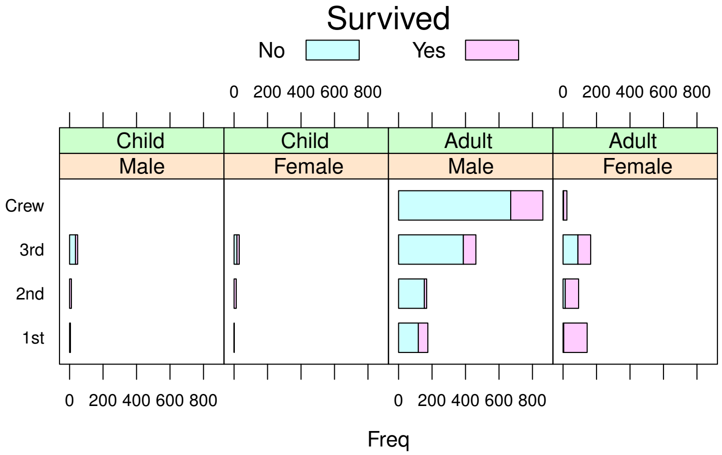
Categorical variables

Stacked bar charts



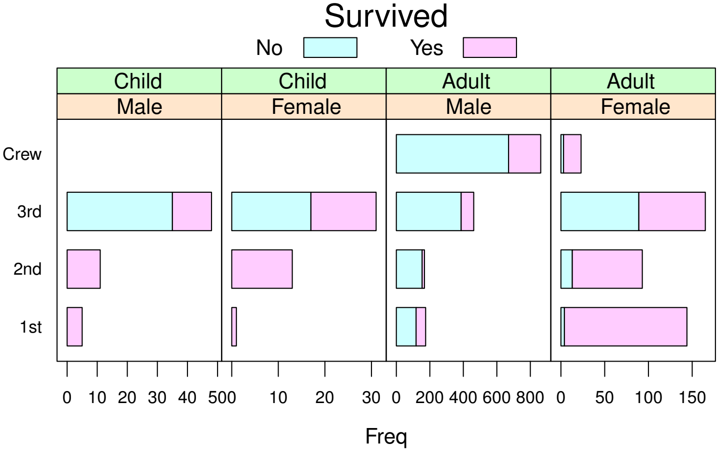
A complex example

The Titanic

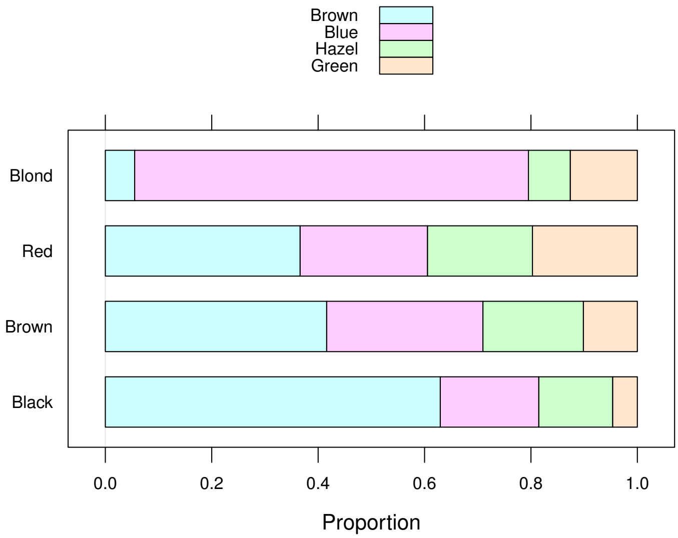


A complex example

The Titanic

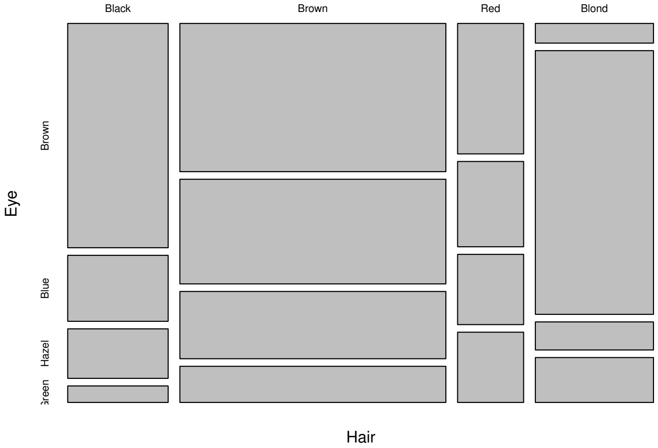


Mosaic plot

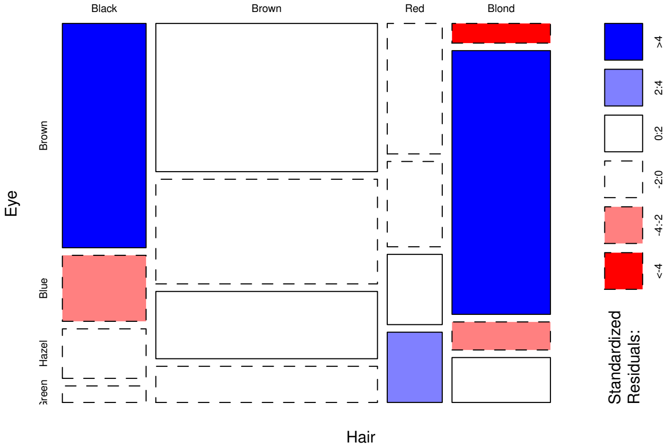


Brown
Blue
Hazel
Green
Blond
Red
Brown
Black
0.0
0.2
0.4
0.6
0.8
1.0
Proportion

Mosaic plot

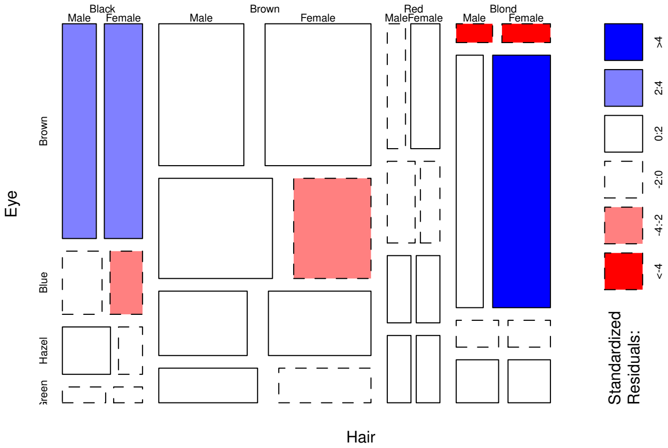


Mosaic plot



Black
Brown
Red
Blond
Eye
Brown
Blue
Hazel
Green
Hair
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized
Residuals:

Mosaic plot in N dim



Mosaic plots

Built-in quality assessment

- mosaic plots show both the data and some statistical indicators
- this limits the risk of false interpretation
- should be used whenever possible
- remains an open challenge for most visualization methods

Interactivity

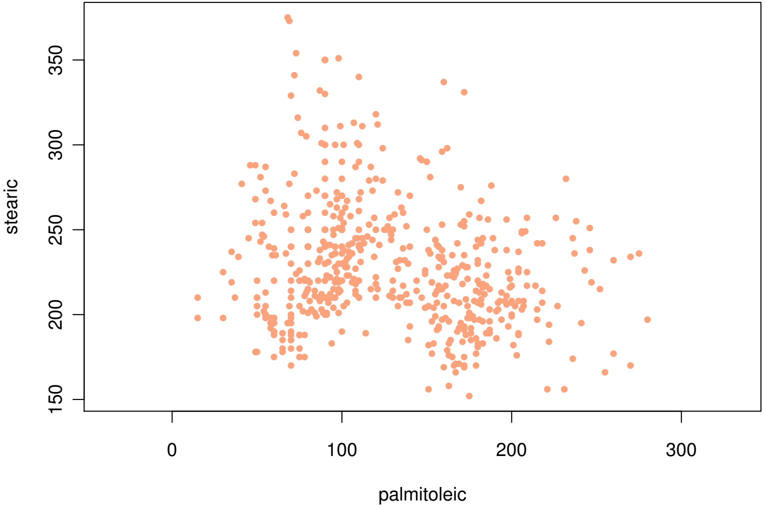
- show information attached to each cell (size, value of the residual)
- change the splitting order
- add or remove variables
- brush and link

Scatter plot

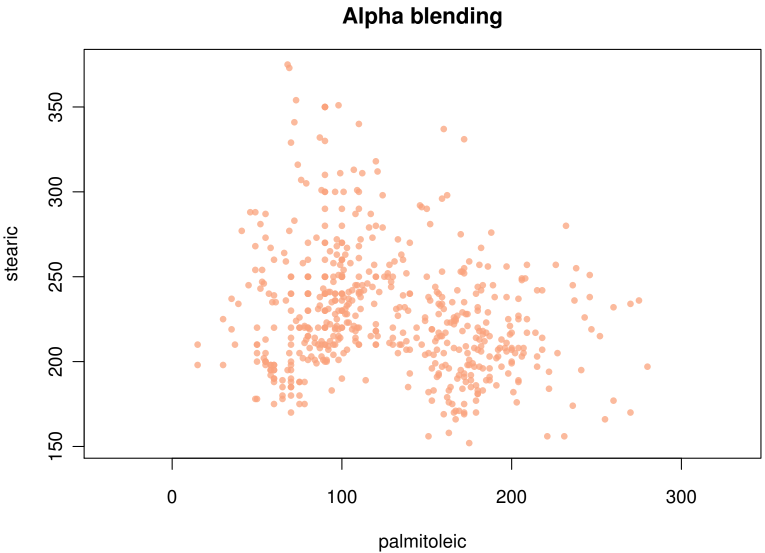
The standard tool for 2 numerical variables

- one point per object
- two dominant numerical characteristics per object
- a few additional characteristics:
  - a nominal variable (hue or shape coded)
  - a numerical variable (lightness or shape coded)
  - a label
- does not scale:
  - low dimension only
  - very sensitive to overlapping

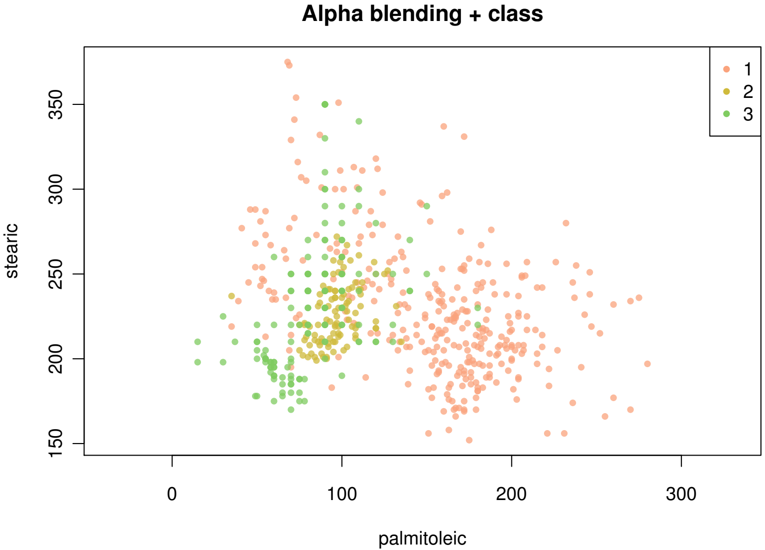
Simple example



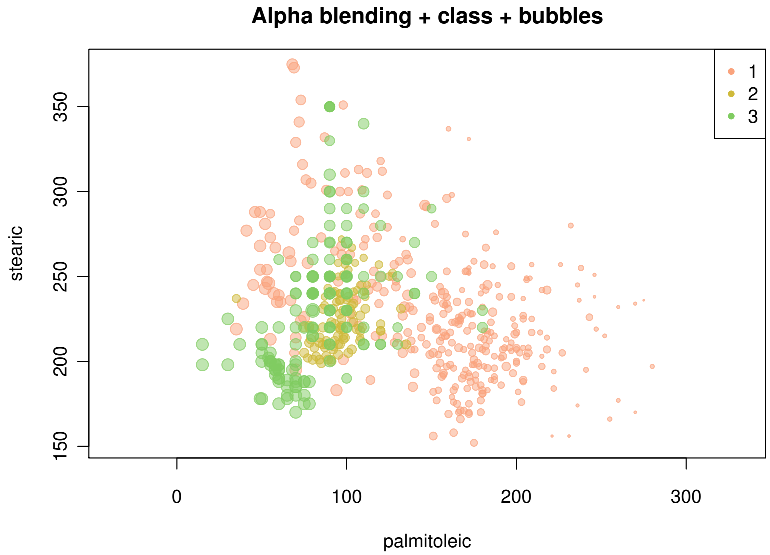
Simple example



Simple example



Simple example



How to improve scalability?

Two scalability problems

1. Dimension (a.k.a. variable or characteristics or feature):
   - Reduce the number of dimension:
     - user choice
     - automated: selection and extraction
   - Display more than 2/3 dimensions at once: visual layout with brushing and linking
2. Object:
   - Reduce the number of objects: clustering and quantization
   - Reduce the size of an object
   - Constrain the display to forbid (or reduce) overlapping

Automated scalability

Reducing the number of variables

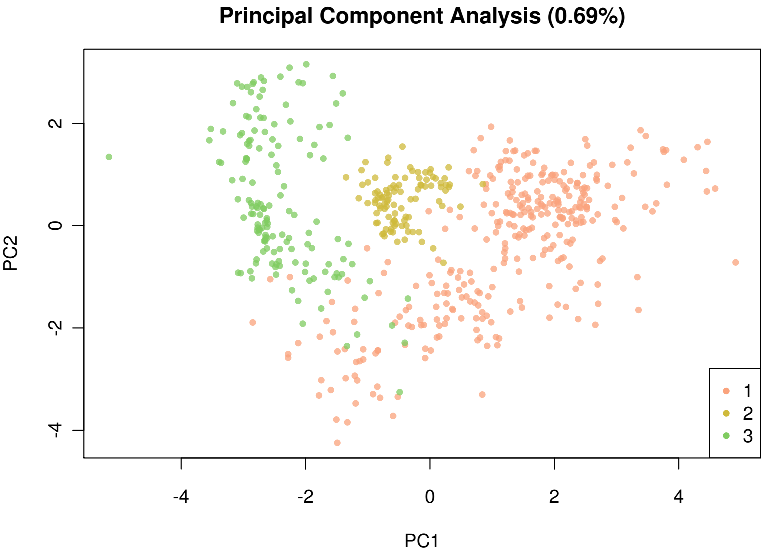
- a major research topic in machine learning
- feature selection: keeping a subset of the original variables
- feature extraction: producing new variables in small quantity
- supervised: variables should “explain” a target variable
- or unsupervised: variables optimize some quality criterion
- some methods:
  - Principal Component Analysis
  - Linear Discriminant Analysis
  - SNE and variants
  - Variable Clustering
  - etc.

Difficulties

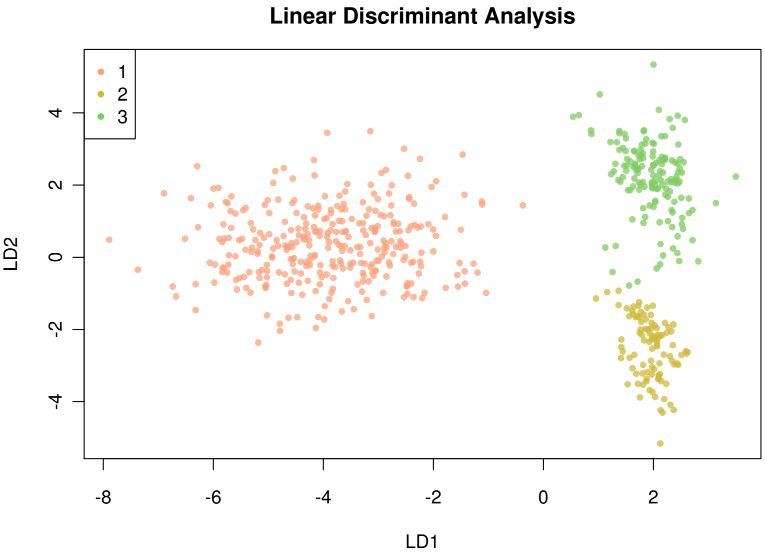
What should be optimized?

- goal dependent:
  - outlier detection $\Rightarrow$ maximize distances between outliers and central objects
  - visual clustering $\Rightarrow$ respect the neighborhood relationships
  - rule finding $\Rightarrow$ keep original variables
  - etc.
- machine learning algorithms optimize abstract quantities (e.g., reconstruction error)
- links between ML optimality and visual usefulness is unclear (at best!)

(Un)Supervised



(Un)Supervised



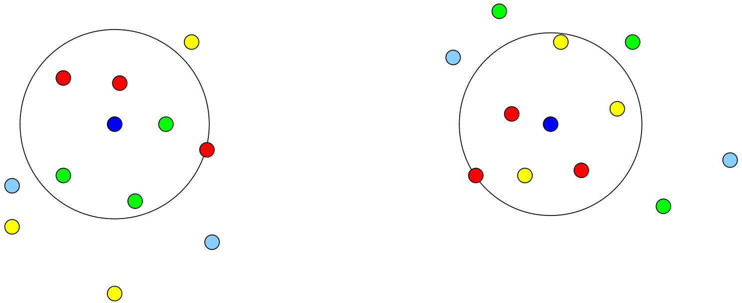
Neighborhood structure preservation

To enable visual cluster analysis on projected dataset, the neighborhood structure of the data must be preserved by the projection.

- Quality measures (Venna & Kaski, 2001 $\rightarrow$ 2007):
  - Trustworthiness: neighbors on the screen are real neighbors
  - Continuity: real neighbors are neighbors on the screen
  - Precision $\simeq$ Trustworthiness
  - Recall $\simeq$ Continuity
- Optimization methods:
  - Stochastic Neighbor Embedding (Hinton & Roweis, 2002)
  - Neighbor Retrieval Visualizer (Venna & Kaski, 2006)

Neighbor Preservation

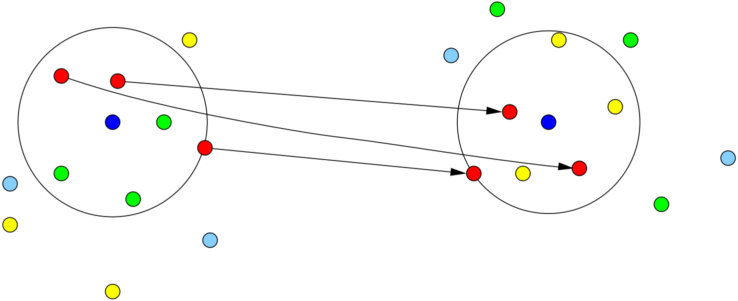
Original space to projection space



6 neighbors

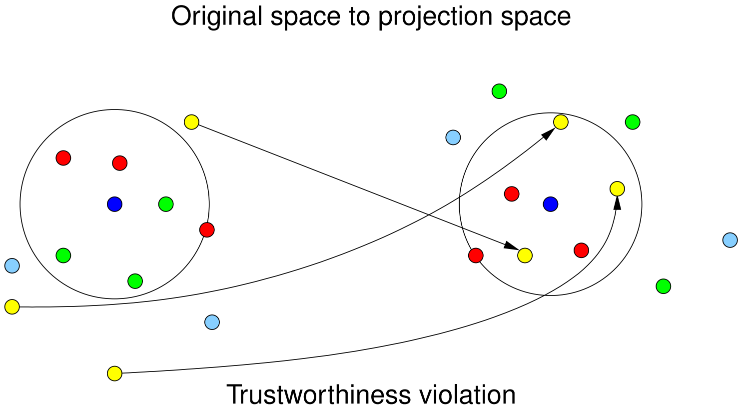
Neighbor Preservation

Original space to projection space

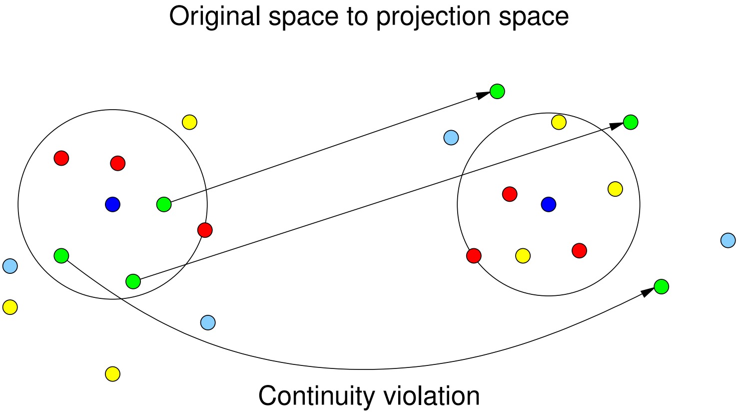


Correct projection

Neighbor Preservation



Neighbor Preservation

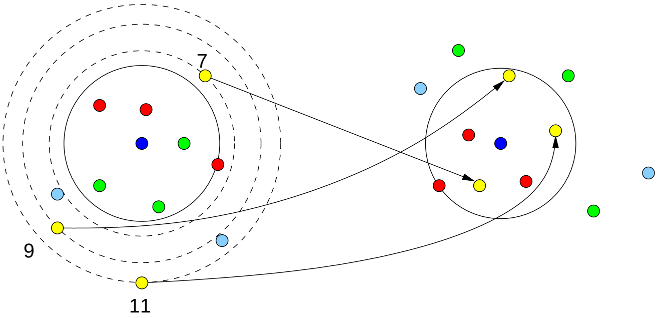


Trustworthiness and Precision

Can you trust neighbors in the projection space?

- $O_k(x_i)$: k-nn of x_i in the original space
- $P_k(x_i)$: k-nn of x_i in the projection space
- $U_k(x_i) = P_k(x_i) \setminus O_k(x_i)$
- Precision
  - maximal precision: $P_k(x_i) \subset O_k(x_i)$
  - mean on i of $1 - \frac{\#U_k(x_i)}{\#P_k(x_i)}$
- Trustworthiness
  - rank preservation: $r^O(x_j, x_i)$ rank of x_j as a neighbor of x_i in the original space
  - $M_1(k) = 1 - \frac{2}{Nk(2N-3k-1)} \sum_{i=1}^{N} \sum_{x_j \in U_k(x_i)} \left(r^O(x_j, x_i) - k \right)$

Trustworthiness and Precision



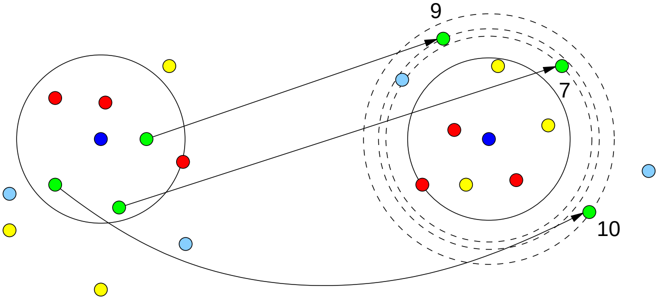
- Contribution to precision: 0.5
- Contribution to trustworthiness: 9

Continuity and Recall

Do you miss neighbors in the projection space?

- $O_k(x_i)$: k-nn of x_i in the original space
- $P_k(x_i)$: k-nn of x_i in the projection space
- $V_k(x_i) = O_k(x_i) \setminus P_k(x_i)$
- Recall
  - maximal recall: $O_k(x_i) \subset P_k(x_i)$
  - mean on i of $1 - \dfrac{\#V_k(x_i)}{\#O_k(x_i)}$
- Continuity
  - rank preservation: $r^P(x_j, x_i)$ rank of x_j as a neighbor of x_i in the projection space
  - $M_2(k) = 1 - \dfrac{2}{Nk(2N - 3k - 1)} \displaystyle\sum_{i=1}^{N} \sum_{x_j \in V_k(x_i)} \left(r^P(x_j, x_i) - k \right)$

Continuity and Recall



- Contribution to recall: 0.5
- Contribution to continuity: 8

Stochastic Neighbor Embedding (SNE)

Approximating the “neighborhood distribution” (Hinton & Roweis, 2002)

- Original space: $p_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq j} \exp(-d_{ik}^2)}$ (d_{ij} dissimilarity between x_i and x_j)
- x_i projected to y_i
- Projection space: $q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq j} \exp(-\|y_i - y_k\|^2)}$
- Minimize the Kullback-Leibler divergence between $p_{i.}$ and $q_{i.}$, i.e.,

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- Corresponds to a smoothed version of the recall (Venna & Kaski, 2006)

Neighbor Retrieval Visualizer

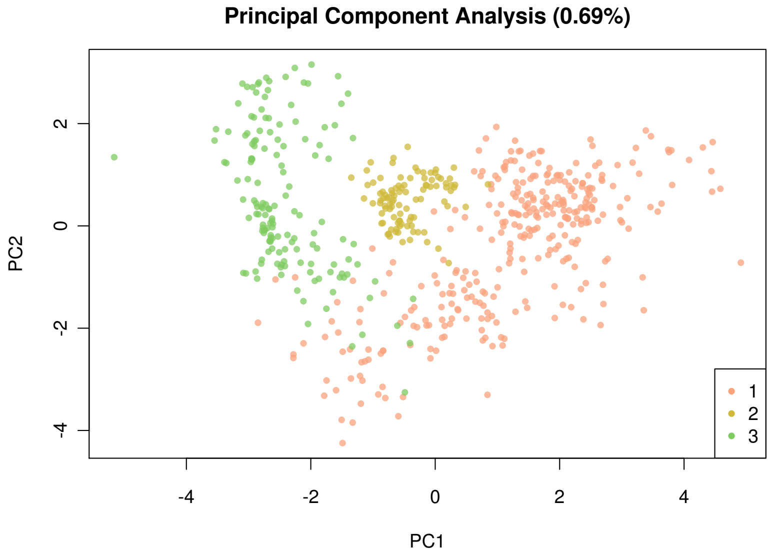
Optimizing smoothed versions of precision and recall (Venna & Kaski, 2006)

- inspired by stochastic neighbor embedding (SNE)
- minimization of

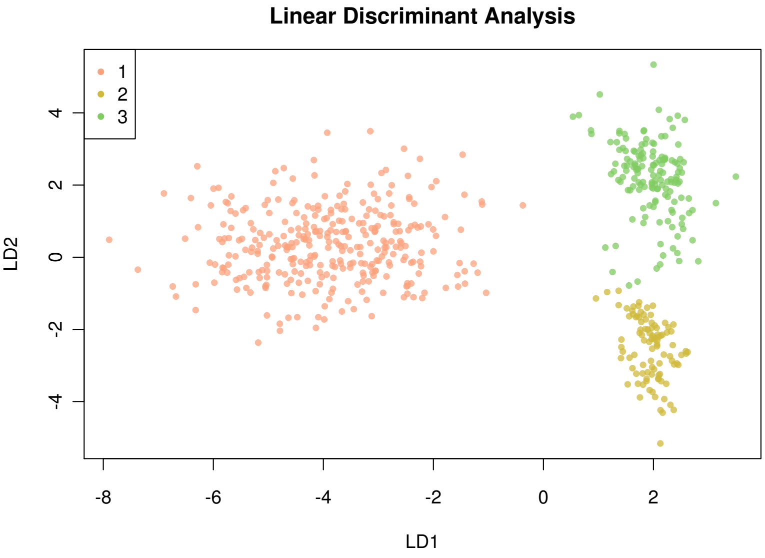
$$\lambda \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1-\lambda) \sum_i \sum_j q_{ij} \log \frac{q_{ij}}{p_{ij}}$$

- $\lambda = 1$ (SNE) $\Rightarrow$ recall
- $\lambda = 0 \Rightarrow$ precision
- λ is user chosen (can be chosen to optimize a quality measure)
- slow ($O(N^3)$ per iteration for N objects, as SNE)

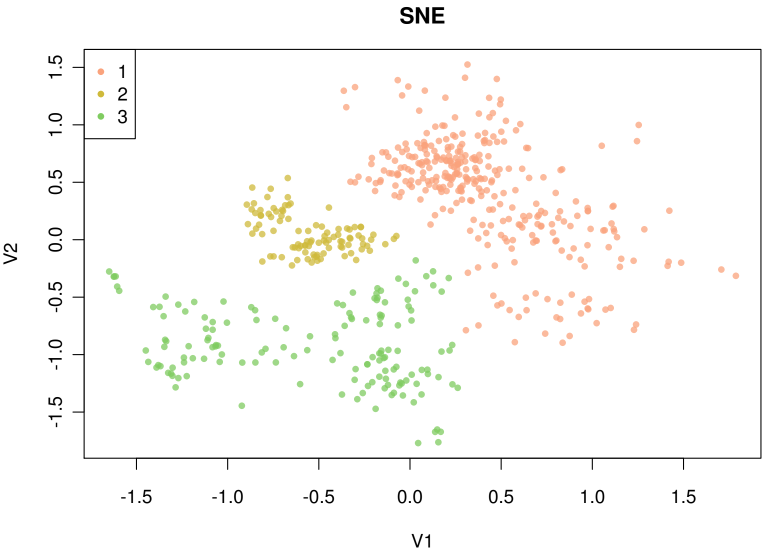
Demonstration



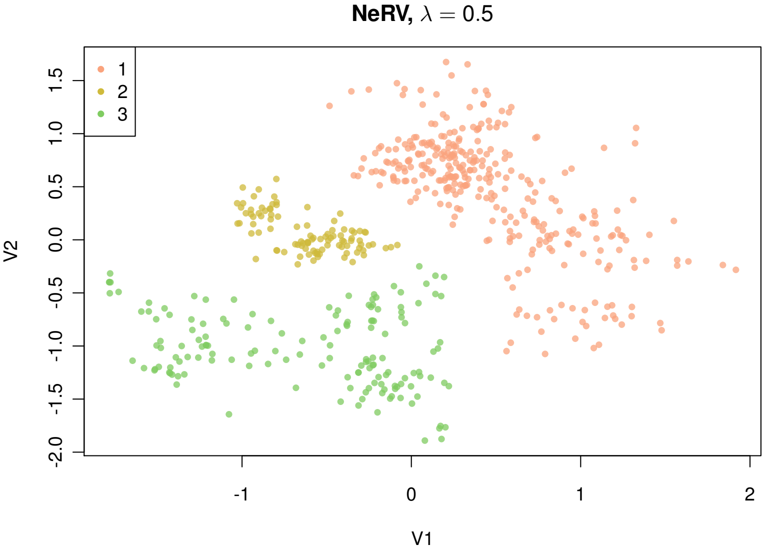
Demonstration



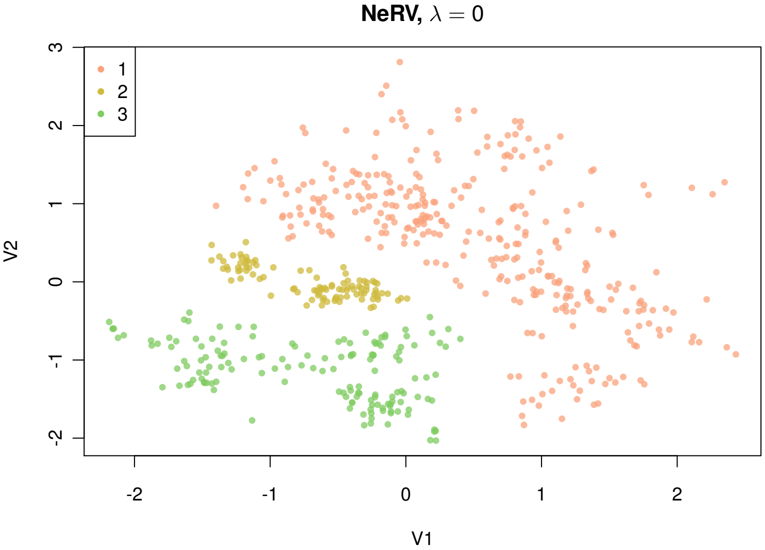
Demonstration



Demonstration



Demonstration



Dimensionality reduction

Status

- very active field
- numerous methods, improving results
- regular breakthroughs: non linearity, emphasizis on small distances, importance of ranks, etc.

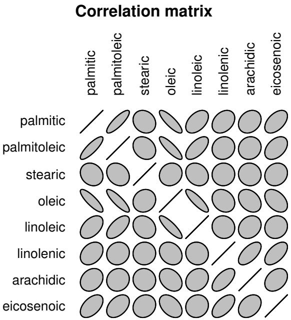
Limitations

- slow methods (nowhere near realtime on realistic datasets)
- highly nonlinear methods:
  - no axis
  - nothing is uniform
- built-in quality assessment is a work in progress
- interactivity?

Picking dimensions manually

Interactive solution

- show hints of possible interesting pairs
- let the user choose



Reducing the overlapping problem

An open research topic

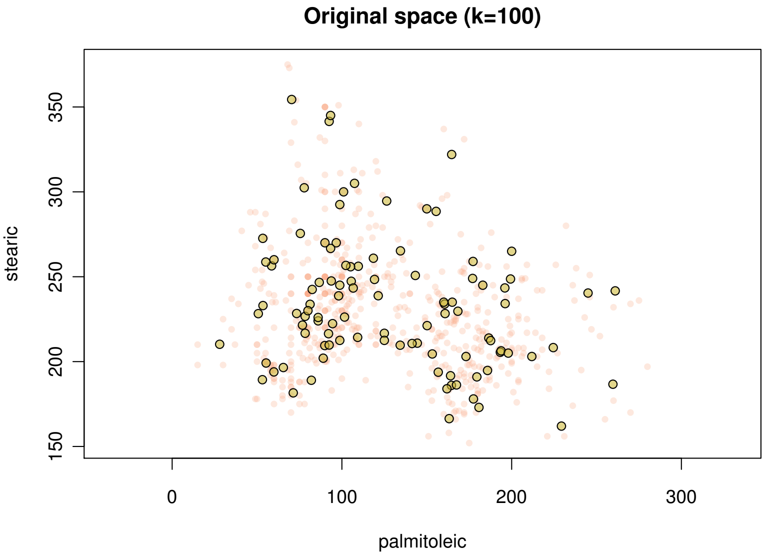
- rendering:
  - transparency (alpha layer)
  - stereo vision
  - overlap counting
- interactivity:
  - sub-sampling
  - zooming and panning
  - excentric labeling
  - magic lenses
  - distortion
- machine learning and related methods:
  - clustering
  - force field approach

Clustering

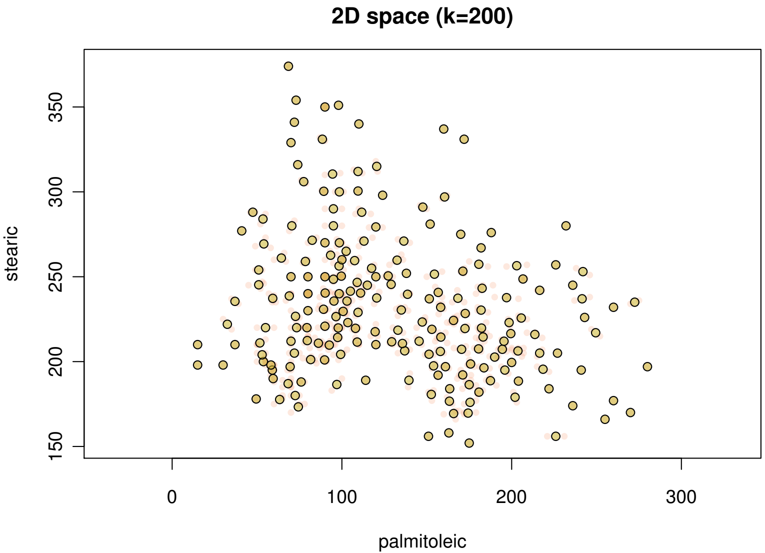
Reduction of the number of objects

- very general solution (not limited to scatter plot)
- clustering $\Rightarrow$ Cluster prototype $\Rightarrow$ Prototype visualization
- interactivity:
  - number of clusters:
    - deterministic clustering
    - minimal surprise principle (e.g., hierarchical)
  - cluster extension (objects or summary)
- important warning:
  - clustering in 2D $\neq$ clustering the original data
  - guaranteed non overlapping: 2D!

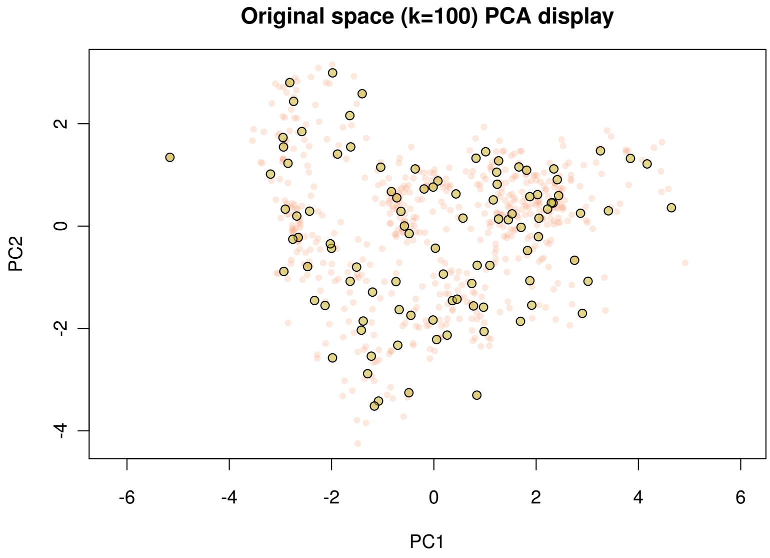
Where to cluster?



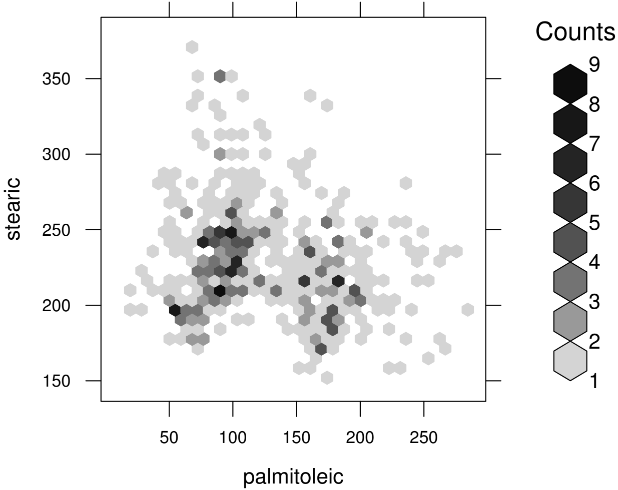
Where to cluster?



Where to cluster?

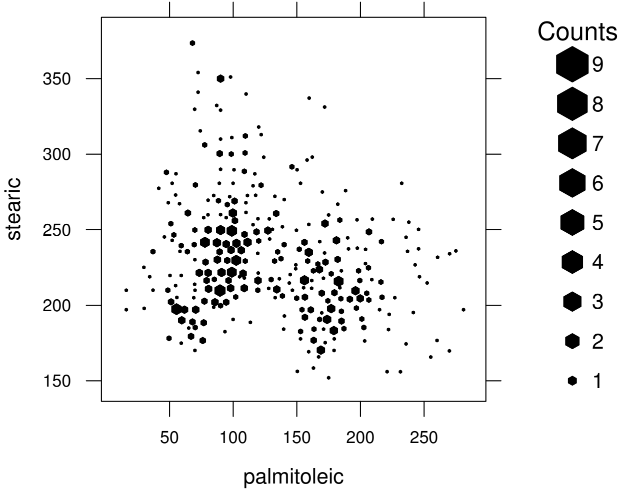


2D density



350
300
250
200
150
stearic
50
100
150
200
250
palmitoleic
Counts
9
8
7
6
5
4
3
2
1

2D density


stearic
350
300
250
200
150
50
100
150
200
250
palmitoleic
Counts
9
8
7
6
5
4
3
2
1

Scatter plot

The workhorse of numerical visualization

- very efficient representation for 2 numerical variables (position on a common scale, best feature according to Cleveland and McGill)
- hue can be used to display an additional categorical variable
- luminance or shape or symbol size can encode another variable (less efficiently)
- overlapping reduces the encoding efficient back to 2 variables

Scalability issues

- are mostly unsolved...
- should be task oriented!

Bibliography I

[CR98] Ed H. Chi and John T. Riedl. An operator interaction framework for visualization systems. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'98)*, pages 63–70, Research Triangle Park, North Carolina, USA, October 1998.