About **Visualization in Bergen.no** and **Interactive Visual Analysis**

Helwig Hauser  
University of Bergen

**In the Following**

1. Briefly about visualization in Bergen, Norway
2. Interactive Visual Analysis (IVA)
3. High-dimensional Data IVA
HH: prof. in visualization (vis)
@ Dept. of Informatics (ii)
@ Univ. of Bergen (UiB)
in Bergen, Norway (.no)

UiB VisGroup
– 2007: group of 3:
– 2009: larger projects start
– 2011: EuroVis in Bergen
– 2013: new prof:

[ranking from NFR’s 10-year evaluation in 2011/2012]

ii.UiB.no/vis Research

➢ Application-oriented basic research in visualization:

1. Researched visualization methodology (how to visualize)
   ➢ Interactive Visual Analysis, nD data (H. Hauser et al.)
   ➢ Visual Knowledge Discovery, 3D data (St. Bruckner et al.)
   ➢ Illustrative Visualization (I. Viola et al.)

2. Applications at which this research is oriented (for whom)
   ➢ Medical Visualization (partner in MedViz Bergen, etc.)
   ➢ GeoSciences / Oil & Gas (e.g., financed by Statoil’s Akademiaavtale)
   ➢ Biology / Bioinformatics (with CBU@ii et al.)
   ➢ Fluid Dynamics (in collab. with FFI.no, for ex.)
   ➢ Engineering (visual analysis of simulation data)
**ii.UiB.no/vis Team**

Two profs. (HH, StBr) and PostDocs, PhD studs., *et al.*

<table>
<thead>
<tr>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Ivan Viola -----------------------------------------------
  - Stefan Bruckner
  - Július Parulek ---------------------------------------------

- Daniel Patel ----------------------------------------------
- Jean-Paul Balabanian -------------------------------------
  - Paolo Angelelli ------------------------------------------
  - Ove Daae Lampe ----------------------------------------
  - Johannes Kehrer ----------------------------------------

- Åsmund Birkeland -----------------------------------------
- Armin Pobitzer -------------------------------------------
  - Veronika Šoltészová ---------------------------------

- Endre Lidal ---------------------------------------------
- Çağatay Turkay ------------------------------------------
  - Andrea Brambilla -------------------------------------
  - Paolo Angelelli --------------------------------------

- Ivan, Andreas --

---

**ii.UiB.no/vis PhDs** (10 so far)

- Daniel Patel (Oct. 2009): Expressive Vis. & Rapid Interpr. of Seismic Volumes
- Jean-Paul Balabanian (Jan. 2010): Multi-Aspect Vis.: from Linked to Integrated Views
- Johannes Kehrer (May 2011): IVA of Multi-faceted Scientific Data
- Ove Daae Lampe (Nov. 2011): IVA of Process Data
- Armin Pobitzer (June 2012): IVA of Time-dependent Flows
- Paolo Angelelli (June 2012): Visual Expl. of Human Physiology
- Åsmund Birkeland (May 2013): Ultrasonic Vessel Vis.: From Extraction to Perception
- Endre Lidal (May 2013): Sketch-based Storytelling for Cognitive Problem Solving
- Çağatay Turkay (Nov. 2013): Interactive Visual Analysis of High-dimensional Data
Interactive Visual Analysis (IVA)

- Given data – *too much* and/or *too complex* to be shown at once:
- IVA is an **interactive visualization approach** to facilitate
  - the **exploration** and/or the **analysis** of data (not necessarily the presentation of data), including
    - **hypothesis generation & evaluation**, **sense making**, **knowledge crystallization**, *etc.*
  - according to the user’s interest/task, *for ex.*, by interactive feature extraction,
  - navigating between **overview** and **details**, *e.g.*, to enable interactive information drill-down [Shneiderman]
- through an **iterative & interactive visual dialog**

Interactive Visual Analysis ↔ Visual Analytics

- **IVA** (“interactive visual analysis”) **since 2000**
- **Tightly related** to **visual analytics**, of course, *e.g.*, **integrating computational & interactive data analysis**
- **A particular methodology** with specific components (*CMV*, *linking & brushing*, *F+C vis.*, *etc.*)
- **General enough to work in many application fields**, but not primarily the VA fields (national security, *etc.*), in particular “**scientific data” fields**…
Integrating Interaction & Computation

- **Goal**: to combine the *best of two worlds* [Keim et al.]:
  - data *exploration/analysis* by the user, based on interactive visualization
  - and data *analysis* by the computer, based on statistics, machine learning, etc.

- State of the art / levels of integration:
  - mostly *no integration*, still
  - some *vis. of results* of computations
  - also: making comp. *semi-interactive* (here called “inner integration”)
  - rare: *tight integration*

- **Outer integration** (here!):
  bundling interaction & computation in a loop

Target Data Model: “Scientific Data”

- **Characterized** by a combination of
  - independent variables, like space and/or time (cf. domain)
  - and dependent variables, like pressure, temp., etc. (cf. range)

- So we can think of this type of data as given as \(d(x)\) with \(x \leftrightarrow \text{domain}\) and \(d \leftrightarrow \text{range}\) – examples:
  - CT data \(d(x)\) with \(x \in \mathbb{R}^3\) and \(d \in \mathbb{R}\)
  - unstead 2D flow \(v(x,t)\) with \(x \in \mathbb{R}^2\), \(t \in \mathbb{R}\), and \(v \in \mathbb{R}^2\)
  - num. sim. result \(d(x,t)\) with \(x \in \mathbb{R}^3\), \(t \in \mathbb{R}\), and \(d \in \mathbb{R}^n\)
  - system sim. \(q(p)\) with \(p \in \mathbb{R}^n\) and \(q \in \mathbb{R}^m\)

- **Common property**:
  - \(d\) is (at least to a certain degree) *continuous* wrt. \(x\)
Interactive Visual Analysis of Scientific Data

- **Interactive visual analysis** (as exemplified in this tutorial) works really well with scientific data, e.g.,
  - results from numerical simulation (spatiotemporal)
  - imaging / measurements (in particular multivariate)
  - sampled models

- When used to study scientific data, **IVA employs**
  - methods from scientific visualization (vol. rend., …)
  - methods from statistical graphics (scatterplots, …), information visualization (parallel coords., etc.)
  - computational tools (statistics, machine learning, …)

- Applications include
  - engineering, medicine, meteorology/climatology, biology, etc.

The Iterative Process of IVA

- Loop / bundling of two complementary parts:
  - visualization – show to the user!
    *Something new, or something due to interaction.*
  - interaction – tell the computer!
    *What is interesting? What to show next?*

- Basic example (**show – brush – show – …**), cooling jacket context:
  1. show a histogram of temperatures
  2. brush high temperatures (>90°[±2°])
  3. show focus+context vis. in 3D
  4. locate relevant feature(s)

- **KISS-principle IVA:**
  - linking & brushing, focus+context visualization, …
Show & Brush

**Tightest IVA loop**
- **show data** (explicitly represented information)
- **one brush** (on one view, can work on >1 dims.)

**Requires:**
- multiple views (≥2)
- interactive brushing capabilities on views (brushes should be editable)
- focus+context visualization
- linking between views

**Allows for different IVA patterns** (wrt. domain & range)

---

**A typical IVA session** of this kind:
- bring up multiple views
  - at least one for $x, t$
  - at least one for $d_i$
  - I see (something)!
  - brush this “something”
  - linked F+C visualization
  - first insight!

---

... leads to...
... requires...
... is realized via...

degree of interest
IVA: Multiple Views

- One dataset, but multiple views
- Scatterplots, histogram, 3D(4D) view, etc.

Interactive Brushing

- Move/alter/extend brush interactively
- Interactively explore/analyze multiple variates
Interactive Brushing

- Move/alter/extend brush interactively
- Interactively explore/analyze multiple variates

[Doleisch et al., '03]
IVA: Focus+Context Visualization

- Traditionally space distortion
  - more space for data of interest
  - rest as context for orientation
- Generalized F+C visualization
  - emphasize data in focus (color, opacity, ...)
  - differentiated use of visualization resources

F+C Visualization in IVA Views

- Colored vs. gray-scale visualization
- Opaque vs. semi-transparent visualization

In a scatterplot (left) or histogram (right): brushed data in red

[Matković et al., '09]
F+C Visualization in IVA Views

[F+ Novotny & Hauzer, ’06]

In parallel coordinates (above): brushed data in red & over

[Muigg et al., ’07]

In 3D (above): less transp. & colored, in illustrative context
IVA: Linked Views

- Brushing: mark data subset as especially interesting
- Linking: enhance brushed data in linked views consistently (F+C)

[Doleisch & Hauser, '02]
IVA: Degree of Interest (DOI)

- **doi(.)**: data items $tr_i$ (table rows) $\rightarrow$ degree of interest
  - $doi(tr_i) \in [0,1]$
  - $doi(tr_i) = 0 \Rightarrow tr_i$ not interesting ($tr_i \in \text{context}$)
  - $doi(tr_i) = 1 \Rightarrow tr_i$ 100% interesting ($tr_i \in \text{focus}$)

- **Specification**
  - explicit, e.g., through direct selection
  - implicit, e.g., through a range slider

- Fractional DOI values: $0 \leq doi(tr_i) \leq 1$
  - several levels (0, low, med., …)
  - a continuous measure of interest
  - a probabilistic definition of interest

(continuation on next slide)

IVA: Smooth Brushing $\rightarrow$ Fractional DOI

- **Fractional DOI values** esp. useful wrt. scientific data: (quasi-)continuous nature of data $\leftrightarrow$ smooth borders

- Goes well with gradual focus+context vis. techniques (coloring, semitransparency)

- **Specification**: smooth brushing [Doleisch & Hauser, 2002]
  - “inner” range: all 100% interesting (DOI values of 1)
  - between “inner” & “outer” range: fractional DOI values
  - outside “outer” range: not interesting (DOI values of 0)
Fuzzy Classification

DOI \([0,1]\), 0… not interesting, 1… 100% interesting

Requires fuzzy logic for combination, we use
\[
c = a \frac{1}{b} \frac{c}{c} = \min(a, b)
\]
\[
c = a \frac{b}{c} = \max(a, b)
\]
\[
c = \frac{a}{c} = 1
\]

Matches the smooth nature of the data

Goes well with F+C visualization, e.g.,
 opacity varies gradually with DOI

[Doleisch & Hauser 2002]
Three Patterns of SciData IVA

- Preliminary: domain $x$ & range $d$ visualized ($\geq 2$ views)

1. **Brushing on domain visualization**, e.g., brushing special locations in the map view
   - "x"
   - "... from $x$ to $d$ ...", local investigation

2. **Brushing on range visualization**, e.g., brushing outlier curves in a function graph view
   - "d"
   - "... from $d$ to $x$ ...", feature localization

3. Relating multiple range variates
   - "d"
   - "... within $d$ ...", multi-variate analysis

IVA – Levels of Complexity

- A lot can be done with basic IVA, already! [pareto rule]
- We can consider a layered information space: from explicitly represented information (the data) to implicitly contained information, features, ...

show & brush

<table>
<thead>
<tr>
<th>temp.</th>
<th>vel.</th>
<th>...</th>
<th>data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>between the lines...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>buried deeper...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>features in application terms</td>
</tr>
</tbody>
</table>

??
A lot can be done with basic IVA, already! [pareto rule]

For more advanced exploration/analysis tasks, we extend it (in several steps):

- IVA, level 2: **logical combinations of brushes, e.g.,** utilizing the **feature definition language** [Doleisch et al., 2003]
- IVA, l. 3: **attribute derivation; advanced brushing,** with interactive formula editor; e.g., similarity brushing
- IVA, l4: **application-specific feature extraction, e.g.,** based on vortex extraction methods for flow analysis

Level 2: like **advanced verbal feature description**
- ex.: “hot flow, also slow, near boundary” (cooling j.)
- brushes comb. with **logical operators** (AND, OR, SUB)
- in a **tree,** or **iteratively** (((b_0 op_1 b_1) op_2 b_2) op_3 b_3) …
IVA – Levels of Complexity

- A lot can be done with basic IVA, already! [Pareto rule]
- For more advanced exploration/analysis tasks, we extend it (in several steps):
  - IVA, level 2: logical combinations of brushes, utilizing the feature definition language [Doleisch et al., 2003]
  - IVA, l. 3: attribute derivation; advanced brushing, with interactive formula editor; e.g., similarity brushing
  - IVA, l. 4: application-specific feature extraction, e.g., based on vortex extraction methods for flow analysis
- Level 3: using general info extraction mechanisms, two (partially complementary) approaches:
  1. derive additional attribute(s), then show & brush
  2. use an advanced brush to select “hidden” relations
IVA (level 3): Advanced Brushing

- **Std. brush**: brush 1:1 what you see
- **Adv. brush**: executes additional function (“intelligent”?)

**Examples:**
- angular brushing [Hauser et al., 2002]
- similarity brushing [Muigg et al., 2008]

**3rd level IVA, adv. brushing example**

- Considering a visualization of a family of function graphs:
  - select the steeply rising graphs

example prepared by Konyha, Zoltán
A simple line brush is not enough

Combining line brushes does not work, either

feature of interest: not explicitly available

3rd level IVA, adv. brushing example

The \textit{angular line brush} (a specialized brush) selects the intended function graphs

- that it intersects, and
- the angle is in a given threshold
IVA (level 3): Attribute Derivation

- **Principle** (in the context of iterative IVA):
  - see some data feature $\Phi$ of interest in a visualization
  - identify a mechanism $T$ to describe $\Phi$
  - execute (interactively!) an attribute derivation step to represent $\Phi$ explicitly (as new, synthetic attribute[s] $d_\Phi$)
  - brush $d_\Phi$ to get $\Phi$

- **Tools** $T$ to describe $\Phi$ from:
  - numerical mathematics
  - statistics, data mining
  - etc.
  - scientific computing

- **IVA w/ T ↔ visual computing**

Attribute Derivation ↔ User Task / example

- The tools $T$, available in an IVA system, must reflect/match the **analytical steps of the user**:

- **Example:**
  - **first vis.**:
    - $\leftrightarrow$ user wishes to select the “band” in the middle
  - so?
    - an advanced brush? a lasso maybe?
  - ah!
    - $\rightarrow$ let’s normalize $y$ and then brush (a)

- **leading to the wished selection:**
What user wishes to reflect?

- Many **generic wishes** – users interest in:
  - something **relative** (instead of some absolute values), example: show me the top-15%
  - change (instead of current values), ex.: show me regions with increasing temperature
  - some **non-local property**, ex.: show me regions with high average temperature
  - **statistical properties**, ex.: show me outliers
  - **ratios/differences**, ex.: show me population per area, difference from trend
  - etc.

- **Common characteristic** here:
  - **questions/tools generic**, not application-dependent!

How to reflect these user wishes?

- Many **generic wishes** – users interest in:
  - something **relative** (instead of some absolute values), example: show me the top-15%
  - change (instead of current values), ex.: show me regions with increasing temperature
  - some **non-local property**, ex.: show me regions with high average temperature
  - **statistical properties**, ex.: show me outliers
  - **ratios/differences**, ex.: show me population per area, difference from trend
  - etc.

- **Common characteristic** here:
  - **questions/tools generic**, not application-dependent!
Some useful tools for 3\textsuperscript{rd}-level IVA

- From analysis, calculus, num. math:
  - **linear filtering** (convolve the data with some linear filter on demand, e.g., to smooth, for derivative estimation, etc.)
  - **calculus** (use an interactive formula editor for computing simple relations between data attributes; +, −, ∙, /, etc.)
  - **gradient estimation, numerical integration** (e.g., wrt. space and/or time)
  - **fitting/resampling** via interpolation/approximation

- From statistics, data mining:
  - **descriptive statistics** (compute the statistical moments, also robust, measures of outlyingness, detrending, etc.)
  - **embedding** (project into a lower-dim. space, e.g., with PCA for a subset of the attribs., etc.)

- **Important**: executed on demand, after prev. vis.

3\textsuperscript{rd}-level IVA – Sample Iterations

- The Iterative Process of 3\textsuperscript{rd}-level IVA:
  - Example 1:
    - you look at some temp. distribution over some region
    - you are interested in raising temperatures, but not temperature fluctuations
    - you use a temporal derivate estimator, for ex., central differences
      \( t_{\text{change}} = (t_{\text{future}} - t_{\text{past}}) / \text{len(future-past)} \)
    - you plot \( t_{\text{change}} \), e.g., in a histogram and brush whatever change you are interested in
    - maybe you see some frequency amplification due to derivation, so you go back and
    - use an appropriate smoothing filter to remove high frequencies from the temp. data, leading to a derived new \( t = t_{\text{smooth}} \) data attribute
    - selecting from a histogram of \( t_{\text{change}} \) (computed like above) is then less sensitive to temperature fluctuations
The Iterative Process of 3rd-level IVA:

- Example exploiting PCA:
  - you bring up a scatterplot of $d_1$ vs. $d_2$: (from an ECG dataset [Frank, Asuncion; 2010])
  - obviously, $d_1$ and $d_2$ are correlated, our interest: the data center wrt. the main trend
  - we ask for a (local) PCA of $d_1$ and $d_2$:
  - then we brush the data center
  - we get the wished selection
  - from here further steps are possible..., incl. study of other PCA-results, etc.

From analysis, calculus, num. math:
- linear filtering (convolve the data with some linear filter on demand, e.g., to smooth, for derivative estimation, etc).
- calculus (use an interactive formula editor for computing simple relations between data attributes; $+$, $-$, $\cdot$, $\div$, etc).
- gradient estimation, numerical integration (e.g., wrt. space and/or time).
- fitting / resampling via interpolation / approximation.

From statistics, data mining:
- descriptive statistics (compute the statistical moments, also robust, measures of outlyingness, detrending, etc).
- embedding (project into a lower-dim. space, e.g., with PCA for a subset of the attributes, etc).

**Brushing of Attribute Clouds for the Visualization of Multivariate Data**

Helke Jänicke, Michael Böttinger, and Gerik Scheuermann, Member, IEEE
A lot can be done with basic IVA, already! [Pareto rule]

For more advanced exploration/analysis tasks, we extend it (in several steps):

- IVA, level 2: **logical combinations of brushes** utilizing the **feature definition language** [Doleisch et al., 2003]
- IVA, l. 3: **attribute derivation**; advanced brushing with interactive formula editor; e.g., similarity brushing
- IVA, l4: **application-specific feature extraction** based on vortex extraction methods for flow analysis

Level 4: **application-specific procedures**
- tailored solutions (for a specific problem)
- “deep” information drill-down
- etc.
Interactive Visual Analysis – delivery

- Understanding data \textit{wrt. range d}
  - which distribution has data attribute $d_i$?
  - how do $d_i$ and $d_j$ relate to each other? (\textit{multivariate analysis})
  - which $d_k$ discriminate data features?

- Understanding data \textit{wrt. domain x}
  - \textbf{where} are relevant features? (\textit{feature localization})
  - \textbf{which} values at specific $x$? (\textit{local analysis})
  - how are they related to parameters?

The Iterative Process of IVA...

...is a \textbf{very useful methodology} for data exploration & analysis

...is \textbf{very general} and can be (has already been) applied to \textbf{many different application fields} (in this talk the focus was on scientific data)

...\textbf{meets scientific computing} as a complementary methodology (with the \textbf{important difference} that in IVA the \textbf{user} with his/her perception/cognition is \textbf{in the loop} at different frequencies, also many fps)

...is \textbf{not yet fully implemented} (we’ve done something, e.g., in the context of \textit{SimVis, ComVis, etc.}) – from here: different possible paths, incl. \textit{InteractiveVisualMatlab, IVR, etc.)}
The **Dual Analysis Framework** for **High-dimensional Data IVA**

Çağatay Turkay, Helwig Hauser  
University of Bergen
High- vs. multi-dimensional data

- multi-dimensional: >3D, 4D, 6D, 12D, ..., 24D(?)

- high-dimensional: ..., 40D, 80D, 240D, 1200D, ...
  - std. tools for multi-dim. data vis. don’t work
  - lots of statistics, etc., do not work properly, esp. when #dims. > #items

Where?

- Biology data (e.g., from genomics/proteomics), astronomy data (e.g., spectral imaging data), survey data (many questions), ...

Understanding \( n \)D for (really) large \( n \)

Curse of dimensionality is a problem, when \( n \) large

- \( n \)D distances become meaningless
- with that distances-based project methods
- statistics of wide tables don’t work

Hypothesis:

- there is valuable information in the «space of dimensions»
- the «space of dimensions» is structured, heterogeneous
- it’s worthwhile to understand this «space of dimensions» in order to do a better informed IVA of the data items

But how to understand this «space of dimensions»?

Can we visualize the dimensions of a dataset?
Almost all of visualization is about visualizing the (multi-dimensional) data items.

A new perspective: Dims. Visualization

Alternatively, and esp., when we have so many dims., we could visualize the data dimensions themselves!
Naïve Approach

Transposing the data table should do it, right? :-)

Not really...

---

### Data Table

<table>
<thead>
<tr>
<th>Car</th>
<th>Cylinders</th>
<th>Displacem.[In^3]</th>
<th>Horsepower</th>
<th>Weight [lb]</th>
<th>Accel.[s@0-60 ModelYear]</th>
<th>Origin</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car001</td>
<td>18</td>
<td>307</td>
<td>130</td>
<td>3504</td>
<td>12</td>
<td>70 US</td>
<td>chevrolet chevelle malibu</td>
</tr>
<tr>
<td>Car002</td>
<td>15</td>
<td>350</td>
<td>165</td>
<td>3693</td>
<td>11.5</td>
<td>70 US</td>
<td>buick skylark 320</td>
</tr>
<tr>
<td>Car003</td>
<td>18</td>
<td>318</td>
<td>150</td>
<td>3436</td>
<td>11</td>
<td>70 US</td>
<td>plymouth satellite</td>
</tr>
<tr>
<td>Car004</td>
<td>16</td>
<td>304</td>
<td>150</td>
<td>3433</td>
<td>12</td>
<td>70 US</td>
<td>amc rebel sst</td>
</tr>
<tr>
<td>Car005</td>
<td>17</td>
<td>302</td>
<td>140</td>
<td>3449</td>
<td>10.5</td>
<td>70 US</td>
<td>ford torino</td>
</tr>
<tr>
<td>Car006</td>
<td>15</td>
<td>429</td>
<td>198</td>
<td>4341</td>
<td>10</td>
<td>70 US</td>
<td>ford galaxie 500</td>
</tr>
<tr>
<td>Car007</td>
<td>14</td>
<td>454</td>
<td>220</td>
<td>4354</td>
<td>9</td>
<td>70 US</td>
<td>chevrolet impala</td>
</tr>
<tr>
<td>Car008</td>
<td>14</td>
<td>460</td>
<td>218</td>
<td>4312</td>
<td>8.5</td>
<td>70 US</td>
<td>plymouth fury iii</td>
</tr>
<tr>
<td>Car009</td>
<td>14</td>
<td>460</td>
<td>218</td>
<td>4312</td>
<td>8.5</td>
<td>70 US</td>
<td>pontiac catalina</td>
</tr>
<tr>
<td>Car010</td>
<td>15</td>
<td>500</td>
<td>150</td>
<td>4425</td>
<td>10</td>
<td>70 US</td>
<td>amc ambassador dpl</td>
</tr>
<tr>
<td>Car011</td>
<td>NA</td>
<td>383</td>
<td>175</td>
<td>4166</td>
<td>10.5</td>
<td>70 US</td>
<td>citroen ds-21 pallas</td>
</tr>
<tr>
<td>Car012</td>
<td>NA</td>
<td>383</td>
<td>170</td>
<td>3563</td>
<td>10</td>
<td>70 US</td>
<td>chevrolet chevelle concour</td>
</tr>
<tr>
<td>Car013</td>
<td>NA</td>
<td>383</td>
<td>170</td>
<td>3563</td>
<td>10</td>
<td>70 US</td>
<td>ford torino (sw)</td>
</tr>
<tr>
<td>Car014</td>
<td>NA</td>
<td>383</td>
<td>170</td>
<td>3563</td>
<td>10</td>
<td>70 US</td>
<td>plymouth satellite (sw)</td>
</tr>
<tr>
<td>Car015</td>
<td>NA</td>
<td>383</td>
<td>170</td>
<td>3563</td>
<td>10</td>
<td>70 US</td>
<td>amc rebel sst (sw)</td>
</tr>
<tr>
<td>Car016</td>
<td>15</td>
<td>430</td>
<td>160</td>
<td>3609</td>
<td>8</td>
<td>70 US</td>
<td>dodge challenger se</td>
</tr>
<tr>
<td>Car017</td>
<td>15</td>
<td>430</td>
<td>160</td>
<td>3609</td>
<td>8</td>
<td>70 US</td>
<td>plymouth 'cuda 340</td>
</tr>
<tr>
<td>Car018</td>
<td>NA</td>
<td>302</td>
<td>140</td>
<td>3353</td>
<td>8</td>
<td>70 US</td>
<td>ford mustang boss 302</td>
</tr>
<tr>
<td>Car019</td>
<td>15</td>
<td>400</td>
<td>150</td>
<td>3761</td>
<td>9.5</td>
<td>70 US</td>
<td>chevrolet monte carlo</td>
</tr>
<tr>
<td>Car020</td>
<td>14</td>
<td>455</td>
<td>225</td>
<td>3086</td>
<td>10</td>
<td>70 US</td>
<td>buick estate wagon (sw)</td>
</tr>
<tr>
<td>Car021</td>
<td>24</td>
<td>113</td>
<td>95</td>
<td>2372</td>
<td>15</td>
<td>70 Japan</td>
<td>toyota corona mark ii</td>
</tr>
</tbody>
</table>

**just for illustration:**

406 cars, 7 numeric dimensions
(not a high-dimensional dataset!)
Naïve Approach

Visualizing rows (cars) from this table is standard InfoVis:

Data transposition makes the dimensions to rows:

So what about visualizing this table?
Naïve Approach

Data transposition makes the dimensions to rows:

<table>
<thead>
<tr>
<th>Car001</th>
<th>Car002</th>
<th>Car003</th>
<th>Car004</th>
<th>Car005</th>
<th>Car006</th>
<th>Car007</th>
<th>Car008</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPG</td>
<td>18</td>
<td>15</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cylinders</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Displacem.[ln']</td>
<td>307</td>
<td>350</td>
<td>318</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horsepower</td>
<td>130</td>
<td>165</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight [lb]</td>
<td>3504</td>
<td>3693</td>
<td>3436</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceler.[s][0-6]</td>
<td>12</td>
<td>11,5</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ModelYear</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>chevrolet chevrolet skylark plymouth sate amc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

So why doesn’t this work?

Naïve Approach

No comparable values in the columns after transposition!
Dual Analysis Framework

What do we do, when we visualize data items?
– per data item $p_i$, we map properties/attributes of $p_i$ to vis.-cues, f.i., $x$ and $y \rightarrow$
– we see, how the $p_i$ relate to each other wrt. to their props.!

Translating this to visualizing dimensions:
– per data dimension $d_j$, we map properties/attributes of $d_j$ to vis.-cues, f.i., $x$ and $y \rightarrow$
– we see, how the $d_j$ relate to each other wrt. to their props.!

Expressive properties of dimensions $d_j$ (selection):
– descriptive statistics, like mean and std.-der.
– measures of outlyingness

So: constructing the properties table for dims. $d_j$:
– normalization first, then feature extraction

<table>
<thead>
<tr>
<th>#Outl</th>
<th>40</th>
<th>0</th>
<th>31</th>
<th>42</th>
<th>44</th>
<th>50</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Outl_high</td>
<td>33</td>
<td>0</td>
<td>31</td>
<td>40</td>
<td>42</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>#Outl_low</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>25</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>high_outl_th</td>
<td>0.698</td>
<td>1.013</td>
<td>0.735</td>
<td>0.638</td>
<td>0.748</td>
<td>0.700</td>
<td>1.008</td>
</tr>
<tr>
<td>low_outl_th</td>
<td>0.076</td>
<td>-0.013</td>
<td>-0.078</td>
<td>0.007</td>
<td>0.029</td>
<td>0.200</td>
<td>0.072</td>
</tr>
<tr>
<td>IQR</td>
<td>0.309</td>
<td>0.800</td>
<td>0.510</td>
<td>0.298</td>
<td>0.396</td>
<td>0.210</td>
<td>0.500</td>
</tr>
<tr>
<td>$q_3$</td>
<td>0.532</td>
<td>1.000</td>
<td>0.605</td>
<td>0.457</td>
<td>0.570</td>
<td>0.548</td>
<td>0.750</td>
</tr>
<tr>
<td>Med.</td>
<td>0.372</td>
<td>0.200</td>
<td>0.214</td>
<td>0.266</td>
<td>0.343</td>
<td>0.446</td>
<td>0.500</td>
</tr>
<tr>
<td>$q_1$</td>
<td>0.223</td>
<td>0.200</td>
<td>0.095</td>
<td>0.159</td>
<td>0.174</td>
<td>0.338</td>
<td>0.250</td>
</tr>
<tr>
<td>Kurt.</td>
<td>-0.511</td>
<td>-1.411</td>
<td>-0.811</td>
<td>0.541</td>
<td>-0.821</td>
<td>0.373</td>
<td>-1.200</td>
</tr>
<tr>
<td>Skew.</td>
<td>0.457</td>
<td>0.506</td>
<td>0.694</td>
<td>1.034</td>
<td>0.506</td>
<td>0.230</td>
<td>0.021</td>
</tr>
<tr>
<td>Std.-Dev.</td>
<td>0.208</td>
<td>0.342</td>
<td>0.271</td>
<td>0.210</td>
<td>0.240</td>
<td>0.167</td>
<td>0.312</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.39</td>
<td>0.50</td>
<td>0.33</td>
<td>0.32</td>
<td>0.39</td>
<td>0.45</td>
<td>0.54</td>
</tr>
<tr>
<td>Max.</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Min.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

| Cylinders | 0.24 | 1.00 | 0.62 | 0.46 | 0.54 | 0.24 | 0.00 |
| Displ | 0.16 | 1.00 | 0.73 | 0.65 | 0.59 | 0.21 | 0.00 |
| Horsepower | 0.24 | 1.00 | 0.65 | 0.57 | 0.52 | 0.18 | 0.00 |
| Weight | 0.19 | 1.00 | 0.61 | 0.57 | 0.52 | 0.24 | 0.00 |
| Acceleration | 0.21 | 1.00 | 0.60 | 0.51 | 0.52 | 0.15 | 0.00 |
| ModelYear | 0.16 | 1.00 | 0.93 | 0.83 | 0.77 | 0.12 | 0.00 |
| Origin | 0.13 | 1.00 | 1.00 | 0.95 | 0.78 | 0.06 | 0.00 |
| Name | 0.13 | 1.00 | 0.96 | 0.92 | 0.77 | 0.03 | 0.00 |

normalized data (1.)

properties table (2.)
Now we can visualize the dims. $d_j$

- mapping coherent and expressive properties to vis.-cues!

### Table (2)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#Out</td>
<td>40</td>
<td>0</td>
<td>31</td>
<td>42</td>
<td>44</td>
<td>50</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#Out_high</td>
<td>33</td>
<td>0</td>
<td>31</td>
<td>40</td>
<td>42</td>
<td>25</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#Out_low</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>25</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high_out_th</td>
<td>0.698</td>
<td>1.013</td>
<td>0.735</td>
<td>0.638</td>
<td>0.746</td>
<td>0.700</td>
<td>1.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low_out_th</td>
<td>0.54</td>
<td>0.5</td>
<td>0.342</td>
<td>0.8</td>
<td>1.034</td>
<td>35</td>
<td>42</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7 dims. in parallel coordinates (wrt. 8 selected properties)

Now the dual analysis can start!

- look up informative properties in the dims.-vis.
- do related items-visualization, accordingly

**Example 1: exploring the most skewed dimension**

1., selecting max(skew) in dims.-vis. (PC)

... it’s Horsepower!
Now the dual analysis can start!

- look up informative properties in the dims.-vis.
- do related items-visualization, accordingly

Example 2: comparing Gaussian & ranking-based stats.

1., seeing that both measures for the spread (std.-dev. vs. IQR) agree for all but one dim.

2., selecting it (not shown) reveals: it’s ModelYear, a discrete data dimension...
Now the dual analysis can start!
– look up informative properties in the dims.-vis.
– do related items-visualization, accordingly

Example 2: comparing Gaussian & ranking-based stats.

next:
comparing a low-spread (Accel.) vs. a high-spread dim. (Displacem.)

... the top-right dim. – max. spread – is Cylinders – basically categorical...
Quickly, we explore the dims. according to their props.
- hundreds or thousands of dims. → no problem! :-) 
- dozens of properties → std. InfoVis is fine!

The Dual Analysis emerges through iteration:
- one key tool: the difference view

Working with the difference view

Higher values for the selection

Change in QQR values

Change in med values
Dual Analysis Framework

Difference View in action:

Dual Analysis Framework

More in the thesis / papers of Çağatay Turkay et al.


Integrating Computational Tools in Interactive and Visual Methods for Enhancing High-dimensional Data and Cluster Analysis

ÇAĞATAY TURKAY

Dissertation for the degree of Philosophiae Doctor (PhD)
Supervised by Helwig Hauser
Co-supervised by Peer Nebert
Institute for Informatics, University of Bergen

November 2014
Acknowledgements

You!

Luis Gustavo!


We’d like to hire ➢ 1 PostDoc