Visualization in the big data era

Dr. Mihael Ankerst
Allianz Deutschland AG
Munich, April, 7th 2016

Allianz

Introduction of myself – something big about me

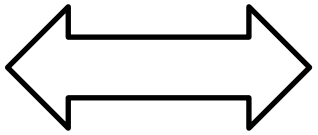
- I have worked for **big** employers

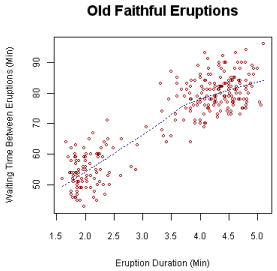
- I have **big** interest in
  data mining, visualization and scalability

- My daughter has **big** expectations

Big data and visualization don‘t seem to be a good match…

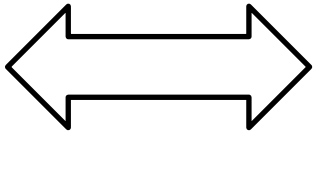
Lots of data !





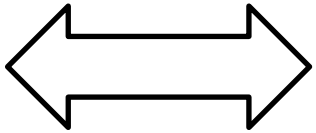
Visualization does not scale easily

increasing variety



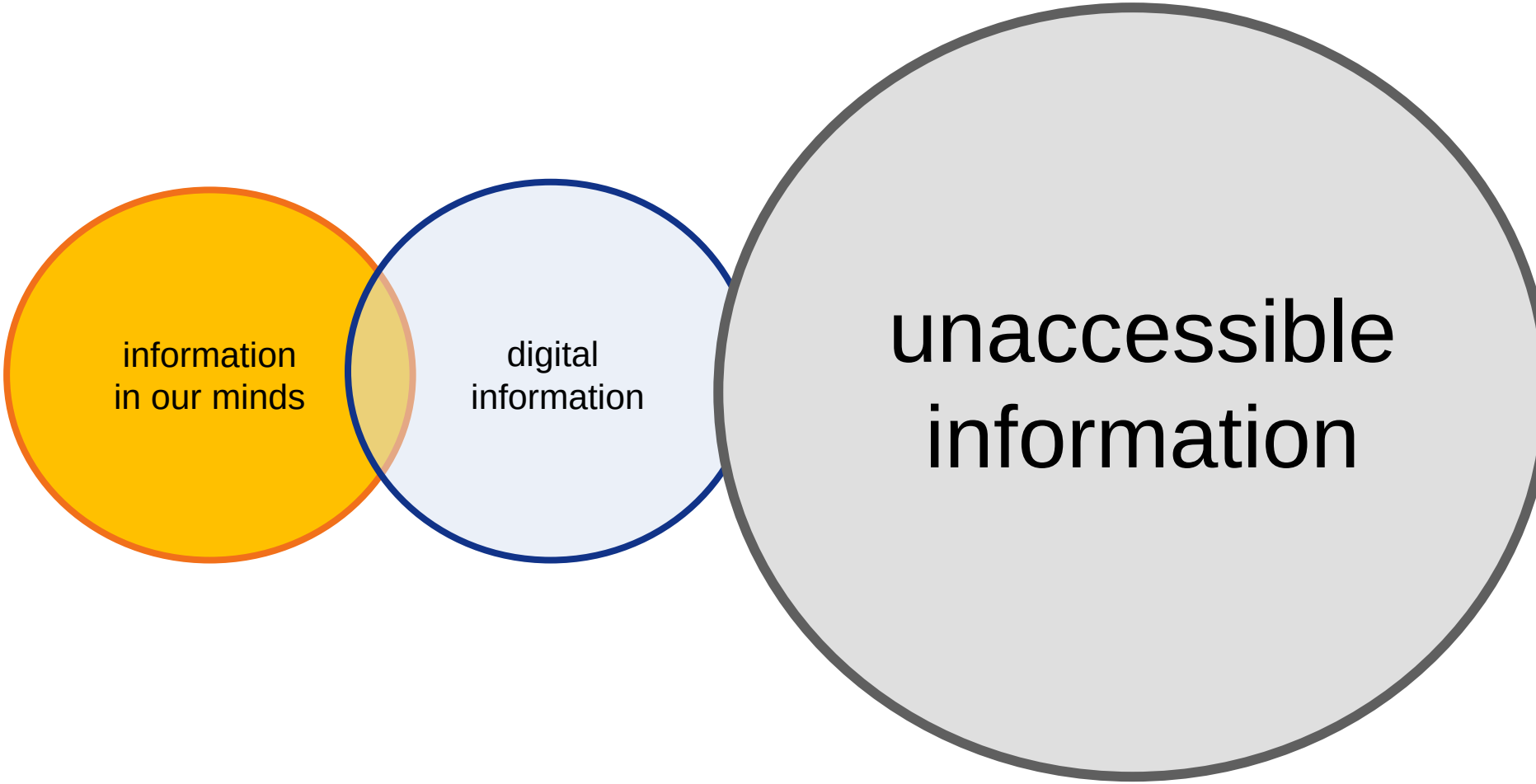
How to represent various data formats?

data is coming fast …



Don‘t we want to automate as much as possible?

The relevant space for data analysis:



The goal of any big data analysis is a result, that is…

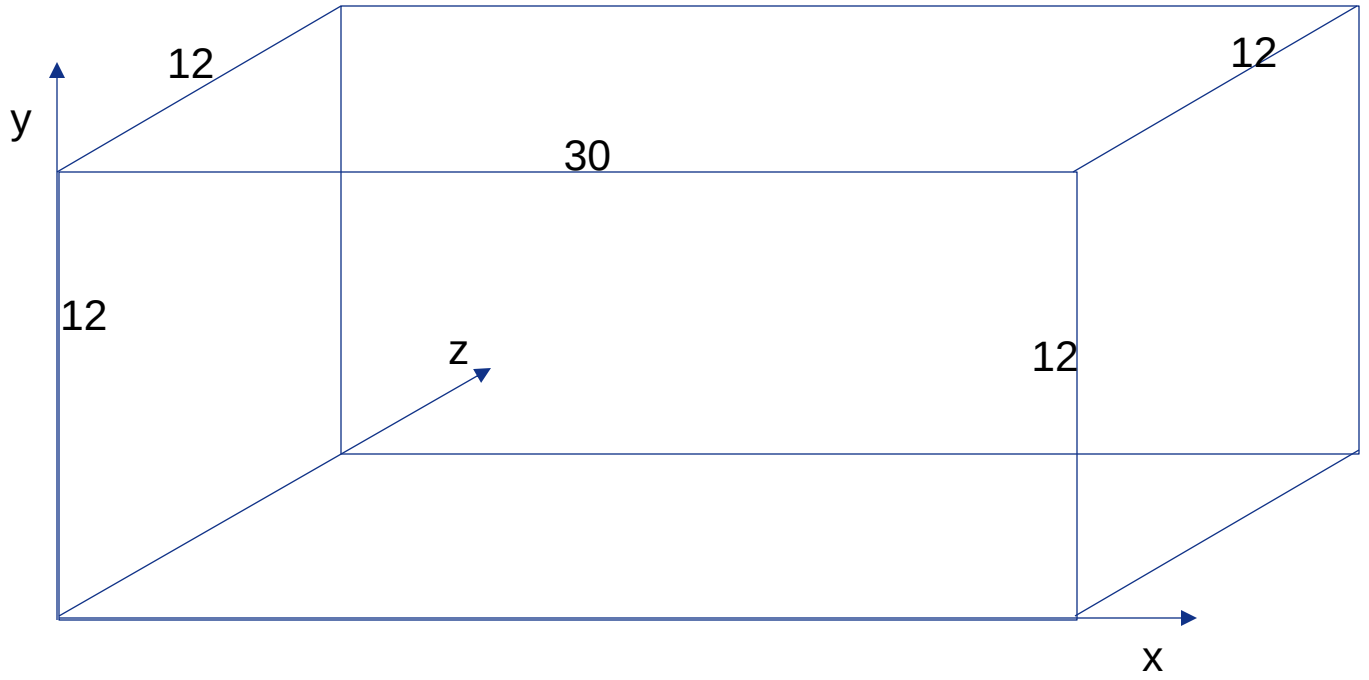
… valid

… new

… and applicable!

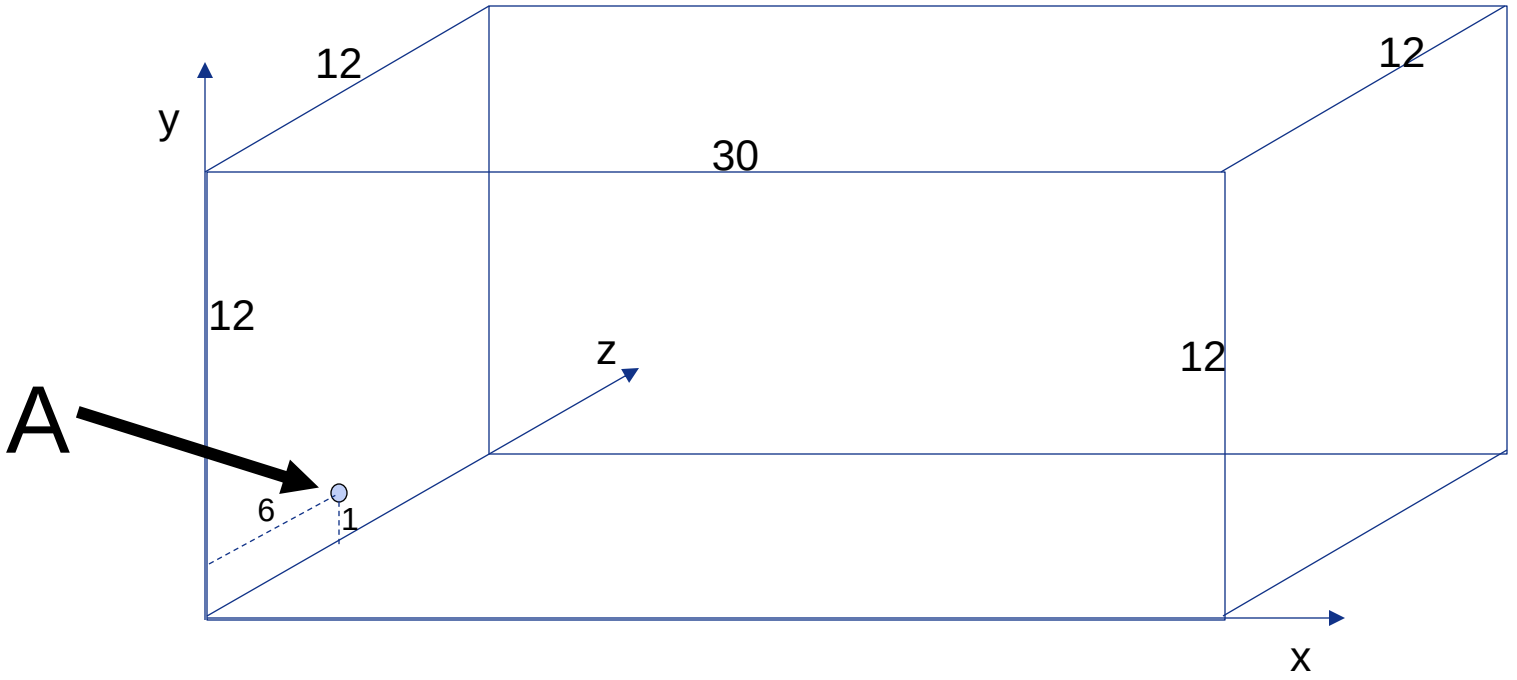
Let‘s look at the following box…

Let‘s look at the following box…



Box has the side lengths: (x, y, z) = (30,12,12)

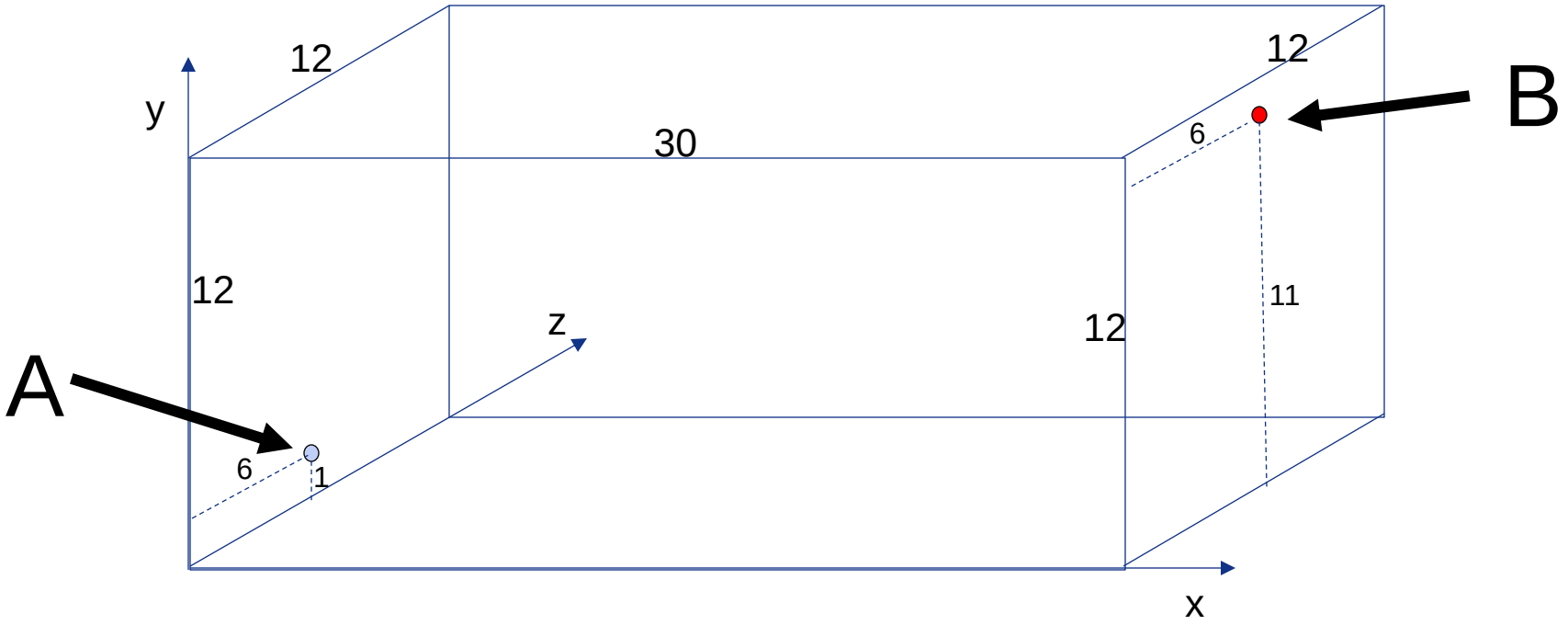
Let‘s look at the following box…



Box has the side lengths: (x, y, z) = (30,12,12)

- Ant A: is standing at (x, y, z) = (0,1,6)

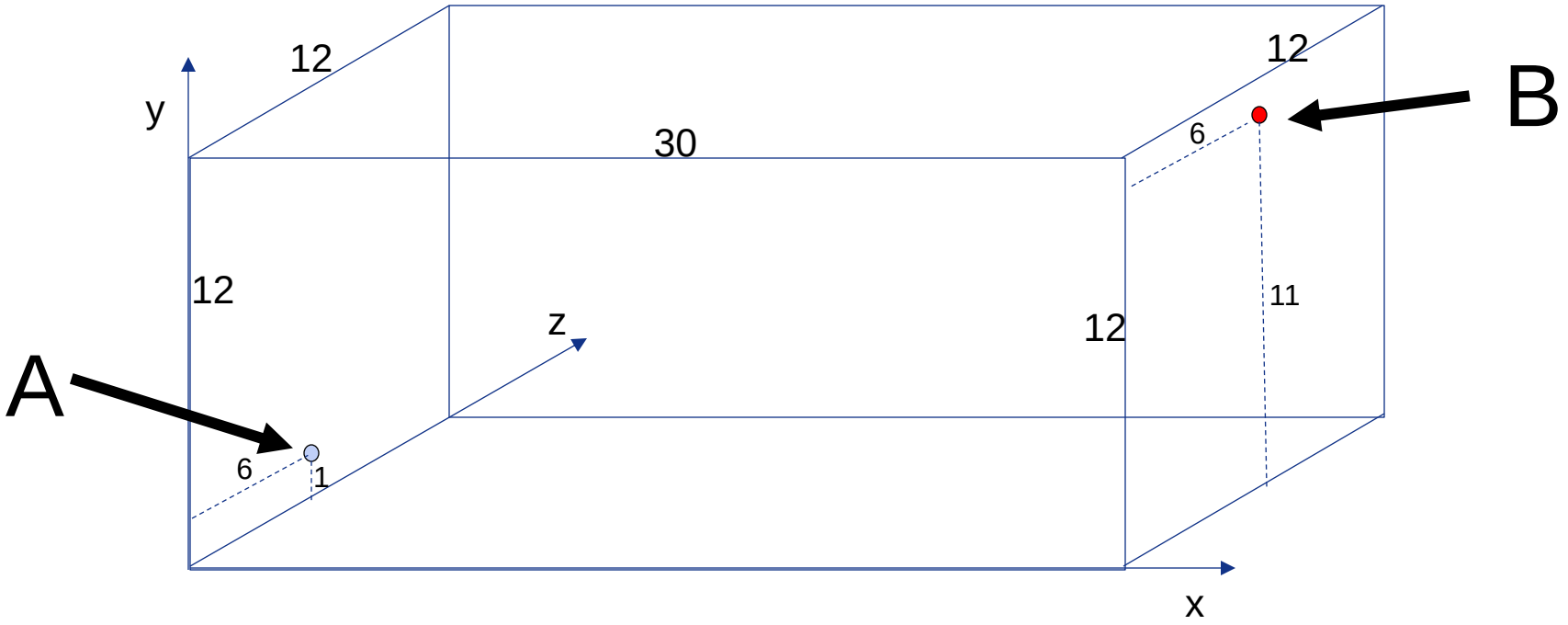
Let‘s look at the following box…



Box has the side lengths: (x, y, z) = (30,12,12)

- Ant A: is standing at (x, y, z) = (0,1,6)
- Ant B: is standing at (x, y, z) = (30,11,6)

What is the shortest path to come from A to B ?



Box has the side lengths: (x, y, z) = (30,12,12)

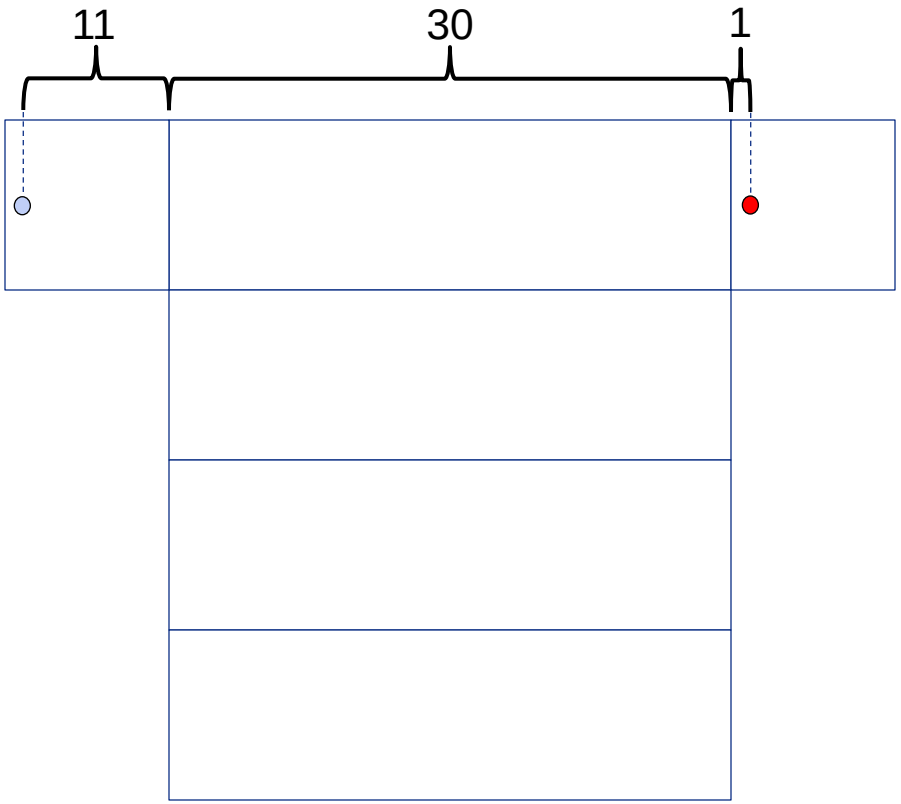
- Ant A: is standing at (x, y, z) = (0,1,6)
- Ant B: is standing at (x, y, z) = (30,11,6)

Question: What is the shortest path for ant A to come to ant B ?

(ant B does not move and moving is just on the surface of box possible - the box is solid)

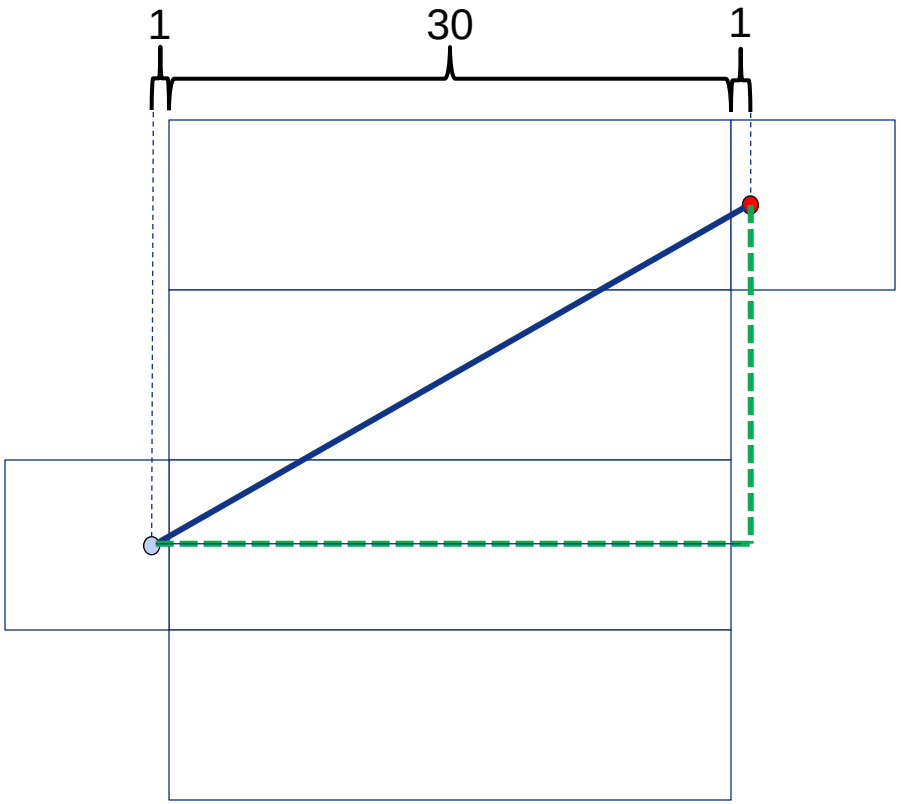
The solution of the puzzle
Part 1…

Unfold the box …



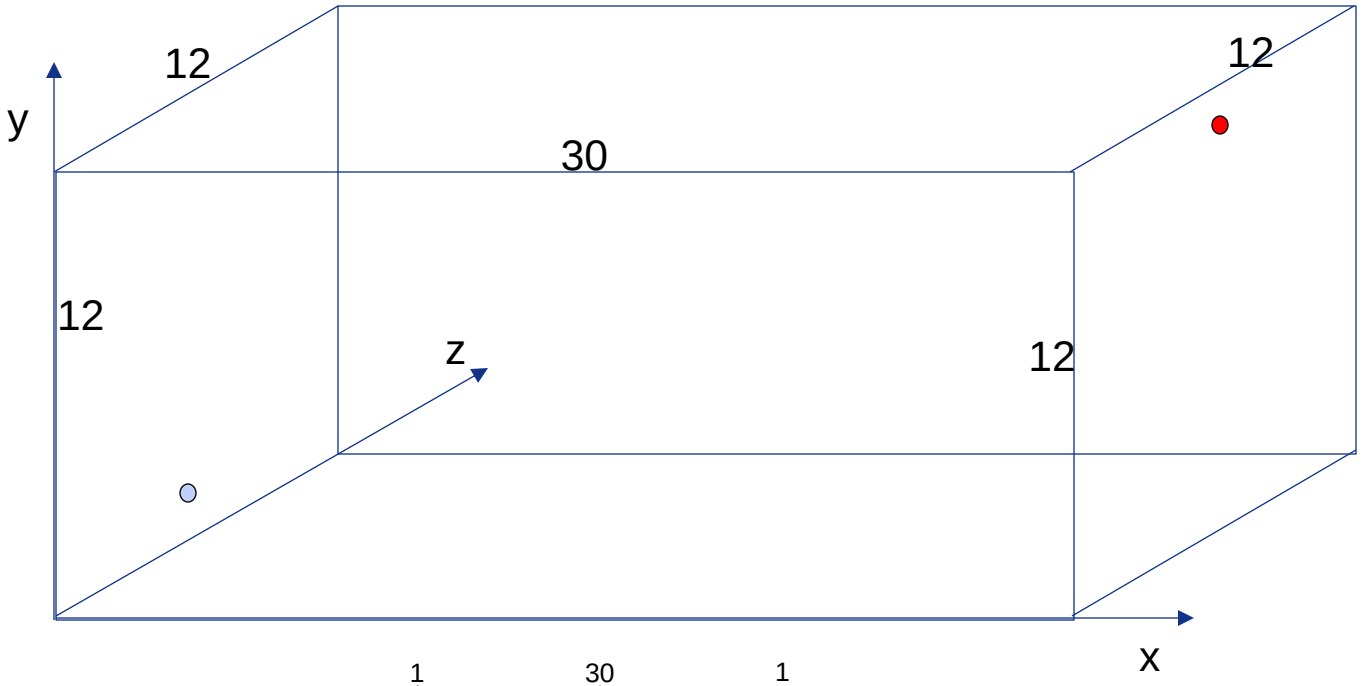
The solution of the puzzle
Part 2…

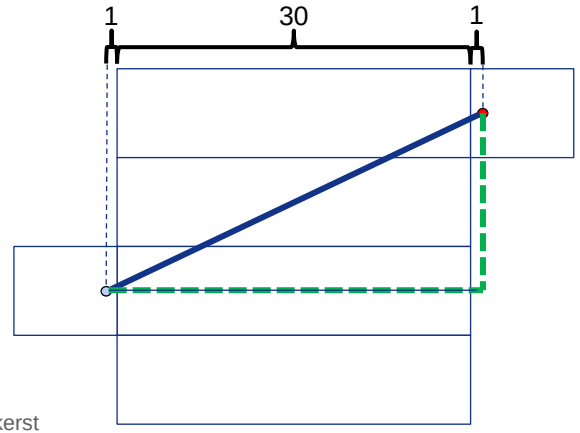
Answer: The shortest path is just 40 units long!



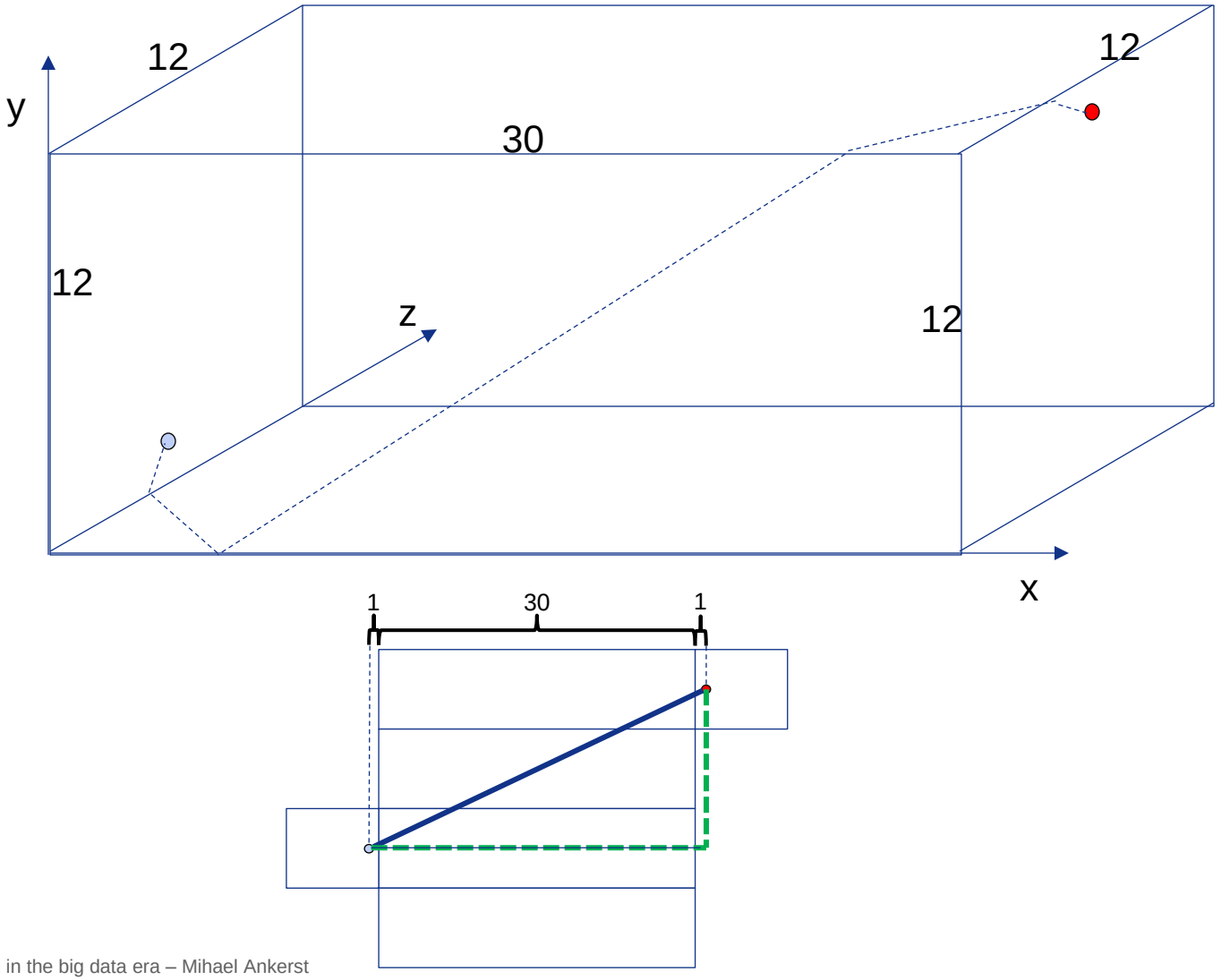
Representation matters!

What is the shortest path to come from A to B ?





What is the shortest path to come from A to B ?



Visualization is the data analysts‘ best friend if …

1) **it is based upon an intuitive representation**

Why should we visualize data ? - 1

Anscombe's Quartet Data Table

Data Set A		*Data Set B*		*Data Set C*		*Data Set D*	
X	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Why should we visualize data ? - 2

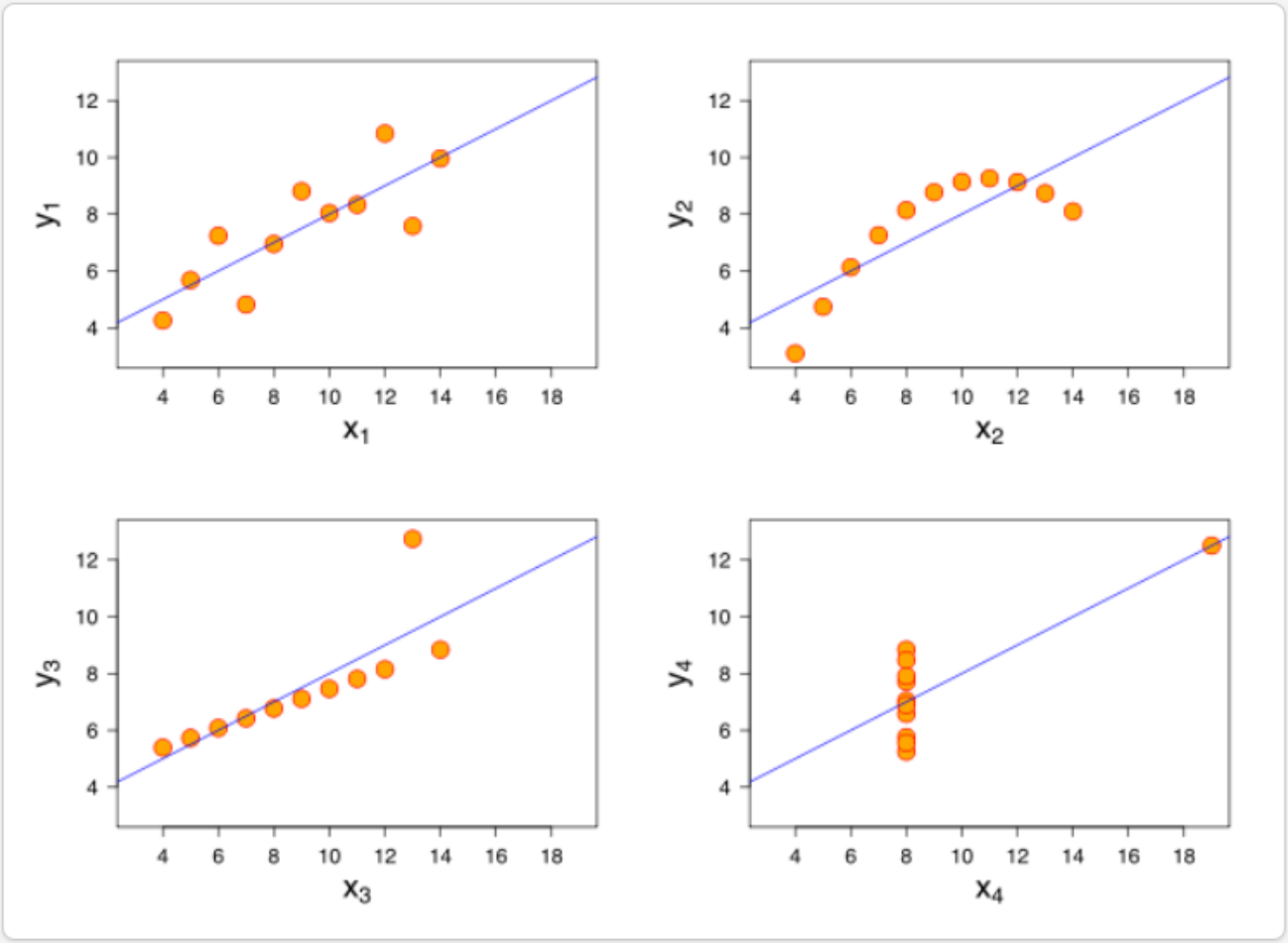
Allianz

Simple Summary Statistics of Anscombe's Quartet Data Table

Property	Value
Mean of x of each data set	9 (exact)
Variance of x in each data set	11 (exact)
Mean of y in each data set	7.50 (to 2 decimal places)
Variance of y in each data set	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each data set	0.816 (to 3 decimal places)
Linear regression line for each data set	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

Why should we visualize data ? - 3

Graph of Anscombe's Quartet Data Table



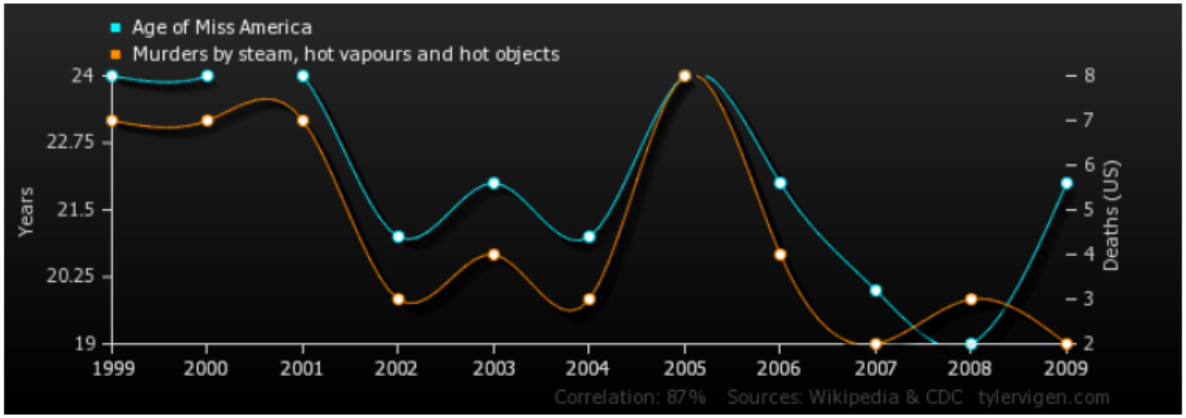
source: Wikimedia Commons

Visualization is the data analysts‘ best friend if …

1) it is based upon an intuitive representation
2) **it leverages the perceptual capabilities of the user**

Correlation is not causation

Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects



Upload this image to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America *Years (Wikipedia)*	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects *Deaths (US) (CDC)*	7	7	7	3	4	3	8	4	2	3	2

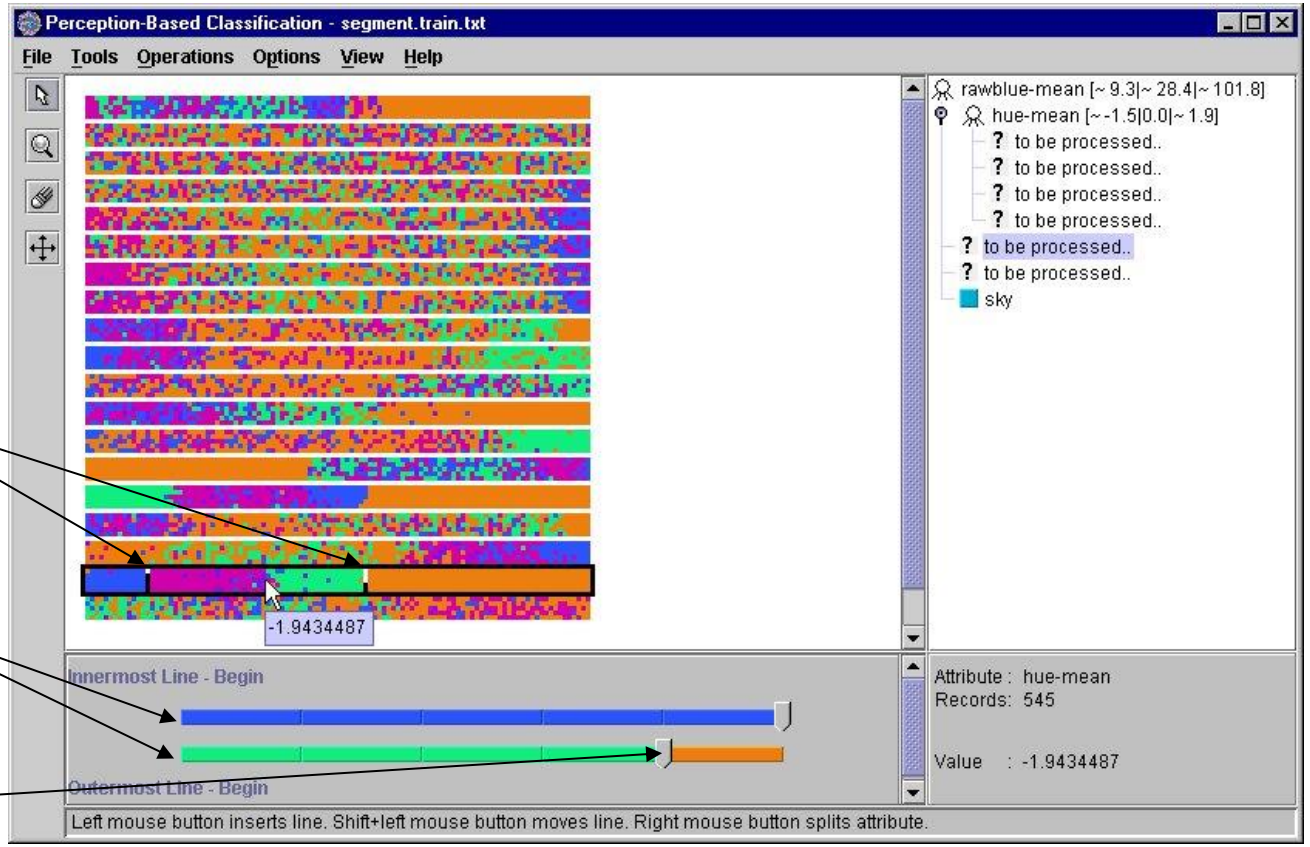
Correlation: 0.870127

→ This and many more examples: http://www.tylervigen.com/

How to incorporate domain knowledge?

Age	Weeks since last purchase	Last purchased product
35	8	P-H
47	6	P-H
20	24	P-K

Illustration 1: Incorporating domain knowledge into the decision tree building process

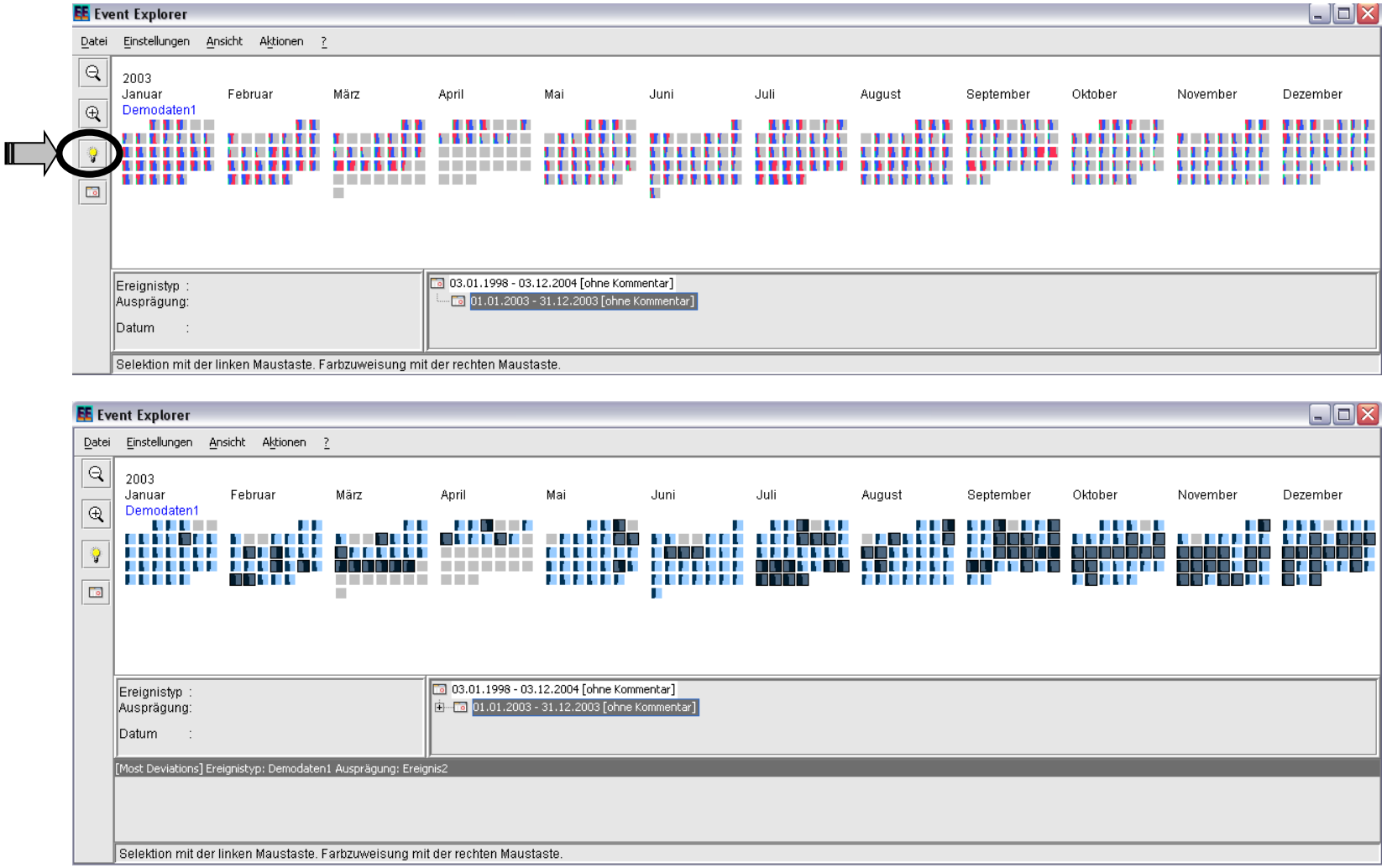


animated split lines

magnified split lines

exact split point

Illustration 2: Incorporating domain knowledge into the analysis of event data



Visualization is the data analysts‘ best friend if …

1) it is based upon an intuitive representation
2) it leverages the perceptual capabilities of the user
3) **it enables the incorporation of domain knowledge**

What kind of products do customers typically buy together in a grocery store?




# customers	fruit	beer	candy	magazines	…
6.388.860	1	0	0	0	
898.973	1	0	1	0	
4.231.452	0	1	0	0	
5.123.433	0	1	1	1	
…	…	…	…	…	

Sorting by frequency …




# customers	fruit	beer	candy	magazines	…
6.567.680	1	1	0	0	
6.549.840	1	1	1	0	
6.488.320	1	0	1	0	
6.388.860	1	0	0	0	
…	…	…	…	…	

… or creating a pivot table ….




Zeilenbeschriftungen	Summe von _customers
0	**20089368**
0	**9965592**
0	4823352
1	5142240
0	**1267040**
0	632864
0	**315776**
0	157232
0	**77960**
0	77960
0	**46840**
0	23020
0	**23020**
0	23020
0	**8482**
1	8482
0	**8482**
0	1213
0	**1213**
0	256
0	256
1	957
0	957
1	7269
0	**7269**
0	3252
0	3252
1	4017
0	4017
1	**14538**
1	14538
0	**14538**
0	4209
0	**4209**
0	1722

… or mining association rules doesn‘t give you the full picture!

```
107 {B.wat}       => {F.fis} 0.2767582 0.5461100  1.0092072
108 {F.fis}       => {B.wat} 0.2767582 0.5114472  1.0092072
109 {B.wat}       => {F.veg} 0.3124506 0.6165395  0.9872194
110 {F.veg}       => {B.wat} 0.3124506 0.5003042  0.9872194
111 {B.wat}       => {F.fru} 0.3120869 0.6158218  1.0069344
112 {F.fru}       => {B.wat} 0.3120869 0.5102953  1.0069344
113 {B.jui}       => {F.mea} 0.2635273 0.5106947  0.9897580
114 {F.mea}       => {B.jui} 0.2635273 0.5107323  0.9897580
115 {B.jui}       => {F.fis} 0.2811916 0.5449267  1.0070205
116 {F.fis}       => {B.jui} 0.2811916 0.5196400  1.0070205
117 {B.jui}       => {F.veg} 0.3157756 0.6119476  0.9798669
118 {F.veg}       => {B.jui} 0.3157756 0.5056283  0.9798669
119 {B.jui}       => {F.fru} 0.3165203 0.6133908  1.0029593
120 {F.fru}       => {B.jui} 0.3165203 0.5175444  1.0029593
121 {F.mea}       => {F.fis} 0.2888593 0.5598271  1.0345563
122 {F.fis}       => {F.mea} 0.2888593 0.5338097  1.0345563
123 {F.mea}       => {F.veg} 0.3048155 0.5907513  0.9459268
124 {F.veg}       => {F.mea} 0.3048155 0.4880787  0.9459268
125 {F.mea}       => {F.fru} 0.3241218 0.6281681  1.0271218
126 {F.fru}       => {F.mea} 0.3241218 0.5299737  1.0271218
127 {F.fis}       => {F.veg} 0.3065035 0.5664161  0.9069605
128 {F.veg}       => {F.fis} 0.3065035 0.4907815  0.9069605
129 {F.fis}       => {F.fru} 0.3417171 0.6314906  1.0325545
130 {F.fru}       => {F.fis} 0.3417171 0.5587439  1.0325545
131 {F.veg}       => {F.fru} 0.3299639 0.5283470  0.8639036
132 {F.fru}       => {F.veg} 0.3299639 0.5395261  0.8639036
133 {F.cer,F.can} => {F.veg} 0.1578878 0.6259526  1.0022920
134 {F.can,F.veg} => {F.cer} 0.1578878 0.5025513  0.9993173
135 {F.cer,F.veg} => {F.can} 0.1578878 0.5024265  0.9994376
136 {F.cer,F.can} => {F.fru} 0.1546667 0.6131823  1.0026184
137 {F.can,F.fru} => {F.cer} 0.1546667 0.5026169  0.9995273
138 {F.cer,F.fru} => {F.can} 0.1546667 0.5025838  0.9995914
139 {F.egg,F.can} => {F.veg} 0.1579408 0.6258022  1.0020511
140 {F.can,F.veg} => {F.egg} 0.1579408 0.5026801  0.9990752
141 {F.egg,F.veg} => {F.can} 0.1579408 0.5024257  0.9994360
142 {F.egg,F.can} => {F.fru} 0.1547359 0.6130835  1.0024896
143 {F.can,F.fru} => {F.egg} 0.1547359 0.5028420  0.9993070
144 {F.egg,F.fru} => {F.can} 0.1547359 0.5025027  0.9995891
145 {B.mil,F.can} => {F.veg} 0.1574137 0.6236467  0.9985098
146 {F.can,F.veg} => {B.mil} 0.1574137 0.5010025  0.9952151
147 {B.mil,F.veg} => {F.can} 0.1574137 0.5020946  0.9987774
148 {B.mil,F.can} => {F.fru} 0.1551516 0.6146844  1.0050746
149 {F.can,F.fru} => {B.mil} 0.1551516 0.5041927  1.0015523
150 {B.mil,F.fru} => {F.can} 0.1551516 0.5024959  0.9995757
151 {B.win,F.can} => {F.veg} 0.1580469 0.6255017  1.0015701
152 {F.can,F.veg} => {B.win} 0.1580469 0.5030177  0.9985919
153 {B.win,F.veg} => {F.can} 0.1580469 0.5024241  0.9994327
154 {B.win,F.can} => {F.fru} 0.1548745 0.6129462  1.0022324
155 {F.can,F.fru} => {B.win} 0.1548745 0.5032923  0.9991369
156 {B.win,F.fru} => {F.can} 0.1548745 0.5025004  0.9995846
157 {B.bee,F.can} => {F.veg} 0.1578409 0.6224635  0.9967851
158 {F.can,F.veg} => {B.bee} 0.1578409 0.5023620  0.9934553
159 {B.bee,F.veg} => {F.can} 0.1578409 0.5020889  0.9987660
160 {B.bee,F.can} => {F.fru} 0.1557057 0.6148432  1.0040262
161 {F.can,F.fru} => {B.bee} 0.1557057 0.5059936  1.0006370
162 {B.bee,F.fru} => {F.can} 0.1557057 0.5024870  0.9995580
163 {B.wat,F.can} => {F.veg} 0.1569873 0.6175769  0.9888807
164 {F.can,F.veg} => {B.wat} 0.1569873 0.4996452  0.9859191
165 {B.wat,F.veg} => {F.can} 0.1569873 0.5024388  0.9994620
```

Output of arules package in R Studio

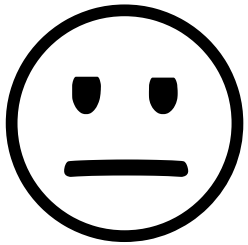
The idea of item explorer was born

- D3.js

- Use bar charts to represent item frequencies!

Development of item explorer – part 1

My daughter‘s face



Munich, March 8th, 2015, 5.15 p.m.

Development of item explorer – part 2

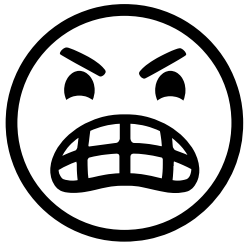
My daughter‘s face

20 minutes later …

Munich, March 8th, 2015, 5.35 p.m.

Development of item explorer – part 3

My daughter‘s face



40 minutes later …

Munich, March 8th, 2015, 5.55 p.m.

Development of item explorer – part 4

44 minutes later …

Munich, March 8th, 2015, 5.59 p.m.

svg element does not show up

0

In my code, I append a 'rect' but for some reason it doesn't show up. The 'rect' I append shows up in the DOM as:

```
▼<g class="altIndicator" style="stroke: rgb(0, 0, 0); stroke-width: 2px; opacity: 0; fill: rgb(255, 255, 0);">
   <circle class="altCircle" r="10" cy="40" style="opacity: 0;"></circle>
   <line x1="0" y1="25" x2="0" y2="30"></line>
   <text class="altText" y="37" dy=".71em" style="text-anchor: middle; opacity: 0;"></text>
   <path class="line" d="M0,25A 1,1 0 0 0 35,25A 1,1 0 0 0 70,25" style="opacity: 1;"></path>
   <rect class="testRect" x="-50" width="700" y="-100" height="425" style="fill: rgb(0, 0, 0); opacity: 1;"></rect>
</g>
```

Here is my d3.js excerpt when I append it:

```
var testRect = d3.selectAll(".altIndicator").append("rect")
  .attr("class", "testRect")
  .attr("x", -50)
  .attr("width", 20*x.rangeBand())
  .attr("y", -100)
  .attr("height", 425).style("fill", "black").style("opacity", 1);
```

What can be the reason it is not shown ? I tried in the console to run this append interactively and it works when I append it anywhere else, so it seems an inheritance issue. However, by setting the style explicitely, it should override the style values inherited from the element or the CSS.

My complete code is below and the append is line 298-303. Any help would be greatly appreciated.

https://gist.github.com/EE2dev/d1c86cc47ad2759d955e

css svg d3.js

share edit close delete flag

asked Mar 8 '15 at 17:59

ee2Dev

661 ● 3 ● 9

Development of item explorer – part 5

…after playing Badminton

Munich, March 8th, 2015, 7.18 p.m.

2

It's contained in a `g` (the `altIndicator`) that has a 0 opacity.

This is being set here:

```
d3.selectAll(".altIndicator")
  .each(function (d) { if (selectedItemSet.has(d) && selectedItemSet.get(d).alternativeId
    d3.select(this).style("fill", "white")
      .style("opacity", 0.5);
    }
    else {
       d3.select(this).style("opacity", 0);
    }
  });
```

Code without that line.

share edit flag

answered Mar 8 '15 at 19:18

Mark

57.3k ● 5 ● 74 ● 125

Great, thanks so much! After the fact, it seems obvious... – ee2Dev Mar 8 '15 at 19:25

add a comment

Demo: item explorer

Info

current filter:

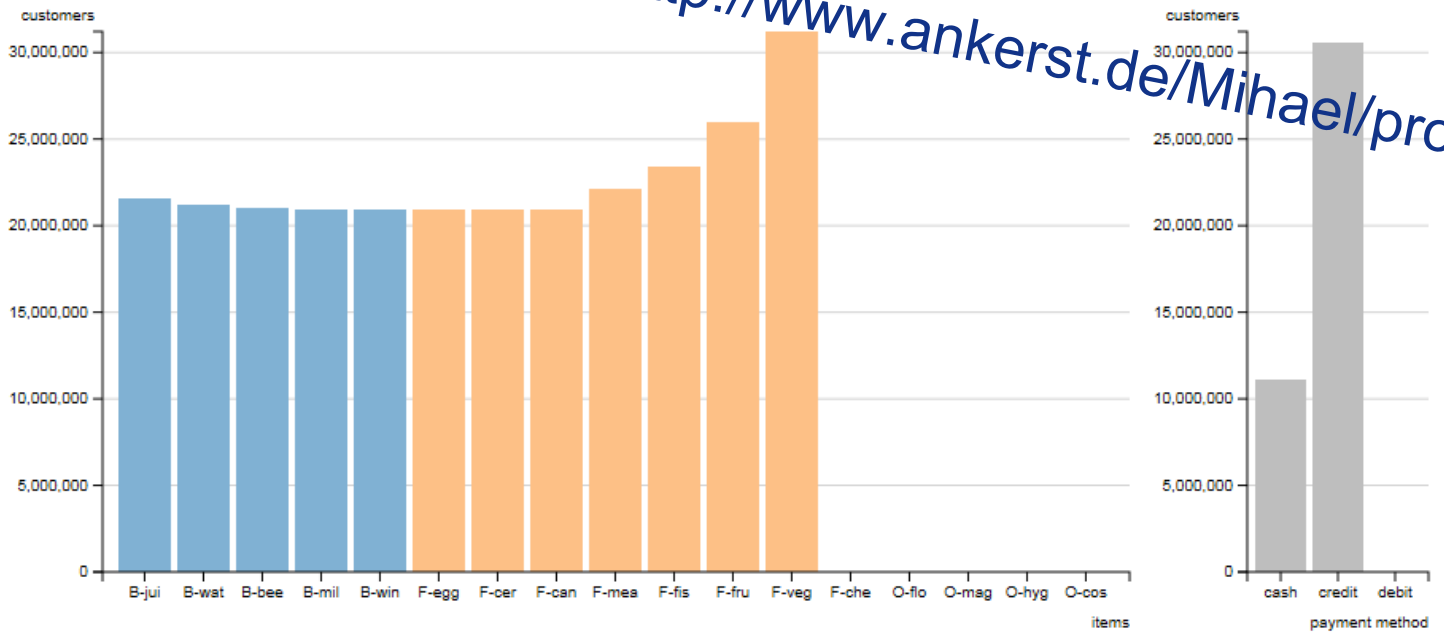
frequency: 41,636,931 percent: 100%

Exploration

item 1	item 2	frequency	percent
F-veg	credit	24,307,588	58.4%
F-fru	credit	19,251,390	46.2%
F-fru	F-veg	18,205,890	43.7%

☐ sort by frequency ☐ update axis help

http://www.ankerst.de/Mihael/proj/mbc/



Visualization is the data analysts‘ best friend if …

1) it is based upon an intuitive representation
2) it leverages the perceptual capabilities of the user
3) it enables the incorporation of domain knowledge
4) **it facilitates the understanding of the data and the results**