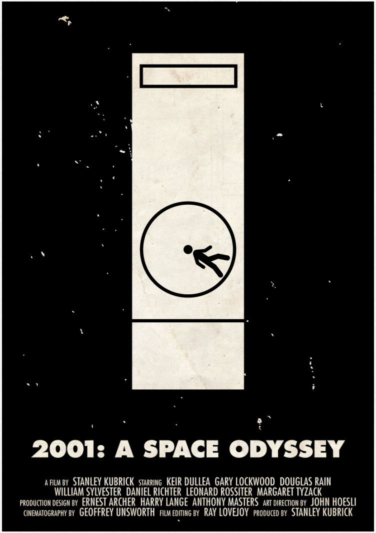
A biased & incomplete introduction to visualization

Bastian Rieck

Dramatis personæ


2001: A SPACE ODYSSEY
A FILM BY STANLEY KUBRICK STARRING KEIR DULLEA GARY LOCKWOOD DOUGLAS RAIN
WILLIAM SYLVESTER DANIEL RICHTER LEONARD ROSSITER MARGARET TYZACK
PRODUCTION DESIGN BY ERNEST ARCHER HARRY LANGE ANTHONY MASTERS ART DIRECTION BY JOHN HOESLI
CINEMATOGRAPHY BY GEOFFREY UNSWORTH FILM EDITING BY RAY LOVEJOY PRODUCED BY STANLEY KUBRICK

Source: Viktor Hertz, Jacob Atienza

What is visualization?

“Computer-based visualization systems provide visual representations of datasets intended to help people carry out some task better.”

— Tamara Munzner, *Visualization Design and Analysis: Abstractions, Principles, and Methods*

Why is visualization useful?

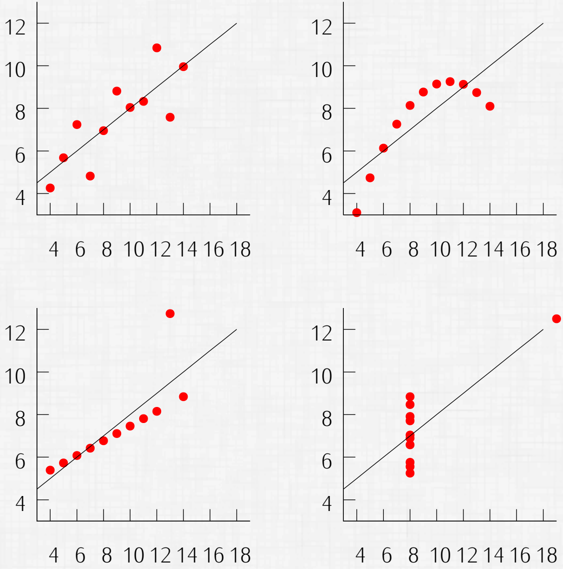
Anscombe's quartet

I		II		III		IV	
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

From the viewpoint of statistics

	x	y
Mean	9	7.50
Variance	11	4.127
Correlation	0.816	
Linear regression line	$y = 3.00 + 0.500x$	

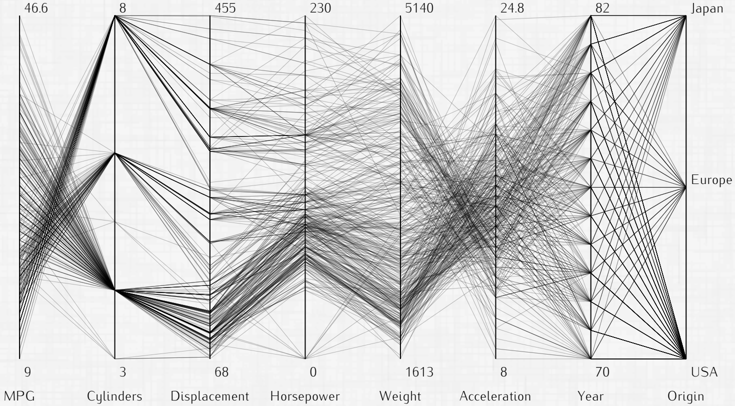
From the viewpoint of visualization



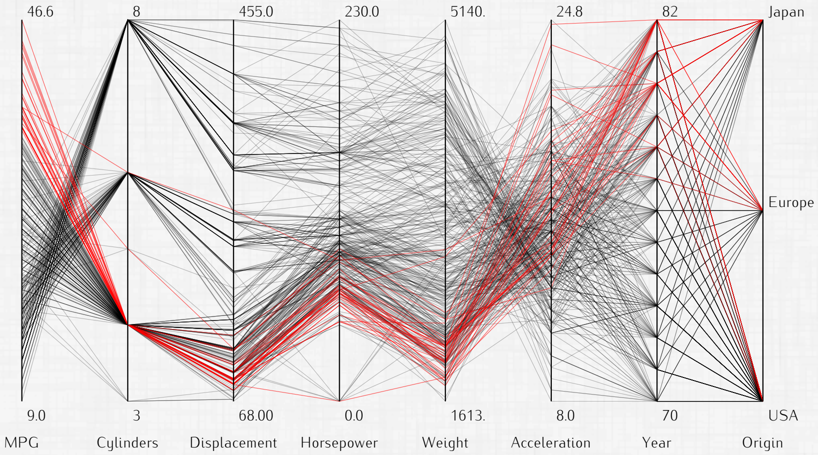
How does it work?

Parallel coordinates

- Tabular data (e.g. attributes in columns, instances in rows)
- Create an axis for each attribute dimension
- Draw a line through these axes to represent an instance



Brushing fuel-efficient cars

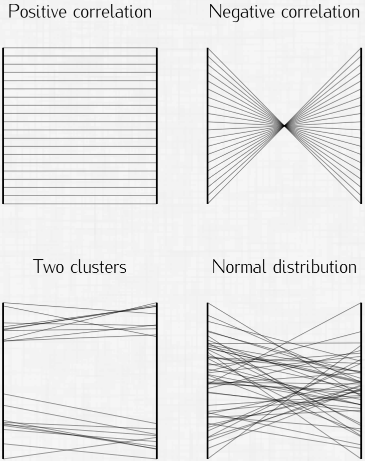


Some drawbacks

- Does not work for dimensions $\gg 10$
- Order of axes matters ($d!$ possibilities)
- Rapid overplotting

(Some of these drawbacks have been solved, others involve workarounds, which in turn cause other drawbacks, …)

Some patterns



Positive correlation

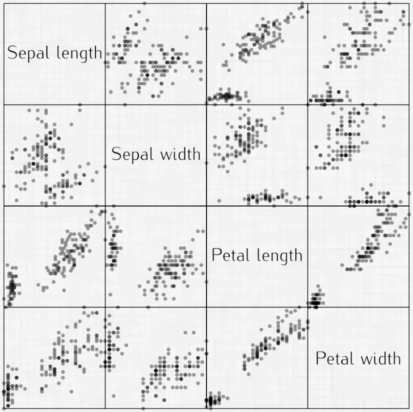
Negative correlation

Two clusters

Normal distribution

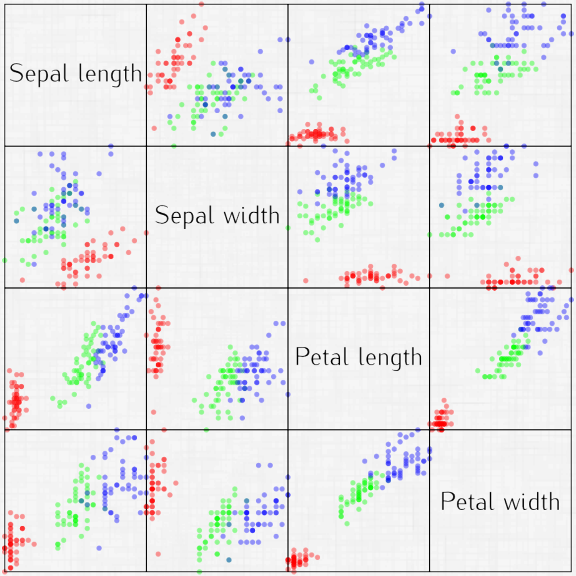
Scatterplot matrices

- Set of vectors from $\mathbb{R}^n$
- Create $(n \cdot n - 1)$ 2-dimensional scatterplots
- Arrange them in a matrix



Brushing by species

Iris setosa, Iris versicolor, Iris virginica



Analysis

Advantages

- Brushing+linking easily possible
- Conceptually simple
- Extendible (histograms, densities, ...)

Drawbacks

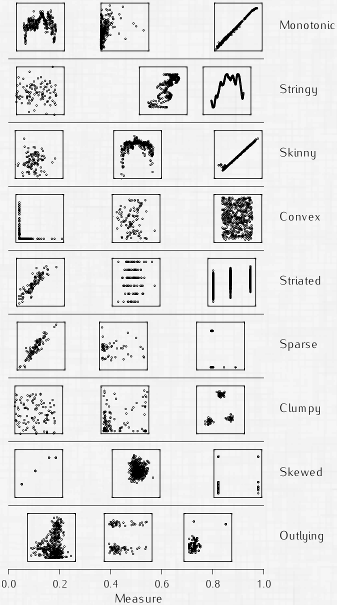
- Quadratic increase in number of plots
- Does not show all *interesting* projections
- Occlusion possible

Scagnostics

A cure for some drawbacks

Procedure

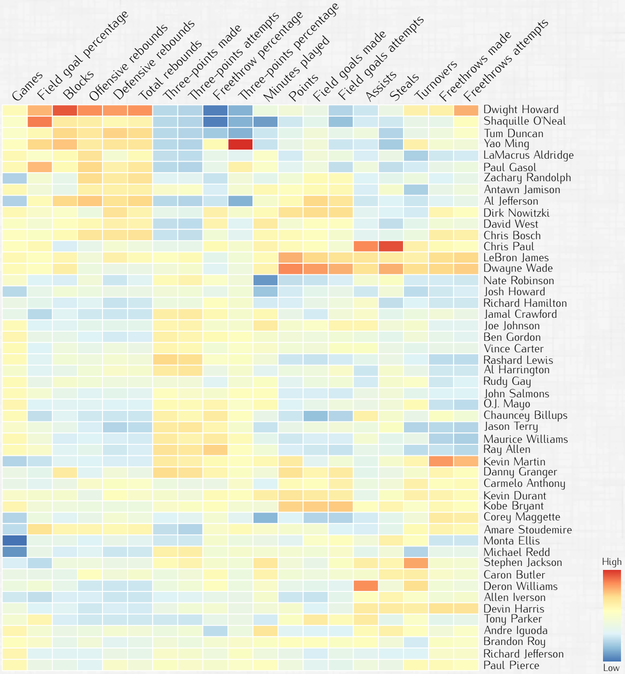
- Calculate $k \ll n$ measures for each scatterplot
- Assign each projection a vector of measures
- Show all vectors in a (smaller!) scatterplot



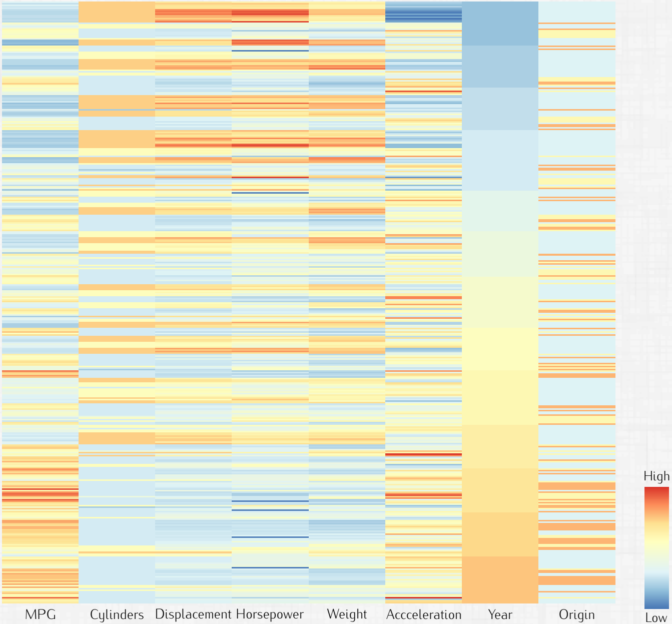
Source: Leland Wilkinson and Graham Wills. “Scagnostics Distributions.” *Journal of Computational and Graphical Statistics (JCGS)* 17:2 (2008), pp. 473–491.

Heatmaps

- “Matrix visualization”
- Assign colours according to value
- Scale *globally*, *per row*, or *per column*


Games
Field goal percentage
Blocks
Offensive rebounds
Defensive rebounds
Total rebounds
Three-points made
Three-points attempts
Freethrow percentage
Three-points percentage
Minutes played
Points
Field goals made
Field goals attempts
Assists
Steals
Turnovers
Freethrows made
Freethrows attempts
Dwight Howard
Shaquille O'Neal
Tum Duncan
Yao Ming
LaMacrus Aldridge
Paul Gasol
Zachary Randolph
Antawn Jamison
Al Jefferson
Dirk Nowitzki
David West
Chris Bosch
Chris Paul
LeBron James
Dwayne Wade
Nate Robinson
Josh Howard
Richard Hamilton
Jamal Crawford
Joe Johnson
Ben Gordon
Vince Carter
Rashard Lewis
Al Harrington
Rudy Gay
John Salmons
O.J. Mayo
Chauncey Billups
Jason Terry
Maurice Williams
Ray Allen
Kevin Martin
Danny Granger
Carmelo Anthony
Kevin Durant
Kobe Bryant
Corey Maggette
Amare Stoudemire
Monta Ellis
Michael Redd
Stephen Jackson
Caron Butler
Deron Williams
Allen Iverson
Devin Harris
Tony Parker
Andre Iguoda
Brandon Roy
Richard Jefferson
Paul Pierce
High
Low

Just what do you think you're doing, Dave?



What should I remember about visualization?

“I am putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do.”
— HAL 9000, 2001: A Space Odyssey

- Model-based approaches help explain data
- Visualization may help arrive at a *description* of the model
- It is challenging to scale methods to larger data sets
- It is easy to get it wrong, but hard to get it right

I want to learn more!

People

Ask your local friendly visualization researchers at INF 368, 5th floor, rooms 528, 529, and 531.

Tools

- D3.js (`http://d3js.org`)
- IBM ManyEyes (`http://www.ibm.com/software/analytics/manyeyes`)
- Tableau Software (`http://www.tableausoftware.com`)

Books

- Stephen Few. *Show Me the Numbers: Designing Tables and Graphs to Enlighten.*
- Tamara Munzner. *Visualization Design and Analysis: Abstractions, Principles, and Methods.*
- Edward R. Tufte. *The Visual Display of Quantitative Information.*
- Robin Williams. *The Non-Designer's Design Book.*

Thank you.