Business Intelligence & Process Modelling

Frank Takes

Universiteit Leiden

Lecture 2 — BI & Visual Analytics

Recap

- **Business Intelligence**: anything that aims at providing actionable information that can be used to support business decision making
  - **Business Intelligence**
  - **Visual Analytics**
  - Descriptive Analytics
  - Predictive Analytics
- Process Modelling (April and May)

Universiteit Leiden

Business Intelligence

Business Intelligence goals

- Operational intelligence
- Corporate governance
- Risk assessment
- Compliance
- Auditing
  - Sarbanes-Oxley (SOX) — role of IT in corporate governance

Management Approaches in BI

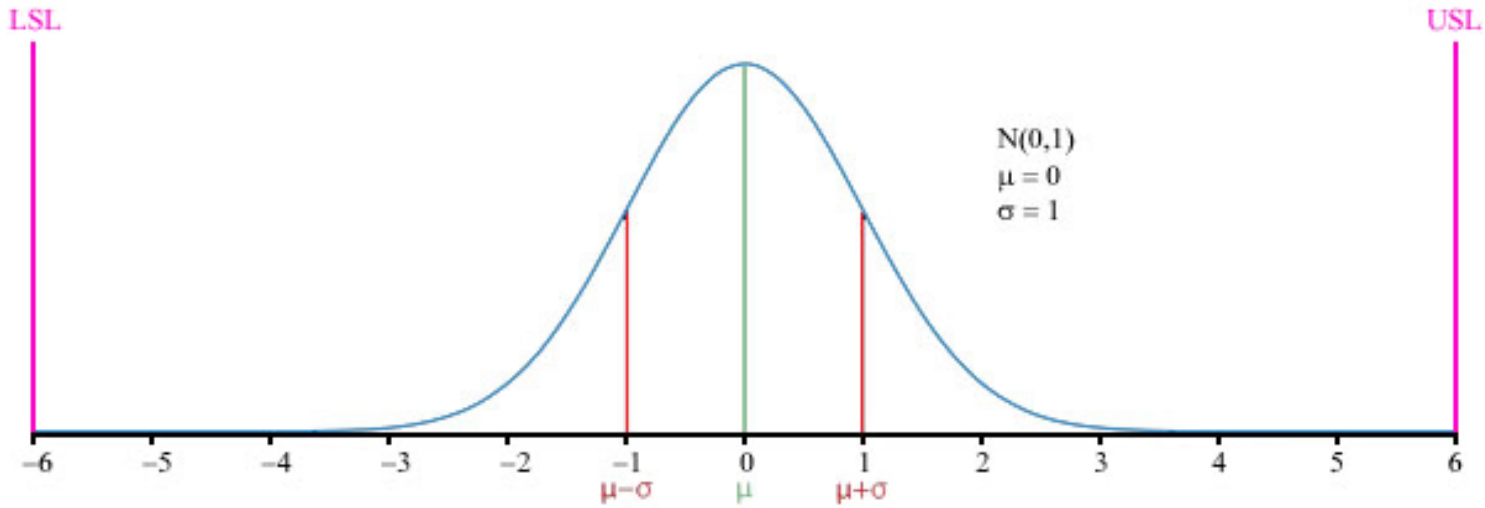
- **Continuous Process Improvement (CPI)**: ongoing effort to improve products, services or processes
  - Incremental improvements vs. Breakthrough improvements
  - Evaluate based on efficiency, effectiveness and flexibility
- **Total Quality Management (TQM)**: improve processes up to the microscopic level, focussing on meeting customer demands and realizing strategic company goals, e.g., **Six Sigma**

Six Sigma

- Originally developed by Motorola in the early 1980s
- Minimize Defective Parts per Million Opportunities (DPMO)
- Mean μ and standard deviaton σ

Quality level	**DPMO**	**% broken**	**% OK**
One Sigma	691.462	69	31
Two Sigma	308.538	31	69
Three Sigma	66.807	6,7	93,3
Four Sigma	6.210	0,62	99,38
Five Sigma	233	0,023	99,977
Six Sigma	3,4	0,00034	99,99966

Normal distribution



DMAIC approach

- **Define** the problem and set targets,
- **Measure** key performance indicators (KPI's) and collect data,
- **Analyze** the data to investigate and verify cause-and-effect relationships,
- **Improve** the current process based on this analysis,
- **Control** the process to minimize deviations from the target.

Key Performance Indicators

- **KPI**: measure, variable or metric to analyze the performance of (part of) an organization
- Strategic goals $\rightarrow$ Measurable variables
- **SMART**
  - **S**pecific
  - **M**easurable
  - **A**cceptable
  - **R**ealistic
  - **T**ime-sensitive

KPI examples

- Operational: increasing market share by 10%
- Financial: increase profit by 10%
- Sales: obtain 10 new customers
- Human resources: attract 10 new sales officers that are part of the world's top 1% in the field
- Customer support: forward no more than 10% of the support calls to second line

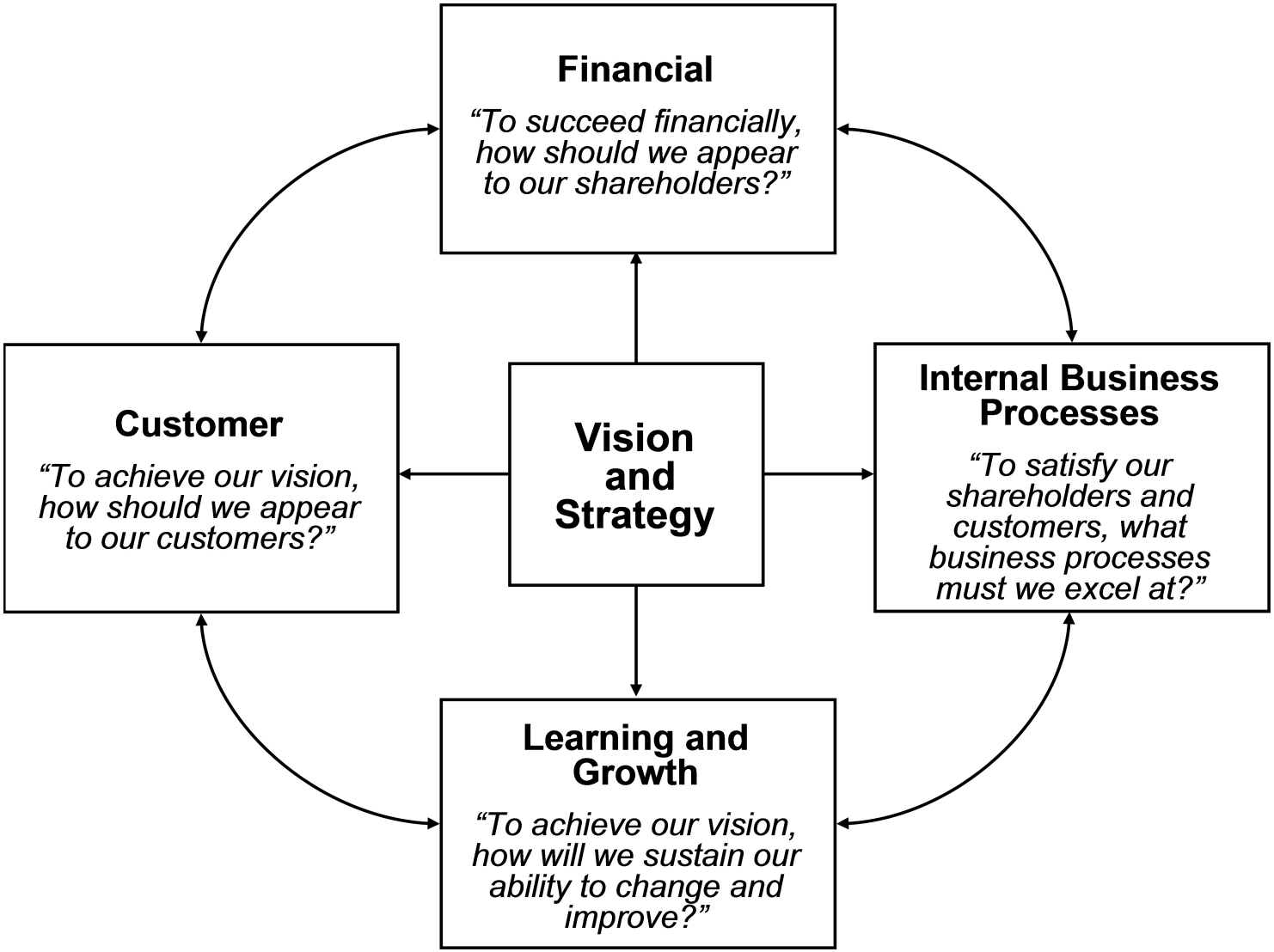
BI in practice

- Codeless reporting
- Instant querying
- Rich visualization
- Dashboards (“Management cockpits”)
- Scorecards

Balanced Scorecards

- R. Kaplan, D. Norton, The balanced scorecard: measures that drive performance, *Harvard business review* 83(7): 172–180, 2005.
- Goal: align business activities to the vision and strategy of the organization
- Financial and nonfinancial goals
- Monitor a relatively small number of summative indicators

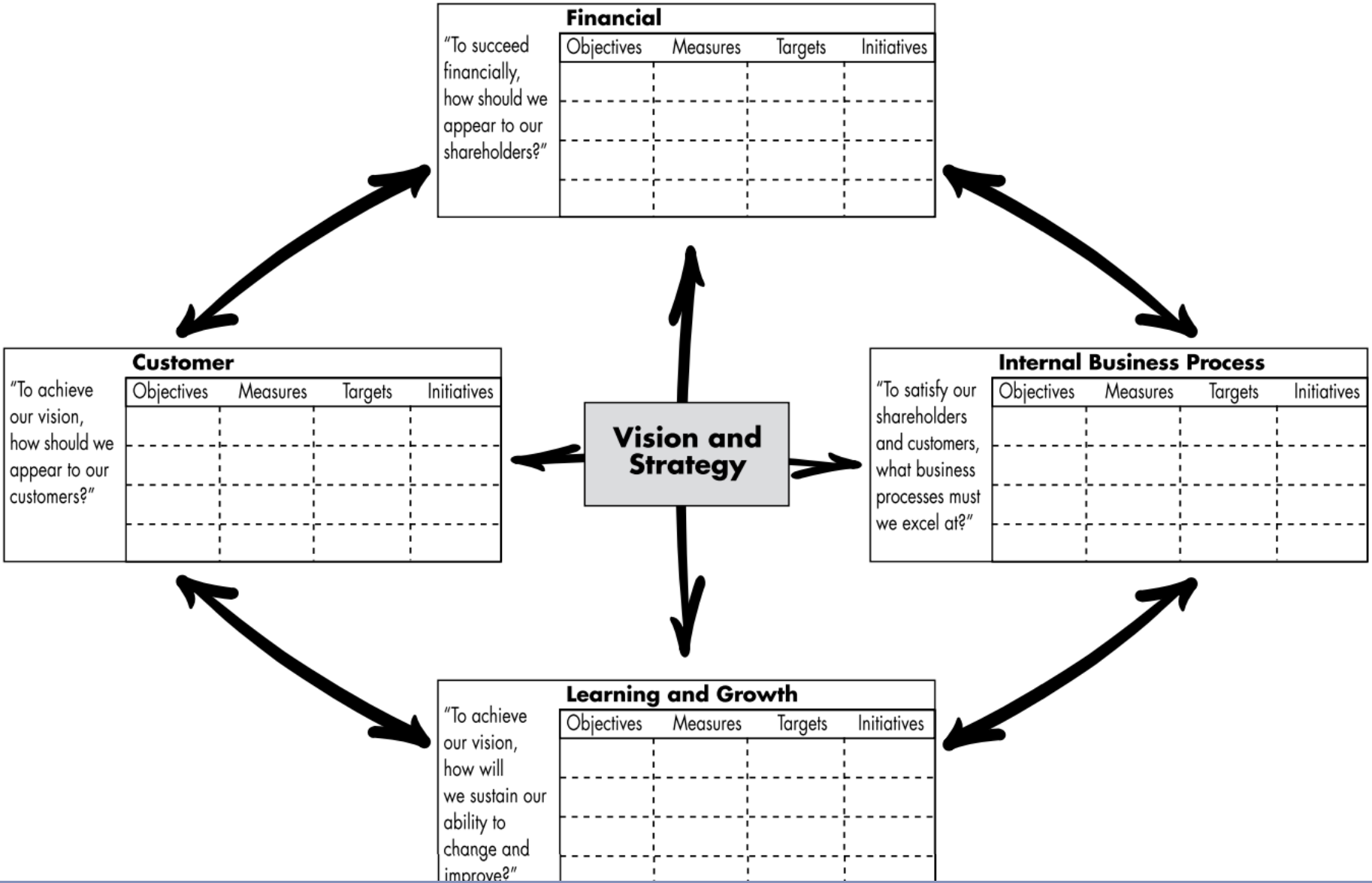
Balanced Scorecard



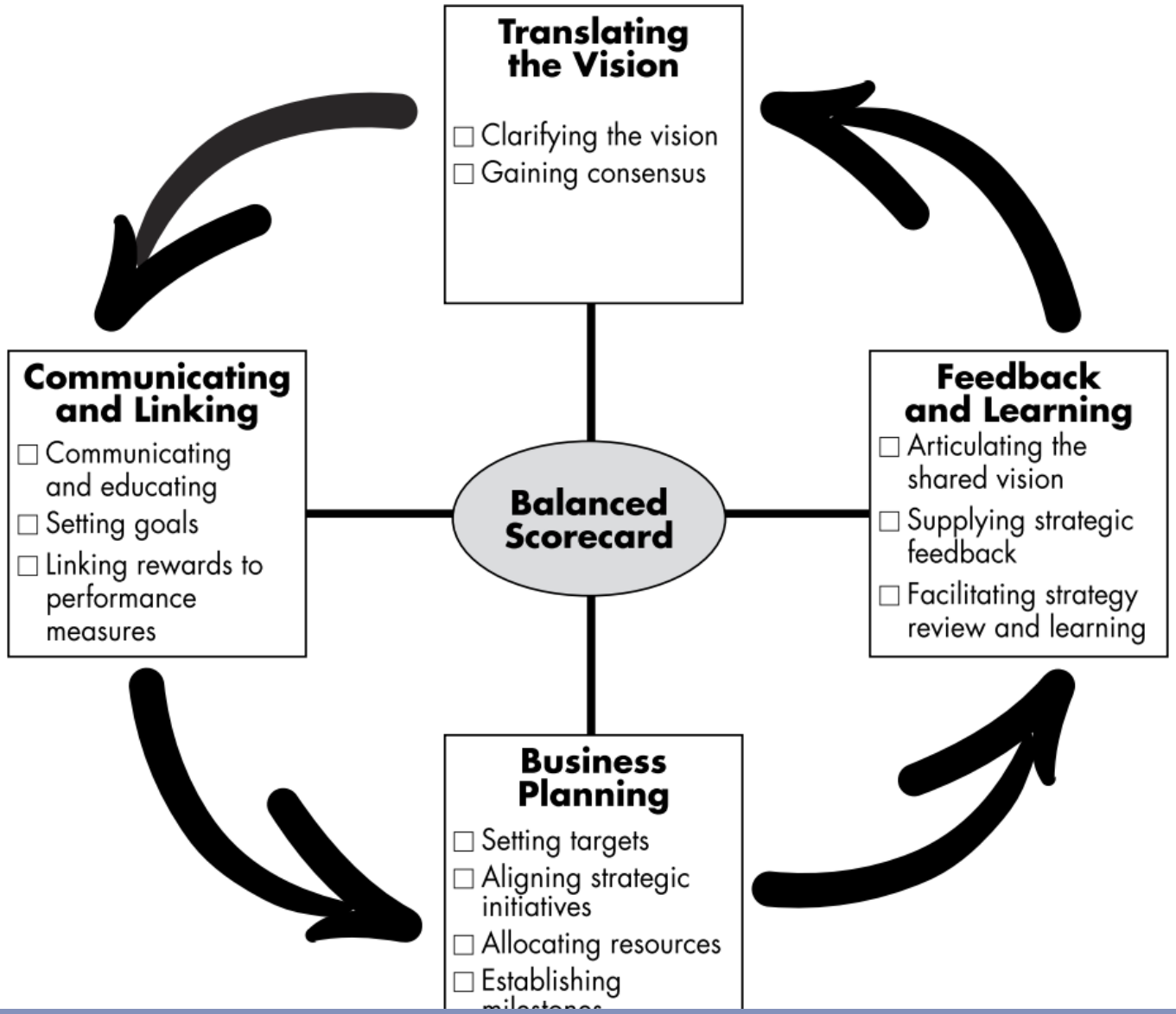
Balanced Scorecard

- Four perspectives:
  - Financial
  - Customer
  - Processes
  - Learning and Growth
- Four elements per perspective
  - Objectives
  - Measures
  - Targets
  - Initiatives

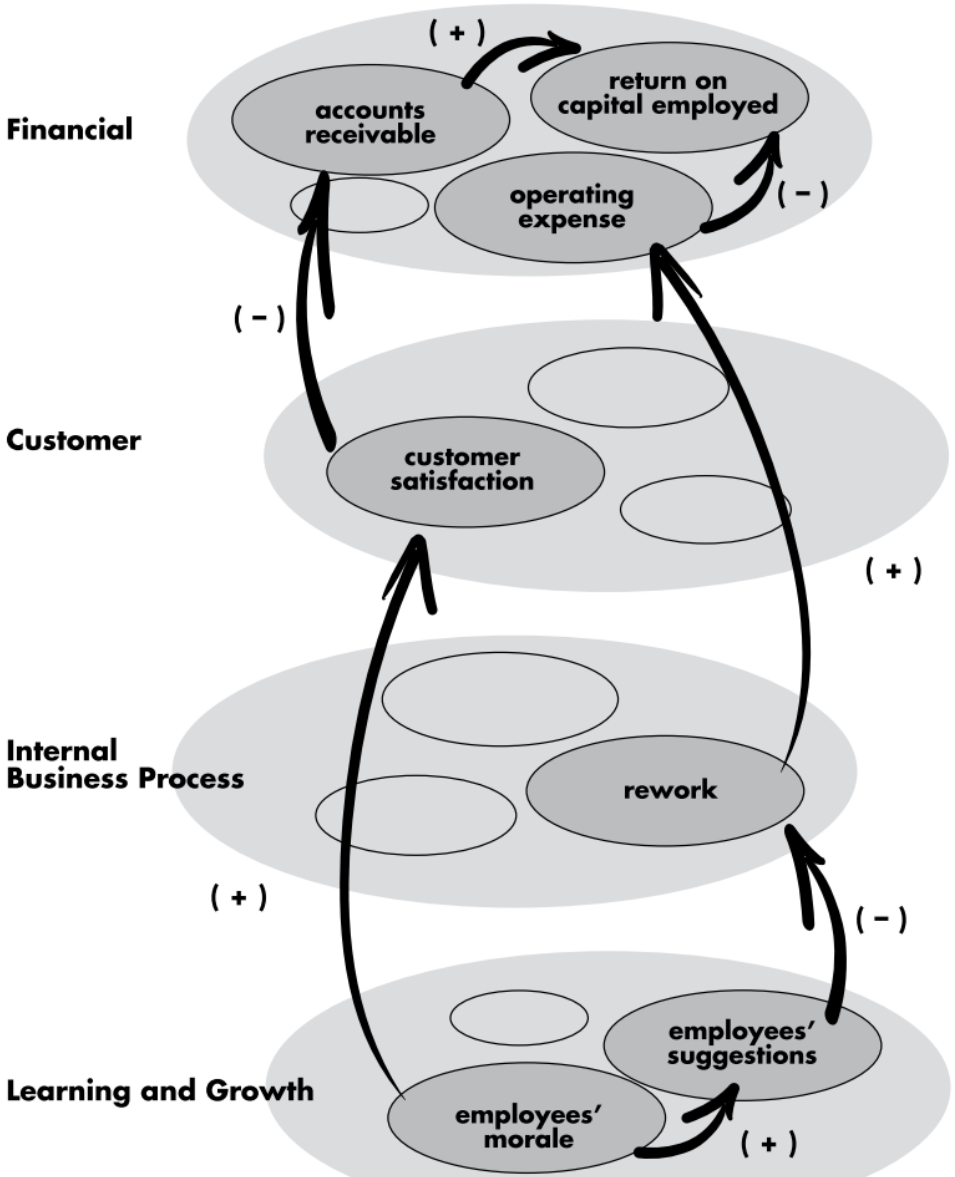
Balanced Scorecard Perspectives



Balanced Scorecard Processes



Relations between Perspectives



Some terms . . .

- **Business Activity Monitoring (BAM)**: insight in operational status and events of a business
- **Complex Event Processing (CEP)**: monitor events and react immediately if a pattern occurs
- **Corporate Performance Management (CPM)**: measuring the (financial) performance of a process or organization

Some systems . . .

- Enterprise Resource Planning (ERP) Systems
- Enterprise Information Systems (EIS)
- Business Information Systems (BIS)
- Management Information System (MIS)
- Executive Information System (EIS)

ETL

- **E**xtract data from source systems: generate dumps, exports, etc.
- **T**ransform data: aggregating, linking, sorting, joining, etc.
- **L**oading data into target system into desired (reporting) format

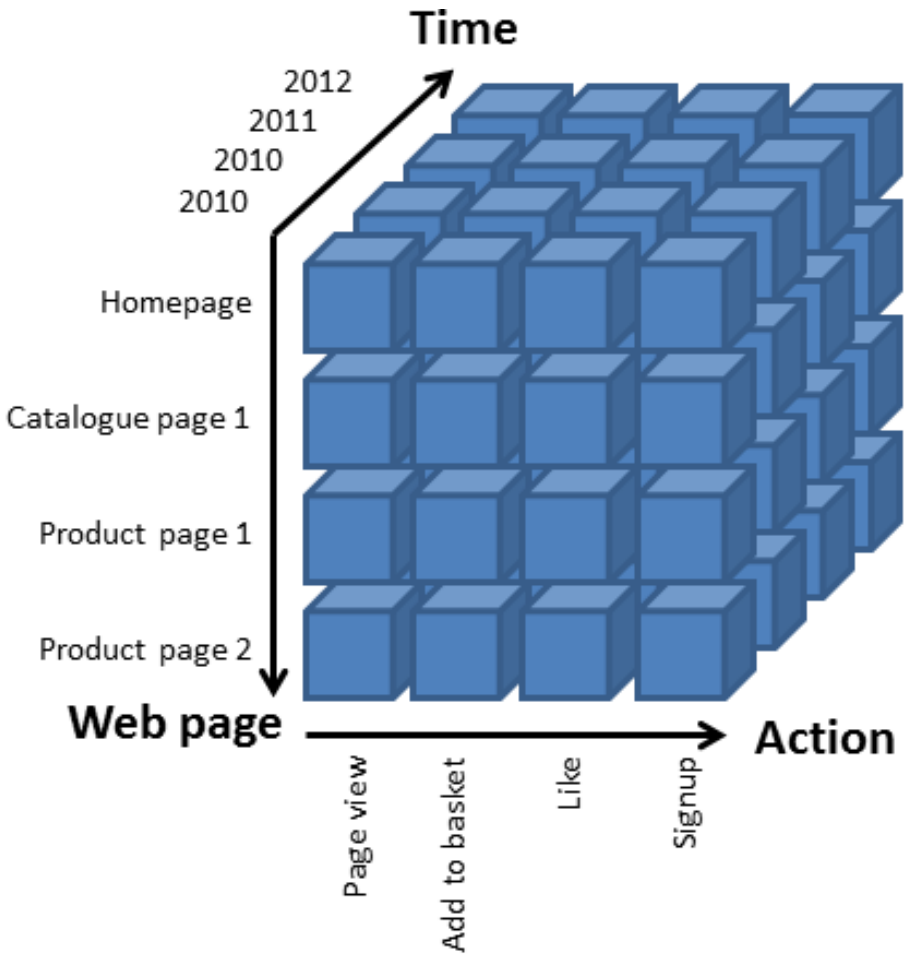
OLAP

- **OnLine Analytical Processing (OLAP)**
- Given a data table with n attributes:
  - Dimensions of an $(n-1)$-dimensional cube represent $n-1$ attributes of the data
  - Value in a cell of the cube represents the remaining attribute
- Use a **slice** or **dice** to get the desired information
- Suitable for, e.g., star schema data

OLAP Example

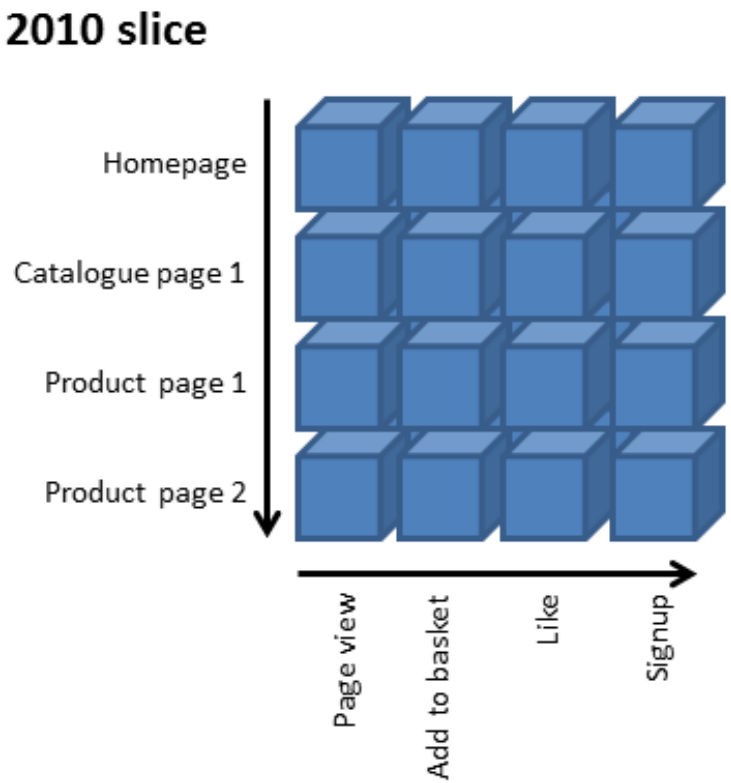
- Example: website visitor logs, storing:
  1. `Time`
  2. `Web page`
  3. `Action`
  4. `Conversion`

OLAP Cube Example



http://snowplowanalytics.com

OLAP Cube Example Slice



http://snowplowanalytics.com

OLAP Cube Dice

http://snowplowanalytics.com

OLAP Formalized

- **OnLine Analytical Processing (OLAP)**
- Given a data table D with 4 attributes W, X, Y and Z
- An OLAP cube can be characterized as a function $f : (X, Y, Z) \rightarrow W$
- An example of a **slice** is a function $g : (Y, Z) \rightarrow W$
- Given subsets $X' \subseteq X$ and $Z' \subseteq Z$ a **dice** is a function $h : (X', Y, Z') \rightarrow W$

Universiteit Leiden

Break?

Universiteit Leiden

Visual Analytics

What is Visualization?

- Intuition: data is more than its raw bits and bytes
- **Visualization**: making something visible to the eye (Oxford dictionary)
- *All visualizations share a common "DNA" — a set of mappings between data properties and visual attributes such as position, size, shape, and color — and customized species of visualization might always be constructed by varying these encodings.*

  Heer et al., A Tour through the Visualization Zoo, CACM 53(6): 59–67, 2010.
- **Visual Analytics**: knowledge discovery (DIKW) based on visualization

What is Visualization?

- **Data properties**: attributes of (groups of) data objects
  Name; Age; City

What is Visualization?

- **Data properties**: attributes of (groups of) data objects
  Name; Age; City
  Frank; 28; "Niels Bohrweg 1, Leiden"

What is Visualization?

- **Data properties**: attributes of (groups of) data objects
  Name; Age; City
  Frank; 28; "Niels Bohrweg 1, Leiden"
- **Visual attributes**: e.g., position, size, shape, label, color, etc.
  Label; Size; Position
- **Visualization**: mapping data properties to visual attributes

What is Visualization?

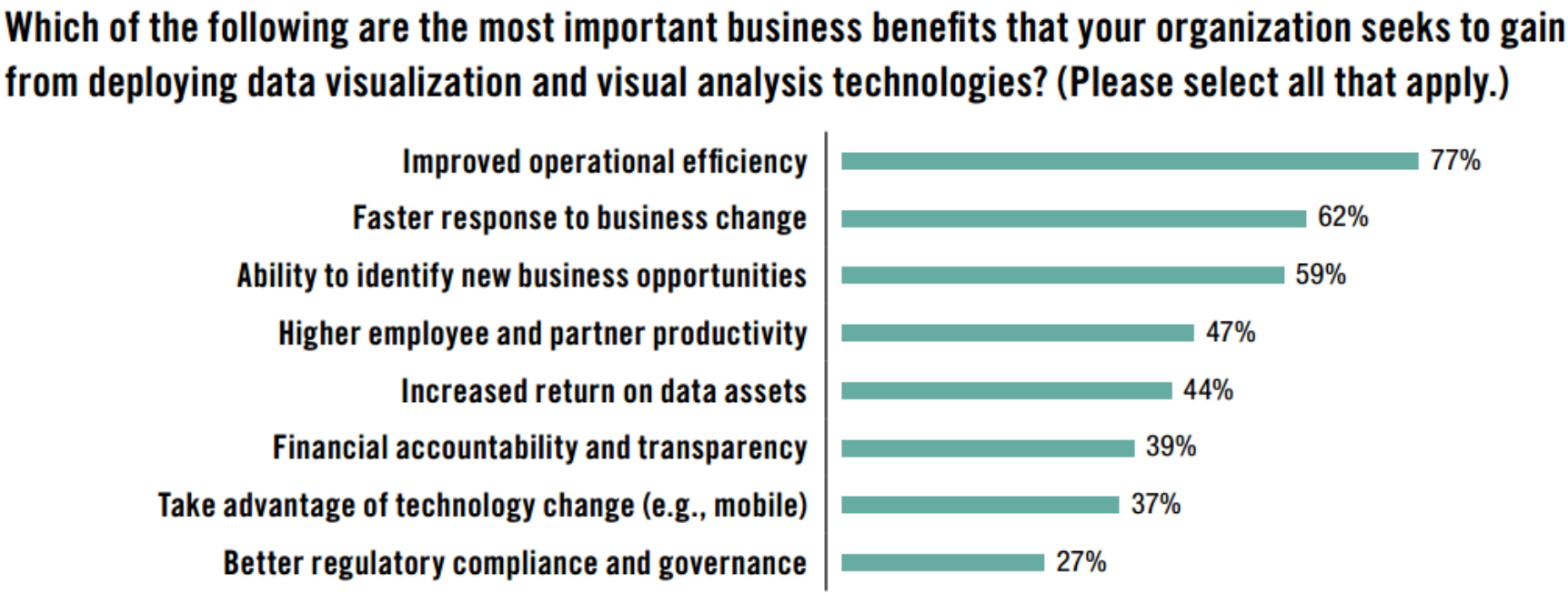
- **Data properties**: attributes of (groups of) data objects
  Name; Age; City
  Frank; 28; "Niels Bohrweg 1, Leiden"
- **Visual attributes**: e.g., position, size, shape, label, color, etc.
  Label; Size; Position
- **Visualization**: mapping data properties to visual attributes
  Name $\rightarrow$ Label
  Age $\rightarrow$ Size
  City $\rightarrow$ Position

What is Visualization?

- **Data properties**: attributes of (groups of) data objects
  Name; Age; City
  Frank; 28; "Niels Bohrweg 1, Leiden"
- **Visual attributes**: e.g., position, size, shape, label, color, etc.
  Label; Size; Position
- **Visualization**: mapping data properties to visual attributes
  Name $\rightarrow$ Label
  Age $\rightarrow$ Size
  City $\rightarrow$ Position
  `"Frank"`, $log_2(28)$, $(52.1603216, 4.4939262)$

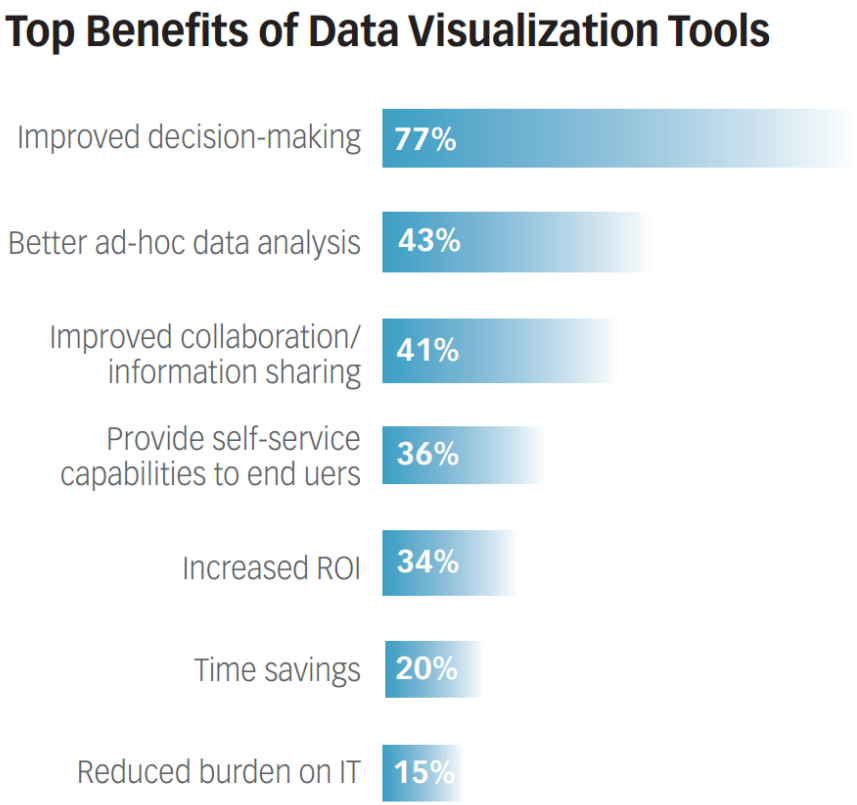
What is Visualization?

Why visualization?



SAS, Data Visualization: Making Big Data Approachable and Valuable, 2014

Why visualization?



SAS, Data Visualization: Making Big Data Approachable and Valuable, 2014

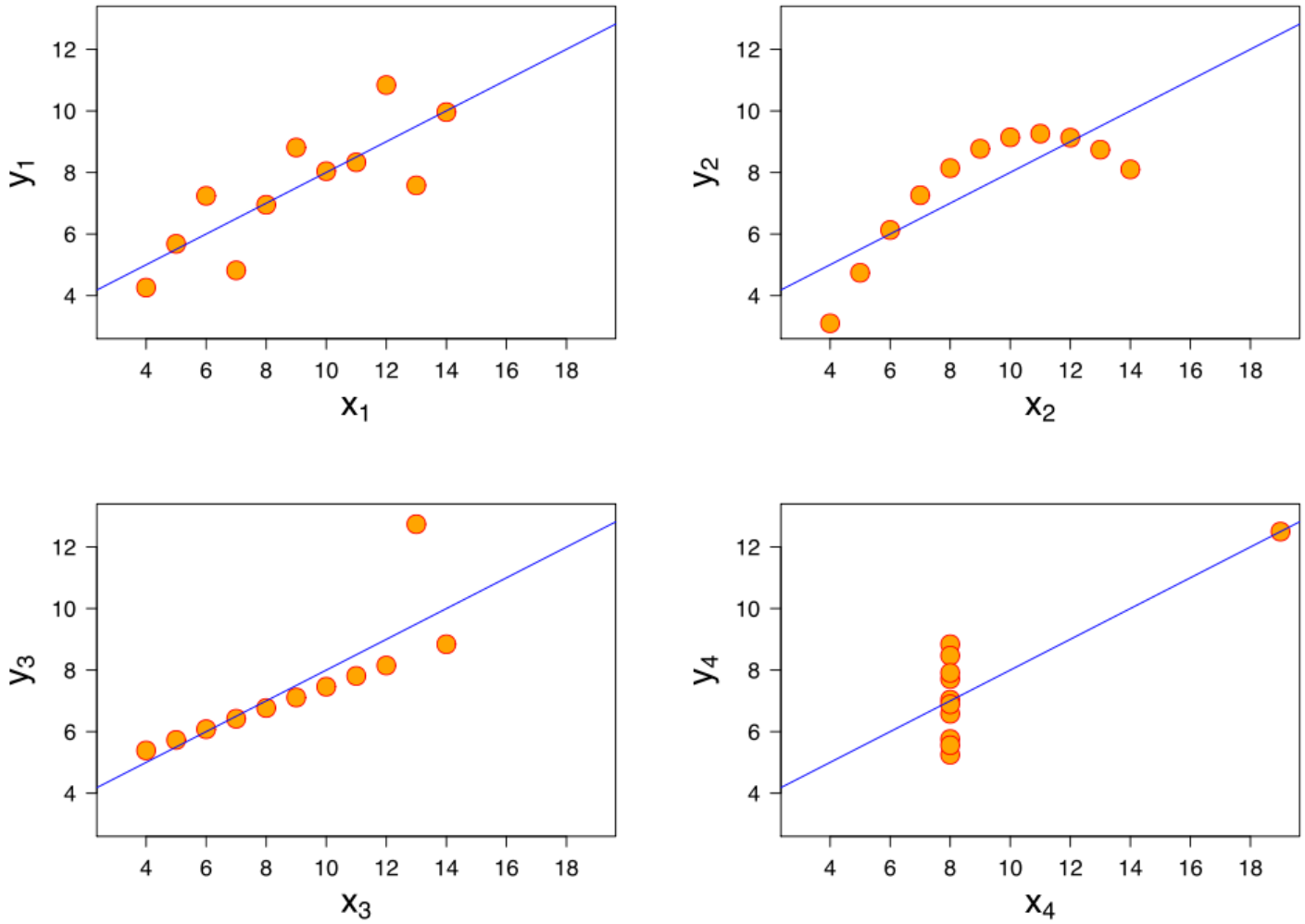
Visualization theory

- Discrete vs. continuous data
- Categorical vs. quantitative data
- Mean or median?
- Variance?
- Correlations? Regression?
- Normal distribution or power law?
- The **correct visualization** depends on the data itself!

Listen to the data to . . .

- Catch mistakes
- See patterns
- Find violations of statistical assumptions
- Generate hypotheses
- Do outlier detection

Anscombe's quartet

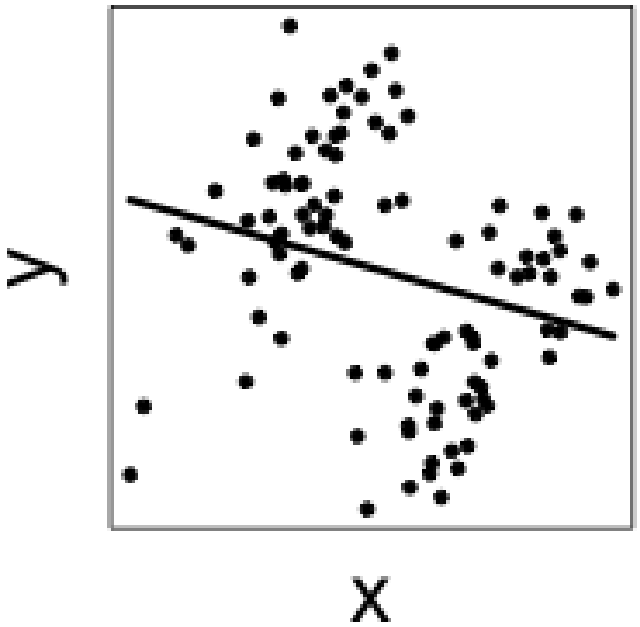


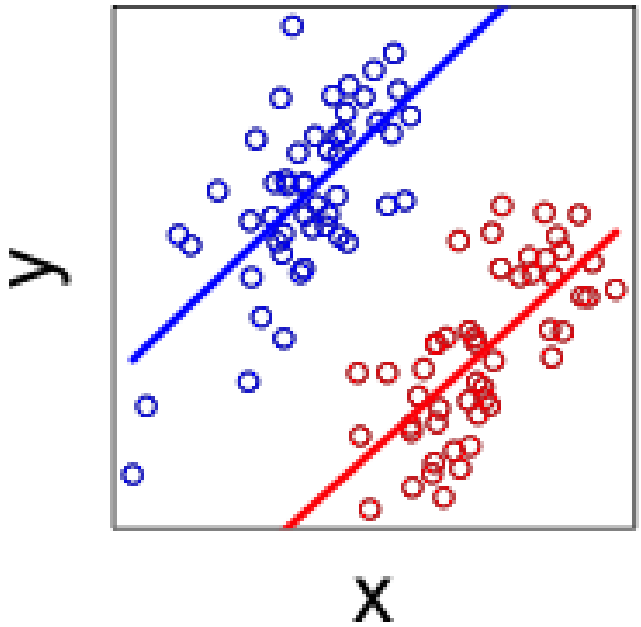
F. J. Anscombe. Graphs in Statistical Analysis. *American Statistician* 27 (1): 1721, 1973

Anscombe's quartet

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.0 + 0.50x$	to 2 decimal places

Simpson's Paradox





Visualization Quality

- When is a certain visualization **“good”**?

Visualization Quality

- When is a certain visualization **“good”**?
- “Proper mapping of data properties to visual attributes”?

Visualization Quality

- When is a certain visualization **"good"**?
- "Proper mapping of data properties to visual attributes"?
- The number of data properties (variables) that is visualized?

Visualization Quality

- When is a certain visualization **"good"**?
- "Proper mapping of data properties to visual attributes"?
- The number of data properties (variables) that is visualized?
- The number of visual attributes that is utilized?

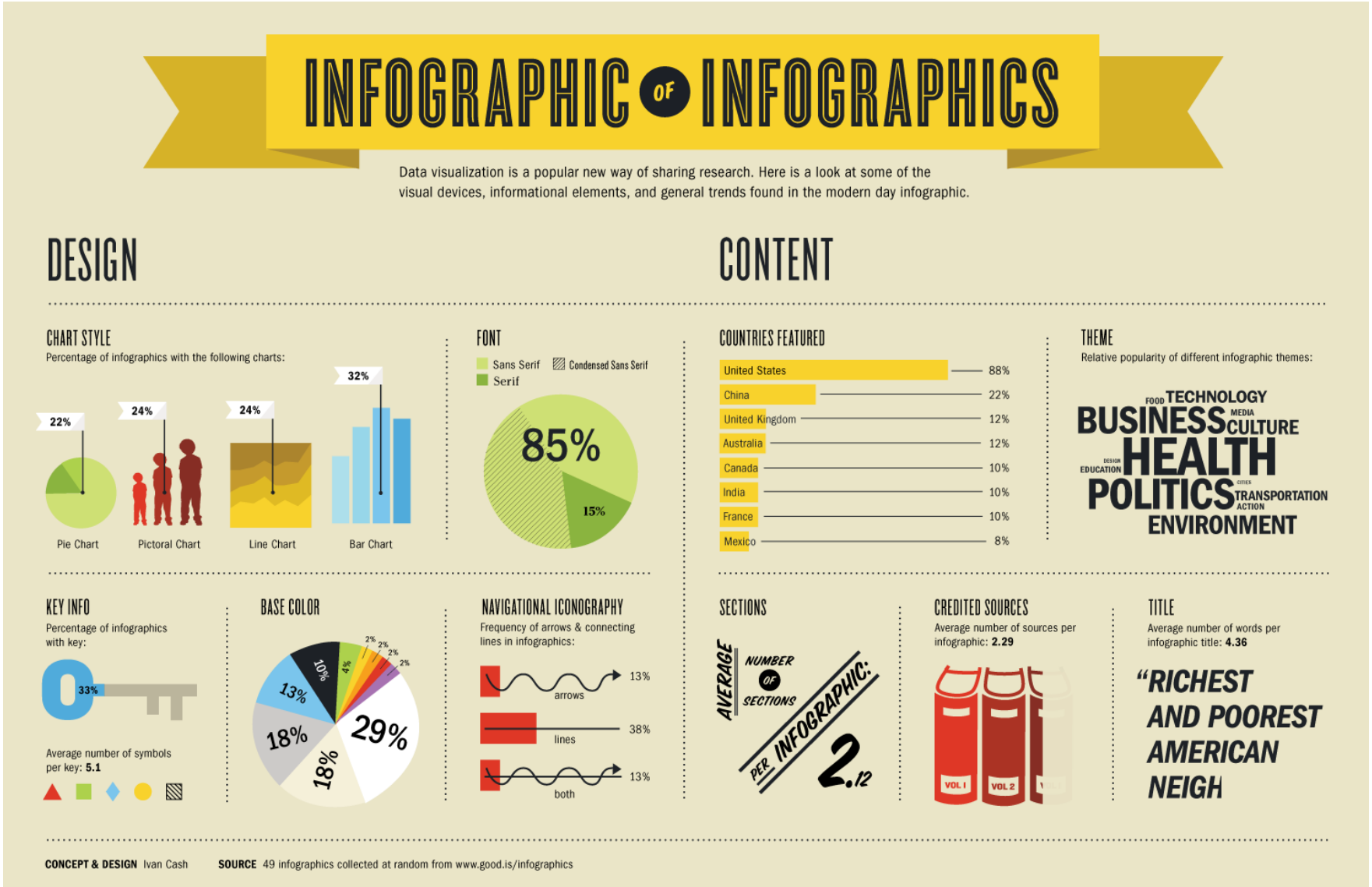
Visualization Quality

- When is a certain visualization **"good"**?
- "Proper mapping of data properties to visual attributes"?
- The number of data properties (variables) that is visualized?
- The number of visual attributes that is utilized?
- Aesthetics?
- ...

Visualization Quality

- When is a certain visualization **"good"**?
- "Proper mapping of data properties to visual attributes"?
- The number of data properties (variables) that is visualized?
- The number of visual attributes that is utilized?
- Aesthetics?
- ...
- Hard to answer objectively!

Infographic of infographics

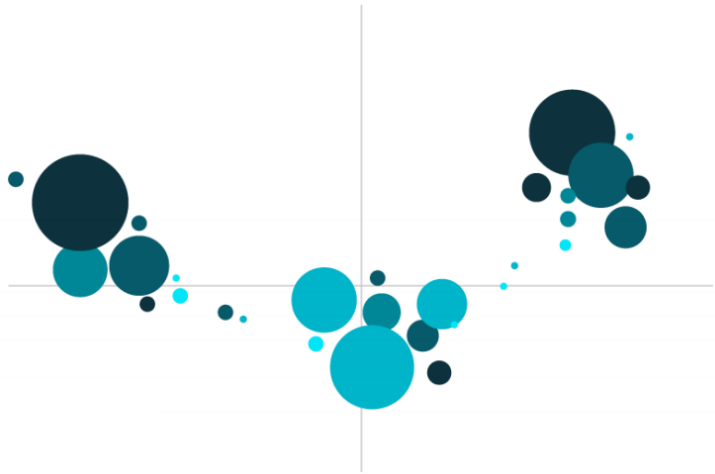


Visualization Metaphors

- Important is Big
- Happy is Up
- More is Up
- Categories Are Containers
- Organization is Physical Structure
- Similarity is Closeness
- Control is Up

Visualization Metaphors

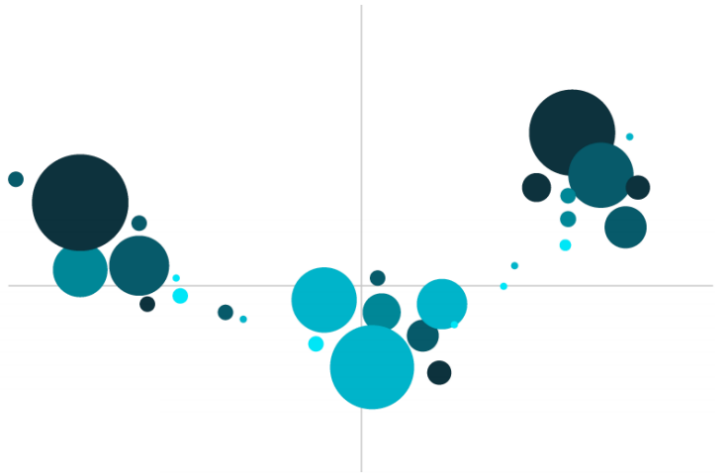
- Important is Big
- Happy is Up
- More is Up
- Categories Are Containers
- Organization is Physical Structure
- Similarity is Closeness
- Control is Up



http://www.bostondatafest.com/wp-content/uploads/2013/11/big_data_viz.pdf

Visualization Metaphors

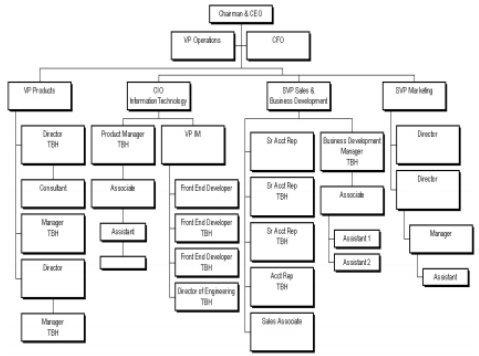
- **Important is Big**
- Happy is Up
- More is Up
- Categories Are Containers
- Organization is Physical Structure
- **Similarity is Closeness**
- Control is Up

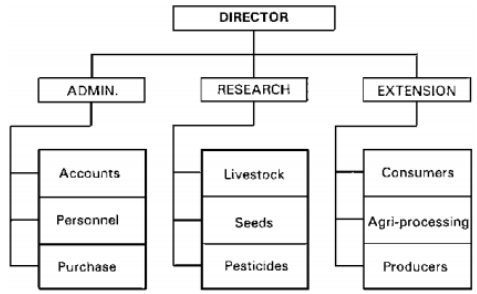


http://www.bostondatafest.com/wp-content/uploads/2013/11/big_data_viz.pdf

Visualization Metaphors

- Important is Big
- Happy is Up
- More is Up
- Categories Are Containers
- Organization is Physical Structure
- Similarity is Closeness
- Control is Up

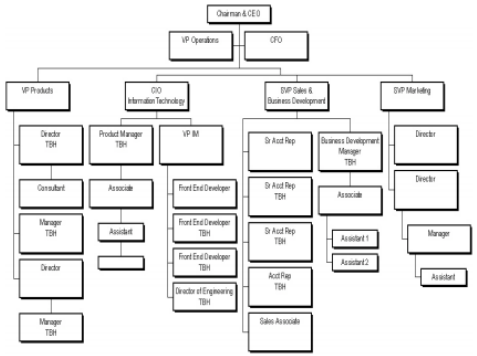


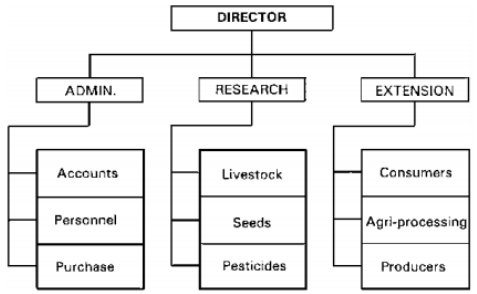


http://www.bostondatafest.com/wp-content/uploads/2013/11/big_data_viz.pdf

Visualization Metaphors

- Important is Big
- Happy is Up
- More is Up
- Categories Are Containers
- **Organization is Physical Structure**
- **Similarity is Closeness**
- **Control is Up**





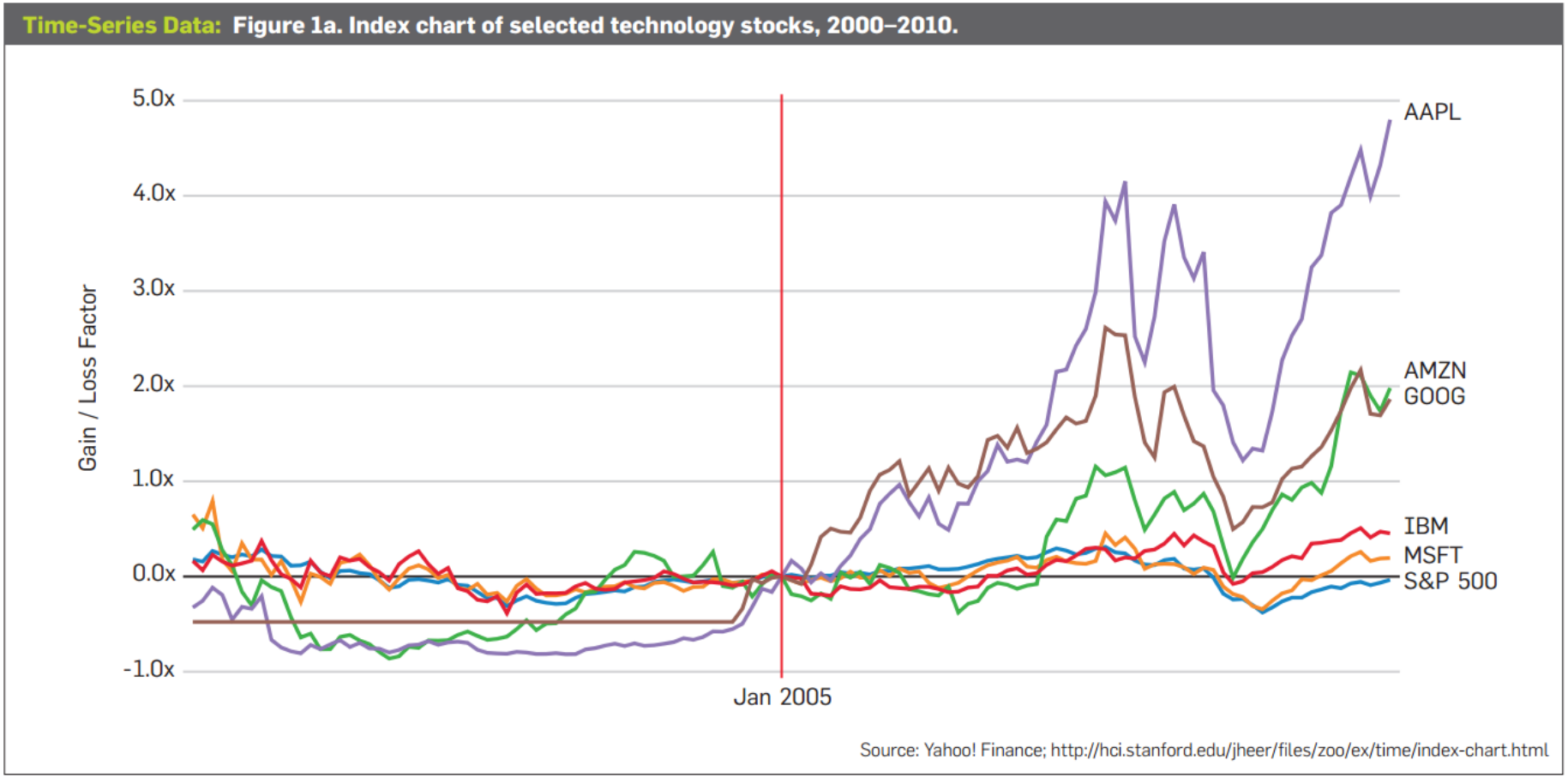
http://www.bostondatafest.com/wp-content/uploads/2013/11/big_data_viz.pdf

Examples

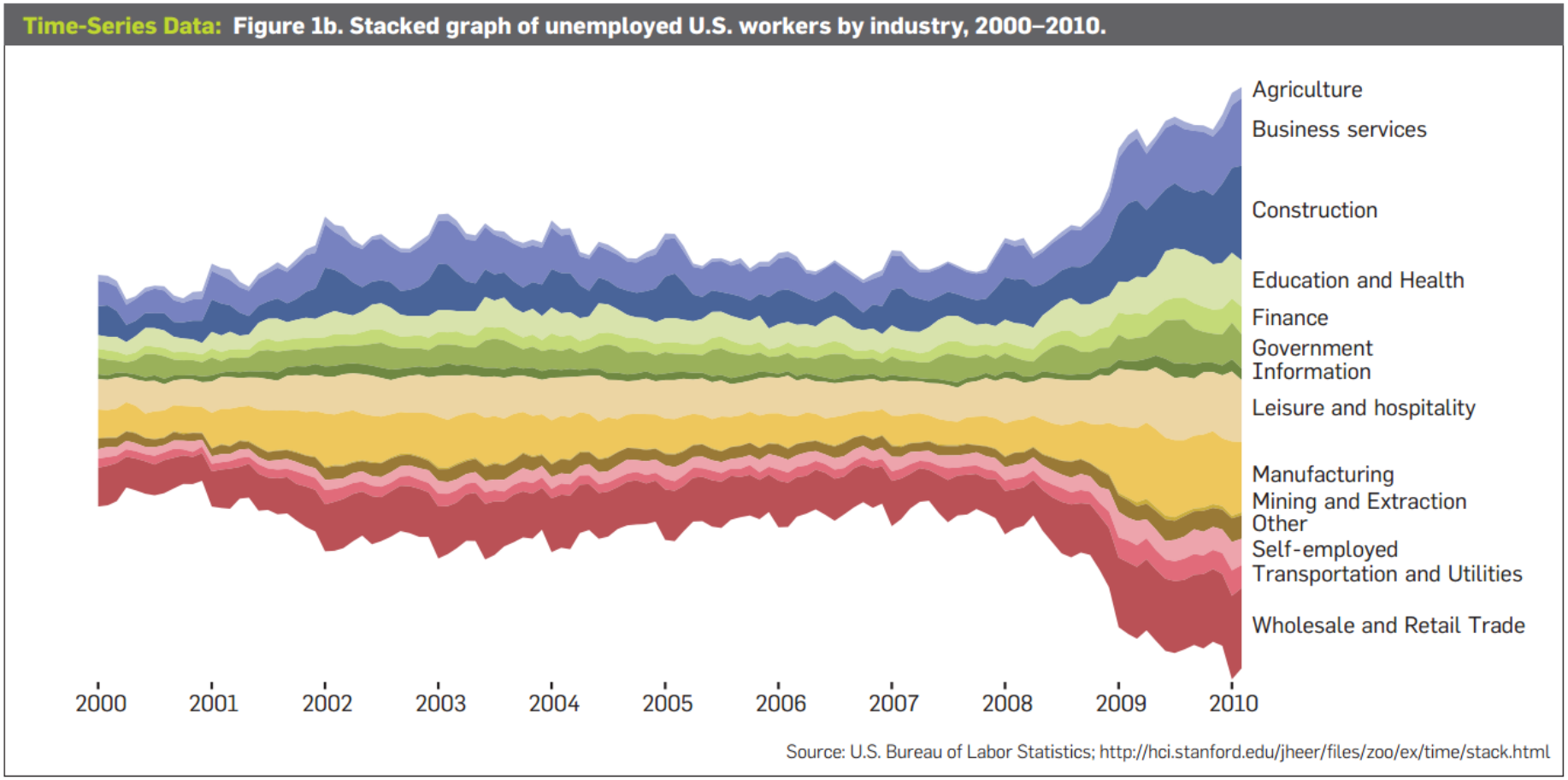
- **Time-series data**
- Statistical data
- Geographical data
- Hierarchical data
- Network data

Heer et al., A Tour through the Visualization Zoo, CACM 53(6): 59–67, 2010.

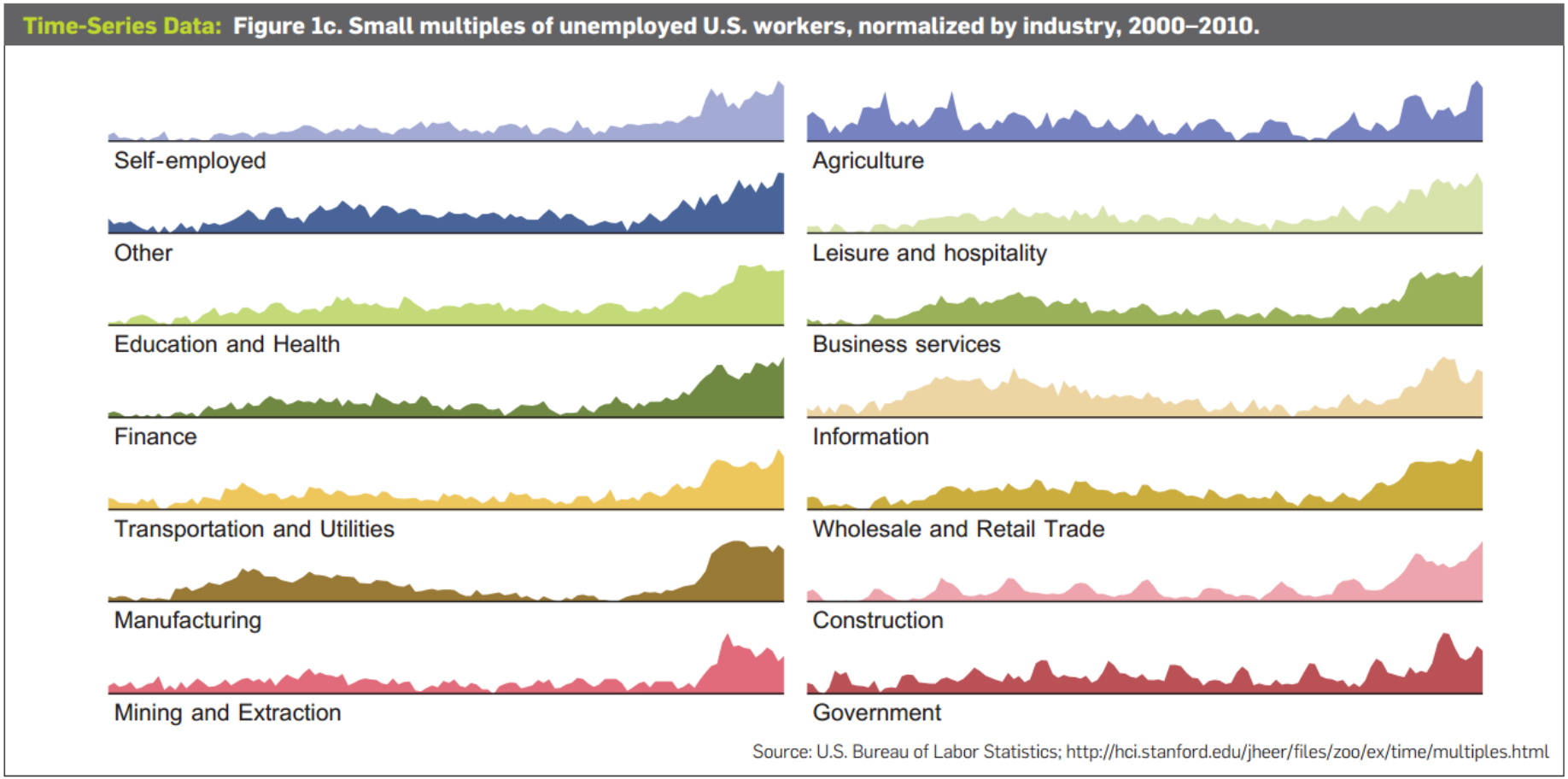
Index chart


Time-Series Data: Figure 1a. Index chart of selected technology stocks, 2000–2010.
5.0x
4.0x
3.0x
2.0x
1.0x
0.0x
-1.0x
Gain / Loss Factor
Jan 2005
AAPL
AMZN
GOOG
IBM
MSFT
S&P 500
Source: Yahoo! Finance; http://hci.stanford.edu/jheer/files/zoo/ex/time/index-chart.html

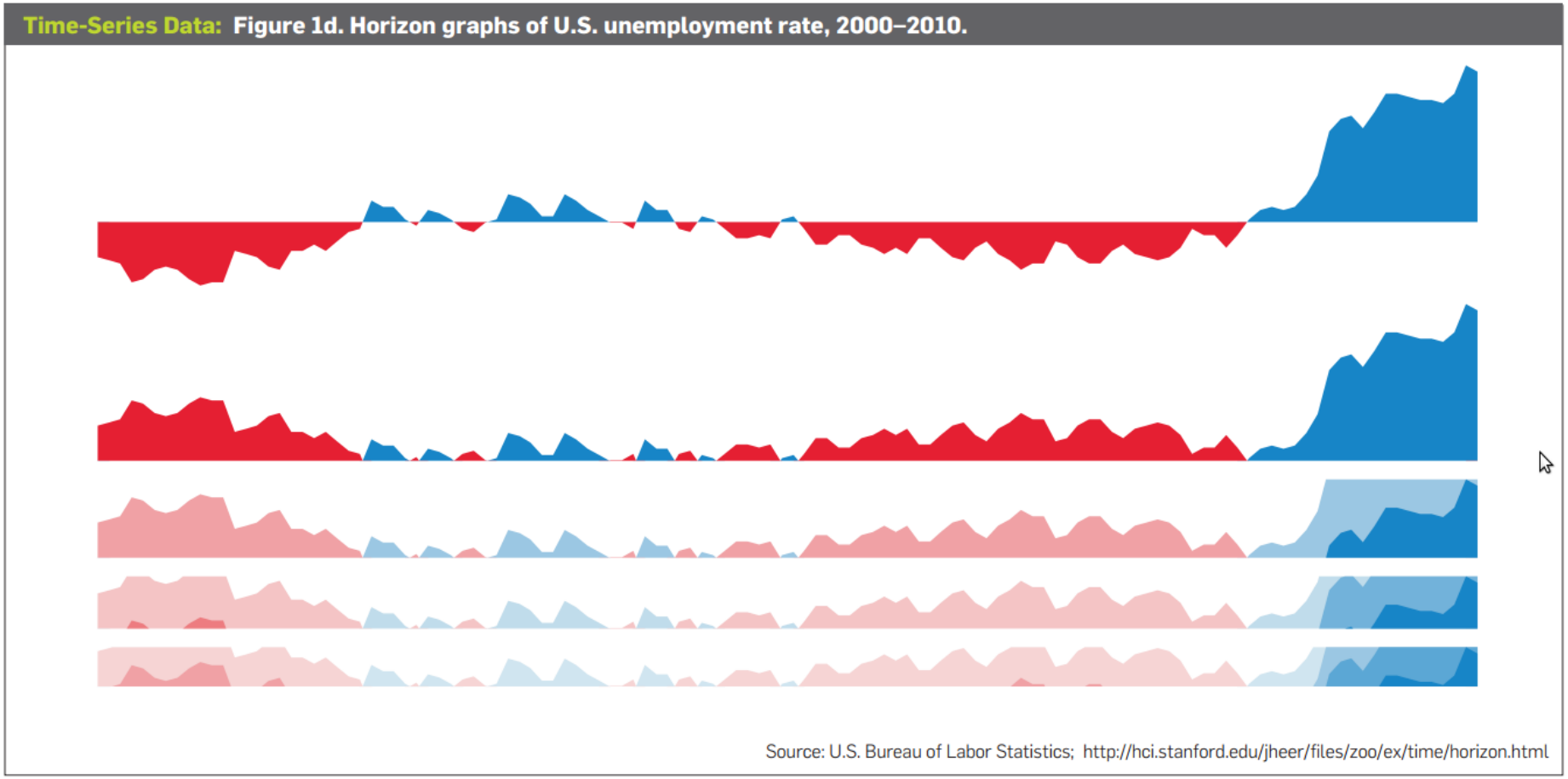
Stacked graph



Small multiples



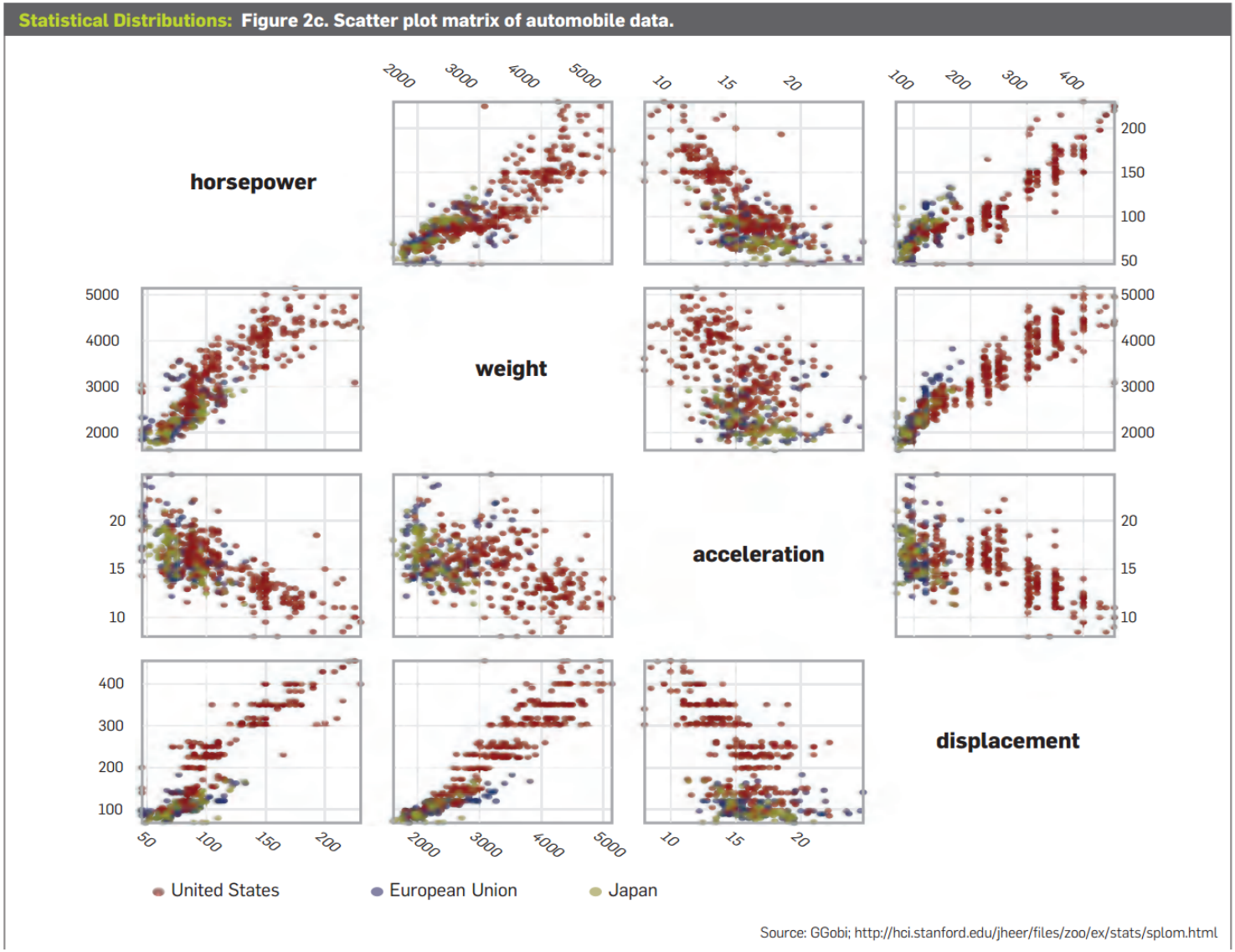
Horizon graphs



Examples

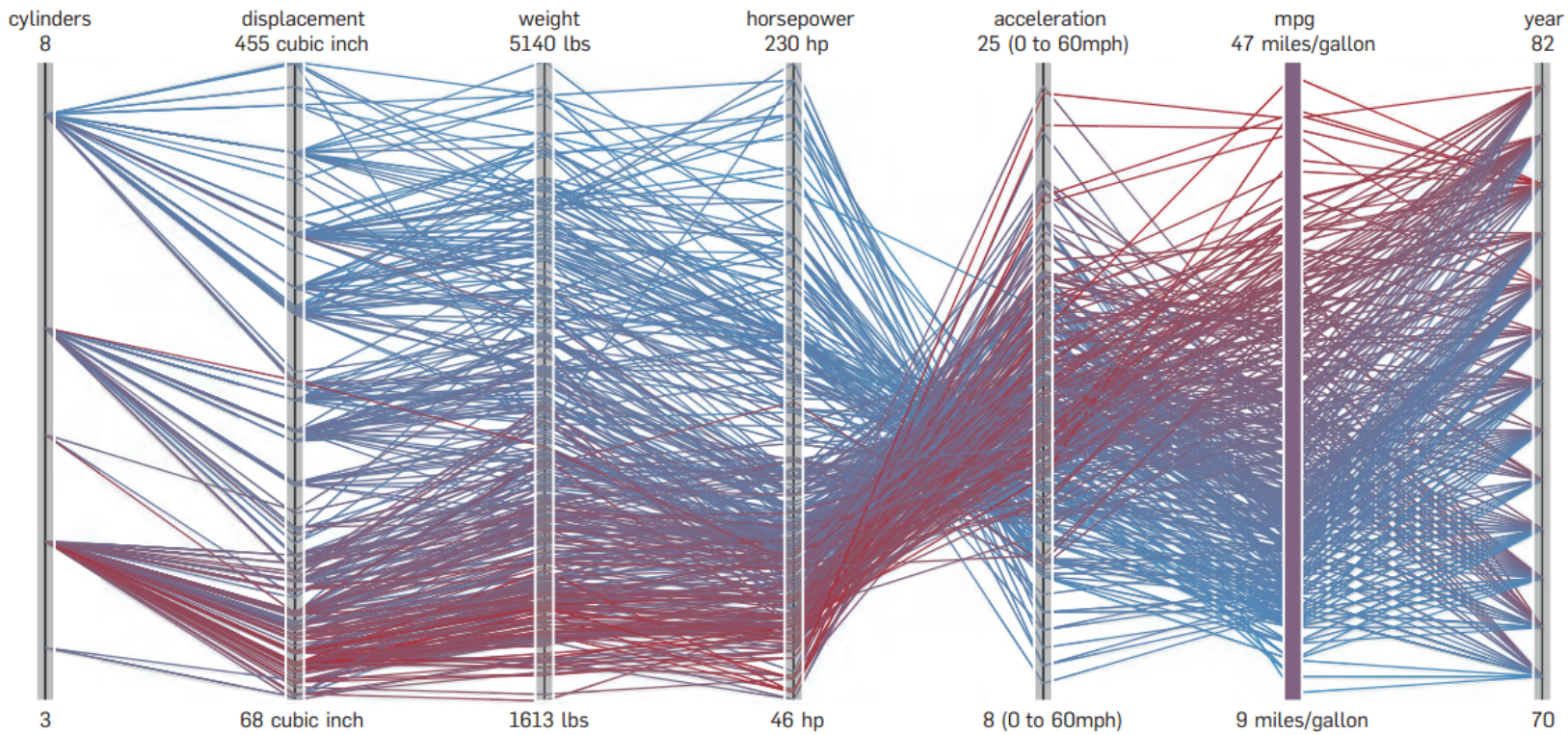
- Time-series data
- **Statistical data**
- Geographical data
- Hierarchical data
- Network data

Scatter plot matrix



Parallel coordinates

Statistical Distributions: **Figure 2d. Parallel coordinates of automobile data.**



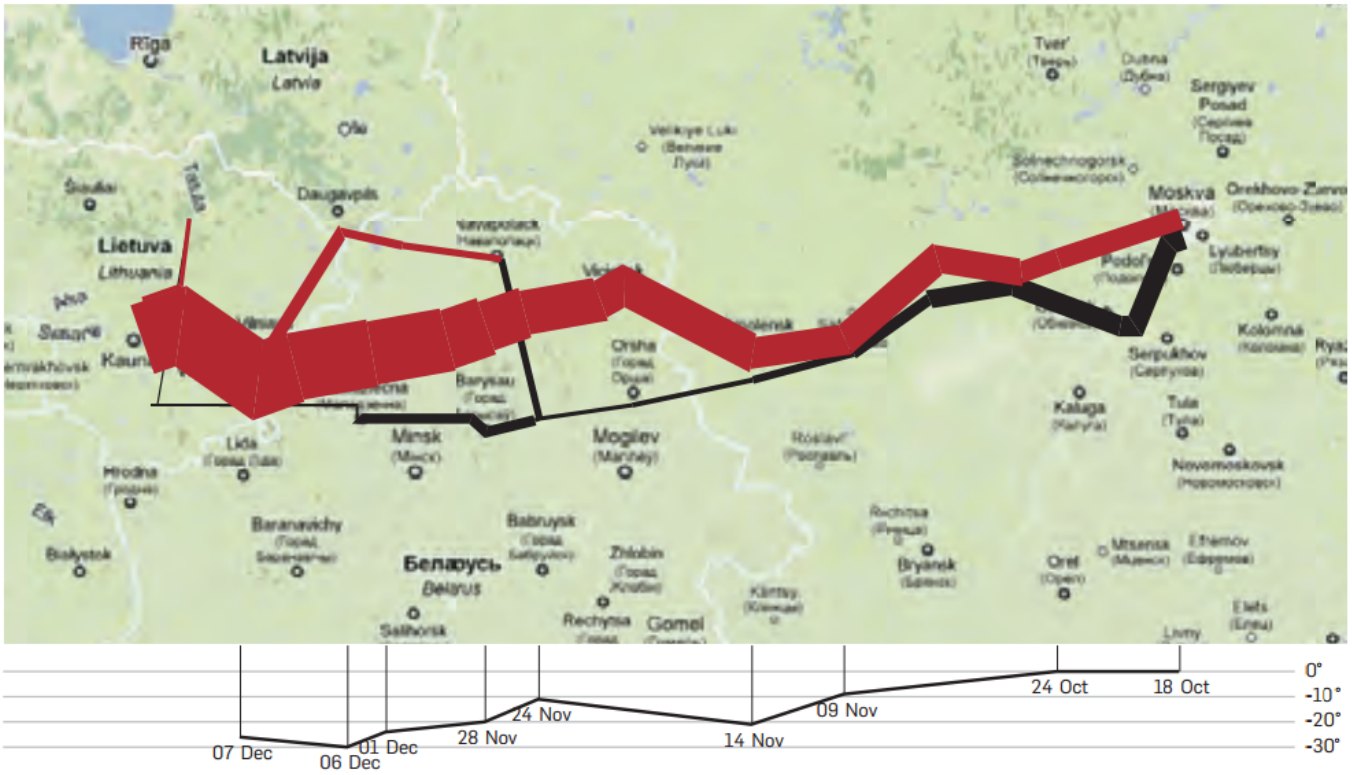
Source: GGobi; http://hci.stanford.edu/jheer/files/zoo/ex/stats/parallel.html

Examples

- Time-series data
- Statistical data
- **Geographical data**
- Hierarchical data
- Network data

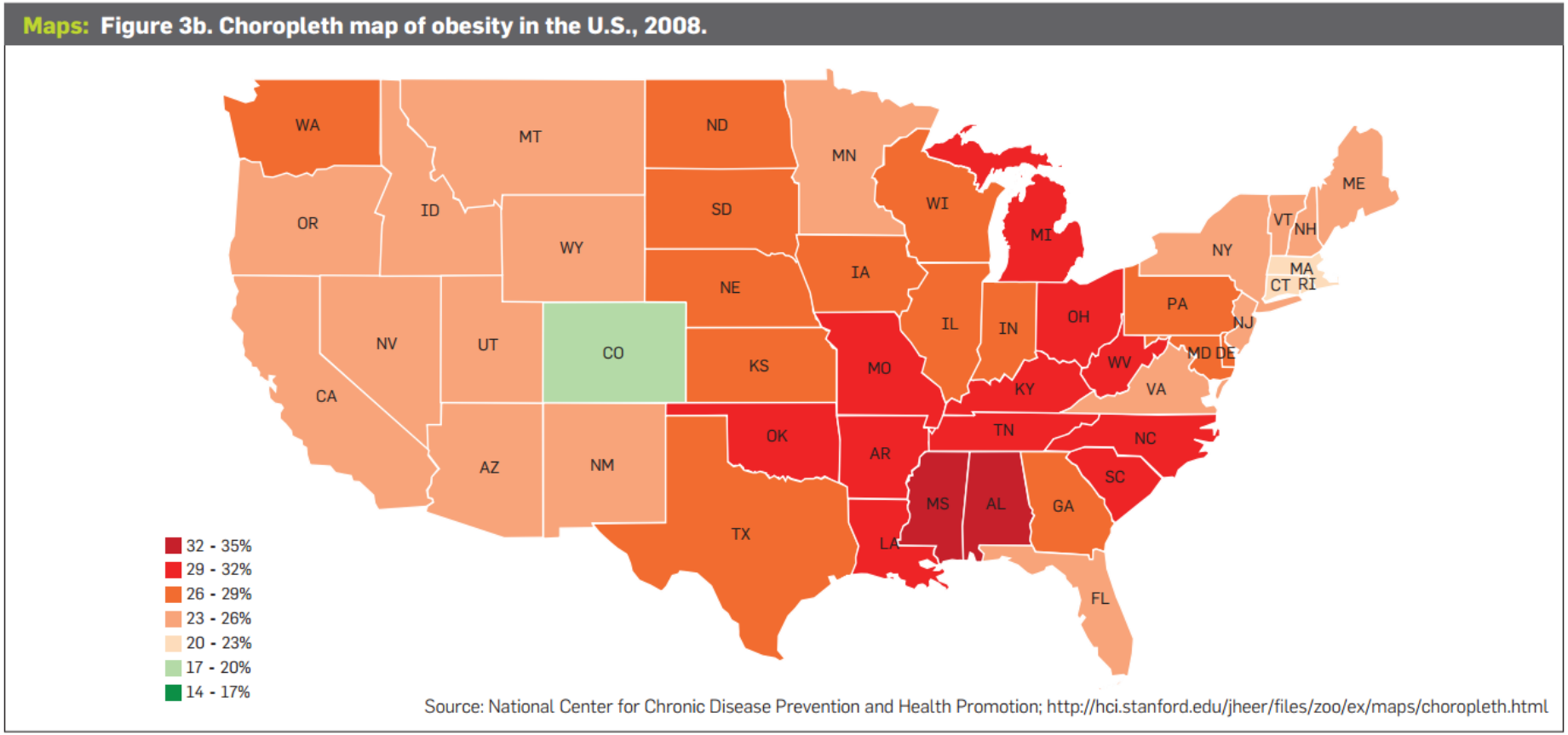
Flow maps

Maps: Figure 3a. Flow map of Napoleon's March on Moscow, based on the work of Charles Minard.

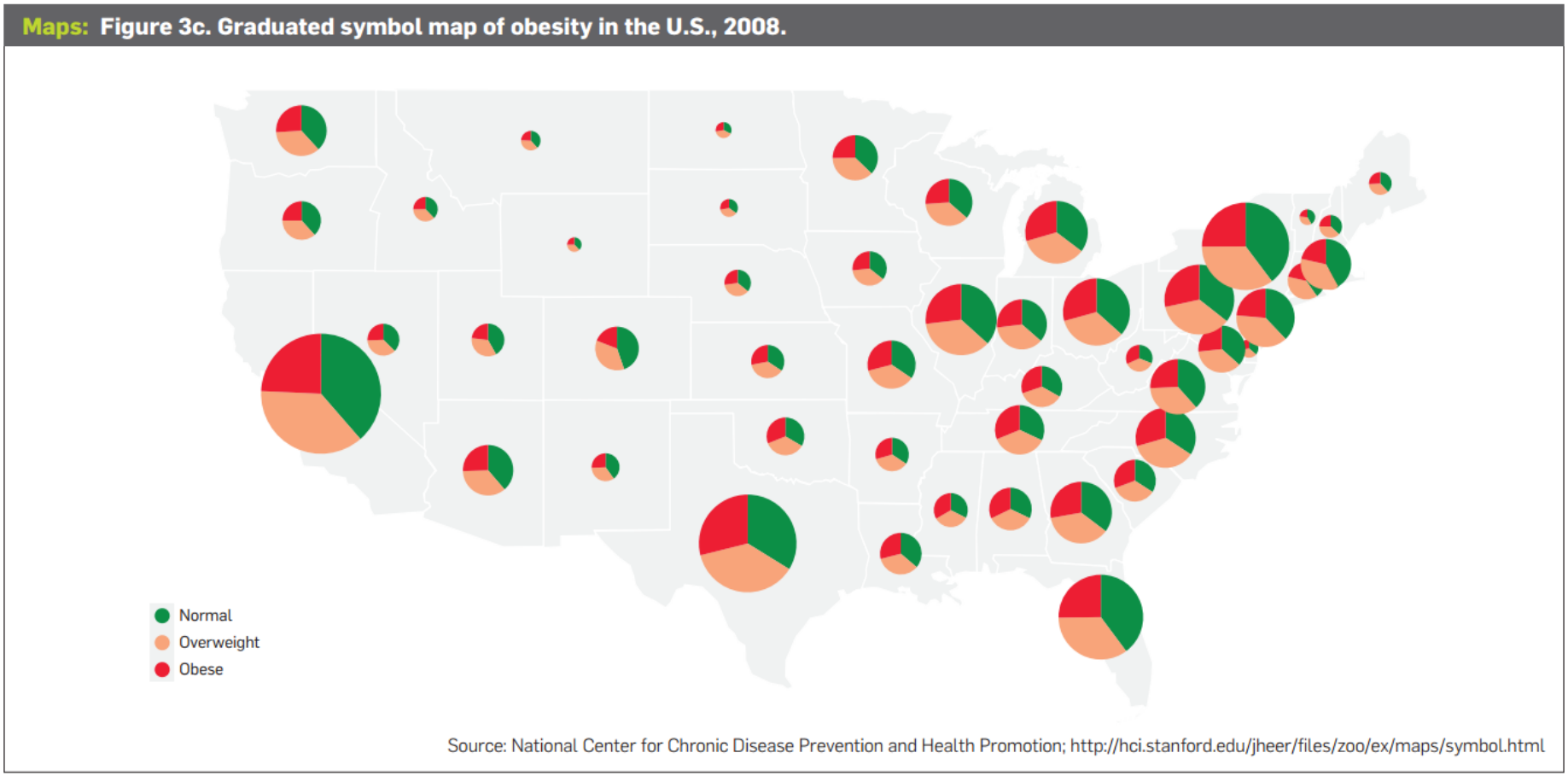


http://hci.stanford.edu/jheer/files/zoo/ex/maps/napoleon.html

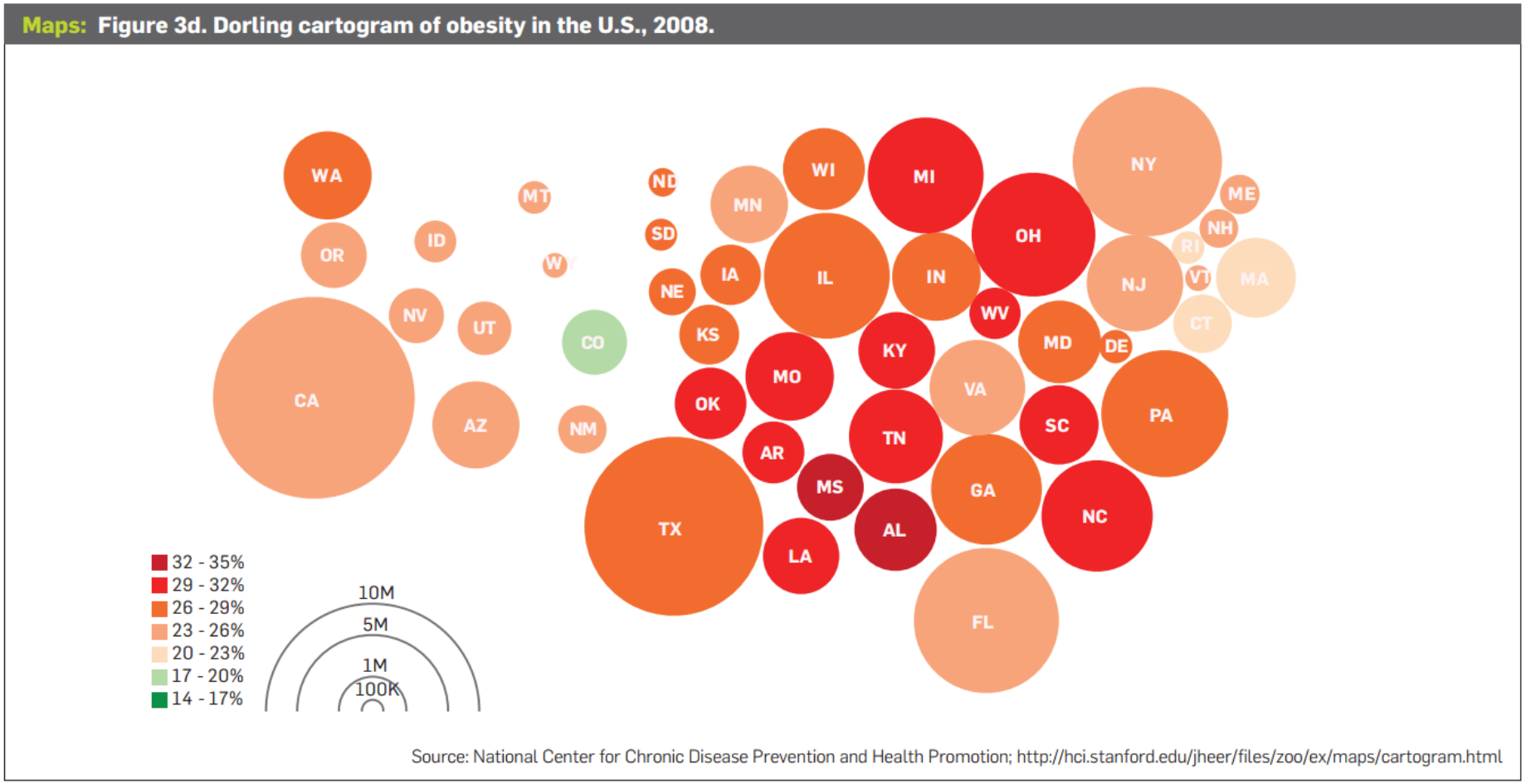
Choropleth maps



Graduated symbol maps



Cartogram

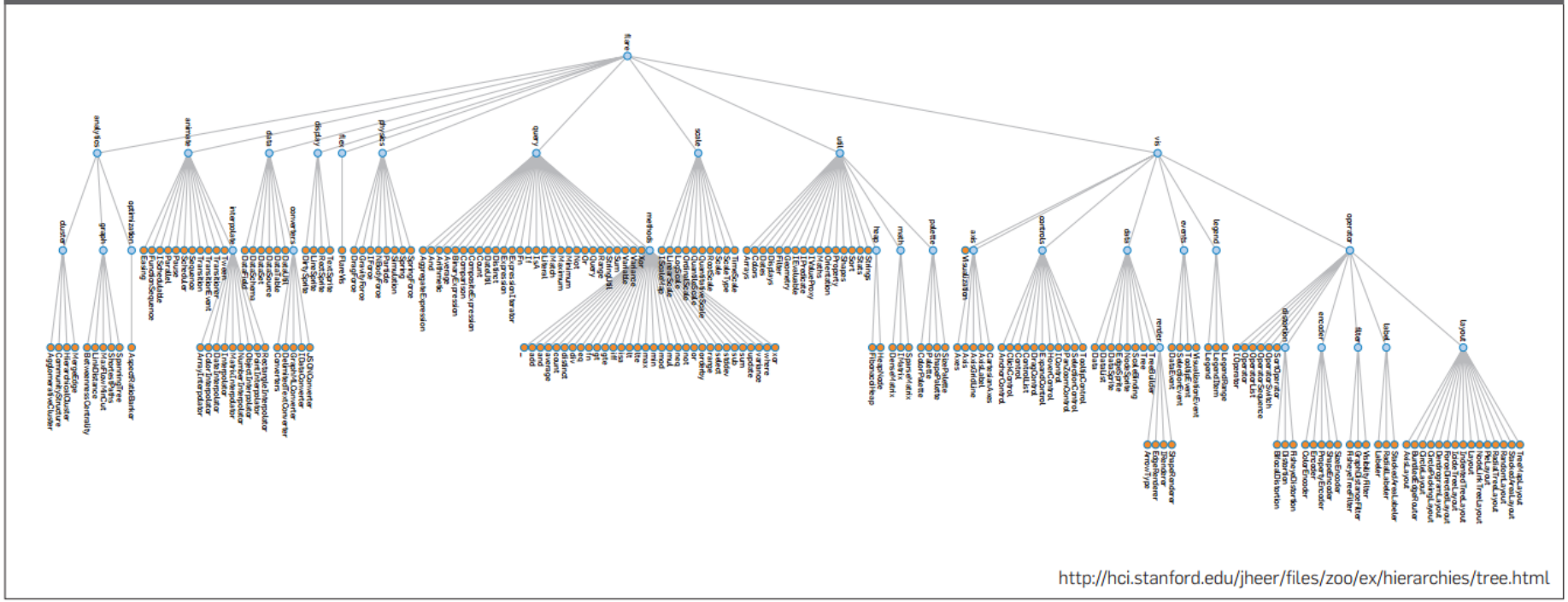

Maps: Figure 3d. Dorling cartogram of obesity in the U.S., 2008.
WA
OR
CA
ID
NV
MT
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
VA
GA
FL
NY
NJ
MD
SC
NC
PA
DE
CT
RI
VT
NH
ME
MA
32 - 35%
29 - 32%
26 - 29%
23 - 26%
20 - 23%
17 - 20%
14 - 17%
10M
5M
1M
100K
Source: National Center for Chronic Disease Prevention and Health Promotion; http://hci.stanford.edu/jheer/files/zoo/ex/maps/cartogram.html

Examples

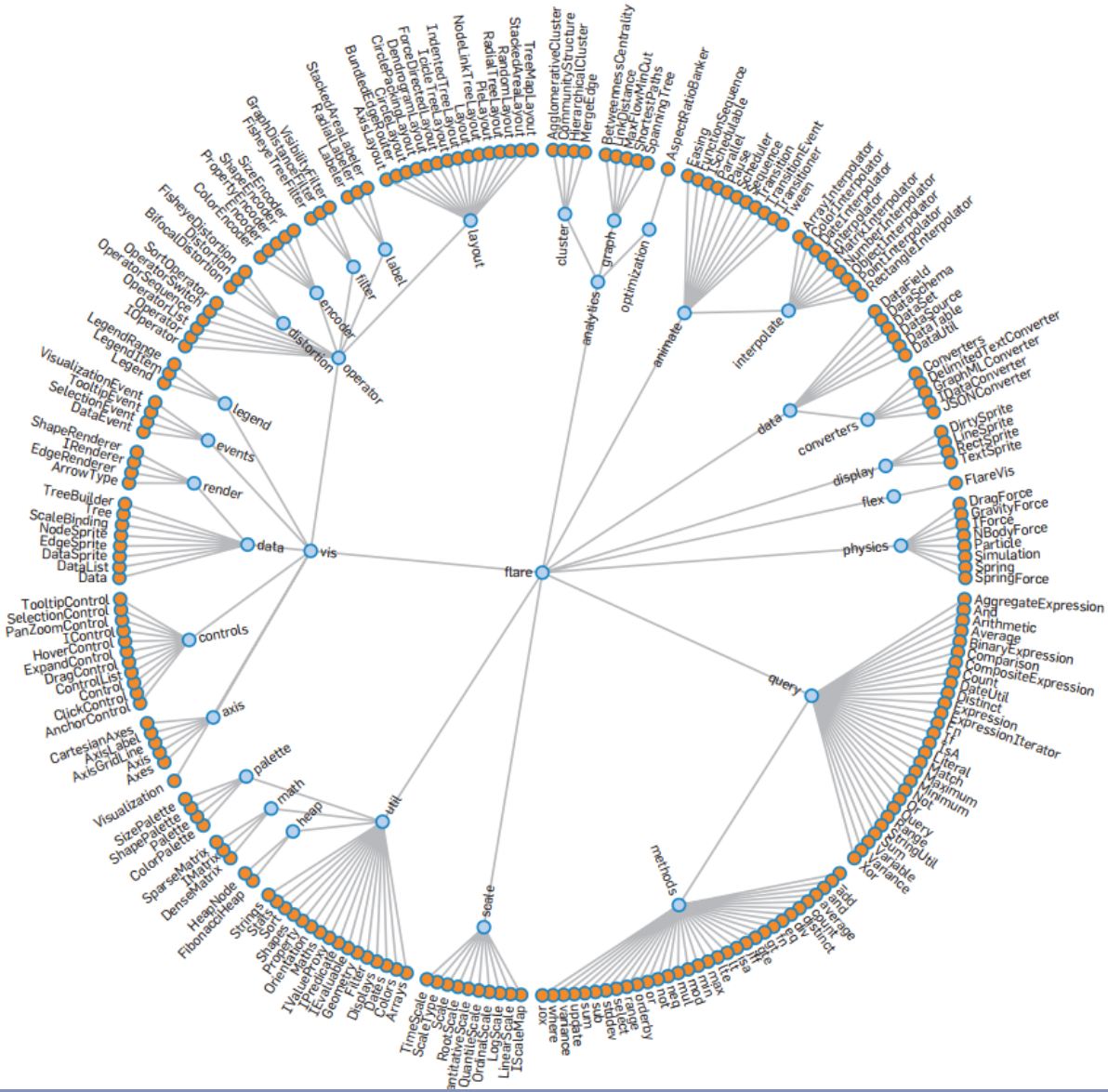
- Time-series data
- Statistical data
- Geographical data
- **Hierarchical data**
- Network data

Node-link diagram

Hierarchies: **Figure 4a. Radial node-link diagram of the Flare package hierarchy.**



Circular dendogram

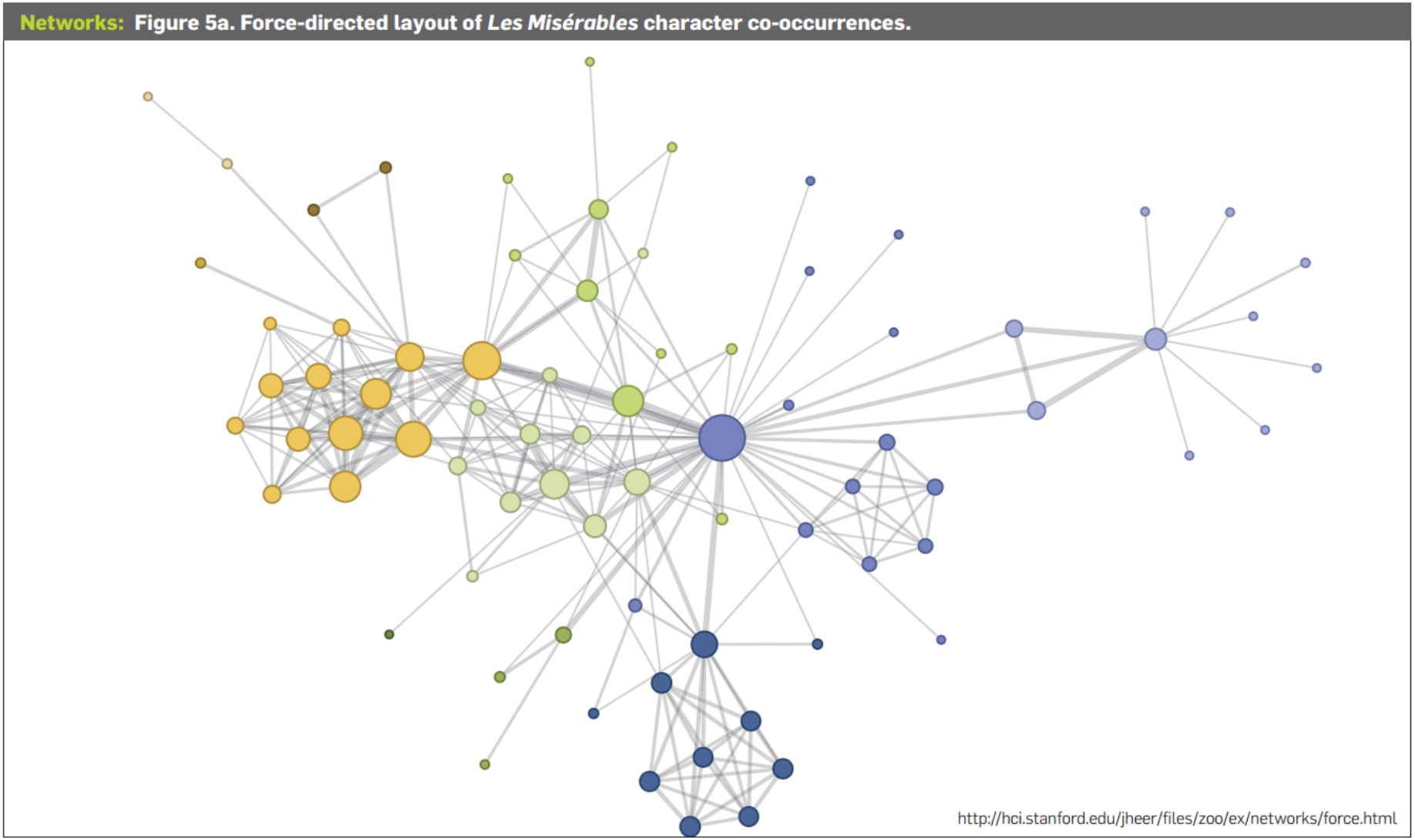


Examples

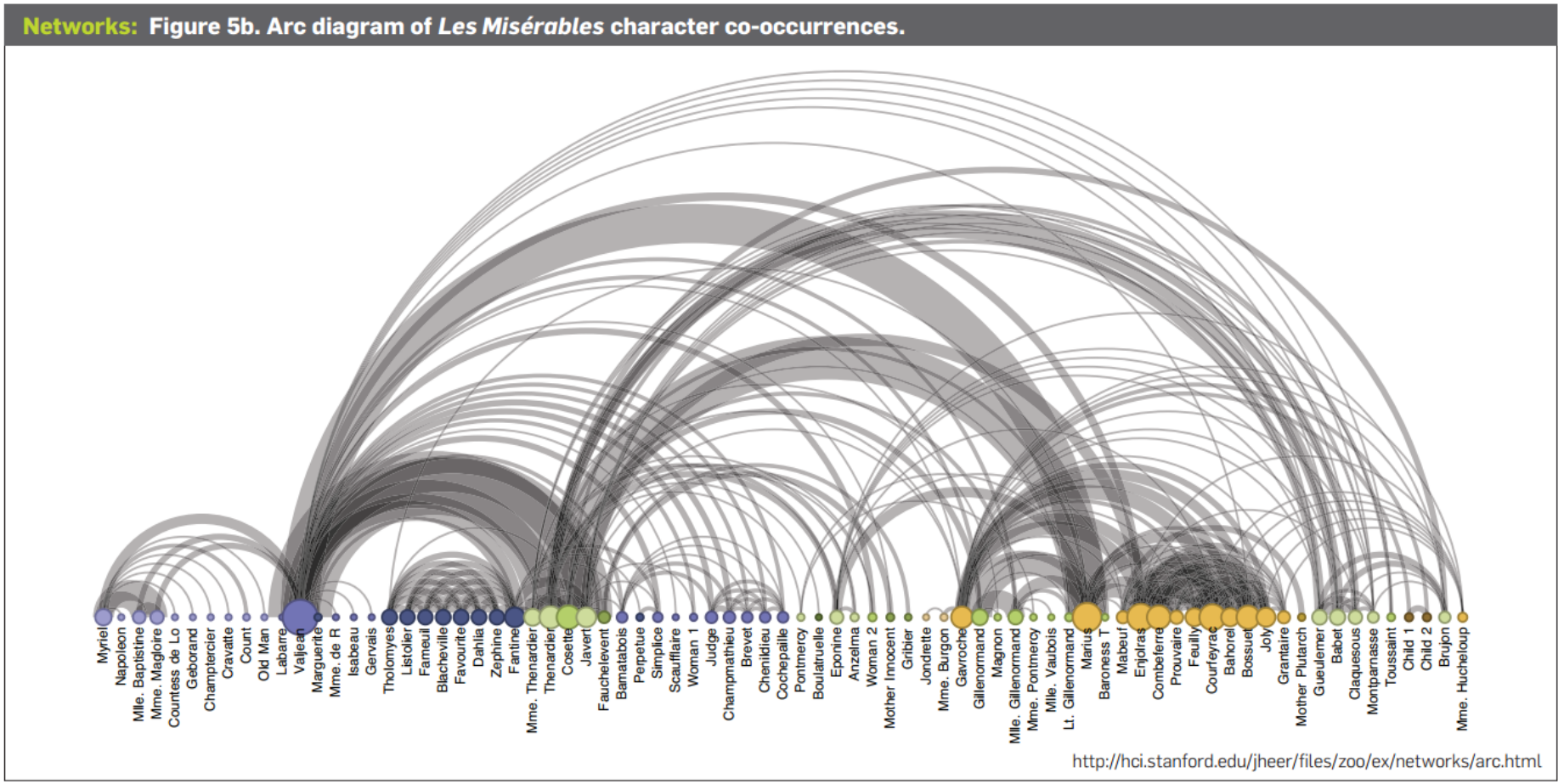
- Time-series data
- Statistical data
- Geographical data
- Hierarchical data
- **Network data**

Force-directed layout

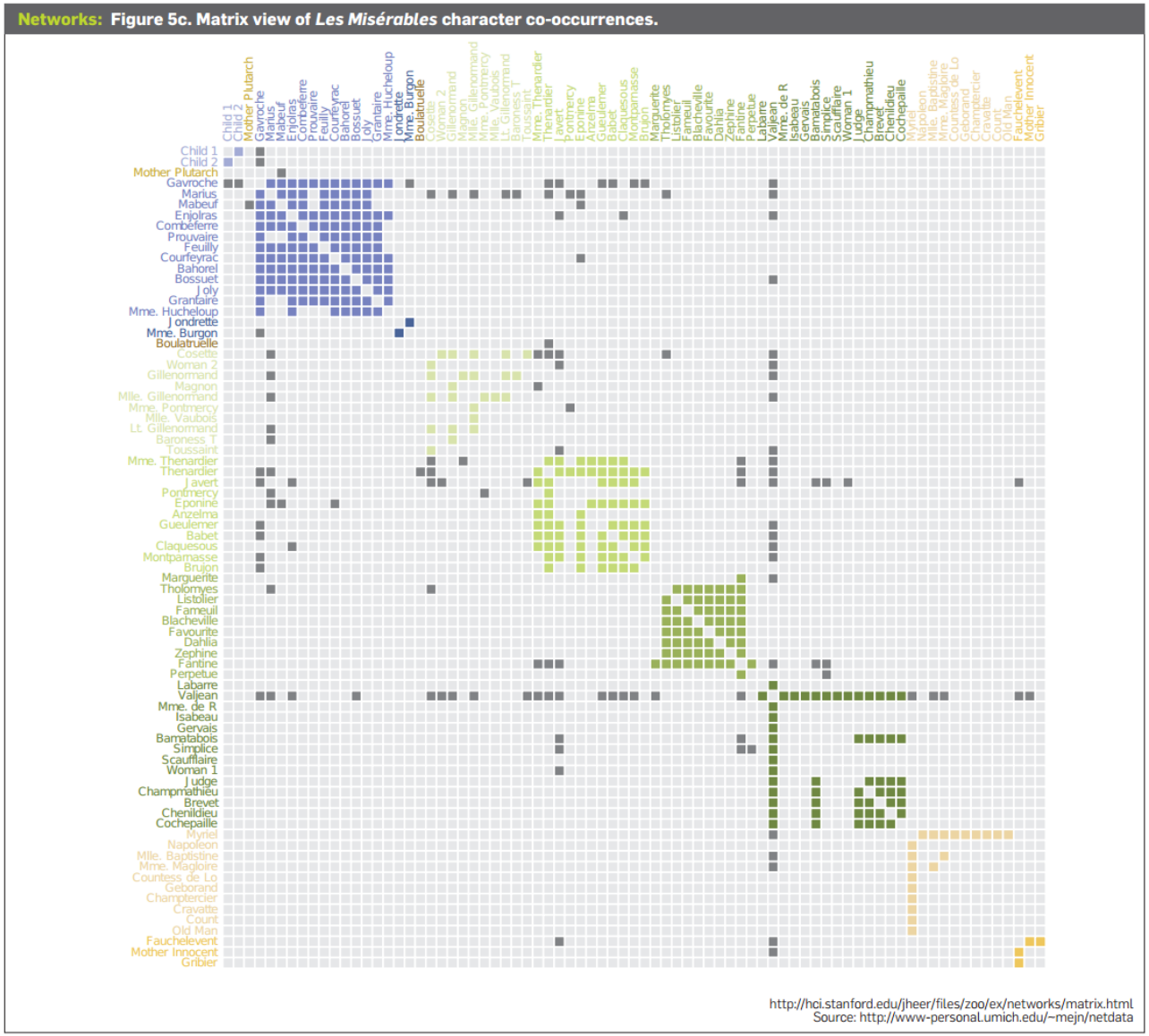
Networks: **Figure 5a. Force-directed layout of *Les Misérables* character co-occurrences.**



Arc diagram


Networks: Figure 5b. Arc diagram of Les Misérables character co-occurrences.
Myriel
Napoleon
Mlle. Baptistine
Mme. Magloire
Countess de Lo
Geborand
Champtercier
Cravatte
Count
Old Man
Labarre
Valjean
Marguerite
Mme. de R
Isabeau
Gervais
Tholomyes
Listolier
Fameuil
Blacheville
Favourite
Dahlia
Zephine
Fantine
Mme. Thenardier
Thenardier
Cosette
Javert
Fauchelevent
Bamatabois
Perpetue
Simplice
Scaufflaire
Woman 1
Judge
Champmathieu
Brevet
Chenildieu
Cochepaille
Pontmercy
Boulatruelle
Eponine
Anzelma
Woman 2
Mother Innocent
Gribier
Jondrette
Mme. Burgon
Gavroche
Gillenormand
Magnon
Mlle. Gillenormand
Mme. Pontmercy
Mlle. Vaubois
Lt. Gillenormand
Marius
Baroness T
Mabeuf
Enjolras
Combeferre
Prouvaire
Feuilly
Courfeyrac
Bahorel
Bossuet
Joly
Grantaire
Mother Plutarch
Gueulemer
Babet
Claquesous
Montparnasse
Toussaint
Child 1
Child 2
Brujon
Mme. Hucheloup
http://hci.stanford.edu/jheer/files/zoo/ex/networks/arc.html

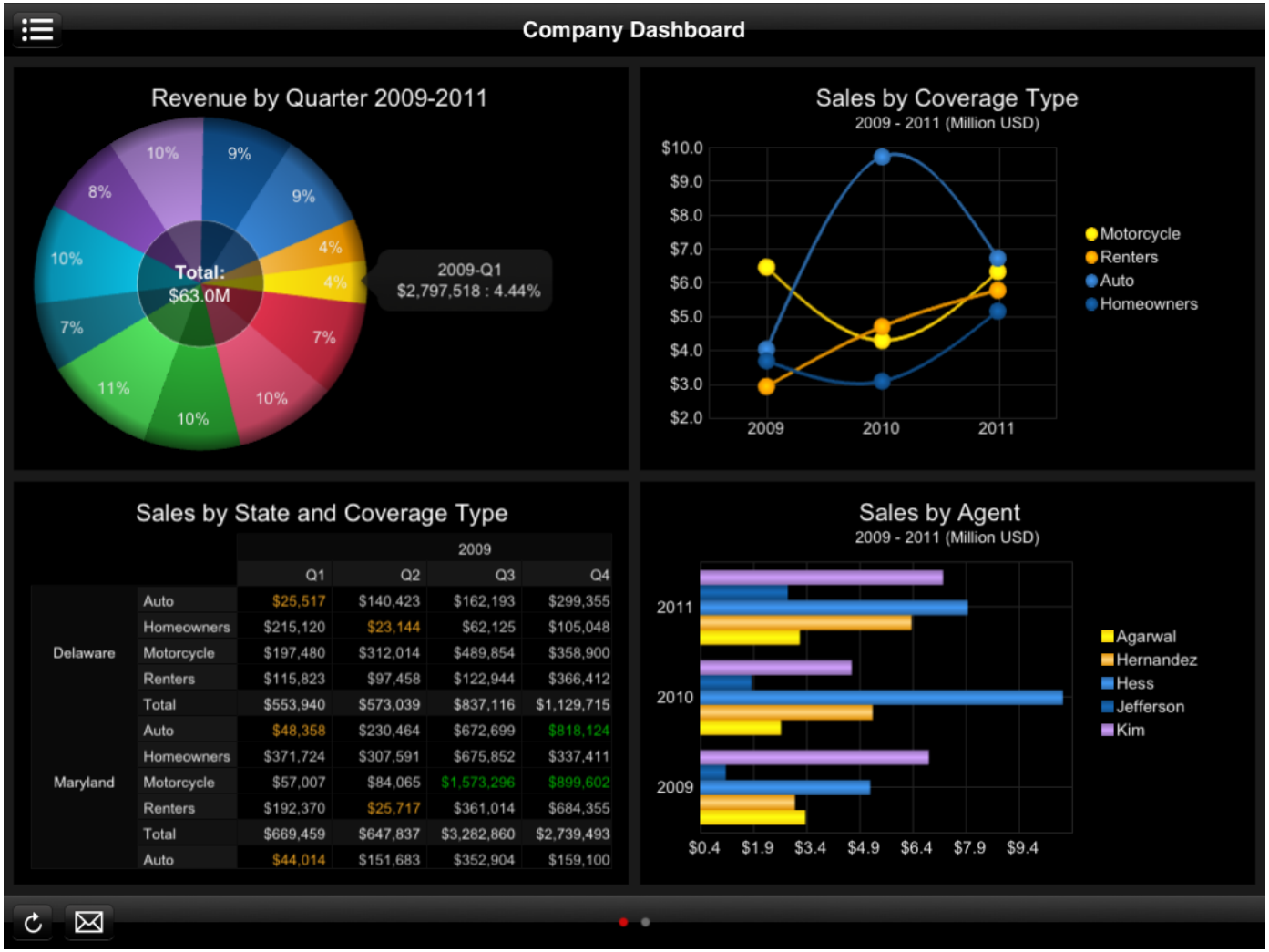
Annotated matrix



Dashboards

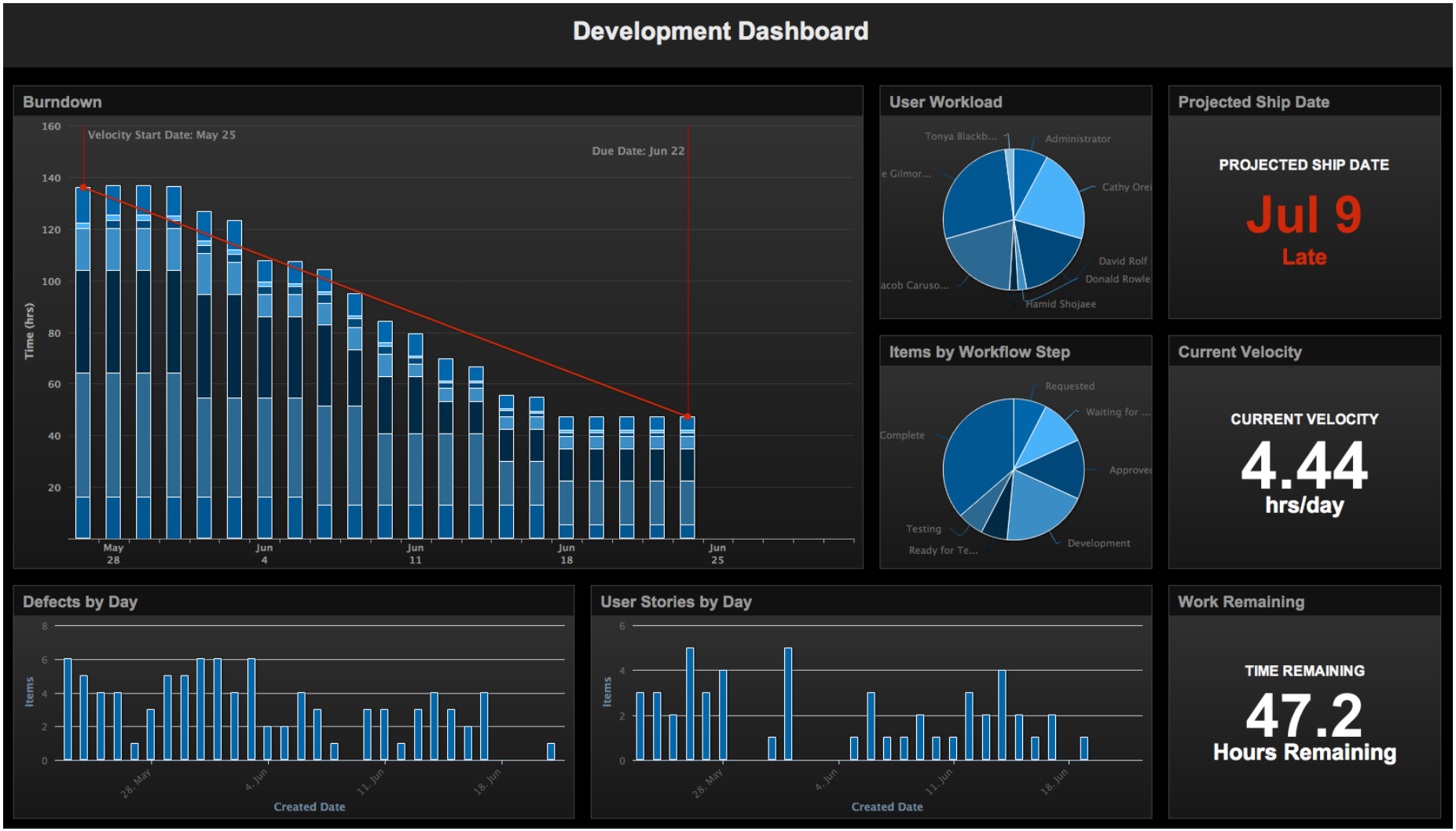
- Multiple **widgets** on one page
- A widget can contain:
  - OLAP slice
  - KPI metric
  - Data mining results
  - ...
- Codeless reporting
- BI in the blink of an eye!

Dashboard (1)



http://blog.jinfonet.com/

Dashboard (2)

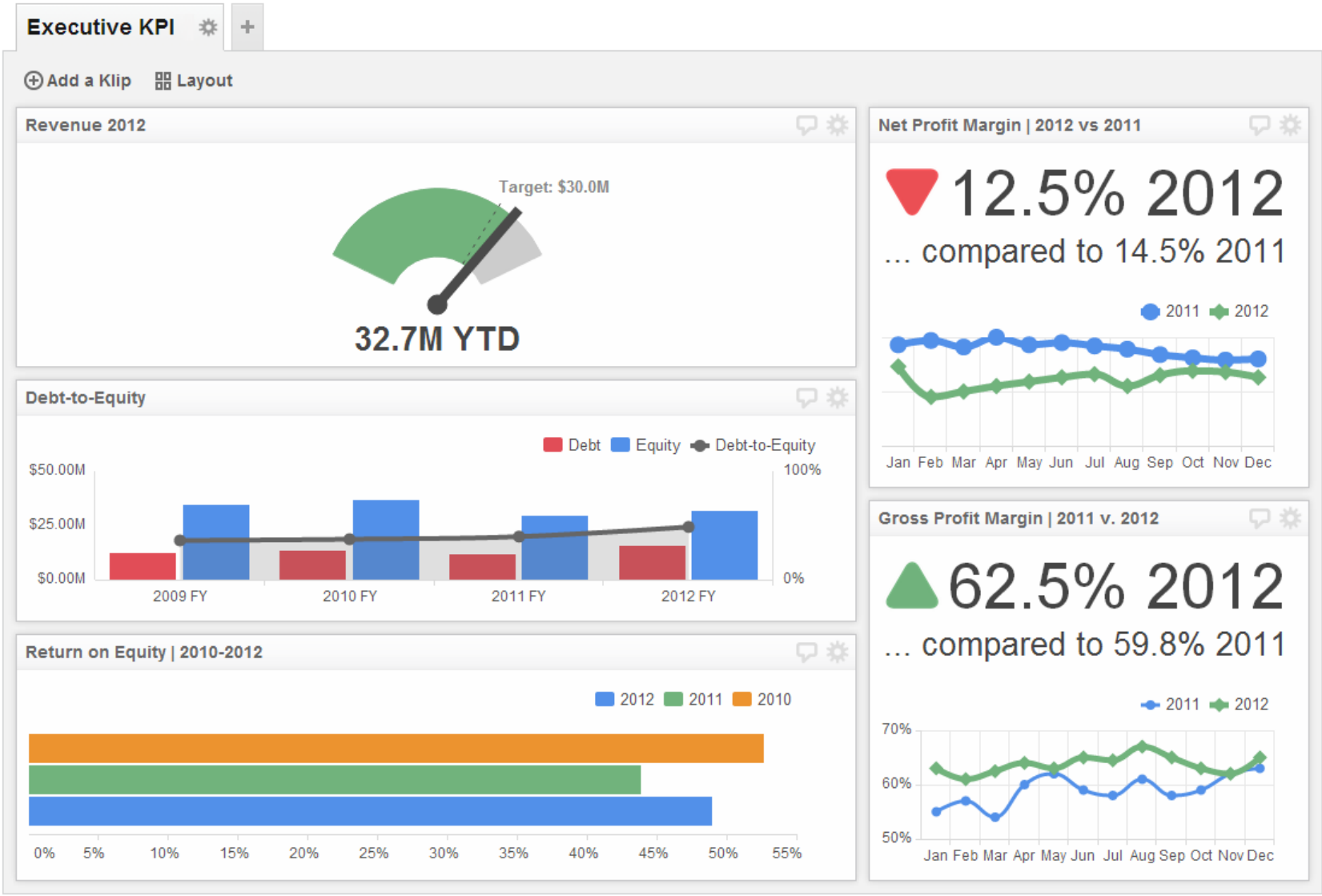


http://www.axosoft.com/

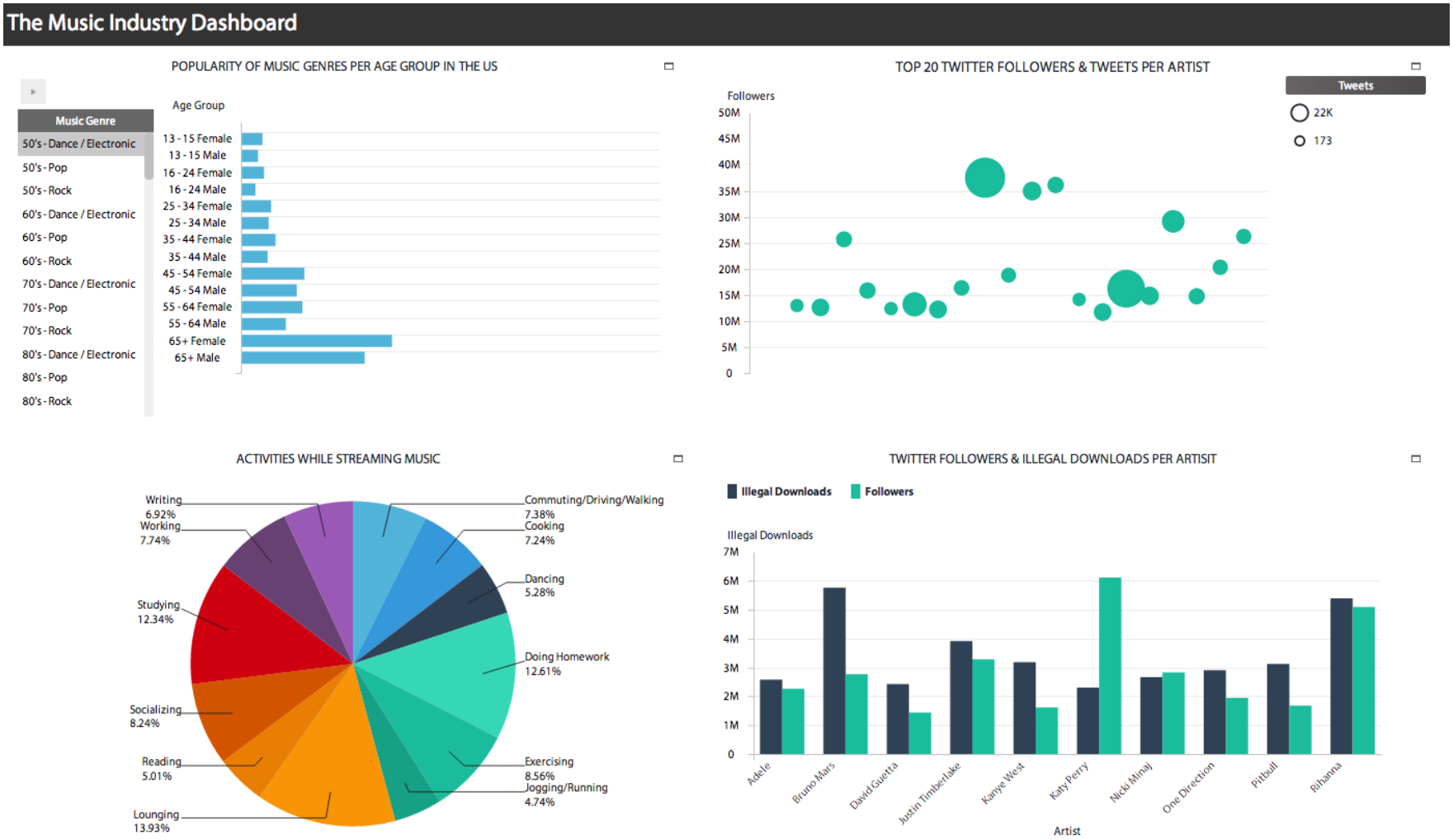
Dashboard (3)

http://www.cyfe.com/

Dashboard (4)


Executive KPI
Add a Klip
Layout
Revenue 2012
Target: $30.0M
32.7M YTD
Net Profit Margin | 2012 vs 2011
12.5% 2012
... compared to 14.5% 2011
2011
2012
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Debt-to-Equity
Debt
Equity
Debt-to-Equity
$50.00M
$25.00M
$0.00M
100%
0%
2009 FY
2010 FY
2011 FY
2012 FY
Gross Profit Margin | 2011 v. 2012
62.5% 2012
... compared to 59.8% 2011
2011
2012
70%
60%
50%
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Return on Equity | 2010-2012
2012
2011
2010
0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55%

Dashboard (5)



http://insideanalysis.com/

Assignment 1

- Gaming industry context
- Sales log spanning 4 years of sales
- Apply and compare BI techniques
- Inspect, visualize, aggregate, segment, score ...
- Deliverables:
  1. Web-based BI Dashboard
  2. Short assignment report in LaTeX

Assignment 1 — Hints

- Model: MySQL database containing the data
- View: HTML page using Javascript that reads JSON
- Controller: PHP outputs relevant data in JSON

Lab session February 17

- Make progress with Assignment 1
- Read paper by Kooti et al.
- Setup a framework for your dashboard
- Load some data into your framework
- Investigate visualization options

Credits

Lecture based on (slides of the (previous edition of the)) course book:
W. van der Aalst, *Process Mining: Data Science in Action*, 2nd edition, Springer, 2016.