VISUAL ANALYTICS

Prof. Dr. Bart Baesens

Department of Decision Sciences and Information Management,

KU Leuven (Belgium)

School of Management, University of Southampton (United Kingdom)

Bart.Baesens@kuleuven.be

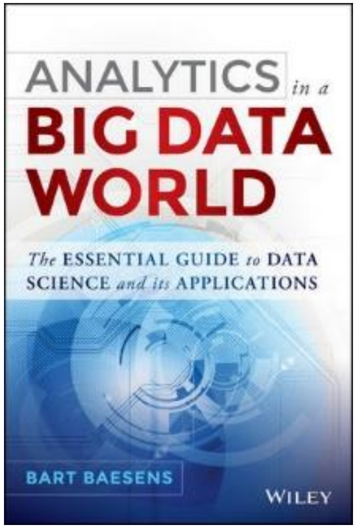
Twitter/Facebook/YouTube: DataMiningApps

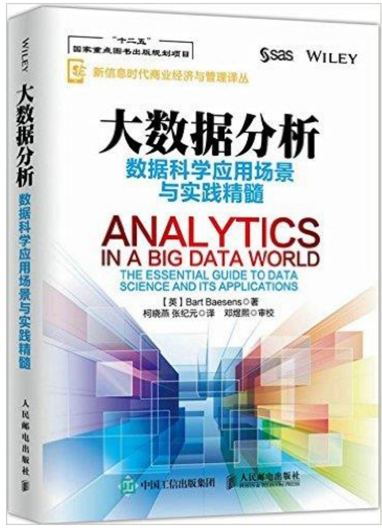
www.dataminingapps.com

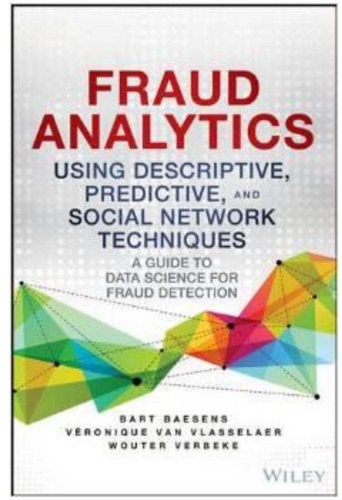
Presenter: Bart Baesens

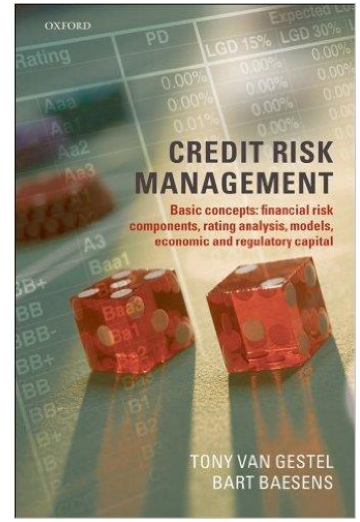
- Studied at KU Leuven (Belgium)
  - Business Engineer in Management Informatics, 1998
  - PhD. in Applied Economic Sciences, 2003
- PhD. : Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
- Professor at KU Leuven, Belgium
- Lecturer at the University of Southampton, UK
- Research: Big Data & Analytics, Credit Risk, Fraud, Marketing, …
- YouTube/Facebook/Twitter: DataMiningApps
- www.dataminingapps.com
- Bart.Baesens@kuleuven.be

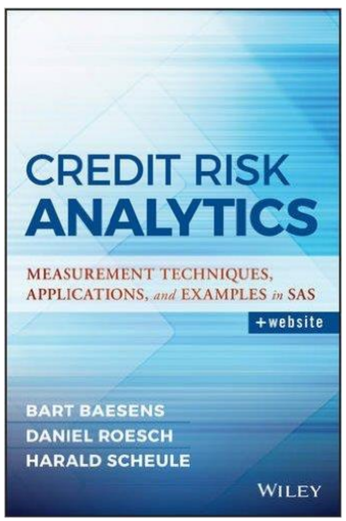
Example Publications

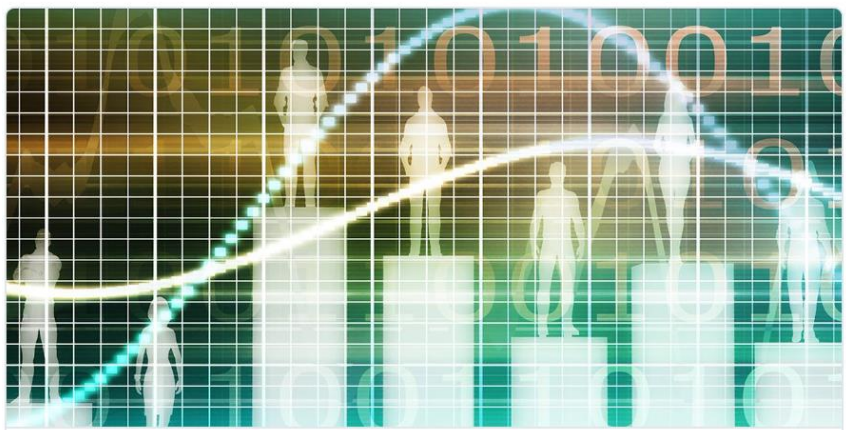








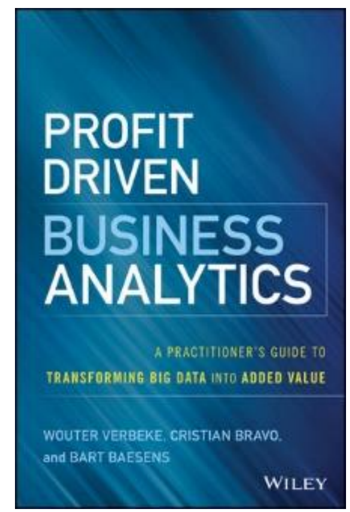




Is Your Company Ready for HR Analytics?

HR analytics is the next big change in human resources management.

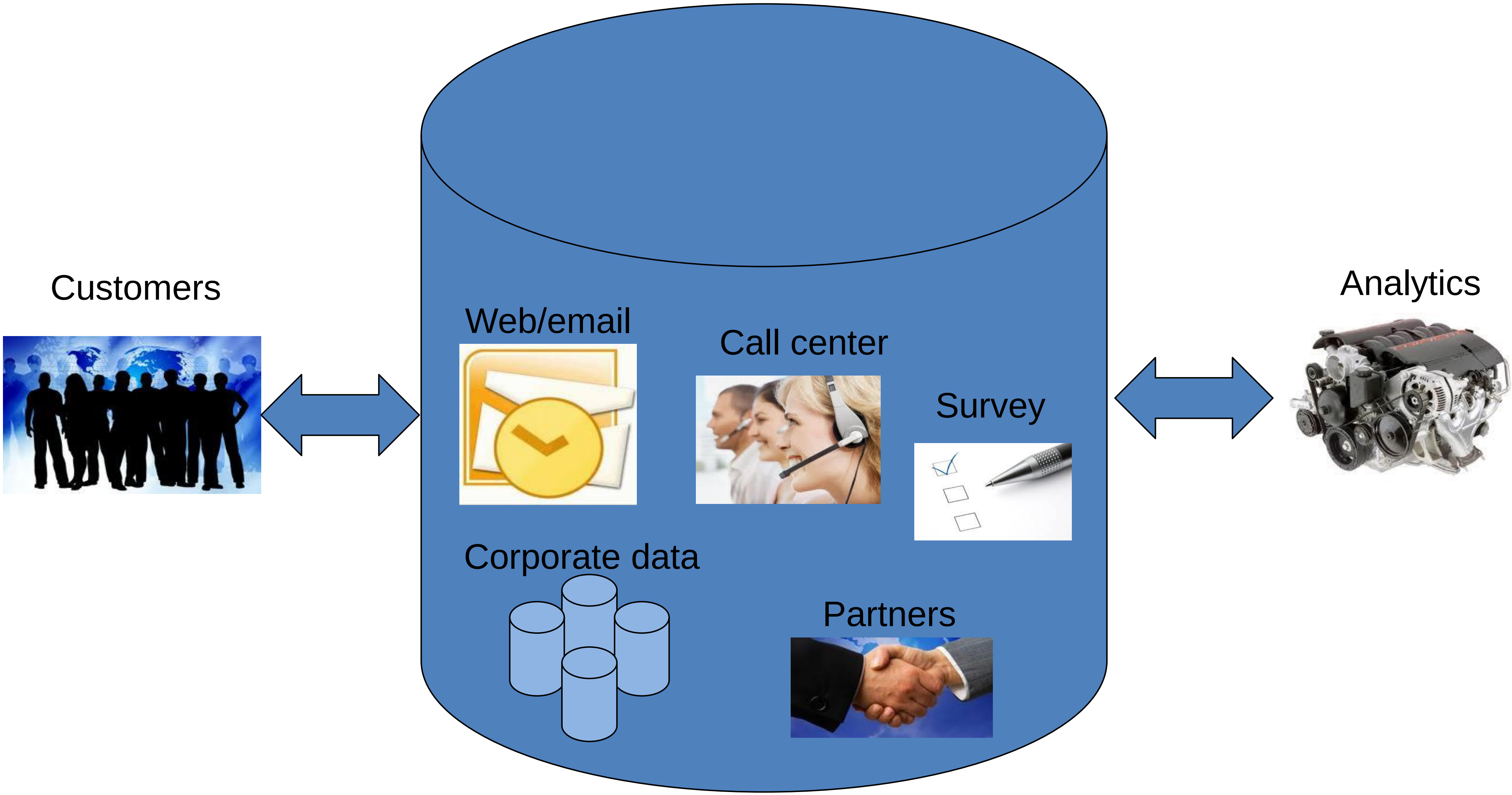
activevoice.us



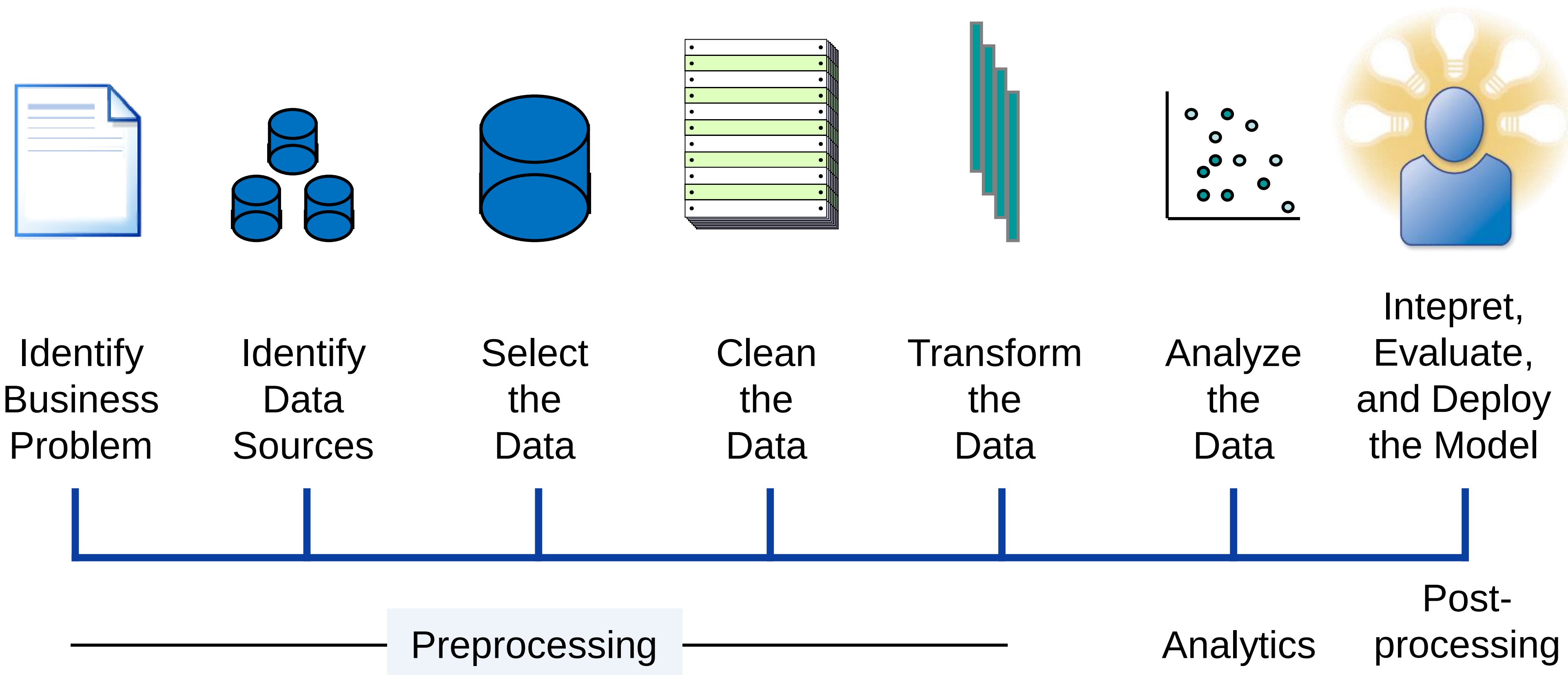
Overview

- Big Data & Analytics: setting the stage
- Power and premise of Visual Analytics
- Visual Analytics and the Analytics process model
  - Data preprocessing
  - Model representation
  - Model usage
  - Model backtesting
- Software
- Guidelines
- Conclusions

Living in a Data Flooded World!

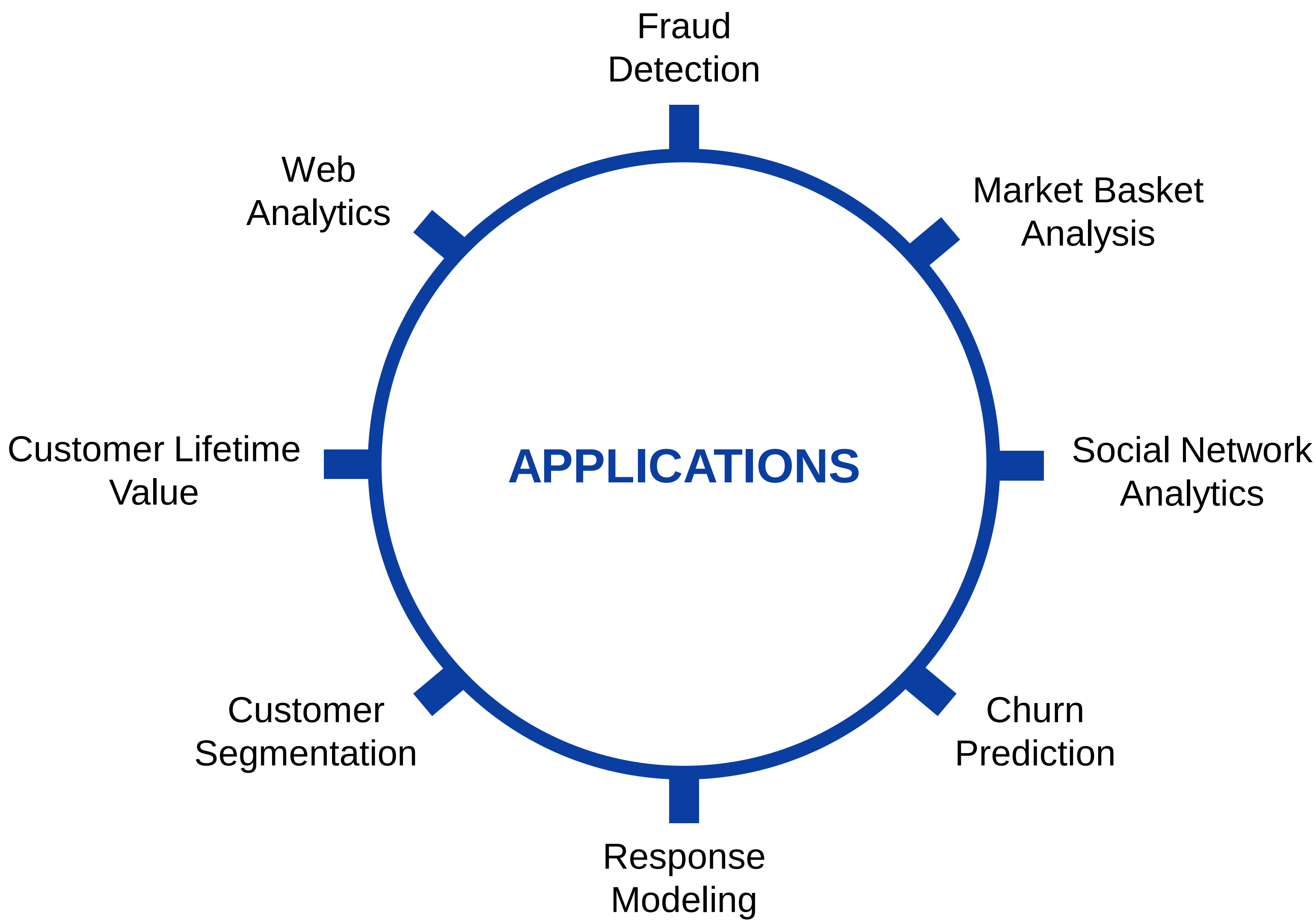


The Analytics Process Model



Baesens, 2015.

Feel the vibe!



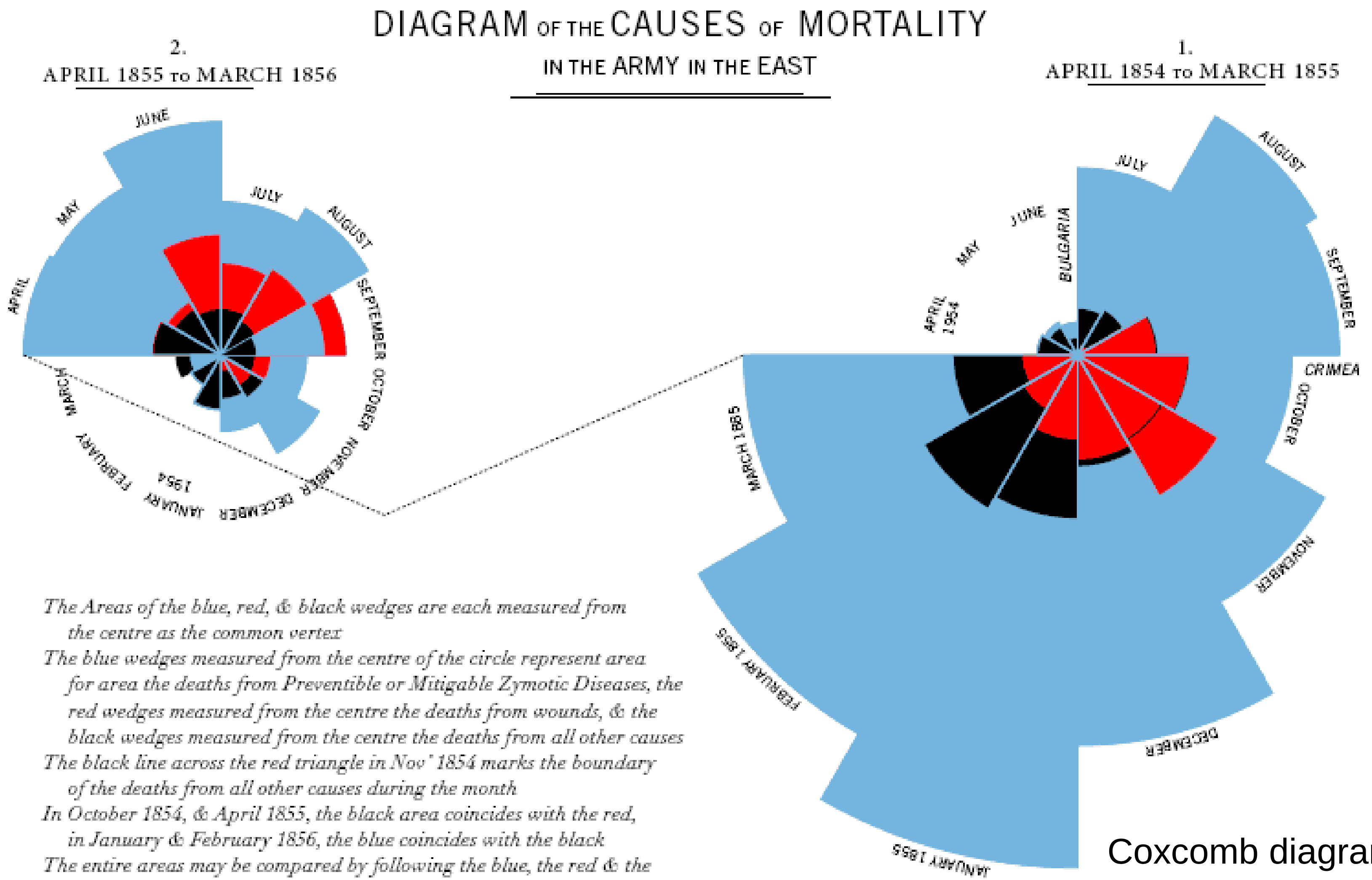
Two Analytical Disconnects

- **Data versus Data Scientist**
  - Data: unstructured, distributed, noisy, time-evolving
  - Data Scientist: patterns in data, statistical significance, predictive power, structure the unstructured!

- **Data Scientist versus Business Expert**
  - Data Scientist: decision trees, logistic regression, random forests, area under ROC curve, top decile lift, R-squared, etc.
  - Business Expert: customers, marketing campaigns, risk mitigation, portfolios, profit, return on Investment (ROI), etc.

Visual Analytics as a mediator!

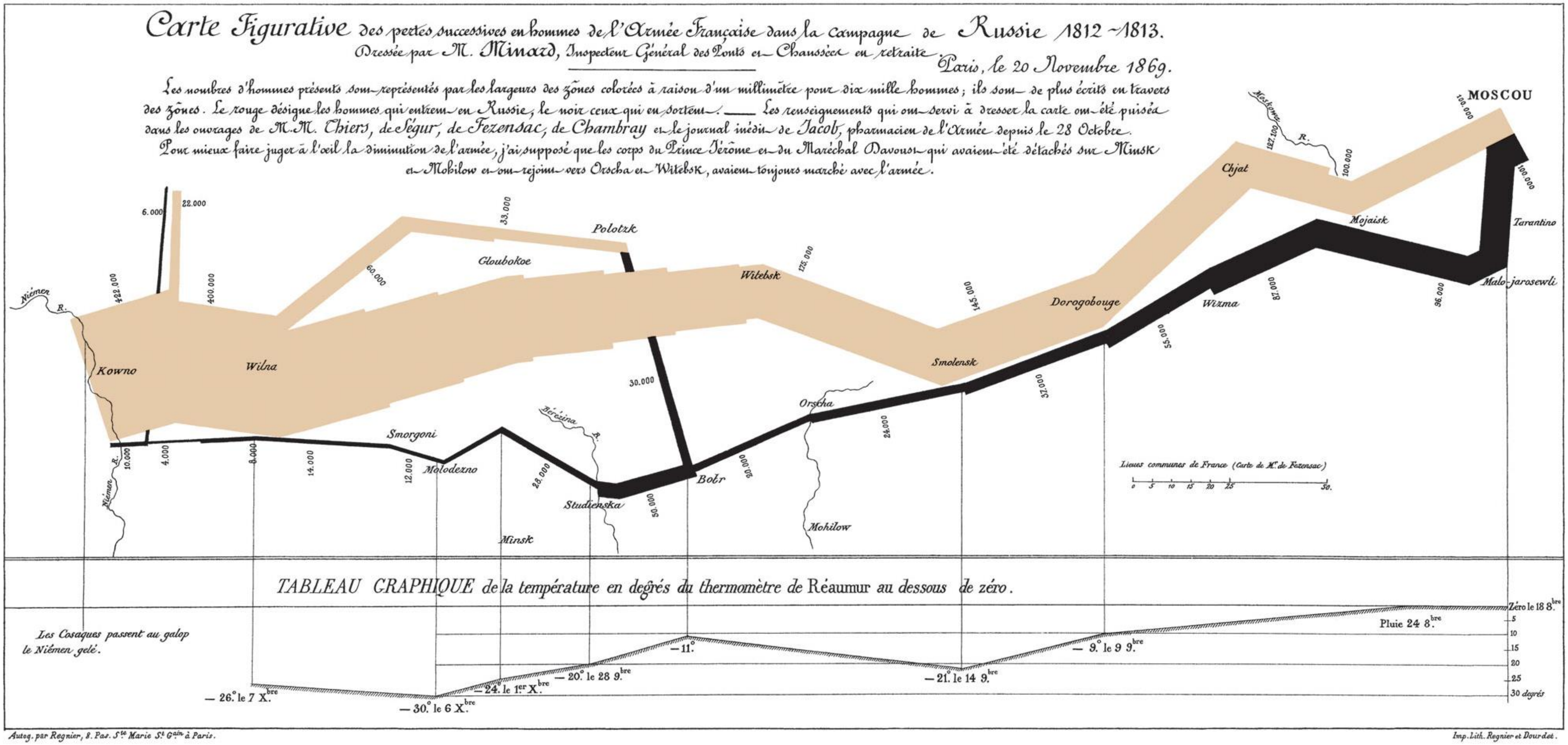
The Power of Visual Analytics



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes
The black line across the red triangle in Nov^r 1854 marks the boundary of the deaths from all other causes during the month
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

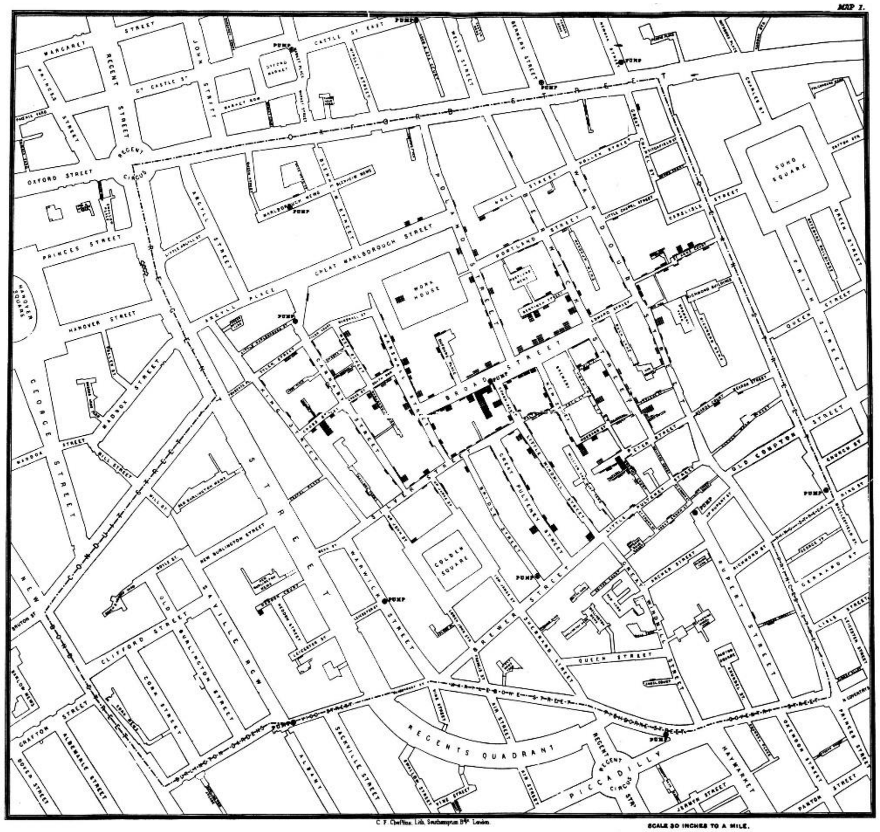
Coxcomb diagram
Florence Nightingale,1858

The Power of Visual Analytics

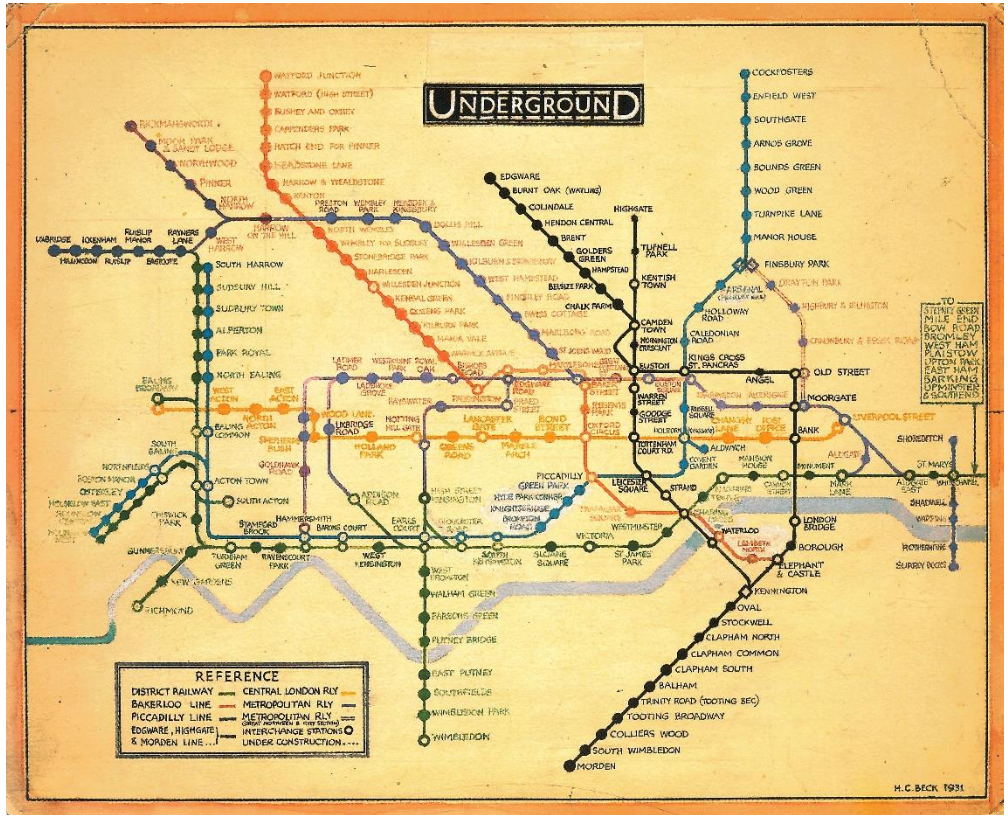


Charles Minnard, 1869

The Power of Visual Analytics

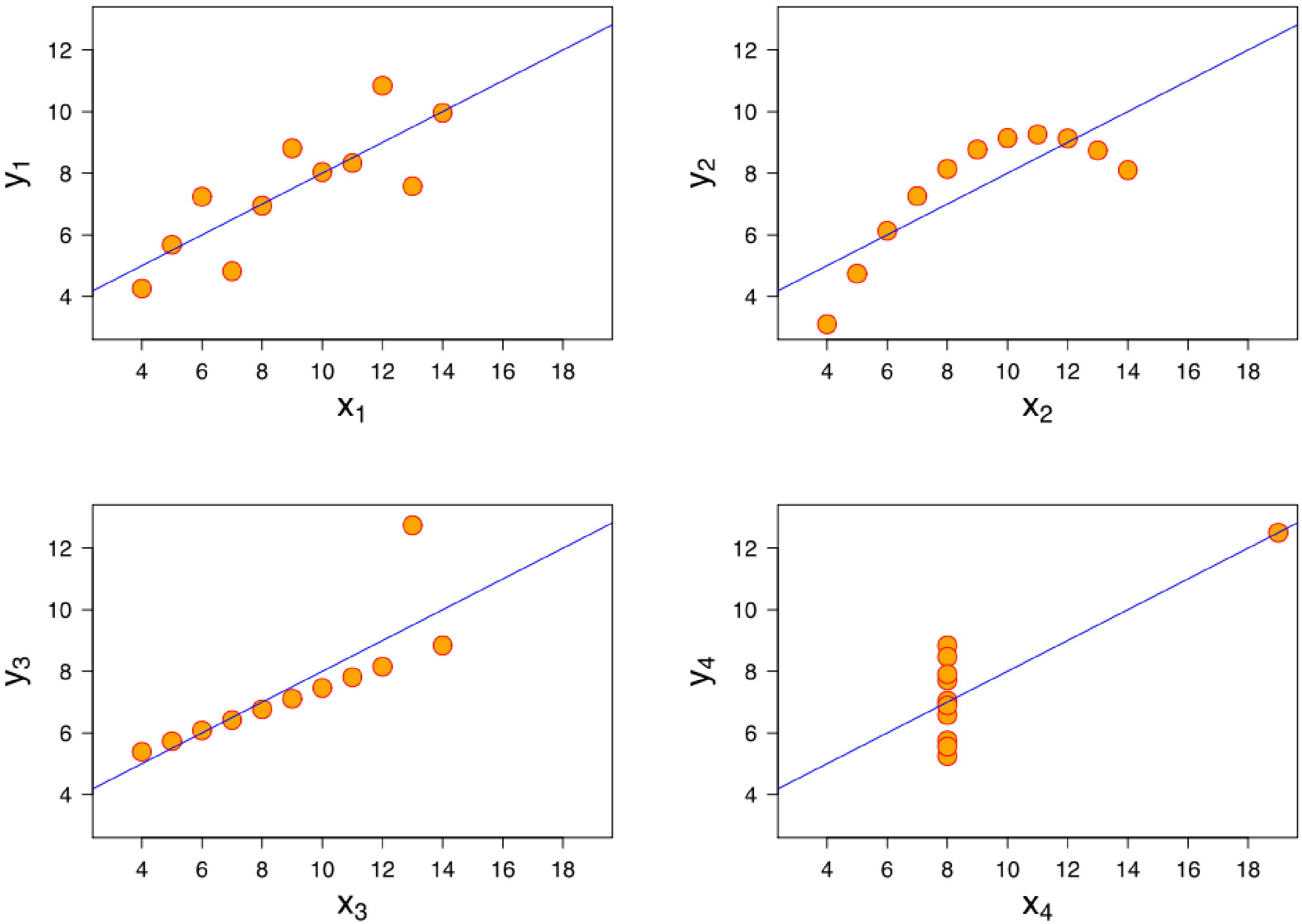


London cholera map
John Snow, 1854



London Tube map
Harry Beck, 1931

Visuals versus Statistics: Anscombe's Quartet



Mean(x)	9
Var(x)	11
Mean(y)	7.50
Var(y)	4.125
Corr(x,y)	0.816

$y = 3.00 + 0.500x$

Anscombe, 1973

Visual Analytics: The Premise

- Reduce cognitive overload by having users interact with data and/or analytical models using visual tools
- “*the science of analytical reasoning facilitated by interactive visual interfaces*“ (Thomas and Cook, 2005)
- Help data scientists + business users to explore and better understad data + models
- "**A picture is worth a thousand words**"

Visual Analytics versus the Analytics Process Model

- **Data preprocessing**
  - Use Visual Analytics to find outliers, missing values, frequent/suspicious/interesting patterns, etc.
  - Visualisation unit: Data!

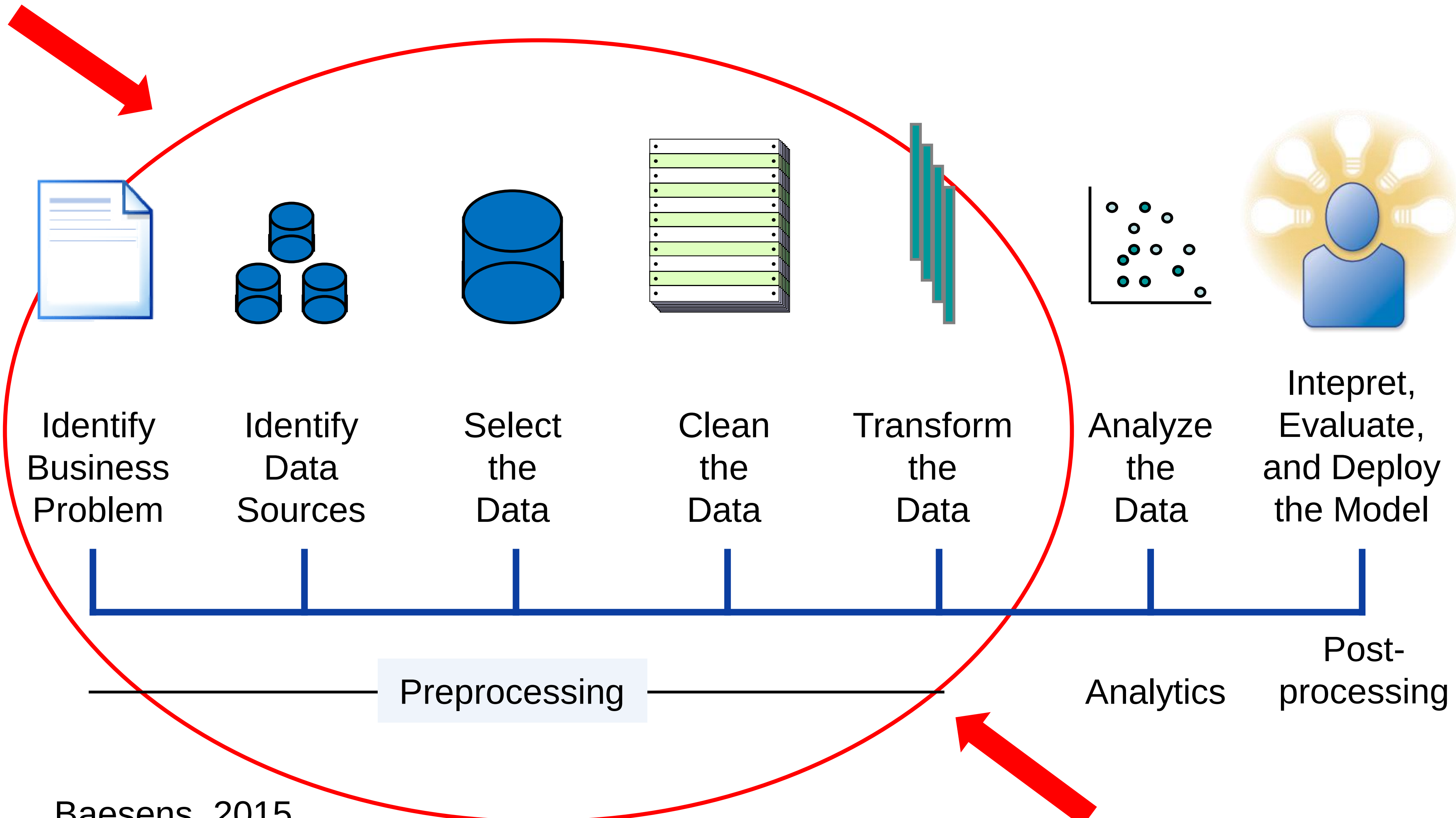
- **Model representation**
  - Use Visual Analytics to represent models in a user-friendly way
  - Visualisation unit: Model formula!

Visual Analytics versus the Analytics Process Model

- **<u>Model usage</u>**
  - Use Visual Analytics to integrate models with other applications (e.g. GIS)
  - Visualisation unit: <u>Model interaction</u>!

- **<u>Model backtesting</u>**
  - Use Visual Analytics to monitor model performance
  - Visualisation unit: <u>Model performance</u>!

The Analytics Process Model



Baesens, 2015.

Data Preprocessing: Statistical plots

Histogram

Radar plot

Bubble plot

Mosaic plot

Pivot chart

Contour plot

Scatter plot

Bar chart

Pie chart

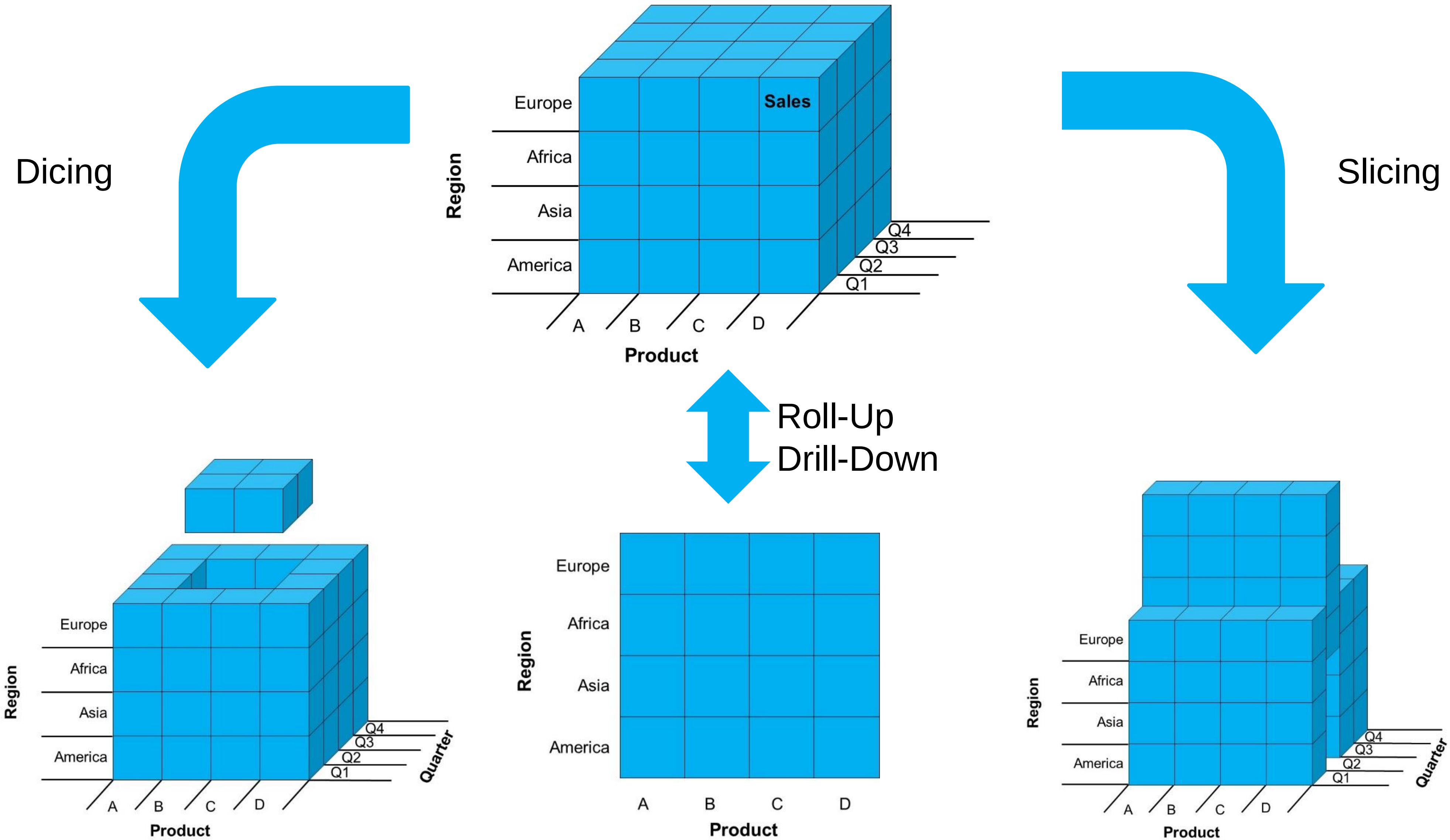
Scatter plot

Violin plot

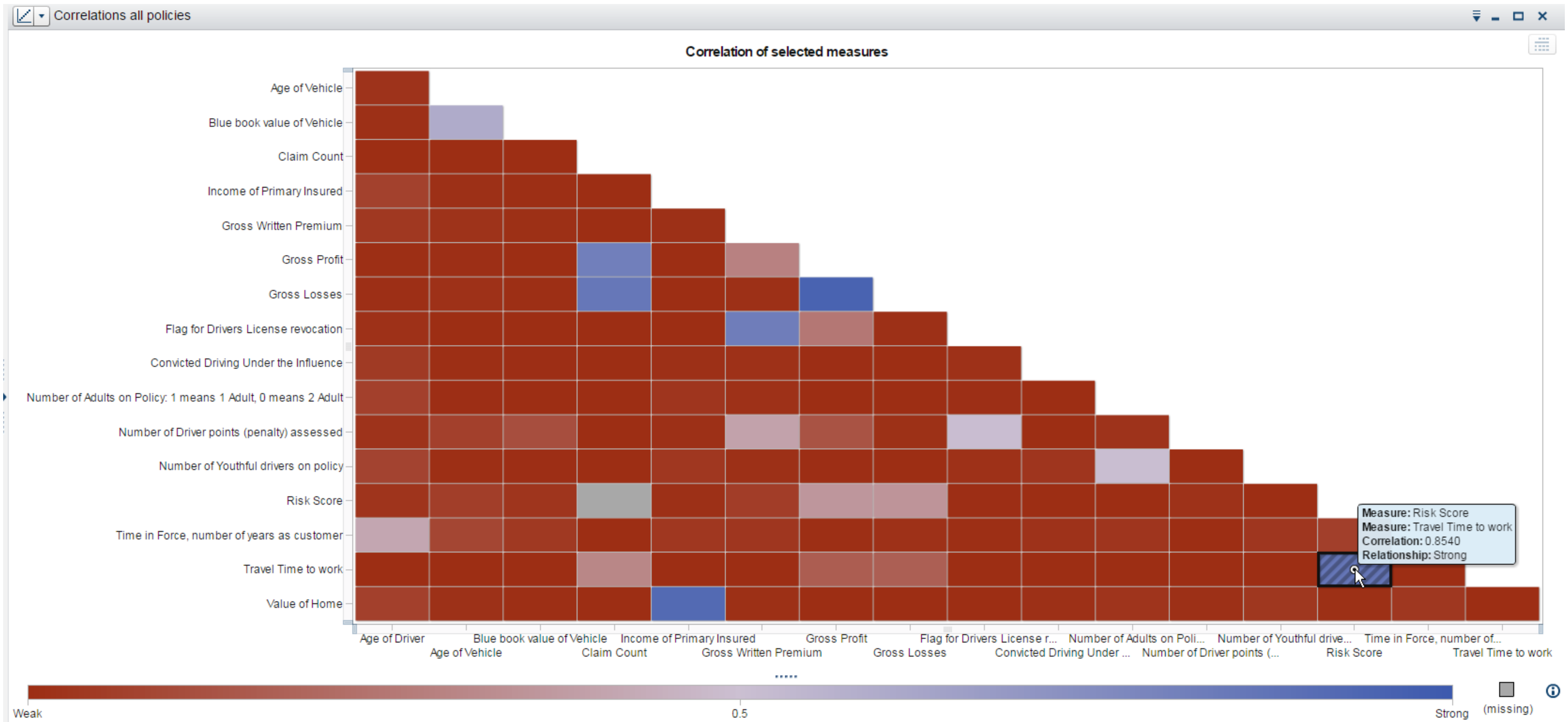
Box plot

Aimed at Exploratory Data Analysis!

Data Preprocessing: OLAP



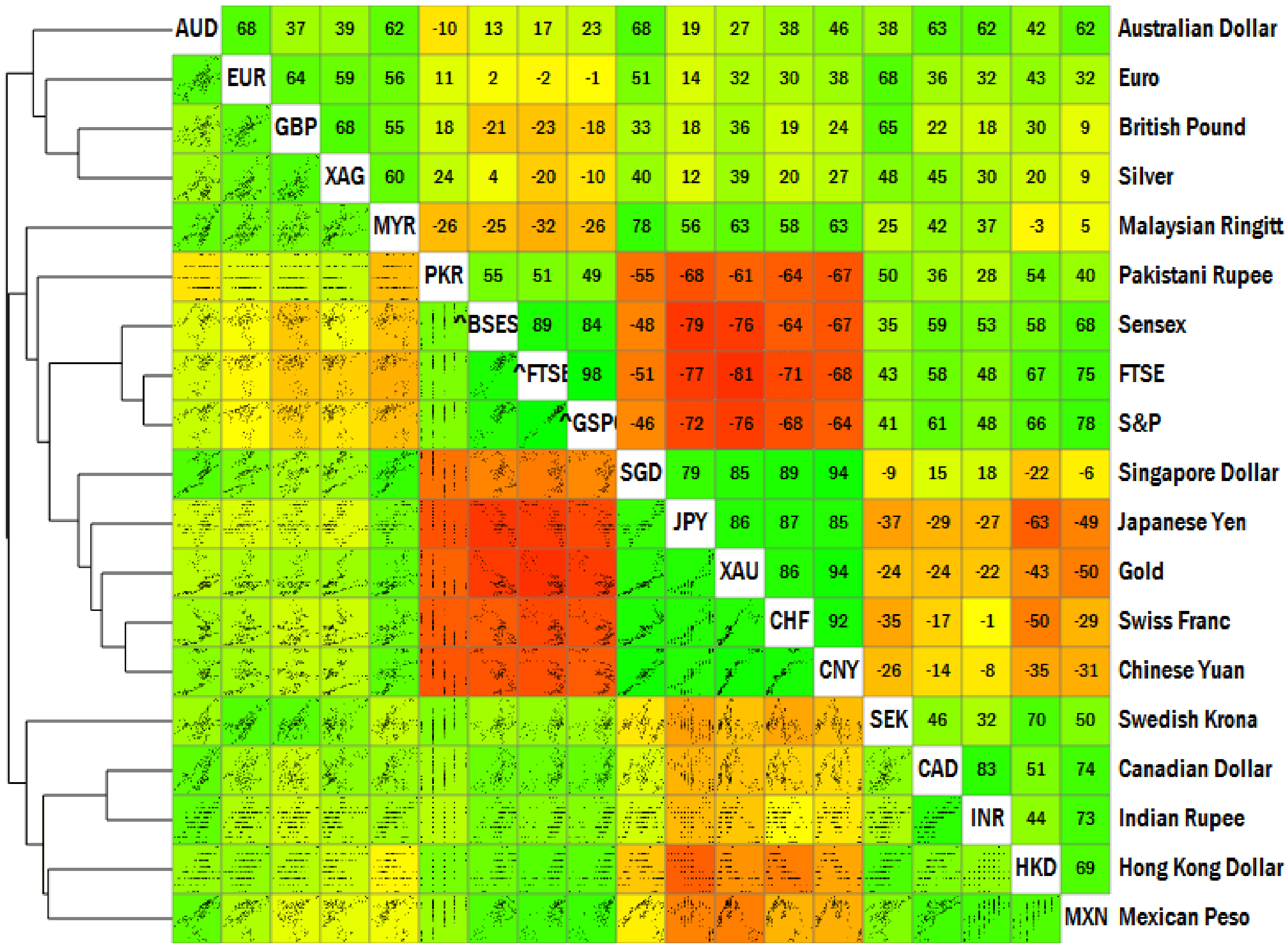
Data Preprocessing: Correlation matrix



First steps towards predictive modeling!

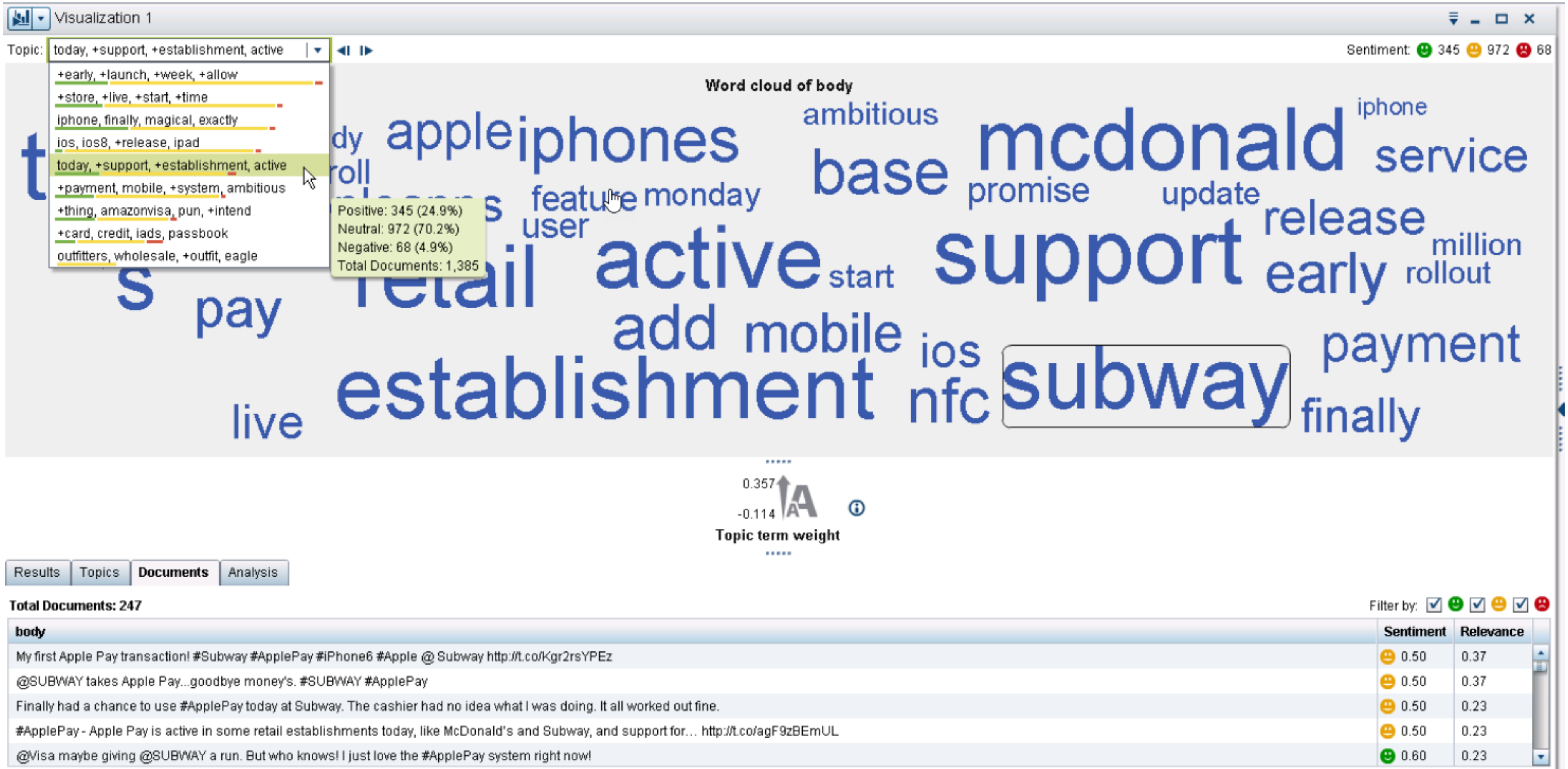
www.sas.com

Data Preprocessing: cluster plot



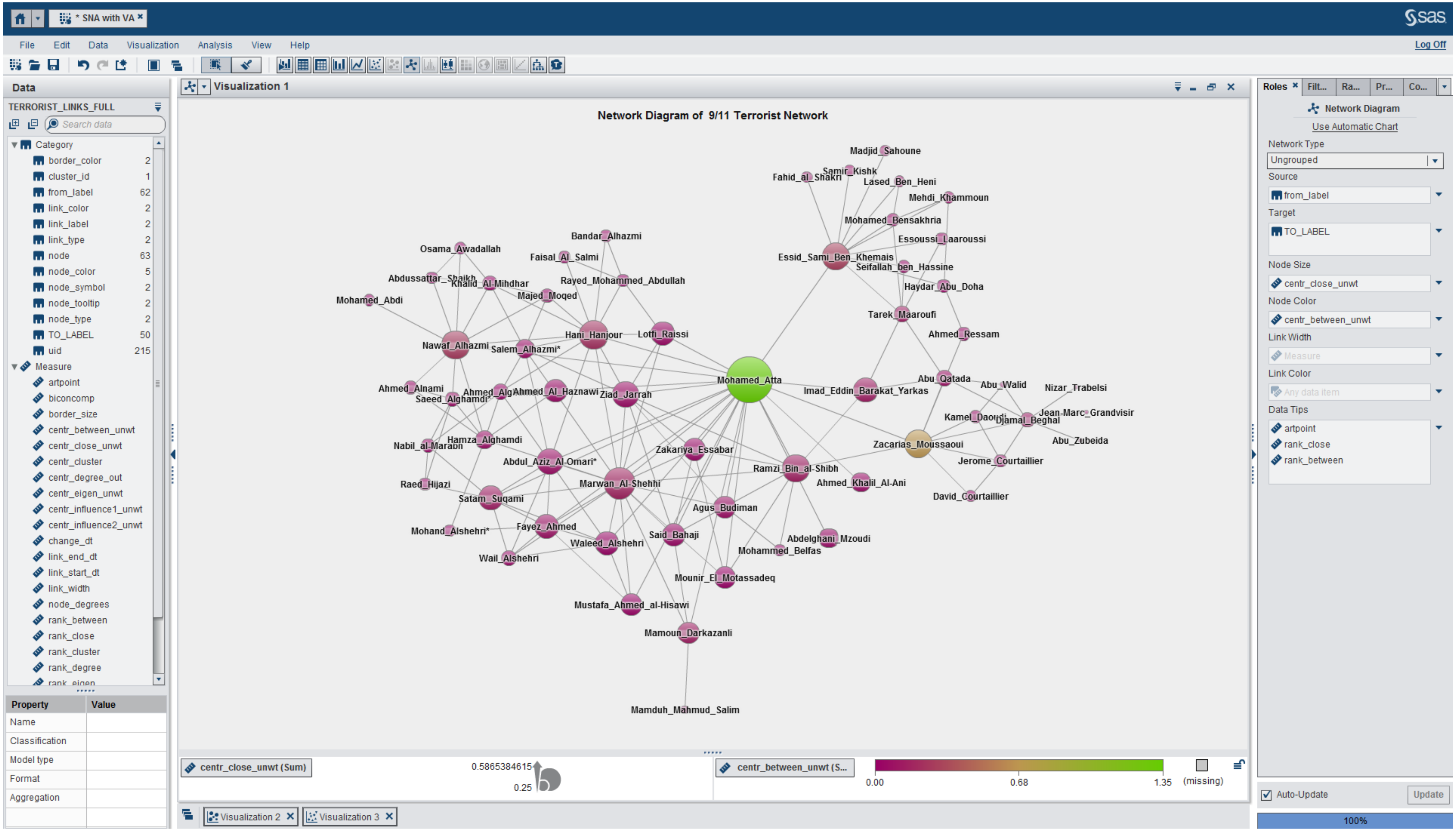
http://blog.gramener.com/18/visualising-securities-correlation

Data Preprocessing: Unstructured Data



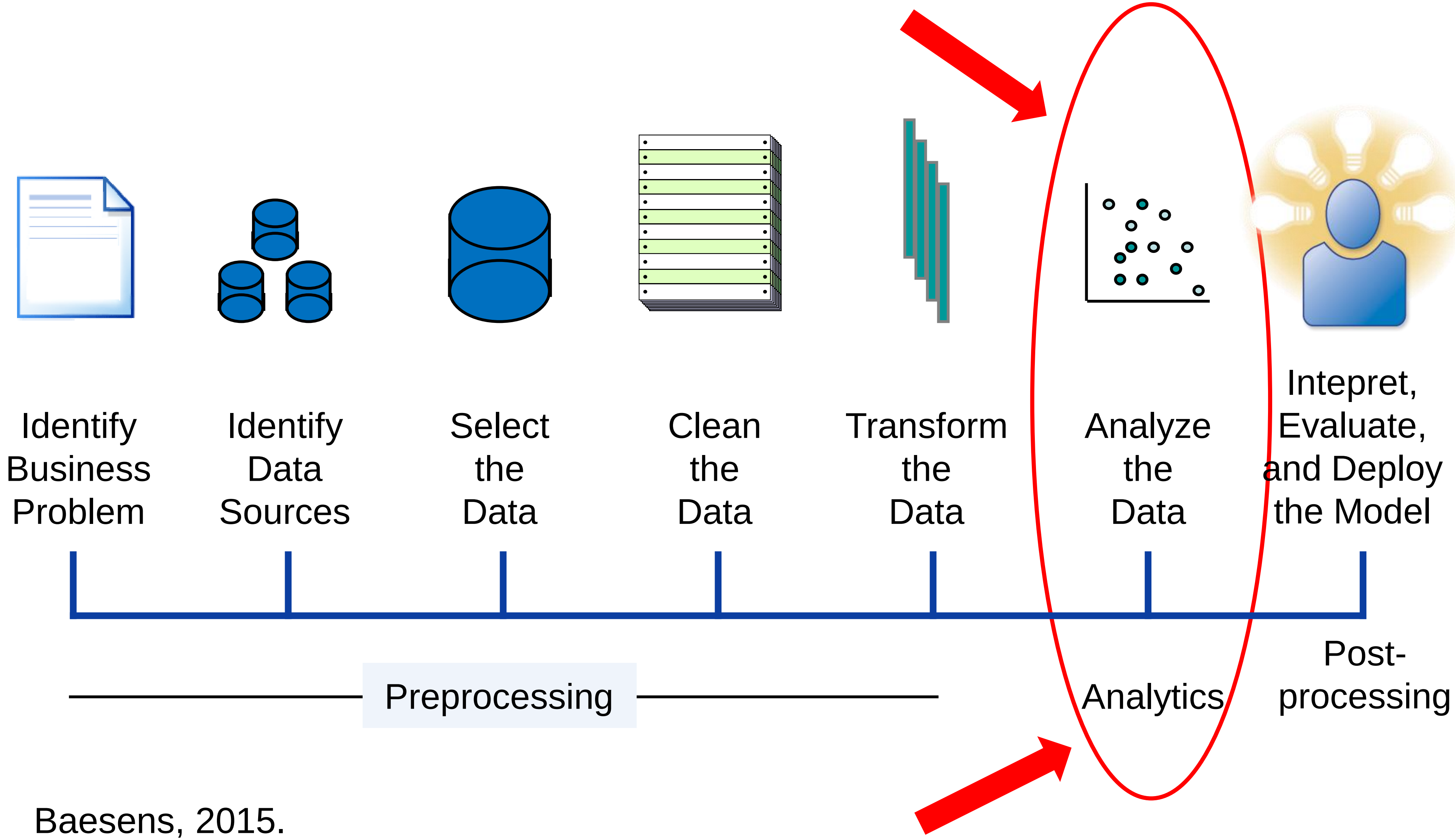
http://blogs.sas.com/content/sascom/2014/11/05/what-a-sentiment-word-cloud-revealed-about-apple-pay/

Data Preprocessing: Unstructured Data



http://journals.uic.edu/ojs/index.php/fm/article/view/941/863

The Analytics Process Model



Baesens, 2015.

Model Representation

- Bridge the gap between the analytical model and the business user
- Minimize information loss between analytical model and visual representation
- Business user engagement to foster **<u>trust</u>**
- Note: model interpretability depends upon business application
  - Credit risk versus medical diagnosis
  - Fraud detection versus fraud prevention

Model Representation: Decision Tables

RULE1: IF **Avg Usage** $<$ 25 AND **Internet Plan** = Y AND **Service Calls** $>$ 3
THEN **Churn**

RULE2: IF **Avg Usage** $<$ 25 AND **Internet Plan** = N THEN **Churn**

RULE3: IF **Avg Usage** $\geq$ 25 AND **Internet Plan** = Y THEN **Not Churn**

RULE4: IF **Avg Usage** $<$ 25 AND **Service Calls** $\leq$ 3 THEN **Not Churn**

Rule Conflicts?

Rule Coverage?

Baesens, Van Vlasselaer, Verbeke, 2015.

Model Representation: Decision Tables

1. Avg Usage	< 25				≥ 25			
2. Internet Plan	Y		N		Y		N	
3. Service Calls	≤ 3	> 3	≤ 3	> 3	≤ 3	> 3	≤ 3	3
1. Churn	-	x	x	x	-	-	-	-
2. Not Churn	x	-	x	-	x	x	-	-
Contributing Rule(s):	R4	R1	R2 R4	R2	R3	R3		

Conflict!

No coverage!

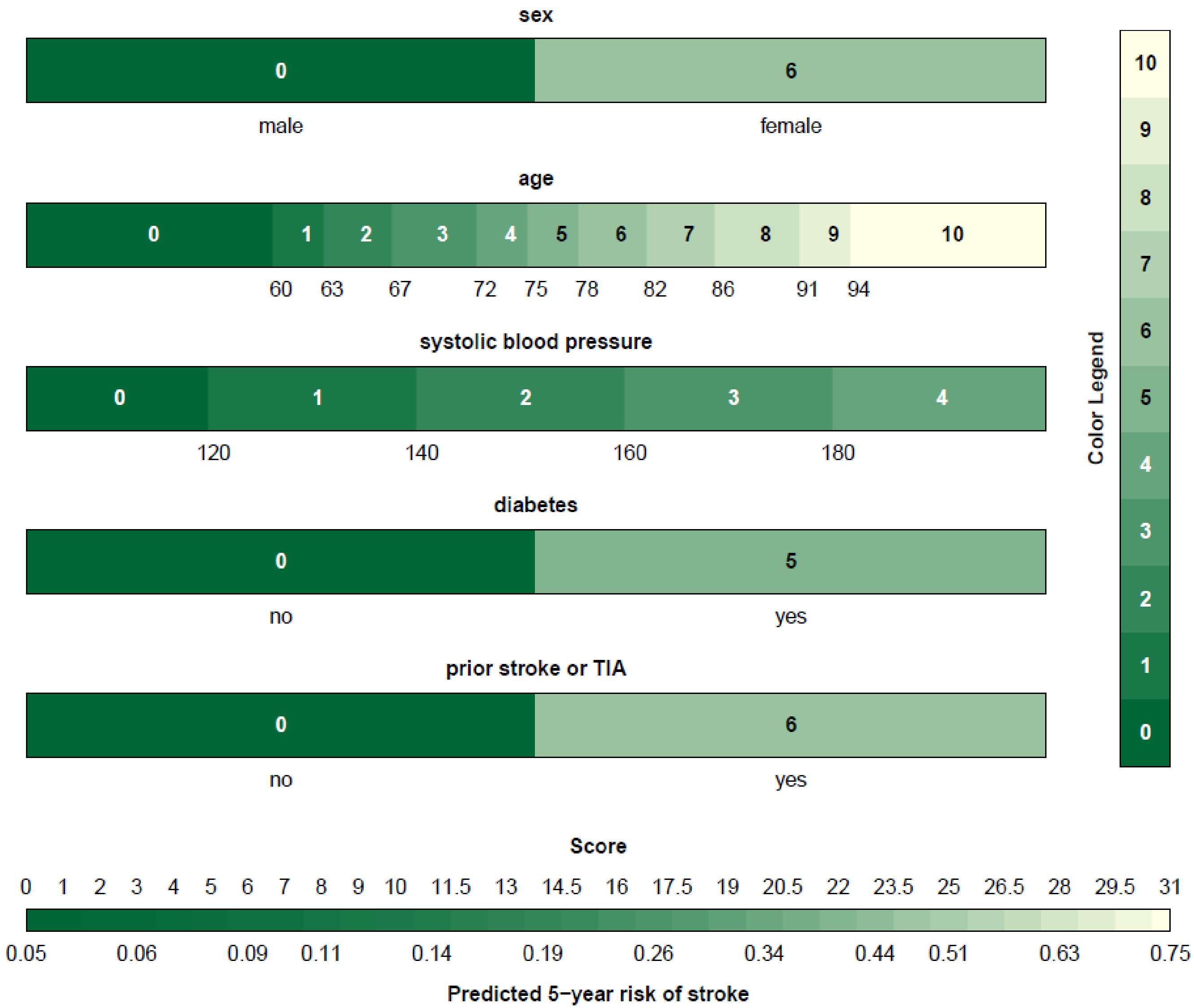
Model Representation: Scorecards

$$P(\text{Good} \mid \text{Age}, \text{Gender}, \text{Salary}, \ldots) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Salary \ldots)}}$$

Characteristic Name	Attribute	Scorecard Points
AGE 1	Up to 26	100
AGE 2	26 - 35	120
AGE 3	35 - 37	185
AGE 4	37+	225
GENDER 1	Male	90
GENDER 2	Female	180
SALARY 1	Up to 500	120
SALARY 2	501-1000	140
SALARY 3	1001-1500	160
SALARY 4	1501-2000	200
SALARY 5	2000+	240

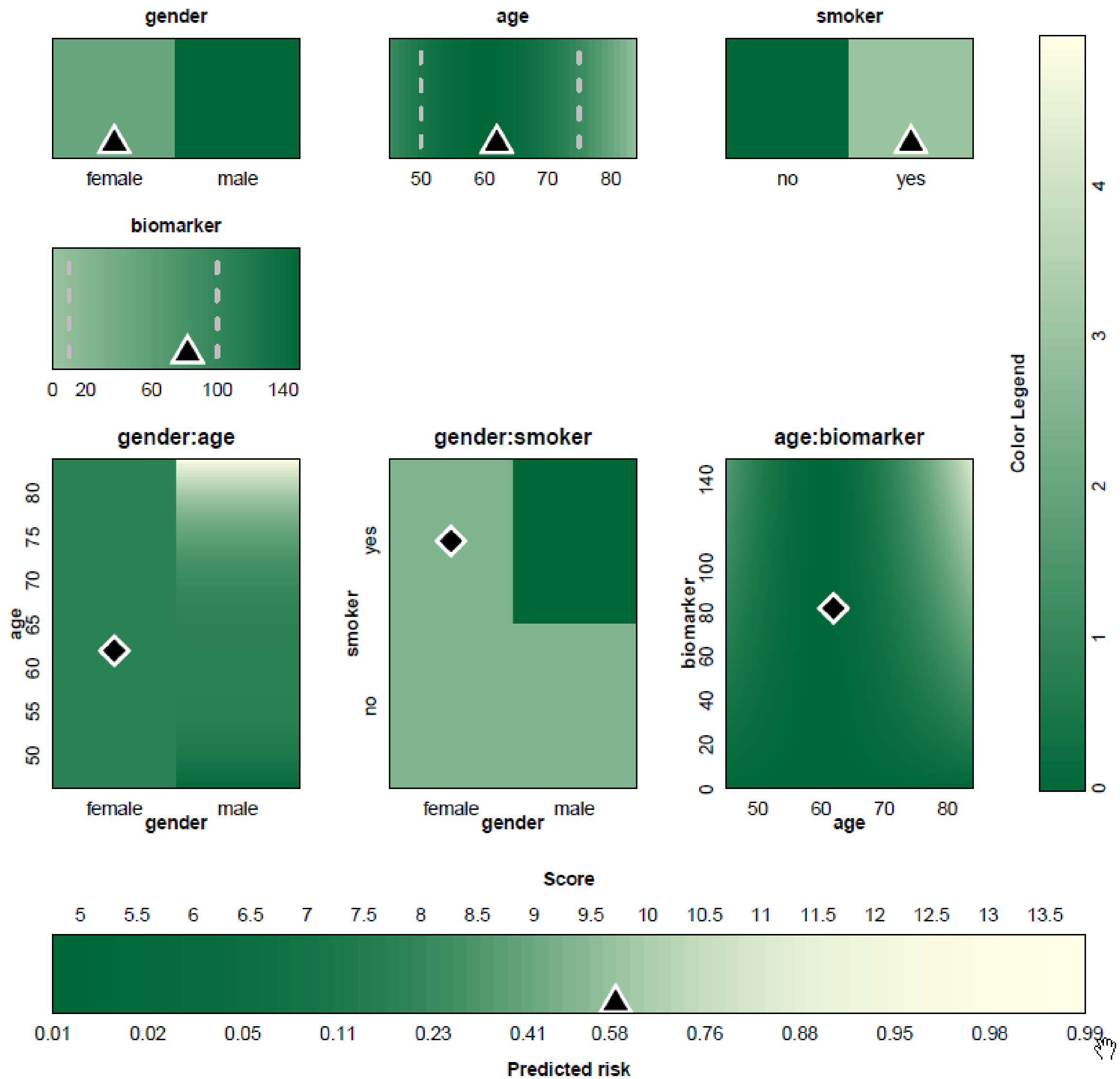
Baesens, Rösch, Scheule, Credit Risk Analytics, Wiley, 2016.

Model Representation: Nomogram



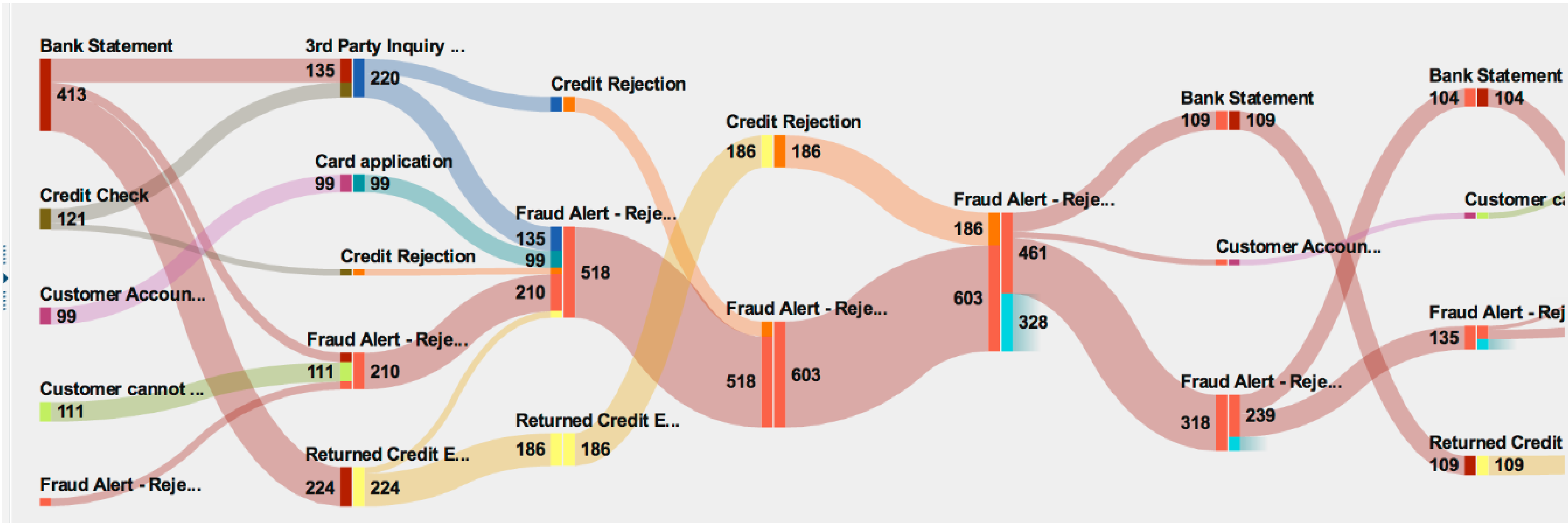
Van Belle and Van Calster (2015)

Model Representation: Nomogram



Van Belle and Van Calster (2015)

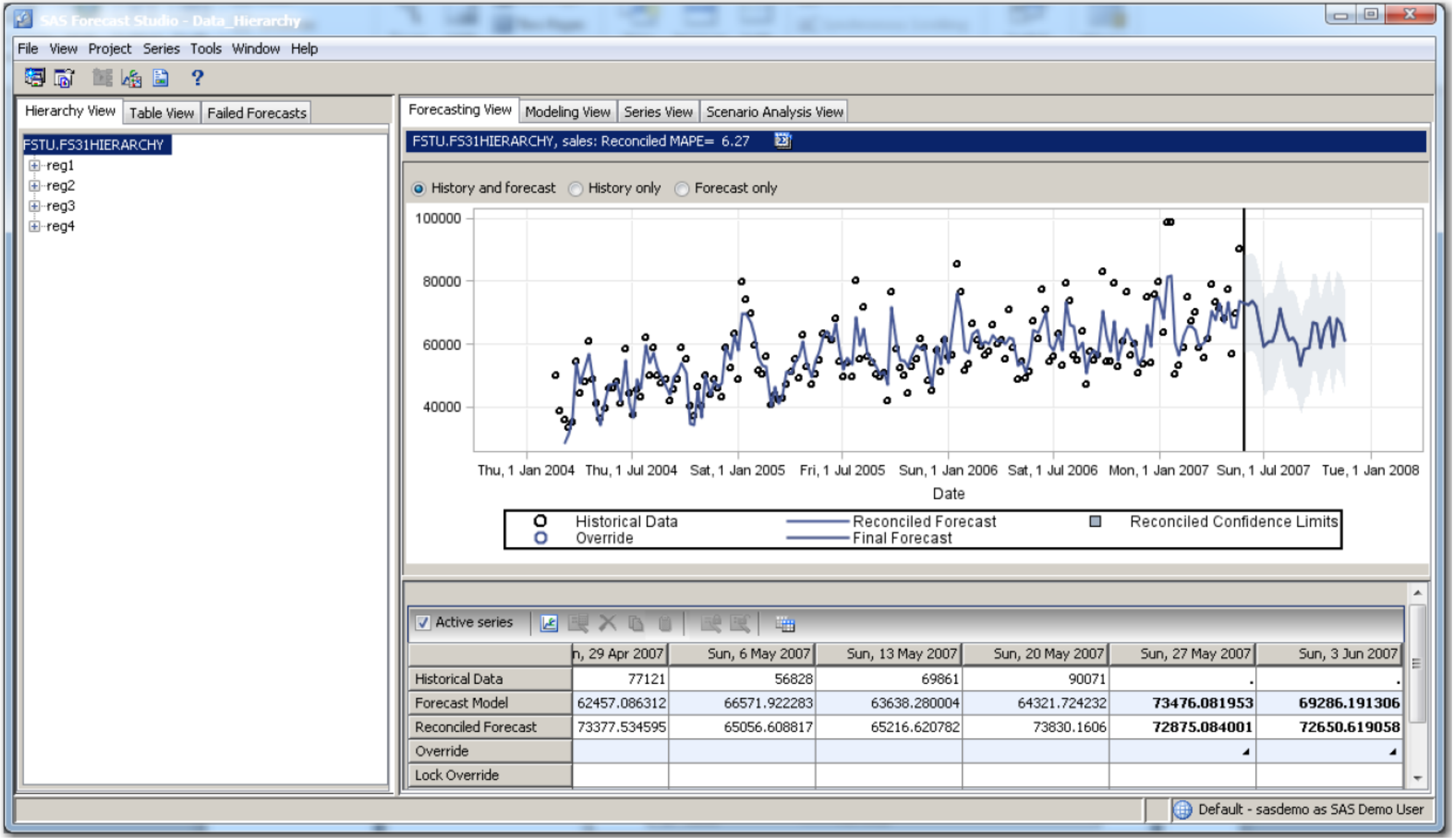
Model Representation: Sankey plot



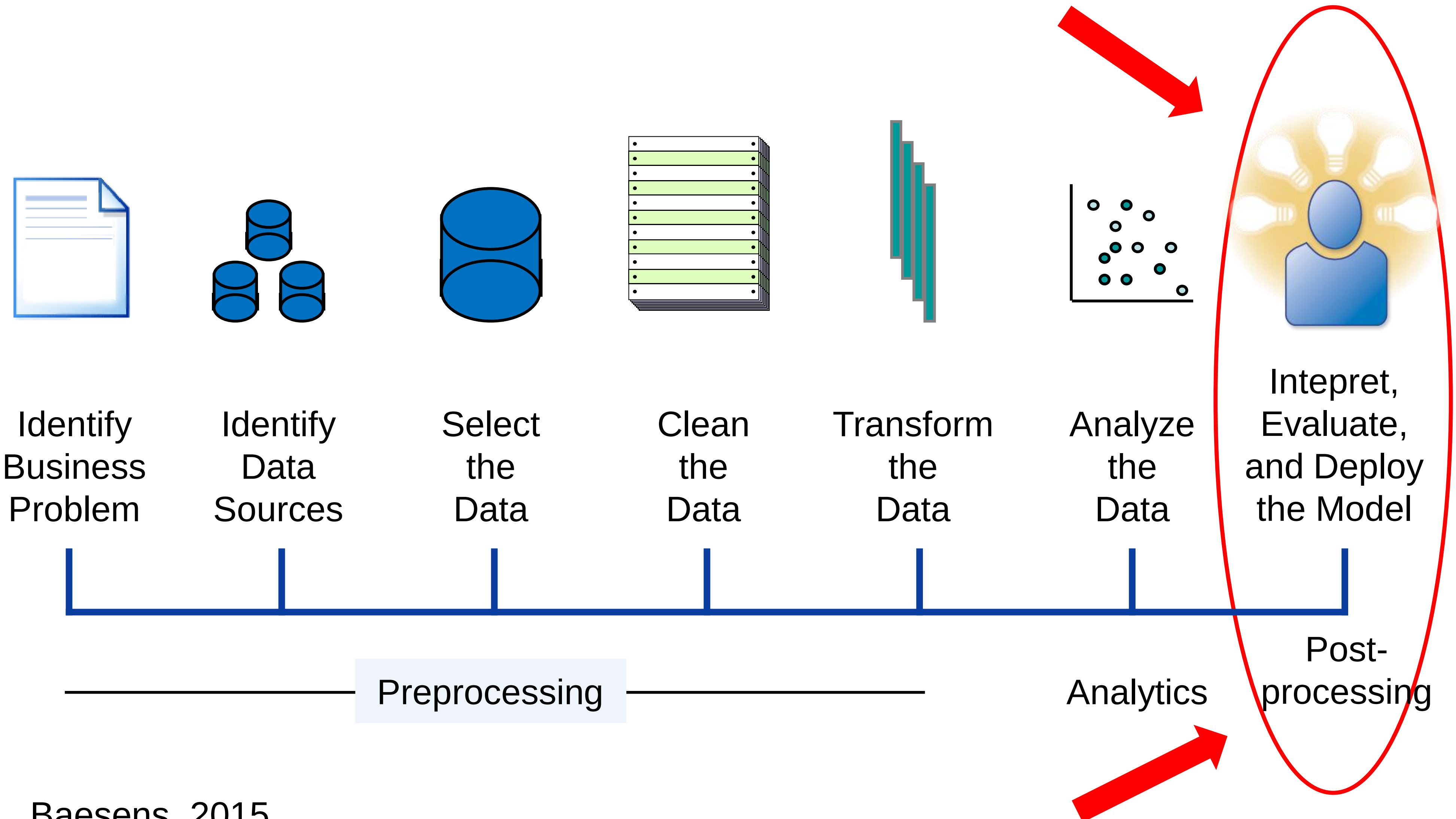
Customer Journey Analytics!

www.sas.com

Model Representation: Time Series

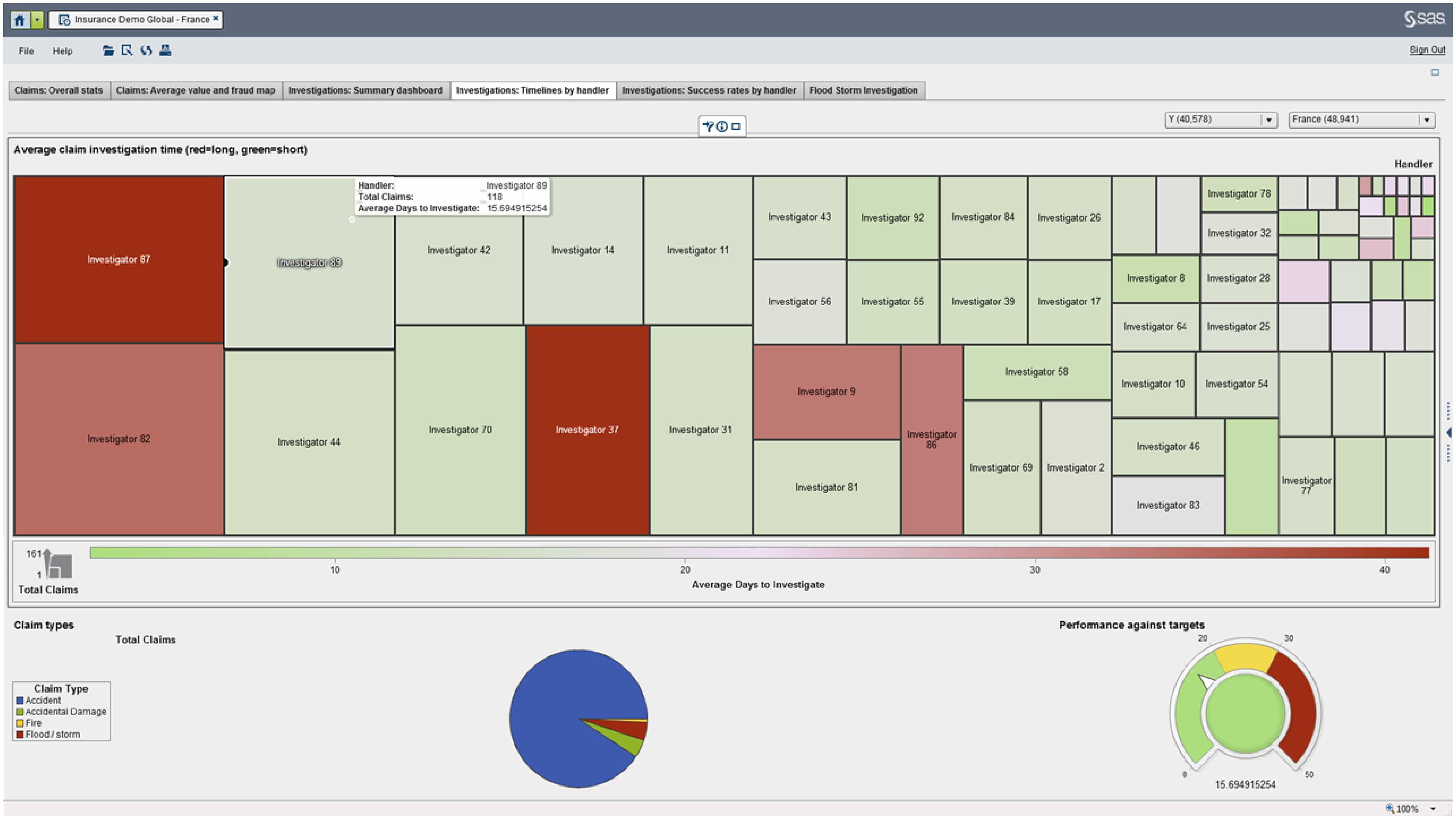


The Analytics Process Model



Baesens, 2015.

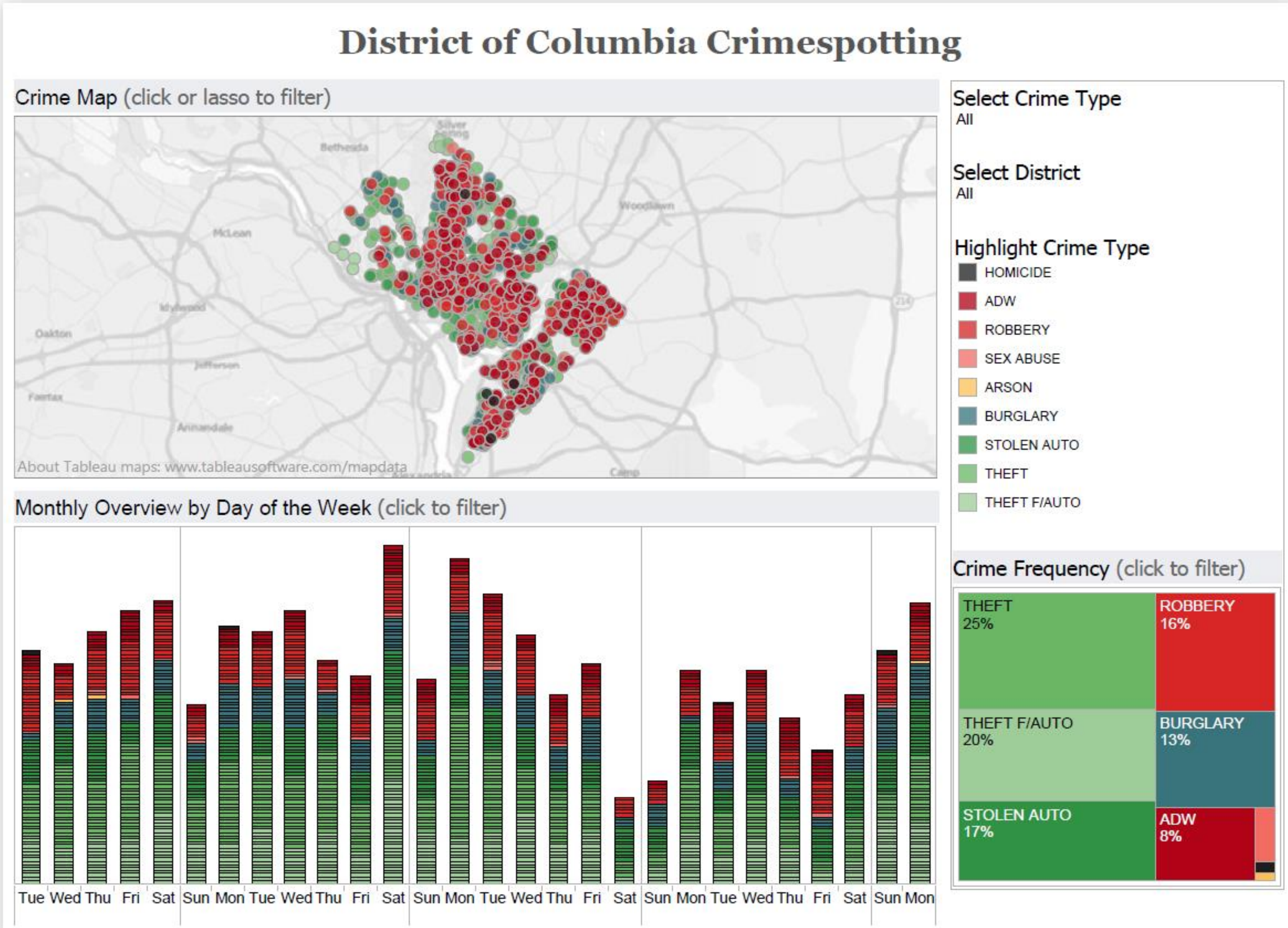
Model Usage: Treemap



www.sas.com

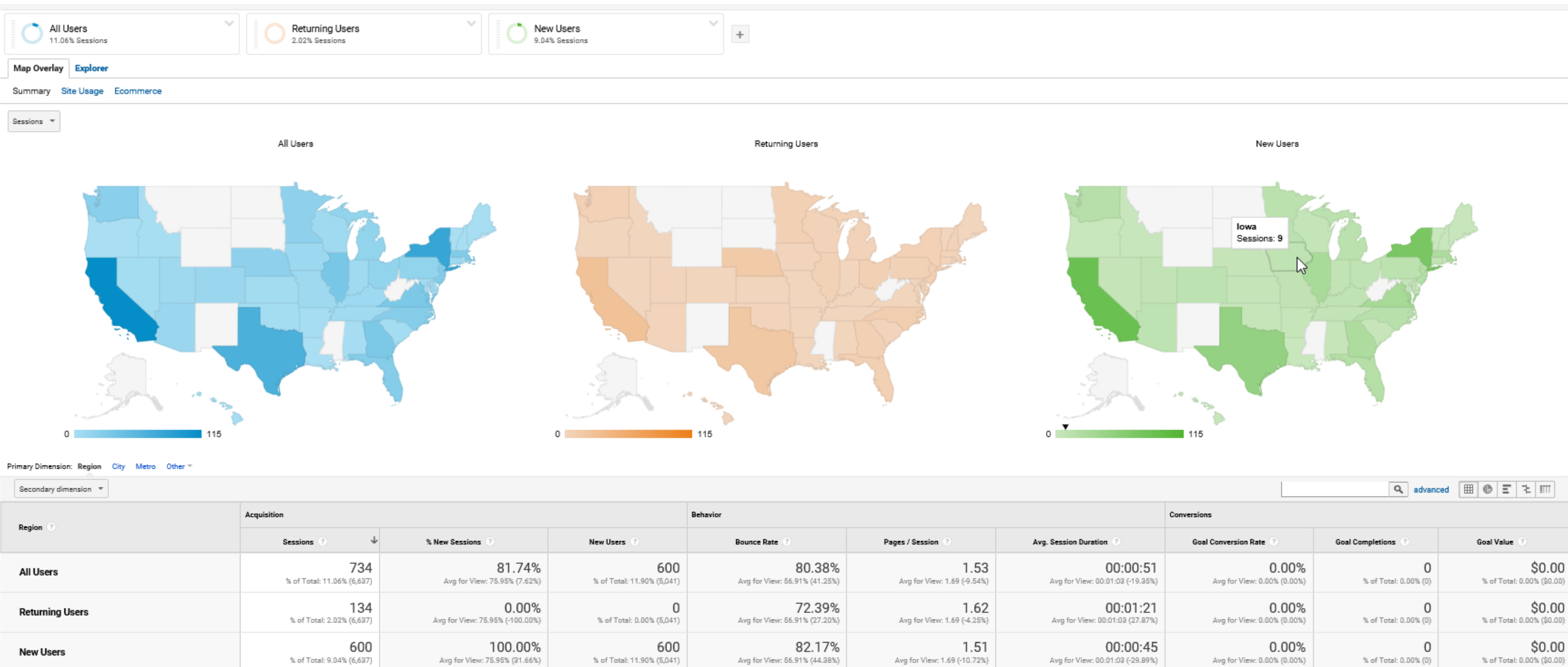
Model Usage: Geospatial plots

DEMO TIME!



https://public.tableau.com/en-us/s/gallery/district-columbia-crimespotting

Model Usage: Segmentation



Region	Acquisition			Behavior			Conversions		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Goal Conversion Rate	Goal Completions	Goal Value
All Users	734 % of Total: 11.06% (6,637)	81.74% Avg for View: 75.95% (7.62%)	600 % of Total: 11.90% (5,041)	80.38% Avg for View: 56.91% (41.25%)	1.53 Avg for View: 1.69 (-9.54%)	00:00:51 Avg for View: 00:01:03 (-19.35%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)	$0.00 % of Total: 0.00% ($0.00)
Returning Users	134 % of Total: 2.02% (6,637)	0.00% Avg for View: 75.95% (-100.00%)	0 % of Total: 0.00% (5,041)	72.39% Avg for View: 56.91% (27.20%)	1.62 Avg for View: 1.69 (-4.25%)	00:01:21 Avg for View: 00:01:03 (27.87%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)	$0.00 % of Total: 0.00% ($0.00)
New Users	600 % of Total: 9.04% (6,637)	100.00% Avg for View: 75.95% (31.66%)	600 % of Total: 11.90% (5,041)	82.17% Avg for View: 56.91% (44.38%)	1.51 Avg for View: 1.69 (-10.72%)	00:00:45 Avg for View: 00:01:03 (-29.89%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)	$0.00 % of Total: 0.00% ($0.00)

Google Analytics

Model Backtesting: Traffic Light Indicator Approach

PD	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
	0.26%	0.17%	0.42%	0.53%	0.54%	1.36%	2.46%	5.76%	8.76%	20.89%	3.05%
DR	**Baa1**	**Baa2**	**Baa3**	**Ba1**	**Ba2**	**Ba3**	**B1**	**B2**	**B3**	**Caa-C**	**Av**
1993	0.00%	0.00%	0.00%	0.83%	0.00%	0.76%	3.24%	5.04%	11.29%	28.57%	3.24%
1994	0.00%	0.00%	0.00%	0.00%	0.00%	0.59%	1.88%	3.75%	7.95%	5.13%	1.88%
1995	0.00%	0.00%	0.00%	0.00%	0.00%	1.76%	4.35%	6.42%	4.06%	11.57%	2.51%
1996	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.17%	0.00%	3.28%	13.99%	0.78%
1997	0.00%	0.00%	0.00%	0.00%	0.00%	0.47%	0.00%	1.54%	7.22%	14.67%	1.41%
1998	0.00%	0.31%	0.00%	0.00%	0.62%	1.12%	2.11%	7.55%	5.52%	15.09%	2.83%
1999	0.00%	0.00%	0.34%	0.47%	0.00%	2.00%	3.28%	6.91%	9.63%	20.44%	3.35%
2000	0.28%	0.00%	0.97%	0.94%	0.63%	1.04%	3.24%	4.10%	10.88%	19.65%	3.01%
2001	0.27%	0.27%	0.00%	0.51%	1.38%	2.93%	3.19%	11.07%	16.38%	34.45%	5.48%
2002	1.26%	0.72%	1.78%	1.58%	1.41%	1.58%	2.00%	6.81%	6.86%	29.45%	3.70%
Av	0.26%	0.17%	0.42%	0.53%	0.54%	1.36%	2.46%	5.76%	8.76%	20.9%	3.05%

Baesens, Rösch, Scheule, Credit Risk Analytics, Wiley, 2016.

Model Backtesting: Traffic Light Indicator Approach

Green		everything is okay
Yellow		decreasing performance, which can be interpreted as an early warning
Orange		performance difference that should be closely monitored
Red		severe problem

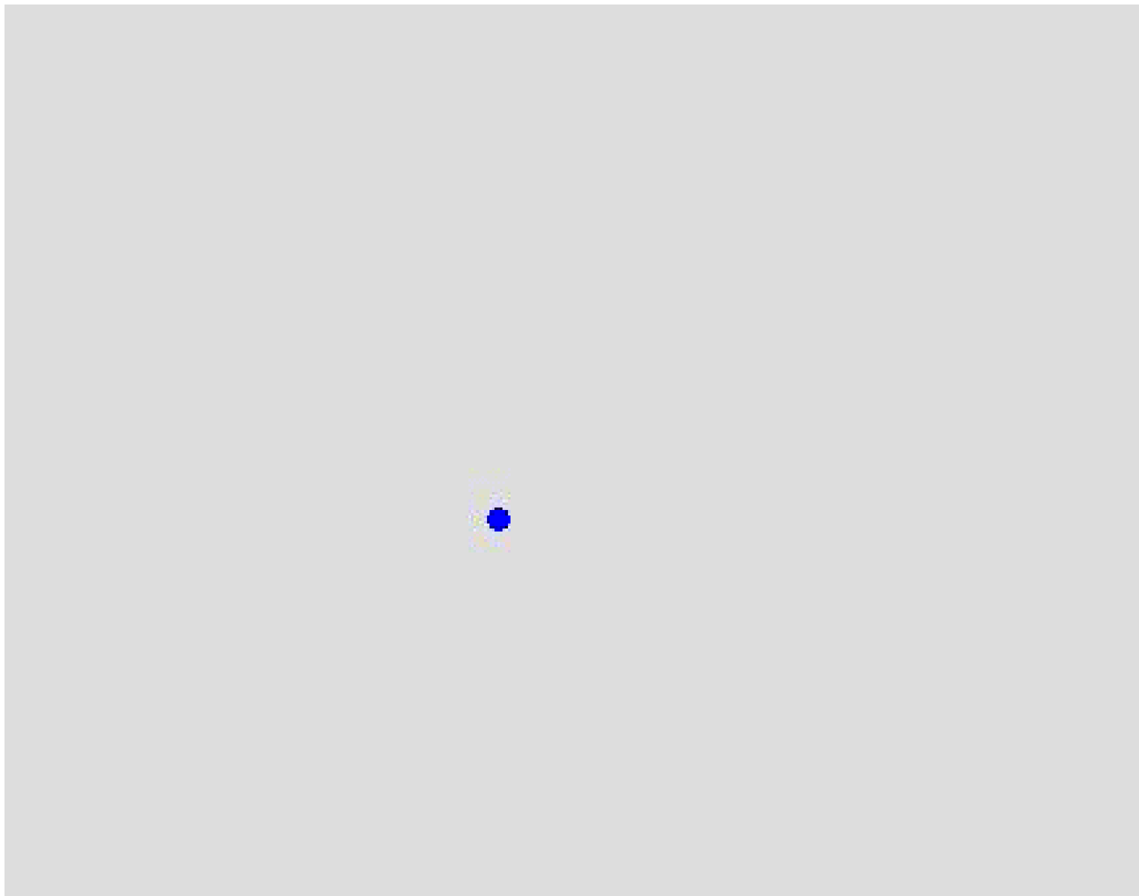
Colors can be defined based on *p*-values.

- *p*-value less than 0.01 = red
- *p*-value between 0.01 and 0.05 = orange
- *p*-value between 0.05 and 0.10 = yellow
- *p*-value higher than 0.10 = green

Baesens, Rösch, Scheule, Credit Risk Analytics, Wiley, 2016.

Visualing Temporal Patterns

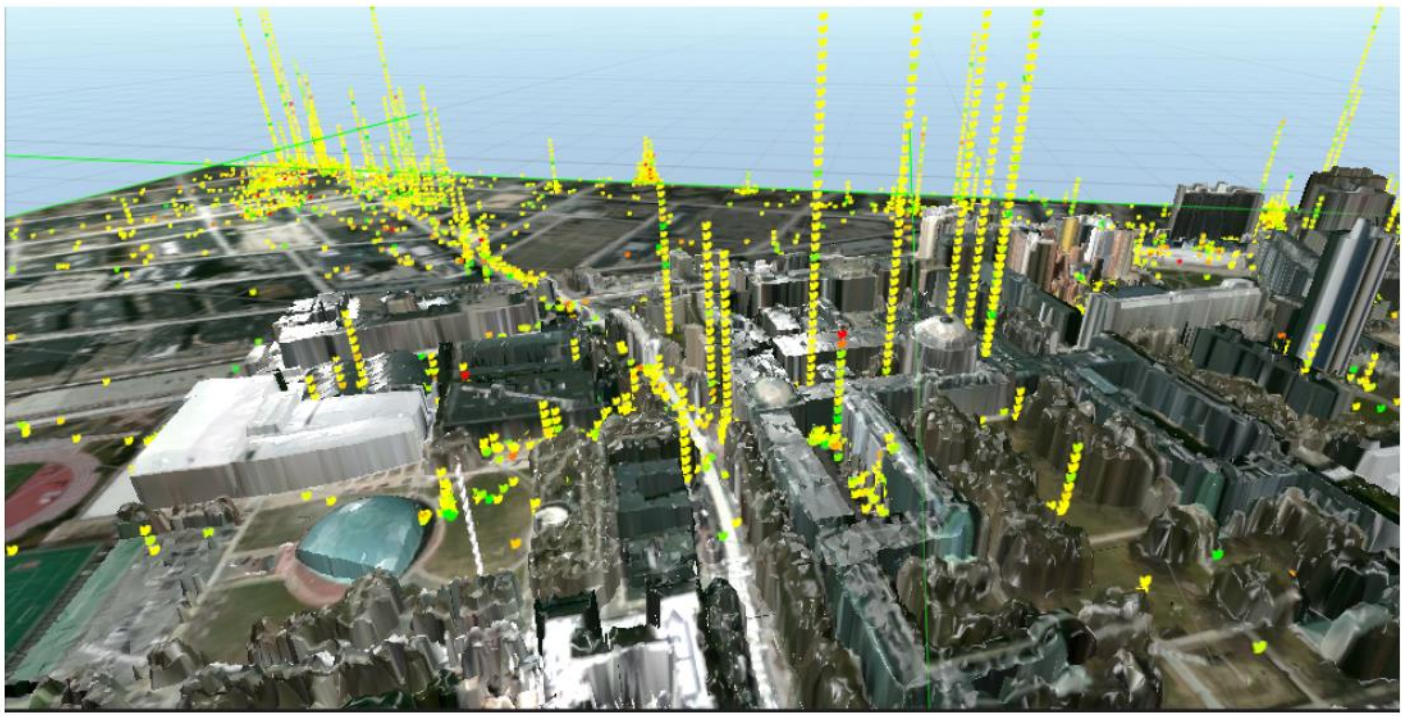
- E.g. Churn Prediction in Telco



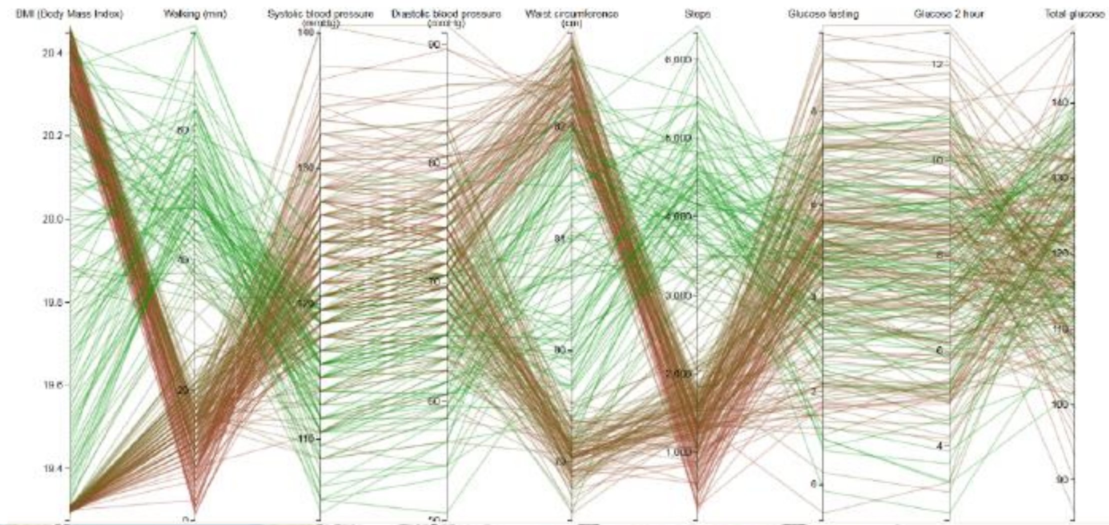
Homophily!

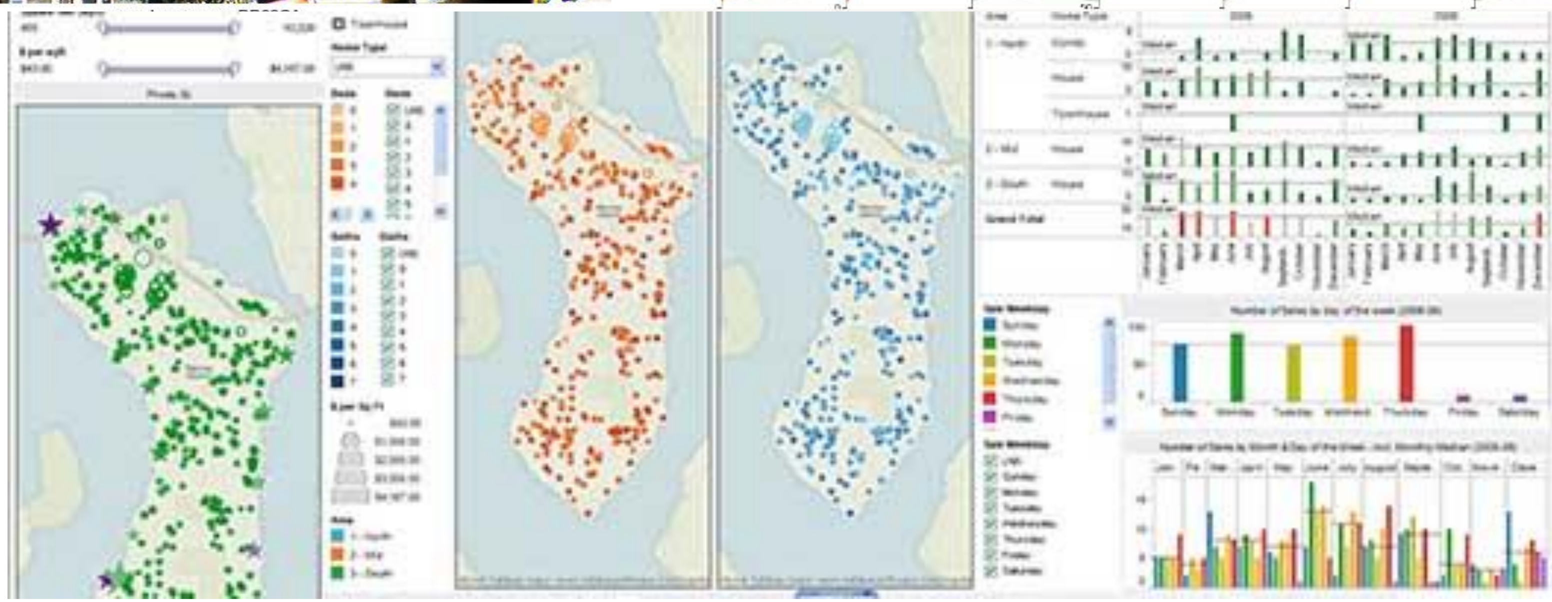
Virtual Reality

- Aim is to create an immersive environment for the user
- E.g. Twitter Sentiment on MIT Campus using geo-tagged Tweets (Moran, 2014)



Visual Clutter





Visual Analytics: Guidelines

- The Visual clutter trap
  - From "information overload" to "visual overload"
  - Humans can only distinguish around 8 colors in 1 visual
- Invoke business user analytical curiosity
- Interactivity
- Consistency
- Avoid scrollbars using range sizing
- Naming, Naming, Naming!
  - e.g. axes, legends, units, currencies, coding schemes, etc.

Software

- SAS Visual Analytics (SAS)
- JMP (SAS)
- Tableau (Tableau)
- QlikView (Qlik)
- Spotfire (Tibco)
- i2 Analyst Notebook (IBM)
- Microsoft BI stack (Excel, PowerPivot, SQL Server)
- Open source: R Shiny, Python igraph, Javascript d3, http://www.infovis-wiki.net

Conclusions

- Visual analytics permeates the entire analytics process!
- Visual analytics catalyzes
  - model discovery
  - model interpretation
  - model monitoring
- Stay focussed; avoid the visual clutter trap!

References

- Anscombe F.J., Graphs in Statistical Analysis, *American Statistician*, 27 (1), pp. 17-21, 1973.
- Baesens B., *Analytics in a Big Data World*, 2014, Wiley.
- Baesens B., Rösch D., Scheule H., *Credit Risk Analytics*, Wiley, 2016.
- Baesens B., Van Vlasselaer V., Verbeke W., *Fraud Analytics*, Wiley, 2015.
- Moran A., Improving Big Data Visual Analytics with Interactive Virtual Reality, *MIT*, 2016.
- Thomas J., Cook K., Illuminating the Path: Research and Development Agenda for Visual Analytics, IEEE-Press, 2005.
- Van Belle V., Van Calster B., Visualizing Risk Prediction Models, *PLoS One, 10* (7), 2015.

Follow-up SAI Events

Courses

- **Analytics: Putting it all to Work (1 day)**
  https://support.sas.com/edu/schedules.html?ctry=us&id=1339

- **Advanced Analytics in a Big Data World (3 days)**
  https://support.sas.com/edu/schedules.html?ctry=us&id=2169

- **Credit Risk Modeling (3 days)**
  https://support.sas.com/edu/schedules.html?ctry=us&id=2455

- **Fraud Analytics using Descriptive, Predictive and Social Network Analytics (2 days)**
  https://support.sas.com/edu/schedules.html?ctry=us&id=1912

More Information

E-learning course: Advanced Analytics in a Big Data World

https://support.sas.com/edu/schedules.html?id=2169&ctry=US

The E-learning course starts by refreshing the basic concepts of the analytics process model: data preprocessing, analytics and post processing. We then discuss decision trees and ensemble methods (random forests), neural networks, SVMs, Bayesian networks, survival analysis, social networks, monitoring and backtesting analytical models. Throughout the course, we extensively refer to our industry and research experience. Various business examples (e.g. credit scoring, churn prediction, fraud detection, customer segmentation, etc.) and small case studies are also included for further clarification. The E-learning course consists of more than 20 hours of movies, each 5 minutes on average. Quizzes are included to facilitate the understanding of the material. Upon registration, you will get an access code which gives you unlimited access to all course material (movies, quizzes, scripts, ...) during 1 year. The E-learning course focusses on the concepts and modeling methodologies and not on the SAS software. To access the course material, you only need a laptop, iPad, iPhone with a web browser. No SAS software is needed.

More Information

E-learning course: Fraud Analytics

https://support.sas.com/edu/schedules.html?ctry=us&id=1912

This new E-learning course will show how learning fraud patterns from historical data can be used to fight fraud. To be discussed is the use of descriptive analytics (using an unlabeled data set), predictive analytics (using a labeled data set) and social network learning (using a networked data set). The techniques can be applied across a wide variety of fraud applications, such as insurance fraud, credit card fraud, anti-money laundering, healthcare fraud, telecommunications fraud, click fraud, tax evasion, counterfeit, etc. The course will provide a mix of both theoretical and technical insights, as well as practical implementation details. The instructor will also extensively report on his recent research insights about the topic. Various real-life case studies and examples will be used for further clarification.