Putting Big Data & Analytics to Work!

Prof. dr. Bart Baesens

Department of Decision Sciences and Information Management, KU Leuven (Belgium)

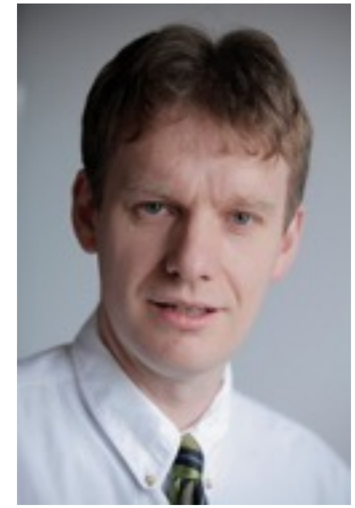
School of Management, University of Southampton (United Kingdom)

Bart.Baesens@kuleuven.be

Twitter/Facebook/YouTube: DataMiningApps

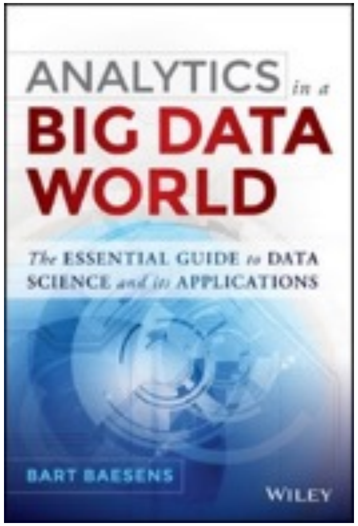
www.dataminingapps.com

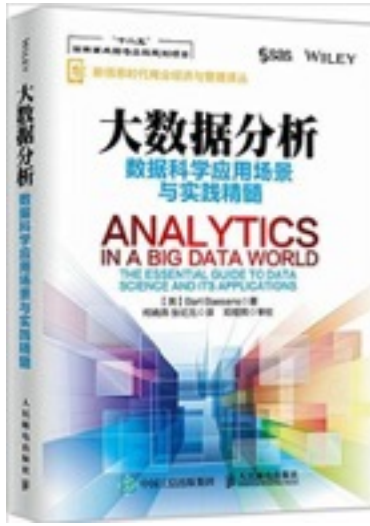
Presenter: Bart Baesens

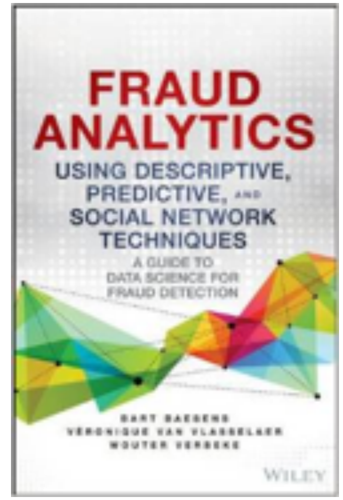


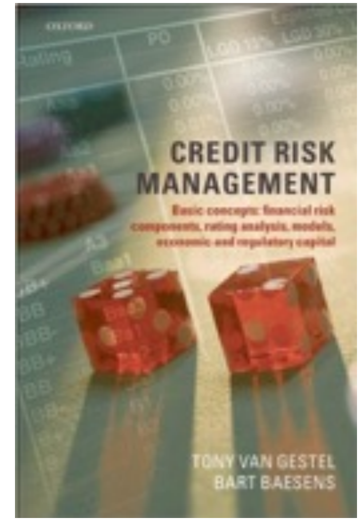
- Studied at KU Leuven (Belgium)
  - Business Engineer in Management Informatics, 1998
  - PhD. in Applied Economic Sciences, 2003
- PhD. : Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
- Professor at KU Leuven, Belgium
- Lecturer at the University of Southampton,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

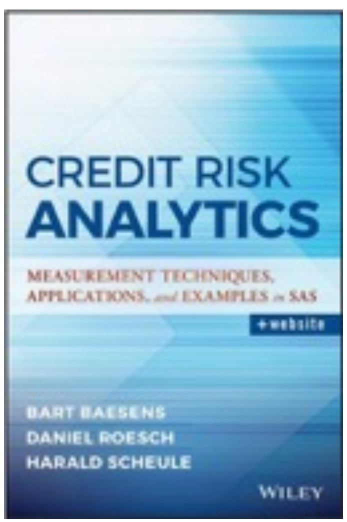
Example Publications

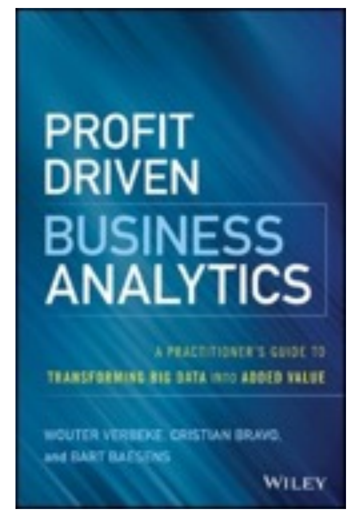




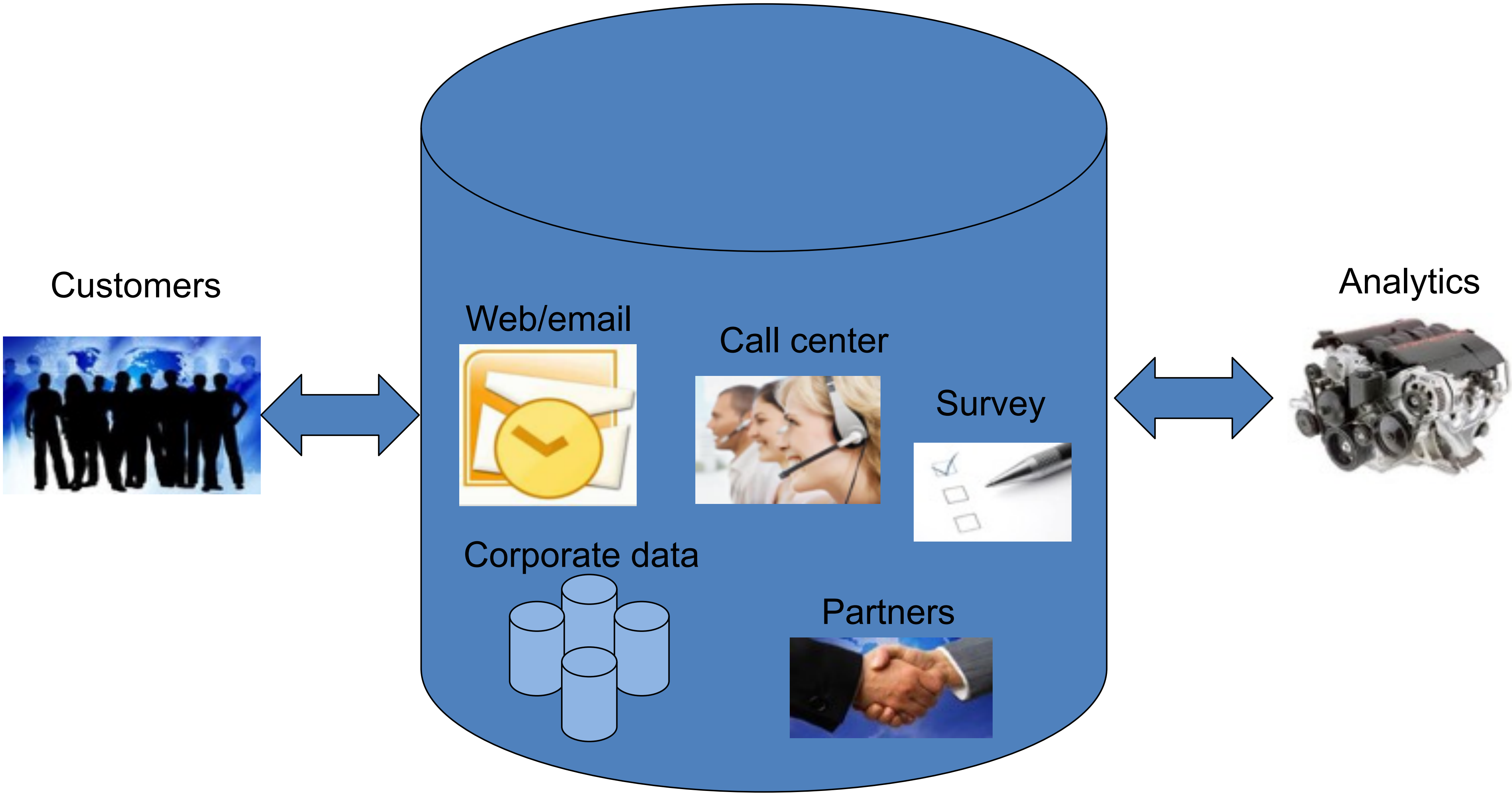




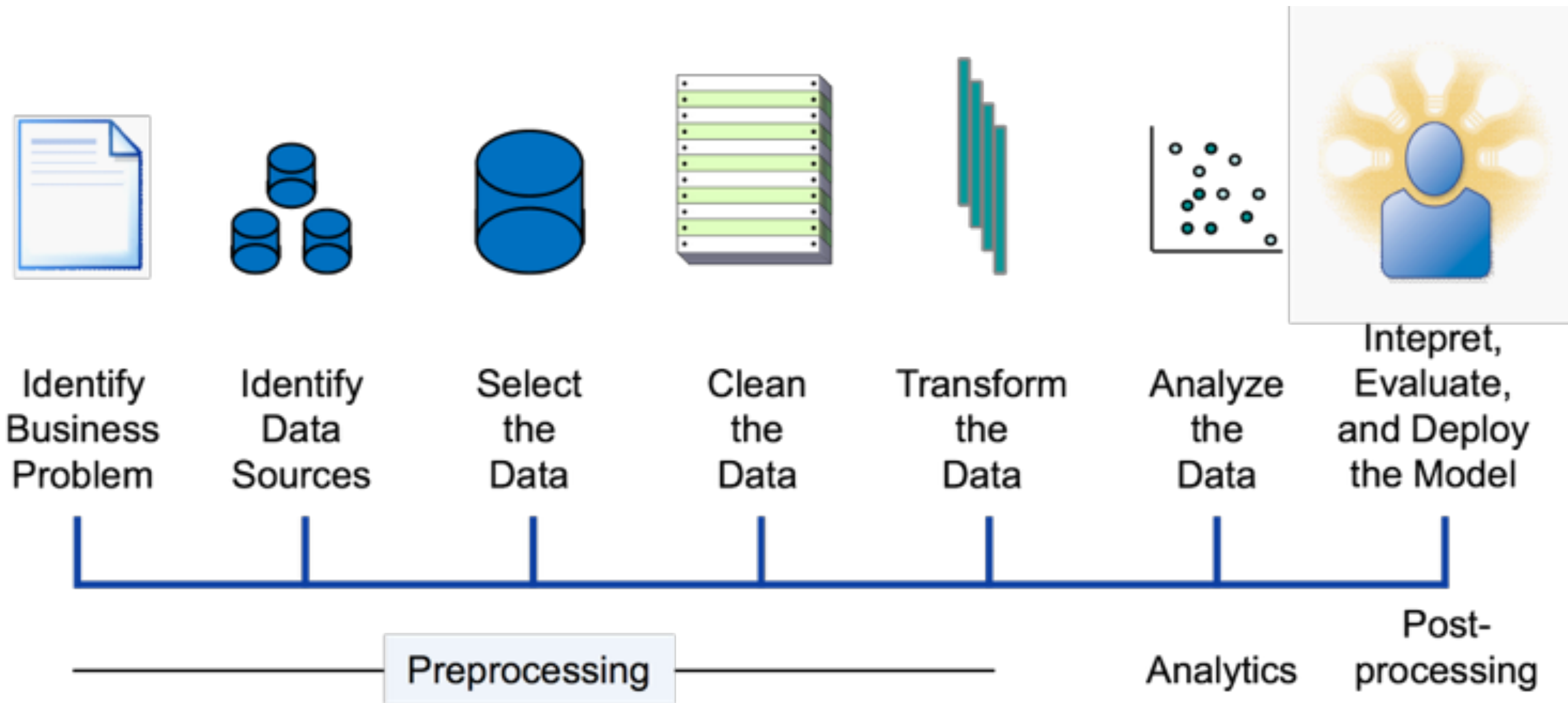




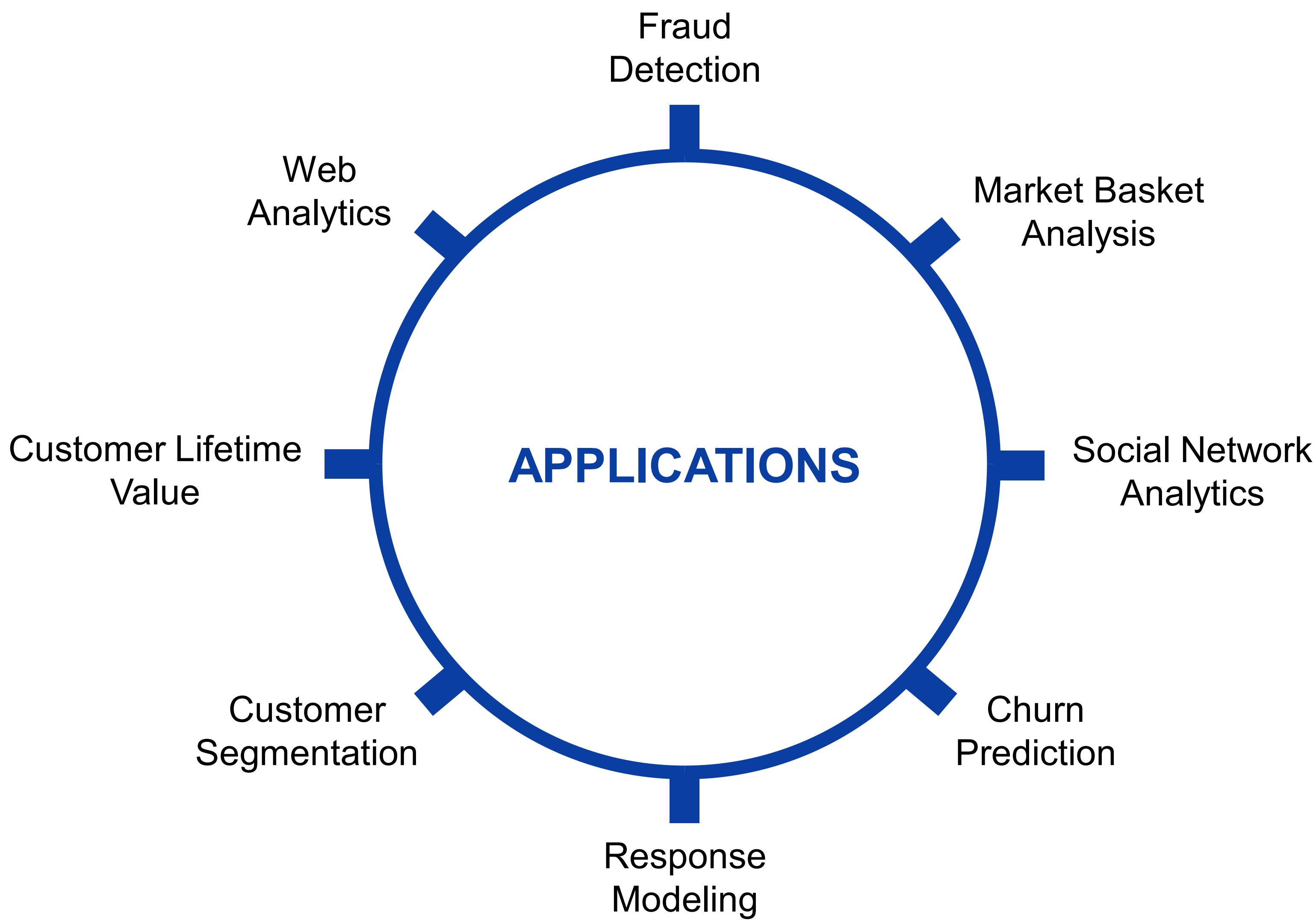
Living in a Data Flooded World!



The Analytics Process Model



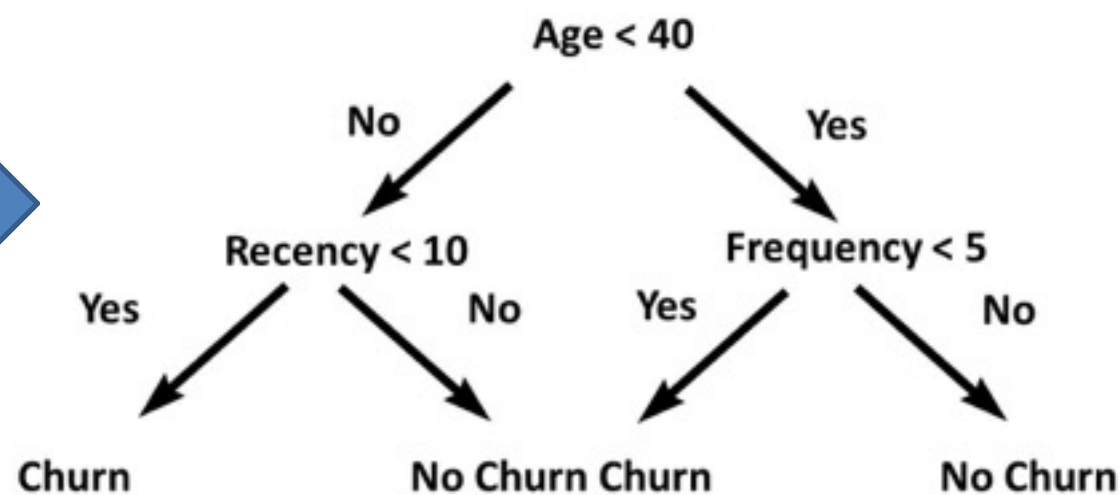
Feel the vibe!



Example: marketing context

Customer	Age	Recency	Frequency	Monetary	Churn
John	35	5	6	100	**Yes**
Sophie	18	10	2	150	**No**
Victor	38	28	8	20	**No**
Laura	44	12	4	280	**Yes**

Analytical Software (SAS, R, Python, ...)



Analytics

- Term often used interchangeably with data science, knowledge discovery, …
- Essentially refers to extracting useful business patterns and/or mathematical decision models from a preprocessed data set
- **Predictive analytics**
  - Predict the future based on patterns learnt from past data
  - Classification (churn, response) versus regression (CLV)
- **Descriptive analytics**
  - Describe patterns in data
  - Clustering, Association rules, Sequence rules

Analytic Model requirements

- **Business relevance**
  - Solve a particular business problem
- **Statistical performance**
  - Statistical significance of model
  - Statistical prediction performance
- **Interpretability + Justifiability**
  - Very subjective (depends on decision maker), but CRUCIAL!
  - Often need to be balanced against statistical performance
- **Operational efficiency**
  - How can the analytical models be integrated with campaign management?
- **Economical cost**
  - What is the cost to gather the model inputs and evaluate the model?
  - Is it worthwhile buying external data and/or models?
- **Regulatory compliance**
  - In accordance with regulation and legislation

Post processing

- Interpretation and validation of analytical models by business experts
  - Trivial versus unexpected (interesting?) patterns
- Sensitivity analysis
  - How sensitive is the model wrt sample characteristics, assumptions and/or technique parameters?
- Deploy analytical model into business setting
  - Represent model output in a user-friendly way
  - Integrate with campaign management tools and marketing decision engines
- Model monitoring and backtesting
  - Continuously monitor model output
  - Contrast model output with observed numbers

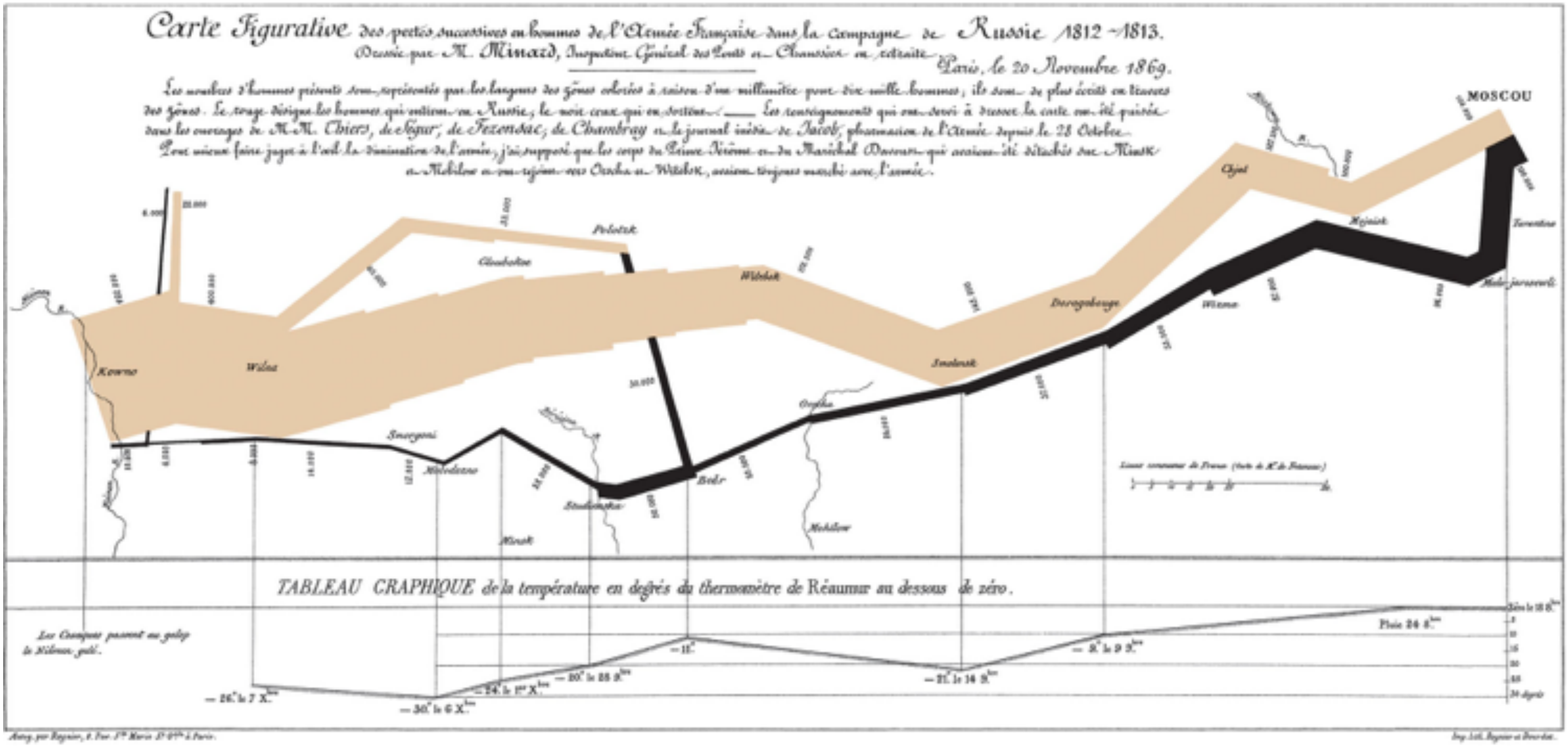
Two Analytical Disconnects

- **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



Charles Minnard, 1869

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!

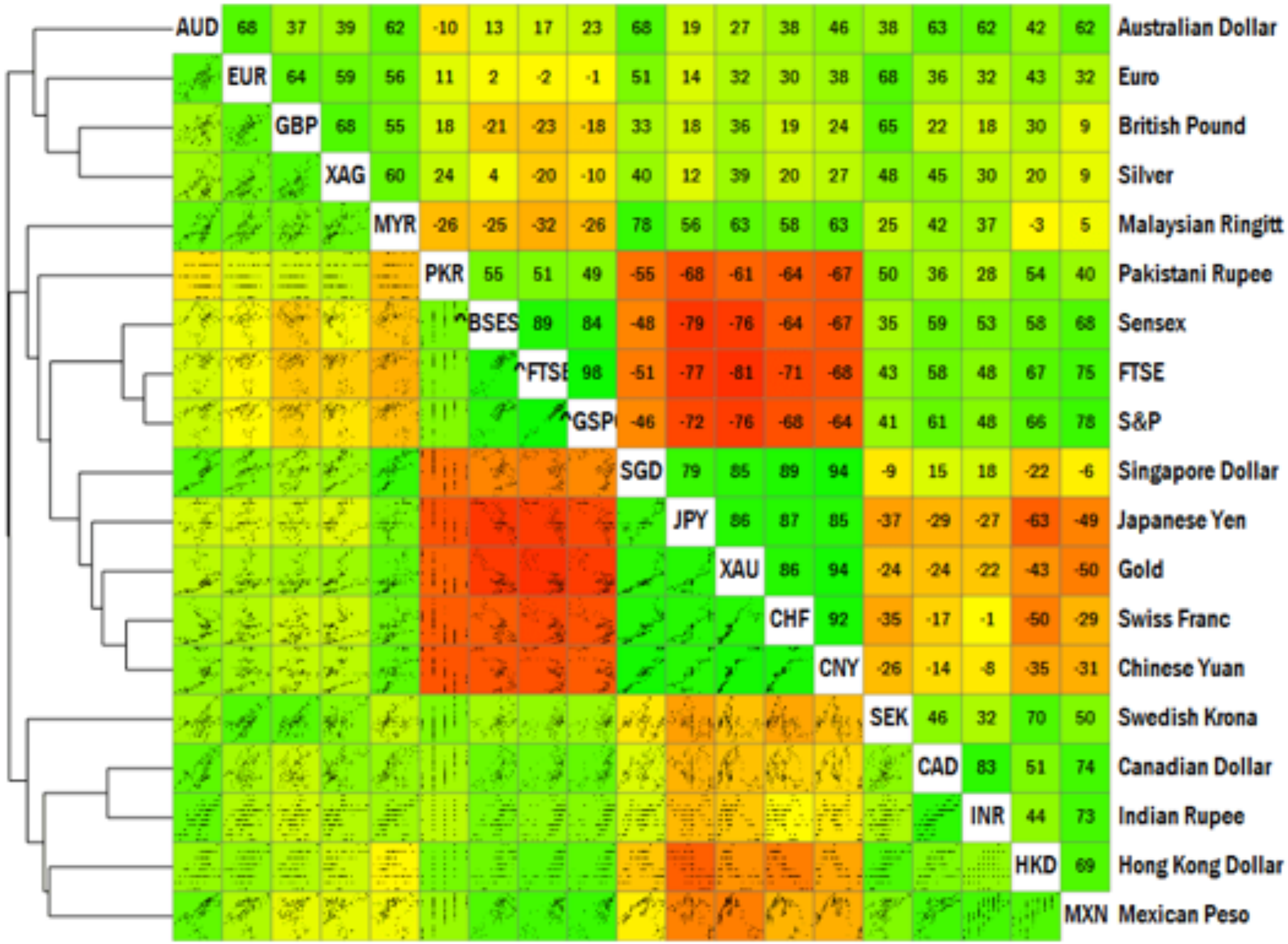
- **Model representation**
  - Use Visual Analytics to represent models in a user-friendly way
  - Visualisation unit: Model formula!

Visual Analytics versus the Analytics Process Model

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

- **Model backtesting**
  - Use Visual Analytics to monitor model performance
  - Visualisation unit: Model performance!

Data Preprocessing: cluster plot



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

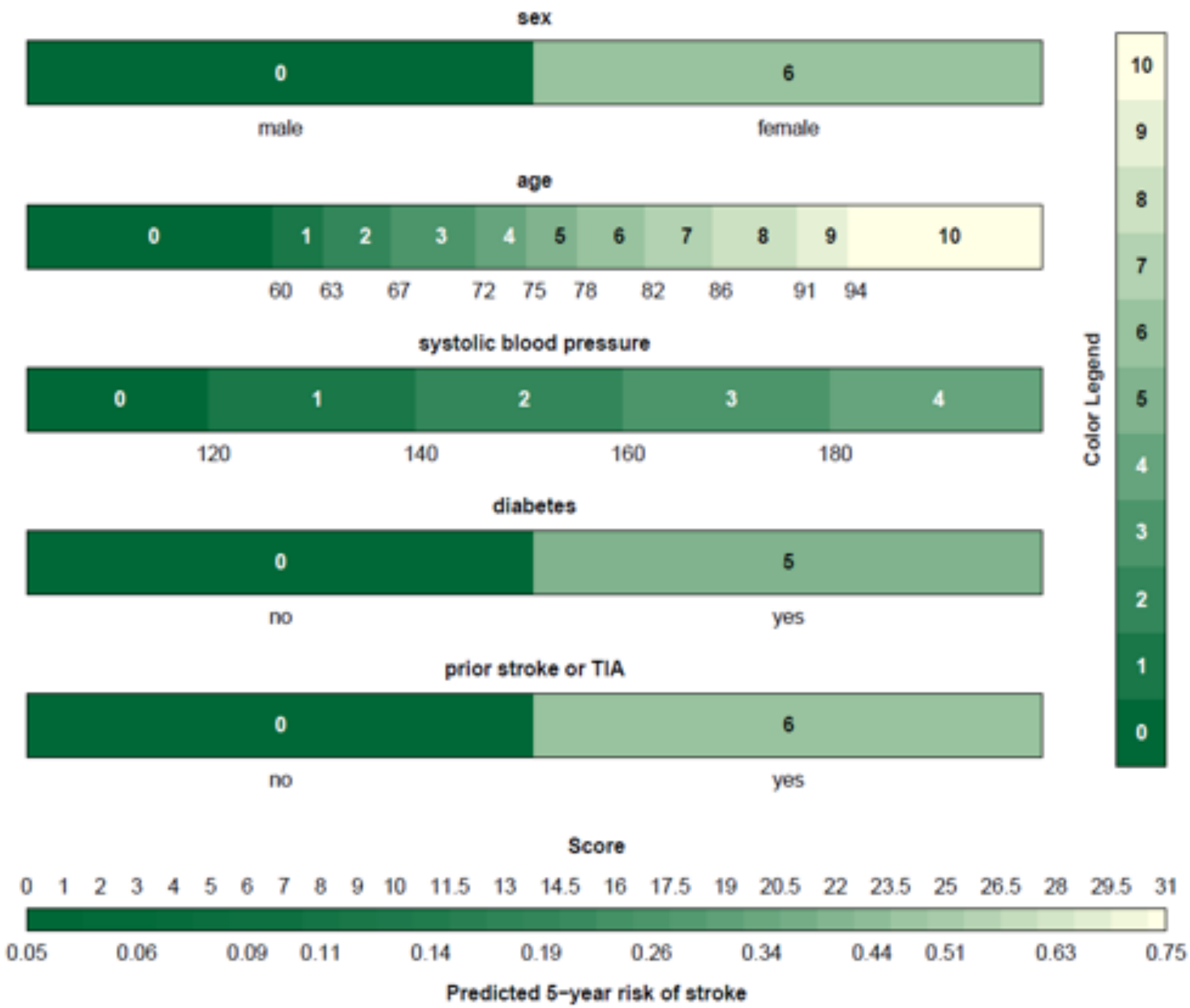
Model Representation: Scorecards

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

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

- 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 **trust**
- 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?

Baesens, Van Vlasselaer, Verbeke, 2015.

Rule Coverage?

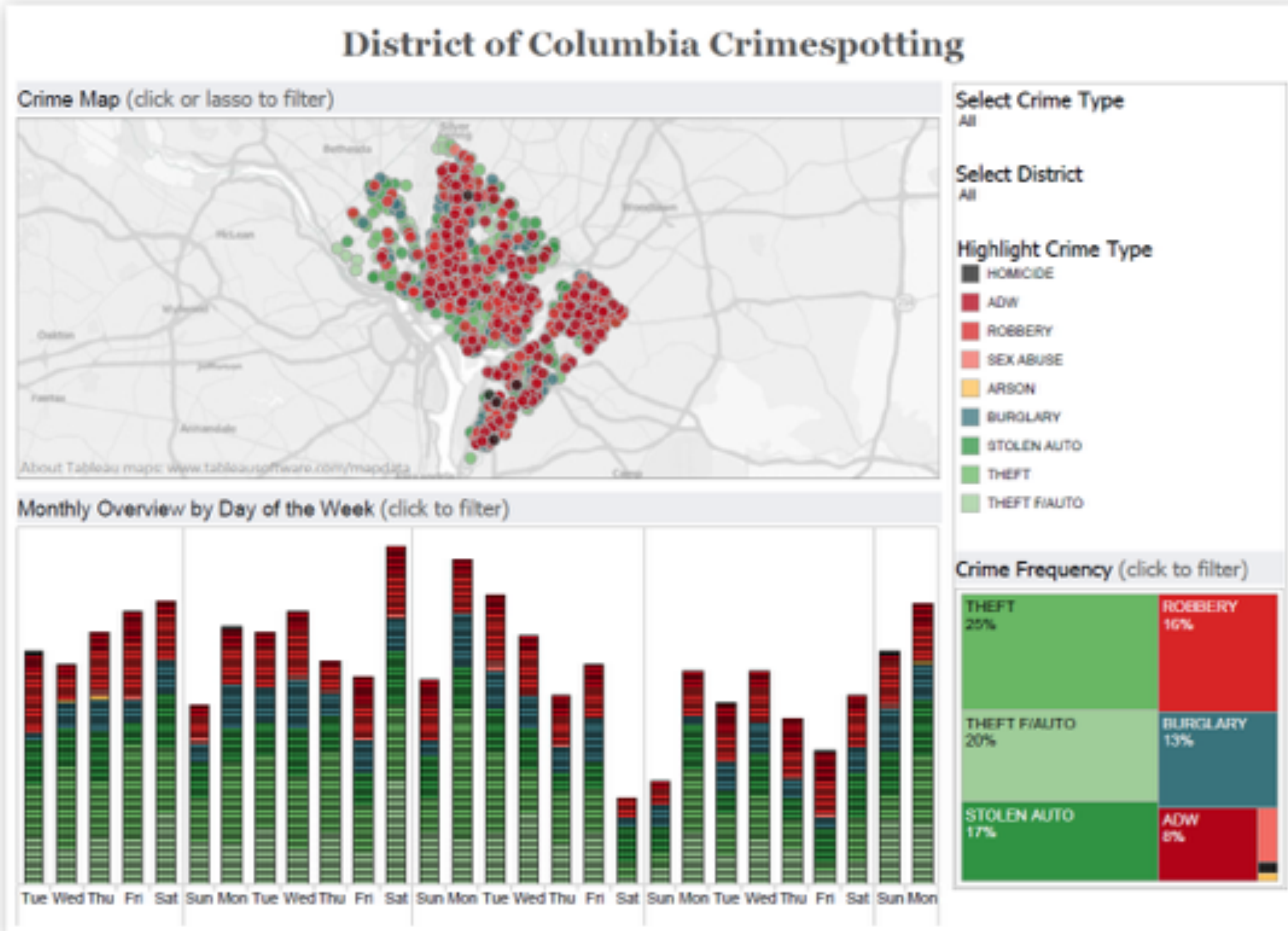
Model Representation: Decision Tables

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!

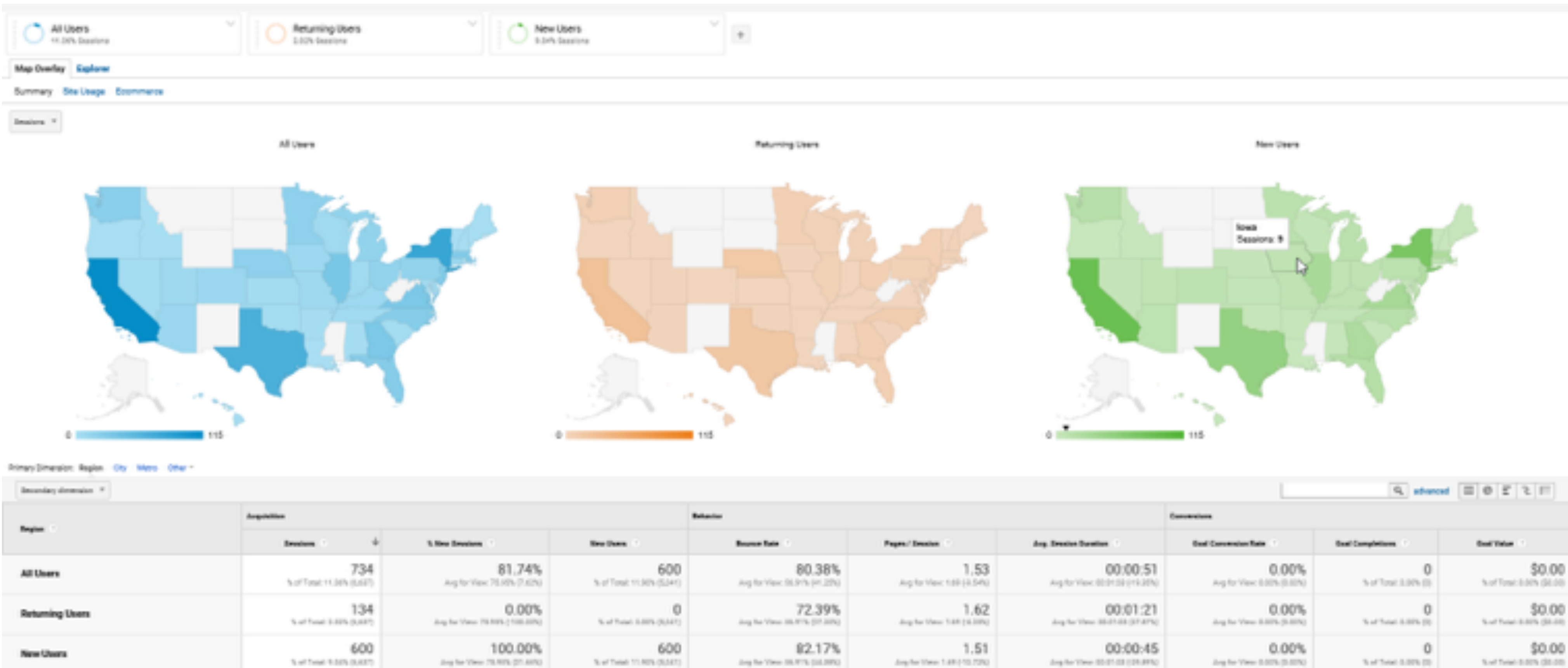
No coverage!

Model Usage: Geospatial plots



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

Model Usage: Segmentation



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

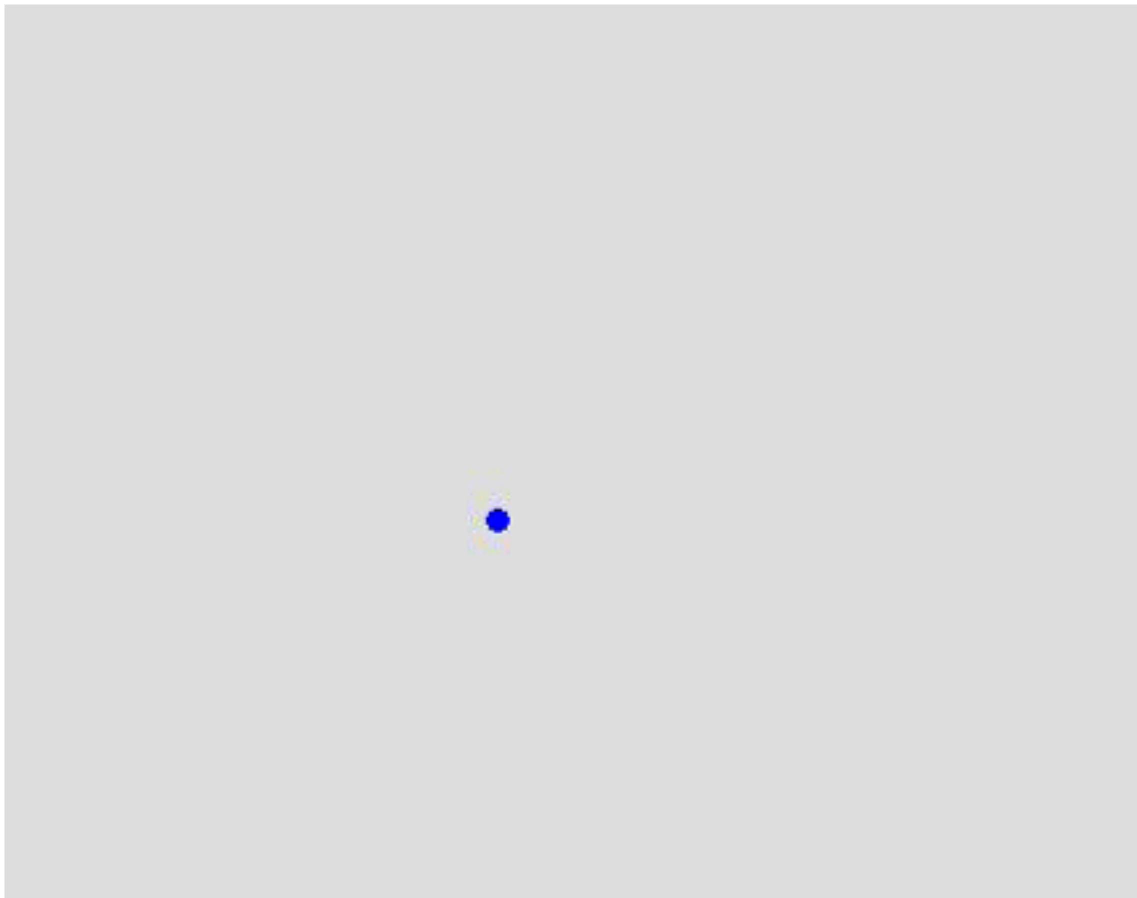
Colors can be defined based on *p*-values.

- *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!

Conclusions

- Be aware but critical about emerging technologies (e.g. deep learning)
- Validation of patterns is key!
- Profit driven analytics (TCO and ROI)
- Visual analytics

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.

More information

[E-learning course: Credit Risk Modeling]

https://support.sas.com/edu/schedules.html?ctry=us&id=2455

The E-learning course covers both the basic as well some more advanced ways of modeling, validating and stress testing Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD) models. Throughout the course, we extensively refer to our industry and research experience. Various business examples and small case studies in both retail and corporate credit 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 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. See https://support.sas.com/edu/schedules.html?ctry=us&id=2455 for more details.