Compact Course in Data Mining

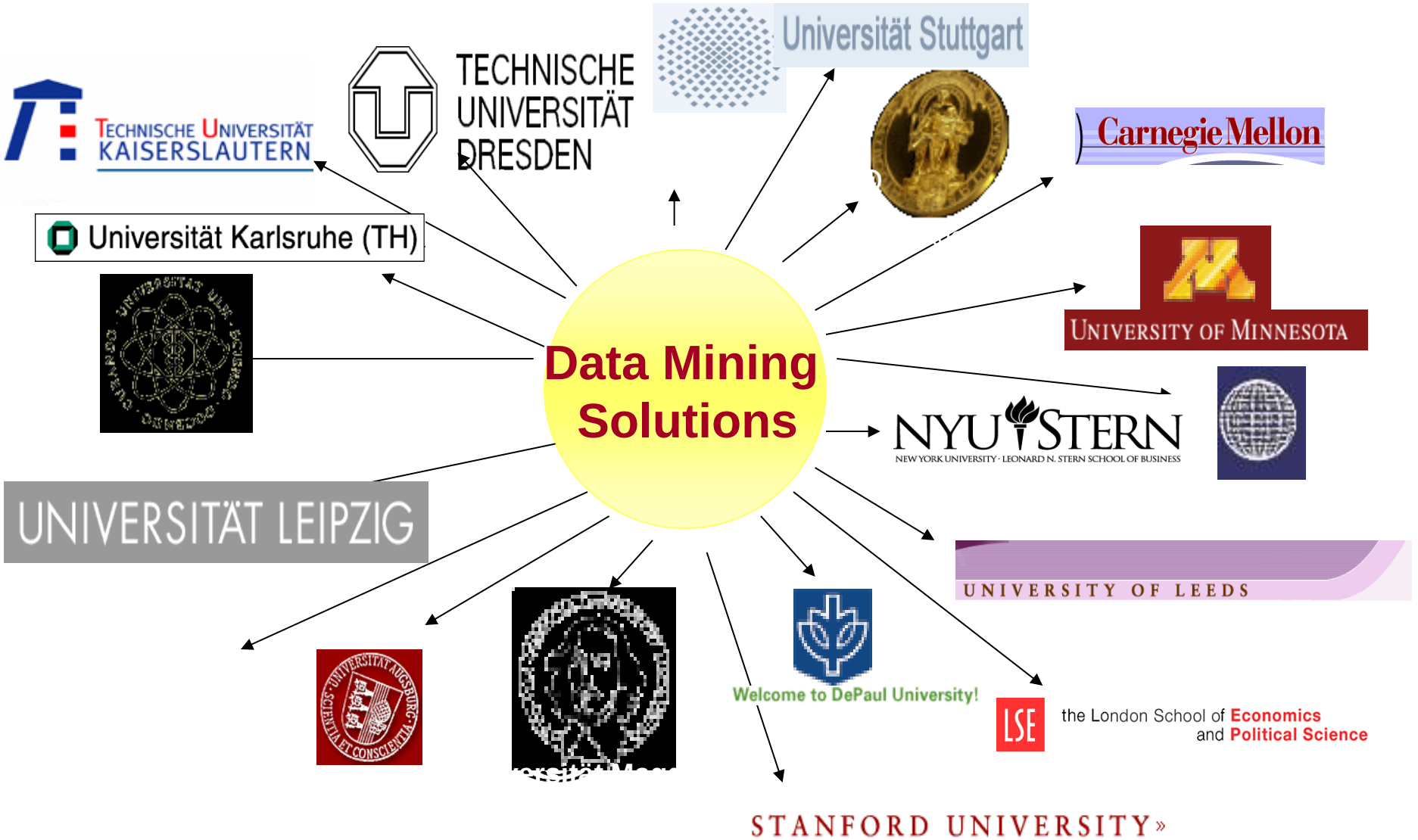
Introduction and General Aspects

Professor Dr. Gholamreza Nakhaeizadeh

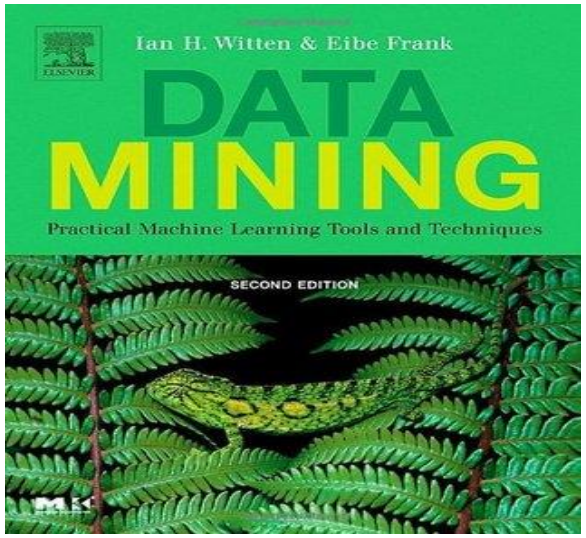
Content

- Why Data Mining?
- What is Data Mining?
- Difference between Data Mining and Knowledge Discovery in Databases
- Interdisciplinary aspects of Data Mining
- Examples of Data Mining Tools
- Short history of Data Mining, Data Mining rapid development
- Some European funded projects on Data Mining
- Scientific Networking and partnership in Data Mining and Machine Learning
- Conducting of Data Mining projects, optimal structure of a Data Mining team
- Success factors of Data Mining projects
- Conferences and Journals on Data Mining

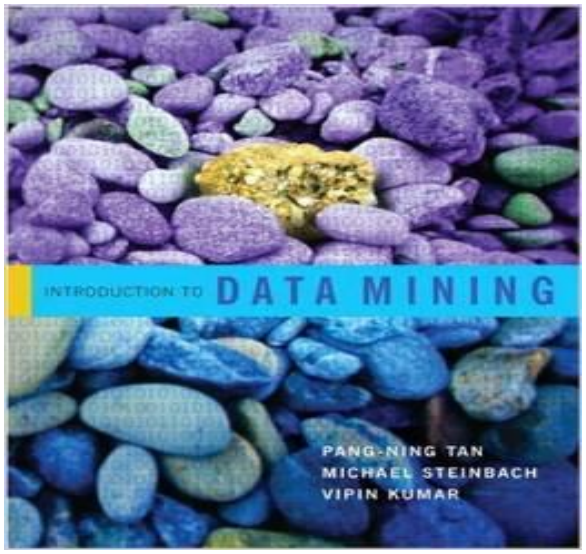
Partnership with universities



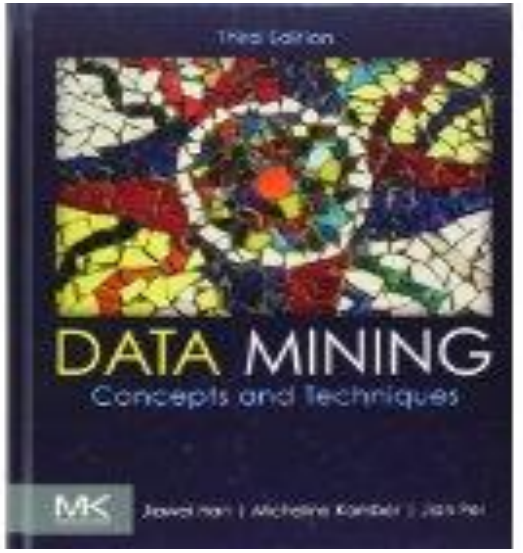
References (Examples)



Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems)
by Ian H. Witten, Eibe Frank



Introduction to Data Mining by
Pang-Ning Tan,
Michael Steinbach ,
Vipin Kumar
ISBN-13: 860-1401421054 ISBN-10: 0321321367
Edition: 1st



Concepts and Techniques,
Third Edition,
The Morgan Kaufmann
by
Jiawei Han
Micheline Kamber
Jian Pei

www.kdnuggets.com

Why Data Mining ?
My own experience: Data Mining in the Automotive Industry

Quality Management

Customer Relationship Management

After Sales

Sales

Supply Chain Management

Production

Credit Risk Management

Development

Suppliers

Data Mining: From high volume data to high value Information

Suppliers Data


saberan.com/product.htm


www.kautex-group.com/.../automotive.html

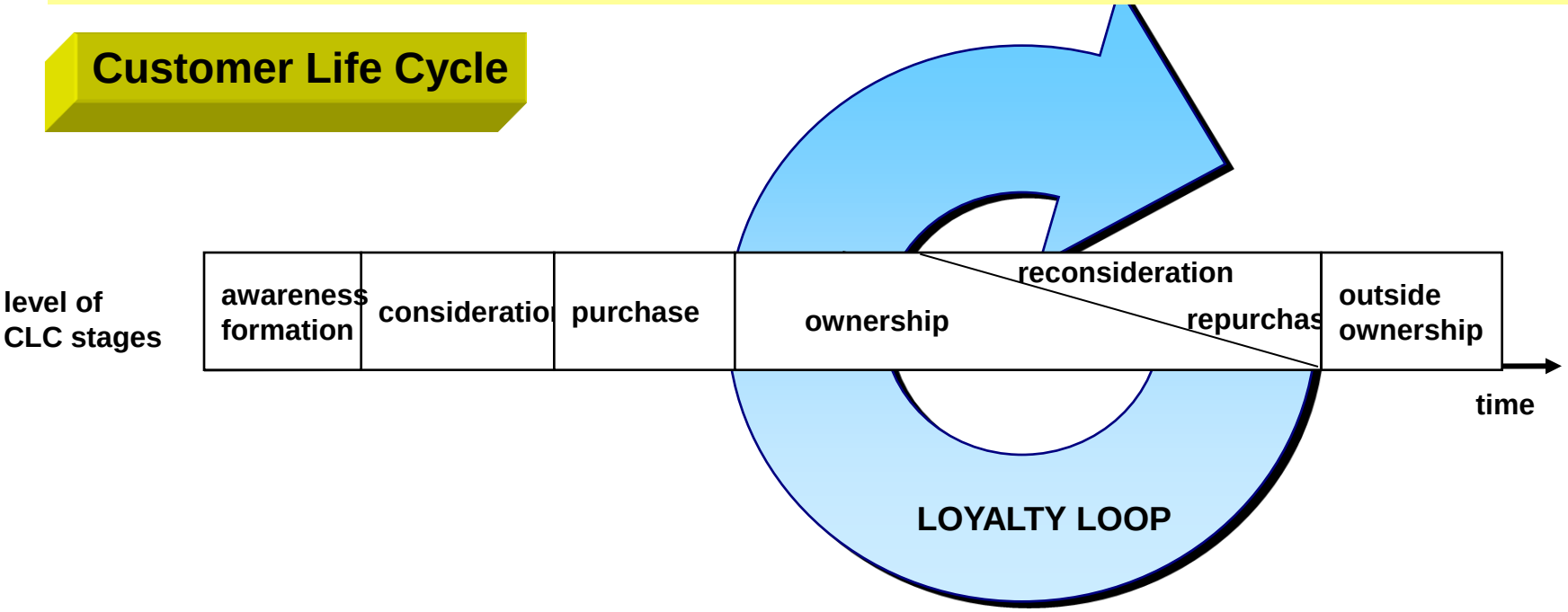
☑	5	12	ST	0056642	SECHSKANTSCHRAUBE
☐	6	2	ST	3007008	PLATTE
☐	7	12	ST	0057130	SECHSKANTMUTTER
☐	8	18	ST	0056668	SECHSKANTSCHRAUBE
☐	9	6	ST	0080731	SCHEIBE
☐	10	6	ST	1781727	LEISTE


picocool.com/go/news/post/car-parts/

Titel	Seite
AUSRUESTUNG	7
LAUFSTEG	11
STEUERBLOCK, VENTIL	82
AUSLEGERZYLINDER	94
STIELZYLINDER	95
LOEFFELZYLINDER	96
LOEFFELZYLINDER	97
MONOAUSLEGER ANBAUTEILE	126
STIEL, BH ANBAUTEILE	130

Why Data Mining in Customer Relationship Management (CRM)?

Customer Life Cycle



acquisition program

loyalty program

acquisition program

Typical aCRM tasks

- data collection
- predictive modeling
- Response analysis
- ..

- customer segmentation
- cross/up-selling analysis
- customer value analysis
- ..

- churn analysis
- modeling for recovery
- response analysis
- ..

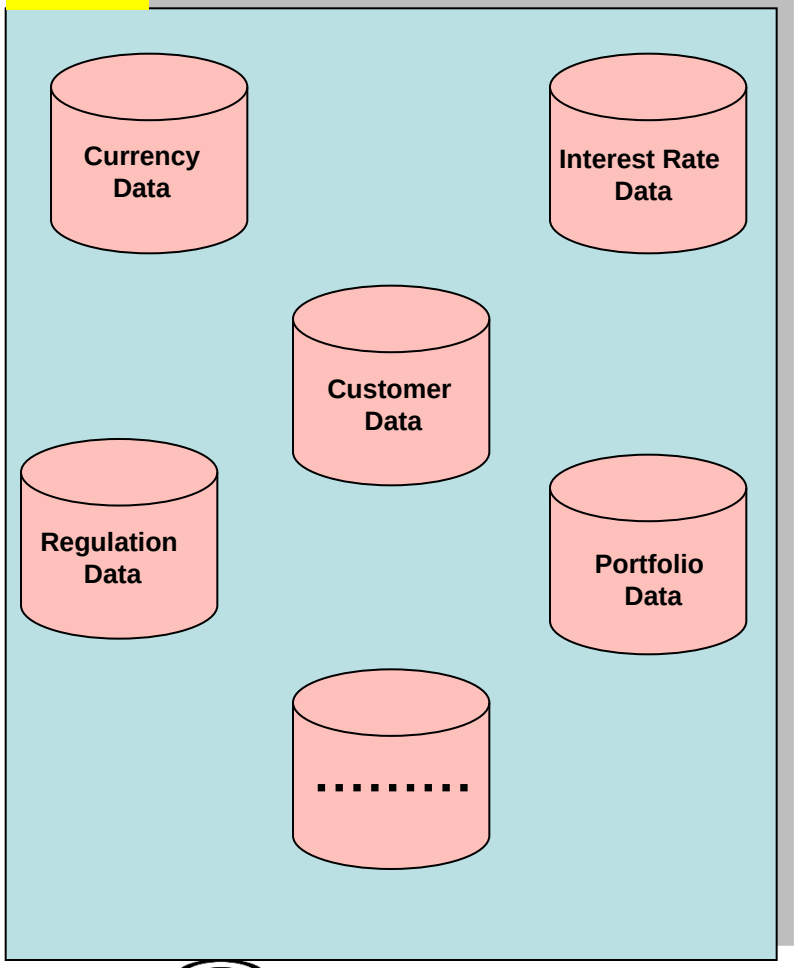
Source: http://www.prudsys.de/Service/Downloads/bin/DMC2003_arndt-daimlerchrysler.pdf

Why Data Mining in Banking ?

Business Issues

- Credit Risk
- Market Risk
- Controlling
- Trading
- Portfolio Manag.
- Investm. Manag.
- CRM
- Regulations& Compliance
- ….

Data



DM- Applications

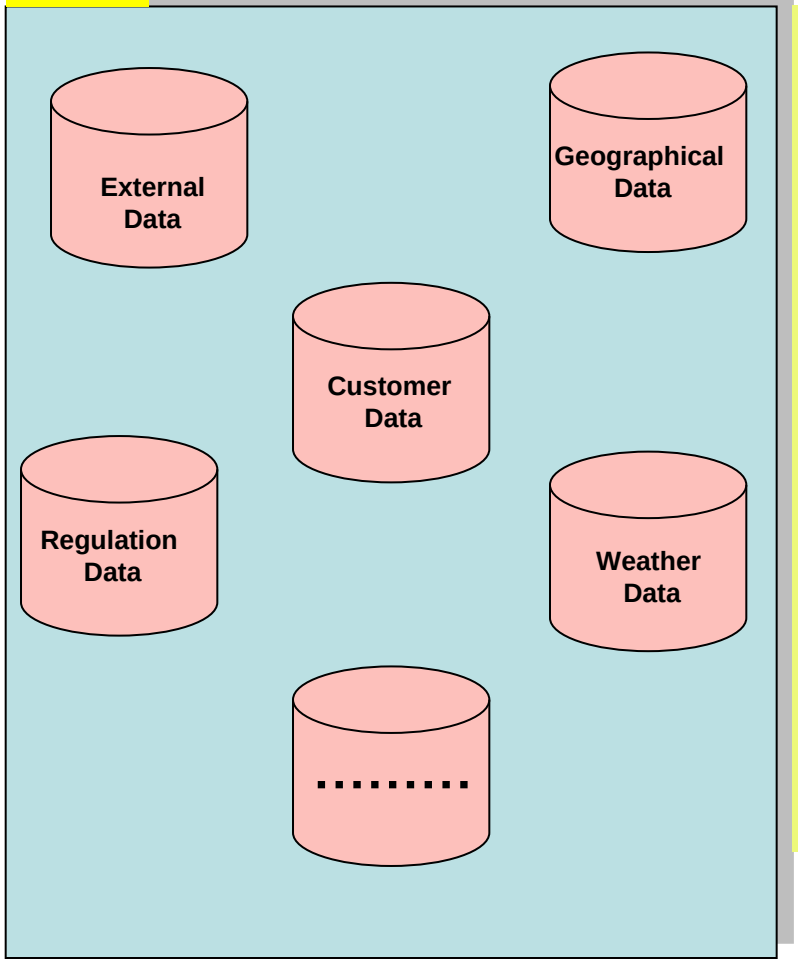
- Credit Scoring
- Market Forecasting
- Cross-Selling
- Up-Selling
- Churn Management
- Fraud Detection
- ……..

Why Data Mining in Insurance Industry ?

Business Issues

- Health
- Fire
- Building
- Transport
- Motor
- Holiday
- Accident
- Legal expenses
- Unemployment
- ….

Data



DM-Applications

- Rate Making
- Commission
- Cross-Selling
- Up-Selling
- Churn Management
- Reinsurance
- Claim Management
- Fraud Management
- ……..

Who produces and collects the Healthcare Data ?

- Hospitals
- Doctors‘ practices & Community health centers
- Pharmacy industry (B2B & B2C)
- Ambulant nursing care centers
- Nursing homes
- Health Insurance companies
- Personal m-Health sensors
- …….

Actions

What kind of actions ?

non-medical

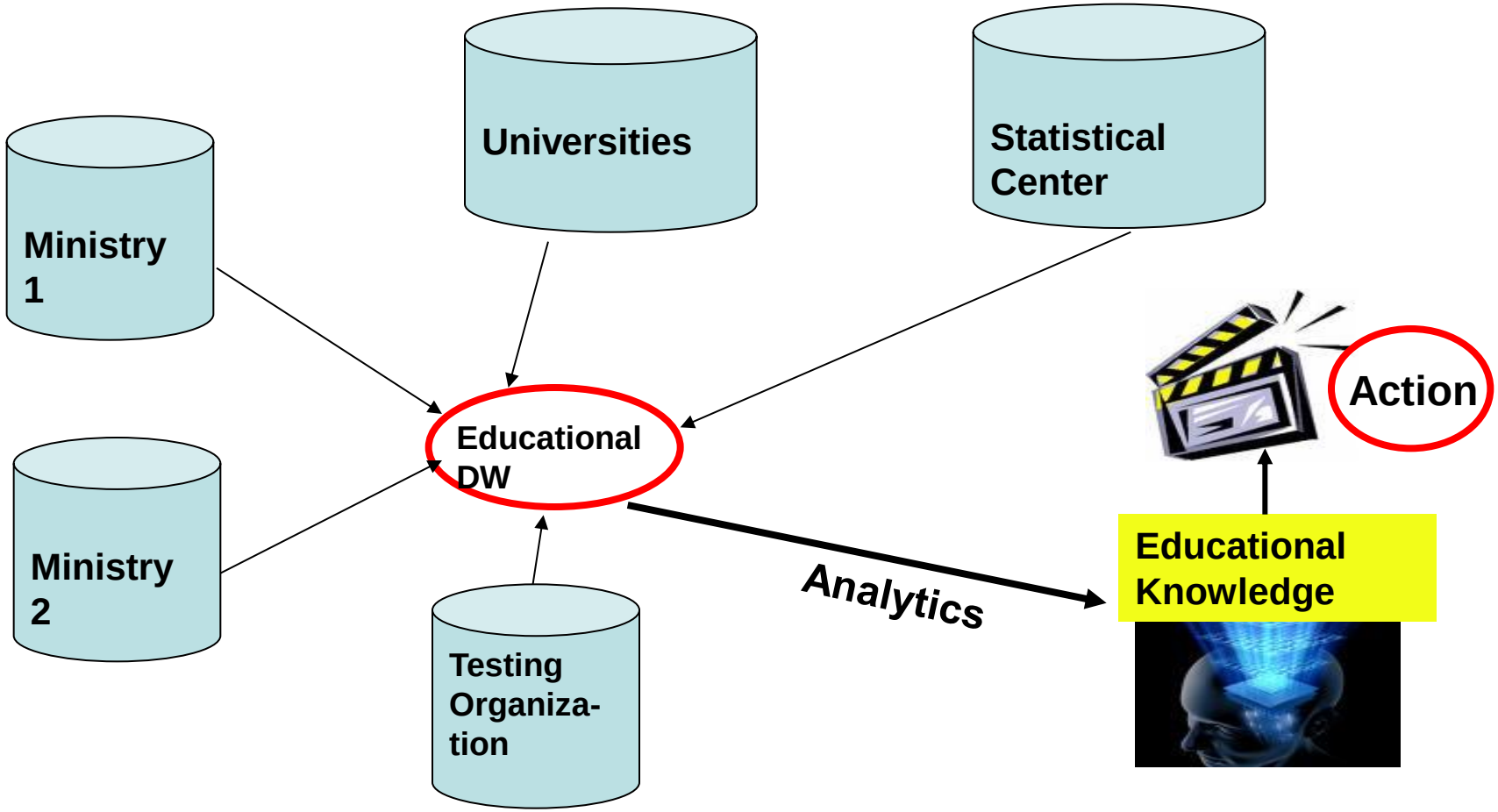
medical

Healthcare Knowledge

Analytics

Healthcare Data

Why Knowledge Discovery in Higher Education ?



Why Data Mining in your organization ?





Are the following issues important in your organization ?

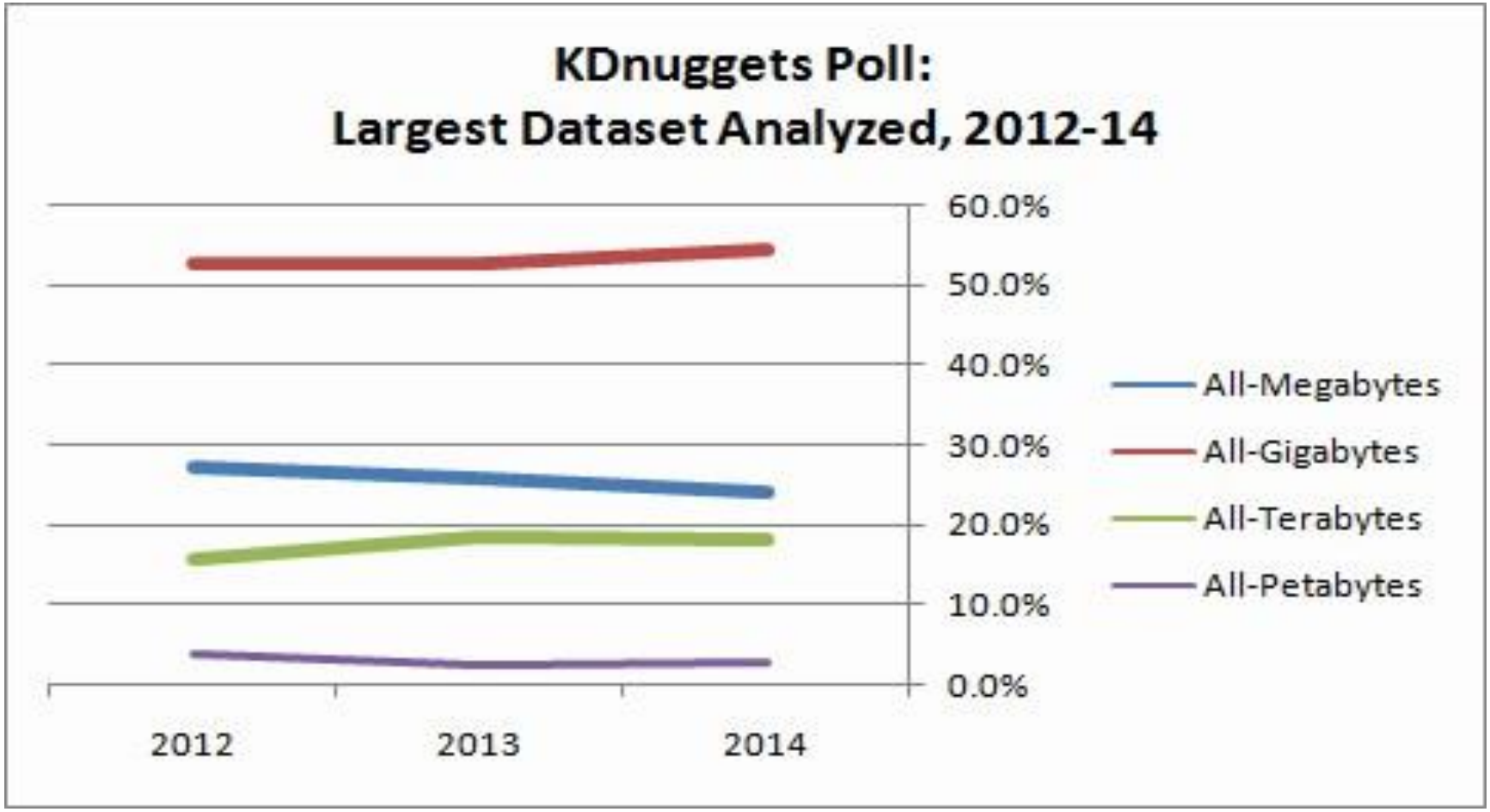
- **satisfaction of your customers**
- **quality of your products and services**
- **identification and optimization of different risks in your organization**
- **optimization of different processes (e. g. production process)**
- **finding of optimal location (e. g. plant location)**

- **optimization of supply chain management**

……

Data Mining can help

What was the largest database or dataset you data-mined ?



Original source: http://www.kdnuggets.com/2014/07/poll-results-largest-dataset-analyzed-surprisingly-stable.html

What is Data Mining ?

One of the most used definition (Fayyad et al 1996):

Knowledge Discovery in Databases (KDD) **is a process that aims at finding:**

- **valid,**
- **useful,**
- **novel and**
- **understandable**

patterns in data

Difference between Data Mining and KDD

KDD and Data Mining:

- KDD comes originally from AI
- Data Mining is a part of KDD
- In the praxis KDD and Data Mining are used as synonyms

Pattern types:

- Y= 2 + 3X (Generality)
- If country= Iran then carpet export= high (Locality)

Implicit and explicit patterns

Understandability

- **In many cases is very important**
- **It depends to DM – algorithm used**
- **Rule based algorithms → High understandability**
- **Artificial Neural Networks → Low understandability**

Remarks about the definition of Data Mining

About Models

- Explicit patterns (Association Rules, Decision Trees,..)
- Implicit patterns (Regression, Artificial Neural Networks,..)

About validity

- Rule validity (specially in Association Rules)
- Model validity (Classification , Prediction)
- Cluster validity (Clustering)

Simple fictive example: Data Mining Application
Claim Classification

	Claim Amount	Contract	Gender	Claim
Doctor 1	low	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	low	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	low	old	F	bad

Simple fictive example: Claim Classification

	Claim Amount	Contract	Gender	Claim
Doctor 1	low	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	low	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	low	old	F	bad

Simple fictive example: Claim Classification

	Claim Amount	Contract	Gender	Claim
Doctor 1	low	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	low	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	low	old	F	bad

Simple fictive example; Claim Classification

	Claim Amount	Contract	Gender	Claim
Doctor 1	old	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	old	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	old	old	F	bad

Simple fictive example; Claim Classification

	Claim Amount	Contract	Gender	Claim
Doctor 1	low	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	low	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	low	old	F	bad

Simple fictive example; Claim Classification

	Claim Amount	Contract	Gender	Claim
Doctor 1	low	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	low	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	low	old	F	bad

Simple fictive example: Claim Classification

Classifier

If Claim Amount= high Claim=good

If Claim Amount= low Claim=bad

If Claim Amount= middle & Contract=new Claim=good

If Claim Amount= middle & Contract=old Claim=bad

This classifier can be regarded as an Inductive expert systems

	Claim Amount	Contract	Gender	Claim
Doctor 1	low	new	F	bad
Doctor 2	middle	old	F	bad
Doctor 3	middle	new	M	good
Doctor 4	low	new	M	bad
Doctor 5	high	new	M	good
Doctor 6	high	new	F	good
Doctor 7	middle	new	F	good
Doctor 8	high	old	F	good
Doctor 9	middle	old	M	bad
Doctor 10	low	old	F	bad

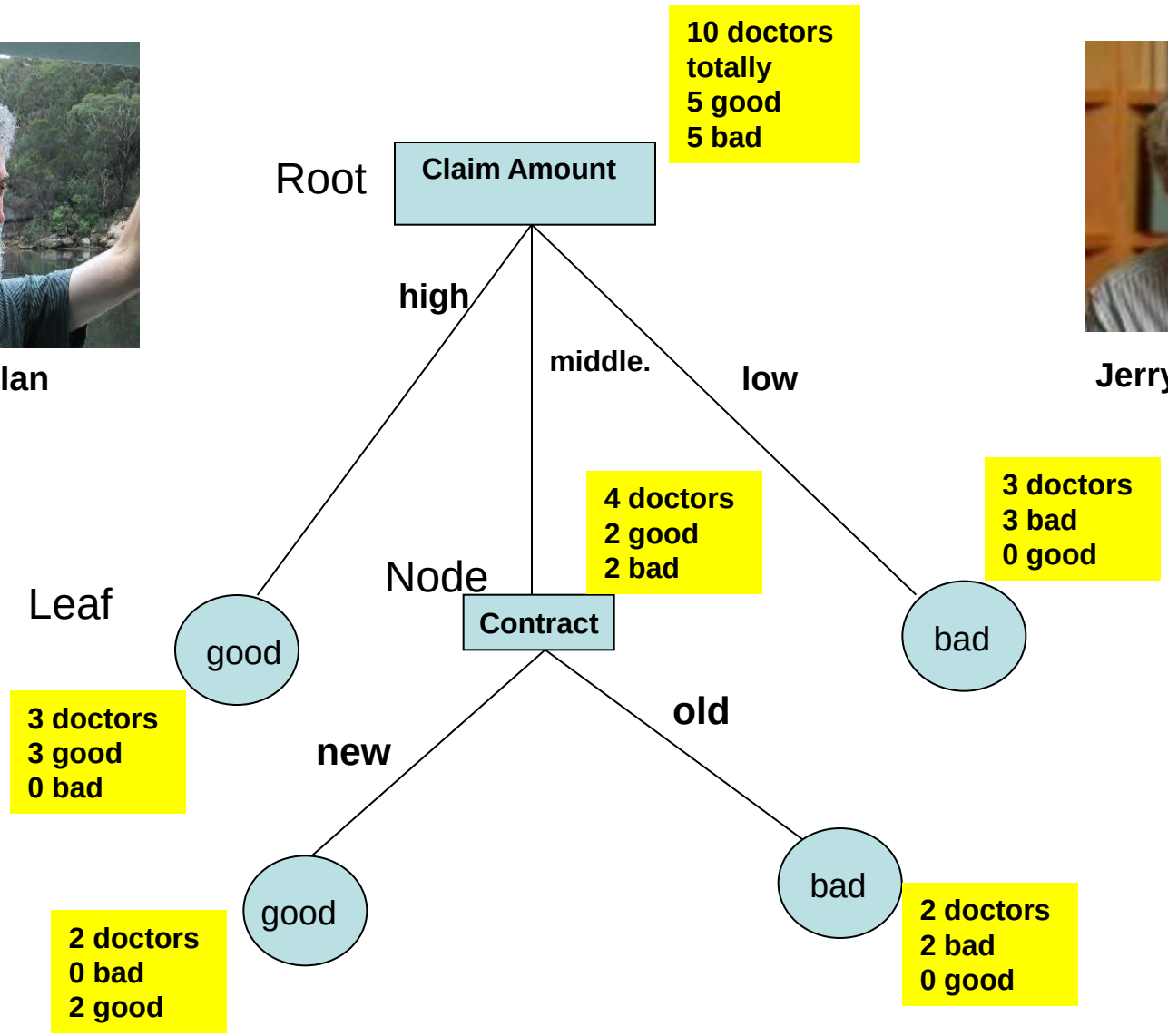
Classifying a new Doctors

- Claim a new Doctor with high Claim Amount = good
- Claim a new Doctor who has old Contract and middle Claim Amount = bad
- .....

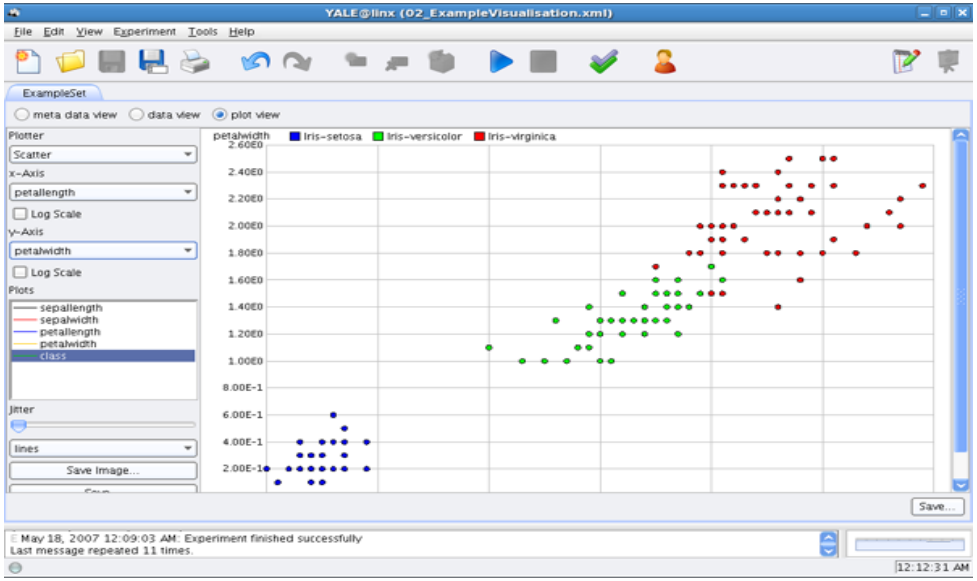
Claim Classification: Decision tree construction

Ross Quinlan

Jerry Friedman



Demo : Construction of a “Claim Miner”



Load Claim_Train, Claim_Predict

Interdisciplinary aspects of Data Mining

Privacy

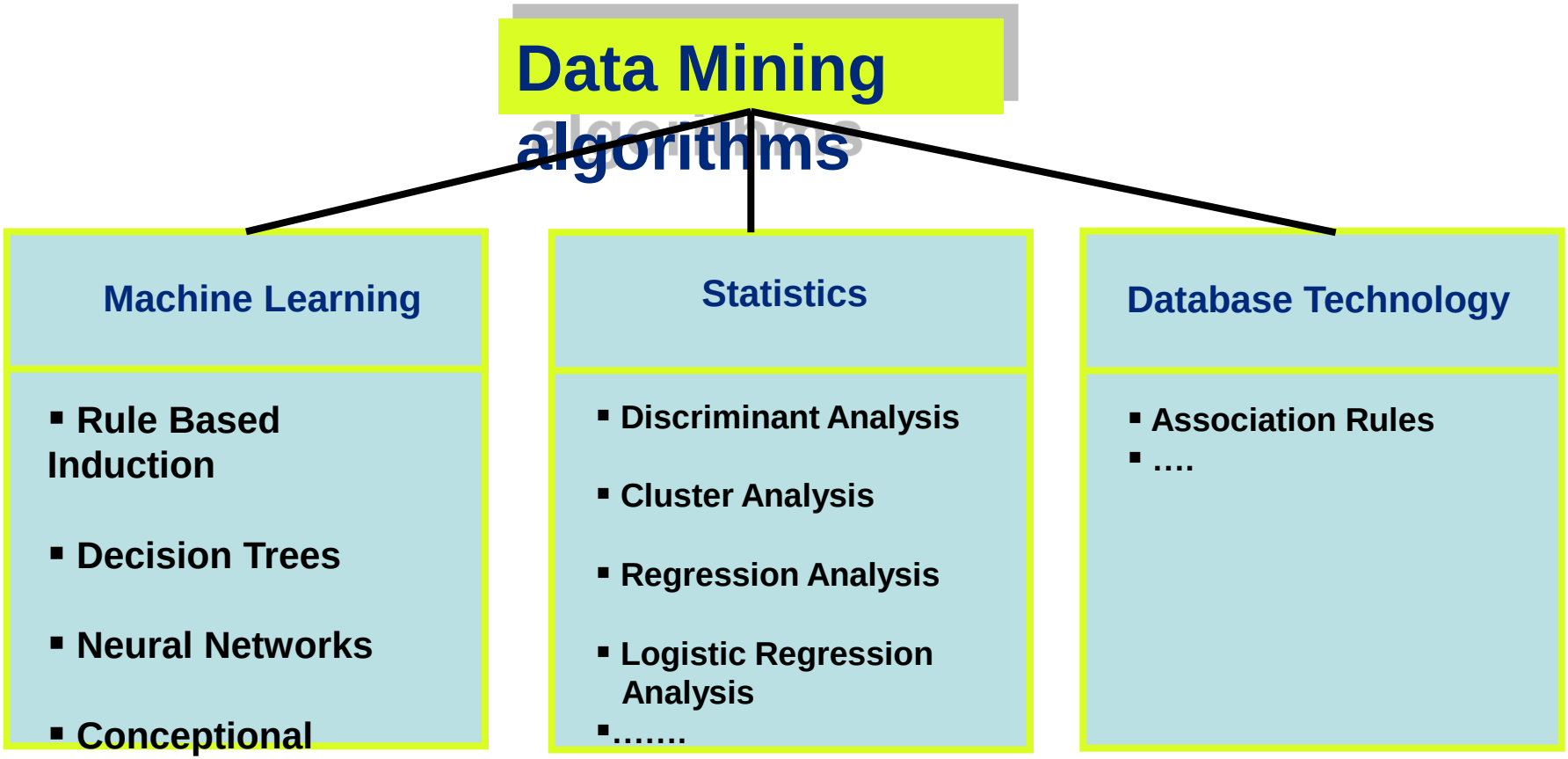
Visualization

Statistics (Biostatistics)

AI (Machine Learning)

Database Technology

Data Mining Algorithms



History of Data Mining: Data Mining rapid development

KDD-89: IJCAI-89 workshop on Knowledge Discovery in Databases
August 20, 1989, Detroit MI, USA

Dr. Gregory Piatetsky-Shapiro,

Google™ Deutschland

"داده کاوی"
About 508.000 Results
(0,36 Sekunden)

Results **about 28.700.000 for "data mining"**

KDD 2015

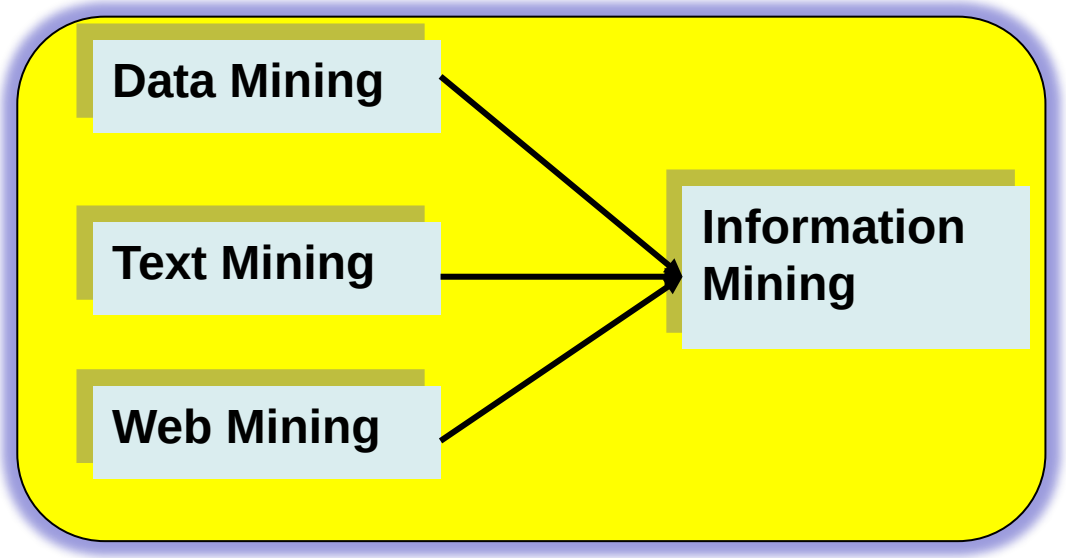
Sydney

KDD 2016

San Francisco

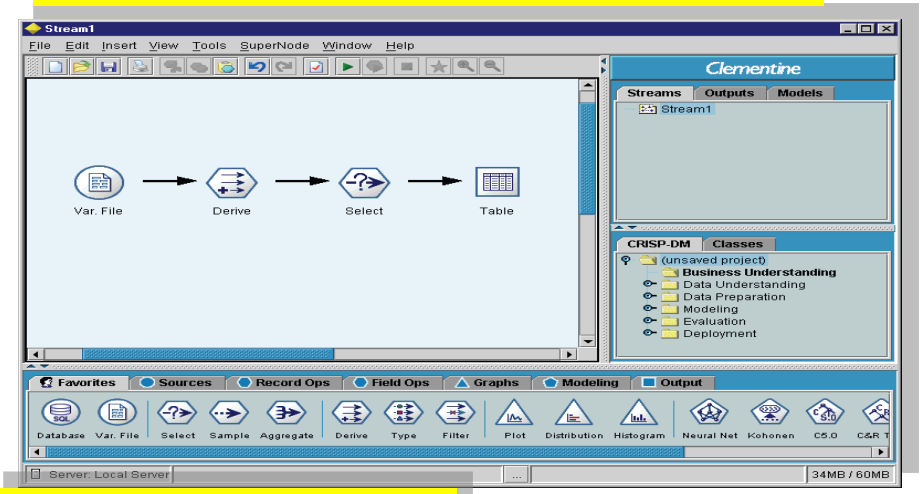
Two most famous children of Data Mining

- Application of Data Mining Methods to text and web driven data
  - Text Mining
  - Web Mining

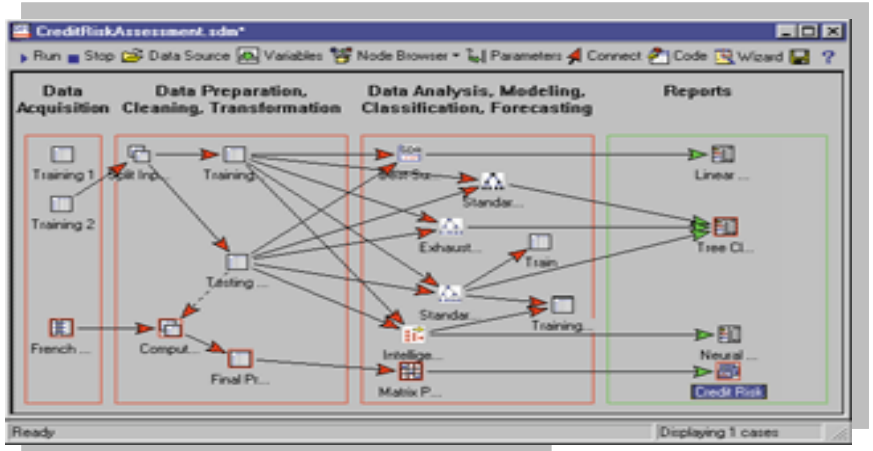


Examples of Data Mining Tools (commercial)

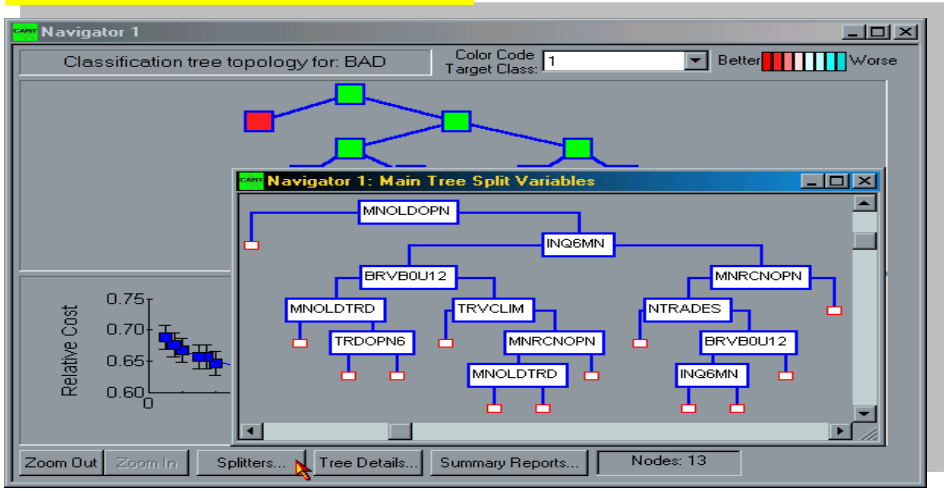
IBM SPSS Modeler (Clementine)



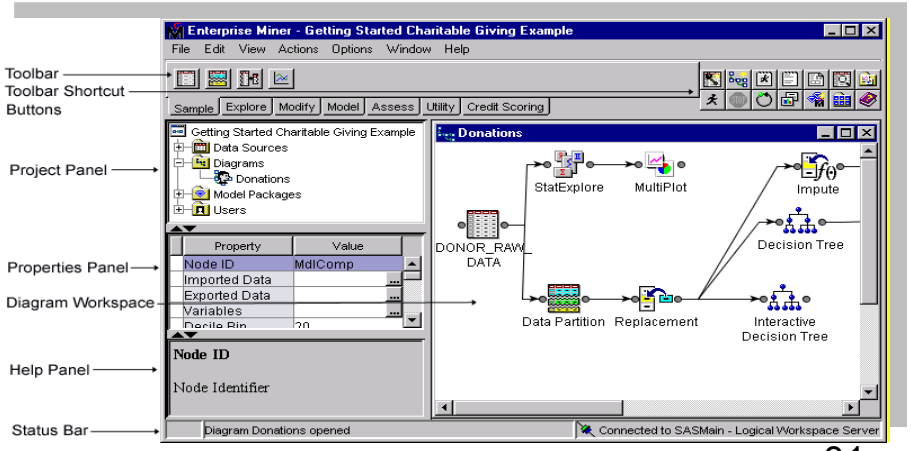
Statistica Data Miner



CART



SAS Enterprise Miner



Examples of other Data Mining Tools

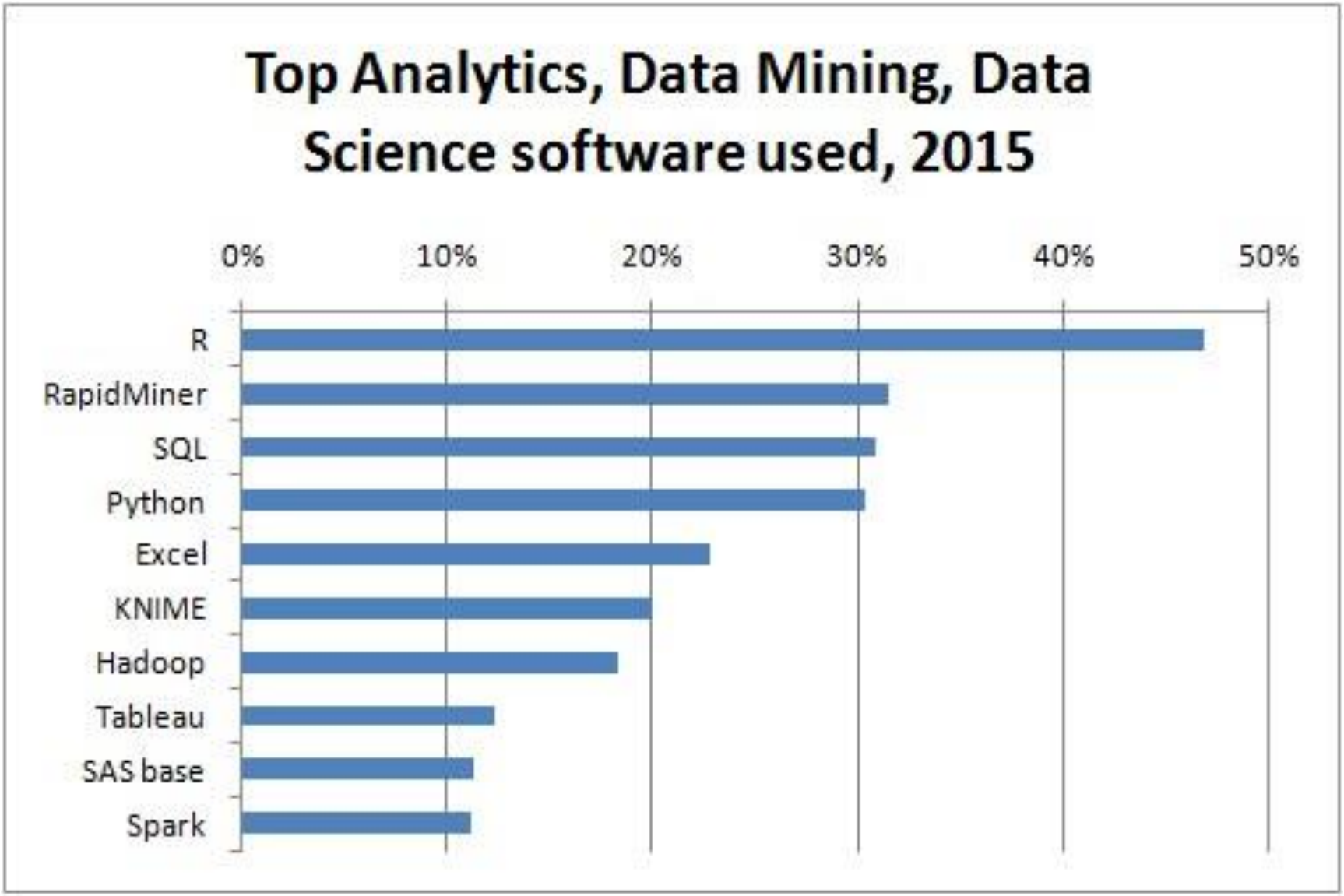
https://rapidminer.com/

https://www.r-project.org/

Department of Statistics

http://www.cs.waikato.ac.nz/ml/weka/

https://www.knime.org/



Tool evaluation

Gartner Magic Quadrant for Advanced Analytic Platforms, 2015

The agony of choice

How to survive in the jungle of Data Mining tools ?

- **Commercial tools ?**
- **non-commercial tools ?**
- **My own tool?**

Most important Selection criterion: NEEDS

- **Present needs**
- **Future needs**

Place of DM in the jungle of abbreviations

Related Technologies

BI, DW, OLAP, MIS, EIS, ERP, EA, …;
How much technology does a manager need?

Selection should be needs oriented

- **Present needs**
- **Future needs**

Related Topics

- **Predictive Modelling**
- **Predictive Analytics**
- **Business Analytics**
- **Service Analytics**
- **Health Analytics**
- **……**

Some European funded Projects

- StatLog
- CRISP-DM
- INRECA
- MetaL
- READ
- Data Mining Grid

Scientific Networking

1994-2001 European Network of Excellence in Machine Learning

2002-2005 European Network of Excellence in Knowledge Discovery

2005-2008 Ubiquitous Knowledge Discovery

Conferences

- KDD
- PKDD-ECML
- SIAM-Data Mining
- ICDM,
- PAKDD
- ICML
- ……

Journals

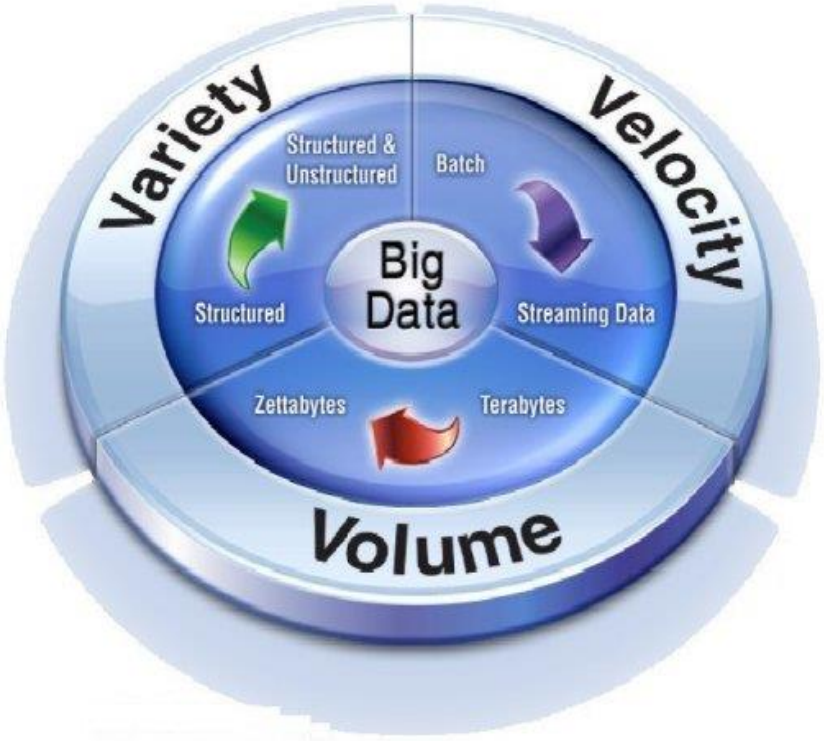
- ACM Transactions on KDD
- IEEE Transactions On Knowledge and Data Engineering
- KDD Explorations
- Data Mining and Knowledge Discovery
- Machine Learning
- …

Definition

Big Data is characterized by three Vs:

1. Volume (large amount of data)
2. Variety (diverse data format: images, texts, videos, audio..)
3. Velocity (high speed in generation)

Doug Laney (2001)



Source IBM

Other Examples

Boeing 747
generates in 10 domestic flights
up to 2.40 PB Data (Mckinsey)

IDC
Analyze the Future

the overall volume of data
generated in 2012
2.7 ZB (International Data Corporations)

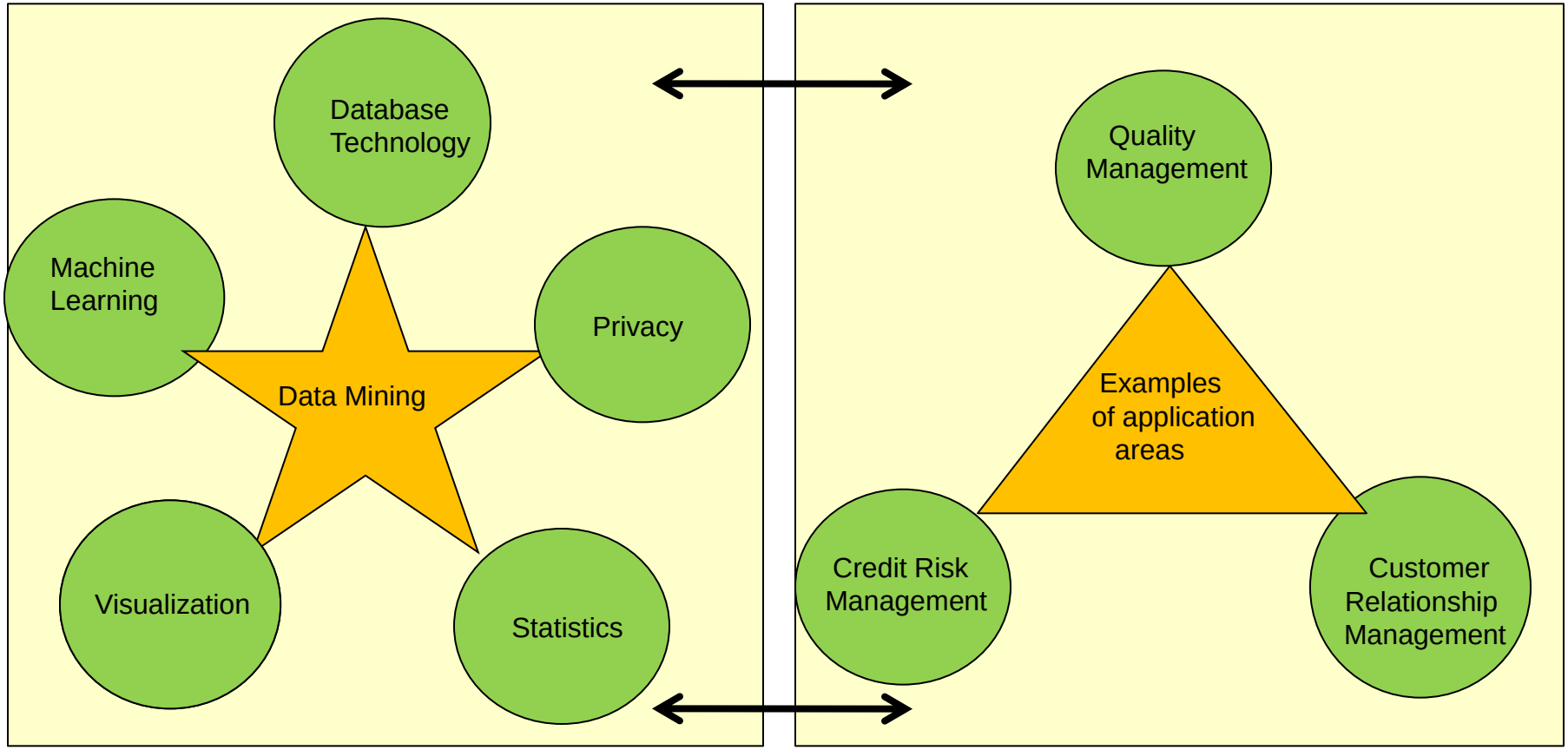
YouTube

In 2014, 300 hours of new videos were uploaded every minute

NSA Utha Data Center Stores ca. 29 PB per day

Big Data

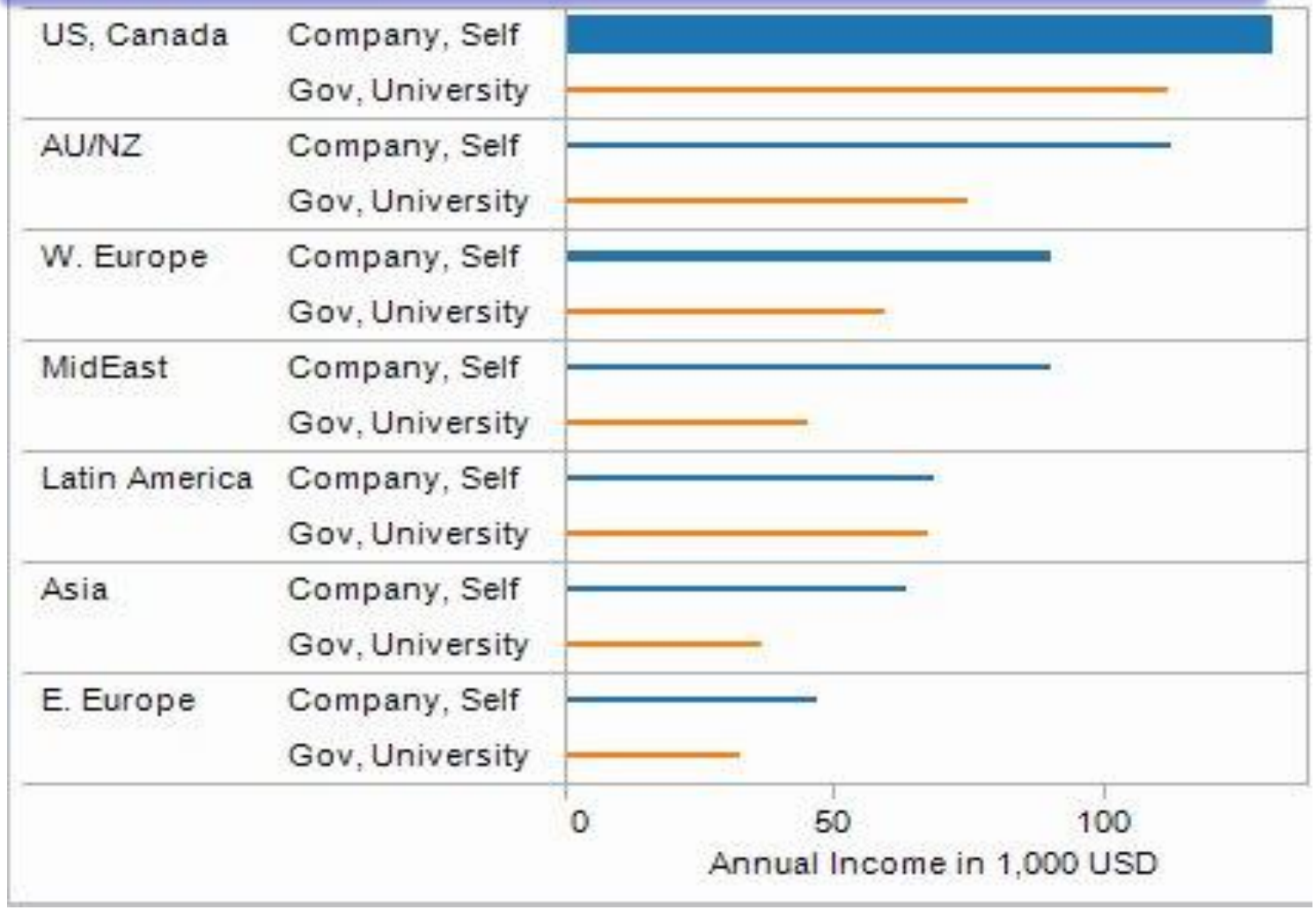
Optimal structure of a Data Mining Team



Success Parameters of Data Mining Solutions

- ☐ **Clear defined goals**
- ☐ **Importance of the business problem**
- ☐ **Management attention and support**
- ☐ **Competence of the Data Mining team**
- ☐ **Data availability and quality**
- ☐ **Close cooperation between the Data Mining team and the end-users**
- ☐ **Integration of the Data Mining Solution in the daily business process of the users**
- ☐ **Other parameters (Please describe briefly)**

Analytics Salary/Income by Region and Employment type excluding students and unemployed.



Sorce:The 2013 KDnuggets Annual Salary Poll