COMPLEMENTARITIES AND DIFFERENCES BETWEEN MACHINE LEARNING AND DATA MINING AND STATISTICS

IN ANALYTICS AND BIG DATA PART I + II

Petra Perner
Institute of Computer Vision and applied Computer Sciences, IBaI, Leipzig Germany

Invited Talk at ENBIS Spring Meeting, Barcelona, Spain, July 4-5, 2015

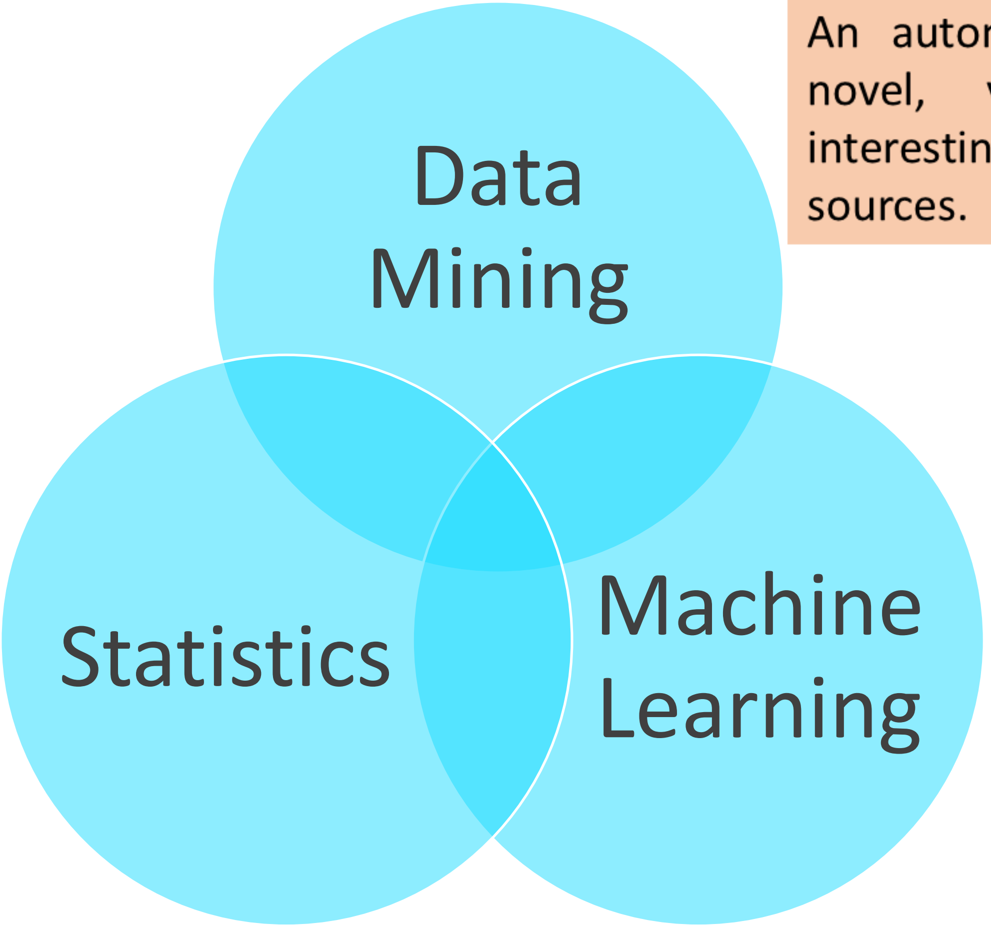
Invited Talk at the Intern. Conference of Knowledge Discovery and Data Analysis, KDDA 2015, November 15-17, 2015, Alger, Algeria

CONTENT

- Overview about Complementaries and Difference between ML, DM & Stat
- The Data Mining Aspects
- What does Big Data mean?
- The Characteristics of Big Data
- Cloud Computing
- Tasks of Cloud Computing
- What new Algorithm do we need?
- Conclusions

COMPLEMENTARIES AND DIFFERENCES

An automated process used to discover novel, valid, useful and potentially interesting knowledge from large data sources.



- Statistical analysis outputs: p-values, standard errors, regression models, principal components, discriminant score functions, ANOVA tables, control charts, descriptive statistics etc…

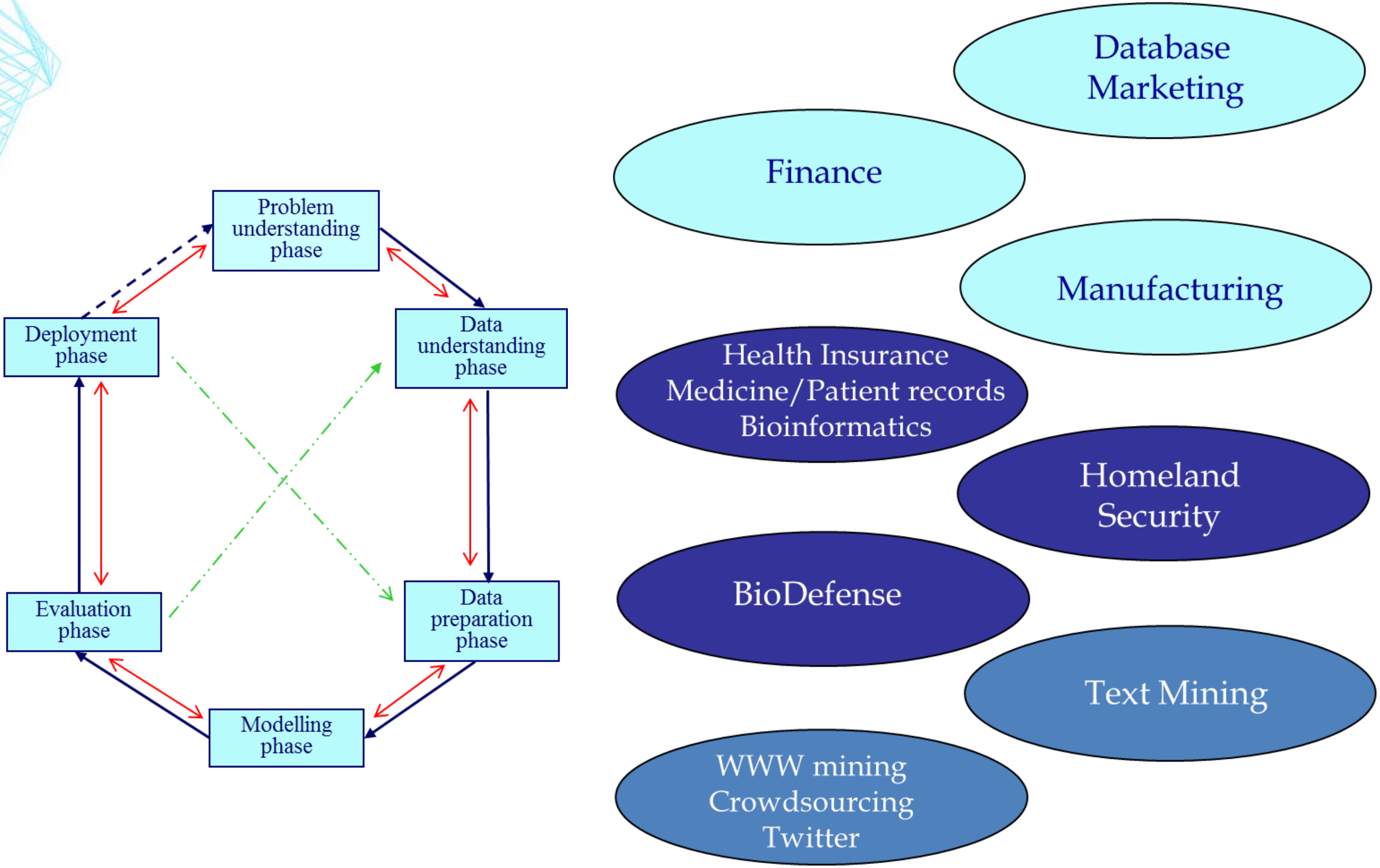
- translate statistical results into relevant information, careful formulation of findings is required

- Deals with representation and generalization
- Representation of data instances and functions evaluated on these instances
- Generalization is the property that the system will perform well on unseen instances

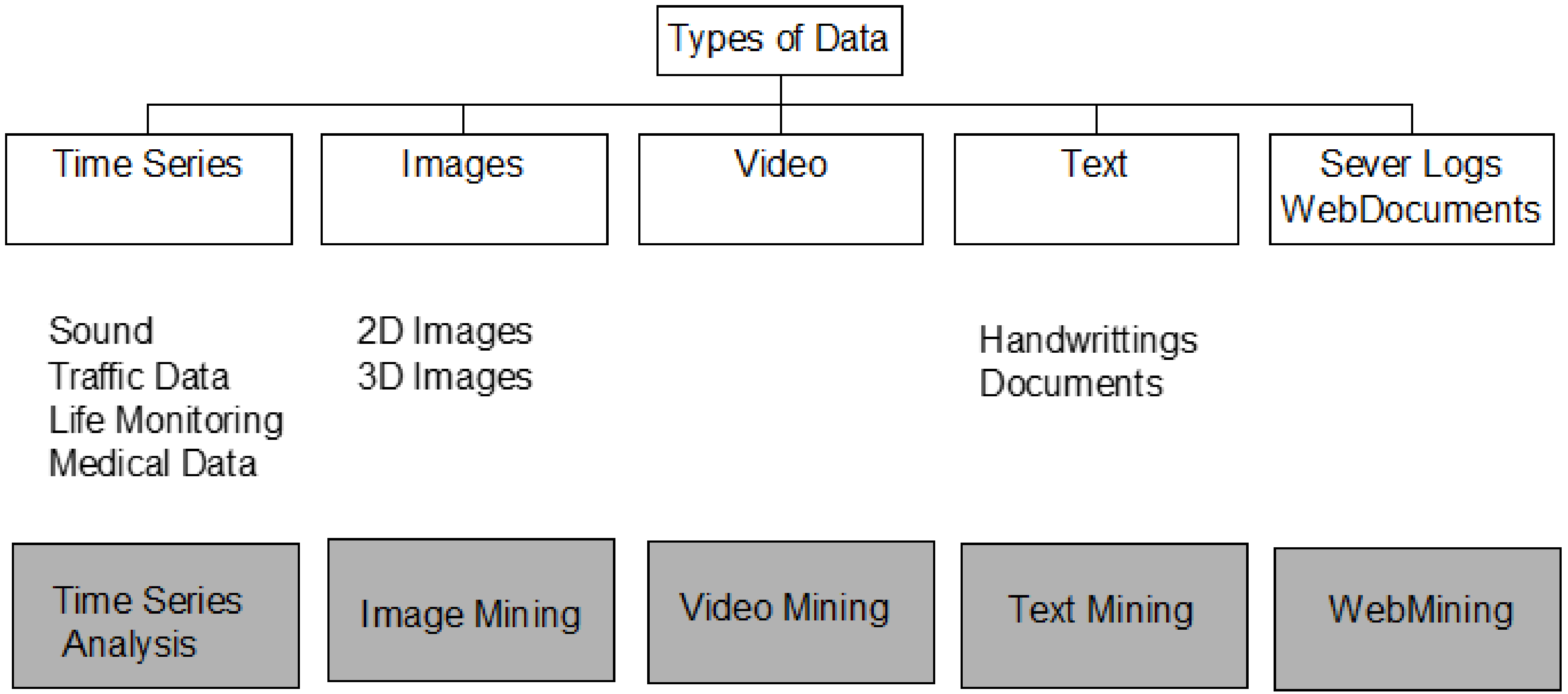
COMPLEMENTARIES AND DIFFERENCES

Type of Method	Statistics	Machine Learning	Data Mining
Descriptive Methods	Statistical Methods	Rule format based	Mixed Types between Stat&ML
	p-values	Decision trees	
	Standard errors		
Clustering	Partitioning Clustering	Conceptual Clustering	All Types
	Hierarchical Clustering	Rule-Based Clustering	
Classification	Discriminat function	Neural nets	Mixed types
	K-NN classifier	Rule-based classifier	
	CART decision tree	Case-based reasoning	
		Decision tree induction	
Regression	Regression Methods	Regression Methods	Regression Methods
Association Rules		Association rule methods	Association rule methods
Visualizations	Dendrogram	Tree Representation Visualization	Tree Representation Visualization

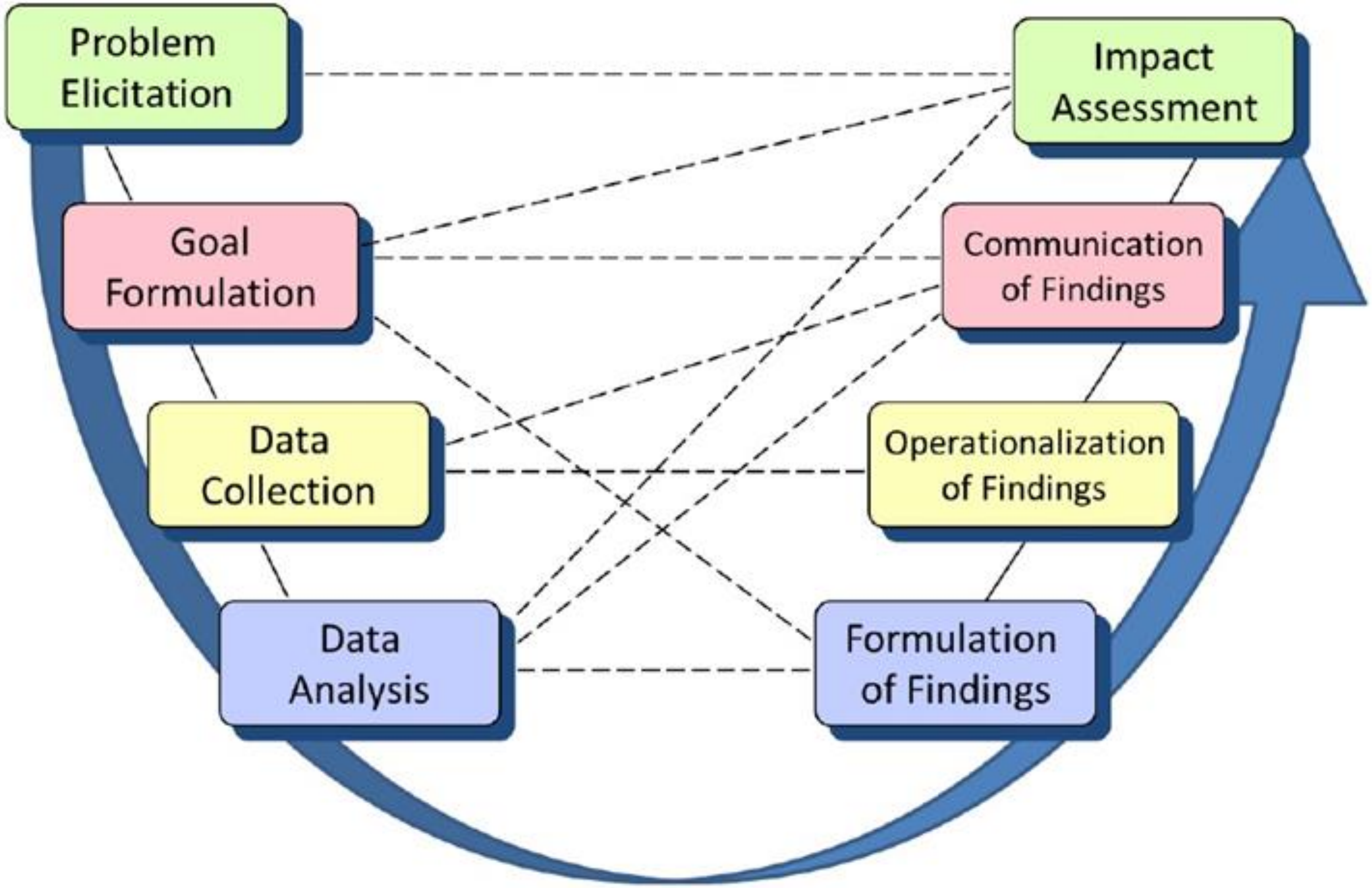
THE DATA MINING CYCLE AND SOME TYPICAL APPLICATION DOMAINS



DATA MINING ON MULTIMEDIA DATA



THE LIVE CYCLE VIEW OF STATISTICS

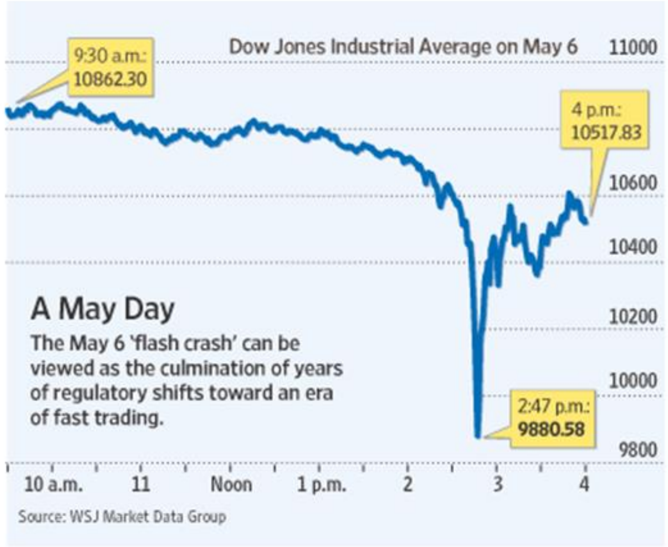


EARLY MOTIVATING APPLICATIONS OF DATA MINING

- Database Marketing
- Trading at Stock Market
- Market Basket Analysis

www.information-drivers.com/market_basket_analysis.php

Better Real-time Behavioral Targeting for Online Advertising



HOW IS DATA MINING DIFFERENT?

- Different from what? (statistics, ML)
- Data can be unstructured (e.g., text) and of different types (e.g. multimedia)
- Data can come from different sources and have conflicts/missing data/outliers
- Usually a data information step is required
- Data pre-processing requires most of the time
- Data might not fit into memory
- Not such a strong requirement to the accuracy of the model

DATA MINING CHALLENGES

- **Large/huge data sets**
  - Data sets can be rich in the number of data
  - Data sets can be rich in the number of attributes
  - Unlabeled data (data labeling might be expensive)
  - Data quality and data uncertainty

- **Data preprocessing and feature definition for structuring data**
  - Data representation
  - Attribute/Feature selection
  - Transforms and scaling

- **Scientific data mining**
  - Classification, multiple classes, regression
  - Continuous, binary, and mixed type attributes
  - Large data sets
  - Nonlinear problems

- **Erroneous data, outliers, novelty, and rare events**
  - Erroneous & conflicting data
  - Outliers
  - Rare events
  - Novelty detection

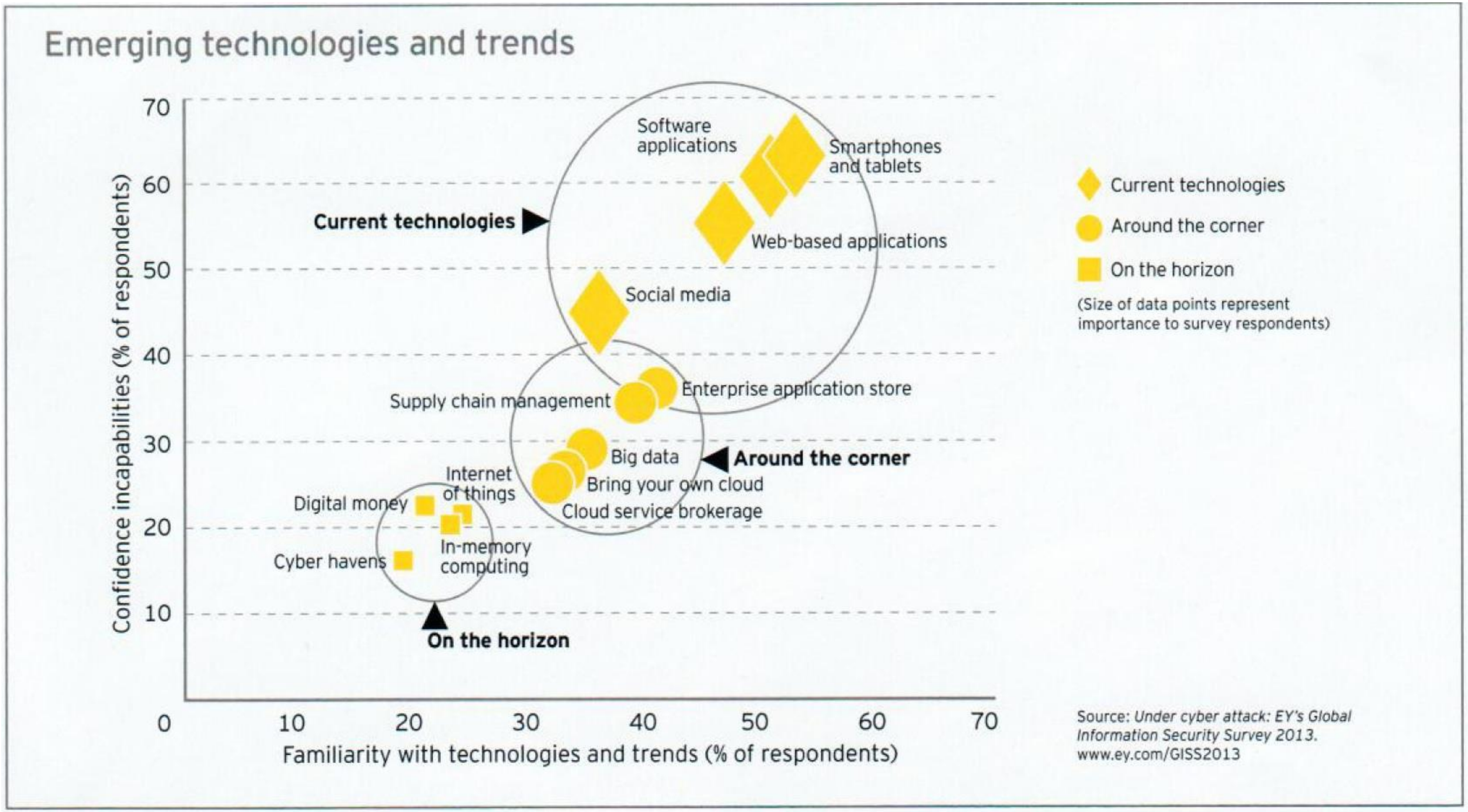
- **Smart visualization techniques**

- **Feature Selection & Rule formulation**

- **Special outcomes: Associations (e.g. NETFLIX), nuggets, enrichment**

- **Recent challenges: causality, active learning, multi-classes, big data**

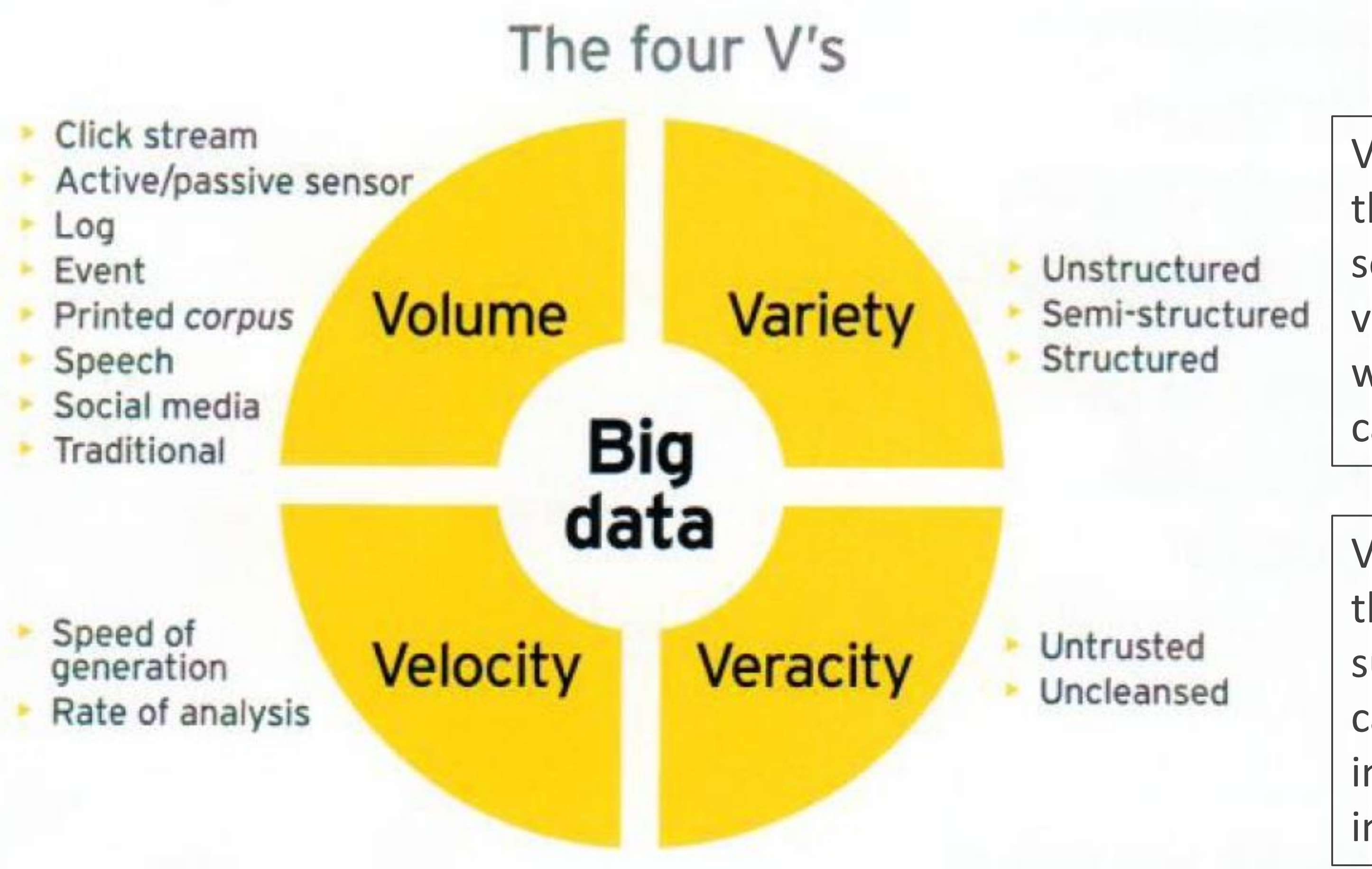
EMERGY TECHNOLOGIES AND TENDS



BIG DATA AND THEIR REQUIREMENTS

Volume represents the actual amount of data available

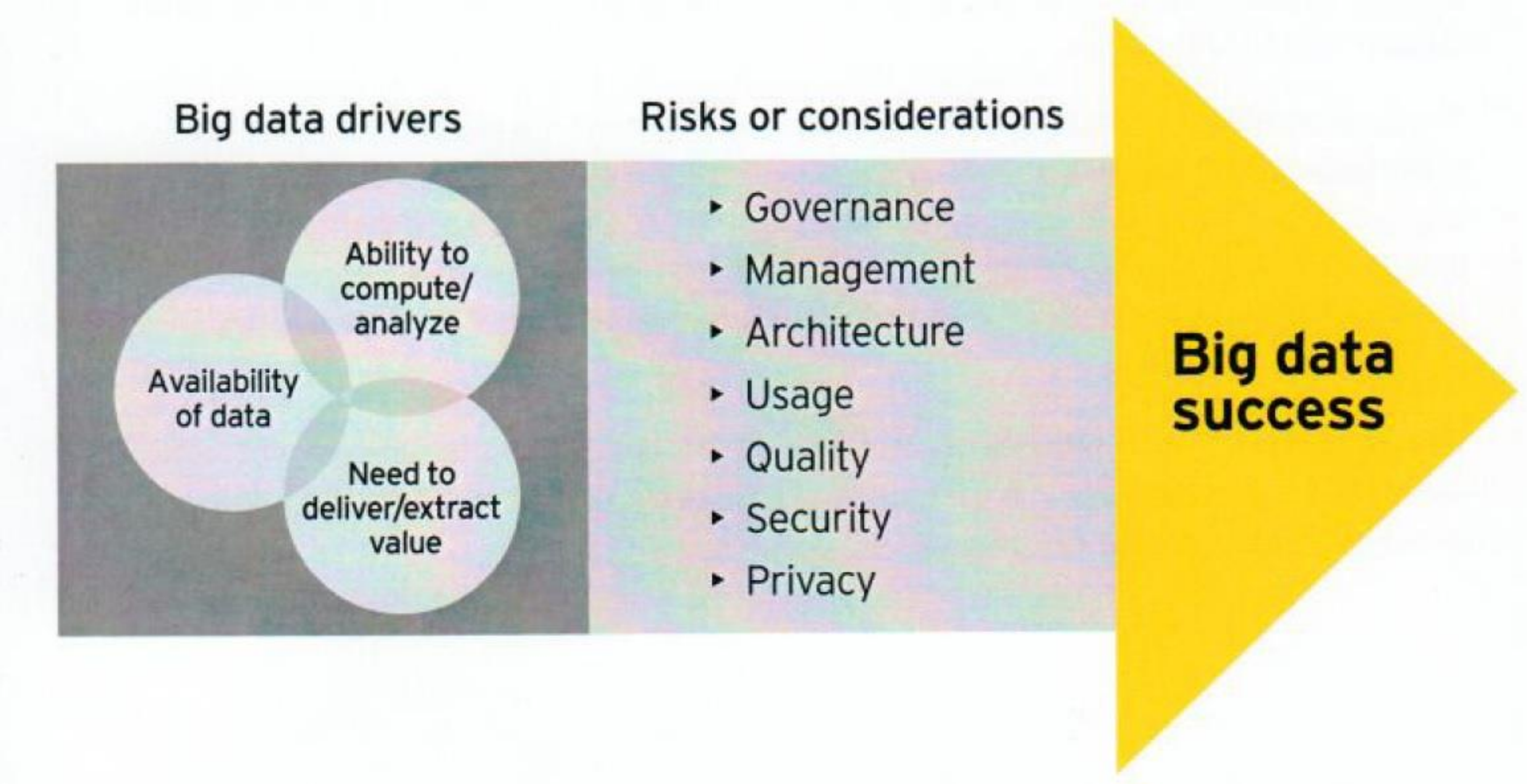
Velocity is the speed at which data is being created and how fast it must be processed to meet business needs.



Variety refers to the multiple data sources and varied formats in which the data can be presented.

Veracity denotes the uncertainty surrounding data caused by inconsistency and incompleteness.

BIG DATA SUCCESS

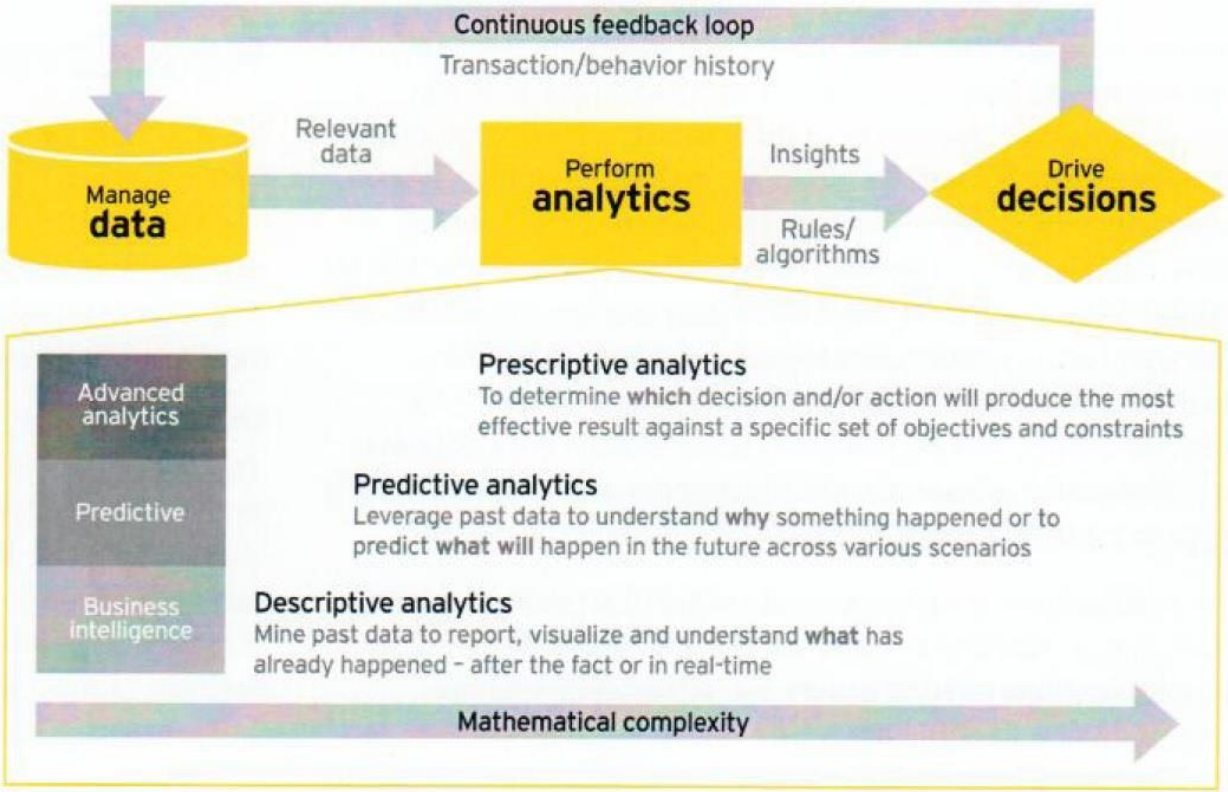


ANALYTICS VALUE CHAIN

EY analytics value chain

The goal is to use analytics to improve the **efficiency** and **effectiveness** of every **decision** and/or **action**.

1. Begin with leveraging leading tools and techniques to manage and extract relevant data from big data sources.
2. Applications of analytics can range from historical reporting, through to real-time decision support for organizations based on future predictions.
3. Use the insight generated by the analysis to drive change.



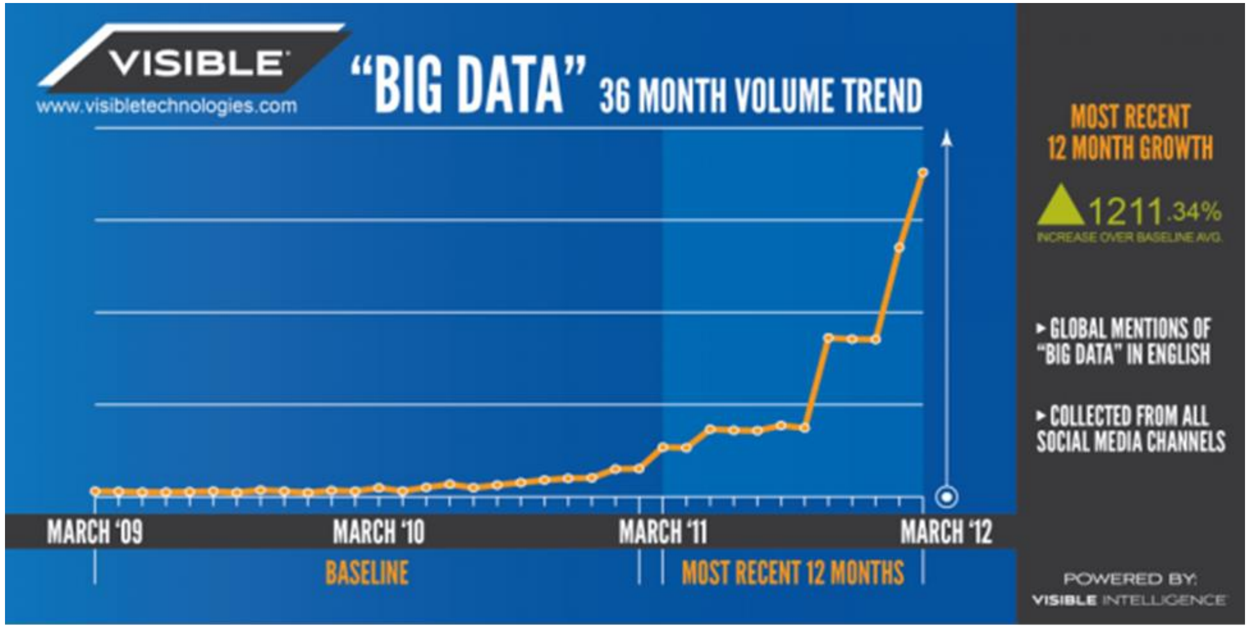
TRENDS IN BIG DATA

Lagging	Industry trends	Leading
	Data	
Functional silos	Users	Spans enterprise
Internal	Sources	Internal/external
Multiple	Technology/repository	One trusted
	Analytics	
Old patterns	Purpose	Exploratory
Several	Variables	Many
Historic trends	Discovery	Predictive
	Decisions	
"Gut feel"	Based on	Data
Tenure	Business case development	Data
Perception	Road map prioritization	Data

DATA MINING AND BIG DATA ANALYTICS

Mystery of Big Data's Parallel Universe Brings Fear, and a Thrill

A HISTORY OF BIG DATA



http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/

- 4-19-2010 - Danah Boyd, "Privacy and Publicity in the context of Big Data." Keynote WWW2010
  http://www.danah.org/papers/talks/2010/WWW2010.html
- May 2011 - James Manyika et al. "Big Data: The next frontier for innovation, competition, and productivity." (McKinsey Global Institute Report).
  http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
- 6-6-2011 - A very short history of Big Data. http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/
- March 27, 2012 – WIRED Magazine: Webcast: Obama goes big on big data.
  http://www.wired.com/cloudline/2012/03/obama-big-data/

HOW IS BIG DATA ANALYTICS DIFFERENT?

- Data is unstructured
- Data comes from different sources and has conflicts/missing data/outliers
- Usually a data fusion step is required
- Data are dynamic
- Often has a crowdsourcing component (e.g. Twitter)
- Often sensor processing steps are required (domain specific)
- Because of the size of processed data things have to be done differently

→ Involves high-performance computing and specialized algorithms

Big data analytics is the process of the automated discovery of potentially actionable/auctionable knowledge from diverse large data sources, where some of these data sources often have a crowdsourcing aspect

ORDERS OF MAGNITUDE OF DATA (SOURCE: WIKIPEDIA)

Multiples of bytes				
SI decimal prefixes		Binary usage	IEC binary prefixes	
Name (Symbol)	Value		Name (Symbol)	Value
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}
See also: Multiples of bits · Orders of magnitude of data				

V · T · E

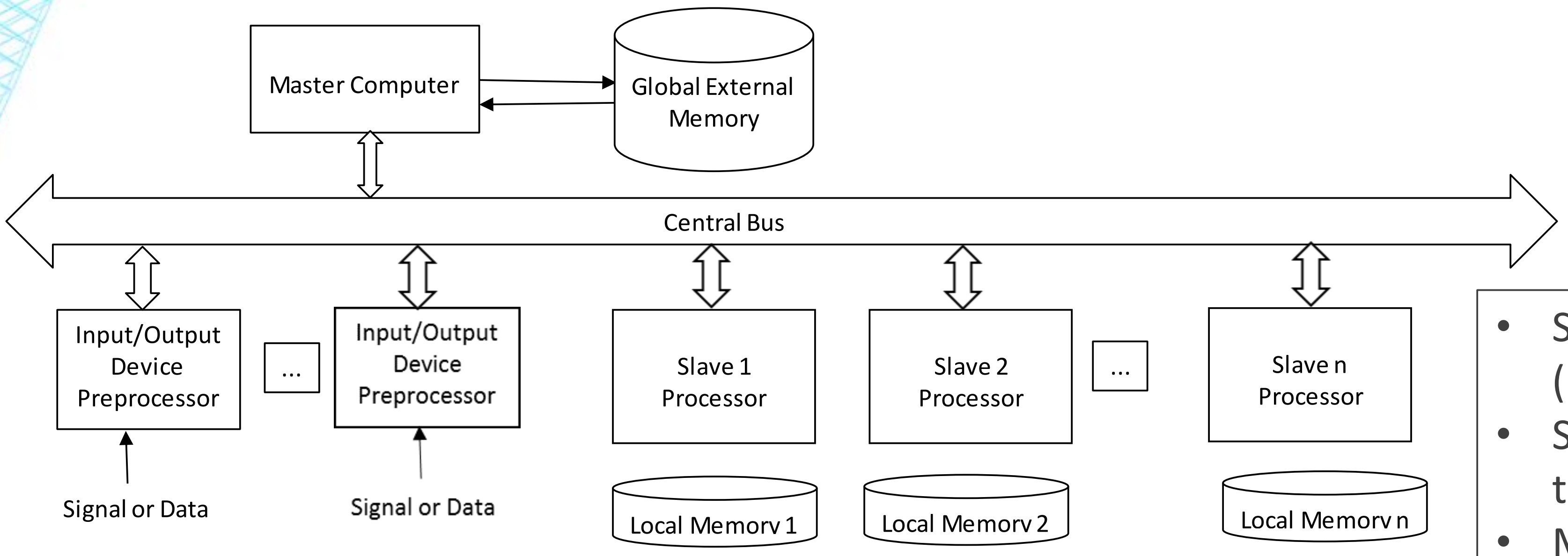
Typical for Big Data

Capacity human memory

Size internet archive 2004
....
All Climate data
Entire Library of congress
Google server farm 2004

1 petabyte = 1000 terabytes = 1000 x

CLOUD COMPUTING (MULTIPROCESSOR COMPUTING AND PARALLEL PROCESSING)



- Scaleable of the Memory (horizontal or vertical)
- Scaleable or Parallelization of the algorithm
- Memory Organization
- Incremental Data Access and Computing

- Scheduling Problem, Protocolls
- Input/Output Problems
- Signal Preprocessing on SP

MEMORY ORGANIZATION

Scaleable Vertical

Data-1	Data_2	…	…	…	Data-n
Data-11	Data_12	…	…	…	Data-1n
…					
…					
…					
…					
…					
Data-n1	Data-n2	…	…	…	Data-nn

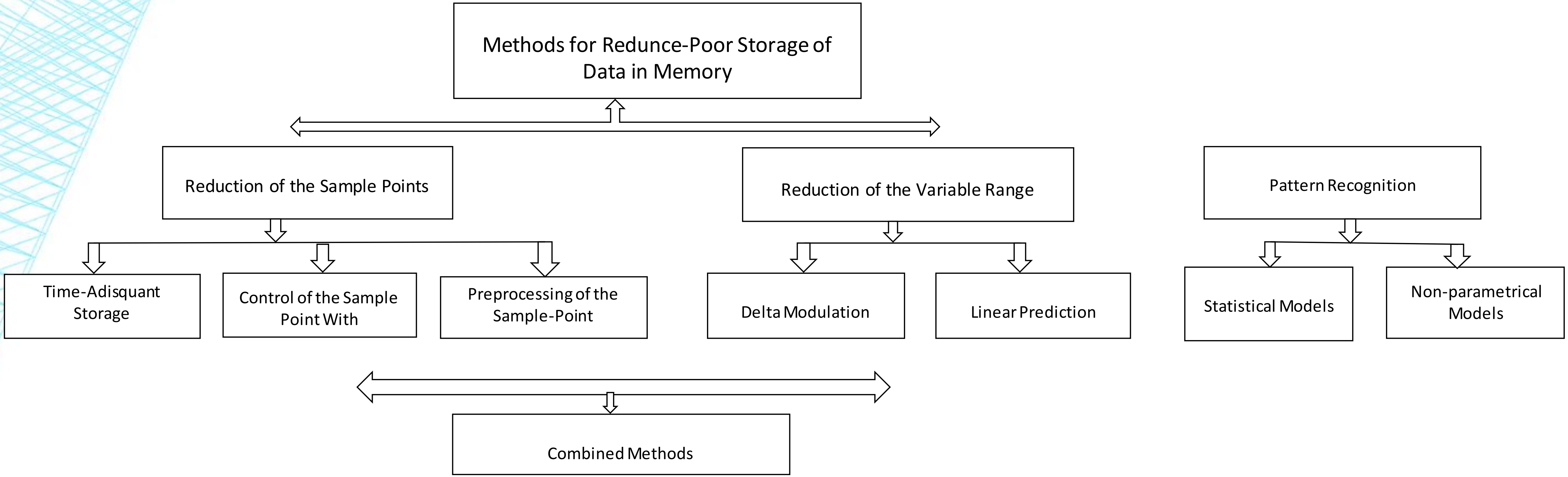
Scaleable Horizontal

Data-1	Data_2	…	…	…	Data-n
Data-11	Data_12	…	…	…	Data-1n
…					
…					
…					
…					
…					
…					
Data-n1	Data-n2	…	…	…	Data-nn

The mode of Scaling has influence to the processing schema!

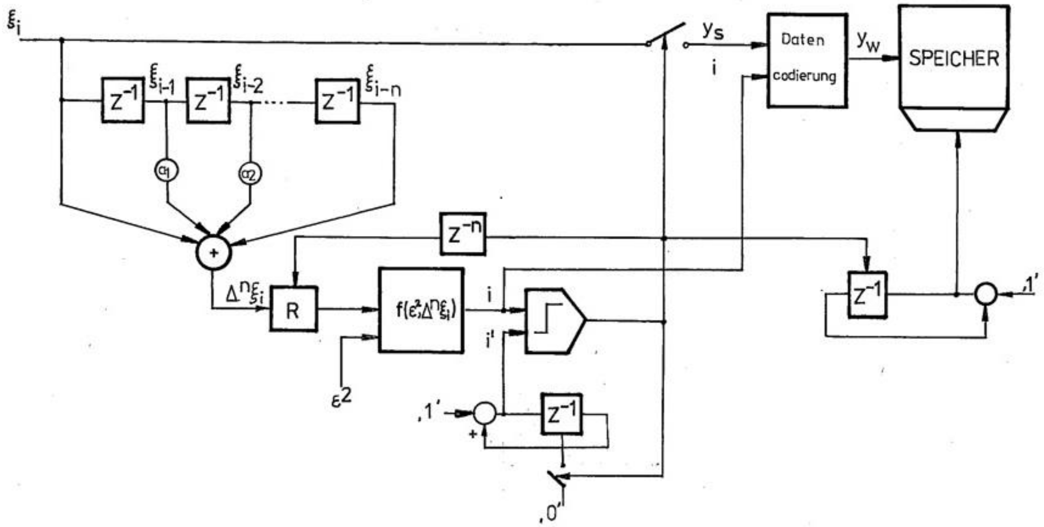
Suppose we want to procress the Distribution Function over a variable Then it is preferable to split the Memory in such a way that one Processor can calculate the the function over the variable.

MEMORY ORGANIZATION (REDUNDANCY-POOR STORAGE)

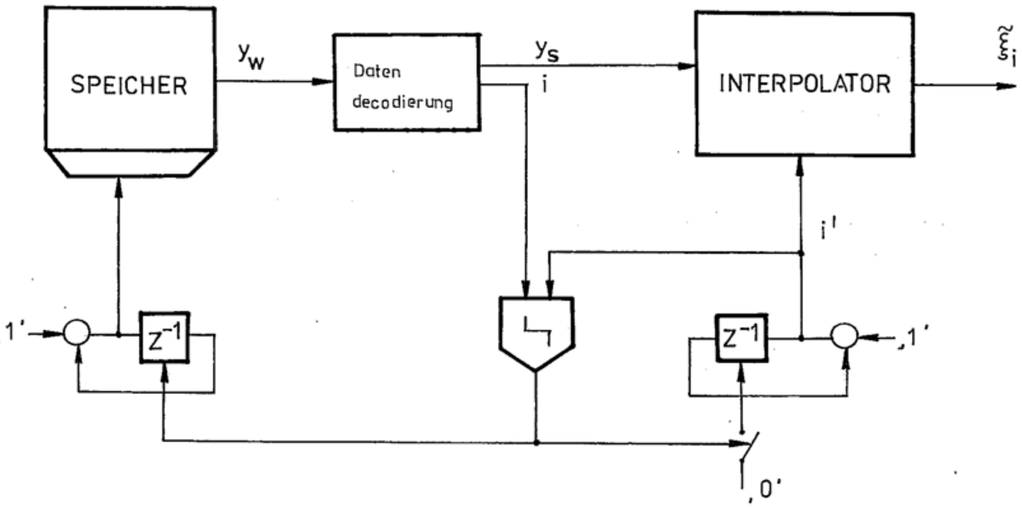


- Model-based organization for Time-Series Data
  - The time series are modelled by Functions and only the coefficient are stored
  - Summarize the data into a statistical model

INTERPOLATION AS DATA REDUCTION



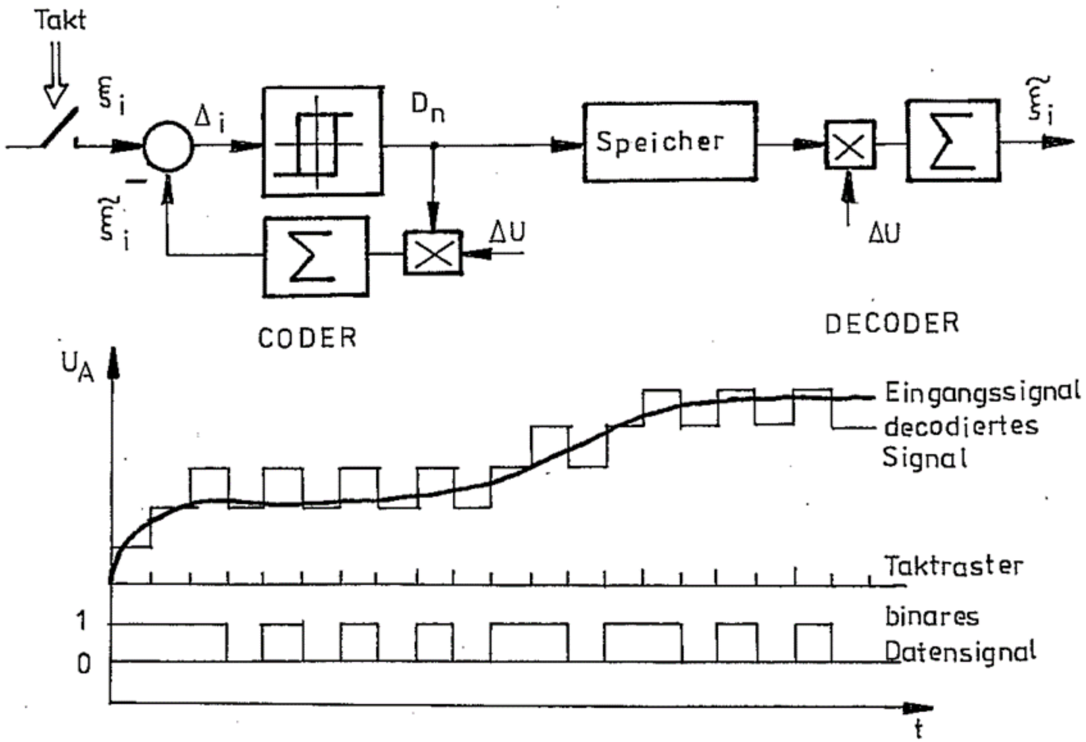
Data Coder



Data Decoder

Special Indexing Meachanism

DELTA MODULATION AS CODER



Memory Organization

		Memory
Data Stream 1	1	1
…		0
…		1
Data Stream 4	4	1
…		1
…		0
…		0
Data Stream 8	8	1
	8 Bits	

MEMORY ORGANIZATION

- The kind of memory organization depends on the type of data.
- Time-series data can be handeled differently than image data.
- Statistical methods can help to summarize the data so that memory capacity is saved!
- This models can be incrementally updated based on MML.

SPECIFIC ALGORITHMS FOR BIG DATA ANALYTICS

Modified algorithms

- Modified algorithms for regression/classification models for parallel computation
- Incremental methods for decision tree learning and other models
- Labeling of unlabeled data, oracle-based methods
- Semi-supervised learning and active learning

New Algorithms

- Outlier detection
- Data cleansing with similarity-based algorithm and statistical methods that can be parallelized
- Data completion algorithms for partially filled data e.g. similarity-based methods
- Data anonymization algorithms

SPECIFIC ALGORITHMS FOR BIG DATA ANALYTICS

- Case-based reasoning algorithm are of specific attraction. The memory can be easily separate horizontally and each processor can handle the data without interaction to any other slave. The master can process the final results.
- CBR can handle incomplete data and can do incremental learning.
- CBR has also to do with indexing and case storage. The CBR methods need to be adapted or further developed for cloud computing.

SPECIFIC ALGORITHMS FOR BIG DATA ANALYTICS

Functions that are separateable and allow calculation of functions in only one direction of the separation and at the end the final result get summarized by all other separate results.

(e.g. filters, descriptive statistics, statistical models that are not based on mixtures or the mixtures can be calculated separately.

CONCLUSION

- Big Data - Mean a Huge Amount of Data is to be processed
- It does not only have to with new methods for data analytics.
- It has to do with Cloud Computing and Parallel Computing.
- Memory Organization is a big Challenge.
- How to keep the data? – As raw data or summarized in a model?
- What algorithm are scaleable?
- Incremental Learning is necessary since processing of big data takes time. We need incremental learning model for all kind of algortihms nontheless if they come from statistics or machine learning.
- Separateable function are welcome.

BIG DATA ANALYTICS: HYPE OR HALELUJA?

Petra Perner

Institute of Computer Vision and applied Computer Sciences, IBaI, Leipzig, Germany

www.ibai-institut.de

pperner@ibai-institut.de

Thank you for your attention!