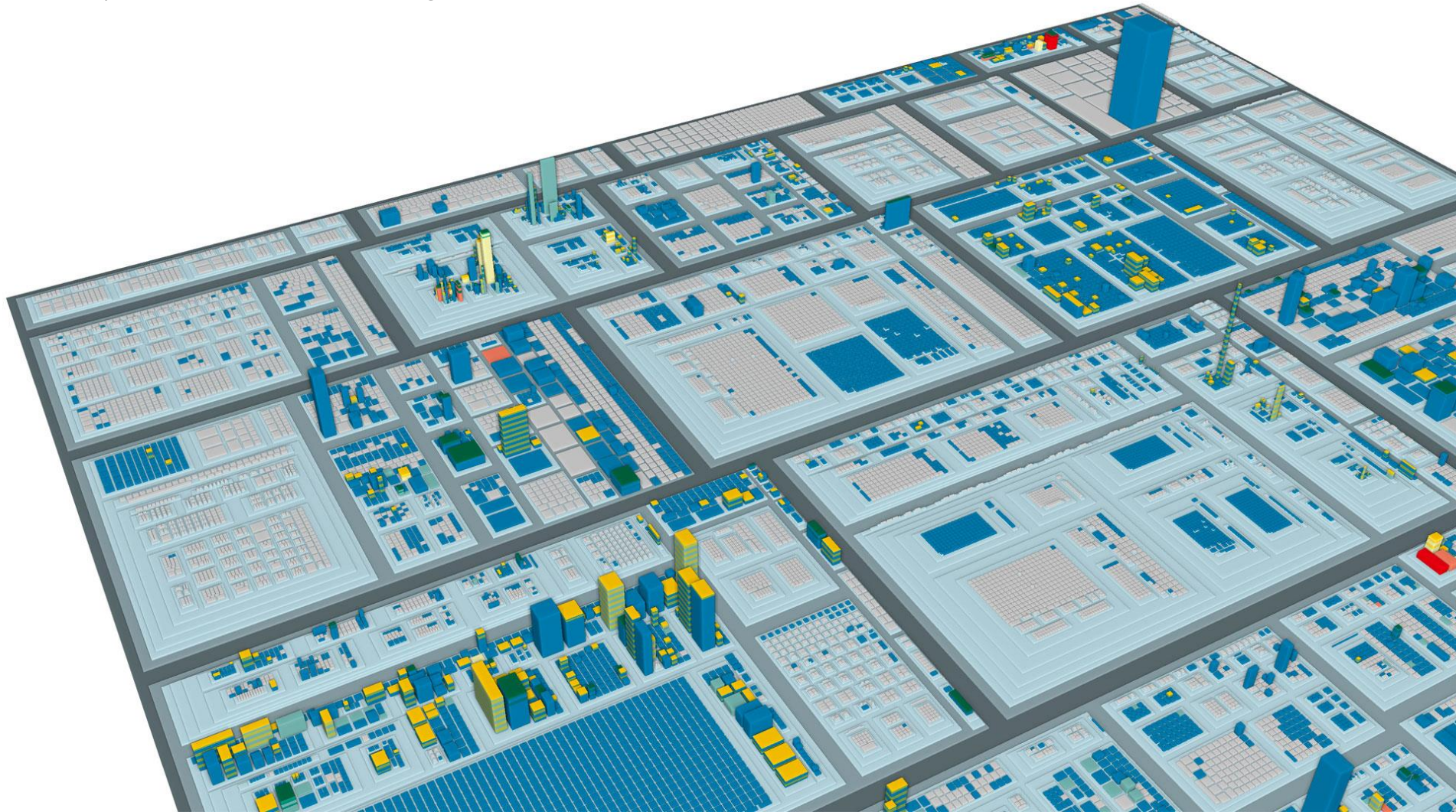


Automated Visual Software Analytics

OpenHPI – Prof. Dr. Jürgen Döllner – 2015



Analytics



Big Data
Big Data Analytics
Analytics Examples

Big Data



"Data is the new Oil. Data is just like crude. It's valuable,
but if unrefined it cannot really be used." Clive Humby, DunnHumby

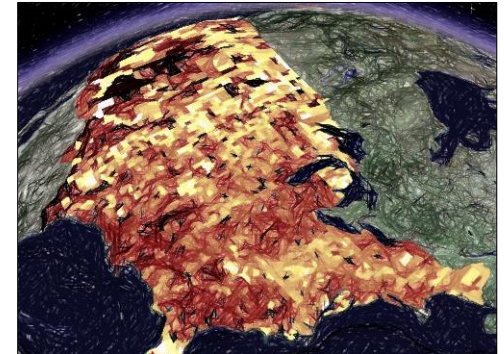
Big Data



www.maritimejournal.com



s.radar.oreilly.com



media.juiceanalytics.com

- **Sensors**, e.g., early-warning systems, automotive systems, assembly lines
- **Business processes**, e.g., transactions, logistics, finance and stock exchange
- **Communication** and digital footprint, e.g., uses of smartphones, media streaming
- **Customer**, e.g., web, online shopping, position tracking
- **Science and research**, e.g., NASA, protein folding simulation
- **Software development**, e.g., large repositories, large software projects, legacy systems
- ...

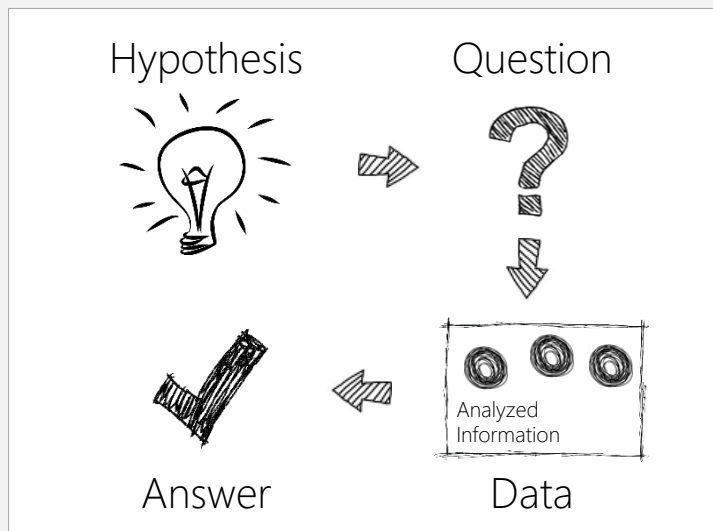
Aspects of Big Data

- **Volume:** high data volume (TB, PB, ZB, ...)
- **Velocity:** high speed of data generation, data streams, and data flows
- **Variety:** high variety such as structured, semi-structured, unstructured, multimedia data
- **Variability:** high variability in data, e.g., inconsistent data flow and flow rates
- **Complexity:** manifold links, relations, and correlations among data
- **Veracity:** high inherent data uncertainty, imprecision, incompleteness

Big Data Analytics

Traditional Analytics

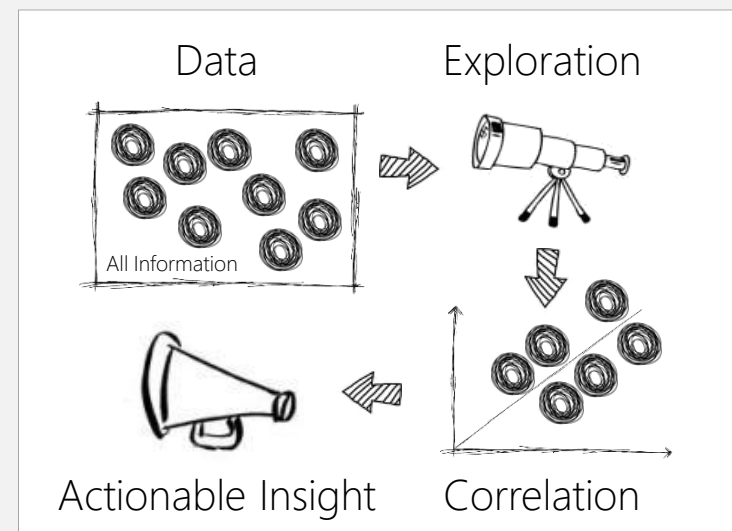
Structured and repeatable
Structure built to store data



Start with hypothesis
Test against selected data

Big Data Analytics

Iterative and exploratory
Data is the structure



Data leads the way
Explore all data, identify correlations

– Adopted from Dr Hammou Messatfa, IBM Europe Government CTO

Big Data Analytics

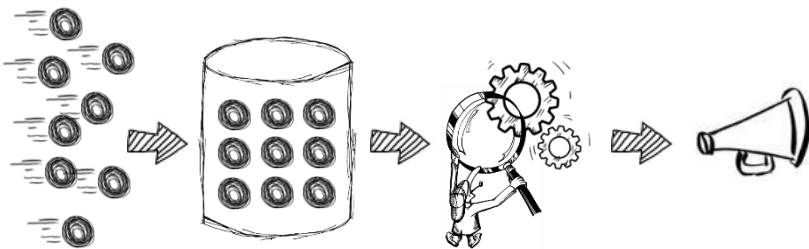
Traditional Analytics

Structured and repeatable
Structure built to store data

Users determine and specify questions



IT builds systems to answer known questions



Analyze after landing...

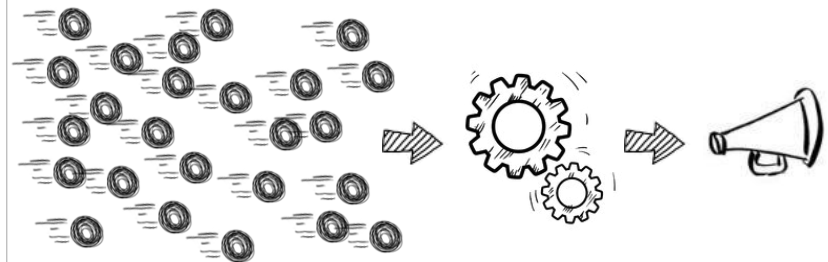
Big Data Analytics

Iterative and exploratory
Data is the structure

IT delivers data from any sources / platform



User asks and explores questions



Analyze while in motion...

– Adopted from Dr Hammou Messatfa, IBM Europe Government CTO

Analytics

- Increasing amounts of large, dynamic, heterogeneous, distributed, and complex data raise the question how applications, systems, and users can use that kind of data.
- Analytics provide effective means to explore, analyze, and use big data.
- Analytics aims at providing methods, techniques, and tools that enable
 - to efficiently get insights into big data,
 - to uncover structures and patterns, and
 - to acquire knowledge by reasoning.

Objectives of Analytics

- discover what is happening,
 - determine why it is happening,
 - predict what is likely to happen and
 - prescribe the best action to take.
-
- “to convert data-driven insights into meaningful actions”
 - “to drive smarter decisions, enable faster actions and optimize outcomes”

– IBM: “Analytics: A blueprint for value”

Analytics in Practice: Examples



Analytics

Measure and boost your impact on Twitter.

Sign in with Twitter



software diagnostics



Analytics versus Analysis

Analytics

- mathematics and statistics, data gathering, data mining, and information visualization technologies, descriptive and predictive strategies (multidisciplinary approach)
- algorithmic rather than heuristic
- derive and deliver quantitative information
- formal and repeatable

Analysis

- systematic examination of structure, composition, and meaning
- heuristic, informal, qualitative
- uncover and understand interrelationships or cause-effect relationships
- providing basis for problem solving and decision making.

- Data Variety: Analytics requires access to a variety of different data sources, typically stored differently, organized differently, and managed by different technologies
- Data Volume: Massive, large data sets demand for efficient access and storage strategies such as out-of-core or external memory algorithms or in-memory databases (e.g., SAP HANA)
- Unstructured Data: It differs from structured data in that its format varies widely and cannot be stored in traditional relational databases
- Data Velocity: To be effective, analytics need to find ways to perform analytics computations in (nearly) real-time, that is, tools and techniques have to find ways to extract, transform, analyze, and visualize information in (nearly) real-time

Analytics represents a major development direction, providing means to effectively use and to take advantage of big data.

The **paradigm** behind **big data analytics** is different to the paradigm in traditional analytics: Big data analytics takes into account the entire data, allows for exploring large data, and provides means to analyze the entire data.

Today's software development processes and software engineering techniques generate big data about software systems, its dynamics, and its evolution. For **software analytics**, principles, concepts, methods, and techniques of analytics can be extended, enhanced, and specialized to get analytics instruments and tools that facilitate to get decision-supporting insights into software development processes and complex computer-based systems.

Analytics



Big Data
Big Data Analytics
Analytics Examples

Big Data

Big data refers to data that is characterized by its high volume, high velocity, high variety, high variability, high complexity, and inhomogeneous veracity. Big data demands cost-efficient and time-efficient approaches of storing and processing for insight and decision making.

As a catchphrase it used to describe both big data itself and related technologies to efficiently handle and manage big data.

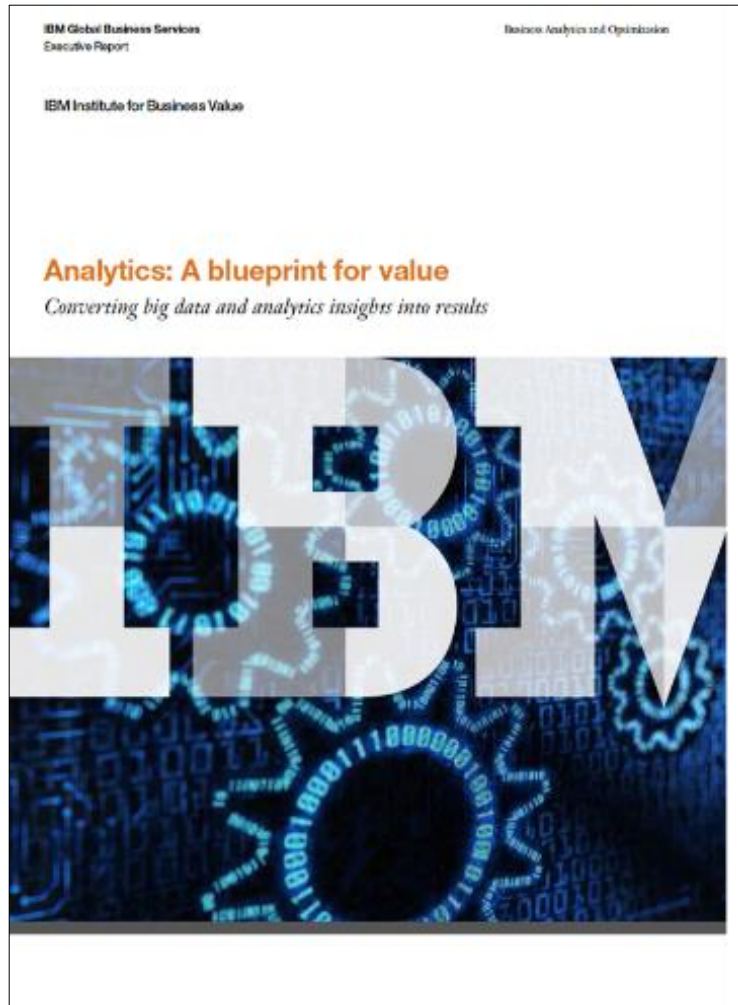
Analytics

Analytics is the science of analytical reasoning that aims at providing concepts, methods, techniques, and tools to efficiently collect, organize, and analyze large sets of data such as big data. Its objectives include to examine data, to draw conclusions, to get insights, to acquire knowledge, and to support decision making.

Analysis

Analysis refers to the examination of a complex, its elements, and their relations (opposite to synthesis). Analytics solutions typically apply, use, and integrate different analysis techniques as building blocks.

Further Reading

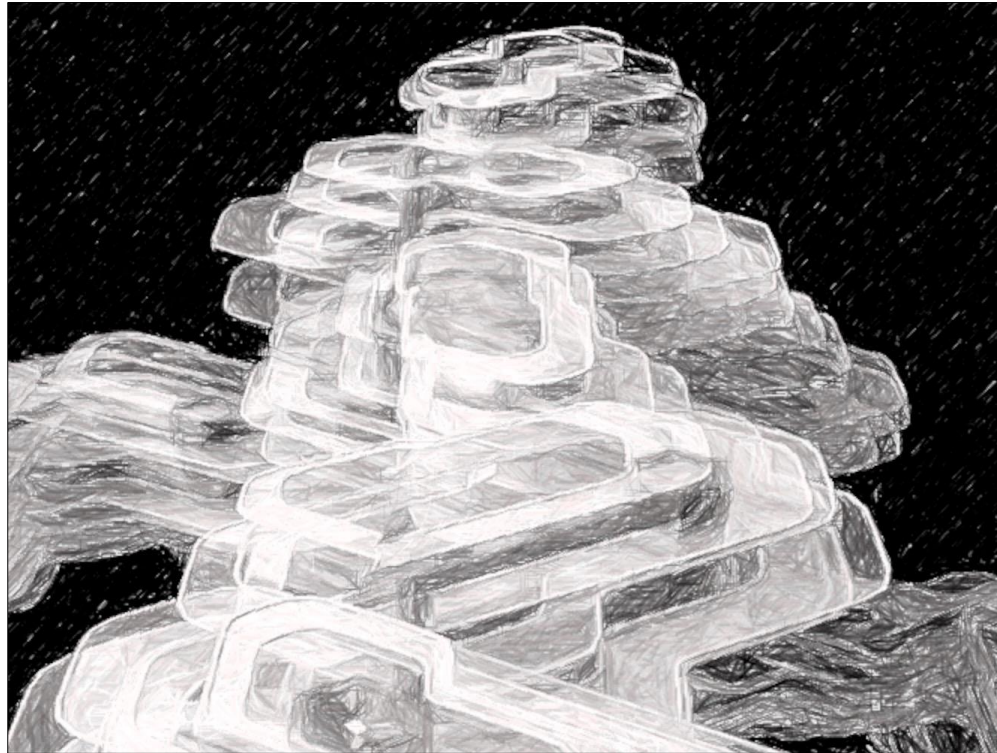


Fred Balboni, Glenn Finch, Cathy Rodenbeck Reese, Rebecca Shockley: "Analytics: A blueprint for value", IBM Institute for Business Value, IBM Global Services, 2013

<http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03575usen/GBE03575USEN.PDF>

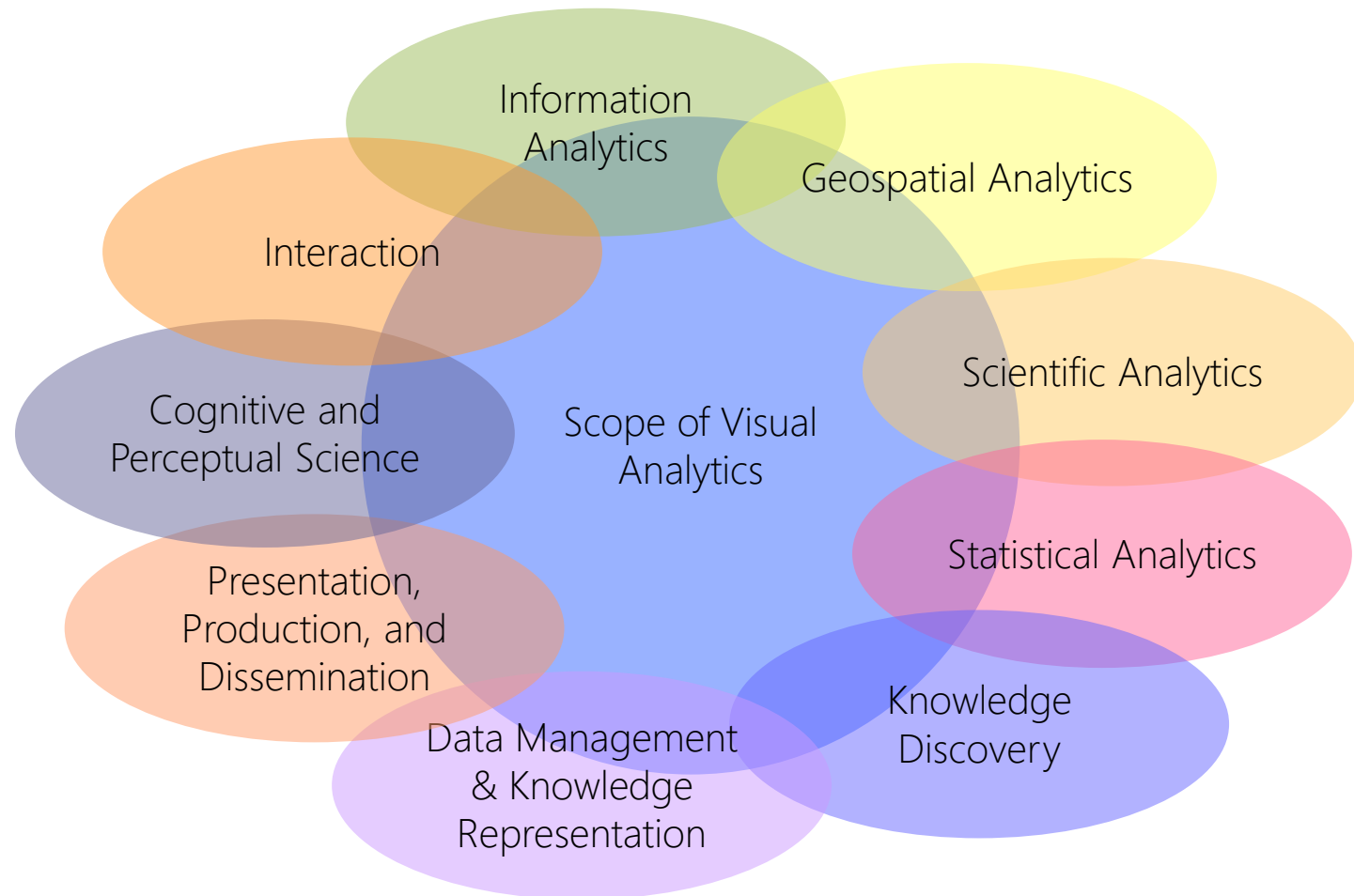
- Big data analytics is considered to be the next key innovation wave in both industry and science. But: There is not too much analytics for software engineering today, both theoretically and practically.
- Analytics will be one of the key “game changing technologies” in software engineering in the future. This course, basically, is exactly about this kind of analytics, called “software analytics”.

Visual Analytics



Principles of Visual Analytics
The Visual Analytics Process
Examples of Visual Analytics

Scope of Visual Analytics



Adopted from Daniel Keim et al.: "Visual analytics: Scope and challenges". Visual Data Mining: 2008, pp. 76-90.

Definition

- Visual analytics combines *concepts of analytics* with *concepts of information visualization and scientific visualization*
- It integrates and exploits capabilities of the human visual system, perception, and cognition to build highly efficient and effective strategies and techniques that enable exploring, analyzing, reasoning, and decision making

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces."

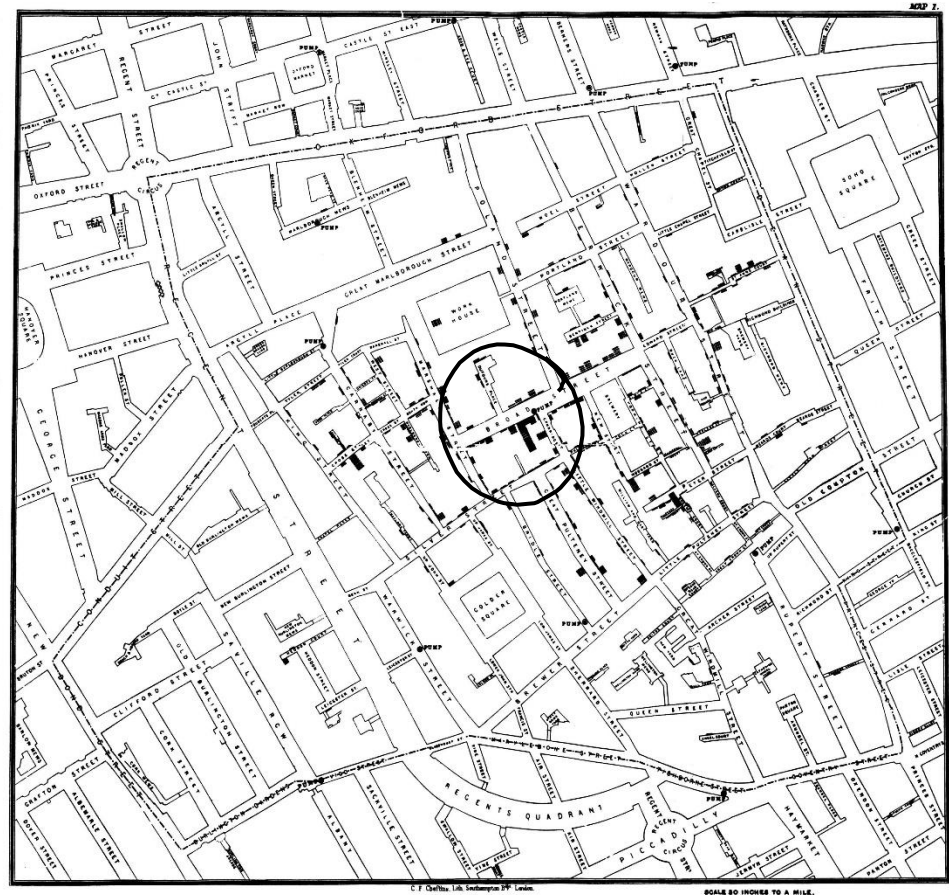
– James J. Thomas, Kristin A. Cook: "Illuminating the Path: The Research and Development Agenda for Visual Analytics", 2005

"Visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis."

– Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, Hartmut Ziegler: "Visual data mining. Chapter: Visual Analytics: Scope and Challenges", 2008

Historic Example of Visual Analytics: John Snow's Map

- London cholera outbreak 1854
- Dot map used to visualize cholera cases on a city map
- Enabled visual exploration and reasoning
- Discovery of relationship between housing and water pumps



http://matrix.msu.edu/~johnsnow/images/online_companion/chapter_images/fig12-5.jpg



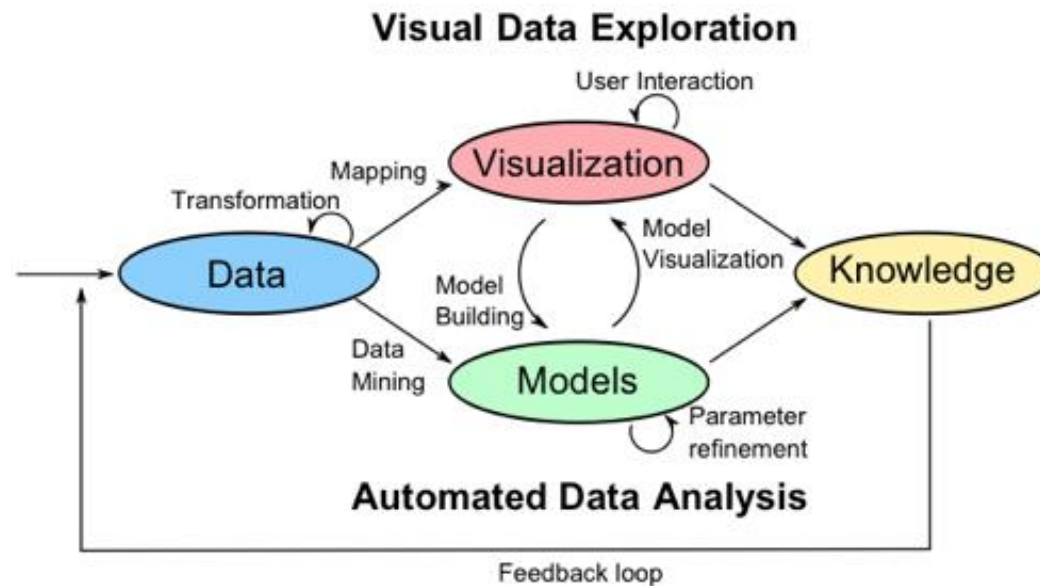
Example: What is being visualized?



Example: Nearest distance to McDonald restaurants in the US

The Visual Analytics Process

- Users have an active role in visual analytics: They can control data transformation, data mapping, model building, model visualization, and all corresponding parameters.
- Users can combine their creativity and background knowledge with analytics techniques.



Adopted from Daniel Keim et al.: "Visual analytics: Scope and challenges". Visual Data Mining: 2008, pp. 76-90.

Visual Analytics vs. Information Visualization

Visual Information-Seeking Mantra

“Overview first, zoom and filter, then details-on-demand.”

– Ben Shneiderman: “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”, 1996

Visual Analytics Mantra

“Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand”

– Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, Hartmut Ziegler: “Visual data mining. Chapter: Visual Analytics: Scope and Challenges”, 2008

General Tasks of Visual Analytics

- Synthesize information from big data or complex data
- Detect the expected, discover the unexpected
- Provide timely, defensible, and understandable assessments
- Communicate assessment effectively for action

Design and Development of Visual Analytics Techniques

Data oriented challenges:

- Find efficient data gathering, fusion, and integration techniques
- Find appropriate data mining and data analysis techniques
- Handle characteristics of big data with respect to volume, velocity, variability, variety, veracity, and complexity.

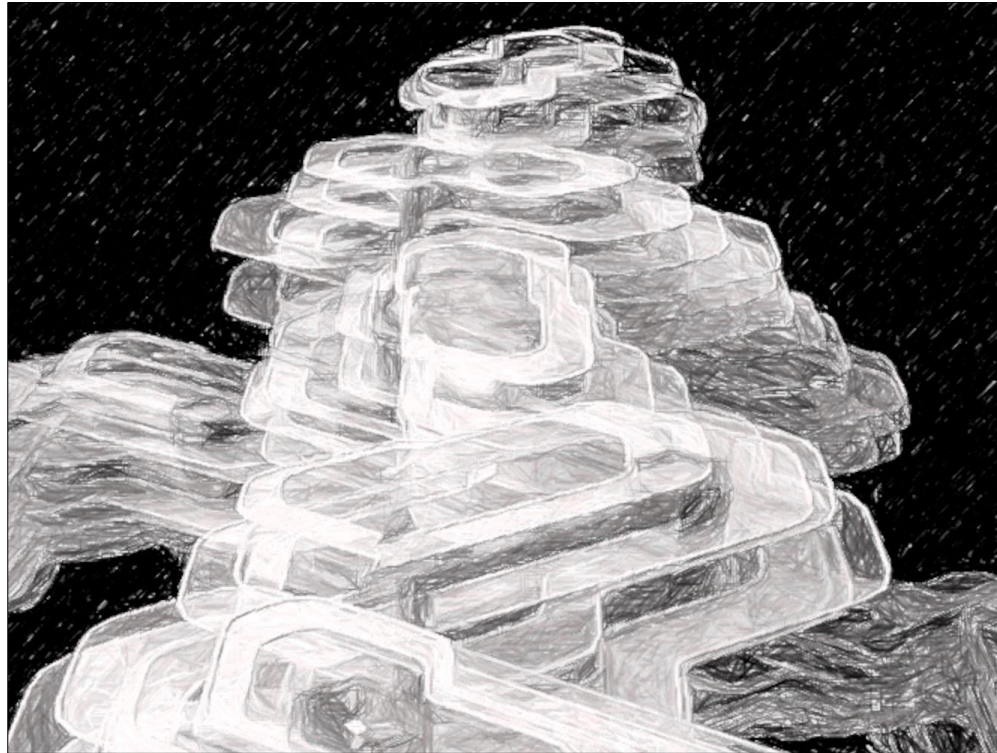
Visualization oriented challenges:

- Find metaphors for unstructured or semi-structured big data
- Find space-efficient and graphics oriented data mappings
- Find comprehensible user interactions

Visual analytics combines algorithmic and technological strengths of analytics with approaches of information visualization and scientific visualization that take advantage of the human visual system to effectively analyze visual media.

Visual analytics offers a framework of principles, concepts, methods and techniques. They are adopted and extended for the scope of software development and software engineering, leading to **visual software analytics**.

Visual Analytics



Principles of Visual Analytics
The Visual Analytics Process
Examples of Visual Analytics

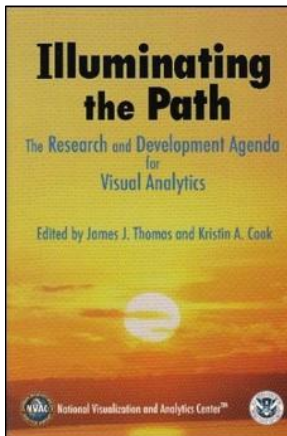
Visual Analytics

Visual analytics combines principles of analytics with principles of information visualization. It integrates visualization as an essential means to explore, analyze, and understand big data supported by the human visual systems.

Further Reading



Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, Hartmut Ziegler: "Visual data mining. Chapter: Visual Analytics: Scope and Challenges", pages 76–90. Springer-Verlag, Berlin, Heidelberg, 2008, <http://bib.dbvis.de/uploadedFiles/55.pdf>

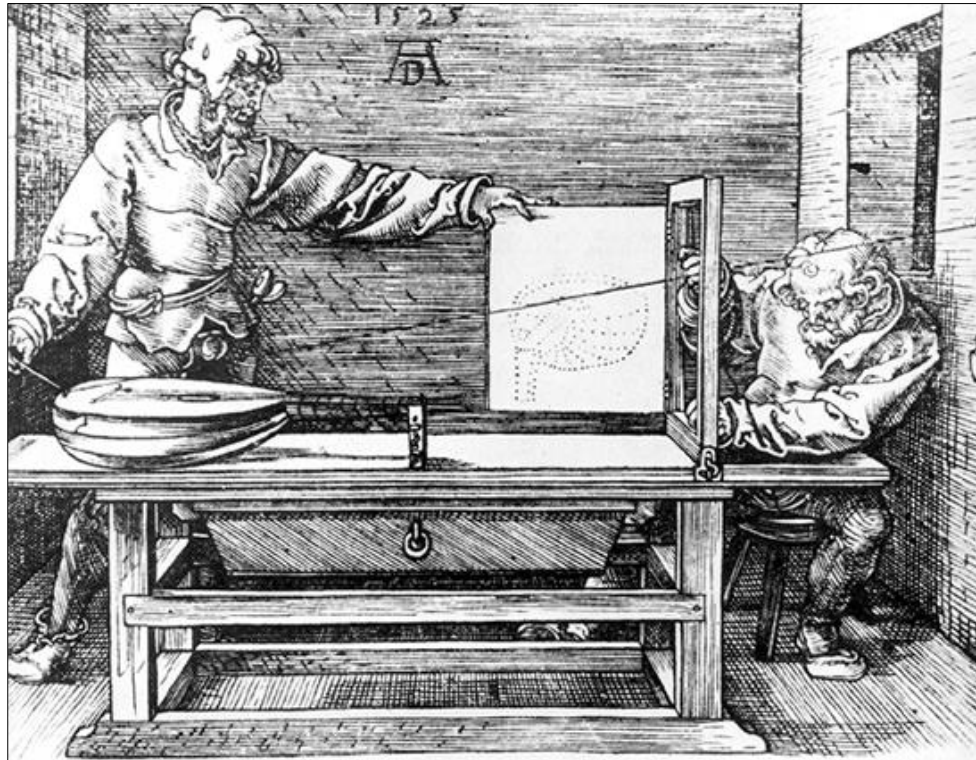


James J. Thomas, Kristin A. Cook: "Illuminating the Path: The Research and Development Agenda for Visual Analytics", National Visualization and Analytics Ctr, 2005

In analytics, visual methods are frequently used and represent key elements, in particular for interactive, visual interfaces of analytics systems and applications.

Visual analytics offers a large "construction kit" that can be used to build domain-specific visual analytics tools and techniques such as in the case of visual software analytics.

The Visualization Pipeline



Objectives of Visualization
Visualization Pipeline Model
Reference Model for Visualization

Objectives of Visualization

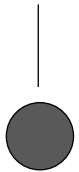
- **Exploration** of large sets of data
- **Analysis** of structure, patterns, and relations
- **Understanding** of concepts and processes
- **Gaining insights** and comprehension
- **Communication medium** and collaboration tool

Visualization

- **“to visualize”: to make something visible to the human mind.**
- takes advantage of and uses the human's visual capabilities

Visualization Pipeline – Raw Data

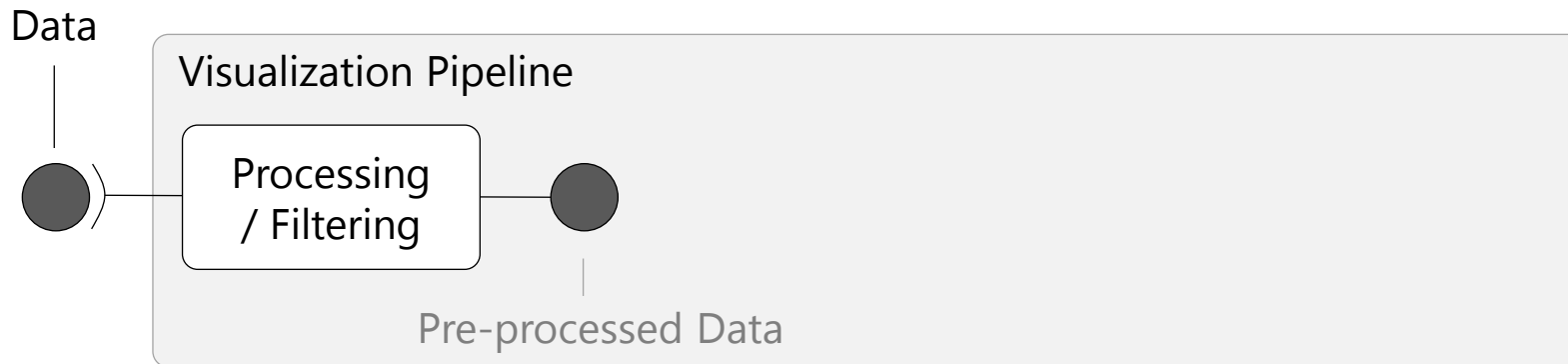
Data



Visualization Pipeline

ID	Name	Costs	Lifetime	Failure Risk
1	Turbine	356000	12	0.40
2	Generator	137000	36	0.50
3	Transformer	120000	42	0.80
4	High-voltage Power Line	245	12	0.05
...		...	...	...
212	Capacitor	530000	3	30

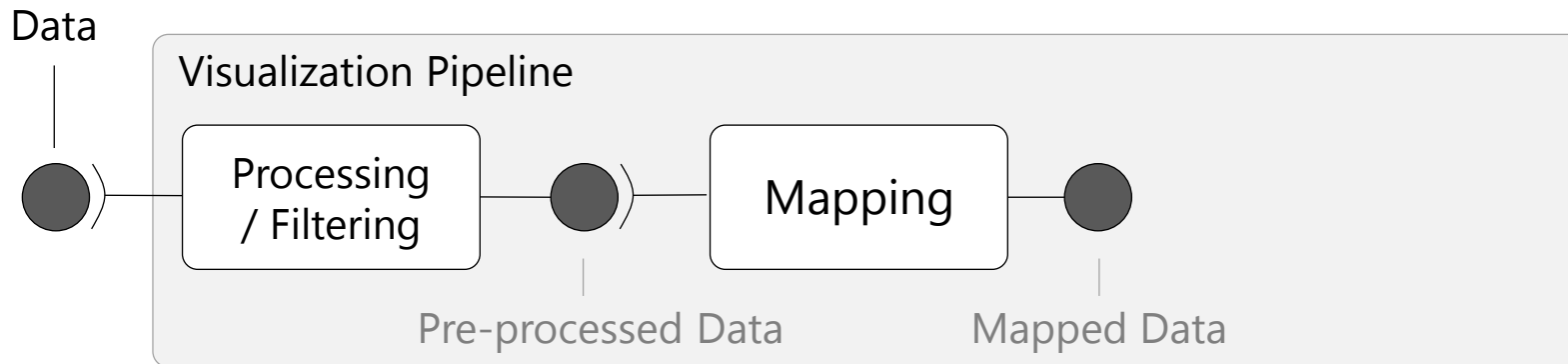
Visualization Pipeline – Processing



ID	Name	Costs	Lifetime	Failure Risk
1	Turbine	356000	12	0.40
2	Generator	137000	36	0.50
3	Transformer	120000	42	0.80
4	High-voltage Power Line	245000	12	0.05
...		...	...	...
212	Capacitor	530000	3	0.30

- Data completion
- Data reduction
- Data normalization
- Data filtering
- Threshold operations

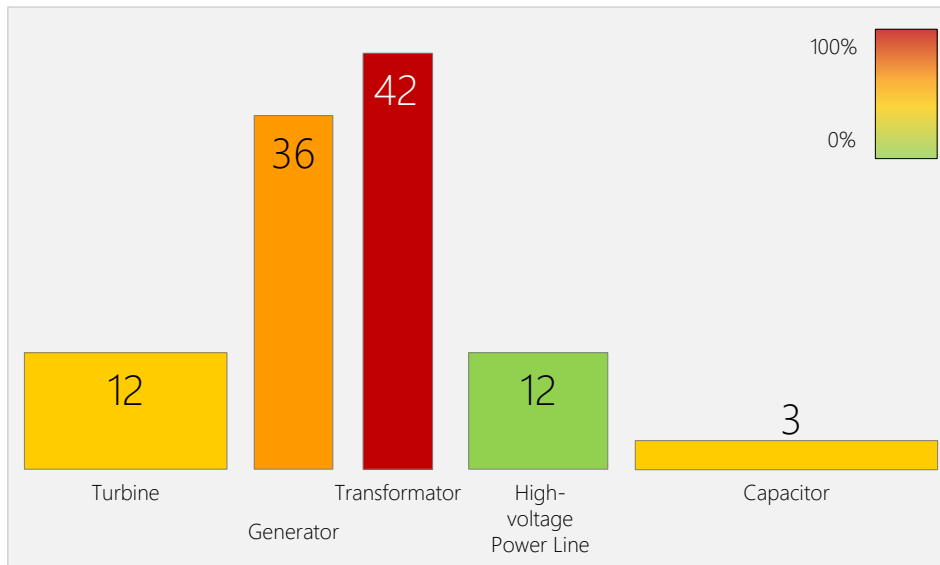
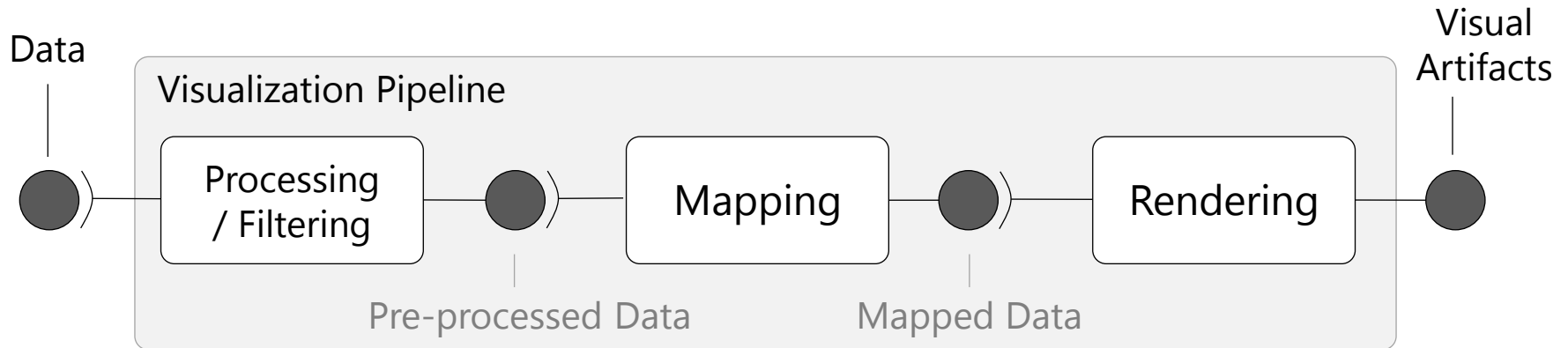
Visualization Pipeline – Mapping



ID	Name	Costs	Lifetime	Failure Risk
1	Turbine	<u>356000</u>	12 	0.40
2	Generator	<u>137000</u>	36 	0.50
3	Transformer	<u>120000</u>	42 	0.80
4	High-voltage Power Line	<u>245000</u>	12 	0.05
...		...	...	...
212	Capacitor	<u>530000</u>	3 	0.30

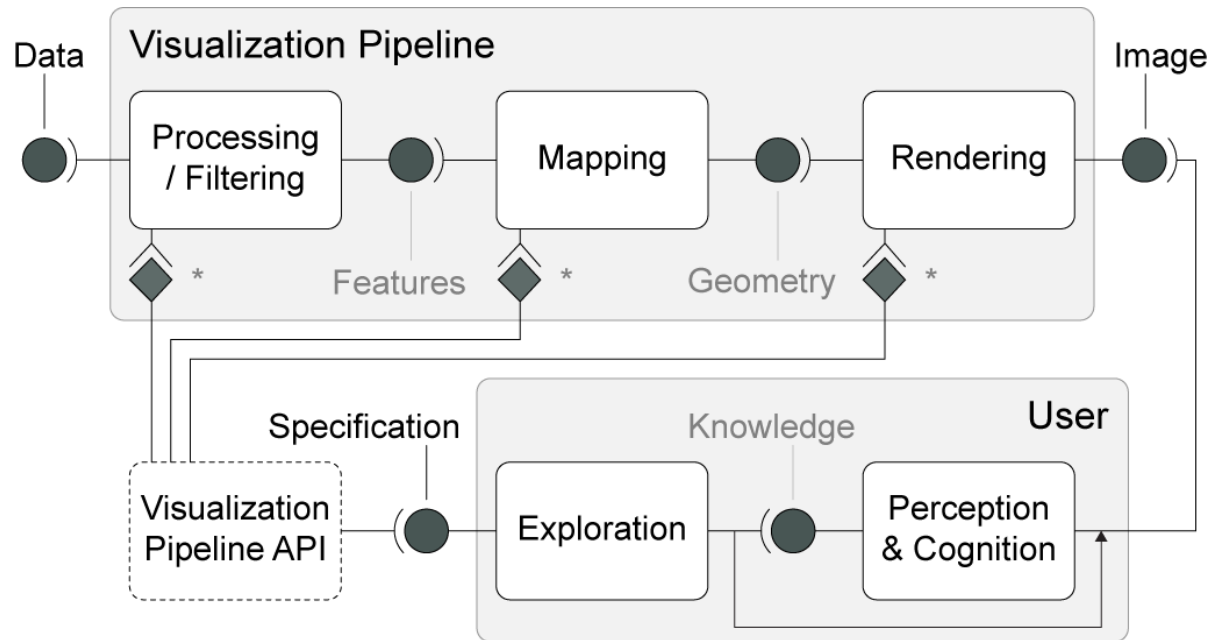
- Mapping data to visual primitives and visual attributes
- Visual variables include form, color, texture, position, size, orientation.

Visualization Pipeline – Raw Data



- Rendering of 2D or 3D graphics
- Common forms include diagrams, plots, and charts.

A Reference Model for Visualization



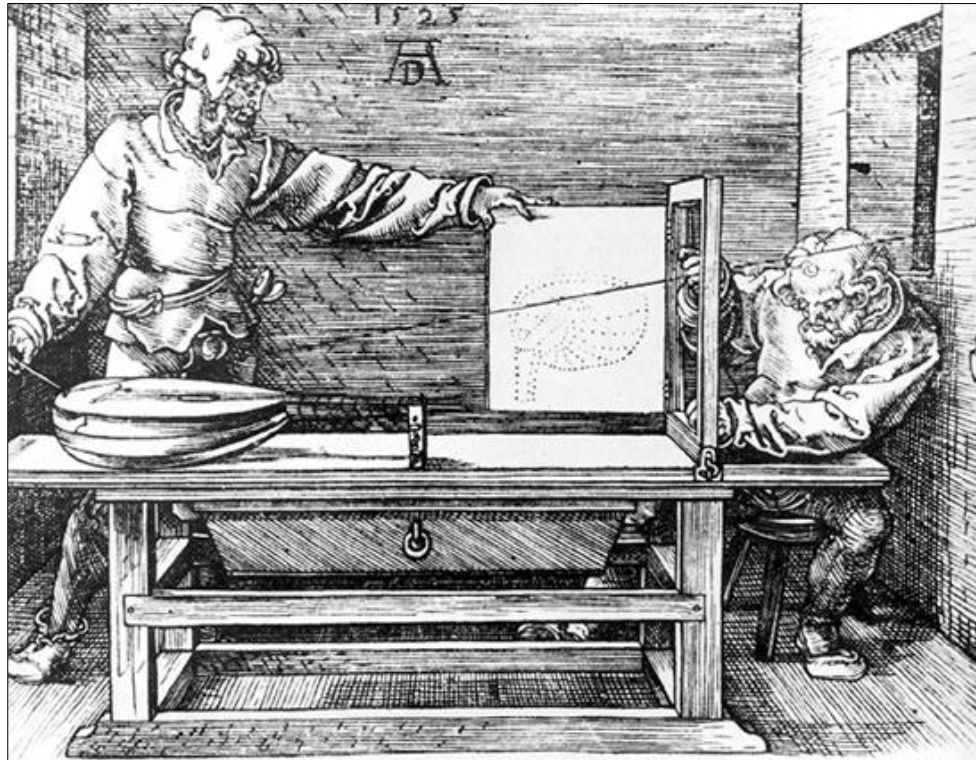
The visualization pipeline is part of the **visualization process** that consists of:

- perceiving images by the human user
- obtaining insights by exploration and analysis
- adjusting data, data processing, data mapping, and rendering
- feedback loop (image → user → image → ...)

Visualization is a powerful method to provide users access to and to facilitate understanding of data. Visualization plays a key role if data becomes massive (in terms of size, number of entities, etc.) and/or if data is complex (in terms of relations, dependencies, hierarchies). Visual analytics applies manifold methods, techniques, and tools of visualization to handle and to cope with massive data and big data.

Software analytics uses visualization to provide interactive access to software-related data. The key challenge is to find appropriate mappings as data about software, software process, and software evolution has no natural “appearance”, geometry or gestalt. Visualization is a defining and essential element of all software analytics approaches.

The Visualization Pipeline



Objectives of Visualization
Visualization Pipeline Model
Reference Model for Visualization

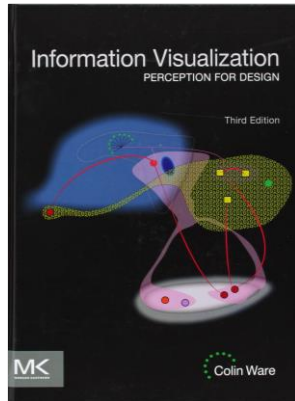
Visualization

The process of filtering, mapping, and rendering data such that it can be perceived, understood, and interpreted by the visual system of the user, aiming at giving comprehension and insight into the underlying data.

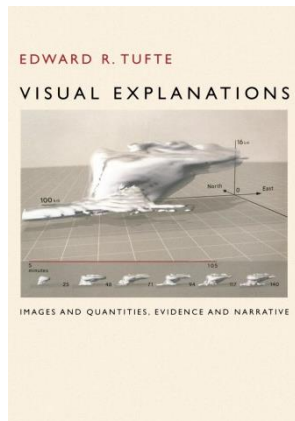
Visual Variables

A visual variable is a parameter to which data properties or values can be mapped. In visual mappings, visual variables include shape (form), position, size, orientation, color, texture, and any further attributes such as animations.

Further Reading



Colin Ware: "Information Visualization, Perception for Design", 3rd Edition, Morgan Kaufman, 2012. (A general, self-contained introduction to the whole field of visualization)

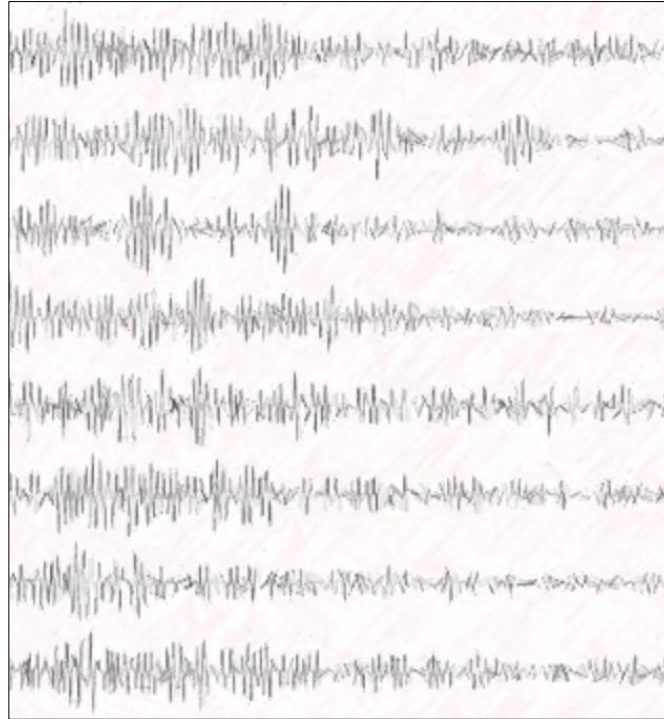


Edward R. Tufte: "Visual Explanations: Images and Quantities, Evidence and Narrative", Graphics Press, 1997.

Visualization is not about making something visible to the eyes – it's about making something visible to the mind.

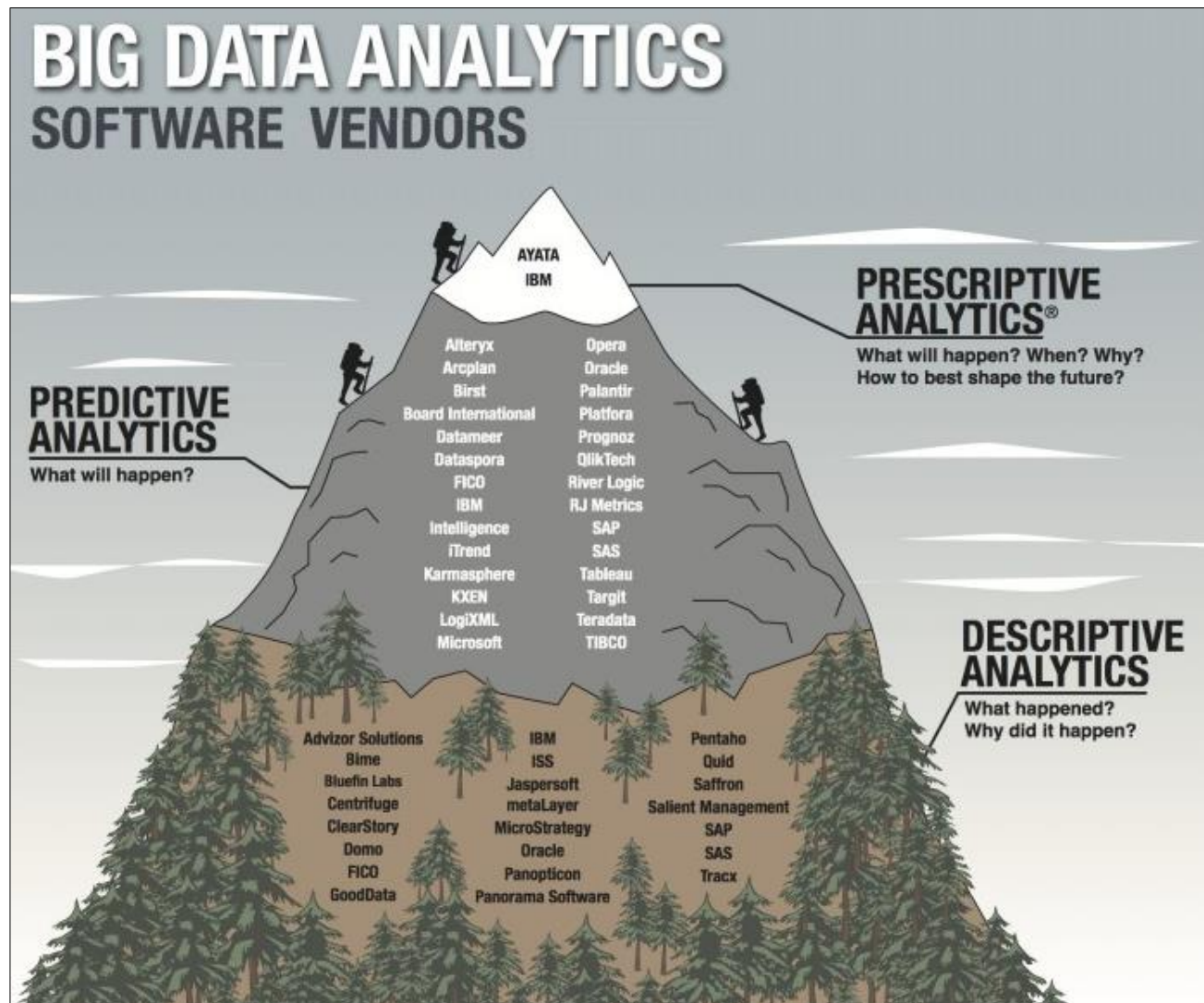
Visualization is considered to be key method and technique to cope with big data, that is, it offers ways to access, grasp, and use that data by means of visual, interactive mappings.

Predictive Software Analytics



Principles of Predictive Analytics
Predictive Software Analytics
Applications of Predictive Software Analytics

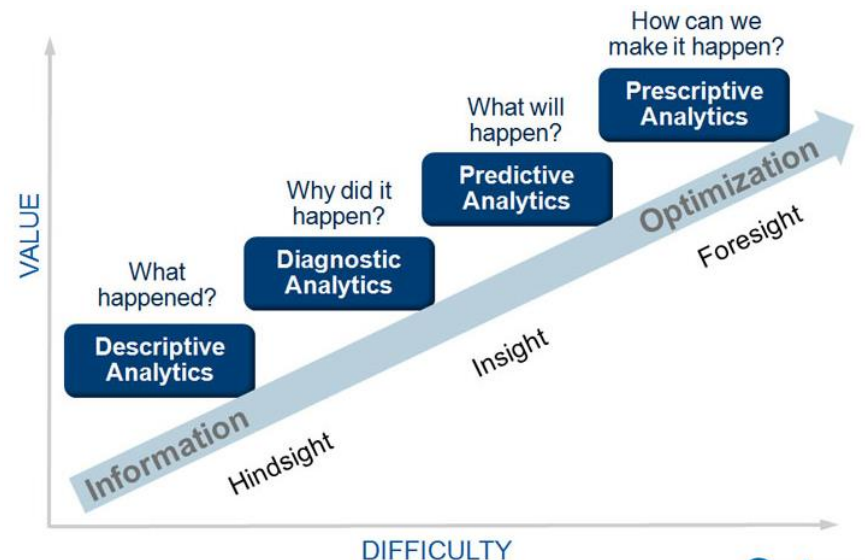
Principles of Predictive Analytics



Principles of Predictive Analytics

Definition of Predictive Analytics

- Predictive analytics denotes analytics used to examine trends and patterns that enable or facilitate to forecast and predict processes, phenomena, or events.
- The core of predictive analytics relies on *capturing relationships between explanatory variables and the predicted variables from past occurrences or from comparable data, and exploiting them to predict the unknown outcome.*
- The “unknown” can be located in the future, in the present, or in the past.



Principles of Predictive Analytics

	Past	Present	Future
Information	What happened? (Reporting)	What is happening now? (Alerts)	What will happen? (Extrapolation)
Insight	How and why did it happen? (Modeling)	What's the next best action? (Recommendation)	What's the best/worst that can happen? (Prediction)

From Davenport et al. "Analytics at Work"

Principles of Predictive Analytics

Examples Predictive Analytics Application Fields

- Clinical decision support
- Cross-selling
- Fraud detection
- Financial risk management



Predictive Modeling

- Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit.
- The objective of the model is to assess the ~~likelihood~~ that a similar unit in a different sample will exhibit the specific performance.
- The available sample units with known attributes and known performances are called "training sample".
- The units in other samples, with known attributes but unknown performances, are called "out of sample" units.
- The out of sample bear no chronological relation to the training sample units.

– Wikipedia

Implementation of Descriptive Data Mining Techniques

- Clustering : division of items into groups of similar objects
- Association Rule Discovery : identifying relations, search for rules
- Sequential Pattern Discovery : finding relevant patterns in items of a sequence

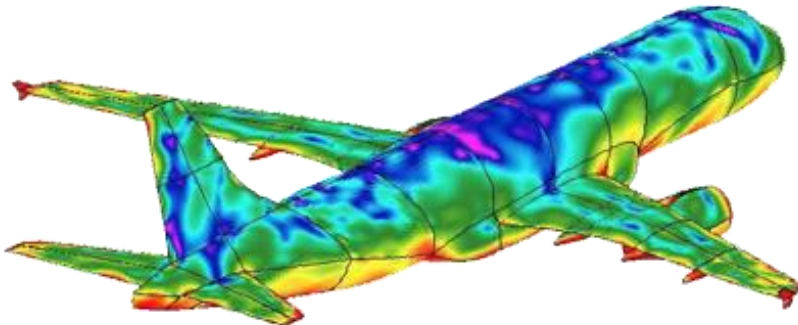
Implementation of Predictive Data Mining Techniques

- Classification : predicting item classes
- Regression : interaction between different variables
- Deviation Detection : finding changes
- Collaborative Filtering : analyzing (e.g., behavior or preference) patterns in user groups

Predictive Analytics in Software Engineering

"In recent years there has been a huge growth in using statistical and machine learning methods to find useful prediction systems for software engineers. Of particular interest is predicting project effort and duration and defect behaviour. Unfortunately though results are often promising no single technique dominates, and there are clearly complex interactions between techniques, training methods and the problem domain."

- Martin Shepperd: "The scientific basis for prediction research", PROMISE 12, 2012



- Predictive Engineering: Example of heat transfer prediction (proto3000.com)

Predictive Analytics in Software Engineering

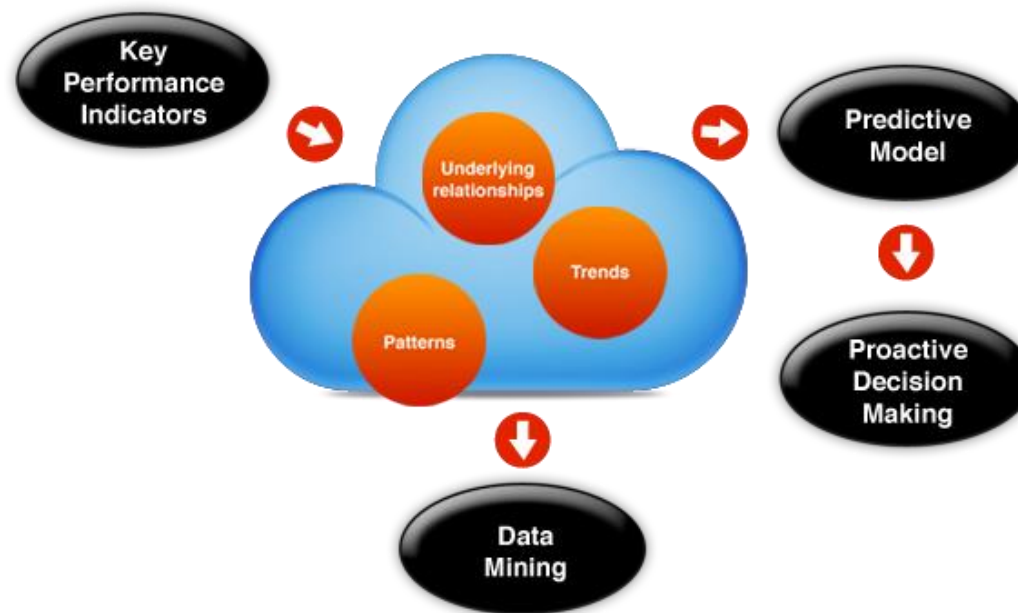
- Software engineering represents “a data rich activity with many sophisticated metrics”
 - the data is recorded but is rarely used for analytics applications
- “many aspects of development, from code repositories to testing frameworks to bug databases, can be measured with a high degree of automation, efficiency, and granularity.”
 - the data that is recorded is perfectly suited for analytics applications

– R- Buse & T. Zimmermann: "Information Needs for Software Development Analytics", 2011

Prediction and Decision Making

- Predictive software analytics constitutes an essential element for decision systems in software engineering
- "Predictive models are sub-routines in a higher level decision process"

–Tim Menzies: "Beyond Data Mining; Towards 'Idea Engineering'", 2013



– <http://www.metrifit.com>

Prediction and Decision Making

- Decision systems generally have to provide a *sense of "comfort"* to managers
- Key features of decision systems:
 - Finding a problem: detection and diagnosis
 - Solving a problem: finding, evaluation of alternatives, and judgment
 - Resolution: monitoring the effect of the solution

Examples of Applications of Predictive Software Analytics for Software Projects

- Estimations on software efforts and costs (budget, time)
- Software fault prediction including bug prediction
- Software quality assurance and predictive maintenance
- Prioritize investments for code with substantial customer impact
- Recommendation systems for software refactoring and redesign
- Recommendation systems for software quality improvements

Examples of Advanced Applications of Predictive Software Analytics

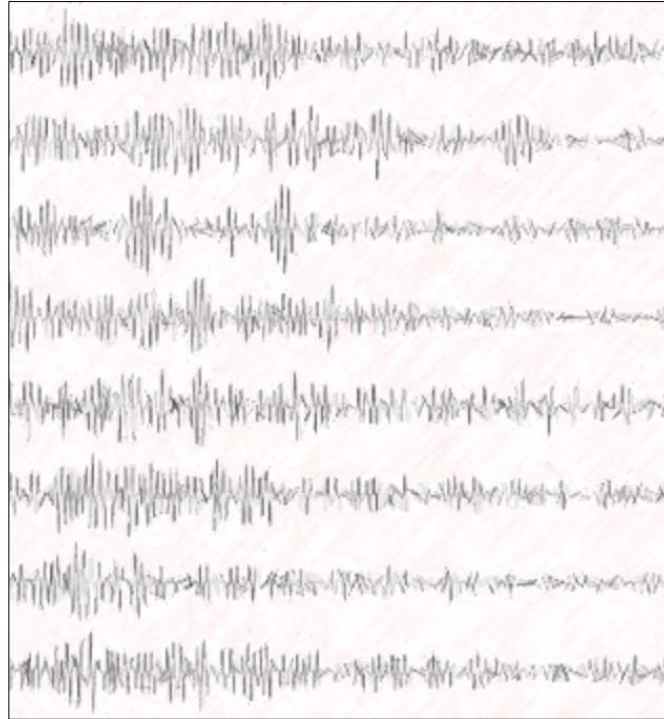
- Comparisons and benchmarking of an individual software project within a group of similar software projects
- “DNA” analytics for a given software project. DNA base is derived and collected from a large number software projects, and constantly updated and extended. What has happened in the past in a specific setting might happen in the future.

Predictive software analytics relies on general prediction analytics and prediction modeling, specialized to data common to large-scale, long-term software development processes and software engineering.

Predictive software analytics examines **trends and patterns in software systems and related development processes** that enable or facilitate to forecast and predict processes, phenomena or events, located in the future, in the present, or in the past. It captures relationships between explanatory variables and the predicted variables from past occurrences or from comparable data.

Predictive software analytics is a **key element to build software engineering decision systems** that provide a sense of “comfort” by detection and diagnosing problems, by providing evaluation of alternatives and judgment, and by monitoring adopted resolutions and observe their effects.

Predictive Software Analytics



Principles of Predictive Analytics

Predictive Software Analytics

Applications of Predictive Software Analytics

Predictive Analytics

Predictive analytics refers to analytics used to examine trends and patterns that enable or facilitate to forecast and predict processes, phenomena or events in the future, in the present, or in the past.

Predictive Software Analytics

Predictive software analytics refers to predictive analytics specific to large-scale, long-term software development projects and software engineering. It provides means, for example, to estimate efforts (costs, time), to predict software quality measures and related deficiencies or weaknesses, and to predict software project changes. Predictions are essential to enable effective, informed decision making in software engineering.

Further Reading

Information Needs for Software Development Analytics

Raymond P.L. Buse and Thomas Zimmermann

Abstract. Software development is a data-rich activity with many opportunities for analytics. To maximize the benefits of such data, we present a taxonomy of information needs for software development analytics. We discuss the challenges of such a taxonomy and propose a taxonomy of information needs for software development analytics. We discuss the challenges of such a taxonomy and propose a taxonomy of information needs for software development analytics.

1. Introduction. Software engineering is a data-rich activity. Many of the activities in software engineering are data-intensive. The data generated by software engineering can be used to improve the efficiency and productivity of the software engineering process. This paper discusses the information needs for software development analytics.

Despite the abundance of data and metrics, development is still a difficult task. The data generated by software engineering can be used to improve the efficiency and productivity of the software engineering process. This paper discusses the information needs for software development analytics.

2. Related Work. There is a large body of work on software development analytics. This paper discusses the information needs for software development analytics.

Raymond P.L. Buse, Thomas Zimmermann: "Information Needs for Software Development Analytics", Proceedings International Conference on Software Engineering (ICSE), IEEE CS, 2012, <http://thomas-zimmermann.com/publications/details/buse-icse-2012>.

Remarks

Predictive software analytics leads to informed, actionable decision making in software engineering projects – a key and ultimate goal of analytics.

Software Analytics



The Need for Software Analytics
Principles of Software Analytics
Automated Visual Software Analytics

The Need for Software Analytics

"In software engineering (SE) there is much we are *seeing*, but little we are *learning*. The sheer volume of SE data is overwhelming. [...] Yet from that data, we have extracted nearly zero general principles- the usual result, across all this data, is that what works on one project may not work on another [...]. Hence there is an urgent need for better analysis of this data. Further, due to the volume of information, that analysis must be (at least partially) automated."

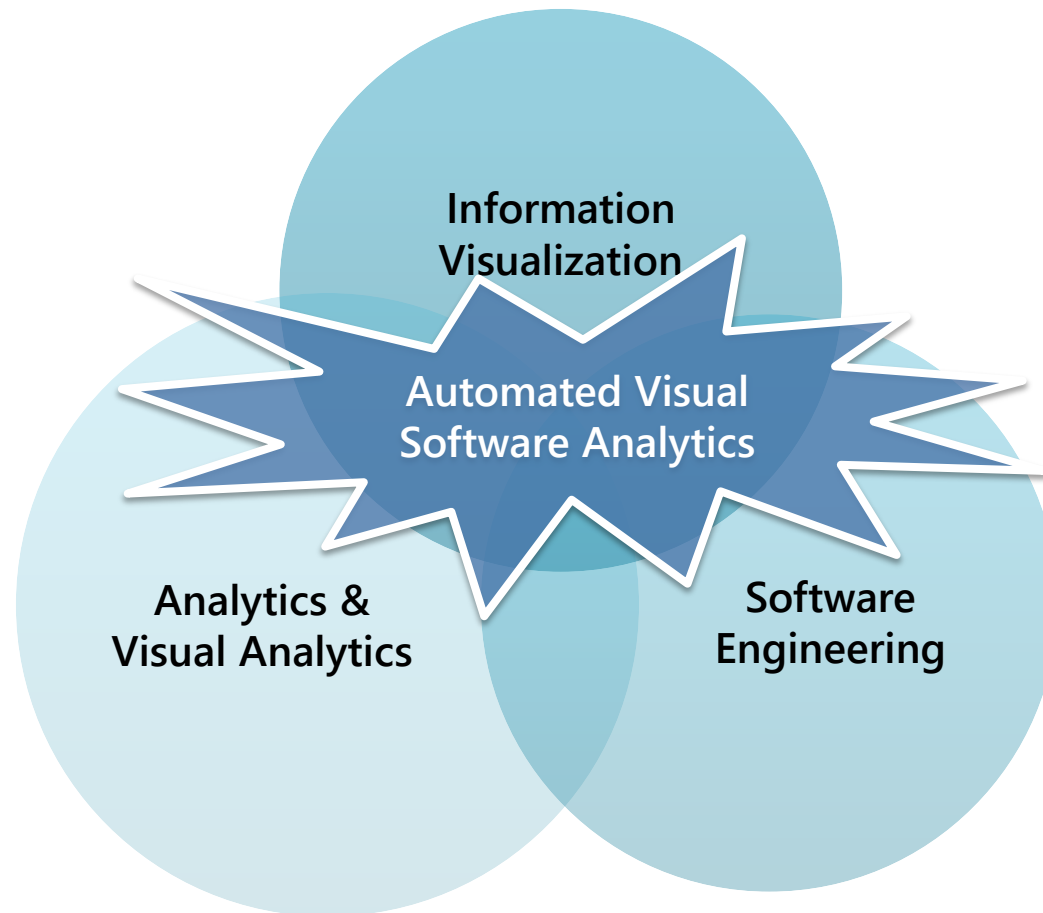
– Tim Menzies: "Predicting the Future of Predictive Modeling", AISE 2012

The Need for Software Analytics

General needs for large-scale, long-term software projects (examples):

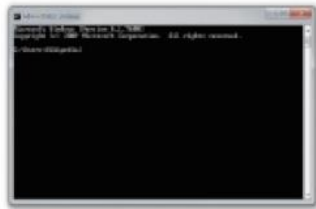
- Generating defect predictors to guide inspection teams
- Estimate technical debts in software projects
- Predicting software power consumption
- Predicting overall project effort
- Finding effective project changes
- Revealing the human networks inside software teams
- Generating normal interface usage patterns by execution traces
- Recording application usage and context data
- ...

Principles of Software Analytics

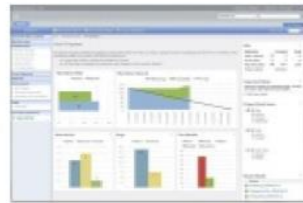


Principles of Software Analytics

Different stakeholders in software development processes



Researcher



Developer



Tester



Dev Lead



Test Lead



Manager



Automated Visual Software Analytics

- refers to analytics as the general discipline
- is specialized to data and information relevant for large-scale software developments
- uses visual and interactive techniques to provide access to complex data
- relies on automated data gathering, integration, and fusion

Visual interfaces for

- Source codes (e.g., hierarchy of system components)
- Source code metrics associated with source code units
- Test data and test results
- Run-time traces
- System evolution data

Constraints

- No “natural” geometry, form, shape or gestalt for most of the data
- Finding appropriate visual metaphors for visualizing software data
- Finding appropriate interactions to handle software data

Automated gathering of data such as

- Repository data
- Software metrics data
- Software test data and run-time traces

Constraints

- No explicit modeling or meta-modeling
- No explicit or specialized software development processes or tools
- No interfere with programming language usage or programming guidelines

Key Criteria for Automated Visual Software Analytics

- It must be “automated”, that is, collecting and gathering the broad variety of data about software systems, their development, dynamics, and evolution *must not depend on explicit or manual steps* (e.g., consultancy, meta models, etc.). For large-scale, long-term projects it is practically impossible to achieve better engineering by “normative” approaches.
- It must rely on visual technologies that master “big software engineering data”, that is, software analytics *requires effective means in terms of interactive interfaces* to the data. As software data is related to an extremely complex but intangible artifact, strong visual metaphors need to be applied that allow to get insight into the data.

Software analytics follows the principles of analytics, specialized to the demands and conditions common to large-scale, long-term software development projects.

There is an enormous **demand for software analytics**. *Engineering* techniques in "software engineering" must grow and evolve to cope with the increasing role of complex computer-based systems in a software-dependent society to ensure system dependability.

Software analytics is most likely the **key step towards strong engineering** in "software engineering", providing a rich framework that enables effective comprehension of complex computer-based systems, their evaluation, and corresponding decision making.

Key criteria for software analytics include **automation** regarding data collection and gathering, and **effective visual interfaces** based on adequate visualization metaphors and techniques.

Software Analytics



The Need for Software Analytics
Principles of Software Analytics
Applications of Software Analytics

Automated Visual Software Analytics

Automated visual software analytics denotes principles, concepts, methods, and techniques of visual analytics that

- a) take into account the specific demands and conditions in large-scale, long-term software development projects and software engineering,
- b) automatically gather data from a wide variety of data source common to software developments and software engineering and
- c) apply effective visual interfaces based on metaphors that allow for mapping the data of intangible artifacts.

Further Reading



Tim Menzies: "Predicting the Future of Predictive Modeling".
NSF Workshop: Planning Future Directions in AI & SE (AISE'12),
September 10, 2012.

Remarks

Automated visual software analytics will be the game changer in software development. And it will significantly contribute to mature the software “engineering” discipline.