Software Requirements
L6. Usage Data Analytics

Daniel Martens

Prof. Walid Maalej

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Outline

1 **Software Analytics**

2 Usage Data

3 Analytics Techniques

How do managers decide?

Nearly half (40 percent) of major decisions are based on the gut feeling

[Accenture, 2008]

What is (data) analytics?

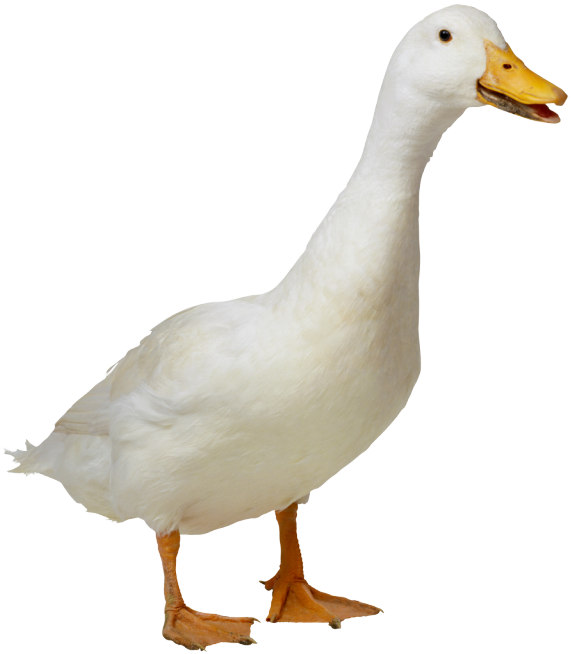
"The use of analysis, data, and systematic reasoning to make decisions"

Davenport et al., *Analytics at Work*:
Smarter Decisions, Better Results, 2000

Many definitions

Why (data) analytics?

- Decisions based on facts instead of assumptions
- Increased transparency and traceability of decisions
- Less subjective decision making
- Data driven engineering
- (Semi) automated way for decision making!
  - Reduced reaction time
- Decreased risk of “overlooking” something



Software analytics…

Enables software practitioners to **explore** and **analyze** data in order to obtain **insightful** and **actionable** information for **data-driven tasks** around Software and Services

[Zhang et al., 2013]

Examples…

- Web analytics:
  - Which pages are visited? By how many? How often?
  - What is the average visit duration?
  - Which devices and browsers are my users using?
  - When do visitors leave the site?
- Mobile apps analytics
  - Which features are popular and which are not?
  - Which views are used, in which sequence?
  - What are the most frequent errors encountered?
  - Which features need to be improved?

Data analytics is insightful

- Analytics results are meaningful and useful for understanding of the task performed
- Typically, not possible by manually inspecting raw data
- Example
  - Experiences of users
  - Usage patterns
  - Quality assurance
  - Bug prediction

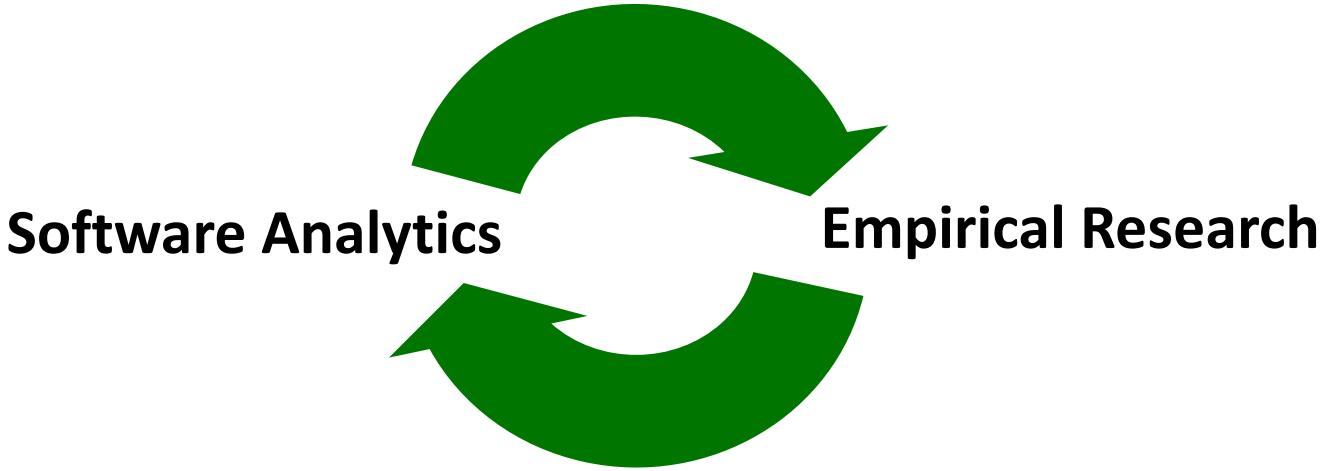
Data analytics is actionable

- Analytics results can be translated into decisions (and actions) to perform the task at hand
- Examples:
  - Identify new requirements
  - Prioritize requirements
  - Plan requirements for future releases

Actionable results? Not enough!

Results of data analytics need to be evaluated by practitioners:

- What do stakeholder think about the results?
- Is it applicable in their context?
- How much would it help them in their daily work?

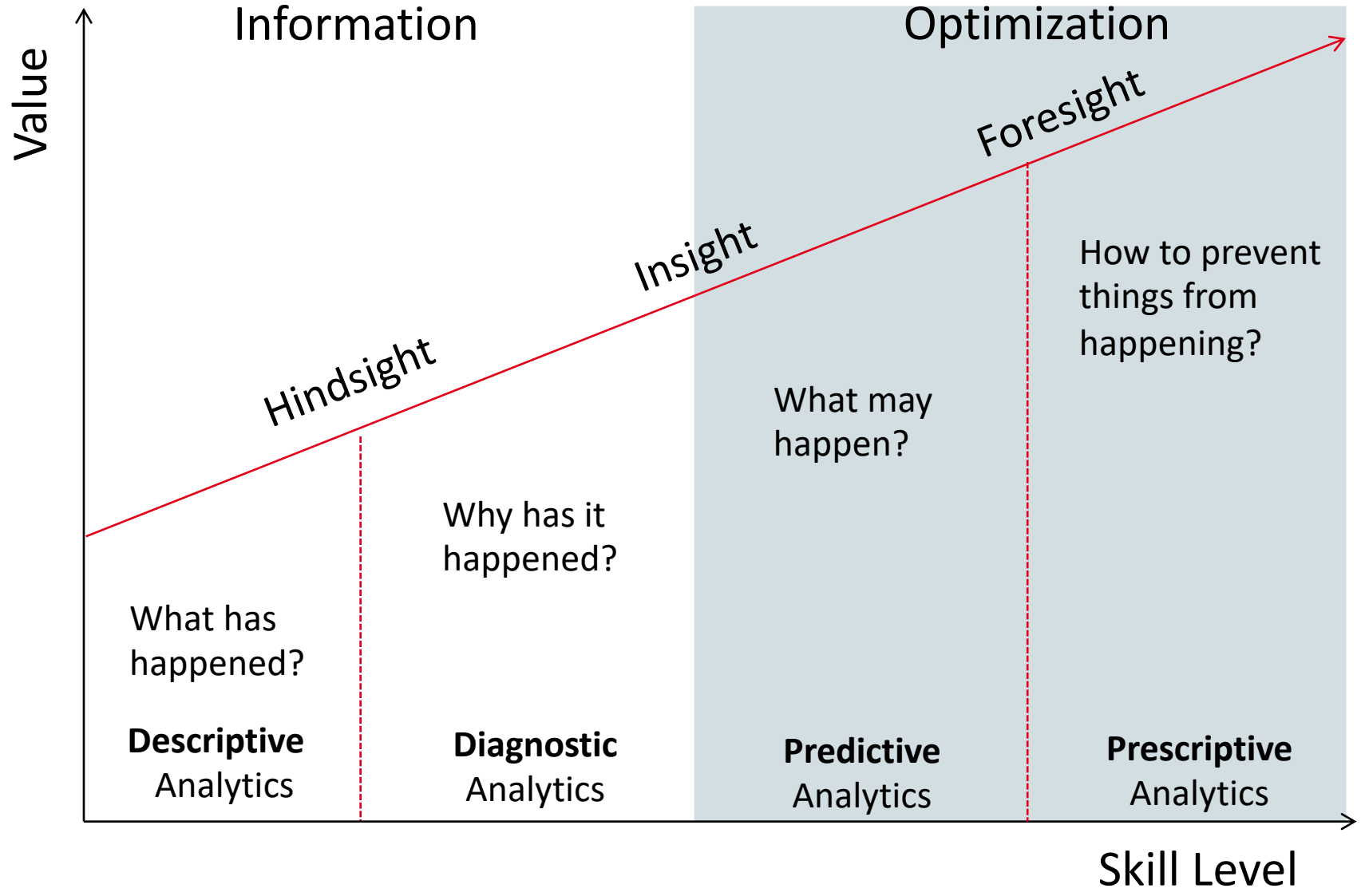


[PROMISE 2014, Zeller et al.]

When to Apply Data Analytics?

- ☑ Complex decisions with lots of variables and steps
- ☑ Simple decisions in which consistency is either desirable or required by law
  - E.g. non-discriminatory, transparency
- ☑ Places where the process need to be optimized
- ☑ Decisions in which connections, correlations, and their significance need to be understood
- ☑ When better forecasts or predictions are needed

Four Types of Analytics



Descriptive analytics

Standard and ad-hoc reporting and **dashboards**

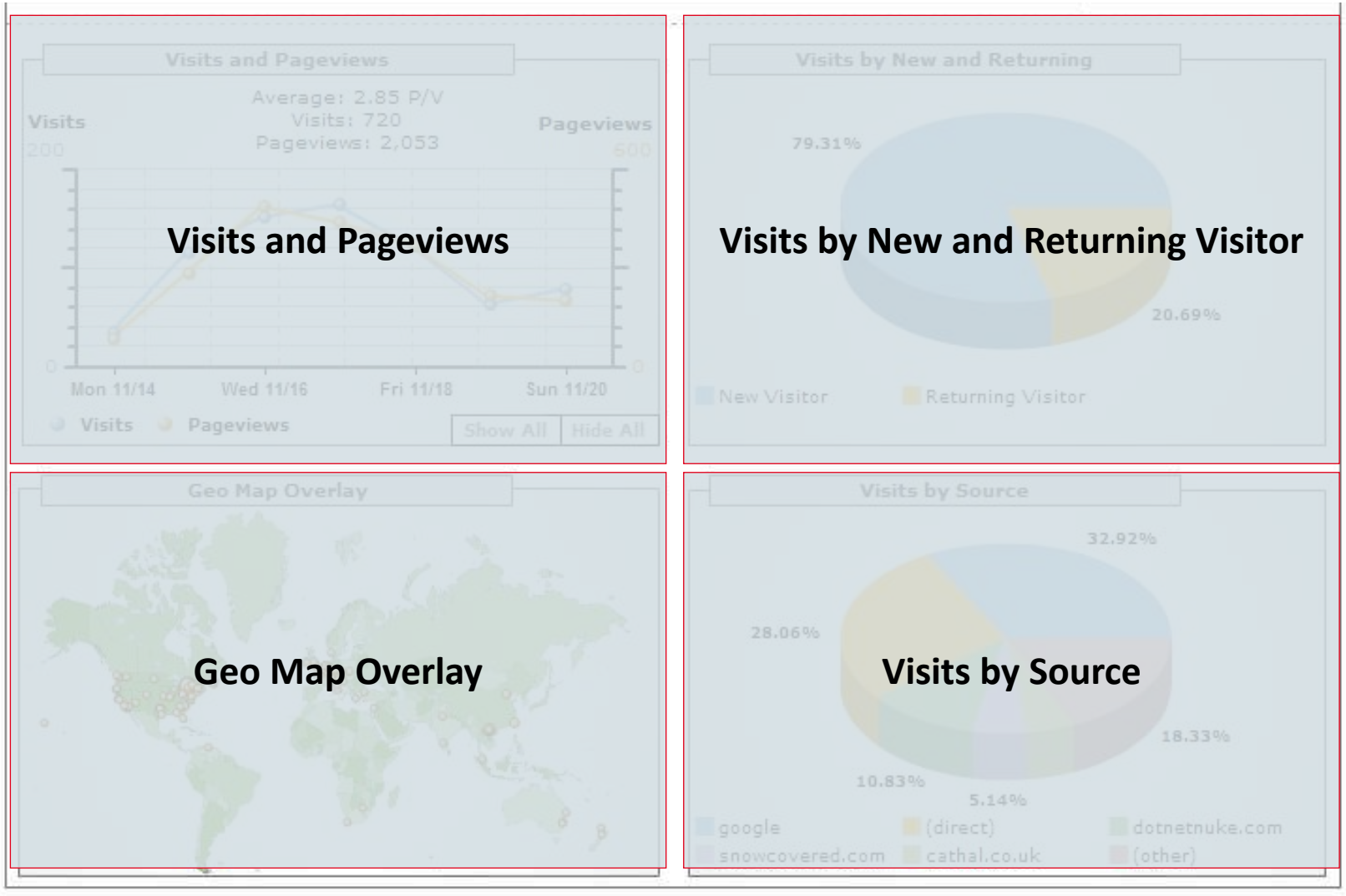
Focuses on the question

- What has happened?
- How does it compare to our plan?
- What is happening now?

Examples

- **Google Analytics**
- Omniture

Google analytics



Descriptive analytics for requirements

Example questions

- How does the system perform?
- How does it scale?
- What are the usage trends?
- What are the exceptions and errors?

Scenarios

- The application performance drops heavily with the increase of number of simultaneous users using a specific feature
  - Implications on the performance requirement
- Billing page is rarely reached from the shopping cart page compared to total views
  - Implications on the usability requirement

Diagnostic analytics

- A deeper analysis on the cause of identified phenomena
- Focuses on the question
  - Why has it happened?
- Context information needed

Challenges

- Laborious work
- More data and tools needed to get some insights

Diagnostic analytics for requirements

Identify issues and by what they are caused

- What is exactly a problem?
  Why is it a problem?

Example

- Problem: Many users abandon the purchase page
- Cause (assumption): Sign up enforcement
- Implication: Usability on registration and authentication need to be improved

Predictive analytics

Focuses on the question

- What may happen?

Data needed

- Contextual Data
- Interaction Data

Techniques needed

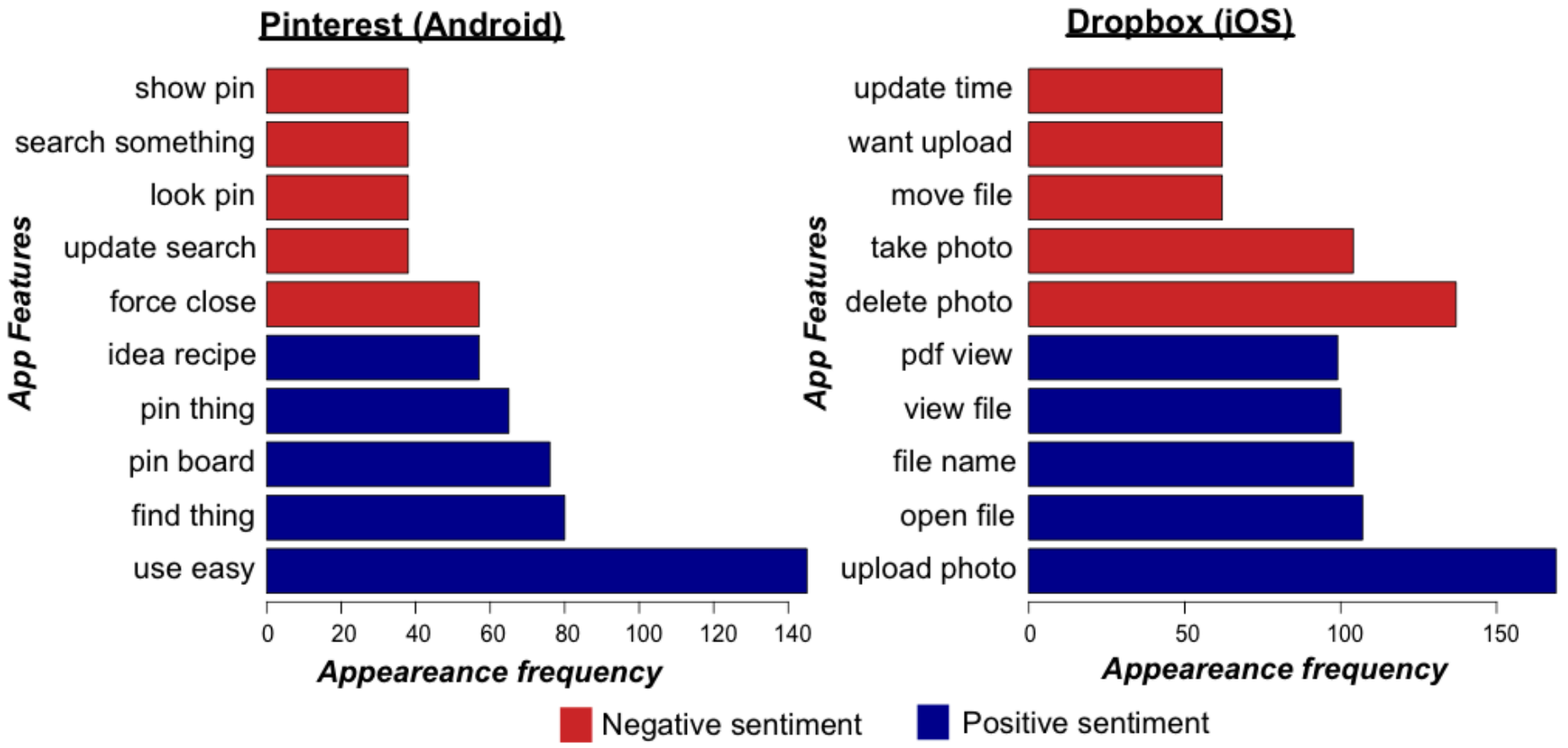
- Advanced statistics
- Modeling
- Data mining
- Machine learning

Predictive analytics for requirements

Example questions

- Who are my best customers and what is the best way to target them?
- Who are the lead users of the application and what is the best way to target them?
- Which features need to be improved in the next released?
- Which users will most likely protest from changing this feature?

Predictive analytics: example using sentiment analysis



Extracted features from Pinterest and Dropbox apps.
Positive features are represented in blue, negative in red

[Guzman & Maalej, 2014]

Prescriptive analytics

Focuses on the question

- How to prevent things from happening?
- How to make sure that specific things happens as we want them to happen?

Data needed

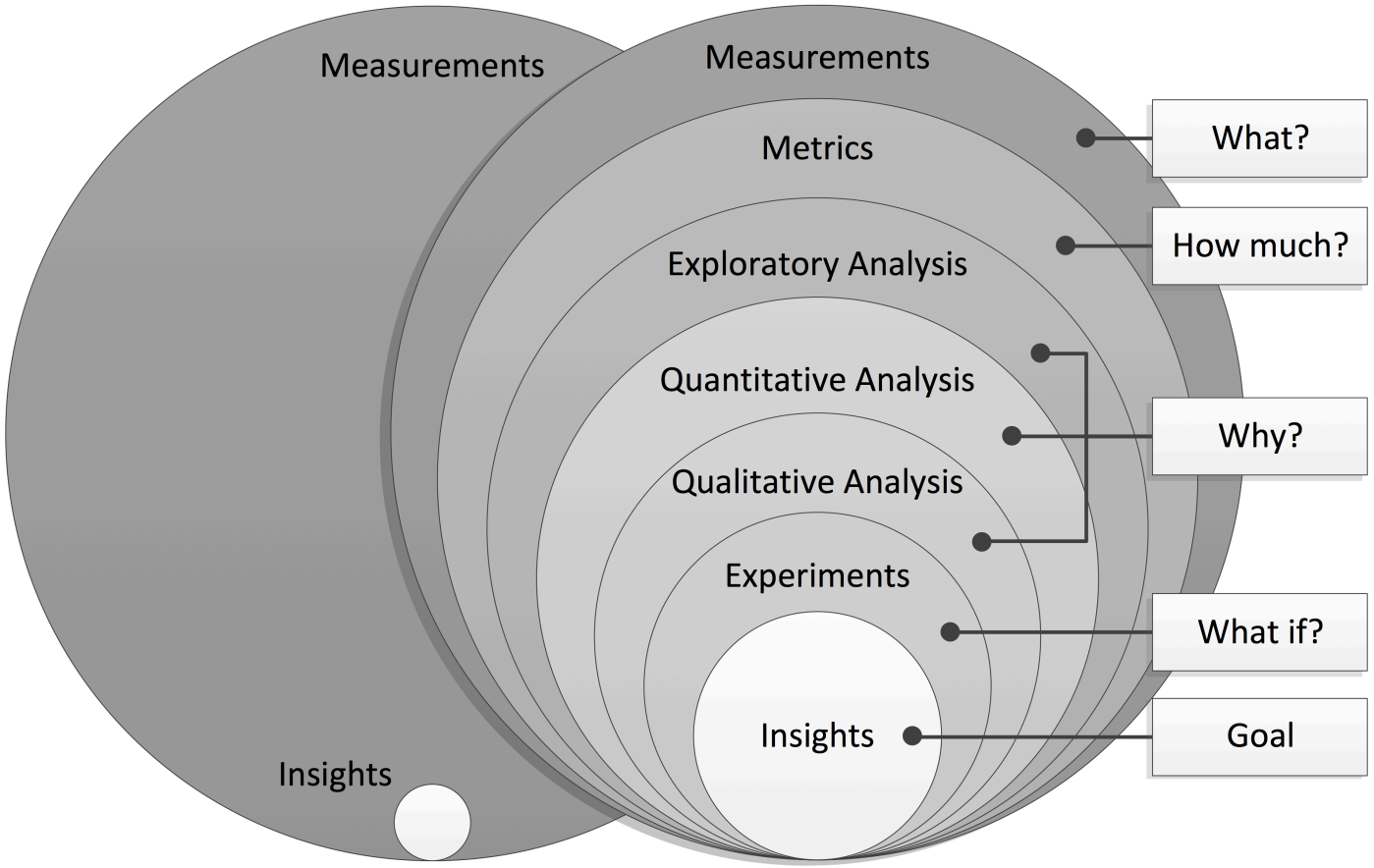
- Contextual Data
- Interaction Data
- Predicted Data

Techniques Needed

- Advanced statistics
- Modeling
- Data mining
- Machine learning

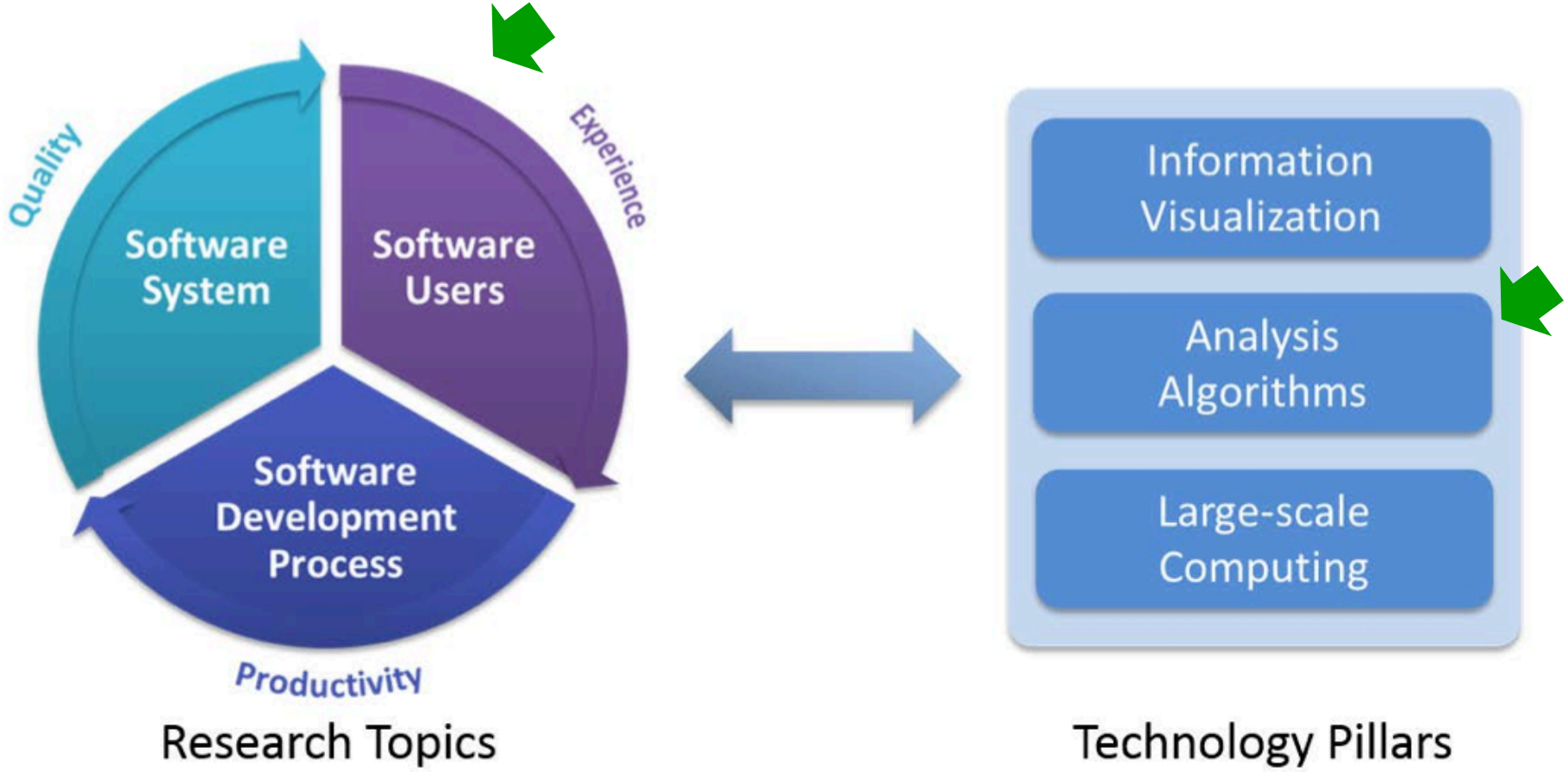
Prescriptive analytics for requirements

Paradigm of analytics



[Raymond P.L. Buse et al., 2011]

Software analytics in practice



[Zhang et al. 2013]

Where to start?

Questions to begin with

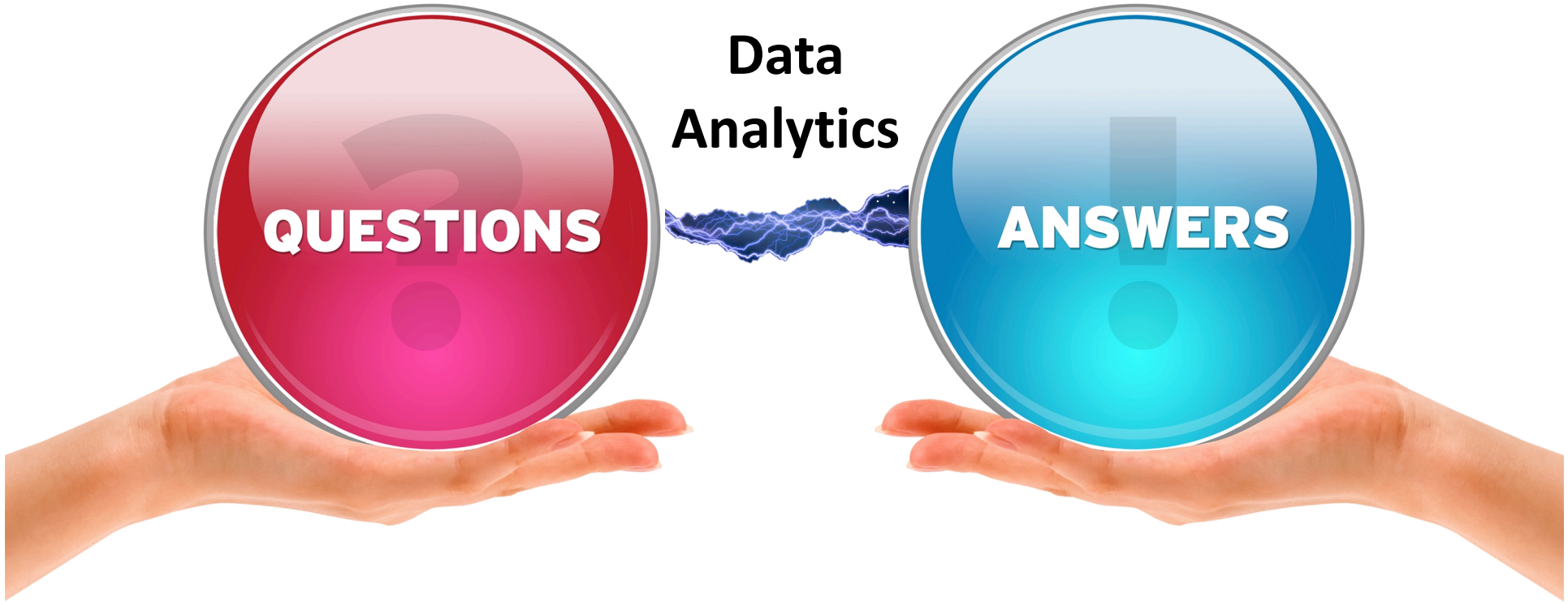
- Who asks the question?
  - Users, Developers, Managers, etc.
- Which questions need to be answered?
- What data sources are needed?
  - Usage logs, Runtime traces, System events, etc.
- What tools and techniques are needed?
  - Data mining, Machine Learning, Sentiment Analysis, etc.

Who asks the questions?

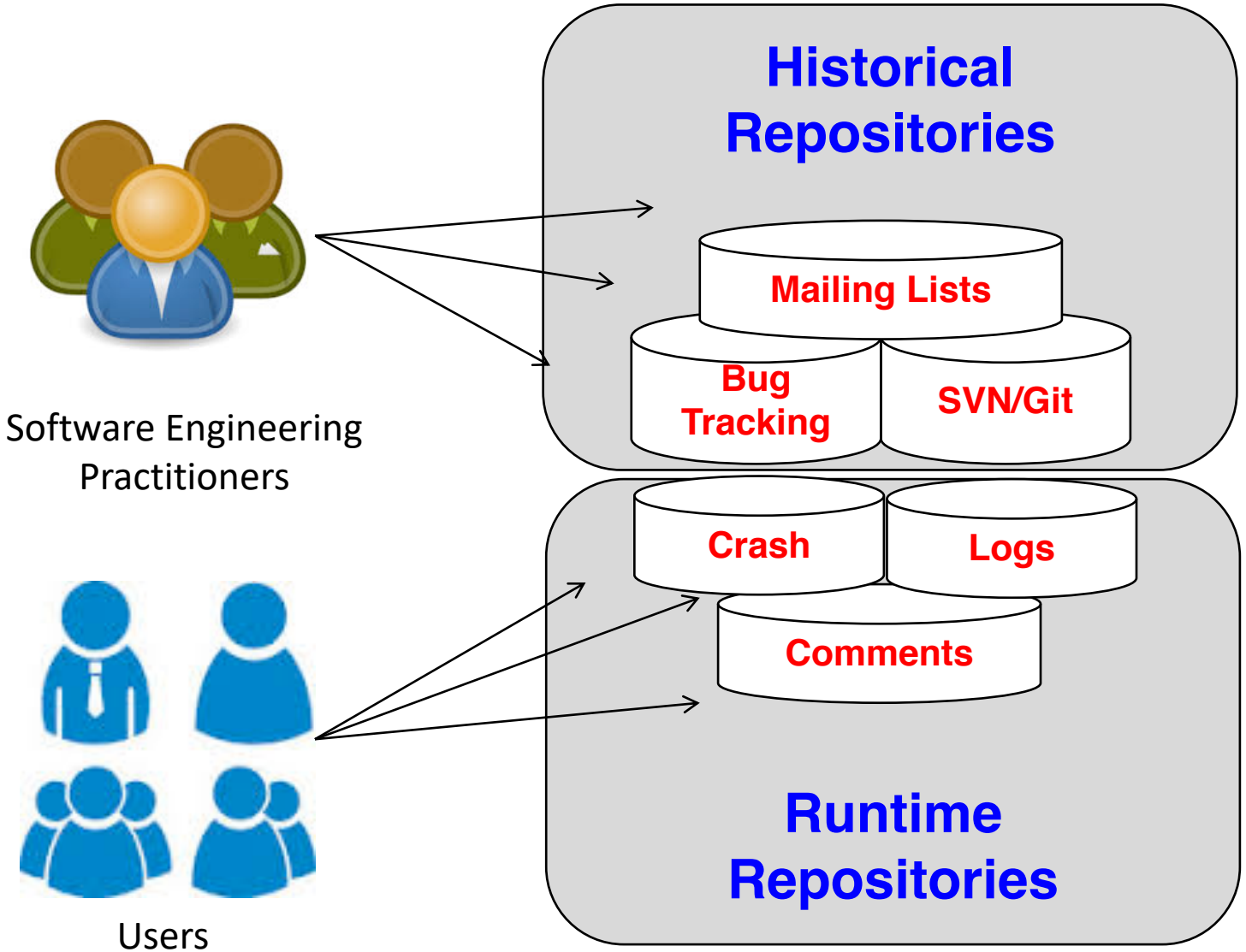
Stakeholders

- Developer
- Manager
- Tester
- Dev. Lead
- Test Lead
- Researcher
- Users
- …

Which question need to be answered?



What data sources are needed?



Guidelines for analytics

- Make easy to use visualization
- Be precise
- Measure many artifacts with many indicators
- Identify important & unusual items automatically
- Relate activity to features/areas
- Focus on past & present over future

[Buse & Zimmermann]

Data analytics best practice

Build tools for frequent questions

Use data scientists for infrequent questions

Outline

1 Software Analytics

2 **Usage Data**

3 Analytics Techniques

What is usage data?

Data about the users while using the system as well as their context

Examples

- Interaction data
  - A record of the actions taken by a user with an application. E.g. clicks, navigation history, comments, shares, likes…
- Context data
  - User context, task context, physical context, thematic context, Meta data etc.

What is context?

Who is the user?	User Context
What does the user do?	Task Context
Where is the user?	Physical Context
What does the user write?	Thematic Context
With whom does the user communicate?	Communication Context

Types of context data: wine-drinking example

1. **Environmental Data (Sensor-Data)**

2. User (Inter)action

3. Metadata

4. Content

1. Sensor data of a modern smartphone

GPS
Location and altitude

Room Temperature Sensor
Thermometer

Accelerometer

Motion detection (tip over, tend..)

Light Sensor

Ambient Light and Brightness

Gyroscope
Rotation and speed recognition

Compass
Magnetisches Feld der Umgebung

1. Sensor Data of a Modern Smartphone

Humidity Sensor

Relative Humidity

Pulssensor

Health apps (e.g. Jogging, Fitness)

Fingerprint Sensor

Authentification

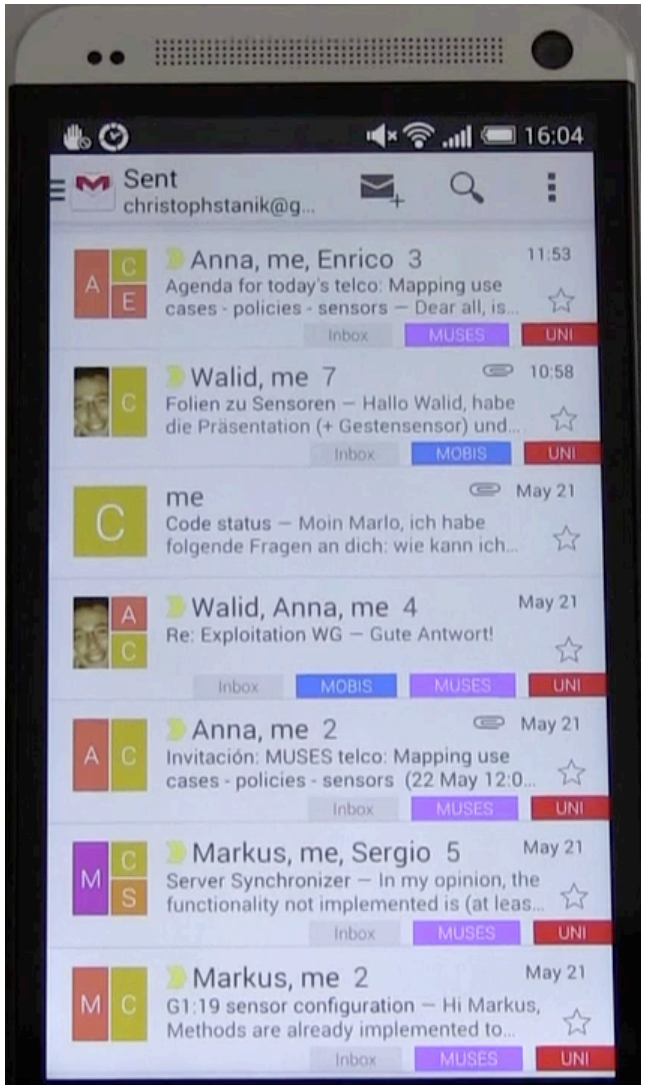
Air Pressure Sensor

Proximity Sensor

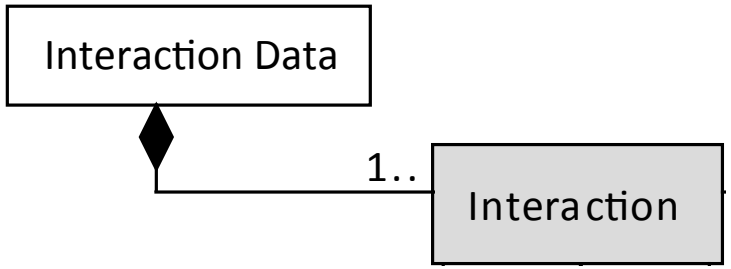
Vicinity of an object relative to the display

2. Interaction data

- **Scroll up**
- **Click on list item**
- **Swipe to left**
- **Click “Answer”**
- **Use keyboad**
- **Click “Send”**

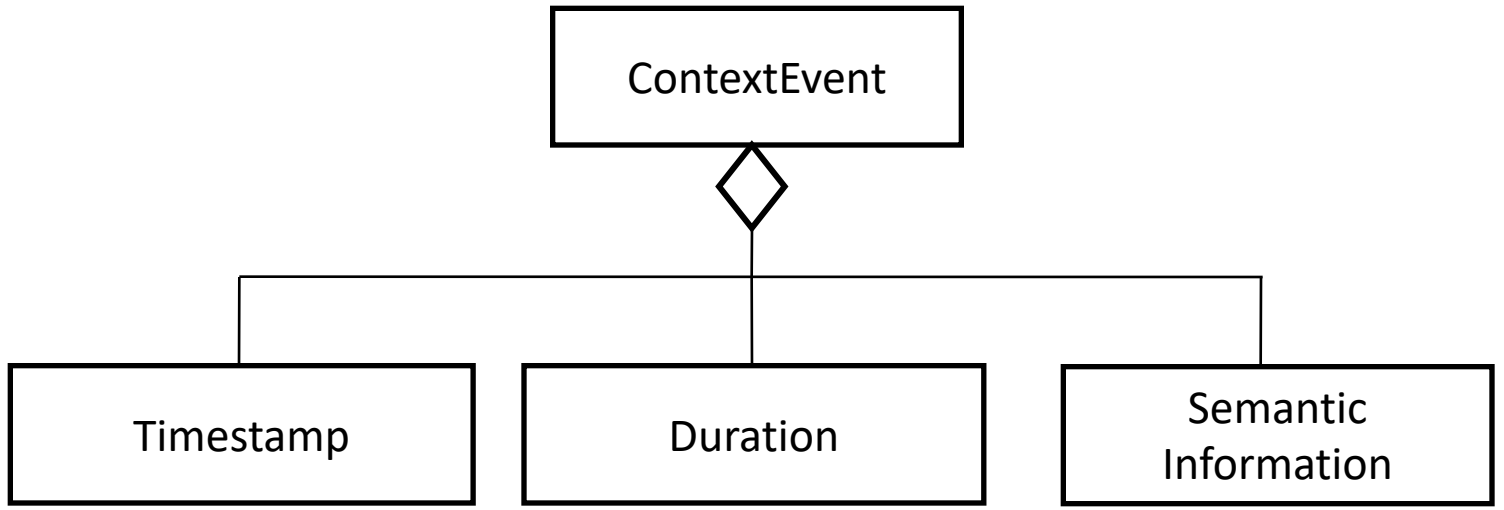


What is interaction data?



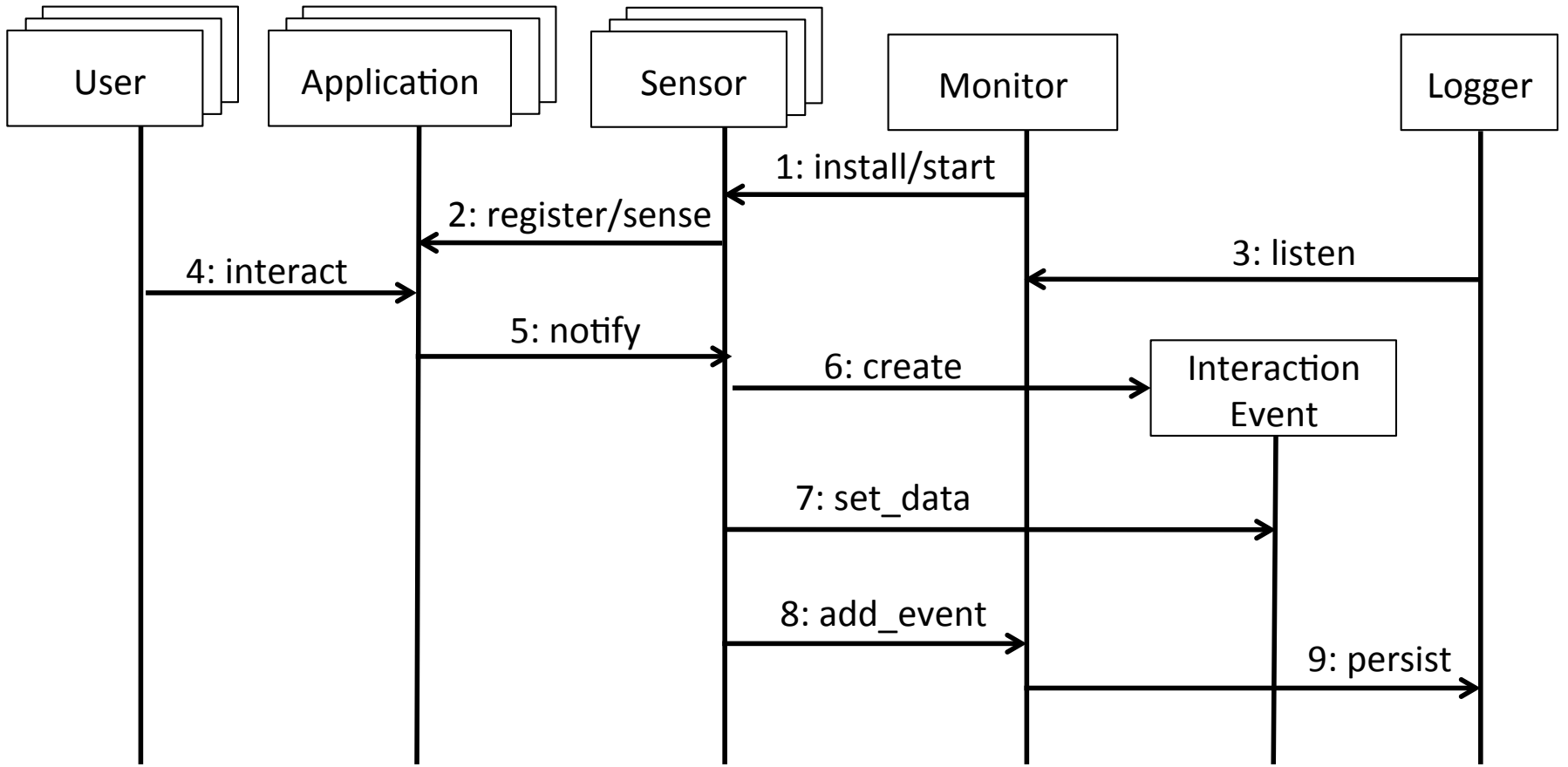
- **Interaction** : Actions taken by a user, such as clicks and changes
- **Artifacts** : Entities users are working with
- **Context** : Additional information, such as task or phases

Structure of interaction events




```
<pre:e1 rdf:type interaction:JavaElementChange />
<pre:e1 interaction:hasTimeStamp 1222002002 />
<pre:e1 interaction:hasDuration 200 />
<pre:e1 interaction:concerns pre:java?name=myMethod />
<pre:java?name=myMethod rdf:type artefact:Method />
<pre:java?name=myMethod artefact:partOf pre:java?name=myProject.myClass />
```

Collecting interaction data



Processing interaction data

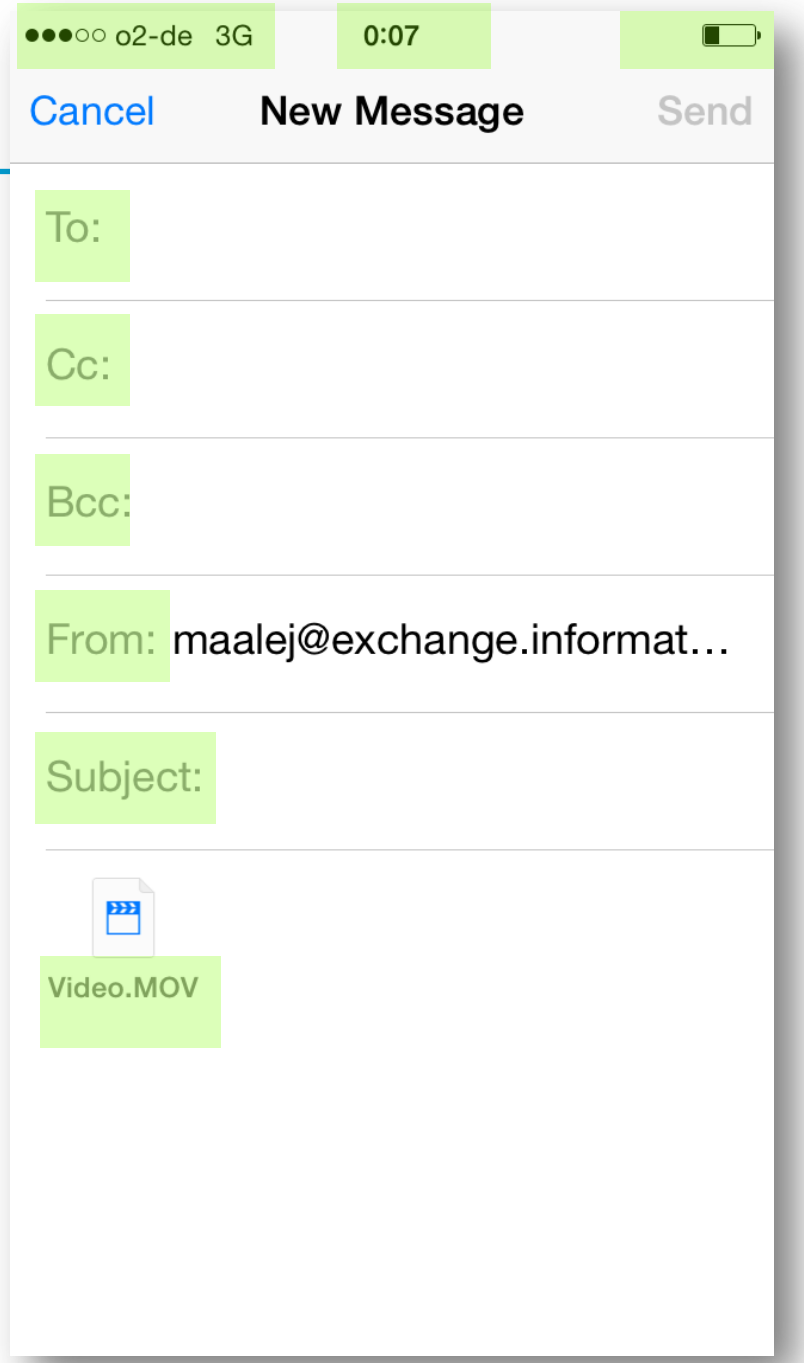
Sessionization of Interaction Events

Filtering of Events

Aggregation of Events & Inference of Context

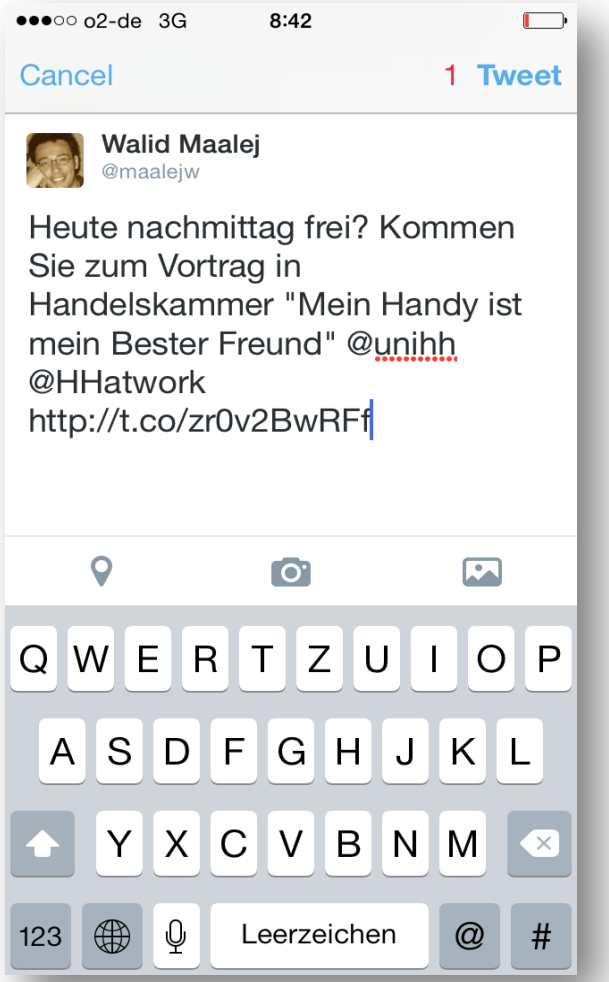
3. Metadata

- E.g. the sender and receiver of a message but not the message itself
- File name but not the file content
- File type, creator, change time
- Time

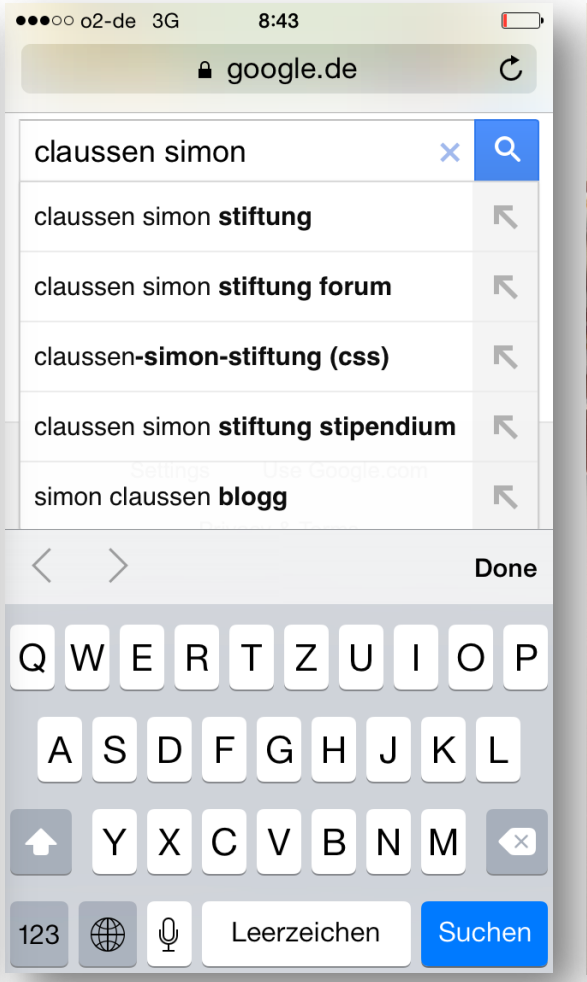


4. Content

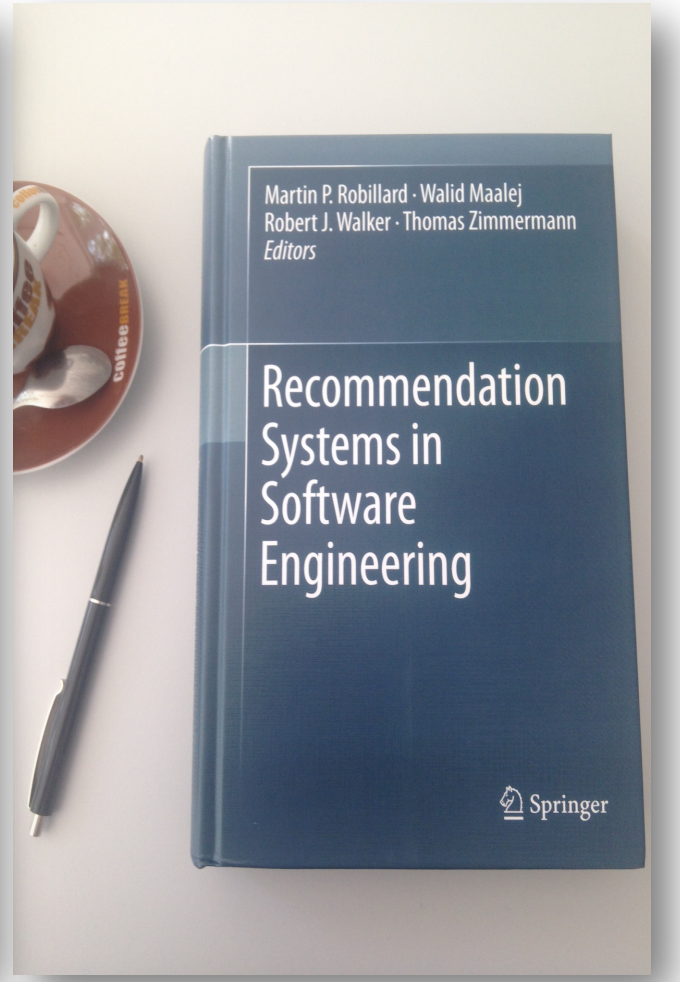
Content of a message or a file



Search keyword or search results



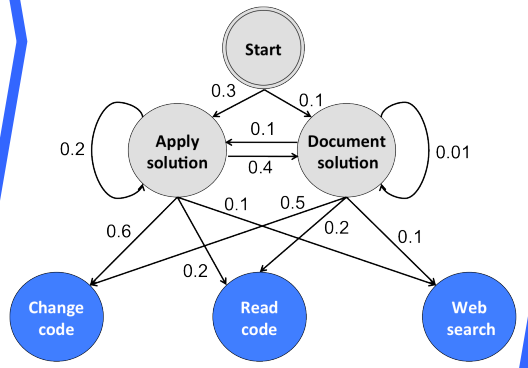
Photo, video ,or other multimedia



Three steps to context sensitivity

Context Monitoring
Collecting Data by Instrumentation

Context Processing
Situation Recognition



Proactiveness
Suitable for the actual situation

[Roehm & Maalej ICSE'12] [Maalej et al. RSSE Buch 2014]

Pitfalls of interaction data analytics

- Each processing step may introduce noise
- Over-processing can negatively influence the precision and recall of data mining algorithms
- Keep original data as long as possible intact, since it is hard to predict when information will be needed
- Each interaction data collection tool has issues. These should be known to avoid false interpretation
- Users inactivity is hard to assess, as it may simply be due to missing interaction data

Challenges

- Efficient, integrated, non-intrusive instrumentation
- Representation of interaction data
- Sessionization of interaction events
- Context prediction and comparison
- Privacy protection
- Ethics (collection, processing, sharing, selling, …)

Outline

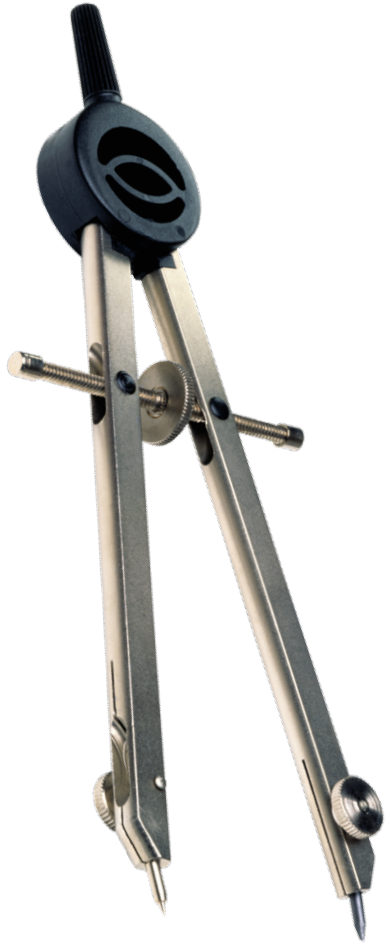
1 Software Analytics

2 Usage Data

3 **Analytics Techniques**

Analytics techniques

- Frequent item sets
- Association rules and frequent patterns
- Classification
- Clustering
- Natural Language Processing
- Visualization
- …



Frequent item sets mining

Discovering frequent item sets in a large dataset

Feature / User	F1	F2	F3	F4
U1	X		X	
U2		X		
U3				X
U4		X	X	X
U5		X	X	
U6		X	X	
U7	X	X	X	X
U8	X		X	
U9	X	X	X	
U10	X	X	X	

Example of market baskets

- B1 = {m,c,b} B2 = {m, p, j} B3 = {m,b} B4 = {c,j}
  B5 = {m,p,b} B6 = {m,c,b,j} B7 = {c,b,j} B8 = {b,c}
- Suppose a minimum support of 3
- Frequent item sets:

{m:5}	{{m,b}:4}
{c:5}	{{c,b}:4}
{b:6}	{{j,c}:3}
{j:4}	

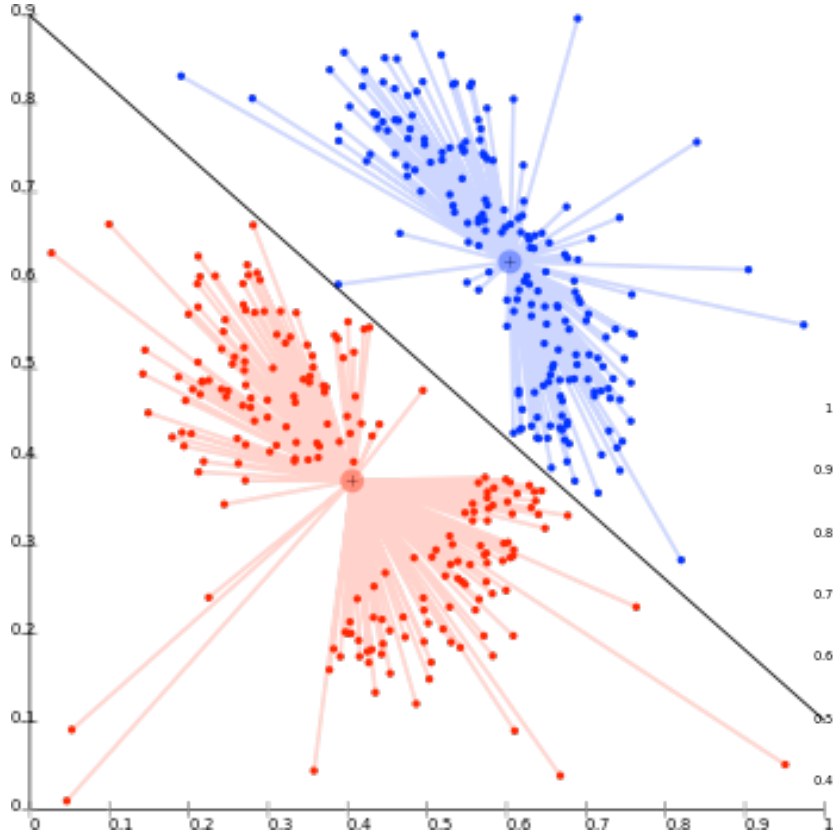
Association rules mining

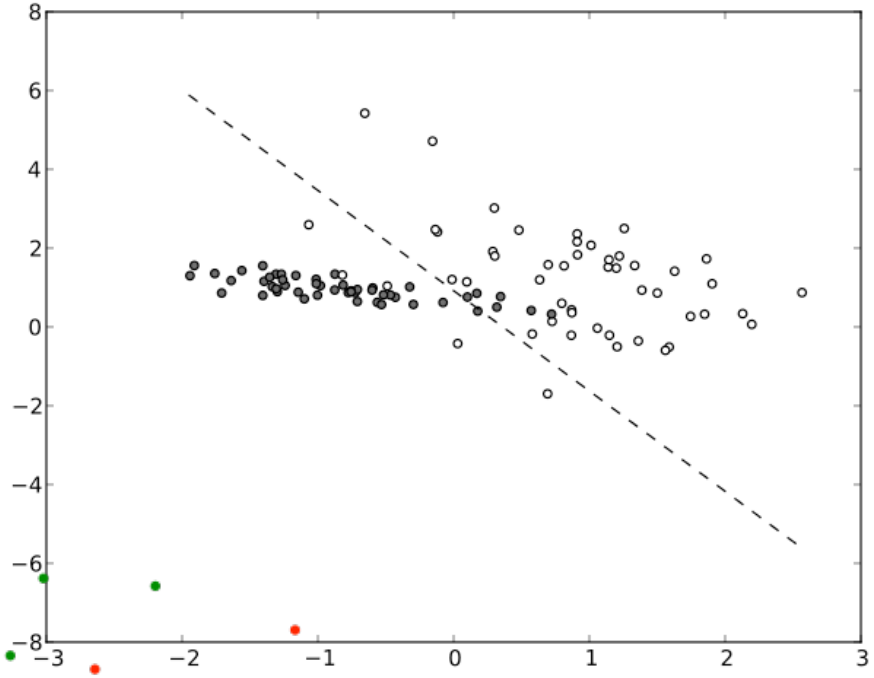
- A special case of frequent item sets
- Discovering interesting relations between data items
- X → Y where X and Y are subsets of the item sets
- If the dataset contains the items i_1, i_2…,i_k then it is likely to contain i_j
- **Confidence** of the rule = the % of finding Y given X
- Algorithms: Apriori, Eclat and FP-Growth
- Another type: Sequential Pattern Mining

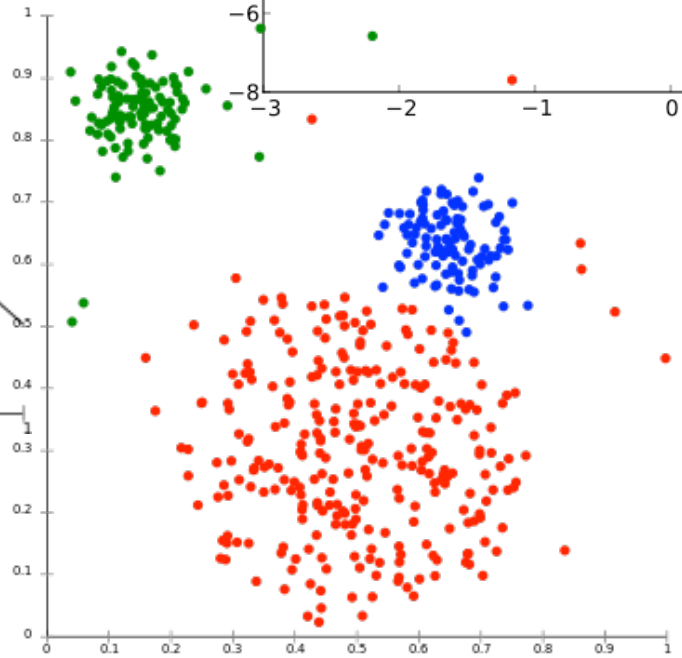
Clustering

- Process of organizing objects into groups whose members are **similar** in some way.
- An unsupervised learning technique
  - It does not rely on predefined classes and manually-labeled training set
- Many algorithms and methods, e.g.:
  - Similarity (distance) measure
  - Hierarchical (agglomerative vs. divisive/top-down)
  - Partitional (e.g. K-Means Clustering)

Clustering examples





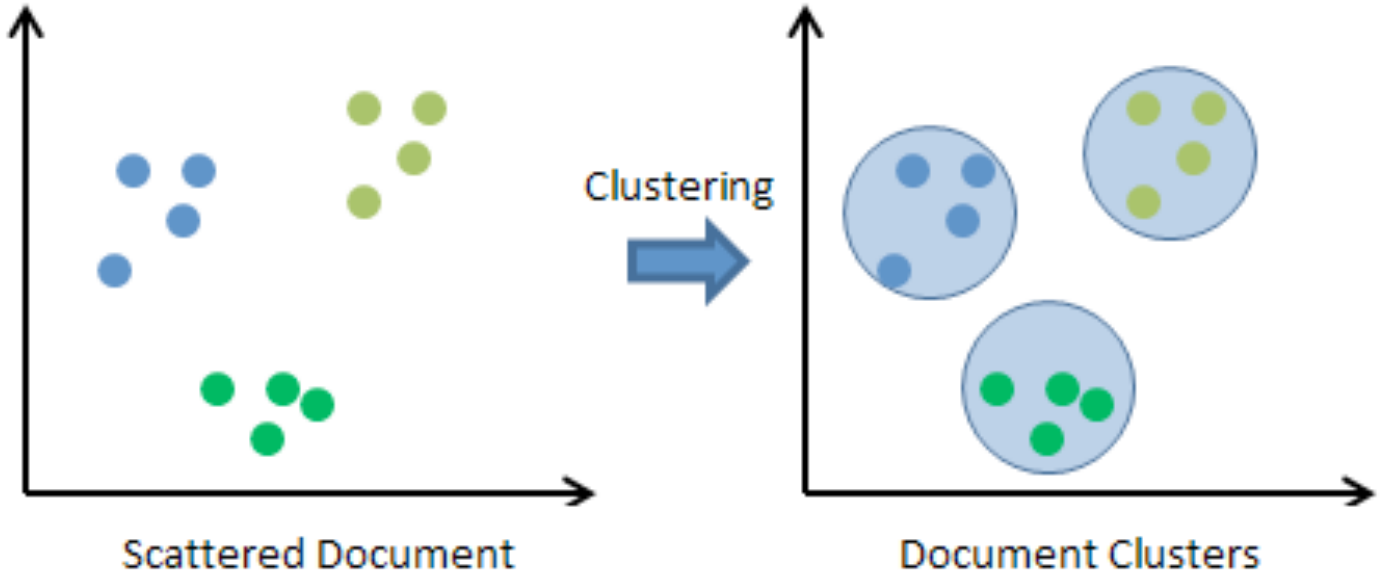


What does similar mean?

We know it when we see it..

Example: identifying homogenous groups of stakeholders

Grouping similar or related requirements



Example of clustering variable: keyword vectors, document owners, time of creation, related subsystem, estimated complexity, dependent requirements…

Natural language processing

Goal: to process natural language texts and to make their information analyzable to computer systems

Techniques:

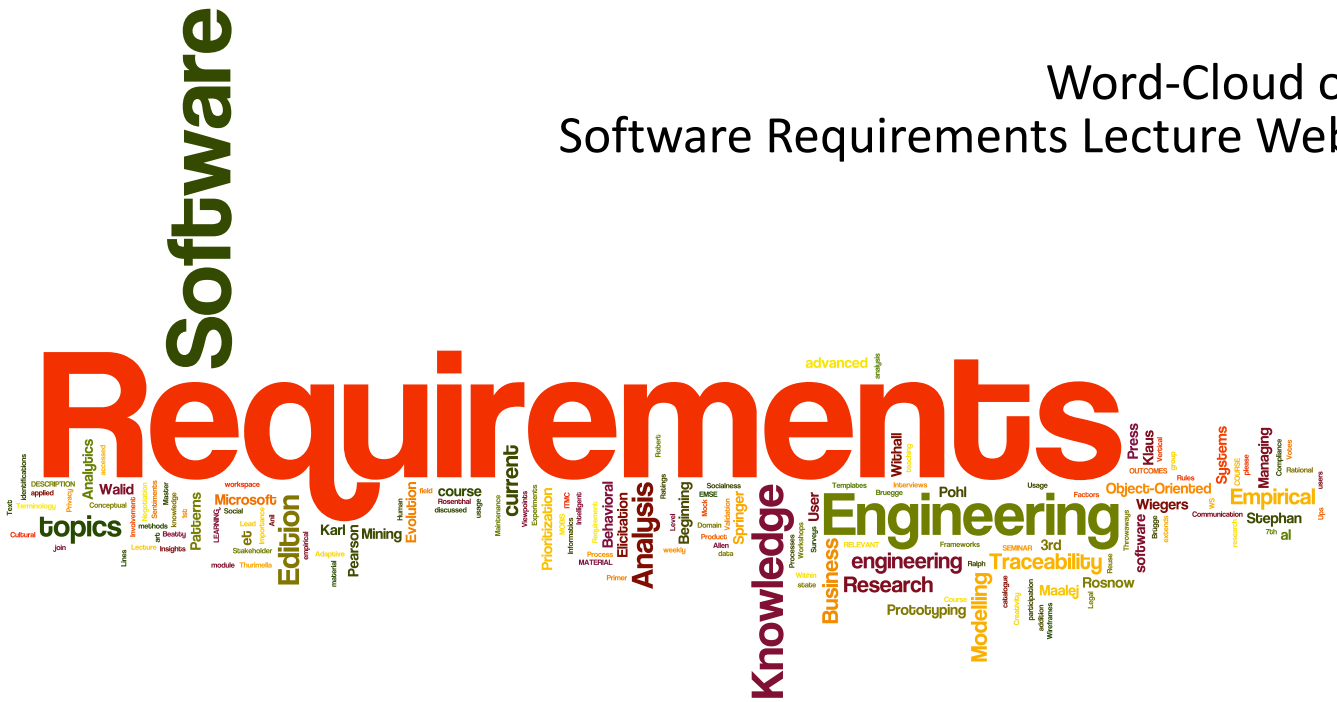
- Indexing
- Lexical Sentiment Analysis
- Part of Speech Tagging (POST), Pattern Analysis
- Collocations
- Latent Dirichlet allocation (LDA)

Data visualization

- Data visualization techniques are used to present analytical results visually
  - Many patterns in data can only be recognized by stakeholders
  - Meaning of data can be understood easier when presented as charts or graphs
- Two types of data visualization
  - **Static** visualization
  - **Interactive** visualization
    - A stakeholder is able to interact with a visualization
    - E.g. zoom in, pop up additional information, etc.

Example: word cloud

- Shows more important words bigger in size
- Importance of a word is usually defined as its frequency of occurrence



Word-Cloud of the
Software Requirements Lecture Website:

Many software tools for analytics

- Weka (machine learning framework, Java, GNU-PL)
- NLTK (natural language processing, Python, Apache)
- Mahoot (scalable machine learning, Java, Apache)
- Lucene (indexing and search, Java, Apache)
- Stanford Parser (NL Parser, Java, GNU-PL)
- R (statistic programming environment, R, GNU-PL)

Summary

1. Analytics allows to make decisions based on facts instead of gut feelings
2. 4 types of analytics (descriptive, diagnostic, predictive, and prescriptive) can support making decision about requirements
3. Usage data includes interaction data (clicks) and context data (sensor, metadata, content…) allow to understand the users
4. Analytics techniques from data mining and natural language processing allow to deal with big data