Software Construction

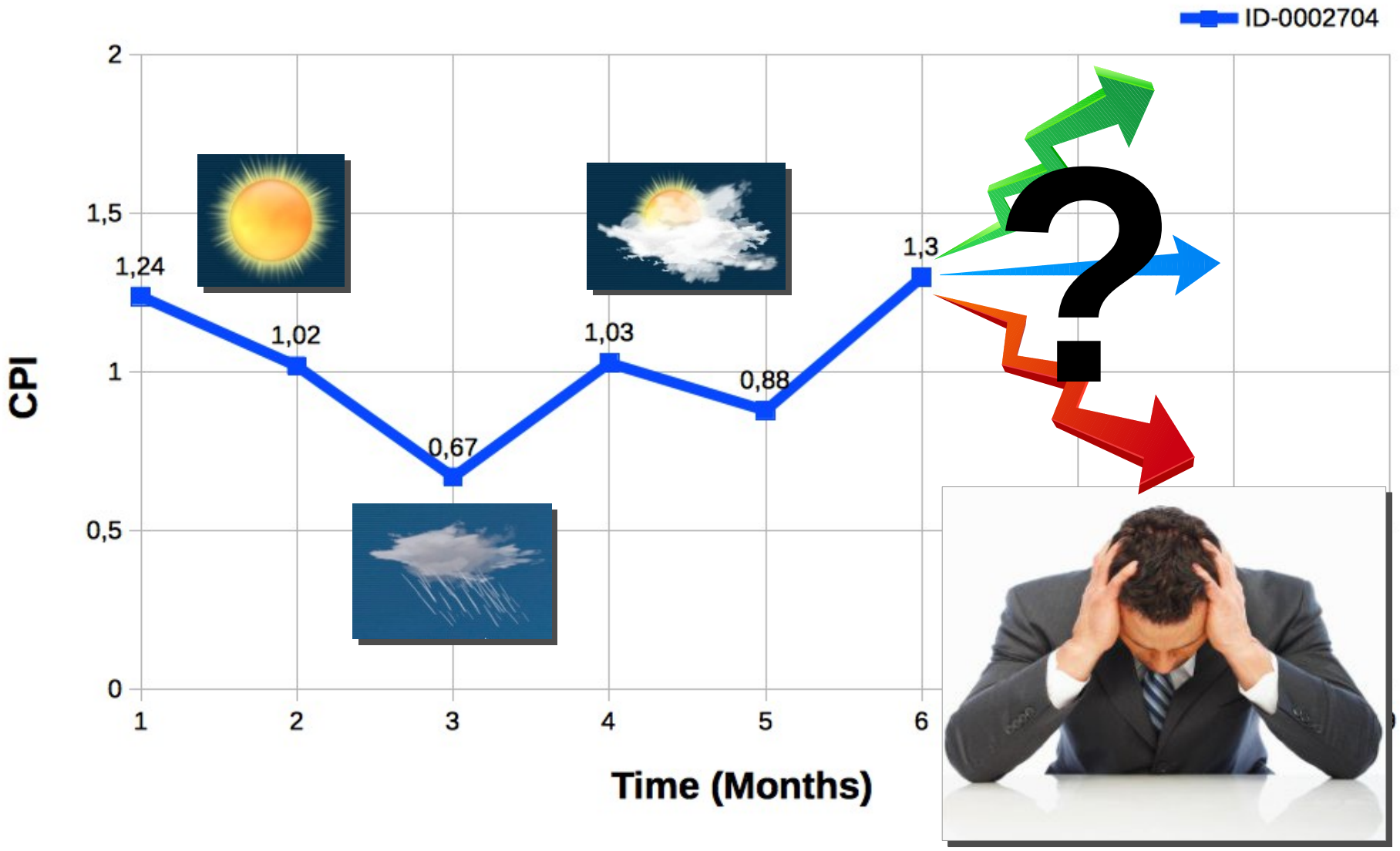
RWTH AACHEN UNIVERSITY

Werkzeugunterstützte Projektprognose

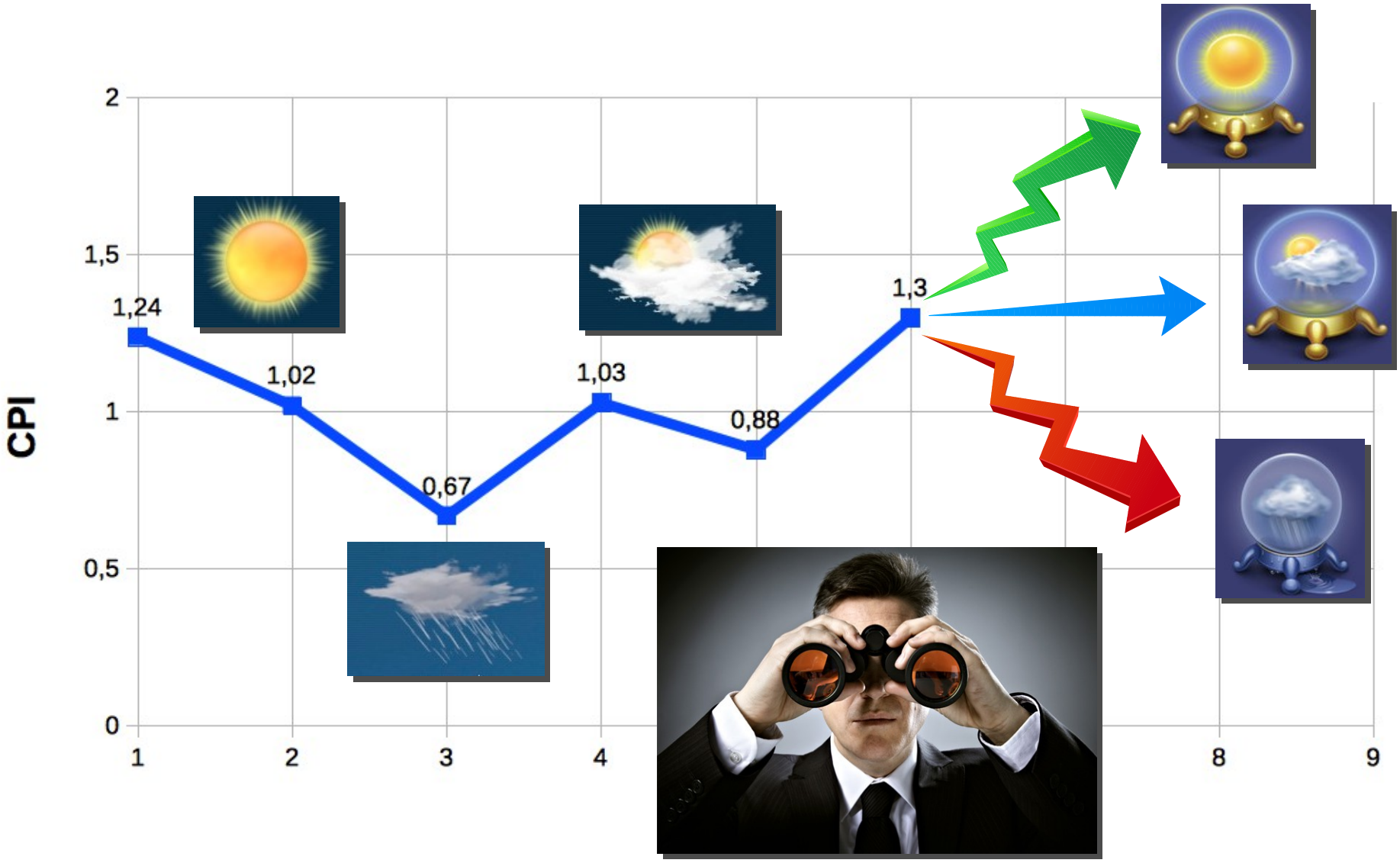
Elena Emelyanova
elena.emelyanova@rwth-aachen.de

26.06.2015

Motivation

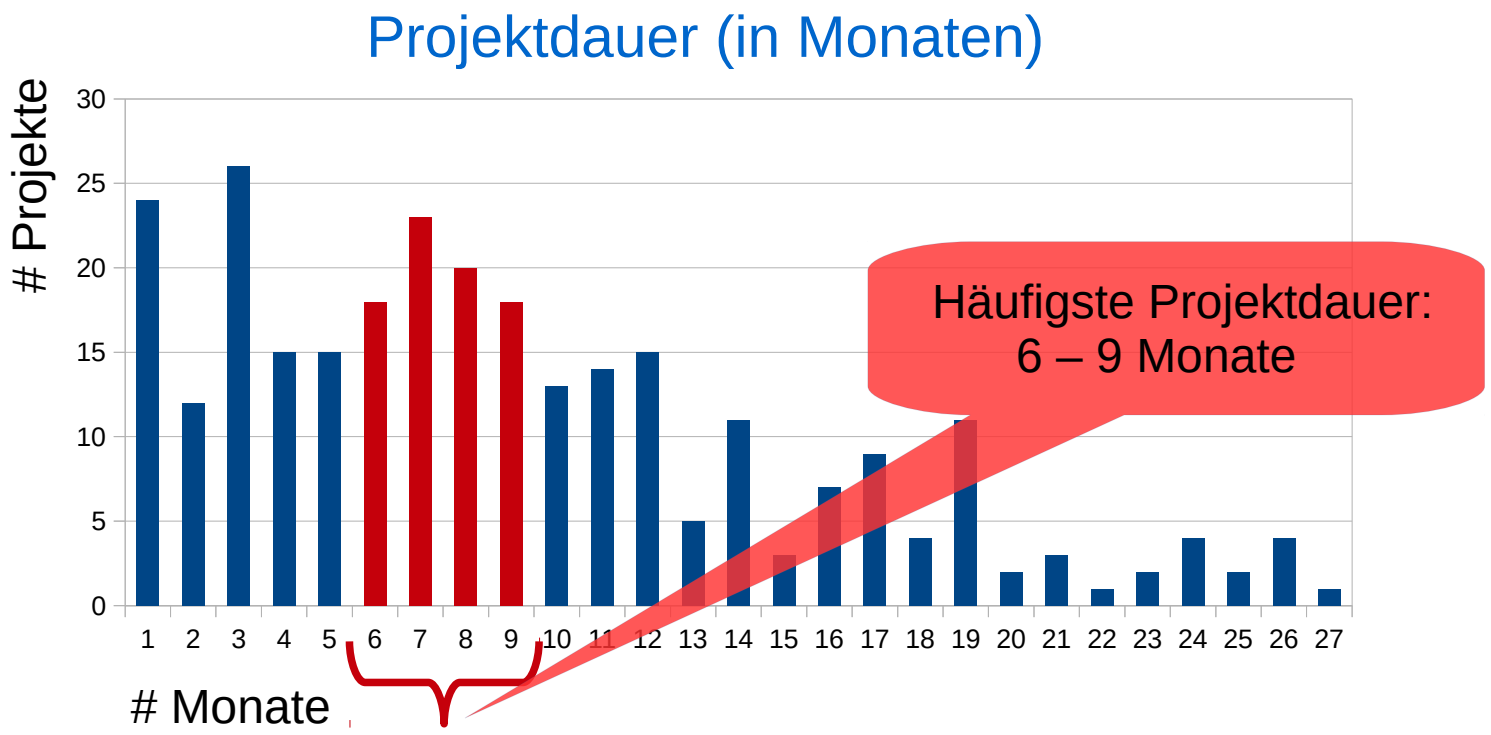

ID-0002704
CPI
Time (Months)
1,24
1,02
0,67
1,03
0,88
1,3
0
0,5
1
1,5
2
1
2
3
4
5
6

Motivation



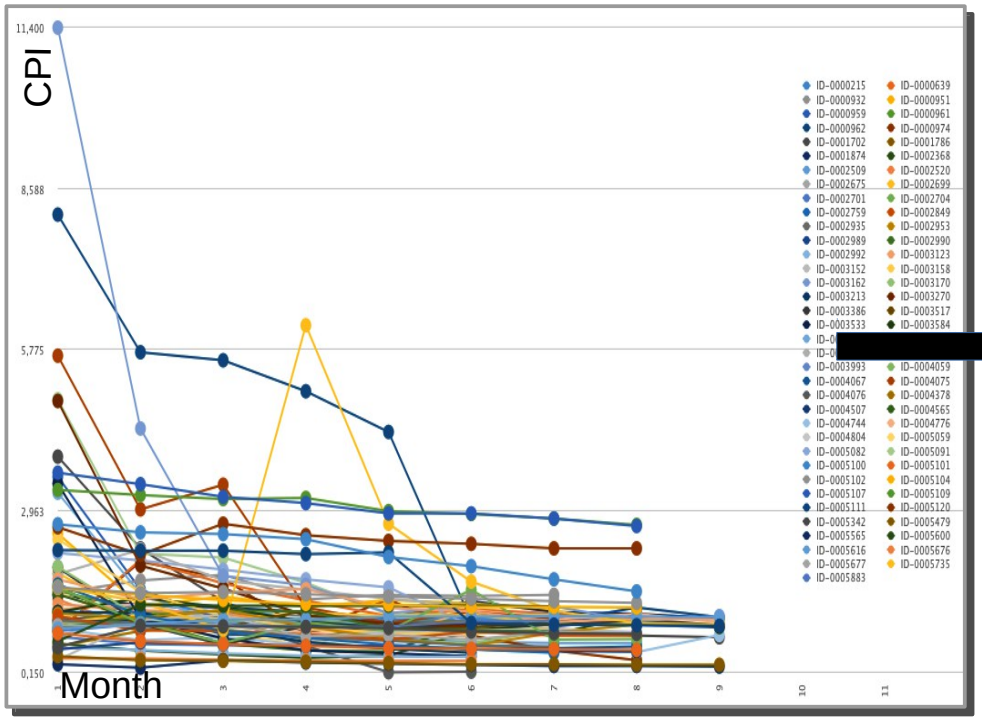
Motivation

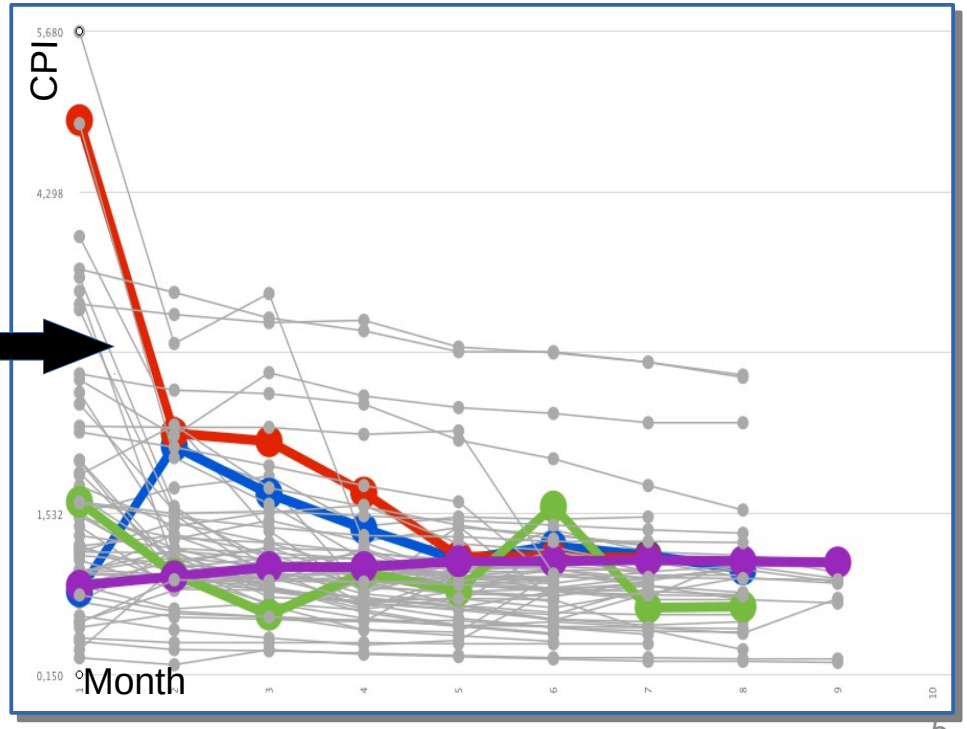
- Daten von externem Kooperationspartner:



Motivation

- Erste Versuch Muster zu identifizieren





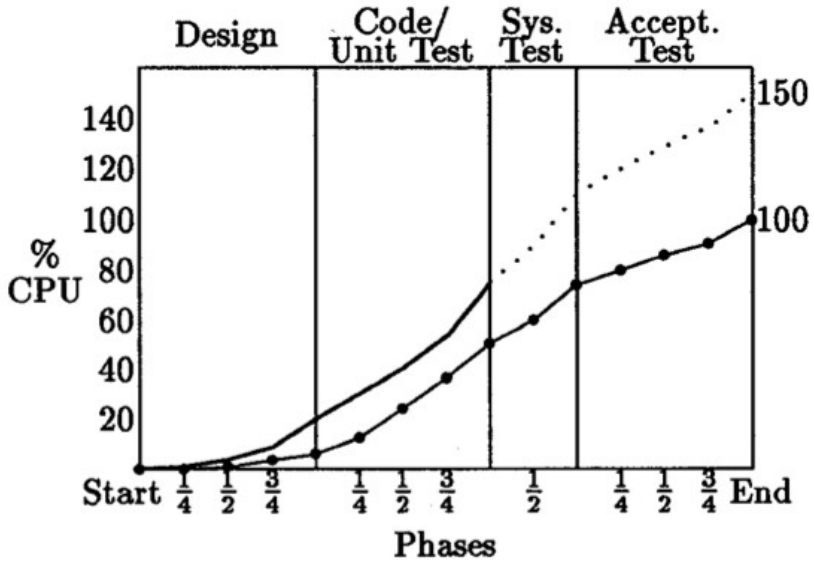
Motivation

Source [1]

Verwandte Arbeiten

- T. Vi Bach, 2014 [2]
  - **Ansatz:** Konzept des Frameworks zur Projektprognose
  - **Daten:** Historische Daten von großem IT-Unternehmen (CPI, SPI)
  - **Implementierung:** nur Konzept, keine Instrumentalisierung!
  - **Evaluation:** Fehlerrate über 100%

- M.V. Zelkowitz et al., 1993 [4]
  - **Ansatz:** Konstruktion des Prognose-modells für Software Projekte
  - **Daten:** über 100 NASA Projekten (LOC, Reported Errors, CPU, etc.)
  - **Implementierung:** keine Information!

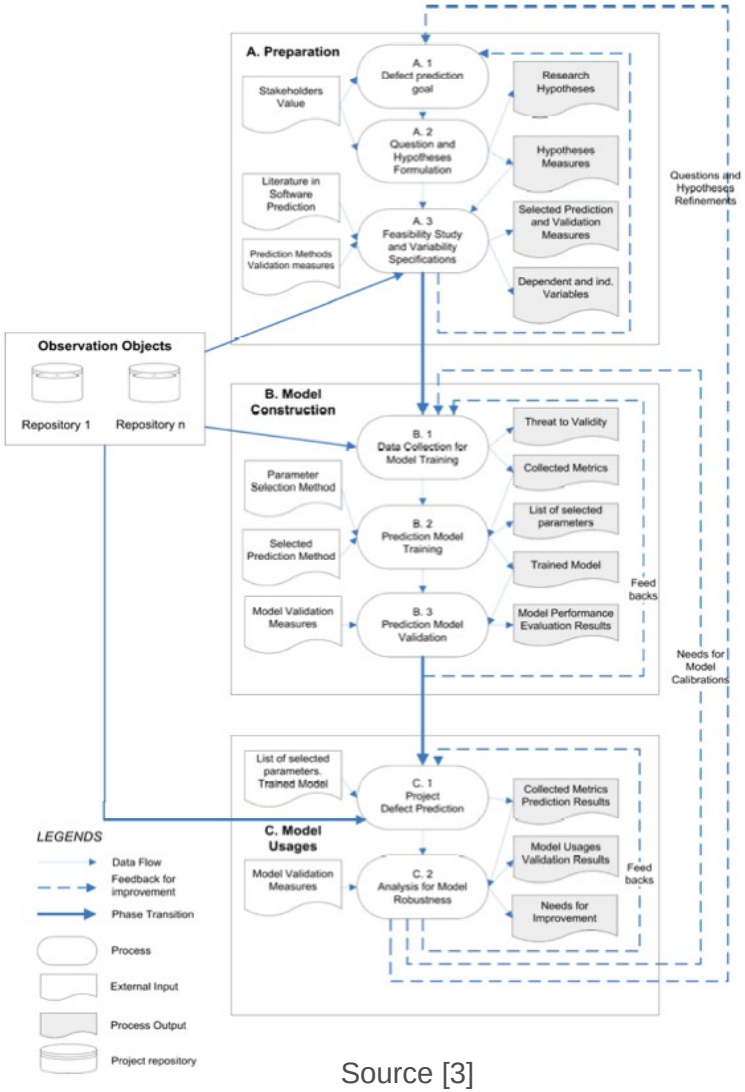


Vorhersage für CPU

Source [4]

Verwandte Arbeiten

- R. Ramler et al., 2008 [3] and 2009 [5]
  - **Ansatz:** Framework für Vorhersage von Software Defekten basierend auf Data Mining Methoden
  - **Daten:** 6 Versionen von Software System, ca.157 Komponenten / 639 KLOCs
  - **Implementierung:** Definiert 3 Phasen, aber keine Info über Instrumentalisierung!
  - **Evaluation:** 22% < Fehlerrate < 33%

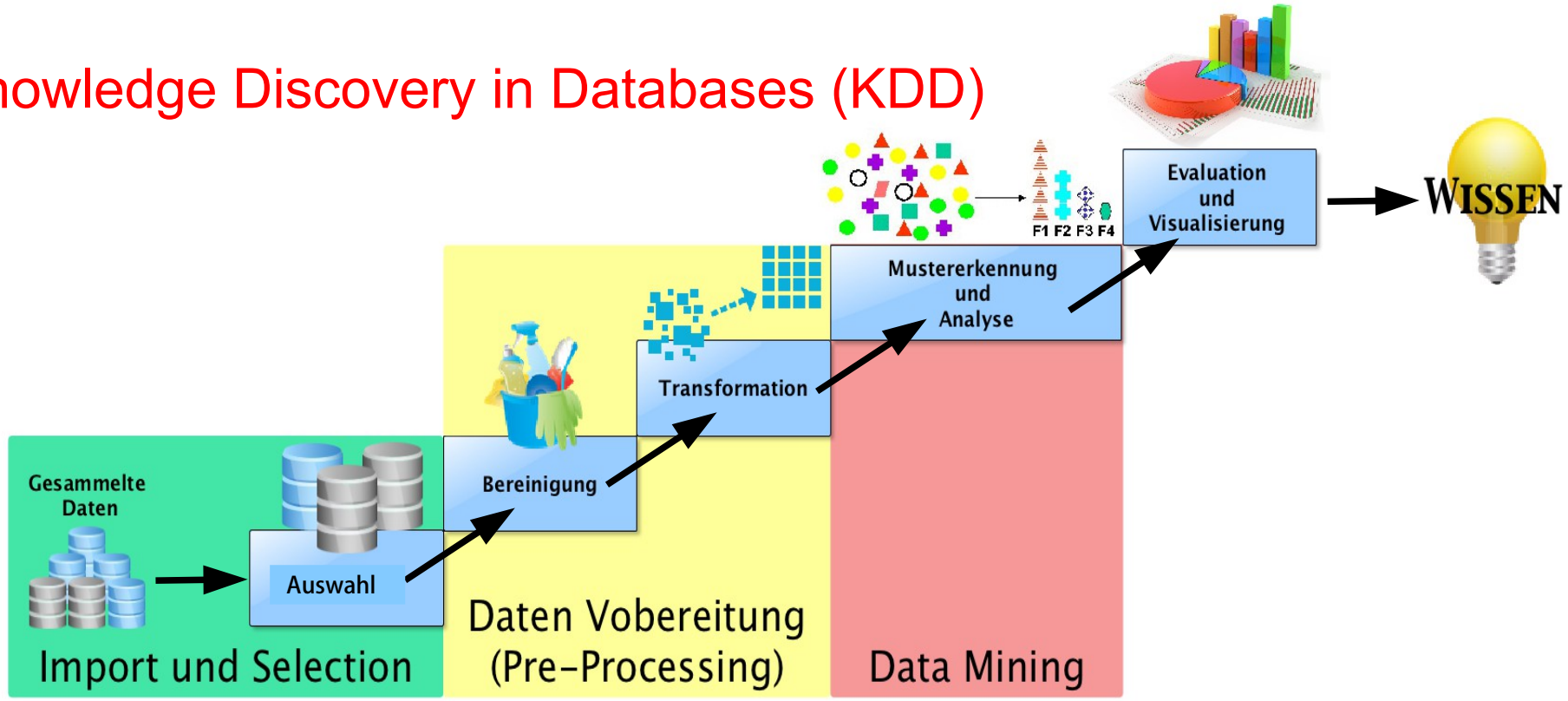


Source [3]

Data Mining Prozess

- Was ist Data Mining?
  - Entdecken und Extrahieren von Information
  - Finden von Muster / Wissensgewinnung

- Knowledge Discovery in Databases (KDD)



Data Mining

Cluster-Analyse (Clustering)

Klassifikation

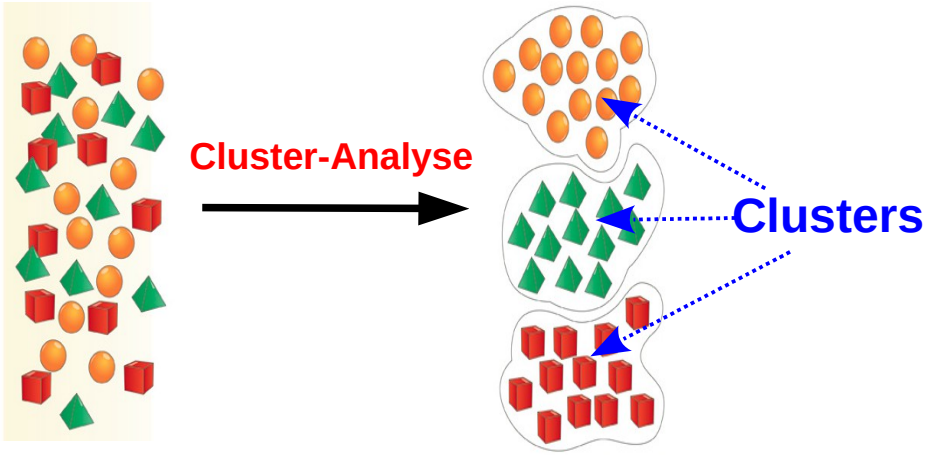
- Zusammenfassen von Objekten zu **Klassen**
- **Klasse** = Gruppe von Objekten mit einer logischen Eigenschaft (Regel)
- Benennung der Klassen

Data Mining

Cluster-Analyse (Clustering)

Klassifikation

- Entdeckung von **Ähnlichkietsstrukturen**
- **Cluster** = Gruppe von “ähnlichen” Objekten gemäß eines Ähnlichkeitsprinzips (Distanzmaß)

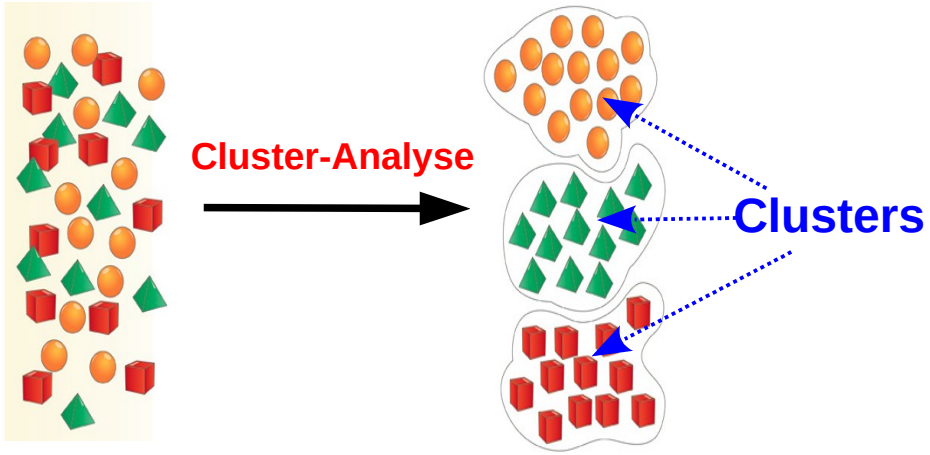


Cluster-Analyse (Clustering)

Klassifikation

Clustering-Parameter

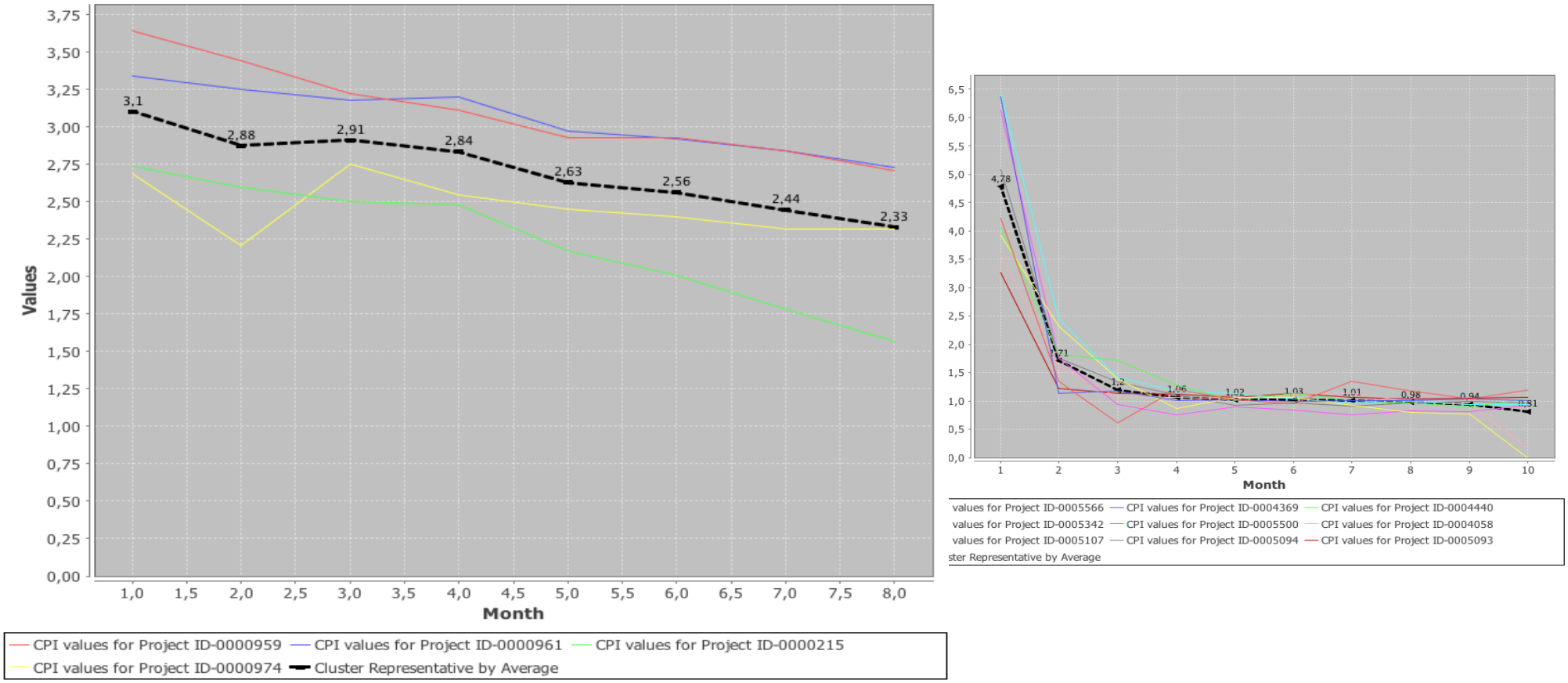
- Clustering-**Verfahren** (k-Means, Hierarchical Clustering)
- Anzahl von Clusters **k**
- **…**



Data Mining

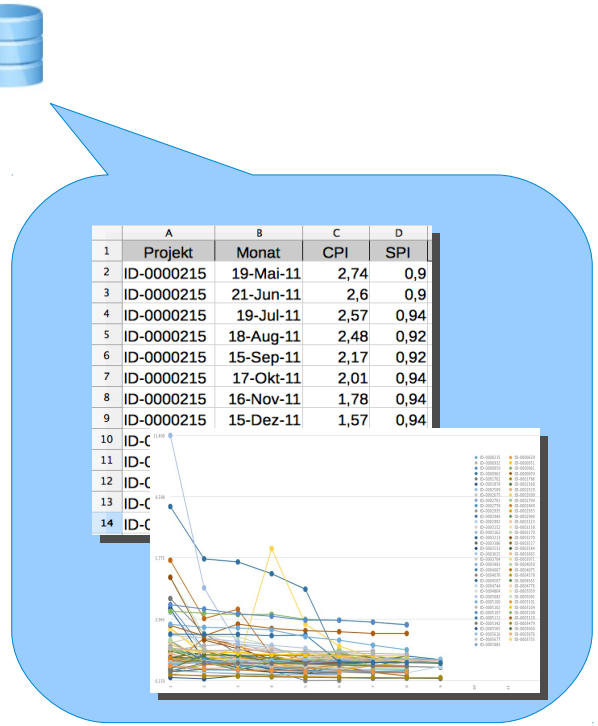
- Ähnlichkeitsprinzip und Cluster-Repräsentant

Beispiel: Ergebnis der Cluster-Analyse



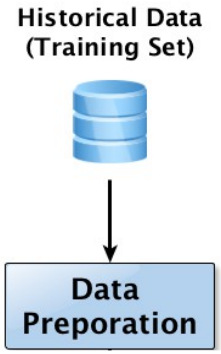
Idee und Konzept

Historical Data
(Training Set)

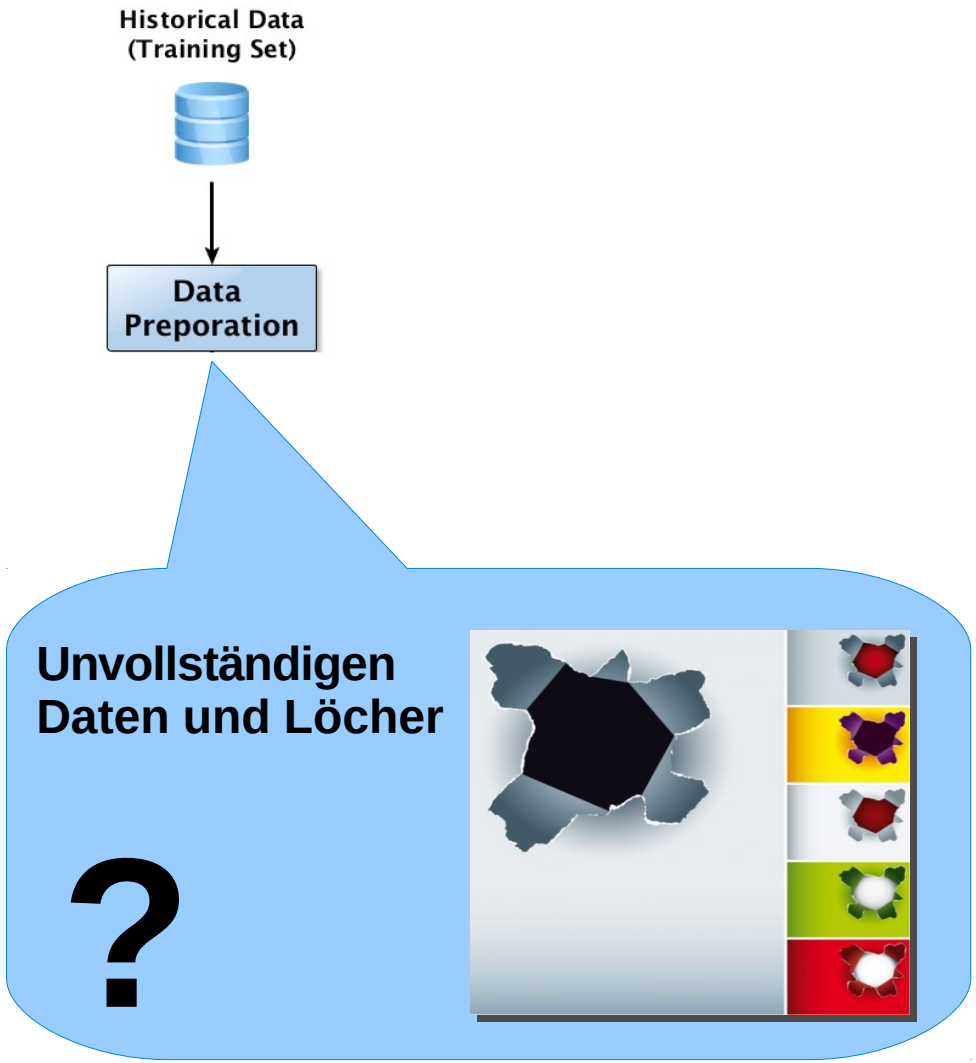


	A	B	C	D
1	Projekt	Monat	CPI	SPI
2	ID-0000215	19-Mai-11	2,74	0,9
3	ID-0000215	21-Jun-11	2,6	0,9
4	ID-0000215	19-Jul-11	2,57	0,94
5	ID-0000215	18-Aug-11	2,48	0,92
6	ID-0000215	15-Sep-11	2,17	0,92
7	ID-0000215	17-Okt-11	2,01	0,94
8	ID-0000215	16-Nov-11	1,78	0,94
9	ID-0000215	15-Dez-11	1,57	0,94
10	ID-(			
11	ID-(			
12	ID-(			
13	ID-(			
14	ID-(			

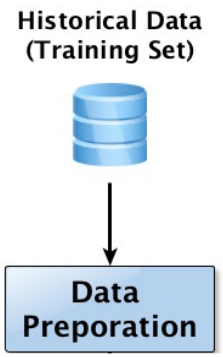
Idee und Konzept

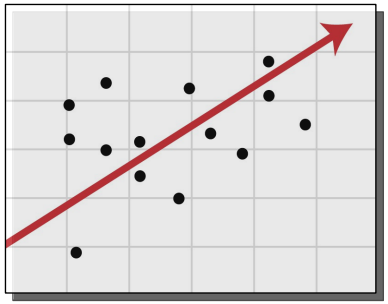


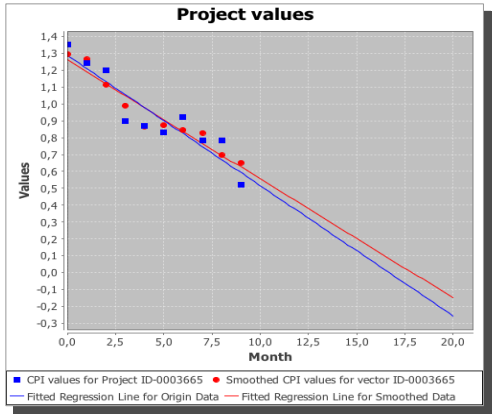
Idee und Konzept



Idee und Konzept





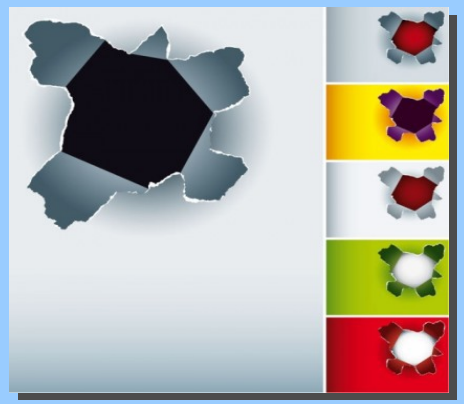


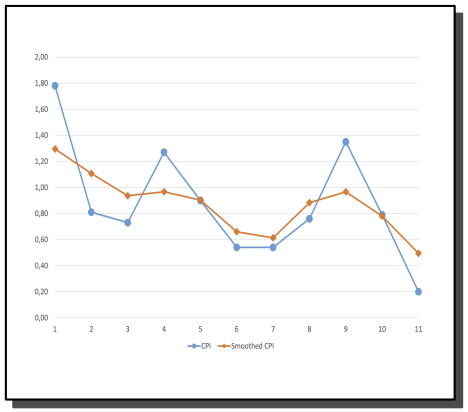
Regression

Smoothing

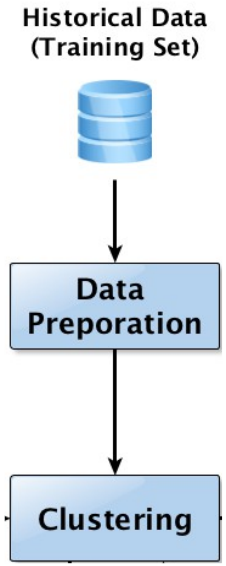
Unvollständigen Daten und Löcher

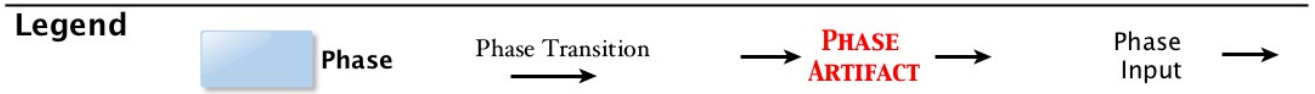
?



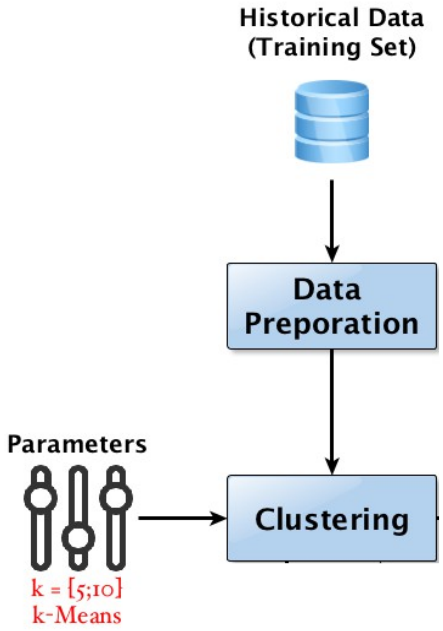


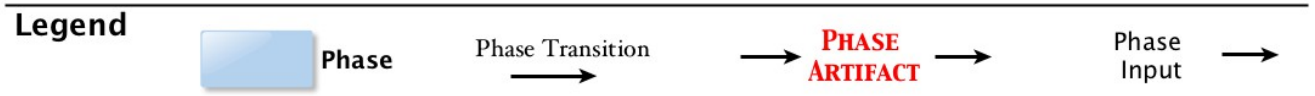
Idee und Konzept



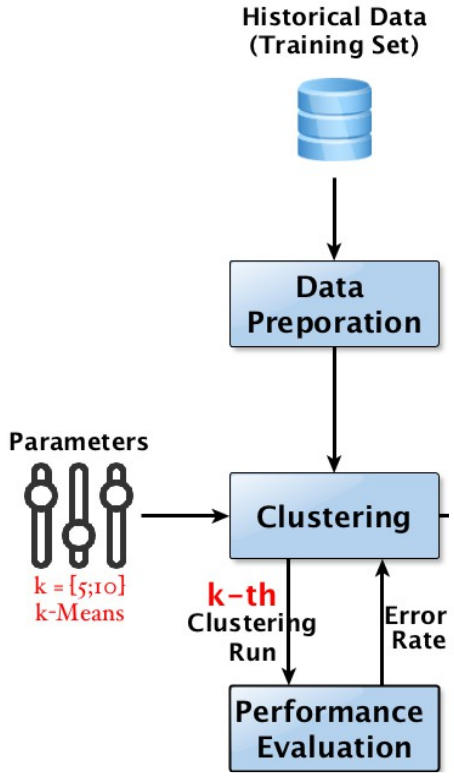


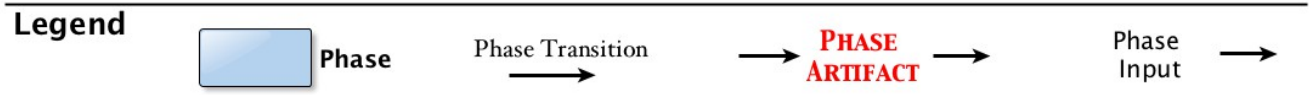
Idee und Konzept



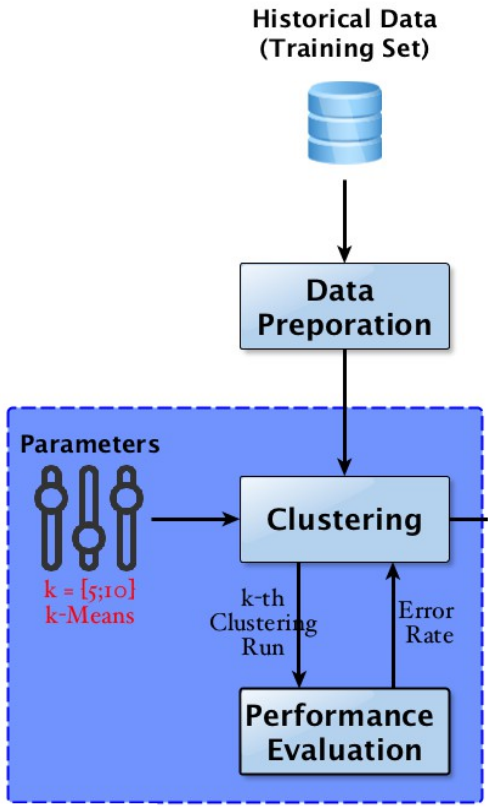


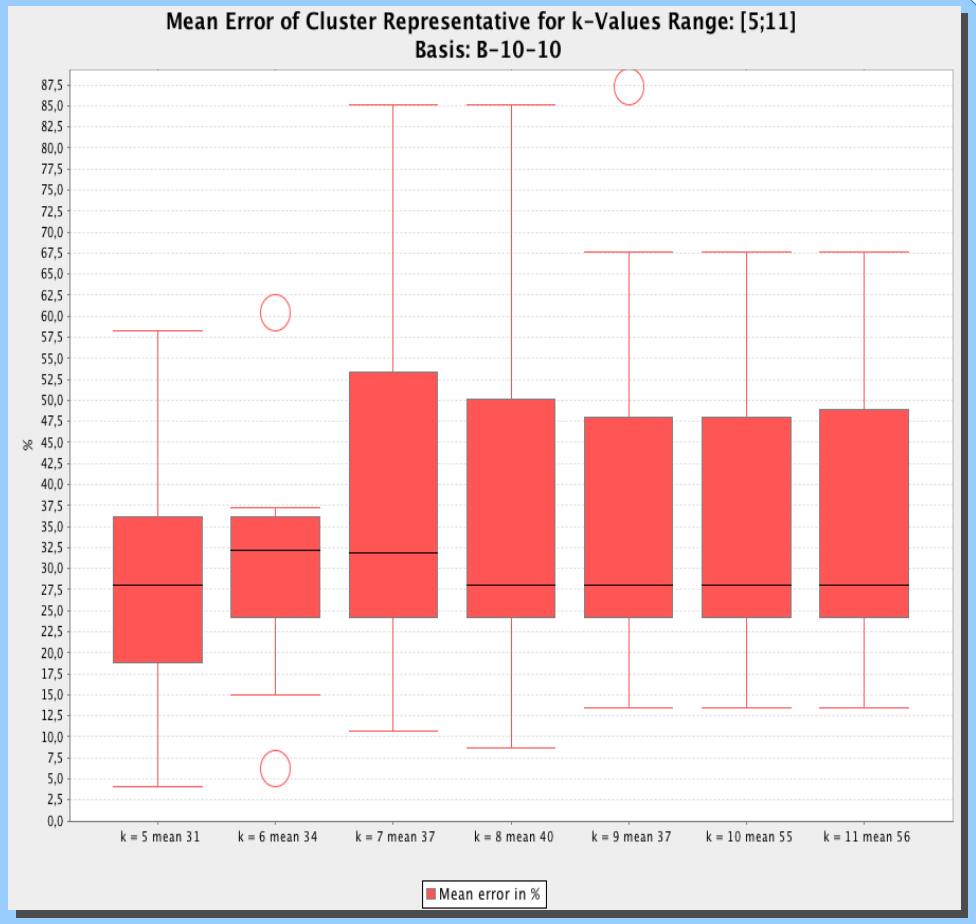
Idee und Konzept



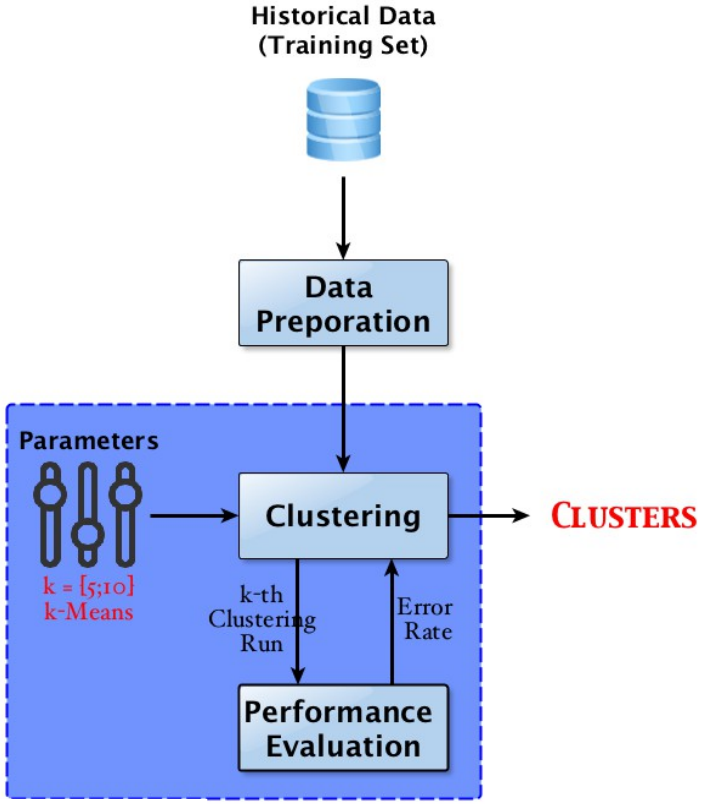


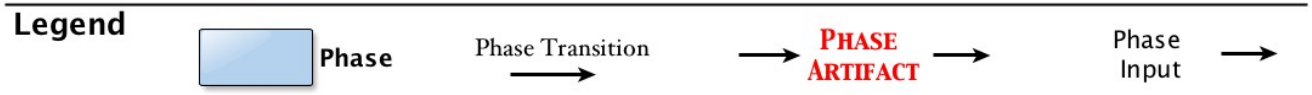
Idee und Konzept



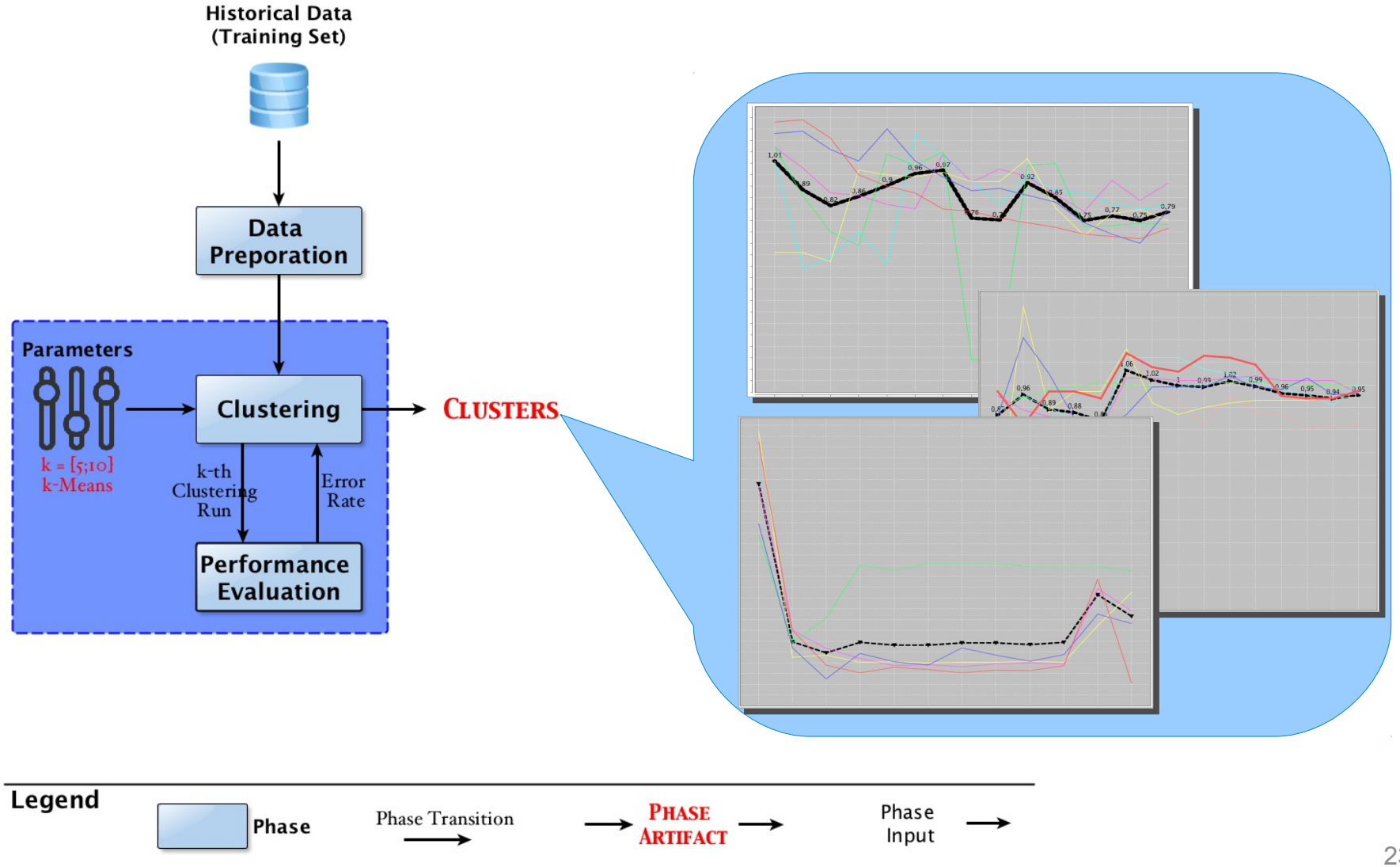


Idee und Konzept

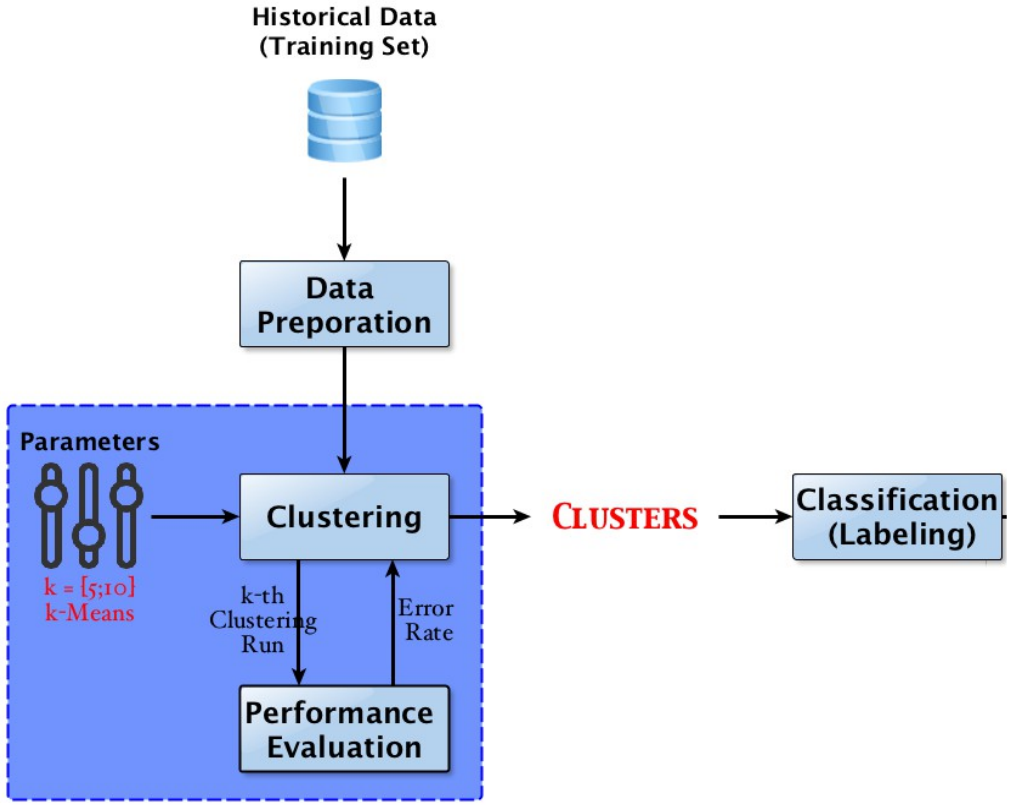


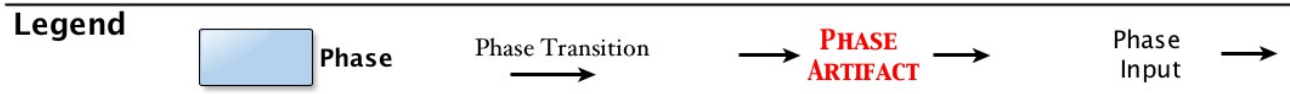


Idee und Konzept

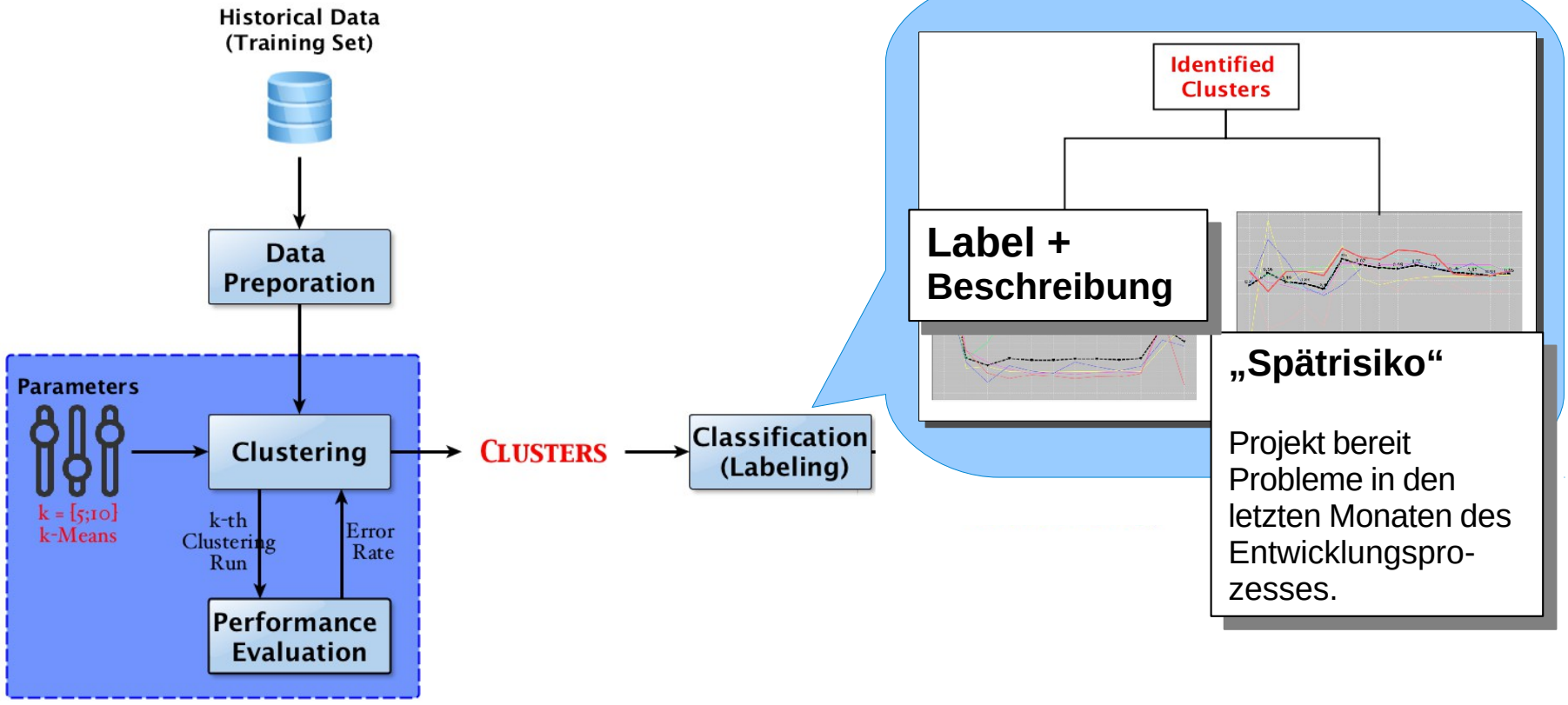


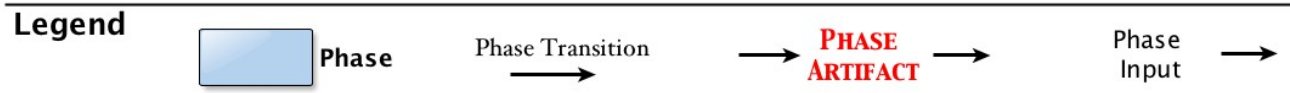
Idee und Konzept



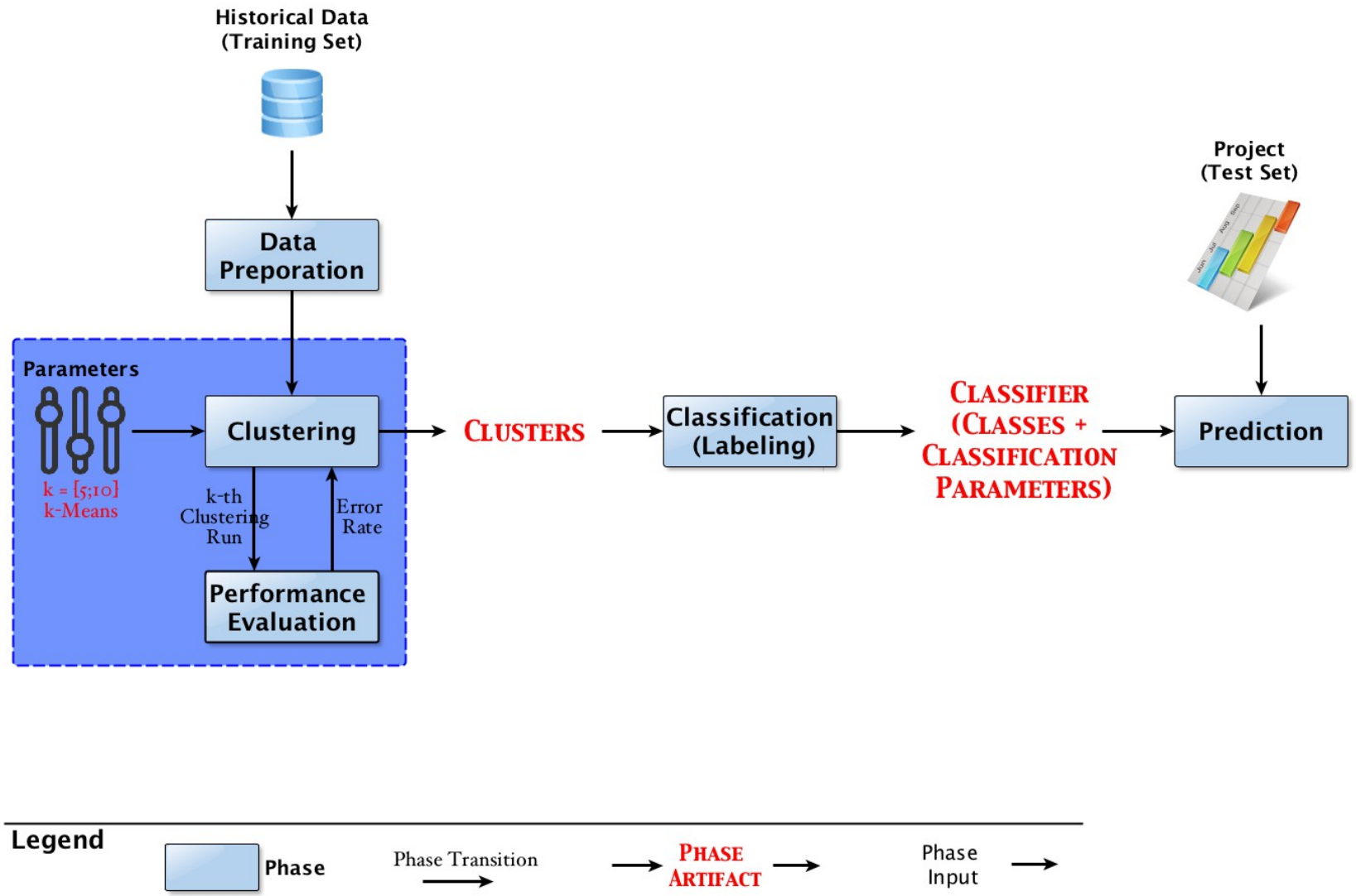


Idee und Konzept

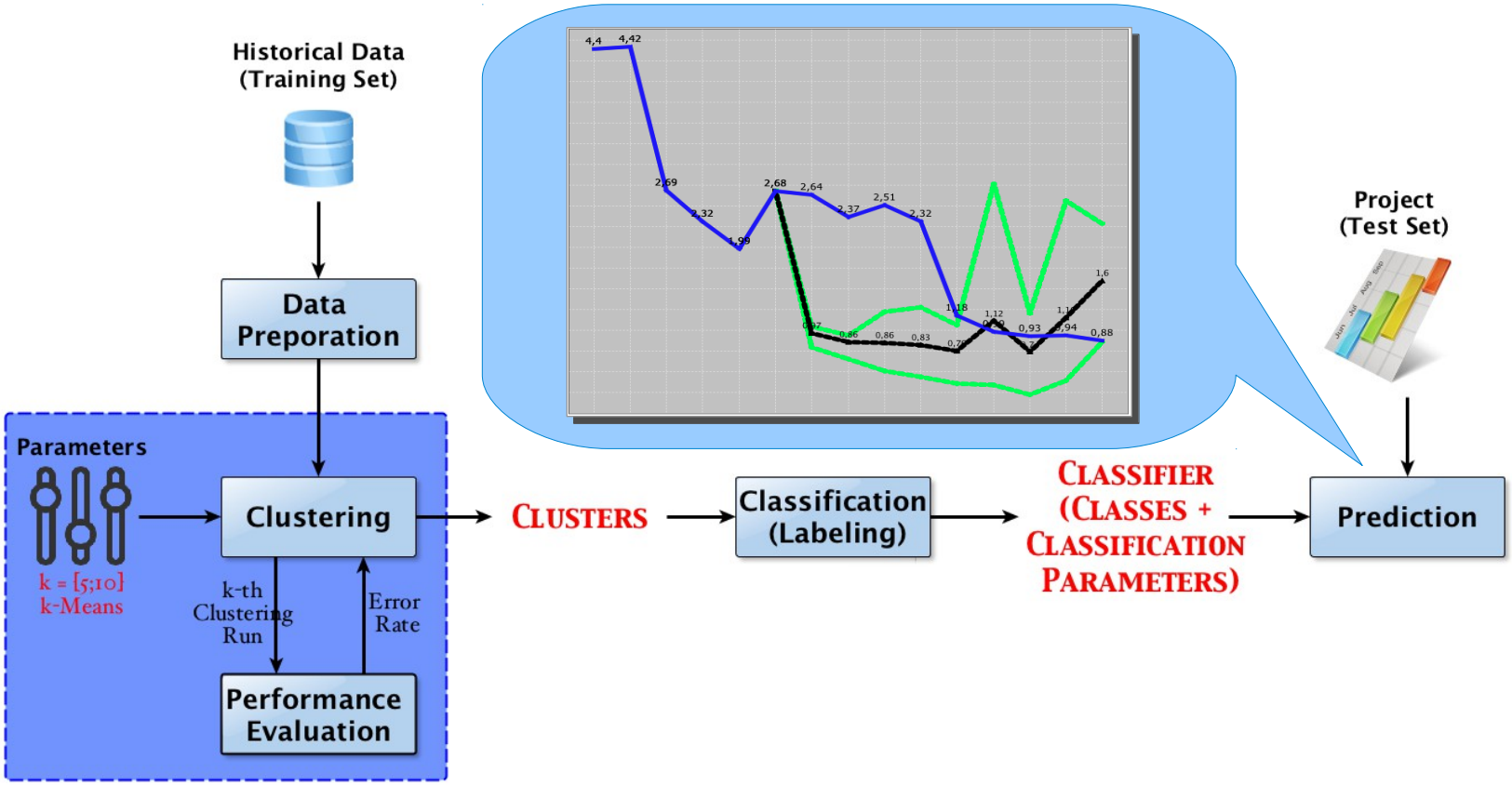




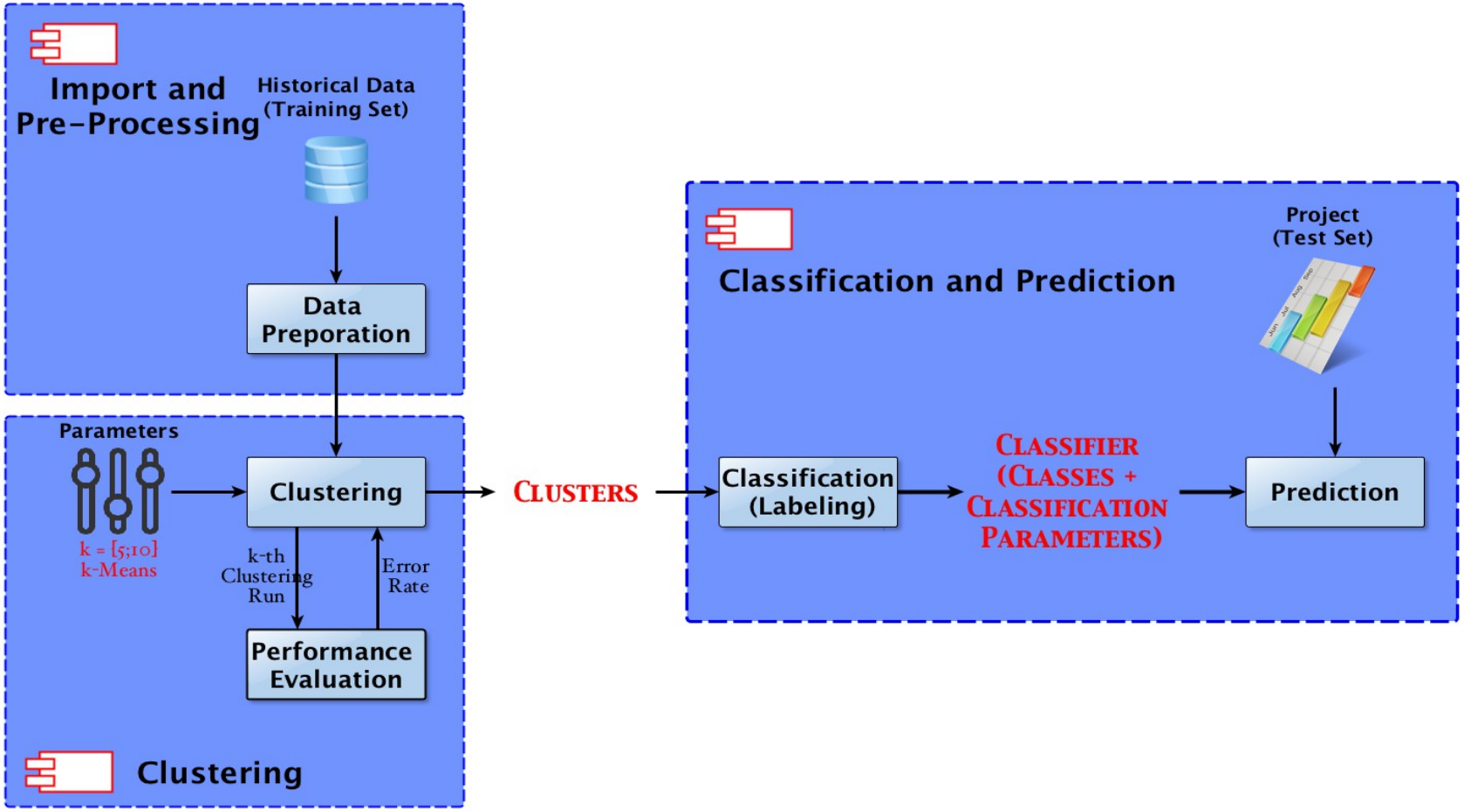
Idee und Konzept



Idee und Konzept



Idee und Konzept

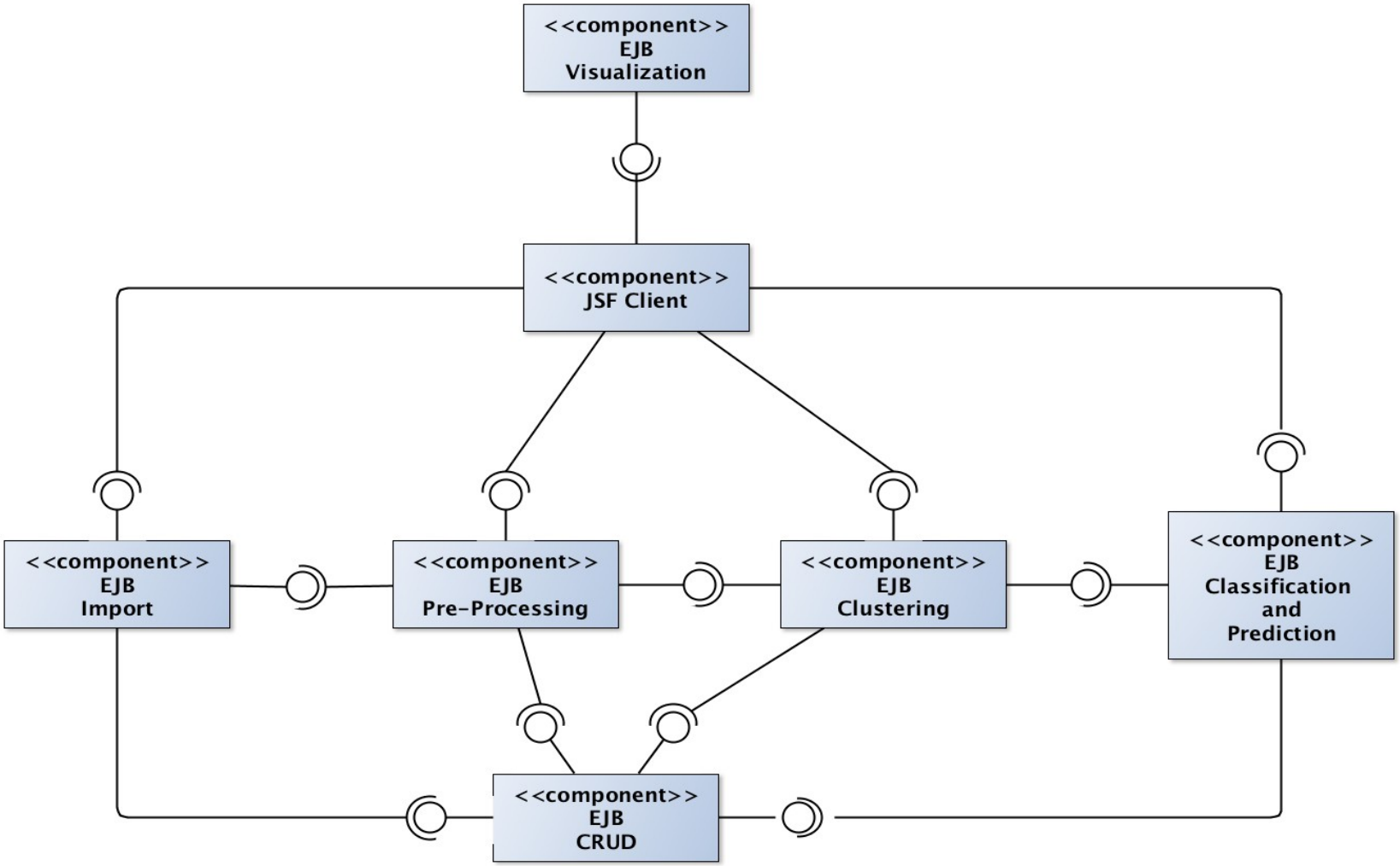


Ziele und Anforderungen

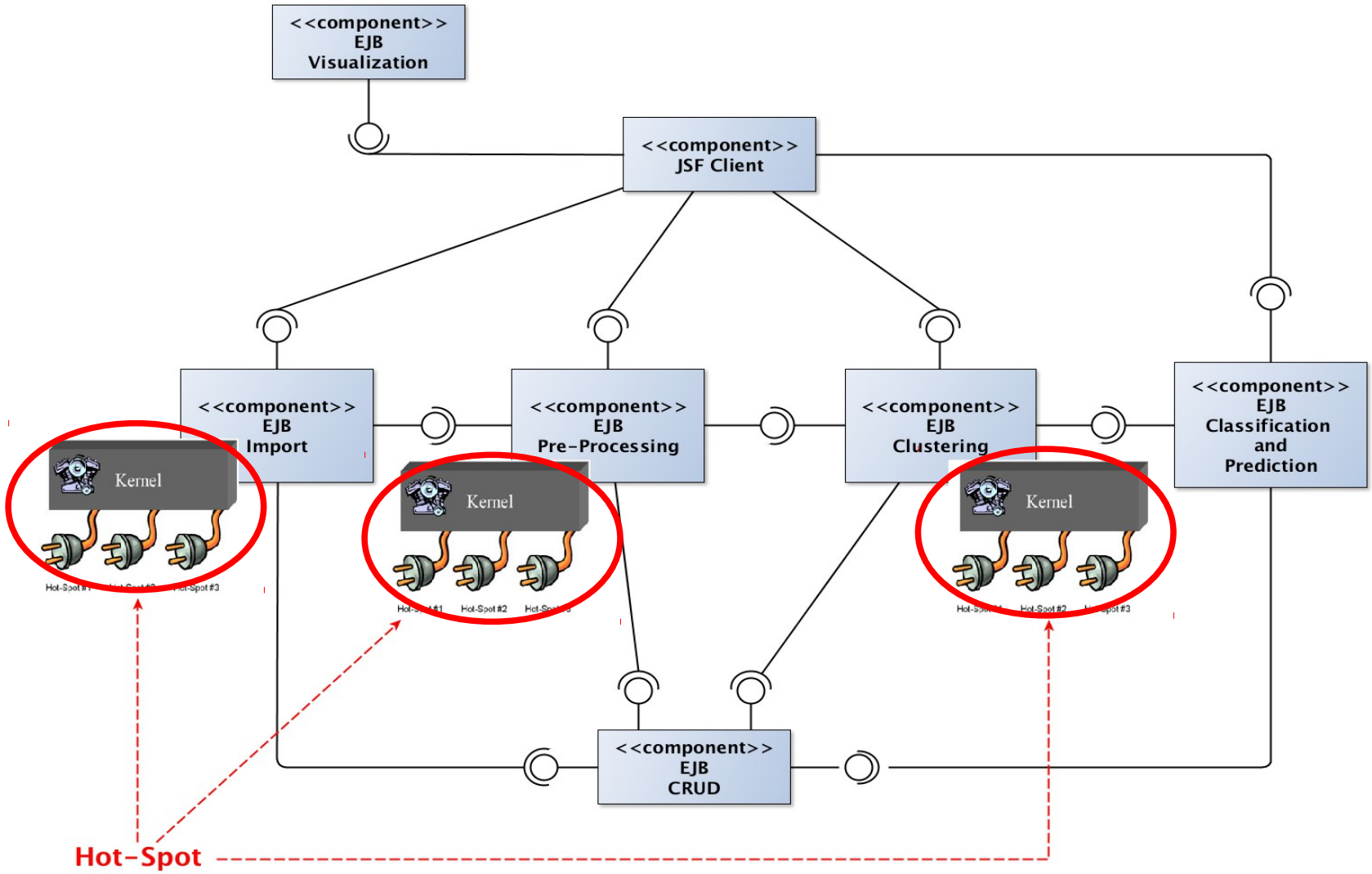
- Entwicklung von Framework
  - Lose gekoppelte Architektur
  - Vollständige Implementierung von Data Mining Prozesses (Wissensextraktion)
  - Jeder Schritt der Wissensextraktion ist gekapselt in eigene Komponente
  - Unterstützung von Variabilität und Erweiterbarkeit
  - Web-Basierte Anwendung + JavaEE

- Prototyp Implementierung von Projekt-Vorhersage-Tool
  - System “versteht” Daten von externem Kooperationspartner
  - Import, Pre-processing, Clustering
  - Benutzbarkeit und UI
  - Konfiguration ohne Expertenwissen

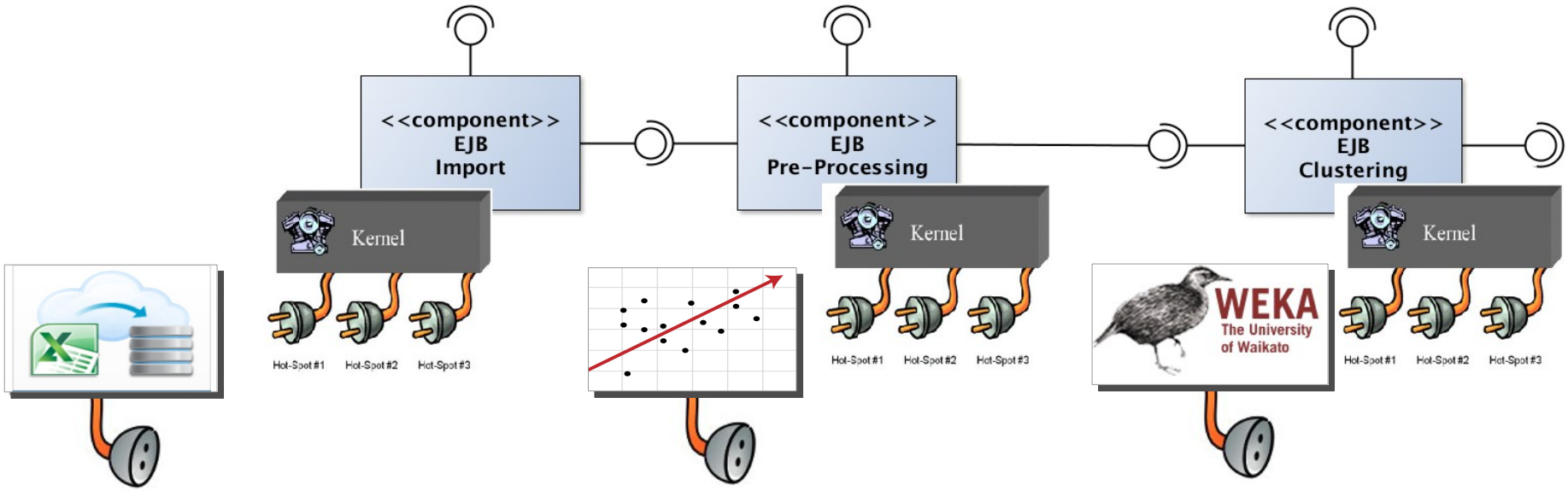
Realisierung



Hot-Spots


<<component>>
EJB
Visualization
<<component>>
JSF Client
<<component>>
EJB
Import
<<component>>
EJB
Pre-Processing
<<component>>
EJB
Clustering
<<component>>
EJB
Classification
and
Prediction
Kernel
Kernel
Kernel
Hot-Spot #1
Hot-Spot #2
Hot-Spot #3
<<component>>
EJB
CRUD
Hot-Spot

Hot-Spots Implementierung



Hot-Spot Implementierung:
Import aus xls/xlsx

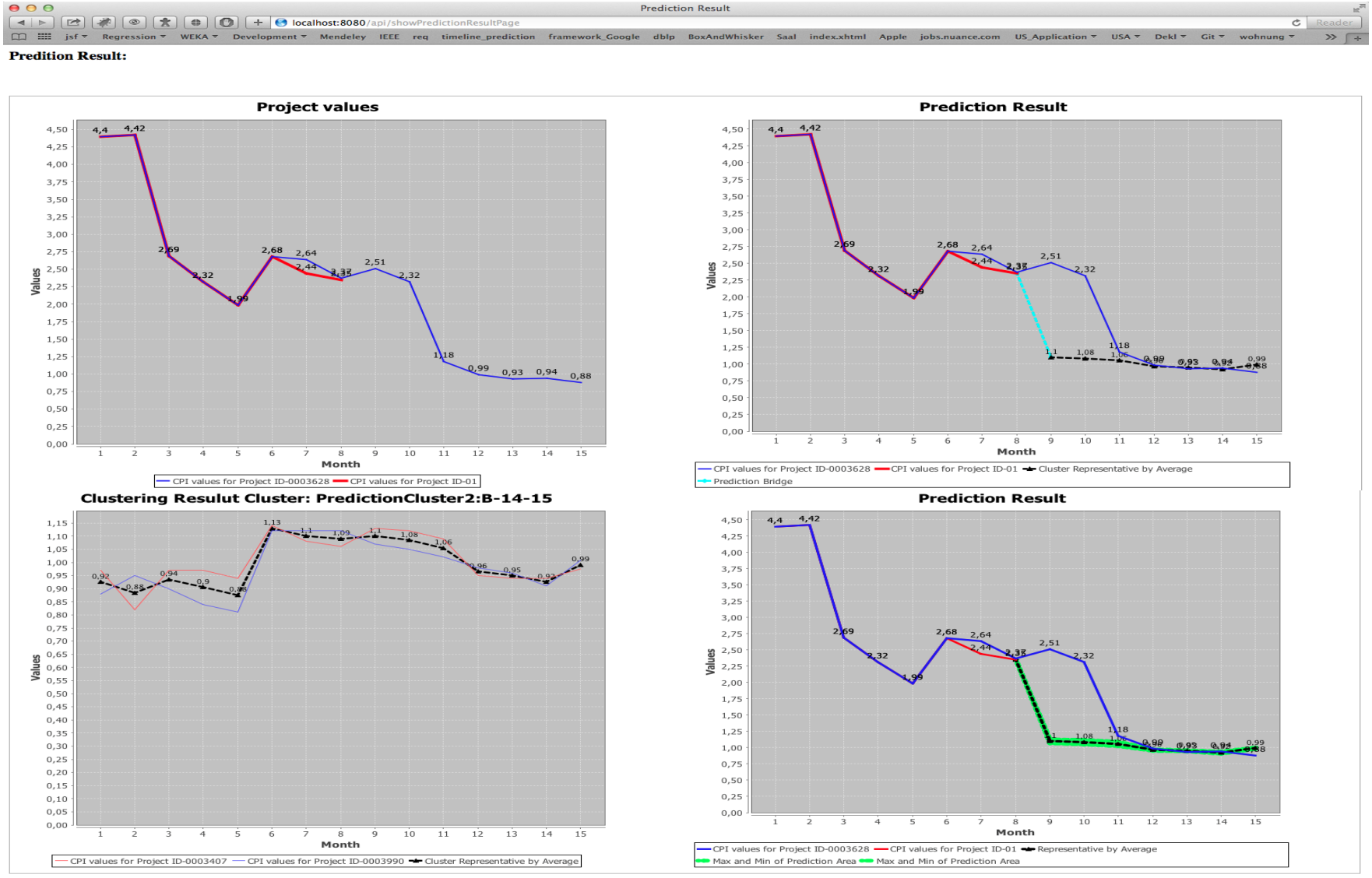
Hot-Spot Implementierung:
Regression Strategies
- Lineare Regression
- Nicht-Lineare Regression

Missing Values Handling

Hot-Spot Implementierung:
Clustering Verfahren
- k-Means
- Hierarchical

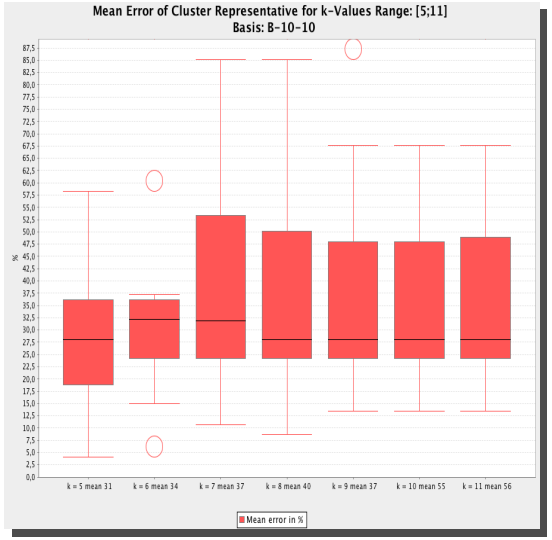
Cluster Repräsentant
- Mittelwert
- Max/Min

Realisierung: Tool Prototyp

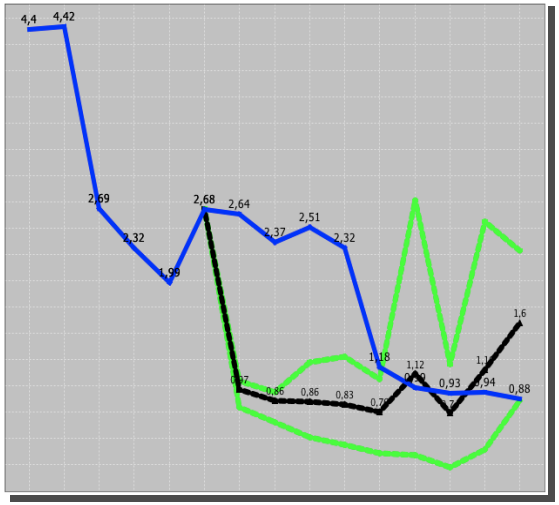


Evaluation

- **Szenario 1:**
  Bewertung von Cluster-Analyse



- **Szenario 2:**
  Bewertung von Vorhersage
  (Prediction)



Evaluation Setup

- Betriebsystem: Mac OS X, Version 10.9.5, Prozessor 2 GHz
- Daten: CPI Werte von 151 Projekten von 5 bis 15 Monaten (≈ 54%)
- Parameter
  - Lineare Regression
  - k-Means
  - Cluster-Zahl *k* aus dem Bereich [5;10]

Evaluation: Szenario 1

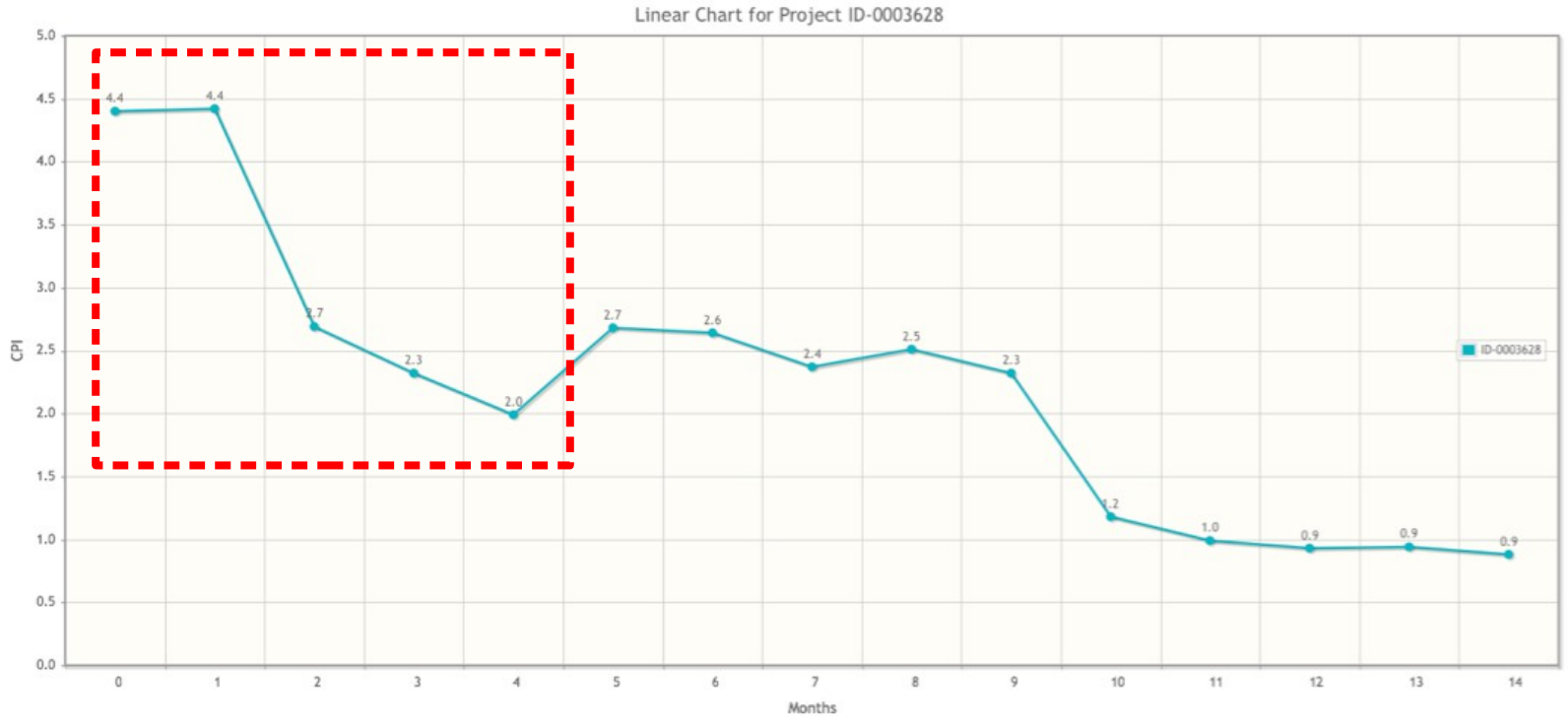
- 10 Datensätze getestet
- **Das beste Ergebnis**
  - Datensatz aus 14 bis 15-monatigen Projekten (12 Proj.)
  - 95% aller Fehler zwischen 3% und 76%
  - Mittelwert der Fehler zwischen 12% und 22%
  - Bestes k = 8
- **Das schlechteste Ergebnis**
  - Datensatz aus 8 bis 15-monatigen Projekten (99 Proj.)
  - 95% aller Fehler zwischen 12% und 322%
  - Mittelwert der Fehler zwischen 94% und 135%
  - Bestes k = 7

Evaluation: Szenario 1
Schlüsse

- Regression als Einflussfaktor auf Qualität des Clusterings
- Für jeden Datensatz existiert einen bestes Cluster-Zahl k
- Sehr heterogene Daten haben negative Auswirkung auf Clustering-Qualität

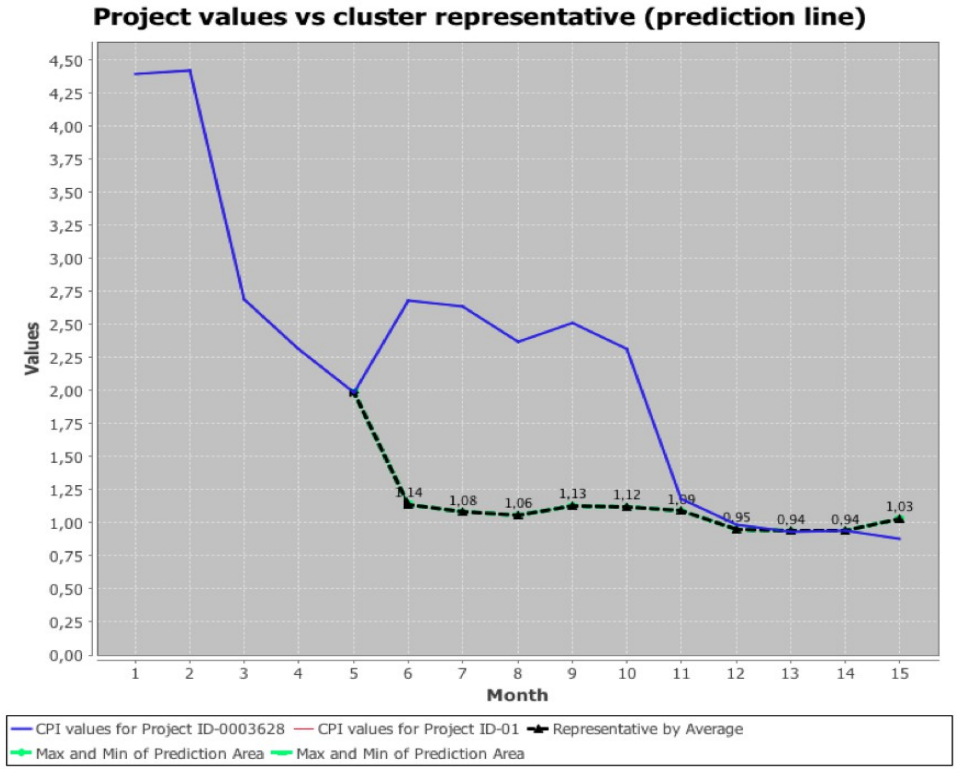
Evaluation: Scenario 2

- 2 Experimente
  - Einfache Vorhersage von einem 15-montigen Projekt
  - Schrittweise Vorhersage



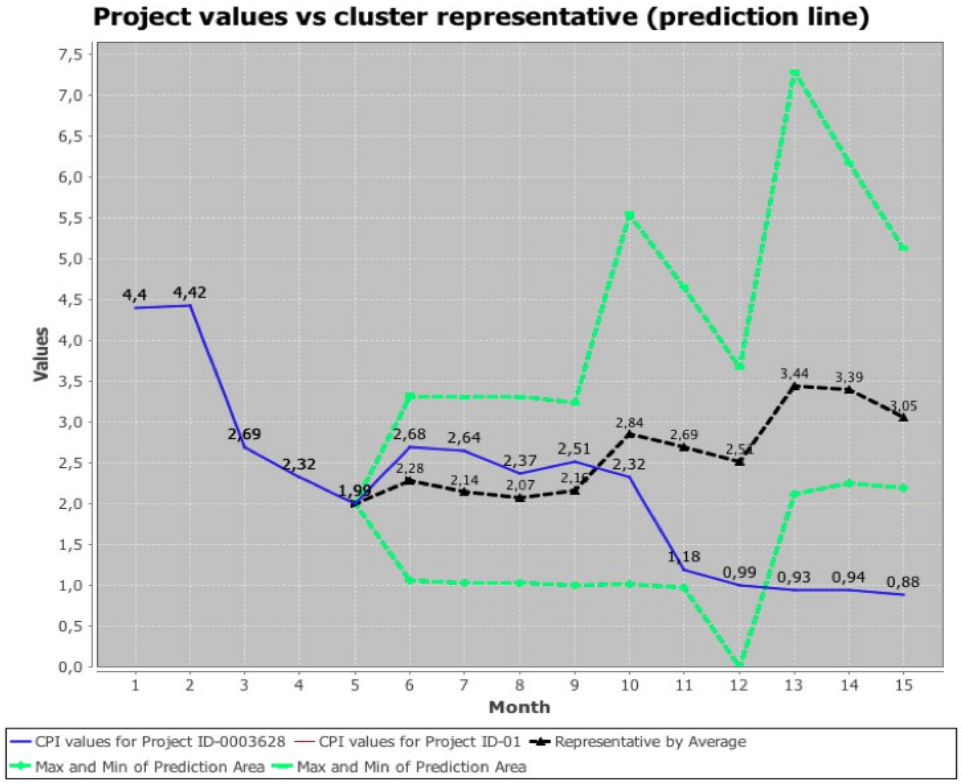
Evaluation: Szenario 2

- Einfache Vorhersage
  - Datensatz: 14 bis 15-monatige Projekte
  - Mittelwert der Fehler: 31,83%



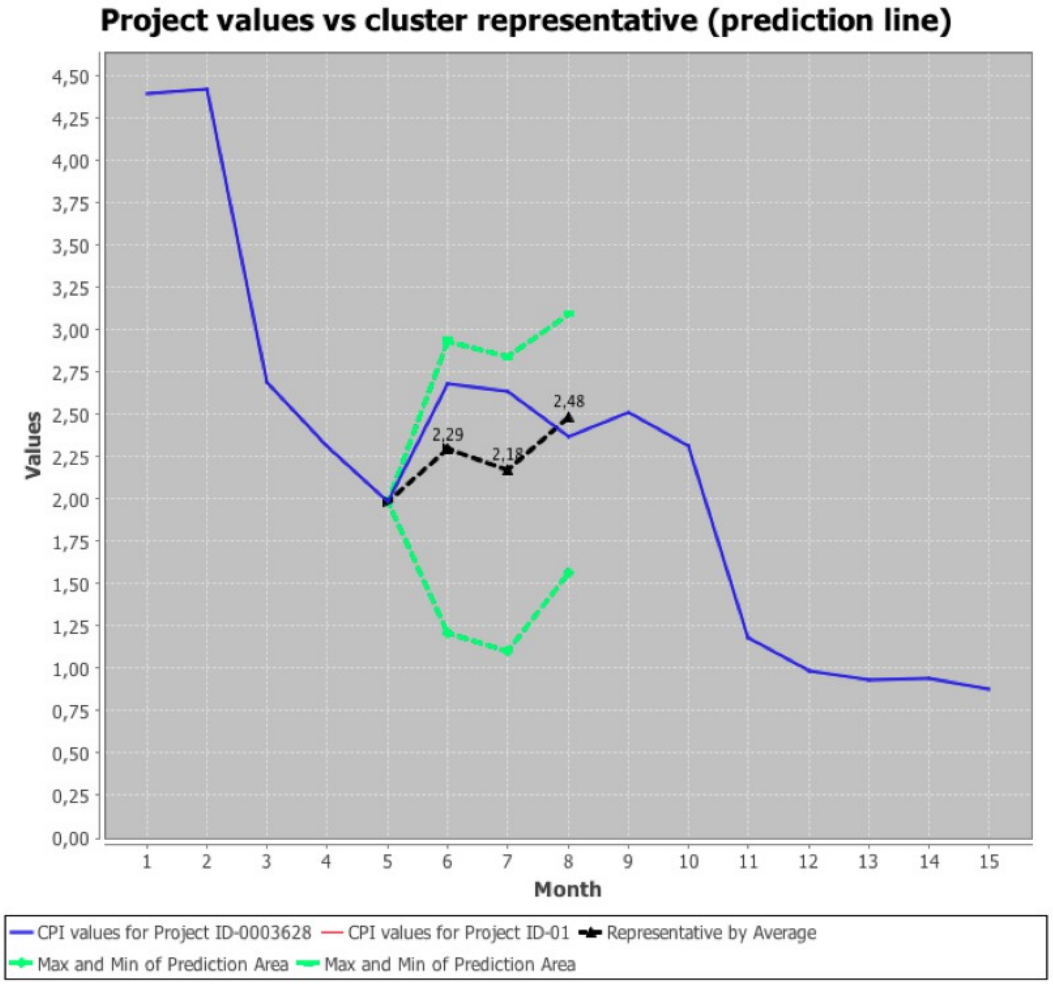
Evaluation: Szenario 2

- Einfache Vorhersage
  - Datensatz: 8 bis 15-monatige Projekte
  - Mittelwert der Fehler: 114,12%



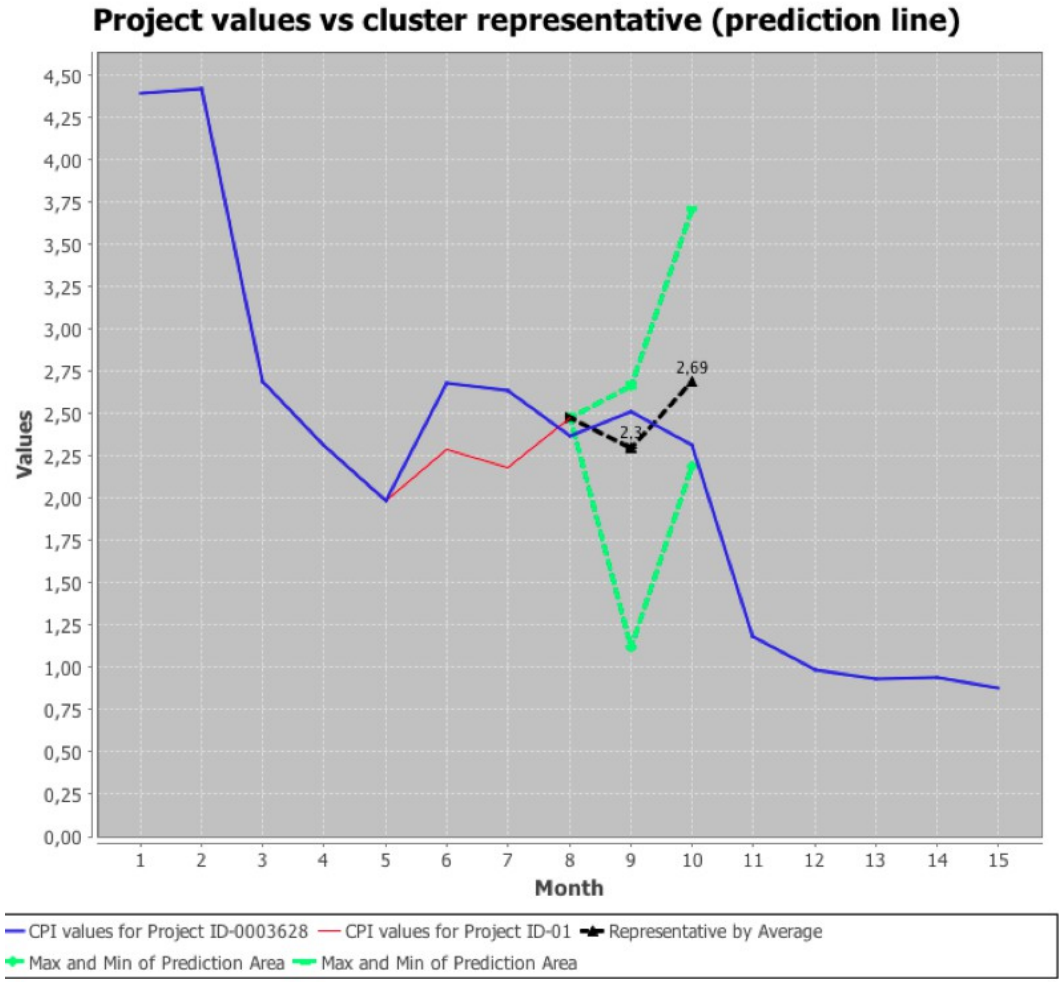
Evaluation: Szenario 2

- Schrittweise Vorhersage



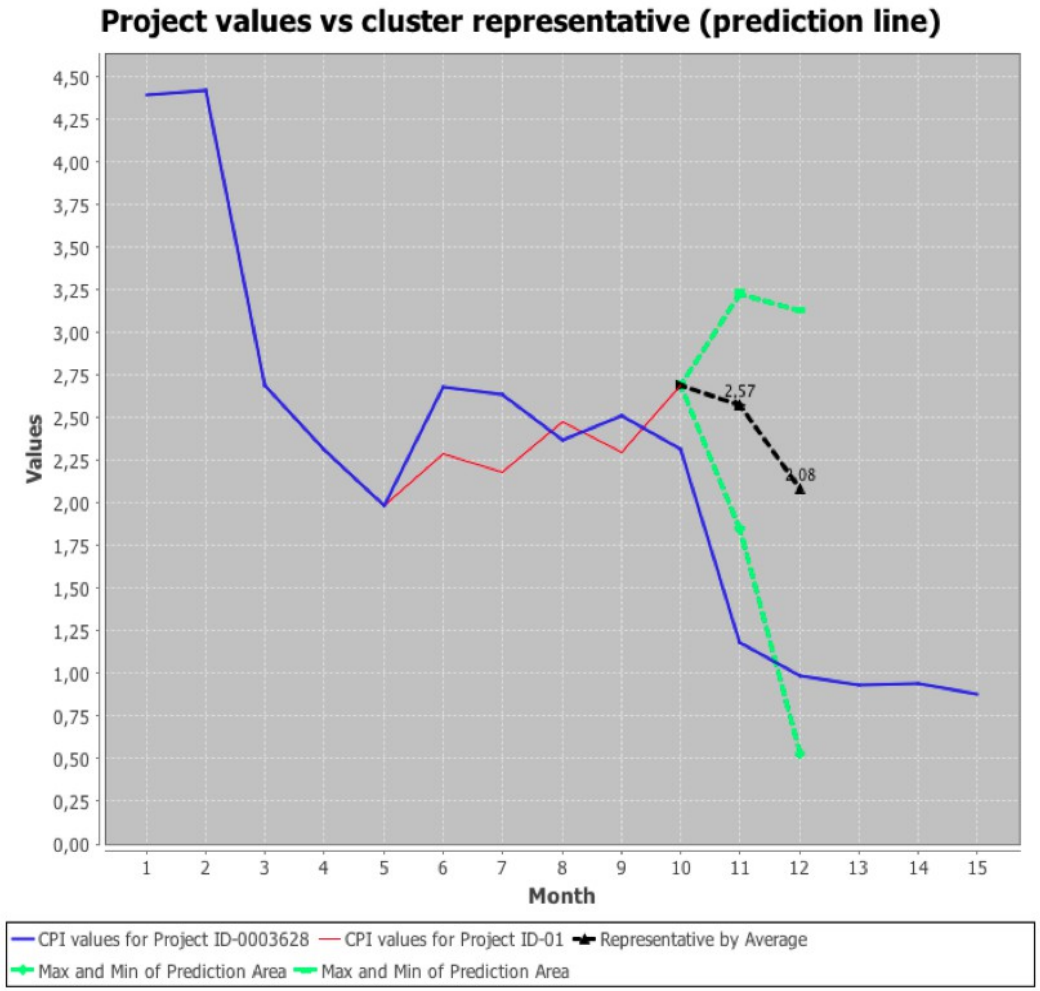
Evaluation: Szenario 2

- Schrittweise Vorhersage



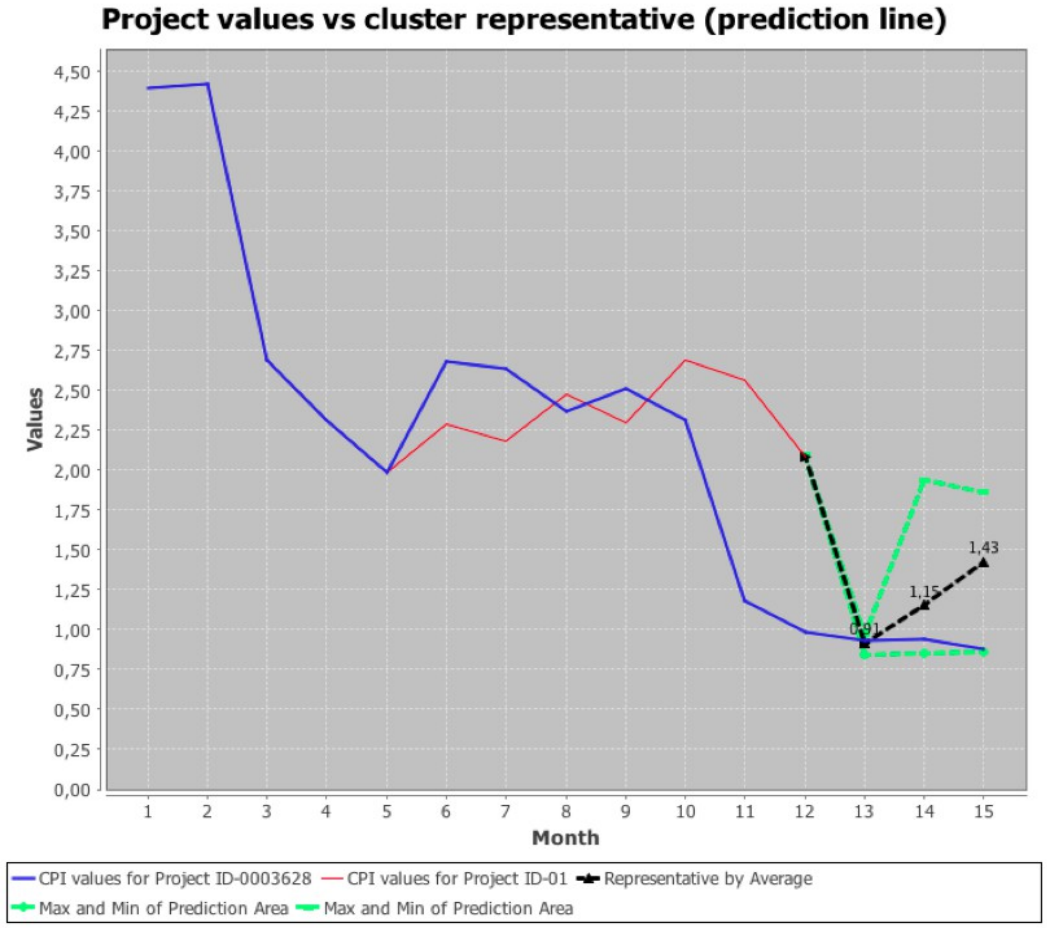
Evaluation: Szenario 2

- Schrittweise Vorhersage



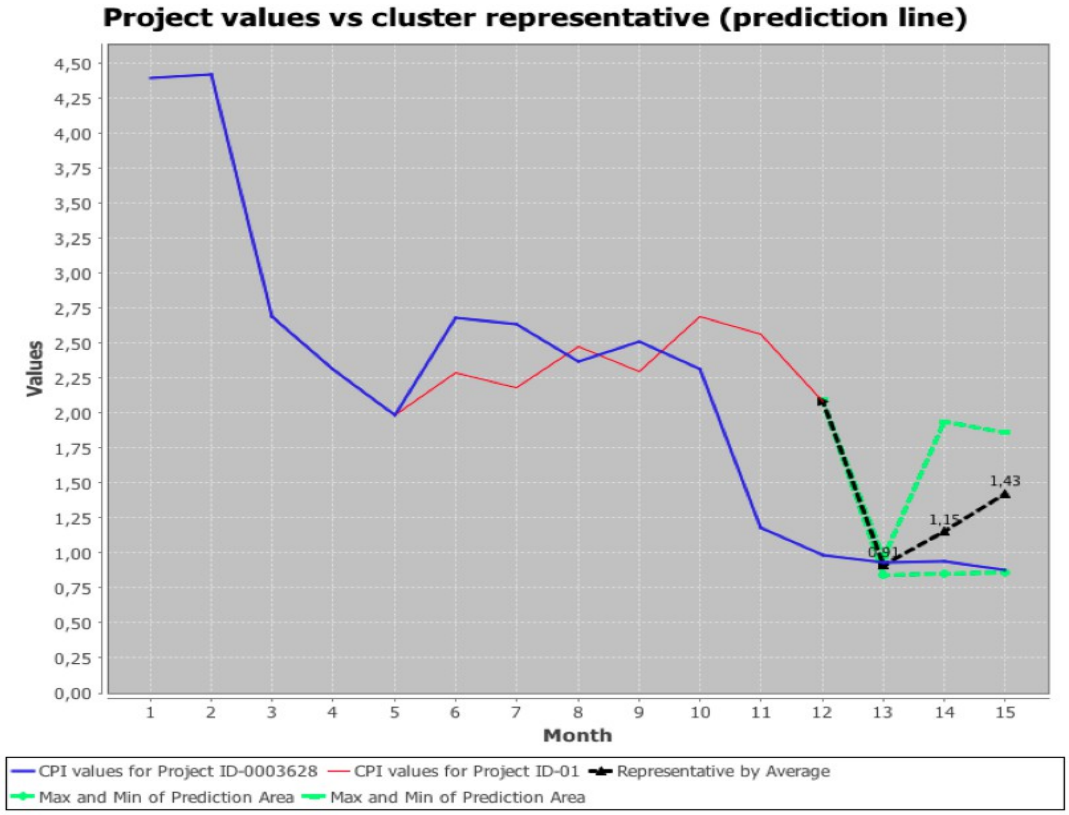
Evaluation: Szenario 2

- Schrittweise Vorhersage



Evaluation: Szenario 2

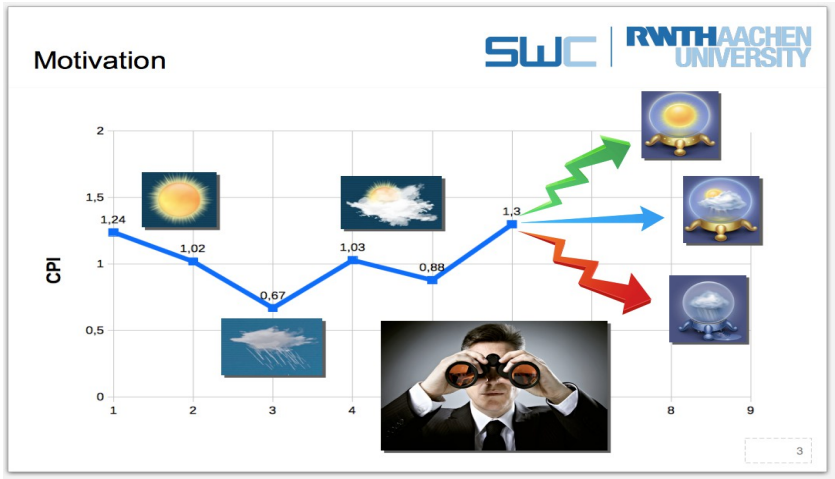
- Schrittweise Vorhersage
  - Mittelwert der Fehler: 37,59%

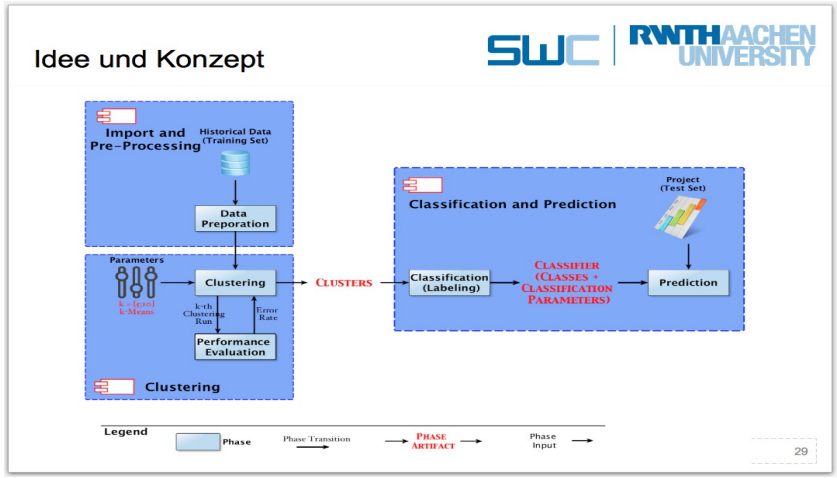


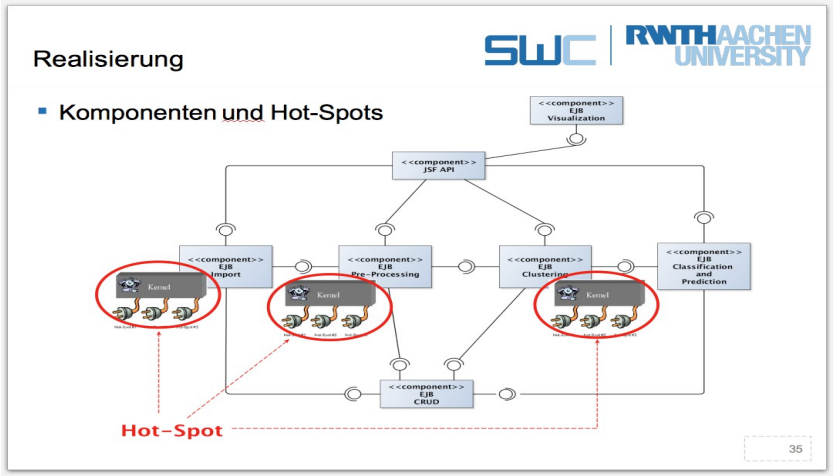
Zukünftige Arbeiten

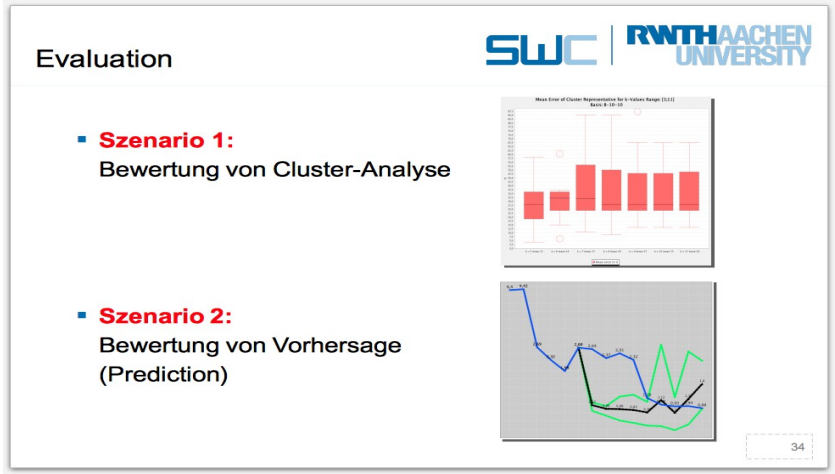
- Weitere Implementierung von Hot-Spots
- Identifizierung von weiteren Einflussfaktoren
- Anwendung und Evaluation in anderen Domänen
- Prototyp-Erweiterung (Client mit AngularJS)
- Integration ins EMI System

Zusammenfassung









Source

- [1] Scafetta Prediction: Scafetta, EPA 2009. – URL http://vademecum.brandenberger.eu/themen/klima/ursache.php. – Access date: 10.11.2014

- [2] Thanh Vi Bach: Entwurf prognostischer Softwareprozessmetriken auf Basis iterativen Clusterings, RWTH Aachen University, Master Thesis, 2014.

- [3] Dindin Wahyudin, Rudolf Ramler, and Stefan Biffl. A framework for defect prediction in specific software project contexts. In Zbigniew Huzar, Radek Kocí, Bertrand Meyer, Bartosz Walter, and Jaroslav Zendulka, editors, CEE-SET, volume 4980 of Lecture Notes in Computer Science, pages 261–274. Springer, 2008.

- [4] Ningda R. Li and Marvin V. Zelkowitz. An information model for use in software management estimation and prediction. In Bharat K. Bhargava, Timothy W. Finin, and Yelena Yesha, editors, CIKM 93, Proceedings of the Second International Conference on Information and Knowledge Management, Washington, DC, USA, November 1-5, 1993, pages 481–489. ACM, 1993.

- [5] Rudolf Ramler, Klaus Wolfmaier, Erwin Stauder, Felix Kossak, and Thomas Natschläger. Key questions in building defect prediction models in practice. In Frank Bomarius, Markku Oivo, Päivi Jaring, and Pekka Abrahamsson, editors, PROFES, volume 32 of Lecture Notes in Business Information Processing, pages 14–27. Springer, 2009.