# DIZ-Konfigurator: Eine Fallstudie zur Empfehlung von Methoden und Modellen zur automatischen Stammdatenklassifikation mittels Semantik

PasDas Summit 2016 - 05.10.2016

Ignacio Traverso, Dr. Benedikt Kämpgen –
Forschungszentrum Informatik am KIT

FZI FORSCHUNGSZENTRUM INFORMATIK

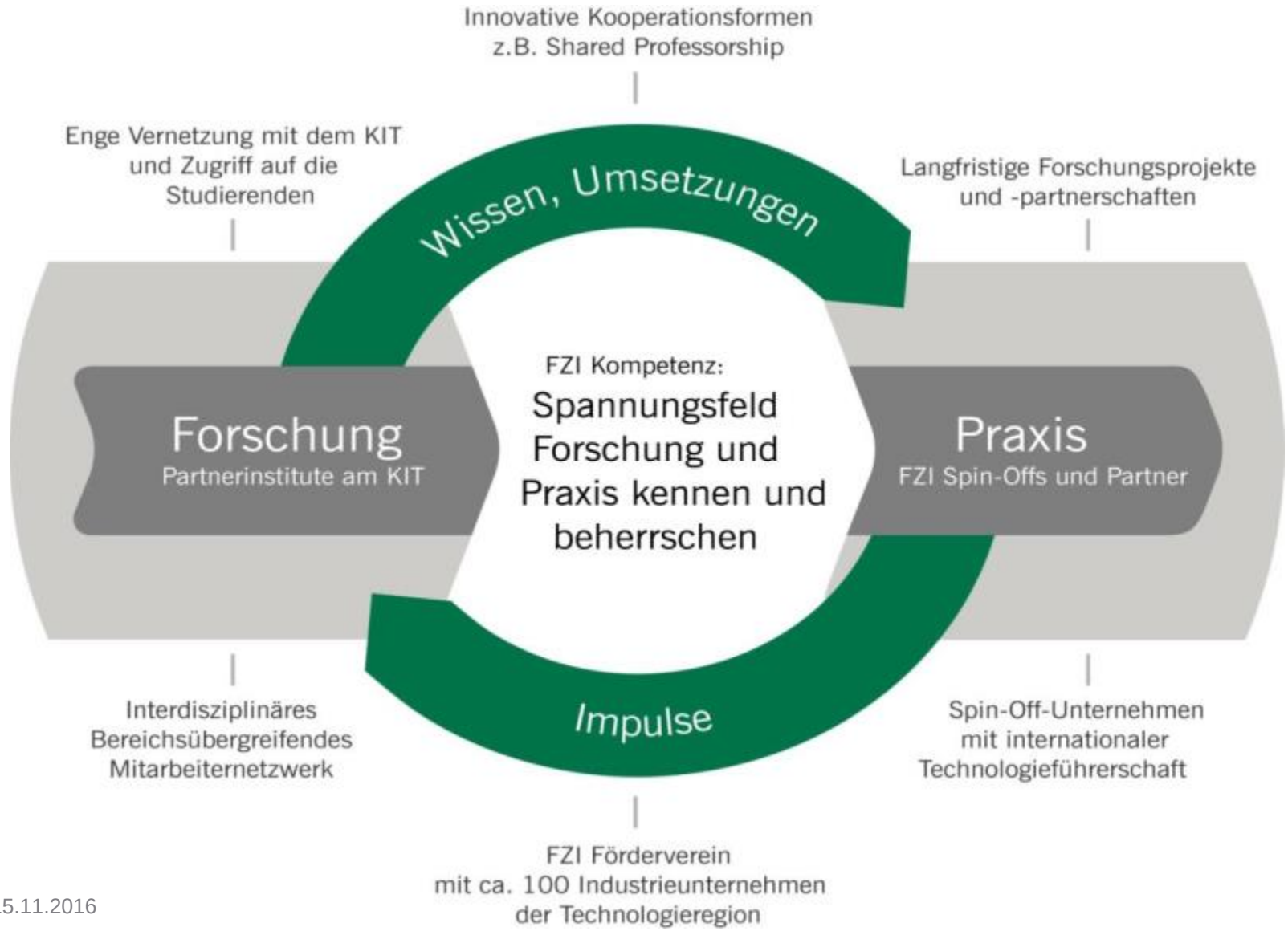
# Agenda

- Überblick FZI
- DIZ-Projekt Konfigurator
- Fallstudie: Automatische Stammdatenklassifikation mit IFCC GmbH
- Zusammenfassung

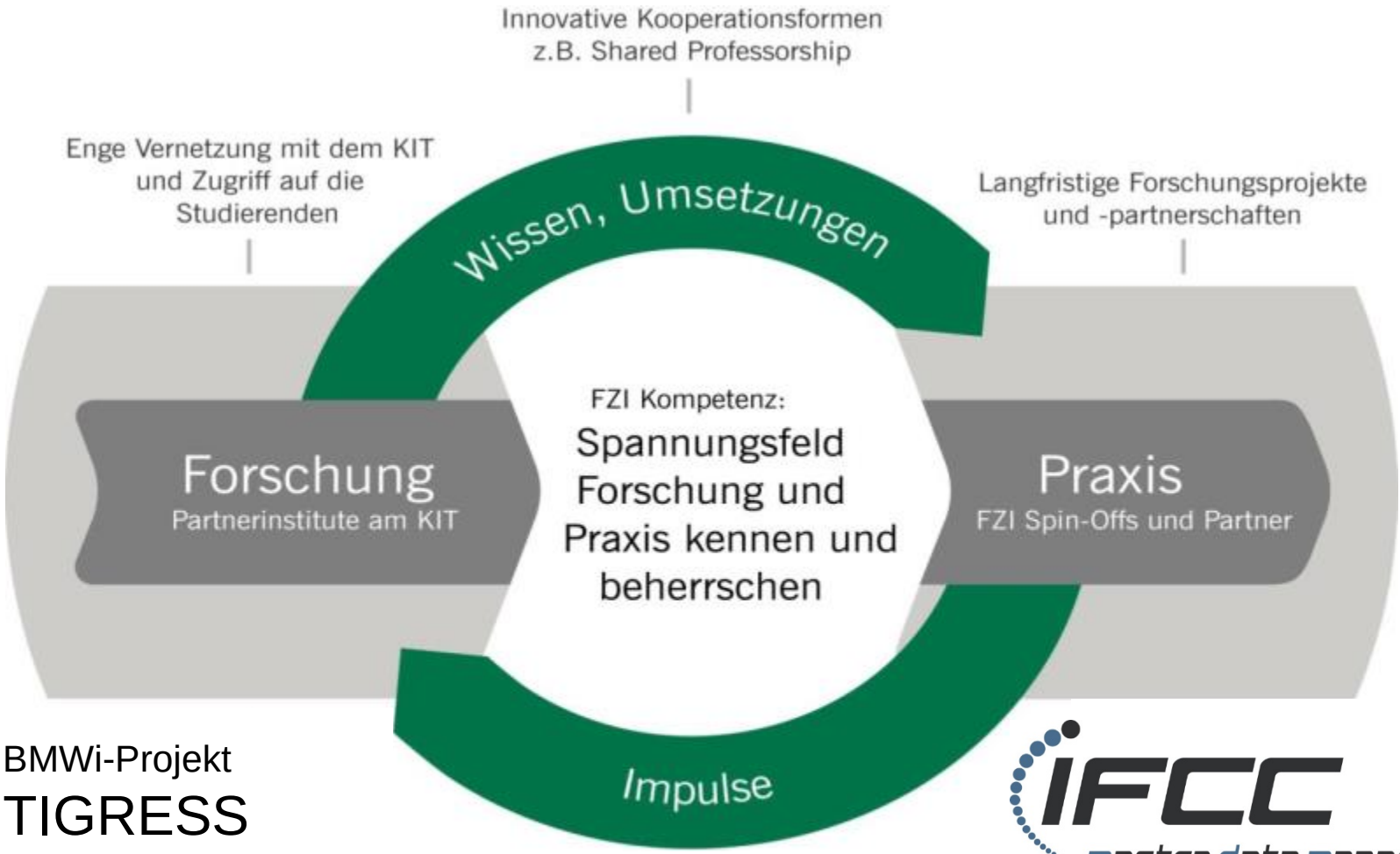
# Die Einrichtung FZI

- Unabhängige gemeinnützige Stiftung des bürgerlichen Rechts für Anwendungsforschung in der Informatik
- Mitglied der Innovationsallianz Baden-Württemberg und Technologieregion Karlsruhe
- Innovationsdrehscheibe in Baden-Württemberg im Bereich Informationstechnologie
- Innovationspartner des KIT im Bereich IT

# FZI Erfolgsrezept:
# Starke Vernetzung und Interdisziplinarität



# FZI Erfolgsrezept:
# Starke Vernetzung und Interdisziplinarität




BMWi-Projekt
TIGRESS

IFCC master.data.management

DIZ DIGITALES INNOVATIONS ZENTRUM

CyberForum
HIGHTECH. UNTERNEHMER. NETZWERK.

# FZI Forschungsbereich
# Information Process Engineering

***Direktoren***





Prof. Dr.-Ing. Kai Furmans

Prof. Dr. Stefan Nickel

Prof. Dr. Rudi Studer

Prof. Dr. York Sure-Vetter

Prof. Dr. Christof Weinhardt

**Leiter Shared Research Group**

**„Corporate Services and Systems“**

Prof. Dr. Thomas Setzer

***Bereichsleiter***

Dr.-Ing. Iris Heckmann

## Forschungsbereich IPE

*Intelligente Informationslogistik und adaptive Infrastrukturen für die vernetzte Welt*

***Abteilungsleiter***

Dr. Boris Amberg

Dr.-Ing. Benedikt Kämpgen

Dr.-Ing. Anne Meyer

Dr. Alexander Schuller

Dr. Stefan Zander






# DIZ Konfigurator

# Motivation

*We believe it's possible to automate certain aspects of data science, and specifically to have machines learn from prior example how to construct new models.*

**DARPA Goes "Meta" with Machine Learning for Machine Learning**

http://www.darpa.mil/news-events/2016-06-17

Defense Advanced Research Projects Agency

# Motivation

Typische Aufgaben der Entscheidungsunterstützung:

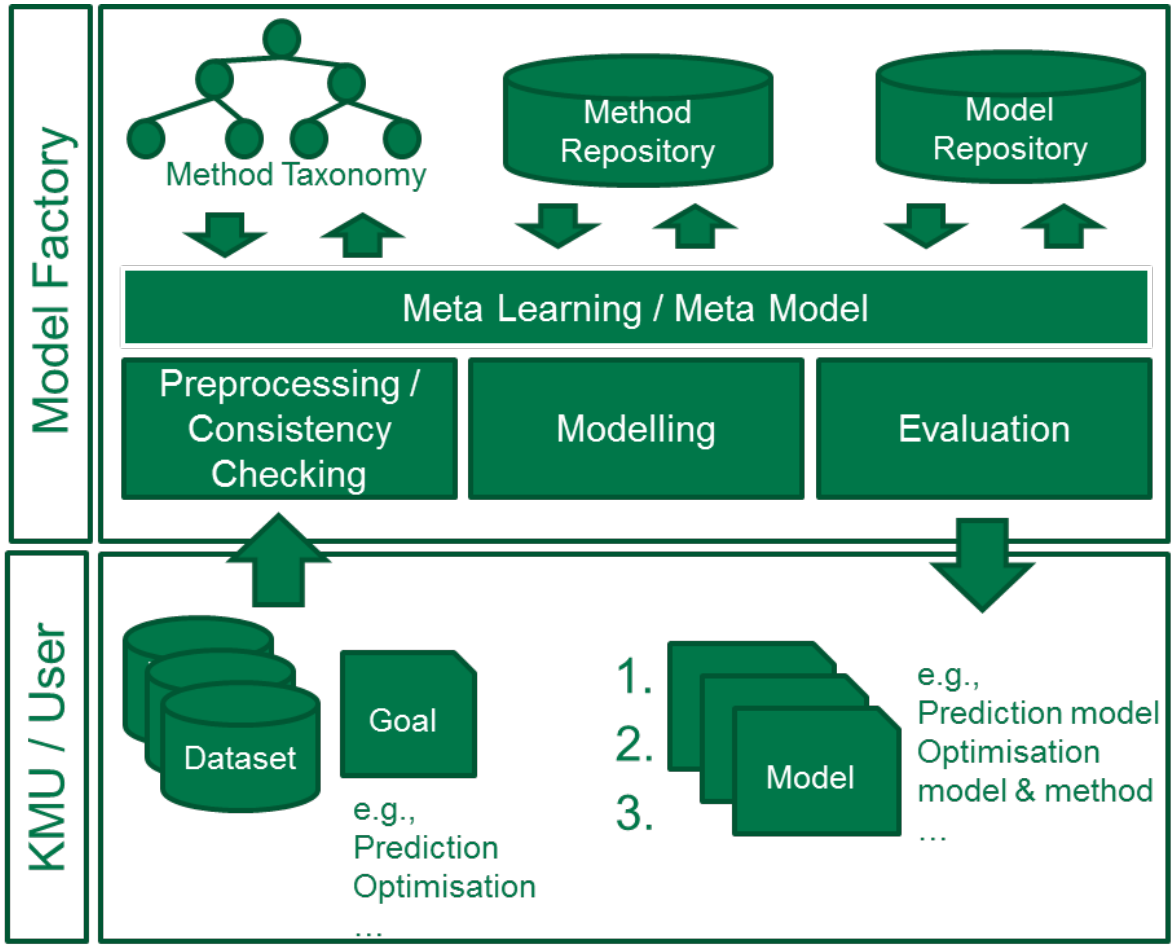
- Optimierung
  - Z.B. Logistik-Firma: Welches ist die schnellste Route? Wie ist ein LKW zu beladen, um das Ausliefern zu vereinfachen?
- Simulation
  - Welche Konsequenzen hat eine Entscheidung? Wie verlässlich ist die Simulation?
- Klassifikation
  - Z.B. Kunden-Meinungen: Welche Produktaspekte werden als positiv oder als negativ wahrgenommen?
- Clustering
  - Wer sind Kunden? In welche Gruppen können Kunden unterteilt werden? Welche Merkmale hat jede Gruppe?
- Vorhersage
  - Wie können Kunden antizipiert werden? Wie kann eine optimale Lagerhaltung vorhergesagt werden?

# Motivation

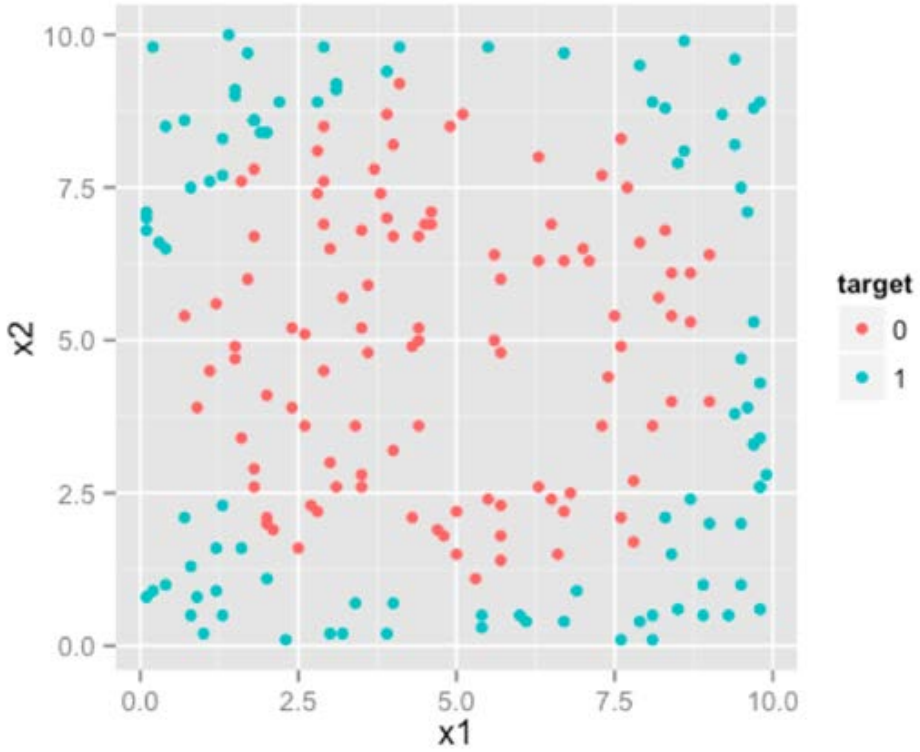
Es gibt aber viele Methoden um solche Aufgaben zu lösen:

- Welche Methoden passen am besten für meine konkrete Aufgabe?
- Welche Methode passt am besten zu meinem Datensatz?
- Gibt es schon fertige Modelle, die meine Aufgabe lösen können?

# DIZ Konfigurator Architektur

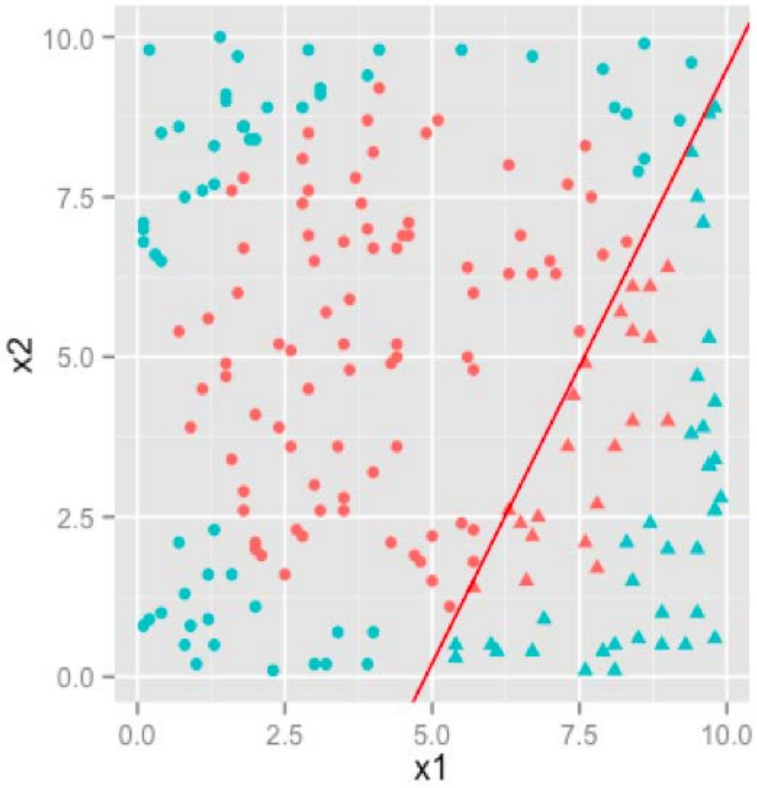


# DIZ Kofigurator Motivation




Fiktiver Feature-Raum (http://www.edvancer.in)

# DIZ Kofigurator Motivation: Logistic Regression



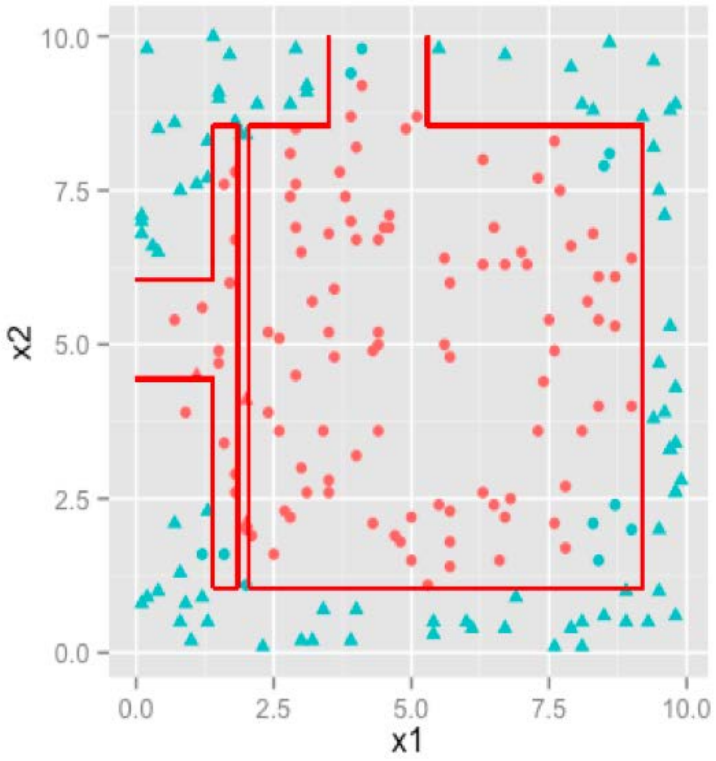

Logistic Regression:

Klassifikatoren trennen den Feature-Raum in zwei Teile (rote Linie).

Problem: Oft sind Feature-Räume nicht linear trennbar.

Fiktiver Feature-Raum (http://www.edvancer.in)

# DIZ Kofigurator Motivation: Entscheidungsbaum



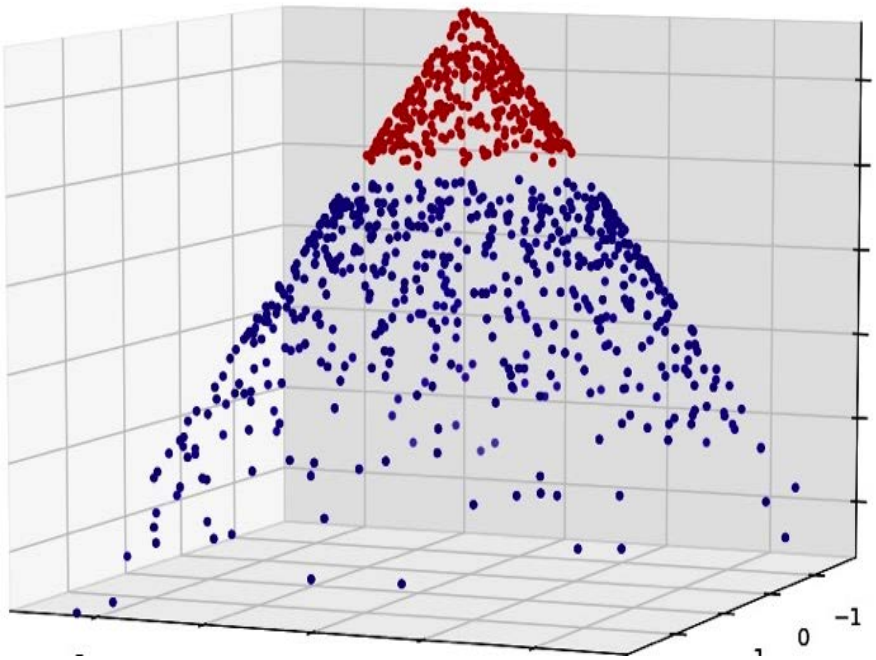

Entscheidungsbäume:

Klassifikatoren trennen den Feature-Raum durch mehrere Linien.

- Vorteile bei nicht linear trennbaren Feature-Räumen, aber …
- langsamer in der Ausführung als Logistic Regression sowie
- höheres Risiko des Overfittings.

Fiktiver Feature-Raum (http://www.edvancer.in)

# DIZ Kofigurator Motivation: Kernel Funktion



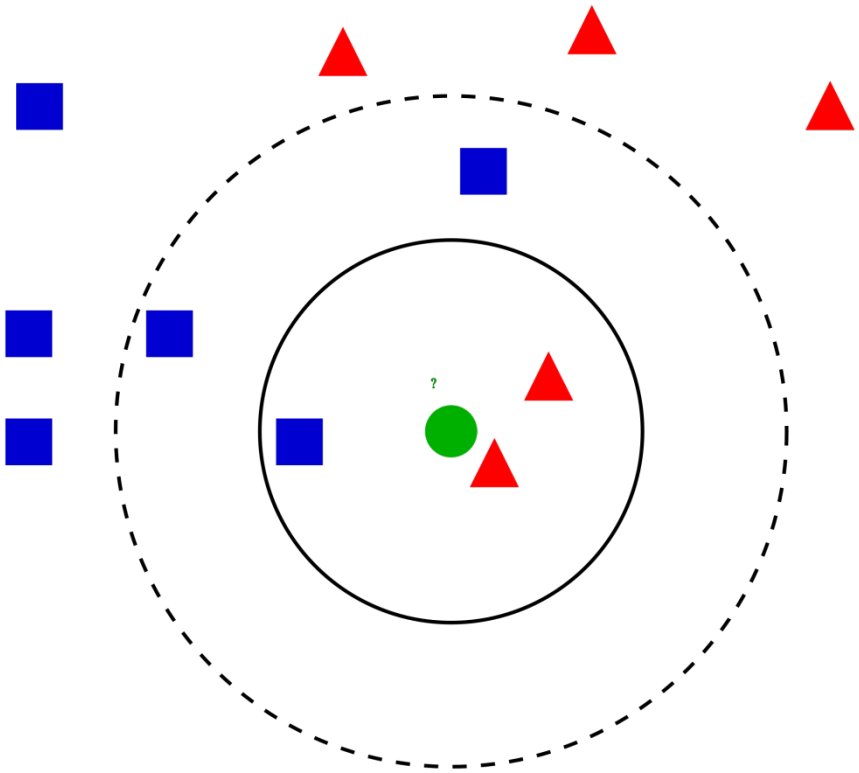

Fiktiver Feature-Raum
(http://www.edvancer.in)

Kernel Funktionen:

Überführen Instanzen in einen höherdimensionalen Raum, in dem die Instanzen linear trennbar sind.

- Vorteile bei nicht linear trennbaren Feature-Räumen, aber …
- langsamer in der Ausführung als Logistic Regression,
- geringeres Risiko von Overfitting.

# DIZ Kofigurator Motivation: KNN




Wikimedia Commons

K Nearest Neighbors:

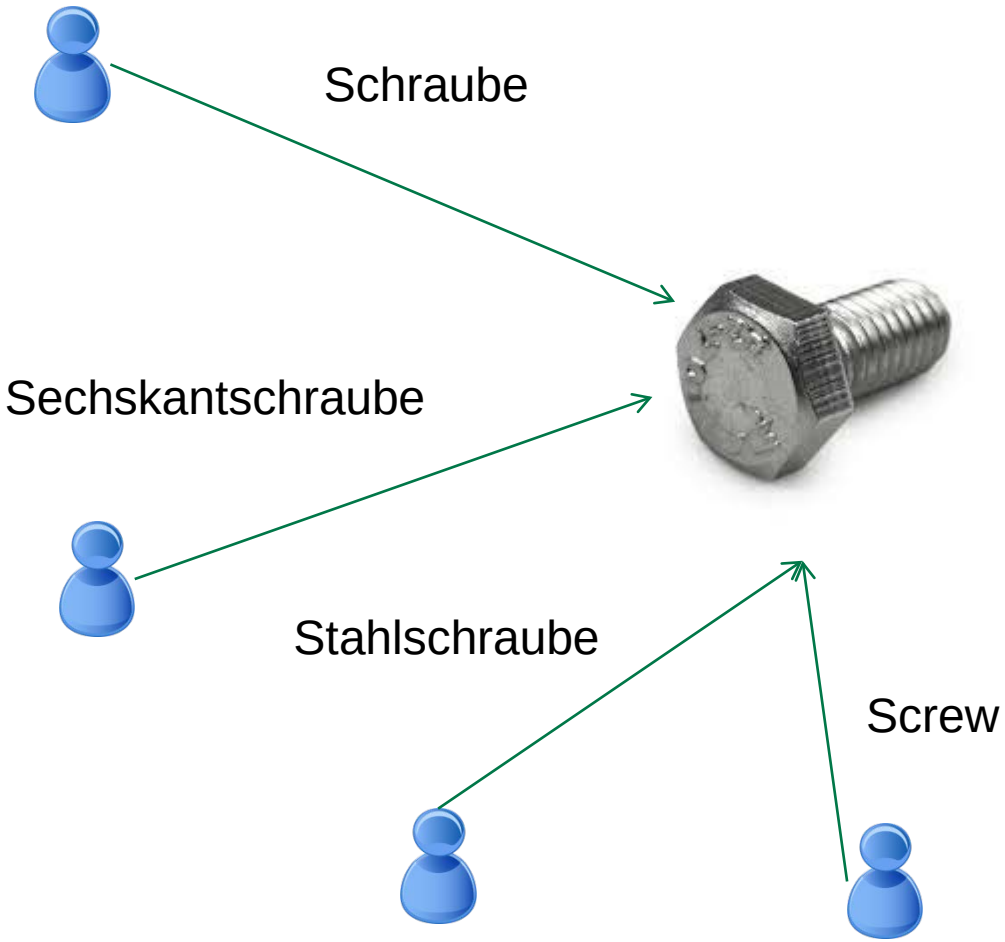
- Kein Training notwendig, aber …
- die Klassifikation ist vergleichsweise langsam (je nach verwendetem Ähnlichkeitsmaß)

# IFCC Use Case: Stammdaten Klassifikation

# Stammdaten: Definition

- Daten, die über einen längeren Zeitraum unverändert bleiben.
- Enthalten Informationen, die in gleicher Weise immer wieder benötigt werden.
- Sowohl bei der Einführung, als auch im laufenden Betrieb der IT-Infrastruktur und Transaktionen eines Unternehmens (wie bspw. eines Warenwirtschafts-Systems, Produktionsplanungs- und Steuerungssystems, etc.) kommt der Pflege von Stammdaten eine große Bedeutung zu.

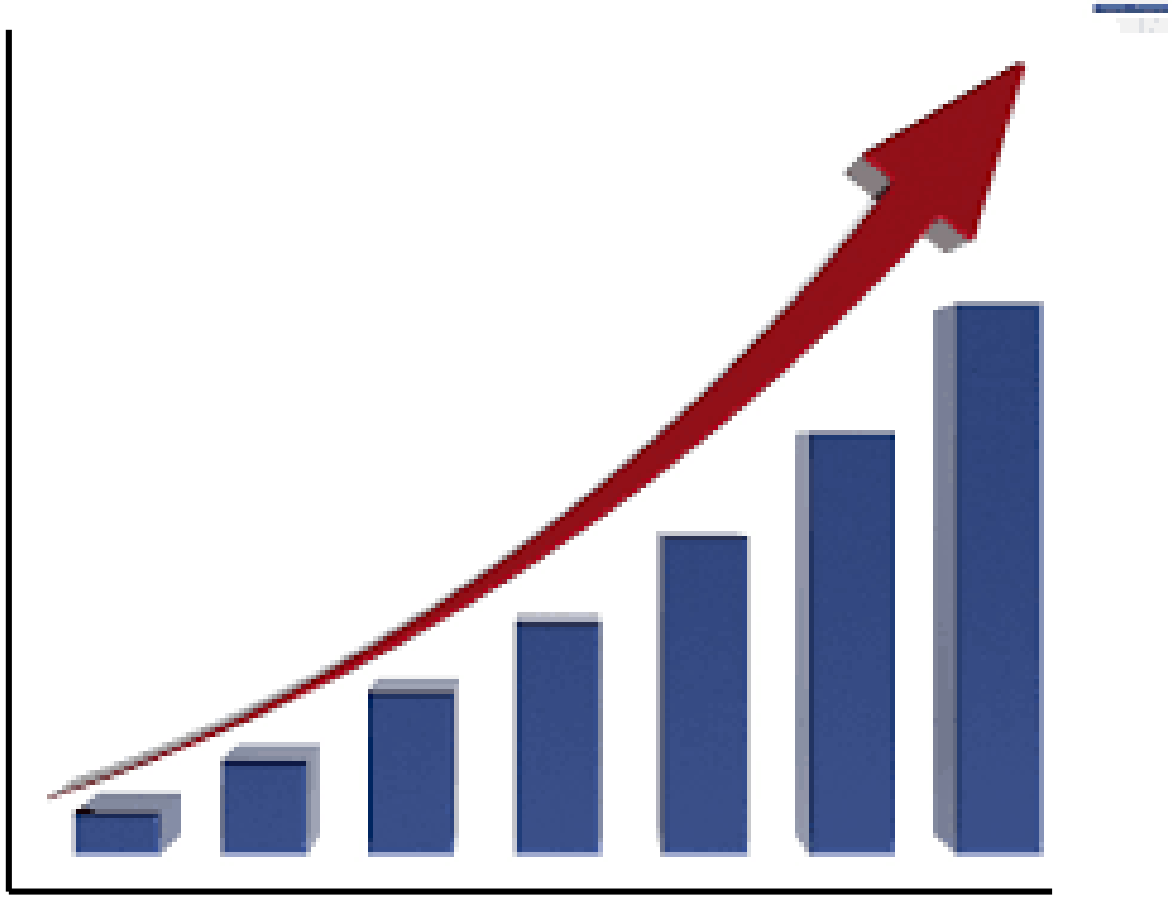
# Motivation: Klassifikation von Stammdaten




Menschen:

- sprechen verschiedene Sprachen und
- benutzen auch innerhalb einer Sprache verschiedene Wörter um gleiche Objekte zu bezeichnen

# Motivation: Klassifikation von Stammdaten



# Lösung

- Auswahl eines Klassifikationsystemes (z.B. eCl@ss)
- Klassifikation der Stammdaten anhand des Systems

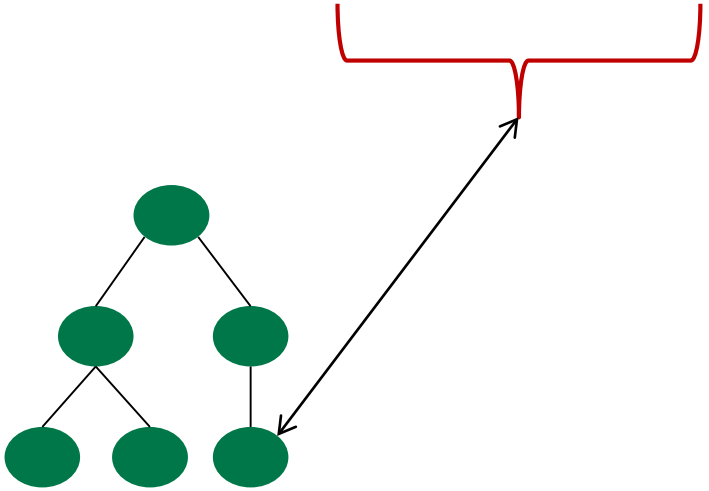
Problem: Manuelle Klassifikation ist sehr aufwändig und kostenintensiv.

# Klassifikation von Stammdaten

## Beispiel von Stammdatum

| SD | LD | Code | CodeDesc |
|---|---|---|---|
| VORLEGIERUNGSMETALL-ZN1 99,99% | ZINK 99,99% HÜTTENZINK IN FORM VON STÜCKEN IN CA. 1 KG GESCHNITTEN, GEM. EINKAUFSSPEZIFIKATION EKS-ASW-000030; REVISION 3, GÜLTIG AB DEM 06.01.2012 | 38150401 | Zink |

eCl@ss®
CLASSIFICATION AND PRODUCT DESCRIPTION



# Klassifikation von Stammdaten

## Aktueller Zustand

Der Kunde liefert eine Menge von Produktbeschreibungen

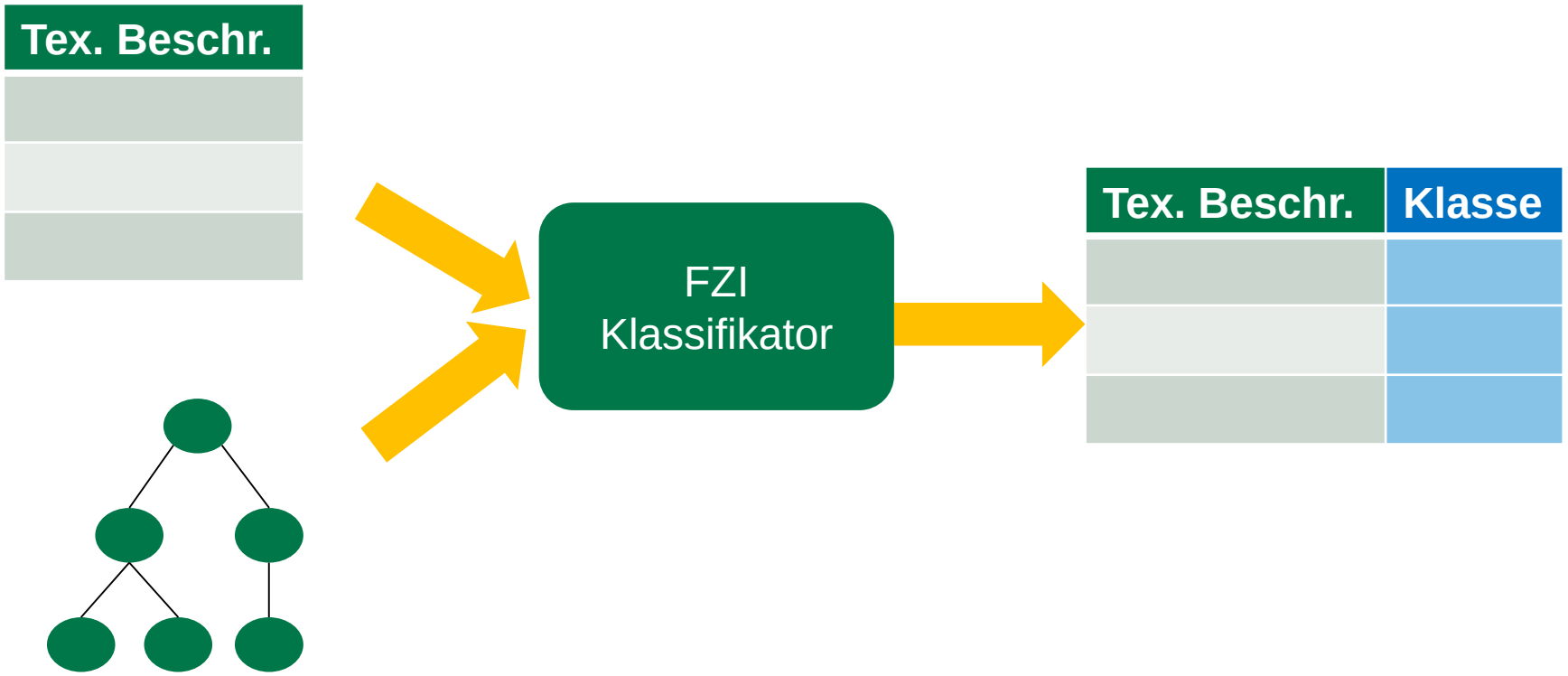
Ein semiautomatischer Ansatz klassifiziert die Produkte

# Klassifikation von Stammdaten

Nachteile

- Die Präzision ist stellt nicht IFCC zufrieden.
- Manuelle Pflege von Regeln ist sehr aufwändig und wird in der Regel nicht mit gleichbleibender Qualität durchgeführt.

# Unser Ansatz



# Unser Ansatz: FZI Klassifikator

FZI
Klassifikator

| Text |
|---|
| |
| |
| |

TFIDF/Count
Vectorizer

| w1 | w2 | w3 | w4 | w5 | w6 | w7 |
|---|---|---|---|---|---|---|
| 1 | 5 | 7 | 4 | 2 | 8 | 0 |
| 0 | 2 | 6 | 3 | 1 | 7 | 4 |
| 9 | 1 | 0 | 8 | 4 | 1 | 0 |

Feature Selection

| w1 | w2 | w3 | w4 |
|---|---|---|---|
| 1,5 | 3,5 | 2,7 | 1,7 |
| -0,8 | -3,6 | 5,4 | 2,8 |
| 2,3 | 1,45 | -1,6 | -0,3 |

Supervised ML
Klassifikator

| Tex. Beschr. | Klasse |
|---|---|
| | |
| | |
| | |

# Evaluation

- Datensatz mit 108.636 Produkten

- 10-Fold Cross Validation

| Hierarchy level | RandomForest | KNN |
|---|---|---|
| 1 | 0,84 | 0,85 |
| 2 | 0,77 | 0,77 |
| All | 0,687 | 0,693 |

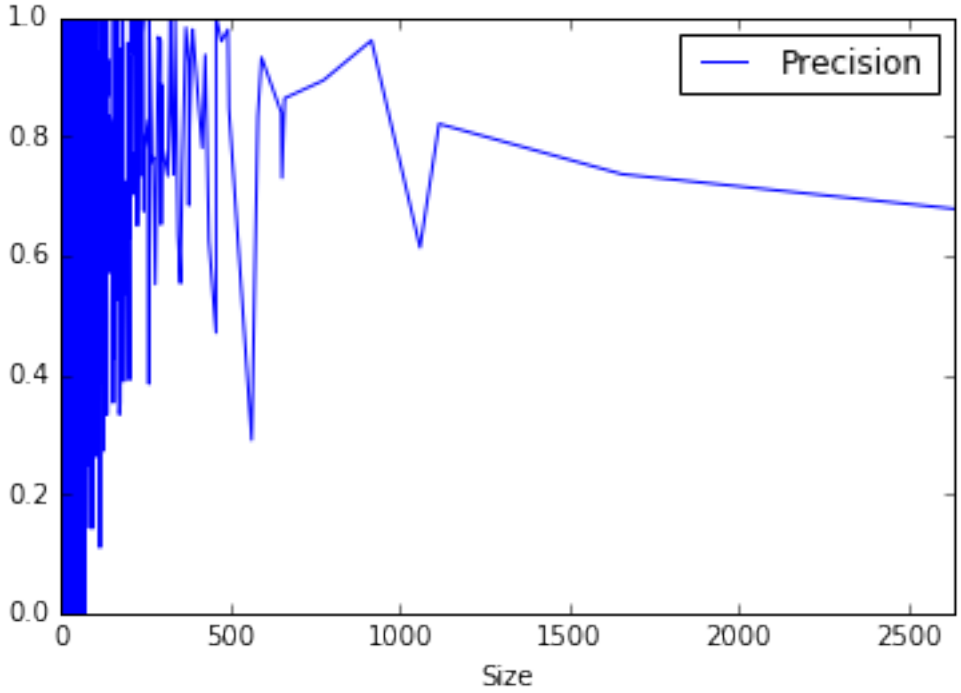
Präzision der Klassifikation in verschiedenen Ebenen der Hierarchie

| Aufgabe | KNN | Random Forest |
|---|---|---|
| Training | 150s | 40 Min |
| Klassifikation | 125s | 7s |

Trainings- und Klassifikationszeit für KNN und Random Forest

# Evaluation

- Präzision – Anzahl an Trainingsdaten



# Evaluation

- Datensatz mit **785.883** Produkten

- 10-Fold Cross Validation

| Hierarchy level | RandomForest | KNN |
|---|---|---|
| 1 | | 0,928 |
| 2 | | 0,90 |
| All | | 0,827 |

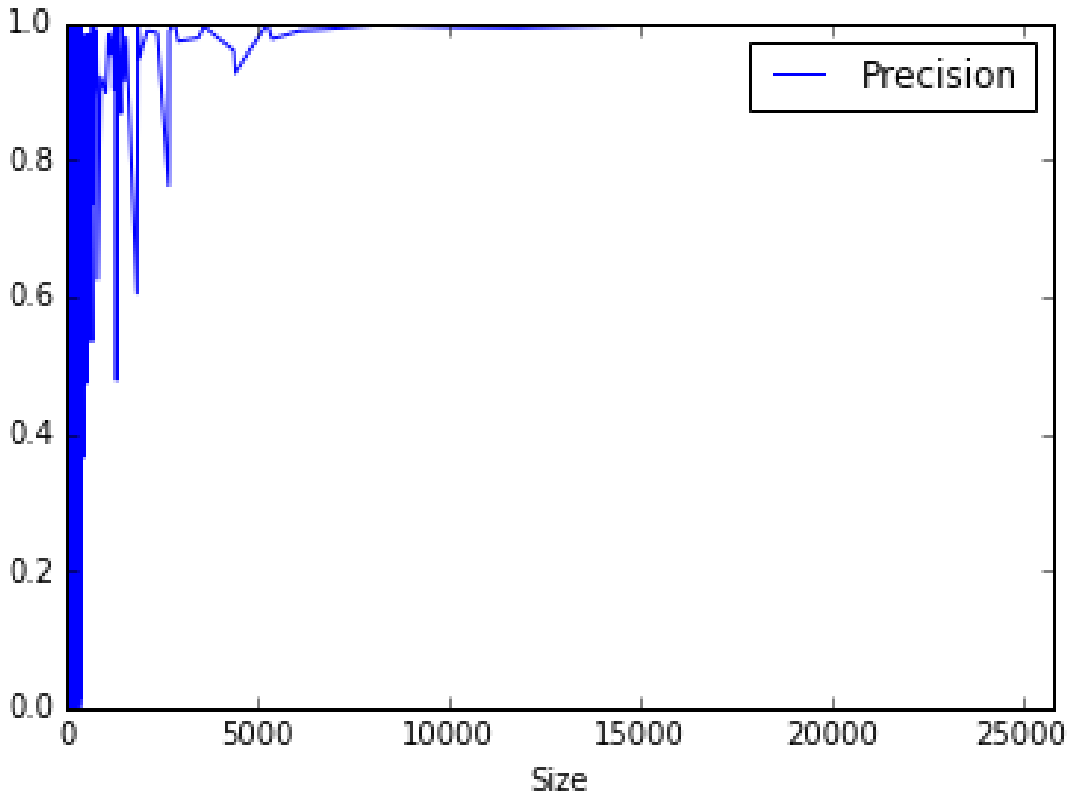
Präzision der Klassifikation in verschiedenen Ebenen der Hierarchie

| Aufgabe | KNN | Random Forest |
|---|---|---|
| Training | 8 Min | Nicht ausführbar |
| Klassifikation | 7 Min | Nicht ausführbar |

Trainings- und Klassifikationszeit für KNN und Random Forest

# Evaluation

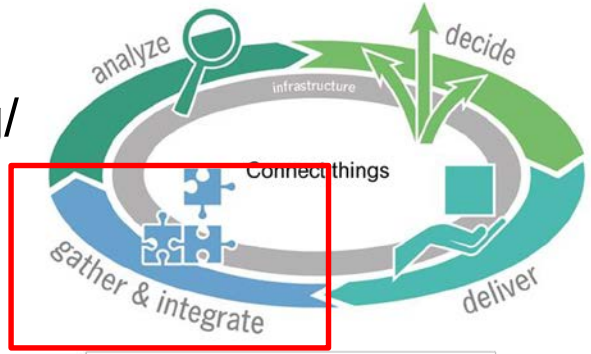
- Präzision – Anzahl an Trainingsdaten



# Lessons Learned

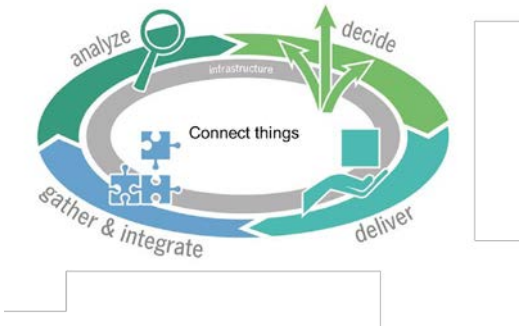
- KNN-/Random-Forest Methoden erreichen zufriedenstellende Präzision, dauern jedoch zu lange oder benötigen zu viel Speicher für iterative/agile Stammdatenklassifikation
  - Heuristiken oder Parallelisierung über Big-Data-Stack notwendig

- Stammdatenklassifikation sehr komplex zum vollständigen Automatisieren im "Konfigurator"
  - Insbesondere bei Datenvorverarbeitung
  - Weitere Fallstudien und Generalisierung/ Formalisierung notwendig



# Zusammenfassung

- Datengetriebene Entscheidungsunterstützung ist in vielen Domänen (Robotik uvm.) ein Thema.




- Im Digitalen Innovationszentrum (DIZ) entsteht aktuell ein "Konfigurator" für die teilautomatische Entscheidungsunterstützung.

DIZ DIGITALES INNOVATIONS ZENTRUM

- Eine Fallstudie zur automatischen Stammdatenklassifikation mit der IFCC GmbH zeigt vielversprechende Ergebnisse und weitere spannende Forschungsthemen

IFCC master.data.management

# Auswahl weiterer Industrieprojekte und öffentlich geförderter Projekte




FZI

DIZ DIGITALES INNOVATIONS ZENTRUM

IFCC master.data.management

# Danke!

**Kontakt**:

Ignacio Traverso-Ribon: traverso@fzi.de

Benedikt Kämpgen: kaempgen@fzi.de

FZI FORSCHUNGSZENTRUM INFORMATIK