DIH SÜD - DIGITAL INNOVATION HUB SÜD

Big Data

Wie man aus Daten Mehrwert für das Unternehmen generiert

19.05.2021

Der DIH SÜD wird unterstützt von:

BIG DATA

Wie man aus Daten Mehrwert für das Unternehmen generiert

DIPL.-ING. HERMANN KATZ

Institut für Wirtschafts- und Innovationsforschung POLICIES

Forschungsgruppe „Datenanalyse und statistische Modellierung“

JOANNEUM RESEARCH Forschungsgesellschaft mbH

hermann.katz@joanneum.at

Graz, 19. Mai 2021

Programm

- Motivation und Vorbemerkungen
- Big Data – Einführung und Begriffsbestimmung
- Von Big Data zu Smart Data
- Chancen / Risiken von Big Data für Unternehmen
- Nutzen der Auswertungen von großen Datenmengen
- Vorgangsweise / Strategie bei datengestützten Fragestellungen
- Anwendungsbeispiele (Predictive Analytics, Predictive Maintenance)
- Zusammenfassung

Zeitplan

09:00 – 09:30: Vorstellung und Erwartungshaltungen

09:30 – 10:00: Einführung und Begriffsbestimmungen

10:00 – 11:00: Von Big Data zu Smart Data

11:00 – 11:30: Kaffeepause

11:30 – 13:00: Chancen, Risiken und Nutzen für Unternehmen

13:00 – 14:00: Mittagspause

14:00 – 15:00: Strategie bei datengestützten Fragestellungen

15:00 – 15:30: Kaffeepause

15:30 – 16:00: Praktische Fallbeispiele

16:00 – 16:30: Zusammenfassung und Resümee

Vorstellung

- Hermann Katz – Studium der Technischen Mathematik / Statistik
- Seit über 25 Jahren im Bereich Datenanalyse tätig
- Direktorstellvertreter des Instituts für Wirtschafts- und Innovationsforschung - POLICIES
- Leitung der Forschungsgruppe Datenanalyse bei JOANNEUM RESEARCH
- Leitung von vielen anwendungsorientierten Projekten
- Autor bzw. Mitautor von zahlreichen Publikationen
- Langjähriger Vortragender an Universitäten und Fachhochschulen

EINFÜHRUNG UND BEGRIFFSBESTIMMUNGEN

DIHSÜD

Der DIH SÜD wird unterstützt von:

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Motivation

- Digitalisierung führt zu Datenüberflutung
- Analyse von Daten wird zum Erfolgsfaktor für Unternehmen
- Sind aber tatsächliche große Datenbestände der Schlüssel zum Erfolg?
- Informationen werden aus Daten extrahiert
- Zur Verbesserung von Prozessen und Produkten sind meist nur wesentliche Informationen nötig!
- Einsatz von statistischen Methoden ist bei der möglichen Reduktion der Datenmenge ein wichtiges Instrument
- Big Data – viele Daten beinhalten oft große Redundanzen!
- Smart Data – Reduktion auf das Wesentliche ist Erfolgsfaktor!

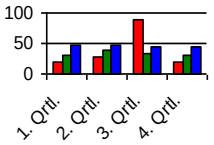
Vorbemerkungen

- Big Data ist Mittel zum Zweck
- Datengewinnung ist aber nur ein Teil bei der Analyse von Daten
- Fragestellung steht im Vordergrund
- Die Reduktion von großen Datenmengen auf aussagekräftige kleiner Submengen keine triviale Angelegenheit

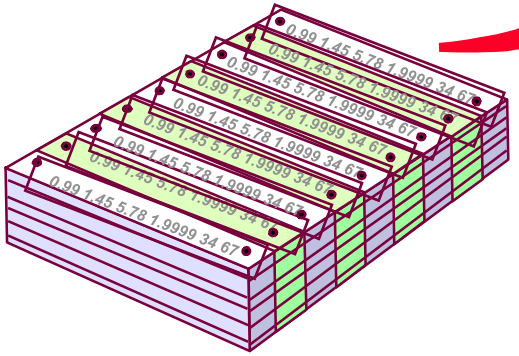
  → Ohne Erfahrung sollte man immer Fachleute einbeziehen

Daten ≠ Information

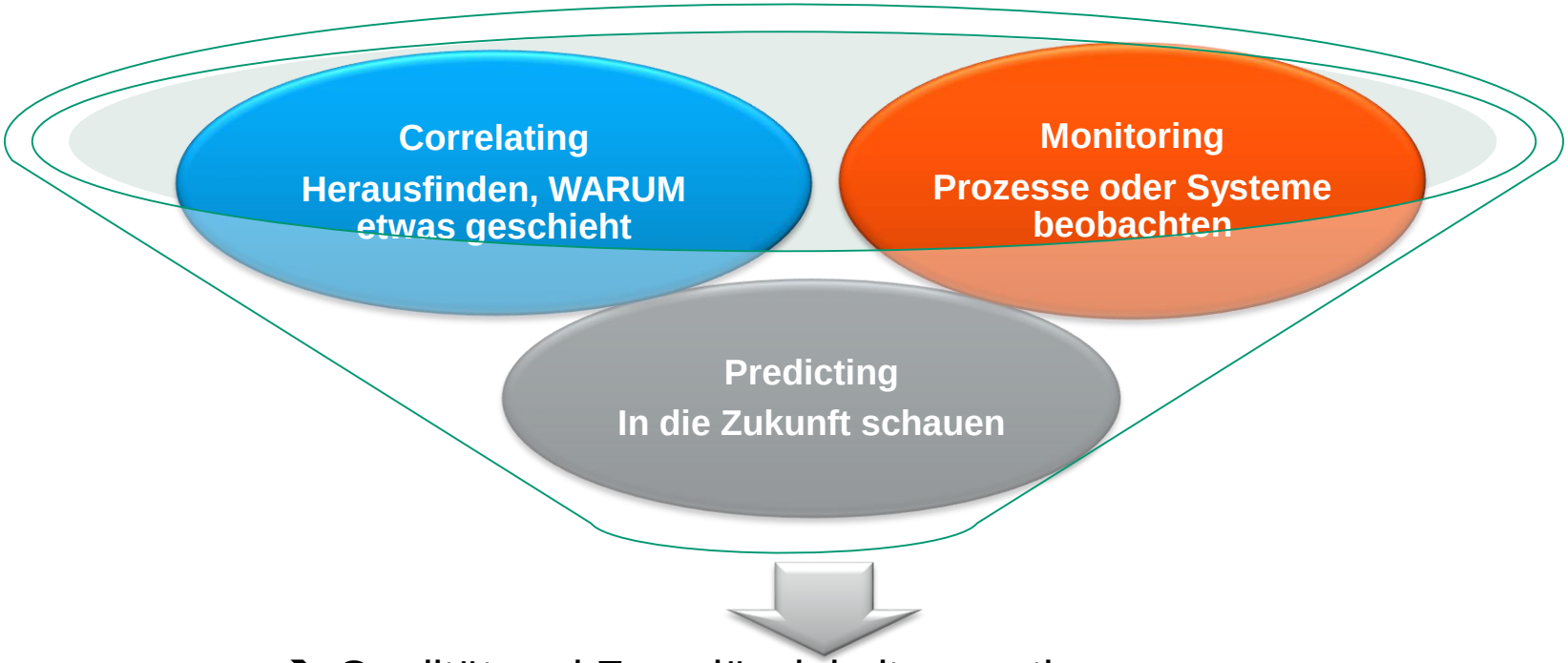
Information



Statistische Werkzeuge



Die Möglichkeiten statistischer Methoden …



- ➔ Qualität und Zuverlässigkeit garantieren
- ➔ Durchsatz erhöhen
- ➔ Kosten reduzieren
- ➔ Verkaufszahlen und Gewinne steigern
- ➔ Bessere Entscheidungen treffen

Charakteristika von Big Data

- Datenvolumen
  - Datenmenge ist anwendungsorientiert zu bewerten
- Geschwindigkeit der Datenverarbeitung
  - Verarbeitungsdynamik
- Veränderungsdynamik
  - Geschwindigkeit mit der sich der Ursachenkomplex verändert
- Datenstrukturen
  - Strukturierte und unstrukturierte Daten
- Unternehmerischer Mehrwert
  - Bewertung im unternehmerischen Kontext
- Validität der Daten
  - Messen die Daten auch das Richtige?

Umgang mit Big Data /1

- Große Datenmengen ergeben sich aufgrund von Umgebungsbedingungen
  - Sensorik bei Anlagen und Maschinen
  - Sicherheitsbedingungen
  - Kundendaten
  - Soziale Medien
  - …
- Daten werden gesammelt, aber nicht immer zielführend ausgewertet
  - Auswertungen erfolgen nur bei Störungen
  - Analyse der Daten wird von IKT-Fachleuten und nicht von Datenanalytikern durchgeführt
  - Möglichkeiten der Datenanalyse wird nur selten ausgeschöpft
- Drang nach Wissen wird aber in den seltensten Fällen erfüllt
  - Bandbreite der statistischen Möglichkeiten ist weitgehend unbekannt
  - Kooperationsbereitschaft mit diesbezüglichen Forschungseinrichtungen ist gering
  - Mehrwert für Unternehmen wird kaum erkannt

Umgang mit Big Data /2

- Systematisierung von großen Datenmengen
  - Welche Daten benötigt man um unternehmerischen Mehrwert zu generieren?
  - Nicht notwendige Daten sollten eliminiert werden
  - Sammlung nur von wesentlichen Datenbeständen
- Detaillierte Analyse der Daten auf Basis von konkreten Fragestellungen
  - Deskriptive und explorative Analyse der Daten
  - Entwicklung von statistischen Modellen
  - Identifikation von Daten mit Mehrwert
- Daten werden verwendet um Wissen zu generieren
  - Kooperation mit Fachexperten
  - Analyse der historischen Daten
  - Einsatz von Prognosemodellen

Datenanalyse als Teil der Unternehmenskultur

- Datenanalyse sollte integraler Bestandteil in Unternehmen werden
  - Analyse von Daten generiert Wissen
  - Wissen verbessert Marktposition
  - Alleinstellungsmerkmale können identifiziert werden
  - Vorteile gegenüber Mitbewerbern
- Detaillierte Analyse der Daten auf Basis von konkreten Fragestellungen
  - Deskriptive und explorative Analyse der Daten
  - Einsatz von statistischen Methoden
  - Entwicklung von zielführenden Softwaretools
- Erworbenes Wissen führt zu Verbesserungen
  - Verbesserung von Produktqualität
  - Einleitung von konkreten Präventivmaßnahmen
  - Steigerung der Ressourceneffizienz

VON BIG DATA ZU SMART DATA

DIHSÜD

Der DIH SÜD wird unterstützt von:

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Gliederung der Statistik

Beschreiben

Deskriptive Statistik

Schlüsse ziehen

Inferenzstatistik

Deskriptive Statistik

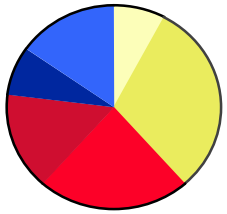
- Deskriptive (beschreibende) Statistik
  - Instrumentarium zur Beschreibung von Daten
  - Vorstufe zur schließenden Statistik

Ziel:

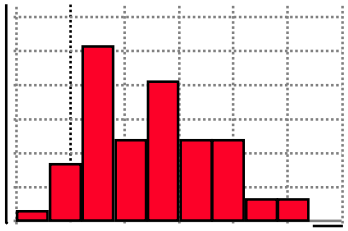
Beschreibung, Strukturierung, Verdeutlichung, Darstellung umfangreichen, unübersichtlichen Datenmaterials

- Methoden:
  - Grafische Darstellungen
  - Kennzahlen (Maßzahlen)

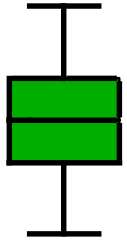
Unterschiedliche Grafiken



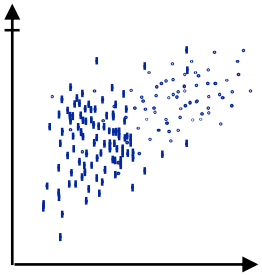
Piechart



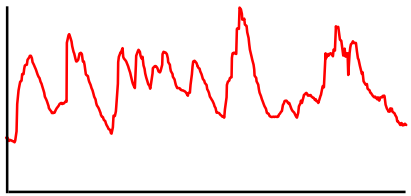
Histogramm



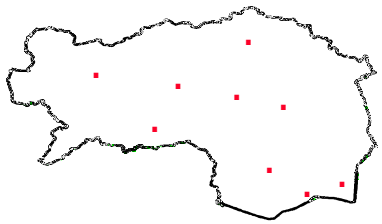
Boxplot



Scatterplot



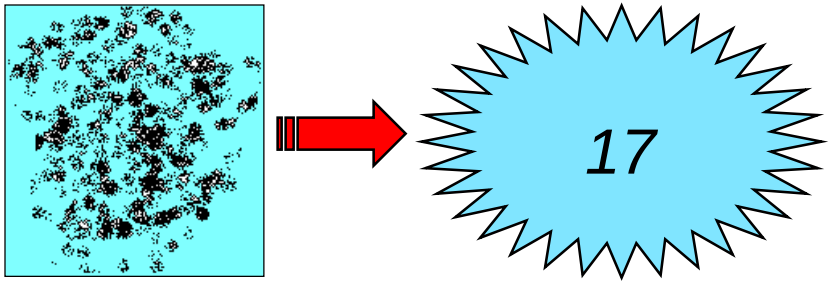
Zeitreihe



Map

Allgemeine Kennzahlen

Sinn von Kennzahlen ...

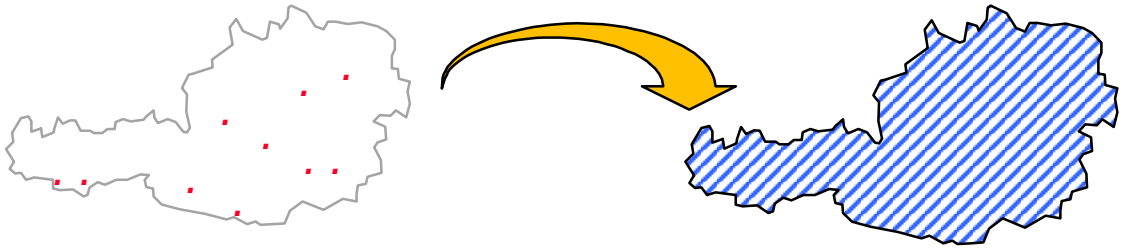


- *Reduzierung von Komplexität*
- *Verdichtung von Information*

... die Sache auf den Punkt bringen!

Schließende Statistik

Schätzen: *Schluss von einer Stichprobe auf die Gesamtheit*



Testen: *Überprüfen einer Vermutung (Hypothese) über die Population mittels einer Stichprobe*

H_0:
Die Österreicher sind im Durchschnitt kleiner als 1,70m

Parsimonitätsprinzip

- Die Größe der Datenbestände ist nicht der Schlüssel zum Erfolg
- Bei der Datenanalyse fokussiert man sich auf die wesentlichen Daten
  - Parsimonitätsprinzip
- Erst durch eine detaillierte Analyse kann man zwischen „wichtigen“ und „unwichtigen“ Daten differenzieren
- Modellhafte Zugänge ermöglichen eine Quantifizierung der Wichtigkeit von Einflussgrößen
- Ursachenkomplex darf sich nicht wesentlich ändern
- Laufende Validierung der erarbeiteten Modelle

Von Big Data zu Smart Data

Datenmanagement

Aus der Datenflut heutzutage ist es unabdingbar, sich mit der Kompetenz des effizienten Datenmanagements zu beschäftigten

- Data Engineering
  - Daten sammeln, speichern, aufzubereiten und zur Verfügung stellen
  - Datenbanken, Data-Warehousing-Tools bzw. Dateninfrastruktur (PostgreSQL) sind zuverlässige Grundlagen
- Data Science
  - Daten aus Datenbank analysieren, visualisieren, modellieren sowie Wissen generieren
- Strukturierte Daten vs. Unstrukturierte Daten
  - Jeder unstrukturierte Datensatz kann in einen Dataframe umgebaut werden (Voraussetzung für Modellierung von Daten)
  - ETL Process aus verschiedensten Datenquellen
    - Automatismus einer Datenaufbereitungsroutine: Extraction → Transformation → Loading
  - Erfahrungsgemäß ist eine qualitative Datenaufbereitung mit Plausibilitätsüberprüfung unerlässlich aber mit Aufwand verbunden
- Zuverlässige Dateninfrastruktur und Datenmanagement ist Grundlage
  - Effizientes Datenmanagement (PostgreSQL, R, Python etc.)
  - Bewältigung und Validierung von größeren Datenmengen

Methoden

- Datenanalyse
  - Statistische Methoden werden eingesetzt, um Zusammenhänge zu identifizieren und zu quantifizieren, Datenreduktion ist Teil des Analysevorgangs.
- Künstliche Intelligenz
  - Teilgebiet der Informatik, welches sich mit der Automatisierung intelligenten Verhaltens und dem maschinellen Lernen befasst. Der Begriff ist insofern nicht eindeutig abgrenzbar, als es bereits an einer genauen Definition von „Intelligenz“ mangelt. Dennoch wird er in Forschung und Entwicklung verwendet.
- Maschinelles Lernen
  - Ein künstliches System lernt aus Beispielen und kann diese nach Beendigung der Lernphase verallgemeinern. Dazu bauen Algorithmen beim maschinellen Lernen ein statistisches Modell auf, das auf Trainingsdaten beruht. Das heißt, es werden nicht einfach die Beispiele auswendig gelernt, sondern es „erkennt“ Muster und Gesetzmäßigkeiten in den Lerndaten.
- Deep Learning
  - Deep Learning unterscheidet sich vom maschinellen Lernen, indem es Maschinen in die Lage versetzt, über die verfügbaren Daten hinaus zu lernen. Das beinhaltet die Fähigkeit, Informationen zu analysieren und zu bewerten, um logische Schlüsse zu ziehen, Lösungswege auszuwählen und aus Fehlern zu lernen.

Verwertung von Daten

- Was kann man mit diesen Daten machen?
  - Auf datenanalytischer Erfahrung beruhend ist eine richtige Auswahl für ein geeignetes Modell entscheidend
    - **Statistische Modellierung** (Interpretierbarkeit, Nachvollziehbarkeit, keine Black-Box, Plausibilität/Ausreißer können einfach detektiert werden, Konfidenz etc.)
    - Lernende Systeme aus dem Bereich des **Machine Learnings** (Automatisierung)
  - Beide Gebiete haben Vor- u. Nachteile (Nachvollziehbarkeit, Black-Box, Automatisierung etc.)
  - Machine Learning: Algorithmus kann trainierte „simple“ Entscheidung schneller treffen als der Mensch (nicht unbedingt besser)
  - Meistens ist algorithmisches Lernen nicht die bessere Wahl → unser Fokus liegt nach wie vor auf statistischer Methodik
- Erfolgreiche Generierung von Wissen aus Daten
  - Handlungsmöglichkeiten abzuleiten, Muster zu erkennen, valide Prognosen treffen zu können oder ein funktionierendes Monitoring-System zu etablieren etc.

Statistische Modellierung

- Lineare Modelle
  - Regressionsanalyse (multiple, multivariate)
  - Generalisierte lineare Modelle (GLMs)
    - Normal, Bernoulli / Binomial (logistische Regr. Logit für Risikoanalysen), probabilistische, Gamma, Poisson etc.
  - Generalisierte Additive Modelle (GAMs)
    - Sehr flexible Modellierung: Parametrisch und mit smoothing splines
  - Erweiterung GAMLSS
    - Modellierung nicht nur von Location (Erwartungswert) und Scale (sigma); auch Shape-Parameter (Schiefe und Kurtosis) sind modellierbar → 4 Verteilungsparameter
      - Filtern von sehr exotischen Verteilungen (light/hard-tailed) → z. B. Modellierung von seltenen Ereignissen; **Risikoanalyse**
  - Reliability/Zuverlässigkeit/Survival Models
    - Weibull, Log-normal etc. oder auch Proportional Hazard (Cox)
    - Verwendung auch von „zensierten“ Daten
  - Objektivierung subjektiver Eindrücke
  - …
- Nichtlineare Modellierung
  - Wird eigentlich nicht benötigt, da sich (fast) alles auf lineare Modelle transformieren lässt bzw. sich mit flexiblen Modellen (GAMs etc.) viel besser modellieren lässt (vor allem auch stabiler)

Selbstlernende Algorithmen (Machine Learning)

- Random Forests (RF)
  - Beliebig viele Variablen können in Betracht gezogen werden
  - Es wird eine beliebige Anzahl von Entscheidungsbäumen generiert
  - Für einen fertigen RF kann für unbekannten Input eine Prognose berechnet werden
    - Je nach Häufigkeit pro Baum wird über alle Bäume (also Forest) eine Prognose mit der größten Häufigkeit ausgegeben
  - Können sowohl zur Klassifikation (diskret) sowie Regression (stetige Variablen) eingesetzt werden
  - Intuitives Prinzip
- Künstliche Neuronale Netze (KNN, Deep learning)
  - Sehr kompakte Modellierung komplexer funktionaler Zusammenhänge
  - Tiefe des Netzes (Anzahl hintereinander geschalteter Schichten/Layer) beliebig wählbar, aber Trade-Off zwischen Transparenz/Interpretierbarkeit und Präzision/Vorhersagekraft
- Geometrisch betrachtet sind RF und KNN eine komplexe Regression aus allen möglichen Features (unabhängige Variablen)
- Support Vector Maching (SVM)
  - Für stetige Zielgrößen als auch für diskrete verwendbar
    - Hauptsächlich aber für **Klassifikationen** geeignet
  - Daten werden (geometrisch betrachtet) durch Hyperebenen größtmöglich separiert

Statistische Modellierung vs. Machine Learning

Statistische Modellierung

- Einfachere Interpretierbarkeit
- Beschreibung und Erklärung von Zusammenhängen einfacher
- Variablenselektion und Dimensionsreduktion erforderlich

Machine Learning

- Zumeist höhere Modellprognose (Gefahr von „overfitting“)
- Weniger Modellvoraussetzungen (mathematische Annahmen)
- Robust für hochdimensionale Daten
- Im Bereich Image Processing überlegen

Statistische Modellierung vs. Machine Learning

Vorhersagegüte vs. Interpretierbarkeit

Statistische Modellierung	Machine Learning
Interpretierbarkeit	Vorhersagegüte
Entscheidungsbaum	Random Forest
Logistische Regression	Support Vector Machine
Lineare Regression	Künstliches neuronales Netz
Generalisiertes lineares Modell (GLM)	Bayes-Netze
Generalisiertes Additives Modell (GAM)	…
…	

Big Data vs. Smart Data

- Die Größe der Datenbestände ist nicht entscheidend
- Smart Data enthalten alle relevanten Informationen zur festgelegten Fragestellung
- Aufwand für Datenmanagement und Speicherung meist geringer
- Quantifizierte Zusammenhänge (Modelle) leicht interpretierbar
- Reproduzierbarkeit der Ergebnisse ist gegeben
- Kausale Zusammenhänge transparent abgebildet

CHANCEN, RISIKEN UND NUTZEN VON BIG DATA FÜR UNTERNEHMEN

DIHSÜD

Der DIH SÜD wird unterstützt von:

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Chancen von Big Data /1

- Eröffnet neue Möglichkeiten in allen Geschäftsbereichen des Unternehmens
  - Marketing
  - Produktentwicklung
  - Ressourcenmanagement
  - Forschung und Entwicklung
  - Produktion und Service
  - Logistik und Transport
- Erschließung neuer Geschäftsfelder und Branchen
- Unterstützung in der Entscheidungsfindung

Chancen von Big Data /2

- Entwicklung eines besseren Kundenverständnisses
  - Präzision im Ansprechen von Kunden
  - Wissen über Kaufhistorie und Kaufverhalten generieren
- Erstellung von Marktprognosen
- Ressourceneffizientes Wirtschaften
  - Energieversorgung (Smart Grids)
  - Effizientere Mobilität (Erstellung von Bewegungsmustern in Verkehrssystemen)
  - Frühzeitige Fehlererkennung in Produktion
- Individualisierung von Anwendungen, Dienstleistungen und Produkten

Chancen von Big Data /3

- Erkennung von versteckten Mustern und Informationen
- Unterstützung beim Treffen von präventiven Entscheidungen
  - Abstände bezüglich Maschinenausfälle erkennen
  - Bestimmung der Faktoren die zum Ausfall führen
- Optimierung von Arbeitsprozessen
- Trenderkennung
- Transparenz der Produktion (steigert Kundenzufriedenheit)

Risiken von Big Data /1

- Durch Datenkapitalismus wird Mensch zur Ware
  - Starker Eingriff in die Privatsphäre der Kunden
  - Datenschutz
  - Eigentumsfrage der Daten nicht eindeutig geklärt
- Risiken durch disruptive Geschäftsmodelle (Uber oder AirBnB)
- Verdrängung von Berufen und Wirtschaftszweigen
  - Ersetzung der menschlichen Arbeitskraft durch Maschine
- Kontrollverlust des Menschen
  - Algorithmen bestimmen über Milliarden an Börsen und handeln in Pikosekunden

Risiken von Big Data /2

- Chancenerhöhung Opfer eines Hackerangriffs zu werden
- Unüberschaubarkeit der Daten kann zu irreführenden Ergebnissen führen
- Falsche Verbindung zwischen unterschiedlichen Daten herleiten und dadurch falsche Entscheidungen treffen
- Wettbewerbsnachteile für kleinere Unternehmen

Nutzen der Auswertungen von großen Datenmengen /1

- Um Wissen zu erlangen, müssen Informationen verarbeitet werden, die wiederum aus Daten generiert werden
  - Verbesserung von Dienstleistungen, Prozessen und Produkten
  - Erhöhung der Zuverlässigkeit von Anlagen
  - Sicherheit erhöhen
  - Entwicklungszyklen verkürzen
- Zusammenhänge identifizieren
- Präventive Aktivitäten rechtzeitig einleiten
- Prognosen erstellen
- Durch Auswertungen müssen zukünftig weniger Daten erhoben werden
  - Weniger Speicherkapazitäten notwendig
  - Kosteneinsparung

Nutzen der Auswertungen von großen Datenmengen /2

- Schaffung von evidenzbasierten Fakten
  - Unterstützung bei Entscheidungen
  - Evaluierung von Maßnahmen
  - Festlegung von quantifizierten Grundlagen
- Reduktion auf eine aussagekräftige Datenmenge
- Rechtzeitig das richtige Wissen generieren
- Reaktionszeiten verkürzen
- Schneller wissen, was der Kunde sich erwartet

VORGANGSWEISE / STRATEGIE BEI DATENGESTÜTZTEN FRAGESTELLUNGEN

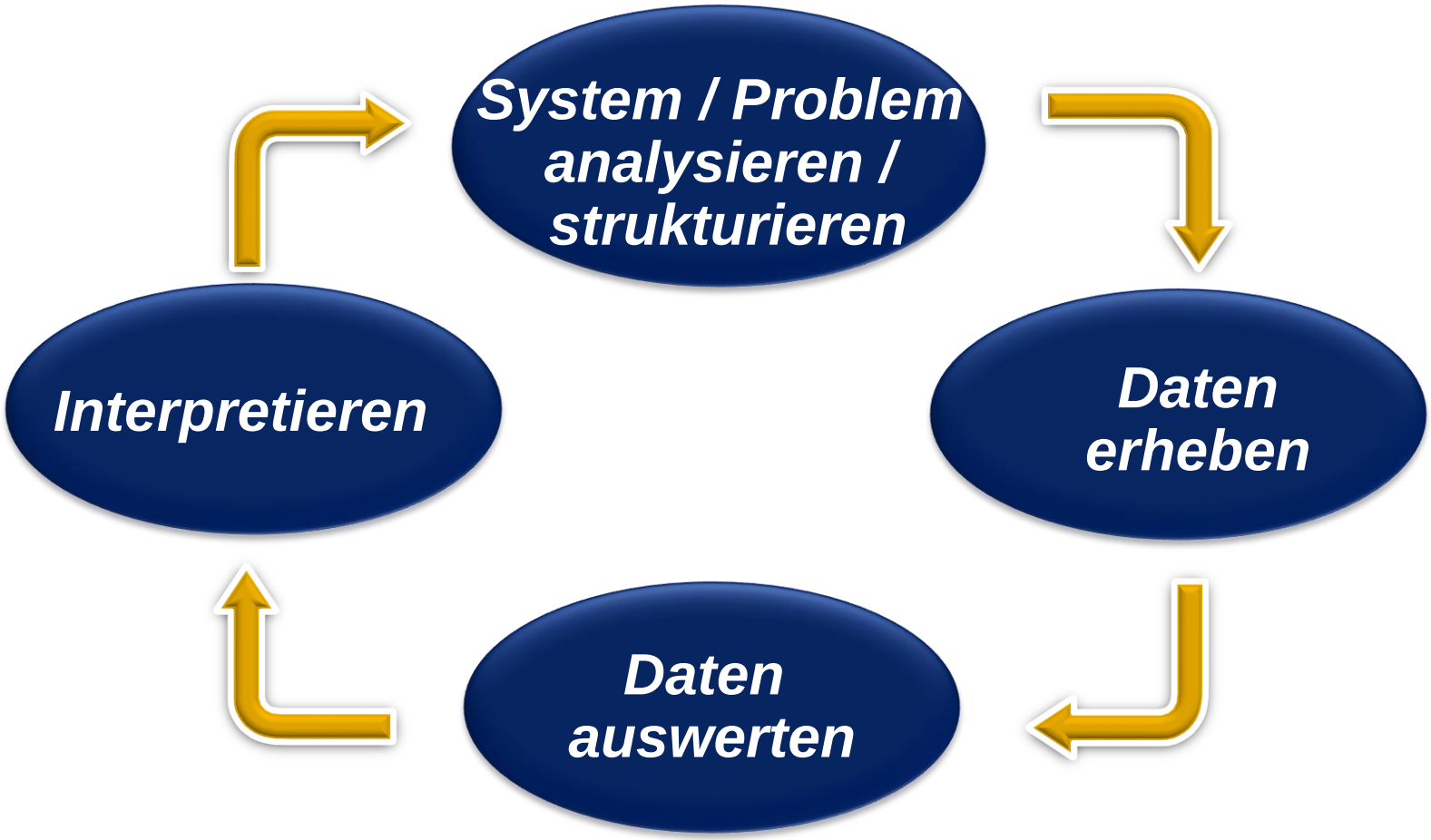
Der DIH SÜD wird unterstützt von:

FFG Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Wissenschaftliche Vorgangsweise in der Angewandten Statistik



Unser Vorgehensmodell …

Ask

- Was ist die spezifische Fragestellung?
- Analysieren der Anforderungen
- Analysieren des Prozesses bzw. des Systems

Get

- Daten aus bestehenden Quellen gewinnen
- Neue Datenquellen erschließen
- Daten fusionieren und prozessieren

Explore

- Plausibilität prüfen
- Analysieren
- Visualisieren

Model

- Detektion
- Klassifikation
- Prognose
- Optimierung
- Validierung

Communicate

- Neue Erkenntnisse vermitteln
- Präsentation
- Report
- Software-Tool

Systemanalyse - Planung

- Formulierung der sachspezifischen Fragestellung
  - Ziele festlegen
  - Fachwissen mit statistischem Know-how verbinden
- Merkmale festlegen
  - Skalierung, Eigenschaften
- Datenquellen konkretisieren
  - Fragebogendesign
  - Pretest
  - Projektdurchführung
- Grundgesamtheit – Stichprobe
  - Auswahl an einer repräsentativen Stichprobe

Datengewinnung

- Erhebungs- oder Versuchsplanung
  - Analyse der Auswahlgrundlage
  - Fragebogendesign
  - Pretest
  - Versuchsplanung
  - Festlegung des Stichprobenumfanges
  - Organisatorischer Ablauf der Datenerhebung – ev. Pilotstudie
- Datensammlung
- Dateneingabe – Online-Befragungen
- Überprüfung der Korrektheit der Daten

Statistische Auswertung

- Kritische Analyse der Urdaten
- Deskriptive und exploratorische Datenanalyse
  - Tabellen
  - Grafiken
  - Kennzahlen
- Inferenzstatistische Aussagen
  - Aussagen bezüglich der Grundgesamtheit
  - Überprüfung von Vermutungen (Hypothesen)
  - Modellierung z. B. Regression

Sachspezifische Entscheidungsfindung

- Aufbereitung der statistischen Ergebnisse für Entscheidungsfindung
- Sachspezifische Interpretation der Ergebnisse
- Ableitung von Maßnahmen
- Ev. Detailstudien

KFV Standard Reporting

Ask

Auswertung und Darstellung bemerkenswerter Aspekte der österr. Unfallstatistiken

Get the Data

KFV, Statistik Austria, internationale Datenbanken

Explore the Data

Tabellen und Grafiken

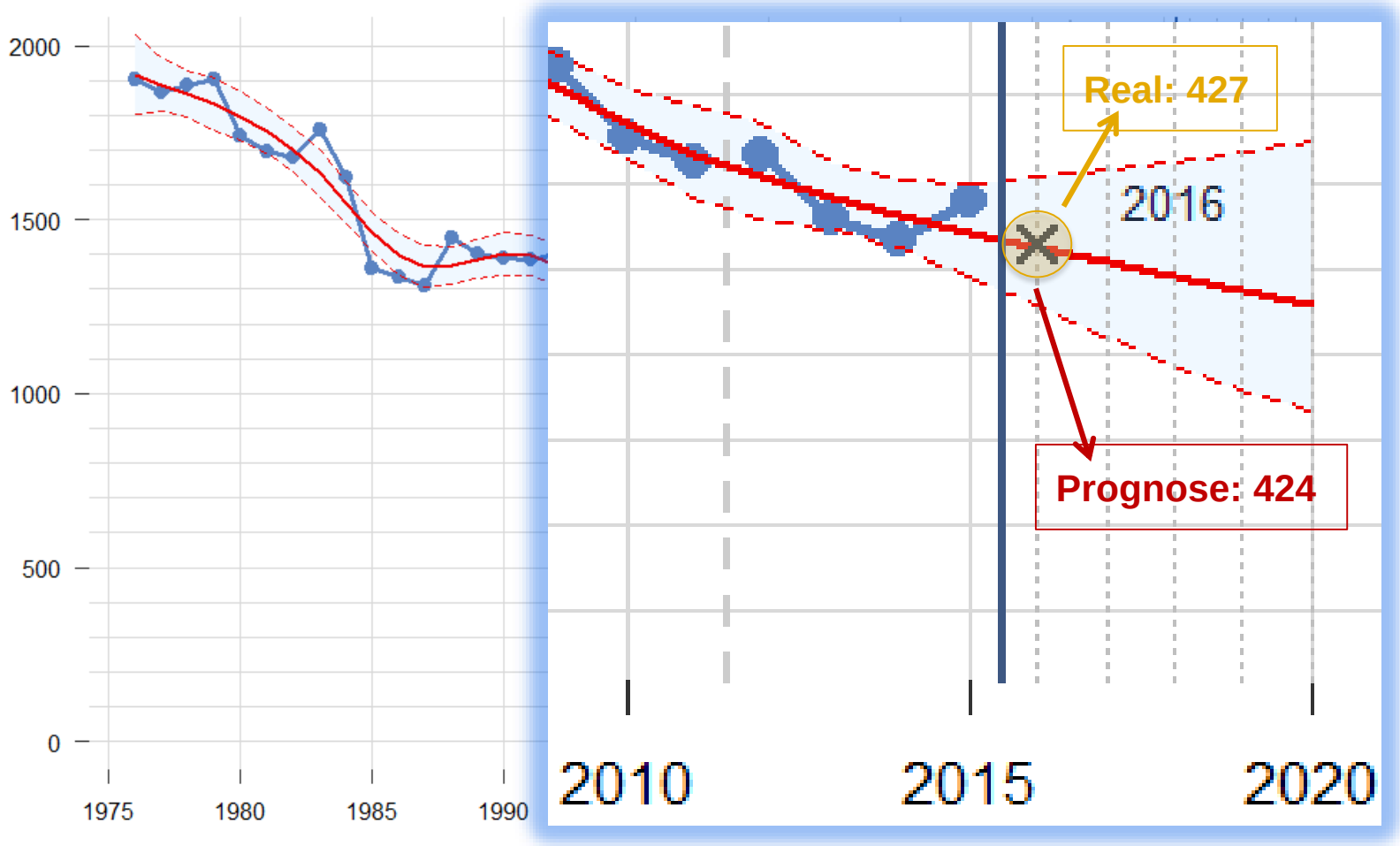
Model the Data

Vorhersagen in Zukunft

Communicate and visualize the result

Bericht

KFV Standard Reporting - Prognose der Todesfälle im Straßenverkehr



ANWENDUNGSBEISPIELE (PREDICTIVE ANALYTICS, PREDICTIVE MAINTENANCE)

DIHSÜD

Der DIH SÜD wird unterstützt von:

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

JOANNEUM RESEARCH POLICIES

.BK REPUBLIK ÖSTERREICH
BUNDESMINISTERIUM FÜR INNERES
BUNDESKRIMINALAMT

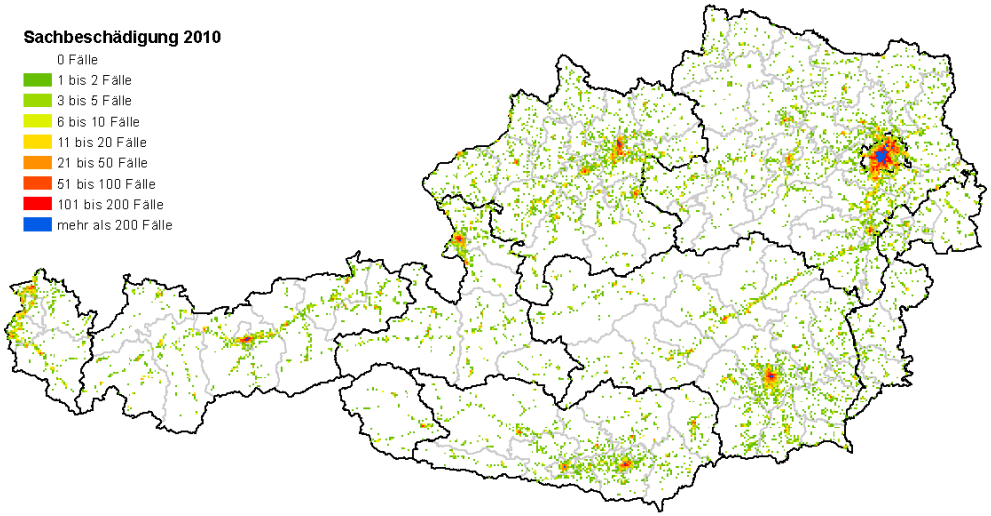
BASE of ACE

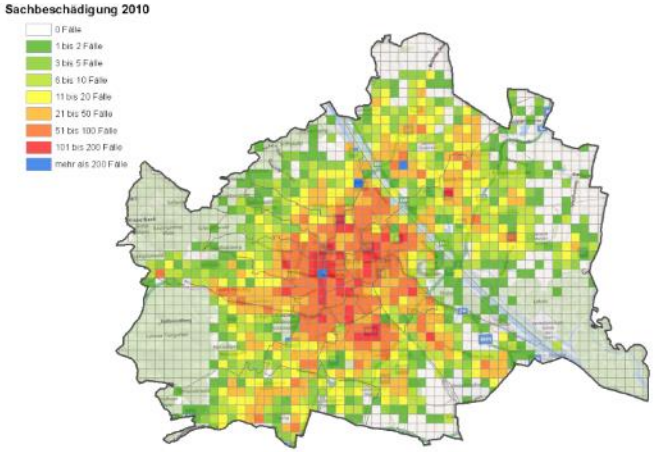
KIRAS Sicherheitsforschung

bmvit Bundesministerium für Verkehr, Innovation und Technologie

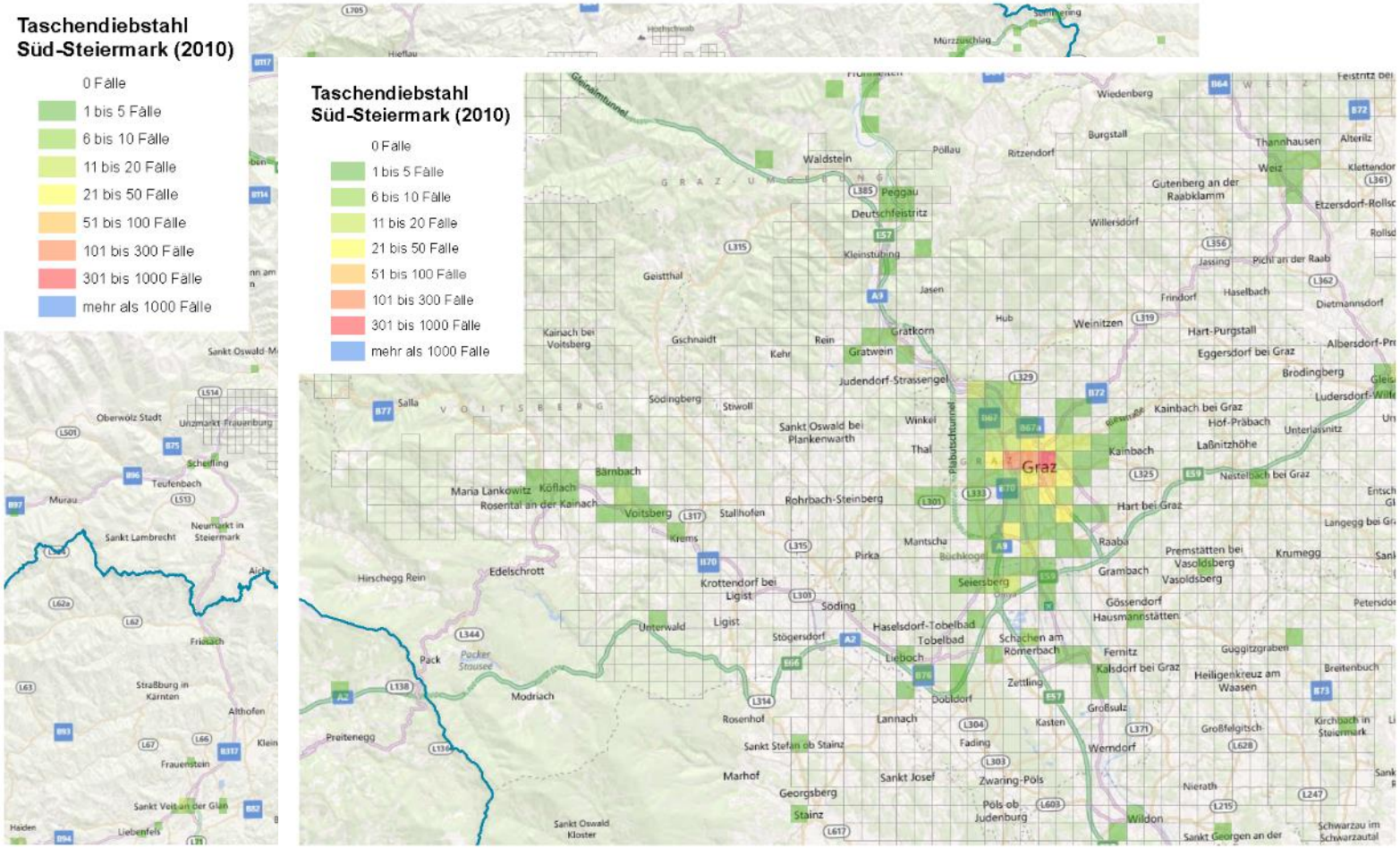
FFG

Austrian Crime Explorer: Analyse, Identifikation und Quantifizierung der Ursachen regionaler Unterschiede in der Kriminalität





Modellergebnisse Taschendiebstahl Region Süd-Steiermark



Modellergebnisse Regionenvergleich

- Taschendiebstahl

Einflussgröße	Region: Ost	Nord	West	Süd	Stmk	Südktn	WestT	WestV	Wien
hws	+	0	+	−	+	0	0		
sport	+	0	+	+	+	0	+		
tagestourismus	0	+	+	0	+	−	0		
hoehere bildung	0	+	0	0	0	0	0		
polizeidienststelle	−	−	−	−	−	−	−		
bahnhof	0	+	+	0	+	0	+		
einkaufszentrum	+	+	+	+	+	0	+		
naechtigungen pro ew sommer	−	−	0	+	0	0	0		
naechtigungen pro ew winter	+	+	+	0	−	0	0		
schueler gesamt pro 100ew	0	0	+	−	0	0	0		
schueler 10bis14 pro 100ew	0	0	−	0	−	0	0		
schueler matura pro 100ew	+	+	−	+	+	+	−		
arbst anz	+	−	0	−	−	+	+		
besch anz	0	+	0	+	+	0	0		
md polizeidienststelle	−	−	−	−	−	−	−		
md bahnhof	0	0	0	0	0	0	−		
geb anz	+	+	+	+	+	+	0		
whg anz	−	+	0	+	−	0	0		
anteil geb handel	0	+	+	+	+	+	+		
anteil geb kultur gesundheit etc	−	0	+	0	0	0	+		
anteil sb eu15	0	−	+	0	0	0	0		
anteil sb osterweiterung	0	0	0	+	0	0	0		
anteil sb afrika	0	0	0	0	+	0	+		
anteil no hws sbeu15	0	0	0	+	−	0	0		
anteil a65 xx	−	+	0	−	0	0	+		
anteil besch dienstl	+	+	+	+	+	+	+		
anteil sbesch	−	−	−	−	−	−	−		
anteil a00 14	0	0	0	0	−	0	0		
anteil a15 18	0	0	0	0	0	+	0		
anteil a19 44	+	+	0	+	+	0	0		
anteil sb balkan hws	−	0	+	+	0	+	+		
anteil sb balkan nohws	0	0	+	−	+	0	0		
R^2	0.89	0.93	0.98	0.97	0.94	0.71	0.85		

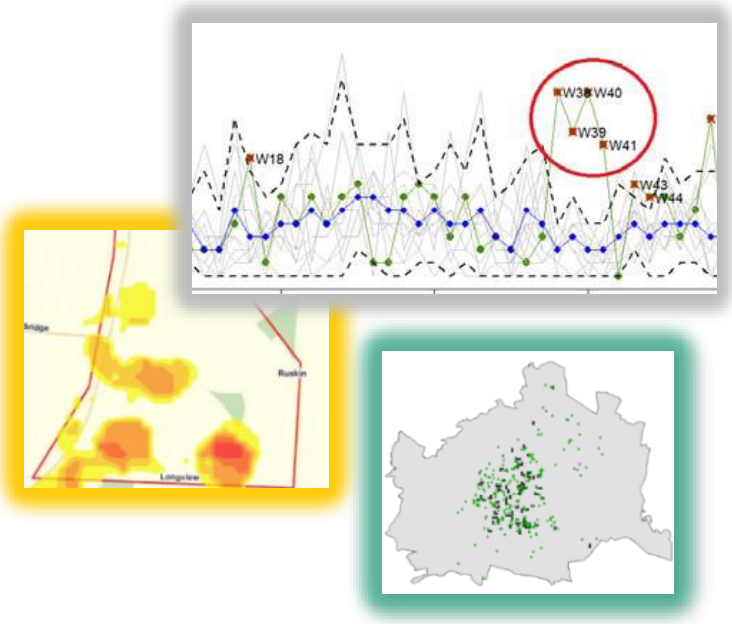
Crime Predictive Analytics (CriPA) Eckdaten

- Kooperatives Forschungsprojekt
- Projektpartner
  - Forschungspartner
    - JOANNEUM RESEARCH (POLICIES, DIGITAL)
    - Universität Salzburg, Z_GIS
    - Institut für Rechts- und Kriminalsoziologie (GSK-Partner)
  - Wirtschaftspartner
    - SynerGIS Informationssysteme
  - Bedarfsträger
    - Bundeskriminalamt
- Laufzeit: Oktober 2013 – Juli 2015

Zielsetzung

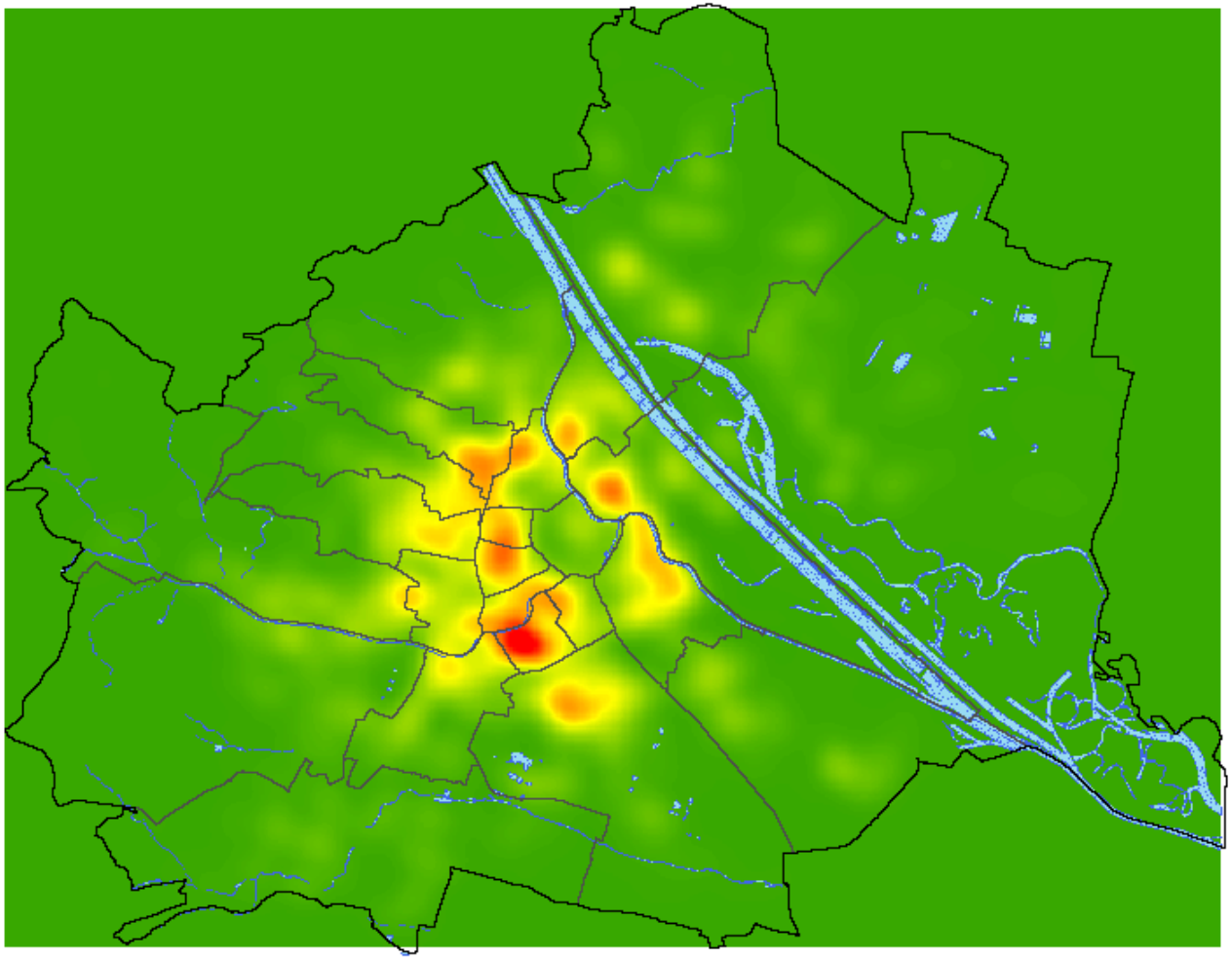
Erforschung quantitativer Methoden um Muster und Zusammenhänge in Kriminalitätsdaten zu identifizieren und Prognosen durchzuführen

- Modelle für mittel-/längerfristige, großräumige Prognosen um zukünftige Trends in der Kriminalität zu schätzen
  - Regressionsmodelle mit erklärenden Größen (Demografie, Landnutzung, Infrastruktur, Events, Wetter, …)
- Methoden für kurzfristige, kleinräumige Prognosen und Risikoabschätzungen
  - Raum-Zeitbezogene Regressionsmodelle
  - Räumliche Analysen (Hot-Spot-Analyse, Risk Terrain Modeling (RTM), Near Repeat)
- Verfahren zur Verbesserung der Prognose, Analyse von unstrukturierten Wissen
  - Text Mining

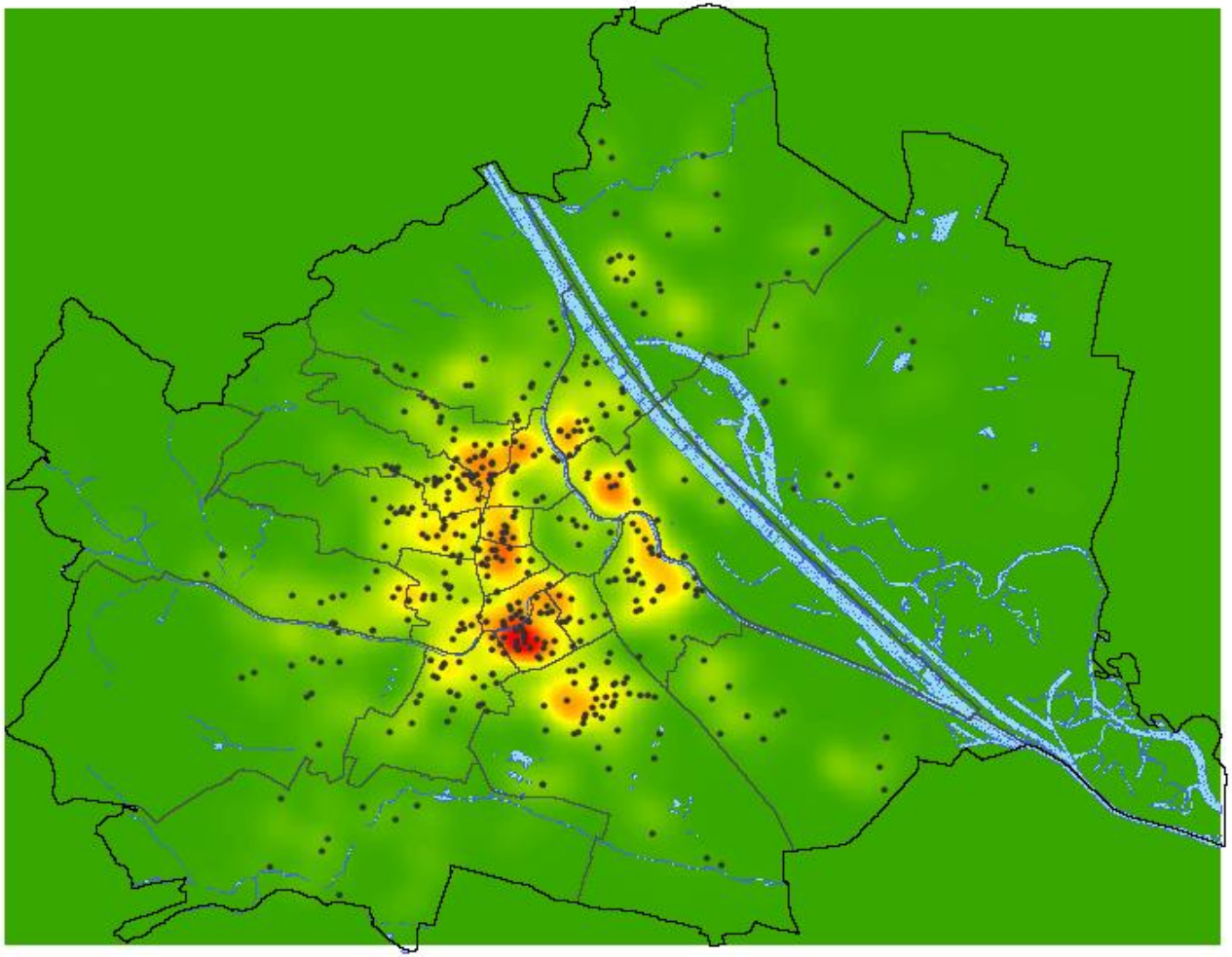


→ **GIS-basiertes Demonstrationssystem**

Prognose Wohnungseinbruch Wien: April 2013

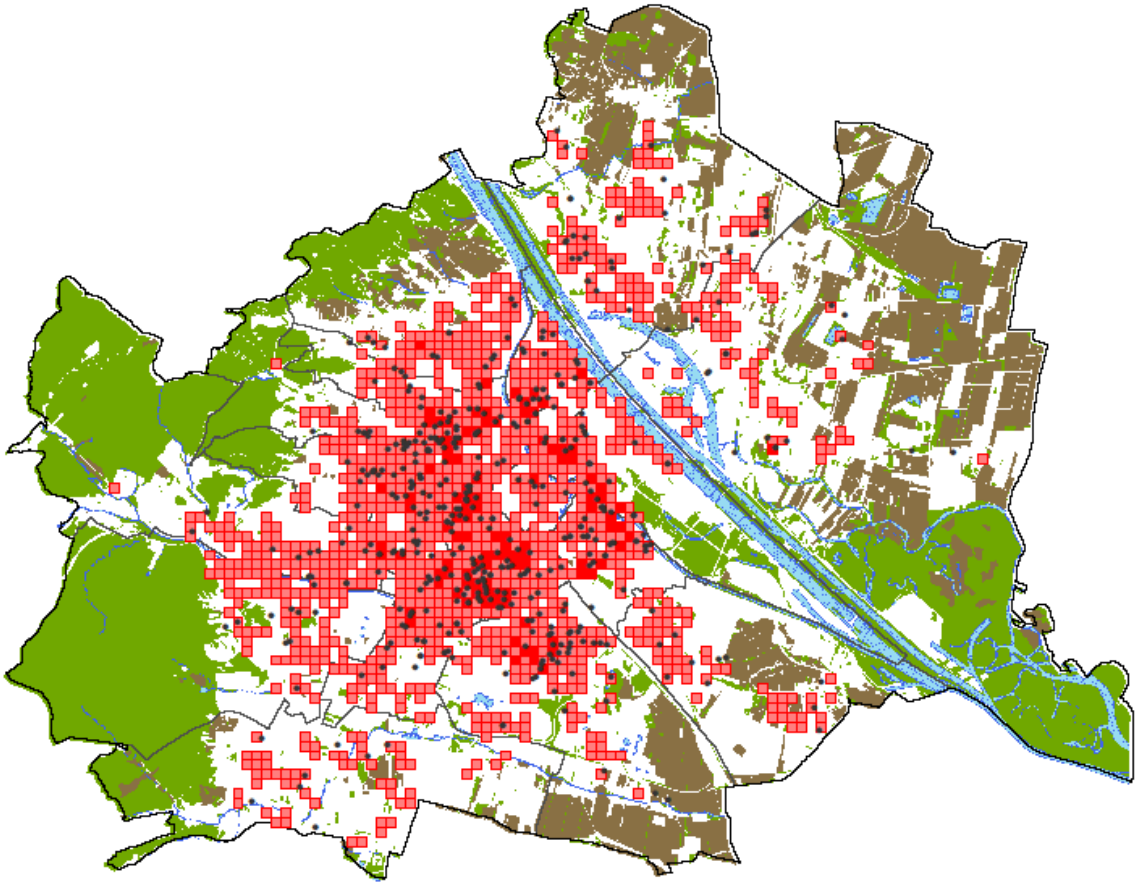


Prognose Wohnungseinbruch Wien: April 2013



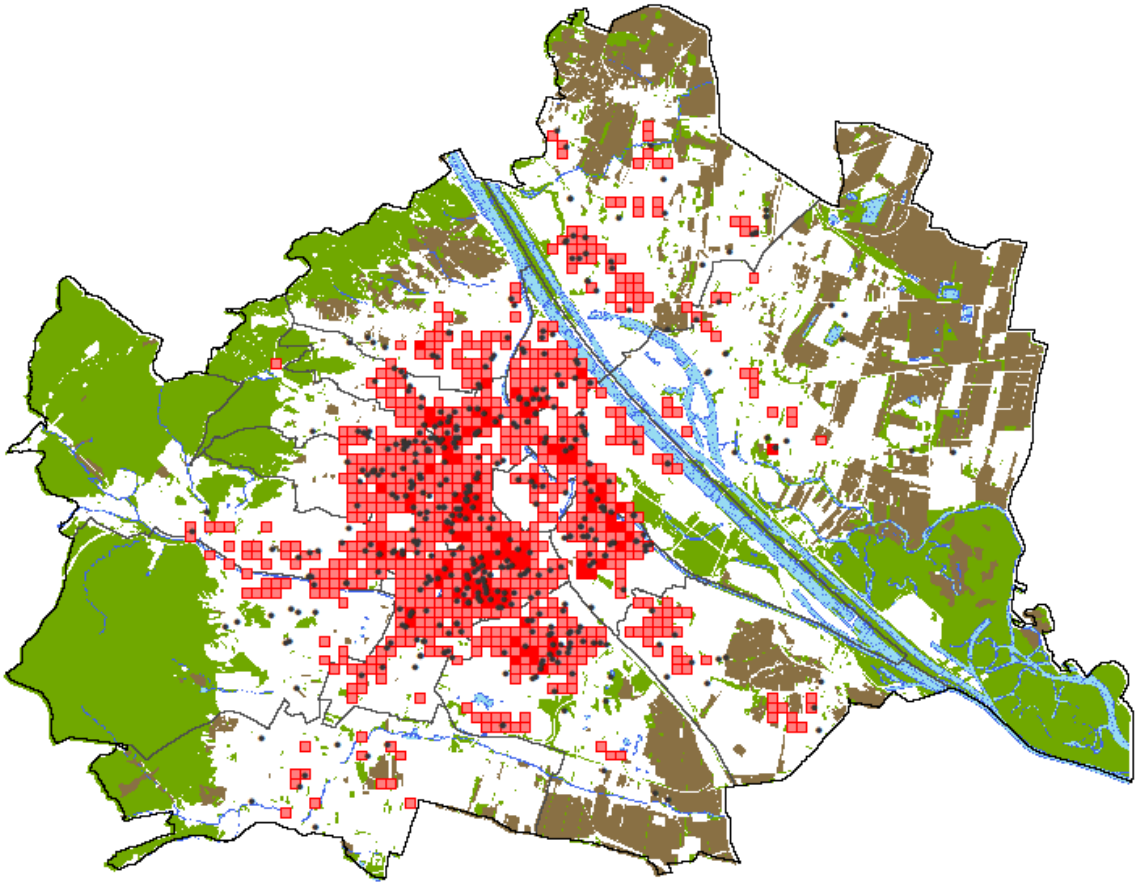
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 1: Risk area = 80.6 km² (Hit-Rate = 89.4%)



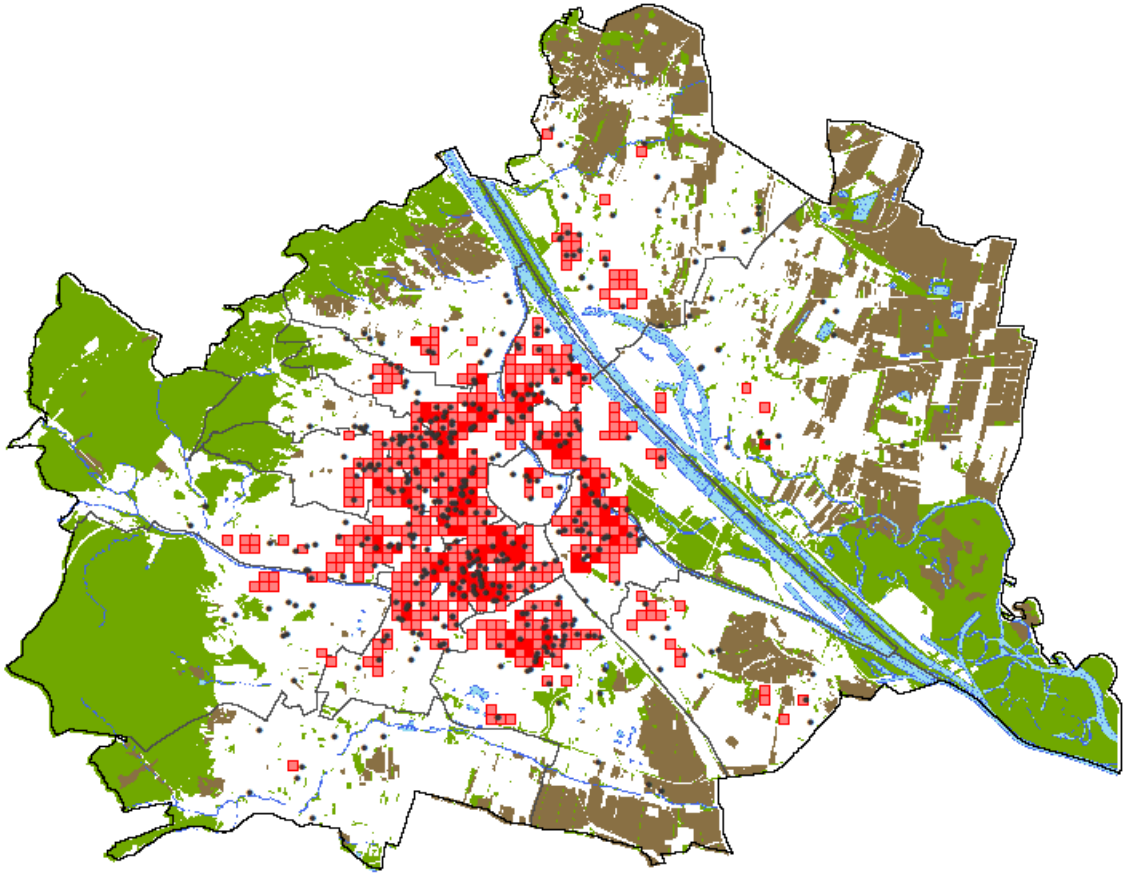
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 2: Risk area = 53.2 km² (Hit-Rate = 77.2%)



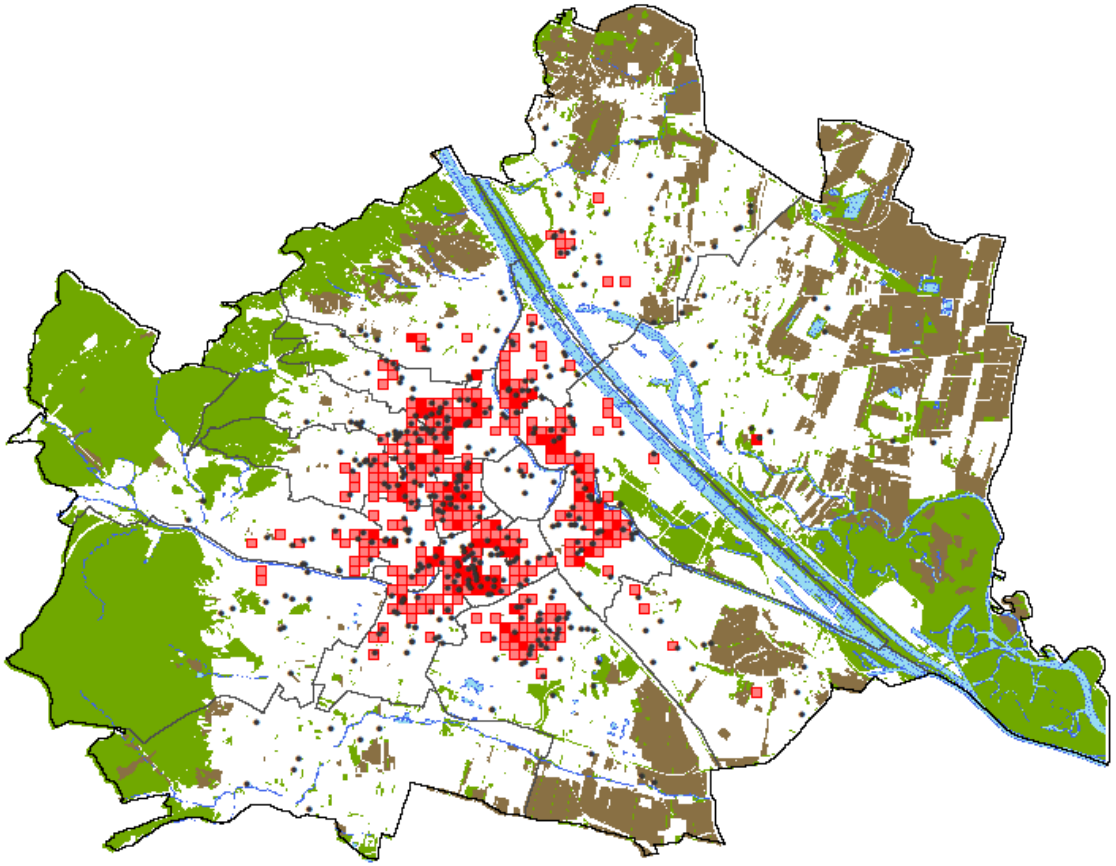
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 3: Risk area = 35.1 km² (Hit-Rate = 63.8%)



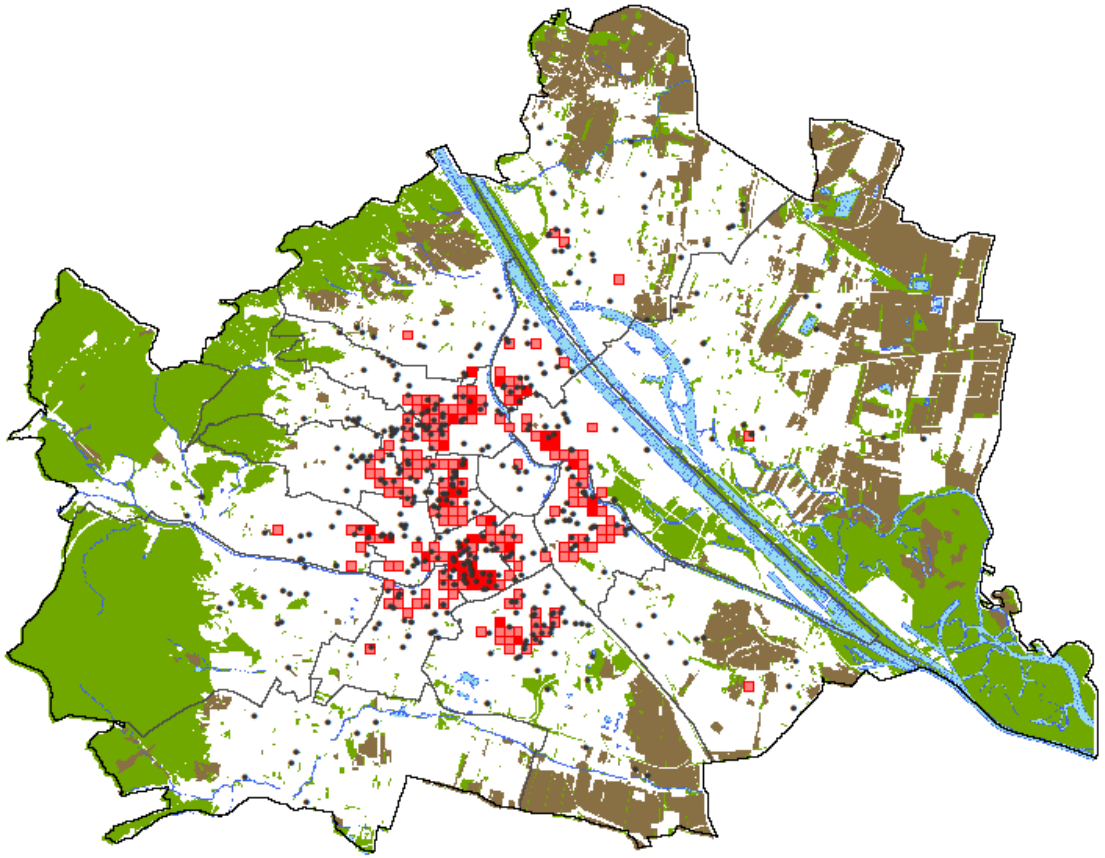
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 4: Risk area = 22.0 km² (Hit-Rate = 48.2%)



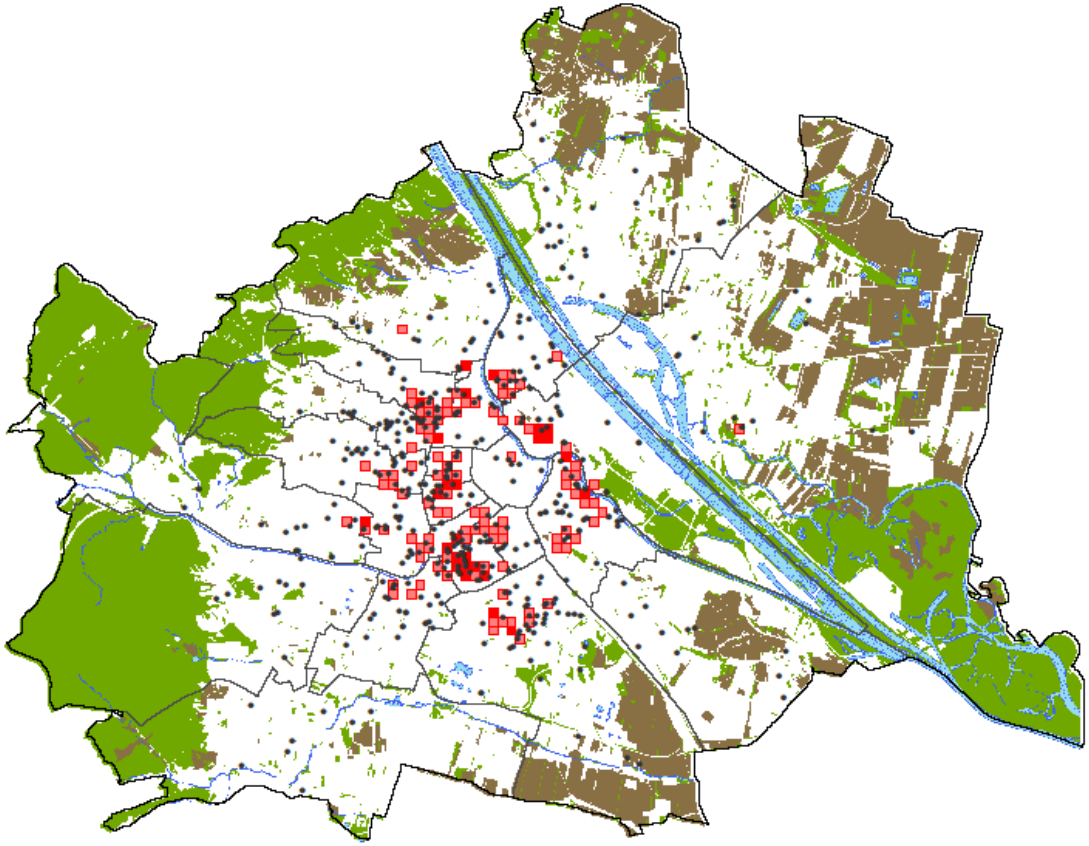
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 5: Risk area = 13.5 km² (Hit-Rate = 32.5%)



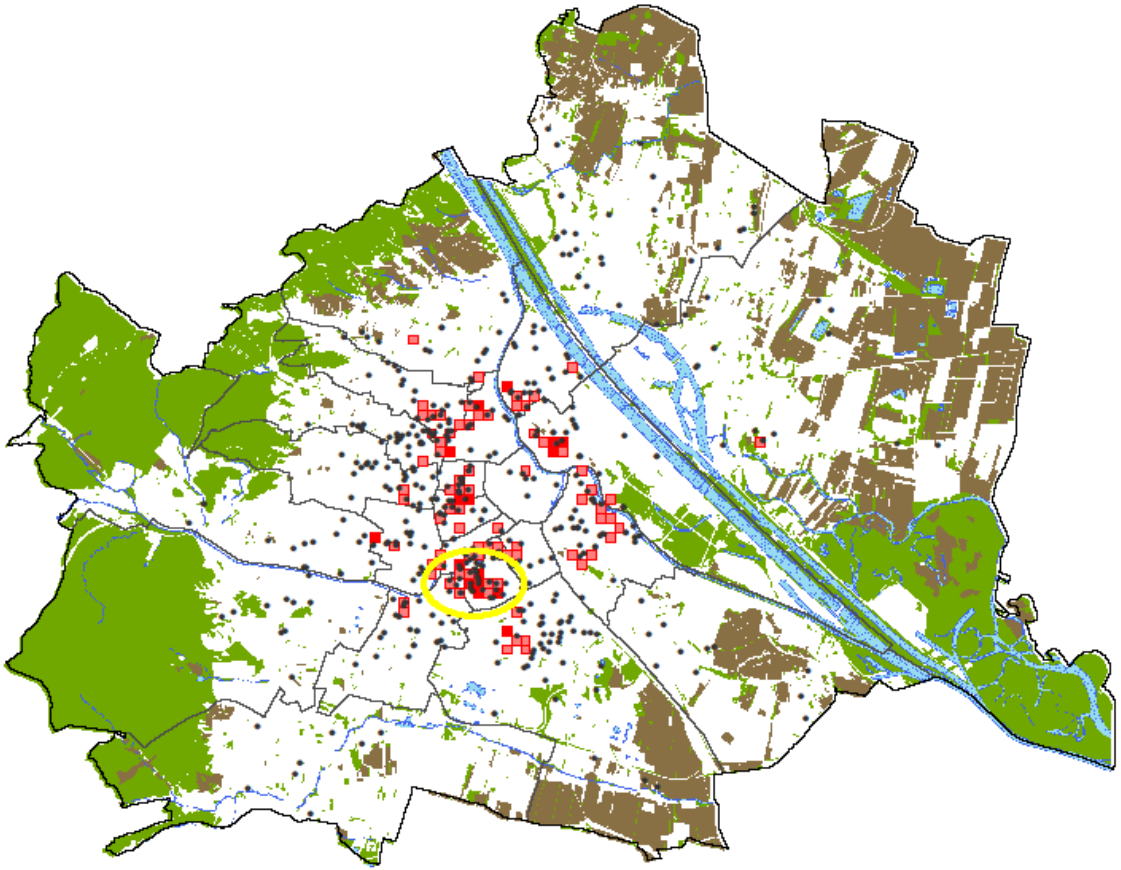
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 6: Risk area = 8,2 km² (Hit-Rate = 21.7%)



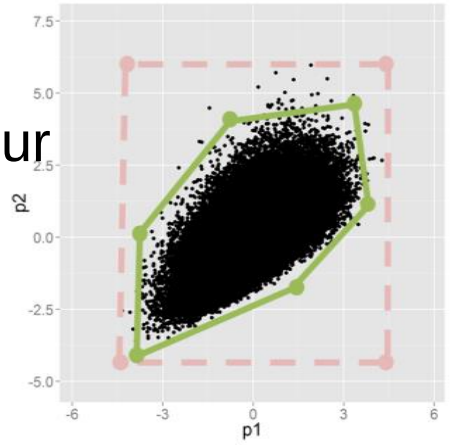
Prognose Wohnungseinbruch Wien: April 2013

Risk Score > 7: Risk area = 5,8 km² (Hit-Rate = 18.4%)



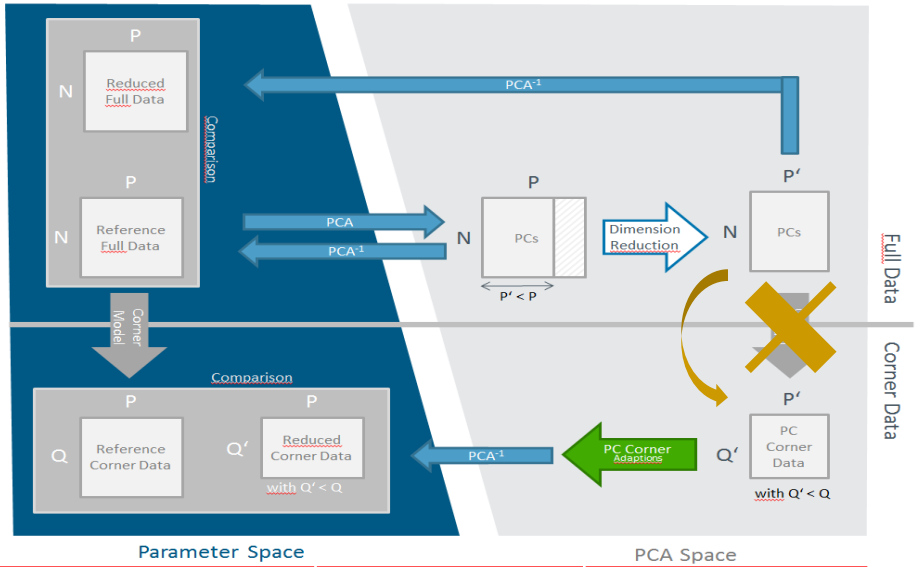
Modellierung der Prozessvariabilität

- Modellierung und Prognose der Prozessvariabilität
  - Chip-Design ist von der Prozessvariabilität abhängig
    - Parameter der Bauteile werden beeinflusst von Schwankungen bei der Herstellung
  - Multivariates Modell dient zur Beschreibung der Variabilität aller Parameter
  - Ziel: optimale Modellierung der Variabilität
  - Subziel: nachvollziehbare Vorgangsweise zur Dimensionsreduktion



Resultat

- Vorgangsweise
  - Signifikante Verbesserung zur ursprünglichen Methode
  - Evaluierung mit realen Produktionsdaten
- Ergebnisse der neuen Methode
  - Dramatische Beschleunigung der Simulation
  - Fähigkeit mit großen Datensätzen zu arbeiten
  - Trotz Dimensions-reduktion bessere Resultate



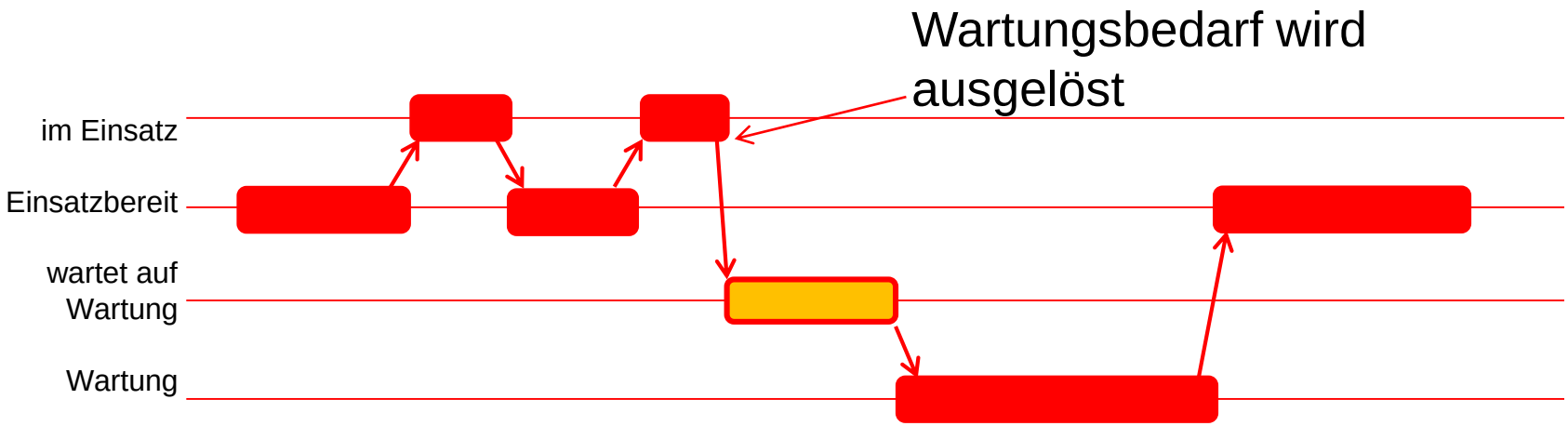
	Original corner model	Principal Corner Model (PCM)
Device	NMOS	NMOS
Number of parameters	6	6
Number of corner points	100%	83%
Confidence	99.57%	97.97%
Univariate confidence	100%	100%
Bivariate confidence	100%	100%
Volume	50242	45059
Correlation deviation	0.09	0.07
Inner limit deviation	0.14	0.00
Outer limit deviation	0.00	0.00
Total computation time	~6500	1.57
Corner computation time (sec)	/	0.19
PCA computation time (sec)	/	1.38
Mahalanobis distance – mean	191.17	182.44
Mahalanobis distance - median	90.31	157.28
Mahalanobis distance – standard deviation	201.96	78.34
Mahalanobis distance – minimum	46.64	111.36
Mahalanobis distance – maximum	648.83	326.59
Mahalanobis distance – range	602.19	215.23

Motivation

- Anlagen müssen regelmäßig gewartet werden
  - Nutzungsbereitschaft zu gewährleisten
  - Lebensdauer zu erhöhen
  - Ungeplante Ausfälle zu minimieren
  - Sicherheitsgründen

- Wartungsmaßnahmen haben unterschiedliche Ursachen
  - Geplante Wartung nach Zeitablauf (kalendarisch)
  - Geplante Wartung nach Nutzungsdauer
  - Ungeplante Wartung wegen Schädigung

Zustandsübergänge eines Objekts



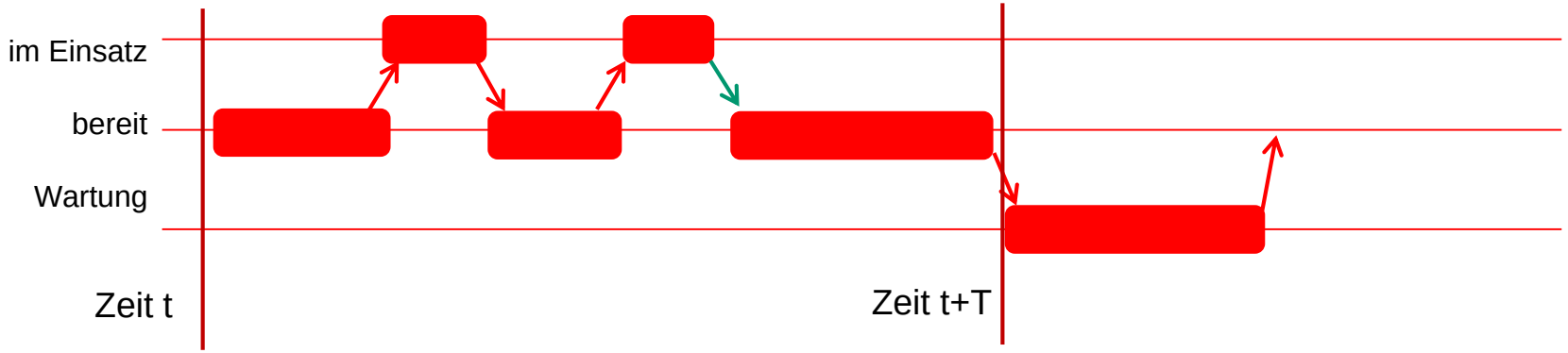
- Wartungsarten
  1. Wartung wegen Zeitablauf (z. B. alle 12 Monate)
  2. Wartung wegen Nutzungsdauer (z. B. nach 2000 Betriebsstd.)
  3. Wartung wegen ungeplantem Ereignis (z. B. Defekt)

Wartungs- und Instandhaltungssystematik

- Systematik soll Wartungsmaßnahmen planen und koordinieren
  - Gesamtziele:
    - Kostenersparnis
    - Erhöhte Verfügbarkeit
- Detailziel Objektbezogen: Wartungszeitpunkte optimieren
- Detailziel Systembezogen: Zahl der einsatzfähigen Objekte maximieren
  - Mathematische Optimierung

Objektbezogene Wartung /1

Beispiel:
Starre Wartungsintervalle: Wartung immer nach T Zeiteinheiten

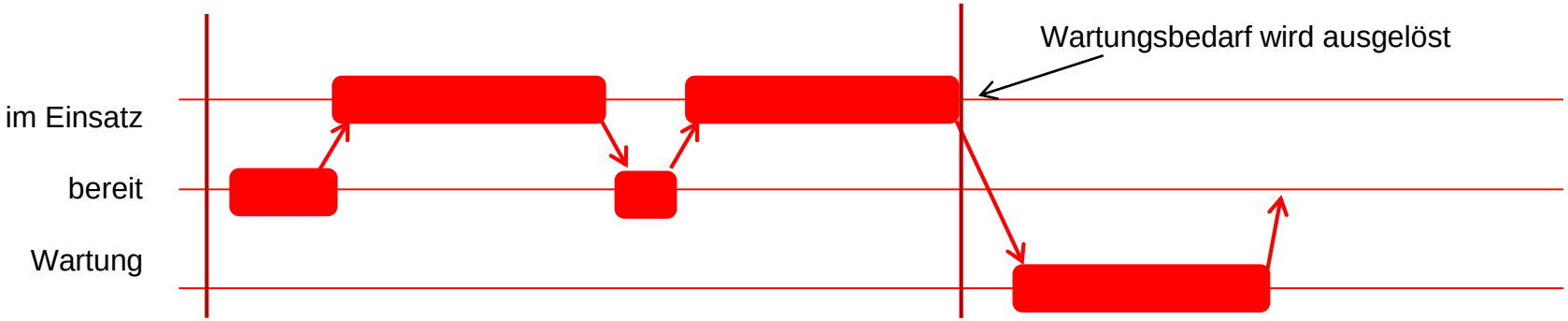


Nachteil: Wartung findet statt, auch bei niedriger Nutzung

- → Zu frühe Wartung
- → Zu hohe Kosten

Objektbezogene Wartung /2

Beispiel:
Anlassbezogene Wartung: Wartung bei Auftreten von Verschleiß, Störung

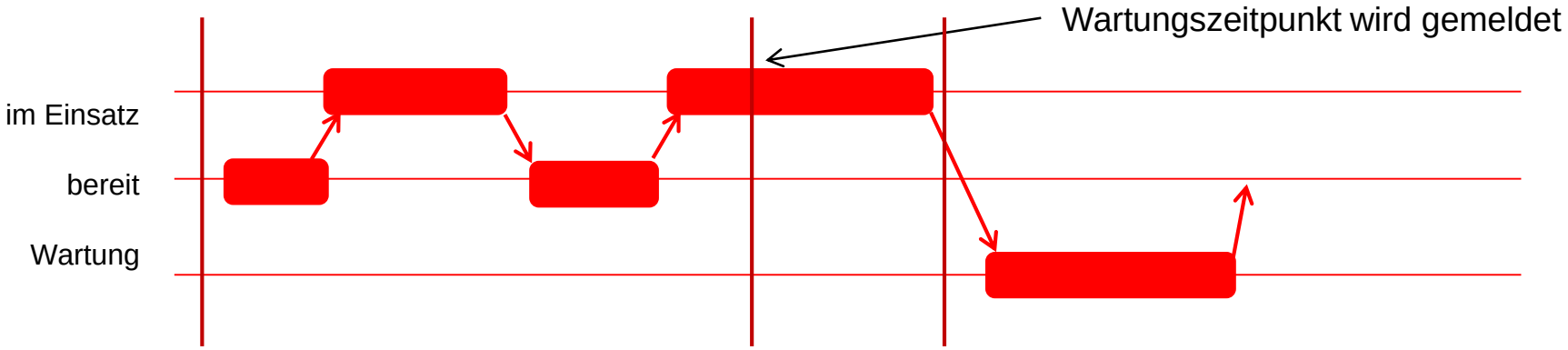


Nachteil: Wartung findet erst bei akutem Bedarf statt, erhöhtes Ausfallsrisiko, ungeplanter Eintritt des Wartungszeitpunkts

- → Kosten des Ausfallsrisikos
- → Unregelmäßige Auslastung der Wartungseinrichtungen, Stehzeiten

Objektbezogene Wartung /3

Beispiel:
Evidenzbasierte Wartungsplanung: Berücksichtigung von Zeitablauf, Nutzungsdauer und Verschleiß in einem statistischem Modell
→ Prognose des „optimalen“ Wartungszeitpunktes



Vorteil: keine unnötige frühe Wartung, dennoch planbare Wartungszeitpunkte bei geringem Ausfallsrisiko

→ Niedrigere Ausfallskosten

→ Niedrigere Wartungskosten

Systembezogene Wartung

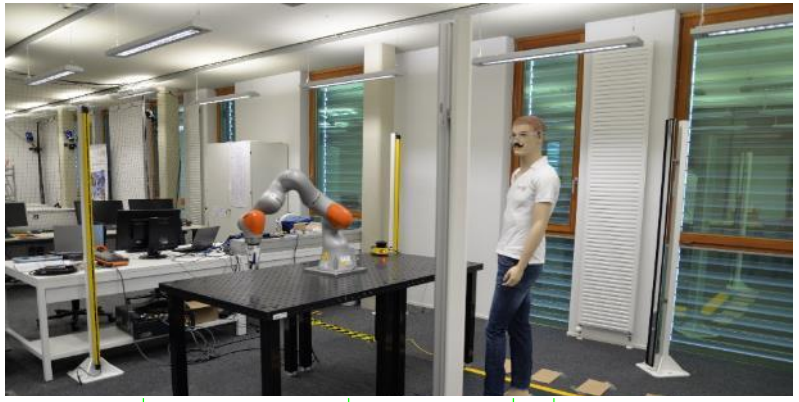
- Kritische Systemparameter / kritische Infrastruktureinrichtungen:
  - Anzahl einsatzbereiter Objekte
    - Wird maximiert durch die Reduktion von
      1. Zu frühen Wartungen
      2. Stehzeiten vor der Wartung („Nicht auf die Wartung warten!“)
- Gleichmäßige Auslastung der Wartungskapazitäten
- Vermeidung von:
  - Überlastung der Wartungseinrichtung (Zusatzkosten)
  - Zu geringe Auslastung der Wartungseinrichtung (Fixkosten)
  - Zukauf von Wartungskapazitäten

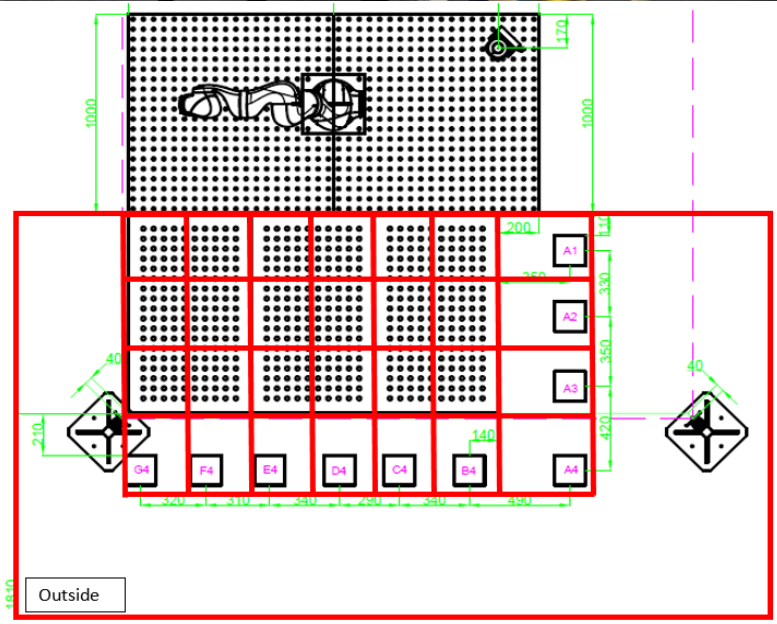
Methodischer Ansatz

- Mathematische Optimierung der Wartungstätigkeiten
  - Entwicklung eines mathematischen Modells in Form eines Integer Linear Programming (ILP) Modells
  - Validierung und Test des ILP-Modells mit unterschiedlichen Solvern
- Ausgabe der Optimierung
  - Wartungsplan für jedes Objekt
    - Wartungsbeginn und Wartungsdauer (mit Toleranzen)
    - Art der Wartung
    - Geplante Wartungsstätte
- Einsatzplan für jedes Objekt
  - Empfohlene Zeitdauer bis zur nächsten Wartung
  - Empfohlene Nutzungsdauer innerhalb dieser Zeit
- Ergebnis der Optimierung
  - Auslastung der Wartungsstätten wird optimiert
  - Nutzungsbereitschaft aller Objekte wird maximiert
  - Stehzeiten der Objekte werden minimiert

Fallbeispiel: Statistische Modellierung vs. Machine Learning - Problemstellung

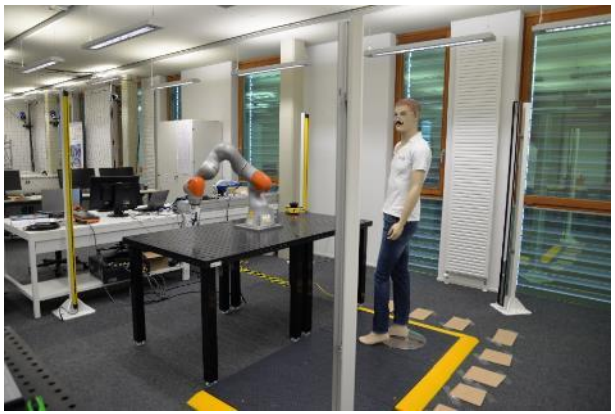
- Verwendete Sicherheitseinrichtungen
  - Laserscanner von Omron
  - Sicherheitslichtvorhang von Keyence
  - Sicherheitsschaltmatte von Pilz
- Ziel: Positionsvorhersage des Menschen
  - Klassifizierungsproblem
- Unterteilung des Arbeitsraumes in 28 Positionen + eine zusätzliche „Position“
  - Zusätzliche Position entspricht Bereich außerhalb des Arbeitsraumes
- Aufzeichnung von Trainingsdaten
- Aufzeichnung von Testdaten
  - 6 unterschiedliche Wege

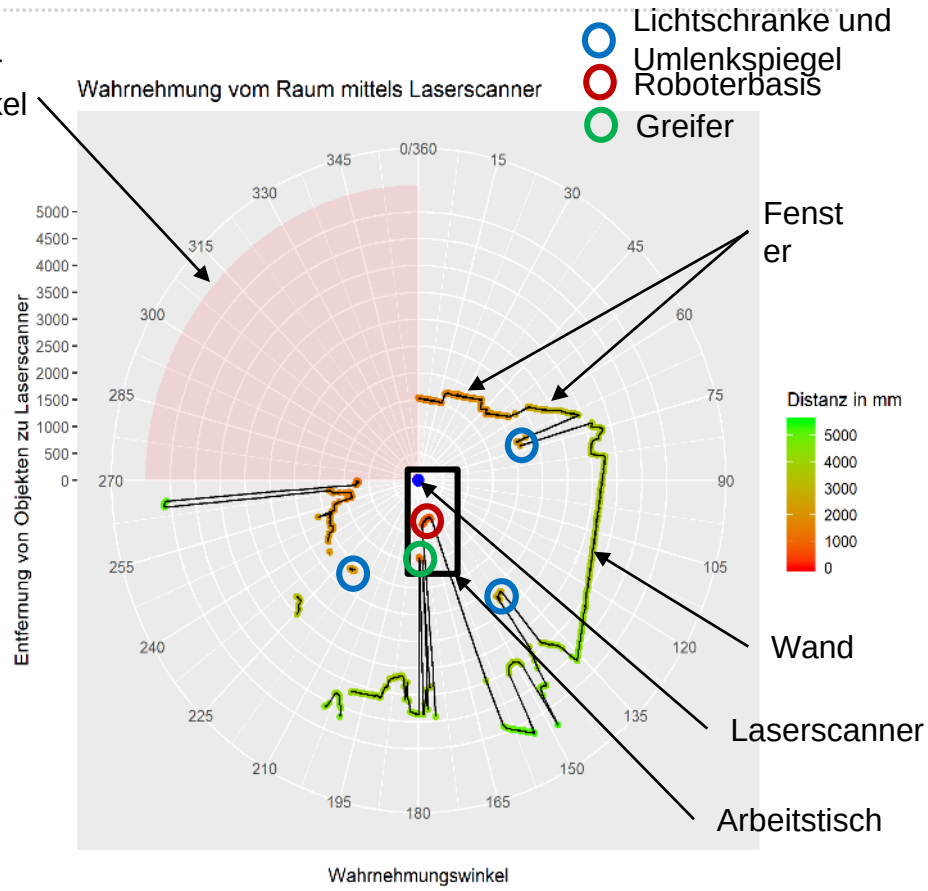




Fallbeispiel: Statistische Modellierung vs. Machine Learning - Datenstruktur I

- Laserscanner von Omron
  - Scannt nur Ebene im Raum
  - Aufzeichnung erfolgt auf 270° eines Kreises
  - Verwendetes Modell sendet alle 0.4° einen Laserstrahl → 677 Laserstrahlen
  - Distanz in mm
  - Taktzeit ca. 70 ms





Fallbeispiel: Statistische Modellierung vs. Machine Learning - Datenstruktur I

- Sicherheitsschaltmatte von Pilz
  - Binäre Variable mit räumlicher Information
  - 1000 x 600 mm
    - Ca. 12,5 x 12 cm
  - Taktzeit: ca. 100 ms

date	time	ms		start mat 0		
17.01.2019	14:53:47	16		0	0	0
17.01.2019	14:53:47	118		0	0	0
17.01.2019	14:53:47	223		0	0	0
17.01.2019	14:53:47	324		0	0	0
17.01.2019	14:53:47	426		0	0	0
17.01.2019	14:53:47	528		0	0	0
17.01.2019	14:53:47	631		0	0	0
17.01.2019	14:53:47	733		0	0	0
17.01.2019	14:53:47	835		0	0	0

- Lichtschranke
  - Binäre Variable ohne räumliche Information
    - Keine Information wo Lichtgitter unterbrochen wurde
  - Taktzeit: ca. 100 ms
- Für Modellierung wird ein gemeinsamer Datensatz benötigt
  - Fusion der einzelnen Sensorinformationen
  - Probleme wegen unterschiedlicher Taktzeiten

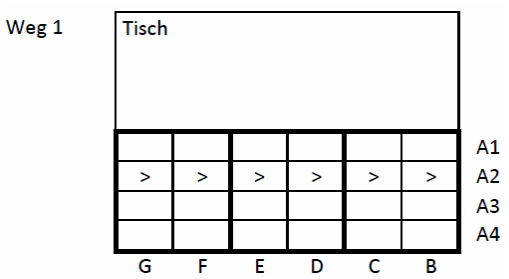
date	time	ms		State of SafetyLightCurtain
17.01.2019	14:53:47	555		0
17.01.2019	14:53:47	656		0
17.01.2019	14:53:47	758		0
17.01.2019	14:53:47	859		0
17.01.2019	14:53:47	961		0
17.01.2019	14:53:48	64		0
17.01.2019	14:53:48	165		0

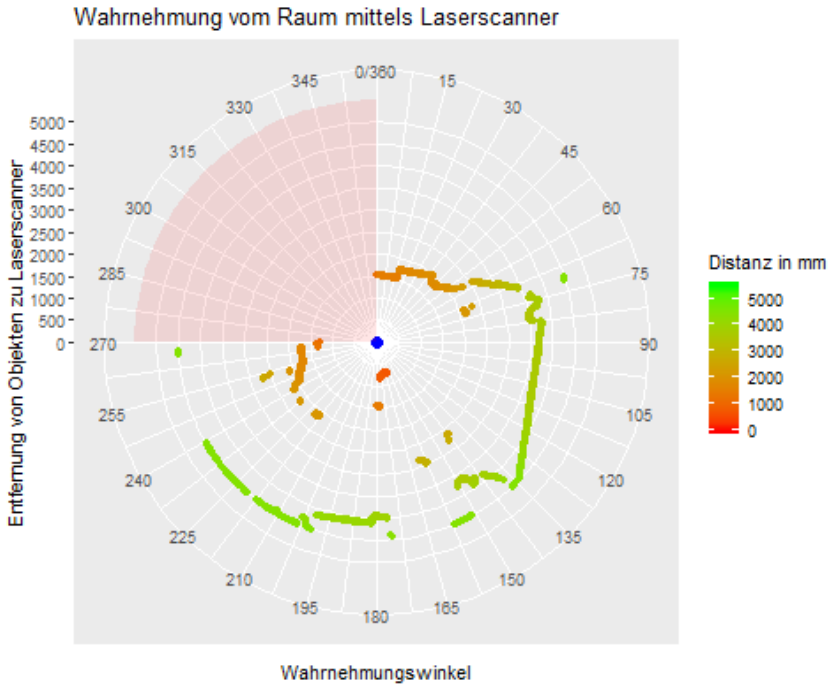
Fallbeispiel: Statistische Modellierung vs. Machine Learning - Methodenübersicht

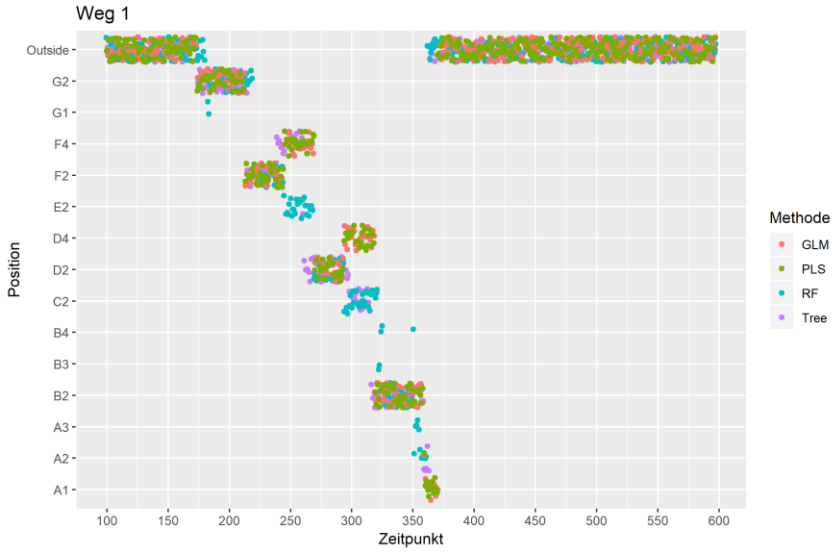
- Statistische Modelle
  - Diskriminanzanalyse (linear und quadratisch)
  - Generalisiertes lineares Modell (GLM)
  - Partial Least Square Modell (PLS)
  - Entscheidungsbaum
- Machine Learning Algorithmen
  - Support Vector Machine
  - Random Forest
- Modellierung unterschiedlichster Kombinationen der Sicherheitseinrichtungen
- Beste Positionsvorhersagen wurden mittels Machine Learning Algorithmus Random Forest erzielt

Fallbeispiel: Statistische Modellierung vs. Machine Learning - Ergebnis

- Ziel: Positionsvorhersage für aufgezeichnete Wege

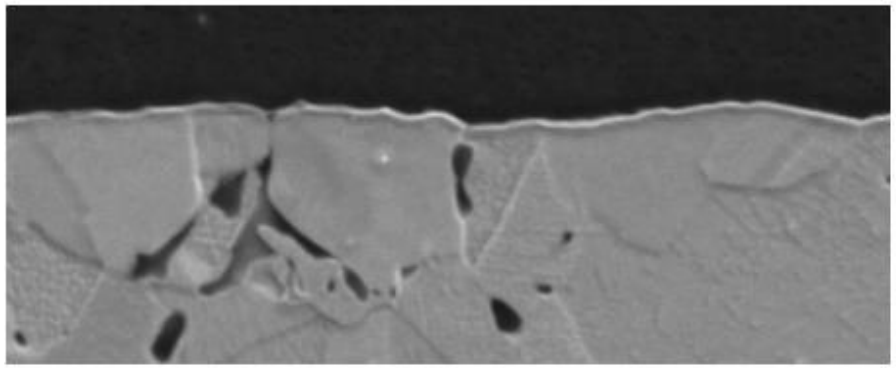






Machine Learning im Bereich von Image Processing - Ausgangssituation

- Zuverlässigkeit von elektronischen Bauteilen in Automobilindustrie von besonderer Bedeutung
- Durchführung von Lebensdauertests
- Durch die thermomechanische Belastung dieser Tests entsteht ein Abbau der Metallschicht
  - Rissbildung bzw. Löcher entstehen
- Visualisiert wird dieser Abbau mit Hilfe eines Rasterelektronenmikroskops
  - Graustufenbild

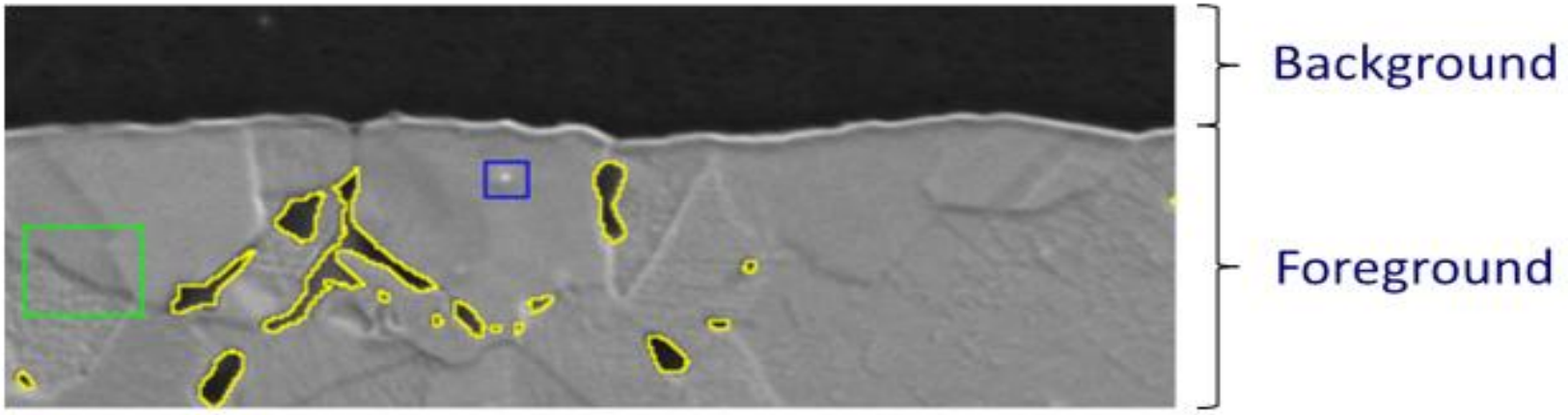


Machine Learning im Bereich von Image Processing - Zielsetzung

- Entwicklung eines Algorithmus welcher Risse bzw. Löcher in einer Metallschicht mit Hilfe von Machine Learning Methoden automatisch findet
- Bei der Problemstellung handelt es sich um Klassifikationsproblem
- Getestet wurden zwei unterschiedliche Machine Learning Ansätze
  - Random Forest
  - Support Vector Machine
- Algorithmus muss zwischen Metallschicht (Foreground) und Nicht-Metallschicht (Background) unterscheiden
- Algorithmus darf nur Risse/Löcher in Metallschicht finden

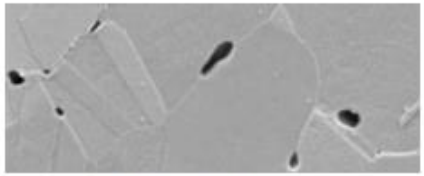
Machine Learning im Bereich von Image Processing - Beispiel

- Gelb markierte Stellen sind Risse/Löcher
- Blau markierte Stelle ist ein Bearbeitungsartefakt
- Grün markierte Stelle wird als Korngrenze bezeichnet (entsteht durch Lebensdauertest + Materialeigenschaft)

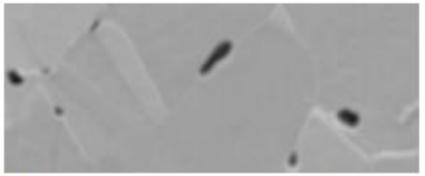


Machine Learning im Bereich Image Processing - Vorgangsweise

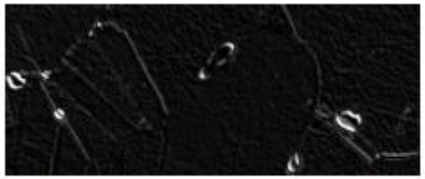
- Aufgrund von wenig Information müssen neue „Features“ bestimmt werden
  - Vorhandene Information sind Pixelkoordinaten und zugehöriger Grauwert
- Berechnung der Features durch Anwendung verschiedener Filtermethoden
  - Kantenfilter (Edge-Detection-Methods)
  - Rauschunterdrückungsverfahren (Noise-Reduction-Methods)
  - Strukturmethoden (Texture-Methods)



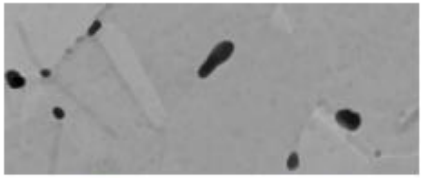
Original Image



Noise reduction method



Edge detection method

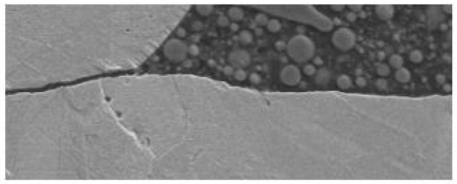


Texture method

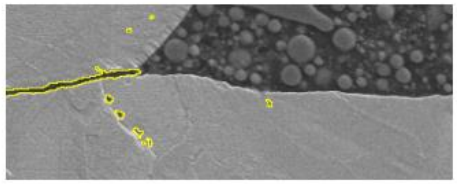
- Umgekehrte Vorgehensweise → von „Small Data“ zu „Big Data“
  - Graustufenwert + 182 zusätzliche „Features“ (Fore-/Background)
  - Graustufenwert + 122 zusätzliche „Features“ (Damage-/No-Damage)

Machine Learning im Bereich von Image Processing - Resultat

- Zwei unterschiedliche Klassifizierer wurden gebildet
  - Ein Klassifizierer unterscheidet zwischen Metallschicht und Nicht-Metallschicht
  - Ein Klassifizierer unterscheidet zwischen Schädigung und Nicht-Schädigung
- Resultate sehr zufriedenstellend
  - Algorithmus markierte Schädigung nur in Metallschicht
  - Unterscheidung zwischen Korngrenze und Risse oftmals nicht eindeutig
  - Qualität vom Output hängt stark von Qualität der Bilder ab!



Input Image



Output Image

Zusammenfassung

- Große Datenmengen sind von sich aus kein Erfolgsgarant
- Systematische Vorgangsweise bei datengestützten Fragestellungen ist von enormer Wichtigkeit
- Daten ≠ Informationen
- Datenanalyse führt meist zur Datenreduktion
  - Einsatz von statistischen Methoden zwingend notwendig
  - Führt meist von Big Data zu Smart Data
  - Basis für Verbesserungen
- Datenanalyse kann in folgenden Bereichen eingesetzt werden
  - Identifikation von Zusammenhängen – statistische Modellierung
  - Monitoring
  - Prognose
- „Mache die Dinge so einfach wie möglich – aber nicht einfacher“. (Albert Einstein)

Resümee und Ausblick

- Datenanalyse sollte integraler Bestandteil im unternehmerischen Umfeld sein
- Sammeln und auswerten der „richtigen“ Daten wird zum Erfolgsfaktor
- Wissen aus den Daten wird Differenzierungsmerkmal am Markt

Angewandte Statistik ist eine Zusammenfassung von Methoden, die uns erlauben, vernünftige Entscheidungen im Falle von Ungewissheit zu treffen.

(Abraham Wald)

Quellen

- Big Data – Fluch oder Segen? (2014)
  Bachmann R., Kemper G. und Gerzer T. – mitp-Verlags GmbH
- Smart Data – Datenstrategien, die Kunden wirklich wollen und Unternehmen wirklich nützen (2015)
  Bloching B., Luck L. und Ramge T. – Redline Verlag
- Big Data – die Revolution, die unser Leben verändern wird (2017)
  Mayer-Schönberger V. und Cukier K. – Redline Verlag

Danke für Ihre Aufmerksamkeit!

DI Hermann Katz

JOANNEUM RESEARCH FORSCHUNGSGESELLSCHAFT MBH

POLICIES – Institut für Wirtschafts- und Innovationsforschung
Datenanalyse und statistische Modellierung

Leonhardstraße 59
8010 Graz

Tel.: +43 316 876-1553
Mobil: +43 664 602 876 1553
PC-Fax: +43 316 8769-1553

Email: hermann.katz@joanneum.at

DIHSÜD

Der DIH SÜD wird unterstützt von:

FFG Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN