Citizen Data Science

Balázs Bárány

Linuxwochen Wien 2016

29. April 2016

Inhalt

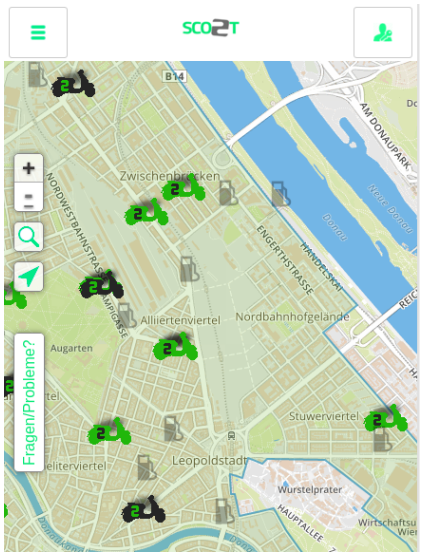
Über mich

Selbständiger Data Scientist – https://datascientist.at

Über mich

Selbständiger Data Scientist – https://datascientist.at

SCO2T – Roller-Sharing in Wien – https://sco2t.com



datascientist.at

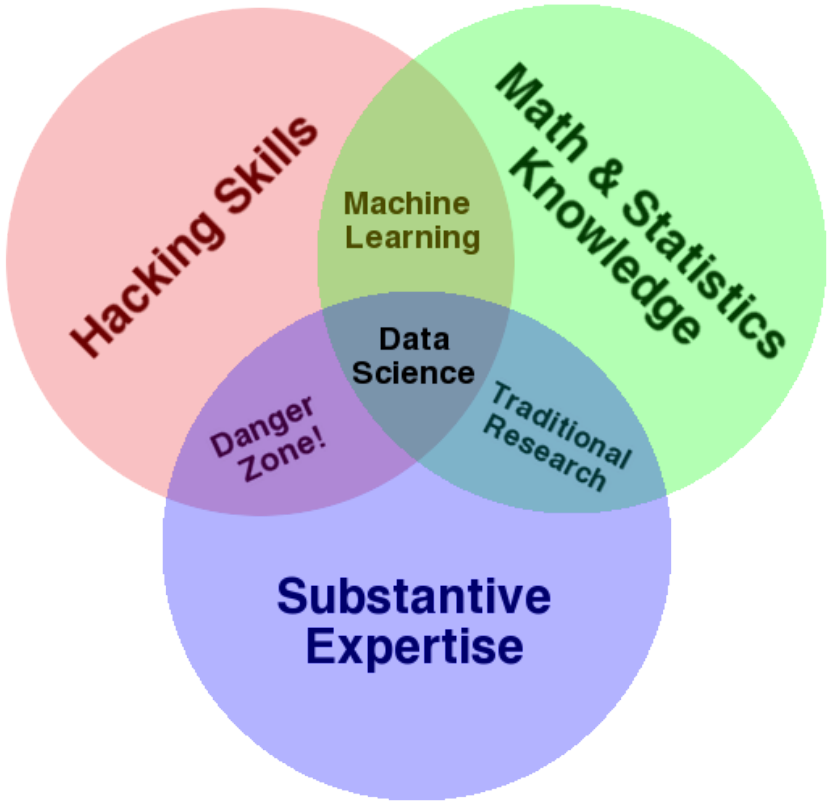
"Sexiest job of the 21st century"

- Sagen Google, LinkedIn, ...

"Sexiest job of the 21st century"

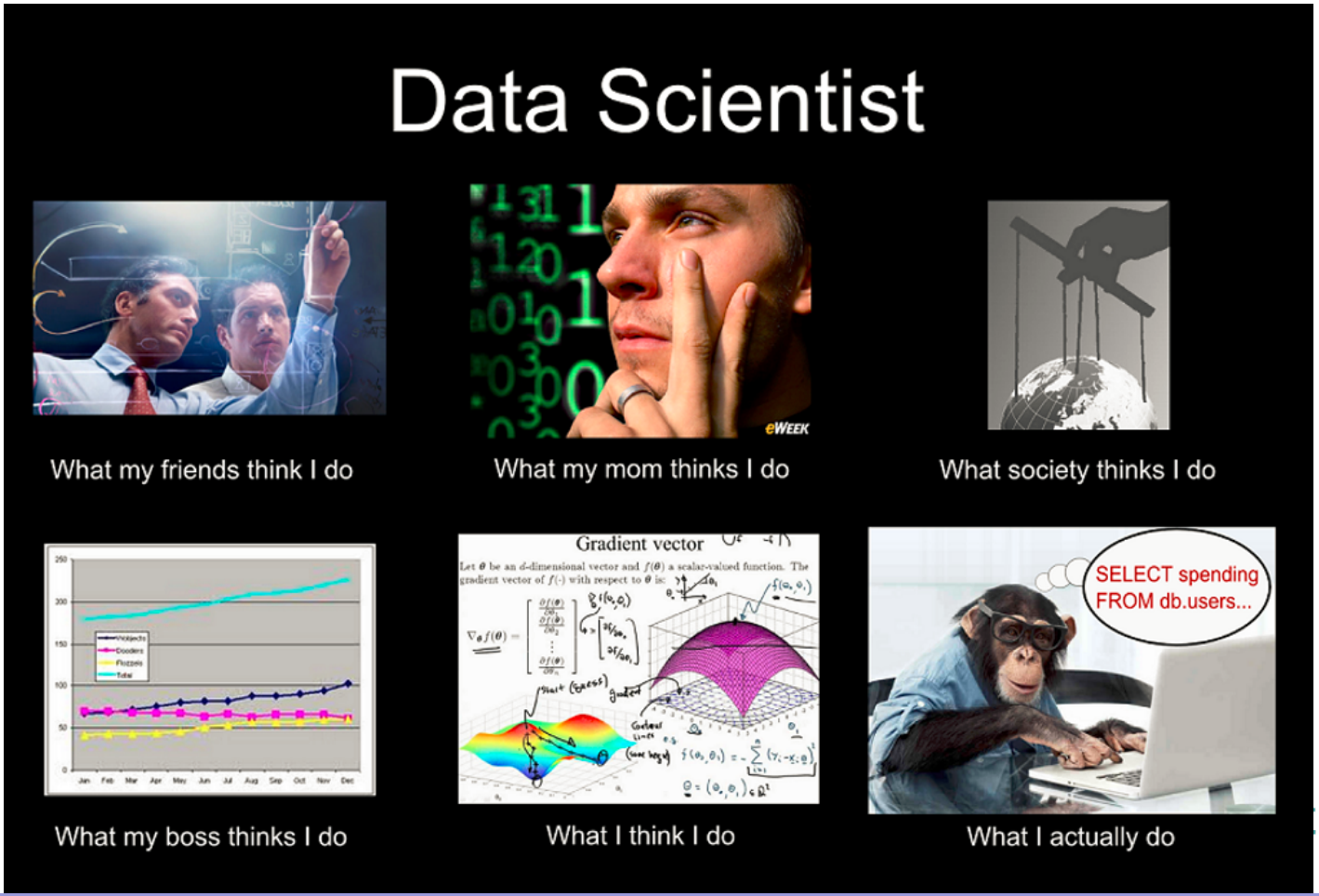
- Sagen Google, LinkedIn, ...
- Wer ist ein Data Scientist?

Data Science Venn Diagram



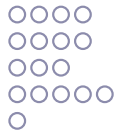
datascientist.at

Was machen Data Scientists?



Aufgaben

- Daten holen und zusammenführen



Aufgaben

- Daten holen und zusammenführen
- Verknüpfen und umformen für Analytik

Aufgaben

- Daten holen und zusammenführen
- Verknüpfen und umformen für Analytik
- Analysieren und visualisieren

Aufgaben

- Daten holen und zusammenführen
- Verknüpfen und umformen für Analytik
- Analysieren und visualisieren
- Vorhersagen und Handlungen empfehlen

Aufgaben

- Daten holen und zusammenführen
- Verknüpfen und umformen für Analytik
- Analysieren und visualisieren
- Vorhersagen und Handlungen empfehlen
- Operationalisieren

Aufgaben

- Daten holen und zusammenführen
- Verknüpfen und umformen für Analytik
- Analysieren und visualisieren
- Vorhersagen und Handlungen empfehlen
- Operationalisieren
- Big Data?

The Data Mining process

Cross Industry Standard Process for Data Mining (Kenneth Jensen/Wikimedia Commons)

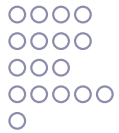
datascientist.at

Fehlende Data Scientists

- Prognosen: 50 % der Stellen nicht besetzbar
- Ausbildung kommt nicht nach

Fehlende Data Scientists

- Prognosen: 50 % der Stellen nicht besetzbar
- Ausbildung kommt nicht nach
- **Citizen Data Scientists**



Werkzeuge und Methoden

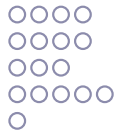
Werkzeuge und Methoden

Scripting und Programmierung

- R
- Python mit Modulen
- Octave/Matlab, andere mathematische Sprachen
- Hadoop, Big Data libraries (Java)
- Cloud services

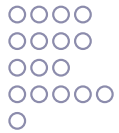
Grafische Werkzeuge

- (teilweise) Open Source: RapidMiner, KNIME, Orange
- Open-Source-Data-Warehouse-Werkzeuge mit Erweiterungen für Analytik: Pentaho, Talend
- Kommerzielle Werkzeuge, z. B. SAS, IBM SPSS
- Hadoop-Newcomer: z. B. Datameer



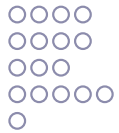
Daten-Infrastruktur

- Datenbanken und Datenspeicher
  - Relational, NoSQL
  - Hadoop-Cluster
  - In-memory
- Datenströme
- Unstrukturiert: Text, Bilder, Video, Audio, ...
- Web APIs
- Open Data



Daten holen und zusammenführen

- Daten im „Rohformat“



Daten holen und zusammenführen

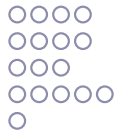
- Daten im „Rohformat“
- Join, Aggregierung, Filterung, Berechnung, ...

Daten holen und zusammenführen

- Daten im „Rohformat“
- Join, Aggregierung, Filterung, Berechnung, ...
- Säuberung
  - Fehlende Werte
  - Ausreißer

Daten holen und zusammenführen

- Daten im „Rohformat“
- Join, Aggregierung, Filterung, Berechnung, ...
- Säuberung
  - Fehlende Werte
  - Ausreißer
- Ergebnis: Für Analytik geeignete Tabelle



Prädiktive Modellierung

- Zielvariable bekannt?
  - Supervised/unsupervised (überwacht/unüberwacht)

Prädiktive Modellierung

- Zielvariable bekannt?
  - Supervised/unsupervised (überwacht/unüberwacht)
- Klassifikation (supervised): Vorhersage einer Kategorie
- Regression (supervised): Vorhersage eines numerischen Wertes

Prädiktive Modellierung

- Zielvariable bekannt?
  - Supervised/unsupervised (überwacht/unüberwacht)
- Klassifikation (supervised): Vorhersage einer Kategorie
- Regression (supervised): Vorhersage eines numerischen Wertes
- Clustering (unsupervised): Automatische Gruppierung
- Assoziationsanalyse, Ausreißererkennung, Zeitreihen-Prognose, ...

Operationalisierung

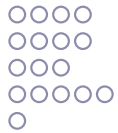
- Anwendung des Modells auf neue Daten ergibt Vorhersage
  - (+ Konfidenz)

Operationalisierung

- Anwendung des Modells auf neue Daten ergibt Vorhersage
  - (+ Konfidenz)
- Im ERP- oder CRM-System speichern
- Aufmerksam machen (E-Mail, Popup)
- Markieren (z. B. E-Mail als Spam)

Operationalisierung

- Anwendung des Modells auf neue Daten ergibt Vorhersage
  - (+ Konfidenz)
- Im ERP- oder CRM-System speichern
- Aufmerksam machen (E-Mail, Popup)
- Markieren (z. B. E-Mail als Spam)
- Transaktion unterbrechen
- Waren nachbestellen
- ...



Citizen Data Science

Data Science für Alle

datascientist.at

Mein Werkzeugkasten

- Datenbank: PostgreSQL
  - Features, Erweiterungen, Ökosystem, ...

Mein Werkzeugkasten

- Datenbank: PostgreSQL
  - Features, Erweiterungen, Ökosystem, ...
- Programmiersprache: R
  - Geschmackssache

Mein Werkzeugkasten

- Datenbank: PostgreSQL
  - Features, Erweiterungen, Ökosystem, ...
- Programmiersprache: R
  - Geschmackssache
- Grafisches Data-Mining-Tool: RapidMiner

Mein Werkzeugkasten

- Datenbank: PostgreSQL
  - Features, Erweiterungen, Ökosystem, ...
- Programmiersprache: R
  - Geschmackssache
- Grafisches Data-Mining-Tool: RapidMiner
- Für Geodaten: QGIS

Daten holen

Datenquellen

Wetterdaten von Weather Underground

- Gratis-API
- Vorhersage, aktuelles Wetter, historische Daten
- JSON- und XML-Format verfügbar

Wetterdaten von Weather Underground

- Gratis-API
- Vorhersage, aktuelles Wetter, historische Daten
- JSON- und XML-Format verfügbar

- Demo mit RapidMiner

Wien: Bezirksgrenzen

- Open Data, in verschiedenen Formaten verfügbar
- Bezirksgrenzen als Polygone; Fläche, Umfang

Wien: Bezirksgrenzen

- Open Data, in verschiedenen Formaten verfügbar
- Bezirksgrenzen als Polygone; Fläche, Umfang

Beispiel

Einlesen in PostgreSQL in einem Befehl:

```
COPY bezirksgrenzen_wien
FROM PROGRAM 'curl -s "http://data.wien.gv.at/daten/geo?..."'
WITH CSV delimiter ',' HEADER;
```

Bezirksgrenzen - Fortsetzung

- Geodaten noch im Textformat
- Umwandlung in echte Geo-Objekte mit PostGIS oder QGIS

Bezirksgrenzen - Fortsetzung

- Geodaten noch im Textformat
- Umwandlung in echte Geo-Objekte mit PostGIS oder QGIS

Beispiel

PostGIS:

```
ALTER TABLE bezirksgrenzen_wien
ADD COLUMN geo geometry;
UPDATE bezirksgrenzen_wien
SET geo = ST_GeomFromText(shape);
```

Bezirksgrenzen - Fortsetzung

- Geodaten noch im Textformat
- Umwandlung in echte Geo-Objekte mit PostGIS oder QGIS

Beispiel

PostGIS:

```
ALTER TABLE bezirksgrenzen_wien
ADD COLUMN geo geometry;
UPDATE bezirksgrenzen_wien
SET geo = ST_GeomFromText(shape);
```

- Demo mit QGIS

Daten verstehen

Data Understanding

Data understanding

- Erster Schritt nach dem Import neuer Daten
- Was ist enthalten?

Data understanding

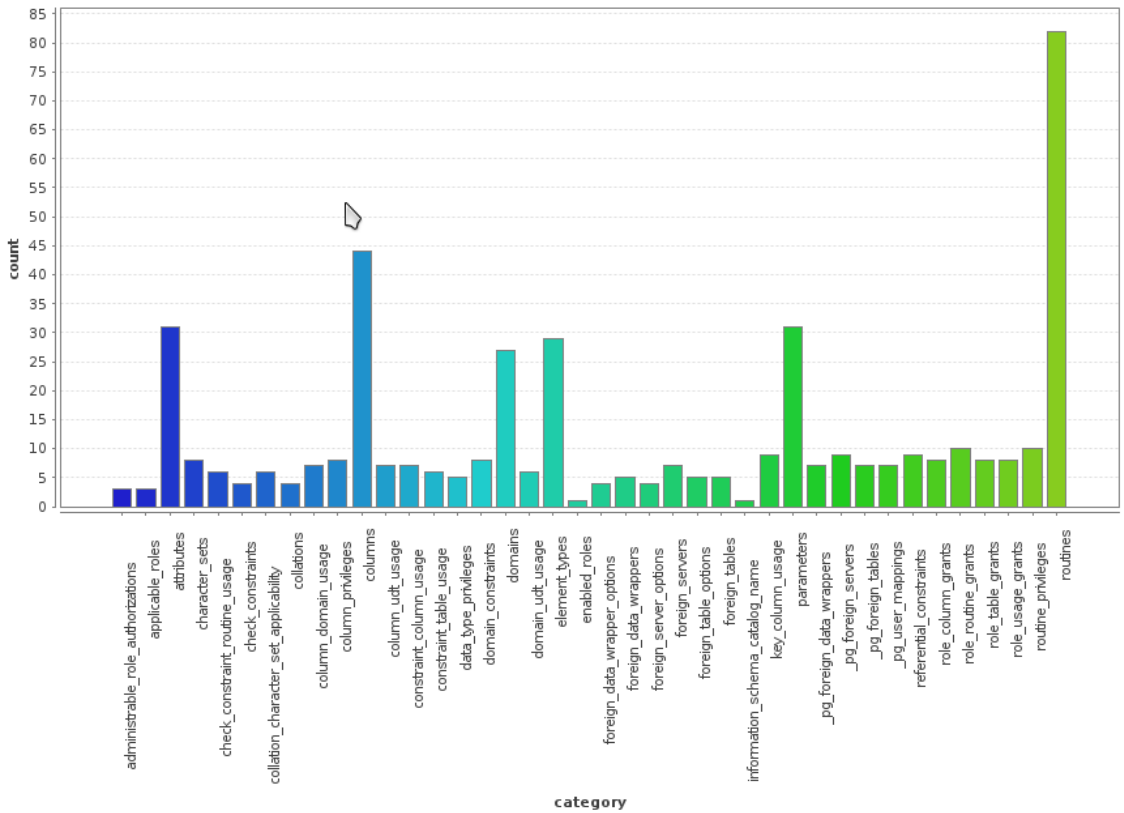
- Erster Schritt nach dem Import neuer Daten
- Was ist enthalten?

- Datenqualität
- Datenmenge

Data understanding

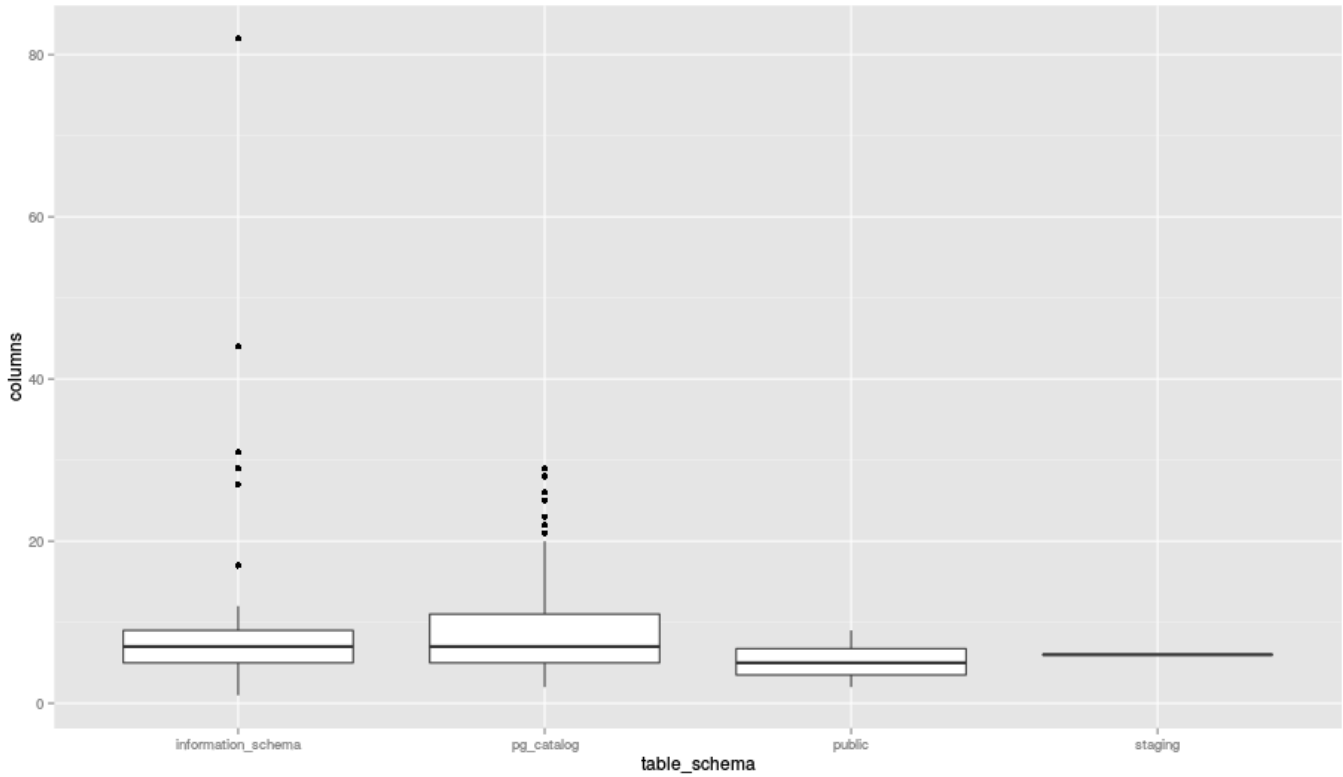
- Erster Schritt nach dem Import neuer Daten
- Was ist enthalten?
- Datenqualität
- Datenmenge
- Schlüssel zu anderen Datenbeständen

Visualisierung (RapidMiner)



datascientist.at

Visualisierung (R)



datascientist.at

Daten-Vorverarbeitung

Vorverarbeitung = Preprocessing

Bis zu 80 % der Projektzeit!

Daten-Vorverarbeitung

- Hereinkommende Daten selten „fertig“
- Zusammenführung unterschiedlicher Datensätze

Daten-Vorverarbeitung

- Hereinkommende Daten selten „fertig“
- Zusammenführung unterschiedlicher Datensätze
- Filtern, Bereinigung

Daten-Vorverarbeitung

- Hereinkommende Daten selten „fertig“
- Zusammenführung unterschiedlicher Datensätze
- Filtern, Bereinigung
- Erstellung/Berechnung neuer Attribute

Daten-Vorverarbeitung

- Hereinkommende Daten selten „fertig“
- Zusammenführung unterschiedlicher Datensätze
- Filtern, Bereinigung
- Erstellung/Berechnung neuer Attribute
- Aggregierung

Beispiel in RapidMiner

- 2 Datensätze von OGD Wien
  - Bevölkerungsprognose pro Bezirk
  - Bezirksdaten

Beispiel in RapidMiner

- 2 Datensätze von OGD Wien
  - Bevölkerungsprognose pro Bezirk
  - Bezirksdaten
- Prognose nach Geschlecht gruppiert => aggregieren

Beispiel in RapidMiner

- 2 Datensätze von OGD Wien
  - Bevölkerungsprognose pro Bezirk
  - Bezirksdaten
- Prognose nach Geschlecht gruppiert => aggregieren
- Über Bezirkscode verknüpfen

Beispiel in RapidMiner

- 2 Datensätze von OGD Wien
  - Bevölkerungsprognose pro Bezirk
  - Bezirksdaten
- Prognose nach Geschlecht gruppiert => aggregieren
- Über Bezirkscode verknüpfen
- Prognostizierte Bevölkerungsdichte berechnen

Beispiel in RapidMiner

- 2 Datensätze von OGD Wien
  - Bevölkerungsprognose pro Bezirk
  - Bezirksdaten
- Prognose nach Geschlecht gruppiert => aggregieren
- Über Bezirkscode verknüpfen
- Prognostizierte Bevölkerungsdichte berechnen
- Demo

Prädiktive Modellierung

Prädiktive Modellierung

Prädiktive Modellierung

- Modell aufbauen (lassen)
  - Zielvariable

Prädiktive Modellierung

- Modell aufbauen (lassen)
  - Zielvariable
- Modell auf neue Daten anwenden
  - Vorhersage, Konfidenz

Prädiktive Modellierung

- Modell aufbauen (lassen)
  - Zielvariable
- Modell auf neue Daten anwenden
  - Vorhersage, Konfidenz
- Validierung

Richtige Validierung

- Modell nicht auf Eingangsdaten anwenden!

Richtige Validierung

- Modell nicht auf Eingangsdaten anwenden!
- Split Validation

Richtige Validierung

- Modell nicht auf Eingangsdaten anwenden!
- Split Validation
- Cross Validation

Richtige Validierung

- Modell nicht auf Eingangsdaten anwenden!
- Split Validation
- Cross Validation
- Demo in RapidMiner

Prädiktive Modellierung – Fortsetzung

- Vergleich verschiedener Lernverfahren

Prädiktive Modellierung – Fortsetzung

- Vergleich verschiedener Lernverfahren
- Parameteroptimierung

Prädiktive Modellierung – Fortsetzung

- Vergleich verschiedener Lernverfahren
- Parameteroptimierung
- Variation der Vorverarbeitung
  - Attributselektion

Prädiktive Modellierung – Fortsetzung

- Vergleich verschiedener Lernverfahren
- Parameteroptimierung
- Variation der Vorverarbeitung
  - Attributselektion
  - Attributgenerierung

Deployment

- Operationalisierung der Ergebnisse

Deployment

- Operationalisierung der Ergebnisse
- Automatisierte Vorverarbeitung und Vorhersagen

Deployment

- Operationalisierung der Ergebnisse
- Automatisierte Vorverarbeitung und Vorhersagen
- Regelmäßige Evaluierung und Optimierung

Anwendungen im privaten Kontext

- Lebensgestaltung: Open Data, OpenStreetMap

Anwendungen im privaten Kontext

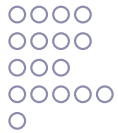
- Lebensgestaltung: Open Data, OpenStreetMap
- Kontrolle: Open Government Data, Firmen-Veröffentlichungen

Anwendungen im privaten Kontext

- Lebensgestaltung: Open Data, OpenStreetMap
- Kontrolle: Open Government Data, Firmen-Veröffentlichungen
- Hobbies
  - Wetter, Geodaten, GPS-Tracks, ...
  - ...

Anwendungen im privaten Kontext

- Lebensgestaltung: Open Data, OpenStreetMap
- Kontrolle: Open Government Data, Firmen-Veröffentlichungen
- Hobbies
  - Wetter, Geodaten, GPS-Tracks, ...
  - ...
- „Egometrics“, „Quantified self“
  - Fitness- und Gesundheitstracker, Smart Meter, Smart Vehicle
  - Internet of Things



Zusammenfassung

- Data Science – ein spannendes Thema

Zusammenfassung

- Data Science – ein spannendes Thema
- Frei verfügbare, einfach bedienbare Werkzeuge

Zusammenfassung

- Data Science – ein spannendes Thema
- Frei verfügbare, einfach bedienbare Werkzeuge
- Vorgehensweise

Zusammenfassung

- Data Science – ein spannendes Thema
- Frei verfügbare, einfach bedienbare Werkzeuge
- Vorgehensweise
- Anwendung im privaten Bereich

datascientist.at

Fragen?

- Balázs Bárány, <balazs@tud.at>

- https://datascientist.at/