UNIVERSITÄT MANNHEIM

RapidMiner
Linked Open Data Extension

Heiko Paulheim

RapidMiner Linked Open Data Extension

RapidMiner:

- Data-Mining- und Datenanalysewerkzeug
- Frei für wissenschaftliche Nutzung
- Datenanalyseprozesse durch Verdrahten von Operatoren
  - d.h., keine Programmierung
- Operatoren für Laden, Transformieren, Analysiere, Visualisierung...
- Skalierbar, Cloud-Adapter (Hadoop etc.) verfügbar
- 200,000 aktive Nutzer

- Eigene Erweiterungen (Extensions) sind möglich

RapidMiner Linked Open Data Extension

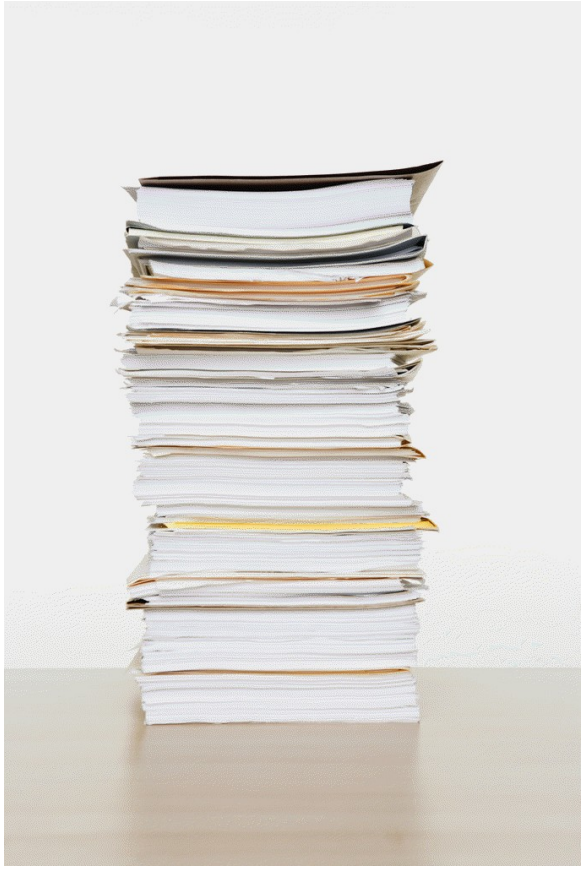
- Die Extension enthält Operatoren für
  - Zugriff auf lokale und Webdaten (RDF, SPARQL, ...)
  - Verknüpfung lokaler Daten mit Linked Open Data
  - Anreichern von lokalen Daten mit Daten aus der LOD Cloud
  - Automatisches Verfolgen von Links
  - Ausnutzung von Schemainformation für bessere Attributauswahl
  - Integration von Daten aus verschiedenen Quellen

- Kann ohne technisches Wissen über LOD verwendet werden

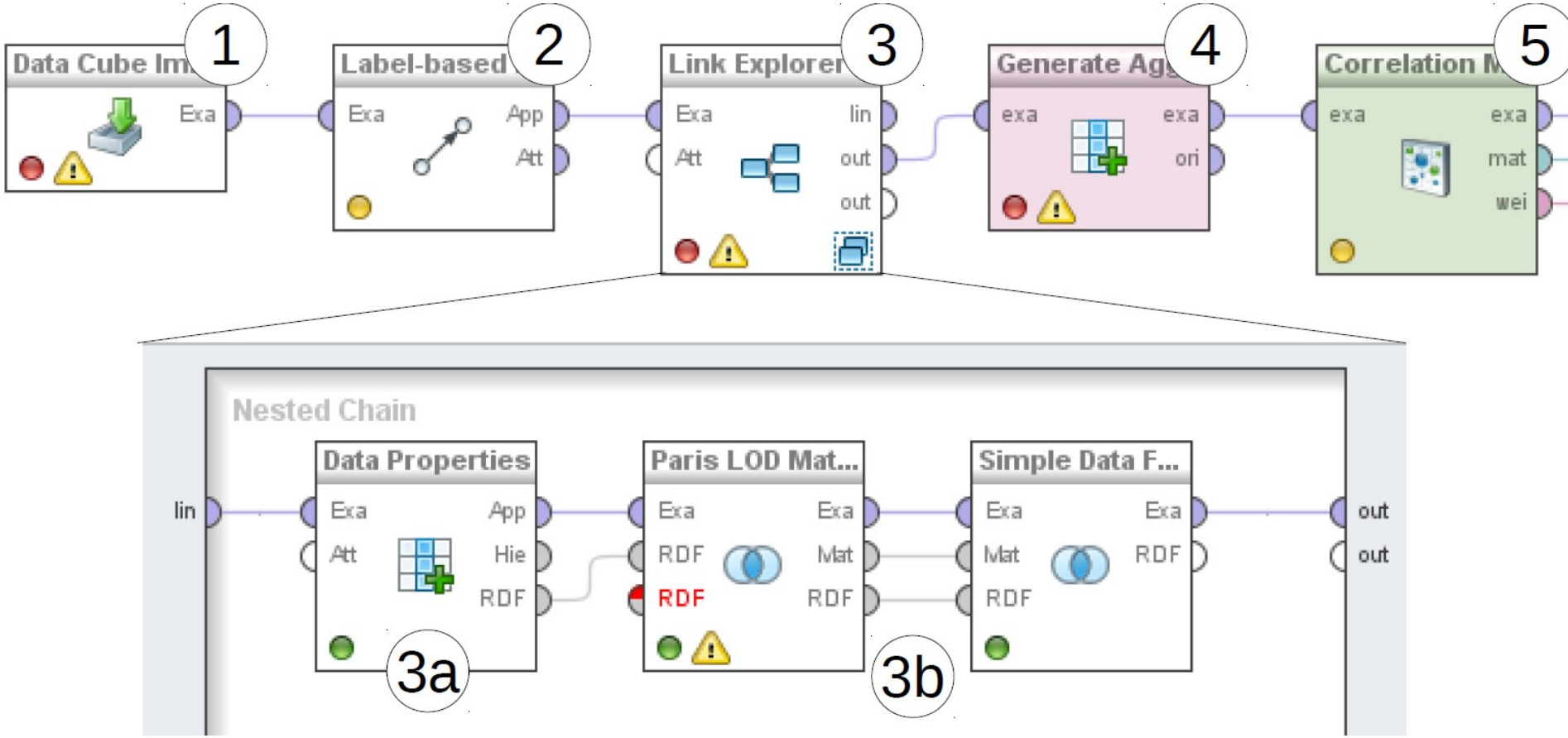
Beispiel

- Welche Faktoren korrelieren mit einer Zunahme wissenschaftlicher Publikationen?

- RapidMiner workflow:
  - Datenimport aus WorldBank RDF Data Cube
  - Länder mit DBpedia verknüfen
  - Zusätzliche Datensets auffinden
  - Attribute generieren und integrieren
  - Resultate analysieren

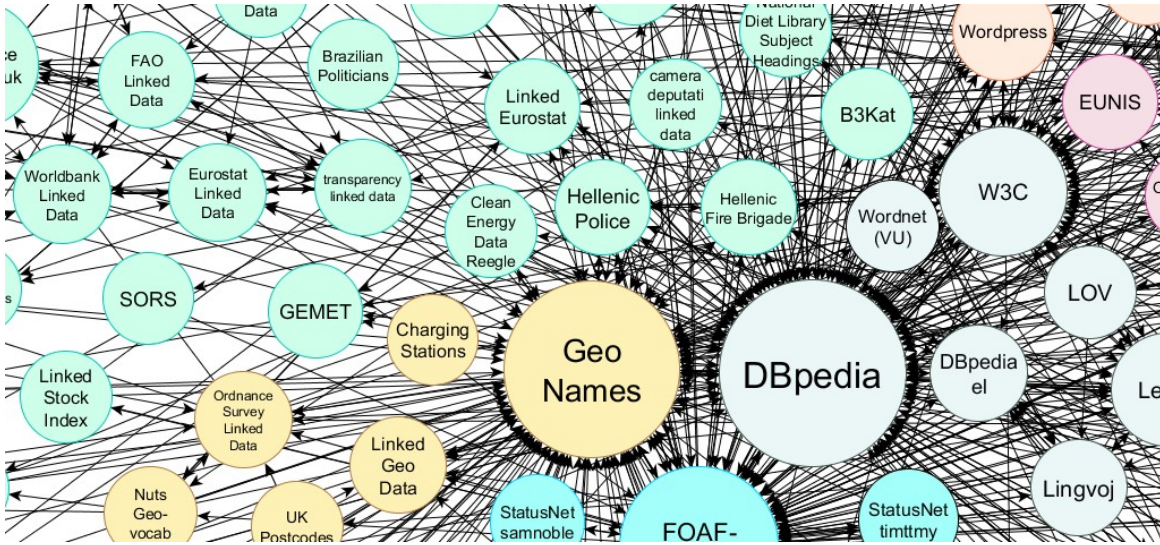


Beispiel: RapidMiner-Prozess



Beispiel: Datensets

- Durch Verfolgen von Links werden verschiedene Datensets gefunden:
  - DBpedia
  - Linked GeoData
  - Eurostat
  - GeoNames
  - WHO's Global Health Observatory
  - Linked Energy Data
  - OpenCyc
  - World Factbook
  - YAGO



- Verwandte Daten werden integriert
  - z.B. Bevölkerungszahlen aus verschiedenen Ländern

Beispiel: Datenerweiterung

- RapidMiner verwendet ein tabellarisches Datenmodell

Input Table | Link | Additional Attributes

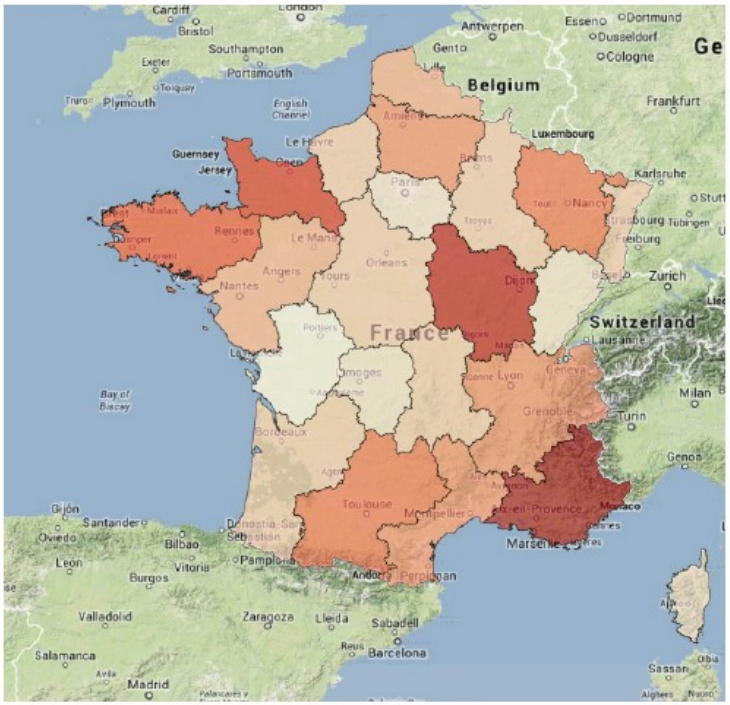
Row No.	cylinders	displacement	horsepower	weight	accelaeration	model	origin	mpg	car	car_uri	http://dbped...	http://dbped...	http://dbped...	http://dbped...	http://dbped...	http://dbped...	http://dbped...	hl
1	8	307	130	3504	12	70	1	18	chevrolet che	http://dbpedi	0	0	0	1	1	0	0	0
2	8	350	165	3693	11.500	70	1	15	buick skylark	http://dbpedi	1	0	0	1	1	0	0	0
3	8	318	150	3436	11	70	1	18	plymouth sa	http://dbpedi	1	0	0	1	1	0	0	0
4	8	304	150	3433	12	70	1	16	amc rebel	http://dbpedi	1	0	0	1	1	0	0	0
5	8	302	140	3449	10.500	70	1	17	ford torino	http://dbpedi	1	0	0	1	0	0	0	0
6	8	429	198	4341	10	70	1	15	ford galaxie	http://dbpedi	1	0	0	1	1	0	0	0
7	8	454	220	4354	9	70	1	14	chevrolet imp	http://dbpedi	1	0	0	1	1	0	0	0
8	8	440	215	4312	8.500	70	1	14	plymouth fur	http://dbpedi	1	0	0	1	1	0	0	0
9	8	455	225	4425	10	70	1	14	pontiac catal	http://dbpedi	1	1	0	1	1	0	0	0
10	8	390	190	3850	8.500	70	1	15	amc ambas	http://dbpedi	1	0	0	1	1	0	0	0
11	8	383	170	3563	10	70	1	15	dodge challe	http://dbpedi	1	0	0	1	1	0	0	0
12	8	340	160	3609	8	70	1	14	plymouth 'cu	http://dbpedi	1	0	0	1	1	0	0	0
13	8	400	150	3761	9.500	70	1	15	chevrolet mc	http://dbpedi	1	0	0	1	1	0	0	0
14	8	455	225	3086	10	70	1	14	buick estate	http://dbpedi	1	0	0	1	0	0	0	0
15	4	113	95	2372	15	70	3	24	toyota coron	http://dbpedi	1	0	0	1	1	0	0	0
16	6	198	95	2833	15.500	70	1	22	plymouth du	http://dbpedi	1	0	0	1	0	0	0	0
17	6	199	97	2774	15.500	70	1	18	amc hornet	http://dbpedi	1	0	0	1	1	0	0	0
18	6	200	85	2587	16	70	1	21	ford maveric	http://dbpedi	1	0	0	1	0	0	0	0
19	4	97	88	2130	14.500	70	3	27	datsun	http://dbpedi	0	0	0	1	0	0	0	0
20	4	97	46	1835	20.500	70	2	26	volkswagen	http://dbpedi	1	0	0	1	0	1	0	0
21	4	110	87	2672	17.500	70	2	25	peugeot 504	http://dbpedi	0	0	0	0	0	0	0	0
22	4	107	90	2430	14.500	70	2	24	audi 100	http://dbpedi	0	0	0	0	0	0	0	0
23	4	104	95	2375	17.500	70	2	25	saab 99	http://dbpedi	0	0	0	0	0	0	0	0
24	4	121	113	2234	12.500	70	2	26	bmw 2002	http://dbpedi	0	0	0	0	0	0	0	0
25	6	199	90	2648	15	70	1	21	amc gremlin	http://dbpedi	1	0	0	1	0	0	0	0
26	8	360	215	4615	14	70	1	10	ford f250	http://dbpedi	1	0	0	1	0	0	0	0
27	8	307	200	4376	15	70	1	10	chevy	http://dbpedi	1	0	0	1	1	0	0	0
28	8	318	210	4382	13.500	70	1	11	dodge d	http://dbpedi	0	0	0	1	0	0	0	0

Beispiel: Ergebnisse

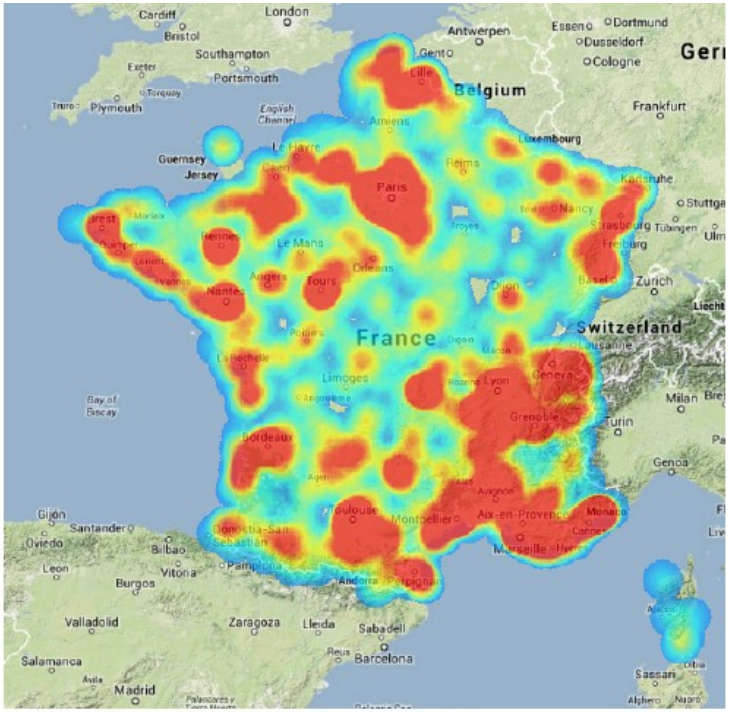
- Faktoren, die mit starker Zunahme an Publikationen korrelieren
  - Fragile State Index
  - Human Development Index
  - Bruttoinlandsprodukt
    - Höhere Investionen in Forschungsförderung?
  - Für EU-Länder: Anzahl der Sitze im EU-Parlament
    - Bessere Interessenvertretung bei Zuteilung von Fördermitteln?
  - Viele Klimafaktoren (Temperatur, Niederschlag...)
    - Ungleiche Reichtumsverteilung entlang der Klimazonen?

Weitere Use Cases

- Analyse von Arbeitslosigkeitsdaten in Frankreich (SemStats'13)
  - Hintergrundwissen von DBpedia, Eurostat, Linked Geo Data
  - Links von DBpedia zu GADM für Visualisierung



(a) Unemployment by region



(b) Heat map of police stations

Weitere Use Cases

- Beispielkorrelationen
  - Afrikanische Inseln, Inseln im Indischen Ozean (positiv)
  - BIP (negativ)
  - Verfügbares Einkommen (negativ)
  - Krankenhausbetten pro Einwohner (negativ)
  - FuE-Ausgaben (negativ)
  - Energieverbrauch (negativ)
  - Bevölkerungszuwachs (positiv)
  - Todesopfer im Straßenverkehr (negativ)
  - Fast-Food-Restaurants (positiv)
  - Polizeistationen (positiv)

Weitere Use Cases

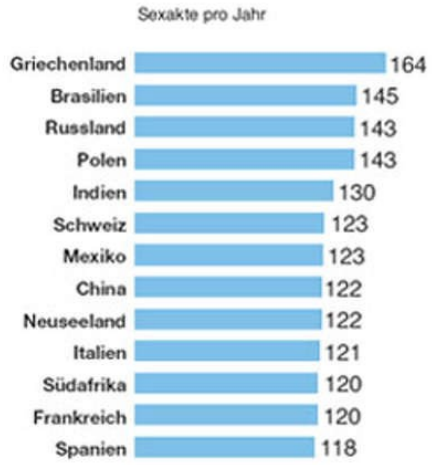
- Recommender-Systeme für Buchempfehlungen (ESWC'14)
- Kombiniert zwei Extensions
  - Linked Open Data extension
  - Recommender system extension
- Hintergrunddaten über Bücher werden für verbesserte Empfehlungen genutzt
  - Bestes System (von 24) in zwei von drei Tasks

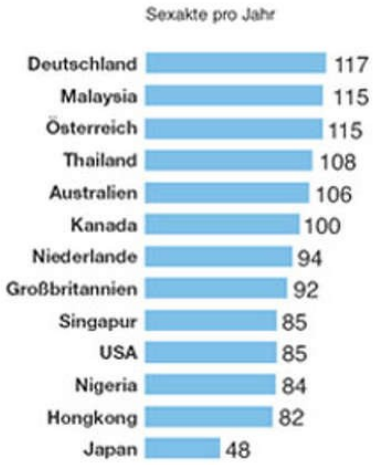
Weitere Beispiele

- Statistik: Suizidraten nach Land
  - http://www.washingtonpost.com/wp-srv/world/suiciderate.html
- Korrelationen
  - Demokratien haben die geringsten Suizidraten
  - Hoher HDI → niedrige Suizidrate
  - Hohe Bevölkerungsdichte → hohe Suizidrate
  - Geographie
    - Am Meer → niedrig
    - In den Bergen → hoch
  - Hoher Gini-Index → niedrige Suizidrate
    - Hoher Gini-Index = ungleiche Einkommensverteilung
  - Hohe Nutzung von Kernergie → hohe Suizidrate

Other Examples

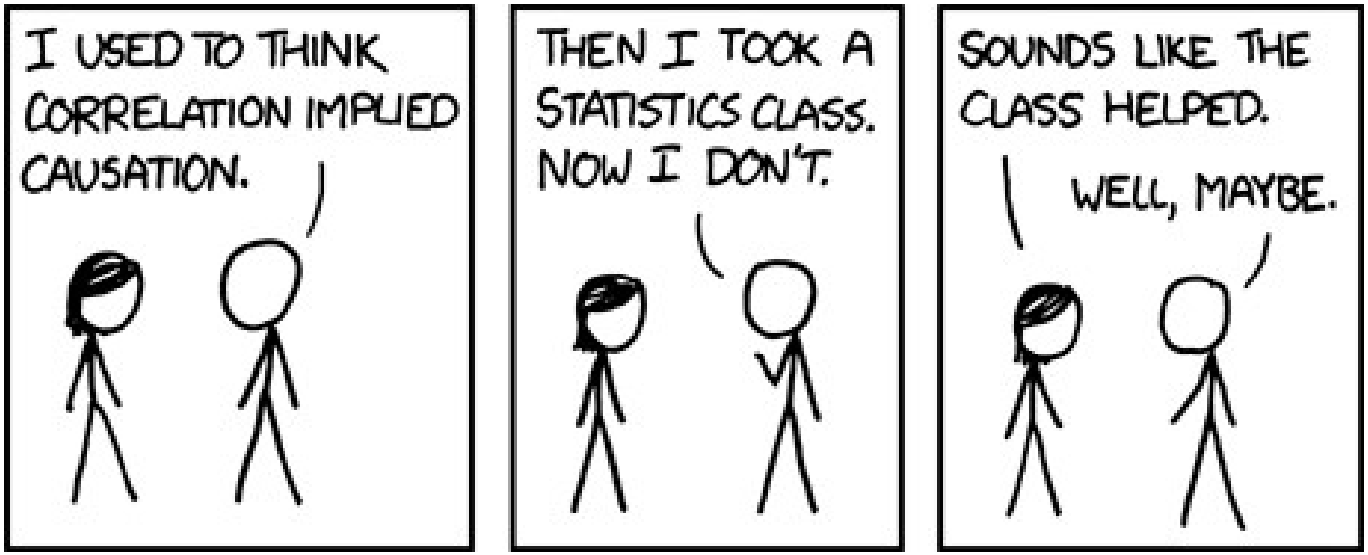
- Data set: Durex-Report zur sexuellen Aktivität
  - http://chartsbin.com/view/uya
- Ergebnisse
  - Nach Geographie:
    - In Europa höher als in Asien
    - Niedrig in Inselstaaten
  - Nach Sprache:
    - Englisch → niedrig
    - Französisch → hoch
  - Niedriges Durchschnittsalter → hoch
  - Hohes BIP pro Kopf → niedrig
  - Hohe Arbeitslosigkeit → hoch
  - Hohe Anzahl von Internet-Service-Providern → niedrig





Warnung

- Wir haben in all diesen Beispielen nach *Korrelationen* gesucht
  - Interpretieren müssen wir die immer noch selbst



UNIVERSITÄT MANNHEIM

RapidMiner
Linked Open Data Extension

Heiko Paulheim