Big Data

Moritz Spindelhirn

Motivation

- **Strategisch**
    - Gewinnoptimierung
    - User Experience verbessern (Nutzergruppen erkennen)
- **Operativ**
    - Verhalten/Lastspitzen prognostizieren

Inhalt

- Definition / Abgrenzung
- Praktischer Einsatz
- Tools
- Fazit

Inhalt

- **Definition / Abgrenzung**
- Praktischer Einsatz
- Tools
- Fazit

Definition

Big ⟷ Data

Definition

Big ⟷ Data

- Genaue Anzahl nicht definiert
- Zu groß für klassische Datenverarbeitungsmethoden
- Datenmenge hoch dynamisch

Definition

Big ⟷ **Data**

- Unstrukturierte Daten
- Daten aus verschiedenen Quellen

Definition

Definition

- Frühe Nennungen ~1996
- Begriff im Wandel
- Viele synonyme Begriffe

Definition

Business Intelligence

- Sicht auf die Prozesse und Abläufe
- Unternehmensziele

Definition

Collective Intelligence

- Fokus auf Daten von vielen Nutzern
- Vorteile aus Diversifizierung

Definition

Machine Learning

- Aus dem Bereich der intelligenten Systeme
- Beschreibt Algorithmen
- Daten als Input, Verhalten als Output

Inhalt

- Definition / Abgrenzung
- **Praktischer Einsatz**
- Tools
- Fazit

Einsatz

Praktischer Einsatz

- Recommendation Engine
- Gruppeneinteilung
- Rankings
- Optimierung

Einsatz

Recommendation Engine

- **Ziel**
    - Interessante Inhalte für Nutzer finden
- **Vorgehen**
    - Inhalte/Dokumente klassifizieren
    - Vorzüge identifizieren
    - Wert klassifizieren / messen

Elektronik

Apple MB707ZM A1300 ...
(26)
~~EUR 16,00~~ EUR 12,49
Warum empfohlen?

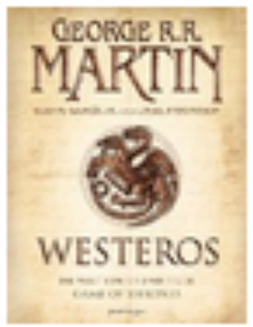
Apple USB Power ...
(41)
EUR 10,95
Warum empfohlen?

Apple MD836ZM / A ...
(84)
~~EUR 19,00~~ EUR 15,99
Warum empfohlen?

› Alle Empfehlungen in Elektronik anzeigen

Bücher

Der leere Thron
Bernard Cornwell
EUR 10,99
Warum empfohlen?



Westeros: Die Welt von ...
George R.R. Martin
(42)
EUR 29,99
Warum empfohlen?

Das Licht der Welt: ...
Daniel Wolf
(78)
EUR 9,99
Warum empfohlen?

› Alle Empfehlungen in Bücher anzeigen

Schönheit

Braun 70S ...
(468)
~~EUR 49,00~~ EUR 33,96
Warum empfohlen?

Braun CoolTec CT5cc ...
(91)
~~EUR 220,00~~ EUR 126,90
Warum empfohlen?

Braun Series 5 5090cc ...
(534)
~~EUR 200,00~~ EUR 159,00
Warum empfohlen?

› Alle Empfehlungen in Schönheit anzeigen

www.amazon.de (Mein Amazon ; Stand 10.05.2015)

Methode 1

- Historie des Nutzers
  - Gekauft
  - Angesehen

2 19 Machine Learning and Data Mining for Computer Security: Methods and Applications [PDF]
(rtsalresearch.googlecode.com)
eingereicht 5 hours ago von galapag0 in /r/netsec
1 Kommentar Weitersagen Speichern Ausblenden melden

3 • Released a Framework Agnostic HTML Helper - Looking for feedback (self.PHP)
eingereicht 59 minutes ago von snsmurf in /r/PHP
2 Kommentare Weitersagen Speichern Ausblenden melden

4 5 Brady's pictures from the Star Wars event (bradyharanblog.com)
eingereicht 2 hours ago von tfofurn in /r/HelloInternet
2 Kommentare Weitersagen Speichern Ausblenden melden

5 7 Cloud cross-browser test automation with Protractor & Browserstack (blog.wishtack.com)
eingereicht 4 hours ago von yjaaldi in /r/node
kommentieren Weitersagen Speichern Ausblenden melden

6 99 OOP, Javascript, and so-called Classes (raganwald.com)
eingereicht 19 hours ago von aeflash in /r/javascript
38 Kommentare Weitersagen Speichern Ausblenden melden

7 8 Wrote an app for people with HVV CC-Karte in Hamburg (self.hamburg)
eingereicht 7 hours ago von egze in /r/hamburg
2 Kommentare Weitersagen Speichern Ausblenden melden

8 42 Introducing gb, a project based build tool for the Go programming language (dave.cheney.net)
eingereicht 18 hours ago von dgryski in /r/golang
8 Kommentare Weitersagen Speichern Ausblenden melden

Reddit (Personalisierte Startseite) Stand 12.05.2015

Methode 2

- Personen mit ähnlichem Interesse/Verhalten
  - Freundeslisten / Gruppen
  - „Kunden, die diesen Artikel gekauft haben, kauften auch"

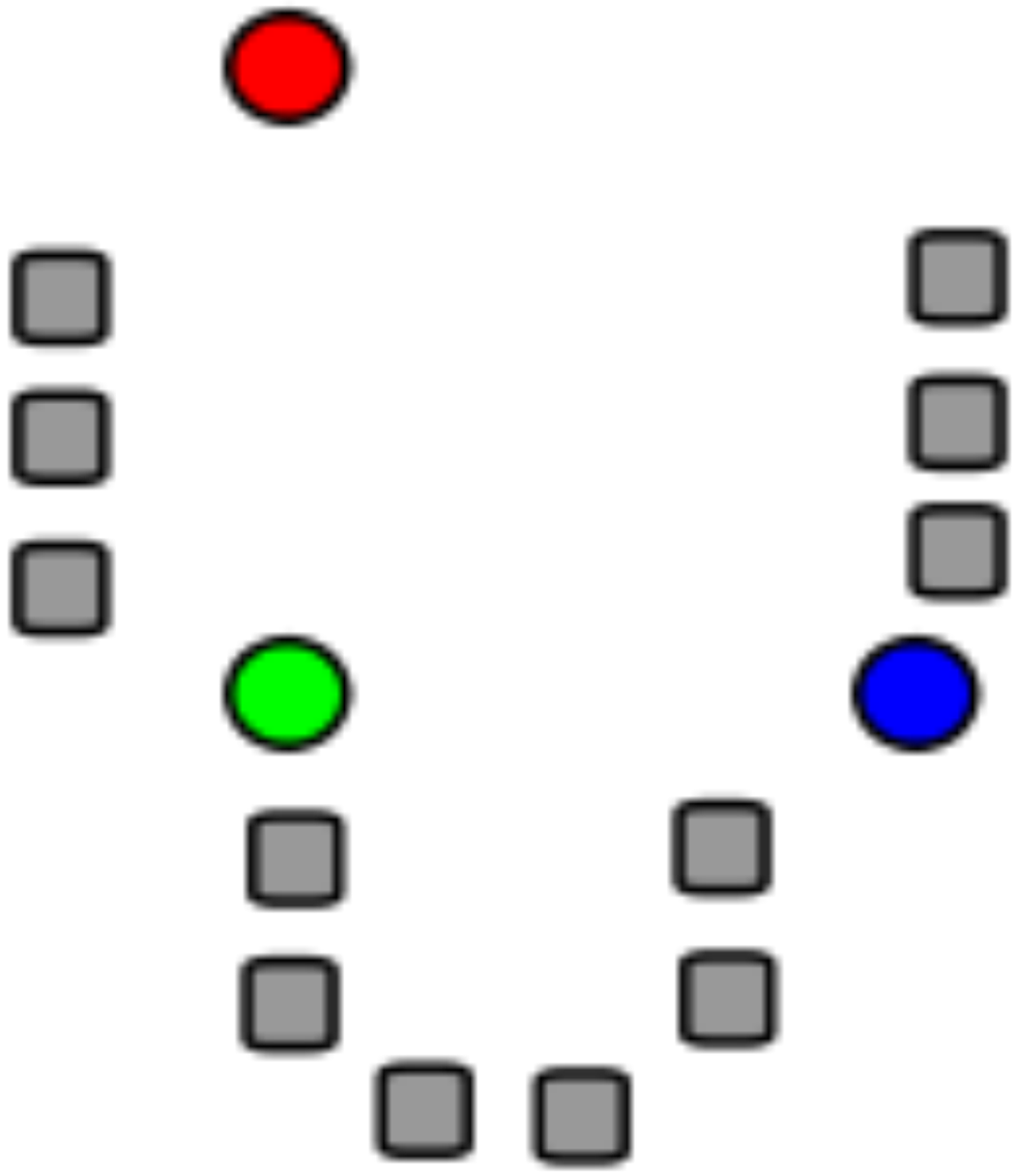
Einsatz

Gruppeneinteilung Clustering

- **Ziel**
    - Datenquellen in Gruppen einteilen
- **Vorgehen**
    - Unterteilungskriterien festlegen
    - Gruppen bilden

Einsatz

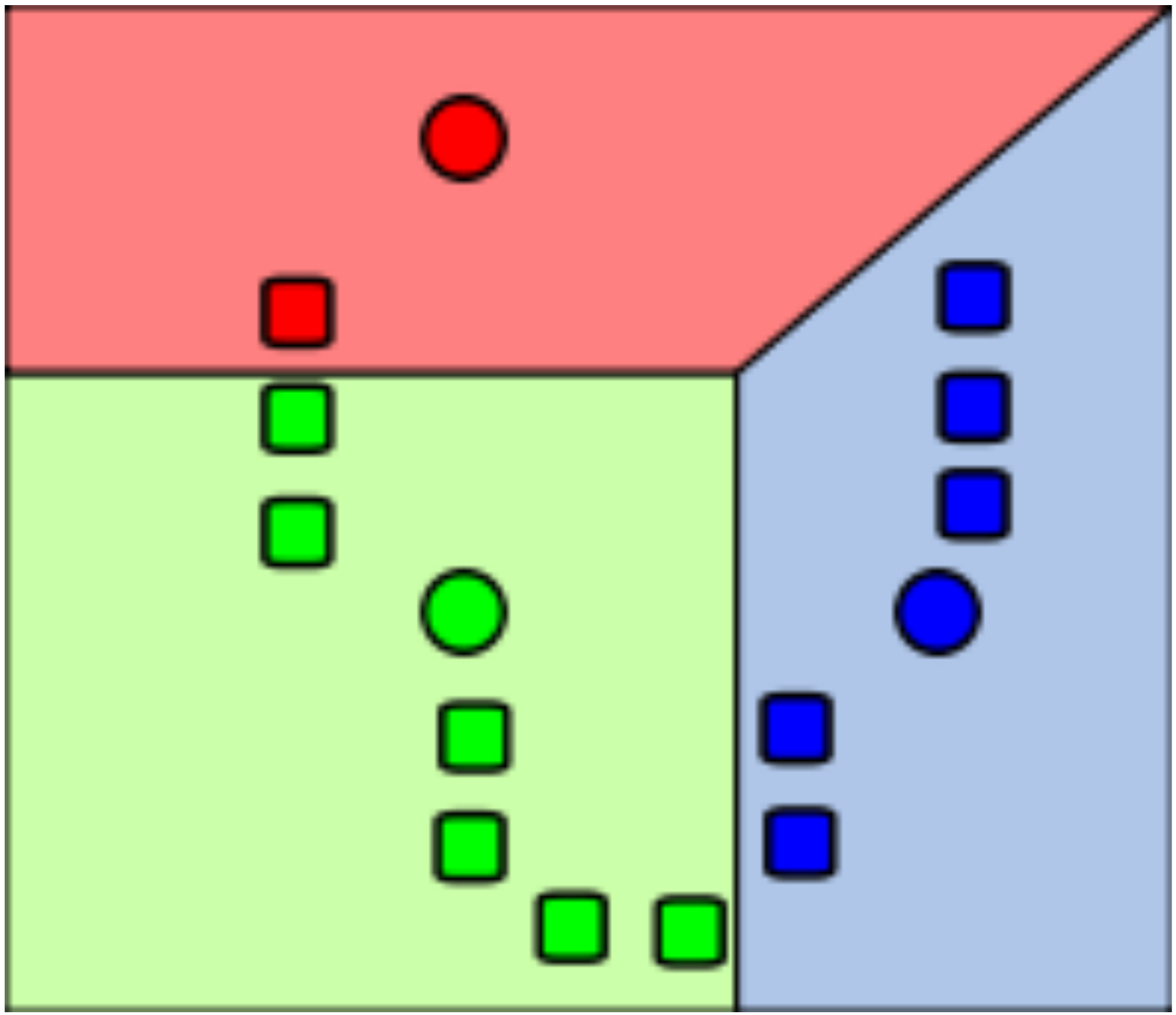
K-Means



http://de.wikipedia.org/wiki/K-Means-Algorithmus (28.04.2015)

Einsatz

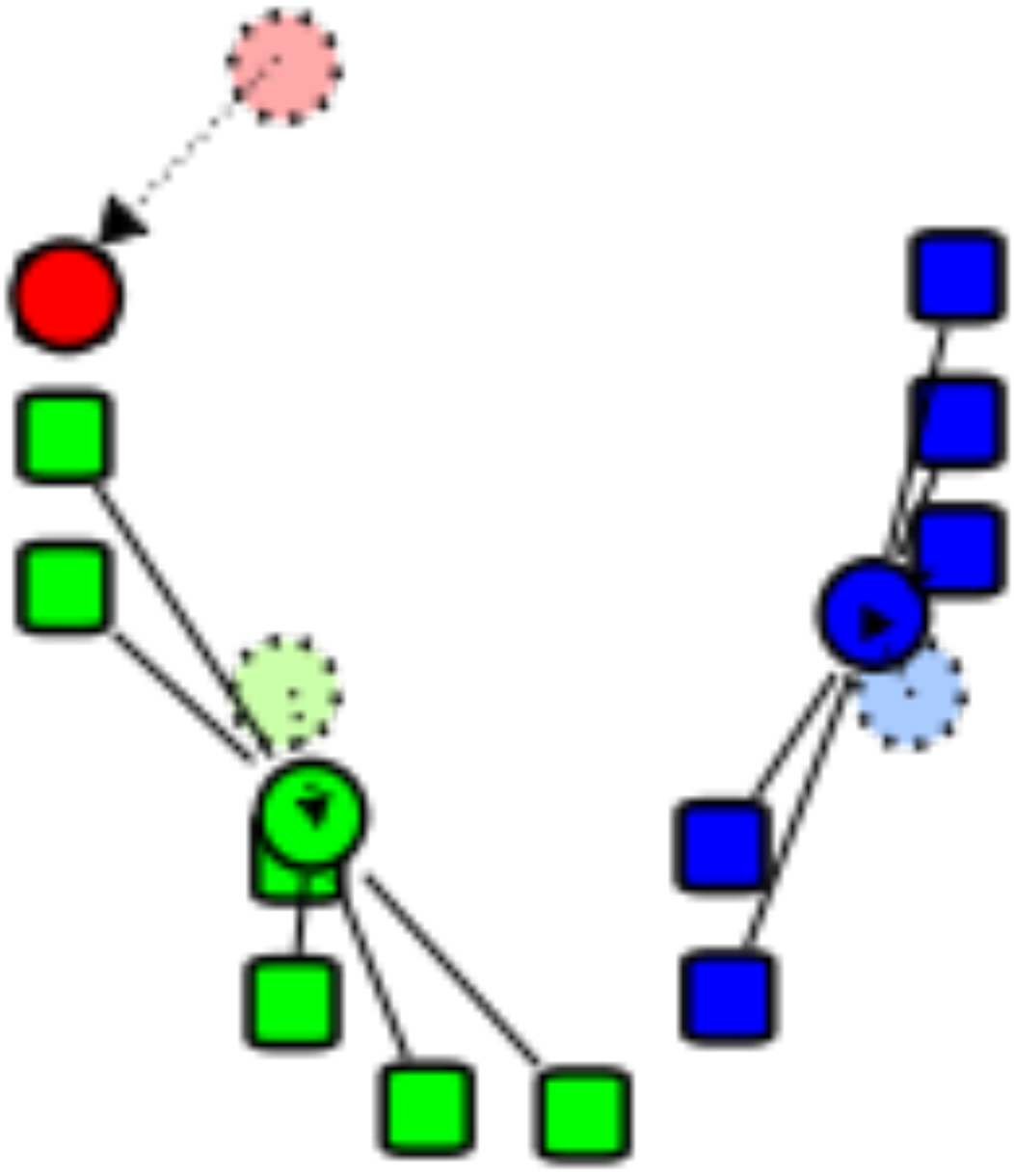
K-Means



http://de.wikipedia.org/wiki/K-Means-Algorithmus (28.04.2015)

Einsatz

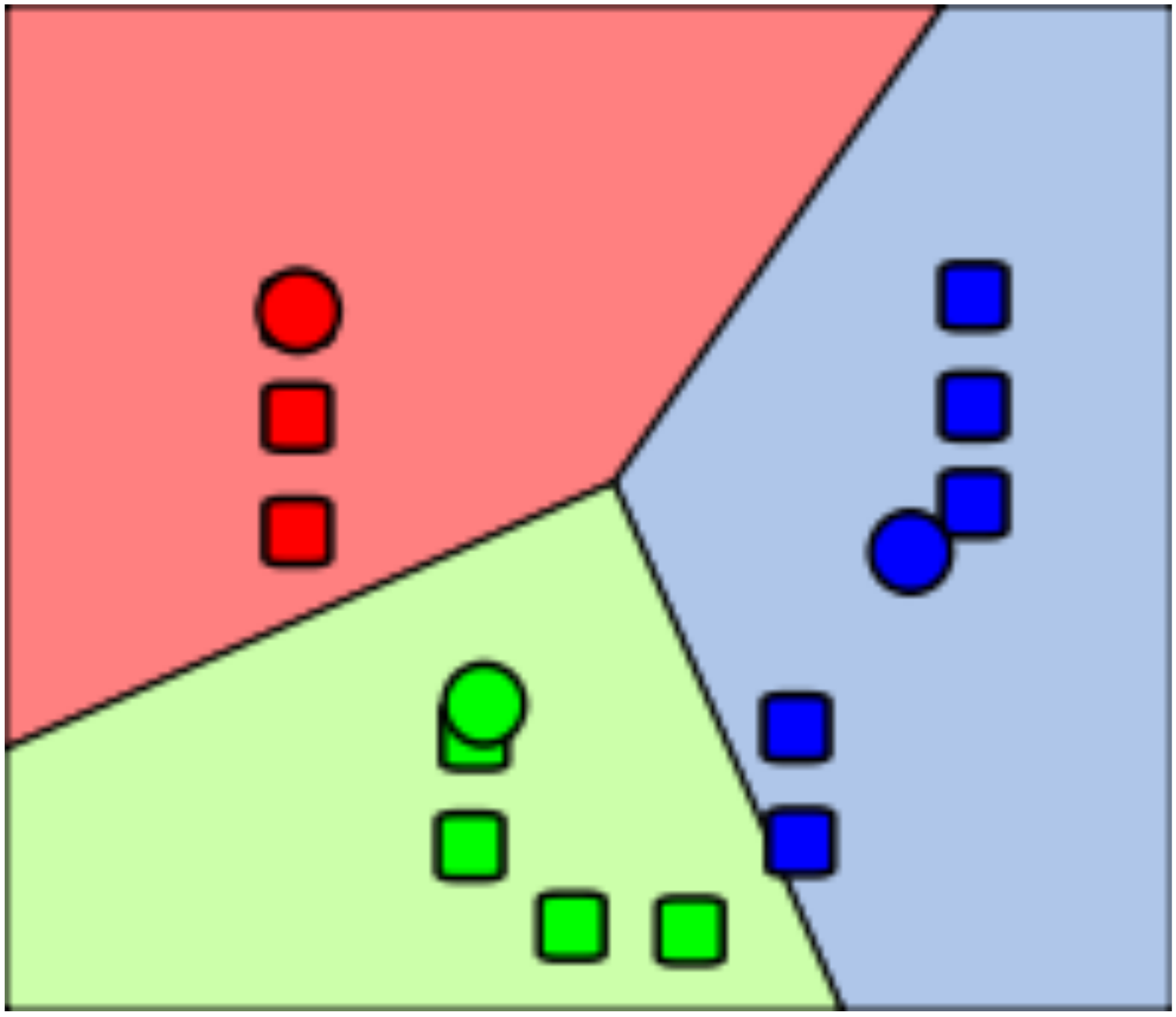
K-Means



http://de.wikipedia.org/wiki/K-Means-Algorithmus (28.04.2015)

Einsatz

K-Means

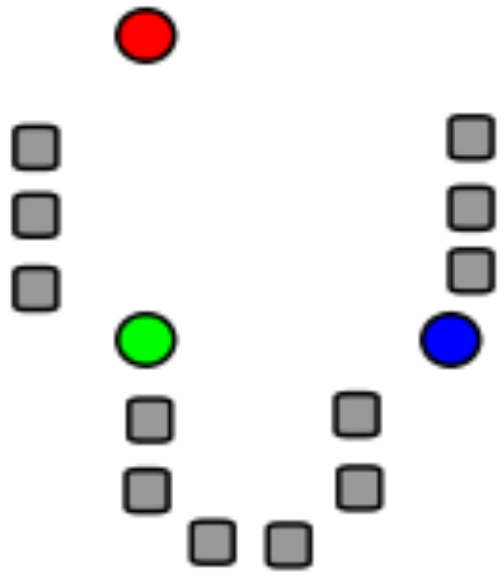


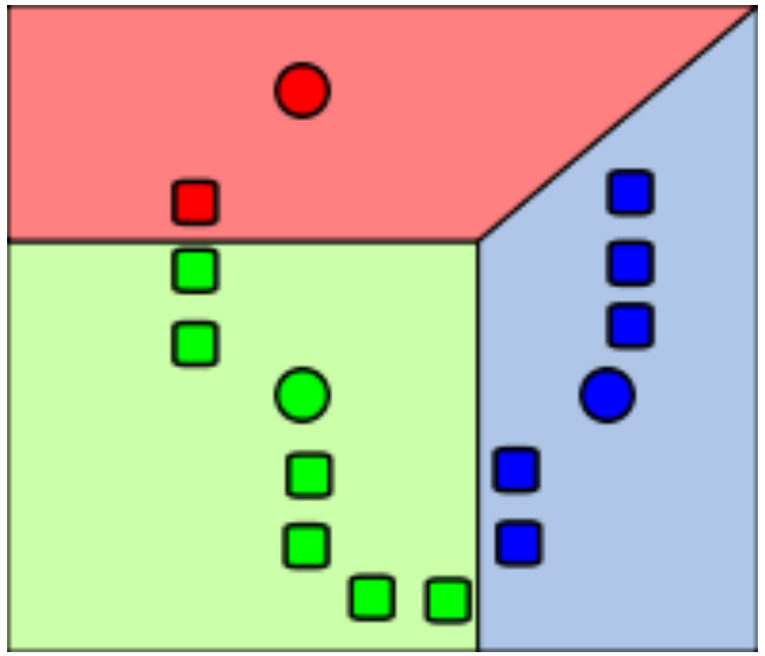
http://de.wikipedia.org/wiki/K-Means-Algorithmus (28.04.2015)

Einsatz

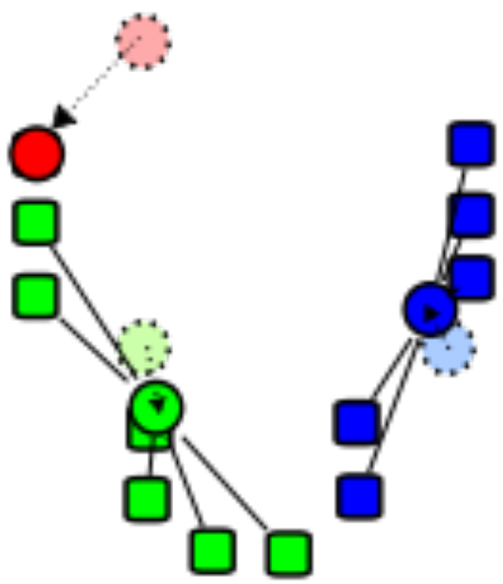
K-Means

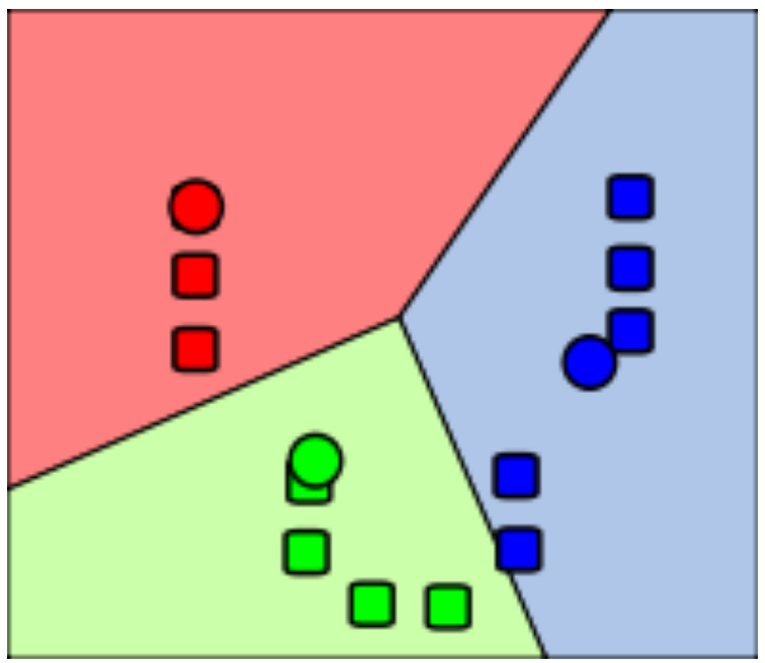
1.





2.





http://de.wikipedia.org/wiki/K-Means-Algorithmus (28.04.2015)

Einsatz

Searching
Ranking

- **Ziel**
    - Inhalte finden die am besten zu Suchbegriff passen
- **Vorgehen**
    - Ggf. Dokumente finden (Crawling) / Indexing
    - Suchanfragen bearbeiten

Einsatz

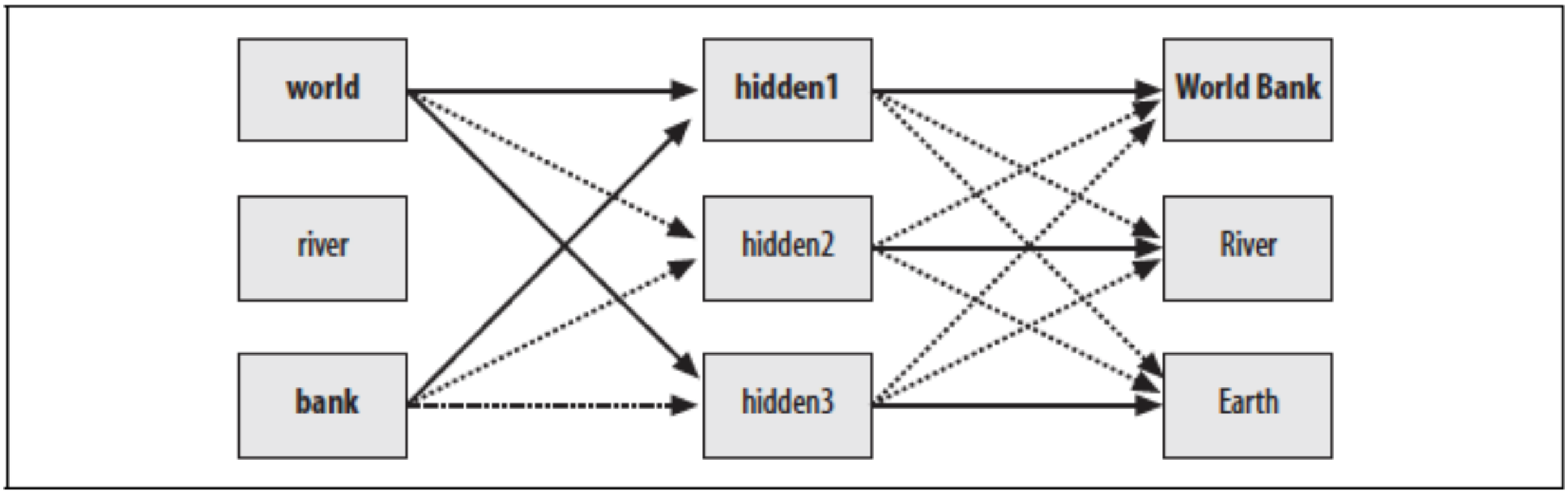
Content-Based Ranking

- **Anzahl der Vorkommen**
    - Je häufiger ein Word im Inhalt vorkommt desto besser
- **Position im Dokument**
    - Je weiter oben ein Begriff auftaucht desto relevanter.
- **Zusammenhang**
    - Mehrere Suchbegriffe sollten enger zusammen sein

Einsatz

Artificial Neural Network

Eingabe: „World Bank“



Toby Segaran (2007): Programming Collective Intelligence

Einsatz

Optimierung

- **Ziel**
    - Bessere Lösung zu komplexen Problemen mit vielen Variablen finden
- **Vorgehen**
    - Verschiedene Lösungen erarbeiten
    - Lösungen bewerten (Cost Function)

Einsatz

Algorithmen

- Hill Climbing
    - Lokales Minimum Problem
- Simulated Annealing
    - Verbessertes Hill Climbing durch erlaubte Sprünge
    - Wahrscheinlichkeit sinkt mit jeder Iteration
- Genetische Algorithmen
    - Population bilden, Über Generationen verbessern

Inhalt

- Definition / Abgrenzung
- Praktischer Einsatz
- **Tools**
- Fazit

Tools

Apache Hadoop

- Verteilte Systeme
- MapReduce
- OpenSource
- Partnervortrag
- HBase (Datenbank)

Tools

Apache Spark

- Verteilte Systeme
- Schnell
- In-Memory computing
- Baut auf Hadoop auf

Tools

prediction.io

- Machine Learning Platform
- „Templates“ für verschiedene Algorithmen
  http://templates.prediction.io/
- Baut wiederum auf Apache Park, HBase und Spray auf

Tools

Elasticsearch

- Datenbank
- Optimiert für Volltextsuche
- Dokumente in JSON Format

Inhalt

- Definition / Abgrenzung
- Praktischer Einsatz
- Tools
- **Fazit**

Fazit

- Breites Spektrum
    - Applikationen rutschen in Big Data
    - Neue Einsatzgebiete im nicht IT Umfeld
- Kein Ende in Sicht
    - Neue Tools
    - Vereinfachen


THIS ONLY TELLS ME IF THEY WERE NAUGHTY OR NICE! WHERE'S THE BIG DATA!
© D.Fletcher for CloudTweaks.com

Literaturverzeichnis

- Toby Segaran (2007): Programming Collective Intelligence

- Jonas Freiknecht (2014): Big Data in der Praxis