UNIVERSITY OF MANNHEIM

Web Mining

Introduction and Course Outline

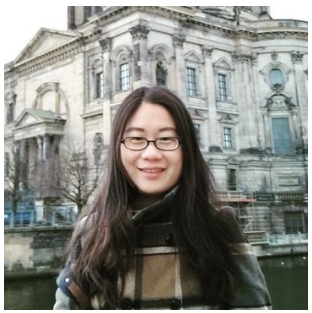
Simone Paolo Ponzetto
Anne Lauscher
Chia-Chien Hung

FSS 2020

Hi!

- Since 2016 full professor of Information Systems (Wifo) in Mannheim
- Main research areas
  - Knowledge Acquisition
  - Natural Language Processing
  - Computational Social Science

- TAs: Anne Lauscher, Chia-Chien Hung
- {anne, chia-chien, simone}@informatik.uni-mannheim.de



Introduction and Course Organization

1. Course Organization
2. The World Wide Web
    1. The Classic Document Web
    2. The Web of Data
    3. Web 2.0 Applications
3. What is Web Mining?
    1. Web Usage Mining
    2. Web Structure Mining
    3. Web Content Mining

1. Course Organization

- **Lecture (IE 671)**
  - covers different types of Web Mining methods
  - presents examples of Web Mining applications
  - discusses how to evaluate learned models
- **Labs**
  - students get their hand dirty with tools and code
- **Evaluation**
  - 60 min final exam
- **Project (IE 684)**
  - teams of four students realize a Web Mining project
  - teams may choose their own data sets and tasks
    (in addition, we will propose some suitable data sets and tasks)
  - write a summary about the project, present the project results
- **Evaluation**
  - Implementation + report + presentation

Course Organization

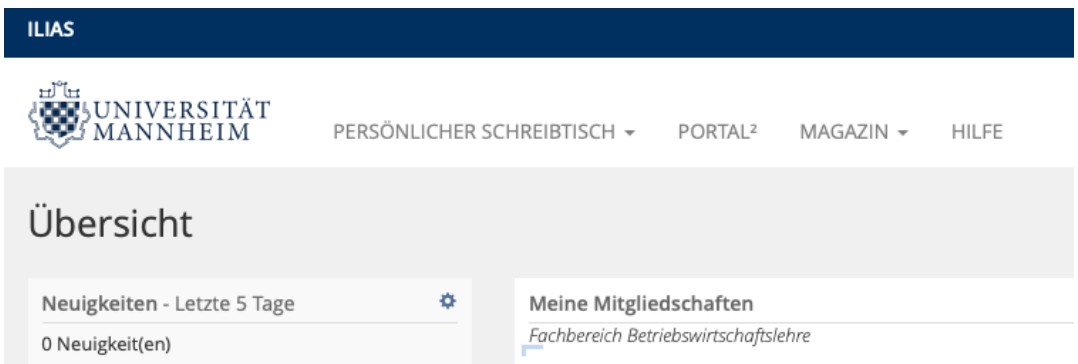
- **Materials**
  - **ILIAS eLearning System, https://ilias.uni-mannheim.de/**

- **Time and Location**
  - **Lecture:**

    Tuesday, 10:15 to 11:45,
    Room: B 6, A104

  - **Labs:**

    Tuesday, 15:30 to 17:00,
    Room: B 6, A104



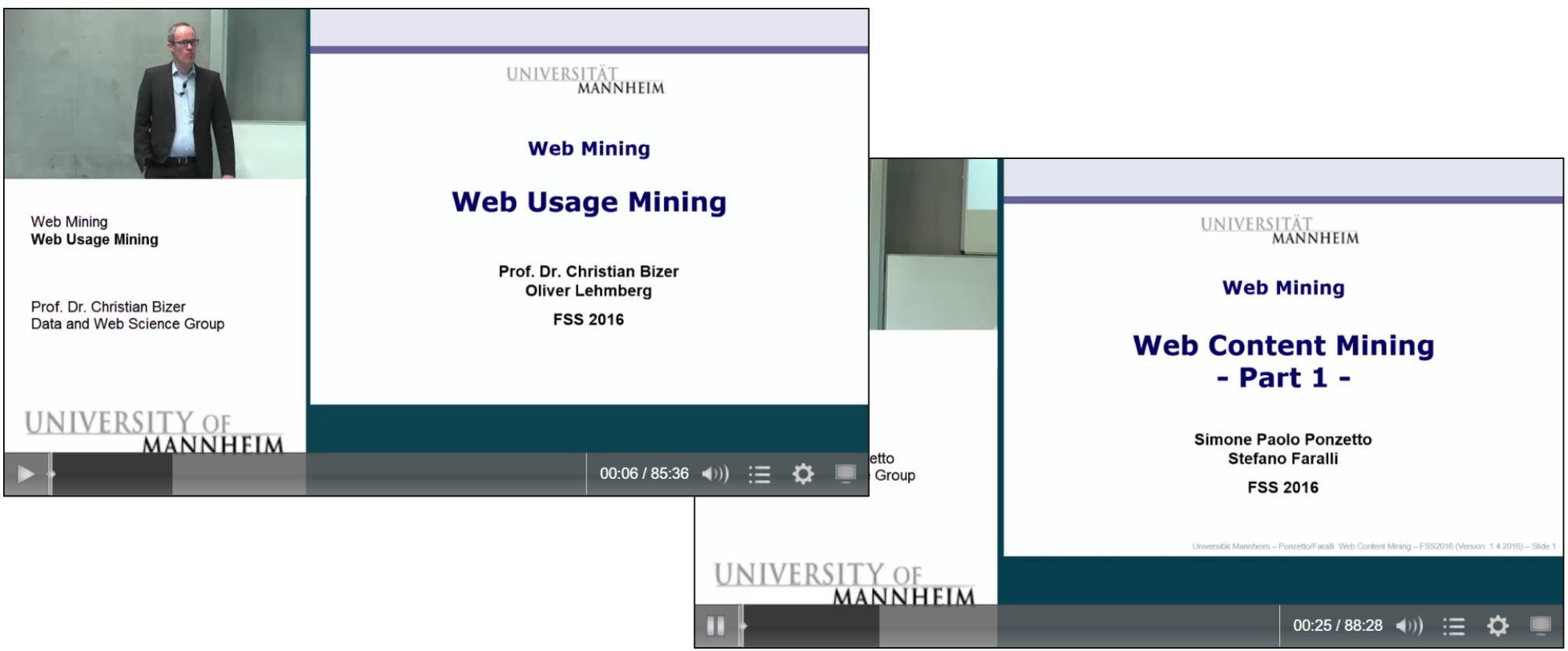
Schedule

Morning session (10:15-11:45)		Afternoon session (15:30-15:00)	
11/2/20	Lecture: Introduction to Web Mining	11/2/20	-
18/2/20	Lecture: Recommender Systems	18/2/20	Lab: Recommender Systems
25/2/20	Lecture: Recommender Systems	25/2/20	Lab: Recommender Systems
3/3/20	Lecture: Social Network Analysis	3/3/20	Lab: Social Network Analysis
10/3/20	Lecture: Web Content Mining	10/3/20	Lab: Web Content Mining
17/3/20	Lecture: Web Content Mining	17/3/20	Lab: Web Content Mining
24/3/20	Lecture: Web Content Mining	24/3/20	Lab: Web Content Mining
31/3/20	Introduction to Student Projects		-
7/4/20	**Osterferien**		
14/4/20	**Osterferien**		
21/4/20	Feedback about Project Oulines	21/4/20	-
28/4/20	-	28/4/20	-
5/5/20	Coaching session (optional)	5/5/20	-
12/5/20	Coaching session (optional)	12/5/20	-
19/5/20	Coaching session (optional)	19/5/20	-
26/5/20	Presentation of the projects	26/5/20	Presentation of the projects

Lecture Videos

Video recordings of past lectures (outdated but useful as extra complementary materials)

http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/



Software/libraries

- You will not succeed in this module if you do no know how to program (we assume some fluency in Python)…

spaCy

surprise

A Python scikit for recommender systems

Home

Documentation

GitHub page

Star Fork

Maintained by Nicolas Hug

Page built with Jekyll and

fast.ai

Making neural nets uncool again

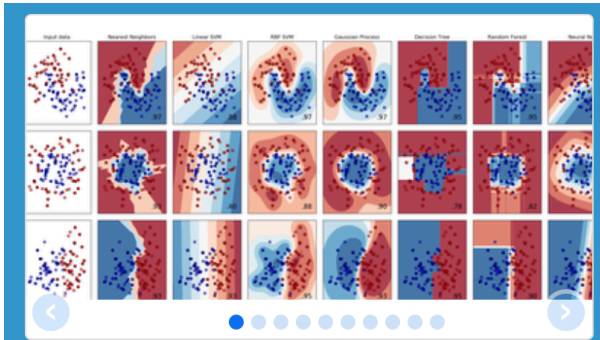
Home

About

Our MOOC

Posts by Topic

© fast.ai 2020. All rights reserved.



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Questions?

2. The World Wide Web

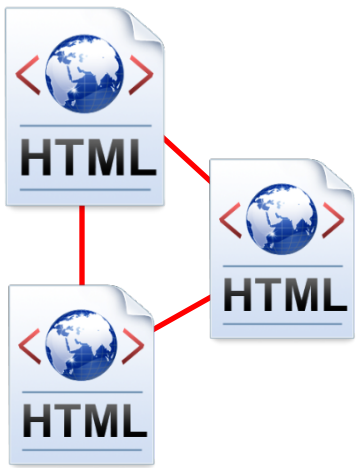
The Web is a global information space build on a set of technical standards for the identification, retrieval and representation of content.

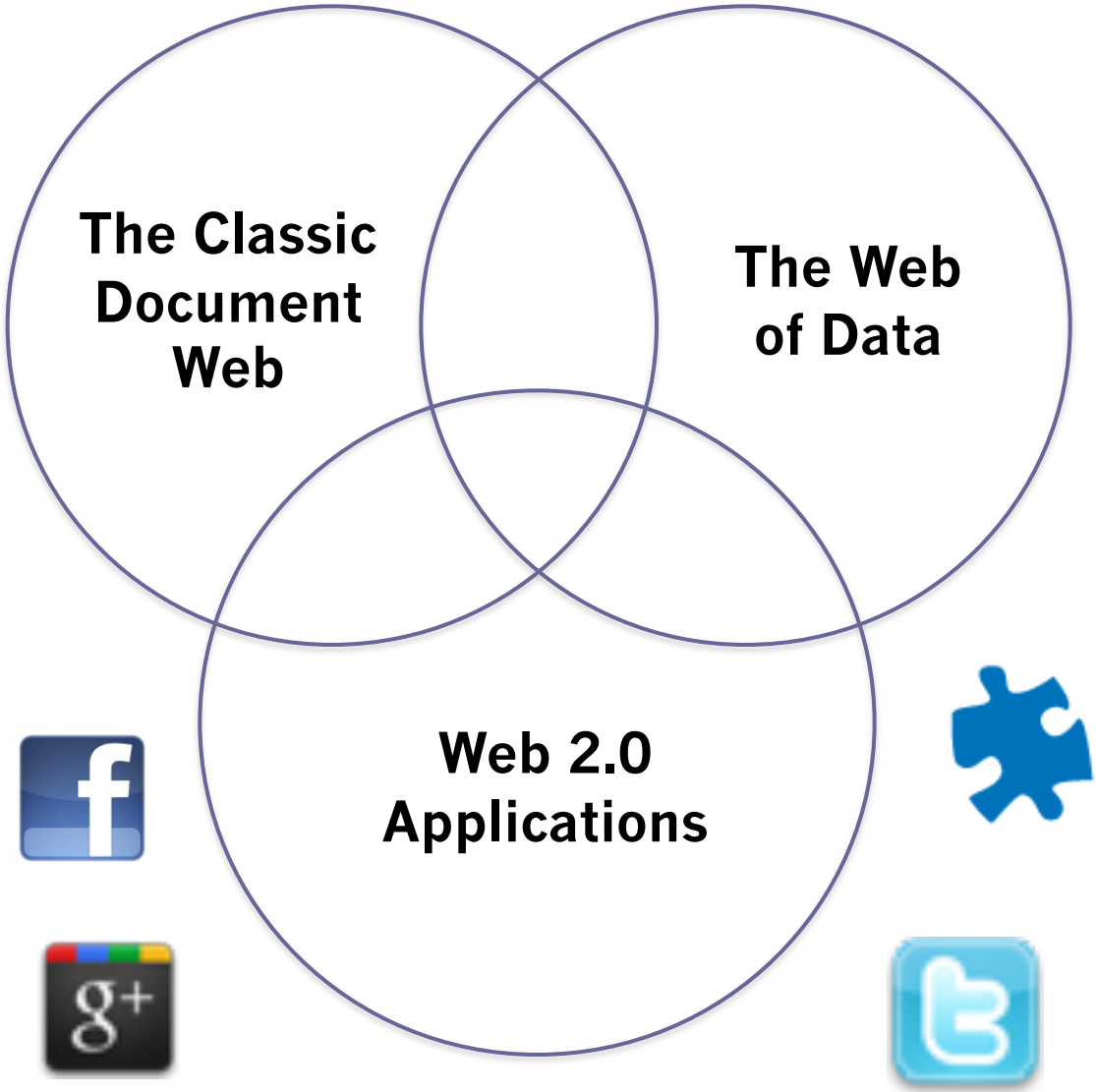
- **Uniform Resource Identifiers (URIs):** Globally unique identification of Web resources.
- **Hypertext Transfer Protokoll (HTTP):** Protocol for interacting with Web resources.
- **Content Formats**: HTML, XML, RDF, ...

- **The Web was invented in 1989 at CERN by Tim Berners-Lee**
- **Architectural Principles of the Web**

Architecture of the World Wide Web, Volume One, W3C Specification, 2004, http://www.w3.org/TR/webarch/

Topology of the Web Today





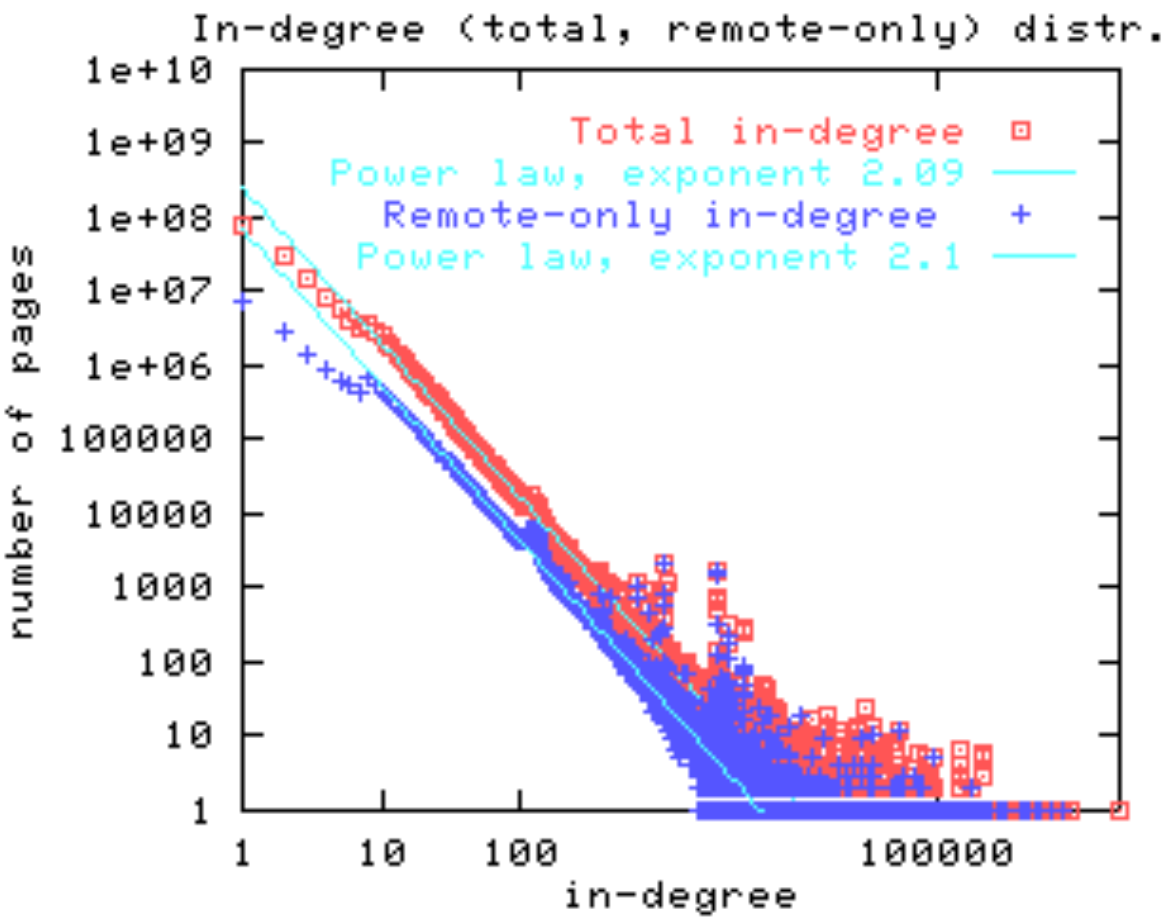
2.1 The Classic Document Web

Global information space consisting of interlinked Web documents (text, images, multimedia).

- Size of the Indexed Web: approx. 60 billion pages
- http://www.worldwidewebsize.com/
  - Estimated on the basis of a method that combines word frequencies from a corpus and search counts returned by the engines
  - 50 words are sent to all four search engines
  - The number of webpages found for these words are recorded
  - Multiple extrapolated estimations are made of the size of the engine's index, which are subsequently averaged
  - Example: The word „the“ appears in 67% of all English pages and has 25.2 billion hits on Google

Link Structure of the Web: In-Degree

- The link distribution follows (kind of) a power law.
  - A small number of pages is target of many links.
  - A large number of pages is target of only a few or no links.
- Classic Paper:
  - Broder at al.: Graph Structure in the Web. WWW2000.
  - AltaVista crawl with over 200 million pages and 1.5 billion links
  - Conclusion: Log-log scale plot shows power-law.

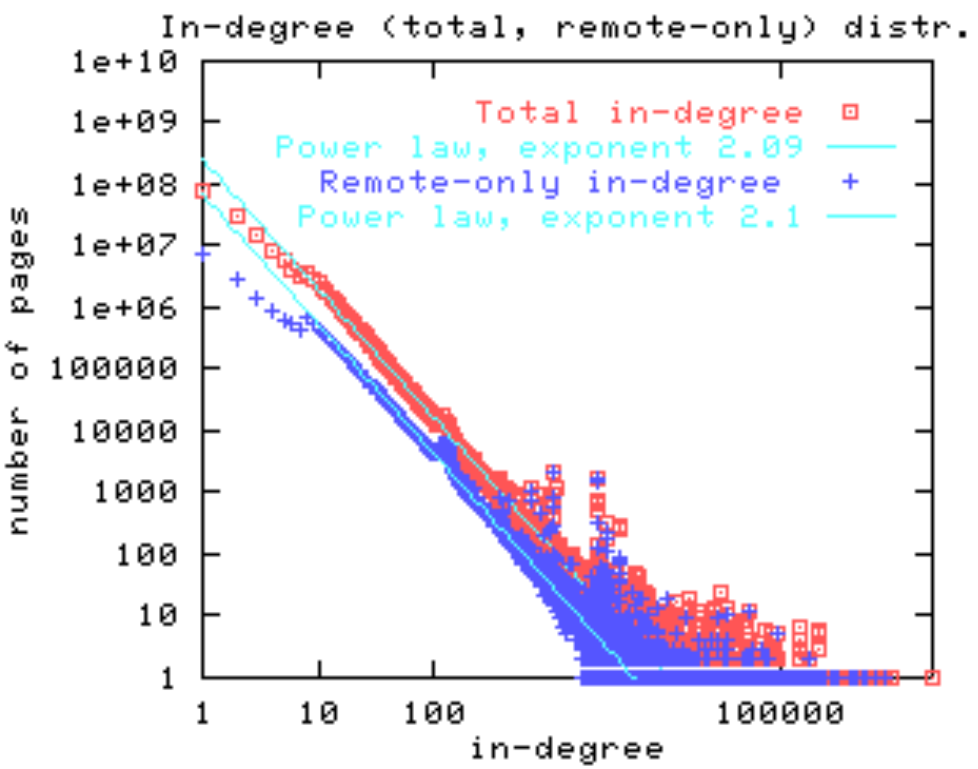


In-Degree Distribution

Broder et al. (2000)

Power law with exponent 2.1

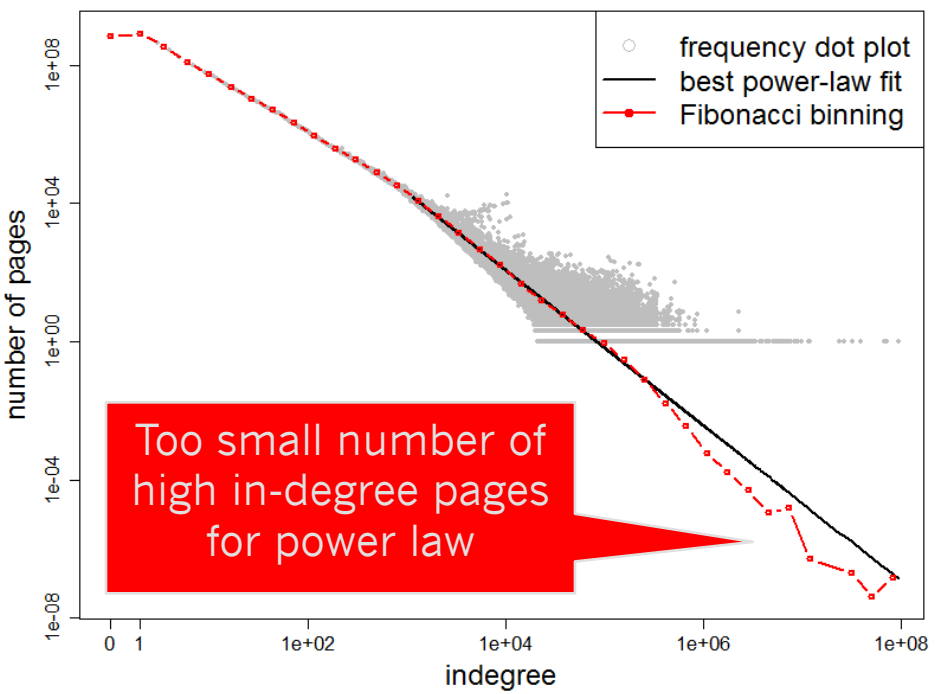
(200 million pages and 1.5 billion links from Altavista crawl 2000)



WDC Hyperlink Graph (2012)

Best power law exponent 2.24

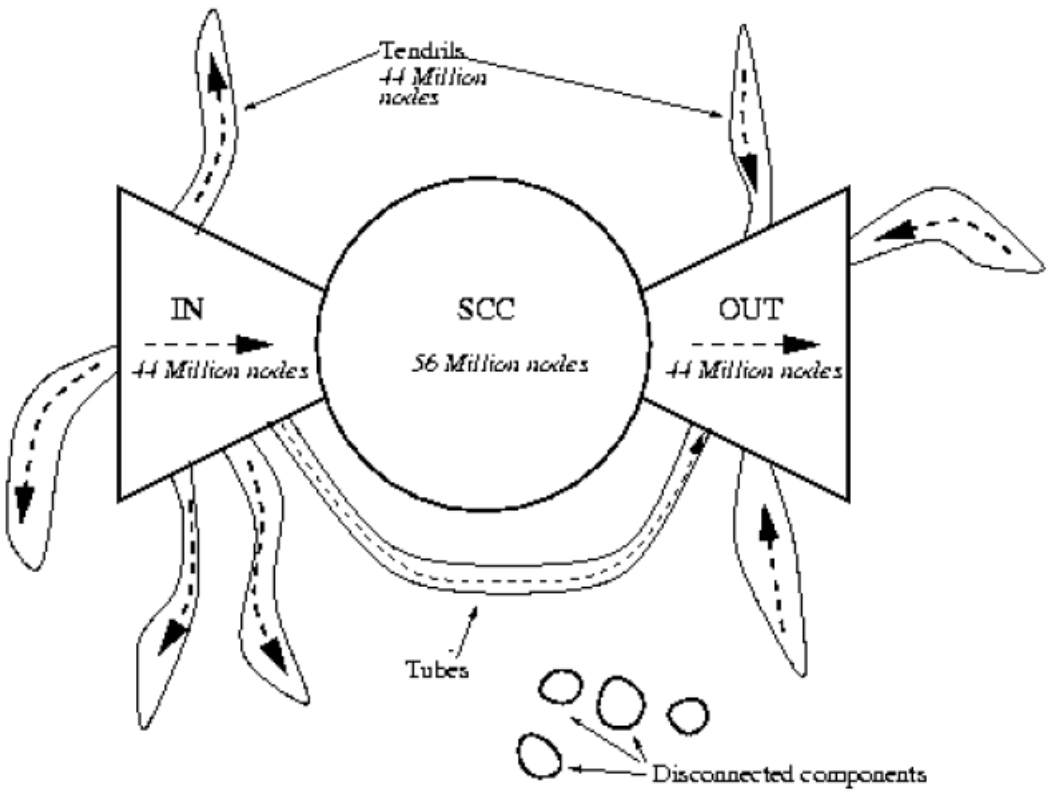
(3 billion pages and 128 billion links from Common Crawl 2012)



Link Structure of the Web: Bow-Tie

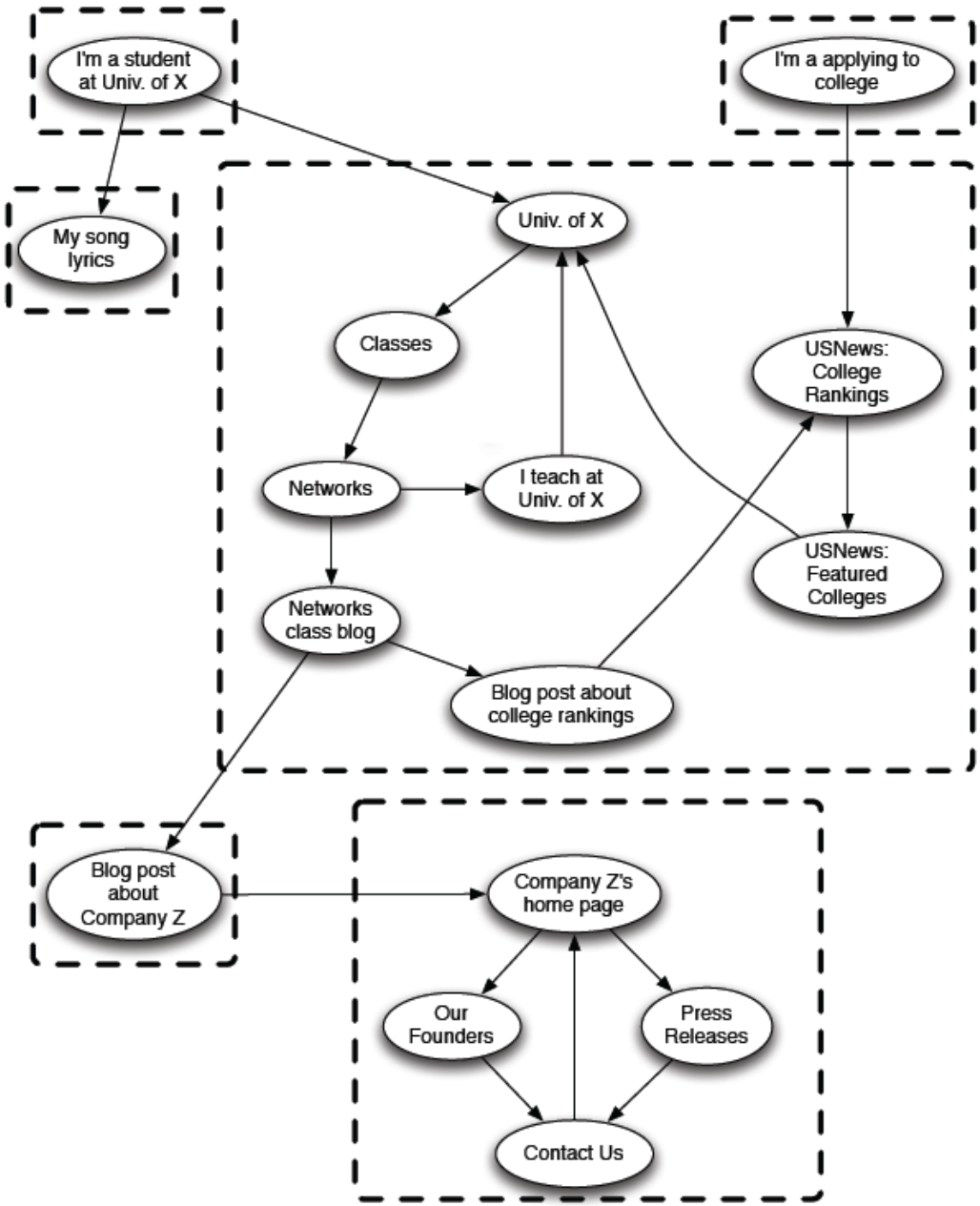
Four mayor components (Border at al., WWW2000)

- **Central Strongly Connected Component (SCC)**
  - pages that can reach one another along directed links
  - about 30% of the Web (normal pages)
- **IN Group**
  - can reach SCC but cannot be reached from it
  - about 20% (maybe new pages or boring ones)
- **OUT Group**
  - can be reached from SCC but cannot reach it
  - about 20% (maybe company pages that don't link)
- **Tendrils**
  - cannot reach SCC and cannot be reached by it
  - about 20%
- **Unconnected**
  - about 10%



Probability of path between nodes is 24%

Web Graph with Strongly Connected Components



A strongly connected component (SCC) in a directed graph is a subset of the nodes such that:

1. **every node in the subset has a path to every other node**
2. **the subset is not part of some larger set with the property that every node can reach every other.**

Largest Strongly Connected Component

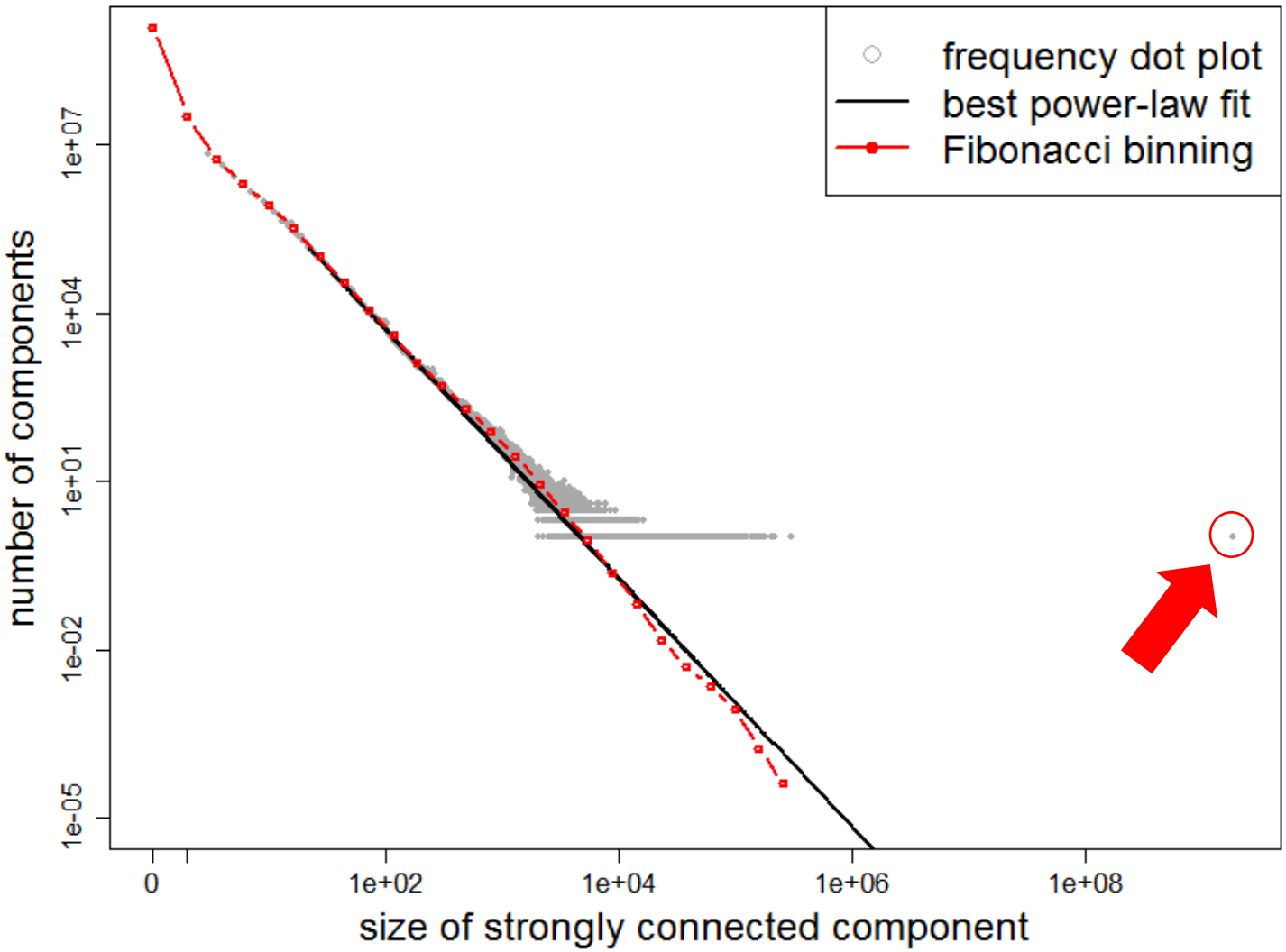
Largest SCC

- Broder, 2000: 27.7%
- WDC, 2012: 51.3 %

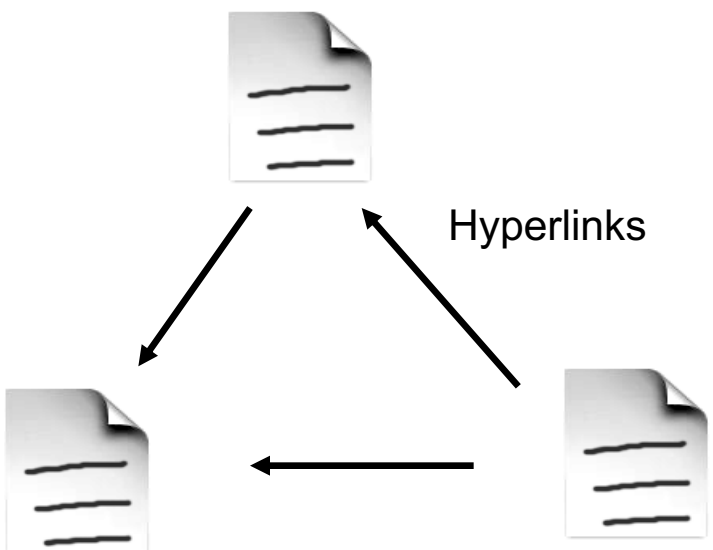
➔ Factor 1.8 larger

➔ Also, factor 4.9 more links/page

➔ The Web has become denser.



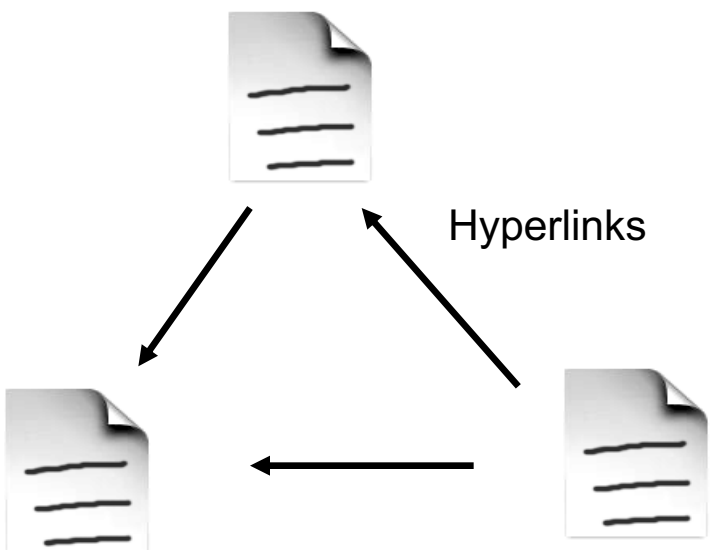
From a Web of Documents to a Web of Data



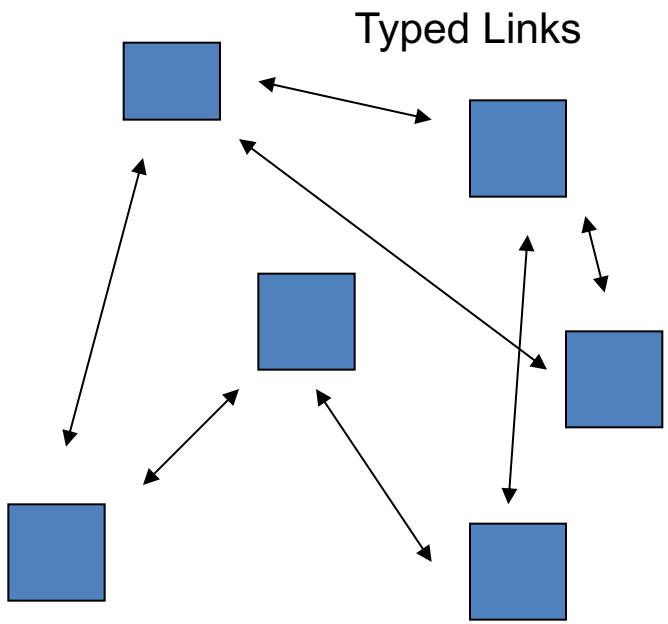
Web of documents

- **Key characteristics:**
  1. **Names (URIs)**
  2. **Documents (Resources) described by HTML, XML, etc.**
  3. **Interactions via HTTP**
  4. **(Hyper)Links between documents or anchors in these documents**

Web of Documents vs. Web of Data



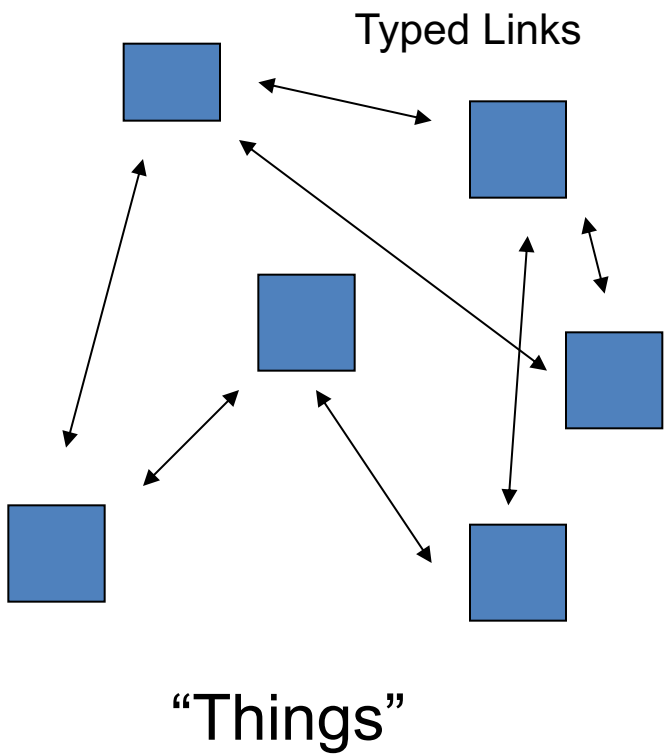
"Documents"



"Things"

Web of data

- **Key characteristics:**
  - **Links between arbitrary things (e.g., persons, locations, events, buildings)**
  - **Structure of data on Web pages is made explicit**
  - **Things described on Web pages are named and get URIs**
  - **Links between things are made explicit and are typed**



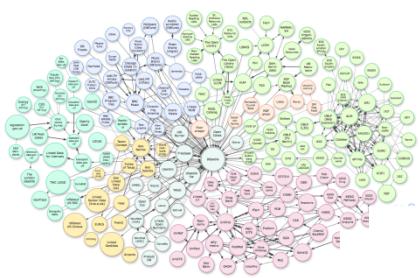
2.2. The Web of Data

Available approaches

1. **semantically markup the content of their HTML pages**
2. **publish structured data in addition to HTML pages**

Microformats

RDFa



Linked Data

Microdata

Microformats

- Microformat effort dates back to 2003
- Small set of fixed formats
  - hcard : people, companies, organizations, and places
  - XFN : relationships between people
  - hCalendar : calendaring and events
  - hListing : small-ads; classifieds
  - hReview : reviews of products, businesses, events
- Key idea
  - use existing HTML attributes to embed structured data types
- Shortcoming
  - Fixed formats means that it can not represent any kind of data

Microformats: example hCard

- **hCard** is a simple, open format for publishing people, companies, organizations on the web, using a 1:1 representation of vCard

```
<div class="vcard">
  <a class="fn org url" href="http://www.commerce.net/">CommerceNet</a>
  <div class="adr">
    <span class="type">Work</span>:
    <div class="street-address">169 University Avenue</div>
    <span class="locality">Palo Alto</span>,
    <abbr class="region" title="California">CA</abbr>  
    <span class="postal-code">94301</span>
    <div class="country-name">USA</div>
  </div>
  <div class="tel">
   <span class="type">Work</span> +1-650-289-4040
  </div>
  <div class="tel">
    <span class="type">Fax</span> +1-650-289-4041
  </div>
  <div>Email:
   <span class="email">info@commerce.net</span>
  </div>
</div>
```

RDFa

- serialization format for embedding RDF data into HTML pages
- proposed in 2004, W3C Recommendation in 2008
- can be used together with any vocabulary

```
1 <html xmlns="http://www.w3.org/1999/xhtml"
2       xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3       xmlns:foaf="http://xmlns.com/foaf/0.1/">
4 ...
5     <div about="http://example.com/Peter" typeof="foaf:Person">
6       <span property="foaf:name">Peter Smith</span> knows
7       <a rel="foaf:knows" href="http://example.com/Paula">Paula
            Jones</a>.
8     </div>
9 ...
```

Microdata

- alternative technique for embedding structured data
- proposed in 2009 by WHATWG as part of HTML5 work
- tries to be simpler than RDFa (5 new attributes instead of 8)

```
1 <div itemscope itemtype="http://schema.org/Person" itemid="http
      ://example.com/Peter">
2   <span itemprop="name">Peter Smith</span>
3   <a href="http://example.com/Paula" itemprop="knows">Paula Jones
        </a>
4 </div>
```

Schema.org

- Problem: which vocabulary to use?
- Schema.org provides a collection of shared vocabularies
- Launched in June 2011 by Bing, Google and Yahoo

- Create a common set of schemas for webmasters to mark-up with structured data their websites.

- 200+ Types: Event, Organization, Person, Place, Product, Review
- Encoding: Microdata, RDFa, JSON-LD

Schema.org: example schema

schema.org

Custom Search

Home | Schemas | Documentation

Person

Thing > Person

A person (alive, dead, undead, or fictional).

[more...]

Property	Expected Type	Description
Properties from Person		
additionalName	Text	An additional name for a Person, can be used for a middle name.
address	PostalAddress or Text	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
alumniOf	EducationalOrganization or Organization	An organization that the person is an alumni of. Inverse property: alumni.
award	Text	An award won by or for this item. Supersedes awards.
birthDate	Date	Date of birth.
birthPlace	Place	The place where the person was born.

Schema.org: examples

RDFa

```
<div vocab="http://schema.org/" typeof="Movie">
  <h1 property="name">Avatar</h1>
  <div property="director" typeof="Person">
  Director: <span property="name">James Cameron</span>
(born <time property="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span property="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" property="trailer">Trailer</a>
</div>
```

Microdata

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span>
(born <time itemprop="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

Usage of Schema.org Data @ Google

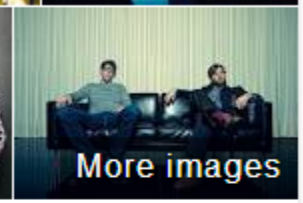
Gramercy Tavern - Flatiron - New York, NY | Yelp
www.yelp.com › Restaurants › American (New)
★★★★½ Rating: 4.5 - 1,288 reviews - Price range: $$$$
Jeff C and I were in **New York** for vacation, and I wanted to treat him to a nice dinner for ..... **Gramercy Tavern** is certainly a legendary **NY** dining establishment.

Gramercy Tavern Restaurant - New York, NY | OpenTable
www.opentable.com › ... › Gramercy restaurants
★★★★½ Rating: 4.7 - 508 reviews - Price range: $50 and over
Book now at **Gramercy Tavern** in **New York**, explore menu, see photos and read 508 reviews: "The menu was so limited but it was worth trying, food was deli..."

Data snippets within search results

Data snippets within info boxes



The Black Keys

Band

The Black Keys is an American rock duo formed in Akron, Ohio in 2001. The group consists of Dan Auerbach and Patrick Carney. Wikipedia

Origin: Akron, Ohio, United States

Members: Dan Auerbach, Patrick Carney

Record labels: Fat Possum Records, Nonesuch Records, V2 Records, Alive Naturalsound Records

Awards: Grammy Award for Best Rock Album, more

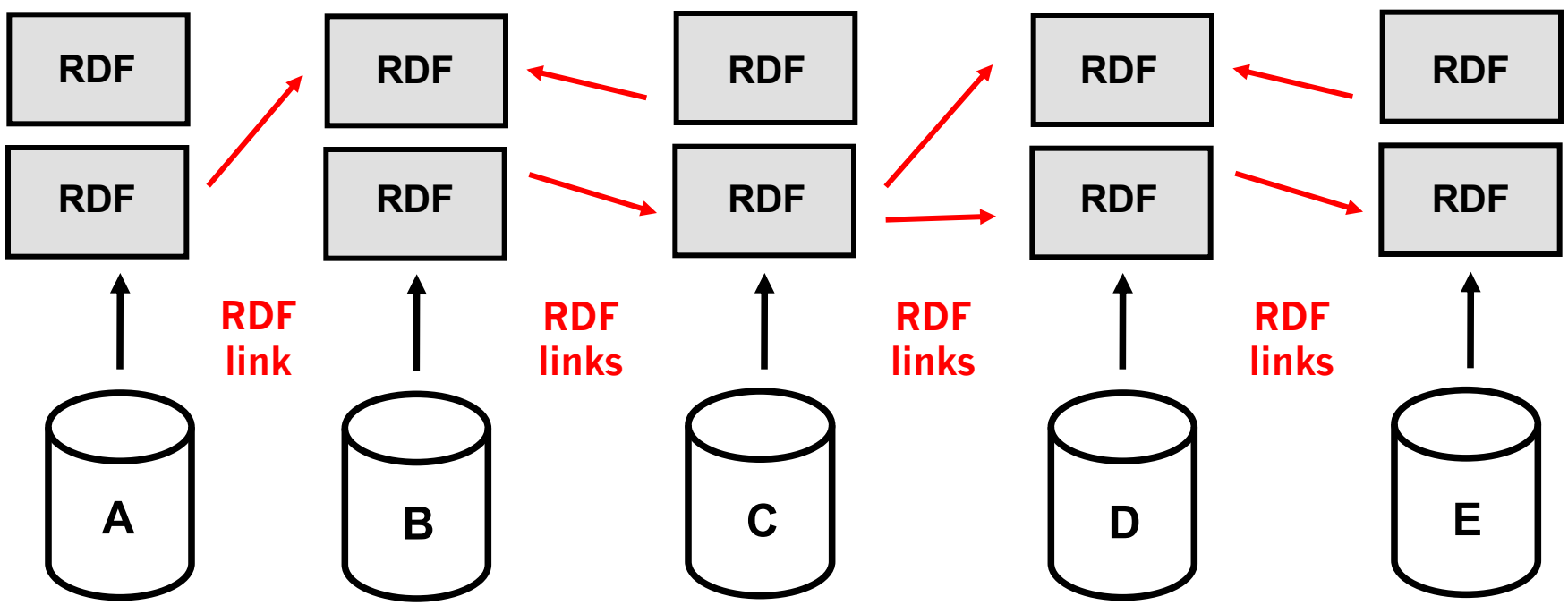
Upcoming events

Jun 20 Fri	The Black Keys Neuhausen ob Eck (near you)
May 16 Fri	The Black Keys Gulf Shores, AL
Jun 22 Sun	The Black Keys Scheeßel

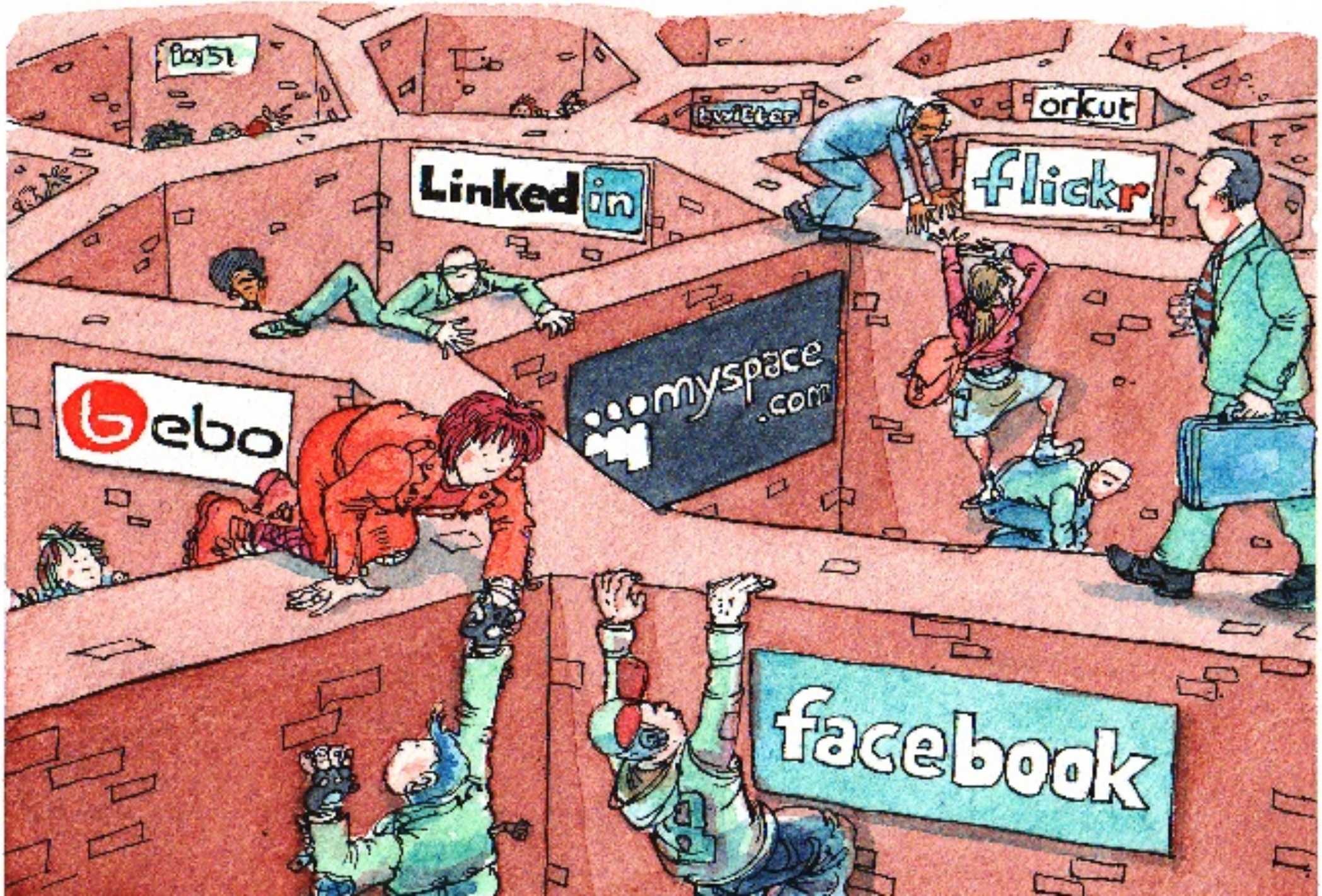
Alternative Approach: Linked Data

Extend the Web with a single global data graph

1. by using RDF to publish structured data on the Web
2. by setting links between data items within different data sources.

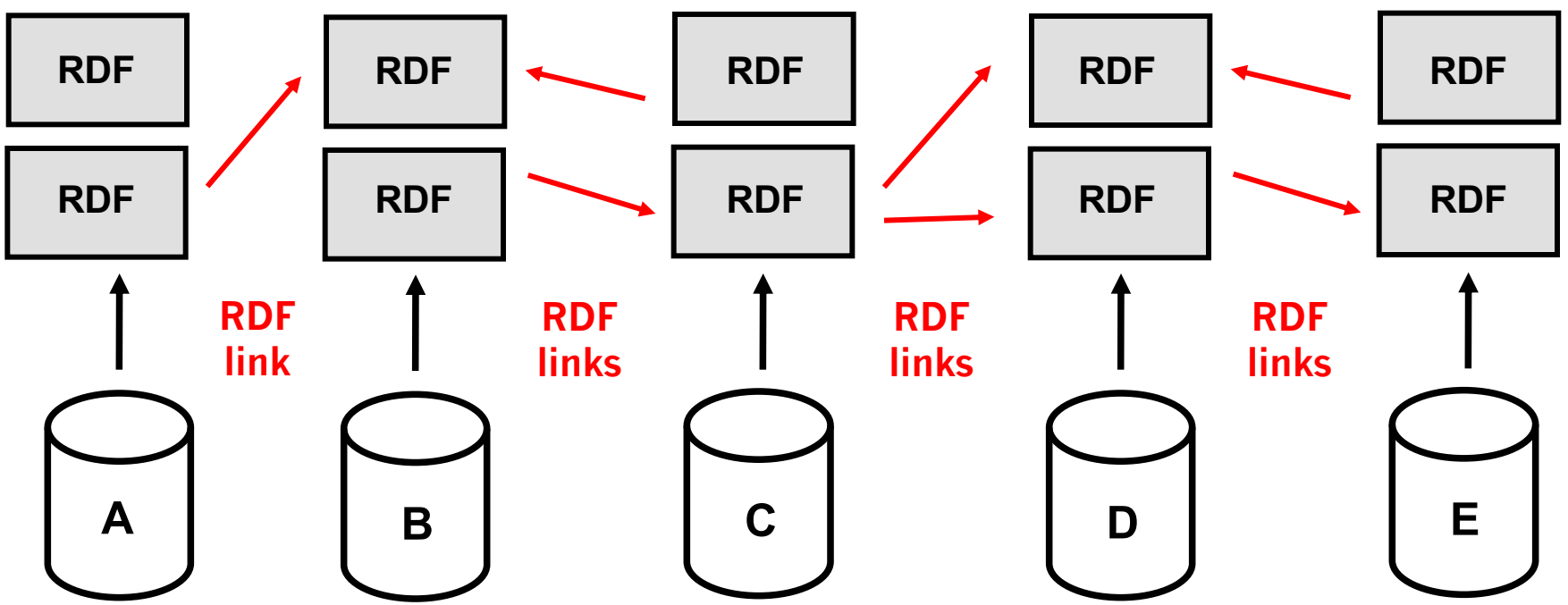


The Web Today as a Multitude of Data Silos

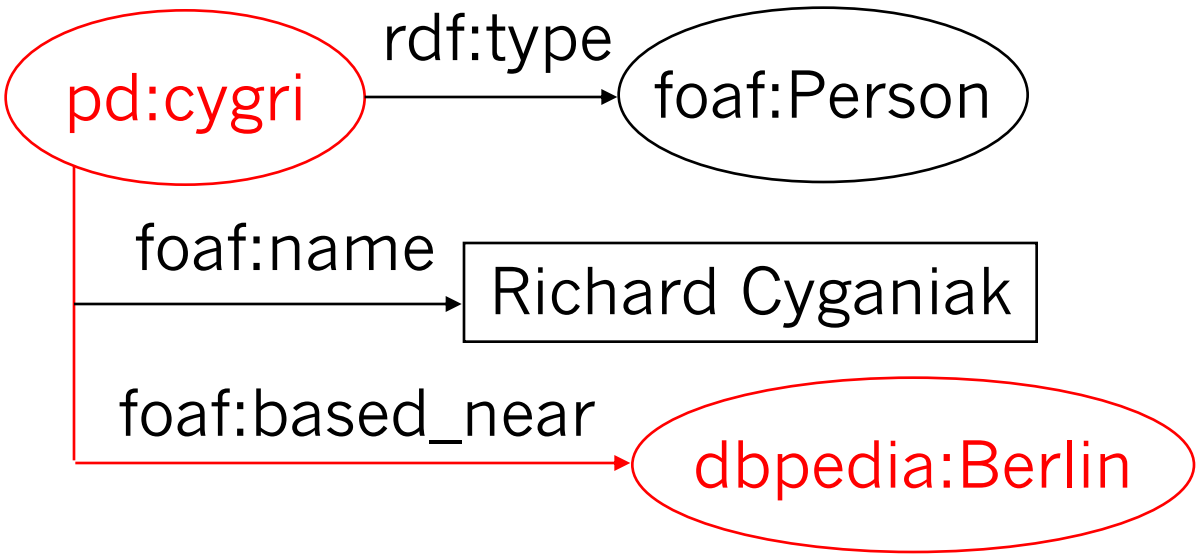


From Data Silos vs. the Web of Data

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful RDF information.
4. Include RDF statements that link to other URIs so that they can discover related things.



Entities are identified with HTTP URIs

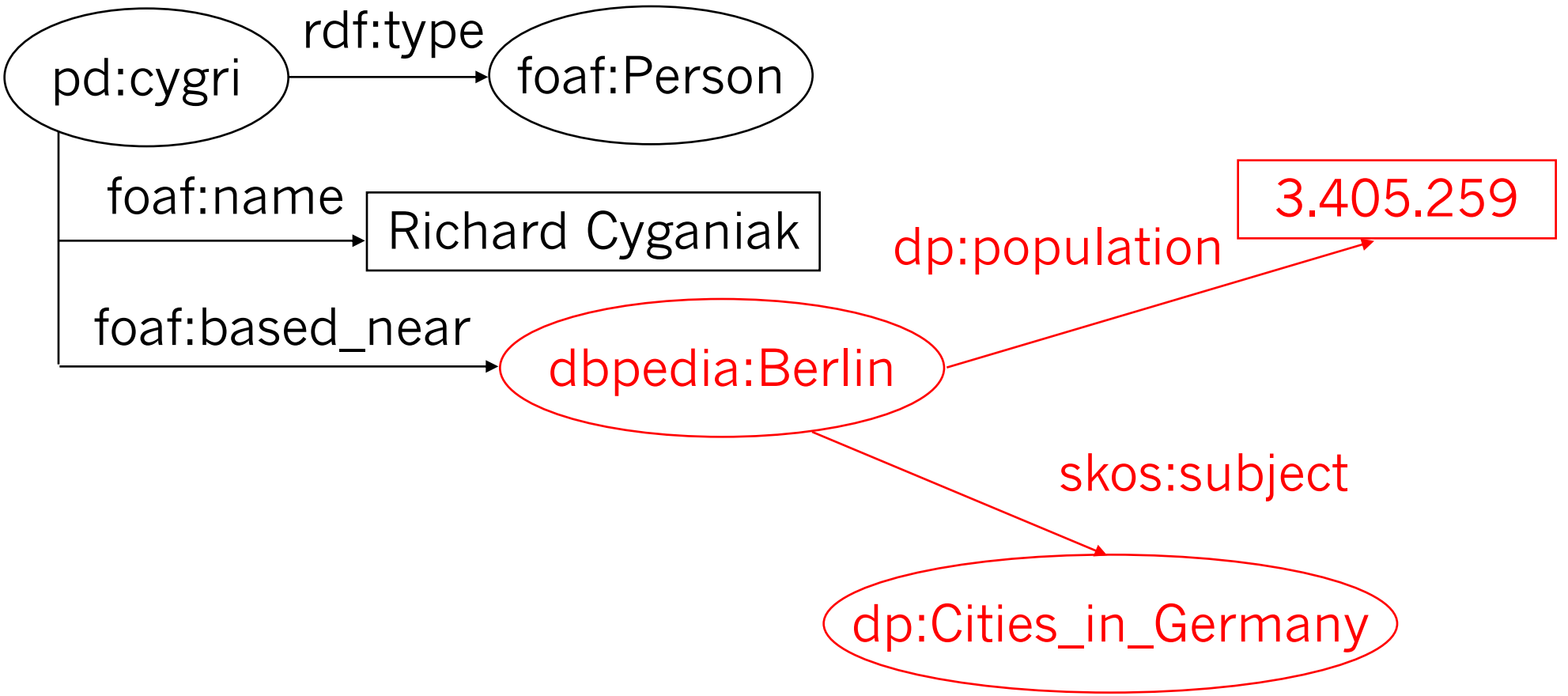


HTTP URIs take the role of global primary keys.

pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri

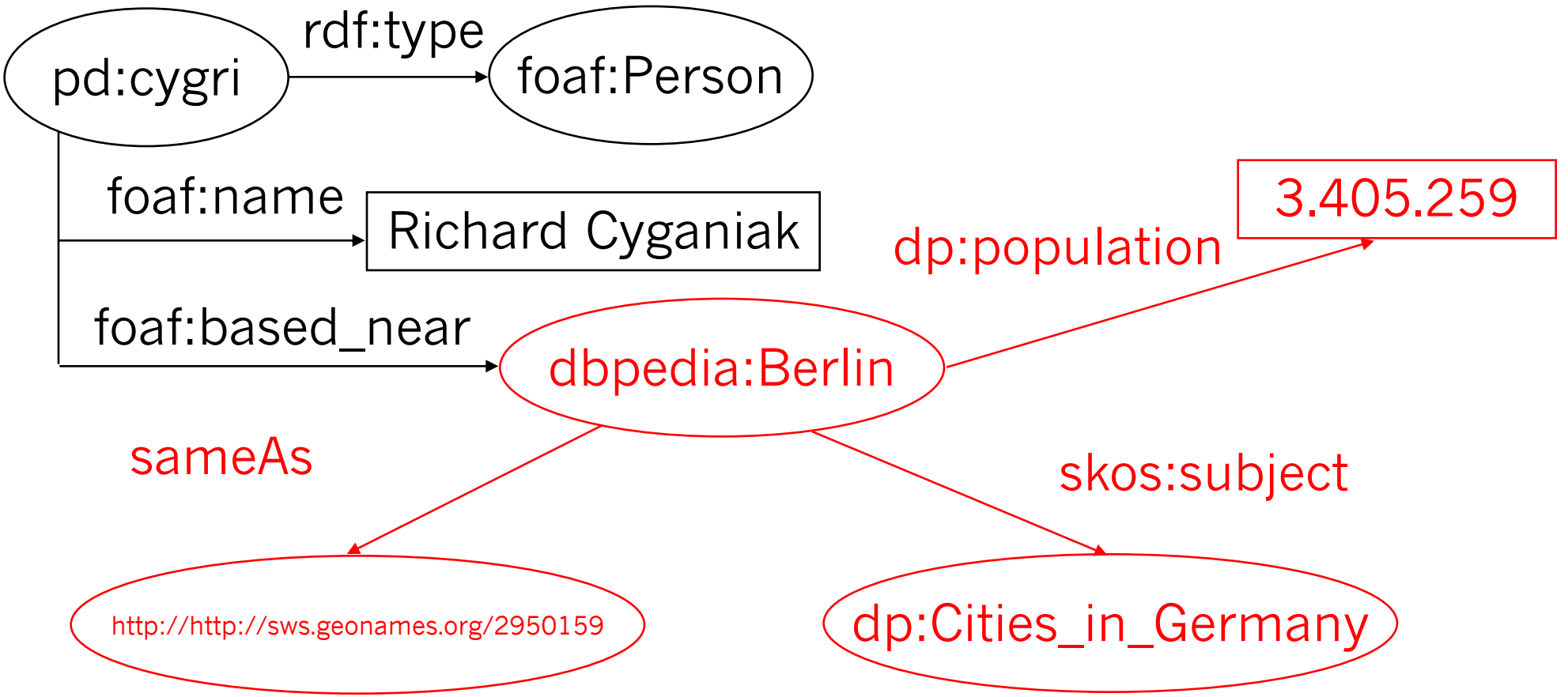
dbpedia:Berlin = http://dbpedia.org/resource/Berlin

URIs can be looked up on the Web



- **By following RDF links applications can**
  - **navigate the global data graph**

URIs can be linked to navigate the data space



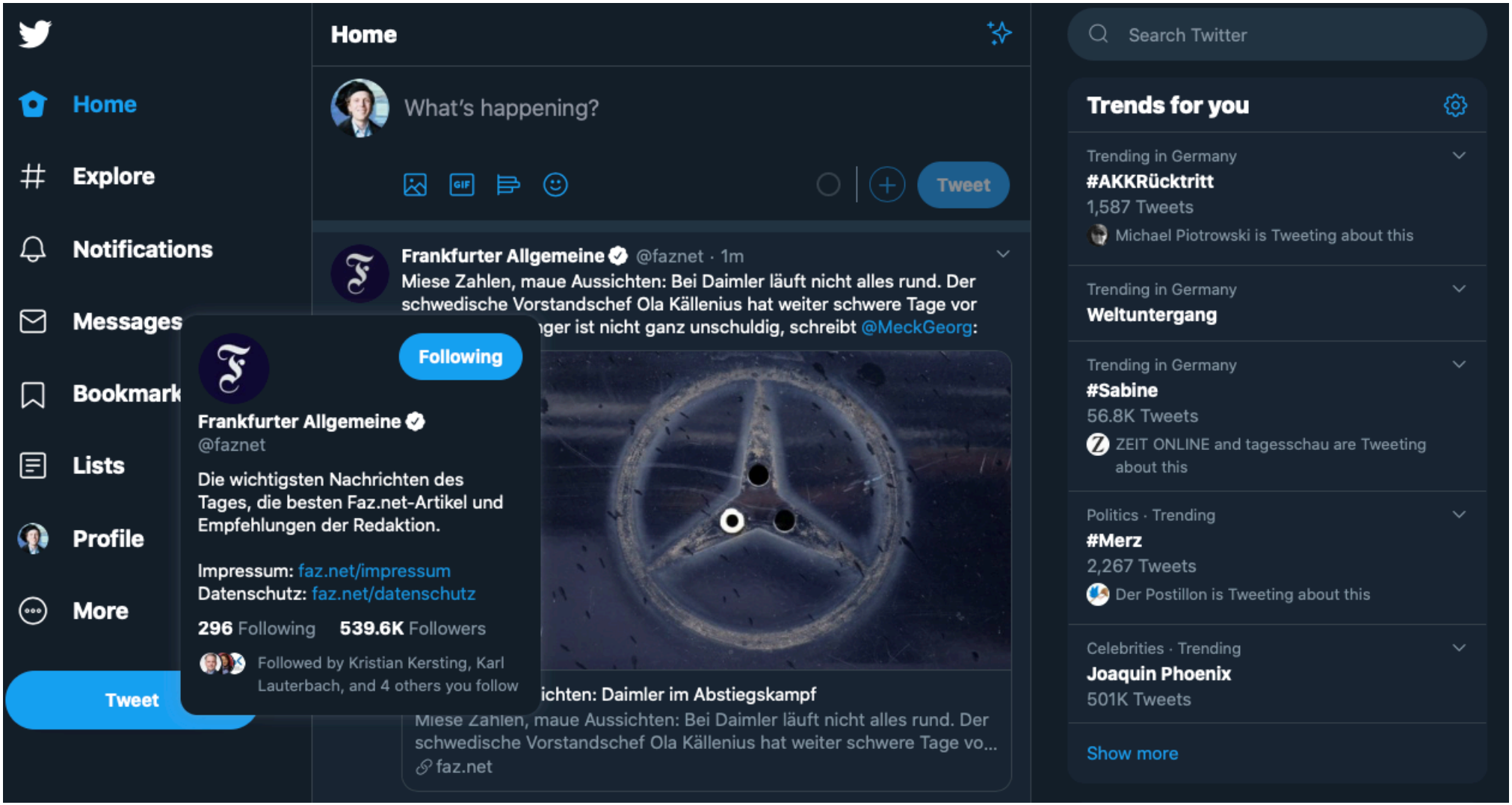
- **By following RDF links applications can**
  - **navigate the global data graph**
  - **discover new data sources**

2.3. Web 2.0 Applications

- A multitude of Web-based applications has sprung up which enable users to share information.
- These applications
  - collect large amounts of data using proprietary schemata.
  - form separate data spaces that are only partly accessible from the Web via:
    1. HTML interfaces
    2. Web APIs

HTML Interfaces

- Allows browser-based access to profile, communication, etc.



Web APIs

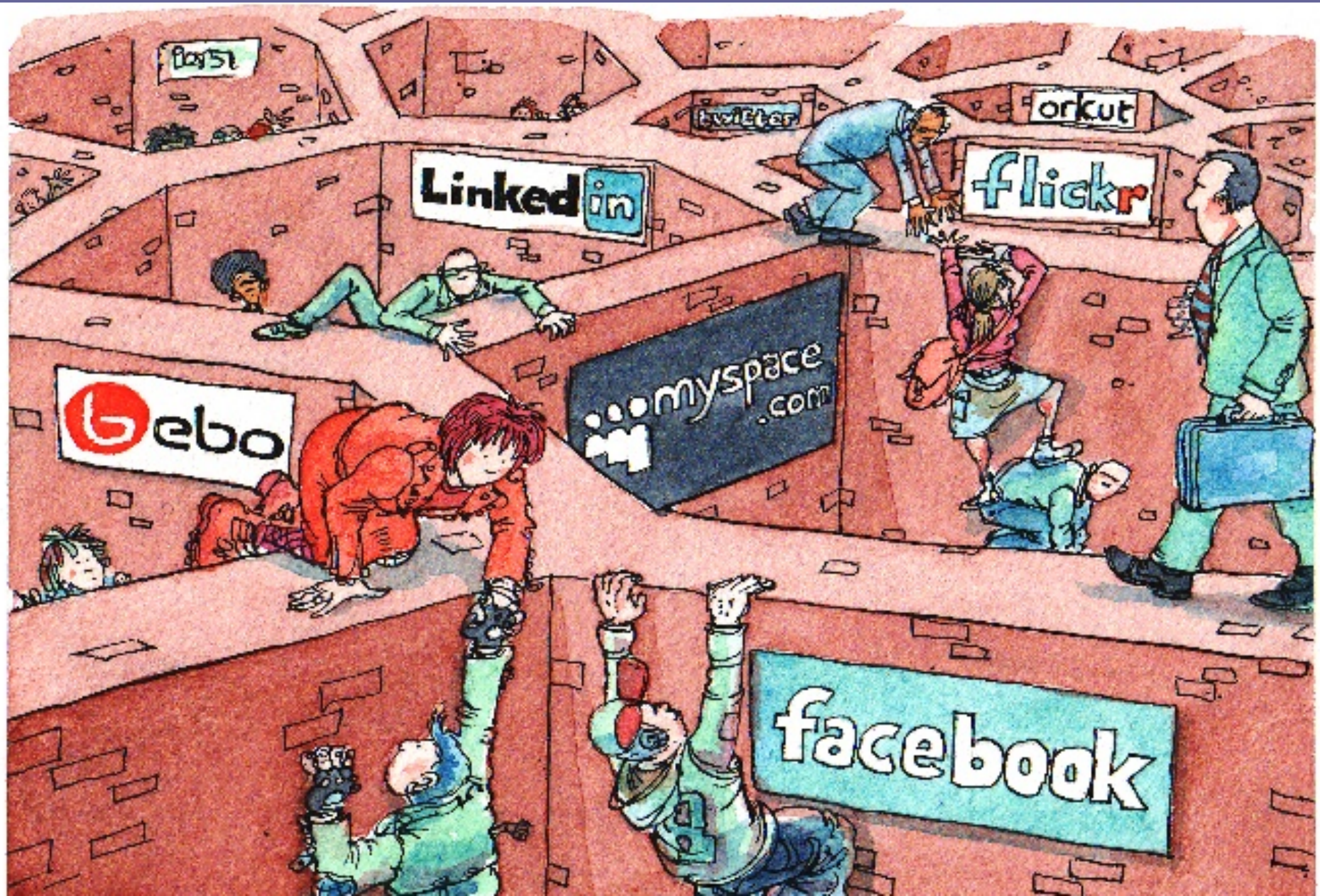
- **Access the data programmatically**

```
import pandas as pd

# Search tweets
dict_ = {'user': [], 'date': [], 'text': [], 'favorite_count': []}
for status in python_tweets.search(**query)['statuses']:
    dict_['user'].append(status['user']['screen_name'])
    dict_['date'].append(status['created_at'])
    dict_['text'].append(status['text'])
    dict_['favorite_count'].append(status['favorite_count'])

# Structure data in a pandas DataFrame for easier manipulation
df = pd.DataFrame(dict_)
df.sort_values(by='favorite_count', inplace=True, ascending=False)
df.head(5)
```

Web APIs slice the Web into Data Silos



3. What is Web Mining?

- **Definition**

Acquiring useful information from

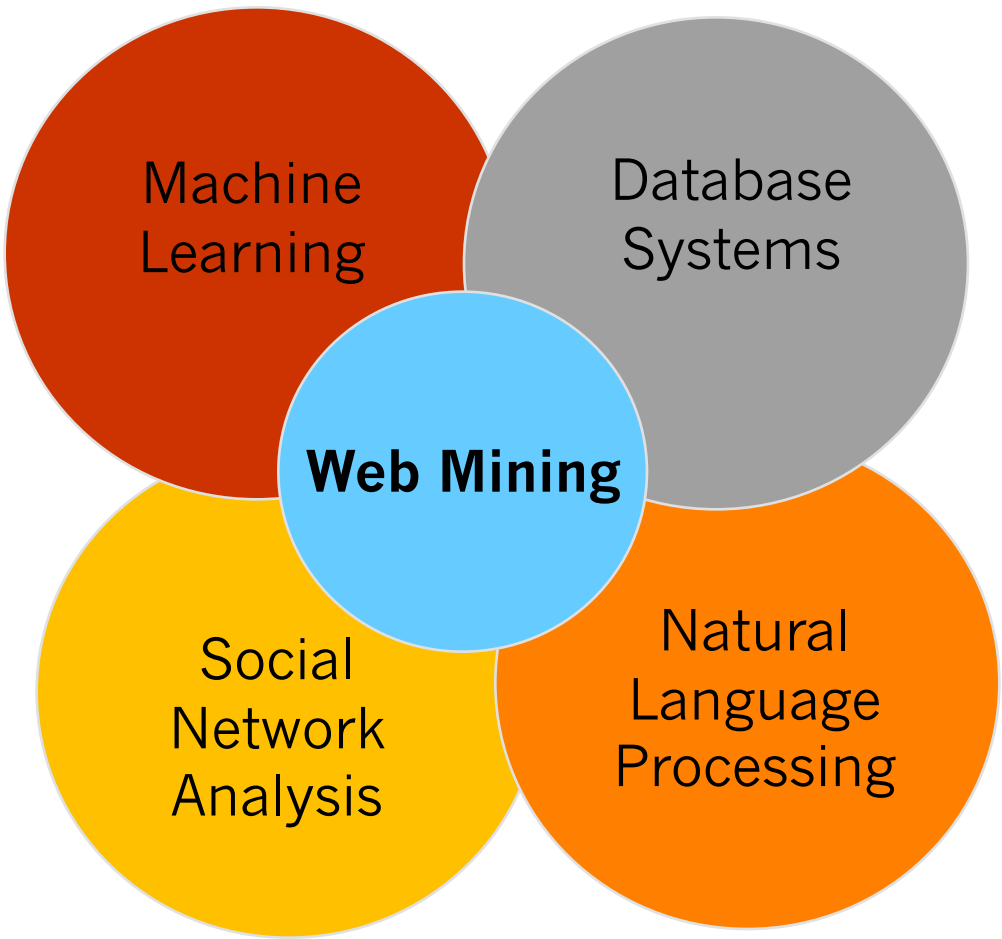
- **Web *content***
- **Web *structure***
- **Web *usage* data**

- **Web data pose unique challenges:**

1. **Large volume**
2. **Semi-structured**
3. **Heterogeneous**
4. **Distributed**

Web Mining is a Multi-Disciplinary Field

- Draws ideas and techniques from



- Sub-Fields
  1. Web Usage Mining
  2. Web Structure Mining
  3. Web Content Mining

3.1 Web Usage Mining

- **Definition**

Discovery of patterns in data collected or generated as a result of user interactions with one or more web sites.

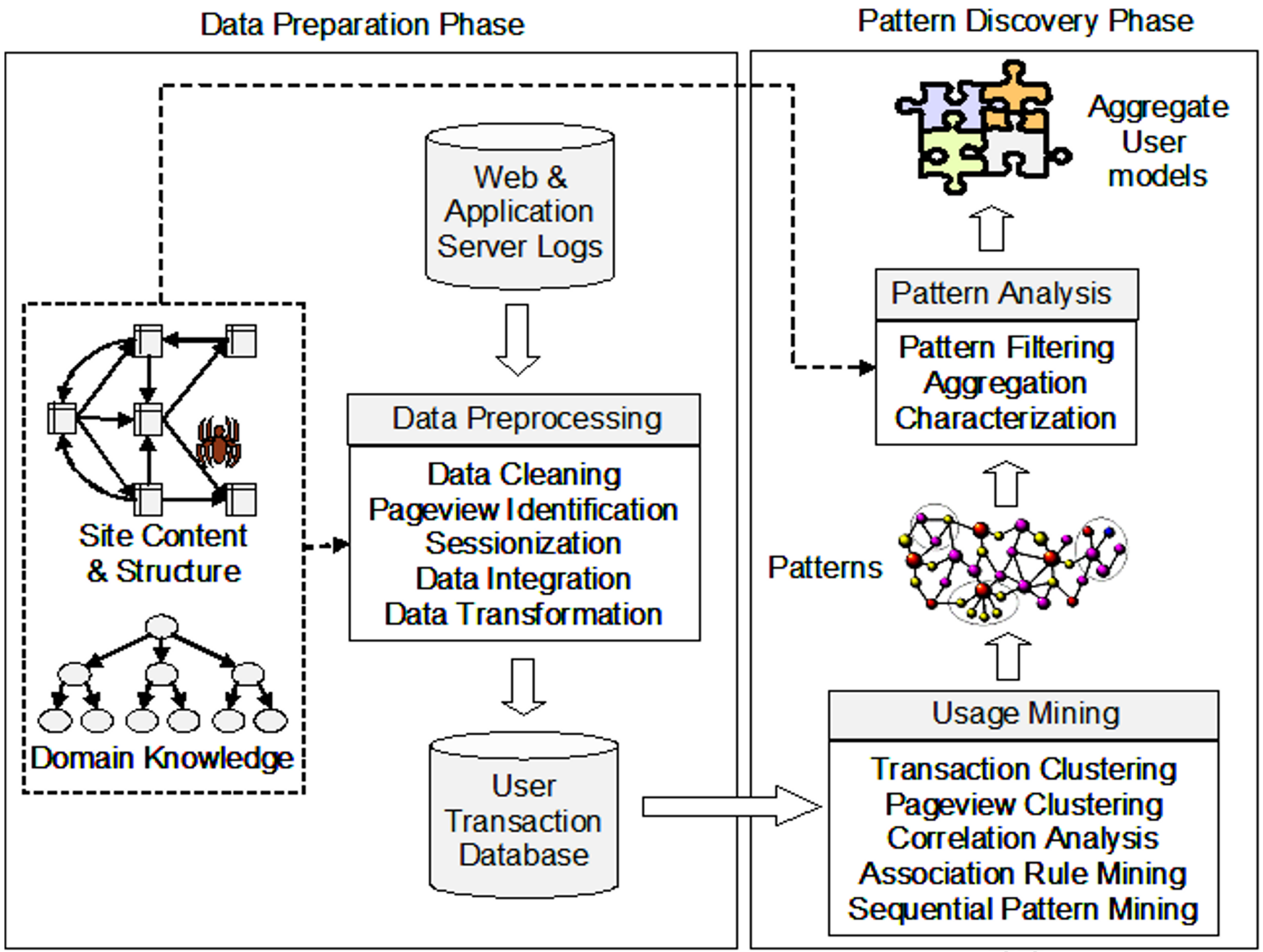
- **Sources of Data**
  1. automatically generated data stored in server access logs
  2. e-commerce and product-oriented user events (e.g., shopping cart changes, ad or product click-throughs, purchases)
  3. user profiles (e.g. Facebook) and/or user ratings (likes)
  4. page attributes, page content, site structure
  5. additional domain knowledge and demographic data

Leading Usage Data Collections

Enable the

- analysis of the current interests and behavior of the world's population.
- identification of suspected terrorists.

The Web Usage Mining Process



Example: product recommendation

amazon.de prime | Alle | DE | Hallo, Simone Konto und Listen | Warenrücksendungen und Bestellungen | Mein Prime | 0 Einkaufswagen

Lieferung an Simone 69115 Heidelberg | Erneut kaufen | Simones Amazon | Angebote | Gutscheine | Verkaufen | Hilfe | Küche, Haushalt & Wohnen | Prime Video | Filme & Serien | Jetzt ansehen

Mein Amazon | Ihre besuchten Seiten | Ihre Empfehlungen | Verbessern Sie Ihre Empfehlungen | Gutscheine | Mein öffentliches Profil | Mehr dazu

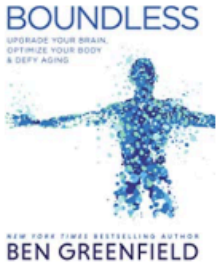
Für Sie empfohlen

Simones Amazon ›

Bücher

Warum empfohlen?

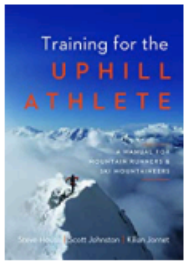
Artikel entfernen



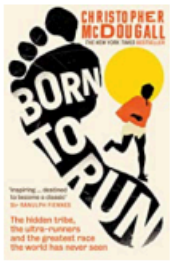
Boundless: Upgrade Your Brain, Optimize Your Body & Defy Aging
Ben Greenfield
4
66,99 € prime
Ähnliches | In den Einkaufswagen

Leadership Strategy and Tactics: Field Manual
Jocko Willink
5
18,99 € prime
Ähnliches | In den Einkaufswagen

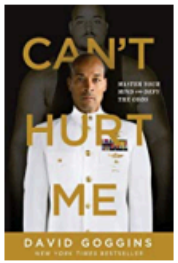
Running Rewired: Reinvent Your Run for Stability, Strength, and…
Jay Dicharry
6
23,59 € prime
Ähnliches | In den Einkaufswagen



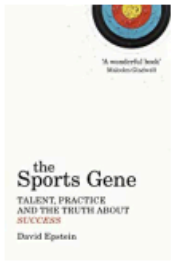
Training for the Uphill Athlete: A Manual for Mountain Runners and…
Steve House
8
22,99 € prime
Ähnliches | In den Einkaufswagen



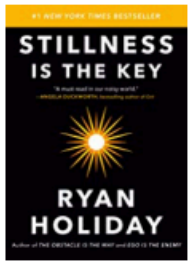
Born to Run: The Hidden Tribe, the Ultra-Runners, and the Greatest…
Christopher McDougall
64
10,29 € prime
Ähnliches | In den Einkaufswagen



Can't Hurt Me: Master Your Mind and Defy the Odds
David Goggins

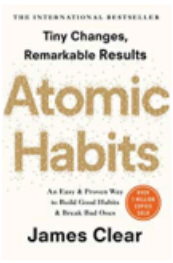


The Sports Gene: Talent, Practice and the Truth About Success
David Epstein



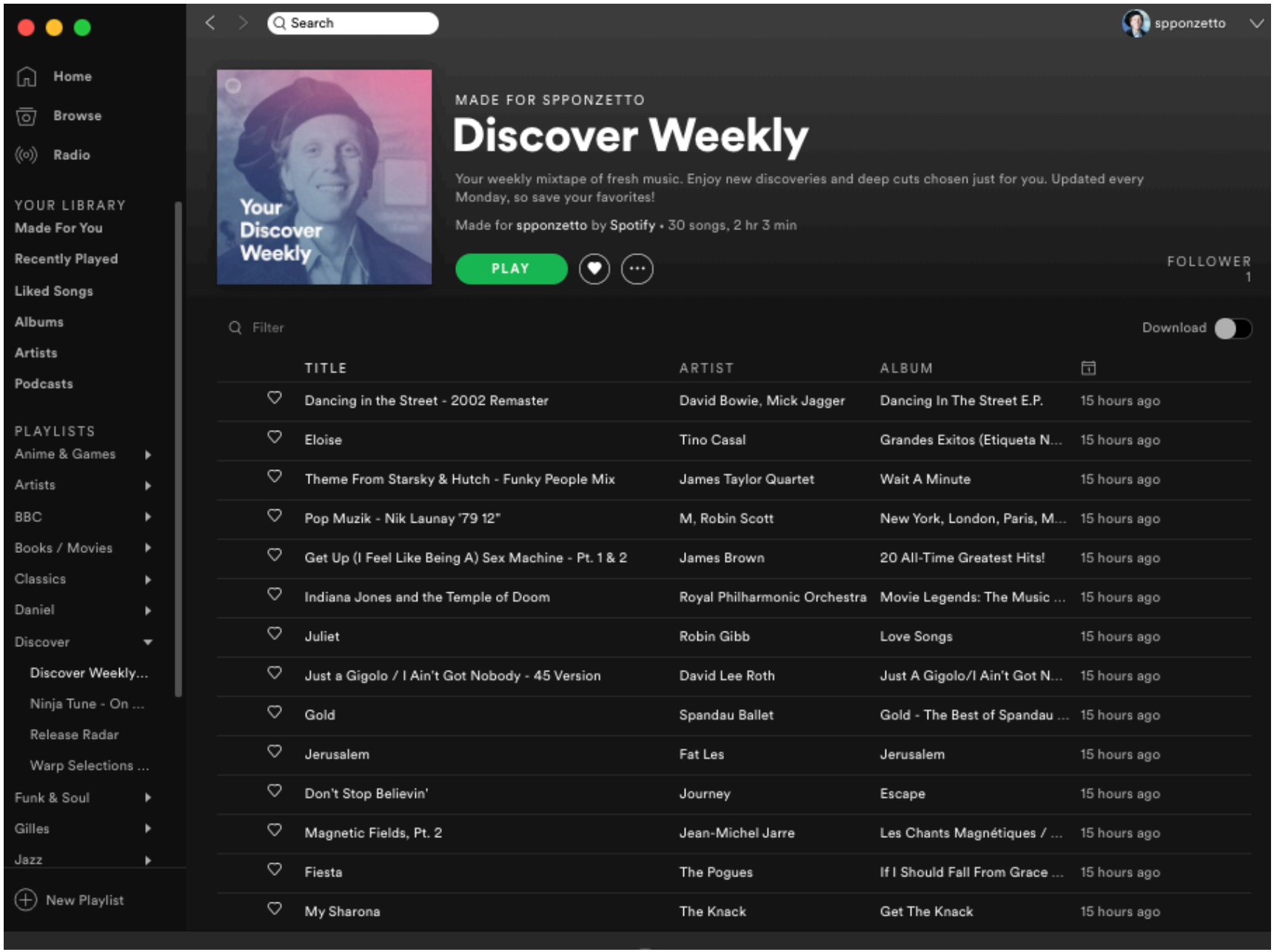
Stillness Is the Key
Ryan Holiday
31

Das große Kettlebell-Trainingsbuch
Dr. Till Sukopp
60



Atomic Habits: The life-changing million copy bestseller
James Clear

Example: product recommendation



Example: personalized search

Google kafka

All Images Videos Maps News More Settings Tools

About 51.700.000 results (0,58 seconds)

Ad · www.confluent.io/

Confluent | Download Apache Kafka® Today | confluent.io

Monitor & Manage Data in Real-Time Using Our Scalable, Reliable, & Flexible Platform. Confluent Platform Includes the Latest Version of Apache **Kafka**, Plus Enterprise Features. Reduce Ops Burden. Streaming Data Service. Deployable in Minutes. 24/7 Support.

What is Kafka?
More Info. About Kafka & Confluent Platform.

Kafka Definitive Guide
Learn All About Kafka From its Original Developers in this eBook.

kafka.apache.org

Apache Kafka

Kafka® is used for building real-time data pipelines and streaming apps. It is horizontally scalable, fault-tolerant, wicked fast, and runs in production in thousands ...
Introduction · Kafka Streams · Documentation · Apache Kafka Documentation

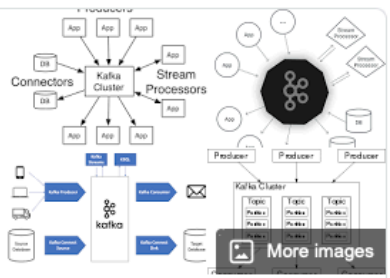
kafka.apache.org › intro

Introduction - Apache Kafka - Apache Software

In **Kafka** the communication between the clients and the servers is done with a simple, high-performance, language agnostic TCP protocol. This protocol is ...

People also ask

- What is Kafka used for?
- What is meant by Kafka?
- What is Kafka and how it works?
- Is Kafka free?



Apache Kafka

Software

Apache Kafka is an open-source stream-processing software platform developed by LinkedIn and donated to the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. Wikipedia

License: Apache License 2.0

Developer(s): Apache Software Foundation

Initial release date: January 2011

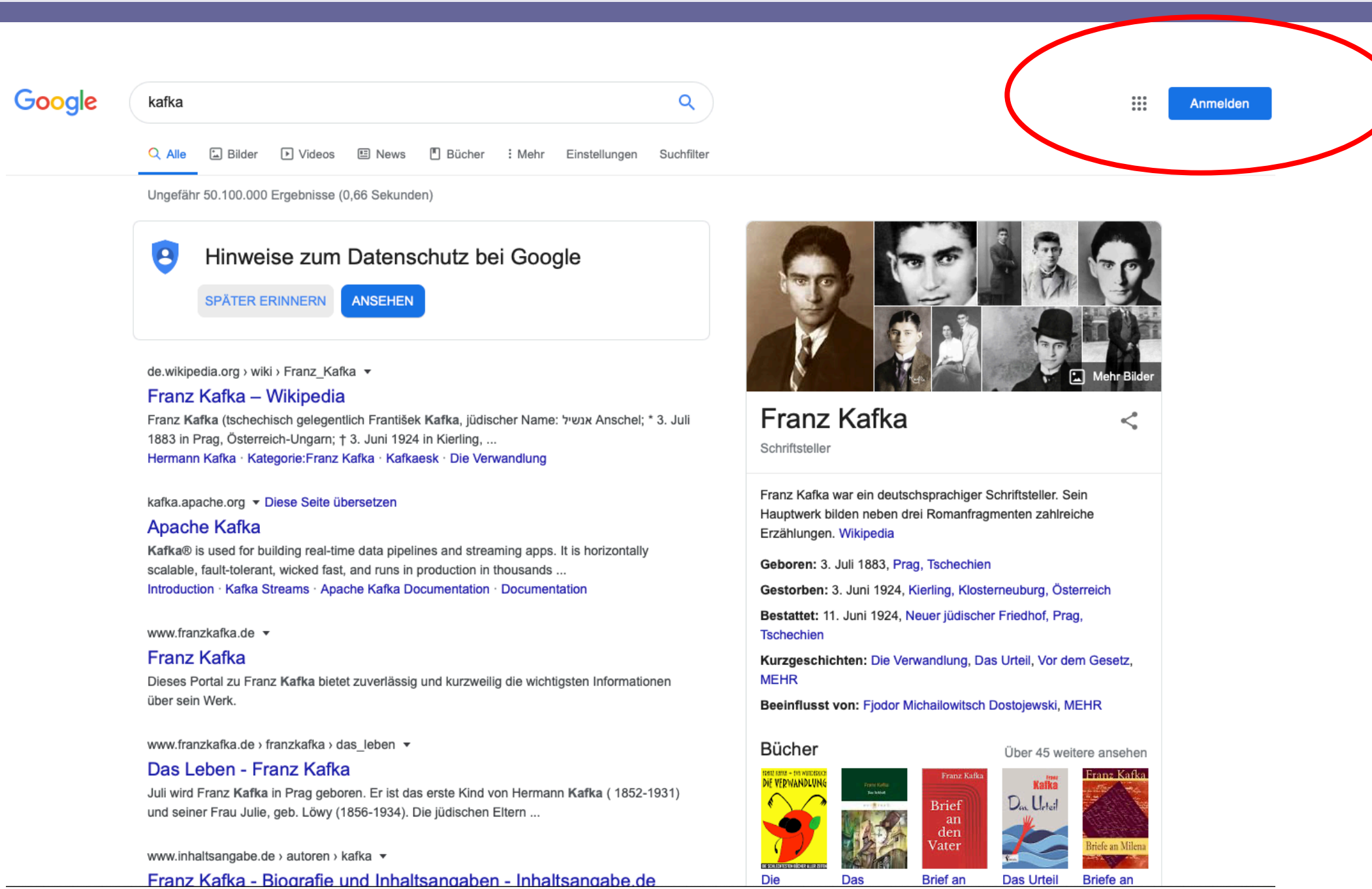
Stable release: 2.4.0 / December 16, 2019; 48 days ago

Written in: Scala, Java

People also search for View 15+ more

Apache Spark · RabbitMQ · Apache ZooKeeper · Apache Cassandra · Apache Hadoop

Example: personalized search



3.2 Web Structure Mining

- Definition

Discovery of patterns in

- **the hyperlink structure of webpages**
- **the structure of communities that interact on the Web**

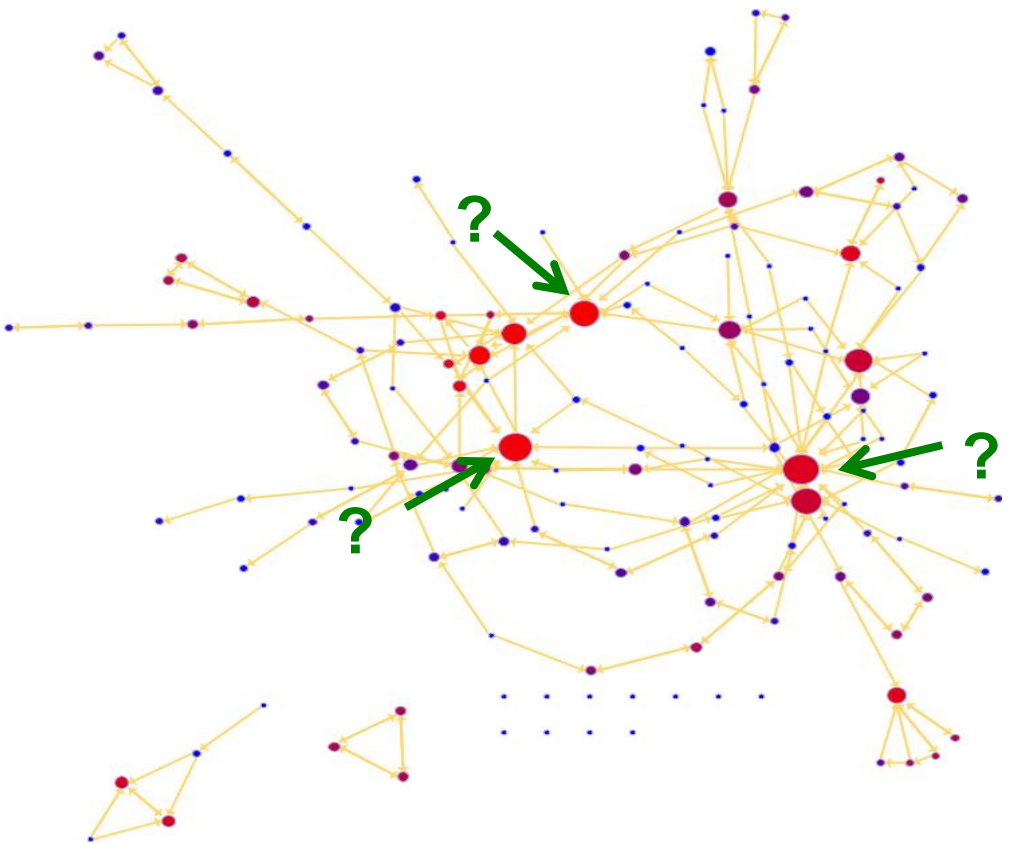
- Exploits the graph structure, but can of course also be combined with content or usage mining techniques.
- Typical Sources of Data
  1. Web crawls including HTML pages and hyperlinks
  2. crawls of the blogosphere
  3. social networks including explicit relations between actors (your Facebook friend network)
  4. other types of community data (discussion forums, email conversations, …)

Identification of Prominent Nodes

Question: Who are the "most important" actors in a social network?

Centrality

- A central actor is one involved in many edges.
- The direction of lines is not considered.



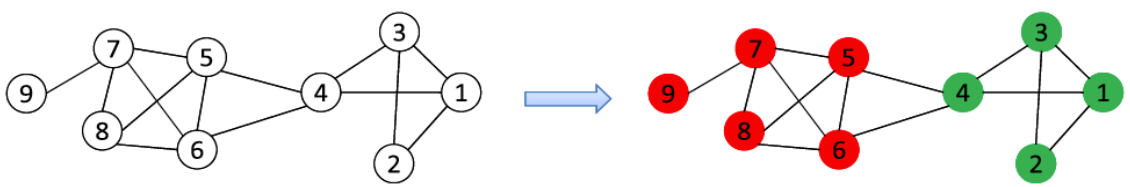
Prestige

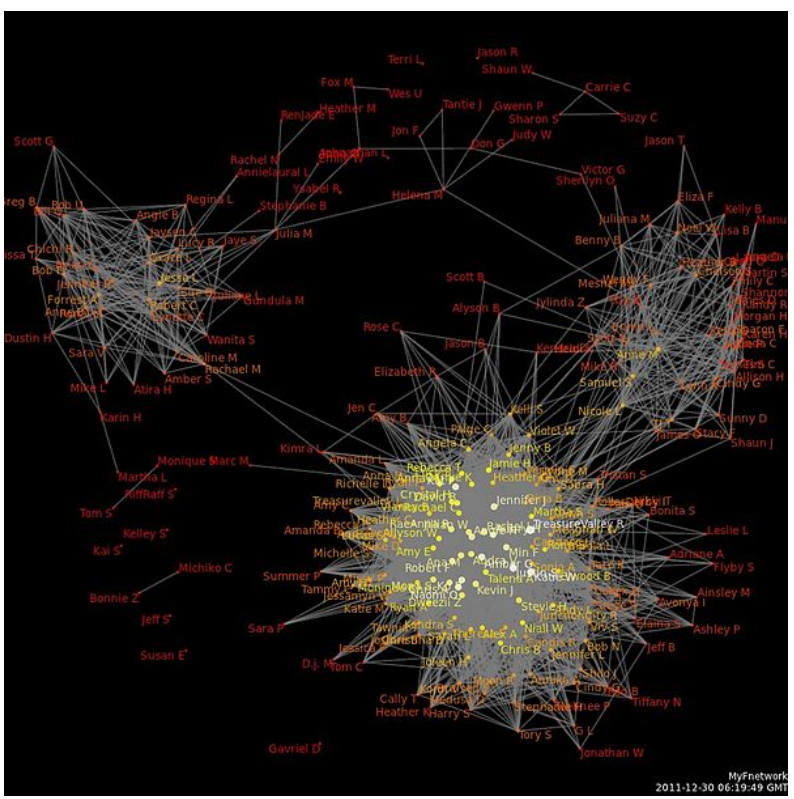
- A prestigious actor is one who is the target of many arcs.
- The direction of arcs is considered.

Community Detection

A community is a set of actors between which interactions are (relatively) frequent.

- Finding a community in a social network is to identify a set of nodes such that they interact with each other more frequently than with those nodes outside the group.





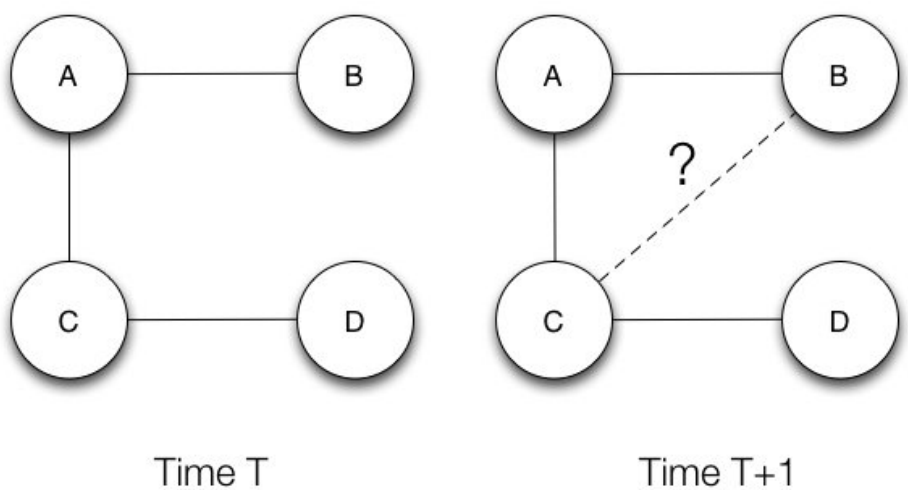
- Methods: Components, K-Cores, Islands, …
- Applications: Recommendation based on communities, visualization of huge networks, network compression

Link Prediction

Question: Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? (Liben-Nowell & Kleinberg, 2007)

Applications

- **Facebook: recommending possible friends**
- **Tinder: recommending potential matches**



D. Liben-Nowell, D. and J. Kleinberg, ***The link-prediction problem for social networks.*** *Journal of the American Society for Information Science and Technology*, 58(7) 1019–1031 (2007)

3.3 Web Content Mining

- **Definition**

Automatic extraction of useful information (facts, patterns) from Web content (text, images, multimedia).

- **Content Mining Tasks**
  - **Content classification and clustering on Web content**
  - **Applications of NLP techniques to social network content**

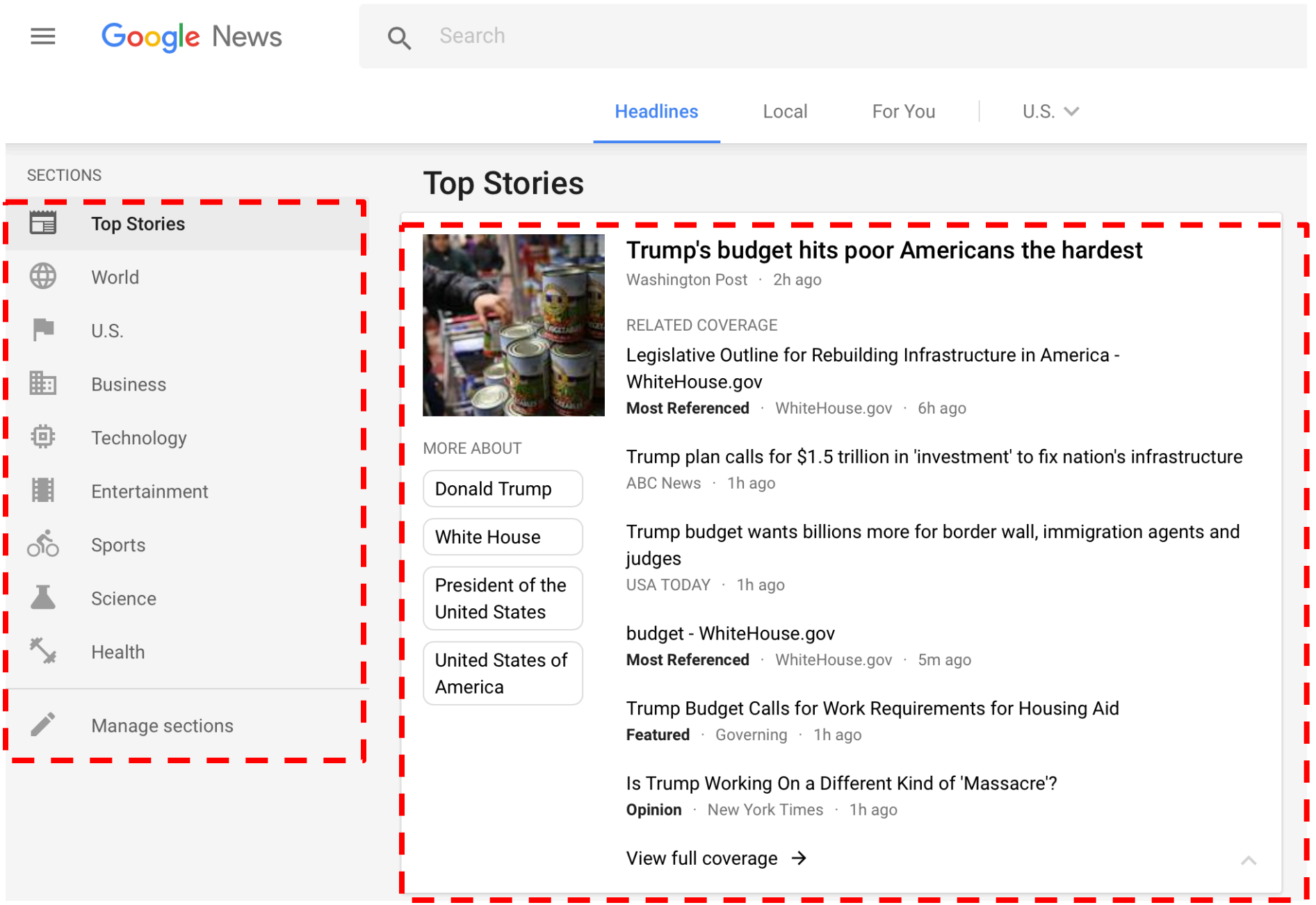
Content Classification

- Supervised Learning: Given a collection of labeled documents/images (training set) find a model for the class as a function of the values of the features.
- Goal: Previously unseen documents/images should be assigned a class as accurately as possible.
- Applications
  - News categorization
  - Topic classification
  - Spam detection
  - Product categorization
- Classification methods commonly used for
  - Naive Bayes, Support Vector Machines, Deep Neural Nets

Content Clustering

- Unsupervised Learning: Given a set of documents and a similarity measure among documents find clusters such that:
  - documents in one cluster are more similar to one another
  - documents in separate clusters are less similar to one another
- Applications
  - Topic discovery
  - Search result clustering
- Techniques
  - Algorithms: K-Means, Expectation-Maximization (EM)
  - Similarity measures: Cosine, Jaccard

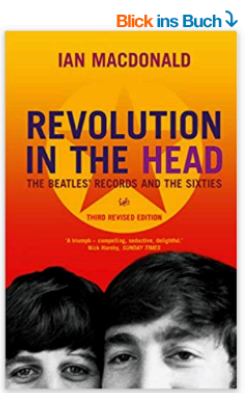
Mixture of Document Clustering and Classification



Sentiment Analysis

The basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level.

- **Polarity Values**
  - Positive, neutral, negative
  - Stars
- **Applications**
  - Document-level: poll prediction from tweets
  - Feature/Aspect-level: analysis of product reviews



Revolution In The Head: The Beatles Records and the Sixties (Englisch) Taschenbuch – 4. Dezember 2008

von Ian MacDonald (Autor)

175 Sternebewertungen

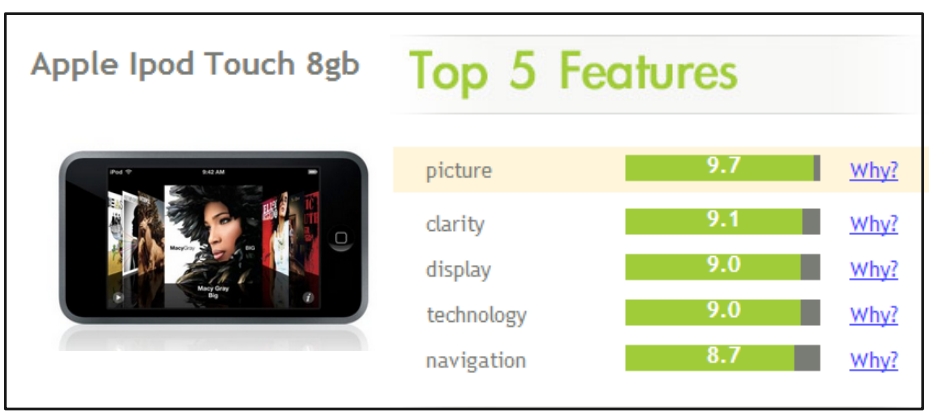
Alle 3 Formate und Ausgaben anzeigen

Taschenbuch
5,69 €

Lieferung 19.-24. Febr., wenn Sie **Standardversand** an der Kasse wählen. Siehe Details.

18 neu ab 5,69 € | 5 gebraucht ab 3,60 €

As dazzling as the decade they dominated, The Beatles almost single-handedly created pop music as we know it. Today, their songs are cited as seminal influences by stars like Oasis and Blur. Eloquently giving voice to their time, The Beatles quite simply changed the world.



Topic tracking / hate speech / offensive language

Applications of NLP to social media data

Twitter	% posts	Whisper	% posts
I hate	70.5	I hate	66.4
I can't stand	7.7	I don't like	9.1
I don't like	7.2	I can't stand	7.4
I really hate	4.9	I really hate	3.1
I fucking hate	1.8	I fucking hate	3.0
I'm sick of	0.8	I'm sick of	1.4
I cannot stand	0.7	I'm so sick of	1.0
I fuckin hate	0.6	I just hate	0.9
I just hate	0.6	I really don't like	0.8
I'm so sick of	0.6	I secretly hate	0.7

Twitter		*Whisper*	
Hate target	**% posts**	**Hate target**	**% posts**
Nigga	31.11	Black people	10.10
White people	9.76	Fake people	9.77
Fake people	5.07	Fat people	8.46
Black people	4.91	Stupid people	7.84
Stupid people	2.62	Gay people	7.06
Rude people	2.60	White people	5.62
Negative people	2.53	Racist people	3.35
Ignorant people	2.13	Ignorant people	3.10
Nigger	1.84	Rude people	2.45
Ungrateful people	1.80	Old people	2.18

Questions?