universität koblenz
weiter:denken

Web Retrieval
Search on the Web

Frank Hopfgartner
Institute for Web Science and Technologies

Intended Learning Outcomes

universität koblenz
weiter:denken

At the end of the lecture, you will be able to:

- Describe the structure of the Web
- Outline personalisation approaches
- Understand how ads are embedded on search result lists
- Describe spamming approaches to improve search ranking

Outline

universität koblenz
weiter:denken

- Motivation
- Web Graph
- Personalisation
- Ads
- Spamming

universität koblenz
weiter:denken

1. Motivation

Classical IR

universität koblenz
weiter:denken

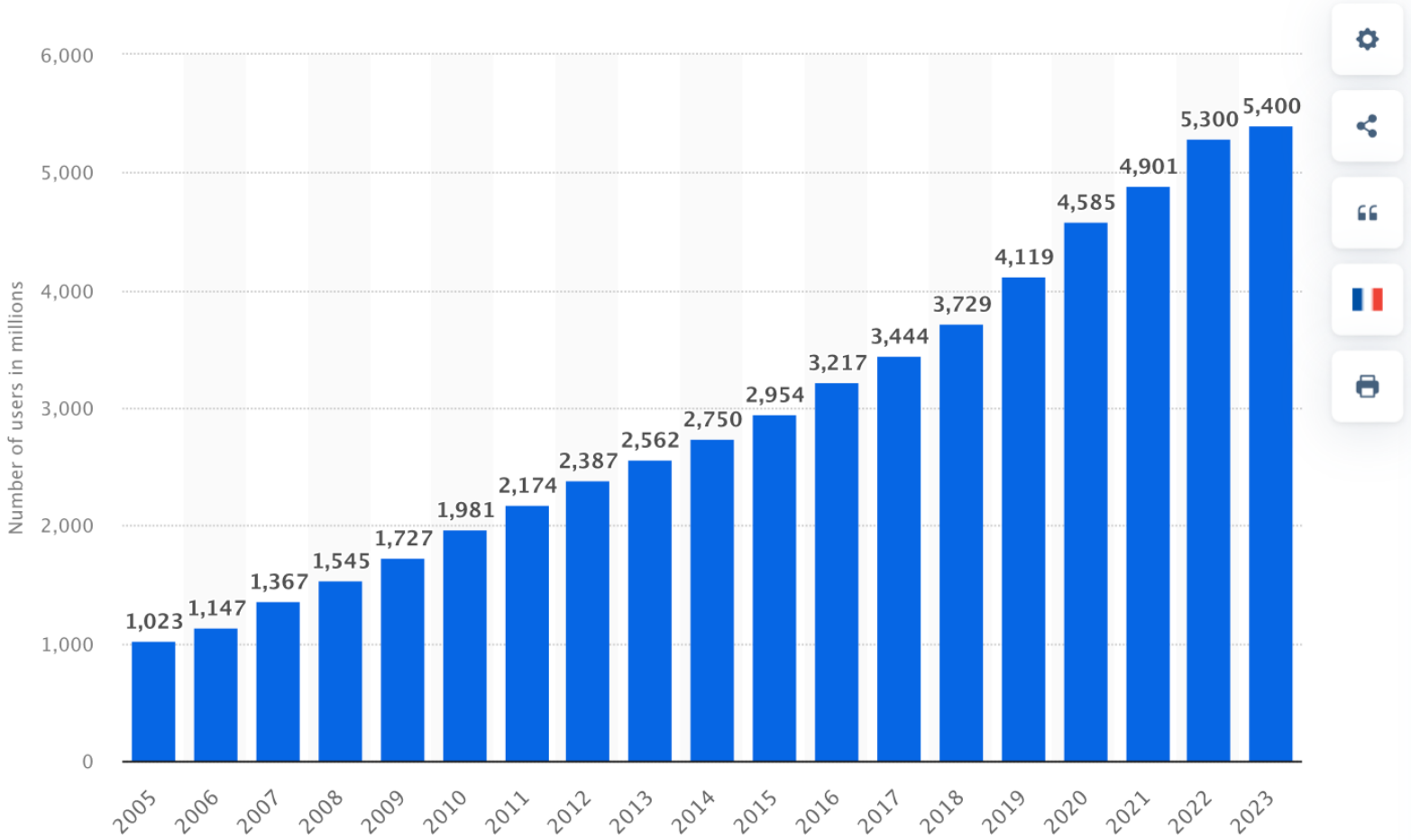
- Corpus
  - Fixed collection
  - Corpus is predetermined
- Goal
  - Retrieve documents with content relevant to user's information need
- Relevance
  - For every query q and a document d, there exists a relevance score $Score(q, d)$
    - Score is context independent
    - Score is user independent

Web Users

universität koblenz
weiter:denken

Number of internet users worldwide from 2005-2023

- Billions of information needs to satisfy
- Relevance is context-dependent



Details: Worldwide; ITU; 2005 to 2023; figures are estimates

Content on the Web

universität koblenz
weiter:denken

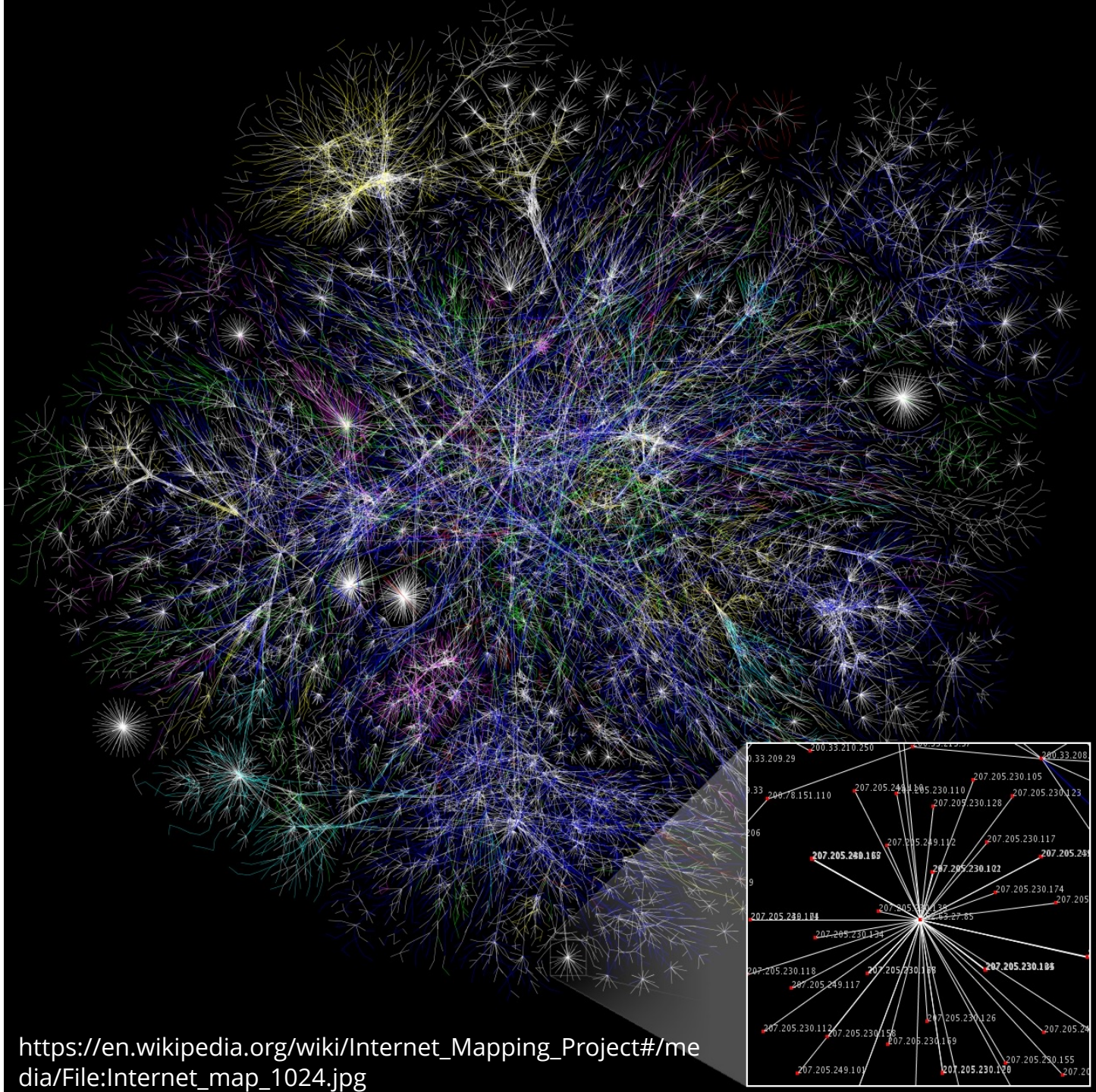
- Editorial content
- User-generated content
- AI-generated content

- Truth, lies, obsolete information, contradictions
- Unstructured, semi-structured, structured
- Can be dynamically generated

The Web has structure

- HTML formatting
- Hyperlinks betweeen web pages

Much more about this in Week 7



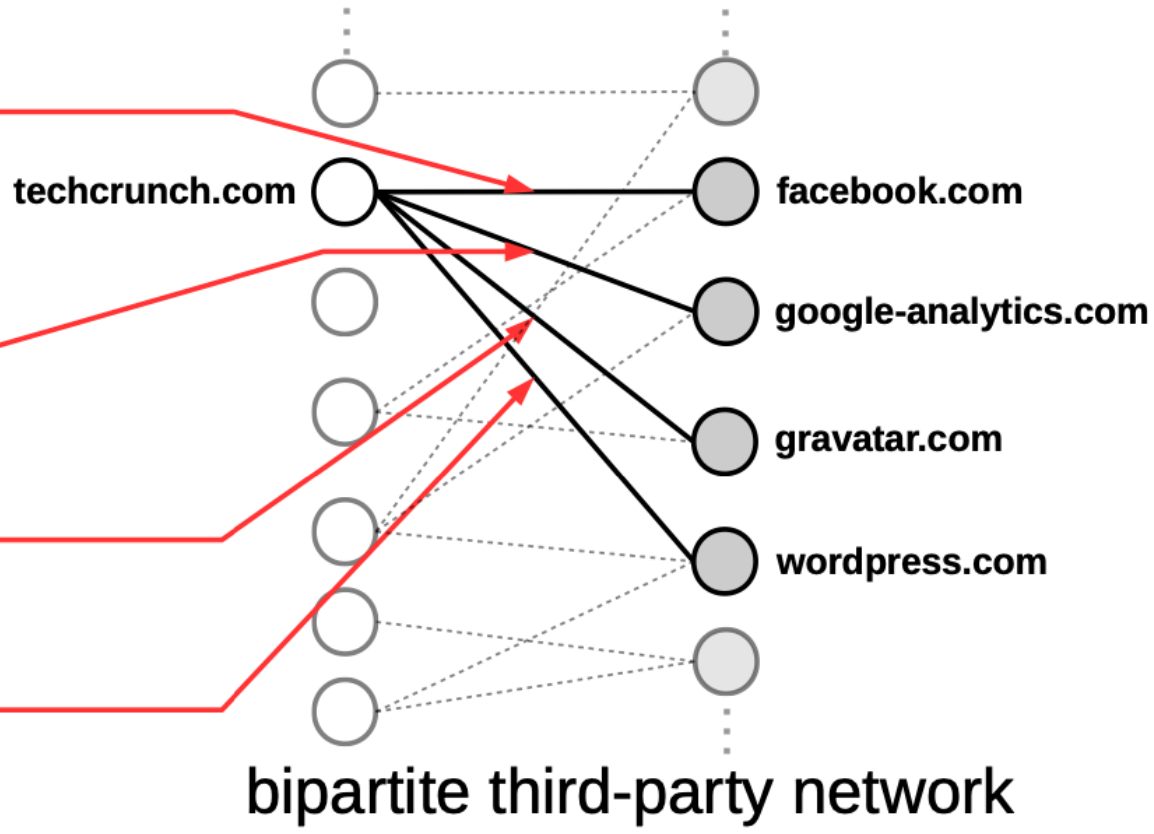
https://en.wikipedia.org/wiki/Internet_Mapping_Project#/media/File:Internet_map_1024.jpg

Web Tracking is happening

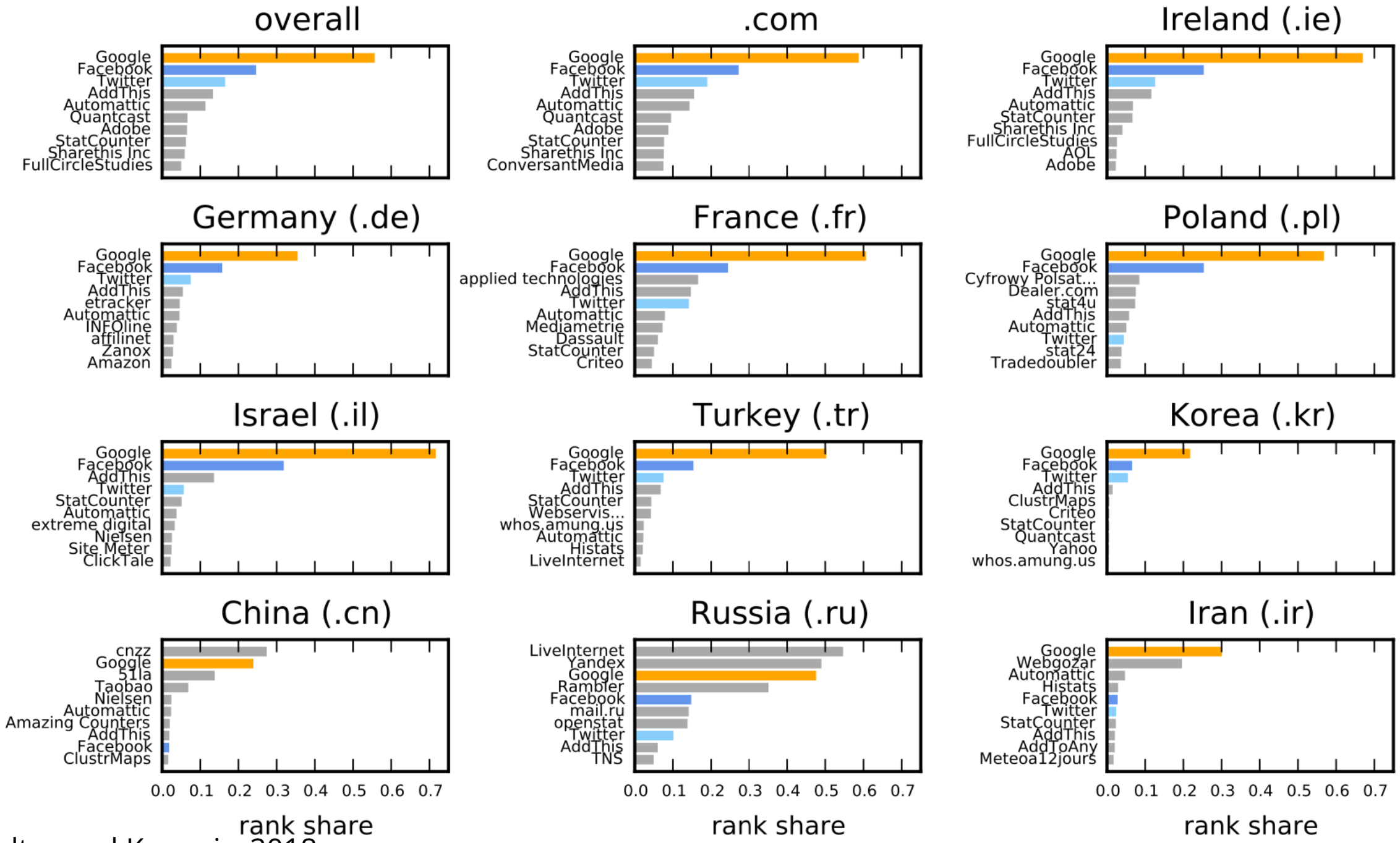
universität koblenz
weiter:denken

http://techcrunch.com

```
...
<div id="rr-social-widget-facebook">
  <iframe src="//www.facebook.com/plugins/
    like.php?href=http%3A%2F%2Fwww.facebook.com
    %2Ftechcrunch&amp;layout=standard…"…>
  </iframe>
</div>
...
<script type="text/javascript">
  var gaJsHost = (("https:" == document.
   location.protocol)?"https://ssl.":"http://www.");
  document.write(unescape("%3Cscript
    src='" + gaJsHost +
    "google-analytics.com/ga.js'
    type='text/javascript'%3E%3C/script%3E"));
</script>
...
<script src='http://s.gravatar.com/js/
  type='text/javascript'
  gprofiles.js?aa&#038;ver=3.5-alpha-21304'>
</script>
...
<noscript>
  <img src="http://stats.wordpress.com/
    b.gif?v=noscript" … />
</noscript>
...
```



Schelter and Kunegis. On the Ubiquity of Web Tracking: Insights from a Billion-page Web Crawl. Journal of Science, 2018, 4: 54-66



Schelter and Kunegis, 2018

Evolution of Web Search

universität koblenz
weiter:denken

- Content (1st Generation)
- Links (2nd Generation)
- Personalisation (3rd Generation)

universität koblenz
weiter:denken

1st generation Web search

- Early 1990s
- Examples: Lycos, Altavista, AlltheWeb,...

- Ranking signals
  - Term frequency (TF)
  - Inverse document frequency (IDF)
  - TF-IDF

universität koblenz
weiter:denken

2nd generation Web search

- Take the link structure of the Web into account
- Second half of 1990s
- Examples: Google (PageRank), Ask! (HITS)

- Ranking signals
  - Website popularity

Personalisation

universität koblenz
weiter:denken

3rd generation Web search

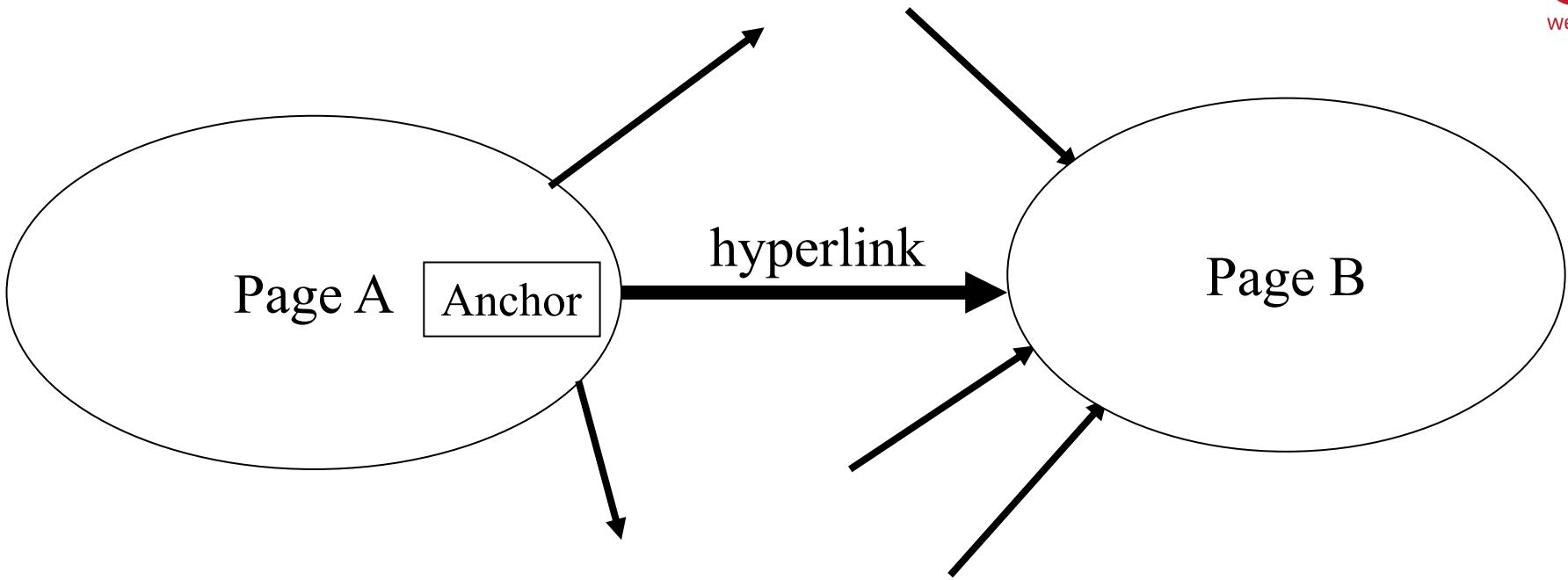
- Provide search results tailored to the individual user
- 2004: Google introduces personalised search

- Ranking signals
  - Users' relevance feedback
  - Context

universität koblenz
weiter:denken

2. Web graph

The Web as a directed graph



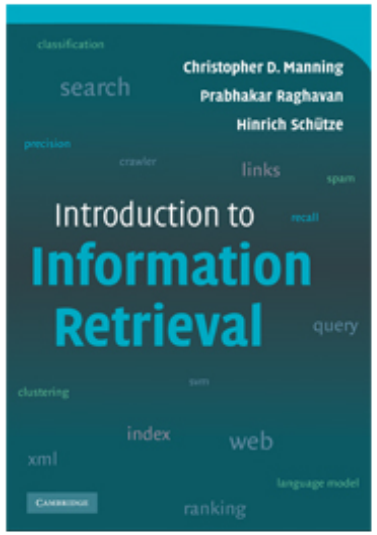
Hypothesis 1: A hyperlink between pages denotes a conferral of authority (quality signal)

Hypothesis 2: The text in the anchor of the hyperlink on page A describes the target page B

Assumption 1: reputed sites

universität koblenz
weiter:denken

Introduction to Information Retrieval



This is the companion website for the following book.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Informat*

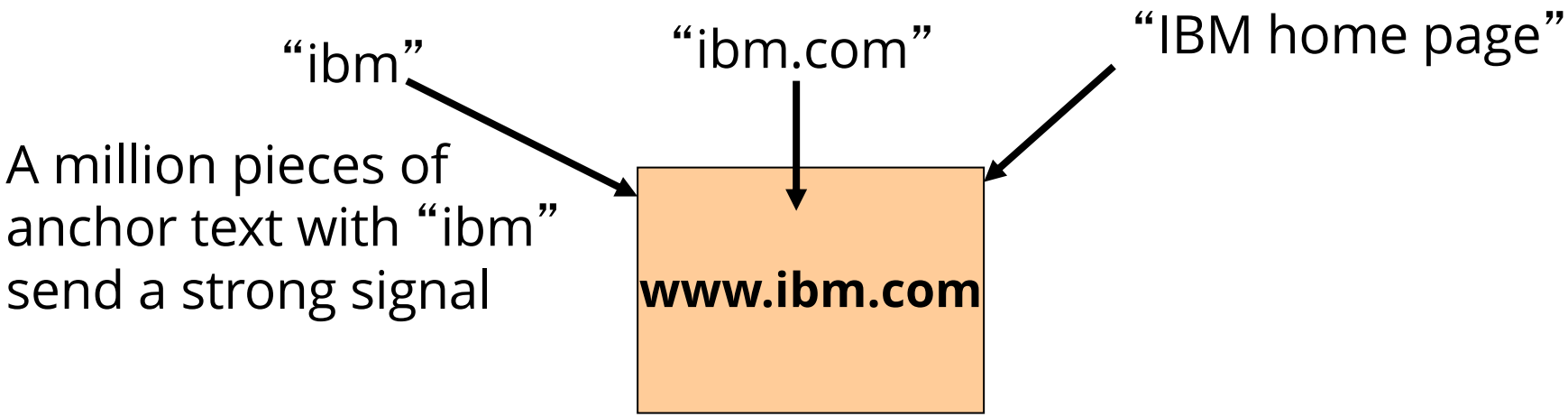
You can order this book at CUP, at your local bookstore or on the internet. The best search

The book aims to provide a modern approach to information retrieval from a computer scie
University and at the University of Stuttgart

We'd be pleased to get feedback about how this book works out as a textbook, what is m
comments to: informationretrieval (at) yahoogroups (dot) com

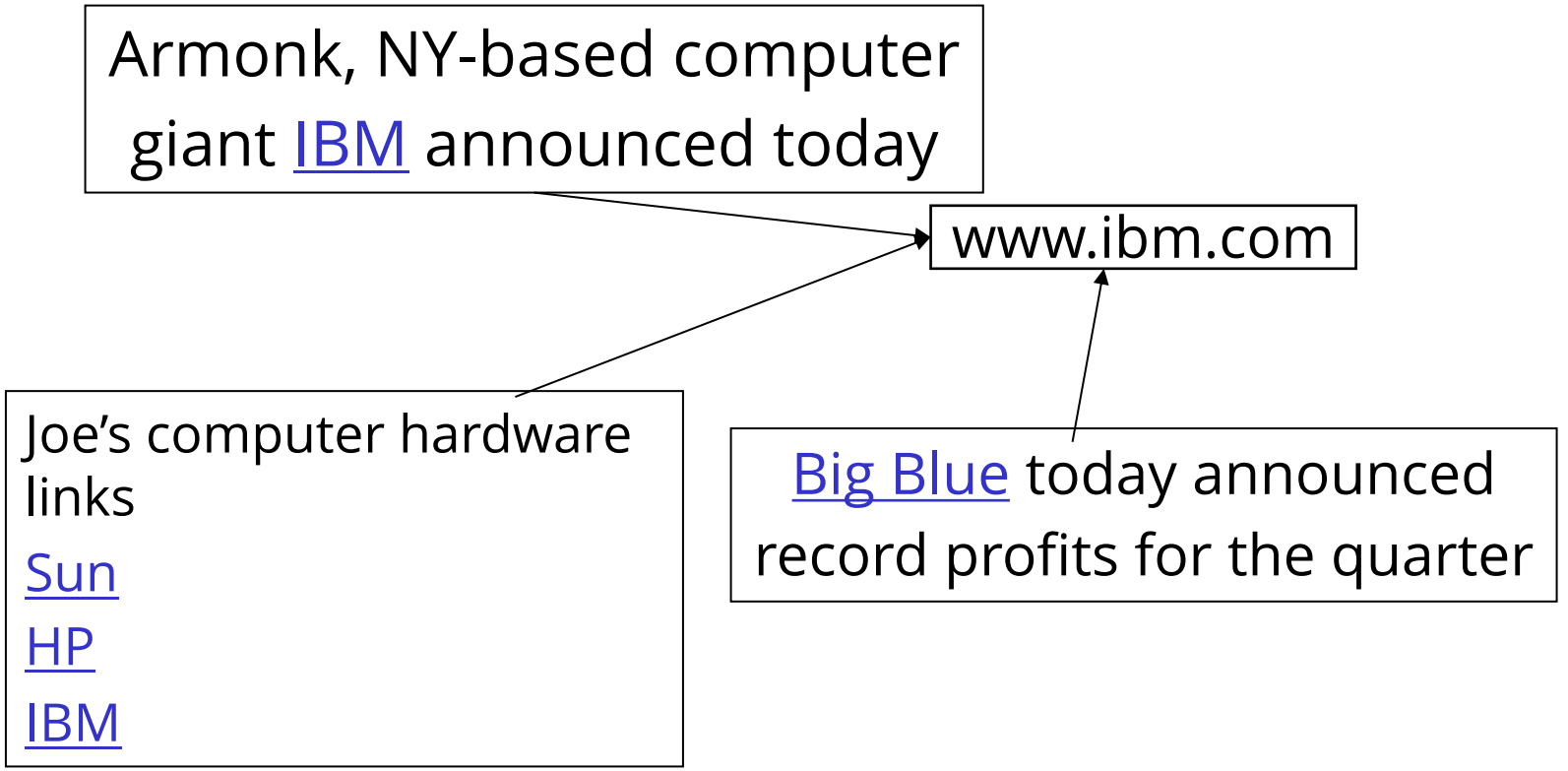
Anchor text

- For ***ibm*** how to distinguish between
  - IBM's home page (mostly graphical)
  - IBM's copyright page (high term freq. for 'ibm')
  - Rival's spam page (arbitrarily high term freq.)



Indexing anchor text

- When indexing a document *D*, include (with some weight) anchor text from links pointing to *D*



Indexing anchor text

universität koblenz
weiter:denken

- Thus: Anchor text is often a better description of a page's content than the page itself
- Anchor text can be weighted more highly than document text

Schema.org

universität koblenz
weiter:denken

- Initiative by Bing, Google, Yahoo!, Yandex, …
- De-facto standard vocabulary for structured data on the Web
- Can be used to describe the meaning of websites

universität koblenz
weiter:denken

3. Personalisation

Pros and Cons

universität koblenz
weiter:denken

- Saves time by reducing number of results to inspect
- Better decision making by filtering out inferior information

- Filter bubble (as much a personal decision as an algorithmic restriction)
- Users are products (using search history for advertisement)

Ranking input

universität koblenz
weiter:denken

Personal details

- Information about the users
- Ranking signals
  - Language
    - Language preferences can be used to filter out results
  - Demographics
    - Usually predicted
    - Results selected by other users from similar cohorts can be ranked higher

Ranking input

universität koblenz
weiter:denken

Social networks

- Information about a user's social network
- Ranking signals
  - Social network connections
    - Results selected by friends for similar searches could be given more weight
    - Web pages shared by friends could be given more weight

Ranking input

universität koblenz
weiter:denken

Prior activities (query logs)

- Information about the queries submitted by the user and other users in the past
- Ranking signals
  - Query suggestion
    - Other users entered queries A and B in the same session -> B might be a good suggestion for a user entering Query A
  - Spellign correction
    - Immediately after query X other query Y -> Y might be the correct version of query X

Ranking input

universität koblenz
weiter:denken

Context

- Information about the context in which the search is performed
- Ranking signals
  - Location
    - Used to prioritise locally relevant results
    - Essential for mobile search
  - Date & Time
    - Seasonal influences, home vs. work, ….

Learning ro rank

universität koblenz
weiter:denken

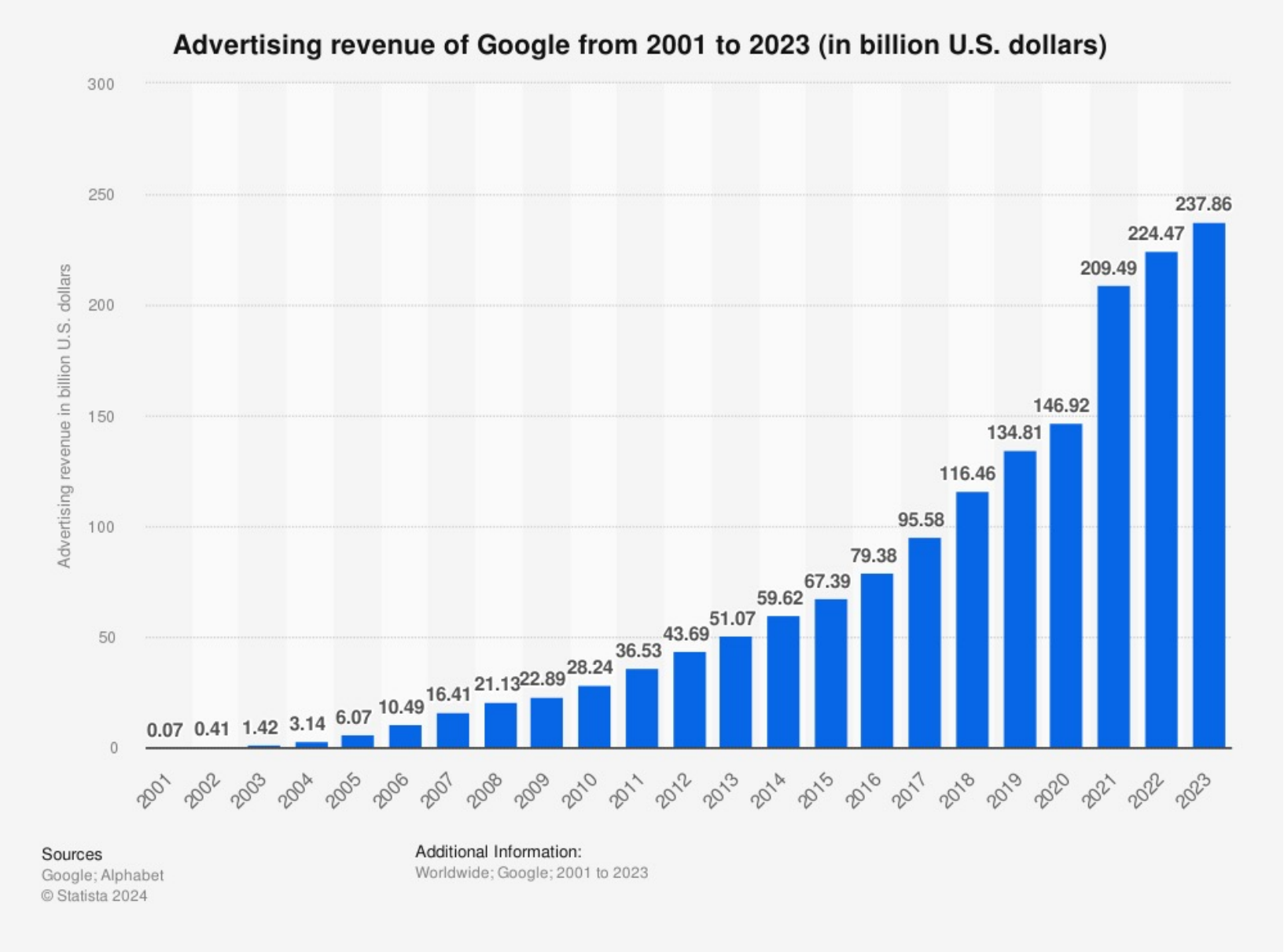
- Learning the optimal combination of all ranking signals
- Goal: to do this continuously and automatically using machine learning
  - Predict for each query-result pair whether the result is relevant for that user's query at this specific time

universität koblenz
weiter:denken

4. Ads

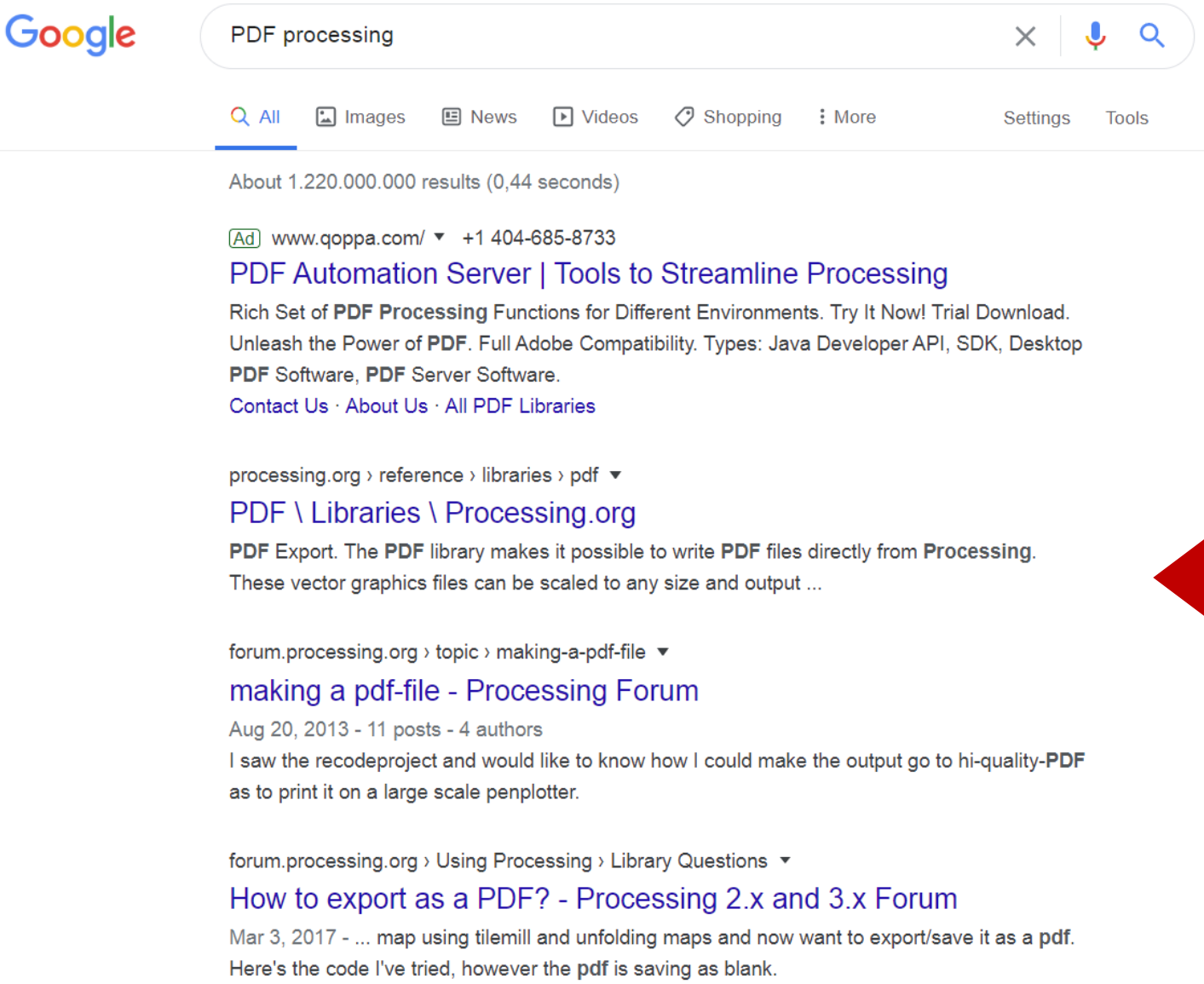
Advertising revenue of Google from 2001-2023

universität koblenz
weiter:denken



Google Ads

universität koblenz
weiter:denken



Ads

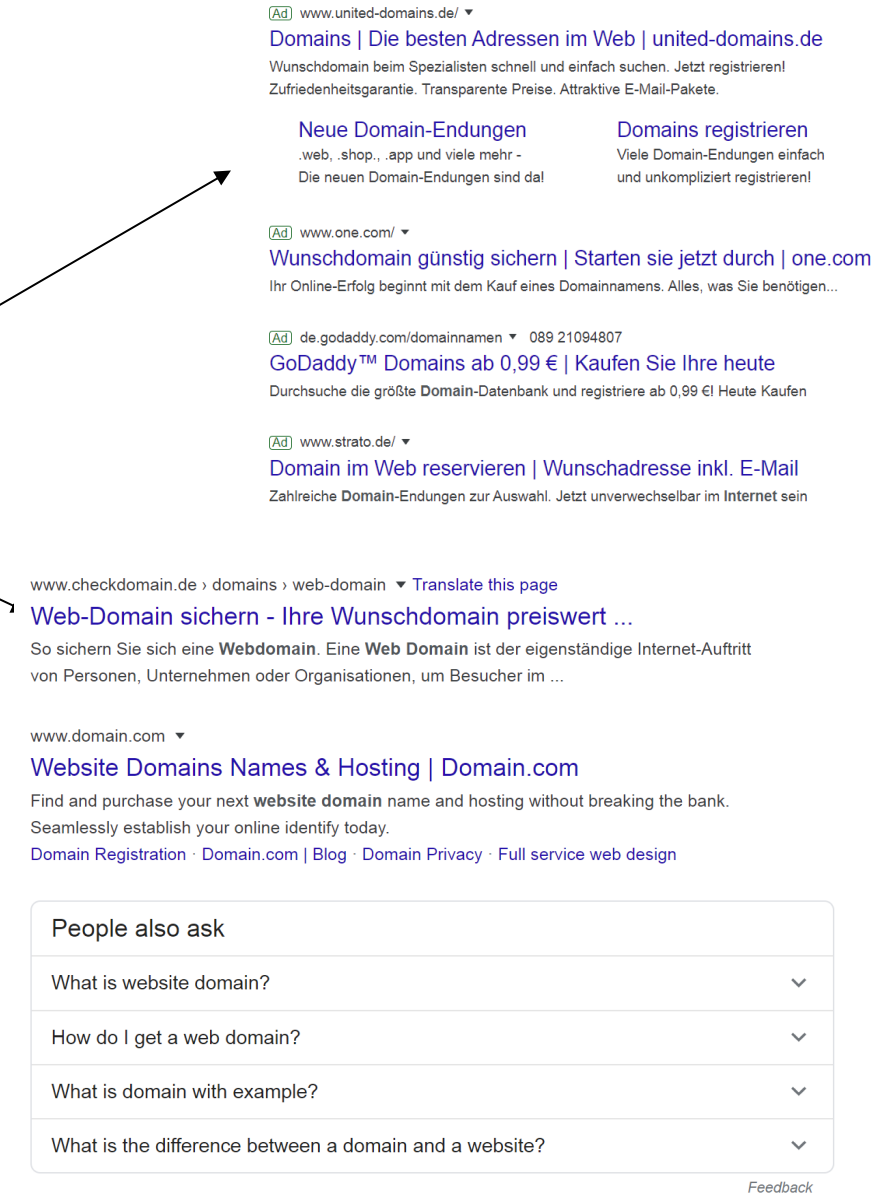
Algorithmic result

Ads vs. search results

universität koblenz
weiter:denken

Google has maintained that ads (based on vendors bidding for keywords) do not affect vendors' rankings in search results

Search = ***web domain***

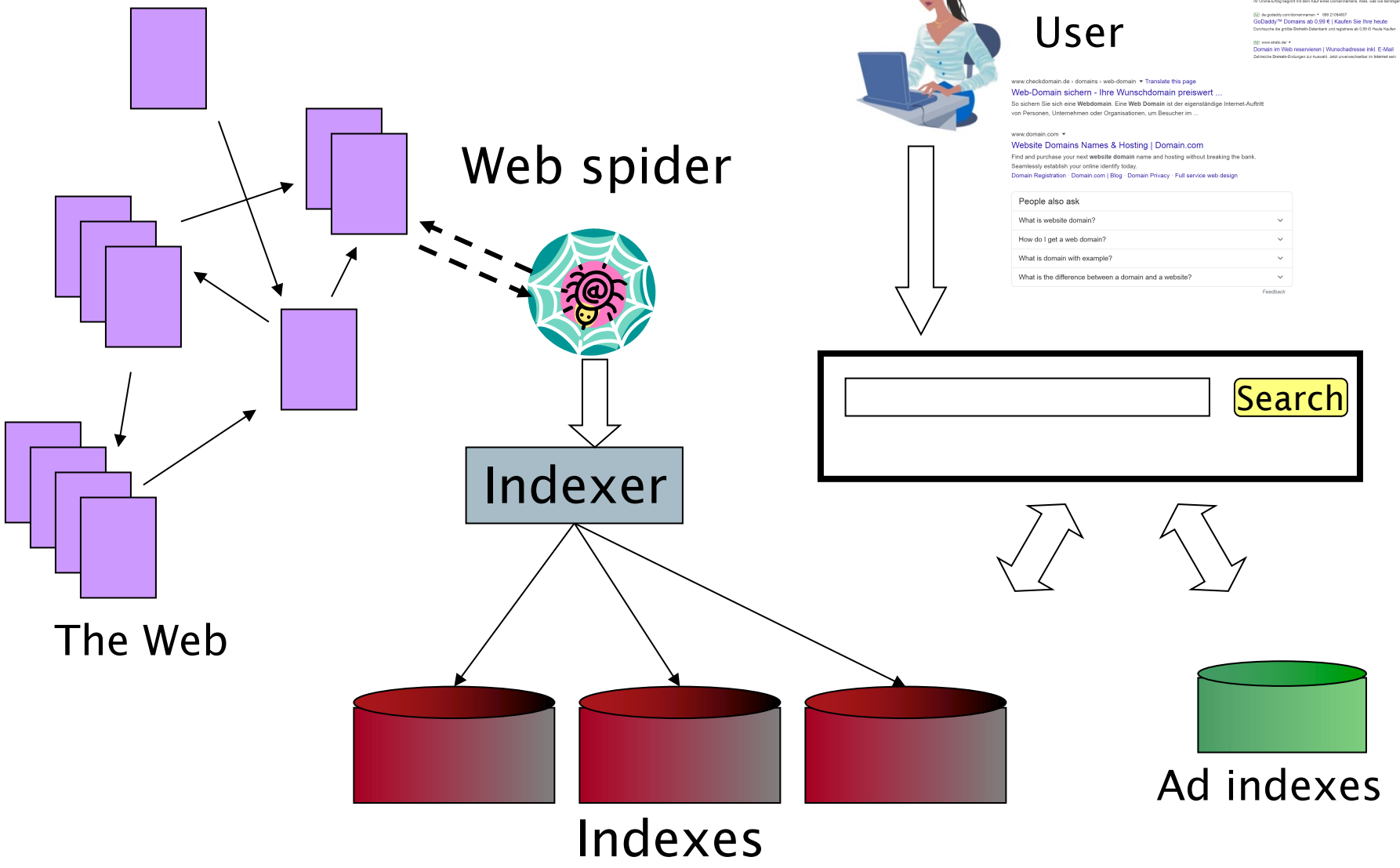


Ads vs. search results

universität koblenz
weiter:denken

- Other search engines (Yahoo, MSN) have made similar statements from time to time
  - Any of them can change anytime
- We will focus primarily on search results independent of paid placement ads
  - Although the latter is a fascinating technical subject in itself

Web search


Web spider
Indexer
The Web
Indexes
User
Search
Ad indexes
People also ask
What is website domain?
How do I get a web domain?
What is domain with example?
What is the difference between a domain and a website?

How are ads ranked?

universität koblenz
weiter:denken

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [Buying fresh Chicken?] → ad for KFC
  - We don’t want to show nonrelevant ads
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions
- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information
- Other ranking factors: location, time of day, quality and loading speed of landing page
- The main ranking factor: the query

Google's second price auction

universität koblenz
weiter:denken

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

- bid: maximum bid for a click by advertiser
- CTR: click-through rate: when an ad is displayed, what percentage of time do users click on it? CTR is a measure of relevance.
- ad rank: bid $\times$ CTR: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- rank: rank in auction
- paid: second price auction price paid by advertiser

Google's second price auction

universität koblenz
weiter:denken

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

Second price auction: The advertiser pays the minimum amount necessary to maintain their position in the auction (plus 1 cent)

- $price_1 = bid_2 \times CTR_2 / CTR_1$
- $p_1 = bid_2 \times CTR_2/CTR_1 = 3.00 \times 0.03/0.06 = 1.50$
- $p_2 = bid_3 \times CTR_3/CTR_2 = 1.00 \times 0.08/0.03 = 2.67$
- $p_3 = bid_4 \times CTR_4/CTR_3 = 4.00 \times 0.01/0.08 = 0.50$

Keywords with high bids

universität koblenz
weiter:denken

- According to https://www.wordstream.com/articles/most-expensive-keywords

Insurance	$54.91
Loans	$44.28
Mortgage	$47.12
Attorney	$47.07
Credit	$36.06
Lawyer	$42.51
Donate	$42.02
Degree	$40.61
Hosting	$31.91
Claim	$45.51
Conference Call	$42.05
Trading	$33.19
Software	$35.29

Search ads: a win-win-win?

universität koblenz
weiter:denken

- The search engine company gets revenue every time somebody clicks on an ad
- The user only clicks on an ad if they are interested in the ad
  - Search engines punish misleading and nonrelevant ads
  - As a result, users are often satisfied with what they find after clicking on an ad
- The advertiser finds new customers in a cost-effective way

Search ads

universität koblenz
weiter:denken

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
- Any way this could be bad for the user?
- Any way this could be bad for the search engine?

Not a win-win-win: keyword arbitrage

universität koblenz
weiter:denken

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user
- Ad spammers keep inventing new tricks
- The search engines need time to catch up with them

universität koblenz
weiter:denken

5. Spam

The trouble with paid placement

universität koblenz
weiter:denken

- It costs money. What's the alternative?
- *Search Engine Optimization*
  - "Tuning" your web page to rank highly in the search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

Simplest forms

universität koblenz
weiter:denken

- First generation engines relied heavily on *tf/idf*
  - The top-ranked pages for the query **maui resort** were the ones containing the most **maui**'s and **resort**'s
- SEOs responded with dense repetitions of chosen terms
  - e.g., **maui resort maui resort maui resort**
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers

Pure word density cannot be trusted as an IR signal

Variants of keyword stuffing

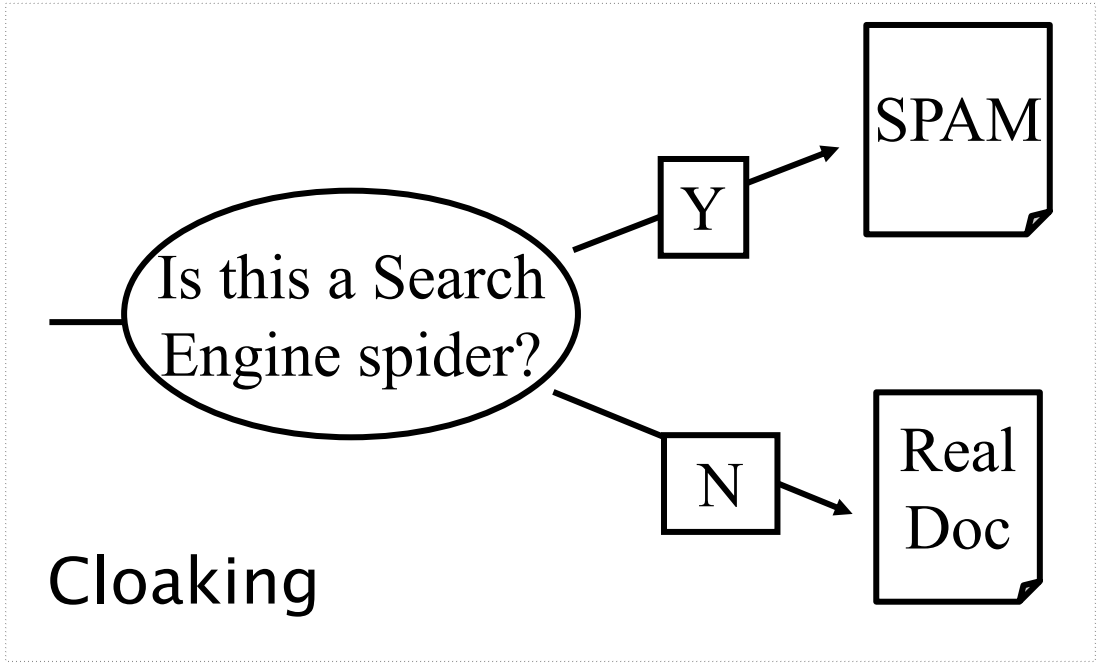
universität koblenz
weiter:denken

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks, etc.

Meta-Tags =
"… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …"

Cloaking

- Serve fake content to search engine spider
- DNS cloaking: Switch IP address. Impersonate



Cloaking

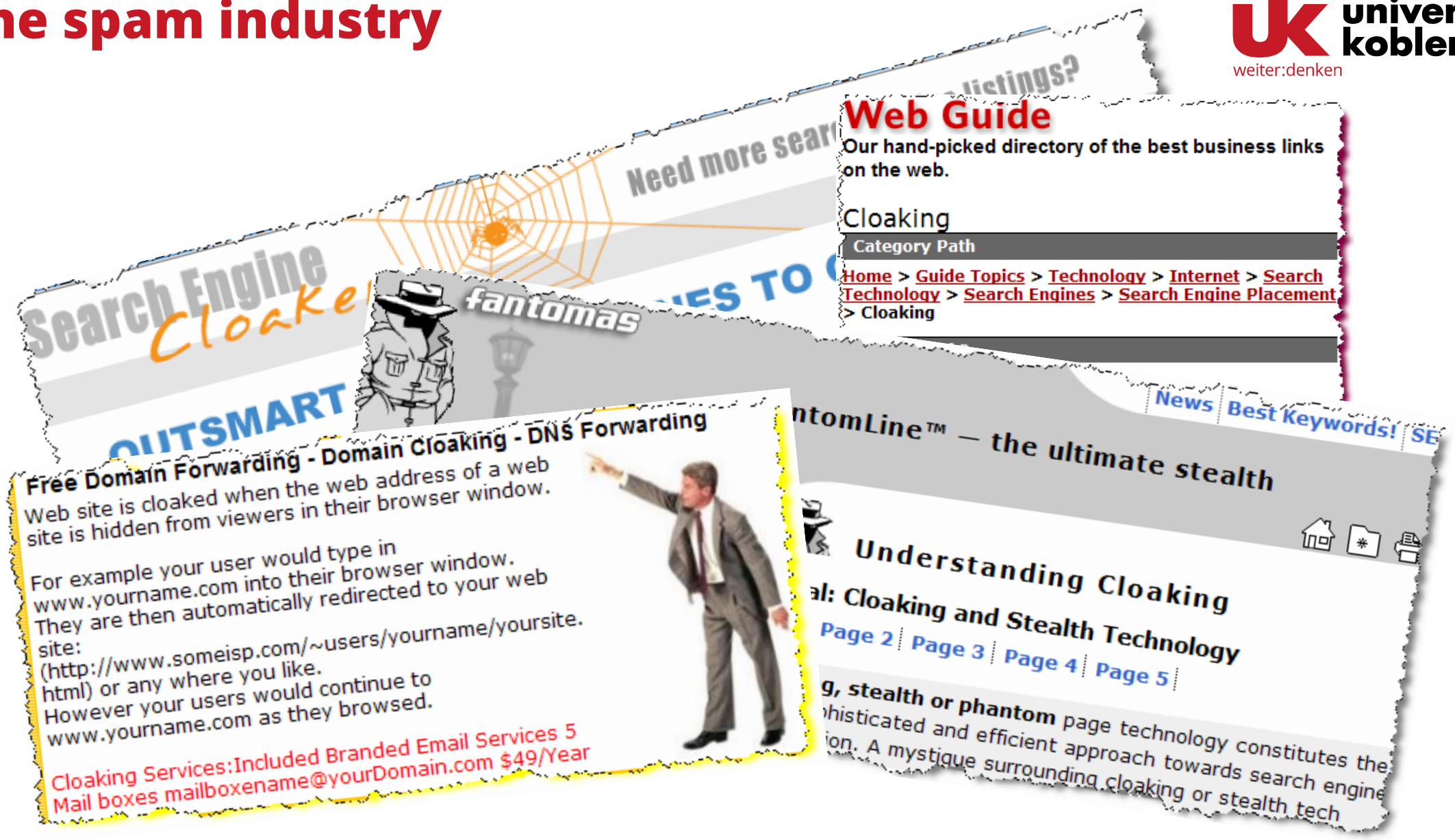
Search engine optimization (Spam)

universität koblenz
weiter:denken

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world (www.webmasterworld.com)
    - Search engine specific tricks

The spam industry

universität koblenz
weiter:denken



The war against spam

universität koblenz
weiter:denken

- Quality signals - Prefer authoritative pages based on
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

More on spam

universität koblenz
weiter:denken

- Web search engines have policies on SEO practices they tolerate/block
  - http://help.yahoo.com/help/us/ysearch/index.html
  - http://www.google.com/intl/en/webmasters/
- Adversarial IR: the unending (technical) battle between SEO's and web search engines
- Research http://airweb.cse.lehigh.edu/

universität koblenz
weiter:denken

6. Summary

Intended Learning Outcomes

You are now able to:

- Describe the structure of the Web
- Outline personalisation approaches
- Understand how ads are embedded on search result lists
- Describe spamming approaches to improve search ranking