Recherchieren im Internet: Wie funktionieren Suchmaschinen?

Steffen-Peter Ballstaedt

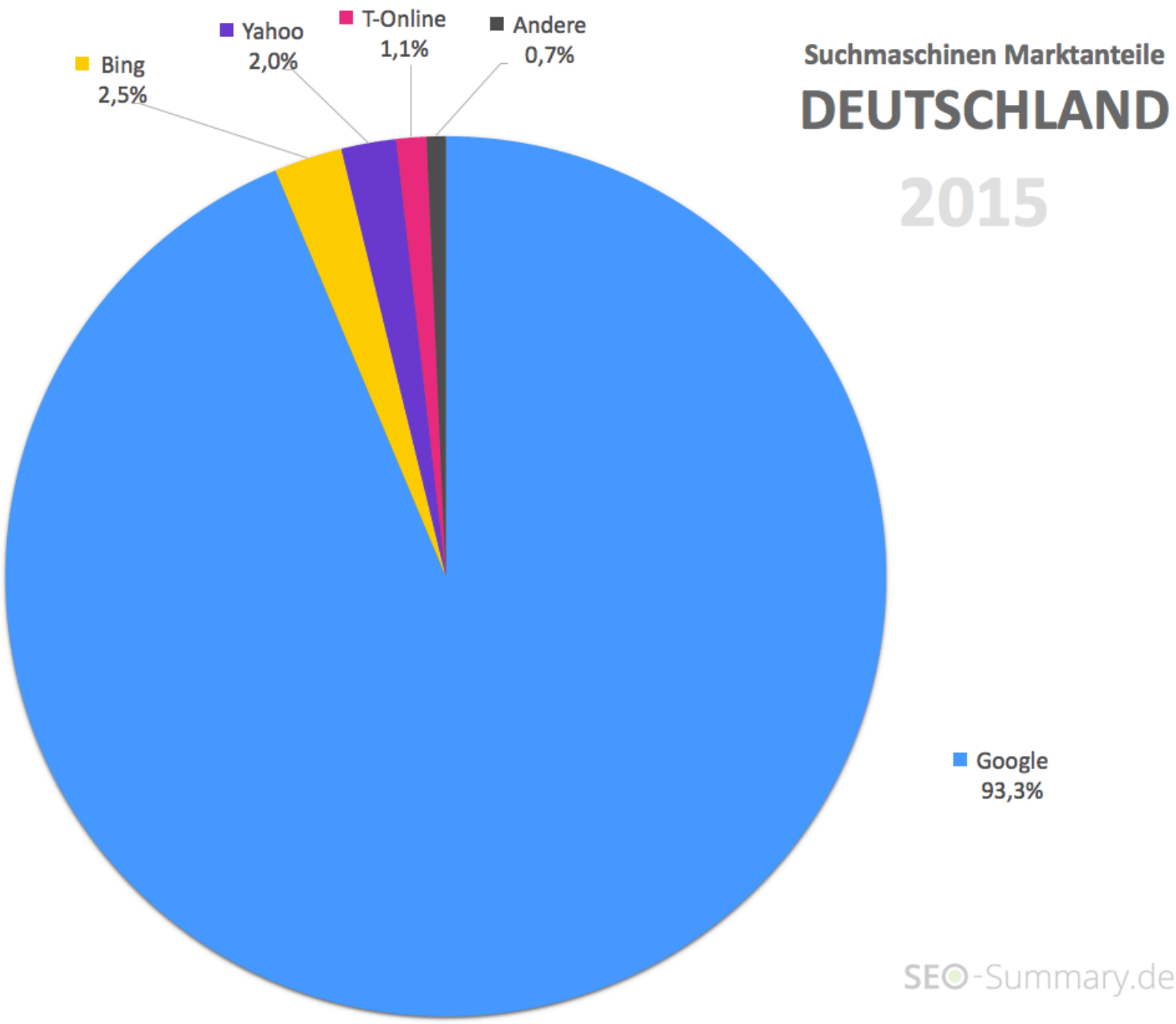
Vortrag vor dem Genealogischen Arbeitskreis Tübingen am 17.2.2016

Zahlen und Fakten

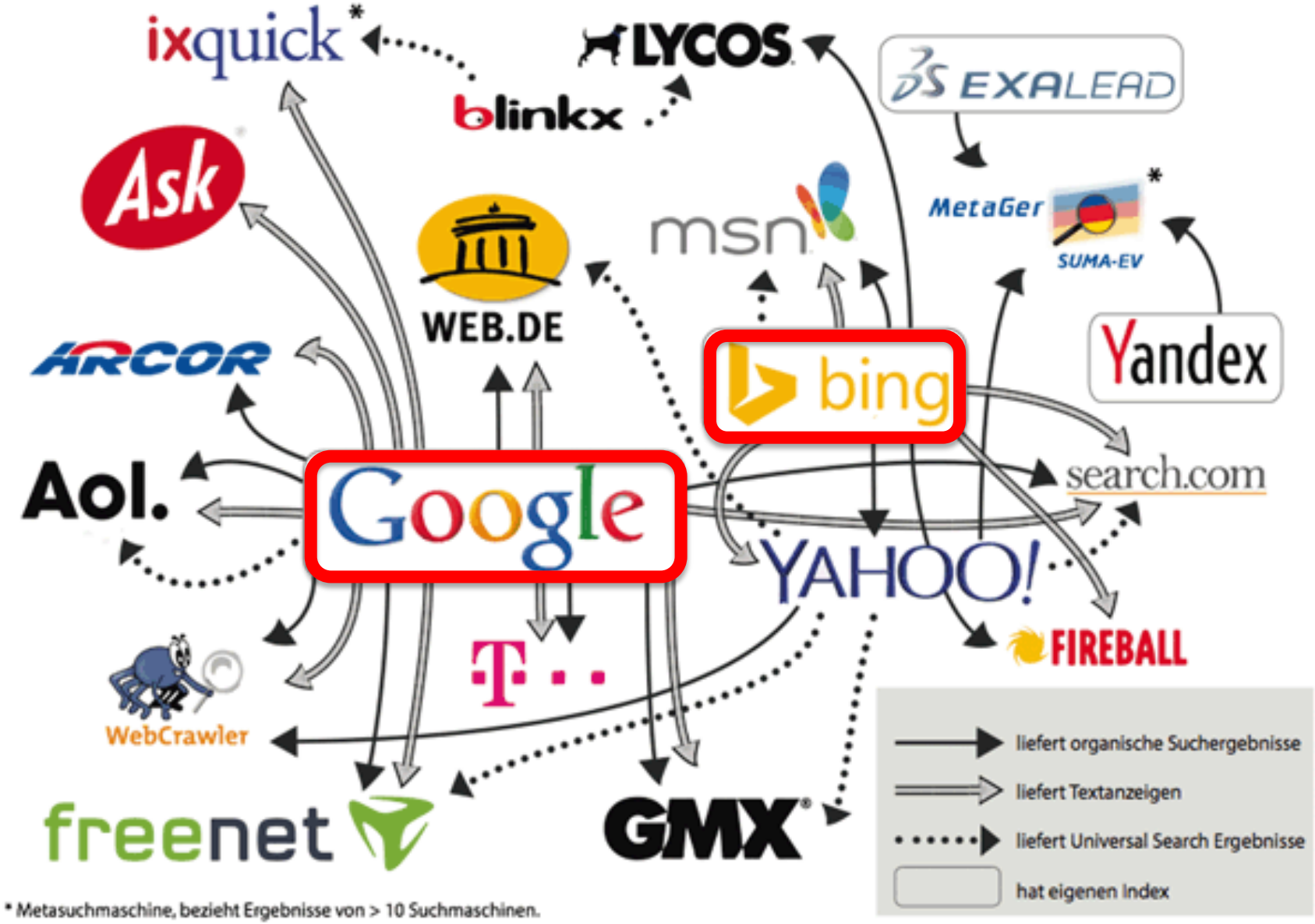
- 82% der Computernutzer suchen, 79% mailen.
- Die häufigste Aktivität im Internet ist Mailen, gleich danach kommt Suchen.
- Über 50.000 Anfragen in der Sekunde bei Google!

internet live stats

Platzhirsch Google



Beziehungsgeflecht der Suchmaschinen



Zahlen und Fakten

- Die Ausdehnung des Internet kann nur grob geschätzt werden!

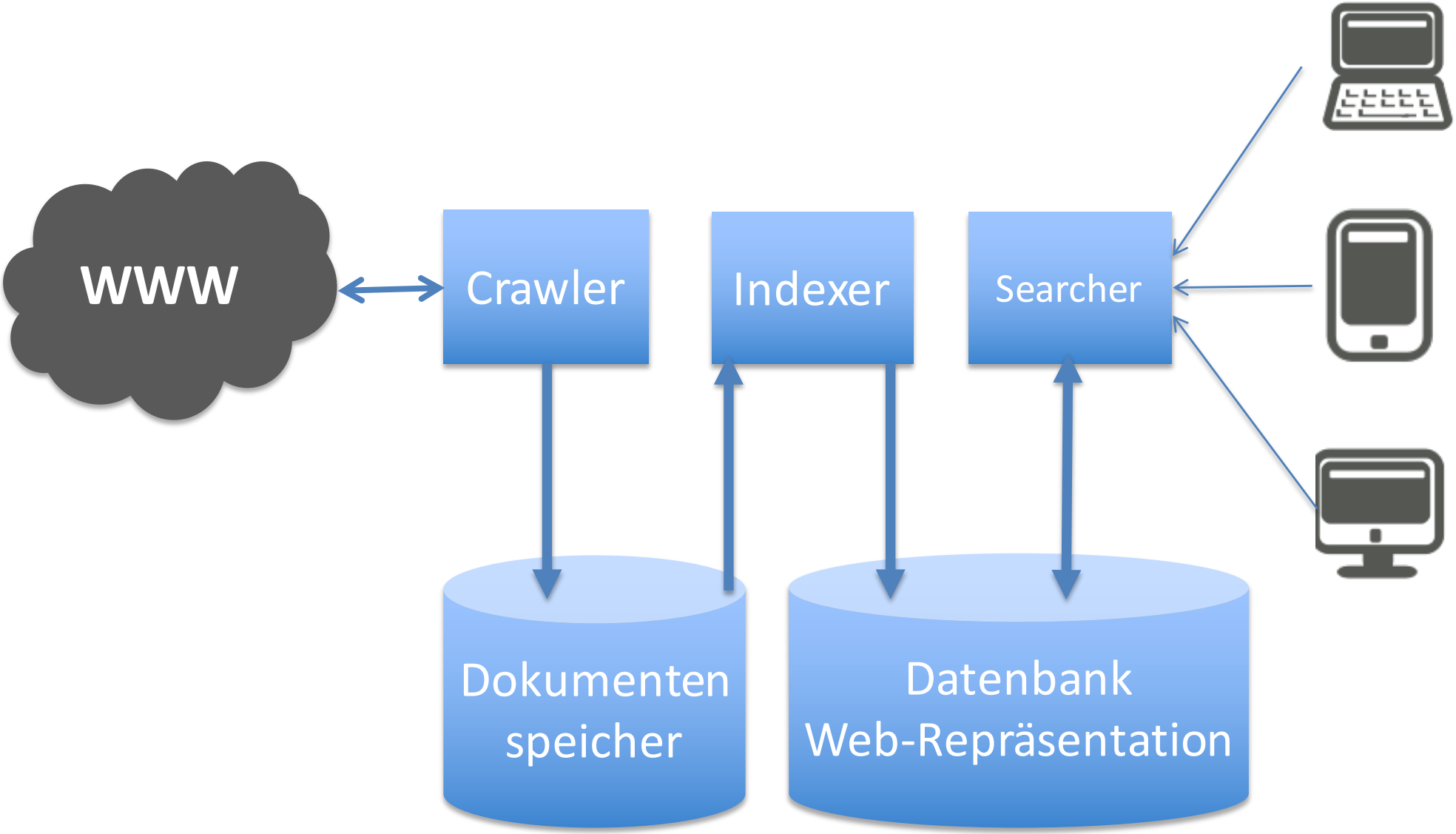
- Weltweit: über 1 Milliarde Websites
- BRD: etwa 15 Millionen Websites
- Das Internet verdoppelt sich alle 5,32 Jahre

Antworten auf folgende Fragen:

- Wie findet eine Suchmaschine Inhalte bzw. Dokumente im WWW?
- Was kann eine Suchmaschine nicht finden?
- Wie kommt die Trefferliste zustande (Ranking)?
- Welche Alternativen zu Google gibt es?
- Wie kann man personenbezogene Daten im Web suchen?

Wie findet eine Suchmaschine Inhalte im WWW?

Aufbau eine Suchmaschine



Crawling

- Ausgehend von bekannten Websites werden die Links verfolgt.
- Die Text-Dokumente werden erfasst und gespeichert (Dokumentenspeicher).
- Crawling findet kontinuierlich statt, nach Aktualität und Popularität.

Indexierung

- Erstellt eine Repräsentation des Dokuments aus sprachstatistischen Daten: Worthäufigkeit, Wortposition, Wort in Titel oder Überschrift.
- Der Index als eine riesige Datenbank ist das Herz einer Suchmaschine.
- Bei einer Suchanfrage wird nicht auf die Dokumente im Web, sondern auf diese Repräsentation des Web zugegriffen.

Beispiel: Indexierung von Text-Dokumenten

D1 Genealogie untersucht verwandtschaftliche Beziehungen zwischen Menschen.

D2 Heimatgeschichte ist eng mit Genealogie verbunden.

D3 Eine Quelle sind Kirchenbücher, eine andere Quelle sind Personenstandsbücher.

D4 Mit dem WWW hat die Genealogie einen Aufschwung erfahren.

Wort	Dokument	Häufigkeit	Position
andere	D3		D3,6
Aufschwung	D4		D4,8
Beziehungen	D1		D1,6
dem	D4		D4,2
die	D4		D4,4
eine	D3	D3:2	D3,1, D3,5
einen	D4		D4,7
eng	D2		D2,3
erfahren	D4		D4,9
Genealogie	D1, D2, D4	D1:1 D2:1 D4:1	D1,1 D2,5 D4,6
hat	D4		D4,4
Heimatgeschichte	D2		D2,1
ist	D2		D2,2
Kirchenbücher	D3		D3,4
Menschen	D1		D1,8
mit	D2, D4	D2:1 D4:1	D2,4 D4,1
Personenstandsbücher	D3		D3,9
Quelle	D3	D3:2	D3,2 D3,7
sind	D3	D3:2	D3,3 D3,8
untersucht	D1	D1:1	D1,2
verbunden	D2		D2,6
verwandtschaftliche	D1		D1,5
WWW	D4		D4,3
zwischen	D1		D1,7

Index der Dokumente

Suchanfragen, Stichworte

Genealogie	D1, D2, D4
Genealogie WWW	D4
Genealogie OR WWW	D1, D2, D4
„Verwandtschaftliche Beziehungen“	D1

Searcher

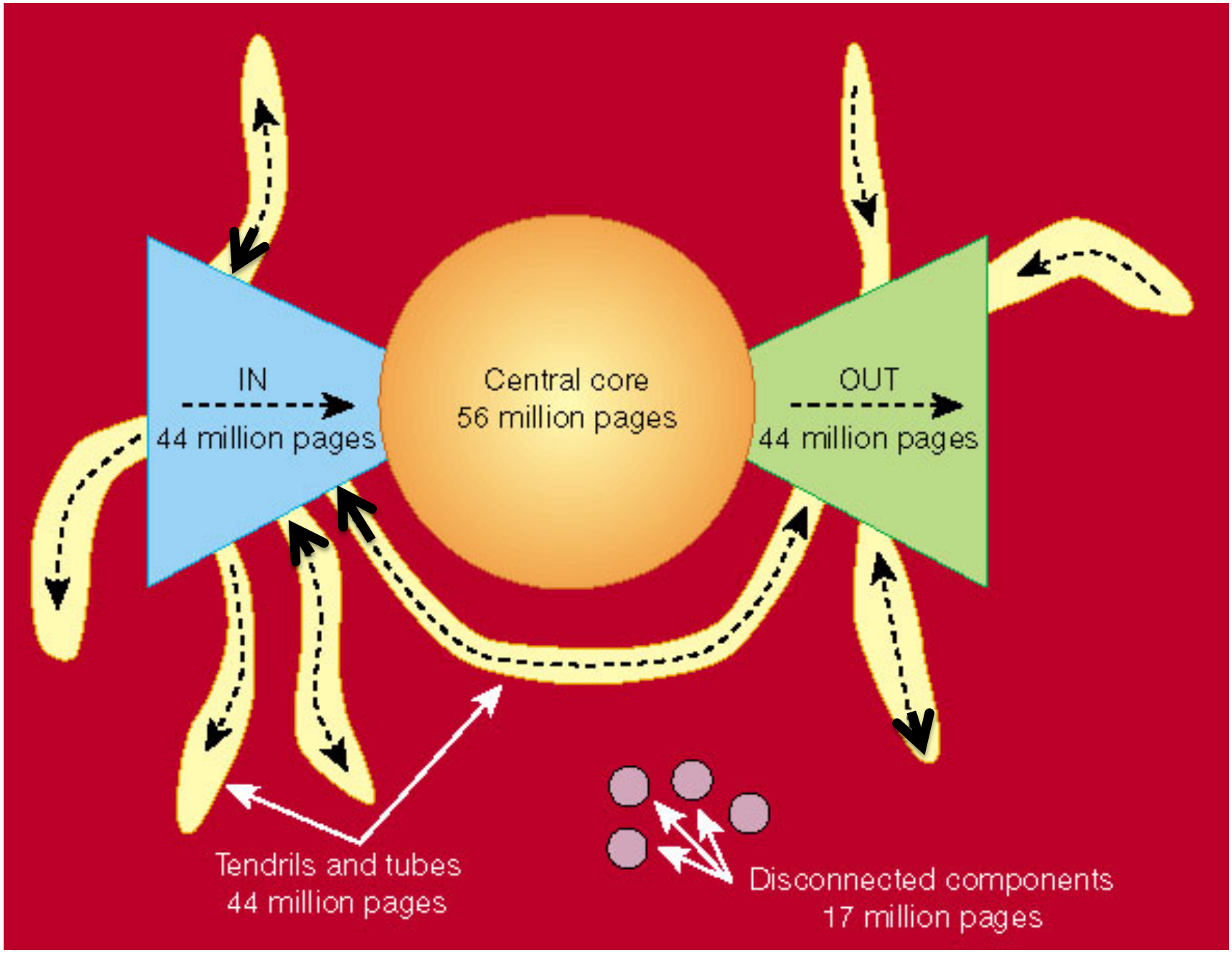
- Interpretation der Suchanfrage (der eingegeben Stichwörter):
  - Suchanfrage in der jeweiligen Session
  - Suchhistorie des Nutzers
  - angeklickte Dokumente
  - Verweildauer auf einem Treffer
  - Suchverhalten vergleichbarer Nutzer

Was kann eine Suchmaschine nicht finden?

Grenzen für die Suchmaschine

- Beschränkung durch die Struktur des WWW: nicht alle Websites sind erreichbar.
- Beschränkung durch die Struktur der Website: nicht alle Dokumente sind erreichbar.

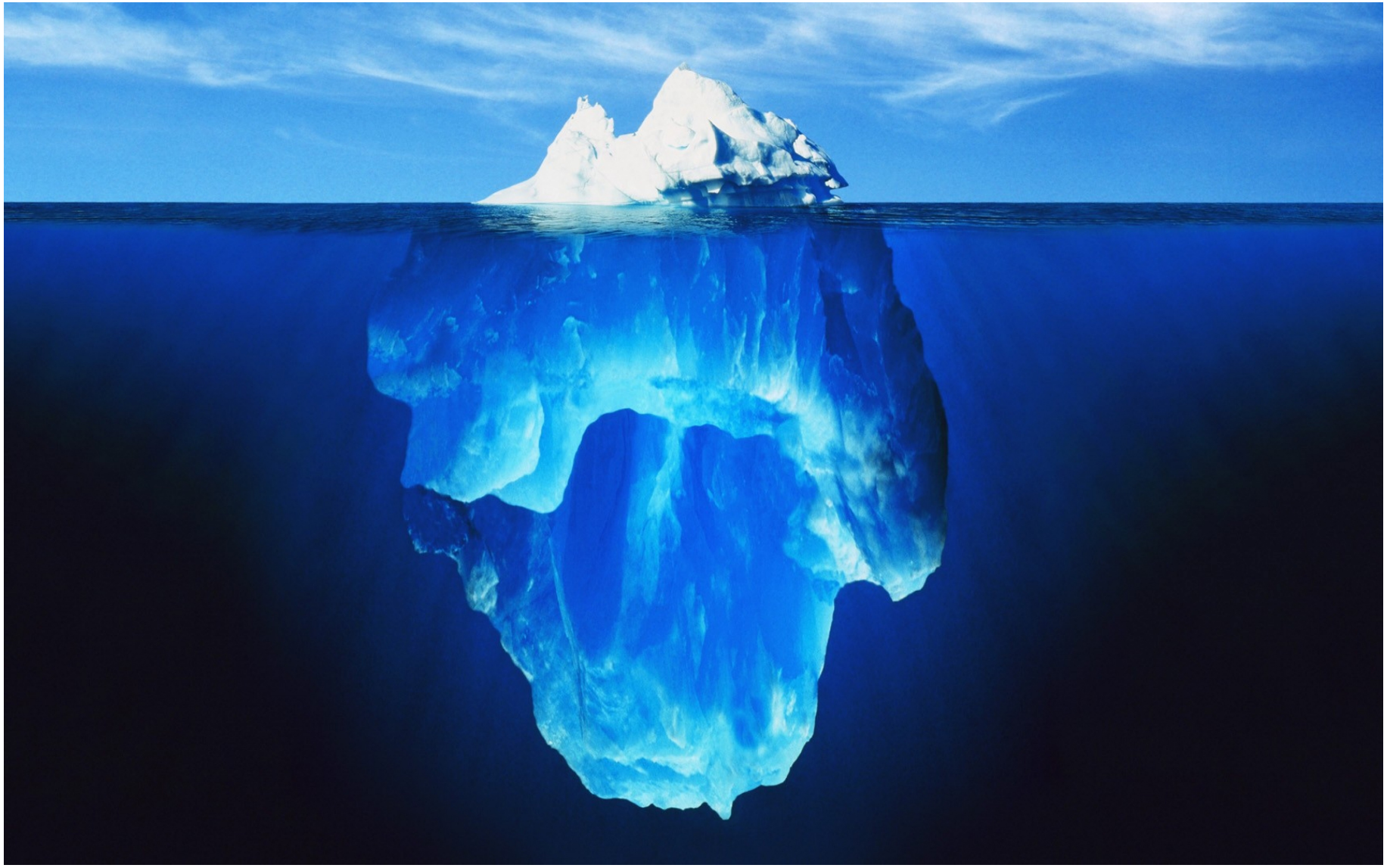
Struktur des WWW



unerreichbare Dokumente

- durch Passwort geschützte Bereiche
  - Beispiel: Intranets
- durch Suchformulare nutzbare Datenbanken
  - Beispiel Telefonbuch
- Social-Media-Dienste
  - Beispiel: Facebook, Twitter
- Inhalte, die für Crawler gesperrt sind
  - gewollter Ausschluss (robots exclusion)
  - Filterung: Zensur oder rechtlich unzulässige Inhalte
- Echtzeitinhalte, die sich ständig verändern
  - Beispiele: Börsenkurse, Wetterdaten

Surface Web und Deep Web



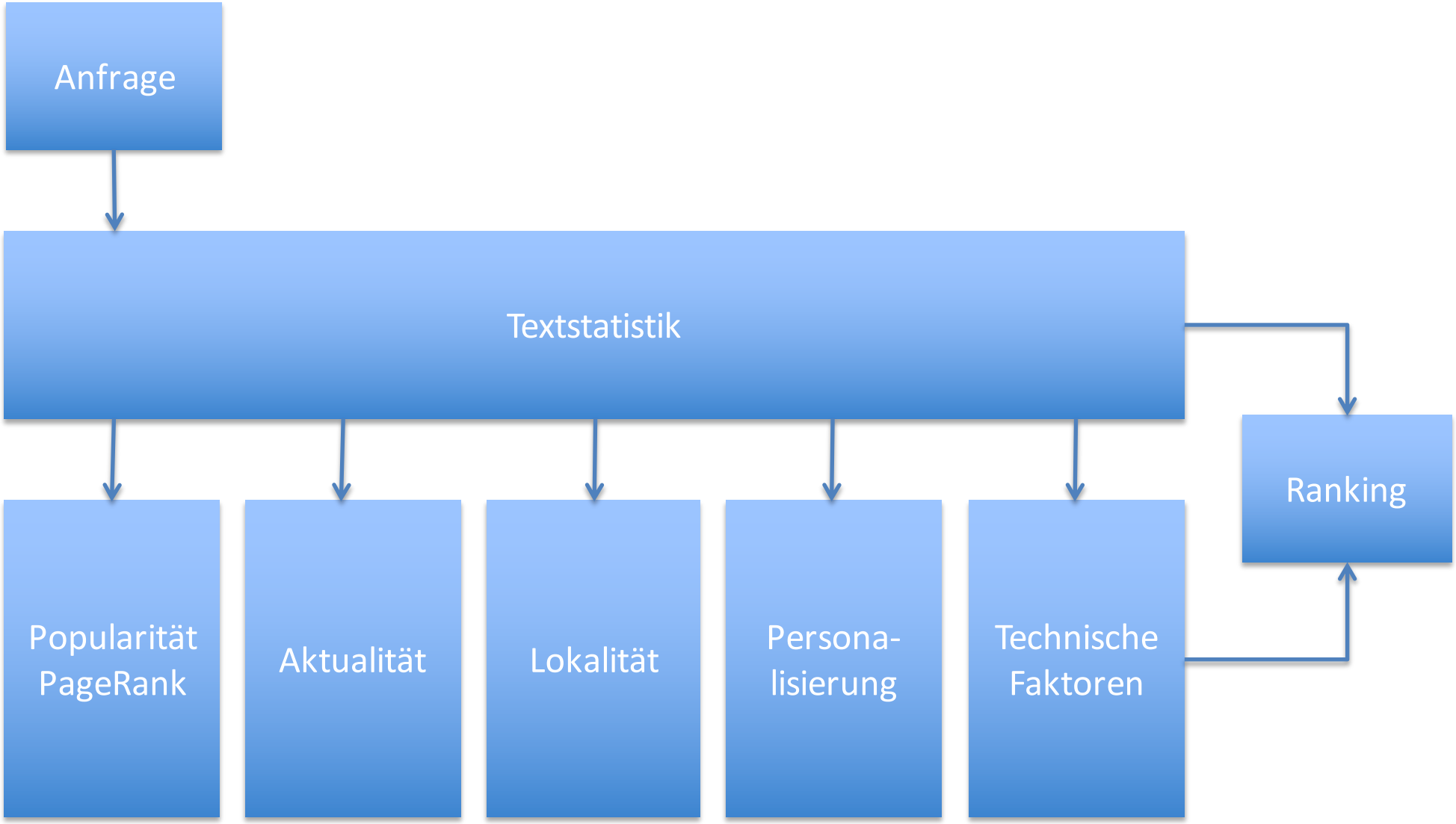
Wie kommt die Trefferliste zustande (Ranking)?

Ranking

- Die Treffer sind nach absteigender Relevanz aufgelistet.
- Die Berechnung des Ranking ist bei jeder Suchmaschine anders und ein Geheimnis (Algorithmus)!
- Das Ranking wird als Kombination und Gewichtung verschiedener Faktoren berechnet (200 bei Google)

shrinktheweb®

Erzeugung der Trefferliste



Ranking-Kriterien

- In das Ranking gehen keine qualitativen, sondern nur quantitative Größen ein:
  - link-statistische Variablen
  - nutzer-statistische Variablen
- Gesponserte Links müssen gekennzeichnet sein.
- Das Ranking kann durch Suchmaschinen-optimierung beeinflusst werden.
- Um das personalisierte Ranking zu erzeugen, speichert die Suchmaschine Daten über den Benutzer, erstellt ein Benutzerprofil.

Diese Daten speichert Google

- die IP-Adresse des Computers
- die Suchanfragen (Suchwörter)
- den Zeitpunkt der Suche
- die angesteuerten Treffer (auch Werbung)
- Verweildauer auf den Treffern

Google- Geschäftsmodell

- Kontextbezogene Werbung zu den Suchwörtern
- Personenbezogene Werbung über angelegte Benutzerprofile
- Verkauf von persönlichen Daten und Nutzerprofilen an Firmen für ihre personenbezogene Werbung

Bewertung der Treffer

- Impressum und Kontaktmöglichkeit vorhanden
- seriöse Institution, z.B. Hochschule, Verlag, Wissenschaftliche Gesellschaft, Qualitätspresse
- Autoren/Autorinnen der Texte sind benannt
- Angabe von Quellen, Belege für Behauptungen
- Verlinkung auf seriöse Sites
- Publikationsdatum und letzte Änderung (last update)
- Vorsicht bei kommerziellen und gesponserten Sites!

Welche Alternativen zu Google gibt es?

Universelle Suchmaschinen

- **Startpage**: Anfragen werden anonymisiert an Google weitergeleitet.
- **Qwant**: französische Suchmaschine mit deutscher Version
- **DuckDuckGo**: „Die Suchmaschine, die Sie nicht verfolgt.“

Meta-Suchmaschinen

- Haben keinen eigenen Index über Crawler angelegt.
- Suchanfragen werden an mehrere andere Suchmaschinen weitergeleitet.
- Treffer aus mehreren Trefferlisten werden neu gerankt.

- Meta-Suchmaschinen ergeben mehr Treffer, aber nicht unbedingt relevantere Treffer!

Meta-Suchmaschinen

- **Ixquik**: keine personenbezogene Daten gespeichert

- **Metager**: deutsche Suchmaschine der Uni Hannover

Spezialsuchmaschinen

- Es gibt Crawler und Suchmaschinen für spezielle Inhalte: Nachrichten, Jobs, Reise, Personen, Bilder, Videos
- Wissenschaftliche Suchmaschinen werten nur wissenschaftliche Dokumente aus:
  - https://scholar.google.de
  - http://www.base-search.net
- Portale/Verzeichnisse für Spezialsuchmaschinen
  - www.suchmaschinen-datenbank.de
    www.suchlexikon.de

Was macht geschicktes und professionelles Suchen aus?

Anfragetypen

- Gezielte Suche nach einem bestimmten Faktum oder Dokument mit klar definiertem Ergebnis.

  Optimal: Der erste Treffer ist richtig

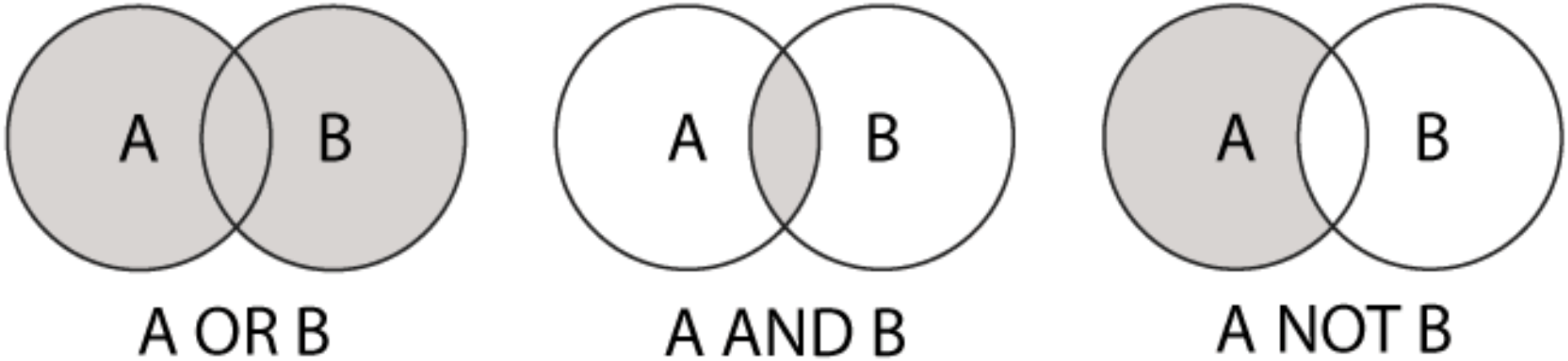
- Recherche zu einem Thema nach bisher unbekannten Informationen ohne klares Ende.

  Optimal: Viele relevante Treffer

Formulieren von Suchanfragen

- Passende Suchwörter
  - nicht zu breite Bedeutung
  - verschiedene Schreibweisen beachten
  - Synonyme beachten
  - Phrasensuche in Anführungszeichen bei Zitaten
- Einengung der Suchanfrage
  - Boolesche Operatoren kombinieren Suchwörter
  - Erweiterte Suche mit Befehlen: Filetyp, Sprache

Suchwort-Verknüpfungen mit Booleschen Operatoren



A OR B	Hund OR Katze	122.000.000 Treffer
A AND B	Hund (AND) Katze	5.800.000 Treffer
A NOT B	Hund NOT Katze	90.300.000 Treffer

(Karotte OR Möhre) (Salat Nüsse) -Koriander

Such-Tipps

- In ausländischen Versionen einer universalen Suchmaschine suchen.
- Gleich eine ausländische Suchmaschine verwenden, z. B. Bing
- Portale und Verzeichnisse sind Linksammlungen zu einem speziellen Thema.
  - Qualitätsprüfung durch ein Redakteursteam
  - nur themenrelevante Treffer
  - Verzeichnis der Verzeichnisse: www.suchlexikon.de

Wie findet man personenbezogene Daten im WWW?

Friedrich Adolf Ballstaedt
* 4.4.1884 in Schrimm
† 13.11.1941 in Stettin

direkte Personensuche

- Phrasensuche mit Vorname und Name in einer universellen Suchmaschine, ev. mit zusätzlichen Stichworten
  - www.google.com
- Phrasensuche mit anderer Länderversion
  - www.google.fr
- Phrasensuche mit ausländischer Suchmaschine
  - https://www.yandex.ru
- Phrasensuche mit anderer Spracheinstellung
  - www.google.de → Einstellungen → erweiterte Suche → Sprache

direkte Personensuche

- Namenseingaben in eine spezielle Personensuchmaschine
  - Liste von sozialen Netzwerken: http://www.yourtraces.com/
- Sammlung von Informationen über lebende Personen ist wegen Datenschutz illegal.

direkte Personensuche

- Suche in Datenbanken, in der eine Person vermutet wird (teilweise registrier- und/oder kostenpflichtig)
- Genealogische Datenbanken
  - https://familysearch.org/family-trees
  - http://gav.azurewebsites.net
- Verzeichnisse/Portale von Datenbanken
  - https://de.wikipedia.org/wiki/Genealogie-Datenbank
  - http://wiki-de.genealogy.net/Genealogische_Datenbank/Linkliste
  - http://www.heimatarchiv.de/genealogie/webseiten.html
- Telefonbücher; Branchenverzeichnisse etc.

direkte Personensuche

- Genealogische Meta-Suchmaschine
  - http://beuss.spdns.de/Crawl/

AKVZ
AncestryTrees
Arkivverket
CG_BremerPL
CG_Verlustlisten
CompGen
CultureFr
Denkmalprojekt
DeutscheBiographie
FTS
Familienanzeigen
FamilySearchRecords
FamilySearchTree
FindAGrave
FreeBMD
GedBas
GenDexNetwork
GenTeam
GeneaNet
Geni
GrabstOstfriesland
Grabsteine
MyHeritage
Other
Roglo
RootsWeb
RyersonIndex
TPO
Volksbund
Werelate
WieWasWie
WikiTree
WikipediaPers

indirekte Personensuche

- Suche nach Dokumenten, in denen die Person vermutet wird.
  - Kirchenbuchportale:
  - www.archion.de
  - http://search.ancestry.de/search/db.aspx?dbid=2116
  - Online-Archive:
  - https://archive.org/details/genealogy&tab=about
  - https://archivalia.hypotheses.org/category/genealogie
- In den meist eingescannten Dokumenten ist keine Volltextsuche möglich.

Fazit

Die genealogische Forschung verlegt sich immer mehr ins Web:

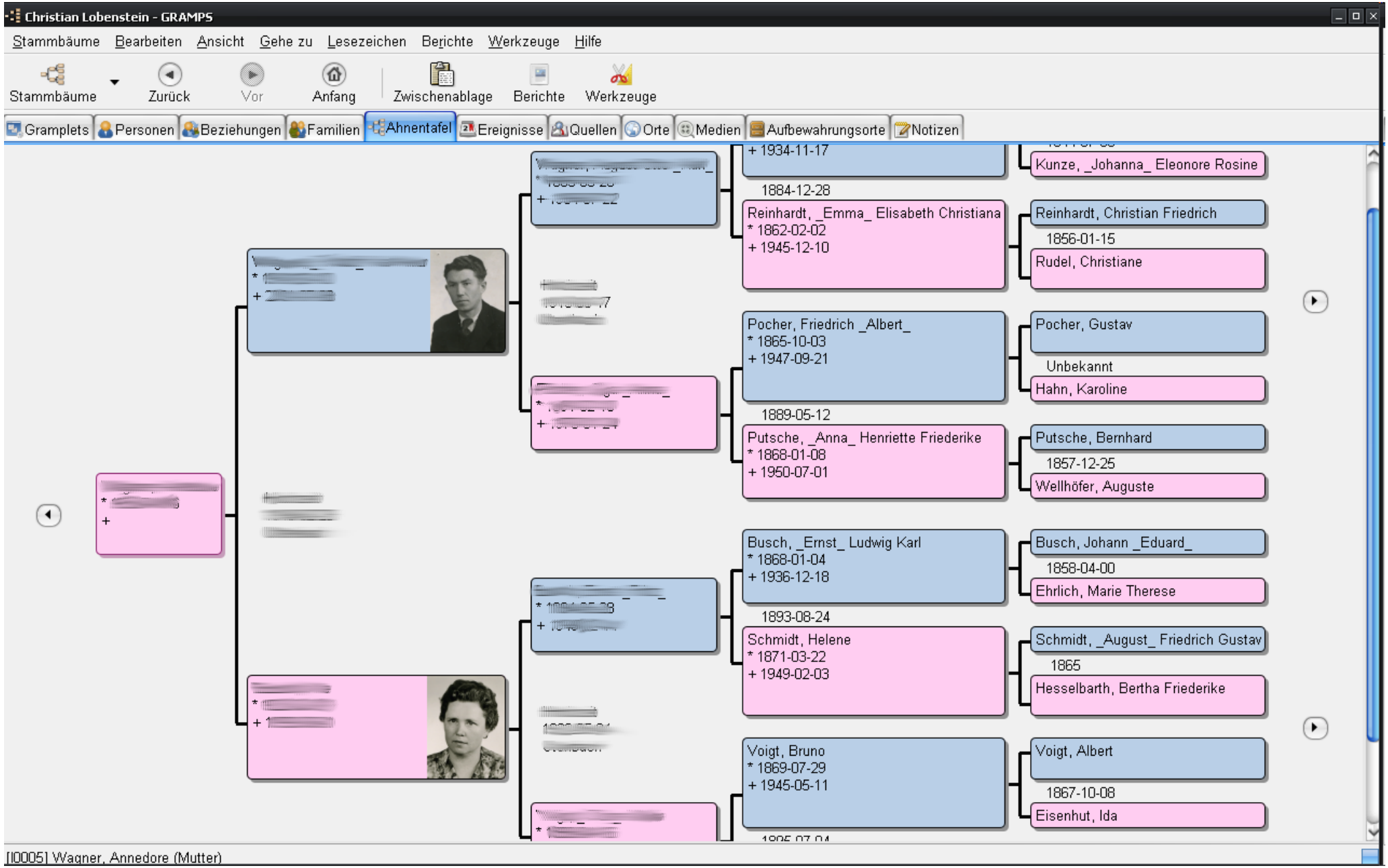
- Erschließung: Primärquellen werden digitalisiert und im Web zur Verfügung gestellt.
- Bearbeitung der Daten erfolgt mit Genealogie-Programmen, z.B. Gen_Plus.
- Die Zusammenführung der Daten findet weltweit über das Web statt. Standard: meist GEDCOM.

Fazit

- Forschungsergebnisse werden immer häufiger im Web präsentiert, da eine Print-Veröffentlichung teuer ist.
- Die Visualisierung der Befunde z.B. in Stammbäumen erfolgt durch spezielle Programme, z. B. GRAMPS.

Schwer durchschaubarer Wildwuchs an Angeboten: Die Grenze zwischen laienhafter Aufarbeitung und wissenschaftlicher Forschung verwischen.

Visualisierung durch GRAMPS



Organisationen

- Eine deutsche Organisation zur Computergenealogie wurde 1989 gegründet.

- Ein Ziel: die EDV-gestützte standardisierte Erschließung, Bearbeitung und Veröffentlichung genealogischer Quellen.

Organisationen

- Family History Information Standards Organization, seit 2013.

- „organisation established for the purpose of developing genealogy and family history information standards on a modern platform“

Vielen Dank für Ihr Interesse!

www.ballstaedt-kommunikation.de ▶ **Downloads**