BINGO!
Ein fokussierender Crawler zur Generierung personalisierter Ontologien

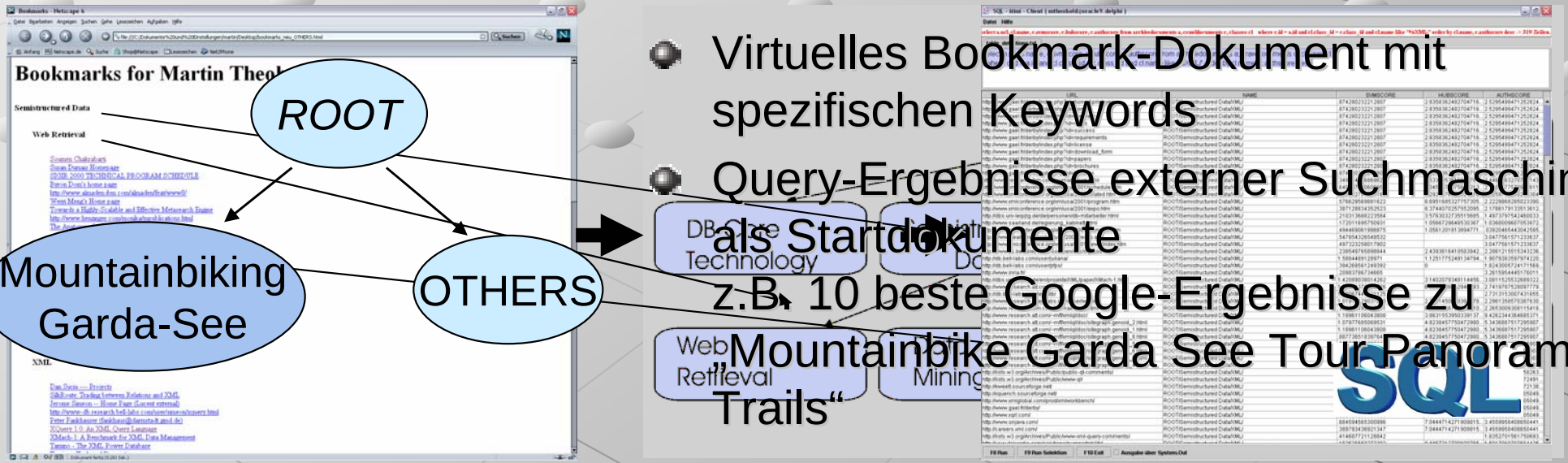
Martin Theobald
Stefan Siersdorfer, Sergej Sizov

Universität des Saarlandes
Lehrstuhl für Datenbanken und Informationssysteme
Prof. Dr.-Ing. G. Weikum

2. Oktober 2002

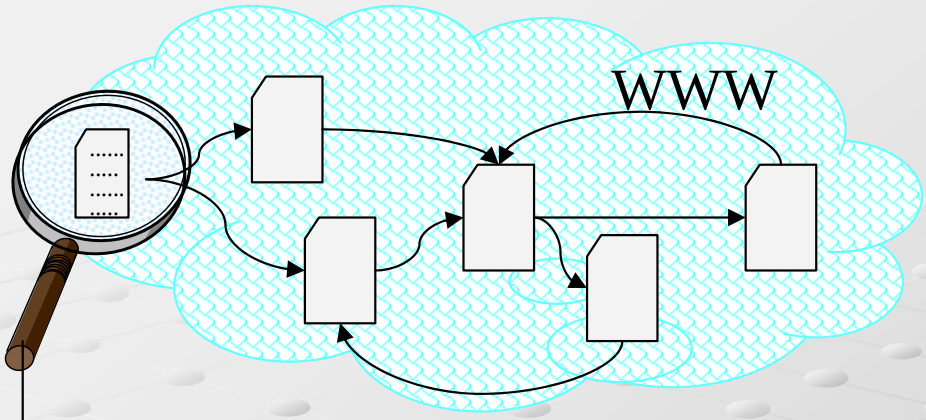
BINGO!
„Bookmark-Induced Gathering of !nformation“

- Fokussierendes Crawling zur Ontologiegenerierung
  - Aufbau und Erweiterung benutzerdefinierter Ontologien über spezifischen Themenhierarchien („*Yahoo-Style*“)
  - Automatisches Aktualisieren und Filtern hierarchischer Themenstrukturen unter Ausnutzung unterschiedlicher Relevanzkriterien (*SVM-Konfidenz, Autorität, Cosinus-Maß*)
  - Automatisierte Bearbeitung von Expertenqueries bzw. Vervollständigung vorhandener Ergebnismengen („Nadel im Heuhaufen“)

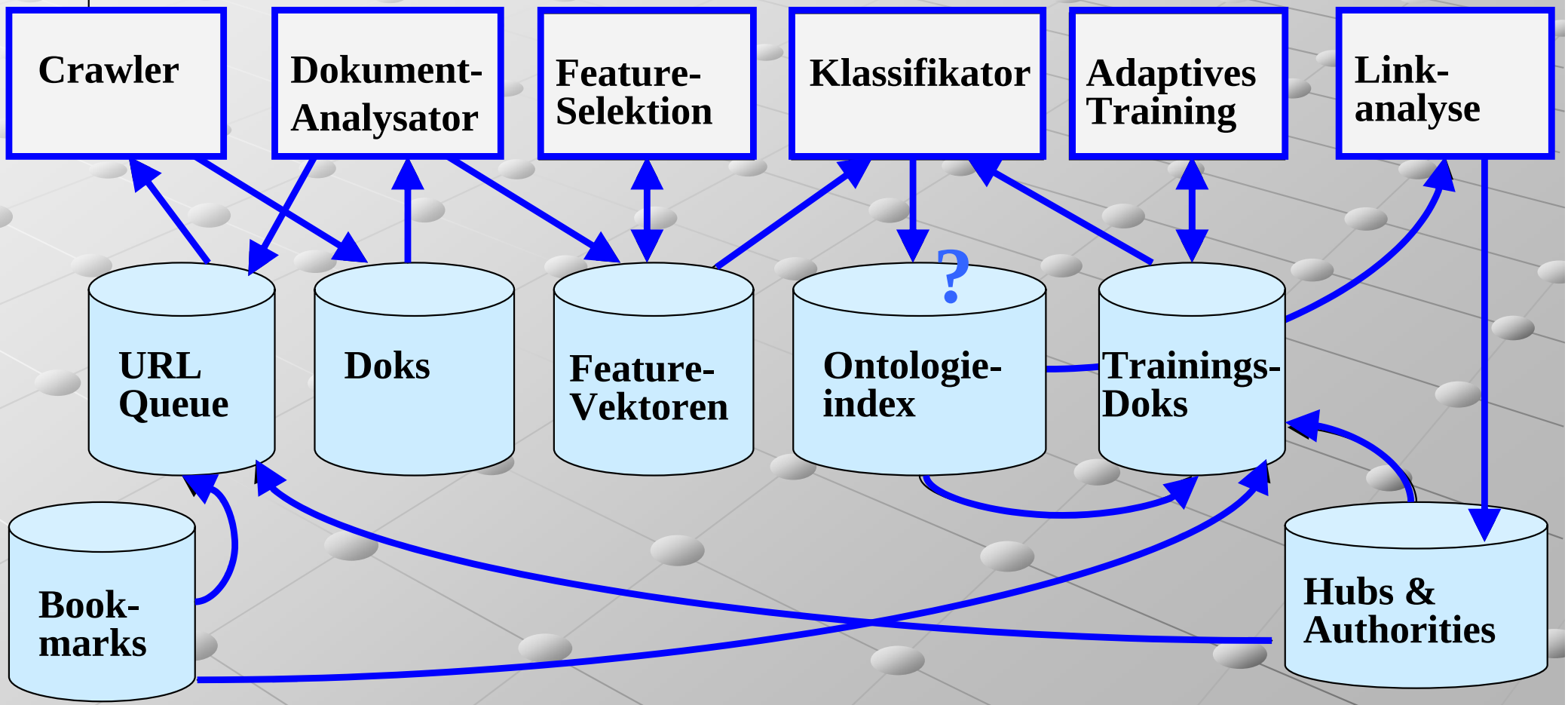


- Virtuelles Bookmark-Dokument mit spezifischen Keywords
- Query-Ergebnisse externer Suchmaschine als Startdokumente z.B. 10 beste Google-Ergebnisse zu „Mountainbike Garda See Tour Panorama Trails“

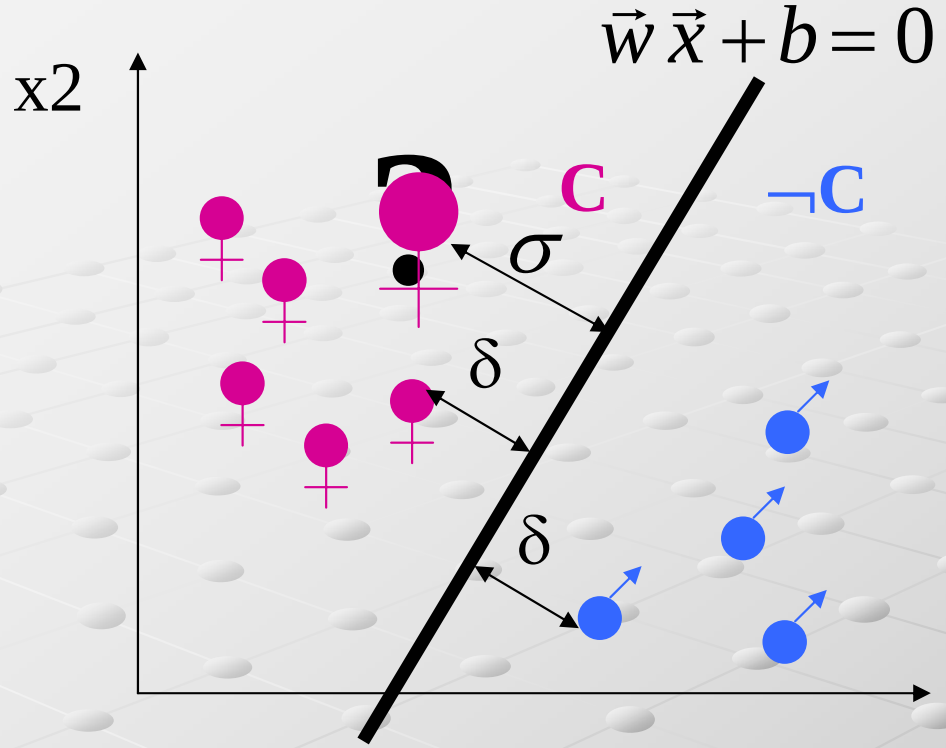
Überblick über den System-Aufbau



Fokussierendes Crawling mit adaptivem Neu-Training auf „Archetypen“



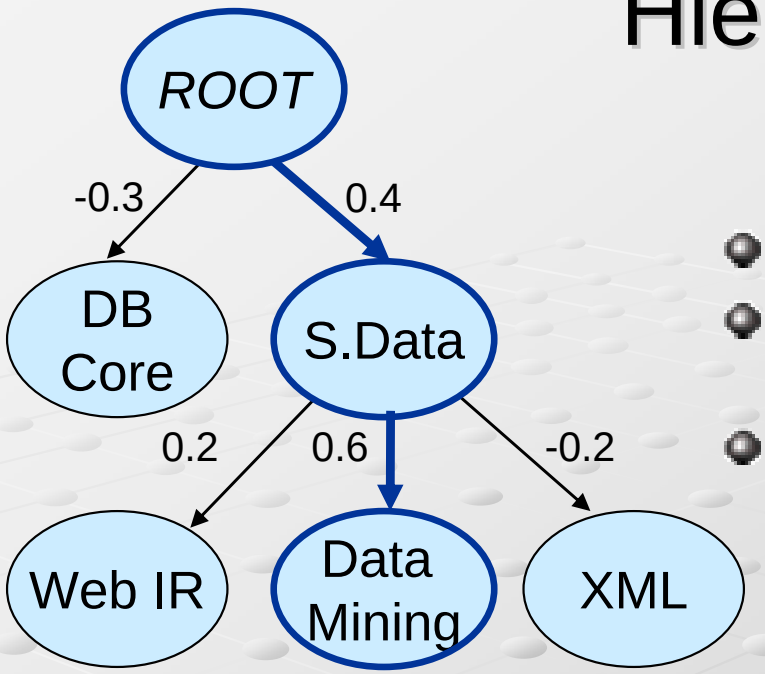
Klassifikation mit Support-Vector-Machines (SVM)



n Trainingsvektoren mit Komponenten $(x_1, \ldots, x_m, C)$ und $C = +1$ oder $C = -1$

- *Training:* Berechne trennende Hyperebene $\vec{w}\vec{x}+b=0$, die die Positiv- von den Negativbeispielen mit maximalem Abstand δ trennt.
  - Löse quadratisches Optimierungsproblem
- *Klassifikation:* Teste unbeschrifteten Vektor *y* auf Lage zur Hyperebene
  - Skalarprodukt $(\vec{w}\,\vec{y}+b)=\sum_{i=1}^{m} w_i y_i + b > 0$ (SVM-Klassifikations-Konfidenz)
  - Sehr effiziente Laufzeit $O(m)$ linear zur Anzahl *m* der Terme in X_i

Hierarchische Klassifikation und Feature-Selektion



- Rekursive Klassifikation entlang der Hierarchie
- Enscheidung basiert auf klassenspezifischen Feature-Räumen
- Beispiel:
  deadlock, recovery, pattern, hypertext
  gut für *DB Core* gegen *Semistructured Data*
  schlecht für *Data Mining* gegen *XML*

- Knotenspezifische Bereinigung der Feature-Vektoren durch Bestimmung der *m* besten Diskriminatoren nach *MI* (Mutual Information bzw. Kullback-Leibler-Distanz)

$$MI(X_i, c_j) := P[X_i \wedge c_j] \log_2 \frac{P[X_i \wedge c_j]}{P[X_i]P[c_j]}$$

- Beste Diskriminatoren für *Data Mining* gegen *Web IR* und *XML* (*m = 200*):

  mine, knowledge, OLAP, pattern, discover, cluster, dataset …

- Termgewichtung nach TF/IDF

Link-Analyse nach Kleinberg‘s HITS Algorithmus

Für einen Webgraphen G=(V,E) und eine themenspezifische Basis B $\subset$ V

finde gute *Authorities* mit Gewichtung

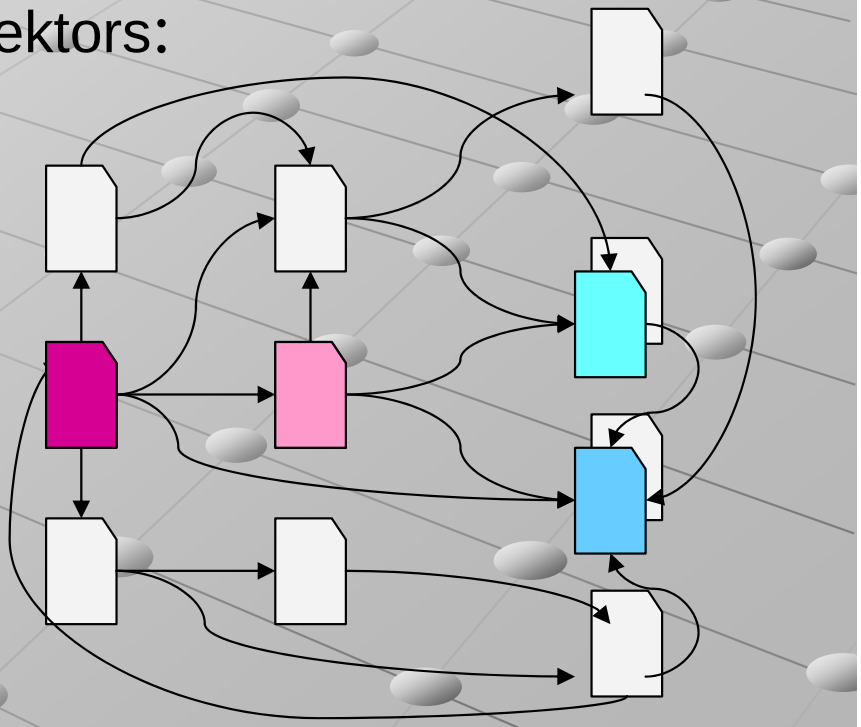
$$x_p = \sum_{q:(q,p)\in E} y_q$$

und gute *Hubs* mit Gewichtung

$$y_p = \sum_{q:(p,q)\in E} x_q$$

Iterative Approximation des dominanten Eigenvektors:

$$\left.\begin{aligned}\vec{x} &= A^T \vec{y}\\ \vec{y} &= A\vec{x}\end{aligned}\right\} \Rightarrow \begin{aligned}\vec{x} &:= A^T \vec{y} := A^T A\vec{x}\\ \vec{y} &:= A\vec{x} := AA^T \vec{y}\end{aligned}$$



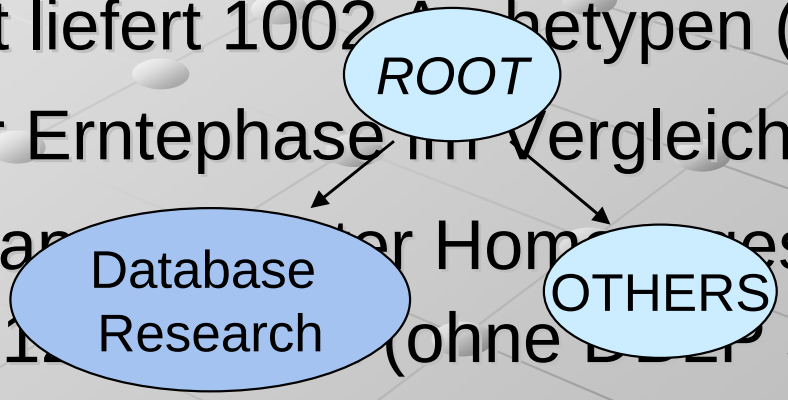
Adaptives Neu-Trainieren auf Archetypen

- *Wachtumsphase:*
  - Iteratives Neutrainieren des Klassifikators ausgehend von Bookmarks und hochwertigen Nachbardokumenten
  - Identifikation von Archetypen:
    - *beste Doks nach SVM-Konfidenz ∩ beste Authorities*
    - *SVM-Konfidenz > mittlere Konfidenz der Bookmarks*
  - Verhindere „Topic-Drift“!
  - Harte Fokussierung des Crawlers:
    - *Akzeptiere nur solche Links (p,q) mit class(p) = class(q)*
- *Erntephase:*
  - Massencrawl nach erweiterter Trainingsbasis
  - Schwache Fokussierung mit erhöhter Ausbeute & Präzision
    - *Akzeptiere Links (p,q) mit class(q) != ROOT/OTHERS/*

Experimentelle Evaluation (I)

- *Portalgenerieung für ein Einzelthema:*
  - Finde möglichst viele Webseiten zu „Database Research“
  - Einzige Quellen: Homepages von *David DeWitt & Jim Gray* gegen 400 Negativbeispiele aus Yahoo Top-Level-Kategorien als initiales SVM-Modell
  - Wachstumsphase mit Crawlingtiefe 4 beschränkt auf den Ausgangshost liefert 1002 Archetypen (inkl. PDF, Word)
  - ***Ausbeute*** der Erntephase im Vergleich zur DBLP-Trier:

    ca. 72% Überlap[illegible]er Hom[illegible]es der Top 1000 Autoren nach [illegible] (ohne DBLP selbst zu crawlen!)



  - ***Präzision:*** ca. 27% der 1000 besten DBLP-Autoren unter 1000 besten nach SVM-Konfidenz

Experimentelle Evaluation (II)

- *Expertensuche*
  - Suche: „*public domain open source implementations of the ARIES recovery algorithm*“ (Shore, MiniBase & Exodus)
  - Keine brauchbaren Ergebnisse unter den Top 10 Google Ergebnissen oder Open-Source Portalen wie sourceforge.net
  - Manuelle Auswahl von 10 Startdokumenten aus Google-Queries zu „*aries recovery algorithm*“ und „*aries recovery method*“ gegen zufällig gewählte Yahoo Top-Level-Kategorien
  - Massencrawl liefert 17.000 URLs mit 2.167 Dokumenten in Bereich „ARIES“ innerhalb von 10 min.
  - Schlüsselwortsuche nach Cosinus-Maß für „*source code release*“ liefert Links zu den Open-Source Projekten „Shore“ und „MiniBase“ unter den Top 10, „Exodus“ wird direkt auf der Shore-Homepage referenziert

Zusammenfassung

- *BINGO! integriert unterschiedliche Techniken des Web-IR wie SVM, MI, HITS mit der Identifikation von Archetypen und adaptivem Neu-Training*
- *Umfassendes und vielseitiges Werkzeug auf dem Weg zu einer neuen Generation der individualisierten Web-Suche / Information-Mining*
- *Erweiterung um einen auf Web-Services basierenden Portal-Explorer mit semantisch reicherem Ontologie-Service*
- *XML: Feature-Generierung, Klassifikation, XPath-Queries…*