Data Integration: From the Enterprise into Your Kitchen

Alon Halevy
C.E.O.
Recruit Institute of Technology

May 14, 2017

Joint work with: *Wang-Chiew Tan, George Mihaila, Behzad Golshan*

Outline

- Throwback to the 90's:
  - Answering queries using views: what and why
- Broader view: data integration
  - Still, mostly 90's
- Data integration today
  - The Web and data lakes
- Two data integration challenges
  - Might sound familiar

not a survey

The View Rewriting Problem

Given a query *Q* and a set of view definitions $V_1,\ldots,V_n$:

Is it possible to find a rewriting of Q using only the V's?

V_1(A,B) :- cites(A,B), cites(B,A)
V_2(C,D) :- sameTopic(C,D), cites(C,C1), cites(D,D1)

Query:
q(x,y) :- sameTopic(X,Y), cites(X,Y), cites(Y,X)

Query rewriting: q'(X,Y) :- V_1(X,Y), V_2(X,Y)

The View Rewriting Problem

Given a query Q and a set of view definitions $V_1,\ldots,V_n$:

Is it possible to find a rewriting of Q using only the V's?

V_1(A,B) :- follows(A,B), follows(B,A)
V_2(C,D) :- FBfriends(C,D), follows(C1,C), follows(D1,D)

Query:
q(X,Y) :- FBfriends(X,Y), follows(X,Y), follows(Y,X)

Query rewriting: q'(X,Y) :- V_1(X,Y), V_2(X,Y)

An Equivalent Query Rewriting

Query:

q(X,Y) :- FBfriends(X,Y), follows(X,Y), follows(Y,X)

Query rewriting: q'(X,Y) :- V_1(X,Y), V_2(X,Y)

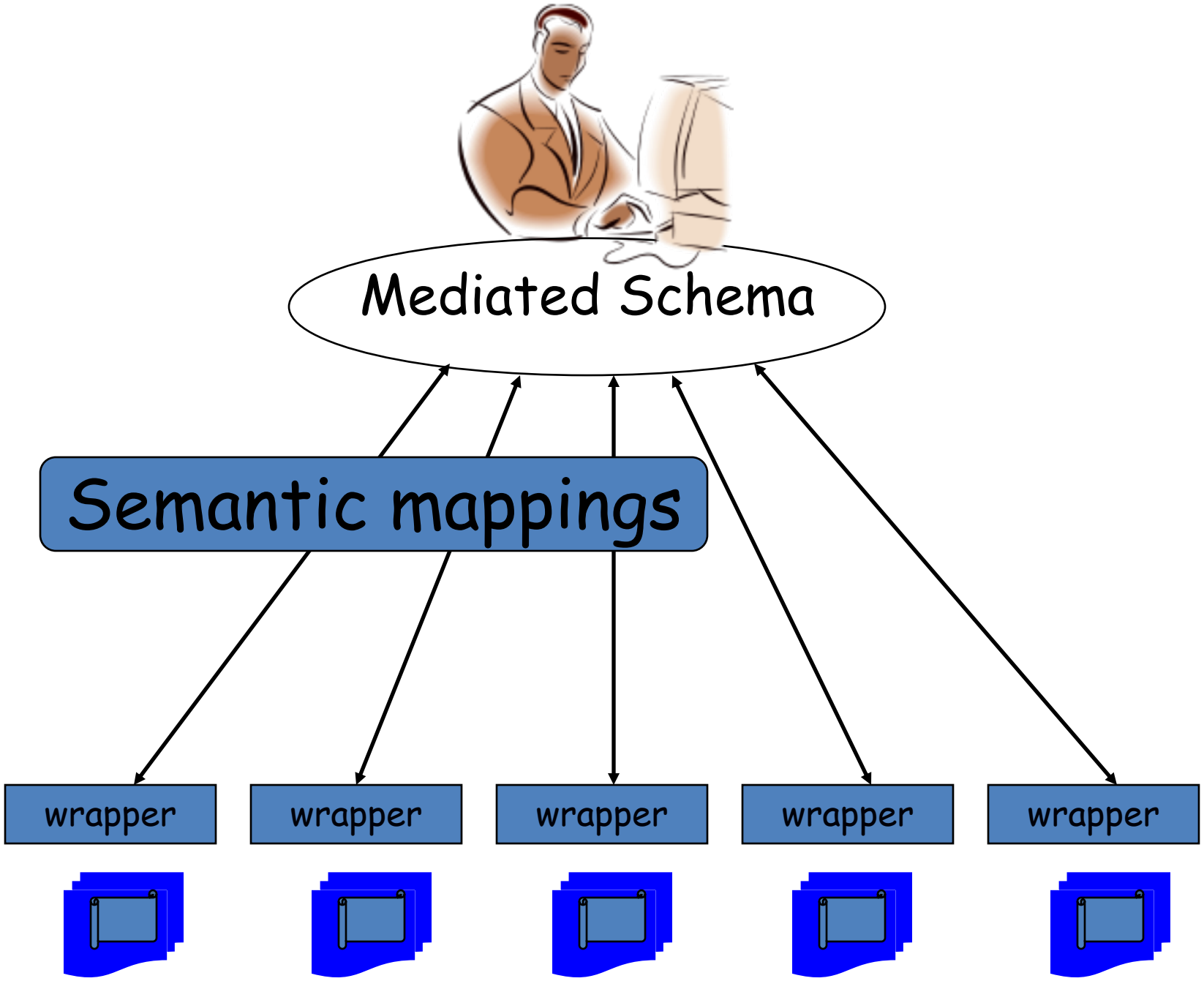
Unfolding of the rewriting:

q''(X,Y) :- follows(X,Y), follows(Y,X),
FBfriends(X,Y),
follows (Z,X), follows(W,Y)

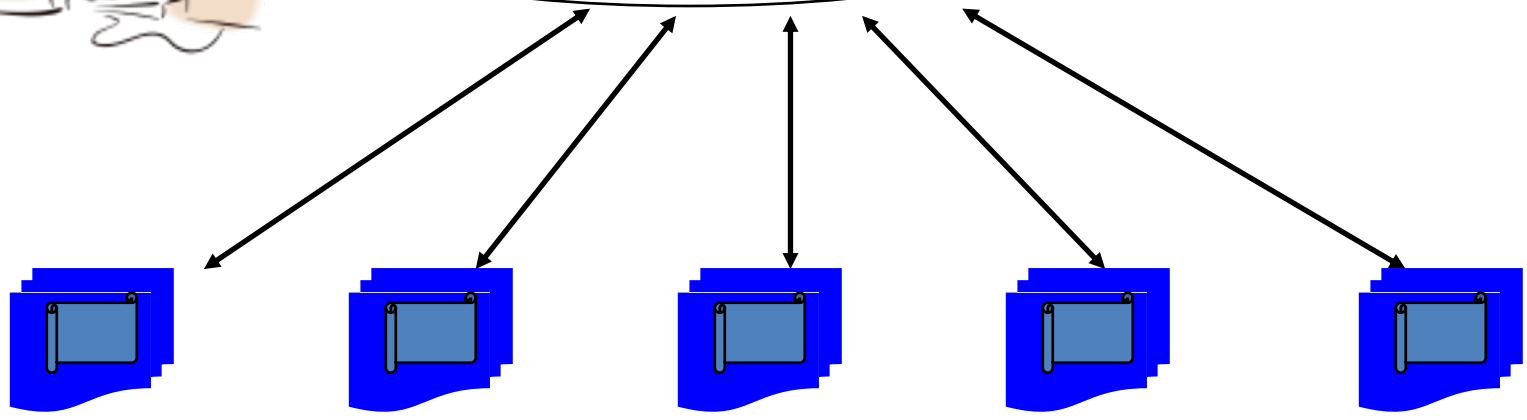
The unfolding is *equivalent* to the original query.

Why Do We Care?

- Query optimization
- Data warehouse design
  - Semantic data caching
- Bridging between storage schema and logical schema
- *Data citation*
- **Data integration**


Mediated Schema
Semantic mappings
wrapper
wrapper
wrapper
wrapper
wrapper

Item(name, category, zip, year)
Review(name, category, review)



S1(name, zip, year) :-
Item(name, category, zip, year),
Category="auto"

S2(name, review) :-
Item(name, category, zip, year),
Category="auto", zip = 94024,
year < 1950

Local As View

- Describe each source as a view
  - Answering a query *requires* view rewriting
  - Unlike GAV, the rewriting algorithm does the hard work of combining sources
- Rewriting may not be equivalent to the query
  - Data sources may not be available or complete
- Goal: find *maximally-contained* rewriting

Q' is a maximally-contained rewriting of Q w.r.t. ***L*** using the V's if there is no other Q'' such that:

Q'' strictly contains Q', and Q'' is contained in Q.

A Basic Decidability Result

Q is a query, *V* is a set of views

- If *Q* has built-in predicates and *V* doesn't, then deciding if there is a rewriting of *Q* using *V* is NP-complete.
- If *V* also has built-in predicates, then the problem is Π^P_2-complete.

Incomplete Info/ Certain Answers

View extensions represent partial information:

Extensions of the views $v_1,\ldots v_n$ define a set of databases ***D*** that are consistent $v_1,\ldots v_n$.

The tuple *t* is a *certain answer* to a query Q if it would be an answer in every database in ***D.*** *[Abiteboul & Duschka]*

- An equivalent rewriting provides *all* certain answers
- A maximally-contained rewriting provides *only* certain answers

A Fertile Ground for Exploration

- View definition language & query language
- Equivalent or maximally contained rewriting
- Semantic constraints (e.g., FD's, inclusions)
- Completeness/soundness of the views
- Binding patterns restrictions
- Data exchange

Outline

- ✓ Throwback to the 90's:
  - ✓ Answering queries using views: what and why
- ➢ Broader view: data integration
  - ➢ Still, mostly 90's
- Data integration today
  - The Web and data lakes
- Two data integration challenges
  - Might sound familiar

The Data Integration Universe

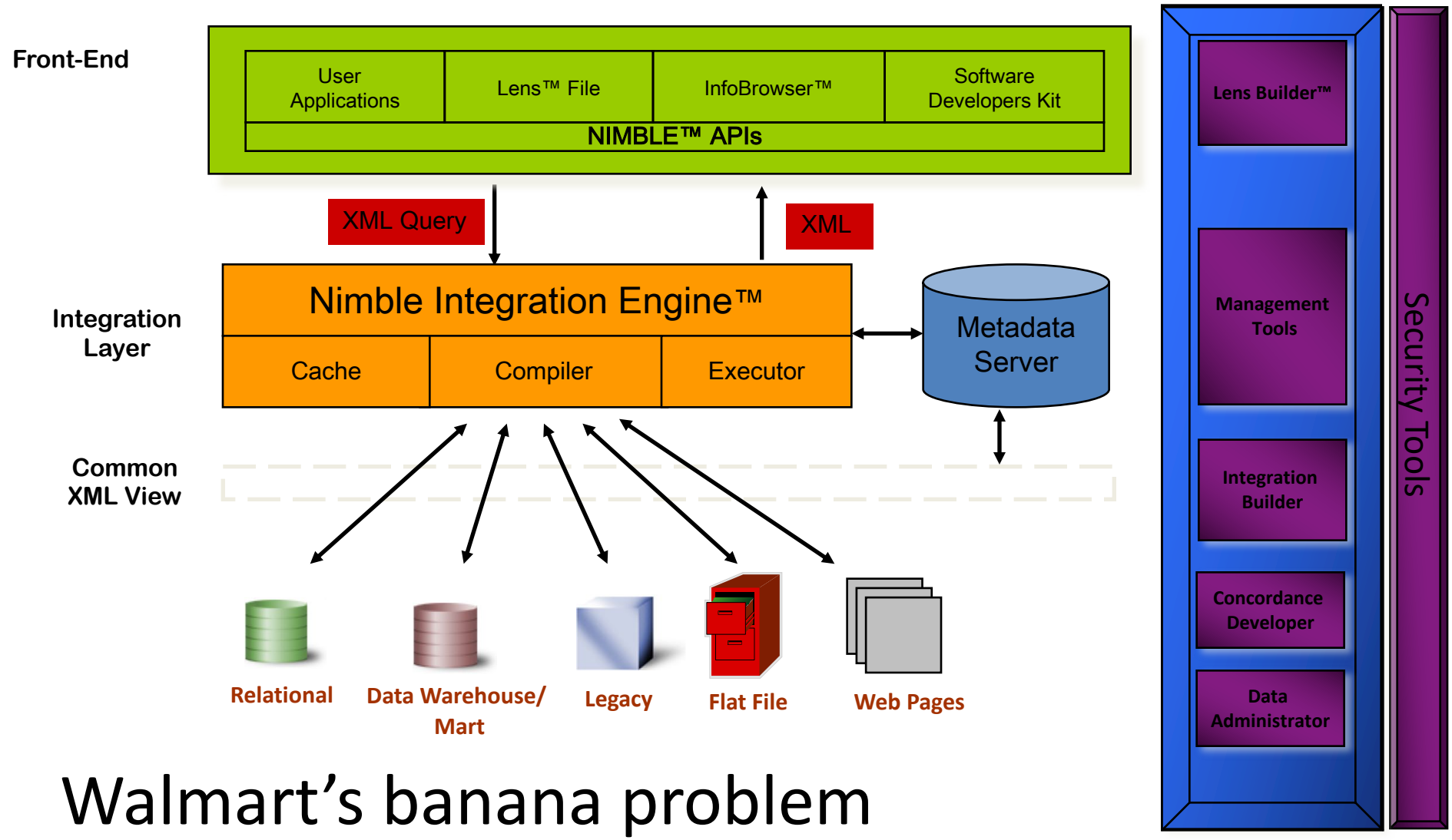
Creating semantic mappings [Clio, LSD, model management]

Language for semantic mappings and reformulation algorithms

Reference reconciliation: "Alon Halevy" = "Alon Levy"?

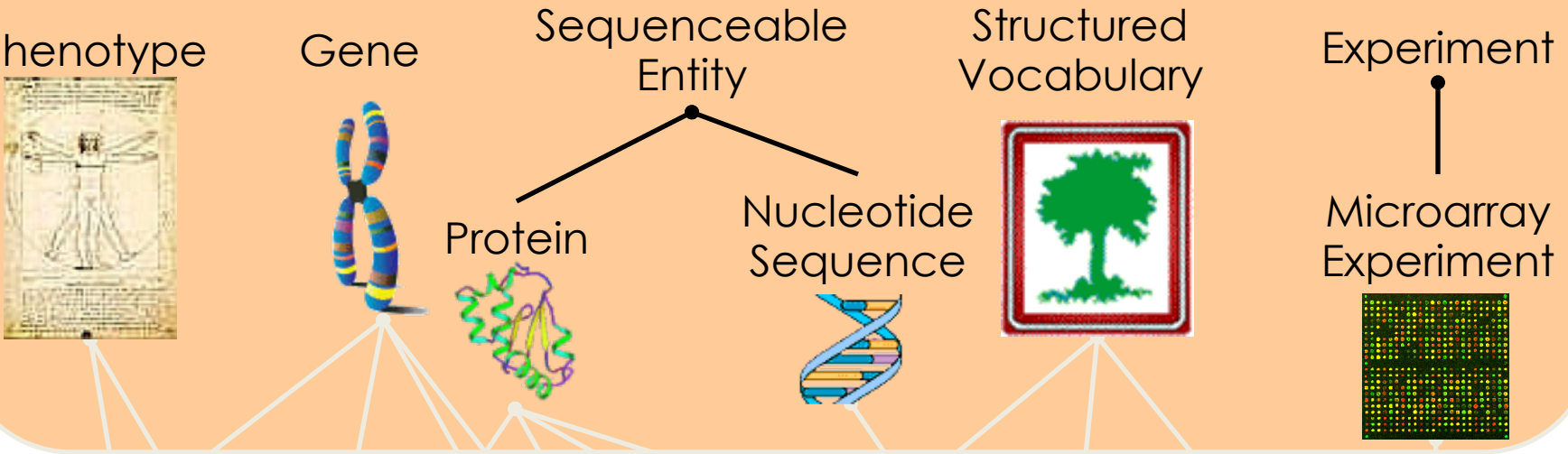
Query optimization and execution: Eddies, adaptive execution

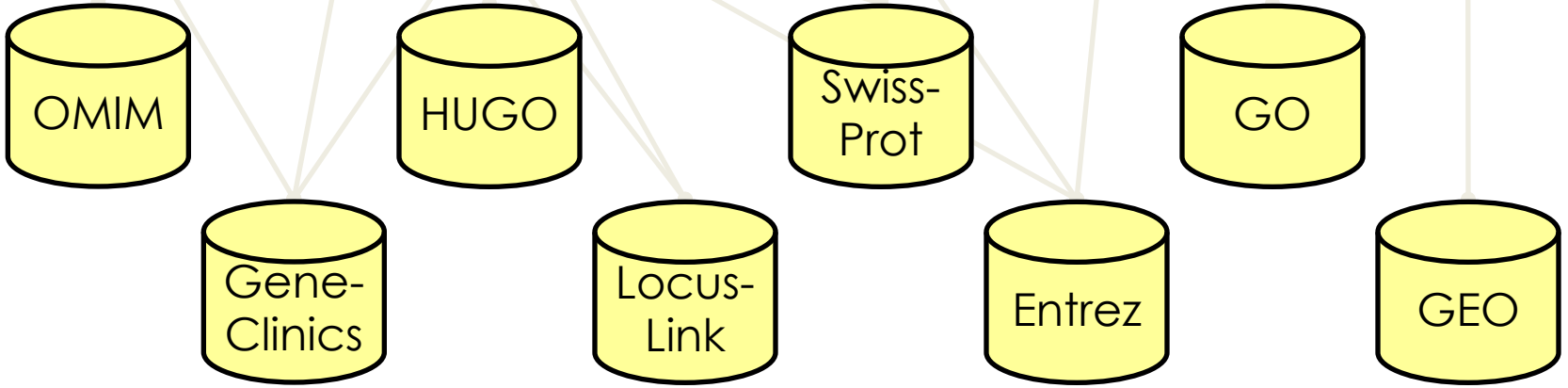
Enterprise Data Integration ($30M Slide)



Walmart's banana problem

Integration of Scientific Data





Data Integration Assumptions: 90's

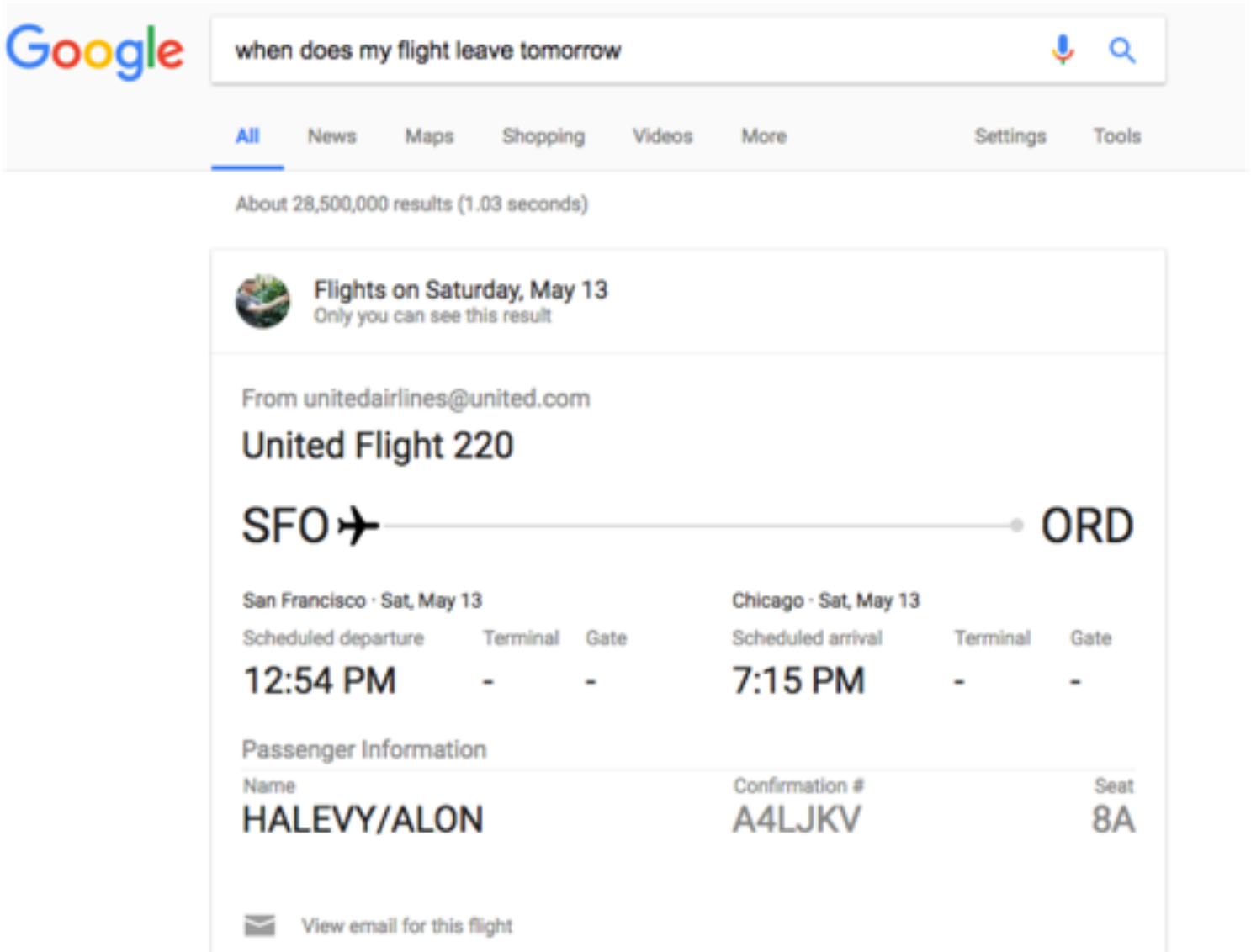
- Mediated schema is of reasonable size
- Data sources are mostly structured
- We *need* to integrate all the sources
- Data is mostly correct and consistent (modulo cleaning)

All these assumptions are reasonable when the number of sources is in the 10's.

Outline

- ✓ Throwback to the 90's:
  - ✓ Answering queries using views: what and why
- ✓ Broader view: data integration
  - ✓ Still, mostly 90's
- ➢ Data integration today
  - The Web and data lakes
- Two data integration challenges
  - Might sound familiar

Today: Data Integration Powers Personal Assistants



And It’s In Your Kitchen & Bedroom

The data integration vision of the 90’s

More “views” Are Being Added

Check out: doneList on YouTube.

Speech recognition and natural language understanding

Creating semantic mappings [Clio, LSD, model management]

Language for semantic mappings and reformulation algorithms

Reference reconciliation:
"Alon Halevy" = "Alon Levy"?

Query optimization and execution:
Eddies, adaptive execution

The Landscape Changed

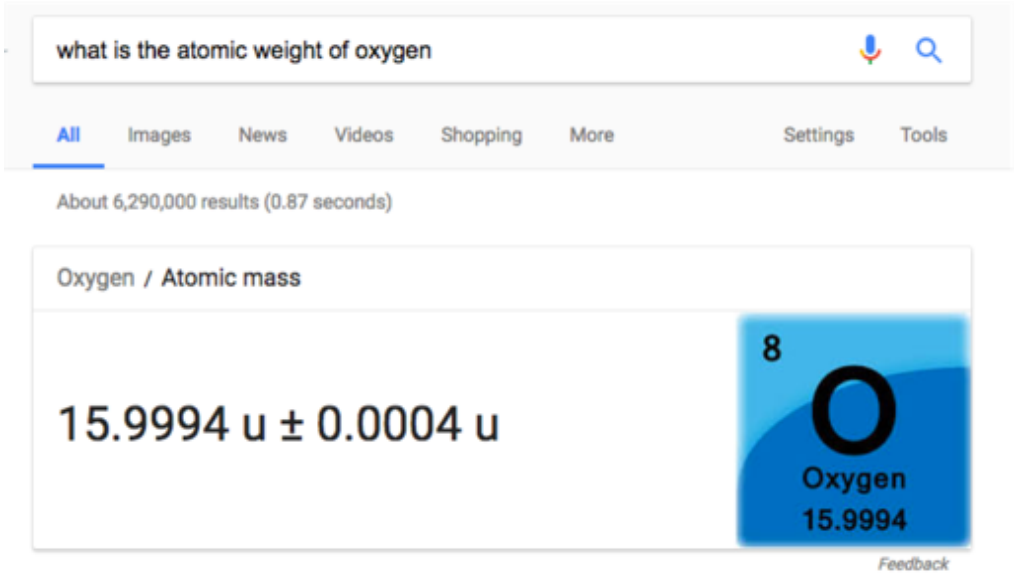
- The Web (& personal agents)
- Enterprise data lakes
- Data science

100's of millions of data sources:

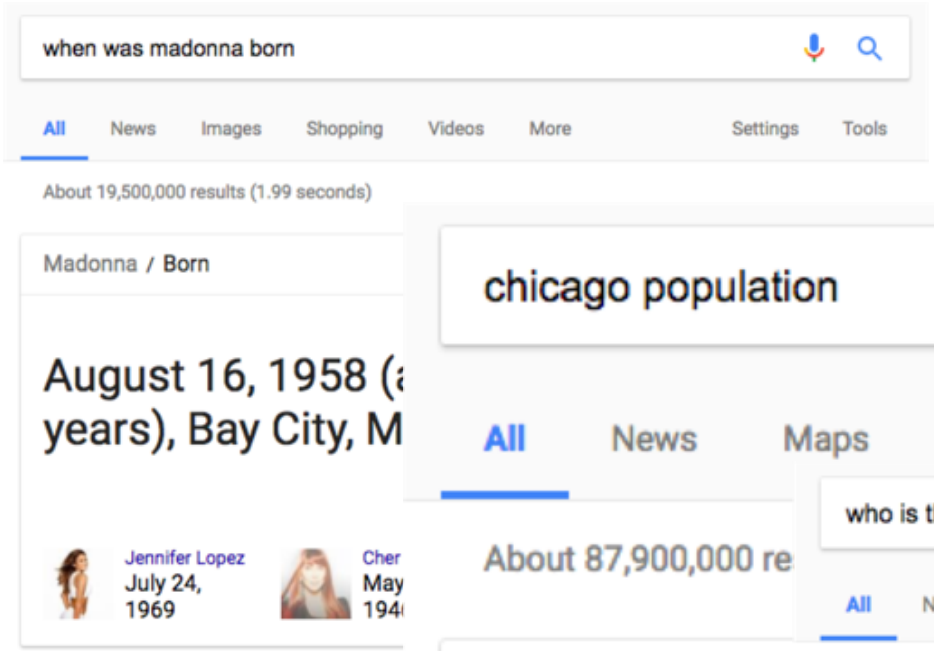
- *Mediated schema* is merely an approximation
- Data sources are of all kinds
- We *cannot* integrate all the sources
- Data is mostly correct and consistent
  - Still valid! Everything on the Web is true, right?

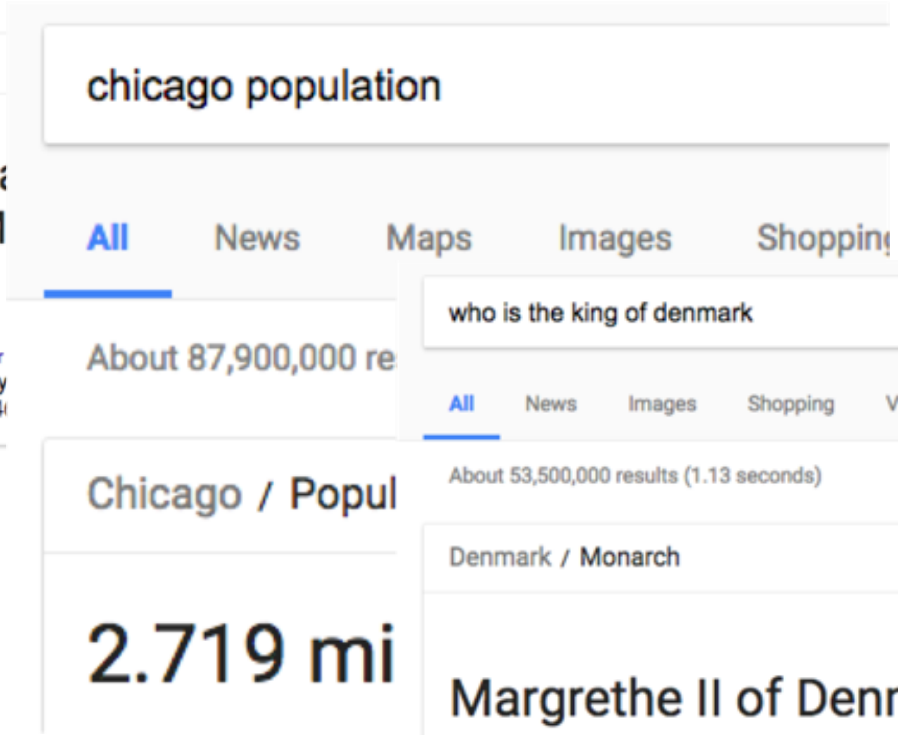
Integration in Search Engines

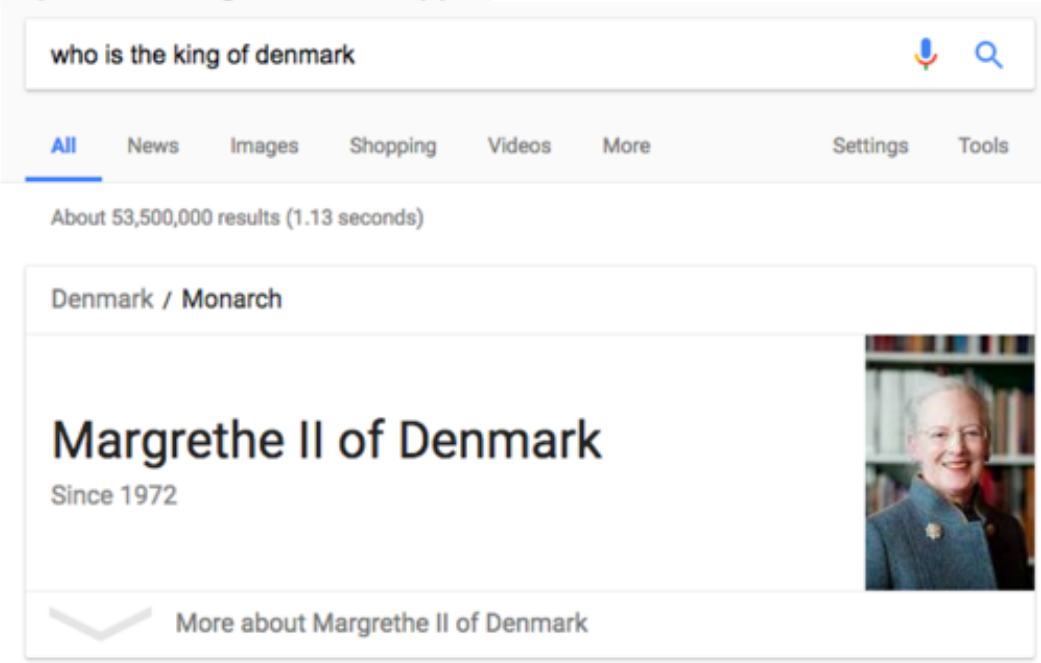
Views map from query intent to queries over the Google Knowledge Graph



But the KG Has to be Very Broad







Repeat x 100 Languages

qui est la femme de zidane

Tous Images Actualités Vidéos Shopping Plus Paramètres Outils

Environ 515 000 résultats (0,85 secondes)

Zinédine Zidane / **Épouse**

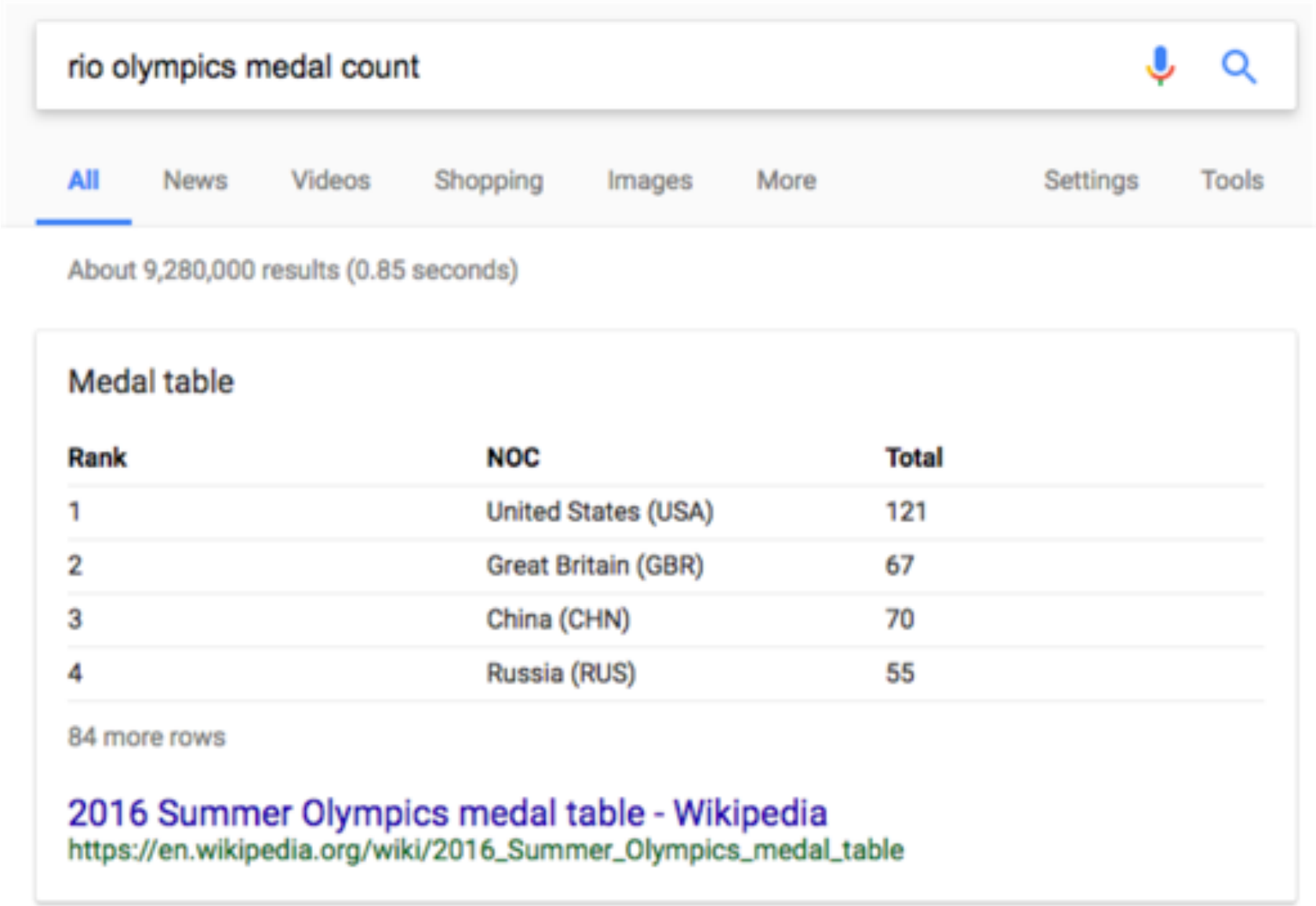
Véronique Zidane

m. 1994

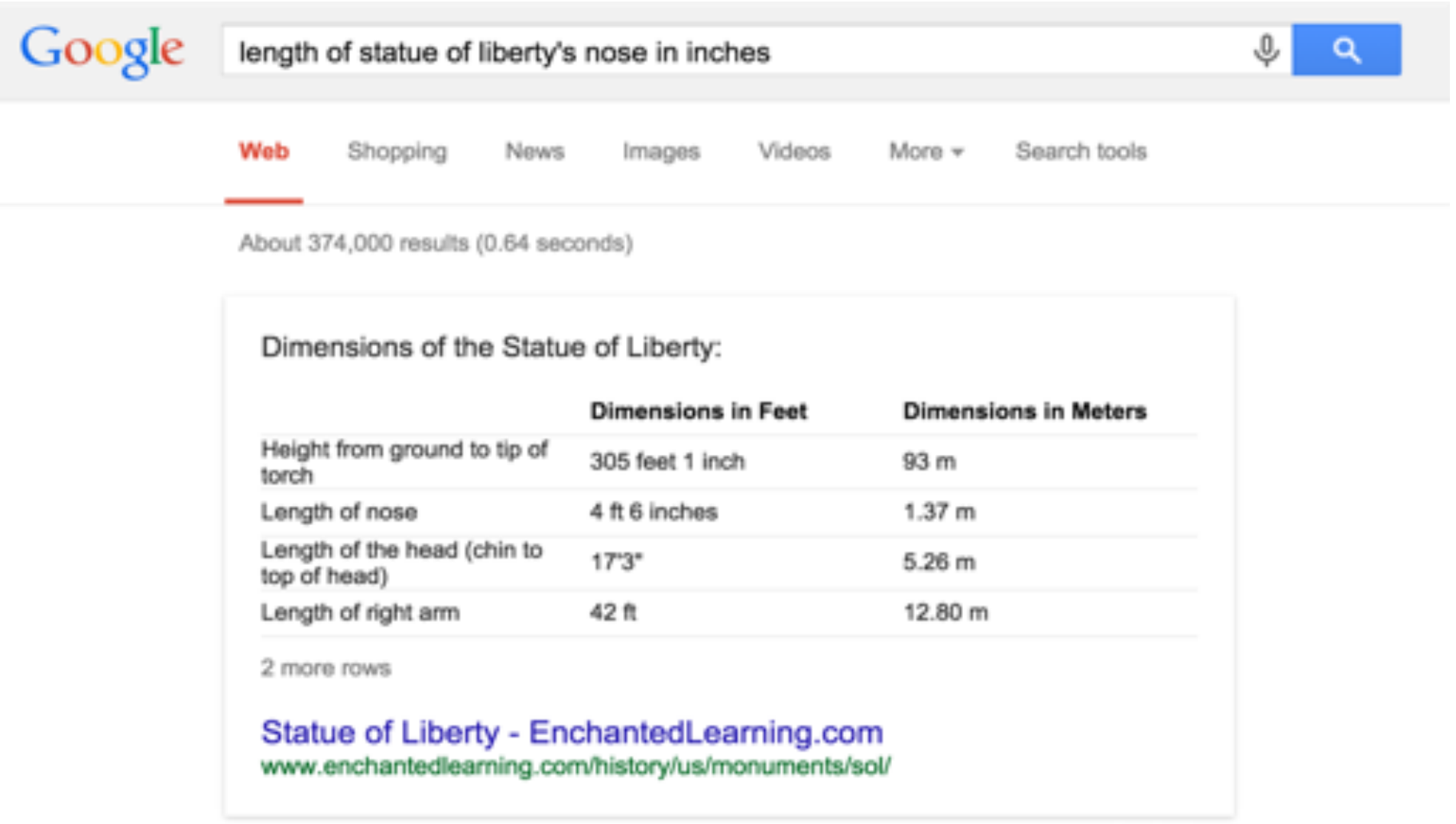
En savoir plus sur Véronique Zidane

Commentaires

Knowledge Graph Rough Edges

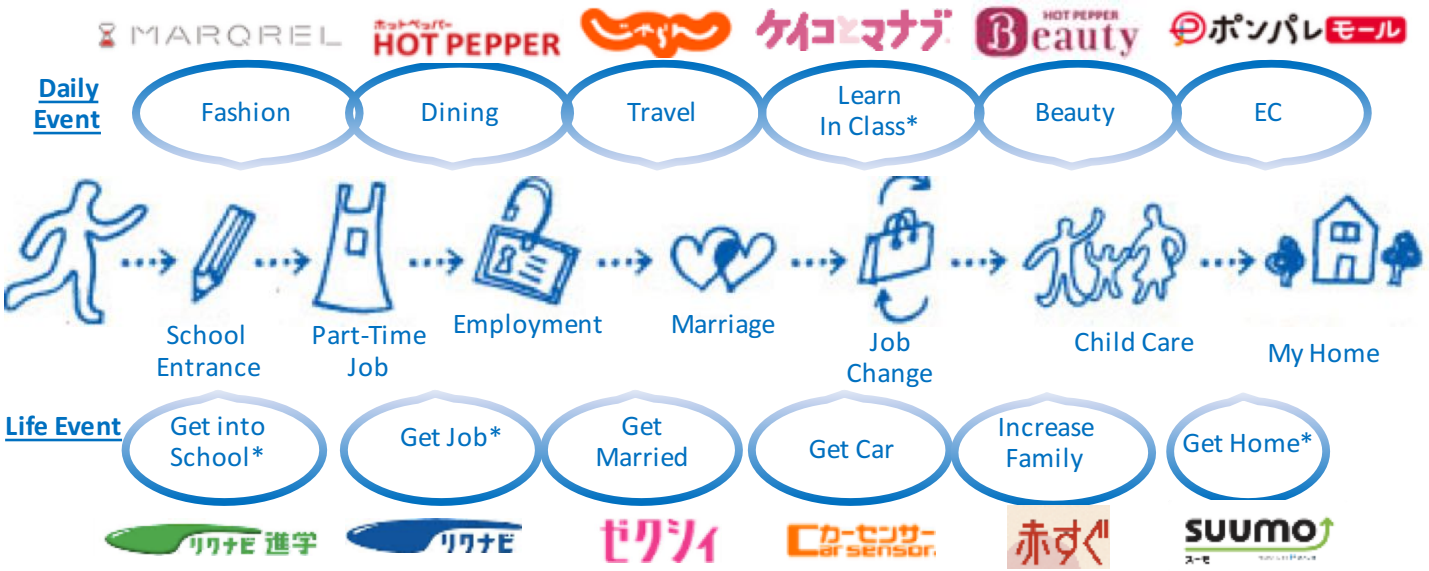


Long Nose Queries



Biperpedia (Whang+, 2014): KG covers a small fraction of structured data queries

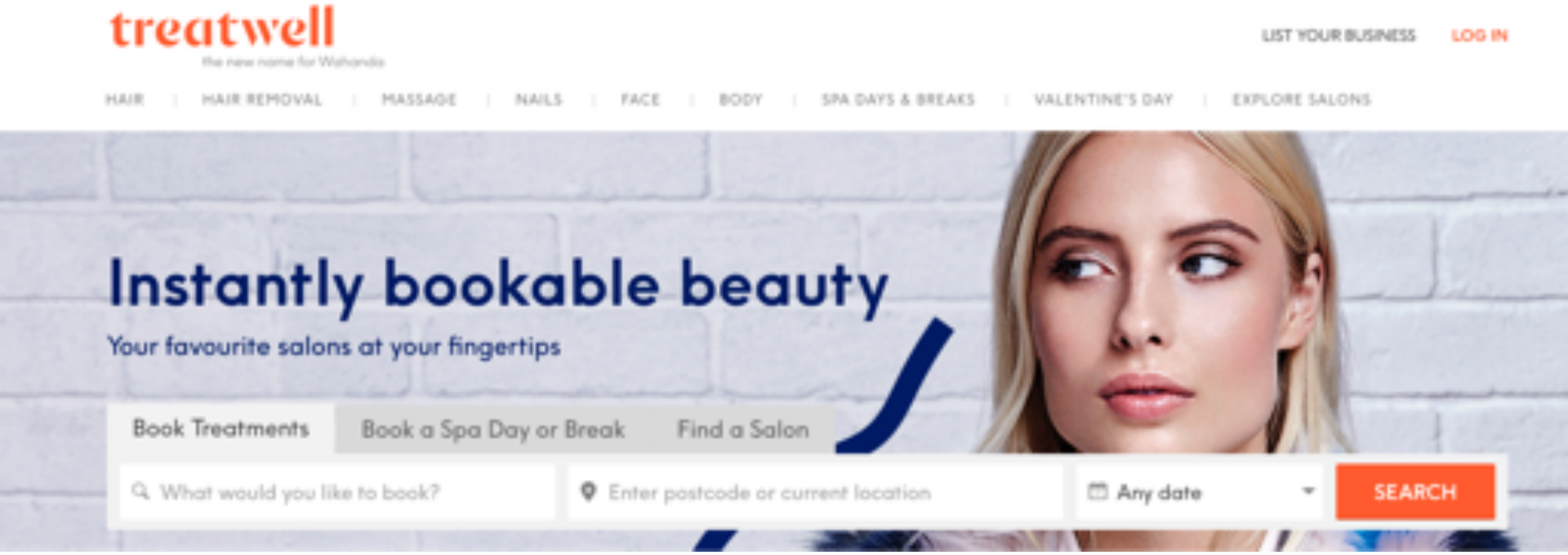
Recruit Holdings: A Lifestyle Company



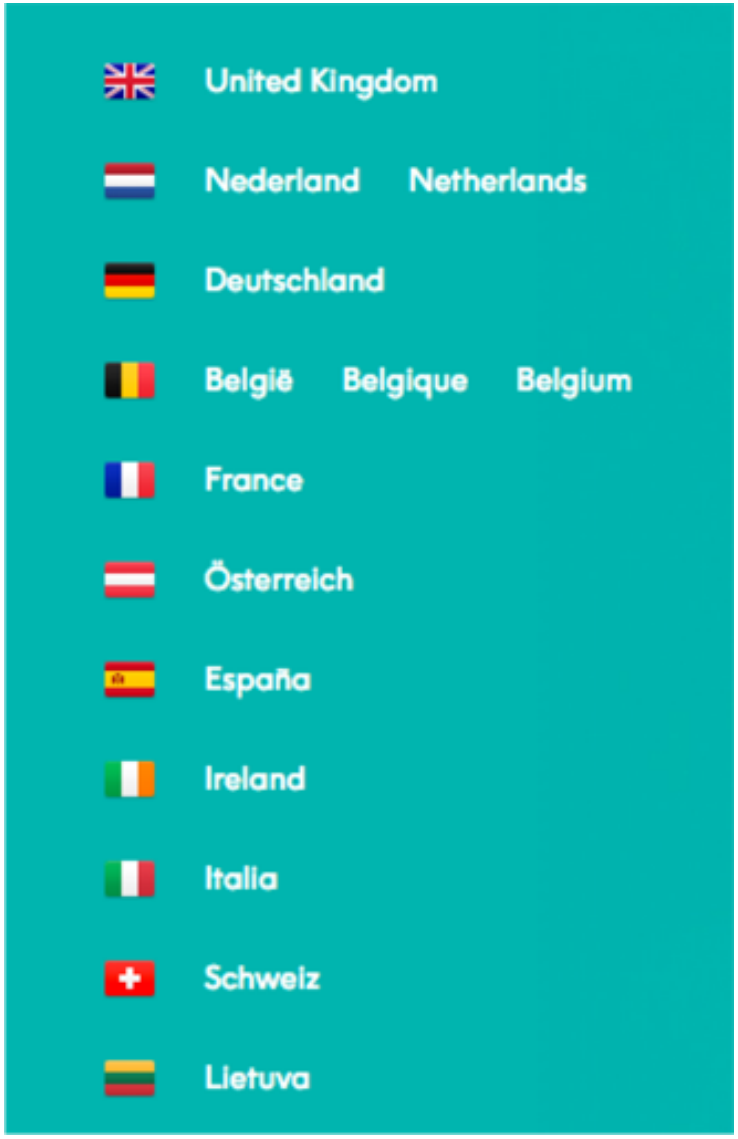
Over 200 online services

Notable subsidiaries: indeed.com, Treatwell

Beauty and Massage



Relevant to Most PODS Authors



treatwell

Book Yourself Fabulous

Relevant to Most PODS Authors

United Kingdom

Nederland Netherlands

Deutschland

België Belgique Belgium

France

Österreich

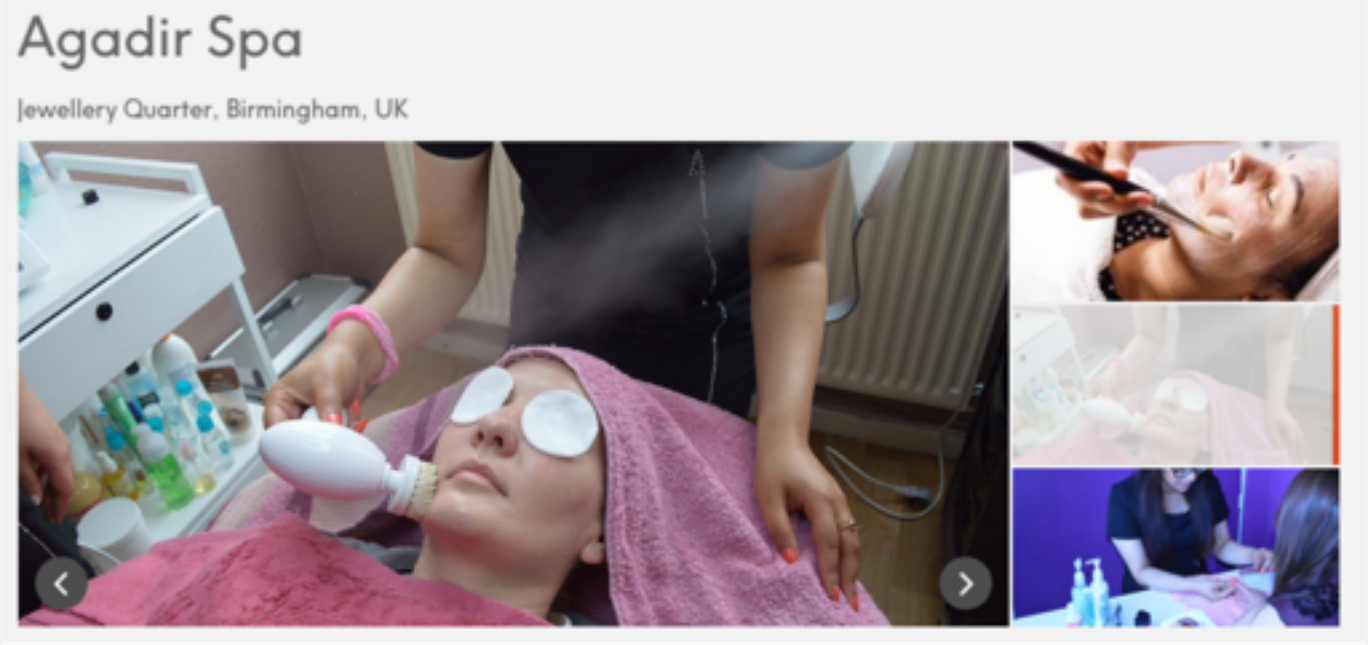
España

Ireland

Italia

Schweiz

Lietuva



Agadir Spa is a an authentic Moroccan beauty centre based in the Jewellery Quarter in Birmingham, where we offer full spa treatment services.

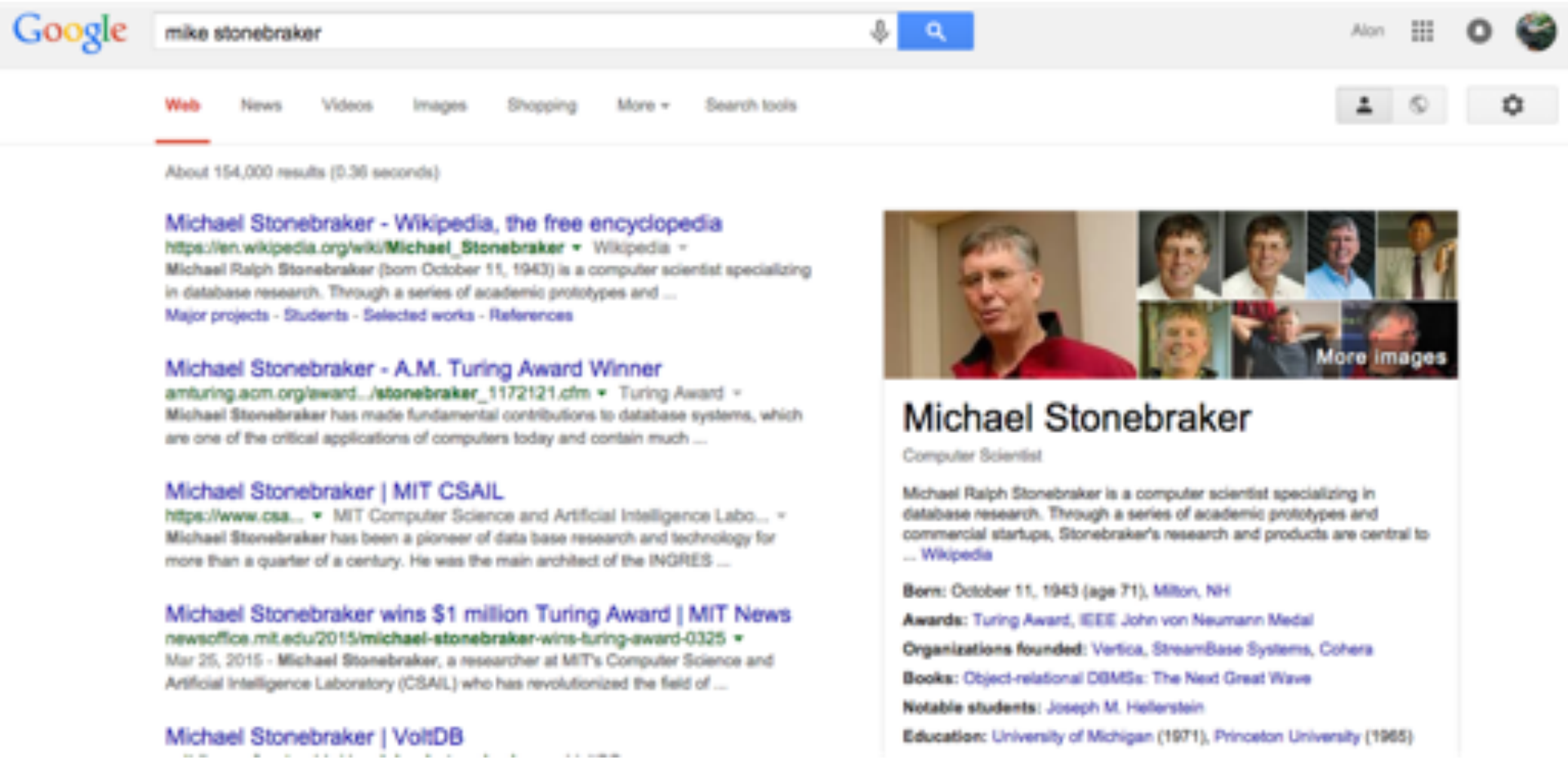
If we could understand that Morocco is not location but a type of beauty treatments we could answer queries such as “Moroccan massage” or “Moroccan spa”

Reformulate Data Integration

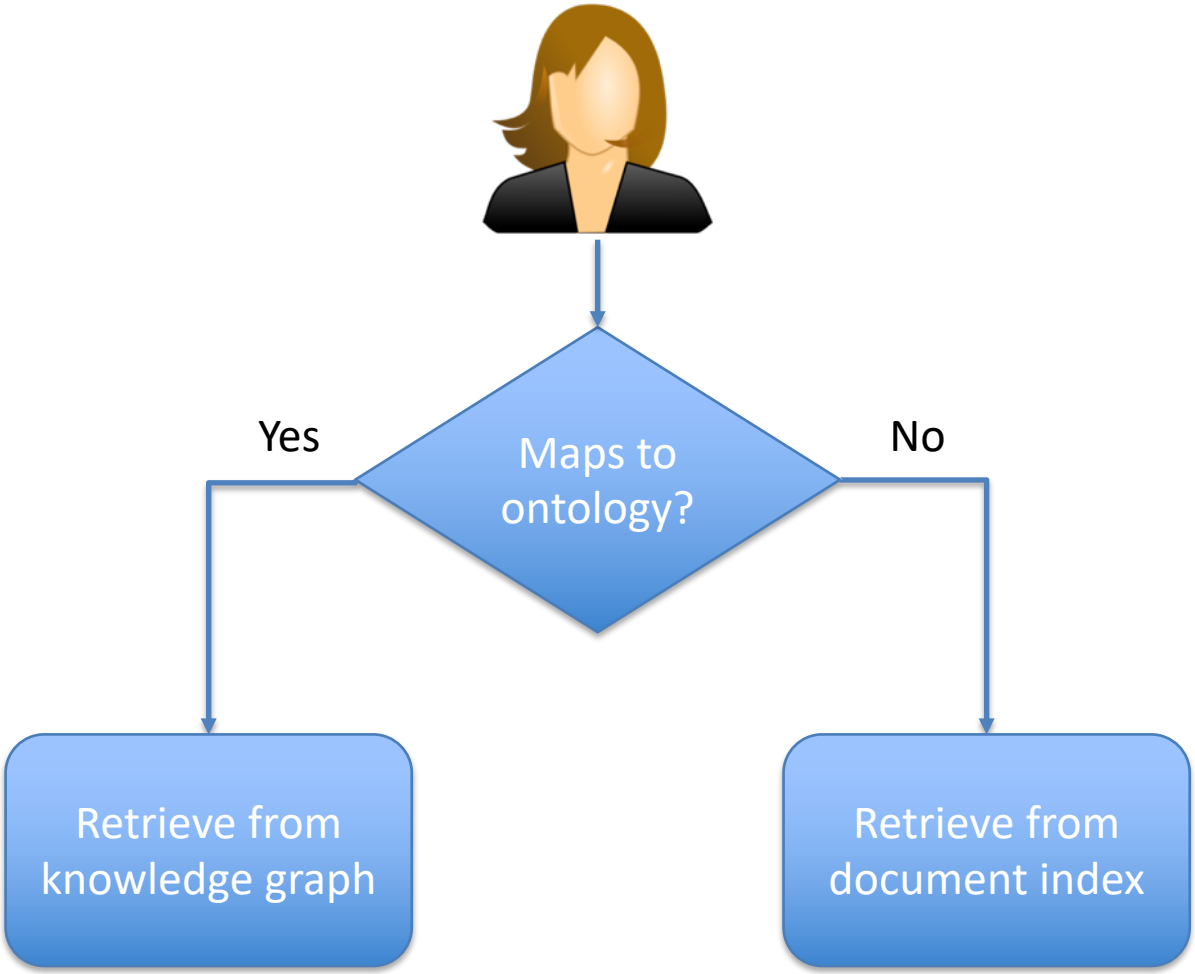
- It's a data space (or data lake), not a data integration system
  - Basic operation to support: search
  - See Goods @ Google (SIGMOD, 2016)

- Pay-as-you-go data integration:
  - For deeper integration, invest where you see the most value (typically measured by query traffic)

Curate Head Content

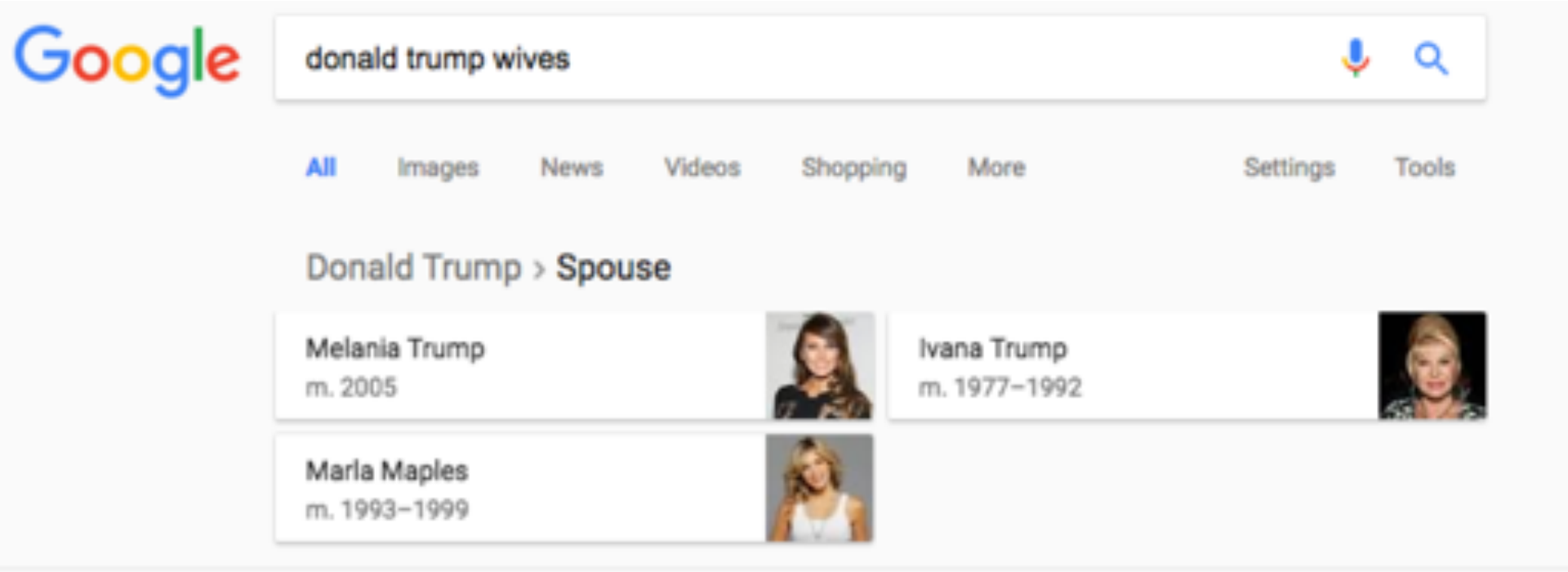


Answering a Query

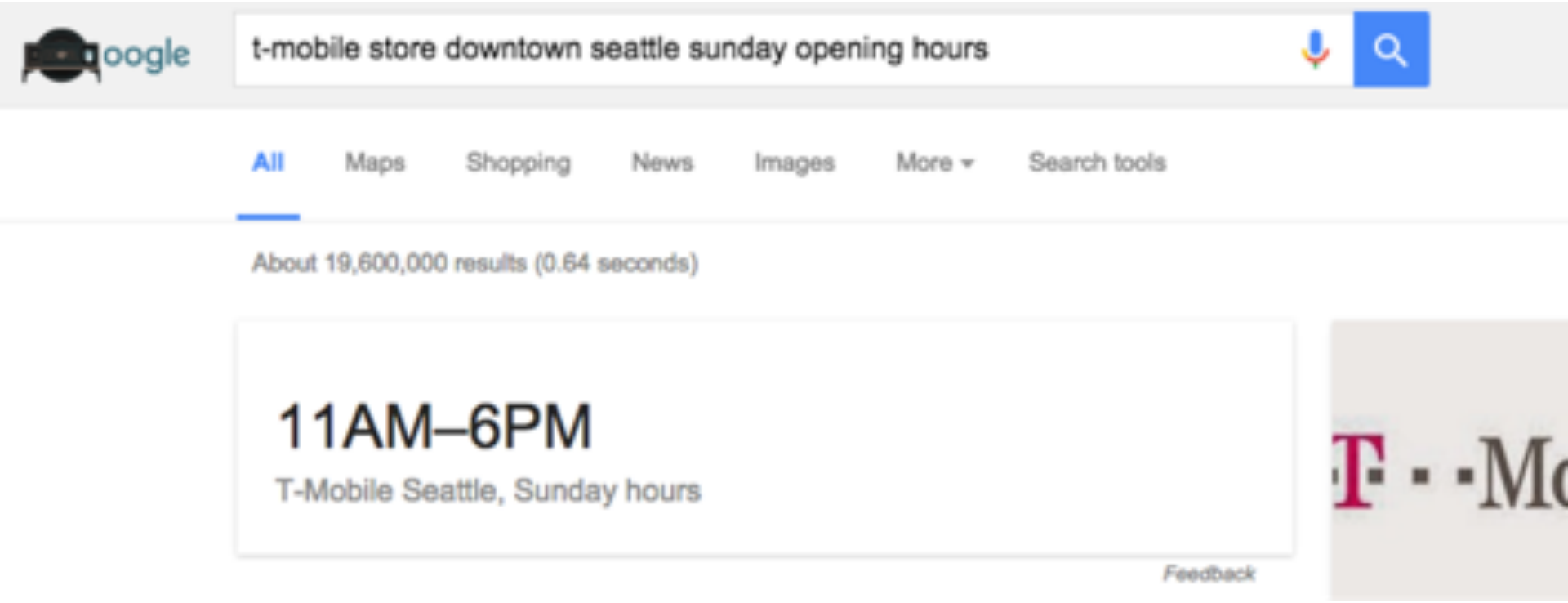


Attribute Queries

Multi-Valued Attributes Supported



Complicated Queries



Structured Data Can Be Funny

how tall is stephen colbert

Web News Images Videos Shopping More Search tools

About 391,000 results (0.33 seconds)

5' 10.5" (1.79m -ish)

Stephen Colbert, Height

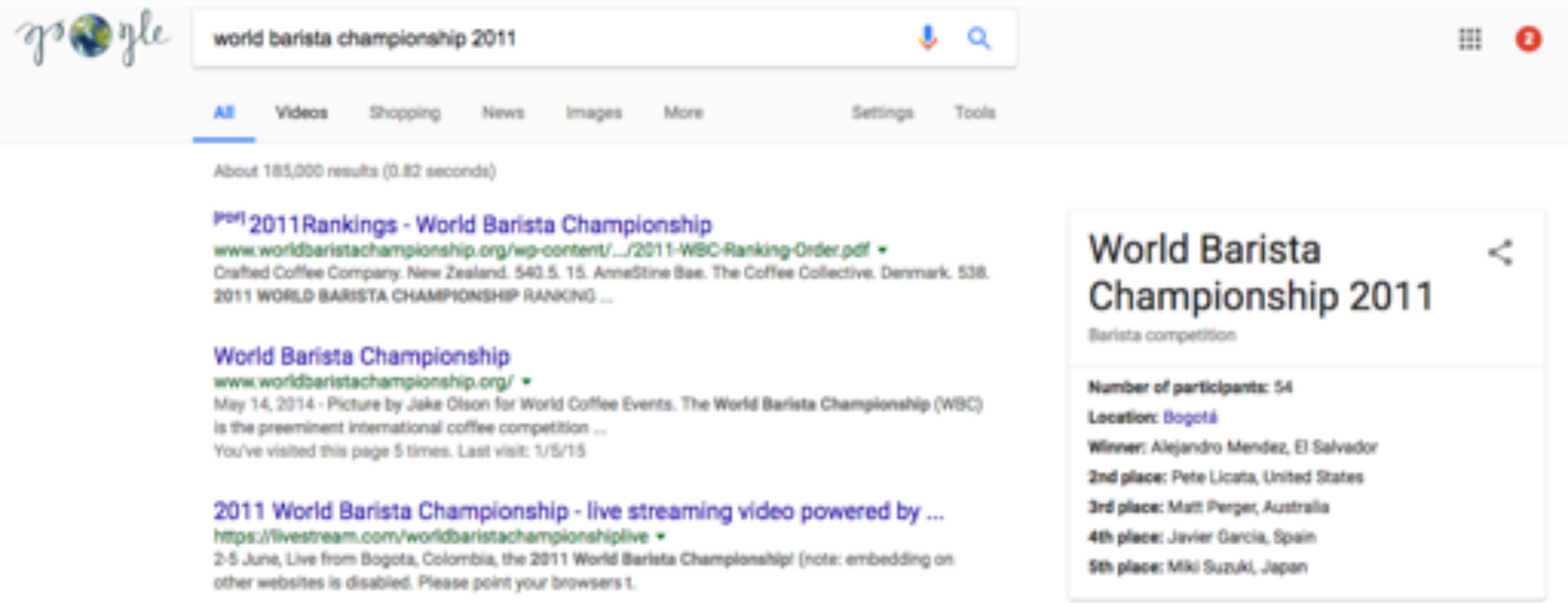
Larry Page
5' 11"

Jon Stewart
Shorter

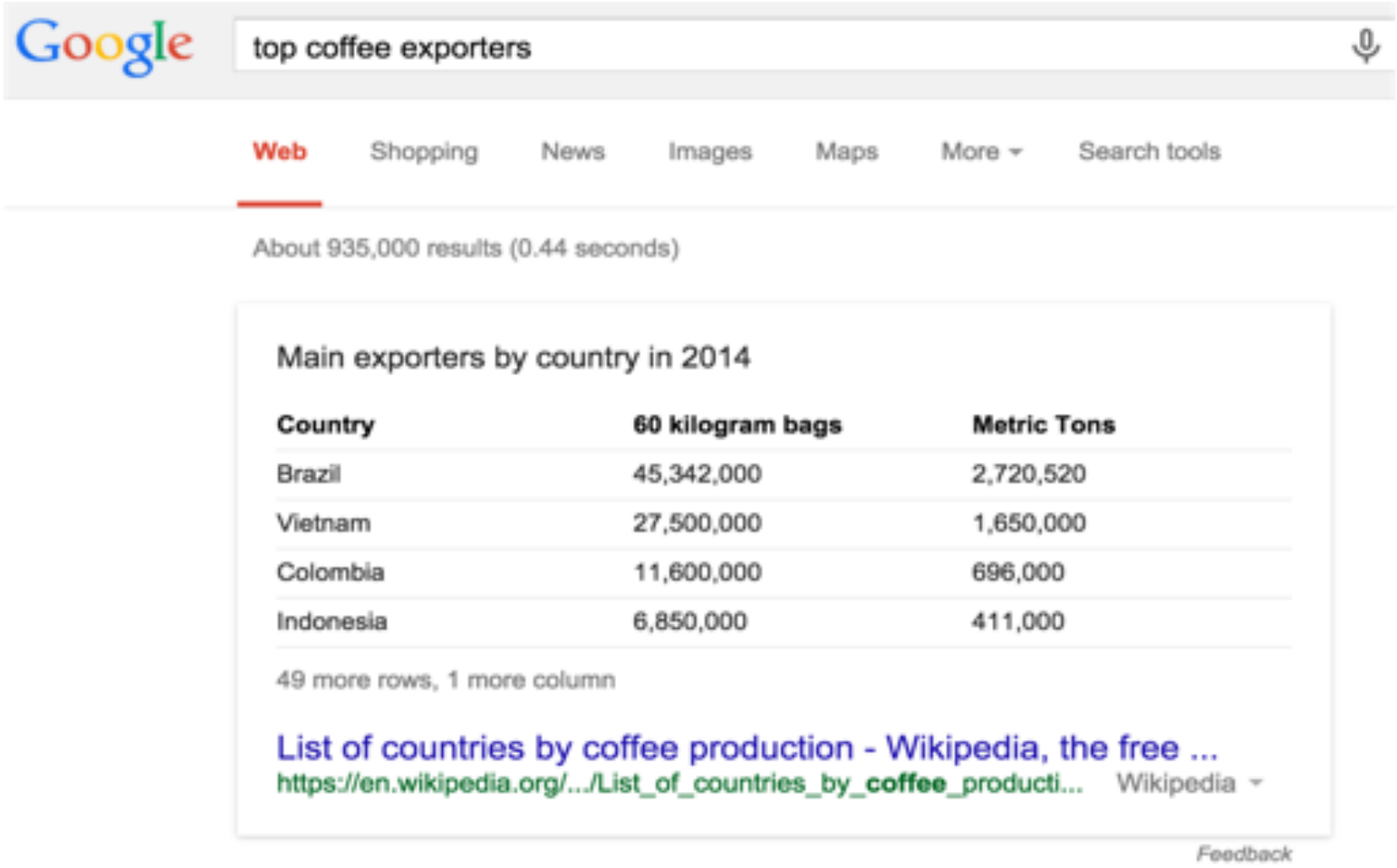
Conan O'Brien
6' 4"

Feedback

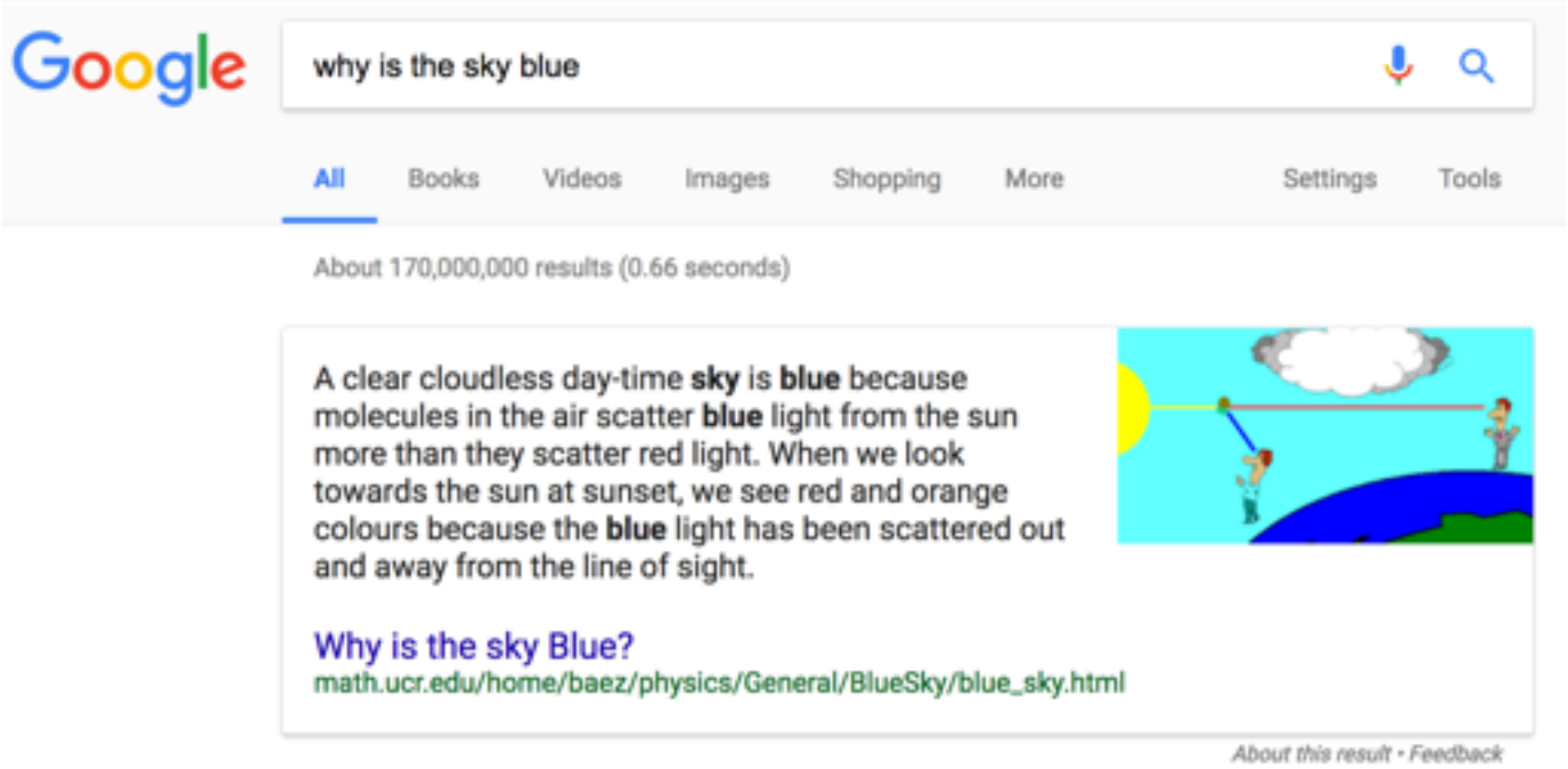
Enable Long-Tail Curators



Fall Back on Unintegrated Content



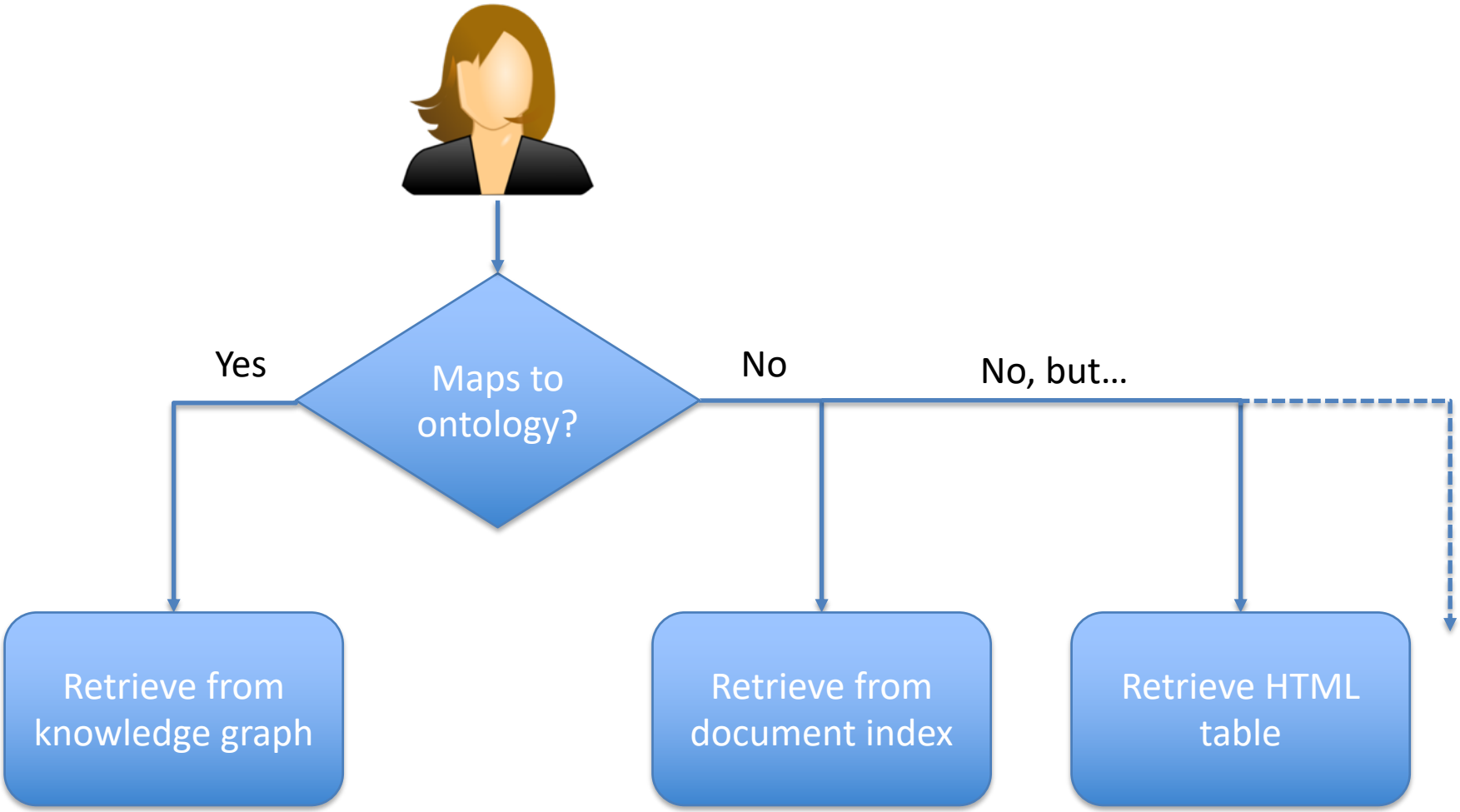
Why Is the Sky Blue?



Outline

- ✓ Throwback to the 90's:
  - ✓ Answering queries using views: what and why
- ✓ Broader view: data integration
  - ✓ Still, mostly 90's
- ✓ Data integration today
  - – The Web and data lakes
- ➢ Two data integration challenges
  - – Might sound familiar

Many Algorithms to Answer Queries



who is the speaker of the house?

About 220,000,000 results (0.59 seconds)

Stand with Speaker Ryan - speakerryan.com
Ad www.speakerryan.com/
ibute $10 to help usher in a positive, conservative agenda!

John Boehner

Speaker of the U.S. House of Representatives

no.	name	party or faction
59	J. Dennis Hastert	Republican
60	Nancy Pelosi	Democratic
61	John Boehner	Republican
62	Paul Ryan	Republican

58 more rows, 3 more columns

peaker of the U.S. House of Representatives | United States ...
https://www.britannica.com/.../**Speaker**-of-the-US-**House**-of-Re... Encyclopædia Britannica

About this result • Feedback

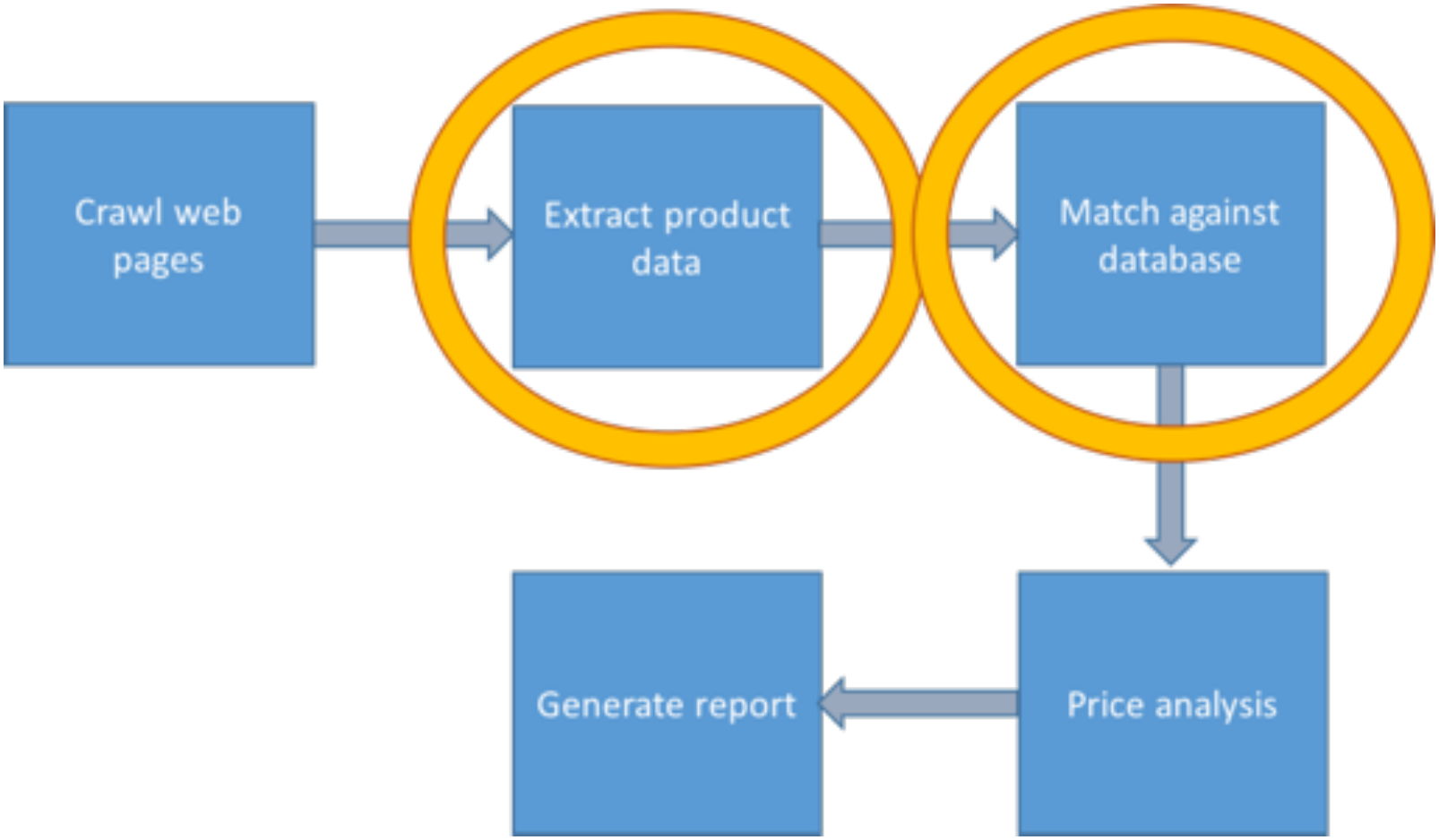
Challenge #1: Combining Structured and Unstructured Data

- We have methods for extracting facts from unstructured data
  - But then we don't know how to reconcile the extractions with structured data
- Two observations about the solution:
  - Combination should be specified declaratively
    - Right now it's hardwired somewhere in code
  - Need to keep the original data (context) around. We might need it later.

Challenge #2: Open-Source Tools

- The field has been around for a while and made impressive progress
- But…
  - No (free) tools available
  - Hard to prototype a data integration application
  - Each research effort needs to start from scratch
  - Missing opportunities for impact on the world (e.g., in data science)

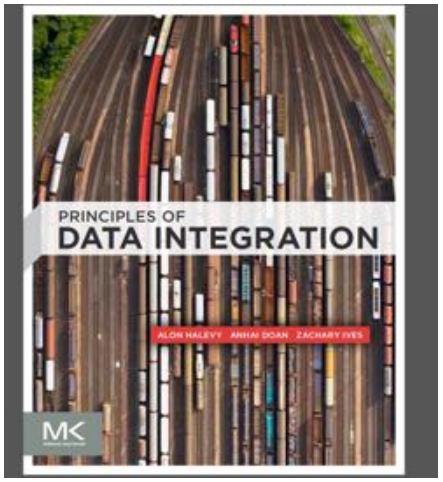
We Need Integration Operators



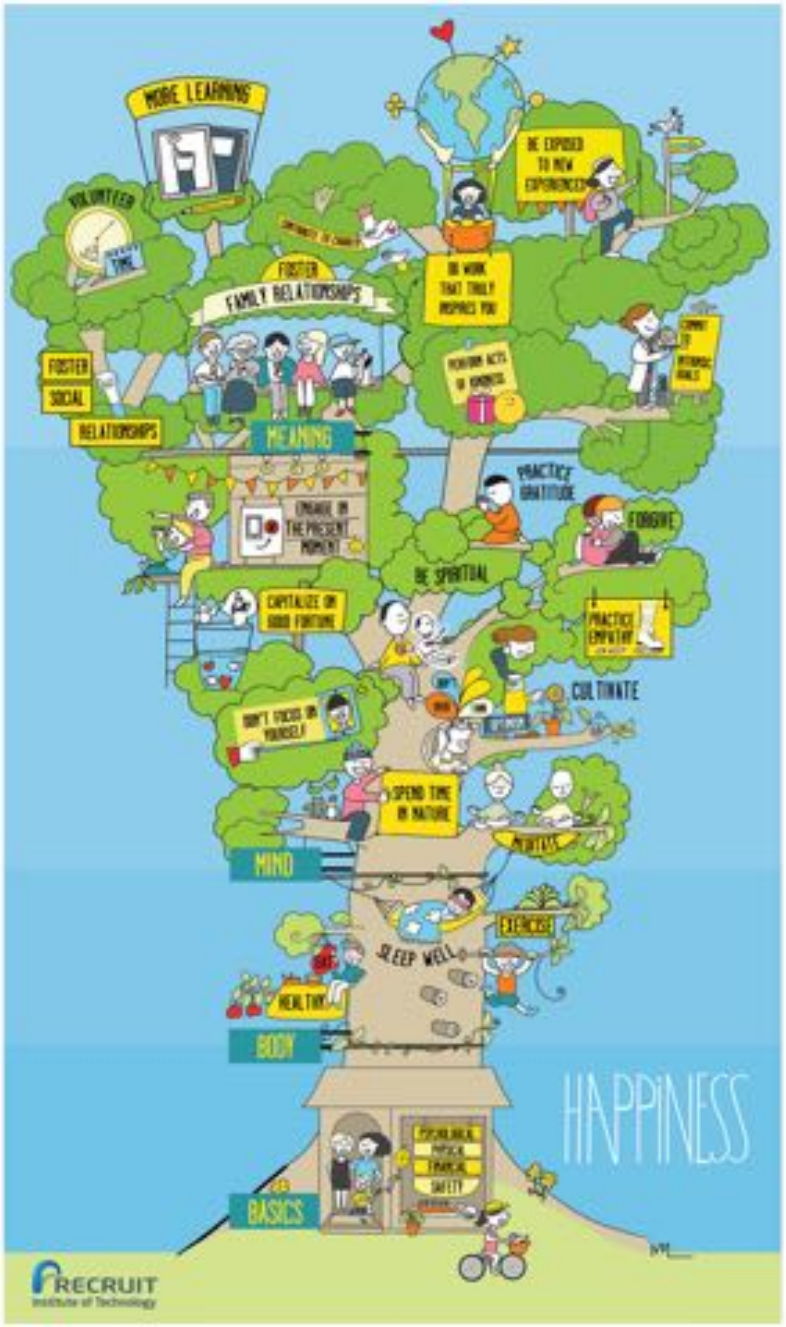
BigGorilla.org

- Goal: open source data integration and preparation eco-system.
- In Python (for easy prototyping)
- Currently: string matching, data matching, schema matching, scraping.
  - See tutorials, code and demos on the site
- Collaboration with U. Wisconsin; looking for additional contributors.

Conclusions



- Data integration is everywhere:
  - Even in my kitchen!
  - Sources described by views
  - AI (NLP & Speech) are enabling new data integration scenarios
  - And enterprises have not solved the problem yet
- But we face two age-old problems:
  - The world is messy and our models are not adequate
  - Data integration is still too hard and labor intensive


MORE LEARNING
VOLUNTEER
FOSTER FAMILY RELATIONSHIPS
BE EXPOSED TO NEW EXPERIENCES
DO WORK THAT TRULY INSPIRES YOU
FOSTER SOCIAL RELATIONSHIPS
MEANING
PRACTICE GRATITUDE
FORGIVE
ENGAGE IN THE PRESENT MOMENT
BE SPIRITUAL
CAPITALIZE ON GOOD FORTUNE
PRACTICE EMPATHY
CULTIVATE
SPEND TIME IN NATURE
MIND
EXERCISE
SLEEP WELL
HEALTHY
BODY
HAPPINESS
PSYCHOLOGICAL
PHYSICAL
FINANCIAL
SAFETY
BASICS
RECRUIT
Institute of Technology

Q & A