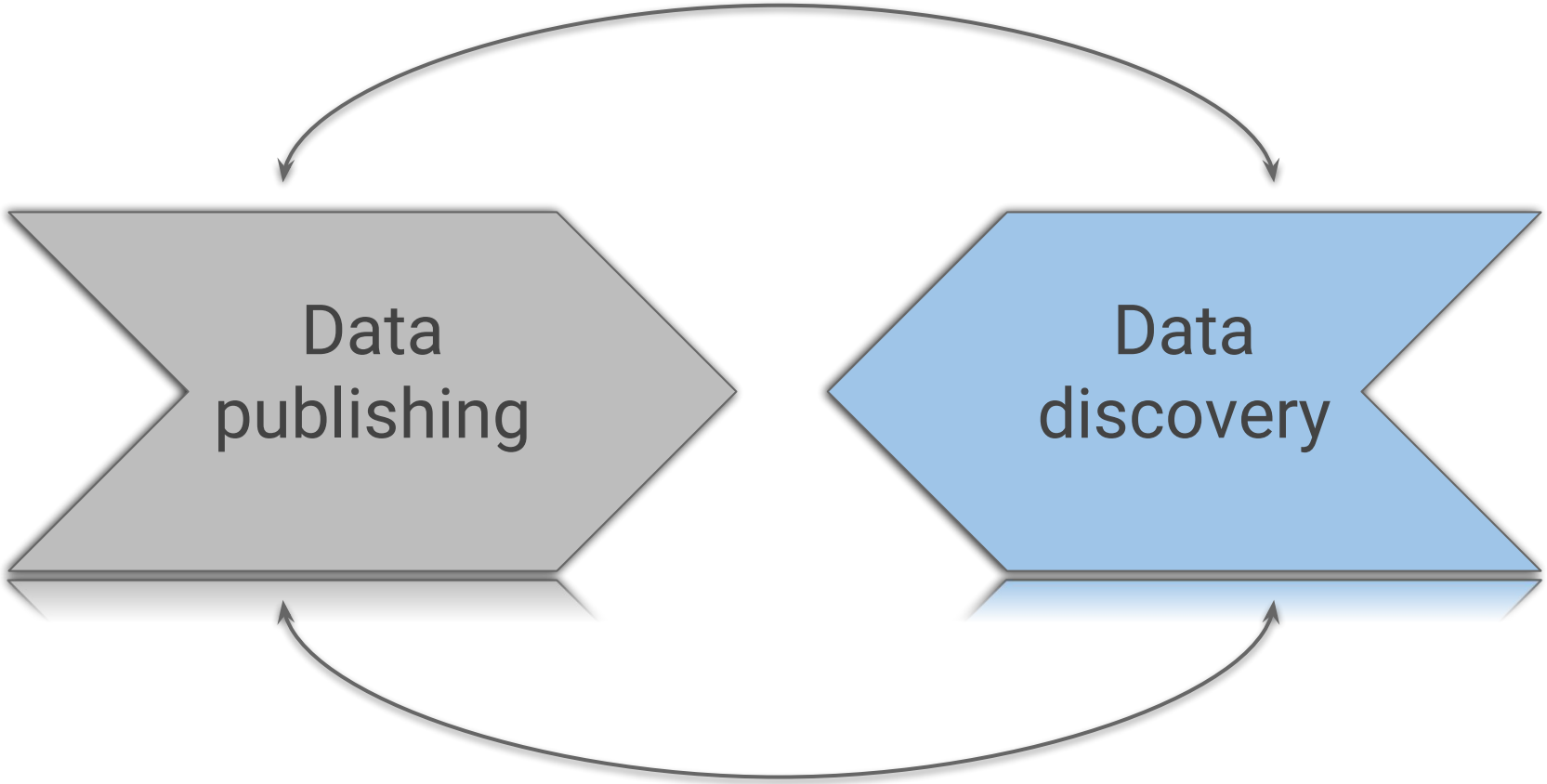
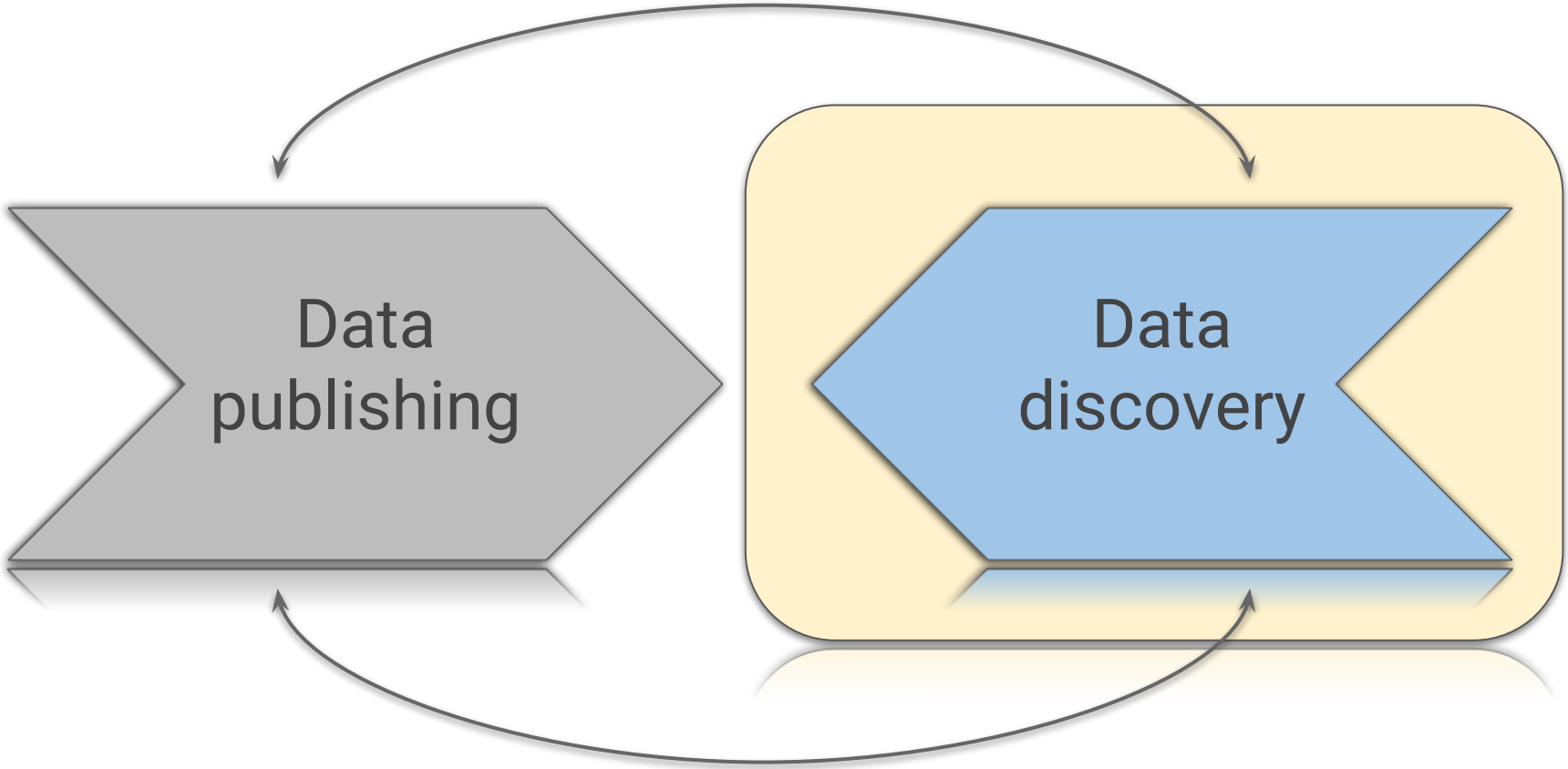
Google

Facilitate Discovery of Scientific Data

Natasha Noy
Google, Inc.
noy@google.com


Data
publishing
Data
discovery

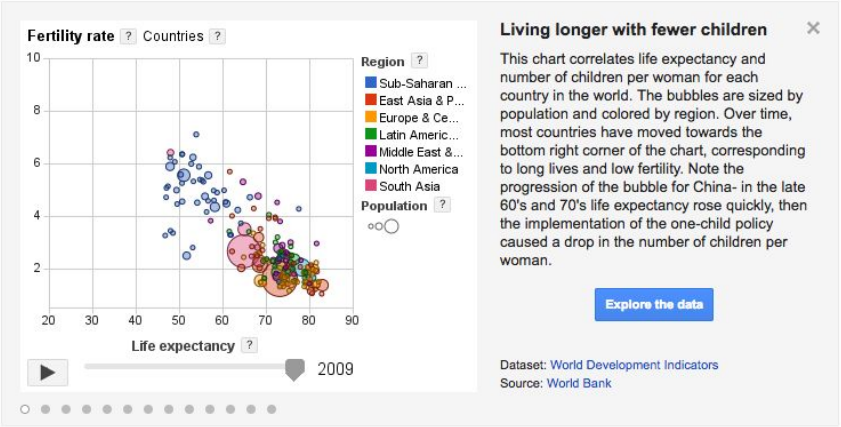

Data
publishing
Data
discovery

Google's mission is to organize the world's information and make it universally accessible and useful.

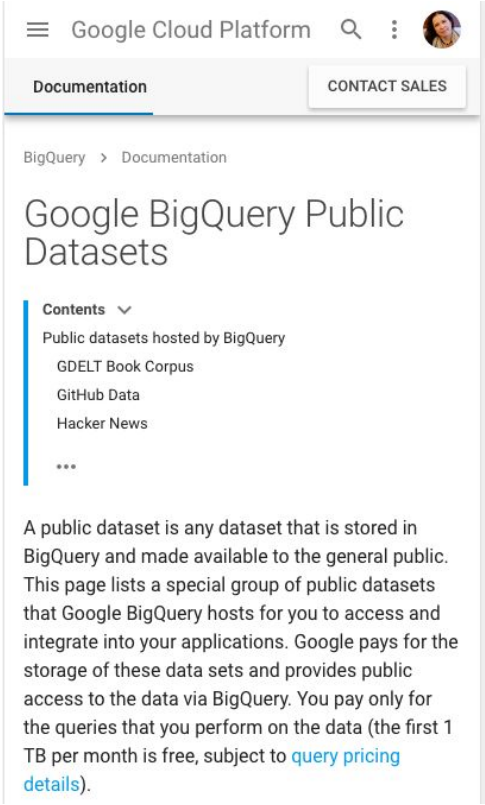
Scientific data is an important part of the world's information

Google’s forays into public data

Public data explorer



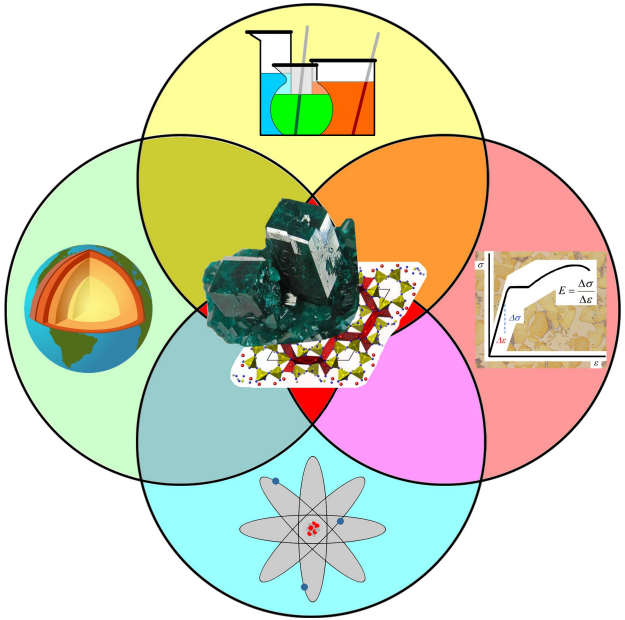
BigQuery public datasets



Public data: a lot more than we can curate

- The good news: There are lots of public data online
  - Funding agencies and journals compel (or mandate) scientists to publish their data
  - Many governments have mandates to publish data
- The bad news: This data is spread through thousands of repositories and largely not searchable.
  - Metadata is not searchable either
- Examples:
  - The Nature publishing group recommends 58 different repositories for their authors
    - Some are generic (Dryad and figshare)
    - Many are domain specific, with several repositories per domain
  - Re3data.org, a registry of research repositories, lists more than 1,300 repositories

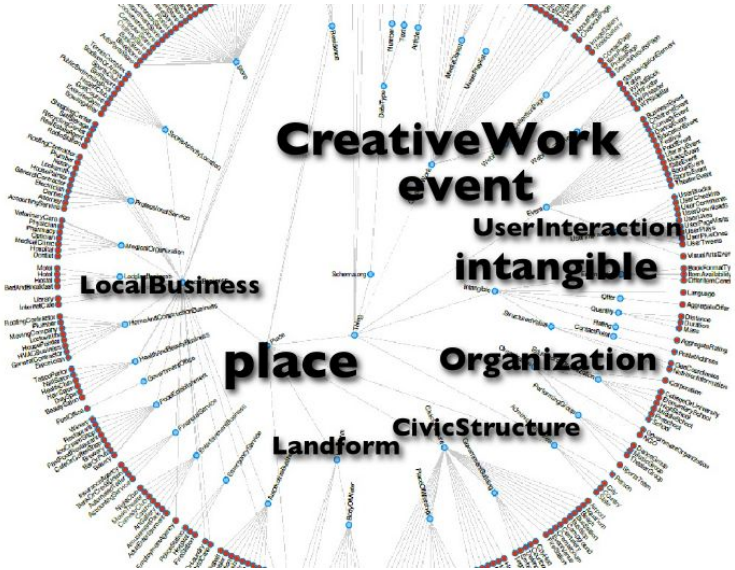
There are a lot more datasets than anybody can, or should, curate.



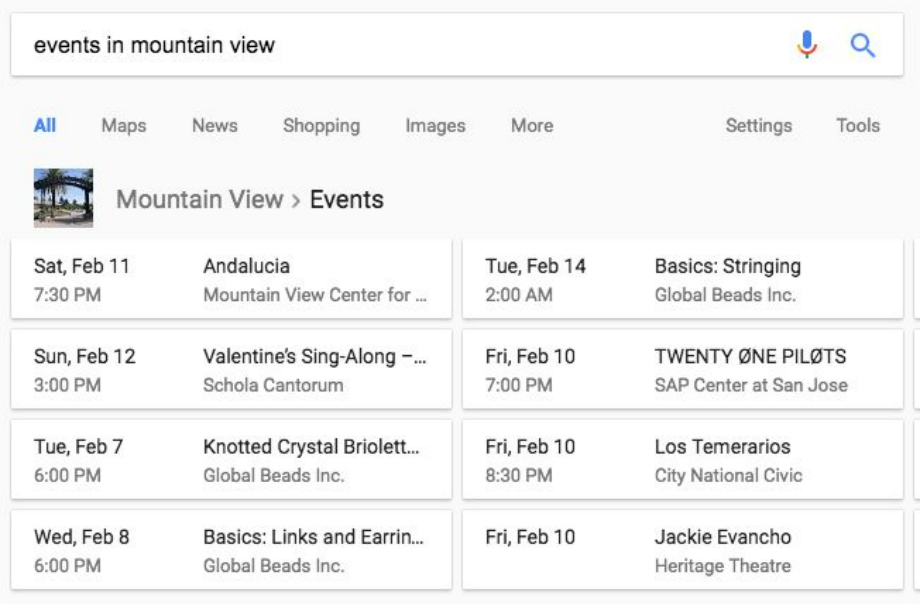
We can enable users to **discover** them.

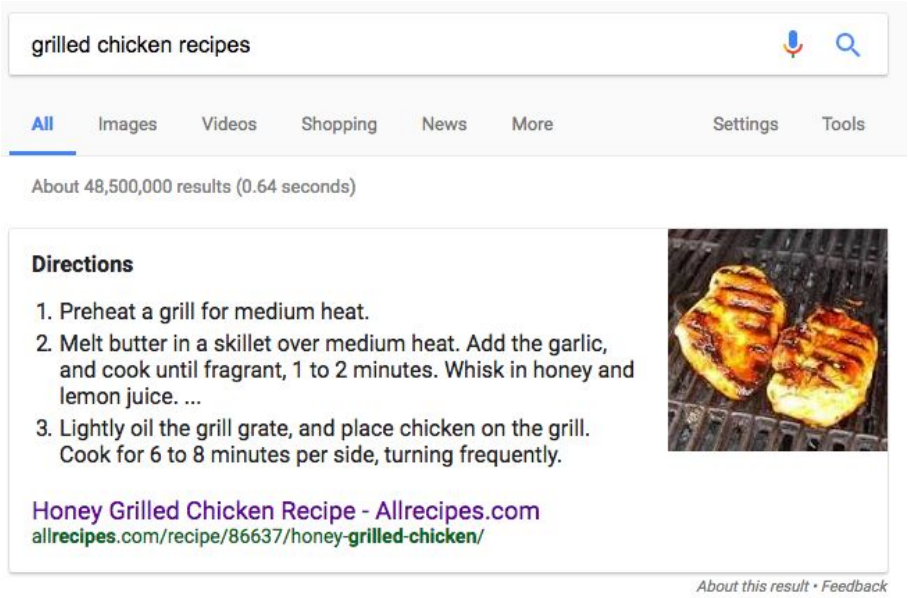
Part of the solution: Structured data (schema.org)

- **Structured data markup on the Web**
  - **Founded by search engines** in 2011
    - Google, Bing, Yahoo!, Yandex
    - Widely used in the Web
- Adoption driven by its use in real search products
  - Initially "**Rich Snippets**"
  - **Knowledge Graph(s)**, Carousels, **Smart assistants** and email-based personalization



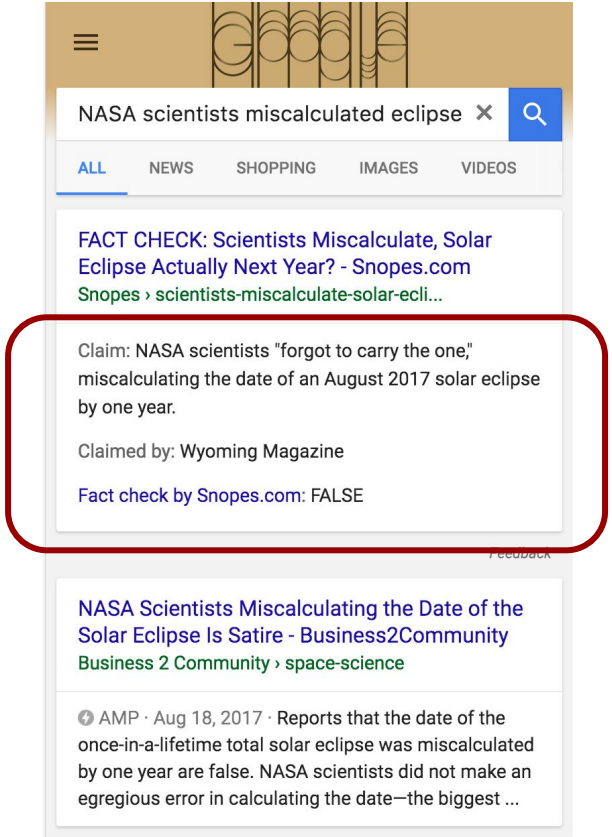
You see the use of structured data for recipes, events, movies...



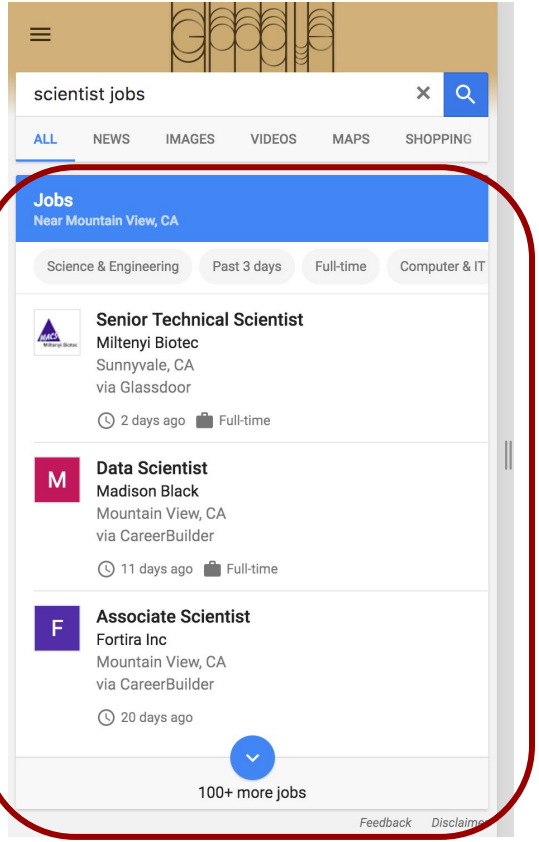


Uses of schema.org in Google products: Recent examples

Fact checking



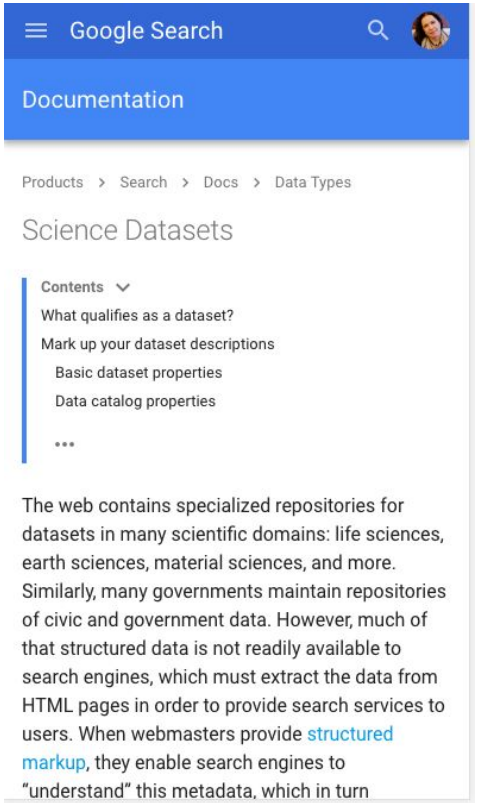
Jobs



Up next: Datasets

- Data providers describe the data
  - in an **open format**
  - that is **web friendly**
  - that search engines **understand**
  - that is **lightweight**
  - that is **easy to implement**, with lots of tools available
- Communities extend schema.org with domain-specific vocabularies
  - Example: bioschemas.org

[Guidelines for describing datasets in schema.org]



Some metadata is better than none

- Metadata about datasets leads to data discovery
  - Descriptions, keywords
  - Digital object identifiers (DOIs for datasets_
  - Spatial and temporal coverage
  - Provenance of the data
  - Download and license information
  - Links to scientific publications
  - Description of the schema

Either stated explicitly by providers or inferred

Next steps

Data providers

publish
structured metadata using schema.org and other community standards

Data consumers

cite
data properly, much as we cite scientific publications

Developers

contribute
to expanding schema.org metadata for datasets

Create a healthy data ecosystem

Google's mission is to organize the world's information and make it universally accessible and useful.

Scientific data is an important part of the world's information

Contact us: dataset-support@google.com