

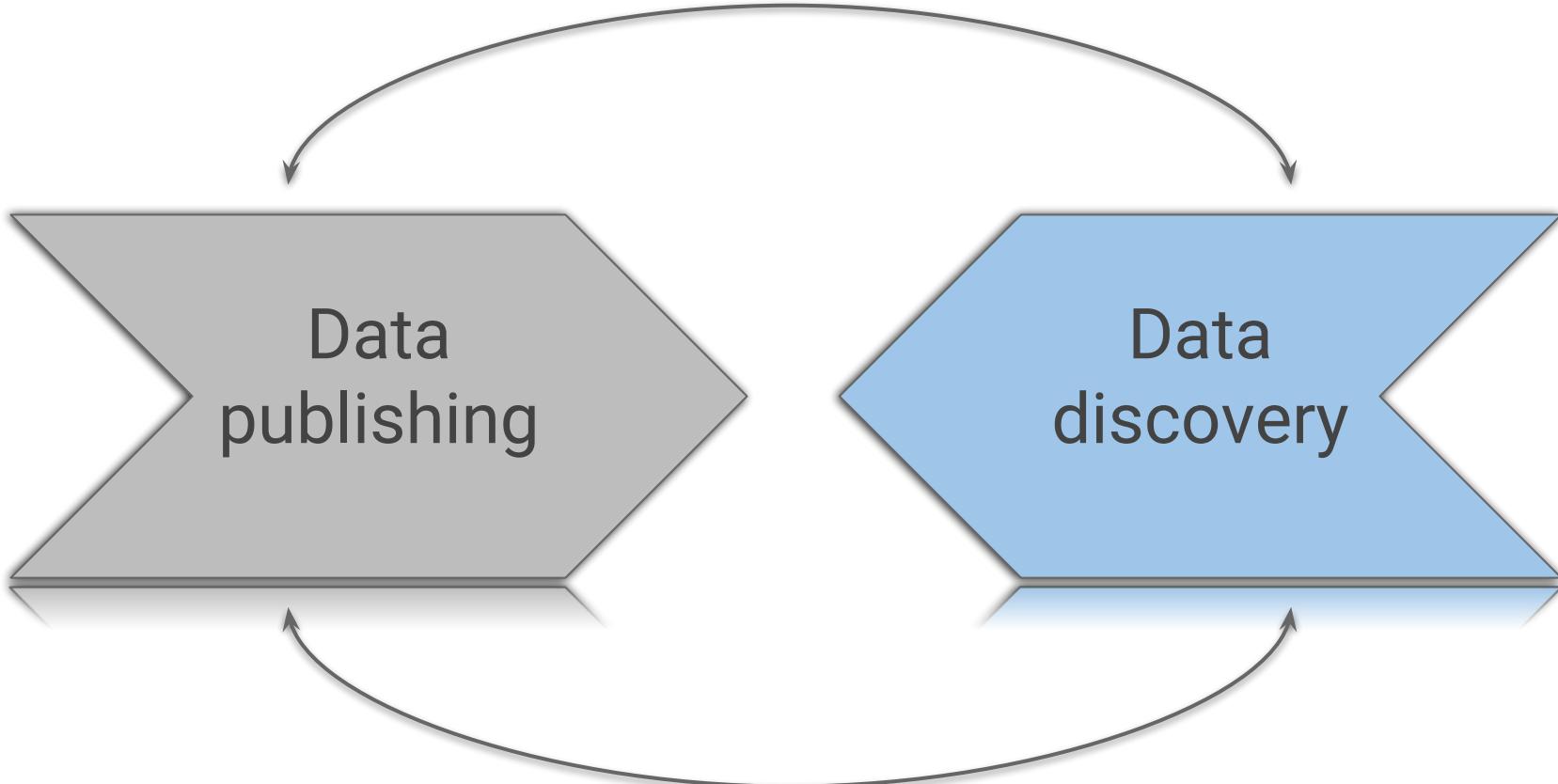


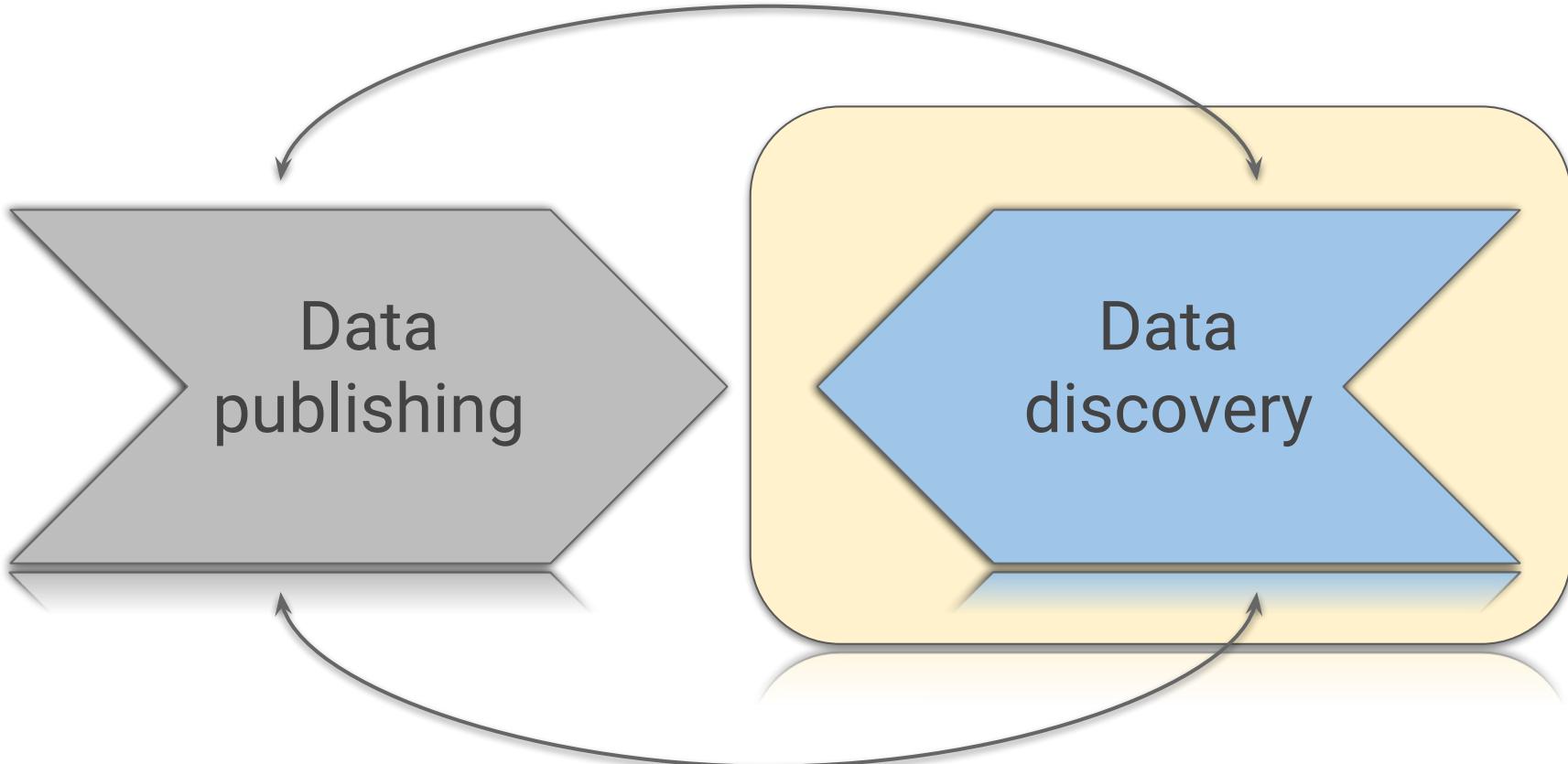
# Facilitate Discovery of Scientific Data

**Natasha Noy**

Google, Inc.

[noy@google.com](mailto:noy@google.com)





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



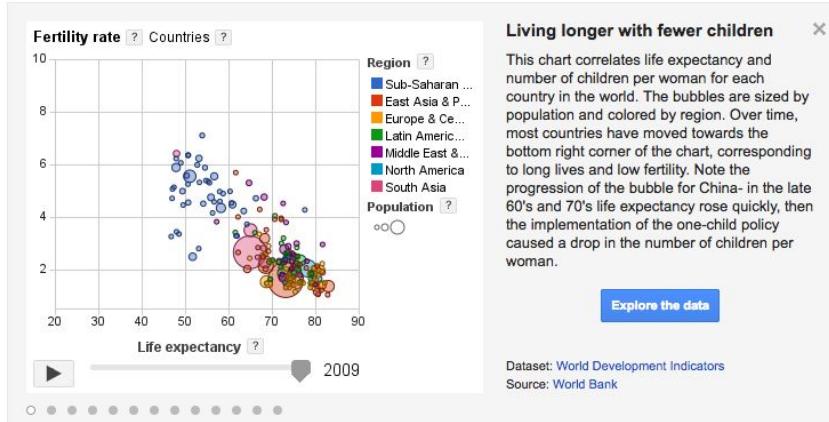
*Scientific data is an important part of the world's information*

# Google's forays into public data

## Public data explorer

- Datasets**
- Metrics
- Any data provider (104)**
  - Eurostat (9)
  - Statistics Iceland (6)
  - U.S. Census Bureau (5)
  - Central Statistics Office, Ireland (4)
  - Data.gov.uk (4)

### My Datasets



## BigQuery public datasets

≡ Google Cloud Platform 🔍 ⋮

Documentation

CONTACT SALES

BigQuery > Documentation

## Google BigQuery Public Datasets

### Contents

- Public datasets hosted by BigQuery
- GDELT Book Corpus
- GitHub Data
- Hacker News

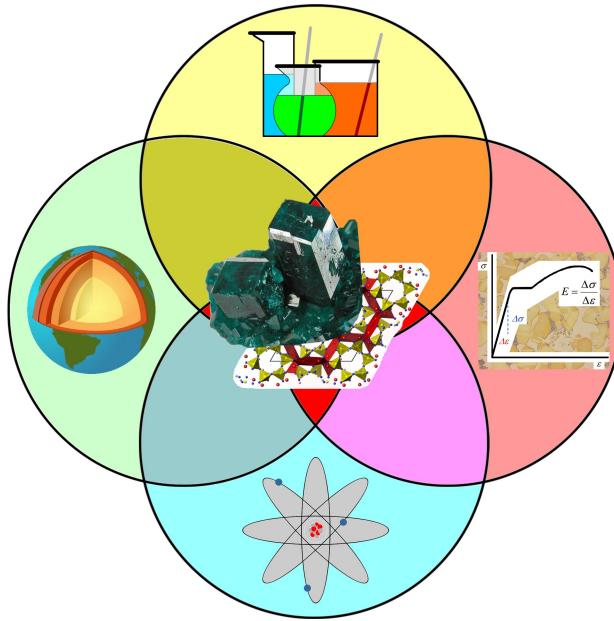
...

A public dataset is any dataset that is stored in BigQuery and made available to the general public. This page lists a special group of public datasets that Google BigQuery hosts for you to access and integrate into your applications. Google pays for the storage of these data sets and provides public access to the data via BigQuery. You pay only for the queries that you perform on the data (the first 1 TB per month is free, subject to [query pricing details](#)).

# 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

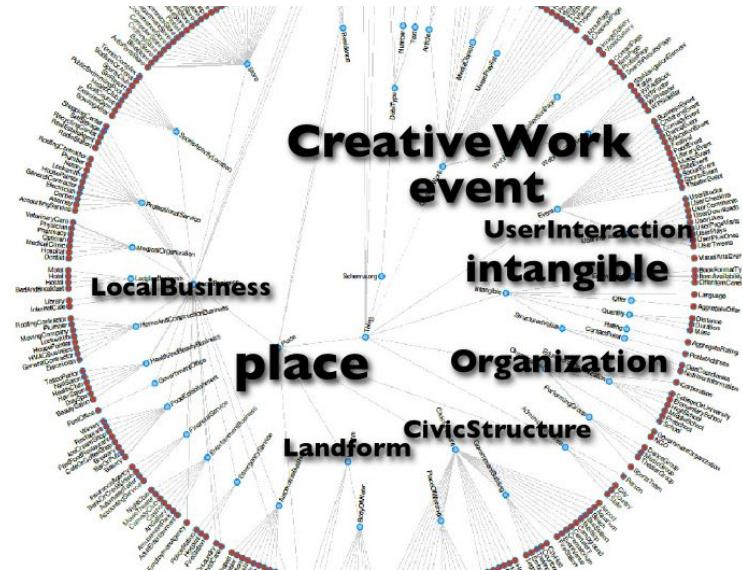
*There are a lot more datasets than anybody can, or should, curate.*



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...

events in mountain view

All Maps News Shopping Images More Settings Tools

Mountain View > Events

Sat, Feb 11 7:30 PM	Andalucia Mountain View Center for ...	Tue, Feb 14 2:00 AM	Basics: Stringing Global Beads Inc.
Sun, Feb 12 3:00 PM	Valentine's Sing-Along —... Schola Cantorum	Fri, Feb 10 7:00 PM	TWENTY ØNE PILØTS SAP Center at San Jose
Tue, Feb 7 6:00 PM	Knotted Crystal Briolett... Global Beads Inc.	Fri, Feb 10 8:30 PM	Los Temerarios City National Civic
Wed, Feb 8 6:00 PM	Basics: Links and Earrin... Global Beads Inc.	Fri, Feb 10	Jackie Evancho Heritage Theatre

grilled chicken recipes

All Images Videos Shopping News More Settings Tools

About 48,500,000 results (0.64 seconds)

**Directions**

1. Preheat a grill for medium heat.
2. Melt butter in a skillet over medium heat. Add the garlic, and cook until fragrant, 1 to 2 minutes. Whisk in honey and lemon juice. ...
3. Lightly oil the grill grate, and place chicken on the grill. Cook for 6 to 8 minutes per side, turning frequently.

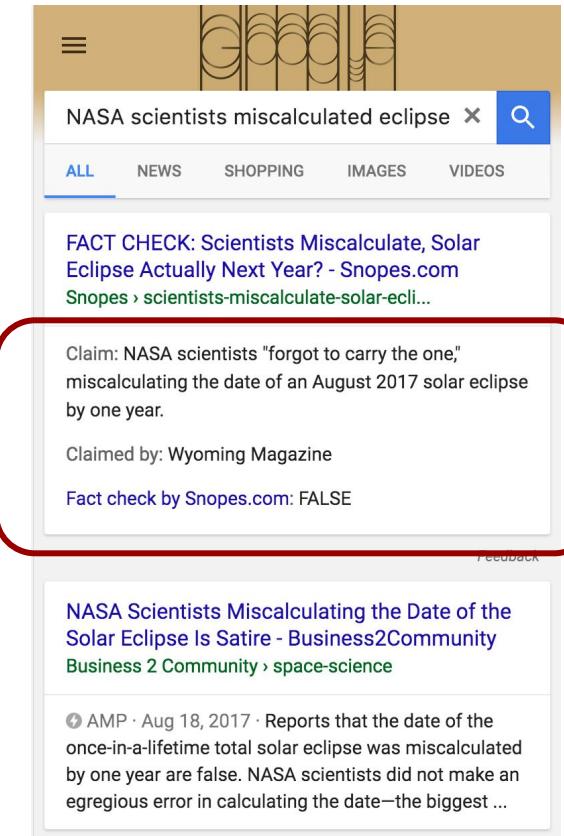
**Honey Grilled Chicken Recipe - Allrecipes.com**  
[allrecipes.com/recipe/86637/honey-grilled-chicken/](http://allrecipes.com/recipe/86637/honey-grilled-chicken/)

About this result • Feedback



# Uses of schema.org in Google products: Recent examples

## Fact checking

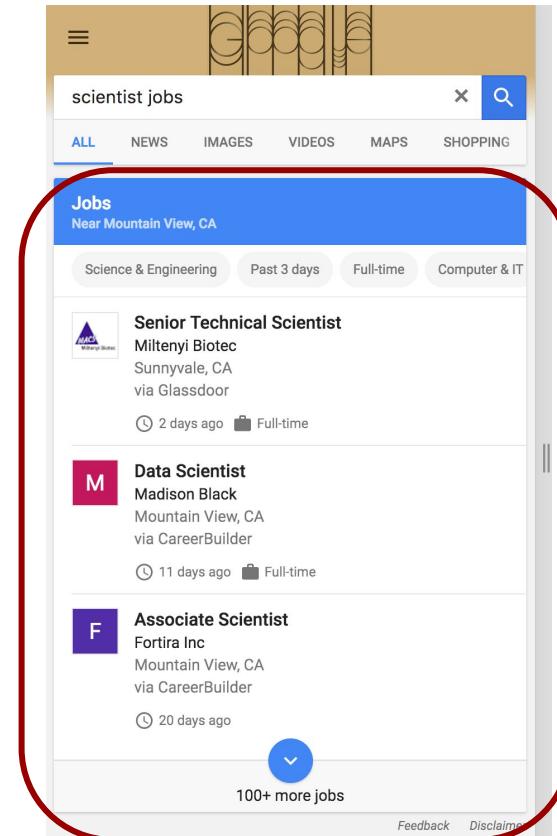


Google search results for "NASA scientists miscalculated eclipse". The top result is a Snopes.com article with a fact-check summary:

**Claim:** NASA scientists "forgot to carry the one," miscalculating the date of an August 2017 solar eclipse by one year.

**Claimed by:** Wyoming Magazine

**Fact check by Snopes.com:** FALSE



Google search results for "scientist jobs" near Mountain View, CA. The results are filtered to show jobs:

**Jobs**  
Near Mountain View, CA

Science & Engineering | Past 3 days | Full-time | Computer & IT

**Senior Technical Scientist**  
Miltenyi Biotec  
Sunnyvale, CA  
via Glassdoor  
2 days ago | Full-time

**Data Scientist**  
Madison Black  
Mountain View, CA  
via CareerBuilder  
11 days ago | Full-time

**Associate Scientist**  
Fortira Inc  
Mountain View, CA  
via CareerBuilder  
20 days ago

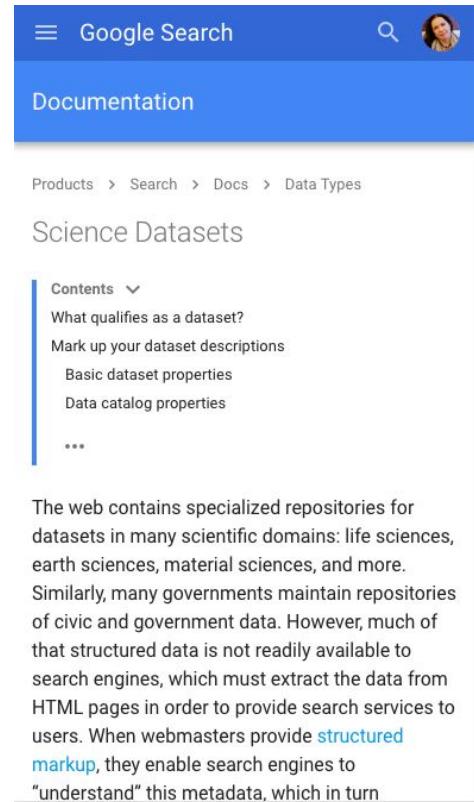
100+ more jobs

## 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](http://bioschemas.org)

## [Guidelines for describing datasets in schema.org](#)



The screenshot shows a navigation bar with 'Google Search' and a user profile icon. Below it, a blue header bar says 'Documentation'. Underneath, a breadcrumb trail shows 'Products > Search > Docs > Data Types'. The main content area is titled 'Science Datasets' and includes a 'Contents' sidebar with links to 'What qualifies as a dataset?', 'Mark up your dataset descriptions', 'Basic dataset properties', 'Data catalog properties', and an ellipsis. The main text area discusses the availability of datasets in various scientific domains and the challenges of extracting structured data from HTML pages using schema markup.

Products > Search > Docs > Data Types

## Science Datasets

Contents

- What qualifies as a dataset?
- Mark up your dataset descriptions
- Basic dataset properties
- Data catalog properties
- ...

The web contains specialized repositories for datasets in many scientific domains: life sciences, earth sciences, material sciences, and more. Similarly, many governments maintain repositories of civic and government data. However, much of that structured data is not readily available to search engines, which must extract the data from HTML pages in order to provide search services to users. When webmasters provide [structured markup](#), they enable search engines to "understand" this metadata, which in turn

# 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](mailto:dataset-support@google.com)