

Google Tools for Data

Hal Varian
16 July 2012



Outline

Tools

Docs: documents, spreadsheets, presentation

Refine: data cleansing

Fusion Tables: database and visualization

Google Scholar

Graphics and APIs

Motion charts: charting

Prediction API: machine learning for data in cloud

Chart Tools: interactive charts for web pages

Data

Insights for Search: query data time series

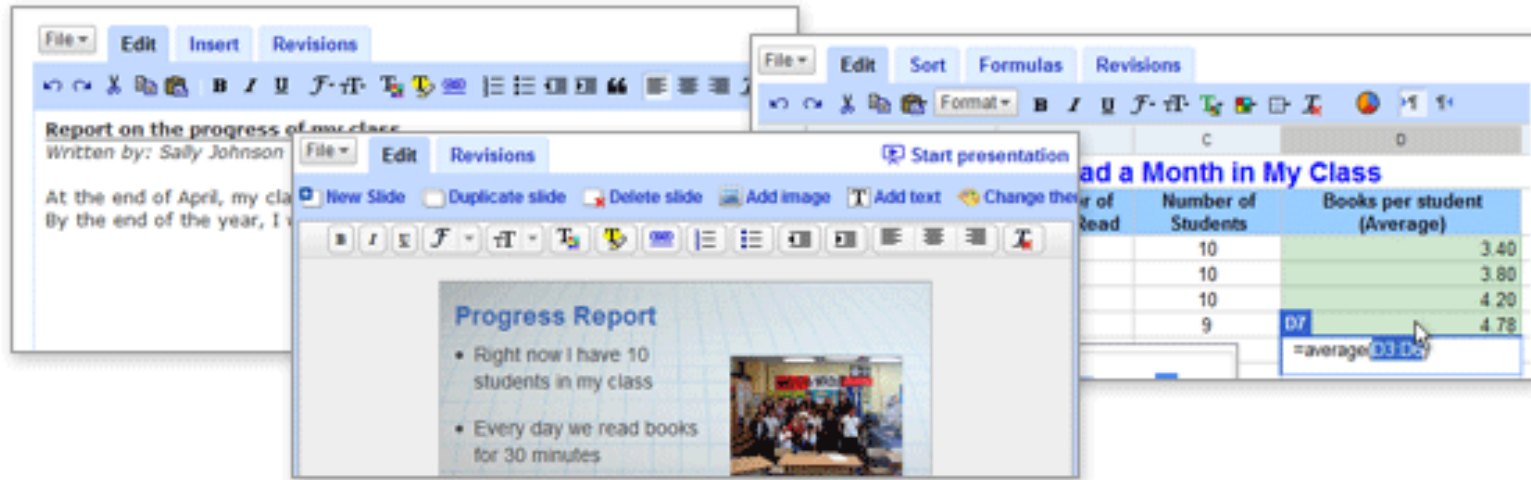
Correlate: query correlations

Ngrams: phrase occurrences in books

Public Data Explorer: find and examine public data

Google Consumer Surveys: run surveys for 10 cents/response

Google Docs



- Integrated suite of word processor, spreadsheet, slides, drawings
- Available in the cloud for multi-authored document creation

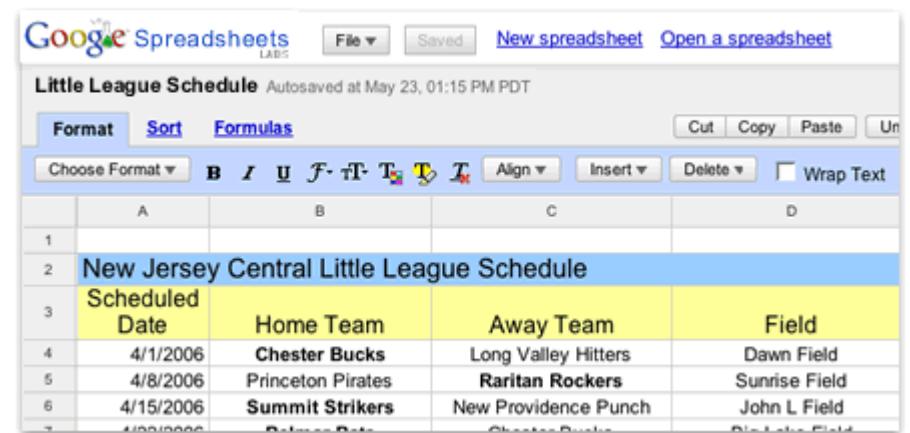
Documents

- Multiauthored documents in cloud
 - Version control
 - Access control
 - Math and drawings
- Move document back and forth between desktop and cloud
- Execute code directly from the cloud
 - LaTeX docs
 - R scripts



Spreadsheets

- Multiauthored calculations in the cloud
 - Collect data from multiple users
 - Collect data from Gmail forms
- Move from desktop to cloud and back
- Access control
- Interactive visualization
- Import data from spreadsheet in R
 - Example of loading R data



The screenshot shows a Google Spreadsheet interface. The title bar indicates the spreadsheet is titled 'Little League Schedule' and was last autosaved on May 23, 2015, at 01:15 PM PDT. The spreadsheet contains a table with the following data:

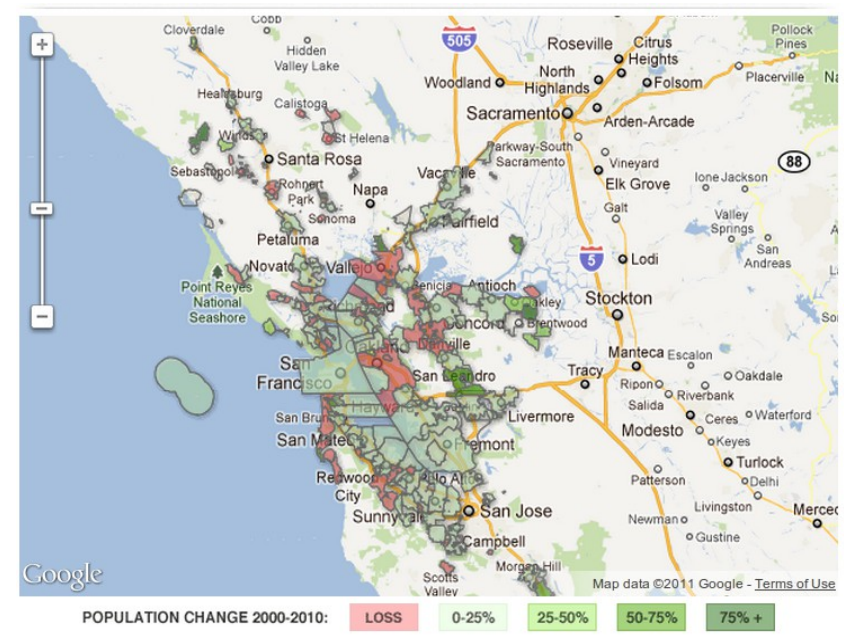
	A	B	C	D
1				
2	New Jersey Central Little League Schedule			
3	Scheduled Date	Home Team	Away Team	Field
4	4/1/2006	Chester Bucks	Long Valley Hitters	Dawn Field
5	4/8/2006	Princeton Pirates	Raritan Rockers	Sunrise Field
6	4/15/2006	Summit Strikers	New Providence Punch	John L Field
7	4/22/2006	Belmont Bats	Chester Bucks	Dawn Field

Presentations

- Multiauthored presentations in the cloud
- Move from desktop to cloud and back
- Access control
- Interactive visualization
- Version control

Fusion Tables

- Capabilities
 - Visualize and publish your data as maps, timelines and charts.
 - Host your data tables online.
 - Combine data from multiple people.
 - Export and import data
- Database in the cloud
 - Access control
 - Annotation



Refine

- Tool for working with messy data
 - Cleaning and normalizing data based on fuzzy matches
 - Transform from one format to another
 - Link to databases and spreadsheets
- Runs on desktop
- See [video](#) for more

Motion charts

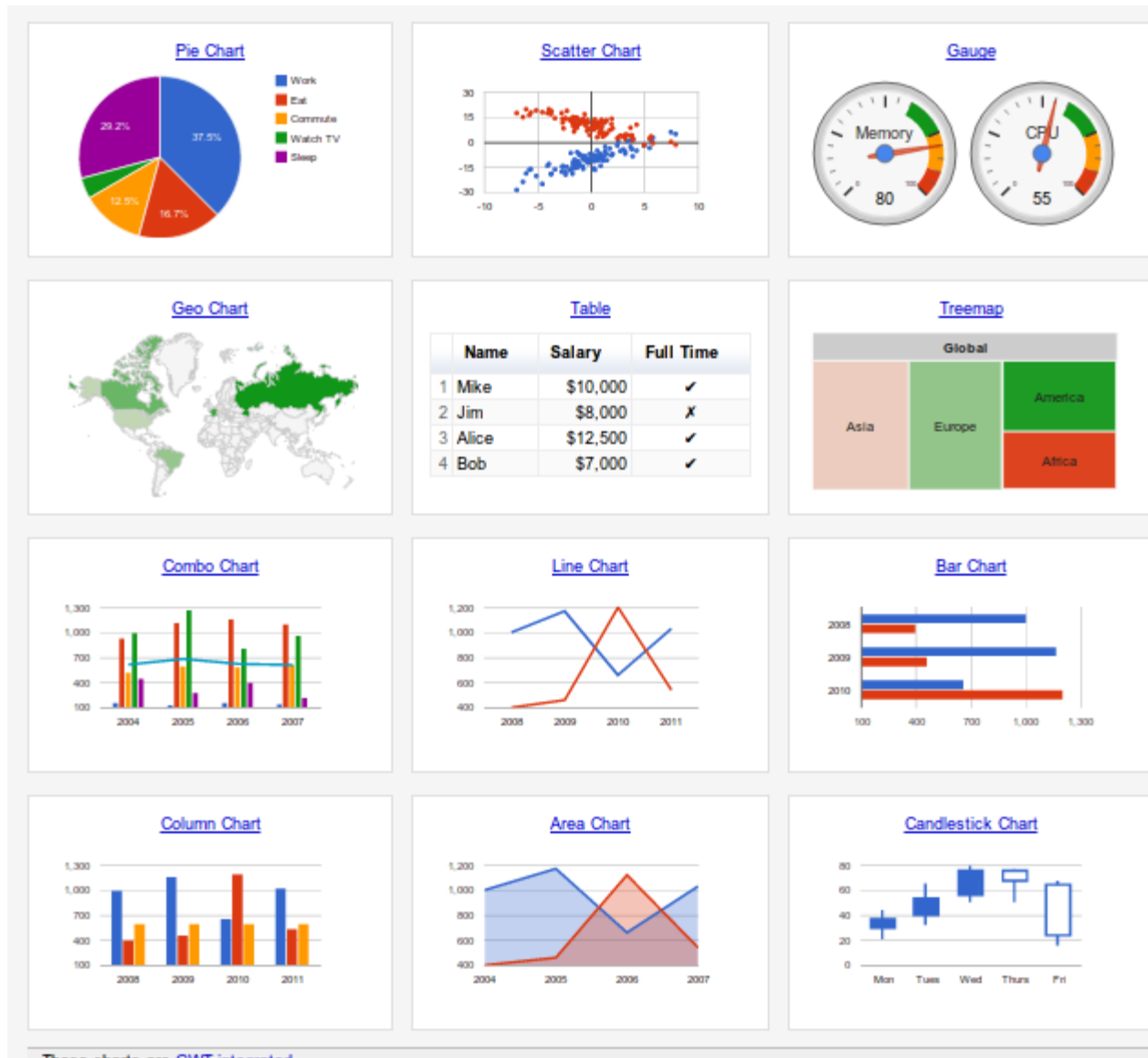
- Accessible from [Spreadsheets](#), [Chart Tools](#), [R](#), and [Public Data Explorer](#)



Prediction API

- Google's cloud-based machine learning tools can help analyze your data to add the following features to your applications:
 - Customer sentiment analysis
 - Spam detection
 - Message routing decisions
 - Upsell opportunity analysis
 - Document and email classification
 - Diagnostics
 - Churn analysis
 - Suspicious activity identification
 - Recommendation systems

Chart Tools for Web sites



Google Search Insights

Google Insights for Search
beta

hal@google.com | [My Account](#) | [Help](#) | [Sign out](#) | [Download as CSV](#) | [English \(US\)](#) ▼

Compare by	Search terms	Filter
<input checked="" type="radio"/> Search terms <input type="radio"/> Locations <input type="radio"/> Time Ranges	Tip: Use quotation marks to match an exact phrase. ("table tennis") <input type="text" value="hangover"/> + Add search term	<div>Web Search</div> <div>United States All subregions All metros</div> <div>Dec 2008 - Feb 2009 Reset</div> <div>All Categories</div> <div>Search</div>

Web Search Interest: hangover

United States, Dec 2008 - Feb 2009

Categories: [Food & Drink \(25-50%\)](#), [Health \(10-25%\)](#), [Entertainment \(0-10%\)](#), [Local \(0-10%\)](#), [more...](#)

Totals [?](#)
hangover 15

Interest over time

☐ Forecast [?](#) ☐ News headlines

[Learn what these numbers mean](#)



[Embed this chart](#)

Google Search Insights

Download search volume index by country/category/term on weekly or daily basis.

Data classified by category, can disambiguate

Can be useful in “nowcasting” economic indicators

- Unemployment

- Auto sales

- Real estate

- Travel planning

Google Correlate

The screenshot shows the Google Correlate web application. The browser's address bar displays the URL `correlate.googlelabs.com/search?e=id:IRcw97A_4R4&t=weekly#default`. The page header includes the Google Correlate logo and a search bar containing the text "initial claims NSA". Below the search bar, there are links for "Compare US states", "Compare time series", and "Shift series" (set to 0 weeks). A sidebar on the left contains links for "Documentation" (Comic Book, FAQ, Tutorial, Whitepaper) and "Correlate Labs" (Search by Drawing). The main content area lists correlations with "Initial claims NSA", starting with a correlation of 0.8841 for "sign up for unemployment". The bottom of the page shows a file manager interface with two files: "correlate-initial....csv" and "herb-simon.jpg", and a button to "Show all downloads...".

initial claims NSA - Google x

correlate.googlelabs.com/search?e=id:IRcw97A_4R4&t=weekly#default

For quick access, place your bookmarks here on the bookmarks bar. [Import bookmark...](#) Other Bookmarks

hal.ucb@gmail.com | [Manage my Correlate data](#) | [Sign out](#)

Google correlate labs

initial claims NSA x Search correlations Edit this data

Compare US states

Compare time series

Shift series 0 weeks

Documentation

- [Comic Book](#)
- [FAQ](#)
- [Tutorial](#)
- [Whitepaper](#)

Correlate Labs

- [Search by Drawing](#)

Correlated with Initial claims NSA

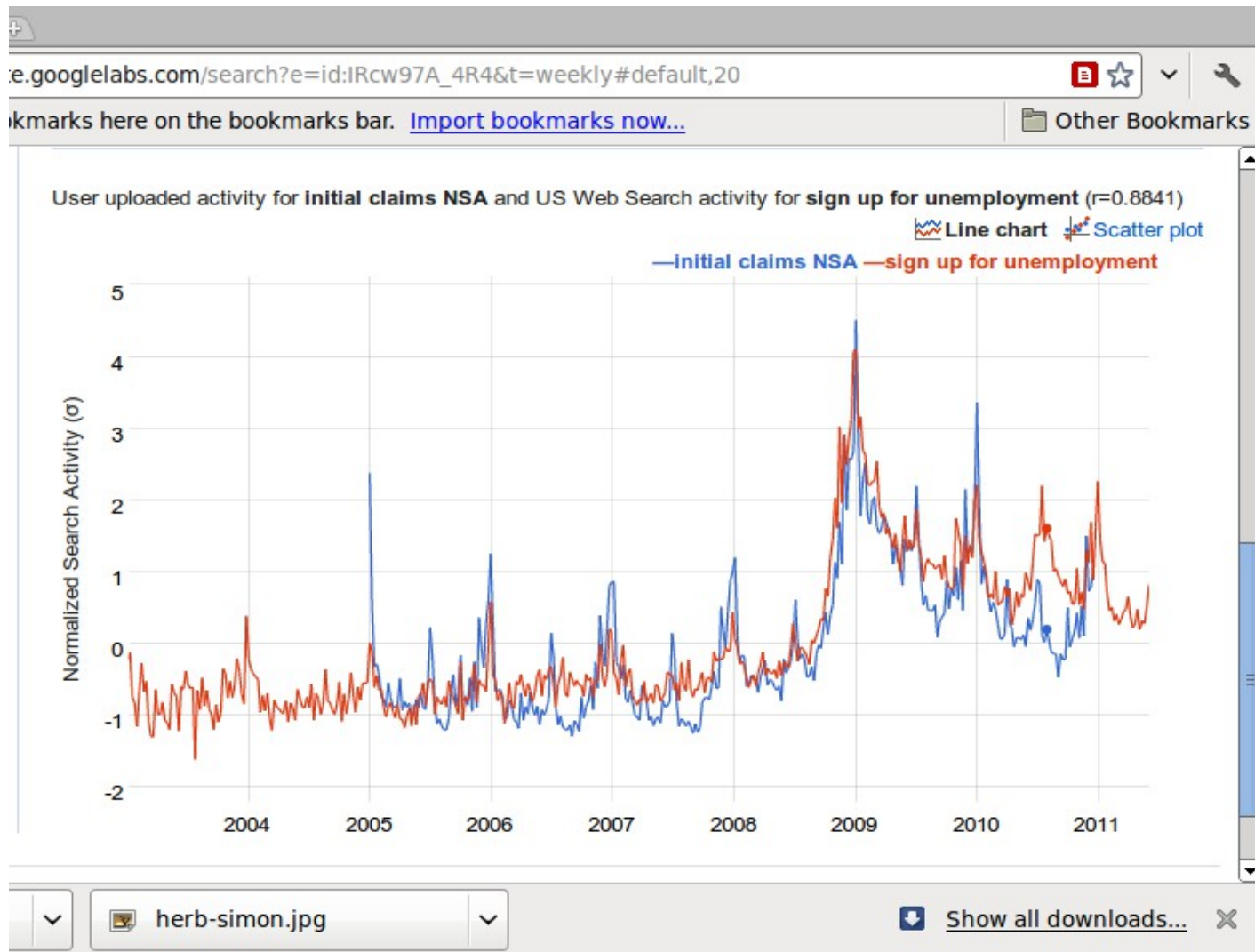
- 0.8841 sign up for unemployment
- 0.8827 idaho unemployment
- 0.8764 michigan unemployment
- 0.8713 unemployment filing
- 0.8678 rhode island unemployment
- 0.8619 unemployment mi
- 0.8610 state of michigan unemployment
- 0.8577 new jersey unemployment
- 0.8573 lalanne juicer
- 0.8547 jack lalanne juicer
- 0.8544 seen on
- 0.8516 unemployment office location
- 0.8508 unemployment application
- 0.8498 department of unemployment
- 0.8497 as seen
- 0.8493 adp flex direct
- 0.8492 rock chucker
- 0.8490 filing unemployment
- 0.8484 peapod coupons
- 0.8477 as seen on

correlate-initial....csv

herb-simon.jpg

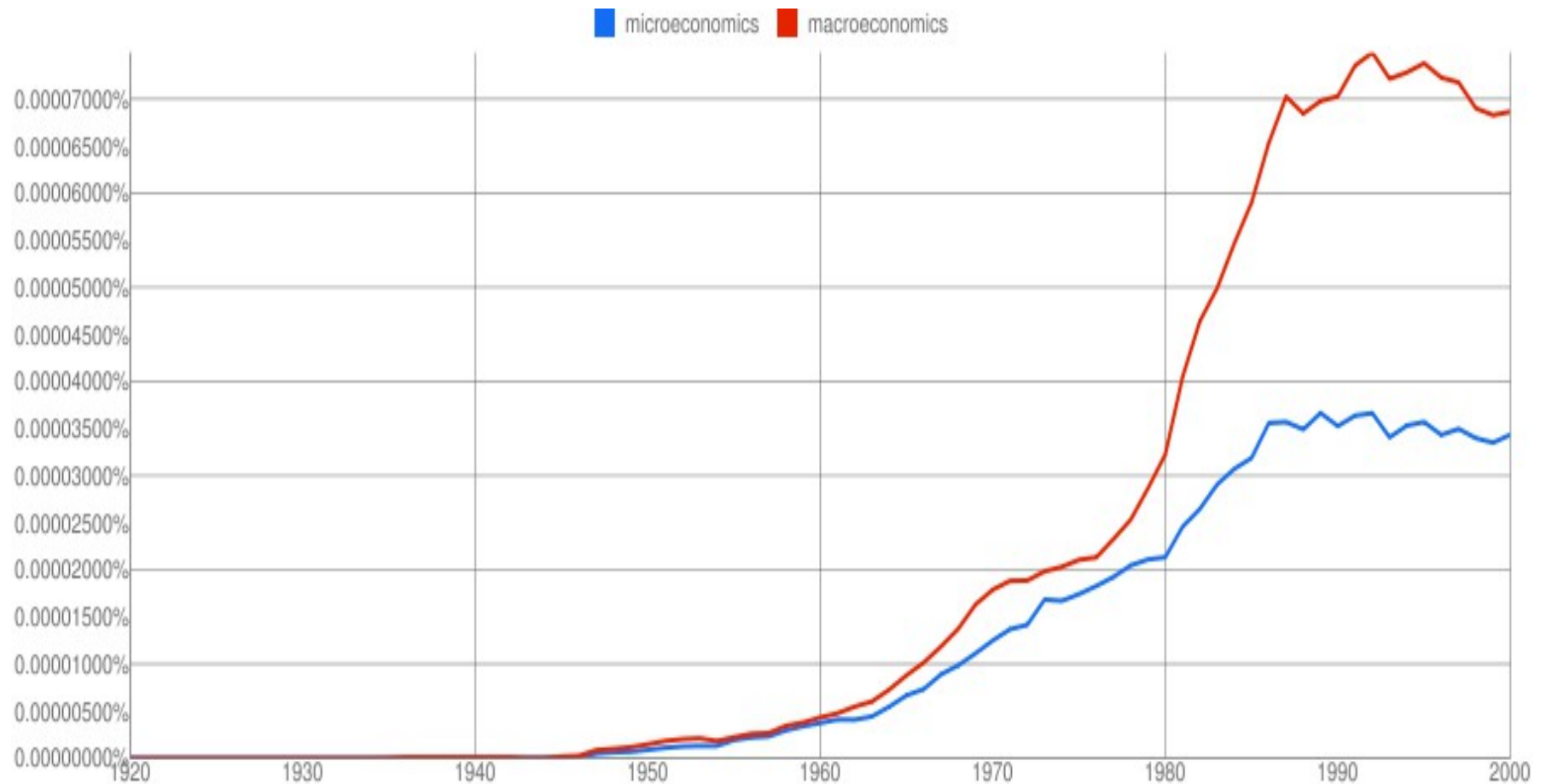
Show all downloads...

Initial claims for unemployment



nGrams

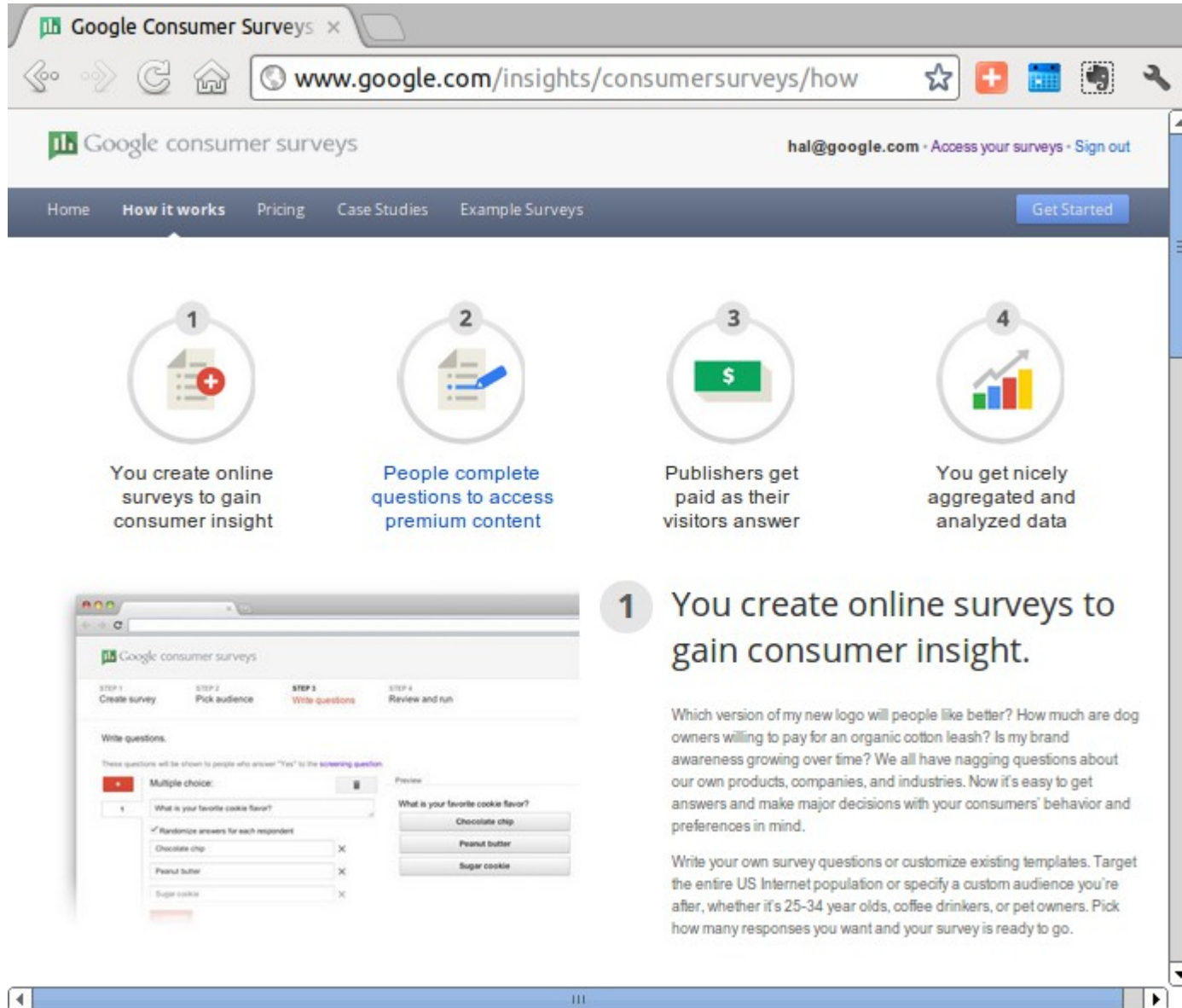
Data from 20 million books



Public Data Explorer

- XML schema for public data
 - World Bank, World Development Indicators
 - International Monetary Fund, September 2011 World Economic Outlook
 - OECD Factbook 2010
 - Unemployment in Europe (monthly)
 - Broadband penetration in Europe
 - Government Debt in Europe
 - Infectious Disease Outbreaks
 - Unemployment in the U.S.
- Visualization tools

Google Consumer Surveys



The screenshot shows the Google Consumer Surveys website. The browser address bar displays www.google.com/insights/consumersurveys/how. The page features a navigation bar with links: Home, How it works, Pricing, Case Studies, Example Surveys, and a Get Started button. Below the navigation bar, four steps are illustrated with icons and text:

- 1** You create online surveys to gain consumer insight. (Icon: Document with a red plus sign)
- 2** People complete questions to access premium content. (Icon: Document with a blue checkmark)
- 3** Publishers get paid as their visitors answer. (Icon: Green dollar sign)
- 4** You get nicely aggregated and analyzed data. (Icon: Bar chart with an upward arrow)

Below the steps, a smaller screenshot shows the 'Write questions' interface. It includes a progress bar with four steps: STEP 1 Create survey, STEP 2 Pick audience, STEP 3 Write questions (active), and STEP 4 Review and run. The 'Write questions' section shows a 'Multiple choice' question: 'What is your favorite cookie flavor?'. The options are: ☒ Randomize answers for each respondent, ☐ Chocolate chip, ☐ Peanut butter, and ☐ Sugar cookie. A 'Preview' section shows the question as it will appear to respondents, with buttons for 'Chocolate chip', 'Peanut butter', and 'Sugar cookie'.

1 You create online surveys to gain consumer insight.

Which version of my new logo will people like better? How much are dog owners willing to pay for an organic cotton leash? Is my brand awareness growing over time? We all have nagging questions about our own products, companies, and industries. Now it's easy to get answers and make major decisions with your consumers' behavior and preferences in mind.

Write your own survey questions or customize existing templates. Target the entire US Internet population or specify a custom audience you're after, whether it's 25-34 year olds, coffee drinkers, or pet owners. Pick how many responses you want and your survey is ready to go.

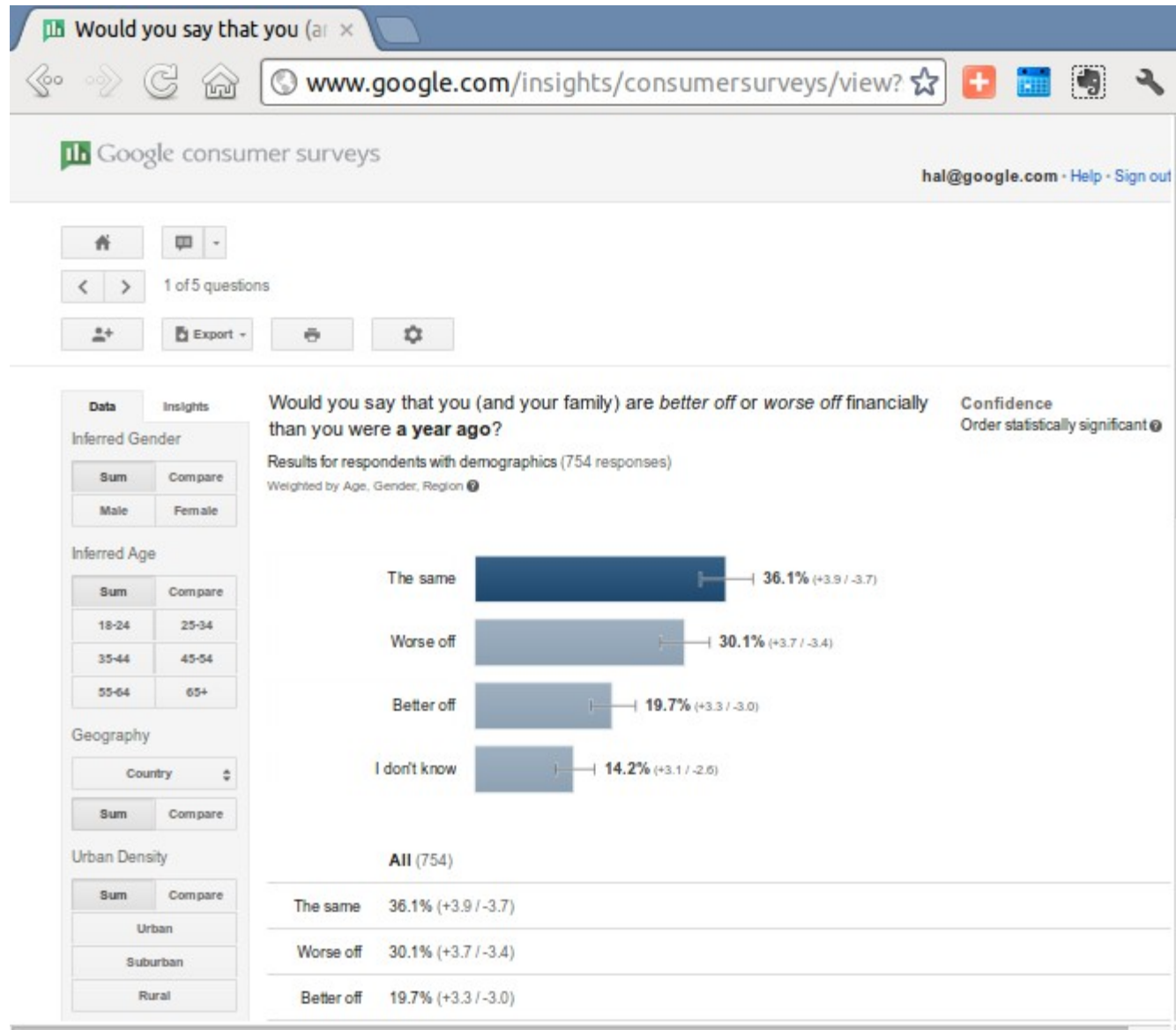
Example survey questions

Consumer sentiment. Would you say that you (and your family) are better off or worse off financially than you were a year ago?

Game theory. A and B are on a TV game show with \$100 to divide. A offers B \$ x . If B accepts the division is carried out. If B refuses each gets zero.

Valuation. Imagine you lived in a world without free search engines like Google and Bing. How much would you pay for a subscription to a search engine?

Consumer Surveys, results



Other web data

Existing

Billion prices project (MIT)
Mastercard Spending Pulse
Monster Employment Index
Intuit Small Business Employment Index
Zillow Real Estate Market Reports

Massachusetts
Institute of
Technology



intuit.

monster



Potential

Wal-Mart, Target, K-Mart retails sales
Price indices from retail data
Package delivery data from UPS, FedEx

WAL★MART



FedEx

Linked in®



Policy issues for discussion

For business “real time” is more important than “historical consistency”

Changes in definitions makes life difficult for researchers

What are incentives for private sector to provide data?

Profit motive (Mastercard, Visa)

Brand identity, thought leadership (Intuit, Monster, Zillow, Google)

Financial reporting to investors (FedEx, UPS, retail)

Some forms of data are subject to manipulation

Search data, non-transactional price data

Econometric issues

Short time series, but many predictors (fat regression)

Model averaging v economic science

Data Requests to Google

Google likes to release data to everyone or no one...

“Managing the world's information...”

Very hard to do custom requests

Clearance, external contracts, engineering resources, privacy, attention, etc.

Very interested in supporting research community

Feature requests for products

Google Faculty Grants – access on-site data (Mt View, Cambridge, New York, Pittsburgh, Ann Arbor, Chicago, Boulder, Irvine, Venice CA, London, Paris, etc.)