Finding Structure in Webpages

Josh Ain, Software Engineer

Google's Mission

To organize the world’s information and make it universally accessible and useful.

Today's Talk

- Either
  - Structured data on the web
  - Building a new product at Google post-acquisition
- Building software at Google

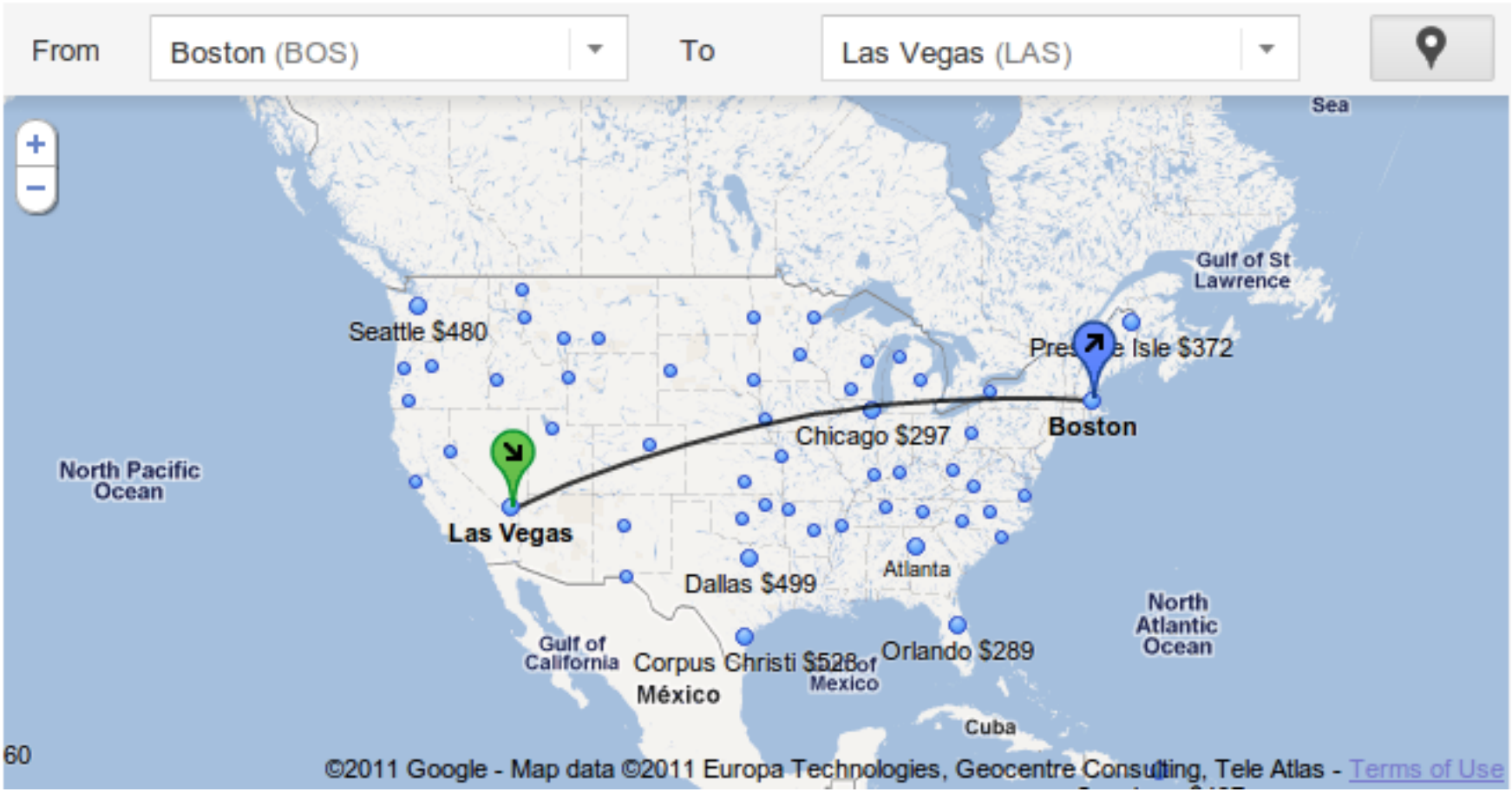
Background

ITA Software by Google

- Handles most online flight search

- Acquired by Google in April 2011 for 700m
- Fourth largest acquisition in Google history
- Run as its own brand

Google Flights



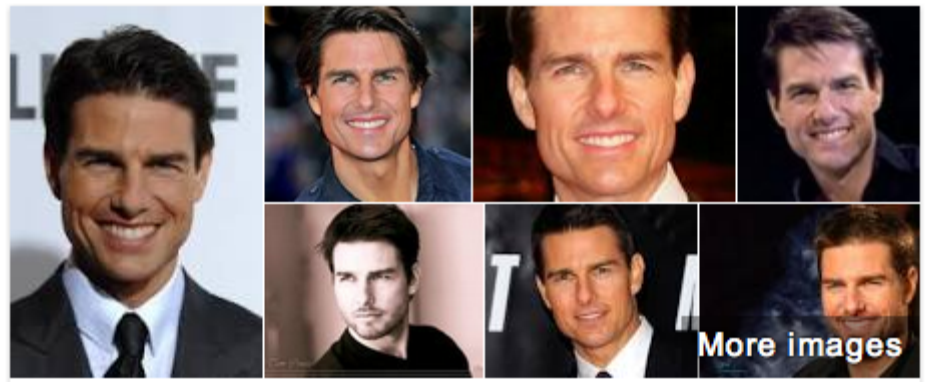
Needlebase

"The WordPress or YouTube of data extraction and manipulation"
– Marshall Kirkpatrick of ReadWriteWeb

- A web-based data platform for non-programmers
- Lets users extract and merge structured data from websites, spreadsheets, and feeds
- Data is easy to dedupe, clean, explore, and publish

Structured data on the web

Knowledge graph



Tom Cruise

Thomas Cruise Mapother IV, widely known as Tom Cruise, is an American film actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Taps. Wikipedia

Born: July 3, 1962 (age 50), Syracuse

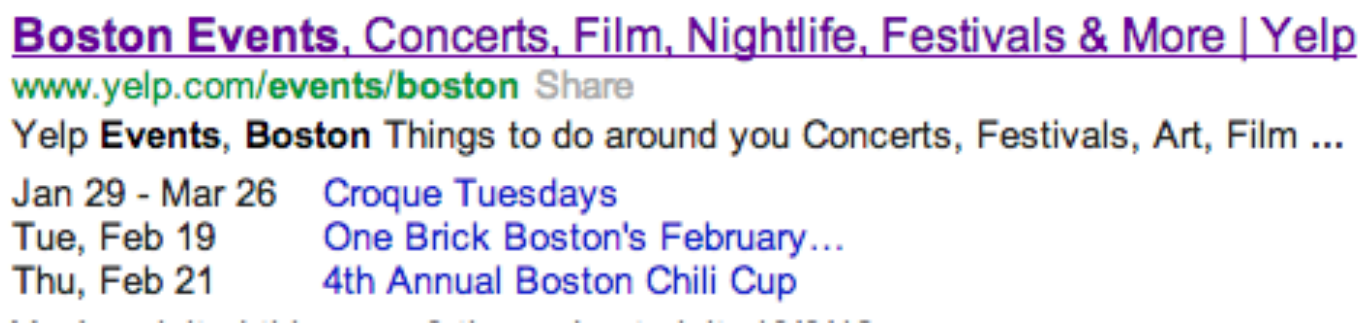
Height: 5' 7" (1.70 m)

Upcoming movies: Oblivion, All You Need Is Kill

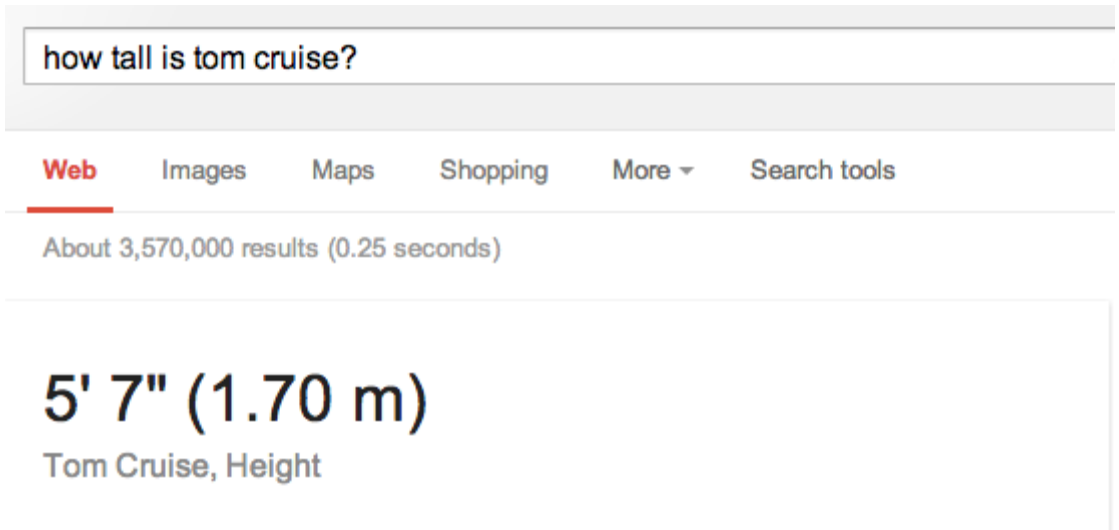
Children: Suri Cruise, Isabella Jane Cruise, Connor Cruise

Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)

Rich snippets



Question answering



Where does this data come from?

- User input
  - Freebase
  - Wikipedia
  - Feedback
  - ...
- Feeds
- Webmaster markup

Webmaster contributions

- Structured data testing tool
- Webmaster tools structured data dashboard
- Webmaster tools Data Highlighter

From Needlebase to Data Highlighter

Finding a place within Google

- Shopped around
- Found those doing compatible work

Reconceptualize product

- Be very good at one thing
- Simplify simplify simplify
- Iterate
- Given Google technology, what is now possible?

Working at Google

Technical work

- C++
- Java
- Python
- Javascript
- Go

- Various types of jobs
- Empowered engineers

My technical work

- Java
  - Eclipse
- Closure templates
- Closure javascript
- Virtual machines

Lots of external dependencies

Lifecycle of a project

- Project definition
- Specs / designs
- Approvals (not what you think)
- Implementation
- User studies / performance studies
- Iterate

Role of an engineer

- Coding
- Product definition
- Product architecture
- Product testing
- Technical specification
- Demonstrating impact

Role of an Engineer

Communication

Slowdowns

- Code standards
- Approvals
- Massive codebase

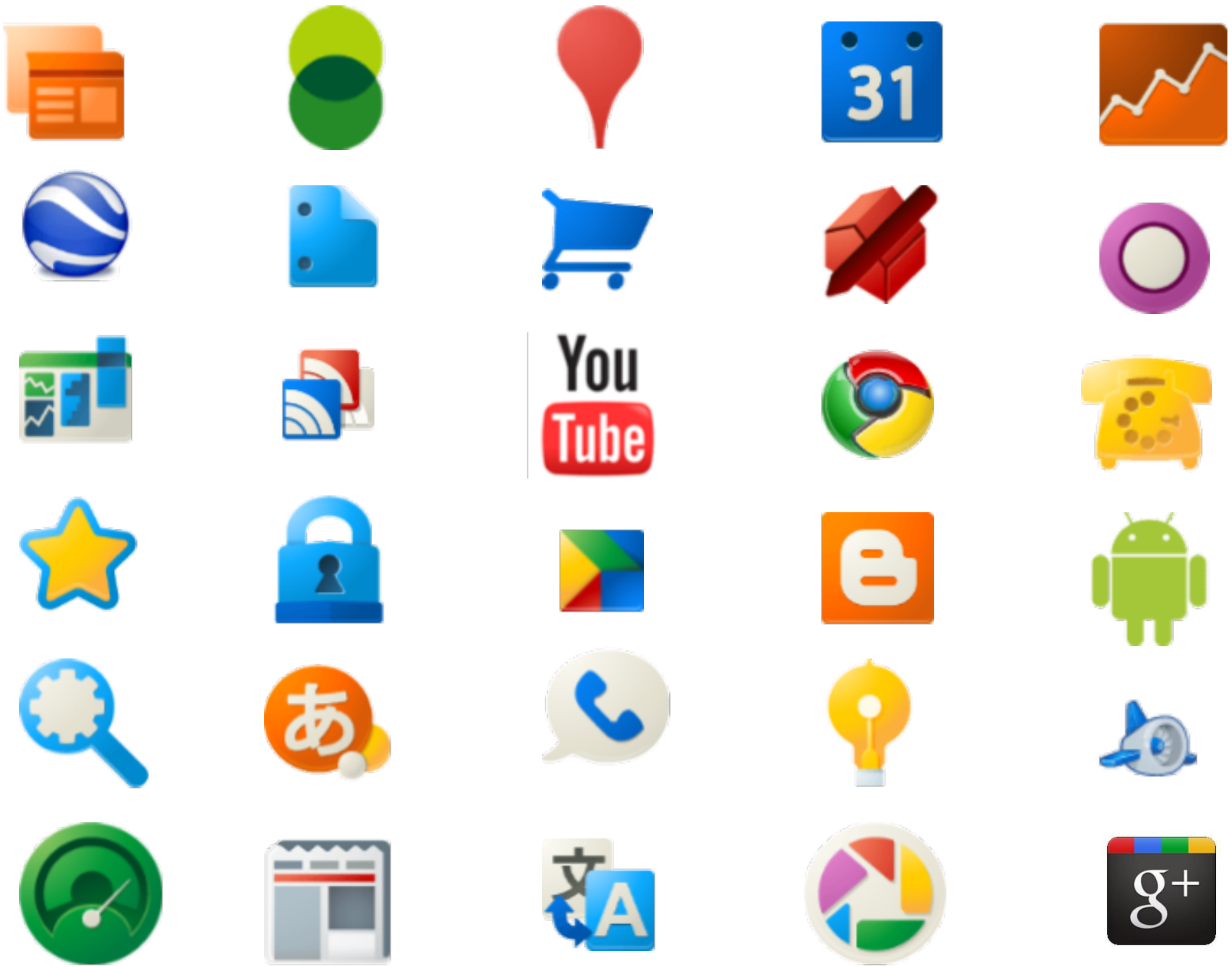
Think as big as you want to

- Can easily analyze every page on the web
- New products have huge reach
- Your work can have impact

Benefits: better than the expected

Below are only some of the amazing benefits you can expect as a full-time employee
Education ● Up to $12000 per year– in pursuit of a degree complementing your job (incl. tuition, books, lab fees, parking etc.)
Fitness & Nutrition ● On-site gyms in various offices (gym subsidies in other locations) ● Intramural sport leagues: basketball, ping-pong, bowling, soccer, etc. ● Free food!
Family ● Maternity leave up to 20 weeks ● Paternity leave
Transportation ● Free wi-fi equipped shuttle to/from work (different Bay Area locations only)
401(k) ● 100% employer match up to $3000; 50% match up to $8,250
Vacation ● Year 1-3: 15 days (3 weeks) ● Year 3-5: 20 days (4 weeks) ● 5+ Years: 25 days (5 weeks) ● Holidays: ~12 paid holidays ● **Sick days**: taken as necessary

This is also Google…



(and much, much more).

Google celebrates diversity

At Google, we don't just accept difference – we thrive on it, we celebrate it, and we support it for the benefit of our employees, our products, and our community.

- **19** Global Employee Resource Groups, such as BGN, HGN, & Gayglers
- Over **3,000** Googlers volunteering to help promote technology education
- **49** RISE Awards were given in 2010 in Europe & the U.S.
- **1,000+** students reached through Google Classroom Outreach Initiative
- **10M** students using Google Apps for Education
- Interacted with **65,000+** students at Science Fairs globally
- **2,100** Google Scholars since 2005, **$8.8M** in scholarship dollars awarded
- **18** global scholarships & awards programs
- **800** Anita Borg Scholars since 2005
- **$500,000+** in cash donations to **6** HBCUs
- **5,000+** Googlers attended Sum of Google at **22** Google offices
- **300** Googlers marched in San Francisco Bay Area Pride Parade

Learn more: google.com/diversity/students

Our North America offices

North America Locations:

- Atlanta, GA
- Boston/Cambridge, MA
- Boulder, CO
- Chapel Hill, NC
- Chicago, IL
- Irvine, CA
- Montréal, QC
- Mountain View, CA
- New York, NY
- Pittsburgh, PA
- San Bruno, CA
- San Francisco, CA
- Santa Monica, CA
- Seattle and Kirkland, WA
- Waterloo, ON

Google Boston/Cambridge

- Located in Kendall Square near MIT with great views of the Charles River and Boston Skyline!
- Boston Googlers are offered boot camp, yoga classes and kickboxing.
- Major Projects: Travel/Flights, Image Search, Google Web Server, Chrome, Mobile, Book Search, Ads Quality

Application Process

Full-Time

1. Apply! Submit your resume and unofficial transcripts. www.google.com/students/eng
2. General fit technical interviews
3. 2nd round technical interviews
4. Offer

Internships

1. Apply! Submit your resume and unofficial transcripts. www.google.com/students/eng
2. Technical interviews on-campus or phone
3. Host & project matching Discuss project & location preferences
4. Offer

Technical Interview Tips

- Review the fundamentals: algorithms and data structures
- Use your strongest language and don't use pseudo-code. We want actual code.
- Prepare to use a whiteboard.
- We're interested in how you approach problem-solving. Think out loud. Ask questions. State assumptions and reasoning. Mention various approaches.

Q&A

Josh Ain, joshain@google.com

Appendix

Machine Learning

- **Induce a page grammar**
- Train page parser (HMM)
- Transform extracted data into target schema

Page grammar transformations:

- Permute
- Factor
- Factor2D
- Interleave
- Lift
- Choice
- Optionalize
- Unloop
- ...

Machine Learning

- Induce a page grammar
- **Train page parser (HMM)**
- Transform extracted data into target schema