



# Making AJAX crawlable

**Katharina Probst**

Engineer, Google

**Bruce Johnson**

Engineering Manager, Google

in collaboration with:

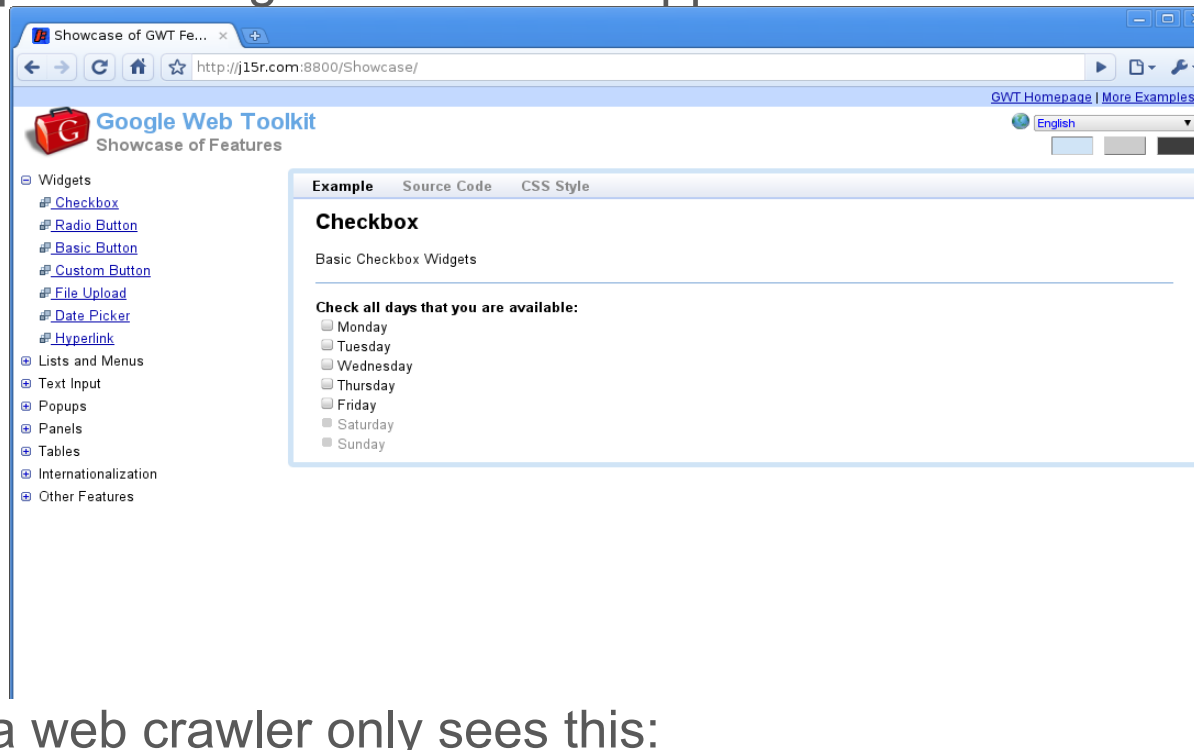
**Arup Mukherjee, Erik van der Poel, Li Xiao, Google**

# The problem of AJAX for web crawlers



Web crawlers don't always see what the user sees

- JavaScript produces dynamic content that is not seen by crawlers
- Example: A Google Web Toolkit application that looks like this to a user...



...but a web crawler only sees this:

```
<script src='showcase.js'></script>
```

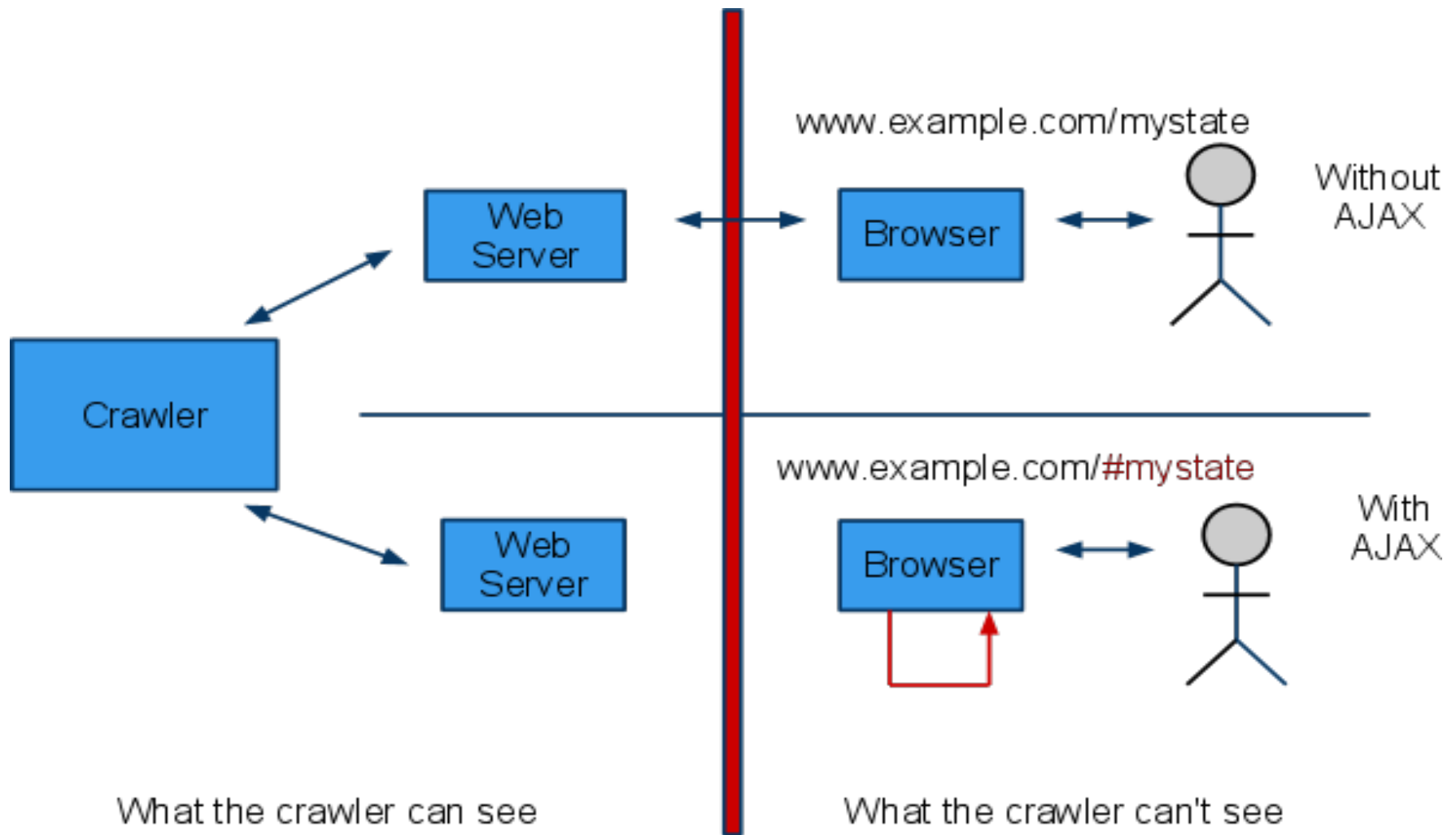
# Why does this problem need to be solved?

---



- **Web 2.0: More content on the web is created dynamically (~69%)**
  - **Over time, this hurts search**
  - **Developers are discouraged from building dynamic apps**
  
  - **Not solving AJAX crawlability holds back progress on the web!**
-

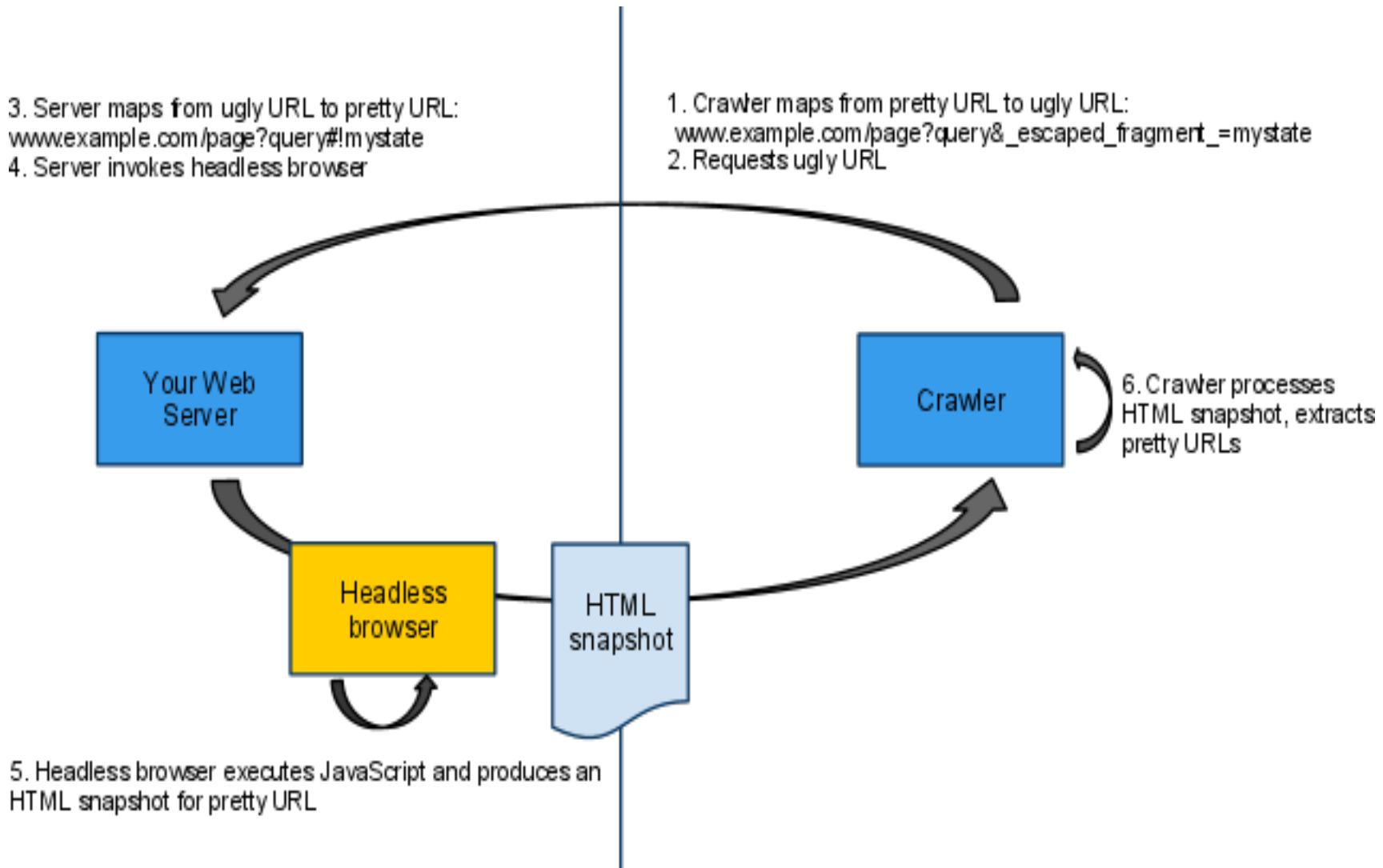
# A crawler's view of the web - with and without AJAX



- **Crawling and indexing AJAX is needed for users and developers**
  - **Problem: Which AJAX states can be indexed?**
    - Explicit opt-in needed by the web server
  - **Problem: Don't want to cloak**
    - Users and search engine crawlers need to see the same content
  - **Problem: How could the logistics work?**
    - That's the remainder of the presentation
-

- **Crawlers execute all the web's JavaScript**
    - This is expensive and time-consuming
    - Only major search engines would even be able to do this, and probably only partially
    - Indexes would be more stale, resulting in worse search results
  - **Web servers execute their own JavaScript at crawl time**
    - Avoids above problems
    - Gives more control to webmasters
    - Can be done automatically
    - Does not require ongoing maintenance
-

# Overview of proposed approach - crawl time

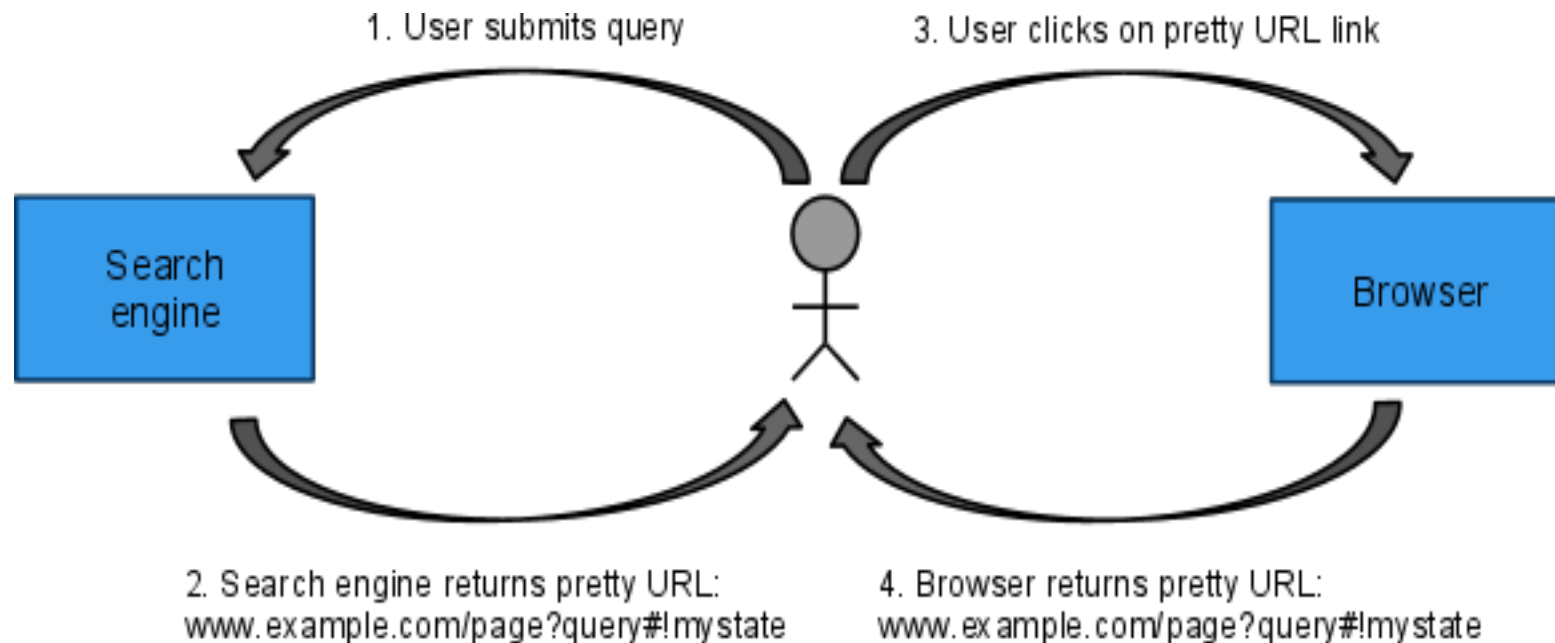


Crawling is enabled by mapping between

- "pretty" URLs: `www.example.com/page?query#!mystate`
- "ugly" URLs: `www.example.com/page?query&_escaped_fragment_=mystate`

# Overview of proposed approach - search time

---



*Nothing changes!*

---

- **Web servers agree to**

- opt in by indicating indexable states
- execute JavaScript for ugly URLs (no user agent sniffing!)
- not cloak by always giving same content to browser and crawler regardless of request (or risk elimination, as before)

- **Search engines agree to**

- discover URLs as before (Sitemaps, hyperlinks)
  - modify pretty URLs to ugly URLs
  - index content
  - display pretty URLs
-

`http://example.com/stocks.html#GOOG`

**could easily be changed to**

`http://example.com/stocks.html#!GOOG`

**which can be crawled as**

`http://example.com/stocks.html?_escaped_fragment_=GOOG`

**but will be displayed in the search results as**

`http://example.com/stocks.html#!GOOG`

---

# Feedback is welcome

---



- We are currently working on a proposal and prototype implementation
  - Check out the blog post on the Google Webmaster Central Blog: <http://googlewebmastercentral.blogspot.com>
  - We welcome feedback from the community at the Google Webmaster Help Forum (link is posted in the blog entry)
-