

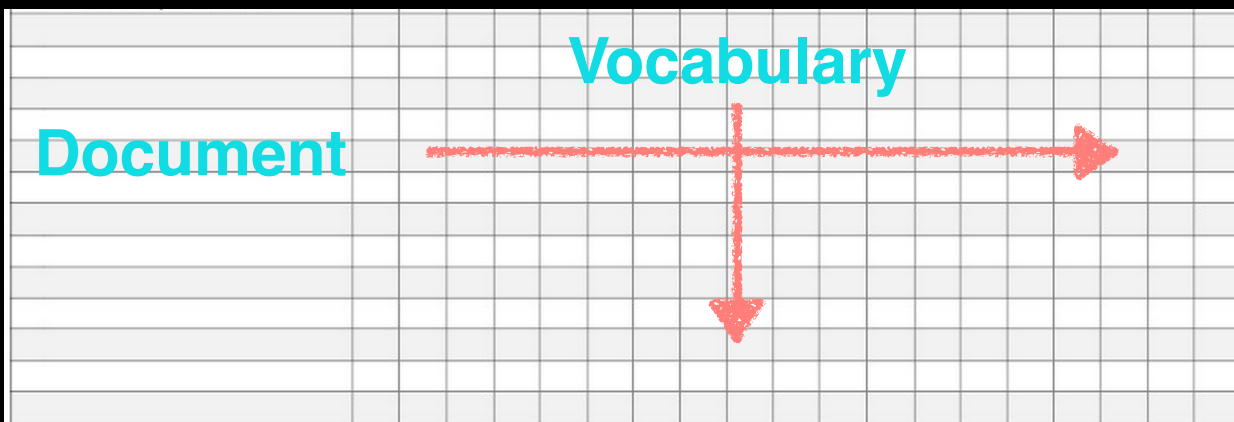
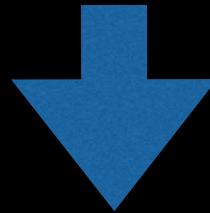
Natural Language Processing (NLP)

Has the Measles Outbreak Changed Your Views on Vaccination? Describe Your Experience

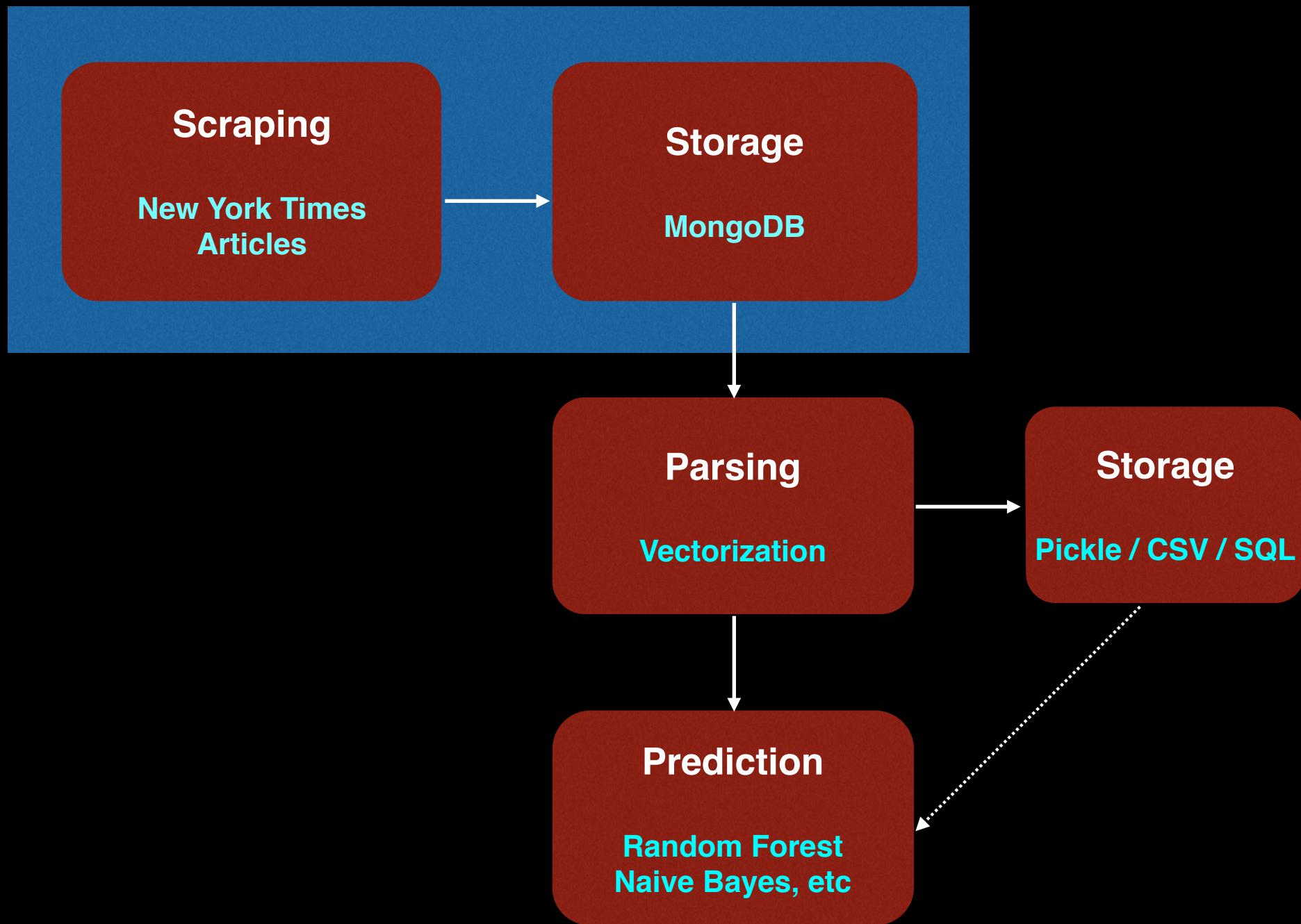
New York Times journalists would like to hear from parents, particularly those in California and Arizona, who have chosen not to vaccinate their children against measles and other diseases.

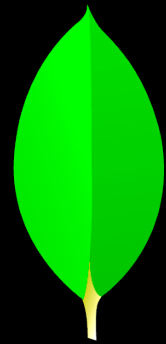


An outbreak of [measles](#) several weeks ago [at Disneyland in Southern California](#) focused minds and deepened concerns. It was as if the amusement park had become the tragic kingdom. Dozens of measles cases have spread across California. Arizona and other nearby states reported their own eruptions of this nasty illness, which officialdom had pronounced essentially eradicated in this country as recently as 2000.



Data Pipeline





mongoDB

NoSQL Database

- Document-based database (json)
- Semi-scalable
- Good for storing unstructured data
- Schema-less
- No joins
- Suboptimal for complicated queries
- No transactions

Transaction

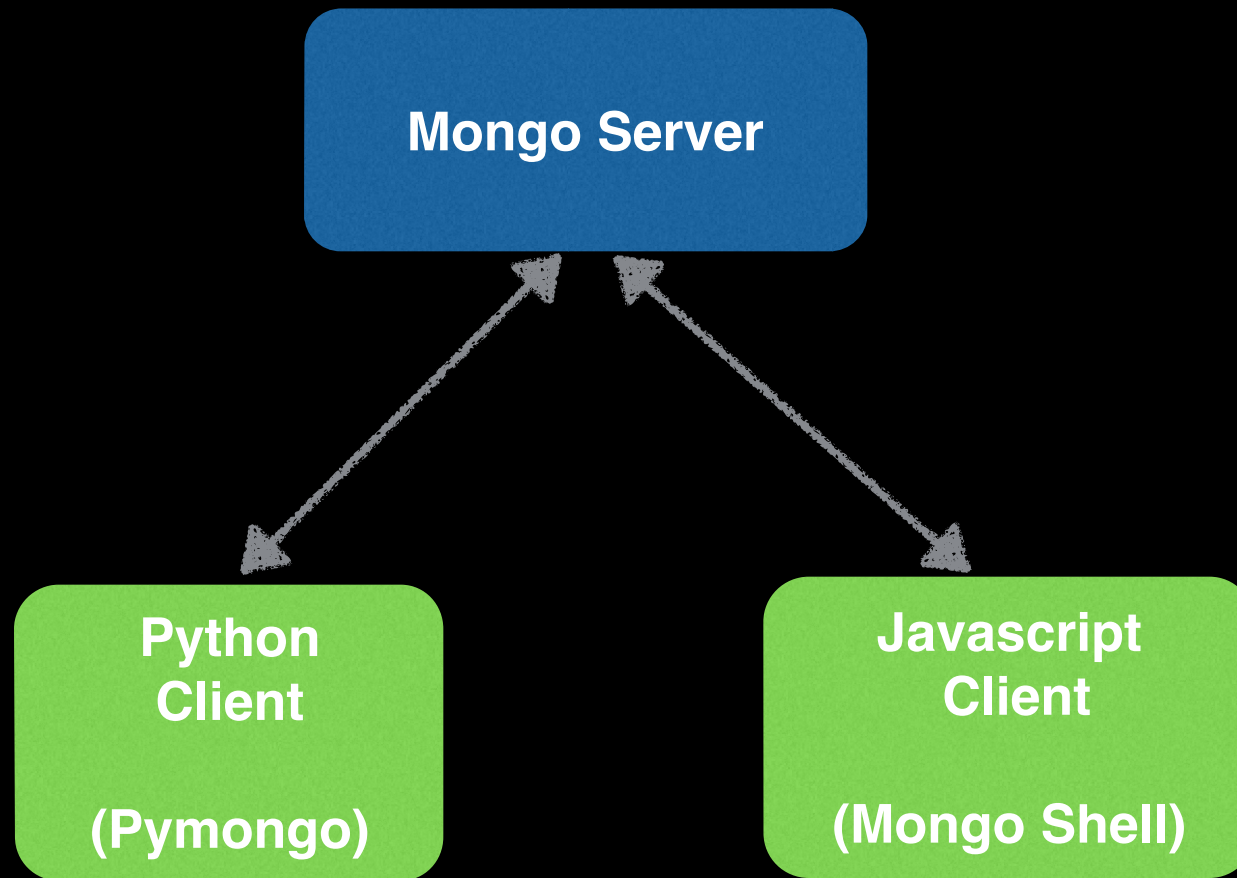
BEGIN;

QUERY: ACCOUNT A - \$300

QUERY: ACCOUNT B + \$300

COMMIT;

Mongo Clients



Live Coding
(See [lecture.md](#))

Web Scraping

Goals

- Appreciate Internet Infrastructure
- Understand Internet Client and Server
- Use of CSS Selector for web scraping

Web VS Internet

- Web is www (World Wide Web)
- Different from Internet
- Web as collection of islands and internet as bridges connecting the islands

HTTP: The language of the Web

protocol



http://mysite.com:80/index.html

URI: Uniform Resource Identifier

file://test.txt

ftp://user@domain.com

protocol



<http://mysite.com:80/index.html>



host

protocol

port

http://mysite.com:80/index.html

host

protocol

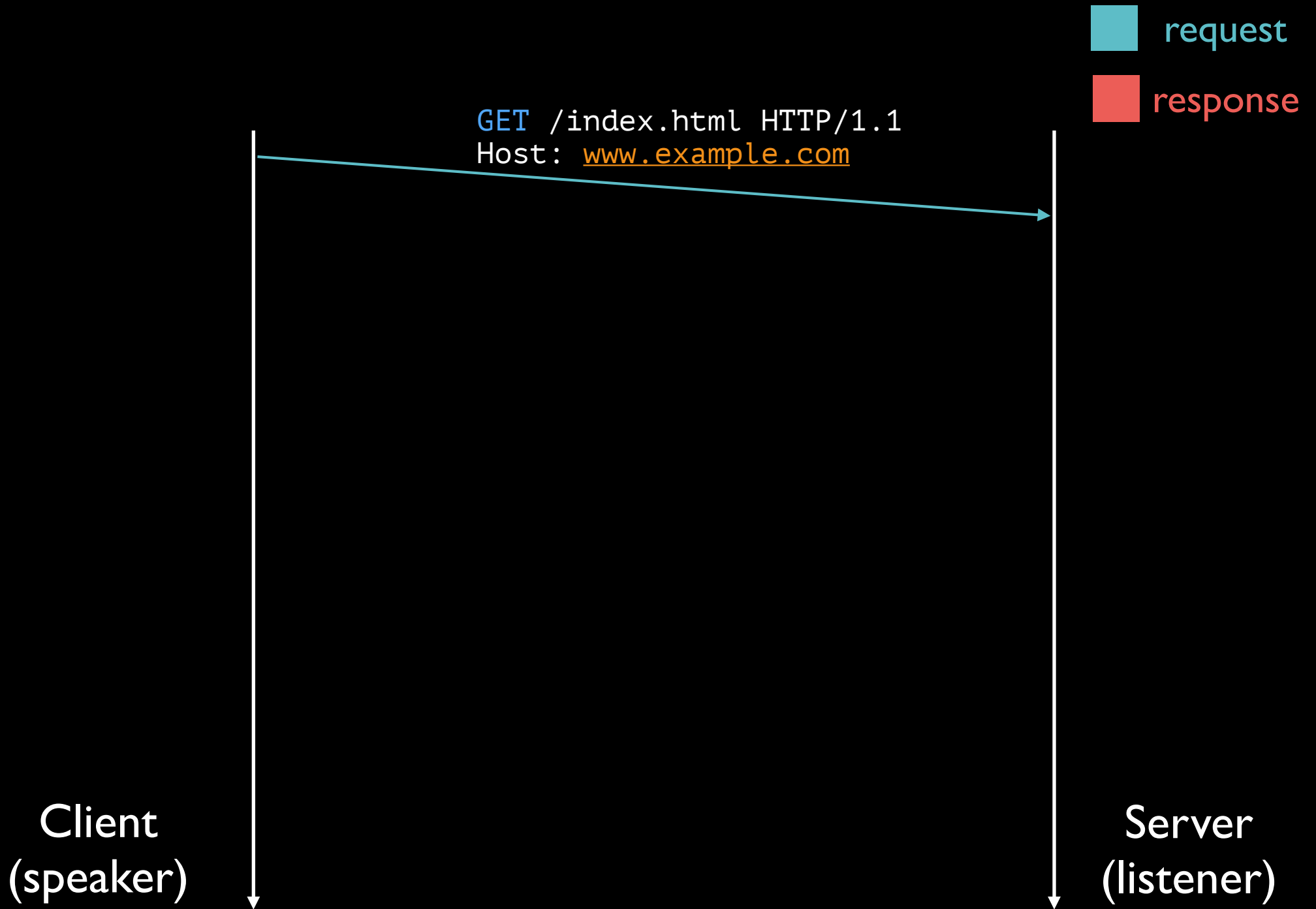
port


http://mysite.com:80/index.html

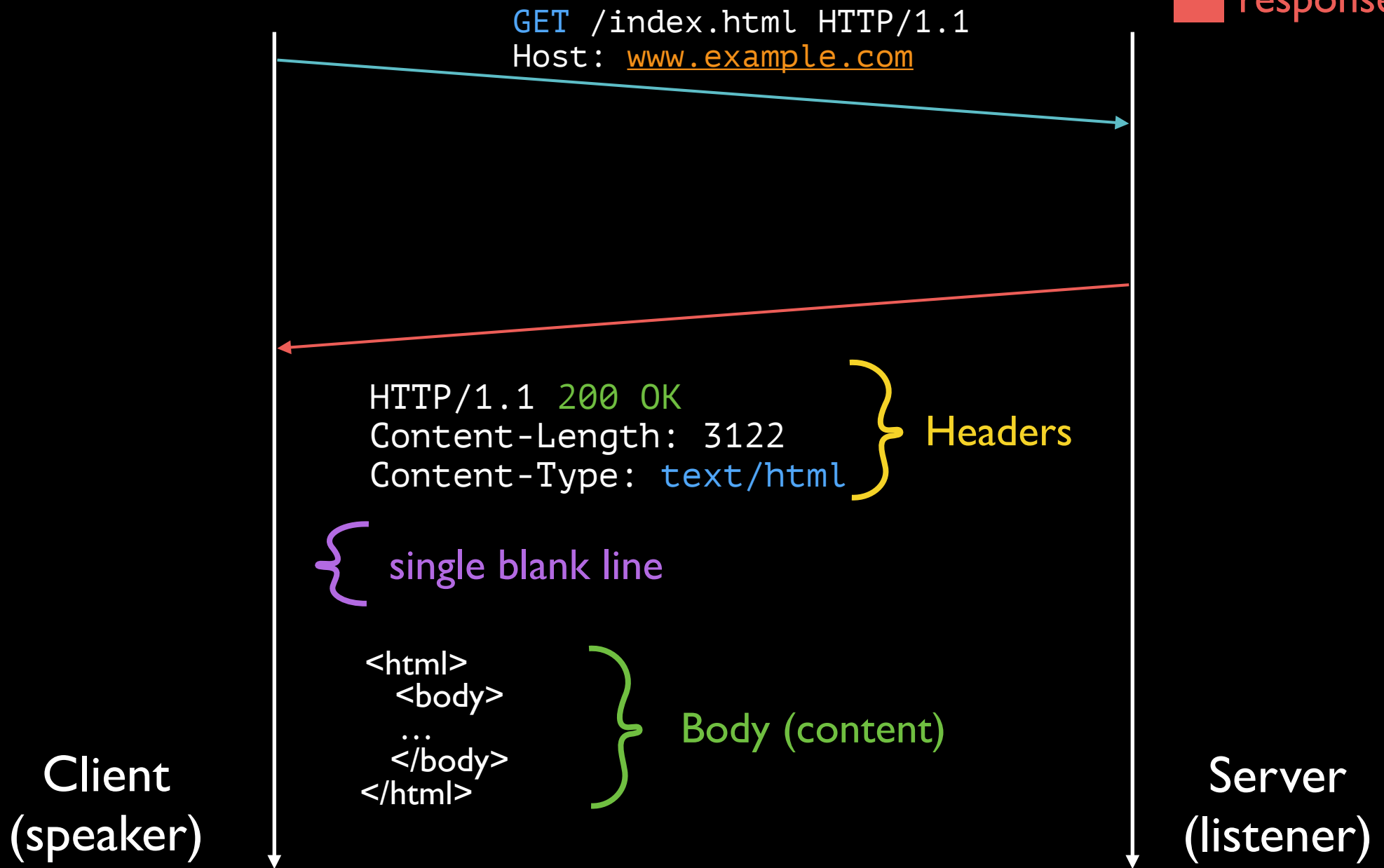
The diagram shows the URL 'http://mysite.com:80/index.html' with four color-coded components and their corresponding labels above and below. The 'http' is yellow, 'mysite.com' is pink, '80' is blue, and 'index.html' is green. A yellow bracket above 'http' points to the 'protocol' label. A blue bracket above '80' points to the 'port' label. A pink bracket below 'mysite.com' points to the 'host' label. A green bracket below 'index.html' points to the 'path' label.

host

path



request
response



request
response

GET /index.html HTTP/1.1
Host: www.example.com

HTTP/1.1 200 OK
Content-Length: 3122
Content-Type: text/html

<html>
<body>
...
</body>
</html>

initial page load

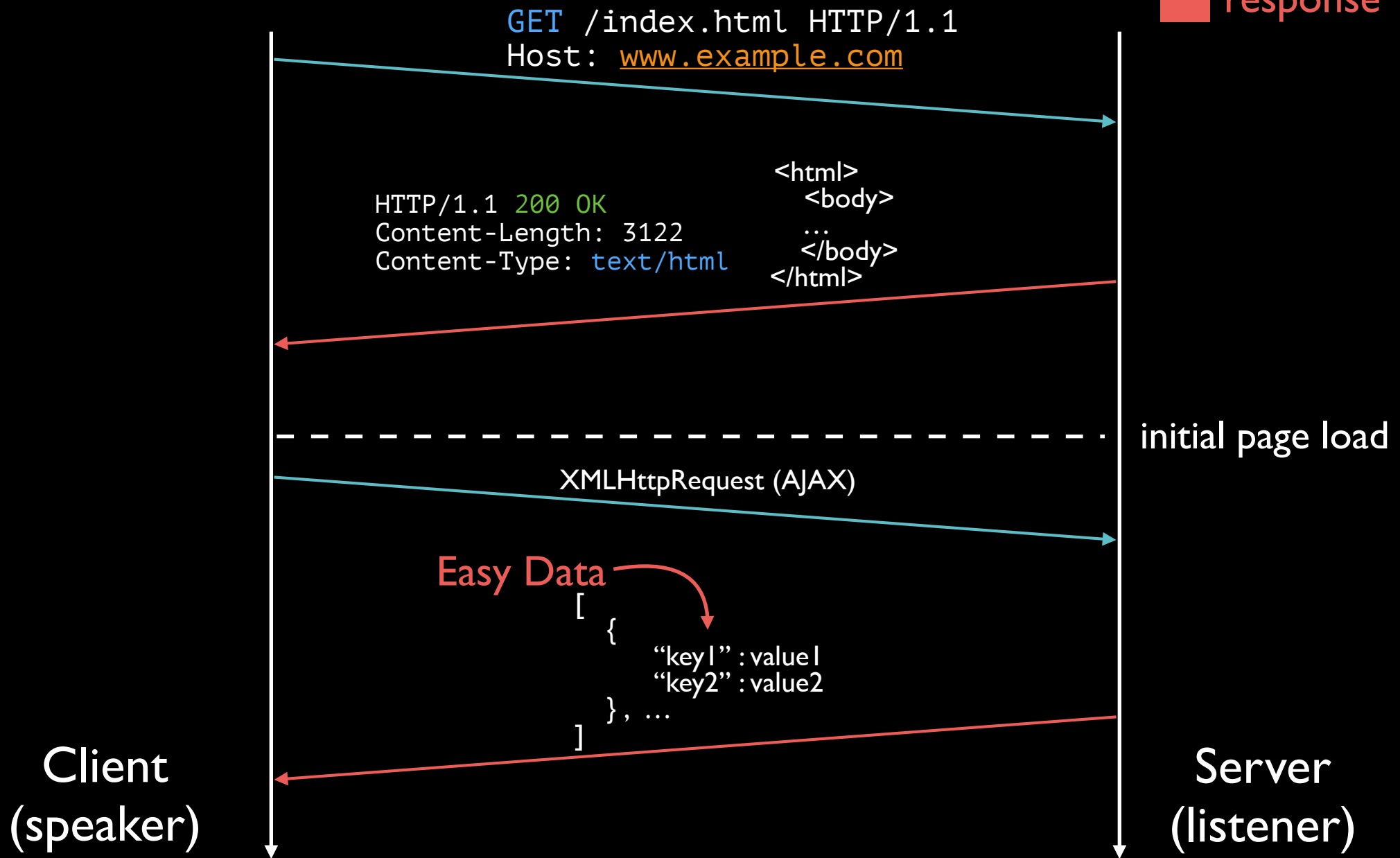
XMLHttpRequest (AJAX)

AJAX = **A**synchronous **J**avaScript **A**nd **X**ML

Client
(speaker)

Server
(listener)

request
response



HTTP: Stateless Protocol

- A website does not remember who you are when you visit again
- Cookies fix that by storing a packet of info on your laptop

CSS Selectors

- CSS = Cascading Style Sheet
- CSS is used for formatting web pages
- CSS Selector selects elements on web pages

Resources for CSS selector

- CSS Cheat Sheet in the repo
- **Game of fruit:** <http://flukeout.github.io/>

Chrome DevTools

Parts & Accessories (191,236)

Automotive Tools & Supplies (317)

Condition [see all](#)

☐ New (183,369)

☐ Used (7,480)

☐ Not Specified (700)

Price

to \$ [>>](#)

Format [see all](#)


191,549 results for shoes [+ Follow this search](#)

[a.img.imgWr2](#) 225px x 225px

NEW LISTING [HARKEN, Classic LEATHER Boat Shoes, Size 9.5](#)

\$39.50

Buy It Now



Elements Network Sources Timeline Profiles Resources Audits Console

<top frame> [▼](#) ☐ Preserve log

Uncaught **SyntaxError: Unexpected token ILLEGAL**

<http://www.ebay.com/itm/HARKEN-Classic-LEATHER-Boat-Shoes-Size-9-5-/291449772901?hash=item43dbc29b65&vxp=mtr> class

http://www.ebay.com/itm/Tech-6-Motorcycle-Leather-Alpinestars-Boots-Si...Condition-/141652631322?_LH_DefaultDomain_0&hash=item20fb279f1a&vxp=mtr class="img imgWr2">...

Scraping Demo