Webscraping 101
(with some Python)

Anna Vassilovski

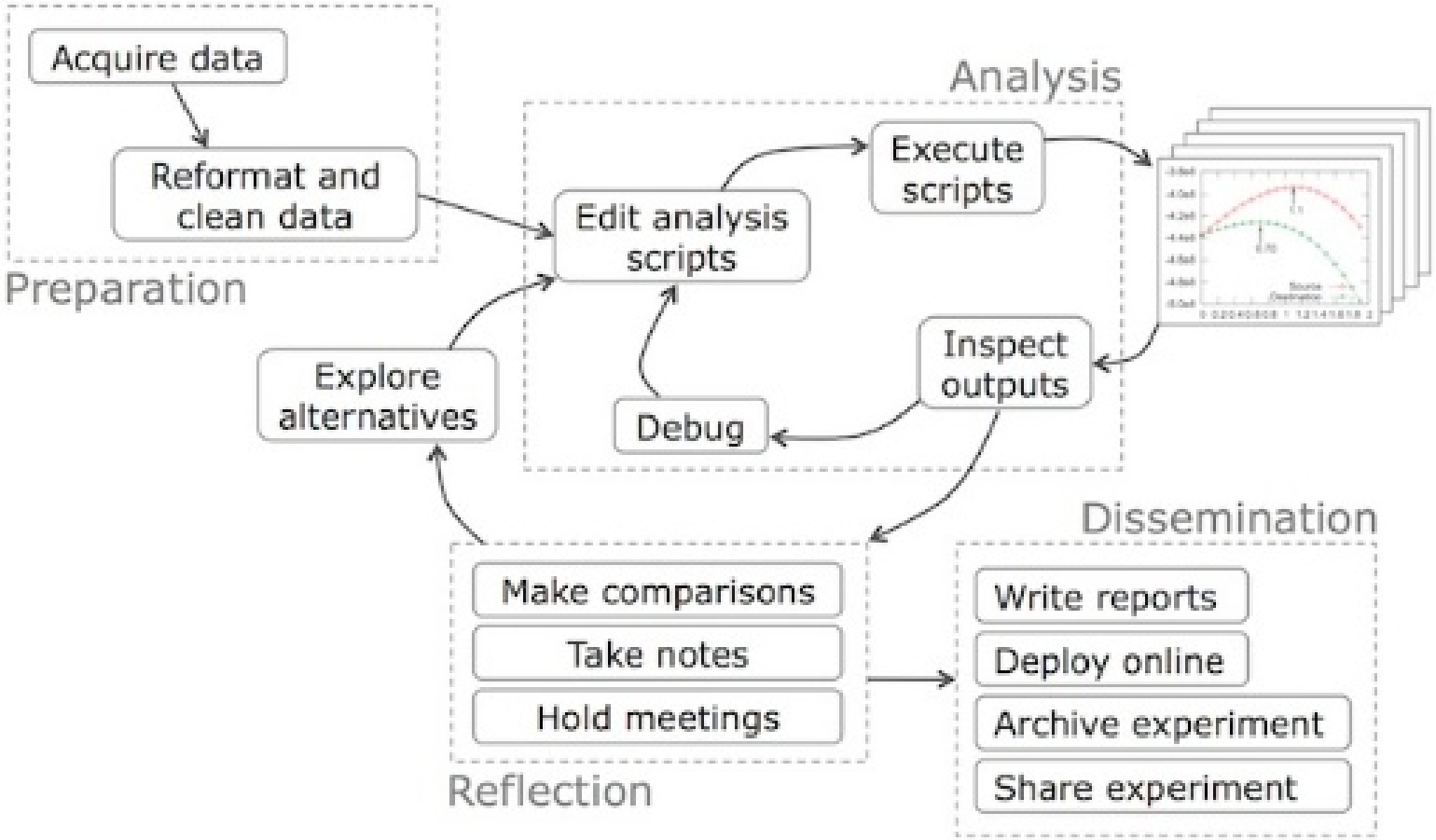
2016

Goals

Walk away from this talk knowing:

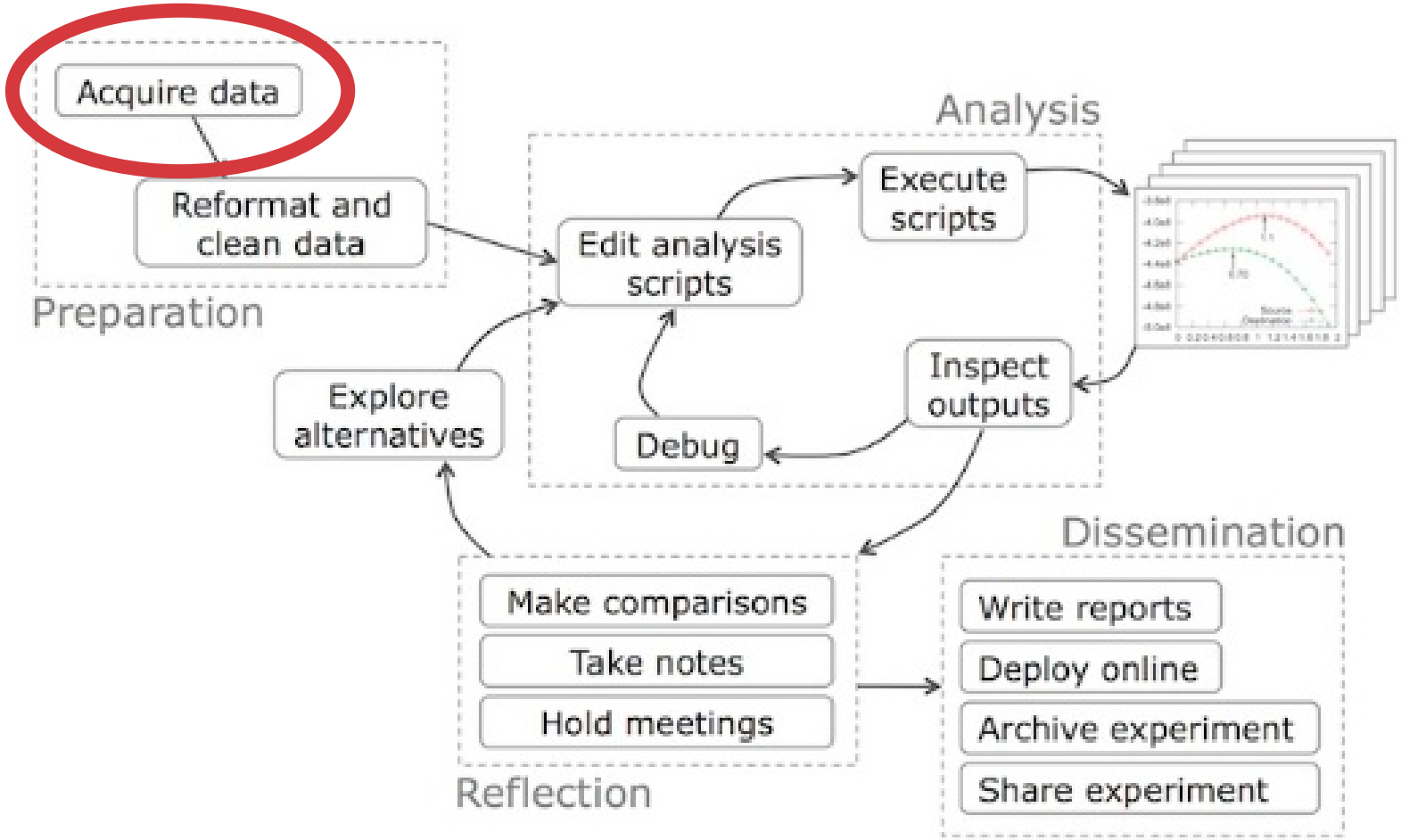
1. What problems scrapers address
2. How they work
3. How to build one (general steps + Py-example)

Data science pipeline



Source: http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

Data science pipeline



Source: http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

What is a scraper?

Tool to download and extract digital content

aka:

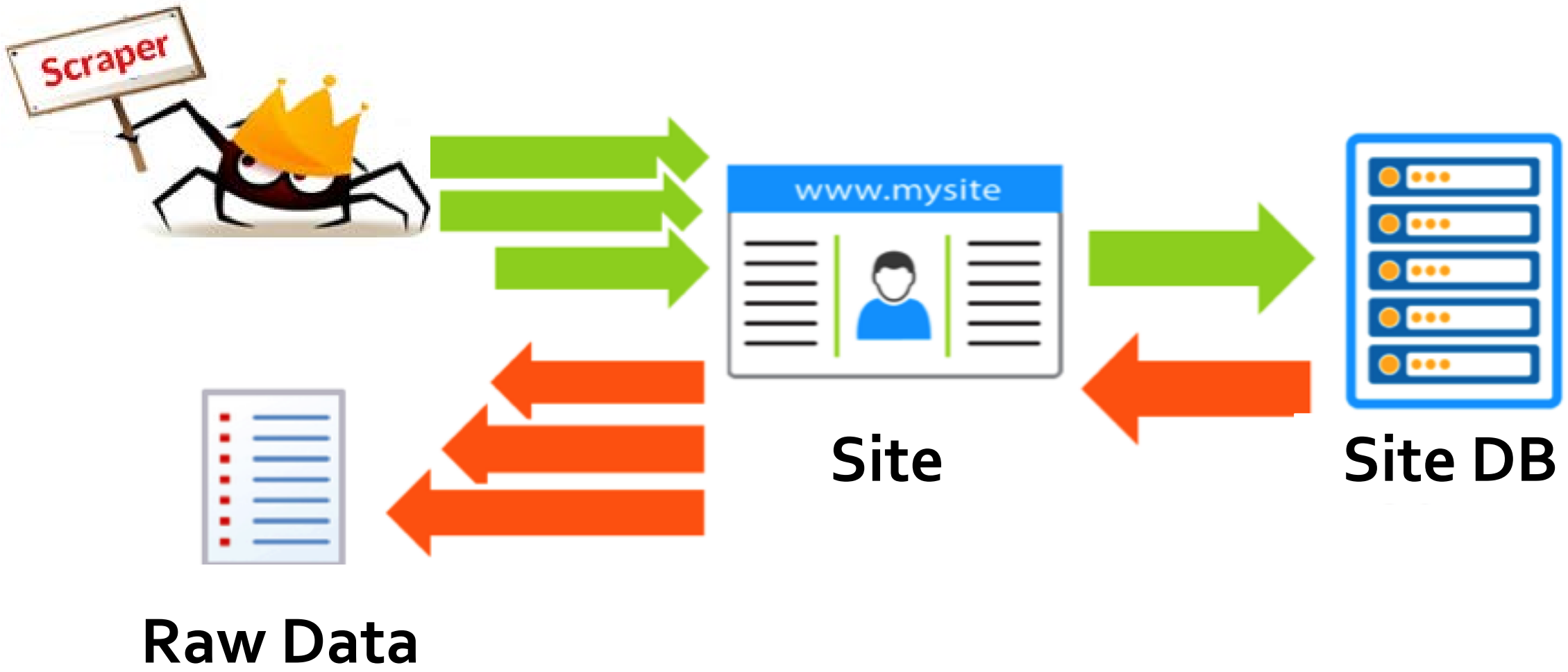
- Machinery behind “Get Data” button
- Provides a custom API

General concept with custom applications

Typical Scenario

Location	Content	Format
City of Toronto Website	News Releases Since 20xx	Table

- Goal: Get insight into news releases
- Output: Create word cloud of news releases
- Problem: How to get the data?


Scraper
www.mysite
Site
Site DB
Raw Data

Who Uses Scrapers?

- Data aggregators
  - search engines (ex Google), job boards (ex Indeed), event aggregators, real estate related (ex WalkScore)
- Businesses
  - price monitoring, reputation monitoring, market research
- Financial firms
  - signal research, "alternative data"
- Academics
- And many more…

- http://blog.datahut.co/how-real-businesses-use-web-scraping/
- https://www.quora.com/What-are-examples-of-how-real-businesses-use-web-scraping

Samsung Galaxy Tab S2 WiFi 8" 32GB

Manufacturer: Samsung

Price Range: **$359.92** to **$718.85** at 11 stores

4.3/5 Read our Review

Recommend 0

Buy Samsung Galaxy Tab S2 WiFi 8" 32GB

softwarecity	$437.69	SEE OFFER ➜
PC Canada	$439.99	SEE OFFER ➜
Mike's Computer Shop	$473.99	SEE OFFER ➜

Samsung Galaxy Tab S2 WiFi 8" 32GB Price

https://www.connectifier.com/product/social-links

Connectifier

SEARCH AUTO SEARCH **SOCIAL LINKS** FAQ

Social Links

Connectifier's free Social Links browser extension provides instant access to a candidate's social media footprint, enabling you to gain deeper insights and personalize your messages.

Log In

Benefits

- Aggregates data from over 1 million public websites
- Gives you more insights into a candidate
- Helps you get more responses through message personalization
- Works with the recruiting tools you already

Connectifier

Ben McCann

Mountain View, CA

Senior Staff Software Engineer at LinkedIn and Co-founder at Connectifier

cf Search

New Applicant

Engineer @ My Company

Find Jobs | Find Resumes | Employers / Post Job | Upload your resume | Sign in

indeed

what: data scientist (job title, keywords or company)

where: Toronto, ON (city or province)

Find Jobs | Advanced Job S

data scientist jobs in Toronto, ON

Sort by: **relevance** - date

Distance:
within 50 kilometers

Salary Estimate
- $60,000+ (228)
- $80,000+ (148)
- $100,000+ (44)
- $120,000+ (11)
- $180,000+ (1)

Job Type
- Full-time (105)
- Permanent (36)
- Contract (16)
- Temporary (2)
- Part-time (1)

Location
- Mississauga, ON
- Data Scientist jobs nationwide

Company
- AMZN CAN Fulfillment Svcs
- Maxxam Analytics
- Michael Page International
- Unata

Upload your resume - Let employers find you

Jobs 1 to 20 of 255

Radiology Data Scientist - 210027
Philips - ★★★★½ 1,201 reviews - Toronto, ON
Radiology Data Scientist. The scientist should be able to work in a multidisciplinary team of software engineers, clinical scientists, physicians, department...
30+ days ago
Sponsored

Senior Business Management Analyst (Data Scientist / Special...
TD Bank - ★★★★☆ 2,062 reviews - Toronto, ON
Experience on Data Modeling / industry Analytical Tools. Experience with data discovery & Analytics identifying insights....
1 day ago
Sponsored

Data Scientist / Data Architect
Lannick Group - ★★★★☆ 5 reviews - Oakville, ON
$110,000 a year
It will be performing data investigations to determine root cause of data quality issues as well answer questions about data and/or data trends....
8 days ago
Sponsored

Data Scientist
RBC - ★★★★½ 461 reviews - Toronto, ON
RBC Machine Intelligence Research is looking for data scientists to join its Machine Learning team. Passion for data, algorithms and statistics!...
2 days ago - save job - more...

Get new jobs for this search b email

My email:

Send me new jobs

You agree to get information about new j for this search by email.You can cancel e alerts at any time.

Company with data scientist jobs

PHILIPS

Philips
Philips is a diversified health and well-being company, focused on improving people's lives through timely innovations.

Systems Administrator - 214325

Bilingual Customer Response Associate (Overnight) - 210114-3

Customer Response Associate - 210114-2

Jobs (14)

dan barker
@danbarker

Follow

.@mattcutts I think I have spotted one, Matt. Note the similarities in the content text: pic.twitter.com/uHux3rK57f

Reply Retweet Favorite Pocket Buffer ••• More

Google

what is a scraper site

Web Shopping Images Videos News More Search tools

About 13,400,000 results (0.56 seconds)

scraper site

Web definitions

A scraper site is a spam website that copies all of its content from other websites using web scraping. The purpose of creating such a site can be to collect advertising revenue or to manipulate search engine rankings by linking to other sites to improve their search engine ranking. ...
http://en.wikipedia.org/wiki/Scraper_site

Scraper site - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Scraper_site** Wikipedia
A **scraper site** is a spam website that copies all of its content from other websites using web scraping. The purpose of creating such a site can be to collect ...
Made for advertising - Legality - Techniques - See also

RETWEETS 33,530 FAVORITES 3,891

11:51 AM - 27 Feb 2014

Flag media

Problem scrapers address

Scenario:

Location	Content	Format
Web	Text Images Video	Table Non Table Spreadsheet RSS Feed

Problem: How to get the data?

Problem scrapers address

Scenario:

Location	Content	Format
Web Disk Email	Text Images Video	Table Non Table Spreadsheet RSS Feed

Problem: How to get the data?

Problem addressed today

Scenario:

Location	Content	Format
Web	**Text**	**Table**

Problem: How to get the data?

Typical Scenario

Location	Content	Format
City of Toronto Website	News Releases Since 20xx	Table

- Goal: Get insight into news releases
- Output: Create word cloud of news releases
- Problem: How to get the data?

How do Web Scrapers* work?

1. Download webpage from a server
2. Process webpage to output data

How do Browsers* work?

1. Download webpage from a server
2. Process webpage to display data

How do Web Scrapers* work?

1. Download webpage from a server
2. Process webpage to output data

* web scraper = customized browser

How do Web Scrapers* work?

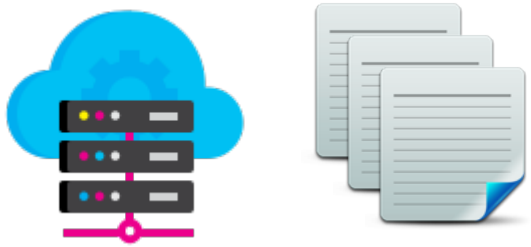
1. Download webpage from a server – HTTP
2. Process webpage to output data – HTML + CSS + JS

* web scraper = customized browser

Downloading: How the web works

- Server
  1. Listen for an incoming request
  2. Send out a response
- Browser
  1. Send a request to a server
  2. Receive and process the server response
- HTTP = Language of Request + Response

Browser



Server

Downloading: How the web works

- Server
  1. Listen for an incoming request
  2. Send out a response
- Browser
  1. Send a request to a server
  2. Receive and process the server response
- HTTP = Language of Request + Response

Downloading: How the web works

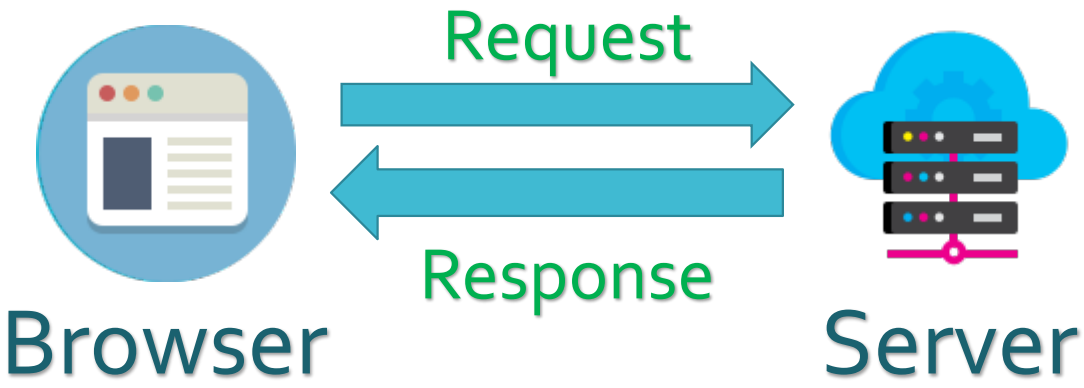
- Server
  1. Listen for an incoming request
  2. Send out a response
- Browser
  1. Send a request to a server
  2. Receive and process the server response
- HTTP = Language of Request + Response

Downloading: How the web works

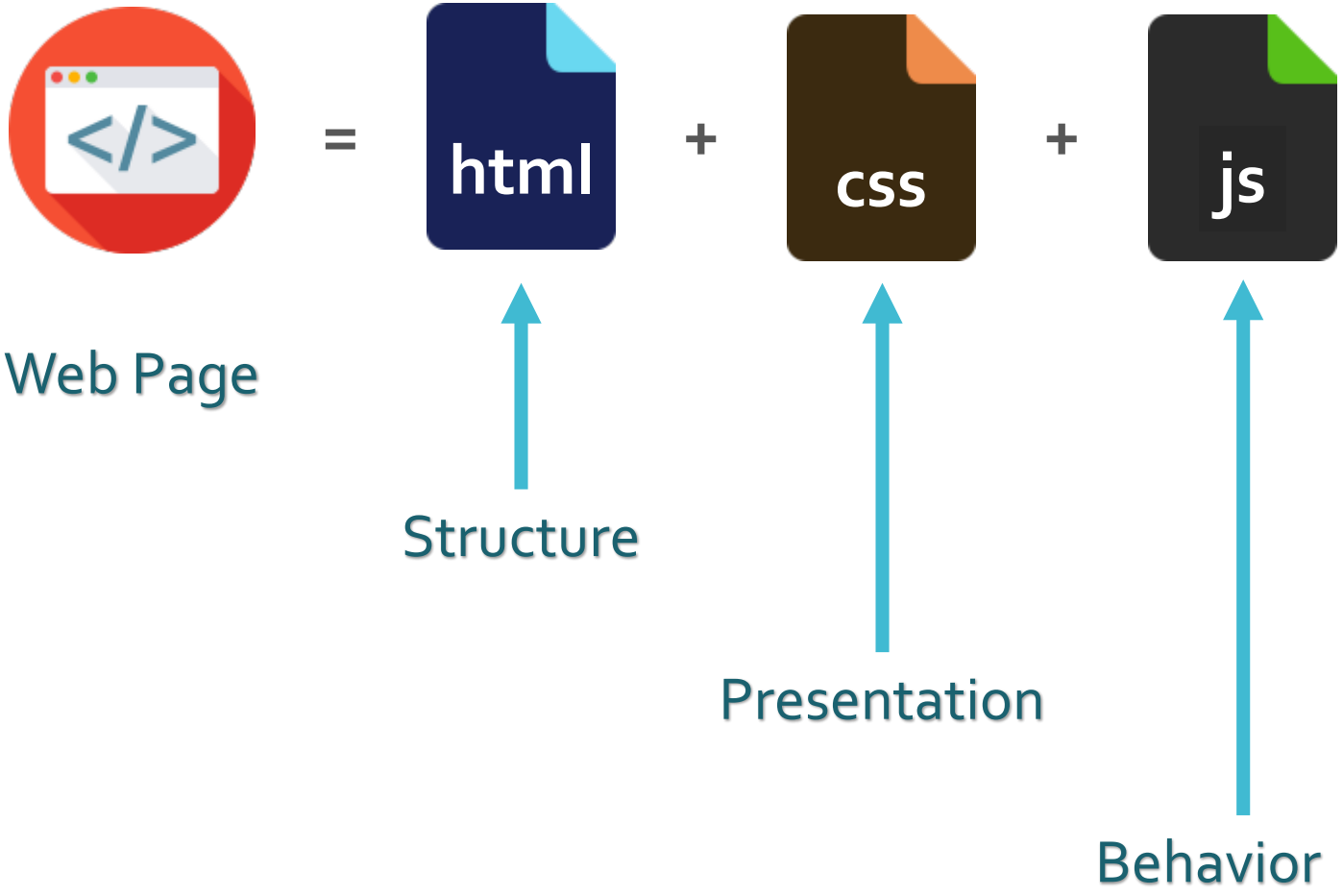
- Server
  1. Listen for an incoming request
  2. Send out a response
- Browser
  1. Send a request to a server
  2. Receive and process the server response
- HTTP = Language of Request + Response

Downloading: How the web works

- Server
  1. Listen for an incoming request
  2. Send out a response
- Browser
  1. Send a request to a server
  2. Receive and process the server response
- HTTP = Language of Request + Response



Processing: How web content gets displayed



How do Web Scrapers* work?

1. Download webpage from a server – HTTP
2. Process webpage to output data – HTML + CSS + JS

How do Web Scrapers* work?

Components

1. Download webpage from a server – HTTP
2. Process webpage to output data – HTML + CSS + JS

=

Downloader

+

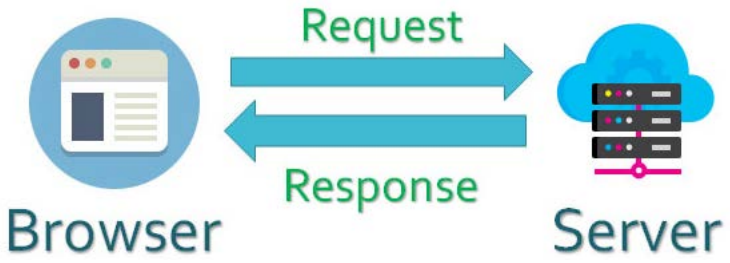
Processor

How do Web Scrapers* work?

Components

1. Targeted Downloader (HTTP)

acts like a



2. Targeted Processor (HTML+ CSS + JS)

acts to process

How do I build a scraper?

1. Identify interesting question
2. Identify target website with data to answer question
3. Investigate website structure
4. Write scraper (downloader + processor)
5. Test scraper
6. Deploy scraper – get data
7. Optional: repeat / refine

How do I write a scraper?

Downloader

1. Identify URL to webpage with data
2. Request webpage with URL
3. Receive response

Processor

1. Read response
2. Extract relevant data from response
3. Output data (to screen, file, db etc.)

Demo

Demo Examples

1. Mississauga:
   - Pure HTML
   - Table format
   - Single page
2. Burlington:
   - Pure HTML
   - Div format (not much different from tables)
   - Multiple pages
3. Toronto:
   - Shell HTML + JavaScript data injection
   - JSON format (after some text wrangling)
   - Multiple pages

Problem scrapers address

Scenario:

Location	Content	Format
Web Disk Email	Text Images Video	Table Non Table Spreadsheet RSS Feed

Problem: How to get the data?

Scraper Setup Reconnaissance Toolkit

1. Chrome DevTools (Examine Content / HTTP)
2. Postman (HTTP Requests)
3. Beyond Compare (Text Diff Tool)

Scraper Implementation Python Libraries

Downloading:

requests for HTTP calls

Processing:

BeautifulSoup
json
other processing tools – Image / PDF / Excel

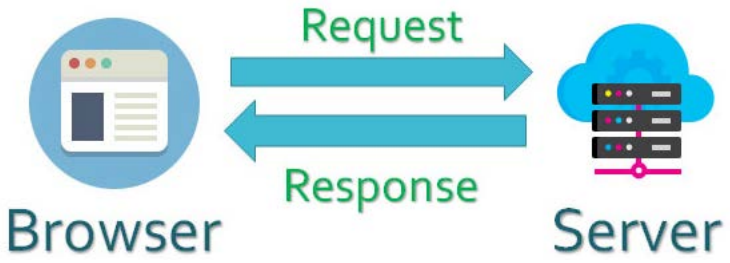
Scraper Considerations

1. Timing of requests
2. Structuring your downloading / processing code
3. What content to extract
4. But different cases add levels of complexity on top of this…

How do Web Scrapers* work?

1. Targeted Downloader (HTTP)

acts like a



2. Targeted Processor (HTML+ CSS + JS)

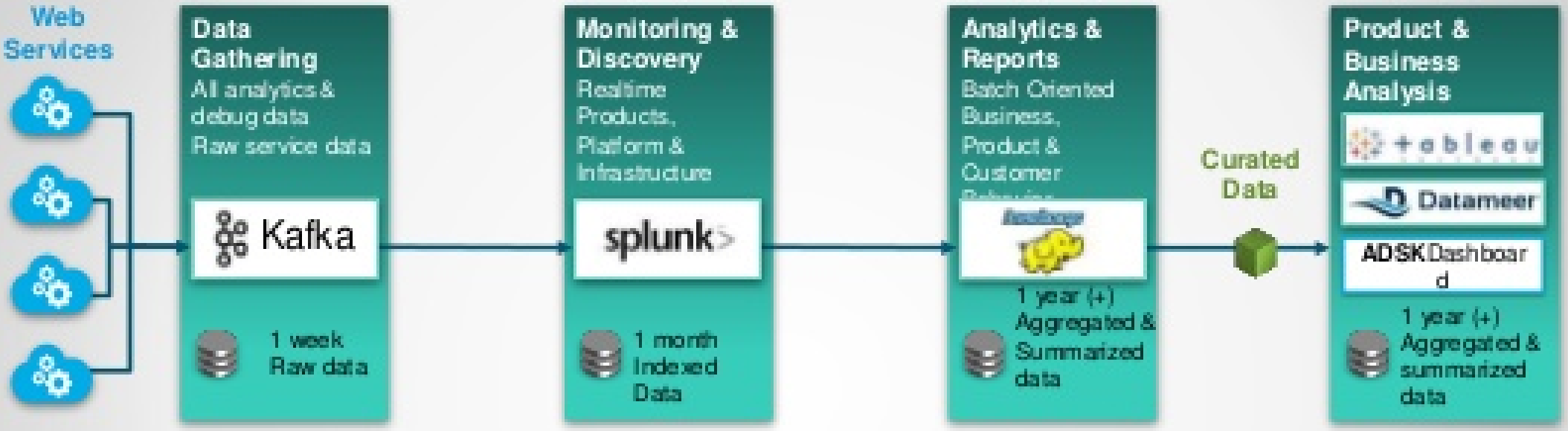
acts to process

Goals

Walk away from this talk knowing:

1. What problems scrapers address
2. How they work
3. How to build one (general steps + Py-example)

Production Big Data Pipeline



AUTODESK.

Q&A