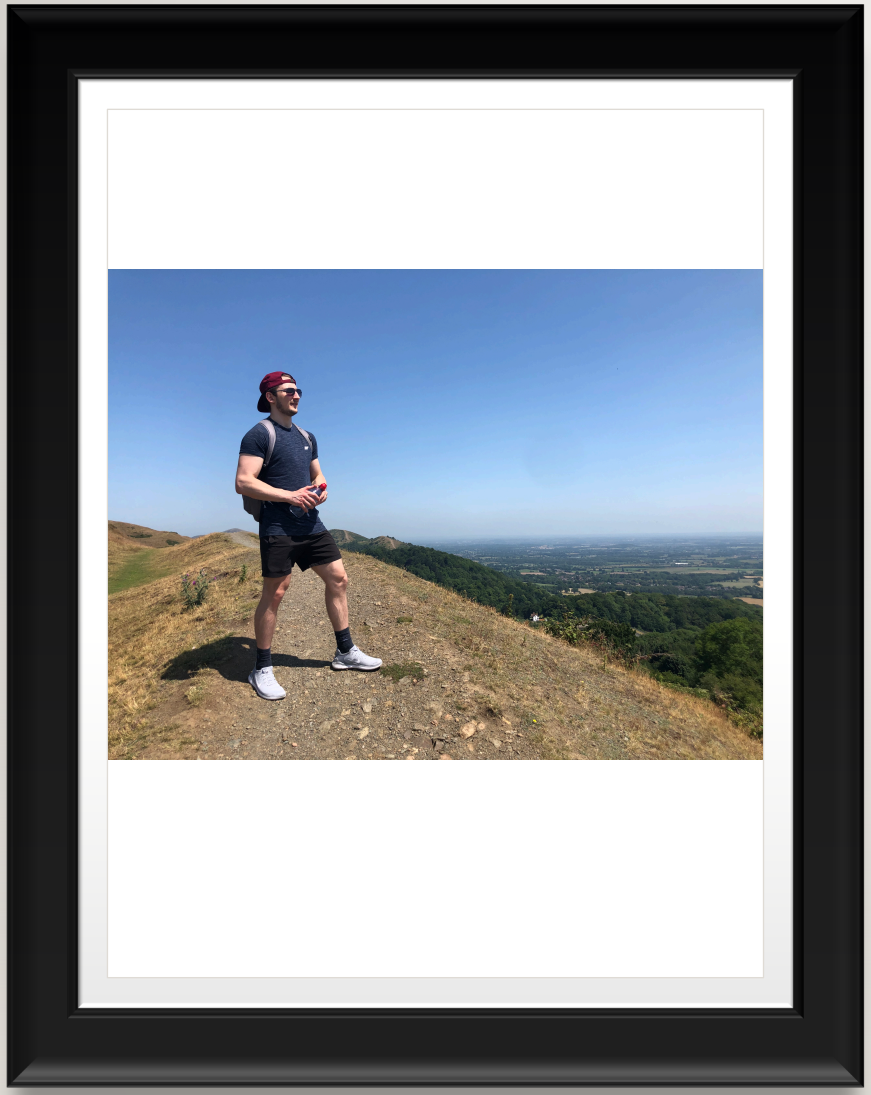


WEB SCRAPING / WEB CRAWLING IN GOLANG

LIAM HAMPTON

Member of the IBM UKI Developer Advocate team, focusing on Golang and Open Source technologies

Tech Jam Podcast Co-Host https://techjam.dev/

Previously a Node.js and Go developer on the Eclipse Codewind Open Source initiative.

@LiamConroyH thetechjam

WHAT IS WEB SCRAPING?

- An automated technique employed to gather vast amounts of data from the web
- Data extraction into a locally stored form – .csv, JSON, Database

WHY IS THIS TECHNIQUE USED?

- SE's use it to index the content of the web
- Price monitoring
- Market research
- Real time analytics
- Machine learning training models
- For fun

METHODS

- Software
  - WebHarvy
  - Visual Web Ripper
  - OutWit Hub
- Programmatically
  - PhantomJS
  - Selenium
  - Gocolly (what you are using today)


@LiamConroyH

thetechjam

DRAWBACKS

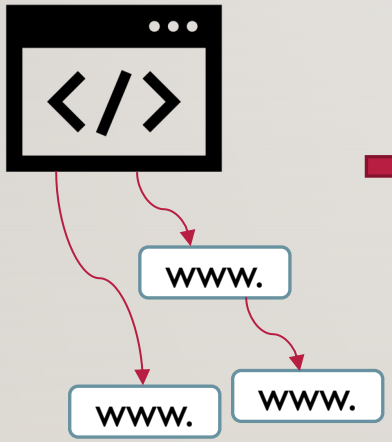
- Web pages are built to be readable by humans and screen readers
- Not built to be easily scraped and have data extracted
- Its hard for one size to fit all – web scrapping software is complex because every web page is built differently
- Not efficient
- Security? – Afterall, this is basically a copy and paste of source code!
  - No validation of the data being collected

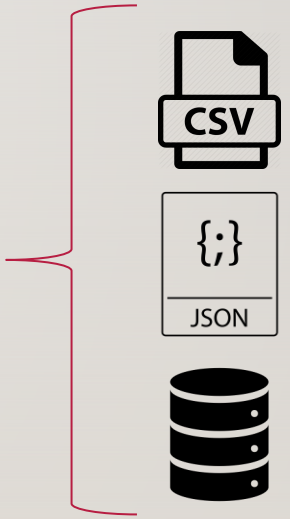
WHAT ARE OTHER ALTERNATIVES?

- Web crawlers / spider / web bots
  - Browser automation
- Dedicated API's
  - Publicly exposed API's to retrieve certain data in a specified format (usually JSON or XML)

WHAT IS WEB CRAWLING?

- An automated technique employed to search documents or embedded links on the web, catagorise the content and then index it
- Often used for repetitive actions





Structured Data

@LiamConroyH thetechjam

WHY MIGHT AN API NOT WORK?

- Not all data is available via an API – you don't decide what data is provided
- Not every website with information exposes a public facing API
- Some websites might make you request access and its at their discretion to give you access or not
- API's sometimes have request limits
- They are not always free

WHERE DOES GOLANG COME INTO THIS?

WHAT IS GOLANG?

- Open source
- Created by Google Engineers and launched in 2009
- Statically typed language for the multi core processor
- Usually used for data intensive tasks – concurrent execution is a staple of the language
  - System migrations
  - Microservices
  - Dubbed to be the networking language of the future

@LiamConroyH thetechjam

HOW DOES GOLANG MAKE WEB SCRAPING EASY?

- Simple to use and write
- Powerful 3rd party libraries help make this easier than writing HTTP requests and analysing streams of data. Two of the most popular libraries used in the community are:
  - gocolly
  - goquery (similar to jQuery)

WHAT ARE YOU DOING IN THIS WORKSHOP?


@LiamConroyH

thetechjam

LAB ELEMENTS

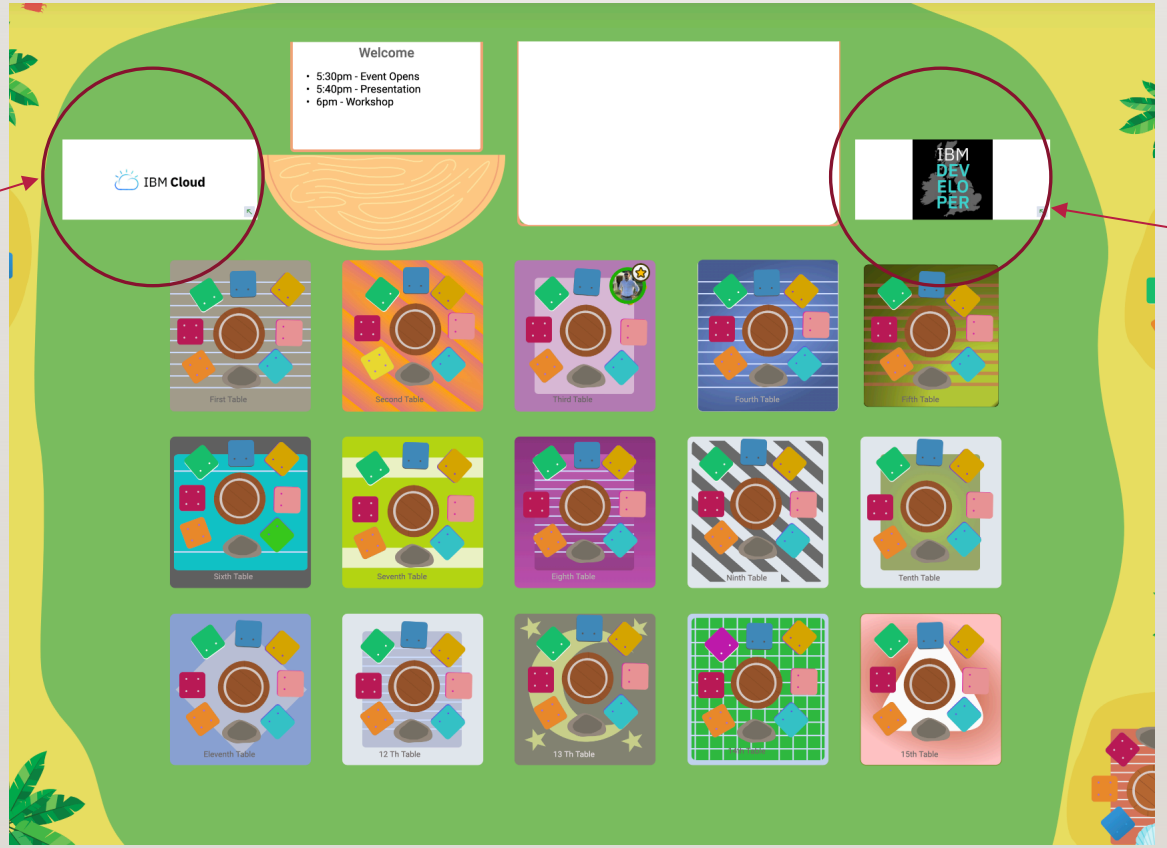
1. A basic Go application
2. Turn the basic app into a HTTP Server
3. Push the app up to Cloud Foundry on IBM Cloud
4. Write a web scraping function
5. Write a web crawling function


@LiamConroyH

thetechjam

WORKSHOP MATERIALS

IBM Cloud sign up



GitHub repository

@LiamConroyH

thetechjam

DISCLAIMER

- Always check the small print
  - Some website prohibit scraping
  - Some websites out right block it
  - Some have certain T&C's regarding it and the way you process their data
- Be kind to their servers – use a sensible rate limit on your requests
- Can get you in trouble if you use the data collected maliciously - repercussions can be costly!
  - Potentially cause a denial-of-service on their servers if you are not careful
  - Blacklist your IP address
  - Block your account if you have one with the site


@LiamConroyH

thetechjam

IBM CODE MEETUP - CODE OF CONDUCT

A safe, respectful, comfortable and harassment-free environment for attendees

ibm.biz/CodeofConduct

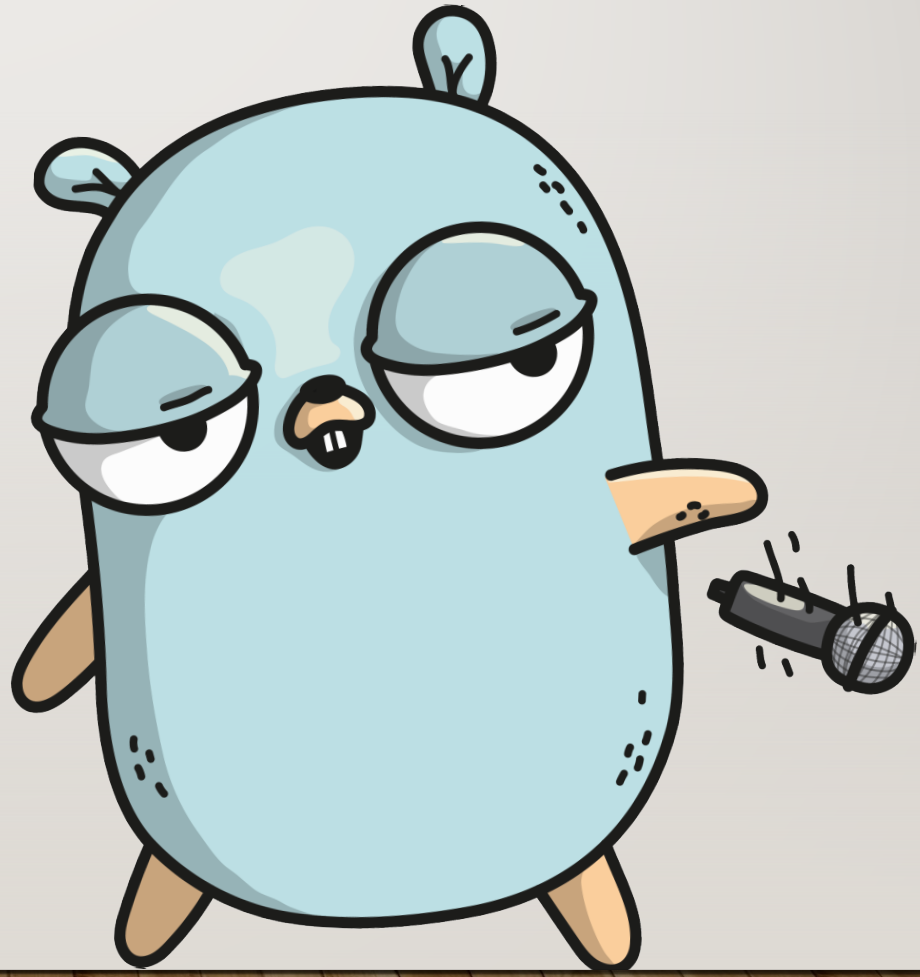
- **Our Code**
- ➢ Be nice
- ➢ Be Respectful
- ➢ Participate but don't disrupt
- ➢ Don't break the law

- **Report Your Concerns**
- Your instructor
- Any IBM'er
- DM via Meetup.com
- DM via @codemeetup
- Email stewaa3@uk.ibm.com
- Phone or text 07802765745

- **We will act on incidents**
- During events – listen & act
- 2 weeks – investigate
- 30 days – further action

- **What we can do**
- ➢ Ask you to leave
- ➢ Written warning
- ➢ 3 months breather
- ➢ Remove from the group

WORKSHOP
LETS CODE



@LiamConroyH thetechjam