How Can I Help?:
Zero-Shot Multi-Modal Automation with QA

Michael Du, Sam Masling
Nancy Xu
The **Average American** Spends **6hrs/day** on the Internet

- **Imagine an agent automated some of those tasks.** And we spent less time!

- **Virtual Personal Assistants (VPA)** ex. Alexa, Google Assistant, Siri, Cortana, and Bixby unable to cover **long tail of user requests**.

- **Programming by Demonstration** systems allow us to demonstrate new skills to agents.
  - 1. Prompting the user to provide a natural language utterance to refer to the skill
  - 2. Asking users to demonstrate the skill in the browser
  - 3. Capturing and name relevant variables and the sequences of clicks.
  - 4. Saving the demonstration to be called by name in the future.
Programming **Dialogue Agents** on the Web is **Hard**

1. Require end-user to **demonstrate full space of possible browser actions** => time-consuming + incomplete.
2. **CSS selectors** are **brittle**.
3. Skills are **not generalizable** to new domains or sites.
4. Training **dialogue systems** is **non-trivial**.
What if you could generate an agent from any website?

Like a human reading a website -- no extensive demonstration needed.
Web Elements Perform 3 Main Purposes: **Inform / Request / Act**

[Image of a web interface with UI hierarchy and semantified UI]
HTML induced questions (with language models?) + UI

Grammar Templates
Zero-Shot **Slot Filling** + Navigation as **Question-Answering**

Please help me book a flight from **SF** to **JFK** departing on **Oct 30, 2020**.
Demo: **SiteBot**, a multi-model conversational interface.

*Book a flight* by navigating through Google -> **OneBox** via Chrome extension chatbot. Powered by QA NLU + Induced Questions.
Project Timeline:

- **Week 4**: Build a simple puppeteer agent that comprehends user utterance -> executes multi-modal automation for Google.
- **Week 5-6**: Study web structure + classify element types. Create question templates w/ ARIA etc. Also experiment with learning questions automatically from HTML with GPT 3 / language models. BoolQA models for actions (or CoQA) + ExQA on content.
- **Week 7**: Finetune Q&A models on synthetic training data generated by UI grammars + paraphrasing. Collect test data (user utterance + slots) on 10 websites using Mechanical Turk.
- **Week 8**: Build chrome extension interface within puppeteer browser for chatting with the agent.
- **Week 9**: Validate results on test data. Compare zero-shot QA technique against known benchmarks for slot-filling etc.
- **Week 10 + Reach**:
  - Identify necessary slots for actions the seed multi-turn dialogue.