
Natural language processing products in the wild

Will Radford - Data Scientist at Red Marker
11th April 2017

—

Plan:
NLP in the wild
Product thinking

Data Scientist: Day one

- You're working for BigCorp
 - They have a system where users contribute messages
 - You need to **classify** these messages
 - For browsing by category
 - For matching to *other* users who want to read messages
 - For indexing in search
 - ...
-

NLP in the wild

Automation philosophy

- Complete automation
 - Moral questions
- Decision support for experts
 - Capture the data for virtuous circles
- Crowd-workers
 - Watch out for latency and quality!

It just depends

Trading-off precision and recall

- High-precision: don't show users wrong results
- High-recall: don't miss anything you shouldn't
- No substitute for real testing
 - A/B Tests
 - [Multiworld testing](#)
 - ...

Models with probabilities are valuable, thresholds are great

Peeking inside the black box

- Iterate quickly
- Decision auditing
 - How does this affect user's trust in the system?
- Can you guarantee to your manager that a statistical model behaves a certain way?

It depends

Moving targets

- Training data isn't right by chance
 - Learning curves - do you need more data?
- If your datasets are temporal, don't time-travel
- New categories emerge
- Pipeline change propagate

Hard area, keep an eye on it

Predicting in production

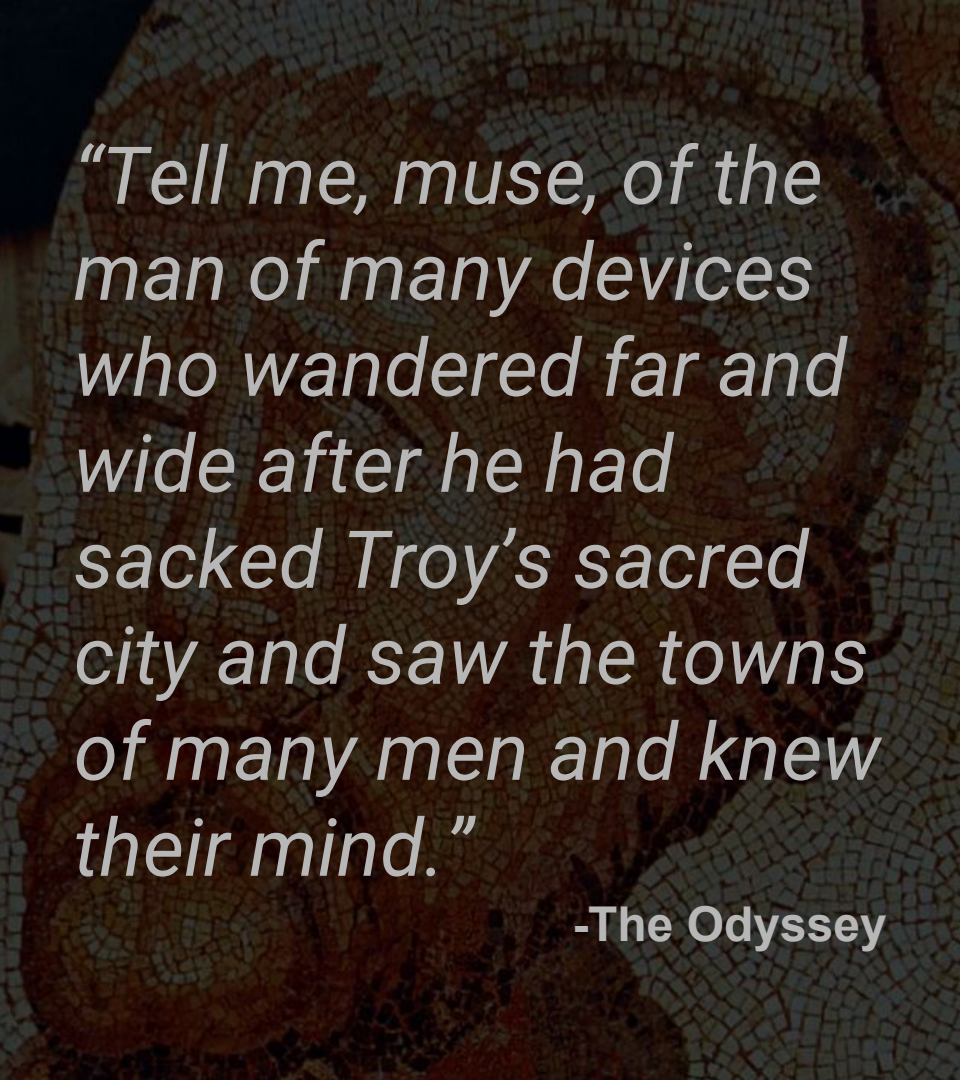
- Ease of deployment is important
- Model-size can be a factor
- Exploit data-parallelism
- Re-use large-scale learning frameworks
- Think about security
- [Rules of Machine Learning \(Zinkevich, 201?\)](#)

Engineering matters

Product thinking

Taken from “Knowledge Base Jumping” by Ben Hachey

<https://twitter.com/benhachey/status/851222298103209984>



*“Tell me, muse, of the
man of many devices
who wandered far and
wide after he had
sacked Troy’s sacred
city and saw the towns
of many men and knew
their mind.”*

-The Odyssey

Entity linking *with* Wikipedia

“Novak Djokovic won an Open Era record fifth men's singles crown by defeating Andy Murray in the finals of the Australian Open. Reigning women's champion Li Na did not defend her title, as she retired from professional tennis in September, 2014.”

Adapted from the 2015 Australian Open Wikipedia article

“**Novak Djokovic** won an Open Era record fifth men's singles crown by defeating **Andy Murray** in the finals of the **Australian Open**. Reigning women's champion **Li Na** did not defend her title, as she retired from professional tennis in September, 2014.”

Adapted from the 2015 Australian Open Wikipedia article

“Andy Murray”

Andy Murray (born 1987), Scottish tennis player

Andy Murray (ice hockey) (born 1951), Canadian ice hockey coach and former player

Andrew Murray (boxer) (born 1982), Irish professional boxer

Andrew Murray (Guyanese boxer) (1971-2002), Guyanese boxer of the 1990s and 2000s

Andrew Murray (golfer) (born 1956), English golfer

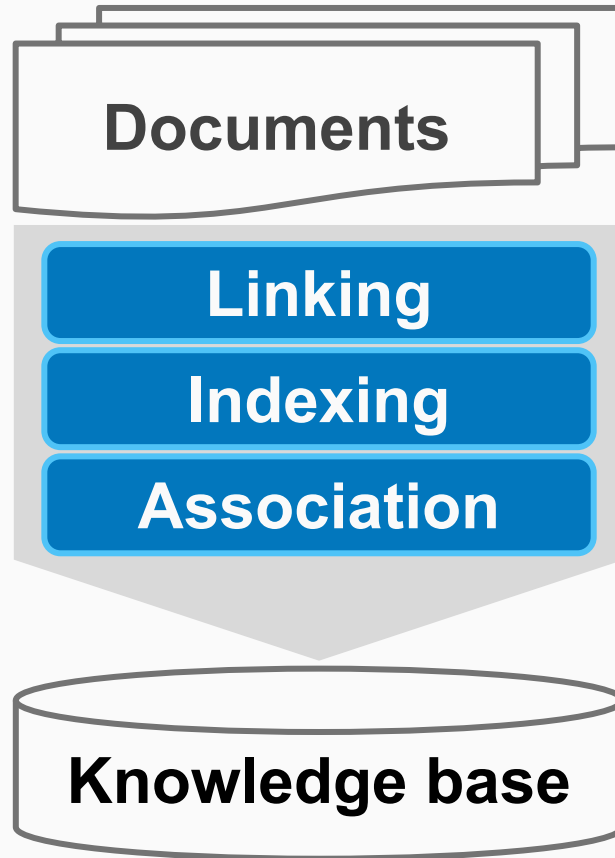
Andrew Murray (ice hockey) (born 1981), Canadian ice hockey player

+ 15 others

“[Novak Djokovic](#) won an Open Era record fifth men's singles crown by defeating [Andy Murray](#) in the finals of the [Australian Open](#). Reigning women's champion [Li Na](#) did not defend her title, as she retired from professional tennis in September, 2014.”

Adapted from the 2015 Australian Open Wikipedia article

Linking applied

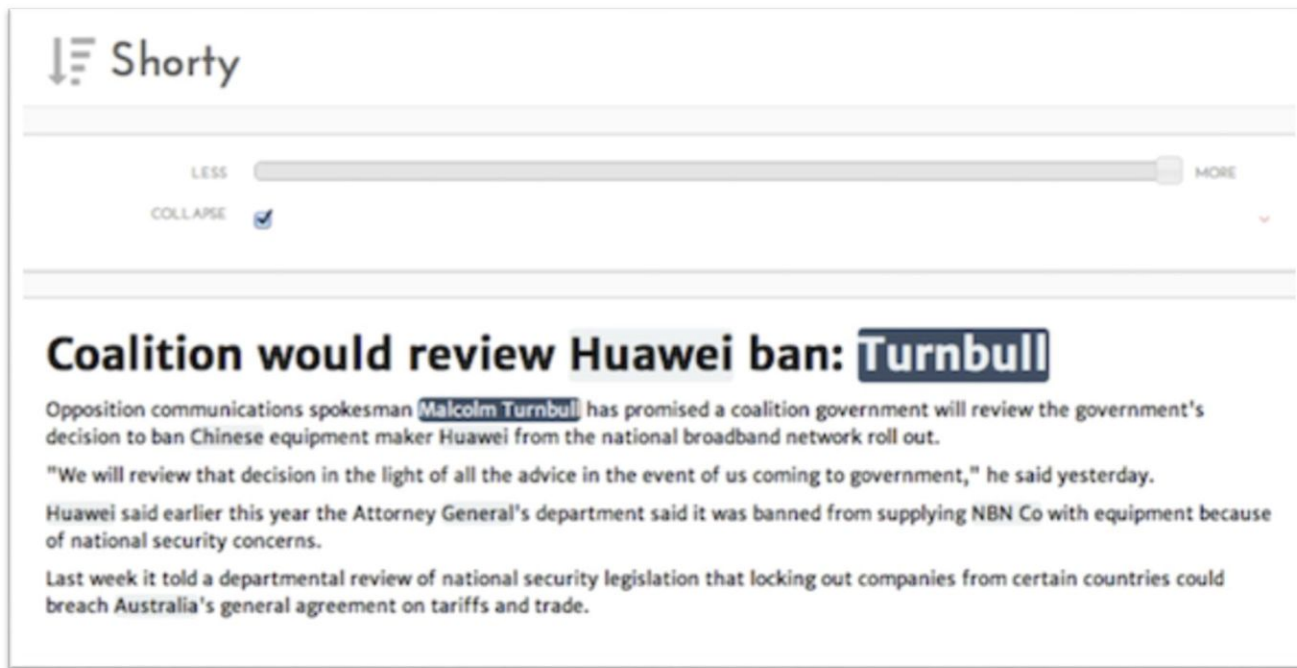


Editor: helping journalists write richer stories

The screenshot shows the 'Compnews Editor' interface. On the left, a text editor with line numbers 1 to 5 contains the following text: [Tony Abbott] (e:Tony_Abbott) announces new budget with [Joe Hockey] (e:Joe_Hockey) . Mr [Hockey] (e:Joe_Hockey) 's wife, Melissa Babbage, also attended the speech. The text is displayed in a monospaced font. In the center, a preview window shows the text with entity links: Tony Abbott announces new budget with Joe Hockey. Mr Hockey's wife, Melissa Babbage, also attended the speech. On the right, a dark grey panel displays entity information. It features two circular profile pictures. The top one is for Tony Abbott, with the text 'Tony Abbott' and 'Anthony John "Tony" Abbott is the Leader of the Opposition in the'. The bottom one is for Joe Hockey, with the text 'Joe Hockey' and 'Joseph Benedict "Joe" Hockey, is an Australian politician and'.

- Automatic text augmentation
 - Entity links [mention]{e:entity_name}
 - Knowledge
 - {partner} → Melissa Babbage
 - {age} → 48 years old
 - {map} → coodinates + maps.google.com
- **Idea: recreate a NER/NEL corpus every week**

Shorty: summarising stories



The screenshot shows the 'Shorty' web interface. At the top left is the 'Shorty' logo with a downward arrow icon. Below it is a horizontal slider bar with 'LESS' on the left and 'MORE' on the right. Under the slider, the word 'COLLAPSE' is followed by a blue checkmark icon. The main content area displays a news headline: 'Coalition would review Huawei ban: **Turnbull**'. The word 'Turnbull' is highlighted in a dark blue box. Below the headline are three paragraphs of text. The first paragraph mentions 'Opposition communications spokesman **Malcolm Turnbull**' and his promise to review the government's decision to ban Huawei. The second paragraph is a quote from him. The third paragraph mentions Huawei's ban from supplying NBN Co and a departmental review of national security legislation.

Shorty

LESS MORE

COLLAPSE ✓

Coalition would review Huawei ban: **Turnbull**

Opposition communications spokesman **Malcolm Turnbull** has promised a coalition government will review the government's decision to ban Chinese equipment maker Huawei from the national broadband network roll out.

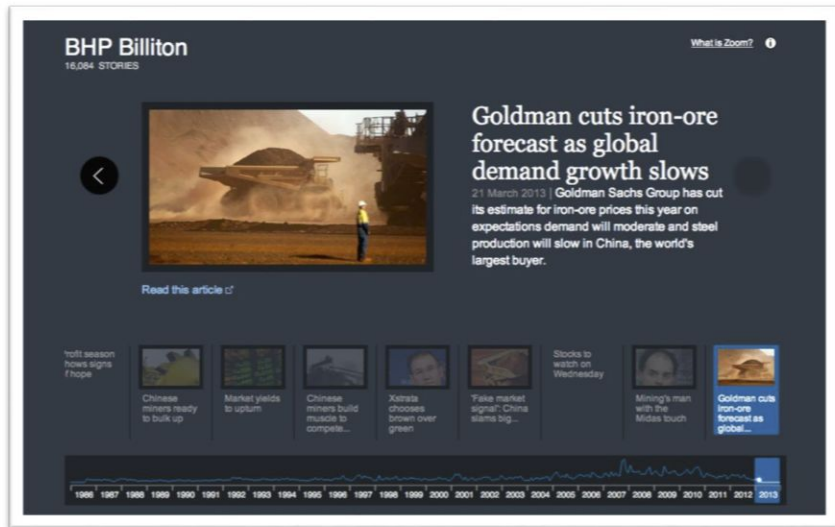
"We will review that decision in the light of all the advice in the event of us coming to government," he said yesterday.

Huawei said earlier this year the Attorney General's department said it was banned from supplying NBN Co with equipment because of national security concerns.

Last week it told a departmental review of national security legislation that locking out companies from certain countries could breach Australia's general agreement on tariffs and trade.

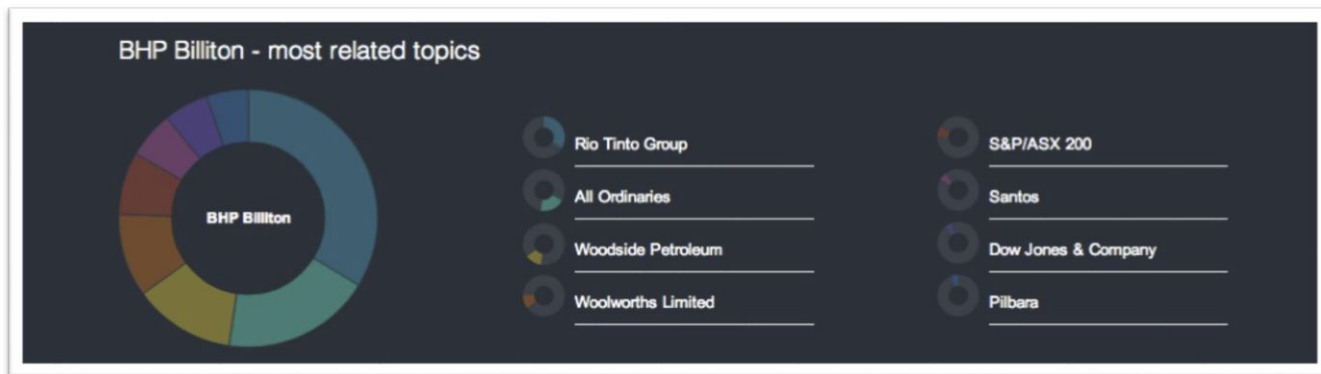
- Baseline extractive summarization
- Slide to reveal more
- **Idea: use coreference chains to repair sentences**

Zoom: accessing the archive through entities



- Landing page for each entity
- Images and snippets
- Mention frequency timeline
- Which stories do we show?
- **Idea: drive traffic deeper into the archive**

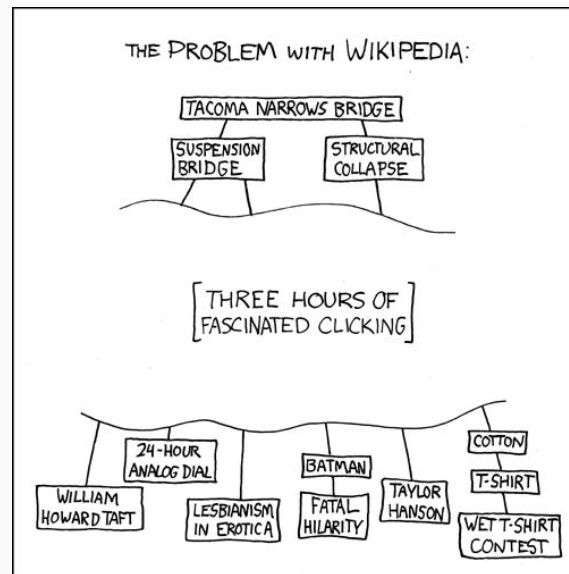
Zoom: accessing the archive through entities

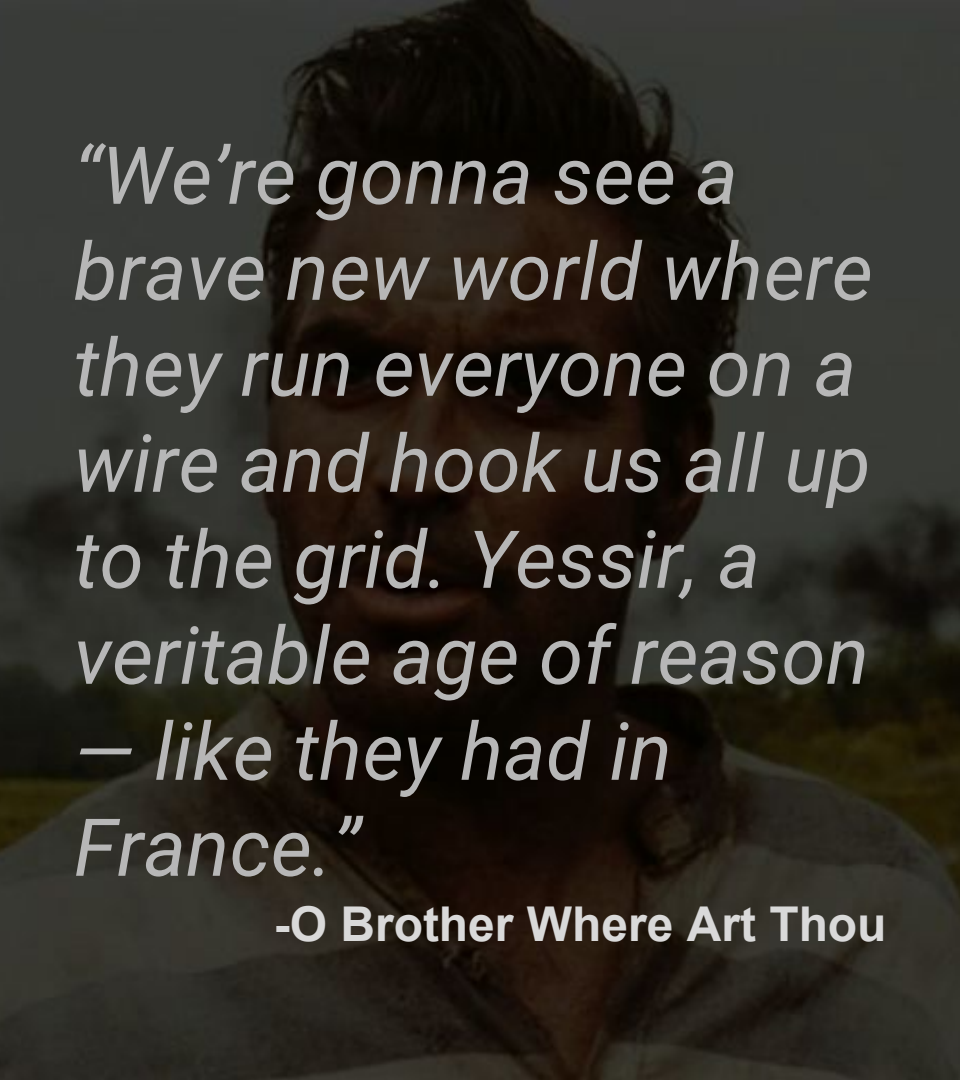


- Balance commonness with surprise
- **Idea: simulate the “lost in Wikipedia” experience**

Lessons and challenges

- Lessons:
 - Browsing provides deep context, facilitates discovery
 - Not great for breaking news
 - Not great for entities outside KB/archive
- Challenges:
 - **Notability:** domain entities, local entities
 - **Changeability:** emerging entities
 - **Scope:** Entity scope context specific





*“We’re gonna see a
brave new world where
they run everyone on a
wire and hook us all up
to the grid. Yessir, a
veritable age of reason
— like they had in
France.”*

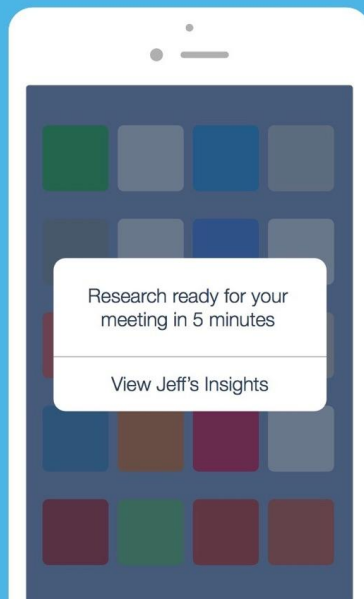
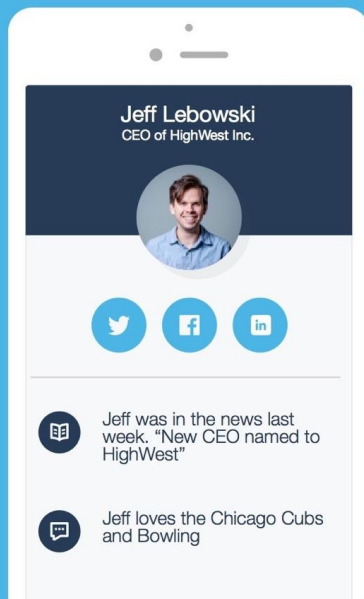
-O Brother Where Art Thou

Interactive intelligence

Charlie App

Get insights on
anyone you meet

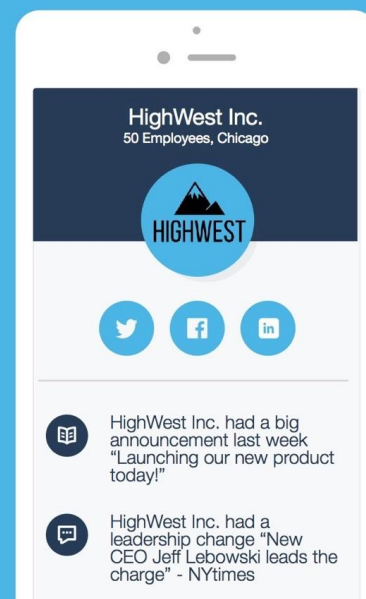
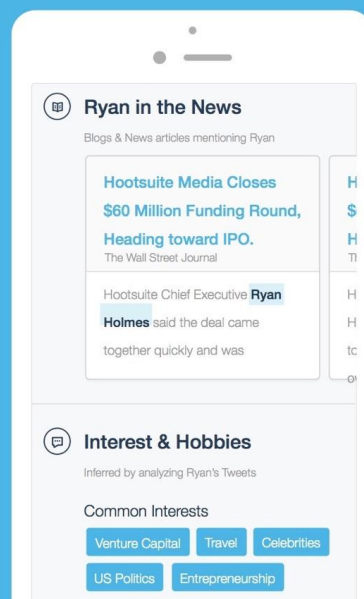
Minutes before every
calendar appointment



Collect and Curate

Always know how to
start the conversation

Deep Company and
Social Intel



Interactive search

1. Do until briefed:
2. Search
3. Select
4. Summarise
5. Update context

Information retrieval over sessions

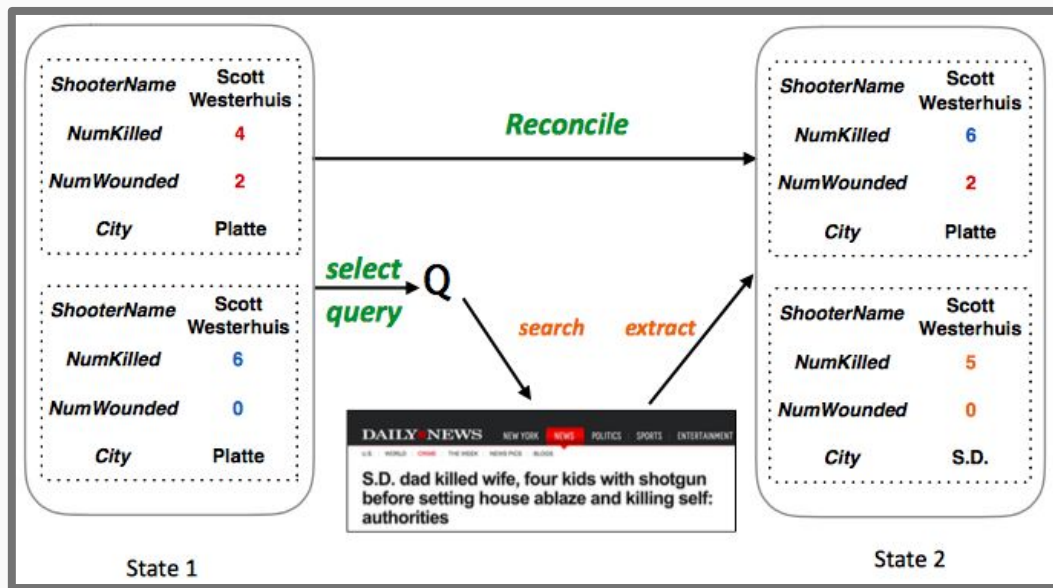
- Users typically refine queries during a session
 - In response to results
 - As need changes
- Need to model and evaluate multi-query sessions

Query: dehumidifiers

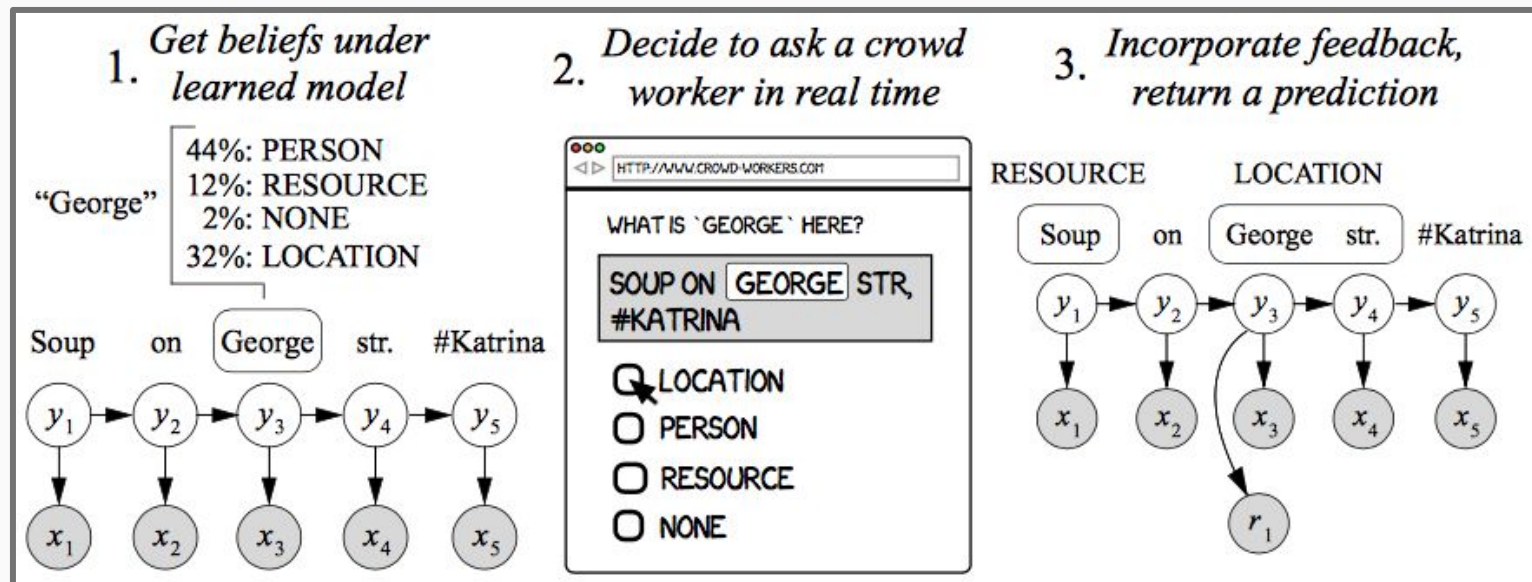
Description: *You would like to buy a dehumidifier. What are some of the technical specifications you should be looking at? What is the price range for dehumidifiers? What makes one dehumidifier more expensive than another?*

External evidence for information extraction

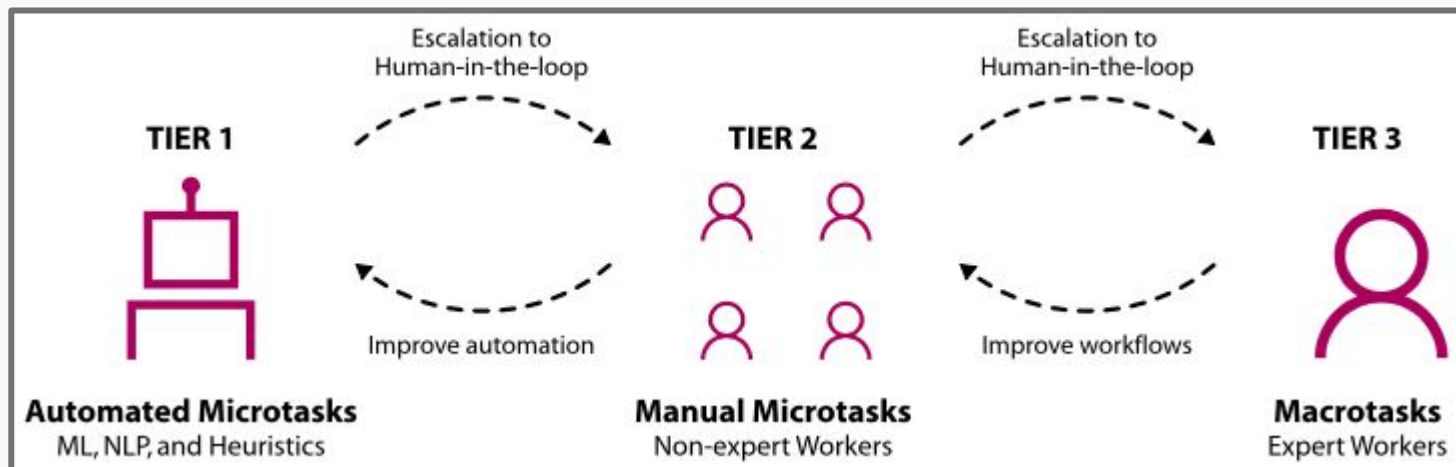
- Move beyond fixed collection
- Acquire and incorporate external evidence from web as required
- Repeat: search, extract, reconcile



On-the-job learning



Workflow agent with humans in the loop



ML + UX = Substantially less human effort

- 1. Do until briefed:**
- 2. Search**
- 3. Select**
- 4. Summarise**
- 5. Update context**

Lessons and challenges

- Lessons:
 - Wiki models require users who are invested in knowledge curation
 - Applications require interfaces that help users complete their work
 - Applications: entity enrichment, systematic review, other information curation tasks
- Challenges:
 - Interactive search and reconciliation
 - Extraction and enrichment across text, tables, DBs, etc
 - Context-aware generation with voice
 - Modelling search, extraction, reconciliation and generation

—

Thanks
Questions?