

The Computable News Project

Research in the newsroom

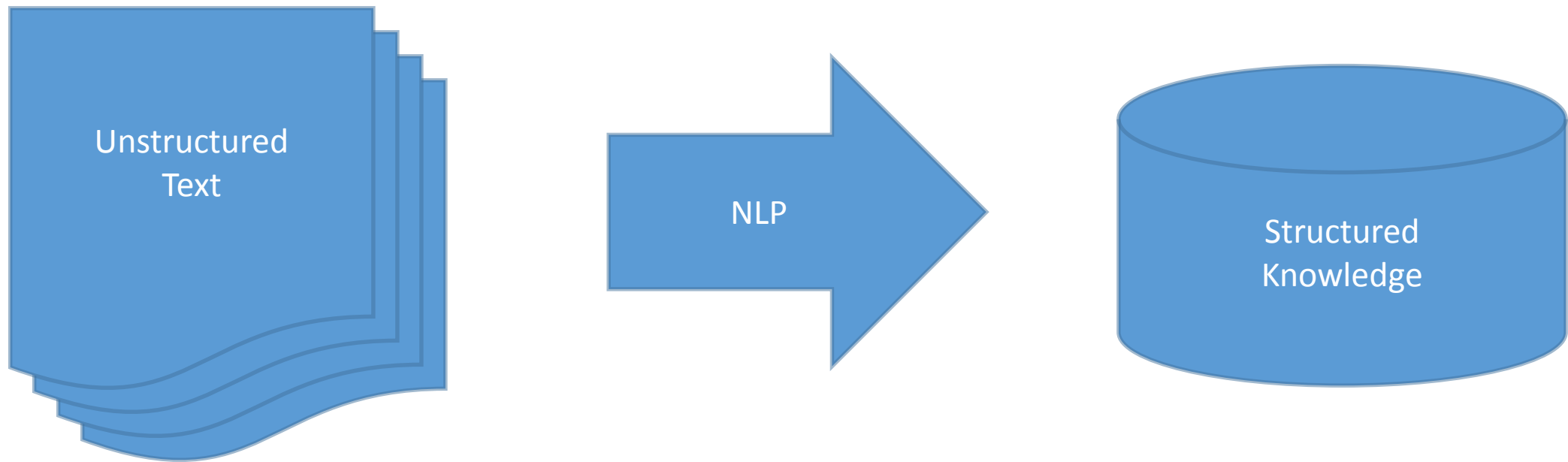
Will Radford, Daniel Tse, Joel Nothman, Ben Hachey, George Wright,
James R. Curran, Will Cannings, Tim O’Keefe, Matt Honnibal,
David Vadas, Candice Loxley

**The University of Sydney, Fairfax Media,
Capital Markets Cooperative Research Centre**

Motivation

Project vision & setup

Most news data is natural language



The Computable News project

Structured layer over unstructured text

Australian Government's Cooperative Research Centre

- Capital Markets CRC
- The University of Sydney (James Curran)
- Fairfax Media (George Wright)

Benefits

- Cost sharing
- Real research problems
- Focussed, time-limited research

Research

Named Entity Linking (Hachey et al., 13; Radford et al, 10-12; Pink et al. 13)

Quotation Extraction and Attribution (O'Keefe, 12-13)

Event Linking (Nothman et al., 12)

Named Entity Linking: text → KB entities

Sports reporter **Scott McIntyre** to sue **SBS**
for sacking over **Anzac Day** tweets

Scott McIntyre (reporter), **Special Broadcasting Service**, **Anzac Day**

- Extract: named entity recognition, heuristic coreference resolution
- Search: search Wikipedia for possible **KB entity** candidates
- Disambiguate: re-rank candidate lists using KB, context and whole document features
- Detect **NILs**: low-scoring entities may not be in our KB

How we use linked entities

Indexing

- Index documents by their entities
- Co-occurrence relation measures for entities
- Associate entities to images via linked captions

Challenges

- Ambiguity: ambiguous mentions, entity aliases
- KB dynamism: emerging entities, disappearing entities
- KB appropriateness: managing local news

Engineering for 25 years of archive

- (Mostly) data-parallel
- Experimented with database backends
 - Solr/Lucene
 - ~~Cassandra~~
 - Hypertable
- Document-representation format: DOCREP (Dawborn, 14)
 - Model data once, use often
 - *NIX pipeline operations

Quotations: who said that?

Mr Brown said “That’s untrue” (direct)

His spokesman said he returned the gift... (indirect)

Different systems

- Heuristic baseline
- CRF model over speaker IDs

Also: Debate opinion framing

Event linking: text → KB events

Rudd returned to the top job in June after challenging
Gillard in a caucus ballot.

More difficult than entities

- Identity
- Granularity

Formalised as hyperlinking a news archive

Applications

Writing news, researching entities, reading distilled news, curating knowledge

Editor: helping journalists write richer stories

Compnews Editø

1 [Tony Abbott] (e:Tony_Abbott)

2 announces new budget with [Joe

3 Hockey] (e:Joe_Hockey) .

4

5 Mr [Hockey] (e:Joe Hockey)'s

6 wife, Melissa Babbage, also

7 attended the speech.

Tony Abbott announces new budget with Joe Hockey.

Mr Hockey's wife, Melissa Babbage, also attended the speech.



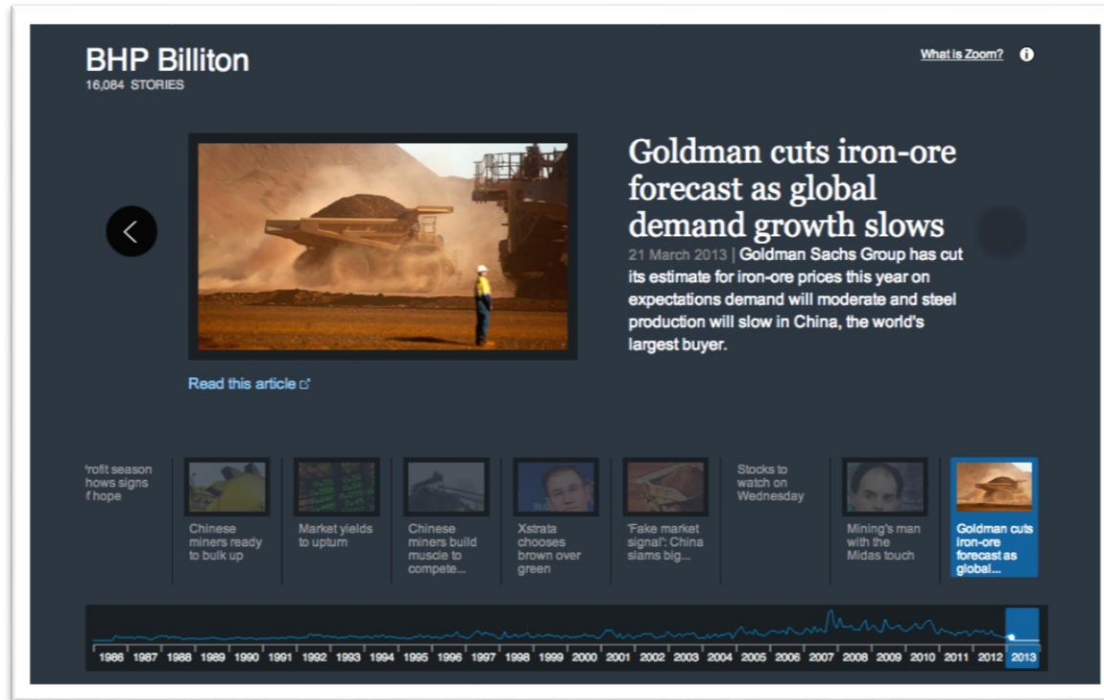
Tony Abbott
Anthony John "Tony" Abbott is the Leader of the Opposition in the



Joe Hockey
Joseph Benedict "Joe" Hockey, is an Australian politician and

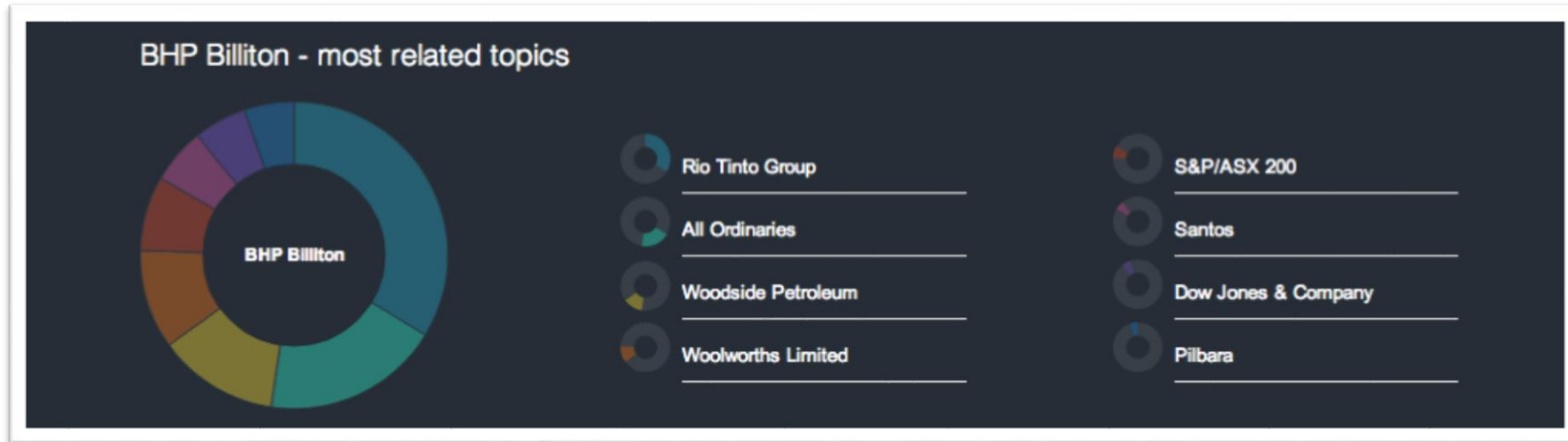
- Automatic text augmentation
 - Entity links `[mention]{e:entity_name}`
 - Knowledge
 - `{partner}` → Melissa Babbage
 - `{age}` → 48 years old
 - `{map}` → `coordinates + maps.google.com`
- **Idea: recreate a NER/NEL corpus every week**

Zoom: accessing the archive through entities



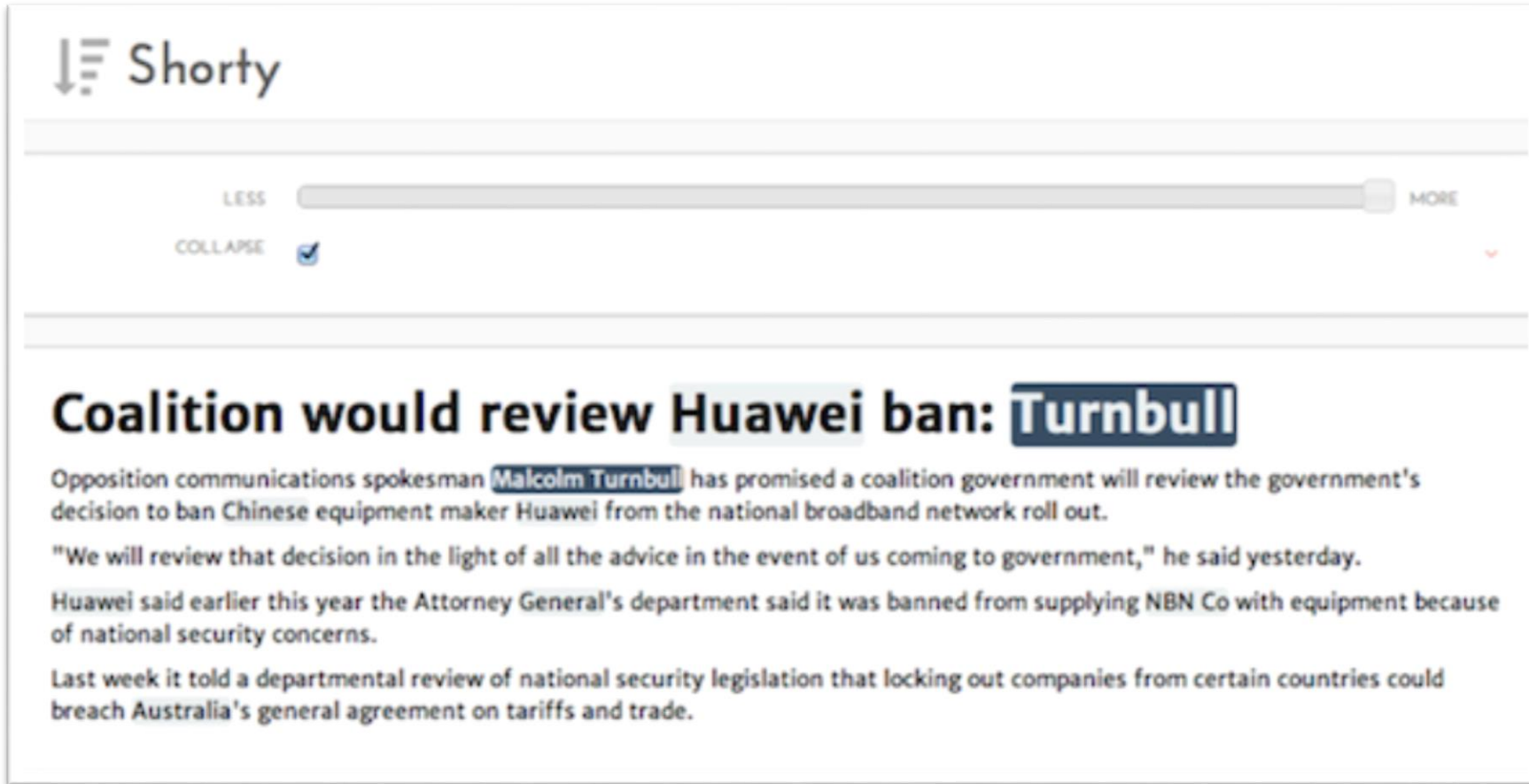
- Landing page for each entity
- Images and snippets
- Mention frequency timeline
- Which stories do we show?
- **Idea: drive traffic deeper into the archive**

Zoom: accessing the archive through entities



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

Shorty: summarising stories



The screenshot shows the 'Shorty' web application. At the top left is the 'Shorty' logo with a downward arrow icon. Below the logo 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 checked checkbox. 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. In the first paragraph, 'Malcolm Turnbull' is highlighted in a blue box. In the second paragraph, 'Huawei' is highlighted in a blue box. In the third paragraph, 'locking out' is highlighted in a blue box.

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.

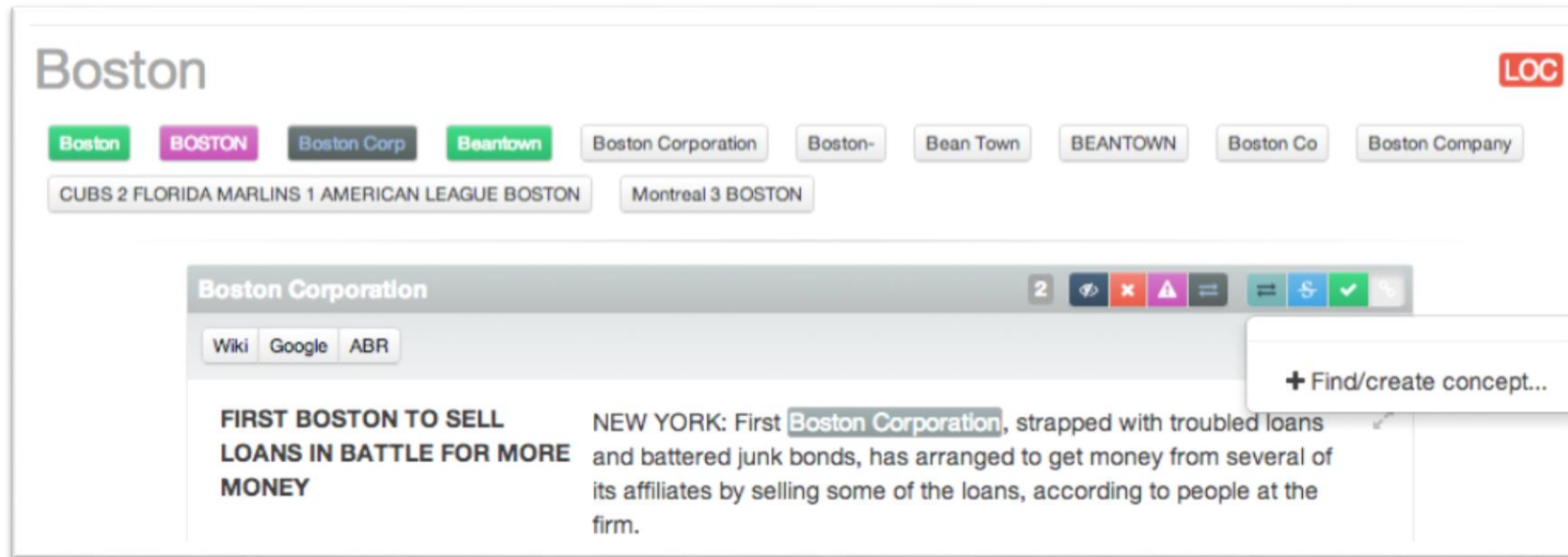
"We will review that decision in the light of all the advice in the event of us coming to government," he said yesterday.

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**

Correction: maintaining the KB



- Fix mention errors in batch or individually
- Add new entities as they become newsworthy
- **Open question: how much do you fix?**

Discussion

What we learned

What the researchers learned

- Shared task participation is useful (i.e. TAC KBP)
- Demonstrations to business > table of figures
- 80% still means 1 in 5 wrong
- Forgivable errors != stupid errors
- Time spent engineering is mostly well spent

What the business learned

From users

- Zoom is good for topic-driven news, not for breaking news
- Journalists like doing less work, exploring related entities
- Editorial staff likes analytics: internal and external

Conclusion

Overview of:

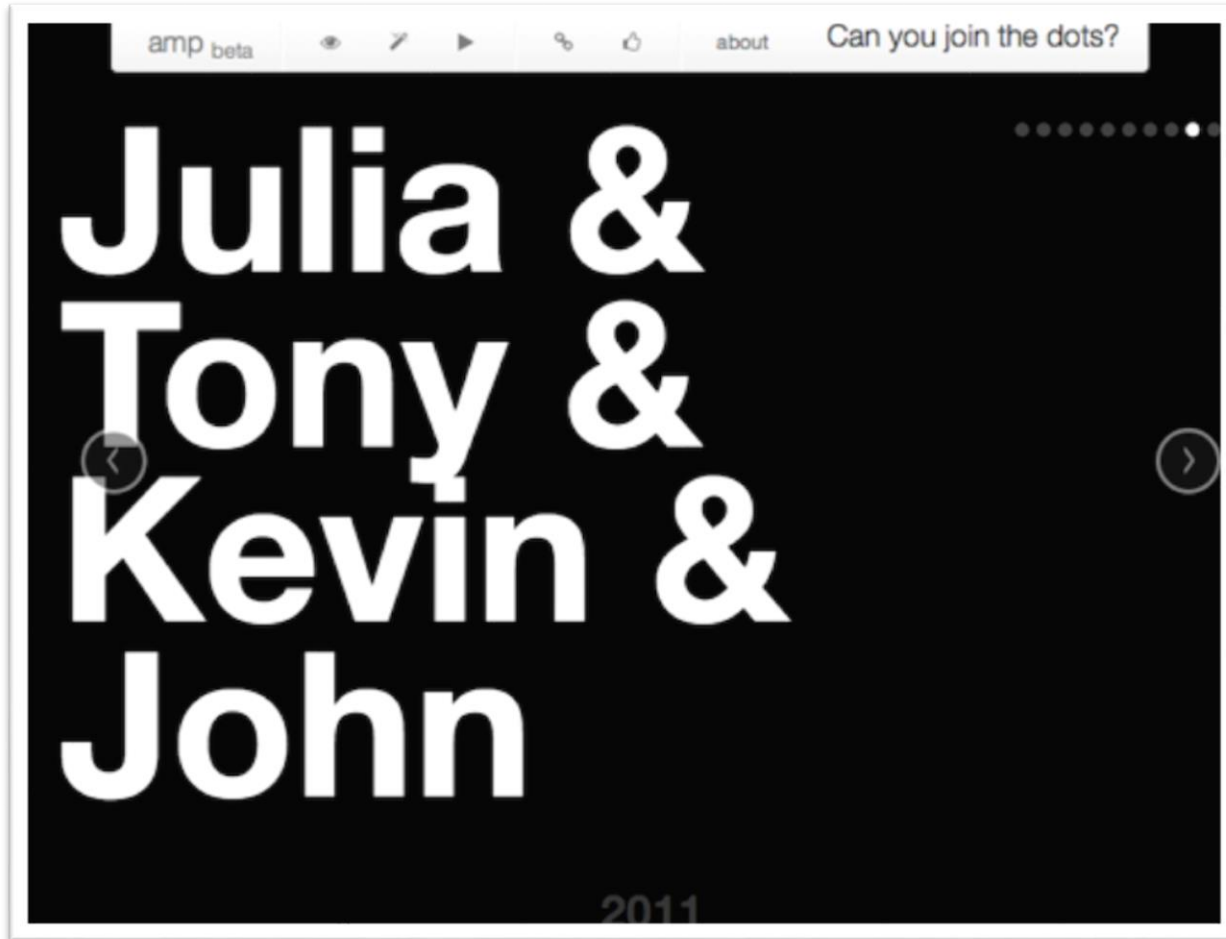
- Entity-centric news research programme
- Wide range of applications

What's Fairfax doing now?

- Early bets on research have increased data literacy
- “Future Services” has become “Intelligent Systems”

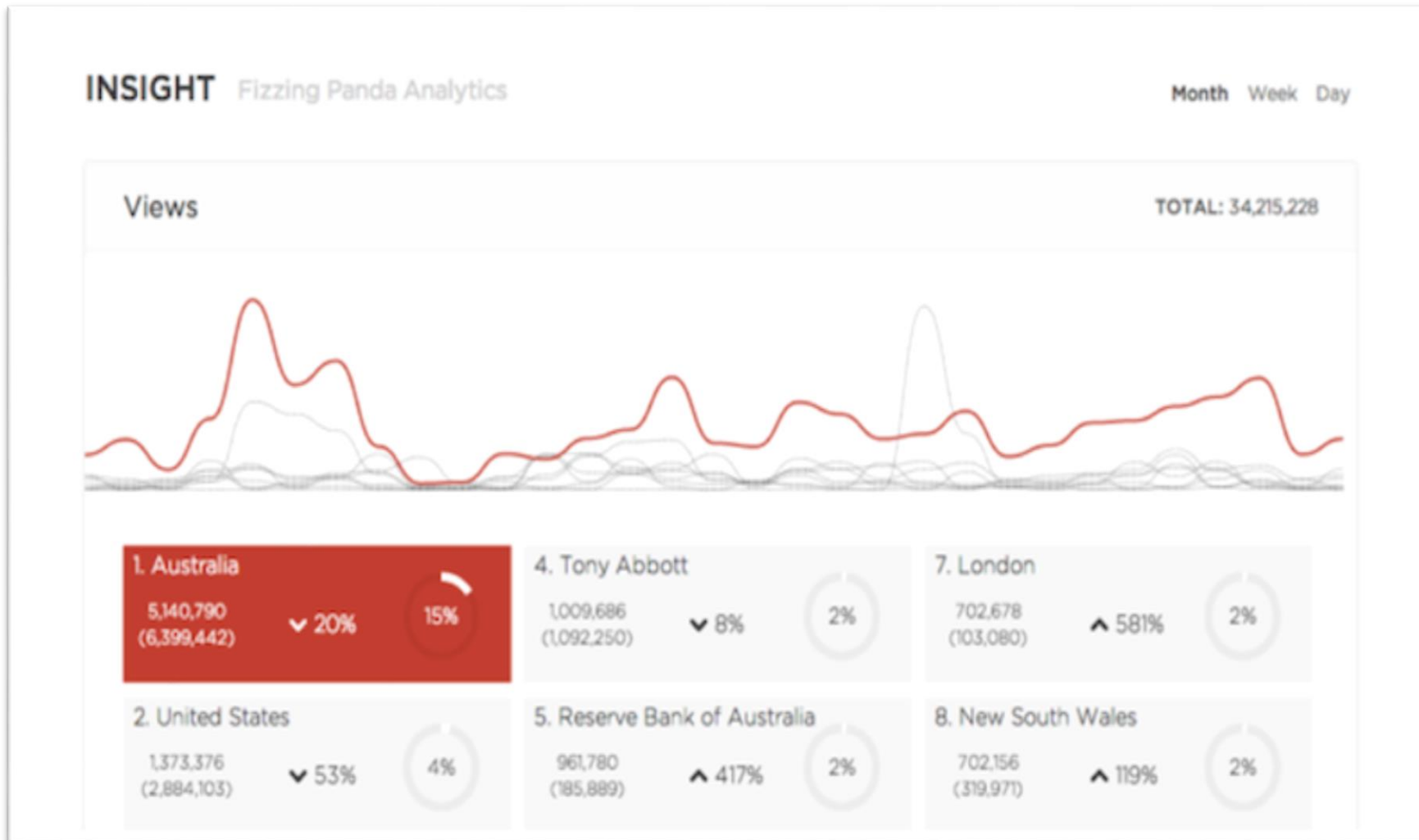
Thanks for listening.
Questions?

Amp: news trivia games



- Show first names of related people
- Guess the last names
- **Idea: KB-sized games, what could you do with feedback?**

Insight: supporting editorial



- What entities do people:
 - read about
 - share
- **Open question: how much should analytics drive editorial?**