Analyzing IBM Watson Experiments With IPython Notebook

Torsten Bittner
IBM Emerging Technologies
@torstenbittner

SciPy 2013, Austin TX, USA

June 27, 2013

Agenda

- IBM Watson in a nutshell

- Domain adaptation challenges

- Case study for analysis tooling:

    - Java, Dojo, D3 - based approach

    - IPython Notebook - based approach

IBM Watson in a Nutshell

1. Ingest

    • Identify, pre-process data, create search indices + other data stores

2. Train

    • Create QA pairs as training set, run in Watson, train ML models
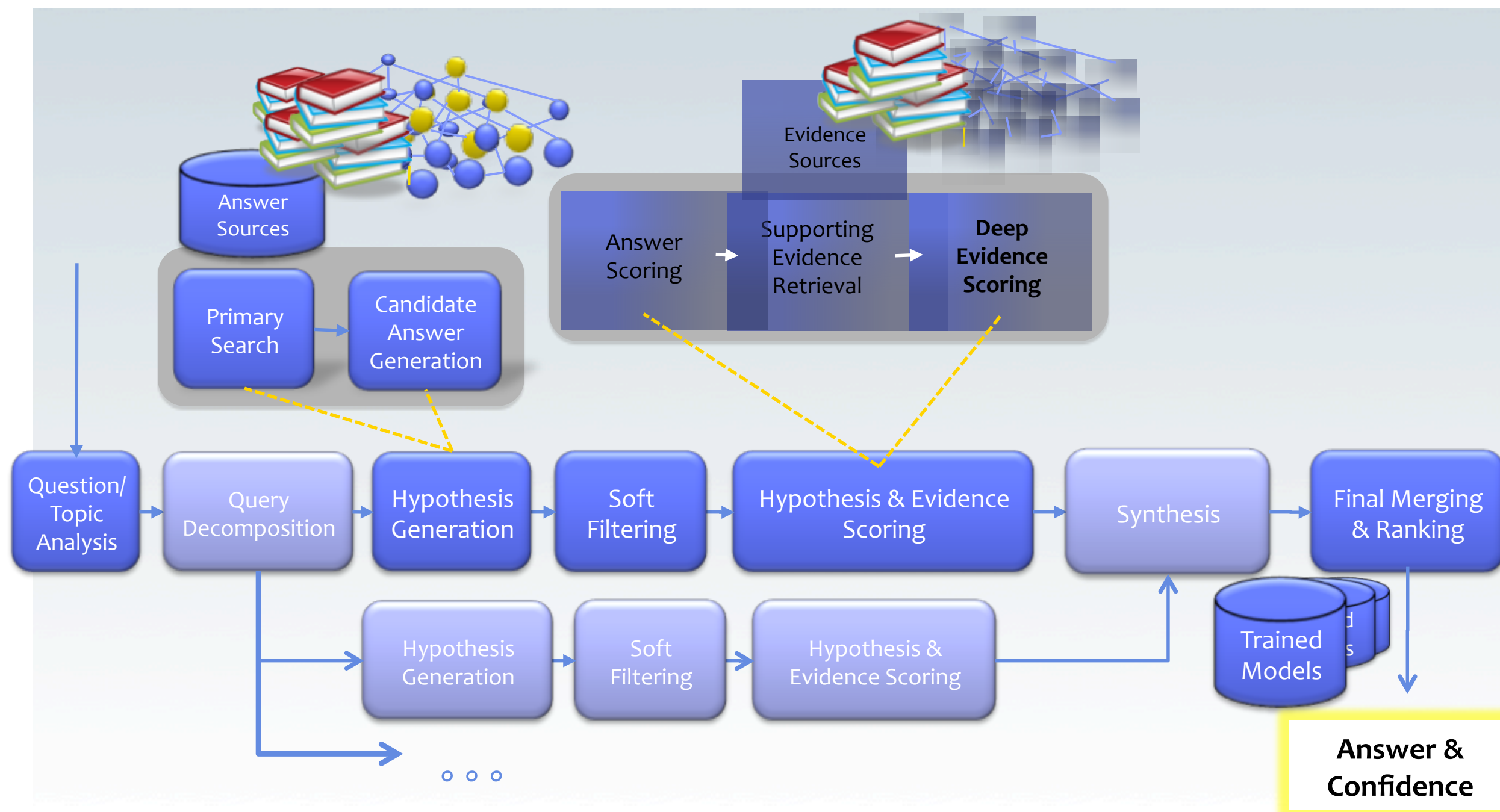
3. Test

   • Run Watson with blind test set and measure accuracy

4. Production

    • Run Watson with unknown questions

IBM Watson in a Nutshell

Domain Adaptation Challenges

- IBM Watson is large and complex, lots of:

  - Knobs - e.g., data, algorithms, models

  - Dependencies:

    - between components
    - between changes

Ingest

Question
Analysis   Search

Candidate
Generation   Scoring   FMR

Information
Dependencies

Pipeline processing

  - Metrics - Which ones are important?

  - Experiments - Which should be run?

- Documentation - How to track changes and attribute results to experiments?
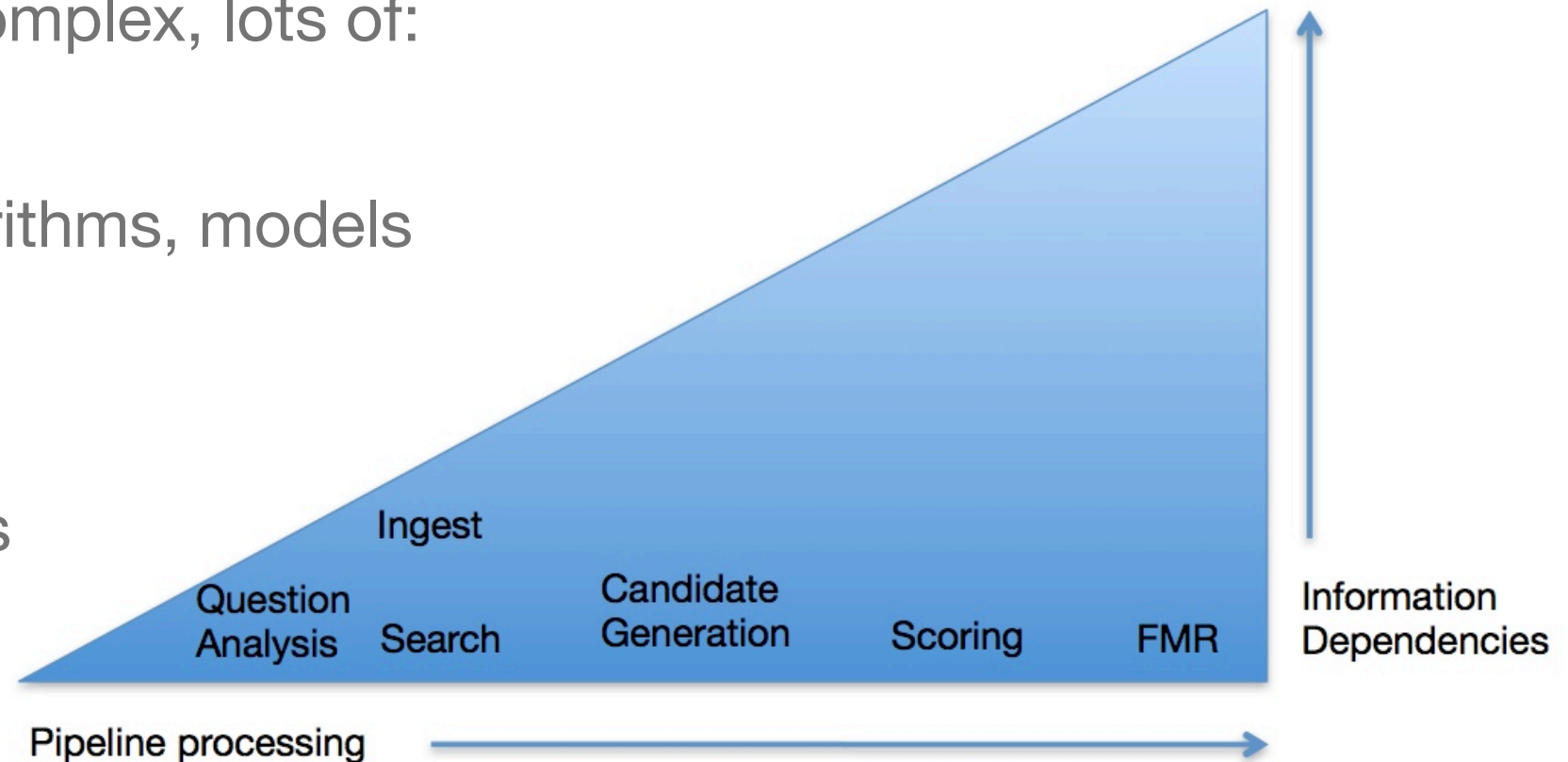
Our Driving Principles

- Apply scientific method.

- Perform ad-hoc analysis.

- Study the knobs, then turn - one at a time.

- Focus on something measurable.

- Develop tooling to support process.

"If you can't solve a problem, then there is an easier problem you can solve: find it."
-George Pólya, *How to Solve It (1945)*

Case study - Question Class Analysis Using Java / Dojo / D3

- Analysis result:
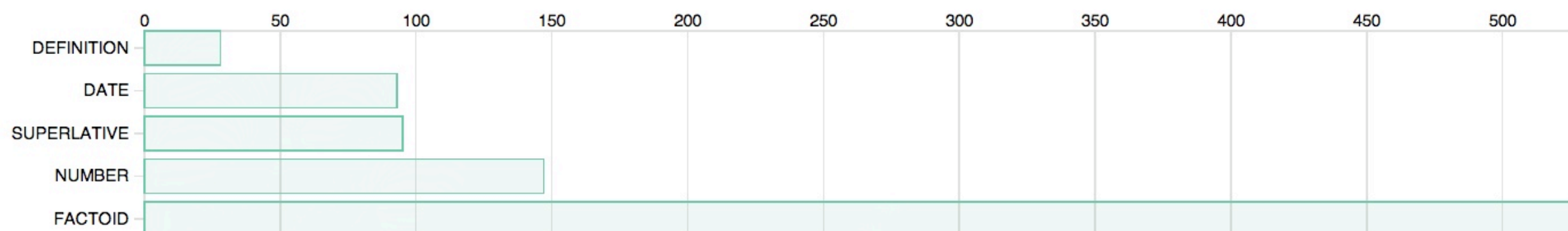
- Effort:

Feature	LOC	Language
Container Web App	150	HTML / JavaScript / Dojo
Extract relevant experiment results	7900	Java
D3 visualization	300	JavaScript / D3
Total:	8350	3

Case study - Question Class Analysis Using IPython Notebook

- Analysis result:

# Case study - Question Class Analysis Using IPython Notebook

- Effort:

Feature	LOC	Language
IPython analysis notebook	20	Python
Extract relevant experiment results (module "result_proc")	200	Python
Total:	220	1

- Setup:

  - IPython Notebook installed in Python virtual environment

    - Packages: Pandas, NumPy, SymPy, Matplotlib, scikit-learn

  - Deployed nbserver for each team member in cluster environment

  - Shared file system for direct access to experiment results

  - Workers deployed across 10 machines for parallel processing
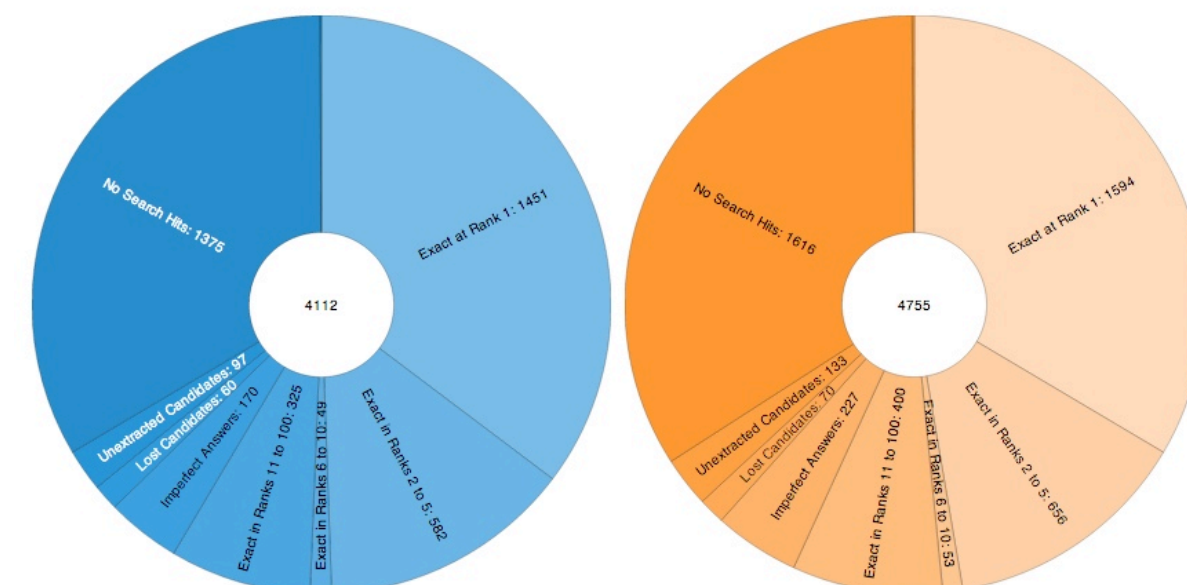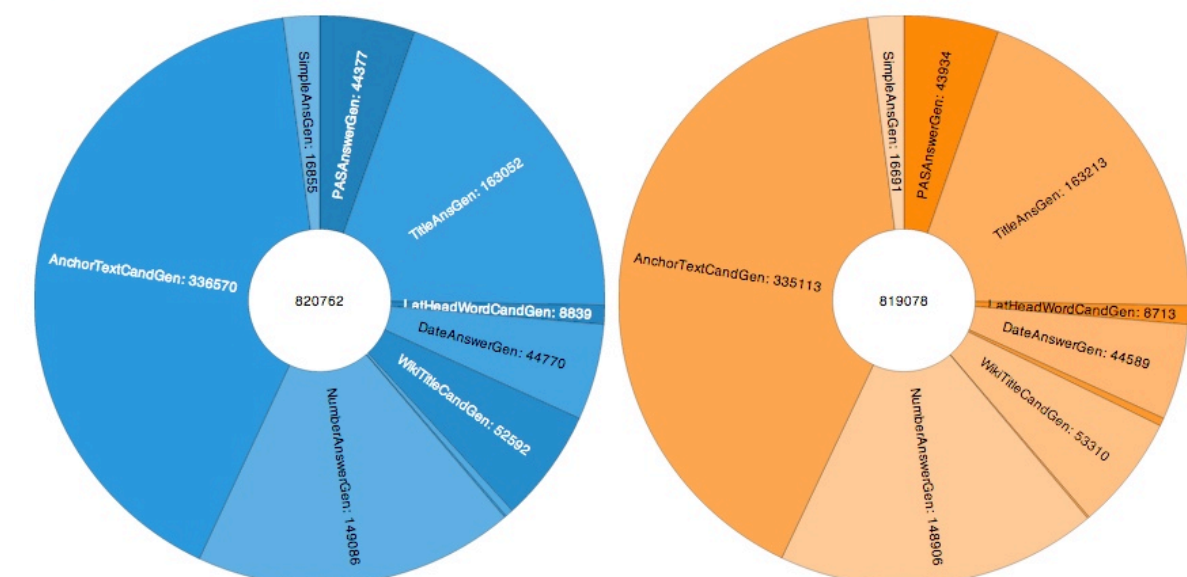
More analysis examples

Experiment Recall Shift

Slice Recall Breakdown

By Candidate Generator
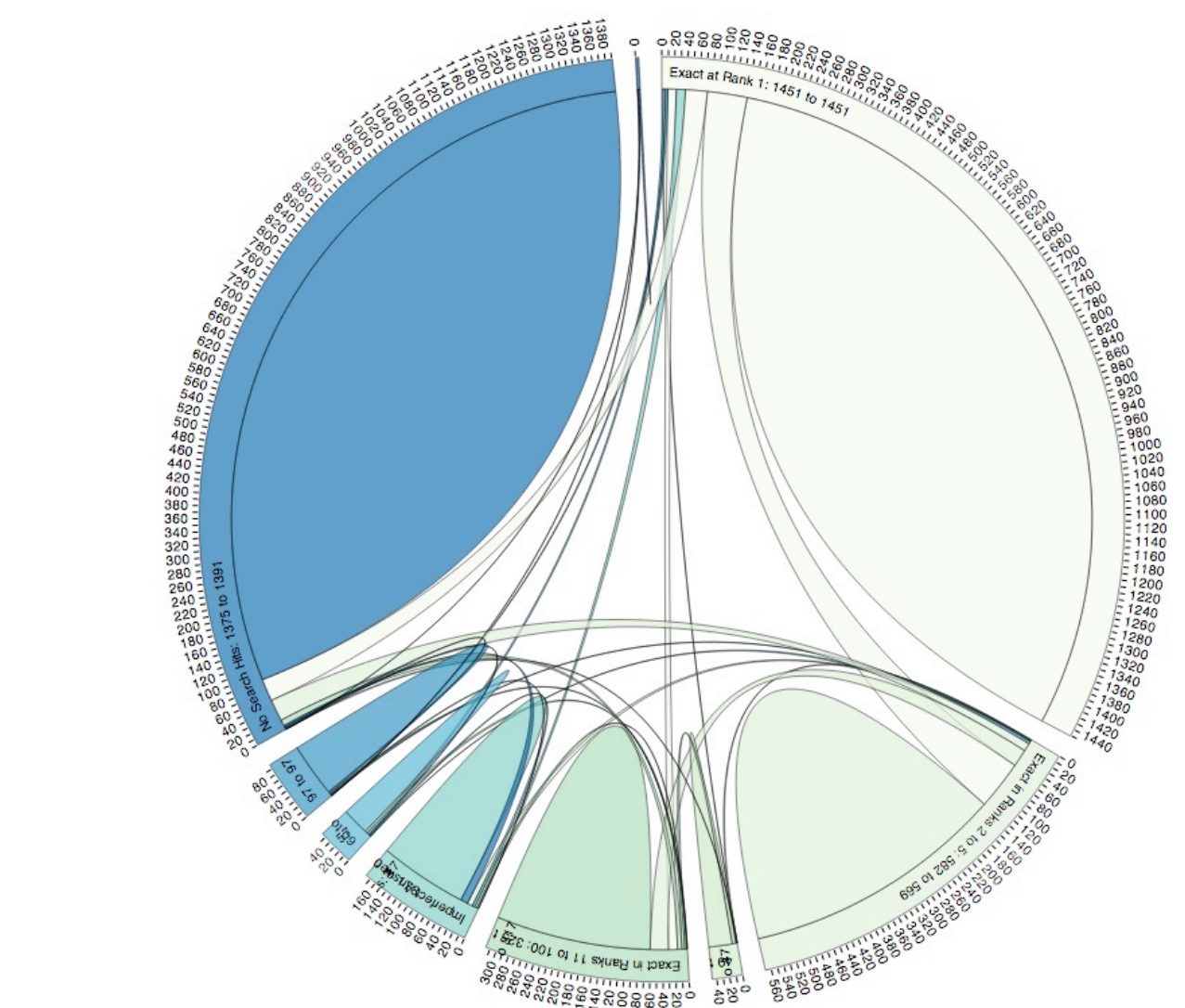
experiment	qid	text	matchesExactInCandidates	score@1	topCorrectRank	hutts
INT13	844783	What is an example of a virus for the Linux operating system?	6	0.16	17	

IPython Notebook pro's

- Less languages, less code, many Python libraries

- Broader population enabled to carry out analyses (data scientists)

- Code = Analysis

- Documentation, hypothesis, experiment description and analysis inline

- Results easy to reproduce, communicate and share

- Scalability though in memory-processing (25 GB/s)

- Simple remote development

- Flexible usage of notebooks for:

  - Experiment analysis

  - Ingestion

  - QA pairs processing

  - ML analysis

Summary

- IBM Watson is a powerful, complex system - domain adaptation requires changes to data, algorithms and models

- IPython Notebook is great tool to analyze IBM Watson experiments:

  - ad-hoc data manipulation

  - ad-hoc analysis and visualization

  - documentation

  - parallel processing

- Try it - http://ipython.org/notebook.html