Principles for Robust Evaluation Infrastructure

Mark Sanderson (RMIT University, Melbourne, Australia)

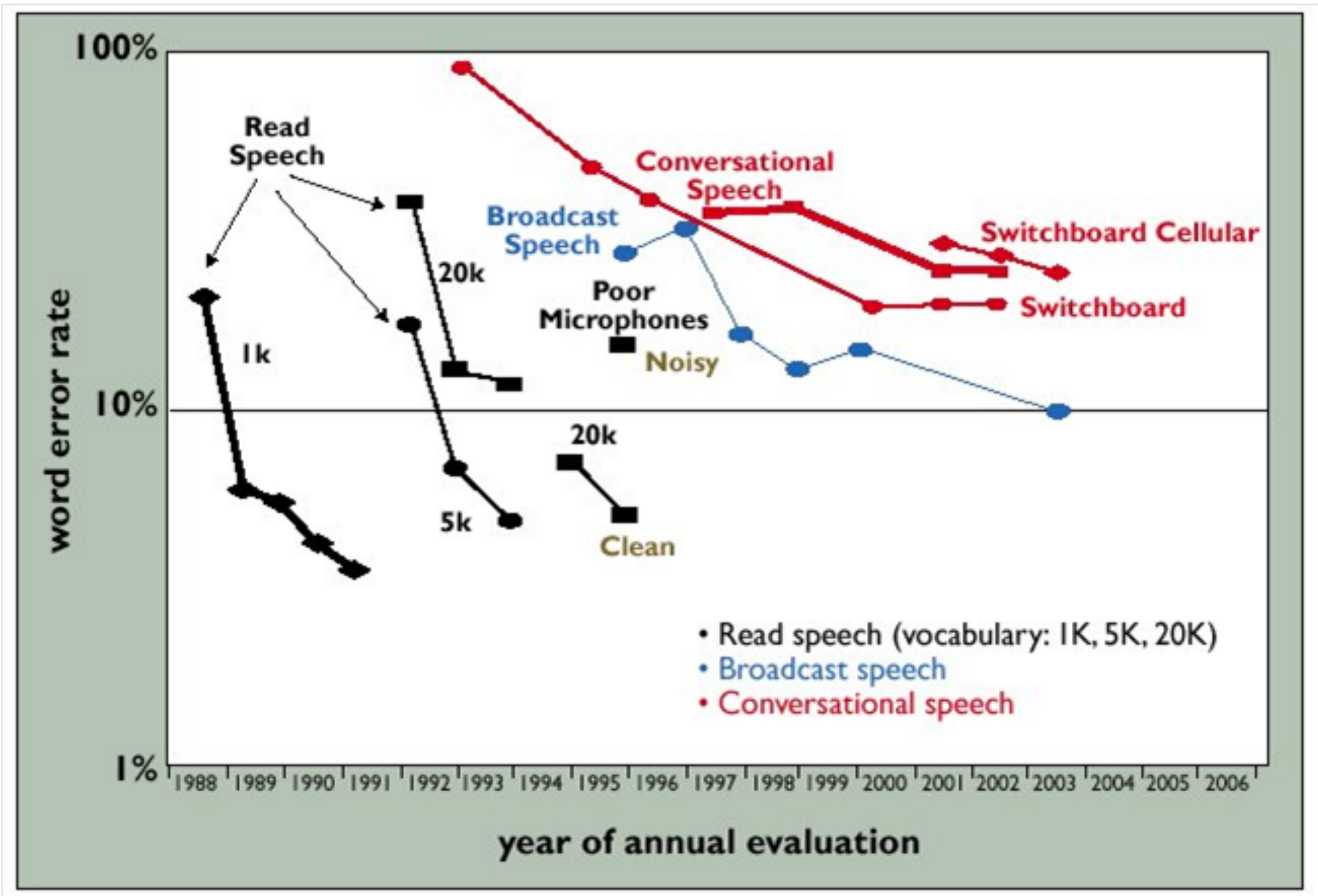
William Webber (University of Maryland, College Park, MD, USA)

NTCIR-9, Tokyo, December 2011

Overview

- Effective reproducibility requires public data and open systems
- Progress needs to be measurable
- Experimental design should be based on statistical principles

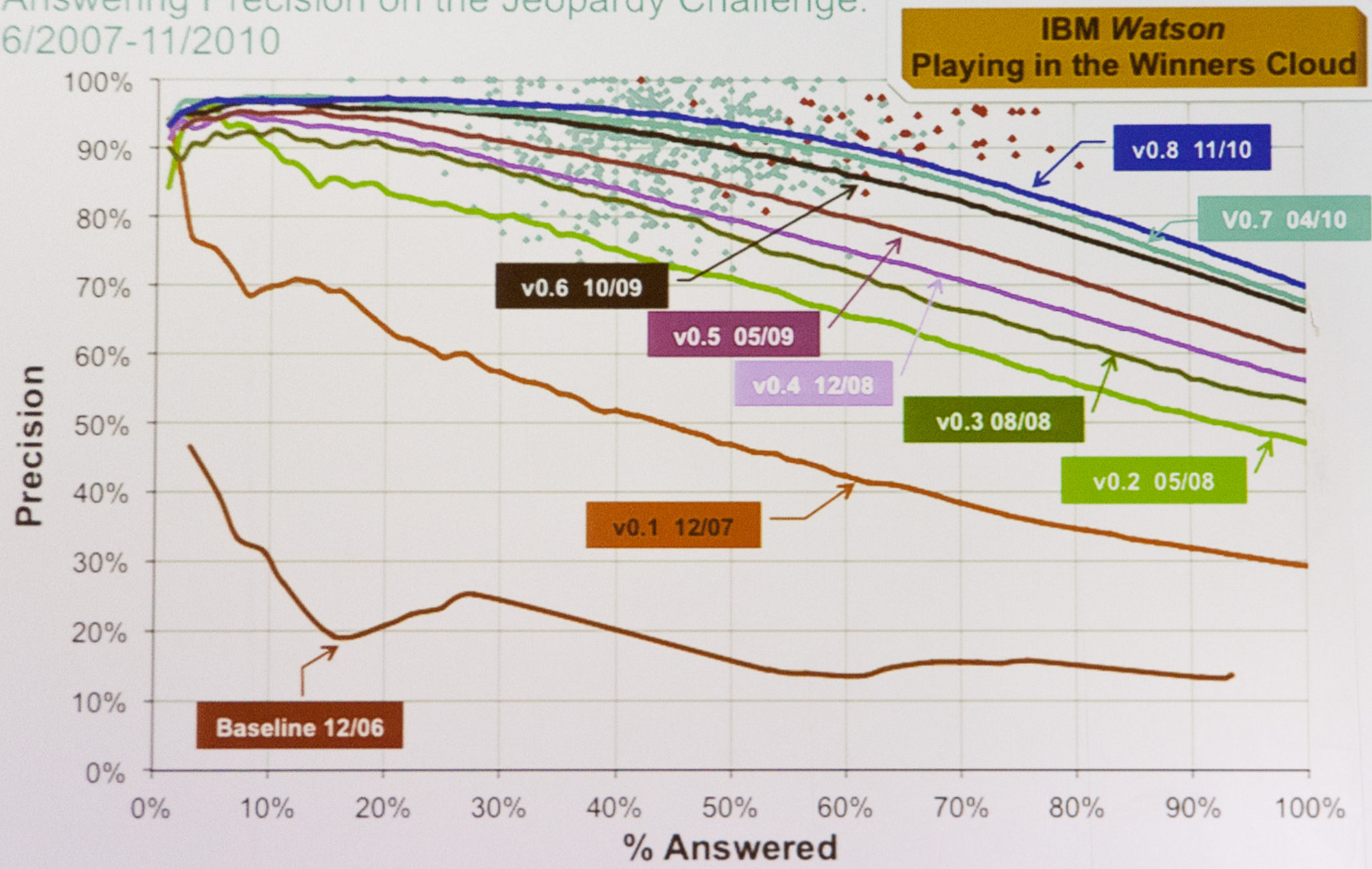
Measurable improvements over time



Deng, L., Huang, X. (2004) Challenges in adopting speech recognition, Communications of the ACM, 47(1), 69-75

DeepQA: Incremental Progress

Answering Precision on the Jeopardy Challenge: 6/2007-11/2010



Implications by not tracking

- Each year we run on new topics
  - Building bigger collection
  - Sometimes learning how to build collection
- But
  - Are we beating last year's systems?
  - Do we have good stopping rule for tracks?

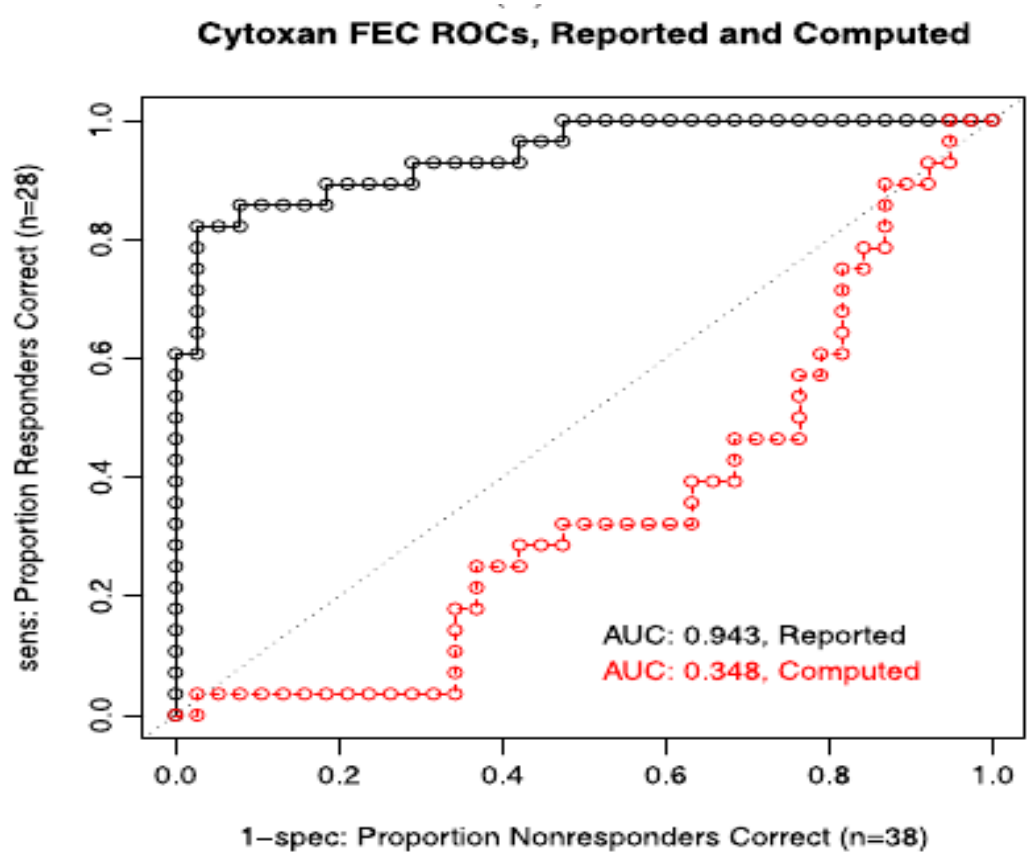
- Is there another way?
  - Repeat on earlier topics
    - trust participants not to cheat
  - Work on very large topic set: incrementally build qrels over years

Reproducibility: a cautionary tale

- Potti et al.,2006: method for microarray screening for sensitivity to treatments for cancer patients.
  - More reliable, more effective, personalized cancer treatment.
  - Patent; research grants; 200+ citations
  - Implementation in clinical trials
- But methods poorly described, difficult to replicate
- Baggerly and Coombs undertook extensive reverse-engineering of results
  - found extensive, basic errors (off-by-one errors, reversed sensitive/resistant labels, etc.)

"Unfortunately, poor documentation and irreproducibility can shift from an inconvenience to an active danger when it obscures not just methods but errors"

Result replicability and verifiability



Keith A. Baggerly, Kevin R. Coombes (2009), Deriving chemosensitivity from cell lines, Annals of Applied Statistics, 3(4), 1309-1334

- Reported sensitivity prediction close to perfect
- Actual predictions worse than random

How reproducible are our results?

- Methods often poorly described in IR papers.
- Private datasets prevent re-running of experiments.
- Lack of intermediate results mean reviewers, readers cannot check results for errors.
- For many experiments, everything is done in code. Why isn't the code provided?

“Most common errors are simple … [and] most simple errors are common” (Baggerly and Coombes, 2009)

"Most Published Research Findings Are False" (Ioannidis, 2005)

Better practice, better standards

- Provision of working code, original and intermediate data, analysis scripts, should be standard, reported.
- Evaluation campaigns should encourage reproducibility:
  - Use closed, not open, corpora
  - Require capture of data derived from external data sources (e.g. translations from Google Translate, pages retrieved from Bing search, etc.)
  - Submission of working code, in form that can be automatically re-run on new collections
- In automated IR, we rarely have a technical excuse for reproducible results.
- We should be leading, but actually are trailing, other sciences in our practice.

Scale of our experiments

- Classic research says 25-200 topics are enough
  - Lots of research to back this up
- But are we checking this well enough?
  - Search engines routinely used thousands of topics
    - Why?

Basing experiments on statistical principles

- Classic significance test make assumptions about data
  - Often IR data breaks these assumptions
  - Evidence tests aren’t working well enough
- Alternatives
  - Only less used historically
  - Require computational time
- Tests
  - Bootstrap
  - Randomisation (permutation) test

References

- Crook, T., Frasca, B., Kohavi, R., & Longbotham, R. (2009). Seven pitfalls to avoid when running controlled experiments on the web. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1105-1114).
- Sakai, T. (2007). Evaluating Information Retrieval Metrics Based on Bootstrap Hypothesis Tests. *Information and Media Technologies*, *2*(4), 1062-1079.
- Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 623-632). ACM New York, NY, USA.

Public, private, and grey data

- Increasing use of private data, particularly from commercial research labs and their collaborators
  - Larger volumes of data than available to public researchers (e.g. 10,000 assessed queries, not 50)
  - Types of data, particularly user behaviour data (query logs, click-through data, browser taskbar)
- Data not publicly released
  - Commercially valuable information
  - Privacy-sensitive data
- Problems with use of private data:
  - Private data leads to private fields of research
  - Results not reproducible by others

Solutions to private data

- Where possible, reproduce experiments on public data sets
- Algorithm deposit models
  - Send in your code, rather than sending out your data
- Statement of data (and code) availability for each published paper
- Educate reviewers to consider whether use of private data vitiates results