Evaluation Campaigns

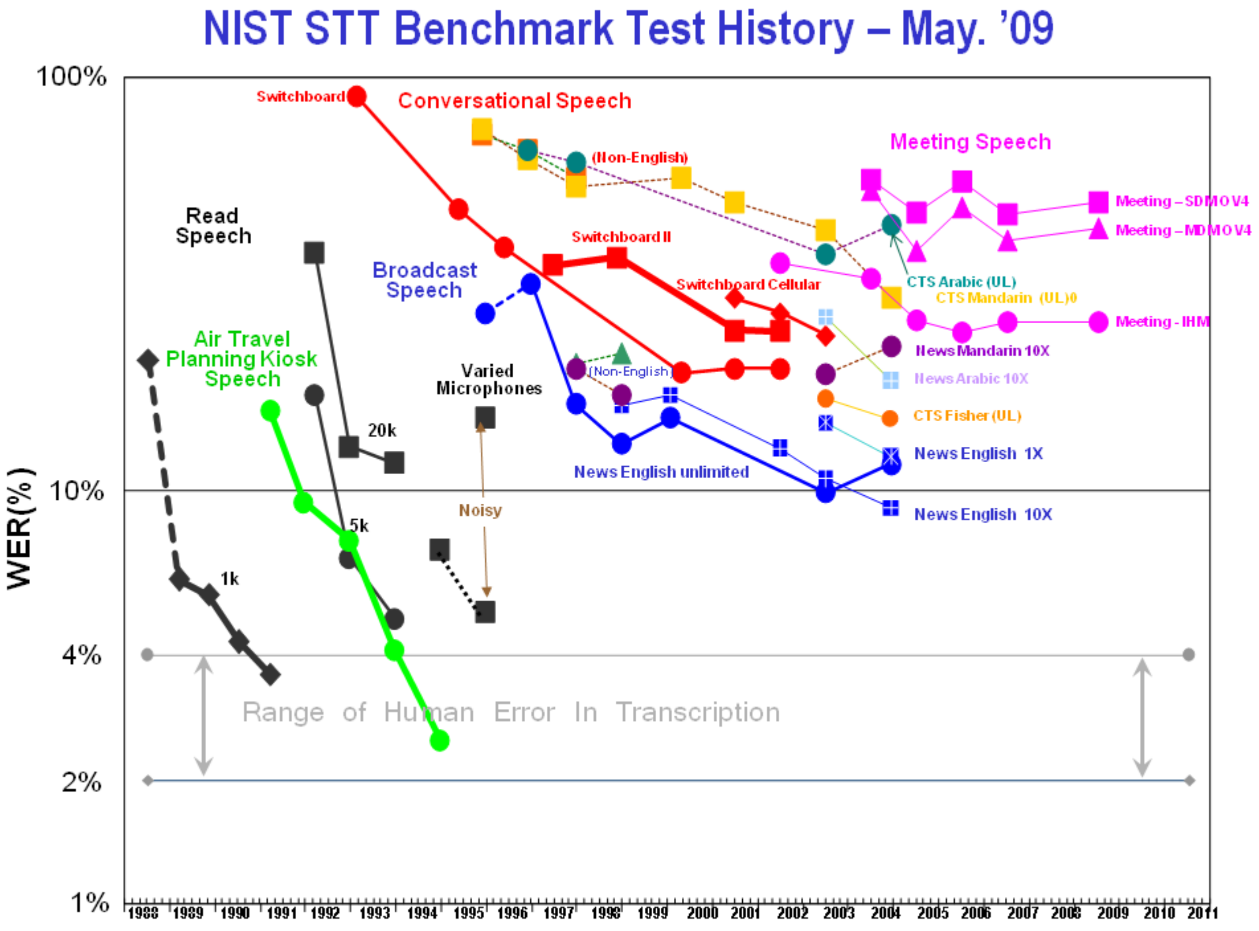
1. Speech
2. Others
3. General Points

Dan Ellis
Dept. Electrical Engineering, Columbia University
dpwe@ee.columbia.edu   http://www.ee.columbia.edu/~dpwe/e4896/

Evaluations

- Systematically evaluating research output with common data & metrics

- DARPA/NIST Speech Recognition as the original & canonical example:


NIST STT Benchmark Test History – May. '09

http://www.itl.nist.gov/iad/mig/publications/ASRhistory/

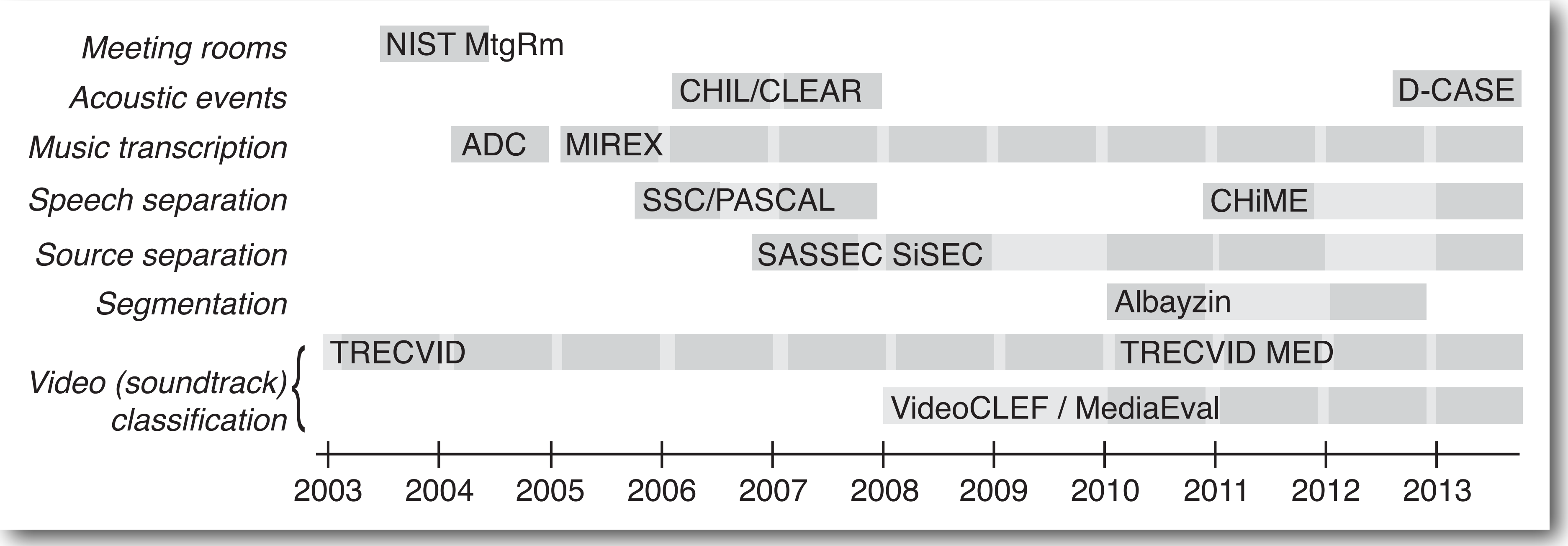
The Origin of Evaluations

- Mark Liberman:
  "Avoiding glamour and deceit"
  - placate funders!

- American Association for the Advancement of Science Meeting, 2011-02-19,
  The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer
  http://languagelog.ldc.upenn.edu/nll/?p=2976
  http://www.stanford.edu/~vcs/AAAS2011/AAAS2011Liberman.pdf

Other Evaluations

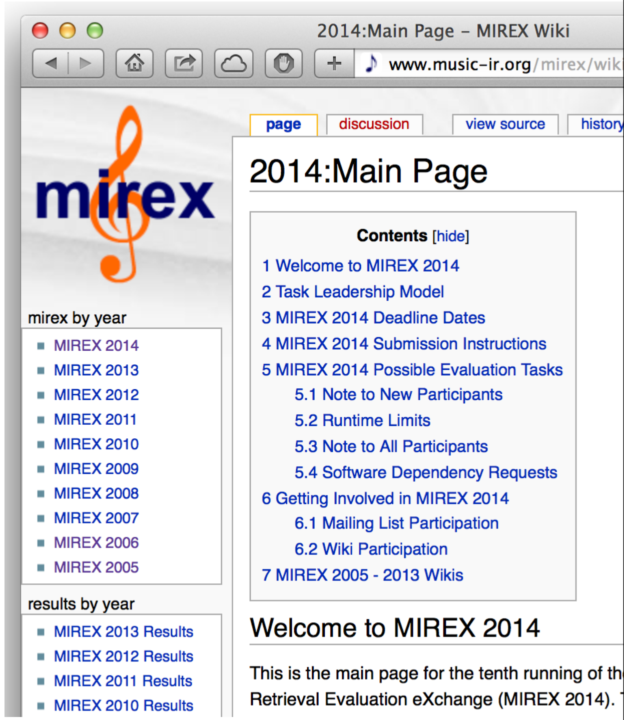
- Benefits to speech led to many copies
  - now standard for DARPA and IARPA programs
  - emulated in many other fields
  - typically volunteer-funded
- Example: Sound Scene Analysis evaluations



- Metrics: SNR, Frame Acc, Event Error Rate, mAP

MIREX

- **Effort to compare Music Information Retrieval algorithms**
  - organized by Stephen Downie, UIUC
  - funded by Mellon Foundation
- **First round in 2005**
  - 5-15 tasks per year, 3-20 participants per trial
    = 2037 algorithms run over 37 datasets in 10 years
- **Impact**
  - organized, solidified research areas - chords, covers
    - focus of community discussion of agenda
  - no open release of data - you have to participate



mirex by year

- MIREX 2014
- MIREX 2013
- MIREX 2012
- MIREX 2011
- MIREX 2010
- MIREX 2009
- MIREX 2008
- MIREX 2007
- MIREX 2006
- MIREX 2005

results by year

- MIREX 2013 Results
- MIREX 2012 Results
- MIREX 2011 Results
- MIREX 2010 Results

2014:Main Page

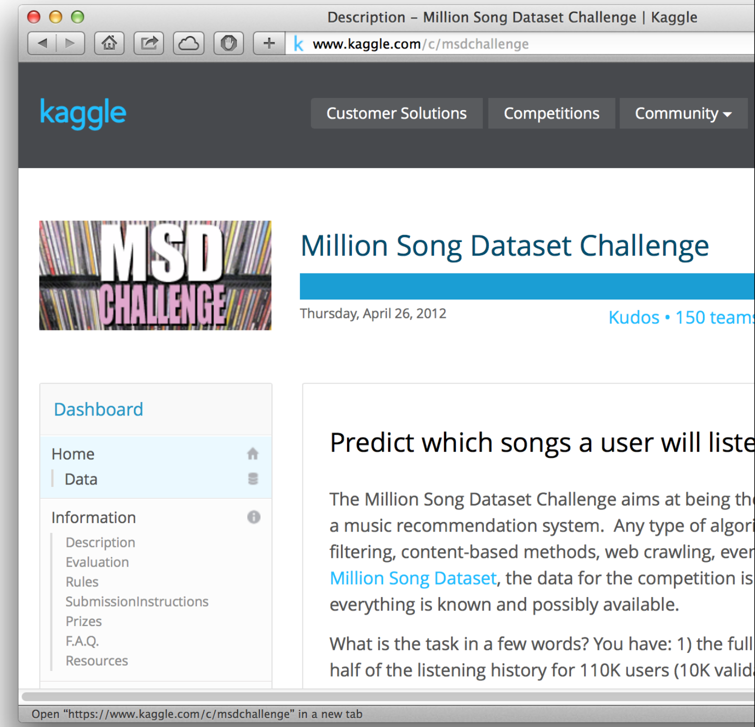
Contents [hide]

Welcome to MIREX 2014

This is the main page for the tenth running of th
Retrieval Evaluation eXchange (MIREX 2014).

Million Song Dataset Challenge

- Listening history data for 1M+ listeners
  - but not time-stamped
  - task is to rank tracks based on partial history
- kaggle.com: "predictive analytics leader"
  - actually, a platform for big-data challenges



Million Song Dataset Challenge

Thursday, April 26, 2012

Kudos • 150 team

Predict which songs a user will liste

The Million Song Dataset Challenge aims at being th
a music recommendation system. Any type of algori
filtering, content-based methods, web crawling, ever
Million Song Dataset, the data for the competition is
everything is known and possibly available.

What is the task in a few words? You have: 1) the full
half of the listening history for 110K users (10K valid

Open "https://www.kaggle.com/c/msdchallenge" in a new tab

- Competition
  - ran for 4 months in 2012; 150 teams participated
  - avg. prec. improved from 0.024 to 0.179
  - .. but no audio features used!

Aspects of Evaluations

- **Relevance of task & metrics**
  - at least you'll solve one task
- **Scale matters**
  - for statistical significance & non-over-fitting
- **Encouraging participation**
  - plusses and minuses of participating
- **Models for distributing the effort**
  - it's a lot of work to run these systems; who pays? +secrets
- **Ensuring the sharing of information**
  - opportunity to share code?
- **Releasing test materials**
  - .. for extensive post-mortems … but next time?

Impact of Evaluations

- Good:
  - direct comparison of techniques
    - invest with confidence!
  - focus community research effort

- Bad:
  - non-evaluated topics are starved of attention
  - leads to conservative monoculture
    - puts off good newcomers?
  - too much focus on one number…

Summary

- Glamour and Deceit
  - common data & tasks provide clarity
- Knowledge and Progress
  - identify the things that work (and how they combine)
- Data and Code
  - systems that conform to a common standard are (more) ready for sharing