Why Human Language Technology (almost) works

Mark Liberman

University of Pennsylvania

http://ling.upenn.edu/~myl

Why Human Language Technology (almost) works

(. . . and what scientists should learn from this)

Mark Liberman

University of Pennsylvania

http://ling.upenn.edu/~myl

Let’s start by establishing that HLT (almost) works…

Questions to **OK Google**, in a quiet room, on an Android Nexus 5:

Question: “OK Google, what is the French word for ‘dog’?”

Transcribed as: “what is the French word for dog?”

Answer: **“**chien”

Question: “OK Google, what is 15 degrees centigrade in Fahrenheit?”

Transcribed as: “what is 15 degrees centigrade in Fahrenheit?”

Answer: **“**15 degree Celsius is 59 degrees Fahrenheit.”

Q: “What’s the name of the student newspaper at the University of Pennsylvania?”

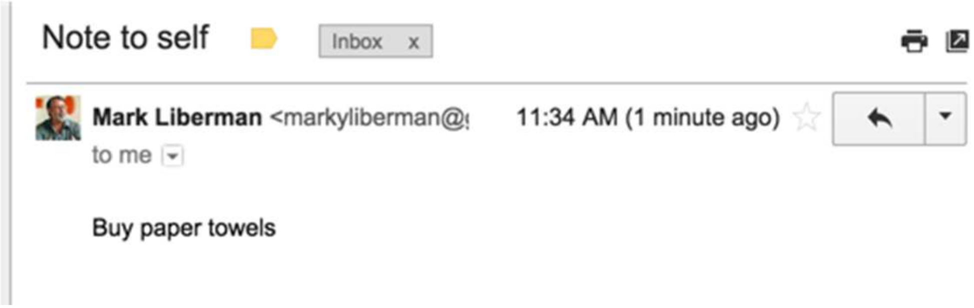
Transcribed: “What’s the name of the student newspaper at the University of Pennsylvania?

Answer: Page of search links, with The Daily Pennsylvanian at the top

Q: “Note to self – buy paper towels.”

Transcribed: “note to self buy paper towels”

Answer:



Note to self

Mark Liberman <markyliberman@

11:34 AM (1 minute ago)

to me

Buy paper towels

Question: “When was Hadley Wickham’s book ggplot2 published?”

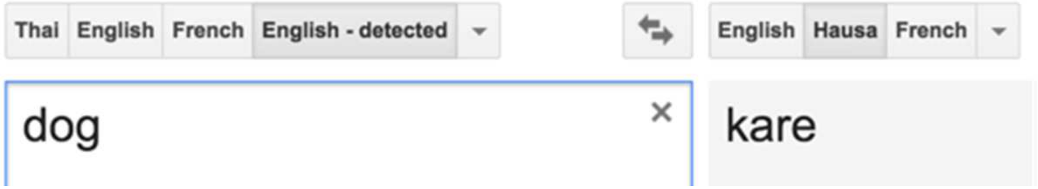
Transcribed: “when was Hadley Wickham zbook ggplot2 published”

Answer: Page of search results with the Amazon listing for ggplot2 at the top

Question: “What is the word for “dog” in Hausa?”

Transcribed: “what is the word for dog in hausa?”

Answer: “Here is your translation:”



dog

kare

Google Translate – from the Centre Cournot's web site:

Le Centre Cournot est une association soutenue par la Fondation Cournot, placée sous l'égide de la Fondation de France. Elle porte le nom du mathématicien et philosophe **franc-comtois** Augustin Cournot (1801-1877), reconnu de longue date comme un pionnier de **la discipline économique**.

The Cournot Centre is an association supported by the Cournot Foundation, under the aegis of the Fondation de France. It is named after the mathematician and philosopher **Franche-Comte** Augustin Cournot (1801-1877), long recognized as a pioneer of **economic discipline**.

Le Centre n’est pas un laboratoire de recherche, il n’est pas non plus un centre de réflexion**. Il** jouit de l’indépendance singulière d’un catalyseur.

The Centre is not a research laboratory, it is not a think tank. **He** enjoys the singular independence of a catalyst.

Pour qu’un débat ait lieu, il faut plus que de la connaissance et de la compréhension. Il faut des préférences, des croyances, des désirs, des objectifs… **C’est en pratique de cela seulement dont les débatteurs disposent** et ils inventent ou ils adoptent les résultats qui leur conviennent.

To have a debate, it takes more than knowledge and understanding. It takes preferences, beliefs, desires, goals ... **In practice this only with the debaters have** and they invent or they adopt the results that suit them.

From Yasmina Khadra, *Le Dingue au Bistouri*, 2013:

> Il y a quatre choses que je déteste.
> Un: qu'on boive dans mon verre.
> Deux: qu'on se mouche dans un restaurant.
> Trois: qu'on me pose un lapin.
> […]

Google Translate:

> There are four things I hate.
> A: we drink in my glass.
> Two: we will fly in a restaurant.
> Three: I get asked a rabbit.
> […]

In the interests of fairness, let’s give Bing Translator a shot:

Il y a quatre choses que je déteste.
Un: qu'on boive dans mon verre.
Deux: qu'on se mouche dans un restaurant.
Trois: qu'on me pose un lapin.
[…]

There are four things that I hate.
One: that one drink in my glass.
Two: what we fly in a restaurant.
Three: only asked me a rabbit.
[…]

So today, HLT (almost) works.

To what do we owe this gift?

Reason #1:

A digital shadow universe

increasingly mirrors real life

in flows and stores of bits.

Society is mostly about communication.

And most communication is text

(or talk, which is just text in fancy calligraphy)

. . . more and more often in digital form.

Simple properties of text

(like the words that make it up)

are a good proxy for content.

Better than anything else we have, anyhow…

Bigger faster cheaper digital everything

(and better programming languages, and . . .)

make it easier and easier

to pull content out of the flows of text

in that digital shadow universe.

There’s an old argument
about whether “Content is King”
or “Communication is King”.

But “the content of communication”
is at least the power behind the throne.

So in that new evolutionary niche:

a host of newly-evolving life forms
have got means, motive, and opportunity
to live off of these flows and stores of text

. . . while adding their digestion products
to the ecosystem.

Reason #2 that HLT (almost) works:

Advances in “Machine Learning”

(i.e. applied statistics)

…and the computer power to apply them

But there's another reason HLT (almost) works today

– a reason that's probably more important than
the new digital ecosystem
or the new machine learning methods –

It's a cultural change that took place half a century ago

. . . and the rest of this talk tells the story.

This talk is based on a presentation to the workshop

“Statistical Challenges
in Assessing and Fostering the Reproducibility of Scientific Results”

Committee on Applied and Theoretical Statistics (CATS),
Board on Mathematical Sciences and their Applications,
National Academy of Sciences

February 26-27, 2015

The NAS reproducibility workshop was alarming –

There's a crisis of credibility
in many areas of scientific research,
as documented elsewhere before and since:

John Ioannidis, "Why Most Published Research Findings Are False",
PLoS Medicine 8/30/2005.

"Amid a Sea of False Findings, the NIH Tries Reform",
Chronicle of Higher Education 3/16/2015:

ALS researchers, seeking a cure for Lou Gehrig's disease, went back and reproduced studies on more than 70 promising drugs. They found no real effects.

"Zero of those were replicable," Dr. [Francis] Collins said. "Zero. And a couple of them had already moved into human clinical trials …"

Today I’ll tell the story
of a crisis of credibility
that afflicted a different research area,
half a century ago.

Once upon a time. . .

there was a Bell Labs executive named John Pierce.

He supervised the team that built the first transistor,
and oversaw development
of the first communications satellite.

Credibility was not a problem for him.

N ASSOCIATES
NMR SPECTROMETER
PALO ALTO CALIFORNIA

In 1966, John Pierce chaired the
“Automatic Language Processing Advisory Committee” (ALPAC)
which produced a report to the National Academy of Sciences,
Language and Machines: Computers in Translation and Linguistics

And in 1969,
he wrote a letter to the Journal of the Acoustical Society of America,
published under the title *Whither Speech Recognition?*

The ALPAC Report

ALPAC noted that MT in 1966 was not very good, and suggested diplomatically that

“The Committee cannot judge what the total annual expenditure for research and development toward improving translation should be. However, it should be spent hardheadedly toward important, realistic, and relatively short-range goals.”

The committee felt that science should precede engineering in such cases:

“We see that the computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe these can be aptly compared with the challenges, problems, and insights of particle physics. Certainly, language is second to no phenomenon in importance. And the tools of computational linguistics are considerably less costly than the multibillion-volt accelerators of particle physics. The new linguistics presents an attractive as well as an extremely important challenge.”

Funders read between the lines, and U.S. MT funding went to zero for more than 20 years.

John Pierce’s views
about automatic speech recognition
were similar to his opinions about MT.

And his 1969 letter to JASA,
expressing his personal opinion,
was much less diplomatic
than that 1966 N.A.S. committee report….

“Whither Speech Recognition?”

“… a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English.”

“Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve ‘the problem.’ The basis for this is either individual inspiration (the ‘mad inventor’ source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach). . . .”

“The typical recognizer ... builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. **No simple, clear, sure knowledge is gained.** The work has been an experience, not an experiment.”

Tell us what you really think, John

“We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn’t attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%.
To sell suckers, one uses deceit and offers glamor.”

“It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds.
Thus, we may pity workers whom we cannot respect.”

Fallout from these blasts

The first idea: **Try Artificial Intelligence . . .**

DARPA Speech Understanding Research Project (1972-75)

Used classical AI to try to "understand what is being said with something of the facility of a native speaker"

DARPA SUR was viewed as a failure; funding was cut off after three years

The second idea: **Give Up.**

1975-1986: No U.S. research funding for MT or ASR

Pierce was far from the only person
with a jaundiced view of R&D investment
in the area of human language technology.

By the mid 1980s,
many informed American research managers
were equally skeptical about the prospects.

At the same time,
many people believed that HLT was needed
and in principle was feasible.

1985: Should DARPA restart HLT?

Charles Wayne, on loan to DARPA from the NSA, had an idea.

He'll design a speech recognition research program that

- protects against "glamour and deceit"

  because there is a well-defined, objective evaluation metric

  applied by a neutral agent (NIST)

  on shared data sets;
- and ensures that "simple, clear, sure knowledge is gained"

  because participants must reveal their methods

  to the sponsor and to one another

  at the time that the evaluation results are presented

Needed: Published data and well-defined metrics

David Pallett, "Performance Assessment of Automatic Speech Recognizers”, *J. of Research of the National Bureau of Standards*, 1985:

Definitive tests to fully characterize automatic speech recognizer or system performance cannot be specified at present. However, it is possible to design and conduct performance assessment tests that make use of widely available speech data bases, use test procedures similar to those used by others, and that are well documented. These tests provide valuable benchmark data and informative, though limited, predictive power.

By contrast, tests that make use of speech data bases that are not made available to others and for which the test procedures and results are poorly documented provide little objective information on system performance.

“Common Task” structure

- A detailed task definition and “evaluation plan”
  developed in consultation with researchers
  and published as the first step in the project.
- Automatic evaluation software
  written and maintained by NIST
  and published at the start of the project.
- **Shared data:**
  Training and “dev(elopment) test” data
  is published at start of project;
  “eval(uation) test” data is withheld
  for periodic public evaluations

Not everyone liked it

Many Piercians were skeptical:
“You can’t turn water into gasoline,
no matter what you measure.”

Many researchers were disgruntled:
“It’s like being in first grade again --
you’re told exactly what to do,
and then you’re tested over and over .”

But it worked.

Why did it work?

1. The obvious: it allowed funding to start

   (because the projects were glamour-and-deceit-proof)

   and to continue

   (because funders could measure progress over time)

Why did it work?

2. Less obvious: it allowed project-internal hill climbing
    - because the evaluation metrics were automatic
    - and the evaluation code was public

This obvious way of working was a new idea to many!

… and researchers who had objected to be tested twice a year began testing themselves every hour…

Why did it work?

3. Even less obvious: it created a culture
   (because researchers shared methods and results
   on shared data with a common metric)

**Participation in this culture became so valuable
that many research groups joined without funding**

What else it did

The *common task method* created a positive feedback loop.

When everyone's program has to interpret the same ambiguous evidence,
ambiguity resolution becomes a sort of gambling game,
which rewards the use of statistical methods,
and has led to the flowering of "machine learning".

Given the nature of speech and language,
statistical methods need the largest possible training set,
which reinforces the value of shared data.

Iterated train-and-test cycles on this gambling game are addictive;
they create "simple, clear, sure knowledge",
which motivates participation in the common-task culture.

The “Common Task Method”

. . . has become the standard research paradigm
in experimental computational science:

- Published training and testing data
- Well-defined evaluation metrics
- Techniques to avoid over-fitting
  (managerial as well as statistical)

Domain: ***Algorithmic analysis of the natural world.***

Over the past 30 years, variants of this method have been applied to many other problems:

machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, OCR, sentiment analysis, image analysis, video analysis, … , etc.

The general experience:

1. Error rates decline by a fixed percentage each year,
   to an asymptote depending on task and data quality
2. Progress usually comes from many small improvements;
   improvement by 1% is a reason to break out the champagne.
3. Shared data plays a crucial role – and is re-used in unexpected ways.
4. Glamour and deceit have mostly been avoided.

There are dozens of current examples –

Some of them are shared-task workshops:

Conference on Natural Language Learning Shared Task for 2015 (CoNLL2015)
Open Keyword Search Evaluation 2015 (OpenKWS2015)
Open Machine Translation Evaluation (OpenMT2014, OpenMT2016)
Reconnaissance de Personnes dans les Emissions Audiovisuelles (REPERE2014)
Speaker Recognition Evaluation (SRE2014, SRE2016)
Text Retrieval Conference (TREC2015)
DiscoMT 2015 Shared Task on Pronoun Translation
TREC Video Retrieval Evaluation (TRECVID2015)
IMAGENET Large Scale Visual Recognition Challenge 2015
. . . and many, many others . . .

Some are just shared datasets and evaluation metrics.

For example, TAC 2014:

“The Text Analysis Conference (TAC) is a series of evaluation workshops organized to encourage research in Natural Language Processing and related applications, by providing a large test collection, common evaluation procedures, and a forum for organizations to share their results.”

“TAC comprises sets of tasks known as ‘tracks,’ each of which focuses on a particular subproblem of NLP. TAC tracks focus on end-user tasks, but also include component evaluations situated within the context of end-user tasks.”

TAC 2014 hosts evaluations in two areas of research:

“**Knowledge Base Population** (KBP): The goal of Knowledge Base Population is to promote research in automated systems that discover information about named entities as found in a large corpus and incorporate this information into a knowledge base. “

“**Biomedical Summarization** (BiomedSumm): The goal of BiomedSumm is to develop technologies that aid in the summarization of biomedical literature.”

Or the CoNLL Shared Task for 2015:

“Since the first CoNLL Shared Task on NP chunking in 1999, CoNLL shared tasks over the years have tackled increasingly complex natural language learning tasks. Early shared tasks focused on identifying text chunks or named entities that typically correspond to single words or short phrases within a sentence. Shared tasks on semantic role labeling are concerned with identifying arguments for individual predicates and characterizing the relationship between each argument and the predicate. Shared tasks on joint dependency parsing and semantic role labeling target the syntactic and semantic structure of the entire sentence, rather than the argument structure of individual predicates. More recently, shared tasks on coreference went beyond sentence boundaries and started to deal with discourse phenomena, and shared tasks on grammatical error correction dealt with detecting and correcting grammatical errors in texts.”

Or TRECVID 2015:

“The main goal of the TREC Video Retrieval Evaluation (TRECVID) is to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation.

In TRECVID 2015 NIST will continue 4 of the 2014 tasks with some revisions […], drop one […], separate out the localization task from semantic indexing, and add a new Video Hyperlinking task previously run in MediaEval:

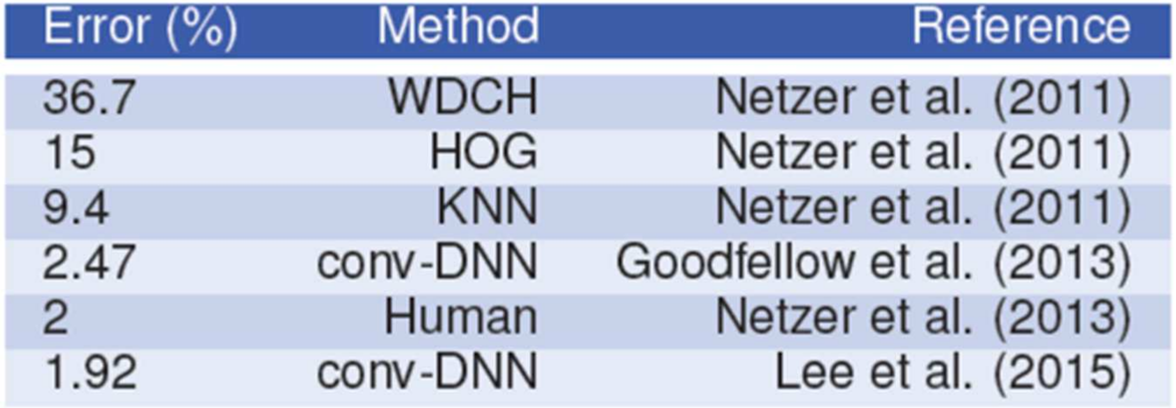
Semantic indexing [IACC]
Interactive surveillance event detection [i-LIDS]
Instance search [BBC EastEnders]
Multimedia event detection [HAVIC]
Localization [IACC]
Video Hyperlinking [BBC for Hyperlinking]

Or the Street View House Numbers (SVHN) dataset:

“SVHN is a real-world image dataset for developing machine learning and object recognition algorithms with minimal requirement on data preprocessing and formatting. It can be seen as similar in flavor to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude more labeled data (over 600,000 digit images) and comes from a significantly harder, unsolved, real world problem (recognizing digits and numbers in natural scene images). SVHN is obtained from house numbers in Google Street View images.”

73257 digits for training,
26032 digits for testing,
and 531131 additional samples to use as extra training data.

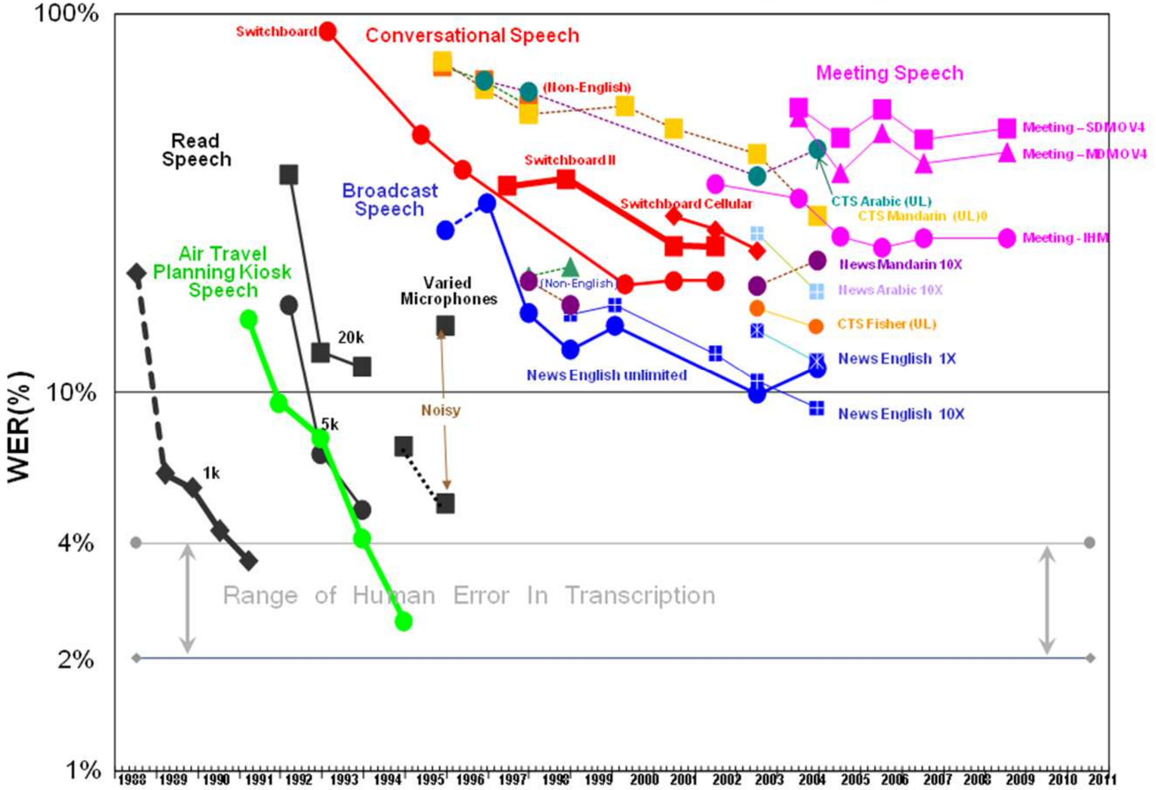
Progress in performance on SVHN:



Error (%)	Method	Reference
36.7	WDCH	Netzer et al. (2011)
15	HOG	Netzer et al. (2011)
9.4	KNN	Netzer et al. (2011)
2.47	conv-DNN	Goodfellow et al. (2013)
2	Human	Netzer et al. (2013)
1.92	conv-DNN	Lee et al. (2015)

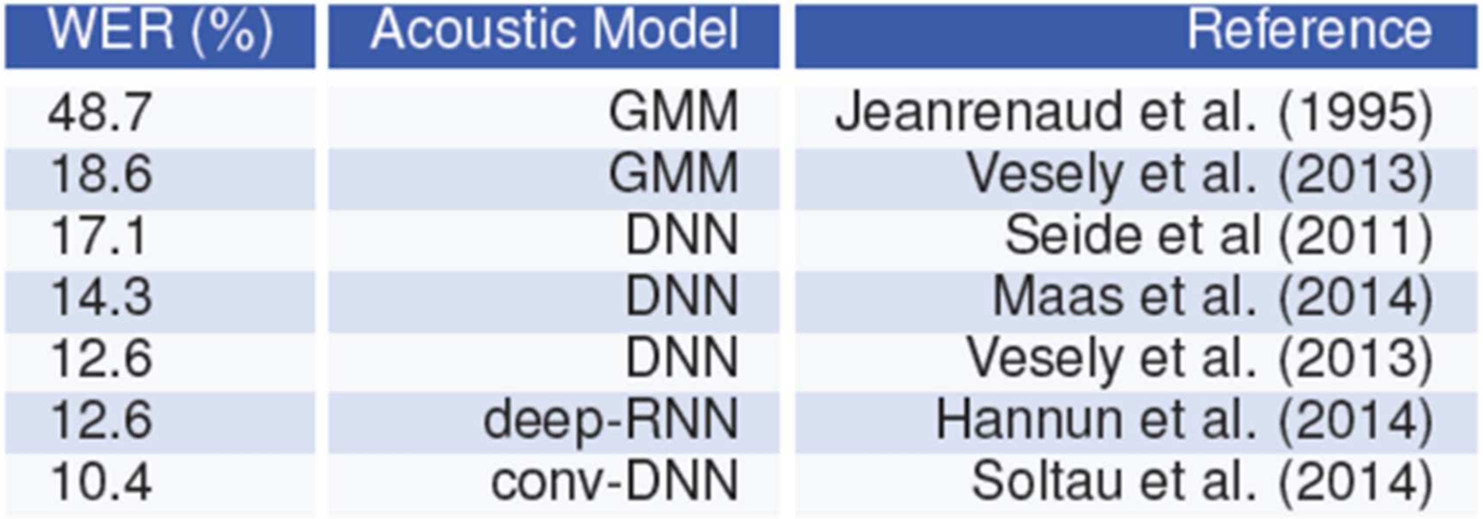
Progress is not always this rapid –
but steady progress almost always happens.

NIST STT Benchmark Test History – May. '09



Continued progress in Speech-to-Text –

Switchboard Corpus of conversational telephone speech,
Stalled at 20-30% word error rate 15 years ago:



WER (%)	Acoustic Model	Reference
48.7	GMM	Jeanrenaud et al. (1995)
18.6	GMM	Vesely et al. (2013)
17.1	DNN	Seide et al (2011)
14.3	DNN	Maas et al. (2014)
12.6	DNN	Vesely et al. (2013)
12.6	deep-RNN	Hannun et al. (2014)
10.4	conv-DNN	Soltau et al. (2014)

Where we were:

ANLP-1983
(First Conference on Applied Natural Language Processing)

34 Presentations:

None used a published data set.
None used a formal evaluation metric.

A more recent sample:

ACL-2010
(48th Annual Meeting of the Association for Computational Linguistics)

274 presentations –
All use published data and published evaluation methods.
(A few describe new data-set creation and/or new evaluation metrics.)

Three random examples from ACL-2010:

Nils Reiter and Anette Frank, "Identifying Generic Noun Phrases”.
Authors are from Heidelberg University; use ACE-2 data.

Shih-Hsiang Lin and Berlin Chen,
"A Risk Minimization Framework for Extractive Speech Summarization”.
Authors are from National Taiwan University;
use Academia Sinica Broadcast News Corpus
and the ROUGE metric (developed in DUC summarization track).

Laura Chiticariu et al., ”An Algebraic Approach to Declarative Information Extraction”.
Authors are from IBM Research; use ACE NER metric, ACE data, ENRON corpus data

Science is different…

But not that different.

Sharing data and problems
lowers costs and barriers to entry
creates intellectual communities
speeds up replication and extension
and guards against “glamour and deceit”
(…as well as simple confusion)

There are some similar scientific initiatives

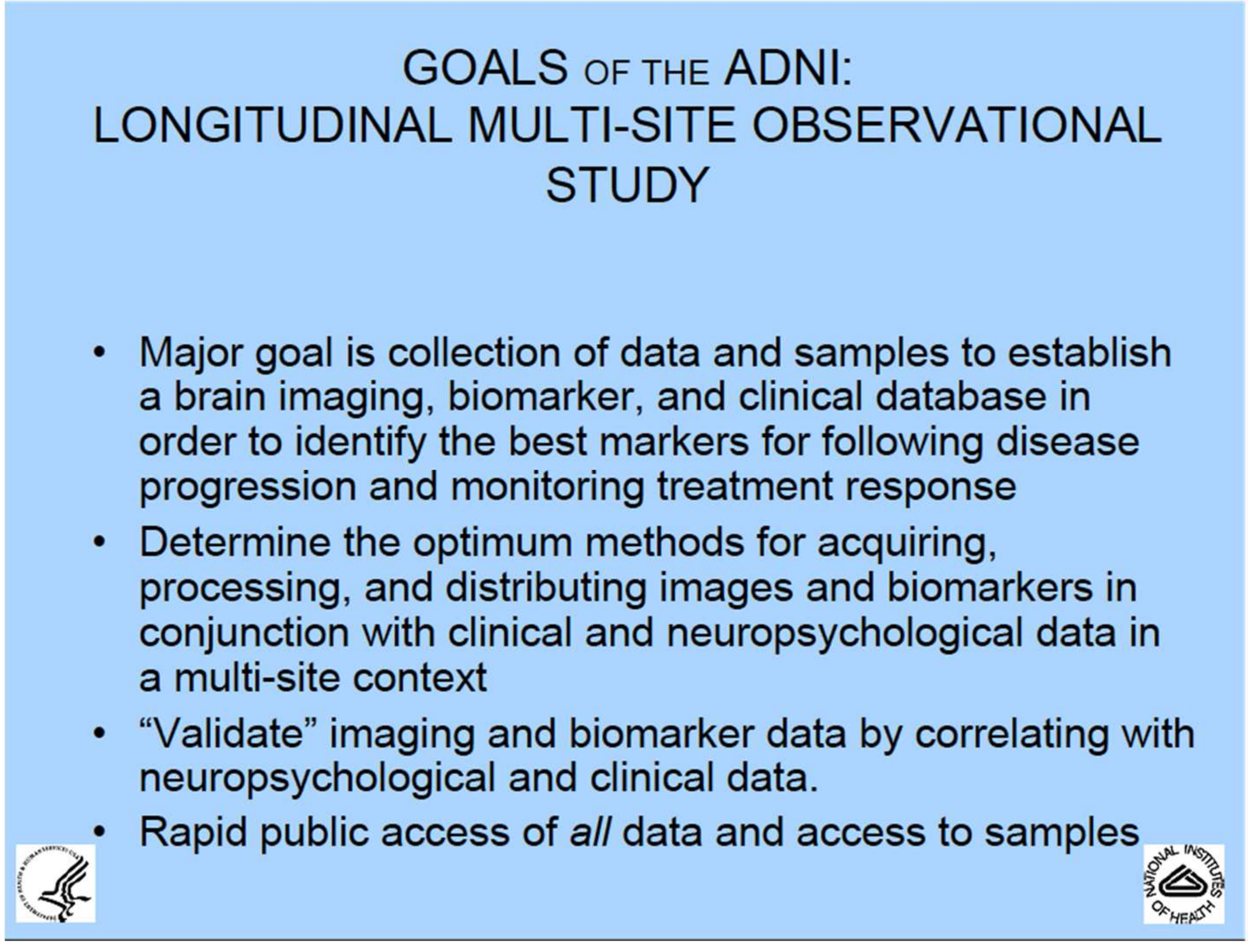
e.g. the **Alzheimer's Disease Neuroimaging Initiative** (ADNI)

organized in 2004 by NIH

According to Neil Buckholtz (National Institute on Aging)

"Transforming Research through Open Access to Discovery Inputs and Outputs"

Berlin 9 Open Access Conference, HHMI, November 2011



GOALS OF THE ADNI: LONGITUDINAL MULTI-SITE OBSERVATIONAL STUDY

- Major goal is collection of data and samples to establish a brain imaging, biomarker, and clinical database in order to identify the best markers for following disease progression and monitoring treatment response
- Determine the optimum methods for acquiring, processing, and distributing images and biomarkers in conjunction with clinical and neuropsychological data in a multi-site context
- “Validate” imaging and biomarker data by correlating with neuropsychological and clinical data.
- Rapid public access of *all* data and access to samples

BUT

- No well-defined versioning of datasets
- No evaluation metric
- No focused workshops

Predicting the time course of Alzheimer’s Disease
is exactly the kind of problem
(“algorithmic analysis of the natural world”)
for which the Common Task method seems to work.

So should we apply such methods
to the large class of similar biomedical problems?

Scientists would mostly be horrified –
But in the face of the reproducibility crisis,
perhaps we should consider it.

Thank you!