

# CSE 599 Empirical Foundations of Machine Learning

University of Washington, Autumn 2021

# Plan for today

Course logistics: project proposal & discussion sessions

Benchmarking history

Applied statistics: confidence intervals for population accuracy

# Plan for today

Course logistics: project proposal & discussion sessions

Benchmarking history

Applied statistics: confidence intervals for population accuracy

# Course projects

Deadline for proposals: one week from now (Thursday October 14)

1 - 2 pages + sketches of main plots

**Today** office hours: 12:30 - 13:30 in CSE2 (Gates) 214 (Ludwig's office)

## Introductions

- Name
- Program & what year
- Research interests & advisor
- ML classes taken
- Project ideas

# Discussion sessions

Format: **Role-Playing Paper-Reading Seminars**

Introduced by Alec Jacobson and Colin Raffel

<https://colinraffel.com/blog/role-playing-seminar.html>

Key idea: spread out paper presentation into many distinct sub-tasks

- Increase participation in discussions
- Reduce load on presenter
- In aggregate deeper engagement with paper from more perspectives

# Roles



**Scientific Peer Reviewer.** The paper has not been published yet and is currently submitted to a top conference where you've been assigned as a peer reviewer. Complete a full review of the paper answering all prompts of the official review form of the top venue in this research area (e.g., *NeurIPS* for Deep Learning and *ACM SIGGRAPH* for Geometry & Animation). This includes recommending whether to accept or reject the paper.



**Archaeologist.** This paper was found buried under ground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work. Find and report on one *older* paper cited within the current paper that substantially influenced the current paper and one *newer* paper that cites this current paper.

# Roles



**Academic Researcher.** You're a researcher who is working on a new project in this area. Propose an imaginary follow-up project *not just* based on the current but only possible due to the existence and success of the current paper.



**Industry Practitioner.** You work at a company or organization developing an application or product of your choice (that has not already been suggested in a prior session). Bring a convincing pitch for why you should be paid to implement the method in the paper, and discuss at least one positive and negative impact of this application.

# Roles



**Hacker.** You're a hacker who needs a demo of this paper ASAP. Implement a small part or simplified version of the paper on a small dataset or toy problem. Prepare to share the core code of the algorithm to the class and demo your implementation. Do not simply download and run an existing implementation – though you are welcome to use (and give credit to) an existing implementation for “backbone” code.



**Private Investigator.** You are a detective who needs to run a background check on one of the paper's authors. Where have they worked? What did they study? What previous projects might have led to working on this one? What motivated them to work on this project? Feel free to contact the authors, but remember to be courteous, polite, and on-topic.

# Roles



**Social Impact Assessor.** Identify how this paper self-assesses its (likely positive) impact on the world. Have any additional positive social impacts left out? What are possible negative social impacts that were overlooked or omitted?

We'll have two more roles

**Connector:** connect & contrast papers if we read more than one

**Summarizer:** Summarize key points of the paper as intro for discussion

# Discussion logistics

One week before the discussion session: we **randomly assign** people to roles

Everyone enrolled in the class is automatically assigned

Everyone else is still **strongly** encouraged to participate → message Mitchell

Everyone is assigned to a role → we will have **multiple people per role**

Each role team decides **presenter** for a given week (load-balance over the quarter)

The day before the discussion session, each role sends **PDF slides** to us

Each role will have **10 min**, usually with a 5 min presentation, 5 min discussion split

Some weeks will have **special instructions** for some of the roles

# First discussion session

**Warm-up:** We'll read two classical papers to get used to the format

Thursday October 14 (a week from now)

You can **vote** what we read!

**Option A:** ImageNet dataset paper (2009)  
ImageNet competition retrospective paper (2015)

**Option B:** AlexNet (2012)  
ResNet (2015)



# Tools channel on Mattermost

Post your favorite Python package, command-line tool, etc.!

## Announcing Streamlit 1.0! 🎉

Streamlit used to be the simplest way to write data apps. Now it's the most powerful.

By Adrien Treuille  
Posted in Announcement, October 5 2021

The image shows a purple-bordered frame containing a cluster of colorful balloons (red, green, blue, orange) floating against a white background. The word "Streamlit" is written in red at the bottom left of the balloon cluster.

# Plan for today

Course logistics: project proposal & discussion sessions

Benchmarking history

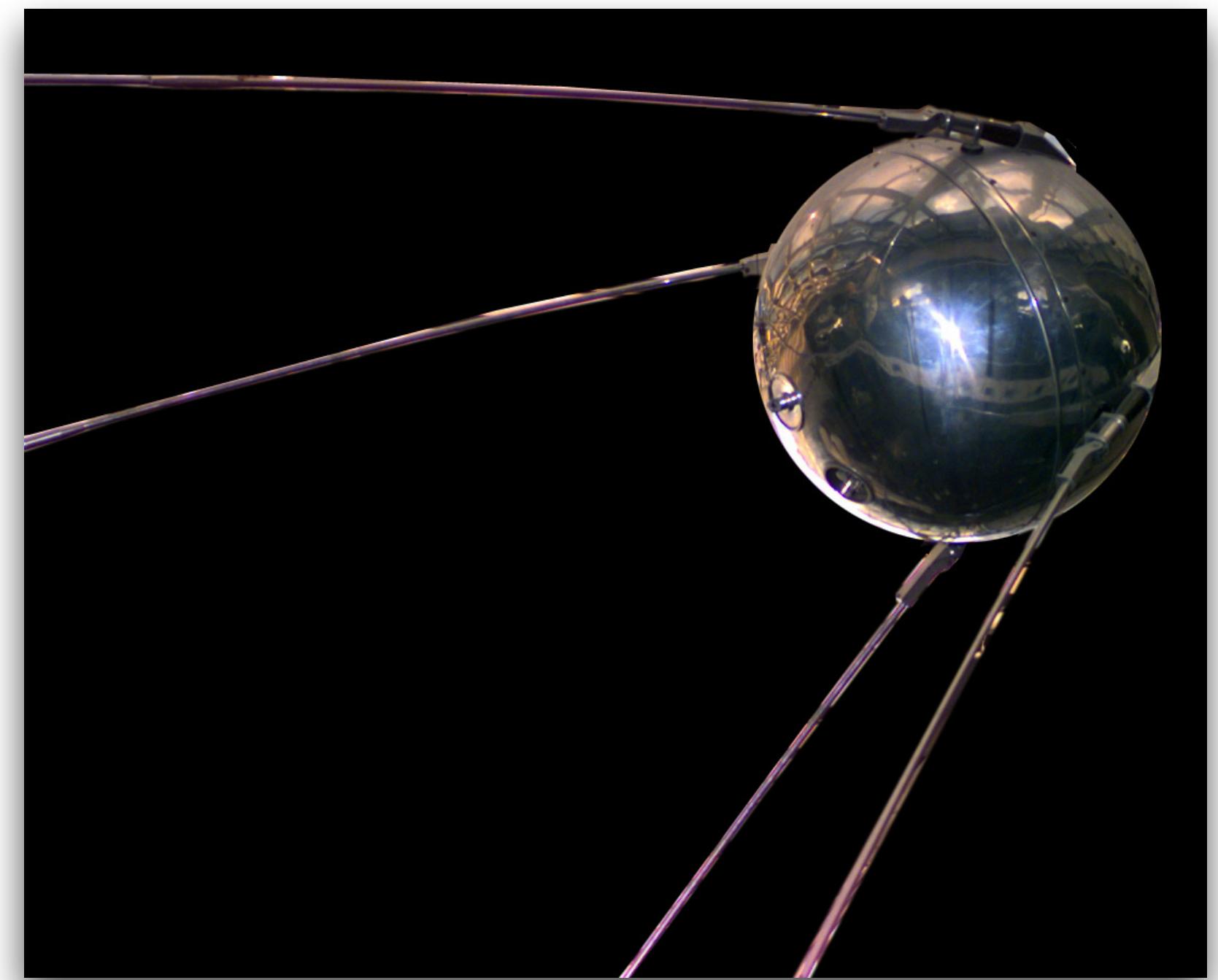
Applied statistics: confidence intervals for population accuracy

# History of Benchmarking in ML

1960s: large investments in science and technology  
(Result of Sputnik, etc.)

Speech recognition and translation get a lot of attention,  
are glamorous fields, and attract funding.

But **results are lacking**



# John R. Pierce (1910 - 2002)

Director of research at Bell Labs

Co-invented **pulse code modulation**, managed the team that invented the **transistor** (and invented the name), led development of first commercial **communications satellite**, etc.



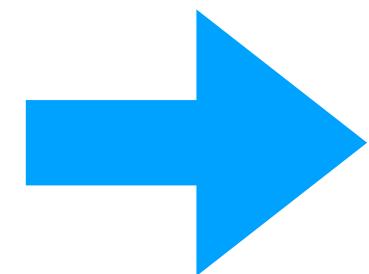
Did not like AI and wrote about it in the ALPAC report and “Whither Speech Recognition?”

# ALPAC Report (1964 - 1966)

**Automatic Language Processing Advisory Committee:** 7 researchers led by Pierce

Established by the US government to evaluate potential of machine translation for various government agencies (mostly defense / science focused (Russian journals)).

**Negative conclusions** for machine translation, recommends investment in computational linguistics instead



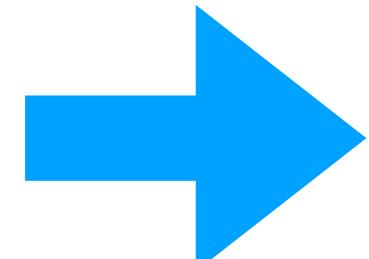
No government funding for machine translation for 10 - 20 years

# “Whither Speech Recognition?” (1969)

Again John Pierce, this time a single-author short 1.5 page letter to the Journal of the Acoustical Society of America

**Very critical** of speech recognition research

*“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 glamour.**”*



No funding for speech recognition for 10 - 20 years

# Quote from “Whither Speech Recognition?”

*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.***

# Quote from “Whither Speech Recognition?”

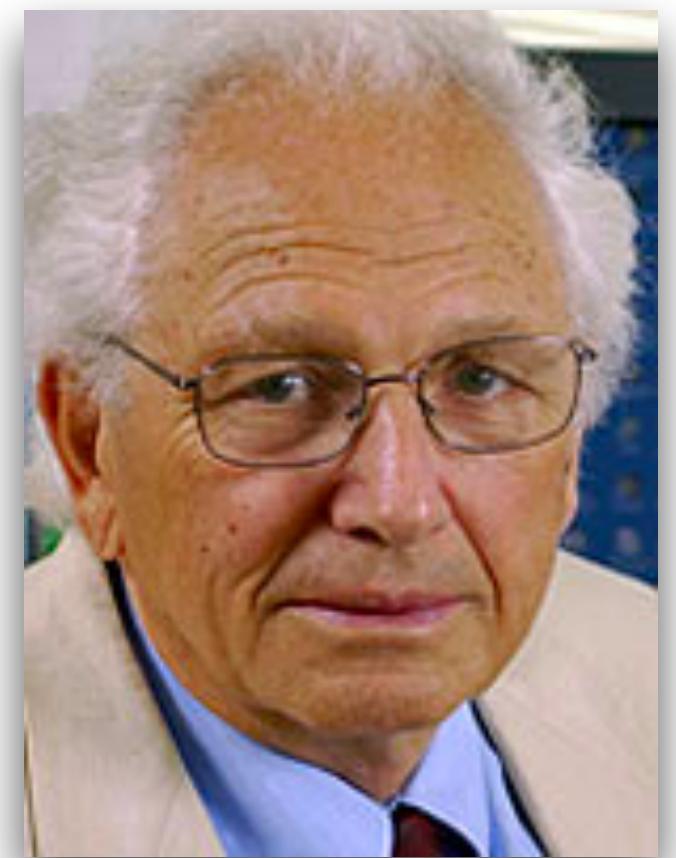
*It is clear that **glamor and any deceit** in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. What particular considerations have led to this enthusiasm? [...]*

*Turing asked, On what basis can we say that a machine thinks? His perfectly rational answer was that if, in conversing with a machine, we cannot tell whether it is a human being or a machine, then we can scarcely deny that the machine thinks. [...]*

*We should consider, however, that **in deception, studied and artful deceit is apt to succeed better and more quickly than science.***

# Bringing Funding for Translation and Speech Recognition Back

Two people were key in resuming government funding for speech and translation in the mid to late 80s:



**Fred Jelinek:** research manager at IBM



**Charles Wayne:** program manager at DARPA



**Key idea:** make evaluations “glamour and deceit”-proof

# Fred Jelinek



PhD in information theory (Fano)

Led IBM's effort on the "**general dictation problem**" from 1972 to 1980

Advocate for comparing the **quantitative performance** of alternative algorithms on **test sets**, using fixed and automatically calculated **evaluation metrics**.

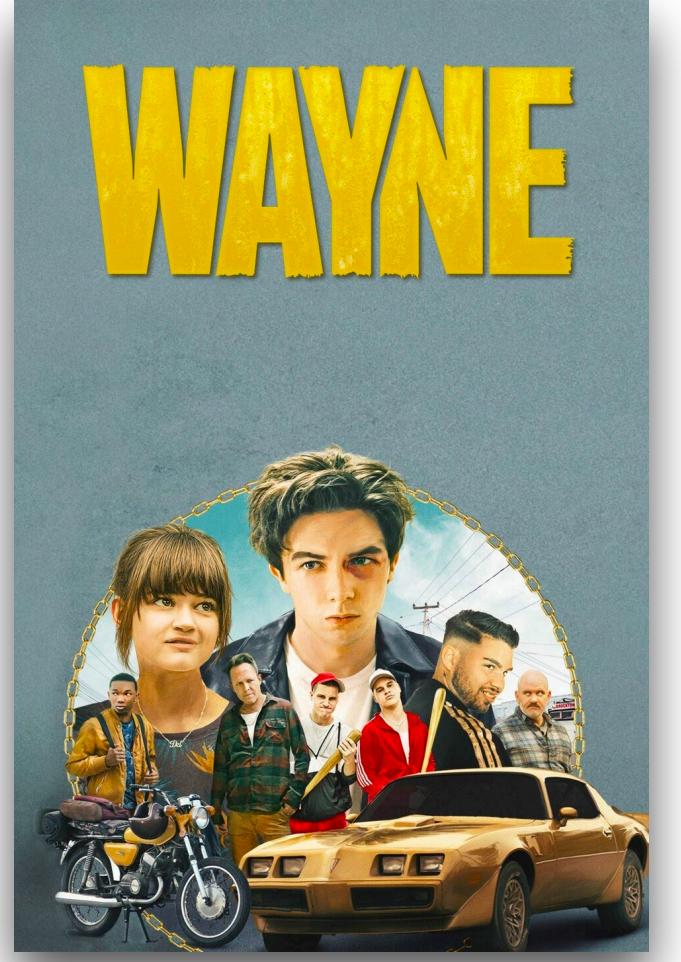
Also strongly in favor of **sharing datasets, evaluation metric, algorithms**, etc.

Same approach for machine translation and other problems in his group.

*"Every time I fire a linguist, the performance of the speech recognizer goes up."*

# Charles Wayne

DARPA program manager responsible for funding restart in 1986



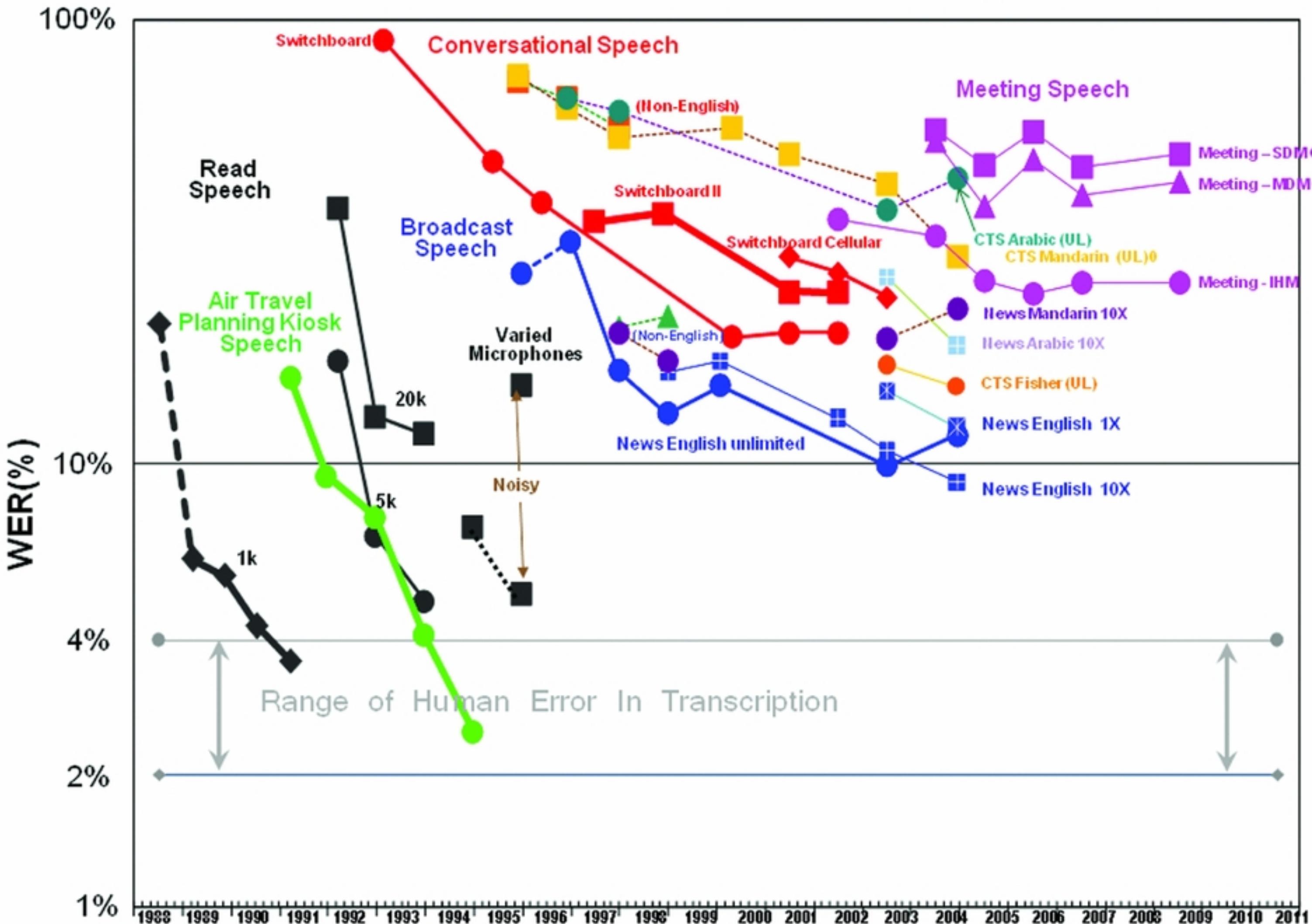
Key idea: emphasize evaluation. **Well-defined objective** evaluation, applied by a **neutral agent** (NIST) on **shared datasets** (often Linguistic Data Consortium)

Initially both Pierce-style engineers and speech researchers were skeptical, but the approach was successful

**“Glamour and deceit”-proof**, funders could measure progress

# Speech Recognition Benchmarks

NIST STT Benchmark Test History – May. '09



Also in 1987:

David Aha creates the **UCI dataset repository**

→ ML community adopts benchmark paradigm

Key benchmark before ImageNet: **PASCAL VOC**

# Plan for today

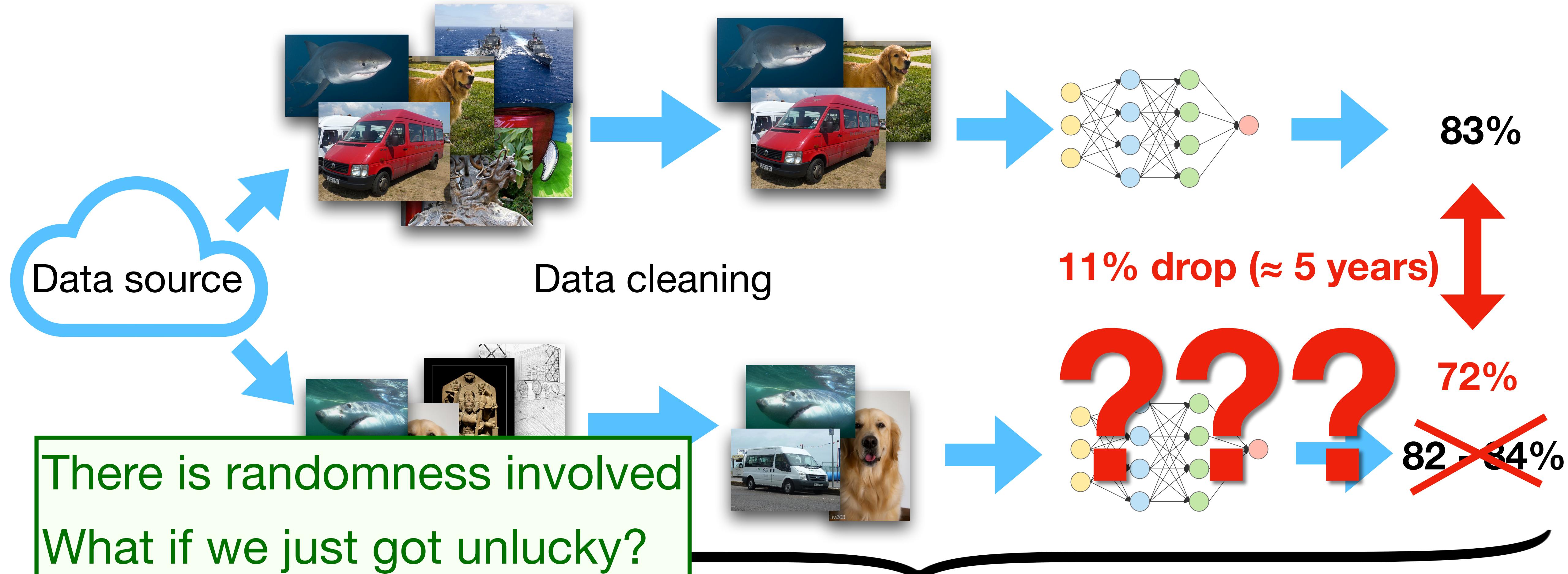
Course logistics: project proposal & discussion sessions

Benchmarking history

Applied statistics: confidence intervals for population accuracy

# Main goal in ML: generalization

At least, the classifiers should perform similarly well on new data from the **same source**.



# Two Possible Causes

$$\widehat{\text{acc}}_S(f) - \widehat{\text{acc}}_{S'}(f) = \cancel{\widehat{\text{acc}}_S(f)} - \cancel{\widehat{\text{acc}}_D(f)} + \text{acc}_D(f) - \text{acc}_{D'}(f) + \text{acc}_{D'}(f) - \widehat{\text{acc}}_{S'}(f)$$

New test accuracy

Overfitting through test set re-use ( $\approx 0\%$ )

Distribution shift

Generalization error ( $\approx 1\%$ )

Original test accuracy (orig. test set S, new S')

$\widehat{\text{acc}}_S(f) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}[f(x) = y]$

$\text{acc}_D(f) = \mathbb{E}_{(x,y) \sim D} \mathbb{1}[f(x) = y] \quad (\text{S is drawn from D})$

$\approx 11\%$

[Submitted on 13 Oct 2020]

# With Little Power Comes Great Responsibility

[Dallas Card](#), [Peter Henderson](#), [Urvashi Khandelwal](#), [Robin Jia](#), [Kyle Mahowald](#), [Dan Jurafsky](#)

Despite its importance to experimental design, statistical power (the probability that, given a real effect, an experiment will reject the null hypothesis) has largely been ignored by the NLP community. Underpowered experiments make it more difficult to discern the difference between statistical noise and meaningful model improvements, and increase the chances of exaggerated findings. By meta-analyzing a set of existing NLP papers and datasets, we characterize typical power for a variety of settings and conclude that underpowered experiments are common in the NLP literature. In particular, for several tasks in the popular GLUE benchmark, small test sets mean that most attempted comparisons to state of the art models will not be adequately powered. Similarly, based on reasonable assumptions, we find that the most typical experimental design for human rating studies will be underpowered to detect small model differences, of the sort that are frequently studied. For machine translation, we find that typical test sets of 2000 sentences have approximately 75% power to detect differences of 1 BLEU point. To improve the situation going forward, we give an overview of best practices for power analysis in NLP and release a series of notebooks to assist with future power analyses.

[Submitted on 30 Aug 2021 ([v1](#)), last revised 5 Oct 2021 (this version, v2)]

# Deep Reinforcement Learning at the Edge of the Statistical Precipice

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, Marc G. Bellemare

Deep reinforcement learning (RL) algorithms are predominantly evaluated by comparing their relative performance on a large suite of tasks. Most published results on deep RL benchmarks compare point estimates of aggregate performance such as mean and median scores across tasks, ignoring the statistical uncertainty implied by the use of a finite number of training runs. Beginning with the Arcade Learning Environment (ALE), the shift towards computationally-demanding benchmarks has led to the practice of evaluating only a small number of runs per task, exacerbating the statistical uncertainty in point estimates. In this paper, we argue that reliable evaluation in the few run deep RL regime cannot ignore the uncertainty in results without running the risk of slowing down progress in the field. We illustrate this point using a case study on the Atari 100k benchmark, where we find substantial discrepancies between conclusions drawn from point estimates alone versus a more thorough statistical analysis. With the aim of increasing the field's confidence in reported results with a handful of runs, we advocate for reporting interval estimates of aggregate performance and propose performance profiles to account for the variability in results, as well as present more robust and efficient aggregate metrics, such as interquartile mean scores, to achieve small uncertainty in results. Using such statistical tools, we scrutinize performance evaluations of existing algorithms on other widely used RL benchmarks including the ALE, Procgen, and the DeepMind Control Suite, again revealing discrepancies in prior comparisons. Our findings call for a change in how we evaluate performance in deep RL, for which we present a more rigorous evaluation methodology, accompanied with an open-source library `rliable`, to prevent unreliable results from stagnating the field.

Comments: NeurIPS 2021 (Oral). Website: [this https URL](#)

# ML and statistical rigor

**Many ML papers do not even attempt to quantify statistical uncertainty**

This would be unacceptable in most other sciences

(Caveat: statistical analyses can easily be flawed or distract from other issues)

The extent to wh

**Computer vision**

→ Simple a

Other fields (e.g.

→ Performa

The screenshot shows a news article from the journal 'nature'. The header includes the word 'nature' in a large serif font, followed by a navigation bar with links for 'Explore content', 'About the journal', 'Publish with us', and 'Subscribe'. Below the header, the breadcrumb navigation shows 'nature > comment > article'. The main title of the article is 'Scientists rise up against statistical significance' in large, bold, black font. Below the title, it says 'COMMENT | 20 March 2019'. A short summary at the bottom states: 'Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.'

# Recall: datasets for ImageNet transfer

Dataset	Classes	Size (train/test)	Accuracy metric
Food-101 [5]	101	75,750/25,250	top-1
CIFAR-10 [43]	10	50,000/10,000	top-1
CIFAR-100 [43]	100	50,000/10,000	top-1
Birdsnap [4]	500	47,386/2,443	top-1
SUN397 [84]	397	19,850/19,850	top-1
Stanford Cars [41]	196	8,144/8,041	top-1
FGVC Aircraft [55]	100	6,667/3,333	mean per-class
PASCAL VOC 2007 Cls. [22]	20	5,011/4,952	11-point mAP
Describable Textures (DTD) [10]	47	3,760/1,880	top-1
Oxford-IIIT Pets [61]	37	3,680/3,369	mean per-class
Caltech-101 [24]	102	3,060/6,084	mean per-class
Oxford 102 Flowers [59]	102	2,040/6,149	mean per-class

ImageNet itself has 50,000 test images

MS COCO has 80,000 test images

# NLP: SuperGLUE

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 <sub>a</sub> /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

# How do we get a handle on uncertainty from sampling?

Statistics has developed **a lot** of answers to these questions in the past 100 years.

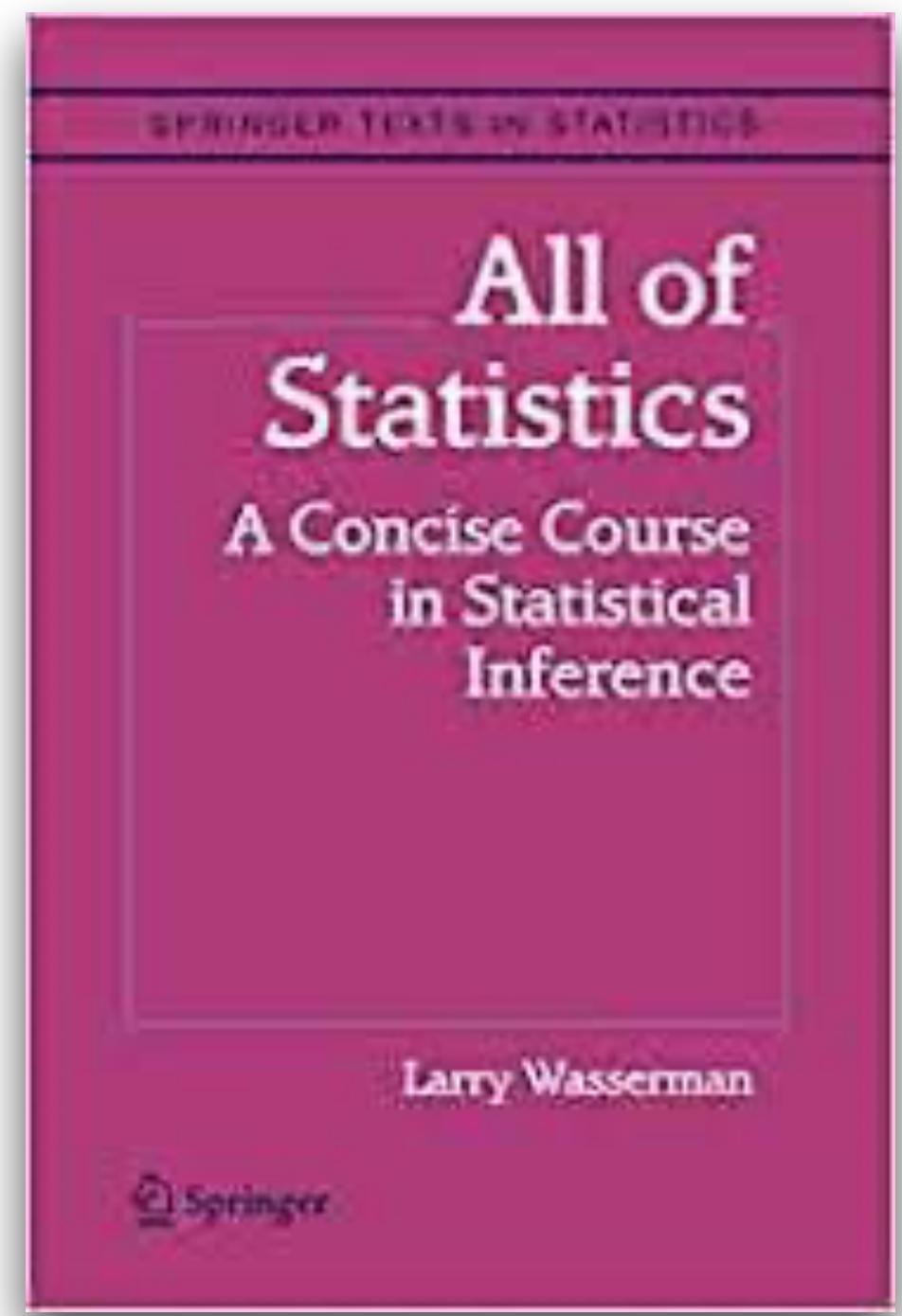
Standard answer:

## 6.3.2 Confidence Sets

A  $1 - \alpha$  **confidence interval** for a parameter  $\theta$  is an interval  $C_n = (a, b)$  where  $a = a(X_1, \dots, X_n)$  and  $b = b(X_1, \dots, X_n)$  are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta. \quad (6.9)$$

In words,  $(a, b)$  traps  $\theta$  with probability  $1 - \alpha$ . We call  $1 - \alpha$  the **coverage** of the confidence interval.



# Interpretation

**Warning!**  $C_n$  is random and  $\theta$  is fixed.

Commonly, people use 95 percent confidence intervals, which corresponds to choosing  $\alpha = 0.05$ . If  $\theta$  is a vector then we use a **confidence set** (such as a sphere or an ellipse) instead of an interval.

**Warning!** There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about  $\theta$  since  $\theta$  is a fixed quantity, not a random variable. Some texts interpret confidence intervals as follows: if I repeat the experiment over and over, the interval will contain the parameter 95 percent of the time. This is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this:

On day 1, you collect data and construct a 95 percent confidence interval for a parameter  $\theta_1$ . On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter  $\theta_2$ . On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter  $\theta_3$ . You continue this way constructing confidence intervals for a sequence of unrelated parameters  $\theta_1, \theta_2, \dots$ . Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

# Single sample interpretation

*"There is a 90% probability that the calculated confidence interval from some future experiment encompasses the true value of the population parameter."*

Probability statement about the confidence interval, not the population parameter.

Pre-experiment point of view: the experimenter sets out the way in which they intend to calculate a confidence interval and to know, before they do the actual experiment, that the interval they will end up calculating has a particular chance of covering the true but unknown value.

Similar to the "repeated sample" interpretation above, except that it avoids relying on considering hypothetical repeats of a sampling procedure that may not be repeatable in any meaningful sense.

Source: Wikipedia and Neyman's 1937 paper

# X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. NEYMAN

*Reader in Statistics, University College, London*

(Communicated by H. JEFFREYS, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

## CONTENTS

	Page
I—INTRODUCTORY . . . . .	333
(a) General Remarks, Notation, and Definitions . . . . .	333
(b) Review of the Solutions of the Problem of Estimation Advanced Hereto . . . . .	343
(c) Estimation by Unique Estimate and by Interval . . . . .	346
II—CONFIDENCE INTERVALS . . . . .	347
(a) Statement of the Problem . . . . .	347
(b) Solution of the Problem of Confidence Intervals . . . . .	350
(c) Example I . . . . .	356

# Confusions

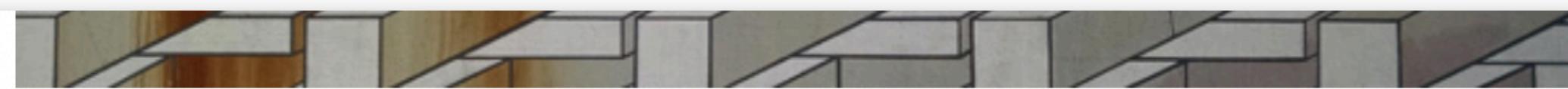
**A 95% confidence level does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval.**

Once an interval is calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability.

The 95% probability relates to the reliability of the estimation procedure, not to a specific calculated interval. Neyman himself (the original proponent of confidence intervals) made this point in his original paper:

Source: Wikipedia and Neyman's 1937 paper

*"It will be noticed that in the above description, the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to a. Consider now the case when a sample is already drawn, and the calculations have given [particular limits]. Can we say that in this particular case the probability of the true value [falling between these limits] is equal to a? The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made..."*



## Experiments as randomized algorithms

Sep 28, 2021. While every statistics course leads with how correlation does not imply causation, the methodological jump from observation to causal inference is small. Using the same algorithmic summarization and statistical analysis tools that we use to... [Continue](#)

---

## Statistics as algorithmic summarization

Sep 28, 2021. Though a multifaceted and complex discipline, Statistics' greatest contribution is a rigorous framework for summarization. Statistics gives us reasonable procedures to estimate properties of a general population by examining only a few individuals from the... [Continue](#)

---

## All statistical models are wrong. Are any useful?

Sep 21, 2021. Though I singled out a mask study in the last post, I've had a growing discomfort with statistical modeling and significance more generally. Statistical models explicitly describe the probability of outcomes of experiments in terms... [Continue](#)

---

## Effect size is significantly more important than statistical significance.

Sep 13, 2021. A massive cluster-randomized controlled trial run in Bangladesh to test the efficacy of mask wearing on reducing coronavirus transmission released its initial results and the covid pundits have been buzzing with excitement. There have already... [Continue](#)

Let's compute some probability bounds