ChaosTalk, 28 April 2026.

THERE'S NO DATA LIKE MORE DATA
The subject of Tuesday's talk is how language modeling and automated speech recognition developed into machine learning, and how language eventually became data.

The history has been collected in a forthcoming book Divination Engines , by Xiaochang Li of Stanford University; available from University of Chicago Press, July of this year.

Three partial presentations of professor Li's upcoming book were available for readings. Two were academic papers by Li, 'Abstractions And Embodiments' and 'There's no Data Like More Data' plus a YouTube panel discussion of the "Data" paper.
 The most complete, a 2023 Osirus article with the TNDLMD title covers the time from the 1970s through the 1980s into the 1990s.

This is the period IBM takes up automated speech recognition (ASR) and natural language modeling from work previously done in the 1950s by AT&T and the department of defense Advanced Research Projects Agency (ARPA).

The Cultural of the Institution Matters
 The AT&T research team focused on communication, persuing efficiencies in band width use and pruning the signal of redundant acoustic information.

 The IBM team would gradually shift the field away from an "expert systems" and knowledge based approach. Redirecting the goal from simulating human facilities, like phonetics and semantics, to a statistical data driven machine learning approach.

The IBM team was less concerned with efficiency. Developments in high speed processing was threatening to reduce their market. IBM Corporate needed to sustain customer demand for hardware and large scale computation. The research team also had ample access to in house hardware and compute-time to develop their projects. Creating new compute-intensive applications was in their interests. Speech recognition slowly morphed into a computational problem, solvable through simple models, powered by massive brute force processing.

Hidden Markov Models (HHMs)

The key concepts enabling this shift was the development of "radical ignorance" and "hidden Markov models". IBM, essentially, bypassed the task of understanding and extracting meaning through language.

Speech became the product of "unknown human mental and psychological processes" that need not be represented directly. The IBM team adjusted their calculation aim to only anticipate the most probable outcome, without any knowledge of the formulas that encoded words into sounds. What the IBM group used statistical functions to quantify was not the complex and variable processes of speech, but the very absence of knowledge about them. They developed a statistical analysis of ignorance.

The hidden Markov modeling required little knowledge of speech, but they required a great deal of data. The differentiation between "the write way" and "the right way" wasn't with grammar or syntax but by the frequency of their occurance in "natural language".
The task for IBM was to now find and train on vast quantities of "language-data".

No Data Like More Data

Initially training data was painstakingly collected, by transcription, to punch cards. Larger training corpora were slowly amassed by collecting from IBM manuals, U S patient office records, a collection of children's novels, the American Printing House for the Blind and eventually the Canadian Parliamentary official proceedings. Specialty Languages required additional editing out of limited usage words.
Luckily, IBM was able to develop the hardware space, staff and budget to sustain the effort. By the early 1990s exponential growth of computer readable data aided the training, and allowed further consolidation of nature language processing, information retrieval, machine learning, and data analysis.

Our Blackboards Look A Lot Like Your Blackboards

In 1993 Robert Mercer and Peter Brown leave IBM to join Renaissance Technologies, an immensely profitable investment firm, and pioneer quantitative trading with founder, Jim Simmons, formerly a code breaker at the National Security Agency.

In 2016, the firm's Medallion Fund became the most profitable hedge fund in history.