Towards mapping contemporary AI The Norvig/Chomsky debate

MAS.S66 New Destinations in Artificial Intelligence

11-16-2015

Yoshihiko Suhara

Brains, Minds and Machines (MIT150, 2011)

Brains, Minds and Machines (MIT150, 2011)

- Pinker: … there is a narrative in which the new direction of both artificial intelligence and cognitive science is one that makes a great deal more use of probabilistic information.

- Chomsky: … It's true there's been a lot of work on trying to apply statistical models to various linguistic problems. **I think there have been some successes, but a lot of failures.** There is a notion of success … which I think is novel in the history of science. **It interprets success as approximating unanalyzed data.**

Peter Norvig's Essay

On Chomsky and the Two Cultures of Statistical Learning

Peter Norvig

Director of Research, Google
pnorvig@google.com

focus

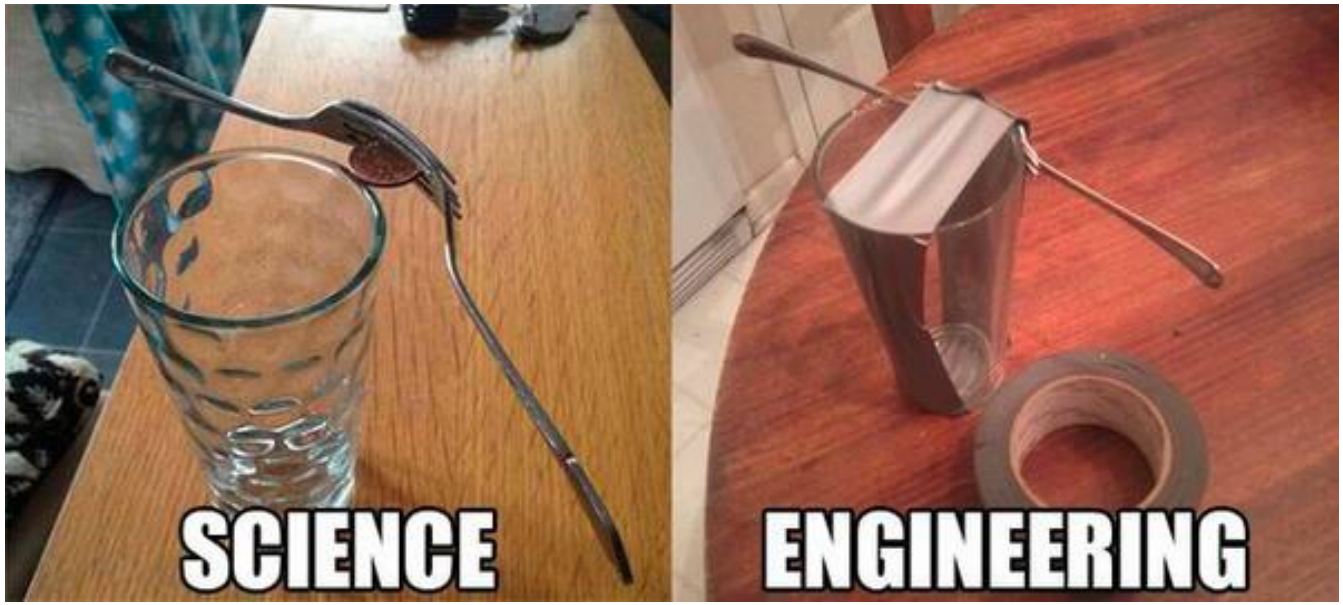
Colorless green ideas
learn furiously
Chomsky and the two cultures of statistical learning

Noam Chomsky on Where Artificial Intelligence Went Wrong

An extended conversation with the legendary linguist

interviewd by Yarden Katz

Science and Engineering



Noam Chomsky on Where Artificial Intelligence Went Wrong

An extended conversation with the legendary linguist

interviewd by Yarden Katz

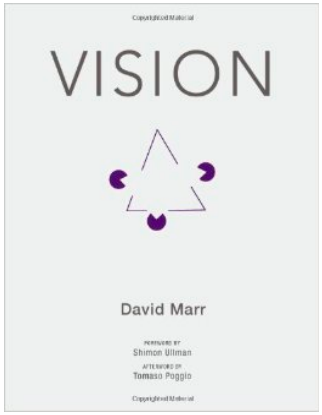
Why has AI been so difficult?

- Chomsky was skeptical about the original work since it was too optimistic
  - assumed we could achieve things that required real understanding of systems that were barely understood

- Neuroscience has been on the wrong track
  - should look at the units of computation
    - e.g., “read”, “write”, and “address” in Turing machine
  - we will **never find computational units** if we look for strengthening of synaptic connections
    - cf. Marr’s model

- The key idea is finding **right units** to describe the problem
  - → Right “level of abstraction”

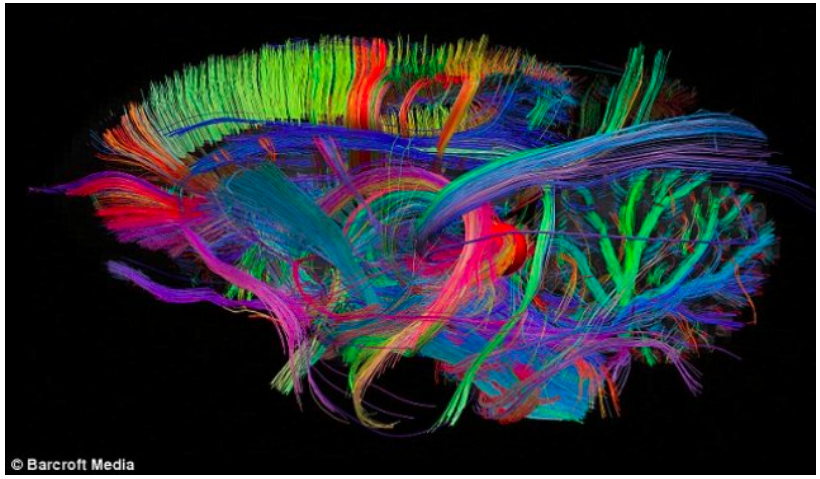
David Marr's model



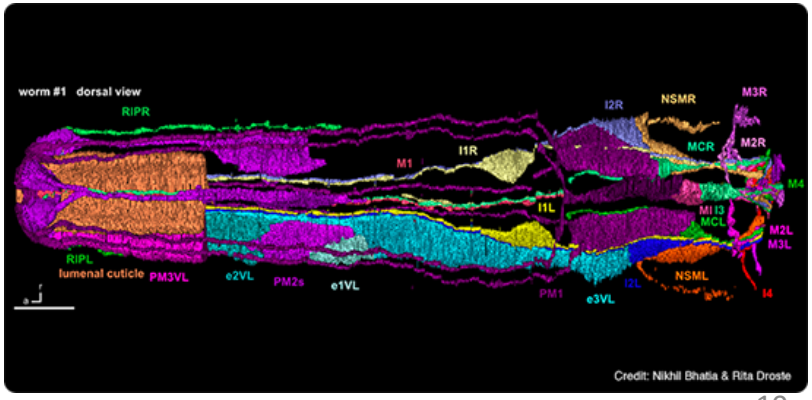
- A general framework for studying complex biological systems
  - (1) Computational level
    - describes the input and output to the system
  - (2) Algorithmic level
    - describes the procedure by which an input is converted to an input
  - (3) Implementation level
    - describes how our own biological hardware of cells implements the procedure described by the algorithmic level

Right “level of abstraction”: Criticism of Connectomics

- The goal of Connectomics is to find **the wiring diagram** of very complex organisms
  - find the connectivity of all the neurons in cerebral cortex
- Example: Nematode Caenorhabditis elegans
  - fails to predict what it is going to do with **only 300 neurons**



Nematode Caenorhabditis elegans



https://biology.mit.edu/people/h_robert_horvitz

Comments on current trend of AI (1/2)
"Good Old Fashioned AI" to statistical approaches

- What we get from probabilistic models is an **approximation to what's happening**
  - more data, better approximation, but we learn nothing about the language

- **A "right" approach** is to see if we can understand what the fundamental principles are
  - that deal with the core properties, and recognize that in the actual stage

Comments on current trend of AI (2/2)
Probability of a sentence is unintelligible

Katz: … there are very rich internal mental representations, comprised of rules and other symbolic structures, and the goal of probability theory is just **to link noisy data in the world with these internal symbolic structures**.

- Chomsky's comments
  - Is there any point in understanding noisy data?
  - Is there some point to understanding what's going on outside the phenomenon?
    - Example: Japanese kids at 9-months will not react to the R-L distinction
      - We do not need to distinguish R-L symbols for them

The unification approach rather than the reductionist approach

- Chemisty-Phyics example
  - Chemistry could not reduce to physics until quantum physics came along
  - We could unify it with **virtually unchanged chemistry**

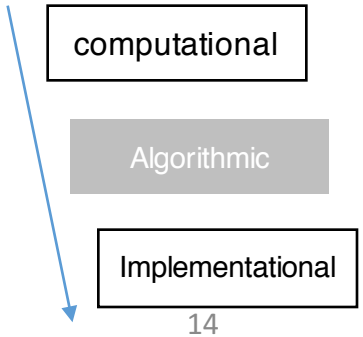
- In the analogy of Marr's approach
  - What we discover at **the computational level** should be unified with what we will find out at **the implementation level**
  - but not in the way of currently understanding the implementation

Language itself is an internal system, which does not have an algorithm

- Marr's conception is designed for information processing systems
  - e.g., vision

- A system of knowledge has no algorithm
  - because it is not a process and there is no calculation of knowledge
  - e.g., Peano axioms and arithmetic operations

- Language is kind of the **arithmetical capacity**
  - only computational level and implementation level



The confilicts between computational efficiency and communicative efficiency

- Example cases
  - Structural ambiguity
    - "Visiting relatives can be a nuisance"
    - it is computationally efficient, but it is inefficient for communication
  - Garden-path sentences
    - "The horse raced past the barn fell"

- Every case of a conflict, **computational efficiency wins**
  - all kinds of ambiguities for simple computational reasons

History of science

- Cognitive sciences are in kind of **pre-Galilean stage**
  - A fact about early science was the recognition that simple things are puzzling

- The difference between reduction and unification
  - chemistry and physics
  - cognitive science and nerosciences

Chomsky: As soon as you allow yourself to be puzzled by it, you immediately find that all your intuitions are wrong.

Peter Norvig's Essay

On Chomsky and the Two Cultures of Statistical Learning

Peter Norvig

Director of Research, Google
pnorvig@google.com

focus

Colorless green ideas
learn furiously
Chomsky and the two cultures of statistical learning

Summary of Norvig's Essay

- (1) What did Chomsky mean, and is he right?
- (2) What is a statistical model?
- (3) How successful are statistical language models?
- (4) Is there anything like their notion of success in the history of science?
- (5) What doesn't Chomsky like about statistical models?
- (6) Two cultures

(1) What did Chomsky mean, and is he right?

- A. Statistical language models have had engineering success, but that is irrelevant to science
- B. Accurately modeling linguistic facts is just butterfly collecting
- C. Statistical models are incomprehensible; they provide no insight
- D. Statistical models may provide an accurate simulation of some phenomena, but the simulation is done completely the wrong way
- E. Statistical models have been proven incapable of learning language

A. Statistical language models have had engineering success, but that is irrelevant to science

- Norvig's comment
  - I agree that engineering success is not the goal or the measure of science
  - **science and engineering develop together**, and that engineering success shows that something is working right, and so is evidence of a scientifically successful model

"Science walks forward on two feet, namely theory and experiment … Sometimes it is one foot that is put forward first, sometimes the other, but continuous progress is only made by the use of both."

by Robert Milikan (physist, 1868-1953)

B. Accurately modeling linguistic facts is just butterfly collecting

- Norvig's comment
  - Science is a combination of gathering facts and making theories; neither can progress on its own

C. Statistical models are incomprehensible; they provide no insight

- Norvig's comment
  - It can be difficult to make sense of a model containing billions of parameters
  - Certainly a human cannot understand such a model by inspecting the values of each parameter individually
  - But one can gain insight by examing the properties of the model—where it succeeds and fails, how well it learns as a function of data

D. Statistical models may provide an accurate simulation of some phenomena, but the simulation is done completely the wrong way

- Norvig's comment
  - Majority of people who study *interpretation* tasks (e.g., speech recognition) see that interpretation is an inherently probabilistic problem
  - Many phenomena in science are stochastic, and the simplest model of them is a probabilistic model

E. Statistical models have been proven incapable of learning language

- Norvig's comment
  - We do not know enough about that capability to rule out probabilistic language representations
  - It is much more likely that human language learning involves something like probabilistic and statistical inference

(2) What is a statistical model?

- A statistical model
  - a function $y = F(x)$
  - given a set of data points $\{(x_1, y_1), (x_2, y_2), \ldots\}$
- A probabilistic model
  - a function $G(x)$ computes a probability distribution
- A model can be statistical or probabilistic or both or neither

Ideal gas laws example

- $P = NkT/V$ is a probabilistic model
  - Ignores the complexity of interactions between individual molecules
  - Summarizes uncertainty about the molecules

- It provides **good predictions and also insight**
  - even though it is a probabilistic model
  - even though it does not completely model reality

- The insight is not available from the true movement of individual molecules

(3) How successful are statistical language models?

- Major application areas:
  - Search engine
  - Machine translation
  - Question answering

- Areas related to the computational linguist:
  - Word sense disambiguation
  - Coreference resolution
  - Part of speech tagging
  - Parsing

(4) Is there anything like [the statistical model] notion of success in the history of science?

- Chomsky meant that the notion of success of **"accurately modeling the world"** is novel, and that the only true measure of success in the history of science is **"providing insight"**

- Norvig's rebuttal
  - Norvig looked at the titles and abstracts from the issues of Science and Cell and 2010 Nobel Prizes
  - He concluded that 100% of these articles are more about **"accurately modeling the world"** than they are about **"providing insight"**

(5) What doesn't Chomsky like about statistical models?

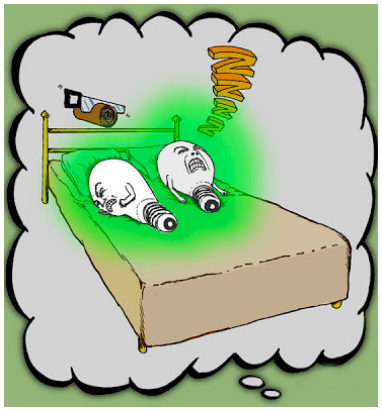
But it must be recognized that the notion of **"probability of a sentence" is an entirely uselss** on, under any known interpretation of this term. (Chomsky, 1967)

I think we are forced to conclude that ... **probabilistic models give no particular insight** into some of the basic problems of syntactic structure. (Chomsky 1957)

"Obvious flaw" by a Markov-chain model

1. **I** never, ever, ever, ever, ... **fiddle** around in any way with electrical equipment.
2. **She** never, ever, ever, ever, ... **fiddles** around in any way with electrical equipment.
3. * **I** never, ever, ever, ever, ... **fiddles** around in any way with electrical equipment.
4. * **She** never, ever, ever, ever, ... **fiddle** around in any way with electrical equipment.

Colorless green ideas sleep furiously



- Famous examples (Chomsky 1957):
  - (A) Colorless green ideas sleep furiously
  - (B) *Furiously sleep ideas green colorless

- Chomsky's claim is that since neither sentence has occured before
  - A statistical model cannot distinguish (A) and (B) since it must assign both a probability of zero
  - A syntactic model can distinguish them

- Norvig says that a simple bigram model with word classes computes (A) is 200,000 times more probable than (B) (Pereira 2001)
  - it also can tell "Effective green products sell well" is more probable than (A)
  - Chomsky's model cannot make this distinction

“Colorless green” in the literature

- "It is neutral green, **colorless green**, like the glaucous water lying in a cellar." The Paris we remember, Elisabeth Finley Thomas (1942).
- "To specify those **green ideas** is hardly necessary, but you may observe Mr. [D. H.] Lawrence in the role of the satiated aesthete." The New Republic: Volume 29 p. 184, William White (1922).
- "**Ideas sleep** in books." Current Opinion: Volume 52, (1912).

Chomsky's objection to learning massive amount of parameters

"we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds."

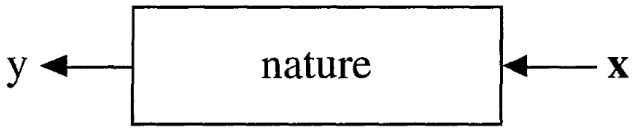
- Nobody is proposing that these parameters are learned one-by-one
- The way to do learning is to set near-zero parameters simultaneously with a smoothing or regularization procedure
- Norvig suggests that probabilistic, trained models are a better model of human language performance than are categorical, untrained model
  - e.g., "big game" rather than "large game"

Statistical Science
2001, Vol. 16, No. 3, 199–231

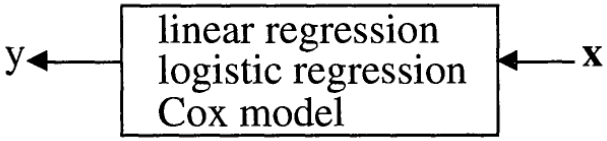
(6) The two cultures

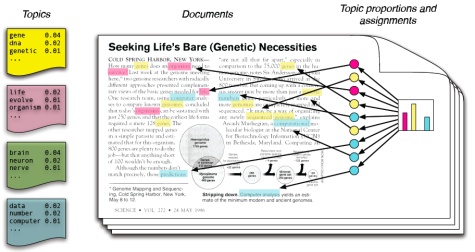
Statistical Modeling: The Two Cultures

Leo Breiman



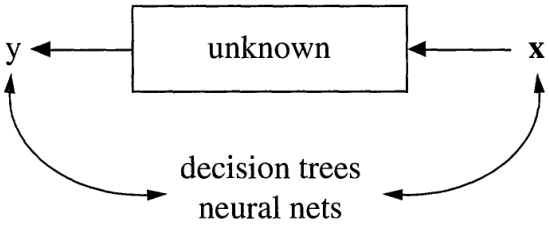
- (1) Data modeling culture

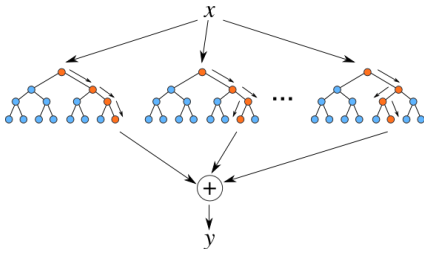




generative model

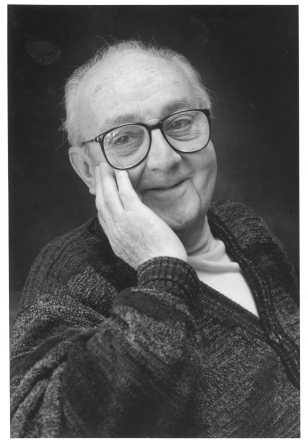
- (2) Algorithmic modeling culture



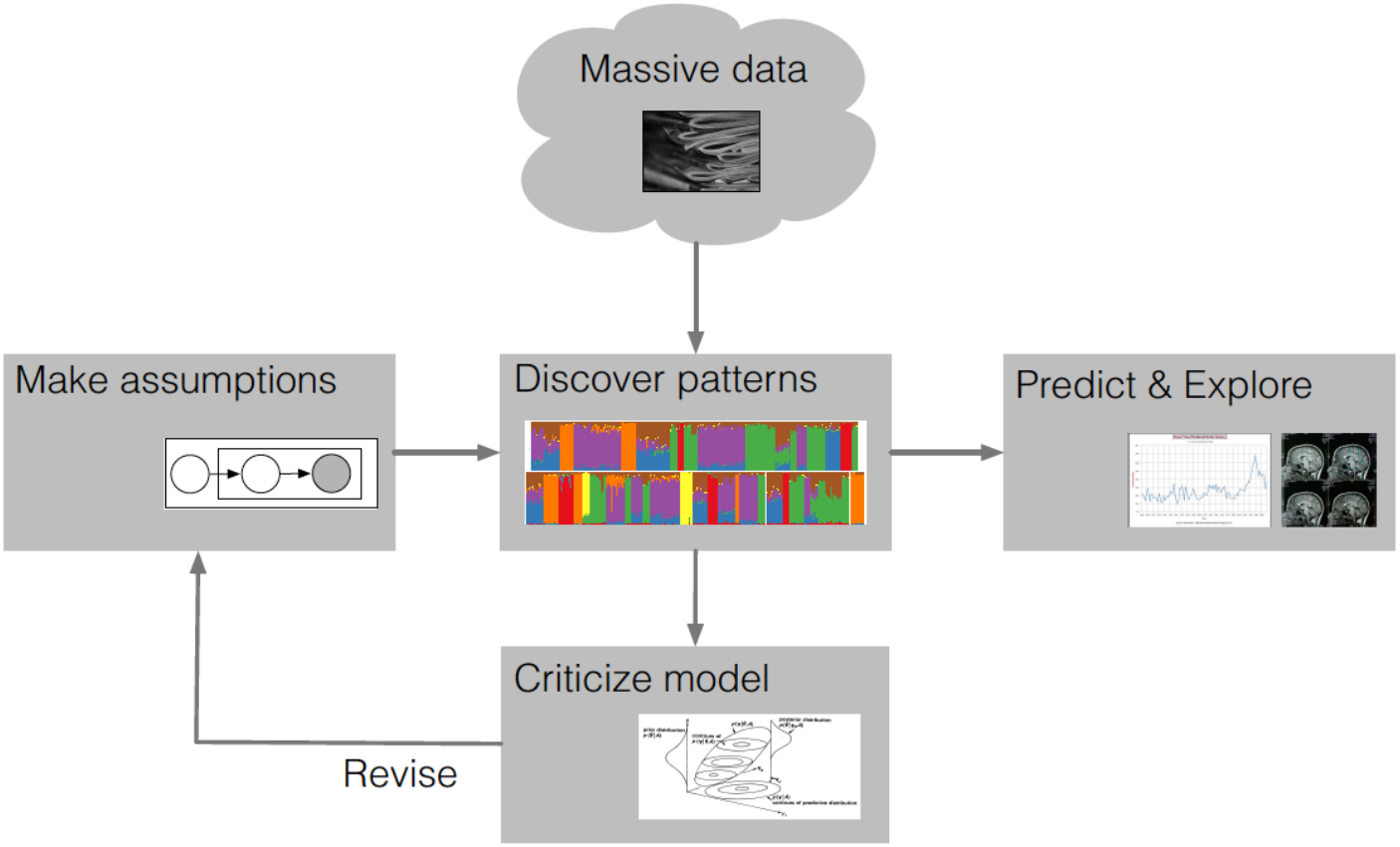


discriminative model

Note: Box's loop in data modeling culture



George E. P. Box



http://www.cs.columbia.edu/~blei/fogm/2015F/notes/introduction-slides.pdf

Horse is not exactly same as "horse" in 15000 B.C. anymore

- Norvig says that Chomsky thinks we should focus on **the ideal, abstract forms** that underlie language
- There is no such thing as a single ideal eternal "horse" form
  - e.g., Lascaux horse and horse
- Languages are complex, random biological processes and must be using something like **probabilistic reasoning**

Reproduction of some Lascaux artwork in Lascaux II

$\neq$