More on Validity

(and the validity of validity)

theories and
methodologies

"All Models Are Wrong"

RICHARD JEAN SO

SEVERAL YEARS AGO, THE FIRST THING I LEARNED IN MY INTRODUC-
TORY STATISTICS CLASS WAS THE FOLLOWING DECLARATION, WHICH THE
instructor had written in capital letters on the blackboard: "all mod-
els are wrong." Models are statistical, graphic, or physical objects, and
their primary quality is that they can be manipulated. Scientists and
social scientists use them to think about the social or natural worlds
and to represent those worlds in a simplified manner. Statistical mod-
els, which dominate the social sciences, particularly in economics,
are typically equations with response and predictor variables. Spe-
cifically, a researcher seeks to understand some social phenomenon,
such as the relation between students' scores on a math test and how
many hours the students spent preparing for the exam. To predict
or describe this relation, the researcher constructs a quantitative
model with quantitative inputs (the number of hours each student
spent studying) and outputs (each student's test score). The researcher
hopes that the number of hours a student spent preparing for the
exam will correlate with the student's score. If it does, this quantified
relation can help describe the overall dynamics of test taking.

Literary scholars have long cast a suspicious and critical gaze
toward modeling, which strikes them as offensively simpleminded
and naive: models run counter to the deep and intensive reading
that literary critics take pride in, the exposing of nuance and sin-
gularity in texts, writers, and human beings. What about gender?
What about race? Don't they influence how well a student does on a
test? And even if you could quantify gender and race and add them
to the model, there are always additional dimensions of identity and
experience to include. In the end, an individual exceeds socially con-
structed identity categories, so what does a model do besides reify
such categories? Historians of finance, such as Mary Poovey and
Donald MacKenzie, have provided many examples of how modeling
imperils the social world; for instance, when economists fixate on the
model as a tool for reasoning, they conflate the model's internal logic
with the logic of the social world, assuming—in a frightful inversion

RICHARD JEAN SO is assistant professor
of English and cultural analytics at McGill
University. His recent articles, both writ-
ten with Hoyt Long, are "Turbulent Flow:
A Computational Model of World Litera-
ture," in *MLQ*, and "Literary Pattern Recog-
nition: Modernism between Close Reading
and Machine Learning," in *Critical Inquiry*.

"All models are wrong, but some are useful."

- "A model allows the researcher to isolate aspects of an interesting phenomenon, and in discovering certain properties of such aspects the researcher can continue revising the model to identify additional properties...

"All models are wrong, but some are useful."

- "A model allows the researcher to isolate aspects of an interesting phenomenon, and in discovering certain properties of such aspects the researcher can continue revising the model to identify additional properties. In this "iterative process," the "truth" of that phenomenon resides at some asymptotic point that can never be reached. But along the way, the modeling process yields productive insights."

"All models are wrong, but some are useful."

- Models are more than simply statistical description; they are mechanisms through which people reason and think.
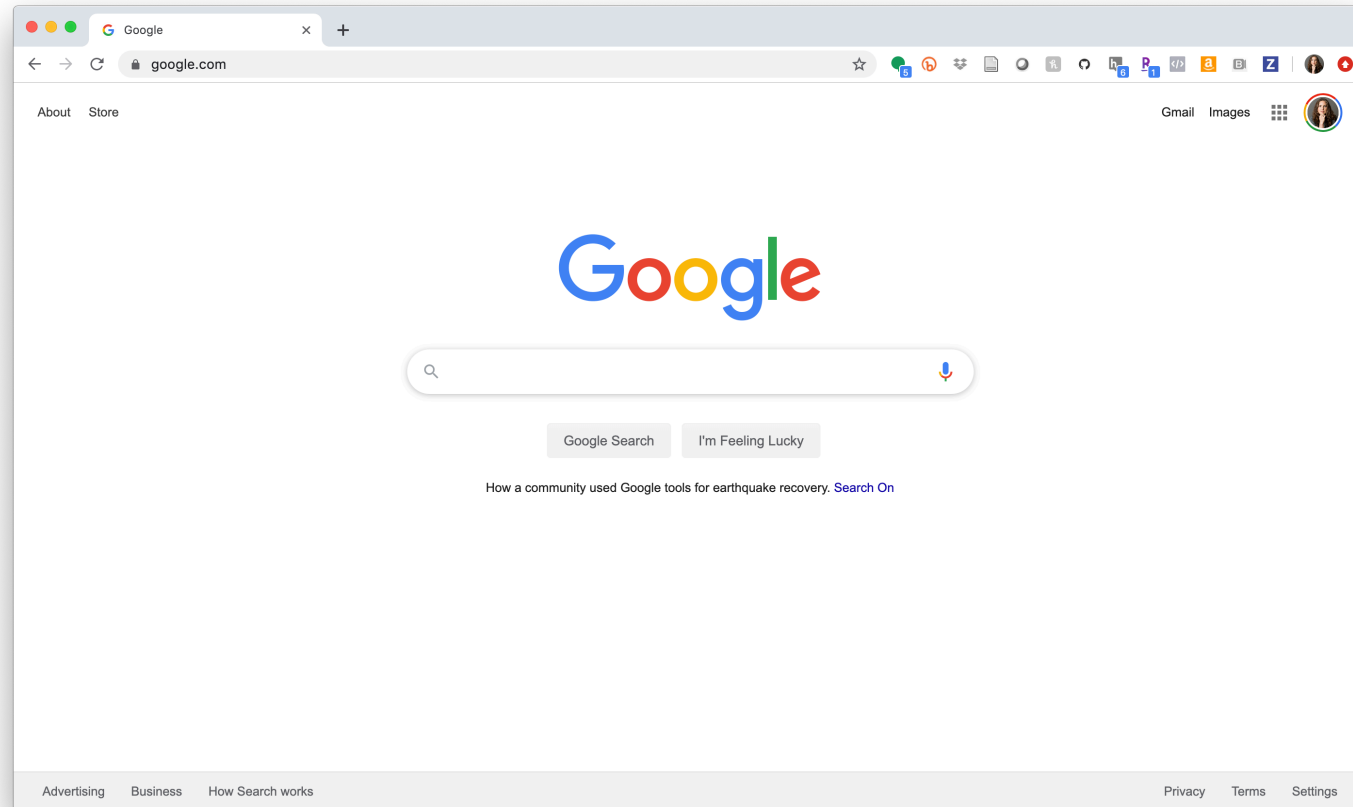
"All models are wrong, but some are useful."

- Danger! Do not conflate the model's internal logic with the logic of the social world! E.g. the 2008 financial crisis.
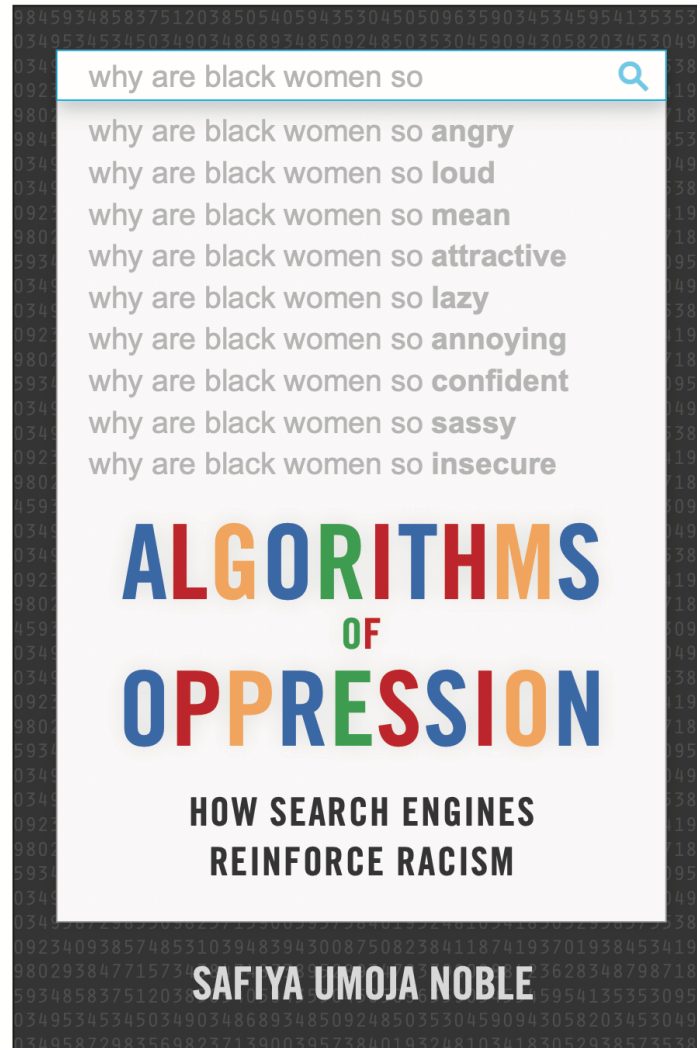
"All models are wrong, but some are useful."

- By focusing on errors, outliers, etc., we can understand *how* it is wrong, and begin to ask *why* it is wrong.

A model that is wrong?

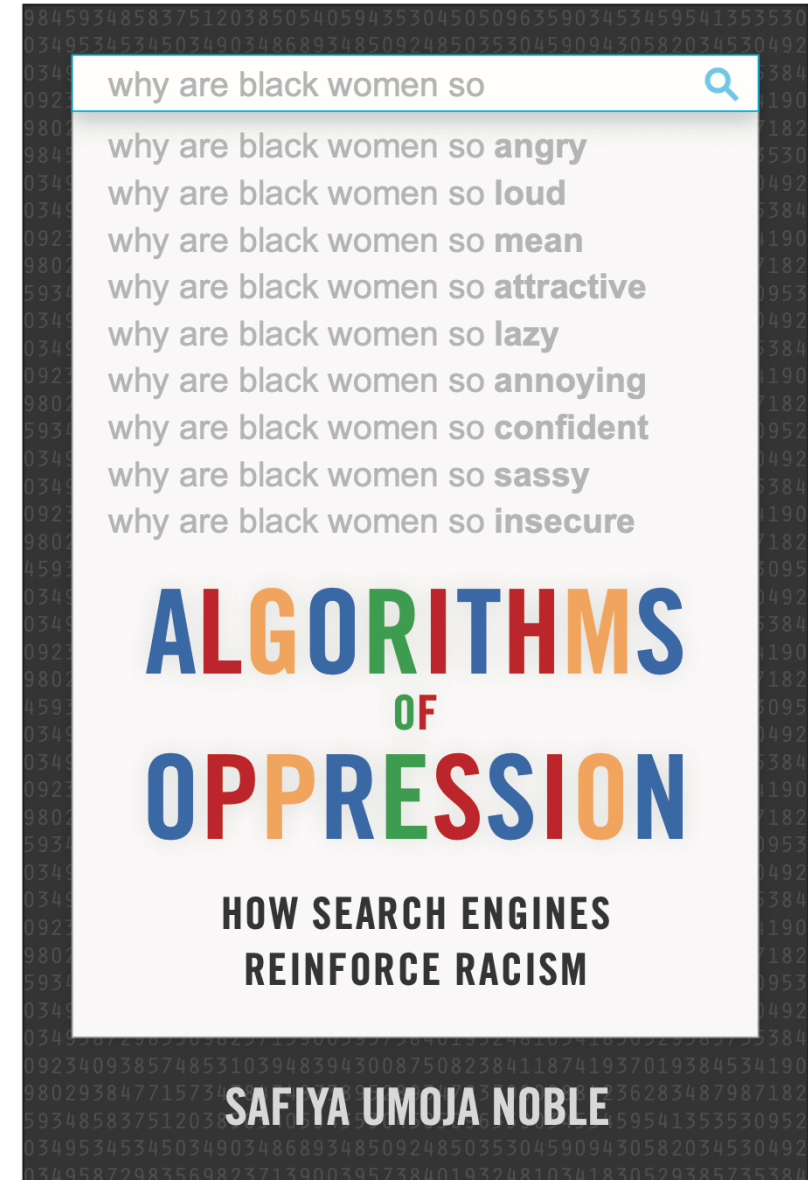# Algorithms of Oppression

Markov Models

- A Markov model is a model that obeys Markov property.

- A Markov property is a process in which the next state depends only on the current state.

- Mathematically speaking, the conditional probability distribution of the next state depends on the current state and not the past states. That is *s(t)* depends only on *s(t-1),* where *s(t)* is the state at time *t*.

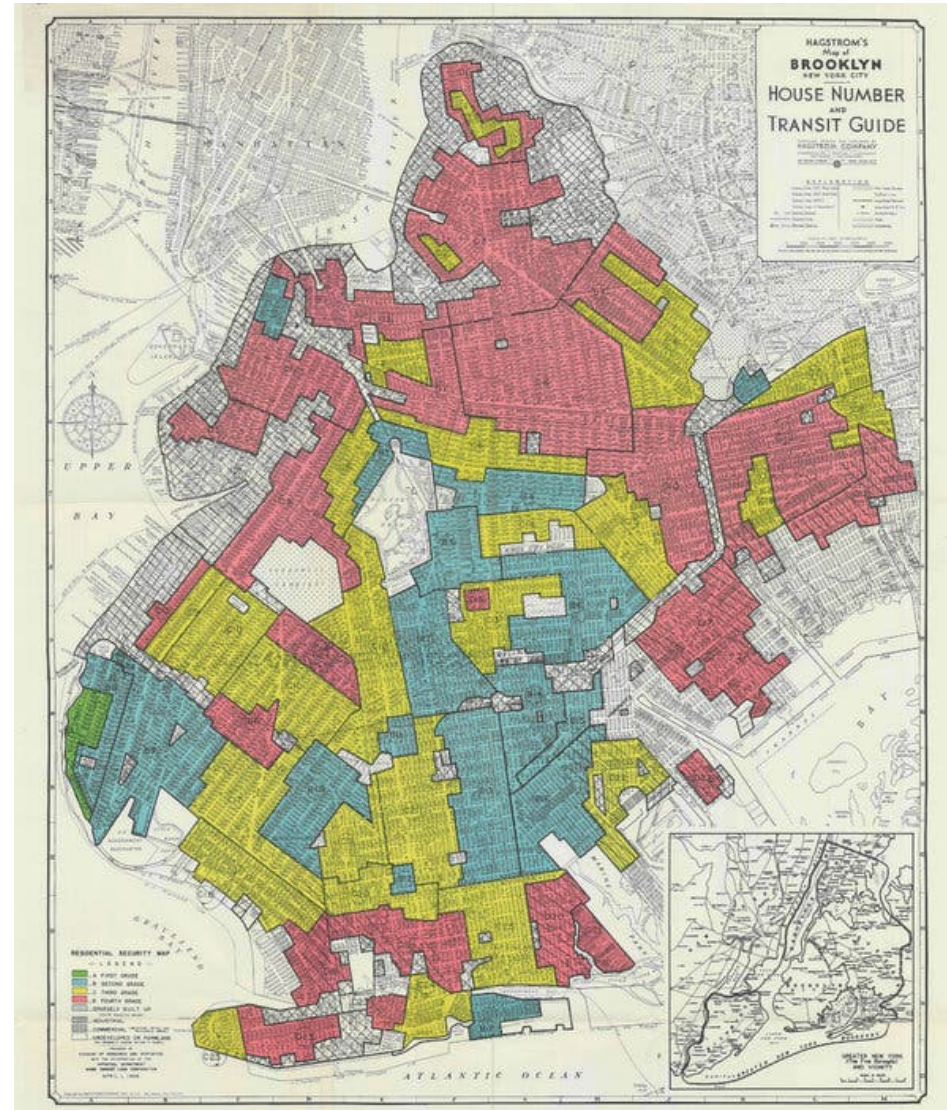- This is what is a "first-order Markov model."

Markov Models

- Writing often modeled as a Markov chain:
  - Next-word prediction (e.g. autocomplete)
  - Random poetry generator
  - Bots, etc.

- Next words are determined on the basis of weighted distributions
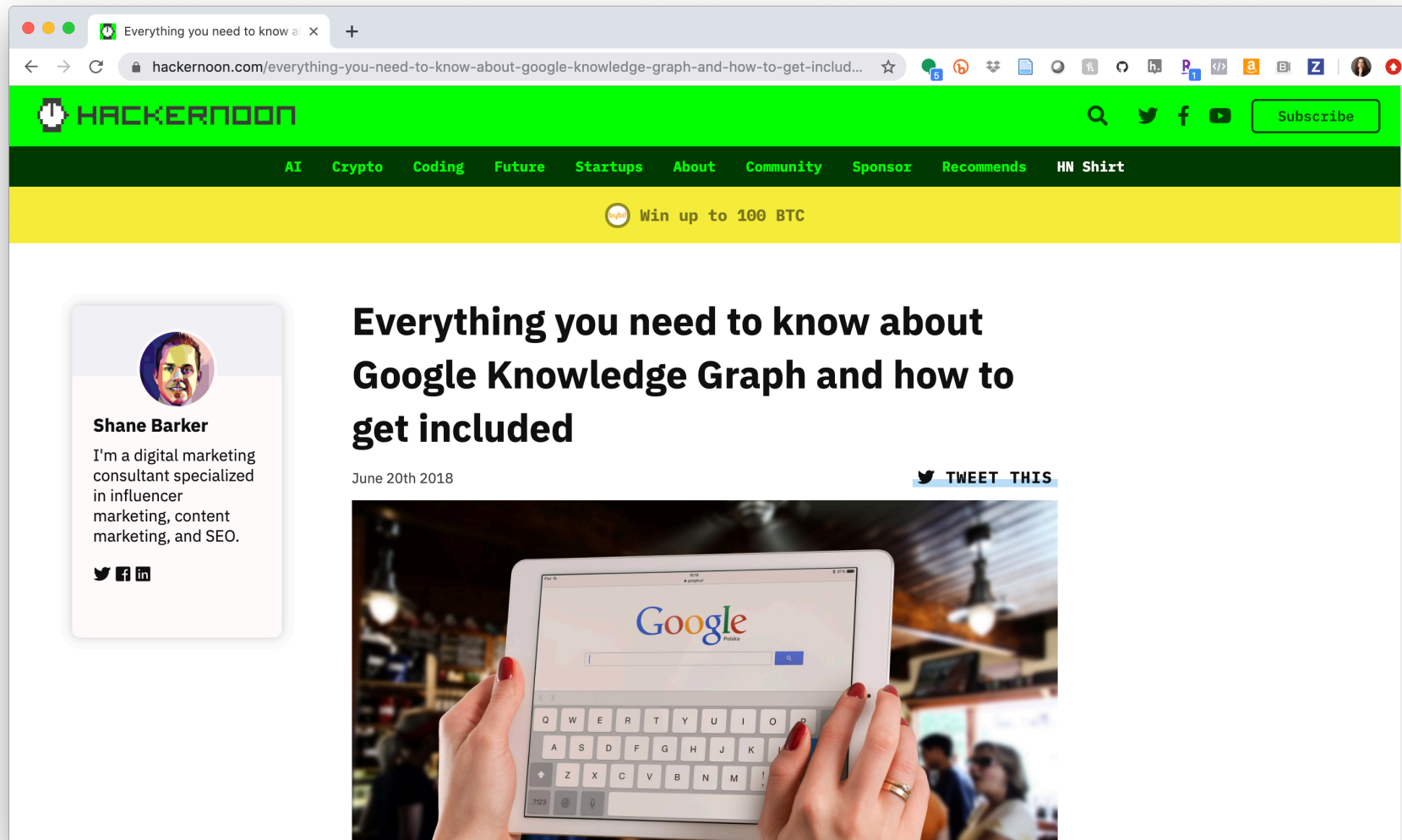 - E.g. probability of word in corpus

Markov Models

- But what is the corpus?
- ML algorithms "predict the past"

What if the past is racist, sexist, etc?

Google Knowledge Graph

Google Knowledge Graph

User Behavior—Google knows and understands what users are searching for, what they are clicking on, and how long they spend on a page after clicking on it. This is when the graph algorithmically connects all of this data to create results about your brand. So, it is the user's queries that contribute in displaying the information through the knowledge graph.

Semantic Search—To produce highly relevant search results, the knowledge graph considers semantic search. This includes the context of research, synonyms, searcher location, and various other points.

Entity Indexing—Google tries to catalog all the related entities through the process of mapping known as Entity Recognition and Disambiguation to produce factual information about each entity.

Algorithms of Oppression

https://canvas.emory.edu/courses/65404/files/3031216?module_item_id=739569