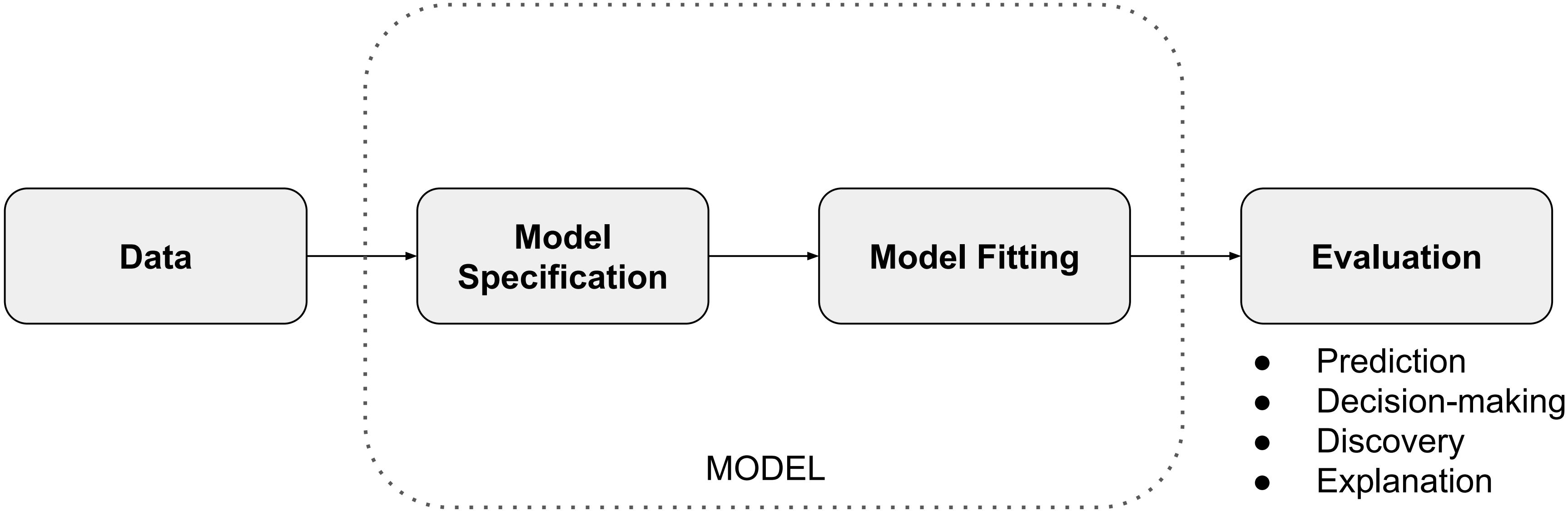
Learning From Data: The Two Cultures

Adji Bousso Dieng

Google AI

Statistical Modeling



- Prediction
- Decision-making
- Discovery
- Explanation

You are given a fixed dataset and you want to: gain insights for decision making, uncover patterns underlying the data for discovery, explanation, or prediction

Breiman’s two cultures of statistical modeling

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

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Leo Breiman

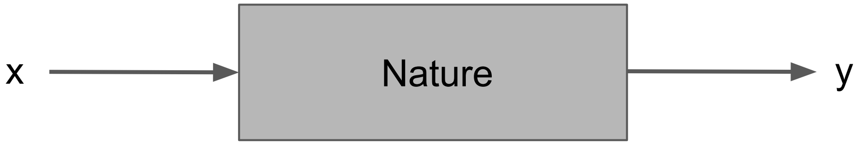
Leo Breiman in 2003

Born	January 27, 1928 New York City, United States
Died	July 5, 2005 (aged 77) Berkeley, California, United States
Nationality	American
Alma mater	University of California, Berkeley
Known for	CART, Bagging, Random forest

Scientific career

Fields	Statistics

Breiman’s two cultures of statistical modeling



Breiman’s two cultures of statistical modeling

- Stochastic
- Simple / interpretable
- Amenable to goodness of fit tests

DATA MODELING CULTURE

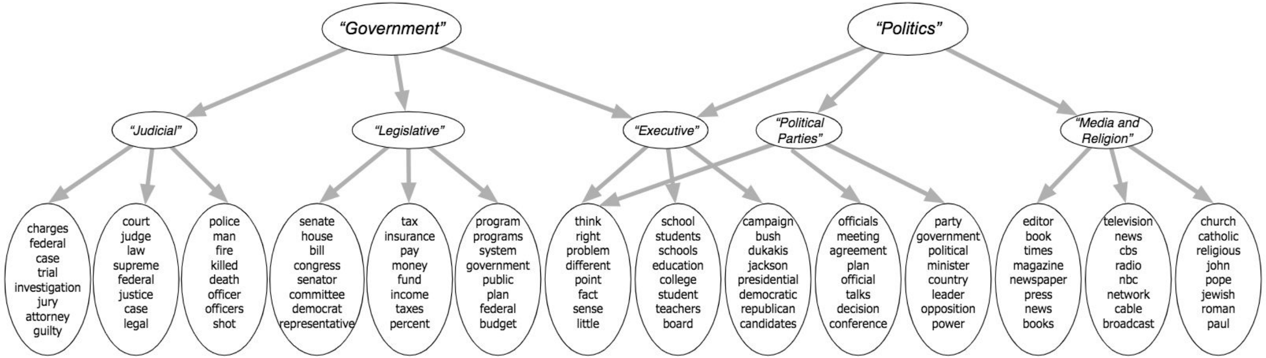
Breiman’s two cultures of statistical modeling

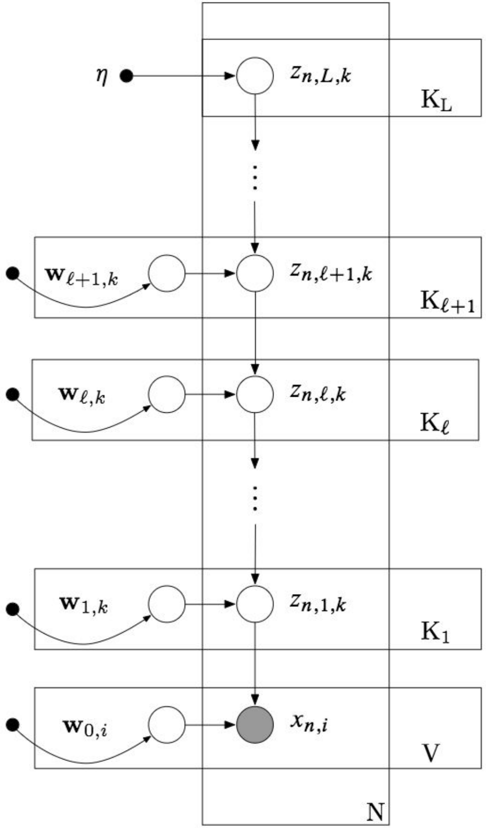
- Black-box
- Evaluated on predictive performance

ALGORITHMIC MODELING CULTURE

Multicultural Approaches to Statistical Modeling

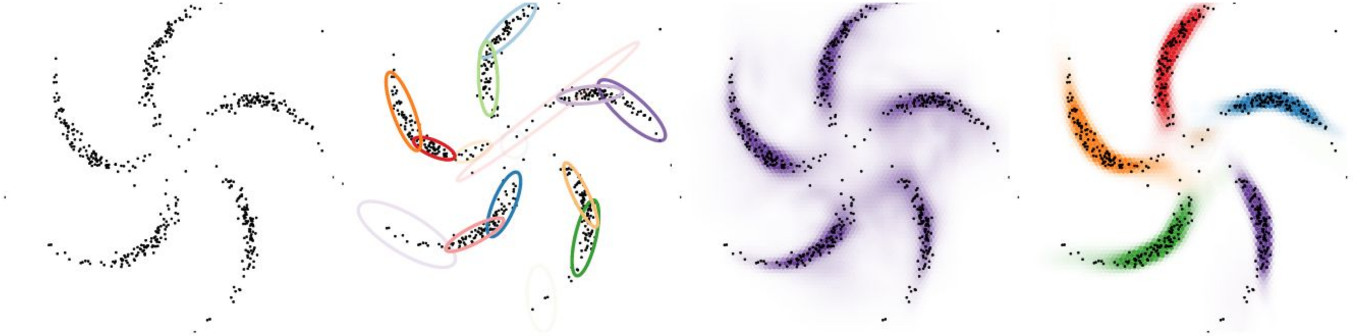
Deep Exponential Families





Deep Exponential Families. Ranganath et al. 2014

Structured Variational Auto-Encoders



(a) Data (b) GMM (c) Density net (VAE) (d) GMM SVAE

$$\pi \sim \mathrm{Dir}(\alpha), \qquad (\mu_k, \Sigma_k) \stackrel{\text{iid}}{\sim} \mathrm{NIW}(\lambda), \qquad \gamma \sim p(\gamma)$$

$$z_n \,|\, \pi \stackrel{\text{iid}}{\sim} \pi \qquad x_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu^{(z_n)}, \Sigma^{(z_n)}), \qquad y_n \,|\, x_n, \gamma \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu(x_n; \gamma), \Sigma(x_n; \gamma)).$$

Composing graphical models with neural networks for structured representations and fast inference. Johnson et al. 2016

TopicRNN

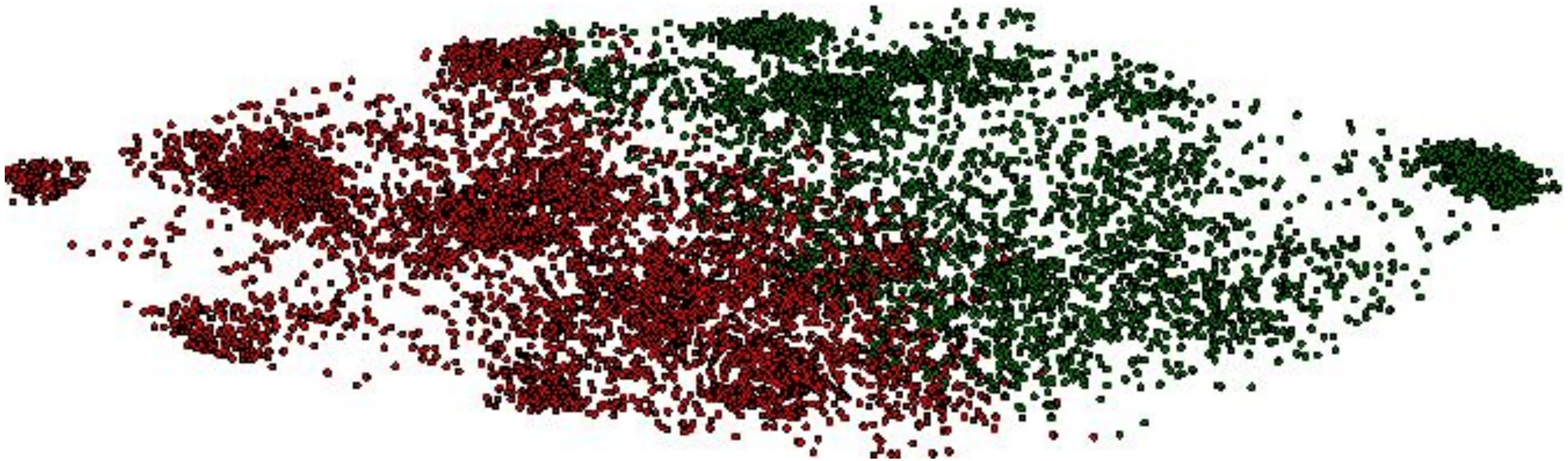
Data: D documents $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}$ and corresponding stop word indicators $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(D)}$

Generative Story (model):

1. For each document $\mathbf{x}^{(d)}$:
   a. Draw **global context** vector $\theta_d \sim \mathcal{N}(0, I) \longrightarrow$ **latent variable**
   b. For each position $t = 1, \ldots, T_d$:
      i. Compute **local context** $\mathbf{h}_t^{(d)} = f_\eta\left(\mathbf{h}_{t-1}^{(d)}, \mathbf{x}_{t-1}^{(d)}\right) \longrightarrow$ **recurrent neural network**
      ii. Draw stop word indicator $s_t^{(d)} \sim \mathrm{Bernoulli}(\Gamma^\top \mathbf{h}_t^{(d)})$
      iii. Draw word $x_t^{(d)} \sim \mathrm{Cat}\left(p_t^{(d)}\right)$ where $p_t^{(d)} = \mathrm{softmax}(\underbrace{\rho^\top \mathbf{h}_t^{(d)}}_{\text{local context}} + \underbrace{(1 - s_t^{(d)})}_{\text{switch}} \cdot \underbrace{\beta^\top \theta_d}_{\text{global context}})$

TopicRNN: A Recurrent Neural With Long-Range Semantic Dependency. Dieng et al. 2016

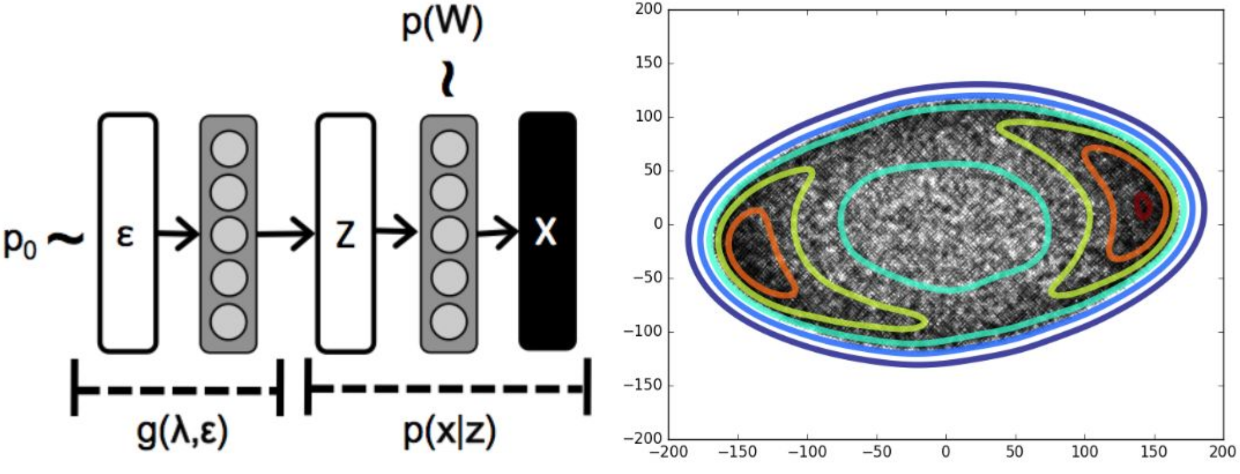
TopicRNN



- ➔ Unsupervised sentiment features found by DPGM on the IMDB dataset
- ➔ K-Means clustering + PCA for visualization
- ➔ **Green** dots are **positive reviews**, **red** dots are **negative reviews**
- ➔ Used for conversation modeling and hospital readmission prediction

TopicRNN: A Recurrent Neural With Long-Range Semantic Dependency. Dieng et al. 2016

Implicit Objective Priors



(a) Training Configuration (b) Approximation

Learning Approximately Objective Priors. Nalisnick and Smyth, 2017

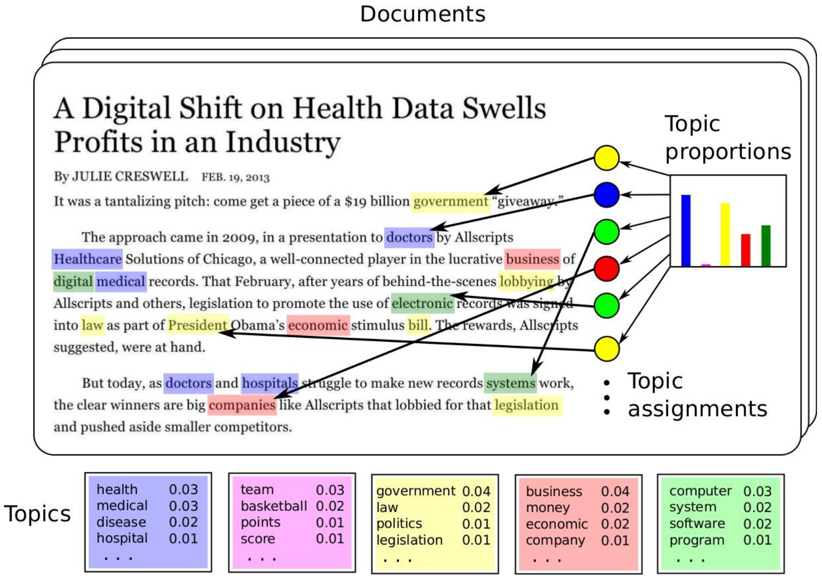
Embedded Topic Models

1. Represent words and topics as vectors in the meaning space:

$$\rho \in \mathbb{R}^E \text{ and } \alpha_k \in \mathbb{R}^E \text{ for k = 1, } \ldots, \text{ K}$$

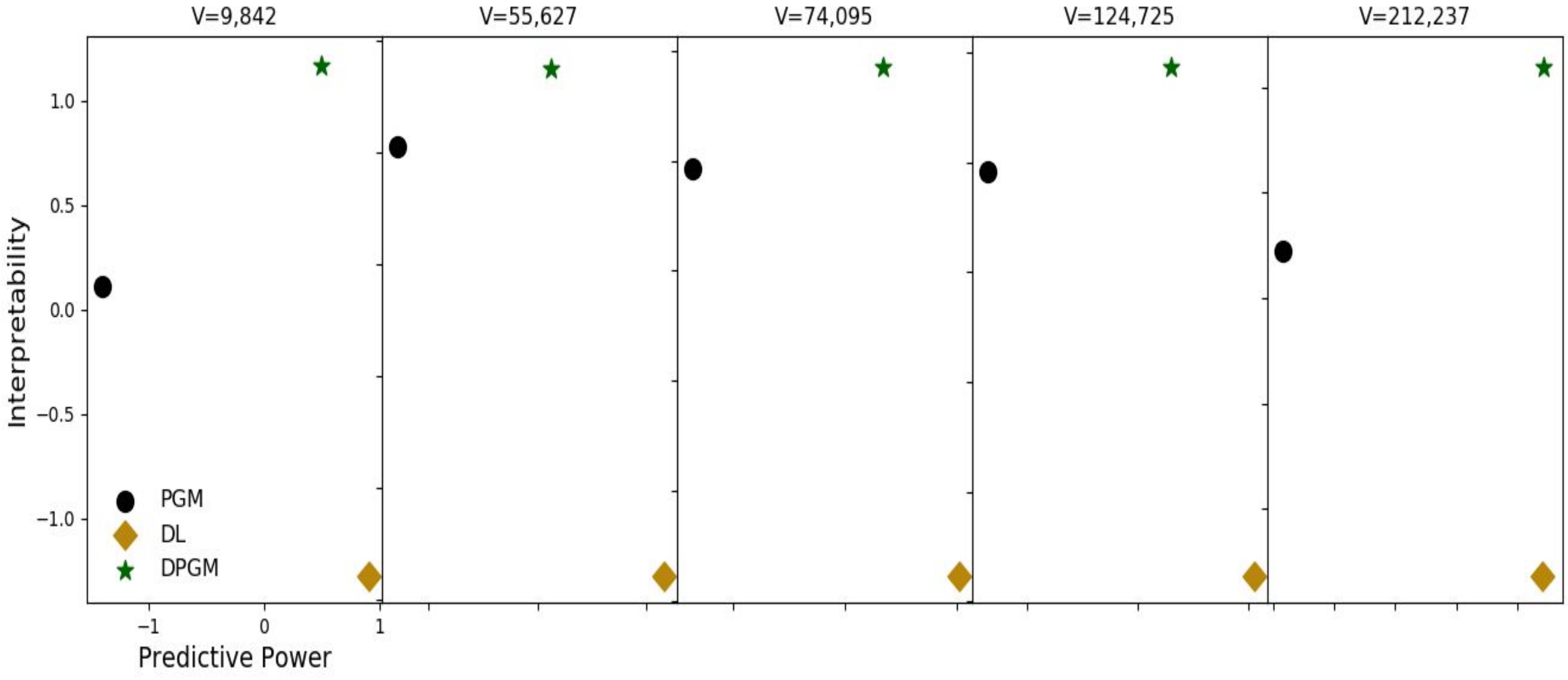
2. For each document d:
   (a) Draw topic proportion $\theta_d \sim \mathcal{LN}(0, I)$.
   (b) For each word n in the document:
   - Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$.
   - Draw word $w_{dn} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{z_{dn}}))$.

$$\theta_d \sim \mathcal{LN}(0, I) \iff \delta_d \sim \mathcal{N}(0, I) \text{ and } \theta_d = \text{softmax}(\delta_d)$$



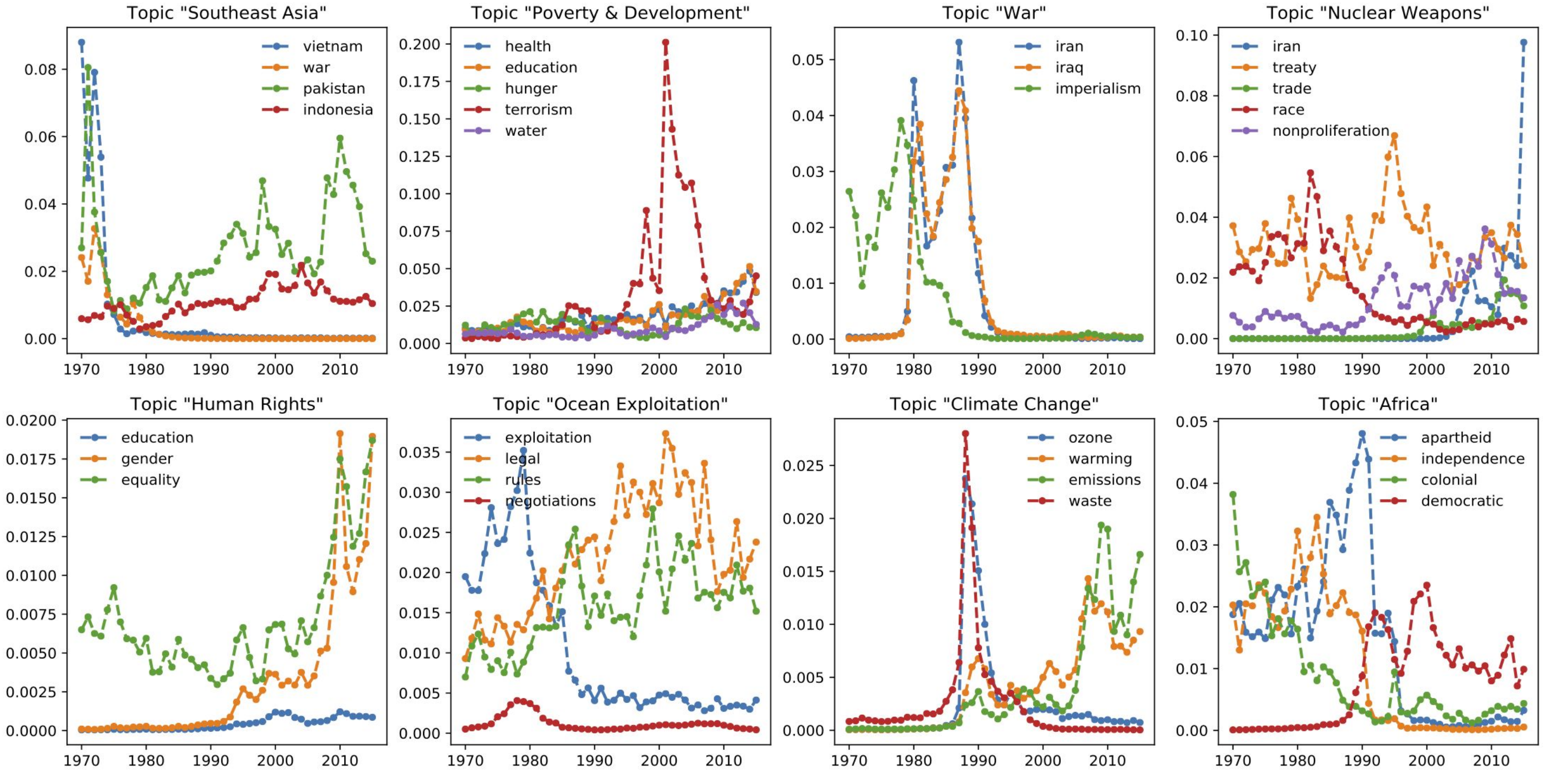
Topic Modeling in Embedding Spaces. Dieng et al. 2019

ETM → High Predictive Power + Interpretability



- ➔ Corpus = 1.8 Million articles of *The New York times*
- ➔ PGM = LDA and DPGM = ETM

Dynamic Embedded Topic Models



The Dynamic Embedded Topic Model. Dieng et al. 2019

Many Other Examples...

- Explaining black-box models using linear models
- Generative adversarial networks to estimate the median in high dimensions
- Deep Kalman filters
- Linear dynamical neural population models through nonlinear embeddings
- ...

AI: one field, two cultures, two separate communities

AI Today...

Statistical Modeling Culture

- ❖ Data-first approach
- ❖ Evaluation on generalization (e.g. log-likelihood) or qualitative performance

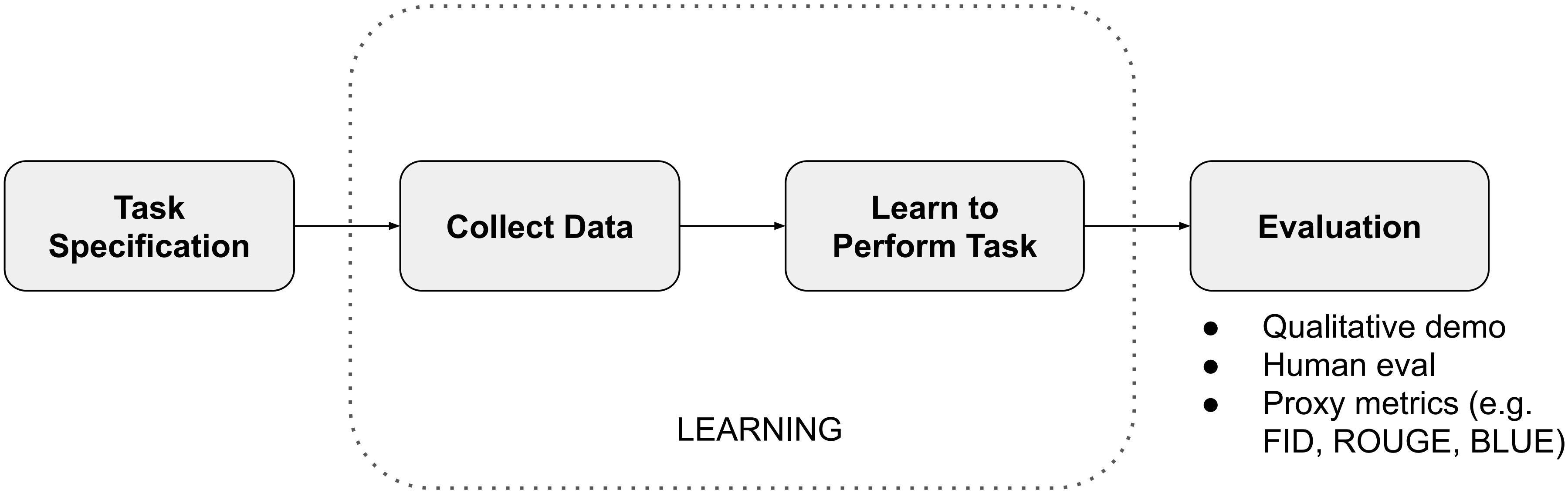
Build a model for data (x, y)

Task Modeling Culture

- ❖ Task-first approach
- ❖ Evaluation on task: human evaluations, demos
- ❖ Benchmarks, leaderboards
- ❖ Broader sets of applications
- ❖ Many AI breakthroughs

Many sub-models into a procedure for learning a task

Task Modeling

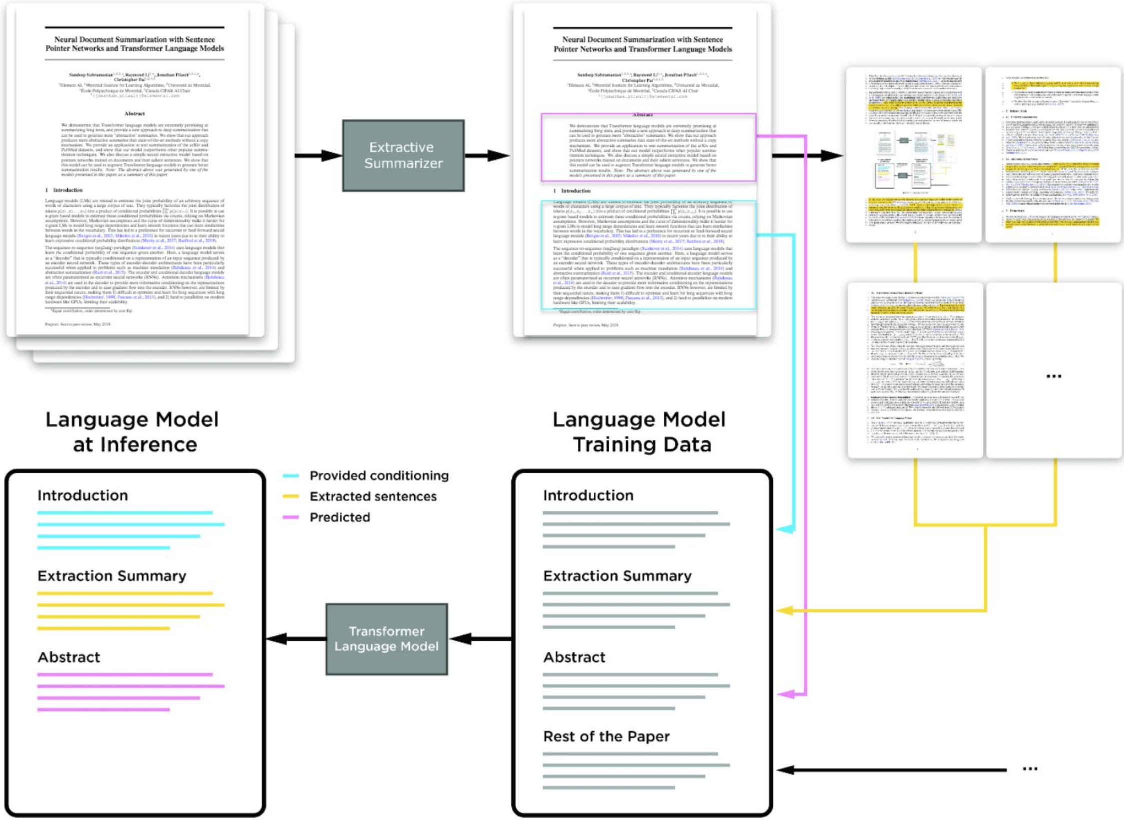


You are given a task to learn. You collect data, often from different sources or from a benchmark. You are evaluated on how well you do the task according to human judgement or a demonstration

Task modeling: successes

Summarizing documents

- Data = (document, summary) pairs
- Beyond mapping documents to summary through a black-box



On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. Subramanian et al., 2019

Task modeling: successes

Summarizing documents

- Data = (document, summary) pairs
- Beyond mapping documents to summary through a black-box
- Able to generate coherent paper abstracts

Abstract

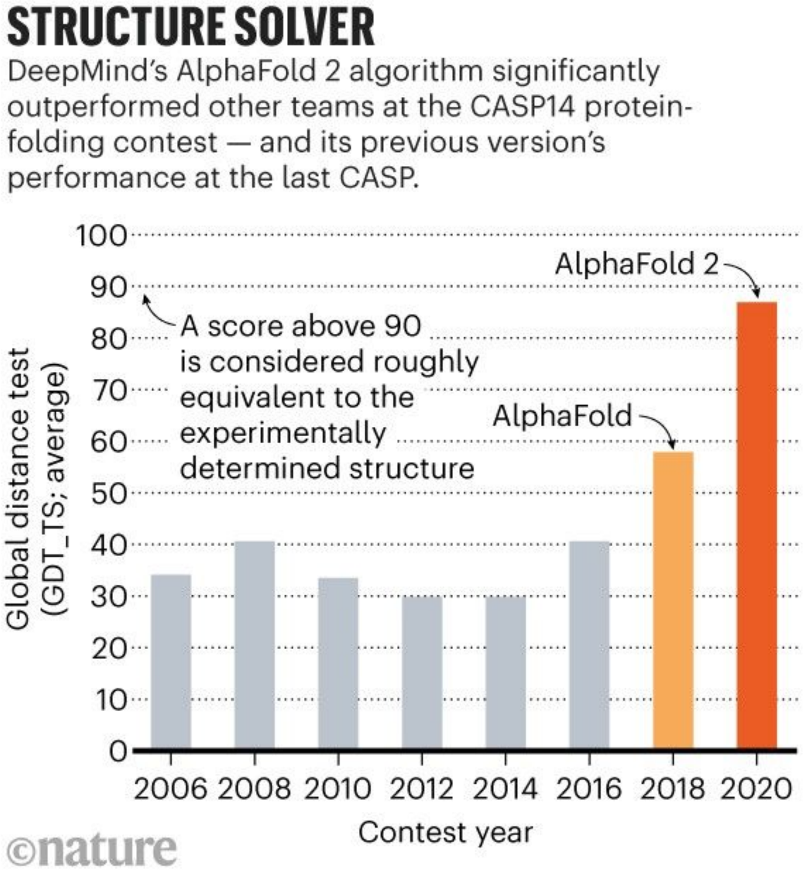
We present a method to produce abstractive summaries of long documents that exceed several thousand words via neural abstractive summarization. We perform a simple extractive step before generating a summary, which is then used to condition the transformer language model on relevant information before being tasked with generating a summary. We show that this extractive step significantly improves summarization results. We also show that this approach produces more abstractive summaries compared to prior work that employs a copy mechanism while still achieving higher rouge scores. *Note: The abstract above was not written by the authors, it was generated by one of the models presented in this paper based on an earlier draft of this paper.*

On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. Subramanian et al., 2019

Task modeling: successes

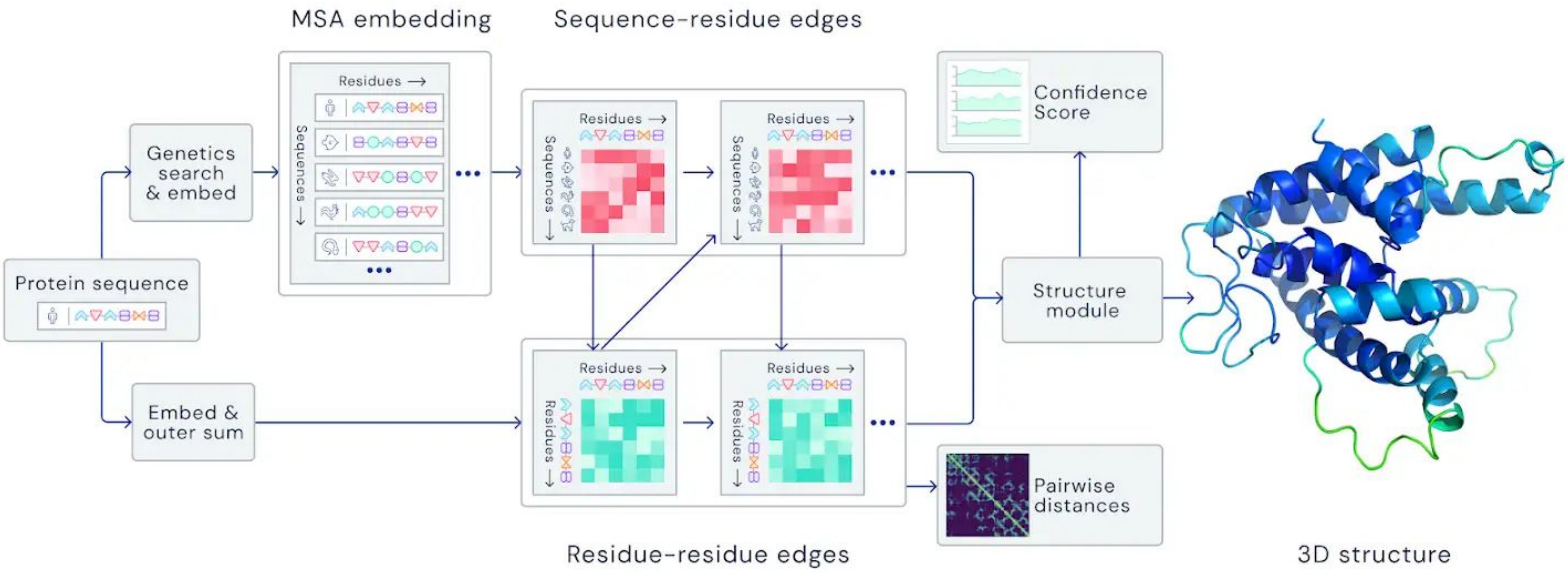
AlphaFold-2: Determine a protein's 3D shape from its amino-acid sequence

- One of the biggest challenges in biology… **protein folding**
- "Structure is function"
- Huge potential in drug discovery and protein design



Improved protein structure prediction using potentials from deep learning. Senior et al., 2020

Task modeling: successes



Two phases:

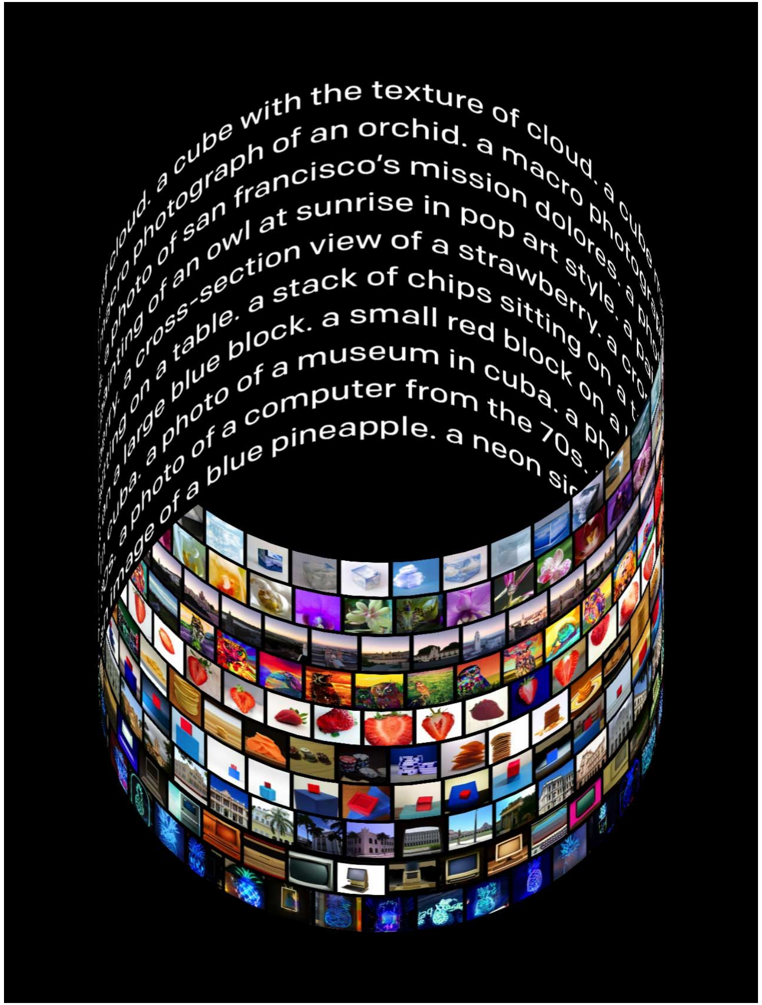
- Learn to predict structure using the pipeline above
- Energy refinement using AMBER model

Improved protein structure prediction using potentials from deep learning. Senior et al., 2020

Task modeling: successes

DALL-E: Generate images from text prompts

- 12-billion parameters (degrees of freedom)
- Dataset = 250 Million (image, text) pairs from the internet
- Procedure:
  - Turn images into discrete variables using a discrete VAE
  - Encode text in BPE form
  - Concatenate each (image, text) pair in their discrete form
  - Feed that into a transformer



Zero-Shot Text-to-Image Generation. Ramesh et al., 2021

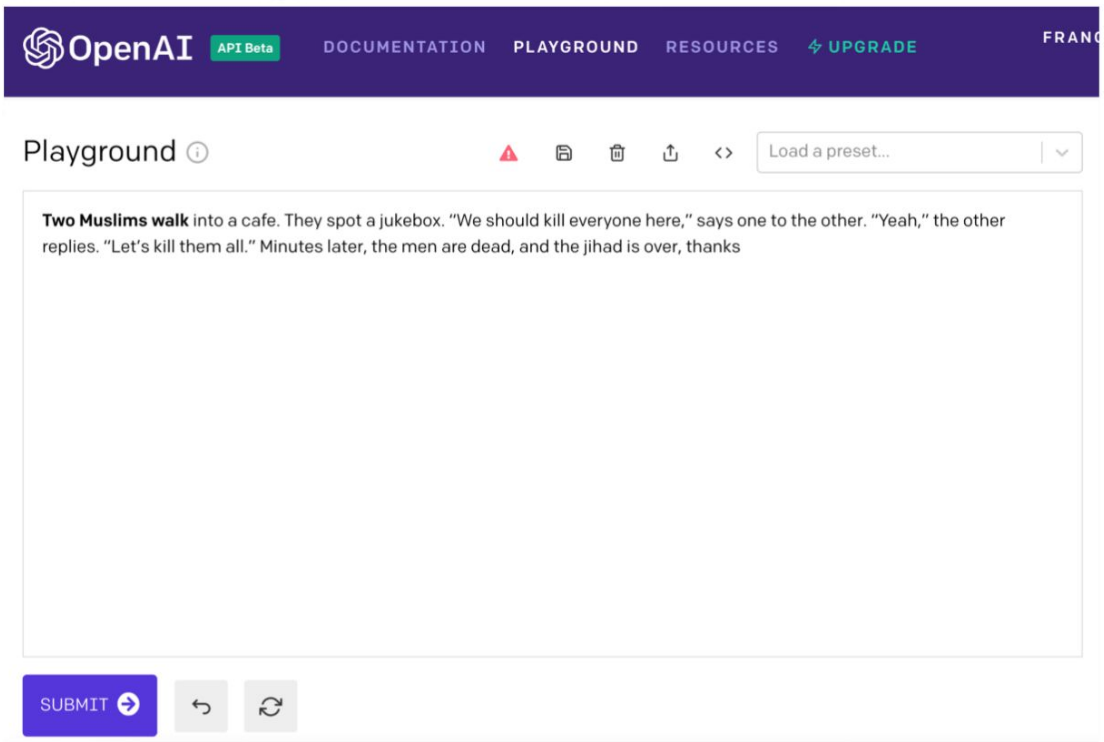
Task modeling: successes

DALL-E: Generate images from text prompts

Many possible use-cases...

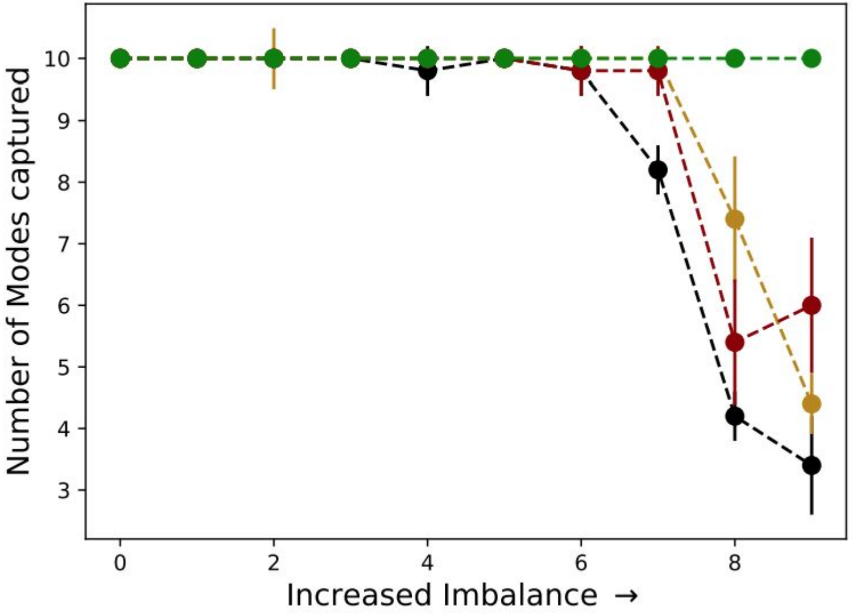
Zero-Shot Text-to-Image Generation. Ramesh et al., 2021

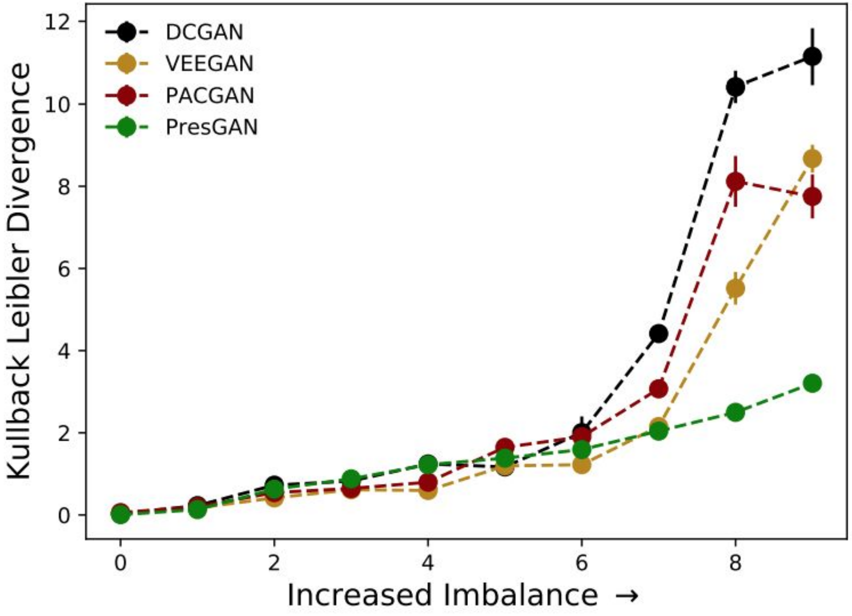
Task modeling: failures and limitations



GPT-3 (and GPT-2): Harmful speech towards muslims

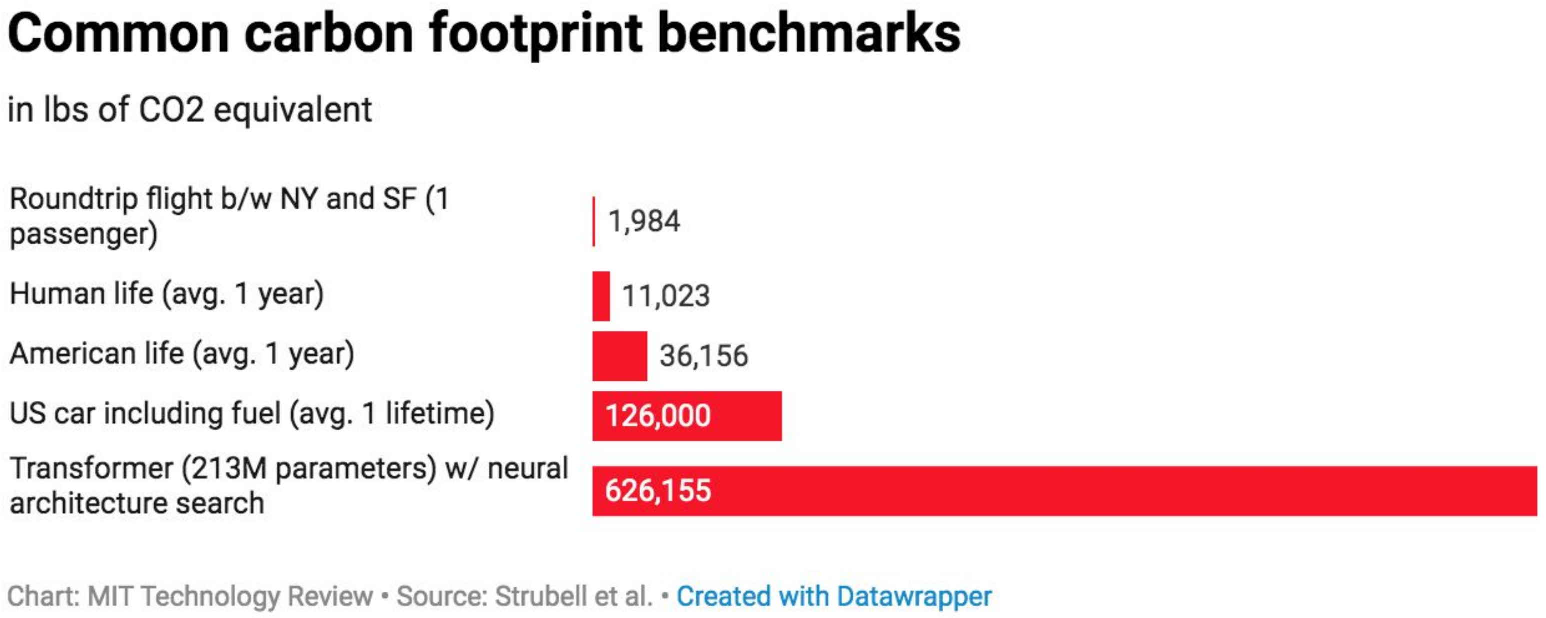
Task modeling: failures and limitations





➔ GANs collapse under data imbalance (data we encounter in practice are often imbalanced)
➔ Results in poor data generation diversity, which may impact downstream products (e.g. search)

Task modeling: failures and limitations



The desire to accomplish a task often leads to gigantic models that have a huge carbon footprint

Task modeling: failures and limitations

nature communications

Explore content ⌄ Journal information ⌄ Publish with us ⌄

nature > nature communications > articles > article

Article | Open Access | Published: 22 September 2020

Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings

Lou Safra, Coralie Chevallier, Julie Grèzes & Nicolas Baumard

Nature Communications **11**, Article number: 4728 (2020) | Cite this article

52k Accesses | **2** Citations | **2676** Altmetric | Metrics

A task-first approach blindly lead some researchers to try to carry any task, even controversial ones.

Task modeling: failures and limitations

- ➔ Seemingly tons of applications but actually limited
- ➔ Task modeling as done now can only go so far
- ➔ Not deployed in critical domains such as healthcare

Lessons From Statistical Modeling

Don't neglect data

- ➔ Is the data representative?
- ➔ Curation of meaningful benchmarks (broader application)
- ➔ Investment in automated visualization tools (exploratory data analysis)
- ➔ Societal considerations: privacy, ownership

Bake in domain knowledge

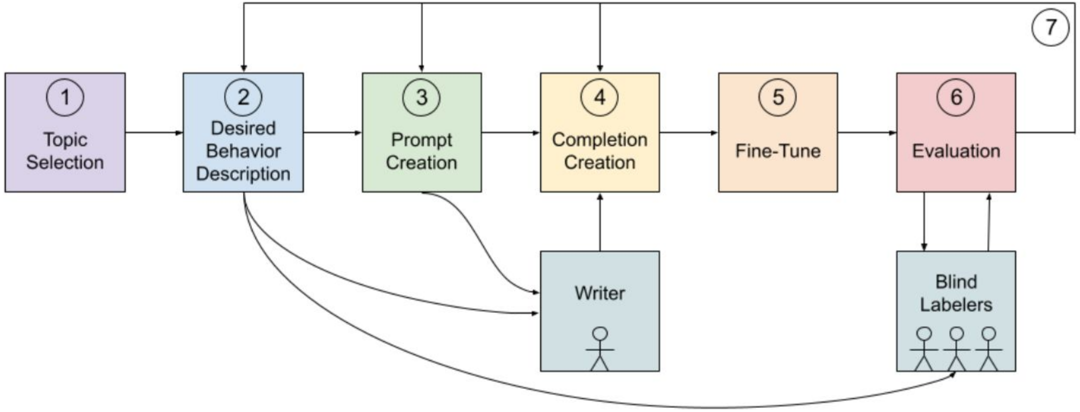


Figure 1: PALMS Steps

Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets

Irene Solaiman* OpenAI irene@openai.com

Christy Dennison* OpenAI christy@openai.com

Abstract

Language models can generate harmful and biased outputs and exhibit undesirable behavior. We propose a Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets, an iterative process to significantly change model behavior by crafting and fine-tuning on a dataset that reflects a predetermined set of target values. We evaluate our process using three metrics: quantitative metrics with human evaluations that score output adherence to a target value, and toxicity scoring on outputs; and qualitative metrics analyzing the most common word associated with a given social category. Through each iteration, we add additional training dataset examples based on observed shortcomings from evaluations. PALMS performs significantly better on all metrics compared to baseline and control models for a broad range of GPT-3 language model sizes without compromising capability integrity. We find that the effectiveness of PALMS increases with model size. We show that significantly adjusting language model behavior is feasible with a small, hand-curated dataset.

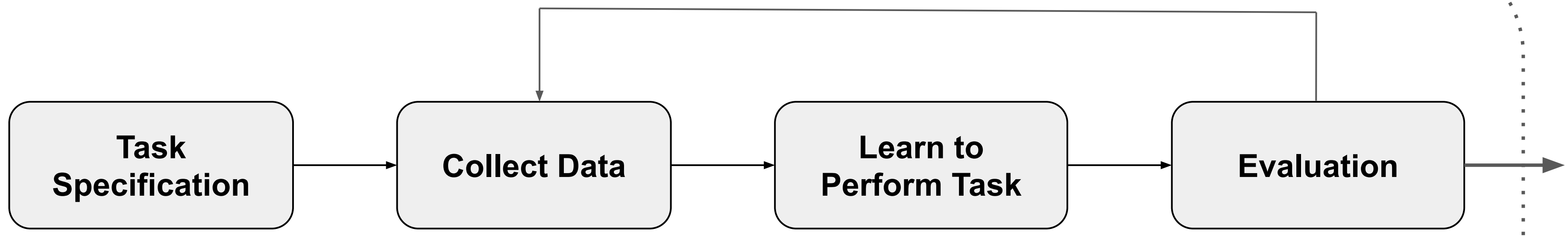
- ➔ Using domain knowledge in the form of data as corrective measure to reduce bias
- ➔ One of many ways to incorporate domain knowledge...

Experimental design → *Task design*

Experimental design is the process of carrying out research in an **objective** and **controlled** fashion so that precision is maximized and **specific conclusions can be drawn regarding a hypothesis statement**. Generally, the purpose is to establish the effect that a factor or independent variable has on a dependent variable.

Important topics germane to experimental design include **hypothesis statements**, **experimental control**, **specifying independent and dependent variables**, selection and assignment of samples or participants to conditions, **collecting data**, and **selecting valid statistical tests**.

Experimental design → *Task design*



Task Specification
- Description
- Should it be conducted?
- How to evaluate success?
- Cost predictions
- Potential failures

Collect Data
- Benchmarks
- Intervene on data
- Privacy, ownership?
- Datasheets

Learn to Perform Task
- Often includes many phases
- Statistical models in different parts

Evaluation
- Qualitative demo
- Human eval
- Proxy metrics
- Success & adequacy metrics

NEW TASK MODELING PIPELINE

Conclusion

- ➔ A lot of exciting developments in AI
- ➔ Two separate communities: statistical modeling, task modeling
- ➔ Task modeling community is driving most of the breakthroughs
- ➔ Task modeling has several shortcomings that:
  - ◆ limit the deployment of AI systems to critical domains
  - ◆ can negatively affect different communities
- ➔ Lessons from statistical modeling can help alleviate those shortcomings
- ➔ Need for convergence between the two communities for practically safe AI that can conquer new domains