Logistics

- Project proposals are due by midnight tonight (two pages)
 - Email to instructors
 - We will send feedback by next week
- ➤ In class proposal presentations on Monday 3/11 (two weeks)
 - 4 minutes each
 - Presentations should briefly answer the following questions
 - > What is the problem and why is it important
 - ➤ What are the key limitations of previous work
 - What is your proposed approach

AUtoML

"Democratizing ML"

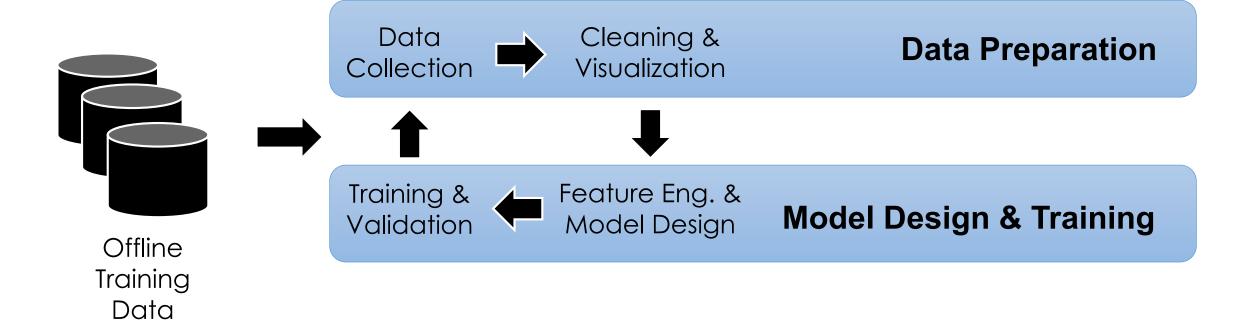
Joseph E. Gonzalez

jegonzal@cs.berkeley.edu

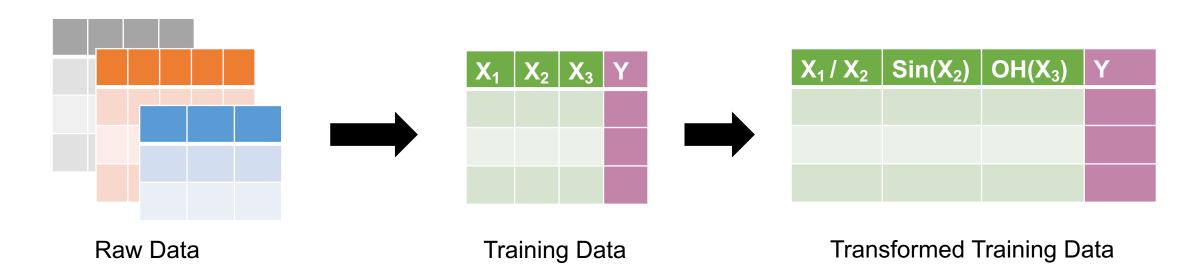
Machine Learning Lifecycle

Training Inference **Model Development Prediction Service** Cleaning & Query Visualization Prediction Trained **End User** Training Pipelines Feature Eng. & Models **Application** Validation Offline Training Feedback Validation Data

Model Development



Model Development: Data Preparation

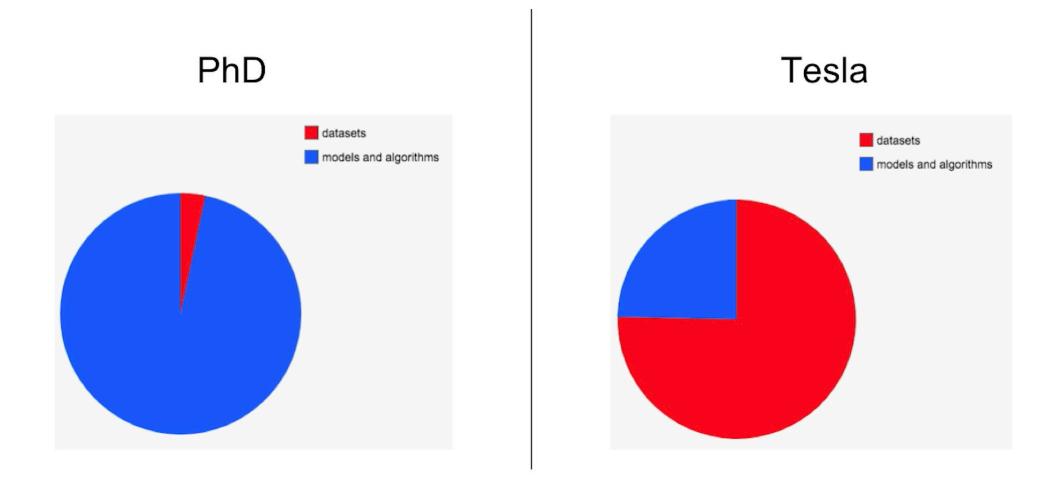


Requires substantial domain expertise:

- Where are the data?
- What do the columns mean?
- How should they be coded?

Andrej Karpathy

Amount of lost sleep over...



Model Development: Design and Training

Model Parameters are "fit" during training.

Model:
$$f_{\theta,s}(x) \to y$$

Tuned Using Expert Knowledge

What Kind of Model

Hyper-Parameters

Classic Models

- Linear Model
- > SVM
- Decision Tree
- Random Forrest

> ...

Neural Network

- > Conv. Net
- > Recurrent Net.
- > Auto Encoder.
- Graph Net.
- **>** ...

Model

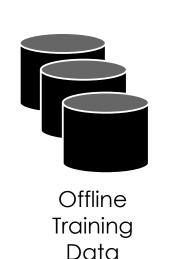
- > Regularization
- > Batch size
- > Structural Parameters
 - > Hidden Units
 - Activations
 - ➤ Layer design

Algorithm

- Training Alg.
- > Learning Rate
- > Parallelism

Tuned using Hyper-Parameter Search

Model Development Technologies



Data Collection



Cleaning & Visualization





Training & Validation

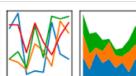


Feature Eng. & Model Design

Data prep. and feature engineering















Model design and Training















Systems Research Opportunities

- Accelerate data collection and preparation
 - > Automatic data discovery
 - Distributed data processing, esp. for image and video data
 - Data cleaning and schema driven auto-featurization
- Accelerate model selection and hyper-parameter search
 - Parallel and distributed execution
 - Data and feature caching across training runs
- Provenance
 - > Track previous model development to inform future decisions
 - Connect errors in production with decisions in model development