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

Model Development

Data Collection
→
Cleaning & Visualization
↓
Feature Eng. & Model Design
→
Training & Validation

Training

Training Pipelines
→
Trained Models
→
Validation

Inference

Prediction Service
Logic
Feedback

Query
Prediction
End User Application

Offline Training Data

Live Data

Trained Models
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...

PhD

Tesla
Model Development: **Design and Training**

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

Model Parameters are "fit" during training.

**What Kind of Model**

- Linear Model
- SVM
- Decision Tree
- Random Forest
- ...  

**Hyper-Parameters**

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

**Tuned Using Expert Knowledge**

**Classic Models**

- Linear Model
- SVM
- Decision Tree
- Random Forest
- ...

**Neural Network**

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

**Algorithm**

- Training Alg.
- Learning Rate
- Parallelism

Tuned using Hyper-Parameter Search
Model Development Technologies

Data Collection → Cleaning & Visualization
Data prep. and feature engineering

Offline Training Data

Training & Validation ← Feature Eng. & Model Design

Model design and Training

Pandas
Dask
Spark
Hive
Presto

PyTorch
TensorFlow
XGBoost
Keras
Caffe2
scikit-learn
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