

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

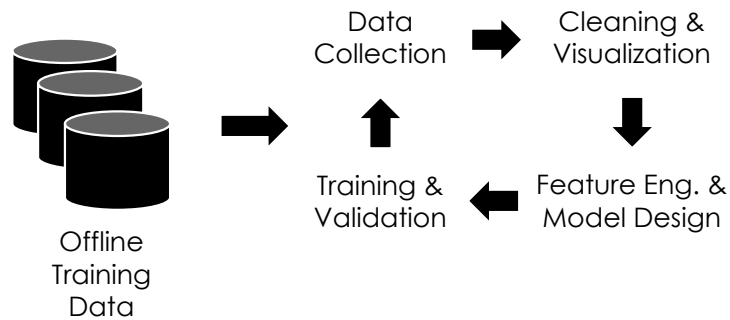
“Democratizing ML”

Joseph E. Gonzalez

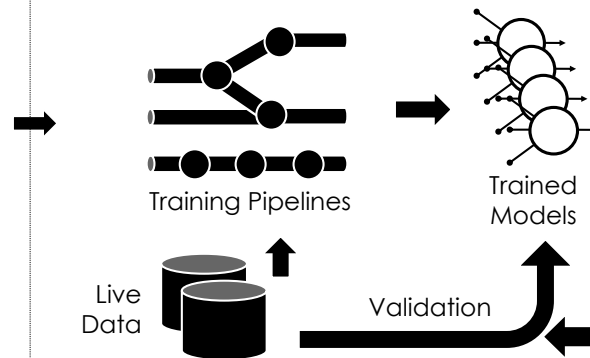
jegonzal@cs.berkeley.edu

Machine Learning Lifecycle

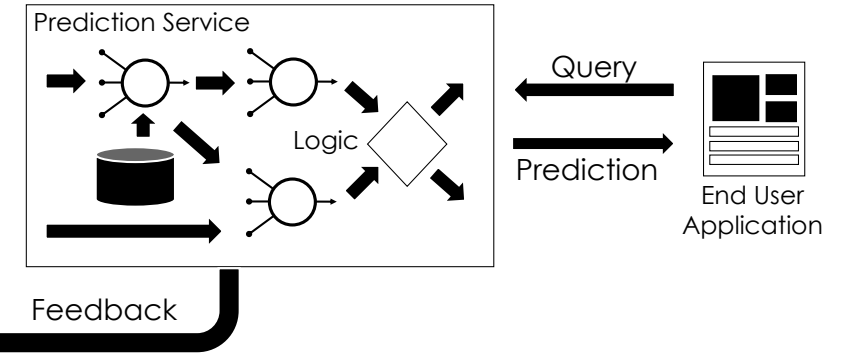
Model Development



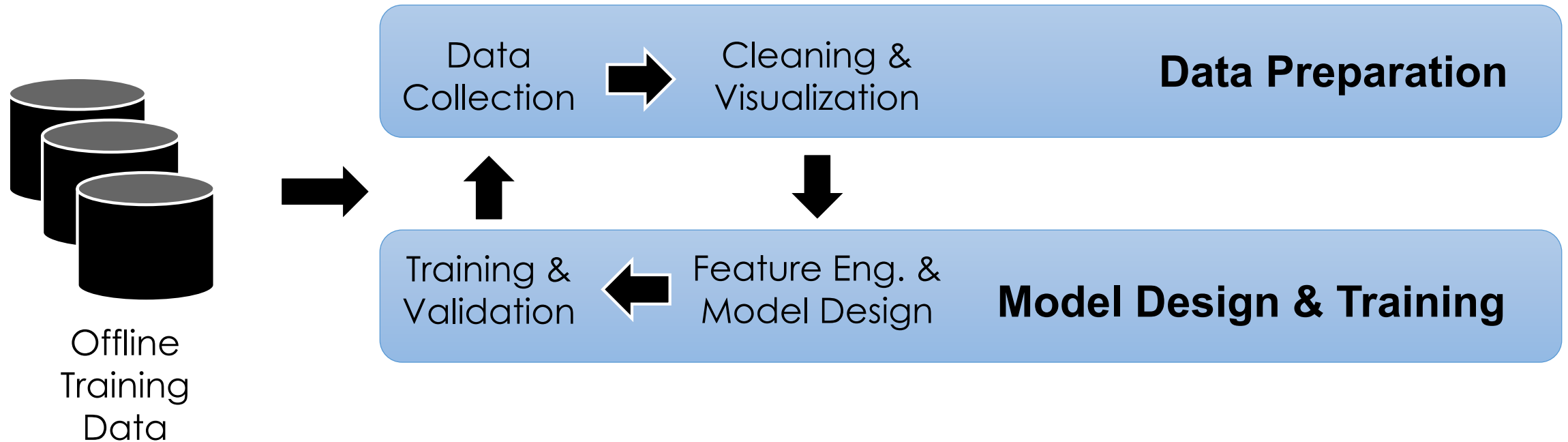
Training



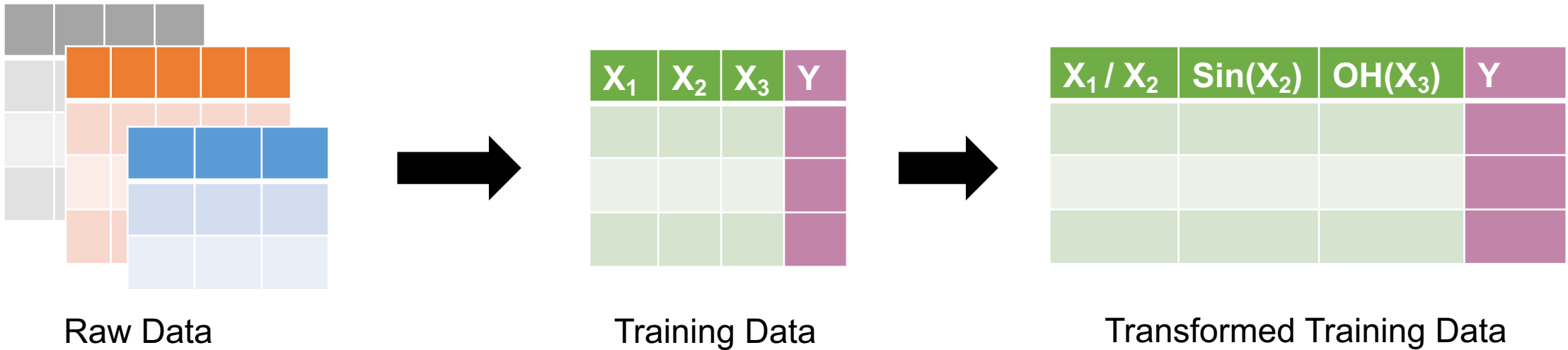
Inference



Model Development



Model Development: Data Preparation



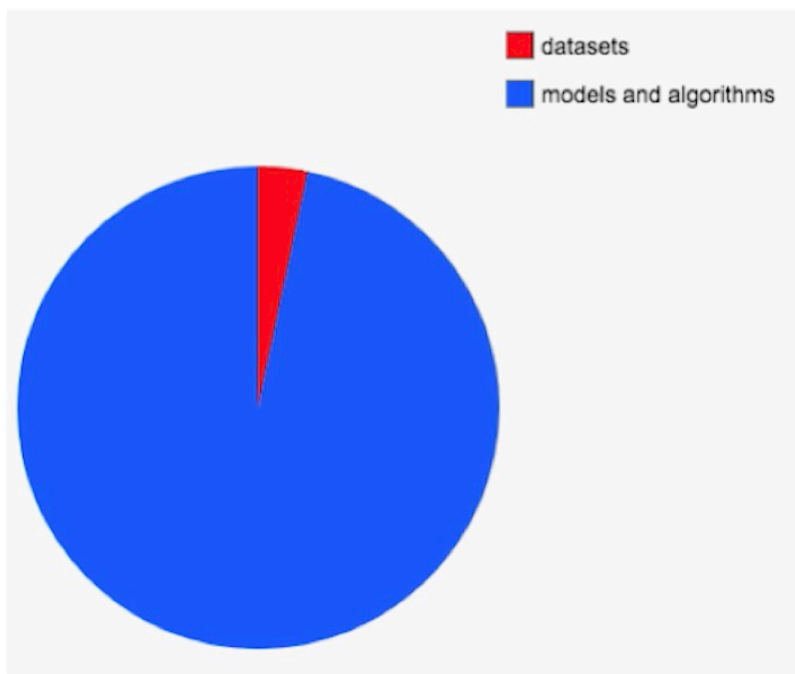
Requires substantial domain expertise:

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

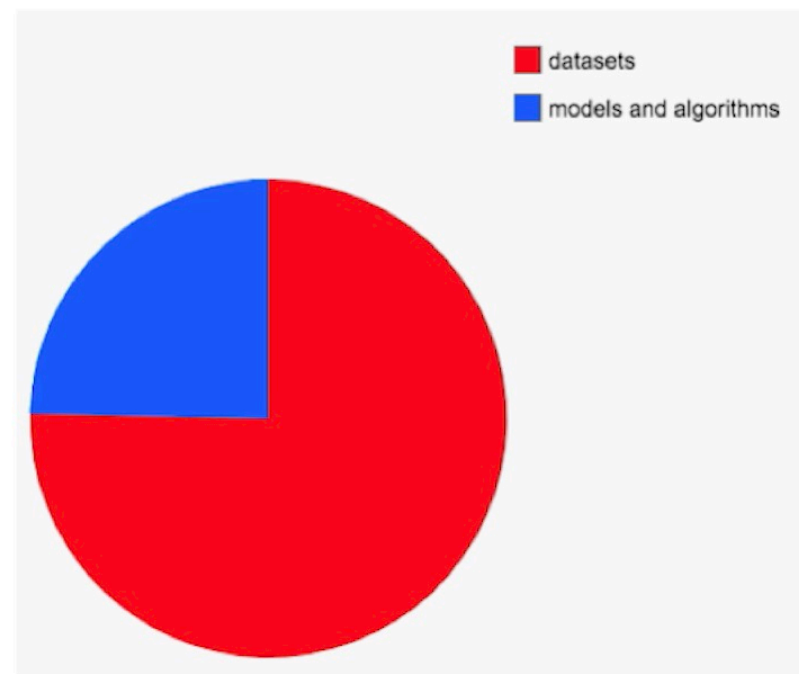
Andrej Karpathy

Amount of lost sleep over...

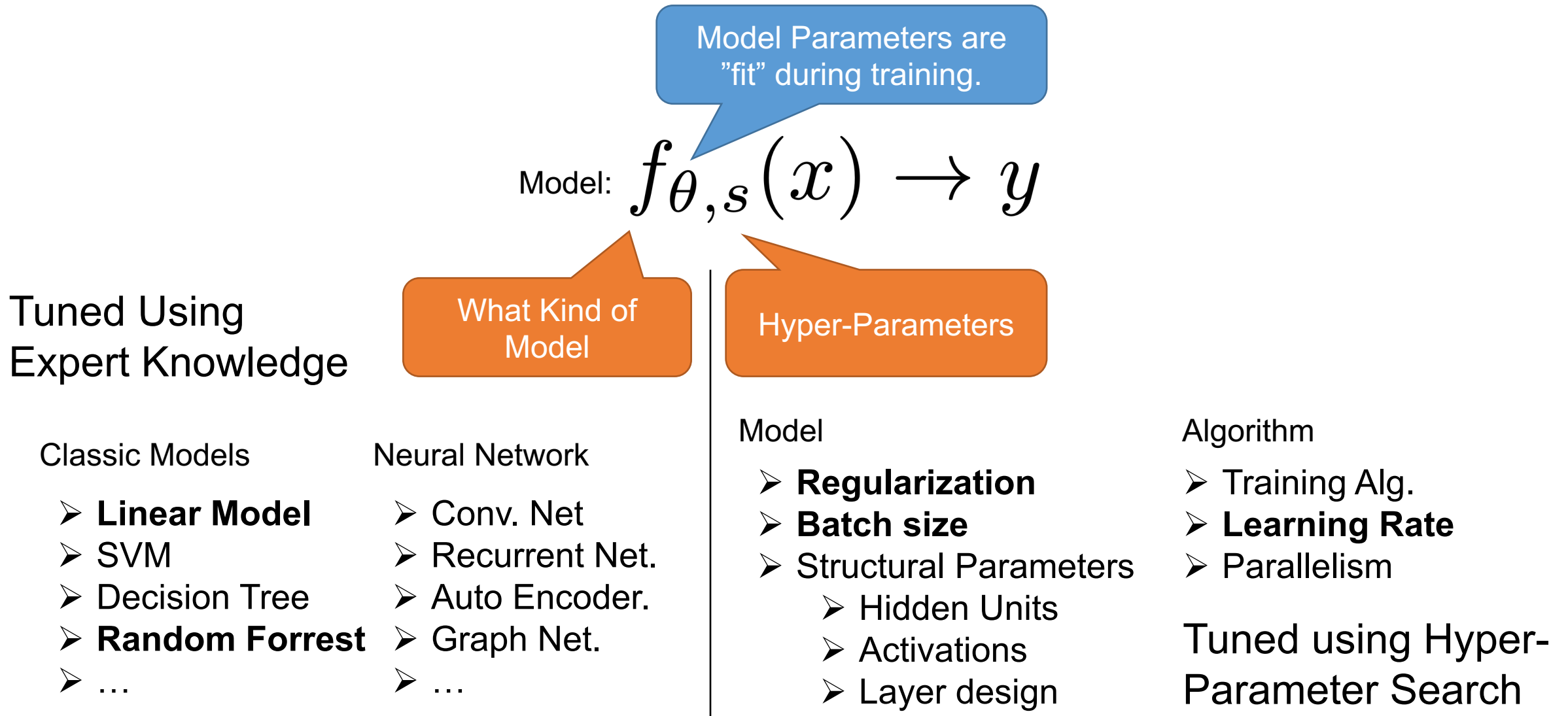
PhD



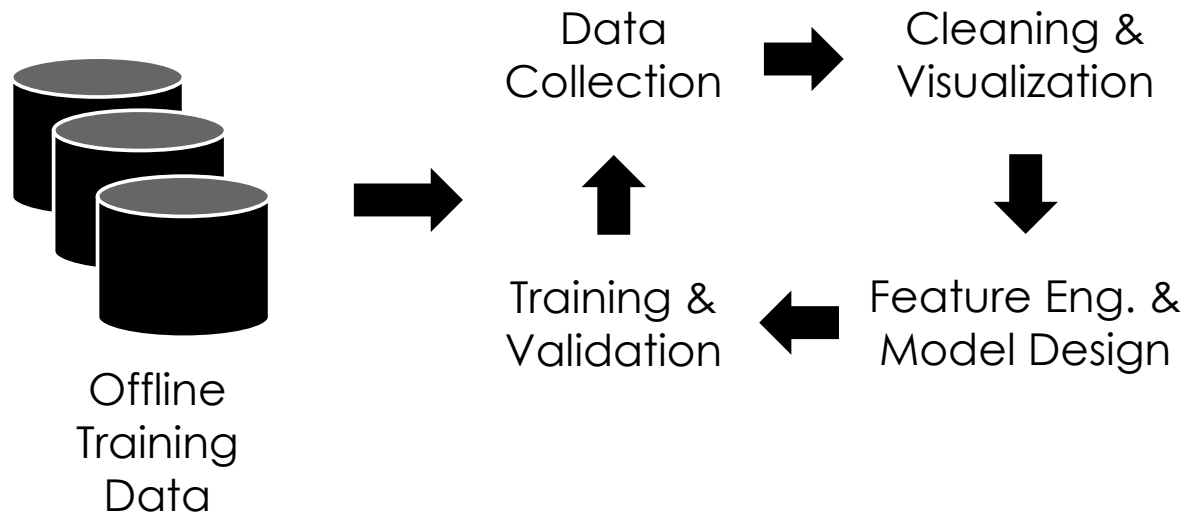
Tesla



Model Development: Design and Training

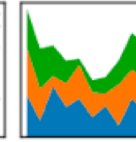
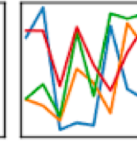


Model Development *Technologies*



Data prep. and feature engineering

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



presto

Model design and Training



PYTORCH



dmlc
XGBoost



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