Scaling ML in Ad Tech

Giri Iyengar

Agenda

Introduction

What are AdTech Platforms?

Big Data in Ad Tech

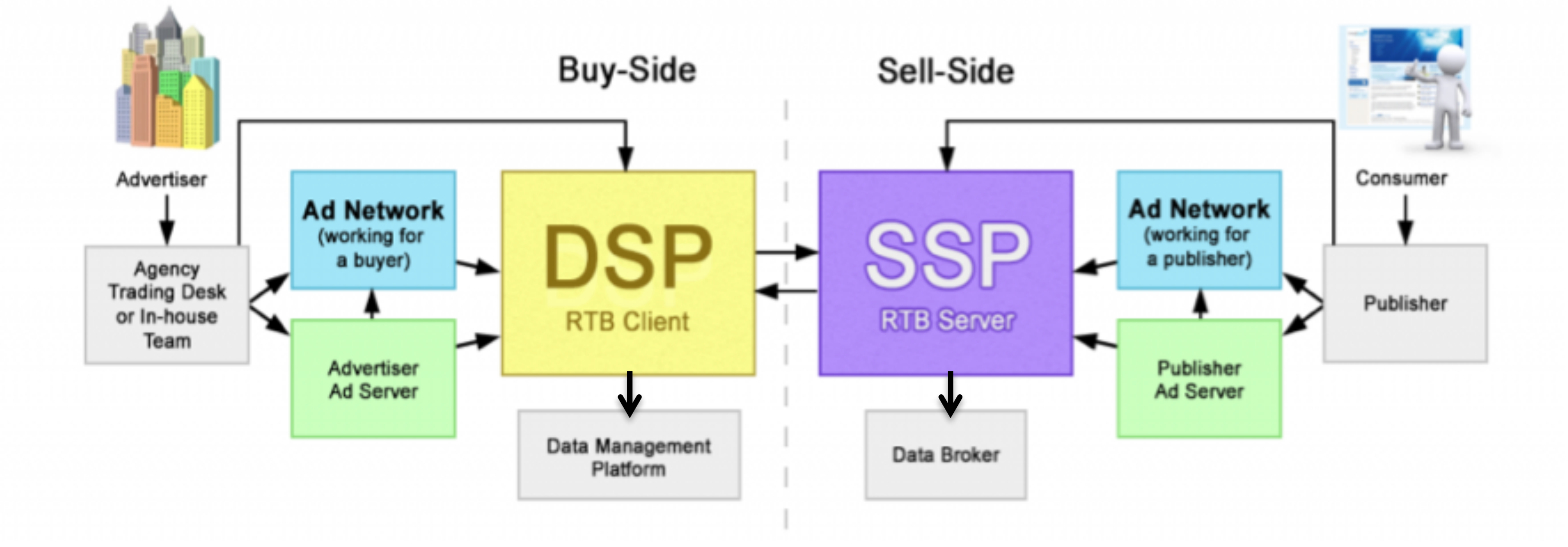
Some Data Science Projects in Ad Tech

Technical & Operational Challenges

In Search of an ML Platform

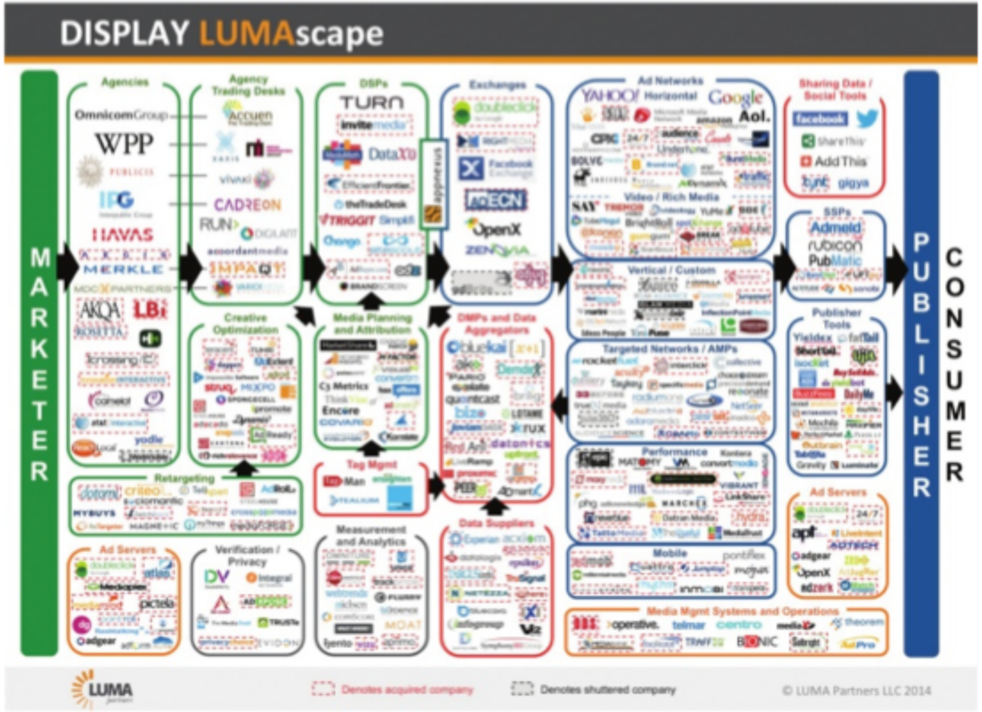
What AdTech Platforms Do

- Service both the Buy Side and Sell Side in Ad Tech Marketplace and Direct Sales
- Ad Network, SSP, DSP, DMP in Display, Video and Mobile
- Pictela, OutBrain and other interesting products and services
- Publishers
  - Huffington Post
  - NYTimes
  - Conde Nast



Big Data In Ad Tech

- Many sources of data across the Buy Side and Sell Side
- Imp/clicks/conversion
- Video events
- Audience data - online, offline, 1st party, 3rd party, BYOD
- RTB Bid Requests & Responses
- Ad information - viewability, sizes, creatives
- Advertiser, Site, and Campaign information

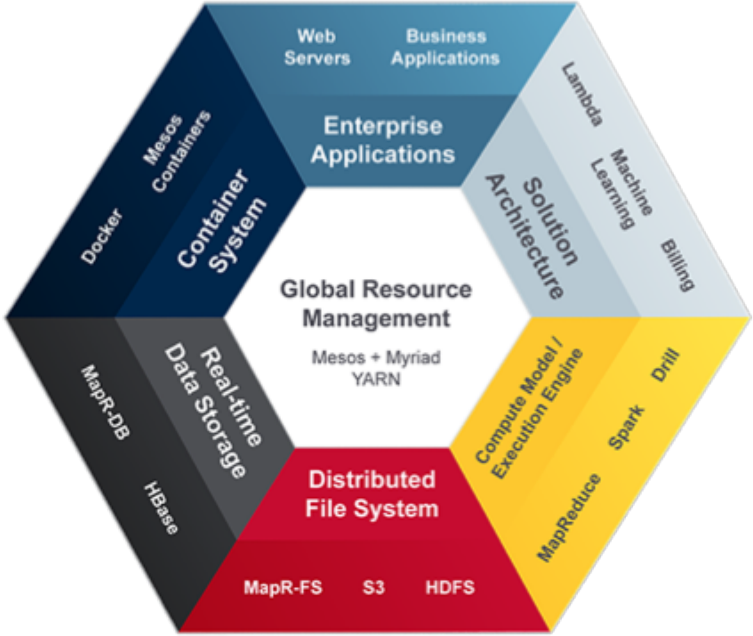


Some Statistics
100s of TBs of data daily
10s of Petabytes of clustered data
Millions of requests per second across systems
Billions of devices reached
ms latencies for responses

Zeta Architecture

- Latest evolution of Big Data best practices.
- Combines architectural practices with operational practices (DevOps & DataOps)
- Goal is to reduce architectural complexity
- Lambda is a sub-architecture within the overall Zeta architecture

- **Distributed File System** - all applications read and write to a common, scalable solution, which dramatically simplifies the system architecture.
- **Real-time Data Storage** - supports the need for high-speed business applications through the use of real-time databases.
- **Pluggable Compute Model / Execution Engine** - delivers different processing engines and models in order to meet the needs of diverse business applications and users in an organization.
- Container and cluster management.



What is Data Science @ AdTech

Scalable Data Science combines:

- Statistics
- Machine Learning
- Big Data Engineering
- Optimization
- Advanced Data Structures

In order to:

- Build business-driven, production-ready machine learning systems at large scale
- Examples include:
  - Buy Side (e.g., real-time bidding optimization, MTA, viewability, predictive segments)
  - Sell Side (e.g., publisher yield optimization, user interest graph)
  - And Everything In Between (e.g., geo-fencing, probabilistic device linking, User-device linking)
- Analyze and model large data sets (100s of TBs to PBs) to pull signal out of the noise

Some Machine Learning Projects in AdTech

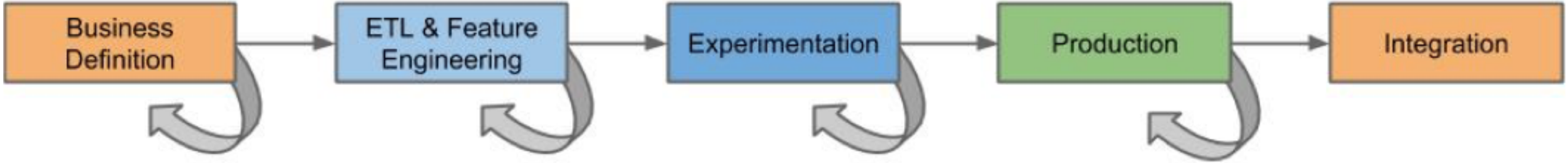
Cross-device User Matching
Process request, impression, click, deterministic link data and probabilistically link devices
Perform connected components analysis on device links to build User-Device graphs
Several billions of rows a day scored daily
Training & Scoring requires ETL and feature engineering across several types of data sets

Price Floor optimization
Process bid data coming into our SSP and recommend price floors by publisher / site / country /demand partner / user
Typically a few hundred billion bid requests & corresponding responses need to be processed daily

Building an Interest Graph
Gravity, an AOL company, provides recommendation widgets on publisher sites
The people behind the popular Goose URL parser
Millions of URLs classified daily

CTR Prediction
Process User Profiles, Bid Request information to predict Click-through-rate (pCTR) for a given request
Has to be **well-calibrated** as our bidding system price the bids based on this
10s Billions of Bid Requests need to be handled daily

Data Science Process



Business Definition

Business Problem Definition
Product-driven
Scoping
Clear Goals/KPIs

ETL & Feature Engineering

Data Cleansing
Normalization
Feature extraction (**This is the Art!**)

Experimentation

Hyperparameter Optimization
Cross-validation
Training Data Transformations
Model Family <-> Feature Sets
Model Tournament
Winning Model Selection

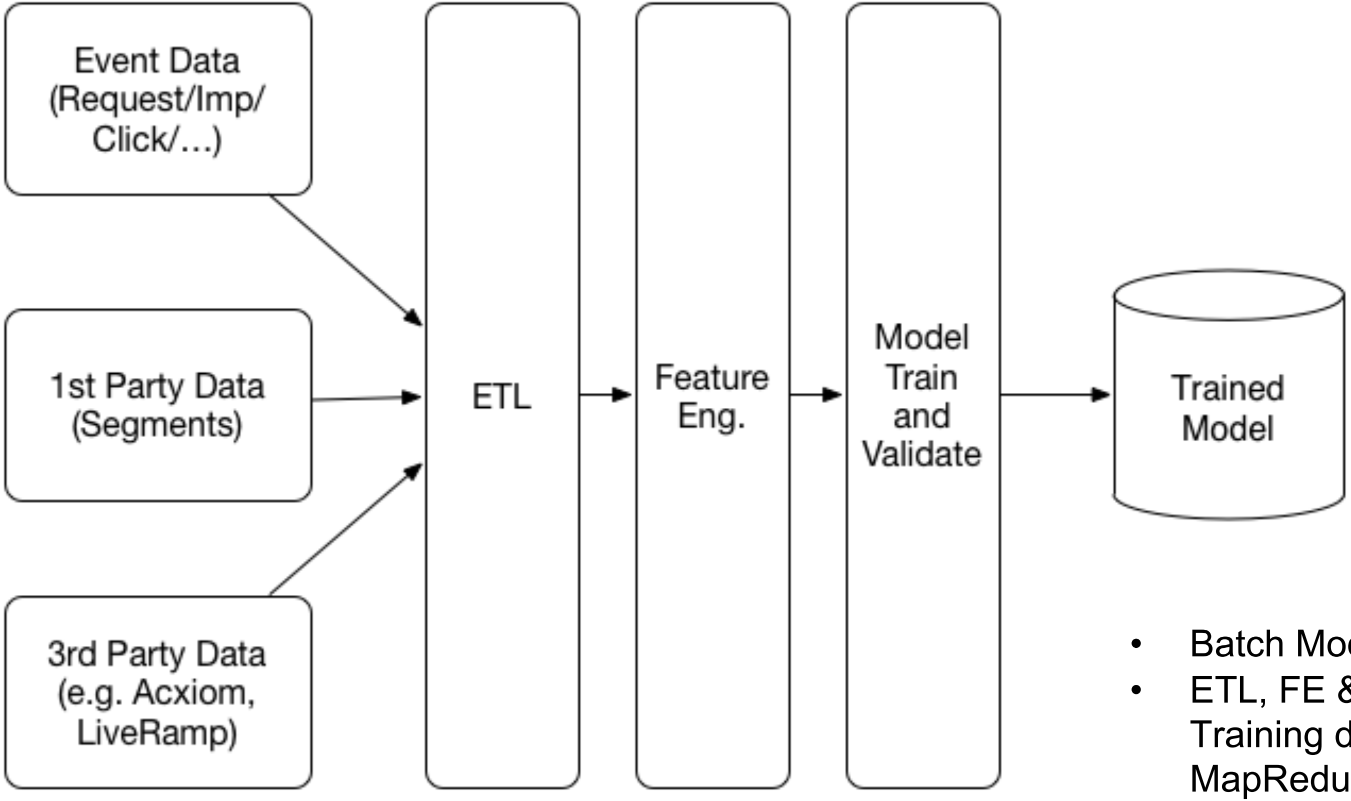
Production

Performance Monitoring
Training workflow
Adaptation and Relearning
Schedules

Integration

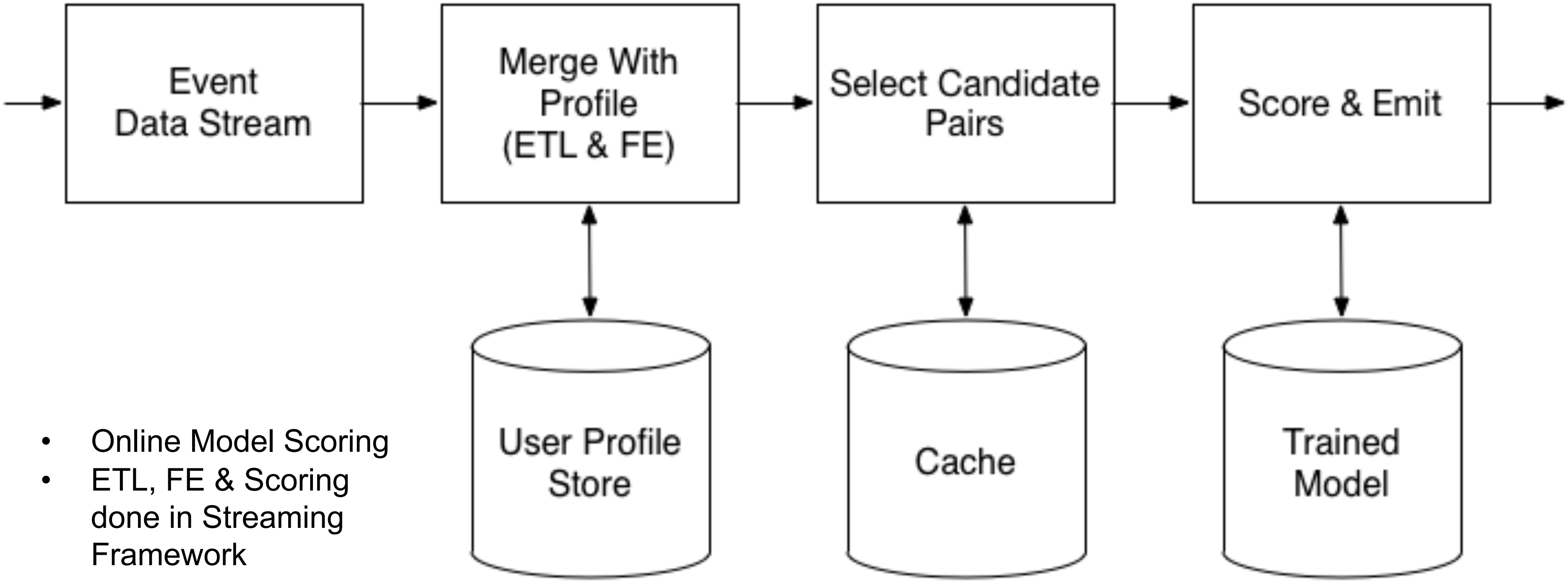
Loose integration with
Decisioning System
Model Files -> Production DB
etc.

Cross-Device User Matching


Event Data (Request/Imp/ Click/...)
1st Party Data (Segments)
3rd Party Data (e.g. Acxiom, LiveRamp)
ETL
Feature Eng.
Model Train and Validate
Trained Model

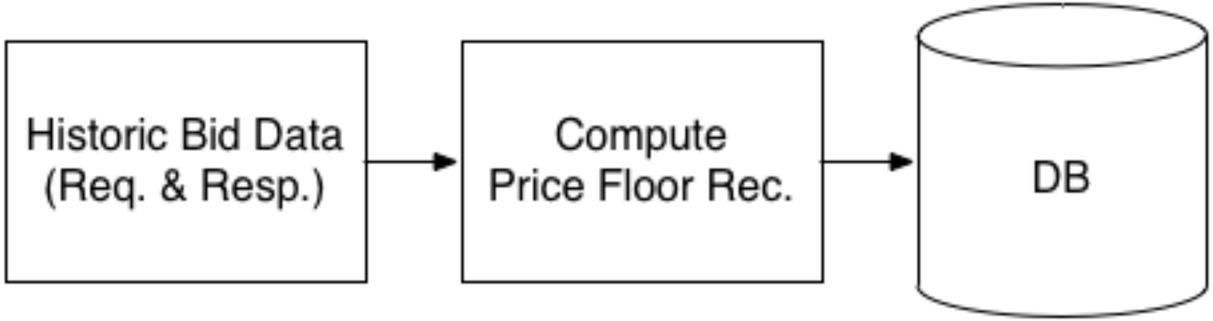
- Batch Model Training
- ETL, FE & Model Training done in MapReduce

Cross-Device User Matching



- Online Model Scoring
- ETL, FE & Scoring done in Streaming Framework

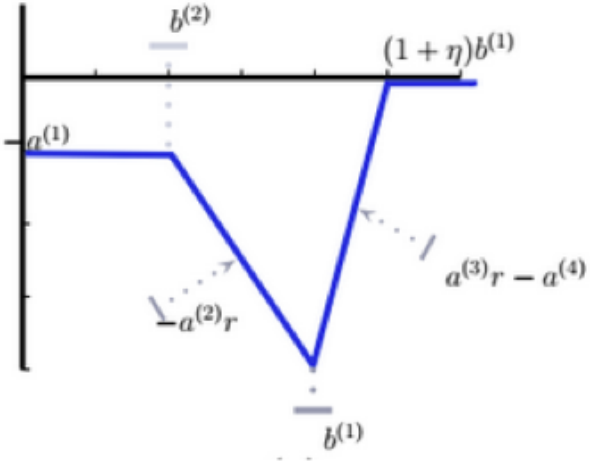
Price Floor Optimization



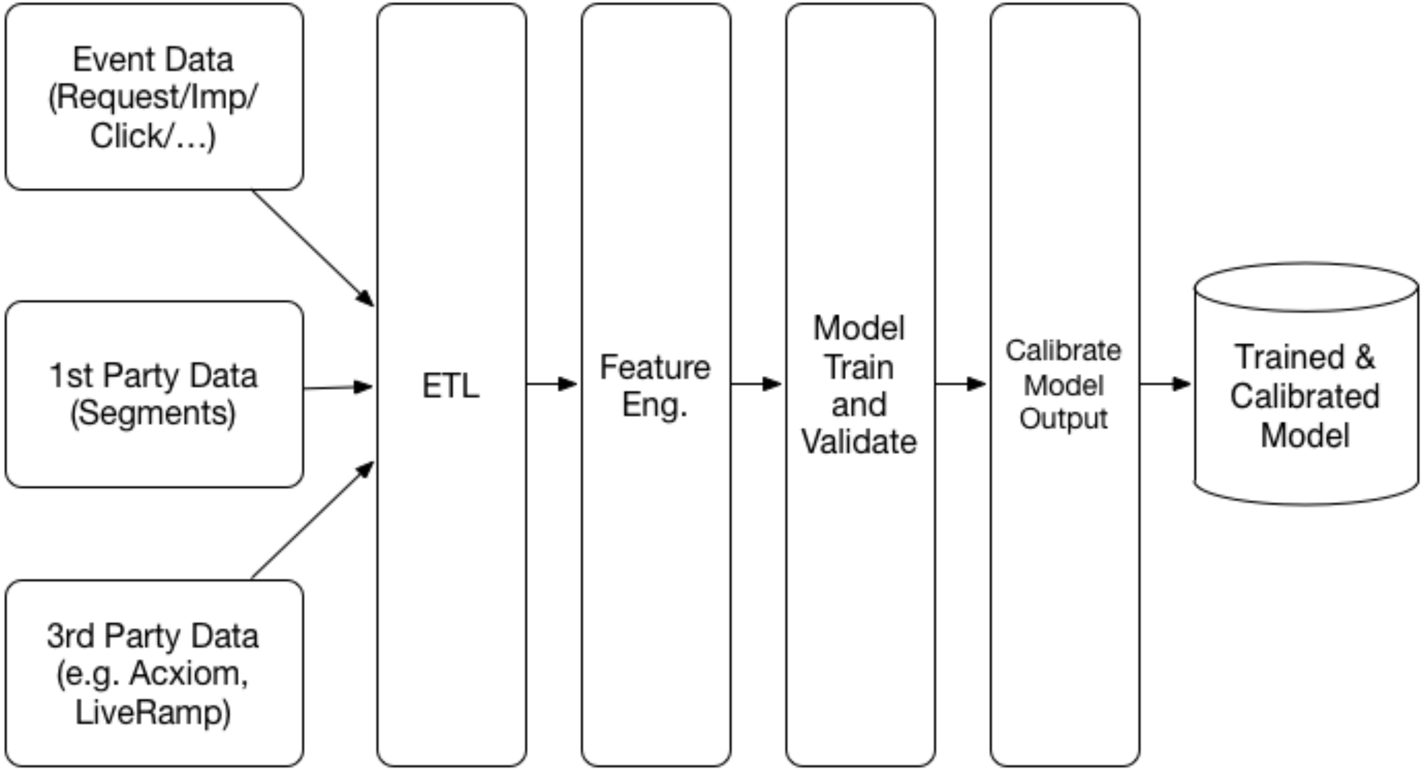
Historic Bid Data (Req. & Resp.)

Compute Price Floor Rec.

DB



Click Prediction


Event Data (Request/Imp/ Click/…)
1st Party Data (Segments)
3rd Party Data (e.g. Acxiom, LiveRamp)
ETL
Feature Eng.
Model Train and Validate
Calibrate Model Output
Trained & Calibrated Model

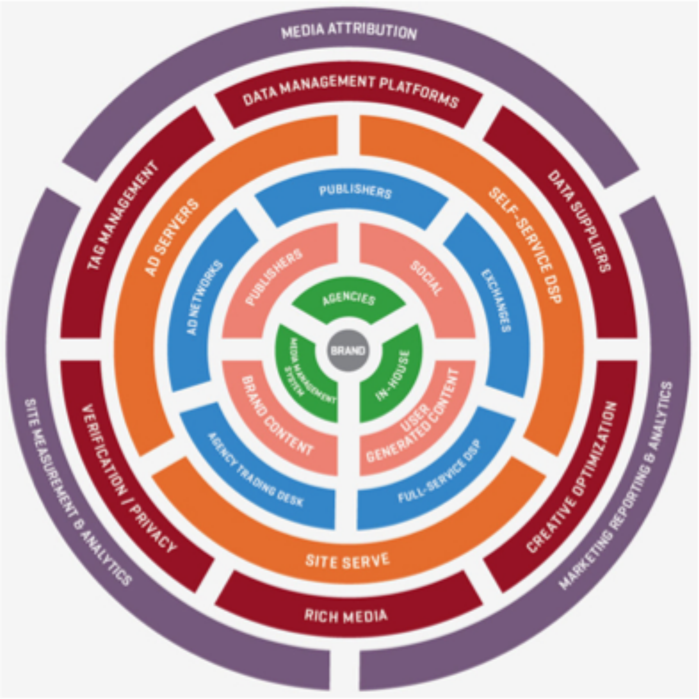
* Very Similar to the UMS Training Flow. When Models are operating in the real world, Calibration becomes important

Technical & Operational Challenges

- Discovering & managing disparate data sources
- Creation and curation of data sets and make them ML-ready
- Feature Engineering -- deep subject matter expertise is often required
- Experimentation & Hyper Parameter Optimization -- the cycle time is too long and so we cannot scale up our search
- Deploying the models to production -- Often times we have to do custom engineering to meet latency requirements
- Ongoing monitoring, retraining models, rolling out new models
- Sharing Features and Outputs across models
- Manual overrides to compensate for revenue target misses

Organizational Challenges

- Ad Tech organization is large
- Upstream data collection practices and a holistic data strategy have a big impact on Data Science agility
  - Data quality/data correctness - 70% of the scientist's time is spent cleansing data!
- Everyone should understand data science process in order for us to build more effective products and systems faster
- Clear definition of the product goal or problem to be solved



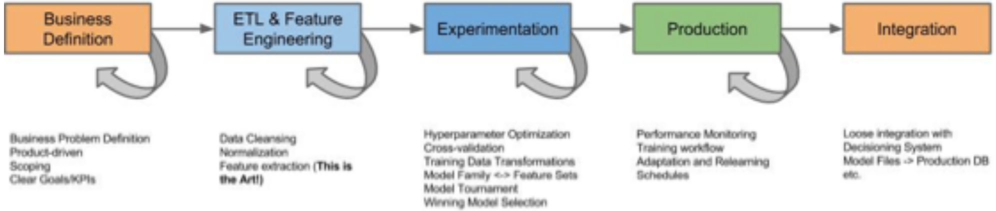
In Search of an ML Platform

Some tools used in AdTech companies

- R, Python 2, Python 3
- Scala, Java, C++, C#
- Vowpal Wabbit, Sofia-ML
- Pig UDFs and scripts
- H2O, Caffe
- Nitro
- Jupyter / R-Markdown / Confluence

Requirements for an Ideal ML Platform

- ETL and Data Exploration
  - Data Source Integrations (e.g. HDFS, HBase, SQL and NoSQL stores)
  - Data Normalizations & ETL Capabilities
  - Exploratory Data Analysis
- Feature Engineering
  - Feature Engineering Support
  - Automatic Feature Detection
- Models and Experimentation
  - Model Building
  - Hyper Parameter Optimization
  - Model Tournaments
- Production & Operations
  - Exporting Models / Rolling Models to Production
  - Model Retraining Automation
  - Big Data Scale
  - Monitoring model performance in Production
  - Closed loop evaluations of models
  - Execution monitoring / logging (e.g. record counts, duration, etc.)
- Reuse and Collaboration
  - Sharing Sub-flows Across Projects
  - Collaborating between teams / Groups on Projects
  - Sharing analysis via Notebooks
  - Sharing analysis to Product and Business



Business Problem Definition
Product-driven
Scoping
Clear Goals/KPIs

Data Cleansing
Normalization
Feature extraction (**This is the Art!**)

Hyperparameter Optimization
Cross-validation
Training Data Transformations
Model Family <-> Feature Sets
Model Tournament
Winning Model Selection

Performance Monitoring
Training workflow
Adaptation and Relearning
Schedules

Loose integration with
Decisioning System
Model Files -> Production DB
etc.

This will be easy. I'll just define my cost function and then minimize it.