Data Science with Big Data in a Corporate Environment

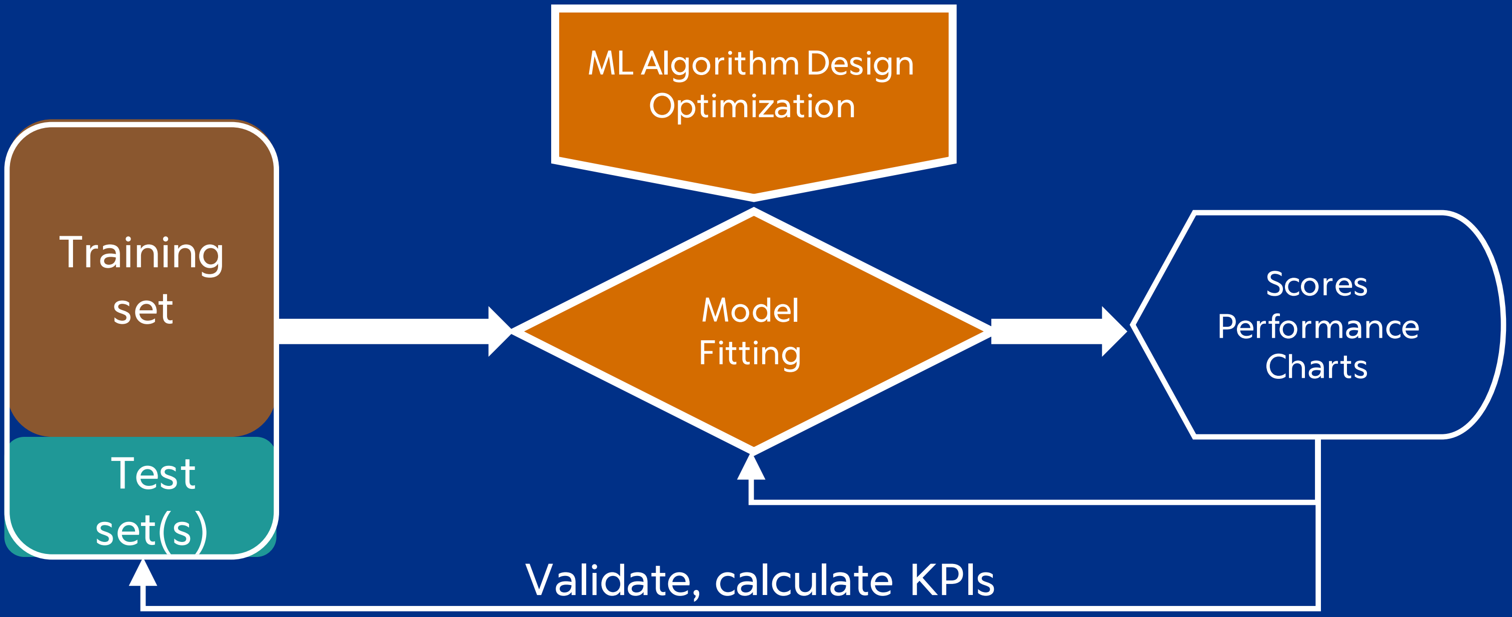
Tasks, Challenges, and Tradeoffs

Nachum Shacham
Principal Data Scientist
PayPal

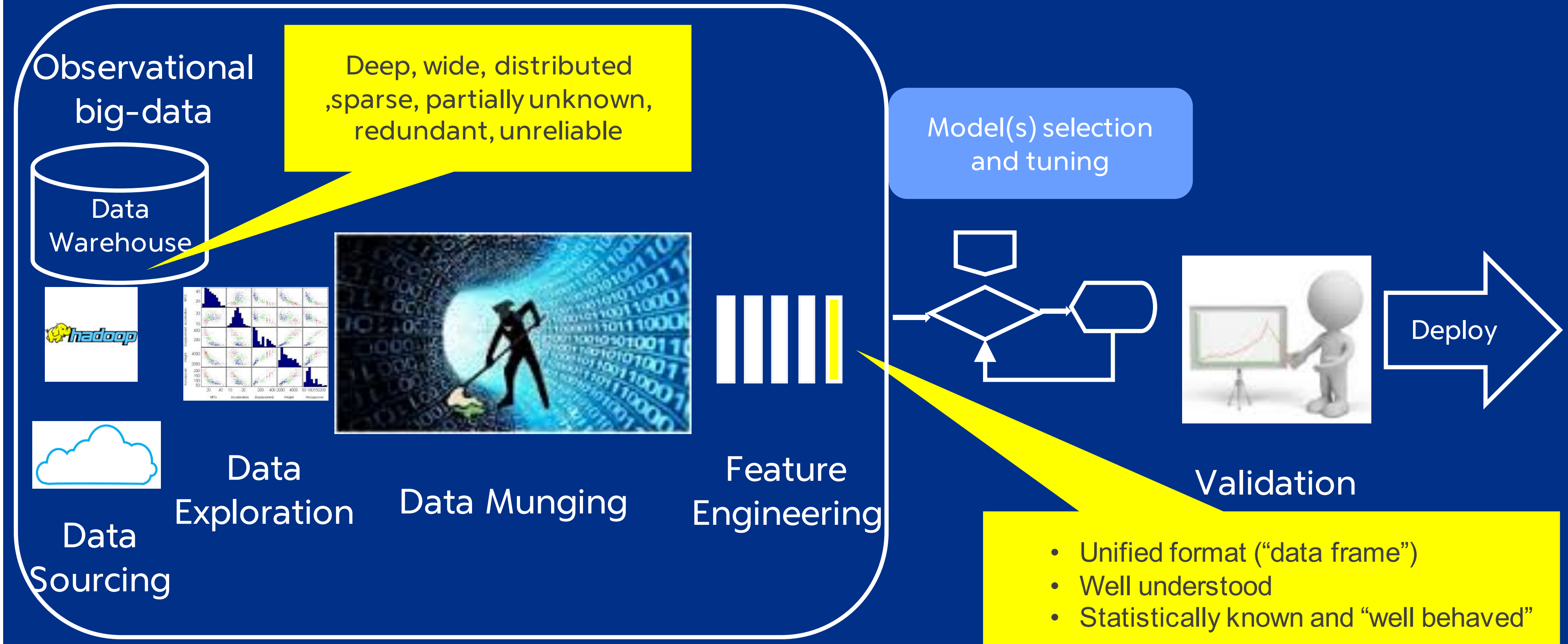
11/10/2015

PayPal

Data Science From 10K ft



Predictive Modeling Meets Big Data: Bridging the Gap



Observational big-data

Deep, wide, distributed ,sparse, partially unknown, redundant, unreliable

Data Warehouse

Data Sourcing

Data Exploration

Data Munging

Feature Engineering

Model(s) selection and tuning

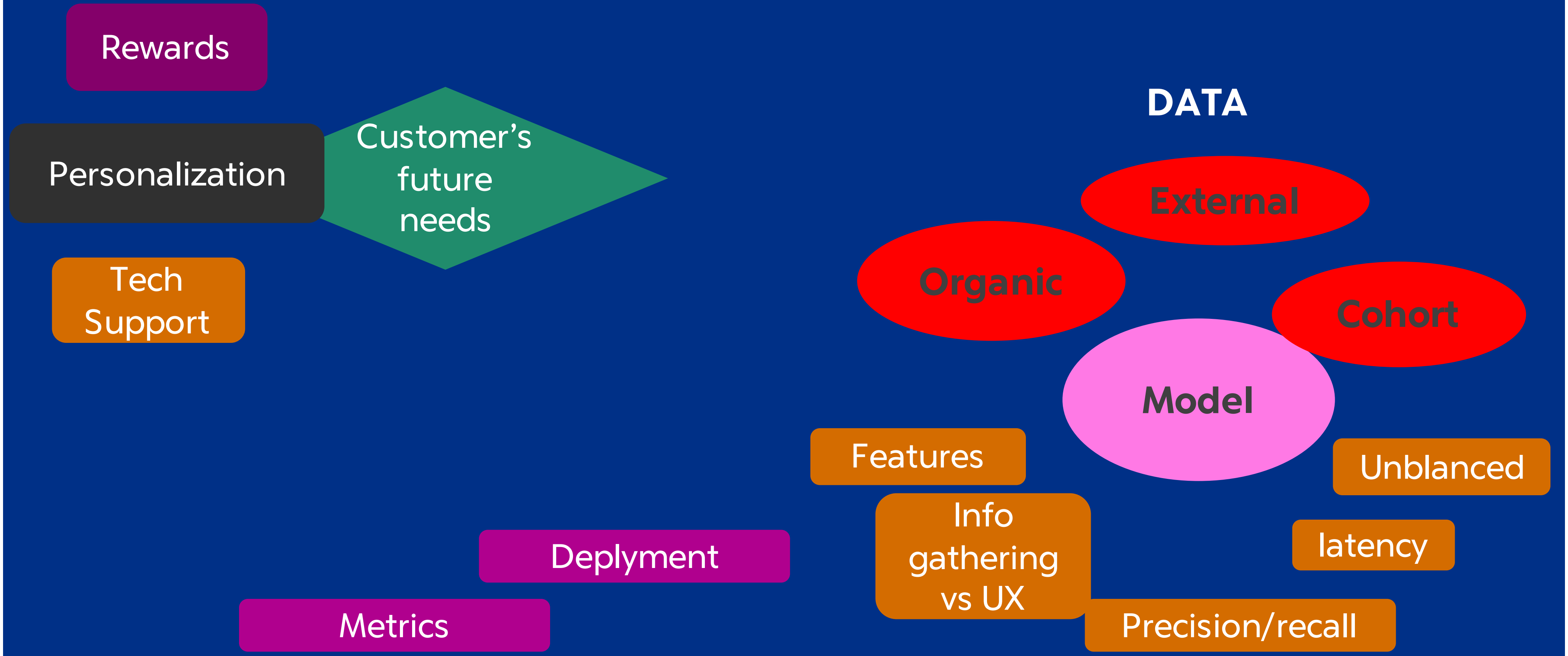
Validation

Deploy

- Unified format (“data frame”)
- Well understood
- Statistically known and “well behaved”

Data Science with Big Data: Quantitative change → Process adjustment

Predictive Models for Serving Customers



Rewards

Personalization

Customer’s future needs

Tech Support

DATA

External

Organic

Cohort

Model

Features

Unblanced

Info gathering vs UX

latency

Deplyment

Metrics

Precision/recall

Data Exploration: Knowing What’s in the Data and Fitting the Pieces Together

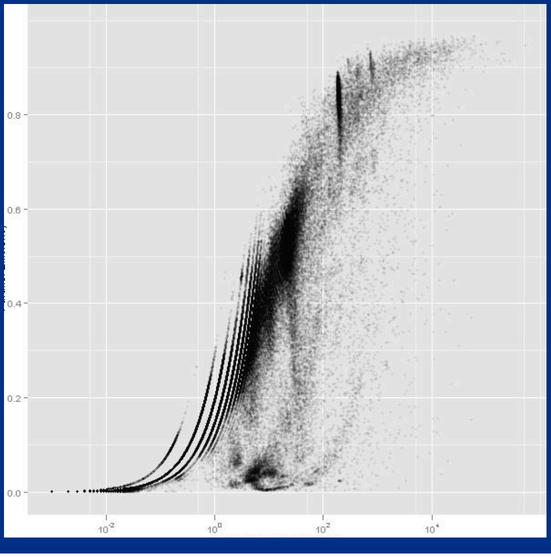
Understaning Data Contents & meaning

Distributions and outliers

Null values: codes, quantities

Language

Join keys

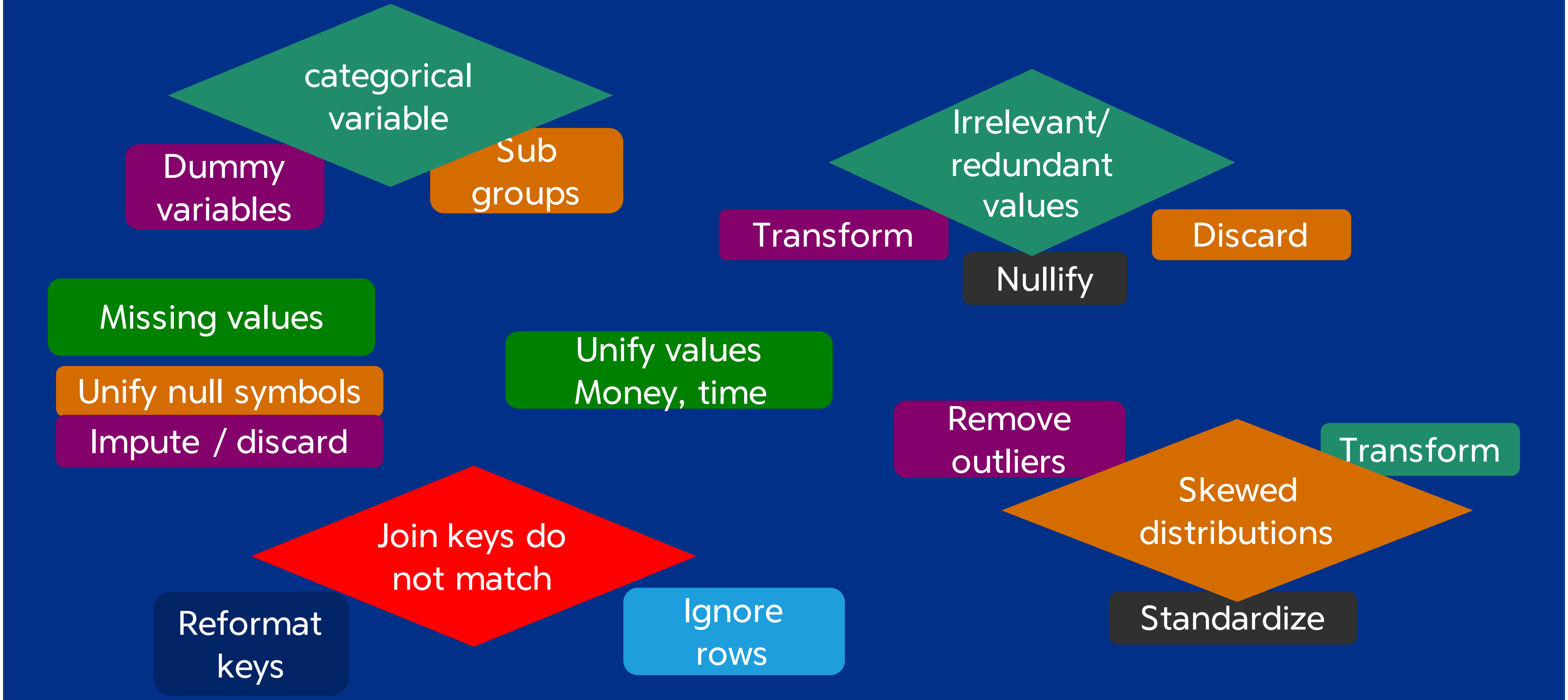


Units

Big data visualization

Correlations

Data Munging: Reshaping the Data → Decisions, Decisions



categorical variable

Dummy variables

Sub groups

Irrelevant/ redundant values

Transform

Nullify

Discard

Missing values

Unify null symbols

Impute / discard

Unify values Money, time

Remove outliers

Skewed distributions

Transform

Standardize

Join keys do not match

Reformat keys

Ignore rows

PayPal

Data Munging Packages & Functions

- R
  - data.table
  - Dplyr
  - Lubridate
  - complete
  - {t, l, s}apply
  - {g}sub
  - grep{l}
  - ggplot2

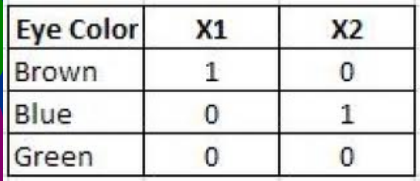
- Python
  - pandas
  - re
  - Numpy
  - Matplotlib
  - csv
  - datetime

Feature Engineering

Transform raw data to better represent the problem to the model

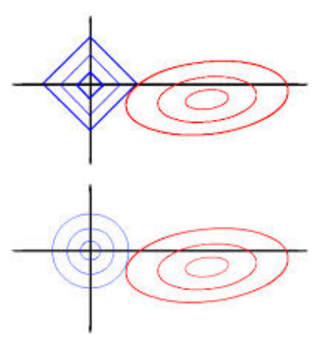
Manual Feature Manipulation

- Dummy variables
- Text integration
- Sub/super sampling (unbalanced sets)

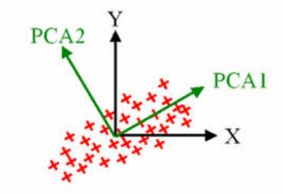


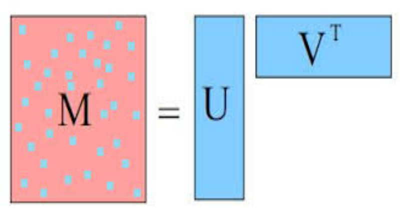
Eye Color	X1	X2
Brown	1	0
Blue	0	1
Green	0	0

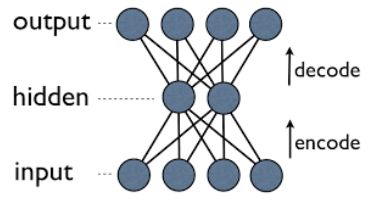
Integrated in model fitting



Feature learning algorithms

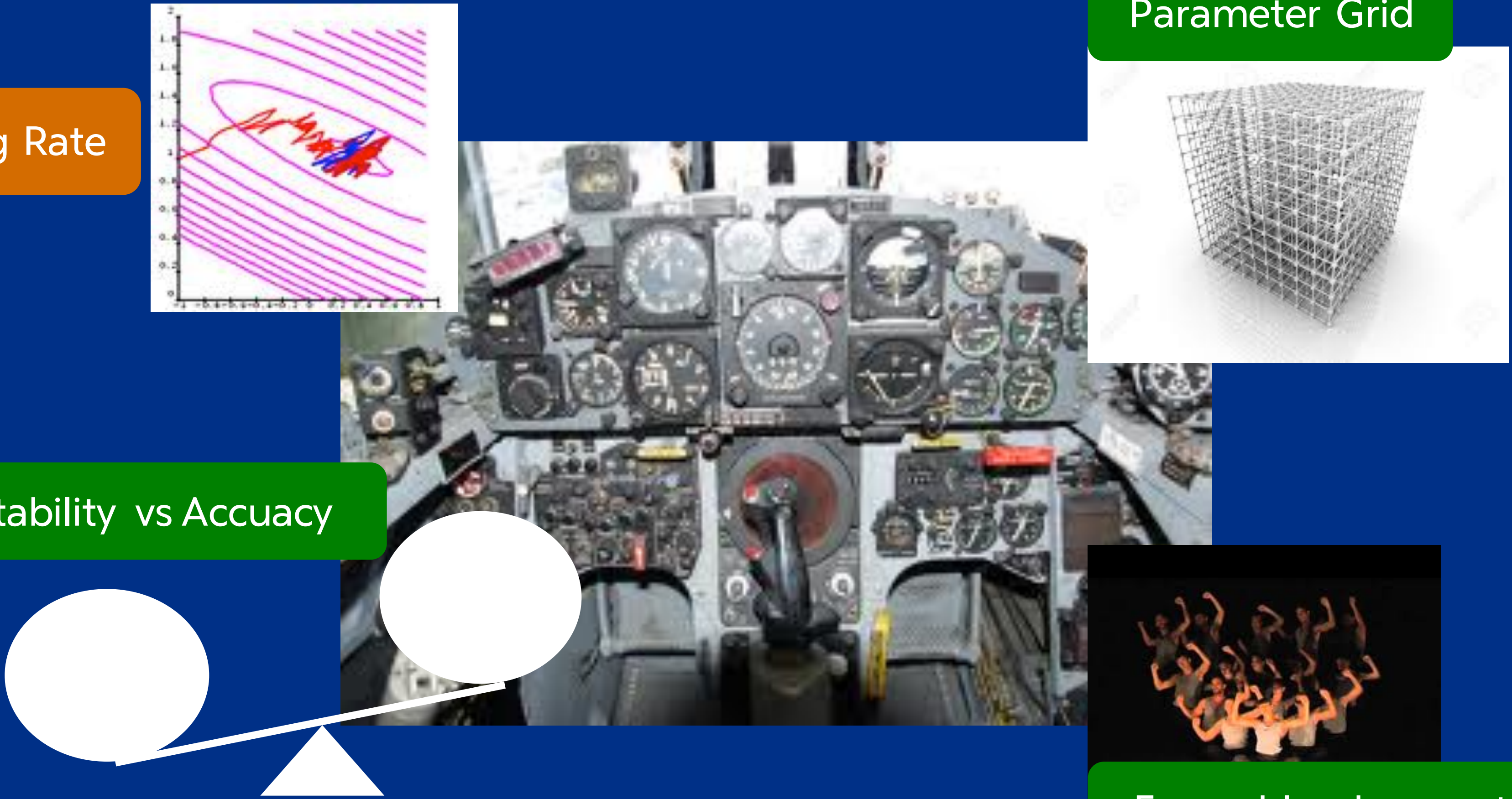






Better features → Simpler models, better results

Model Tuning



Learning Rate

Parameter Grid

Interpretability vs Accuacy

Ensemble: degree, type

Model Building Using H2O

Rich library of ML algorithms

Auto detect feature format

Interoperate with ordinary R: split work for complete munging

Runs through parameter grid

Model Evaluation

- Who should judge the value of the results?
  - Evaluation criteria
- What are the cost/value of FP, TP, FN, TP?
  - FP: PR cost
  - FN: Financial loss
  - TP: Intended gain
- Continuous Model evaluation in production

Model Deployment

- Model results' deployment destinations
  - Humans (decision makers, analysts)
  - Computers (recommendations)
  - Databases (scores, leads)
- Latencies: hours to milliseconds
  - From events to model
  - From model output to action

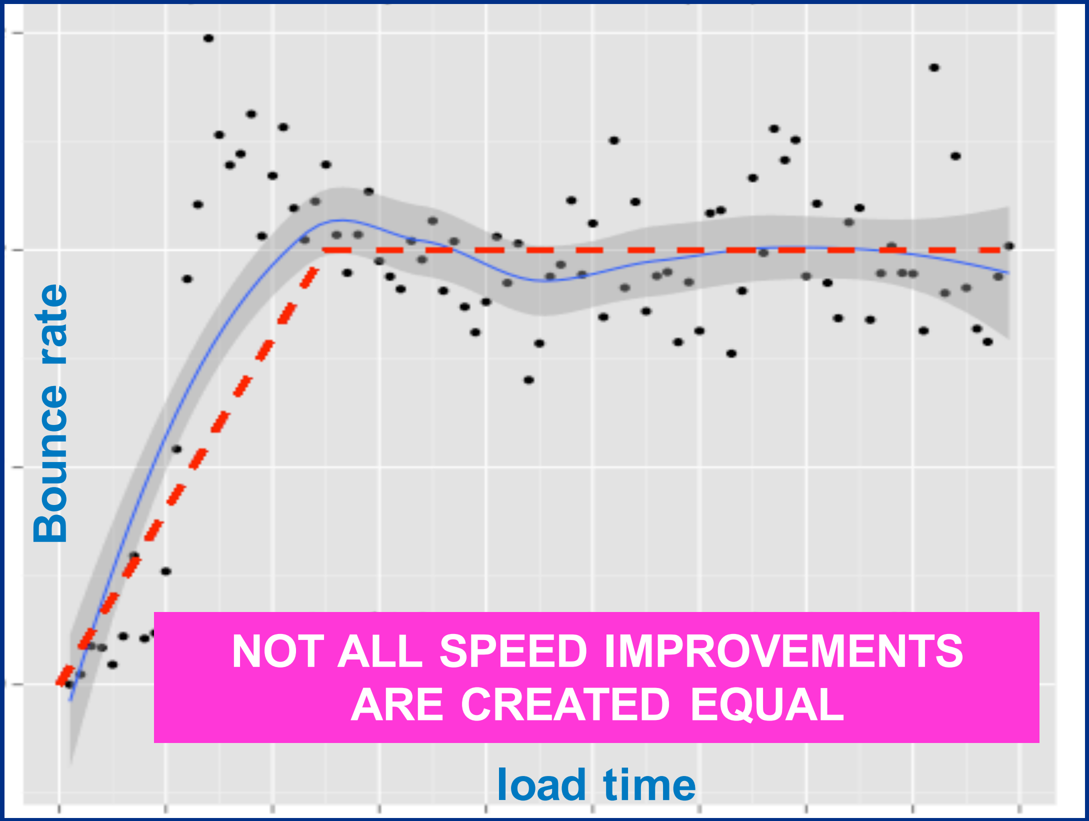
Big Data Visualization

- Multiple aligned plots
- Large number of data points on frame
- Statistical plots
- Network plots
- Model fitting on plot data

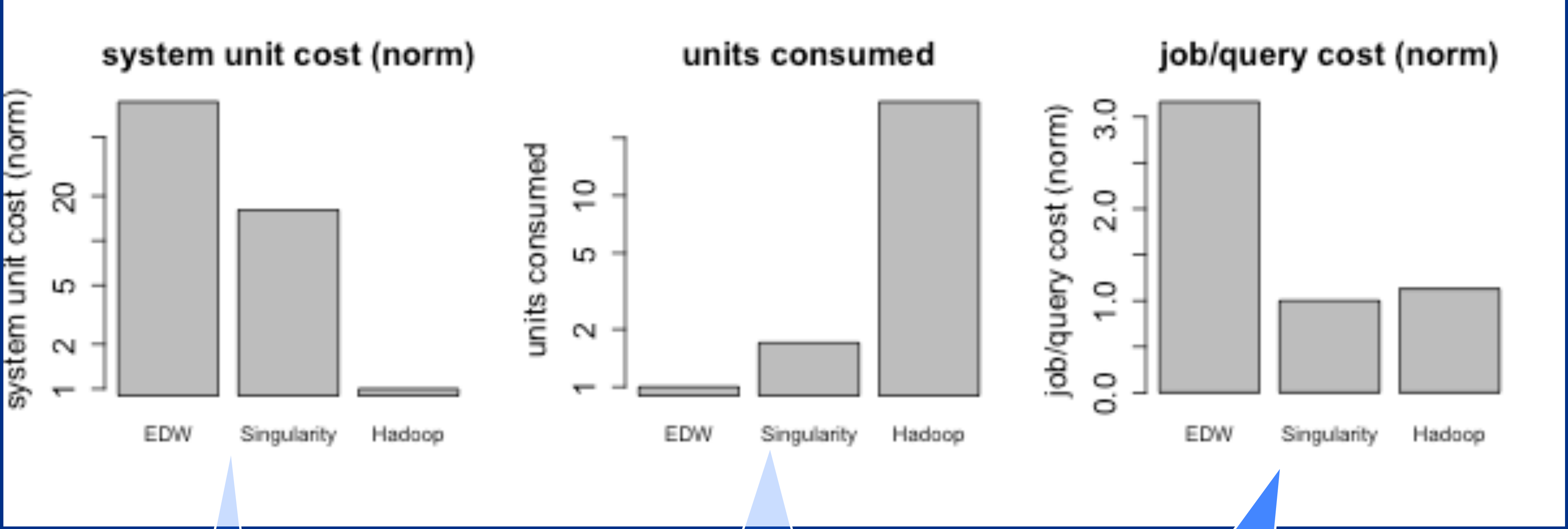
Model Drift

- Detection of concept drift post deployment
  - Change in feature set distribution
  - Change in conditional probability P(y | X)
- Auto correction for concept drift
  - Repeat training on time window of recent data
  - Time window: fixed vs. variable size
- Change detection
  - CUSUM
  - Statistical Process Control

Transaction & Behavioral: Site Speed impact



System Log Data: Query Cost Comparison

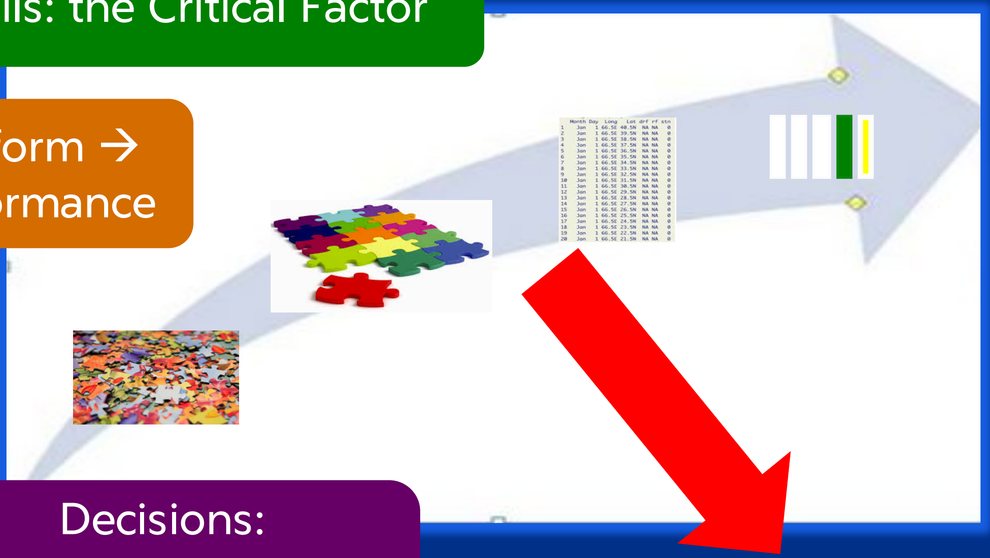


TCO × CONSUMPTION = EXEC COST

Summary: The “Extra” Steps Make The Difference

Details: the Critical Factor

Platform → Performance



Prediction

Decisions: understand the impact

Delayed
Incomplete
Irrelevant

THANK YOU