

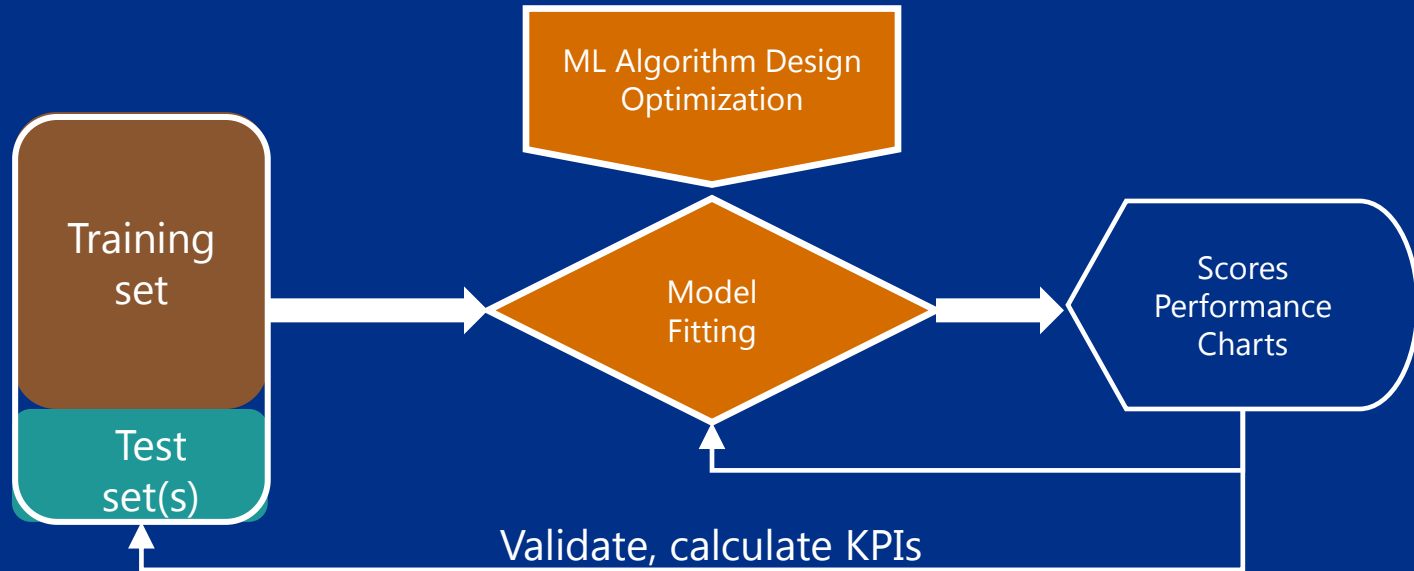
On the Practice of Predictive Modeling with Big Data: The Extra Steps that Make the Difference

*Nachum Shacham
PayPal*

*XLDB
5/19/2015*



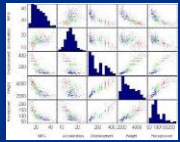
Data Science From 10K ft



Predictive Modeling Meets Big Data: Bridging the Gap

Observational big-data

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



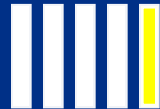
Data Sourcing

Data Sourcing

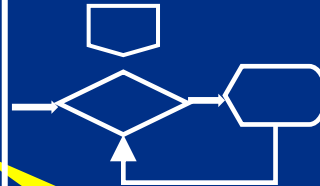
Data Exploration

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

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

Understanding Data Contents & meaning

Distributions and outliers

Null values: codes, quantities

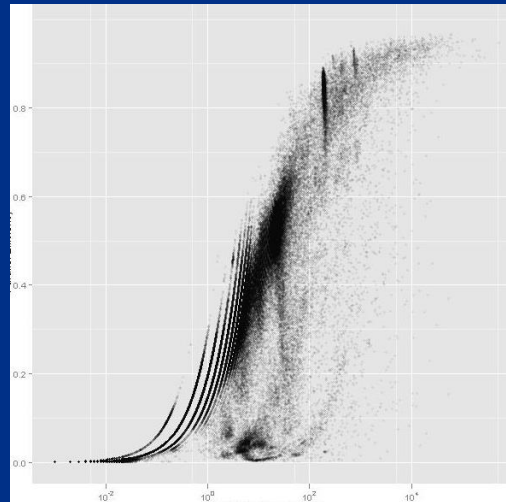
Language

Join keys

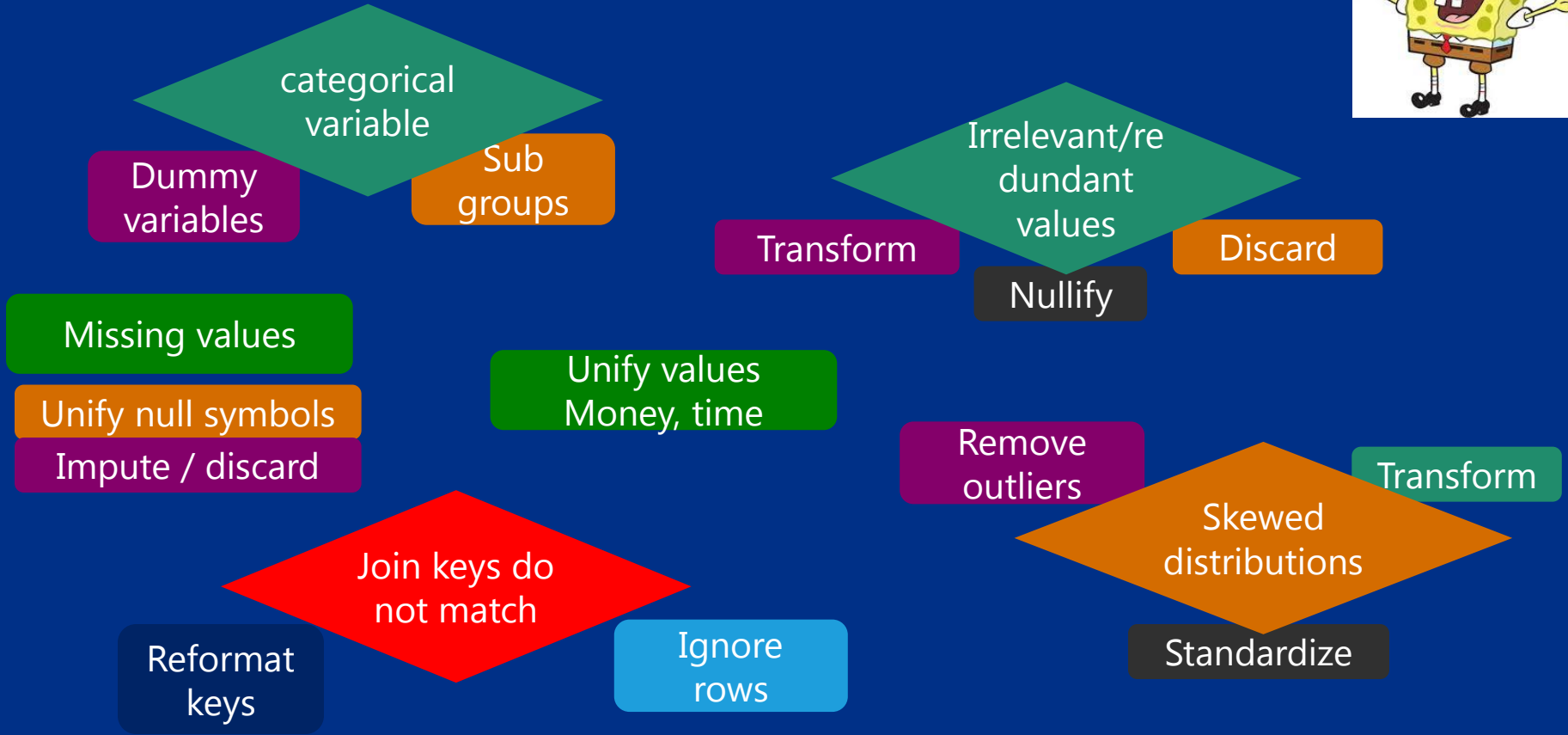
Units

Correlations

Big data visualization



Data Munging: Reshaping the Data → Decisions, Decisions



Feature Engineering

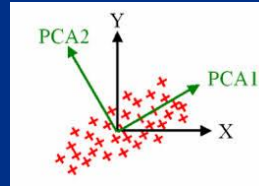
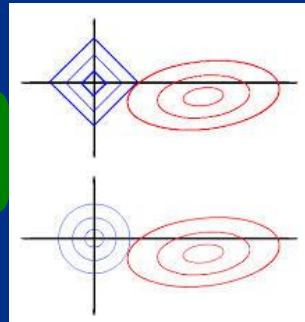
Transform raw data to better represent the problem to the model

Manual Feature Manipulation

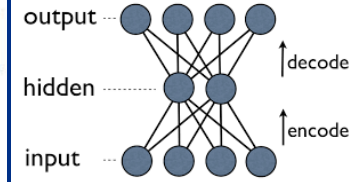
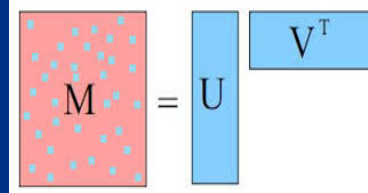
- Dummy variables
- Text integration
- Sub/super sampling (unbalanced sets)

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

Integrated in model fitting



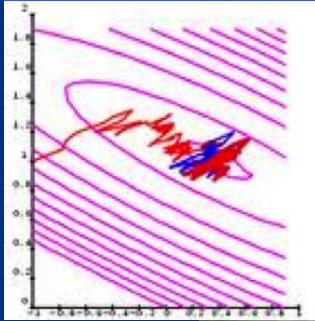
Feature learning algorithms



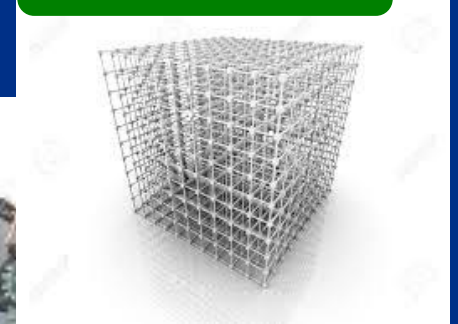
Better features → Simpler models, better results

Model Tuning

Learning Rate



Parameter Grid



Interpretability vs Accuracy



Ensemble: degree, type

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



Big Data Machine Learning in A Cloud

Rich library of ML algorithms

Auto detect feature format

Interoperate with ordinary R: split work for complete munging



Runs through parameter grid

Summary: The "Extra" Steps Make The Difference

Details: the Critical Factor

Platform →
Performance



Month	Day	Category	Unit	Price	Qty	Revenue
Jan	1	Electronics	Smartphones	\$100	100	\$10,000
Jan	2	Electronics	Smartphones	\$100	100	\$10,000
Jan	3	Electronics	Smartphones	\$100	100	\$10,000
Jan	4	Electronics	Smartphones	\$100	100	\$10,000
Jan	5	Electronics	Smartphones	\$100	100	\$10,000
Jan	6	Electronics	Smartphones	\$100	100	\$10,000
Jan	7	Electronics	Smartphones	\$100	100	\$10,000
Jan	8	Electronics	Smartphones	\$100	100	\$10,000
Jan	9	Electronics	Smartphones	\$100	100	\$10,000
Jan	10	Electronics	Smartphones	\$100	100	\$10,000
Jan	11	Electronics	Smartphones	\$100	100	\$10,000
Jan	12	Electronics	Smartphones	\$100	100	\$10,000
Jan	13	Electronics	Smartphones	\$100	100	\$10,000
Jan	14	Electronics	Smartphones	\$100	100	\$10,000
Jan	15	Electronics	Smartphones	\$100	100	\$10,000
Jan	16	Electronics	Smartphones	\$100	100	\$10,000
Jan	17	Electronics	Smartphones	\$100	100	\$10,000
Jan	18	Electronics	Smartphones	\$100	100	\$10,000
Jan	19	Electronics	Smartphones	\$100	100	\$10,000
Jan	20	Electronics	Smartphones	\$100	100	\$10,000
Jan	21	Electronics	Smartphones	\$100	100	\$10,000
Jan	22	Electronics	Smartphones	\$100	100	\$10,000
Jan	23	Electronics	Smartphones	\$100	100	\$10,000
Jan	24	Electronics	Smartphones	\$100	100	\$10,000
Jan	25	Electronics	Smartphones	\$100	100	\$10,000
Jan	26	Electronics	Smartphones	\$100	100	\$10,000
Jan	27	Electronics	Smartphones	\$100	100	\$10,000
Jan	28	Electronics	Smartphones	\$100	100	\$10,000
Jan	29	Electronics	Smartphones	\$100	100	\$10,000
Jan	30	Electronics	Smartphones	\$100	100	\$10,000
Jan	31	Electronics	Smartphones	\$100	100	\$10,000
Feb	1	Electronics	Smartphones	\$100	100	\$10,000
Feb	2	Electronics	Smartphones	\$100	100	\$10,000
Feb	3	Electronics	Smartphones	\$100	100	\$10,000
Feb	4	Electronics	Smartphones	\$100	100	\$10,000
Feb	5	Electronics	Smartphones	\$100	100	\$10,000
Feb	6	Electronics	Smartphones	\$100	100	\$10,000
Feb	7	Electronics	Smartphones	\$100	100	\$10,000
Feb	8	Electronics	Smartphones	\$100	100	\$10,000
Feb	9	Electronics	Smartphones	\$100	100	\$10,000
Feb	10	Electronics	Smartphones	\$100	100	\$10,000
Feb	11	Electronics	Smartphones	\$100	100	\$10,000
Feb	12	Electronics	Smartphones	\$100	100	\$10,000
Feb	13	Electronics	Smartphones	\$100	100	\$10,000
Feb	14	Electronics	Smartphones	\$100	100	\$10,000
Feb	15	Electronics	Smartphones	\$100	100	\$10,000
Feb	16	Electronics	Smartphones	\$100	100	\$10,000
Feb	17	Electronics	Smartphones	\$100	100	\$10,000
Feb	18	Electronics	Smartphones	\$100	100	\$10,000
Feb	19	Electronics	Smartphones	\$100	100	\$10,000
Feb	20	Electronics	Smartphones	\$100	100	\$10,000
Feb	21	Electronics	Smartphones	\$100	100	\$10,000
Feb	22	Electronics	Smartphones	\$100	100	\$10,000
Feb	23	Electronics	Smartphones	\$100	100	\$10,000
Feb	24	Electronics	Smartphones	\$100	100	\$10,000
Feb	25	Electronics	Smartphones	\$100	100	\$10,000
Feb	26	Electronics	Smartphones	\$100	100	\$10,000
Feb	27	Electronics	Smartphones	\$100	100	\$10,000
Feb	28	Electronics	Smartphones	\$100	100	\$10,000
Feb	29	Electronics	Smartphones	\$100	100	\$10,000
Feb	30	Electronics	Smartphones	\$100	100	\$10,000
Feb	31	Electronics	Smartphones	\$100	100	\$10,000



Prediction



Decisions: understand
the impact

Delayed
Incomplete
Irrelevant

THANK YOU