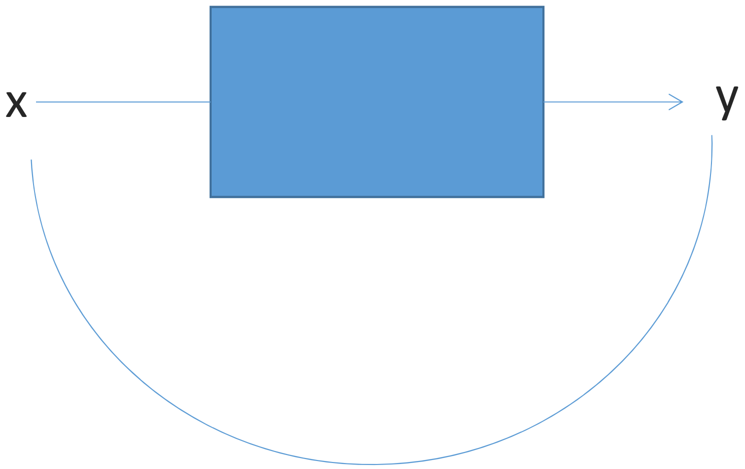
The Difference

	Data Mining/Machine Learning	Statistics (classical)
Goal	Prediction	Infer relationships from a sample to the population
Theme	Precise prediction (Out-of-Sample)	Model interpretation
Model validation	Precise prediction (Out-of-Sample)	Hypothesis test, goodness-of-fit, residuals
Models	Regression trees, random forest, neural networks, support vector machines, ..	linear regression, generalized linear models,, analysis of variance …

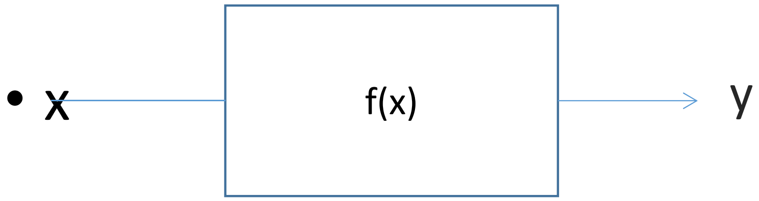
	Data Mining/Machine Learning	Statistics (classical)
Modelling	Flexible – few, if any assumption of underlying distribution	Particular model and error structure
Computational Aspects	Computational efficiency important and considered	Computational efficiency usually not primary focus
Variable/model fitting	Variable selection and overfitting can be problematic	Variable and model selection issues must be considered

In short …



x

y



- x

f(x)

y

Examples (Caffo, Leek, Peng 2016)

Machine learning

- build an automated movie recommender system
- success - anything that produces reliable recommendations

Statistical analysis

- build a parsimonious and interpretable model to better understand why people choose the movies that they do
- success - anything true learned about movie choices

Prediction versus Explanation

Many applications of classical statistics have focused on explaining how or possibly why certain predictors are related to outcomes we care about

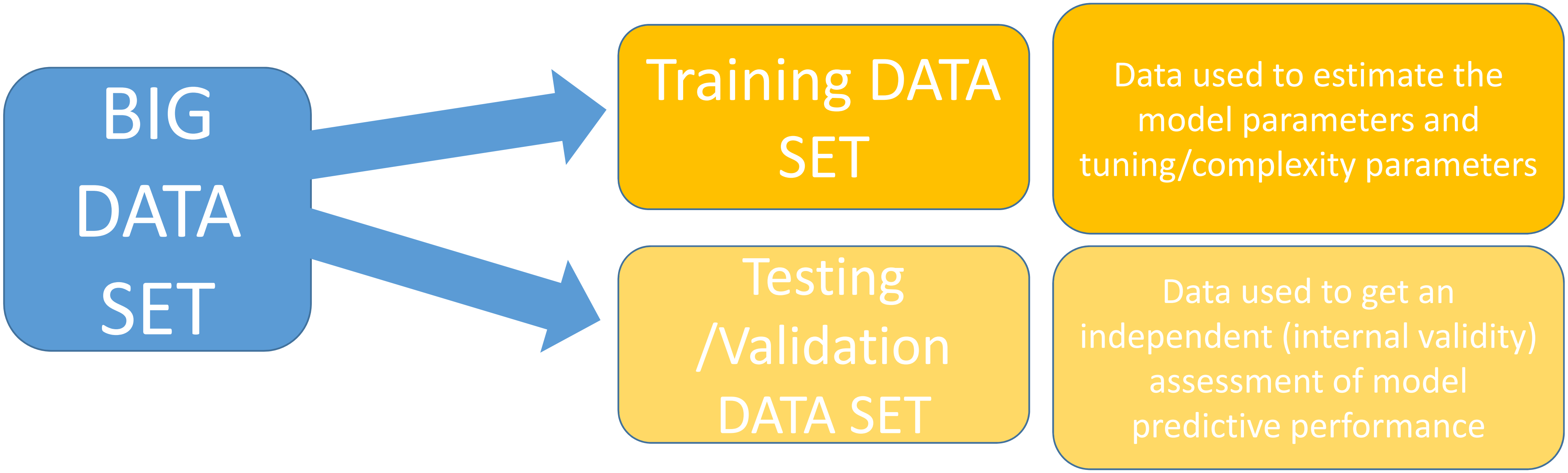
- Descriptive statistics quantify the degree of association or relationship.
- Inferential statistics provide a way to confirm that the relationship goes beyond random error.
- We evaluate models based on their *explanatory power* (R^2).

Machine learning algorithms can also describe relationships between predictors and outcomes

- But many times, this is not the end goal.
- End goal is to develop well tuned predictive “models” that can be applied to new data.
- Here models are evaluated based on their *predictive power.*

An excellent overview of the model building process - Shmueli (2010) http://bit.ly/1F27hGT

Model Evaluation Strategy: Split Sample



BIG DATA SET

Training DATA SET

Data used to estimate the model parameters and tuning/complexity parameters

Testing /Validation DATA SET

Data used to get an independent (internal validity) assessment of model predictive performance

Machine Learning Overview 2 types of learning

Supervised Learning –

- *Dataset has clear set of Predictors (Xs) and Outcomes (Y)*
- *Goal is to understand the relationship between predictors and outcomes*
- *End goal is either <u>Inference</u> or <u>Prediction</u>*

Regression

Classification

Unsupervised Learning –

- *Dataset consists of Predictors (Xs) only*
  - Think of Sampling Frame Data – all auxiliary variables – no survey outcome variables (i.e. Ys)
- *We seek to understand the relationships between the variables/observations in the dataset*

Clustering

Dimensionality Reduction