

Biostat 202: Opportunities and Challenges of “Big Data”: Machine Learning 1: Introduction & Regression

Aaron Wolfe Scheffler

Division of Biostatistics

Department of Epidemiology and Biostatistics

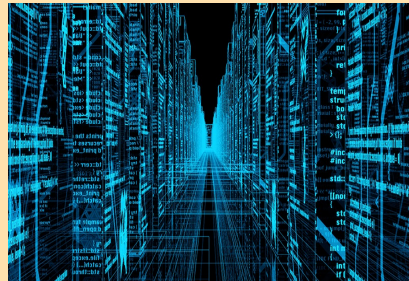
- **Goal:** Use machine learning algorithms for the purpose of *prediction* or *identifying patterns* in data
- **Case study:** dataset from a heart disease study, want to predict systolic blood pressure

1. Data

2. Train/fit model

3. Output

4. Interpret



Algorithm(s)

- Prediction
- Dimension Reduction
- Clustering

= INFORMATION?



Pop quiz: Do machine learning/data mining techniques generate information?

Overview of machine learning (ML) lectures

- ML Lecture 1: Introduction to ML and supervised learning [regression]
 - Machine learning/data mining “definition”
 - Structured vs unstructured data
 - Supervised vs unsupervised machine learning
 - Linear regression
 - Validation
- ML Lecture 2: Supervised learning [classification]
 - Introduction to classification/evaluating classifiers
 - Validation continued
 - Logistic regression
- ML Lecture 3: Supervised learning, continued [classification]
 - Classification, continued: K nearest neighbors, decision trees, support vector machines, neural networks
- ML Lecture 4: Misc supervised learning and ensemble learning
 - Cross validation
 - Feature importance
 - Combining models: boosting, bagging, deep learning (ensemble learning)
- ML Lecture 4: Unsupervised learning [clustering/dimension reduction]
 - Clustering
 - Data reduction and PCA
- Orange throughout

Overview of this lecture

- Machine learning/data mining “definition”
- Structured vs unstructured data
- Supervised vs unsupervised machine learning
- Linear regression
- Validation (central to evaluating predictions)



Pop quiz: name that data!

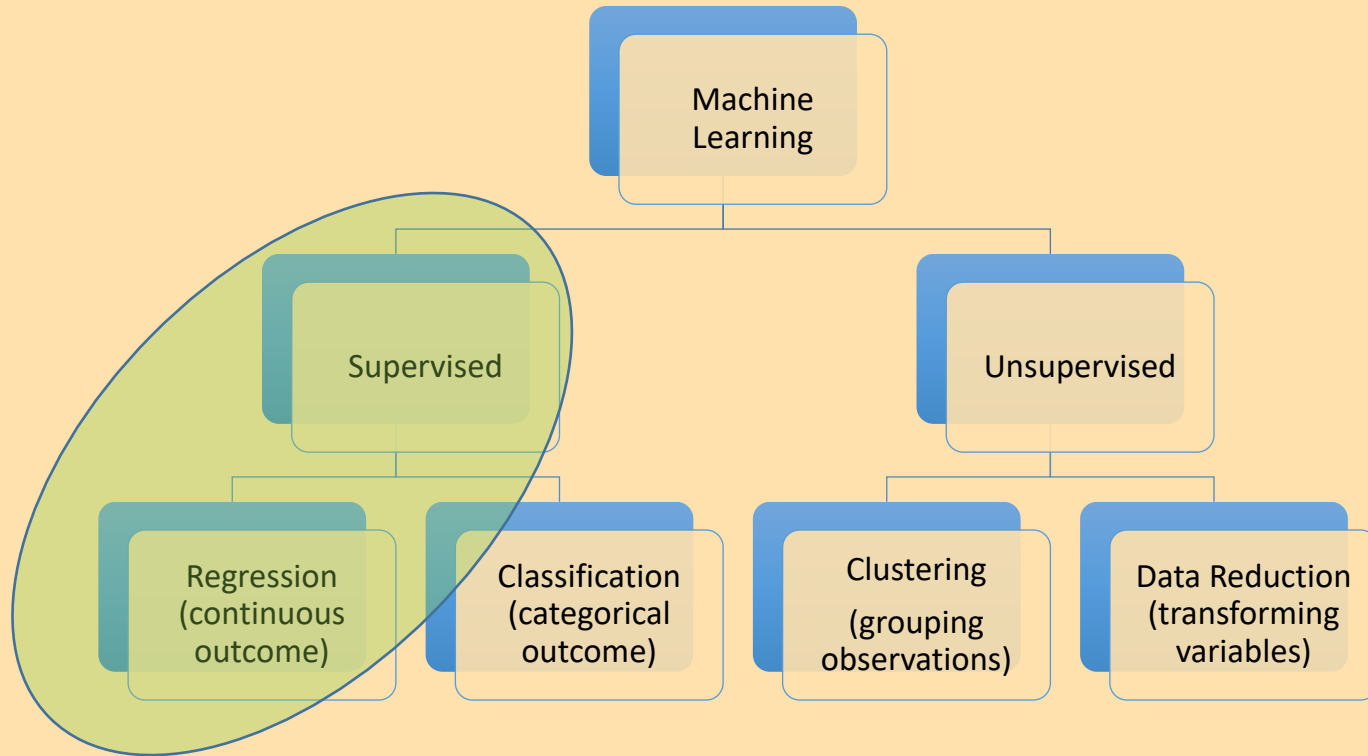
Structured or unstructured?

- Blood glucose levels
- MRI/X-ray image
- Census data
- Physician notes
- Electronic medical records

Pop quiz: name that data!

- Blood glucose levels (**structured**)
- MRI/X-ray image (**unstructured**)
- Census data (**structured**)
- Physician notes (**unstructured**)
- Electronic medical records (**both!**)

Supervised vs unsupervised



Fitting linear regression

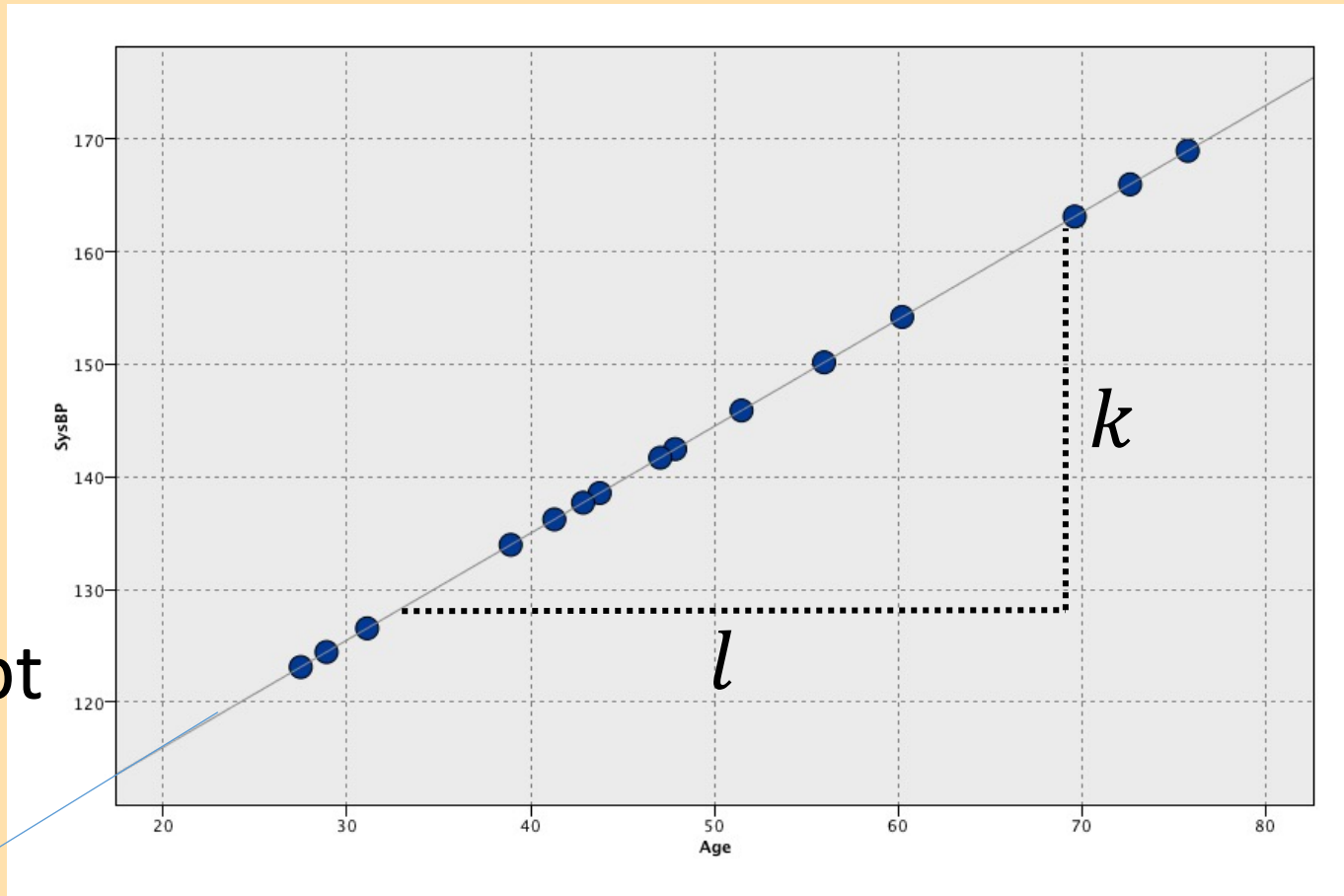
y ↑

SysBP

intercept

c

Age = 0



slope

$$a = \frac{k}{l}$$

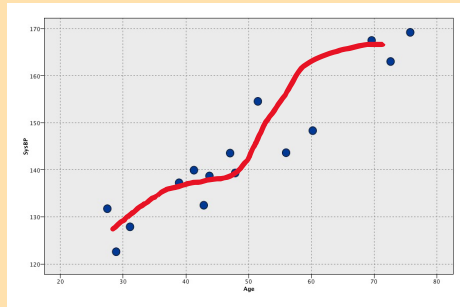
Age

x

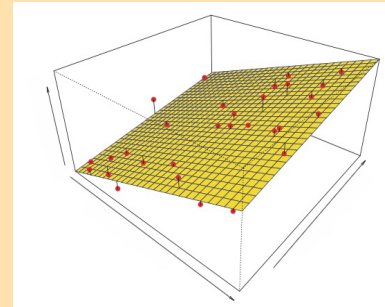
Model selection

- Want to find “true” predictive patterns in the data

Nonlinear trends

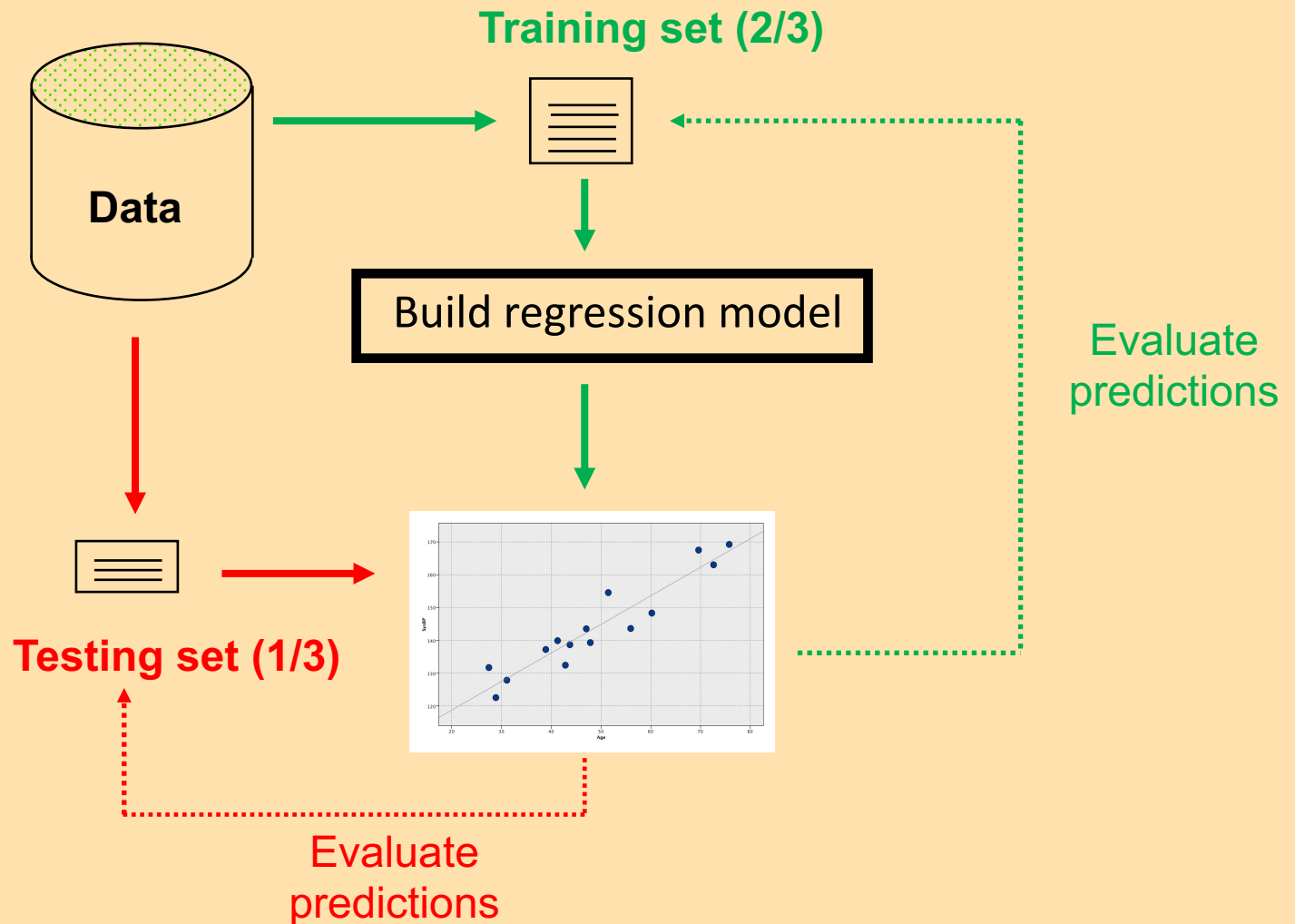


Many predictors

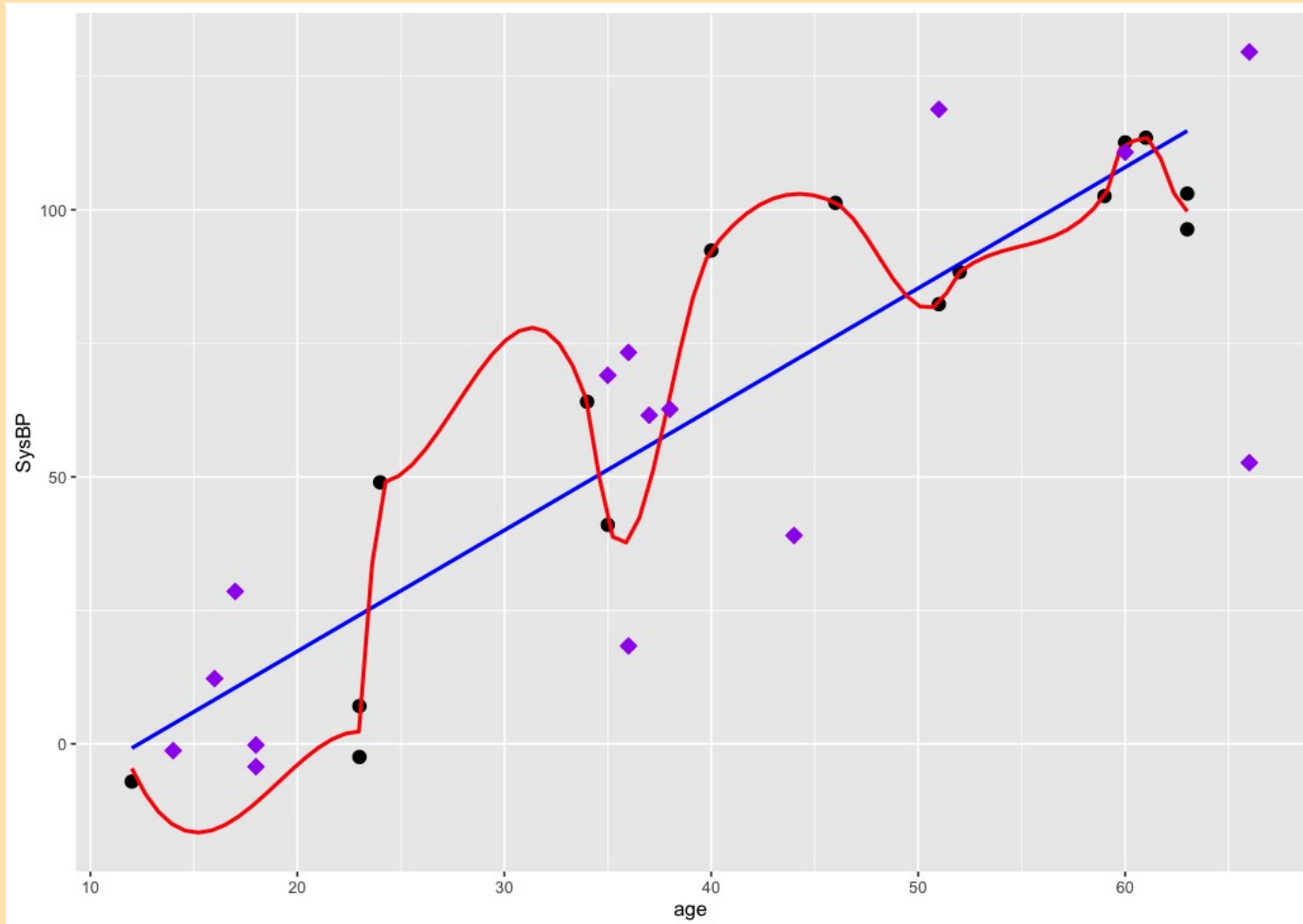


- Want to avoid overfitting spurious detail – do not confuse noise for signal
- Solution – **independent validation** – reserve some data that has not been used in model fitting for the purpose of model evaluation and comparison. Choose the model that performs the best on the testing data.

Training/Test

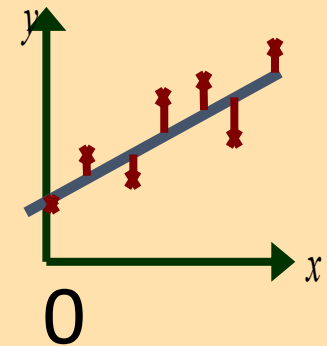
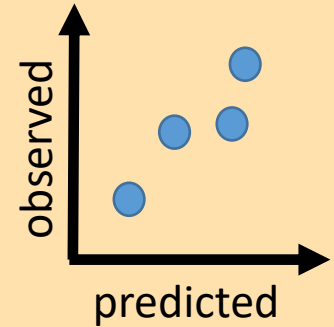


Simulation of overfitting



Evaluation metrics for continuous predictions

- Linear correlation² (r^2) – square of correlation coefficient between the observed target values and predicted target values
- Mean absolute error (MAE) -- the average absolute deviation between observed target values and predicted target values



Case study – Heart data

- Data:
 - $n = 303$
 - 14 attributes
- **Target:**
 - **SysBp: systolic blood pressure**
- Predictors:
 - Age: age (years)
 - Sex: gender
 - ChestPain: chest pain type
 - Chol: cholesterol
 - Fbs: fasting blood sugar >120 mg/dL
 - restecg: resting electrocardiographic results
 - MaxHR: maximum heart rate achieved
 - ExAng: exercise induced angina
 - Oldpeak: ST depression induced by exercise
 - Slope: slope of peak exercise ST segment
 - Ca: # of major vessels colored by flouroscopy
 - Thal: Thallium stress test
 - AHD: diagnosis of heart disease (1 = yes, 0 =no)

Goal: Predict systolic blood pressure using multiple linear regression.

Review of this lecture

- Machine learning/data mining “definition”
- Structured vs unstructured data
- Supervised vs unsupervised machine learning
- Regression (supervised)
- Validation (central to evaluating predictions)

Thursday's Lecture

- Introduction to classification
- Evaluation metrics for binary/flag predictions
- Validation and testing, continued
- Logistic regression