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?

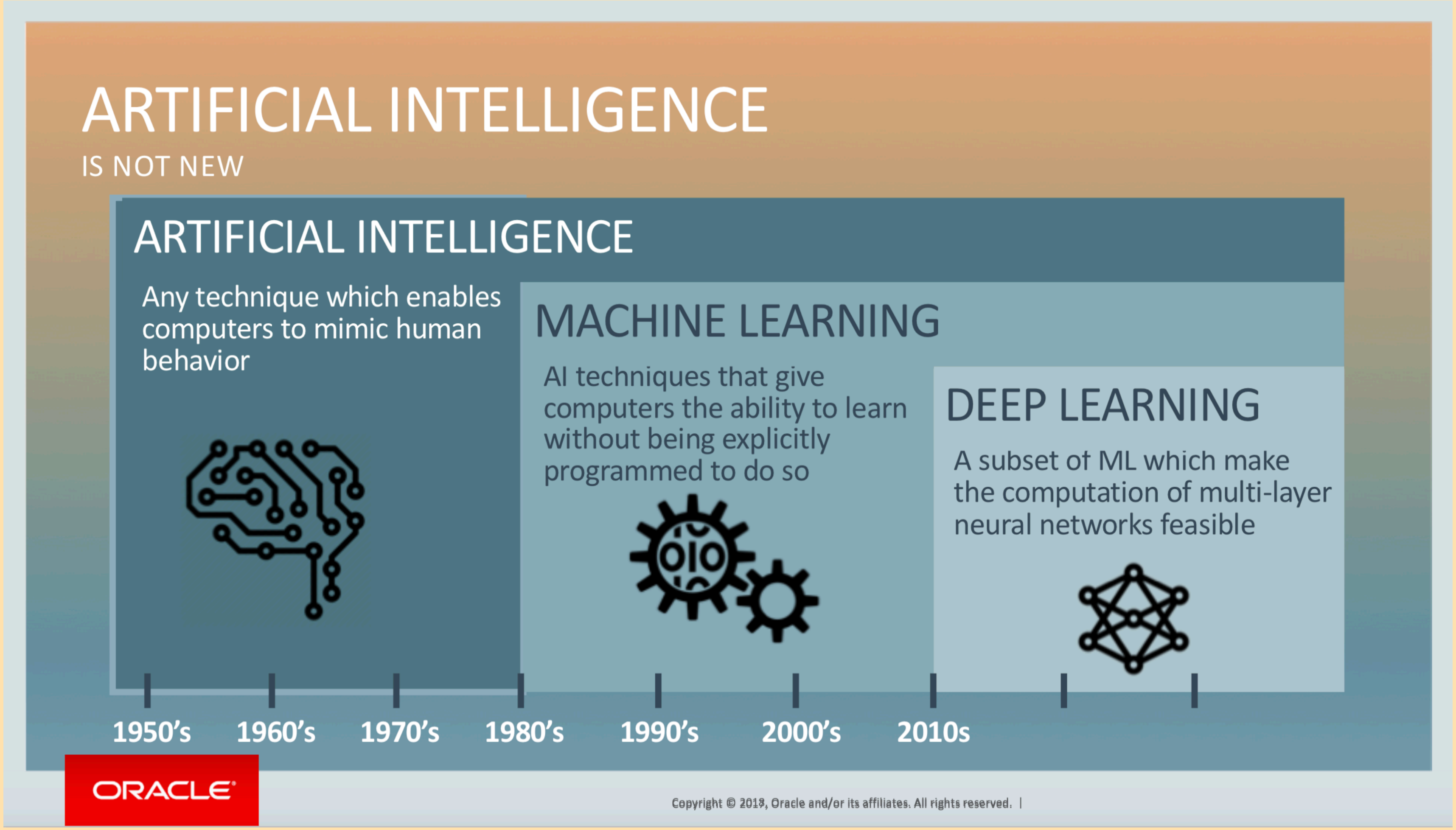
Overview of machine learning (ML) lectures

- 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
  - Cross validation
  - Combining models: boosting, bagging, deep 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)

A historical view



https://tinyurl.com/y37tgvt6

Machine learning definitions

- *Artificial intelligence (AI)* is any technique which enables computers to mimic human behavior.
- *Machine learning* is a subfield of artificial intelligence centering on the application of algorithms and statistical models. In 1959, Arthur Samuel defined machine learning as a “Field of study that gives computers the ability to learn without being explicitly programmed”.
  - *Data Mining* is the process of identifying patterns and associations in large data sets.
- *Deep learning* is a subfield of machine learning that utilizes artificial neural networks. This technology underlies many recent successful applications of machine learning.

Local headlines

Research · February 24, 2016

Using Big Data to Chart Cancer's Hidden Genetic Weaknesses

By Nicholas Weiler

Nevan Krogan, PhD, thinks genomics has brought us closer to a revolution in cancer treatment than most geneticists even realize.

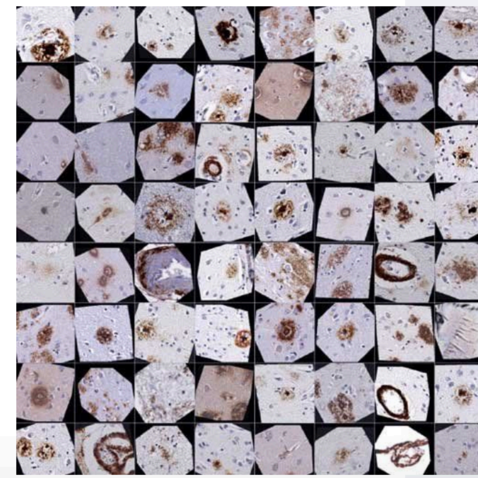
"There's been a tsunami of genetic data about different cancers," says Krogan, a UC San Francisco professor of Cellular and Molecular Pharmacology and member of the UCSF Helen Diller Family Comprehensive Cancer Center. But despite this rising tide of knowledge, Krogan says, breakthroughs

Artificial Intelligence Could Vastly Scale Up Alzheimer's Research

Machine Learning Tool Automates Pathologists' Work to Identify Disease Markers

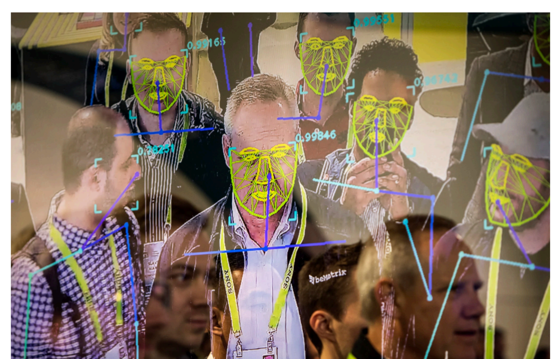
By Nicholas Weiler and Dorsey Griffith

Researchers at UC Davis and UC San Francisco have found a way to teach a computer to precisely detect one of the hallmarks of Alzheimer's disease in human brain tissue, delivering a proof of concept for a machine-learning approach capable of automating a key component of Alzheimer's research.



Amyloid plaques are clumps of protein fragments in the brains of people with Alzheimer's disease that destroy nerve cell connections. Much like the way Facebook recognizes faces based on captured images, the machine learning tool developed by a team of University of California scientists can "see" if a sample of brain tissue has one type of amyloid plaque or another — and do it very quickly.

San Francisco Bans Facial Recognition Technology



Machine learning and statistics

- The purpose of machine learning is to form predictions or detect patterns in data
- Statistics is an older field which overlaps with machine learning but focuses more on describing associations and quantifying uncertainty
- Many techniques (such as linear regression) are utilized in both fields while others (neural networks) tend to be considered strictly machine learning

Analysis pipeline

1. Data

2. Train/fit model

Algorithm

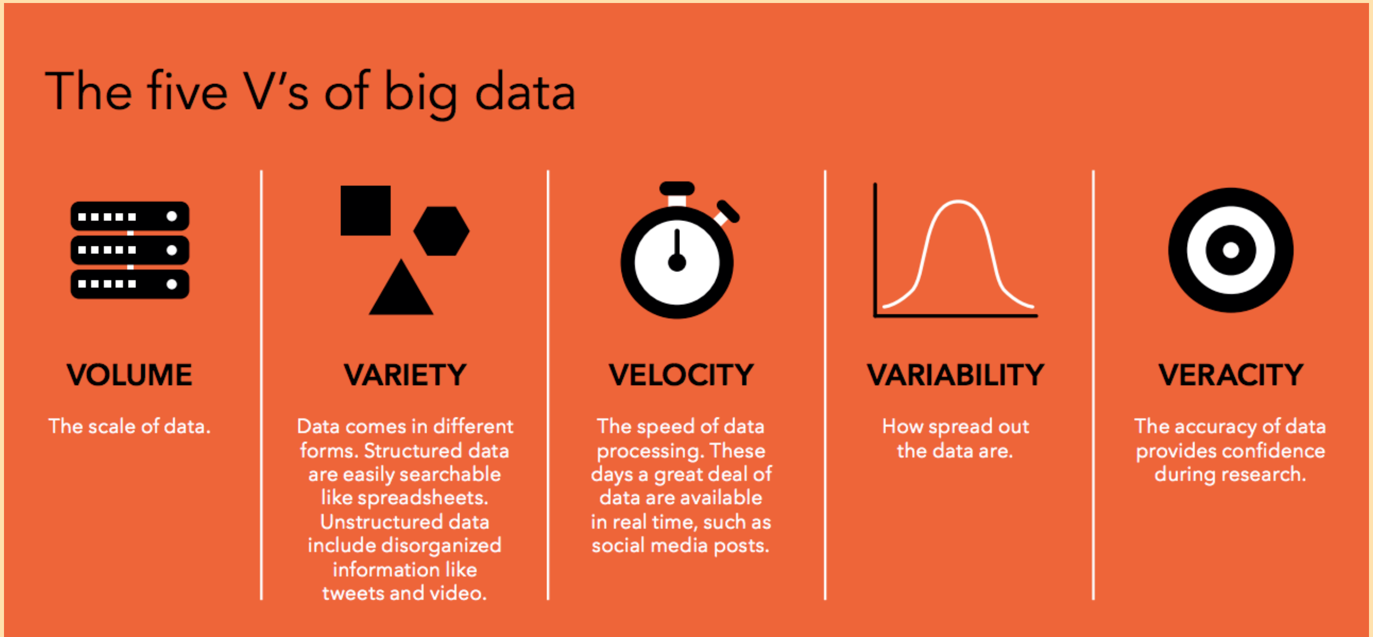
3. Output

- Prediction
- Dimension Reduction
- Clustering

4. Interpret

= INFORMATION?

“Big Data”



https://hac.nursing.umich.edu/2018/02/26/big-data-finding-its-mark/

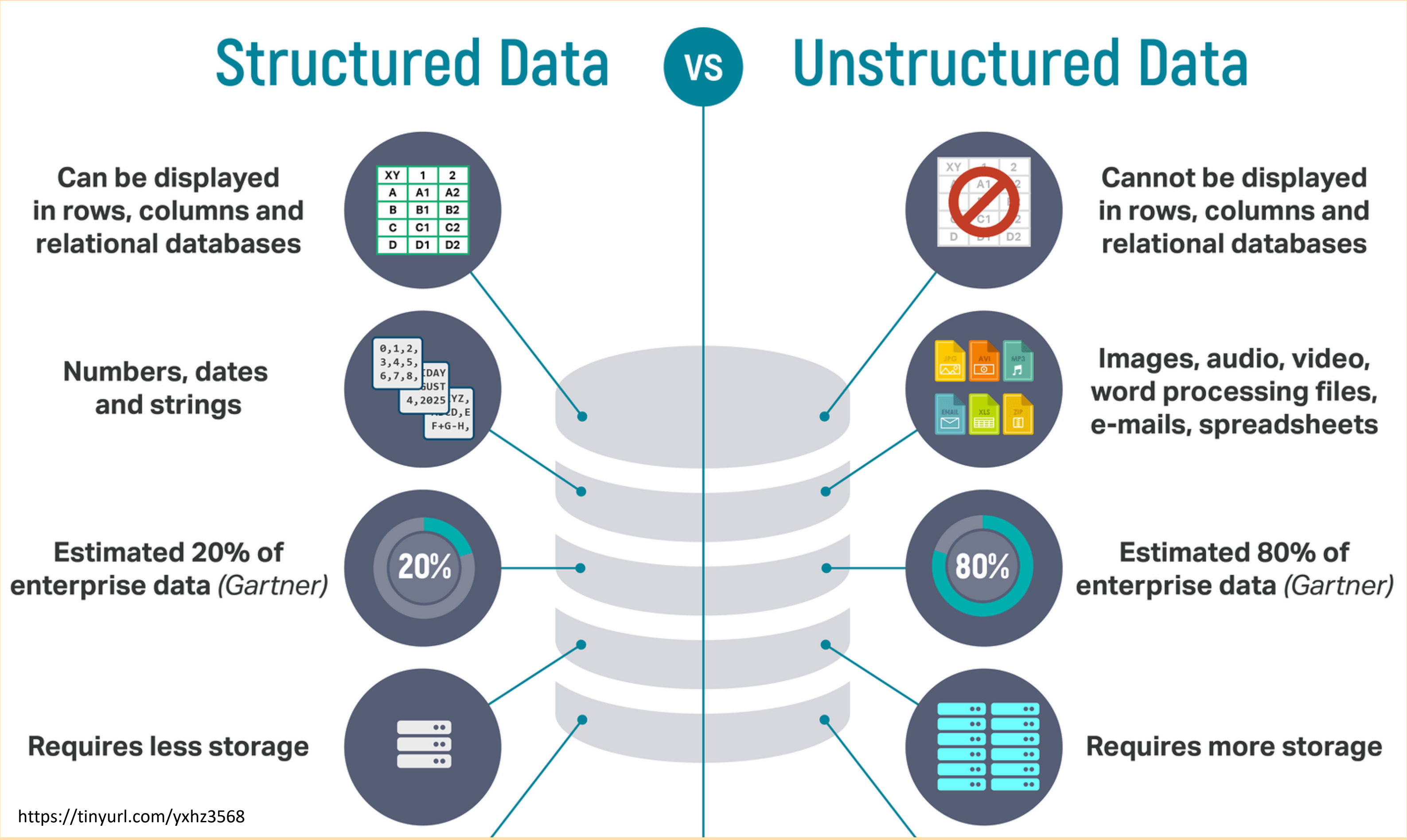
Big Data

All Data

Big Data – large N and large P

- Big data can be thought of in two ways
- Large N = large number of subjects – implies more information for fitting complex models or ones with many predictors
  - E.g. Assignment 2 - NEISS Injury database (360K records)
- Large P = large number of predictors, or complex relationships between predictor and outcome
  - E.g. Project 2 Cancer survival (243 records, 21K features)

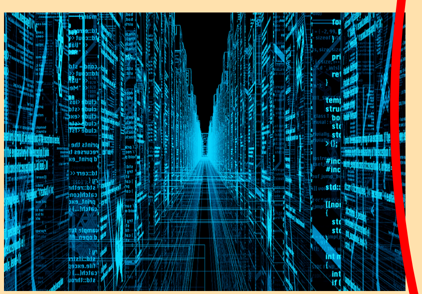
Structured vs unstructured data



Our focus

Analysis pipeline

1. Data



2. Train/fit model

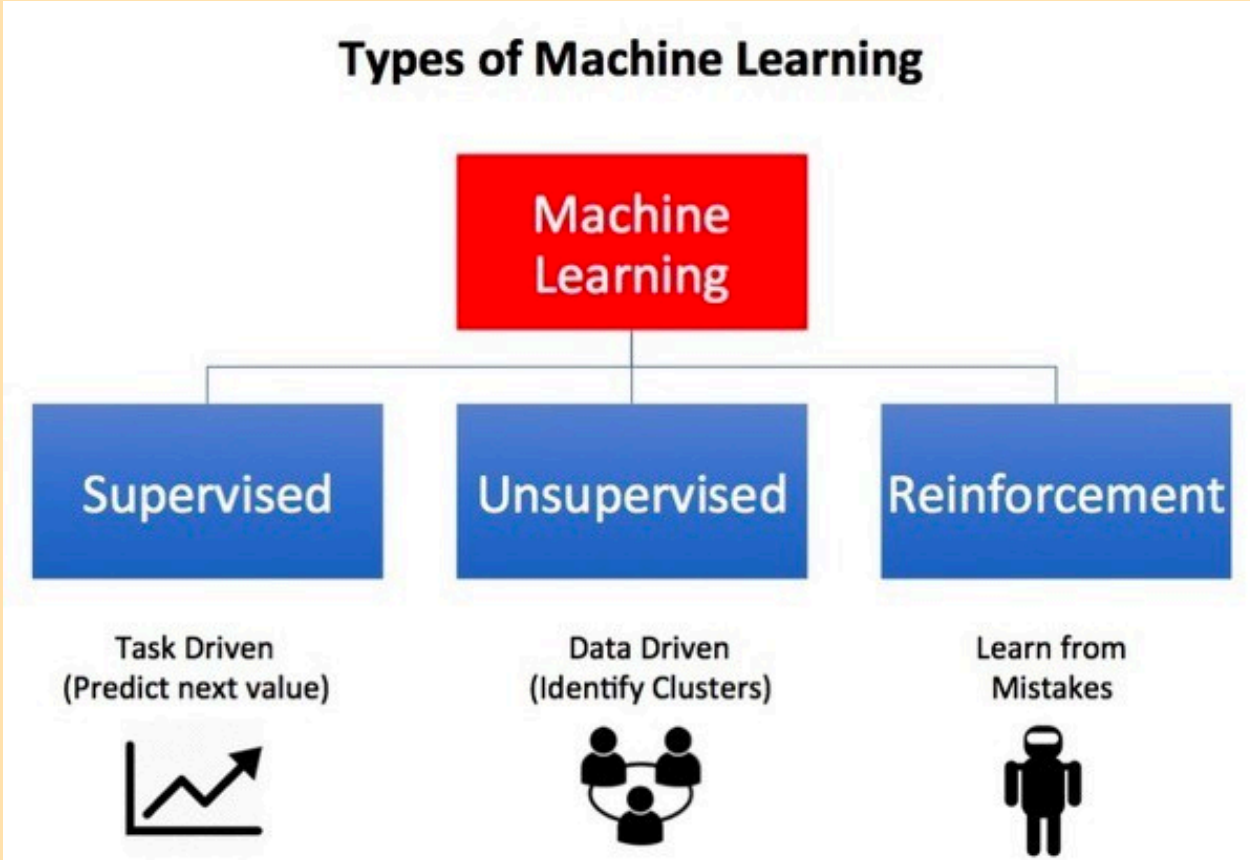
Algorithm(s)

3. Output

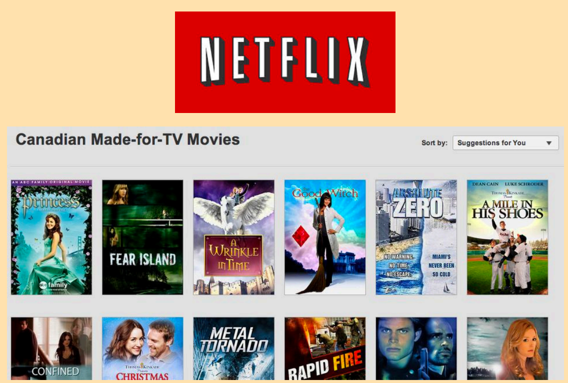
- Prediction
- Dimension Reduction
- Clustering

4. Interpret

= INFORMATION?


Types of Machine Learning
Machine Learning
Supervised
Unsupervised
Reinforcement
Task Driven
(Predict next value)
Data Driven
(Identify Clusters)
Learn from
Mistakes


BRCA
BREAST
CANCER


NETFLIX
Canadian Made-for-TV Movies
Sort by:
Suggestions for You
FEAR ISLAND
A MILE IN HIS SHOES
METAL TORNADO
RAPID FIRE
CONFINED

Supervised vs unsupervised

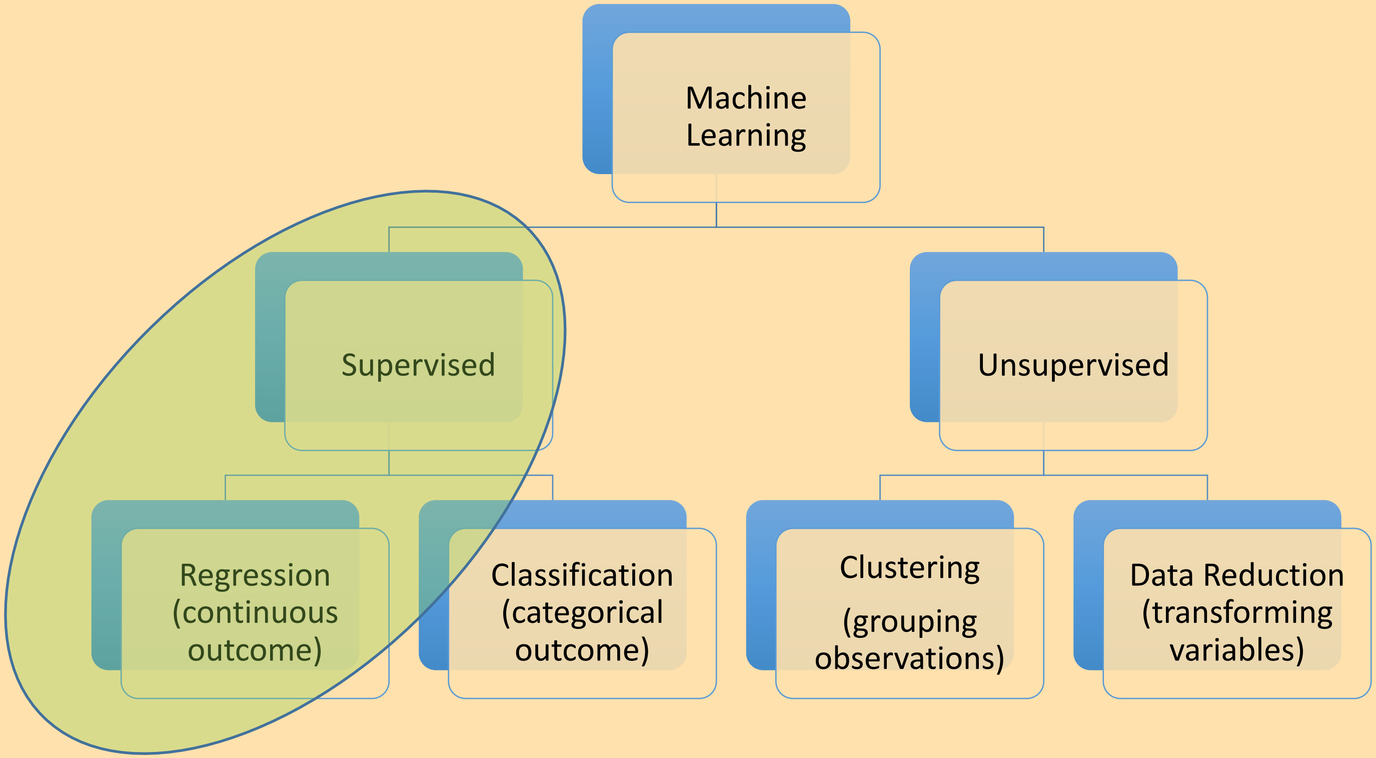
- **Supervised Learning**
- Have data where you know outcome in order to *train* the model
- Train the model to predict future outcomes from sets of predictors

Prediction: Regression (continuous outcomes) and classification (categorical outcomes)

- **Unsupervised Learning**
- No specified outcomes
- Cluster data points into groups with "similar" properties (clustering)
- Or reduce data to fewer variables in some way that still captures important structure of the data

Clustering and data reduction

Supervised vs unsupervised

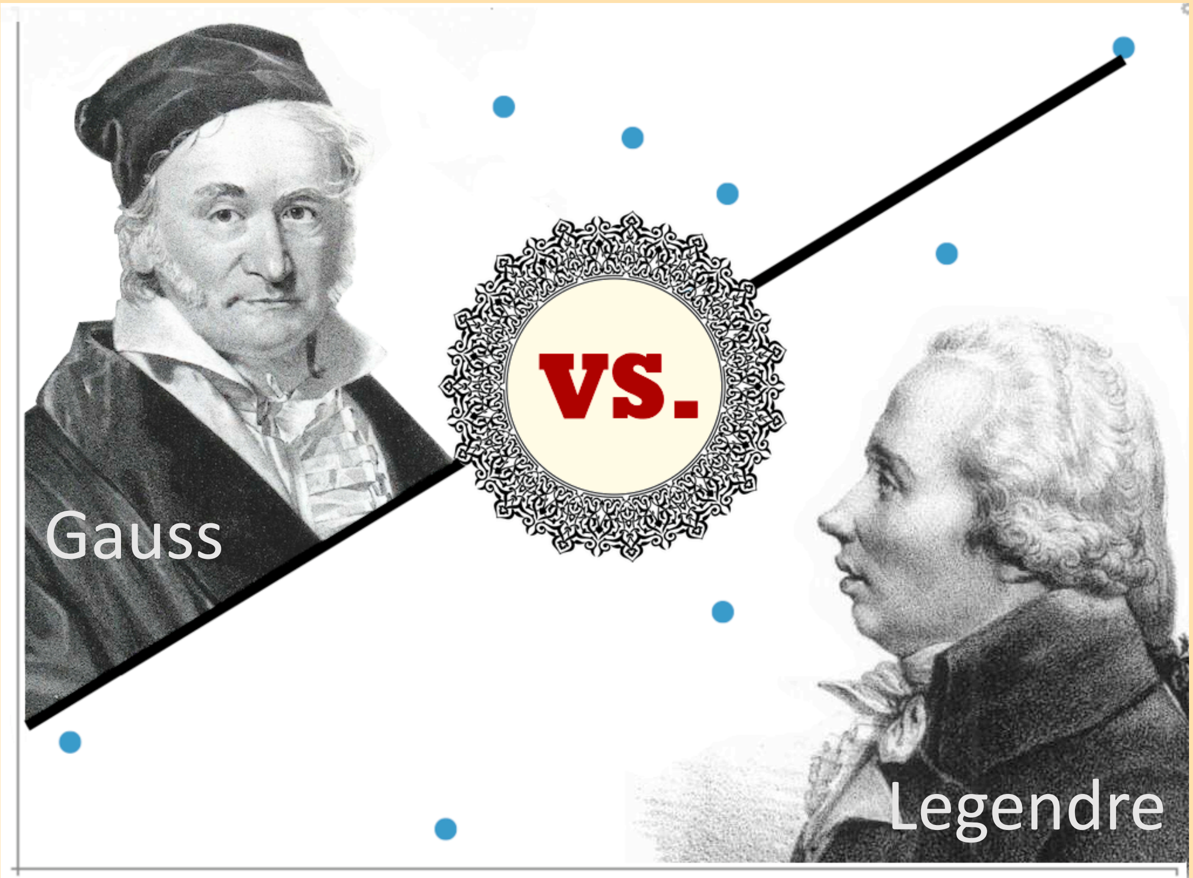


Prediction problems

- Predict the CD4 count in a patient 6 months from now based on current CD4, viral load, demographic and lifestyle variables
- Predict whether a patient will develop breast cancer based on genetic, demographic, and lifestyle variables
- Predict whether or not, and what type of dementia a patient has based on MRI of the brain

Linear regression

- Start simple with one feature/predictor
- We will move onto multiple predictors later…



Linear regression

Consider a target variable y as linearly related to an feature/predictor x, then:

$$y = ax + c$$

Question: Is Age predictive of SysBP?

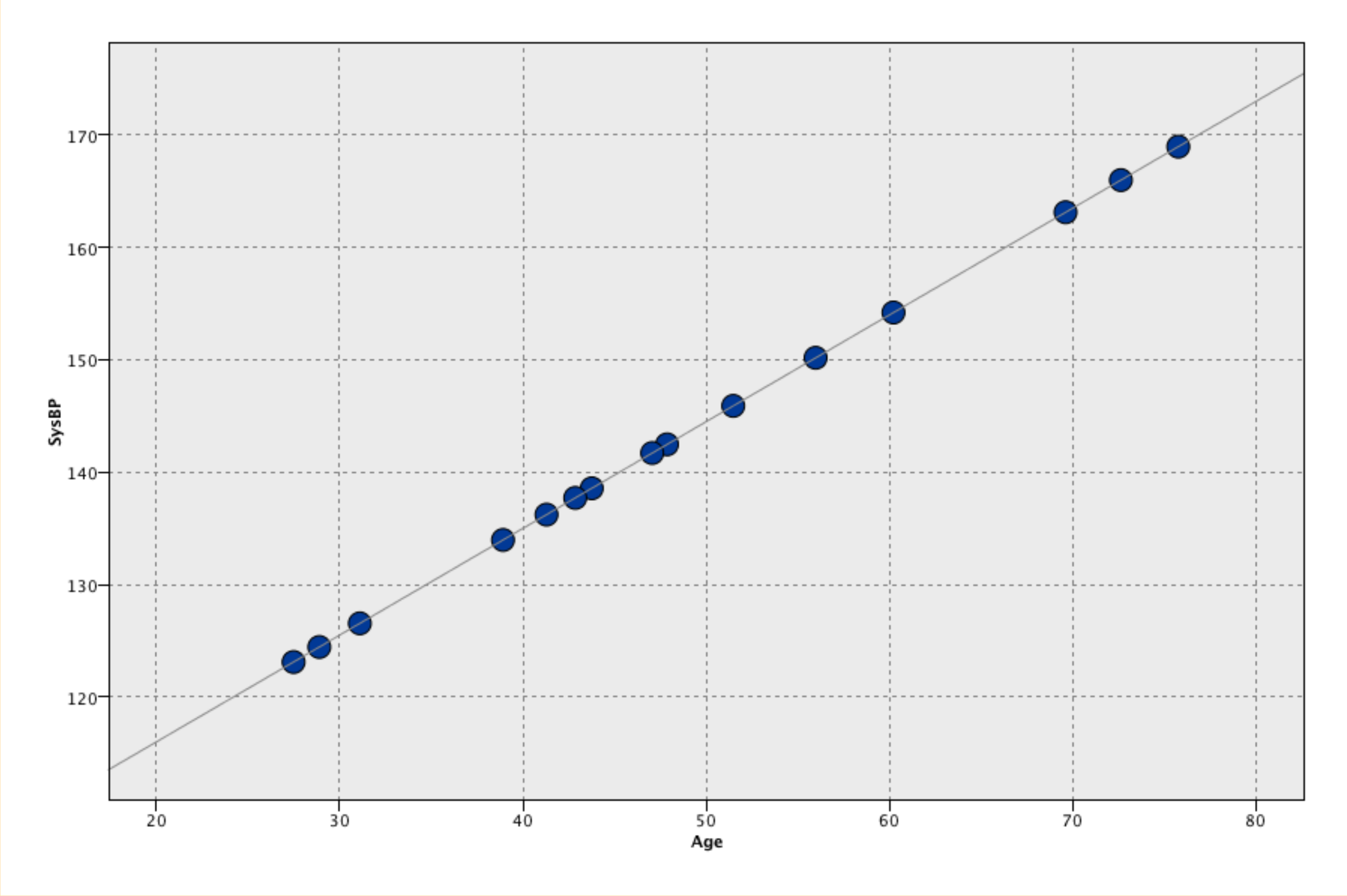
y = Systolic BP (SysBP)	[target]
x = Age	[feature]
a = change SysBP per year in Age	[parameter]
c = intercept (SysBP at age 0)	[parameter]

$$SysBP = a * Age + c$$

Fitting linear regression

y

SysBP



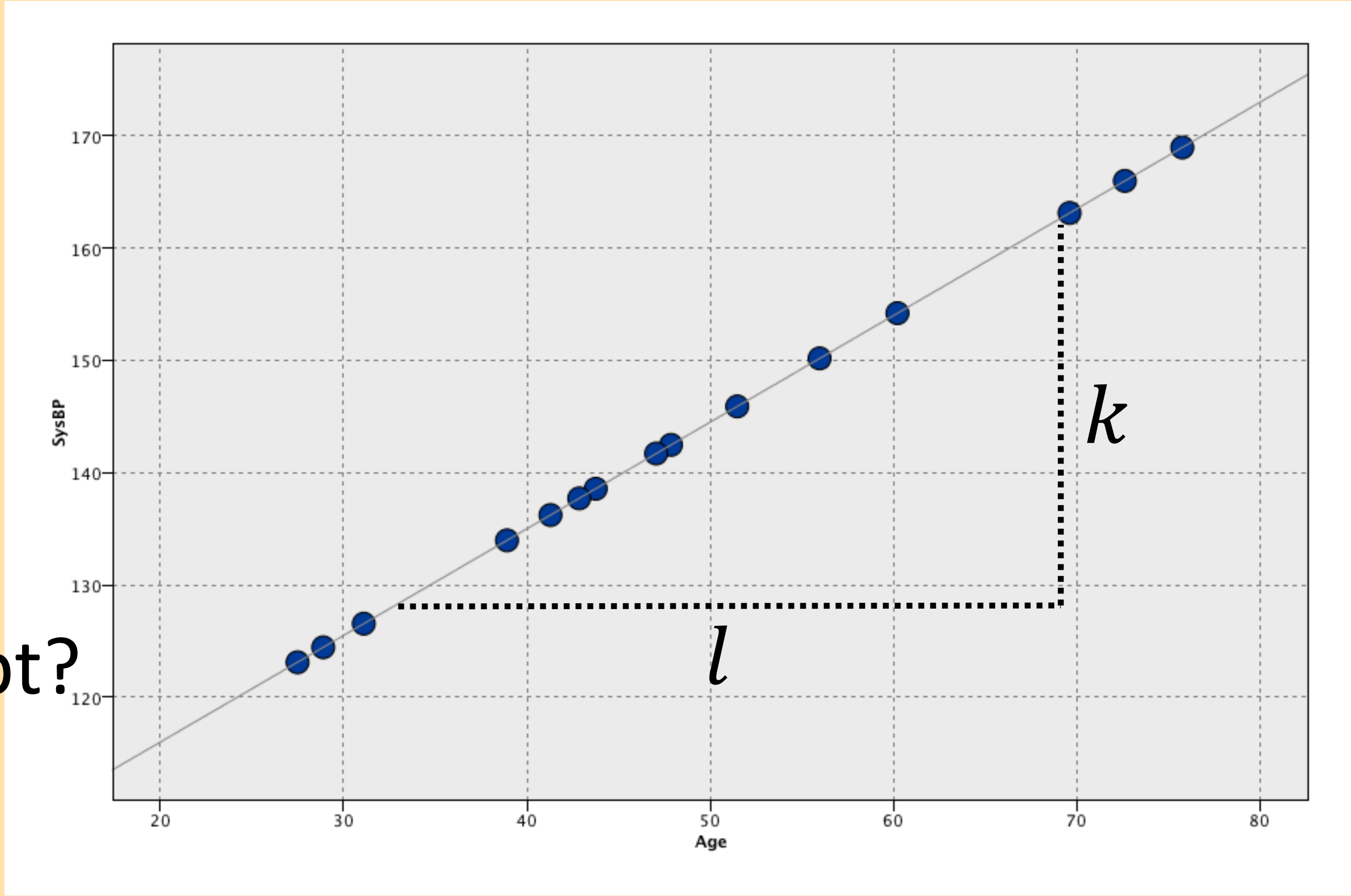
Age

x

Fitting linear regression

y

SysBP



Intercept?

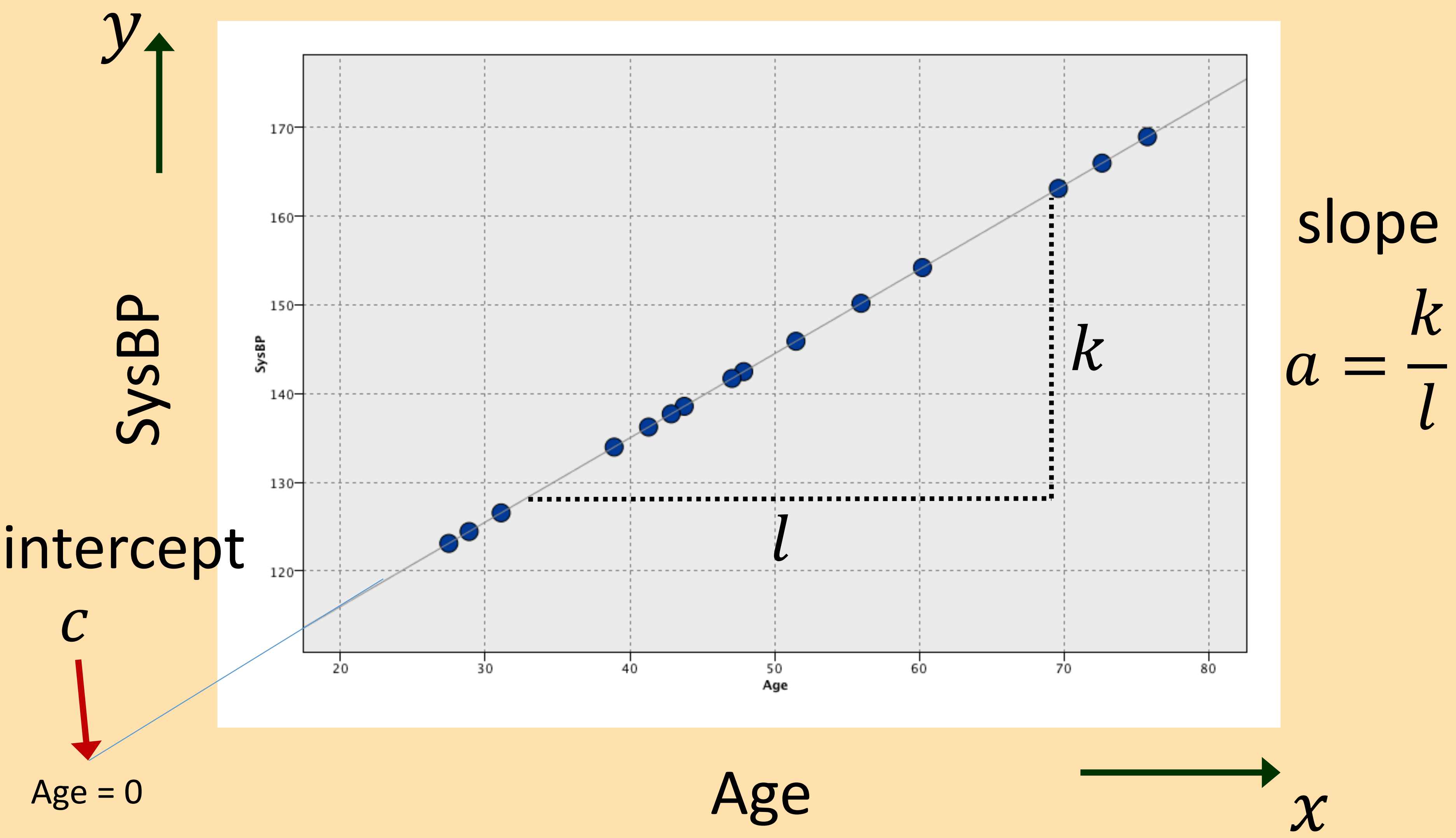
slope

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

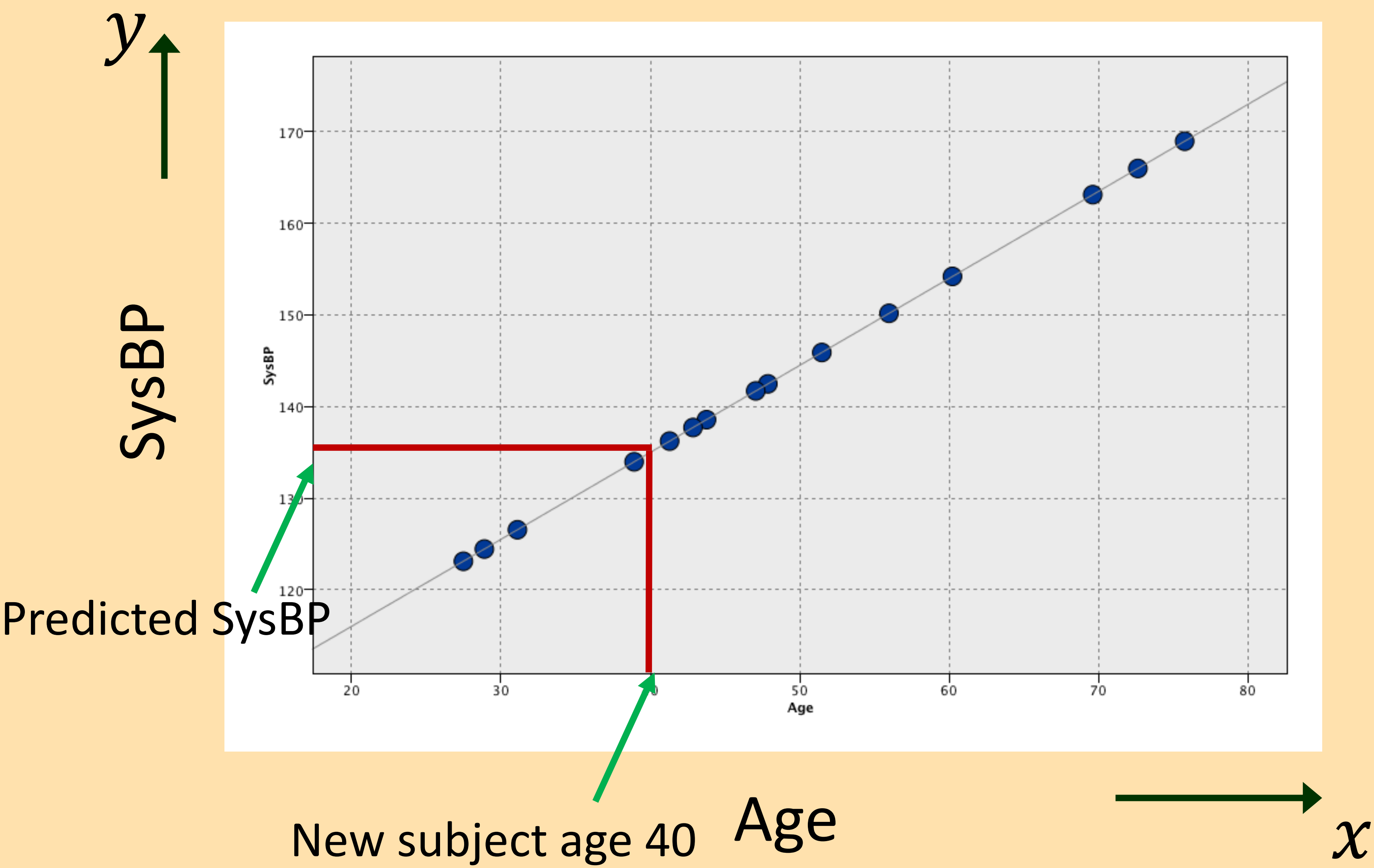
Age

x

Fitting linear regression



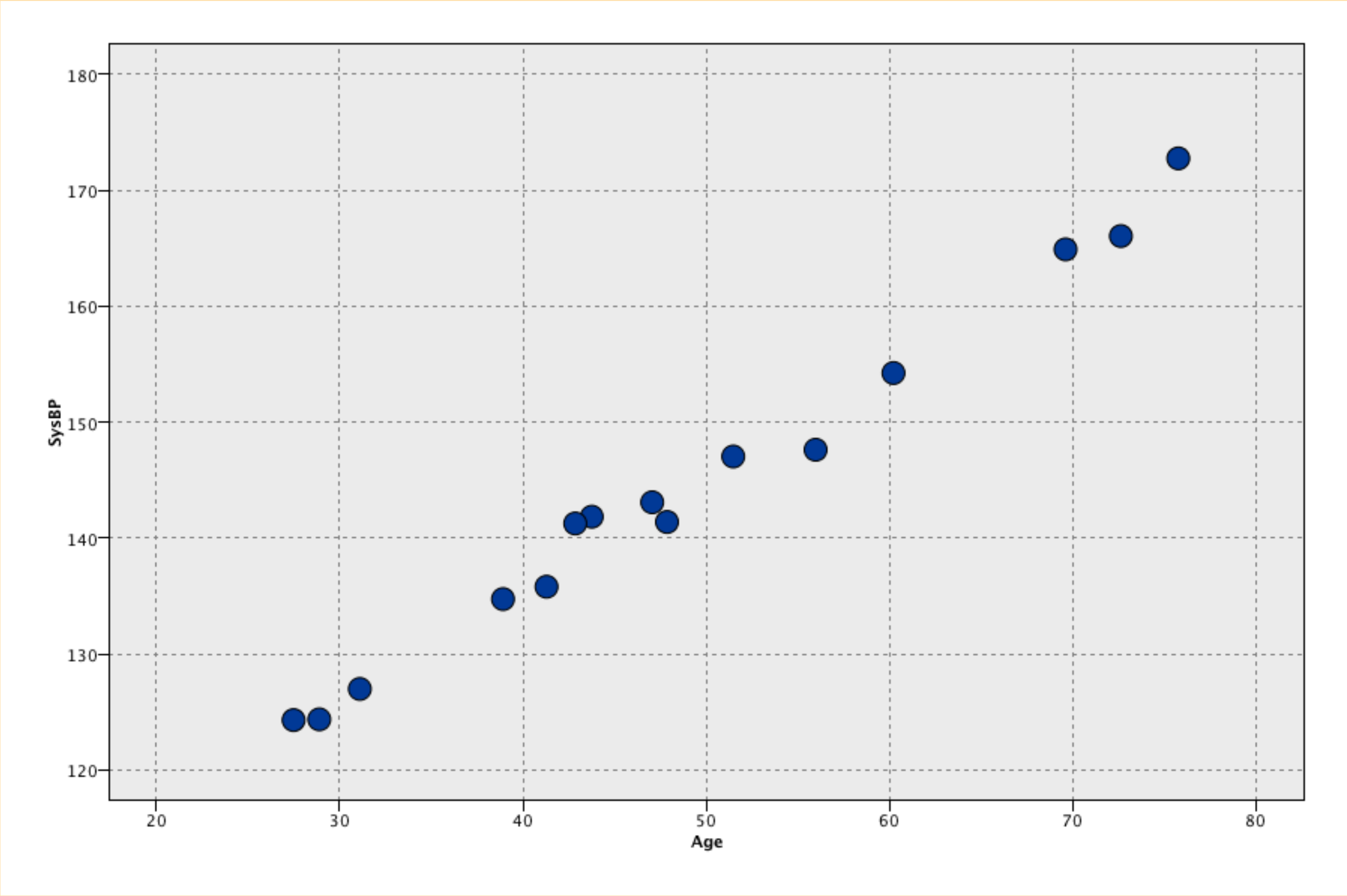
Making a prediction



Fitting linear regression

- But data (almost) never lie on a perfect straight line. The data also contain:
  1. Measurement error
  2. Between subject variability (e.g. not all people of a certain age have the same Systolic BP)

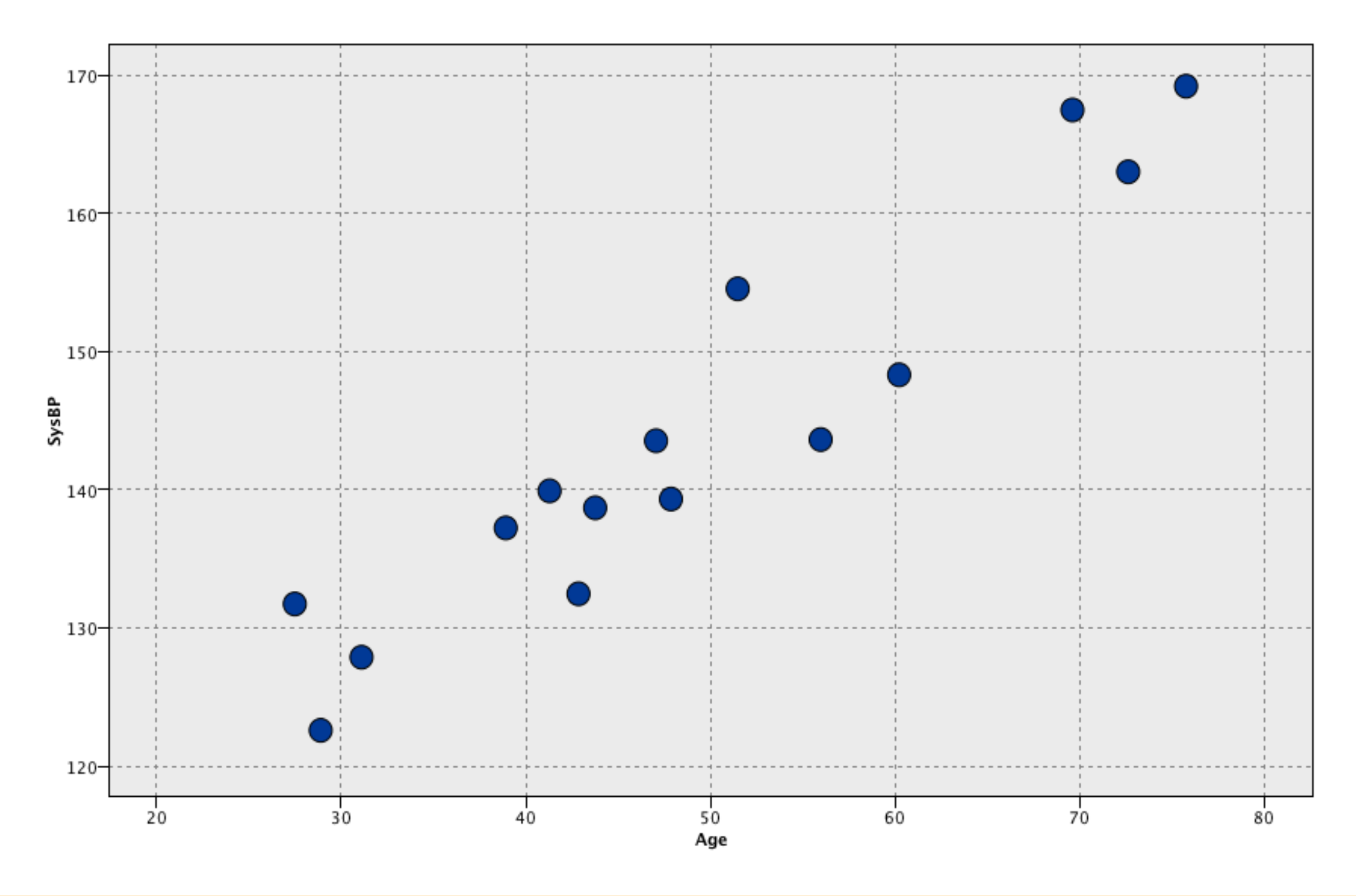
Data with measurement error



SysBP

Age

Data with measurement error plus biological variability

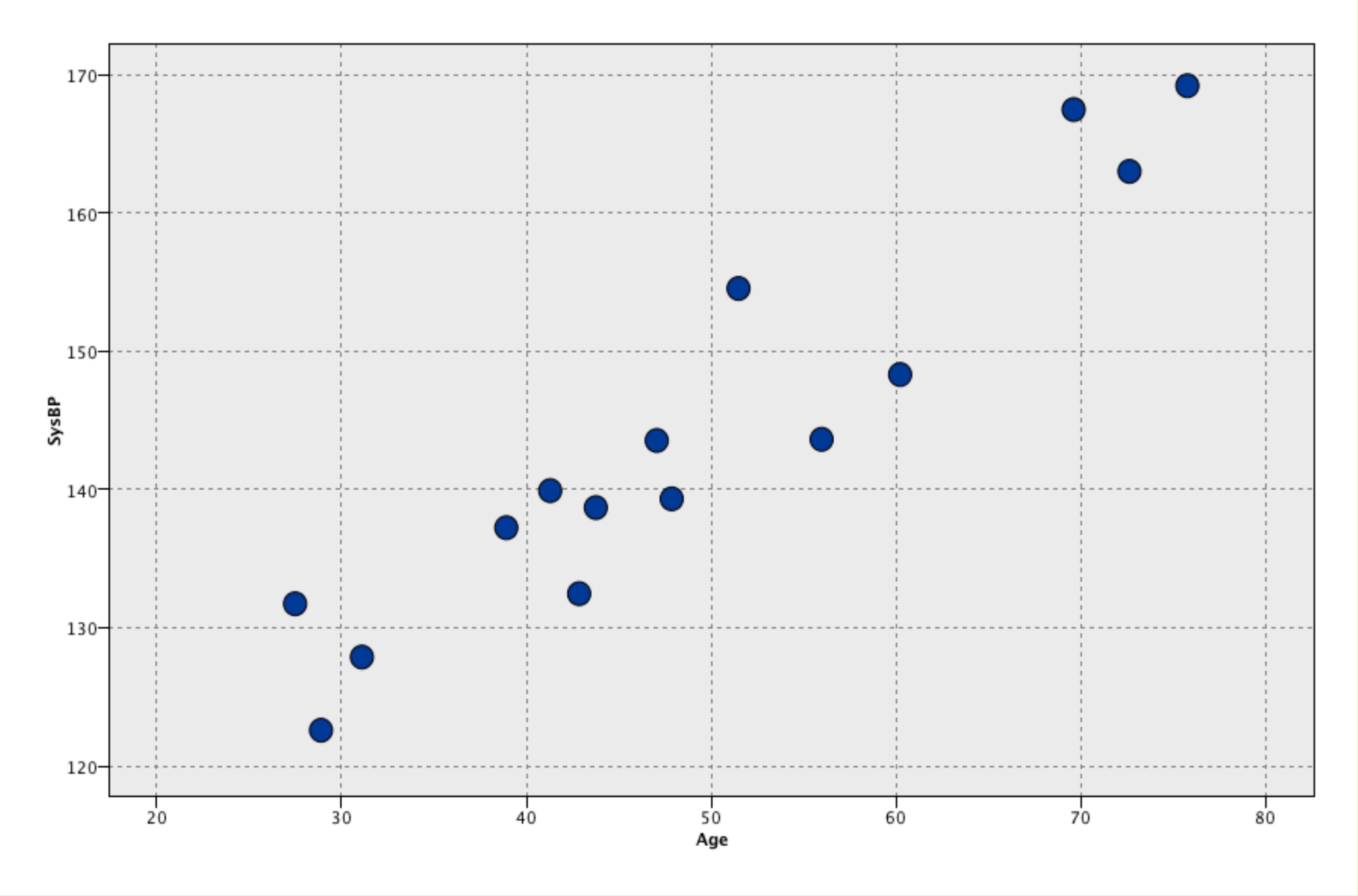


SysBP

Age

How to choose a fitted line?

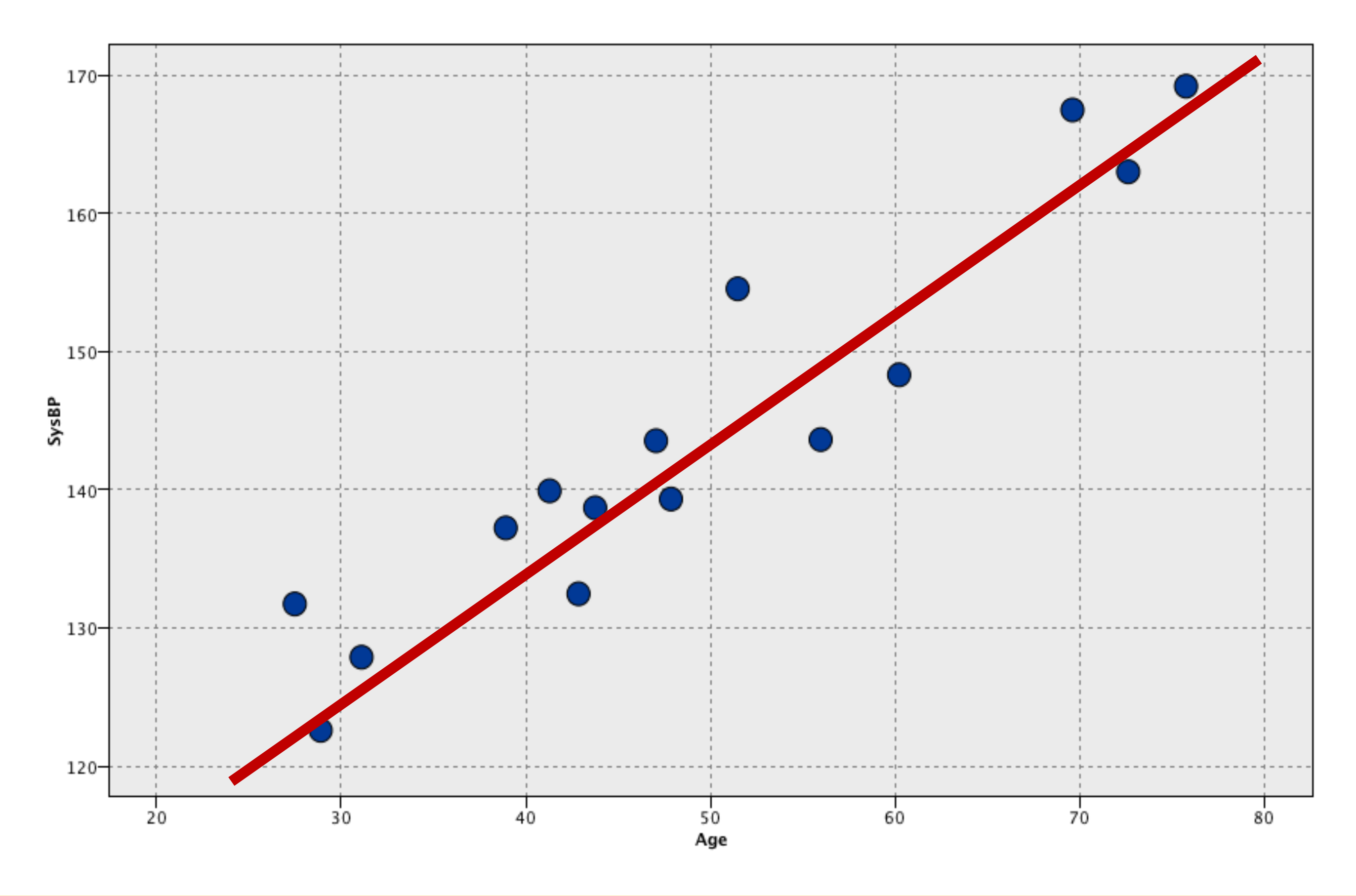
SysBP



Age

How to choose a fitted line?

SysBP

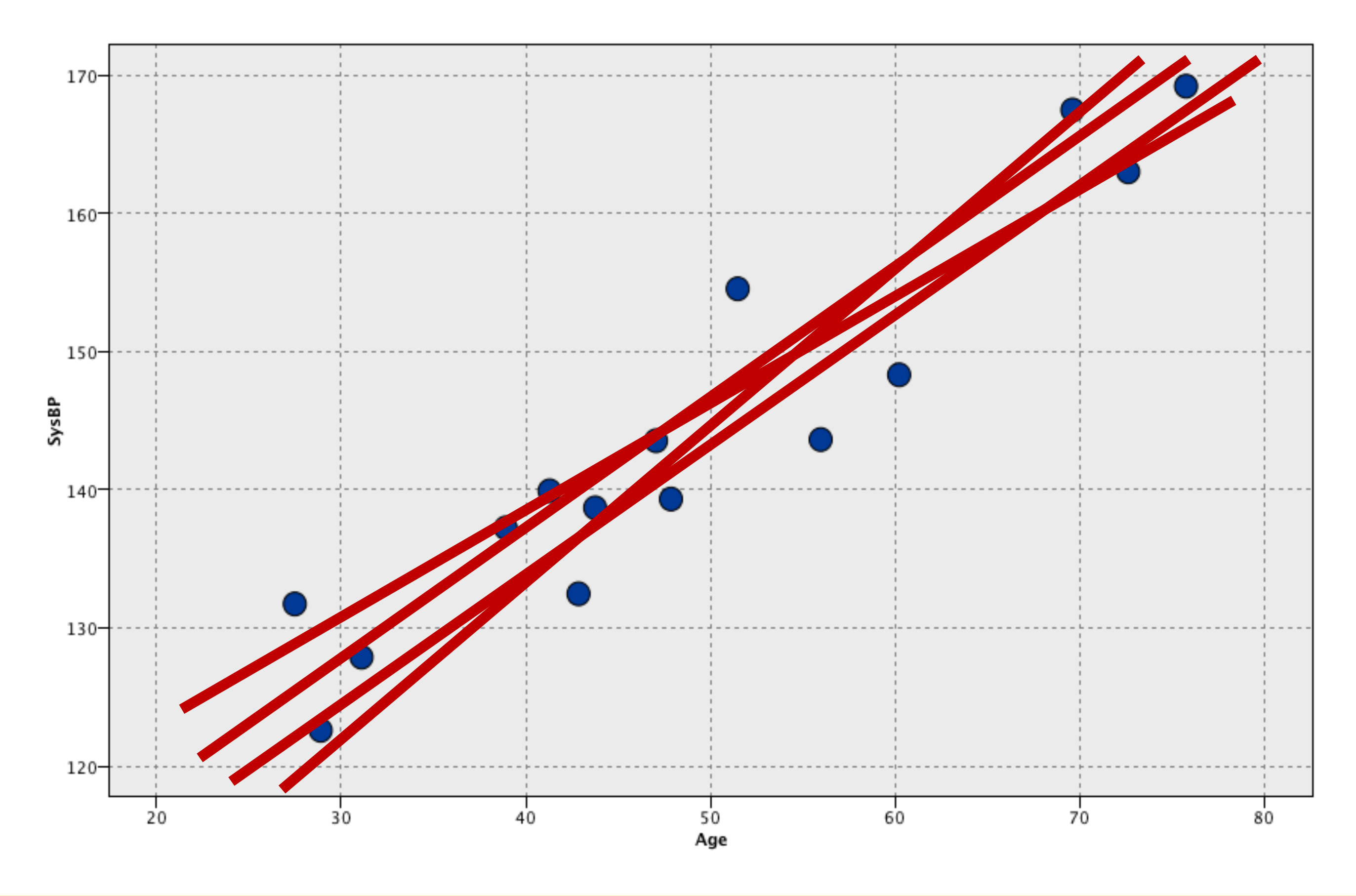


Age

One possible line

How to choose a fitted line?

We want to fit a straight line to give us a relationship, but which one is 'best'?

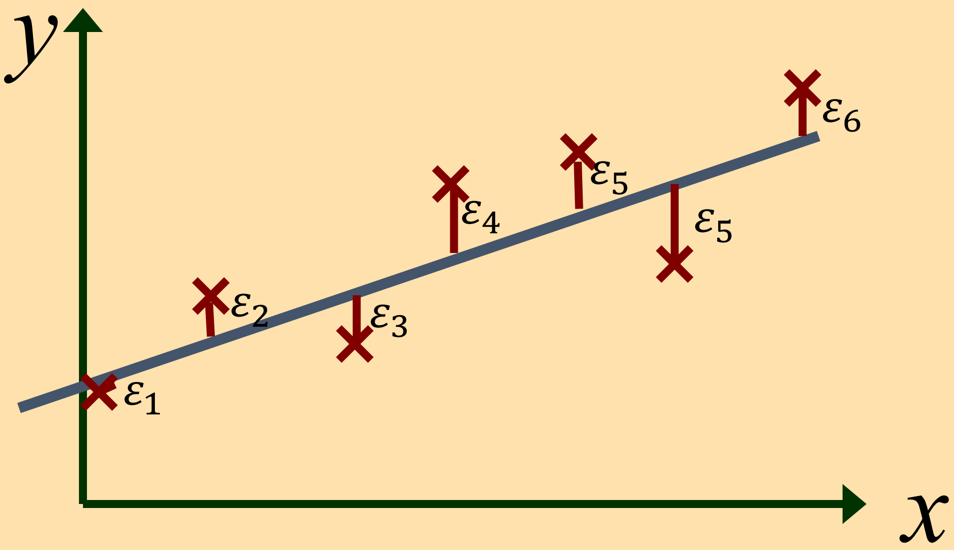


SysBP

Age

Best linear relationship

- Model relationship between x and y (e.g. linear)
- There will be some errors for each subject (ε)
- Find the line such that it gives you the smallest *sum of squared* errors



Add the squared values of the vertical red lines together:

$$SS = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2$$

choose line that minimizes this total SS

Model training

- *Least squares fitting* for linear regression/model

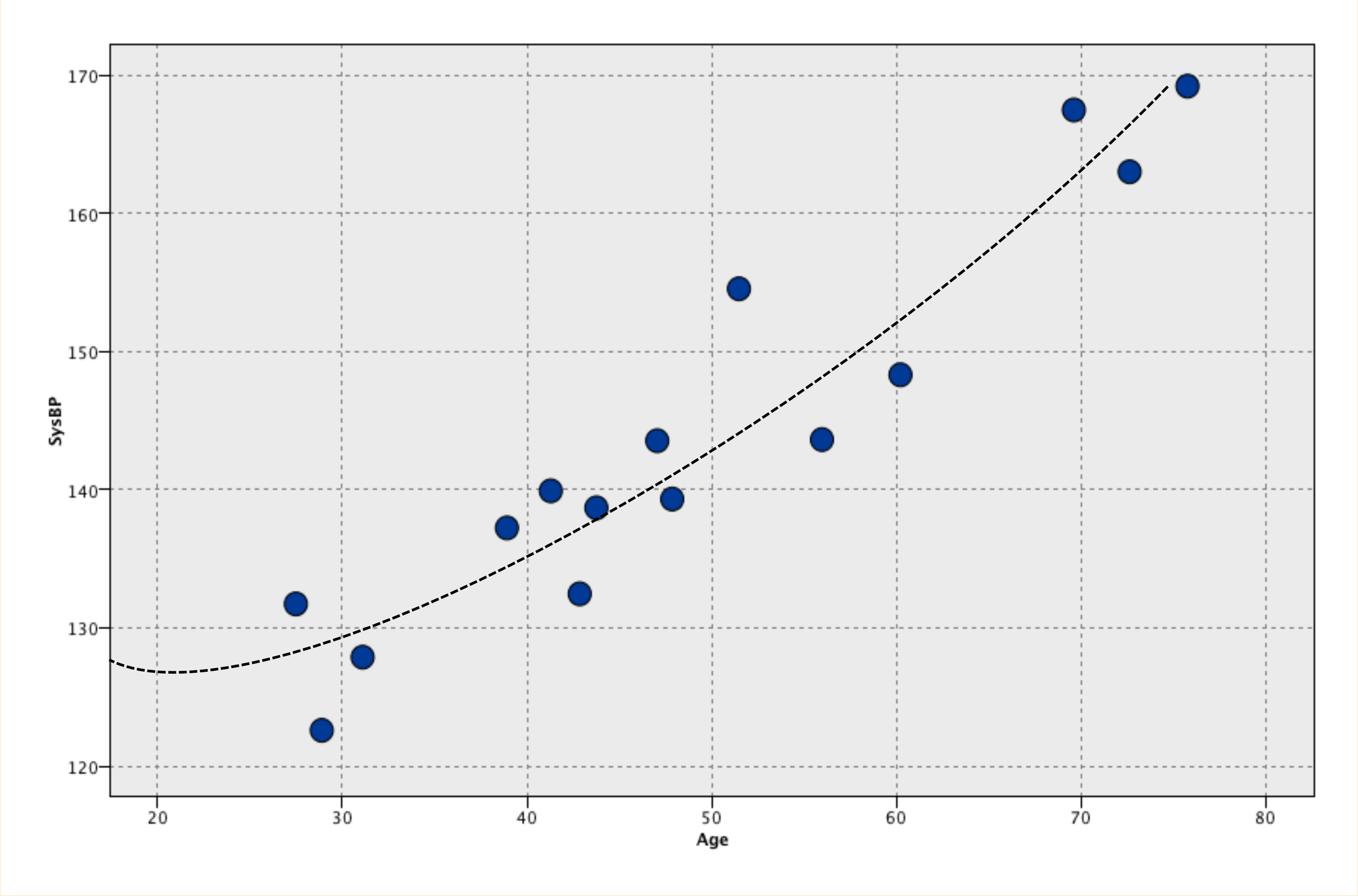
  Find the line such that the sum of all the square vertical distances from data points to the line is minimized (previous slide)

- Classical statistical inference focuses on the regression relationship between x and y -- usually concerned with whether there is evidence that the slope a is non-zero and what is its estimated value (and uncertainty)?

- In machine learning *emphasis* is more on using estimates of a and c to predict future values of y from future values of x -- interested in predictors that add predictive value, not evidence for non-zero relationship

Why straight line? What about?

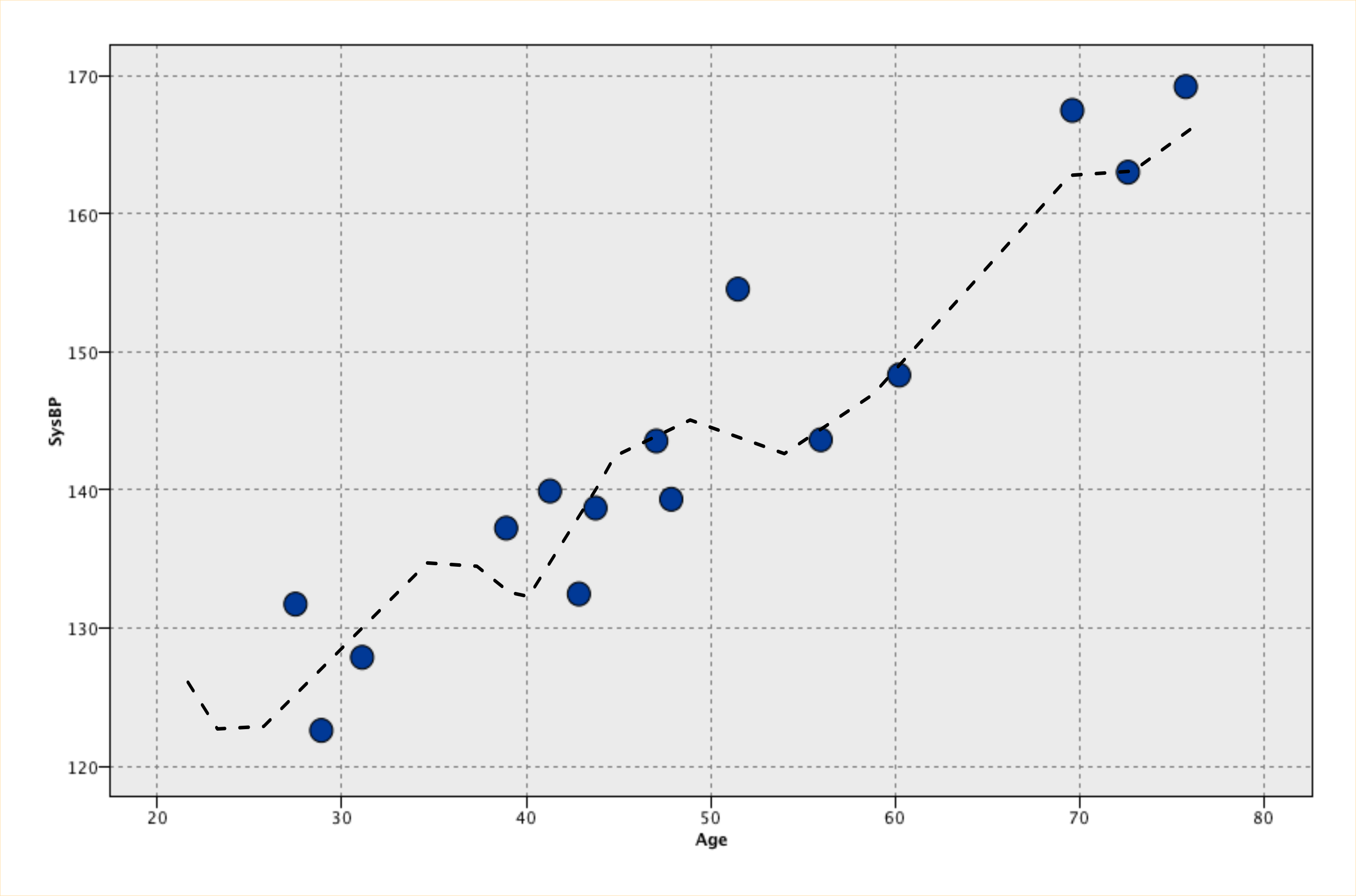
SysBP



Age

Or even?

SysBP



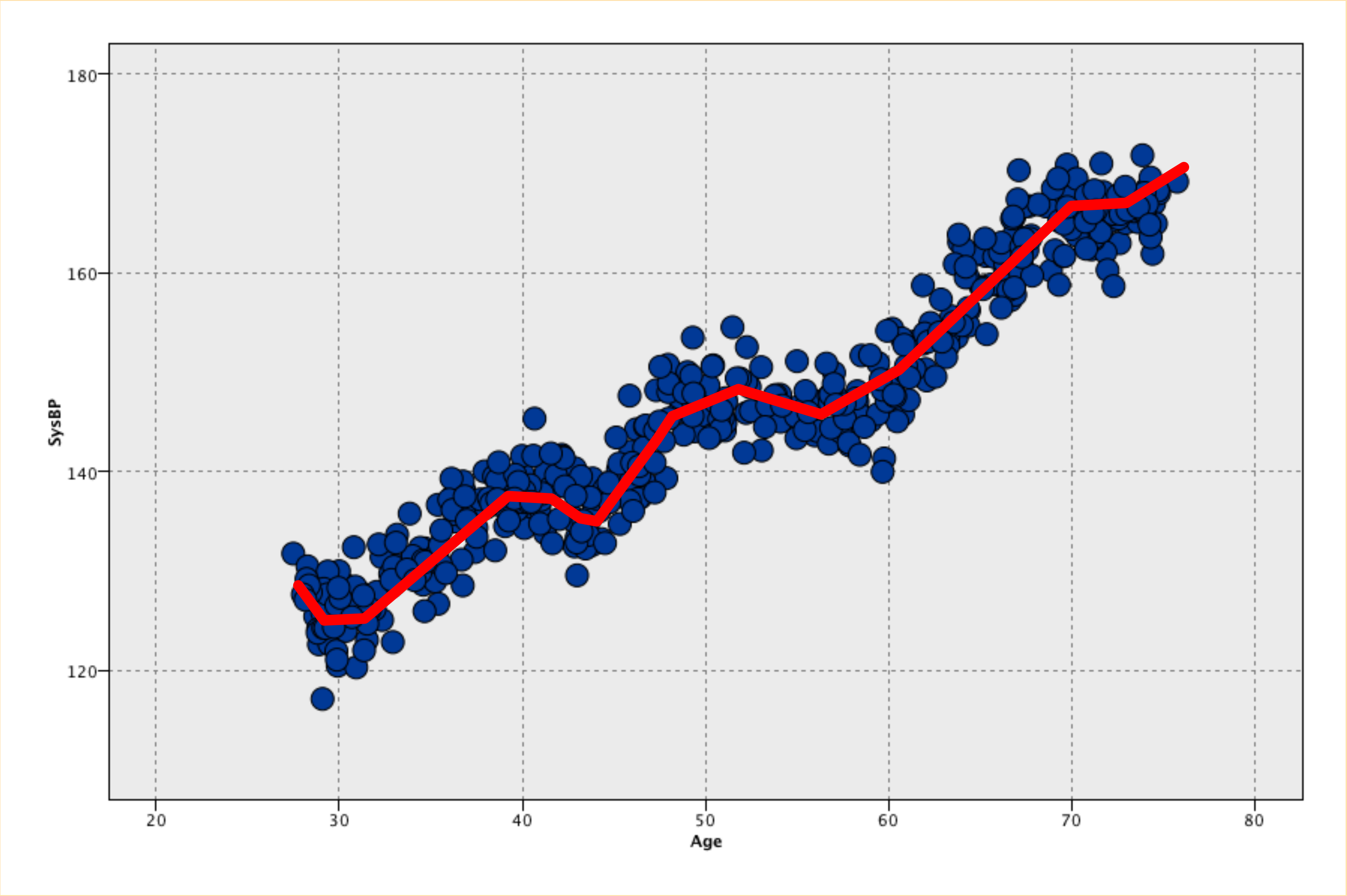
Age

Issue? Overfitted?

Overfitting

What about with more data?

SysBP

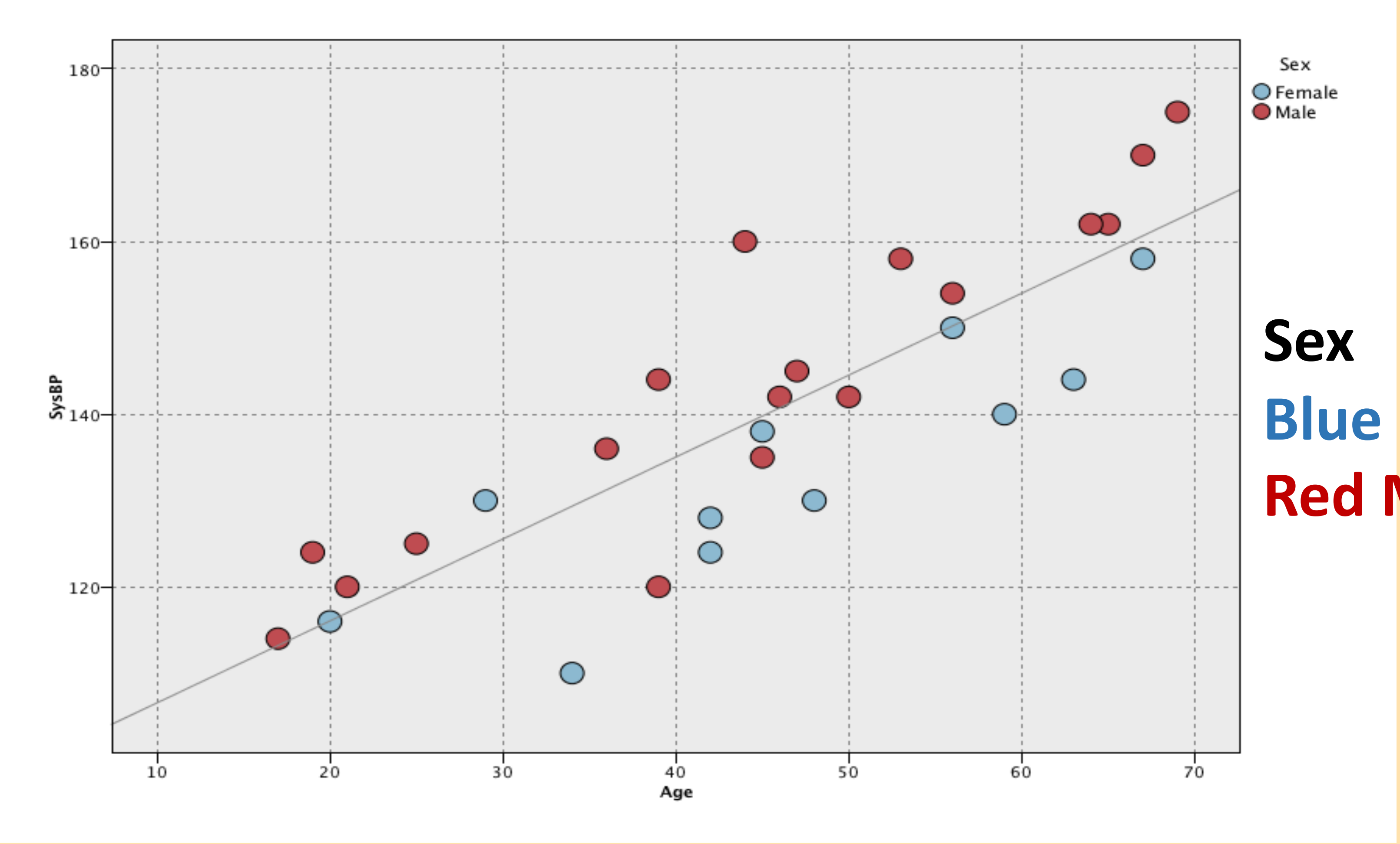


Age

Separate by Sex

What if we consider sex, in addition to age, as a predictor of SysBP?

SysBP



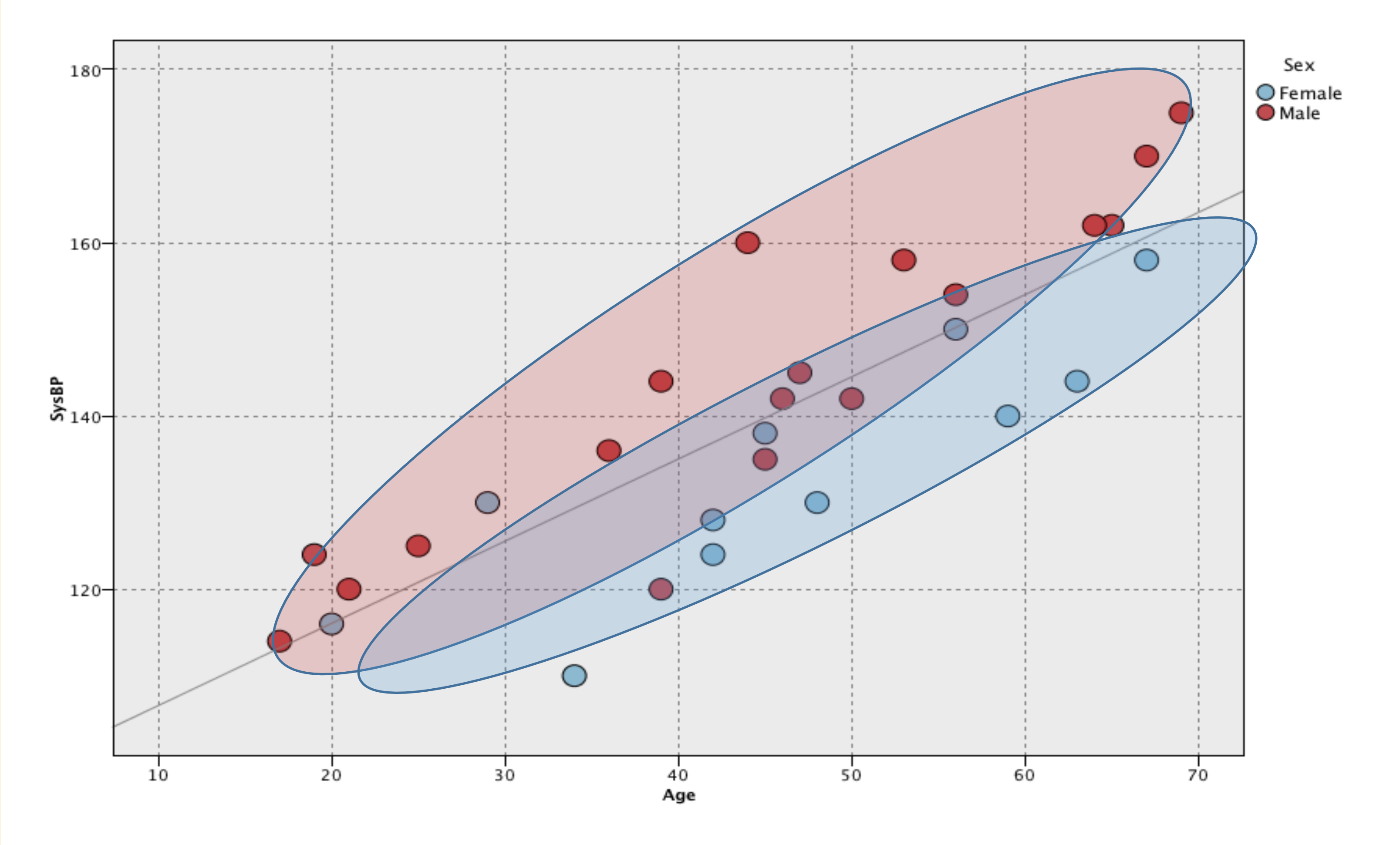
Age

Sex
Blue F
Red M

Separate by Sex

Do trends in SysBP vs Age differ based on sex?

SysBP



Age

Multi-predictor models

Now we might entertain a more complicated model:

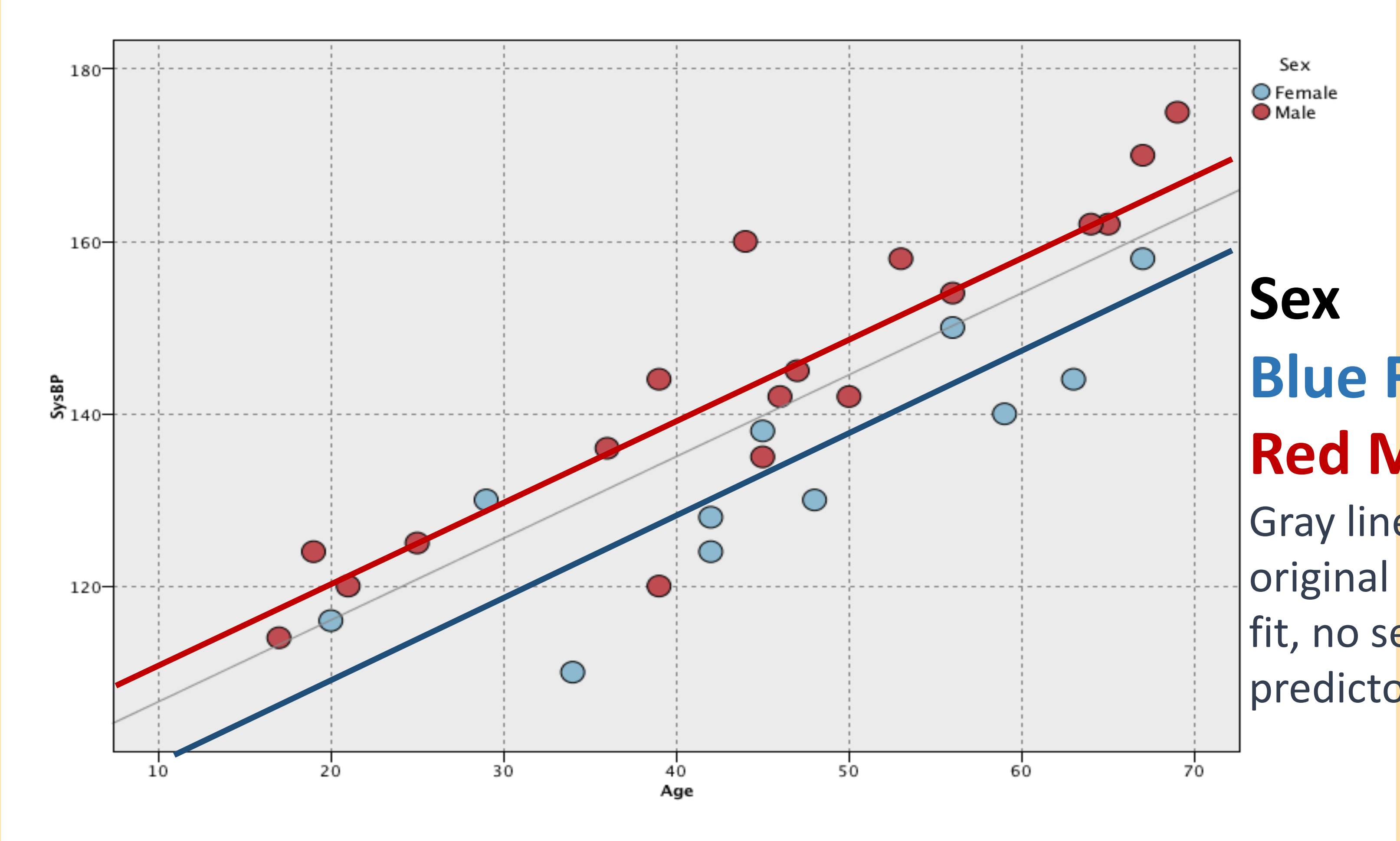
$$SysBP = a * Age + b * Sex + c$$

Sex , (1 = M, 0 = F)
c = intercept (SysBP at age 0 and Sex = F)
a = change SysBP per year in Age
b = difference in SysBP for M vs. F

Question: Are Age and Sex predictive of SysBP?

Age *and* Sex as predictors

SysBP



Sex

Blue F

Red M

Gray line = original fit, no sex predictor

Age

And interactions

But why not even more complicated such as:

$$SysBP = a * Age + b * Sex \ + w \ * Age * Sex + c$$

Sex, (1 = M, 0 = F)

c = intercept (SysBP at age 0 and Sex = F)

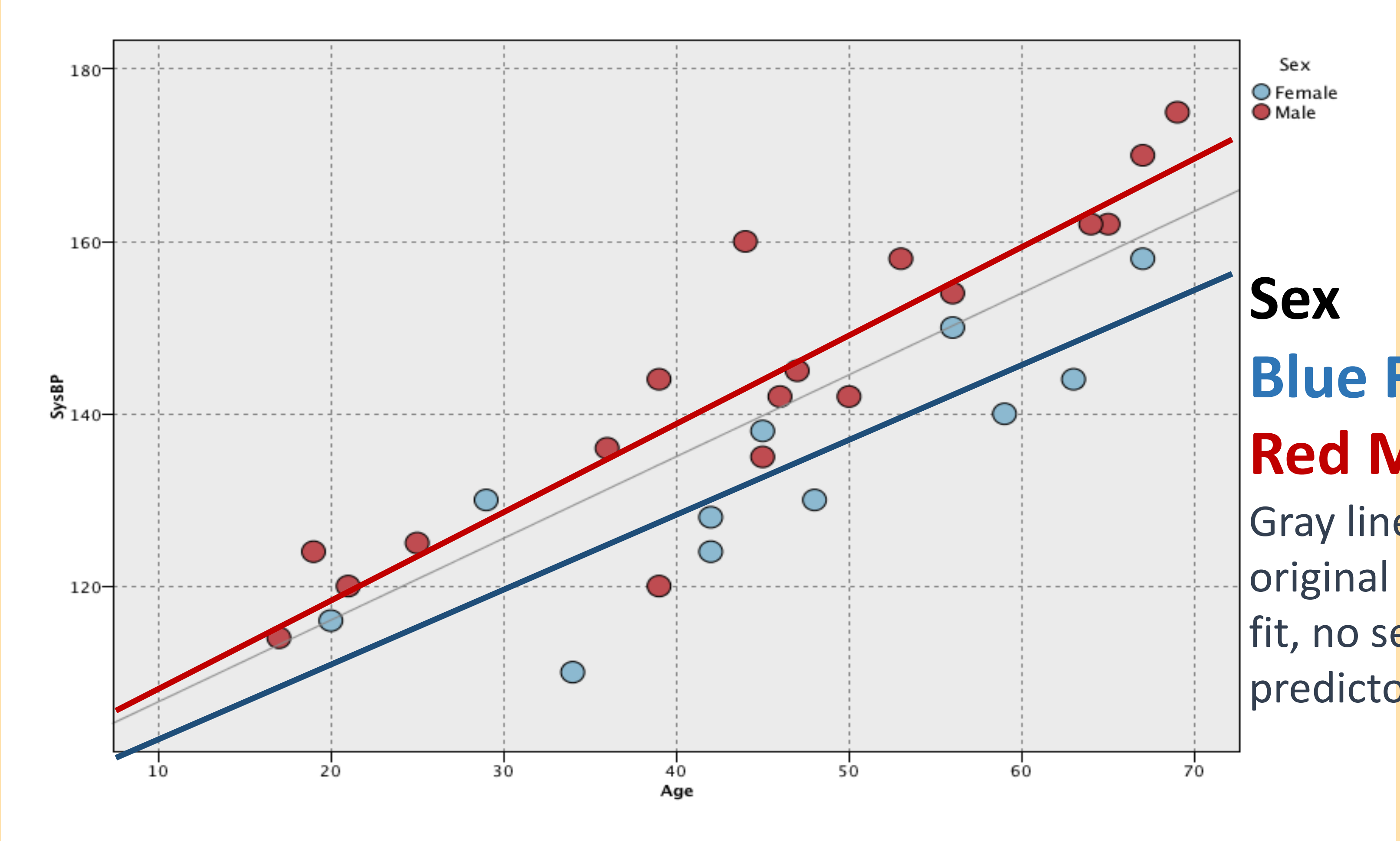
a = change in SysBP per year in Age for Sex = F

b = difference in SysBP for M vs. F at Age = 0

w = interaction effect - difference in changeSysBP per year in age for Sex = M vs. Sex = F

Question: Are Age and Sex and interaction of Age and Sex predictive of SysBP?

Age and Sex interaction



SysBP

Sex

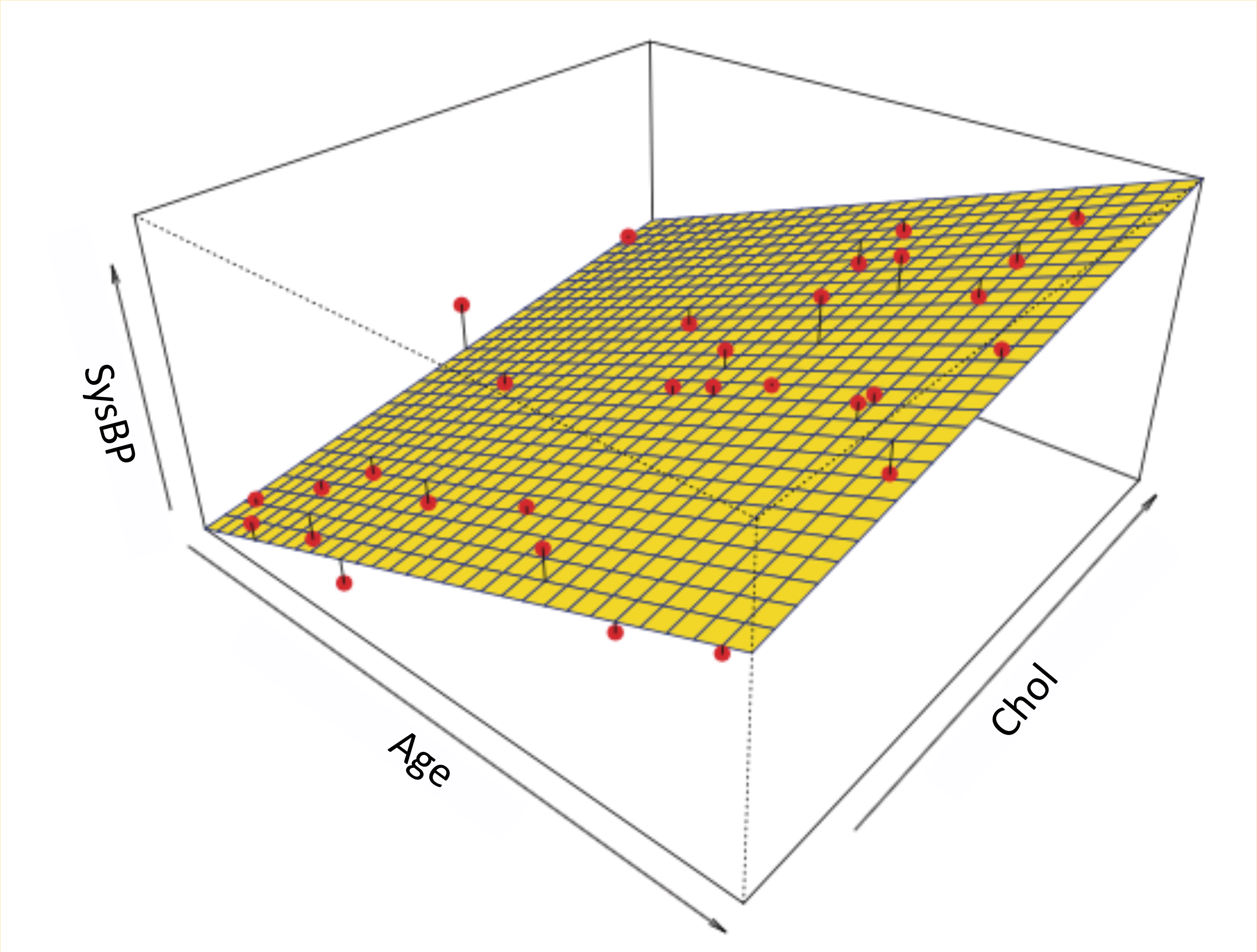
Blue F

Red M

Gray line = original fit, no sex predictor

Age

Linear regression with two (continuous) predictors



Multiple (predictor) linear regression

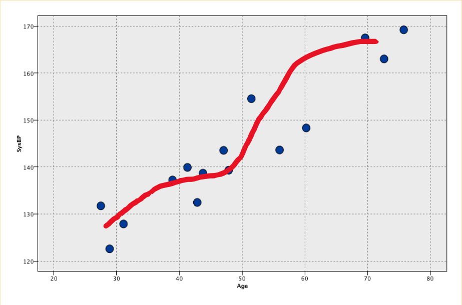
$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

- Question of interest: Which of the features $x_1, x_2, \ldots, x_n$ are predictive of y and how do we model their effect on y?

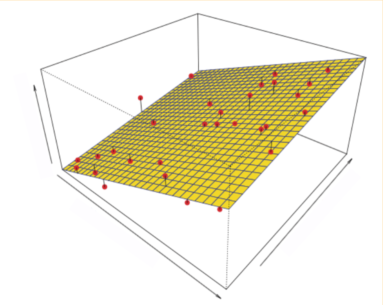
Model selection

- 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.

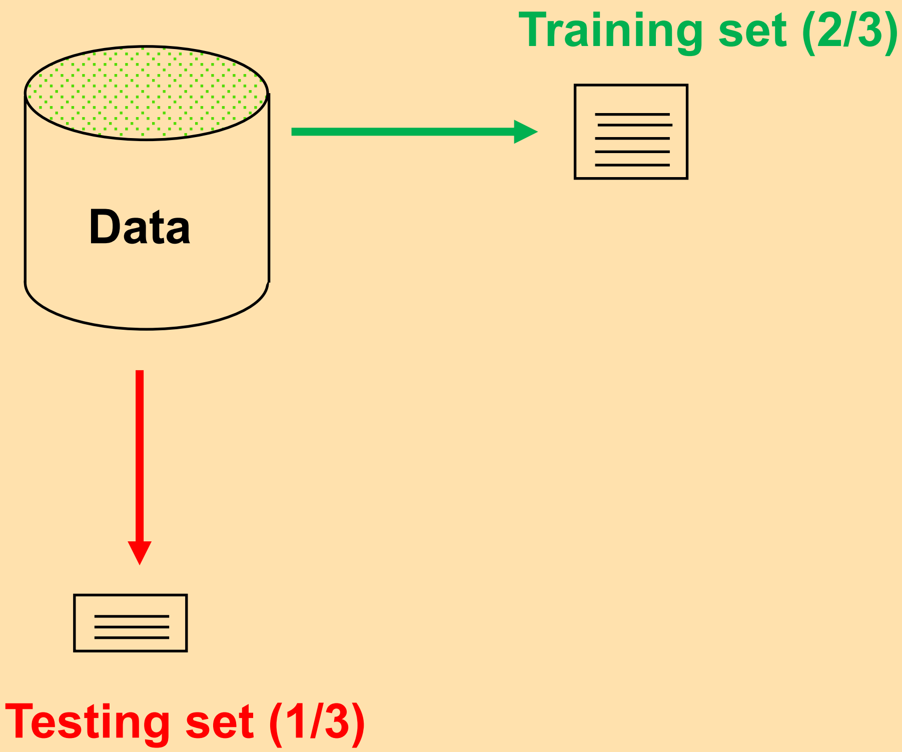
Validation

- **Validation**: an approach for model evaluation and comparison
- **Select** the model that produce the lowest predictive error in independent validation data.
- But we can’t use the same data we trained the model to in order to assess how well it did – could be perfect fit but not generalize to other data
- So we deliberately leave data aside for “testing”…

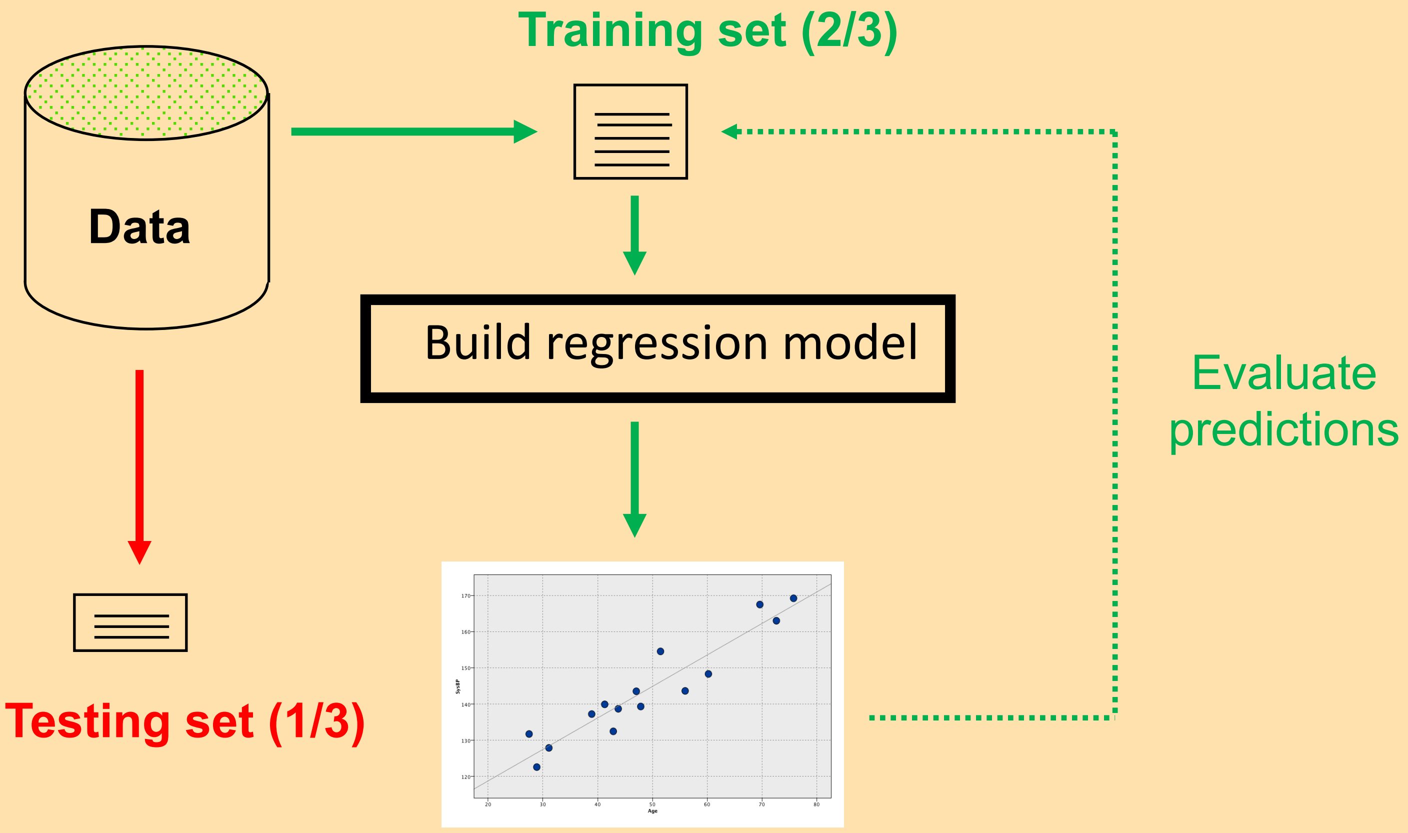
Validation using training/test data

- Randomly split data into training and test sets (often ~2/3 for train, 1/3 for test)
- Build a regression model using the *training* set and evaluate it using the *test* set
- Strategies vary based on the # instances/observations (at least 100, preferably > 1000) – depends on model
- Note: we will return to this theme continuously, it is a critical aspect of machine learning

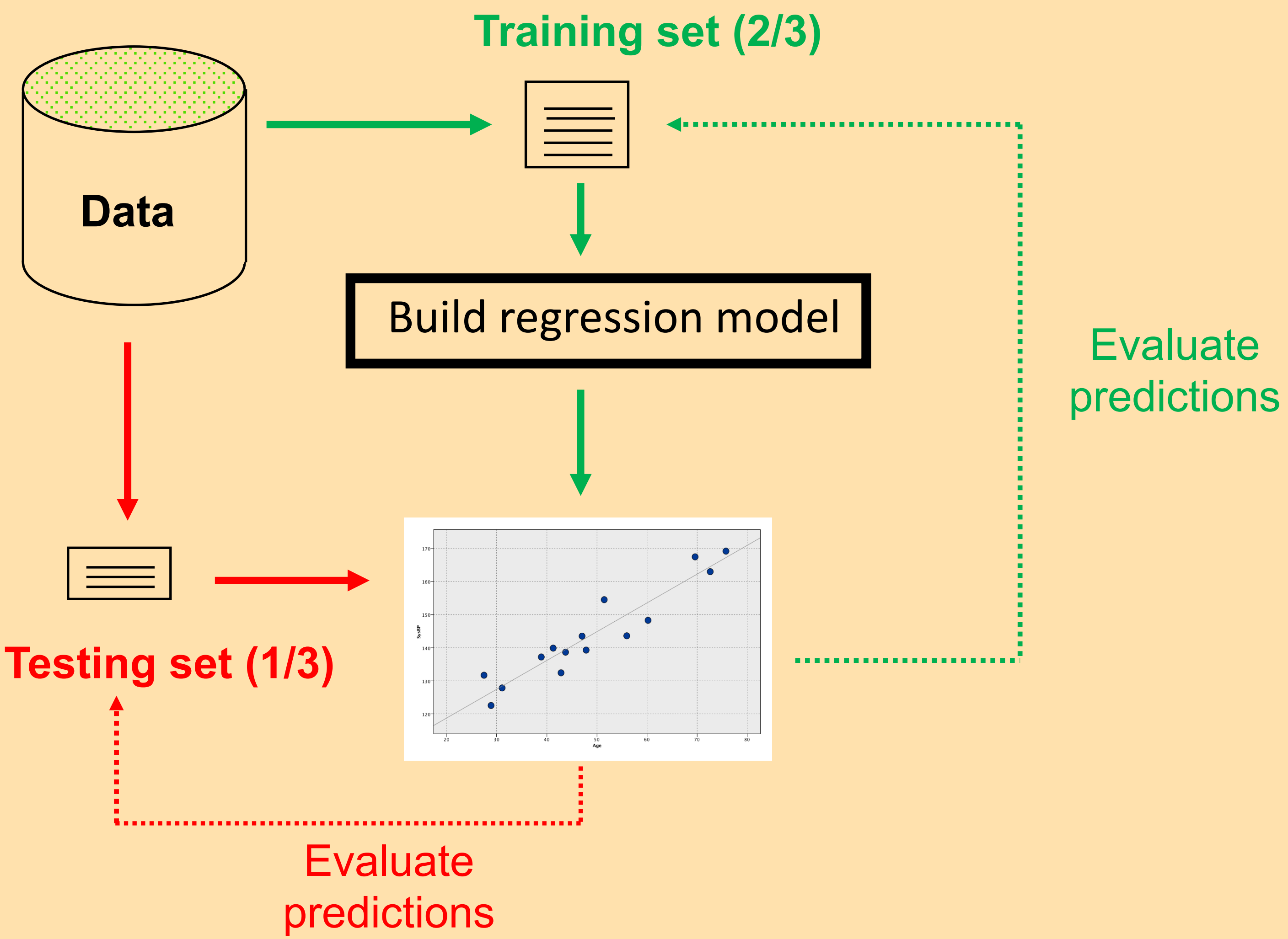
Training/Test



Training/Test

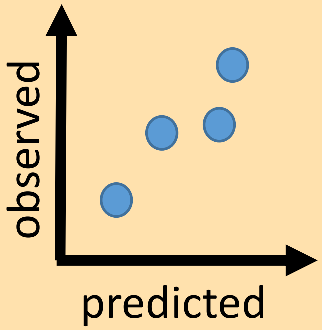


Training/Test

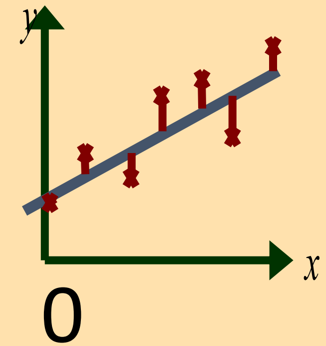


Evaluation metrics for continuous predictions

- Linear correlation^2 (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



Analysis pipeline

1. Data

2. Train/fit model

Algorithm(s)

3. Output

- Prediction
- Dimension Reduction
- Clustering

4. Interpret

= INFORMATION?

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 categorical predictions
- Validation and testing, continued
- Logistic regression

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)

<u>Goal: Predict systolic blood pressure using multiple linear regression.</u>