ECE5984 – Applications of Machine Learning
Lecture 1 – Introduction; Data

Creed Jones, PhD

Today's Objectives

Course Introduction
- Syllabus
- Course Objectives
- Your Instructor

Chapter 1 – Introduction
- 1.1 What Is Predictive Data Analytics?
- 1.2 What Is Machine Learning?
- 1.3 How Does Machine Learning Work?
- 1.4 What Can Go Wrong with Machine Learning?
- 1.5 The Predictive Data Analytics Project Lifecycle: CRISP-DM
- 1.6 Predictive Data Analytics Tools
- 1.7 The Road Ahead

ECE 5984 SPECIAL STUDY: APPLICATIONS OF MACHINE LEARNING

T Th 6:30–7:45 p.m. (Durham 261, and Zoom)

Description

Introduction to Machine Learning (ML) for predictive data analytics. Probability for ML including conditional probability, the product and chain rule, and the Theorem of Total Probability. Data preparation for ML algorithms, normalization, cleaning, and imputation of missing values. Information-based learning using decision trees. Similarity-based methods, data classification and clustering. Probability-based learning, conditional probability and Bayes' theorem, and applications. Linear and logistic regression and optimization-based learning. Performance evaluation of ML systems. Real-world applications of ML and case studies. Pre: Graduate Standing. (3H, 3C).

The complete syllabus is on the Canvas site.

Applications of Machine Learning: Course Learning Objectives

Having successfully completed this course, the student will be able to

1. Apply standard Machine Learning (ML) approaches in real-world scenarios using software tools for predictive data analysis.
2. Prepare raw data sets for use by ML algorithms and software using appropriate techniques.
3. Formulate decision-tree solutions in information-based learning applications.
4. Perform data classification and clustering for ML applications using similarity metrics.
5. Compute probability-based solutions for inference and prediction using Bayes' theorem.
6. Apply optimization-based learning and regression techniques to engineering applications.
7. Evaluate ML approaches and systems using standard performance measures for specific case studies.

Please review the syllabus carefully – I plan to follow the schedule, grade breakdown and course objectives as closely as possible

			Advance	ECE5984 SP22 Daily Schedule	
Day	**Module**	**Lec**	**Reading**	**Topics**	**Due**
18-Jan	I - Foundations	1	1.1 - 1.7	Course introduction	
20-Jan		2	App. D	Review of linear algebra	
25-Jan		3	App. B & A	Review of statistics	
27-Jan	II - Data Prep	4	3.1 - 3.5	Data exploration	quiz 1
1-Feb		5		More on data exploration and presentation	
3-Feb		6		Python and sklearn	
8-Feb		7	3.4	Missing values	hw 1
10-Feb		8	3.6	Data preparation	quiz 2
15-Feb	III - Information and Similarity	9	4.1 - 4.3	Introduction to decision trees	hw 2
17-Feb		10	4.4 - 4.5	More on decision trees	
22-Feb		11	5.1 - 5.23	Similarity measures	
24-Feb		12	5.4	Classification	quiz 3
1-Mar	IV - Probability	13	6.1 - 6.2	Probability-based learning; Bayes' theorem	hw 3
3-Mar		14	6.3 - 6.4	Bayesian prediction	quiz 4
8-Mar	*No Class - Spring Break*				
10-Mar					
15-Mar	V - Gradient-based methods	15	7.1 - 7.2	Gradient-based methods	
17-Mar		16	7.3 - 7.4	Multivariate linear regression	
22-Mar		17	5.4.6	Variable selection	prj 1
24-Mar		18	7.4	Logistic regression	quiz 5
29-Mar	VI - Performance	19	9.1 - 9.3	Performance evaluation; misclassification	
31-Mar		20	9.4	ROC curves; other performance measures	
5-Apr		21	4.4.5	Model selection / ensemble models	hw 4
7-Apr	VII - Neural networks	22	8.1 - 8.3	Neural networks	quiz 6
12-Apr		23	8.4	More on neural networks	
14-Apr		24		Deep learning	
19-Apr		25	7.4.7	Support vector machines	hw 5
21-Apr	VIII - Other methods	26		Other modeling techniques	quiz 7
26-Apr		27	10.1 - 10.5	Unsupervised learning	
28-Apr		28	11.1 - 11.5	Reinforcement learning	hw 6
3-May		29		Course review	quiz 8 / prj 2
5-May	*No Class - Reading Day*				
7-May	**FINAL EXAM (7:00 to 9:00 PM)**				

VIRGINIA TECH.

ece
BRADLEY DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

We will have six homework assignments, eight brief quizzes, two team projects, a technical paper review and a final exam

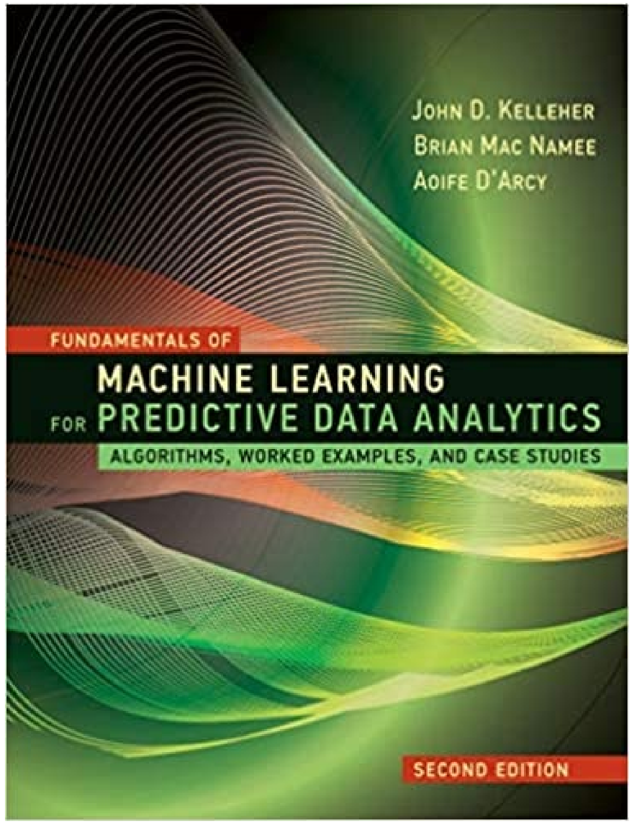
- Most homework assignments and the projects will require you to implement a model in Python
- Quizzes are short (ten questions) and will take less than fifteen minutes

Graded Item	# of Items	Points per Item	Total Points	Percentage
Homework Assignments	6	25	150	30%
Projects	2	60	120	24%
Technical Paper Review	1	50	50	10%
Final Exam	1	100	100	20%
Quizzes	8	10	80	16%
			500	100%

Required Resources

- **Textbook:** Kelleher, Mac Namee and D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 2nd edition, MIT Press, 2015, ISBN 978-0-2620-4469-1

- **Software:** We will be doing several programming assignments in this course, using Python with numpy and scikit-learn. For Python, I recommend PyCharm from JetBrains (freely available using your VT email address).
  We will also be using Tableau, also available to you using an edu email address.

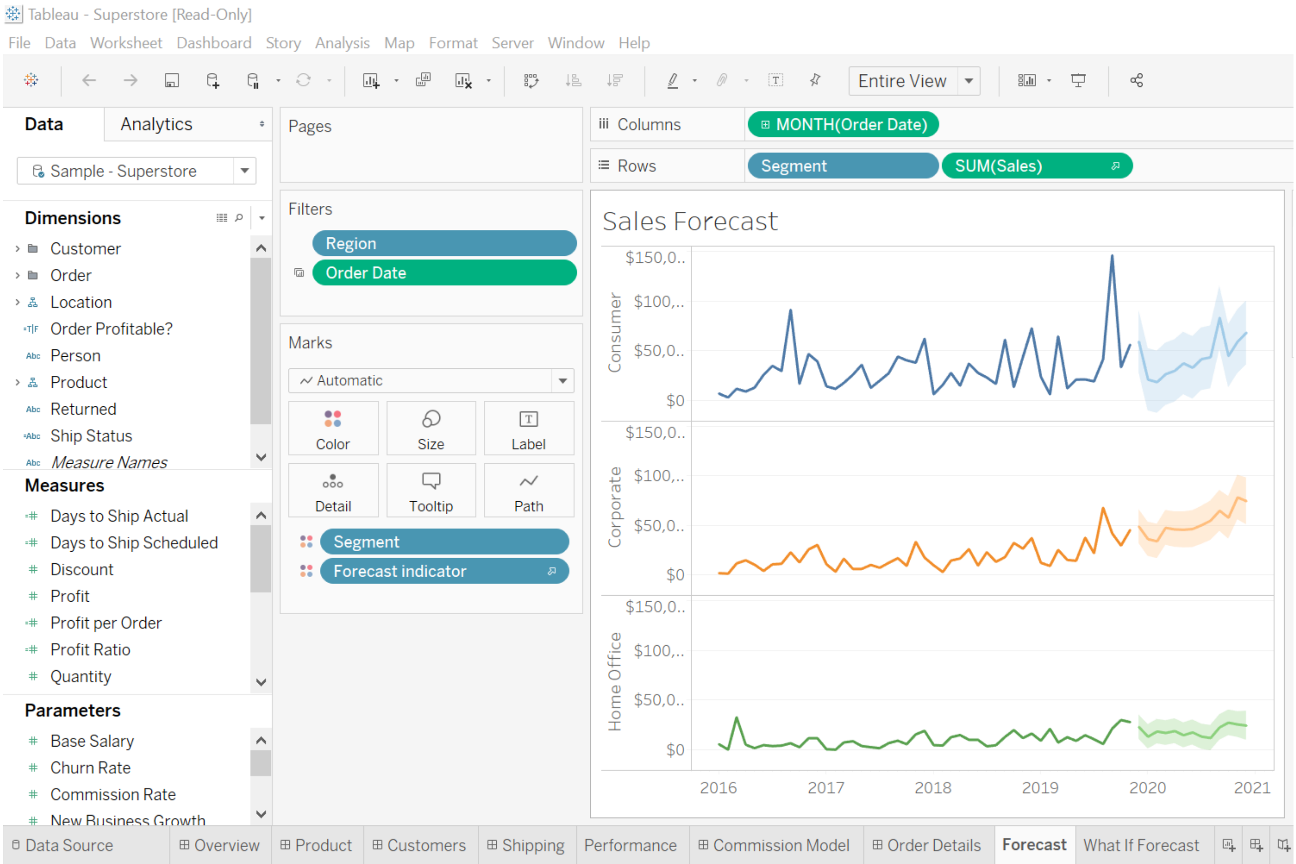


We will be using two software environments: Tableau for data exploration Python/numpy and scikit-learn for modeling

VIRGINIA TECH

ece BRADLEY DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

- With your VT student email, you can get a free license for Tableau
- Visit https://www.tableau.com/academic/students
- In the next week, download and install Tableau on your laptop (or desktop if need be)
- More info on Python and scikit coming soon

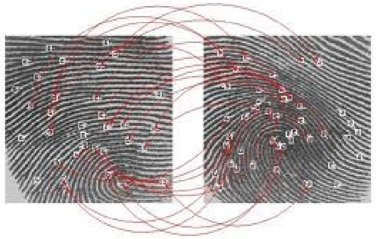


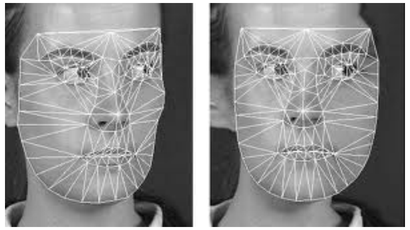
As a student in this course, I expect you to adhere to the ten points of the Hokie Community Wellness Commitment. If necessary, I will take steps to ensure your adherence.

Community Wellness Commitment:

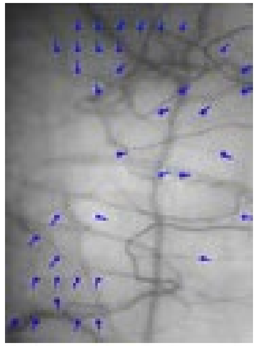
- We will affirm our commitment to the safety, health, and well-being of our campuses and local communities.
- We will affirm that we will support the mental well-being of all community members.
- We will wear face coverings/masks in public areas.
- We will practice physical distancing by maintaining at least 6 feet of distance from others.
- We will practice good hygiene, including frequent handwashing and covering coughs or sneezes.
- We will stay home and avoid public spaces when not feeling well.
- We will contact a health care provider or an urgent care facility if we believe we are sick or have been exposed to the coronavirus.
- We will support but avoid contact with those who are sick.
- We will follow public health guidelines and medical recommendations to be tested and self-isolate as necessary.
- We will make a list of all others with whom we have had close contact, if necessary, to aid in contact-tracing efforts.

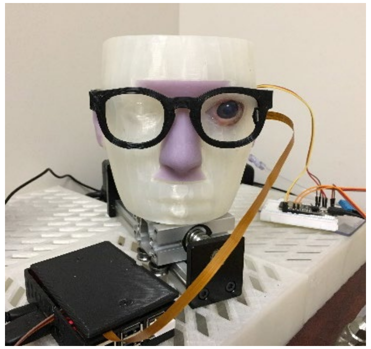
Creed Jones, PhD – Collegiate Professor of ECE





- First-generation college student
- Engineering scholarship; BS/MS
- 25 years in industry
- Image Processing at GM – Perceptics – Optimas – Avereon – Sagem Morpho…
- PhD at Virginia Tech, 2005 (Use of Color for Face Recognition)
- Faculty at Seattle Pacific Univ / California Baptist Univ
- Predictive Modeling at Humana
- Joined Virginia Tech in Fall 2019
- Globe Biomedical – startup in imaging for eye health
- Eight patents (to date)
- International industry standards
- Seven years' industry experience in machine learning and predictive modeling for medical and health care applications





(12) **United States Patent**
Jones, III et al.

(10) **Patent No.: US 8,543,428 B1**
(45) **Date of Patent: Sep. 24, 2013**

(54) **COMPUTERIZED SYSTEM AND METHOD FOR ESTIMATING LEVELS OF OBESITY IN AN INSURED POPULATION**

(75) Inventors: **Creed Farris Jones, III**, Louisville, KY (US); **Diana J. Beasley**, Louisville, KY (US); **Farooq Azam**, Porspect, KY (US); **John Louis Kucera**, Louisville, KY (US); **Carol Jeanne McCall**, Libertyville, IL (US)

(73) Assignee: **Humana Inc.**, Louisville, KY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 853 days.

(21) Appl. No.: **12/635,043**

(22) Filed: **Dec. 10, 2009**

(51) **Int. Cl.**
G06Q 40/00 (2012.01)

(52) **U.S. Cl.**
USPC **705/4**

(58) **Field of Classification Search**
USPC 705/4
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,322,504 B1 *	11/2001	Kirshner	600/300
7,194,301 B2 *	3/2007	Jenkins et al.	607/2
8,024,204 B1 *	9/2011	Goral	705/4
8,388,532 B2 *	3/2013	Morgan	600/301
2008/0051679 A1 *	2/2008	Maljanian	600/587
2008/0294370 A1 *	11/2008	Kriger	702/173
2011/0105852 A1 *	5/2011	Morris et al.	600/300
2012/0116801 A1 *	5/2012	Hu et al.	705/2

OTHER PUBLICATIONS

Kuriyama, S. et al. "Medical Care Expenditure Associated with Body Mass Index in Japan: The Ohsaki Study." International Journal of Obesity and Related Disorders 26.8 (2002): 1069-74 (6 pages).*
Rowald, Laura A. "Relationships Among Body Mass Index, Physical Activity Status, and Health-Related Quality of Life in Employed Adults", Diss. Southern Illinois University Carbondale, 2006. 3244485 (124 pages).*
Libann: Creating a Feature Vector. May 15, 2003. http://www.nongnu.org/libann/doc/libann_3.html>(2 pages).*

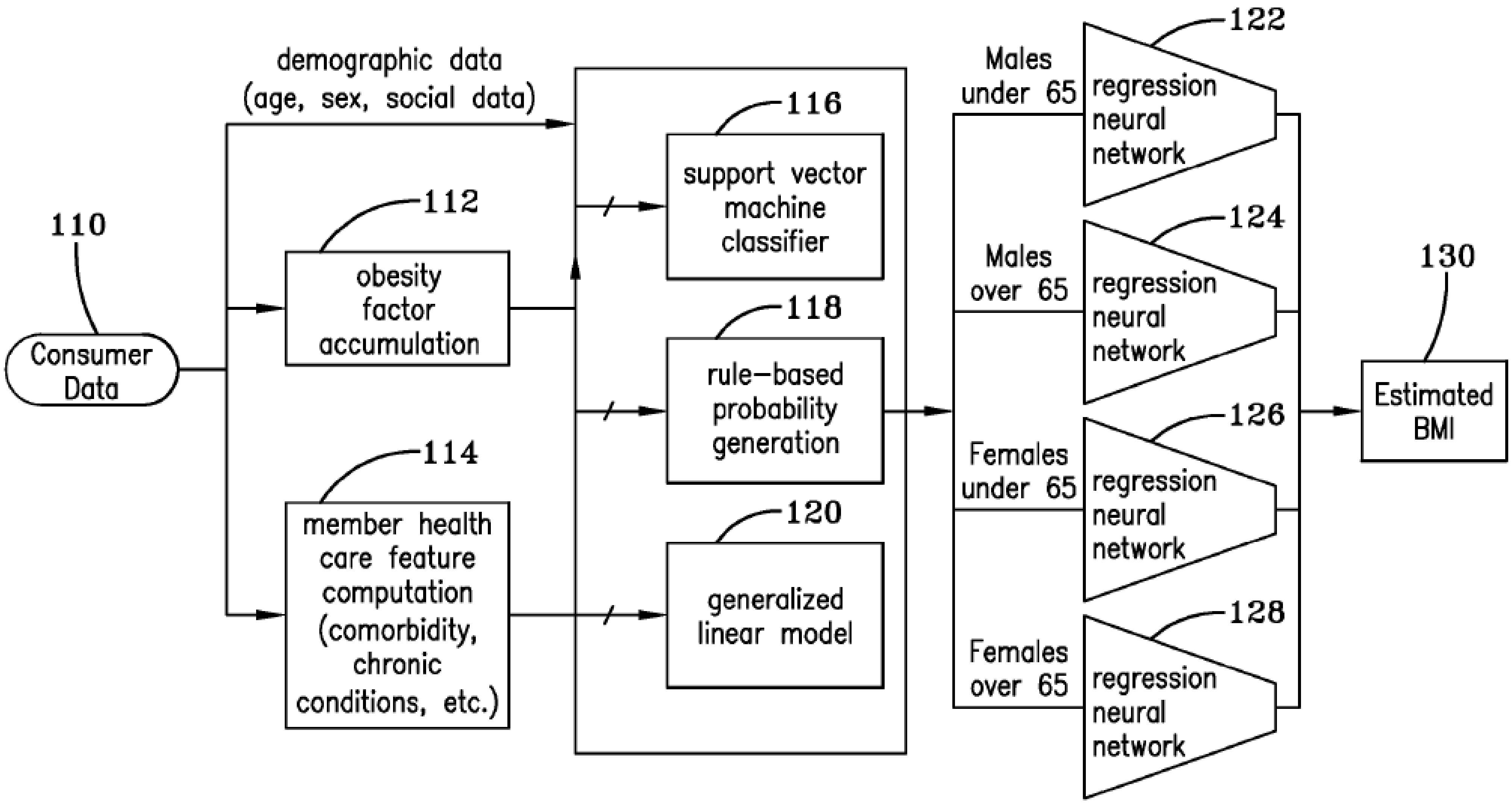
* cited by examiner

Primary Examiner — Elda Milef
(74) *Attorney, Agent, or Firm* — Standley Law Group LLP

(57) **ABSTRACT**

A computerized system and method for estimating levels of obesity in an insured population using claims data. The model uses health risk assessment data comprising age, height, and weight information as well as information about health conditions and health behaviors for a member population. Claims data is used to train a two-stage model on the member population. The first stage comprises a support vector machine, a rule-based module, and a generalized linear model that estimates the probability of obesity. The second stage comprises a regression neural network that operates on the output of the first stage and a subset of the input feature vector. Cost and utilizations in these areas, along with overall health measures as well as demographics and social factors, are inputs to a set of pattern recognition engines that perform regression. The output is the estimated body mass index of the member.

25 Claims, 4 Drawing Sheets


demographic data
(age, sex, social data)
110
Consumer Data
112
obesity factor accumulation
114
member health care feature computation (comorbidity, chronic conditions, etc.)
116
support vector machine classifier
118
rule–based probability generation
120
generalized linear model
Males under 65
122
regression neural network
Males over 65
124
regression neural network
Females under 65
126
regression neural network
Females over 65
128
regression neural network
130
Estimated BMI

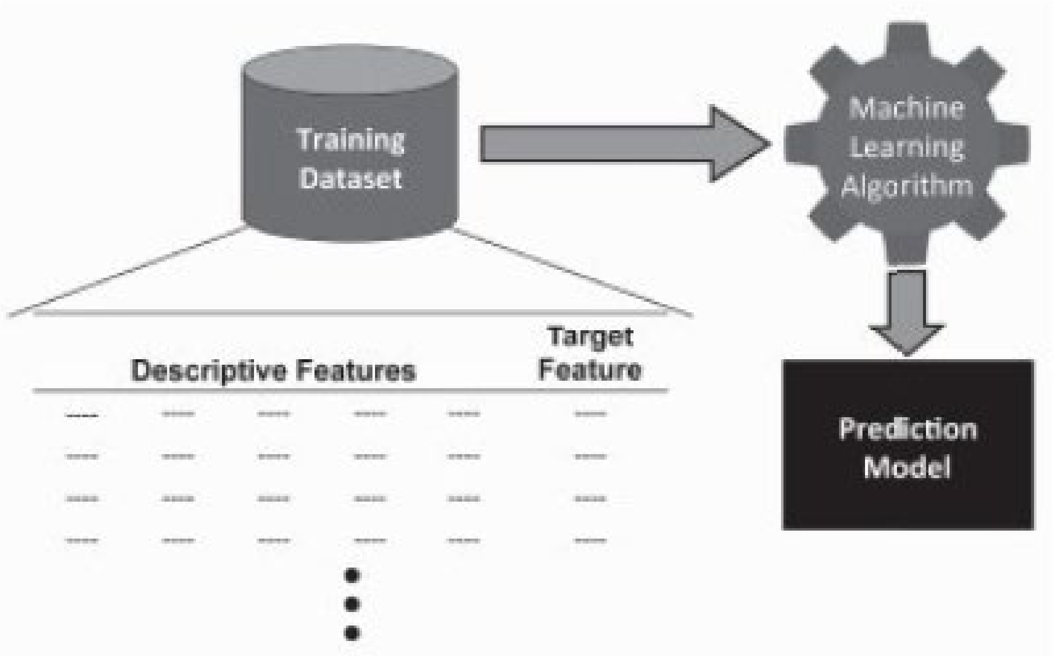
Predictive data analytics is the art of building and using models that make predictions based on patterns extracted from historical data

- We observe past relationships between observable variables
- Assuming that these relationships persist, we measure some variables and predict others

- What will the stock price of Amazon be next week?
- What is the risk of some particular action?
- What is the proper dose of a medication for a patient, given that we know their vital signs and health history?
- From measurements of a tree's leaves, can we predict what fruit it will bear?
- Which students in an incoming freshman class will graduate on time?

In supervised machine learning, we automatically learn a model of the relationship between a set of descriptive features and a target feature based on a set of historical examples.

We can then use this model to make predictions for new instances.



(a) Learning a model from a set of historical instances

(b) Using a model to make predictions

Consider a simple dataset of past history of clients, some of who defaulted on their loan

ID	OCCUPATION	AGE	LOAN-SALARY RATIO	OUTCOME
1	industrial	34	2.96	repay
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.8	default
6	industrial	61	2.52	repay
7	professional	37	1.5	repay
8	professional	40	1.93	repay
9	industrial	33	5.25	default
10	industrial	32	4.15	default

- From this past data, we want to develop a relationship between the *descriptive features* (age, etc.) and the *target feature* (outcome)
- This is a *training set;* each line is an *instance*
- How we derive a relationship is our machine learning algorithm
- For this simple dataset, we can say:

  if LOAN-SALARY RATIO > 3 **then**

  OUTCOME = *default*

  else

  OUTCOME = *repay*
- It's likely that, if we observed more clients that we would see instances for which the model is not *consistent*

Machine learning works by searching through a set of potential models to find the prediction model that best *generalizes* beyond the dataset

- Machine learning is an *ill-posed problem*.
  - If our model is consistent with noisy data, then it may not be consistent with new data with (different) noise.
  - The training set represents only a small sample of the possible set of instances in the domain.
- A predictive model that makes the correct predictions for unseen instances must capture the underlying relationship between the descriptive and target features.
  - It is said to *generalize* well.
- There are two types of *inductive bias* that a machine learning algorithm can use:
  - A *restriction bias* constrains the set of models that the algorithm will consider during the learning process.
  - A *preference bias* guides the learning algorithm to prefer certain models over others.

We must avoid *sampling bias*

- "Sampling bias arises when the sample of data used within a data-driven process is collected in such a way that the sample is not representative of the population the sample is used to represent.
- If a sample of data is not representative of a population, then inferences based on that sample will not generalize to the larger population."

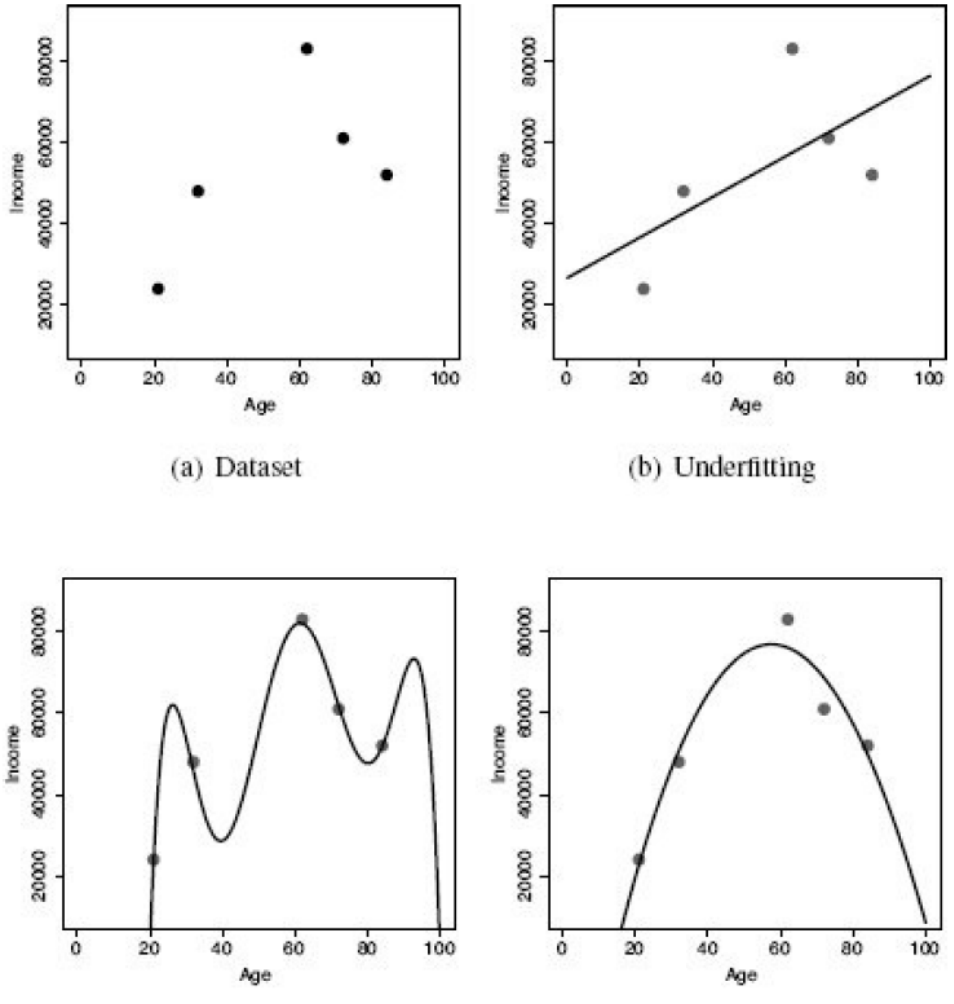
- A sample can fail to represent the population because:
  - It was collected at a different time
  - It was collected in a biased manner (phone polling, for example)
  - There is a biased source of noise
  - The sample is just too small

A famous example of sample bias – the 1948 US Presidential Election

- Nearly all of the pollsters predicted that the Republican candidate, Thomas Dewey, would defeat the Democratic candidate, Harry Truman (the sitting President)
- They based their analysis on polls conducted weeks before the election
  - They made a conscious decision that few voters would change their mind in the last part of the campaign
- Truman was elected by a margin of 49.6% to 45.1%

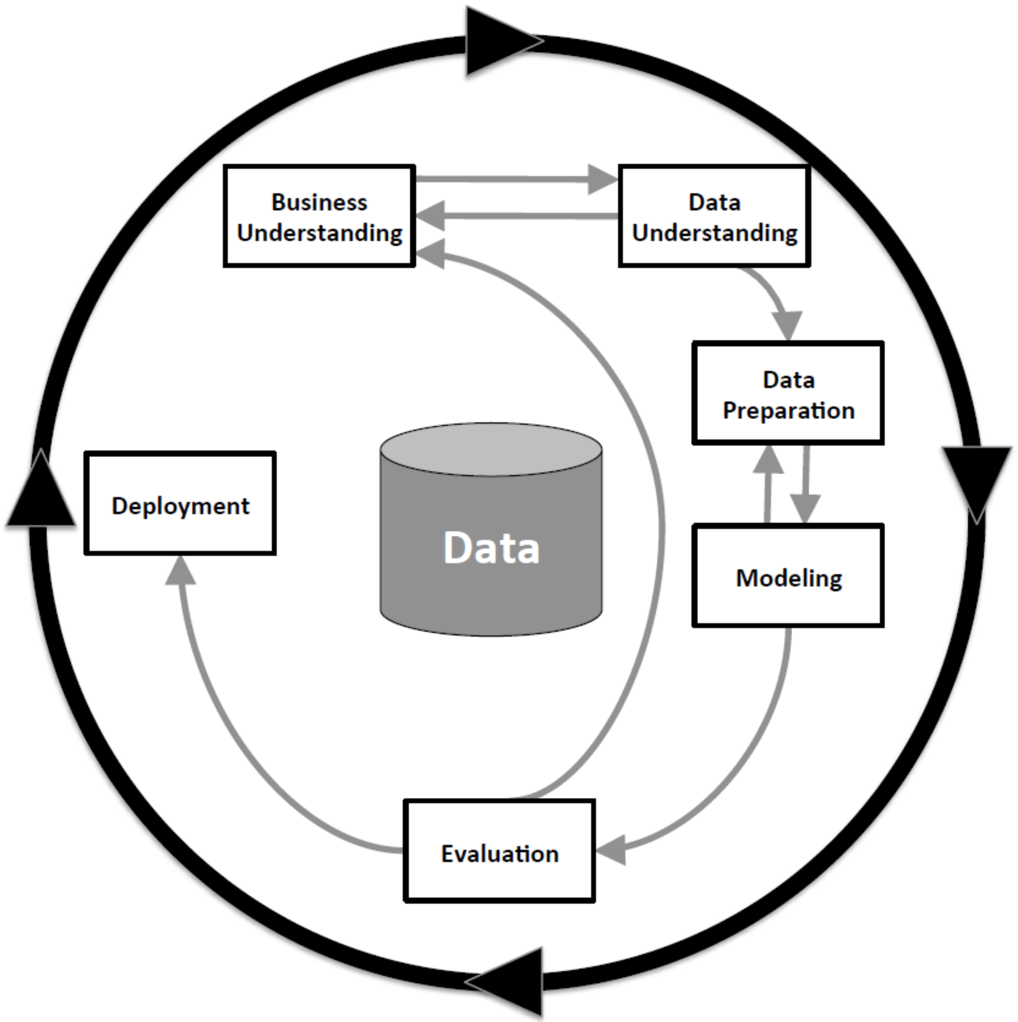
A machine learning system can experience two types of poor performance: *underfitting* and *overfitting*

- Underfitting occurs when the prediction model selected by the algorithm is too simplistic to represent the underlying relationship in the dataset between the descriptive features and the target feature.
- Overfitting, by contrast, occurs when the prediction model selected by the algorithm is so complex that the model fits to the dataset too closely and becomes sensitive to noise in the data.



(a) Dataset
(b) Underfitting
(c) Overfitting
(d) Just right

1.5 The Predictive Data Analytics Project Lifecycle: CRISP-DM



Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

1.6 Predictive data analytics tools

Analytics Application Suites

- IBM SPSS
- Knime Analytics Platform
- RapidMiner Studio
- SAS / SAS Enterprise Miner
- Weka

Program Development Environments

- R
- Python / scikit – PyTorch – tensorflow – etc.
- Java / dl4j – etc.
- C++ / various

1.7 A map of the textbook

Chapters 2 and 3 covers the Business Understanding, Data Understanding, and Data Preparation phases of the process.

The second part of the book covers the Modeling phase of CRISP-DM. We consider:

- Information-based learning (Chapter 4)
- Similarity-based learning (Chapter 5)
- Probability-based learning (Chapter 6)
- Error-based learning (Chapter 7)
- Neural networks (deep learning) (Chapter 8)

The third part of the book covers the evaluation and deployment of machine learning.

- Chapter 9 discusses the evaluation of predictive models.
- Chapters 10 and 11 preview some advanced topics – unsupervised learning and reinforcement learning.
- Chapters 12 and 13 present case studies describing specific predictive analytics projects from Business Understanding right up to Deployment.
- Chapter 14 provides some overarching perspectives on machine learning for predictive data analytics.

Today’s Objectives

Course Introduction
- Syllabus
- Course Objectives
- Your Instructor

Chapter 1 – Introduction
- 1.1 What Is Predictive Data Analytics?
- 1.2 What Is Machine Learning?
- 1.3 How Does Machine Learning Work?
- 1.4 What Can Go Wrong with Machine Learning?
- 1.5 The Predictive Data Analytics Project Lifecycle: CRISP-DM
- 1.6 Predictive Data Analytics Tools
- 1.7 The Road Ahead