Principles and Practices of Data Science

Lecture 2

Melvin Ayala

Lecture 2: Data Science Domains, Categories, and Roles

Sections:

1. Data Science Domains
    1.1. The Intersection of Science, Technology, and Data
    1.2. The Three Domains of Data Science
2. Categories behind Data Science
3. Data Science Roles

2.1. Data Science Domains

The Intersection of Science, Technology, and Data

Data, Science, and Technology are interconnected

Science

Numerical Analysis
Approximation methods, series expansion, ...

Matrix Algebra
vector, matrixes, matrix multiplication, rotation, eigenvalues, eigenvectors, ...

Statistics
descriptive and inferential

Probability Theory
random, Markov chains, Bayes theorems, ...

Technology

Business Intelligence

Data Mining

Big Data

Predictive Analytics

Machine Learning

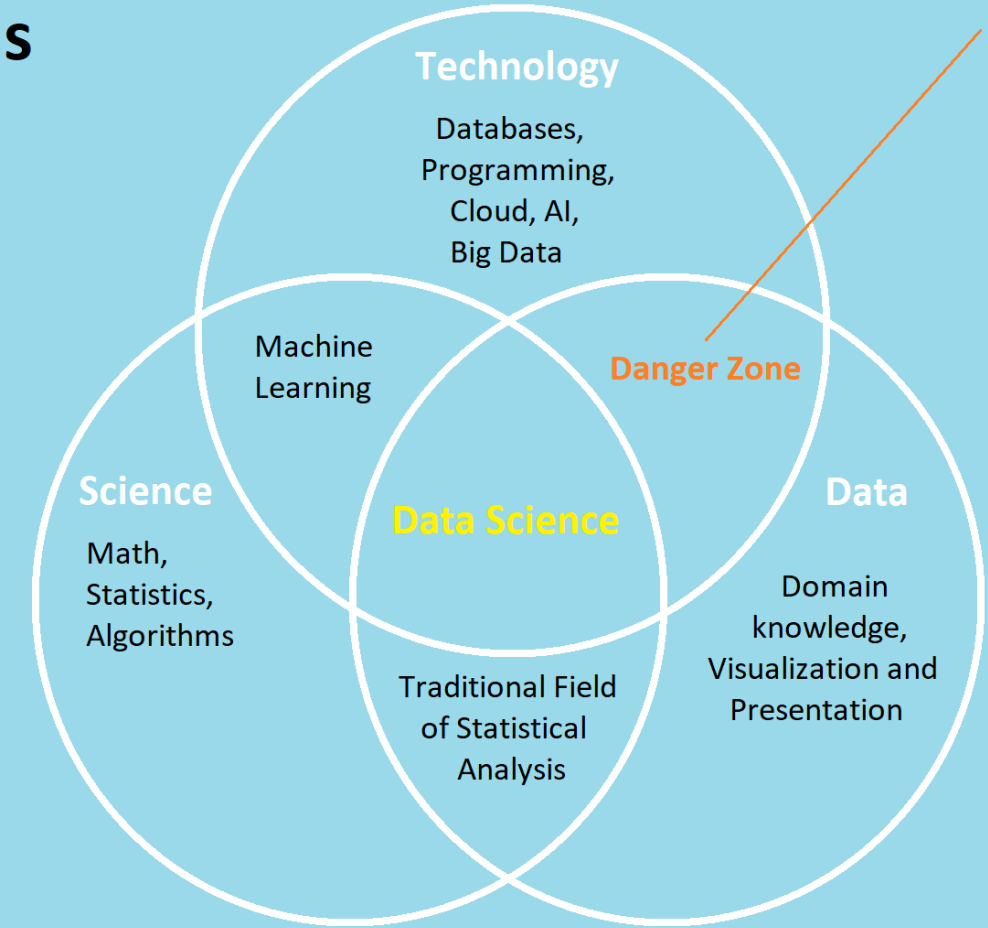
Data

Structured:
data bases, data sets, files, libraries, ...

Unstructured:
text, images, video, music, articles, books, newspapers, ...

The Three Domains of Data Science

Three Domains involved in a Data Science project



Danger Zone:

Engaging in data intensive projects with no Analytics expertise

2.2. Categories behind Data Science

Categories Behind Data Science

2. DATA SCIENCE DOMAINS

Categories behind data science

Data Science

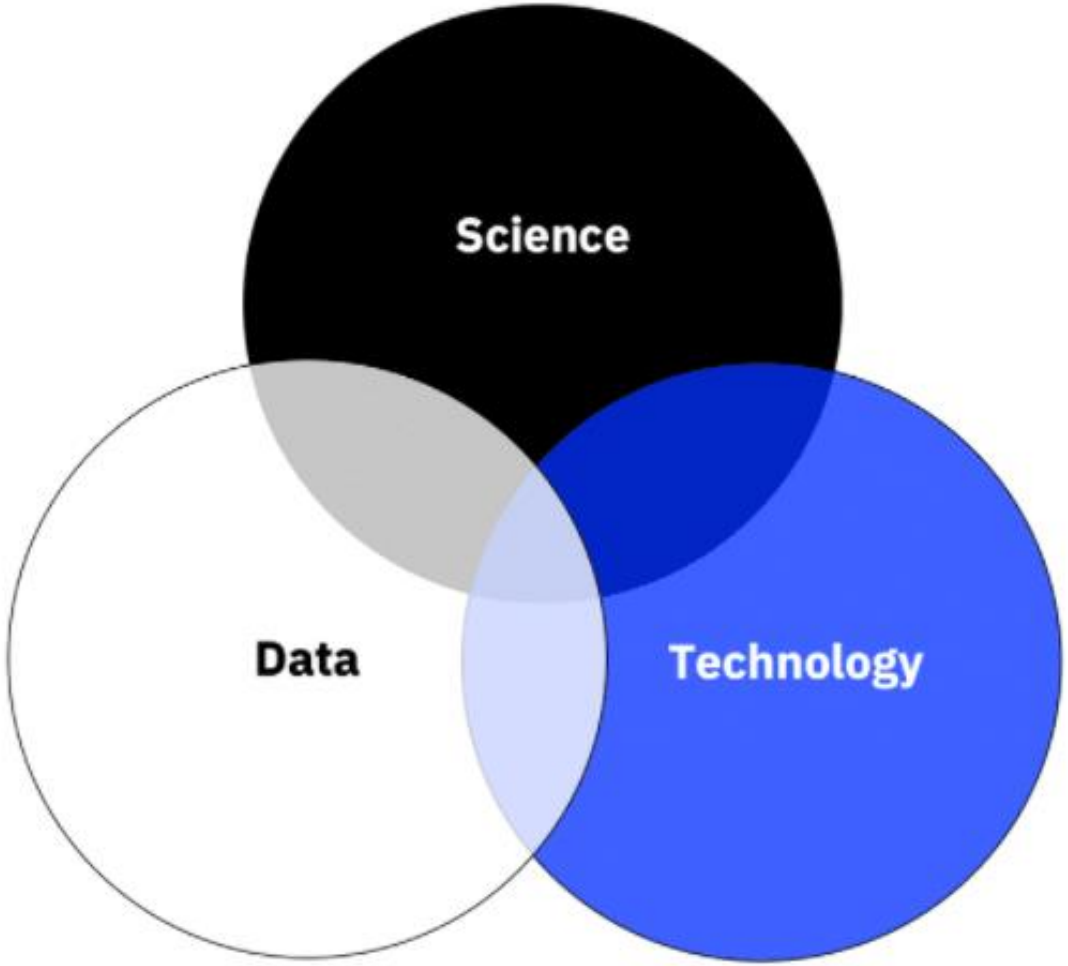
- Machine learning
- Statistical modeling
- Experiment design
- Statistics, research, mathematics

Data Journalism

- Domain expertise
- Strategic problem solving
- Business acumen
- Communication skills
- Visualization skills
- Decision making based on insights

Data Engineering

- Database and data storage
- Scripting language
- Artificial Intelligence
- Cloud Infrastructure
- Statistical computing



2.3. Data Science Roles

The Data Science Team

Data Analyst:

- gathers, cleans, and studies data sets to help solve problems.

Data Scientist:

- applies algorithms and designs experiments and Machine Learning models.

Data Engineer:

- manages the data infrastructure, servers, repositories

Who is a Data Analyst?

Data Analyst:

Roles:

- Collects, cleans, study and interprets data sets in order to answer a question or solve a problem.
- Explains and visualize data
- Works in many industries, including business, finance, criminal justice, science, medicine, and government.
- Explores and understands business domain
- Familiar with visualization tools

Characteristics:

- Great communicator
- Good presentation skills
- Critical thinking and agile design
- Familiar with visualization tools

Who is a Data Scientist?

Data Scientist:

Roles:

- Analyzes, interprets extremely large amounts of data.
- Works closely with business stakeholders to understand their goals and determine how data can be used to achieve those goals.
- Designs data modeling processes, creates algorithms and predictive models

Characteristics:

- Applies several traditional technical roles, including mathematician, scientist, statistician and computer professional
- Performs data investigation and exploratory data analysis
- Chooses potential models and algorithms
- Applies data science techniques, such as machine learning, statistical modeling, and artificial intelligence
- Measures and improves results
- Presents final results to stakeholders
- Makes adjustments based on feedback

Who is a Data Engineer?

Data Engineer:

Roles:

- Prepares data for analytical or operational uses
- Responsible for building data pipelines to bring together information from different systems
- Integrates, consolidates and cleanses data and structure it for use in analytics applications.
- Utilizes advanced programming techniques
- Makes data easily accessible and to optimize their organization's big data ecosystem.
- Manages the data infrastructure
- Tests and deploys Machine Learning models

Characteristics:

- Skilled in programming languages (C#, Java, Python, R, Ruby, Scala and SQL)
- Tech savvy
- Uses Machine Learning API calls
- Familiar with infrastructure architecture

Different Functions of the Data Science Team

Function	Skill	Role
Define the problem and build a hypothesis	Subject matter expertise	Product owner
Acquire, transform and clean data	Data Engineering	Data Engineer
Build models	Machine Learning or Decision Optimization Engineering	Data scientist: Machine Learning Engineer Decision Optimization Engineer
Communicate the results	Data journalism, web development	Data Analyst

A Data Science Story

Problem with insurance company ABC:

- experiencing a large number of fraudulent insurance claims (for example, vandalized incidents submitted as legitimate accident claims)
- having high losses due to fraud
- wants to reduce that number to a minimum.

Solution:

- reduce fraudulent insurance claims

Approach:

- identify fraudulent claims

A Data Science Story (cont'd)

- Questions to ask:
  - How to identify fraudulent claims? Not easy
  - What data is available? Need to look at the data
  - Are there actual cases of demonstrated fraud?

- Identify = predict (with a probability)

- Probability for a prediction to be correct:
  - 50%: very bad (just by chance, coin flipping)
  - 70%: better
  - > 95%: ideal

A Data Science Story (cont'd)

- What **historical** data is available?
- Spreadsheet (tabular data is the preferred way)
- Some of the columns are:
  - first and last name
  - age
  - address
  - insurance ID
  - insurance claim amount
  - insurance claim date
  - driver's license expiration date
  - Insurance policy expiration date
  - Times claims made
  - …
  - Actual fraud? (0 or 1)

- First step: Identifying the critical columns
- Are those column correlated to the last column (actual fraud)?
- Need to compute correlation

A Data Science Story (cont'd)

Performance of the prediction depends on many factors:

- amount of available data
- amount of data that represents actual legitimate claims
- amount of data that represents actual fraudulent claims
- presence of a pattern that clearly shows:
  - a pattern for the legitimate claims and
  - a different pattern for the fraudulent claim.

If two different patterns are found (overlap possible, but not too much), the problem can be represented with as logsig function.

A Data Science Story (cont'd)

How to determine which claims are legitimate and which aren't?

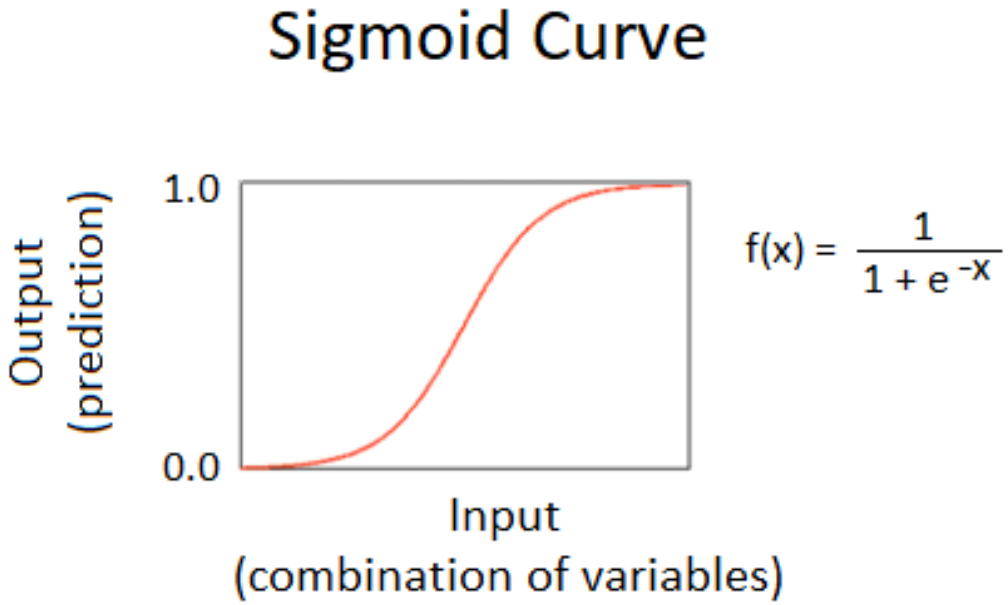
- Need to look at the data available (no data? Need to collect first)
- Need to determine the most critical variables (features)
- Most critical:
  - Claim amount
  - Days claim made before policy expiration
  - Days claim made before driver's license expiration
  - etc.

Each variable can correlate with the prediction (legitimate or fraudulent claim) in a very different way.

Sigmoid Curve

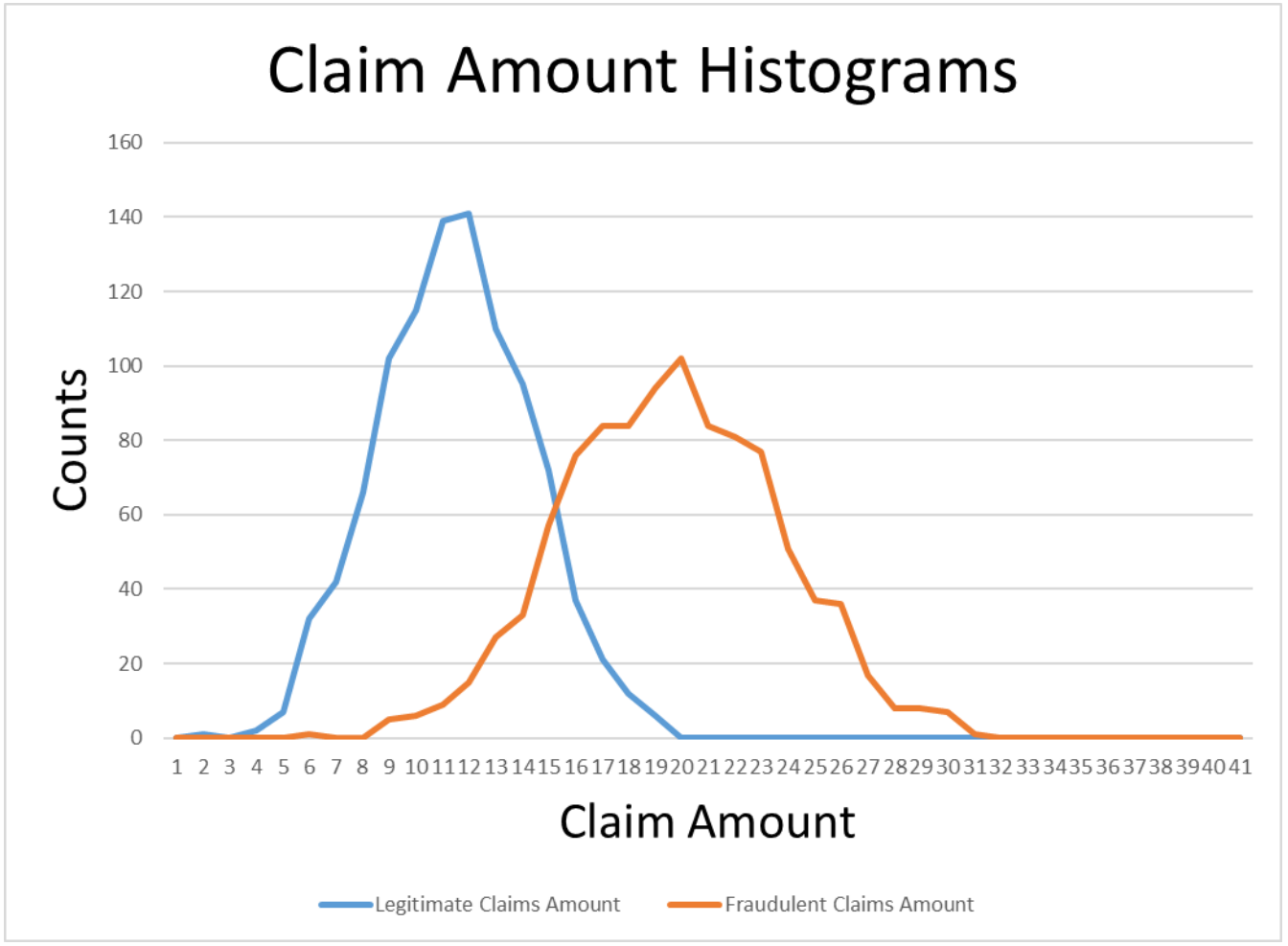
- Very convenient for binary classification
- Can represent a probability relationship between a set of inputs and a binary variable:
  - No/Yes
  - 0/1

- In a multivariable problem, represents a black box.



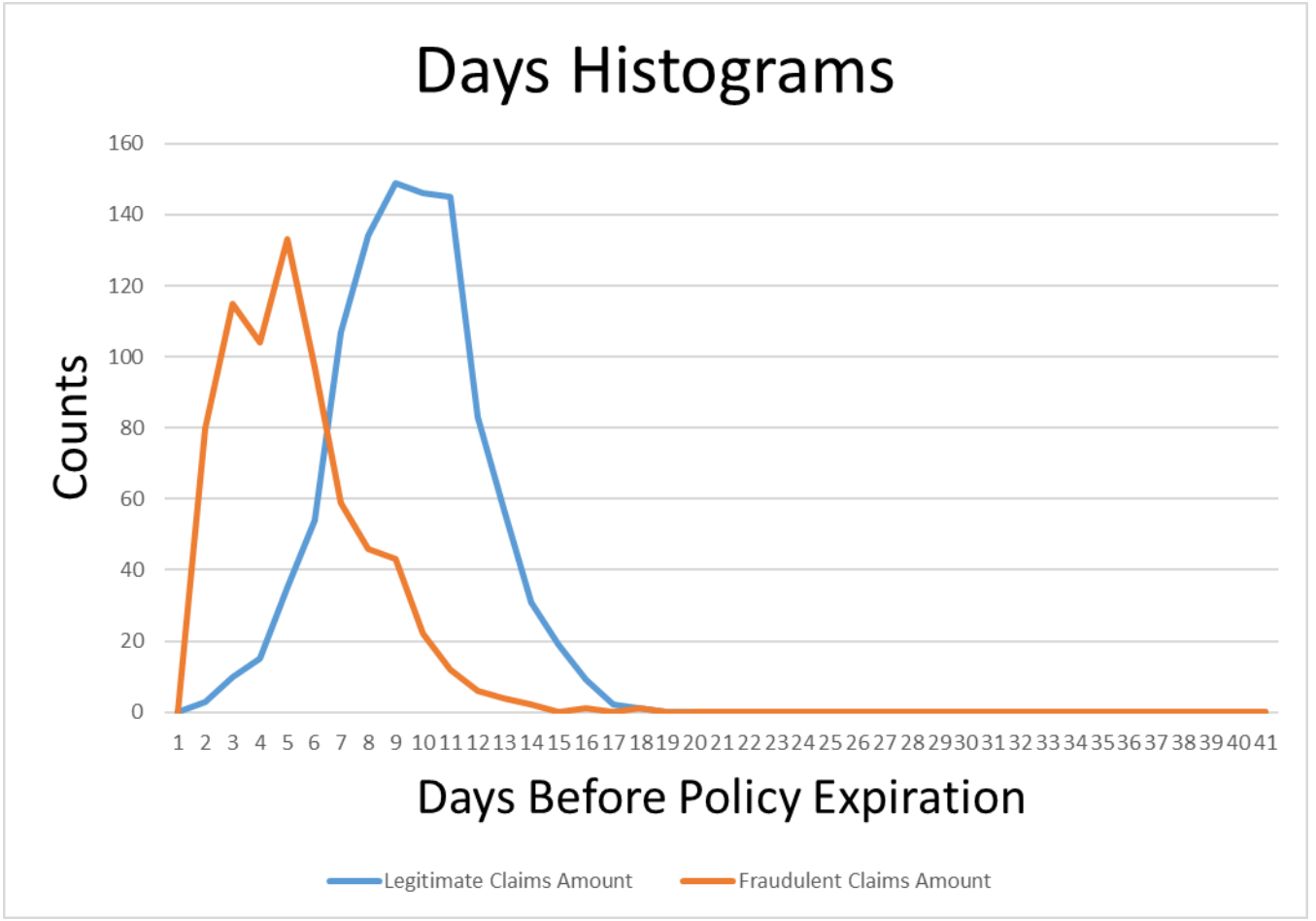
Interpreting Data to Make a Classification: Claim Amounts

Case #	Legitimate Claims Amount	Fraudulent Claims Amount
1	2558.37	3905.72
2	1985.10	4509.77
3	3158.64	3914.68
4	3683.62	4146.68
5	2415.06	6188.90
6	2074.36	4509.14
7	3603.33	4240.84
8	842.66	6586.27
9	1588.92	5187.63
10	2903.00	4814.23
11	2010.70	7120.74
12	1589.55	4946.97
13	1567.28	2989.41
14	2771.69	5376.76
15	3289.86	5067.72
16	2831.30	4517.05
17	1788.62	5766.31
18	2947.63	5112.68
19	3081.45	4946.53
20	2419.34	4901.88
21	…	…



Interpreting Data to Make a Classification: Days Claim Made before Policy Expiration

Case #	Legitimate Claims Days Before Policy Expiration	Fraudulent Claims Days Before Policy Expiration
1	22	1
2	39	26
3	63	3
4	56	2
5	51	4
6	18	43
7	37	20
8	27	34
9	28	1
10	40	7
11	47	2
12	44	5
13	48	17
14	42	4
15	39	1
16	42	4
17	41	11
18	57	36
19	12	33
20	28	56
21	…	…



End of Lecture