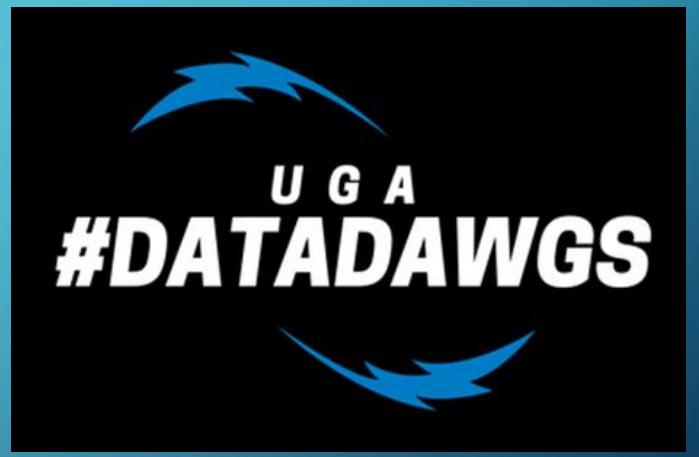
INTRODUCTION TO DATA SCIENCE AND MACHINE LEARNING

#DATADAWGS PRESENTATION BY JONATHAN WARING



PLUG TO CERTIFICATE PROGRAMS

- The Informatics certificate at UGA provides a pathway for UGA students in any major to obtain broadly marketable skills in informatics and the specific expertise to use those skills in their chosen domain.
  - Foundational Course is INFO 2000, which will be taught at TR 2:00-3:15 (CRN: 45318) during the Spring 2018 semester
- The Applied Data Science certificate at UGA develops expertise in the collection, storage, analysis, visualization, and interpretation of data. It is more focused for students in a variety of mathematical, scientific, and engineering fields.
  - One of the course courses, CSCI 3360, will be taught at TR 12:30-1:45/W 12:20-1:10 (CRN: 41607) during the Spring 2018 semester (CSCI 1301 or 1360 required as pre-req)

BUZZWORD AMBIGUITY

- The terms informatics, data analytics, machine learning, data mining, and data science are often used interchangeably without many people really knowing what they mean
- Let's start by trying to establish the differences between the fields

INFORMATICS

- The term “informatics” broadly describes the **study** and **practice** of creating, storing, finding, manipulating and sharing information.
- The field considers the interaction between humans and information alongside the construction of interfaces, organizations, technologies, and systems.
- The term was coined as a combination of "information" and "automatic" to describe the science of automating information interactions

DATA ANALYTICS

- Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain.
- Data analytics is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software
- This term is perhaps the most ambiguous one of them all

MACHINE LEARNING

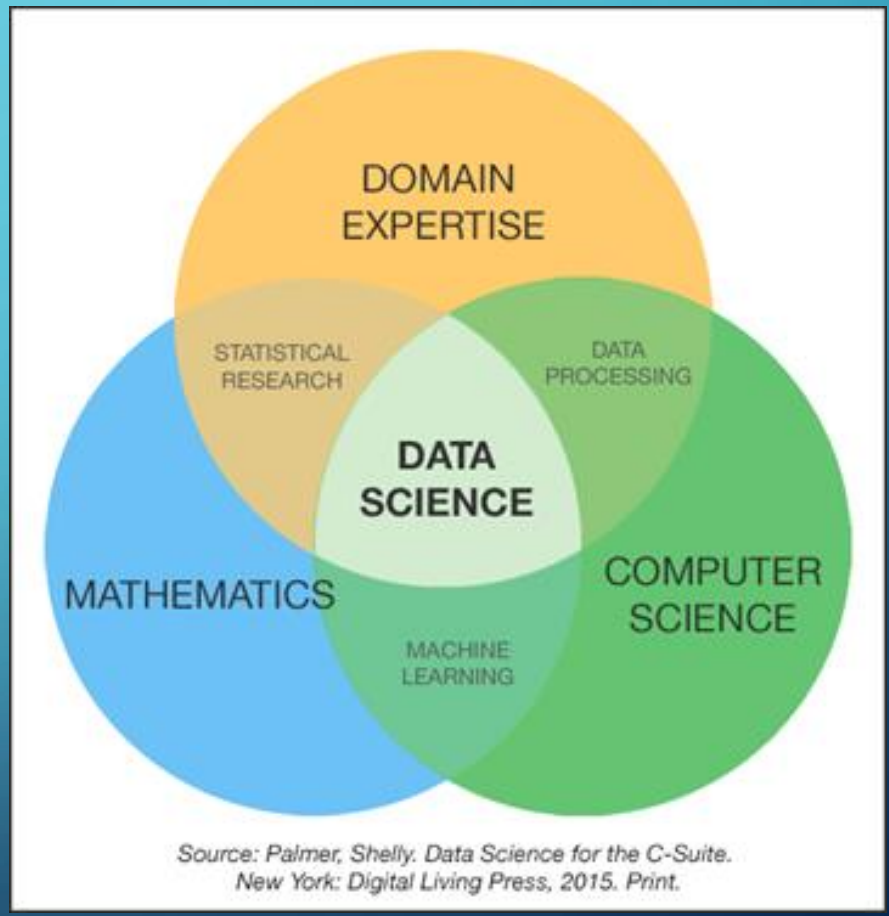
- Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed
- Machine learning is a broad term to refer to methods used to devise complex models and algorithms that lend themselves to pattern recognition
- Machine learning can be further classified into different categories depending on the nature of the data being inputted to the system **OR** by considering the desired output of the system → we'll look at these later

DATA MINING

- Finding patterns in data that provide insight or enable fast and accurate decision making
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- The term is bit of a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself

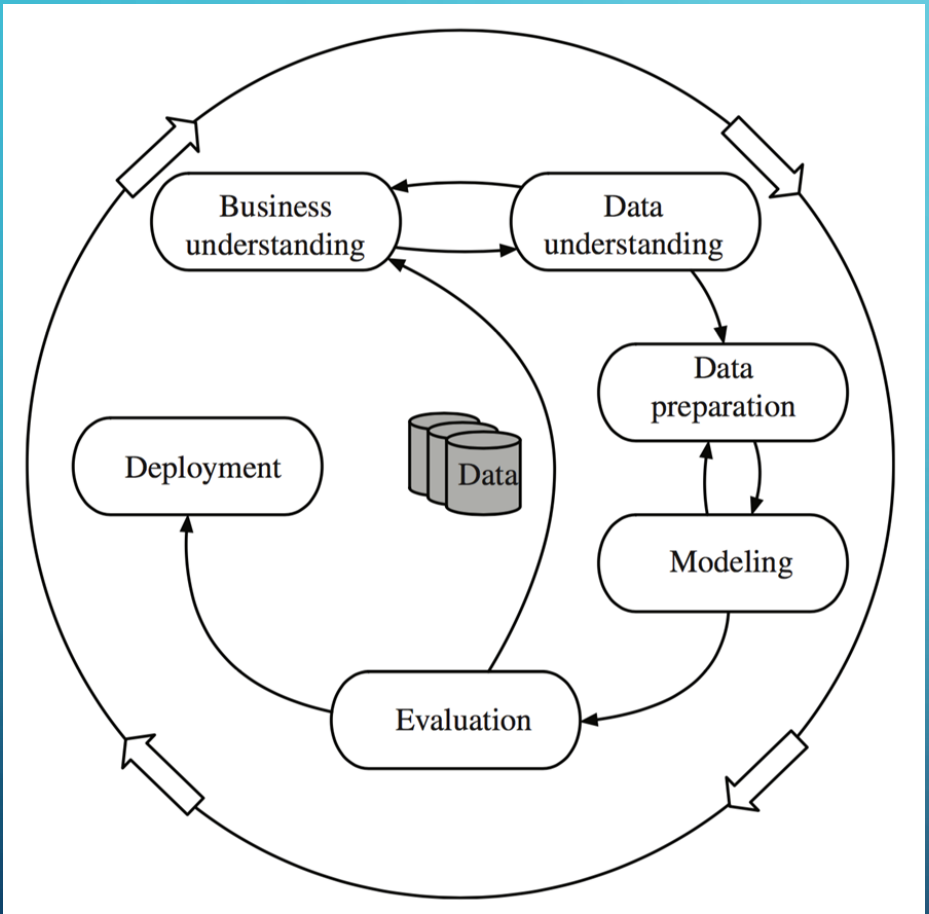
DATA SCIENCE

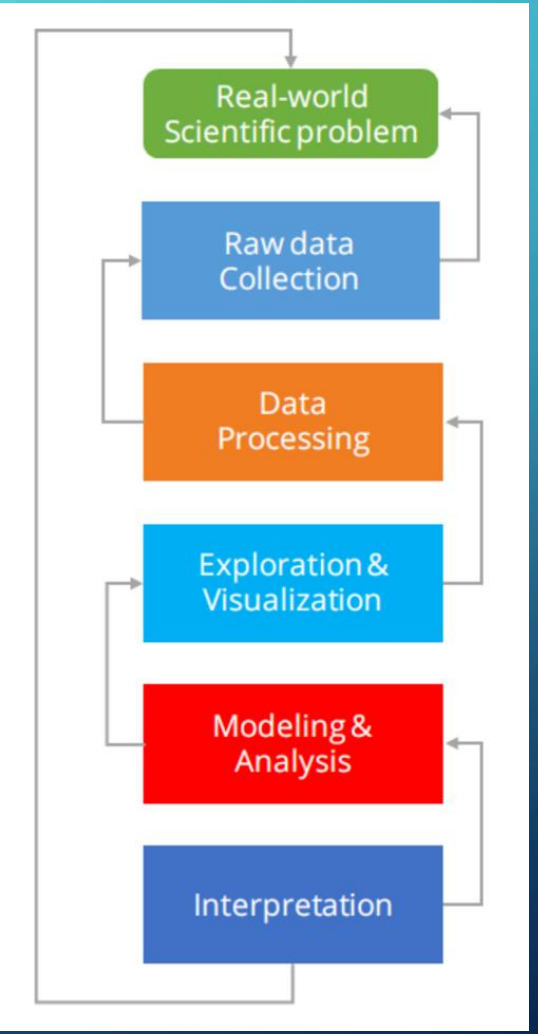
- Data science is often described as the intersection of mathematics/statistics, computer science, and domain expertise
- Data Science really encompasses the **entire problem stack**:
  - Problem definition (domain expertise)
  - Data collection & cleaning (informatics)
  - Exploration (data analytics)
  - Modeling (machine learning)
  - Interpretation & insights (data mining)



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

DATA SCIENCE PROCESSES





FROM DATA TO INFORMATION

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
- This data is a potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
  - Data: recorded facts
  - Information: patterns underlying the data
- We are concerned with machine learning techniques for automatically finding patterns in data
- Patterns that are found may be represented as *structural descriptions* or as black-box models

WHAT DOES IT MEAN FOR MACHINES TO "LEARN"?

- Definitions of learning from the dictionary:
  - To get knowledge of by study, experience, or being taught
  - To become aware by information or from observation
  - To commit to memory
  - To be informed of, ascertain; to receive instruction
- The first two definitions are hard to measure, while the last two are trivial
- We define learning as:
  - Things learn when they change behavior in such a way that makes them perform better in the future → quite a few ways to measure this (cost function, accuracy, precision, F1 score, etc.)

THREE BROAD CATEGORIES OF MACHINE LEARNING

- As mentioned earlier, machine learning can be broken down into three categories depending on the nature of the input data given to the learner:
  - **Supervised**: This means that the computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs
  - **Unsupervised**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input
  - **Reinforcement Learning:** A computer program interacts with a dynamic environment in which it must perform a certain goal. The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

WHAT INFORMATION CAN WE LEARN?

- It might surprise you, but there are only five questions that data science answers:
  - Is this A or B?
  - Is this weird?
  - How much or How many?
  - How is this organized?
  - What should I do next?
- Each one of these questions is answered by a separate family of machine learning methods, called algorithms

IS THIS A OR B? → CLASSIFICATION

- Determining whether an example is A or B is called a two-way classification problem.
- You can extend this question to ask is this A or B or C (or D, etc.), and this is known as multiclass classification.
- Classification is a **supervised** machine learning problem.
- Spam filtering is an example of classification, where the inputs are emails and the classes are "spam" and "not spam".

IS THIS WEIRD? → ANOMALY DETECTION

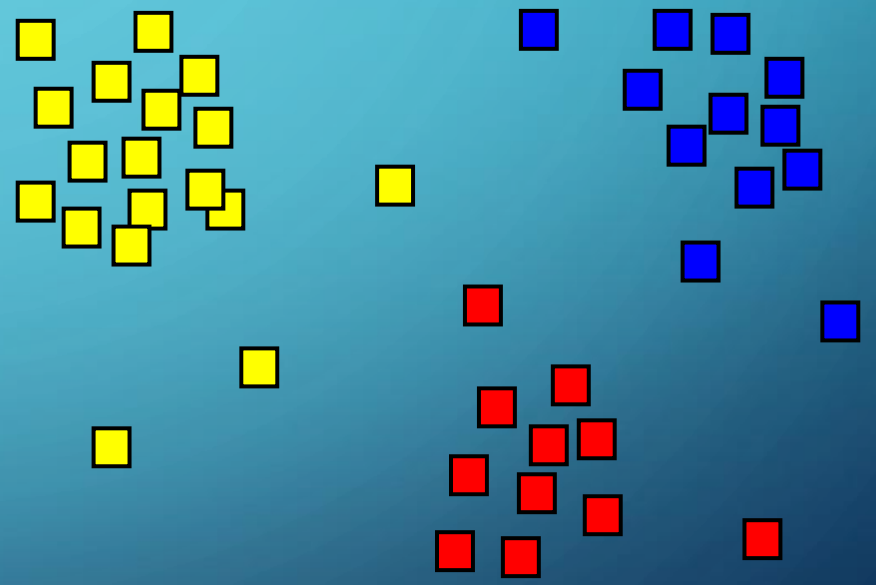
- Anomaly detection is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset
- Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions
- Anomaly detection may be **unsupervised** or **supervised**
- If you have a credit card, you've already benefited from anomaly detection.
  - Your credit card company analyzes your purchase patterns, so that they can alert you to possible fraud.

HOW MUCH OR HOW MANY? → REGRESSION

- Regression allows you to make predictions from data by learning the relationship between features of your data and some observed, numeric response
- Regression makes numerical predictions for everything from what will the temperature be next Tuesday to stock prices to understanding gene regulatory networks (really any question that asks for a number)
- Regression is a **supervised** machine learning problem
- If you have a credit card, you've already benefited from anomaly detection.
  - Your credit card company analyzes your purchase patterns, so that they can alert you to possible fraud.

HOW IS THIS ORGANIZED? → CLUSTERING

- Sometimes you want to understand the structure of a dataset, but you don't have any examples that you already know the outcome for
  - We therefore know that clustering is an **unsupervised** machine learning problem
- Clustering separates data into natural "clumps," for easier interpretation.
- With clustering, there is no one right answer.
- By understanding structure, you can better understand - and predict - behaviors and events.
  - Common example: Which viewers like the same movies?



WHAT SHOULD I DO NOW? → REINFORCEMENT LEARNING

- Reinforcement learning was inspired by how the brains of rats and humans respond to punishment and rewards. These algorithms learn from outcomes, and decide on the next action.
- Typically, reinforcement learning is a good fit for automated systems that have to make lots of small decisions without human guidance
- Example: If I am a self-driving car, should I brake or accelerate at a yellow light?
- Reinforcement learning algorithms gather data as they go, learning from trial and error.

DATA PRE-PROCESSING

- Real world data is typically:
  - **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - **Noisy**: containing errors or outliers
  - **Inconsistent**: containing discrepancies in codes or names
- Tasks in data preprocessing
  - **Data cleaning**: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
  - **Data integration**: using multiple databases, data cubes, or files.
  - **Data transformation**: normalization and aggregation.
  - **Data reduction**: reducing the volume but producing the same or similar analytical results.
  - **Data discretization**: part of data reduction, replacing numerical attributes with nominal ones.

MACHINE LEARNING ALGORITHMS

- Representation: language for patterns/models, expressive power
- Evaluation: scoring methods for deciding what is a good fit of model to data
  - Beware of overfitting → your model may be "too good" with your training data, but may not generalize well during testing
- Search: method for enumerating patterns/models
  - Optimization techniques → most commonly used optimizer is gradient descent, which iteratively searches for the minimization of some error function (will not go into detail of how this works today)

LOOKING FORWARD

- We will try to provide some basic programming and mathematical knowledge that is necessary for understanding more about how these algorithms work
- We will start to look at how to implement these techniques using the Python programming language and the scikit-learn library
- Look out for future workshops and events!

REFERENCES

- Some material was taken from Ben Manning's INFO 2000 lecture material
- Some material was taken from Dr. Rasheed's CSCI 4380 lecture material
- The 5 data science questions material came from Microsoft
- I also used Joel Grus' *Data Science From Scratch*, which is a book I highly recommend