Data Science
AI, ML, DL

INSC 702: Advanced Topics in Information Science

Shimelis Assefa

Agenda

- How to get started with Data Science
- Artificial Intelligence
- Machine Learning
- Deep Learning
- Case Studies

How to get started with Data Science

- First – the big picture
  - Asking good question
  - Getting to know your datasets
  - Coding, programming language
  - Version control software such as GitHub
  - Exploratory data analysis
  - Visualization
  - Communicate – telling the story

How to get started …

- First thing
- Business Understanding
  - What are the business needs you are trying to address?
  - What are the objectives of the Data Science project?
  - For example, if you are at a telecommunications company, that needs to retain its customers, can you build a model that predicts churn?
  - Maybe you are interested in using live data to help better predict what coupons or incentives to offer what customers at the grocery store.

How to get started …

- Data Understanding
  - What kind of data is available to you?
  - Is it stored in a relational or NoSQL database?
  - How large is your data?
  - Can it be stored and processed on your hard drive or will you need cloud services?
  - Are the any confidentiality issues or NDAs involved if you are working in partnership with a company or organization?
  - Can you find a new data set online that you could merge and increase your insights

How to get started …

- Data Preparation
  - This stage involves doing a little Exploratory Data Analysis and thinking about how your data will fit into the model that you have.
  - Is the data in data types that are compatible with the model?
  - Are there missing values or outliers?
  - Are these naturally occurring discrepancies or errors that should be corrected before fitting the data into a model?
  - Do you need to create dummy variables for categorical variables?
  - Will you need all the variables in the data set are some dependent on each other?

How to get started …

- Modeling
  - Choose a model and tune the parameters before fitting it to your training set of data.
  - Python’s scikit learn library is a good place to get model algorithms.
  - With larger data, consider using Spark ML.

How to get started …

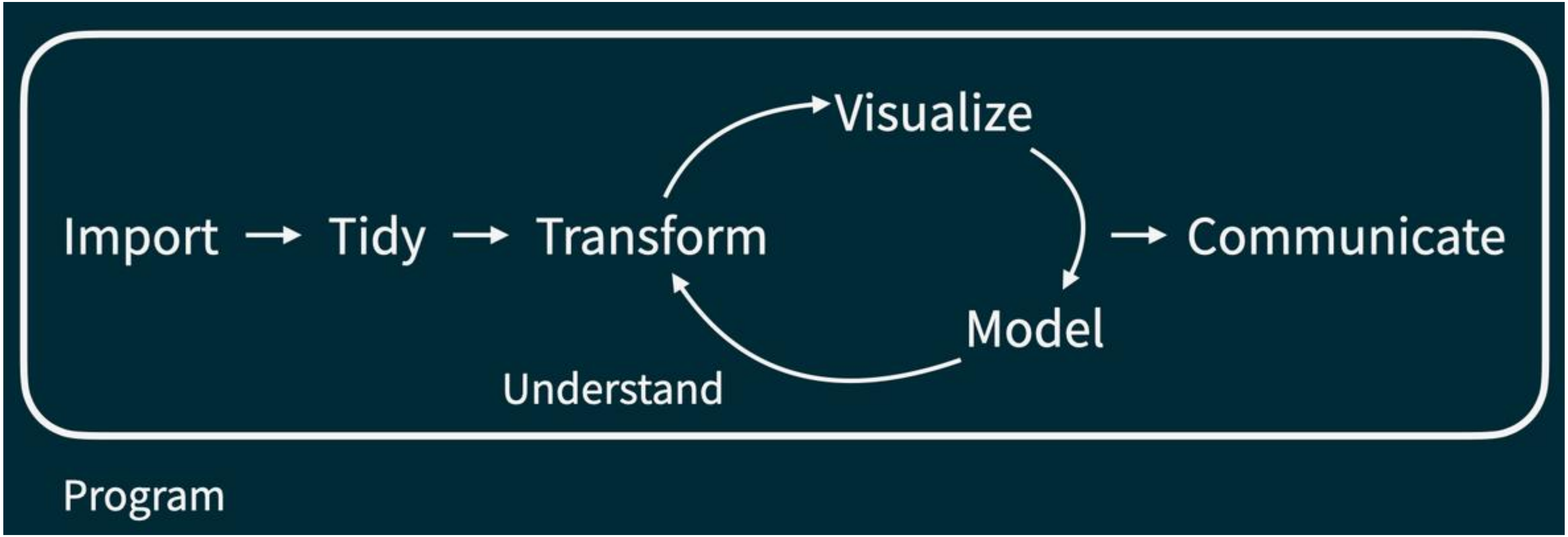
- Evaluation
  - Withhold a test set of data to evaluate the model performance.
  - Data Science Central has a great post on different metrics that can be used to measure model performance.
  - The Confusion Matrix can help with considering the cost-benefit implications of the model’s performance.

How to get started …

- Deployment/Prototyping
    - Deployment and implementation are some of the key components of any data driven project.
    - You have to get past the theory and algorithms and actually integrate your data science solution into the larger environment.
    - Flask and bootstrap are great tools to help you deploy your data science project to the world.

How to get started …

- Data Science Lifecycle – check Data Science in a Box resource - https://datasciencebox.org



Introduction…

- A few concepts first
- Classification
  - A Machine Learning task which seeks to classify data points into different groups (called targets or class labels) that are pre-determined by the training data.
  - A problem in which the output can be only one of a fixed number of output classes known apriori like Yes/No, True/False – binary or multi-class classification problem

Introduction…

- Regression
  - A supervised learning task that tries to predict a numerical result given a data point. For example, giving the description of a house (location, number of rooms, energy label) and predicting the market price of the house
- Underfitting
  - A phenomenon in which a Machine Learning algorithm is not fitted well enough to the training data, resulting in low performance on both the training data and similar but distinct data.
  - A common example of underfitting occurs when a neural network is not trained long enough or when there is not enough training data

Introduction…

- Overfitting
    - A phenomenon in which a Machine Learning algorithm is too fitted to the training data, making performance on the training data very high, but performance on similar but distinct data low due to poor generalizability.
    - A common example of overfitting occurs when a neural network is trained for too long.
- Cost function
    - This is what Machine Learning algorithms are trying to minimize to achieve the best performance.
    - It is simply the error the algorithm makes over a given dataset. It is also sometimes referred to as “loss function.”

Introduction…

- Loss function
  - A (generally continuous) value that is a computation-friendly proxy for the performance metric. It measures the error between values predicted by the model and the true values we want the model to predict.
  - During training, this value is minimized.
- Validation data
  - A subset of data that a model is not trained on but is used during training to verify that the model performs well on distinct data.
  - Validation data is used for hyperparameter tuning in order to avoid overfitting.

Introduction…

- Neural Network
    - A specific type of Machine Learning algorithm which can be represented graphically as a network, inspired by the way that biological brains work.
    - The network represents many simple mathematical operations (addition, multiplication, etc.) that are combined to produce a complex operation that may perform a complicated task

Introduction…

- Parameter
    - Refers to the numbers in a neural network or Machine Learning algorithm that are changed to alter how the model behaves (sometimes also called weights).
    - If a neural network is analogous to a radio, providing the base structure of a system, then parameters are analogous to the knobs on the radio, which are tuned to achieve a specific behavior (like tuning in to a specific frequency).
    - Parameters are not set by the creator of the model, rather, the values are determined by the training process automatically

Introduction…

- Hyperparameter
    - A value that takes part in defining the overall structure of a model or behavior of an algorithm.
    - Hyperparameters are not altered by the model training process and are set ahead of time before training.
    - Many potential values for hyperparameters are generally tested to find those that optimize the training process. E.g, in a neural network, the number of layers is a hyperparameter (not altered by training), whereas the values within the layers (“weights”) themselves are parameters (altered by training).
    - If the model is a radio, then a hyperparameter would be the number of knobs on the radio, while the values of these knobs would be parameters.

Introduction…

- Generative Models
  - Generative Models are a subset of AI Models that can be used to generate data that is similar to a set of training data.
  - For example, if a well-performing generative model is trained on a dataset of human faces, then it can be used to generate entirely novel images of new human faces.
  - Generative Models have become popular in recent years and include DALLE-2, Imagen, Stable Diffusion, and Poisson Flow Generative Models (PFGMs)

Introduction…

- Reinforcement Learning
  - Reinforcement Learning (RL) is an area of Machine Learning inspired by real-world environments where learning happens by rewarding desired behaviors and punishing undesired ones.
  - It first gained popularity when computers learned to play different video games through RL approaches, but it’s now a powerful technique that’s widely used in many Deep Learning fields.

Introduction…

- Large Language Models
  - Language Models are mathematical models that give the probability of a certain sequence of words occurring in a given language.
  - Large Language Models are very big and trained on huge amounts of data, therefore encapsulating significant amounts of information about a language.
  - Large language models are widely used in applications relevant to natural language, powering things like text-prediction.

Introduction…

- Transformers
  - Transformers are a type of Deep Learning model that are widely used in state-of-the-art applications.
  - Their main benefit is that they are able to distribute their computations across many machines, making the complicated computations required for e.g. automatic speech recognition feasible to do in a reasonable amount of time.

Introduction…

- Transfer learning
  - Transfer learning is a Machine Learning technique where a trained model for one task is reused as the starting point for a new model on a different task.
  - The new model can then utilize the already existing knowledge and often needs less training.
  - It is a powerful technique that is widely used and often yields good results.

Introduction…

- Accelerators
  - Accelerators are specialized computer hardware that are useful for AI and can accelerate the computation speed of models.
  - Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) are different types of accelerators.

Introduction…

- GPUs
  - A graphics processing unit (GPU) is a computer chip designed to perform rapid mathematical calculations.
  - Traditionally, GPUs are responsible for rendering graphics and images, although today, they have a wider use range.
  - With the emergence of Deep Learning, the importance of GPUs has increased.
  - Training of deep neural networks can be more than 100 times faster with GPUs than with CPUs (Central Processing Units).

Introduction …

- To get started, finally, prepare on the following:
    - Programming - R, Python, SQL, Git
    - Math Fundamentals – statistics, linear algebra, differential calculus
    - Data analysis – feature engineering, data wrangling, EDA
    - Web scraping – Beautiful SOAP, Scroppy, URLLIB
    - Visualization – ggplot2, D3.js, Tableau, Bokeh, Plotly
    - Machine Learning – classification, regression, reinforcement learning, deep learning, clustering
    - Deployment – AWS, Microsoft Azure, Google Cloud Platform

Artificial Intelligence

- Artificial Intelligence is comprised of computing systems that are able to perform human-like processes to achieve specific goals and tasks by using data to learn, adapt, synthesize, and self-correct.
- AI includes:
  - Analytics
  - Big data
  - Data visualization
  - Decision logic
  - Machine learning
  - Natural language processing

ARTIFICIAL INTELLIGENCE: 7 QUESTIONS (AND ANSWERS)

1

WHAT IS ARTIFICIAL INTELLIGENCE?

Artificial intelligence (AI) refers to the leveraging of multiple technologies that together create a device or construct that accomplishes certain tasks formerly requiring human input. In higher education, the principles of AI underlie a range of innovative systems, including analytics, robot writers, virtual experiences, and intelligent tutoring systems.

2

HOW DOES IT WORK?

To exhibit intelligence, computers apply algorithms to find patterns in large amounts of data—a process called machine learning, which plays a key role in a number of AI applications. AI systems often incorporate human feedback to help calibrate the system's learning.

3

WHO'S DOING IT?

Many colleges and universities are developing AI projects that aid teaching and learning, such as the Pennsylvania State University, Georgia Tech, MIT, and Harvard.

4

WHY IS IT SIGNIFICANT?

AI opens the possibility of individual tutoring to students who could never otherwise have access to it. AI learning agents have the potential to function like adaptive learning but at a much more sophisticated and nuanced level. AI allows faculty and students to do their work more effectively by providing not just tutors but AI assistants for scheduling, interactive immersive simulations, and human-machine partnerships.

5

WHAT ARE THE DOWNSIDES?

Considerable misunderstanding exists about what AI can and cannot do, resulting in inflated expectations and a risk that users could assign inappropriate kinds and amounts of authority to AI systems. For AI developers, one key issue is an emerging lack of transparency among corporate entities that see their AI programming and algorithmic development as intellectual property.

6

WHERE IS IT GOING?

AI will trend toward devices and constructs that conform more closely to human behavior and systems that are better able to handle conflicting or false information. The use of AI systems and devices will expand further into routine activities. Users may come to understand AI as a system that enhances human capabilities in a partnership between humans and machines, leveraging what each does best.

7

WHAT ARE THE IMPLICATIONS FOR TEACHING AND LEARNING?

AI bots can respond to student questions when access to the instructor and teaching assistants is limited or unavailable. AI has the potential to give every student a computer-simulated personal mentor and provide better communication between classrooms worldwide by offering translation services and cultural context. Lectures may be accompanied or augmented by immersive virtual reality environments populated by AI personalities that offer safe opportunities to practice emerging skills.

This infographic was adapted from *7 Things You Should Know About Artificial Intelligence in Teaching and Learning* (Louisville, CO: EDUCAUSE Learning Initiative, April 2017).

Machine Learning

- Learning - as a generic process is about acquiring new, or modifying existing, behaviors, values, knowledge, skills, or preferences
- Machines rely on data
- Humans: learn from experience
- (ML) is a category of artificial intelligence that enables computers to think and learn on their own.
- The term was coined by Arthur Samuel in 1959, who defined ML as a field of study that provides learning capability to computers without being explicitly programmed

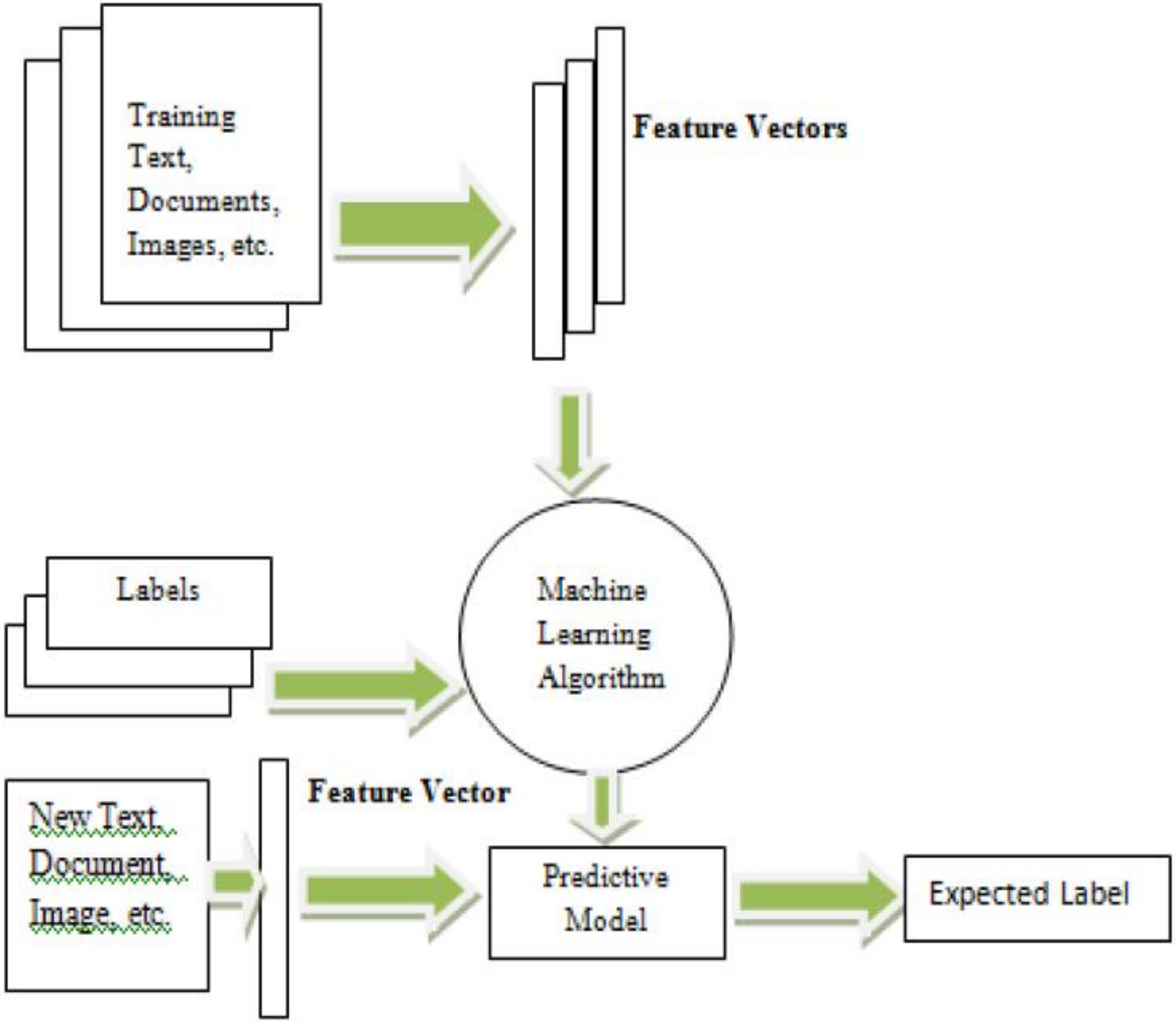
Machine Learning …

- ML is used to solve various problems that require learning on the part of the machine. A learning problem has three features:
    - Task classes (The task to be learnt)
    - Performance measure to be improved
    - The process of gaining experience
- For example, in a game of checkers, the learning problem can be defined as:
    - Task T: Playing the game
    - Performance Measure P: number of games won against the opponent.
    - Experience E: practicing via playing games against itself and consistently improving the performance

Machine Learning …

- Supervised Learning
    - a set of examples or training modules are provided with the correct outputs and on the basis of these training sets, the algorithm learns to respond more accurately by comparing its output with those that are given as input.
    - Supervised learning is also known as learning via examples or learning from exemplars.
    - Supervised learning finds applications in prediction based on historical data.

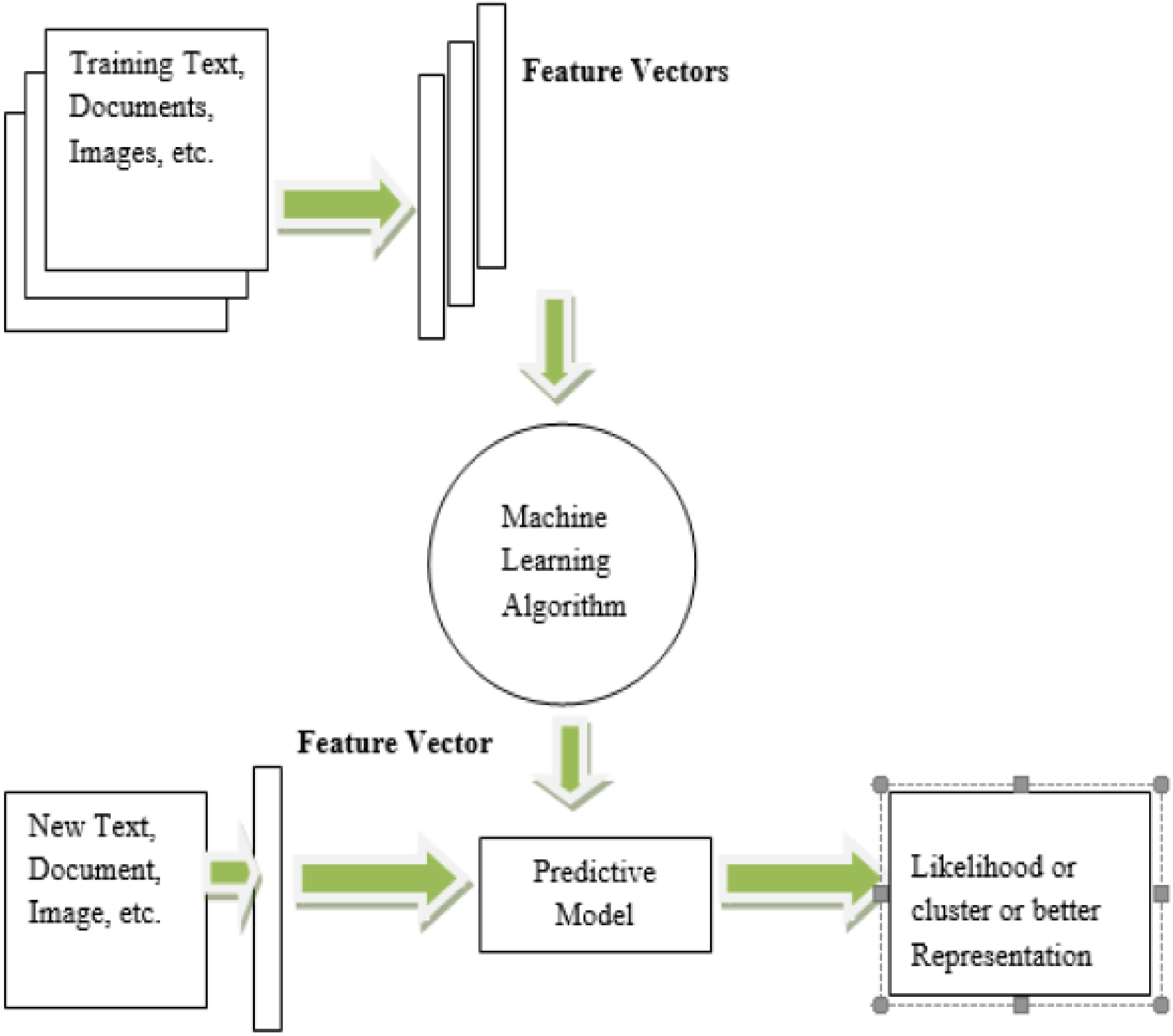
Machine Learning …


Training Text, Documents, Images, etc.
Feature Vectors
Labels
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Machine Learning …

- Unsupervised Learning
  - The unsupervised learning approach is all about recognizing unidentified existing patterns from the data in order to derive rules from them.
  - This technique is appropriate in a situation when the categories of data are unknown.
  - Here, the training data is not labeled.
  - Unsupervised learning is regarded as a statistic based approach for learning and thus refers to the problem of finding hidden structure in unlabeled data.

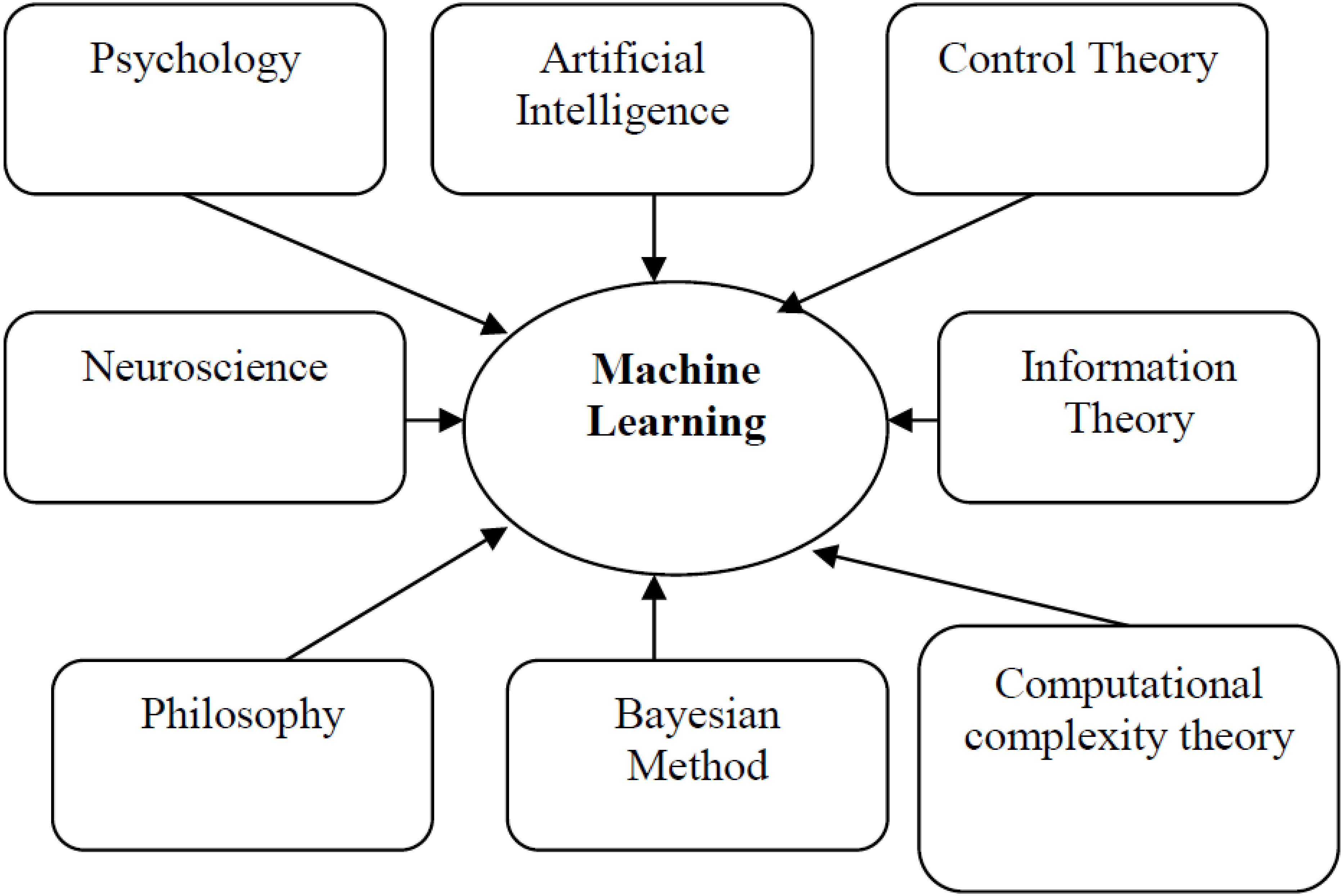
Machine Learning …



Machine Learning …

- Tom Mitchell gave a “well-posed” definition that has proven more useful to engineering set-up: “A computer program is said to learn from experience ***E*** with respect to some task ***T*** and some performance measure ***P,*** if its performance on ***T***, as measured by ***P***, improves with experience *E*
- Machine Learning Applications
  - Email spam filtering,
  - fraud detection on social network,
  - online stock trading,
  - face & shape detection,
  - medical diagnosis,
  - traffic prediction,
  - character recognition
  - product recommendation
  - self-driving Google cars,
  - Netflix showcasing the movies and shows a person might like
  - online recommendation engines—like friend suggestions on Facebook, “more items to consider” and “get yourself a little something” on Amazon
  - credit card fraud detection,

Machine Learning …



Machine Learning…

- Machine Learning starter pack
  - Google Colab: Code editor
  - Pandas: Importing and manipulating data
  - Numpy: For performing linear algebraic functions
  - Scikit learn: Make machine learning models
  - TensorFlow/ PyTorch: Making deep learning models
  - Matplotlib: Visualizing data

Deep Learning

- The term “Deep Learning” was coined in 2006 by Geoffery Hinton which referred to a new architecture of neural networks that used multiple layers of neurons for learning
- 2011, IBM’s Watson, built to answer questions posed in a natural language, defeats a Human Competitor at Jeopardy Game
- 2012, Jeff Dean from Google, developed GoogleBrain, which is a Deep Neural Network to detect patterns in Videos and Images

Deep Learning

- 2014, Facebook invented the “DeepFace” algorithm based on Deep Neural Networks capable of recognizing human faces in photos
- 2016, Google proposed DeepMind which is regarded as the most complex Board Game. Google AlphaGo program becomes the first Computer Go program to beat a professional human player
- 2017, Nvidia proposed NVIDIA GPUs- The Engine of Deep Learning.

Case Studies

- Talk to Books - https://books.google.com/talktobooks/
- Excel Formula Bot https://excelformulabot.com/
- Remove unwanted things from photo - https://magicstudio.com/magiceraser
- Generative Adversarial Network https://thispersondoesnotexist.com/
  - Access codes and train your own image - https://github.com/NVlabs/stylegan2
- Create Videos from tex https://www.synthesia.io/

Case Studies

- Papers with code https://paperswithcode.com/
  - Image classification
  - Object detection
  - Image generation
  - Language modelling
  - Question answering
  - Machine translation
  - Drug discovery
  - Medical diagnosis

Case Studies

- AI-Powered writing assistants
  - https://writesonic.com/
  - https://tribescaler.com/
  - https://tweethunter.io/
  - https://twemex.app/
  - https://answerthepublic.com/
  - https://buzzsumo.com/
  - https://mailbrew.com/