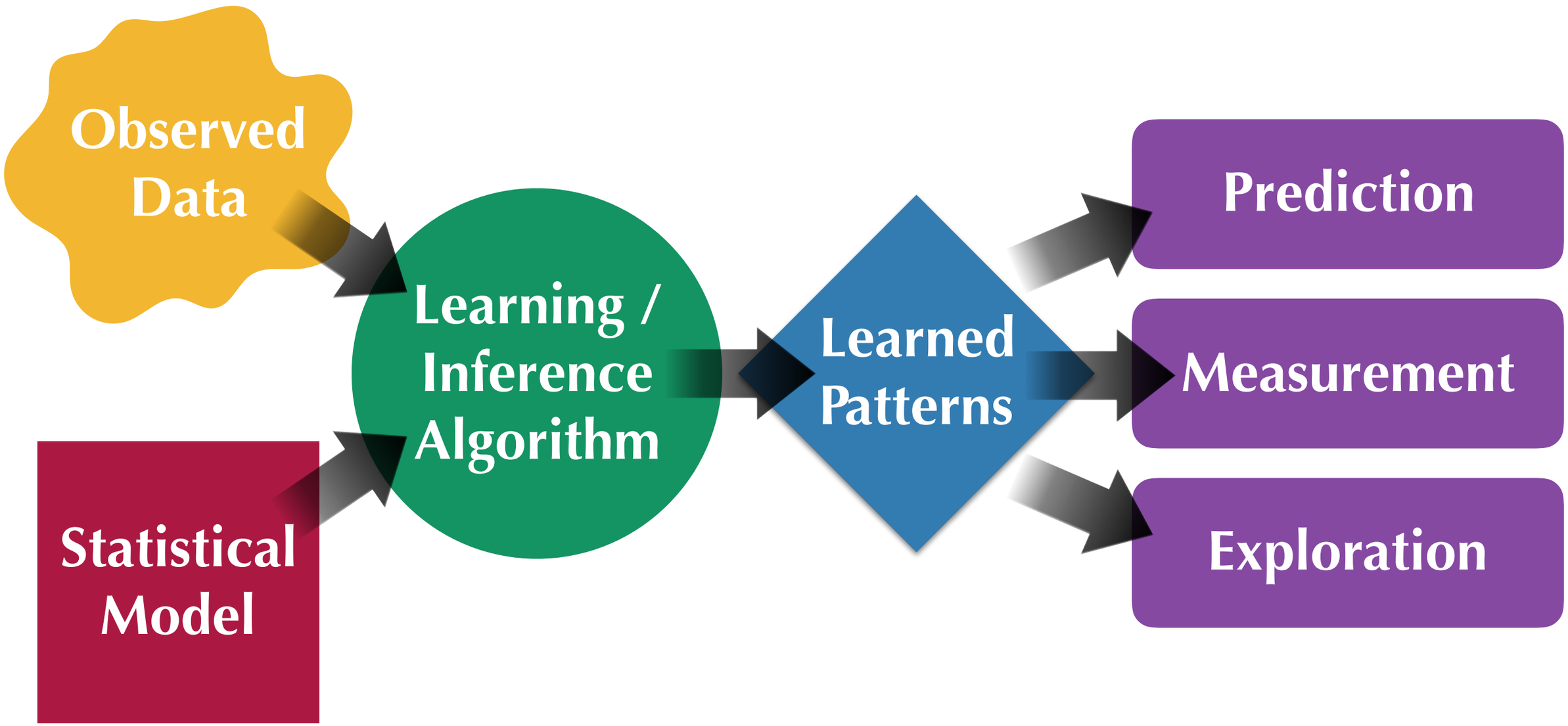
Introduction to **Machine Learning Methods**: What you Need to Know to **Conduct** and **Interpret** Research with ML

Allison J.B. Chaney
ajbc.io/MLintro

DUKE FUQUA SCHOOL OF BUSINESS

What is Machine Learning?



ajbc.io/MLintro

How do I want to use ML?

Prediction

Does person *A* belong to segment *B*?

What will revenue be if we change *X*?

Measurement

What products are perceived as *Y*?

How many people care about idea *M*?

Exploration

How many communities in network *N*?

What themes exist in reviews for *P*?

Outline

Part 1: Overview of Machine Learning

- Survey of model types
- Algorithms
- Software

Outline

Part 2: Case Studies (Methods and Challenges)

- **K-means Clustering** and **Choosing K**
- **Topic Models** and **Data Processing & Exchangeability**
- **Matrix Factorization** and **Evaluation Metrics**
- **Decision Trees & Ensemble Methods** and **Overfitting & Model Selection**
- **Deep Learning** and **Learning Rates**

Part 1:
Overview of Machine Learning

Types of ML Models

Supervised vs. Unsupervised

Types of ML Models

Supervised vs. Unsupervised

Types of ML Models

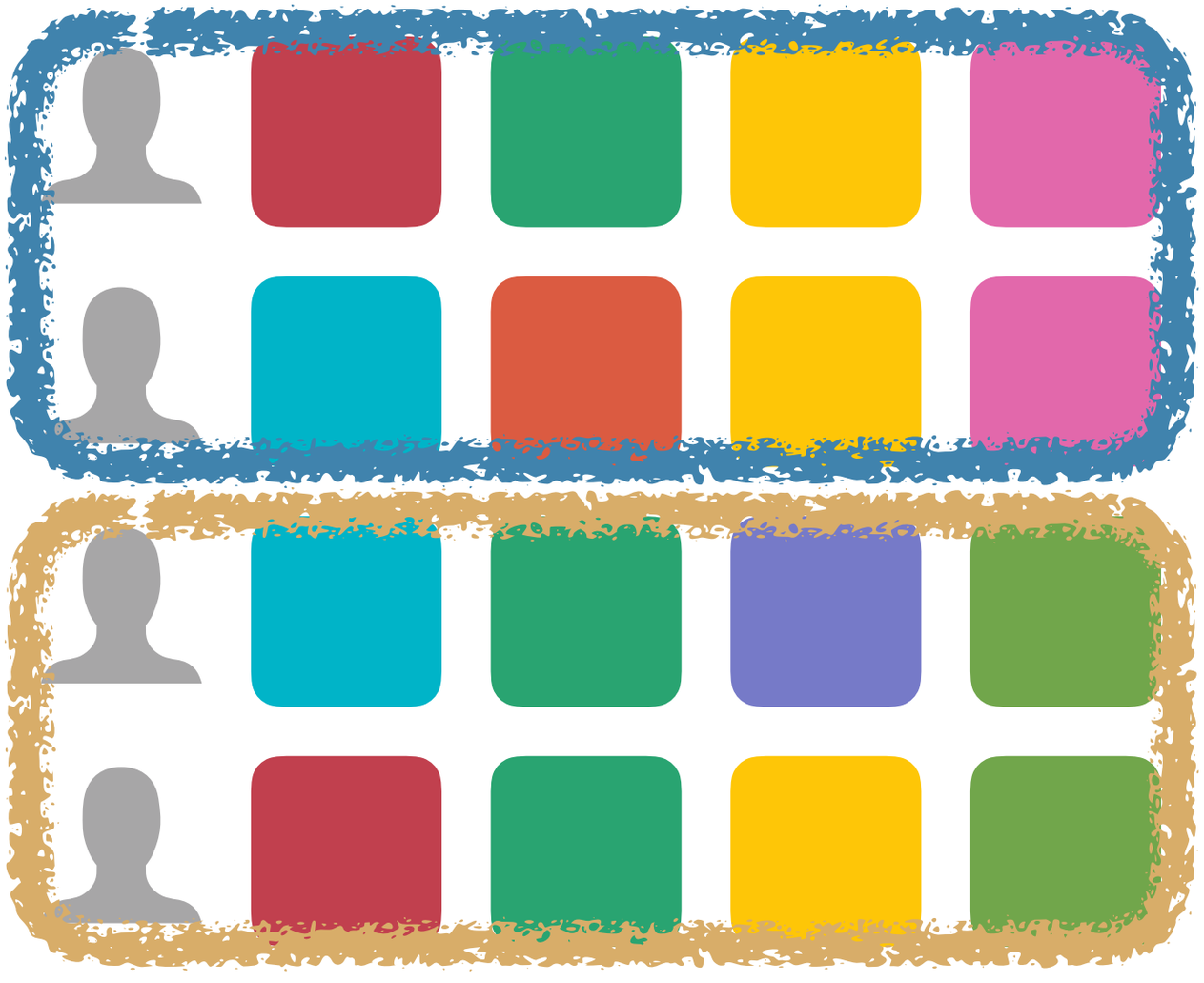
Supervised vs. Unsupervised

Types of ML Models

Supervised vs. Unsupervised

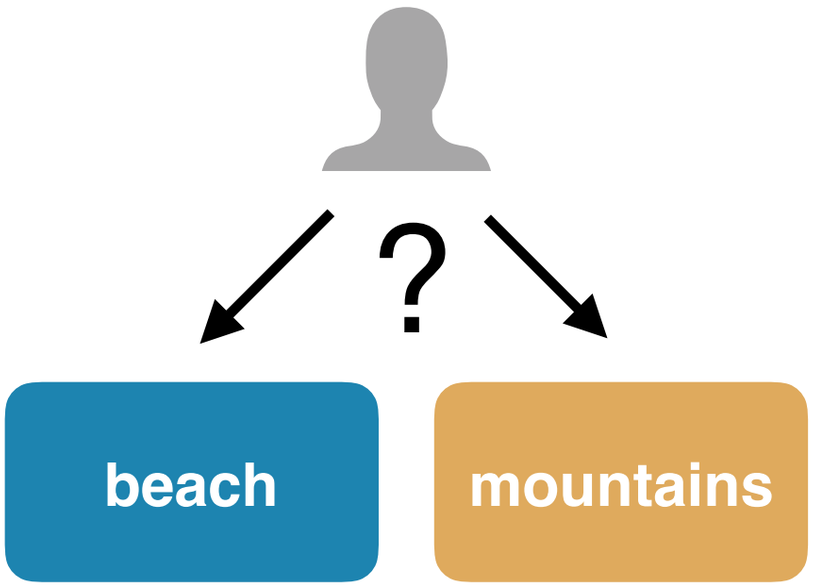
Types of ML Models

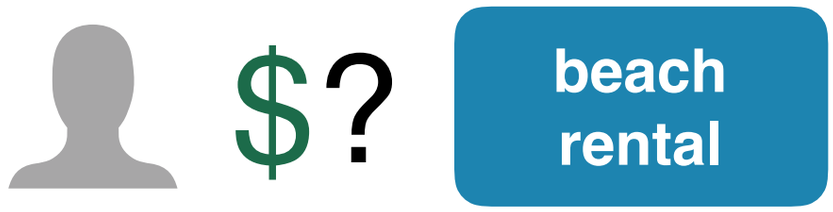
Supervised vs. **Unsupervised**



Types of ML Models

Discrete vs. Continuous





Types of ML Models

One Useful Grouping

	discrete	continuous
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Algorithms

	discrete	continuous
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

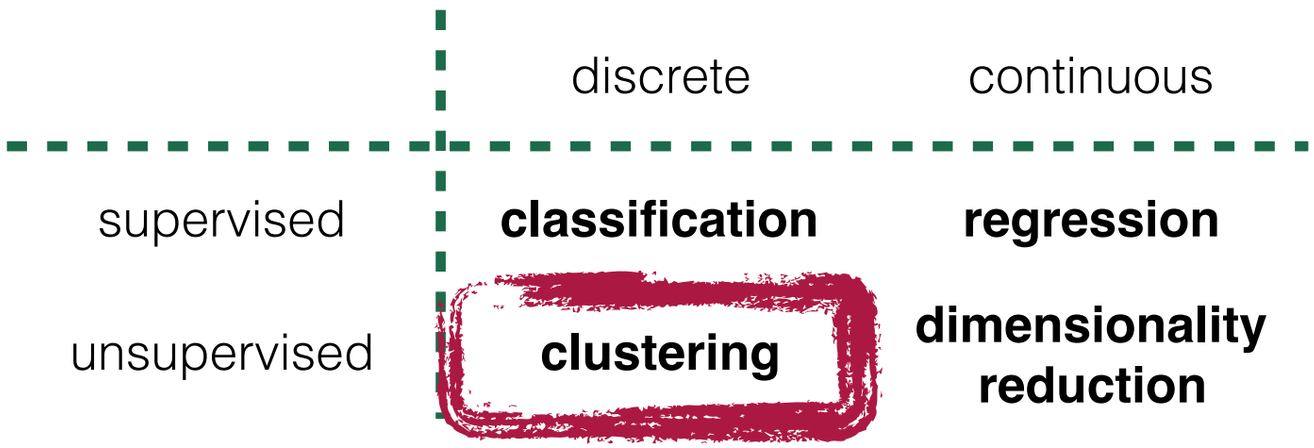
- Neural Networks / Deep Learning
- Decision Trees / Random Forests
- Boosting (ensemble method)
- Support Vector Machines
- ...

Algorithms

	discrete	continuous
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

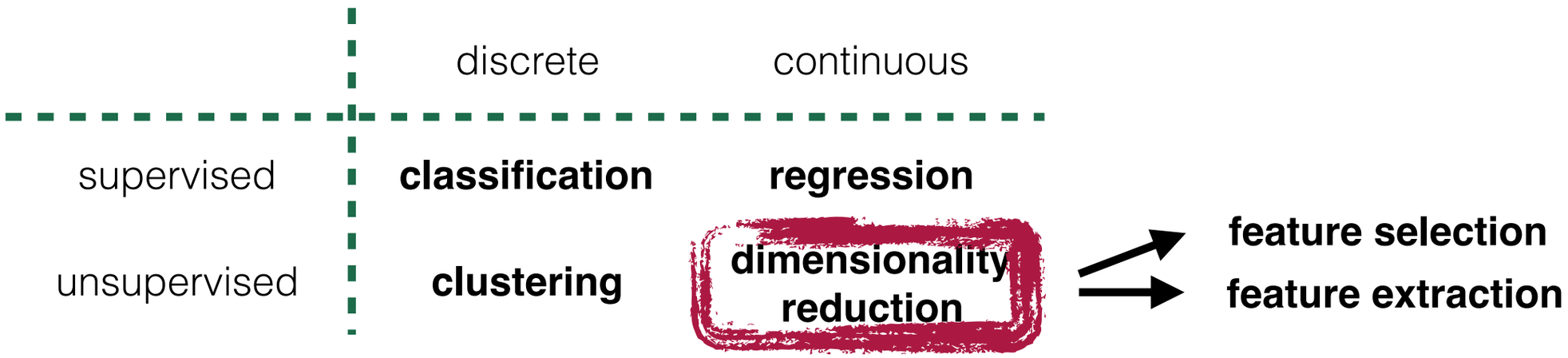
- Least squares
    - Regularization: Ridge, LASSO, ElasticNet
- Neural Networks & Support Vector Machines (again!)
- …

Algorithms



- *k*-means & fuzzy *k*-means (centroid-based)
- Expectation–Maximization (EM) using Gaussian Mixture Models (GMM) (distribution-based)
- DBSCAN (density-based)
- ...

Algorithms

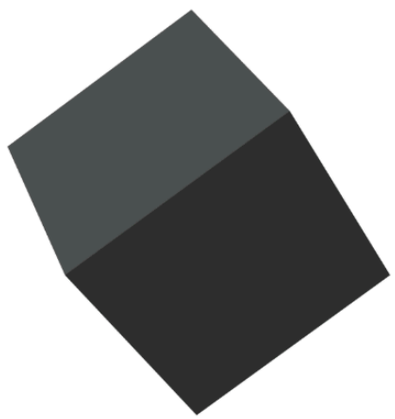


- Principal component analysis (PCA)
- Non-negative matrix factorization (NMF)
- Linear discriminant analysis (LDA)
- Autoencoder (neural network variant)
- ...

Software

VOWPAL WABBIT

Edward



https://scikit-learn.org/stable/
https://www.tensorflow.org/
http://edwardlib.org/
http://hunch.net/~vw/
https://cran.r-project.org/web/views/MachineLearning.html
...

Part 2:
Case Studies

Outline

Part 2: Case Studies (Methods and Challenges)

- **K-means Clustering** and **Choosing K**
- **Topic Models** and **Data Processing & Exchangeability**
- **Matrix Factorization** and **Evaluation Metrics**
- **Decision Trees & Ensemble Methods** and **Overfitting & Model Selection**
- **Deep Learning** and **Learning Rates**

Outline

Part 2: Case Studies (Methods and Challenges)

- **K-means Clustering** and **Choosing K**
- **Topic Models** and **Data Processing & Exchangeability**
- **Matrix Factorization** and **Evaluation Metrics**
- **Decision Trees & Ensemble Methods** and **Overfitting & Model Selection**
- **Deep Learning** and **Learning Rates**

Let’s head over to a Jupyter notebook…

Take Care!

Define the problem before choosing an ML method.

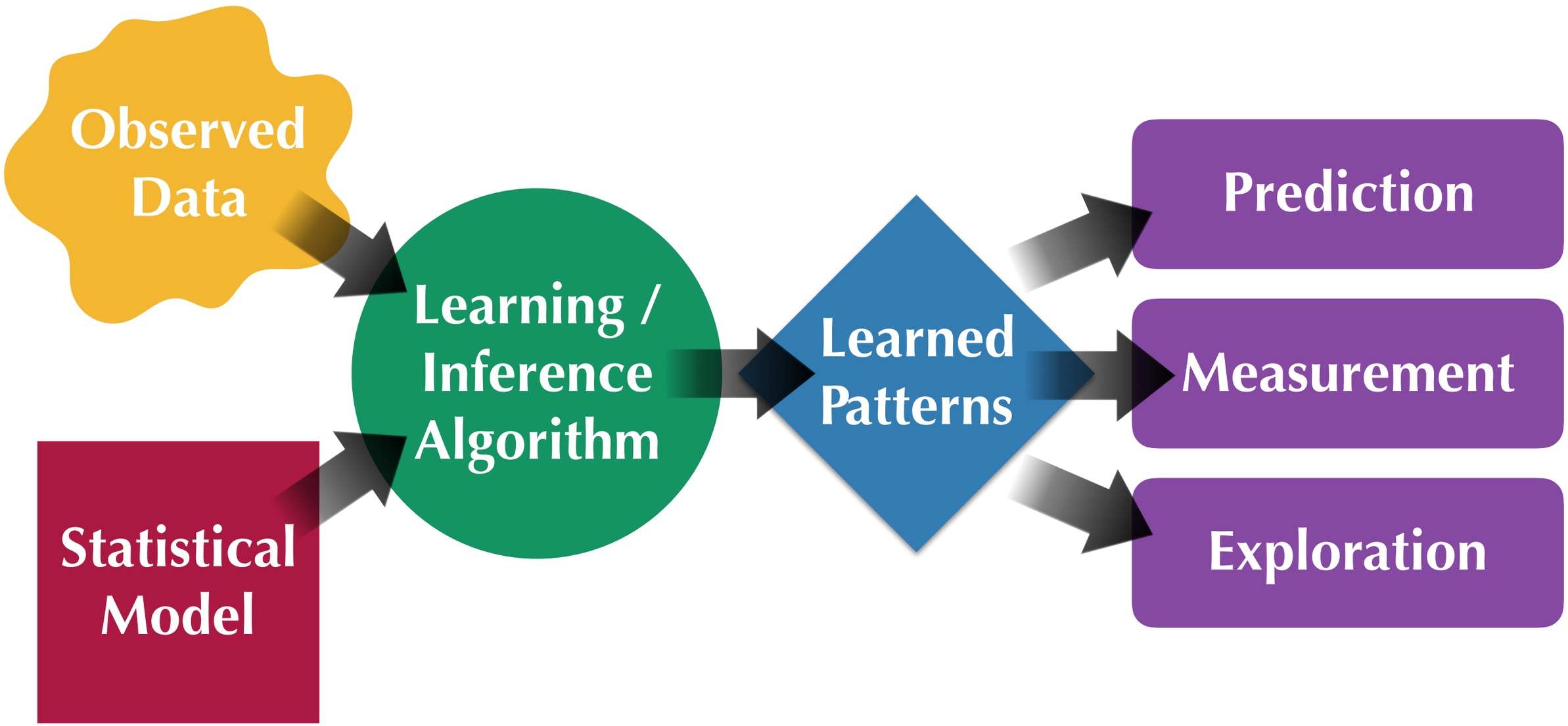
Be critical of ML model assumptions. Are a method's assumptions acceptable in the context of your work?

Ask yourself: What is the source of your data? What biases might be created by the data-generating process or your curation of the data?

Beware of tuning parameters, train/validate/test splitting, evaluation metrics, and overfitting.

What are you comparing against? Will a simple method do better?

Conducting and Interpreting Research with ML



ajbc.io/MLintro