# Introduction to Machine Learning Applications

Spring 2023

## Machine Learning tasks

**Minor Gordon**

gordom6@rpi.edu


Rensselaer

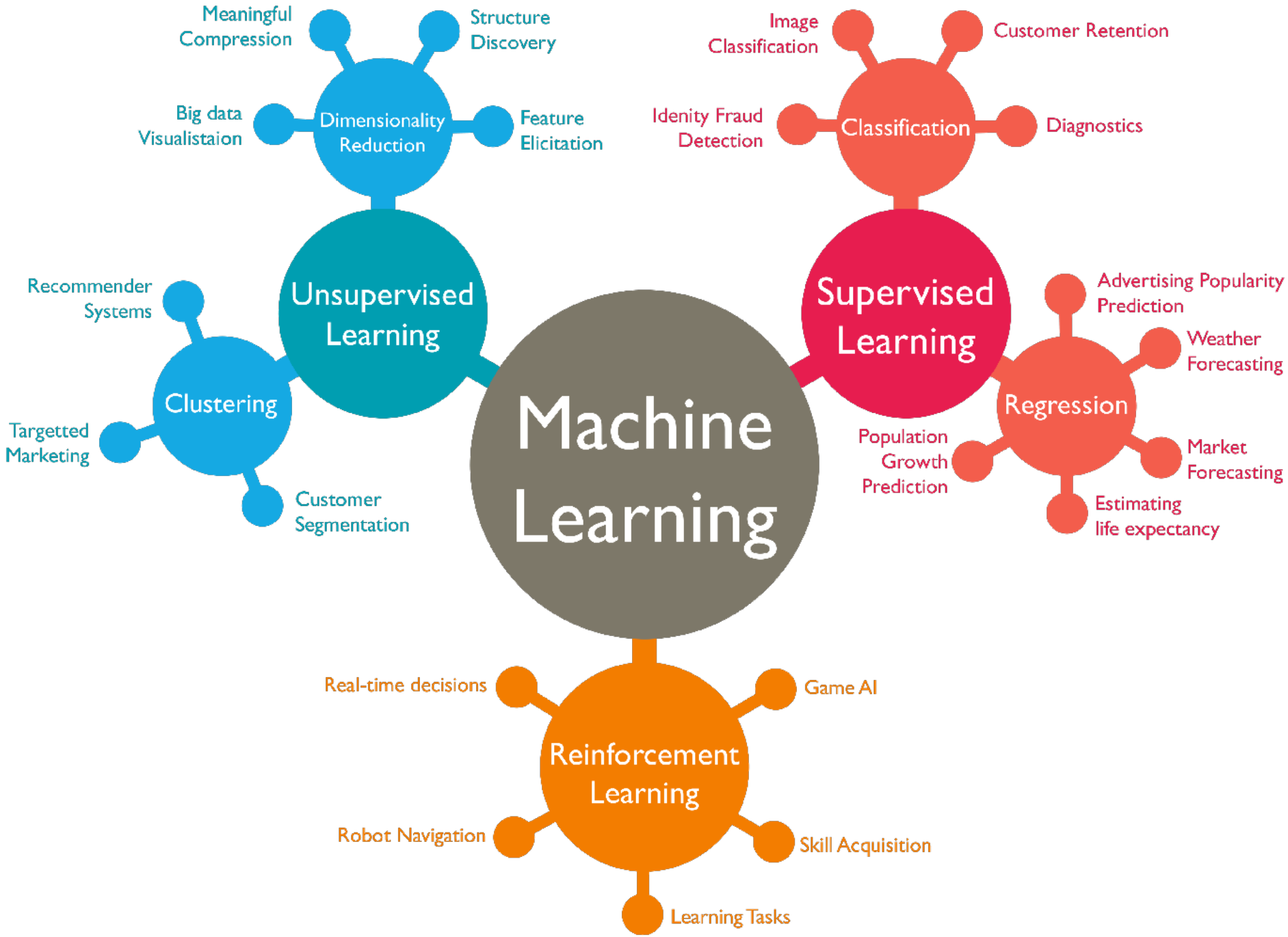
# Machine Learning

According to Tom Mitchell (1998):

Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E

Well-defined learning task: <P, T, E>


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

# Supervised Learning

“**Supervised learning** is the machine learning task of inferring a function from *labeled training data*.The training data consist of a set of *training examples*.” - Wikipedia

# Supervised Learning

- Prediction with focused target variable
- Training data provided
- Example:
  - Iris Example
  - Titanic Example
  - Housing prices
  - Nearly every Kaggle (there are some exploratory visualization tasks that wouldn’t be supervised)

# Supervised Learning

- Must have a defined problem, dataset, ideal solution

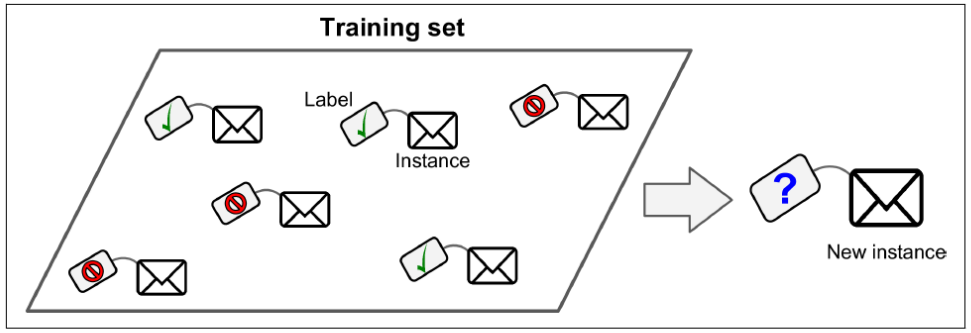



*Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)*


*Figure 1-6. Regression*

# Can you predict species from characteristics given a training set?

Iris setosa

Iris versicolor

Iris virginica

https://en.wikipedia.org/wiki/Iris_flower_data_set

# Sample Data

| sepal_length | sepal_width | petal_length | petal_width | species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | versicolor |
| 4.9 | 3 | 1.4 | 0.2 | versicolor |
| 4.7 | 3.2 | 1.3 | 0.2 | versicolor |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |

# MNIST Dataset

Each has a real label of the actual number

# Unsupervised Learning

“**Unsupervised machine learning** is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations). Since the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the relevant.”
- Wikipedia

# How can we do an analysis if we don’t know the dependent variables?

# Unsupervised Learning

- Finding hidden structures in unlabeled data
- No target dependent variable is provided
- Example: Clustering
  - K-Means
  - DBSCAN
  - Hierarchical Cluster Analysis
- Anomaly detection
  - Once class SVM
  - Isolation Forest

# Unsupervised Learning

- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - t-distributed stochastic neighbor embedding (t-SNE)

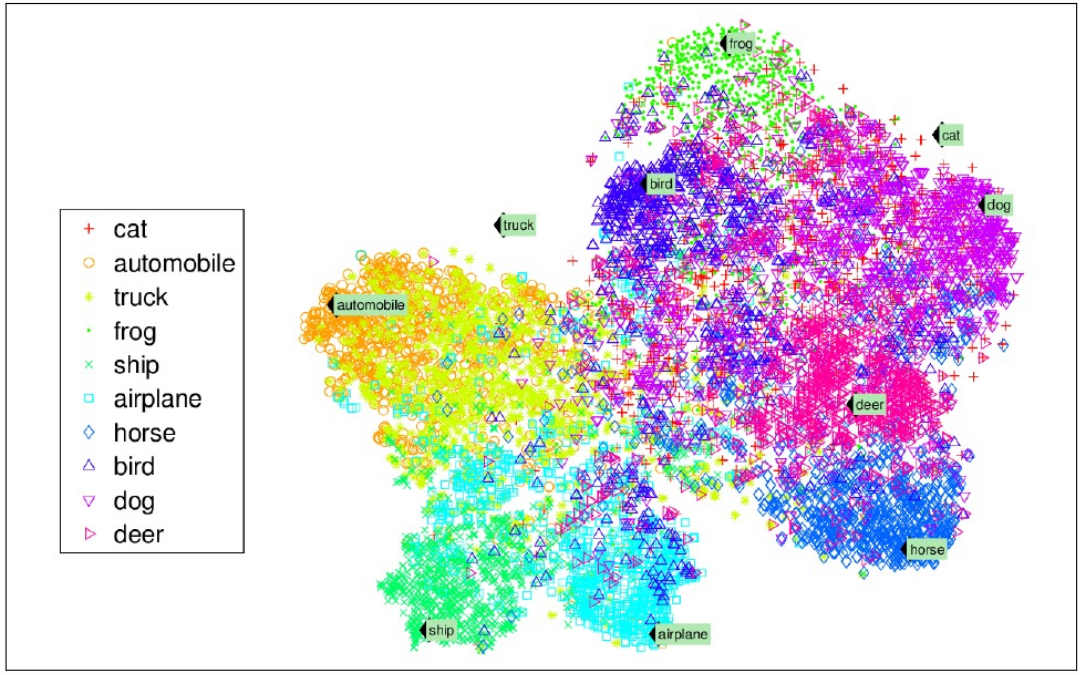



*Figure 1-9. Example of a t-SNE visualization highlighting semantic clusters*[3]

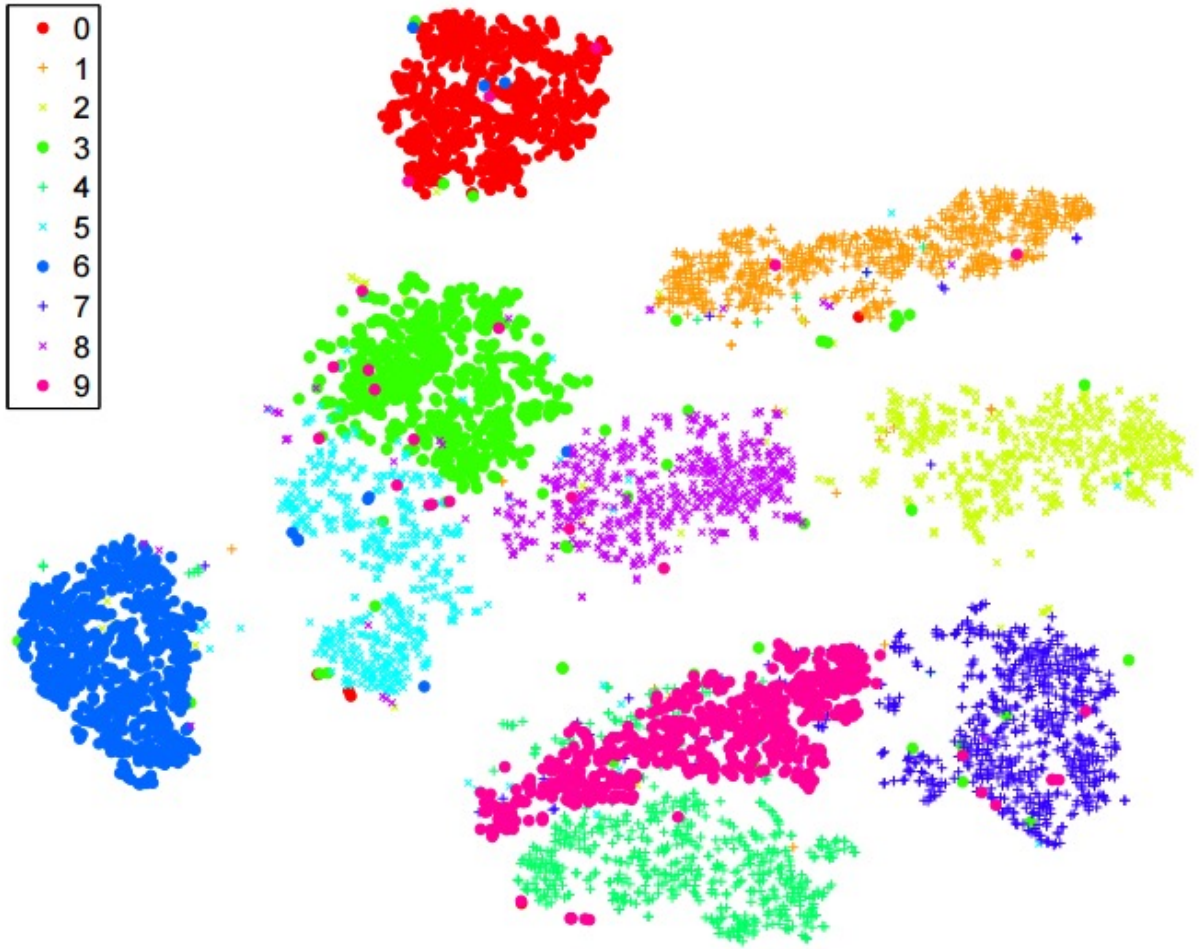
Lets say you develop and algorithm to immediately measure all flowers in a field via image data.

# Sample Data

| sepal_length | sepal_width | petal_length | petal_width |
|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 |
| 4.9 | 3 | 1.4 | 0.2 |
| 4.7 | 3.2 | 1.3 | 0.2 |
| 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 3.6 | 1.4 | 0.2 |
| 5.4 | 3.9 | 1.7 | 0.4 |
| 4.6 | 3.4 | 1.4 | 0.3 |

How many different species might there be in the flower patch?

# Clustering




Can apply clustering techniques even when working with things that could be supervised datasets

https://stats.stackexchange.com/questions/270391/should-

Unsupervised Learning + Supervised Learning

# Combining Supervised and Unsupervised

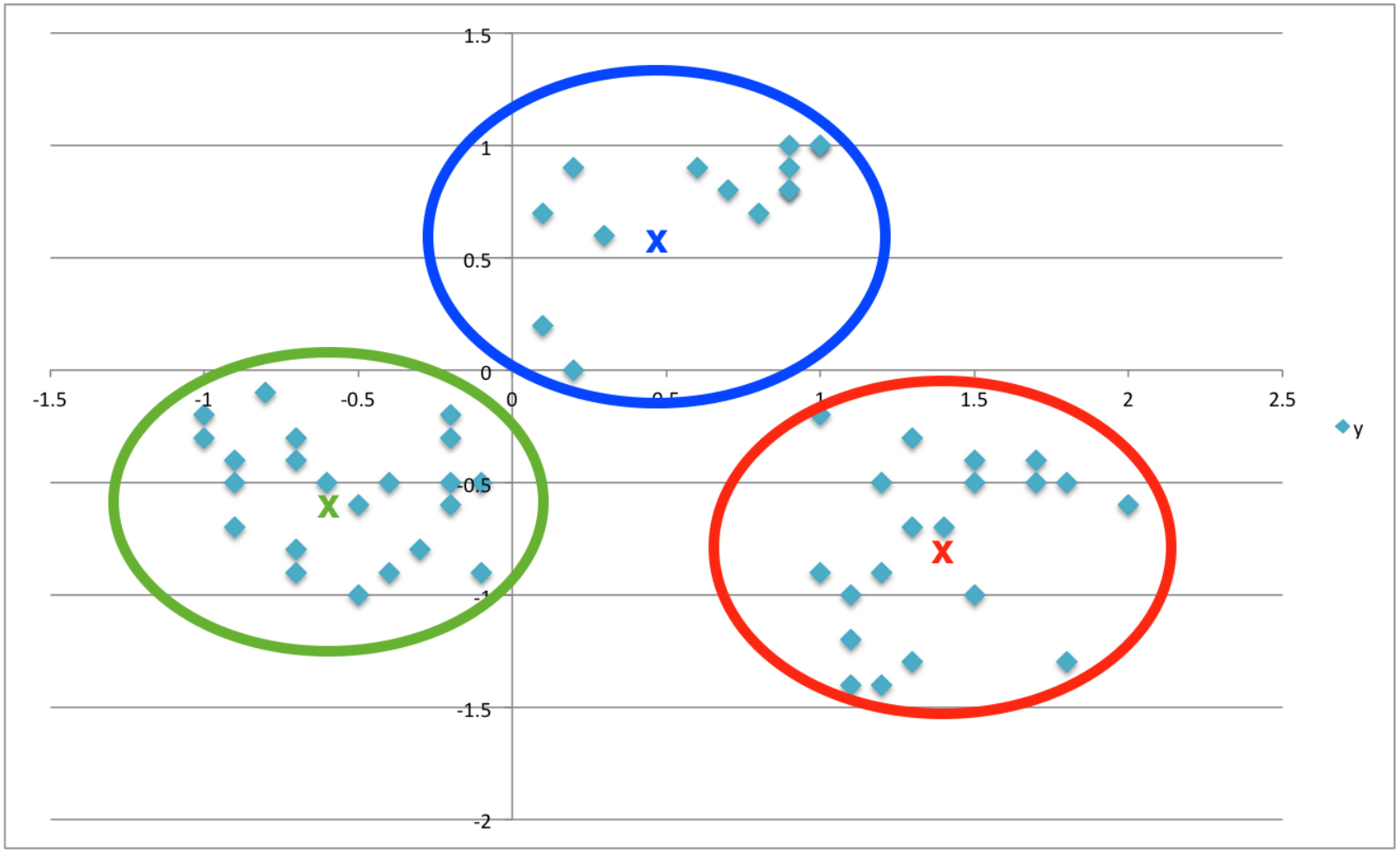
Two methods:

- Utilize unsupervised learning for feature creation
- Semi-supervised learning (advanced technique) may combine supervised and unsupervised learning, a single labeled data point could make it clear the class of many

# Customers and Overall Sales?

| offer_id | campaign | varietal | min_qty | discount | origin | past_peak | |
|---|---|---|---|---|---|---|---|
| 0 | 1 | January | Malbec | 72 | 56 | France | False |
| 1 | 2 | January | Pinot Noir | 72 | 17 | France | False |
| 2 | 3 | February | Espumante | 144 | 32 | Oregon | True |
| 3 | 4 | February | Champagne | 72 | 48 | France | True |
| 4 | 5 | February | Cabernet Sauvignon | 144 | 44 | New Zealand | True |

# Different Types of Customers



# Customers and Overall Sales?

- Use unstructured data to identify clusters
- Clusters become factor variables (dummy variables), (we know this stuff, feature creation!)