COMPSCI 389
Introduction to Machine Learning

Introduction to Supervised Learning

Prof. Philip S. Thomas (pthomas@cs.umass.edu)

Data & Supervised Learning

- Different subfields of ML assume access to different kinds of data.
- During the first part of the course, we will focus on **supervised learning** problems.
- These are problems where the data is a set of points, and so it is called a **data set** or **dataset.**
  - Data set ≠ Database
- Each point consists of a pair of **inputs** and **outputs**.
- Given a data set of such input-output pairs, a supervised learning algorithm learns to predict the output given the input, even for points not in the data set.

Data Set Notation

- X: **Input** (also called **features**, **attributes**, **covariates**, or **predictors**)
  - Typically, X is a vector, array, or list of numbers or strings.
- Y: **Output** (also called **labels** or **targets**)
  - Typically, Y is a single number or string.
- An input-output pair is (X, Y).

- **Note**: We will *frequently* flip between terms for X and Y.
  - Different sources use different terms, and it's important to be comfortable with all of them.

Example Input-Output Pairs

- Predict university student GPAs from entrance exam scores.
  - Features = scores on 9 entrance exams.
  - Labels = GPA
  - Example input-output pair:

$$((622.6, 491.56, 439.93, 707.64, 663.65, 557.09, 711.37, 731.31, 509.8), 1.33333)$$

9 exam scores
X

GPA
Y

Example Input-Output Pairs

- Predict whether a sentence is a lie.
  - **Input**: A statement made by a person.
  - **Output:** A label indicating whether the sentence was truthful or a lie.
  - Example input-output pair:

("I am not a crook.", "lie")

Statement	Truth/Lie
X	Y

Data Set Notation (Revisited)

- X: **Input** (also called **features**, **attributes**, **covariates**, or **predictors**)
  - Typically, X is a vector, array, or list of numbers or strings.
- Y: **Output** (also called **labels** or **targets**)
  - Typically, Y is a single number or string.
- An input-output pair is (X, Y).
- Let n, called the **data set size** or **size of the data set**, be the number of input-output pairs in the data set.
- Let (X_i, Y_i) denote the i^{th} input output pair.
- The complete data set is

$$(X_i, Y_i)_{i=1}^{n} = \big((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\big).$$

Data Set Notation

$$(X_i, Y_i)_{i=1}^{n} = \big((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\big)$$

- Different sources use different notation.
  - X, X_i, x, or x_i can denote one input vector.
  - X can denote all the input vectors:
    $$X = (X_i)_{i=1}^{n}.$$
  - Y, Y_i, y, or y_i can represent one label.
  - Y or y can represent all the labels:
    $$Y = (Y_i)_{i=1}^{n}.$$
- Upper and lower case can mean different things:
  - Upper case = matrix (2-dimensional table), lower case = vector.
    - (X, y) denotes a complete data set. (We'll see this later in our code!)
  - Upper case = random variable, lower case = constant.

Review: Countable, Uncountable

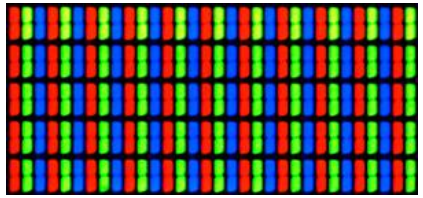
- Many (infinite!) different infinities
- The two most common are *countable infinity* and *uncountable infinity*.
- **Countable Infinity $\aleph_0$**
  - The number of integers.
  - The cardinality (size) of any set that has a one-to-one correspondence with the integers.
- **Uncountable Infinity $\geq \aleph_1$**
  - The number of real numbers. ($\aleph_1$)
  - The cardinality (size) of any set that has a one-to-one correspondence with the real numbers.
  - **Note**: Uncountable infinity actually refers to any infinity at least as large as the cardinality of the reals.
- **Countable Set (cardinality $\leq \aleph_0$)**
  - A set with finite or countable cardinality (size).
  - E.g., days of the week or even numbers
- **Uncountable Set**: Not countable.
- **Question**: Are there more natural numbers or even natural numbers?

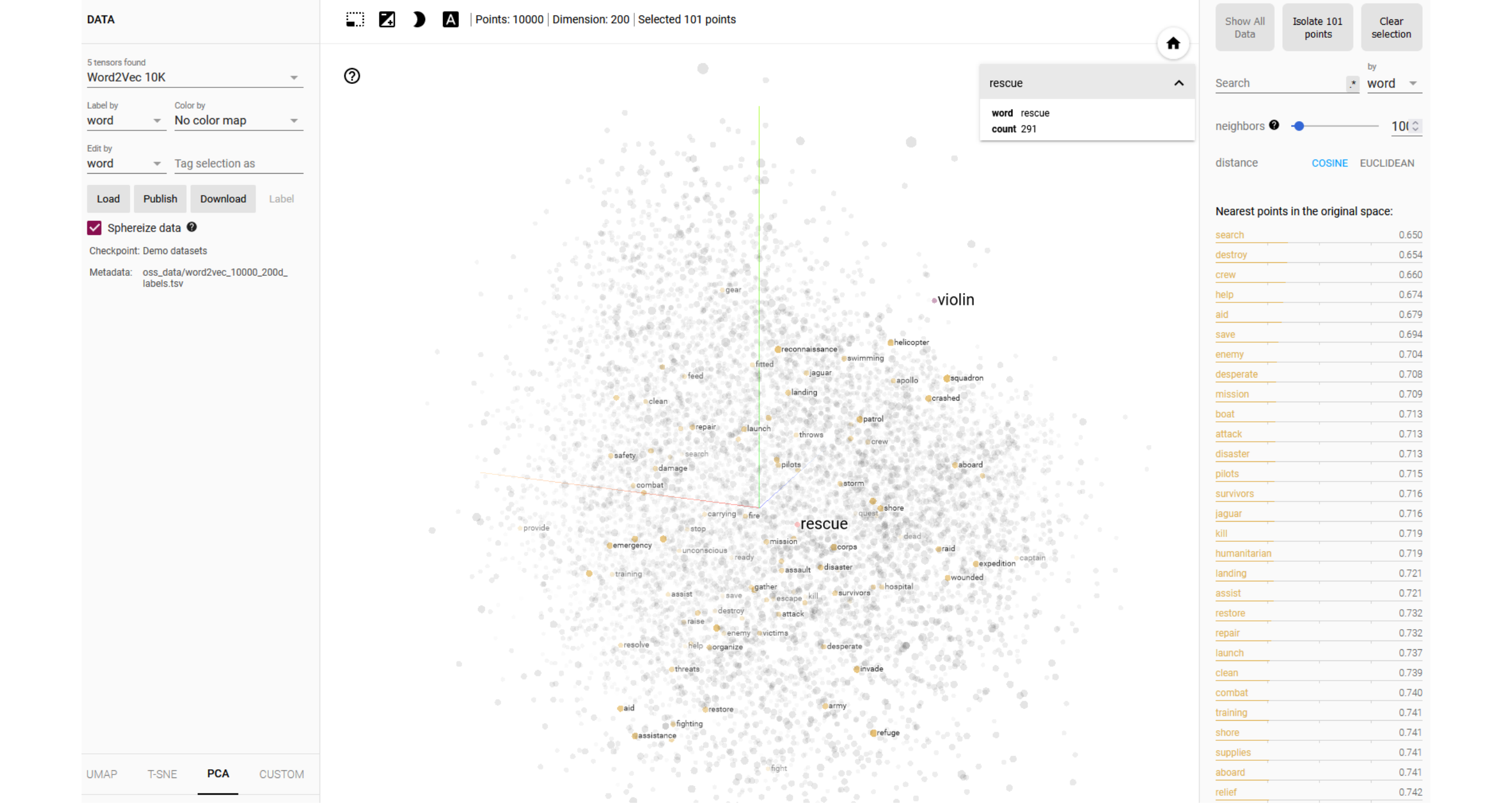
Feature Types

- **Numerical**
  - **Continuous**: Features that can take any value in a range, like temperature or velocity. These features have an uncountable set of possible values.
  - **Discrete**: Features that take a countable number of distinct values, like the number of cats a person owns. (**Binary** features are a special case.)
- **Categorical** (discrete, but not numbers)
  - **Nominal**: Unordered categories like colors (red, green, blue) or genre (drama, comedy, science fiction, etc.).
  - **Ordinal**: Categories with a specific order like educational level (high school, bachelor's, master's) or military rank (private, specialist, corporal, etc.)
- **Text/String**
- **Image**
- Others?

Feature Types

- Non-numerical features are often converted into numerical features to make them easier to work with.
  - Categorical features map to integers: “Sunday”→0, “Monday”→1, “Tuesday”→2, etc.
  - Images can be converted to sequences of (r,g,b) values describing each pixel.
  - Text can be converted to discrete or continuous features
    - **Discrete**: Each word (or part of a word) maps to a unique integer.
      - Each basic unit of text (word, character, or subword) is called a **token**.
    - **Continuous**: Each word can be mapped to a vector of real numbers. This is called a **word embedding**. Ideally, similar words are mapped to similar vectors of numbers. Word embeddings are themselves learned from data.





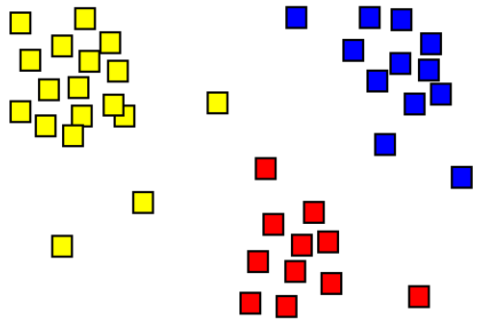
https://projector.tensorflow.org/

Why “Supervised”?

- In **supervised learning**, each data point includes a label Y indicating what the ML algorithm should provide as output when presented with input X.
  - This label provides supervision for the ML algorithm, telling it what it should do.
- In **unsupervised learning**, data points do not have labels.
  - The ML algorithm sees inputs but has no supervision telling it what it should or should not do when presented with different inputs.
- In **reinforcement learning**, the ML algorithm is told how good its outputs were, but not what the correct outputs would have been.

Unsupervised Learning

- Learning word embeddings is one example of unsupervised learning.
- **Clustering** is another common example of unsupervised learning.
- Clustering algorithms try to identify groups of similar inputs.
- Example: Given images of hand-written letters, we may want to identify the number of different letters in the alphabet and learn to distinguish between them.



Regression and Classification

- Within supervised learning, recall that a data set is a set of input-output pairs (X, Y).
- **Regression**: Y is a continuous number.
  - Multivariate Regression: Y is a vector. That is, $Y \in \mathbb{R}^m$ and $m > 1$.
- **Classification**: Y is categorical (mapped to an integer).
  - Binary Classification: $Y \in \{0,1\}$ or $Y \in \{-1,1\}$.
  - Multi-Class Classification: $Y \in \{0,1, \ldots, k\}$.

Regression ($Y \in \mathbb{R}$) or Classification ($Y \in \mathbb{Z}$)?

- Predict the location of the nearest pedestrian from an image or video taken from a car.
  - Multivariate regression
- Predict how long a person will live based on their age, gender, address, and other health indicators.
  - Regression or classification
- Predict whether a person will repay a loan.
  - Binary classification
- Predict the rating that a person would give to a movie.
  - Depends on the rating scale.

Data Set Storage

- There is no agreed upon format for storing data sets.
  - Sometimes they are in plaintext, other times they are not.
  - When in plaintext, they are often in CSV (comma separated values) files.
    - In other cases, they use semicolons or other symbols to separate values.
  - Sometimes separate files store headers saying what each feature is, other times this header is included at the start of the file.

```
physics_exam,biology_exam,history_exam,second_language_exam,geography_exam,literature_exam,portuguese_essay_exam,math_exam,chemistry_exam,gpa
622.6,491.56,439.93,707.64,663.65,557.09,711.37,731.31,509.8,1.33333
```

```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
```

Data Set Representation

- There is no agreed upon way of storing data in software.
  - The entire data set could be one large matrix (two-dimensional array).
  - The data set could be stored as an array of points, each having an X component and a Y component.
  - More commonly, the X values can be stored separately from the Y values.
    - The X values can be stored as a matrix.
    - The X values can be stored as an array of arrays (vector of vectors).
    - These structures could be built in structures in your programming language, or structures built for efficient linear algebra operations.
- One *common* way in python is using the **pandas** library.