**1.3**

**A History of Data Mining**

Data mining

Builds on statistics, artificial intelligence, and machine learning

Is the marriage of statistics and computer science

John Tukey: a founder of modern data science and data mining

Coined the term *bit*

Was coinventor of fast Fourier transform

Popularized exploring data, with visualization

Looked at preprocessing techniques

**A History of Data Mining (cont.)**

Gregory Piatesky-Shapiro

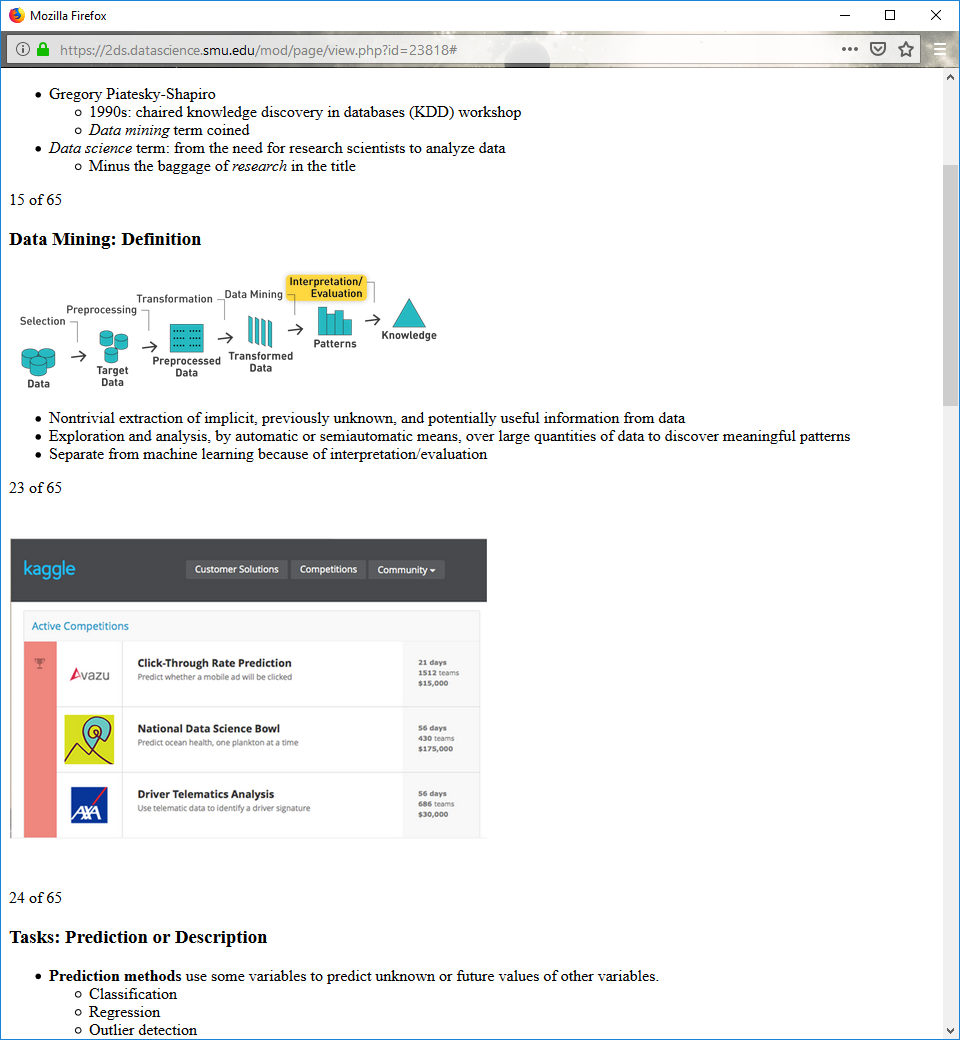
1990s: chaired knowledge discovery in databases (KDD) workshop

*Data mining* term coined

*Data science* term: from the need for research scientists to analyze data

Minus the baggage of *research* in the title

**Data Mining: Definition**



Nontrivial extraction of implicit, previously unknown, and potentially useful information from data

Exploration and analysis, by automatic or semiautomatic means, over large quantities of data to discover meaningful patterns

Separate from machine learning because of interpretation/evaluation

**Tasks: Prediction or Description**

**Prediction** methods use some variables to predict unknown or future values of other variables.

Classification

Regression

Outlier detection

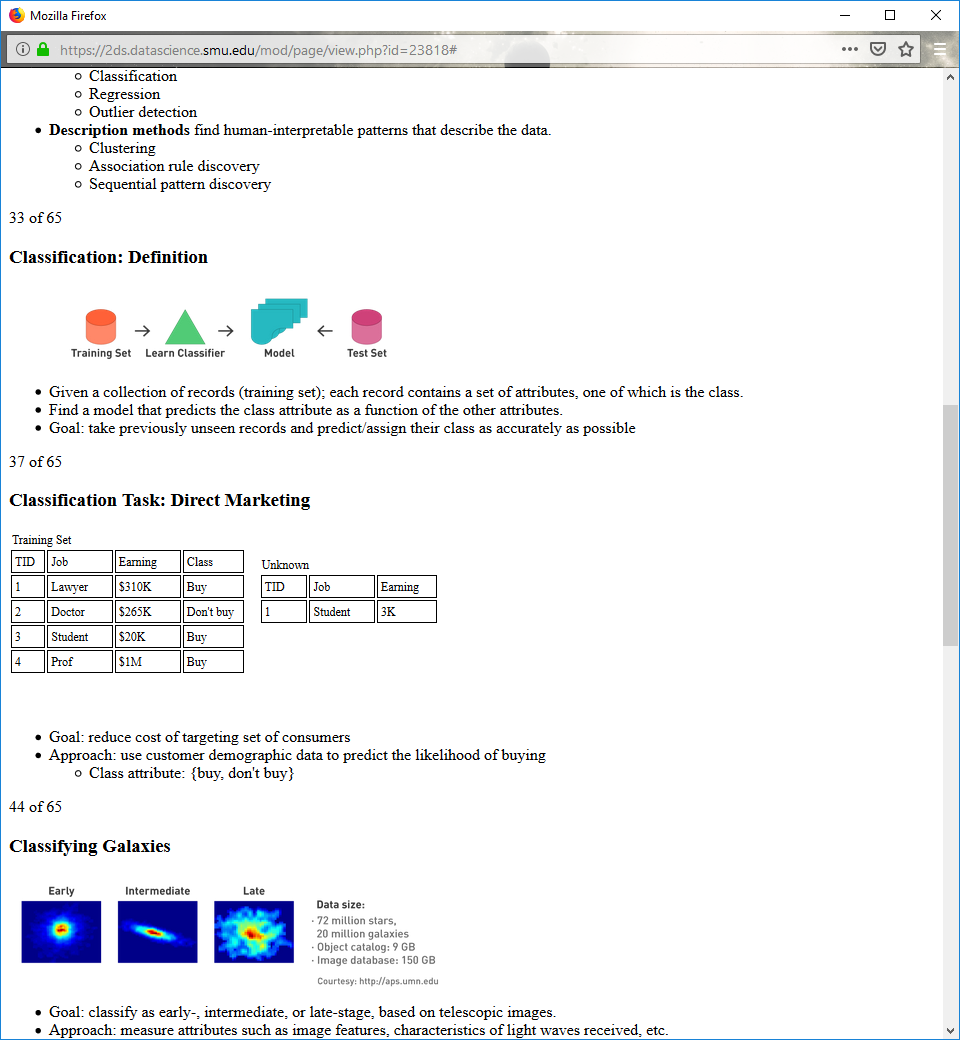
**Description** methods find human-interpretable patterns that describe the data.

Clustering

Association rule discovery

Sequential pattern discovery

**Classification: Definition**

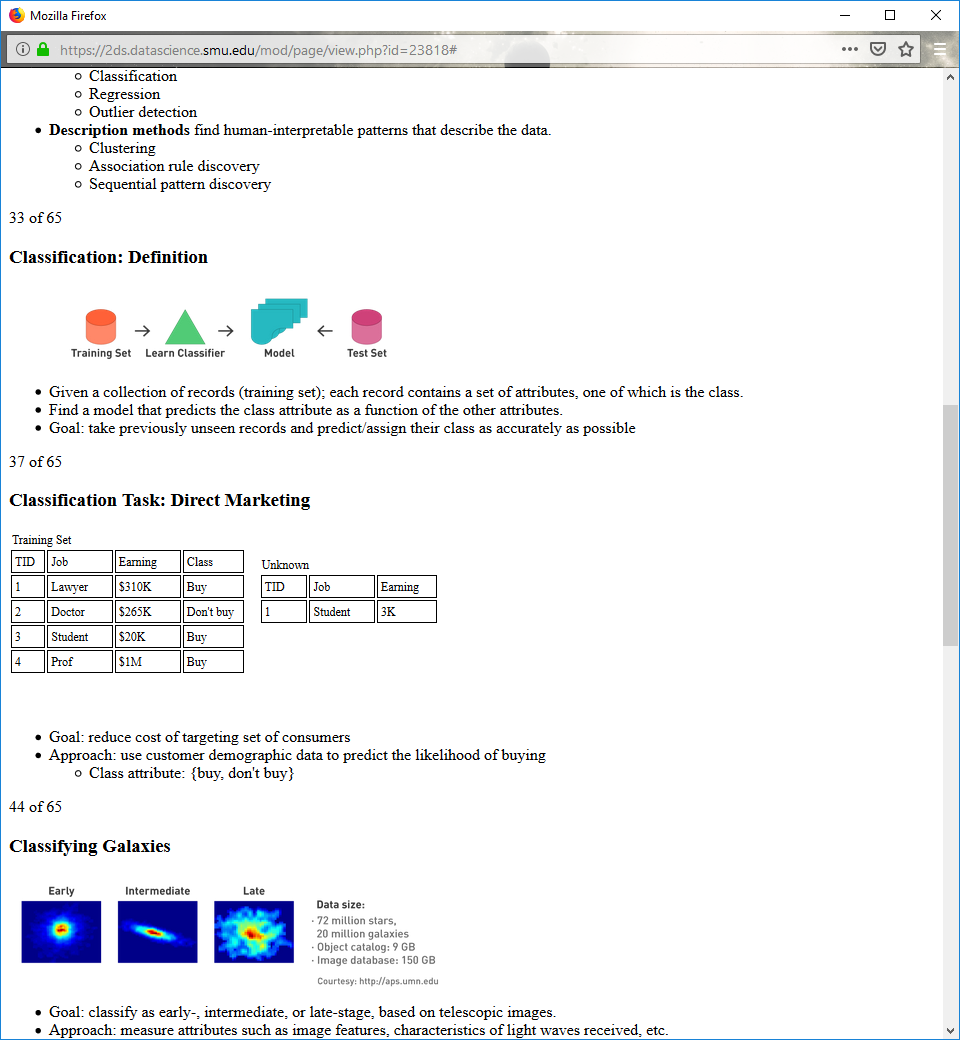


Given a collection of records (training set); each record contains a set of attributes, one of which is the class.

Find a model that predicts the class attribute as a function of the other attributes.

Goal: take previously unseen records and predict/assign their class as accurately as possible

**Classification Task: Direct Marketing**

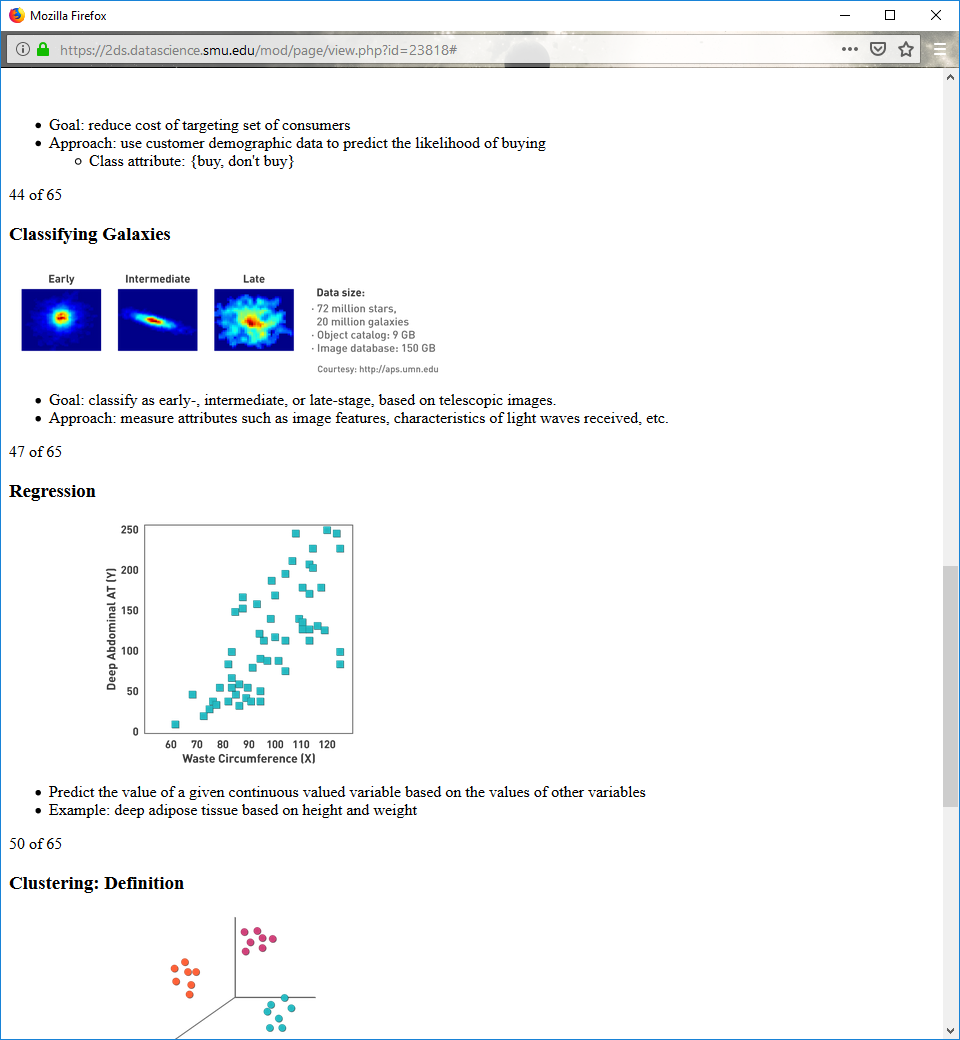


Goal: reduce cost of targeting set of consumers

Approach: use customer demographic data to predict the likelihood of buying

Class attribute: {buy, don't buy}

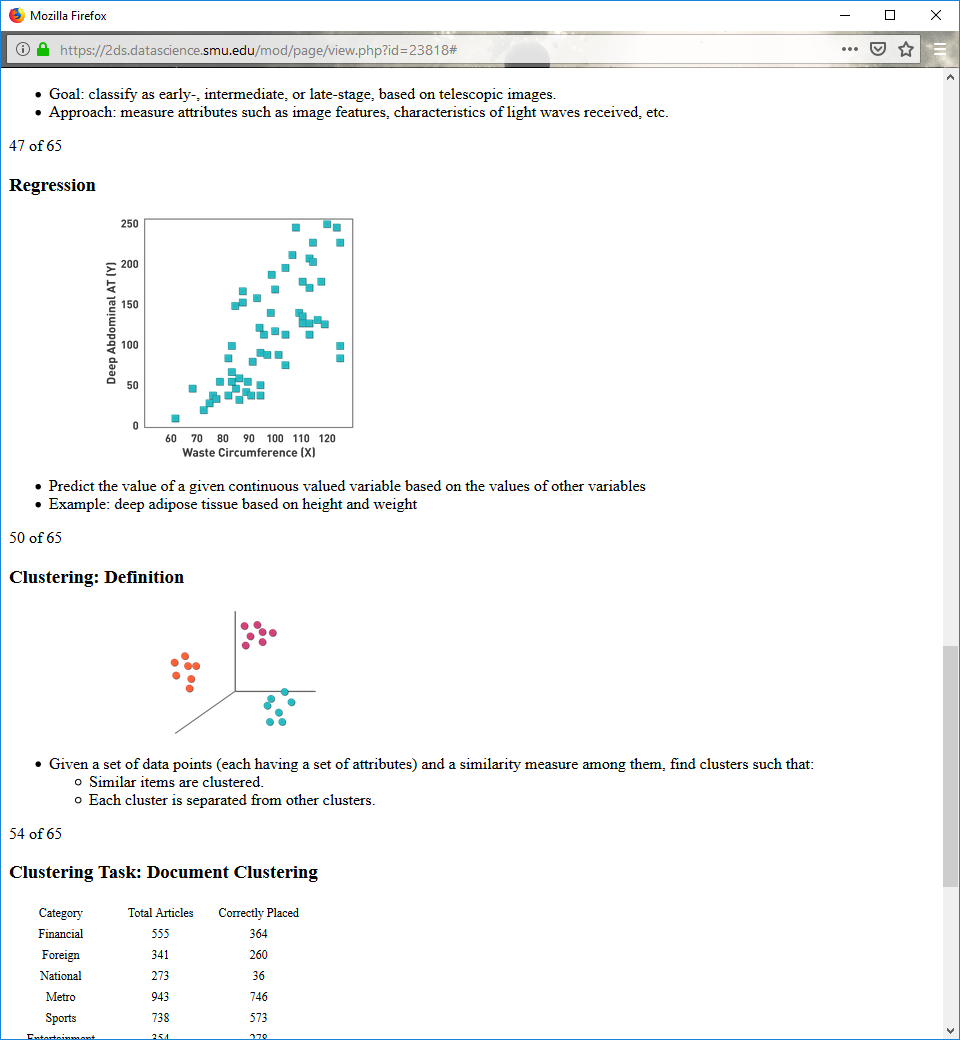
**Classifying Galaxies**



Goal: classify as early-, intermediate, or late-stage, based on telescopic images.

Approach: measure attributes such as image features, characteristics of light waves received, etc.

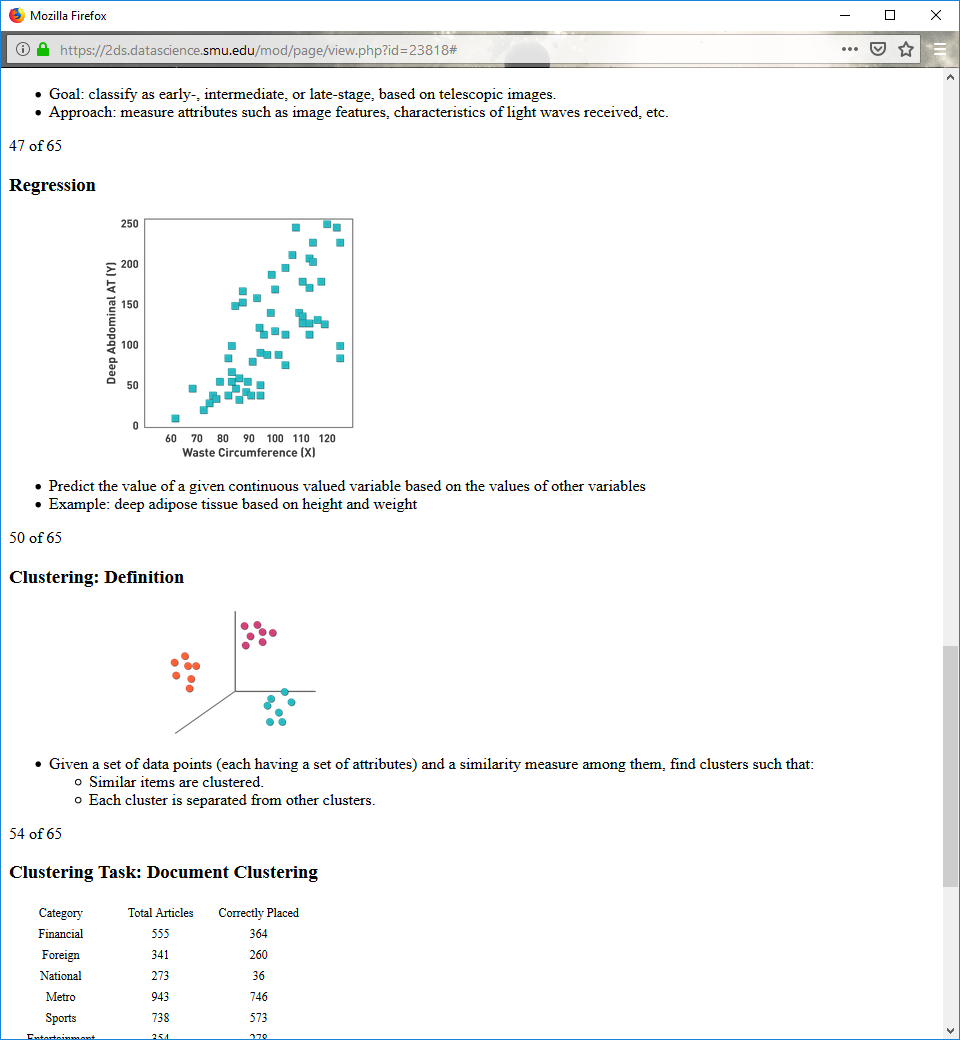
**Regression**



Predict the value of a given continuous valued variable based on the values of other variables

Example: deep adipose tissue based on height and weight

**Clustering: Definition**

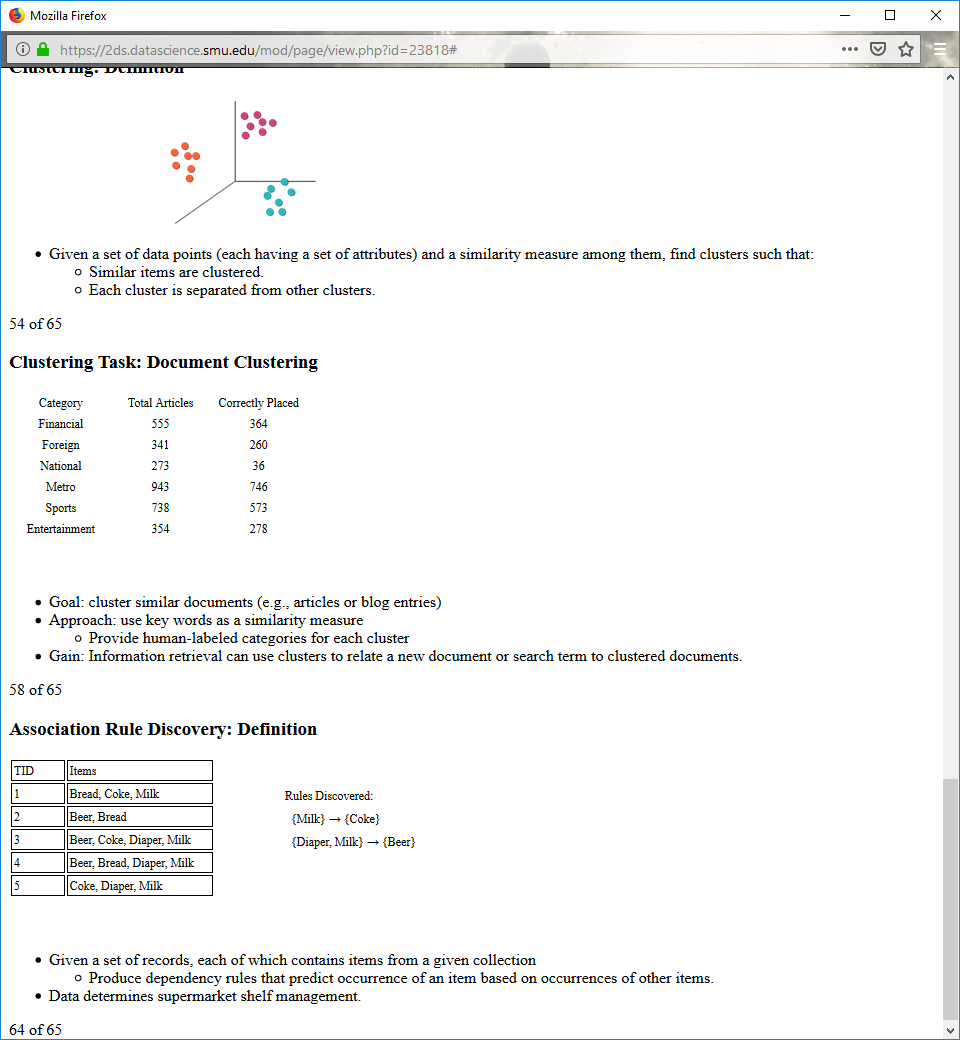


Given a set of data points (each having a set of attributes) and a similarity measure among them, find clusters such that:

Similar items are clustered.

Each cluster is separated from other clusters.

**Clustering Task: Document Clustering**



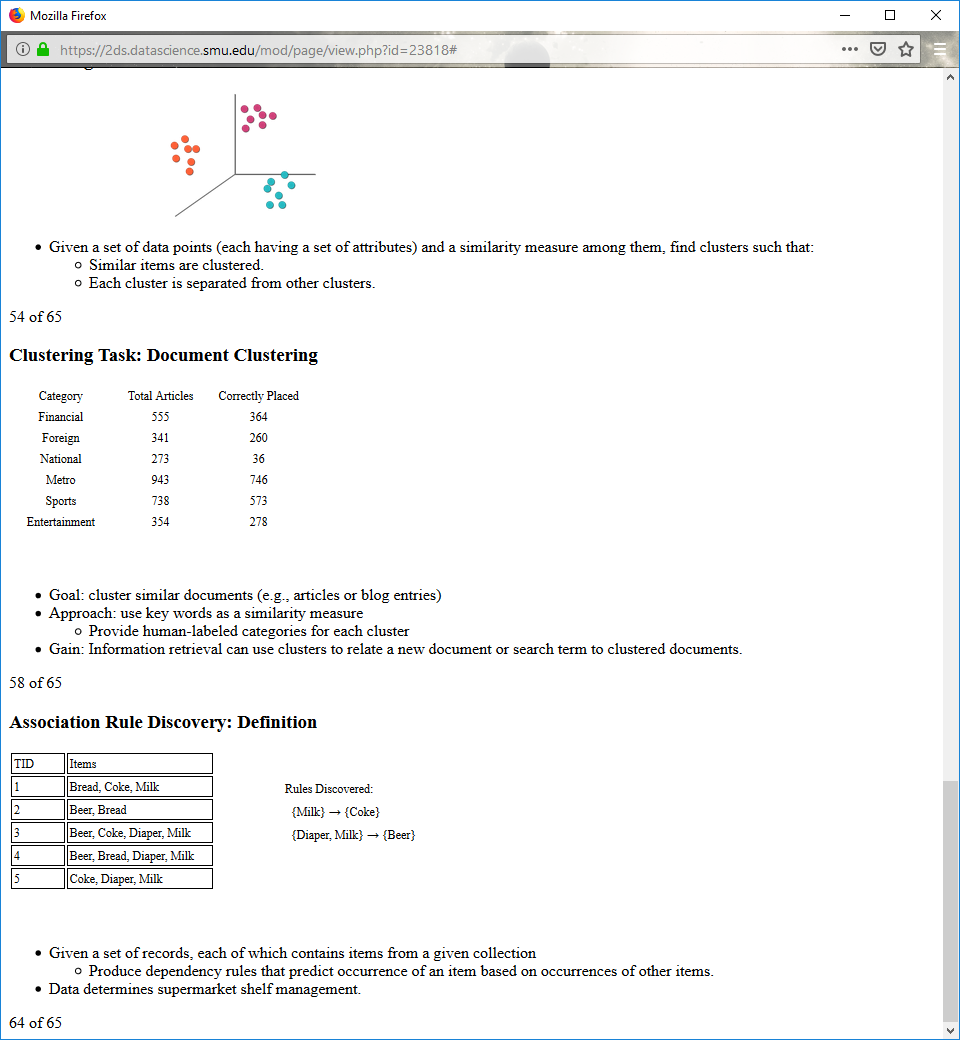
Goal: cluster similar documents (e.g., articles or blog entries)

Approach: use key words as a similarity measure

Provide human-labeled categories for each cluster

Gain: Information retrieval can use clusters to relate a new document or search term to clustered documents.

**Association Rule Discovery: Definition**



Given a set of records, each of which contains items from a given collection

Produce dependency rules that predict occurrence of an item based on occurrences of other items.

Data determines supermarket shelf management.

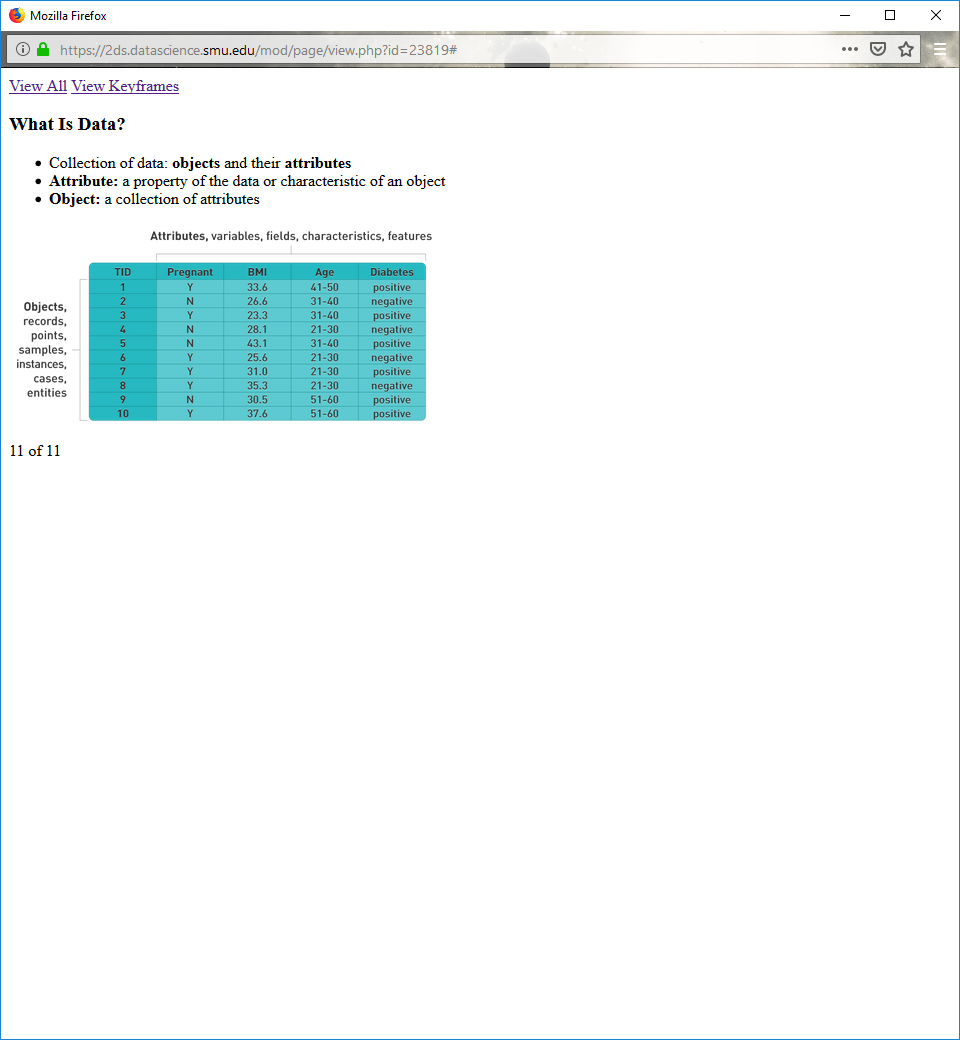
**1.4**

**What Is Data?**

Collection of data: **objects** and their **attributes**

**Attribute:** a property of the data or characteristic of an object

**Object:** a collection of attributes



**Types of Attributes**

There are four different types of attributes:

**Nominal**

**Ordinal**

**Interval**

**Ratio**

Attributes can be divided into two types:

**Discrete:** nominal or ordinal

**Continuous:** interval or ratio

**Properties of Attribute Values**

**Nominal:** distinctness

**Ordinal:** distinctness and order

**Interval:** distinctness, order, and addition

**Ratio:** distinctness, order, addition, and multiplication

**Examples of Attribute Values**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Examples of Attribute Values*** | | | |
| *Attribute Type* | **Description** | **Examples** | **Operations** |
| **Nominal** | The values are different names, i.e., only enough information to distinguish one object from another.  (=, ≠) | Zip codes, employee ID numbers, eye color, sex: {male, female} | Mode, entropy, contingency correlation, x2 test |
| **Ordinal** | The values of an ordinal attribute provide enough information to order objects. (<, >) | Hardness of minerals, {good, better, best}, grades, street numbers | Median, percentiles, rank correlation, run tests, sign tests |
| **Interval** | For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, – ) | Calendar dates, temperature in Celsius or Fahrenheit | Mean, standard deviation, Pearson's correlation,  t and F tests |
| **Ratio** | For ratio variables, both differences and ratios are meaningful. (\*, /) | Temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current | Geometric mean, harmonic mean, percent variation |

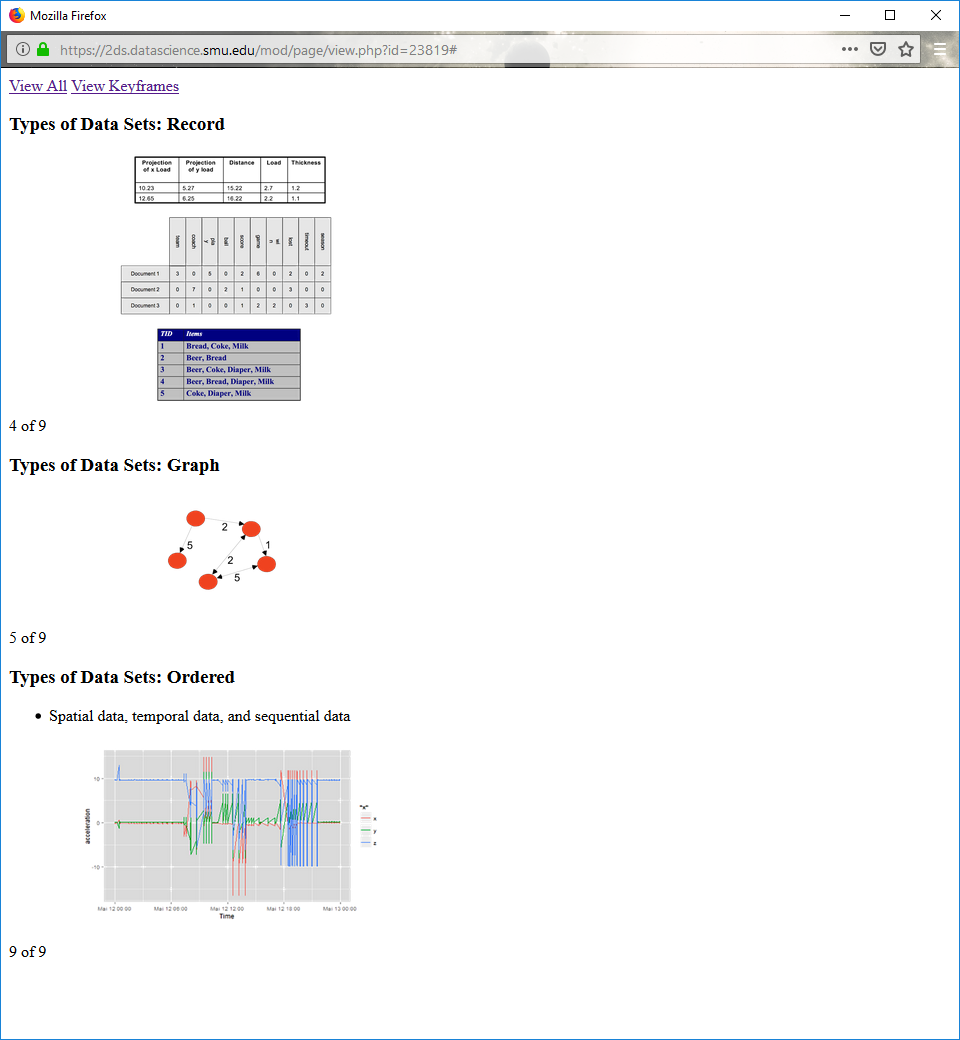
Interval and ratio typically measured with floating point value

Nominal and ordinal typically measured with a one-hot encoding, or integer

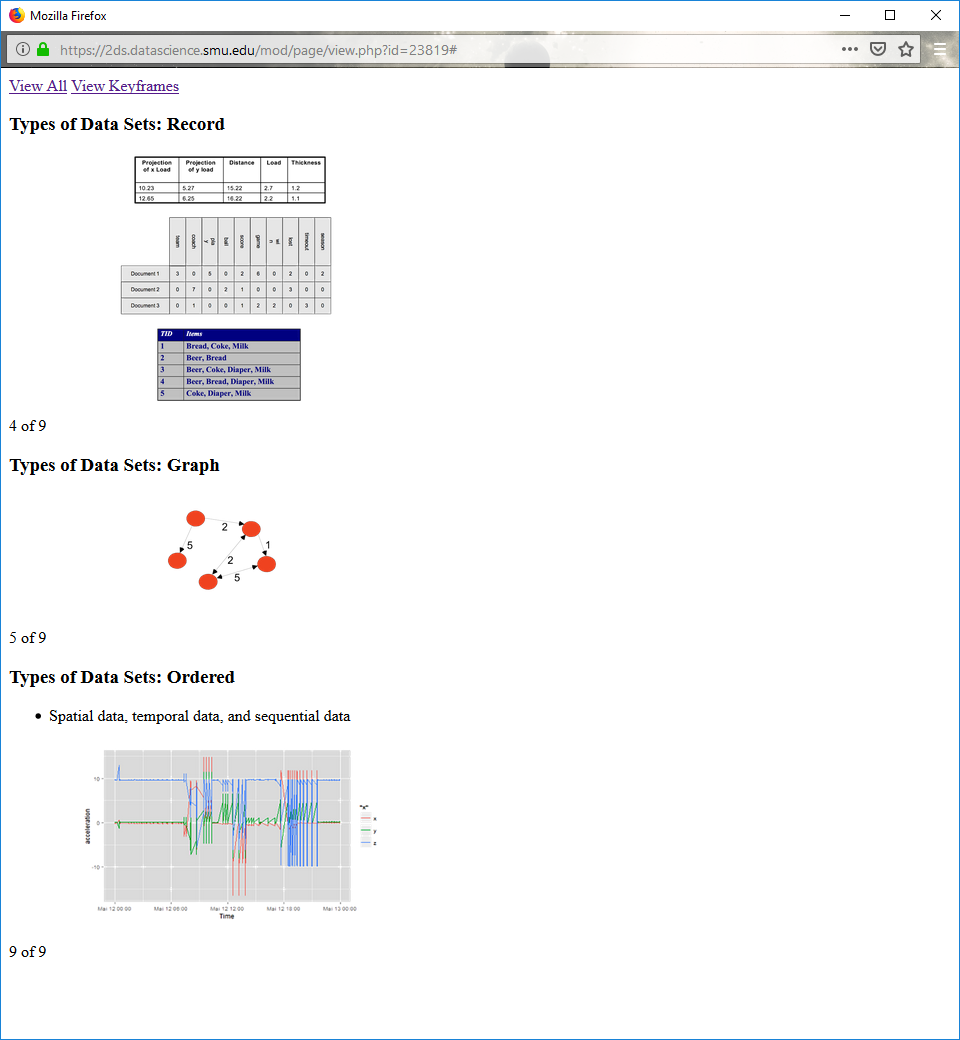
**Possible Transforms for Attribute Types**

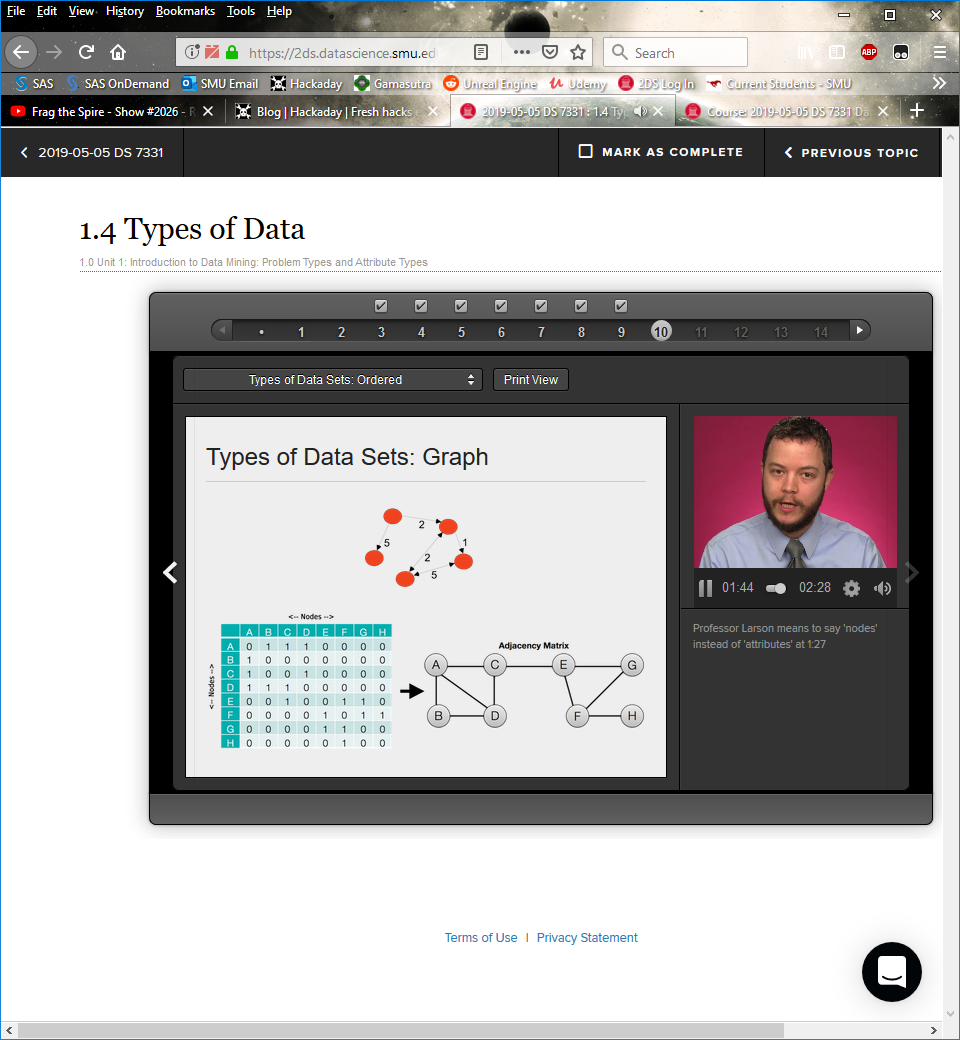
|  |  |  |  |
| --- | --- | --- | --- |
|  | Attribute Level | Transformation | Comments |
| Discrete | Nominal | Any permutation of values | If all employee ID numbers were reassigned, would it make any difference? |
| Ordinal | An order preserving change of values, i.e., new\_value = f(old\_value) where f is a monotonic function. | An attribute encompassing the notion of good, better, best can be represented equally well by the values {1, 2, 3} or {0.5, 1, 10}. |
| Continuous | Interval | Affine or linear, i.e., new\_value =a \* old\_value + b where a and b are constants | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| Ratio | Multiplication, i.e., new\_value = a \* old\_value | Length can be measured in meters or feet. |

**Types of Data Sets: Record**



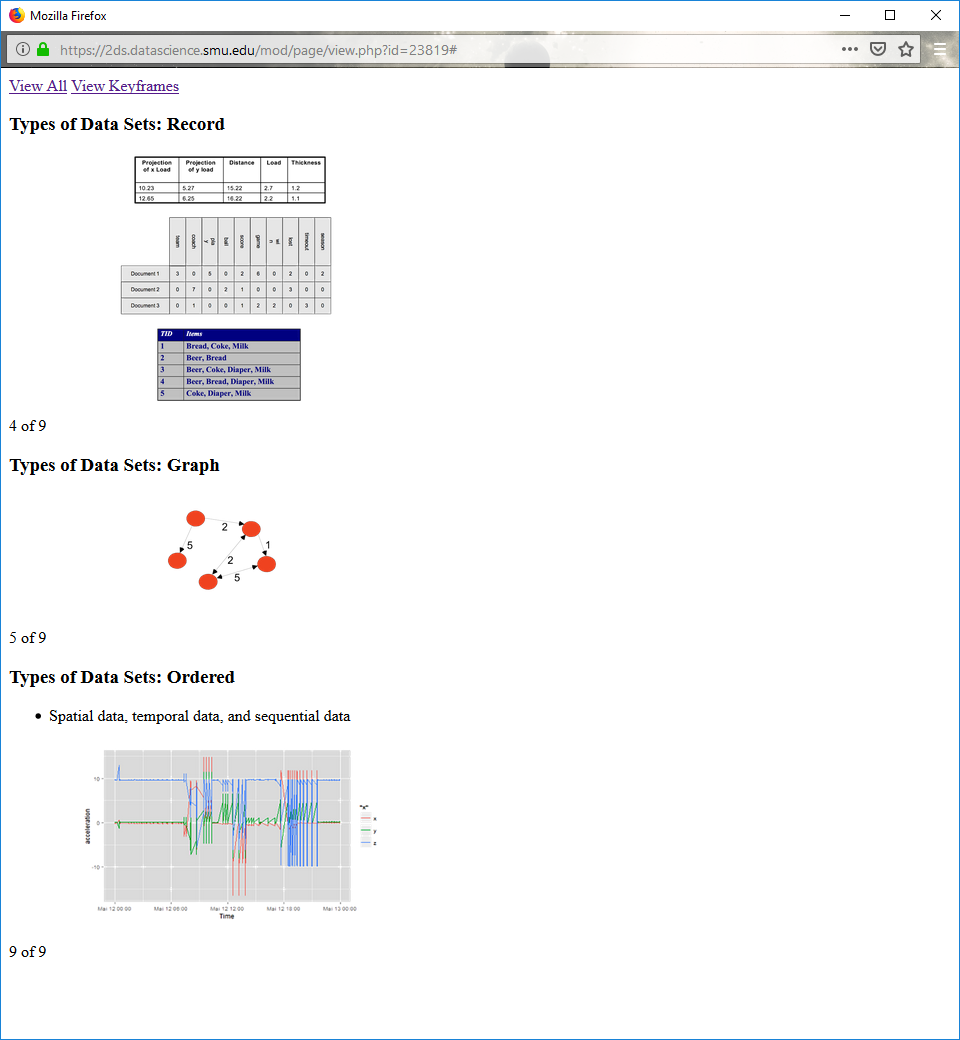
**Types of Data Sets: Graph**





**Types of Data Sets: Ordered**

Spatial data, temporal data, and sequential data



**1.6**

**Missing Values**

Causes

Not collected: e.g., third-class details

Not applicable: e.g., children's income

Imputation: lambda function

Group data by class

Fill in missing values with group medians

Eliminate entries that still have empty values

**Normalize Attributes**

Normalize values by subtracting the minimum then dividing by the maximum.

All values are now between zero and 1.

Divide values by the standard deviation.