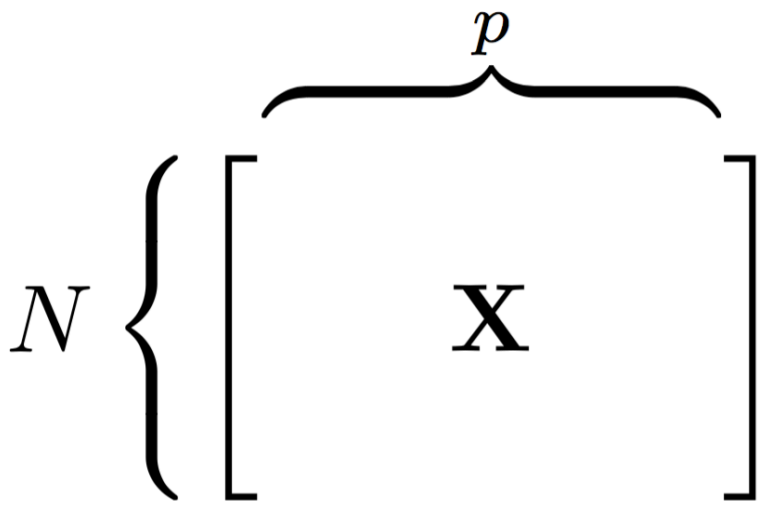
Machine Learning Basics

- Supervised Learning
- Unsupervised Learning

Data an Algorithm Understands

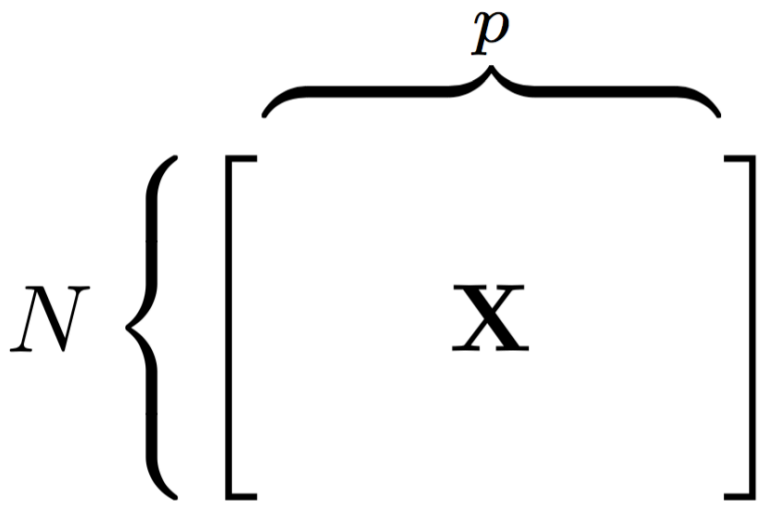
Columns



Rows

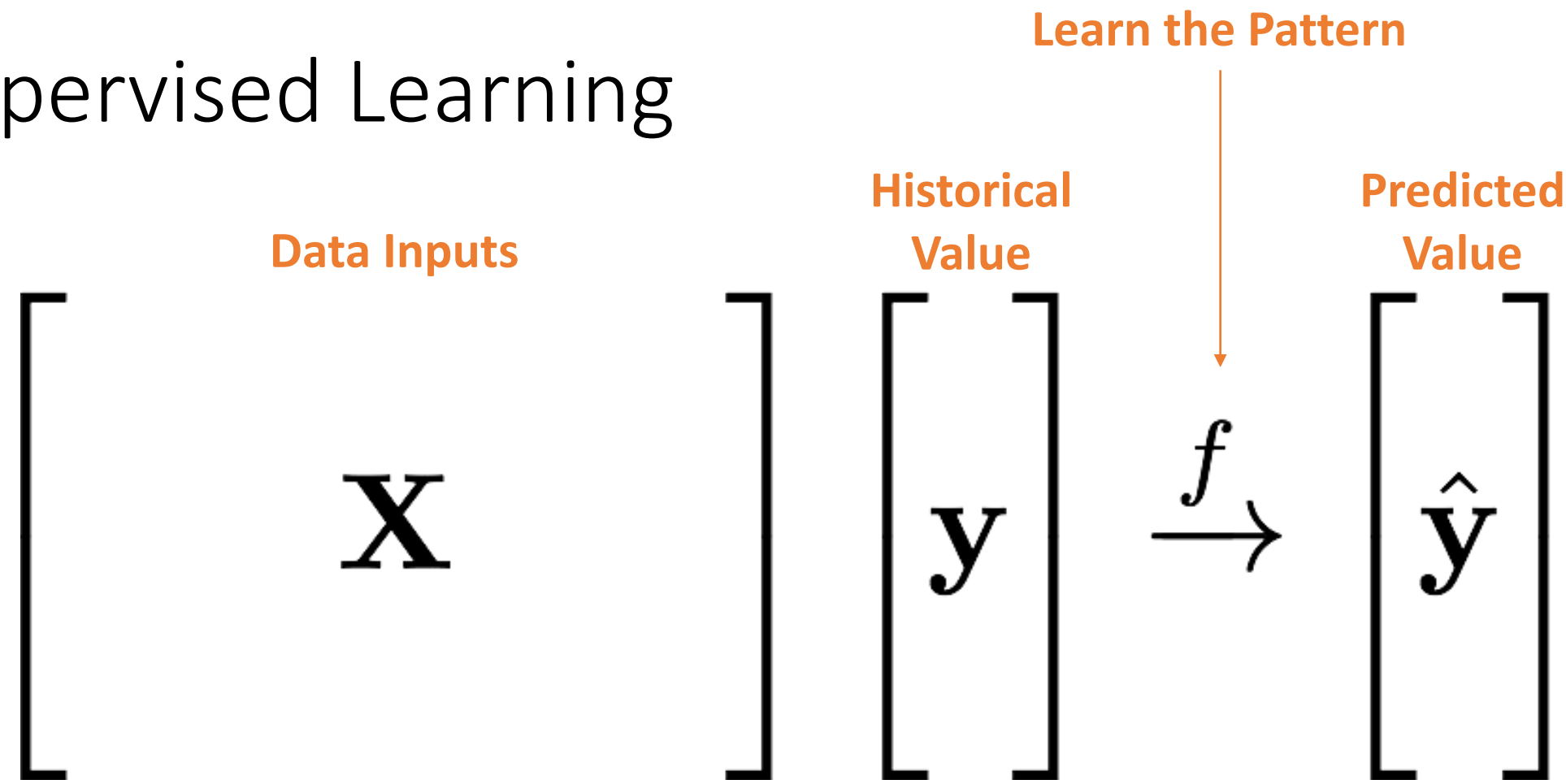
Data an Algorithm Understands

Columns ← Columns can go up to 100,000



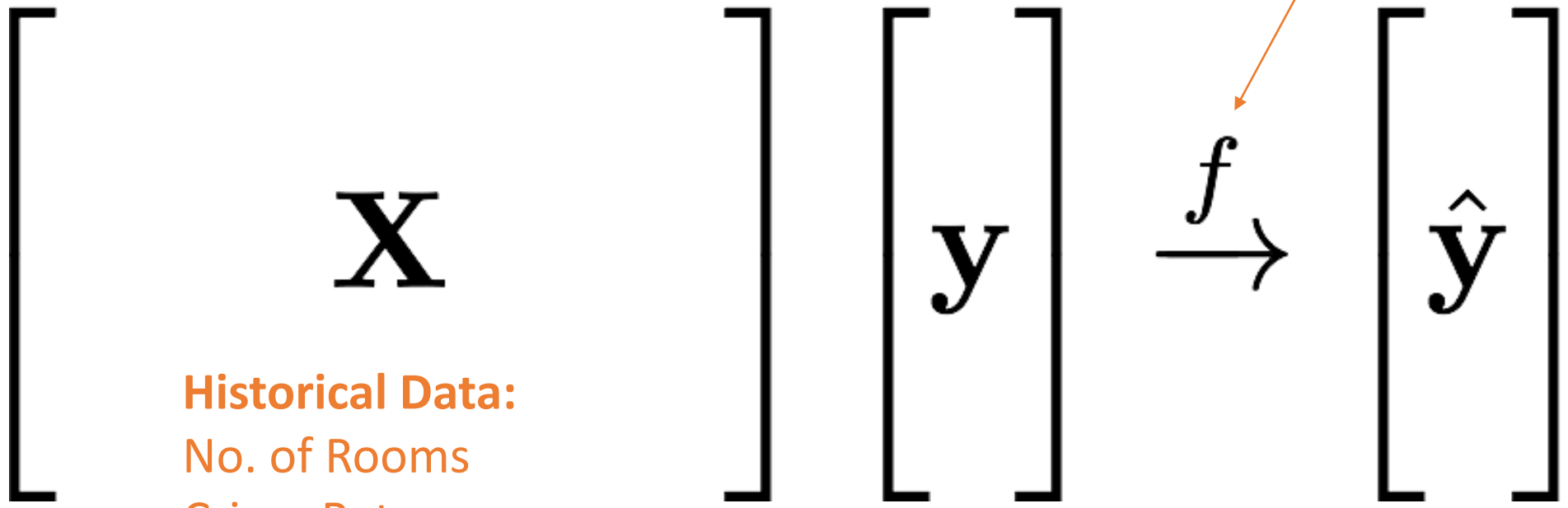
Rows ← Rows can go into **Billions** with H_2O

Supervised Learning



Supervised Learning Example

Machine Learning:
Learn Patterns
from Data



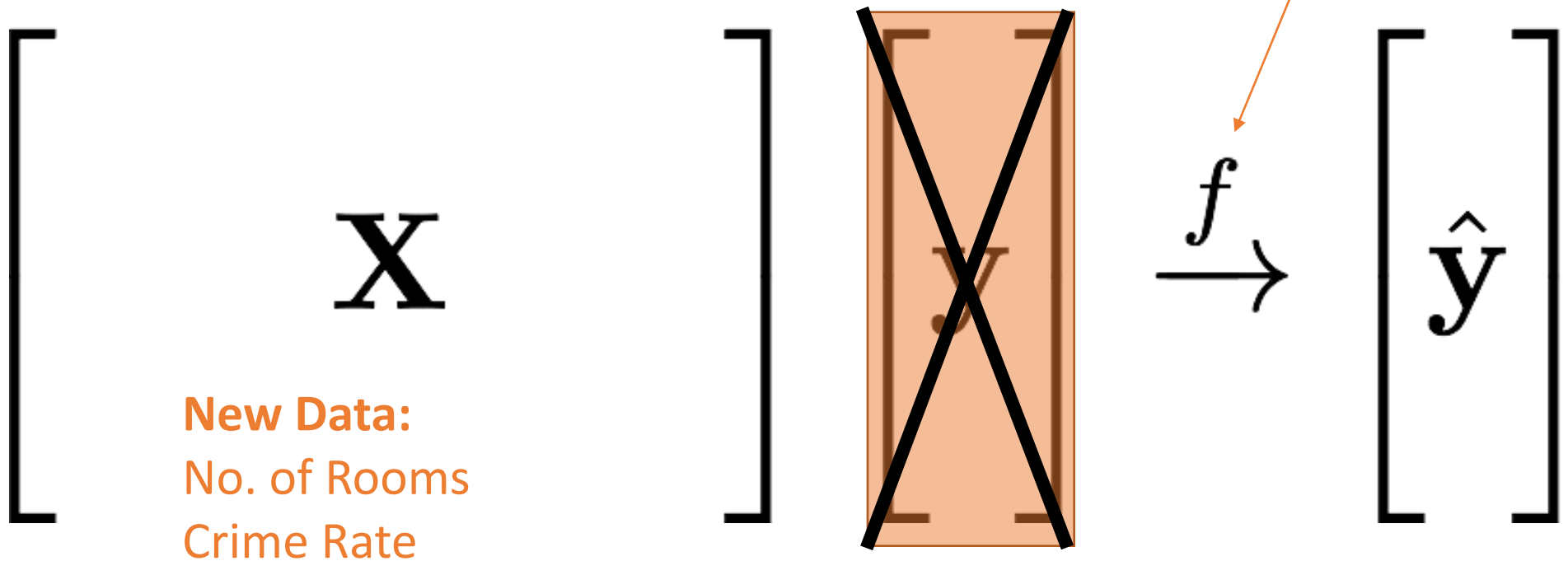
Historical Data:
No. of Rooms
Crime Rate
Pupil-Teacher Ratio
…

Target:
House Value

Predicted Value
(for evaluation)

H2O.ai

Supervised Learning Example



New Data:
No. of Rooms
Crime Rate
Pupil-Teacher Ratio
…

Target:
Unknown

Predicted Value
(for decision making)

Current Algorithm Overview

Statistical Analysis

- Linear Models (GLM)
- Naïve Bayes

Ensembles

- Random Forest
- Distributed Trees
- Gradient Boosting Machine
- Stacking / Super Learner

Deep Neural Networks

- MLP
- Autoencoder
  - Anomaly Detection
  - Deep Features

Clustering

- K-Means (Auto-*K*)

Dimension Reduction

- Principal Component Analysis
- Generalized Low Rank Models

Word Embedding

- Word2Vec

Time Series

- iSAX

Machine Learning Tuning

- Hyperparameter Search
- Early Stopping

H2O's Supervised Algorithms

Supervised Algorithms

Statistical Analysis

- Linear Models (GLM)
- Naïve Bayes

Ensembles

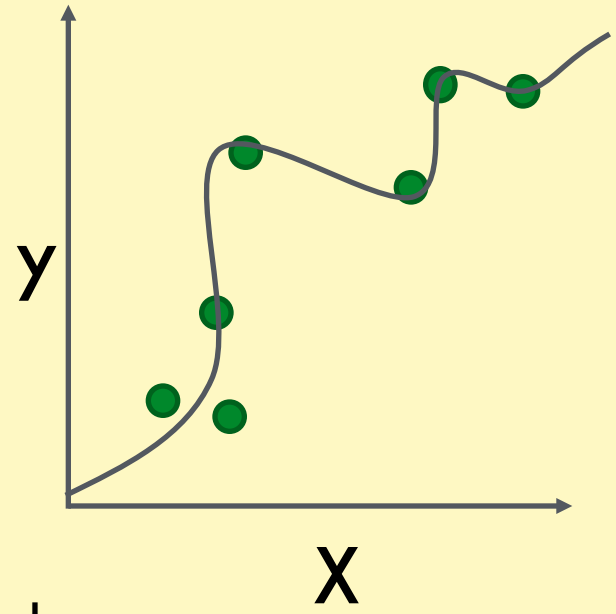
- Random Forest
- Distributed Trees
- Gradient Boosting Machine
- Stacking / Super Learner

Deep Neural Networks

- MLP

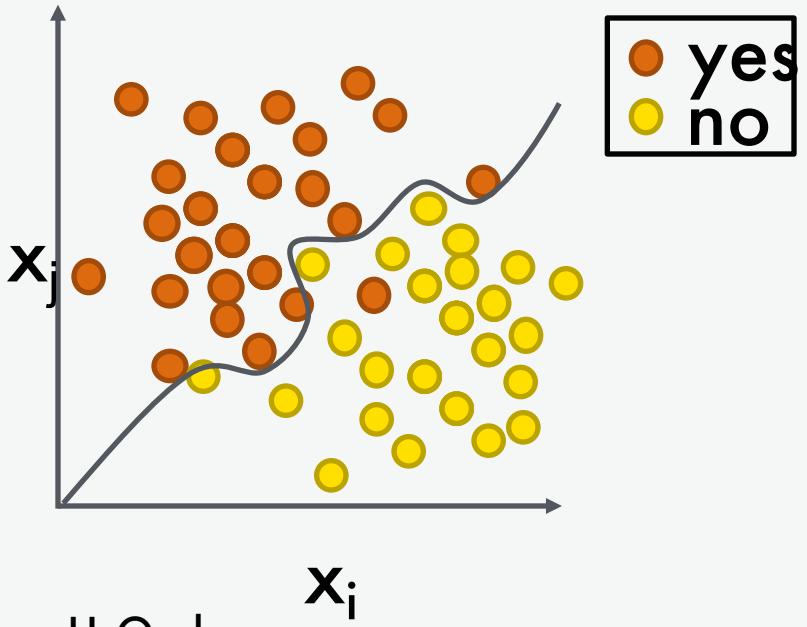
Supervised Learning Algorithms

Regression:
How much will a claim cost?



H_2O algos:
Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:
Will a physician commit fraud? Yes or No



H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

H2O's Unsupervised Algorithms

Unsupervised Algorithms

Clustering

- K-Means (Auto-*K*)

Dimension Reduction

- Principal Component Analysis
- Generalized Low Rank Models

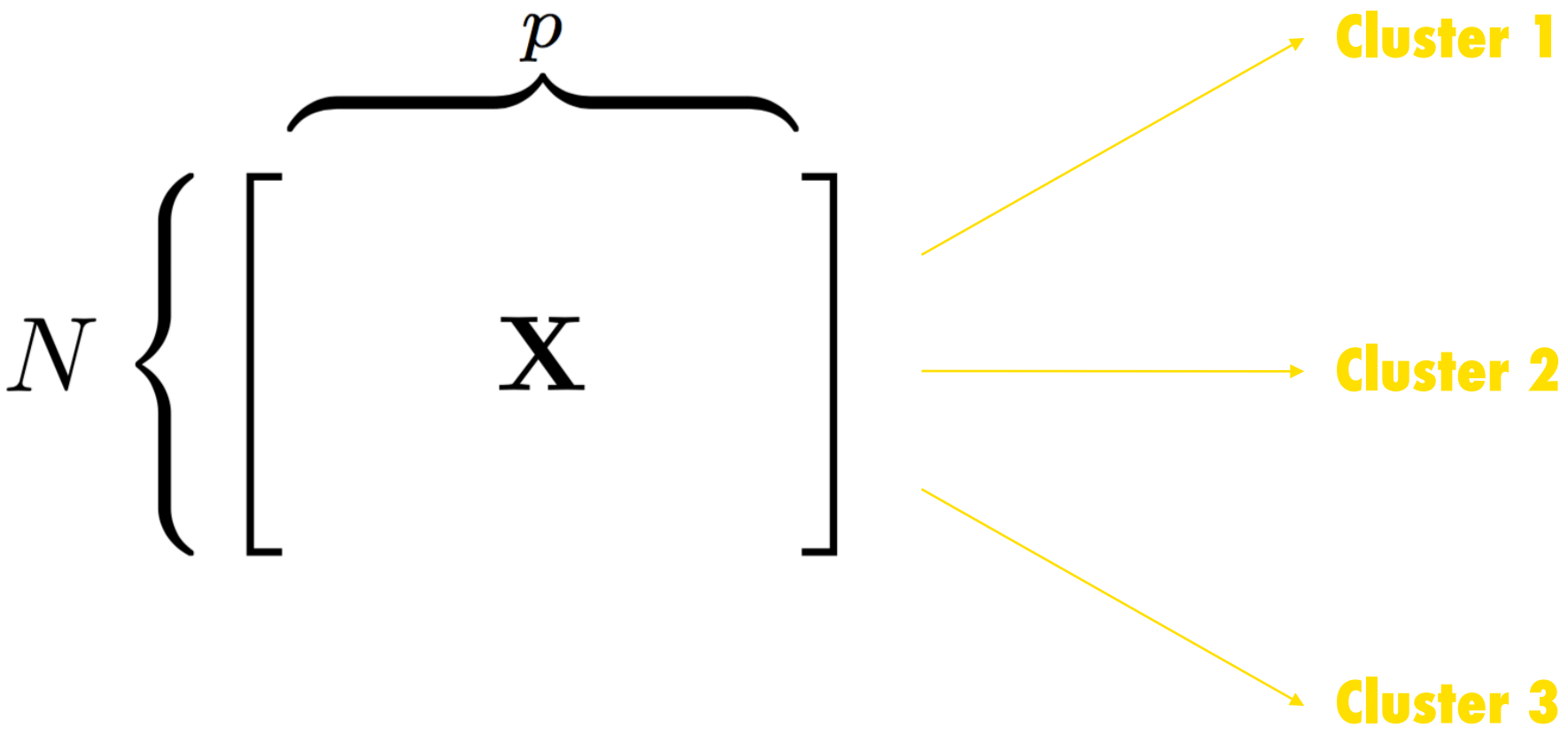
Word Embedding

- Word2Vec

Time Series

- iSAX

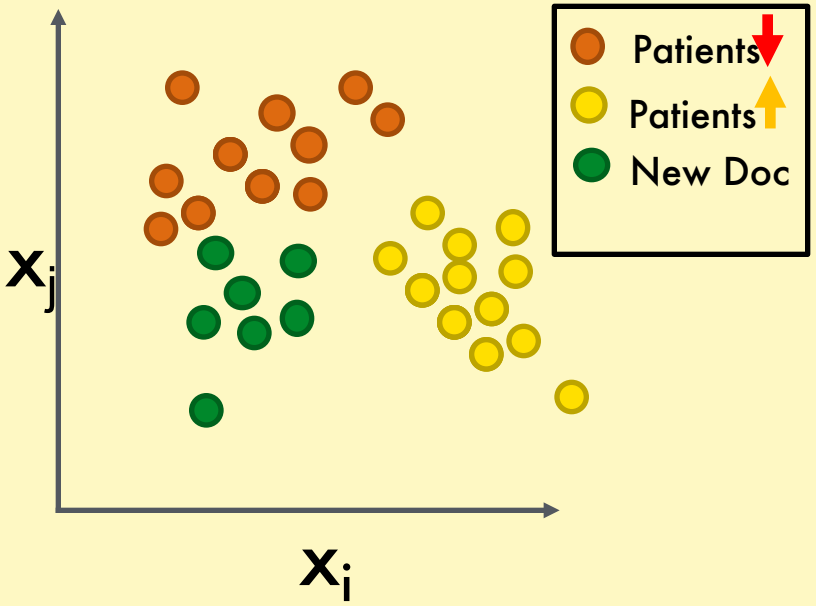
How to Group Customer Claims?



Unsupervised Learning Algorithms

Clustering:

Grouping rows – e.g. creating groups of similar physicians

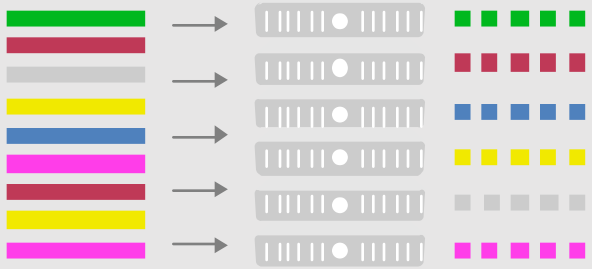


H_2O algos:

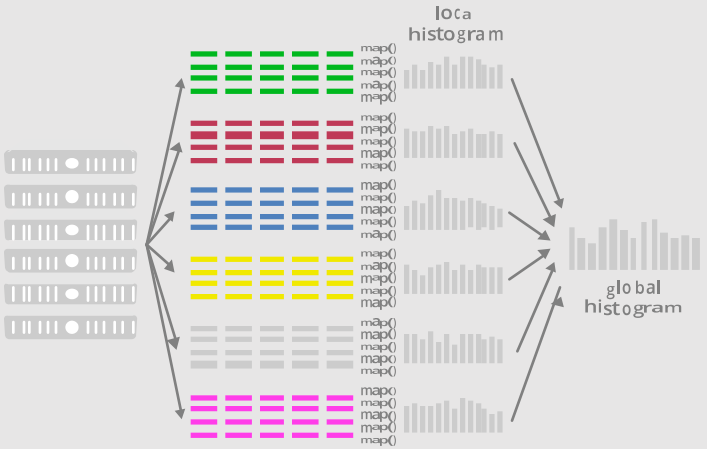
K-Means

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable Distributed Histogram Calculation for GBM

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**