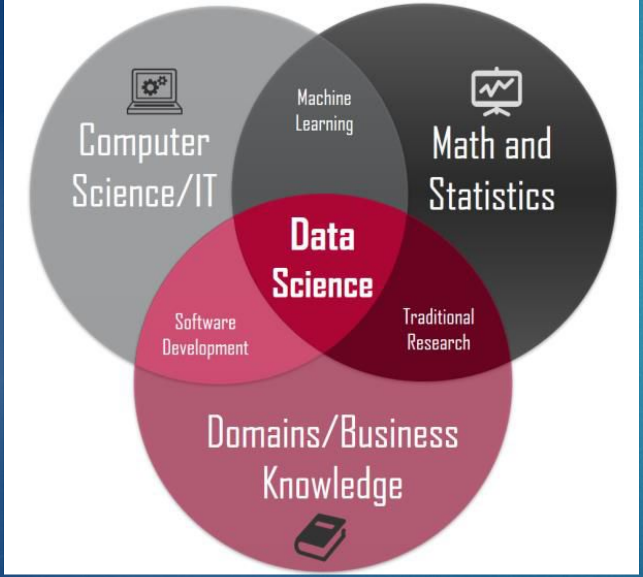
MS in data science a joint work of the Statistics and Computer Science departments

Cristina Tortora, Bem Cayco, and Guangliang Chen

SAN JOSE STATE UNIVERSITY

Definition



According to Josh Wills, of Cloudera, Google, and Slack fame, a Data Scientist is

“A Person who is better at statistics than any software engineer and better at
software engineering than any statistician”.

Need of a new program

- Market demand for Data Scientists
  - number of job listings for core Data Science and Analytics (DSA) is growing
  - Currently have an average annual salary of $94,576 and are projected to see demand spike by 28% (from 48,347 to 61,799)
  - DSA jobs remain open for an average five days longer than the market average.
- Lack of rigorous, balanced data science programs in today's universities
- Student interest
- Equity and affordability

Data science programs at SJSU

MS Data Science (Math & Stats, and CS)

MS Statistics, Specialization in Machine Learning

MS Data Analytics

MS Software Engineering, Specialization in Data Science

MS Software Engineering, Specialization in Cybersecurity

MS Computer Engineering, Specialization in Data Science

MS Artificial Intelligence (Computer Engineering)

MS Informatics (School of Information)

MS Bioinformatics (Bio, and CS)

Program highlights

- First launched in Fall 2020
- **Jointly held between Math & Computer Science**
- Balanced training in theory, computing and data analysis
- Rigorous curriculum
- Affordable tuition (we are a state-funded degree program)
- Dedicated faculty (Math and CS combined)
- Proximity to Silicon Valley

Program Learning Outcomes (PLOs)

Upon successful completion of the MS Data Science program, students will be able to

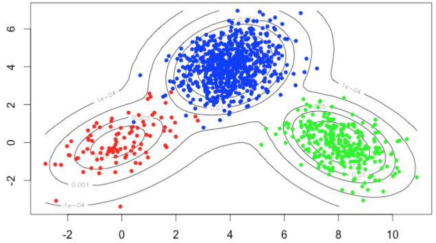
- **PLO 1** Apply computer science knowledge and tools to assist in performing data science tasks
- **PLO 2** Summarize and evaluate statistical and machine learning concepts, models and techniques
- **PLO 3** Integrate multidisciplinary knowledge and software to tackle challenging, complex data science tasks
- **PLO 4** Communicate effectively, both orally and in writing, data science concepts, algorithms, and results to a broad audience.
- **PLO 5** Identify ways in which data scientists can contribute to the cultural and economic well-beings of diverse societies in local, national and global scopes

New classes developed for the MS

- Math 250 Mathematical Methods for Data Visualization
- Math 251 Statistical and Machine Learning Classification
- Math 252 Cluster analysis

Math 252 Cluster analysis

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- Applied in several different fields: biology, marketing, psychology, etc.



The **goal** of this course is to give the students a strong theoretical base and applied skill on cluster analysis to be able to work on real problems.

Prerequisites of the course:

- Math 32 multivariable calculus
- Math 39 linear algebra
- Math 163 probability theory
- Math 167 R Statistical programming with R or CS 122 Advanced Programming with Python

Math 252 Cluster analysis

Final project: International Federation of Classification Society data competition

- In 2017

  1 group of students
  - Won the data competition

  3 groups of students
  - presented their work at IFCS conference in Japan
  - Published the results in Archives of data science series B

ifcs

IFCS BENCHMARK DATA ANALYSIS
CHALLENGE WINNER 2017
FENGMEI LIU, SUCHARA GUPTA,
CRISTINA TORTORA

- In 2019
  - 3 groups of students presented their work at IFCS conference in Greece
  - 2 groups published the results In Studies in Data Analysis and Rationality in a Complex World

Math 250: Mathematical Methods for Data Visualization

Rational of this course

- ***Central topic***: dimension reduction (specially, feature transformation)
- ***Main motivation and application***: Data visualization (with online large, complex data sets)
- ***Supporting tools:***
  - *Mathematical*: advanced linear algebra, such as positive definite matrices, Rayleigh quotient, SVD, matrix norm and pseudoinverse, low-rank approximation, and constrained optimization
  - *Applied*: Matrix computing and 3D data plotting in MATLAB

It prepares students with mathematical, computing, and data foundations for machine learning.

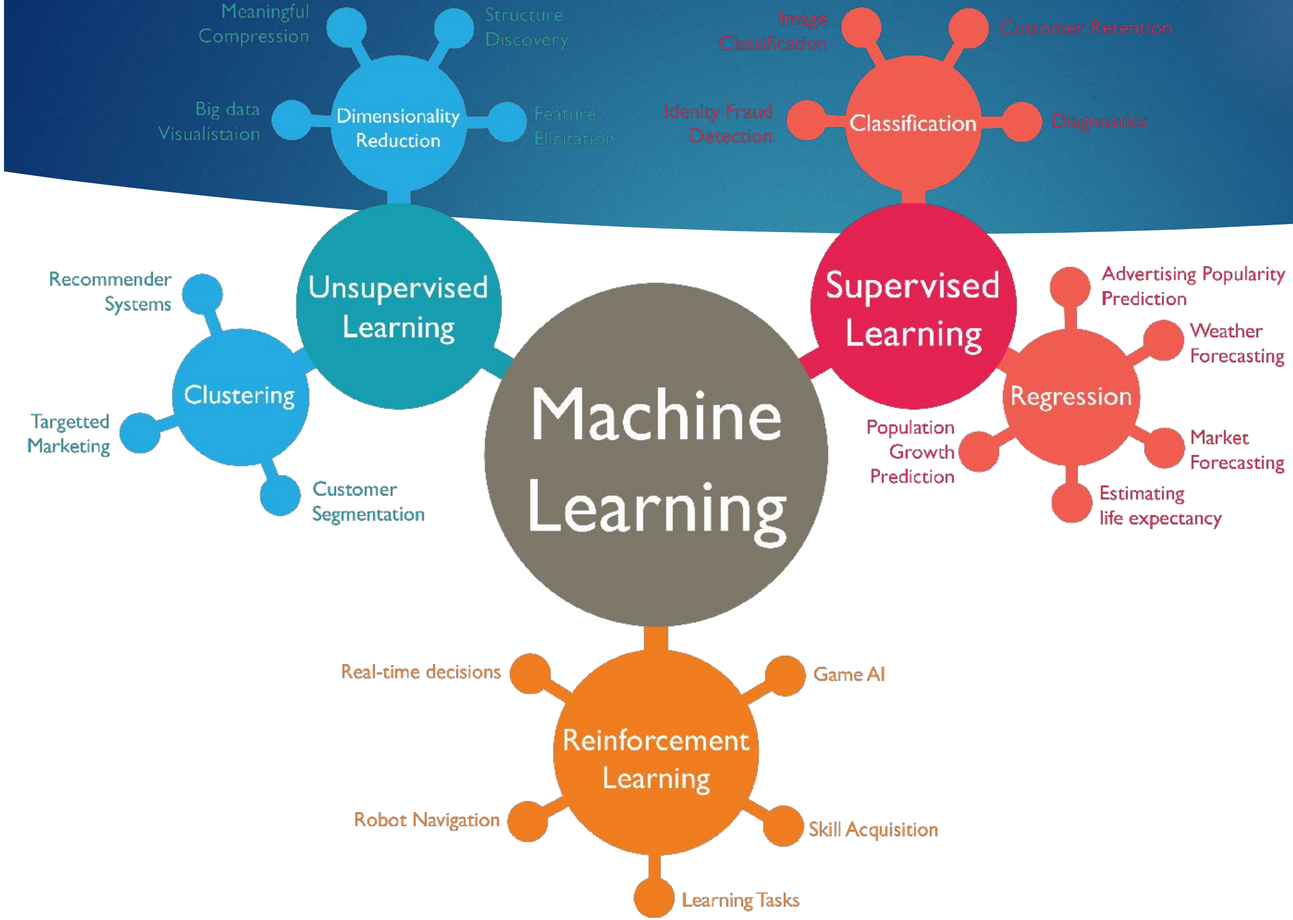
Prerequisites of the course:

- Math 32 multivariable calculus (method of Lagrange multipliers)
- Math 39 linear algebra (strong linear algebra knowledge and skills are crucial)
- Math 163 probability theory (mathematical maturity)

Math 250: Mathematical Methods for Data Visualization

Dimension reduction methods covered in this course:

- **Linear projection methods:**
  - PCA (for unlabeled data),
  - LDA (for labeled data)
- **Nonlinear embedding methods:**
  - Multidimensional scaling
  - ISOmap
  - LLE
  - Laplacian eigenmaps
  - Nonnegative matrix factorization



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

Math 251: Statistical and Machine Learning Classification

Rational of this course

Teach a machine learning topic (classification)

… through an application (digits recognition)

… using a benchmark dataset (MNIST Handwritten Digits)

… assisted by a technical computing language (MATLAB or Python)

… enhanced by a hands-on project (data science competition)

It aims to provide a balanced training in machine learning theory, computing, & project experience.

Prerequisites of the course:

- Math 250 Mathematical Methods for Data Visualization* (advanced linear algebra, optimization, dimensionality reduction, data plotting and visualization, and coding)
- Math 164 Mathematical Statistics (statistics, MLE, Bayesian inference)

Math 251: Statistical and Machine Learning Classification

Classifiers covered in this course:

- **Instance-based classifiers**: kNN and its variants
- **Bayes classifiers**: LDA/QDA, Naive Bayes
- **Logistic regression:** binary/multiclass, multinomial
- **Support vector machine:** binary/multiclass, kernel SVM
- **Ensemble methods**: decision trees, bagging, random forest, and boosting
- **Neural networks and deep learning**: ANN, and CNN

For each method, I cover the underlying mathematics/statistics, computing, and practical issues (such as memory/speed, dimension reduction, and parameter estimation)

Full program (30+6=36 units)

Catalog number	Title
CS 156	Introduction to Artificial Intelligence
CS 157A	Introduction to Database Management Systems
CS 200W	Graduate Technical Writing
CS 274	Topics in Web Intelligence
MATH 164	Mathematical Statistics
MATH 261A	Regression Theory and Methods
MATH 250	Mathematical methods for data Visualization
MATH 252	Cluster Analysis
CS 271 or MATH 251	Machine Learning
Elective	Subject to approval by program coordinator

Culminating experience:

Must complete one of the following two 6-unit options:

- **Plan A (thesis):**
  - Math 297A and 299, or
  - CS 297 and 299
- **Plan B (project):**
  - Math 297A and 298, or
  - CS 297 and 298

Program prerequisites

- *Math 32 Multivariable Calculus (*with a grade of B or better)
- *Math 39 Linear Algebra
- *Math 161A Applied Probability & Stats I
- Math 163 Probability Theory
- *CS 146 Data Structures
- CS 151 Object-oriented Programming

For more information, see the admissions page at

https://www.sjsu.edu/science/special-programs/ms-data-science.php

Target student populations

Students with a Bachelor's degree in the sciences or engineering (e.g., **applied math, statistics, computer science, and software engineering**) from a regionally accredited institution with a minimum GPA of 3.0

Examples of ideal applicants:

- Dual major in math/statistics and CS/software engineering
- Math/Statistics major with a CS/software engineering minor
- CS/software engineering major with a math minor

Questions?

Thank you for your time and attention!

Please encourage your students to

- **apply to our program**, and
- **send any admission-related questions** to
  - Program inbox: *sci-ms-datascience@sjsu.edu*, **or**
  - My SJSU email: *guangliang.chen@sjsu.edu*