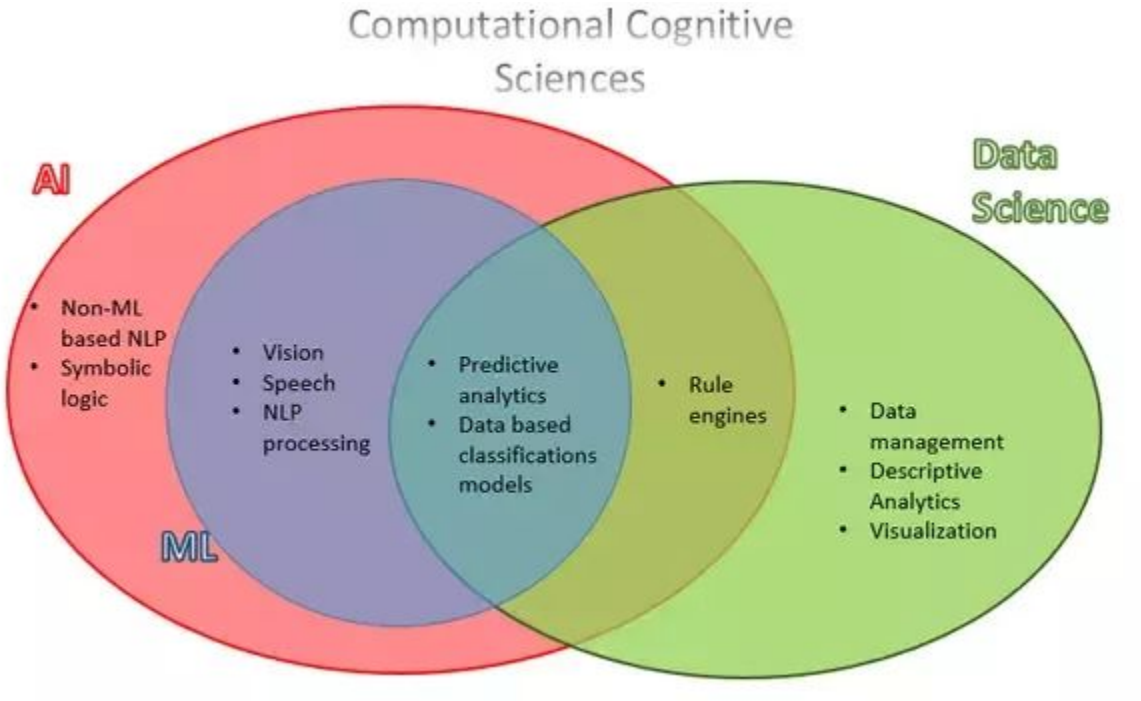
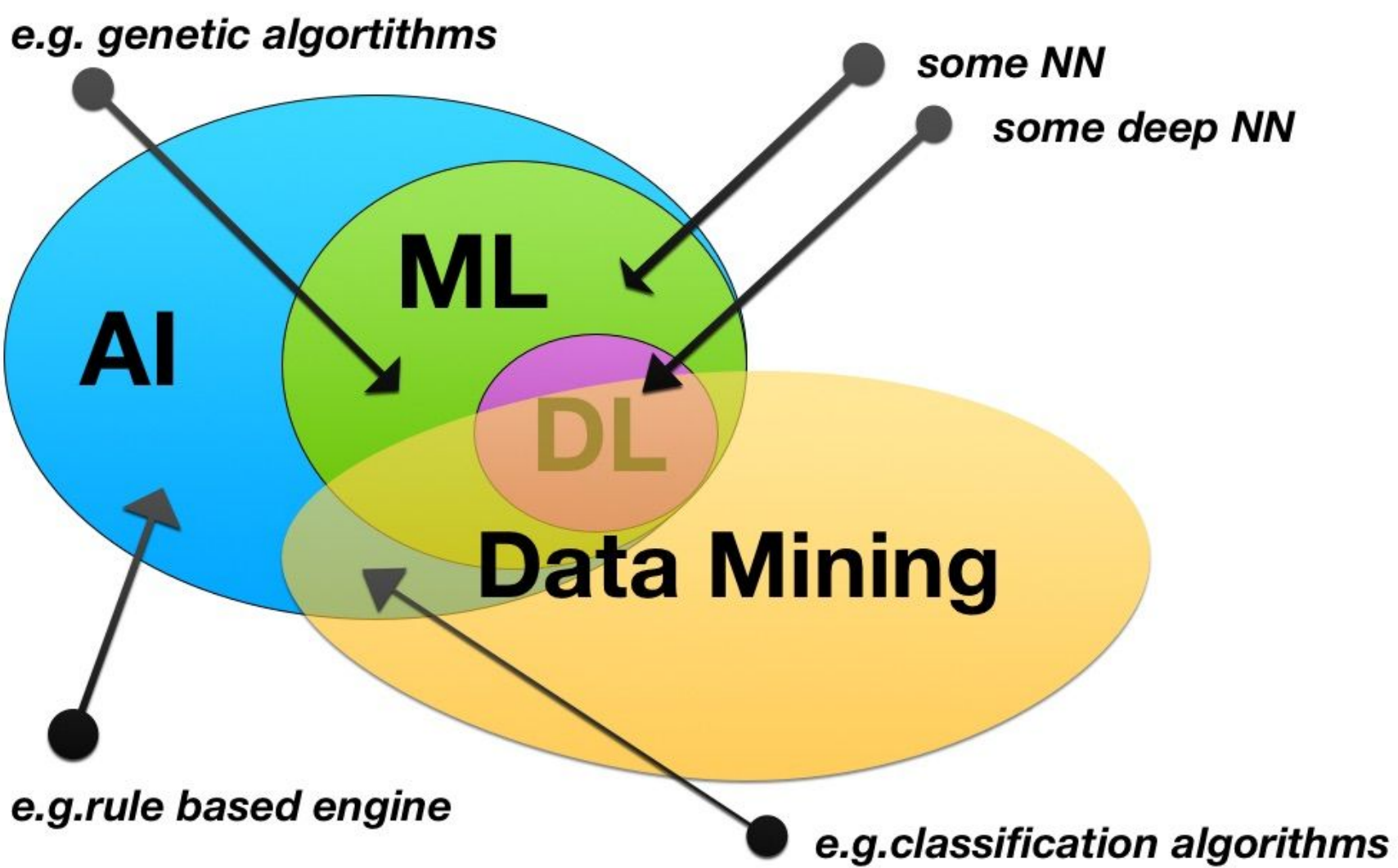
ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Основные понятия машинного обучения и анализа данных

к.ф.-м.н., заместитель руководителя по научной работе, доцент ДАДиМО
Корчагин С.А.
SAKorchagin@fa.ru


Computational Cognitive Sciences
AI
Data Science
ML
Non-ML based NLP
Symbolic logic
Vision
Speech
NLP processing
Predictive analytics
Data based classifications models
Rule engines
Data management
Descriptive Analytics
Visualization


e.g. genetic algortithms
some NN
some deep NN
AI
ML
DL
Data Mining
e.g.rule based engine
e.g.classification algorithms

Говорят, что компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E.

Области применения ML

Распознавание речи.

Компьютерное зрение.

Компьютерная лингвистика и обработка естественных языков.

Медицинская диагностика.

Техническая диагностика.

Рубрикация текстов.

Интеллектуальные игры.

Главные вопросы ML

Какое количество и какой информации необходимо для обучения?

Какие данные лучше выбирать для обучения и почему?

Какой алгоритм решает поставленную задачу наилучшим образом?

Как свести какую-либо задач обучения к аппроксимации или оптимизации некоторой функции?

Основные понятия

Признак (feature)

Объект (object)

Чистые данные (tidy data)

Набор данных (dataset)

Модель

Шкала

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Определение чистых данных (tidy data)

Каждая переменная соответствует колонке

Каждое измерение соответствует строке

Каждая таблица\файл содержит данные об одном виде наблюдений\экспериментов

Обзор данных (data exploration)

Отсутствующие данные

Значения вне разумного диапазона

Ошибки в единицах измерения (шкалах)

Ошибки в подписях переменных (колонок)

Ошибки в классах переменных

Предварительная обработка данных

Создание новых переменных

Слияние наборов данных

Трансформация переменных

Удаление несогласованных данных

Этапы анализа данных

Определить вопрос
Определение идеального набора данных
Определение доступного набора данных
Получение данных
Очистка данных
Исследовательский анализ данных
Статистическое моделирование
Интерпретация результатов
Проверка результатов
Описание результатов
Создание воспроизводимого кода

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Этапы процесса машинного обучения

Получение данных

Трансформация данных

Очистка данных

Визуализация данных

EDA

Выбор модели

Обучение модели

Верификация результата

Основные типы шкал

Бинарные (Пол, наличие боли в спине, в сознании ли пациент).

Номинальные (Тип боли: колющая, режущая, ноющая).

Порядковые (Общее состояние больного: удовлетворительное, средней тяжести, тяжелое, крайне тяжелое).

Количественные (Температура тела, пульс, артериальное давление).

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Основные форматы хранения наборов данных

CSV

XML

JSON

XLSX

DB

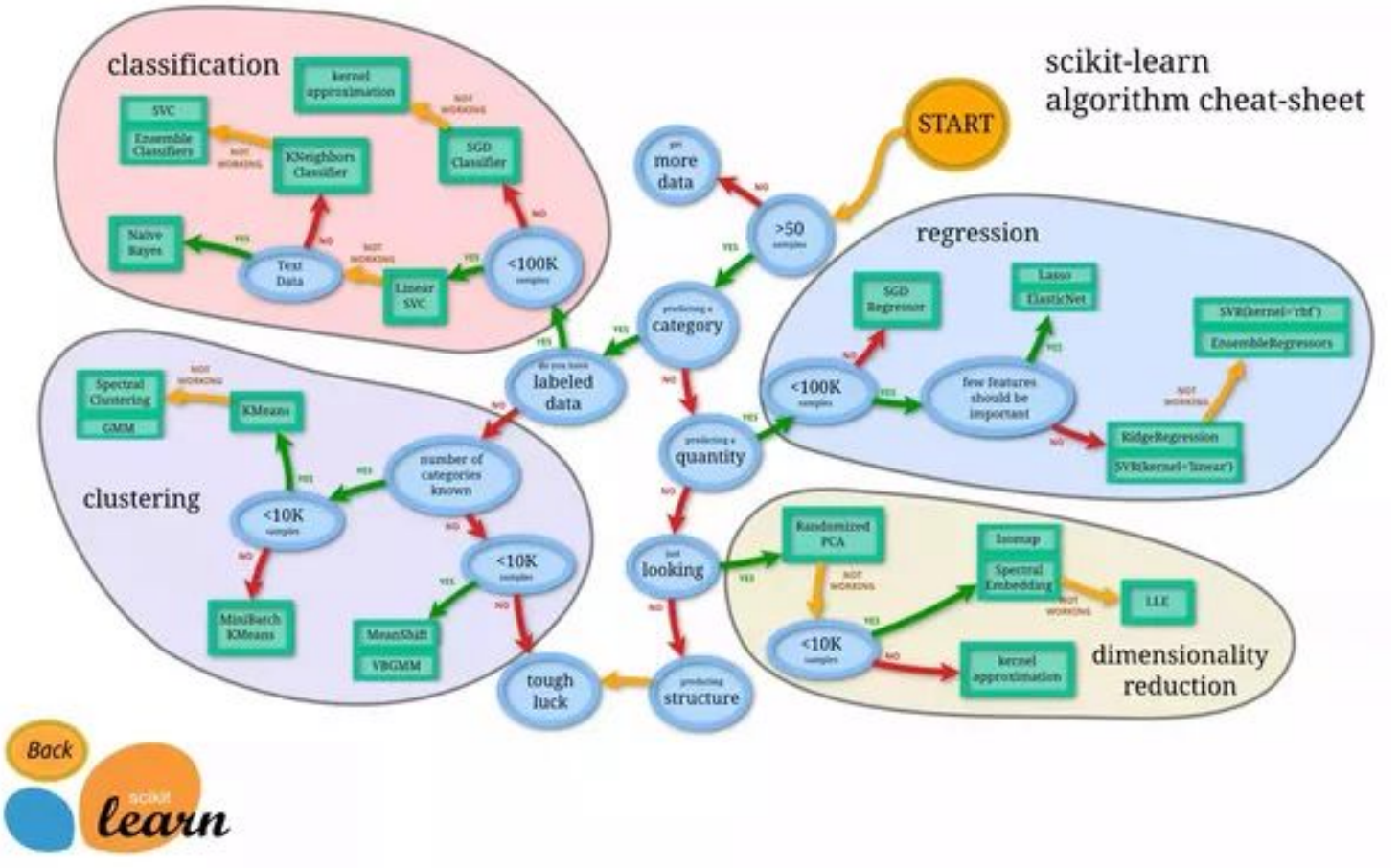
Главные задачи машинного обучения

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

- Обучение с учителем (supervised learning)
  - Регрессия
  - Классификация
- Обучение без учителя (unsupervised learning)
  - Понижение размерности
  - Обнаружение аномалий
  - Кластеризация
- Рекомендательные системы
- Обучение с подкреплением (reinforecement learning)

Основой машинного обучения является оптимизация некоторой функции ошибки


learn
matplotlib
pandas
PyMC
SciPy
NumPy
ython
SymPy
IP[y]:
IPython
python
jupyter


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K
more data
>50
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K
few features should be important
RidgeRegression
SVR(kernel='linear')
predicting a category
labeled data
predicting a quantity
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K
MiniBatch KMeans
<10K
MeanShift
VBGMM
looking
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit learn