



PhD student | Computer Science
AutoML | Meta-learning | RML
Research & Business Experience

Data Scientist (5 years)
Research assistant - MI2 Lab (3 years)
Coordinator of Case Studies 2020/2021



MS Teams
katarzyna.woznica.dokt@pw.edu.pl

DATA SCIENCE

DATA SCIENCE EVERYWHERE





Public speaking



Database

Statistics

Machine learning

Visualization

Data mining

Coding

Communication



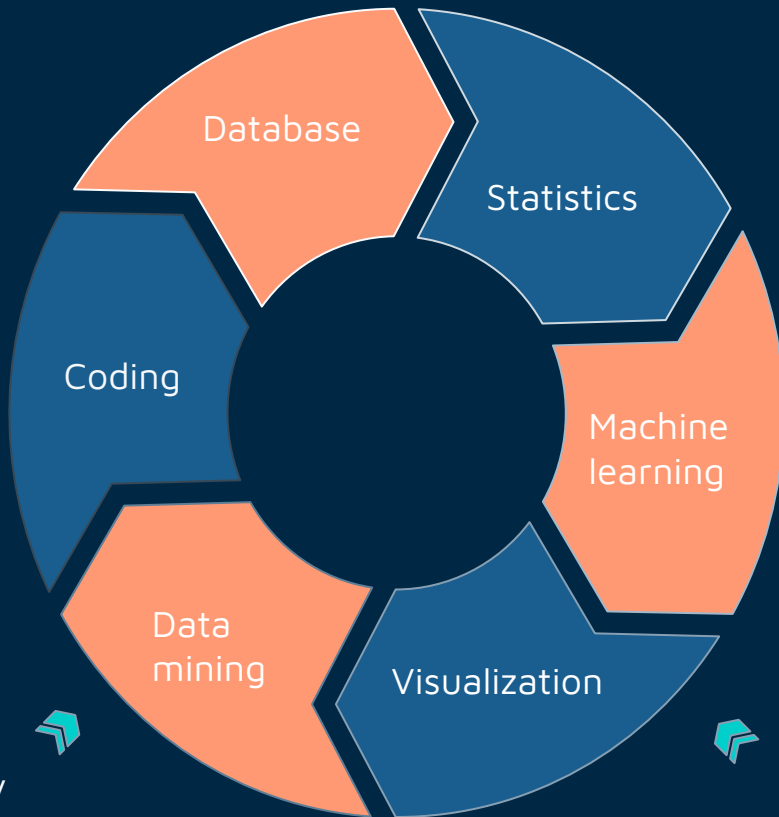
Domain Expertise

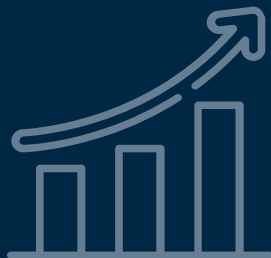
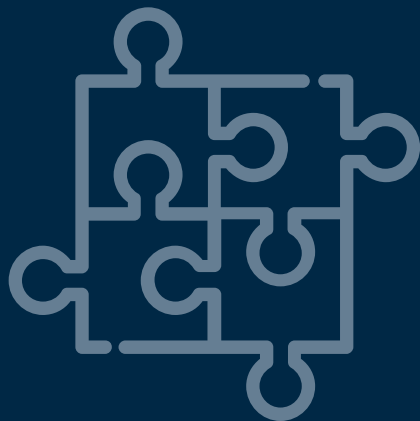


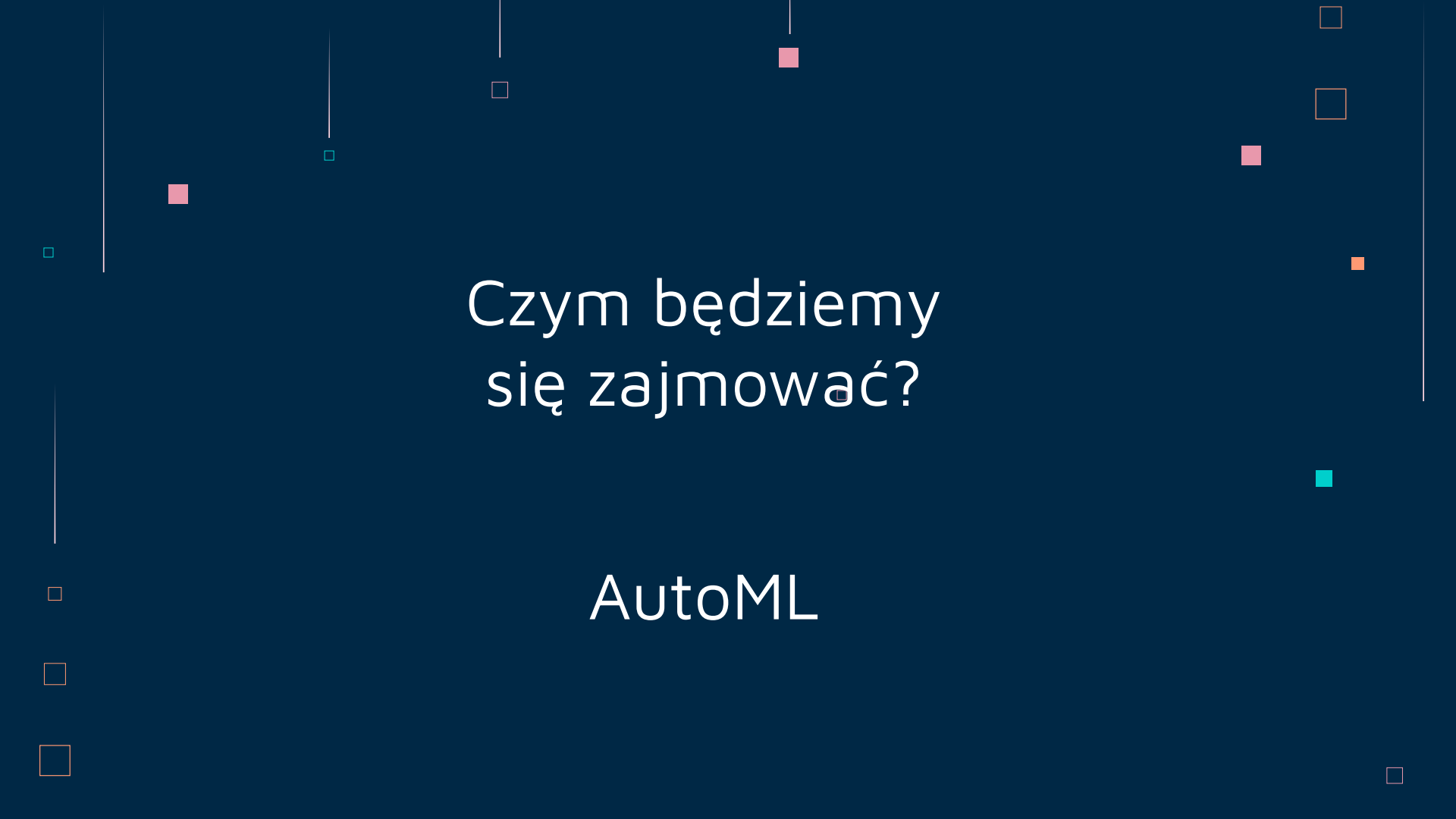
Real problems



Creativity



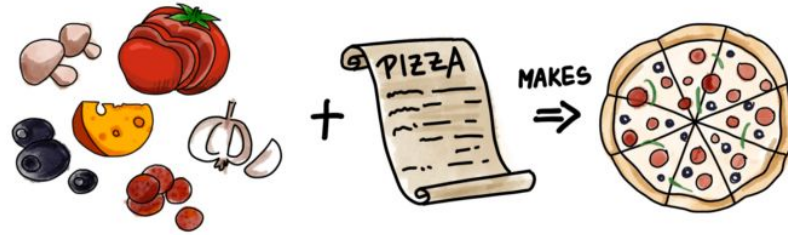




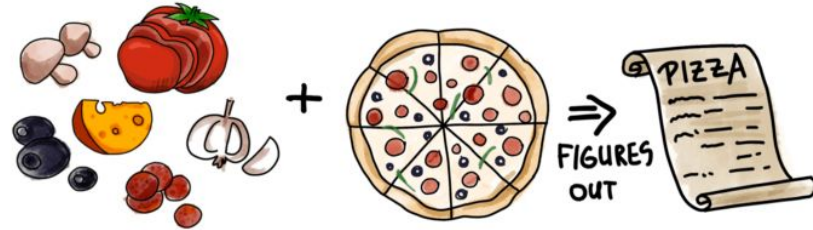
Czym będziemy
się zajmować?

AutoML

TRADITIONAL / PROGRAMMING



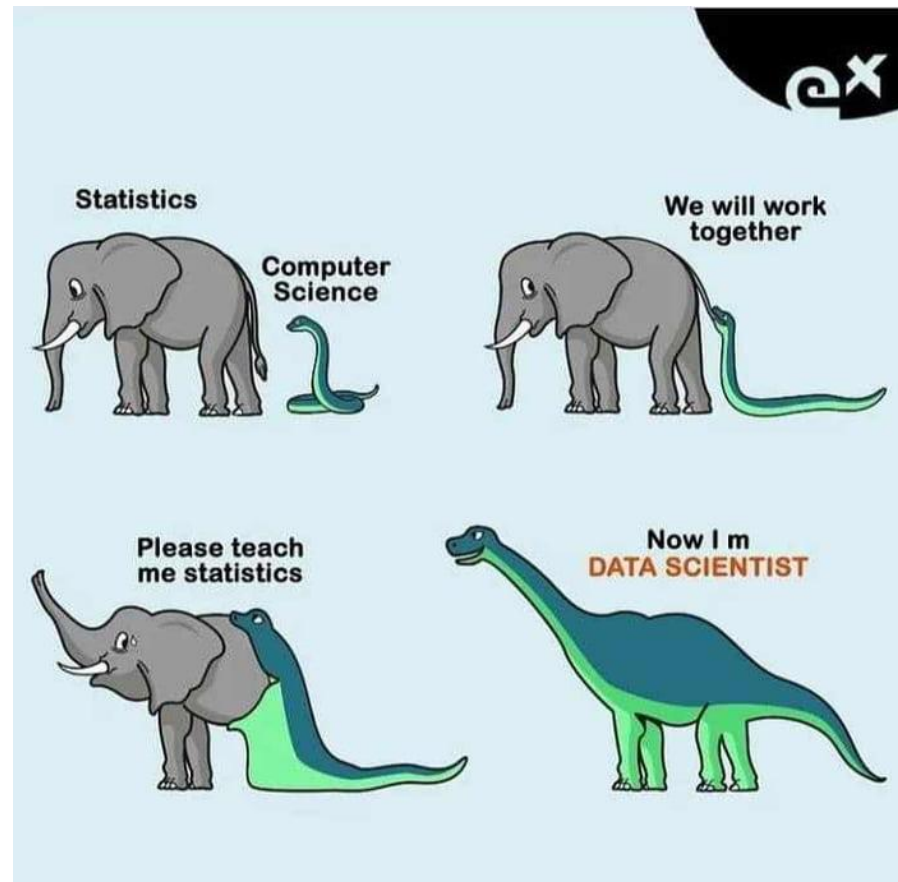
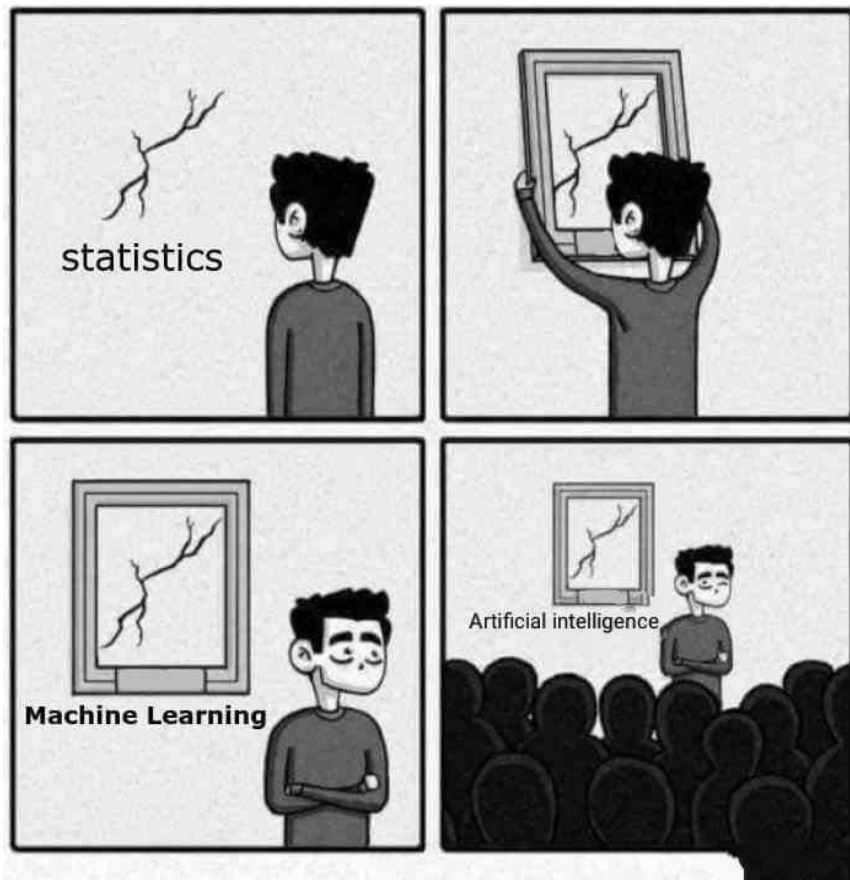
MACHINE LEARNING ALGORITHM

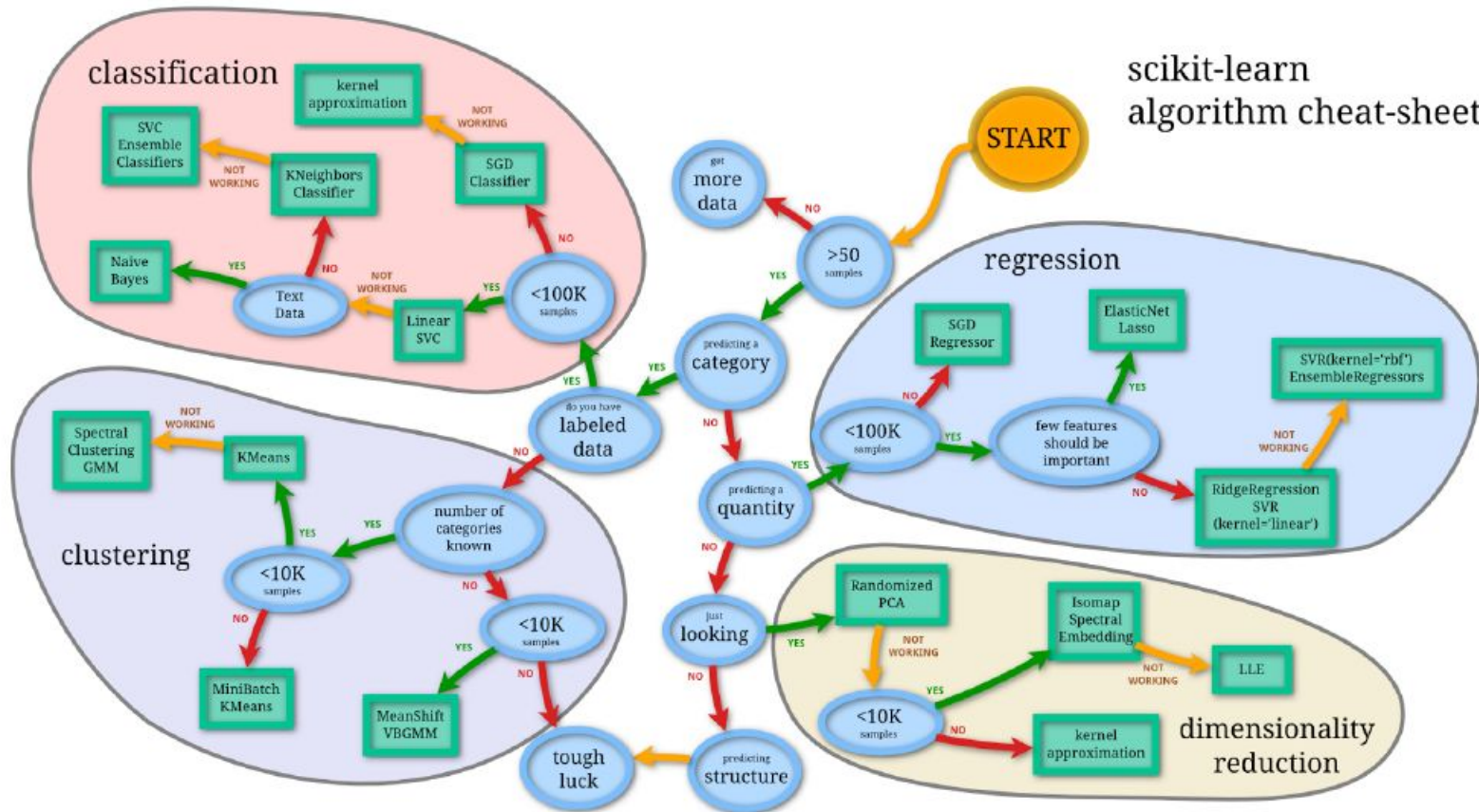


@luminousmen.com

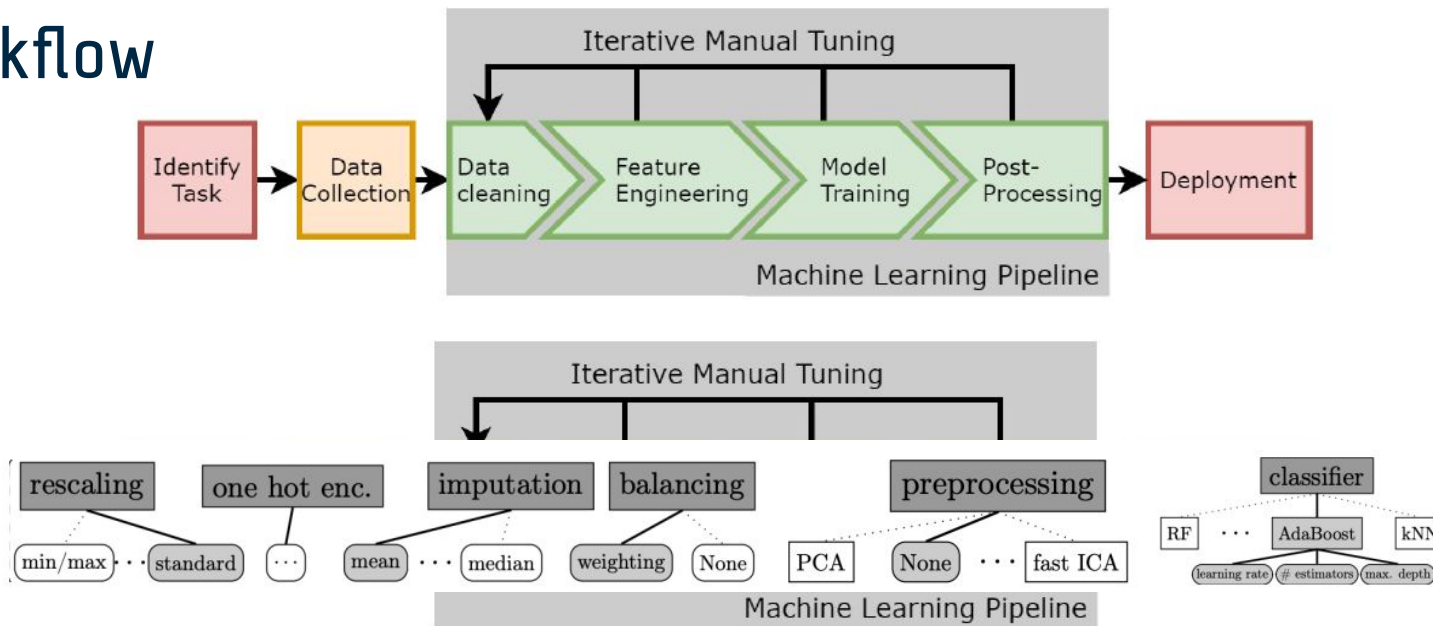
Machine learning is the science of getting computers to act without being explicitly programmed.

Andrew Ng



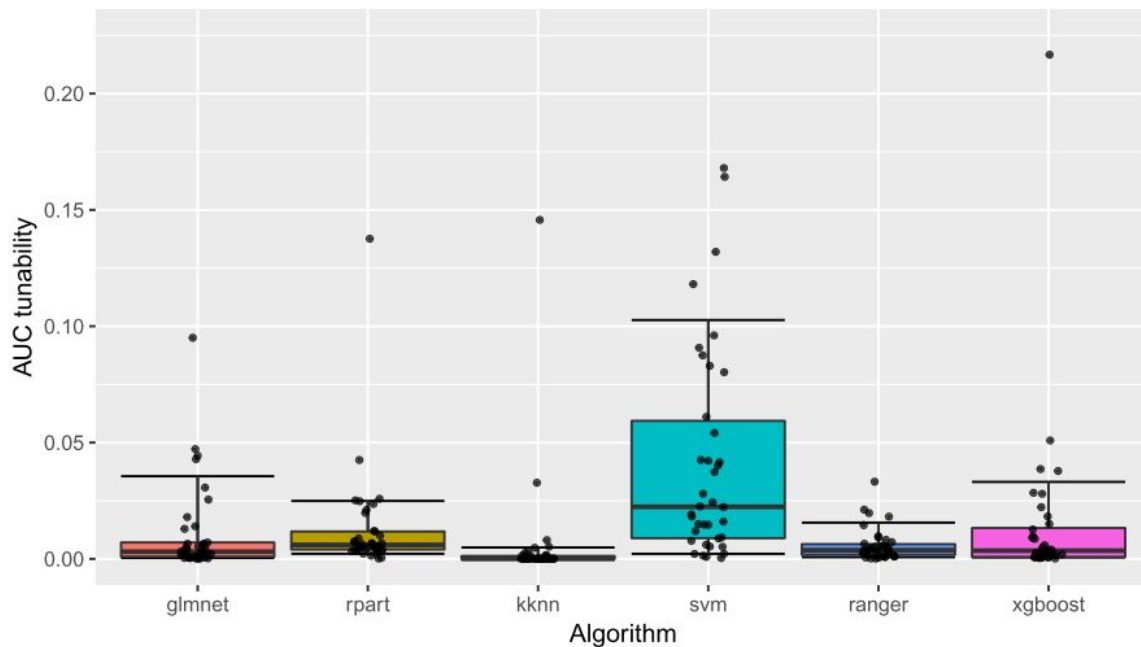
scikit-learn
algorithm cheat-sheet

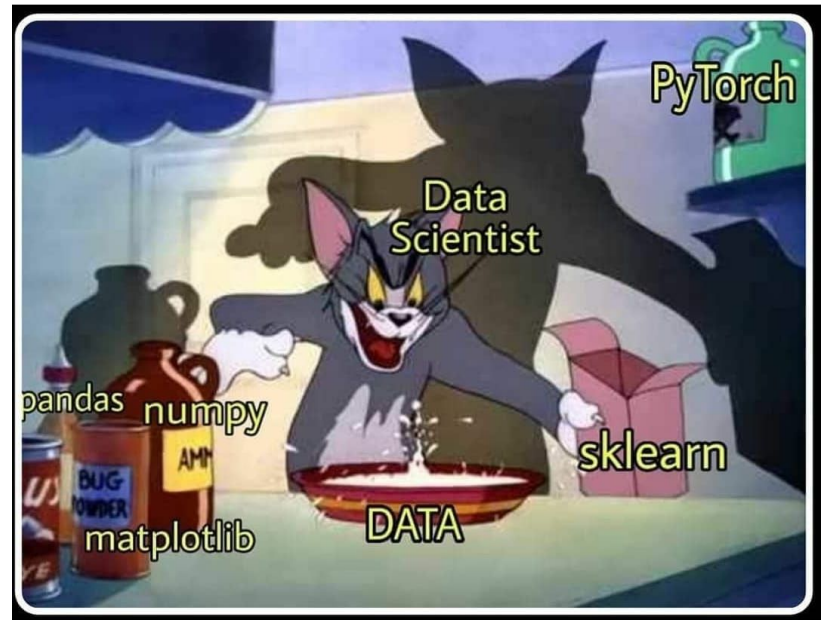
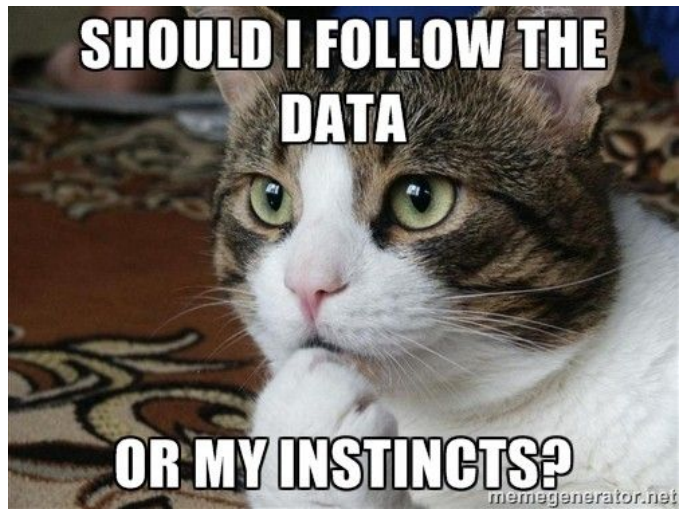
ML workflow



<https://learn.ki-campus.org/courses/automl-luh2021/items/6os022oUVHsYcvb29yGFjY>

Wybór preprocessingu, algorytmów i ich konfiguracji jest kluczowe dla mocy predykcyjnej wytrenowanych modeli ML





- Podstawy ML są łatwe do zrozumienia
- Osiągnięcie state-of-the-art performance jest raczej trudne
- Decyzje dotyczące procesu trenowania modeli są nieintuicyjne i wymagają dużo wiedzy eksperckiej
 - są one powtarzalne i podatne na błędy
- Na rynku pracy brakuje ekspertów ML
- Rozwój modeli ML wymaga czasu

SHOULD I FOLLOW THE
DATA

ML model returns above 99% accuracy on real-world data

Junior Data Scientist

Senior Data Scientist

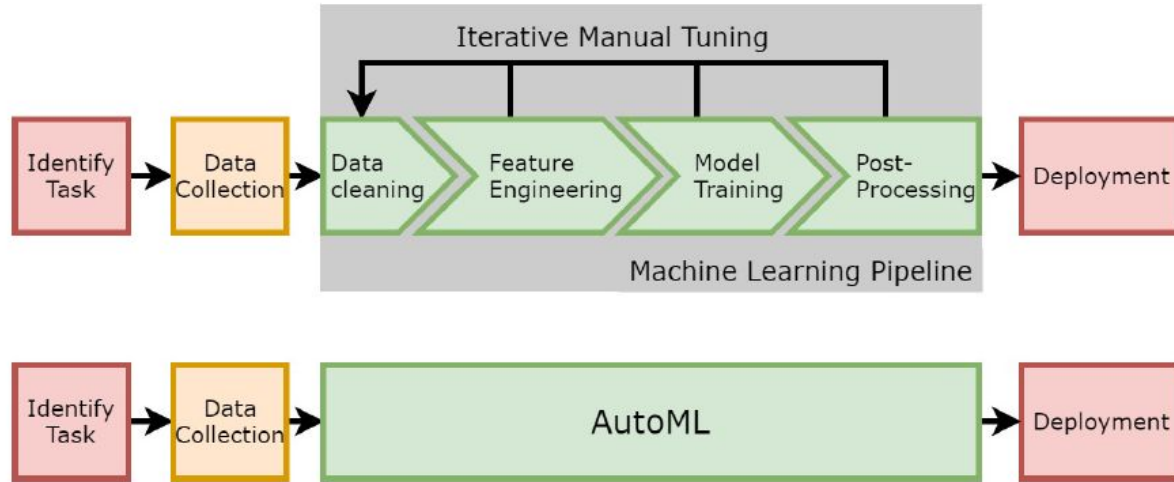


PyTorch

learn

matplotlib

DATA



- support ML users
- improve the efficiency of developing new ML applications
- reduce the required ML-expertise
- might achieve better performance than developers w/o AutoML

STARTED USING AUTO-ML

NOW I HAVE ALL DAY TO CREATE AI MEMES

imgflip.com

DATAFTTER.COM

Dostępne frameworki dla danych tabelarycznych

- [Autosklearn](#)
- [AutoWEKA](#)
- [Autogulon](#)
- [AutoKeras](#)
- [AutoPytorch](#)
- [GAMA](#)
- [FLAML](#)
- [Hyperopt](#)
- [naiveAutoML](#)
- [ML-Plan](#)
- [TPOT](#)
- [AutoPrognosis](#)
- [H2O AutoML](#)
- [OBOE](#)
- [LightAutoML](#)
- [AutoXGBoost](#)

Tool	Platform	Input data sources		Data pre-processing	Data types detected					Feature engineering				ML Tasks		Model selection and Hyperparameter optimization					Quick start / early stop			Model evaluation / Result analysis/ Visualization		
		Spreadsheet datasets	Image, text		Numerical	Categorical	Datetime	Time-series	Other (Hierarchical types) (7*)	Datetime, categorical processing	Imbalance, missing values	Feature selection, reduction	Advanced feature extraction (8*)	Supervised learning (9*)	Unsupervised learning (10*)	Ensemble	Genetic algorithm	Random search	Bayesian search	Neural architecture search	Quick finding of starting model	Allow maximum limit search time	Restrict time consuming combination of components	Model dashboard	Feature importance	Model explainability and interpretation, and reason code (11*)
TransmogrifAI	Apache Spark	Y	N	Y(1*)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N	N			Y	Y	
H2O-AutoML	AWS, GCP, Azure	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y	N	Y	N	Y	N	N	N	Y	Y	Y	Y	Y
Darwin (+)	GCP	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y	Y
DataRobot (+)	AWS, GCP, Azure	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y(12*)	Y		Y	Y	Y
Google AutoML (+)	Google Cloud	N	Y	Y						N	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Auto-sklearn		Y	N	N	N	N	N	N	N	Y(2*)	Y	Y	Y	Y	N	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y
MLjar (+)	MLJAR Cloud	Y(3*)	N	Y	Y	Y	N	N	N	Y	Y(4*)	N	N	Y(5*)	N	Y	N	Y	N	N	N	N	N	Y	Y	N
Auto_ml		Y	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N	N	N	N	Y	Y	Y
TPOT		Y	N	N	N	N	N	N	N	N	Y	N	Y	Y	N	Y	Y	N	N	N	N	Y	N	Y	Y	N
Auto-keras		Y	Y	N	N	N	N	N	N	N	Y	Y	N	Y	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y
Ludwig		Y	Y	Y(1*)	Y	Y	N	Y	Y	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	N	N	Y	Y	N
Auto-Weka		Y	N	N	Y	Y	N	N	N	N	Y	Y	N	Y	N	Y	N	Y	Y	N	N	Y	Y	Y	N	N
Azure ML (+)	Azure	Y	Y	Y(6*)	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N		Y	Y	Y	Y	
H2O-Driverless AI (+)	AWS, GCP, Azure	Y(3*)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y	Y	Y	Y

Organizacja pracy

- pierwszy model ML (PD1*)
- optymalizacja hiperparametrów (PD2*)
- studium literaturowe, prezentacja jednego z pakietów i pierwsza praca z nim (KM1)
- przygotowanie kodu automatyzującego użycie wybranego pakietu i przetestowanie go na danych z OpenML (KM2)
- ocena pakietu pod kątem wykorzystania meta-learningu (KM3)
- code review (PD3*)
- modyfikacja pipeline-u (dodanie meta-learningu, ensembling modeli) (KM4)

Benchmark



Wasz pipeline



nieznane dane



Ranking modeli

komponent
oceny z KM4

Rezultat

- Raport opisujący wybrany pakiet i zmodyfikowany pipeline
- Prezentacje rozwiązań na wykładzie w formie nagranej prezentacji