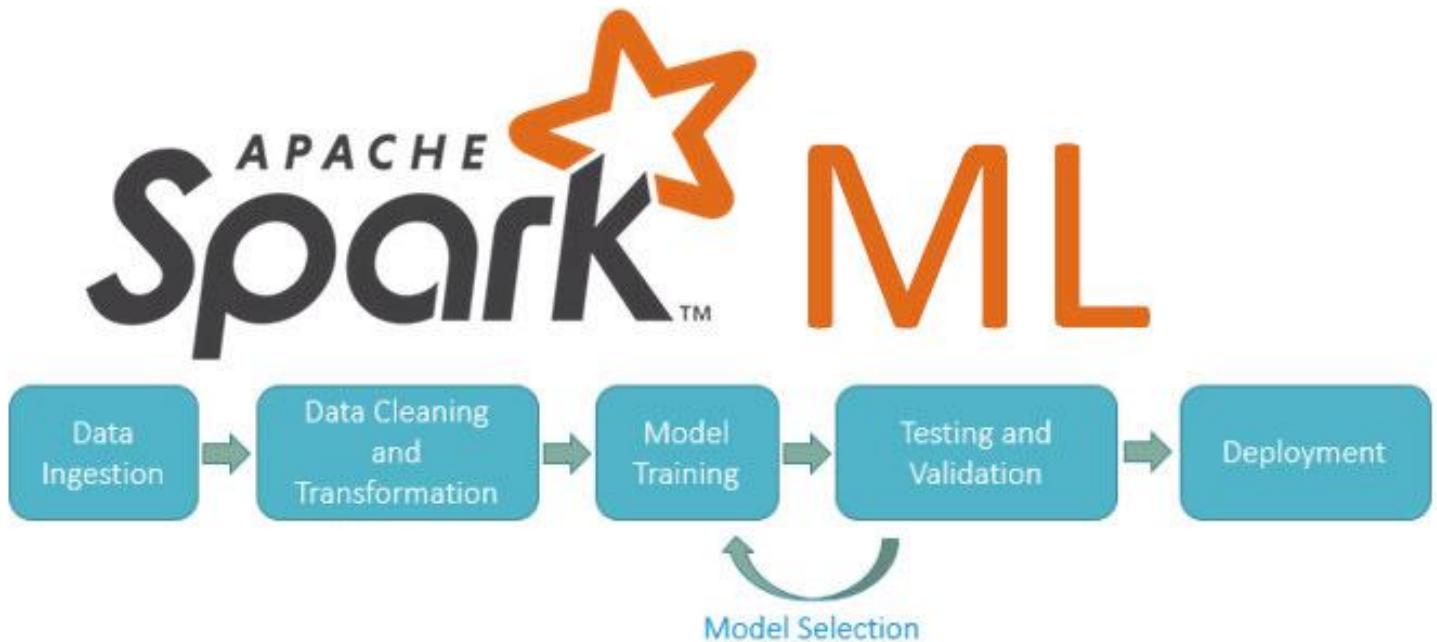
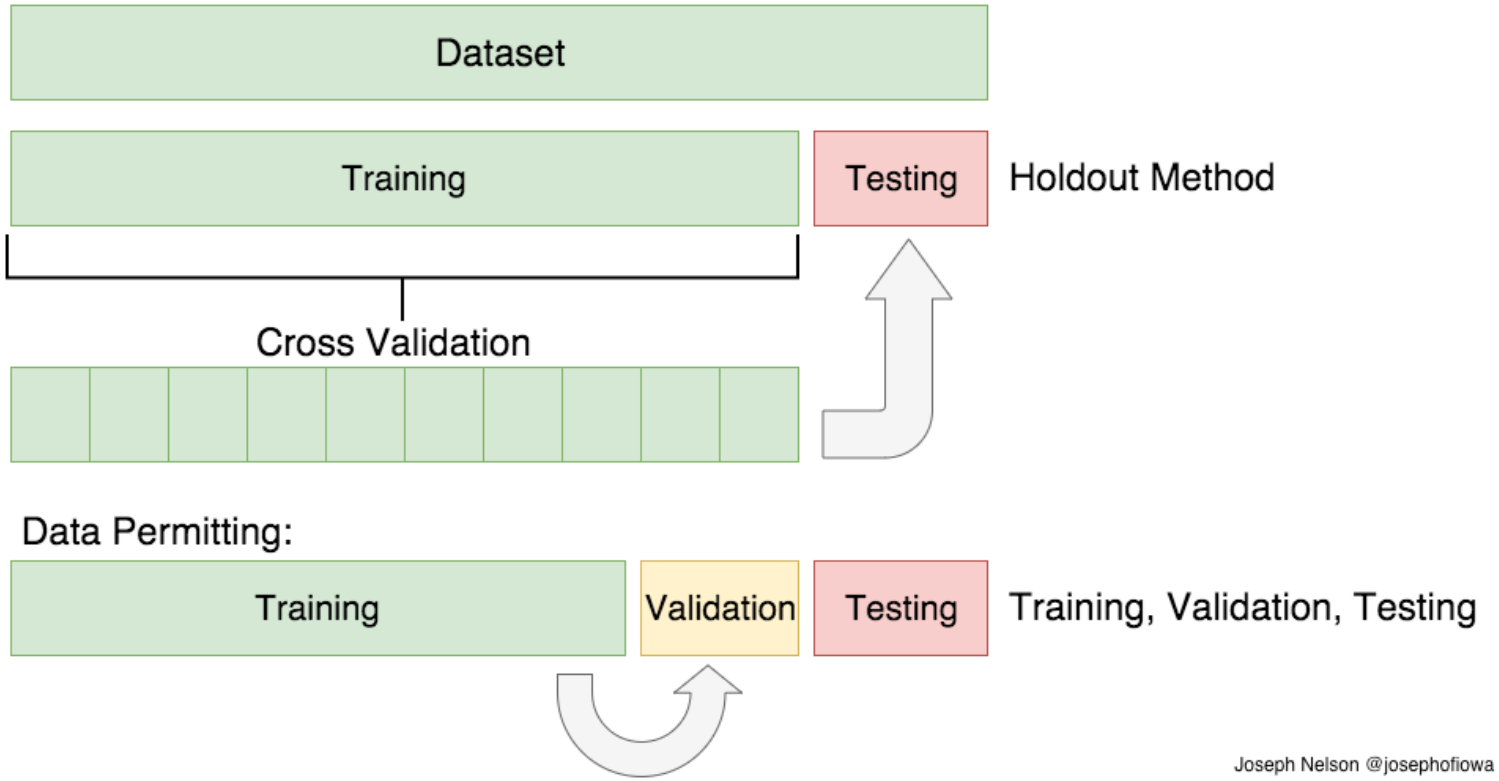


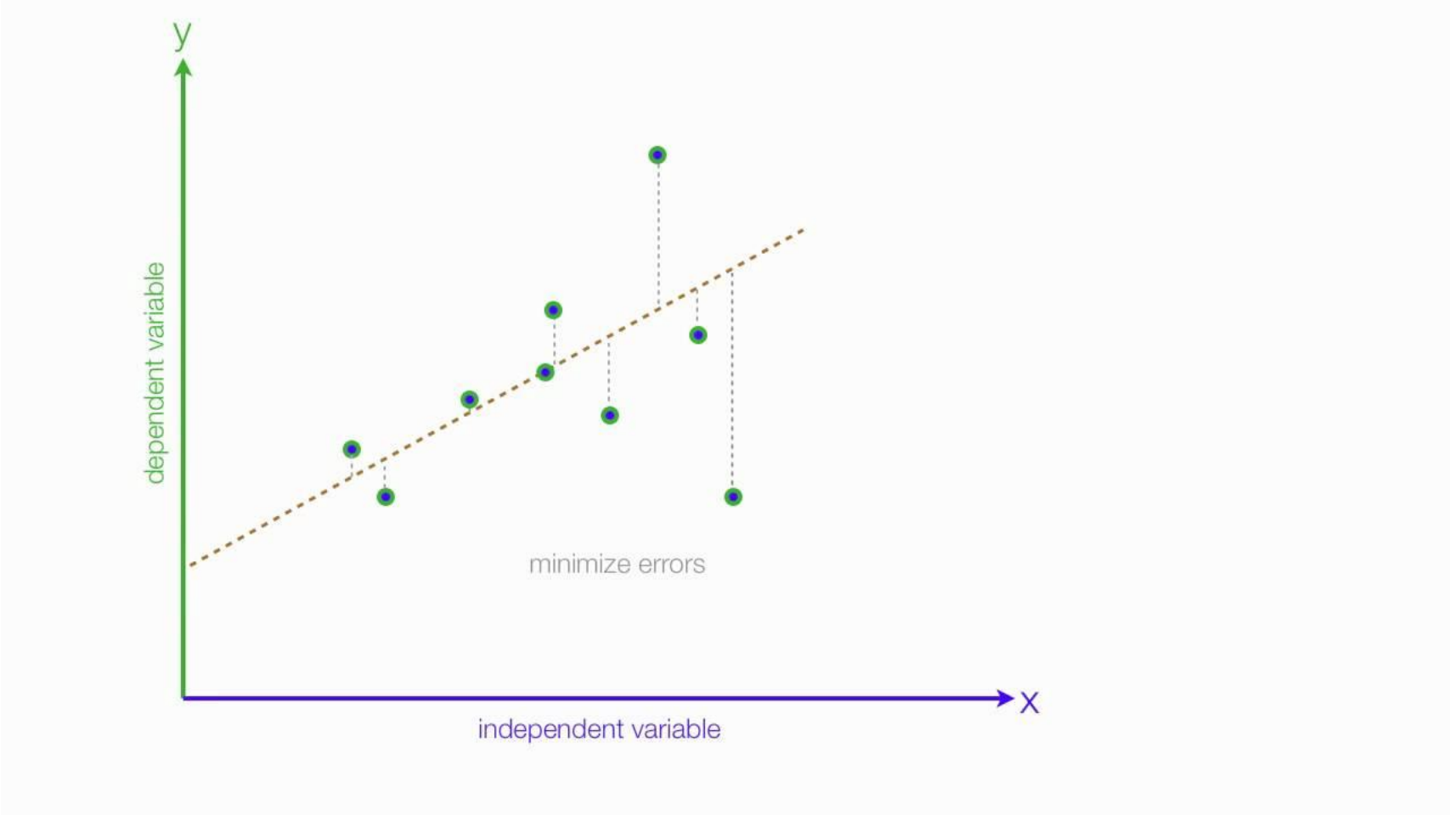
Vladimir Osin

Cross-Validation

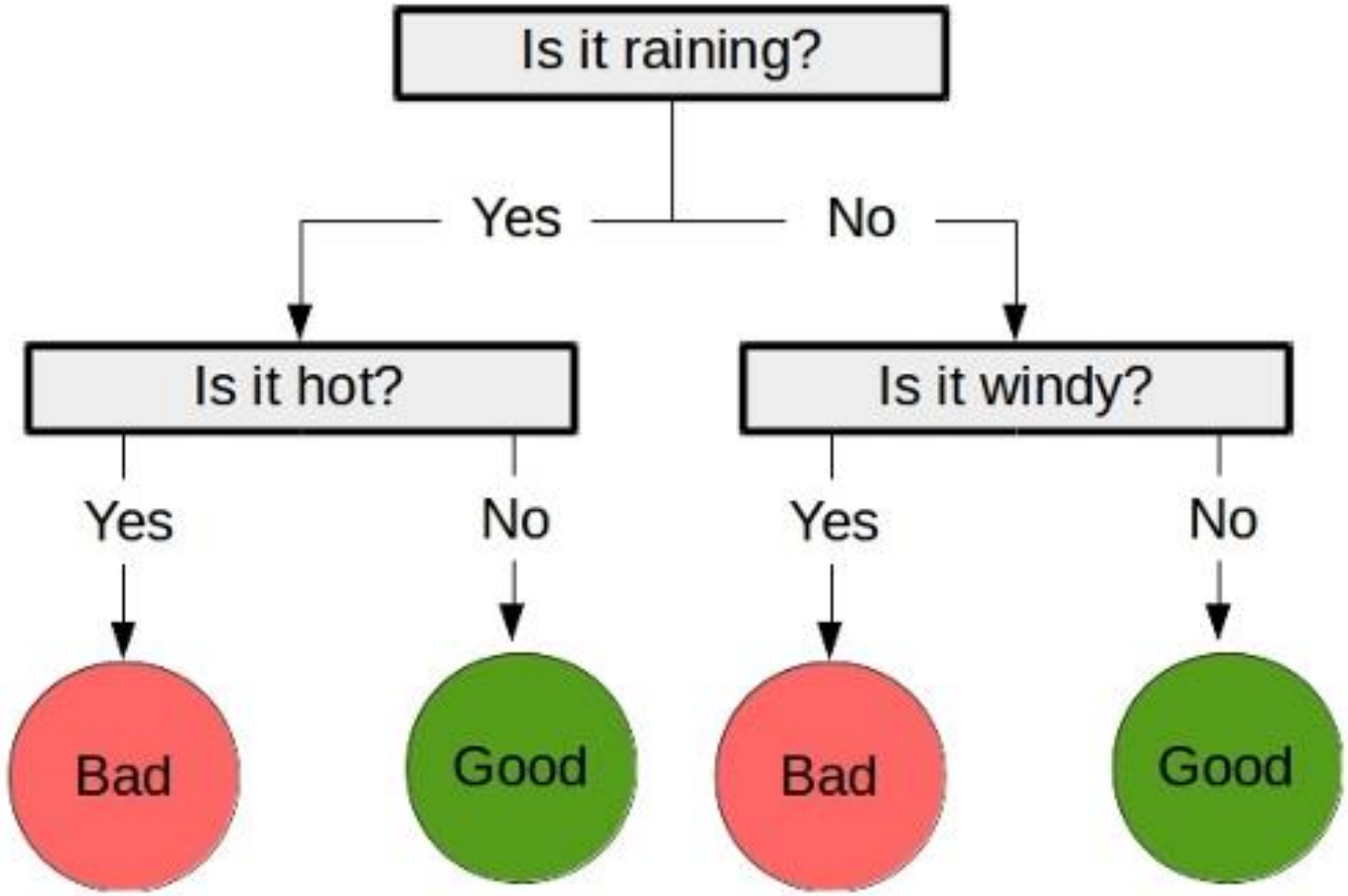
- Split your data
- Control overfitting
- Control underfitting
- Hyperparameters tuning



Linear Regression



Decision Tree

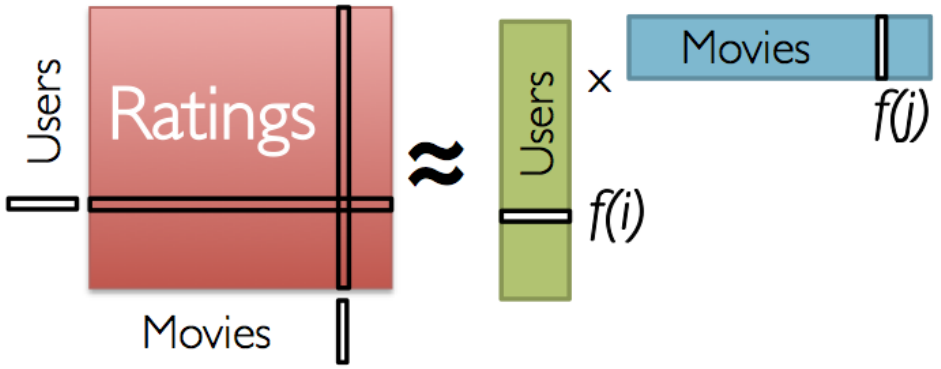


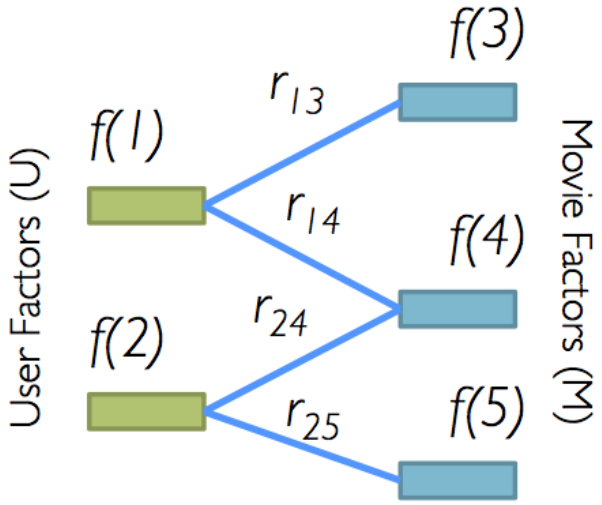
Some Results

Model	RMSE	R2
Linear Regression	0.455340	0.363434
Linear Regression with CV	0.446816	0.387042
Decision Tree with CV	0.372874	0.573129
Random forest	0.366229	0.588207
Gradient-Boosting DT	0.355057	0.612949

Recommendation system

Low-Rank Matrix Factorization:





Iterate:

$$f[i] = \arg\min_{w \in \mathbb{R}^d} \sum_{j \in \mathrm{Nbrs}(i)} \left(r_{ij} - w^T f[j]\right)^2 + \lambda ||w||_2^2$$

What else?

- https://docs.databricks.com/index.html

aws

Amazon Machine Learning

Amazon SageMaker

Build

Connect to other AWS services and transform data in Amazon SageMaker notebooks

Train

Use Amazon SageMaker's algorithms and frameworks, or bring your own, for distributed training

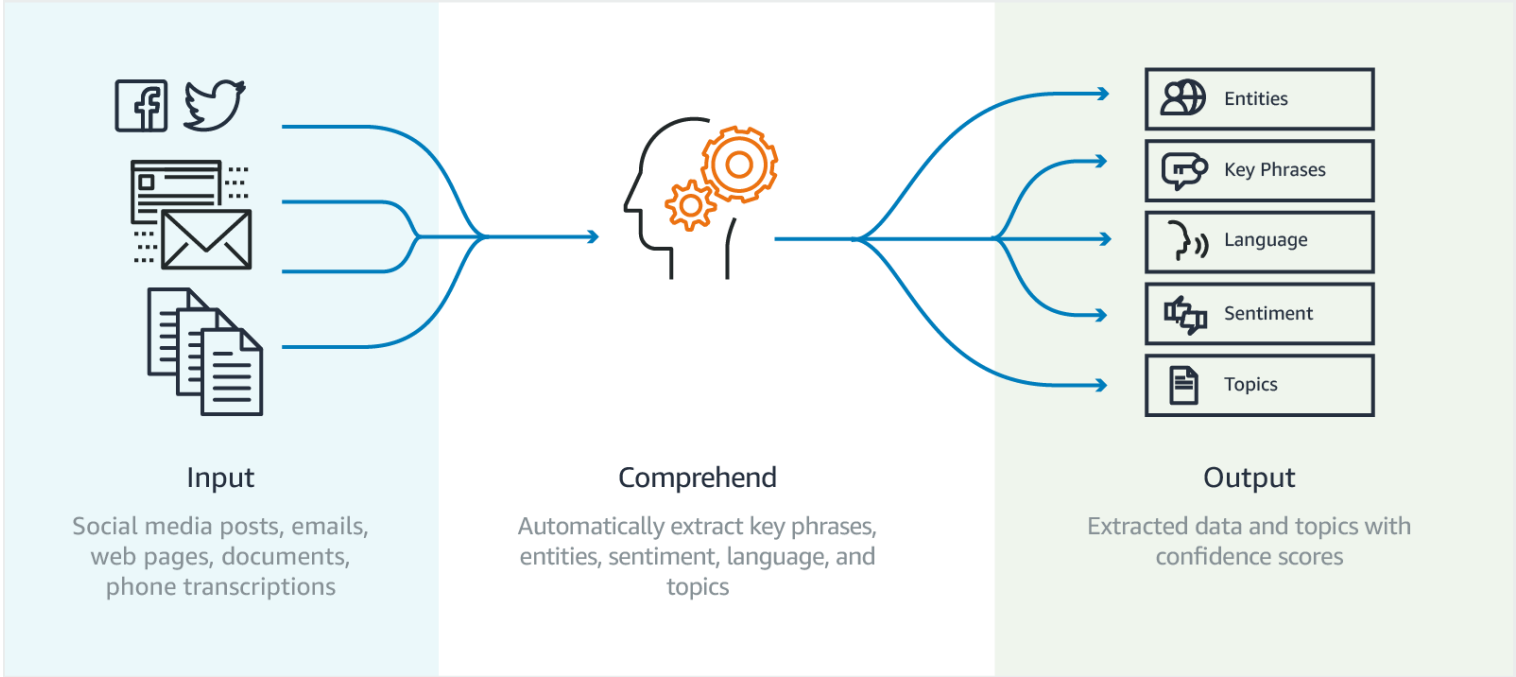
Tune

Amazon SageMaker automatically tunes your model by adjusting multiple combinations of algorithm parameters

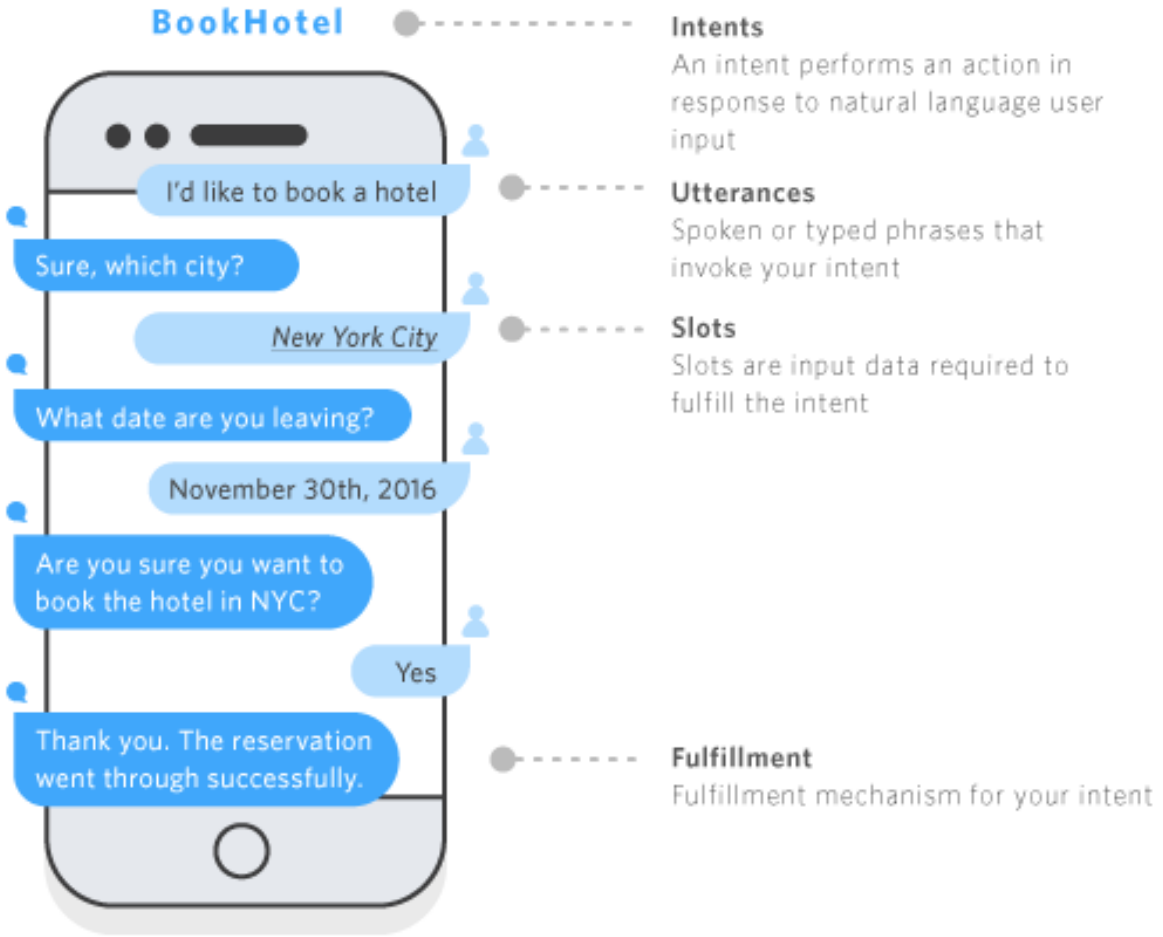
Deploy

Once training is completed, models can be deployed to Amazon SageMaker endpoints, for real-time predictions

Amazon Comprehend

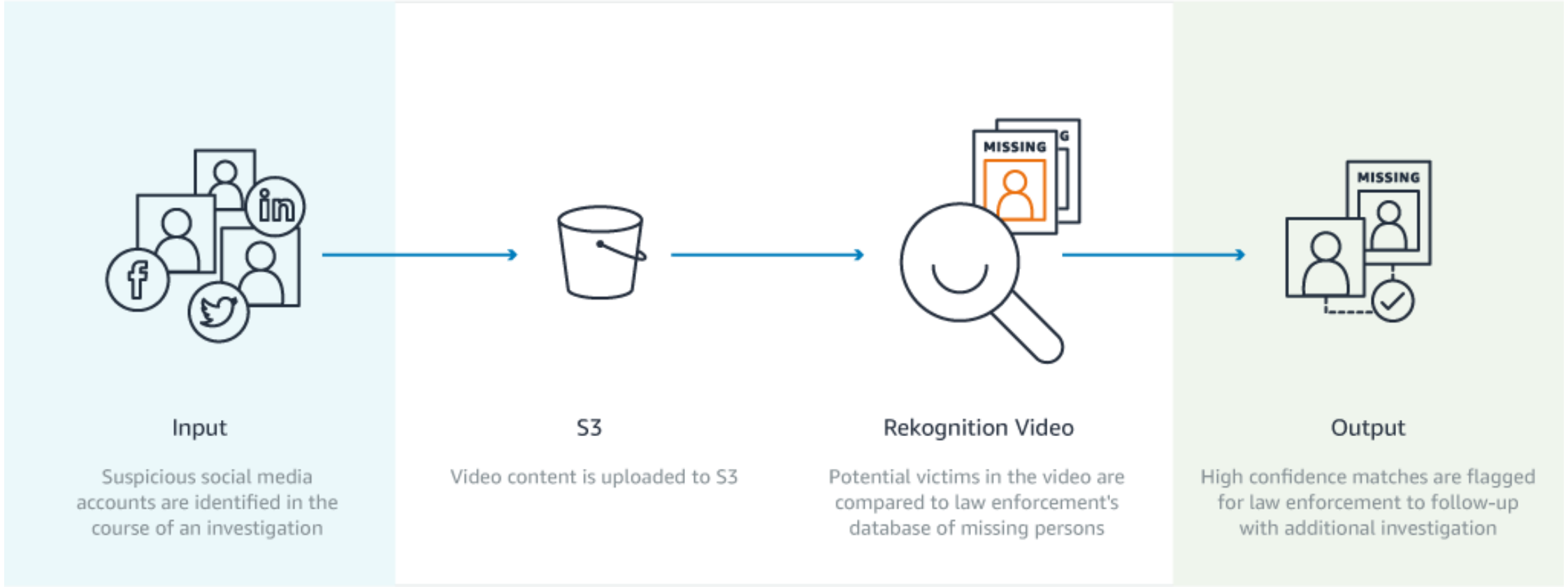


Amazon Lex


BookHotel
I'd like to book a hotel
Sure, which city?
New York City
What date are you leaving?
November 30th, 2016
Are you sure you want to book the hotel in NYC?
Yes
Thank you. The reservation went through successfully.
Intents
An intent performs an action in response to natural language user input
Utterances
Spoken or typed phrases that invoke your intent
Slots
Slots are input data required to fulfill the intent
Fulfillment
Fulfillment mechanism for your intent

Input
Customer calls the contact center to check account balance
Contact Center
Contact Center looks up balance and sends it to Amazon Polly
Amazon Polly
Amazon Polly receives text and streams speech audio back to Contact Center
Contact Center
Contact Center plays voice audio response
Output
Customer receives account balance information via audio response

Amazon Poly



Amazon Rekognition

PERSON
99.3%
OUTDOORS
83.1%
CREST
83.0%
MOUNTAIN BIKE
99.1%
ROCK
82.8%

Female
100%
Eyes are open
100%
Happy
82.8%
Smiling
75.8%

PERSON 2
PERSON 4
PERSON 3
PERSON 1

Machine learning

Amazon Transcribe
Automatic Speech Recognition

Amazon Transcribe provides high-quality and affordable speech-to-text transcription for a wide range of use cases.

Machine Translation Use Cases

Enable multilingual sentiment analysis of social media content

With Amazon Translate, you are not restricted by language barrier. Understand the social sentiment of your brand, product, or service while monitoring online conversations in different languages. Simply translate the text to English before using a natural language processing (NLP) application like Amazon Comprehend to analyze textual content in a multitude of languages.

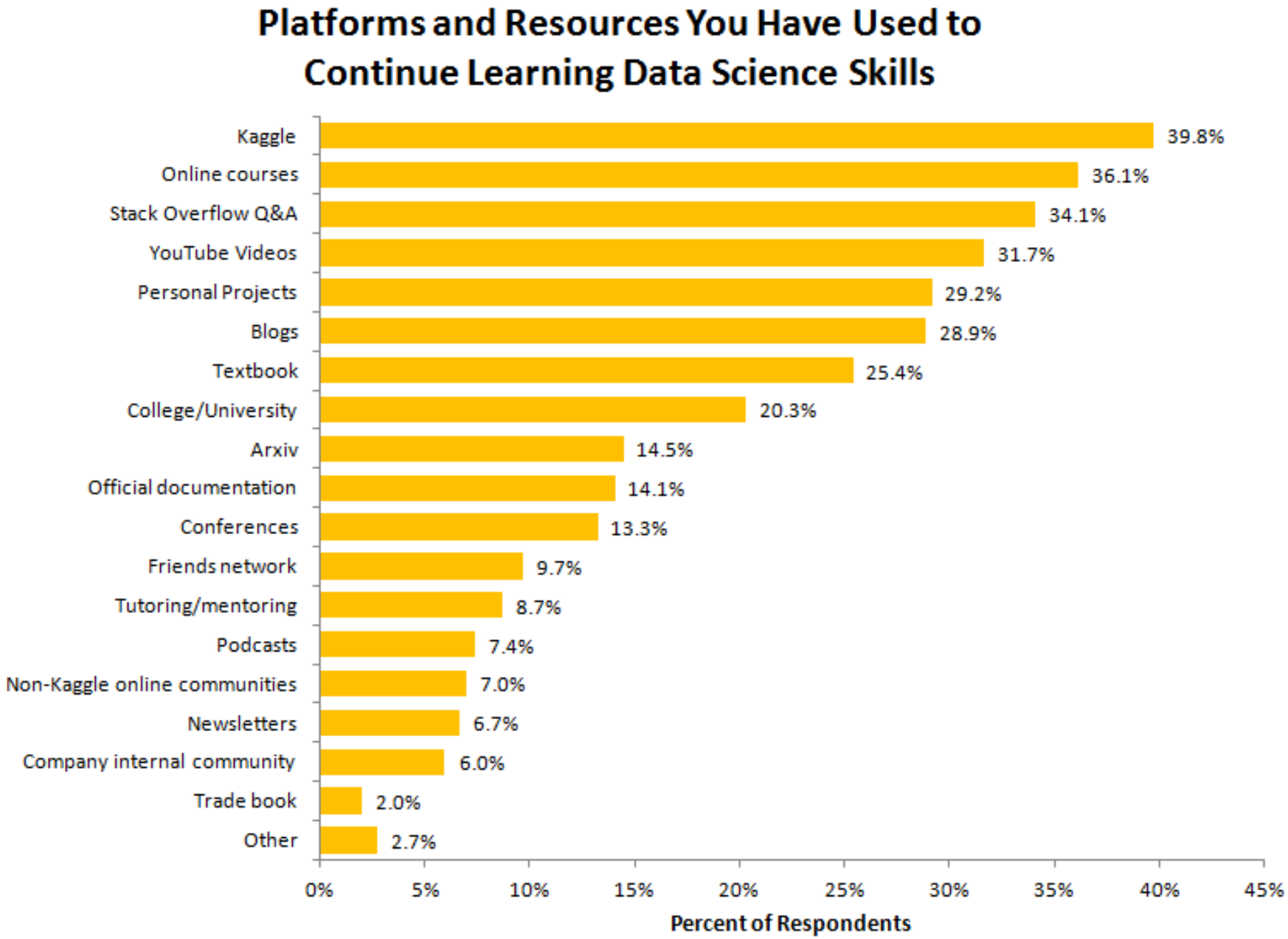
Provide on-demand translation of user-generated content

It's very difficult for human translation teams to keep up with dynamic or real-time content. With Amazon Translate, you can easily translate massive volumes of user-generated content in real-time. Websites and applications can automatically make content such as feed stories, profile descriptions, and comments, available in the user's preferred language with a click of a "translate" button.

Add real-time translation for communications applications

Amazon Translate can provide automatic translation to enable cross-lingual communications between users for your applications. By adding real-time translation to chat, email, helpdesk, and ticketing applications, an English-speaking agent or employee can communicate with customers across multiple languages.

kaggle™


Platforms and Resources You Have Used to Continue Learning Data Science Skills
Kaggle
Online courses
Stack Overflow Q&A
YouTube Videos
Personal Projects
Blogs
Textbook
College/University
Arxiv
Official documentation
Conferences
Friends network
Tutoring/mentoring
Podcasts
Non-Kaggle online communities
Newsletters
Company internal community
Trade book
Other
39.8%
36.1%
34.1%
31.7%
29.2%
28.9%
25.4%
20.3%
14.5%
14.1%
13.3%
9.7%
8.7%
7.4%
7.0%
6.7%
6.0%
2.0%
2.7%
0%
5%
10%
15%
20%
25%
30%
35%
40%
45%
Percent of Respondents

Competitions Platforms

- Kaggle
- DrivenData
- CrowdAnalityx
- CodaLab
- DataScienceChallenge.net
- DataScience.net
- Single-competition sites (like KDD, VizDooM)

Why to participate?

- Great opportunity to **learn** and networking
- Interesting non-trivial tasks and state-of-the-art approaches
- A way to become famous inside data science community
- A way to earn some money

- Wrap up competition results as:
  - LinkedIn project
  - GitHub repository
  - Medium blog

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

5,058 teams · 2 years to go

Overview

Description
Evaluation
Frequently Asked Questions
Tutorials

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

scikits
learn
machine learning in Python