- Inferencing at the edge:

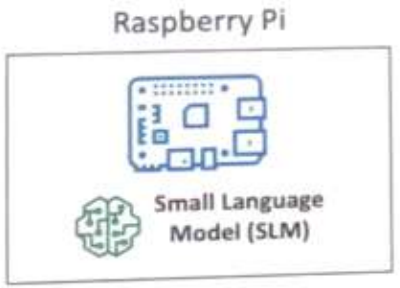
↳ Edge devices are the devices with less computing power that are close to where the data is generated, in places where internet connections can be limited.

i. Small language Model (SLM):

↳ Runs on an edge device.

↳ Very low latency, low compute footprint

↳ Offline Capability and local inference

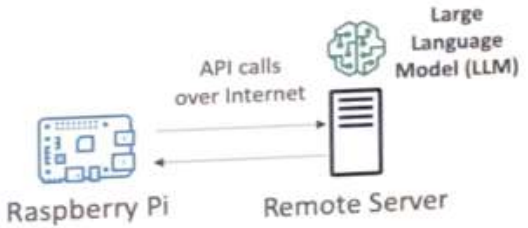


ii. Large language model (LLM):

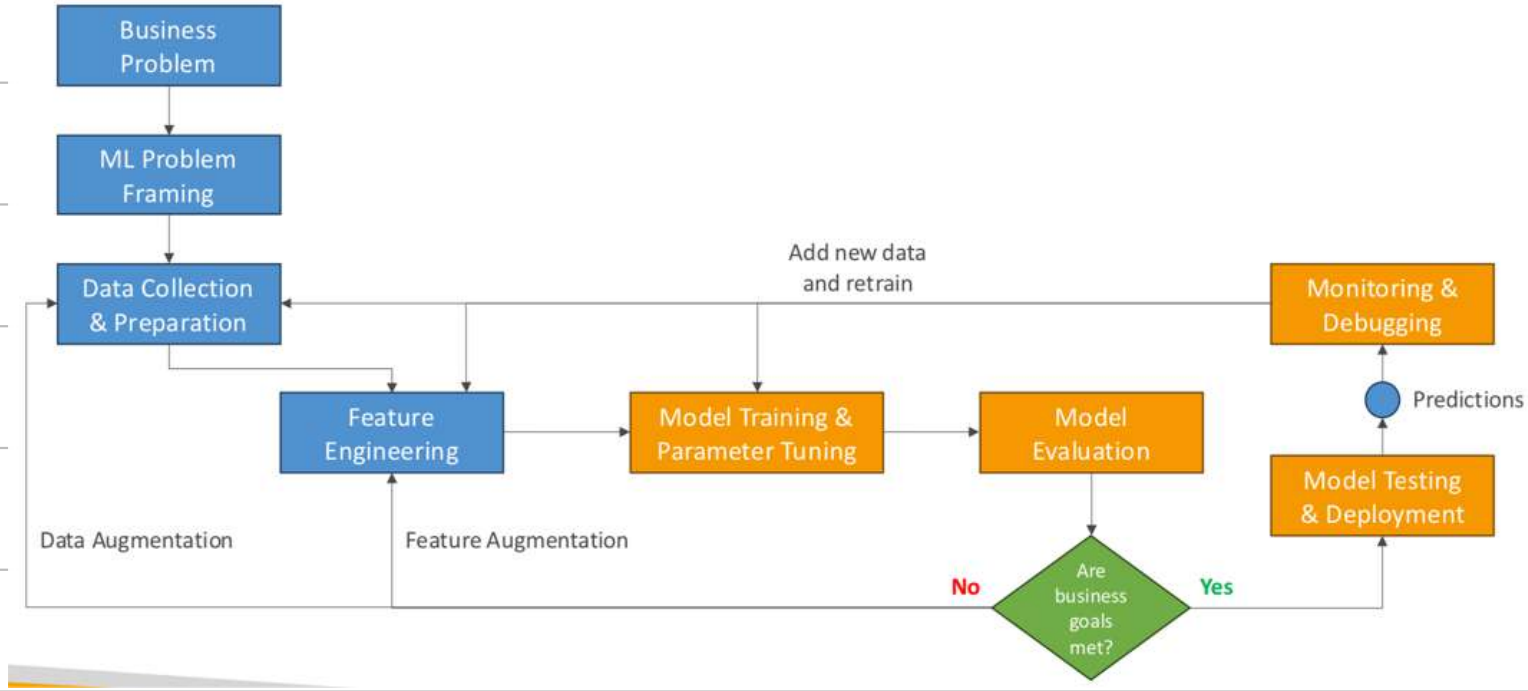
↳ runs on a remote server.

↳ more powerful model, high latency

↳ must be online to be accessed



- Phases of Machine Learning Project:



↳ Exploratory Data Analysis:

- Visualize the data with graphs.
- Correlation Matrix: look at the correlations b/w variables and decide which features can be important to the model.

- Hyperparameter Tuning:

↳ hyperparameter is a setting that define the model structure and learning algorithm and process and set before the training begins

Ex: learning rate, batch size, number of epochs & regularization.

↳ tuning is finding the best hyperparameter values to optimize the model performance.

↳ it improves model accuracy, reduces overfitting and enhances generalization.

↳ It is done using Grid search, Random Search, SageMaker Automatic Model Tuning (AMT)

↳ Learning Rate: how large or small steps when updating model's weight during training.

higher leads to faster convergence, lower result in more precise but slower convergence.

↳ Batch size: No. of training examples used to update model weight. Small - more stable but takes more time to compute, larger - faster but less stable updates.

↳ No. of epochs: No. of times the model will iterate over the entire training dataset. Too few - underfitting, too many - overfitting.

AWS Managed AI Services

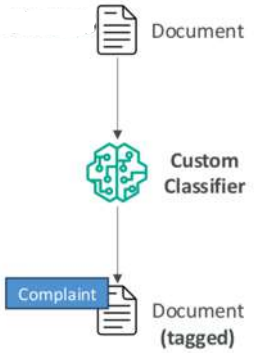
- These are pre-trained ML services for your usecase.
- Responsiveness and Availability.
- Redundant & Regional Coverage
- Performance
- Token-Based pricing
- Provisioned throughput (for predictable workloads).
- Amazon Comprehend:

↳ for NLP, fully managed and serverless service.

↳ Uses ML to find insights and relationships in text:

1. language of text
2. Extracts key phrases, people, places, brands or events
3. How positive or negative is the text
4. Analyzes text using tokenization and parts of speech
5. Automatically organizes a collection of text files by topic.

↳ It can organize documents into categories of custom classes at both real-time and asynchronous mode and supports different document types.



↳ Named Entity Recognition (NER): extracts predefined general-purpose entities like people, places, organization, dates and other standard categories from text.

↳ Custom Entity Recognition: Analyze text for specific terms and noun-based phrases. To do this, we can train the model with custom data like list of entities & documents that contain them. Ex: policy number or customer escalation etc

- Amazon Translate:

↳ Natural and accurate language translation

↳ It allows you to localize content (websites/applications) for international users.

↳ It easily translates large volumes of text efficiently.

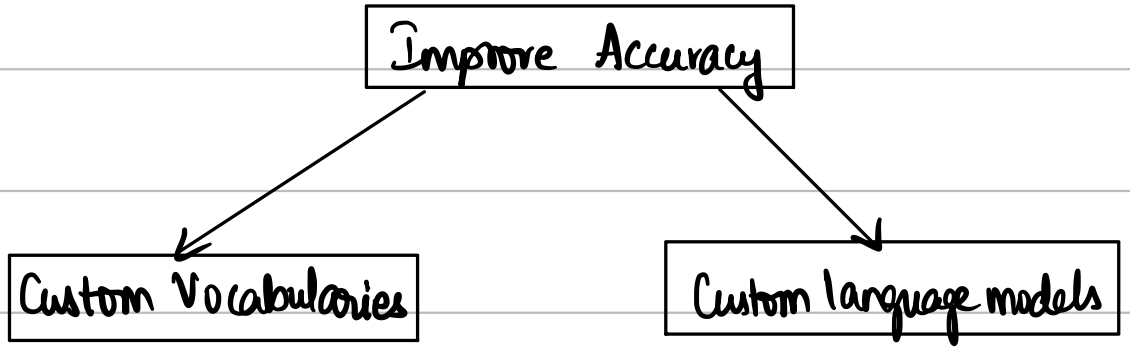
- Amazon Transcribe:

↳ Automatically convert speech to text.

↳ Uses a deep learning process called automatic speech recognition (ASR)

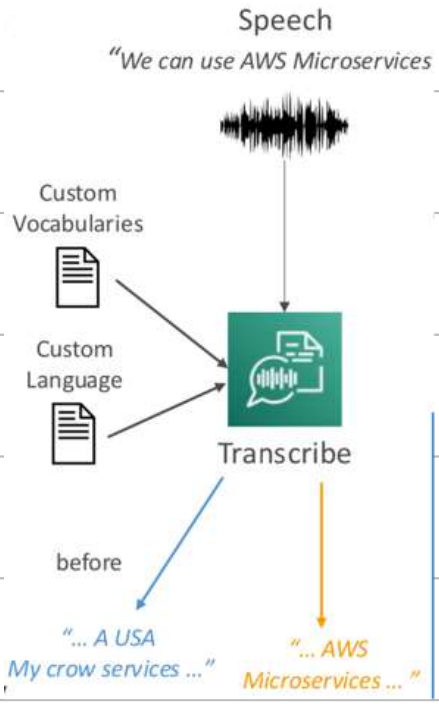
↳ automatically removes PII using Redaction

↳ supports automatic language identification for multi-lingual audio.



- for words
- domain specific terms, brands...
- provide pronunciation

- for context
- good for domain-specific data
- learn context associated with it.



↳ use both of these for high accuracy

- Amazon Polly:

↳ Turn text into lifelike speech using deeplearning.

↳ Allows you to create applications that talk.

i, Lexicons: Define how to read specific pieces of text.

"AWS" - Amazon Web Services

ii, SSML (Speech Synthesis Markup language):

↳ Markup for you text to indicate how to pronounce it.

Hello, <break> how are you?

iii, Voice Engine: generative, long-form, neural, standard ...

iv, Speech Mark:

↳ Encode where a sentence/word starts or ends in audio

↳ Helpful for lip-syncing or highlighting

Action	SSML tag
Adding a pause	<break>
Emphasizing words	<emphasis>
Specifying another language for specific words	<lang>
Placing a custom tag in your text	<mark>
Adding a pause between paragraphs	<p>
Using phonetic pronunciation	<phoneme>
Controlling volume, speaking rate, and pitch	<prosody>
Setting a maximum duration for synthesized speech	<prosody amazon:max-duration>
Adding a pause between sentences	<s>
Controlling how special types of words are spoken	<say-as>
Identifying SSML-enhanced text	<speak>
Pronouncing acronyms and abbreviations	<sub>
Improving pronunciation by specifying parts of speech	<w>
Adding the sound of breathing	<amazon:auto-breaths>
Newscaster speaking style	<amazon:domain name="news">
Adding dynamic range compression	<amazon:effect name="drc">
Speaking softly	<amazon:effect

- Amazon Rekognition:

↳ find objects, people, text, scenes in images and videos using ML.

↳ can perform facial analysis and facial search.

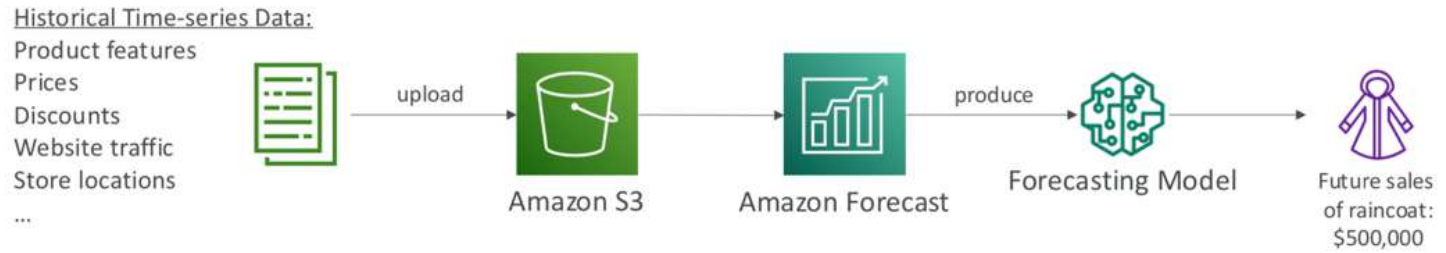
↳ you can also create a db of "familiar faces".

↳ Use cases include: labeling, content moderation, text detection, face detection & analysis, face search & verification, celebrity recognition, pathing (sports game analysis)

i, Custom labels: identify specific images you needed (ex: logos) in images or videos. Just create a custom model by giving few hundred images or less.

- Content Moderation: Automatically detect inappropriate, unwanted or offensive content. It brings down human review to 1-5% of total content volume. If you integrate that with Amazon Augmented AI (A2I), you get human review. You can also create custom moderation Adaptors by providing your own labelled data & train the model.

- Amazon forecast: fully managed service that uses ML to deliver highly accurate forecasts. It is 50% more accurate than looking at the data itself.



- Amazon Lex: It builds chatbots quickly for your application using voice and text. It integrates with AWS Lambda, Connect, Comprehend, Kendra to understand the intent

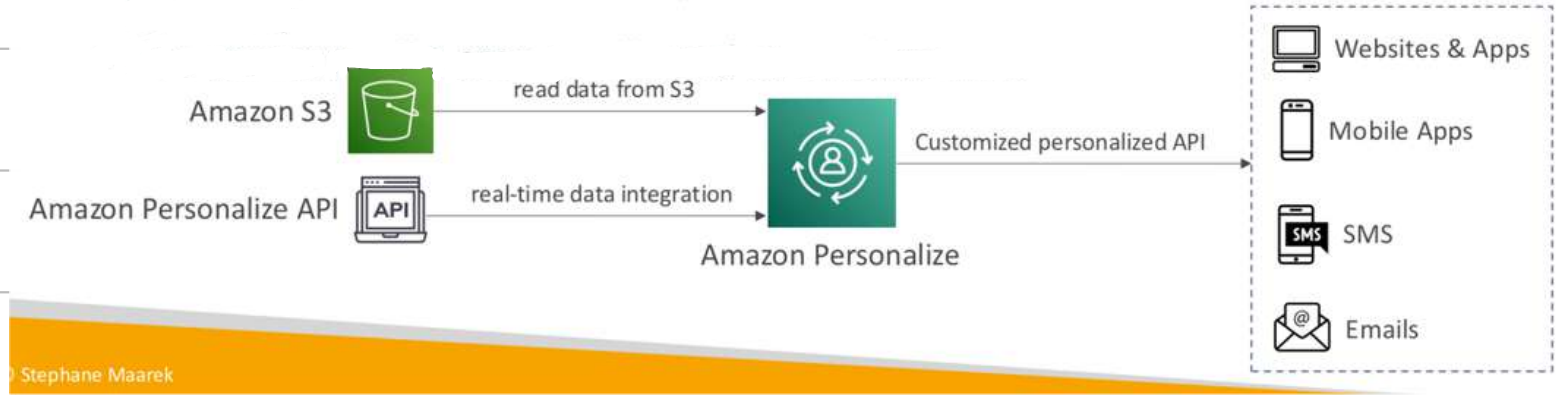
and invoke the correct lambda function after asking for slots (i/p parameters)

- Amazon Personalize:

↳ fully managed ML service for apps with personalized recommendations.

↳ Same technology used by Amazon.com

↳ integrates into existing websites, apps, SMS, email marketing systems etc.



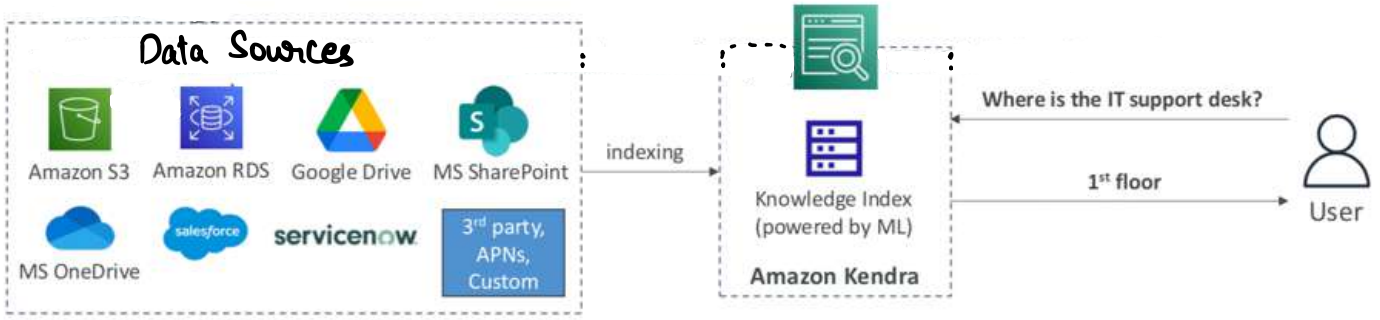
↳ Recipes: Algorithms for specific usecase, but need training configuration.

Ex: User_personalization, Personalized_Ranking, Popular_items, related_items etc.

- Amazon Textract: It automatically extracts text, handwriting and data from any scanned documents (any type) using AI and ML.

  ↳ Used in financial services, healthcare and public sector.

- Amazon Kendra: It is a fully managed document search service. It extracts answers from within the document with natural language search. It can learn from human interactions and can fine-tune the answers as well.



- Amazon Mechanical Turk: It is a crowdsourcing marketplace to perform simple human tasks like having a distributed virtual workforce. It deeply integrates

with Amazon A2I, SageMaker Ground Truth etc.

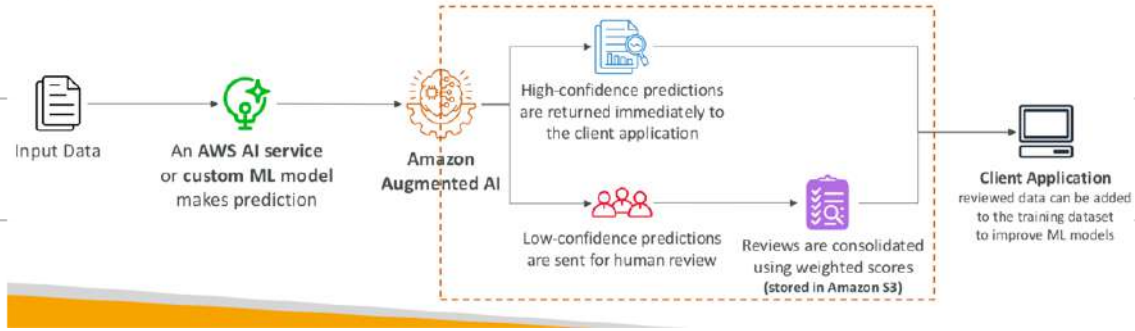
Uses: Image classification, data collection, business processing etc.

- Amazon Augmented AI (A2I):

↳ Human oversight of ML predictions in production.

↳ These people can be own employees, contractors, Mechanical Turk, pre-screened vendors etc.

↳ The ML model can be built on AWS or elsewhere.



- AWS DeepRacer:

↳ fully autonomous 1/18th scale car race driven by Reinforcement Learning (RL).

↳ DeepRacer Console to train & evaluate model in 3D.

↳ If needed, can be deployed in a real vehicle too.

- AWS Comprehend Medical & Transcribe Medical:

↳ Automatically convert medical-related speech to text (HIPAA compliant)

↳ Can transcribe medical terminology & supports real-time & batch transcriptions.

↳ Comprehend Medical detects and returns useful info. in unstructured clinical text.

↳ Uses NLP to detect Protected Health Information (PHI)

↳ Use Amazon S3 or Kinesis Data firehouse for real-time.

- Amazon's Hardware for AI:

i, Amazon EC2: Elastic Compute Cloud

↳ Renting virtual machines (EC2)

↳ Storing data on virtual drives (EBS)

↳ Distributing load across machines (ELB)

↳ Scaling the services using an auto-scaling group (ASG)

↳ Can configure: OS, CPU, RAM, storage, network card, security group, bootstrap script

- GPU based EC2 instances (P3, P4, P5..., G3... G6...)
- AWS Trainium: ML chip built to perform DL on 100B+ parameter models (Trn1). It reduces cost by 50% while training the model.
- AWS Inferentia: ML chip built to deliver inference at high performance & low cost. It gives upto 4x throughput and 70% cost reduction (Inf1, Inf2)
- Trn & Inf have the lowest environmental footprint.