AMRITA VISHWA VIDYAPEETHAM | School of Engineering

What happens before and after writing code for data science projects in the real world?

Kaushik Shakkari

Introduction - Kaushik Shakkari

- Senior Data Scientist, Cognistx
- Subject Matter Expert / AI course contributor, CMU
- Alumni Mentor at Insight Data Science Program
- Masters in Computer Science (Data Science) at USC
- B.Tech in CSE at Amrita University, Coimbatore
- Part time blogger - Medium
- LinkedIn Profile

Huge Gap: Academia VS Industry

	Academia	Industry
Skill Set	Intellectual understanding of concepts	Technical proficiency, AI and ML tools
End Goal	Learn theoretical concepts, research paper	Deploy AI service to customers for $$$
Metrics	Improving model metrics	Improving business metrics + ROI
Type of data	Often clean and available	Often messy, complex, varied and raw
Scale of data	Often small datasets	Relatively large datasets

Note

What happens before and after writing code in industry might not be taught in an academic setting

Agenda

- Before Writing Code
- Experiment Tracking with MlFlow and Optuma (Colab)
- After Writing Code
- Conclusion

Before writing code

Tasks before writing the base code

- Requirements Gathering
- Customer Value Proposition
- Effort Estimation and Timeline
- Machine Learning System Design
- Project and Version Control Setup

Requirements Gathering

- **Goals:**
  - What customers ***want*** vs ***need***?
  - Is problem **feasible** to solve?
  - Break down problem into tasks + prioritize tasks
  - What are **benefits** vs **risks** for company?
  - Get details for avoiding bad assumptions for machine learning system design
- **Initial questions on problem**
  - Why we need to solve this problem? *eg: maximize user engagement (Recsys)*
  - Why we need to solve **now**? -> *Competitors? Customer churn?*
  - What is in-scope and out of scope? *(assumptions and constraints)*
  - How does the end user use the system? *(If possible, get a walkthrough of the user flow)*

More Questions -> More Clarity

- **Questions on defining the success of the project**
  - What are business success metrics? (Extensic Metric)
  - How costly are errors: False Negative vs. False Positives?
  - Offline Metrics (Development) & Online Metrics (Production)
- **Questions on data**
  - Are all the data sources identified?
  - What type of data do you have access to?
  - What data sources do you need to collect?
  - How will the data be used? (data security and privacy related compliance)
  - Do we have enough annotations for training and evaluation?
    - If not, do we have resources to collect and manage annotations?
    - Define annotation process to ensure good annotation quality

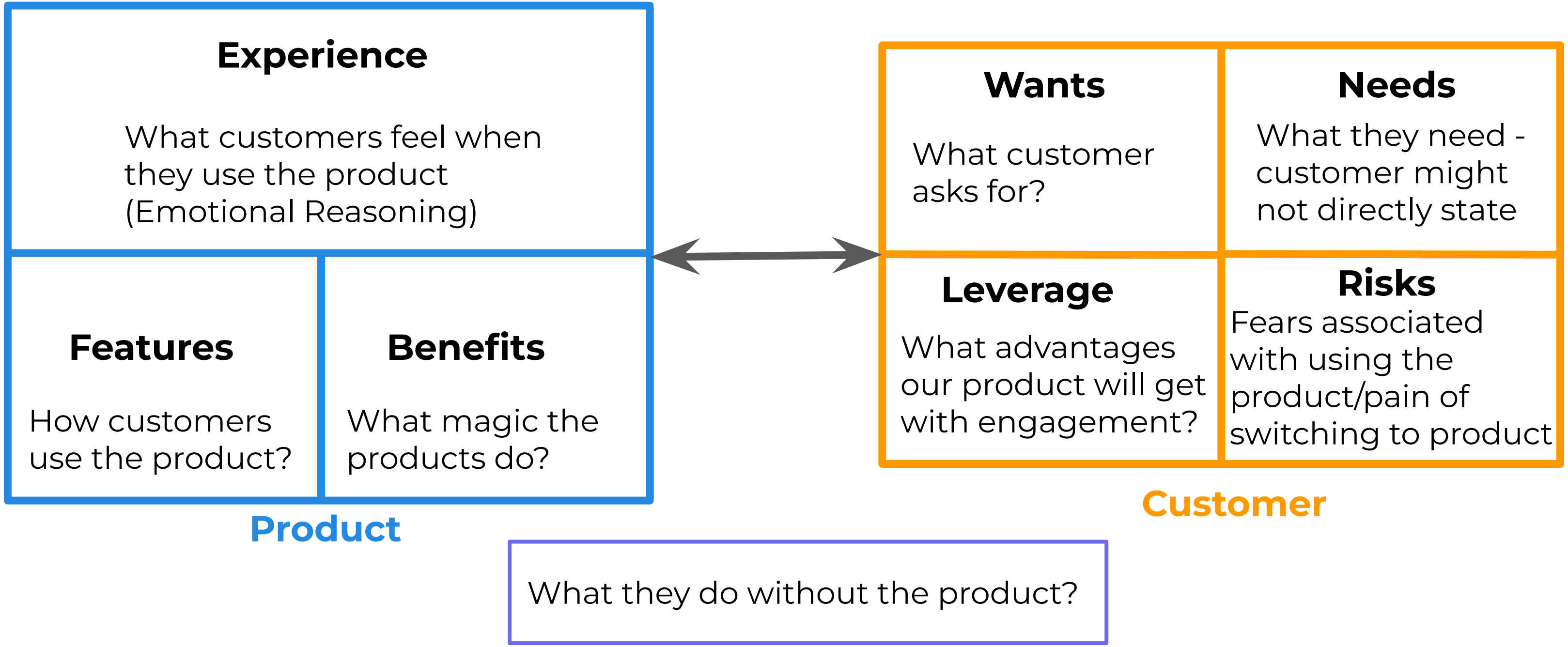
More Questions -> More Clarity

- **Questions on infrastructure and deployment**
  - Any cloud and resource limitations?
  - Do customers have the expected number of users? Concurrent users?
  - Do predictions need to be real-time or batch? How often? How fast?
- **Questions on maintenance and monitoring**
  - What data should be monitored to ensure the model behaves as expected?
  - Will domain expert gives continuous feedback?
    - How often you get feedback and models need to be evaluated?
    - How often models need to be retrained and deployed?
    - Manual retraining vs auto train - CI/CD + CT

Customer Value Proposition

- CVP is a place where company product interest intersects customer’s requirements.
- Gives minimum viable clarity to start the project
- Answers why people needs to buy your product in x domain
- Helps to define business success metrics
- How non-AI solutions performs against the success metric
- To ensure models have business impact
- Business metric X DS/SWE metric Association
- Avoids overselling the product -> longer customer engagement (recurring customer)

My template for CVP



Experience

What customers feel when they use the product (Emotional Reasoning)

Features

How customers use the product?

Benefits

What magic the products do?

Product

Wants

What customer asks for?

Needs

What they need - customer might not directly state

Leverage

What advantages our product will get with engagement?

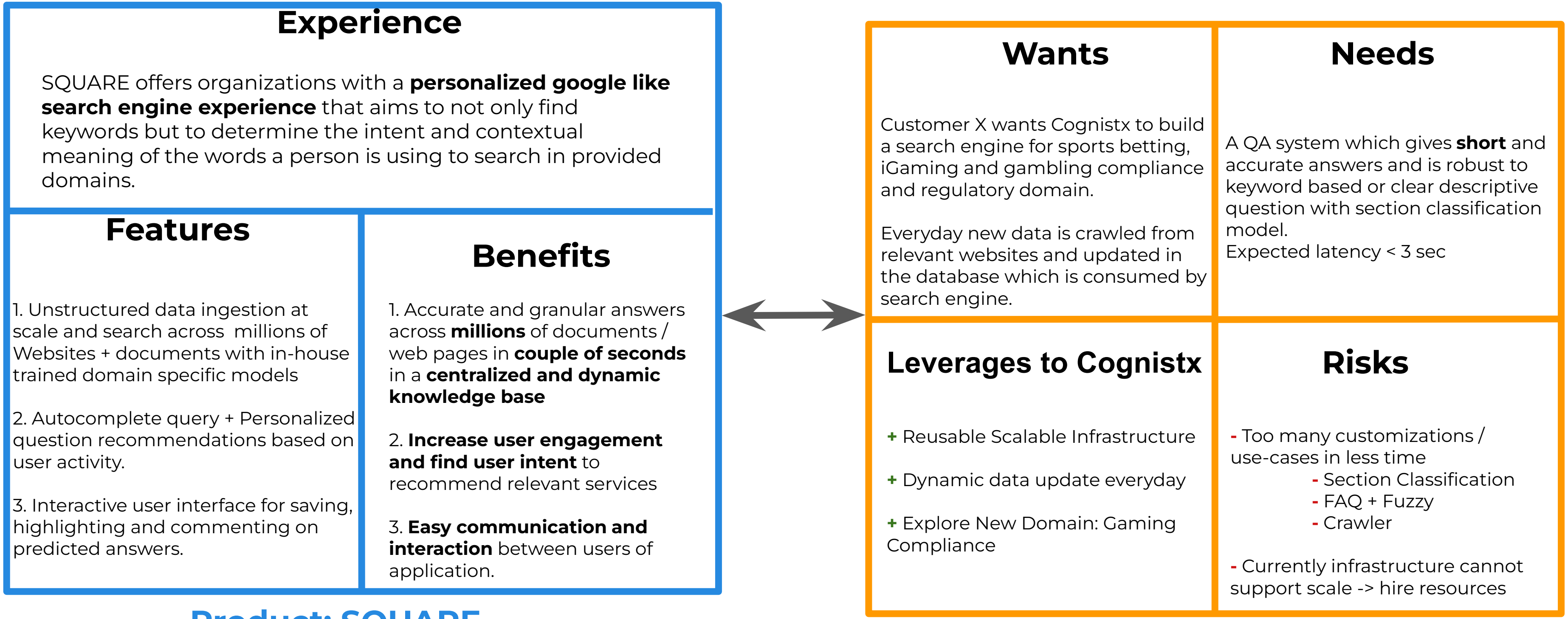
Risks

Fears associated with using the product/pain of switching to product

Customer

What they do without the product?

Sample CVP of Cognistx's SQUARE



Experience

SQUARE offers organizations with a **personalized google like search engine experience** that aims to not only find keywords but to determine the intent and contextual meaning of the words a person is using to search in provided domains.

Features

1. Unstructured data ingestion at scale and search across millions of Websites + documents with in-house trained domain specific models

2. Autocomplete query + Personalized question recommendations based on user activity.

3. Interactive user interface for saving, highlighting and commenting on predicted answers.

Benefits

1. Accurate and granular answers across **millions** of documents / web pages in **couple of seconds** in a **centralized and dynamic knowledge base**

2. **Increase user engagement and find user intent** to recommend relevant services

3. **Easy communication and interaction** between users of application.

Product: SQUARE

Wants

Customer X wants Cognistx to build a search engine for sports betting, iGaming and gambling compliance and regulatory domain.

Everyday new data is crawled from relevant websites and updated in the database which is consumed by search engine.

Needs

A QA system which gives **short** and accurate answers and is robust to keyword based or clear descriptive question with section classification model.
Expected latency < 3 sec

Leverages to Cognistx

+ Reusable Scalable Infrastructure

+ Dynamic data update everyday

+ Explore New Domain: Gaming Compliance

Risks

- Too many customizations / use-cases in less time
- Section Classification
- FAQ + Fuzzy
- Crawler

- Currently infrastructure cannot support scale -> hire resources

Customer: X

Effort Estimation and Timeline

- **Goals:**
  - Estimates resources needed for full time (100%) or part time of their work
    - Sample: 2 analysts (SME), 1 data scientist and 1 machine learning engineer
  - Milestones and deadlines to track the progress in future

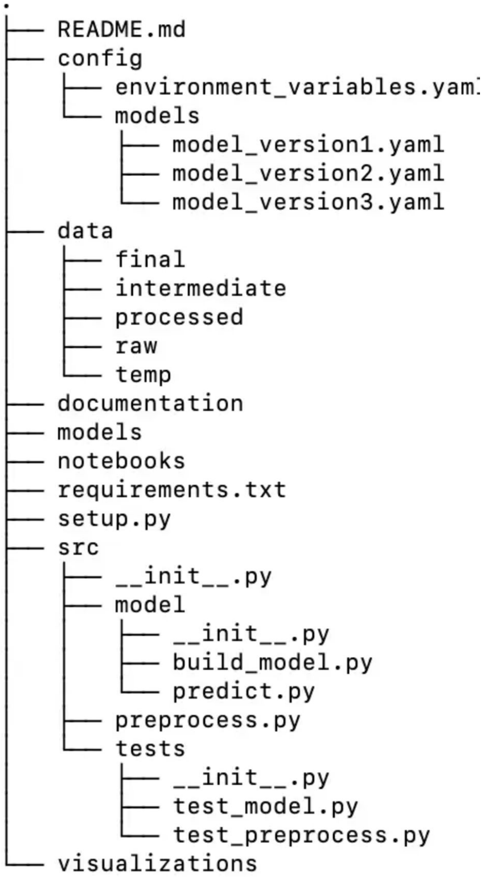
- **Steps:**
  1. Break down the project into smaller tasks
  2. Assess the duration of each task
  3. Put together the timeline
  4. Finalize the timeline with stakeholders

Note

This task is often performed by senior technical member like senior data scientist with a project manager

Project Setup and Version Control

- Setup repository (GIT)
- Installations
  - cloud, languages, tools and technologies
- Store and track project related files for reproducibility
  - Code Versioning
  - Data Versioning
- Hands on article for quick setup for code and data versioning with DVC and AWS S3 : [Medium]

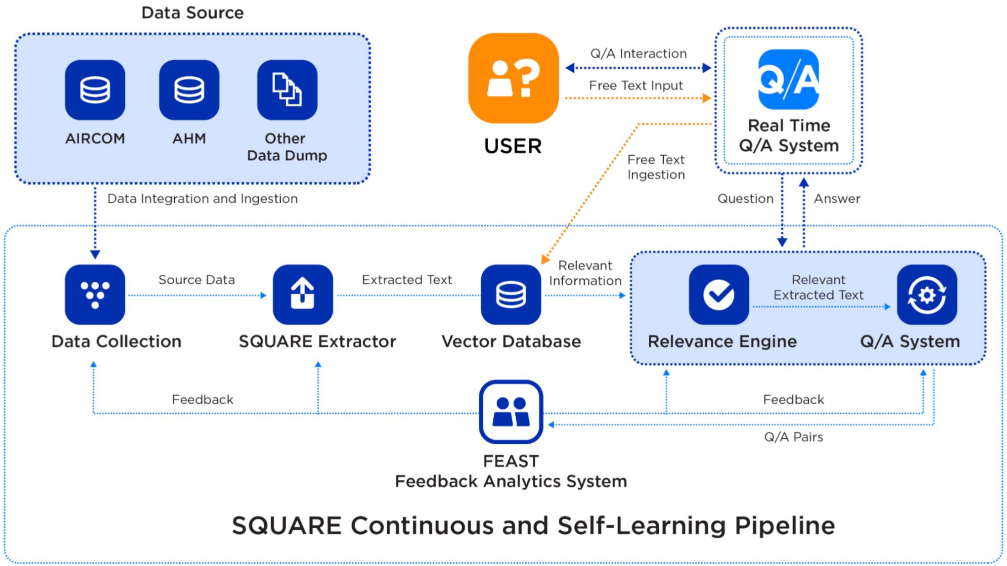


```
.
├── README.md
├── config
│   ├── environment_variables.yam
│   └── models
│       ├── model_version1.yaml
│       ├── model_version2.yaml
│       └── model_version3.yaml
├── data
│   ├── final
│   ├── intermediate
│   ├── processed
│   ├── raw
│   └── temp
├── documentation
├── models
├── notebooks
├── requirements.txt
├── setup.py
├── src
│   ├── __init__.py
│   ├── model
│   │   ├── __init__.py
│   │   ├── build_model.py
│   │   └── predict.py
│   ├── preprocess.py
│   └── tests
│       ├── __init__.py
│       ├── test_model.py
│       └── test_preprocess.py
└── visualizations
```

Sample Code Base Structure

Machine Learning System Design

- There is no one-size-fits-all design
- Start with a simple rough system flow design



Note

The architect or senior technical member will approve the proposed architecture

Note

The diagram gets more polished with specific service details over time (Iterative)

Experimentation Tracking

- Skipping Typical data science life cycle process - Data Extraction, Exploratory Data Analysis, and Data Processing, modeling and evaluation etc for webinar.
- Start with simple Jupyter Notebook
- Experimentation Tracking using MLFlow (Open Source and Free)
  - Log and Analyse parameters and metrics for experiments
  - Model Repository: Run IDs -> Load best model from Run IDs
  - Tune hyperparameters using Optuna
  - Colab

After writing code

Note

I have skipped data cleaning and modeling (code part)

Tasks after writing the base code

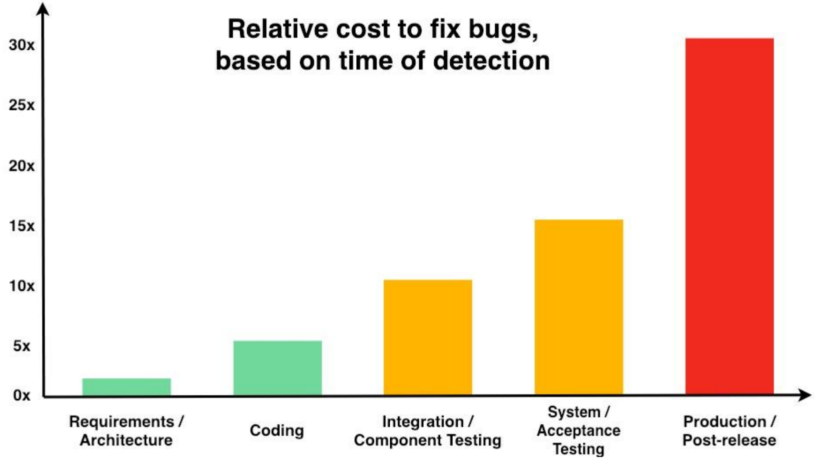
- Skipping Packaging and Documentation
- Logging
- Testing
- Deployment
- Monitoring and Maintance

Logging

- Record important information which later can be used to debug
- Logging Levels: (Highest Severity to Lowest) ; Set Logger Level
    - **CRITICAL:** Need to handle immediately
    - **ERROR:** There is a break in some component in code
    - **WARNING:** Not a bug but just we aware of this log
    - **INFO:** Just an information log
    - **DEBUG:** Just a debugging statement
- Multiple instances -> logs (Cloudwatch)

Testing

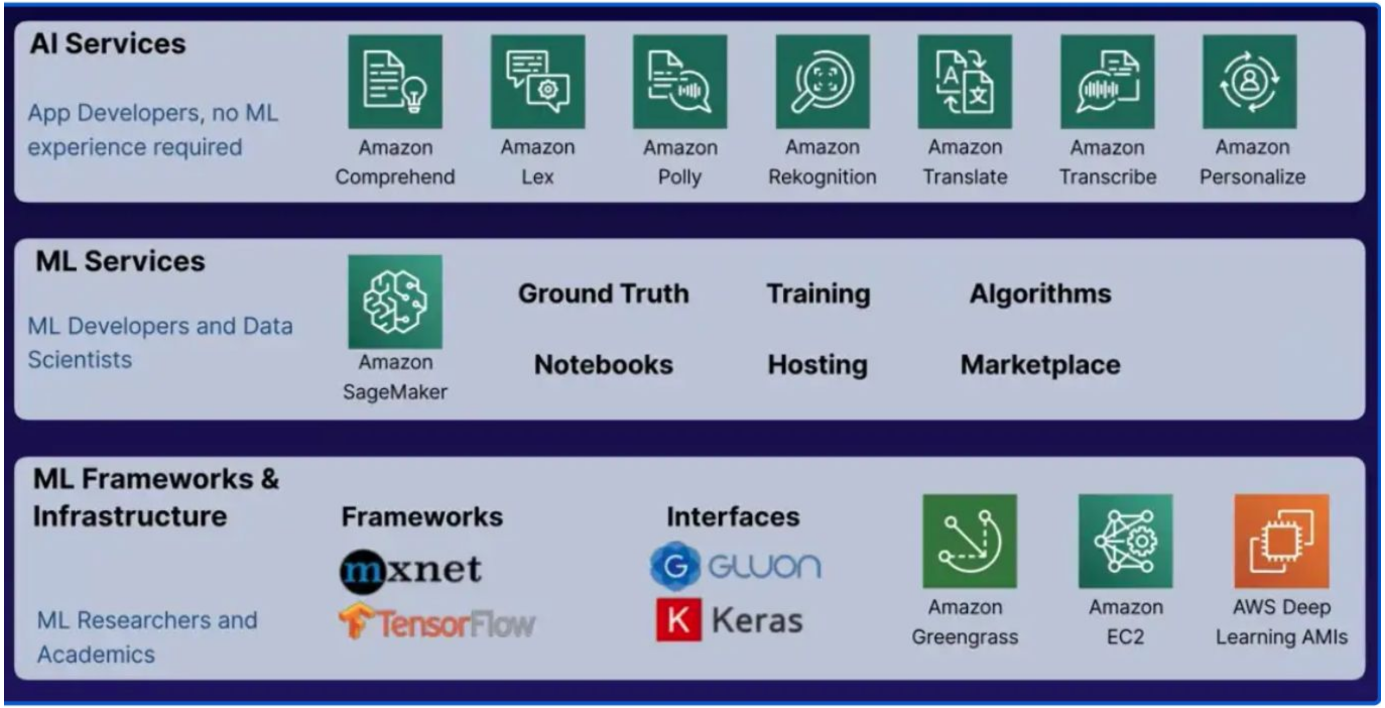
- Earlier you find and fix bugs -> More money can be saved
- What we test - data, model and code
- **Main types of testing ML systems:**
    - Unit Testing
    - Regression Testing
    - Integration Testing
    - UAT
    - Maintenance & Monitoring



Logging

- Record important information which later can be used to debug
- Logging Levels: (Highest Severity to Lowest) ; Set Logger Level
  - **CRITICAL:** Need to handle immediately
  - **ERROR:** There is a break in some component in code
  - **WARNING:** Not a bug but just we aware of this log
  - **INFO:** Just an information log
  - **DEBUG:** Just a debugging statement
- Multiple instances -> logs (Cloudwatch)

AWS AIML Platforms and Services



AI Services

App Developers, no ML experience required

Amazon Comprehend

Amazon Lex

Amazon Polly

Amazon Rekognition

Amazon Translate

Amazon Transcribe

Amazon Personalize

ML Services

ML Developers and Data Scientists

Amazon SageMaker

Ground Truth

Training

Algorithms

Notebooks

Hosting

Marketplace

ML Frameworks & Infrastructure

ML Researchers and Academics

Frameworks

mxnet

TensorFlow

Interfaces

GLUON

Keras

Amazon Greengrass

Amazon EC2

AWS Deep Learning AMIs

Source

Basic ways to deploy models on AWS

- Deployment architecture depends on requirements
- Different ways to deploy service on AWS:

1. Deploy with ***EC2*** service
    - Quick & Easy
    - Not a production grade - not scalable and not fault tolerant
2. Deploy with ***lambda*** (Serverless)
    - Managed, Auto Scalable, relatively cheap, Good for batch processing
    - Limit 15 min, No GPU support, hard start, Limited memory, less flexible
3. Deploy with ***Lambda + EC2***
    - Lambda triggers EC2 -> create new vm + execute task + terminate vm
    - Bad for busy real-time applications

Dockers and Containers

- **Docker:** Virtualization platform to package applications into isolated containers
- **Container:** Provides environment to run applications. It is an instance of an image
- **Image:** Image is a set of instructions written by developer to spin containers
- **Advantages:**
    - Easy to test and deploy
    - Easy to scaling
    - Versioning -> easy rollback
    - Isolation -> security
- **Challenge:** Managing multiple containers in different environments
- **Solution:** Container Orchestration tools like Kubernetes
- **AWS managed solutions:** ECS, EKS and beanstalk

Deploy using Sagemaker

- 7. **Sagemaker**
  - Designed specifically to develop, tune and deploy Machine Learning models
    - Managed Notebook (code -> Deployment instance/API)
  - Every model on sagemaker is dockerized - pull image to use it
  - Pipelines is a CI/CD/CT service that helps create and automate ML workflows
  - Sub Services:
    - Feature Store: Store and reusuage features. Ensure quality
    - Model Registry: Track versions of models

Deploy using AI services

- Less or no code use case specific AI services
  - **Comprehend:** Understand text. Detect entities, key phrases, and sentiment etc.
  - **Lex:** Build custom chatbot models
  - **Polly:** Text to Speech conversion. Supports multi-languages.
  - **Transcribe:** Speech to Text conversion service
  - **Translate:** Translate text from one language to another
  - **Personalize:** Build custom recommendation system models
  - **Rekognition:** Identify objects, people, text, activities in images and video (CV)
  - **Panorama**: Computer Vision applications on edge devices
  - **DeepLens:** Deep Learning applications on video
  - **Forecast:** Build AI applications for time-series data
  - **Omics: Lex:** Build AI models genome related data

Monitoring and Maintenance

- Models degrade over time
    - Negatively impact user experience and business value.
- Metrics for monitoring
    - Resource Metrics (SWE metrics)
        - incoming traffic, CPU/GPU memory usage or utilization, prediction latency, throughput , and cost
    - Data Metrics
        - Anomaly Check, DQ Issues, Data Drift
    - Model Metrics
        - Relationship between features (Covid: Birth and Death Rate Ratio)

Three environments for Data Science Professionals - Research, Development & Production

- Research
    - Analyze and Build prototype
- Development
    - Build MVP (Minimum Viable Product)
- Production
    - Deploy in production where actual users use service
- Resource: Medium

Conclusion and Notes

- Feasibility Check - Prototypes or Minimum Viable Products are created
  - Research, Development and production Environment
- Requirements change - hence the entire process need to be iterative
  - After feedback loops
    - Might collect more data
    - Might perform additional data quality check and resolve inconsistencies
    - Might migrate to different system design

Extras

What You Can Do Now To Protect Your Family

If you suspect that your house has lead hazards, you can take some immediate steps to reduce your family's risk:

- **If you rent, notify your landlord of peeling or chipping paint.**
- **Clean up paint chips immediately.**
- **Clean floors, window frames, window sills, and other surfaces weekly.** Use a mop or sponge with warm water and a general all-purpose cleaner or a cleaner made specifically for lead. REMEMBER: NEVER MIX AMMONIA AND BLEACH PRODUCTS TOGETHER SINCE THEY CAN FORM A DANGEROUS GAS.
- **Thoroughly rinse sponges and mop heads after cleaning dirty or dusty areas.**
- **Wash children's hands often, especially before they eat and before nap time and bed time.**
- **Keep play areas clean.** Wash bottles, pacifiers, toys, and stuffed animals regularly.
- **Keep children from chewing window sills or other painted surfaces.**
- **Clean or remove shoes before entering your home to avoid tracking in lead from soil.**
- **Make sure children eat nutritious, low-fat meals high in iron and calcium,** such as spinach and dairy products. Children with good diets absorb less lead.

Reducing Lead Hazards In The Home

Removing lead improperly can increase the hazard to your family by spreading even more lead dust around the house.

Always use a professional who is trained to remove lead hazards safety.

In addition to day-to-day cleaning and good nutrition:

- You can **temporarily** reduce lead hazards by taking actions such as repairing damaged painted surfaces and planting grass to cover soil with high lead levels. These actions (called "interim controls") are not permanent solutions and will need ongoing attention.
- To **permanently** remove lead hazards, you must hire a certified lead "abatement" contractor. Abatement (or permanent hazard elimination) methods include removing, sealing, or enclosing lead-based paint with special materials. Just painting over the hazard with regular paint is not permanent removal.

Always hire a person with special training for correcting lead problems-someone who knows how to do this work safely and has the proper equipment to clean up thoroughly. Certified contractors will employ qualified workers and follow strict safety rules as set by their state or by the federal government.

Once the work is completed, dust cleanup activities must be repeated until testing indicates that lead dust levels are below the following:

- 40 micrograms per square foot (?g/ft²) for floors, including carpeted floors;
- 250 ?g/ft² for interior windows sills; and
- 400 ?g/ft² for window troughs.

Call your state or local agency (see bottom of page 11) for help in locating certified professionals in your area and to see if financial assistance is available.

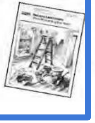
Remodeling or Renovating a Home With Lead-Based Paint

Take precautions before your contractor or you begin remodeling or renovations that disturb painted surfaces (such as scraping off paint or tearing out walls):

- **Have the area tested for lead-based paint.**
- **Do not use a belt-sander, propane torch, heat gun, dry scraper, or dry sandpaper** to remove lead-based paint. These actions create large amounts of lead dust and fumes. Lead dust can remain in your home long after the work is done.
- **Temporarily move your family** (especially children and pregnant women) out of the apartment or house until the work is done and the area is properly cleaned. If you can't move your family, at least completely seal off the work area.
- **Follow other safety measures to reduce lead hazards.** You can find out about other safety measures by calling 1-800-424-LEAD. Ask for the brochure "Reducing Lead Hazards When Remodeling Your Home." This brochure explains what to do before, during, and after renovations.

If you have already completed renovations or remodeling that could have released lead-based paint or dust, get your young children tested and follow the steps outlined in the section labeled "What you can do now to protect your family."

If not conducted properly, certain types of renovations can release lead from paint and dust into the air.



Other Sources of Lead

While paint, dust, and soil are the most common lead hazards, other lead sources also exist.

- **Drinking water.** Your home might have plumbing with lead or lead solder. Call your local health department or water supplier to find out about testing your water. You cannot see, smell, or taste lead, and boiling your water will not get rid of lead. If you think your plumbing might have lead in it:
  - Use only cold water for drinking and cooking.
  - Run water for 15 to 30 seconds before drinking it, especially if you have not used your water for a few hours.
- **The job.** If you work with lead, you could bring it home on your hands or clothes. Shower and change clothes before coming home. Launder your work clothes separately from the rest of your family's clothes.
- Old painted **toys** and **furniture**.
- Food and liquids stored in **lead crystal** or **lead-glazed pottery** or **porcelain**.
- **Lead smelters** or other industries that release lead into the air.
- **Hobbies** that use lead, such as making pottery or stained glass, or refinishing furniture.
- **Folk remedies** that contain lead, such as "greta" and "azarcon" used to treat an upset stomach.