2024

Building ML solutions:

From System Design to Deployment

Alexander Guschin
Mikhail Rozhkov

OUTLINE

1. Introduction to ML System Design
2. ML System Design Doc
3. Design ML Product
4. Problem Statement
5. Customers & Value Proposition
6. Business Metrics (Success)

MODULE DETAILS

Goals

- Understand the critical role of system design in successful ML projects
- Learn to bridge the gap between business needs and technical solutions
- Develop skills to effectively communicate with stakeholders

Learning Outcomes

- Grasp the differences between system design and ML system design.
- Be able to brief a client effectively.
- Understand how to set measurable goals and success metrics.
- Collect and formalize requirements for an ML project.

HANDS-ON ACTIVITIES

Exercise

- Stakeholder interview simulation: gather information about the business problem, stakeholders and requirements.
- Goal: Understand requirements for ML Product

Output

1. Interviewing a client to gather requirements.
2. Drafting the first part of a design document: **ML Product Design**
    a. Problem Statement (Motivation)
    b. Value Proposition
    c. Customers
    d. Business Metrics (Success)
    e. Assumptions and Constraints

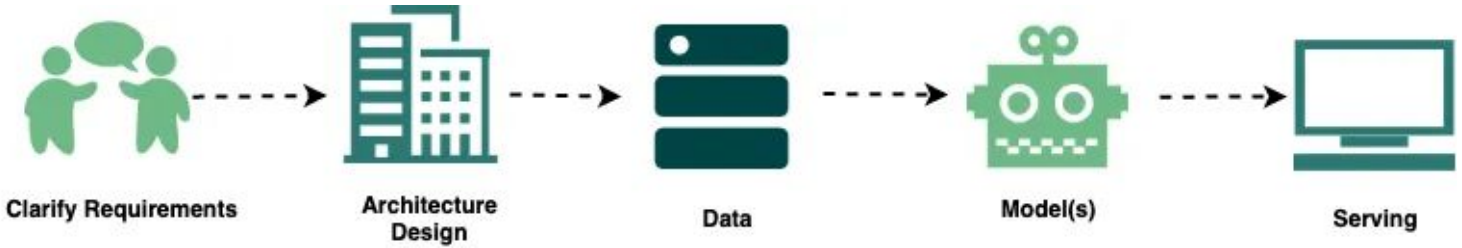
Introduction to ML System Design

System design <> ML system design

Introduction to ML System Design

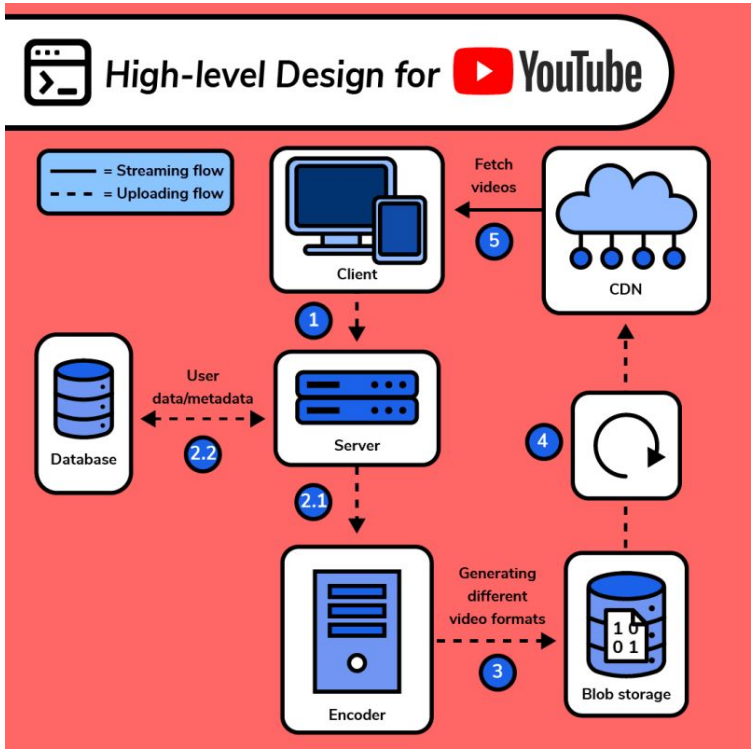
I'm a designer now!

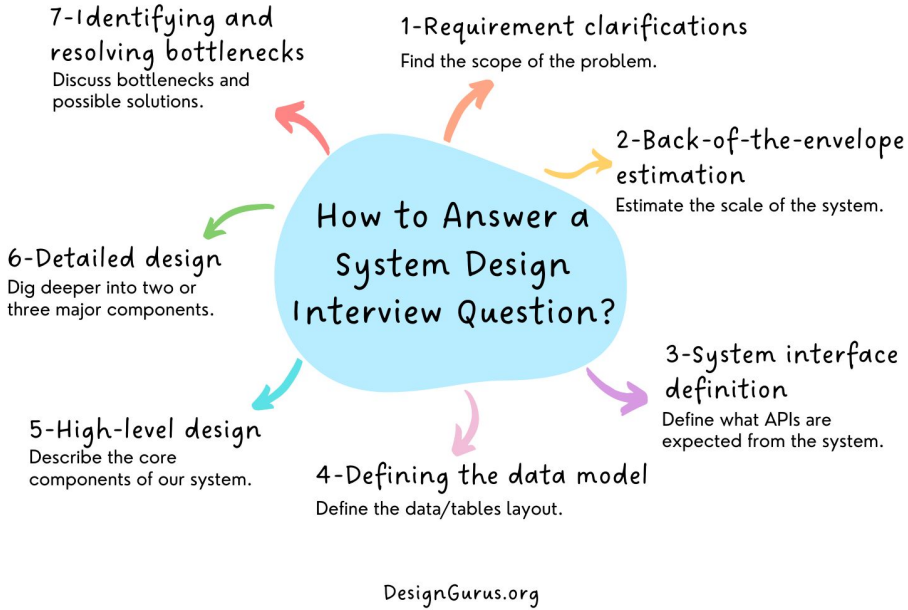
- What is ML system design?
- Differences between system design and ML system design.



Source: How To Answer Any Machine Learning System Design Interview Question

What is System Design?



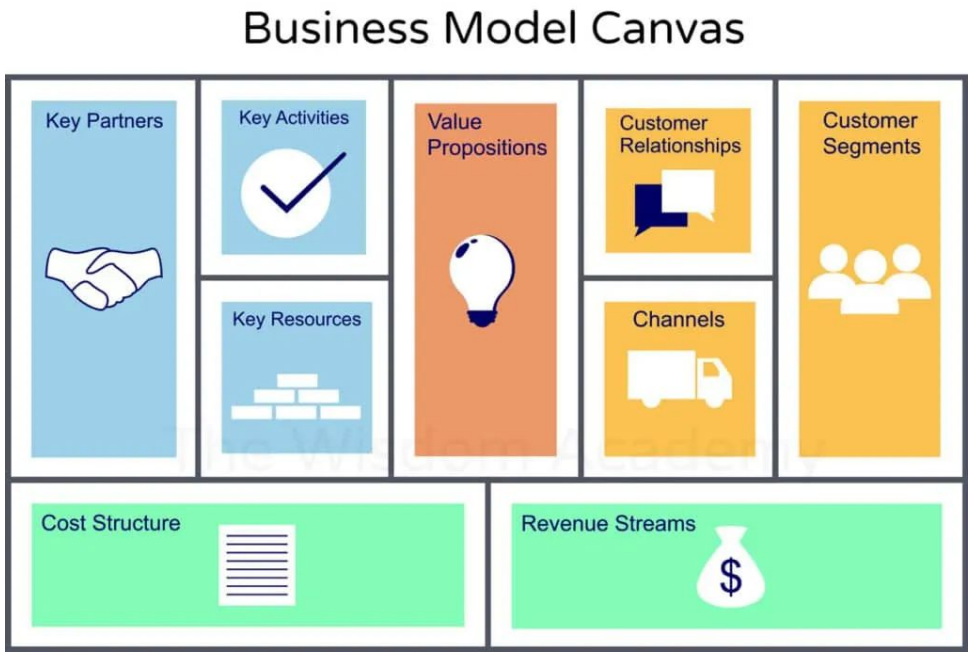


Source:
https://www.educative.io/courses/grokking-modern-system-design-interview-for-engineers-managers

ML System Design Document

System design <> ML system design

Business Model Canvas is a way to understand your business

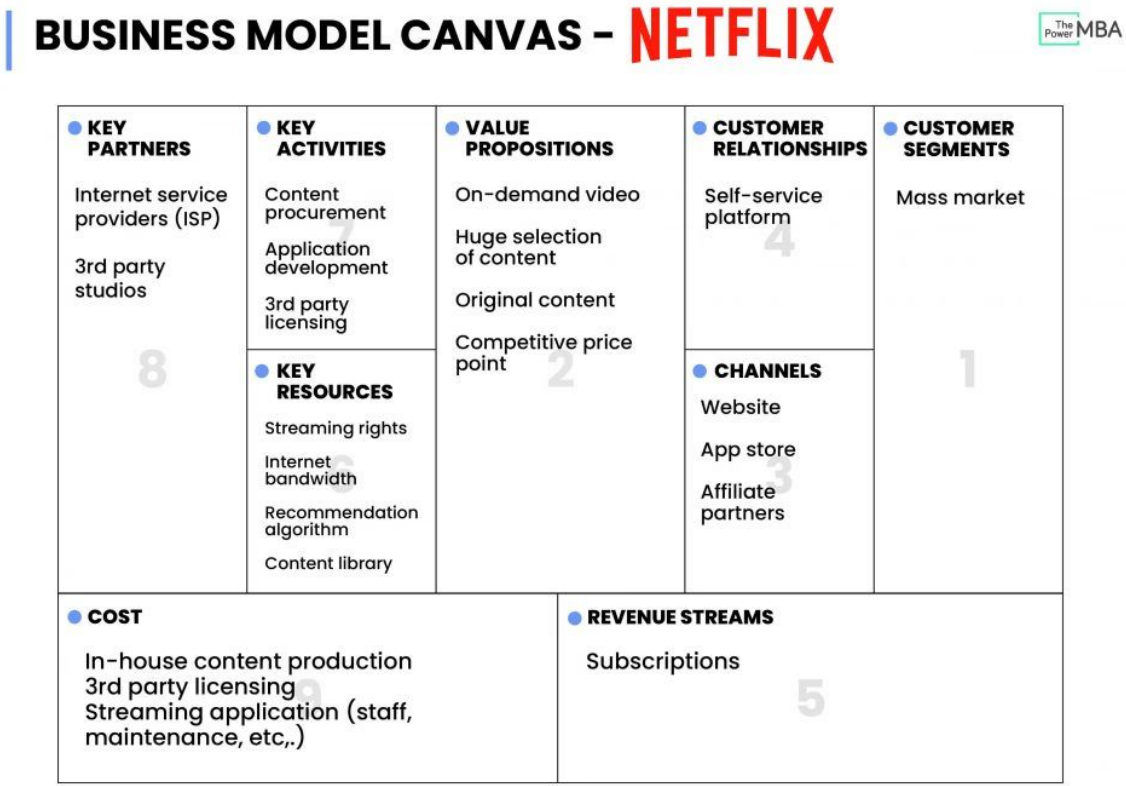


- Simple
- Informative
- Efficient

Source:

- https://medium.com/@niloal361/what-is-the-business-model-canvas-e0f3e7816a4f
- https://hbr.org/2013/05/a-better-way-to-think-about-yo

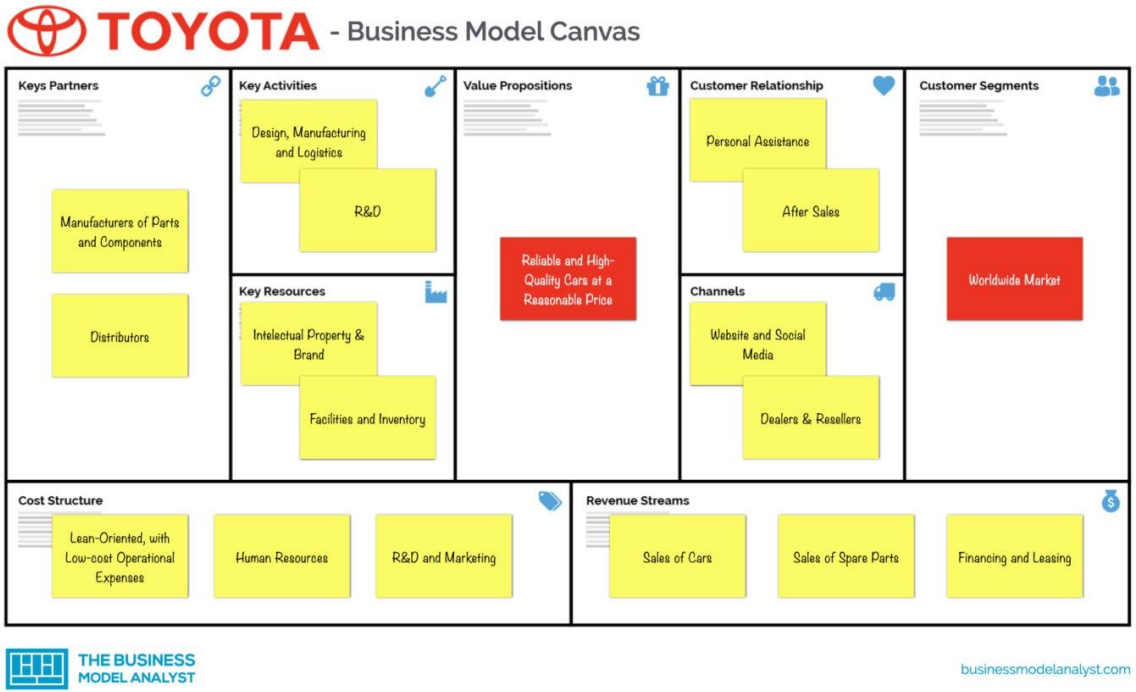
Example: Netflix



Source:

- https://global.thepower.education/blog/business-model-canvas

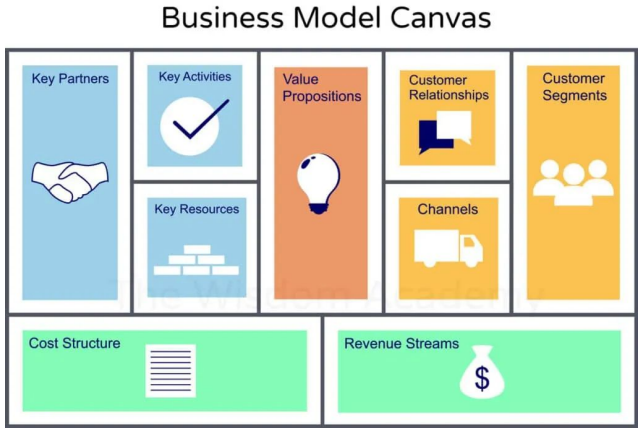
Example: Toyota



Source:

- https://www.upgrad.com/blog/business-model-canvas-explained-with-examples/

Can it be useful for ML/AI projects?

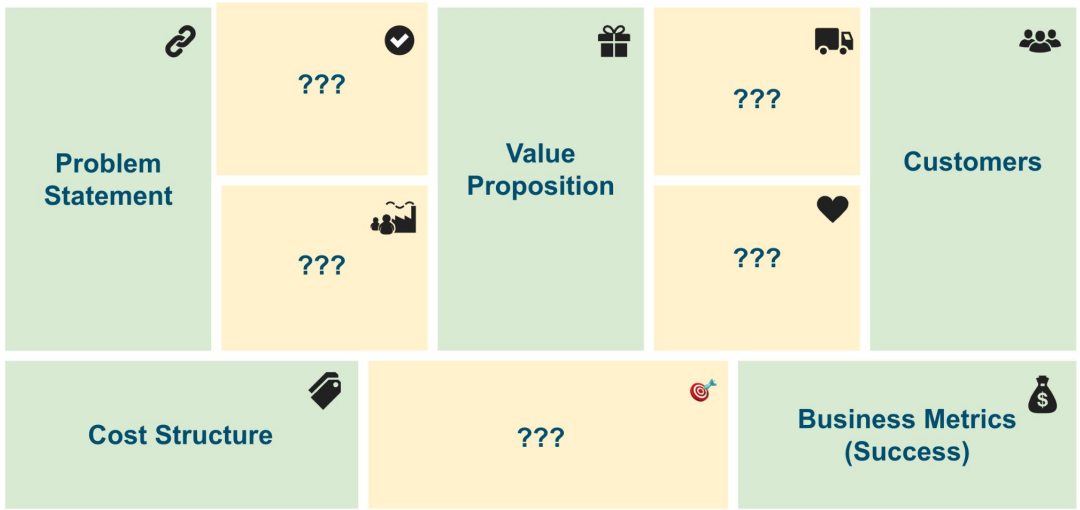


- Clear problem statement
- Proposed ML solution
- Expected business impact
- Operational requirements

Source:

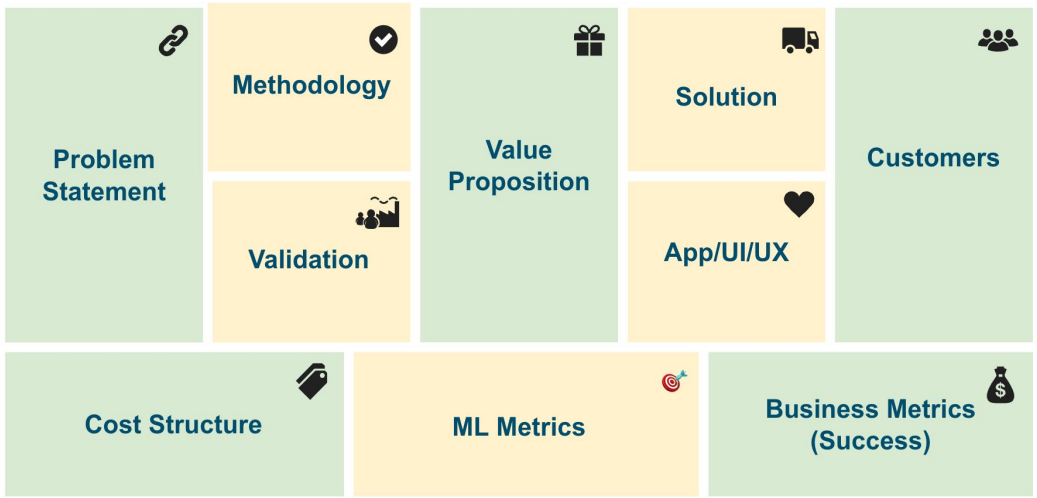
- https://madewithml.com/courses/mlops/product-design/

ML goal is to drive the business impact



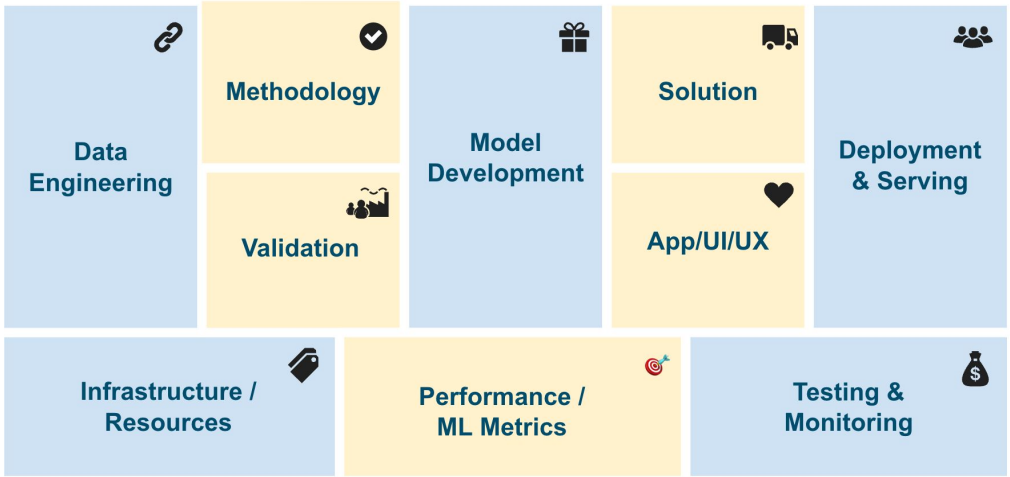
- Who is the customer?
- What problem do they have?
- How ML can help?
- How to measure business impact?

1 - Design ML Product to achieve the business goals



- What is the ML solution look like?
- How can prove it works?
- How to deliver the value?
- How do we measure the effect?

2 - Design ML System to enable the ML Product



- How do we prepare the data?
- How do we train and evaluate the model?
- How do we deploy the ML system?
- How do we operate and monitor the ML system?

Design ML Product

> ML Product Design

> Guide: 1 - Overview and Impact

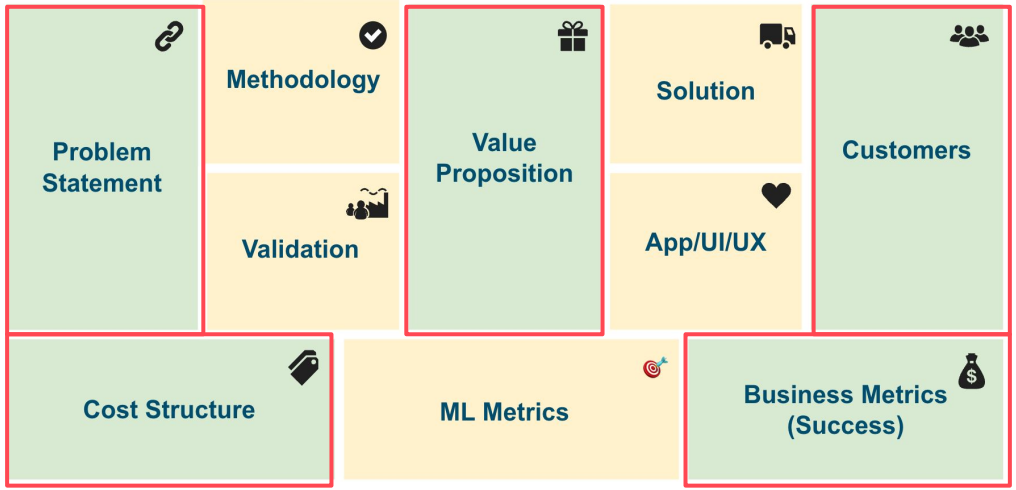
2.x – ML Product Design: Business Understanding

Purpose

- Define business problem & metrics

Guiding questions:

- Is it the right problem to solve?



Case: EasyRide Taxi – Predict Arrival Time

Purpose

- Discuss every step of the ML System Design process
- Example for your own projects

Example

Overview: Purpose and Impact

Overview

- This project aims to improve the prediction accuracy of taxi trip durations for our company, EasyRide. We currently rely on an external provider with a MAPE > 30%.

Key points:

- **Problem:** High MAPE in current trip duration predictions.
- **Solution:** Develop an in-house ML model.
- **Business Impact:** Reduce revenue loss and customer churn by improving prediction accuracy.
- **Timeline:** 1 week for POC, 1 week for testing, decision point thereafter.

Example

ML-powered trip duration prediction for EasyRide Taxi optimizes dispatching, improves pricing accuracy, and reduces customer churn, boosting revenue and satisfaction.

Problem Statement

> ML Product Design

> Guide: 2.1 - Problem Statement

2.1 – Problem Statement (Motivation)

Purpose

- Clearly define the business problem and its relevance to the organization.

Guiding questions:

- Why the problem is important to solve, and why now?
- What are the costs of not solving this problem?
- How does this align with our overall business strategy?

Case: NewPizza – long waiting time

What is the problem?

Overview

- NewPizza, a franchise pizzeria, experiences challenges with managing peak times, leading to long wait times in queues. The goal is to predict customer queues a week in advance to optimize staff schedules and reduce service time.
- Accurate queue prediction allows for efficient staff allocation, improving customer experience by minimizing wait times and enhancing operational efficiency.
- Currently, the inability to predict queues results in suboptimal staffing, longer wait times for customers, and potential loss of sales due to customer dissatisfaction.

Example

Key points:

- **Problem:** Significant variations in queue times, with more than 10% of customers waiting over 5 minutes.
- **Current Approach:** Manual planning of staffing by the owner.
- **Industry Context:** Fast-paced food service where quick turnaround is essential for customer satisfaction.
- **Alignment:** Directly supports NewPizza's goal of enhancing customer service by ensuring a fast and efficient ordering process.

Exercise: Client Interview

What information do we need? Who should we ask?

Group task:

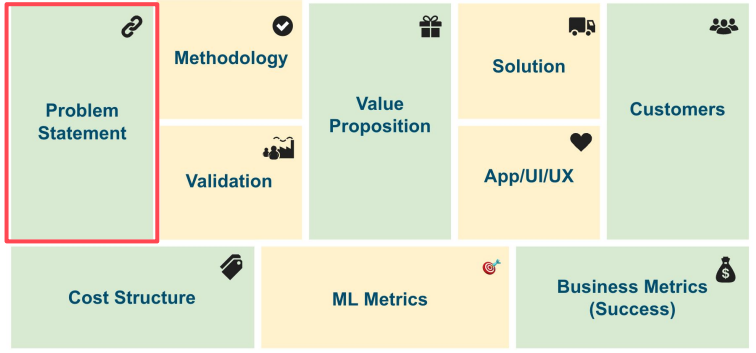
- Copy Practice - EasyRide Taxi - PUBLIC for your team
- Conduct an interview with a "client" using a template on the next slide
- Understand the problem and make a draft of the “2.1 - Problem Statement” description
- 15 min

Practice

Team 1 - Michael Lee
Team 2 - Emily Chen
Team 3 - David Wong

Key points:

- Detailed description of the business problem
- Current approaches and their limitations
- Market or industry context
- Alignment with business strategy



Case: EasyRide Taxi – Predict Arrival Time

What is the problem?

Overview

- EasyRide Taxi, a leading ride-hailing service in New York City, faces challenges in efficiently dispatching taxis and providing accurate estimated arrival times (ETAs) to customers.
- Accurate trip duration prediction is crucial for optimizing fleet management, improving customer satisfaction, and maximizing driver utilization.

Key points:

- **Problem:** High MAPE > 30%.
- **Current Approach:** External provider's service.
- **Industry Context:** Competitive market with accurate pricing as a differentiator.
- **Alignment:** Critical to EasyRide's strategy of superior customer service.

Example

ML Product Design: EasyRide Taxi

Problem Statement

- High MAPE > 30%.
- External provider's prediction service (we can't improve).
- Competitive market with accurate pricing as a differentiator.
- Critical to EasyRide's strategy of superior customer service.

Methodology

-
-

Validation

-
-

Value Proposition

-
-

Solution

-
-

App/UI/UX

-
-

Customers

-
-

Cost Structure

-
-

Performance / ML Metrics

-
-

Business Metrics (Success)

-
-

Customers & Value Proposition

> ML Product Design
> Guide: 2.2 - Customers
> Guide: 2.3 - Value Proposition

2.2 - Customers

Who are our customers?

Purpose

- To ensure all relevant perspectives are considered and to clarify who will be using or impacted by the system.

Guiding questions:

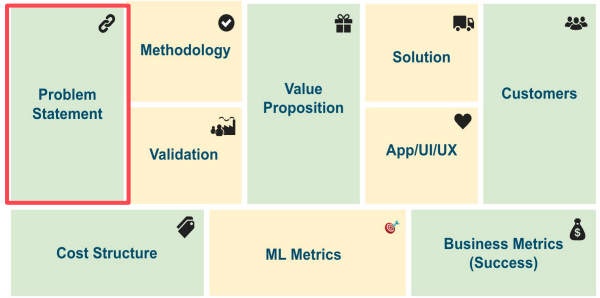
- Who will be directly using the ML system?
- Whose work or processes will be affected by the system?
- Who needs to be involved in the decision-making process?

2.3 – Value Proposition

Why AI/ML is required?

Purpose

- To justify the use of AI/ML over traditional approaches and highlight its unique benefits.



Guiding questions:

- How does AI/ML solve this problem better than traditional methods?
- What new capabilities does AI/ML bring to our business?
- How does this solution position us for future growth?
- Why AI/ML is required?

Exercise: Customers & Value Proposition

What information do we need? Who should we ask?

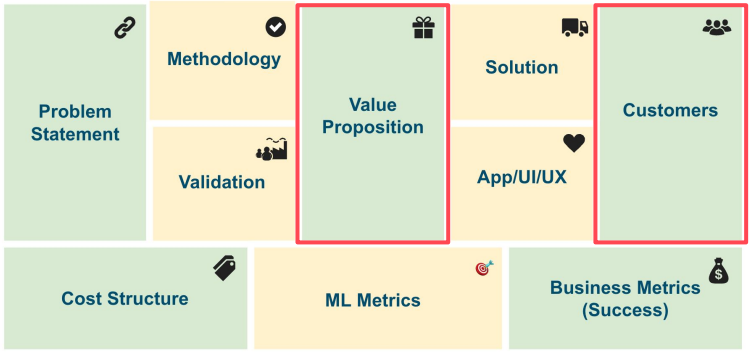
Group task:

- Brainstorm and complete the ML Product Design sections:
  - 2.2 - Customers
  - 2.3 - Value Proposition
- 5 min

Key points:

- Who are our customers (stakeholders)?
- Primary end users and their needs
- Potential impact on each group
- Unique advantages of using AI/ML
- Potential improvements over current methods

Practice



Customers

Who are our customers?

Brainstorm on customers and their pains:

- **Taxi app customers**: unpredictable fares, long wait times, Inaccurate ETAs, lack of pricing transparency.
- **Taxi drivers**: inefficient trip allocations, excessive idle time, unpredictable earnings, wasted fuel during empty rides.
- **Business Leaders:** revenue loss, difficulty in maintaining market share, customer churn.
- **Internal Teams**: inaccurate dispatching decisions, lack of real-time data insights, difficulty in optimizing operations, challenges in improving service quality.

Example

Key points:

- **Taxi app customers**
- **Taxi drivers**

Notes:

- ETA stands for Estimated Time of Arrival.
- Instead of Customer Pains you may want to write down Customer Needs.

Value Proposition

Why AI/ML is required?

EasyRide's AI-powered trip duration prediction offers:

- Enhances pricing accuracy, minimizing revenue loss from over/underpricing
- Optimizes dispatching efficiency, reducing idle time and fuel costs
- Improves customer retention by providing more reliable ETAs
- Increases driver satisfaction through better trip allocation
- Adapts continuously to NYC's dynamic urban environment

Key points:

- **Minimize revenue loss**
- **Improve customer retention**
- **Improve driver retention**

Example

ML Product Design: EasyRide Taxi

Problem Statement

- High MAPE > 30%.
- External provider's prediction service (we can't improve).
- Competitive market with accurate pricing as a differentiator.
- Critical to EasyRide's strategy of superior customer service.

Methodology

-
-

Validation

-
-

Value Proposition

- Minimize revenue loss
- Improve customer retention
- Improve driver retention

Solution

-
-

App/UI/UX

-
-

Customers

- Taxi app customers
- Taxi drivers

Cost Structure

-
-

Performance / ML Metrics

-
-

Business Metrics (Success)

-
-

Business Metrics (Success)

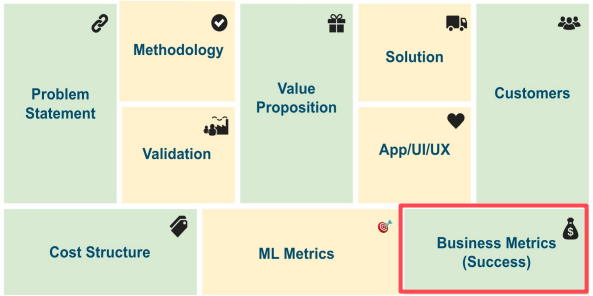
> ML Product Design
> Guide: 2.4 - Business Metrics (Success)
> Guide: 2.5 - Cost Structure & ROI
> Guide: 2.6 - Assumptions and Constraints

2.4 – Business Metrics (Success)

How do we measure success?

Purpose

- To establish clear, quantifiable goals that align business objectives with technical performance.



Guiding questions:

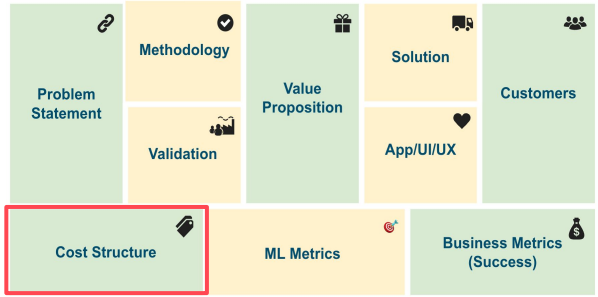
- How will we measure the success of this ML system?
- What metrics align with our business objectives?
- How do we balance technical and business performance?

2.5 – Cost Structure & ROI

How much does it cost?

Purpose

- To justify the investment in the ML system and set realistic expectations for financial returns



Hints:

- Estimate costs after a Solution draft is complete

Guiding questions:

- What are all the costs associated with this project?
- How the costs will change with time?
- When do we expect to see a return on our investment?
- Consider cost in money and in time spent by dev/ml specialists

Exercise: Business Metrics

How do we measure success?

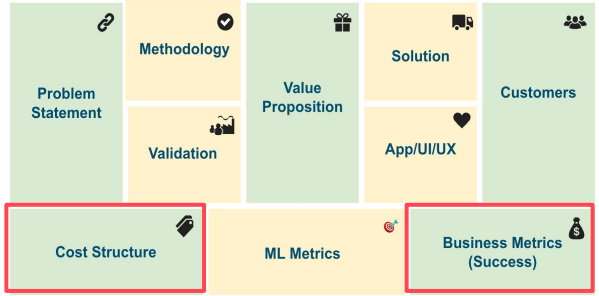
Group task:

- Brainstorm and complete the ML Product Design sections:
  - 2.4 - Business Metrics (Success)
  - 2.5 - Cost Structure & ROI (name key factors that affect costs)
- 10 min

Key points:

- Specific, quantifiable business metrics
- Name key factors that affect costs

Practice



Business Metrics (Success)

How do we measure success?

Success is measured by improved dispatching efficiency, pricing accuracy, and reduced customer churn. Primary Metrics and improvements:

1. **Daily Revenue Increase** (by dispatching efficiency Improvement):
    a. Current: $0 (baseline)
    b. Target: $24,000 per day
2. **Pricing Loss Reduction** (by pricing accuracy improvement):
    a. Current: $35,500 per day
    b. Target: $18,000 per day
3. **Booking Rate Improvement** (by customer churn reduction):
    a. Current: 88%
    b. Target: 94%

Key points:

- Daily Revenue Increase by $24,000
- Pricing Loss Reduction by $17,000
- Booking Rate Improvement by 6%
- Timeline: Evaluate metrics daily, with quarterly reviews

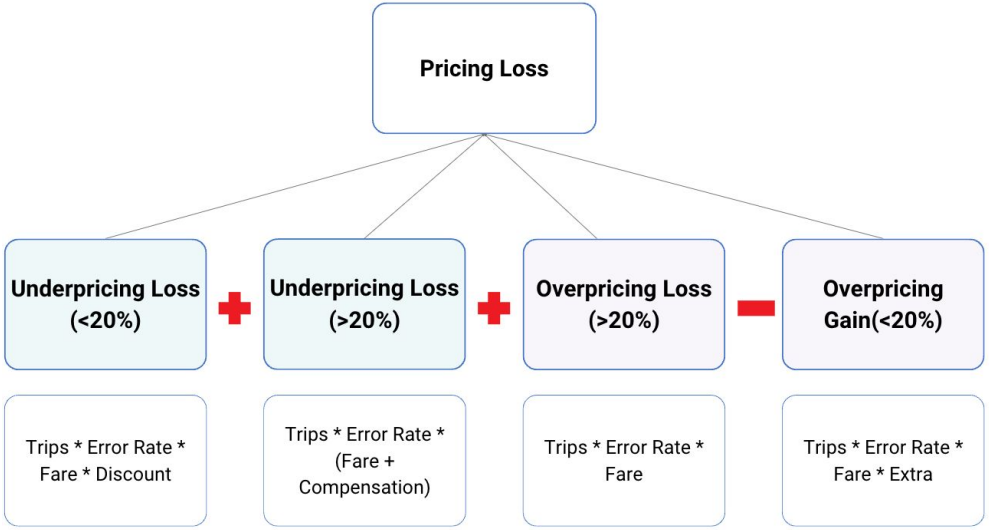
Example

Example Formulas

How do we calculate metrics?

Pricing Accuracy Improvement:

- **Pricing Loss Reduction =** Current Pricing Loss - New Pricing Loss



Example

Assumptions & Expectations

How do we measure success?

Assumptions:

- Average fare per trip: $20
- Average trip duration: 20 minutes
- Price per minute: $1
- Number of trips per day: 10,000
- Current average idle time: 20 minutes
- Fuel cost per hour: $3
- Taxi driver salary per hour: $15
- Current external provider's MAPE: 30%
- Overpricing Rate (>20%) = 0.10
- Underpricing Rate (<20%) = 0.20

Expectations of ML Solution Implementation and Operations:

- Model expected accuracy improvement: 15%
- Idle time reduction due to better prediction: 10%
- ML model's expected MAPE: 15%

- Initial Development Cost: $214,083
- Annual Operational Cost: $386,000

Example

Cost Structure & ROI

How do we measure success?

Costs involve initial development and ongoing operations. Substantial financial benefits expected from improved dispatching and pricing accuracy

1. **Costs:**
    a. Initial Development: $214,083
    b. Annual Operations: $386,000
    c. First Year Total Cost: $600,083 (Initial Development + First Year Operation)
    d. Subsequent Annual Cost: $386,000
2. **Financial Benefits:**
    a. Annual Benefit: $15,147,500
        - Dispatching: $8,760,000
        - Pricing: $6,387,500
3. **ROI:**
    a. First Year: 2,424%
    b. Subsequent Years: 3,825%

Example

Key points:

- Initial Development: $214,083
- Annual Operations: $386,000
- Annual Benefit: $15,147,500
- ROI:
    - First Year: 2,424%
    - Subsequent Years: 3,825%

ML Product Design: EasyRide Taxi

Problem Statement

- High MAPE > 30%.
- External provider's prediction service (we can't improve).
- Competitive market with accurate pricing as a differentiator.
- Critical to EasyRide's strategy of superior customer service.

Methodology

-
-

Validation

-
-

Value Proposition

- Minimize revenue loss
- Improve customer retention
- Improve driver retention

Solution

-
-

App/UI/UX

-
-

Customers

- Taxi app customers
- Taxi drivers

Cost Structure

- Initial Development: $214,083
- Annual Operations: $386,000
- Annual Benefit: $15,147,500
- ROI: First Year: 2,424% / Subsequent Years: 3,825%

Performance / ML Metrics

-
-

Business Metrics (Success)

- Daily Revenue Increase by $24,000
- Pricing Loss Reduction by $17,000
- Booking Rate Improvement by 6%
- Evaluate metrics daily, with quarterly reviews

Project Assignment

Start with ML System Design!

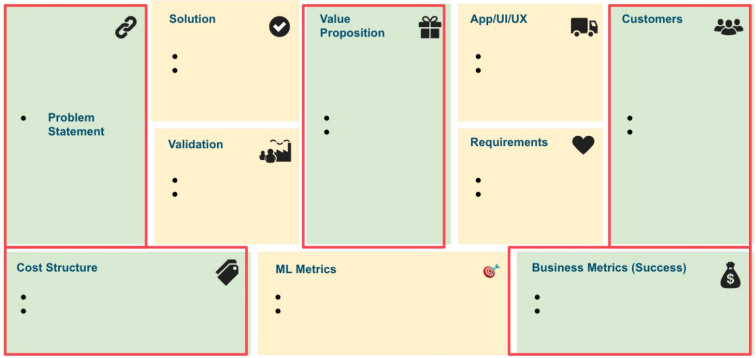
Prerequisites

- Select a project to work on during the week 1
  - Your own project
  - Join other project (max 2 people per team)

Business Problem Understanding

ML Product Design

1. Describe the business problem for your project
2. Follow the guide to describe each section:
    - 1 - Overview: Purpose and Impact
    - 2.x - ML Product Design: Business Understanding
3. Update the ML Product Design Canvas



Materials & Links

- Course Materials: Google Drive
- Practice - EasyRide Taxi - PUBLIC
- Guide - ML System Design - Canvas