UNLEASHING ML POTENTIAL

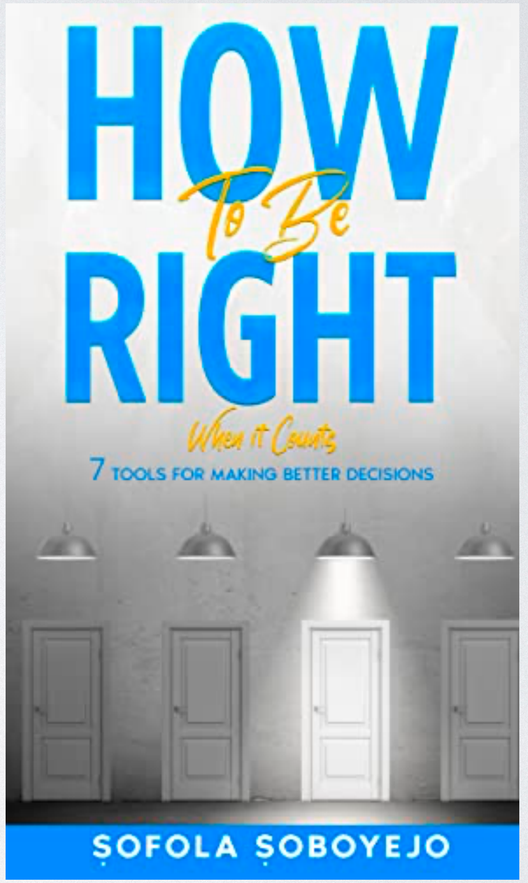
Essential Design Patterns and MLOps Strategies for Business Solutions

Sho Fola Soboyejo

@shoreason linkedin.com/in/shofola

INTRO

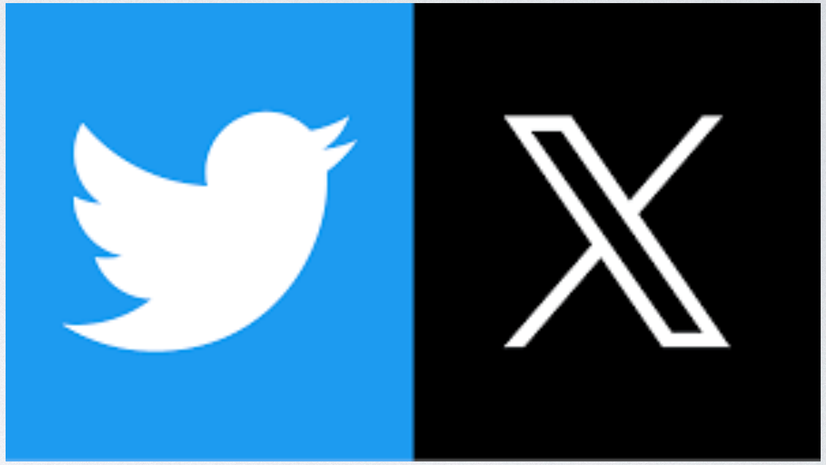
- VP Engineering for Data at *84.51*
- Industry Experience in Banking, Retail, Marketing, Advertising and Social Media
- Author of "*How to be Right, When it Counts*"



INTRO

- Customer segmentation
- Creative brand assistant

- Audience Targeting
- Modeling Attribution



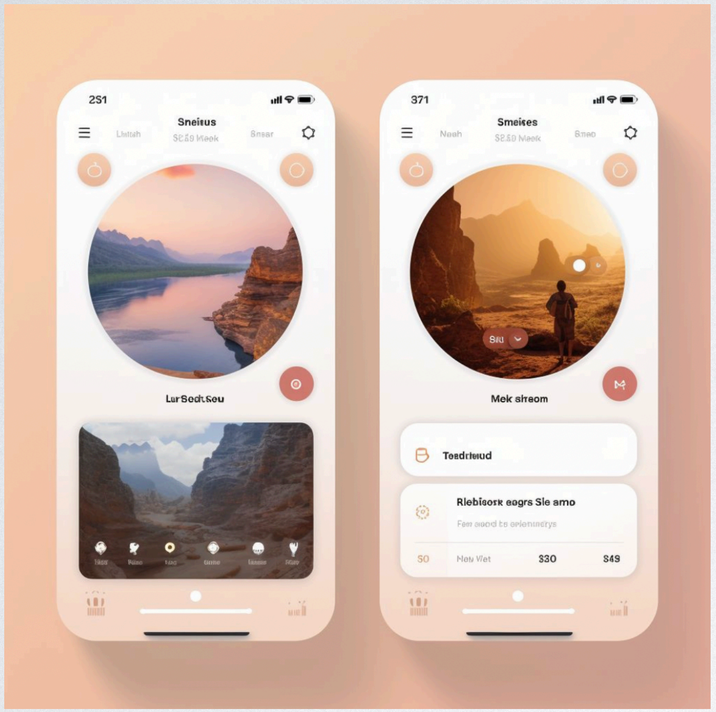
- Home timeline relevant tweets
- Managing Content

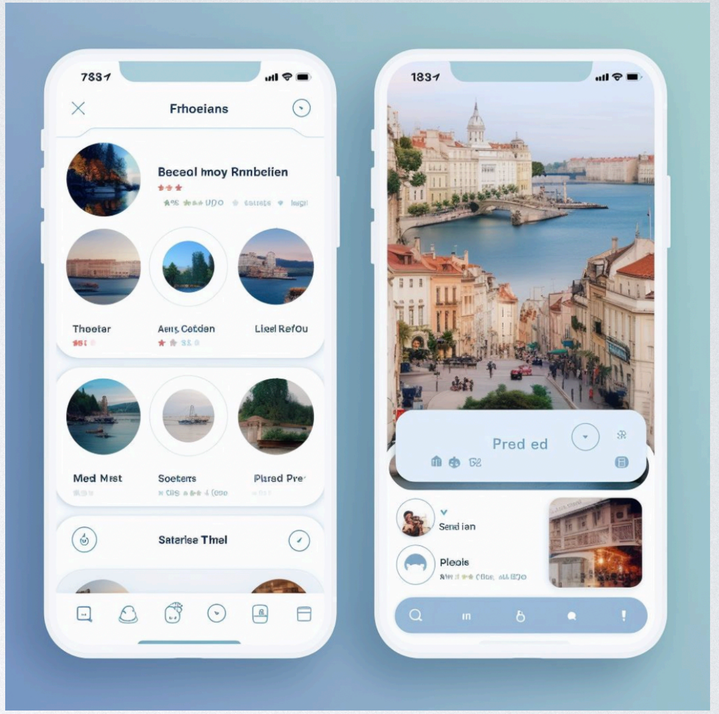
GEN AI

CURRENT STATE OF ML

Created using Bing and Dall-E

CURRENT STATE OF ML

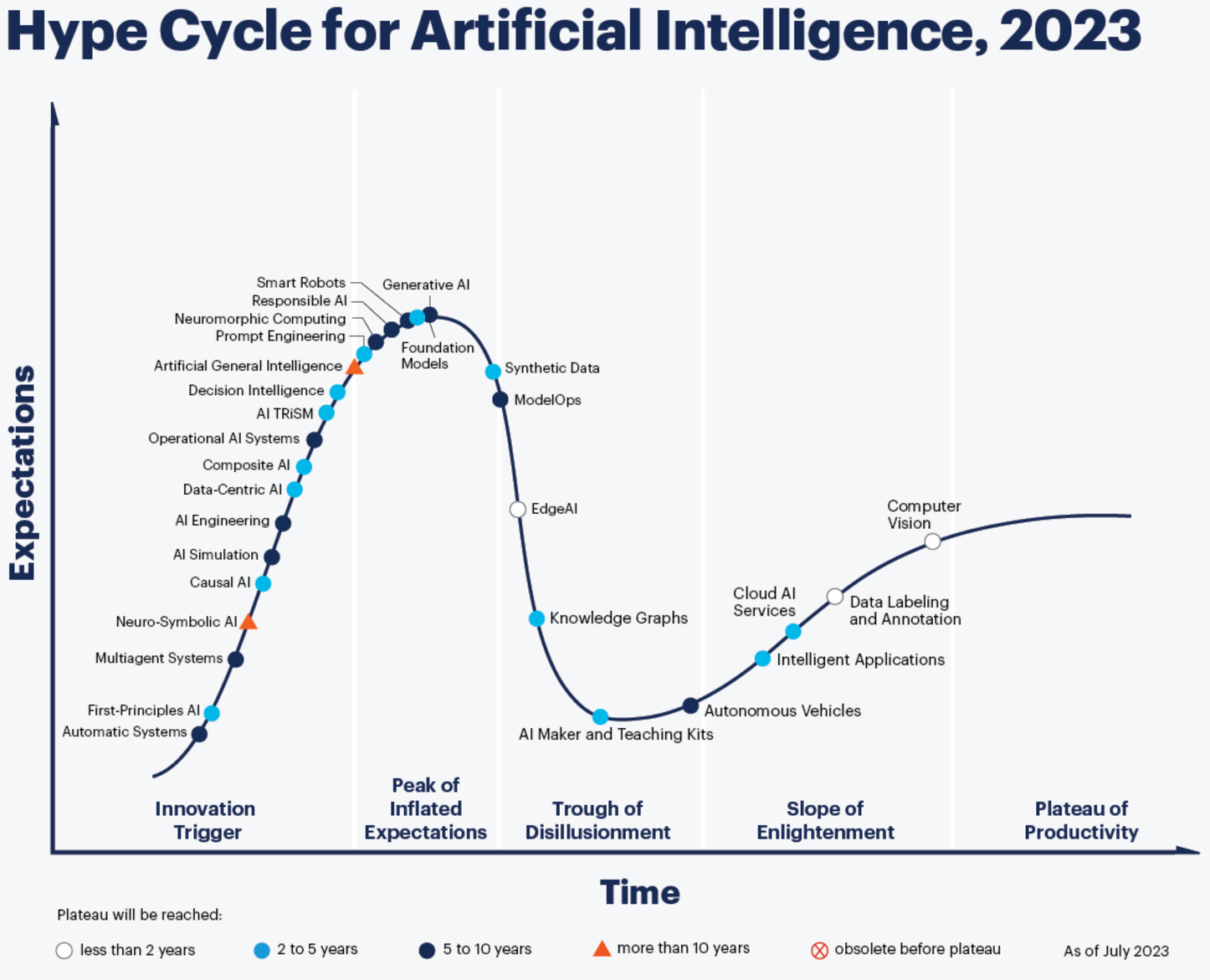




Created using Midjourney

ON THE HORIZON

- ModelOps



gartner.com

Gartner

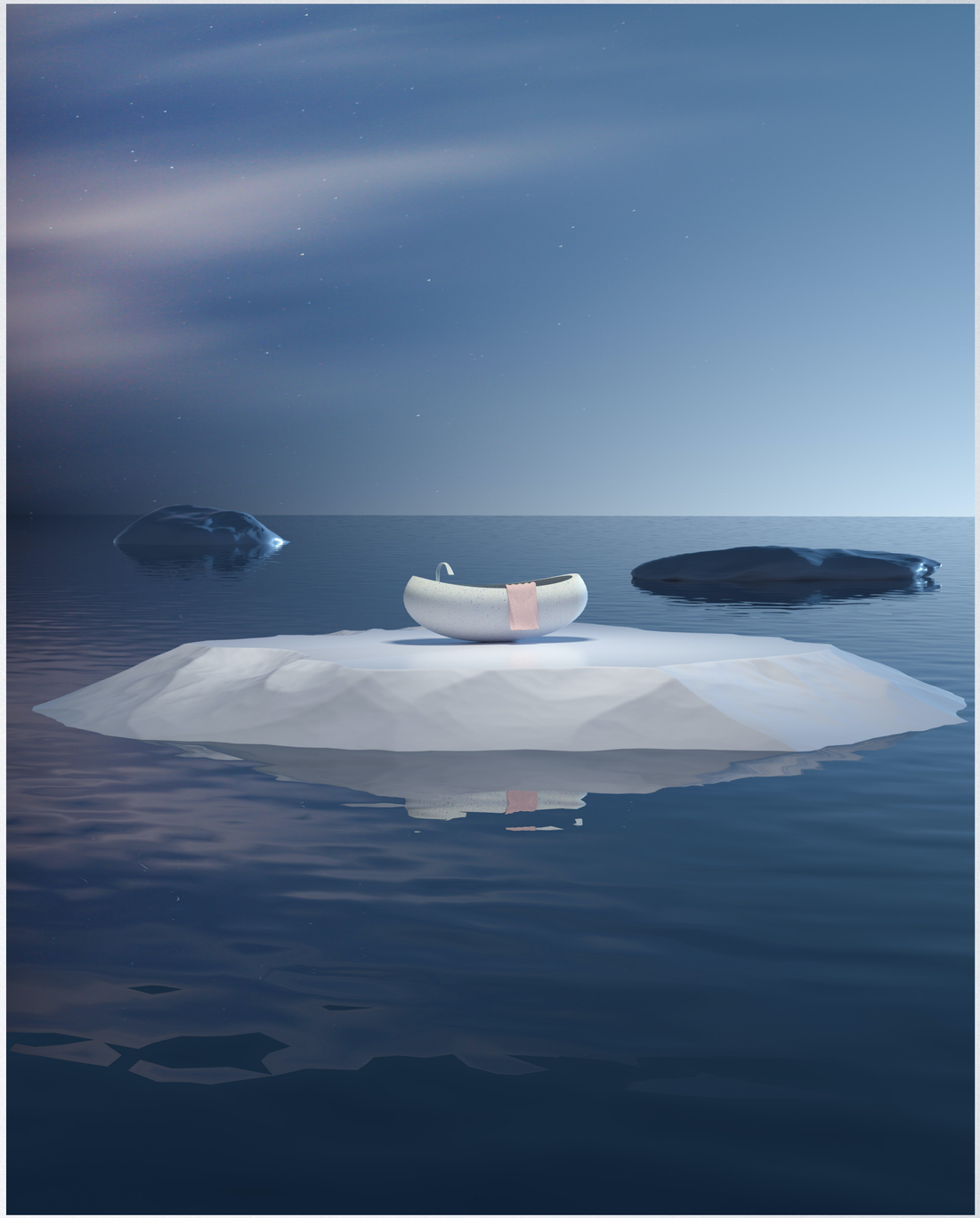
THE REAL JOURNEY

54%

The Gartner survey revealed that on average, 54% of AI projects make it from pilot to production. This is a slight increase from the Gartner 2019 AI in Organizations Survey, which reported an average of 53% of AI projects that make it to production.

BEYOND MODELING AND UX

- Scalability
- Maintainability
- Monitoring & Observability
- Security
- Integration



BEYOND MODELING AND UX

Source: The Verge

Source: CNN Business

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr Follow

Updated July 1, 2015 at 3:41 pm ET

Source: WSJ

How a Self-Driving Uber Killed a Pedestrian in Arizona

By TROY GRIGGS and DAISUKE WAKABAYASHI UPDATED MARCH 21, 2018

Source: NY Times

BEYOND MODELING AND UX

- No Focus on Models: LLM > GPT-4, LLaMA, PaLM, Claude, Falcon
- No Focus on Interfaces: Chat, Voice, Text, Images or Video

ML DESIGN PATTERNS

"You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future "

- Steve Jobs (Stanford 2005 Commencement Speech)

WHAT ARE THEY?

- A **software design pattern** is a general, reusable solution to a commonly occurring problem within a given context in software design - Wikipedia
- A template for how to solve a problem
- Singleton, Decorator, Command, Factory

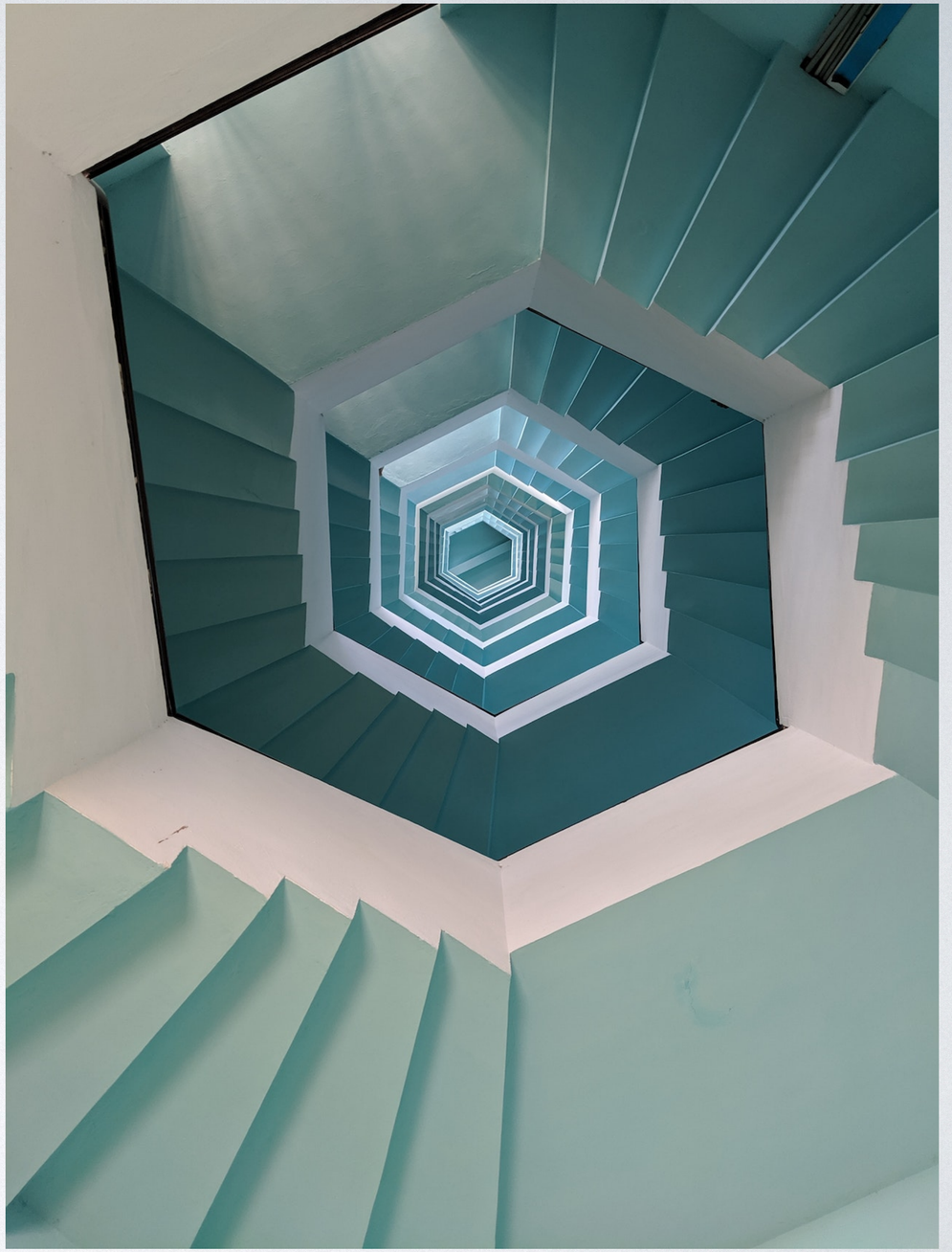


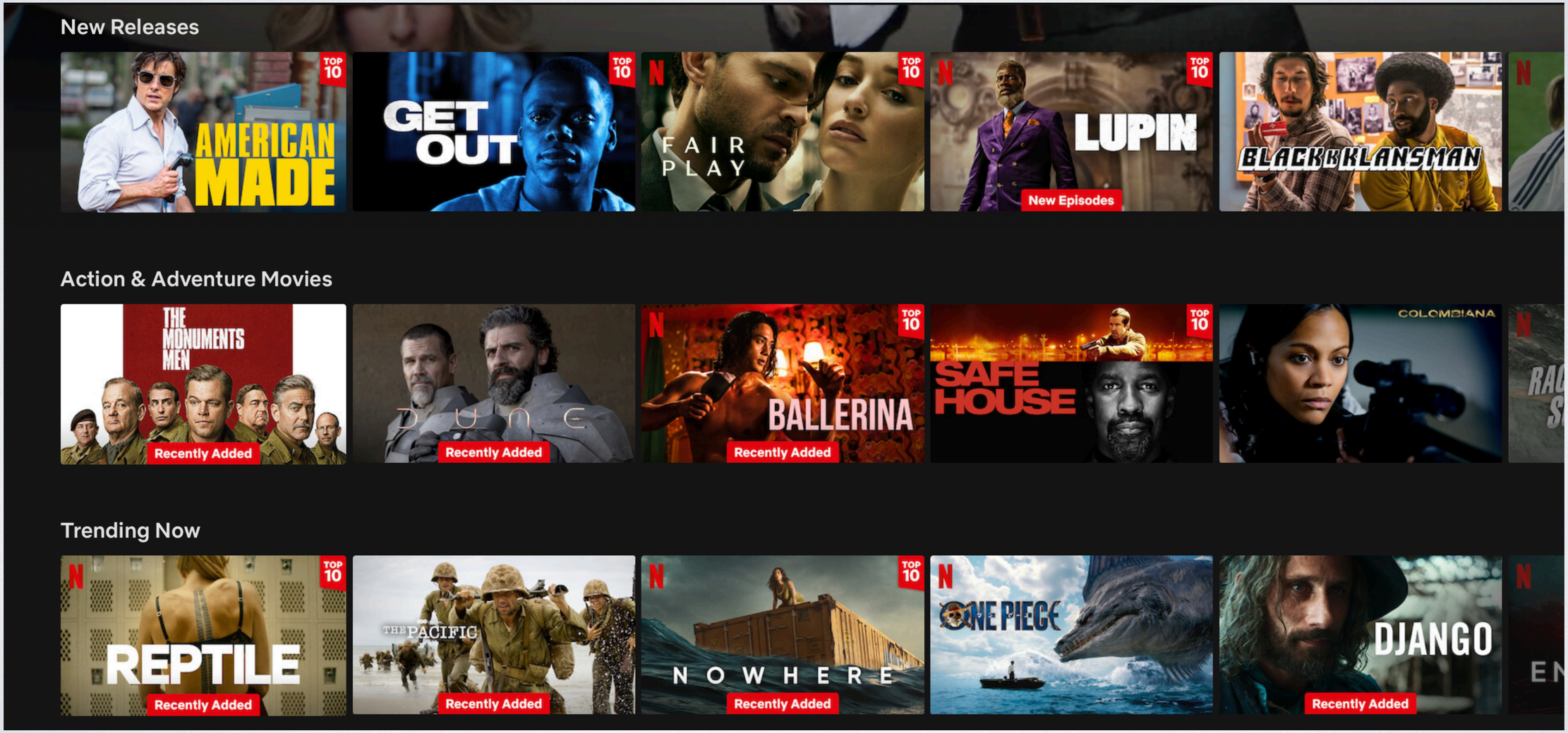
Photo by Mithul Varshan

ML FIST OF FURY?

- Recommender Systems
- Feed Ranking
- Ad Predictions
- Search Systems
- Entity Linking

Created using Dall-E

RECOMMENDER SYSTEMS



RECOMMENDER SYSTEMS

- Problem Statement: Display Recommendations for a User of [X]

RECOMMENDER SYSTEMS

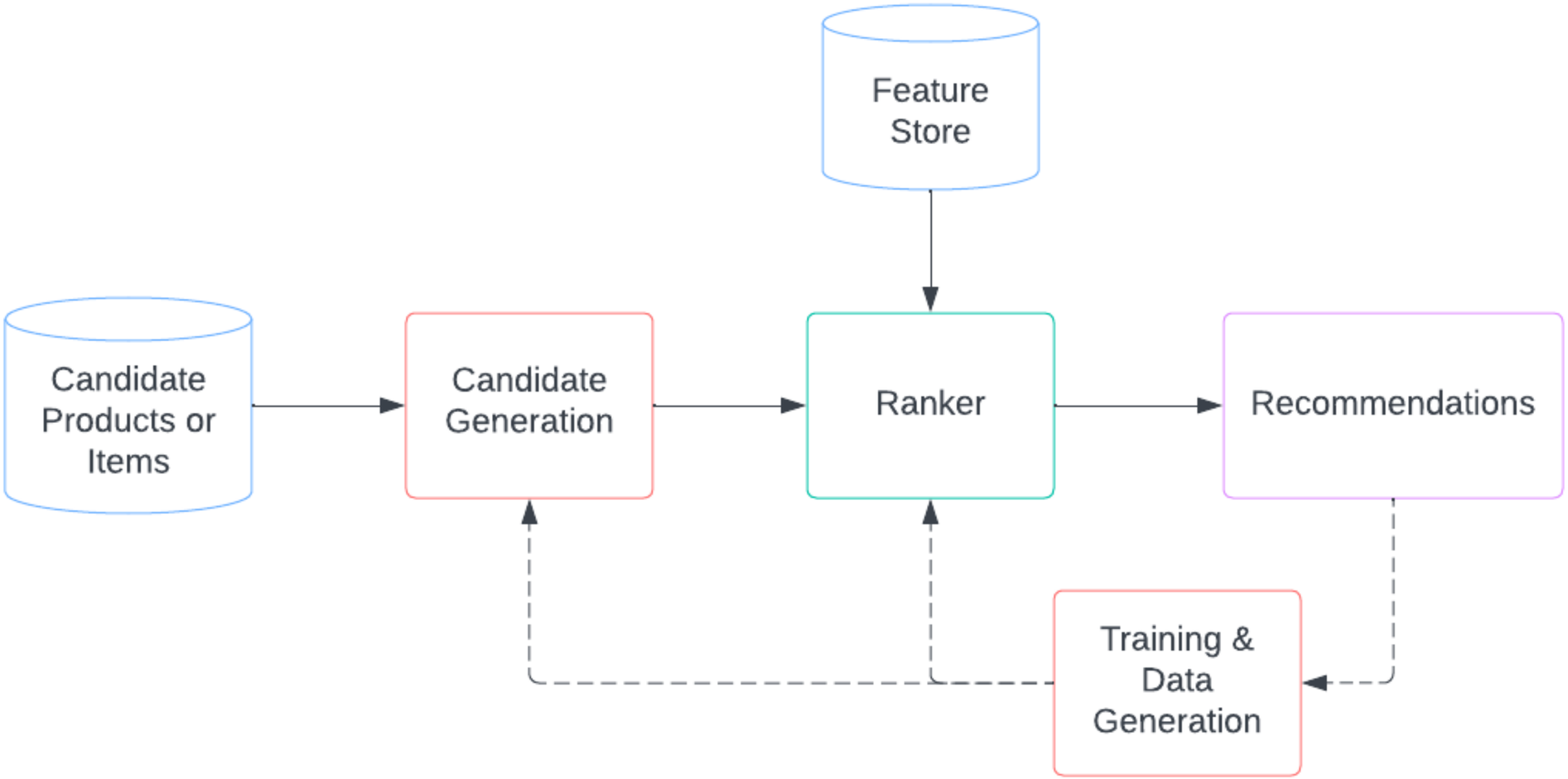
Objective: Distilling a large sample of options to a select relevant few

Photo by Ketut Subiyanto

RECOMMENDER SYSTEMS METRICS

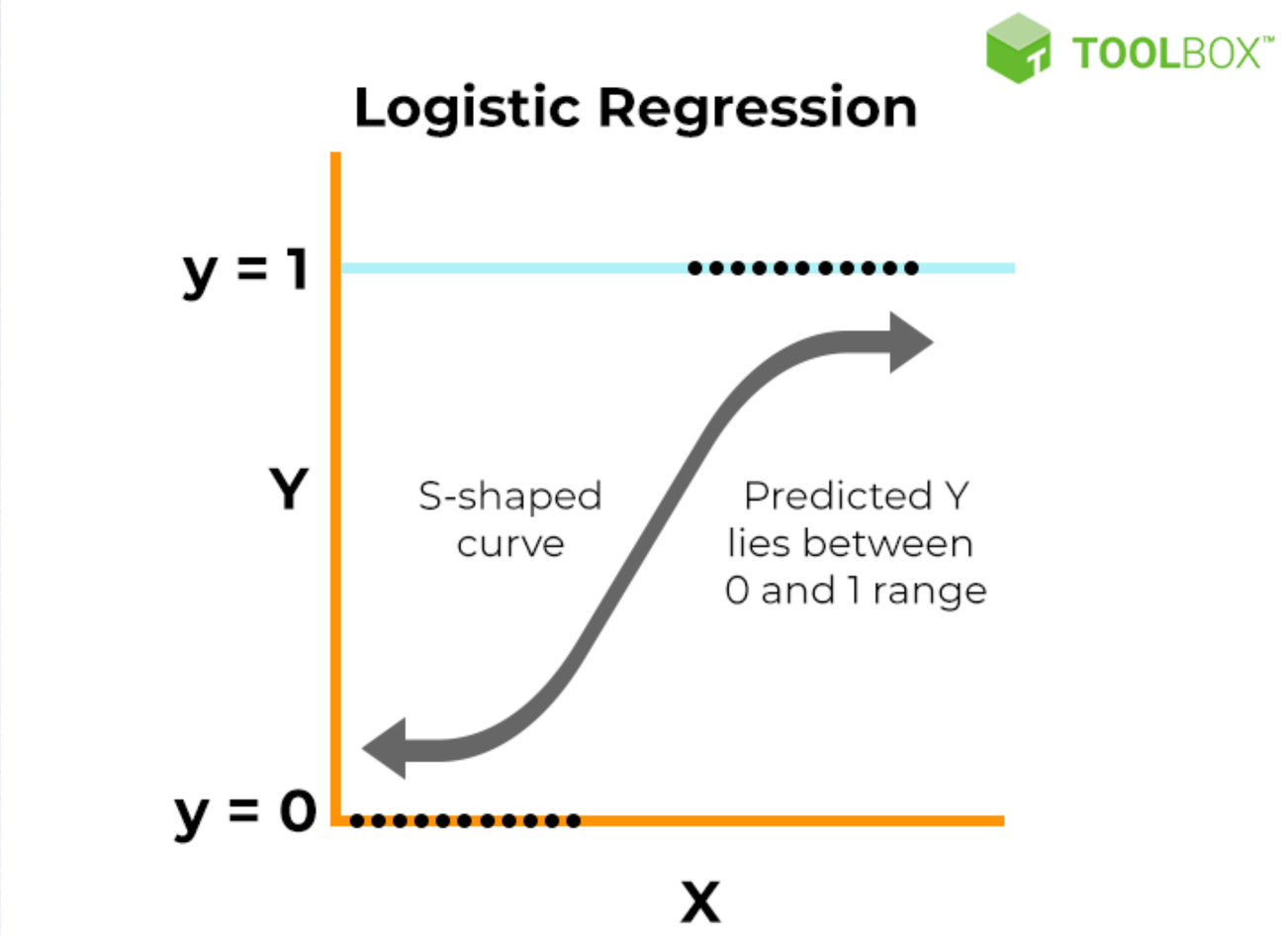
- Online Metrics: **Engagement Rate**, Session Engagement Time. **Purchases** or Add to Cart, Click/Detailed Page View
- Offline Metrics: **F-1 Score (precision, recall), AUC**, RMSE (optimizing for explicit feedback like ratings)

RECOMMENDER SYSTEMS



RANKER: LOGISTIC REGRESSION

- Predicted probability > .5 then its group membership is 1 (+ve) and vice versa
- Uses maximum likelihood to fit the line to the data (not least squares like linear)



Source: spiceworks.com

RANKER: LOGISTIC REGRESSION

- When data is limited
- Limited Training/Model Eval Capacity
- Require Initial Baseline
- Exploring explainability of feature importance
- Fast to train

FEED RANKING SYSTEMS

- Problem Statement: Show the Most Relevant List of Content to a User of [X]
- ML Model that predicts the likelihood a user clicks on a Post

FEED RANKING SYSTEMS

- **Objective**: Provide a ranked list of feed content
- Ex: Design a ML System to provide a list of relevant Reddit Posts

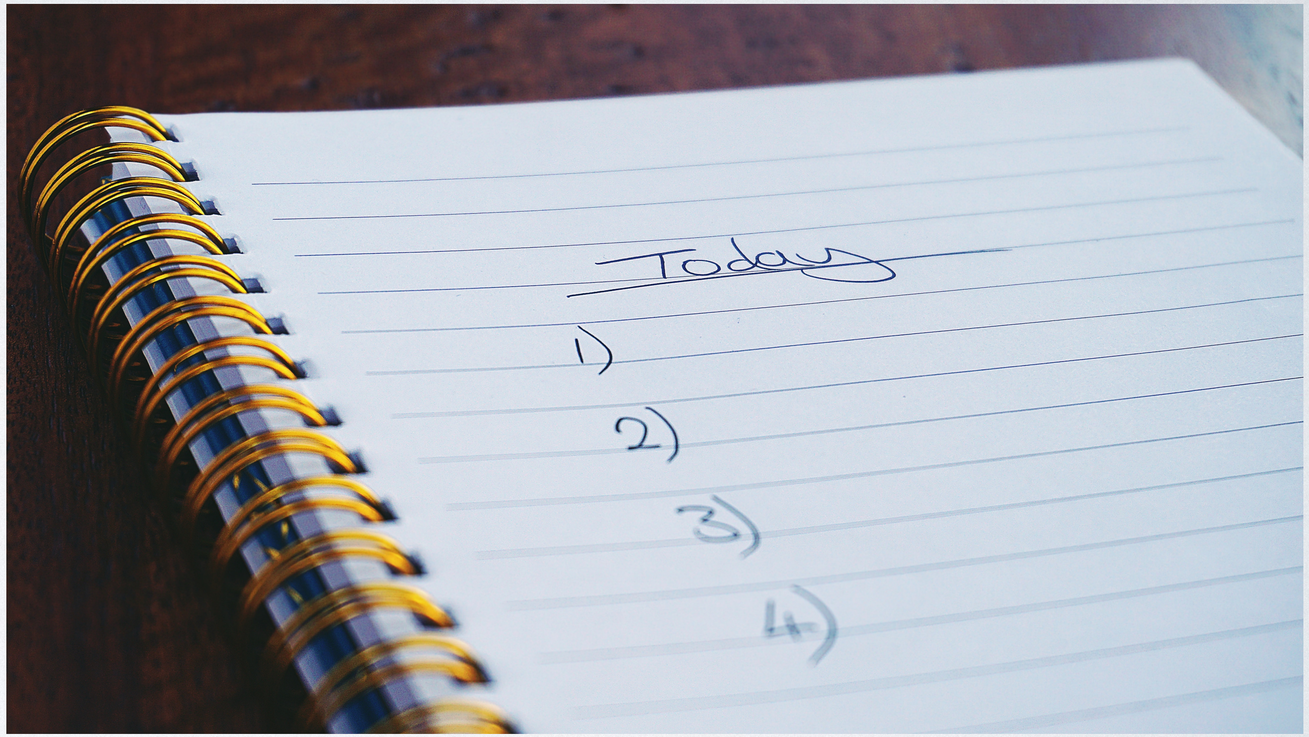


Photo by Suzy Hazelwood

SCALE & SCOPE

- 50MM Daily Active Users (user_id)
- Provide a list of 20 most relevantly ranked posts (post_id)
- 1Billion posts need to be presented daily = 20 * 50MM
- 1 Month => 30 Billion Posts (30 days)
- System needs to serve 12 K posts/sec (assuming 2.5MM secs per month)

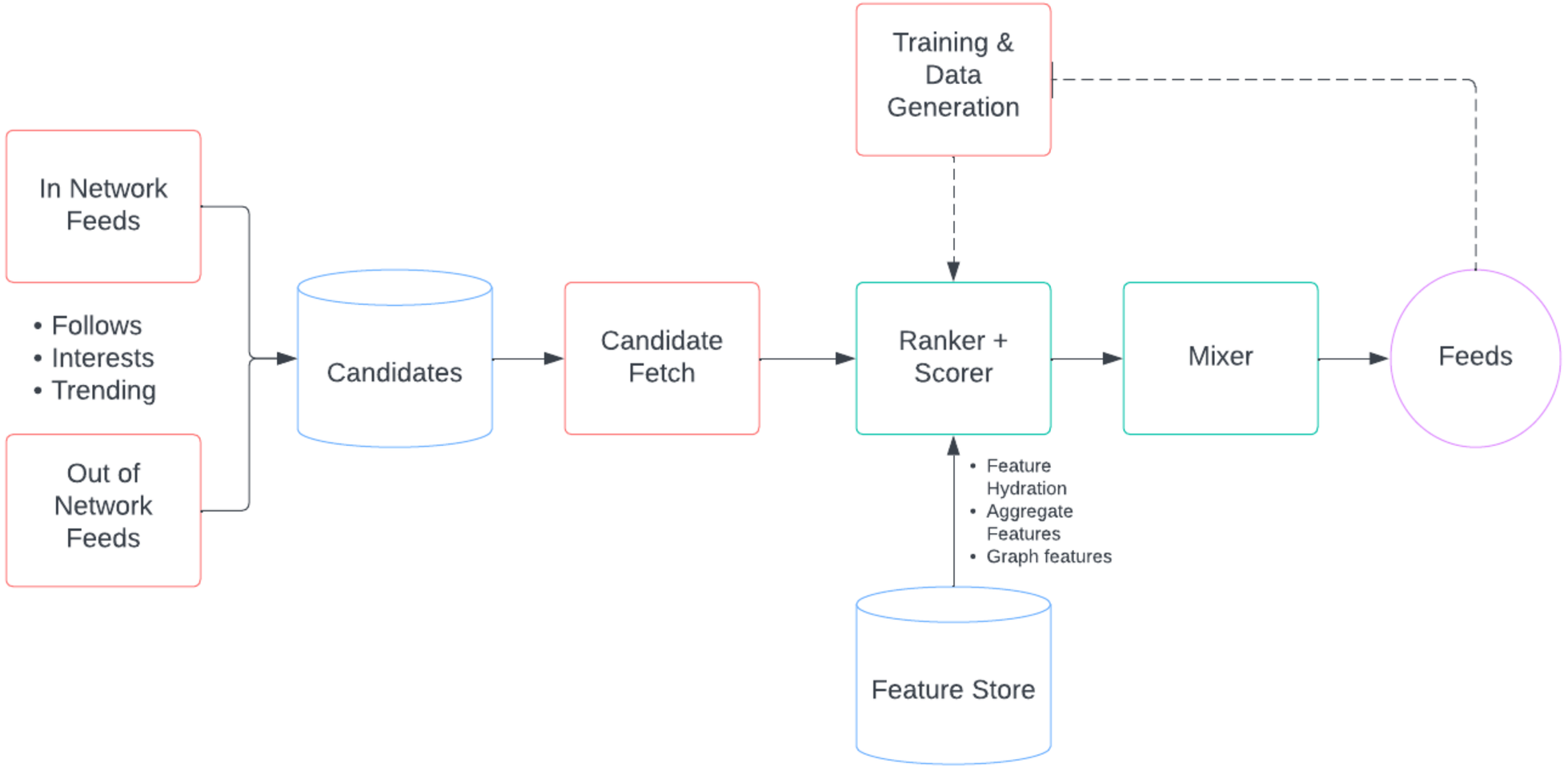
FEED RANKING SYSTEMS METRICS

- Online Metrics: Engagement Rate, Weighted Engagement
  - Positive User Actions: Time Spent, Like, Sharing, Commenting
  - Negative User Actions: Hiding, Blocking, Reporting
- Offline Metrics: AUC, F1 Score => (Precision/Recall)

FEED RANKING SYSTEMS

- Content Diversity
  - Media Type (Video, Text, Image)
  - Source (In Network & Out of Network)
- Repetition Penalty

FEED RANKING SYSTEMS



AD PREDICTION SYSTEMS

- Problem Statement:
  Display Relevant Ads to User of [X]

Photo by Jose Francisco Fernandez Saura

AD PREDICTION SYSTEMS

- **Objective**: Show a list of relevant Ads which maximize monetization
- Ex: Design a ML System to display 20 relevant Ads to a user while shopping online

Created using Dall-E

SCALE & SCOPE

- 50 MM Daily active users
- 1 Billion Ads displayed per day (Assume they see 20 Ads a day)
- 30 Billion Ads/Month
- 2.5 MM secs in a month
- 12k Ad Requests/Sec

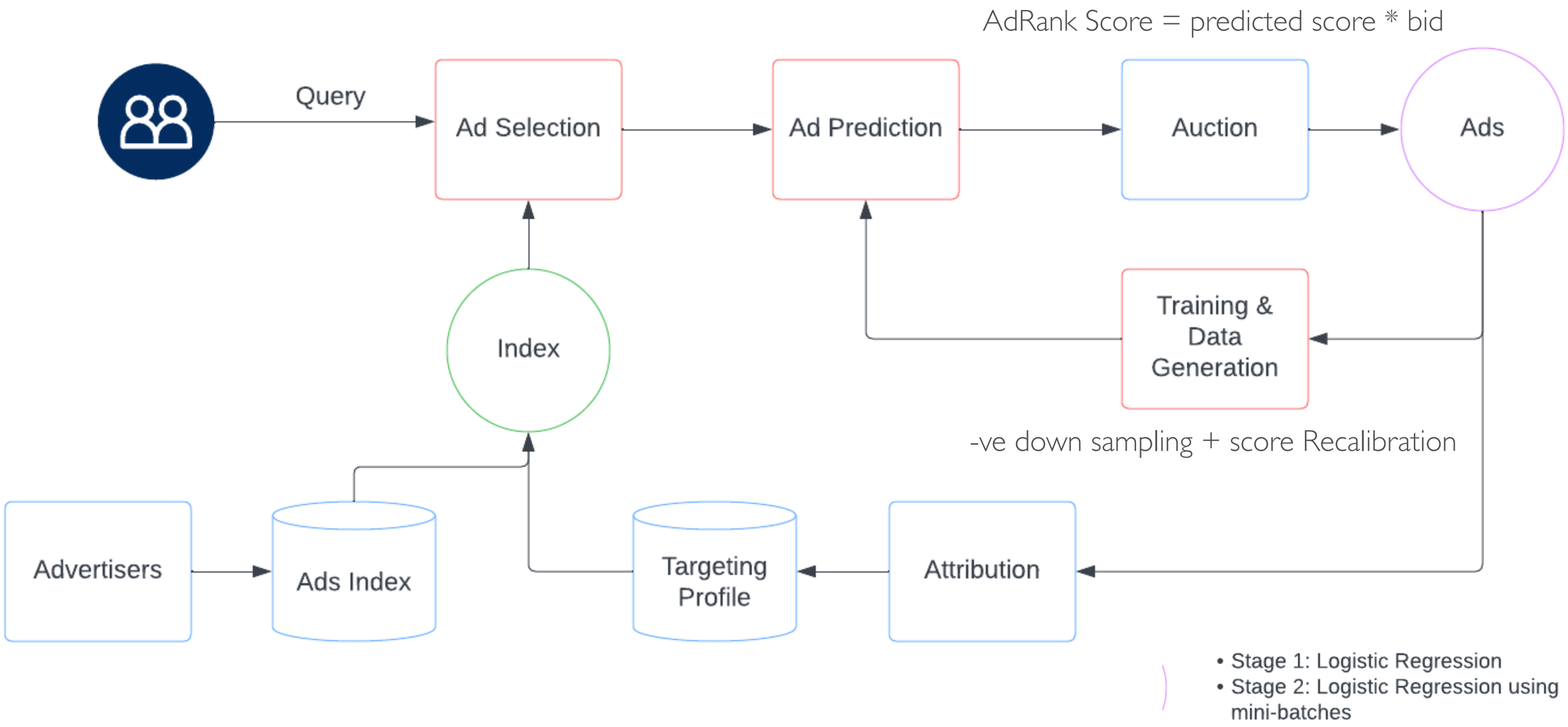
AD PREDICTION SYSTEMS METRICS

- Online Metrics:
  - Engagement Rate: 1) **CTR**, 2) Downstream Action Rate (Add to Cart, **Purchase** etc)
  - Revenue Increase
  - Counter Metrics: Ad Hides, Blocks, Reported
- Offline Metrics: **Cross Entropy Loss/Log Loss** (not AUC)

CROSS ENTROPY/LOG LOSS

- Used for classification with probability between 0 and 1
- Is key with Ads because how far off the predicted score is from ground truth impacts the Auction
- Log Loss better than AUC for Ad Prediction because AUC does not penalize a model for how far off the prediction is from the actual label
- In AUC a score of 0.51 and 0.8 (threshold 0.5) contributes equally even though one is closer to the threshold than the other

AD PREDICTION SYSTEMS



SEARCH SYSTEMS

- Problem Statement:
  Display Relevant Search Results to User of [X]

SEARCH SYSTEMS

- **Objective**: Find content that best match my search keywords or phrase

Created using Dall-E

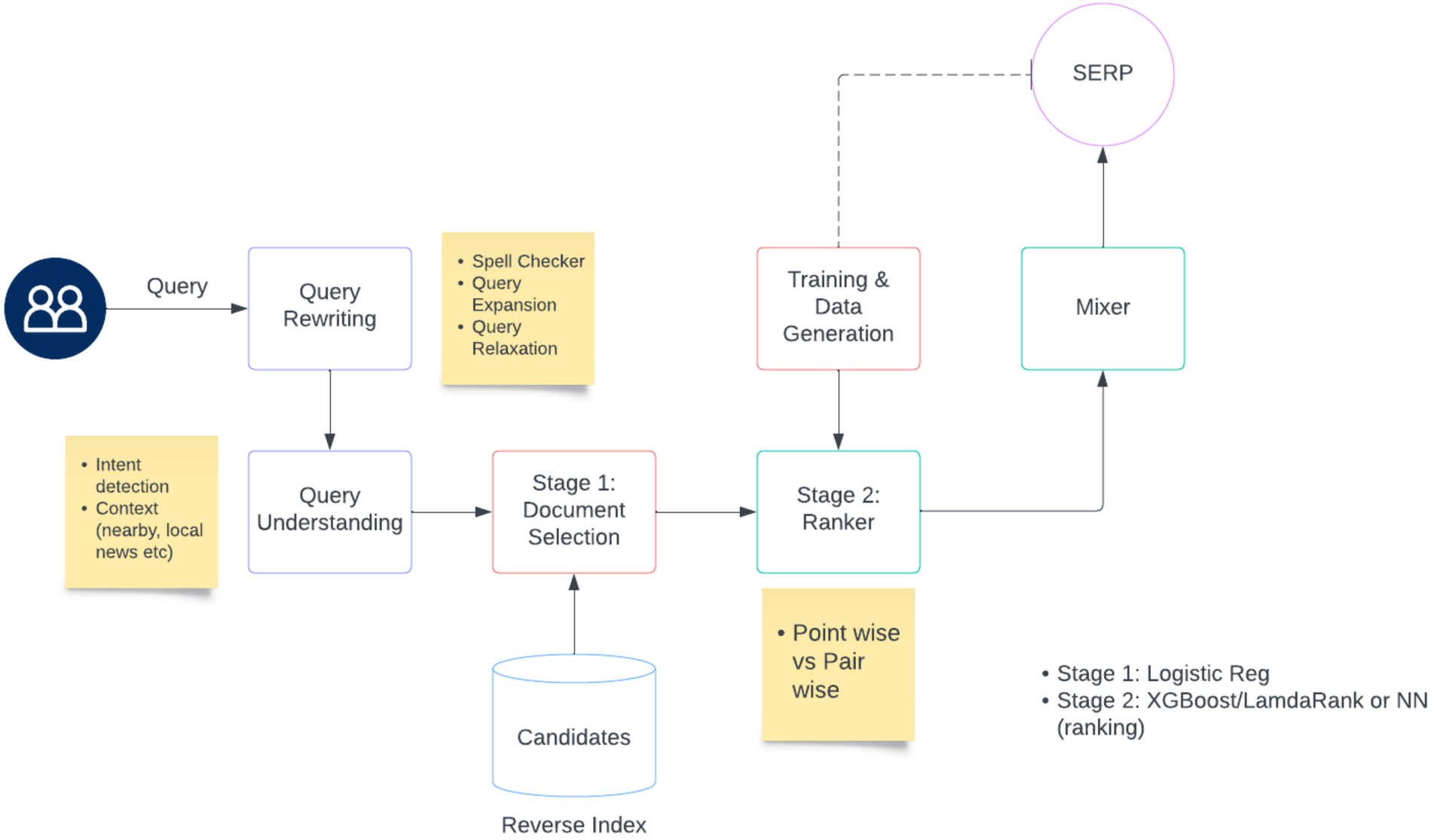
SEARCH SYSTEMS: SCALE & SCOPE

- 100 MM active users
- 100 searches/month
- 10 Billion searches/month
- 2.5 MM secs in month
- 4000 search req/sec

SEARCH SYSTEMS METRICS

- Online Metrics:
  - **CTR** (Click Through Rate)
  - Successful Session Rate (Dwell Time)
  - **Zero-Click Searches**
- Offline Metrics: Normalized Discounted Cumulative Gain (**NDCG**)

SEARCH SYSTEMS DESIGN



NDCG

- Normalized Discounted Cumulative Gain
- Accounts for the overall quality of search results
- Sum up total relevance of each item in result set, then position of each item is discounted
- Penalize lower ranked more relevant item

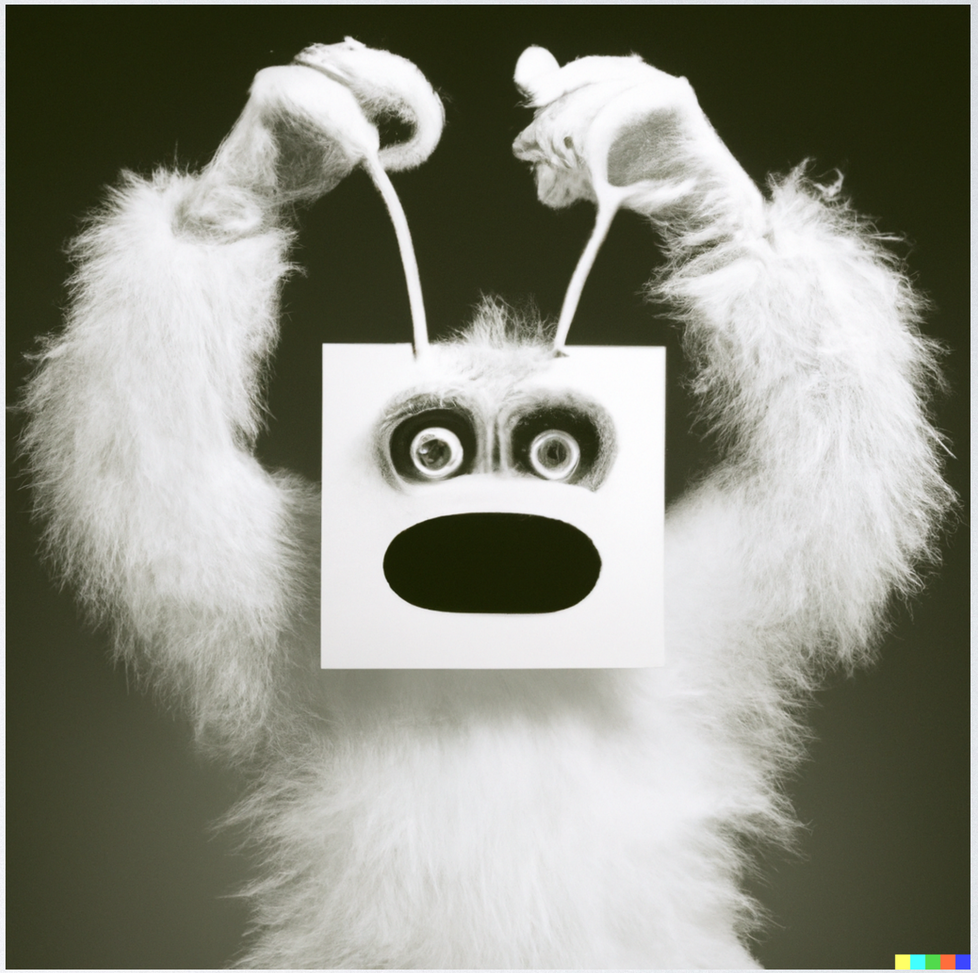
ENTITY LINKING SYSTEM

- Problem Statement:
  Detect and Link Entity in Text to Entities in a Target Knowledge Base [X]

ENTITY LINKING SYSTEM

- **Objective**: Make the best association to an entity given a body of text and link to mode details

- Ex: Determines if which "Michael Jordan" was in the movie "Creed"

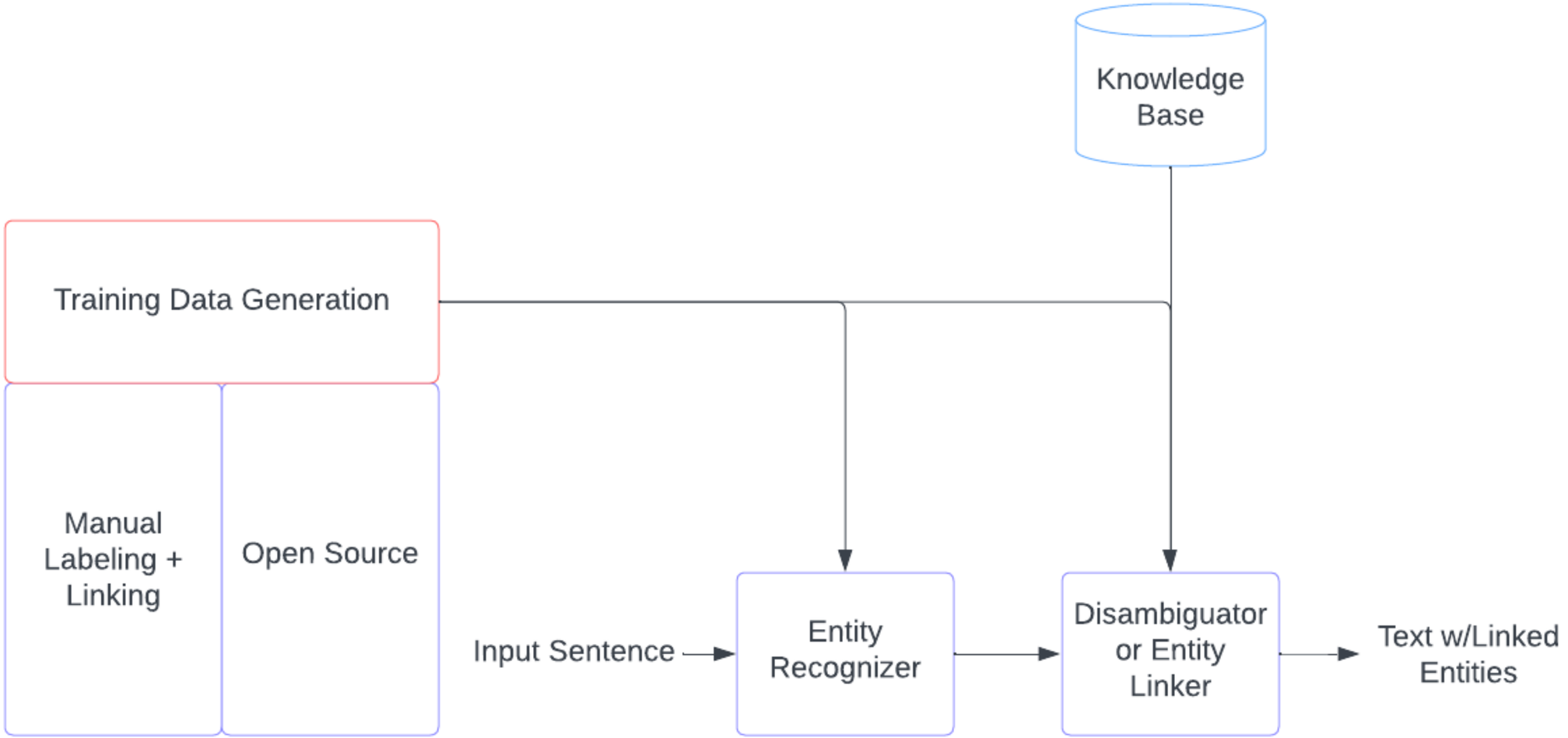


Created using Dall-E

ENTITY LINKING SYSTEM METRICS

- Online Metrics:
  - Successful Session Rate (Search Engine)
  - % of Questions Answered Correctly (Virtual Assistant)
- Offline Metrics: Recall, Precision, F1-Score

ENTITY LINKING SYSTEM



MENTAL MODELS FOR YOUR DESIGN

- Be clear on the problem statement
- Understand and articulate your required scale and scope
- Define and select your offline and online metrics

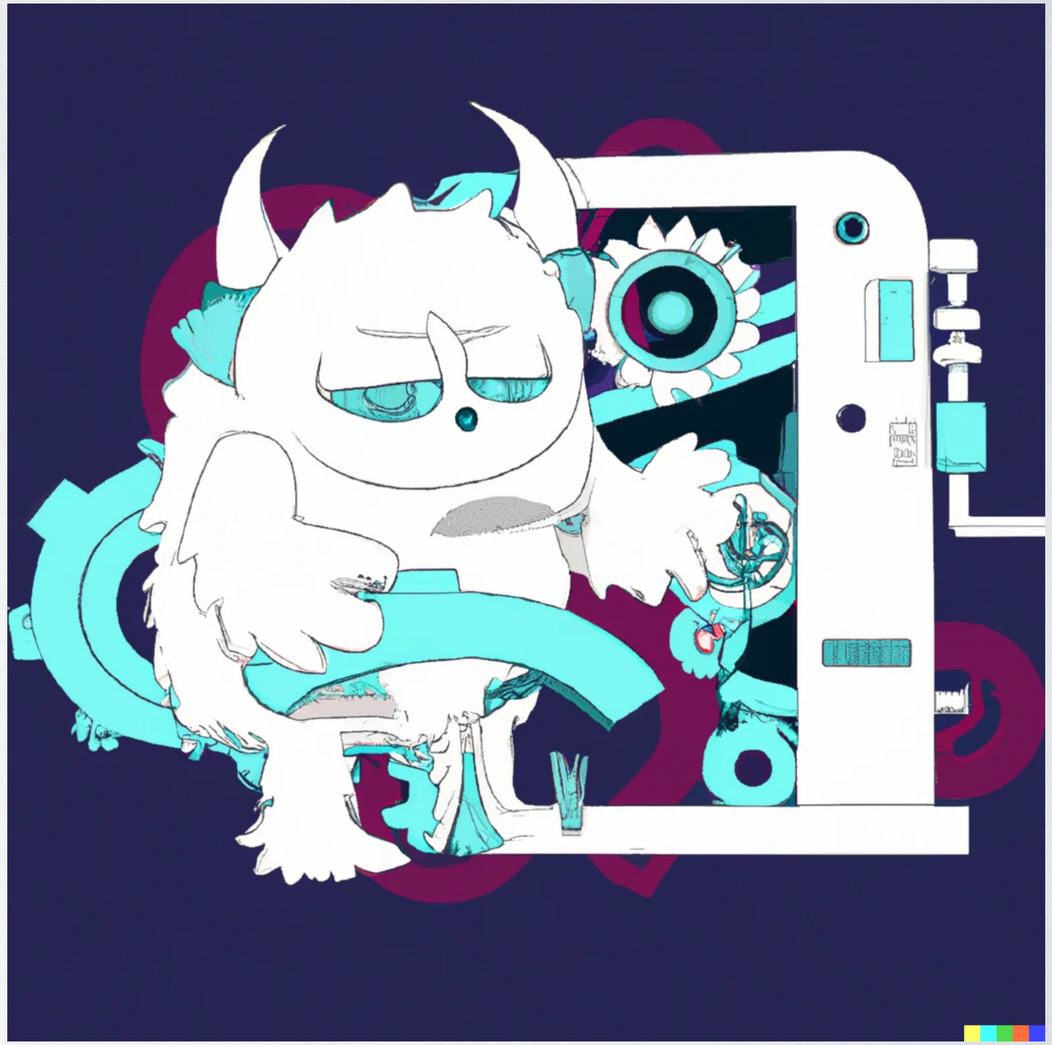
ML OPS FOR BUSINESS

"In theory, there is no difference between theory and practice. In practice, there is"

– Yogi Berra

ML OPS: WHY AND WHAT

- Manage, Deploy and Monitor ML Systems
- Tooling and Best Practices



Created using Dall-E

ESSENTIALS

- Data and Model versioning
- Continuous Integration and Deployment + Automated testing
- **Monitoring and Logging**
- **Scalable Infrastructure and Reproducibility**

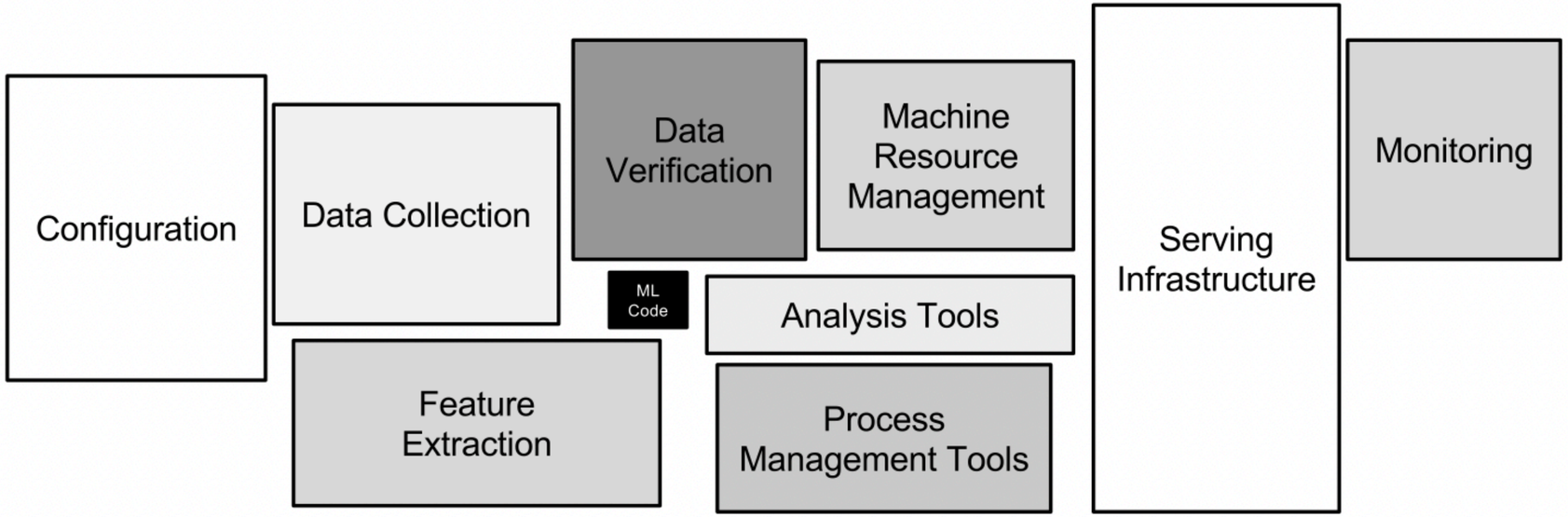
ESSENTIALS

- **Experiment Tracking and Management**
- **Feedback Loops**
- Data Management and DataOps
- Security and Model Bias Checks

DEALING WITH PRODUCTION

- Rolling releases
- Experiment
  - 1% of live traffic
- Failure : anomaly detection, feature distributions, data loss
- Rollbacks

STILL BEYOND MODELING AND UX



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

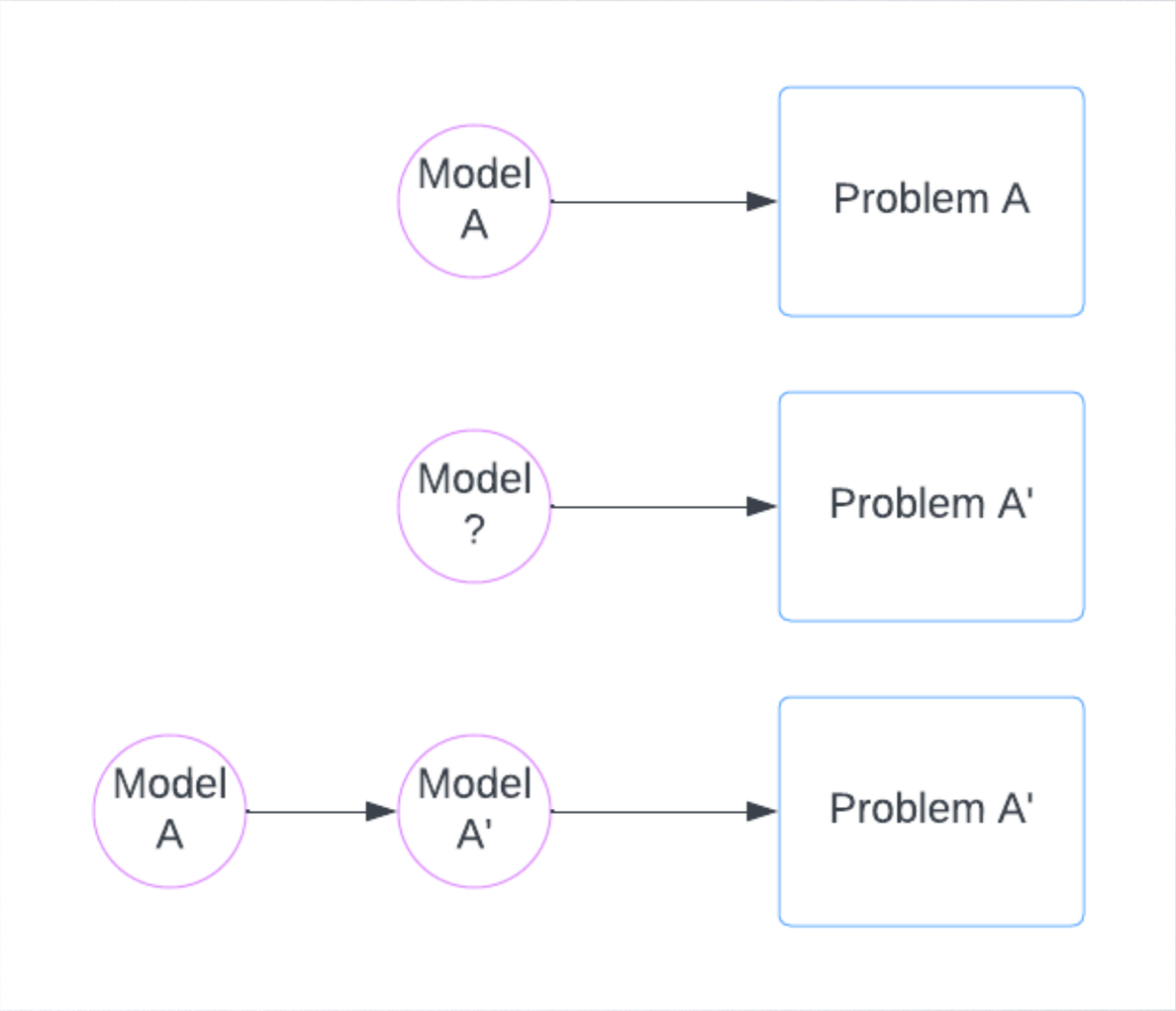
Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

ENTANGLEMENT

- Messy interfaces in ML vs Software -> Isolation of Improvements
- CACE - Change Anything, Change Everything
  - Signals, Hyper Parameters, learning settings etc
- Managing Entanglement
  - Isolating Models + Serving Ensembles
  - Detecting changes in prediction behavior early

CORRECTION CASCADES

- Improving accuracy of individual component could lead to system level detriments
- Manage by making Model A aware of the unique use cases



ANTI-PATTERNS

- Glue Code
  - Generic Packages (Buyer beware)
  - Managing with Black Box Wrappers (Consistent Interfaces)
- Pipeline Jungle
  - Ever growing new signals and data sources
  - Managing with design reset

MORE ML TECH DEBT

- Dead experimental code -> Like dead feature flags
- Data dependencies -> Like code dependencies
- Hidden feedback loops

CONNECTING THE DOTS

- Understand Problem & get clarity of scope (use case)
- Outline system requirement and scale (features to target and NFRs)
- Define metrics for performance (Online and Offline)
- Work out the component Architecture
- Figure out training data generation

CONNECTING THE DOTS

- Feature Engineering
- Modeling Technique
- Experimentation Plan
- Address failure scenarios and plan for ML Observability and Ops

CONNECTING THE DOTS

Ask : How precisely can the impact of a new change to the system be measured?

THANK YOU

QUESTIONS?