RECOMMENDATION ENGINES

MASTER IN BUSINESS ANALYTICS AND BIG DATA

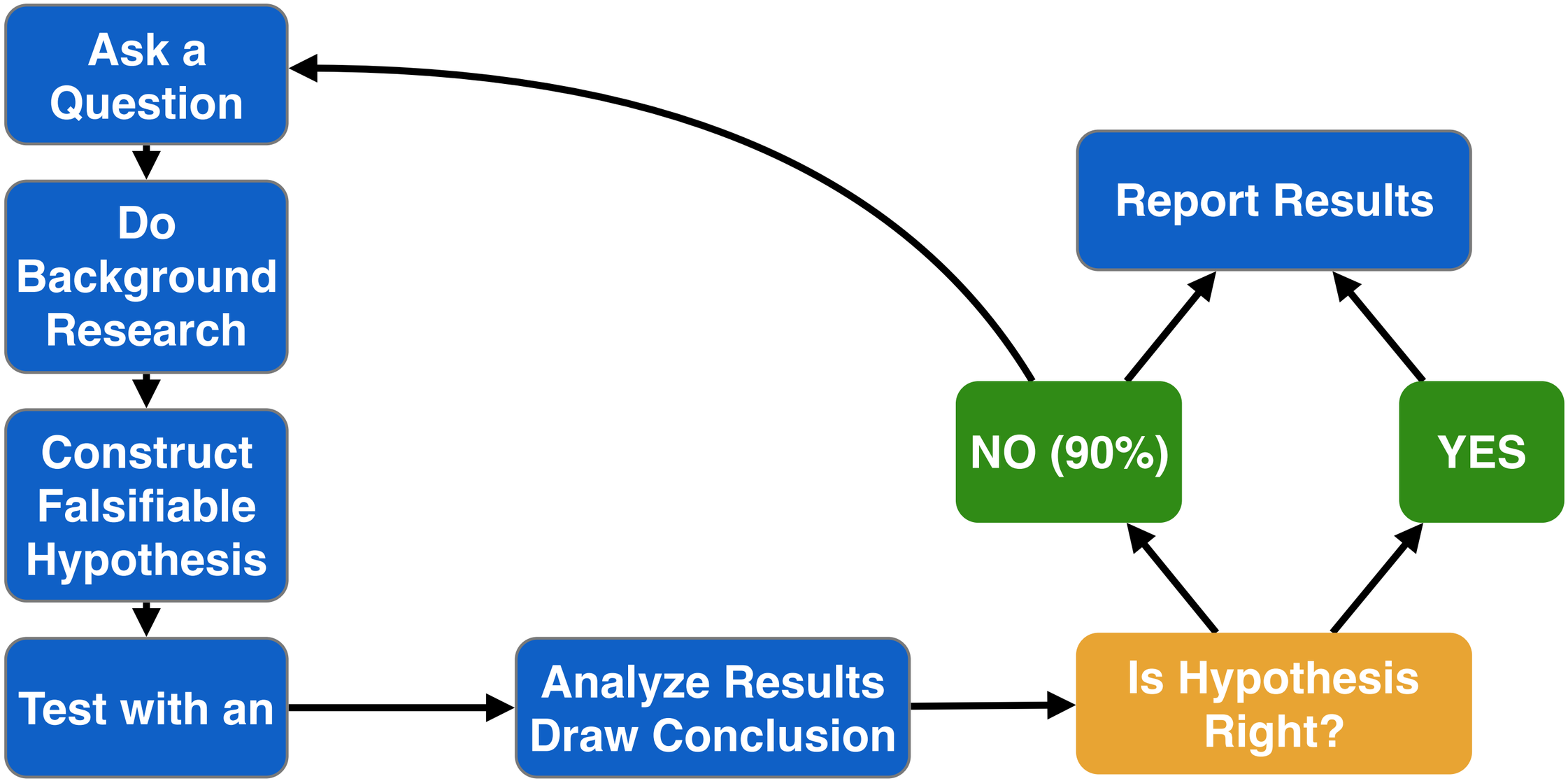
SESSION 7-8, **BUILDING A RECOMMENDATION ENGINE**

- The Scientific Process for Business
- Evaluation Methods
- In-Class Exercises
- Real World Cases
- Top 10 Lessons Learned

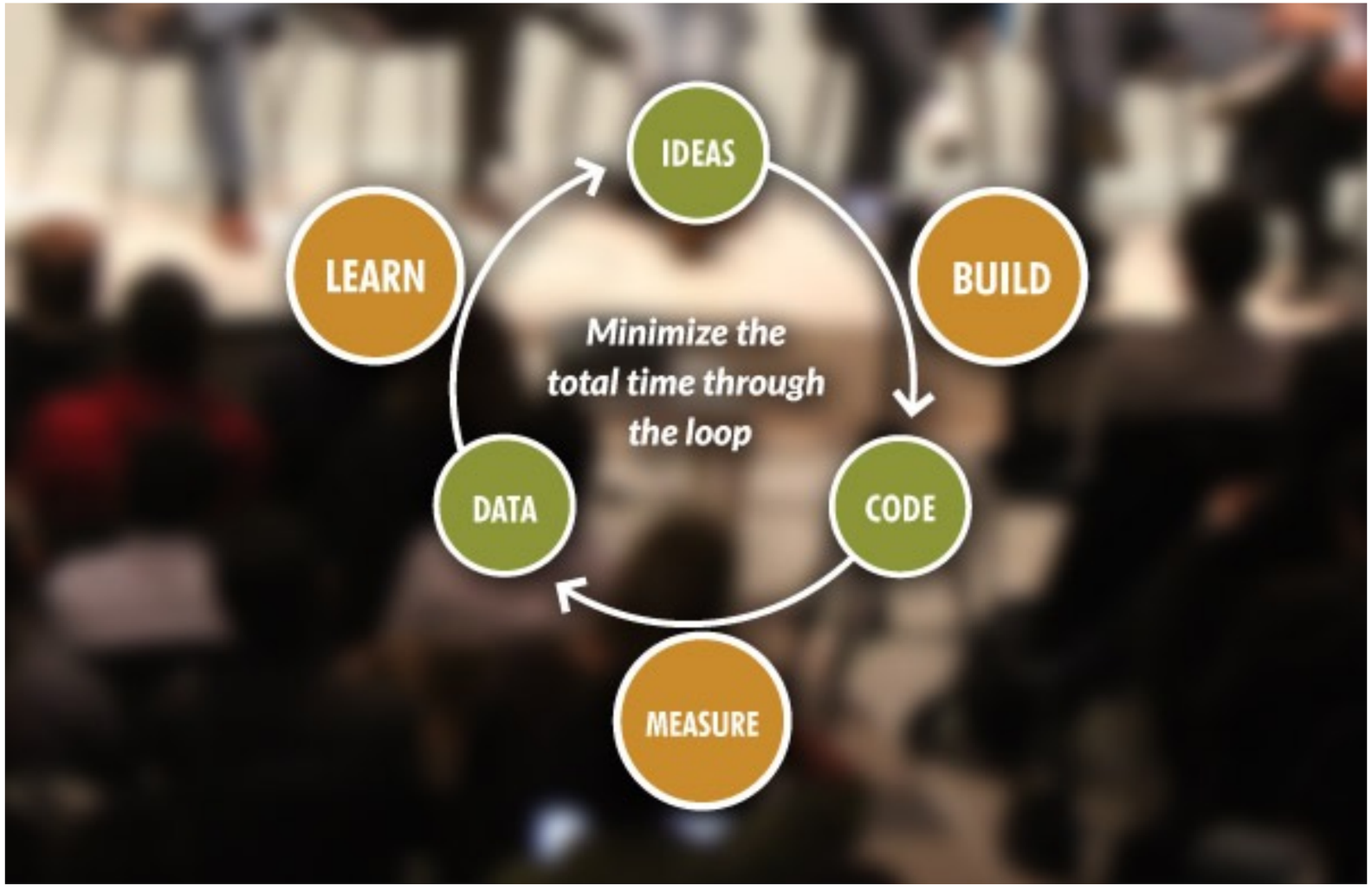
CONSUMER (DATA) SCIENCE

1. Start with an hypothesis
   - Solution **X** solves problem **Y**
   - E.g. : We have a low customer retention and algorithm/feature/UX/design X will increase the customer retention.
2. Validate the hypothesis
   - Two variables X,Y! First validate problem, solution afterwards
   - Define the experiment
   - Define success metric (i.e. problem: customer retention is 10% lower than competitors, i.e. solution: customer retention increases more than 10%)
3. Execute the Test
4. Let **Data** speak for itself!

THE SCIENTIFIC PROCESS

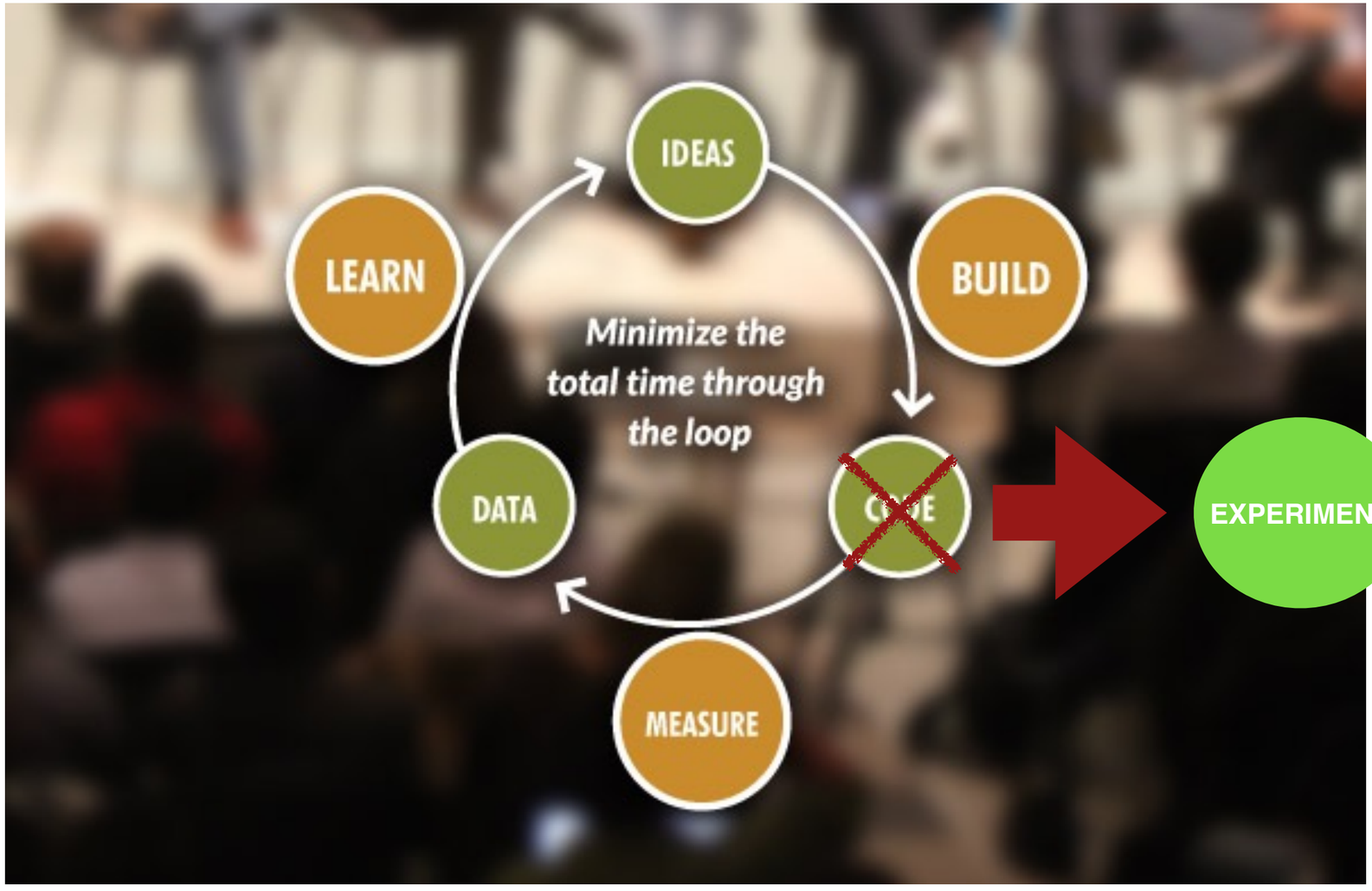


IN BUSINESS : THE LEAN STARTUP

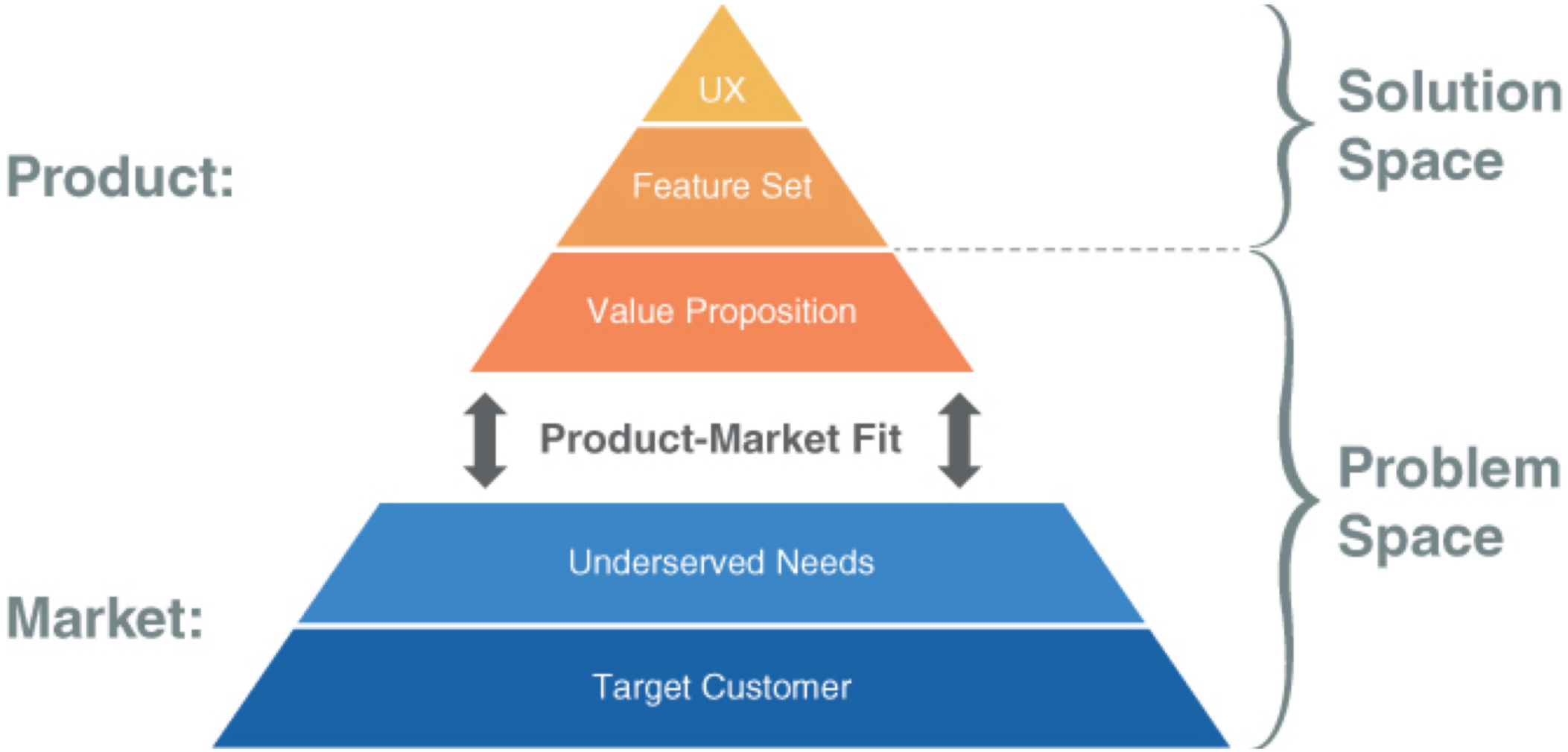


*Lean Startup, Eric Ries - 2011

THE FEEDBACK LOOP



PROBLEM VS SOLUTION SPACE



*The Lean Product Playbook, Dan Olsen

PROBLEM/SOLUTION FIT

VALIDATING THE PROBLEM

Pay Ads (Copy validation, A/B Testing)*

Landingpages (A/B Testing)*

CUSTOMER INTERVIEWS

Online chats (Leads for interviews)

By Phone

Micro-surveys

Cold Calling

VALIDATING THE SOLUTION

MINIMU VIABLE PRODUCT (Building the engine)

Sketches

Wireframes

Mockups (Design)

Prototypes (Online, close to real, may use data)

Fake doors (incomplete new features + A/B Testing)

Testing other people's solution (discovering key features)

Micro-surveys

Online chats

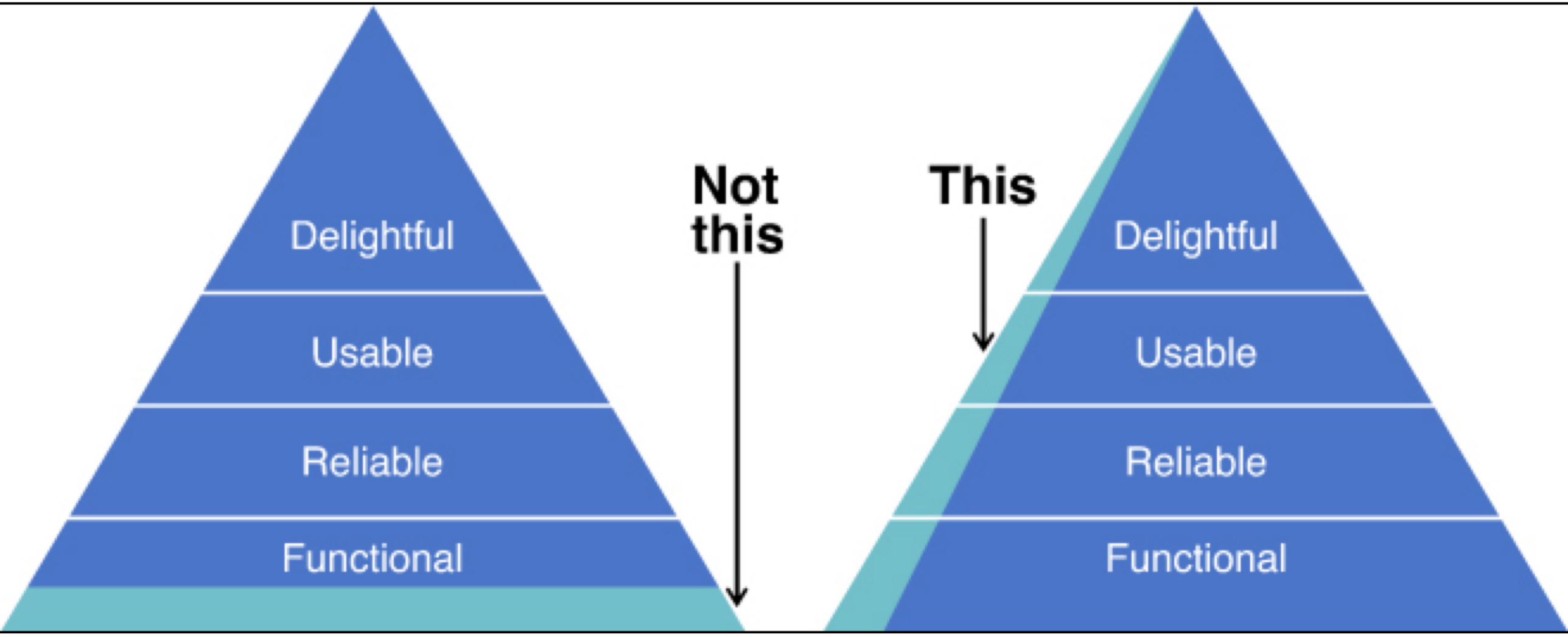
Customer Interviews

Do it yourself usability testing

Site visits (heatmap, session video)

*Be careful, you are also testing Marketing / Copy here

THE MVP



*The Lean Product Playbook, Dan Olsen

CONSUMER (DATA) SCIENCE

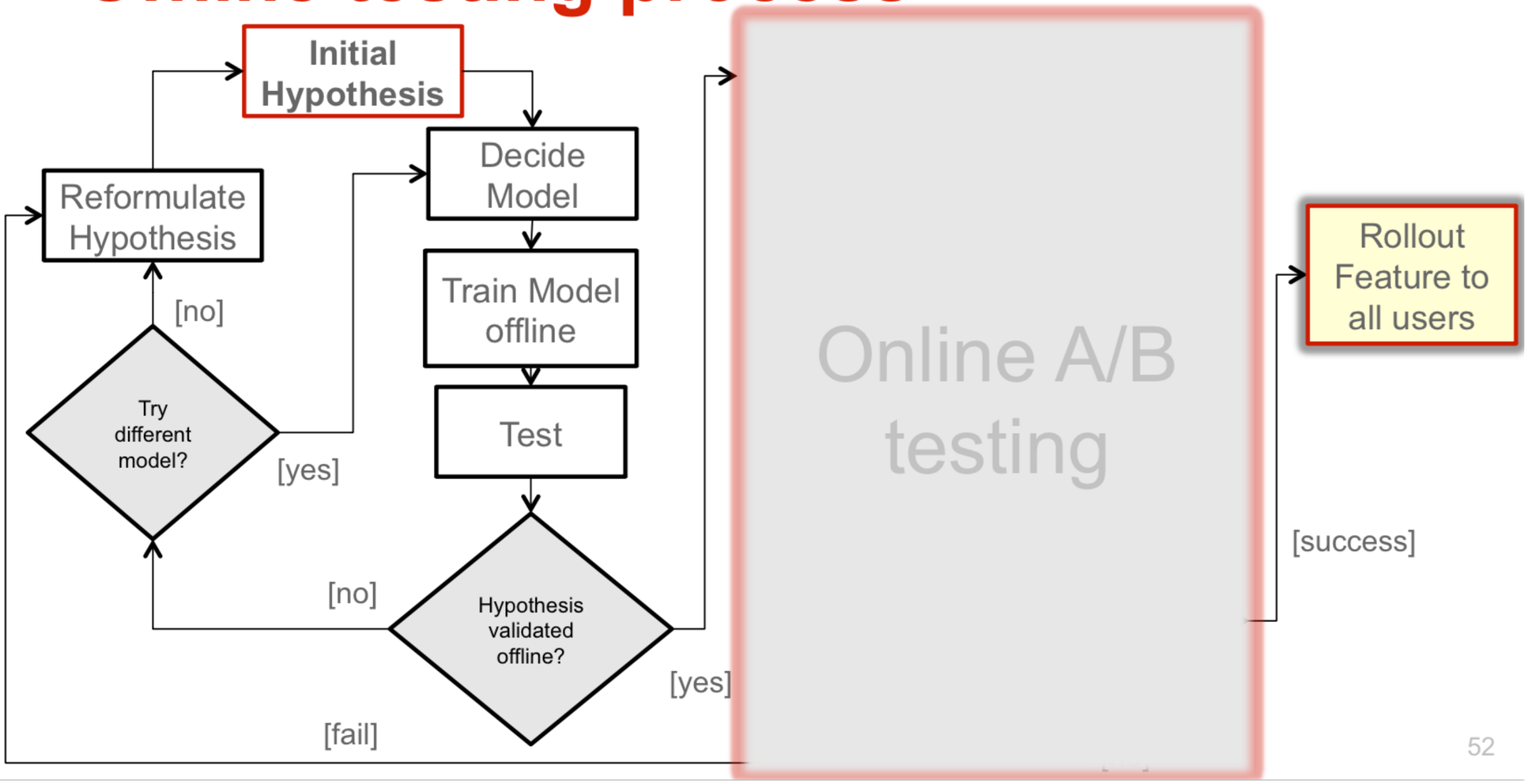
- Offline Testing
- Online Testing

*Xavier Amatriain- 2011

OFFLINE TESTING

(DAYS)

Offline testing process



52

*Xavier Amatriain- 2011

OFFLINE TESTING

thinking…

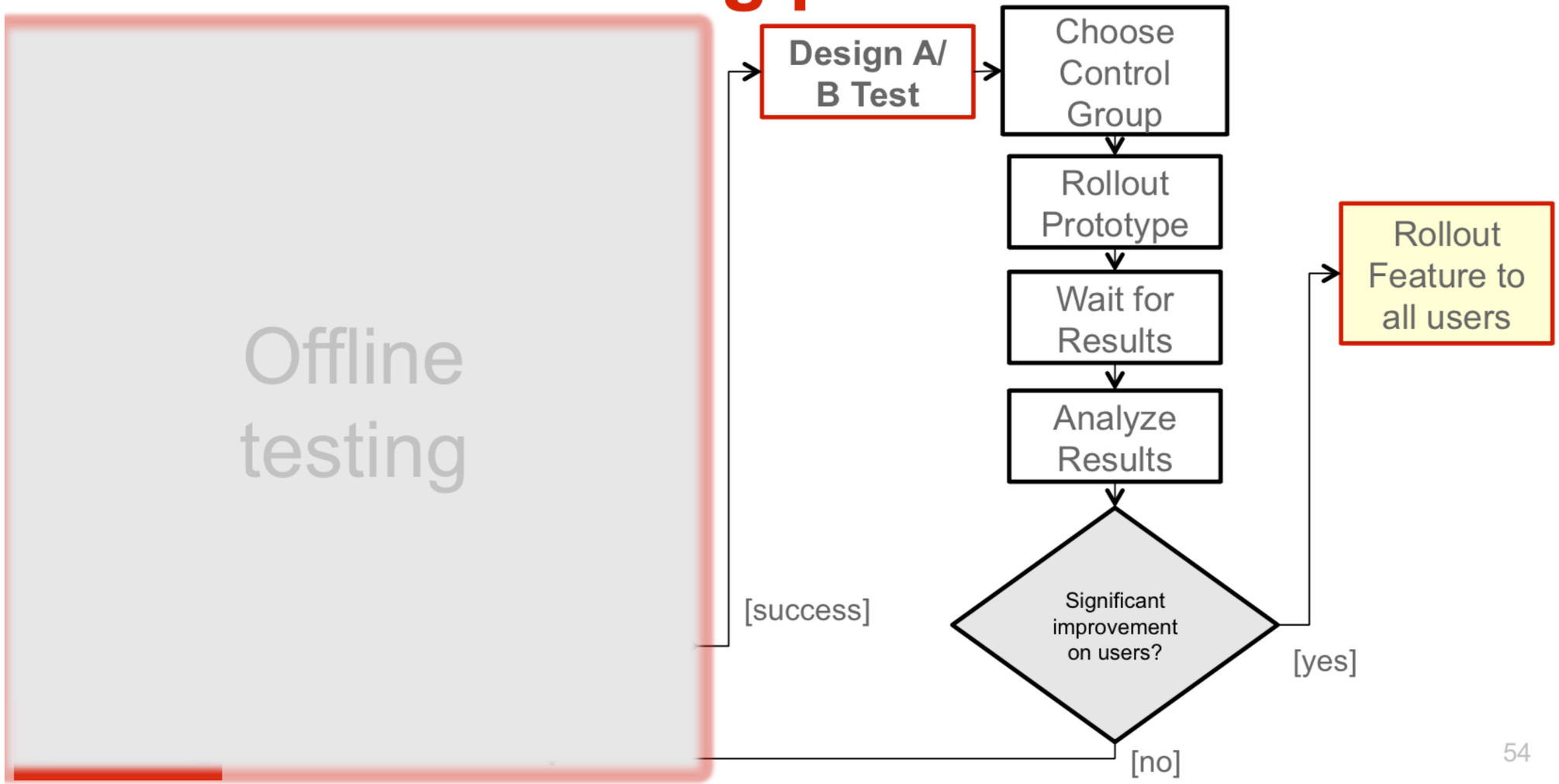
- The likelihood of a customer to buy shoes in the next month.
- What is your intuition?

*Xavier Amatriain- 2011

ONLINE TESTING

(WEEKS TO MONTHS)

Online A/B testing process



*Xavier Amatriain- 2011

VALIDATING THE SOLUTION

DATA AVAILABLE?

FEW

QUALITATIVE

ANALYSE MOTIVATORS (WHYS)

MANY

QUANTITATIVE

ANALYSE ACTIONS (WHATS)

- **RMSE**
- **PRECISSION / RECALL**

EVALUATION TYPES

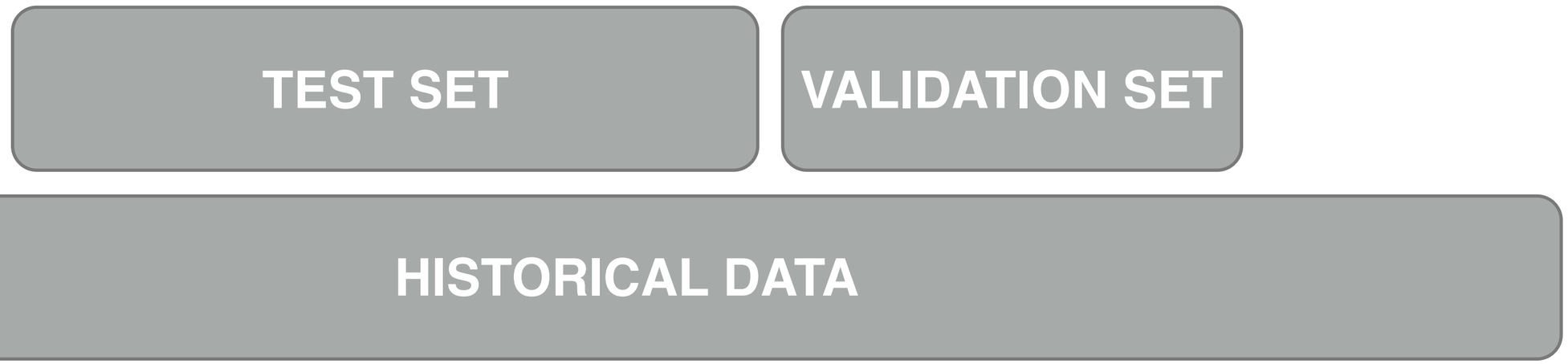
1. PREDICTION
2. RECOMMENDATION ENGINE
3. OVERALL PRODUCT
4. BUSINESS

DATA SETS FOR PREDICTION EVALUATION

Use a subset of historical data
to test the algorithm predictions:
i.e. known movie ratings, past purchases …

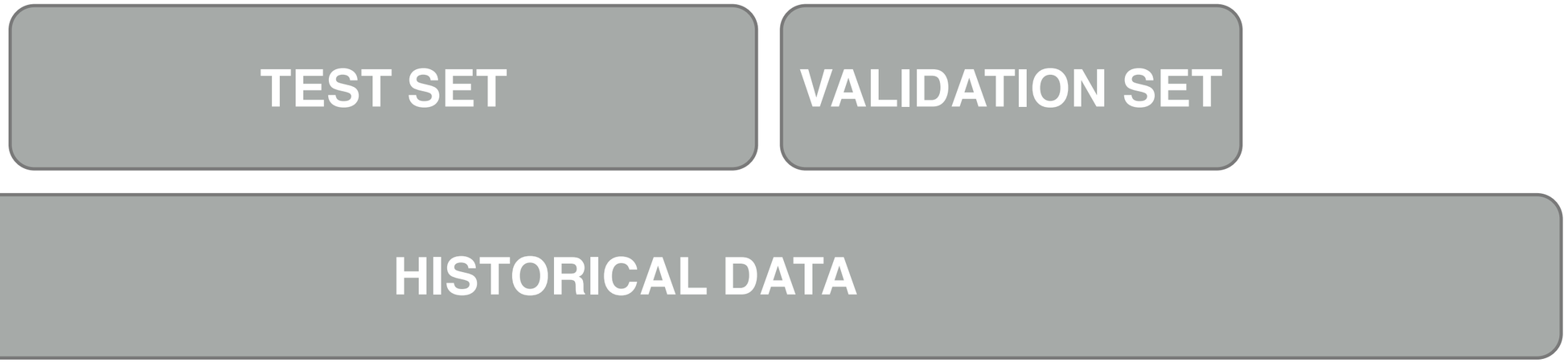
DATA SETS FOR PREDICTION EVALUATION

Use a validation set to assure that the algorithm evaluation is not dependent only of the test set



DATA SETS FOR PREDICTION EVALUATION

Select randomly the data to test and validate.
It has to be representative not a subgroup.



EVALUATING THE PREDICTIONS

how accurate the predictions are?
(error ratings average vs users ratings average)

(Mean Absolute Error) → $MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|$ → *Removes direction*

(Mean Squared Error) → $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ → *+ Penalizes large errors*

(Root Mean Squared Error) → $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$ → *Harder to misinterpret*

EVALUATING THE RECOMMENDATIONS

how good the recommendations to take good decisions?
(not considering only numerical accuracy as before, top-n ranking)

- **ROC AUC**
- **BREESE SCORE**
- **PRECISSION / RECALL**

how wrong the recommendations are when are wrong?
(evaluates large mistakes that lead to loss of confidence in the system)

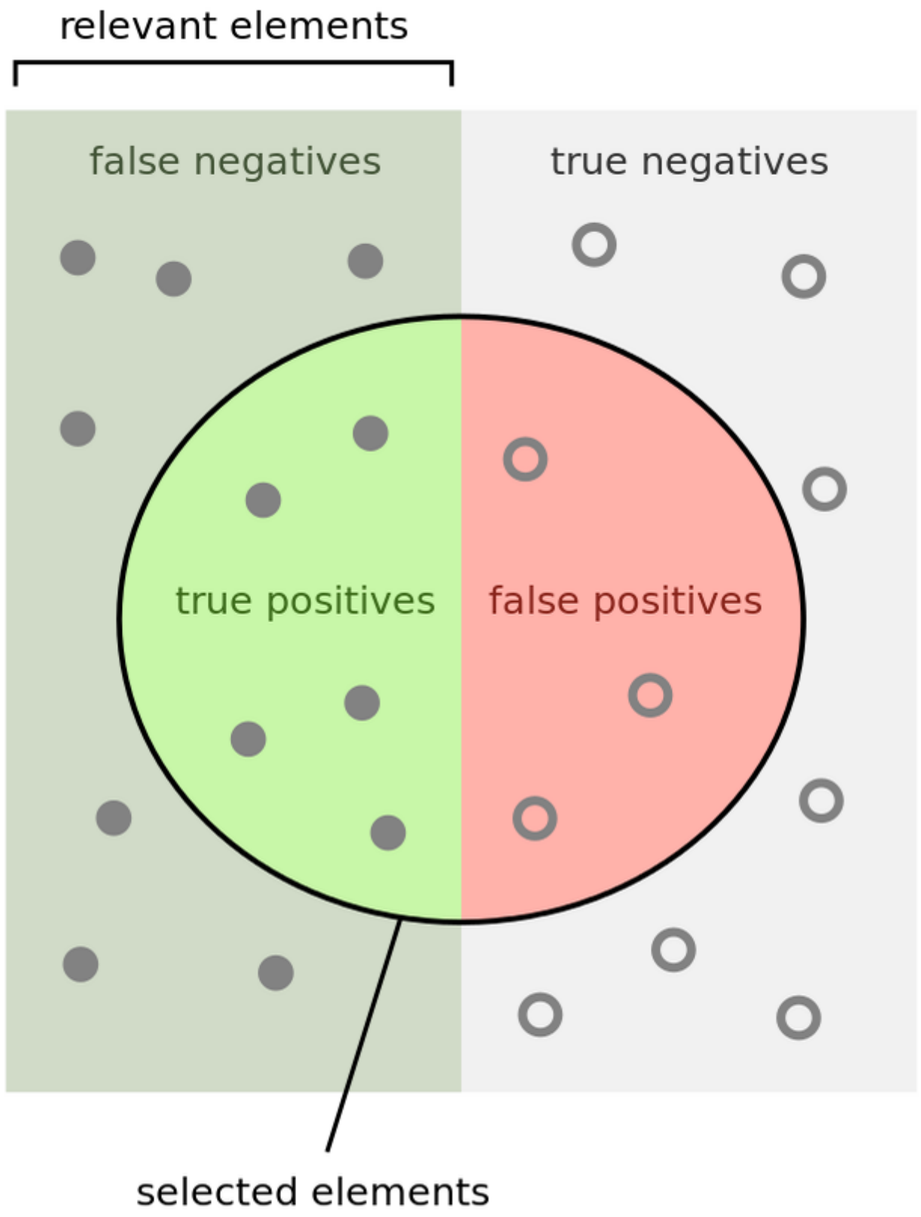
- **REVERSALS**

how good the recommendations help to navigate throughout items?

- **SERENDIPITY**
- **DIVERSITY**

EVALUATING THE PREDICTIONS

PRECISION / RECALL



How many selected items are relevant?

Precision = 

$$P = \frac{TP}{TP + FP}$$

How many relevant items are selected?

Recall = 

$$R = \frac{TP}{TP + FN}$$

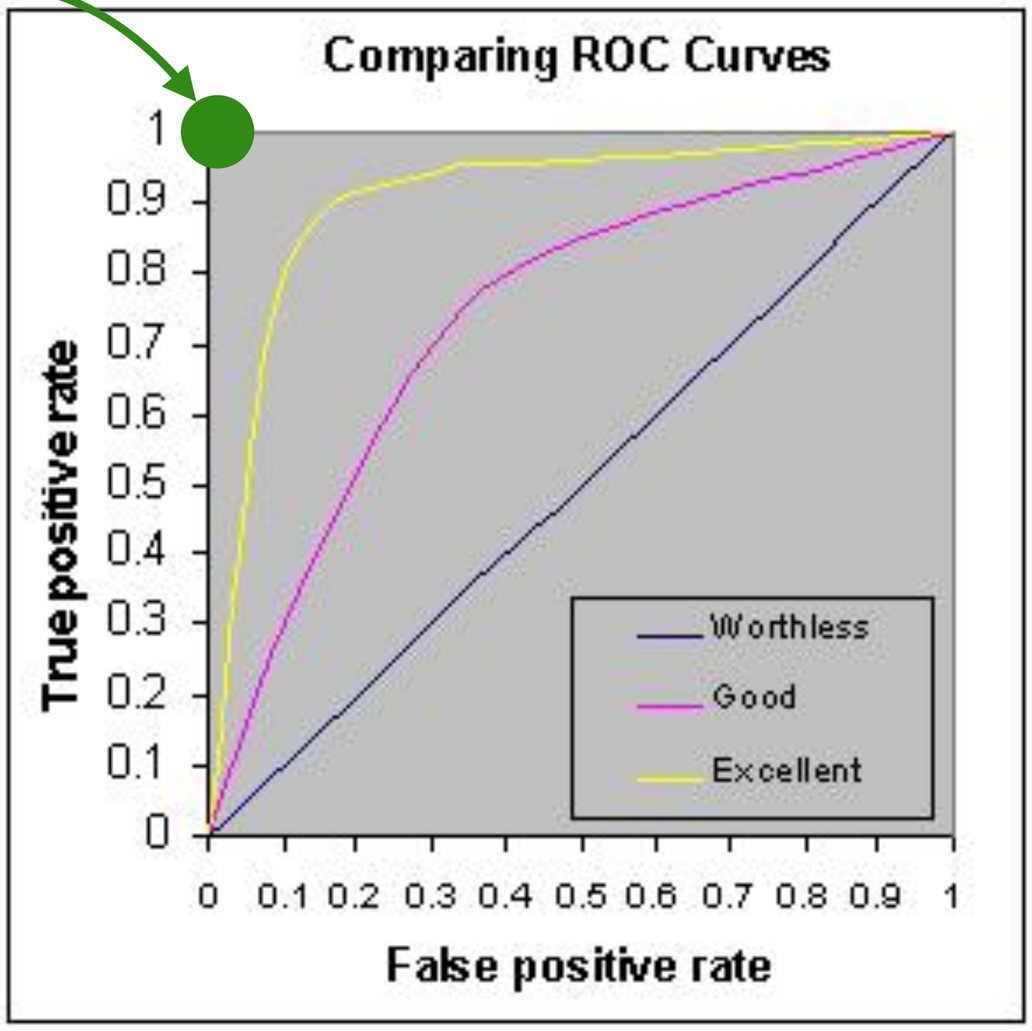
RELEVANT: good recommendations
SELECTED: given recommendations

i.e. 100% of recommendations correct (P=1), but only 10% of the available (R=0.1). Case: Top-1 recommendation was displayed, when 10 correct recommendations were available.

EVALUATING THE PREDICTIONS

ROC AUC

Perfect recommendation



RECALL ->

(Selected relevant recommendations)

<-FALL-OUT
(Selected irrelevant recommendations)

EVALUATING THE RECOMMENDATIONS

*How good the recommendations **rank**?*
(How good a recommender is ordering staff)

*How good is puting **first good stuff**?*

- **MEAN RECIPROCAL RANK (MRR)**

*How good **all rank order** is?*

- **SPEARMAN RANK CORRELATION**

*How good **initial rank order** is?*

- **DISCOUNTED CUMULATIVE GAIN (DCG)**

*How good pairwise **order accuracy** is?*

- **FRACTION OF CONCORDANT PAIRS (FCP)**

EVALUATING THE RECOMMENDATIONS

MEAN RECIPROCAL RANK (MRR)

1/i formula, where i is the first "correct" item (i.e. liked by the user)

	Recommender u_j	
Position i	***Top-5 List***	***Feedback***
Position 1	*itemA*	✗
Position 2	*itemB*	👍
Position 3	*itemC*	👍
Position 4	*itemD*	✗
Position 5	*itemE*	👍

→ *MRR: 1/2 -> 0.5*

Final Result is the mean of all MRRs

EVALUATING THE RECOMMENDATIONS

SPEARMAN RANK CORRELATION

u_j

Position i	*Recommender Top-5 List*	*Ground Truth* Top-5 List*
Position 1	*itemA (4.9) -> 1*	*itemB (4.5) -> 1*
Position 2	*itemB (4.8) -> 2*	*itemA (4.4) -> 2*
Position 3	*itemC (4.7) -> 3*	*itemC (4.3) -> 3.5***
Position 4	*itemD (4.5) -> 4*	*itemE (4.3) -> 3.5***
Position 5	*itemE (4.4) -> 5*	*itemD (3.8) -> 5*

	REC	GT	PEARSON
A	1	2	
B	2	1	
C	3	3.5	**0,7182** ***
D	4	5	
E	5	3.5	

**i.e. Based on known user j ratings (multi-level) for movies*

*** Average position of the positions where the rating is equal -> 3+4 / 2 -> 3.5*

**** Pearson correlation of the two vectors, that are the positions of the items in the predicted vs correct recommendations rank*

1 = perfect correlation
0 = no correlation at all
-1 = inverse correlation

EVALUATING THE RECOMMENDATIONS

DISCOUNTED CUMULATIVE GAIN (DGR)

*Define **utility function.***
Utility of an item for a user -> i.e. By rating, by click/not click, time listening …

*Define **discount function.***
Probabilty of user to click on item based on position -> i.e. exponential decay model

***Normalize** DGR with perfect DGR*

$$DGR(R) = \sum_{i}^{n} u(i) * d(i) \qquad nDGR(R) = \frac{DGR(R)}{DGR(R_{perfect})}$$

EVALUATING THE RECOMMENDATIONS

DISCOUNTED CUMULATIVE GAIN (DGR)

The Utility:

When Items have ratings -> Just the rating number itself
When items have Like / Dislike -> 1 Like, 0 dislike

The Discount:

Common: 1 / min (1, log_2 i) -> First the two items is 1 (no discount) then decays
Other: Half life, based on probabilistic model, with exponential decay function

You can choose your owns for each function.

EVALUATING THE RECOMMENDATIONS

FRACTION OF CONCORDANT PAIRS (FCP)

What number of pairs of items are in the right order?

	REC	GT
A	1	2
B	2	1
C	3	3.5
D	4	5
E	5	3.5

10 pairs: *AB,AC,AD,AE,BC, BD,BE,CD,CE,DE*

Right order: *AC(2<=3,5),AD(2<=5),AE(2<=3,5) BC(1<=3,5),BD(1<=5),BE(1<=3,5) CD(3,5<=5),CE(3,5<=3,5)*

FCP = 8 / 10

LAB 3

IN-CLASS

LAB 3 - EVALUATION METHODS

1) Evaluating predictions (MAE, MSE, RMSE) using Error ratings average

2) Evaluating Users ratings average predictions, only considering RMSE

3) Evaluating Ranking - Mean Reciprocal Rank (MRR)

4) Evaluating Ranking - Spearman Rank Correlation (SPR)

5) Evaluating Ranking - Discounted Cumulative Gain (DCG)

EVALUATING THE PRODUCT

what's the status of the product by the end user?

(whenever possible long term metrics much better than short term metrics)

- **COVERAGE (how often a recommendation is available?)**
- **RECOMMENDATION UPTAKE**
- **SATISFACTION**
- **LTV (LIFE TIME VALUE)**
- **METRICS FOR PIRATES (AARRR)**
  - **ACQUISITION**
  - **ACTIVATION**
  - **RETENTION**
  - **REVENUE**
  - **REFERRAL**
- **ONE METRIC THAT MATTERS (OMTM)**

*Lean Analytics. AARRR Metrics for Pirates.

EVALUATING THE BUSINESS

FINAL RECOMMENDATION GOAL

HAS A BUSINESS PURPOSE

It is bringing more money?

- **CROSS-SALES**
- **UP-SALES**
- **CONVERSIONS**

REAL-WORLD CASES

BOOKING CASE

Building a Testing Culture

Stuart Frisby - Director of Design at Booking.com

https://www.youtube.com/watch?v=_sx5LV23hIE

A/B testing principles

- If it can be a test, test it
- No platform goes untested
- No HIPPOs
- Teams are made for testing
- Everyone gets 100% access
- Hypotheses > ideas
- Test Small
- Guidelines, not Rules
- 9/10 tests fail
- You still don’t know WHY!
- You still have to spend on research
- Customers Drive the Product

IN CLASS EXERCISE

In Group exercise

Given a set of retail banking data

Volume: 2 years of transactions

Available data:

- **Customer Profile :** Customer Id, Age, Gender, Postal Code
- **Merchants Profile :** Merchant Id, Industry Id, Name, Postal Code, Brand Id (opt. FK to Merchant)
- **Customer Transactions :** Customer Id, Merchant Id, Tx Description, Amount, Currency, Date
- **Merchant Transactions :** Tx Description, Amount, Date

DESIGN A RECOMMENDATION ENGINE

IN CLASS PRESENTATION

FIN-TECH CASE

STRANDS

CLO

TOP 10 LESSONS LEARNED DEVELOPING, DEPLOYING AND OPERATING REAL-WORLD RECOMMENDER SYSTEMS

THANKS!