Data Science by Example

Problems & Solutions

Statistics & Machine Learning

The Risk of Return

The Problem

Argonaut sell: 100 Toys / yr

Profit / Toy: €10.00 given 1x Costs

Sale Price: €100.00

Costs:
- Delivery: €5.00 ← Paid 2x on Return & Resale
- Admin Costs: €5.00 fixed

Problem: 20% toys returned (r)

Impact

Max. Profit: $100 * £10$

$= £1000$

Actual: $(1 - r) * £10 = £800$

$r = 0.2$

Goal: Minimize r

How can we affect τ?

- Modify Prices?
- Reduce Return Window?
  - lets Explore...

Algorithm Proposal

What is this?

if $P(\text{Return} \mid \text{Customer Info}) > \text{RISK}$:

ReturnWindow := 7 DAYS

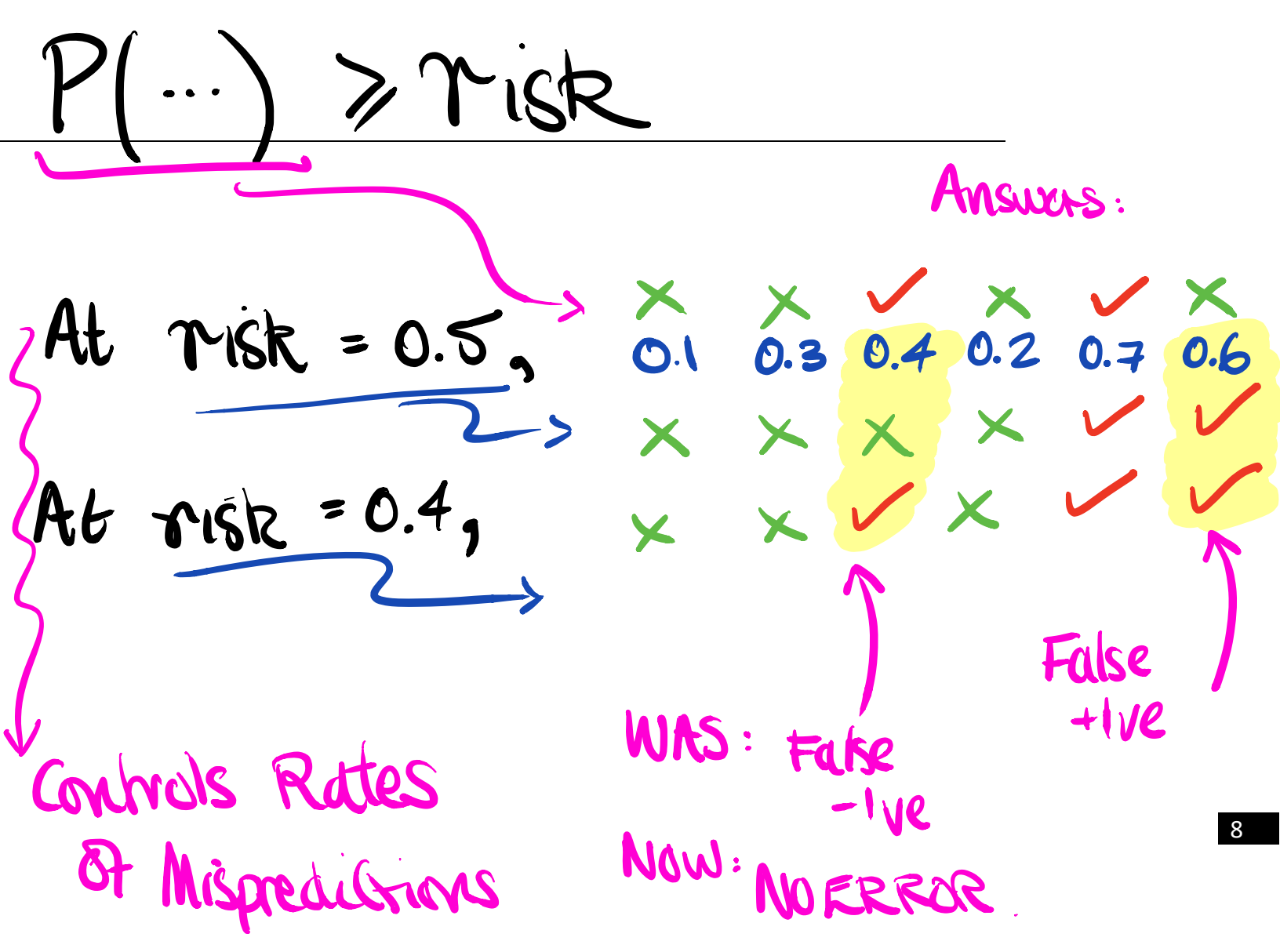
else:

ReturnWindow := 30 DAYS

$$P(\text{Return} \mid X_0 \dots X_n)$$

...?

- Age $\in [0, 100]$
- Location $\in \{UK, FR \dots\}$
- History of Returns (AVG./DAY) $\in [0, 10]$
- Freq. of Visits $\in [0, 100]$
- Gift $\in \{0, 1\}$
  2> is it a gift?


P(...) ≥ risk
Answers:
At risk = 0.5,
At risk = 0.4,
0.1 0.3 0.4 0.2 0.7 0.6
Controls Rates
of Mispredictions
WAS: False -Ive
NOW: NO ERROR
False +Ive

How do we find P(Return | X...)?

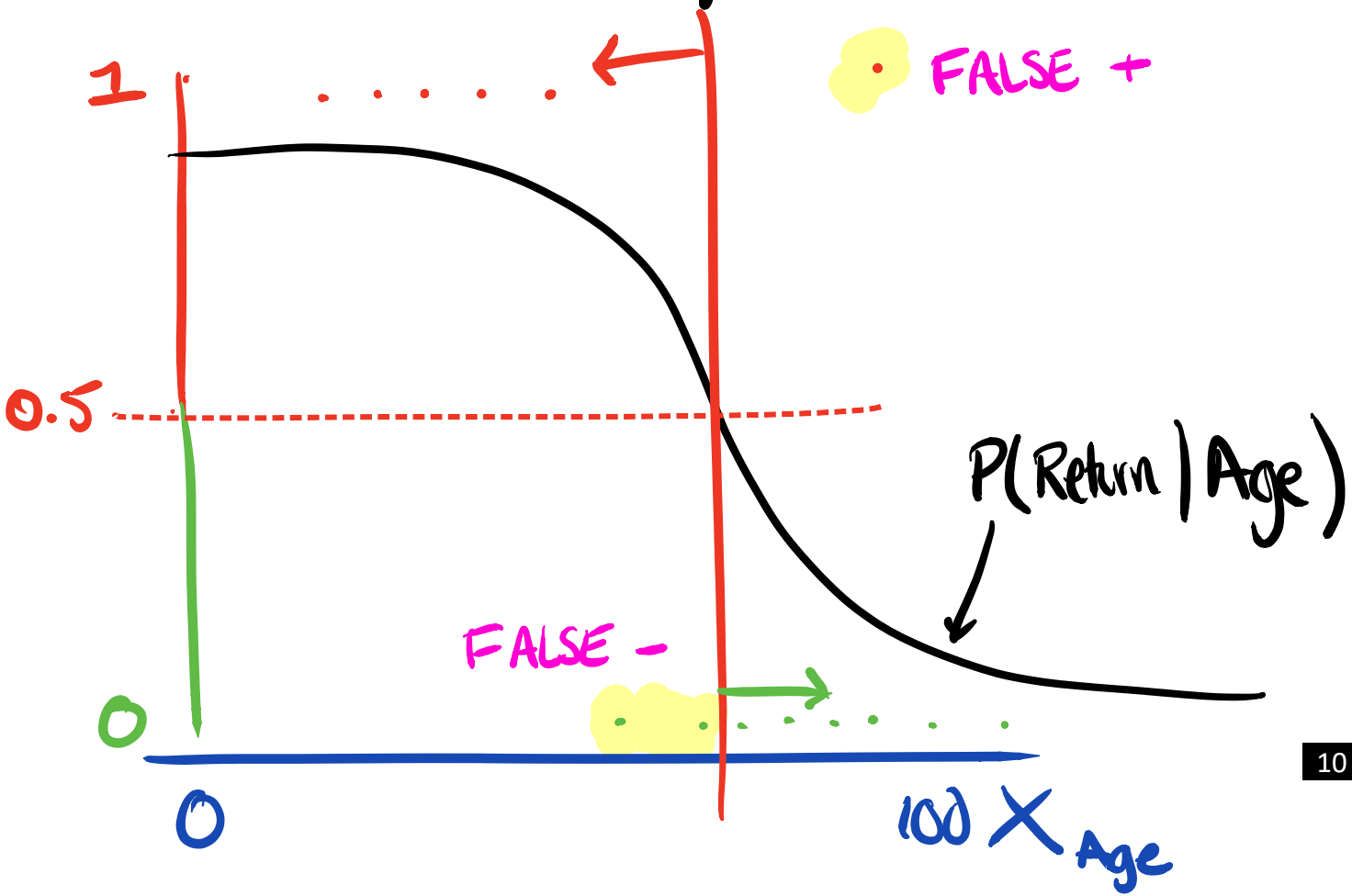
Experiment

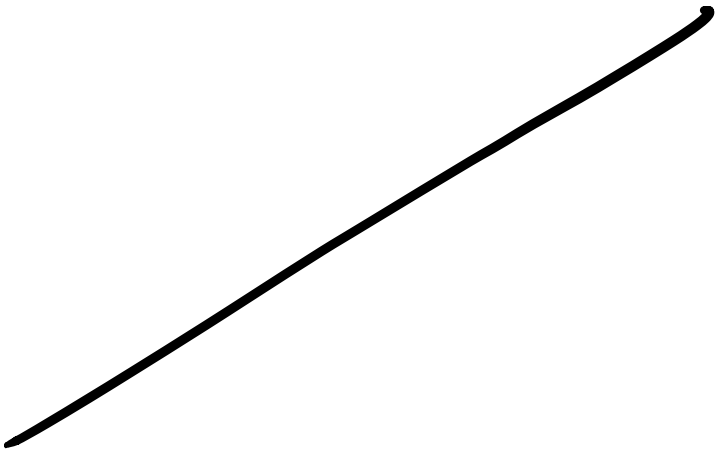
Age	Fq Return	Fq Visit	Gift?	...	Returned
30	0.01	1.2	0		0
21	0.2	3.2	0		1
...	...	...	...		...

PREDICTIVE FEATURES

RESPONSE TARGET

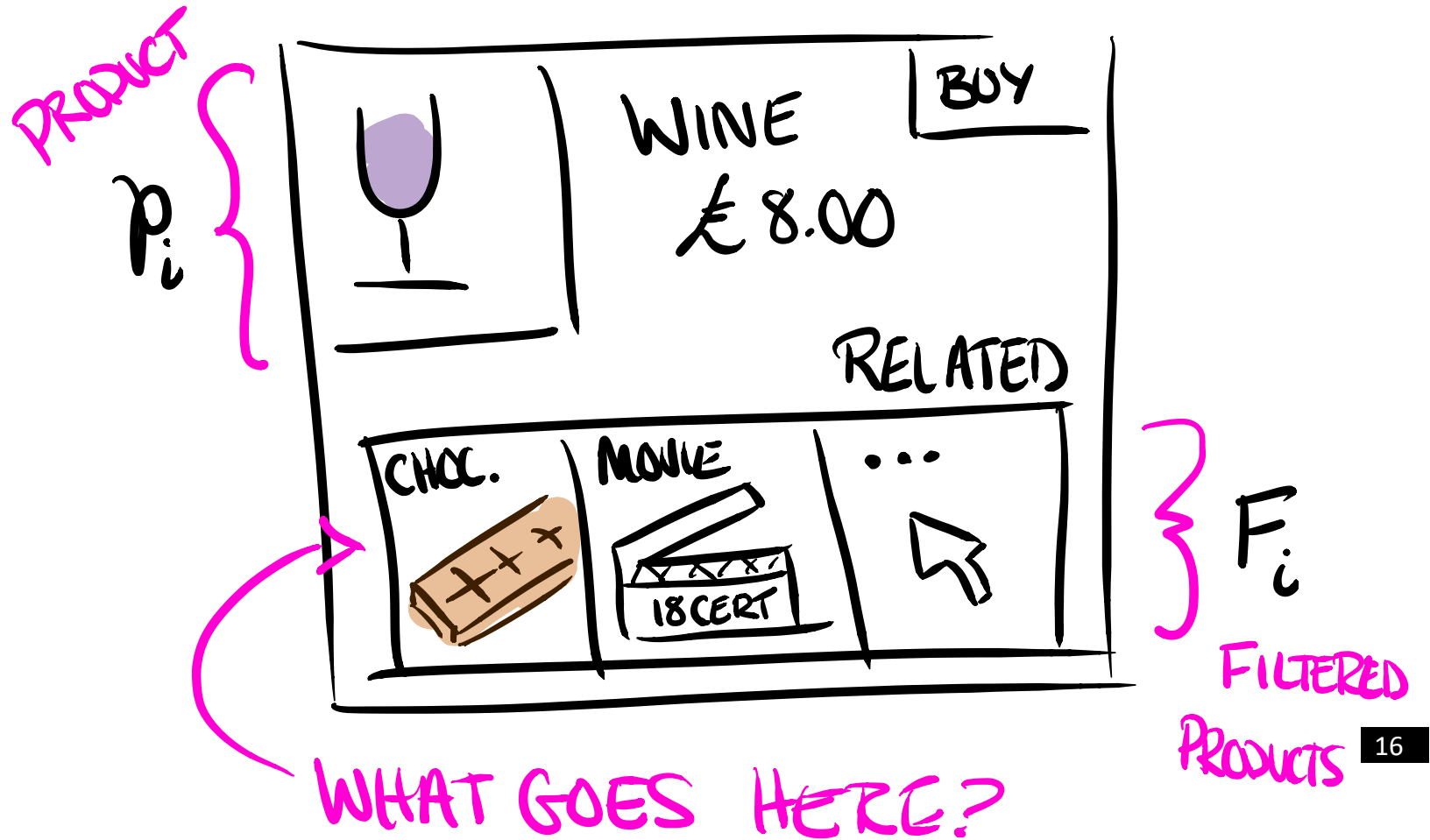
Find a Relationship





Algorithms & Graph Theory

Product Recomendation


PRODUCT
p_i
WINE
€8.00
BUY
RELATED
CHOC.
MOVIE
18CERT
...
F_i
FILTERED PRODUCTS
WHAT GOES HERE?

The Problem

Sales Team sets $F_i \subset P$ (subset)

F_i: filtered set of Related products

P: total set

Products in F_i are scored, $p_i \in F_i$ (in)

p_i: product ID

$\text{Score}(p_i) = \text{avg. Review}$

Eg.

$P = [\underset{0}{\text{Chocolate}}, \underset{1}{\text{Wine}}, \underset{2}{\text{Carrots}} \ldots]$

$F_0 = \{\text{Wine}, \text{Movie}, \ldots\}$ Related Products

$F_1 = \{\text{Chocolate}, \text{Movie}, \ldots\}$

$F_2 = \{\text{Bread}, \text{Parsnips}, \ldots\}$

Eg. Scoring

$F_1 = \{ \text{Chocolate}, \text{Movie}, \cdots \}$

Score (Chocolate) = 4.5 #1

Score (movie) = 3.8 #2

$\therefore$

When visiting "Wine"

Recommend

How Can We Improve F?

USE PROFILE:

Prob (Purchase | Age, Gender, ...)

Prob. Customer purchases

given their age, etc.

USE "SIMILARITY":

SET P := LIKED PRODUCTS OF SIMILAR USERS

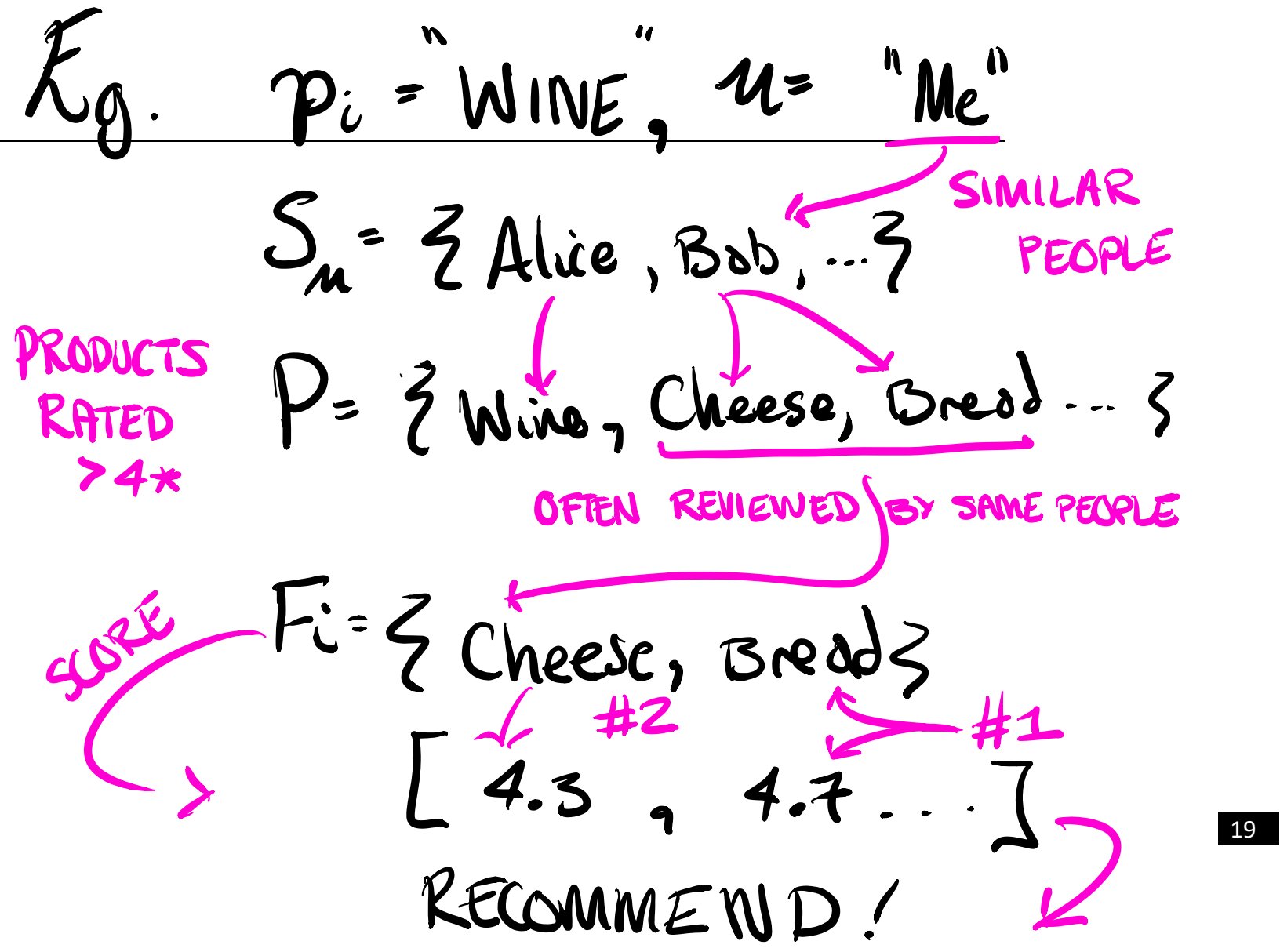
How 'alike' is customer to other purchasers?

Algorithm:

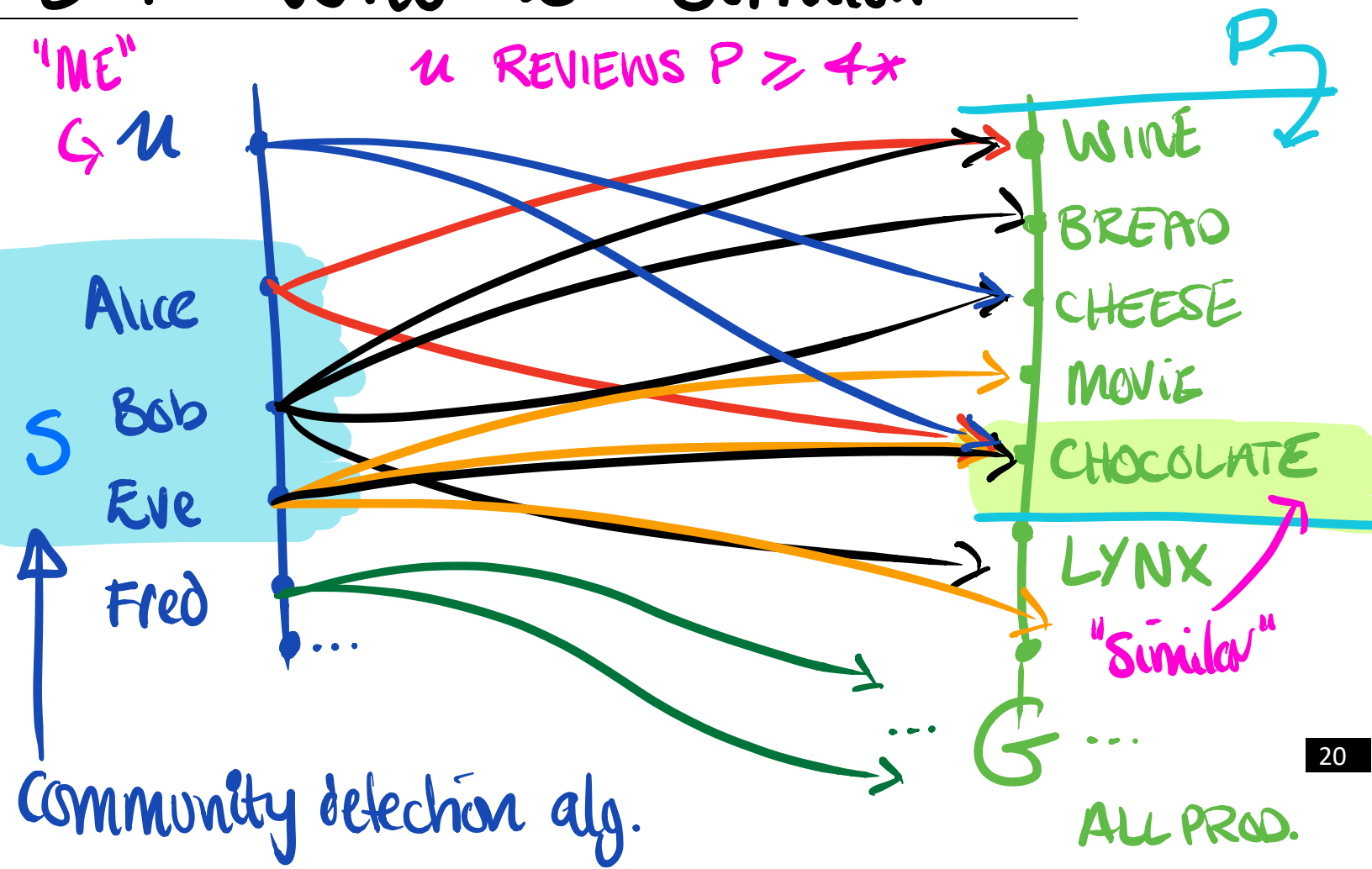
FOR USER u LOOKING AT p_i

How?

* FIND SIMILAR USERS S_u
* Set P := PRODUCTS REVIEWED BY S_u
* Set F_i := PRODUCTS OF REVIEWERS OF p_i IN P
* Score F_i BY REVIEWS & RECOMMEND.


Eg. p_i = "WINE", u = "Me"
S_u = { Alice, Bob, ... }
SIMILAR PEOPLE
PRODUCTS RATED >4*
P = { Wine, Cheese, Bread ... }
OFTEN REVIEWED BY SAME PEOPLE
SCORE
F_i = { Cheese, Bread }
#2
#1
[4.3 , 4.7 ...]
RECOMMEND!

But... Who is 'Similar'?

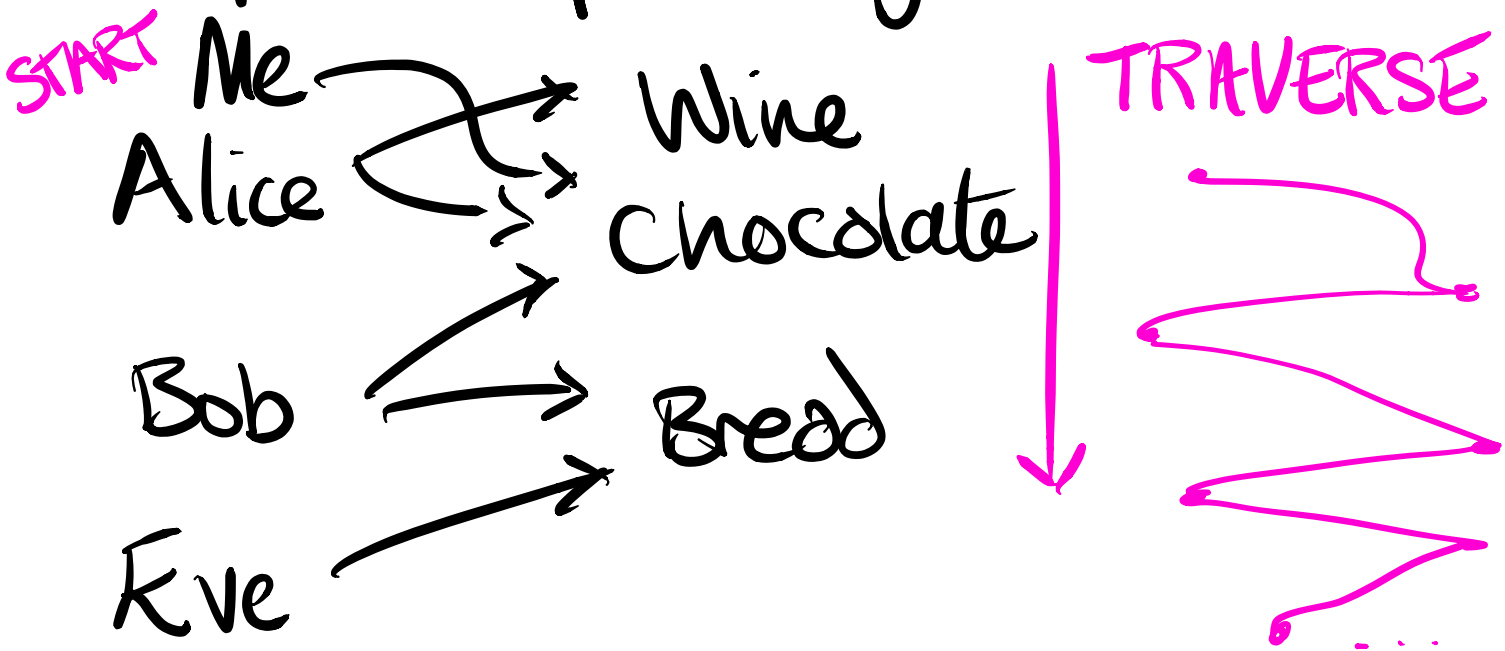


Community Detection

A 'similar' to B by $J = \frac{\#\text{Shared Reviews}}{\#\text{Reviews}}$

$\{A, B, C, \dots\}$ is a community S if $J(S) \geq$ minimum

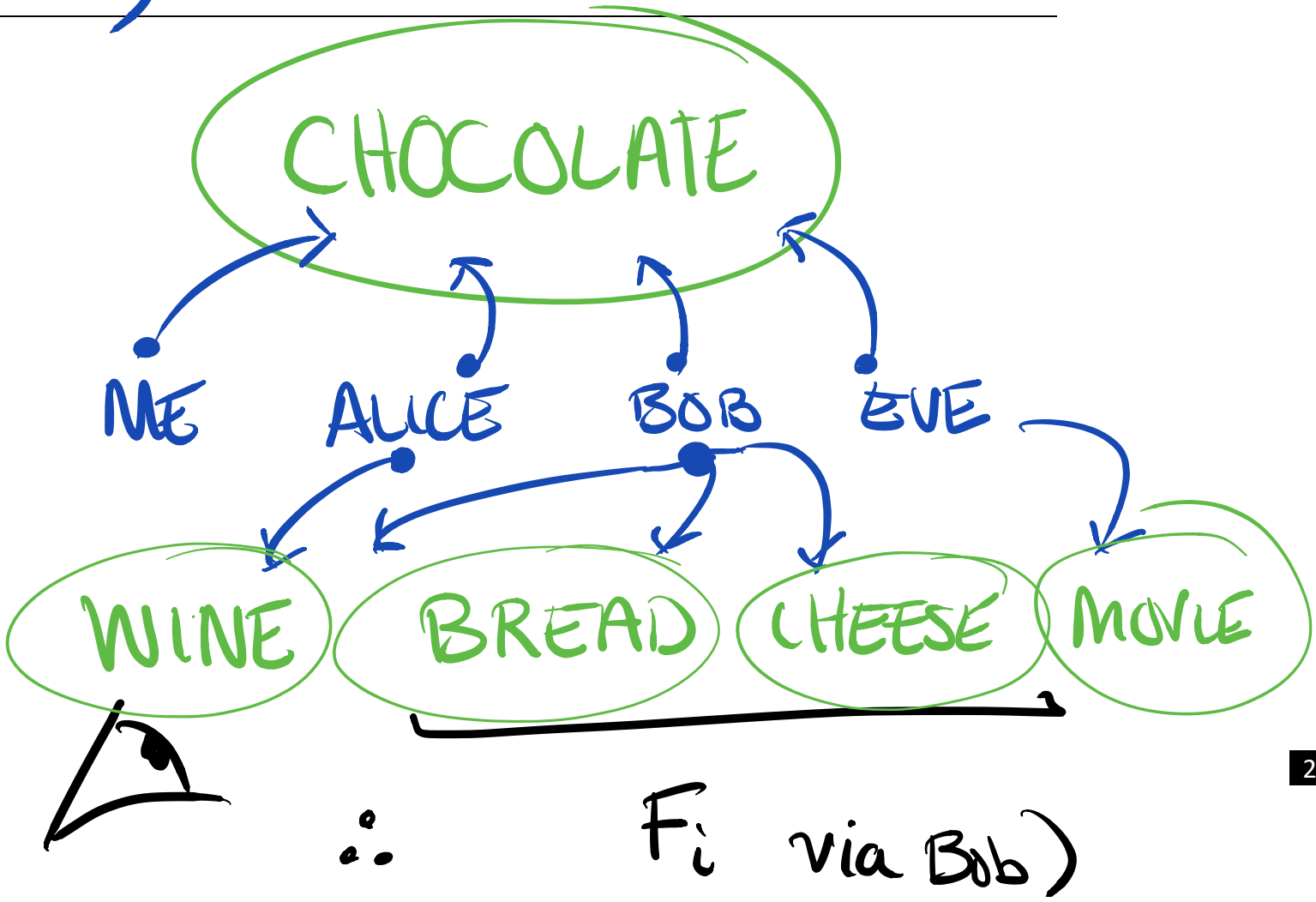
Simple Graph Alg.

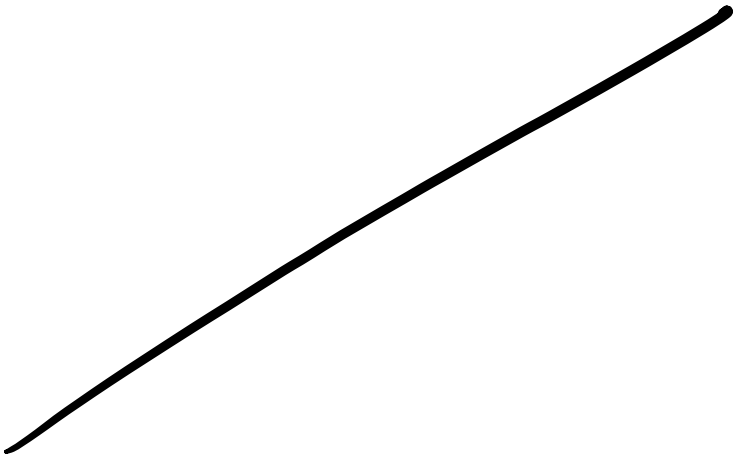


Add user u to S iff

$J(S) \geq$ minimum

So,


CHOCOLATE
ME
ALICE
BOB
EVE
WINE
BREAD
CHEESE
MOVIE
∴
Fi via Bob)



Big Data

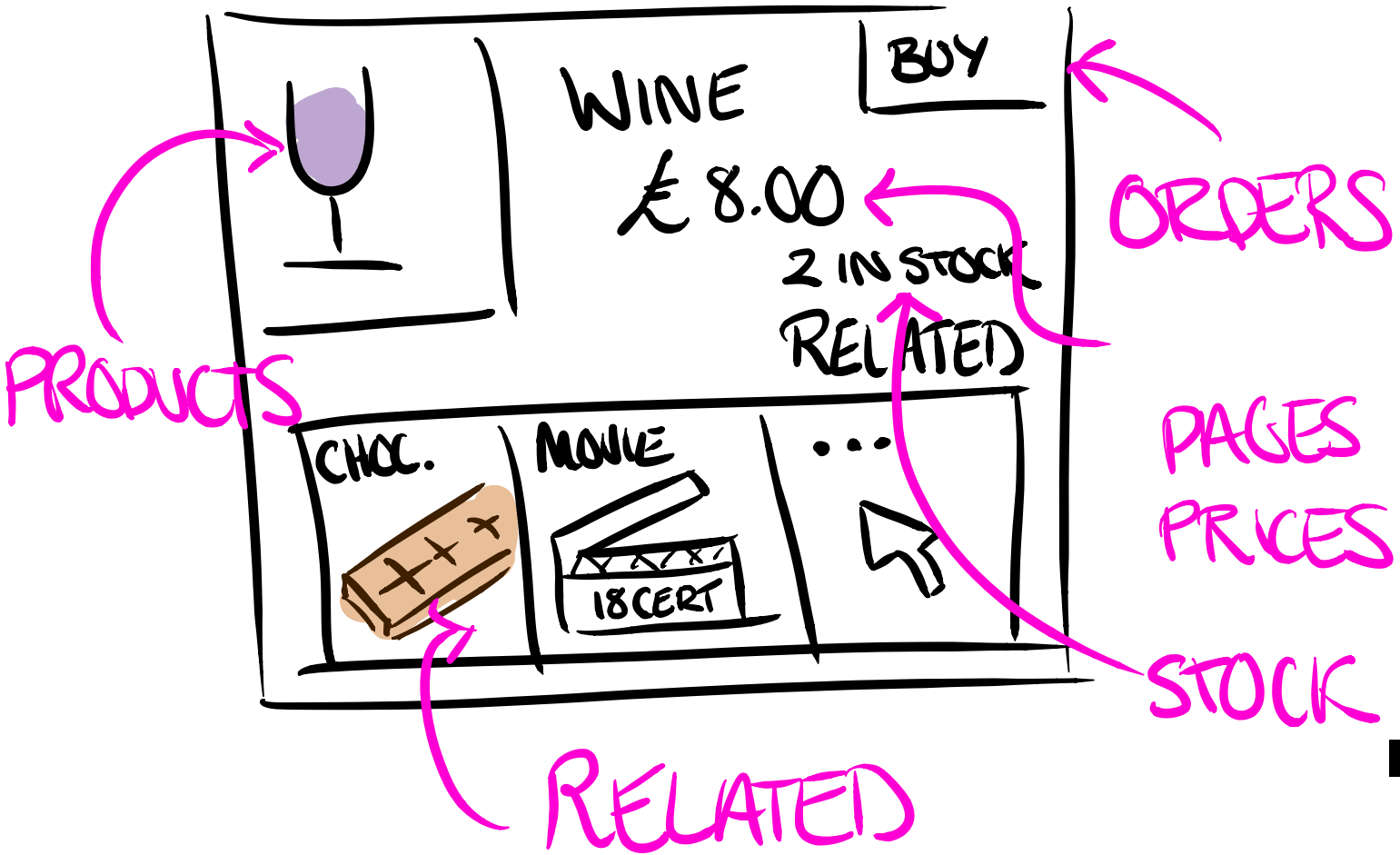
User Behavior & Event Systems

The Problem

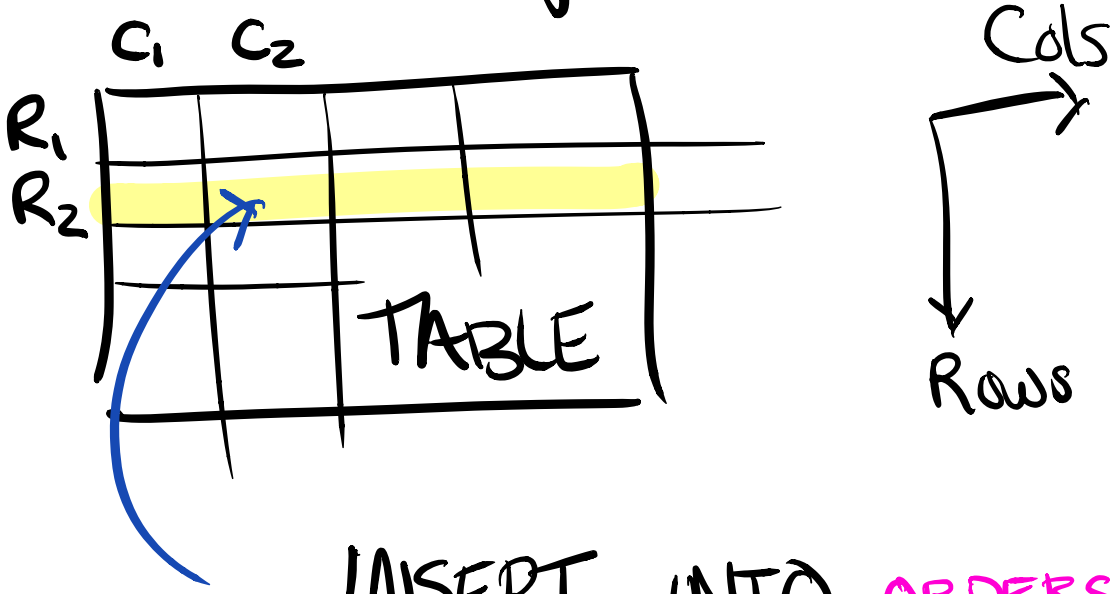
Existing data infrastructure does not track "analytical data"

Why?

Data infrastructure was designed for Retail Transactions


WINE
BUY
€ 8.00
2 IN STOCK
RELATED
CHOC.
MOVIE
...
18CERT
PRODUCTS
ORDERS
PAGES
PRICES
STOCK
RELATED

The Existing System



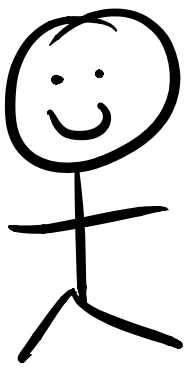
INSERT INTO ORDERS VALUES

Row = (ID, PRODUCT, QUANTITY, ...)

Essential Retail Information

What's Missing?

Event Context



12pm CAMPAIGN

→ SEARCH

→ READ REVIEW

→ ...

→ ORDER

12:10pm

Realtime Events

How do we store 'Events'?

Event =

(Subject, Verb, Object, Context)

time

CUSTOMER CLICKS AD

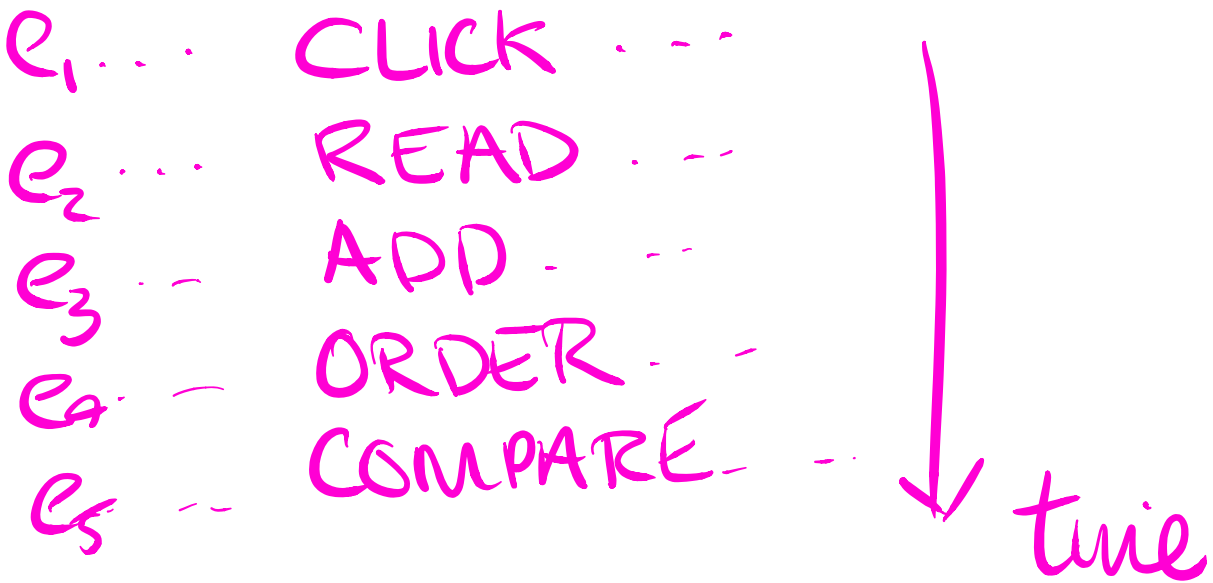
ID: 35

TIME: 12pm

⋮

Event log

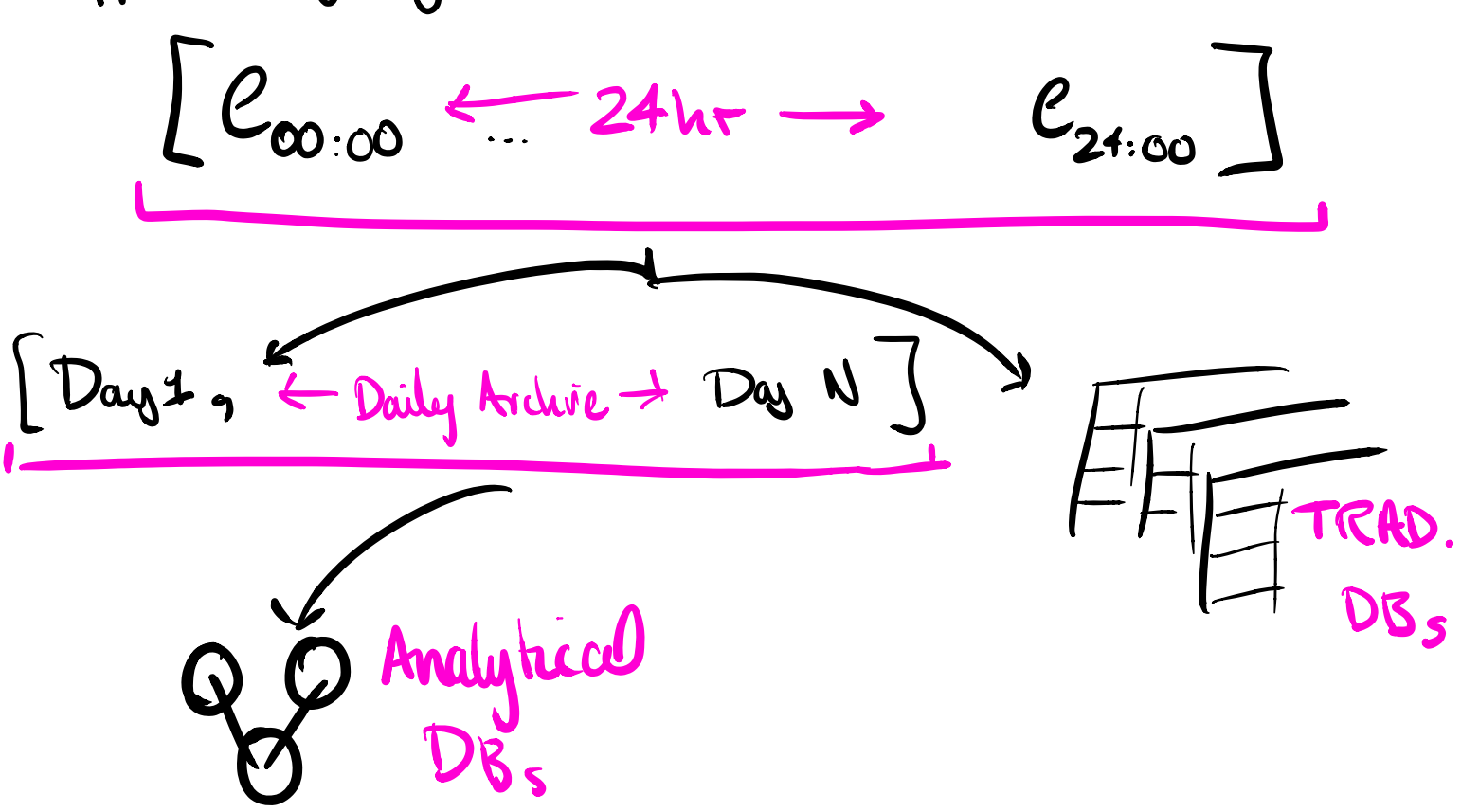
$$log = [e_1, e_2, e_3 \dots e_N]$$



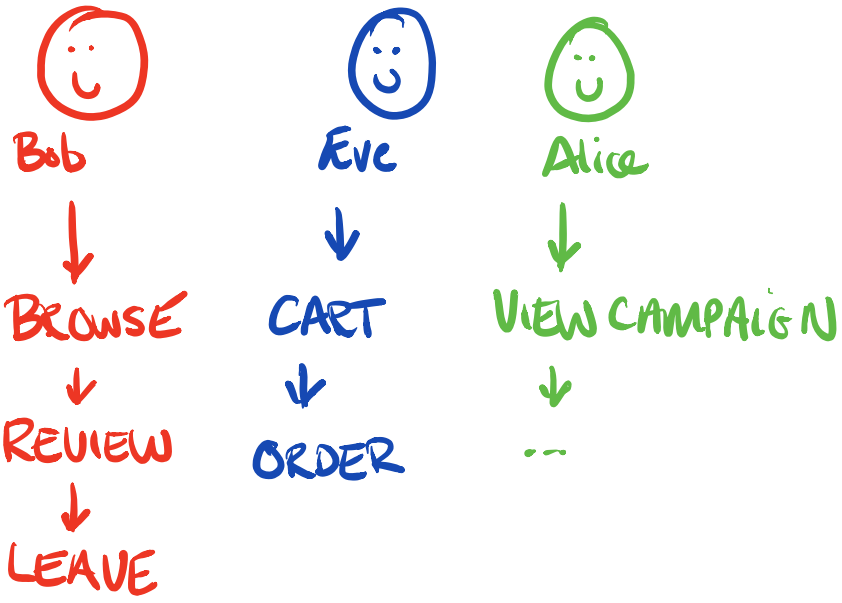
The New Data System

append-only log

CLEARED DAILY

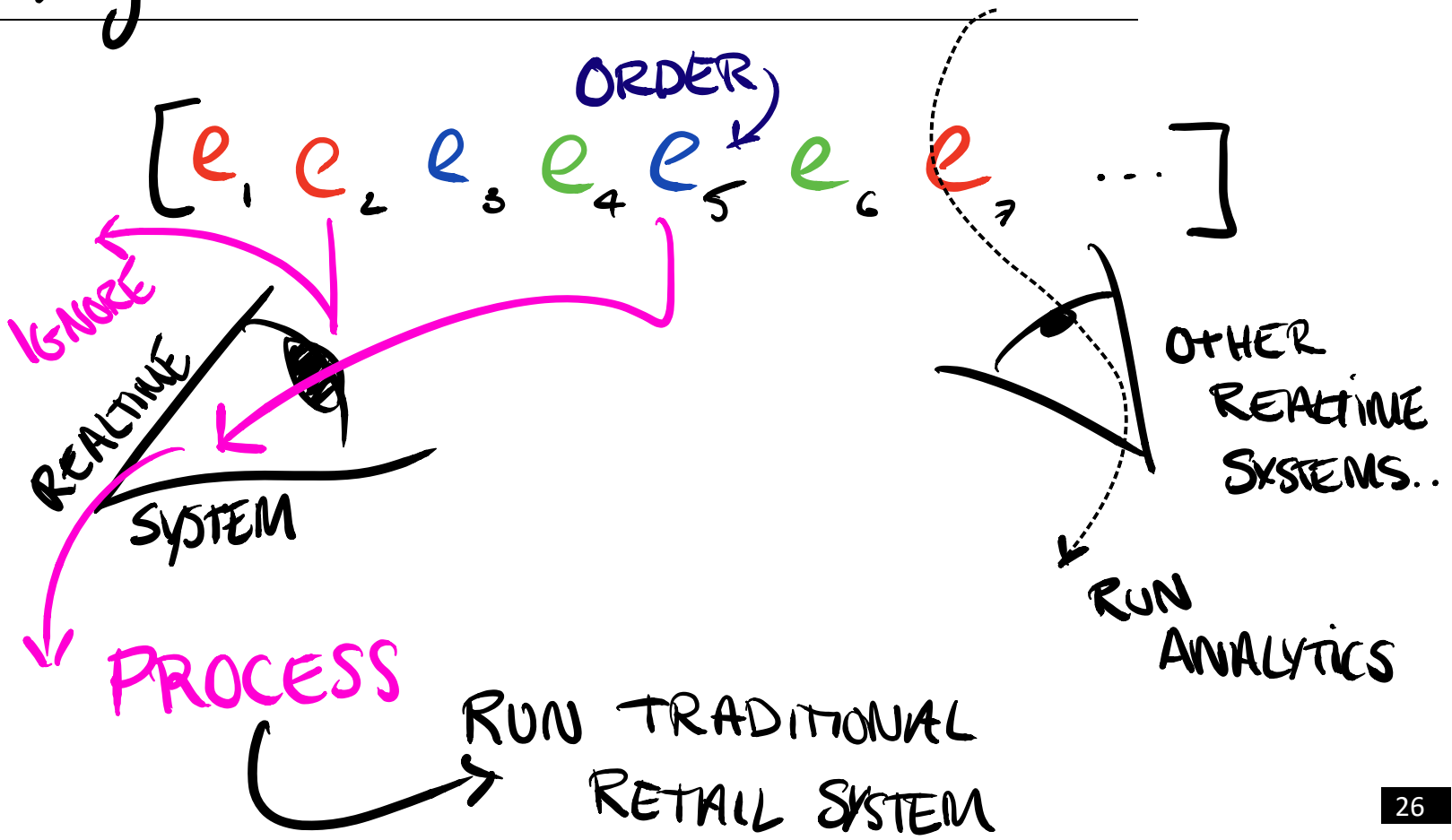


Eg.



$[e_1\ e_2\ e_3\ e_4\ e_5\ e_6\ e_7\ \ldots]$

Eg.


ORDER
e_1 e_2 e_3 e_4 e_5 e_6 e_7 ...
IGNORE
REALTIME
SYSTEM
OTHER
REALTIME
SYSTEMS..
RUN
ANALYTICS
PROCESS
RUN TRADITIONAL
RETAIL SYSTEM