Data Science By Examples

THREE AREAS

- STATS & ML → Risk $P(y|x)$?
- ALGORITHMS & GRAPHS
  - ↳ FRAUD
- BIG DATA
  - ⇒ BANKING APP.
  - "REALTIME" RESP.

APPROACH

EXISTING BIZ : BANK (LOCAL)

↳ WHAT ARE "TRADITIONAL" SOLNs?

↳ WHAT VALUE DOES A "DATA SCIENCE" SOLN ADD?

↳ WHAT IS A DSCI SOLN?

BANK

RISK

LOCAL

FRAUD

- DEFAULT RISK
- LACK OF RESERVES
- BANK RUN
- LIQUIDITY

LOANs : ₹1,000

6mo , 12mo

AA
BB
DF

RISK OF DEFAULT.

PREDICT WILL DEFAULT ?

X Y

$$P(Y \mid X)$$

WHAT INFO ABOUT A CUSTOMER "RELEVANT TO" WHETHER THEY WILL DEFAUT?

$X_{CUSTOMER}$

INCOME

CREDIT SCORE

DEBTS

EXISTING CREDIT LIMITS

DEFAUTS

CCS, BANKRUPCY.

DEPENDENTS.

$X_{ECONOMY}$

INFLATION, ..

HOUSE PRICES

CAUSAL

$X_{INCOME} \downarrow \qquad Y_{DEFAUT} \uparrow$

ASSOCIATIVE

AGE
PROFFESSION
LOCATION

MACHINE LEARNING

$$P(\underbrace{Y}_{\text{TARGET}} \mid X_{\text{CUSTOMER}})$$

OBSERVABLES

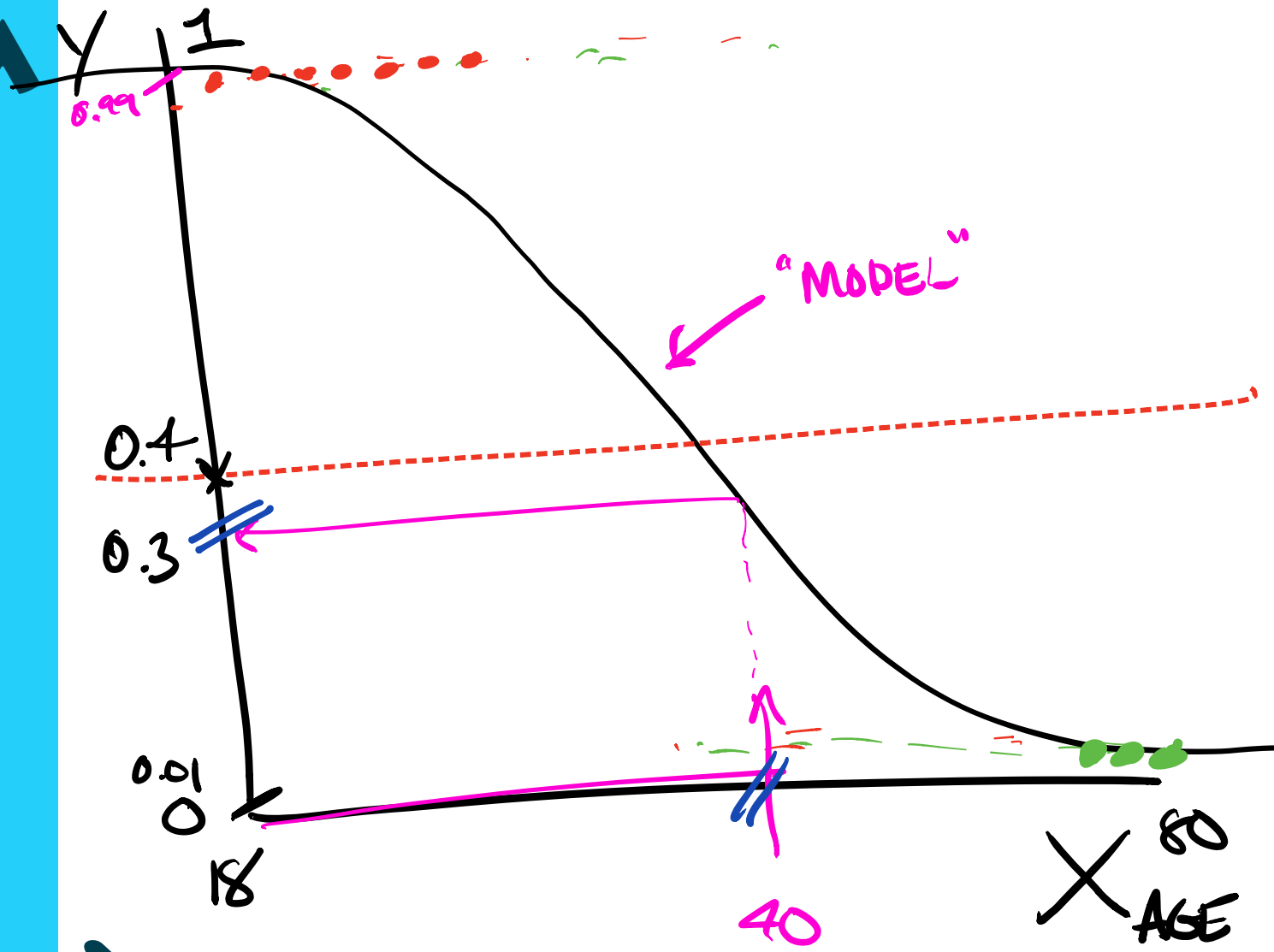
'FEATURES'

HISTORICAL DATASET

UNOBSERVABLE 'TARGET'

X_{AGE}	...	X_{SCORE}	...	Y
30	...	600	...	1
32		600		0
31		550		0

HISTORICALLY OBSERVABLE


Y
1
0.99
"MODEL"
0.4
0.3
0.01
0
18
40
60
X
AGE

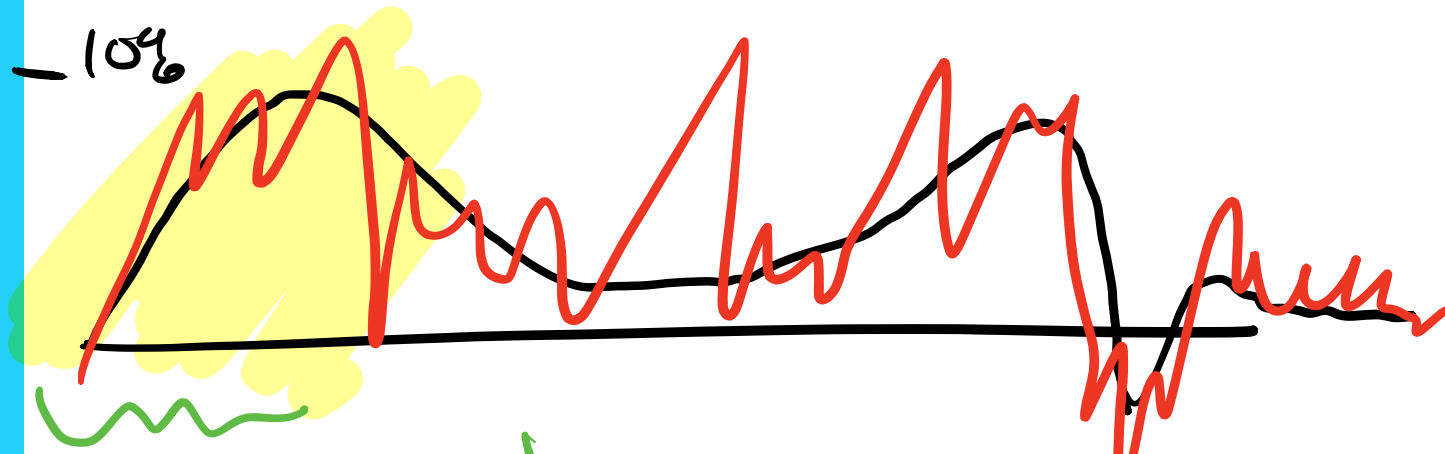
IF $P(\text{DEFAULT} \mid \text{AGE}) \geq 0.4$:

DEFAULT

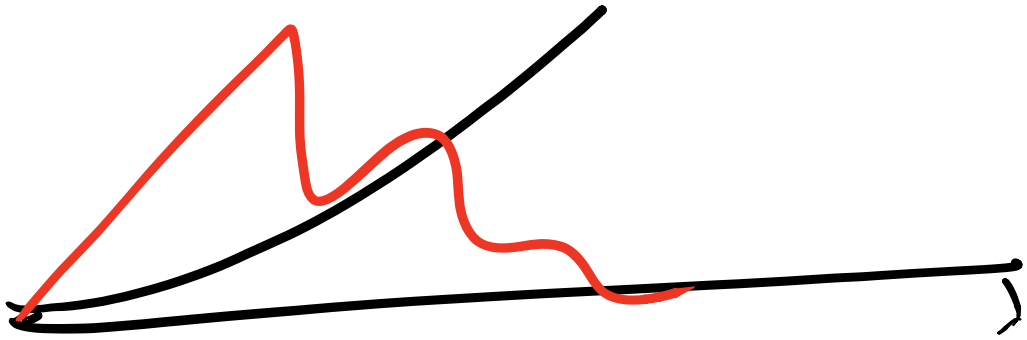
ELSE :

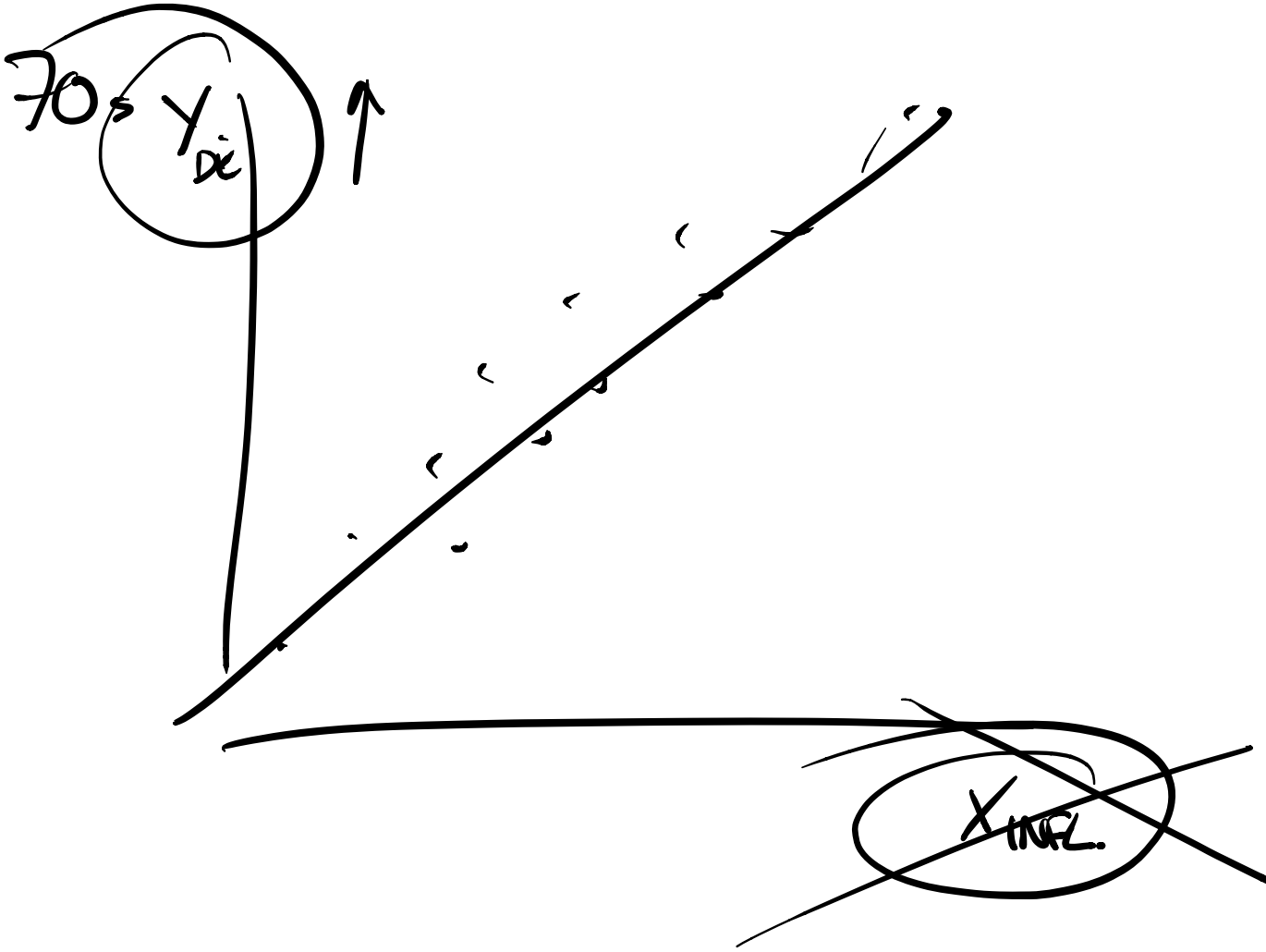
OK

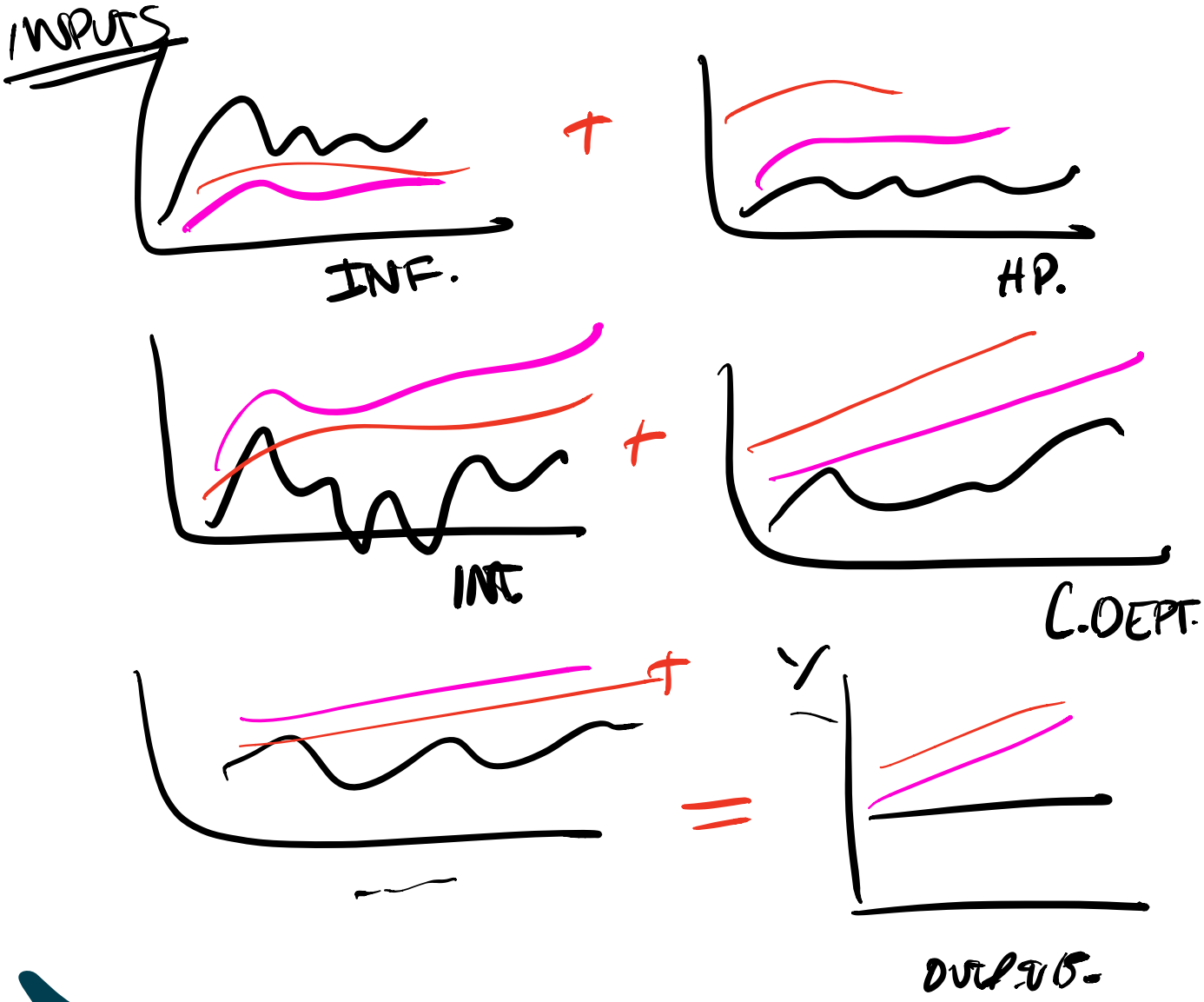
"MACRO ECONOMIC EVENTS"



UNIQUE ECONOMIC COND.




70s
Y_{DC}
$X_{INFL.}$


INPUTS
INF.
HP.
INT.
C.DEPT.
Y
OUTPUTS

ML

⇒ "AUTOMATED STATS"

∴ HAS EXPERT KNOWLEDGE

∴ LACKS ANY ABILITY TO "SUPPOSE"

QUESTIONS?

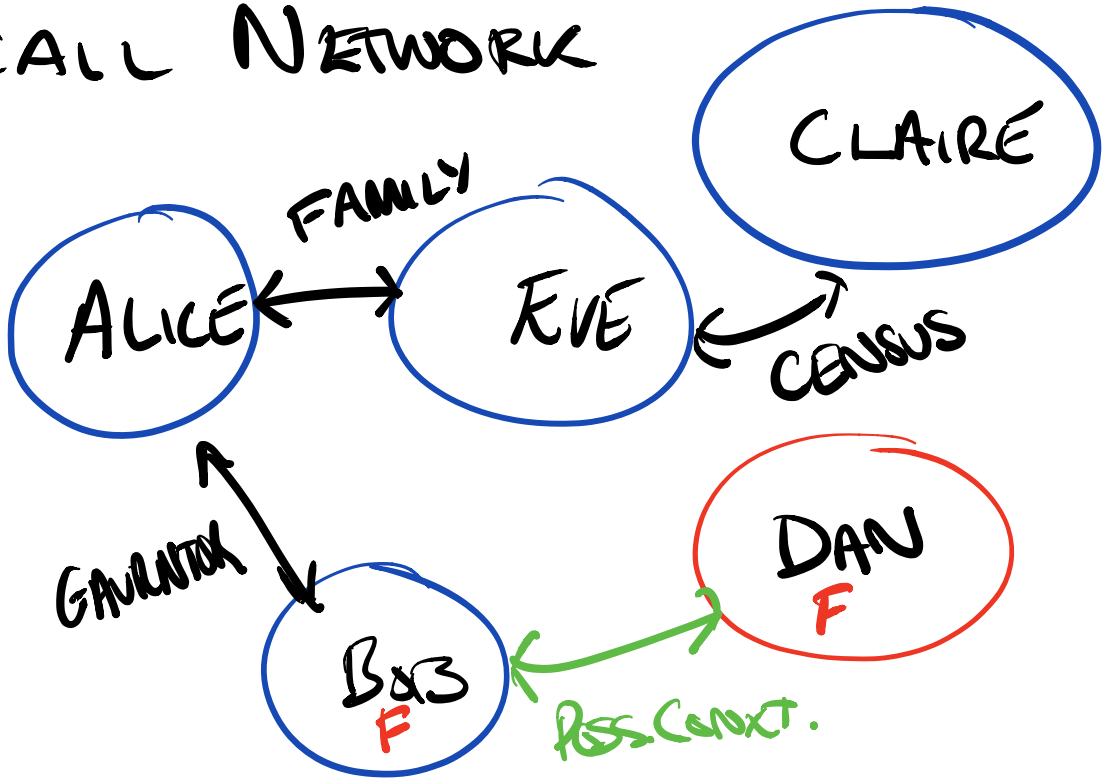
GRAPH THEORY.

BANK CUSTOMERS.

LOAN X INCOME ...

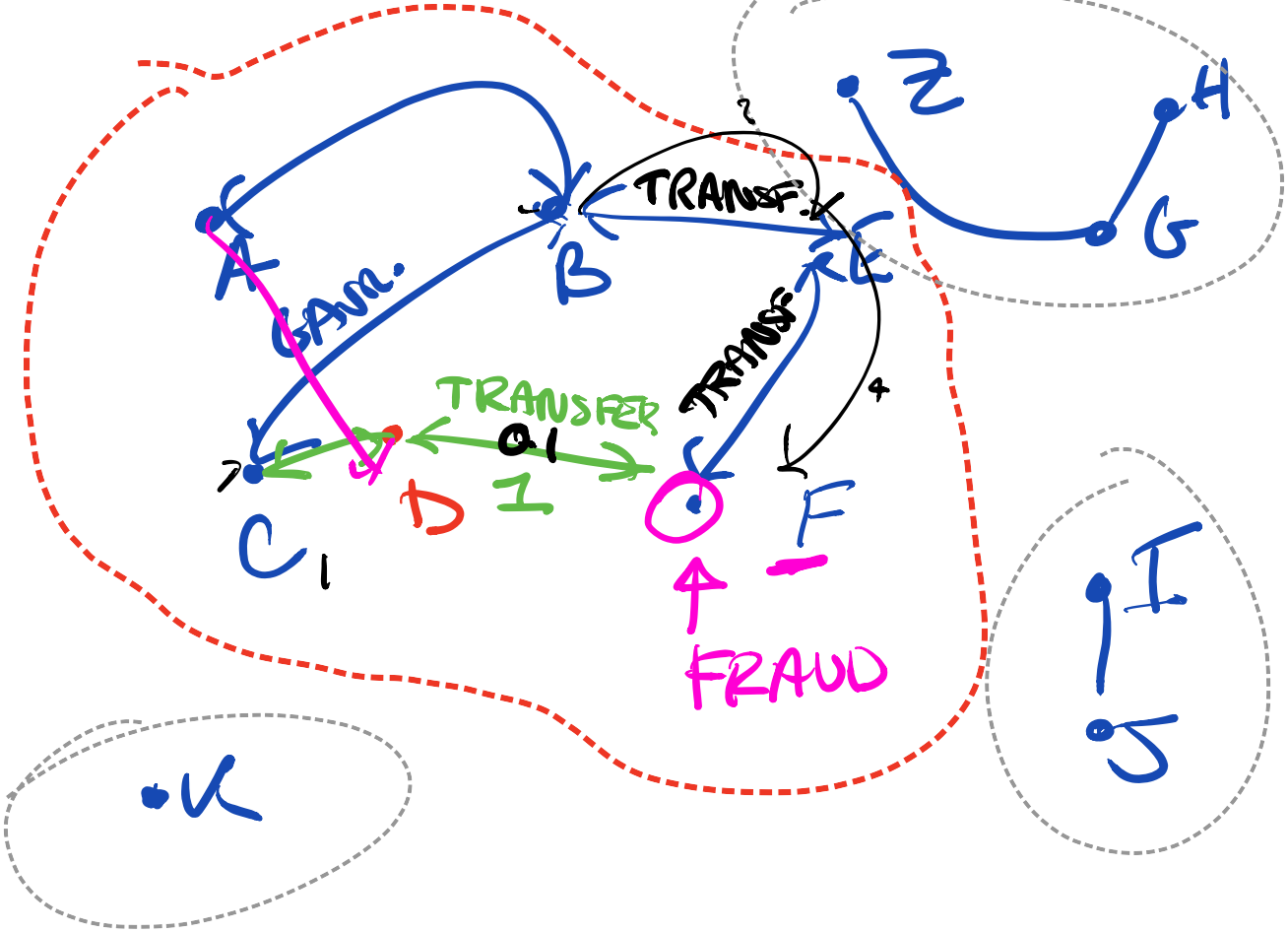
MISREPRESNT

SOCIAL NETWORK



Q1

X =



SURNAME
+
POST CODE

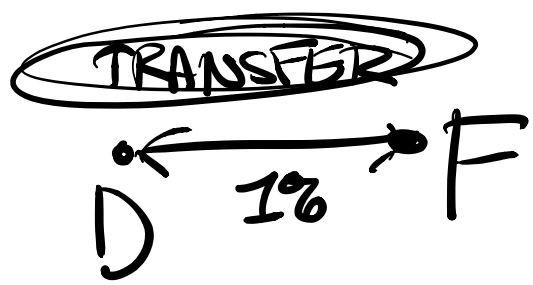
$$P(Y_{D\,FRAUD} \mid X_{GRAPH}) = P(Y \mid X_F)$$

$$X_F = \text{Connected Component}(X, F)$$

$$P(Y \mid X_F)$$

$$X_{F4} = X_F - \text{Edges}(F, D, 4)$$

$$X_{F4} =$$



TRANSFERE	FRAUD
1500	✓
200	✗
50	✗
1000	✓

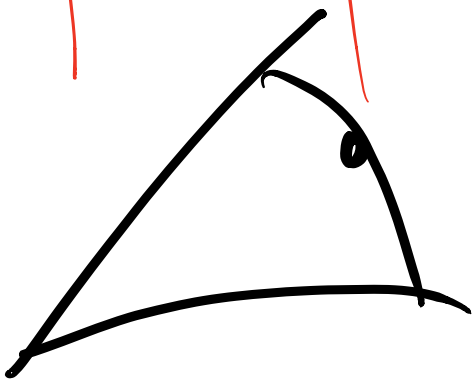
< 1000	✗
> 1000	0.01

BIG DATA

~~EXCEL~~
~~SQL~~

↳ REALTIME ANALYTICS

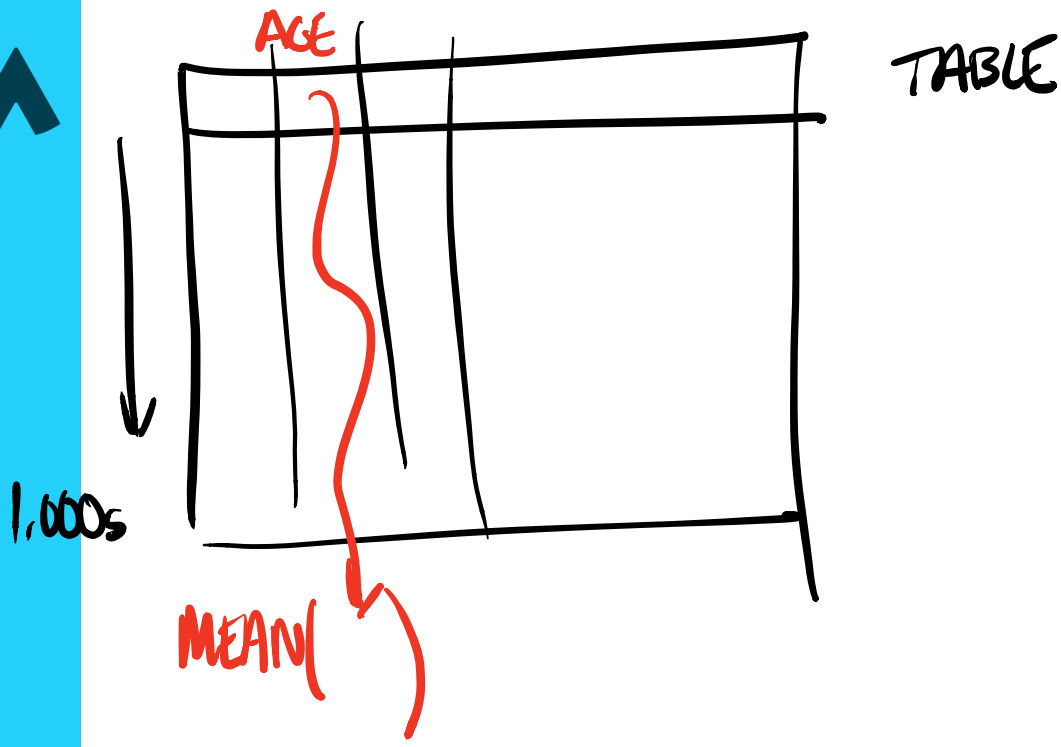
EVENT$_1$, EVENT$_2$, ...



STREAMING

INGEST

Stream.


AGE
TABLE
1,000s
MEAN(

$t = 0$

18 30 40 27 ...

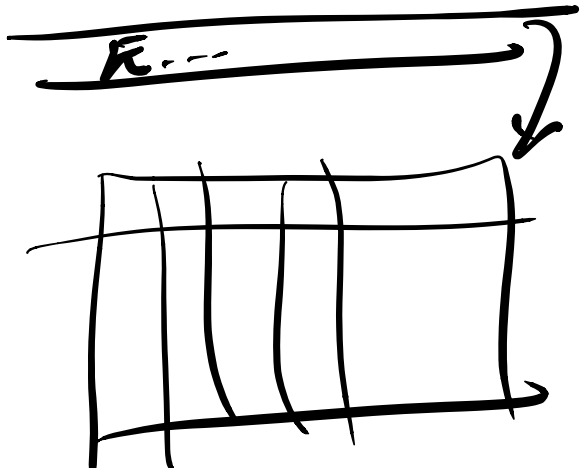
18

$\frac{18+30}{2}$ $\frac{18+30+40}{3}$ $\frac{18+30+40+27}{4}$.

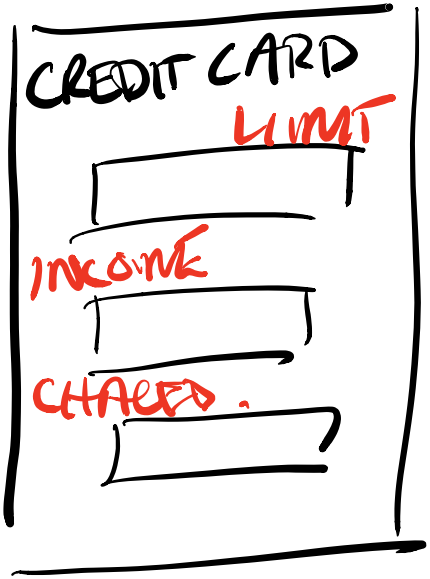
mean

18, 26, 31, 29

BATCH

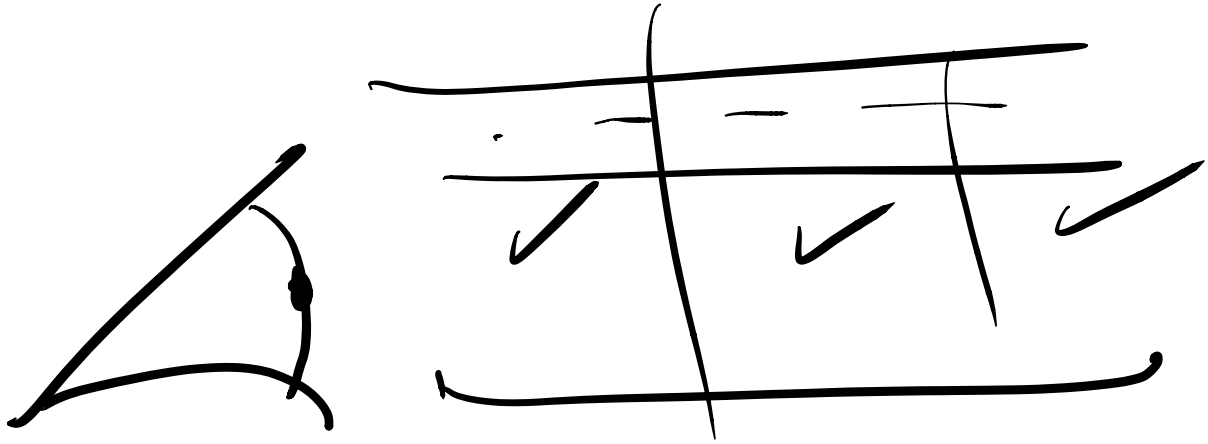


BANKING APP.



FIRED

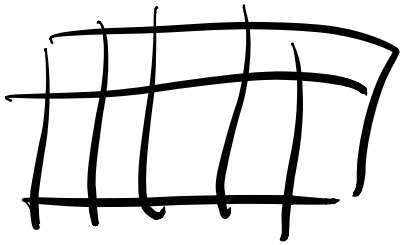
$EVENT_{LIMIT\ INC}$	E_{LIMIT}	E_{LIMIT}



$Count_A = 3$

$P(Y|A)$ IF $Count > 3$

$E_c \; E_c \; E_c \; E_{sys}$



BATCH