CRACK THE LOVE CODES

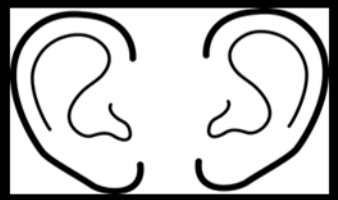
WTF is Machine Learning – Non-technical discussion

VietStartup – Hoa Giang

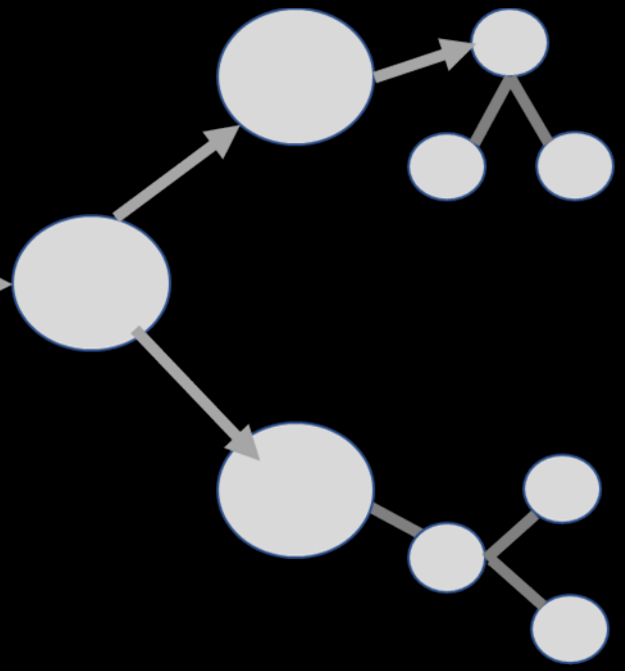
MACHINE LEARNING IS MORE HUMANE THAN YOU THINK

- ARTIFICIAL INTELLIGENCE OR NATURAL STUPIDITY?
- WHAT IF HUMAN IS JUST AN ADVANCED MACHINE?
- EVOLUTION OF MACHINE LEARNING TAKES AFTER HUMAN LEARNING PROCESS?

Infant brain learning modelling



"IF-then-DO"
REACTION RULES



Sensors to recognize conditions

Knowledge about how to react

Motors to carry out actions

Recognize a problem type

Active a way to think

Source: Emotion Machine, Marvin Minsky (2006)

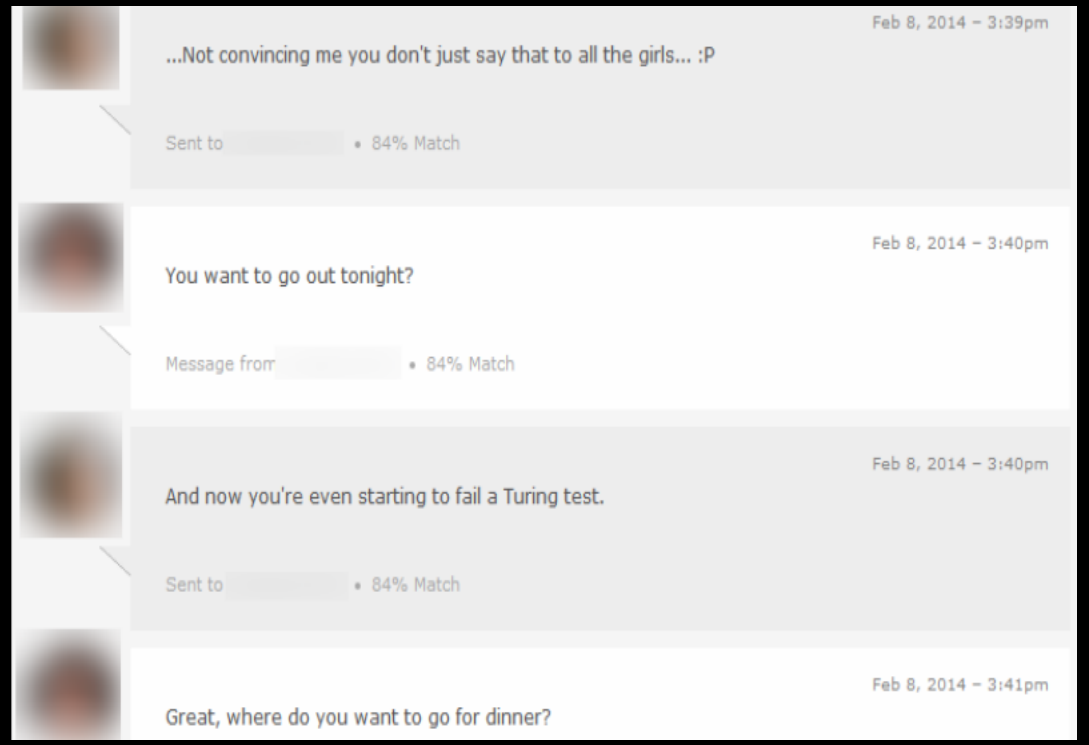
Critics

Critic-Selector based Machine

Selectors

MACHINE LEARNING IS MORE APPLICABLE THAN YOU THINK (A.K.A YOU USE THIS MORE OFTEN THAN YOU THINK)

- Have you ever tried to classify a boy/girl as friendzone/potential dating zone?
- Have you ever tried to cluster a group of people into smaller subsets?
- Have you ever tried to predict what are the next words that your crush will text you?
- Have you ever tried to perform pattern recognition to recognize signs of cheating in your relationship?
- Have you ever chatted with a chatbot?



BECAUSE WE TALKED ABOUT LOVE. LET CRACK THE LOVE CODES

BLACK EYED PEAS ASKED:
WHERE IS THE LOVE?

WELL, WE GONNA ANSWER IT TODAY.

SORT OF…

CASE STUDY: DATING BUSINESS - WHERE IS THE LOVE? (1)

Disclaimer: The data is from OKCupid and not for commercial purposes

I love to ask questions. And we could answer them together?

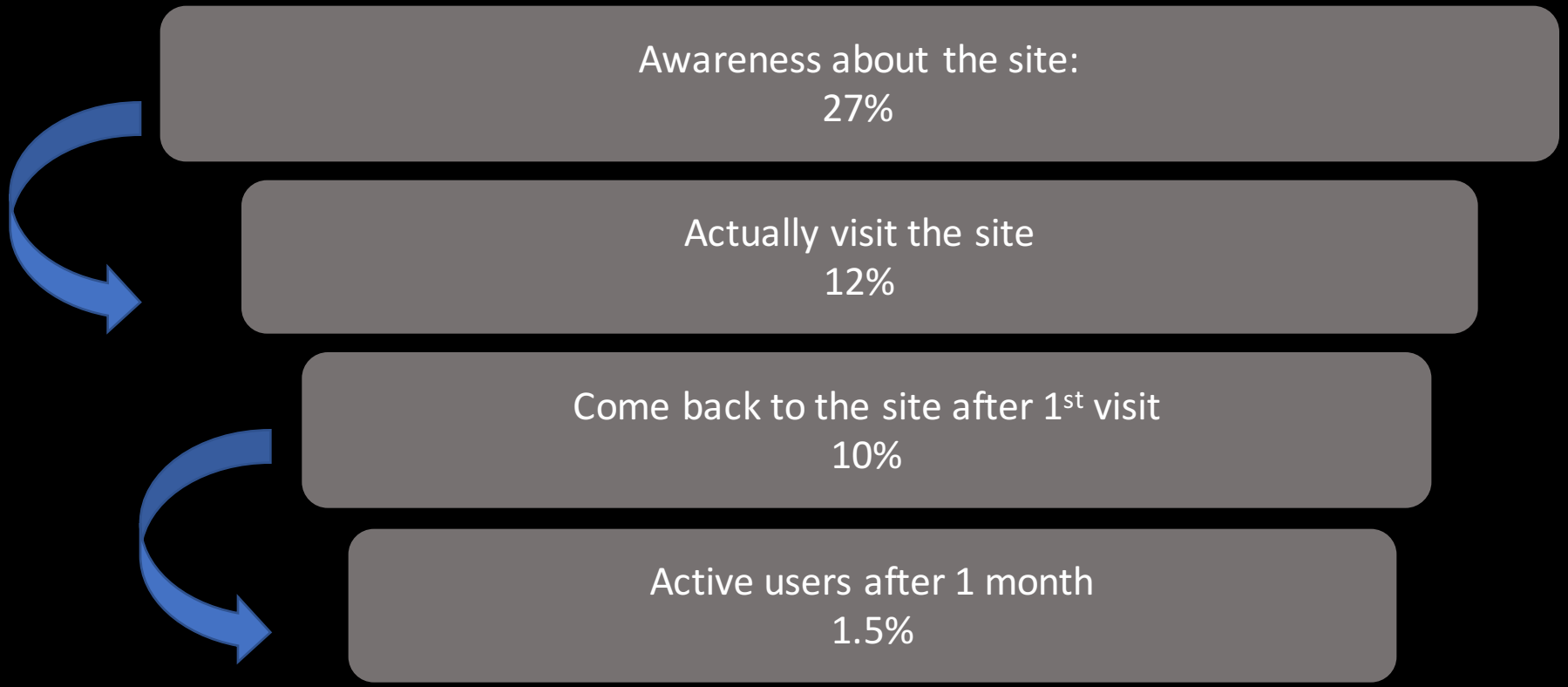
BUSINESS CASES:

- We are now working for a dating company which enjoy benefits from successful matches among users – indicating by the change of status.

  (assuming all the changes of status are recorded properly)

- The market is super competitive because the idea is easy to be cloned and users are very much swayed by advertising activities.
- We did a user survey and found out that we suffered from bad tunnel analysis. Site awareness and number of visitors are good but users mostly left after one month if they could not find the best matches.

CASE STUDY: FUNNEL ANALYSIS (2)



WHAT IS THE BUSINESS QUESTION HERE?
- SO I WOULD BE ABLE TO COME AND IMPRESS MY BOSS?

USER RETENTION

Tadahh! Let's talk in bosses' language!
"We need to identify our problem and find a way to tackle it"
"YES SIR"

WAIT…
WE ARE IN DATING BUSINESS.
THAT MEANS WE NEED TO CREATE MORE LOVE FOR THIS WORLD.
I FEEL LIKE CUPID FOR REAL!!!!!

more love, please ..

HOW TO TRANSLATE BUSINESS QUESTION INTO A DATA QUESTION?
- SO THAT THE @#$%^&* DATA SCIENTIST COULD UNDERSTAND?

- WE NEED TO UNDERSTAND WHICH FEATURES THAT MIGHT AFFECT THE STATUS OF AN USER (SINGLE/NOT SINGLE) AFTER A MONTH SO THAT WE COULD PAY ATTENTION TO HIGH CHURNING RISK USERS?

Now you know why data scientists are called data wizards?
Because we try to kill Snow White before she meets the prince – We need to tackle the HIGH RISKS!!!!

HOW TO PREPARE AND MODELLING THE DATA? A.K.A HOW TO DEAL WITH THIS MESS

- We have the data as provided in the case sheets

user_ID	age	height	income
id1	35	70	80000
id2	23	71	20000
id3	28	72	40000
id4	40	71	60000
id5	22	65	20000
id6	35	73	150000
id7	21	71	20000
id8	34	69	100000
id9	29	67	40000
id10	25	78	80000
id11	41	69	1000000
id12	39	70	100000
id13	21	72	500000
id14	21	70	1000000
id15	27	68	150000
id16	27	68	150000
id17	50	73	80000
id18	59	73	80000
id19	24	67	60000

user_ID	status_aft
id1	1
id2	1
id3	0
id4	1
id5	1
id6	1
id7	1
id8	1
id9	1
id10	1
id11	1
id12	1
id13	1
id14	1
id15	1
id16	1
id17	1
id18	1
id19	1

We create a training data with *features* & *labels*

age	height	income	status_afte
35	70	80000	1
23	71	20000	1
28	72	40000	0
40	71	60000	1
22	65	20000	1
35	73	150000	1
21	71	20000	1
34	69	100000	1
29	67	40000	1
25	78	80000	1
41	69	1000000	1
39	70	100000	1
21	72	500000	1
21	70	1000000	1
27	68	150000	1
27	68	150000	1
50	73	80000	1
59	73	80000	1

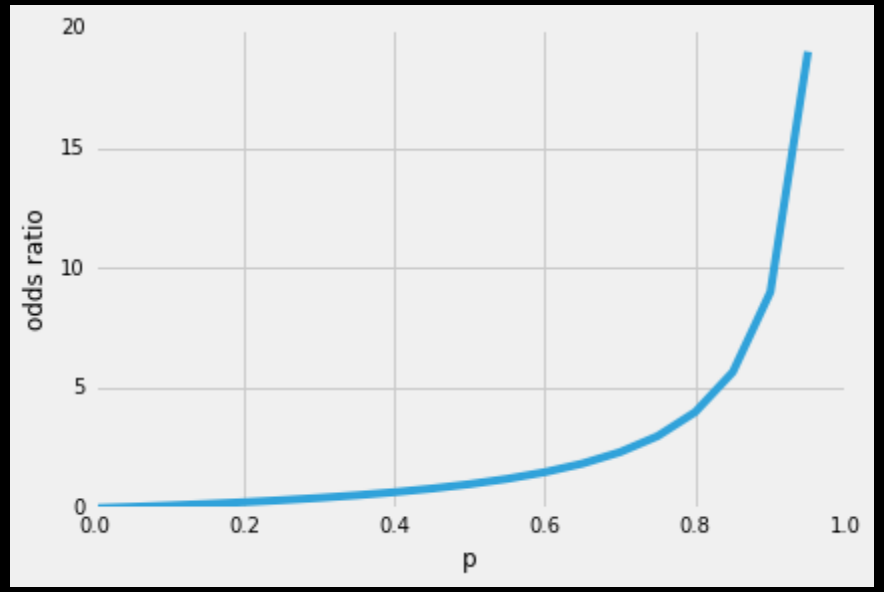
BINOMIAL LOGISTIC REGRESSION

WHY?

- Rule of thumb:

In love, there are only two choices.
When will be our wedding?
OR Will you come to my wedding?

➔ So it's binary in terms of probability of getting a match after 1 month



- And WTH is logistic regression?

Good news is your probability of getting a match can't be -50% (I'm not in hell). Bad news is the probability can't be 200% either (even if I'm Brad Pitt??????)

→ Not linear regression but logistic regression

WE COULD PRETEND THAT WE RAN THE MODEL SUCESSFULLY
(WHICH YOU WILL HAVE CHANCE TO PRACTISE IN AFTERNOON SESSION)

TIME FOR CANDY CRUSH! PUT ALL THE PIECES TOGETHER

USER RETENTION

PREPARE THE DATA WITH LABELS

BINOMIAL LOGISTIC REGRESSION

Did we kill the Snow White?

Results and interpretation

- Log_odd(Match_or_not) = 0.1367*income – 0.016*height + 0.018*age
- However, ony income is significant in Wald test.
- Income is the best predictor for probability of being matched against not being matched.
- We should perform user segmentation by income to perform better matching algorithm.