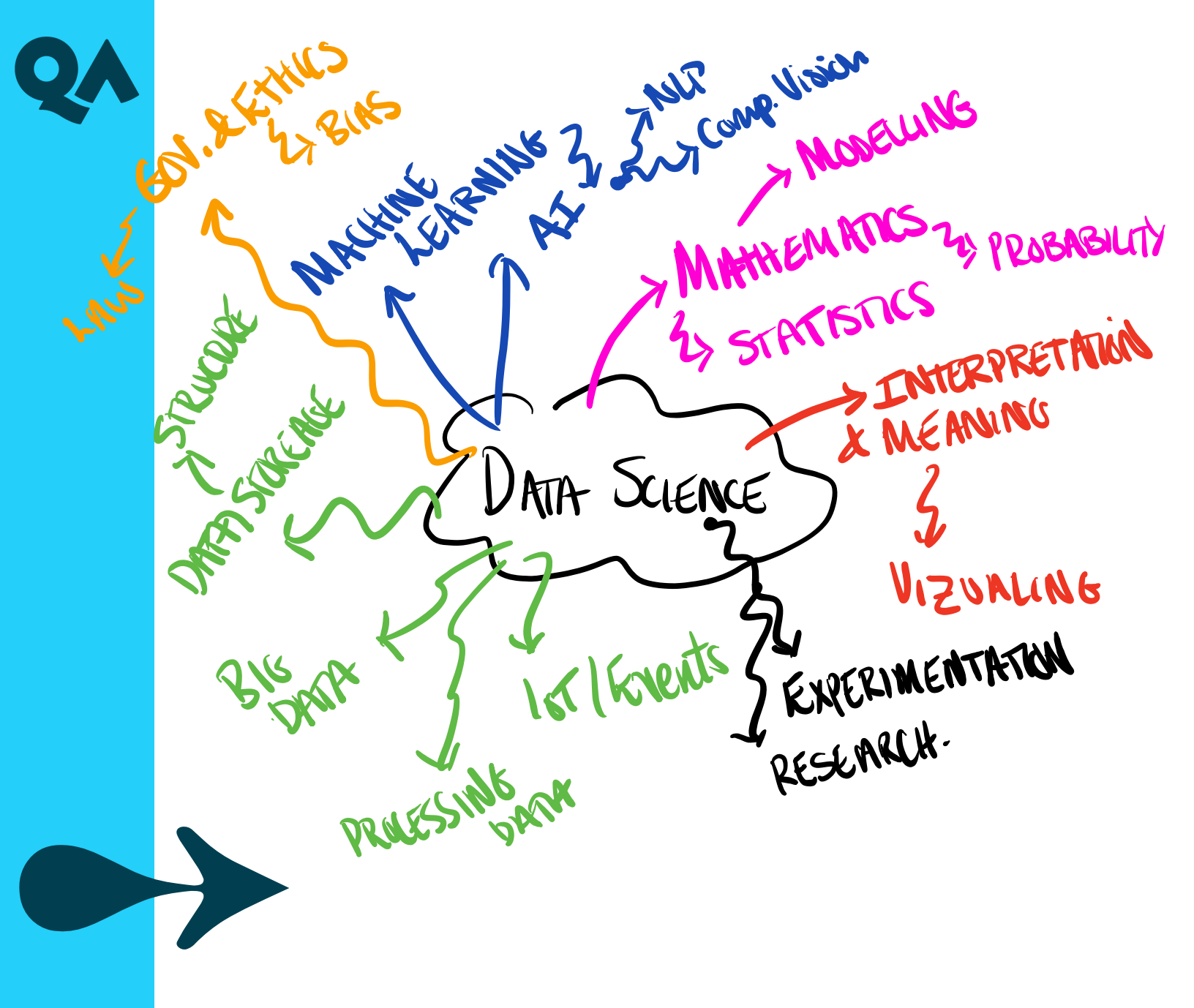
Intro to Data Science


GOV. & ETHICS
BIAS
LAWS
MACHINE LEARNING
AI
NLP
Comp. Vision
MODELLING
MATHEMATICS
PROBABILITY
STATISTICS
STRUCTURE
DATA STORAGE
INTERPRETATION & MEANING
DATA SCIENCE
VIZUALING
BIG DATA
IOT / Events
EXPERIMENTATION
RESEARCH.
PROCESSING DATA

✱ WHY NOW?

vs. 1990, 2005... ?

→ MORE DATA

↓

WHERE?

↓

INTERNET

↳ USER BEHAVIOUR (People!)

"MORE COMPUTE"

~~Algs~~

↓

CPU

✱ WHAT'S DIFFERENT?

↳ vs. DATA ANALYSIS?

→ SCIENCE (EXPERIMENT, EXPLANATION)

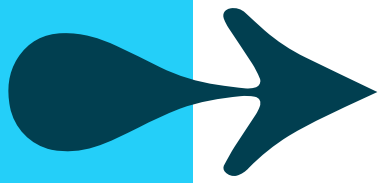
→ DATA ANALYSIS

↳ HISTORICAL ("CERTAIN") QA. ANSLA

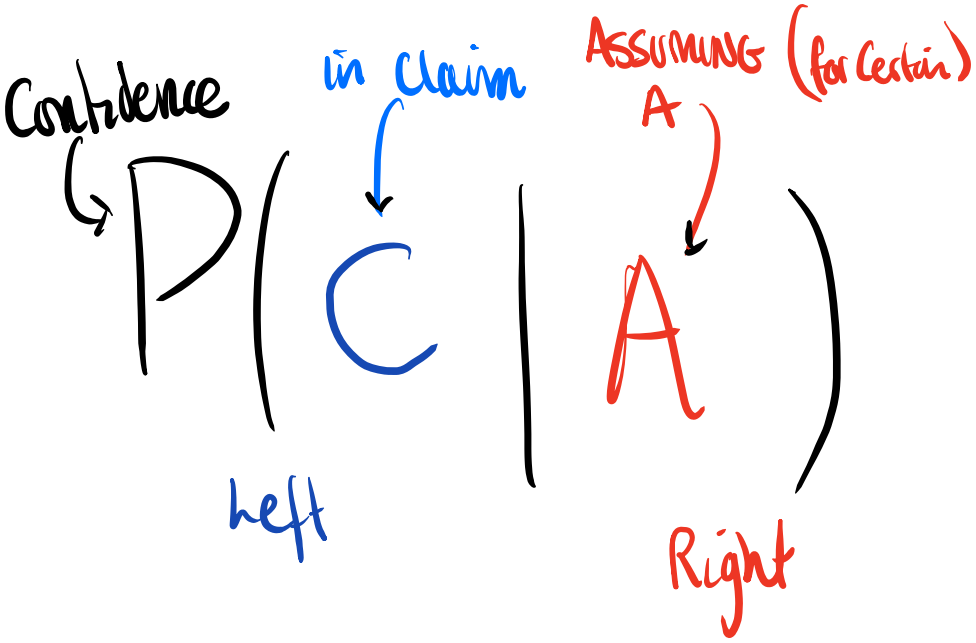
Eg. mean(profit$_{1990}$) = £1m.

ALL DATA SCIENCE QUESTIONS ARE PROBABALISTIC

Eg. What profit is most likely M 2030 - AND - How confident are we in that profit?



PROBABILITY



P("it will rain tomorrow" | "it has rained today")

VARIATIONS IN NOTATION

$P(Y|X)$

Future

Y - "amount of rain"

X - "amount of rain" PAST

A Claim about Y

Assumption about X

$P(Y)$ ← RHS is missing

"ASSUMING IMPLIED BACKGROUND 'knowledge'"

Eg.

$P(H)$ ← doesn't make sense...

$P(H \mid \{H, T\}) = \frac{1}{2}$

CLEARER

Eg.

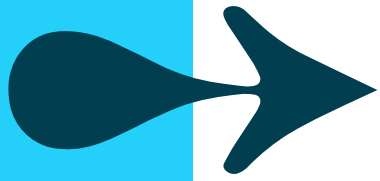
$$P(\text{Fraud} \mid \text{Age}, \text{Amount})\Big|_{0}^{1}$$

Question ← → Assumptions

Data Science Examples

Industries & datasets

- Healthcare - screening
- Finance - stocks
- Utilities - power supply
- Insurance - fraud
- Retail - products



RETAIL – PRODUCTS

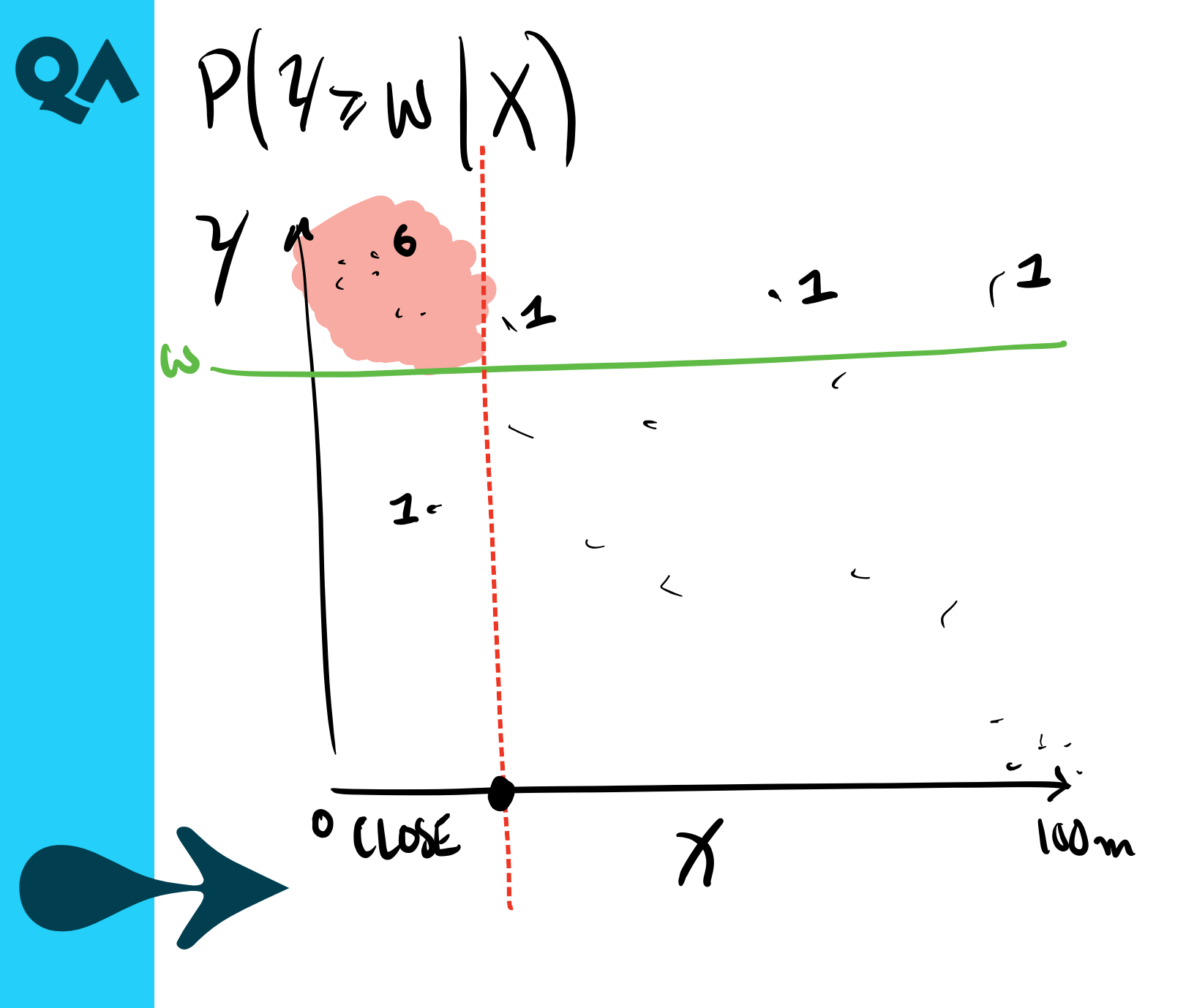
↳ "HOW WELL WILL A PRODUCT SELL?" ← QUANTITATIVE?

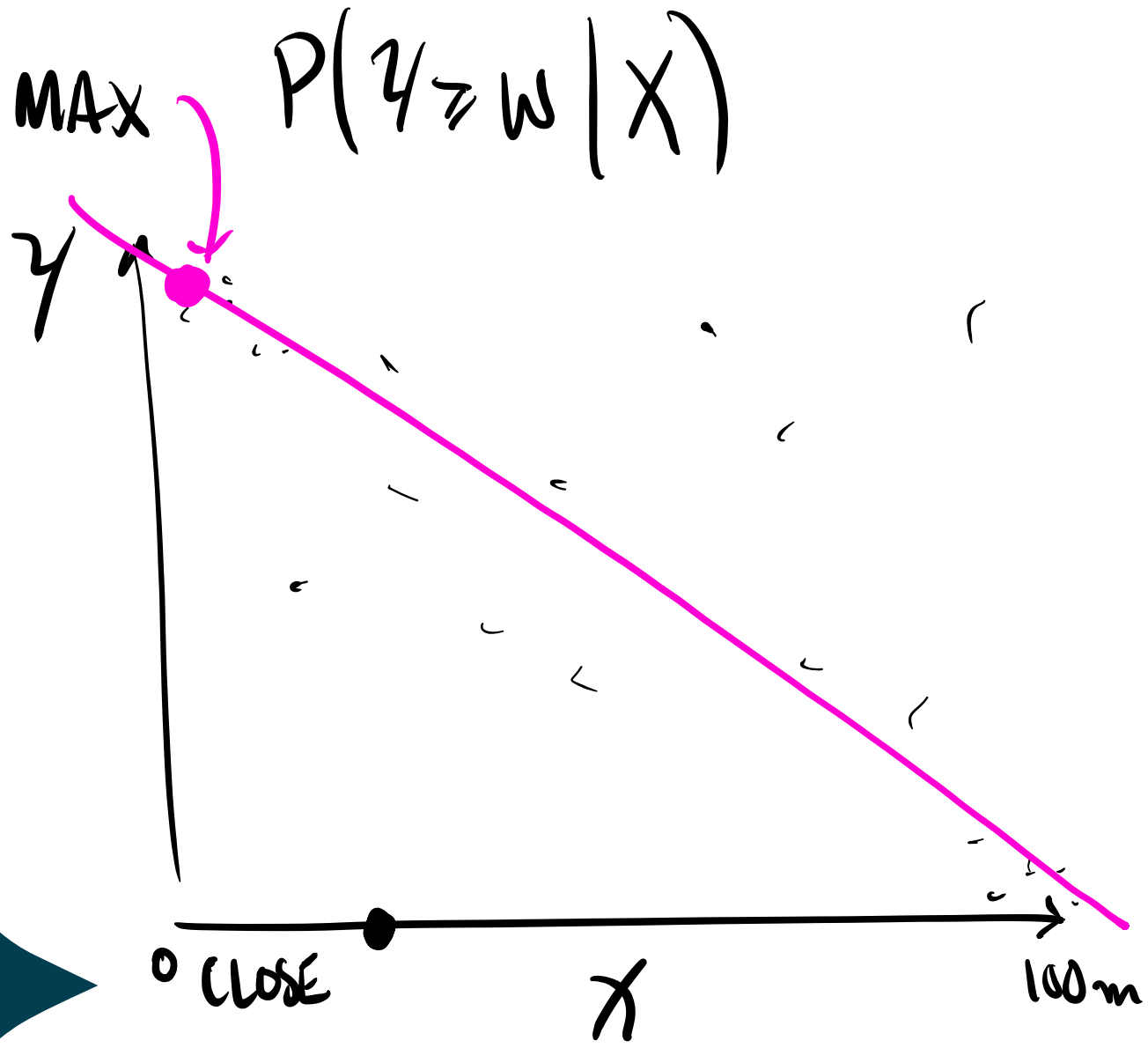
↳ Y (question)

BASED ON "LOCATION IN STORE" X (assumption)

WELL def. No. Units/Month $\geq W$ ← cutoff

$$P(Y \geq W \mid X)$$


P(Y≥w|X)
Y
6
1
1
1
w
1
0 CLOSE
X
100m


MAX
$P(Y \geq w \mid X)$
Y
0 CLOSE
X
100m

Aside: Eg. Probability Rule

"Emphasising the significant connections"

$$P(C|E) = \frac{P(E|C)\,P(C)}{P(E)}$$

↓ "MORE ACCURATE"

$$P(C|E,B) = \frac{P(E|C,B)\,P(C|B)}{P(E|B)}$$