NAÏVE BAYESIAN CLASSIFICATION

LBG INCUBATION PREPARATION

Yobi Livingstone

Contents

Intro to Bayes Theorem

Intro to Bayes' Theorem

→ **Imagine you overhear a conversation:**

→ **Steve's great but he's often shy and withdrawn, he likes to talk about books and not much else.**

→ **Now, which of these two statements do you find to be more likely?**

→ **1) Steve is a Programmer.**   **2) Steve is a Librarian**

Intro to Bayes' Theorem

→ **Imagine you overhear a conversation:**

→ **Steve's great but he's often shy and withdrawn, he likes to talk about books and not much else.**

→ **Now, which of these two statements do you find to be more likely?**

→ **1) Steve is a Programmer.**   **2) Steve is a Librarian**

→ **Many might say Steve's more likely to be a Librarian because of the evidence given by 'likes to talk about books', however this would be an irrationality.**

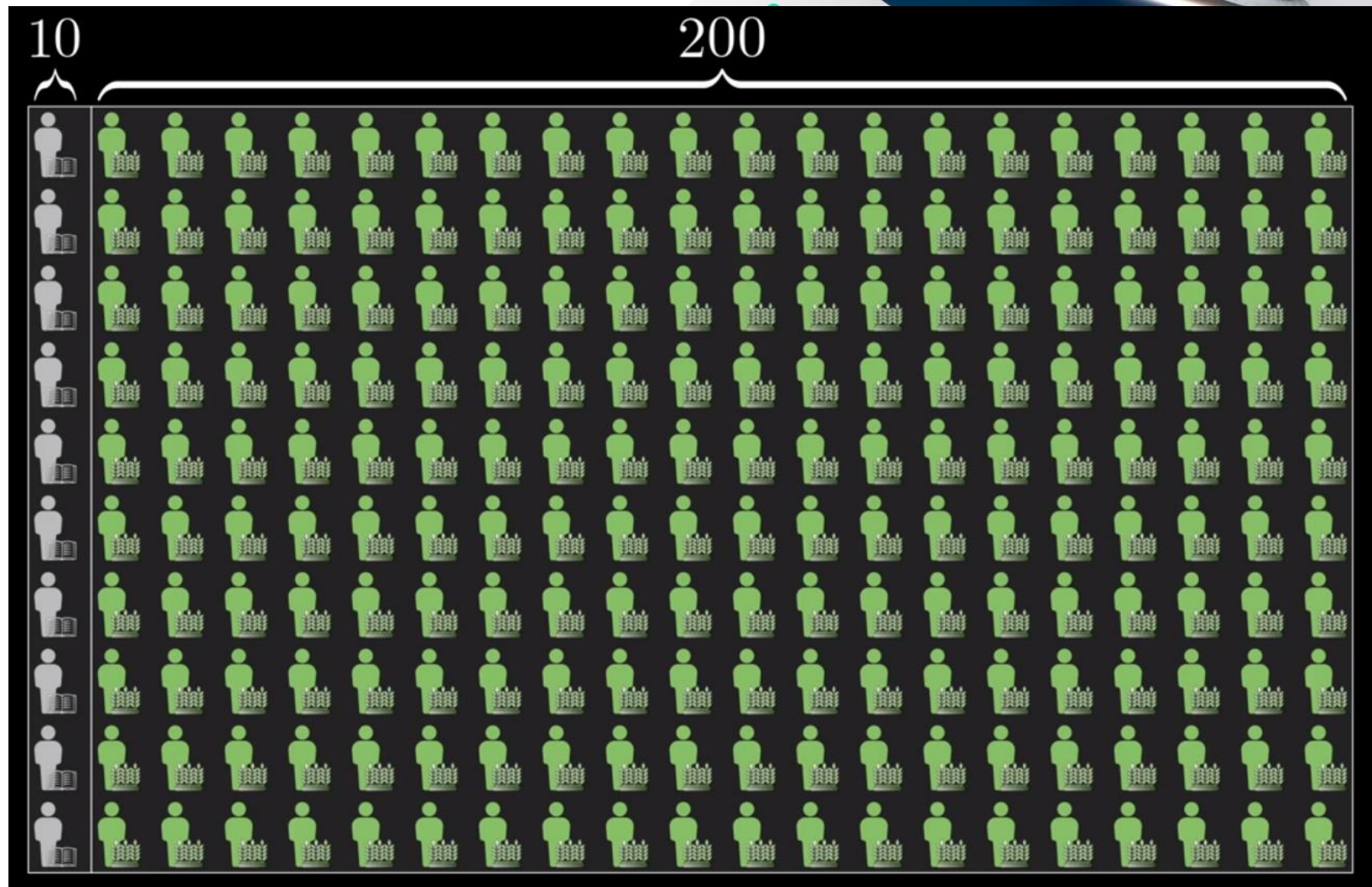→ **Why is that the case? – Discuss with one of your colleagues**

Bayes Theorem

There are approx. 338,000 programmers in the UK
Source Statista

There are approx. 16,000 librarians in the UK
Source Readingagency

Therefore, there's **roughly 20 times more programmers than librarians.** Rationality is about realising which facts are relevant, given new evidence.
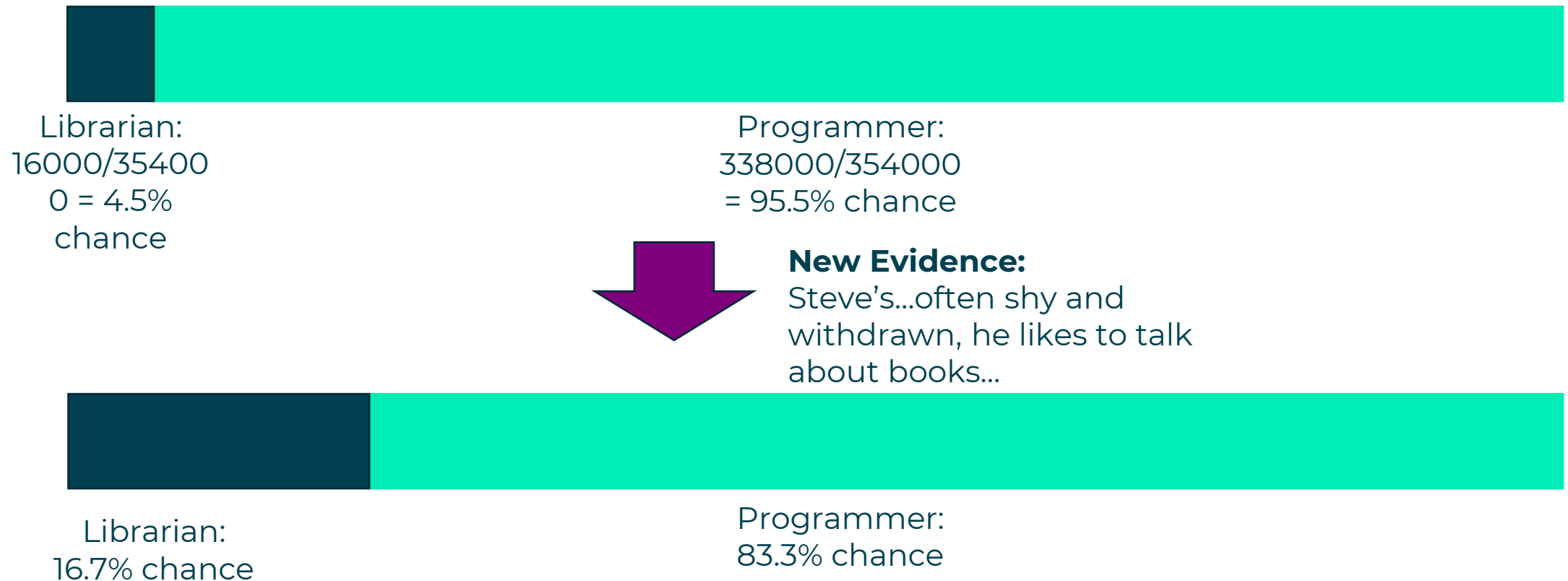
Bayesian Hypothesis testing

Belief that 40% of librarians like to talk about books, not much else

Belief that 10% of programmers like to talk about books, not much else

Bayesian Hypothesis testing

→ **The Bayesian mantra is, that new evidence should not completely determine your beliefs in a vacuum, new evidence should update your existing beliefs.**

→ **Total population of librarians and programmers : 338,000 + 16,000 = 354,000**

Librarian:
16000/35400
0 = 4.5%
chance

Programmer:
338000/354000
= 95.5% chance

New Evidence:
Steve's...often shy and withdrawn, he likes to talk about books...

Librarian:
16.7% chance

Programmer:
83.3% chance

Rationality

- **The experiment is <u>NOT</u> about stereotyping people with more accuracy (i.e. librarians are 2.5x more likely to be *bookish introverts* than programmers)**

- **Its about paying attention to and adjusting the original likelihood with new evidence.** **Even if people may not know the original likelihood, most rarely make the effort to think about including it at all!**

- **Nobel Prize winners**: **Daniel Kahneman and Amos Tyersky** –popularised this irrational thinking concept where we often make judgements based on new evidence without including the original premise (aka Base Rate Fallacy).

- Written about in books like Thinking Fast and Slow.

THINKING,
FAST AND SLOW

DANIEL
KAHNEMAN

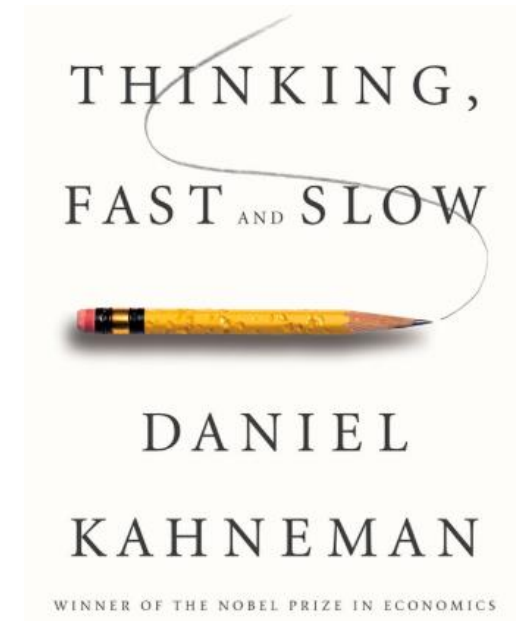WINNER OF THE NOBEL PRIZE IN ECONOMICS

Bayes' mantra

→ **The Bayesian mantra is, that new evidence should not completely determine your beliefs in a vacuum, new evidence should update your existing beliefs.**

When to use Bayes' rule:

Steve's great but he's often shy and withdrawn, he likes to talk about books and not much else.

Bayes' Theorem: **The Maths**

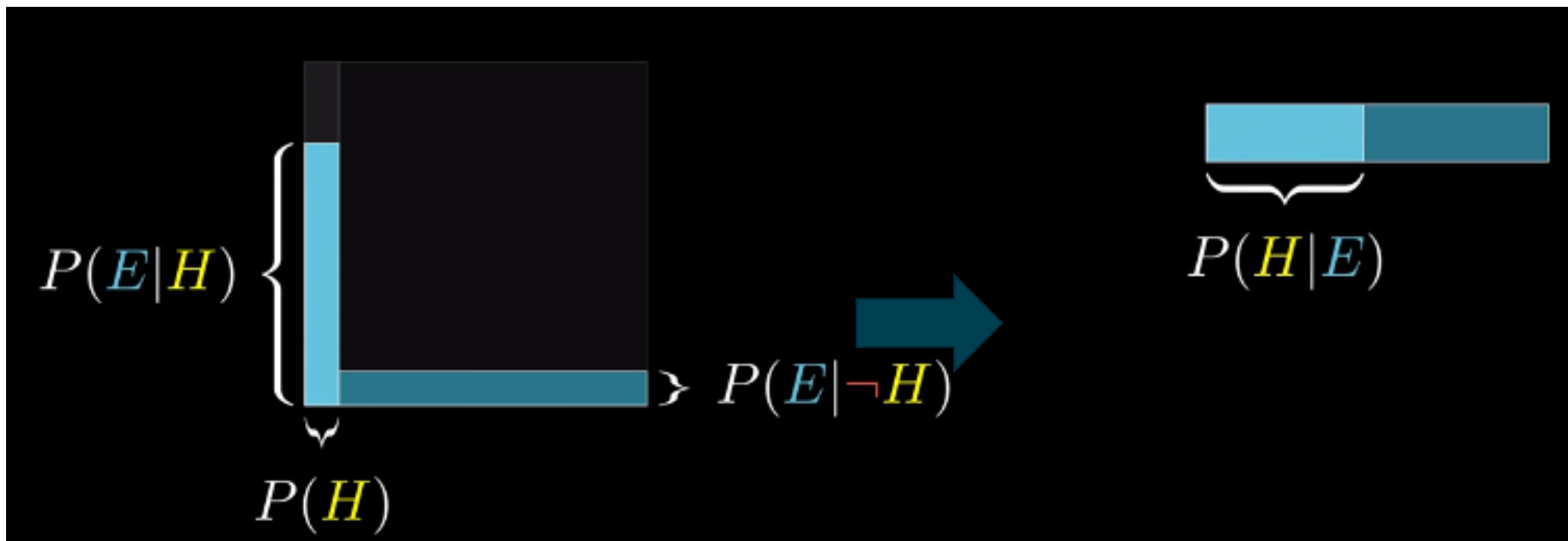$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)}$$

$P(E|H)$

$P(E|\neg H)$
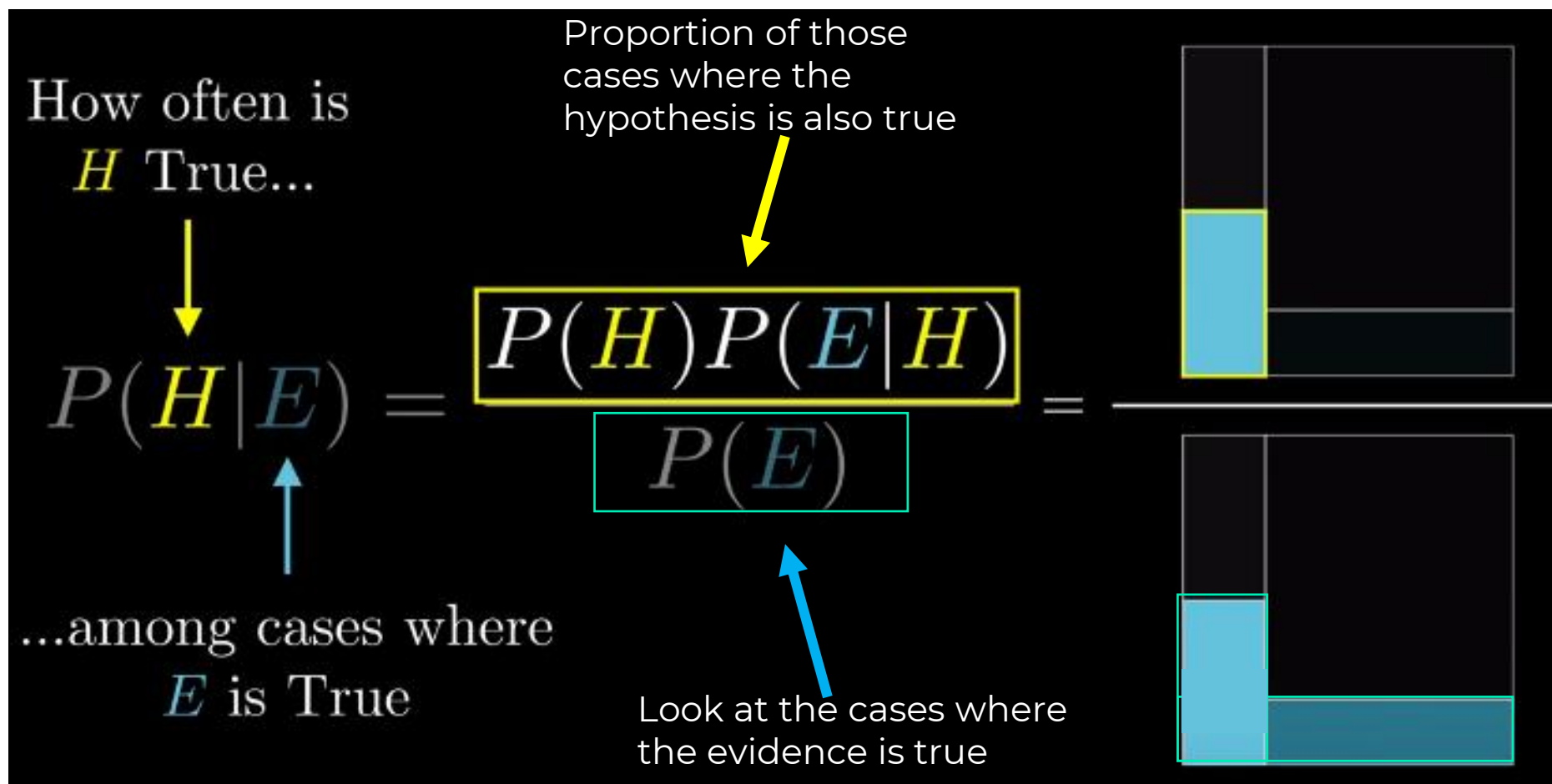
$P(H)$

$P(H|E)$

Bayes' Theorem: **The Maths**

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

$P(E|H)$

$P(E|\neg H)$
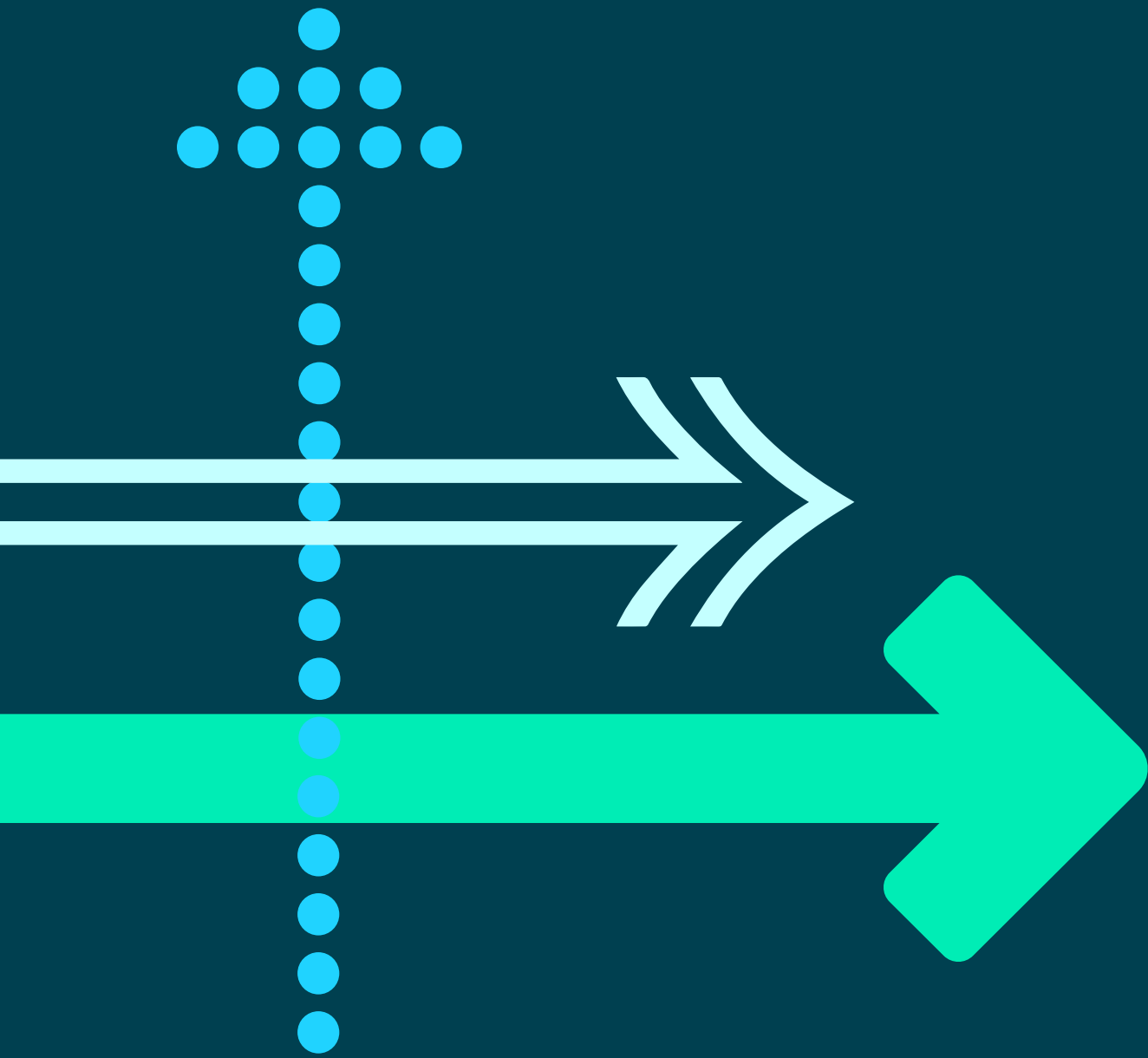
$P(H)$

$P(H|E)$
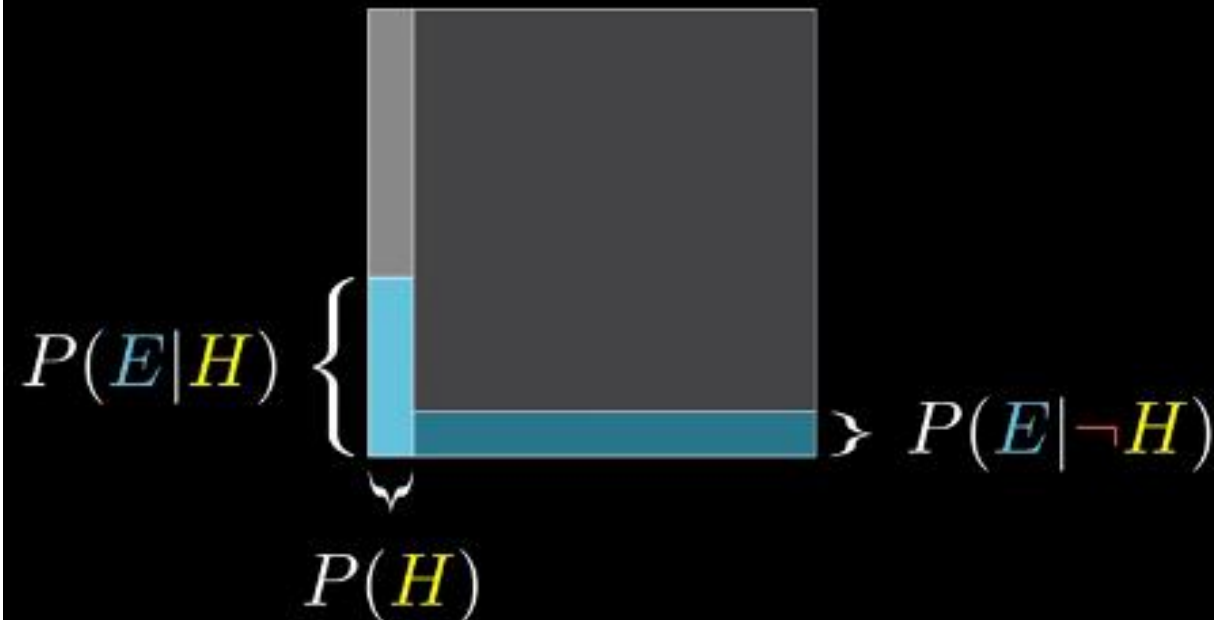
Bayes' Theorem: **The Maths**

Testing your hypothesis
with Bayes' Theorem

Using Bayes' Theorem for hypothesis testing

→ **The theorem converts the results from your test into the real probability of the event.**

- **What is the likelihood that a customer will buy this cereal when exposed to this advertising campaign?**

- **What is the likelihood that a vegan would vote for the Green Party?**

- **Given that this email came from outside the company and has words "Free", "transfer money", "gift vouchers", whats the probability that its Spam?**

→ **Using Bayes Theorem, you can find probability of H (hypothesis)being True, given E (evidence): Pr(H|E). By starting with Pr(E|H), the chance the evidence appears, when the hypothesis is True**

$$\Pr(H|E) = \frac{\Pr(E|H)\,\Pr(H)}{\Pr(E|H)\,\Pr(H) + \Pr(E|\text{not } H)\,\Pr(\text{not } H)}$$

Using Bayes' Theorem for hypothesis testing

→ **Imagine the following scenario, Michelle is 47 years old and is due to get screened for breast cancer with mammography.**

- **Lets assume that 1% of women submitted to screening at 47years old have breast cancer (and therefore 99% do not).**

- **87% of mammograms detect breast cancer when it is there (and therefore 13% miss it). – source: Breast Cancer Surveillance Consortium (BCSC)**

- **9.6% of mammograms detect breast cancer when it's not there (and therefore 90.4% correctly return a negative result). – see above**

→ **Organised the probabilities in a table:**

Mammogram test results	Likelihood of breast-cancer in a 47yr old woman	
	Cancer (1%)	No Cancer (99%)
Test Positive	87%	9.6%
Test Negative	13%	90.4%

Anatomy of a Hypothesis test

→ **The theorem acknowledges:**

- **Tests are not the event: a cancer test is separate from the event of having cancer. A test for spam is separate from the message actually being spam.**

- **Tests are flawed: False positives (positive on test, but no cancer) and False Negatives(negative on test, but has cancer). Its important to adjust test's results for test errors**

- **False positives skew results: suppose you are searching for something very rare (1 in a million) Even with a good test of 99% accuracy, you will identify numerous false positives.**

QA

Anatomy of a test

→ **Michelle gets a positive result indicating breast cancer, what is the probability that a positive result leads to Michelle having breast cancer. 87%? 90.4%? 1%?**

Mammogram test results	Likelihood of breast-cancer in a 47yr old woman	
	Cancer (1%)	No Cancer (99%)
Test Positive	87%	9.6%
Test Negative	13%	90.4%

→ **A positive test result means we should be concerned with the top of the table.**

Mammogram test results	Likelihood of breast-cancer in a 47yr old woman	
	Cancer (1%)	No Cancer (99%)
Test Positive	True Pos. 1% x 87% = 0.0087	False Pos. 99% x 9.6% = 0.095
Test Negative	False Neg. 1% x 13% = 0.0013	True Neg. 99% x 90.4% = 0.895

QA

Calculating the actual probability

$$Pr(H|E) = \frac{Pr(E|H)\,Pr(H)}{Pr(E|H)\,Pr(H) + Pr(E|\text{not } H)\,Pr(\text{not } H)}$$

→ **The chance of an event is the number of ways it could happen given all possible outcomes:**

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)} = \frac{\blacksquare}{\blacksquare + \rule{2em}{0.8em}} = \frac{\text{True Pos. 1\% x 87\% = 0.0087}}{\text{True Pos. 1\% x 87\% = 0.0087} + \text{False Pos. 99\% x 9.6\% = 0.095}}$$

→ **Pr(Having cancer from positive result)** = 0.0087 / (0.0087 + 0.095) = 8.3%

→ **Pr(NOT having cancer from positive result)** = 0.095 / (0.0087 + 0.095) = 91.7%

→ **This may seem low, but consider the base rate fallacy! The test gives false positives 9.6% of the time, applied over a population where 99% don't have cancer. This would lead to many false positives in a given population. For a rare disease, most positive test results would be wrong!**

→ **Experiment with Variations on the calculation: Bayes Theorem | InstaCalc Online Calculator**

QA

Context for Clinical tests

1) **87% mammography screening accuracy goes up with older age groups(source).**

2) **Tests are used in conjunction with the display of other symptoms. Hard lump, redness, skin thickness etc.**

3) **Risk factors such as age, family history, diabetes/obesity play a part in indicating the viability of a test result.**

QA
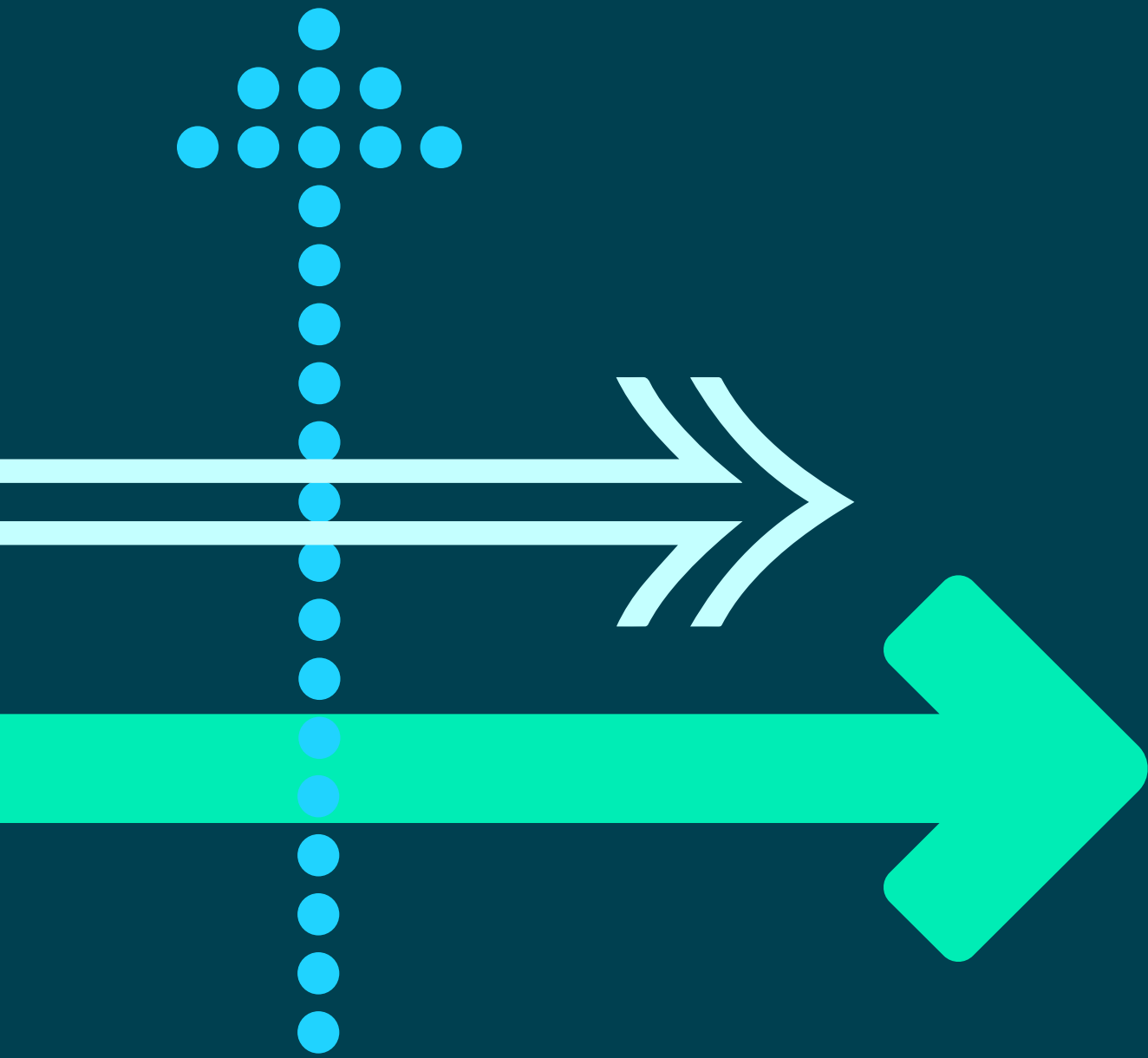
NBC : Spam Filters

Naïve Bayesian Classification

Naïve Bayes Classification remains one of popular methods to solve text categorisation problem, the problem of judging documents as belonging to one category or the other, such as email spam detection.

Naïve Bayes classifiers are a family of simple probabilistic classifiers, based on applying Bayes' theorem with strong(Naive) independence assumptions between features. i.e. the word 'invest' is a feature assumed to be evident in much of spam. This feature would increase the likelihood of a spam classification. When enough features have indicated a higher likelihood of spam classification, the message would be filtered.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.

NBC: Features and Labels

In Bayesian classification, we're interested in finding the probability of a label given some observed features, which we can write as Pr(Label | Features) . i.e. probability that this label: Spam, can be given when the following features have been observed "massive", "double", "£££", "buy". Each feature, continues to adjust the probability

Typical phrases (features) of a Spam email (label):

- As seen on. Buy. Buy direct. ...
- Dig up dirt on friends. Meet singles. Score with babes.
- Additional Income. Be your own boss. Compete
- $$$ Affordable. Bargain. ...
- Accept Credit Cards. Cards accepted. ...
- Avoid bankruptcy. Calling creditors. ...
- Acceptance. Accordingly. ...
- Dear [email/friend/somebody] Friend.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

$$P(L \mid \text{features}) = \frac{P(\text{features} \mid L)P(L)}{P(\text{features})}$$

NBC: Probability of each Label

→ **What's the probability that this message is SPAM? Given these informative features? The classify, predicts the probability of SPAM and Legitimate email, classifies the message with whichever label is most probable.**
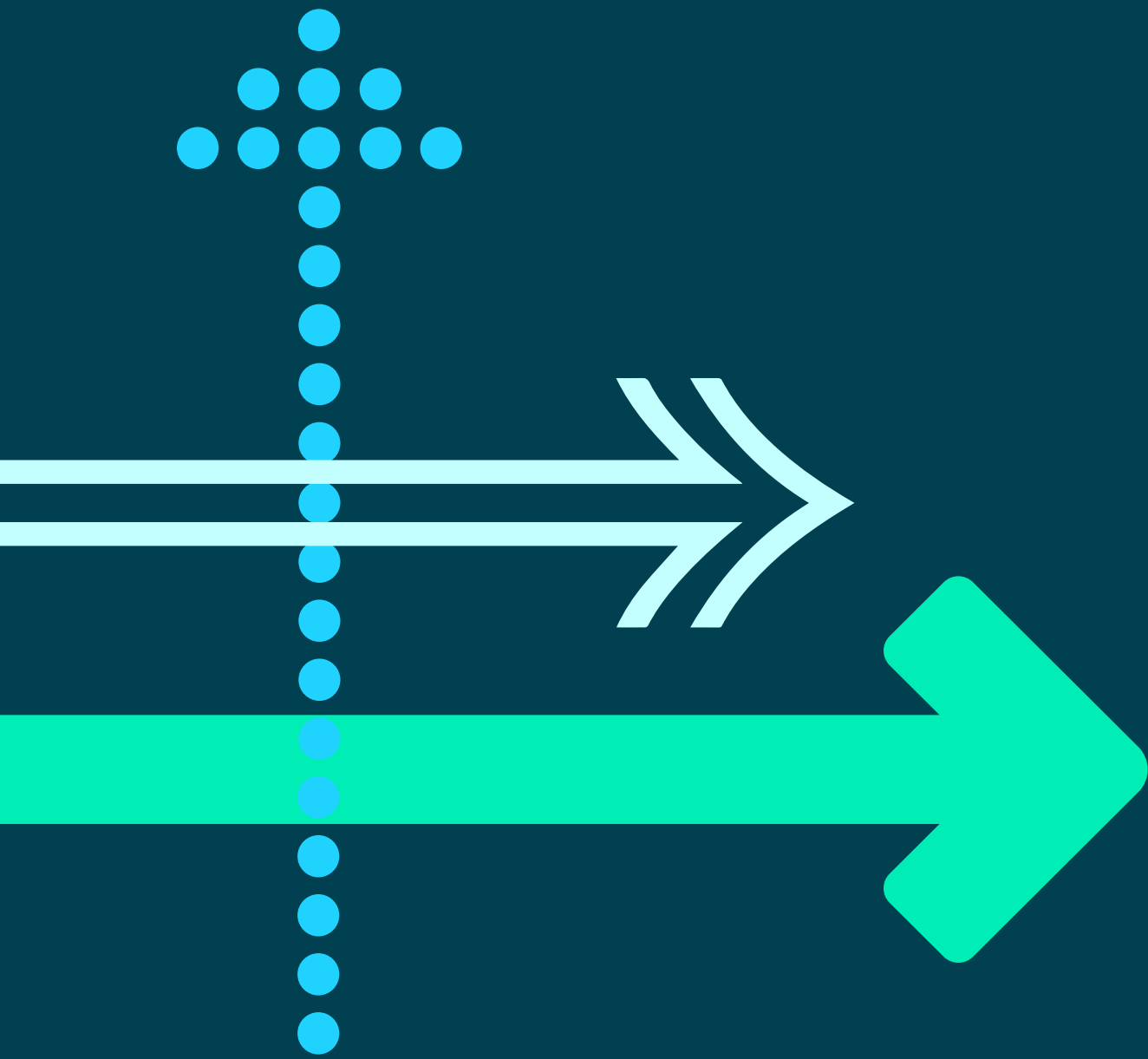
Dear Jeremy L,

Thanks for signing up to our investors centre, a select group of like-minded and market savvy individuals with a financial growth mindset. We are still awaiting your deposit of £200 to get in on this 10% growth opportunity to invest Aerodyne shares.

Kind Regards, Jordan.

$$P(L \mid \text{features}) = \frac{P(\text{features} \mid L)P(L)}{P(\text{features})}$$

P (*SPAM* | "invest", unknown email, "your deposit", "£") **=** $\dfrac{\textbf{P} (\text{"Invest", unknown email, "your deposit", "£"} \mid SPAM) * \textbf{P} (SPAM)}{\textbf{P} (\text{"Invest", unknown email, "your deposit"})}$
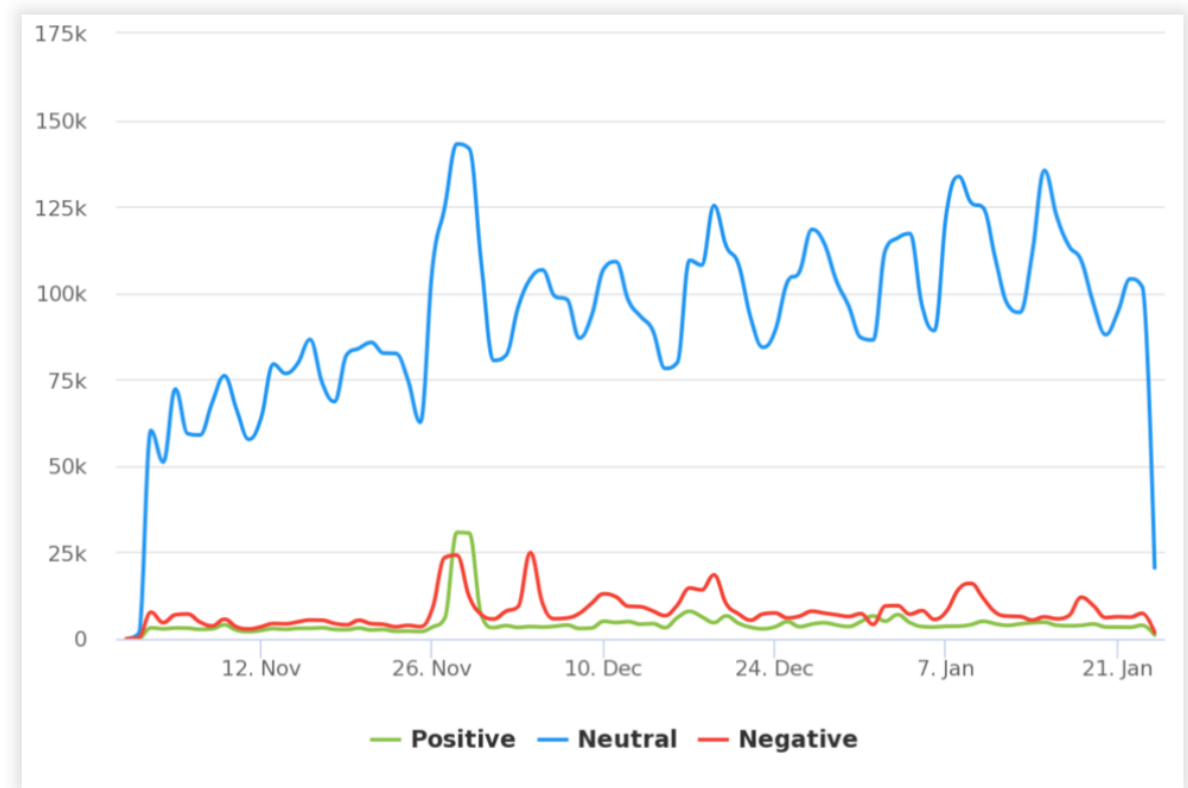
NBC : Sentiment Analysis

NBC: Sentiment Analysis

→ **This is also highly applicable for sentiment analysis that takes a long strings like an email, a review, a social media post, a news article, a doctors report etc ect.**

→ **And provides metrics such as a sentiment score, a polarisation score and a subjectivity score.**

→ **It allows us to analyse qualitative data effectively by identifying key words associated with these attitudes. For example. More use of the word "I think", "I believe" is associated with opinion articles that are high in subjectivity. This can be used to also tag topics**

Central Bank interest rates
- Twitter mention Volume & Sentiment

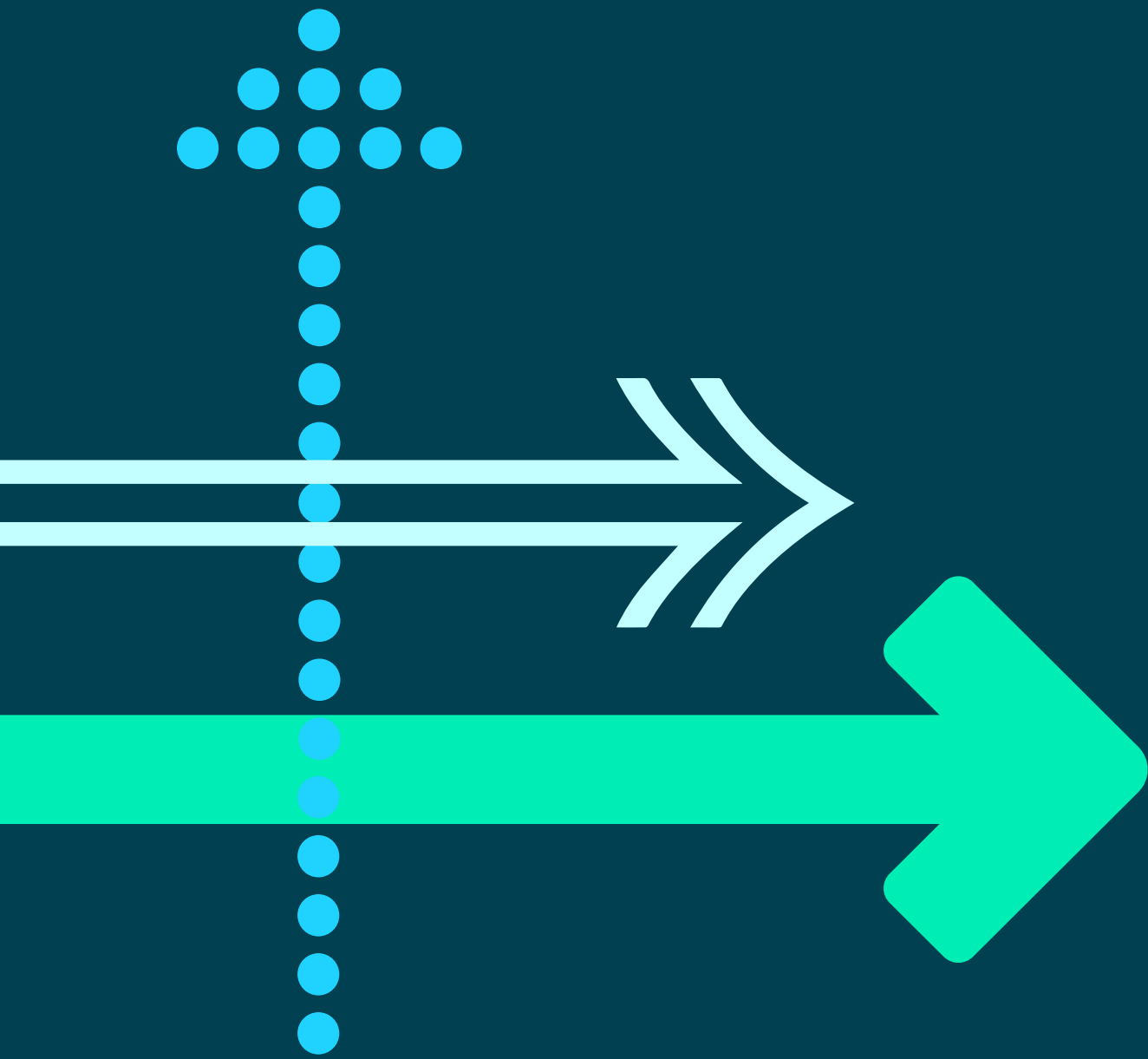# Sentiment Analysis Methods

Lexicon-based Approach:

• Uses a sentiment dictionary that associates words with sentiment scores. Explains how scores are aggregated to determine overall sentiment.

Examples: Vader (trained on Social Media posts), TextBlob(Customer Reviews etc.)

Machine Learning Approach:

• Uses supervised learning models (e.g., **Naive Bayes**, SVM, deep learning) trained on labelled datasets to predict sentiment.

• The advantage is training it on your own Qualitative Data : Financial, Banking etc

QA

data preparation and
feature identification

Data Preparation – creating a manageable training dataset

1. **Tokenization**: This is the process of breaking down text into smaller units called tokens, which can be words, phrases, symbols, or other meaningful elements. Tokenization helps in simplifying text analysis by reducing it to manageable pieces. For instance, a sentence is tokenized into individual words.

Input:	"I received this product in a broken state, I want a refund"
Output:	'I' , 'received' , 'this' , 'product', 'in', 'a', 'broken', 'state', 'I', 'want', 'a', 'refund'

2. **Lemmatization**: Lemmatization involves reducing words to their base or root form. Unlike stemming, which crudely chops off word endings, lemmatization considers the context and uses the lexical knowledge of words to transform them into their lemma or dictionary form. For example, "running", "ran", and "runs" would all be lemmatized to "run".

Input:	Broke, Break, Broken
Output:	Break

QA

Data Preparation – creating a manageable training dataset

3. Stop Words Removal: Stop words are common words like "and", "the", "is", etc., that are usually removed from the text as they offer little value in understanding the meaning of the text for analysis.

Input:	"I received this iPhone in a broken state, I want a refund"
Output:	'I' , 'received' , ~~'this'~~ , 'iPhone', ~~'in'~~, ~~'a'~~, 'broken', 'state', 'I', 'want', ~~'a'~~, 'refund'

4. Normalization: This includes converting all text to the same case (upper or lower), removing white spaces, and correcting typos or spelling variations to ensure consistency in the text data.

Input:	"I received this product in a broken state, I want a refund"
Output:	'i' , 'received', 'iphone', 'broken', 'state', 'i', 'want', 'refund'

Data Preparation – creating a manageable training dataset

5. **n-grams Creation**: n-grams are continuous sequences of n items from a given sample of text or speech. Creating n-grams involves combining adjacent tokens to capture more context than individual words alone, which can be useful for text classification or language modelling.

2-gram moving window	'i' , 'received', 'iphone', 'broken', 'state', 'i', 'want', 'refund'
	'i' , 'received', 'iphone', 'broken', 'state', 'i', 'want', 'refund'
	'i' , 'received', 'iphone', 'broken', 'state', 'i', 'want', 'refund'
	'i' , 'received', 'iphone', 'broken', 'state', 'i', 'want', 'refund'

→ **- Computationally demanding as you expand the window.**

→ **- Essential for text contextualise by preceding word:**

→ **"not bad", "not great", "cried with joy", "want a refund", "got a refund"**

Feature Selection for modelling

TF-IDF – Term Frequency Inverse Document Frequency -

statistical measure used to evaluate the importance of a word in a document relative to a collection of documents or corpus.
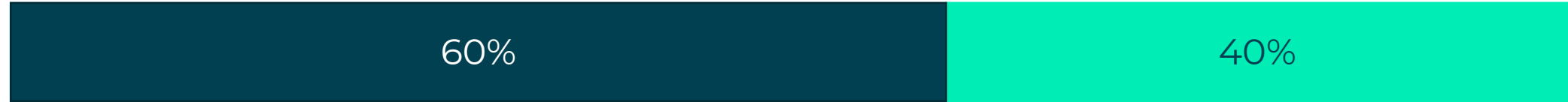
5-Star Positive Reviews	1-Star Negative Reviews	Informative Feature for Sentiment?
Broken (0.2% of reviews)	Broken (12% of reviews)	Very informative towards negative
Received (52%)	Received (57%)	No information
Battery (24%)	Battery (48%)	Some information, nudges the probability towards negative

→ Informative features **worth reviewing before fitting (indicates likely informative features for model)** <u>and</u> **after fitting your model (confirm that these features were not overlooked).**

NBC: Training the model

Without any information, we see that there are more positive reviews than negative. So here is the initial probability.

60%	40%

Positive:

Negative:

Some standardise the reviews so its equal amounts of positive and negative reviews. Depends on your goal for the model

New Social Media post:
"I received this iPhone in a broken state, I want a refund" #amazon #apple

15%	85%

Your NBC Sentiment model will determine the percentage likelihood of either class and then classify it as negative

QA

Summary

By connecting to a social media or product review API, you can track sentiment in real-time or simply evaluate the language and historical sentiment regarding services.

These can provide meaningful insights for opportunities and decision makers respond with much more agility to customers.

This is one of the ways in which is business becomes more data-driven and responsive to their customers needs as a result.

Its not simply limited to positive or negative. Recall polarisation and subjectivity with it too.
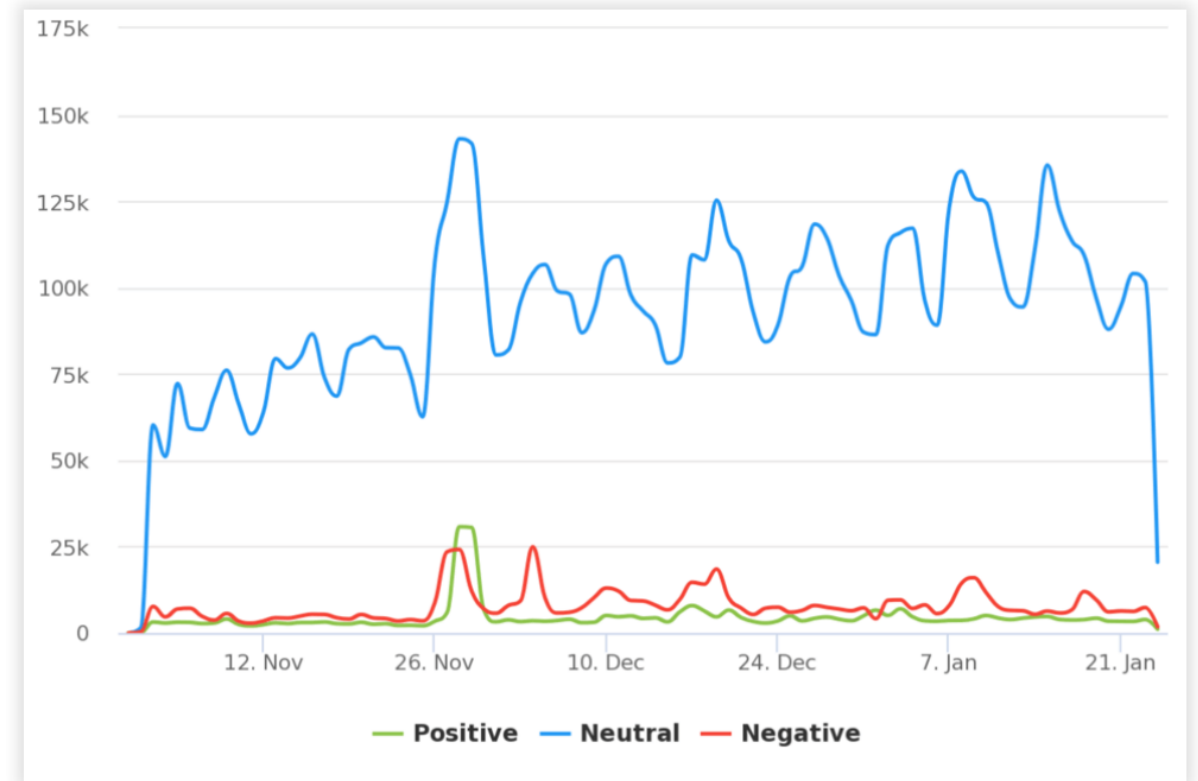
Open and run through the
Jupyter Notebook for
Amazon Food reviews as a
practical Demo of these
concepts

```
#import libraries

import pandas as pd

filepath = r'C:\Users\Yobi-Work\Documents\LBG\Sprint 2\NLP\Amazon Food Reviews\Ama

df = pd.read_csv(filepath)
df
```
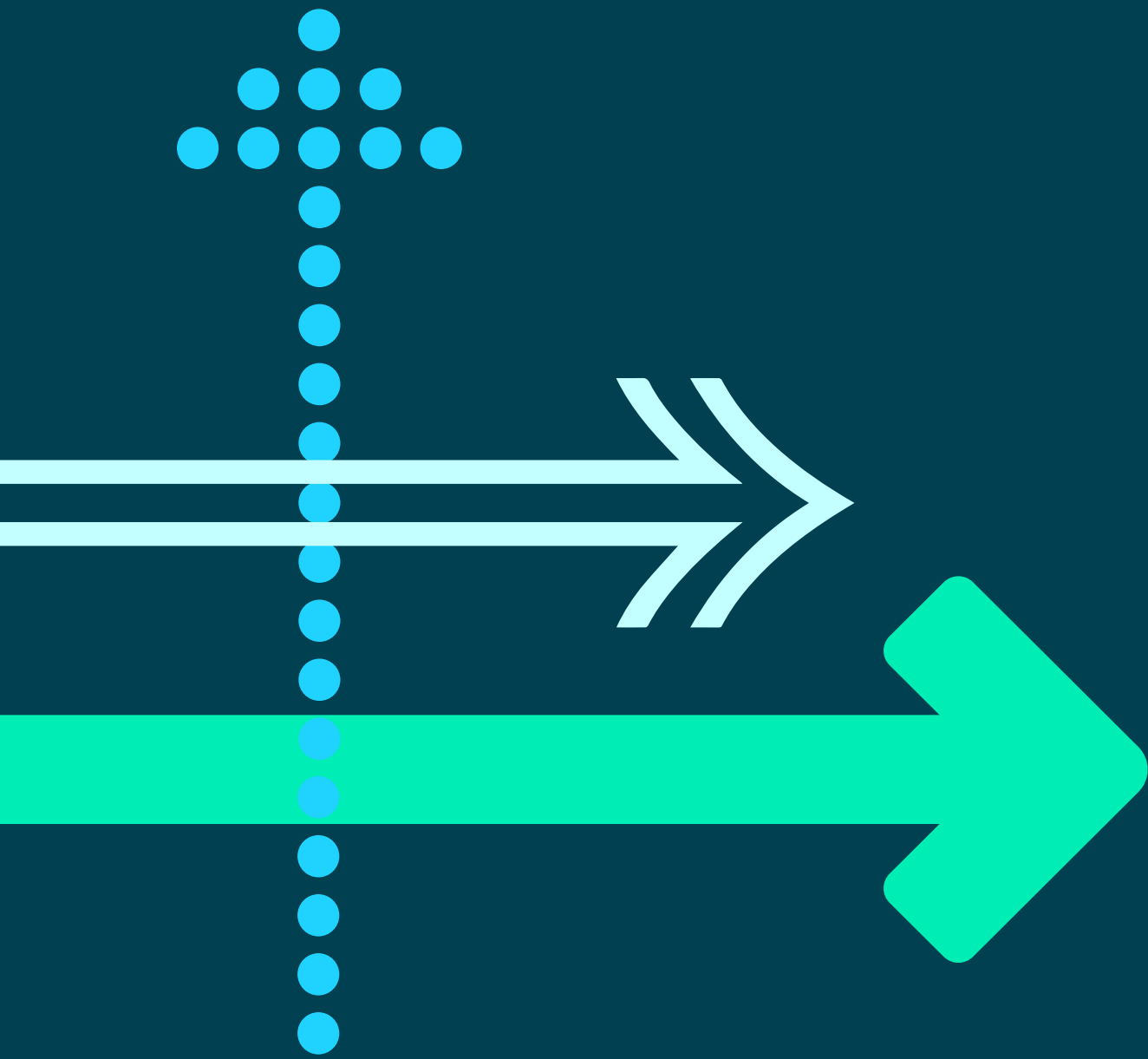
	Score	Summary	Text
0	5	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	5	Great taffy	Great taffy at a great price. There was a wid...
3	5	Great! Just as good as the expensive brands!	This saltwater taffy had great flavors and was...
4	5	Wonderful, tasty taffy	This taffy is so good. It is very soft and ch...
...	...	...	...
30061	1	Not what I expected.	I certainly could do better by these if I just...
30062	5	Great!!	Table is 35 inches in height. It is heavy dut...
30063	5	Black tea v Coca Cola	I have been one of those people that always fe...
30064	1	Never Received the product	I was charged for this product by Liquid Natio...
30065	5	Great taste - and gluten free	Everyone in my family loves this chip - from t...

30066 rows × 3 columns

THANK YOU

for preparing for your forthcoming sprint!