WHAT HAS DATA SCIENCE GOT TO DO WITH IT?

Thordis Thorsteins

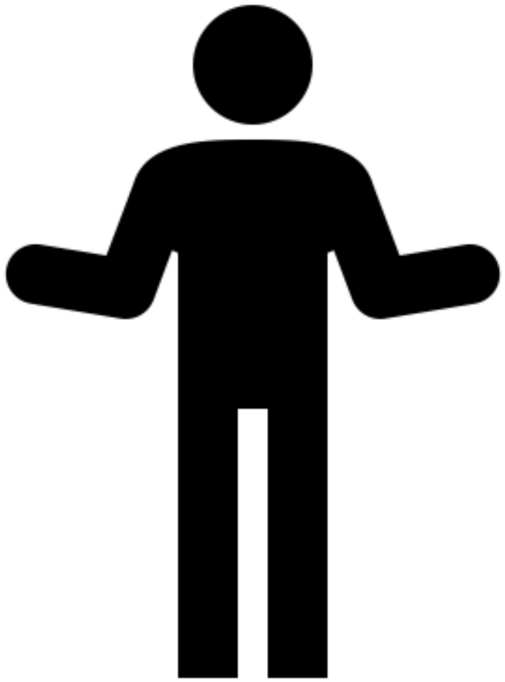
Data Scientist @ Panaseer

@thordisstella

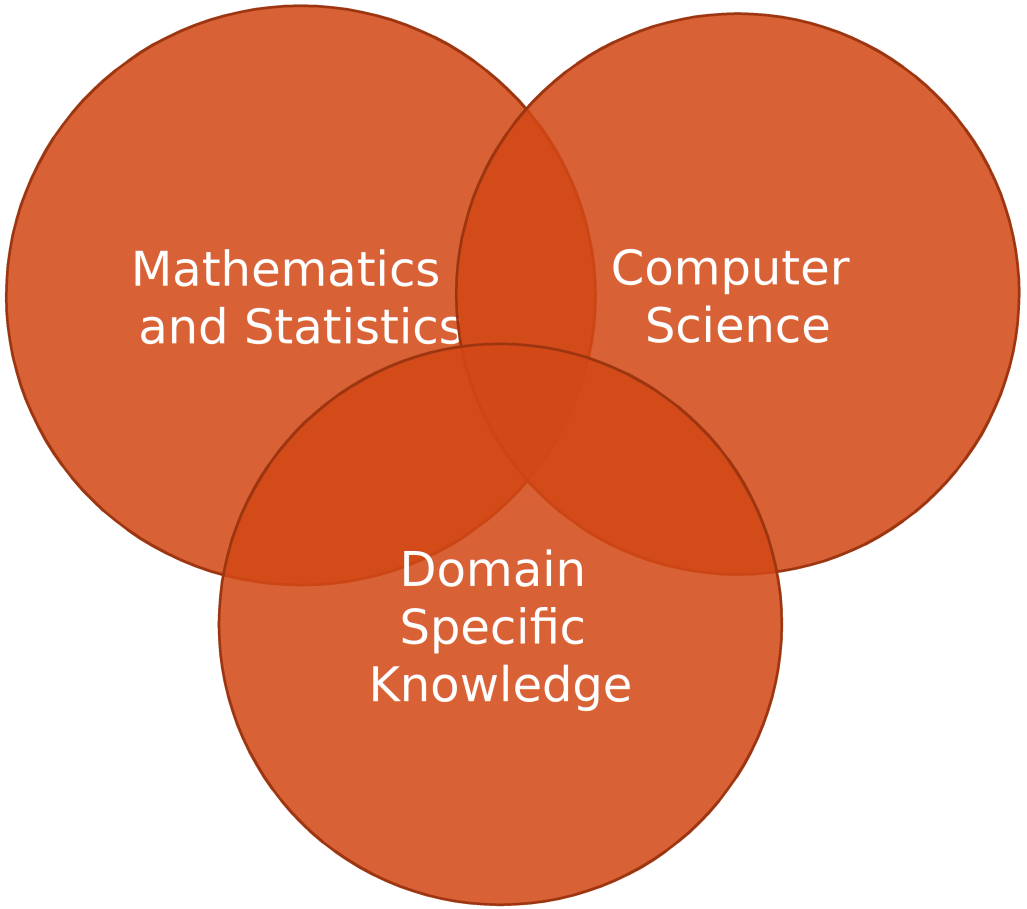
THE PLAN

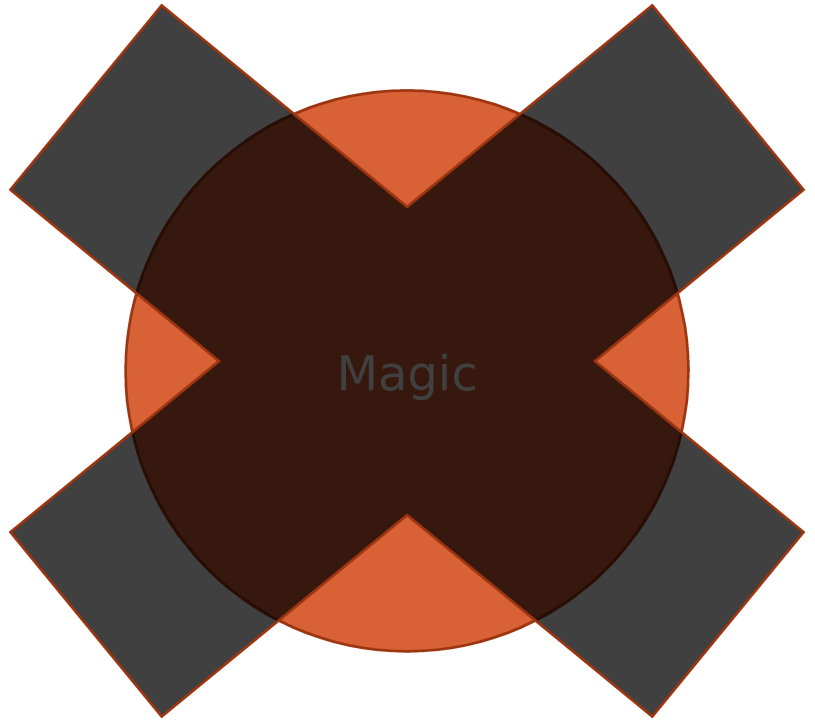
1. What is data science?
2. Using data science to tackle a security problem
   - Machine learning only makes sense when a list of requirements is met
3. Assessing data science solutions

1. WHAT IS DATA SCIENCE?



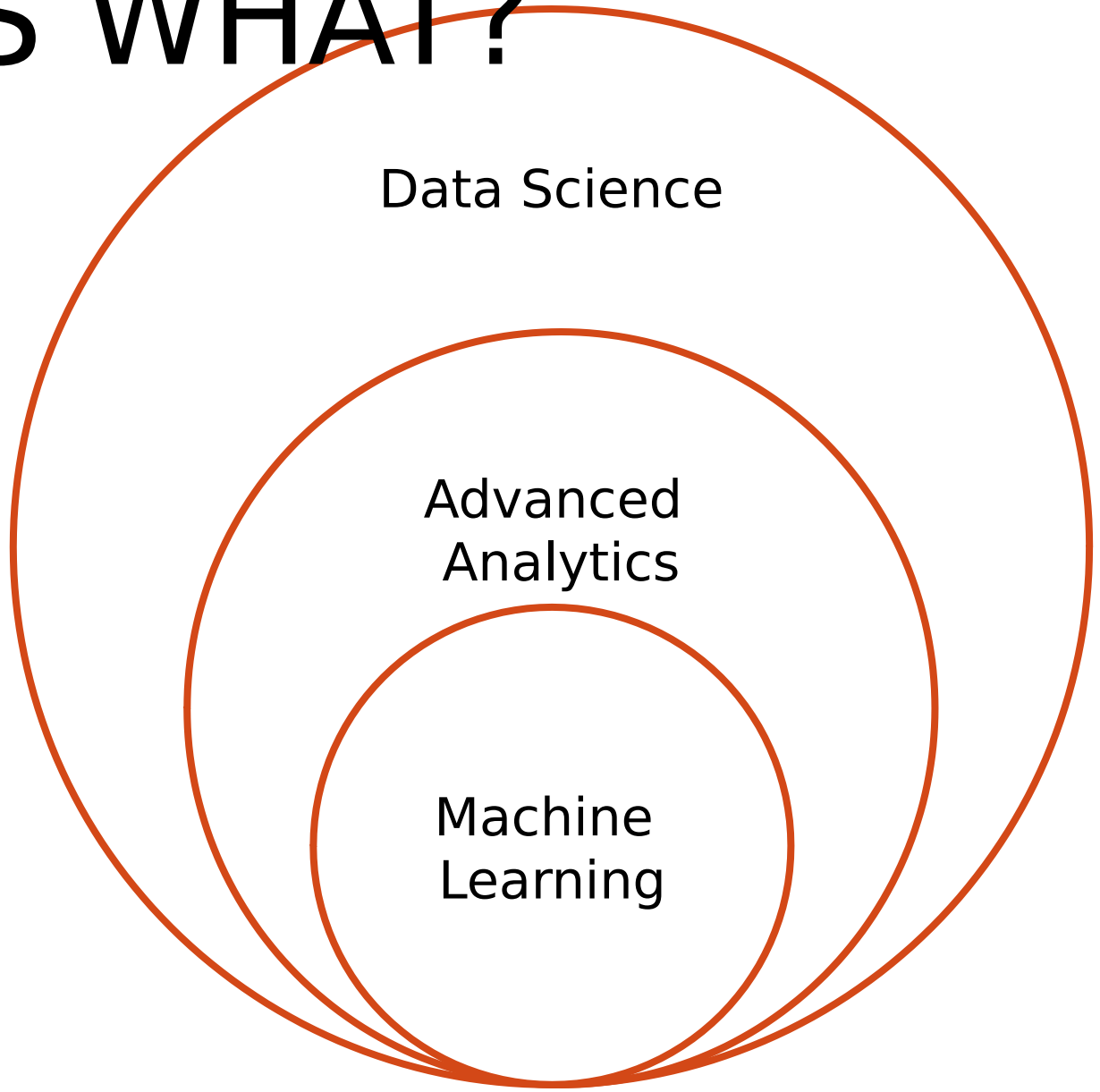
WHAT IS DATA SCIENCE?





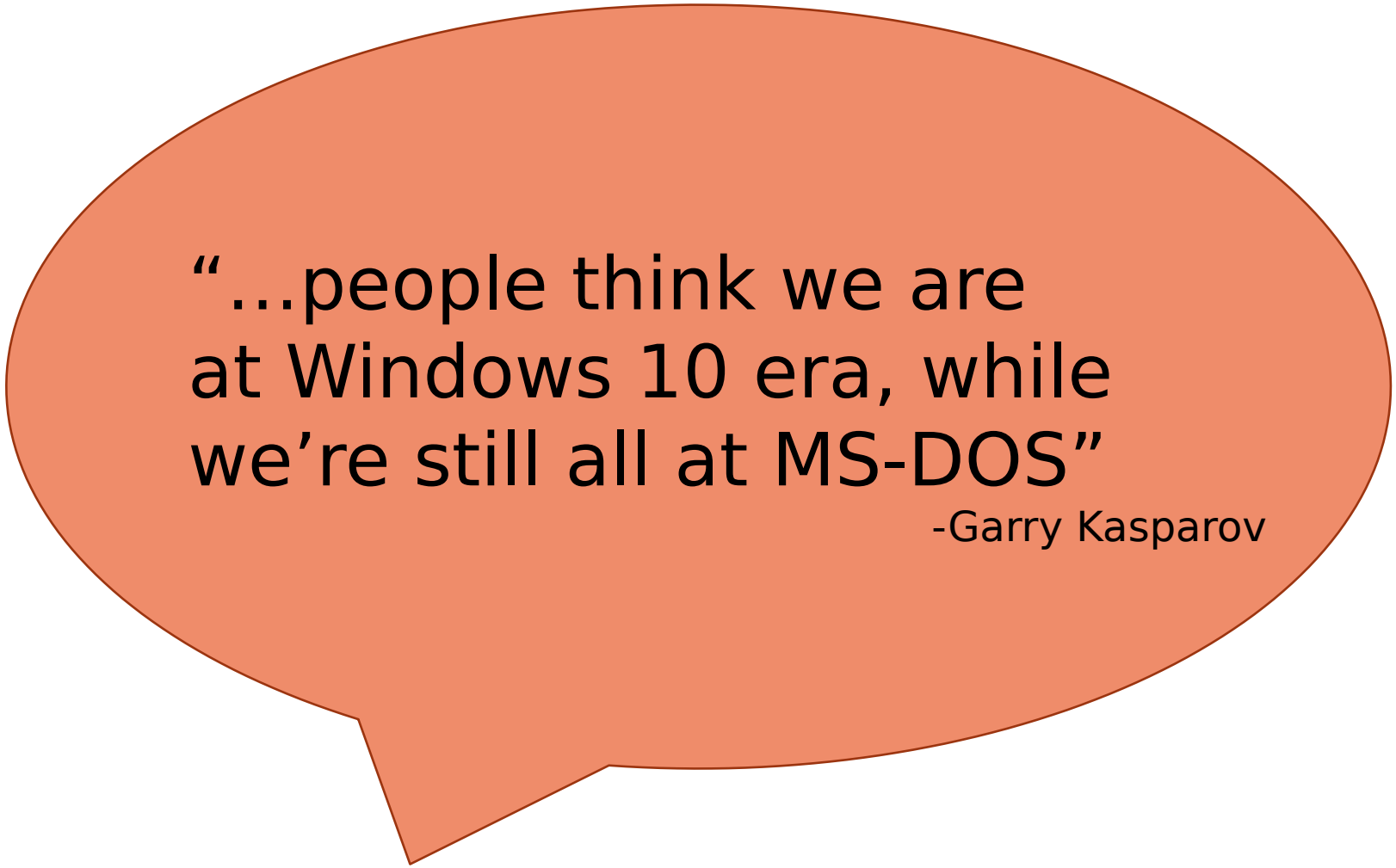
@thordisstella

WHAT IS WHAT?



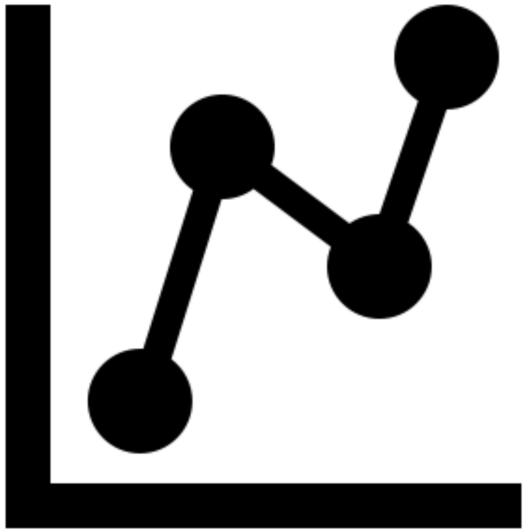
"Buzzword jargon buzzword, hyperbole buzzword buzzword, trite rhyming platitude… Yep, looks good."

Image credit: Anderson https://andertoons.com/



@thordisstella

2. USING DATA SCIENCE TO TACKLE A SECURITY PROBLEM



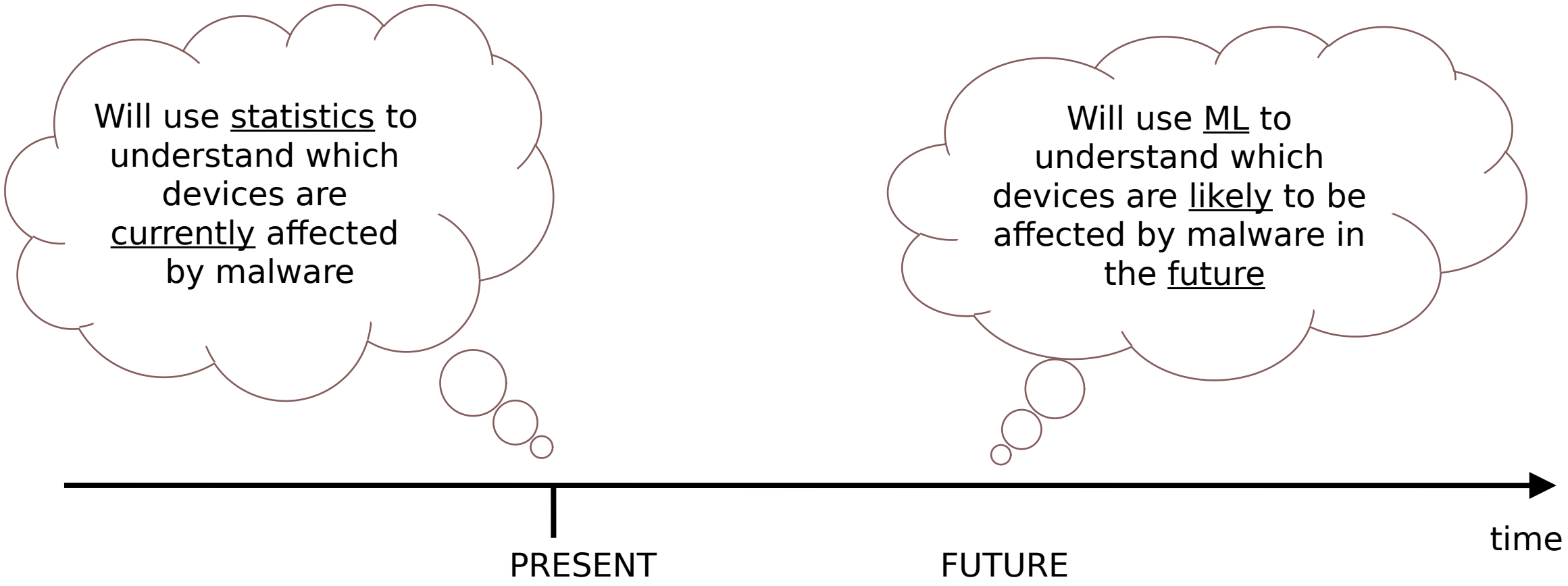
@thordisstella

PROBLEM

Want to reduce the number of devices that get infected with malware

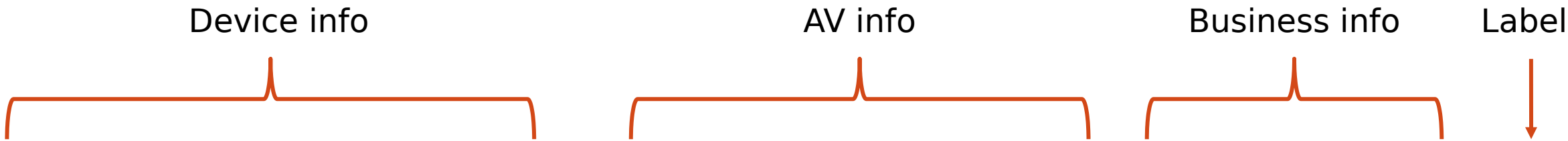
SOLUTION

1. Not machine learning
2. Machine learning



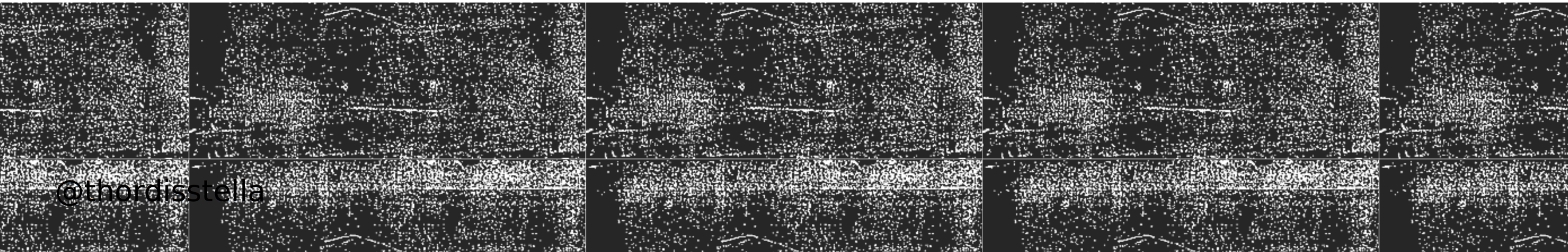
@thordisstella

- Data: https://www.kaggle.com/c/microsoft-malware-prediction



MachineIdentifier	Platform	Census_DeviceFamily	Firewall	AVProductsInstalled	AVProductsEnabled	ProductName	OrganizationIdentifier	GeoNameIdentifier	HasDetection
d8ejkl2alHuj	windows10	Windows.Desktop	1	2	1	win8defender	18	35	0

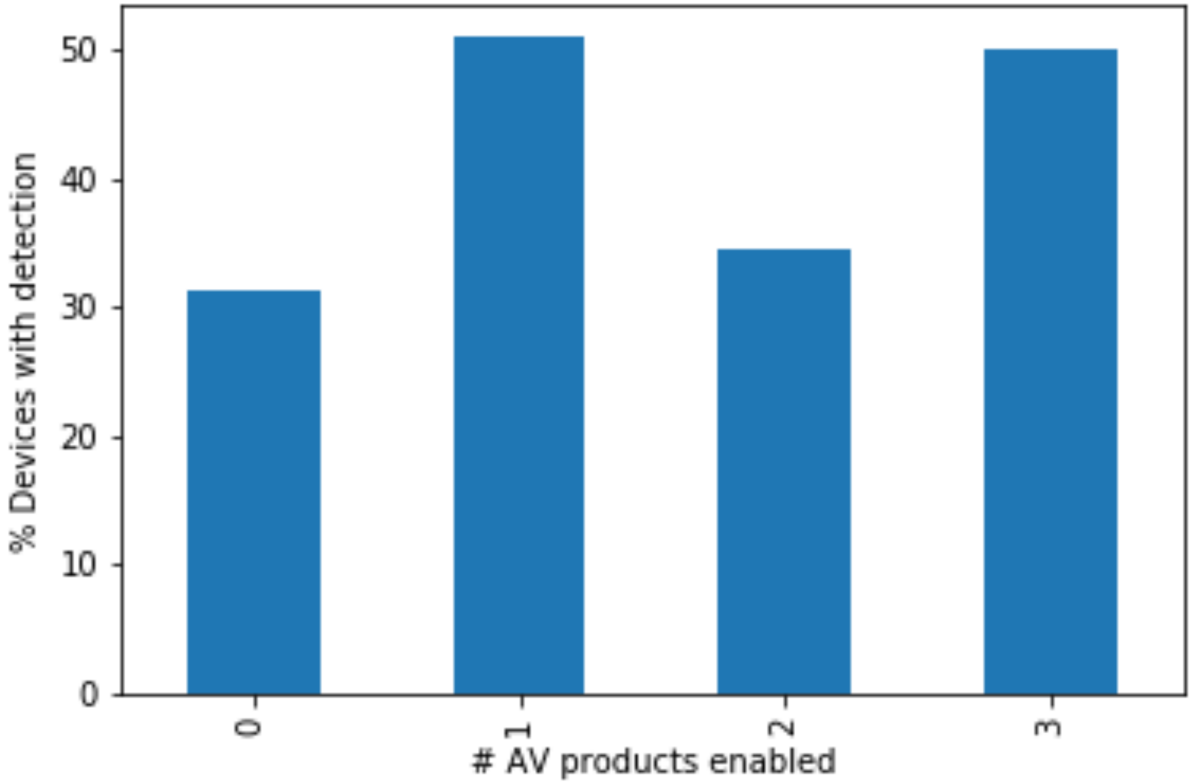
APPROACH 1: NOT MACHINE LEARNING



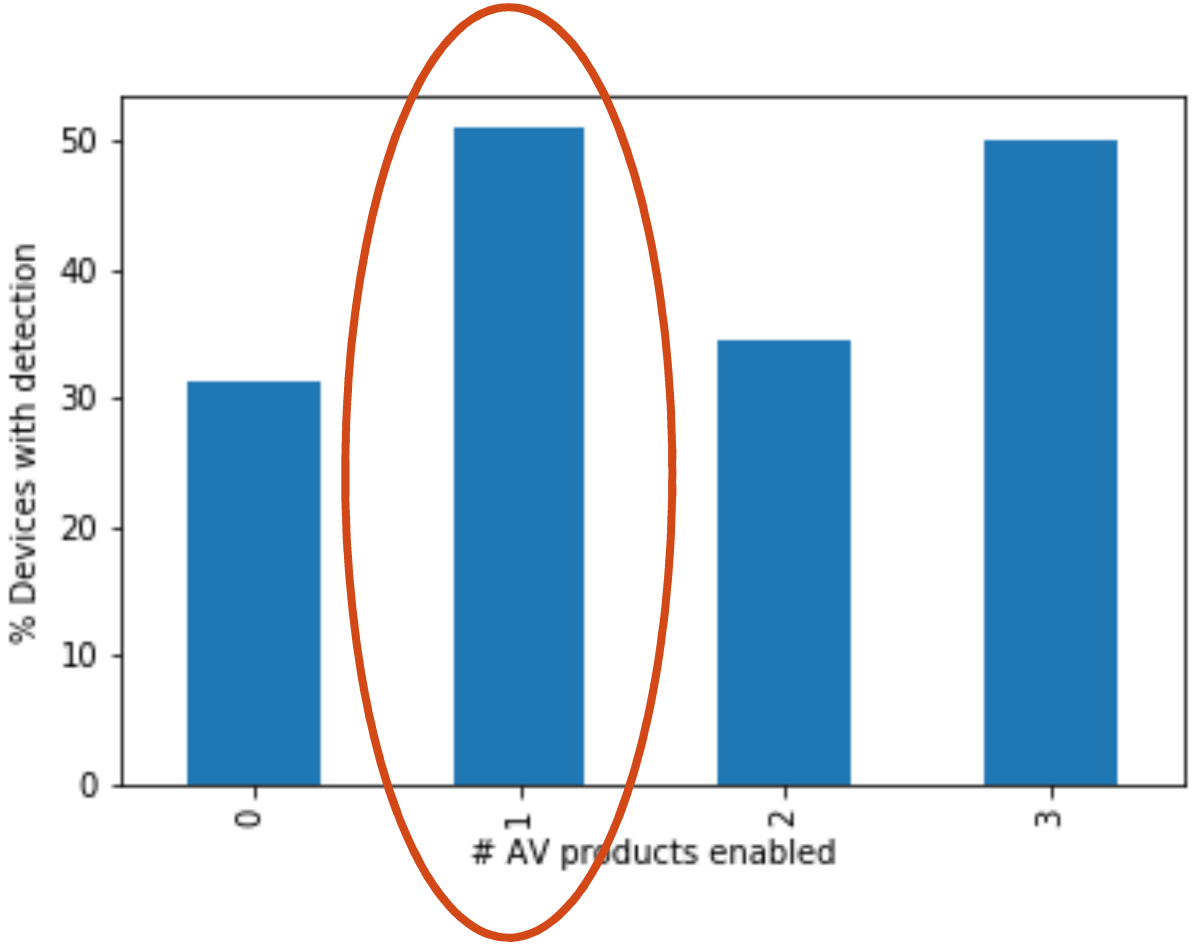
@thordisstella

AVProductsEnabled	HasDetections	count
0	0	44
0	1	20
1	0	4761
1	1	4954
2	0	143
2	1	75
3	0	1
3	1	1
4	0	1

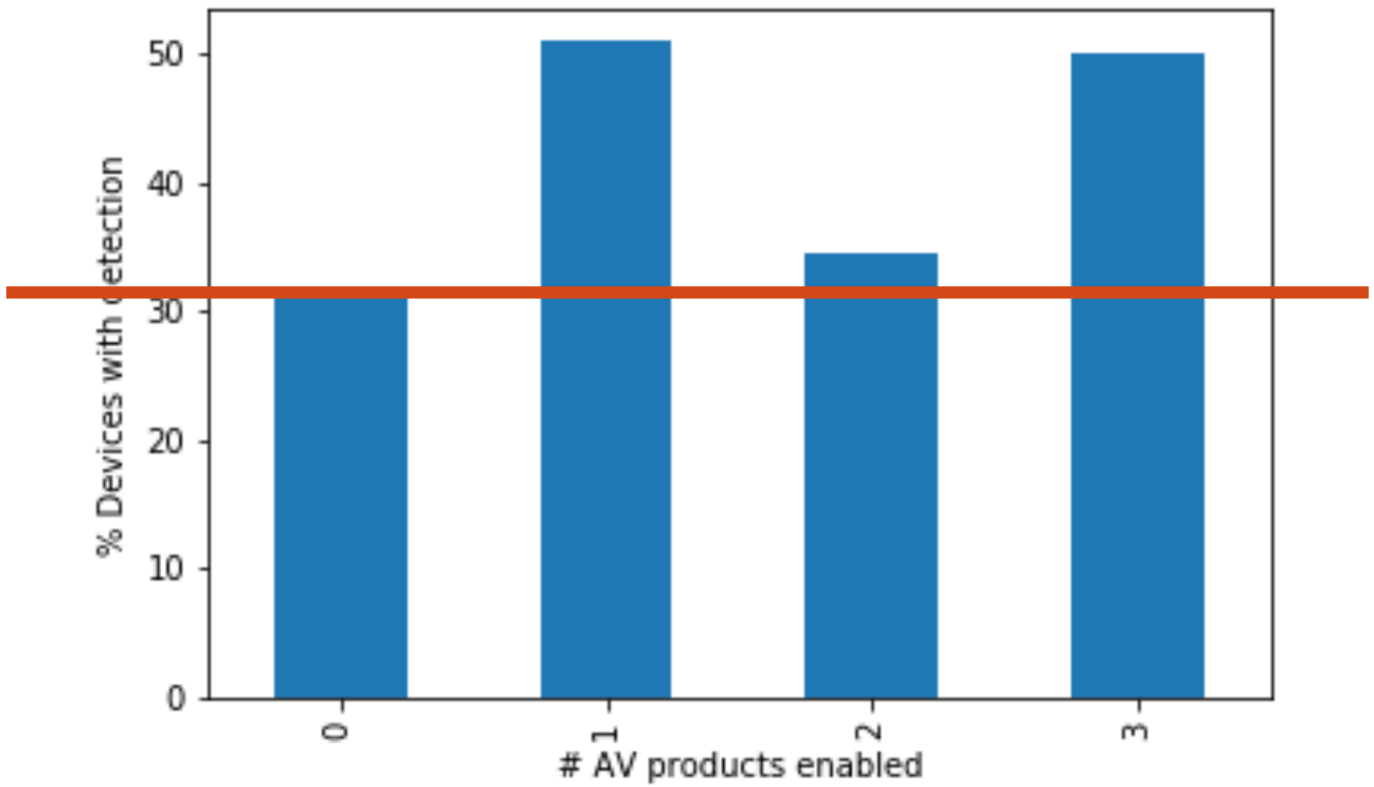
AVProductsEnabled	HasDetections	count	perc
0	0	44	68.8
0	1	20	31.2
1	0	4761	49.0
1	1	4954	51.0
2	0	143	65.6
2	1	75	34.4
3	0	1	50.0
3	1	1	50.0
4	0	1	100.0



@thordisstella

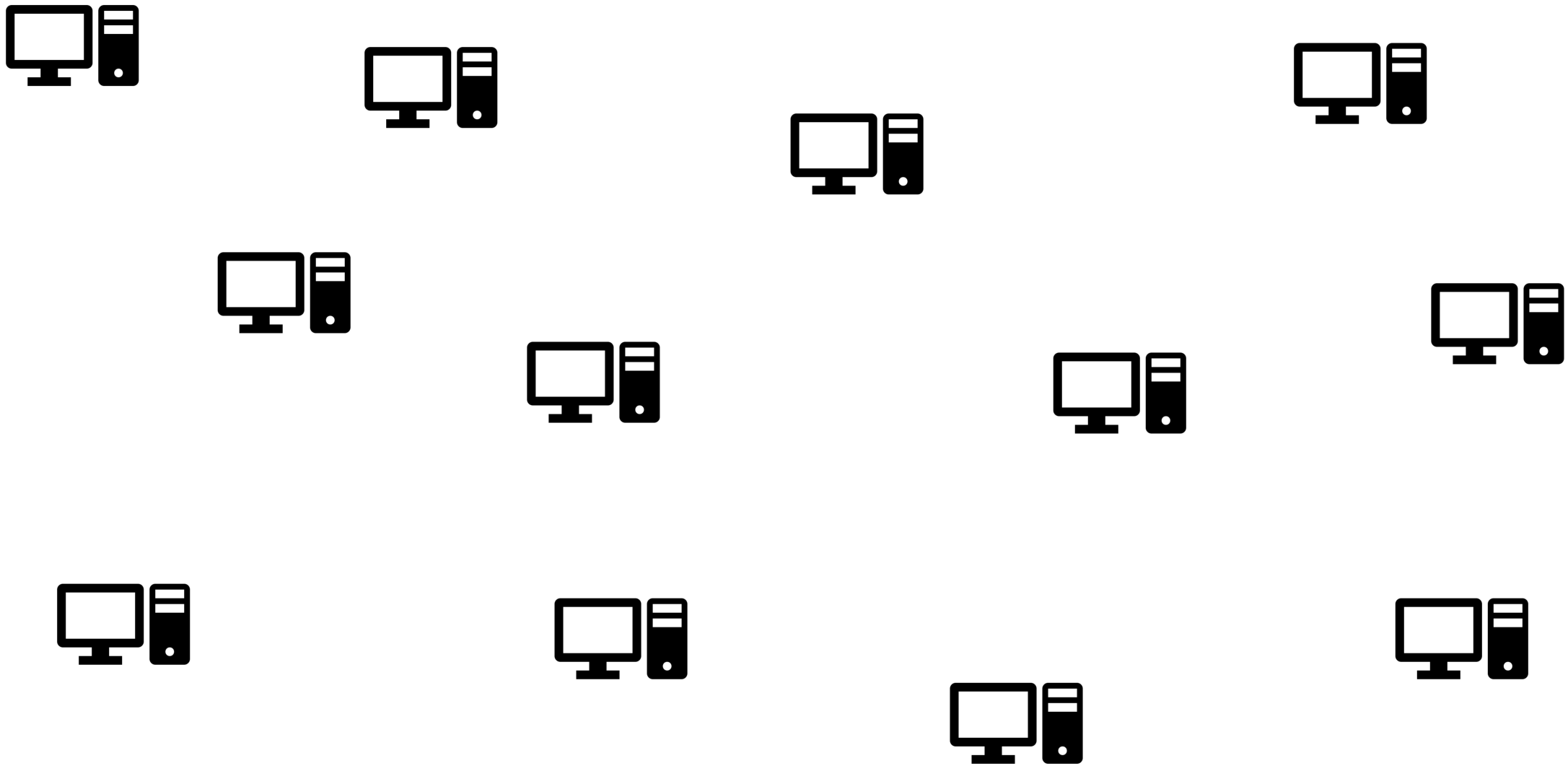


@thordisstella



Are AV products correctly configured?

@thordisstella



@thordisstella

Applying business context

OrganizationIdentifier	AVProductsEnabled	HasDetections	count
18	1	0	3
27	1	0	14
37	1	1	3

Applying business context

OrganizationIdentifier	AVProductsEnabled	HasDetections	count
18	1	0	3
27	1	0	14
37	1	1	3

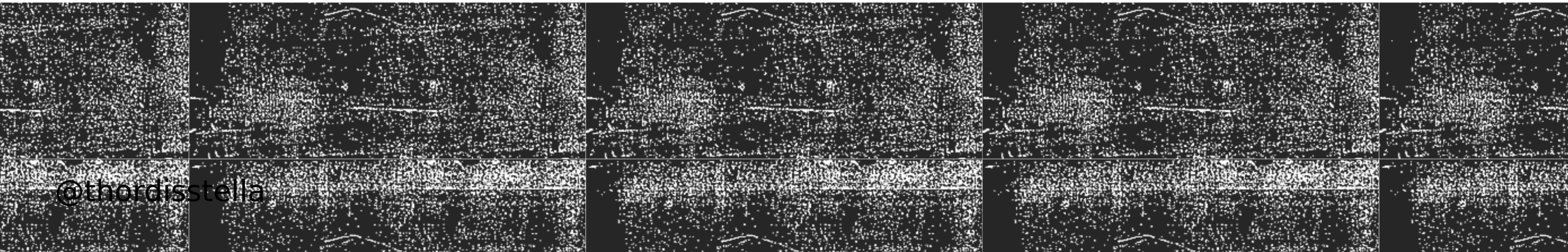
Applying business context

OrganizationIdentifier	AVProductsEnabled	HasDetections	count
18	1	0	3
27	1	0	14
37	1	1	3

KEY TAKEAWAYS

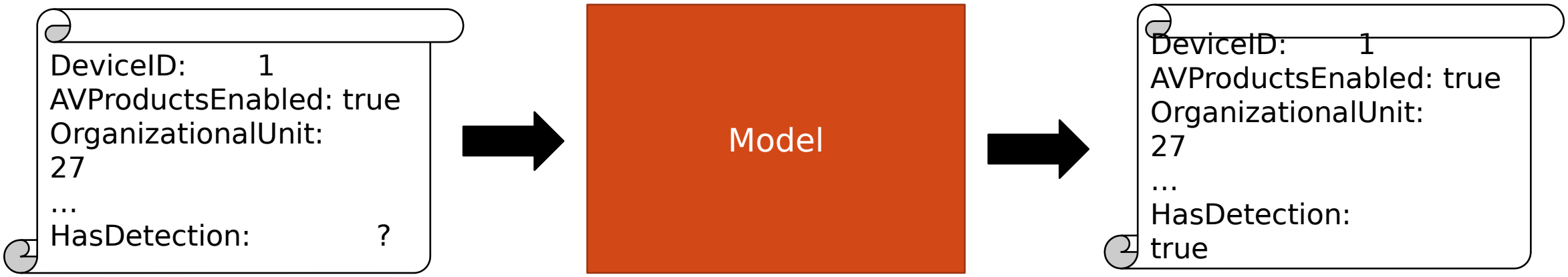
- Look at how key characteristics affect outcome
- Focus on the part of the problem you care most about

APPROACH 2: MACHINE LEARNING



APPLYING ML TO OUR PROBLEM

- Goal:
  - Be able to predict whether a device is going to have malware detections, or not
    - Train a model to give us a true/false outcome, given information about the device (binary classification)



	Logistic Regression	Binary Neural Network

	Logistic Regression	Binary Neural Network
Transparent	✔	✖

	Logistic Regression	Binary Neural Network
Transparent	✔	✖
Simple	✔	✖

	Logistic Regression	Binary Neural Network
Transparent	✔	✖
Simple	✔	✖
Efficient to train	✔	✖

	Logistic Regression	Binary Neural Network
Transparent	✔	✖
Simple	✔	✖
Efficient to train	✔	✖
Works without much data pre-processing	✖	✔

	Logistic Regression	Binary Neural Network
Transparent	✔	✖
Simple	✔	✖
Efficient to train	✔	✖
Works without much data pre-processing	✖	✔
Flexible (works well when true and false values can’t be “easily” separated)	✖	✔

	Logistic Regression	Binary Neural Network
Transparent	✔	✖
Simple	✔	✖
Efficient to train	✔	✖
Works without much data pre-processing	✖	✔
Flexible (works well when true and false values can't be "easily" separated)	✖	✔
Performance can generally be improved with more training data	✖*	✔

* Improves up to a certain point, after which more training data has no significant impact

	Logistic Regression	Binary Neural Network
Transparent	✔	✖
Simple	✔	✖
Efficient to train	✔	✖
Works without much data pre-processing	✖	✔
Flexible (works well when true and false values can't be "easily" separated)	✖	✔
Performance can generally be improved with more training data	✖*	✔
Not prone to overfitting	✖	✖

* Improves up to a certain point, after which more training data has no significant impact

1. Logistic Regression

2. Neural Networ

@thordisstella

Image credit: Anderson https://andertoons.com/

REQUIREMENTS

- Clearly formed research question
  - Bad: “What can I do to improve my AV program?”
  - Good: “Does having an AV product enabled decrease the chances of having a malware detection?”
- Good quality data
  - Bad: Only data from the UK, but solution will also be applied in the US
  - Good: Making sure that training data is representative for the world where the solution will be applied
- Understanding overall potential to cause harm
  - Bad: Overwhelming the security analyst with alerts
  - Good: Consider impact of false positive and false negative predictions (e.g. overestimating number of malware detections vs. missing a critical detection)

WHEN REQUIREMENTS WEREN'T MET

Fail: [redacted] Face ID Defeated by a 3D Mask

Source: Lexalytics

[redacted] ditched AI recruiting tool that favored men for technical jobs

Source: Guardian

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Fail: [redacted] for Oncology" Cancelled After $62 million and Unsafe Treatment Recommendations

Source: Lexalytics

@thordisstella

KEY TAKEAWAYS

- Not all machine learning algorithms are created equal
  - Pros and cons must align with your solution requirements
- Machine learning can only be applied when a list of requirements is met
  - Clearly formed research question
  - Good quality data
  - Understanding overall potential to cause harm

3. ASSESSING DATA SCIENCE SOLUTIONS

@thordisstella

“Our solution provides unlimited information and insights driven by leading artificial intelligence.”

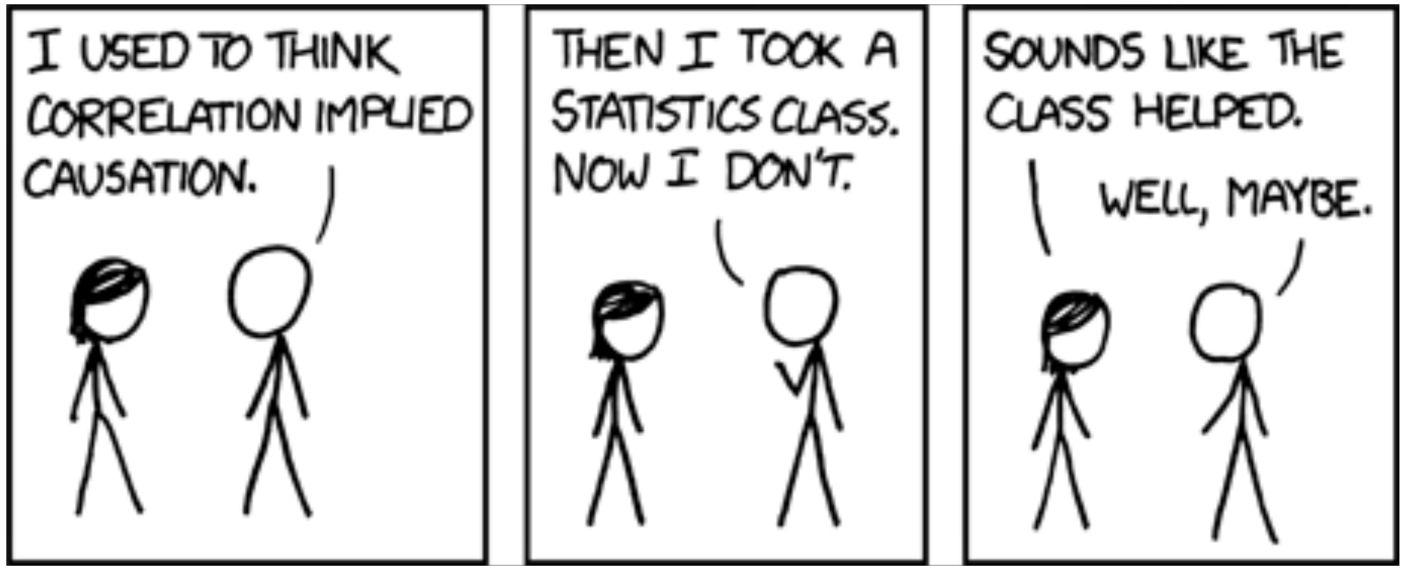
CHEAT SHEET FOR DATA SCIENCE CONSUMERS

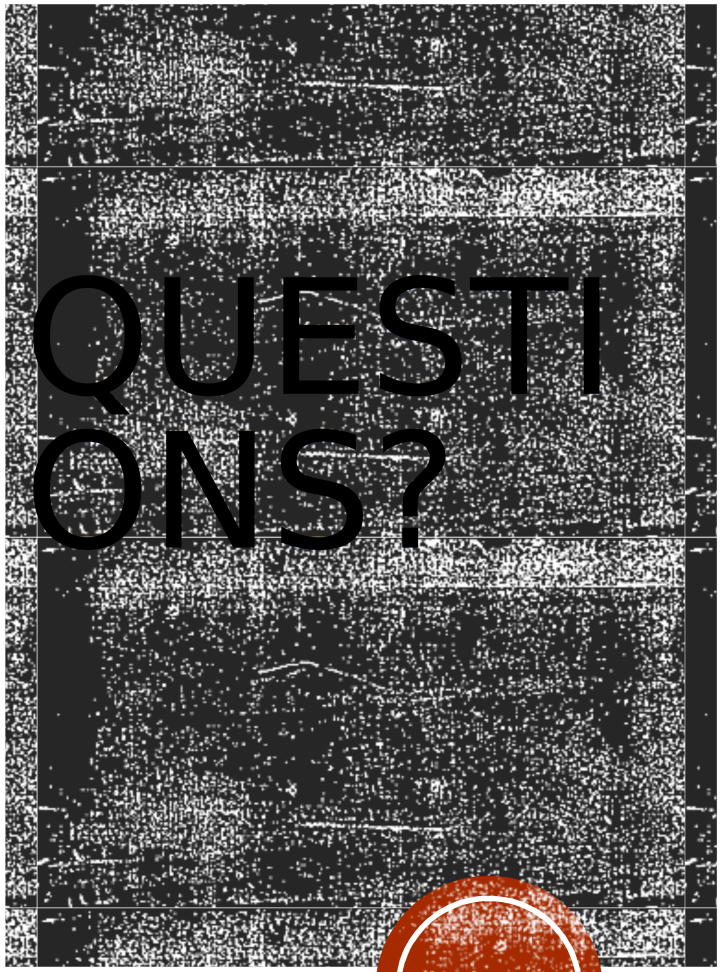
- “What do you mean?”
- ”How specifically do you use [insert buzzword here]?”
- “Why did you choose to use [insert buzzword here]?”
- “How was the model trained?”
- ”Can I validate the decisions being made?”
- ”How do you minimize false positives/false negatives?”

Don’t worry about asking naïve questions - ask about anything that is unclear

WHAT IT ALL BOILS DOWN TO

- Data science provides powerful tools to use in security
  - Stats can often give you great insight
  - Machine learning only makes sense when:
    - Requirements are met
    - Pros and cons of algorithm align with your use case
- You should ask about anything that’s unclear for data science solutions you consider using
  - Many of the marketing statements you hear are meaningless
  - Use cheat sheet to assess solutions





Thordis Thorsteins
thordis.thorsteins@panaseer.com
@thordisstella

Image credit: XKCD https://m.xkcd.com/552/