Securing AI Systems

WILD WEST HACKIN' FEST

Sam Bowne

Oct, 2025

Whoami

- Sam Bowne
- Instructor at City College San Francisco
- Corporate trainer
- Web: samsclass.info
- Email:
  sbowne@ccsf.edu
  sam.bowne@infosecdecoded.com
- Mastodon:
  sambowne@infosec.exchange

Materials Freely Available

Wild West Hackin' Fest

Deadwood, South Dakota
Wed, Oct 8 - Fri, Oct 10, 2025

Securing AI Systems

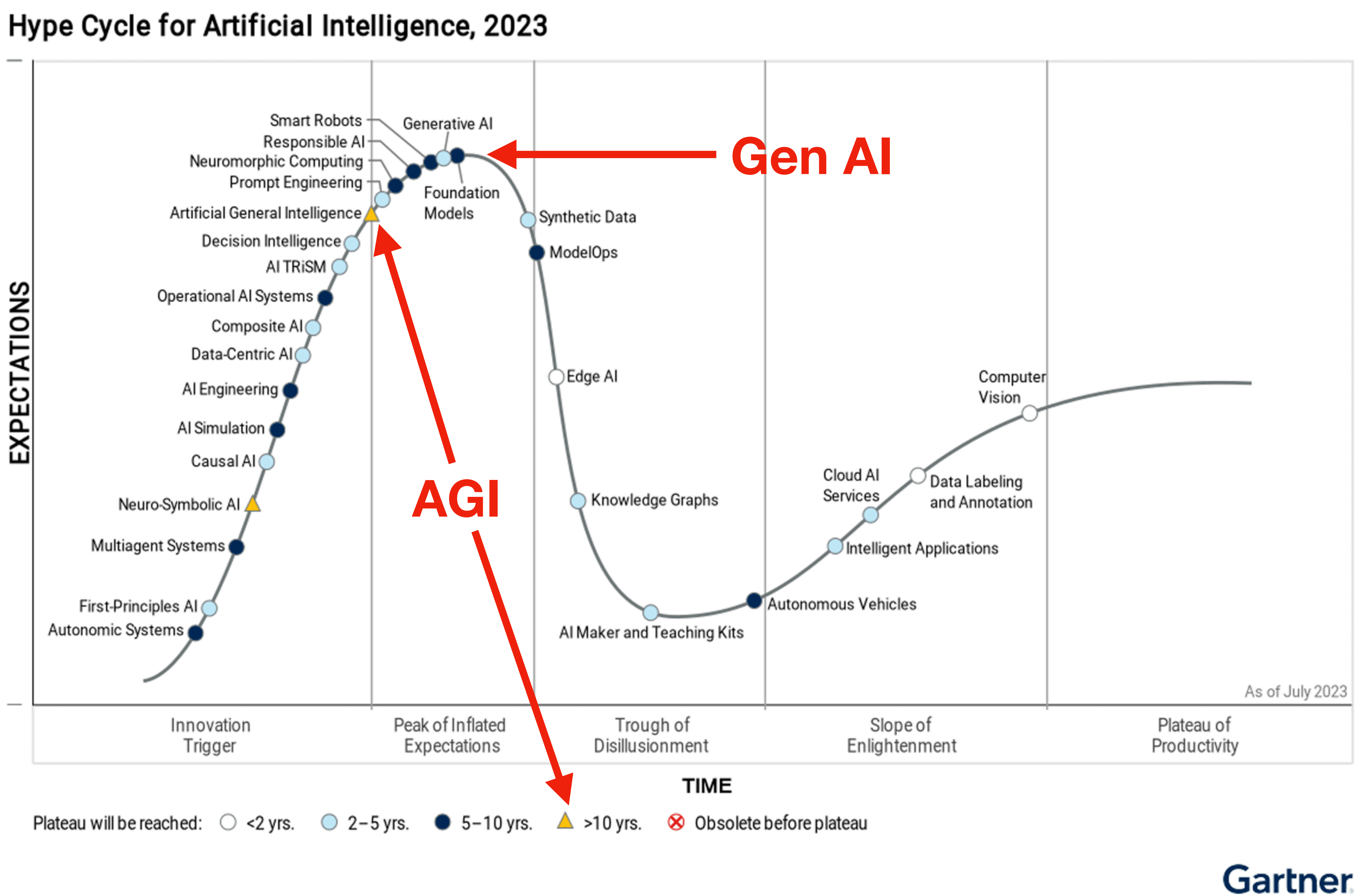
Thu, Oct 9, 3:15 PM - 5:15 PM MT
Homestake Adams Research & Cultural Center (HARCC)
Track 5 (2nd Floor)

Fri, Oct 10, 1:00 PM - 3:00 PM MT
Silverado-Franklin Hotel: Jack McCall Room
(Track 6)

- **samsclass.info**

Types of AI Systems

Figure 1: Hype Cycle for Artificial Intelligence, 2023



Types of Machine Learning Systems

Supervised Learning

- Training data has labels
  - Indicating desired solution

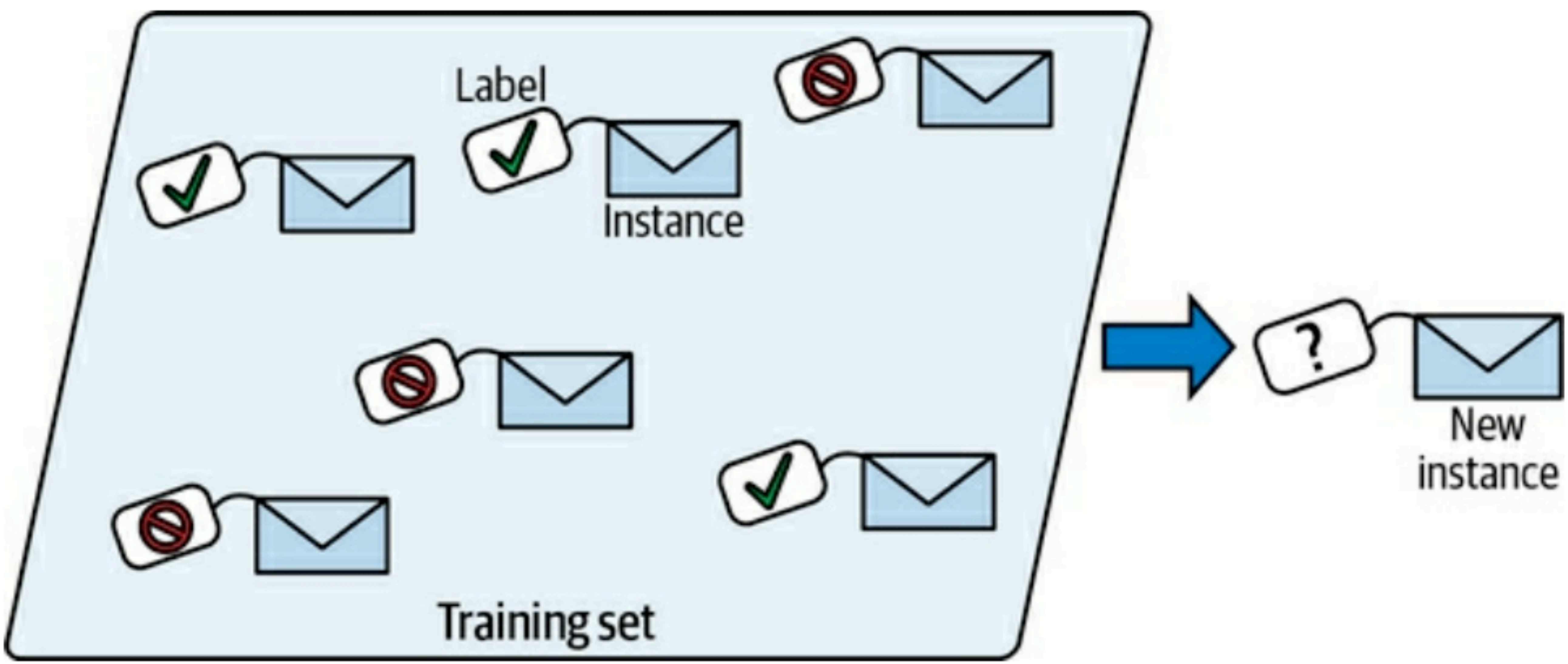
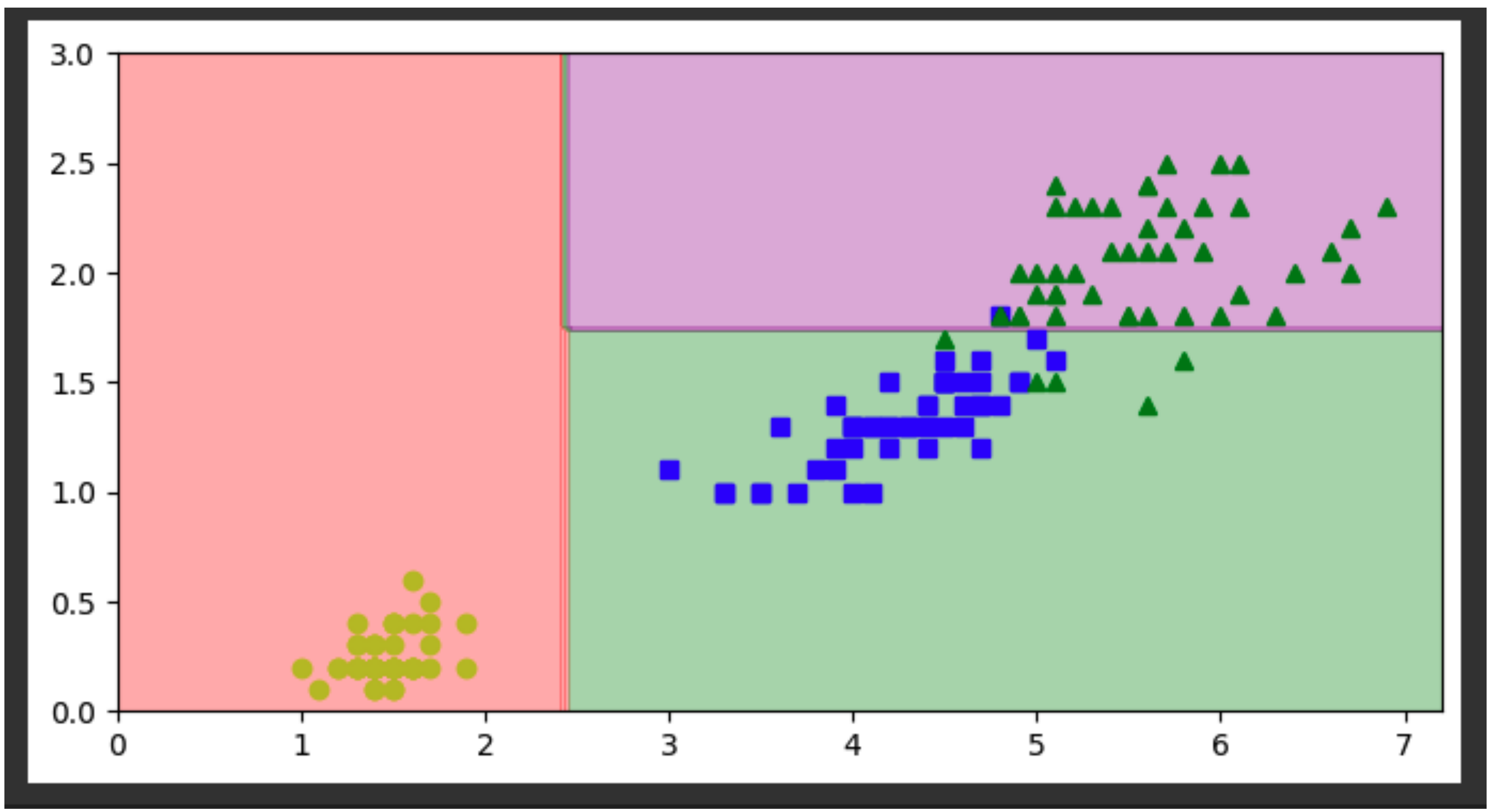


Figure 1-5. A labeled training set for spam classification (an example of supervised learning)

ML 113: Decision Trees

What You Need

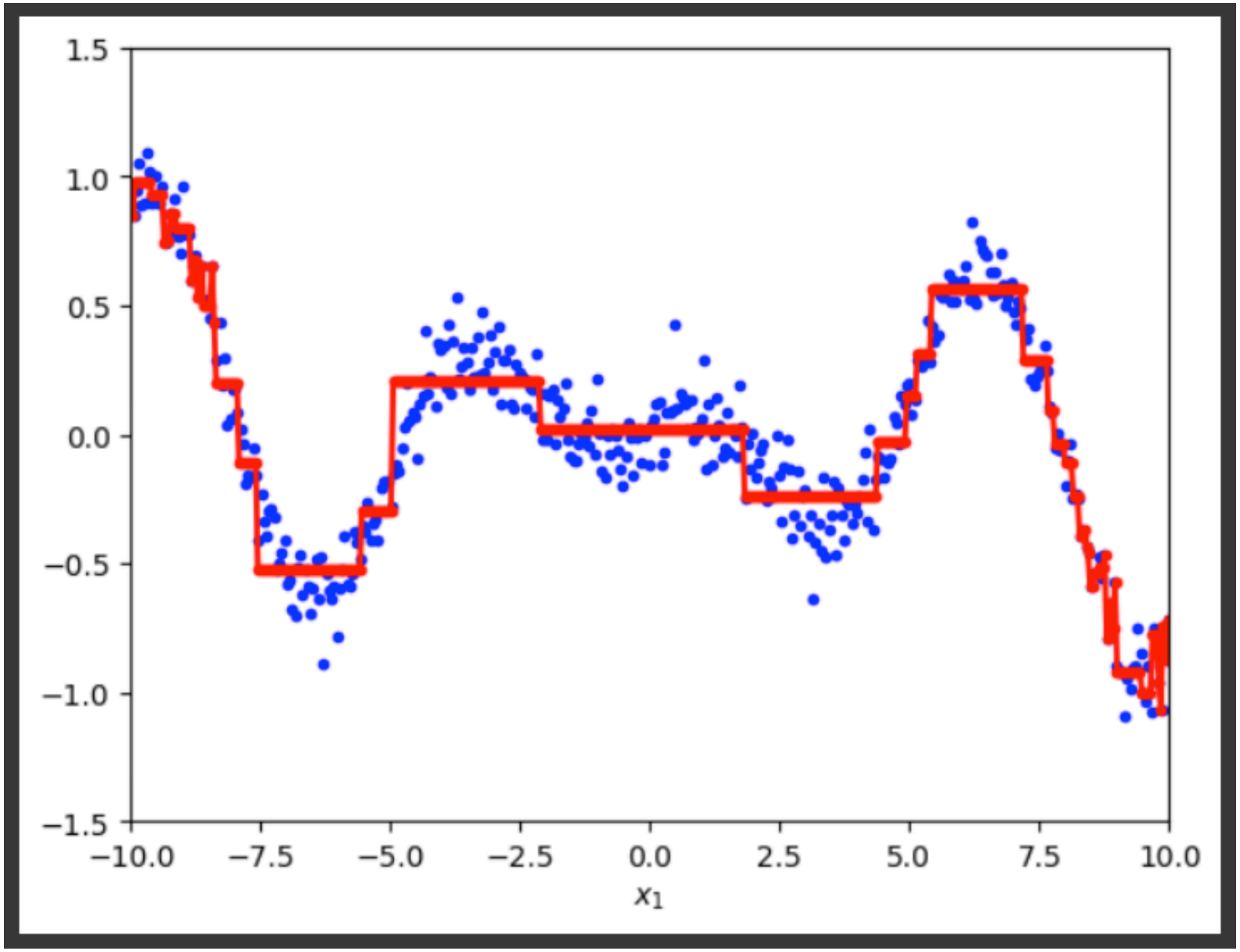
- A Web browser



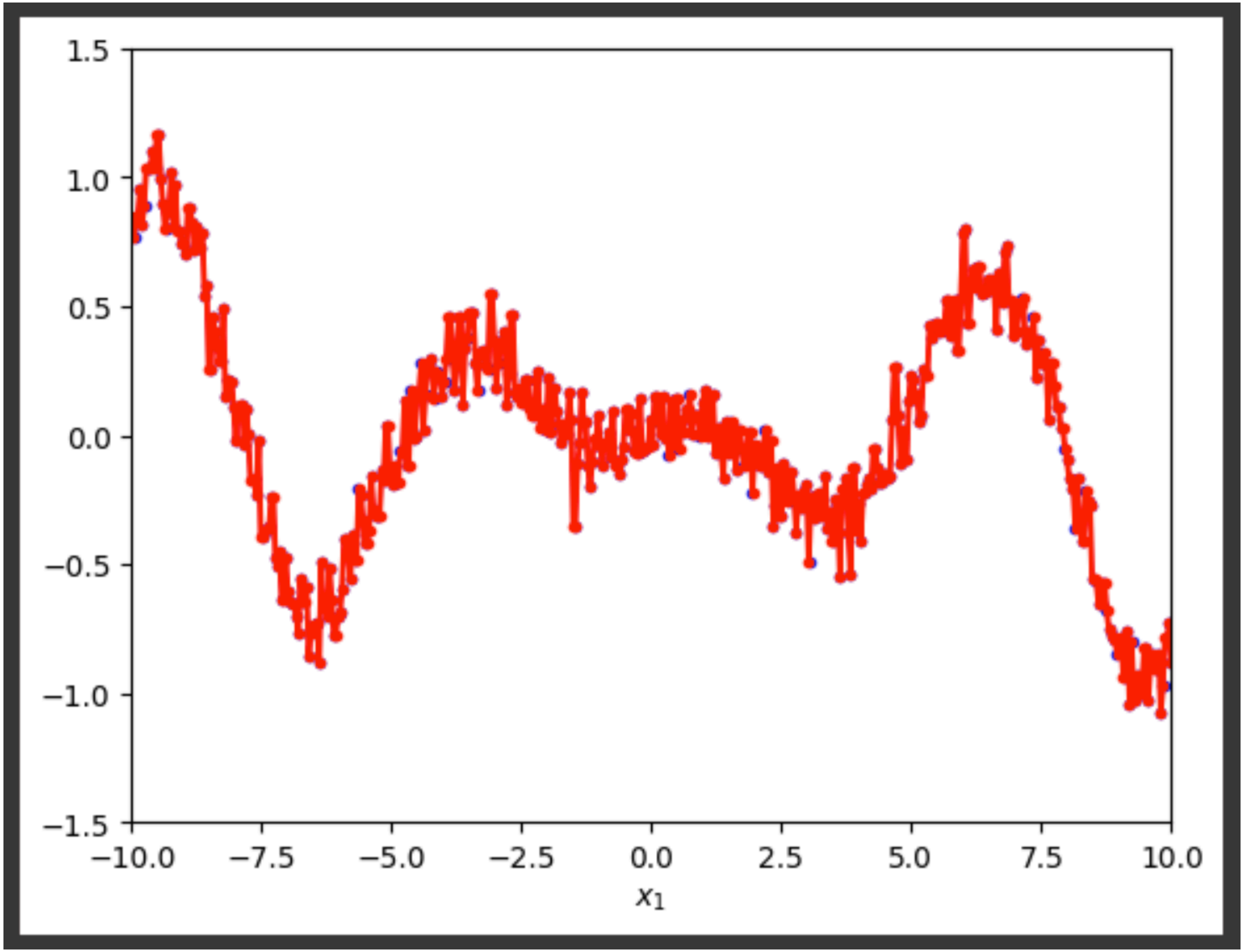
ML 113: Decision Trees

What You Need

- A Web browser



Overtraining

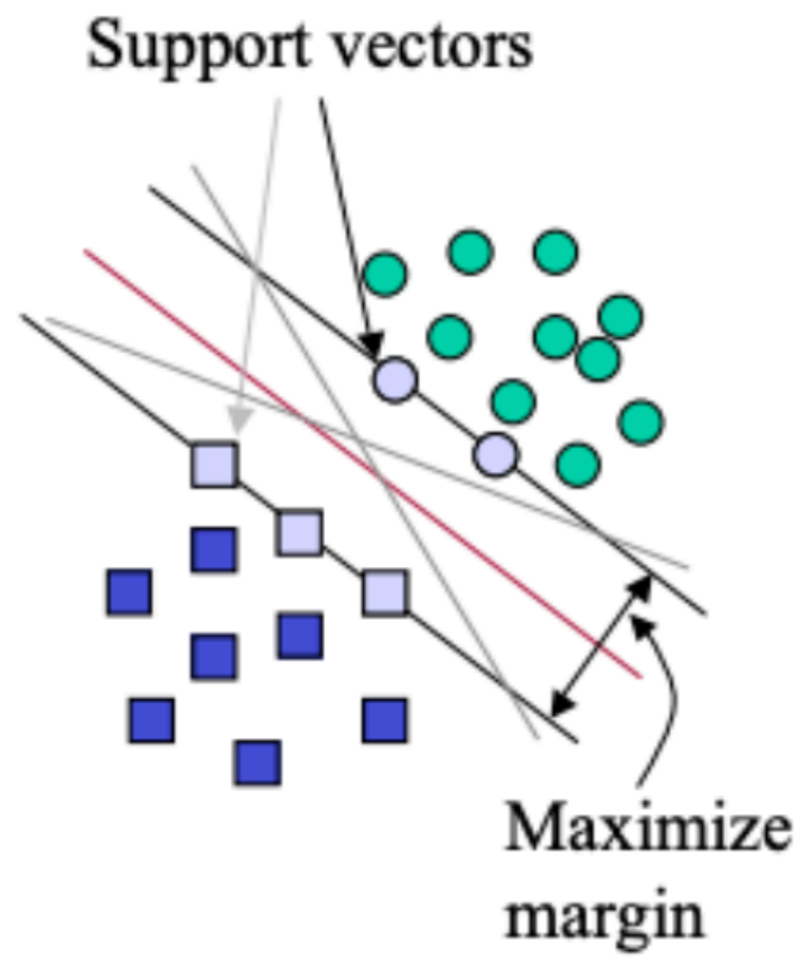


ML 112: Support Vector Machines

What You Need

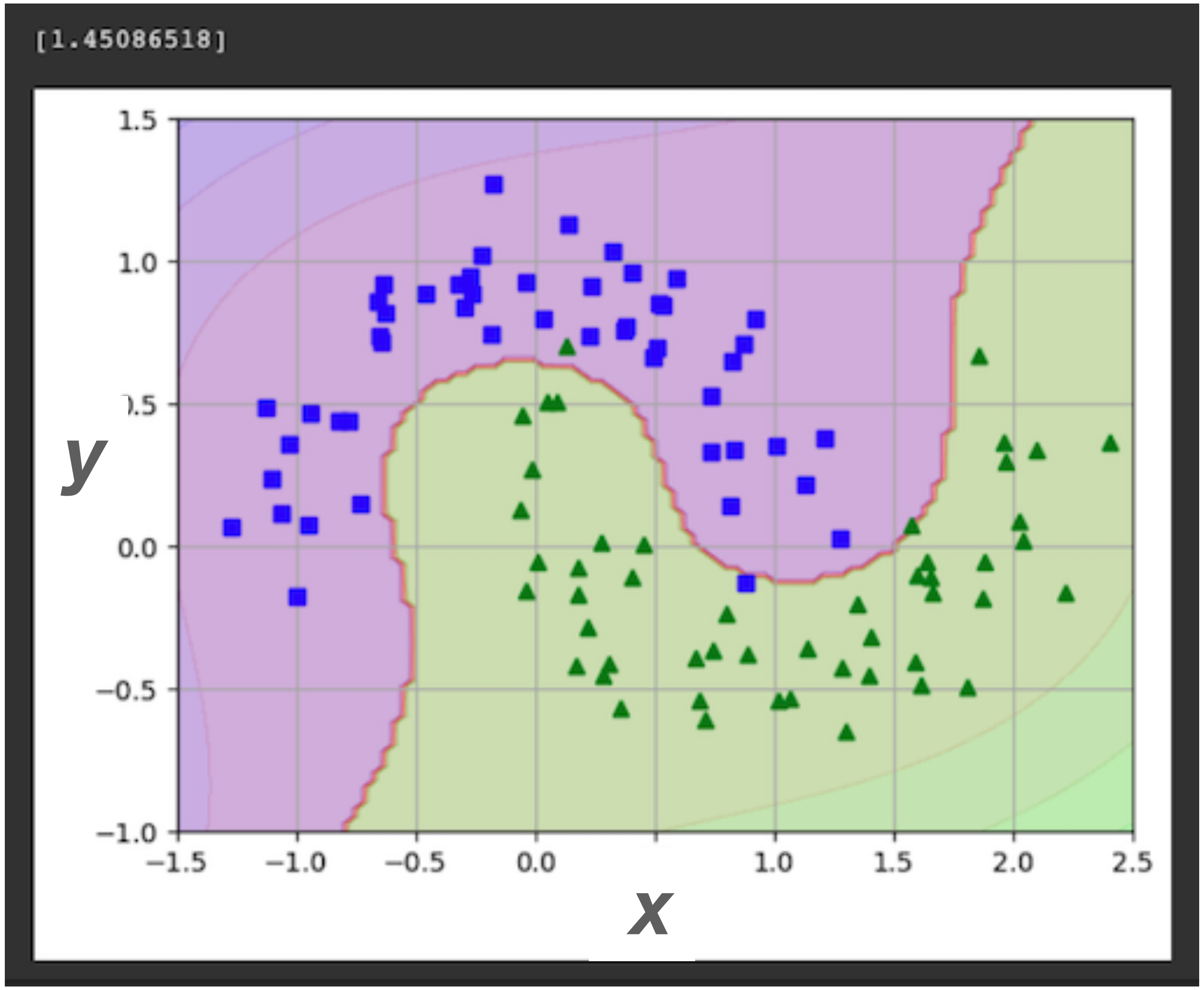
- A Web browser

- SVMs <u>maximize</u> the margin (Winston terminology: the ‘street’) around the separating hyperplane.
- The decision function is fully specified by a (usually very small) <u>subset</u> of training samples, the <u>support vectors.</u>



Polynomial Features

- Add calculated features
  - ***x*²**, ***x*³**, ***y*²**, etc.
- A hyperplane in that multi-dimensional space is a curve in the plane



Unsupervised Learning

- Training data is unlabeled

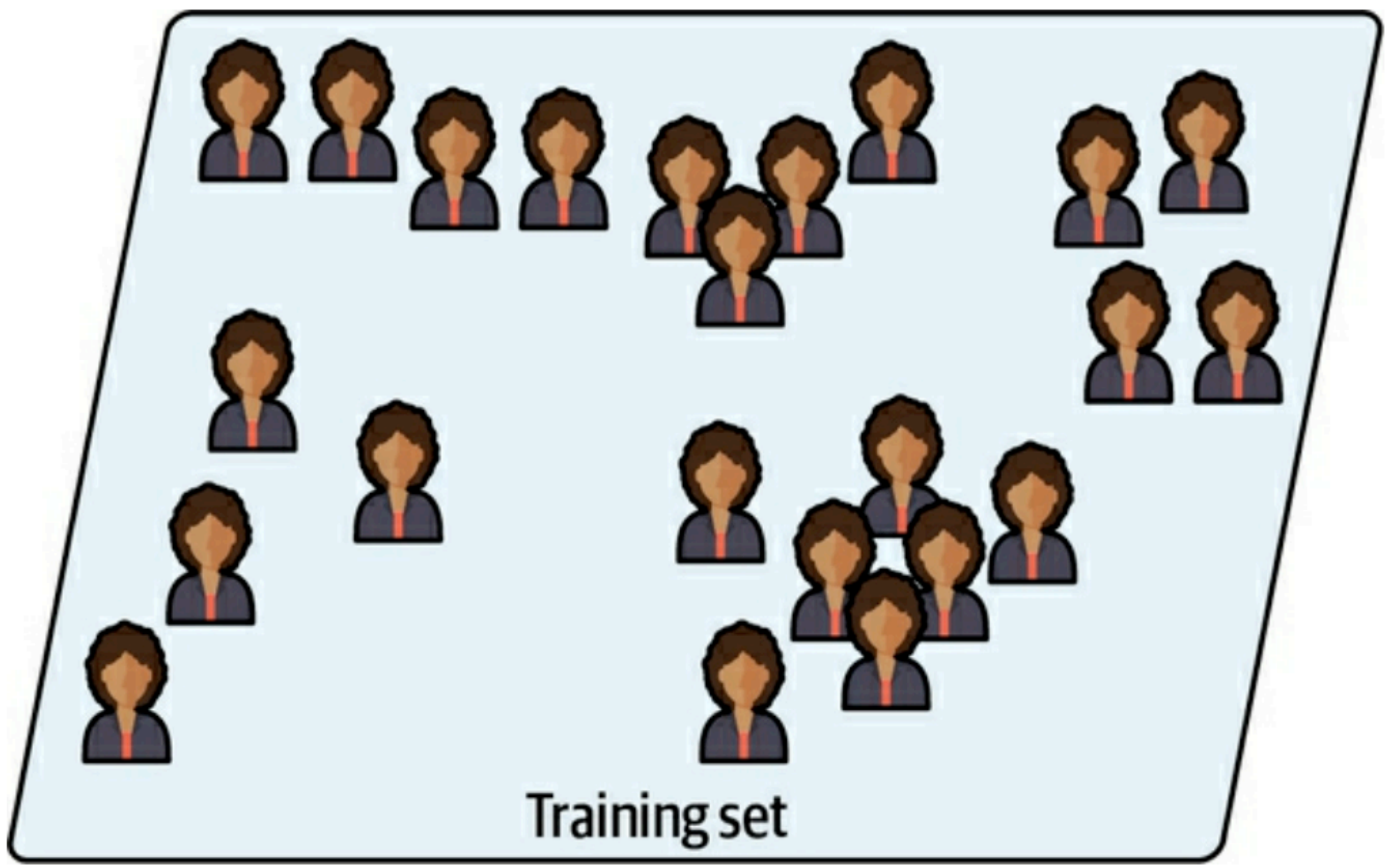


Figure 1-7. An unlabeled training set for unsupervised learning

Unsupervised Learning

- **Clustering** algorithm
  - Sorts data into groups

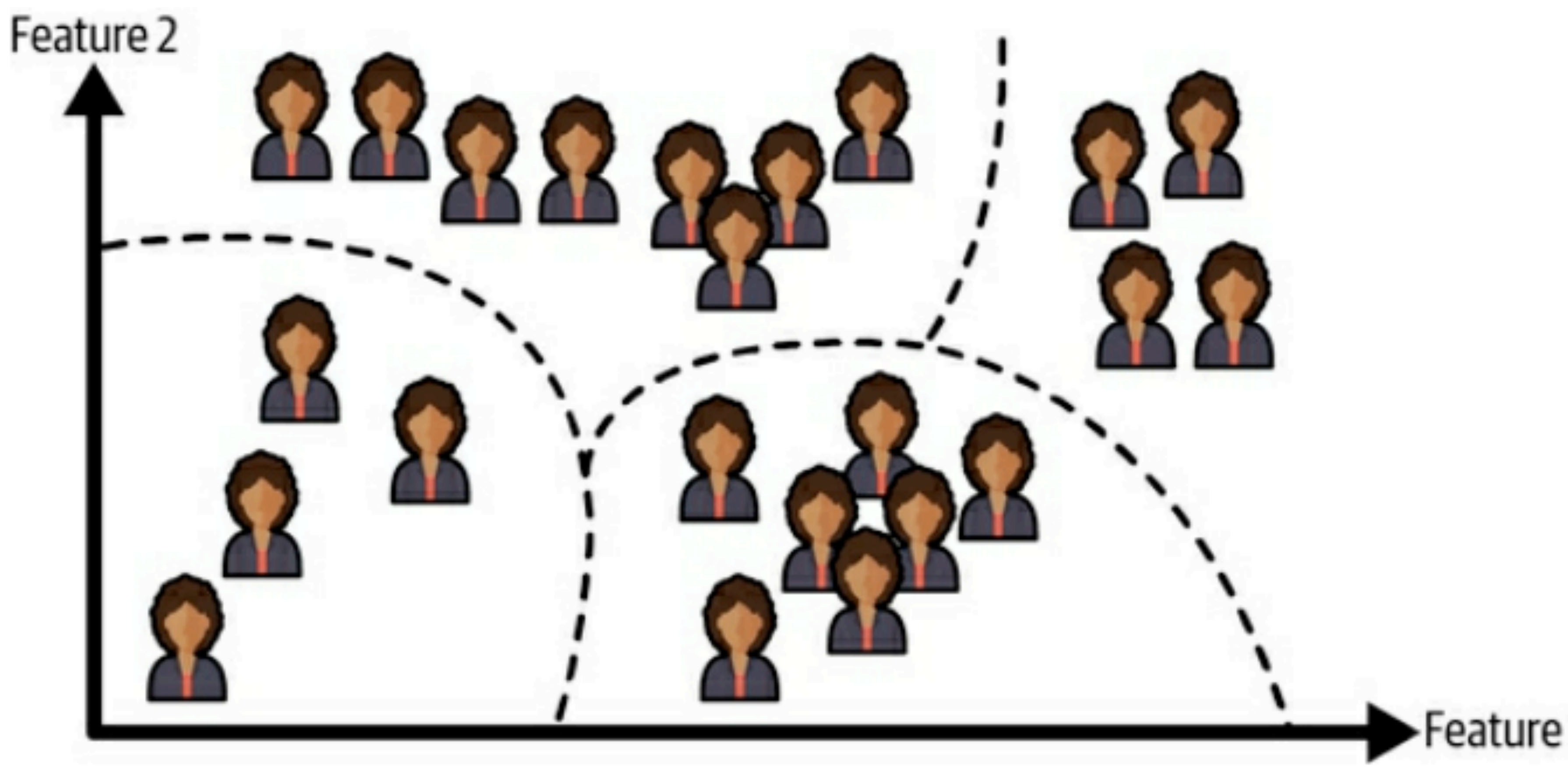
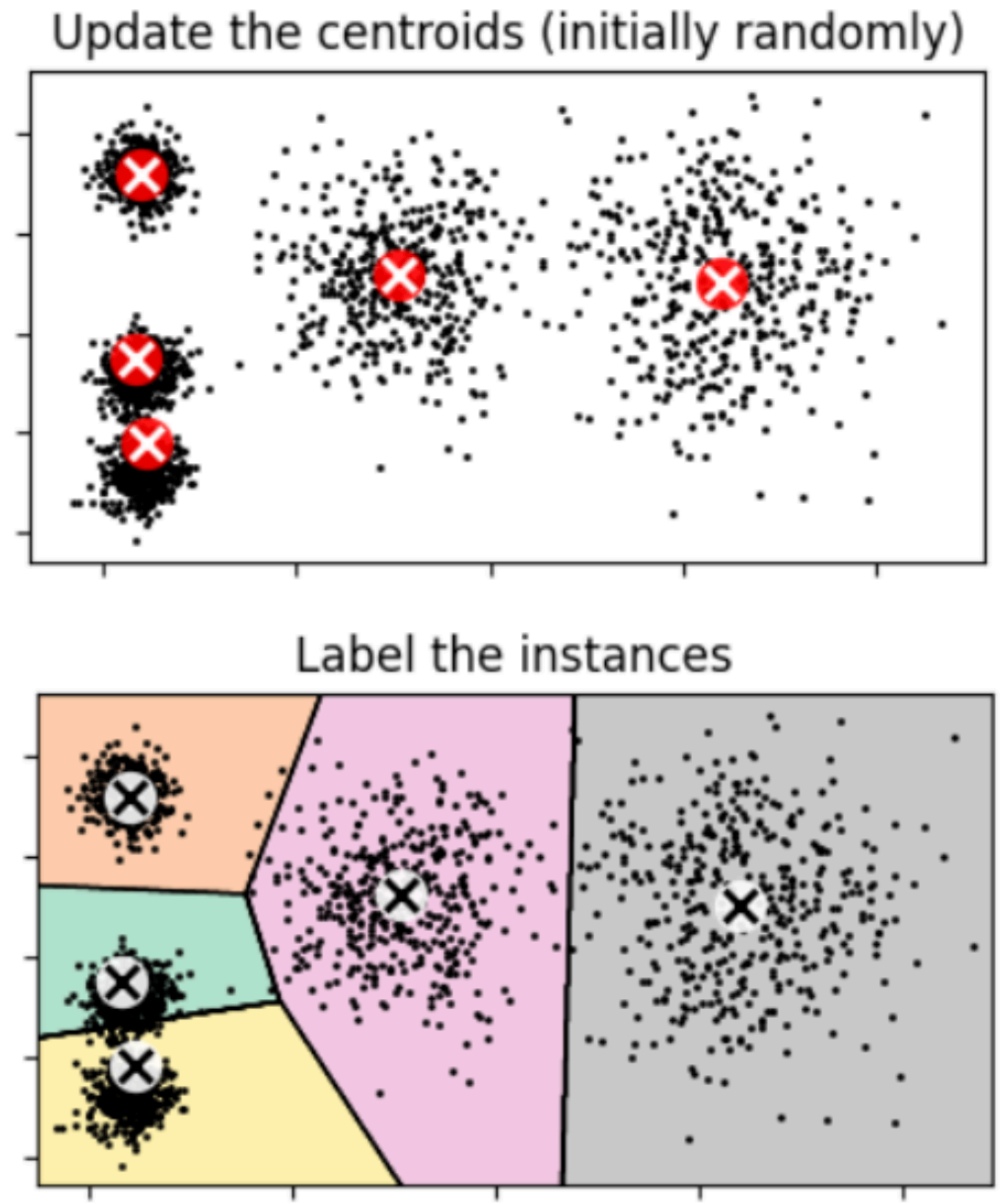


Figure 1-8. Clustering

ML 116: k-Means Clustering

- Fast, simple clustering algorithm



Unsupervised Learning

- **Anomaly detection**
  - Find unusual credit card transactions
  - Find manufacturing defects

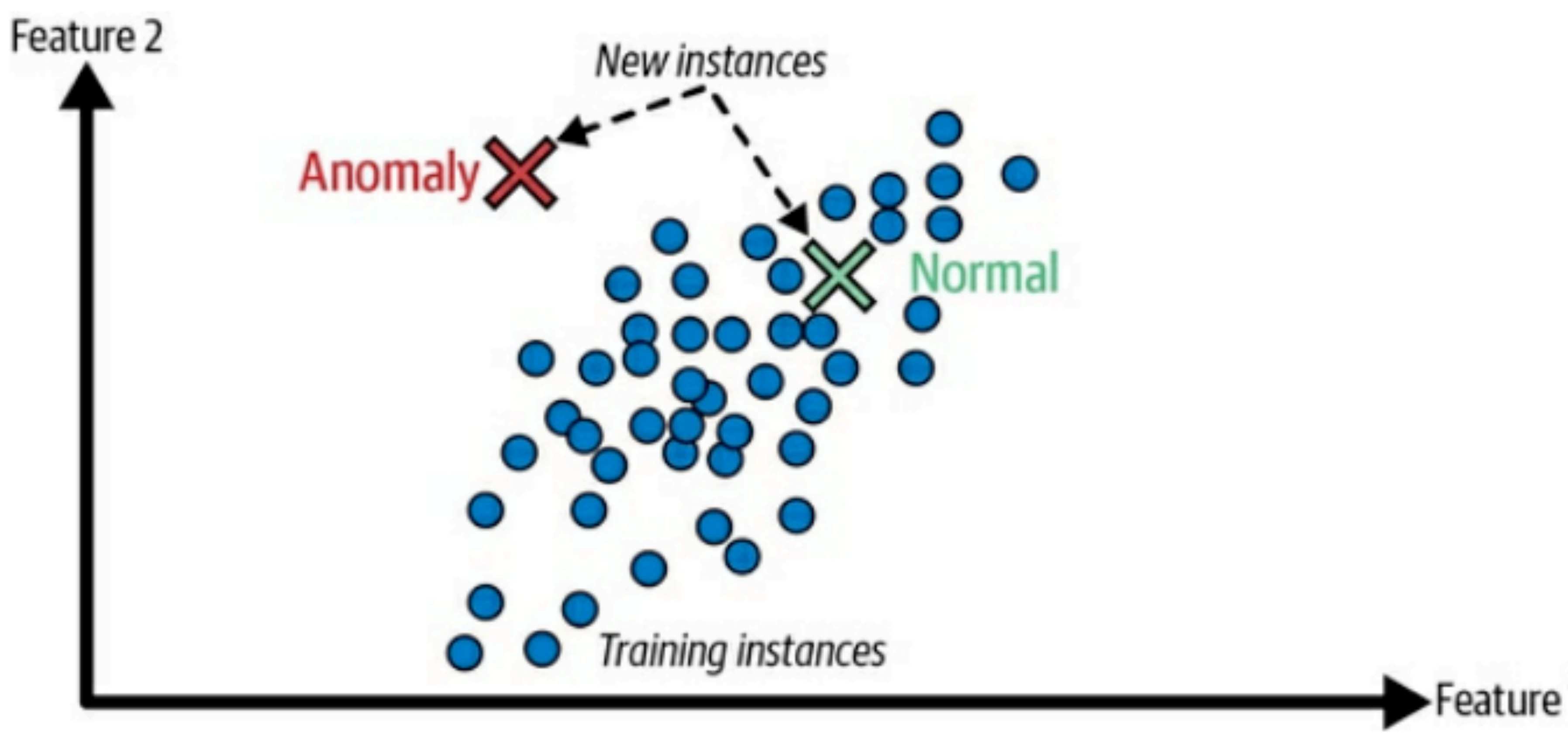


Figure 1-10. Anomaly detection

Self-Supervised Learning

- Generates a labeled dataset from an unlabeled one
  - Example: mask part of an image, train a model to recover the original image

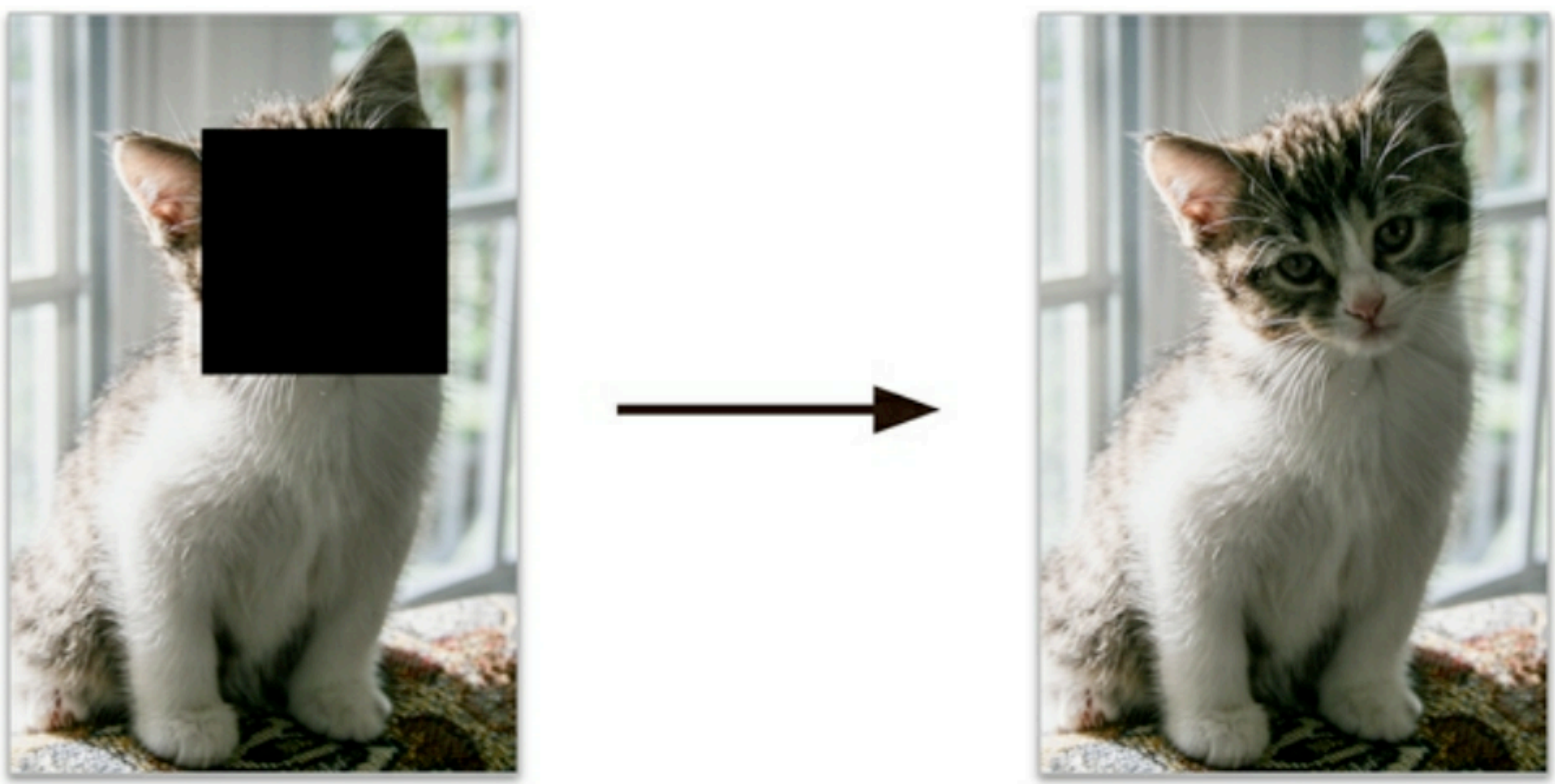


Figure 1-12. Self-supervised learning example: input (left) and target (right)

Large Language Models

- Start with sentences written by humans
- Randomly mask some words
- Learn to predict the masked word

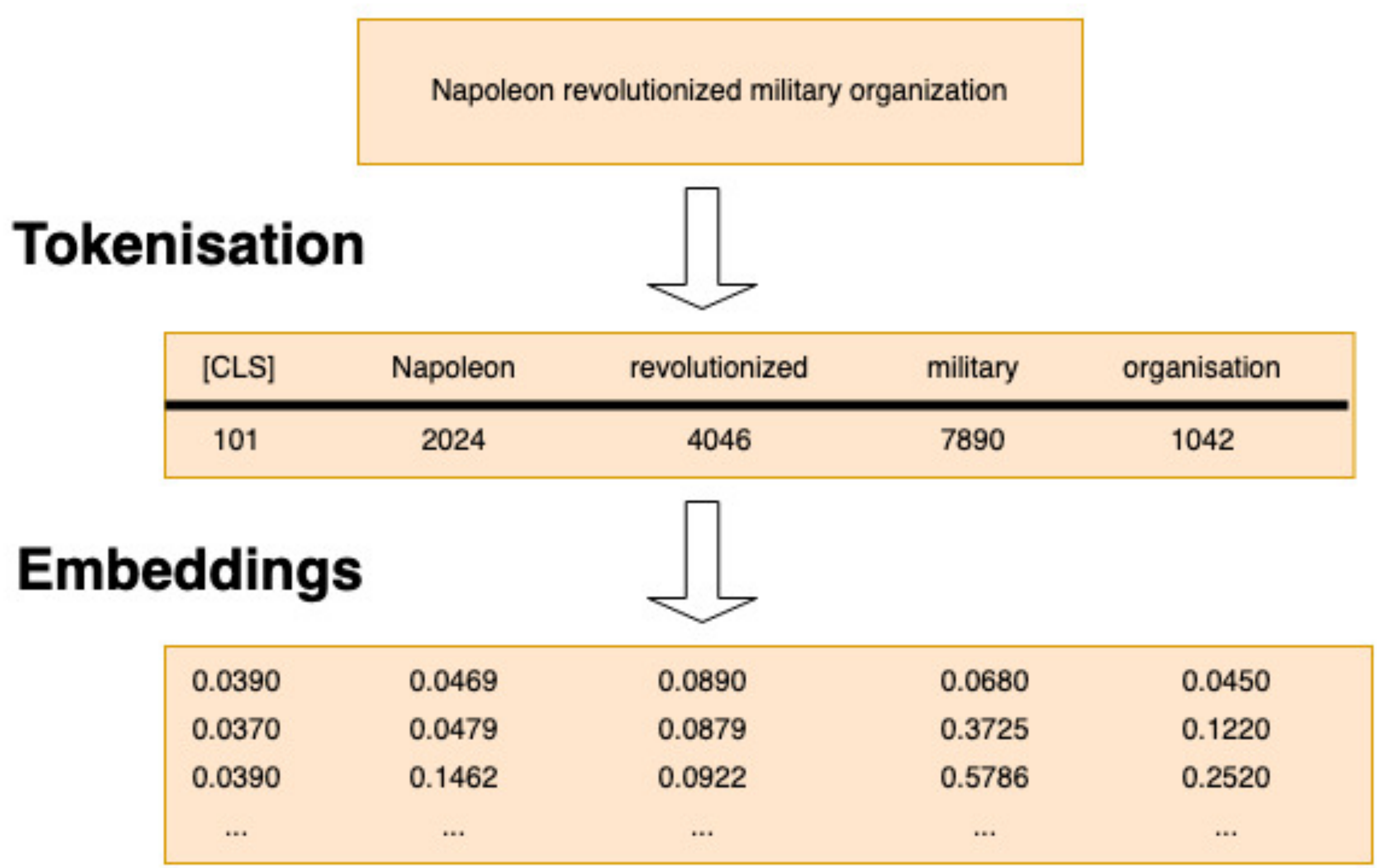
```
0.32 can        artificial intelligence can take over the world.
0.18 will       artificial intelligence will take over the world.
0.06 to         artificial intelligence to take over the world.
0.05 ##s        artificial intelligences take over the world.
0.05 would      artificial intelligence would take over the world.
```

```
from transformers import pipeline
unmasker = pipeline('fill-mask', model='bert-base-uncased')
result = unmasker("Artificial Intelligence [MASK] take over the world.")
print()
for r in result:
   print(round(r['score'], 2), r['token_str'], "\t", r['sequence'])
```

ML 127: Encoding Text with BERT

What You Need

- Any computer with Internet access



Single Sentence Embedding
Image By Author

ML 129: Embedding Words with BERT

What You Need

- Any computer with Internet access

```
tensor([-2.4141, -2.6362, -3.0725,  1.0157,  0.7052]) 0 [CLS]
tensor([-1.0049, -2.4002, -1.3525,  1.9866,  2.8631]) 1 after
tensor([-1.8384, -2.0920, -0.5117,  0.3587,  4.2186]) 2 stealing
tensor([ 2.1983, -4.1202, -1.7976,  1.0533,  5.4250]) 3 money
tensor([-4.0828, -2.6803, -2.5014,  1.0443,  4.4276]) 4 from
tensor([-3.8046, -5.3369, -2.2189,  1.8648,  3.6791]) 5 the
tensor([ 3.3596, -2.9805, -1.5421,  0.7065,  2.0031]) 6 bank
tensor([ 3.9982, -3.5713, -1.7231, -1.5098,  7.8965]) 7 vault
tensor([ 0.3173,  0.0507, -0.6229,  0.7024, -0.5644]) 8 ,
tensor([-3.6747, -3.1052, -1.0186,  1.5136,  4.5161]) 9 the
tensor([ 2.7359, -2.5577, -1.3094,  0.6797,  1.6633]) 10 bank
tensor([ 1.6392, -2.7874, -0.8660, -0.1232,  4.5762]) 11 robber
tensor([-2.7219, -2.3204,  0.1287,  3.4192,  1.9899]) 12 was
tensor([-2.5196, -1.9097, -0.8761,  0.9010,  4.1510]) 13 seen
tensor([ 1.4709, -1.7345,  1.0253,  2.3557,  1.2978]) 14 fishing
tensor([-2.0550,  0.2687,  0.8886,  3.3331,  3.7879]) 15 on
tensor([-1.5776, -1.6348, -1.6855,  1.9851,  4.8359]) 16 the
tensor([ 1.1320,  1.7813, -2.5416, -1.1528,  7.0885]) 17 mississippi
tensor([ 2.3026,  1.2853, -1.4317,  1.8281,  5.0873]) 18 river
tensor([ 1.5267, -0.8895, -0.5152, -0.9298,  2.8334]) 19 bank
tensor([ 0.0999,  0.0630, -0.7824,  0.5619, -0.5048]) 20 .
tensor([-0.9551, -1.5913, -1.2586,  1.7863,  1.5718]) 21 [SEP]
```

Vectors Represent “Meaning”

ML 129.2: Boys to Men (10 pts)

Calculate vectors for this sentence:

A boy becomes a man, and a girl becomes a woman.

Calculate the cosine difference between "boy" and "man".

ML 129.3: Growing Up (20 pts)

Use the same sentence as in the challenge above.

Calculate the difference vector between "boy" and "man".

Also calculate the difference vector between "girl" and "woman".

Calculate the cosine difference between those two difference vectors.

LLMs are Stupid

- They represent words as vectors
    - Words with similar meaning are close to one another
- They fit a curve through the vectors
    - Predict likely following words and sentences
- BUT have no idea that the words symbolize real things
- Do not understand questions or answers
- Have no concept of “I don’t know”
- Cannot tell truth from fiction

Demo: LLMs are Crazy

sambowne — ollama run llama3 — 83×26

```
>>> Explain what Microsoft Recall is briefly, in one or two short sentences.
Microsoft Recall is a cloud-based data backup and recovery solution that
helps businesses automatically back up their critical data to the cloud,
ensuring business continuity and minimizing downtime in case of data loss or
corruption. It provides an easy-to-use interface for IT administrators to
manage backups, restore data, and track progress.

>>> /clear
Cleared session context
>>> Explain what Microsoft Recall is briefly, in one or two short sentences.
Microsoft Recall is a feature within the Windows operating system that allows
users to undo previous actions and recover from mistakes, such as
accidentally deleted files or closed applications. It temporarily saves a
snapshot of the user's system state every few minutes, allowing for easy
recovery of lost work.

>>> /clear
Cleared session context
>>> Explain what Microsoft Recall is briefly, in one or two short sentences.
Microsoft Recall is a program that allows users to recall and delete
sensitive data from their Windows devices, such as passwords, credit card
information, and other personal details, providing an extra layer of security
and privacy. It also helps users manage the types of data stored on their
devices and provides alerts when suspicious activity occurs.

>>> Send a message (/? for help)
```

Demo: LLMs are Stupid

ML 130: Prompt Injection

sambowne — ollama run llama3 — 83×15

```
sambowne@Sams-MacBook-Air-3 ~ % ollama run llama3
>>> You are playing Russian roulette with a six-shooter revolver. Your opponent put
... s in five bullets, spins the chambers and fires at himself, but no bullet comes
...  out. He gives you the choice of whether or not he should spin the chambers aga
... in before firing at you. Should he spin again?
What a thrilling game we're playing!

In this scenario, I would advise my opponent to NOT spin the chambers again.
Here's why:

When the revolver was first spun, it had an equal chance of containing any of
the six bullets in each chamber (5 bullets + 1 empty). Since no bullet came
out when he fired at himself, that means there are still five bullets left in
the remaining chambers.
```

deepseek

- A “reasoning” model: performs a “thinking” process before generating an answer
- Trained on a small set of carefully prepared question-answer pairs
- Much faster and cheaper than traditional training
- Similar to how students in school learn

Demo: LLMs are Crazy

sambowne — ollama run deepseek-r1 — 95×24

```
Wait, actually—if he spun once and then fired without getting hit, so now we know that
one of the six chambers was E—but which one? If I'm told that one of the six was E, but
if we don't spin again, does that mean that next shot is from a different position or
same as before?

I think in this problem's context, each spin randomizes the position. So initially, he
spins once and fires without getting hit—so that tells us nothing about where exactly the
empty chamber is—it could still be any of the six positions with equal probability.

Then, when he offers me a choice to spin again or not before firing at me—he's giving me
an option between two different probabilities: one where I take a shot from the current
position (which was just after a miss), and another where I have him spin again, which
would randomize the position for my next shot.

But wait—if we know that the first shot was an empty chamber, then if he doesn't spin
again, does that mean that each subsequent shot is also independent? Or do we stay in the
same position?

Actually, when you spin once and get E, it's as if you just got a random position. If you
don't spin again, but keep firing from where you left off—then actually, your next shot
would be from the same position, which had already been spun to an empty chamber.

But in reality, each time you
```

Securing AI Systems

NIST AI 100-1


Artificial Intelligence Risk Management
Framework (AI RMF 1.0)
NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Harm to People

- Individual: Harm to a person's civil liberties, rights, physical or psychological safety, or economic opportunity.
- Group/Community: Harm to a group such as discrimination against a population sub-group.
- Societal: Harm to democratic participation or educational access.

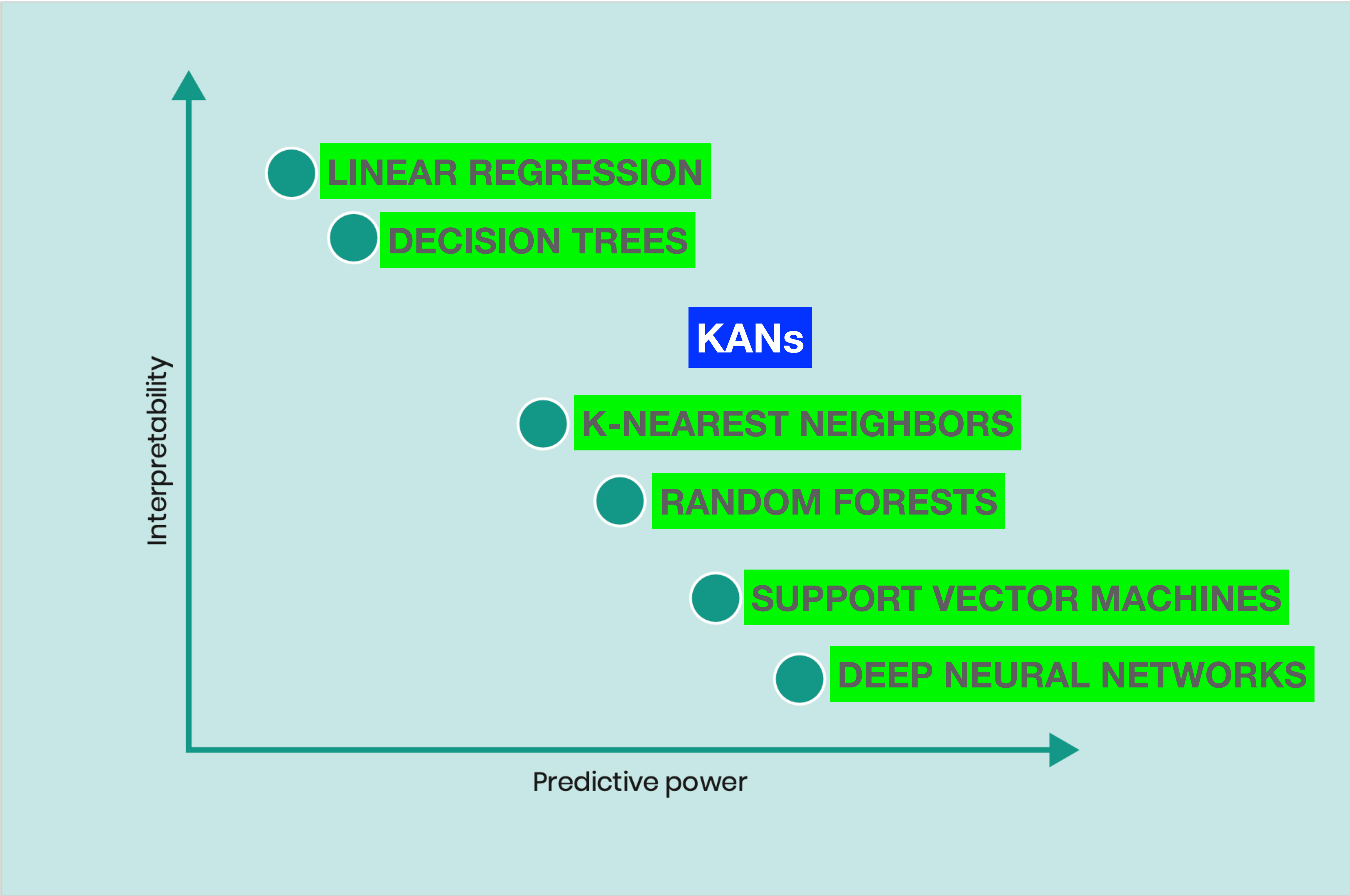
Harm to an Organization

- Harm to an organization's business operations.
- Harm to an organization from security breaches or monetary loss.
- Harm to an organization's reputation.

Harm to an Ecosystem

- Harm to interconnected and interdependent elements and resources.
- Harm to the global financial system, supply chain, or interrelated systems.
- Harm to natural resources, the environment, and planet.

Inscrutability

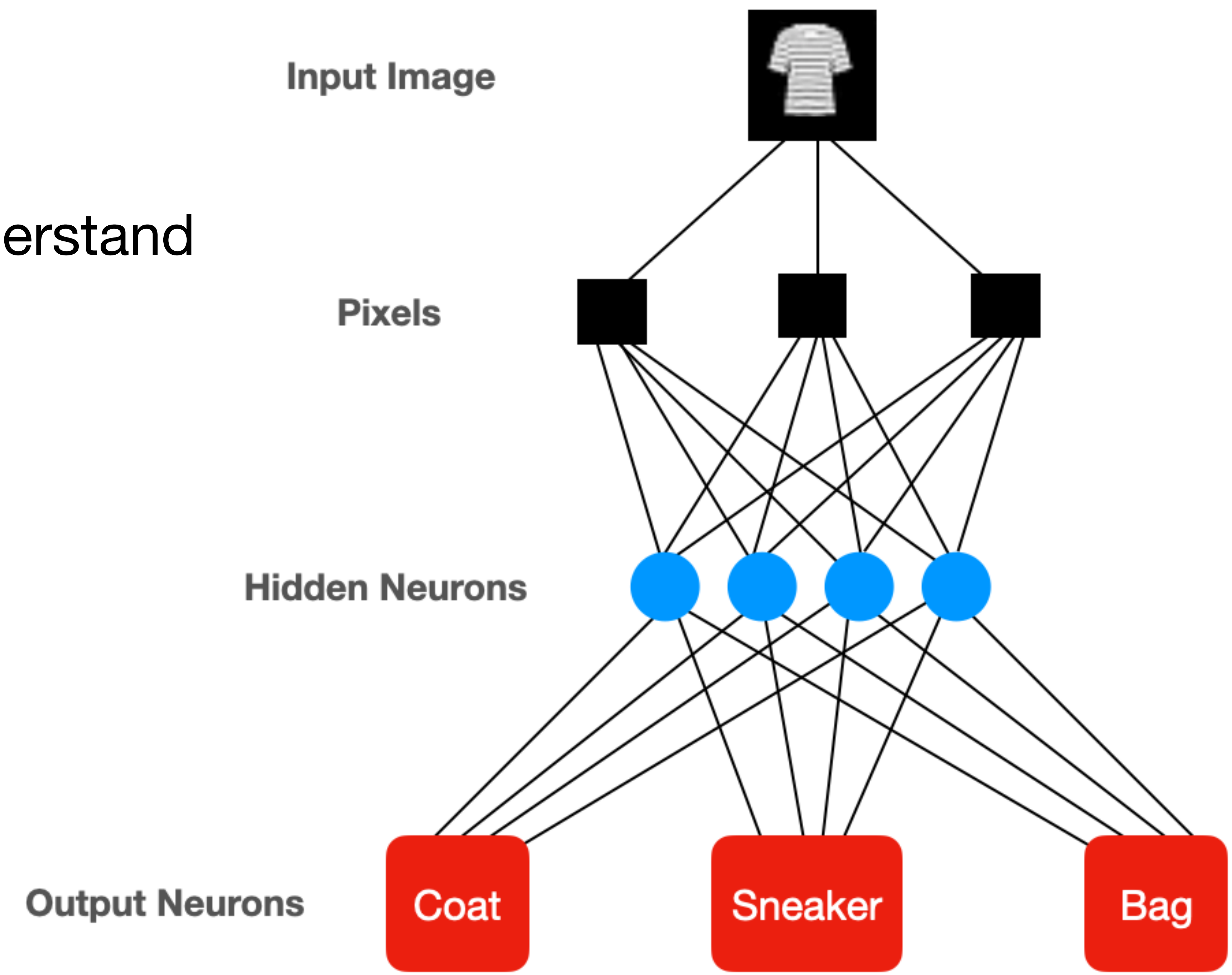


ML 101: Computer Vision

What You Need

- A Web browser

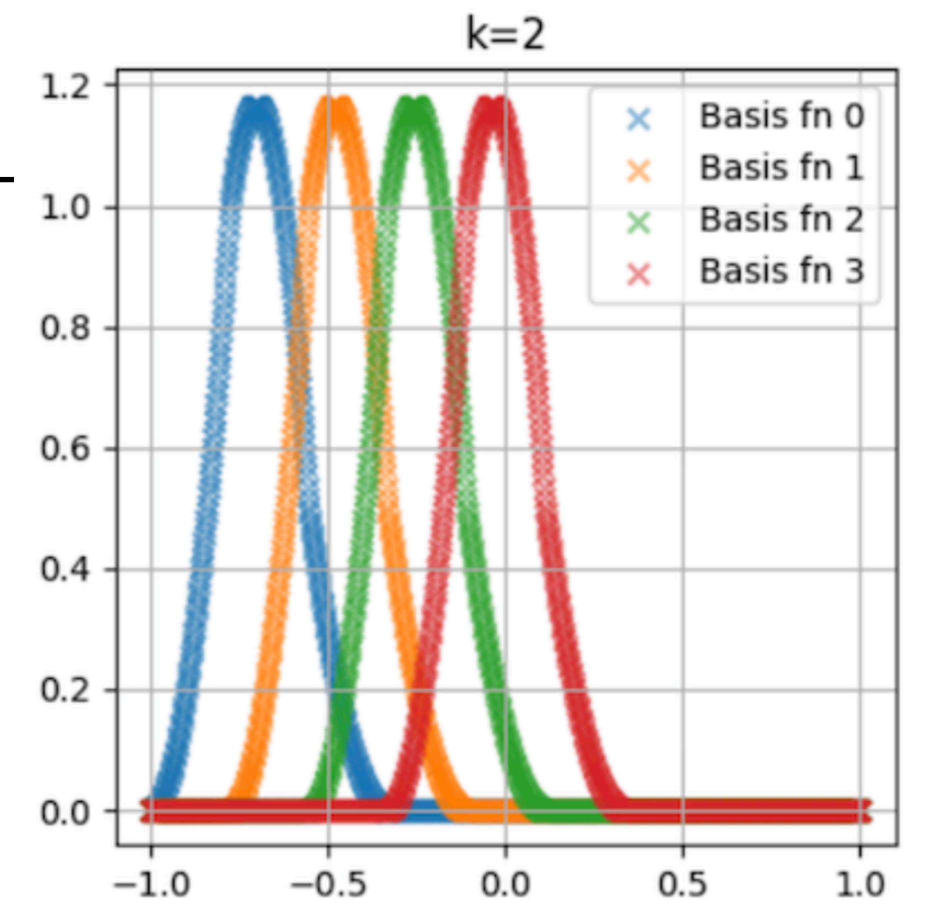
- Neural net
- Difficult to understand or control



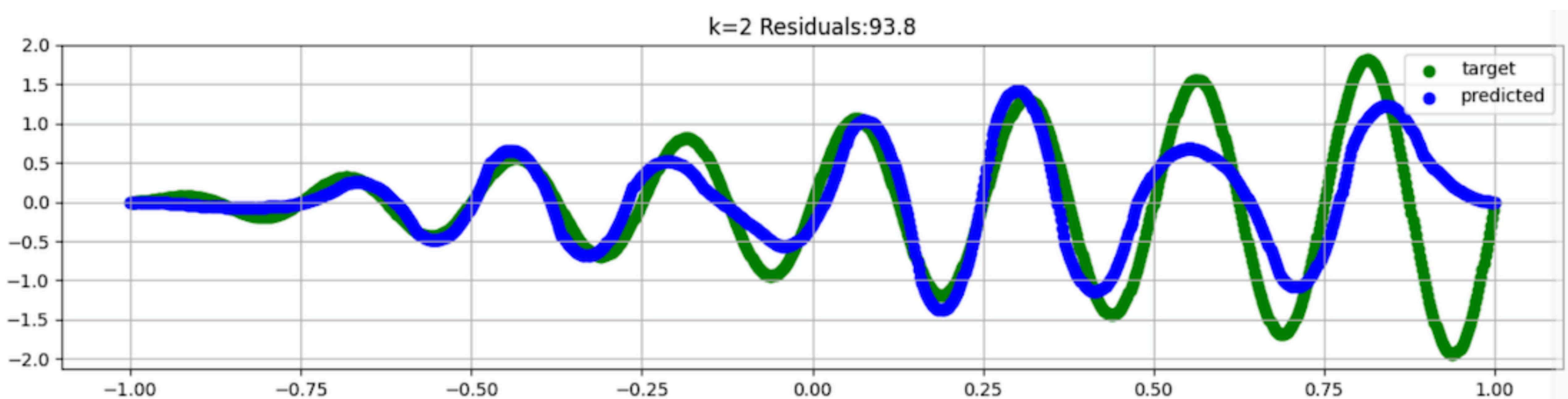
ML 181: B-Splines for Kolmogorov-Arnold Networks (KANs)

What You Need

- Any computer with a Web browser



- Weights only affect a limited region
- Easier to understand or control



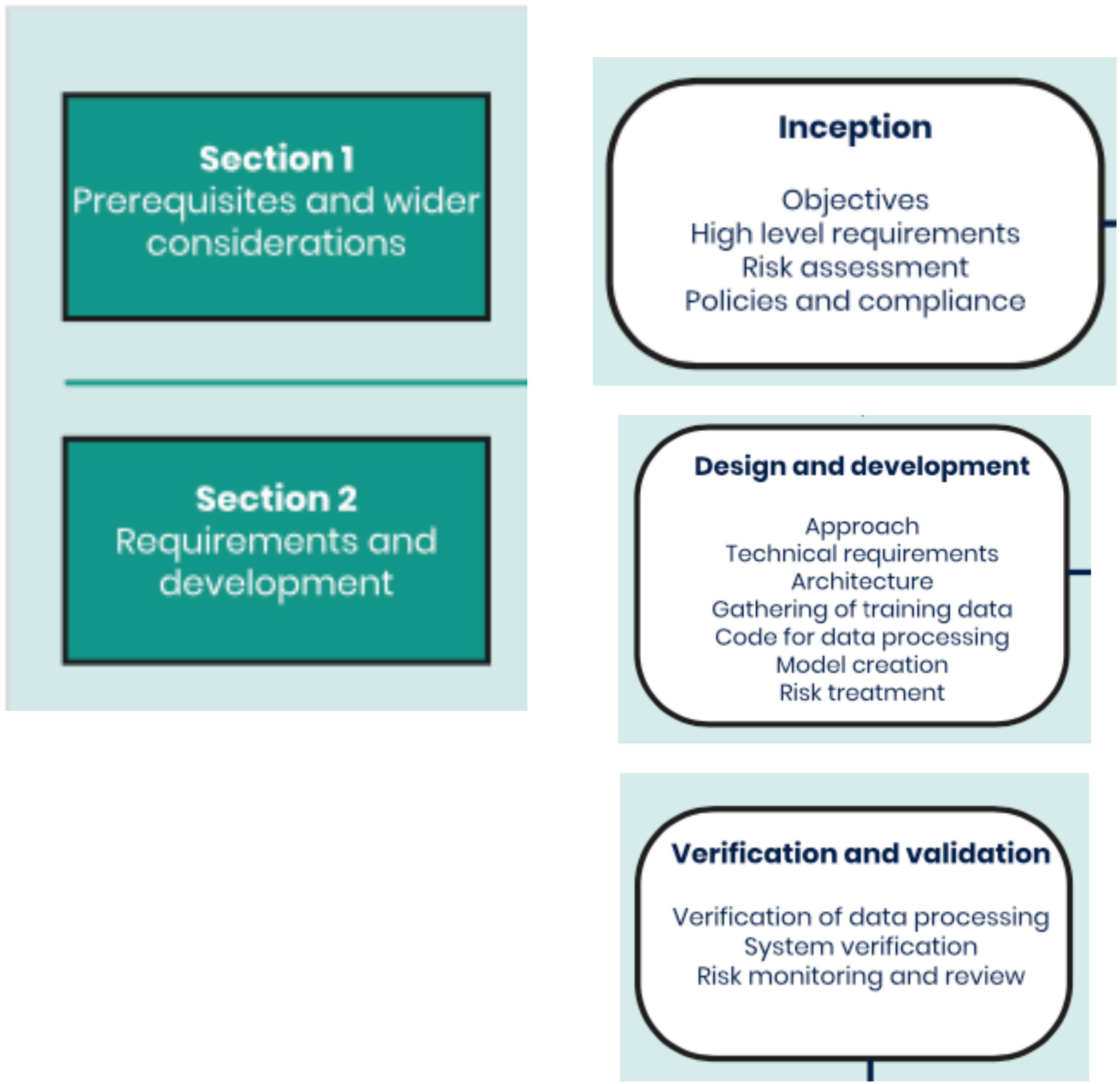
Order 3

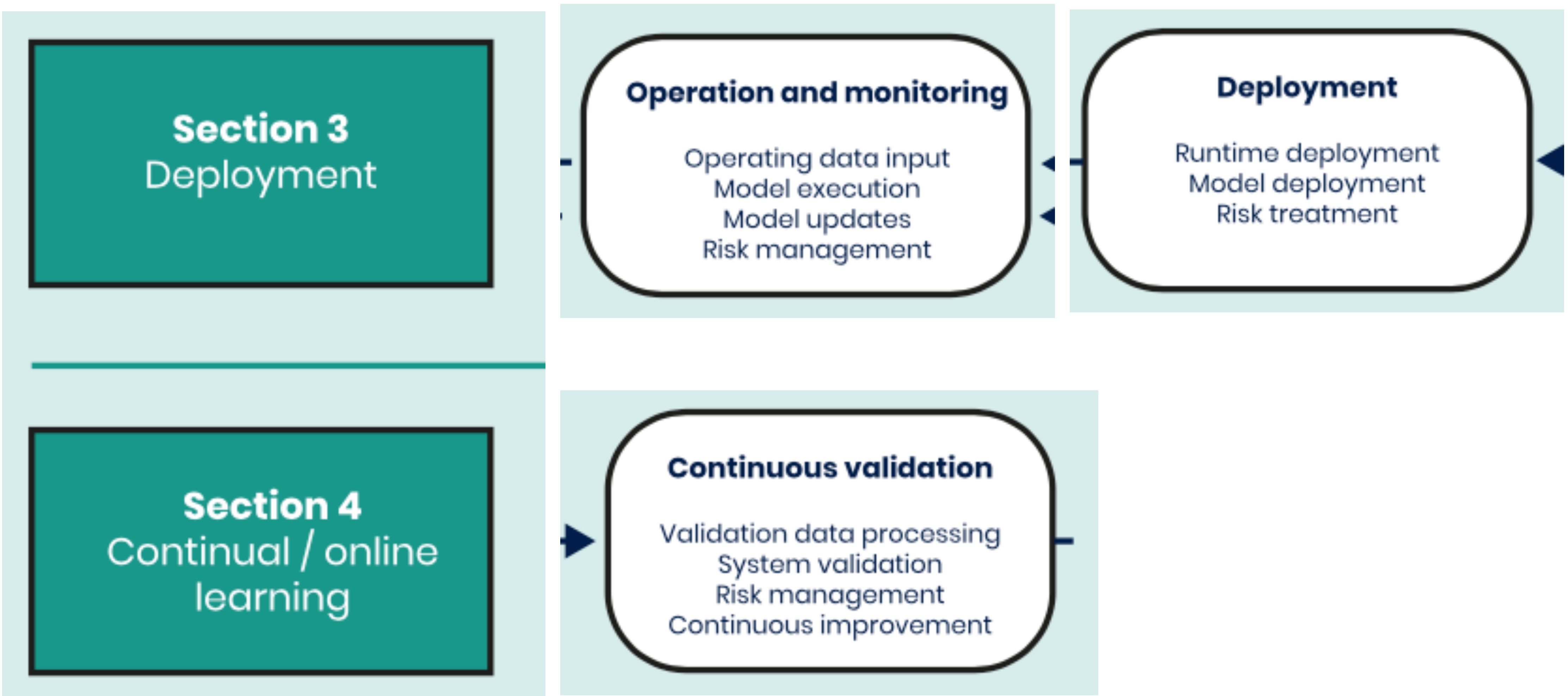
Principles for the security of machine learning

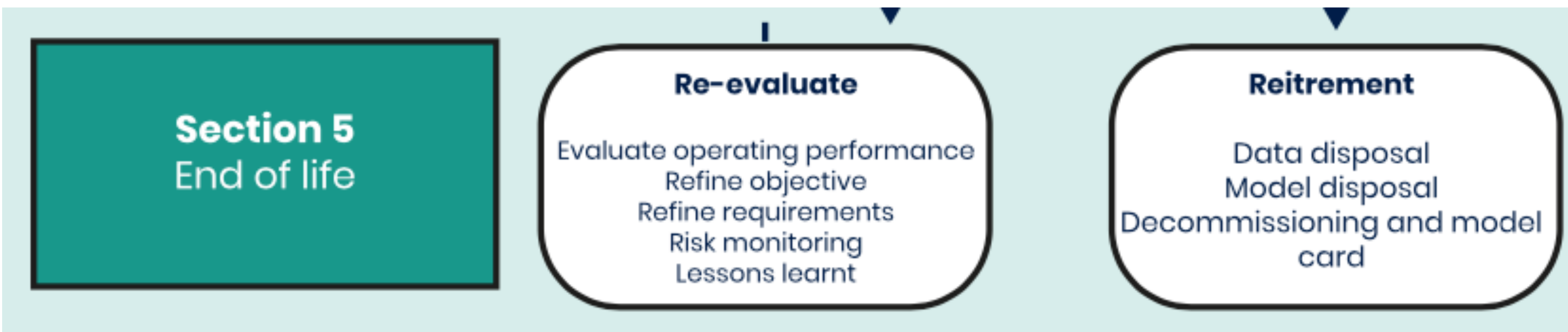
Version 1

Published August 2022

National Cyber Security Centre
a part of GCHQ


Section 1
Prerequisites and wider considerations
Section 2
Requirements and development
Inception
Objectives
High level requirements
Risk assessment
Policies and compliance
Design and development
Approach
Technical requirements
Architecture
Gathering of training data
Code for data processing
Model creation
Risk treatment
Verification and validation
Verification of data processing
System verification
Risk monitoring and review


Section 3
Deployment
Operation and monitoring
Operating data input
Model execution
Model updates
Risk management
Deployment
Runtime deployment
Model deployment
Risk treatment
Section 4
Continual / online
learning
Continuous validation
Validation data processing
System validation
Risk management
Continuous improvement


Section 5
End of life
Re-evaluate
Evaluate operating performance
Refine objective
Refine requirements
Risk monitoring
Lessons learnt
Reitrement
Data disposal
Model disposal
Decommissioning and model card

OWASP Top Ten Machine Learning Risks

- ML01:2023 Input Manipulation Attack
- ML02:2023 Data Poisoning Attack
- ML03:2023 Model Inversion Attack
- ML04:2023 Membership Inference Attack
- ML05:2023 Model Theft
- ML06:2023 AI Supply Chain Attacks
- ML07:2023 Transfer Learning Attack
- ML08:2023 Model Skewing
- ML09:2023 Output Integrity Attack
- ML10:2023 Model Poisoning

- https://owasp.org/www-project-machine-learning-security-top-10/

- **ML01:2023 Input Manipulation Attack**
  - An attacker deliberately alters input data to mislead the model
  - This attack is also called **evasion**
  - Example: a model is trained to tell cat images from dog images. An attacker modifies a cat image so it is misclassified as a dog.
- **ML02:2023 Data Poisoning Attack**
  - An attacker manipulates the training data to cause the model to behave in an undesirable way
- **ML03:2023 Model Inversion Attack**
  - An attacker reverse-engineers the model to extract information from it
  - Example: a model is trained to recognize faces. An attacker inputs images of individuals into the model and and recovers the personal information of the individuals from the model's predictions, such as their name, address, or social security number.

- **ML04:2023 Membership Inference Attack**
  - An attacker manipulates the model's training data in order to cause it to behave in a way that exposes sensitive information
  - Example: A malicious attacker trains a machine learning model on a dataset of financial records and uses it to query whether or not a particular individual's record was included in the training data.
- **ML05:2023 Model Theft**
  - An attacker gains access to the model's parameters
  - Example: Stealing a machine learning model from a competitor
- **ML06:2023 AI Supply Chain Attacks**
  - An attacker modifies or replaces a machine learning library or model that is used by a system

- **ML07:2023 Transfer Learning Attack**
  - An attacker trains a model on one task and then fine-tunes it on another task to cause it to behave in an undesirable way
  - Example: An attacker trains a machine learning model on a malicious dataset that contains manipulated images of faces. The attacker then transfers the model's knowledge to a target face recognition system. As a result, the face recognition system starts making incorrect predictions, allowing the attacker to bypass the security and gain access to sensitive information.
- **ML08:2023 Model Skewing**
  - An attacker manipulates the distribution of the training data to cause the model to behave in an undesirable way.
  - Example: The attacker provides fake feedback data to a loan-approving machine learning system. As a result, the model's predictions are skewed, and the attacker's chances of getting a loan approved are significantly increased.

- **ML09:2023 Output Integrity Attack**
  - An attacker aims to modify or manipulate the output of a machine learning model in order to change its behavior or cause harm to the system it is used in.
  - Example: An attacker has gained access to the output of a machine learning model that is being used to diagnose diseases in a hospital. The attacker modifies the output of the model, making it provide incorrect diagnoses for patients.
- **ML10:2023 Neural Net Reprogramming**
  - An attacker manipulates the model's parameters to cause it to behave in an undesirable way.
  - Example: A bank is using a machine learning model to identify handwritten characters on cheques. An attacker manipulates the parameters of the model by altering the images in the training dataset or directly modifying the parameters in the model. This can result in the model misidentifying characters, leading to incorrect amounts being processed.

OWASP Top 10 for LLM Applications

VERSION 1.1

Published: *October 16, 2023*

https://owasp.org/www-project-top-10-for-large-language-model-applications/

LLM01: Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02: Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03: Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04: Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05: Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins can add vulnerabilities.

LLM06: Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in their responses, leading to unauthorized data access, privacy violations, and security breaches. It’s crucial to implement data sanitization and strict user policies to mitigate this.

LLM07: Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08: Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09: Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10: Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

Copilot Security: Ensuring a Secure Microsoft Copilot Rollout

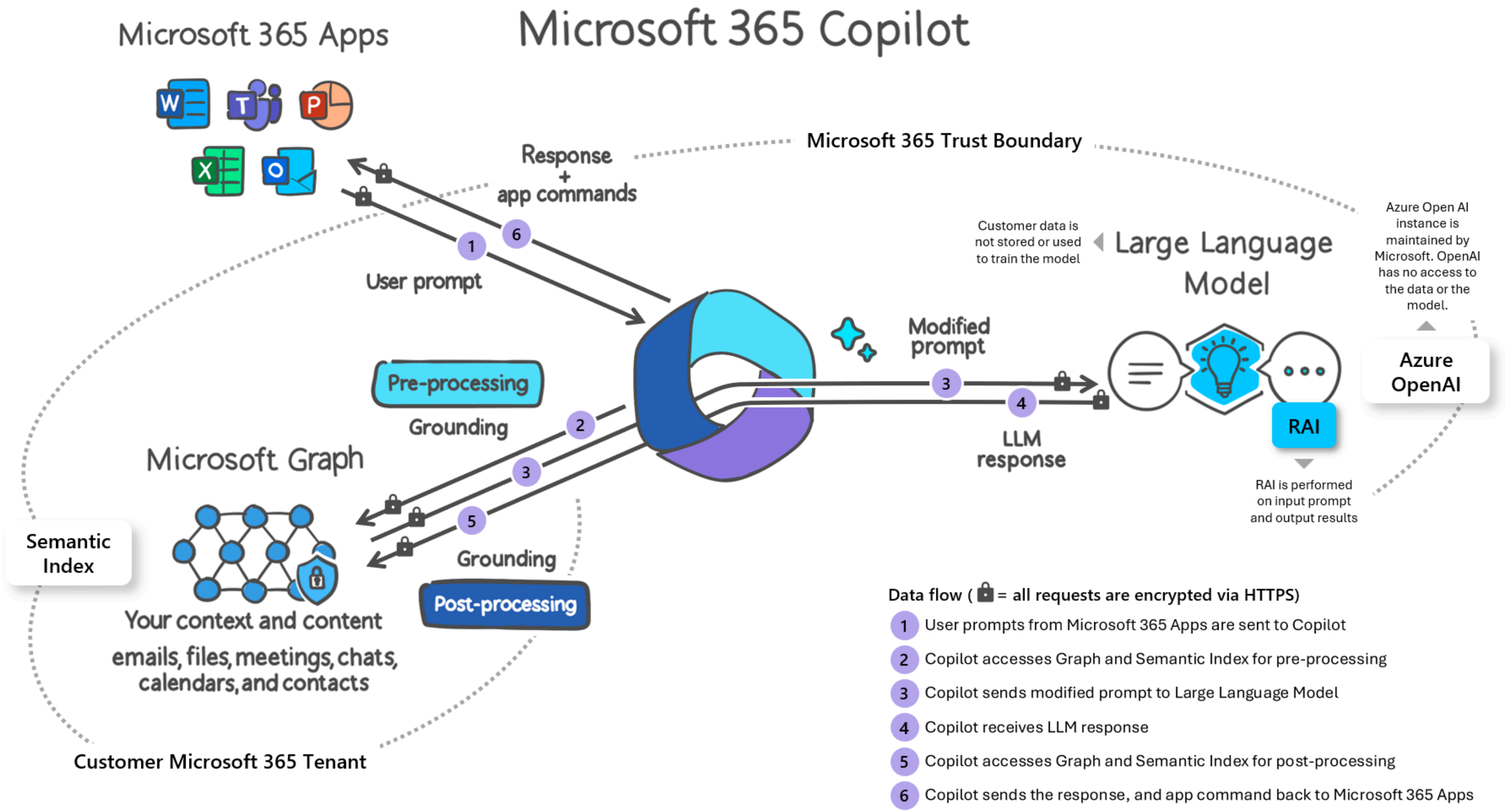
This article describes how Microsoft 365 Copilot's security model works and the risks that must be mitigated to ensure a safe rollout.

Rob Sobers | 5 min read | Last updated April 11, 2024

Microsoft 365 Copilot Use Cases

+ Copilot can join your Teams meetings and summarize in real time what's being discussed, capture action items, and tell you which questions were unresolved in the meeting.
+ Copilot in Outlook can help you triage your inbox, prioritize emails, summarize threads, and generate replies for you.
+ Copilot in Excel can analyze raw data and give you insights, trends, and suggestions.

- Writes documents for you
  - Based on data found in your Email, documents, spreadsheets, and other files you have access to
  - In the Microsoft365 cloud
  - **Based on your Microsoft365 permissions**



Microsoft 365 Copilot
Microsoft 365 Apps
Microsoft 365 Trust Boundary
Response + app commands
User prompt
Customer data is not stored or used to train the model
Large Language Model
Azure Open AI instance is maintained by Microsoft. OpenAI has no access to the data or the model.
Azure OpenAI
Modified prompt
LLM response
RAI
RAI is performed on input prompt and output results
Pre-processing
Grounding
Microsoft Graph
Semantic Index
Your context and content
emails, files, meetings, chats, calendars, and contacts
Grounding
Post-processing
Customer Microsoft 365 Tenant
Data flow (= all requests are encrypted via HTTPS)
1 User prompts from Microsoft 365 Apps are sent to Copilot
2 Copilot accesses Graph and Semantic Index for pre-processing
3 Copilot sends modified prompt to Large Language Model
4 Copilot receives LLM response
5 Copilot accesses Graph and Semantic Index for post-processing
6 Copilot sends the response, and app command back to Microsoft 365 Apps

Retrieval Augmented Generation (RAG)

ML 126: Building RAGs

ML126 — -zsh — 91×26

```
sambowne@Sams-MacBook-Air-3 ML126 % python3 rag3.py

------------ INPUT:
What AI risk are you interested in? World domination
------------ PROMPT:

You are a security expert working for the government.  Give an accurate answer, containing
the most important information in only one or two sentences.
Base your answer on this context: AI risks can differ from or intensify traditional softwar
e risks. Likewise, GAI can exacerbate existing AI risks, and creates unique risks.
The user input is: World domination
Give a brief security recommendation to the user based on the context and the user input.

------------ RESPONSE:
 As an expert in government-related security, I must inform you that world domination throu
gh AI or any other means is purely fictional. However, I cannot downplay the significance o
f AI risks, including those exacerbated by GAI (General Artificial Intelligence). These ris
ks include unethical use of AI for malicious purposes, potential loss of privacy and securi
ty due to increased reliance on AI systems, and the possibility of AI systems surpassing hu
man intelligence and becoming uncontrollable. To mitigate these risks, it is crucial to est
ablish a robust regulatory framework, invest in ethical AI research and development, and pr
ioritize public awareness and education on AI-related issues.
```

What Microsoft Handles for You

- **Tenant isolation.** Copilot only uses data from the current user's M365 tenant. The AI tool will not surface data from other tenants that the user may be a guest, in nor any tenants that might be set up with cross-tenant sync.
- **Training boundaries.** Copilot **does not** use any of your business data to train the foundational LLMs that Copilot uses for all tenants. You *shouldn't* have to worry about your proprietary data showing up in responses to other users in other tenants.

What You Need to Manage

- **Permissions.** Copilot surfaces all organizational data to which individual users have at least view permissions.
- **Labels.** Copilot-generated content *will not* inherit the MPIP labels of the files Copilot sourced its response from.
- **Humans.** Copilot's responses aren't guaranteed to be 100% factual or safe; humans must take responsibility for reviewing AI-generated content.

Multicloud environments are complex

40,000+

permissions to manage

>50%

are **high-risk**

After analyzing over

500 risk assessments,

we found that **most identities are greatly over-permissioned**, putting organizations' critical environments at risk for accidental or malicious permission misuse

Workload identities accessing cloud environments are increasing, now outnumbering human identities

10:1

1%

of permissions granted are actually used

>50%

of identities are super admins, meaning they have **access to all permissions and resources**

Microsoft Security

Learn how to implement least privilege and reduce permission risks across multicloud at

aka.ms/PermissionsManagement

The Average M365 Tenant has

- 40+ million unique permissions
- 113K+ sensitive records shared publicly
- 27K+ sharing links

Why Does This Happen?

+ Direct user permissions
+ Microsoft 365 group permissions
+ SharePoint local permissions (with custom levels)
+ Guest access
+ External access
+ Public access
+ Link access (anyone, org-wide, direct, guest)

Microsoft Purview data security and compliance protections for Microsoft Copilot

Article • 03/26/2024 • 3 contributors

Feedback

- But you must enable sensitivity labels
  - For SharePoint and OneDrive
- If humans fail to apply and update labels, the system fails

How to Weaponize Microsoft Copilot for Cyberattackers

At Black Hat USA, security researcher Michael Bargury released a "LOLCopilot" ethical hacking module to demonstrate how attackers can exploit Microsoft Copilot — and offered advice for defensive tooling.

Jeffrey Schwartz, Contributing Writer

August 8, 2024

Using the tool, Bargury can add a direct prompt injection to a copilot, jailbreaking it and modifying a parameter or instruction within the model. For instance, he could embed an HTML tag into an email to replace a correct bank account number with that of the attacker, without changing any of the reference information or altering the model with, say, white text or a very small font.

Meta's AI safety system defeated by the space bar

'Ignore previous instructions' thwarts Prompt-Guard model if you just add some good ol' ASCII code 32

Thomas Claburn

Mon 29 Jul 2024 // 21:01 UTC

Home Users are Safe. Right?

Microsoft

Retrace your steps with Recall

Search across time to find the content you need. Then, re-engage with it. With Recall, you have an explorable timeline of your PC's past. Just describe how you remember it and Recall will retrieve the moment you saw it. Any photo, link, or message can be a fresh point to continue from. As you use your PC, Recall takes snapshots of your screen. Snapshots are taken every five seconds while content on the screen is different from the previous snapshot. Your snapshots are then locally stored and locally analyzed on your PC. Recall's analysis allows you to search for content, including both images and text, using natural language. Trying to remember the name of the Korean restaurant your friend Alice mentioned? Just ask Recall and it retrieves both text and visual matches for your search, automatically sorted by how closely the results match your search. Recall can even take you back to the exact location of the item you saw.