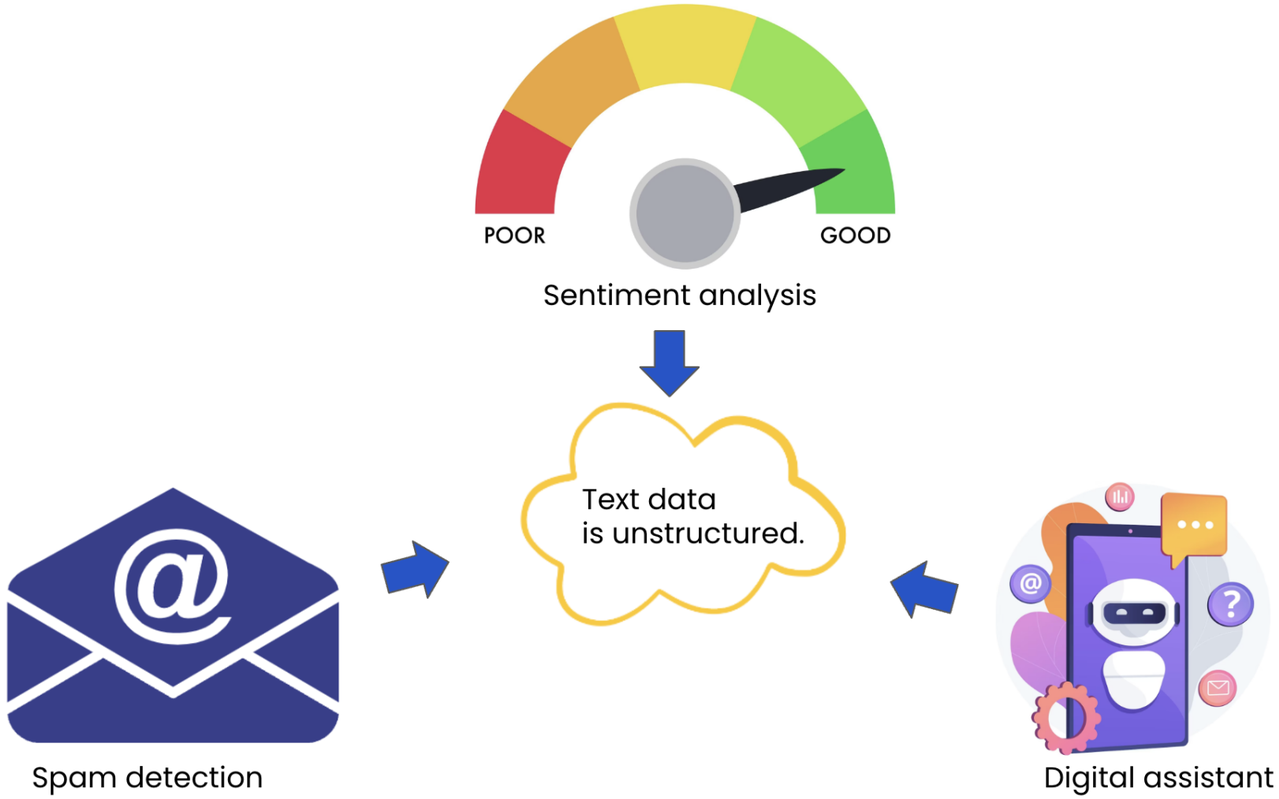
# Novelty of LLMs

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

**Vidhi Chugh**
AI strategist and ethicist

# Using text data




- Unstructured data - messy and inconsistent

[1] Freepik

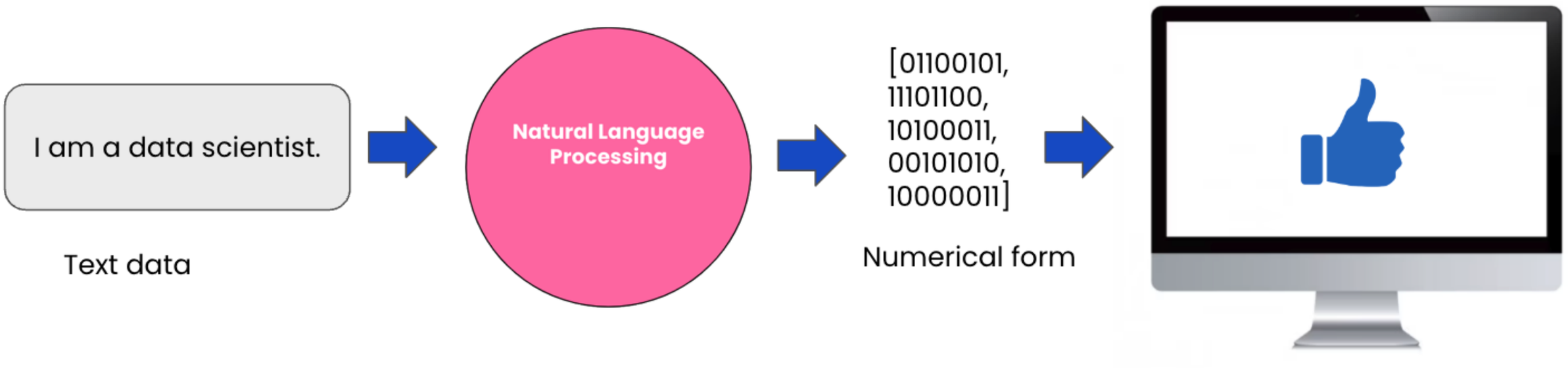
# Machines do not understand language!

I am a data scientist.

Text data

[1] Freepik

# Need for NLP




[1] Freepik

# Unique capabilities of LLMs

- **Linguistic subtleties**
  - Irony
  - Humor
  - Pun
  - Sarcasm
  - Intonation
  - Intent

[1] Freepik

# What's your favorite book?

- **Natural response:** "Oh, that's a tough one!"
- **Personal opinion:** "My all-time favorite book is To Kill a Mockingbird by Harper Lee."
- **Supporting statement:** "It's a powerful story about prejudice, justice, and the human experience."
- **Follow-up question:** "Have you read it?"

# Linguistic subtleties

**Sarcasm:** "Oh great, another meeting."

**Traditional language model:**

- Response: "What's the meeting about?"
  - Neutral
  - Does not pick up sarcasm

**Large language model:**

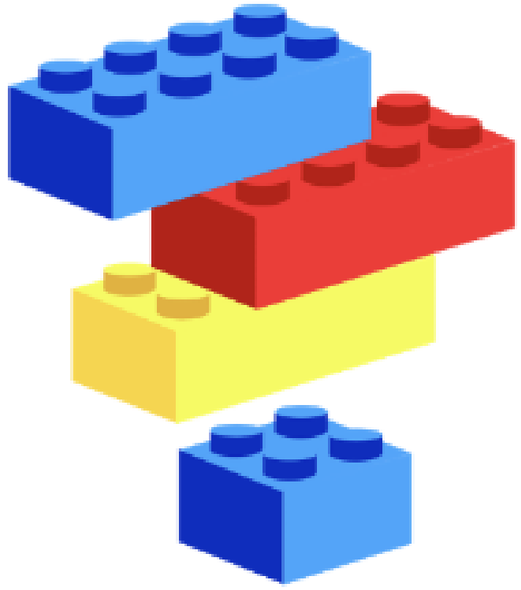
- Response: "Sounds like you're looking forward to it!"
  - Playful
  - Engaging
  - Matches the sarcasm

# How do LLMs understand

- Trained on vast amounts of data
- Largeness of LLMs: parameters
- Parameters represent the patterns and rules
- More parameters -> complex patterns
- Generates sophisticated and accurate responses

# Parameters

- Small number of bricks -> limited structures

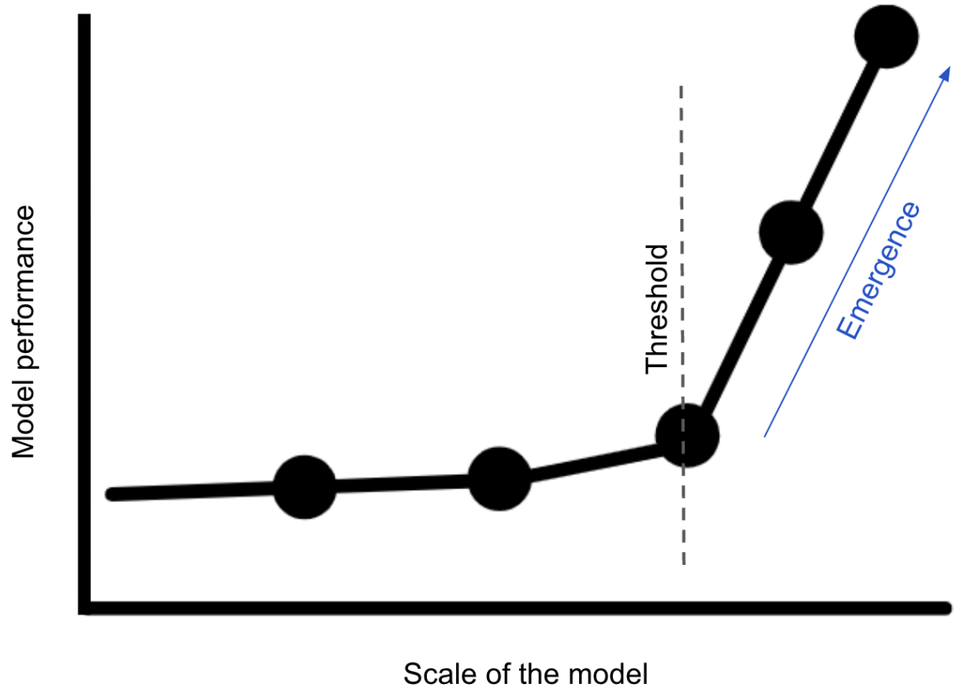


- Larger number of bricks -> complex and detailed structures

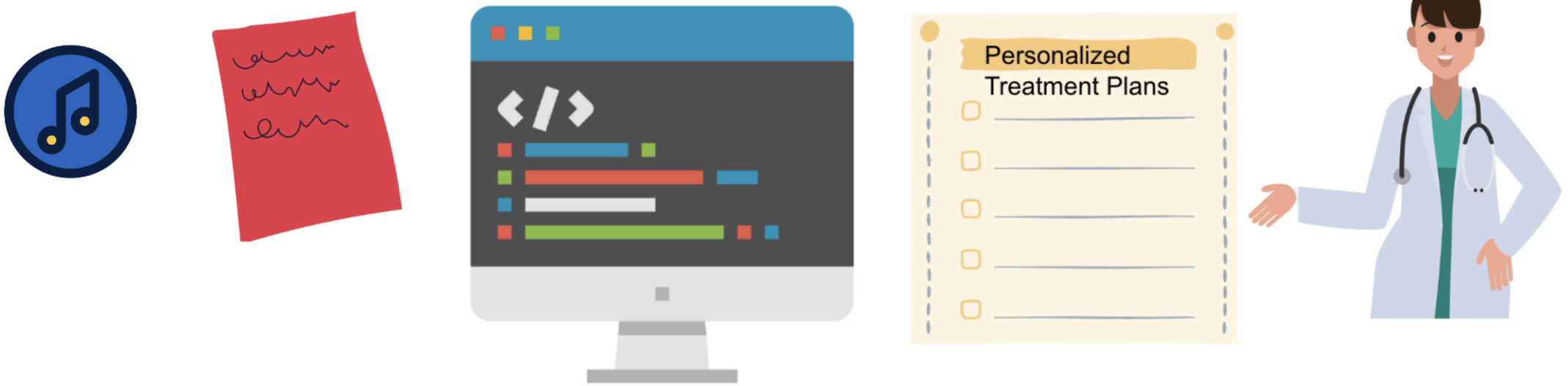
# Emergence of new capabilities

- Emergent abilities
  - only present in large-scale models

- Scale:
  - The volume of training data
  - The number of model parameters



# Emergence of new capabilities



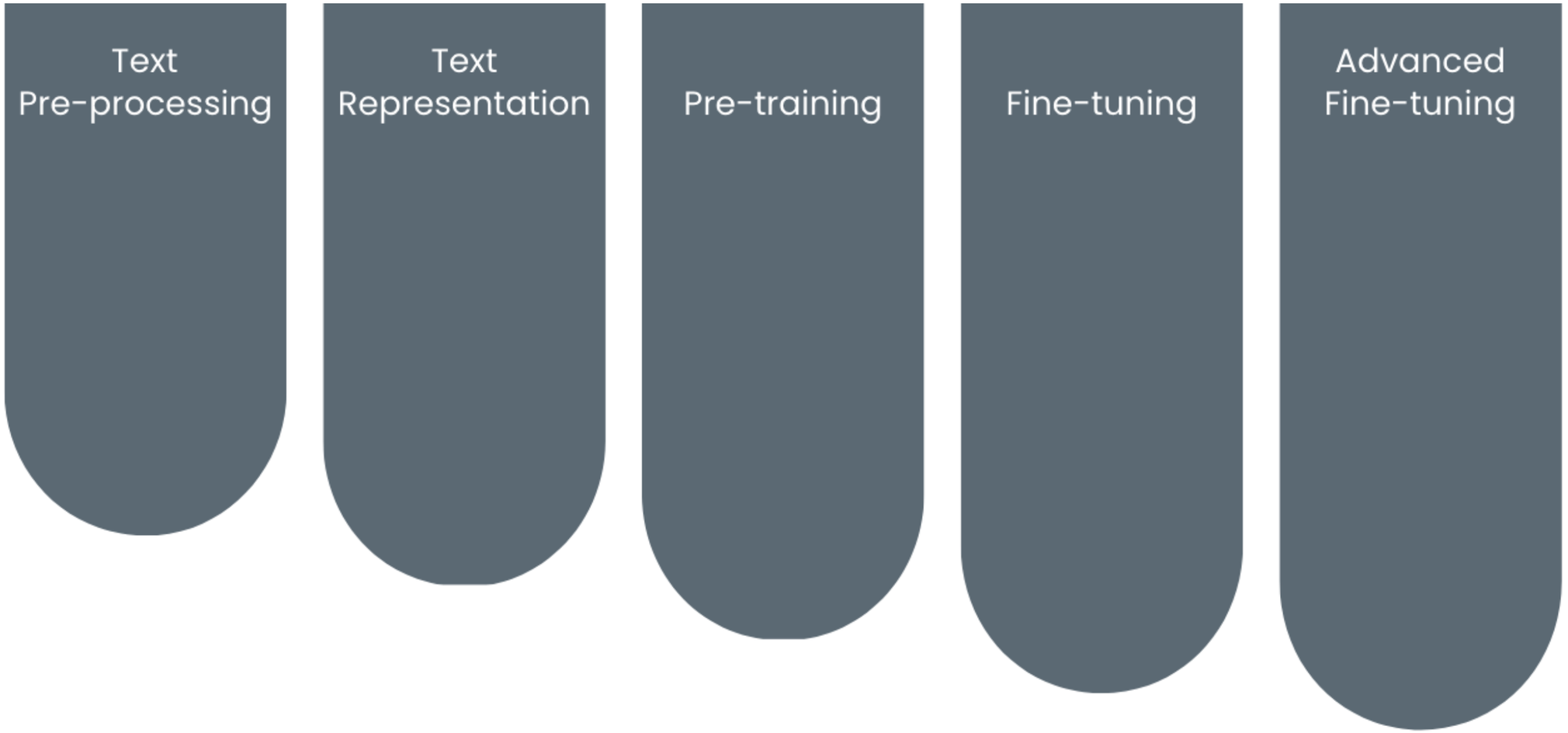

Music

Poetry

Code generation

Medical diagnosis and treatment plans

# Building blocks of LLMs



# To recap

**LLMs:**

- Overcome data's unstructured nature
- Outperform traditional models
- Understand linguistic subtleties

**How?**

- LLMs' "largeness"
- Extensive training data
- Many parameters
- Emergent abilities

# Let's practice!

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

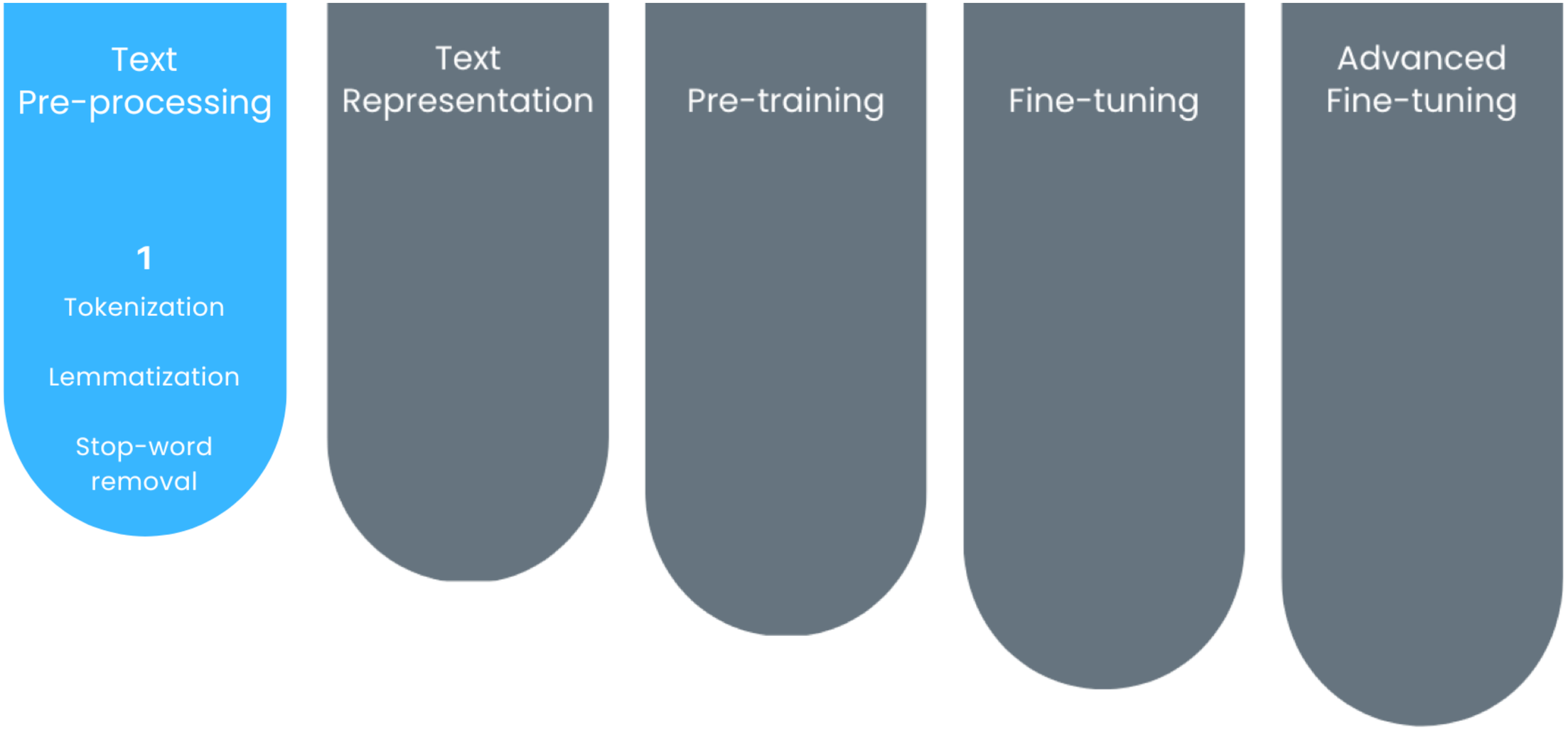
# Generalized overview of NLP

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

**Vidhi Chugh**
AI strategist and ethicist

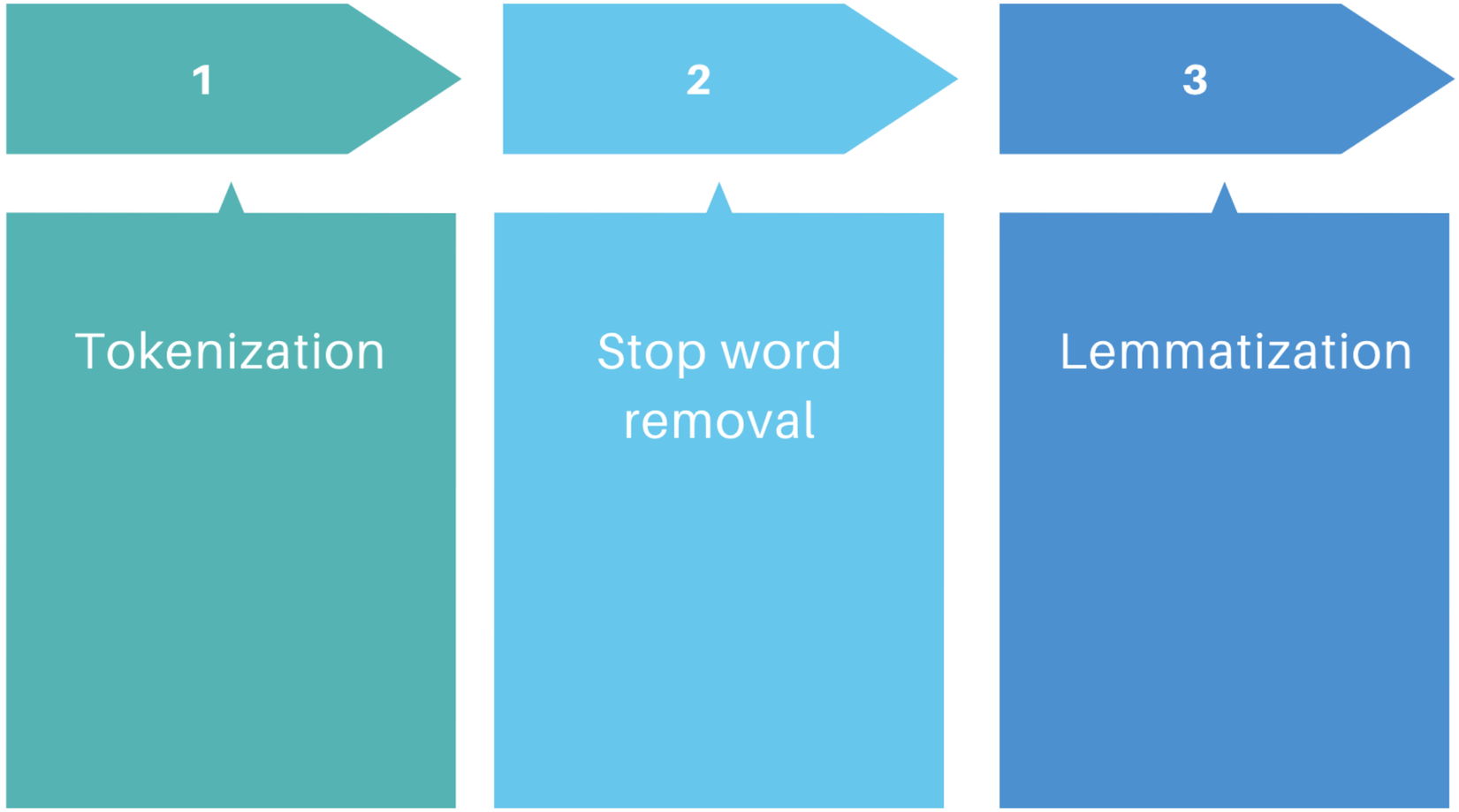
datacamp

# Where are we?


Text Pre-processing
1
Tokenization
Lemmatization
Stop-word removal
Text Representation
Pre-training
Fine-tuning
Advanced Fine-tuning

# Text pre-processing

- Can be done in a different order as they are independent



# Tokenization

- Splits text into individual words, or **tokens**

- Text:
  - "Working with natural language processing techniques is tricky."

- Tokenization:
  - ["Working", "with", "natural", "language", "processing", "techniques", "is", "tricky", "."]
  - Converts into a list

# Stop word removal

- Stop words do not add meaning
- Eliminated through stop word removal

- **Before** stop word removal:
  - ["Working", "with", "natural", "language", "processing", "techniques", "is", "challenging", "."]

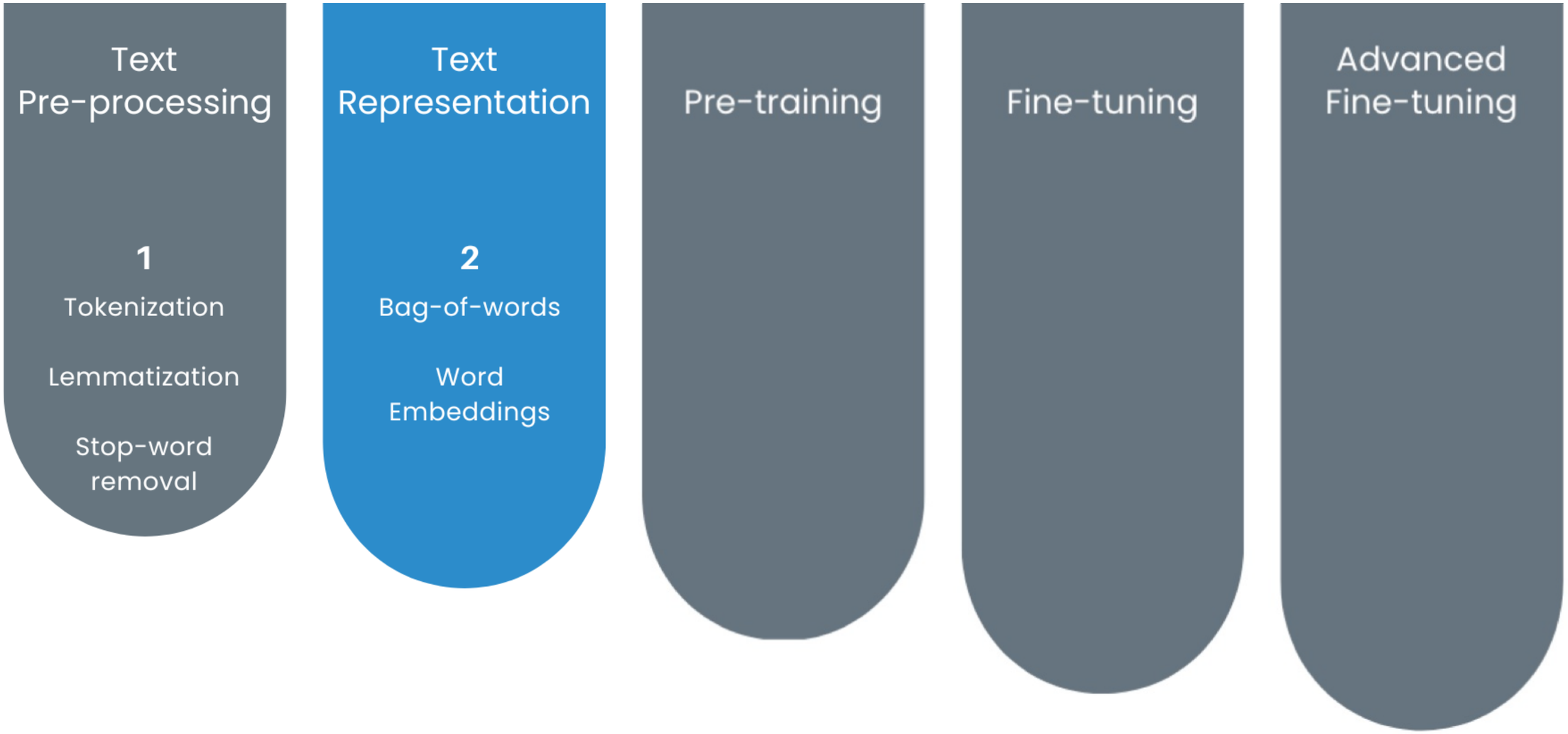
- **After** stop word removal:
  - ["Working", "natural", "language", "processing", "techniques", "challenging", "."]

# Lemmatization

- Group slightly different words with similar meaning
- Reduces words to their base form
- Mapped to root word

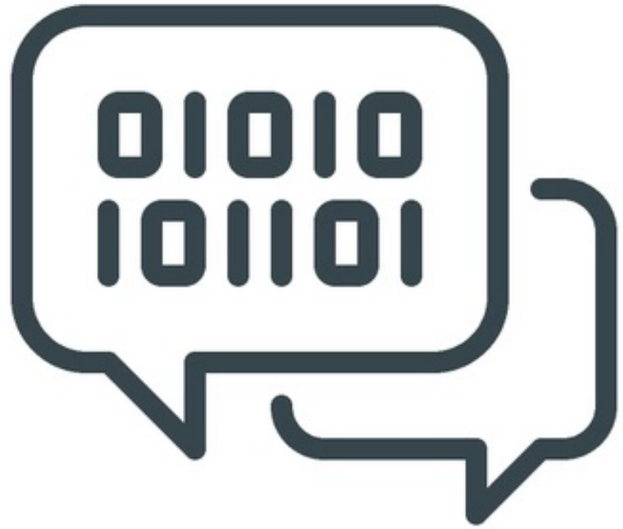
- Talking -> Talk
- Talked -> Talk
- Talk -> Talk

# Text representation



# Text representation

- Text data into numerical form



- Bag-of-words
- Word embeddings

# Bag-of-words

- Text into a matrix of word counts

| Sentence | “cat” | “chased” | “mouse” | “swiftly” |
|---|---|---|---|---|
| "The cat chased the mouse swiftly." | 1 | 1 | 1 | 1 |
| "The mouse chased the cat." | 1 | 1 | 1 | 0 |

- 0 represents the absence of a word

# Limitations of bag-of-words

- Does not capture the order or context
  - Can lead to incorrect interpretations
  - Similar sentences but opposite meaning
    - "The cat chased the mouse swiftly."
    - "The mouse chased the cat."
- Does not capture the semantics between the words
  - Treats related words as independent
  - Like "cat" and "mouse"

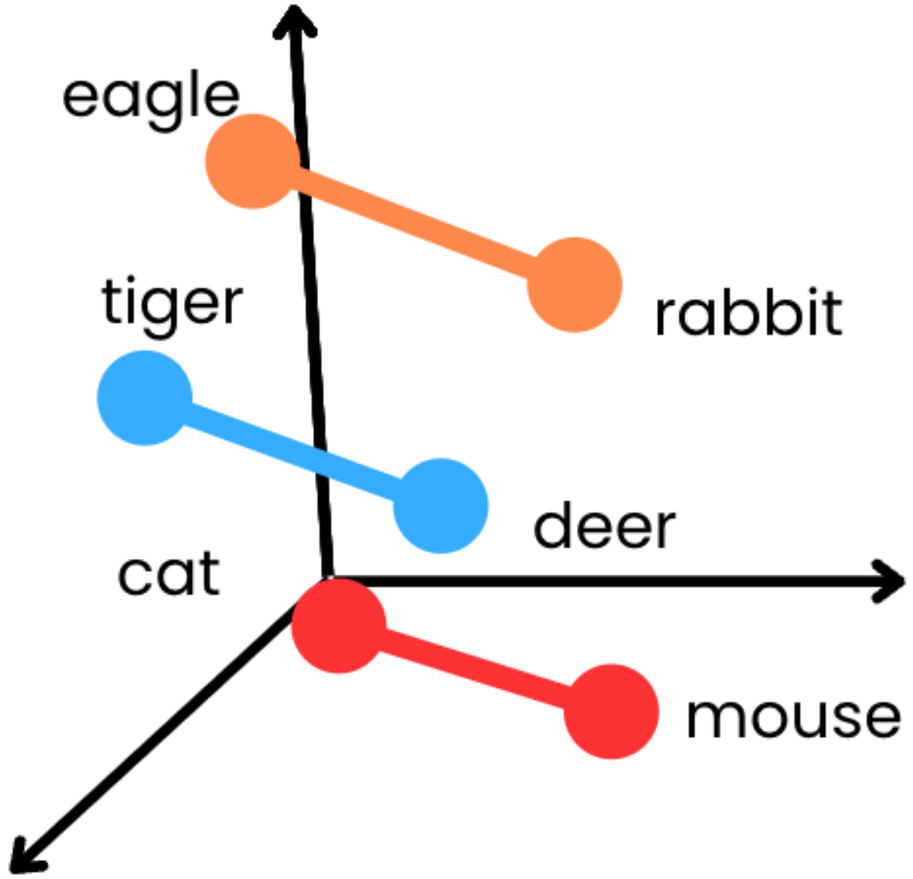
# Word embeddings

- Capture the semantic meanings as numbers

| | Cat | Mouse |
|---|---|---|
| Plant | -0.9 | -0.8 |
| Furry | 0.9 | 0.7 |
| Carnivore | 0.9 | -0.8 |

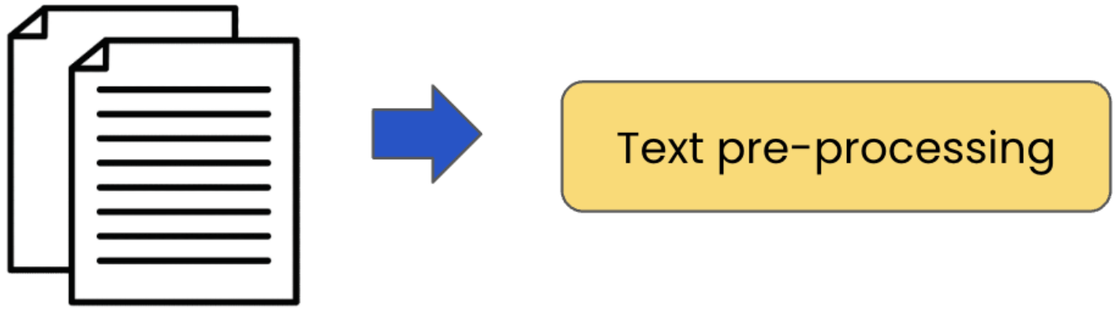
- Cat [-0.9, 0.9, 0.9]

- Predator-prey relationship:



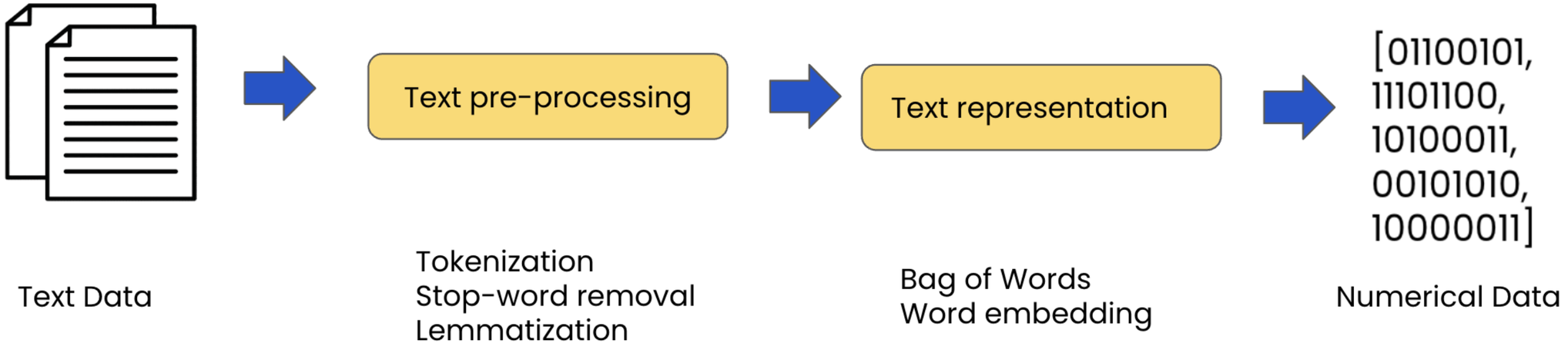
# Machine-readable form

- Start with text pre-processing



# Machine-readable form

- Convert pre-processed text to numerical format



# Let's practice!

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

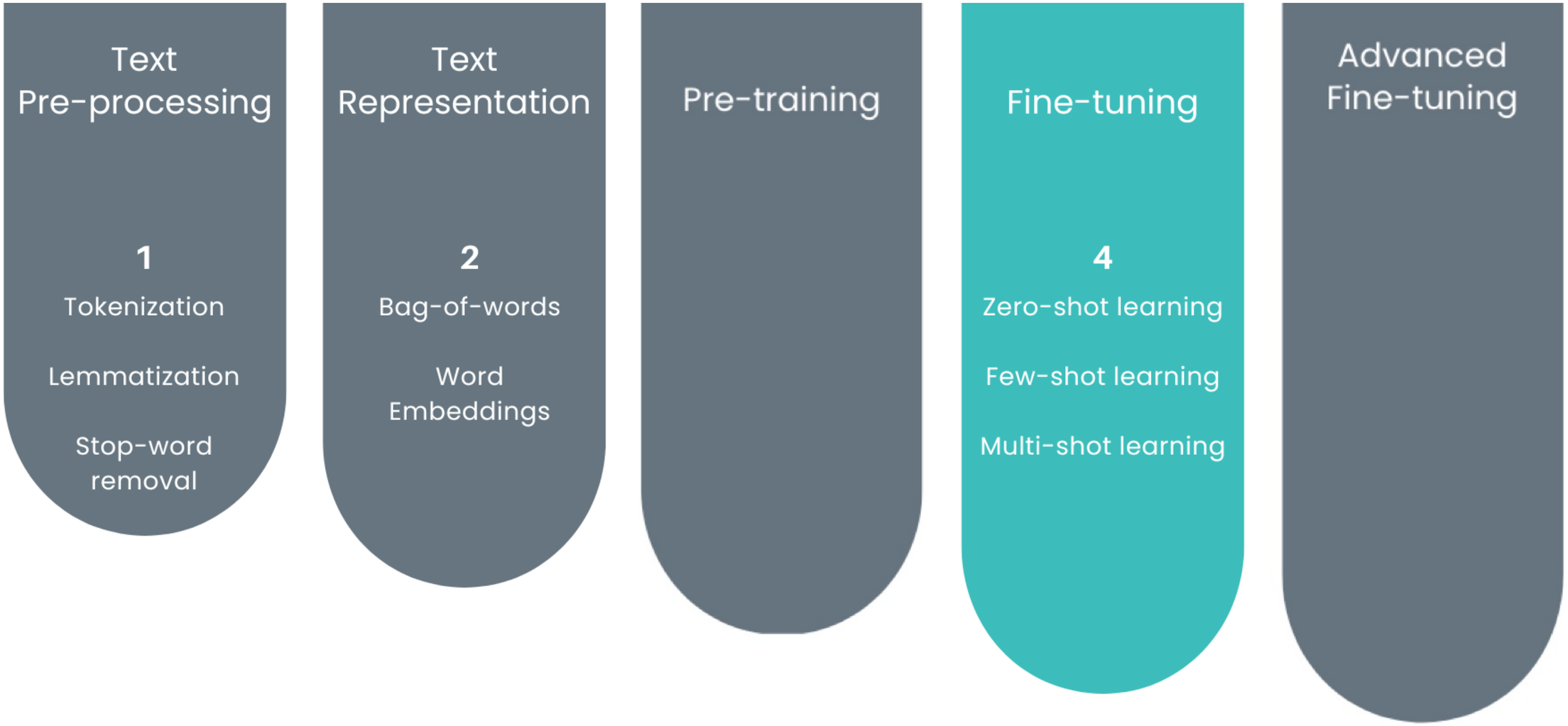
# Fine-tuning

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

**Vidhi Chugh**
AI strategist and ethicist

datacamp

# Where are we?


Text Pre-processing
1
Tokenization
Lemmatization
Stop-word removal
Text Representation
2
Bag-of-words
Word Embeddings
Pre-training
Fine-tuning
4
Zero-shot learning
Few-shot learning
Multi-shot learning
Advanced Fine-tuning

- **Pre-training**

School education

- **Fine-tuning**

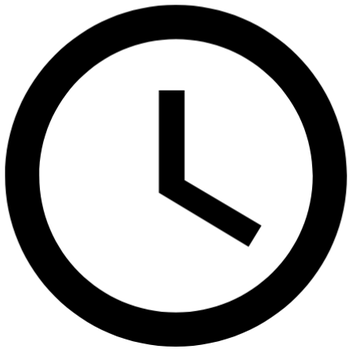
University specialization

[1] Freepik

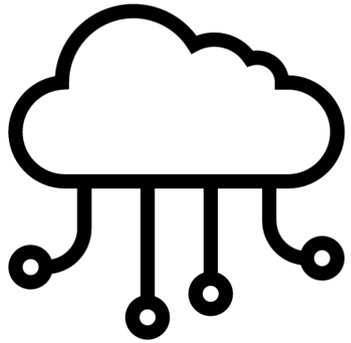
# "Largeness" challenges

- Fine-tuning can help
- Powerful computers
- Efficient model training methods
- Large amounts of training data

data



train



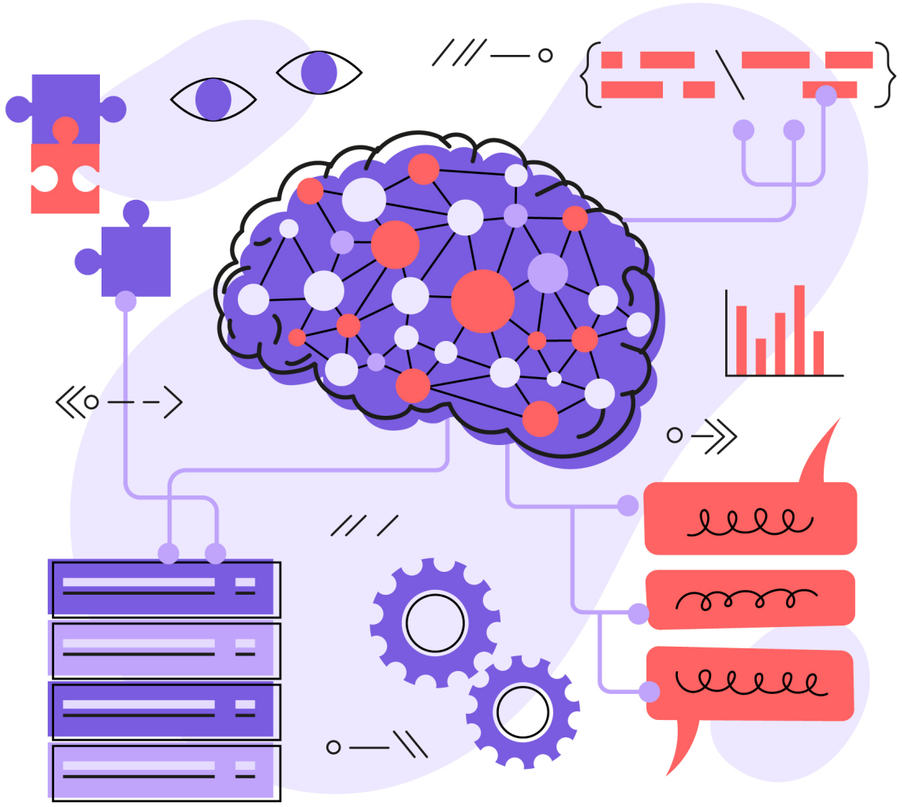
compute

# Computing power

- Memory
- Processing power
- Infrastructure
- Expensive
- **LLM:**
  - 100,000's Central Processing Units (CPUs)
  - 10,000's Graphic Processing Units (GPUs)
- **A personal computer:** 4-8 CPU and 1-2 GPUs

[1] Freepik

# Efficient model training



- Training time is huge
- May take weeks or even months
- Efficient model training = faster training time
- 355 years of processing time on a single GPU

# Data availability

- Need of high-quality data
- To learn the complexities and subtleties of language
- A few hundred gigabytes (GBs) of text data
  - More than a million books
- Massive amount of data

# Overcoming the challenges

- **Fine-tuning**
  - Addresses some of these challenges
  - Adapts a pre-trained model

- **Pre-trained model**
  - Learned from general-purpose datasets
  - Not optimized for specific-tasks
  - Can be fine-tuned for a specific problem

# Fine-tuning vs. Pre-training

- **Fine-tuning**
- Compute
  - 1-2 CPU and GPU
- Training time
  - Hours to days
- Data
  - ~1 gigabyte

- **Pre-training**
- Compute
  - Thousands of CPUs and GPUs
- Training time
  - Weeks to months
- Data
  - Hundreds of gigabytes

# Let's practice!

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

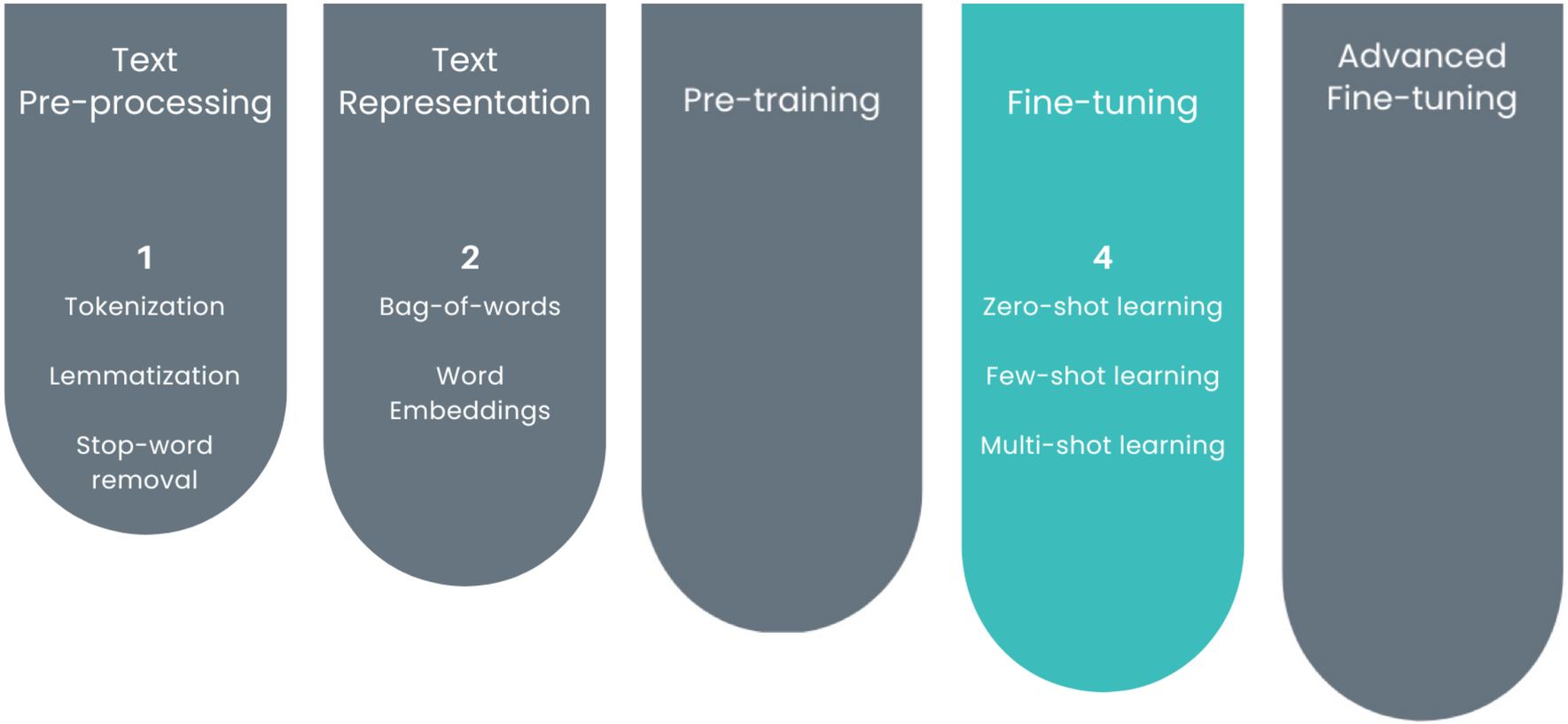
# Learning techniques

LARGE LANGUAGE MODELS (LLMS) CONCEPTS

**Vidhi Chugh**
AI strategist and ethicist

datacamp

# Where are we?


Text Pre-processing
1
Tokenization
Lemmatization
Stop-word removal
Text Representation
2
Bag-of-words
Word Embeddings
Pre-training
Fine-tuning
4
Zero-shot learning
Few-shot learning
Multi-shot learning
Advanced Fine-tuning

# Getting beyond data constraints

- **Fine-tuning**: training a pre-trained model for a specific task
- But, what if there is little to no labeled data?
- **N-shot learning**: zero-shot, few-shot, and multi-shot

# Transfer learning

- Learn from one task and transfer to related task
- Transferring knowledge from piano to guitar
  - Reading musical notes
  - Understanding rhythm
  - Grasping musical concepts
- **N-shot learning**
  - Zero-shot - no task-specific data
  - Few-shot - little task-specific data
  - Multi-shot - relatively more training data

# Zero-shot learning

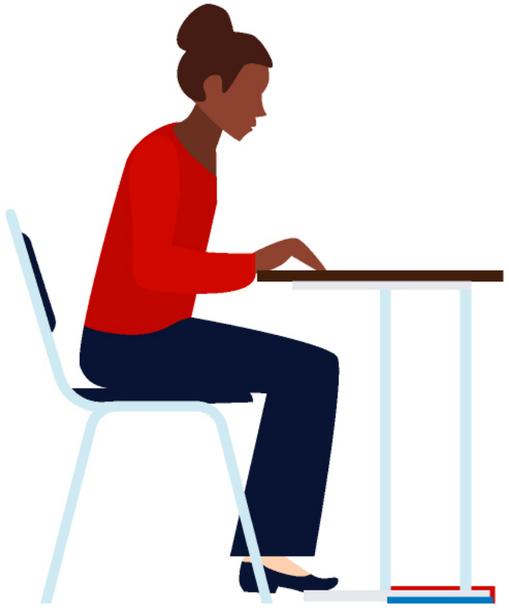
- No explicit training
- Uses language understanding and context
- Generalizes without any prior examples

[1] Freepik

# Few-shot learning

- Learn a new task with a few examples

- Prior knowledge to answer new question



- One-shot learning: fine-tuning from one example

# Multi-shot learning

- Requires more examples than few-shot

- Previous tasks, plus new examples

- For example, a model trained on Golden Retriever

[1] Freepik

# Multi-shot learning

- **Model output:** Labrador Retriever

- Saves time in collecting and labeling data

- No compromise on accuracy

[1] Freepik

# Building blocks so far

- Data preparation workflow
- Fine-tuning
- N-shot learning techniques
- Next up: pre-training

# Let's practice!

LARGE LANGUAGE MODELS (LLMS) CONCEPTS