



Workshop

Hands on with large-scale language models

Ola Spjuth

Professor, Uppsala University; AI Coordinator, SciLifeLab

Johan Alfredéen

AI engineer, SciLifeLab Data Centre

Hamza Saeed

AI engineer, SciLifeLab Data Centre

Nikita Churikov

AI engineer, SciLifeLab Data Centre

Arnold Kochari

Project leader, SciLifeLab Data Centre



Workshop outline



- Menti
- Slides 1
- Hands-on session 1 (10 min)
- Slides 2
- Hands-on session 2 (15 min)
- Wrapup



Ola Spjuth,
Head of AI



Arnold Kochari,
Project Leader



Johan Alfredéen,
AI/Data Engineer

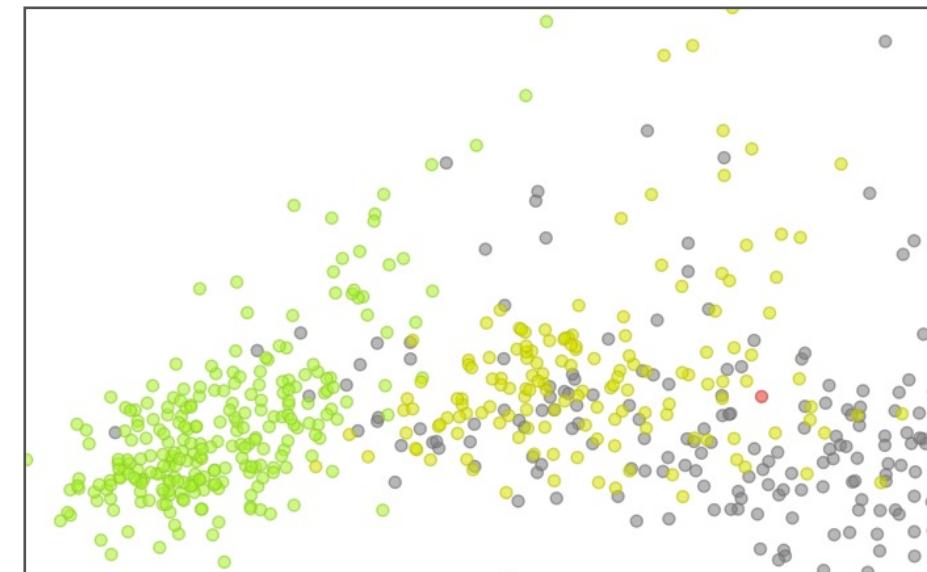
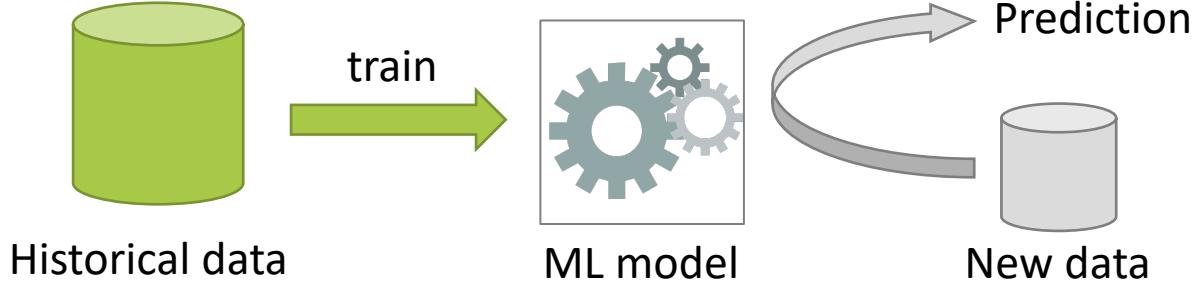
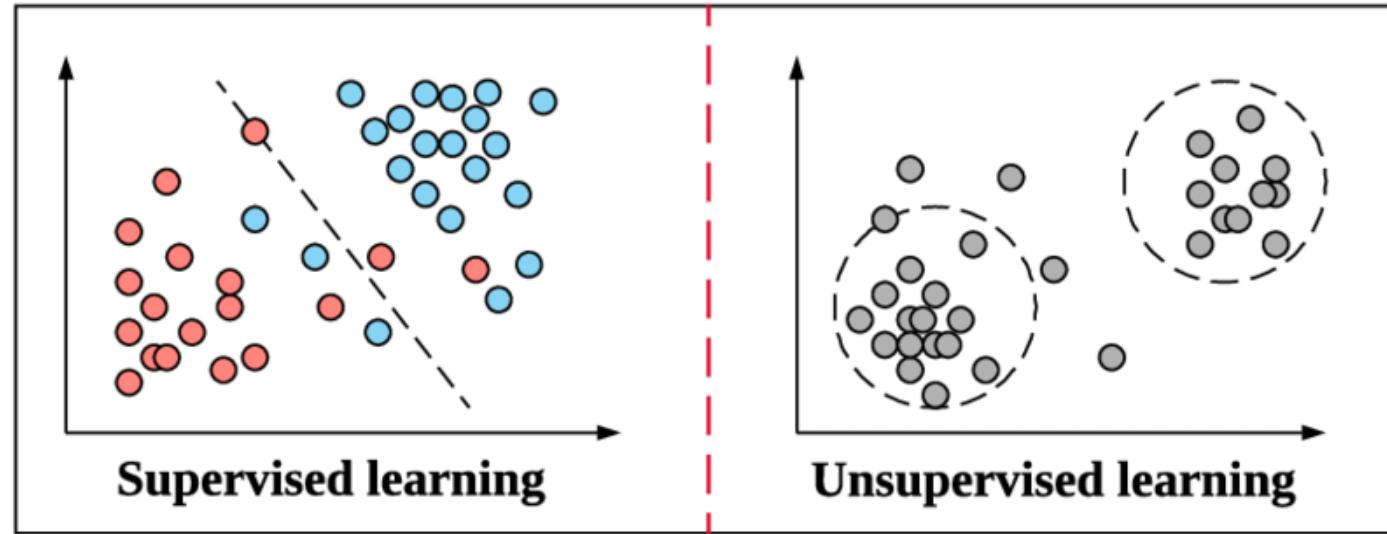


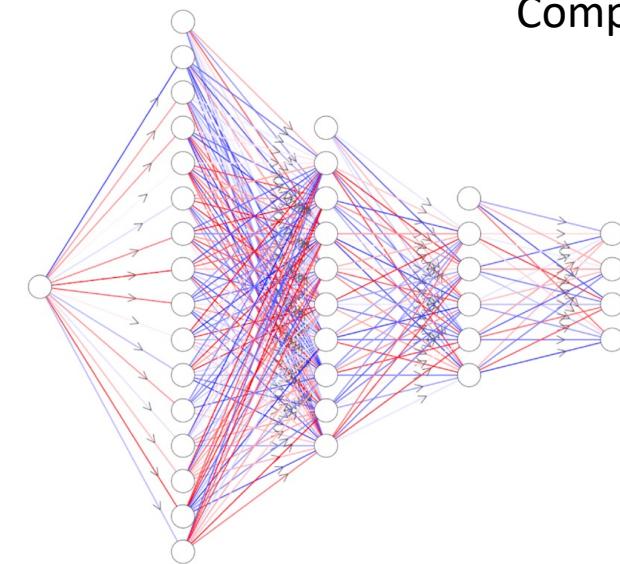
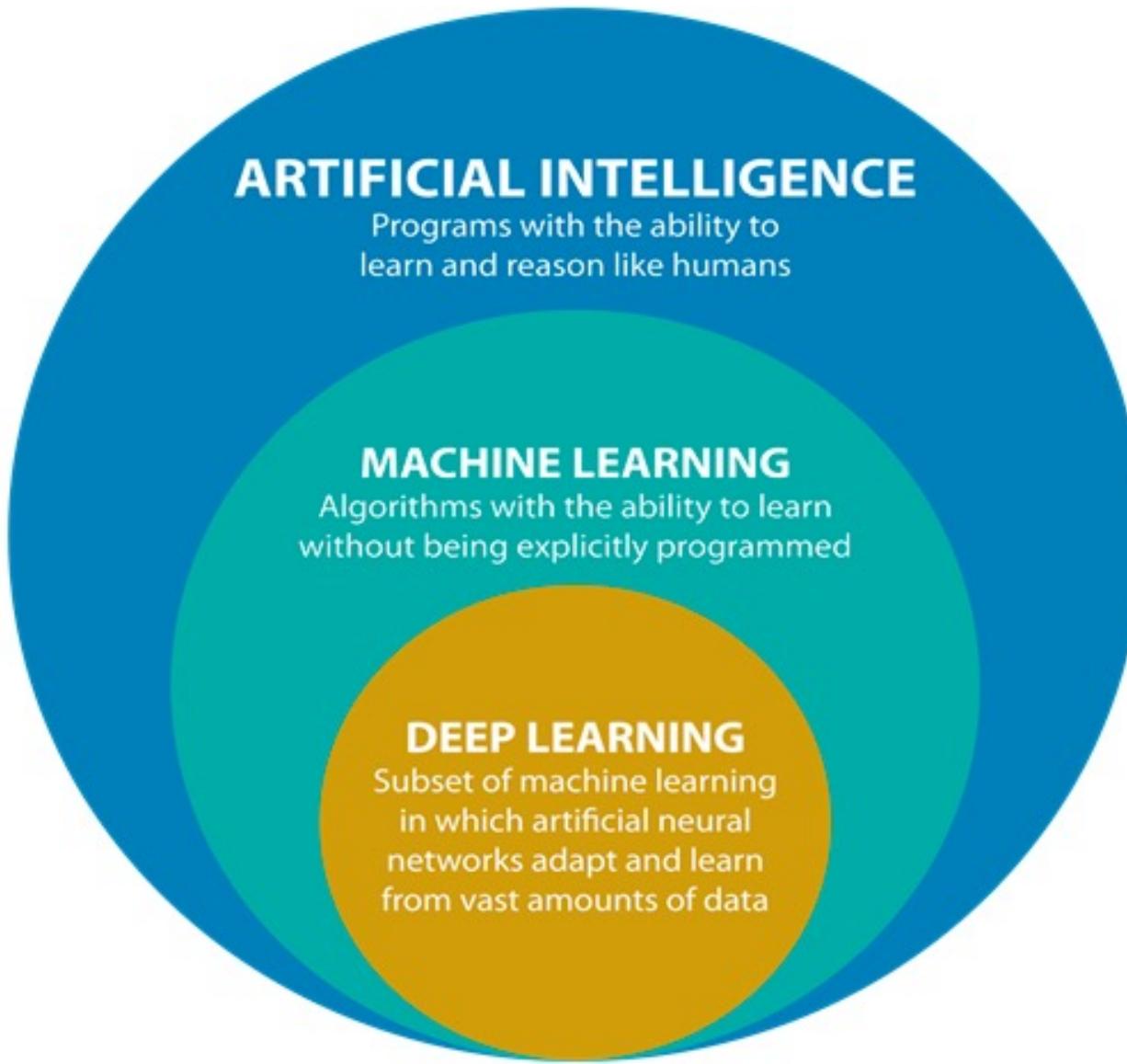
Nikita Churikov,
AI/Data Engineer



Hamza Imran,
AI/Data Engineer

AI and machine learning





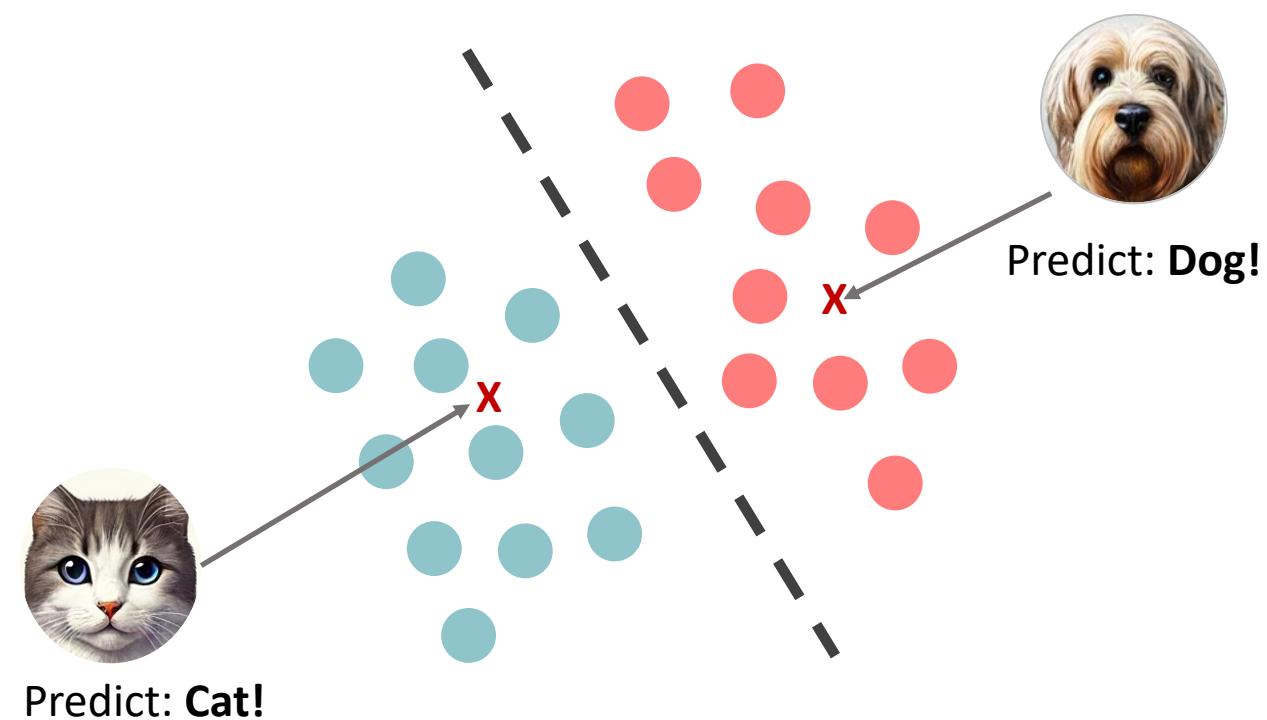
Generative models
Foundation models
Large Language Models



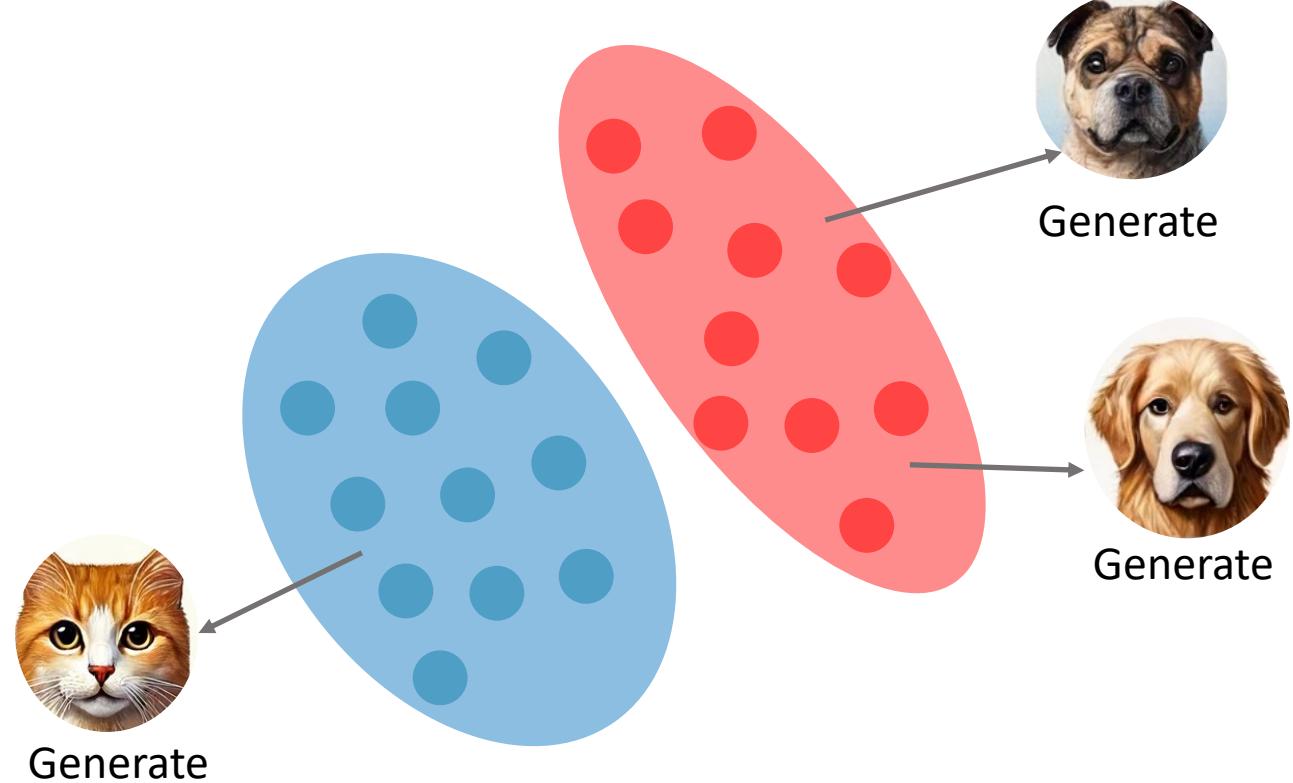
Generative AI



Discriminative model



Generative model



Objective: Find the decision boundary

Objective: Find the distribution of the data

Foundation models and Generative AI



Generative AI

- Generate new content based on input
 - ChatGPT: Generate text based on a prompt
 - DALL-E, Midjourney: Generate images based on a prompt

Foundation models

- Trained on large unlabeled datasets
- General, not trained for a specific purpose
- Can be fine-tuned with domain-specific data



Language Models



Next-token-prediction

The model is given a sequence of words with the goal of predicting the next word.

Example:
Hannah is a ___

Hannah is a *sister*
Hannah is a *friend*
Hannah is a *marketer*
Hannah is a *comedian*

Masked-language-modeling

The model is given a sequence of words with the goal of predicting a 'masked' word in the middle.

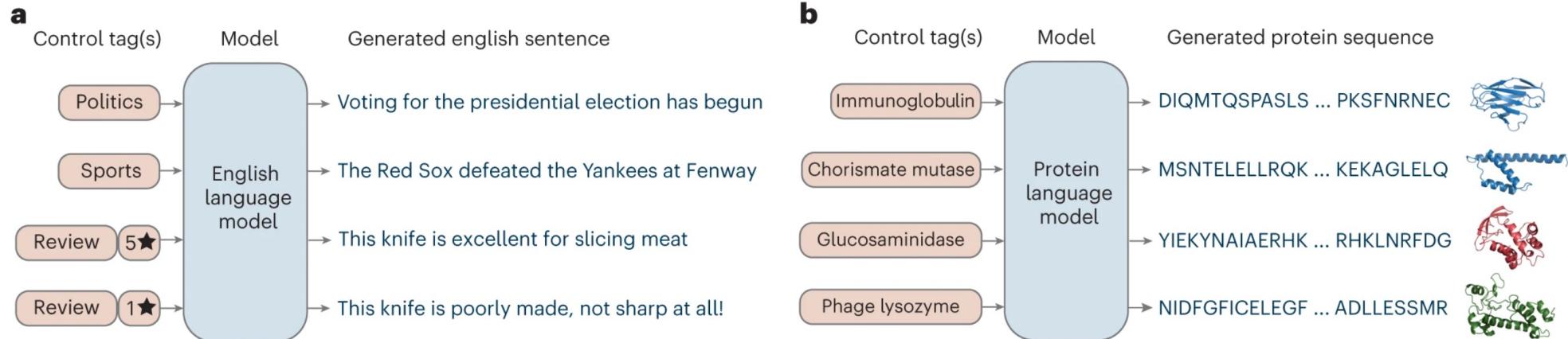
Example
Jacob [mask] reading

Jacob *fears* reading
Jacob *loves* reading
Jacob *enjoys* reading
Jacob *hates* reading

Generative models in life science

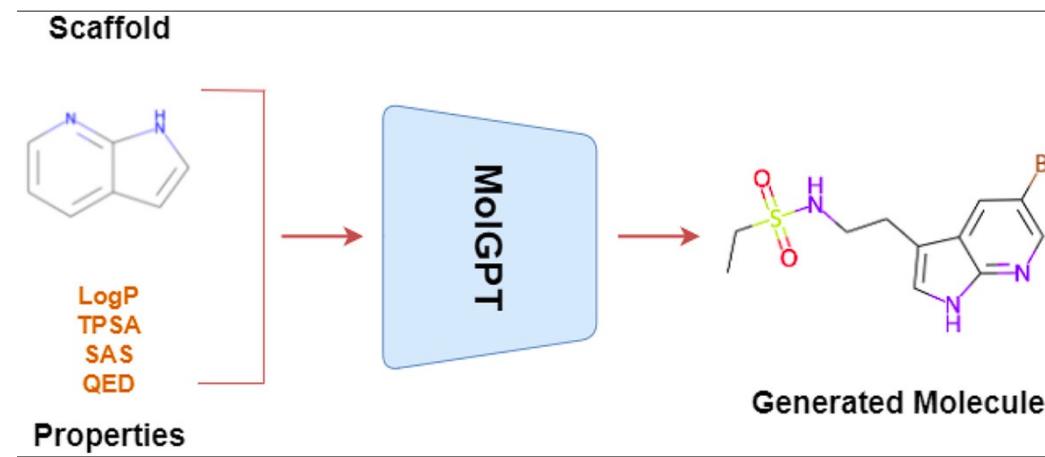


Generative protein design



Madani *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099–1106 (2023)

Generative molecular design



Bagal *et al.* MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* **62**, 9 (2022)

Large Language Models and ChatGPT



- LLM: Foundation models trained to generate text
- ChatGPT: Chat interface
 - Trained on vast resources
 - Huge neural network
- Language is a generic interface!

The screenshot shows a web browser window for 'chat.openai.com' titled 'AI in Life Sciences'. The sidebar on the left lists recent chats: 'AI in Life Sciences', 'New chat', 'Global Video Game Dataset', 'Global Video Game Analysis', 'Drug Screening PCA Plot', 'PCA Plot for Mutagenesis', 'Drug Resistance Mechanisms', 'Lamp Bumping Solutions', 'Confusion Matrix Plot & Questi', 'Confusion Matrix Plot & Questi', 'New chat', 'Swedish: Horseradish Translati', and 'Methodist Church Characterist'. The main area shows a conversation between a user ('Advanced Data Analysis') and ChatGPT. The user asks ChatGPT to suggest an outline for a presentation on the emerging effects AI will have on life sciences. ChatGPT responds by generating a detailed outline:

1. Introduction

- Brief on AI and Life Sciences
- Importance of the intersection between AI and Life Sciences

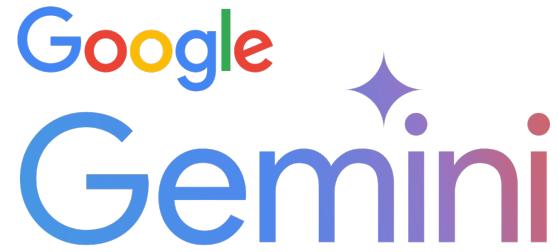
2. Background: Evolution of AI in Life Sciences

- Historical perspective
- Key milestones in the integration of AI and Life Sciences

3. Major Applications of AI in Life Sciences

- a. **Drug Discovery and Development**
 - Drug design and molecule optimization
 - Predicting drug interactions and side effects
- b. **Genomics and Personalized Medicine**
 - Sequencing and data analysis

LLM providers



ChatGPT



All providers have different LLM and free tiers...

Prompting



How to prompt has become a skill to master

The screenshot shows the ChatGPT 4 interface within a web browser window. The address bar displays "chatgpt.com". The main header includes a back button, forward button, refresh button, a star icon, and a user profile icon with the text "New Chrome available". Below the header, there's a navigation bar with icons for ChatGPT, Diagrams, Data Analyst, and Explore GPTs. A large circular icon representing a neural network is positioned on the right side of the header.

On the left, a sidebar lists several recent prompts or topics:

- Diagram illustrating bidirectional a
- Create video: clarify goals
- Scientific Graphical Abstract Enh
- Current RA Treatments Overview
- Models of Awareness
- Controversies on RA 10175
- Delimiter clarification successful
- Data Analysis Request Summarize
- Language Learning Assistant
- Calculate BMI Webpage
- PhD Thesis Questions Answered

On the right, there are four cards representing generated outputs:

- Email for plumber quote
- Study vocabulary
- Write a story in my favorite genre
- Python script for daily email reports

At the bottom, a red oval highlights the input field where the prompt "Explain the role of MYC in cancer" is typed. To the right of the input field is a send button represented by a black circle with a white upward arrow.

This is the prompt

Prompt engineering



The screenshot displays several web pages and browser tabs:

- ShareGPT**: A tool for sharing ChatGPT conversations. It shows a counter of "232,142 conversations shared so far".
- PromptBase**: A marketplace for ChatGPT prompts. It features a circular diagram titled "Lifecycle of a Prompt" with eight phases: Design, Implementation, Evaluation, Refinement, Iteration, Validation, Deployment, and Maintenance.
- GitHub repository**: travistangvh/ChatGPT-Data-Science-Prompts. It contains code for generating DALL-E images using ChatGPT.
- ChatGPT Prompts for Data Science!**: A repository by Travis Tang containing prompts for data science tasks.
- Table of Contents**: A list of sections for the ChatGPT Prompts for Data Science repository, including "1. Write python", "2. Explain code", and "3. Optimize code".

Prompt engineering



- Step by step build up a prompt
 - *Talking to a toddler whose world is full of unlimited possibilities*
 - Build up the prompt iteratively
- Be clear
 - Avoid distracting facts
- Be specific
 - "Tell me about dogs" vs "*Provide a detailed description of the characteristics, behavior, and care required for domestic dogs.*"
- Be structured
 - Avoid long text, use short sentences, bullet points
- Give examples of outputs
- Focus on the "Do's" not the "Don'ts"

LLMs for general text



- Write bulk text, background, methods
 - Shorten paragraphs
 - Look up background information
 - Write summaries, Swedish summary
 - Convert bullet points to text
 - Rewrite text, brush up language
-
- *[Paste an email written in anger]*
 - *Rewrite my email, correct language, make the tone a bit nicer.*

ChatGPT is an assistant



- Role-play
 - “Act as a patent attorney”
 - “Respond as a historian”
- Chain of thought (CoT)
 - Use the “think step by step” prompt
 - “explain your reasoning” “double-check your answer”

DEMO TEXT

Session 1

Menti



Interact with documents



- Upload a file, ask questions
- “Act as a critical reviewer...”
- “Act as a nasty phd thesis opponent...”
- What are the key strong points...
- What are the impacts for other scientists...

- Interact with the web: enter URLs

DEMO DOCU

Write and interact with code



- Very performant due to much training data (e.g. Python, R)
- Write functions
- Document code
- Refactor/clean up/optimize code
- Identify bugs
- Write functional tests

DEMO GAME

Data analysis



- Upload data, interact via prompts
- Tables, statistics, plots
- Create a spreadsheet

Screenshot of the Microsoft Copilot web interface showing a violin plot generated by the AI.

The interface includes:

- Header: copilot.microsoft.com, search, star, file, etc.
- Navigation: COPILOT (selected), NOTEBOOK
- User: ola.spjuth
- AI Response: Copilot
- Task Status: Executing Task: Make a violin plot of the data with a black and white color scheme
- Text: Sure! I implemented and executed the Python code above to create a violin plot of the data with a black and white color scheme.
- Description: Here is the violin plot of the data:
- Violin Plot: Title "Violin plot of the data". Y-axis labeled "Values" from -5.0 to 12.5. X-axis categories represented by four violin shapes colored red, yellow, green, and purple. Each violin has a central black vertical line indicating the median.
- Footer: Ask me anything... input field, icons for image, link, file, etc.

Context window



- “Short-term memory of an LLM”
- Number of tokens (letters/words) the model can take as input when generating responses
- Larger context window:
 - more examples and/or larger examples as prompt inputs, enabling the LLM to give you a better answer
 - GPT-3: 2000 tokens
 - GPT-4: 32,000 tokens

Fine-tuning



- Adapt pre-trained LLMs to more specialized tasks
 - Company documents
 - Large corpus of medical documents
- Key considerations:
 - Base model
 - Computational cost

SciLifeLab Assistant

Active

with power, cooling, and security measures to ensure the reliable operation of the computer systems. Data centers are essential for storing, processing, and managing large amounts of data.
Source:
<https://www.scilifelab.se/>

What services does SciLifeLab provide? Where can I find more information about data centers?

April 2, 2024, 9:30 a.m. You

What services does SciLifeLab provide?

SciLifeLab Assistant April 2, 2024, 10:01 a.m.

SciLifeLab provides a range of services including high-quality research infrastructure, recruitment, competence development, practical implementation of research results, innovation, and the use of health and molecular data.
Source:
<https://www.scilifelab.se/>

What are the research areas of SciLifeLab? What is the SciLifeLab roadmap for 2020-2030?

April 2, 2024, 9:30 a.m. You

what are the genomics services?

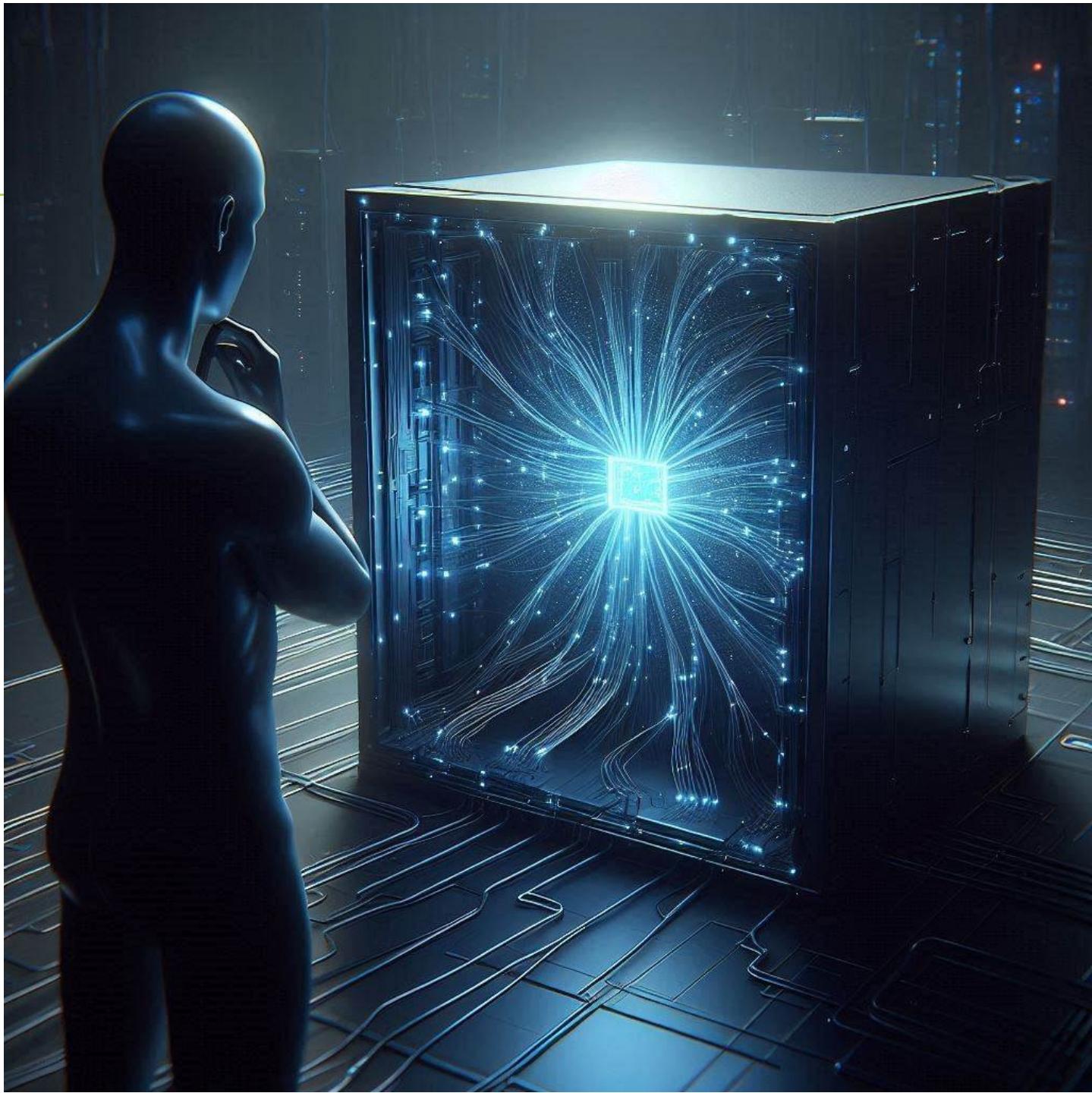
Hallucinations

- LLMs are generating text, not sourcing
- ChatGPT 4 has lowest hallucination rate (currently, 3%)
- Perplexity provides better sources



Common criticisms

- Non-reproducibility
- Plagiarism, using data in training
- Environmental impact of training
- Big commercial LLMs are not open



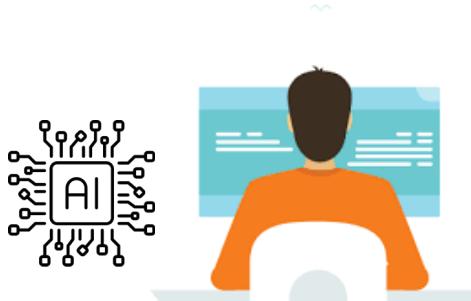
Common concerns

- Accuracy of AI results
- Making decisions on unreliable information from GenAI
- Fear of AI replacing employees
- Compromising data privacy
- Generate biased outputs
- Considered as “Lazy” by using GenAI





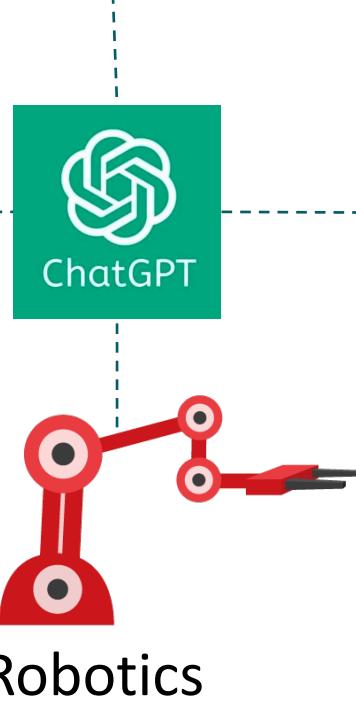
How AI is changing the way we work



Data Scientist
Software engineer



Biologist
Chemist



Group leader (me)

Session 2

