



# **Introduction to Data-driven Life Sciences**

Wei Ouyang - August 29th 2023

# What are the data types in life science?

---



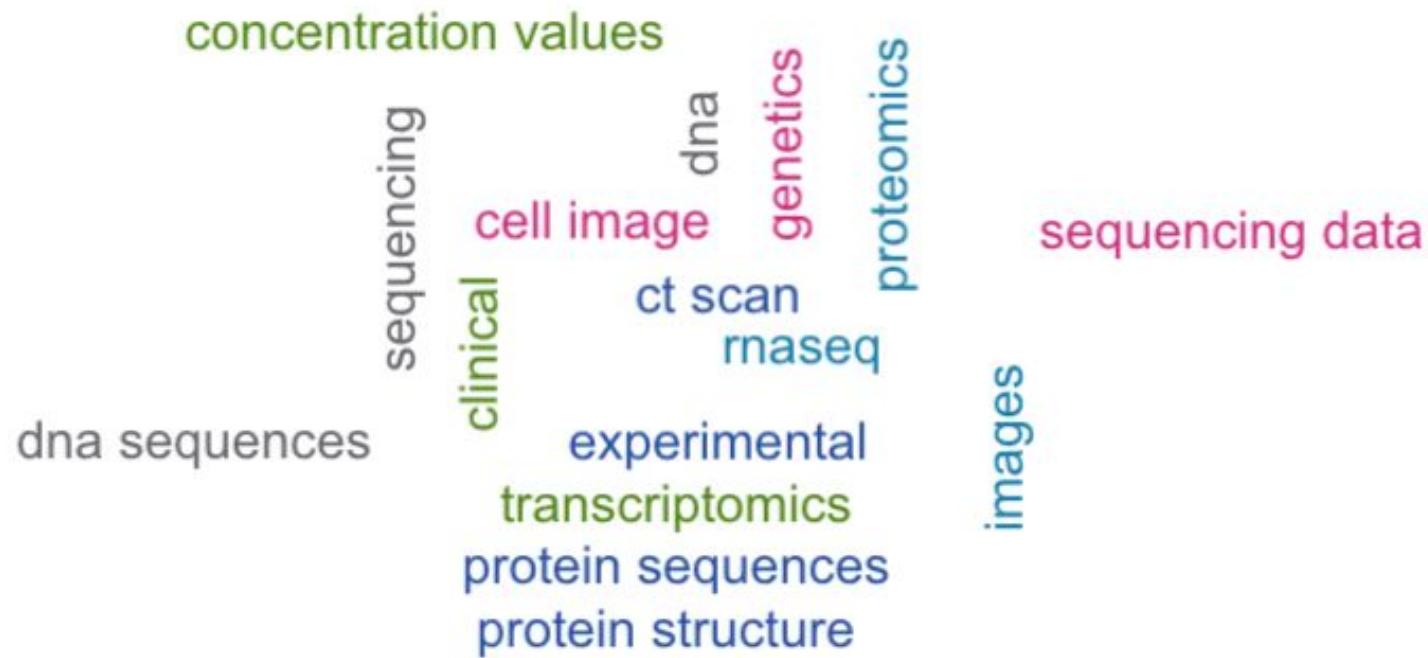
A QR code is centered within a white square frame, which is itself centered within a red rounded rectangle. Below this is a blue rounded rectangle containing the text "Menti for Google Slides" and a descriptive note.

**Menti for Google Slides**

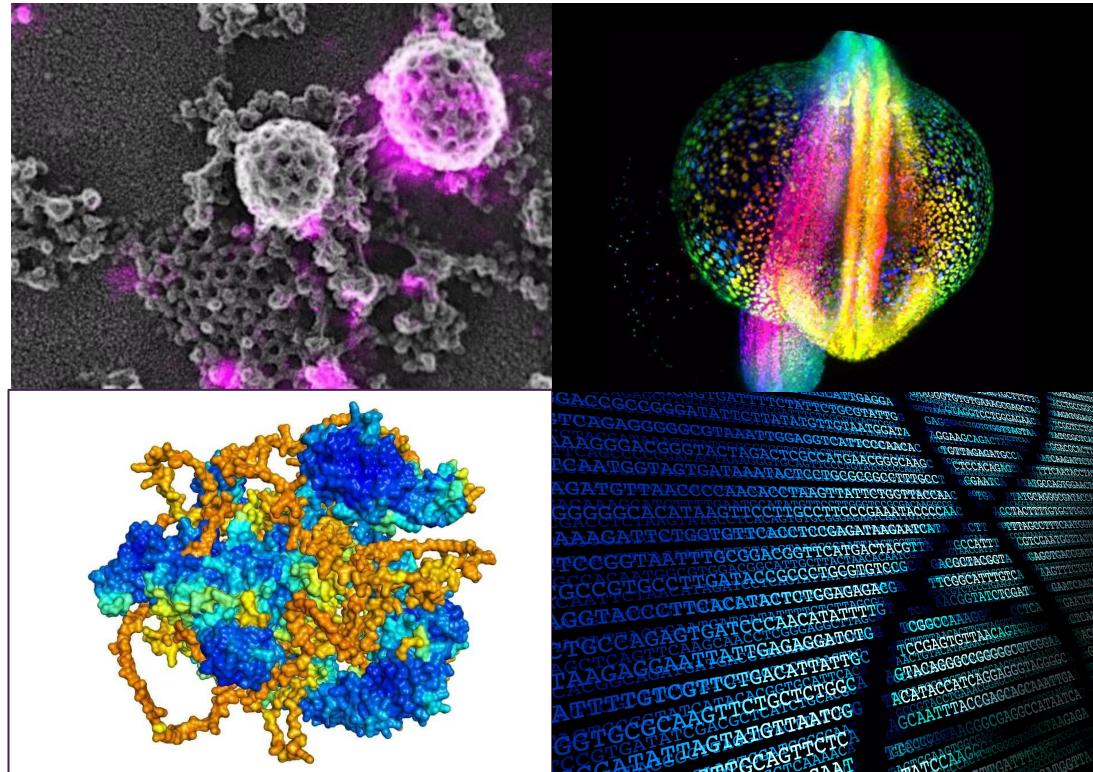
Only the Menti presentation  
will be shown for this slide

## The data types in life science?

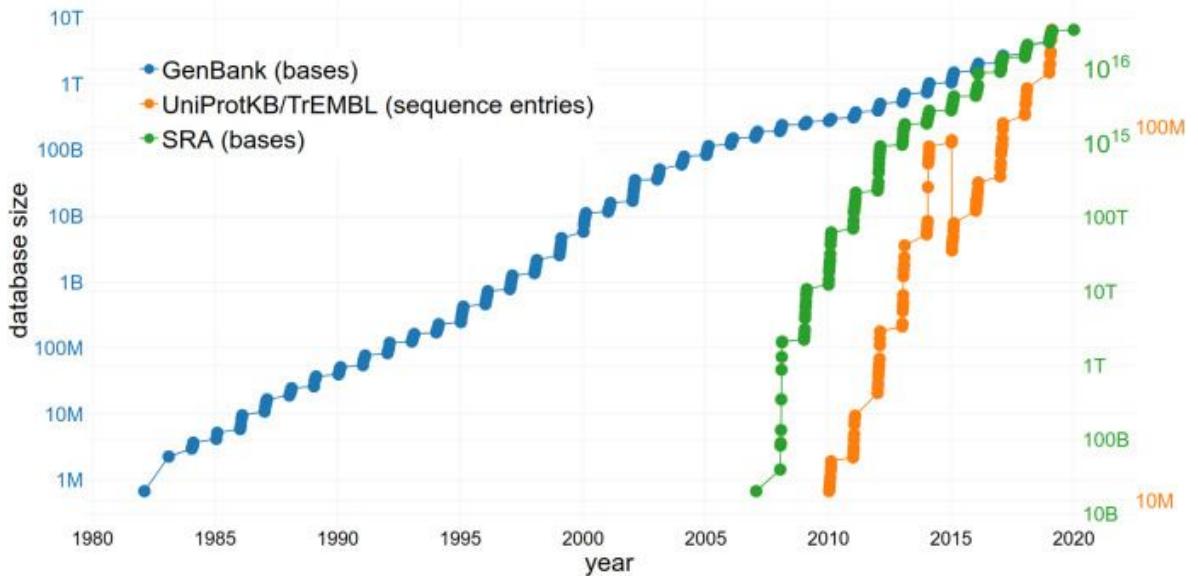
16 Responses



# Paradigm Shift: Big Data in Life Science

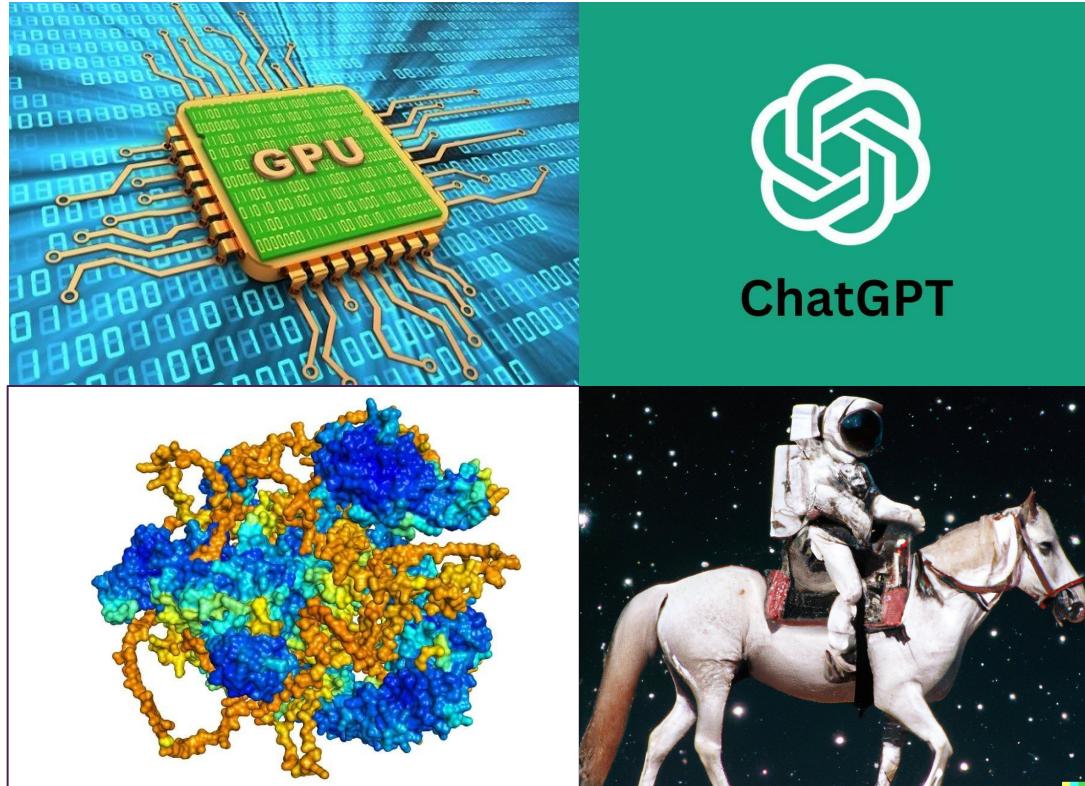


The rapidly increasing databases



Sielemann et al, 2020

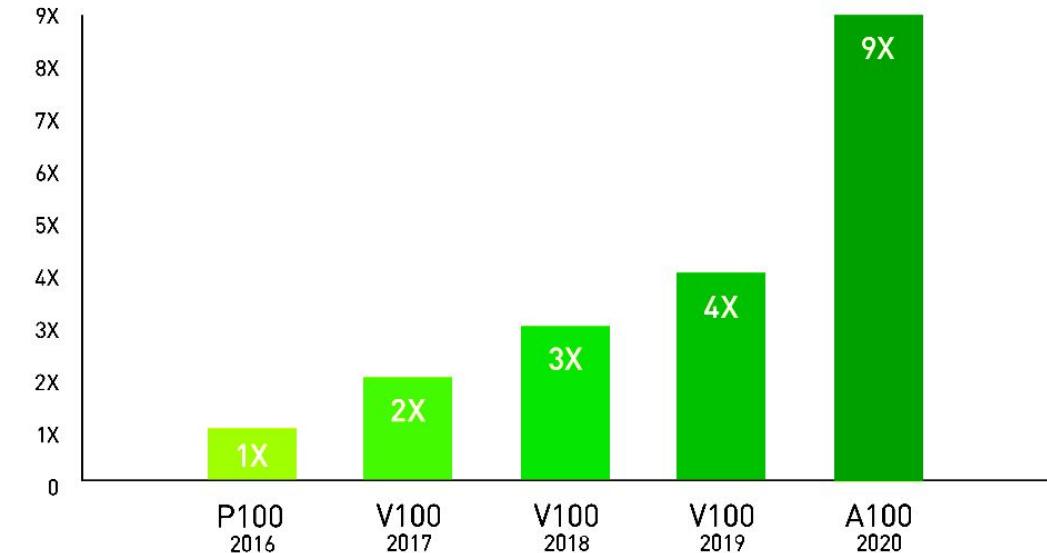
# Paradigm Shift: Compute Power & AI



9X More HPC Performance in 4 Years

[nvidia.com](https://nvidia.com)

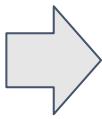
Throughput for Top HPC Apps



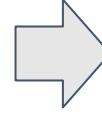
# Paradigm shift: Data-driven Life Science



Hypothesis  
driven  
research



Data driven  
research

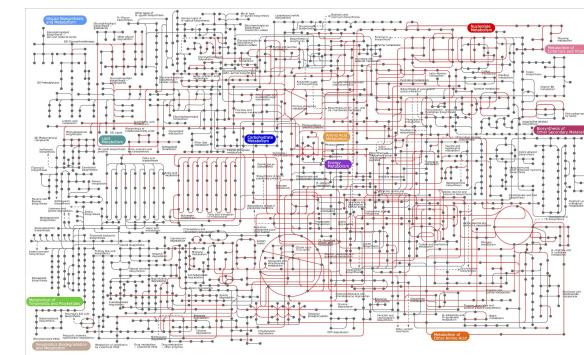
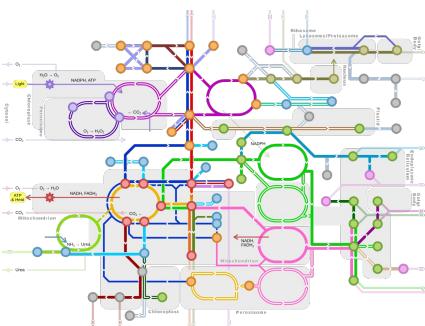
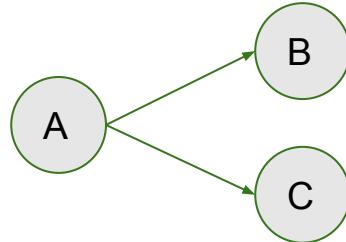


AI-powered  
data-driven  
research

1. Hypothesis
2. Experiment
3. Accept or reject

1. Acquire data
2. Find pattern
3. Evaluate pattern

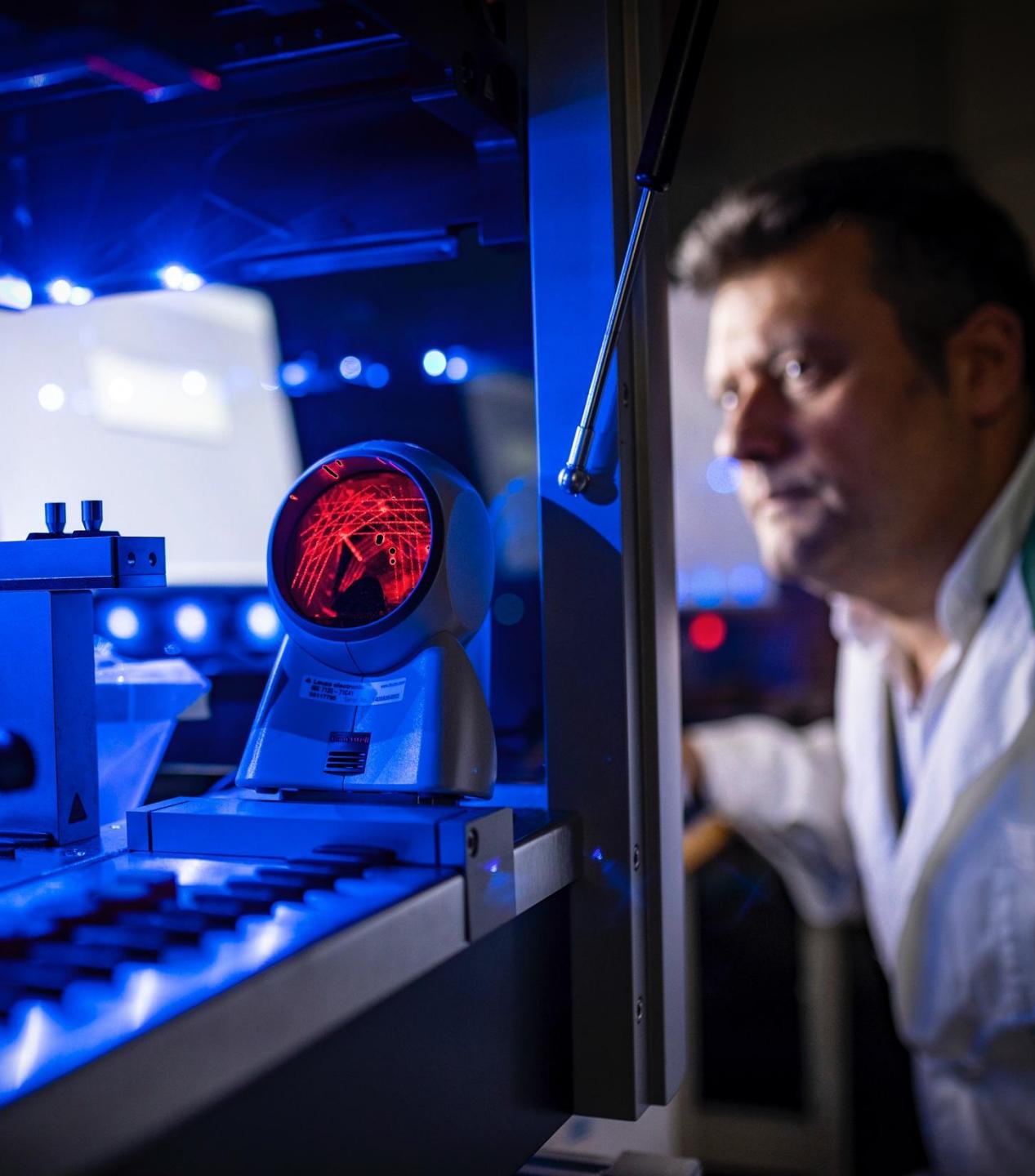
1. Acquire big data
2. Fit AI model
3. Evaluate AI model





# About SciLifeLab

Our vision, mission and strategic objectives.



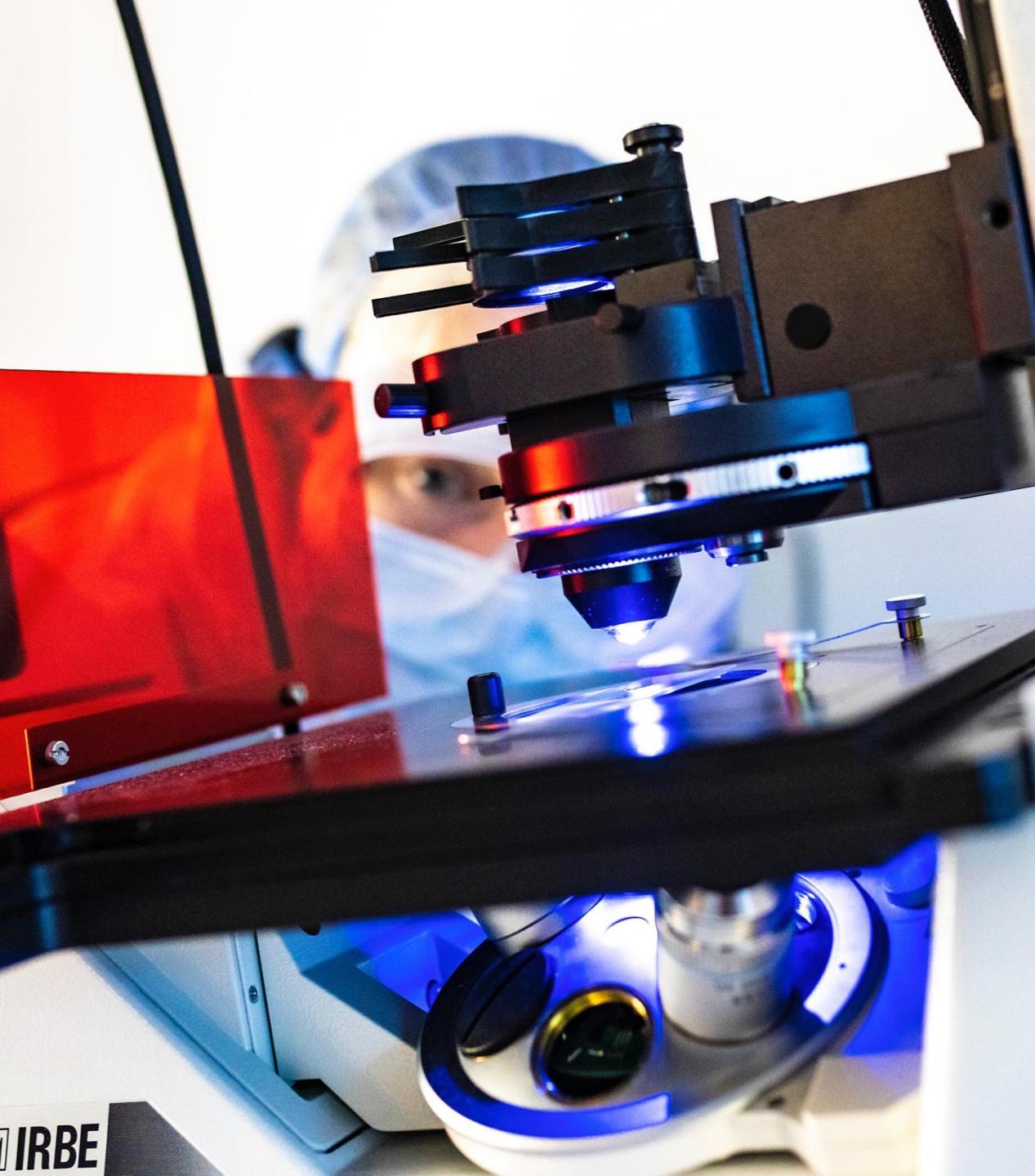
# What is SciLifeLab?

National hub **enabling** life science research that would otherwise not be possible

Government appointed mission as a **national research infrastructure**.

Started in 2010 by Karolinska Institutet, KTH Royal Institute of Technology, Stockholm University and Uppsala University.

Today, activities at **all major Swedish universities**.



## Vision and mission

**Vision:** for Sweden to be a world-leading nation in life science

**Mission:** Enable life science research that would otherwise not be possible

# Three dimensions of SciLifeLab



## Research environment

Approx. 190 affiliated research groups

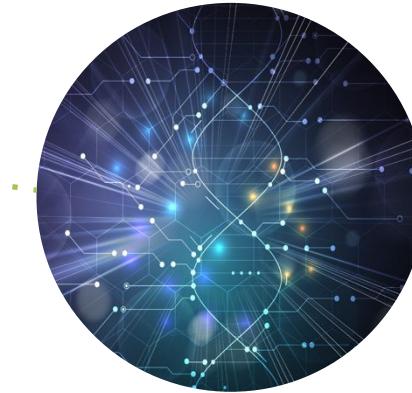
- Environment and climate change
- Farming and forestry
- Evolution and biodiversity
- Gene editing
- Biofuels and biomaterials
- Microbiology and microbiome
- Drugs and biomedicine
- Healthcare and aging



## Infrastructure

Service to ~ 1400 Swedish researchers annually (2020)

- Bioinformatics
- Cellular and molecular imaging
- Clinical diagnostics
- Single cell biology
- Genomics
- Chemical biology and gene editing
- Drug development
- Proteomics and metabolomics



## Data-driven life science

3.1 billion SEK, 12-year-program

Putting Sweden at the forefront of data-driven life science research and fostering the next generation of life scientists

- Four strategic research areas
- Recruiting talent from across the globe
- Academic and industry PhD and postdoc programs
- Sparking collaborations, innovation and interdisciplinary team science
- Building a strong computational and data science base for open, real-time data



# **SciLifeLab and Wallenberg National Program for Data-Driven Life Science**

Changing the way life science is carried out

# SciLifeLab and Wallenberg National Program for Data-Driven Life Science



Knut och Alice  
Wallenbergs  
Stiftelse

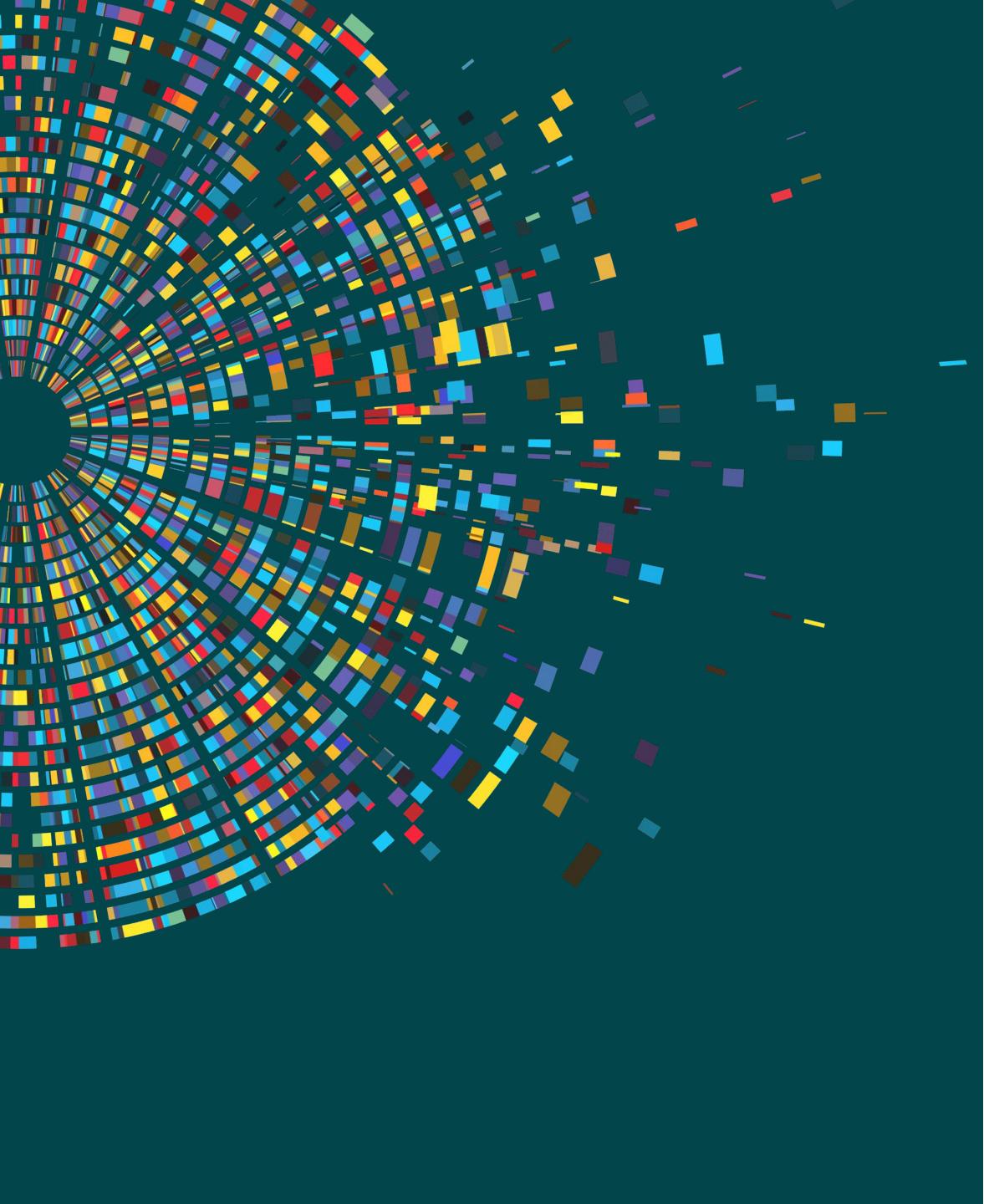
WASP | WALLENBERG AI,  
AUTONOMOUS SYSTEMS  
AND SOFTWARE PROGRAM



WALLENBERG CENTRES FOR MOLECULAR MEDICINE

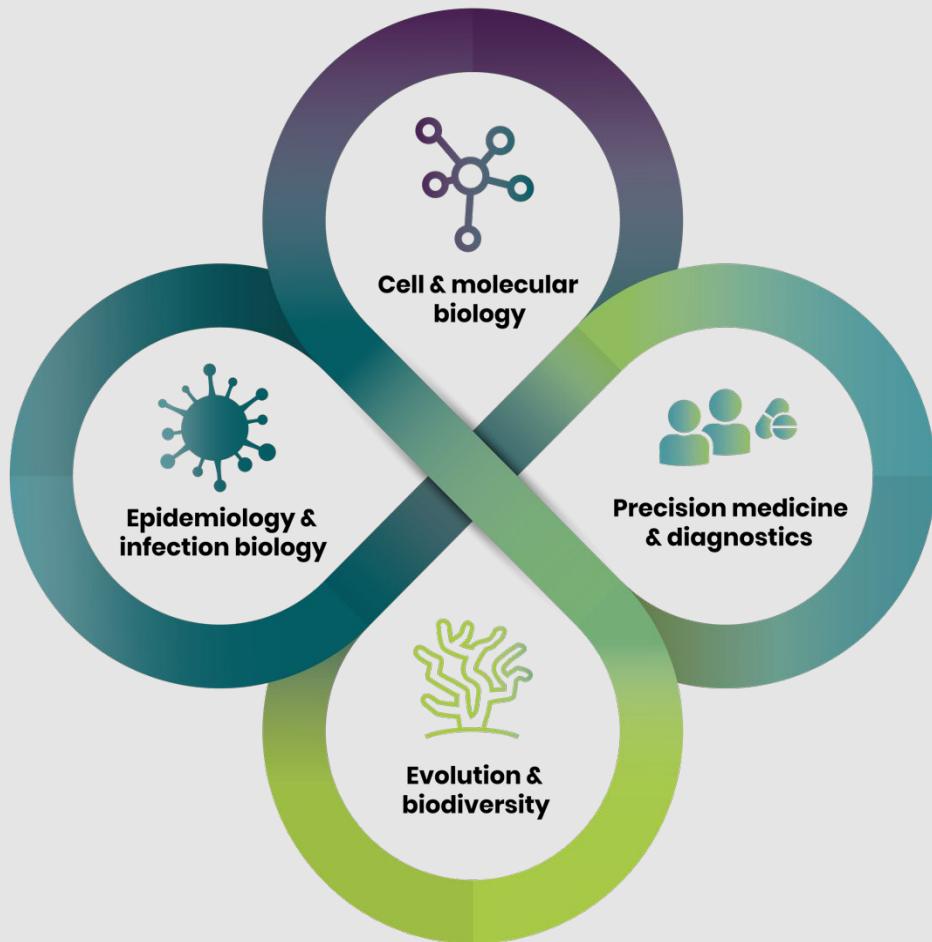
WACQT | Wallenberg Centre  
for Quantum Technology





## Data has a central place in life science

- Practice of life science is more and more data-dependent
- The amount of data grows exponentially
- Data becomes more complex, continuous, and needs to be openly available accessible and reusable in real-time to all



*The four strategic research areas of the DDLS program*

## Promoting a paradigm shift in life sciences

**Launching a research program to:**

- Foster the next generation of life scientists
- Enable every biologist to better analyze and interpret data patterns and integrate their own data seamlessly with the global life science data streams
- Create a strong computational and data science base
- Enable deep understanding of life through data

---

Advancing competence and innovation in **four research areas**

Using new **technologies** and **strategies** to utilize open and real-time data

Engaging in education, training, recruiting new talent, sparking collaborations, and in innovation activities

# The data life cycle



## Data Centre

Created to **maximize impact** of SciLifeLab generated data

Assists in communication between platforms, users, and research community

Acts as a point of contact for data management questions relating to SciLifeLab generated data

Assists platforms with data tracking and statistics

Facilitates providing SciLifeLab generated data with SciLifeLab funded bioinformatics and data management support

Assists with planning the handling of SciLifeLab generated data throughout projects



# FAIR principles

---



**Findable** – assigning a globally unique and eternally persistent identifier (like a DOI or Handle), describing the data with rich metadata, and making sure it is findable through disciplinary discovery portals.

**Accessible** – data and metadata should be retrievable in a variety of formats that are sensible to humans and machines using persistent identifiers.

**Interoperable** – the description of metadata elements should follow community guidelines that use an open, well defined vocabulary.

**Reusable** – the data should maintain its initial richness. The description of essential, recommended, and optional metadata elements should be machine processable and verifiable, use should be easy and data should be citable to sustain data sharing and recognize the value of data.

# Group Discussion

---

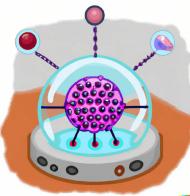


Discuss in groups of 3-4 on the following questions (5 minutes):

- What are the obstacles for implementing FAIR in Life Science?
- How can the FAIR principles be implemented more effectively in life sciences?

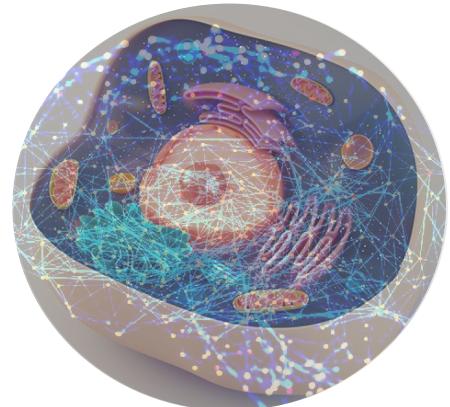
Reminder: FAIR = Findable, Accessible, Interoperable, Reusable

# Getting prepared for Data-Driven Life Science Efforts in the AICell Lab



# AICell Lab (<https://aicell.io>)

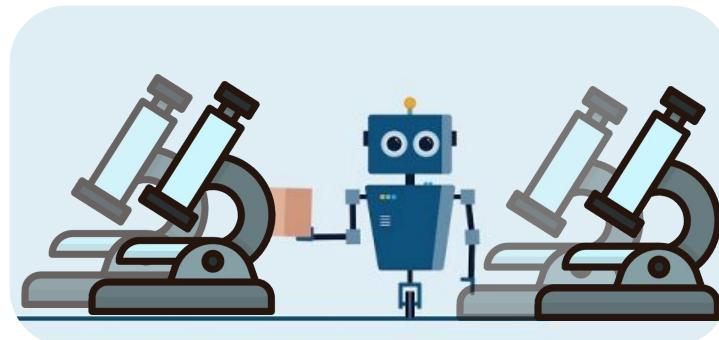
Human Cell Simulator



**Data  
Generation**

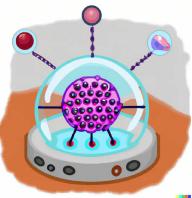
**AI Models  
Execution**

Microscopy Imaging Farm

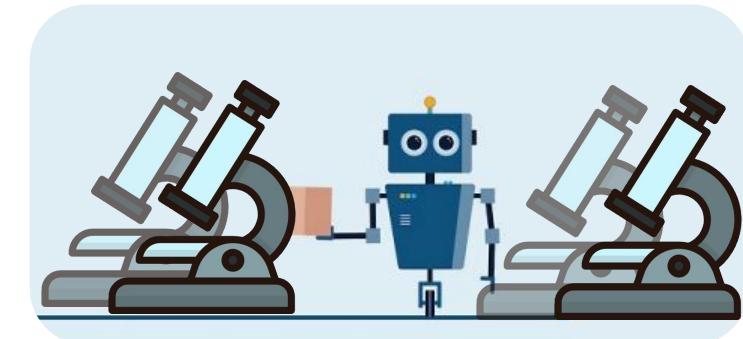
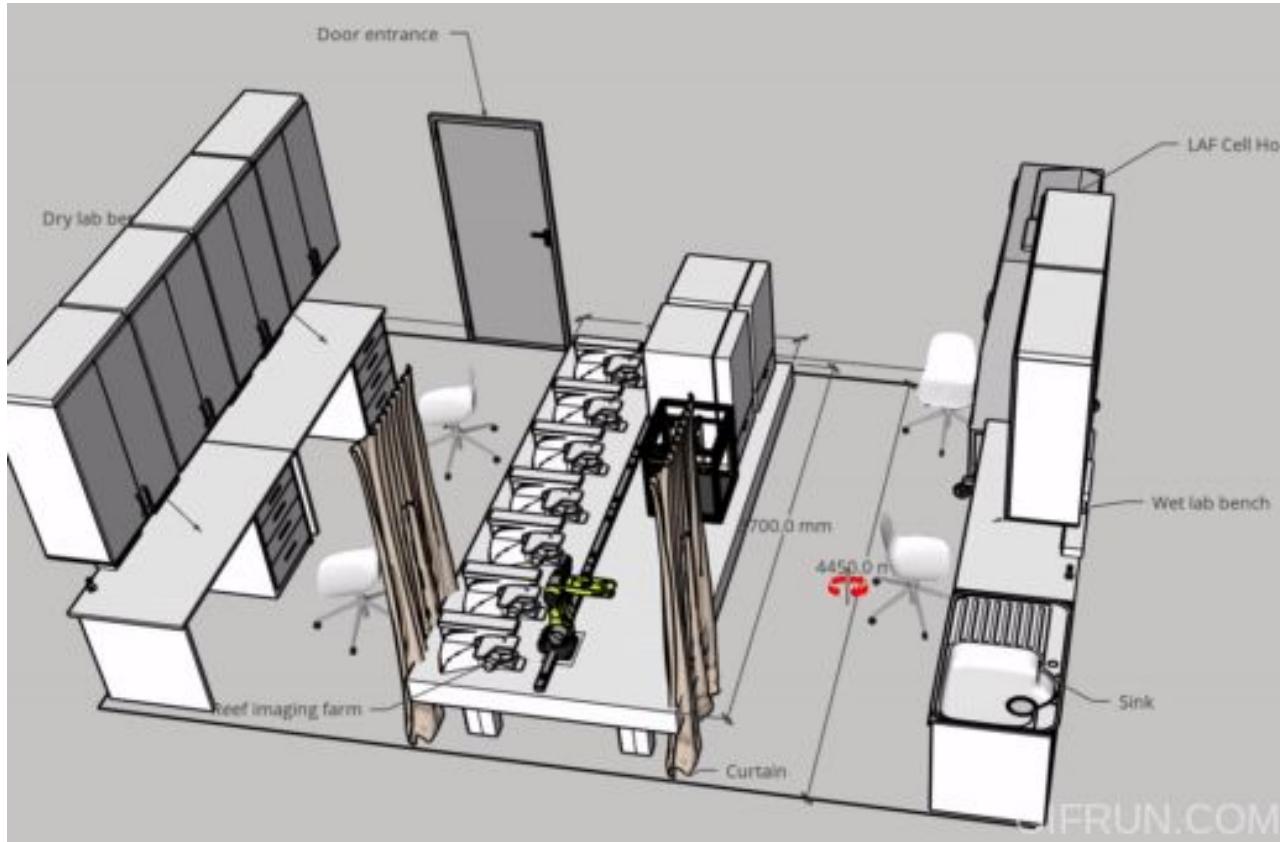


AI Cloud Infrastructure



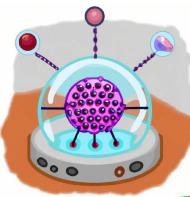


# REEF: Smart Microscopy Imaging Farm

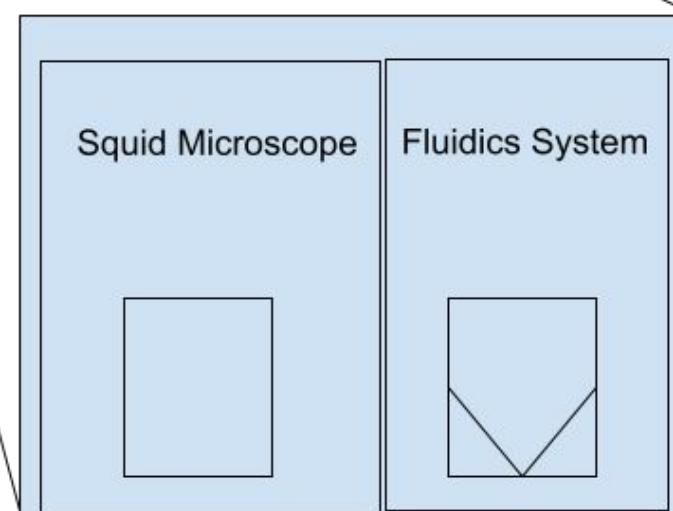
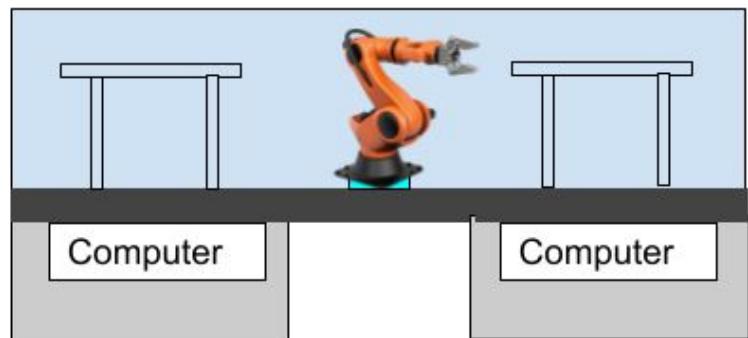
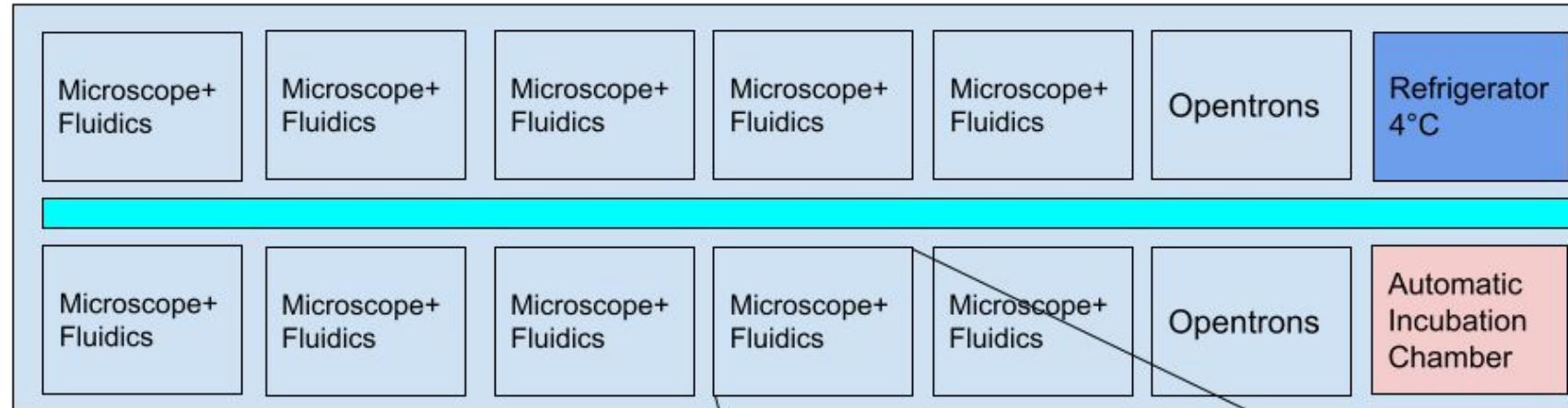


**For Massive Dataset Generation**

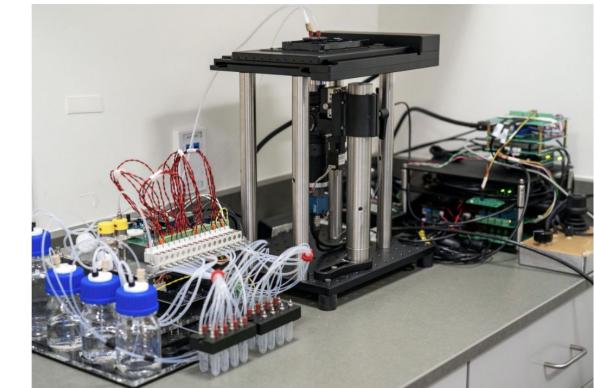
REEF Imaging Farm @ SciLifeLab



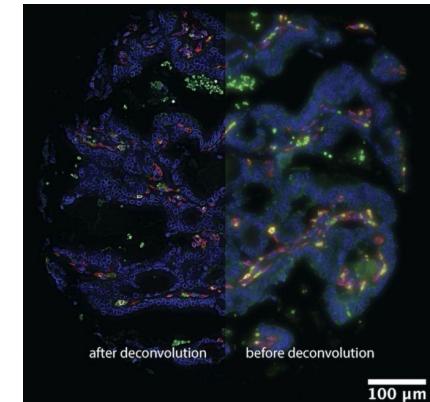
# REEF: Smart Microscopy Imaging Farm



Squid Microscope Prototype



CODEX multiplex imaging

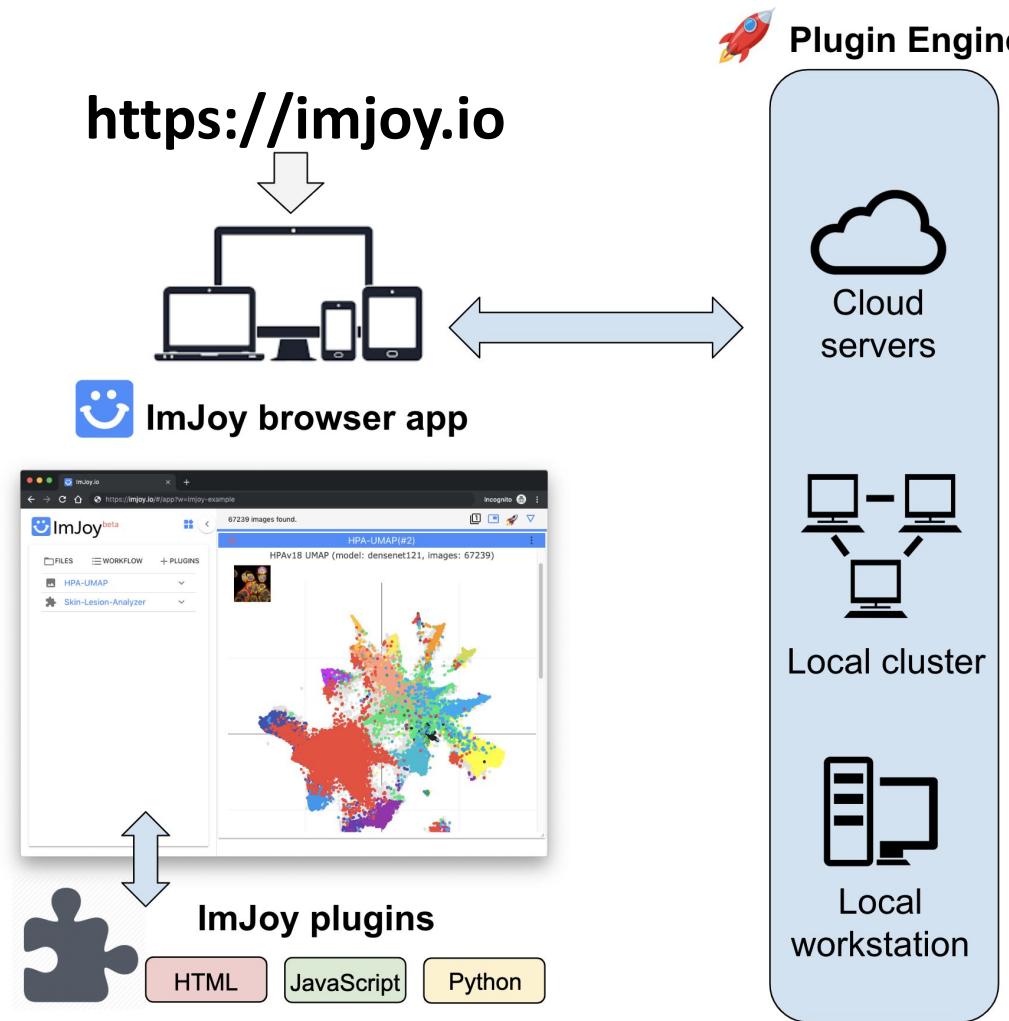


In collaboration with Manu Prakash group at Stanford and Heidstar CO., LTD



# ImJoy AI computing platform

Supercharging interactivity and scalability in AI-powered life science



**ImageJ.JS: ~1000 users / day**



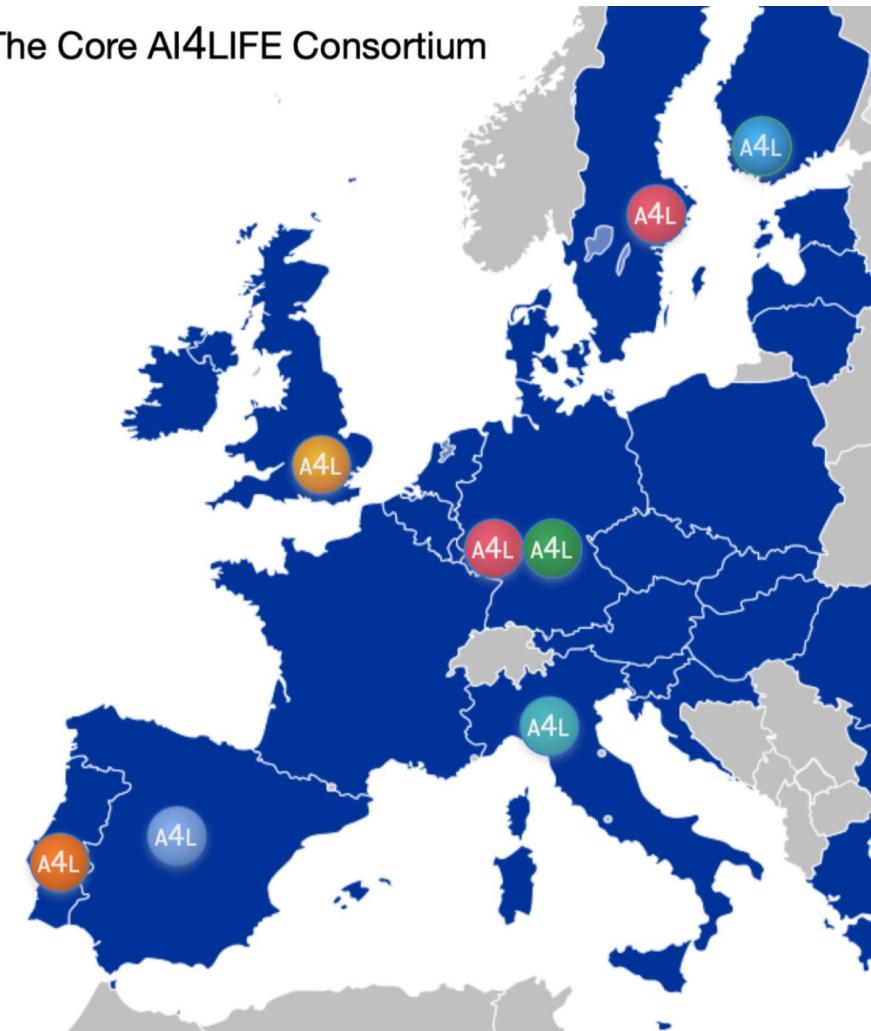
# AI4Life Consortium



- Making AI models more FAIR
- Supported by EU Horizon grant
- Since 2019, with EMBL, HT, Euro Bioimaging...



The Core AI4LIFE Consortium



<https://ai4life.eurobioimaging.eu/>



# Biolmage.IO

- Model description file standard
  - Repository for sharing models
  - Continuous integration for model testing
  - Cloud-based model serving and test run



**Join us as a community partner!**



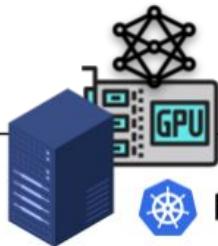
# Preprint for the BioImage Model Zoo

<https://doi.org/10.1101/2022.06.07.495102>

<https://bioimage.io>

The screenshot shows the BiolImage Model Zoo homepage. At the top, there's a navigation bar with icons for upload, documentation, and about. Below the header, the title "BiolImage Model Zoo" is displayed in large letters, followed by the subtitle "Advanced AI models in one-click". There are several promotional text blocks: "Integrate with Fiji, Ilastik, ImJoy and more", "Try model instantly with BioEngine", "Contribute your models via Github", and "Link models to datasets and applications". A prominent button labeled "Explore the Zoo" is visible. The page features a "Community Partners" section with logos for #ZeroCostDL4Mic, Ilastik, BioEngine, Smiley, DeepImageJ, and a plus sign icon. Below this is a search bar with filters for "All", "models", "applications", and "datasets", along with a keyword input field and a "Tags & Filters" dropdown. The main content area displays eight AI model cards, each with a thumbnail image, a title, a brief description, tags, and a license link. The models shown are:

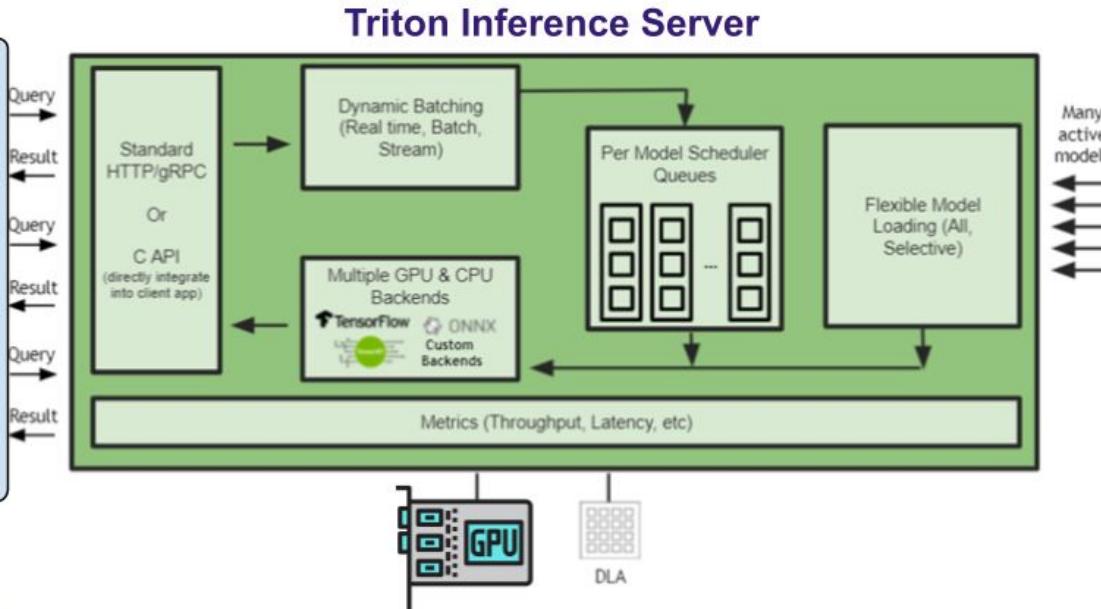
- Mitochondria Segmentation ... (Electron microscopy, mitochondria, 3d, CC-BY-4.0)
- Neuron Segmentation in EM ... (Neuron segmentation in EM, trained on CREMI challenge data, unet, neurons, instance-segmentation, electron-microscopy, CC-BY-4.0)
- Nuclei Segmentation Bound... (Fluorescence-light-microscopy, nuclei, unet, instance-segmentation, CC-BY-4.0)
- LiveCellSegmentationBound... (Cell segmentation for phase-contrast microscopy, 2d, transmission-light-microscopy, label-free, cells, CC-BY-4.0)
- PlatynereisEMcellsSegment... (Cell segmentation in EM of platynereis, unet, cells, instance-segmentation, electron-microscopy, CC-BY-4.0)
- PlatynereisEMnucleiSegment... (Organelle segmentation in EM of platynereis, unet, nuclei, instance-segmentation, electron-microscopy, CC-BY-4.0)
- HPA Cell Segmentation (DP...) (Cell segmentation model for segmenting images from the Human Pro..., nucleus-segmentation, editable, CC-BY-4.0)
- Cell Segmentation from Me... (A 3D U-net to predict cell membranes in plant tissues, trained on ..., zero-costdl4mic, deepimagej, segmentation, 3d, CC-BY-4.0)



# BioEngine: Scalable AI Model Serving

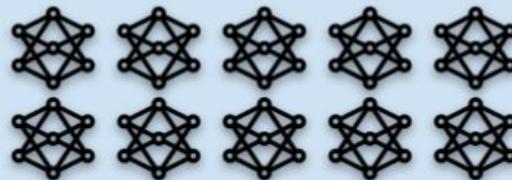
kubernetes

**Hypha**



## Model Repository

Biolimage.IO



## Desktop Clients



Users with thin client computers

## Web Clients



No Installation Required!



# About the course

More information at <https://ddls.aicell.io>

# What do you do with ChatGPT?



## What do you do with ChatGPT?

17 Responses

code faster

- writing reports know things quickly
- learn language writing my thesis
- explaining difficult stuf rewrite emails
- write thesis history to study
- write emails math
- coding summarize papers
- summarize materials
- learn abt hard topics
- get instruction on method

# Tips for using ChatGPT

---



- Set a role
- Be specific
- Provide examples
- Use formatting
- Avoid assumptions
- Check limitations
- Try different approaches
- Watch for repetition
- Limit prompt length
- Correct gently
- Give feedback



# Assignments

See here: <https://ddls.aicell.io/course/ddls-2023/module-1/#lecture-tuesday>