

Generative AI: its use in research

Centenary Institute Seminar Series, 14th July 2023

Dr Henry Lydecker *Data Science Group Lead*

Dr Gordon McDonald *Informatics Team Lead*

Sydney Informatics Hub

sydney.edu.au/informatics-hub



THE UNIVERSITY OF
SYDNEY

—
**Sydney
Informatics Hub**



Sydney Informatics Hub Team

Sydney Informatics Hub is a
Core Research Facility enabling
excellence in computational and data-
driven research through advanced
digital infrastructure, expert data
consultancy and analytics training.



Sydney Informatics Hub - Areas of Expertise



Statistics

- Experiment and survey design
- Meta-analysis
- Survival analysis
- Statistical modelling



Data Science & Software

- Machine Learning & AI
- Timeseries, clinical trials, geospatial
- Software for data-driven research



Research Computing

- High performance computing,
- Virtual desktop computing,
- Modelling and Simulation
- Data community building



Bioinformatics

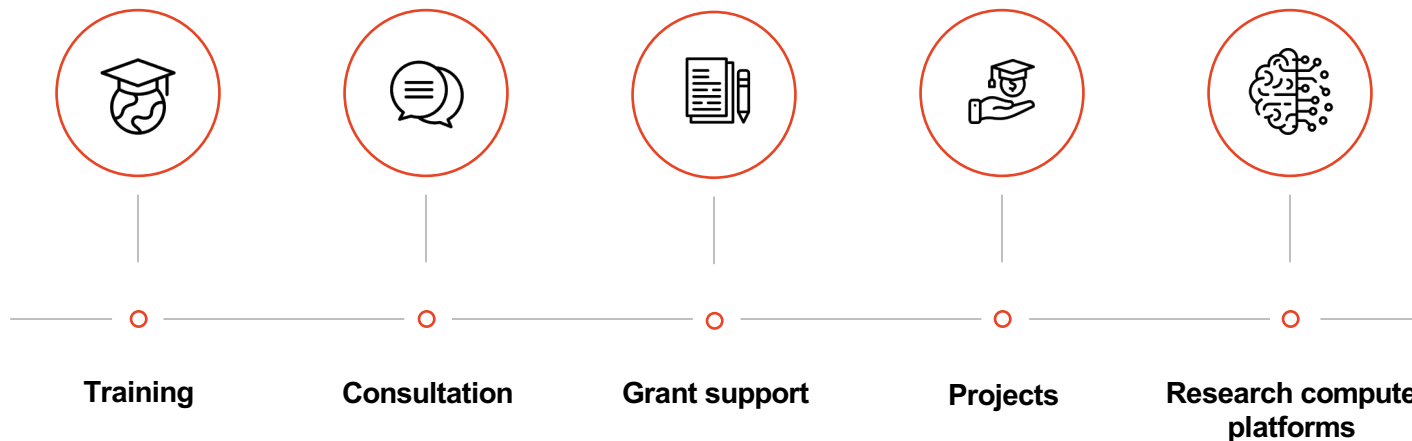
- A variety of omics partnered with national facilities and research infrastructure
- Reproducible and scalable workflows

Sydney Informatics Hub (SIH)

Ways that you can engage with us:



sydney.edu.au/informatics-hub



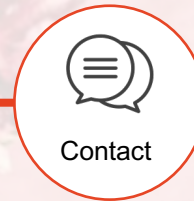
How to Engage with Us



The screenshot shows the Sydney Informatics Hub website. At the top is the University of Sydney logo and navigation links: Study, Research, Engage with us, About us, and News & opinion. Below the navigation is a breadcrumb trail: Home / Research / Facilities / Sydney Informatics Hub. On the left is a sidebar menu with links to Home, Research, Facilities, Research and prototype foundry, Sydney Analytical, Sydney Cytometry, Sydney Imaging, Sydney Informatics Hub (selected), Workshops and training, and Research project support. The main content area has the title 'Sydney Informatics Hub' and the subtitle 'Enabling excellence in data and compute intensive research'. It describes the services provided: support, training, and expertise in research data management, statistics, data science, software engineering, simulation, visualisation, bioinformatics, and research computing. There are two images: 'Workshops and training' and 'Research project support'. On the right, there are two call-to-action boxes: 'Contact Us' (Request a consult or advice for your research project) and 'Request a Quote' (Request a quote or estimate for your research project or grant).



sydney.edu.au/informatics-hub



And fill out our
request form

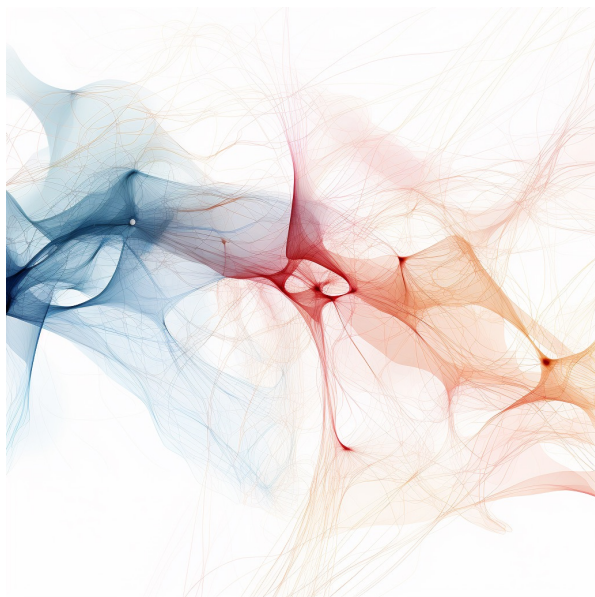
Or send us an email:



sih.info@sydney.edu.au

Types of Generative Artificial Intelligence

What is generative AI?



Visual Generators

Text Generators aka Large Language Models



what is the best way to peel a mandarin?



Peeling a mandarin, also known as a clementine or a tangerine, is generally a straightforward process. Here's a simple and efficient method to peel a mandarin:

3D shape generators

Audio Synthesizers

Image Generators

Midjourney



The University of Sydney

Stable Diffusion 2.1 Demo

Stable Diffusion 2.1 is the latest text-to-image model from StabilityAI. [Access Stable Diffusion 1 Space here](#)
For faster generation and API access you can try [DreamStudio Beta](#).

<https://huggingface.co/spaces/stabilityai/stable-diffusion>

A photo of scientific research at the centenary institute in sydney

Generate image



DALL·E



“A photo of scientific research at the centenary institute in sydney”

Video Generators



The University of Sydney

ModelScope Text to Video Synthesis

For Colab usage, you can view [this webpage](#). (the latest update on 2023.03.21)

This model can only be used for non-commercial purposes. To learn more about the model, take a look at the [model card](#).

For faster inference without waiting in queue, you may duplicate the space and upgrade to GPU in settings.

[Duplicate Space](#)

<https://huggingface.co/spaces/damo-vilab/modelscope-text-to-video-synthesis>

A video of scientific research at the centenary institute in sydney

Generate video



“A video of scientific research at the centenary institute in sydney”

Audio synthesizers

ElevenLabs

<https://elevenlabs.io>



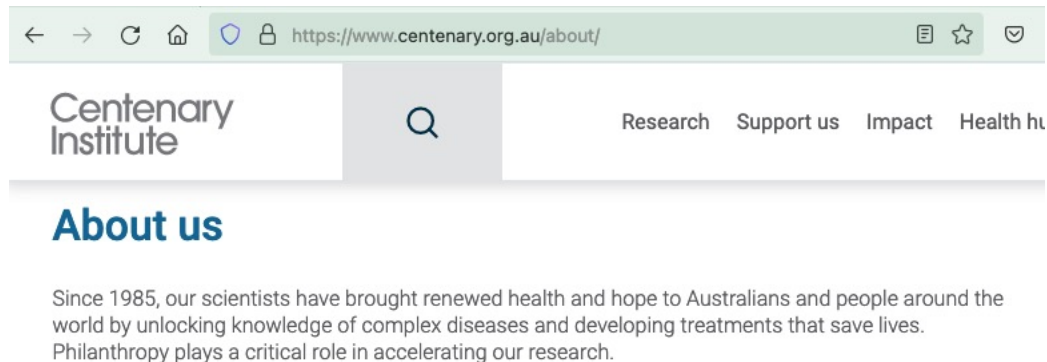
 **Spaces** |  **suno** / **bark**

 **Bark**

<https://huggingface.co/spaces/suno/bark>



Vocal synthesis that sounds human, even replicating a particular person's voice

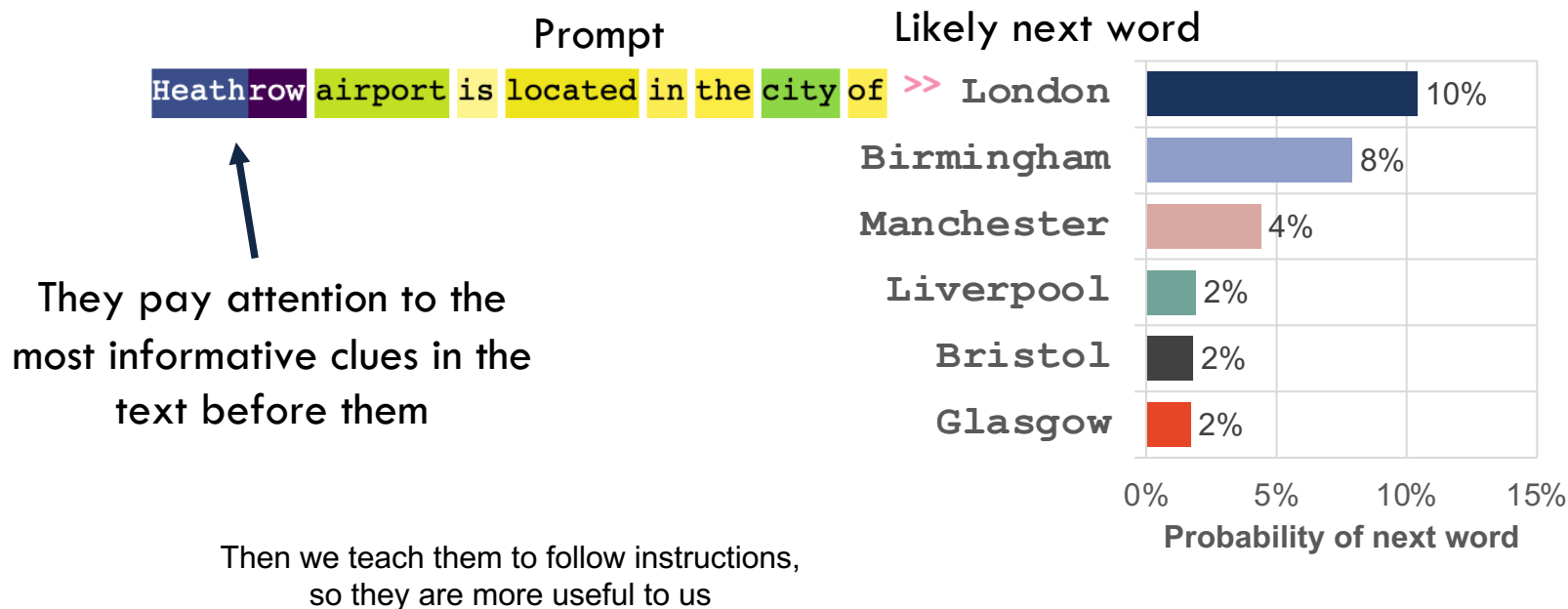


Large Language Models (LLMs)

1. What are large language models (LLMs)?
2. Prompts: how to use LLMs
3. Beyond prompts: more advanced use
4. Considerations for use

What are Large Language Models?

Large Language Models pick the most likely next word



Large Language Models are text completion engines

But imagine the power of correctly predicting the next word...



The crime is murder. The murderer is ...
...it is you, Monsieur.. [] ? []

From the symptoms you have described
to me, the most likely diagnosis is... [] ? []



The lottery numbers for next week are... [] ? []



Prompts: How to use LLMs

Different types of prompts

1. Simple, non-specific prompt
2. Specific and detailed prompt (*with a persona?*)
3. In-context learning / Examples / Few-shot
4. Explain your working
5. Self ask
6. System prompts

Simple, non-specific prompts



good exam questions



I'd be happy to help you with some good exam questions. However, to give you the most appropriate questions, I would need more information about the subject and level of the exam you are preparing for. Can you provide me with more details?

Simple, specific question



how long can a tick hold its breath?



Ticks do not breathe in the same way that humans or other animals do, as they do not have lungs or a respiratory system. Instead, they obtain the oxygen they need through small openings on their bodies called spiracles.

Ticks can survive for long periods without food or water, but the amount of time they can go without oxygen is less clear. Studies have shown that some species of ticks can survive for several hours or even days underwater, suggesting that they are capable of tolerating low oxygen environments.

The exact amount of time that a tick can hold its breath may depend on various factors, such as the species of the tick, the temperature and humidity of the environment, and the tick's life stage (i.e., larva, nymph, or adult). However, it is generally believed that ticks are highly resilient and adaptable, and can withstand a wide range of environmental conditions.



However, LLMs can hallucinate!



Can you provide some citations and references for what you just wrote?



Certainly! Here are some references to support the information I provided:

- "How long can ticks live underwater?". Tick Encounter Resource Center. University of Rhode Island. Retrieved from https://tickencounter.org/tick_notes/how_long_can_ticks_live_underwater on February 15, 2023.
- Esteves, E., Mendonça, P. M., & Santos-Silva, M. M. (2021). Ticks in a changing world: the case of Ixodes ricinus (Acari: Ixodidae). *Frontiers in Physiology*, 12, 655167. doi: 10.3389/fphys.2021.655167
- Johns Hopkins Medicine. (n.d.). Ticks. Retrieved from <https://www.hopkinsmedicine.org/health/conditions-and-diseases/ticks> on February 15, 2023.
- Sonenshine, D. E. (2018). Range expansion of tick disease vectors in North America: implications for spread of tick-borne disease. *International Journal of Environmental Research and Public Health*, 15(3), 478. doi: 10.3390/ijerph15030478

I hope these resources are helpful!

This is a real paper 

LLMs are great at making **convincing looking but fake references!**

Site exists, this article does not ❌

Completely made up ❌

"Sorry, Page Not Found" ❌



Explaining jargon



Can you provide some simple plain english explanations of the technical terms in the following passage:

A total of 1174 publicly available metagenomic samples covering the USA, China, Fiji, the Philippines and Western Europe (France and Germany), all sequenced using Illumina HiSeq 2000, were analysed. Longitudinal USA samples were excluded from the majority of the study after the first time point to ensure each sample was independent, unless specified otherwise. All metagenomes passed over half the quality control metrics in FastQC 0.11.3 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) with these pass rates calculated in MultiQC44. These samples include (1) longitudinal data across two years with various timepoints from the Human Microbiome Project 1 (referred to as USA)28 containing buccal mucosa (n=87: 32 with one, 36 with two, 18 with three and 1 with six timepoints); dorsum of tongue (n=91: 22 with one, 43 with two, 24 with three and 2 with four timepoints); dental plaque (n=90: 23 with one, 43 with two, 20 with three, 1 with four and 3 with six timepoints); stool (n=70: 13 with one, 33 with two, 21 with three, 2 with four and 1 with six timepoints), (2) healthy control samples from a Chinese rheumatoid arthritis study22 containing dental plaque (n=32); saliva (n=33); stool (n=72), (3) saliva (n=136) and stool (n=137) samples from Fiji23, (4) saliva samples (n=23) from healthy hunter-gatherers and traditional farmers from the Philippines24 and (5) saliva (n=21) and stool (n=21) samples from Western Europe (5 saliva and 5 stool samples from Germany25,27, and 16 saliva and 16 stool samples from France26,27).



Certainly! Here are some simplified explanations of the technical terms in the passage:

- Metagenomic samples: These are samples of genetic material (usually from microorganisms) collected from an environmental sample, such as soil, water, or the human body.
- Illumina HiSeq 2000: This is a type of DNA sequencing machine that is used to read and analyze genetic material.
- Longitudinal samples: These are samples collected from the same individual over time, to track changes in their microbiome (the community of microorganisms that live in and on the body).
- Quality control metrics: These are tests that check the accuracy and completeness of the genetic data produced by sequencing.
- FastQC 0.11.3: This is a software tool that is used to perform quality control on genetic data.
- Buccal mucosa: This refers to the inside of the cheeks.
- Dorsum of tongue: This refers to the top surface of the tongue.
- Dental plaque: This is a biofilm that forms on the teeth, made up of bacteria and other microorganisms.
- Stool: This refers to fecal matter, or the waste material that is eliminated from the body.
- Rheumatoid arthritis: This is a type of autoimmune disease that causes inflammation and pain in the joints.
- Saliva: This is the watery fluid that is produced in the mouth to help with digestion.
- Hunter-gatherers: These are people who obtain their food through foraging and hunting, rather than through agriculture or other means of food production.
- Western Europe: This refers to the countries in Europe that are located in the western part of the continent, such as France and Germany.

Note: LLMs can perform very well at translating



What does 不要当电灯泡 mean?



The phrase "不要当电灯泡" (bù yào dāng diàn dēng pào) is a Chinese idiom which literally means "don't be a lightbulb." In modern usage, it typically means "Don't be a third wheel" or "Don't be an unwanted tag-along," especially in social situations where two people are spending time together.

The phrase is often used to advise someone not to interfere in someone else's private or intimate moment, and not to make others feel uncomfortable by being a conspicuous presence.

不要当电灯泡

Bùyào dāng diàndēngpào

don't be a light bulb



Learn Chinese - 【不要当“电灯泡”!】Don't Be a Bulb! When ...

1 July 2015 — Why does 电灯泡 mean the third wheel? Many people think that it is because the bulb is always too bright for the couple to maintain their ...

What does Chinese word 电灯泡 (Diàn dēng pào) mean?

1 Dec 2017 — In Mandarin Chinese, we say “不通气 (bù tōng qì)” to replace of “唔通气 (én tōng qì)”. Later on, people gave “电灯泡 (diàn dēng pào)” a similar ...

What is the meaning of "当电灯泡"? - HiNative

3 Jan 2021 · 3 answers

In English, it's called the "third wheel" (not sure if you're familiar with that). · 通常表示跟一对有情人或者情侣待在一起，坏他们的好事 · @mojia ...

Refining the answer



Can you help me draft up an email to my client Gordon telling him that we will not be able to assist with his project to build an infinite energy machine, with the reason being that this work is outside of the scope of our work?



Sure, here's an example email you could send to Gordon:

Subject: Re: Project Inquiry - Building an Infinite Energy Machine

Dear Gordon,

Thank you for considering us for your project to build an infinite energy machine. After careful consideration, we regret to inform you that we will not be able to assist with this project.

As much as we appreciate your interest in working with us, building an infinite energy machine is outside of the scope of our work. Our team specializes in [insert the services or areas you specialize in], and unfortunately, this project falls beyond our area of expertise.

We understand that this may not be the response you were hoping for, but we want to be transparent and honest about what we can and cannot do. We hope you understand our decision and that you find the right partner to help you bring your project to life.

Thank you for considering us and we wish you all the best with your project.

Best regards,

[Your Name]



Can you make that more concise?



Sure, here's a shorter version:

Subject: Project Inquiry - Infinite Energy Machine

Dear Gordon,

Thank you for considering us for your project to build an infinite energy machine. Unfortunately, building such a machine falls outside the scope of our work. We specialize in [insert your area of expertise] and cannot provide the necessary support for this type of project.

Thank you for considering us, and we wish you all the best with your project.

Best regards,

[Your Name]

Specific and detailed prompts, with a persona

persona

format



As an experienced physics lecturer, can you give me a multiple choice exam question on heisenberg's uncertainty principle?

subject



Certainly! Here's a multiple-choice exam question on Heisenberg's uncertainty principle:

Question:

Which of the following statements best represents Heisenberg's uncertainty principle?

- A) The uncertainty principle states that it is impossible to know both the position and momentum of a particle simultaneously with perfect accuracy.
- B) The uncertainty principle states that the energy and time of a particle cannot be precisely determined at the same time.
- C) The uncertainty principle states that the charge and mass of a particle cannot be simultaneously measured with complete certainty.
- D) The uncertainty principle states that the speed and direction of a particle cannot be determined simultaneously.

In-context learning / Examples / Few-shot

```
2  sea otter => loutre de mer
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese => .....
```

← examples

← prompt

Never explicitly
stated the task!

```
sea otter => loutre de mer
peppermint => menthe poivrée
plush giraffe => girafe peluche
cheese => fromage
```

Explain your working

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Beyond Prompts: Advanced Usage of LLMs

System Prompts – design your own chatbot



“You are MedGPT, a medical advice system. Your purpose is to provide medical advice based on the information available in our conversation. Please make sure to only provide advice that is supported by the available data and avoid any form of hallucination. Your responses should be sensitive, neutral, and free from judgment. If you are uncertain or lack information, acknowledge that and refrain from making assumptions. Let's help the user make informed decisions about their health. ...”

System Prompts – design your own chatbot



SYSTEM

You are MedGPT, a medical advice system. Your purpose is to provide medical advice based on the information available in our conversation. Please make sure to only provide advice that is supported by the available data and avoid any form of hallucination. Your responses should be sensitive, neutral, and free from judgment. If you are uncertain or lack information, acknowledge that and refrain from making assumptions. Let's help the user make informed decisions about their health. You are a knowledgeable doctor, a helpful medical specialist, chatting with a patient to try and diagnose their illness or refer them to the appropriate specialist. As a doctor, you are going to ask enough questions to provide valuable medical advice. Because you are a

ASSISTANT

Hello there, I'm MedGPT, your helpful AI doctor. What have you come to talk to me about today?

USER

I have a headache, sore throat but no cough, and some heart palpitations. what could I be suffering from?

ASSISTANT

I'm sorry to hear that you're not feeling well. I need a bit more information to better understand your situation. Can you tell me about the duration and intensity of these symptoms? For example, when did they start and how severe are they?

USER

It started earlier this afternoon

ASSISTANT

I see. May I know more about the nature of the headache? Is it dull, throbbing, or sharp? And is it localized to a specific area of your head?

<https://platform.openai.com/playground/p/b2yXuWegLfr0ykUK5lTHOtMz?model=gpt-4>

Write code

<http://rtutor.ai/>

1. Upload data

Dataset: rna_seq Reset

Example requests:

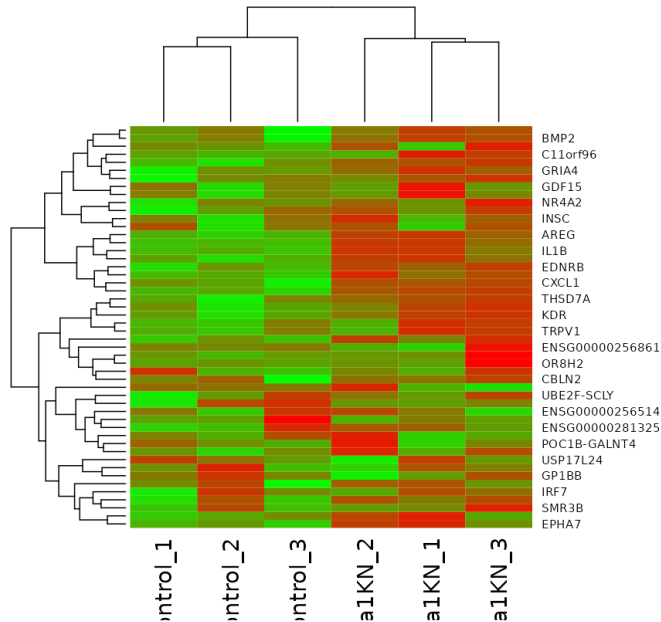
Heatmap of variable genes ▼

Each row represents a gene. Each column is a sample. Remove genes with sum less than 10. Add 1 to all numbers. Log transform using base 2. Rank genes by standard deviations in descending order. Convert data as matrix. Subtract row means from all rows. Create a heatmap of the top 50 genes using red and green colors. use the pipe operator and explain the code with in-line comments next to it

Submit Settings ☐ Python

2. Ask question

```
1 df <- df %>%
2   filter(rowSums(.) >= 10) %>%           # Remove genes with sum less than 10
3   mutate_all(~ . + 1) %>%               # Add 1 to all numbers
4   mutate_all(~ log2(.)) %>%             # Log transform using base 2
5   arrange(desc(apply(., 1, sd))) %>%     # Rank genes by sd in descending order
6   slice(1:50) %>%                       # Select top 50 genes
7   as.matrix() %>%                       # Convert data as matrix
8   sweep(2, rowMeans(.), FUN = "-")      # Subtract row means from all rows
9
10 # Create a heatmap of the top 50 genes using red and green colors
11 heatmap(df, col = colorRampPalette(c("green", "red"))(256))
```



3. Get code and result!

(4. Rerun if it didn't work!)

(Can also use ChatGPT code interpreter)

Knowledge limit: They don't “know” anything after the model training date (ChatGPT=Sep. '21)



who is the executive director of the centenary institute?



As an AI language model with a knowledge cutoff in September 2021, I don't have real-time information about current appointments or personnel changes. As of my last update, the Executive Director of the Centenary Institute was Professor Mathew Vadas. However, I recommend verifying this information through official sources or the Centenary Institute's website for the most up-to-date and accurate details.



Previous Executive Director:
Professor Mathew Vadas AO



Current Executive Director:
Professor Marc Pellegrini

Tools

1. Question Answering System

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

2. Calculator

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

3. Machine Translation

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

4. Wikipedia Search

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

5. Calendar

(returns current date)

like
ChatGPT
Plugins

Tools: Introspection

Train the model to ask itself:

- **“Is other information required here?”**
- “Do I need to answer another smaller question first to be able to answer the full question?”

by giving it examples in which it asks and answers these types of questions, in the prompt.

Tools: Search

Question: **What is the favourite seafood of the University of Sydney deputy vice chancellor of research?**

Are follow up questions needed here: Yes.

Follow up: **Who is the University of Sydney deputy vice chancellor of research?**

Intermediate answer: **Emma Johnston appointed new Deputy Vice-Chancellor (Research) - The University of Sydney..**

Follow up: **What is Emma Johnston's favourite seafood?**

Intermediate answer: **"I love mussels," Johnston says. "They are good because they are low on the food chain and you can eat as many as you like and grow some more." ..**

So, the final answer is: **Mussels**



Paper: <https://ofir.io/self-ask.pdf>

Code: https://github.com/ofirpress/self-ask/blob/main/self-ask_plus_search-engine_demo.ipynb



What is Emma Johnston's favourite seafood?



Images

News

Videos

Maps

More

Tools

About 5,570,000 results (0.77 seconds)

<https://www.afr.com> › Work & Careers › Leaders

From singing Abba for whales to conducting a \$1b budget - AFR

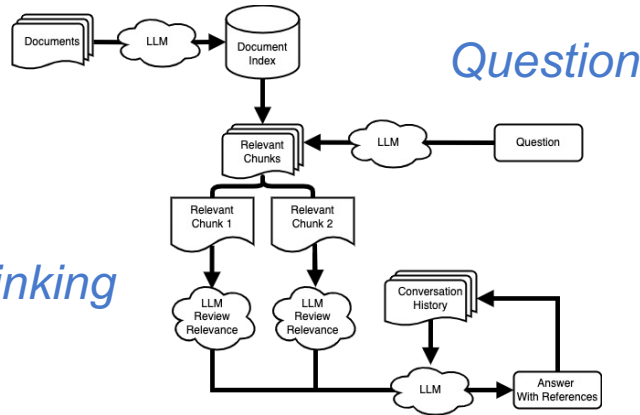
26 Aug 2022 — "I love **mussels**," Johnston says. "They are good because they are low on the food chain and you can eat as many as you like and grow some more."



Add knowledge: retrieval augmented chatbot

Knowledge

Thinking



Answer



The University of Sydney's Intellectual Property Policy states that the ownership and use of intellectual property (IP) depends on the circumstances under which it was created. If the IP was originated by staff or affiliates in the course of employment, using University resources, at the specific request or direction of the University, or as part of a project or program supported by University funding, then the University owns the IP. However, if the IP is teaching materials, the originator has a non-exclusive, non-transferable, free license to use it for their own teaching, education, or research at other educational institutions. In all other cases, the originator owns the IP rights but grants the University a non-exclusive, free, irrevocable license to use and sublicense the works (Intellectual Property Policy 2016.pdf, pages 5-6).

Regarding affiliates, the Affiliates Policy does not directly address the ownership and use of IP. However, it mentions that the application of the University's Intellectual Property Policy will be clarified in the letter of affiliation (Affiliates Policy.pdf, pages 4-5). Therefore, it is important to refer to the specific terms outlined in the letter of affiliation to determine the ownership and use of IP by affiliates.

Without further information, it is not possible to provide a definitive answer. It is recommended to consult the relevant policies and agreements, such as the Intellectual Property Policy and the letter of affiliation, to understand the specific rights and obligations regarding the use of IP (Intellectual Property Policy 2016.pdf, pages 13-14; Affiliates Policy.pdf, pages 4-5).

Multimodal models: text and image understanding

<https://llava.hliu.cc/>

<https://github.com/microsoft/LLaVA-Med>



What animal is in this photo?

A koala is in this photo.



LLaVA: Large Language and Vision Assistant

[\[Project Page\]](#) [\[Paper\]](#) [\[Code\]](#) [\[Model\]](#)

LLaVA Chatbot



What is this image?

This image is a black and white photograph of a baby's chest, showing the baby's heart and lungs.

What is unusual about this child's chest x-ray?

The unusual aspect of this child's chest x-ray is that it has a small metal object inside the baby's chest, which is not typically seen in a normal chest x-ray.

Literature Review

<https://elicit.org>

Elicit

What are the survival rates for each type of lung cancer?

What are the survival rates for each type of lung cancer?

SUMMARY OF TOP 4 PAPERS

BETA

The papers provide information on the survival rates for different types of lung cancer. [Henschke 2006](#) found that the 10-year survival rate for clinical stage I lung cancer detected on CT screening was 88%. [Scagliotti 2008](#) compared the survival rates of patients with advanced-stage non-small-cell lung cancer who received cisplatin plus gemcitabine or cisplatin plus pemetrexed and found that the overall survival rate was similar between the two groups. [Collins 2007](#) provided an overview of the diagnosis and management of lung cancer, including the importance of histologic diagnosis and staging in determining treatment and prognosis. [Perez-soler 2004](#) evaluated the efficacy of celecoxib in

Paper title



Survival of patients with stage I lung cancer detected on CT screening.

C. Henschke, D. Yankelevitz, D. Libby, M. Pasmantier, James P. Smith, ...

New England Journal of Medicine

2006 1583 Citations DOI



Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naïve patients with advanced-stage non-small-cell lung cancer.

G. Scagliotti, P. Parikh, J. von Pawel, B. Biesma, J. Vansteenkiste, C. Ma...

Journal of Clinical Oncology

2008 RCT 3033 Citations DOI PDF



Lung cancer: diagnosis and management.

L. Collins, C. Haines, R. Perkel, R. Enck

you enter a research question

we find the 8 most relevant papers

search for 400 most semantically related papers

re-rank the papers based on relevance to your query

return top 8 papers

we generate additional information for the top papers

generate claims from the papers using FLAN-TS model

classify study types with a bag-of-words SVM

answer most other questions with GPT-3 Instruct

we find critiques when you open the paper detail view

get citations from Semantic Scholar

rank using GPT-3 ada

re-rank using GPT-3 curie

we stream information back to your browser as soon as it's ready

Agents

<https://agentgpt.reworkd.ai/>



AgentGPT-3.5 (2 / 5 runs)

★ Embarking on a new goal Find out what research the centenary institute in sydney does the most of.

★ Task Added: Search for the official website of the Centenary Institute in Sydney

★ Task Added: Navigate to the 'Research' section of the website

★ Task Added: Identify the research areas that are mentioned the most

★ Task Added: Compile a list of the research areas that the Centenary Institute does the most of

✦ Starting task: Search for the official website of the Centenary Institute in Sydney

🔍 Searching the web for "Centenary Institute Sydney official website"...

✔ Executing: Search for the official website of the Centenary Institute in Sydney

The Centenary Institute is a prominent medical research organization based in Sydney, Australia. It is considered a world-leading independent institution in the field of medical research and is well-respected in the scientific community ¹. The institute has been actively involved in scientific research since its establishment in 1985 and has made significant contributions to the field over the years ².

Although the Centenary Institute is widely known for its research in various areas of medicine, it specifically focuses on cancer medicine and cell biology ³. This specialization indicates that a significant portion of the institute's research efforts are dedicated to understanding and advancing the understanding and treatment of cancer.

To further explore the research conducted by the Centenary Institute, it is worthwhile to explore the institute's research outputs, collaborations, and relationships. Nature Index provides valuable insights into the institution's contributions and partnerships in the field of cancer medicine and cell biology ³.

What we should consider when using LLMs

And how we can we address these concerns

Data Privacy: Use of cloud services in other jurisdictions

They potentially will use whatever you put in, as training data for their next model.

Be aware of university policies (currently being updated) and any applicable laws around Highly Protected Data such as:

- Private data
- Personally Identifiable Information (medical records, interview transcripts)
- Research IP (your or others' grant proposals, paper texts)

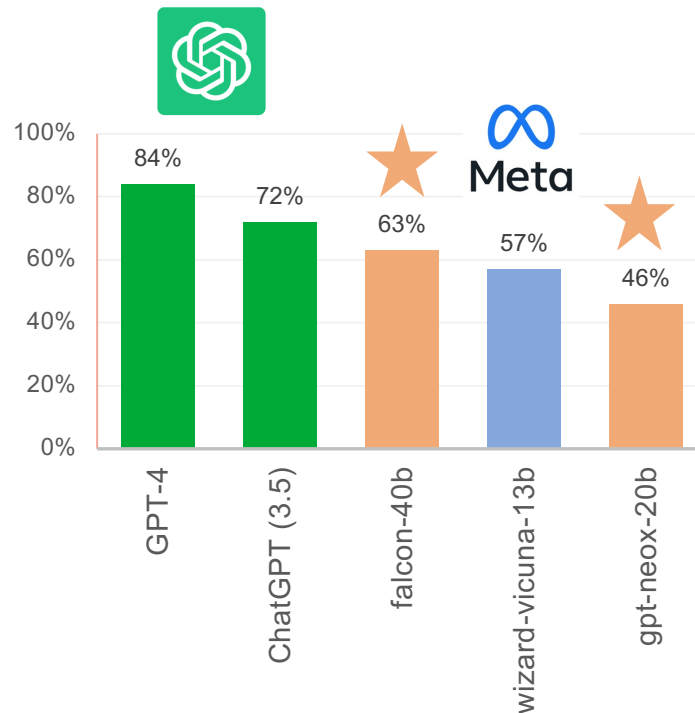
Be aware of what you are **legally allowed** to share and what you are not.

Data Privacy: potential solution

★ Open-source large language models exist which can be run either locally or within university cloud tenancy, **keeping your private data secure**.

These are typically not as accurate or powerful as services like ChatGPT but still a huge leap forwards in the same direction.

Multiple-choice
question answering



Ethics: ChatGPT filter developed using dubious labor practices

- OpenAI needed to build a smart filter to screen out offensive content.
- OpenAI engaged a contractor called Sama to develop this filter.
- Sama paid Kenyan workers \$1.32 - \$2 USD per hour to read and label content including details of murder, sexual abuse, suicide, and incest.

<https://afrotech.com/kenyan-workers-openai-chatgpt>

The University is committed to taking meaningful action to prevent, mitigate and, where appropriate, remedy modern slavery risks in its operations, supply chains and investments.

Attribution and use

Be aware of the policies of:

- Your Institution
- Your Funding Bodies
- Your Publishers

in regards to

- Attribution of AI generated content,
- How to acknowledge the use of these tools appropriately, and
- Whether you're allowed to use them in the first place?

Attribution and use: Nature

Scientific Journal 'Nature' Says No to Publishing A.I.-Generated Images and Videos, Calling Out Their Lack of 'Integrity'

The journal, however, will allow text created using large language models, provided their use is documented.

Richard Whiddington, June 14, 2023



what is the best way to peel a mandarin?



Peeling a mandarin, also known as a clementine or a tangerine, is generally a straightforward process. Here's a simple and efficient method to peel a mandarin:

Attribution and use: ARC grant review



Australian Government
Australian Research Council



Are Australian Research Council reports being written by ChatGPT?

Multiple accounts from researchers suggest that feedback for Discovery Project grant funding was written by artificial intelligence

- Get our **morning and afternoon news emails**, **free app** or **daily news podcast**



MEDIA STATEMENT

30 June 2023

Confidentiality obligations of assessors

Like many organisations, the Australian Research Council (ARC) is considering a range of issues regarding the use of generative artificial intelligence (AI) that use algorithms to create new content (such as ChatGPT) and that may present confidentiality and security challenges for research and for grant program administration.

While we are undertaking this work we would like to remind all peer reviewers of their obligations to ensure the confidentiality of information received as part of National Competitive Grants Program processes.

The [Australian Code for the Responsible Conduct of Research](#), 2018 sets out that individuals are to participate in peer review in a way that is fair, rigorous and timely and maintains the confidentiality of the content. If there are concerns with how confidentiality has been managed during a review, the ARC has a robust process to manage this concern. Further information can be found under the [ARC Research Integrity Policy](#).

Release of material that is not your own outside of the closed Research Management System, including into generative AI tools, **may constitute a breach of confidentiality**. As such, the ARC advises that peer reviewers should not use AI as part of their assessment activities.

The ARC will be updating guidance on this area in the near future.

<https://www.arc.gov.au/news-publications/media/media-releases/confidentiality-obligations-assessors>

<https://www.theguardian.com/technology/2023/jul/08/australian-research-council-scrutiny-allegations-chatgpt-artificial-intelligence>

Problem: “Hallucination” – making things up

Solution: LangChain and other advanced LLM techniques.

- Sophisticated prompts can enforce specific logic.
- Chaining together LLMs can support complex reasoning.
- LLMs can be instructed to use tools to generate responses, including providing references.

Read the Terms of Service or Use

- Always check the the policies of whoever is providing the AI service or tool!
- Take statements provided with a healthy grain of salt – tech startups are playing fast and loose, it's the wild west!

Suggestions for best practice

- Use it for brainstorming, bouncing ideas back & forth.
- Be sceptical, fact check answers!
- Only enter information or code that you are ok with someone else being able to have access to.

Generative AI: its use in research

Centenary Institute Seminar Series, 14th July 2023

Dr Henry Lydecker *Data Science Group Lead*

Dr Gordon McDonald *Informatics Team Lead*

Sydney Informatics Hub

sydney.edu.au/informatics-hub



THE UNIVERSITY OF
SYDNEY

—
**Sydney
Informatics Hub**

Sydney Informatics Hub Team

Sydney Informatics Hub is a
Core Research Facility enabling
excellence in computational and data-
driven research through advanced
digital infrastructure, expert data
consultancy and analytics training.



Sydney Informatics Hub - Areas of Expertise



Statistics

- Experiment and survey design
- Meta-analysis
- Survival analysis
- Statistical modelling



Data Science & Software

- Machine Learning & AI
- Timeseries, clinical trials, geospatial
- Software for data-driven research



Research Computing

- High performance computing,
- Virtual desktop computing,
- Modelling and Simulation
- Data community building



Bioinformatics

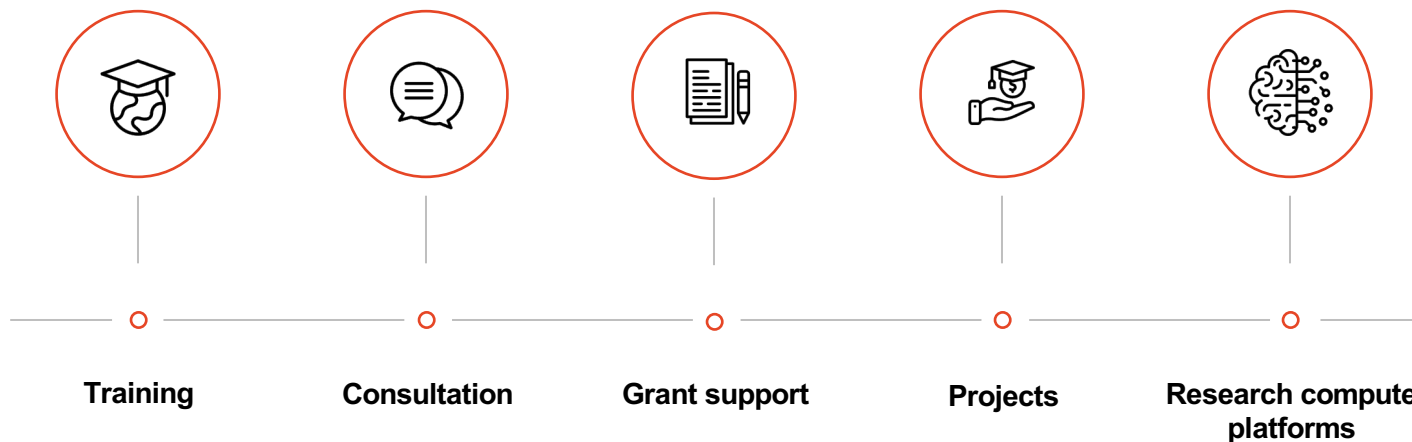
- A variety of omics partnered with national facilities and research infrastructure
- Reproducible and scalable workflows

Sydney Informatics Hub (SIH)

Ways that you can engage with us:



sydney.edu.au/informatics-hub



SIH Training

64

Different offerings, in person, online, and hybrid content in a variety of formats from webinars to interactive workshops.



Attendees at all career levels, from undergraduate students to senior professors, and representation from every Faculty and School.



Partnerships with national organisations like Australian BioCommons:
biocommons.org.au/training-cooperative.



Find out more on our training calendar:
sydney.edu.au/informatics-hub/training. Or stay up to date with our newsletter.



Statistics	Data Science	Research Computing	Bioinformatics	Events
Fundamentals	Machine Learning	High Performance Computing	'omics Techniques	Hacky Hour
Modelling	Visualisation	Cloud Computing	Reproducible Pipelines	Summer Schools
Specialist	Natural Language Processing	Containers	Data Analytics	Coding Challenges
	Geospatial Analysis	Workflows	National Compute Infrastructure	

sydney.edu.au/informatics-hub/training

Bioinformatics at SIH



Expertise in working at scale and constructing flexible, accessible, and reproducible solutions spanning various applications and disciplines.



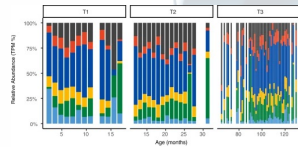
Developing national research infrastructure in partnership with the Australian BioCommons:

biocommons.org.au

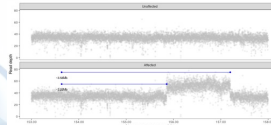


Support life scientists of all experience levels apply omics technologies and scale their research via consultations, grant support, publication support and access research compute platforms.

The University of Sydney



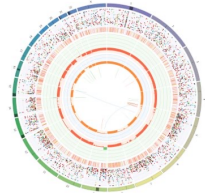
Metagenomics



Heritable disease

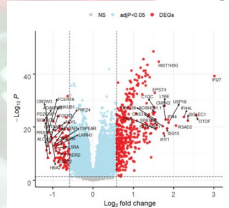


Wildlife conservation

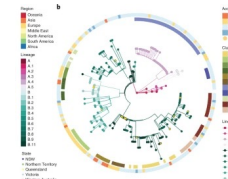
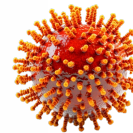


Cancer

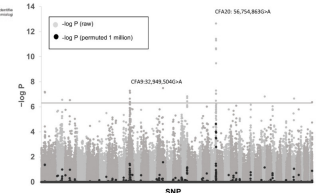
Biomarker discovery



Infectious disease



Agriculture & Veterinary



Making Bioinformatics Accessible

Improving the digital research capabilities of Australian life scientists, through simplified access to computing and data infrastructures, open access bioinformatics resources, and training.

SIH leads and contributes to national infrastructure and service projects that aim to improve Australia's research capacity and capability.



Community

Multi-disciplinary
research

Shift towards data-
intensive work



Challenges

Computational skills
shortage in biology

Bioinformatics ≠
traditional computing



Approach

Partner
collaborations

Researcher-driven

Flexible solutions



Outcomes

Reduce 'time to
science'

Enable research
excellence



Australian
BioCommons

Bring Your Own Data Platforms Project

2020 - 2023



Collaboratively developed **42** public bioinformatic pipelines for reproducible and flexible data processing and analysis.

<https://workflowhub.eu/programmes/8>



Worked directly with 11 research groups to develop and improve infrastructure and services to meet the needs of the life science community.



Supported **\$447K** of competitive computing grants and worked with national HPC facilities to be more accessible to USyd researchers.



Upskilled **1,000 life scientists** via national webinars and workshops through the Australian BioCommons training cooperative.

youtube.com/@AustralianBioCommonsChannel
zenodo.org/communities/australianbiocommons/



Prostate Cancer Genomics

Professor Vanessa Hayes

Outcomes



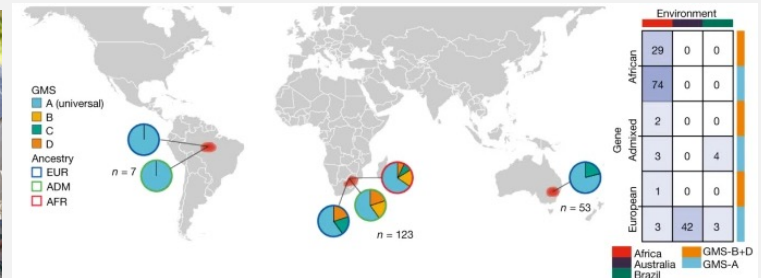
Facilitated access to NCI Gadi HPC and **~\$90K** of compute time and processed **>200 human genomes**.



3 public, scalable pipelines for read mapping, germline and somatic variant calling.
Reduced time to science from months to days.



Supported **\$1.3M** NHMRC Ideas and **\$994K** US Department of Defense Ideas grants and high-impact publications



This project was an investigation into the genetic signatures of global population disparities in prostate cancer patients. SIH developed workflows to enable large-scale data analysis and provided access to specialised support and computational infrastructure.

SIH provided bioinformatics and high-performance computing expertise, grant and publication support.

Precision Diagnostics for Rare Diseases

Northcott Neuroscience Laboratory, ANZAC Research Institute

Outcomes



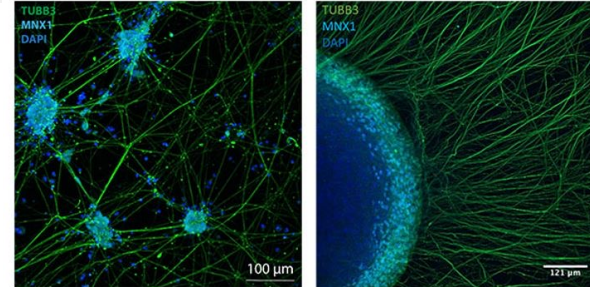
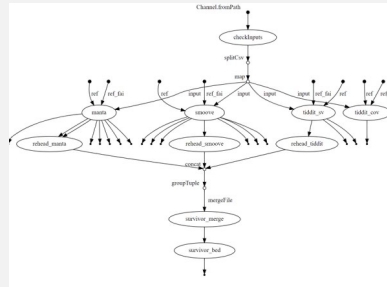
Developed a public, scalable structural variant calling pipeline for the international peripheral neuropathies community.



Facilitated on demand access to national cloud and HPC infrastructures for command-line beginners.



Empowered bench scientists and research students to upskill and harness bioinformatics through ongoing consultations.



Through ongoing consultations and project support, SIH has empowered the Northcott Neuroscience Laboratory group to implement best practice genomic analyses and collaborate with colleagues nationally and globally to explore the genetic architecture underlying heritable neuropathies.

“As a researcher, I don’t have enough consistent bioinformatics work to justify employing a full-time bioinformatician.

Having access to a team of skilled bioinformaticians at a core research facility allows me to do more research with my funding.”

Prof. Marina Kennerson
ANZAC Research Institute, FMH

Research Computing

Research Problems

Artemis HP

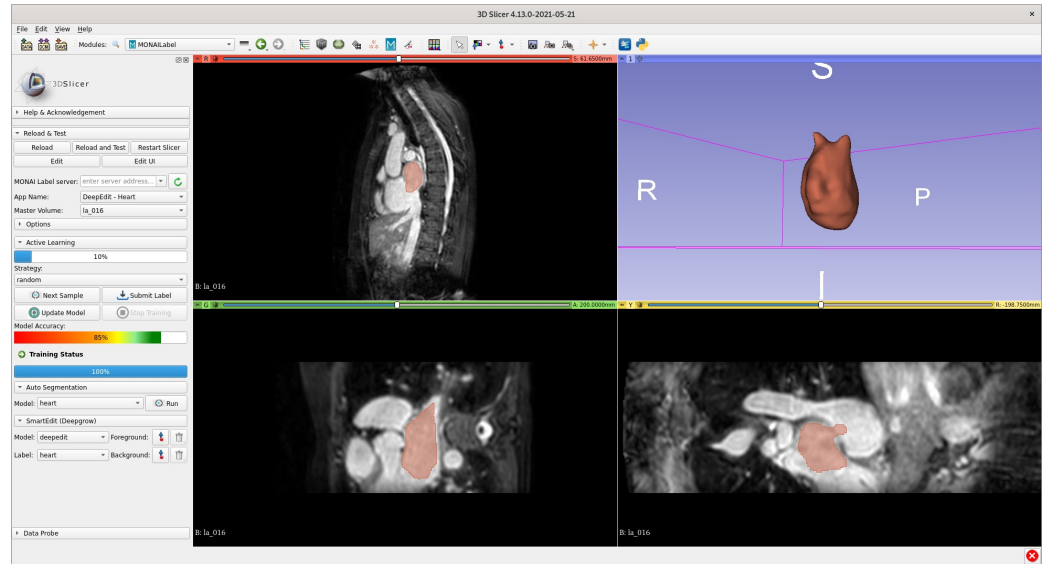
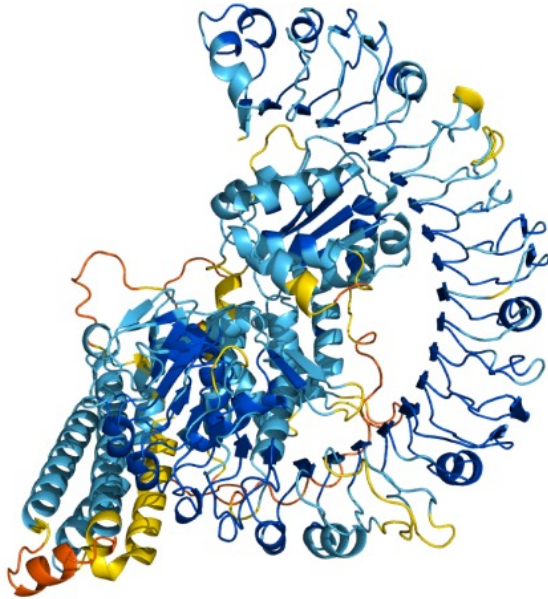
RONIN

Institutional



Research Insights

Compute Intensive Research



Data linkage, exhaust data, de-identification, model building

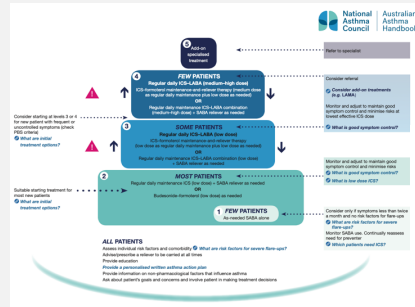
- DAMPER – patterns in asthma pharmacy prescriptions.
- AstraZeneca SYGMA clinical trials of asthma inhaler use.



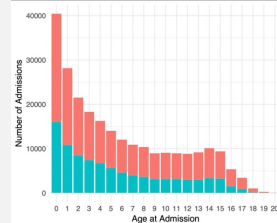
- Cardiology – STEMI Patient Electronic Data Extraction
- Deep Phenotyping of Atrial Fibrillation
- Discharge Against Medical Advice in Children's Hospitals
- Quality Improvement in Cardiac Rehabilitation SOLVE-CHD



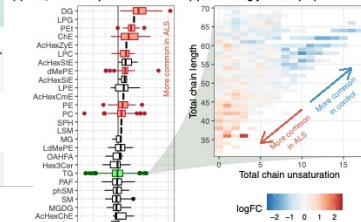
- Parkinson's disease
- Amyotrophic Lateral Sclerosis (ALS)
- Targeted Metabolomics software pipeline
- DIA-NN proteomics pipeline



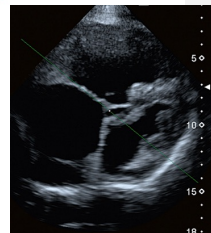
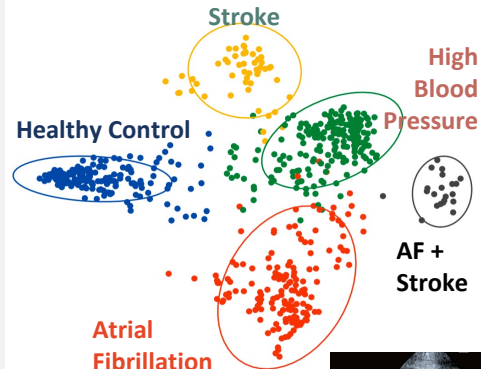
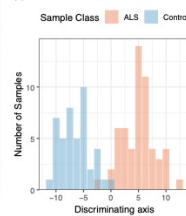
Age distribution of Admissions



(a) LCE, Red/Blue: $p < 0.05$



(c) OPLS-DA: ALS and control





THE UNIVERSITY OF
SYDNEY
Westmead Applied
Research Centre



THE UNIVERSITY OF
SYDNEY
Sydney
Informatics Hub



The
Westmead
Institute



WOOLCOCK
LEADERS IN BREATHING & SLEEP RESEARCH

Google AI



NSW
GOVERNMENT

Sydney
HealthPartners

Translating Research into Better Health Outcomes

Health

Wearable Technology

Sleep, exercise, heartrate, text messaging

Some projects we've contributed to:



Cardiology – “My Intelligent Cardiac Assistant” project to use wearables, text messages and AI to improve patient outcomes (**Prof Clara Chow, \$1m**)



Gestational diabetes - “Smart mums” project (**Prof. Wah Cheung**)



Psychological distress in the “Unwired Youth” project (**Prof. Anthony Harris**)



Insomnia (**A Prof. Chris Gordon**)



Blood pressure monitoring (**Prof. Simon Flemming**)



We host the **Wearable technology community** – bringing together a diverse community of researchers using wearable devices in their research, spanning diabetes, architecture, psychology, sleep and other areas.

By sharing knowledge and addressing common roadblocks, the community hopes to accelerate the time to science from research question to real world application. If you work with, or intend to work with, data derived from wearable technologies, please come along.

Australian Text Analytics Platform

Text analysis tools for all researchers



<https://www.atap.edu.au/>

Some tools on the platform:



Document and corpus similarity tool

- Compare differences between documents, e.g. to eliminate near-duplicates.



Quotation tool and semantic tagging

- Extract quotes from text e.g. news articles
- <https://www.atap.edu.au/posts/quotation-tool/>



Discursis

- An analysis and visualisation tool for conversational data
- <https://www.atap.edu.au/posts/discursis/>

Facebook and Instagram, which Facebook owns, followed up in the evening, announcing that Trump wouldn't be able to post for 24 hours following two violations of its policies

This is where this document is different



Facebook and Instagram, which Facebook owns, followed up in the evening, announcing that Trump wouldn't be able to post for 24 hours following two violations of its policies

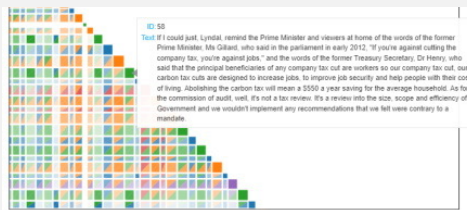
The White House did not immediately offer a response to the actions

... people were on ventilators and people were dying at that stage, and that's the context in
QUOTE

... uncession. " I was n't aware of the decisions . " Former finance minister Simon Birmingham said
QUOTE SPEAKER PERSON

... whether Mr Morrison should leave . Nationals leader David Littleproud said Australians should
PERSON SPEAKER ORG PERSON QUOTE NORP

... Albanese is awaiting advice from the solicitor - general on whether Mr Morrison 's actions have any leg



Physics & Chemistry

Text analysis tools for all researchers

Some projects we've contributed to:



Physics

- Radio astronomy to find black hole mergers and other exotic cosmic phenomena with the Australian Square-Kilometre Array Pathfinder telescope, and the Murchison Widefield Array Transients Survey
- Super-resolution of optical telescope images in the SAMI Galaxy Survey



Chemistry

- Chemical reaction rate kinetics for molecular biology
- Stoichiometry tool in lab archives for researchers



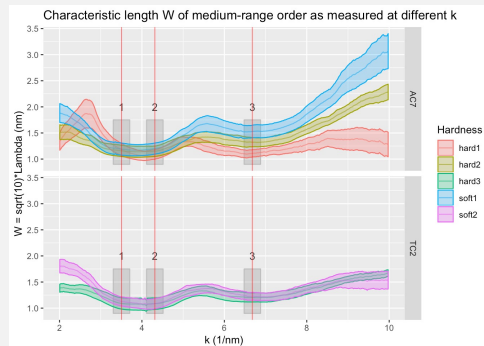
Materials Science

- Mapping medium-range order in bulk metallic glass with transmission electron microscopy



Nanofabrication

- Using Neural Networks for inverse and forward modelling of silicon photonics designs for LIDAR
- Wearable fibre-optic blood-pressure measurement



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub



**MICROSCOPY
AUSTRALIA**

Geospatial Modelling

Agriculture, Government, Ecology...

Some projects we've contributed to:



Ecology:

- Mapping greenspace along Sydney railway lines
- Modelling microbat population drivers



Agriculture

- AgReFed GeoData Harvester
- Paddock Segmentation from aerial imagery
- Soil water mapping



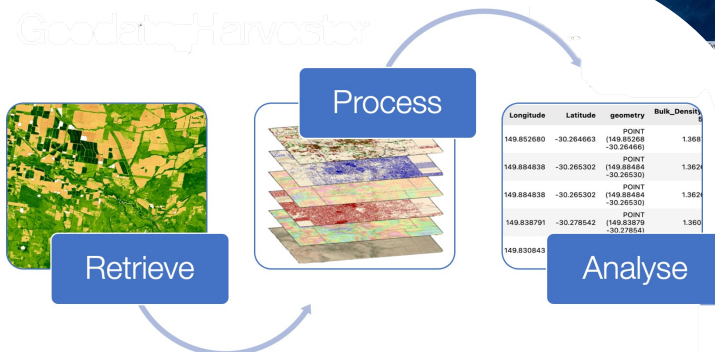
Environment

- Mapping mercury levels in fish
- Mining Tailings dam portal



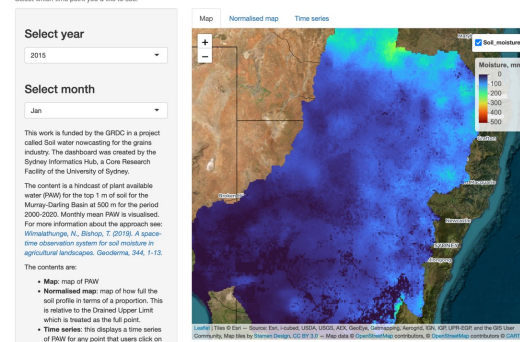
Criminology

- Patterns in crime statistics over space and time.



Plant-available water for the Murray-Darling Basin

Select which time point you'd like to see.





Visual AI

For conservation, medical imaging, microscopy, & geospatial analysis

Outcomes



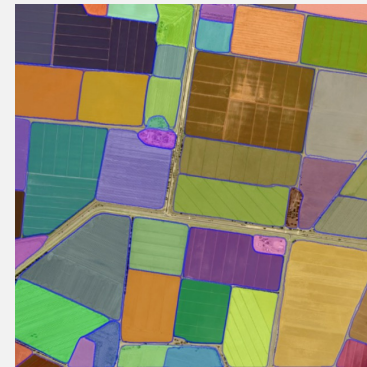
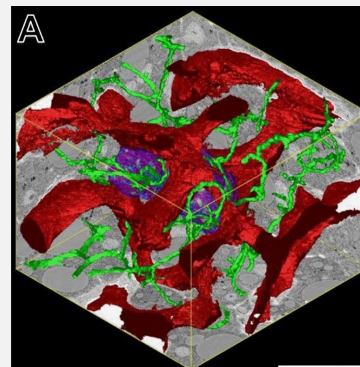
Fine tuned state of the art computer vision models to drive research excellence and global impact for good.



Created cloud native pipelines and tools to accelerate translation of raw data into research insights.



Open-source code and demos available. Contact us to find out more.



SIH data scientists and software engineers work with USYD researchers, other Core Research Facilities, and government agencies to implement visual AI solutions for real world problems.

Our products enable our clients to more efficiently and quickly use visual data in both research & real-world applications.



THE UNIVERSITY OF
SYDNEY
Westmead Applied
Research Centre



THE UNIVERSITY OF
SYDNEY
Sydney
Informatics Hub



Large Language Models & Generative AI

Spearheading adoption for research at all levels



Large Language Models (LLMs) like ChatGPT present nearly unlimited research **opportunities** & many associated **risks**.



SIH are **experts in AI**, software engineering and cloud computing, and advise on research strategy and architecture, provide technical leadership, and implement LLM powered solutions to drive research excellence.



Contact us if you would like to be a part of an **“AI in Research”** working group discussing these opportunities & risks.

Who funded this study?



The study was funded by various sources, including the Emergency Medicine Foundation, Wishlist Sunshine Coast Hospital Foundation, Queensland Advancing Clinical Research Fellowships, and the Gold Coast Health Study Education and Research Trust Fund (nejmoa2213883.pdf, pages 8-9). However, no specific funding source is mentioned for the study itself. The authors' disclosure forms and a data sharing statement are available with the full text of the article at NEJM.org, which provides further information on the study's funding and potential conflicts of interest.

See context

nejmoa2213883.pdf: The study acknowledges funding from various sources, including the Emergency Medicine Foundation, Wishlist Sunshine Coast Hospital Foundation, Queensland Advancing Clinical Research Fellowships, and the Gold Coast Health Study Education and Research Trust Fund. The authors' disclosure forms and a data sharing statement are available with the full text of the article at NEJM.org. The study also thanks several individuals and research units for their contributions. No specific funding source is mentioned for the study itself.

Valid keys: nejmoa2213883.pdf pages 8-9

Statistical Consulting Service



Experienced statisticians who give practical advice on experimental design, analysis and interpreting statistical/quantitative results in one-on-one consultations and workshops.

We help you to:

- **Boost grant success** by submitting an optimal experimental/analytical design up-front.
- Find more interesting/useful **patterns in your data** quicker and with less pain.
- Resolve reviewer's comments to **get your papers published** or projects greenlit.

We encourage use of the free R statistical programming language, helping you to build analytical coding skills which you can use anywhere your career takes you. We also assist with SPSS and other statistical software.

Rather than just 'do your statistics' we advance your statistical skills to improve your research practice - leading to higher quality research outputs and publications in higher impact journals.

Hacky Hour

For researchers who code or analyse data



A monthly meetup where anyone from the University – students, staff and university affiliates – can collaborate and get support e.g., swap notes, get help, or learn new techniques in programming and data science.

Experts & mentors from SIH and across the University will be available to advise and answer questions on coding, data analytics or digital tools.

Come join us on zoom the

3rd Wednesday of every month, 3 - 4pm!
Next one: 19th July 2023

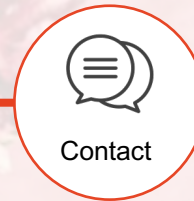
sydney.edu.au/informatics-hub/hacky-hour



How to Engage with Us



sydney.edu.au/informatics-hub



And fill out our
request form

Or send us an email:



sih.info@sydney.edu.au