AI in Research

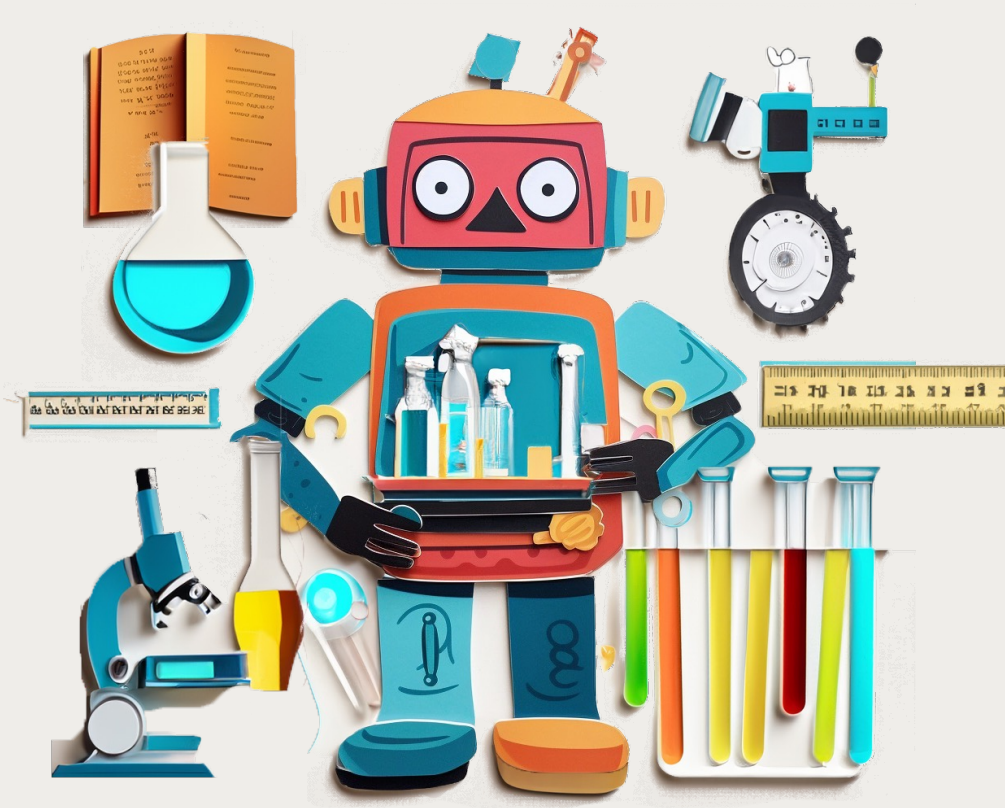
21st February 2024
Research Portfolio Researcher Spotlight

Dr Gordon McDonald
Dr Henry Lydecker

Sydney Informatics Hub

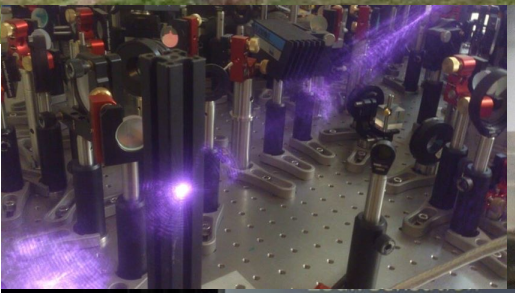
sydney.edu.au/informatics-hub

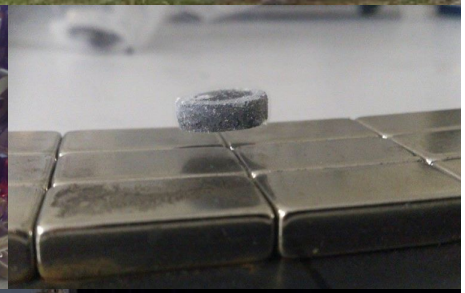
THE UNIVERSITY OF SYDNEY
Sydney Informatics Hub



We recognise and pay respect to the Elders and communities – past, present, and emerging – of the lands that the University of Sydney's campuses stand on. For thousands of years they have shared and exchanged knowledges across innumerable generations for the benefit of all.

THE UNIVERSITY OF SYDNEY



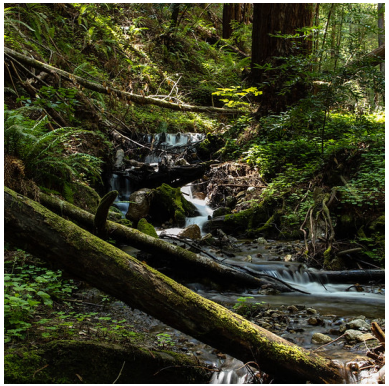


Gordon McDonald

Six years building Quantum Sensors with atom interferometry at ANU
2016 – Now, University of Sydney, Sydney Informatics Hub.

In my time here we've had five thousand consults and projects in data science, AI, statistics, bioinformatics and high performance computing for academic researchers in all disciplines, and in collaboration with a myriad of national and international stakeholders. Together with my team, we've unleashed our tools and experience solving problems in conservation, medicine, law, geology, business, physics, education, genomics, archaeology and many more. Lately we're working out how to channel state-of-the-art AI models into research in a bunch of different ways which we'll tell you about in this talk.

Henry Lydecker



Redwood Forest Conservation Ecology
Coffee-ant Agroecology
Urban Ecology of Parasites
Zoonotic Disease Risk Modelling

Henry is data scientist specialising in using artificial intelligence techniques including large language models and computer vison to solve problems in the real world. While his background is in ecology and public health research, he has worked on a range of projects from many different fields including agriculture, biology, ecology, geology, health, and law and is an expert in deep learning, machine learning, GIS, statistics, and software development. He leads a team of data scientists working to provide data science solutions to research problems, and to uplift the digital skills of the research community through training and outreach.

Sydney Informatics Hub

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

- *Statistics*
- *Data Science, AI & Software*
- *Research Computing*
- *Bioinformatics*

The University of Sydney

sydney.edu.au/informatics-hub

What is AI?

Generates…

“A friendly koala cyborg”

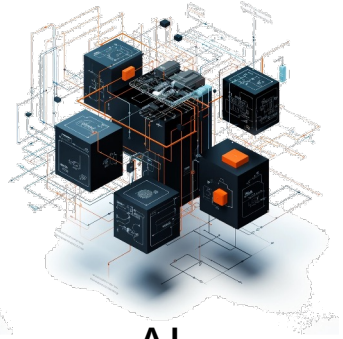
Images, Video

what is the best way to peel a mandarin?

Text

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:

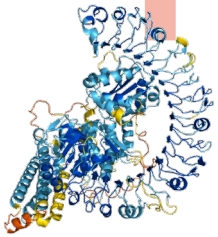
“Hello”



AI

“你好”

Audio



has an internal representation of meaning or content

[0, 1, .5, .2]

3D structures or molecules

Blessing

Amazing results in far less time/effort

Well-suited to poorly defined and ambiguous problems

Curse

Sometimes results will be unfeasible, wrong, ‘hallucinated’, biased etc.

You need a way to check the answer.

Some approaches are very data limited.

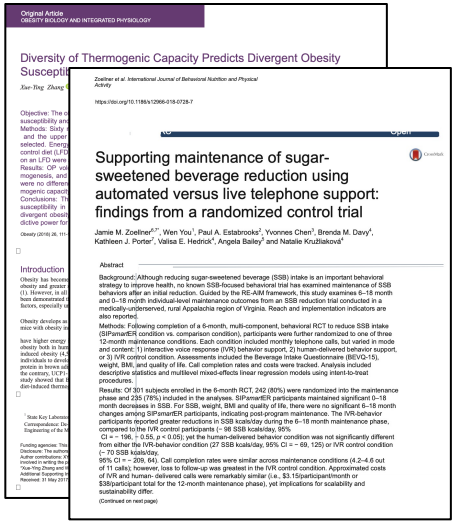
Modern Work

Information harvesting, integrating, and storytelling.

Large language models for literature review, synthesis, and text analysis.

Extracting information from text with language models

Academic papers on nutrition



Original Article
OBESITY BIOLOGY AND INTEGRATED PHYSIOLOGY

Diversity of Thermogenic Capacity Predicts Divergent Obesity Susceptibility

Zoellner et al. International Journal of Behavioral Nutrition and Physical Activity

https://doi.org/10.1186/s12966-018-0728-7

Supporting maintenance of sugar-sweetened beverage reduction using automated versus live telephone support: findings from a randomized control trial

Jamie M. Zoellner[6,7], Wen You[1], Paul A. Estabrooks[2], Yvonnes Chen[3], Brenda M. Davy[4], Kathleen J. Porter[5], Valisa E. Hedrick[4], Angela Bailey[4] and Natalie Kruzliakova[6]

Abstract

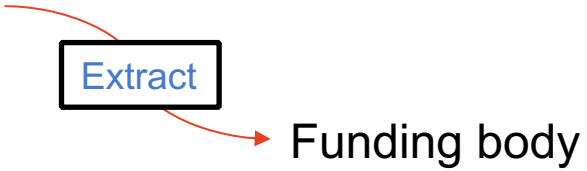
Background: Although reducing sugar-sweetened beverage (SSB) intake is an important behavioral strategy to improve health, no known SSB-focused behavioral trial has examined maintenance of SSB behaviors after an initial reduction. Guided by the RE-AIM framework, this study examines 6–18 month and 0–18 month individual-level maintenance outcomes from an SSB reduction trial conducted in a medically-underserved, rural Appalachia region of Virginia. Reach and implementation indicators are also reported.

Methods: Following completion of a 6-month, multi-component, behavioral RCT to reduce SSB intake (SIPsmartER condition vs. comparison condition), participants were further randomized to one of three 12-month maintenance conditions. Each condition included monthly telephone calls, but varied in mode and content: 1) interactive voice response (IVR) behavior support, 2) human-delivered behavior support, or 3) IVR control condition. Assessments included the Beverage Intake Questionnaire (BEVQ-15), weight, BMI, and quality of life. Call completion rates and costs were tracked. Analysis included descriptive statistics and multilevel mixed-effects linear regression models using intent-to-treat procedures.

Results: Of 301 subjects enrolled in the 6-month RCT, 242 (80%) were randomized into the maintenance phase and 235 (78%) included in the analyses. SIPsmartER participants maintained significant 0–18 month decreases in SSB. For SSB, weight, BMI and quality of life, there were no significant 6–18 month changes among SIPsmartER participants, indicating post-program maintenance. The IVR-behavior participants reported greater reductions in SSB kcals/day during the 6–18 month maintenance phase, compared to the IVR control participants (− 98 SSB kcals/day, 95% CI = − 196, − 0.55, $p = 0.05$); yet the human-delivered behavior condition was not significantly different from either the IVR-behavior condition (27 SSB kcals/day, 95% CI = − 69, 125) or IVR control condition (− 70 SSB kcals/day, 95% CI = − 209, 64). Call completion rates were similar across maintenance conditions (4.2–4.6 out of 11 calls); however, loss to follow-up was greatest in the IVR control condition. Approximated costs of IVR and human- delivered calls were remarkably similar (i.e., $3.15/participant/month or $38/participant total for the 12-month maintenance phase), yet implications for scalability and sustainability differ.

(Continued on next page)

"Who were the authors funded by?"



Identifying biases in nutrition research - Dr. Fabian Held
SIH: Dr. Henry Lydecker, Dr. Joel Nothman, Dr. Chao Sun

Extracting information from text with language models

NSW Supreme Court
Inheritance law judgments

Extracting Factors such as:

- Value of estate
- Relationship of parties to deceased
- Misconduct
- Drug or alcohol use

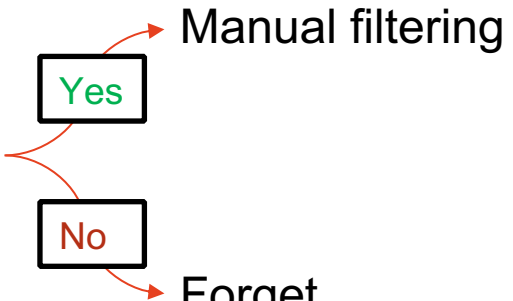
Extract → Structured data

Inheritance Case Law – Dr. Ben Chen
SIH: Mike Lynch, Xinwei Luo

Extracting information from text with language models

PubMed articles

Does this paper report on an inherited trait?



OMIA

Online Mendelian Inheritance in Animals

OMIA Text Mining

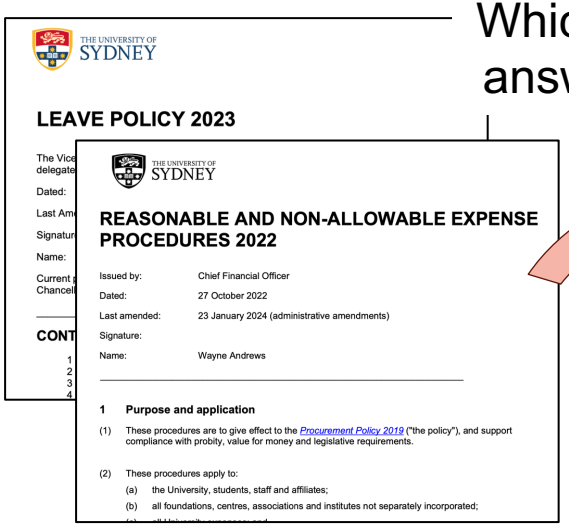
Professor Frank Nicholas, Associate Professor Imke Tammen

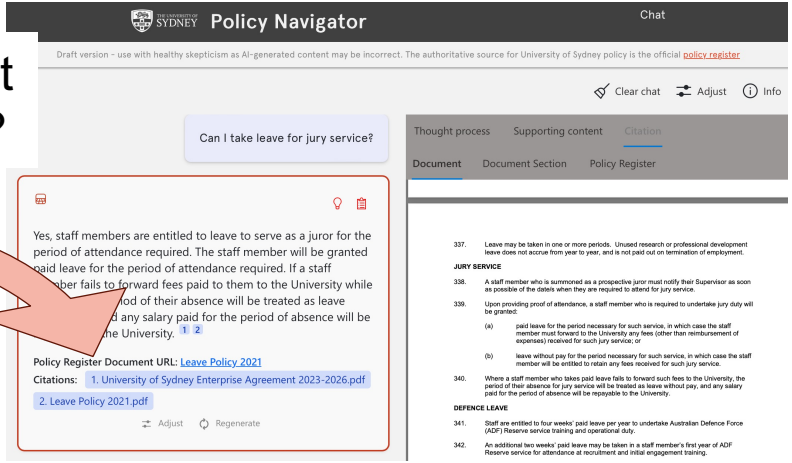
SIH: Marius Mather, Sony Jufri, Joel Nothman, Di Lu

Searching for information

University of Sydney policies

Which policy document answers the question?



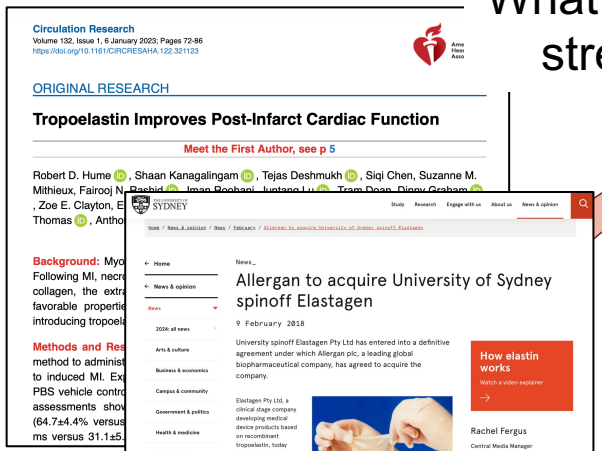


https://policy-navigator.techlab.works

Policy Navigator (alpha)
SIH: Henry Lydecker, Gordon McDonald
TechLab: Iqbal Chowdhury, Jim Cook
OGC: Lauren Myers, Deb Hook

Contextualizing information and storytelling

Papers, patents, PR, …



What is the impact of this stream of research?

Research Impact Assessment Framework Case Study

Organisation Details		
Organisation name:	University of Sydney Charles Perkins Centre	
Title of case study:	**Elastagen**	
Period when research was undertaken:	2013 – 2023	
Period when the claimed impact occurred:	2013 – 2023	
Names and roles of staff:	**Staff name**	**Role**
	Anthony Weiss	McCaughey Chair in Biochemistry, University of Sydney, NHMRC Leadership Fellow, Founder of Elastagen, and leader of Tissue Engineering & Regenerative Medicine in the Charles Perkins Centre, University of Sydney.

1. What is the problem your research seeks to address? Why is it significant? (250 word limit)

The research by Anthony Weiss and his team on Elastagen seeks to address the significant challenge of replicating the properties of human elastin, a key protein in the human body that provides elasticity and resilience to tissues such as skin, lungs, and arteries[1]. Human elastin's unique mechanical properties and biocompatibility make it difficult to mimic synthetically, which is a significant barrier in tissue engineering and regenerative medicine[2]. The significance of this research lies in its potential to revolutionize the field of biomedical engineering by providing a synthetic alternative to human elastin that can be used in a variety of medical applications, including wound healing, tissue repair, and the improvement of skin quality[1]. The development of Elastagen represents a breakthrough in the creation of biomaterials that can closely match the properties of natural human tissues, thereby enhancing the effectiveness of medical treatments and improving patient outcomes[2].

Research Impact Assessment Framework

SIH: Sebastian Haan, Nathaniel Butterworth, Gordon McDonald

FMH: Mona Shamshiri, Janine Richards, Prof. Robyn Ward

NSW Health

AI for Life Sciences

Genomics and Protein Design

AlphaFold
Dia-NN
RF Diffusion
AI laboratories

Enabling access to AlphaFold

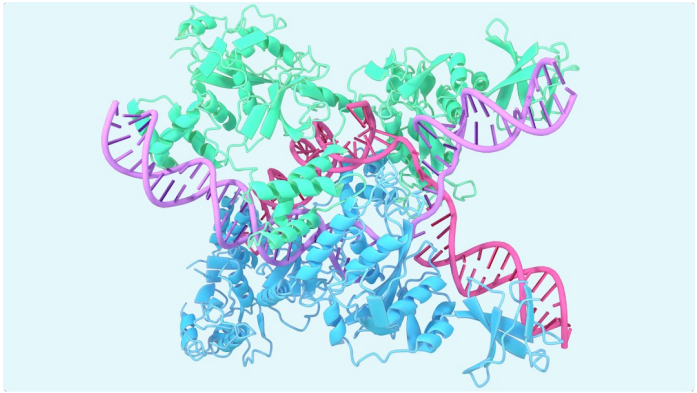
SIH: Nate Butterworth, Chris Le, Nandan Deshpande
FMH: Jake Chen

Key Highlights

Deployed on Artemis HPC, NCI Gadi, Galaxy Australia, Ronin, & Google Colab.

Training materials publicly available on GitHub & YouTube.

USYD researchers now able to easily apply AlphaFold in their work.



Google Deepmind's AlphaFold2 has revolutionized the way that researchers study protein structures.

We've provided USYD researchers many ways to access and use AlphaFold, and developed training courses & materials to make it easier for a diverse range of researchers to use this cutting-edge tool in their work.

https://sydney-informatics-hub.github.io/training.alphafold/

Scaling Proteomics with Dia-NN

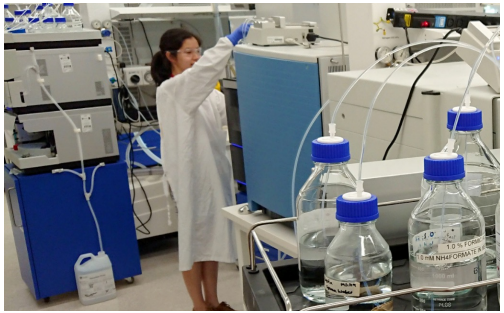
SIH: Cali Willet, Tracy Chew, Nathaniel Butterworth, Georgie Samaha
Sydney Mass Spec: Ben Crossett, A/Prof Carsten Schmitz-Peiffer (SOLES)

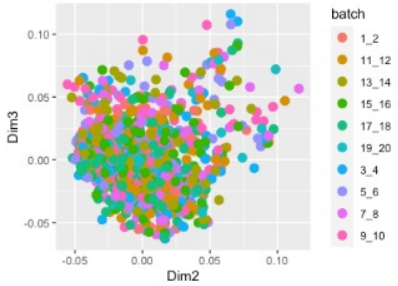
Key Highlights

Deployed state-of-the-art tool for identifying proteins in mass spec data on NCI GADI.

Eliminated data processing bottleneck, speeding up by orders of magnitude (46 days down to 18 hours).

Strategic collaboration between SIH, Sydney Mass Spec, Australian Biocommons, and the CPC.





Dia-NN is the leading tool to identify peptides and proteins in mass spectrometry data. We built a scalable Dia-NN workflow, dramatically speeding up processing and solving a major bottleneck in large scale proteomics studies.

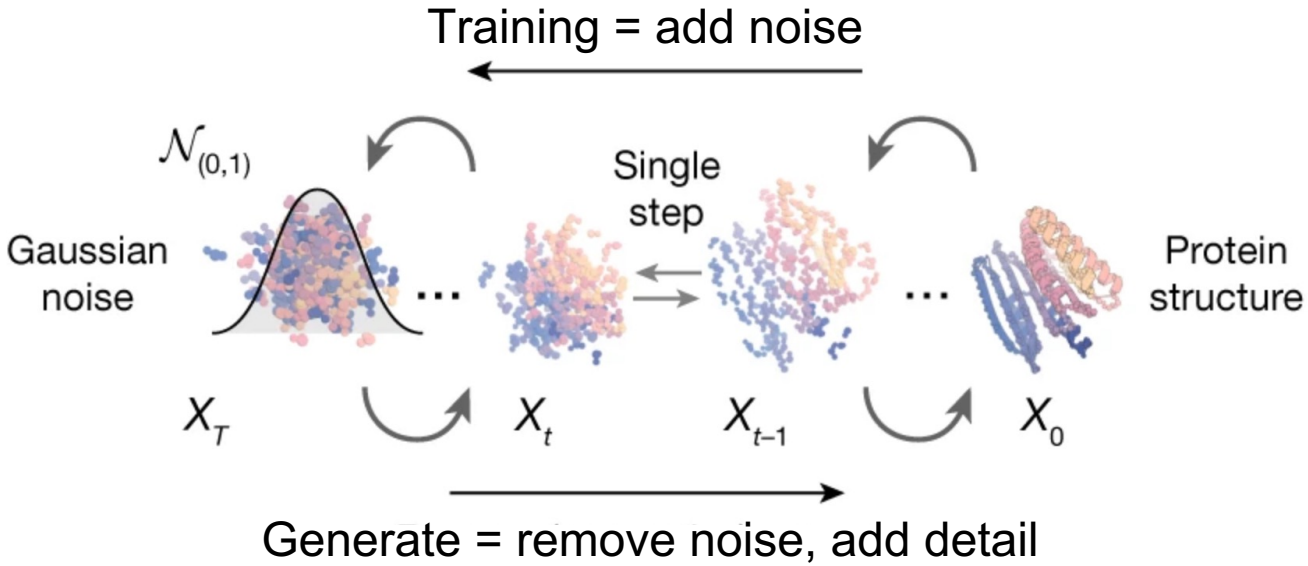
Diffusion models

Training = add noise

Generate = remove noise, add detail

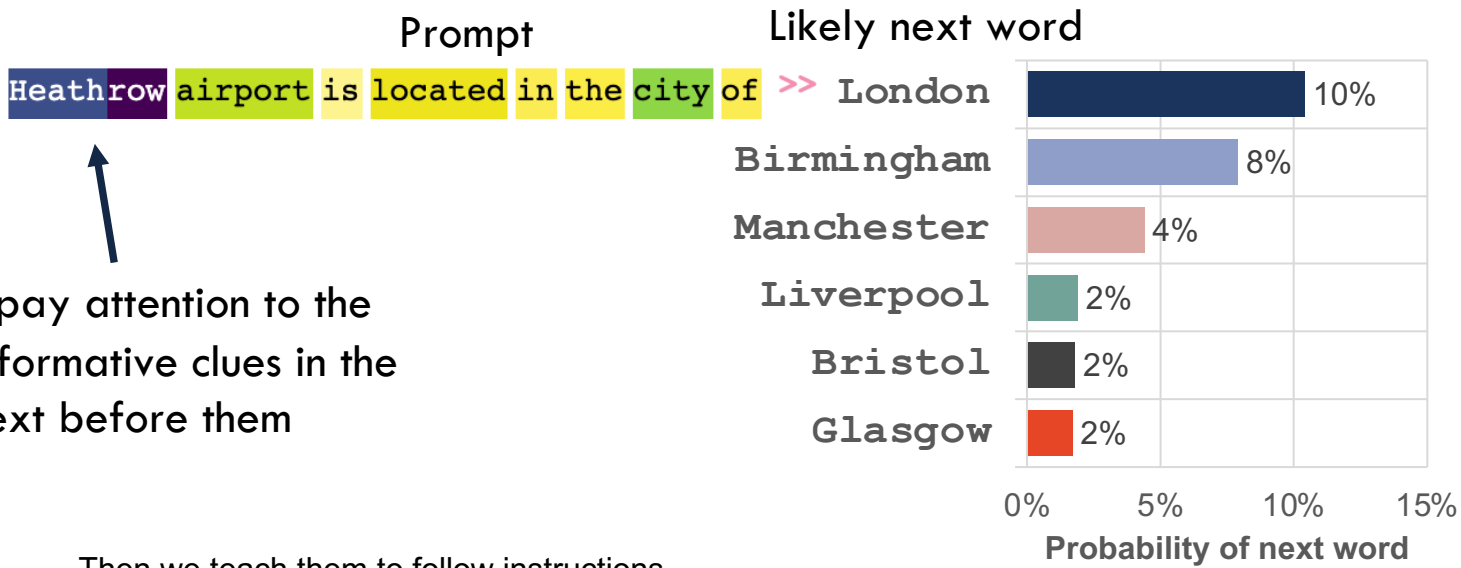
https://wandb.ai/wandb_gen/audio/reports/A-Gentle-Introduction-to-Dance-Diffusion--VmlldzoyNjg1Mzky

Diffusion models



Watson, J.L., Juergens, D., Bennett, N.R. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). https://doi.org/10.1038/s41586-023-06415-8

Language models (like ChatGPT)



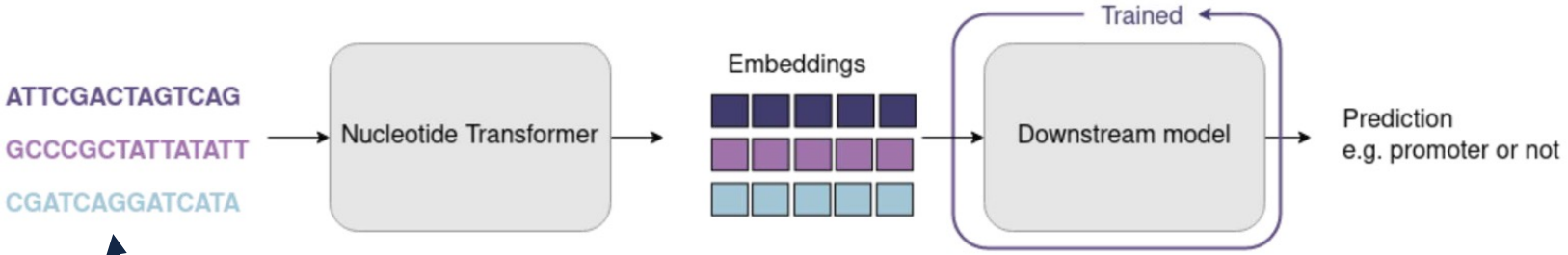
They pay attention to the most informative clues in the text before them

Then we teach them to follow instructions, so they are more useful to us

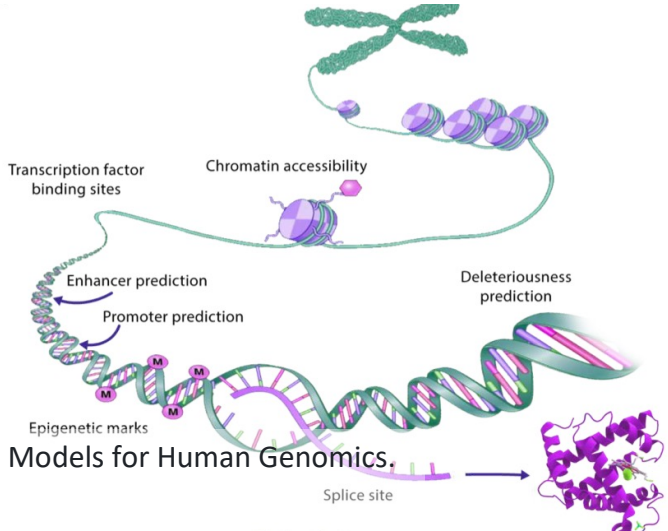
https://jalammar.github.io/explaining-transformers/

Ecco Collab notebook

Genomics models



They pay attention to the most informative clues in the text before them



Then we fine tune them, so they are more useful to us for particular tasks

Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics.
https://www.biorxiv.org/content/10.1101/2023.01.11.523679v2

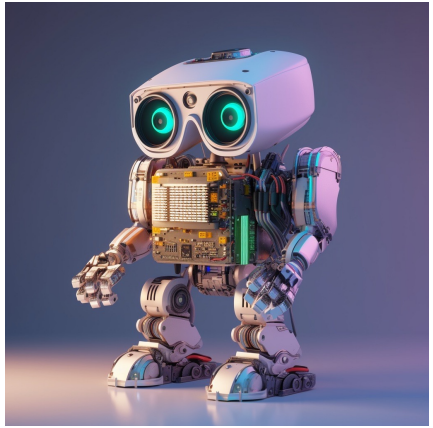
Agents

LLMs can simulate individual roles in a project

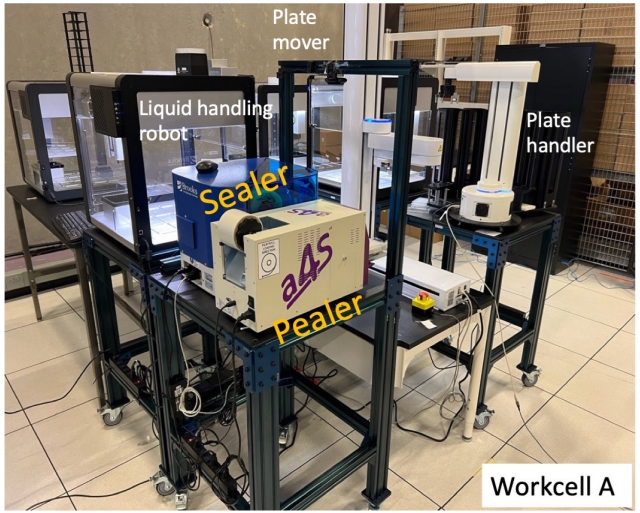
- Project manager
- Research assistant
- Professor
- Software engineer
- Statistician
- Subject matter expert.

Make them talk to each other, and you can **create a team of specialist AI agents** working together towards your goal.

Get results in minutes instead of days!



Agent-based lab automation



Self Driving Laboratories @ Argonne
https://github.com/AD-SDL

A-Lab, a facility at Berkeley Lab where artificial intelligence guides robots in making new materials created by GenAI

Szymanski, et al (2023). *An autonomous laboratory for the accelerated synthesis of novel materials*, Nature
Merchant, et al (2023). *Scaling deep learning for materials discovery*, Nature

Picture This

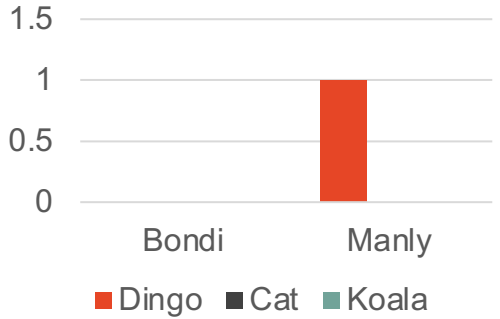
AI in Geospatial, Microscopy, Medical, Agriculture, & Conservation imagery

Image Classification
Object Detection
Instance Segmentation
Multi-Modal Models

“We have a lot of images, and we want a way to automate analyzing them”

Problem: Ecologist has 100,000 images and wants to know what is in them

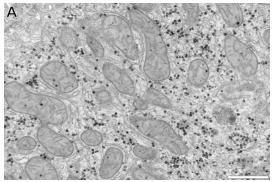
Photo	Species	Count	Site
1	NA	NA	Manly
2	Dingo	1	Manly
3	NA	NA	Manly
4	NA	NA	Manly
5	NA	NA	Manly
6	NA	NA	Manly
7	NA	NA	Manly

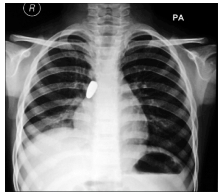


Dr Aaron Greenville, SOLES

The University of Sydney

Computer Vision





Model

What is this image of?

Is there anything in it?

Where are the things in it?

Detecting wildlife with NSW National Parks

SIH: Henry Lydecker, Nathaniel Butterworth, Gordon McDonald

Key Highlights

Fine-tuned family of YOLOv5 models on **2.16 million** camera trap images.

Detect and identify up to **72 different species**, with **94% precision** and **95% recall** on the top 33 most common.

GPU-accelerated prediction pipeline **1371x faster** than manual labeling.

Ecological researchers like Dr. Aaron Greenville, conservation agencies, and governments make extensive use of camera traps to monitor wildlife. Processing this data requires an immense amount of time.

We implemented an AI pipeline for automatically detecting & identifying animals in camera trap images, dramatically speeding up speed of data processing.

MICROSCOPY AUSTRALIA

Speeding up microscopy data processing

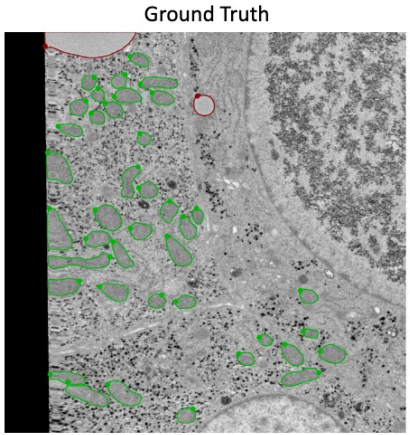
SIH: Sebastian Haan, Nathaniel Butterworth

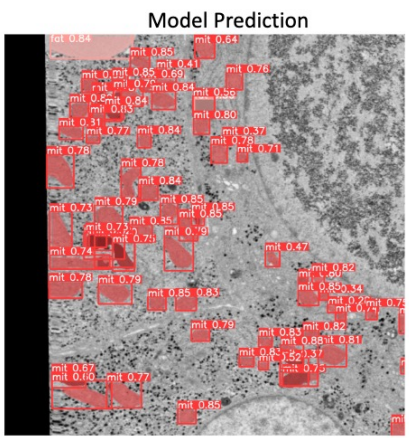
Key highlights

Sped up annotation process by up to 10x existing approaches

Full technology review of existing solutions with best practice solutions provided to clients

Strategic collaboration between SIH and SMM





Sydney Microscopy & Microanalysis collects immense amounts of data at the micro, nano, & atomic scales. Processing this data is manual and time consuming. We developed AI pipelines to speed up cell segmentation in electron microscopy images, automating 3D cell tomography.

SMM: Gerry Shami, Chad Moore, Andre Venne, Eleanor Kable, Prof. Filip Braet

Using urban morphology to model thermal comfort

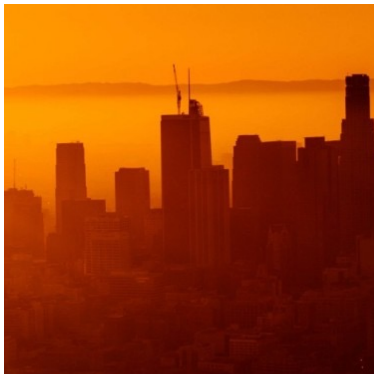
SIH: Henry Lydecker, Sahand Vahidnia, Thomas Mauch, Xinwei Luo

Key Highlights

Open-source Python tools for spatial imagery processing, annotation, and analysis using AI.

Model training dataset: 55k tree patches, 75k buildings, 3,900 aerial images.

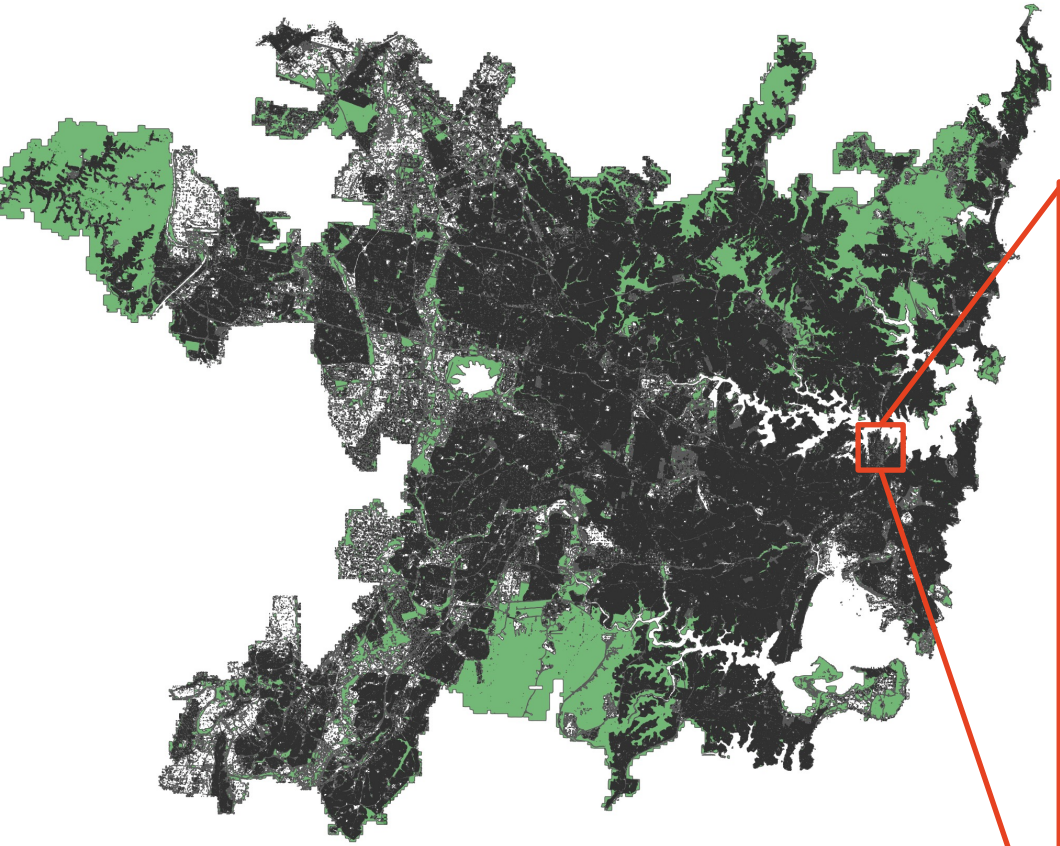
Predicted Dataset: 2 million tree patches, 1 million buildings, 37k aerial images over 12k km^2. Freely available in GeoParquet format.



Professor Ben Eggleton's group are studying how heat waves are harming cities, increasing water demand and health risks.

We developed AI models for extracting key variables for understanding human thermal comfort from spatial imagery, paving the way for future solutions to support urban cooling and water demand management.

Greater Sydney Dataset



THE UNIVERSITY OF SYDNEY

Sydney Informatics Hub

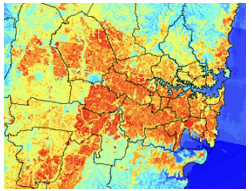
>2 million tree patches
1 million buildings

Understanding cities with computer vision

Prof. Ben Eggleton, NSW Smart Sensing, SIH: Henry Lydecker, Sahand Vahidnia, Thomas Mauch, Xinwei Luo

NSSN
NSW Smart Sensing Network

Urban Heat ML Model



Geospatial Features

Local Climate Zone = “2: Compact Midrise”

(49% class confidence)

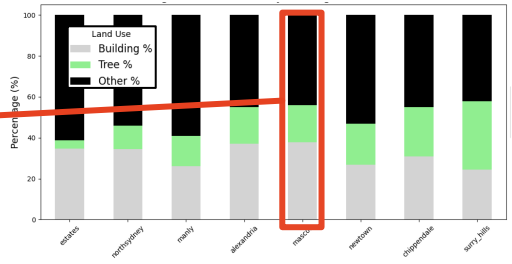
Input Image

RGB Visual Light Satellite Imagery
Aerial Imagery (Best)

Computer Vision

Insights, ML, Analysis

Compare neighborhoods

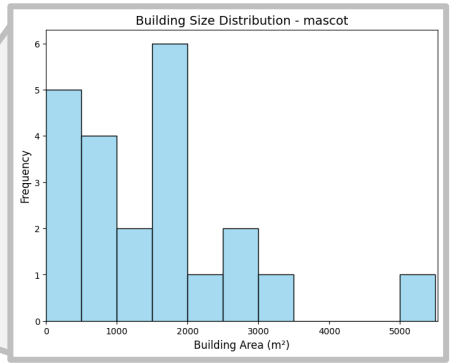


Tree Patches

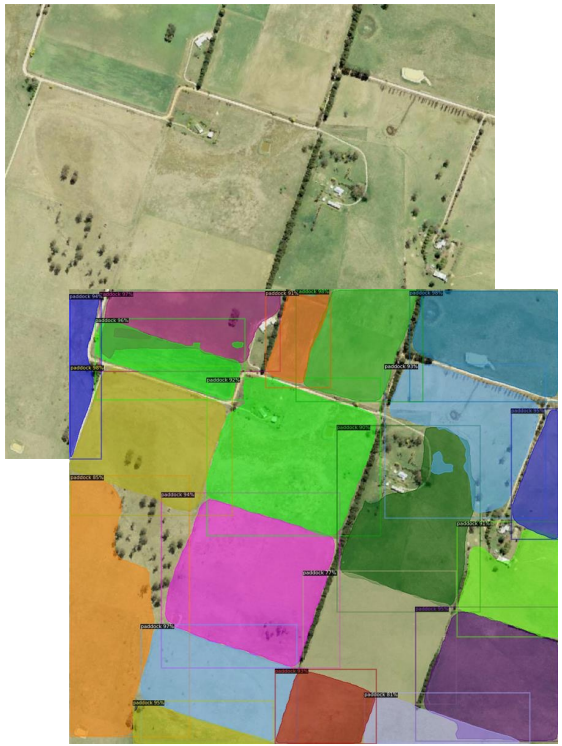
- 101 tree patches
- 15% coverage

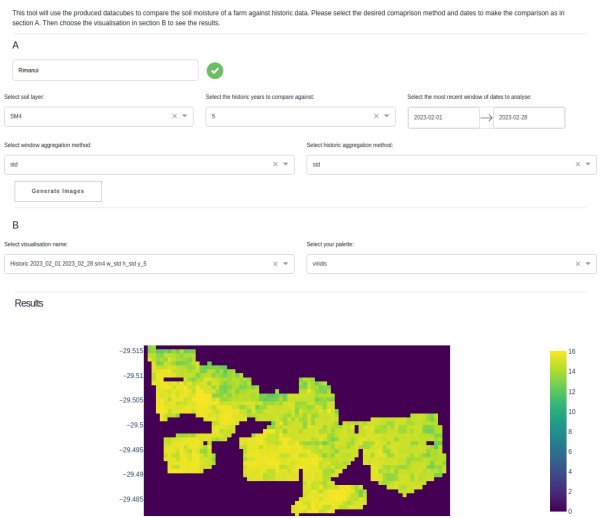
Building Outlines

- 22 buildings
- 38% coverage
- building footprint area distribution



Agricultural land use models, urban ecology





Recommended Tools and Platforms

Computer Vision Models

- YOLOv5 – models for classification, object detection, and segmentation
- Detectron2 – powerful tools, using vision transformers
- Segment Anything Model – general purpose zero-shot panoptic segmentation
- Grounding DINO – make your models multi-modal using text

Platforms & Tools

- Weights & Biases – Model training and evaluation diagnostics
- Roboflow – Very easy to use dataset utilities
- Huggingface – Open & free tools for AI proof of concepts
- PaddlePaddle - frameworks for AI use cases
- Papers with code – compare state of the art models' performance
- Colab – free Jupyter notebook interface from Google
- ARDC Jupyter notebook service - free Jupyter notebook interface from ARDC

Literature Search

Research Rabbit – free for researchers, operates on donation model
Consensus Free basic usage, $7/month for pro features
Paper Digest Literature Review Free to try but $7-$11/month for pro features
Elicit free trial, then $10/month
Rayyan $8/month researcher, $4/month student
Scite Free trial then $13-$22/month
SciSpace
Scopus-AI

Coding assistance

Free (ish)
Jupyter-AI
Open Interpreter
Tabby
MetaGPT

Paid
ChatGPT code interpreter
Github Copilot

Agent building

Free (ish)
Cogniti
Flowise
MetaGPT

Paid
ChatGPT Custom GPTs
Azure Promptflow

Sydney Informatics Hub
Dr Gordon McDonald, Informatics Team Lead
Dr Henry Lydecker, Data Scientist
sih.info@sydney.edu.au
first.last@sydney.edu.au

sydney.edu.au/informatics-hub

THE UNIVERSITY OF SYDNEY
Sydney Informatics Hub