

> AI and research

(& how to get stuff done)

@ SOLES AI Workshop
Thursday 28th Nov 2024

Dr. Gordon McDonald
Informatics Team Lead,
Sydney Informatics Hub



THE UNIVERSITY OF
SYDNEY
—
Sydney
Informatics Hub

informatics.sydney.edu.au/

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



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.



sydney.edu.au/informatics-hub/training

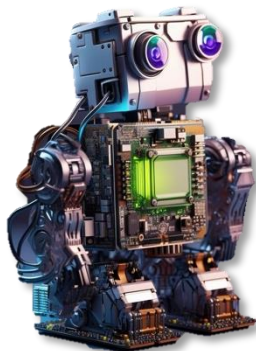
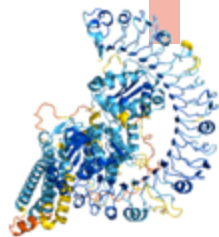
Statistics	Data Science	Research Computing	Bioinformatics	Events
Fundamentals	Machine Learning	High Performance Computing	'omics Techniques	Hacky Hour
Modelling	Visualisation	Cloud Computing	Reproducible Pipelines	Summer Schools
Specialist	Natural Language Processing	Containers	Data Analytics	Coding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

What is AI?

“A friendly koala cyborg”

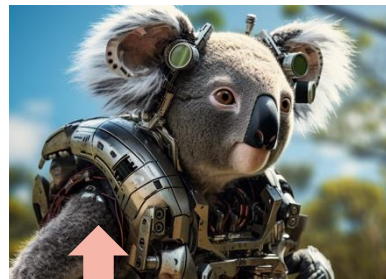


what is the best way to peel a mandarin?



AI

has an internal representation
of meaning or content
 $[0, 1, .5, .2]$



Generates...

Images, Video

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:

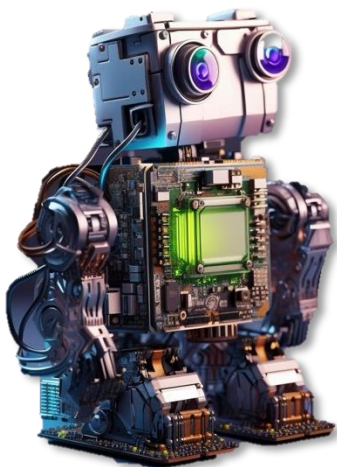


Audio



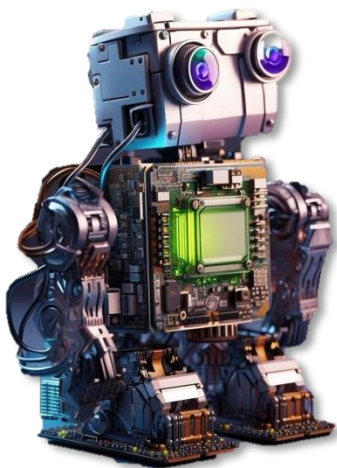
3D structures
or molecules

You're trusting this guy



- He's doing his best
- But mistakes happen.
- Maybe you should pay him more

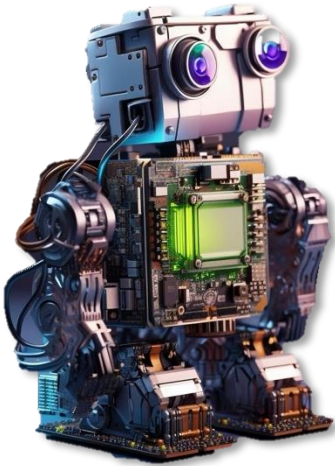
What can he do well?



The boring mindless repetitive bits
you give to an RA like

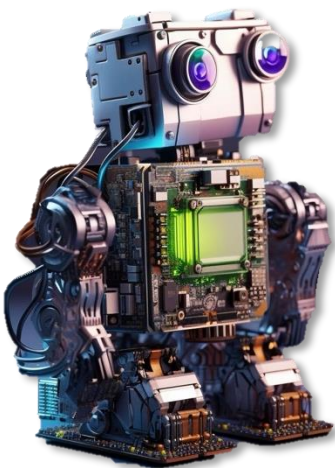
- labelling stuff
- classifying stuff
- turning your vague instructions
into concrete actions
- If you can see it, he might. If you
can't, he might have trouble too.

What's going to go wrong?



- Over complication
- Oversimplification
- "best recollection"
not absolute truth
- nothing -> average
- lack of diversity of
ideas and perspectives

What's not changing?

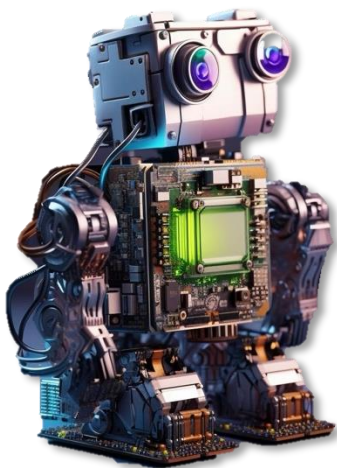


We want to **build** our ideas, **explore** our data, **understand** the world better

We want tools that:

- help us do this efficiently,
- minimize information overload
- don't miss critical parts,
- help us make the important decisions, but not the irrelevant ones.
- let us be the final judge and don't replace our agency

Top tips

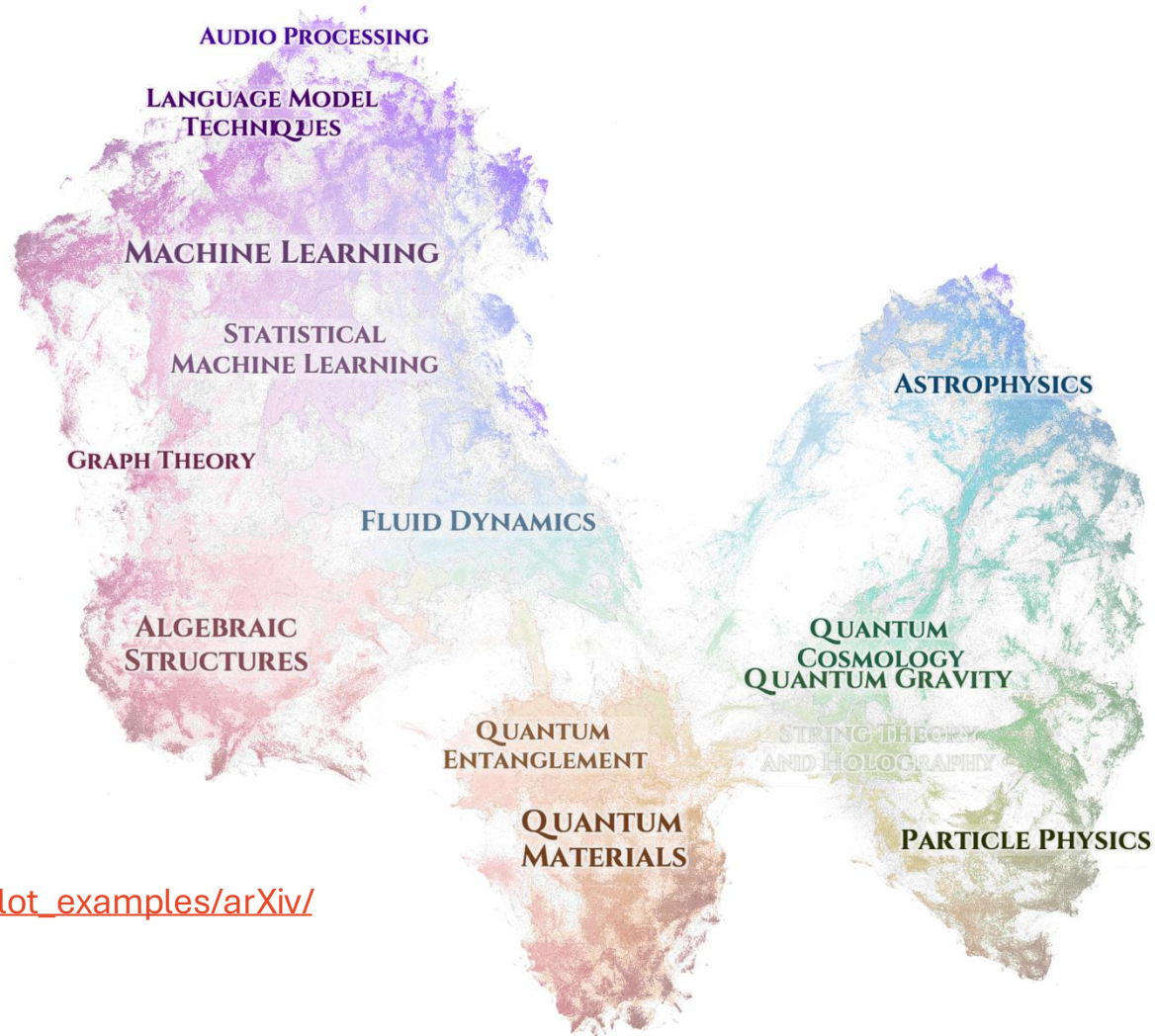


- Be as **specific** as you can about exactly what you want (or ask it to help you refine)
- make sure **private** stuff stays private!
- think carefully about what you **don't** want to automate
- small, **incremental** changes
- **test/check** as you go

Research examples with AI

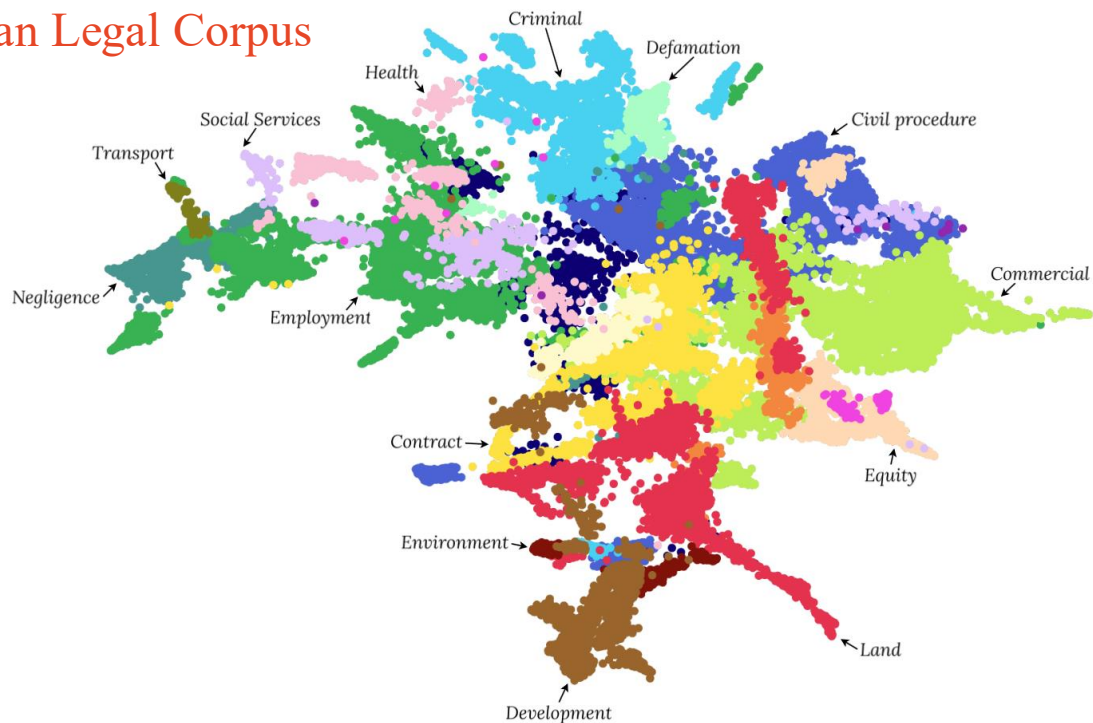
Semantic map of ArXiv

(by the guy who invented UMAP)



lmcinnes.github.io/datamapplot_examples/arXiv/

Semantic map of the Open Australian Legal Corpus



Extracting information from text with language models



OpenAI GPT-4o
(May '24)

Caselow

Home Browse About Contact us Enter a search term Search

Advanced Search Search Tips

Back Email PDF Print

Listen

New South Wales Supreme Court

CITATION : Dunn v McCarthy [2010] NSWSC 675

HEARING DATE(S) : 21/06/2010

JUDGMENT DATE : 7 July 2010

JURISDICTION : Equity Division

JUDGMENT OF : Macready AsJ at 1

DECISION :
1. In lieu of the legacy of \$30,000 in favour of the plaintiff contained in the will of the late Frederick Milton Gear dated 20 February 2008, the plaintiff will receive a legacy in the sum of \$160,000.
2. Interest to be paid on the balance of such legacy if it is not paid within three months from today's date and if not so paid, to bear interest at the rate prescribed for under the Probate and Administration Act 1898.
3. That the plaintiff's costs to be paid on the ordinary basis and the defendant's costs on an indemnity basis be paid or retained from the estate of the deceased.

CATCHWORDS : Family Provision. Application by stepson given modest legacy. Legacy increased. No matter of principle.

SOURCES : James Nevill Dunn v Maureen John McCarthy
All rights reserved.

NSW Supreme Court
Inheritance law judgments

Name	James Nevill Dunn
Role in trial	Plaintiff
Costs	\$40,000 (p6)
Natural person?	true
Relationship to deceased	stepson (p1)
Was dependent on deceased?	partly dependent (p32)
Alleged misconduct?	false
Estranged from deceased?	briefly estranged due to a misunderstanding (p43)
Financial circumstances	receives a pension of \$520 per fortnight, wife earns \$800 per fortnight, owns a house with a \$15,000 mortgage (p39-40)



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

Semantic search on private data

Digital Criminal Justice Project–
Dr. Carolyn McKay



SIH: Nathaniel Butterworth,
Gordon McDonald, Marius Mather
bit.ly/sih_search

Local language model
(not in the cloud)

Survey and Interview transcripts
⚠️ Highly protected data

Semantic Query
food



Find

100%

Before the sun had risen, the wife came and awakened the two children, saying, "Get up, you lazy bones; we are going into the forest to cut wood." Then she gave each of them a piece of bread, and said, "That is for dinner, and you must not eat it before then, for you will get no more." Grethel carried the bread under her apron, for Hansel had his pockets full of the flints. Then they set off all together on their way to the forest. When they had gone a little way Hansel stood still and looked back towards the house, and this he did again and again, till his father said to him, "Hansel, what are you looking at? take care not to forget your legs."

6 "O father," said Hansel, "I am looking at my little white kitten, who is sitting up on the roof to bid me good-bye." - "You young fool," said the woman, "that is not your kitten, but the sunshine on the chimney-pot." Of course Hansel had not been looking at his kitten, but had been taking every now and then a flint from his pocket and dropping it on the road. When they reached the middle of the forest the father told the children to collect wood to make a fire to keep them, warm; and Hansel and Grethel gathered brushwood enough for a little mountain; and it was set on fire, and when the flame was burning quite high the wife said, "Now lie down by the fire and rest yourselves, you children, and we will go and cut wood; and when we are ready we will come and fetch you."

7 So Hansel and Grethel sat by the fire, and at noon they each ate their pieces of bread. They thought their father was in the wood all the time, as they seemed to hear the strokes of the axe: but

Greater Sydney Dataset



THE UNIVERSITY OF
SYDNEY
—
Sydney
Informatics Hub

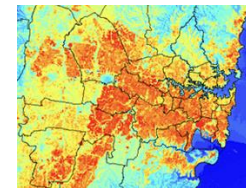
>2 million tree patches
1 million buildings

Understanding cities with computer vision

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



Urban Heat ML Model

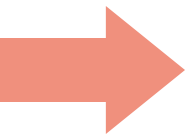


Geospatial Features

Local Climate Zone = "2: Compact Midrise"

(49% class confidence)

Input Image



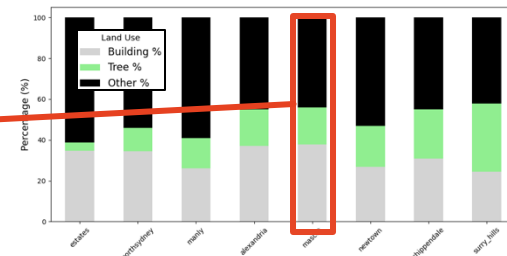
Computer Vision



bit.ly/aigis

Insights, ML, Analysis

Compare neighborhoods



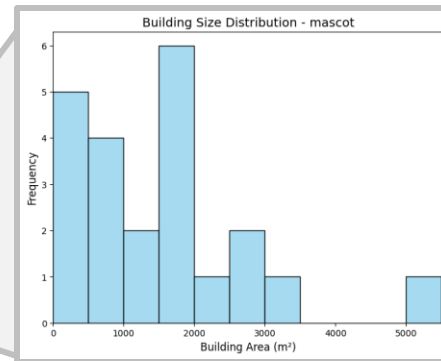
RGB Visual Light
Satellite Imagery
Aerial Imagery (Best)

Tree Patches

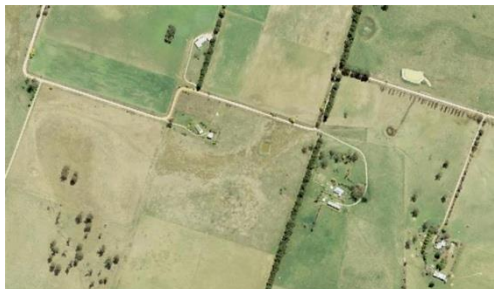
- 101 tree patches
- 15% coverage

Building Outlines

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



Agricultural land use models, urban ecology



This tool will use the produced datasets to compare the soil moisture of a farm against historic data. Please select the desired comparison method and dates to make the comparison as in section A. Then choose the visualisation in section B to see the results.

A

☒

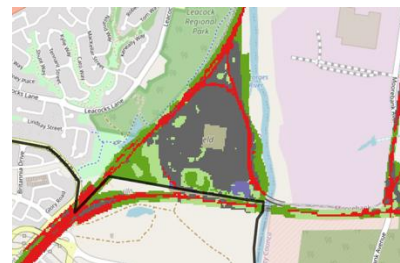
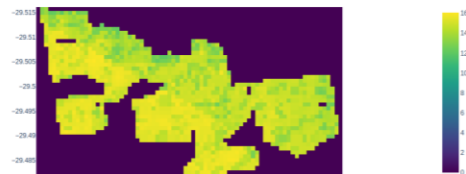
Select soil layer:

Select window aggregation method:

B

Select visualisation name:

Results



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.

Speeding up microscopy data processing

SIH: Sebastian Haan, Nathaniel Butterworth

bit.ly/sih-micro

Key highlights



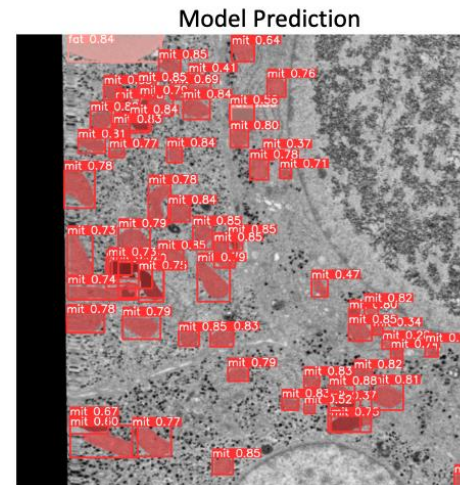
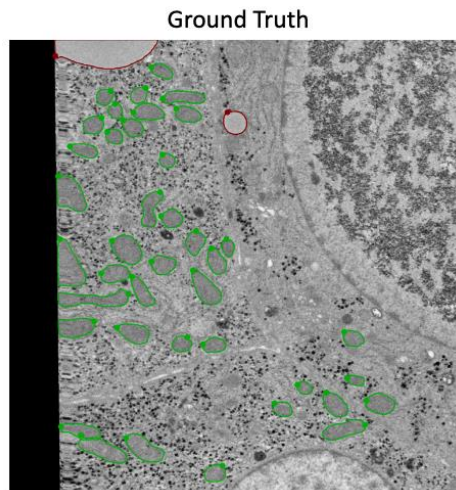
Sped up annotation process by up to 10x existing approaches



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. We developed AI pipelines to speed up cell segmentation in electron microscopy images, automating 3D cell tomography.

Australian Text Analytics Platform

Text analysis tools for all researchers



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
- www.atap.edu.au/posts/quotation-tool/



Discursis

- An analysis and visualisation tool for conversational data
- 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

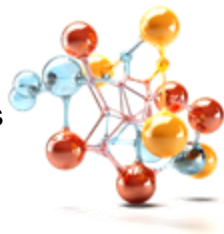
... people were on ventilators and people were dying at that stage, and that's the context in
QUOTE
... I was n't aware of the decisions. " Former finance minister Simon Birmingham said
QUOTE SPEAKER PERSON
/ whether Mr Morrison should leave. Nationals leader David Littleproud said Australians should
PERSON SPEAKER ORG PERSON QUOTE NORP
Albanese is awaiting advice from the solicitor - general on whether Mr Morrison's actions have any legal



Generative X

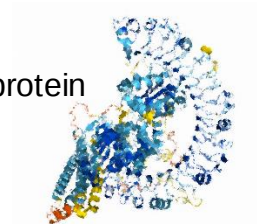
Drug discovery

Design your own chemicals



Biology

Design your own protein
synthetic biology

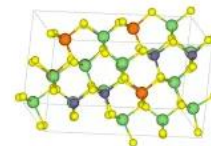


Manufacturing

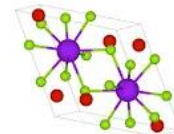
Design me a diffraction grating / micro resonator / widget with these properties...

Material discovery

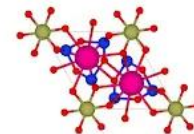
AI mineral discovery and
autonomous material synthesis



$\text{Li}_4\text{MgGe}_2\text{S}_7$



KV_2Se_3



$\text{Rb}_2\text{HfSi}_3\text{O}_9$



Sydney Informatics Hub

Dr Gordon McDonald, Informatics Team Lead

gordon.mcdonald@sydney.edu.au

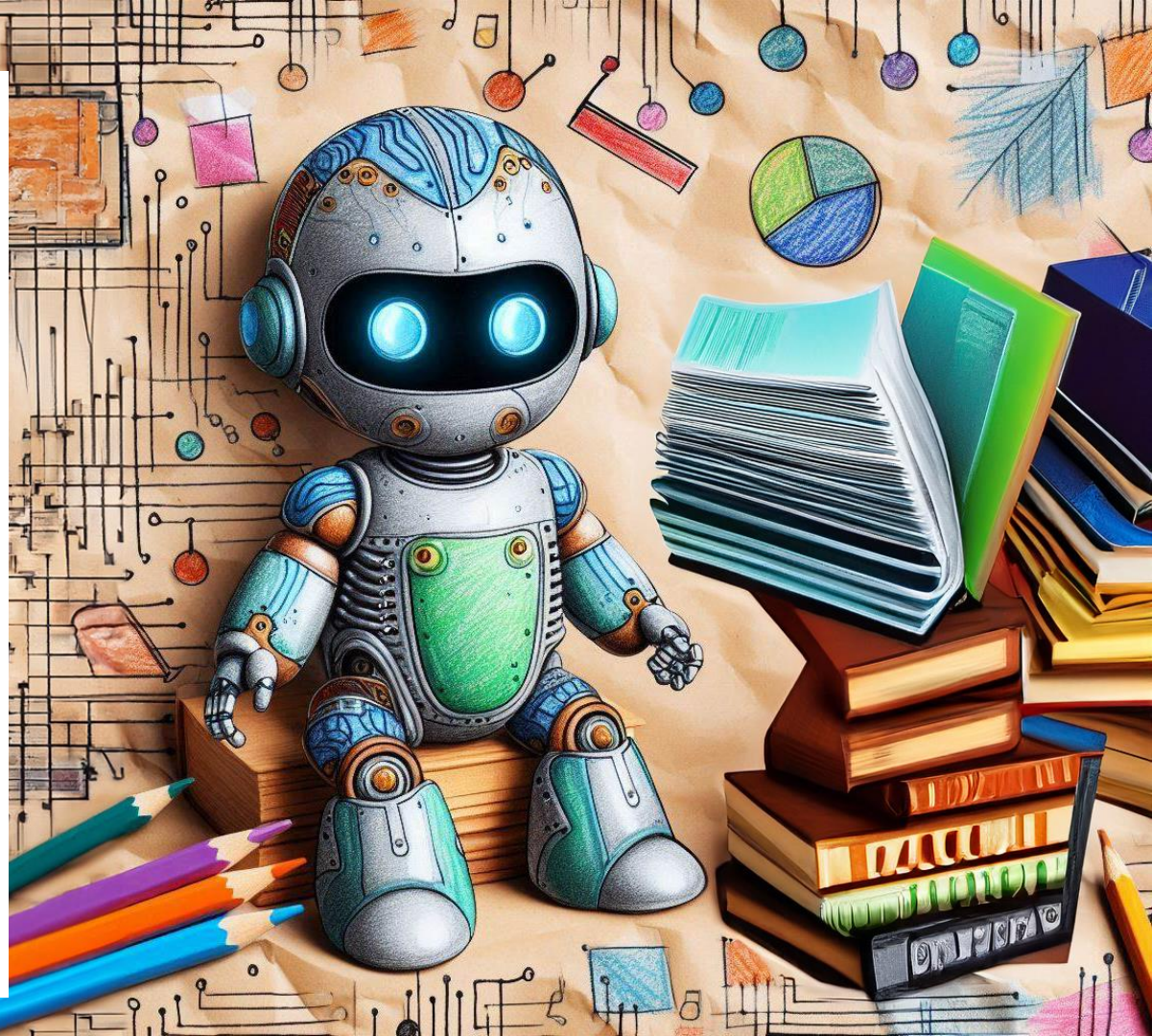
sydney.edu.au/informatics-hub

sih.info@sydney.edu.au



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub



Some Tools and Platforms

Audio

Translate audio between different languages, keeping your tone of voice

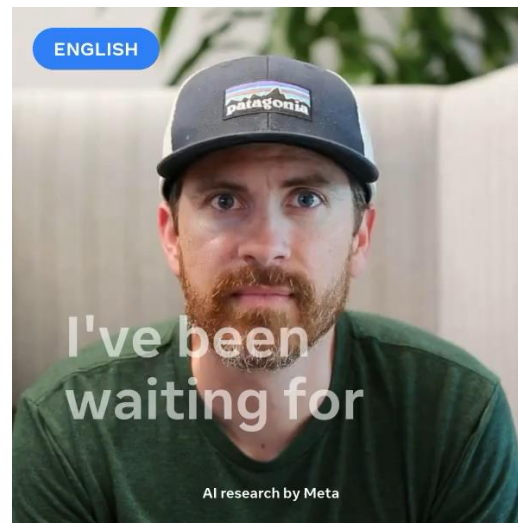
<https://seamless.metademolab.com/expressive>

Generate Speech

<https://elevenlabs.io>

Generate Music

<https://suno.com>



On-device video transcription

If you're on a mac

sindresorhus.com/aiko

Data and model runs in your browser – word level timestamps

huggingface.co/spaces/Xenova/whisper-word-level-timestamps

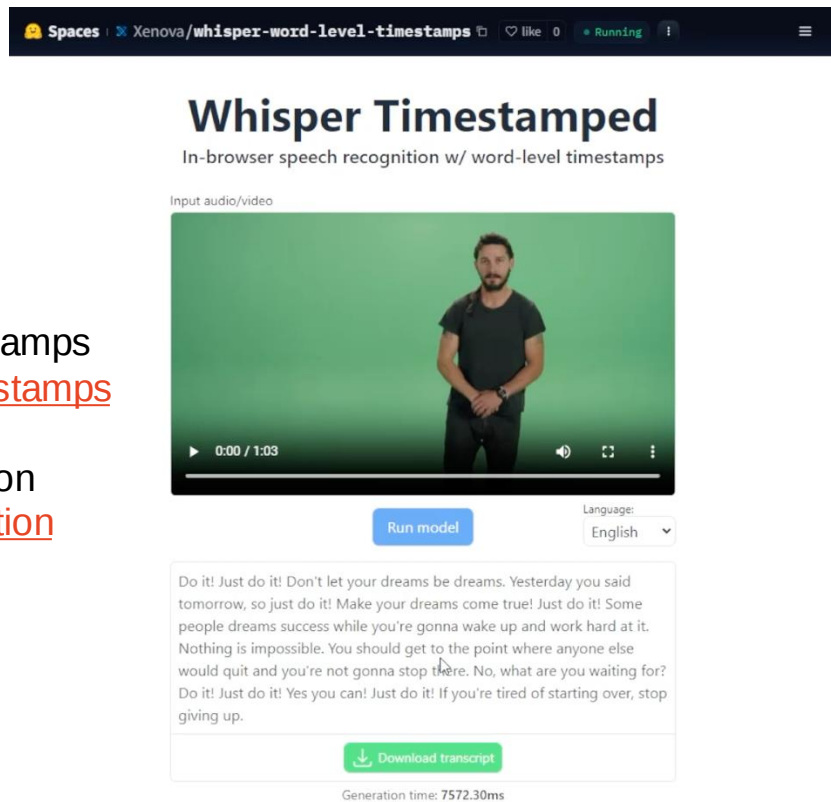
Data and model runs in your browser – speaker diarization

huggingface.co/spaces/Xenova/whisper-speaker-diarization

If you're ok with code and installs

github.com/ggerganov/whisper.cpp

(Not on device) Sharepoint automatic video transcription



The screenshot shows the Hugging Face Spaces interface for the 'Whisper Timestamped' demo by Xenova. The header bar includes the Spaces logo, the user 'Xenova', the space name 'whisper-word-level-timestamps', and a 'Running' status indicator. The main title 'Whisper Timestamped' is prominently displayed, followed by the subtitle 'In-browser speech recognition w/ word-level timestamps'. Below this, a video player shows a man speaking against a green background. The video player has a progress bar at 0:00 / 1:03 and standard playback controls. To the right of the video player is a 'Run model' button and a language dropdown menu set to 'English'. Below the video player, a text box displays the transcribed speech: 'Do it! Just do it! Don't let your dreams be dreams. Yesterday you said tomorrow, so just do it! Make your dreams come true! Just do it! Some people dreams success while you're gonna wake up and work hard at it. Nothing is impossible. You should get to the point where anyone else would quit and you're not gonna stop there. No, what are you waiting for? Do it! Just do it! Yes you can! Just do it! If you're tired of starting over, stop giving up.' Below the text box is a green 'Download transcript' button. At the bottom, the generation time is shown as '7572.30ms'.

Image segmentation

Segment a human

huggingface.co/spaces/facebook/sapiens-seg

Segment images

sam2.metademolab.com

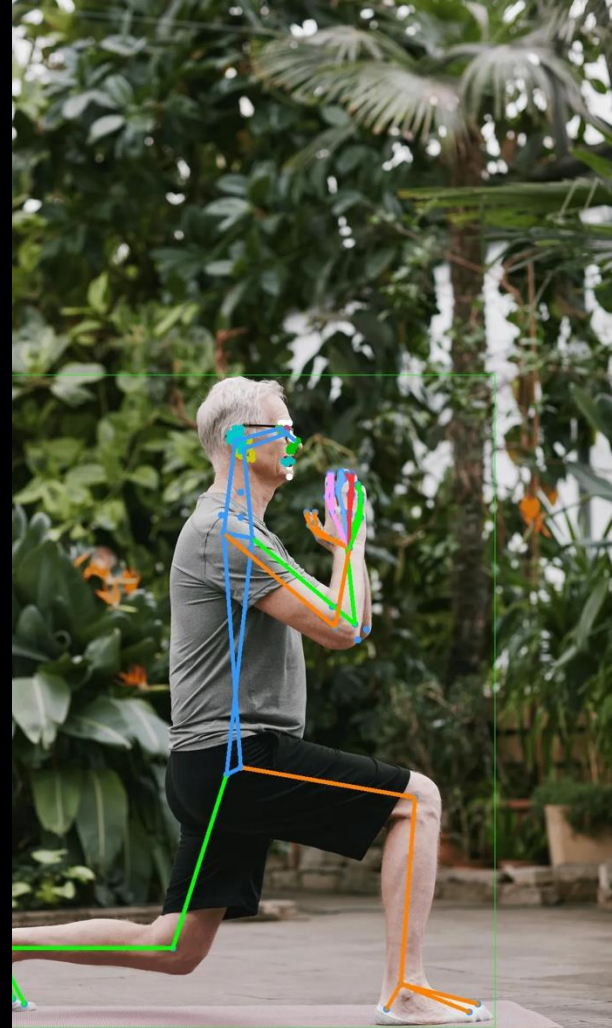
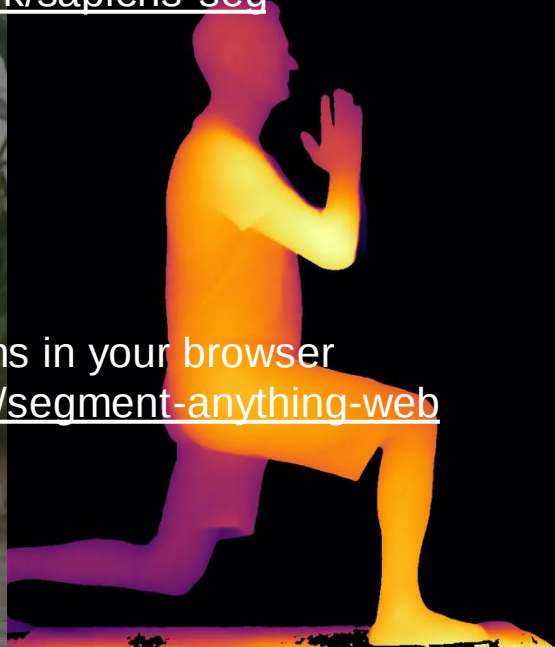
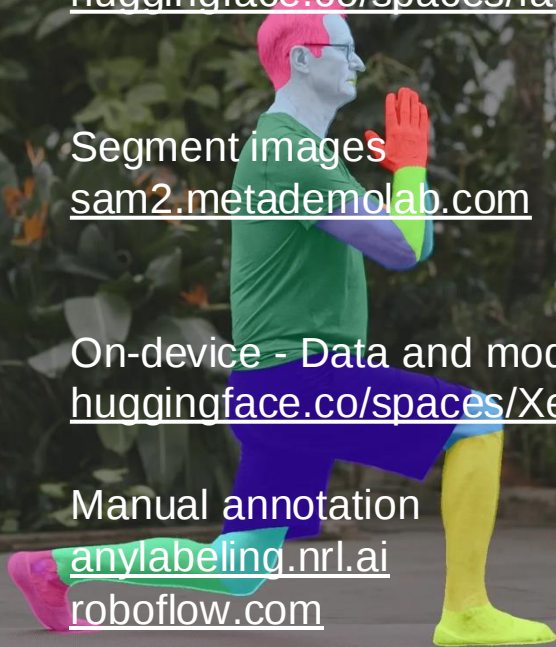
On-device - Data and model runs in your browser

huggingface.co/spaces/Xenova/segment-anything-web

Manual annotation

anylabelling.nrl.ai

roboflow.com



Platforms & Tools

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

Literature Search

Research Rabbit – free for researchers, operates on donation model

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)

Aider

Jupyter-AI

Open Interpreter

MetaGPT

Copilot

Paid

ChatGPT

Github Copilot

Best language models to use
with these tools now (10-09-24):

- claude-3.5-sonnet (via API)
- gpt-4o (via API)
- DeepSeek 2.5 (via API)
- Local models? (none of them are quite good enough yet to be useful) ☹️

Agent building



Agents

to reduce
time to science

Free (ish)

Cogniti

Flowise

MetaGPT

Paid

ChatGPT Custom GPTs

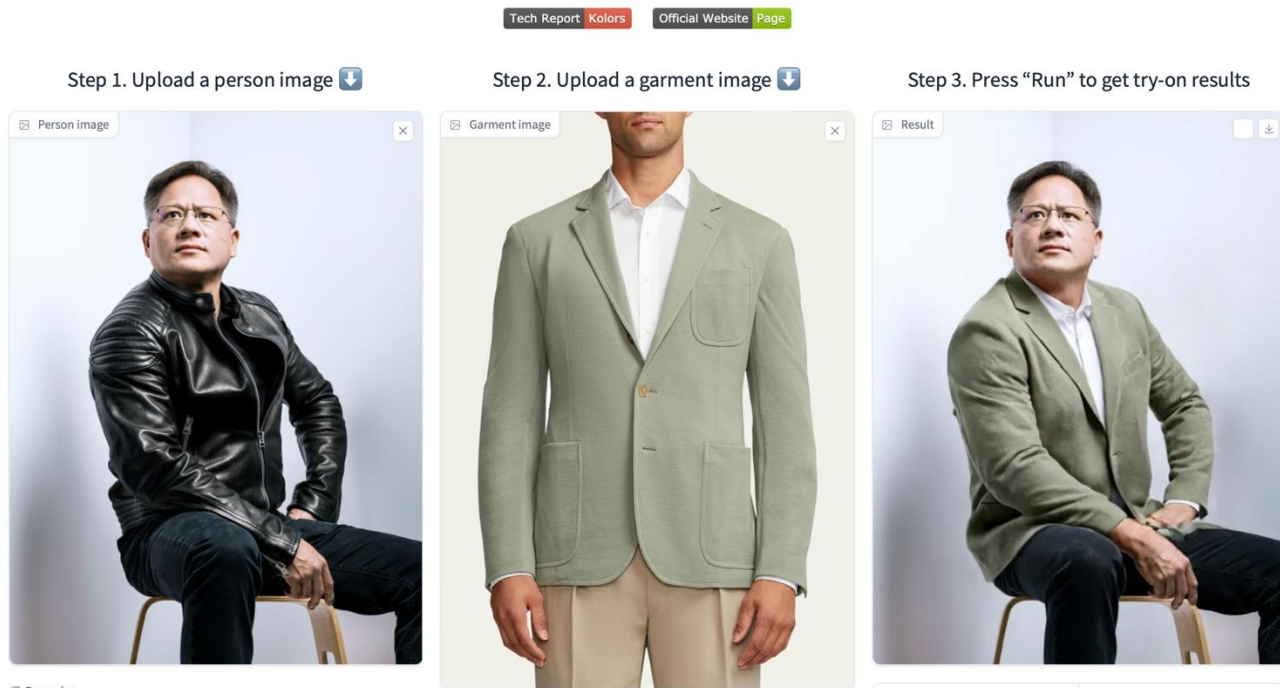
Azure Promptflow

Try on virtual clothes!

To find more new models as they are released:

<https://huggingface.co/spaces>

Kolors Virtual Try-On in the Wild





Sydney Informatics Hub

Dr Gordon McDonald, Informatics Team Lead

gordon.mcdonald@sydney.edu.au

sydney.edu.au/informatics-hub

sih.info@sydney.edu.au



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

