AHLI SUMMER CAMP 2026 · DAY 2 — AFTERNOON LECTURE

Beyond EHR

Finding and accessing diverse health data.

Solly Sieberts · Jineta Banerjee
Sage Bionetworks

Sage Bionetworks

WHAT THIS TALK COVERS

Five questions, one idea: data is part of the problem, not an afterthought

01	**The health data landscape** How varied it is, and how big it gets	Solly
02	**Why it lives behind walls** Consent, access tiers, trusted research environments	Solly
03	**Working with it responsibly** Re-identification, redistribution, and agents	Jineta
04	**Where next** Context, confounders, and when you generate data	Jineta
05	**What to take to your project** The questions for your small group this afternoon	Jineta

THE MENTAL MODEL YOU ARRIVED WITH

For most of machine learning, data is a URL

WHAT YOU'RE USED TO

`pip install` the dataset, or click download

It's on Hugging Face, Kaggle, or Google Dataset Search

If you can't find it there, it doesn't exist

WHAT HEALTH LOOKS LIKE

That URL is a request, a review, and a data-use agreement

The data may never leave a secure environment

Behind every record is a person who consented

01

The health data landscape

Before we talk about finding data, it helps to see how varied — and how large — health data really is.

HEALTH DATA IS NOT ONE THING

Six families of data, each with its own shape and tooling

Clinical / EHR

Structured records, labs, billing codes, vitals — tabular and longitudinal.

MIMIC, eICU

Clinical notes

Free-text documentation — the natural-language layer of the record.

MedAlign, discharge summaries

Genomics & omics

Sequence, expression, single-cell — large, binary, pipeline-processed.

VCF, CRAM, RNA-seq

Medical imaging

Radiographs, pathology slides, retina — not just MRI.

DICOM, whole-slide images

Patient-generated

Wearables, phone sensors, self-report — continuous and messy.

Actigraphy, app streams

Multi-omic & linked

Several modalities on the same person — the hardest to assemble.

Linked cohorts, biobanks

START FROM WHAT YOU KNOW

Most of you know MIMIC best — and that shapes your instincts

An ICU EHR dataset is tabular, mostly text, a few gigabytes, and de-identified enough to download. That's a perfectly reasonable picture of "health data" — it's just a small corner of it.

READING Rajkomar et al., *Scalable and accurate deep learning with EHR* , npj Digital Medicine (2018)

A SENSE OF SCALE

56 GB

is one whole-genome CRAM file —
for a single person.

MIMIC-IV (EHR)
~a few GB

Chest X-ray
~10 MB each

Pathology slide
1–4 GB each

One genome (CRAM)
~56 GB / person

Genomics cohort
petabytes

A PRACTICAL CATCH

Most of it isn't text — it needs processing before you can model it

An LLM can read a clinical note out of the box. It cannot read a CRAM, a DICOM series, or a single-cell matrix. Each modality carries its own formats, pipelines, and domain conventions between the raw file and anything you'd put in a model.

HEALTH DATA IS NOT ONE THING

Basic -omics you should know

Genomics

Genetic sequence variants (SNPs/SNVs), insertion/deletions, structural variants, etc..

A/T/C/G, dosages, CNVs

Epigenomics

Methylation, histone modifications, 3D DNA structure measure chromatin accessibility.

Some text

Transcriptomics

Amount of (coding and non-coding) RNA produced by a cell.

Some text

Proteomics

Protein abundances measured from using a variety of different technologies.

Metabolomics

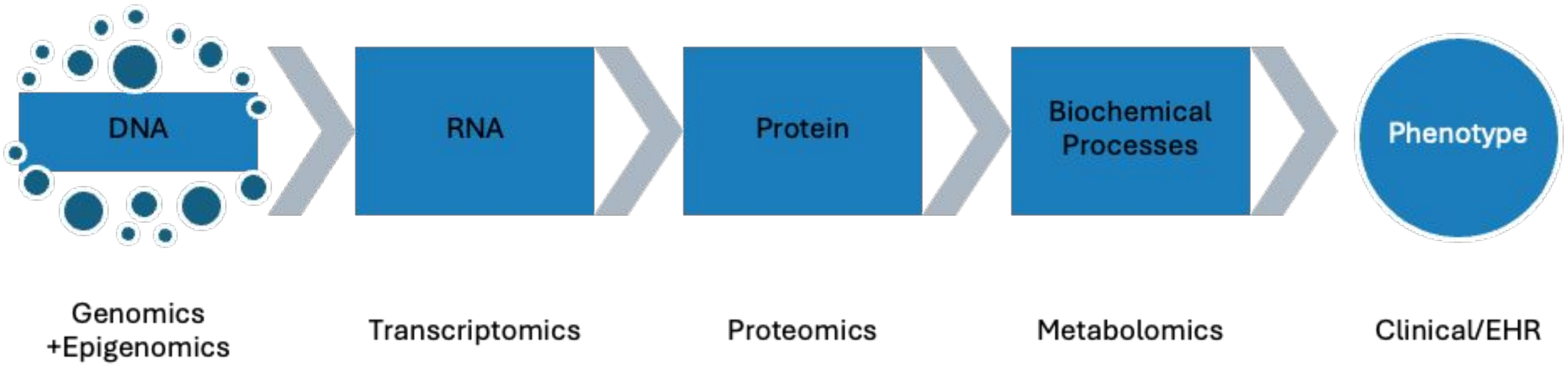
End product of cellular metabolism (amino acids, lipids, sugars, etc).

Metagenomics

Sequencing of microbial communities to understand microbiome makeup.

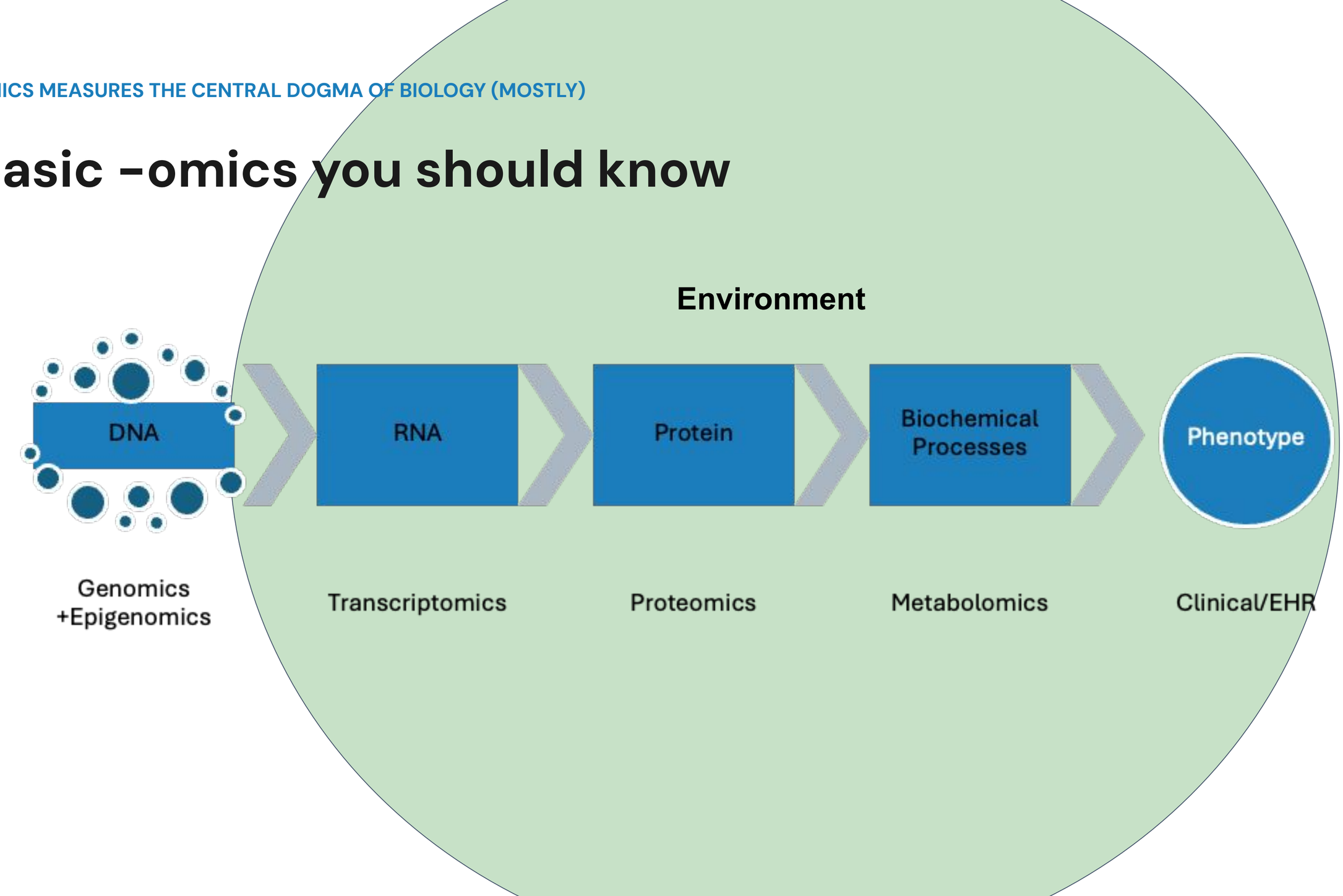
OMICS MEASURES THE CENTRAL DOGMA OF BIOLOGY (MOSTLY)

Basic –omics you should know



OMICS MEASURES THE CENTRAL DOGMA OF BIOLOGY (MOSTLY)

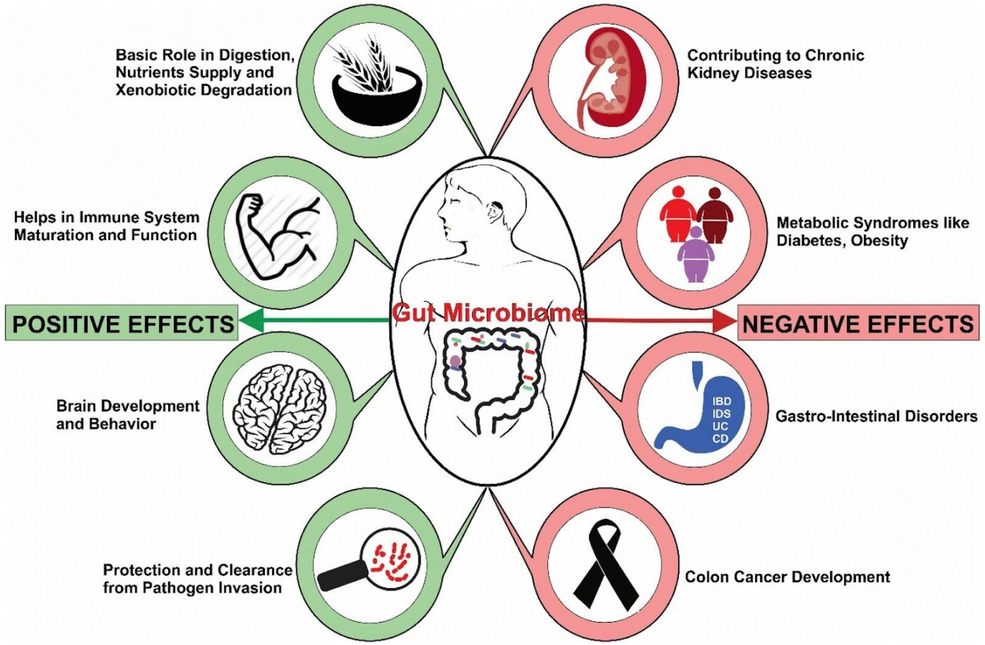
Basic –omics you should know



MICROBIOME: WE'RE NOT JUST US

Basic -omics you should know

Our bodies are host to a variety of bacterial, fungal and viral species in roughly equal numbers to our own cells. The microbiome has been linked to everything from neurological disorders to cancer. Different species can have positive or negative effects.

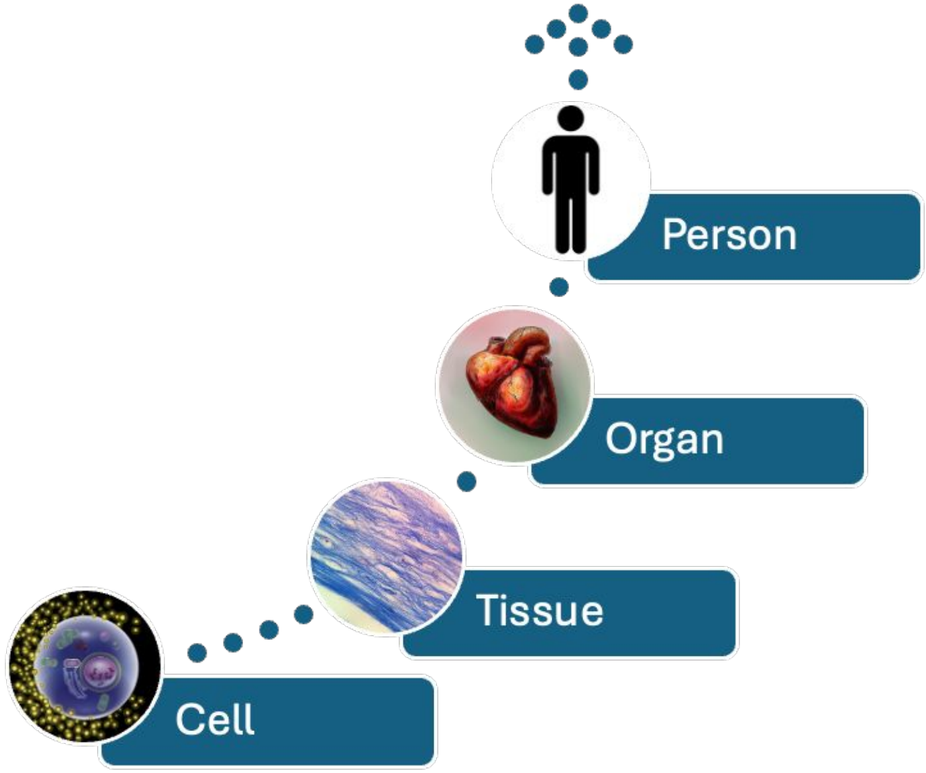


CONTEXT MATTERS

Be aware of tissue/cell type

For most -omics, the tissue of origin matters.

Technologies can assay at the single-cell or bulk tissue scale.



WHERE THE DATA ACTUALLY LIVES

The repositories worth knowing

Synapse

Open-science platform for sharing data, code, and models with fine-grained governance.

We help run this one

dbGaP

NIH's controlled-access archive for genotype–phenotype studies.

Controlled access

All of Us

1M+ participant cohort; analysis happens inside the Researcher Workbench.

Enclave

AD Knowledge Portal

Alzheimer's data, analyzed in connected environments like AD Workbench.

Controlled access (TRE integration)

UK Biobank

Deep multi-modal cohort — imaging, genomics, EHR, accessed under approval.

Controlled access

PhysioNet

Where MIMIC lives — credentialed access for clinical signals and records.

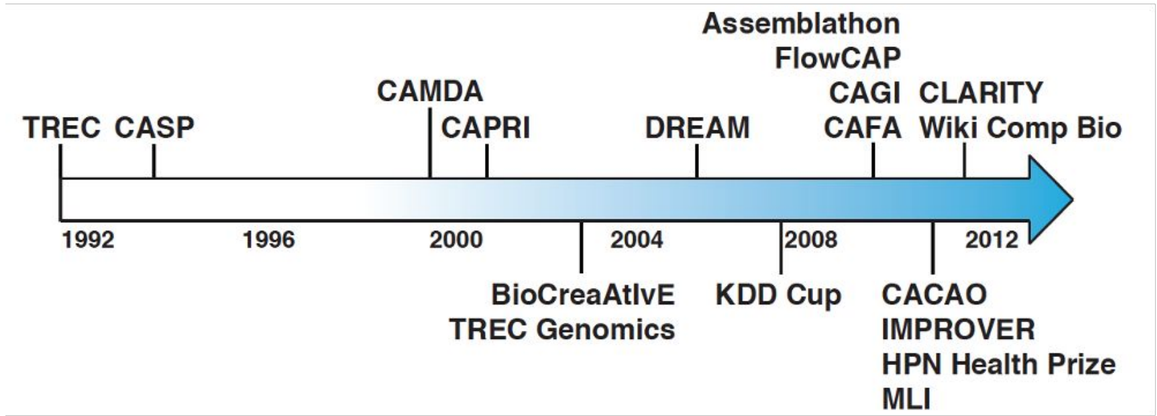
Registered / credentialed

And hundreds more: https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html

BEYOND DATA: CHALLENGES

Beyond Kaggle: Finding important problems

Challenges are a great way to find important biomedical problems (and data). There are a variety of active biomedical challenges communities.



CHALLENGES: THE SELF-ASSESSMENT TRAP

The ML literature is like Lake Wobegon

“Where the women are strong, all the men are good-looking, and all the children are above average.”

CHALLENGES DRIVE IMPORTANT INNOVATION

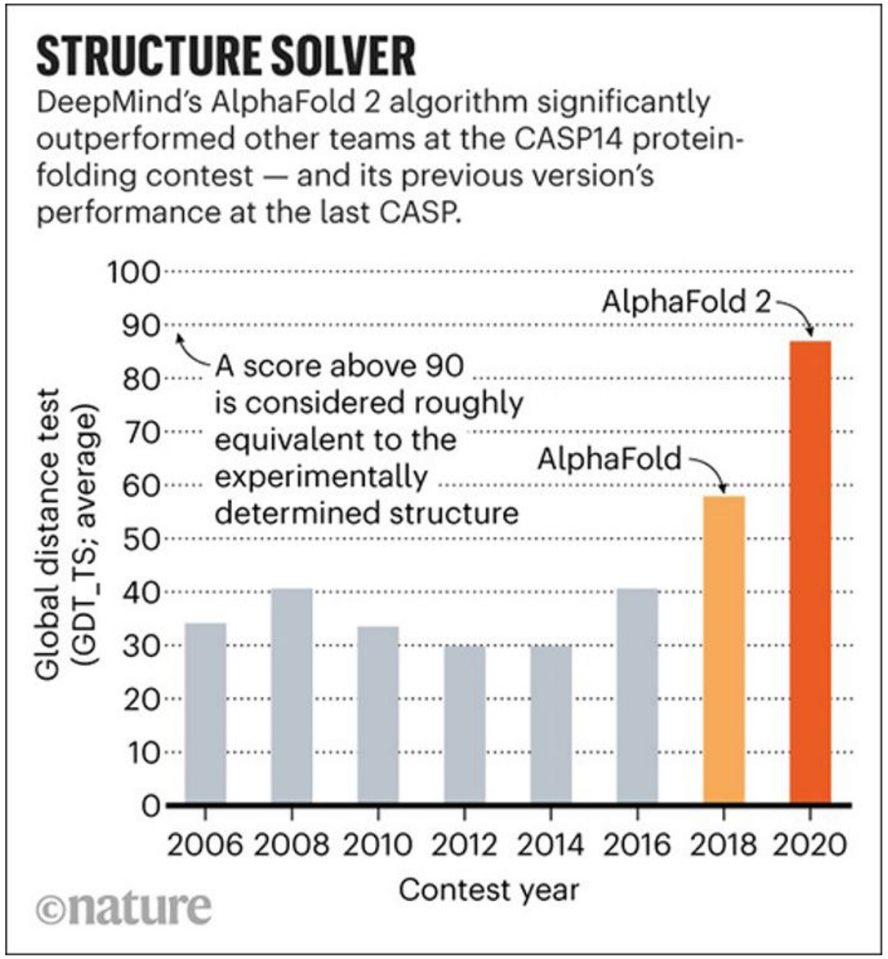
Challenges can be Nobel Prize-worthy

NEWS | 09 October 2024

Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures

This year's prize celebrates computational tools that have transformed biology and have the potential to revolutionize drug discovery.

https://www.nature.com/articles/d41586-024-03214-7



CHALLENGES PLATFORMS

Finding Challenges

From Kaggle to Synapse to Hugging Face to Grand Challenges, there are a number of challenges platforms. Open Challenges (openchallenges.io) is a one-stop location for finding active challenges.

BENCHMARKING WHEN YOU CAN'T SHARE THE DATA

How do you compare methods on data nobody can hold?

Kaggle's model doesn't fit. You can't hand everyone a download and a leaderboard when the data is sensitive.

Challenges still work. A shared task, a held-out truth, and a fair scoreboard — run against private data.

Reproducibility is the real prize. Benchmarks tell the field which methods actually generalize.

MODEL-TO-DATA

Bring the model to the data, not the data to the model

No one touches the raw data — so even highly sensitive cohorts can drive an open, competitive benchmark.

02

Why it lives behind walls

In health, the friction is a feature. Understanding it is how you actually get to the data.

WHY YOU CAN'T JUST DOWNLOAD IT

The friction protects the people in the data

Consent is specific. Participants agreed to certain uses, not to "anything, by anyone, forever."

It's protected health information. Re-identification can cause real harm, so the law and the IRB set the terms.

Governance is a person, not a checkbox. A data access committee reviews who you are and what you intend to do. * Use of AI may need to be disclosed.

The cost is real. We lose researchers at this step — many requests need more than one submission to get through.

THE TWO WORLDS OF DATA

There's the indexed world — and a much larger one you can't search

THE WORLD YOU CAN SEE

Hugging Face, Kaggle, Google Dataset Search
One click from a model to a download
Great for benchmarks; thin on real patients

THE WORLD YOU CAN'T

Controlled-access repositories and biobanks
Discoverable by metadata, not by the data itself
Where most consented health data actually lives

TIERS OF ACCESS

Access is a spectrum, not a yes/no

Open	Anyone can download. De-identified or aggregate, low re-identification risk.	**e.g. public reference data**
Registered	Free, but you log in and agree to terms — identity and intent on record.	**e.g. click-through DUA**
Controlled	A data access committee reviews your request and approved use.	**e.g. dbGaP, Synapse controlled**
In a TRE	Data never leaves; you bring your analysis to it inside a secure enclave.	**e.g. All of Us, AD Workbench**

HOW ACCESS ACTUALLY WORKS

The path from "I found it" to "I can use it"

01

Discover by metadata

You browse descriptions and schemas — not the records themselves

02

Request access

State who you are, your institution, and your intended use

03

Review & agreement

A committee approves; you sign a data-use agreement (sometimes an IRB)

04

Download — or compute in place

The data comes to your machine, or you go to it inside a secure environment

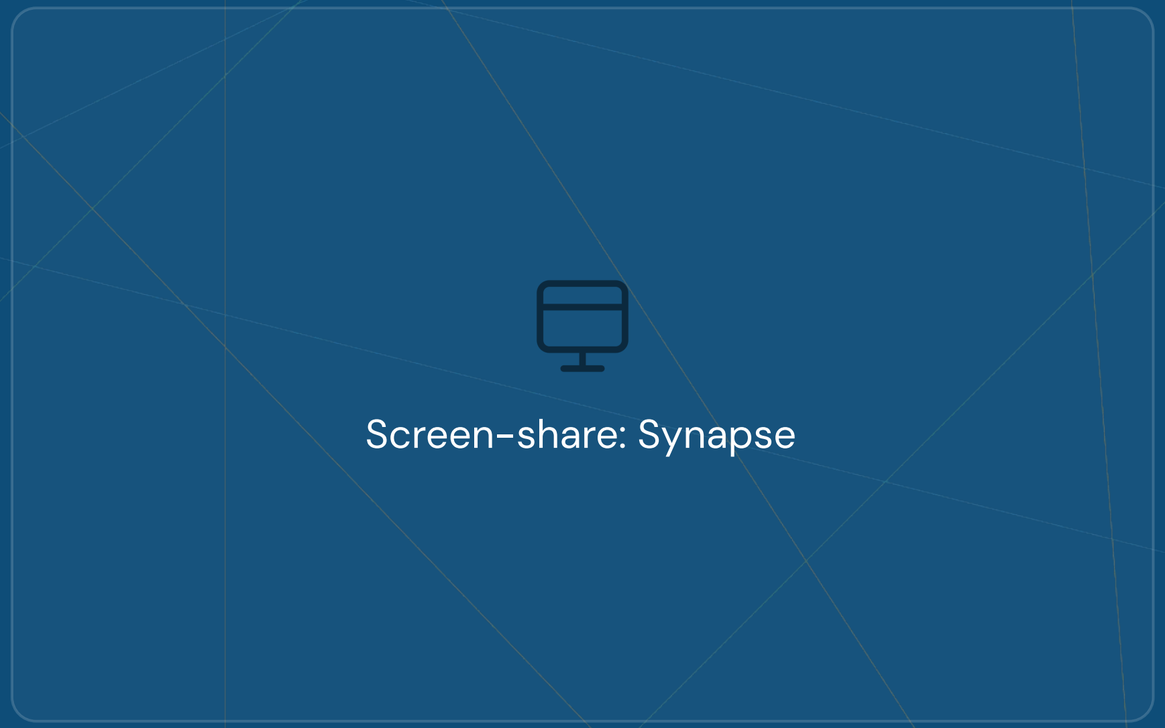
TRUSTED RESEARCH ENVIRONMENTS

When the data can't come to you, you go to the data

A TRE (trusted research environment) is a secure enclave with the data already inside. You log in, get compute and tools, run your analysis there, and take only approved results out. For sensitive cohorts this is increasingly the default — so it's worth being comfortable working inside one.

LIVE DEMO

Requesting access on Synapse



Find a dataset by its **metadata and schema**

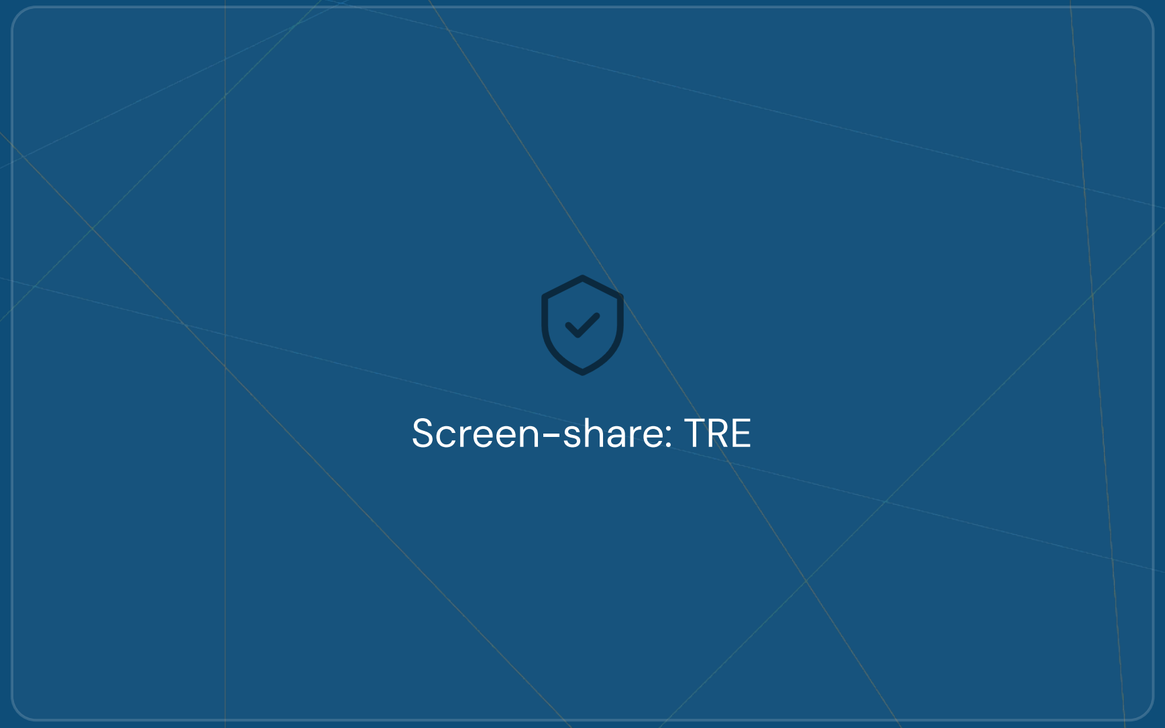
Read the **conditions for use** on the data page

Submit an **access request** and see what review looks like

Show the difference between **open** and **controlled** tiers

LIVE DEMO

Working inside a Trusted Research Environment



Log into the **All of Us Researcher Workbench**

Build a cohort and notice the data **never leaves**

Compare with **AD Workbench / Cavatica** for omics

See what you can — and can't — **take out**

Platforms: All of Us · AD Workbench · Cavatica

03

Working with data responsibly

Access comes with obligations — and they get sharper the moment you hand the keys to an agent.

WHAT RESPONSIBLE USE MEANS

Two promises you make when you get access

DON'T REDISTRIBUTE

The data was shared with *you*, for *your* approved use. You can't repost it, hand it to a collaborator who wasn't approved, or push it to a public bucket — even to be helpful.

DON'T RE-IDENTIFY

You can't try to link records back to real people, and you can't combine datasets in ways that would. The de-identification only holds if everyone respects it.

WHAT RESPONSIBLE USE MEANS

Can molecular data have re-identification risk?

YES

Genomic variants called from a genome can act as fingerprints for an individual

YES

Multiple -omics data like variants, gene expression, others when linked together can be identifying of an individual.

WHAT RESPONSIBLE USE MEANS

Can synthetic data have re-identification risk?

YES

If synthetic data generation does not include privacy preserving mechanisms, they can be identifying - e.g. low tolerance to Membership Inference Attacks

YES

Synthetic omics data can also be prone to re-identification when combined with other omics data - e.g. low tolerance to linkage attacks.

THE AGENTIC ERA

Can an agent work on this data?
Yes — but with caution

The same rules apply to your tools. An agent acting for you inherits your obligations — including "don't redistribute."

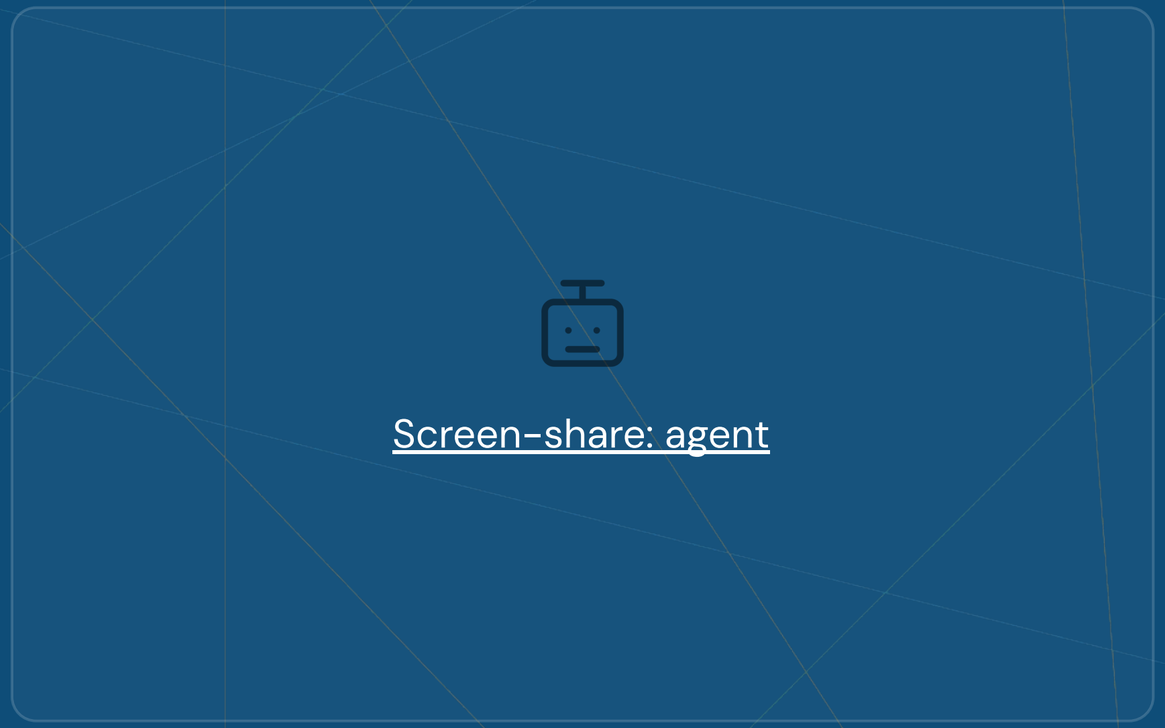
Egress is the risk. A coding agent that can hit the open internet can exfiltrate data without meaning to.

So keep the work inside the walls. Run agents within TRE, only with agents that come with assurance that they dont transmit anything to their developer organization (i.e. enterprise level agents, not Claude Code with your personal API Key)

This is where the field is going. Just as software engineers now code *with* agents, computational biology is starting to as well.

LIVE DEMO

An agent working inside the data



Screen-share: agent

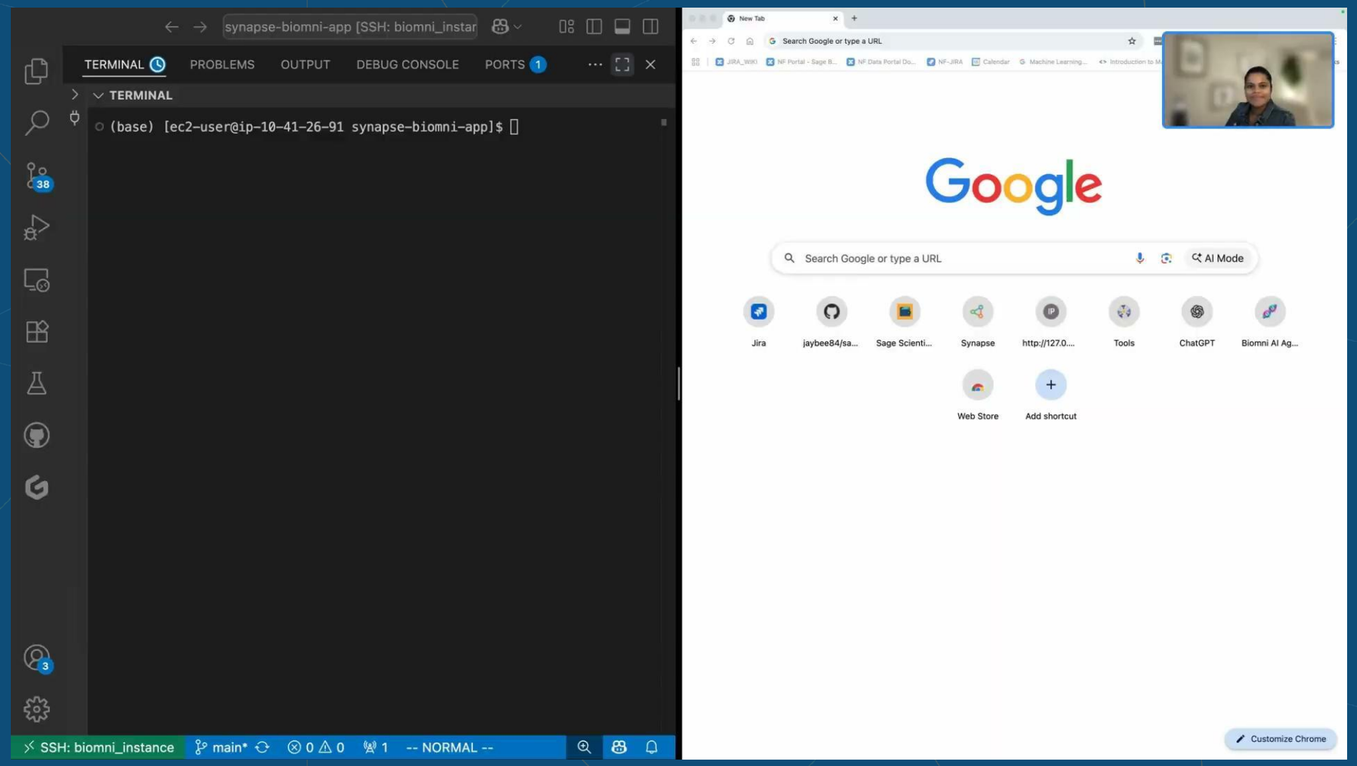
Point an agent at an **open dataset** and ask a real question

Watch it **pick tools** and write analysis code

Tools: Biomni · co-scientist workflows

LIVE DEMO

An agent with controlled access



Point an agent at a restricted **dataset** and ask a real question

Watch it **pick tools** and write analysis code

Note it stays **inside the environment** — no egress

Tools: Biomni · co-scientist workflows

04

Where this is heading

Two bets on multimodal health AI — and why the bottleneck isn't what you'd guess.

TWO BETS ON MULTIMODAL HEALTH AI

Orchestrate specialist tools — or train one model on everything

BET ONE · TANGIBLE TODAY

Reasoning orchestrators

An LLM reasons through a problem and calls specialist tools — AlphaGenome, ESM, AlphaFold — stitching a workflow across modalities.

BET TWO · FURTHER OUT

Multimodal foundation models

One model trained to align every modality in a shared space, CLIP-style. Powerful in principle — and very hungry for data and compute.

TWO CAMPS ON DATA

Ask around and you'll hear two opposite complaints

CAMP ONE

"We already have the data."

It's out there — we just can't find it. The problem is discovery: petabytes of *dark data* sitting unindexed in repositories and hard drives.

CAMP TWO

"What's out there is unusable."

The public data is trash for our question — we have to generate new data under proper controls before we can train anything.

WHY BOTH CAMPS ARE PARTLY RIGHT

Most public data was generated to answer one question

Traditional hypothesis driven science has led to data that is currently available in the public domain

Each dataset is usually built for a single hypothesis, in a single experiment.

It measures exactly what that study needed — and, reasonably, not much else. Anticipating every future question is simply too hard at collection time.

Incorporates various biases of the scientist

THE PART COMPUTATIONAL FOLKS MISS

The controls you'd need for *your* question were never measured

Great for its question, not comparable to anything else. Different conditions, batches that aren't controlled across studies.

Unmeasured confounders are everywhere. What wasn't recorded can't be adjusted for later.

No experimental background, no instinct for this. If you've never run the bench experiment, the gaps are invisible.

This is "problems and data are not independent." The data already encodes someone else's problem definition.

Exceptions: High throughput unbiased datasets like whole genome sequencing - unbiased from the perspective of the sequencing, but may have other sampling biases like being non-representative of all demographics

READING Yuan et al., *Temporal bias in case-control design* , Nature Communications (2021)

Treat data as a first-class citizen in defining your problem — not as a step you reach at the end.

Problems and data are not independent. The dataset you choose has already made choices for you.

WHEN FINDING ISN'T ENOUGH

How about synthetic data?

When the data you need don't exist — or the real data is too sensitive to move — synthetic data is an option: data you create to behave like the real thing, without being anyone's record.

But jury is still out on its utility, specially in the bioinformatic domain

A WORKED EXAMPLE

Generating synthetic omics, end to end

01

Start from a real controlled dataset

e.g. variant calls (VCF) or RNA-seq from an approved cohort

02

Train a generative model

Learn the joint distribution — including the structure that matters biologically

03

Produce synthetic replicas

New samples that look real but map to no real person

HOW YOU DECIDE TO TRUST IT

Three tests every synthetic dataset must pass

Fidelity

Do the distributions match the real data? The statistics should be hard to tell apart.

Utility

Does a downstream biological task still work? A classifier ML model is generally not a biological utility task

Privacy

Does it resist a membership-inference or re-identification attack?

The catch is the **privacy–utility trade-off** : push one and you strain the other. Getting that balance right is the whole game.

Data cards Data sheets

The bottleneck isn't compute. It's well-controlled, well-conceptualized data.

Compute is expensive. The gold-standard data needed to train these models is more expensive still.

WHEN FINDING ISN'T ENOUGH

Sometimes the honest answer is: go generate it

Most of the time you can find data you can work with. But when the controls you need don't exist sometimes you have to generate new data that is adequately unbiased to train an unbiased model

FOR SMALL GROUP THIS AFTERNOON

Take these two questions back to your project

1

What is the data generation process for your project?

2

What data collection already exists, independent of your AI system?