



National Institute of
Biomedical Imaging
and Bioengineering

NIH Bridge2AI Program

-- Creating trustworthy open data
for scientific discovery

NIEHS Superfund Research Program

Risk e-Learning Webinar Series: Advancing Environmental Health Research with Artificial Intelligence and Machine Learning

Session III — ML & AI Applications to Understand Omics, Metabolomics, & Immunotoxicity and Optimize Bioengineering Using Datasets, Models, and Mass Spectrometry

November 22, 2024

Grace C.Y. Peng, PhD

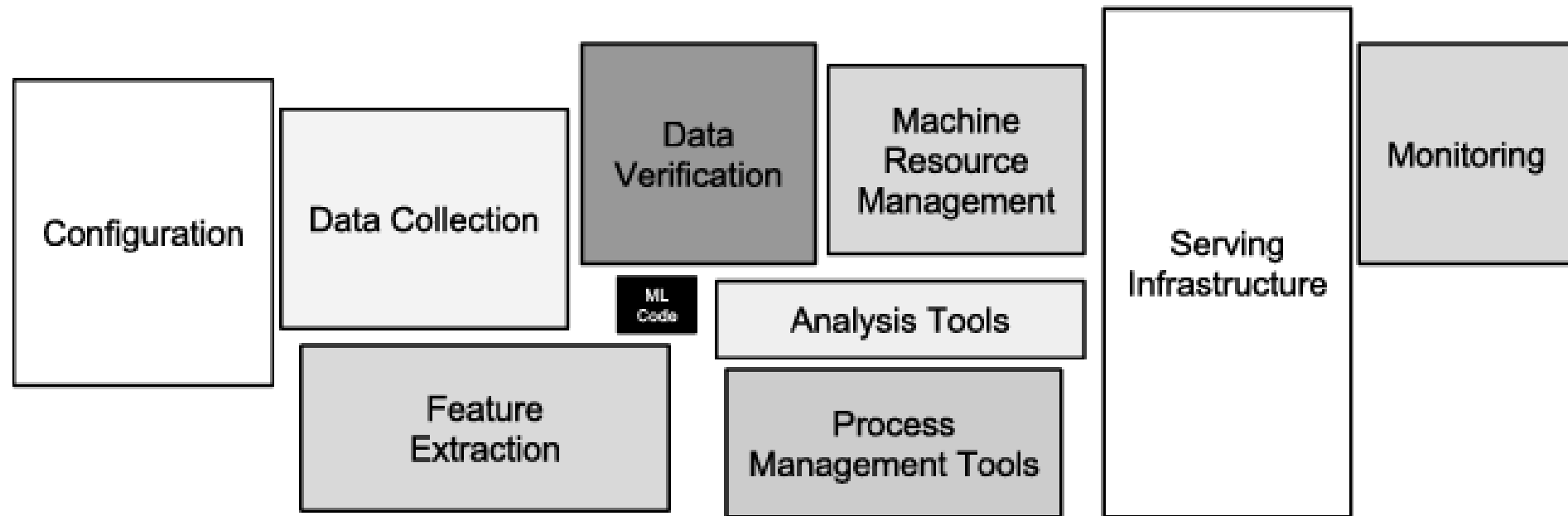


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Scully et al. (2015): Hidden technical debt in Machine learning systems [doi: 10.5555/2969442.2969519]

Artificial Intelligence Working Group Update

119th Meeting of the Advisory Committee to the Director (ACD)
December 13, 2019



David Glazer
Engineering Director, Verily

Lawrence A. Tabak, DDS, PhD
Principal Deputy Director, NIH
Department of Health and Human Services



- **December 6, 2019 ACD AI WG Report**
- https://acd.od.nih.gov/documents/presentations/12132019AI_FinalReport.pdf
- **December 13, 2019 ACD presentation**
- <https://acd.od.nih.gov/documents/presentations/12132019AI.pdf>

Report of the ACD AI WG

December 6, 2019

TABLE OF CONTENTS

| | |
|--|----|
| Fusing Biomedicine and Machine Learning | 2 |
| Opportunities | 3 |
| Challenges | 7 |
| Data Challenges | 7 |
| Consent Challenges | 7 |
| Ethics Challenges | 8 |
| People Challenges | 9 |
| Recommendations | 11 |
| Recommendation 1: Support flagship data generation efforts to propel progress by the scientific community. | 12 |
| Recommendation 2: Develop and publish criteria for ML-friendly datasets. | 14 |
| Recommendation 3: Design and apply “datasheets” and “model cards” for biomedical ML. | 16 |
| Recommendation 4: Develop and publish consent and data access standards for biomedical ML. | 17 |
| Recommendation 5: Publish ethical principles for the use of ML in biomedicine. | 18 |
| Recommendation 6: Develop curricula to attract and train ML-BioMed experts. | 19 |
| Recommendation 7: Expand the pilot for ML-focused trainees and fellows. | 21 |
| Recommendation 8: Convene cross-disciplinary collaborators. | 22 |
| Conclusion | 23 |
| Acknowledgements | 23 |

The NIH Bridge2AI Program

Supported by the NIH Common Fund

Bridge2AI Program Management Team

Co-Chairs

Michael Chiang
Eric Green
Helene Langevin
Steve Sherry
Bruce Tromberg

Common Fund Program

Leader

Haluk Resat

Common Fund Program

Officers

Chris Kinsinger
George Papanicolaou

Working Group Coordinators

James Gao, NEI
Lanay Mudd, NCCIH
Grace Peng, NIBIB
Shurjo Sen, NHGRI

Common Fund Staff

Natalie Vineyard (Comm)
David Dzamashvili (Ops)
Karen Kellton (Prog Mgmt)
Kristina Faulk (Prog Coord)

Awards Management

Kristen Kreuter (DOTM)
Erna Petrich (DOTM)

Federal Working Group (+100 Members)

CC, CIT, FIC, NCATS, NCI, NCCIH, NEI, NHGRI,
NIA, NIAID, NICHD, NIBIB, NIDA, NIDDK,
NIAMS, NIGMS, NIMHD, NINDS, NLM

Other Federal Agencies:

DARPA, DOE, FDA, NIST, NSF



Bridge to Artificial Intelligence

Vision: to propel biomedical and behavioral research forward by setting the stage for widespread use of artificial intelligence (AI) technologies

Goals:

- Use biomedical and behavioral research grand challenges to generate **flagship datasets**
- **Prepare** AI/ML-friendly data
- Prioritize **ethical** best practices
- Promote **diverse perspectives**



DATA

Diverse

FAIR

AI-ready



ETHICS

Accurate

Reliable

Ethically-sourced

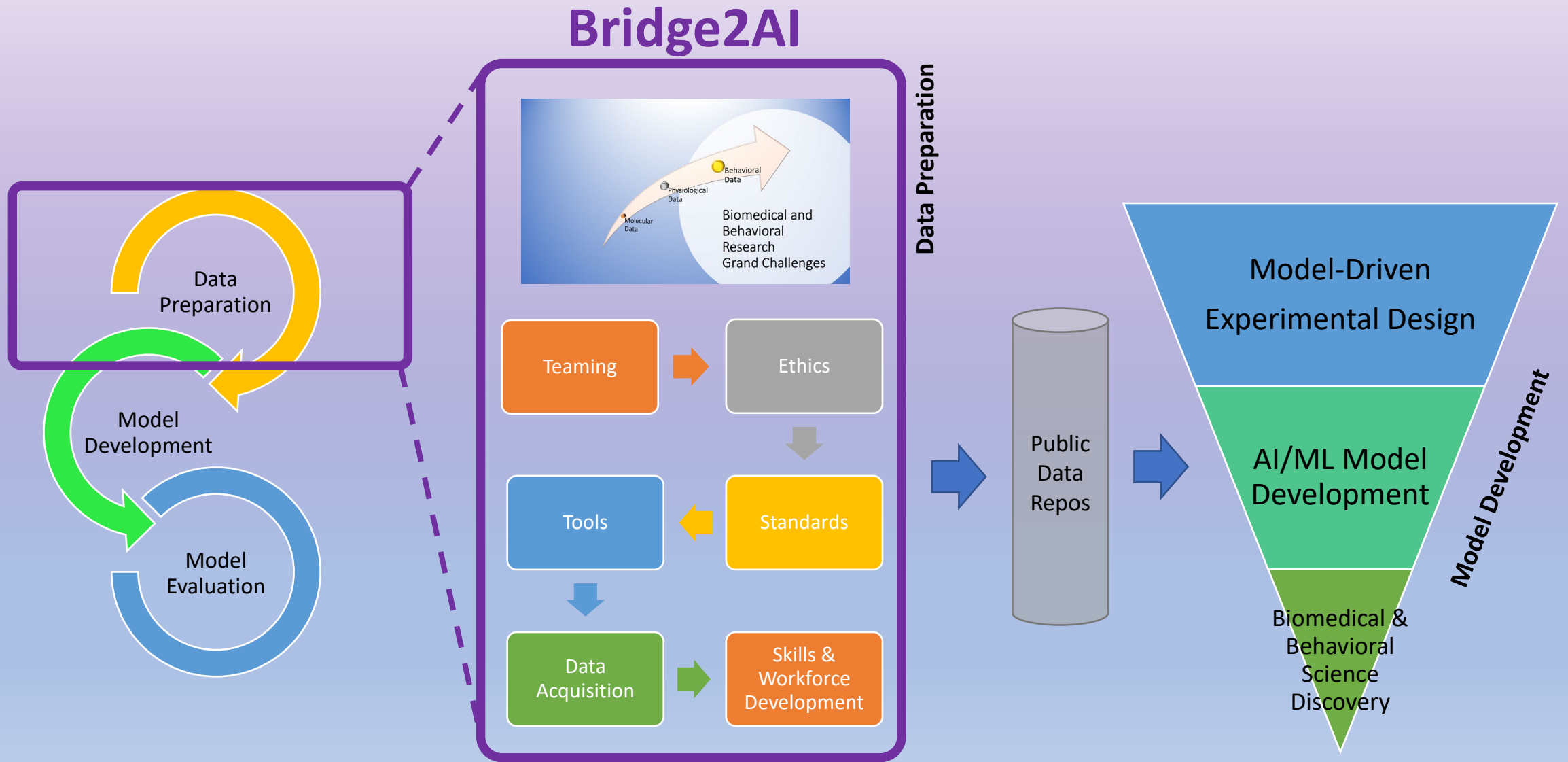


PEOPLE

Diverse teams
&
research cohorts

Training

Scientific Discovery Pipeline



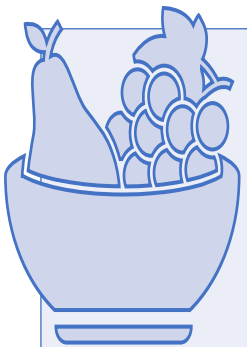
Grand Challenges -- Data Generation Projects



Clinical Care - Using imaging, clinical, and other data collected in an **ICU setting** for diagnosis and risk prediction



Precision Public Health - Using **voice as a biomarker** for human health, revealing how genomic variation, human development, behavioral, and environmental factors affect individual and population health

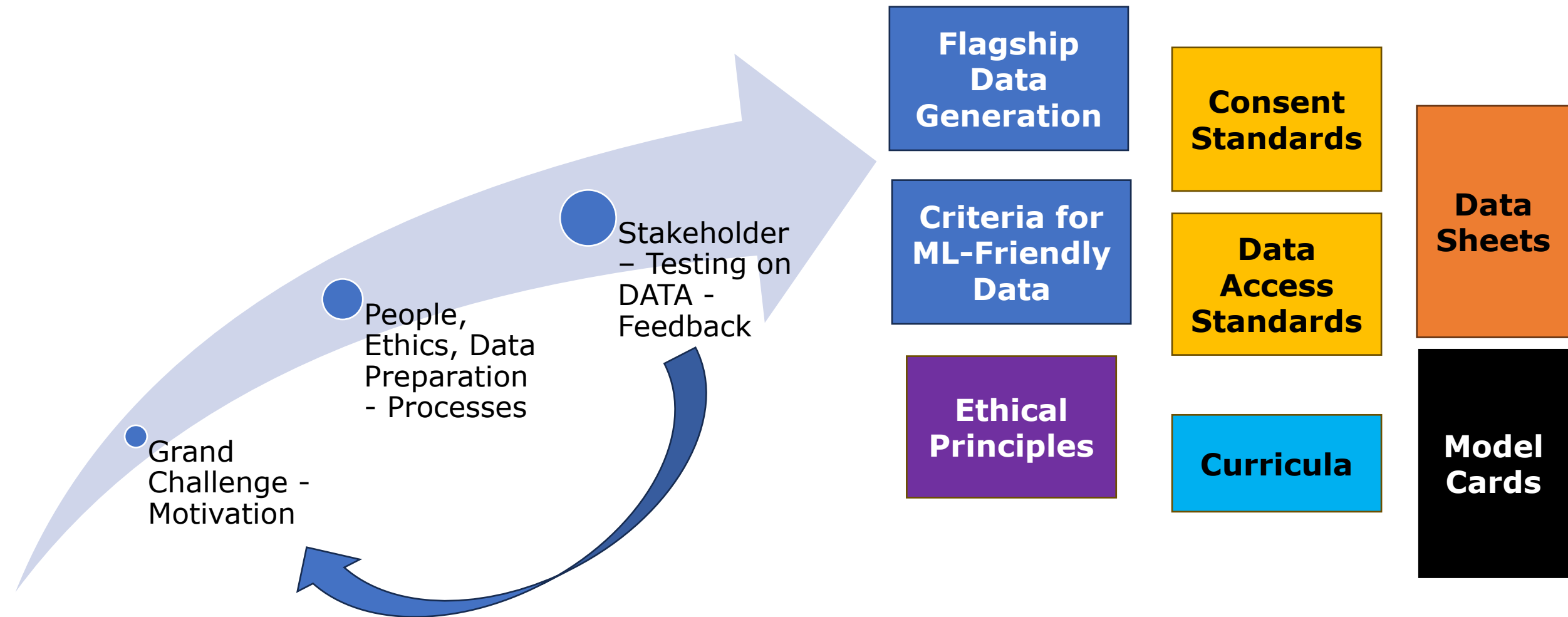


Salutogenesis (Return to Health) - Uncovering the details of how human health is restored after disease, using **type 2 diabetes** as a model

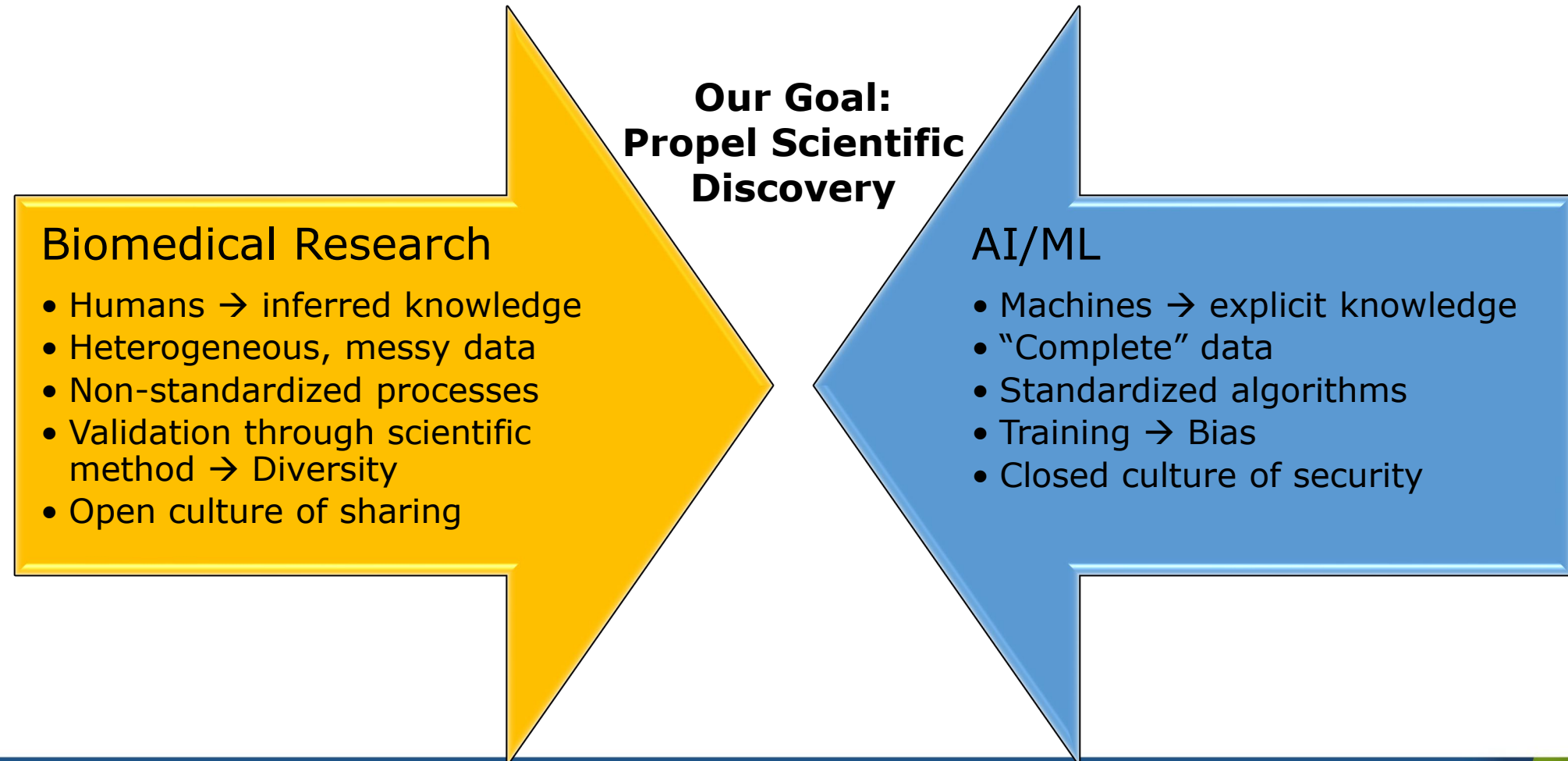


Functional Genomics - Mapping spatiotemporal architecture of human cells to interpret cell structure/function in health and disease

From Vision to Deliverables



What make Bridge2AI challenging?



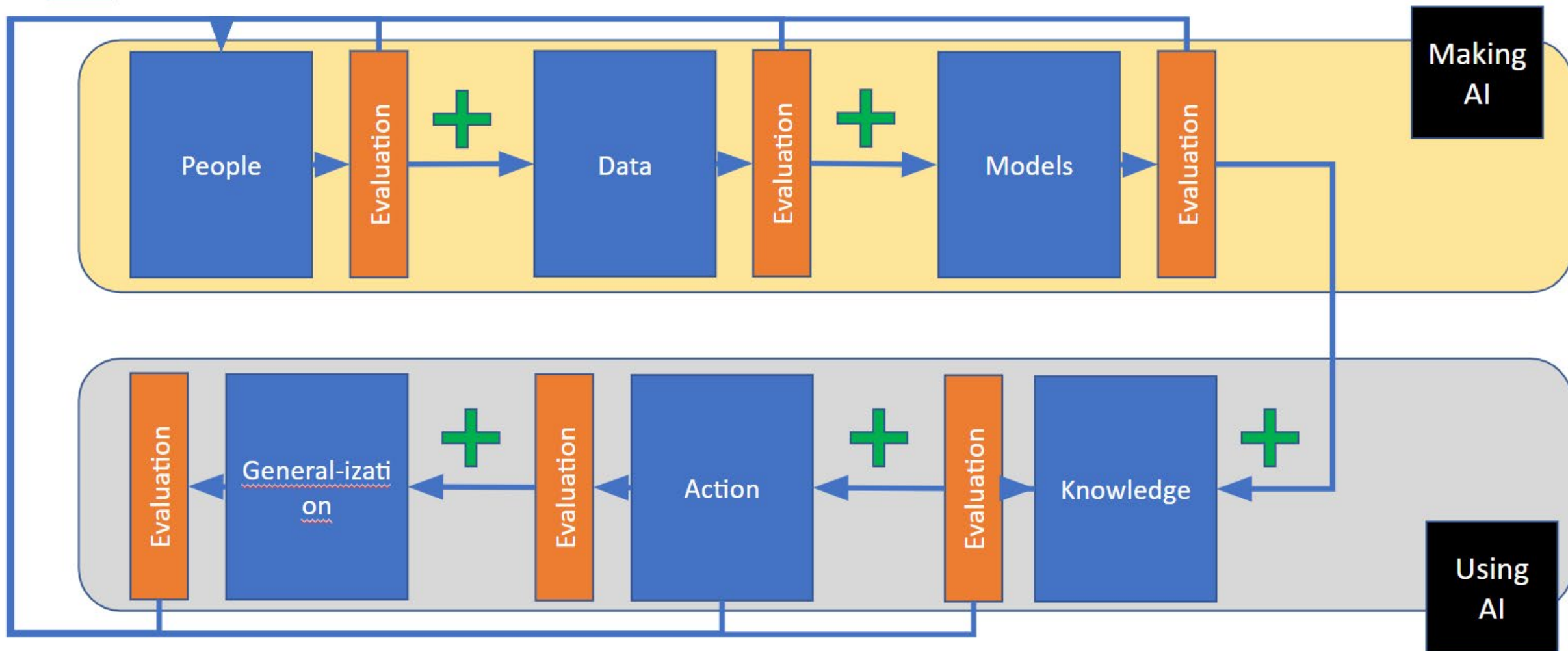
Ethical Challenges → for Open Science

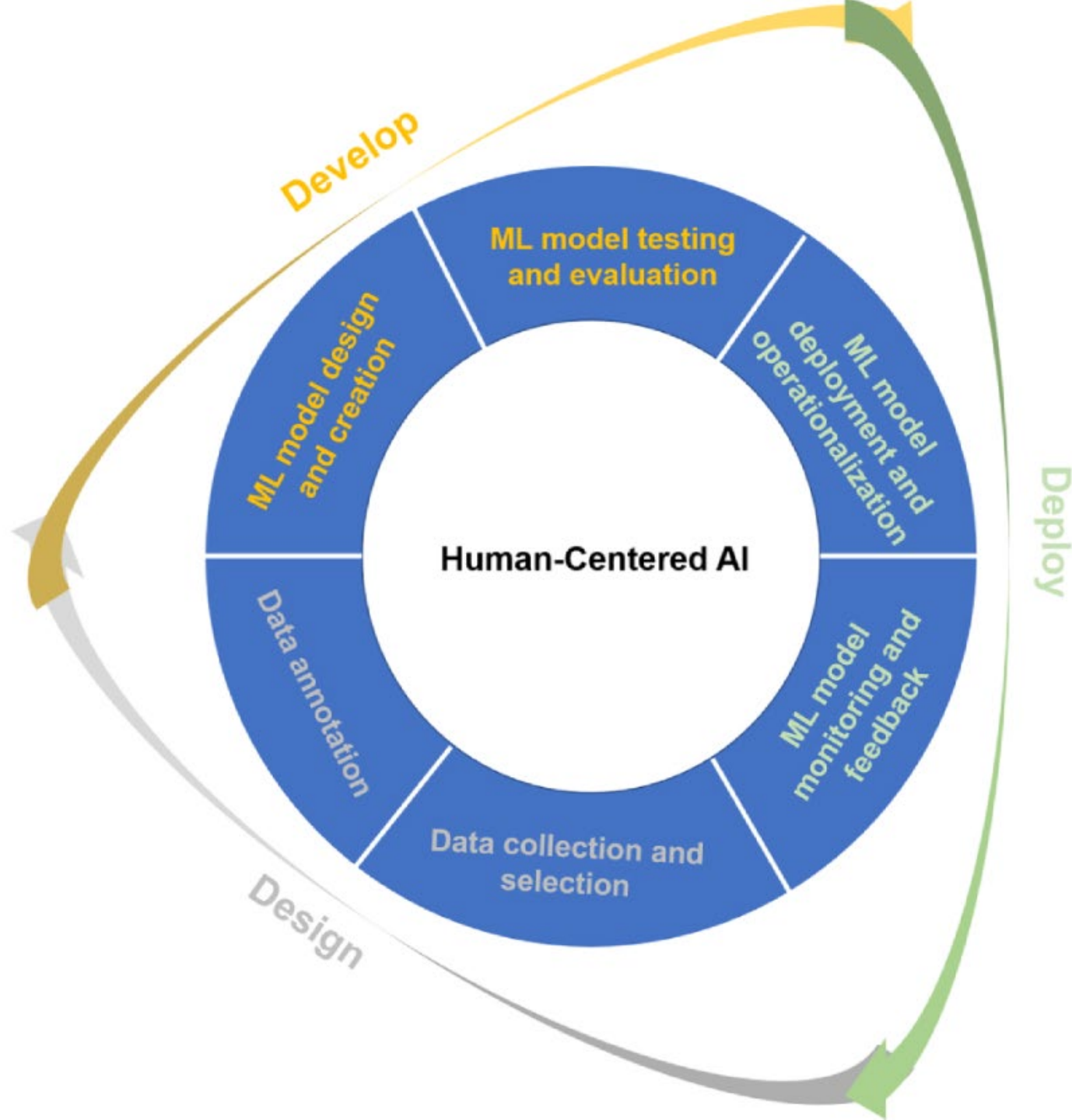
- **Biases:** Issues related to inherent biases of the data
- **Informed Consent:** Going beyond a legal consent form
 - How do we ensure consent given the evolving landscape of AI/ML?
- **Re-identification:** Navigating the risk of re-identification with multi-modal data
- **Unauthorized Use:** How do we prevent unauthorized secondary use?

Bridge2AI

Generating ethically sourced data and best practices

Instilling a culture of ethical inquiry





Chen, Clayton, Novak, Anders, Malin. Human-Centered Design to Address Biases in Artificial Intelligence. JMIR. 2022.

PRECISION
PUBLIC
HEALTH

BRIDGE2AI



Bridge2AI Voice

Cloud environment
Microsoft Azure

SALUTOGENESIS



Cloud environment
Microsoft Azure

CLINICAL
CARE



Cloud environment
Microsoft Azure

FUNCTIONAL
GENOMICS



CM4AI

Cell Maps for AI

Cloud environment
Google Cloud

| | EHR/CLINICAL | SURVEYS | IMAGING | SENSOR-BASED | OMICS | WAVEFORM |
|---|---|--|---|--|--|---|
| <p>A database of 10,000 diverse bioacoustic waveforms is being established to establish voice biomarkers in mental health, respiratory, neurological, and other areas.</p> | <ul style="list-style-type: none"> Demographics Diagnosis (ICD) Severity of disease Treatment information Social history (smoking, alcohol) | <ul style="list-style-type: none"> 12 validated questionnaires (e.g., MOCA, GAD-7, VHI-10, PANAS, DI, etc.) | <ul style="list-style-type: none"> Brain MRI/CTs Chest/neck CTs Laryngoscopy | | <ul style="list-style-type: none"> Whole genome sequencing | <ul style="list-style-type: none"> Bioacoustic data tasks of voice and non-voice sounds, shared as waveforms, Mel spectrograms, features |
| | OMOP | OMOP | Brain imaging: DICOM; laryngoscopy: MP4 | | CRAM & VCFs with metadata | Waveform database (WFDB); creating new standard for bioacoustics |
| <p>Creating a temporal atlas from 3,000 individuals around pathogenesis and salutogenesis to expand applications of AI in clinical care, focusing on Type 2 diabetes</p> | <ul style="list-style-type: none"> Demographics, SDoH Diet Social history Lab tests (blood, urine) Monofilament test Physical assessment Medications Vision testing | <ul style="list-style-type: none"> Multiple validated self-reporting surveys (CES-D, PAID-5, etc.) | <ul style="list-style-type: none"> Retinal imaging (undilated/dilated fundus photography, pupillary dilation, FLIO, optical coherence tomography (OCT), OCT angiography) | <ul style="list-style-type: none"> Continuous glucose monitoring (CGM) Physical activity monitoring (heart rate, steps, sleep phases) Environmental sensors (air quality and particulate measures, temperature) | <ul style="list-style-type: none"> Whole genome sequencing | <ul style="list-style-type: none"> Electrocardiogram (ECG) |
| | OMOP, LOINC | OMOP, LOINC | DICOM | CGM, physical activity: open mHealth; Air: Earth Science Data Spec | CRAM & VCFs with metadata | Waveform database (WFDB) |
| <p>Establishing a set of >100,000 patients from 14 ICU sites across the United States to improve recovery from acute illnesses</p> | <ul style="list-style-type: none"> Demographics, SDoH Clinical notes Lab tests Medications Encounters Procedures | | <ul style="list-style-type: none"> All imaging acquired during ICU setting and captured in PACS (MR, CT, US, x-ray) | | | <ul style="list-style-type: none"> Physiological data (ECG; electroencephalogram, EEG) |
| | OMOP, LOINC | | DICOM | | | Waveform database (WFDB) |
| <p>Creating a library of large-scale maps of cellular structure, function, and disease contexts using cell lines. 200 genes/proteins are the subject of coordinated experiments in three modalities</p> | | | <ul style="list-style-type: none"> Immunofluorescence imaging data for cell imaging | | <ul style="list-style-type: none"> Proteomic mass spectrometry CRISPR perturbation scRNA-Seq Datasets Cell maps | |
| | | | Cell imaging: RO-Crate with JPEG 4-channel (red, green blue, yellow) and metadata | | Mass spec: RO-Crate w/TSV & metadata; CRISPR: RO-Crate with h5ad file & metadata; Cell maps: RO-Crate with Cytoscape CX & metadata | |

Towards Best Practices

What type of Data are you collecting?

Identifier under HIPAA

Non-Identifier under HIPAA

High/low Risk of Re-identification under Common Rule

Do you need consent and what should it contain?

Consent exempt

Consent

Assent

Blanket Consent, Opt-in/Opt Out, Menu?

Who will be using your data and how?

Only PIs from Academia

Only academic researchers

Public population

Companies

What kind of regulatory contracts do you need?

Codes of Conducts

DUA

License

What are the risk vs benefits of your data being released?

Technology available

Type of similar data online

Public health context

Urgency

Can technology be used to diminish risk?

Blockchain technology

Watermarking

Others!

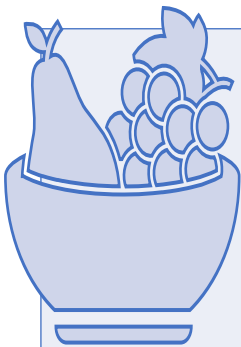
Grand Challenges -- Data Generation Projects



Clinical Care - Using imaging, clinical, and other data collected in an **ICU setting** for diagnosis and risk prediction



Precision Public Health - Using **voice as a biomarker** for human health, revealing how genomic variation, human development, behavioral, and environmental factors affect individual and population health



Salutogenesis (Return to Health) - Uncovering the details of how human health is restored after disease, using **type 2 diabetes** as a model

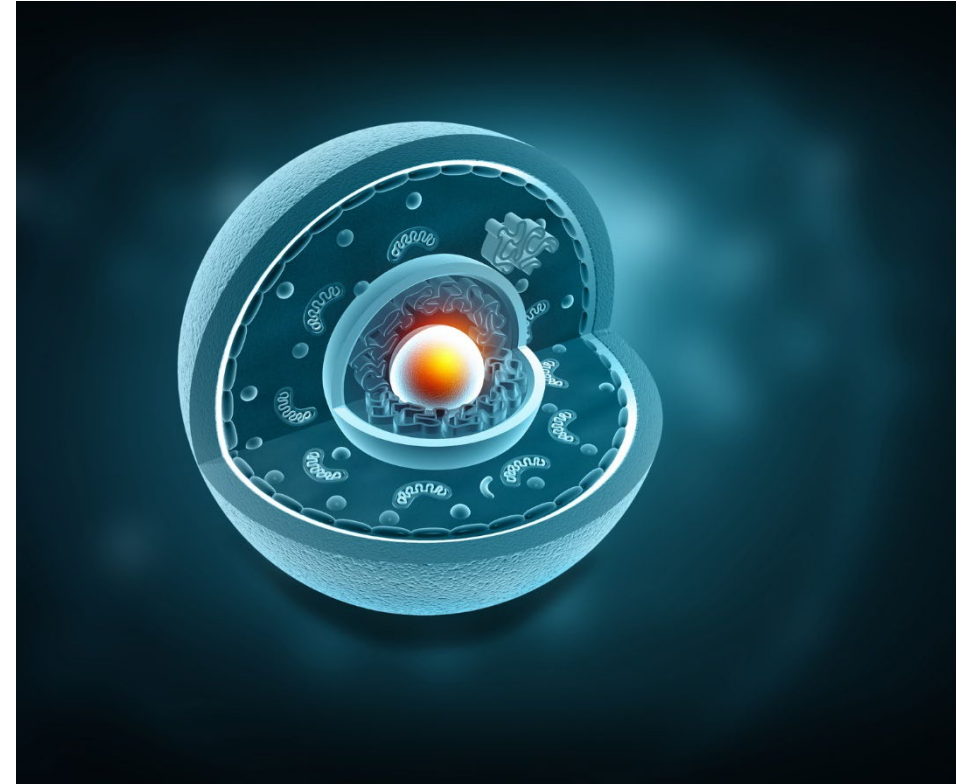


Functional Genomics - Mapping spatiotemporal architecture of human cells to interpret cell structure/function in health and disease

Please welcome Dr. Trey Ideker

Functional Genomics Grand Challenge

- Map the spatiotemporal architecture of human cells and use these maps toward the grand challenge of interpretable genotype-phenotype learning.
- 3 complementary mapping approaches:
 - proteomic mass spectrometry,
 - cellular imaging,
 - genetic perturbation via CRISPR/Cas9
- Create a library of large-scale maps of cellular structure/function and disease contexts using cell lines



Contact PI: *Trey Ideker*, UC San Diego