Stanford MEDICINE

Deep Data Needs and Challenges in Precision Health

Amir Bahmani, PhD

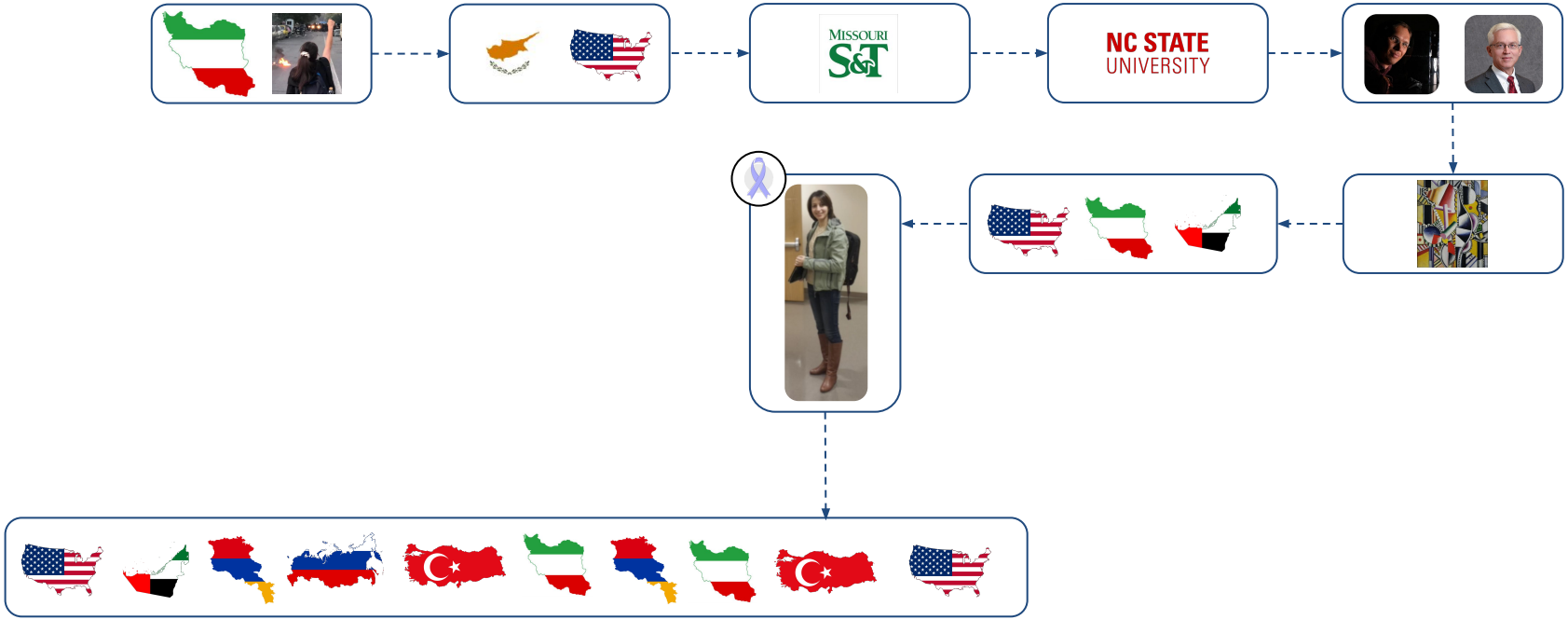
Outline

- **Personal Story 1:** International Students
- **Precision Medicine:** Needs and Challenges
- **Disruptive Solutions:**
  - Deep Data Collection and Analysis at Scale
    - **Real Time Health Framework**
      - Personal Health Dashboard (PHD)
      - Real-time Alert System
- **Personal Story 2:** Collaboration
  - Optimizing Computation and Storage
    - **Hummingbird:** Performance Prediction
    - **Swarm**: Federated Learning
- **Personal Story 3:** Serendipity
  - Unlocking Deep Data for Precision Medicine Applications
    - **Serverless System:** Stanford Data Ocean

Stanford MEDICINE

Timeline

Life of an International Student

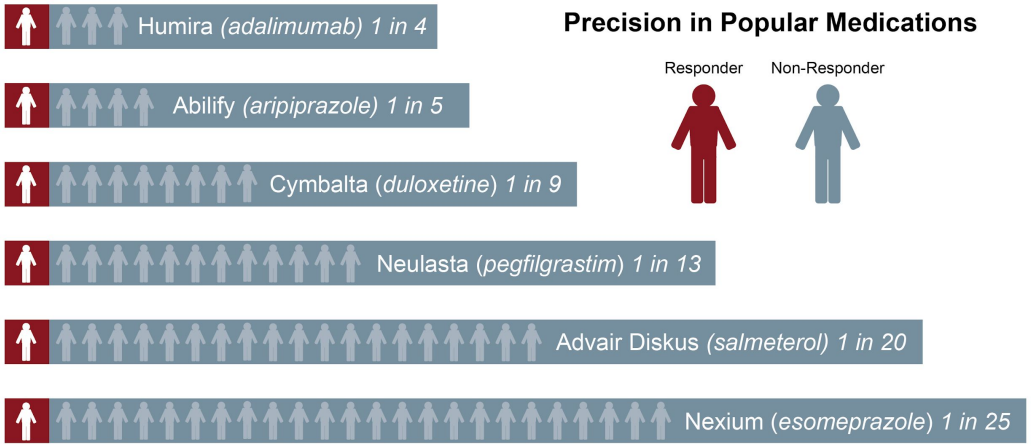


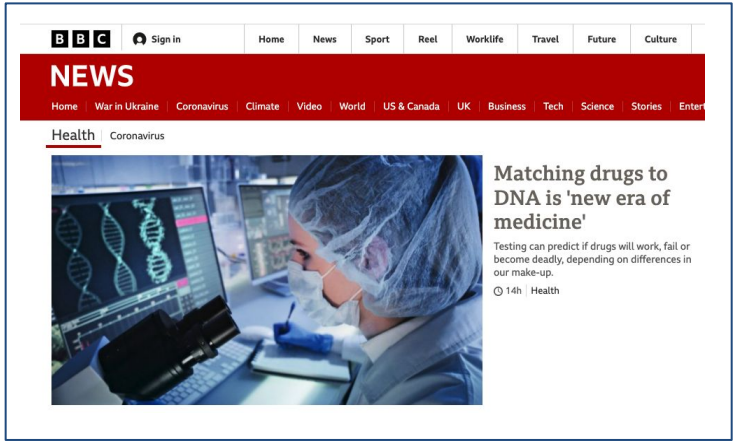
Stanford MEDICINE

From Shallow Medicine to Deep Medicine

"The top ten highest-grossing drugs in the United States help between 1 in 25 and 1 in 4 of the people who take them."

Schork, Nicholas J. "Personalized medicine: time for one-person trials." Nature 520.7549 (2015): 609-611.



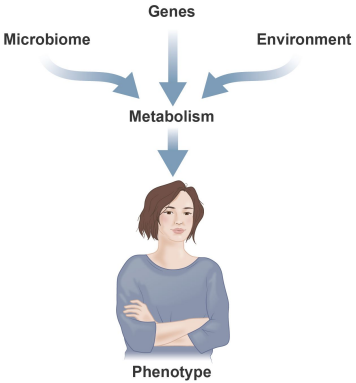


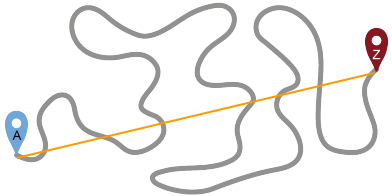
"Some drugs are completely ineffective or become deadly because of subtle differences in how our bodies function." BBC News, 29 March 2022.

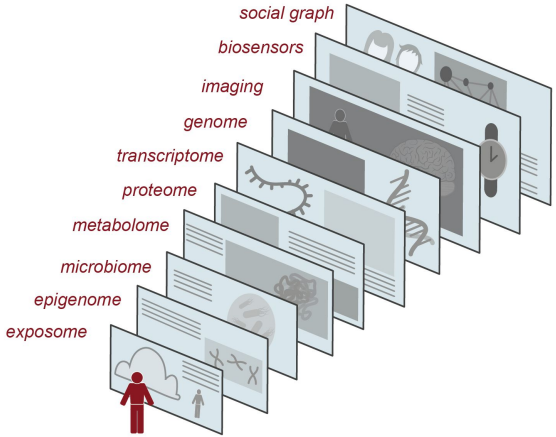
Source: https://www.bbc.com/news/health-60903839

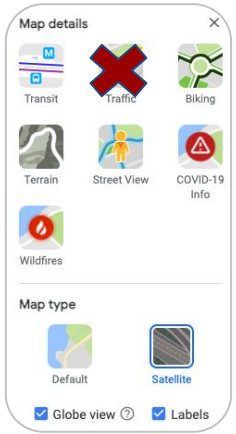
Stanford MEDICINE

From Shallow Medicine to Deep Medicine (cont'd)









"It is more important to know what sort of person has a disease than to know what sort of disease a person has"

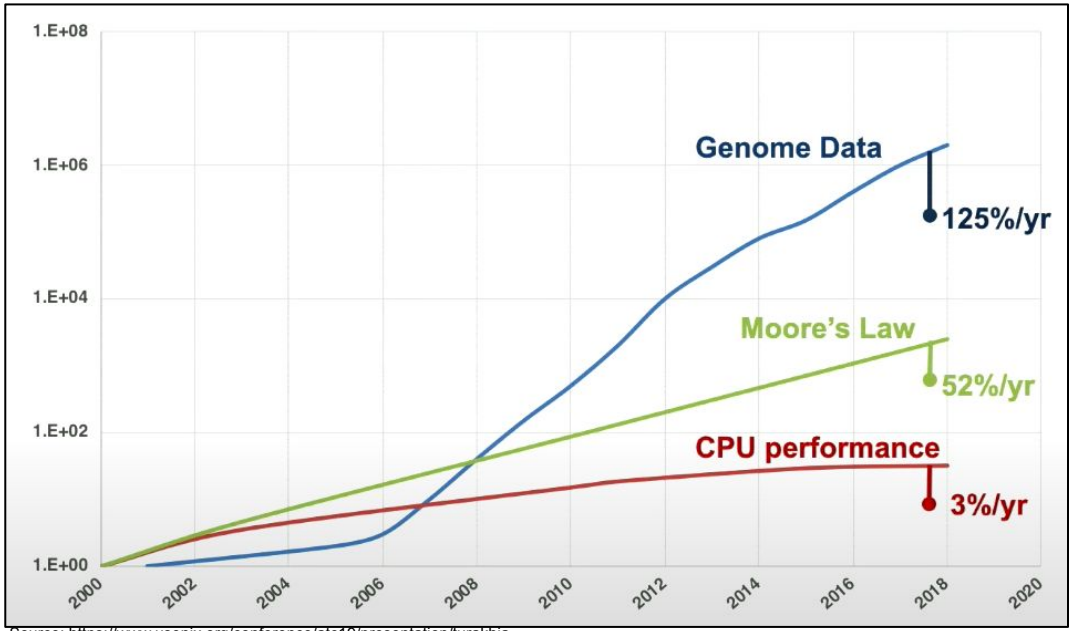
~ Hippocrates, a Greek physician, 2500 years ago

Stanford MEDICINE

A Tsunami of Data is Approaching the Healthcare Industry

At Stanford Center for Genomics and Personalized Medicine (SCGPM), we have collected over 2 PB of Data around Dr. Michael Snyder

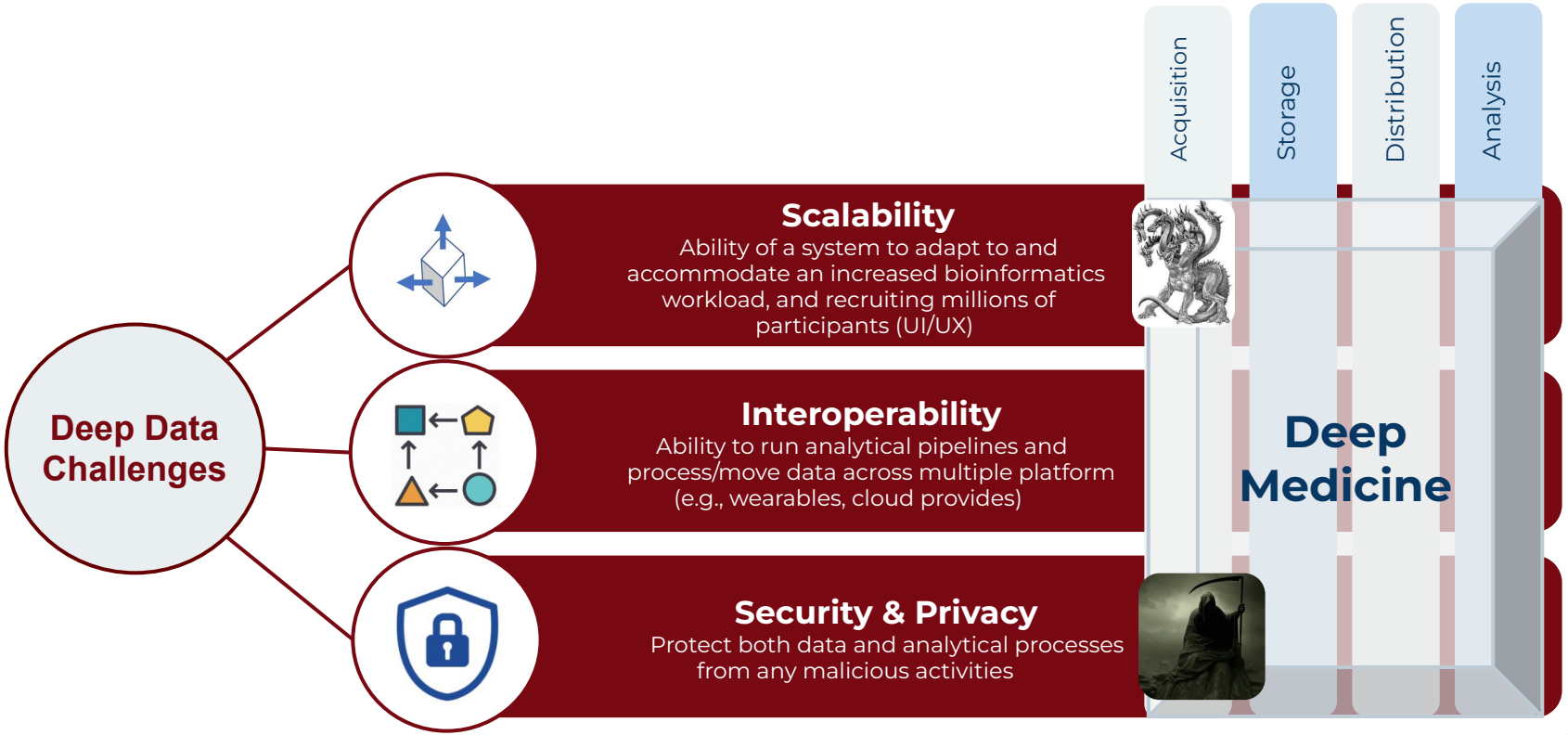
Cost of Standard Storage Per Month for **1PB** (**pay-as-you-go**) as of March 2022	
GCP	$20000 US West1
Azure	$19100 US West
AWS	$21000 US West (Oregon)



Source: https://www.usenix.org/conference/atc19/presentation/turakhia

Stanford MEDICINE

Deep Data Needs and Challenges

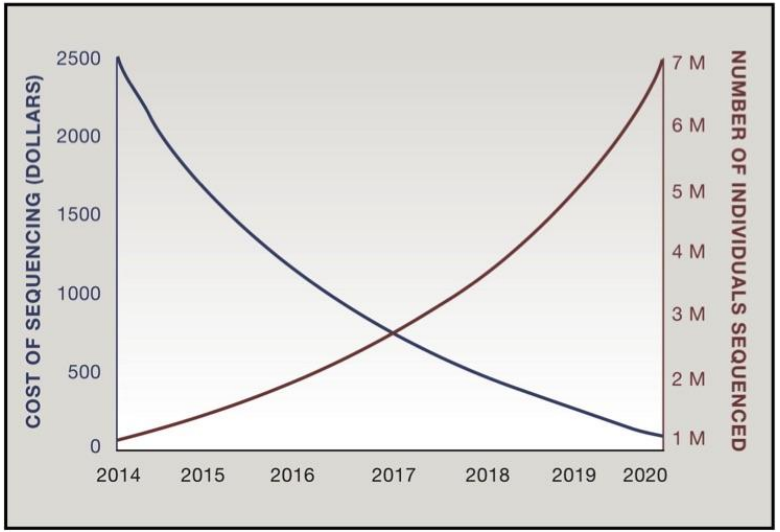


Stanford MEDICINE

The Gap

How can I test my scalable system with a real-world application?

Computer Scientists/Engineers



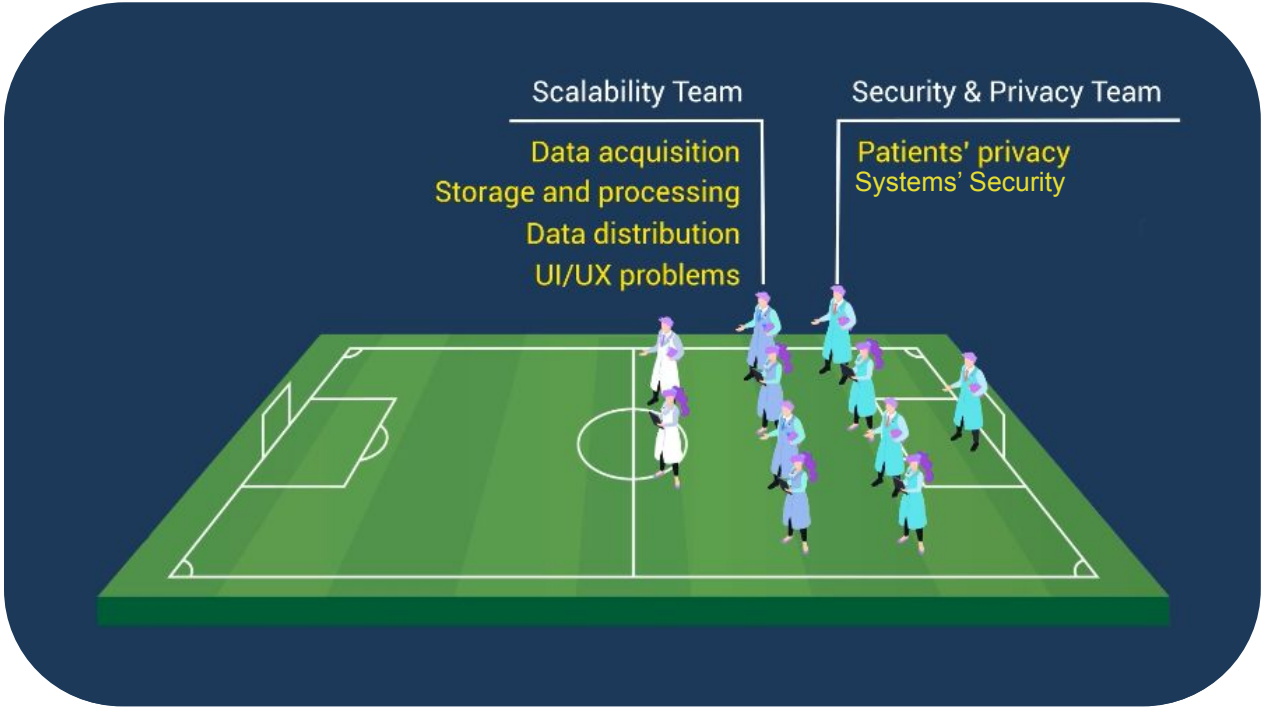
Topol, Eric J. "Individualized medicine from prewombto tomb." *Cell* 157.1 (2014): 241-253.

How am I going to scale my gene model?

Biologists/Physicians

Stanford MEDICINE

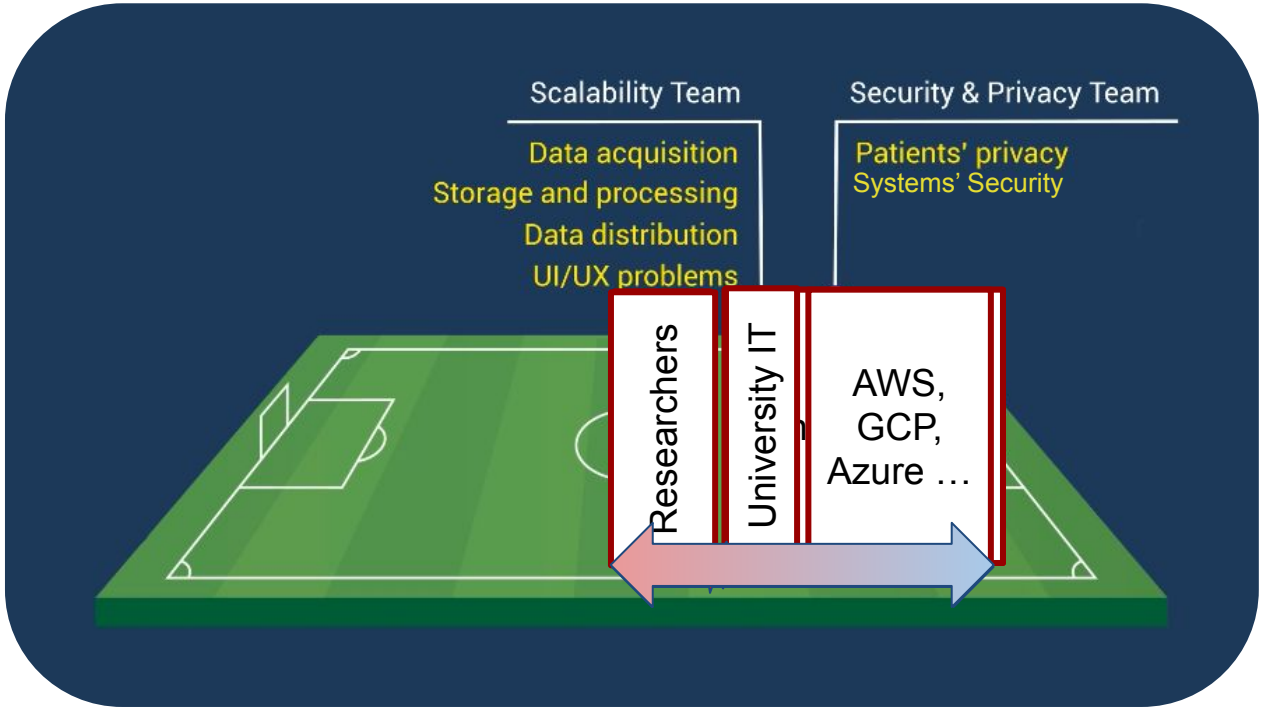
Problem #1: Team Formation in Schools of Medicine

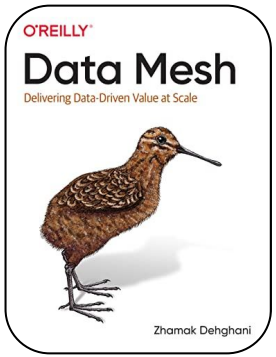


Stanford MEDICINE

Problem #2: Global Governance

Maintain Dynamic Equilibrium b/w Domain Autonomy and Global Governance





Stanford MEDICINE

Outline

- **Personal Story 1:** International Students
- **Precision Medicine:** Needs and Challenges
- **Disruptive Solutions:**
  - Deep Data Collection and Analysis at Scale
    - **Real Time Health Framework**
      - Personal Health Dashboard (PHD)
      - Real-time Alert System
- **Personal Story 2:** Collaboration
  - Optimizing Computation and Storage
    - **Hummingbird:** Performance Prediction
    - **Swarm**: Federated Learning
- **Personal Story 3:** Serendipity/Mentorship
  - Unlocking Deep Data for Precision Medicine Applications
    - **Serverless System:** Stanford Data Ocean

Stanford MEDICINE

Wearables can track many things

Fitbit Charge

Fitbit Ionic

Apple Watch

Garmin

SensOmics

Dexcom G6

Biostrap

Motiv Ring

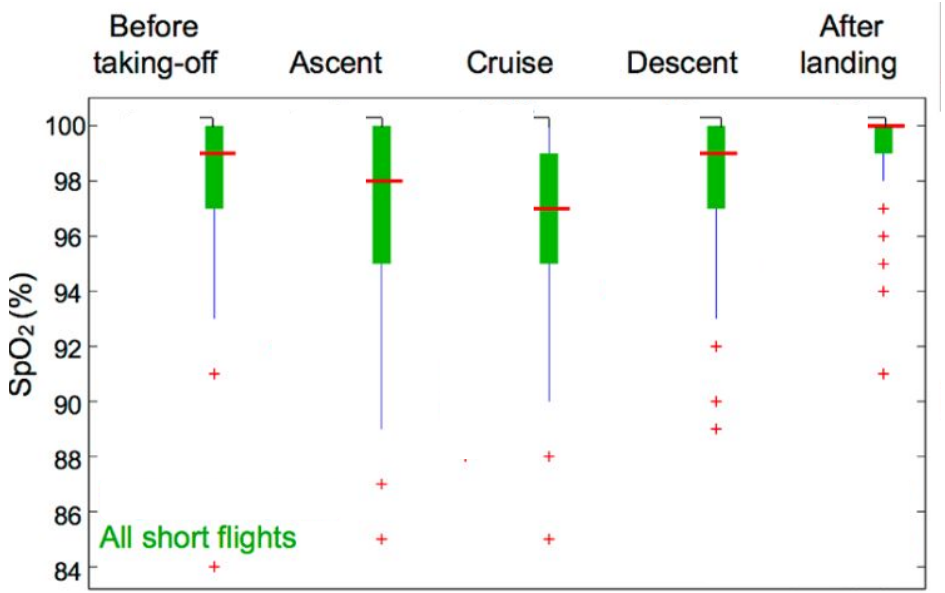
Oura Ring

Empatica E4

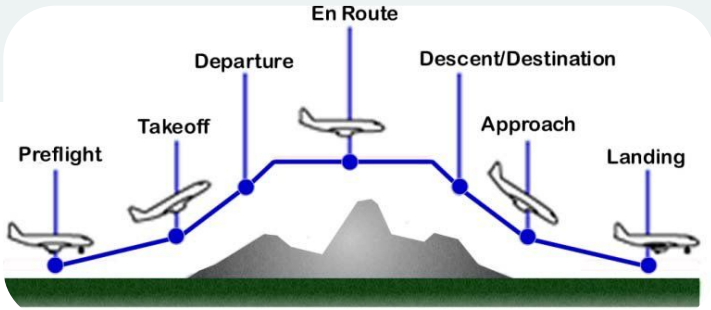
Heart Rate, Heart Rate Variability, Blood Oxygen Level, Sleep, Skin Temperature, Steps, Respiration Rate, etc.

Stanford MEDICINE

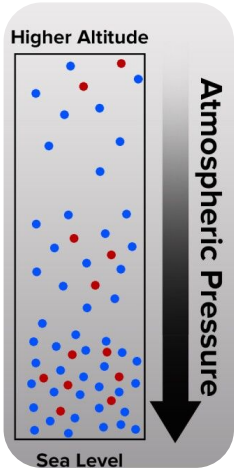
Flight Study: SpO2



* Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information (PLOS BIOLOGY 2017)



Wu, Paul, et al. "Fuzzy multi-objective mission flight planning in unmanned aerial systems." *2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*. IEEE, 2007.



Source:https://www.koaa.com/news/covering-colorado/oxygen-and-altitude-why-do-you-get-so-winded-when-you-climb

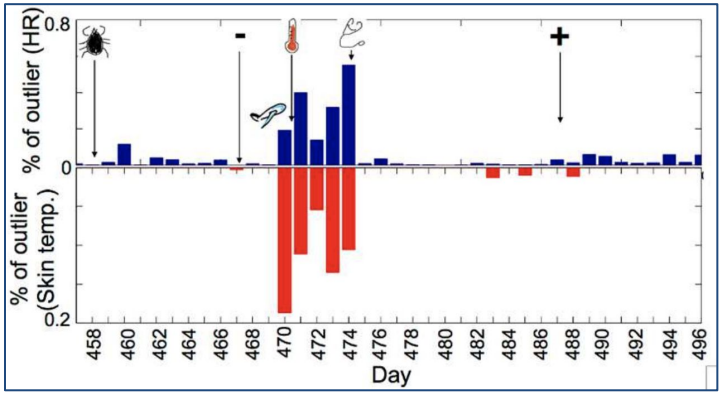
*"The current **FAA** regulations for limiting **cabin pressures** to **8,000-feet** equivalent altitudes allow for mildly hypoxic conditions."*

Source:https://www.faa.gov/data_research/research/med_humanfacs/cer/media/HealthEffectsVulnerablePassengers.pdf

Stanford MEDICINE

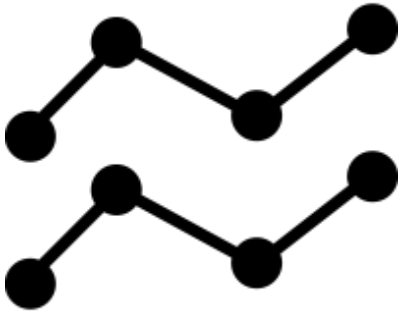
Wearables Data <> Infectious Disease

Lyme Disease



* Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information (PLOS BIOLOGY 2017)

Heart rate / HRV / Temperature / SpO2



Infectious Diseases

2017

Need: Wearable Data Acquisition at Scale

Stanford MEDICINE

Personal Health Dashboard (PHD)

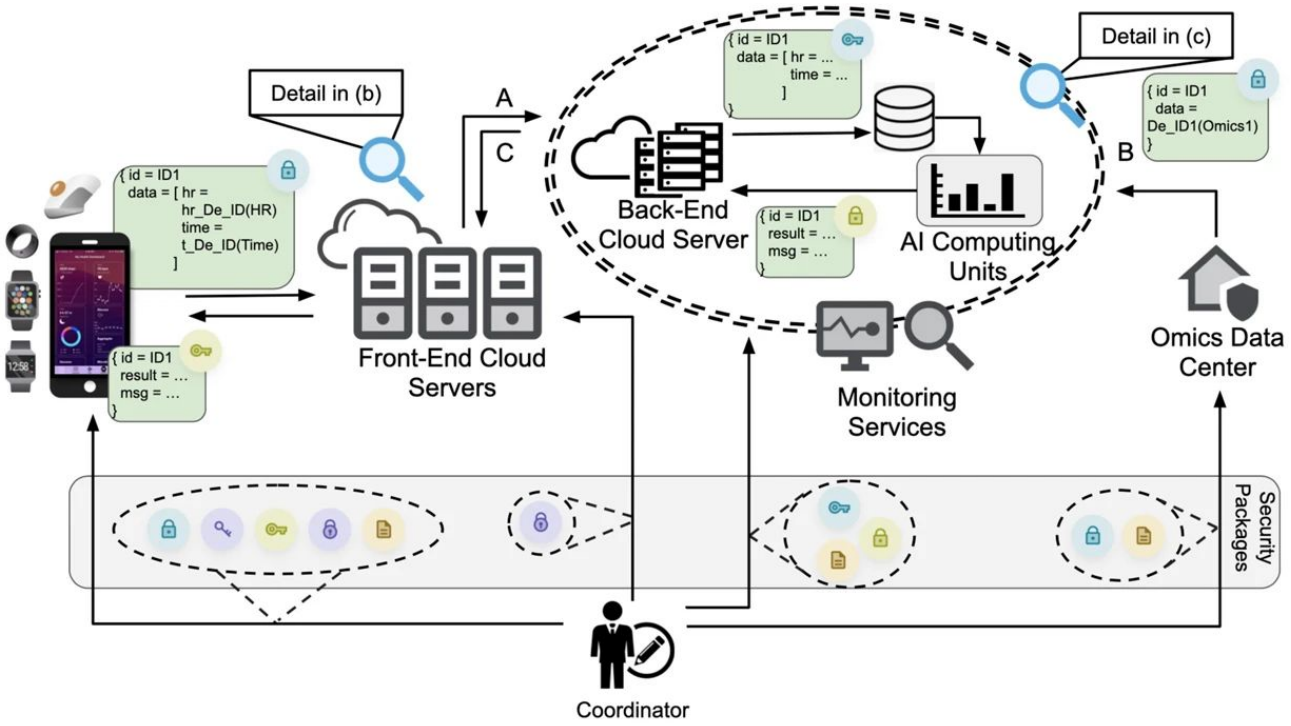
A scalable, secure, and interoperable platform for deep data-driven health management

Amir Bahmani[1,2,3,7], Arash Alavi[1,2,3,7], Thore Buergel[3,7], Sushil Upadhyayula[1,3,4], Qiwen Wang[1,3,4], Srinath Krishna Ananthakrishnan[3], Amir Alavi[3], Diego Celis[1,3,4], Dan Gillespie[3], Gregory Young[1,3], Ziye Xing[1,2], Minh Hoang Huynh Nguyen[1,2], Audrey Haque[1,2], Ankit Mathur[1,3,4], Josh Payne[1,3,4], Ghazal Mazaheri[1,3], Jason Kenichi Li[1,3,4], Pramod Kotipalli[1,3,4], Lisa Liao[1,3,4], Rajat Bhasin[3], Kexin Cha[1,3], Benjamin Rolnik[1,3], Alessandra Celli[1], Orit Dagan-Rosenfeld[1], Emily Higgs[1], Wenyu Zhou[1,2], Camille Lauren Berry[1,3], Katherine Grace Van Winkle[1,3], Kévin Contrepois[1], Utsab Ray[1,2,3], Keith Bettinger[1,2], Somalee Datta[5], Xiao Li[1,6]✉ & Michael P. Snyder[1,2,3]✉

Nature Communications 2021

Stanford MEDICINE

PHD Overview



Stanford MEDICINE

PHD Case Study 1: COVID-19

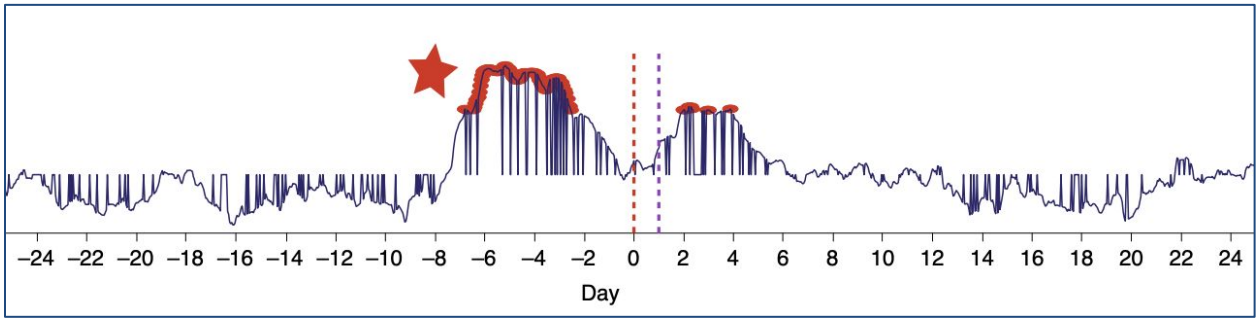
Pre-symptomatic detection of COVID-19 from smartwatch data

Tejaswini Mishra [#1], Meng Wang [#1], Ahmed A Metwally [#1], Gireesh K Bogu [#1], Andrew W Brooks [#1], Amir Bahmani [#1], Arash Alavi [#1], Alessandra Celli [1], Emily Higgs [1], Orit Dagan-Rosenfeld [1], Bethany Fay [1], Susan Kirkpatrick [1], Ryan Kellogg [1], Michelle Gibson [1], Tao Wang [1], Erika M Hunting [1], Petra Mamic [1], Ariel B Ganz [1], Benjamin Rolnik [1], Xiao Li [2], Michael P Snyder [3]

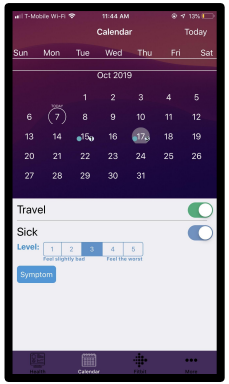
[Nature Biomedical Engineering 2021]

Stanford MEDICINE

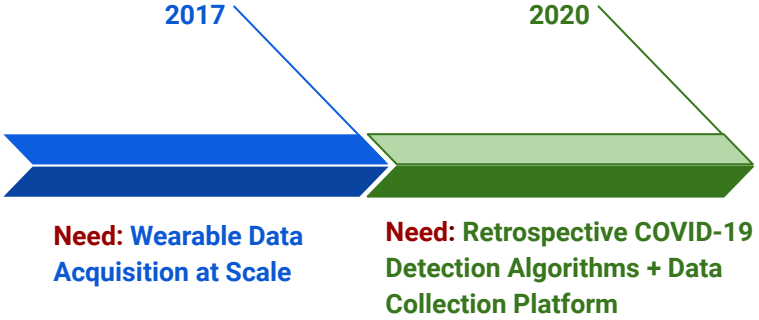
PHD Case Study 1: COVID-19



* Pre-symptomatic detection of COVID-19 from smartwatch data (NATURE BIOMEDICAL ENG 2020)



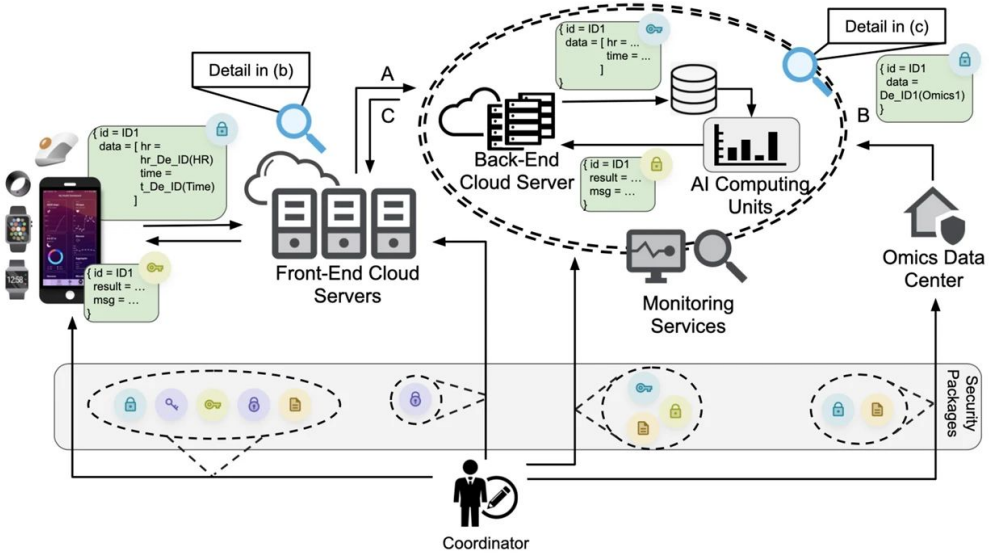
MyPHD Phase 1



- **5** months
- **30+** COVID-19 patients
- **5300+** MyPHD users
- Algorithm successfully detects COVID-19 up to **10** days and a median of **4** days before first reported symptom.

Stanford MEDICINE

PHD: Secure Data Collection and Analysis at Scale



2017

Need: Wearable Data Acquisition at Scale

2020

Need: Retrospective COVID-19 Detection Algorithms + Data Collection Platform

2021

Need: Secure Data Collection and Analysis at Scale

MyPHD

Stanford MEDICINE

PHD COVID-19 Alert System

Real-time alerting system for COVID-19 and other stress events using wearable data

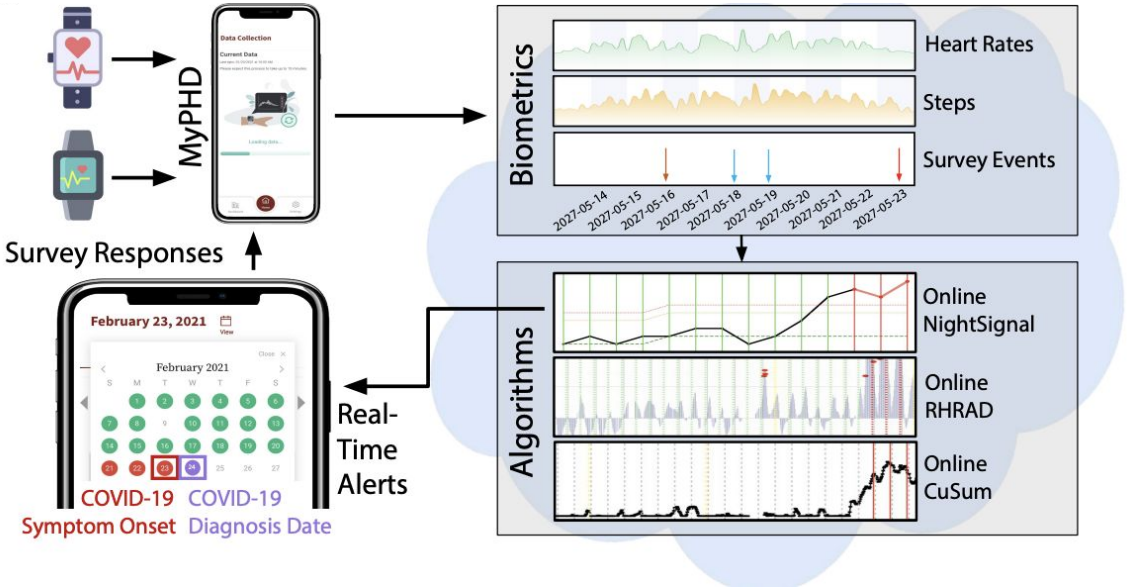
Arash Alavi, Gireesh K. Bogu, Meng Wang, Ekanath Srihari Rangan, Andrew W. Brooks, Qiwen Wang, Emily Higgs, Alessandra Celli, Tejaswini Mishra, Ahmed A. Metwally, Kexin Cha, Peter Knowles, Amir A. Alavi, Rajat Bhasin, Shrinivas Panchamukhi, Diego Celis, Tagore Aditya, Alexander Honkala, Benjamin Rolnik, Erika Hunting, Orit Dagan-Rosenfeld, Arshdeep Chauhan, Jessi W. Li, Caroline Bejikian, Vandhana Krishnan, Lettie McGuire, Xiao Li, Amir Bahmani & Michael P. Snyder

[Nature Medicine 2022]

NightSignal

Stanford MEDICINE

Real-time Infectious Disease Alerting



80%

early detection for pre-symptomatic and asymptomatic COVID-19 cases with a median of 3 days before symptom onset

2017 — **Need:** Wearable Data Acquisition at Scale

2020 — **Need:** Retrospective COVID-19 Detection Algorithms + Data Collection Platform

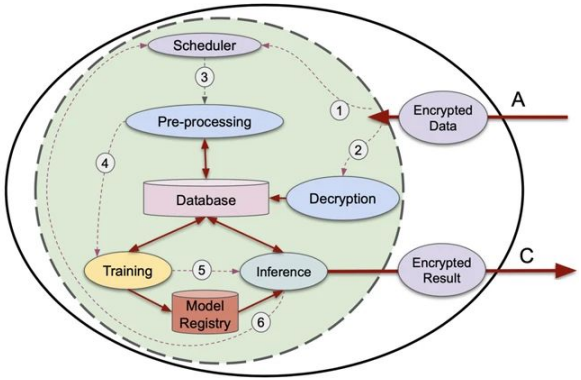
2021 — **Need:** Secure Data Collection and Analysis at Scale MyPHD

2021 — **Need:** First Real-time Alert System

Stanford MEDICINE

COVID-19 Phase 2: Technological Advancements

	NightSignal	Anomaly detection (EllipticEnvelope)	CuSum Statistical Algorithm
Sensitivity	80%	69%	72%
Specificity	87.8%	87.6%	82.1%
Complexity	LightWeight	Heavy	Heavy
Sensitive to resolution	No	Yes	Yes

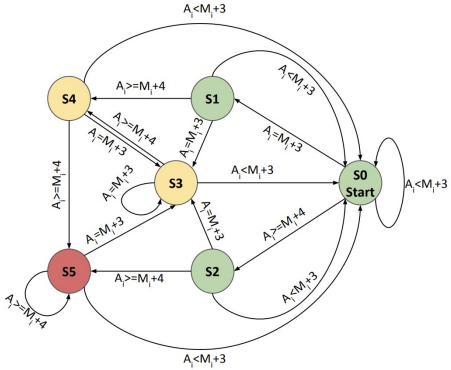


Machine Learning Cluster

it scales any AI/ML algorithms for wearable/multi-omics datasets on the cloud

Lightweight Algorithm

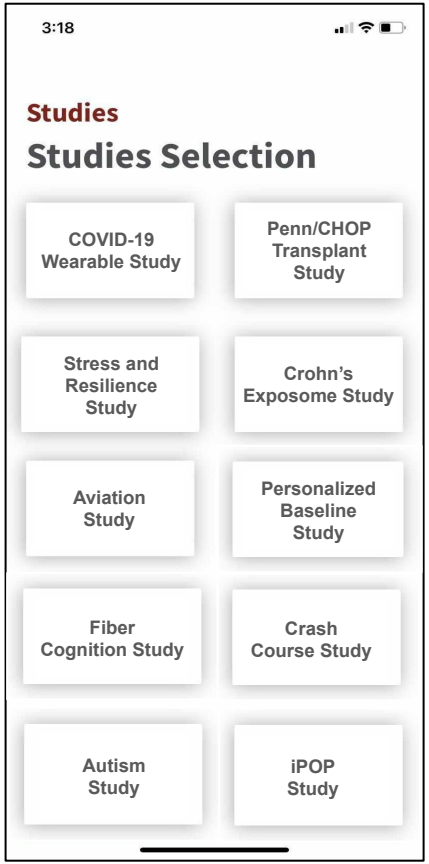
NightSignal can be executed on the phone
Low cost & useful for countries/places w/ slow Internet speeds

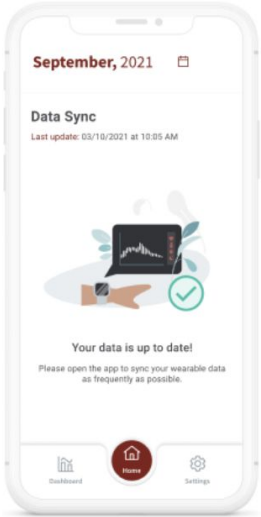


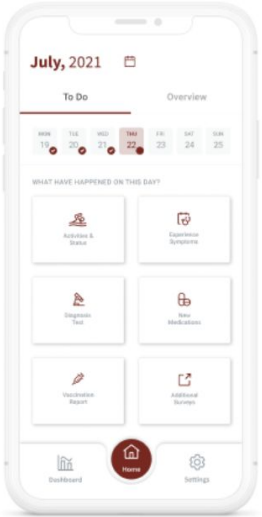
Alert State Machine

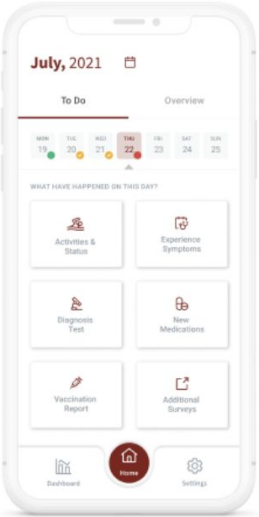
Stanford MEDICINE

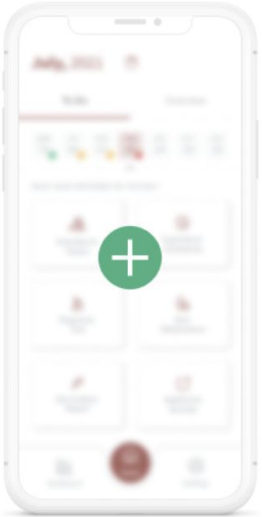
MyPHD Supports Multiple Research Studies











Package 1

- [x] Wearable data collection
- [ ] Survey collection
- [ ] Data analysis
- [ ] Customized requests

Package 2

- [x] Wearable data collection
- [x] Survey collection
- [ ] Data analysis
- [ ] Customized requests

Package 3

- [x] Wearable data collection
- [x] Survey collection
- [x] Data analysis
- [ ] Customized requests

Package 4

- [x] Wearable data collection
- [x] Survey collection
- [x] Data analysis
- [x] Customized requests

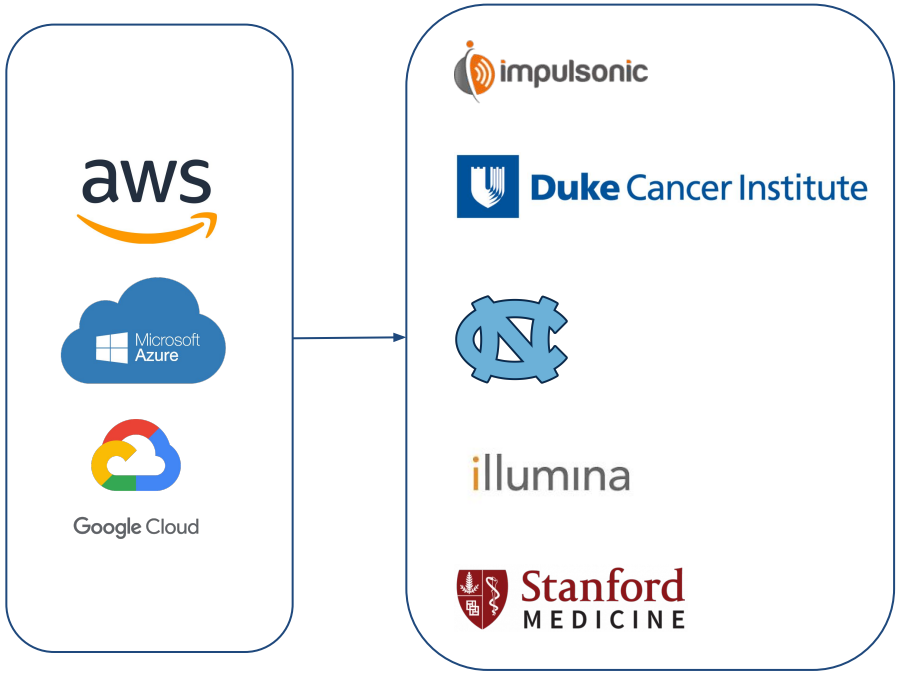
Outline

- **Personal Story 1:** International Students
- **Precision Medicine:** Needs and Challenges
- **Disruptive Solutions:**
  - Deep Data Collection and Analysis at Scale
    - **Real Time Health Framework**
      - Personal Health Dashboard (PHD)
      - Real-time Alert System
- **Personal Story 2:** Collaboration
  - Optimizing Computation and Storage
    - **Hummingbird:** Performance Prediction
    - **Swarm**: Federated Learning
- **Personal Story 3:** Serendipity/Mentorship
  - Unlocking Deep Data for Precision Medicine Applications
    - **Serverless System:** Stanford Data Ocean

Stanford MEDICINE

Timeline
Collaboration

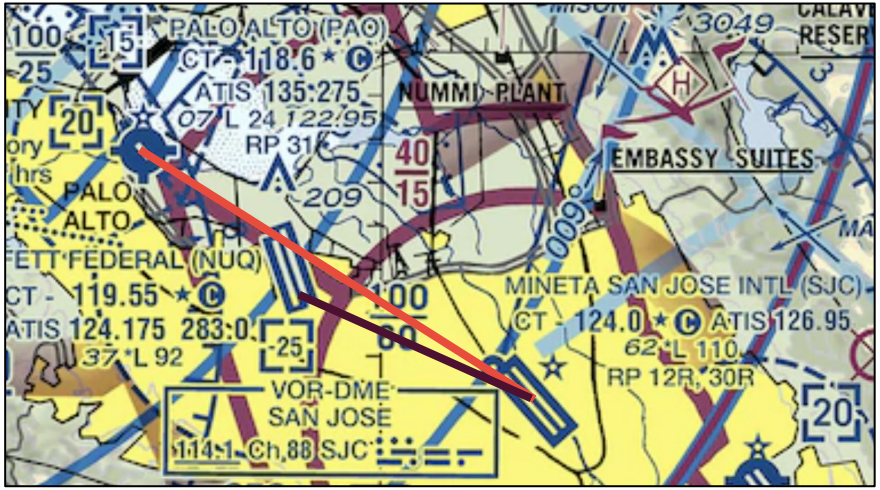
John C. Maxwell:
"**Collaboration is multiplication**."



Stanford MEDICINE

Problem #3: Understand the behaviour of our solution!

Think about scalability at an early stage



$25,000

$250

100X

1. Excessive Cost
2. If your implementation only works for B747, then you are restricted to only special runways

Stanford MEDICINE

Hummingbird: Efficient Performance Prediction

https://www.weforum.org/agenda/2020/11/formula-one-f1-innovation-ventilators-fridges/

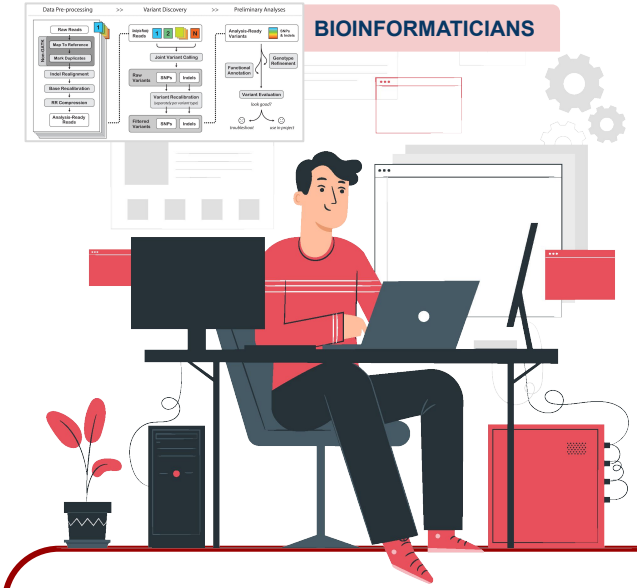
Hummingbird: efficient performance prediction for executing genomic applications in the cloud FREE

Amir Bahmani, Ziye Xing, Vandhana Krishnan, Utsab Ray, Frank Mueller, Amir Alavi, Philip S. Tsao, Michael P. Snyder ✉, Cuiping Pan ✉ Author Notes

[Bioinformatics 2021]

Stanford MEDICINE

Optimizing Computation and Storage



For a computational pipeline, what machine type is the:

1) Cheapest
2) Fastest
3) Cost-efficient

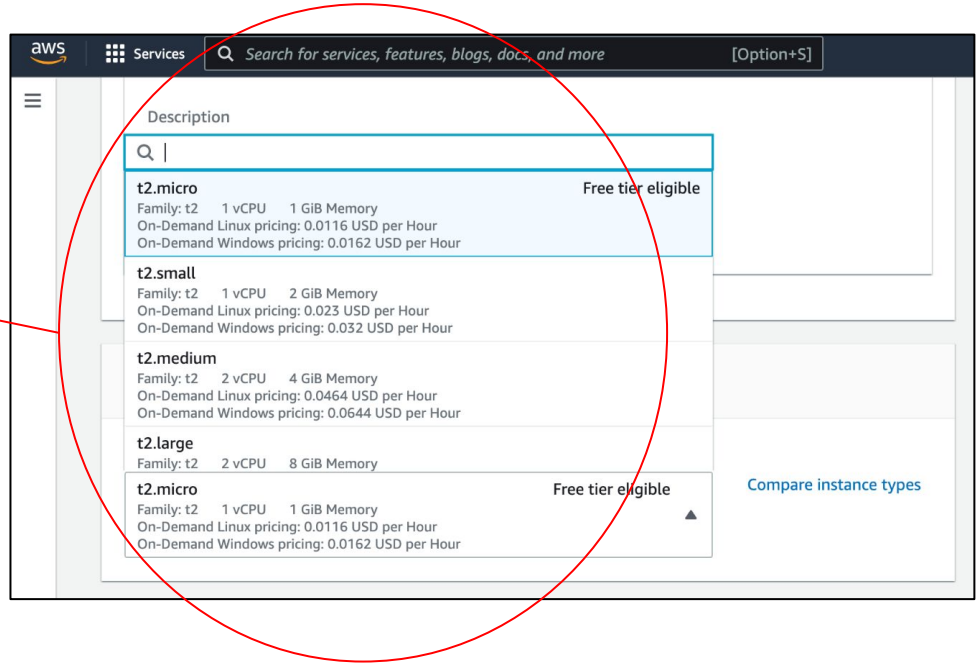
Stanford MEDICINE

Optimizing Computation and Storage (cont'd)

This is the problem we're solving...

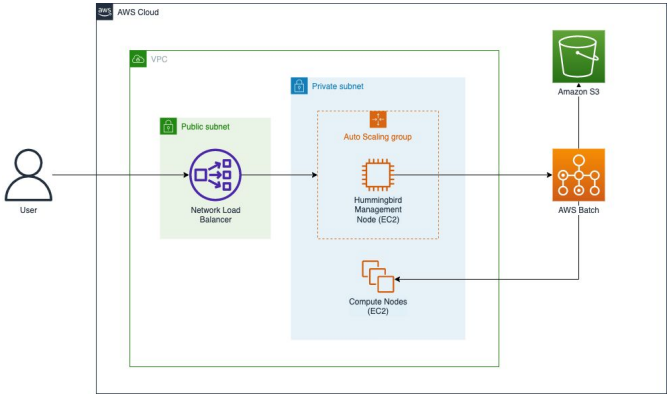
What do I choose?

- Machine Configuration is complex *(CPU, Memory, I/O)*
- Even highly trained informaticians cannot optimize their configurations adequately



Stanford MEDICINE

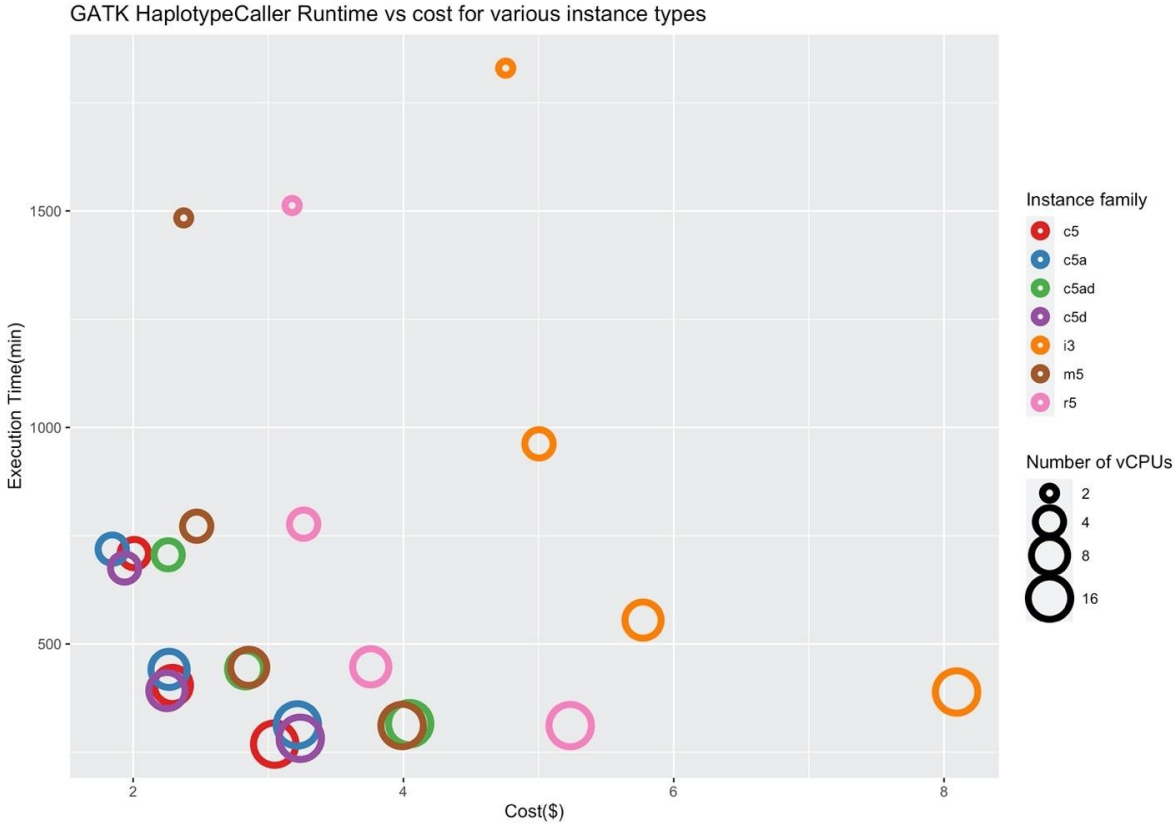
Hummingbird – a tool for effective prediction of performance and costs of genomics workloads on AWS - 2022



Goutam Nistala

Paul Saxman

Sujaya Srinivasan



Stanford MEDICINE

Problem #4:Centralized Databases

Interoperability and Federated Computing

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

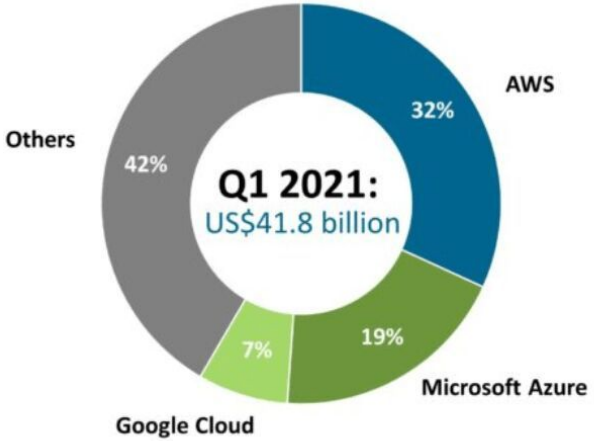
The Illusion of Inclusion — The "All of Us" Research Program and Indigenous Peoples' DNA

Keolu Fox, Ph.D.

Article Figures/Media Metrics

5 References 43 Citing Articles Letters 1 Comment

RAW DATA, INCLUDING DIGITAL SEQUENCE INFORMATION DERIVED FROM HUMAN genomes, have in recent years emerged as a top global commodity. This shift is so new that experts are still evaluating what such information is worth in a global market. In 2018, the direct-to-consumer genetic-testing company 23andMe sold access to its database containing digital sequence information from approximately 5 million people to GlaxoSmithKline for $300 million. Earlier this year, 23andMe partnered with Almirall, a Spanish drug company that is using the information to develop a new antiinflammatory drug for autoimmune disorders. This move marks the first time that 23andMe has signed a deal to license a drug for development.



Source: Public cloud market share statistics in 2022 by Enterprise Engineering Solutions, Inc. (EES)

Stanford MEDICINE

Swarm: A Federated Cloud Framework

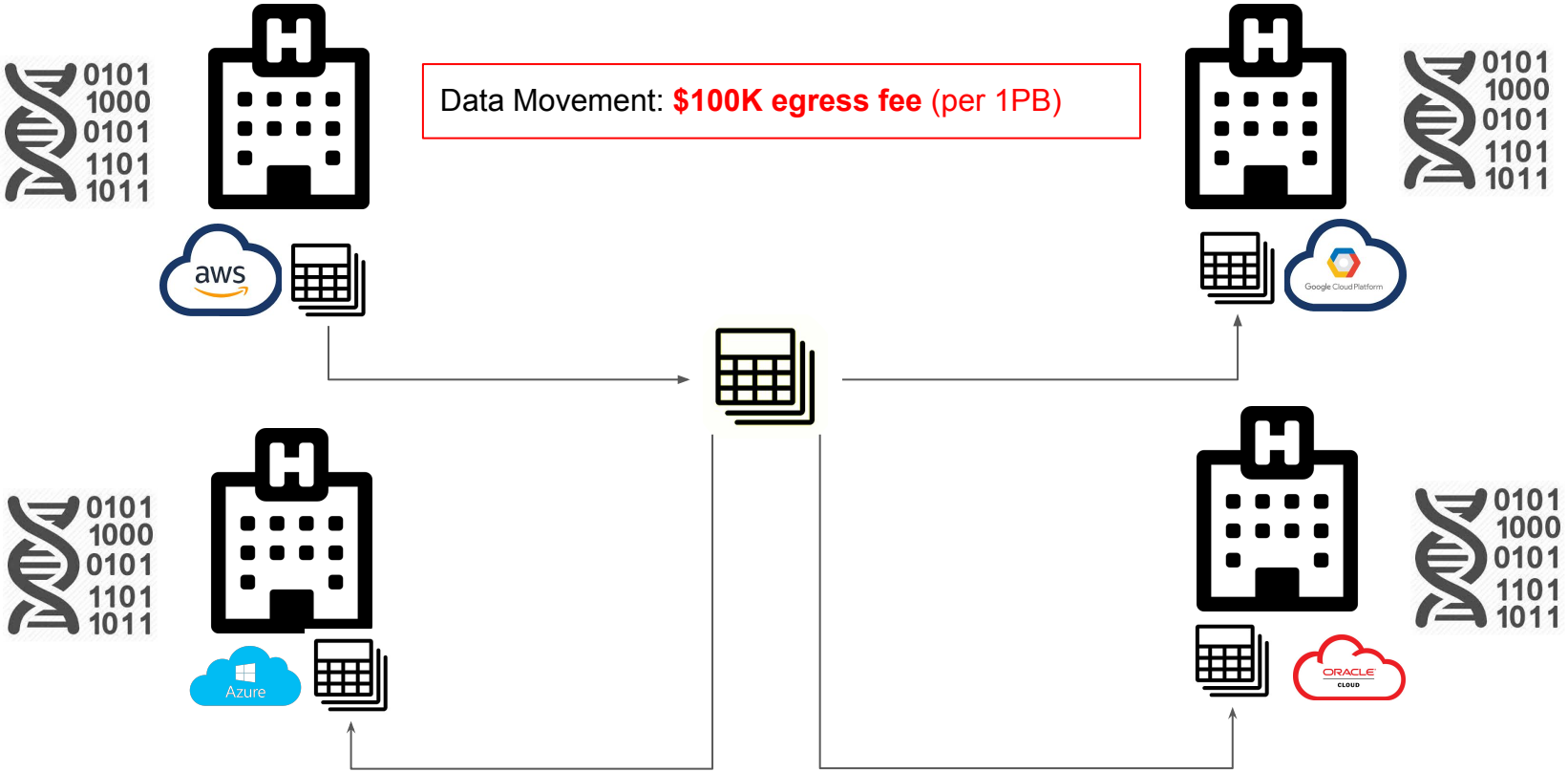
RESEARCH ARTICLE

Swarm: A federated cloud framework for large-scale variant analysis

Amir Bahmani[1,2,3☯], Kyle Ferriter[2,3☯], Vandhana Krishnan[2,3], Arash Alavi[2,3], Amir Alavi[2,3], Philip S. Tsao[4,5], Michael P. Snyder[1,2,3*], Cuiping Pan[5*]

[PLOS Computational Biology 2021]

Stanford MEDICINE

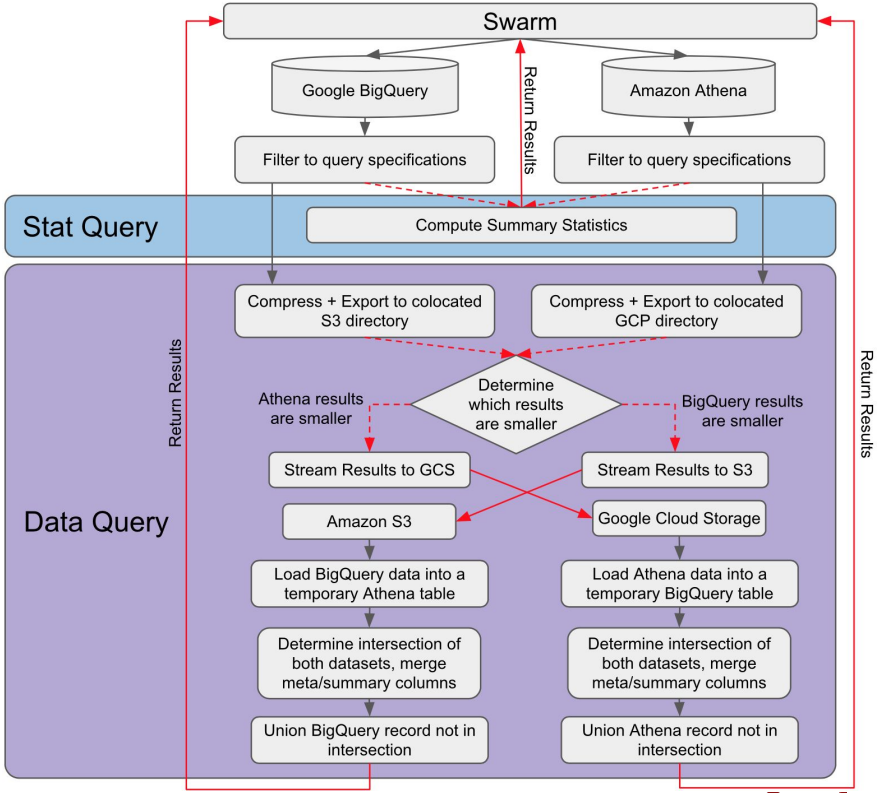


Stanford MEDICINE

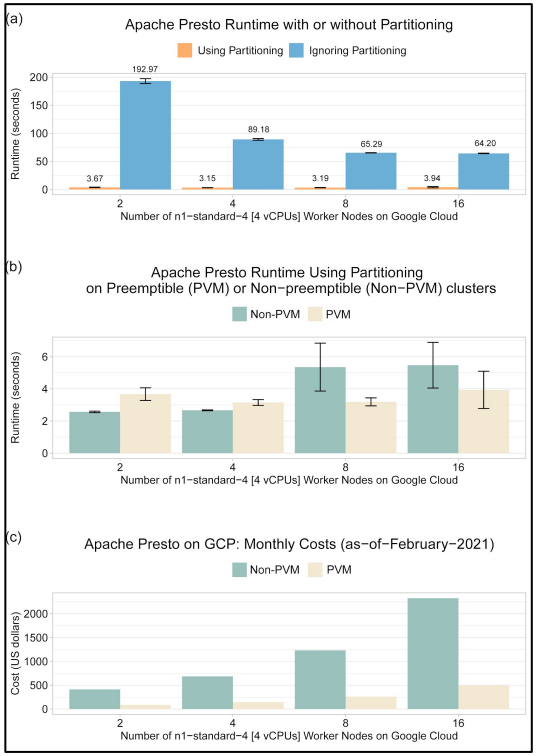
Swarm Overview

- Minimal data motion
- Reduction of costs, delays, and security and privacy risks.
- Applications:
  - Searching for particular genes or variants
  - Annotating large VCF tables
  - Allele frequency discrepancies between subject populations

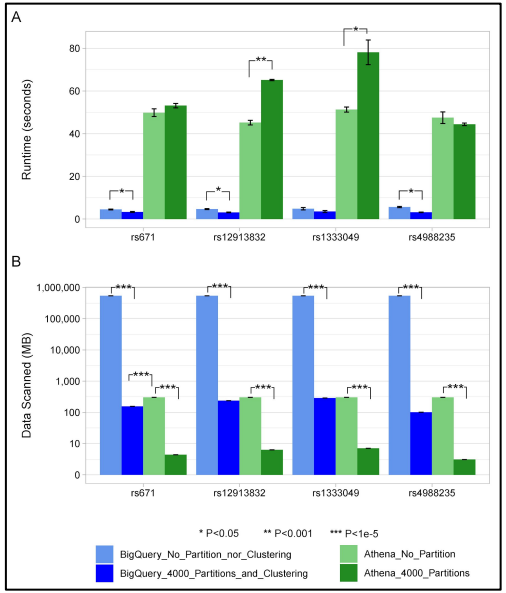
Swarm: Federated computation promoting **minimal data motion** and facilitates crosstalk between genomic datasets stored on various cloud platforms



Stanford MEDICINE



Average execution time and amount of data processed for computing allele frequency for an input set of rsIDs in Google BigQuery, Amazon Athena, Apache Presto



Description	rsID	Chr	Pos
Attention-deficit/hyperactivity disorder (ADHD)	rs671	12	112241766
Blue Eye Color (BEC)	rs12913832	15	28365618
Coronary Heart Disease (CHD)	rs1333049	9	22125503
Lactose Intolerance	rs4988235	2	136608646

- Minimize data motion: **500GB** vs. **4KB**
- Facilitates model training without the need of sharing raw data, and therefore strengthens privacy protection.

Stanford MEDICINE

Outline

- **Personal Story 1:** International Students
- **Precision Medicine:** Needs and Challenges
- **Disruptive Solutions:**
  - Deep Data Collection and Analysis at Scale
    - **Real Time Health Framework**
      - Personal Health Dashboard (PHD)
      - Real-time Alert System
- **Personal Story 2:** Collaboration
  - Optimizing Computation and Storage
    - **Hummingbird:** Performance Prediction
    - **Swarm**: Federated Learning
- **Personal Story 3:** Serendipity/Mentorship
  - Unlocking Deep Data for Precision Medicine Applications
    - **Serverless System:** Stanford Data Ocean

Stanford MEDICINE

Timeline

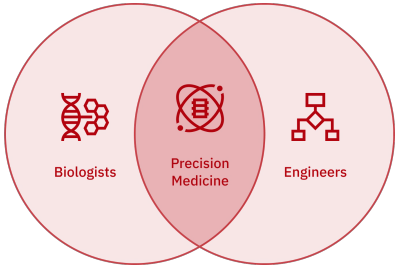
Serendipity: Meeting Late Prof. James L. Morrison

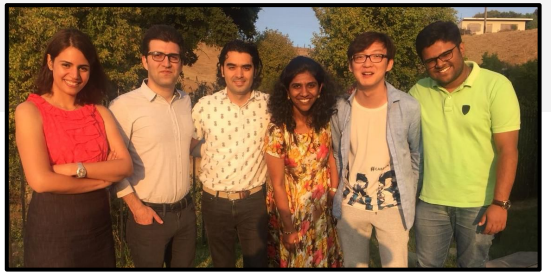
Stanford MEDICINE

Precision Medicine and Education

Stanford Data Ocean (SDO)

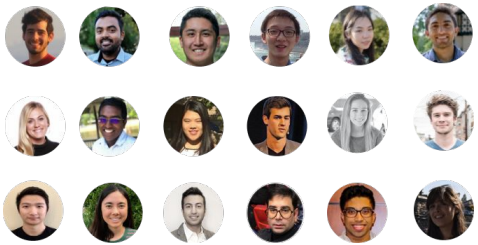
Educating a New Generation of Computer Scientists and Biologists





Summer Internship 2018

Summer Internship 2019



Summer Internship 2020

Stanford MEDICINE

Stanford Data Ocean (SDO)

Steps:

1) Download datasets
   -) Large datasets?
2) Configuring the computing environments
   -) Download the code
   -) Install dependencies
   -) Configure input parameters/environmental variables
   -) Computing capabilities (e.g., GPU)?

Article | Published: 18 November 2020

Pre-symptomatic detection of COVID-19 from smartwatch data

Tejaswini Mishra, Meng Wang, Ahmed A. Metwally, Gireesh K. Bogu, Andrew W. Brooks, Amir Bahmani, Arash Alavi, Alessandra Celli, Emily Higgs, Orit Dagan-Rosenfeld, Bethany Fay, Susan Kirkpatrick, Ryan Kellogg, Michelle Gibson, Tao Wang, Erika M. Hunting, Petra Mamic, Ariel B. Ganz, Benjamin Rolnik, Xiao Li & Michael P. Snyder

Nature Biomedical Engineering **4**, 1208–1220 (2020) | Cite this article

Data availability

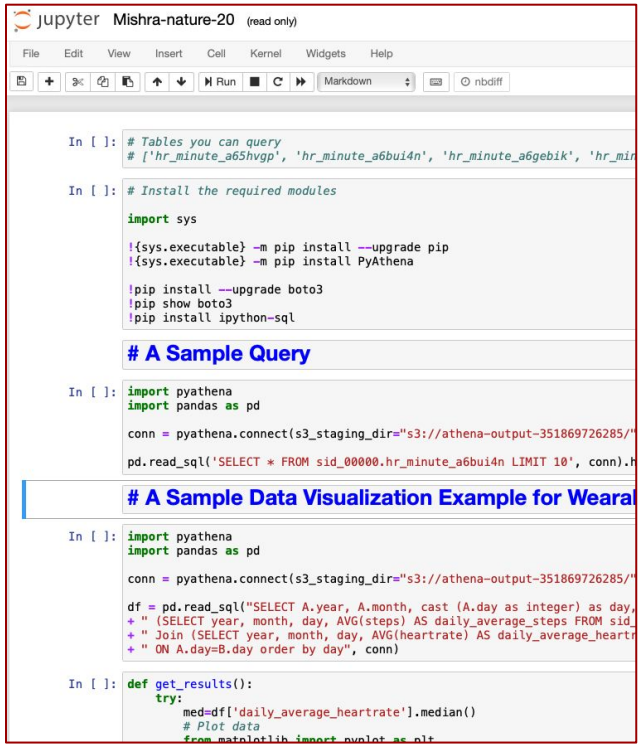
The de-identified raw heart rate, steps and sleep data used in this study can be downloaded from the study data repository (https://storage.googleapis.com/gbsc-gcp-project-ipop_public/COVID-19/COVID-19-Wearables.zip). Processed data, including algorithm outputs and the data used for plotting the figures are provided as Supplementary Data 1.

Code availability

Code for the algorithms used in this manuscript is available at https://github.com/mwgrassgreen/WearableDetection (RHR-Diff and CuSum) and https://github.com/gireeshkbogu/AnomalyDetect (HROS-AD).

Stanford MEDICINE

Stanford Data Ocean (SDO) (cont'd)



1-click

Article | Published: 18 November 2020

Pre-symptomatic detection of COVID-19 from smartwatch data

Tejaswini Mishra, Meng Wang, Ahmed A. Metwally, Gireesh K. Bogu, Andrew W. Brooks, Amir Bahmani, Arash Alavi, Alessandra Celli, Emily Higgs, Orit Dagan-Rosenfeld, Bethany Fay, Susan Kirkpatrick, Ryan Kellogg, Michelle Gibson, Tao Wang, Erika M. Hunting, Petra Mamic, Ariel B. Ganz, Benjamin Rolnik, Xiao Li & Michael P. Snyder

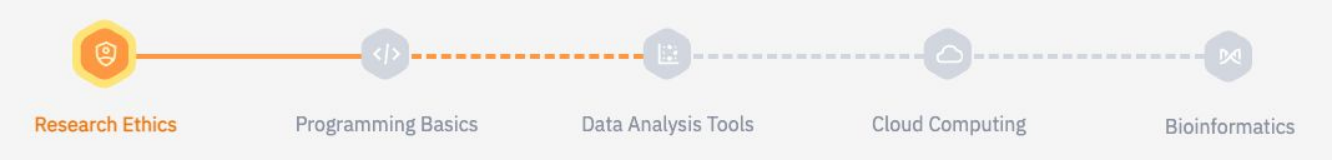
Nature Biomedical Engineering **4**, 1208–1220 (2020) | Cite this article

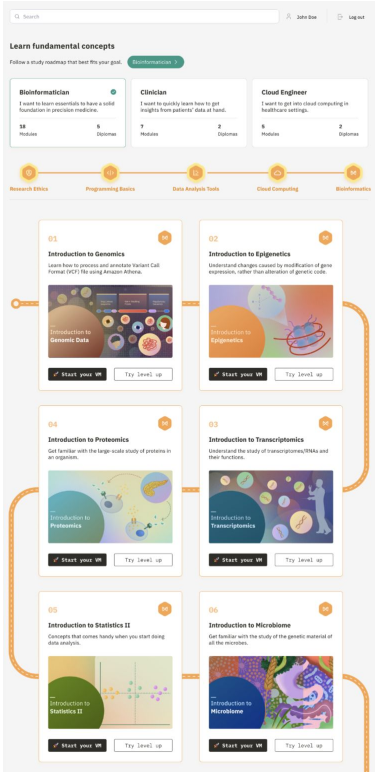
Code and Data Availability

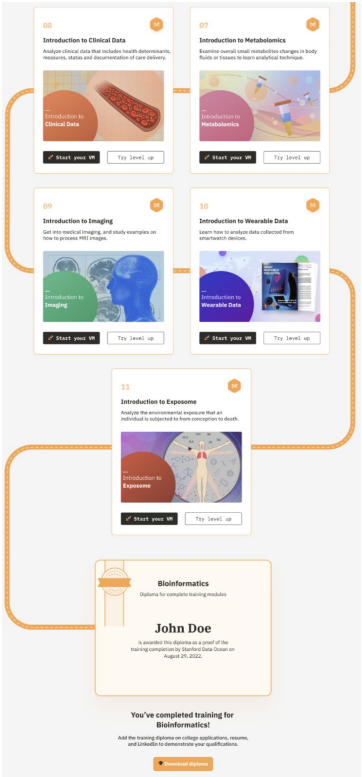
Code and data for the algorithm used in this manuscript are available at (https://dataocean.stanford.edu/wearables/mishra-naturebioen-20).

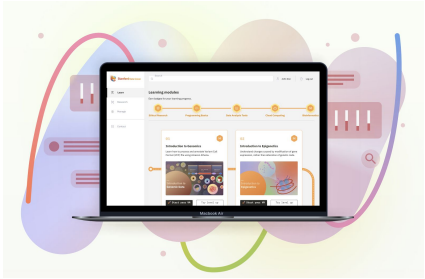
Stanford MEDICINE

Stanford Data Ocean (SDO): Bioinformatics









EARLY ACCESS

Stanford MEDICINE

Internship + Course Impact

Personal Health Dashboard

- Started Working on the Security and Privacy Concerns of MyPHD Apps

Summer 2018

2019

- Finished PhD in Computer Science
- Joined Stanford as a Full-time Employee

2020

- Helped Scaling the COVID-19 Study on MyPHD Apps (Nature Biomedical Engineering)
- Presented at the CyberSecurity Festival

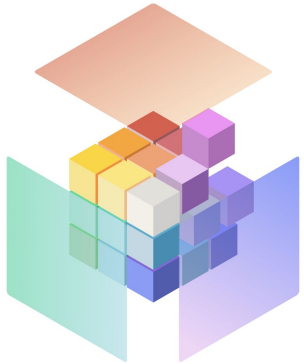
2021

- MyPHD Paper (Nature Communications)
- Real-time Alert System (Nature Medicine)

Stanford MEDICINE

Stanford Deep Data Research Center (deepdata.stanford.edu)

Visit our website and learn about our projects! We are always looking for collaborators!



Stanford MEDICINE

Special Thanks to:

Aaron Horning
Alessandra Celli
Amir Alavi
Amit Dixit
Andrew W Brooks
Andrew Crabb
Ankit Mathur
Ankit Sethia
Antony Ross
Anshul Kundaje
Arash Alavi
Arshdeep Chauhan
Ashok Ganguli
Audrey Haque
Ben Rolnik
Camille Berry
Claire Muscat
Cuping Pan
Dan Gillespie
Diego Celis
Dominic Young
Ekanath Srihari Rangan
Emily Higgs
Gireesh K Bogu
Gregory Young
Goutam Nistala
Hariharan Swaminathan
Jason Li
Josh Payne
Katherine Van Winkle
Keith Bettinger
Ken Harris
Kerre Dubinsky
Kevin Contrepois
Kevin Galens
Kexin Cha
Kouros Owzar
Kyle Ferriter
Lek Tin
Lettie McGuire
Lisa Liao
Mahmoud Parsian
Mark Schreiber
Meng Wang
Mehrzad Samadi
Minh Nguyen
Negin Forozesh
Nicholas Midler
Nick Kittmito
Orit Dagan-Rosenfeld
Paul Billing-Ross
Paul Saxman

Somalee Datta Director of Research IT, Stanford

Michael Snyder, Chair & Professor of Genetics, Stanford

Frank Mueller, Professor of Computer Science, NCSU

Phillip Tsao, Professor of Medicine, Stanford
Associate Chief of Staff for Precision Medicine, VA Palo Alto Health Care System

Late James L. Morrison, Professor of Educational Leadership, UNC

Peter Knowles
Pramod Kotipalli
Jessi W Li
Qiwen Wang
Quentin Hall
Rajat Bhasini
Ramesh Nair
Shrinivas Panchamukhi
Srinath Krishnan
Sujaya Srinivasan
Sushil Upadhyayula
Tadeu Perillo
Tao Wang
Tejaswini Mishra
Utsab Ray
Vandhana Krishnan
Wenyu Zhou
Xiao Li
Ziye Xing

NIH

Google Cloud

aws

Microsoft Azure

FLU LAB

SCHMIDT FUTURES

MVP MILLION VETERAN PROGRAM

NVIDIA

Stanford MEDICINE

Stanford Deep Data Research Center (deepdata.stanford.edu)

Deep Data Research Center

MyPHD Platform

SDO Platform

Swarm

PHD

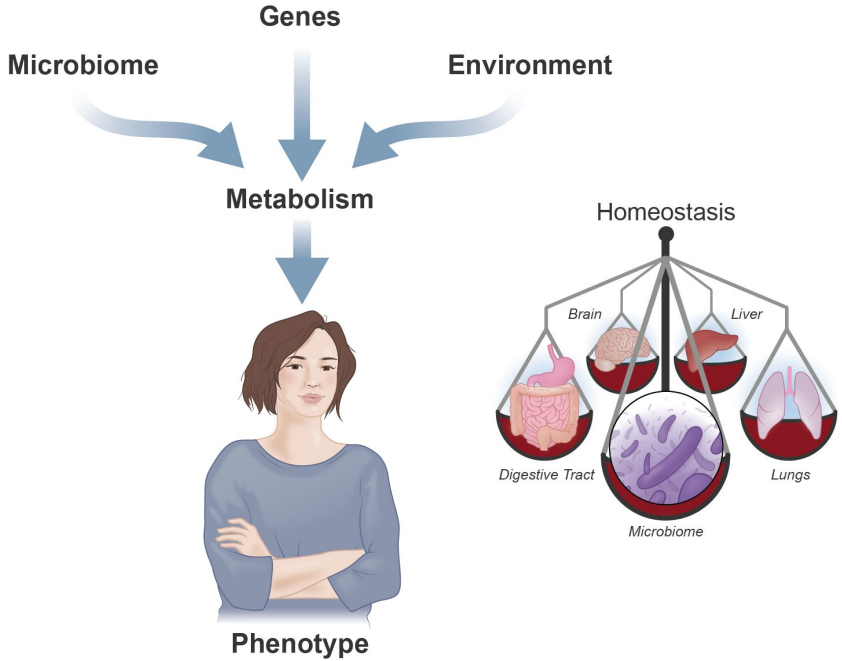
COVID19 Phase 1

Alert COVID-19 Phase 2

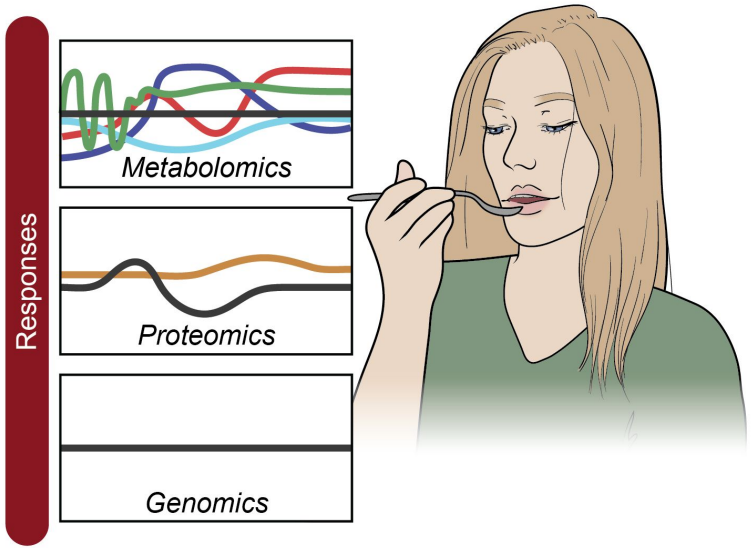
Hummingbird

Stanford MEDICINE

Find the effective resolution+window



Metabolism: Conversion of food/fuel to energy to run cellular processes



Metabolomics is more time sensitive than other “omics”

Stanford MEDICINE

Latency-sensitive vs. Throughput-oriented

Applications can be grouped into two broad classes:

Latency-sensitive workload
(User-facing front-end applications)

A job might require a latency of 100 seconds for uploading daily heart-rate data

Throughput-oriented workload
(Internal batch analytics frameworks)

A batch job might require a throughput of 10000 genomes per day

Stanford MEDICINE