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
Timeline
Life of an International Student
“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.”

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

<table>
<thead>
<tr>
<th>Platform</th>
<th>Cost (USD per month)</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCP</td>
<td>$20000</td>
<td>US West1</td>
</tr>
<tr>
<td>Azure</td>
<td>$19100</td>
<td>US West</td>
</tr>
<tr>
<td>AWS</td>
<td>$21000</td>
<td>US West (Oregon)</td>
</tr>
</tbody>
</table>

Source: https://www.usenix.org/conference/atc19/presentation/furakhia
Deep Data Needs and Challenges

Deep Data Challenges

**Scalability**
Ability of a system to adapt to and accommodate an increased bioinformatics workload, and recruiting millions of participants (UI/UX)

**Interoperability**
Ability to run analytical pipelines and process/move data across multiple platforms (e.g., wearables, cloud providers)

**Security & Privacy**
Protect both data and analytical processes from any malicious activities
The Gap

How am I going to scale my gene model?

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

Problem #1: Team Formation in Schools of Medicine

Scalability Team
- Data acquisition
- Storage and processing
- Data distribution
- UI/UX problems

Security & Privacy Team
- Patients’ privacy
- Systems’ Security
Problem #2: Global Governance
Maintain Dynamic Equilibrium b/w Domain Autonomy and Global Governance
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
Wearables can track many things

Heart Rate, Heart Rate Variability, Blood Oxygen Level, Sleep, Skin Temperature, Steps, Respiration Rate, etc.
Flight Study: SpO2

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

“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/media/healtheffects_vulnerable_passengers.pdf

Wearables Data <> Infectious Disease

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

Need: Wearable Data Acquisition at Scale

Infectious Diseases

Heart rate / HRV / Temperature / SpO2
A scalable, secure, and interoperable platform for deep data-driven health management

Amir Bahmani1,2,3,7, Arash Alavi1,2,3,7, Thore Buergel3,7, Sushil Upadhyayula1,3,4, Qiwen Wang1,3,4, Srinath Krishna Ananthakrishnan3, Amir Alavi3, Diego Celis1,3,4, Dan Gillespie3, Gregory Young1,3, Ziye Xing1,2, Minh Hoang Huynh Nguyen1,2, Audrey Haque1,2, Ankit Mathur1,3,4, Josh Payne1,3,4, Ghazal Mazaheri1,3, Jason Kenichi Li1,3,4, Pramod Kotipalli1,3,4, Lisa Liao1,3,4, Rajat Bhasin3, Kexin Cha1,3, Benjamin Rolnik1,3, Alessandra Celli1, Orit Dagan-Rosenfeld1, Emily Higgs1, Wenyu Zhou1,2, Camille Lauren Berry1,3, Katherine Grace Van Winkle1,3, Kévin Contrepois1, Utsab Ray1,2,3, Keith Bettinger1,2, Somallee Datta5, Xiao Li1,6,8, & Michael P. Snyder1,2,3,8

Nature Communications 2021
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]
PHD Case Study 1: COVID-19

Need: Wearable Data Acquisition at Scale

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

- 2017
- 2020

2017

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

2020

- 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.
PHD: Secure Data Collection and Analysis at Scale
PHD COVID-19 Alert System

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


[ Nature Medicine 2022 ]
Real-time Infectious Disease Alerting

2021
Need: First Real-time Alert System

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

80%
early detection for pre-symptomatic and asymptomatic COVID-19 cases with a median of 3 days before symptom onset
**COVID-19 Phase 2: Technological Advancements**

### Table: Comparison of NightSignal and CuSum Statistical Algorithm

<table>
<thead>
<tr>
<th>Feature</th>
<th>NightSignal</th>
<th>Anomaly detection (EllipticEnvelope)</th>
<th>CuSum Statistical Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>80%</td>
<td>69%</td>
<td>72%</td>
</tr>
<tr>
<td>Specificity</td>
<td>87.8%</td>
<td>87.6%</td>
<td>82.1%</td>
</tr>
<tr>
<td>Complexity</td>
<td>LightWeight</td>
<td>Heavy</td>
<td>Heavy</td>
</tr>
<tr>
<td>Sensitive to resolution</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Diagram: Alert State Machine

- **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.

- **MyPHD**
  - Data collection
  - Encrypted Data
  - Pre-processing
  - Encrypted Result
  - Database
  - Model Registry
  - Training
  - Inference
  - A: Average resting heart rate overnight for night / M: Median of averages of resting heart rate overnight for all nights upto night /
MyPHD Supports Multiple Research Studies

- COVID-19 Wearable Study
- Penn/CHOP Transplant Study
- Stress and Resilience Study
- Crohn's Exposome Study
- Aviation Study
- Personalized Baseline Study
- Fiber Cognition Study
- Crash Course Study
- Autism Study
- iPOP Study

Package 1
- Wearable data collection
- Survey collection
- Data analysis
- Customized requests

Package 2
- Wearable data collection
- Survey collection
- Data analysis
- Customized requests

Package 3
- Wearable data collection
- Survey collection
- Data analysis
- Customized requests

Package 4
- Wearable data collection
- Survey collection
- Data analysis
- 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
Timeline
Collaboration

John C. Maxwell:
“Collaboration is multiplication.”
Problem #3: Understand the behaviour of our solution!
Think about scalability at an early stage

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

- $25,000
- $250
- 100X
Hummingbird: Efficient Performance Prediction

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

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

[ Bioinformatics 2021 ]
For a computational pipeline, what machine type is the:
1) Cheapest
2) Fastest
3) Cost-efficient
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
Hummingbird – a tool for effective prediction of performance and costs of genomics workloads on AWS - 2022
Problem #4: Centralized Databases
Interoperability and Federated Computing

Perspective
The Illusion of Inclusion — The “All of Us” Research Program and Indigenous Peoples’ DNA
Keshu Fox, Ph.D.

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)
Swarm: A Federated Cloud Framework

RESEARCH ARTICLE

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

Amir Bahmani, Kyle Ferriter, Vandhana Krishnan, Arash Alavi, Amir Alavi, Philip S. Tsao, Michael P. Snyder, Cuiping Pan

[ PLOS Computational Biology 2021 ]
Data Redundancy: 4x Storage Fees

Data Movement: $100K egress fee (per 1PB)
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
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.

- Minimize data motion: 500GB vs. 4KB
- Facilitates model training without the need of sharing raw data, and therefore strengthens privacy protection.
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 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
Timeline
Serendipity: Meeting Late Prof. James L. Morrison
Educating a New Generation of Computer Scientists and Biologists
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)?

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 (IJROS-AD).
Pre-symptomatic detection of COVID-19 from smartwatch data


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

103K Accesses | 78 Citations | 1391 Altmetric | Metrics

1-click

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 Data Ocean (SDO): Bioinformatics
Internship + Course Impact

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

Summer 2018

2019
- Helped Scaling the COVID-19 Study on MyPHD Apps (Nature Biomedical Engineering)
- Presented at the CyberSecurity Festival

2020
- Finished PhD in Computer Science
- Joined Stanford as a Full-time Employee

2021
- MyPHD Paper (Nature Communications)
- Real-time Alert System (Nature Medicine)
Visit our website and learn about our projects! We are always looking for collaborators!
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
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

Michael Snyder, Chair & Professor of Genetics, Stanford
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
Metabolism: Conversion of food/fuel to energy to run cellular processes

Metabolomics is more time sensitive than other “omics”
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