Title: The Future of Personal Health Assistant: Multimodal Health Data Management Agents

Xuhai “Orson” Xu (xx2489, DBMI, CUIMC VP&S)
Eugene Wu (ew2493, CS, Fu Foundation School of Engineering)

Need and Opportunity for this Course

We are currently witnessing an explosion of personal health data. An average individual generates terabytes of health-related data annually through consumer-grade wearables (e.g., Apple Watch, Oura Ring, Fitbit), mobile sensing apps, and genomic services. Simultaneously, their clinical history resides in Electronic Health Records (EHR) systems. Currently, these data streams are fragmented; a physician seeing a patient for depression has no access to their sleep patterns from the last month, and a patient managing diabetes has to manually correlate their diet logs with glucose readings.

The recent rise of Large Language Models (LLMs) and Agentic Workflows offers a transformative solution: the ability to act as a universal "glue" that can reason across heterogeneous data sources. However, building a reliable, privacy-preserving, and medically grounded "Personal Health Assistant" requires a unique intersection of skills that is currently absent in the Columbia curriculum. Computer Science students often lack an understanding of health data standards (FHIR), clinical workflows, and the ethical nuances of patient agency. Conversely, Biomedical Informatics students often lack the systems engineering skills to build robust data pipelines and complex multi-agent orchestrations.

We propose a new interdisciplinary course that merges **Data Management** and **Human-Centered AI for Health**. This course will teach students to engineer the full stack of a personal health assistant: from the ingestion and cleaning of noisy sensor data to the architectural design of data pipelines, and finally, to the human-computer interaction (HCI) design of the agent itself. This supports the Collaboratory’s mission by embedding deep computational data literacy into the domain of personal health.

Learning Objectives

This course integrates the principles of database systems, AI agents, and health informatics. Students will gain expertise in the following areas:

- **Multimodal Data Engineering:** Ingesting, cleaning, and normalizing heterogeneous data (time-series sensor data, unstructured clinical notes, structured FHIR records)
- **Agentic Architectures:** Designing LLM-based agents capable of tool use (function calling), planning, and memory management (LangChain, LlamaIndex).
- **Retrieval-Augmented Generation (RAG):** Implementing vector databases and semantic search strategies optimized for personal health contexts.
- **Privacy and Ethics:** Understanding HIPAA compliance, differential privacy, and the ethical implications of AI "hallucinations" in health advice.
- **HCI for Health:** Designing applications and interfaces that foster trust, explainability, and engagement in self-tracking and patient-provider communication.
- **Domain Knowledge:** Understanding the basics of chronic disease management, personal health monitoring, and clinical decision support.

Prereqs

Students should satisfy one of the following prerequisites to ensure a balanced cohort of technical builders and domain experts.

Computer Science / Data Science (any one of the following):

- COMS 4111: Introduction to Databases
- COMS 4170: User Interface Design
- COMS 4701: Artificial Intelligence
- COMS 4705: Natural Language Processing

Biomedical Informatics / Health (any one of the following):

- BINF G4000: Introduction to Biomedical Informatics

- BINF G4002: Methods II: Machine Learning for Healthcare
- PHIN P8105: Data Science I (Public Health)
- OR equivalent experience in clinical research or digital health

Format and Outcomes

This course is designed as a systems-building studio. It bridges the gap between static data analysis and dynamic, interactive agents. The course will move beyond theoretical lectures; students will work in mixed teams (pairing CS with Health students) to build a functional **Personal Health Assistant**.

The semester culminates in a "Demo Showcase Day" where students demonstrate their agents interacting with real (anonymized) or synthetic longitudinal health data. We plan to invite researchers and practitioners across the health and healthcare industry (e.g., Pfizer, Google, Oura) to attend and judge the showcase.

Potential examples:

- **The "Contextual" Medical Scribe:** An agent that ingests a patient's wearable data (heart rate, sleep) prior to a clinic visit and generates a summary report for the doctor, highlighting anomalies correlated with the patient's self-reported stress logs.
- **The Metabolic Coach:** A system connecting Continuous Glucose Monitor (CGM) data with photo-based food logs (analyzed by vision models) to answer user queries like, "Why did my sugar spike yesterday afternoon?"
- **The Mental Health Companion:** A privacy-first agent running locally on a device that monitors passive sensing features (screen time, mobility) to detect depressive episodes and offer Just-In-Time Adaptive Interventions (JITAI).
- **The Lab Results Translator:** An agent that pulls raw FHIR data from patient portals and translates complex lab reports into plain language, answering follow-up questions based on verified medical guidelines.

Industry Connection and Integration

We will partner with industry partners, such as Pfizer, Google, and Oura, who will provide guest lectures to discuss the practical challenges of deploying health AI at scale. We have already identified, researchers at Google to offer technical mentorship and feedback on student projects during key design reviews and the final showcase. Crucially, this partnership will support dataset curation and sharing, where our Google contacts will assist in identifying high-quality, de-identified open-source health datasets (e.g., benchmark sensor and metabolic data, public health records databases) to ensure students build their agents on realistic, representative data foundations. We will spend the course preparation time soliciting additional industry experts to provide additional perspectives.

Evaluation criteria

The success of this course will be evaluated through the following measures:

- **Student Project Fidelity**
  - Can the system handle dirty, real-world data? Does the agent hallucinate, or does it correctly cite its sources? Success is defined by the *reliability* of the data pipeline and the *safety* of the agent's interaction.
- **Technical & Domain Critique**
  - Architecture Reviews: Evaluated by Eugene Wu on system efficiency, latency, and data schema design.
  - Clinical/HCI Reviews: Evaluated by Orson Xu on user experience, clinical utility, and safety guardrails.
- **Interdisciplinary Fluency** (measured via pre- and post-course surveys):
  - CS students' understanding of health data standards (FHIR) and clinical privacy.
  - Health students' confidence in understanding LLM limitations, vector stores, and API integrations.

Instructor Roles

- **Orson Xu (DBMI):** Focuses on the "Front of House" and "Domain": Covers Human-Computer Interaction, clinical workflow integration, health sensing modalities (wearables), ethical safety guardrails for AI in medicine, and user evaluation methodologies.
- **Eugene Wu (CS):** Focuses on the "Back of House" and "Infrastructure": Covers data cleaning pipelines, database management, vector embedding strategies, RAG system optimization, and agentic frameworks (state management, tool orchestration).

Ethics Statement

Health data is historically biased, and AI models often propagate these biases. This course places ethics at the center of the technical curriculum, not just as a footnote.

Addressing Algorithmic Bias in Health:

Students will explicitly learn how LLMs can exhibit bias against marginalized populations (e.g., under-diagnosing conditions in specific demographics). We will teach "Red Teaming" for health agents to identify and mitigate these biases before deployment.

Access for Non-CS Majors:

By pairing health/informatics students (who may have biology or public health backgrounds) with CS students, we lower the barrier to entry for building complex AI systems. We will provide "scaffolding" code modules for the heavy infrastructure parts, allowing non-CS students to focus on prompt engineering and interaction design.

Diverse Data Representation:

Course datasets will be curated to represent diverse populations. We will challenge students to build agents that are culturally competent and can handle health data that may vary across socioeconomic statuses (e.g., data from a high-end Apple Watch vs. manual SMS text logs).

Syllabus

The course is expanded to 16 weeks to allow for a more paced absorption of complex technical concepts. Instead of heavy, separate homework assignments, we utilize a Living Lab model.

- Week 1: Students receive a loaner Fitbit. They wear it for the duration of the course.
- Data Usage: The exercises involve querying and processing their own data generated from previous weeks. EHR data will be prepared separately.

Week	Theme	Lecture (Topic)	Studio (Lightweight Exercise)
1	**Kickoff**	**The Quantified Self & Health Data:** Introduction to the fragmented health ecosystem.	**Device Onboarding:** Hand out Fitbits. Account setup, API key generation, and consent forms for self-tracking.
2	**Data Sources**	**Anatomy of a Health Record:** Difference between clinical data (snapshot) vs. sensor data (continuous).	**Data Export 101:** A walk-through script to export the first week of their own Fitbit JSON data.
3	**Phones & Wearables**	**Sensors & Signals:** How accelerometers become "steps" and PPG becomes "heart rate." How to extract nutrition data from smartphone photos.	**Visualization:** Use a pre-made Python notebook to plot your own heart rate from the last 7 days, and generate a food diary from a series of photos.
4	**Industry Talk**	**Health Research at Scale:** The opportunities and challenges of doing health research and deployment in industry	**Project Ideation I:** Students will start to brainstorm their project idea and form teams

5	**Clinical Data**	**The Language of Medicine (FHIR):** Understanding Interoperability, HL7, and standard ontologies.	**FHIR Sandbox:** Query a public synthetic FHIR server (e.g., HAPI) to retrieve a "patient's" medication list.
6	**Databases**	**Storing Health Data:** Timeseries databases vs. Relational/Document stores for health apps.	**Project Ideation II:** Teams will come up with multiple specific project ideas and conduct initial assessment
7	**LLM Intro**	**Generative AI in Health:** Capabilities, limitations, and the risk of hallucination in medical advice.	**Prompt Engineering:** Experiment with prompts to summarize a complex medical abstract into "patient-friendly" language.
8	**RAG**	**Embeddings & Retrieval:** How to ground AI in facts using vector representations.	**Simple Search:** Create embeddings for 5 of their own Fitbit daily summary logs and search for "My most active day."
9	**Project Proposal & Feedback**	**Project Proposal:** Students will present their project ideas and collect feedback	**Project finetuning:** Students will discuss with TA to finetune their project ideas.
10	**Agents I**	**Tool Use & Function Calling:** Teaching LLMs to use "calculators" and "APIs."	**The "Step Count" Tool:** Modify a starter-code agent so it can answer "How many steps did I take yesterday?".
11	**Agents II**	**Planning & Reasoning:** How agents break down complex health questions (e.g., "Why am I tired?").	**Logic Chains:** Paper-prototyping the logic chain an agent would need to correlate sleep data with activity data.
12	**Privacy**	**Ethics, HIPAA & Governance:** Who owns the data? Discussion on privacy-preserving AI.	**Red Teaming:** Try to trick a peer's agent into revealing "private" (synthetic) information or giving dangerous advice.
13	**HCI**	**Interface & Trust:** Designing for vulnerability. How to display uncertainty to patients.	**UI Sketching:** Sketch the interface for a "Health Assistant." focus on how to display AI confidence levels.
14	**Integration**	**System Stitching:** Strategies for connecting the separate modules (Data, Agent, UI).	**Open Studio:** Teams merge their lightweight scripts into a single unified Python script (The "MVP").
15	**Refinement**	**User Testing & Feedback:** Rapid feedback cycles.	**Peer Swap:** Let a classmate ask questions to your agent about *your* (anonymized) activity patterns.
16	**Finale**	**Final Showcase.**	**Demo Day:** Presentation of the lightweight agents to invited guests.

Implementation Timeline

- **Spring 2026:** Curriculum refinement and dataset curation (collecting open-source de-identified wearable/EHR datasets). Hiring a CA (Course Assistant) to develop the starter-code repositories.
- **Summer 2026:** Pilot workshop with a small group of students to test the Agent coding modules.
- **Fall 2026 / Spring 2027:** First official offering of the course.

Resources

One time

- $6000 for Device Pool: Purchase of 50 Fitbit Trackers (e.g., Fitbit Inspire 3 or Charge 6, ~$120/each). This covers a cohort of ~40 students + 10 spares for replacements. These devices will be reused for at least 4-5 semesters.

Each offering (each year, 3 years in total)
- 1 TA per 15 students (Critical for debugging complex agentic code).
- $2,000 for Cloud Compute Credits: OpenAI/Anthropic API credits or Azure/AWS credits.
- $2,000 for Materials & Maintenance: Replacement of lost charging cables, broken straps, sanitization supplies, and printed poster materials for the final showcase.

Relationship with Existing Courses
This course fills a critical void between the **School of Engineering** and the **Medical Center**.
- **Biomedical Informatics:** *BINF G4001 (Intro to Computer Applications in Health Care)* provides a high-level overview of the field but is not a "build" course. *BINF G4002 (Machine Learning for Healthcare)* focuses heavily on training predictive models (e.g., predicting sepsis from EHR data) and evaluating them via AUC/ROC. It does not cover the deployment of these models into user-facing tools. *BINF G4003 (Symbolic AI for Health Care)* covers ontologies (SNOMED, RxNorm) but focuses on classical symbolic reasoning rather than modern Generative AI/LLM architectures. DBMI students rarely get hands-on engineering experience with Vector Databases, RAG pipelines, or API orchestration.
- **Computer Science / Data Science Institute:** *COMS 4111 (Databases)* covers SQL but not unstructured vector retrieval. *COMS 4721 (Machine Learning for Data Science)* focuses on the analysis of static datasets and predictive modeling. *COMS 4170 (UI Design)* covers interfaces but not the backend logic of AI agents. There is no CS/DSI course that covers the Multimodal nature of health data (mixing time-series sensors with unstructured text) nor the Agentic workflows required to build interactive assistants.

This course is the first to combine **Systems**, **AI**, and **Health** into a single applied curriculum.

Long Term Plans
Teaching
We view this course as a foundational pillar for a potential future dual-degree or certification program between DBMI and CS. We are committed to integrating this into their standard teaching rotations. We anticipate high demand and will explore scaling the course via multiple sections or a MOOC adaptation for the Columbia Plus platform.
Funding
The prototypes developed in this class will serve as preliminary data for federal funding proposals. Specifically, we intend to target: **NSF Smart and Connected Health (SCH):** focusing on the data infrastructure for personalized health; **NIH/NLM R01:** focusing on LLM-driven patient communication tools; **Industry Partnerships:** We will leverage the course showcase to build relationships with health-tech companies (e.g., Google Health, Apple Health, Oura, Pfizer) for future student recruitment and research sponsorship.

Support
There is support for this interdisciplinary course from
- Chair of Biomedical Informatics (Noémie Elhadad)
- Chair of Computer Science (Luca Carloni)
- Dean of CUIMC (Katrina Armstrong)
- Dean of Engineering (Shih-fu Chang)