BMEG 3105: Data analytics for personalized genomics and precision medicine

Lecturer: Yu LI (CSE)

Email:liyu@cse.cuhk.edu.hk

Scriber: HAN Mingyu(1155213539)

Lecture 1: Course Introduction

Agenda

- Review the pre-course survey results
- Course logistics
- Brief overview of DATA in personalized genomics and precision medicine

Pre-course Survey Results:

Response: 28/46 (Mostly BME UG students from U2-U4)

Drive of taking the course: 1. For the biological/ genomics/ health applications 2. For the data analytics/ machine learning techniques 3. For research project experience 4. For the three credits and degree requirement 5. Just exploring a new field	Help maybe needed for: 1. Implementation the methods with programming 2. Understand the mathematical background of the data analytical methods 3. Understand concepts in data analytics 4. Apply the methods learned from the course to new data 5. Finding additional materials and resources to learn more about the concepts
Interested topics: 1. Neural networks 2. Cancer genomics 3.Data visualisation 4. Protein-protein / RNA interaction 5. Convolution neural networks	What would we provide? 1. basic Python programming tutorials 2. not touch too much math 3. fundamental concept 4. learn to apply the techniques into the data 5. practical usage and project topics 6. Additional resources and materials will be

	provided

ESTR3605: ESTR mirror of BMEG3105.Additional 45-min lecture each week

Minimum number of students to open the course:5.

Course Logistics

Lectures: Wed 9:30am-11:15am (11:05am), SC L4 Fri 9:30am-10:15am, MMW703

Tutorial: Fri 10:30-11:15am, MMW703

- Slides will be available the day before the lecture day
- there is no video recording, as required by the school
- TA will have several sessions to help you on Python programming
- All the materials will be available on https://lim-zq.github.io/BMEG3105-Fall-2025/

Teaching team:

Yu Li (Instructor) liyu@cse.cuhk.edu.hk Office hour: 3-5pm, Friday Location: SHB-106

ZiqianLin(TA) linziqian@link.cuhk.edu.hk Office hour:3pm-5pm,Tuesday Location: SHB-904
Xinyuan Liu(TA) 1155246738@link.cuhk.edu.hk Office hour:3pm-5pm,Thursday Location: SHB-904

Software and communications:

Blackboard	Piazza (https://piazza.com/cuhk.edu.hk/fall2025/bmeg3105)
The main software to manage the course Grading will be through Blackboard	You can ask questions through Piazza For a personal matter, please use the private post to the instructor and the TA

Grading:

❖ Homework (20%): Three grading homework(5%+5%+5%) and one non-grading programming assignment (5%, make sure you learn something from it)

❖Scribing (10%): Grading scribing. Summarize one of the lectures. Submit it one week after the course. Each student should do at least one lecture. Notice that your note and scribing will be posted online, for others reference. You can choose to remove your name or not. You can sign for at most two, for additional 1%

❖In-class quiz (10%): Two in-class quizzes. The questions will be simple. Mainly for checking the participation. The exact dates are on the website: Oct 15, and Nov26

❖Midterm (20%): A grading midterm exam. One bonus question (2%)

❖Project (20%): A grading project. You can give us your project and seek our help, or we will predefine some projects for you to choose (You should submit a mid-term report (5%), a final report (7%) + presentation (3%) together with the implementation (5%).)

❖Final (20%): A grading final

All the exams or quizzes will be open book!

Programming: Non-grading assignment (5%), Grading programming included in the project (5%)

Bonus:

One bonus question in Midterm (2%)

One additional scribing: 1%

Pre-course survey + Post-lecture survey: 0.5% for each, and the maximum is 3%.

The bonus is sufficient to cover all the programming credit.

Attendance:

do not check the attendance except for the two participation quizzes as well as the mid-term and final

Programming:

Language: Python

Assignment:

	Posted	Due
Programming Assignment 0: Programming environment setup	Sep 5	Sep 17
Assignment 1: About the basic concept of data analytics-1	Sep 12	Sep 24
Assignment 2: About the basic concept of data analytics-2	Oct 3	Oct 15
Programming Assignment 1: About application of DA to the biology	Oct 24	Nov 14
Assignment 3: DA in Personalized Genomics and Precision Medicine	Nov 12	Nov 21

Do be serious to the assignments. They can be very helpful for you to prepare for the mid-term and final.

Scribing:

Summarize one of the lectures, Submit it within one week after the lecture

Grades will be deducted by 25% for each additional late day

Lecture date: 5 Sep. Deadline: 12 Sep, 11:59pm

First two lectures: additional one week for submission

Midterm Date:

Oct 17 (Fri) In class, open-book

Project

Project milestone report(Proposal) : 1-page report (Due: Nov 7)

1. Title, author

2. What problem do you want to do? Why is the problem interesting? (1%)

3. What data are you going to process? The source, the size, the sample of the data(1%)

4. What' s the output of your method? (1%)

5. How are you going to do it? Describe the method step by step, from input to output(1%)

6. What are the expected results? How are you going to evaluate the results? (1%)

7. What have you done?

Project report : No length requirement, submit the report together with codes (5%, whether it is correct or not) (Due: Dec2)

1. Title, author

2. What problem do you want to do? Why is the problem interesting and important?(0.5%)

3. What data have you processed? The source, the size, the sample of the data (0.5%)

4. What have you done to resolve the problem? Describe the method step by step, from input to output (2%)

5. What are the results? (1.5%)

6. Result evaluation (1.5%)

7. Any idea for further improvement? (1%)

Project presentation (Date: Nov 21, Nov 28):7 mins for each student

Will be graded in the following way:

1.Logic (1%)

What is the problem?

Why is it important?

How do you resolve it?

The overview of your idea

The overview of the results

2.Clarity (1%)

Whether the audience can understand and follow the presentation Slides

3.Preparation (1%)

Clear illustration No typos, no grammar errors

Late days

Can be used on Assignment 1,2,3 and programming assignment, project mid-term report

Cannot be used on final project report and scribing report

6 late days total, 2 max for assignments

Grade will be deducted by 25% for each additional late day

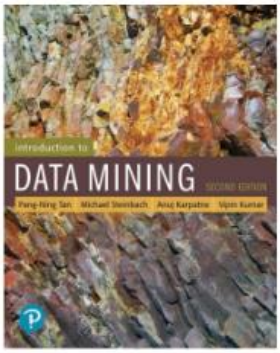
Let TA know when it is necessary to use the late dates

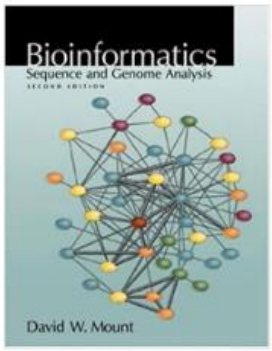
Course schedule

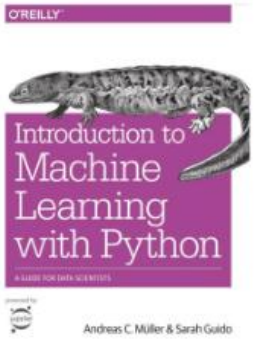
Lecture	Date	Location	Topic	Slides	Notes	Reading	Important dates (All due at 11:59 pm)
1	Sep 3 (Wed)	SC L4	Introduction				
2	Sep 5 (Fri)	MMW703	Data & Python				PA0 posted
3	Sep 10 (Wed)	SC L4	Data & Python & Sequence				Scribing preference registration due
4	Sep 12 (Fri)	MMW703	DP				A1 posted
5	Sep 17 (Wed)	SC L4	Assembly & Mapping				PA0 due
6	Sep 19 (Fri)	MMW703	Data exploration				
7	Sep 24 (Wed)	SC L4	Clustering				A1 due
8	Sep 26 (Fri)	MMW703	Classification				
9	Oct 3 (Fri)	MMW703	Classification & Perf evaluation				A2 posted
10	Oct 8 (Wed)	SC L4	Perf evaluation				
11	Oct 10 (Fri)	MMW703	Dim reduction				
12	Oct 15 (Wed)	SC L4	Midterm review				Quiz, A2 due
13	Oct 17 (Fri)	MMW703	Midterm				8:30am-11:15am, Midterm exam
			Module 2 start				
14	Oct 22 (Wed)	SC L4	Multi-omics overview				
15	Oct 24 (Fri)	MMW703	Cancer genomics overview				PA1 posted
16	Oct 31 (Fri)	MMW703	Genomics data analysis				
17	Nov 5 (Wed)	SC L4	Single cell genomics				
18	Nov 7 (Fri)	MMW703	Data visualization				Project M-report (Proposal) due
			Module 3 start				
19	Nov 12 (Wed)	SC L4	Deep learning				A3 posted
20	Nov 14 (Fri)	MMW703	CNN				PA1 due
21	Nov 19 (Wed)	SC L4	EHRs & Text				
22	Nov 21 (Fri)	MMW703	Project Presentation				A3 due
23	Nov 26 (Wed)	SC L4	Course review				Quiz
24	Nov 28 (Fri)	MMW703	Project Presentation				Project report due on 2 Dec

Reference book

Module 1







Module 2 + Module 3

Cancer genomics, Single cell, Multi-omics, EHRs, Protein-RNA interaction are all cutting-edge research topics. There is no such a book covering all of them.

provide the reading materials and reference books when we are there.

Overview of DATA in personalized genomics and precision medicine

1.Why data analytics?

1. Massive of data is being collected and warehoused

1.1 Web data: Meta, Google, Amazon, X, TikTok

1.2 Biological data: DNA sequences, protein sequences

1.3 Bank/credit card transaction data: Alipay, PayPal

1.4 Mobile data: China Mobile, CSL

2. Computers have become cheaper and more powerful

3. Data analytics are useful

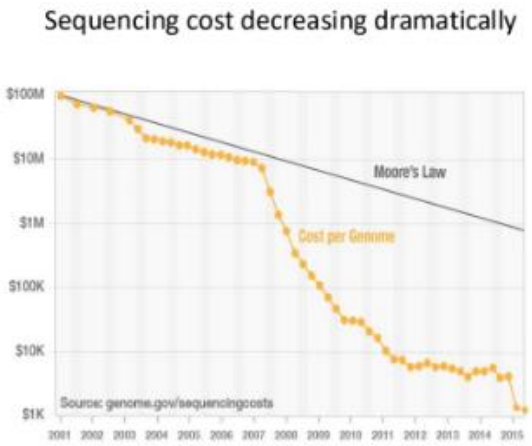
3.1 Aggregate data

3.2 Generate hypothesis

3.3 Support the conclusion

2.Why data analytics in personalized genomics and precision medicine?

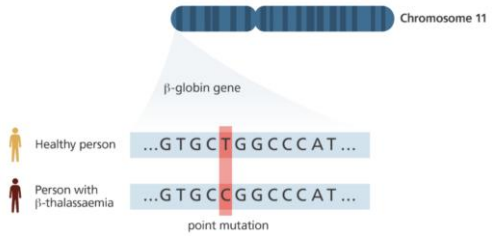
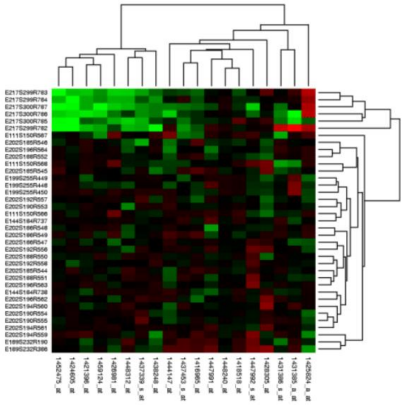
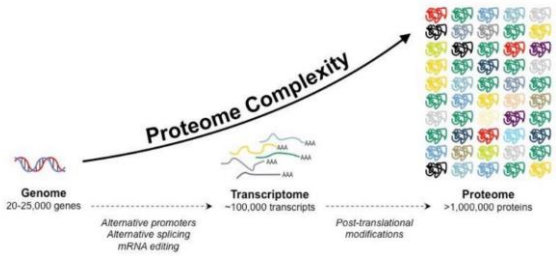
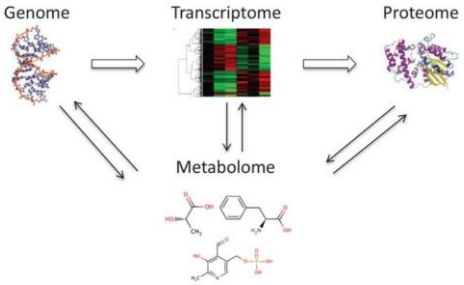
Tons of sequencing and health data available and waiting to be analyzed

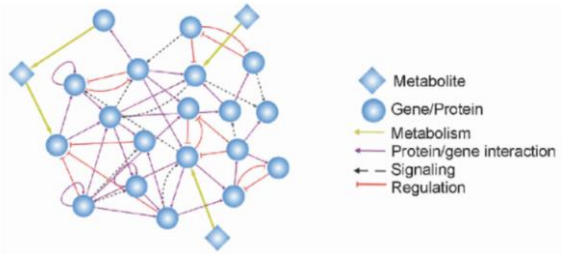
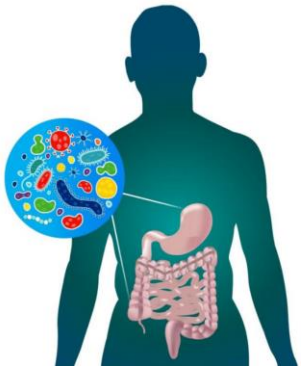
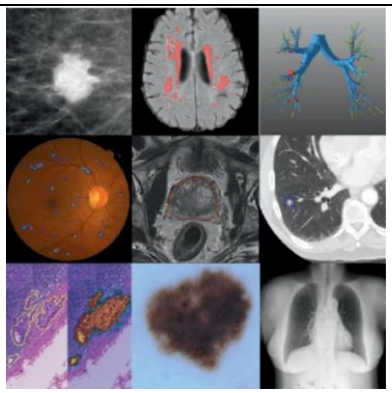


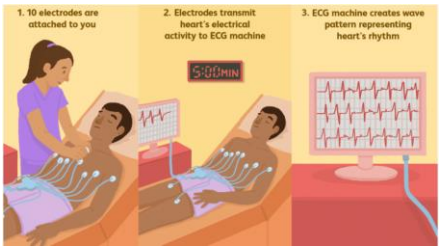
Global efforts in building biobank

Location	Biobank	N (goal)
Canada	CARTaGENE biobank[119]	43,000
USA	All of Us[33] Million Veteran Program[49]	1,000,000 > 600,000
Mexico	The Mexico City Prospective Study[92]	150,000
Iceland	deCODE Genetics	500,000
UK	UK Biobank[36] Avon Longitudinal Study of Parents and Children (ALSPAC)[120]	500,000 > 15,000
Netherlands	Lifelines Biobank[120]	> 167,000
Denmark	Danish National Biobank[121]	
Norway	HUNT - Nord-Trøndelag Health Study[122]	125,000
Sweden	Biobank Sweden	
Finland	FinnGen	500,000
Estonia	Estonian Biobank[123]	52,000
Israel	Project 10K	10,000
Saudi Arabia	Saudi Biobank	200,000
Qatar	Qatar Biobank[124]	60,000
China	China Kadoorie Biobank[51] Guangzhou Biobank[125]	> 500,000 30,000
Japan	BioBank Japan[126]	200,000
Korea	National Biobank of Korea[127]	500,000
Taiwan	Taiwan Biobank[128]	200,000

3. What data can we have to measure a person?

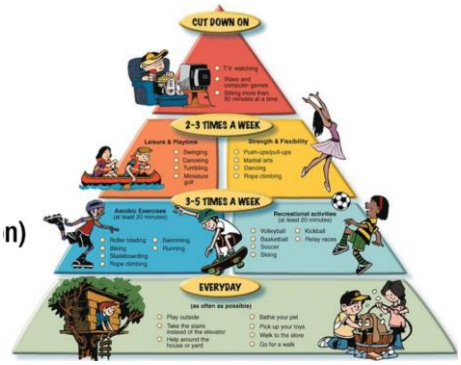
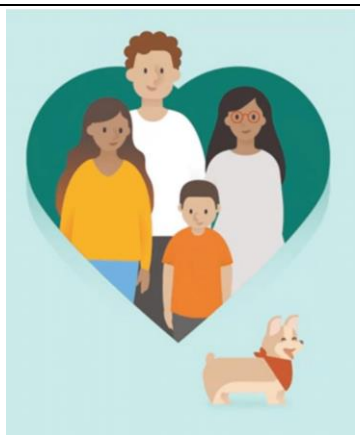
Data	Example
Gene and mutations	 Whole Genome Sequencing (WGS), SNP array (e.g., 23andMe), BRCA1/2 gene mutation test
Gene expression(Transcriptome)	 RNA-Seq, Microarray analysis, Single-cell RNA sequencing (scRNA-seq)
Proteome	 Mass spectrometry-based proteomics, Antibody arrays, Blood protein biomarkers (e.g., CRP, IgG)
Metabolome	 Mass spectrometry (LC-MS/GC-MS), NMR

	spectroscopy, Blood glucose, cholesterol, amino acids levels
Molecular network & cellular network	 Protein-protein interaction maps, Metabolic pathways (e.g., KEGG), Gene regulatory networks
Microbiome(Oral and gut)	 16S rRNA sequencing (to identify gut bacteria composition), Stool test for pathogens
Organ(Biomedical imaging)	 MRI, CT Scan, X-Ray, Ultrasound, Echocardiogram

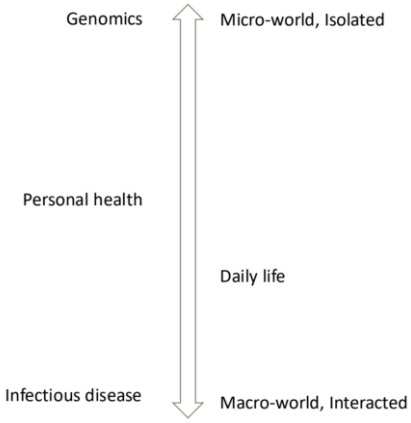
<table>
<tr>
<td>Hospital test(Blood test and so on)</td>
<td>
<table>
<tr><th>Blood Test</th><th>Result</th><th>Normal Value</th></tr>
<tr><td>WBCs (billion/L)</td><td>8.00</td><td>3.5 to 10.5</td></tr>
<tr><td>Neutrophils (%)</td><td>62</td><td>40 to 70</td></tr>
<tr><td>Lymphocytes (%)</td><td>28</td><td>25 to 45</td></tr>
<tr><td>Monocytes (%)</td><td>10</td><td>2 to 8</td></tr>
<tr><td>Eosinophils (%)</td><td>1</td><td>1 to 5</td></tr>
<tr><td>Basophils (%)</td><td>0</td><td>0 to 1</td></tr>
<tr><td>RBCs (trillion/L)</td><td>3.84</td><td>4.3 to 5.7</td></tr>
<tr><td>Hb (g/dL)</td><td>11.7</td><td>13 to 17</td></tr>
<tr><td>Hematocrit (%)</td><td>37</td><td>37 to 52</td></tr>
<tr><td>Platelets (billion/L)</td><td>262</td><td>150 to 450</td></tr>
</table>
MRI, CT Scan, X-Ray, Ultrasound, Echocardiogram
</td>
</tr>
<tr>
<td>Electrocardiography</td>
<td>


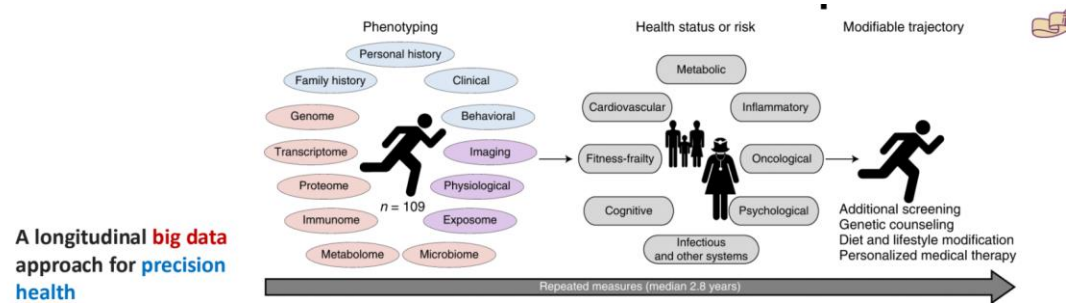
Resting ECG
</td>
</tr>
<tr>
<td>Demographic information(Age, gender,location and so on)</td>
<td>
<table>
<tr><th>Demographic Categories</th><th>Frequency</th><th>Valid Percentage</th><th>U.S. National Census Data (2012), %</th></tr>
<tr><td>Gender</td><td></td><td></td><td></td></tr>
<tr><td>Female</td><td>150</td><td>49.5</td><td>51.1</td></tr>
<tr><td>Male</td><td>153</td><td>50.5</td><td>48.9</td></tr>
<tr><td>Age</td><td></td><td></td><td></td></tr>
<tr><td>18–24</td><td>24</td><td>9.5</td><td>11.2</td></tr>
<tr><td>25–34</td><td>53</td><td>20.9</td><td>13.4</td></tr>
<tr><td>35–44</td><td>36</td><td>14.2</td><td>12.9</td></tr>
<tr><td>45–54</td><td>67</td><td>26.5</td><td>14.2</td></tr>
<tr><td>55–64</td><td>61</td><td>24.1</td><td>12.3</td></tr>
<tr><td>65+</td><td>12</td><td>4.7</td><td>13.5</td></tr>
<tr><td>Not specified</td><td>50</td><td>—</td><td>—</td></tr>
<tr><td>Ethnicity</td><td></td><td></td><td></td></tr>
<tr><td>Euro-American/Caucasian</td><td>241</td><td>79.5</td><td>63.0</td></tr>
<tr><td>African American</td><td>26</td><td>8.6</td><td>13.1</td></tr>
<tr><td>Hispanic/Latino(a)</td><td>17</td><td>5.6</td><td>16.9</td></tr>
<tr><td>Asian American</td><td>11</td><td>3.6</td><td>5.1</td></tr>
<tr><td>Other</td><td>8</td><td>2.6</td><td>1.9</td></tr>
<tr><td>Marital status</td><td></td><td></td><td></td></tr>
<tr><td>Married</td><td>168</td><td>55.4</td><td>56.4</td></tr>
<tr><td>Never married</td><td>70</td><td>23.1</td><td>26.9</td></tr>
<tr><td>Divorced/separated/widowed</td><td>68</td><td>21.5</td><td>16.7</td></tr>
<tr><td>Employment</td><td></td><td></td><td></td></tr>
<tr><td>Full time</td><td>144</td><td>47.5</td><td>64.7</td></tr>
<tr><td>Part time</td><td>45</td><td>14.9</td><td>25.7</td></tr>
<tr><td>Unemployed/retired</td><td>114</td><td>37.6</td><td>9.6</td></tr>
<tr><td>Highest level of education</td><td></td><td></td><td></td></tr>
<tr><td>4-year college or graduate degree</td><td>140</td><td>46.2</td><td>28.5</td></tr>
<tr><td>High school degree or other</td><td>163</td><td>53.8</td><td>57.2</td></tr>
<tr><td>Annual household income</td><td></td><td></td><td></td></tr>
<tr><td>Less than US$20,000</td><td>39</td><td>12.9</td><td></td></tr>
<tr><td>US$20,000–US$39,999</td><td>84</td><td>27.7</td><td></td></tr>
<tr><td>US$40,000–US$59,999</td><td>63</td><td>20.8</td><td></td></tr>
<tr><td>US$60,000–US$79,999</td><td>43</td><td>14.2</td><td>N/A*</td></tr>
<tr><td>US$80,000–US$99,999</td><td>33</td><td>10.9</td><td></td></tr>
<tr><td>US$100,000 or up</td><td>41</td><td>13.5</td><td></td></tr>
<tr><td>Total</td><td>303</td><td>100</td><td></td></tr>
</table>
Note. N/A = not available.
*The category of annual household income published by the U.S. Census Bureau is different from the one in this study and the distribution is as follows: under US$15,000 (13.0%), US$15,000–US$24,999 (11.9%), US$25,000–US$34,999 (11.1%), US$35,000–US$49,999 (14.1%), US$50,000–US$74,999 (18.1%), US$75,000–US$99,999 (11.5%), and US$100,000 and over (20.1%).
Age, Gender, Postal code, Ethnicity, Education level, Occupation
</td>
</tr>
<tr>
<td>Personal statement and doctor diagnosis</td>
<td>


</td>
</tr>
</table>

	Patient-reported symptoms, Electronic Health Record (EHR) diagnoses (e.g., ICD-10 codes: I10 for hypertension)
Living habit(Exercise)	 Wearable data (steps, active minutes, heart rate from Fitbit/Apple Watch), Self-reported workout logs
Diet	 Food diary apps (MyFitnessPal), Nutritional intake analysis (calories, macros, micronutrients)
Family history	 History of heart disease, cancer, or diabetes in first-degree

	relatives
Communication and social media data	 Language sentiment analysis, social network size, frequency of interactions (for mental health research)
Environment(Pollution)	 PM2.5 exposure level (from local air quality monitors), Lead level in drinking water
Travel history(Global pandemic)	 Travel itinerary during a global pandemic



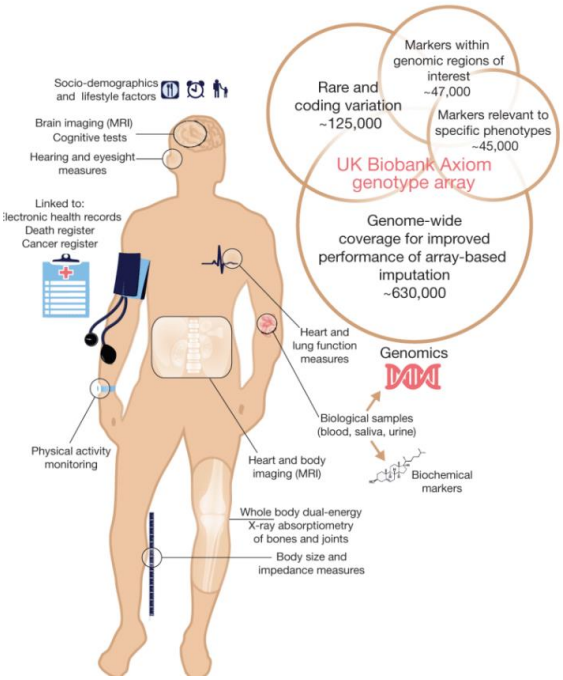


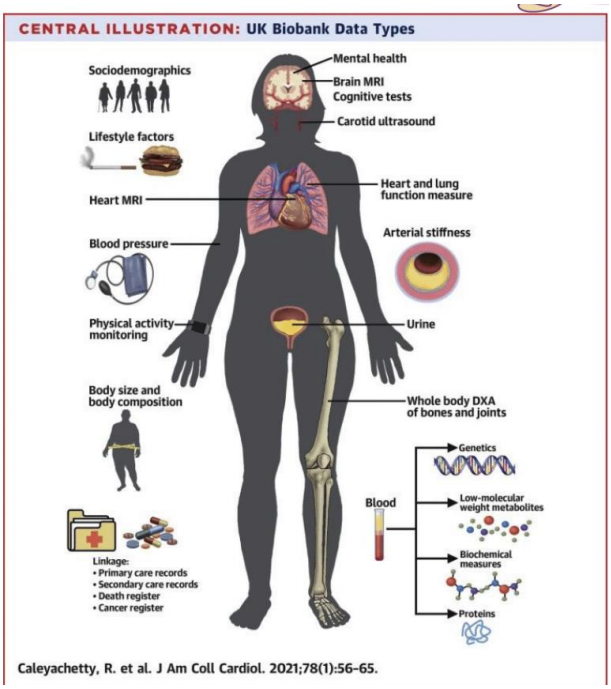
Goal: Use long-term, frequent data tracking for personalized health.

Method: Tracked 109 people for 4-8 years, collecting a massive amount of data.

Data Collected: Everything from genes and proteins to blood tests, wearables, and lifestyle questionnaires.

Purpose: To build a complete health profile and assess risks for diseases like diabetes, heart disease, and cancer. Enable personalized actions like diet plans, genetic counseling, and custom medical treatment.





The UK Biobank is a major biomedical database with deep genetic and health data from 500,000 people. It combines genomics, imaging (MRI, DEXA), physical measures, and long-term health records to study disease causes and prevention.

Learning outcome

1.Learn thefundamental conceptof data analytics

2.Know the various data ingenomicsandmedicine

3.Apply the data analytics techniques toprocess the data and resolveproblemsin biology