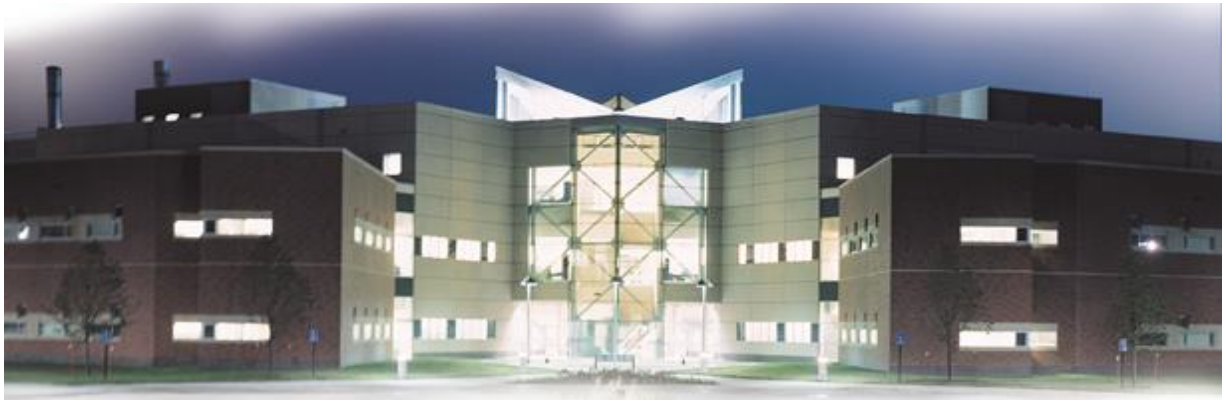
University of Nebraska Omaha

Big Data Analytics and AI Approaches for Advancing Biomedical Research and Personalized Healthcare



September 2023

Hesham H. Ali, PhD

UNO Bioinformatics Core Facility

College of Information Science and Technology

Hesham H. Ali, PhD

Background and Professional Experience:

- Professor of Computer Science, College of Information Science and Technology, University of Nebraska Omaha (UNO)
- Director of UNO Bioinformatics Core Facility
- Served as the Lee and Wilma Seemann Distinguished Dean of UNO College of Information Science and Technology between 2006 and 2021
- Published over 200 articles, books and book chapters in various IT areas including scheduling, distributed systems, data analytics, wireless networks, and Bioinformatics
- Has been leading a Research Group that focuses on developing innovative computational tools to analyze all health-related data with the goal of advancing next generation of biomedical research and contributing to the preventive and personalized healthcare initiatives

Biomedical Informatics: The new major revolution in Sciences!

- Each generation, a scientific discipline emerges with a bang and promises to change the way we do things – a game changer.
- The last major new discipline was Computer Science over 50 years ago.
- Is Biomedical Informatics (BMI) the new major discipline for this generation?
- The connection to Human Health add another layer of significance to BMI

Innovation/Discovery for All!!

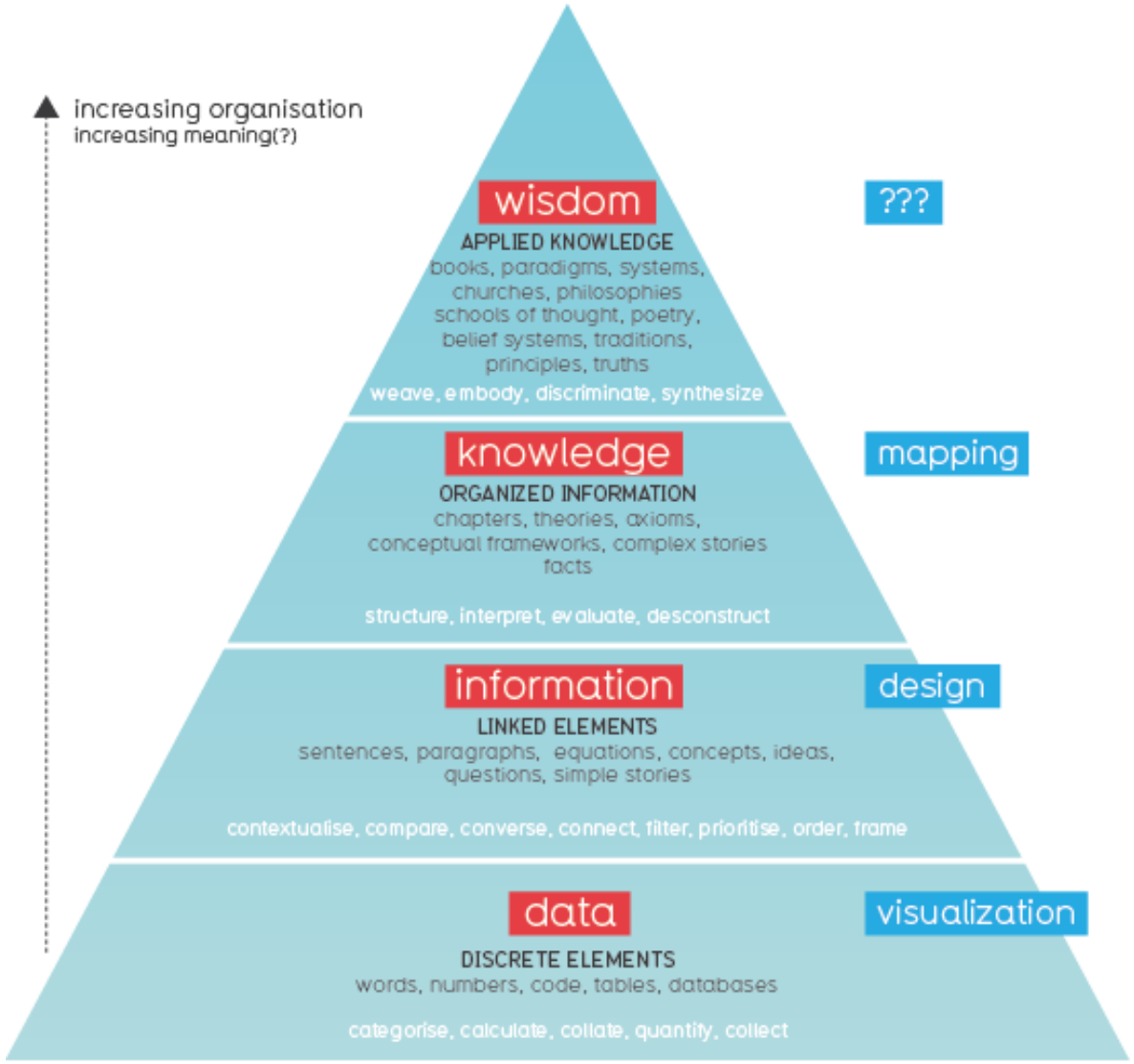
- Innovation and research is no longer conducted by select group in isolated labs in few large institutions
- Research is becoming a driver to the education and industry – a process that is leading to discovery-based learning and innovation
- Availability of all sorts of data to everyone is a key to this revolution
- In addition, computational tools and computational facilities have evolved significantly recently

Is it all about the Data?

- How it all began?
  - Advances in medical instruments and computational technologies led to new new research directions
  - Massive accumulation of Biomedical data led to investigating new potential discoveries
  - The availability of enormous various types of public/private Biomedical data
  - All sorts of data: Bioinformatics - Health Informatics - Imaging - Public Health Informatics – Biomedical/Mobility Devices – etc.
  - How to take advantage of the available data

Is having such rich data enough to advance biomedical Research?

Data-Information-Knowledge-DSS



http://www.ritholtz.com/blog/wp-content/uploads/2010/11/data_info_knowledge_wisdom.png

Bioinformatics and Human Health

- Integrating the exploding computational revolution and the increasing availability of biological/biomedical data promises so much advancements associated with biomedical research and healthcare.
- The impact of Bioinformatics research on healthcare practices or human health remains somehow unsatisfactory.
- It can be argued however that advances in Bioinformatics has bigger impact on biomedical research.
- How about the impact of biomedical research on healthcare and human health?

Biomedical Informatics in 2023

- Every generation, a scientific discipline emerges with a bang and promises to change the way we do things and dominates the scene and dictates the terms – Bioinformatics or Biomedical Informatics has that potential to dominate the scene and significantly impact biosciences and healthcare.

- However, back in the mid nineties, one would have expected BMI with all its components to be further along after almost 30 years, in terms of development and impact.

BMI and the Current Debate

On the one hand, BMI has delivered the goods: recent advances in data availability and informatics tools have advanced biomedical research in a big way.

On the other hand, some argue that BMI has not delivered at the big stage:

the impact on the critical aspects of biomedical research is relatively minimal and the contribution in advancing healthcare remains limited.

The Trendiness of BMI

- With parents in computing and biosciences, Bioinformatics inherited genetic disposition to high degree of trendiness
  - Mostly due to newer biological data but due to computational tools in some cases
  - The availability of new types of data due to new data generating technologies
  - The availability of computational tools and advanced computational facilities
- A current relevant question: Is AI the trend of the day (this period) or it is here to stay – what is the long-term impact.
- AI tools and the Black Box model

Current Challenges for BMI

- On the biomedical side:
  - Too much focus on data collection
  - Competition to own the latest technology
  - Excitement associated with New technologies – which leads to more raw data
  - The black box syndrome
- On the computational side:
  - Certain level of casualness remain a major concern – just another application domain
  - Inconsistent results – lack of robustness and reproducibility
  - Heuristics and thresholds
  - Lack of Biomedical-rich integration

Big Data Analytics is a Big Deal

- Extracting useful (sometimes critical) knowledge from the available raw data can be considered as the single most outstanding research area of this generation
- Many advancements in numerous application domains highly depend on our success in developing data analytics tools

How to Extract Knowledge from Raw Data?

- Just the data is not enough
- A robust, powerful and flexible model is needed
- The model needs to have the ability to learn and get smarter, hence the need for artificial intelligence and machine learning
- We propose a population analysis model that utilizes:
  - Graphs modeling
  - network theory and
  - data analytics

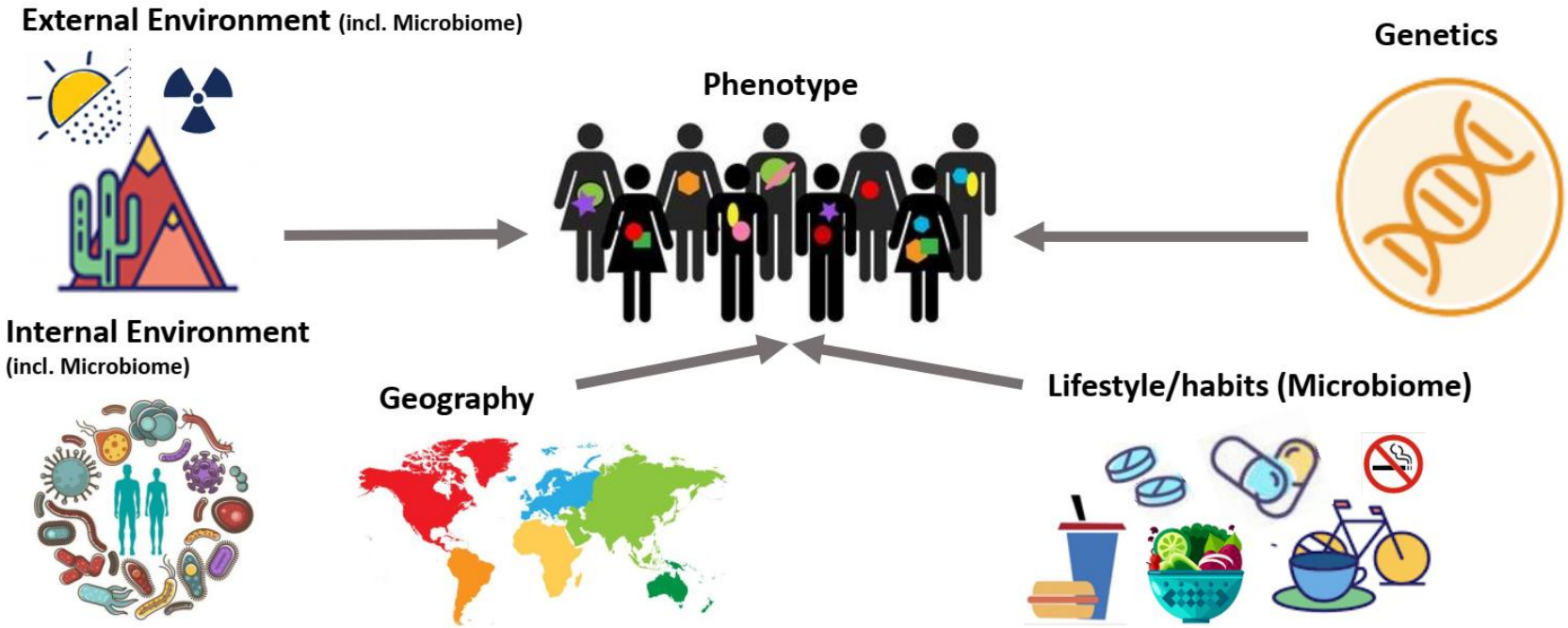
Taking Advantage of Available Data

- Every time the continuously evolving biomedical technologies make it possible for bioscience researchers to have access to new type of biological data, similar research questions attract new studies:
  - Would it possible for the new available data to provide new biological signals that can be used to classification purposes?
  - Can we use the new data types to profile groups with common phenotypes?
  - Would the biological signals associated with the new data be robust enough to be used for the purpose of disease diagnosis and/or the assessment of different treatments for certain health conditions.

History: Taking Advantage of Available Data

- Previous and on-going efforts:
  - Sequences: Alignment tools
  - Transcriptomics: gene expressions – Microarrays - RNASeq
  - Genomics: Genome assembly, genetic variants, SNPs, copy numbers.
  - Proteomics, metabolomics, etc. CNB
- Recent and new efforts:
  - Microbiomes data
  - Mobility data
  - Nano particles data

Weak Signals and Data Integration



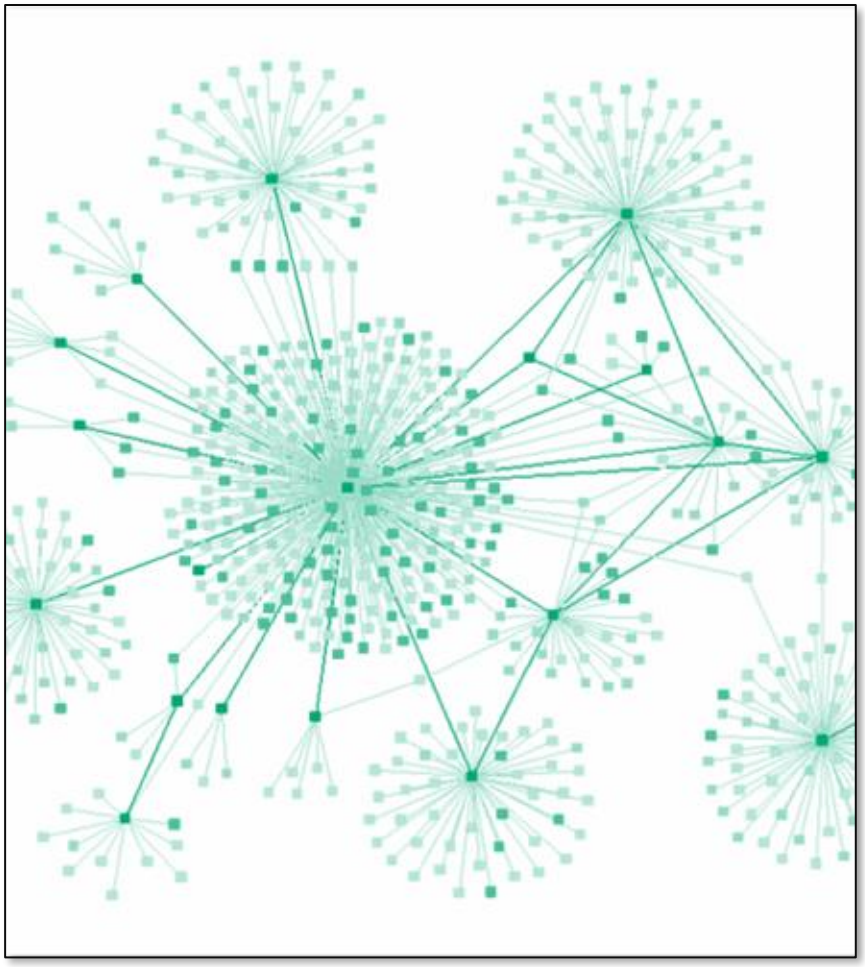
Analytics using Graph Modeling, and Complex Networks

- Since almost all available data in biomedical systems are made of elements and their interrelationships, graphs and networks represent a powerful and flexible tool to model such systems and solve associated problems.
- It is difficult to provide useful analysis or assessment elements in isolation. Many analysis-related studies are conducted by comparing elements to its population
- We approach big data analytics by building networks (graphs) of elements under study using different types of inter-relationships among the elements – such as correlations or cooccurrences
- We then use graph theoretic properties of constructed networks to mine useful knowledge associated with big data

Networks and Population Analysis

- A network represents elements and their inter-relationships
- Nodes → elements

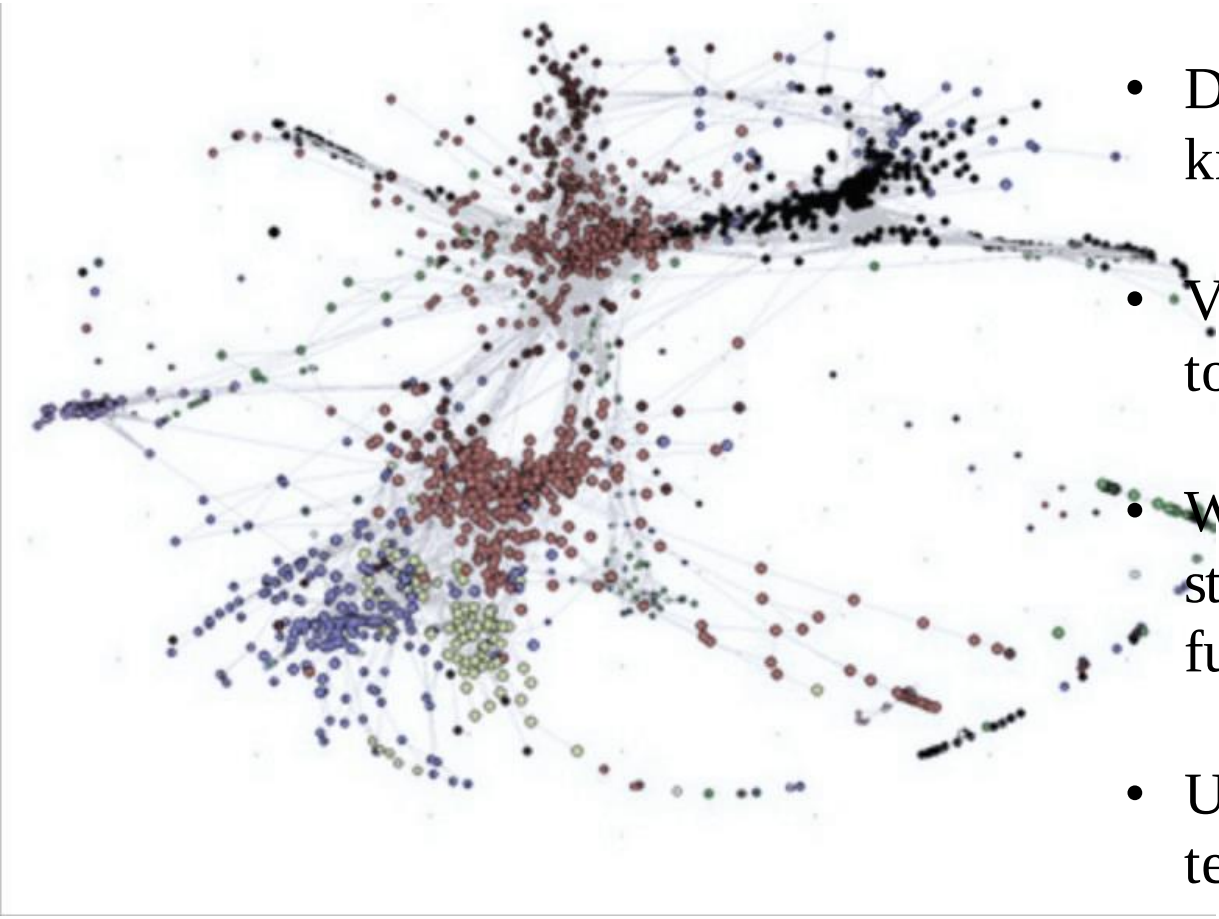
  Edges → relations
- Can represent multiple types of elements and relationships
- Analysis conducted at the:
  - Element level (nodes
  - Group Level (clusters)
  - Entire network level



AI and Network Models

- AI is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans
- AI is often used to describe machines (computers) that mimic "cognitive" functions associated with the human mind, such as "learning" and "problem solving"
- Machine learning is a form of AI that enables a system to learn from relevant available data rather than through explicit programming or deterministic algorithms
- AI tools still need friendly structure of the data

Population Analysis



- Able to handle ‘big’ data
- Draws from centuries of knowledge in graph theory
- Visually appealing and easy to understand
- When built correctly, structures can be tied to function
- Used in social, biological, technical applications

CS1: Gene Expression Levels as Signals

5 sets of temporal gene expression data

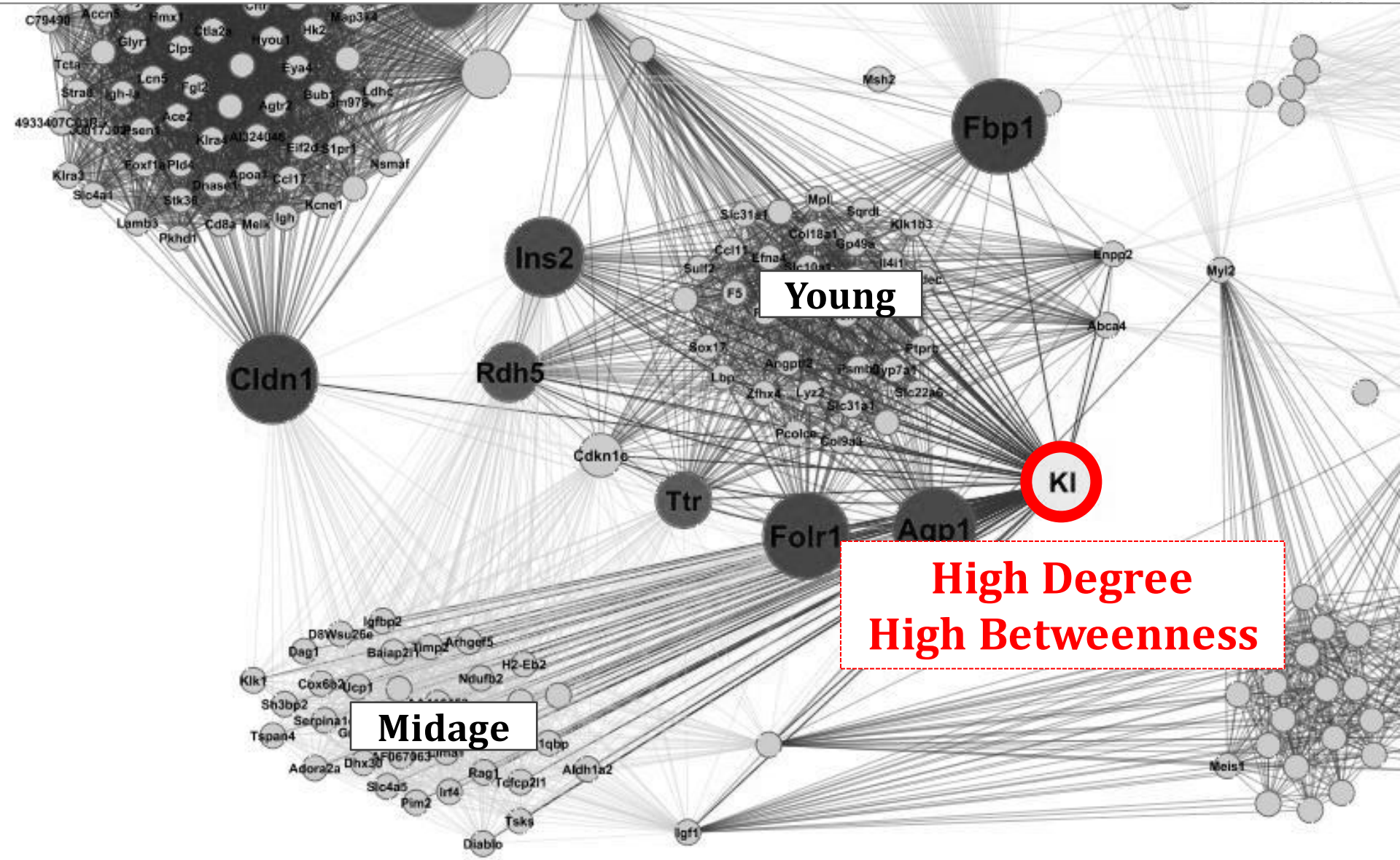
Strain	Gender	Tissue Type	Ages
BalbC	Male	Hypothalamus	Young, mid-age, aged
CBA	Male	Hypothalamus	Young, mid-age, aged
C57_J20	Male	Hypothalamus	Young, aged
BalbC	Female	Hypothalamus	Young, aged
BalbC	Female	Frontal cortex	Young, aged

Hub Lethality

- Young Male BalbC Mouse
  - 12/20 hubs tested for *in vivo* knockout
    - 8/12 lethal phenotype pre-/peri-natally
    - 4/12 non-lethal but system-affecting
    - 0/12 no observed phenotype

- Aged Male BalbC Mouse
  - 11/20 hubs tested for *in vivo* knockout
    - 7/11 lethal phenotype pre-/peri-natally
    - 3/11 non-lethal but system-affecting
    - 1/11 no observed phenotype (Aldh3a1)

Critical Node: Klotho



HIV and Drug Addiction

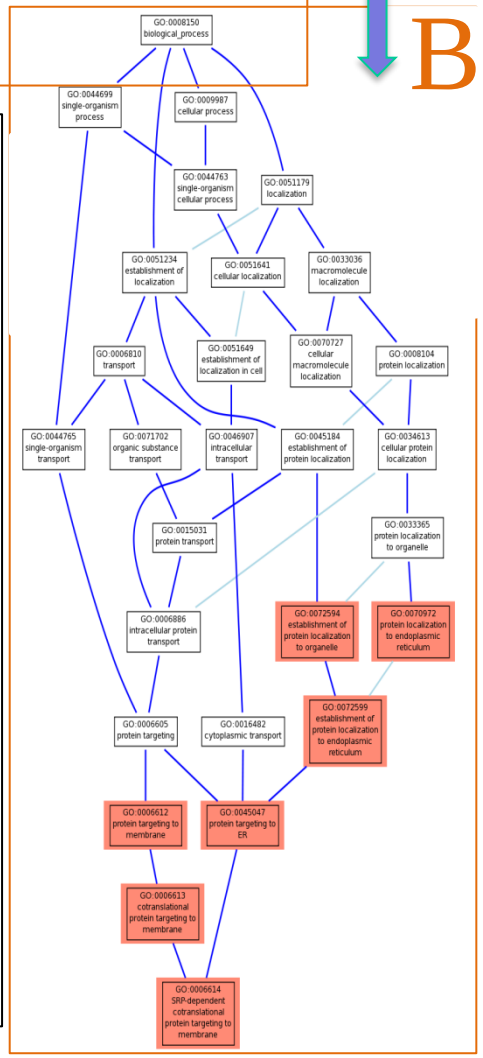
- Methamphetamine is a major drug of abuse with reported high use by HIV-infected groups
- Methamphetamine users have higher risk of getting HIV infection
- Impact on nervous system is higher when Methamphetamine is used by HIV infected individual (neuronal injury)

Clusters Enriched in Specific Functions

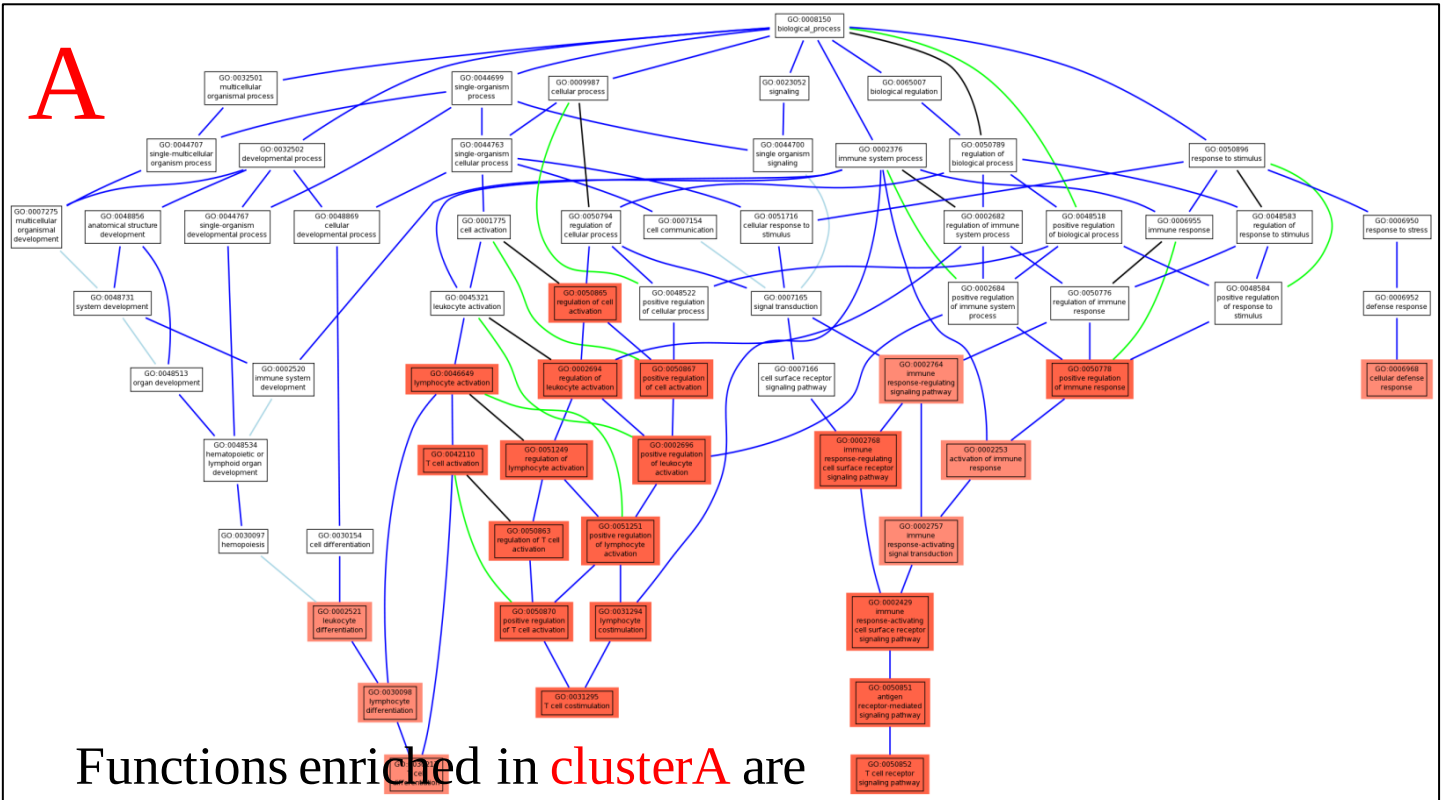
UNO BIOINFORMATICS

Functions enriched in clusterB are protein targeting and localization.

B



A



Functions enriched in clusterA are all immune response specific.

CS2: Mobility Data as Signals

- The explosive and widespread use of sensors and wearable devices
- Many studies connected the way people move to their health (physical and mental), safety concerns and overall states of people and environments
- We can collect and store mobility levels parameters
- Data Analytics, AI and machine learning are critical in extracting knowledge from mobility data

Mobility & Health

- Inherent asset of every human being

Healthy **Mobility skills** conditioned

Mobility Monitoring

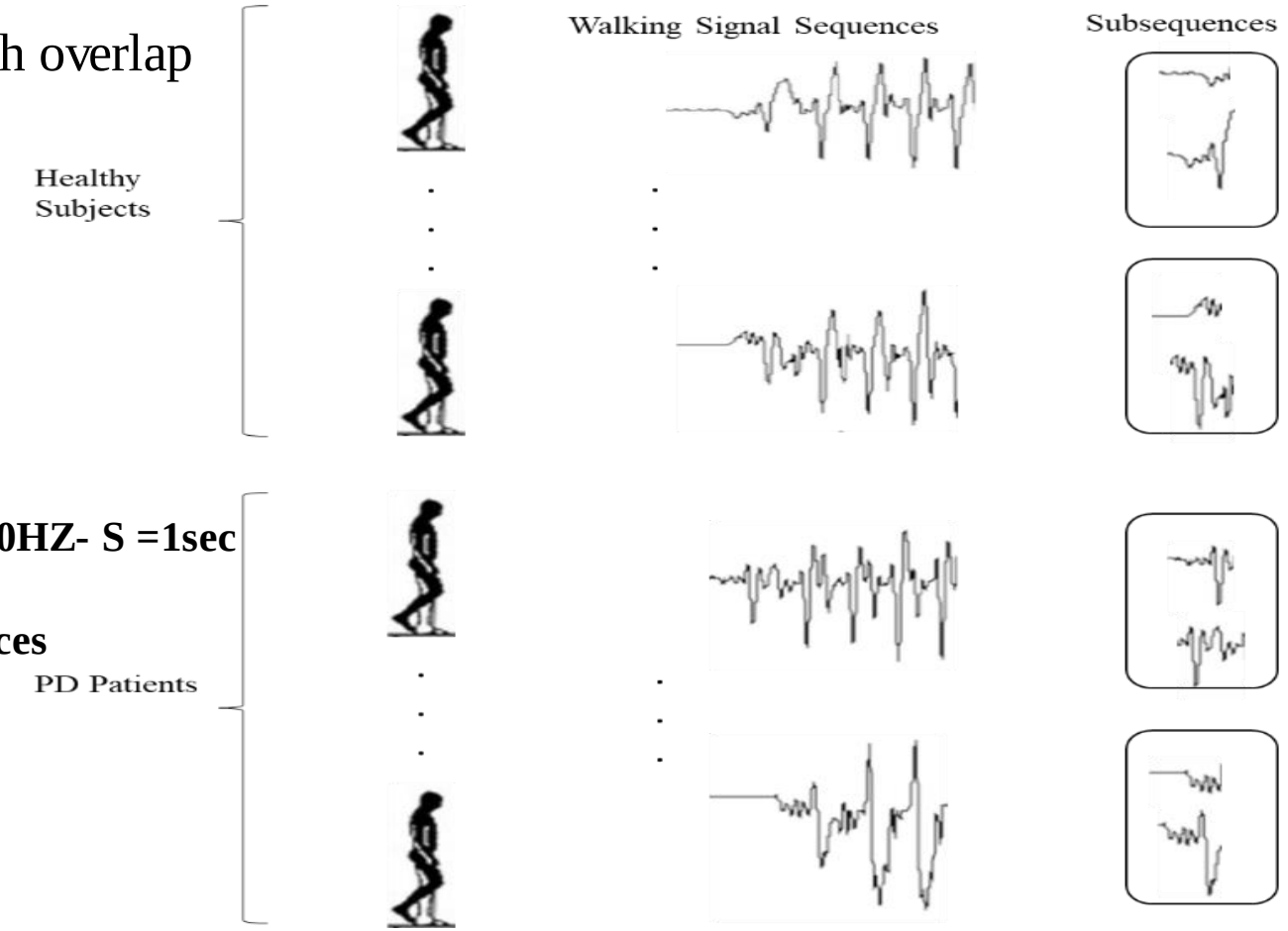
- Availability of many large useful devices – focus on collecting relevant data
- Availability of numerous helpful software packages
- Lack of data integration and trendiness of the discipline
- Fragmented efforts by computational scientists and domain experts
- Lack of translational work – from the research domain to engineering and healthcare applications
- Increasing interest among researchers, industry and educators

How to Compare Mobility Patterns -Movement Words Coding Scheme

Generating Subsequences of Signal

Window Size = S sec with overlap

$$N = \sum_{i=1}^{n} 2 * \left(\frac{ti}{s}\right) - 1$$

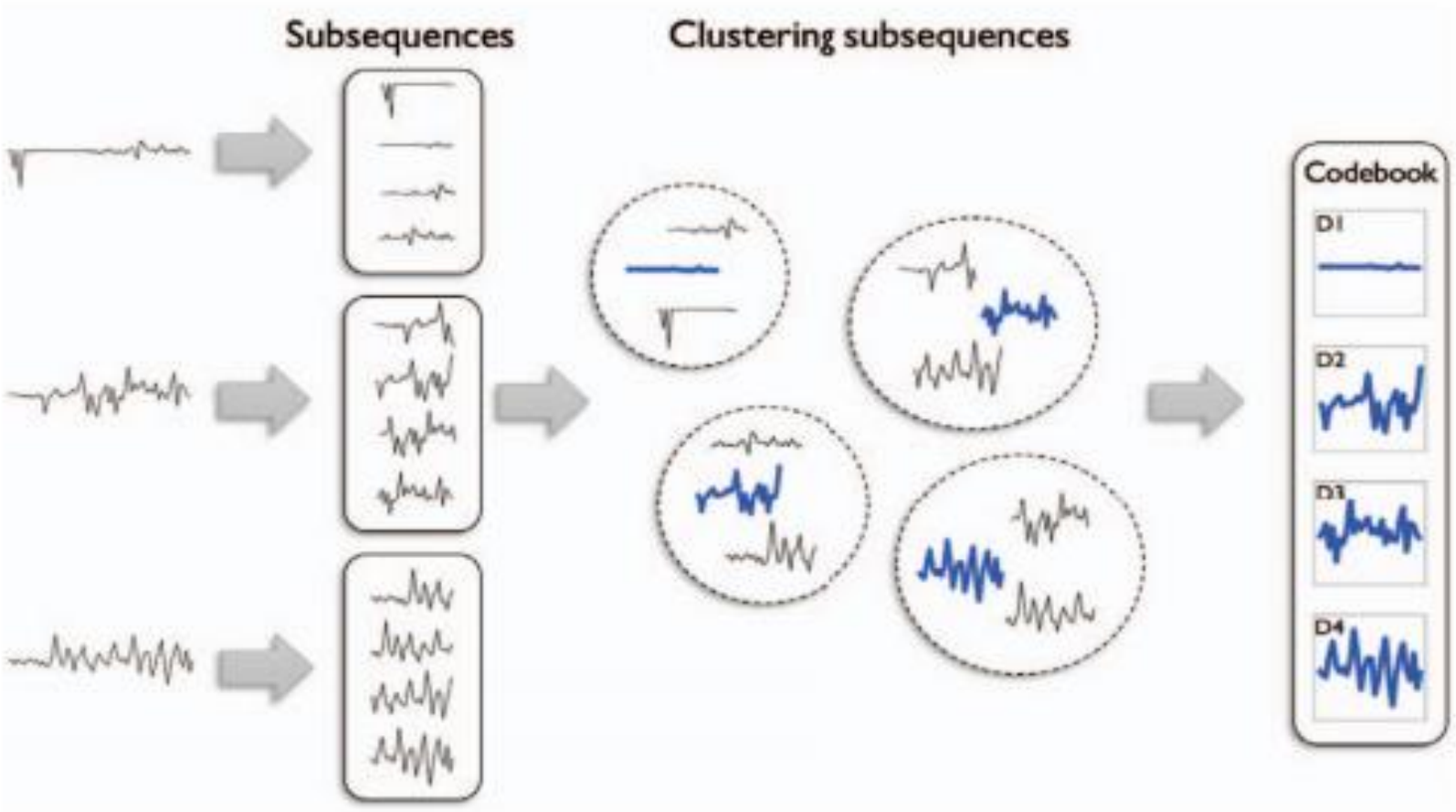


1 Person-1 Day- Frequency 100HZ- S =1sec

8,640,000 subsequences

Feature Engineering- Movement Words Coding Scheme

Vocabulary Generation



Can we use Mobility/Movement Data to Assess Health?

- Can we use mobility parameters or patterns to assess health?
- Would mobility data be useful to adjust or tweak rehabilitation or physical therapy programs?
- Can we use mobility data to determine the right time to discharge patients after hospital procedures?
- Would mobility data be helpful in predicting health hazards before they happen?

Key Applications for connecting mobility and Health:

- Aging Applications
  - Early detection of neurodegenerative diseases
  - Objective evaluation of treatments
- Early childhood development
  - Early and objective detection of developmental disorder
- Depression and Autism Disorder
  - Early and objective detection of depression and autism
  - Objective evaluation of degrees of severity

Dataset: Participants and Protocol (Wrist Data)

- Three phases of data collection (6-months period between each two phases)- One week of data per individual-per week
- Sampling frequency:100
- Mild, moderate, and sever PD (overall mild PD)

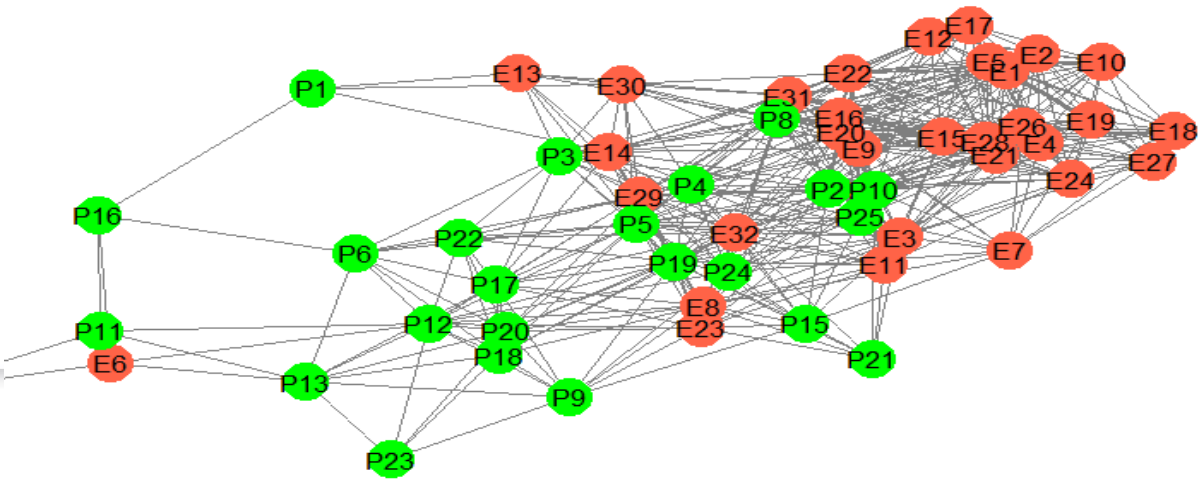
	Healthy young	Healthy elderlies	PD
Number of subjects	3	3	3
Gender (M/F)	2:1	1:2	1:2
Age	23 ± 3.6	65.3 ± 16.2	66 ± 5.0
UPDRS III		------	
H & Y		2.16 ± 0.88	

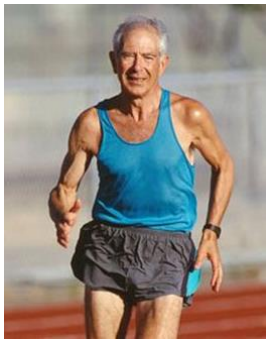
Modeling: Machine Learning

- **Standard Features:**
  - All features (32)
  - First reduced set of features (22)
    - Using Information Gain and Ranker methods
  - Second reduced set of features (8)
    - Using Pearson Correlation coefficient and ANOVA table
  - Third reduced set of features (7)
    - feature sets with one feature less than the optimal number of features
- **Document-of-Words Features:**
  - 10 Features for wrist data and 4 features for ankle data
- **Various Machine Learning Techniques:**
  - SVM, Random Forest, Naïve Bayes, AdaBoost, and bagging
- **Validation:**
  - K-Fold Cross validation
- **Accuracy measures:**
  - F-measure, Precision, Recall

Similarity Network Model – Wrist Data-Word Features

Similarity Network Model for the data from the first phase of wrist dataset- Threshold at **90%-** PD and HE





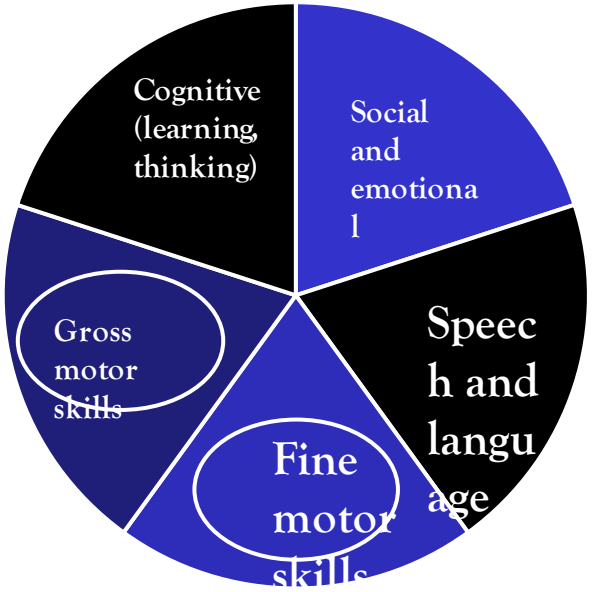
Subject	Gender	Age	MoCA	FoG	FAB	TUG	GDS	H&Y	MFES	Lawton
PD8	Male	69	28	2	39	6.7	0	1	10	8
PD10	Male	71	28	1	39	6.7	1	1	9.3	8
PD1	Male	83	26	8	39	11.2	0	1	8.6	8
PD21	Male	54	25	0	39	9.0	0	1	10	8

Post-operative Nursing Care

- A *post-operative* assessment is very important to a full and speedy *recovery from* any type of *surgery*.
  - a full assessment and an individualized treatment plan based upon the patient's needs and level of function, coupled with clinician expectations

Mobility and Developmental Disorder

- Developmental disorder during early childhood influences motor/mobility
  - Autism
  - Cerebral Palsy
- Certain diseases at any stage in the lifetime may alter mobility
  - Mental disorders - Depression



5 types of skills developed during early childhood

Clinical Diagnosis

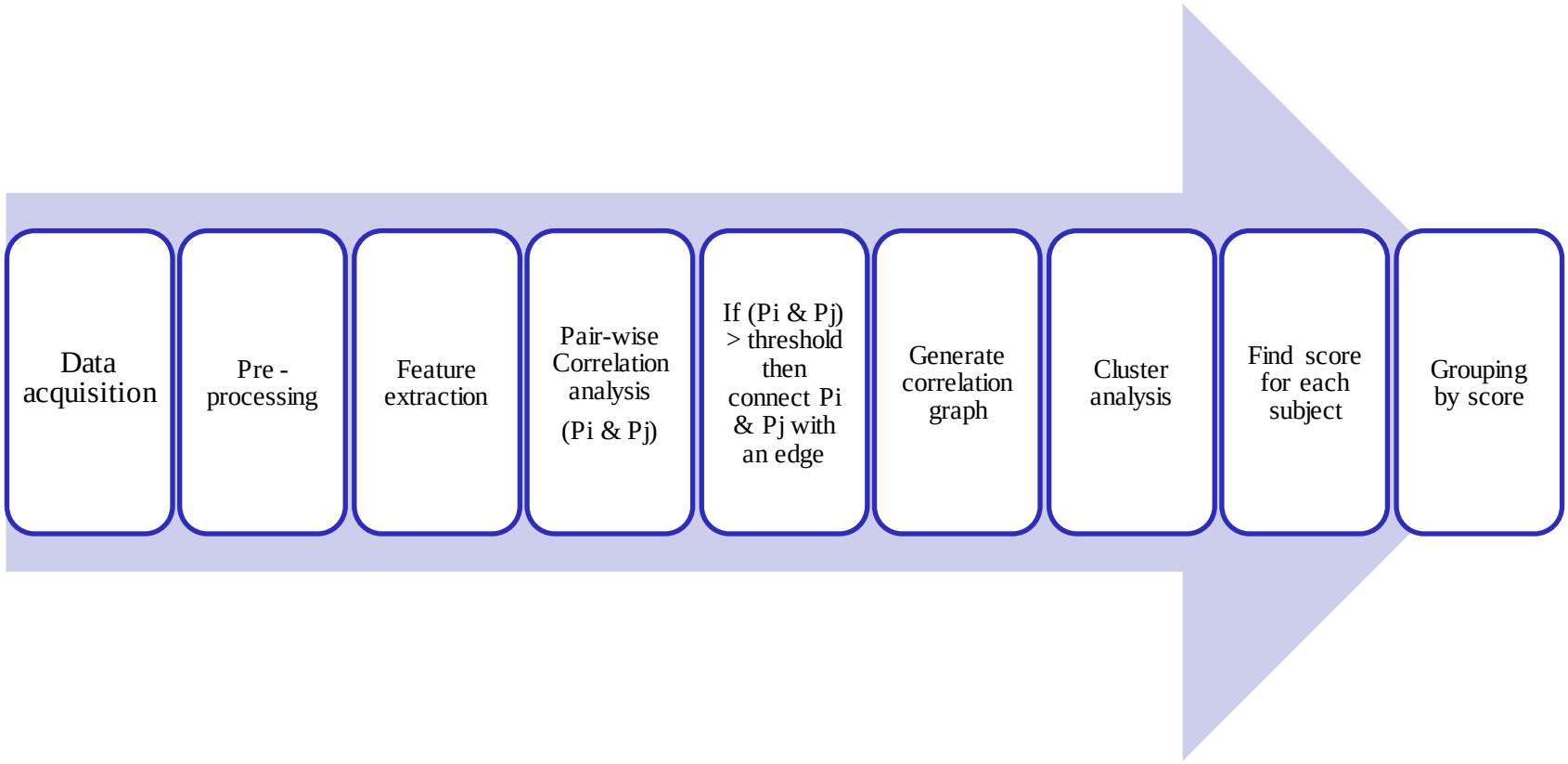
- Developmental disorders (Autism, CP)
- Mental disorders (Depression)
- No pathology test
- Self reporting assessment
- Interview based diagnosis
- Observational scale
  - MADRS score for Depression
  - Pretchl's assessment for CP

Analysis of Depression Episodes

- Depression
  - 280 million people suffering from depression across the world[1]
  - a deep sorrow that lasts for more than two weeks
  - Serious mental disorder
  - Poor performance in work, school
  - Affects quality of life, relationships

World Health Organization https://www.who.int/news-room/fact-sheets/detail/depression

Correlation Network Model



Depression Public Dataset

- Depression dataset
- 55 Subjects- 23 condition (depressed) & 32 control (normal)
- Duration – avg 12.6 days by each subject
- Sensor – ActiWatch accelerometer-based sensor worn on right wrist
- Data - each person activity is recorded in separate csv file with time stamp

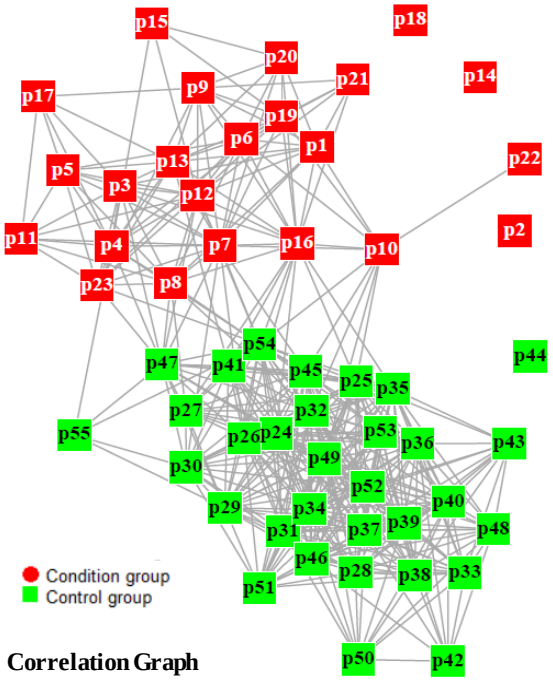
3. Enrique et all, Depression: A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients, In MMSys'18 Proceedings of the 9th ACM on Multimedia Systems Conference, Amsterdam, The Netherlands, June 12 - 15, 2018.

Demographic details

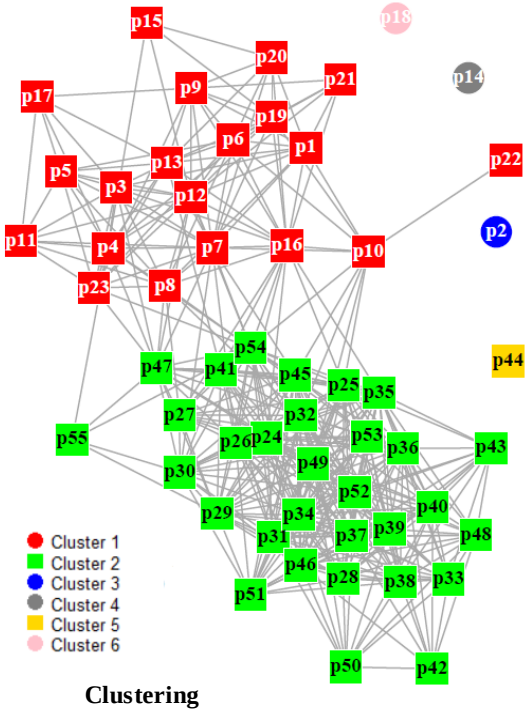
- Demographic details of each patient
  - Number of days monitored by each patient
  - Gender
  - Age
  - Unipolar or bipolar disorder
  - MADRS score at day 1 and day last

number	days	gender	age	afftype	melanch	inpatient	edu	marriage	work	madrs1	madrs2
condition_	11	2	35-39	2	2	2	10-Jun	1	2	19	19
condition_	18	2	40-44	1	2	2	10-Jun	2	2	24	11
condition_	13	1	45-49	2	2	2	10-Jun	2	2	24	25
condition_	13	2	25-29	2	2	2	15-Nov	1	1	20	16
condition_	13	2	50-54	2	2	2	15-Nov	2	2	26	26
condition_	7	1	35-39	2	2	2	10-Jun	1	2	18	15
condition_	11	1	20-24	1	NA	2	15-Nov	2	1	24	25
condition_	5	2	25-29	2	NA	2	15-Nov	1	2	20	16

Correlation graph and clustering

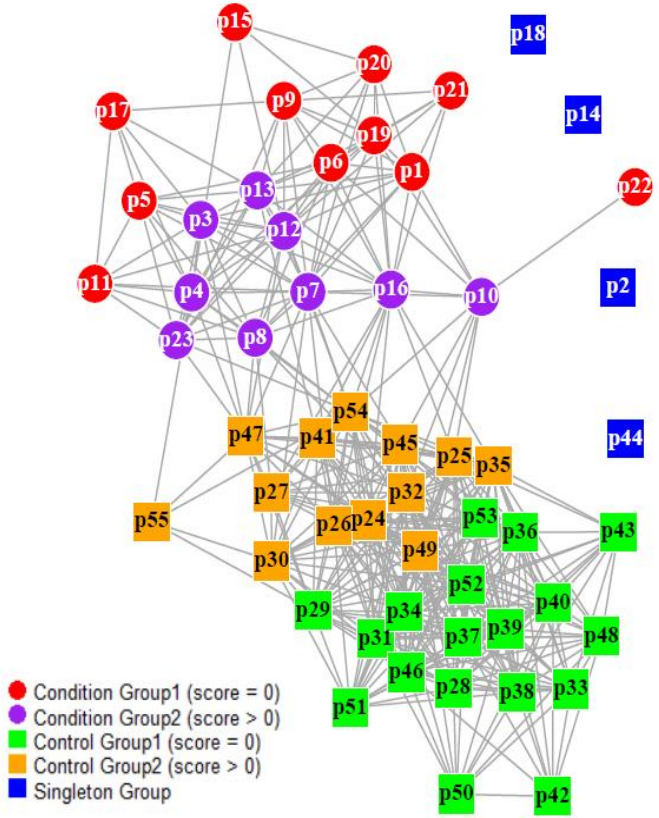


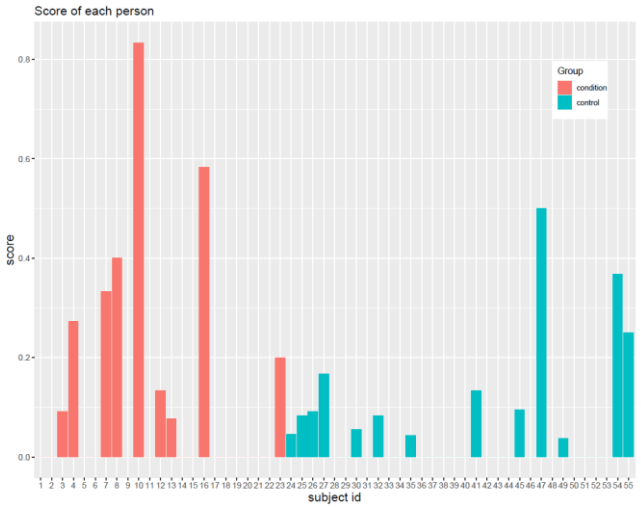
Correlation Graph



Clustering

Groups by score





Group no	Original group	score	Color code
Group 1	Condition	0	Red
Group 2	Condition	>0	Purple
Group 3	Control	>0	Orange
Group 4	Control	0	Green

Summary

- Ability to identify groups with different health conditions using mobility as a feature by utilizing population analysis-based Correlation network
- Identifying group of persons with similar mobility features without using class labels
- Mobility as a quantitative method for early diagnosis of developmental disorders
- YES – we can use wireless sensors and mobility data for health assessment.

CS3: Microbiomes as Signals

The microbiome is the community of microorganisms (such as fungi, bacteria and viruses) that exists in a particular environment.

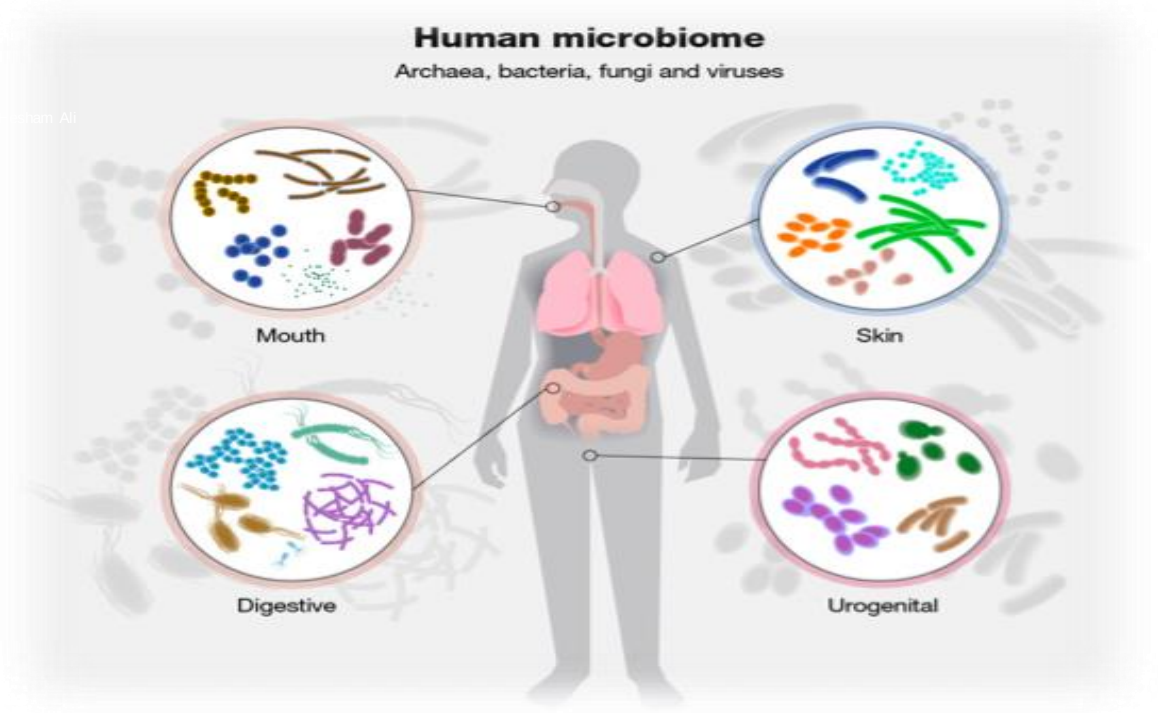


Figure by National Human Genome Research Institute

Microbiome-based interventions and future of healthcare

- Success in treating Clostridioides difficile infections of the gut through fecal microbiota transplantation (FMT) procedures.
  - FMT are now standard care for recurrent C.difficile infections
- Hopefully more targeted microbiome-based interventions will be available to treat complex diseases.
- Can we identify specific microbial signatures associated with host health phenotype?

Problem statement

Can we identify specific microbial signatures associated with host health phenotype?

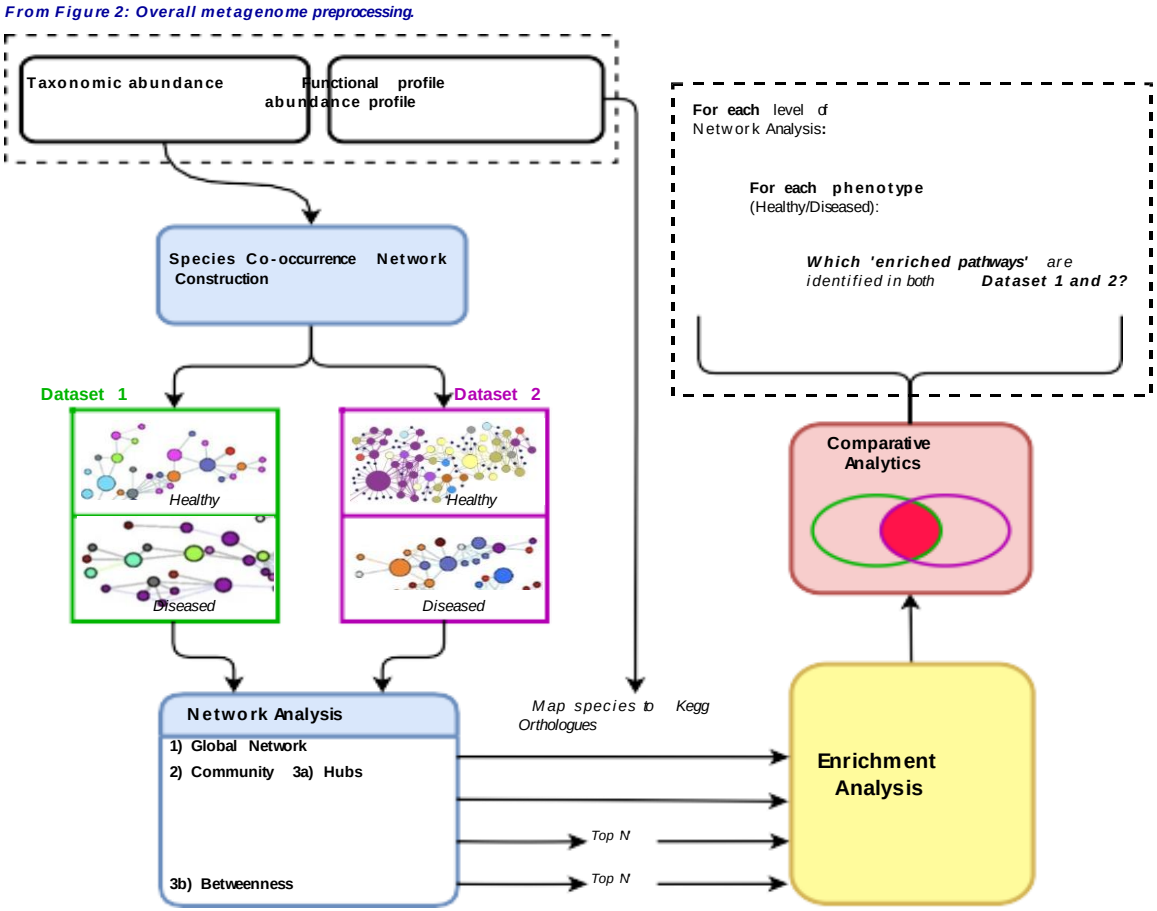
- Various computational tools exist for identification of differentially abundant microbes from abundance data.
- Microbes engaged with one another through various interactions to form ecological communities.
- Several computational tools offers functionality for constructing and analyzing microbiome association networks (i.e., NetCoMi and iNAP).
- There are few computational approaches for the identification of signature pathways based on co-occurring microbial species.

Proposed Approach

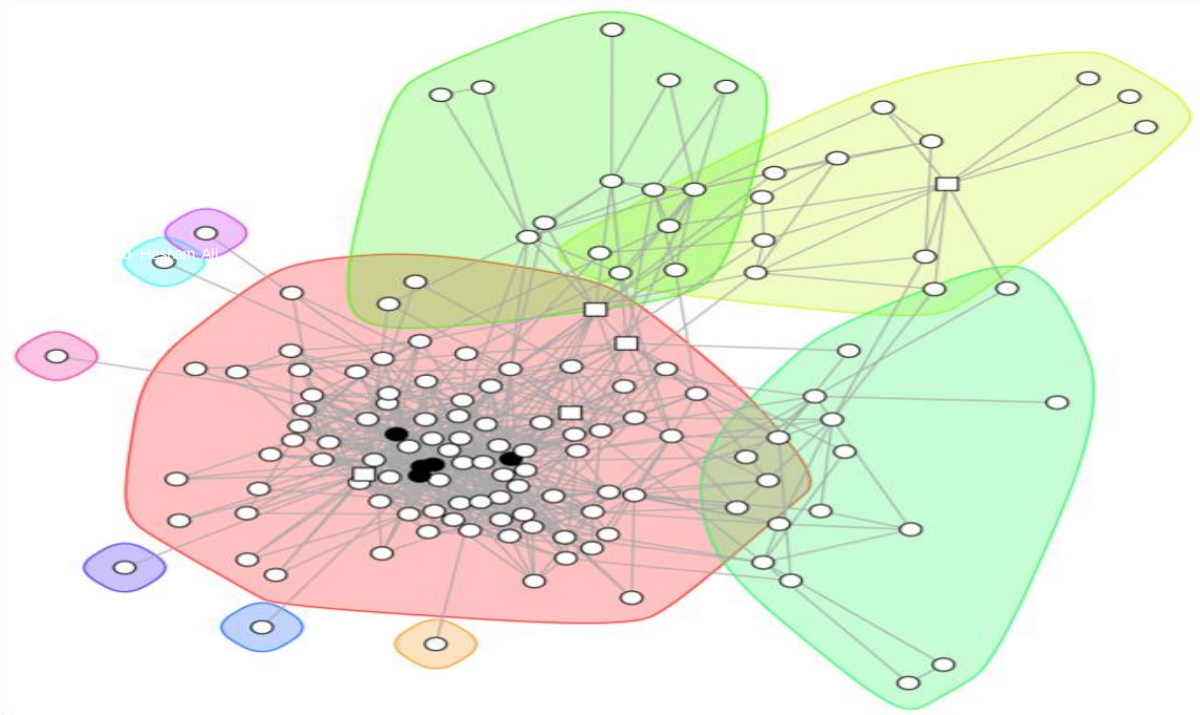
Develop a computational approach to identify microbiome-based functional enrichment patterns associated with host health:

- To characterize species-level microbial co-occurring communities with enrichment of their pathways that are associated with host health.
- To capture the phenotype-relevant pathways through multiple granularity network analysis from the metagenome abundance datasets.
- To provide microbiome researchers with easy-to-implement analysis tool (https://github.com/skimicrobe/GutNetMining)

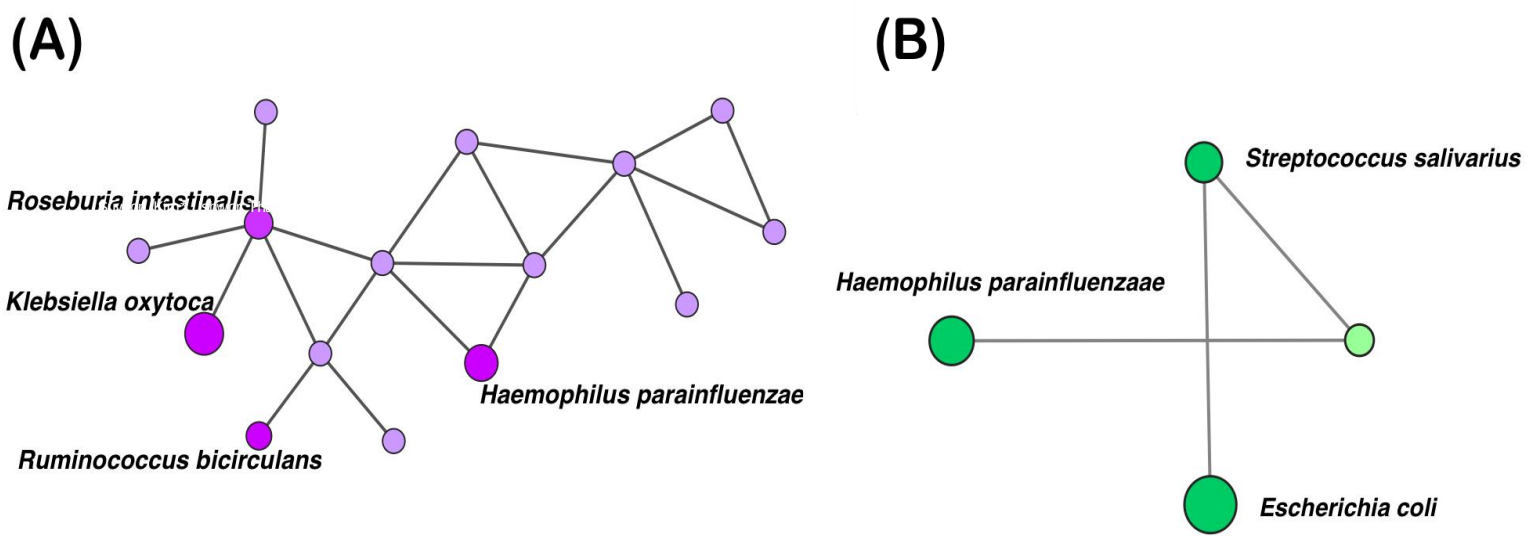
A computational pipeline for mining signature pathways with host health conditions



Species co-occurrence network

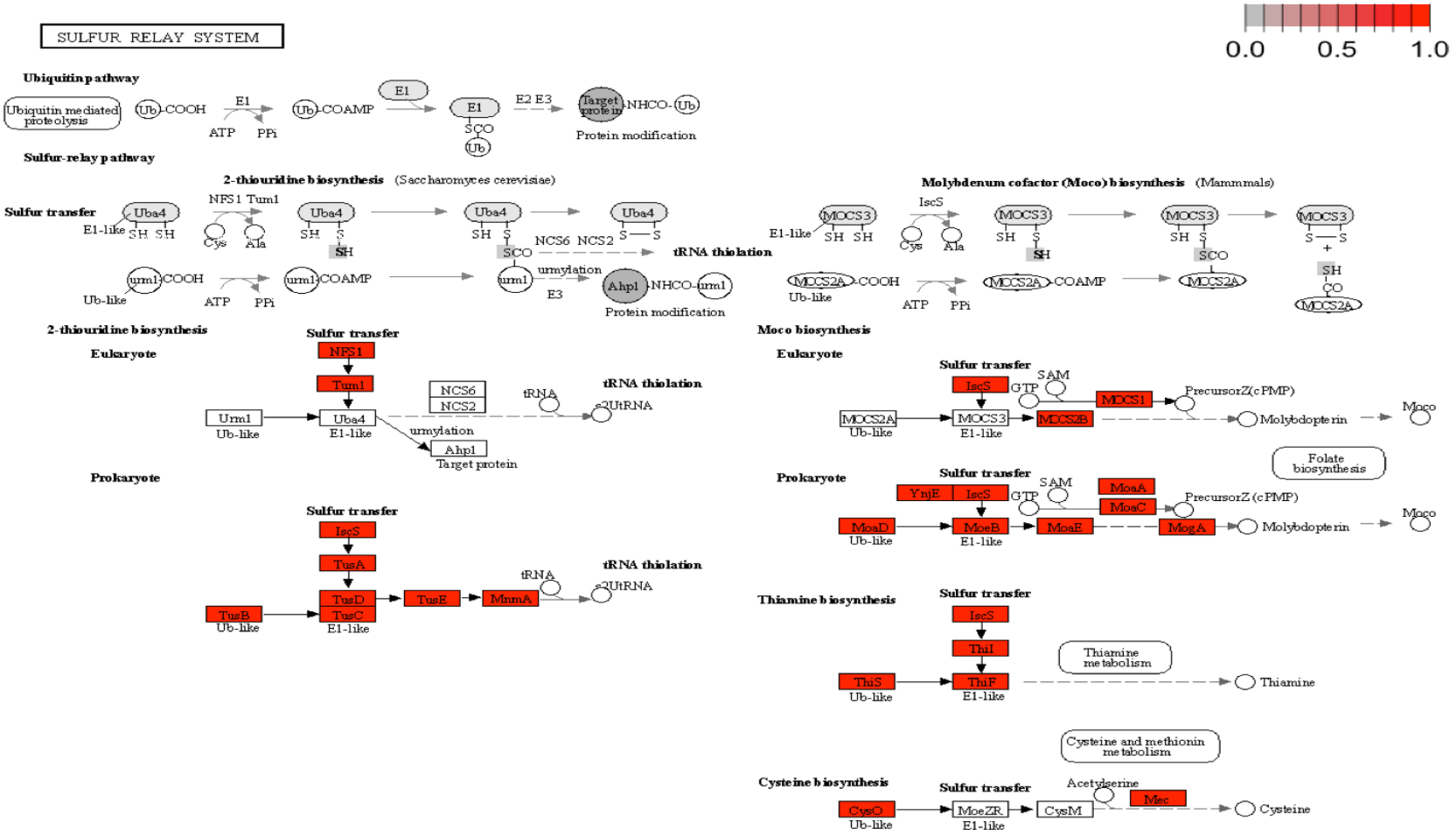


Sulfur relay system pathway for each microbial community in CD and UC group



- Identified clusters enriched in sulfur relay pathways in CD and UC.
- The size of nodes corresponds to the number of KOs each bacteria contributes in the sulfur relay pathway.

Sulfur relay system pathway with KOs (gene) from co-occurring species in CD group



Literature-based validation (Pathways for CD)

Global	Community	Key-Elements
ABC transporters (map02010)	ABC transporters (map02010) Two-component system (map02020) Sulfur relay system (map04122)	Biosynthesis of cofactors (map01240) Cell-cycle - Caulobacter (map04112) Homologous recombination (map03440) Mismatch repair (map03430) Protein export (map03060) Pyrimidine metabolism (map00240) Streptomycin biosynthesis (map00521)

Literature-based validation (Pathways for UC)

Global	Community	Key-Elements
Cysteine and methionine metabolism (map00270)	Sulfur relay system (map04122)	Valine, leucine and isoleucine biosynthesis (map00290) Cell-cycle - Caulobacter (map04112) Drug metabolism - other enzymes (map00983) Homologous recombination (map03440) Protein export (map03060) Lysine biosynthesis (map00300) Mismatch repair (map03430) One carbon pool by folate (map00670) Peptidoglycan biosynthesis (map00550) Protein export (map03060) 2-Oxocarboxylic acid metabolism (map01210)

Summary

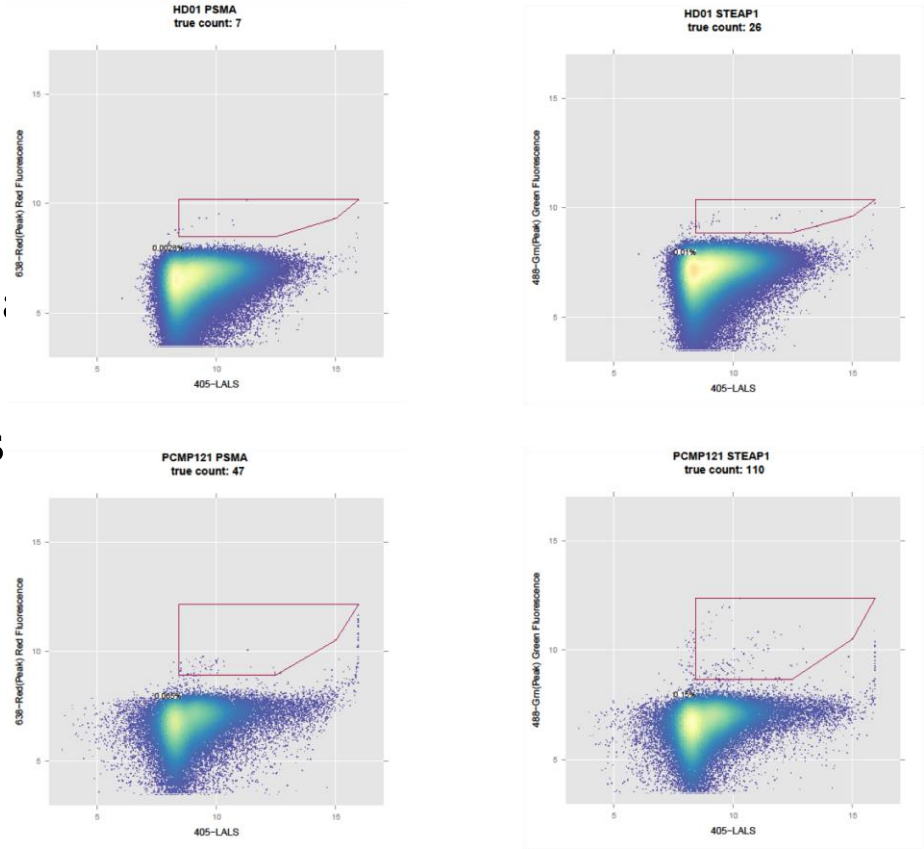
- Demonstrated the viability of identifying the phenotype-relevant pathways through multi-granularity network analysis from the metagenome abundance datasets.
- Developed an easy-to-implement computational approach for microbiome researchers.
- We have general-purpose pre-processing tools to obtain functional/taxonomic abundances from multiple datasets.
- The obtained results are important since they point to a deeper understanding of the roles in the microbial community.
- Here, we showcased the utility of our pipeline using IBD datasets, the tool can be applied to analyze metagenomic data from any conditions.

CS4: Extracellular Vesicles as Signals

- Extracellular vesicles (EVs) are submicron particles (< 1 μm) that may contain proteins (peptides), fragments of DNA, mRNA, non-coding-RNAs and metabolites.
- EVs are released by all cell types including tumor cells in body fluids such as blood and urine
- As next-generation biomarkers, circulating EVs may be useful in improving diagnosis of cancer, patient stratification and disease recurrence prediction.

How to measure EVs?

- Flow cytometry (FCM) is a commonly used technique for single EV analysis.
  - allows for high-speed detection of millions of particles (≥100 nm) within a few minutes from a small volume.
- Provides light-scatter detection of particles with fluorescence measurements for specific markers.
- The light-scatter and fluorescence are saved in a table like FCS files (shown as scatter plots).
- The gates capture the EVs positive for the markers (PSMA and STEAP1).

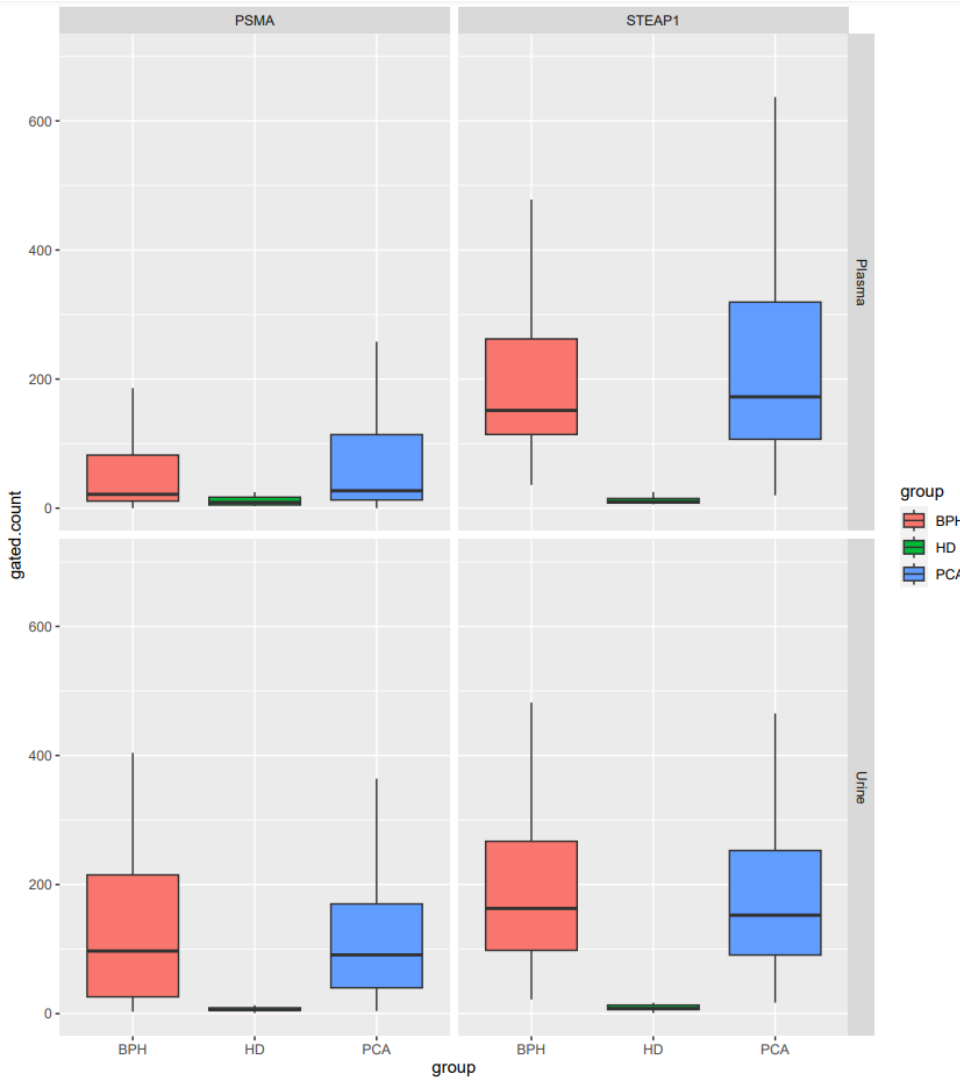


Automated quantification of EVs from healthy donors (top) and cancer patients (bottom)

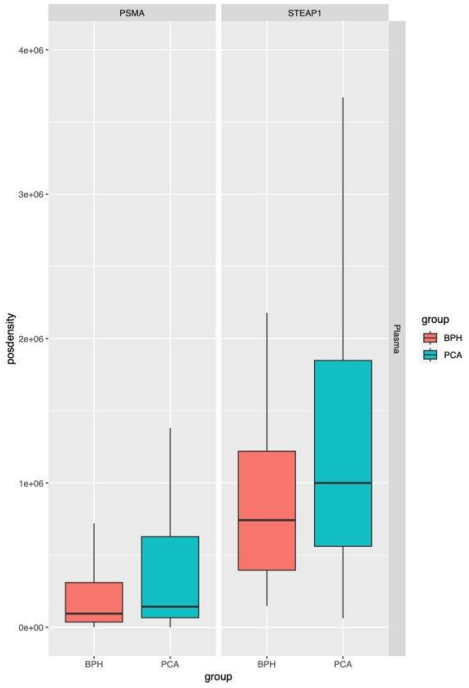
The scattering of the laser beam corresponds to the size and color of the particle. The size is plotted on X axis of the scatter plot and Y axis is the fluorescence of a particular wavelength.

EVs counts as a biomarker?

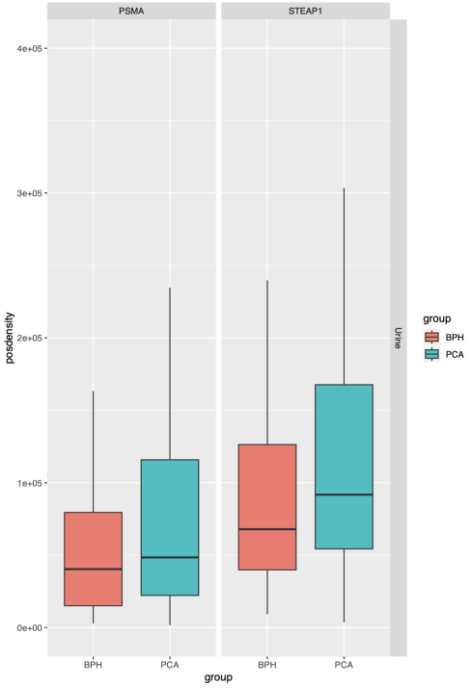
- Elevated levels of PSMA positive and STEAP1 positive EV counts in samples from Prostate Cancer (PCa) patients and benign hyperplasia (BPH) group as compared to healthy donors (in middle).
- Top panel Blood (Plasma)
- Bottom panel Urine
- Left = PSMA +ve EV counts
- Right = STEAP1 +ve EV counts



Or is it EV density?



- With the EV counts divided by the size of Prostate (EV density), the PCA (cancer group) has higher levels of EV density than the benign group.
  - With both markers and in both sample sources (blood and urine)



Sciences at Crossroads

- Many Scientific disciplines are now at crossroads
- The proper penetration of IT represent tremendous challenges and great opportunities
- The importance of interdisciplinary approach and knowledge integration to problem solving
- The need for in-depth analysis and problem solving rather than the surface-level approaches
- Revolution is data collection requires a revolution in data Analytics – Complex Data demand Complex Methods

Acknowledgments

- UNO Bioinformatics Group

  Kiran Bastola
  Sanjukta Bhoomwick
  Kate Cooper
  Dario Ghersi
  Suyeon Kim
  Elham Rastegari
  Ishwor Thapa
  Ling Zhang

- UNO Big Data Group

  Prasad Chetti
  Zahra Hatami
  Rama Krishna Thelagathoti
  Saiteja Malisetty

- Domain Experts

  Biomedical Researchers
  Civil Engineering Researchers

- Funding Sources

  NIH
  NSF
  Nebraska Research Initiative

- Former Group Members

  Alexander Churbanov
  Xutao Deng
  Huiming Geng
  Xiaolu Huang
  Daniel Quest
  Julia Warnke-Sommer
  Sean West