Biomedical Data Mining with Matrix Models

SDM 2016 Tutorial Part II

May 5, 2016

Ping Zhang

Center for Computational Health

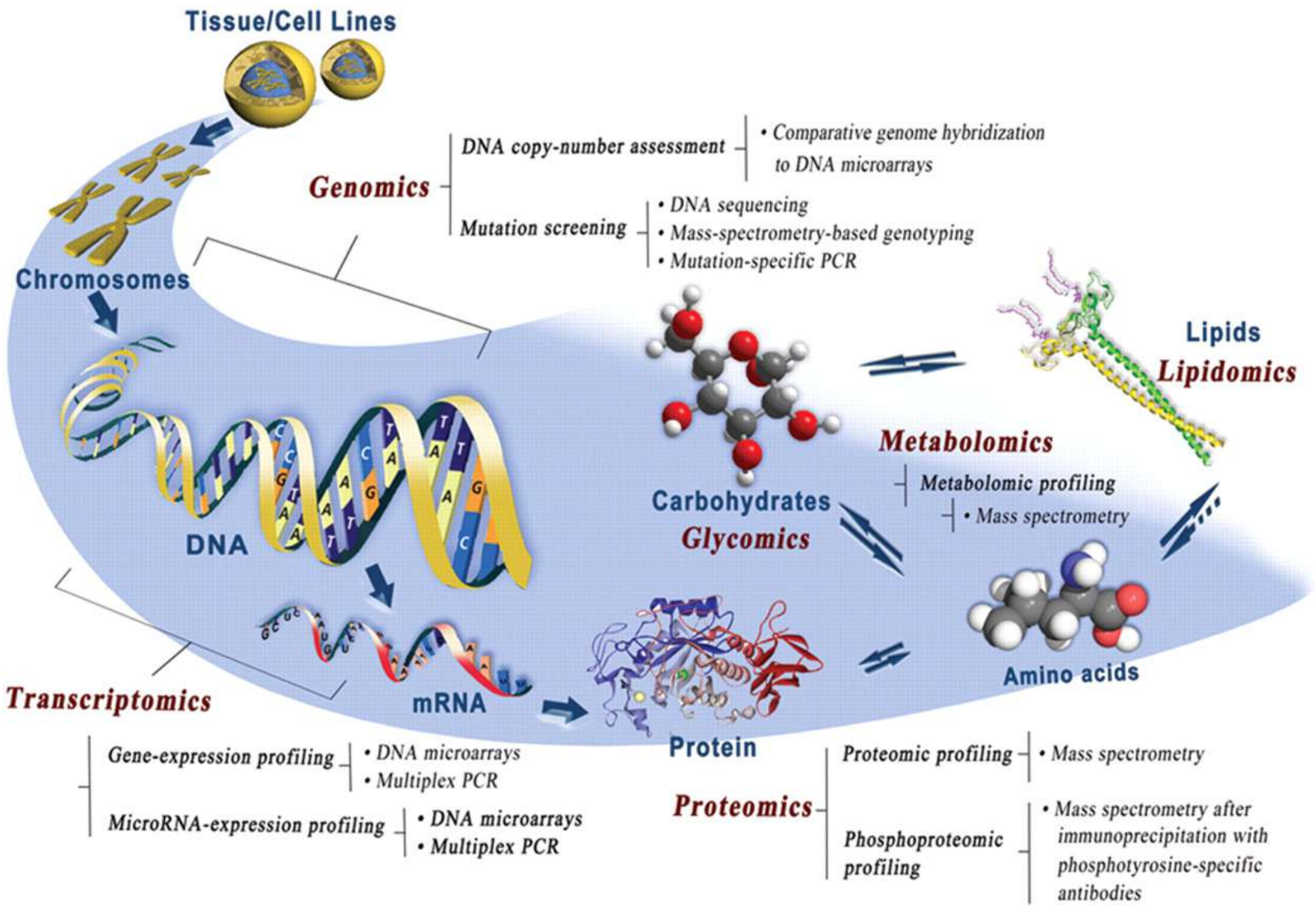
IBM T.J. Watson Research Center

pzhang@us.ibm.com

Recent Applications in Biomedicine

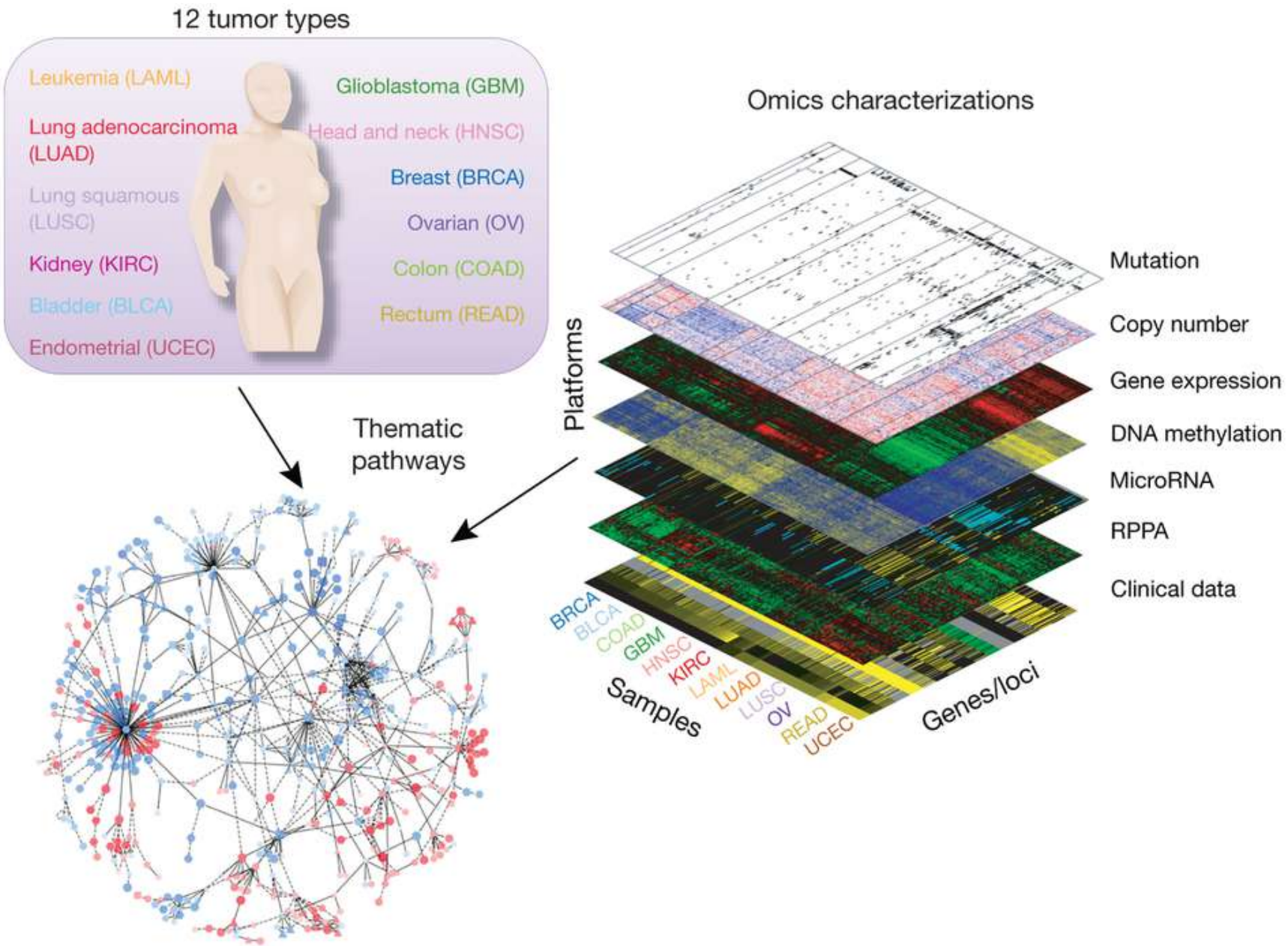
- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Nonnegative Matrix Tri-Factorization and Patient-Specific Data Fusion
- Tensor Factorization and Patient Phenotyping

Omics technologies in biomedicine



R. Wu, et al. Novel Molecular Events in Oral Carcinogenesis via Integrative Approaches. *JDR*, 90(5):561-572, 2010.

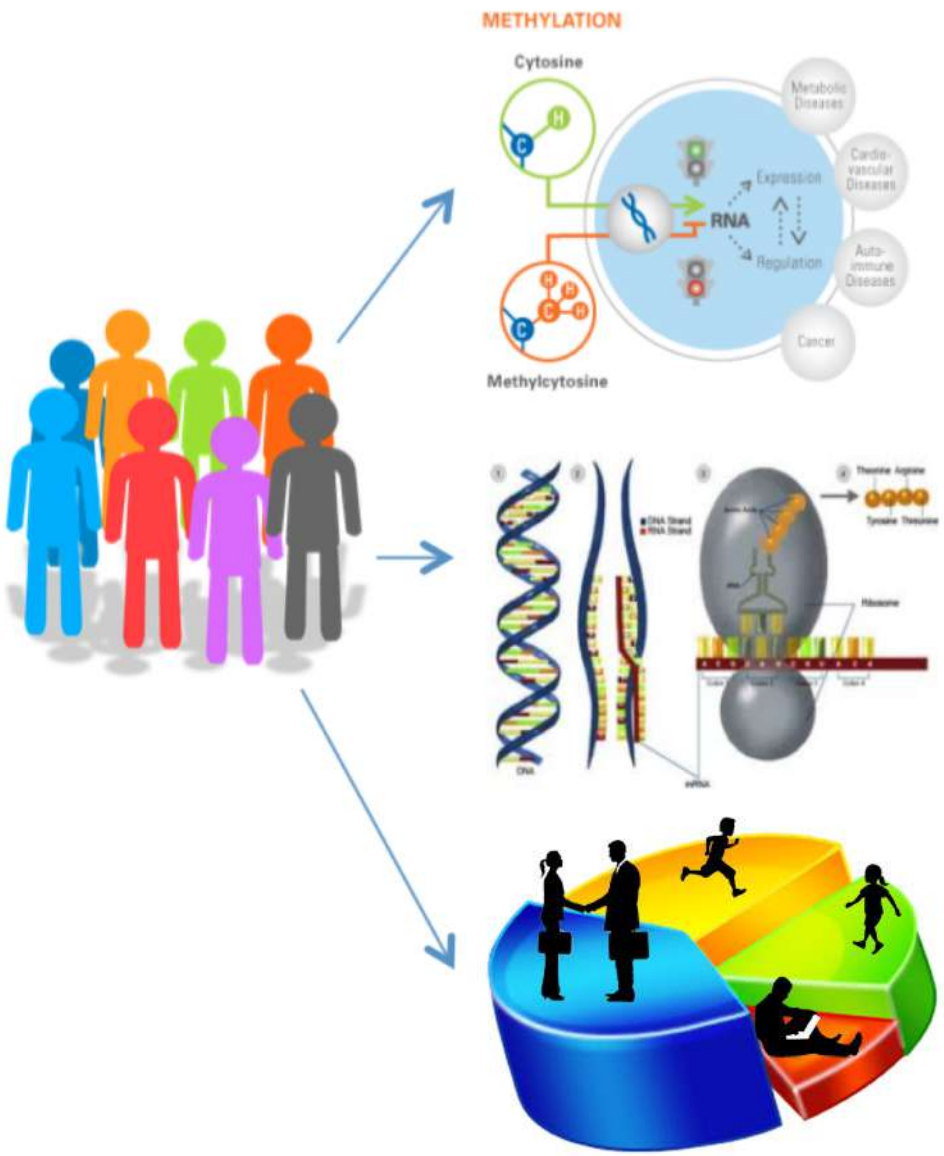
The Cancer Genome Atlas Pan-Cancer analysis project



The Cancer Genome Atlas Research Network, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45:1113-1120, 2013.

Patient similarity networks

How to combine different networks?



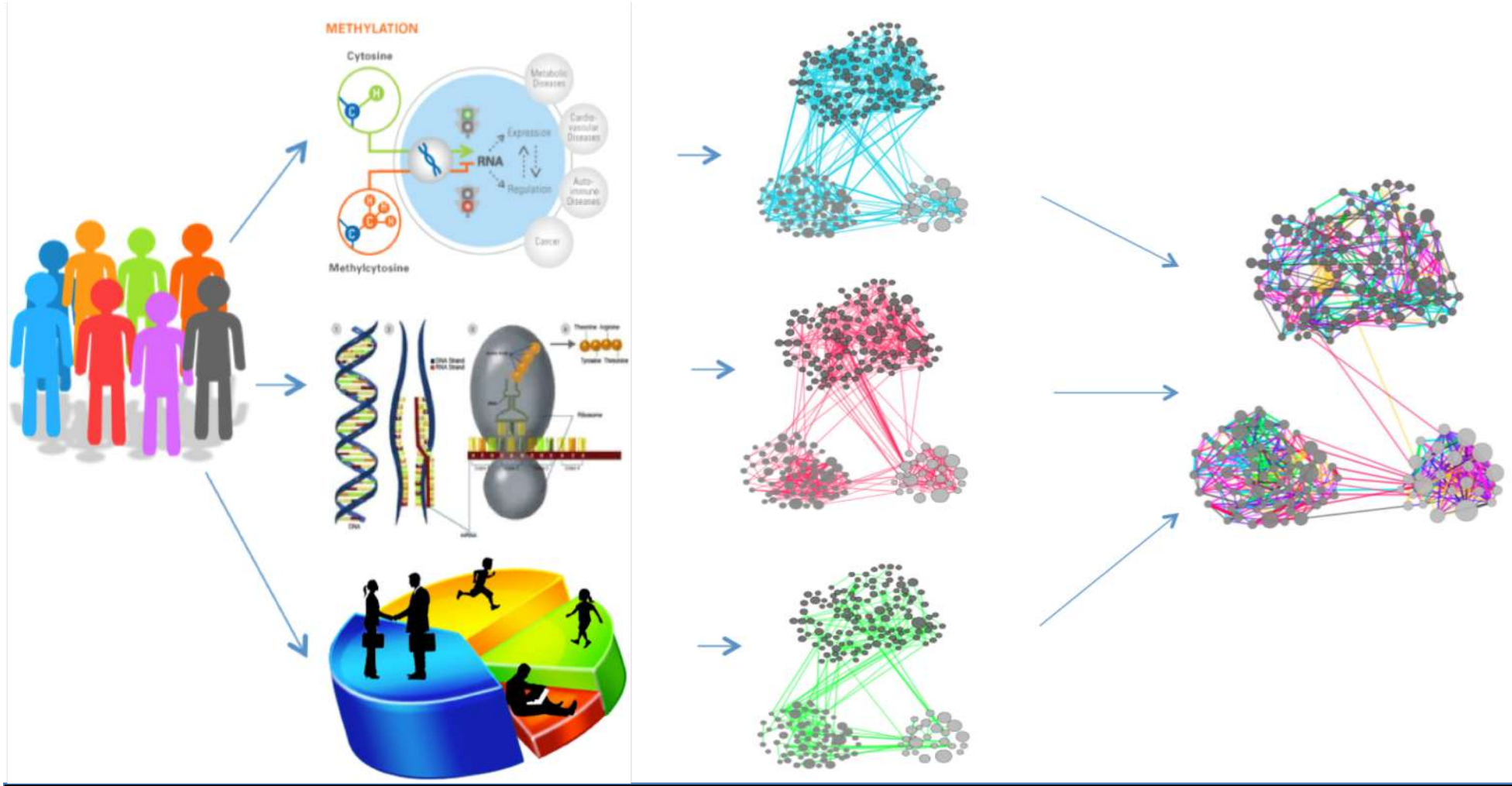
Issues:

- Large number of measurements, small sample sizes (p>>n)
- Need to integrate common and complementary information
- Not all measurements can be normalized and mapped to the same unit

Similarity network fusion

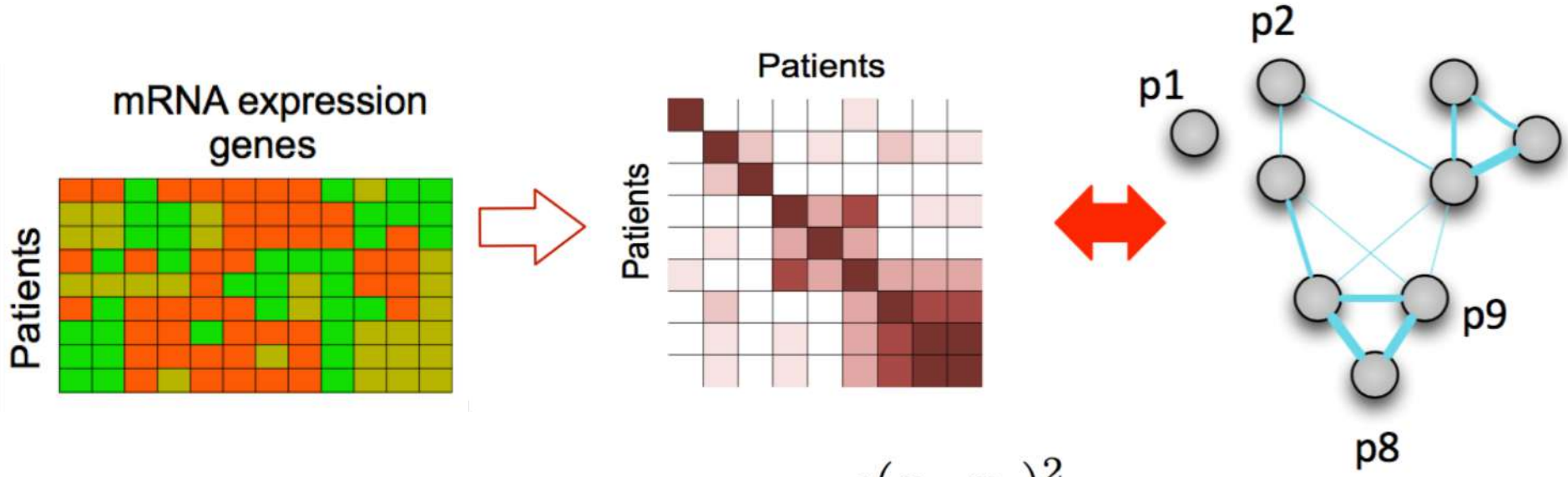
Step 1. Construct a similarity network for each data source

Step 2. Integrate networks using data fusion method



Wang B, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11:333-337, 2014.

Construct similarity networks (1)



Patient similarity: $$W(i,j) = exp(\frac{\rho(x_i, x_j)^2}{\eta \xi_{ij}^2})$$

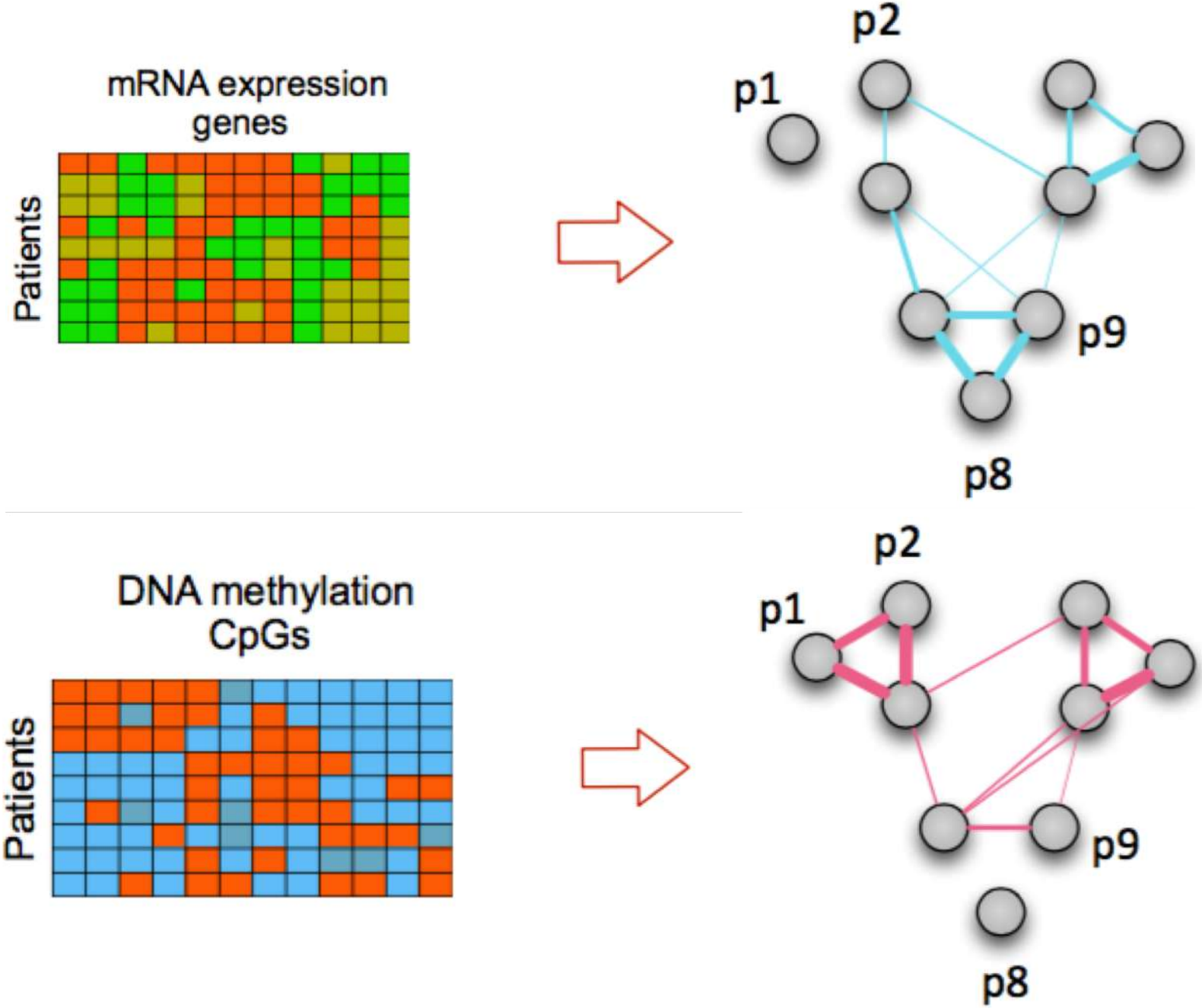
Adjacency matrix: $$P(i,j) = \frac{W(i,j)}{\sum_{k \in V} W(i,k)}$$

Sparsification

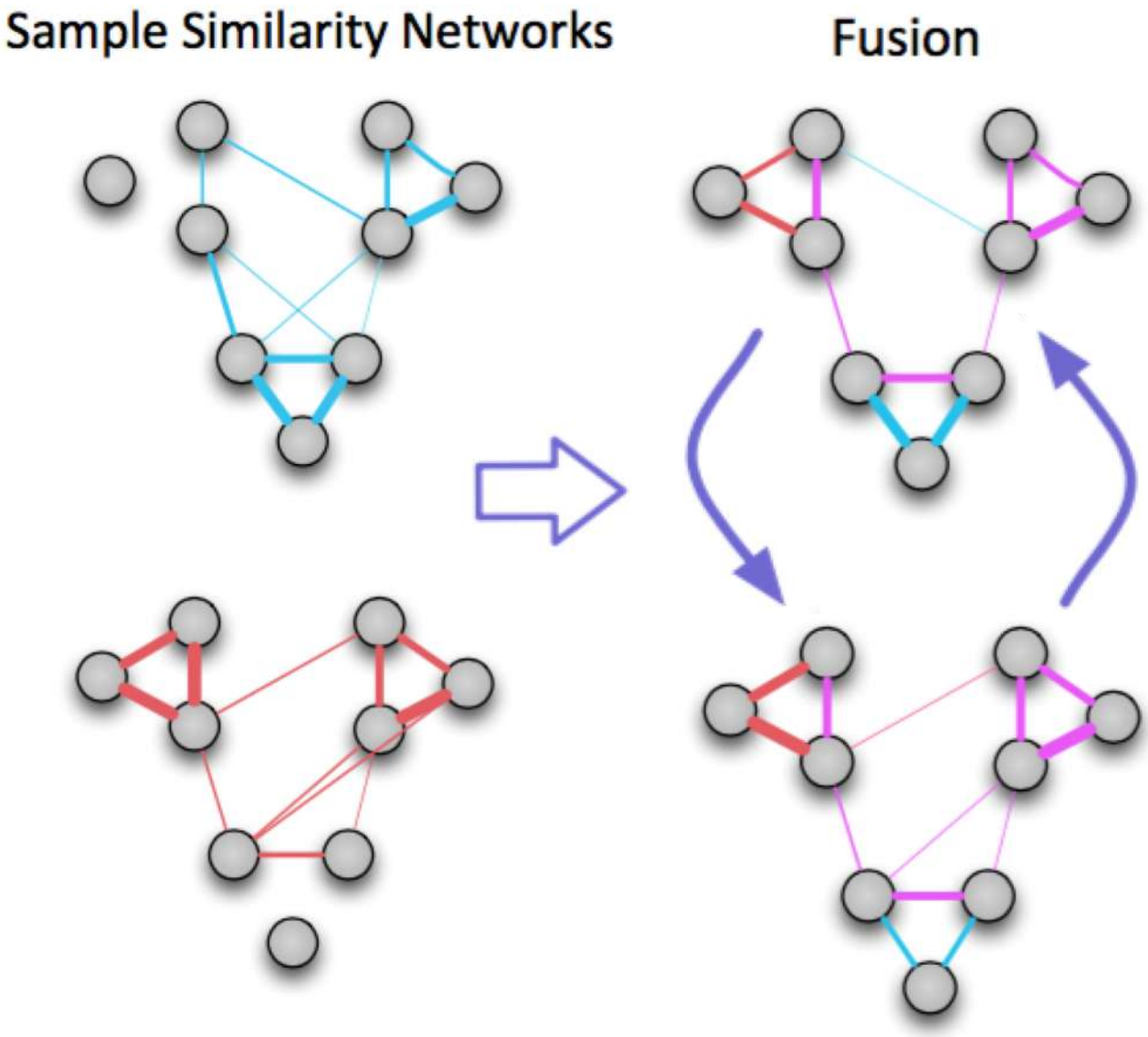
1) $$\mathcal{W}(i,j) = \begin{cases} W(i,j) \text{ if } x_j \in KNN(x_i) \\ 0 \text{ otherwise} \end{cases}$$

2) $$\mathbf{S}(i,j) = \frac{\mathcal{W}(i,j)}{\sum_{x_k \in KNN(x_i)} \mathcal{W}(i,k)}$$

Construct similarity networks (2)



Combine networks (1)



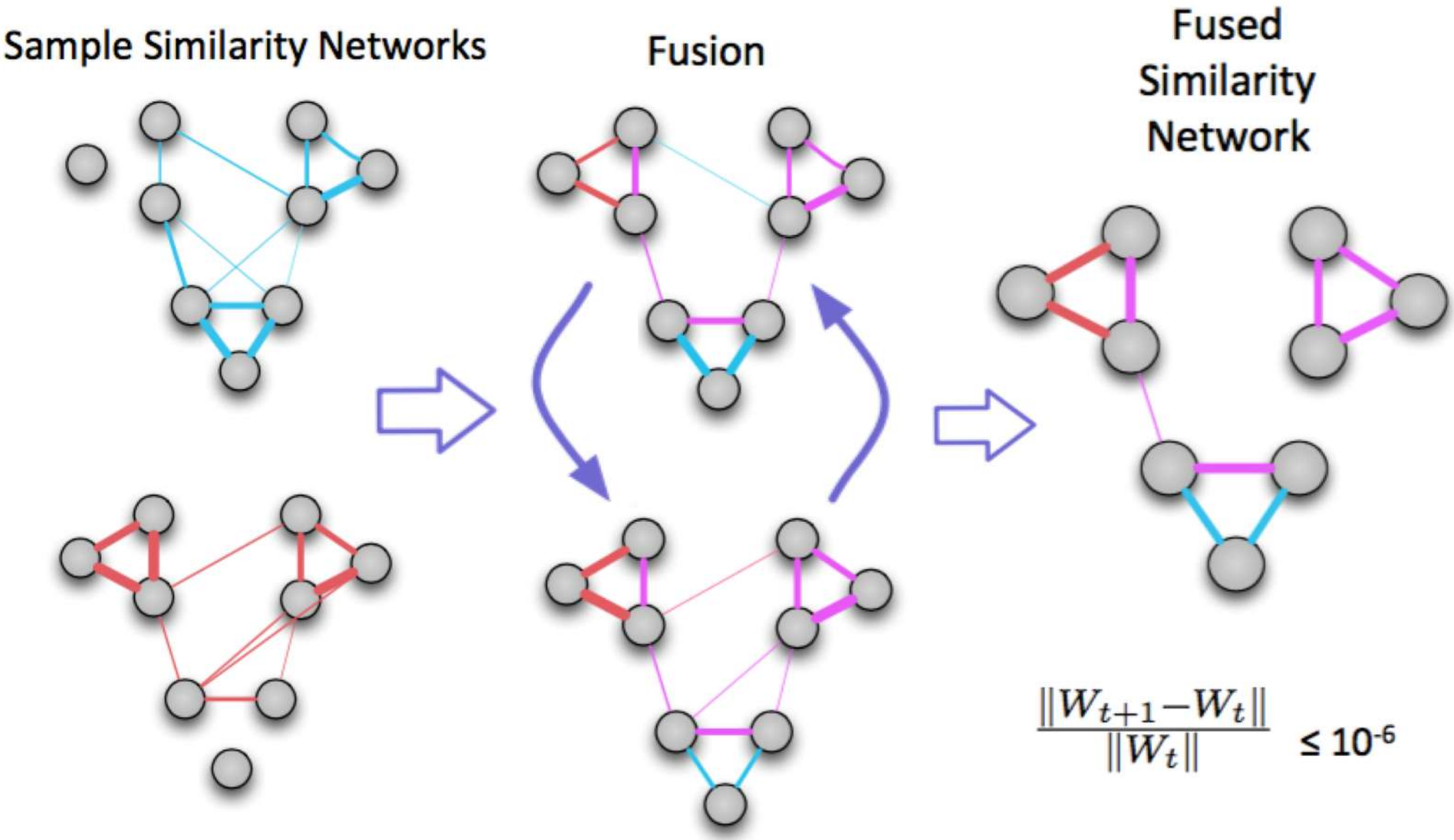
$$\mathbf{P}_{t+1}^{(1)} = \mathbf{S}^{(1)} \times \mathbf{P}_{t}^{(2)} \times (\mathbf{S}^{(1)})^{T}$$

$$\mathbf{P}_{t+1}^{(2)} = \mathbf{S}^{(2)} \times \mathbf{P}_{t}^{(1)} \times (\mathbf{S}^{(2)})^{T}$$

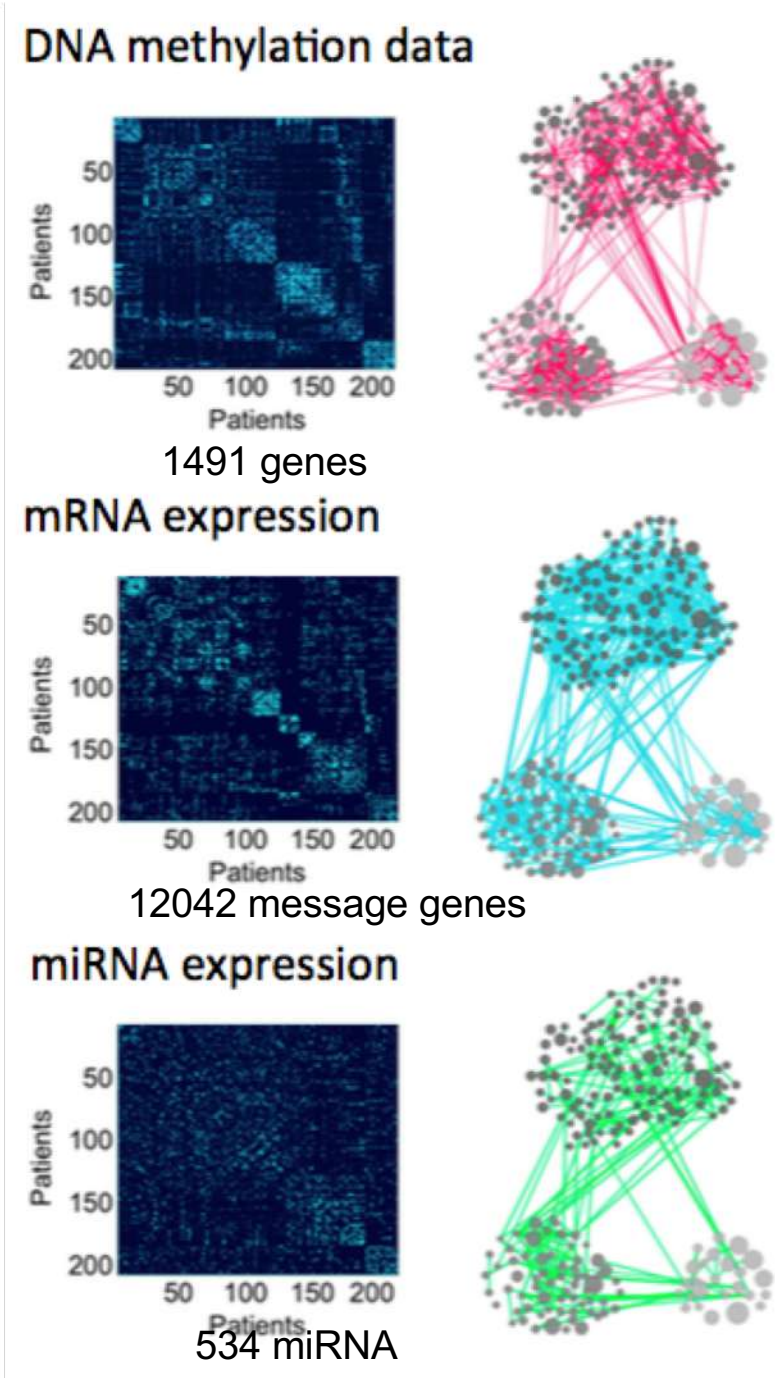
Can also be extended to more than 2 data types

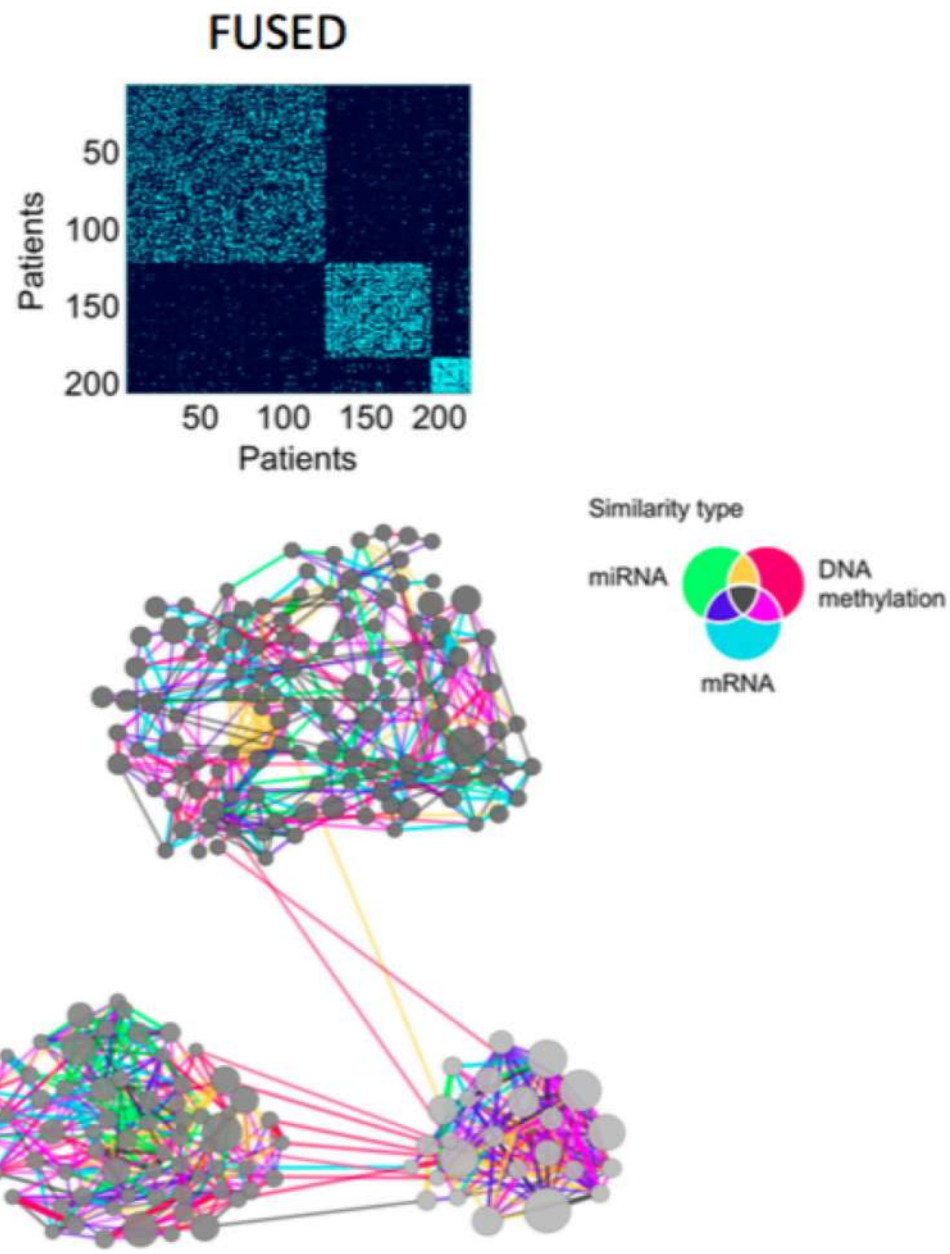
Patient Patient similarity: mRNA-based DNA Methylation-based Supported by all data

Combine networks (2)

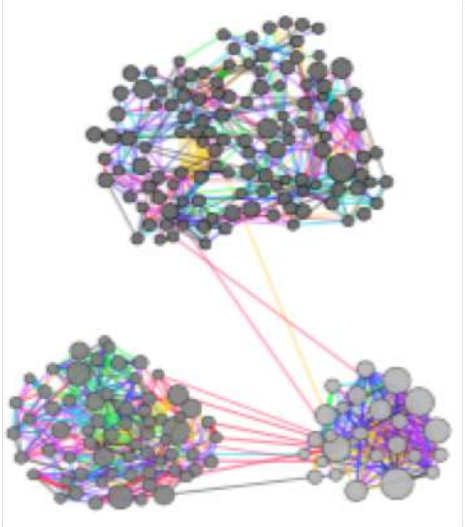


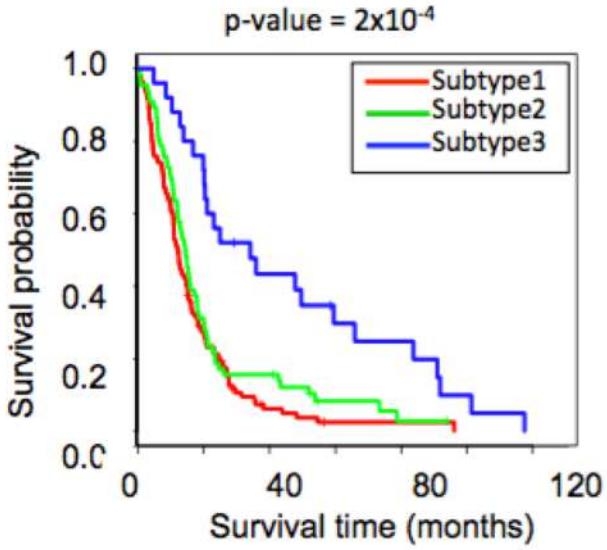
Case study: glioblastoma multiforme (GBM)

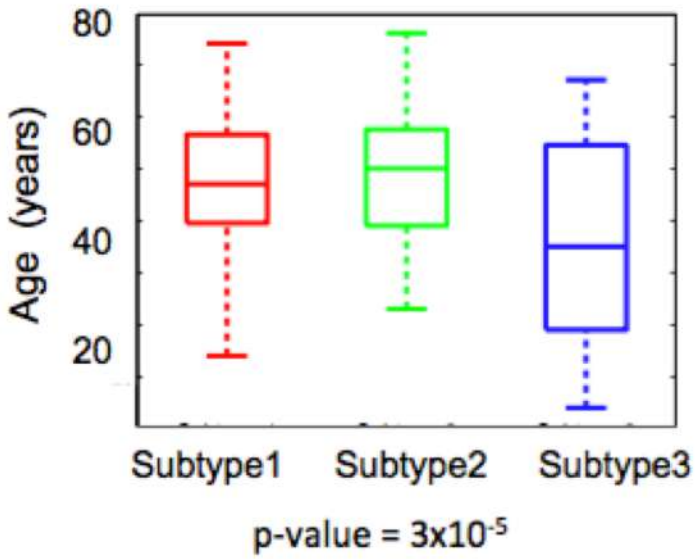


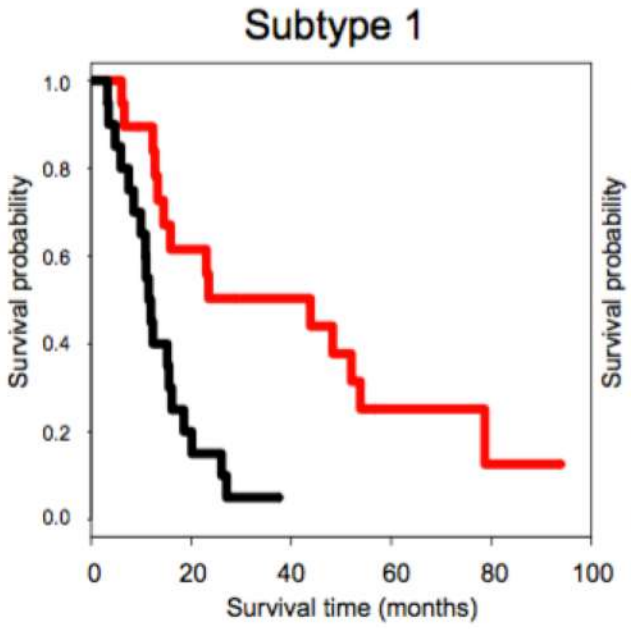


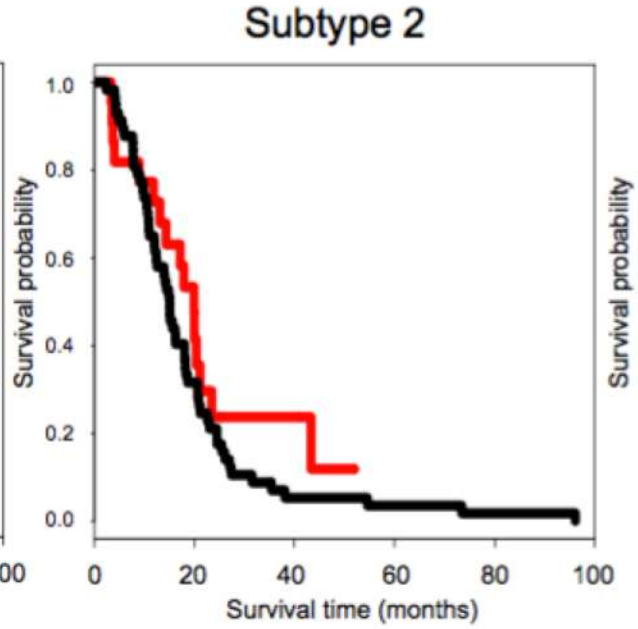
Clinical properties of the subtypes

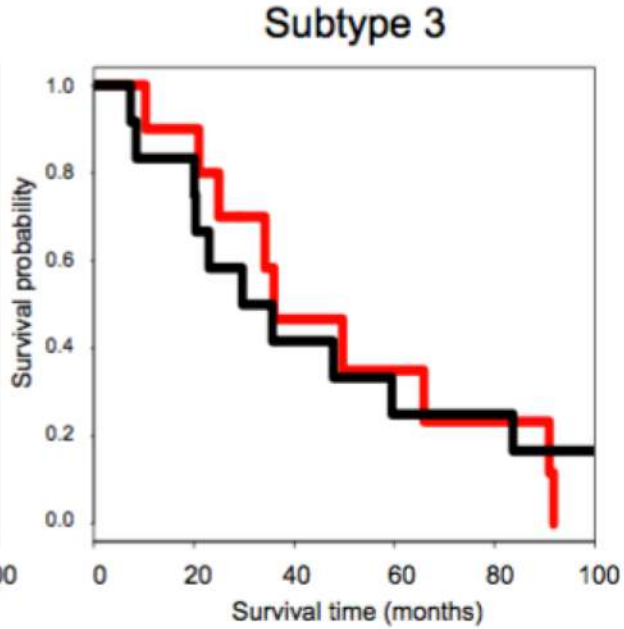




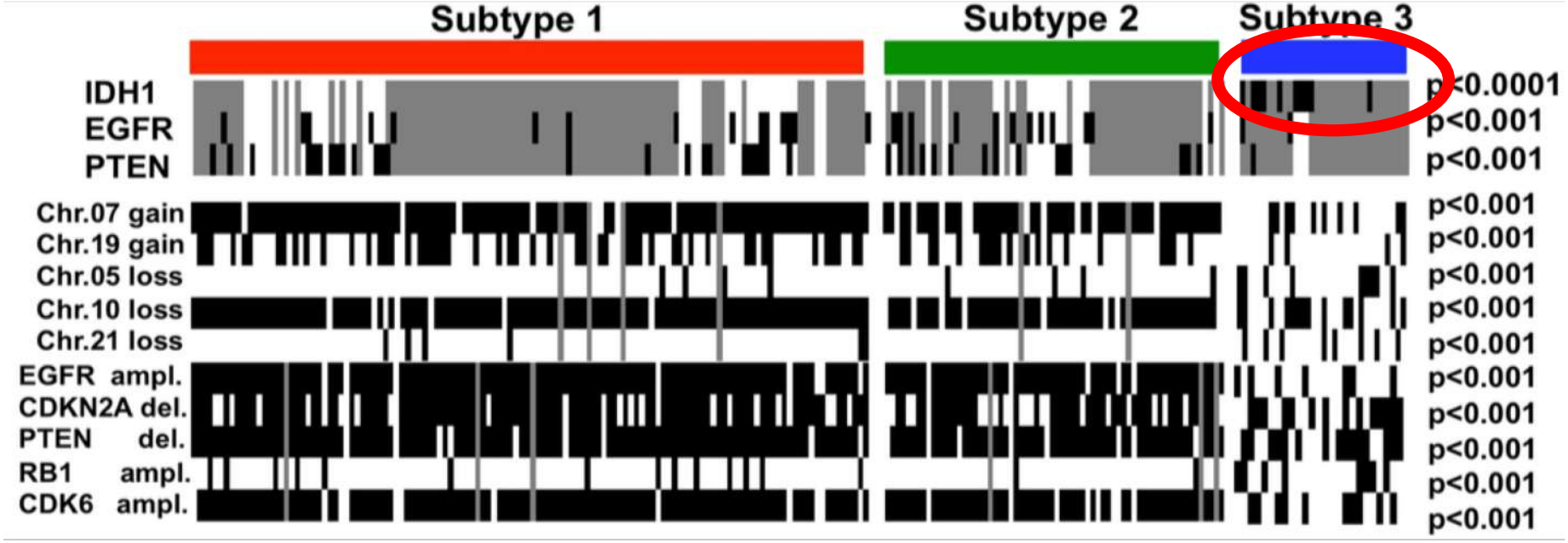


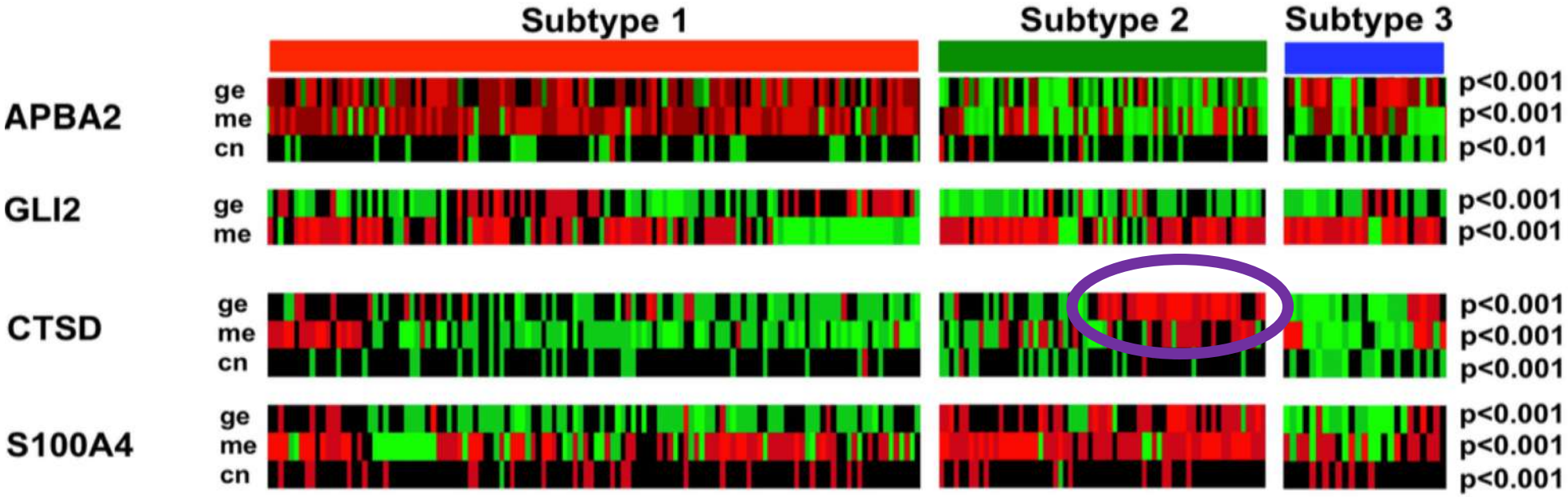




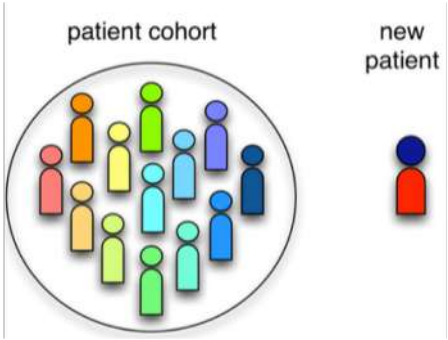


Biological characterization of the subtypes

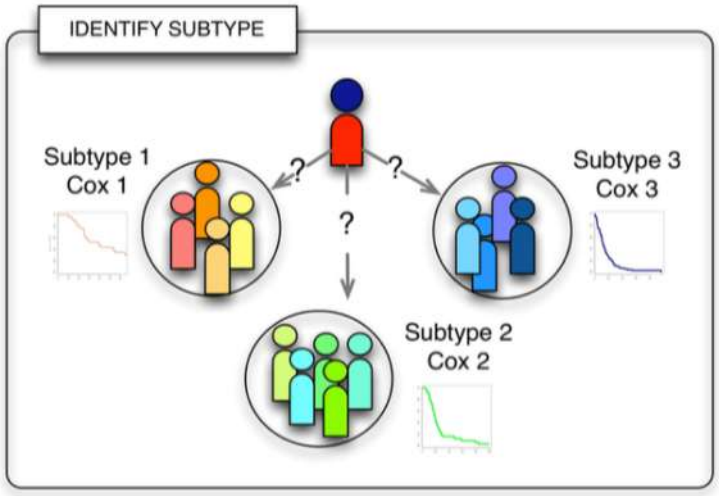


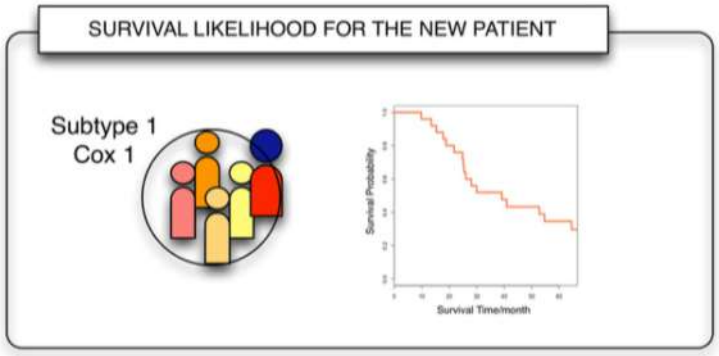


From subtype-based to network-based outcome prediction

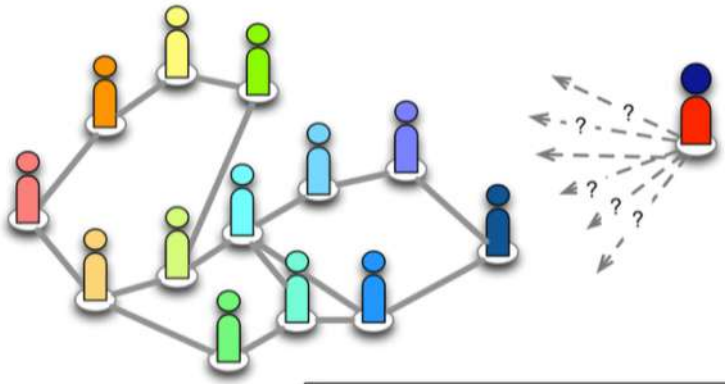


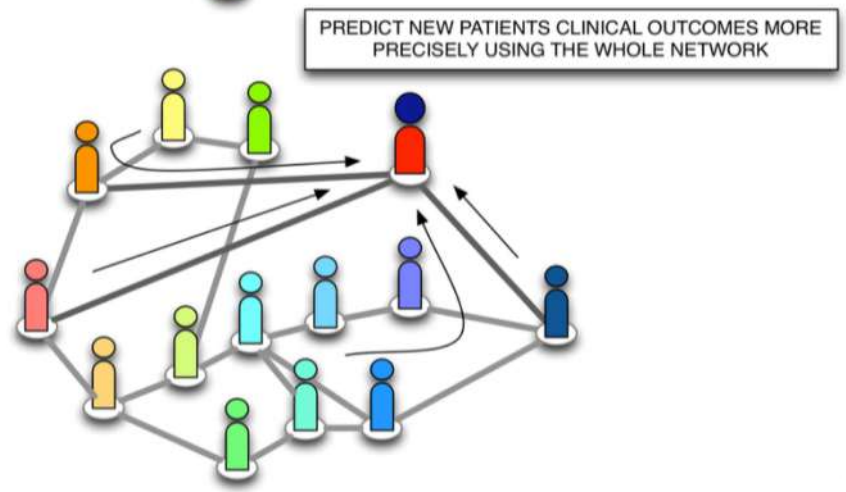
Current Analytics





Future Analytics





Comparisons on an METABRIC breast cancer data

Cox objective

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(\mathbf{x}_i^T z - \log \left(\sum_{j \in R(t_i)} \exp\left(\mathbf{x}_j^T z \right) \right) \right)$$

Network-regularized objective

Incorporate fused patient network structure

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(X_i^T z - \log \left(\sum_{j \in R(t_i)} \exp(X_j^T z) \right) \right) - \lambda \sum_i \sum_j (X_i^T z - X_j^T z)^2 w_{ij}$$

CNV and expression data
Discovery: 997 patients, Validation: 995 patients

	PAM50 (5 clusters)	iCluster (10 clusters)	SNF (5 clusters)	SNF (10 clusters)	Network
P value discovery cohort	3.0×10^{-9}	1.2×10^{-14}	6.10×10^{-11}	3.31×10^{-12}	–
P value validation cohort	1.7×10^{-9}	2.9×10^{-11}	5.12×10^{-13}	7.86×10^{-12}	–
CI discovery cohort	0.560	0.621	0.638	0.638	0.720
CI validation cohort	0.551	0.605	0.633	0.633	0.706

Summary of patient networks framework

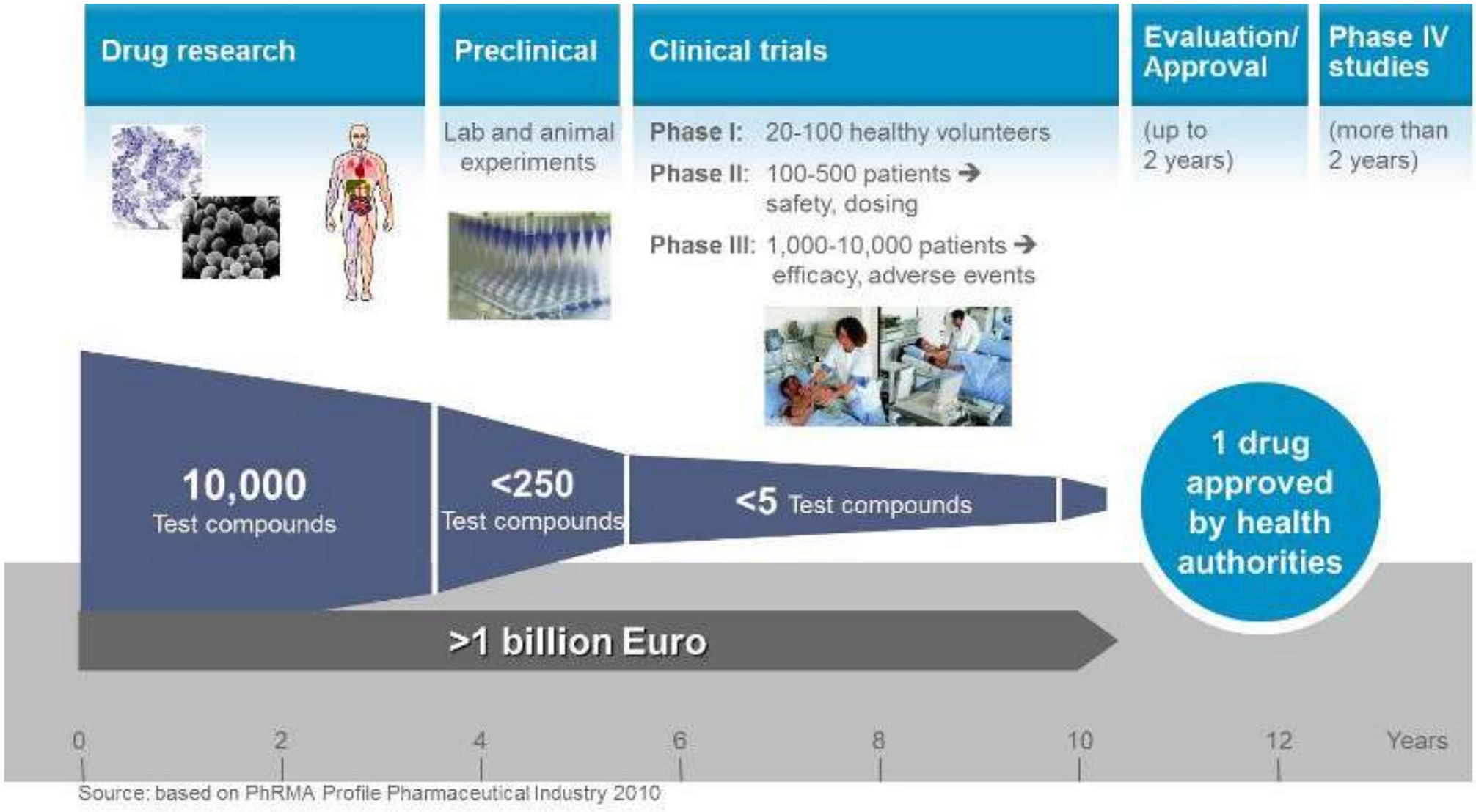
- Creates a unified view of patients based on multiple heterogeneous sources
- Integrates gene and non-gene based data
- Robust to different types of noise
- Obtain superior results on regular tasks such as subtyping and outcome prediction
- Scalable

Wang B, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11:333-337, 2014.

Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Nonnegative Matrix Tri-Factorization and Patient-Specific Data Fusion
- Tensor Factorization and Patient Phenotyping

The Challenge of Drug Discovery



High cost, long time, and low success rate

Reichert JM. Trends in development and approval times for new therapeutics in the US. *Nature Reviews Drug discovery*. 2003;2(9):695-702.

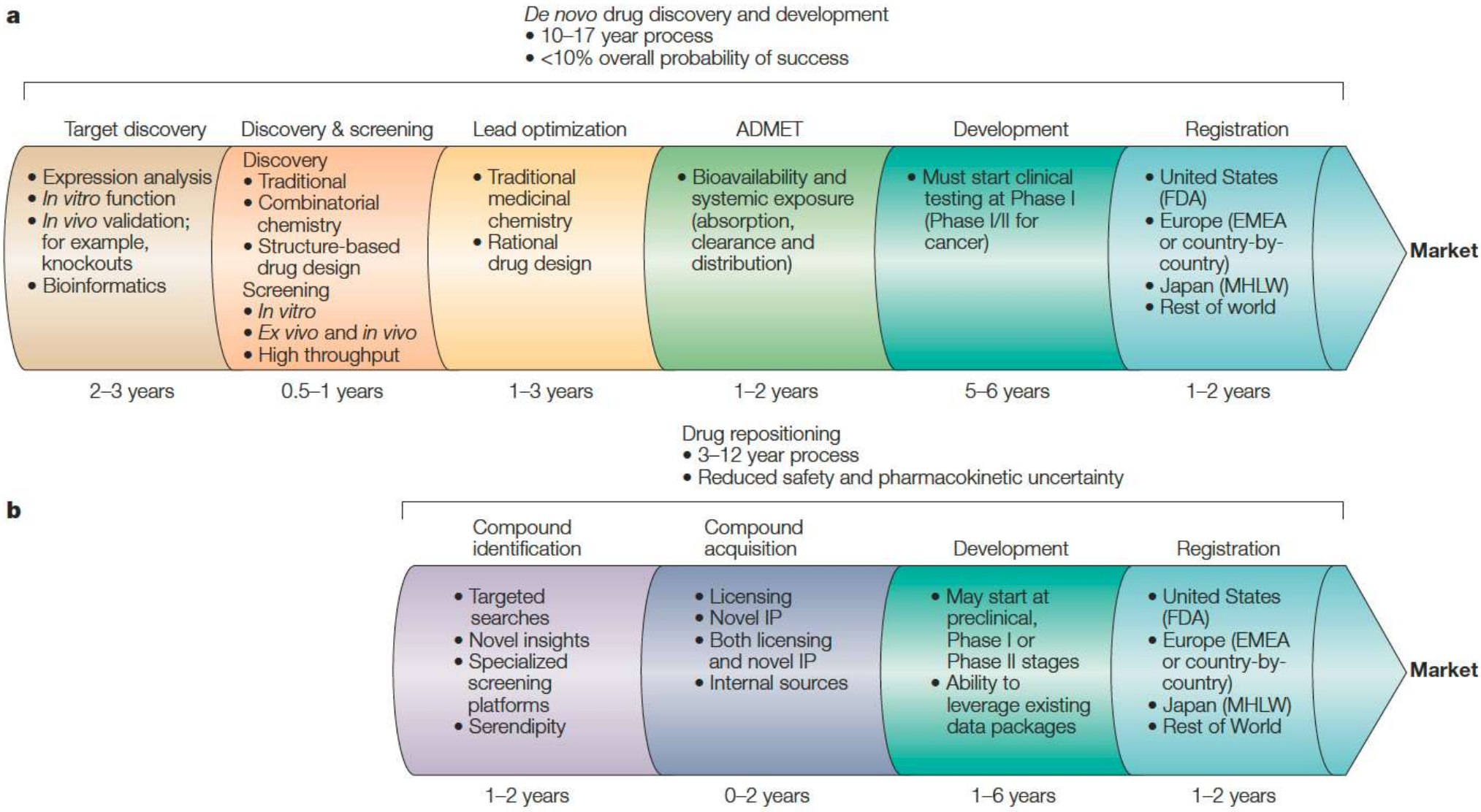
Drug repositioning

- **Drug repositioning** (also known as **Drug repurposing**, **Drug re-profiling**, **Therapeutic Switching** and **Drug re-tasking**) is the application of known drugs and compounds to new indications (i.e., new diseases).

Drug	Original indication	New indication
Viagra	Hypertension	Erectile dysfunction
Wellbutrin	Depression	Smoking cessation
Thalidomide	Antiemetic	Multiple Myeloma

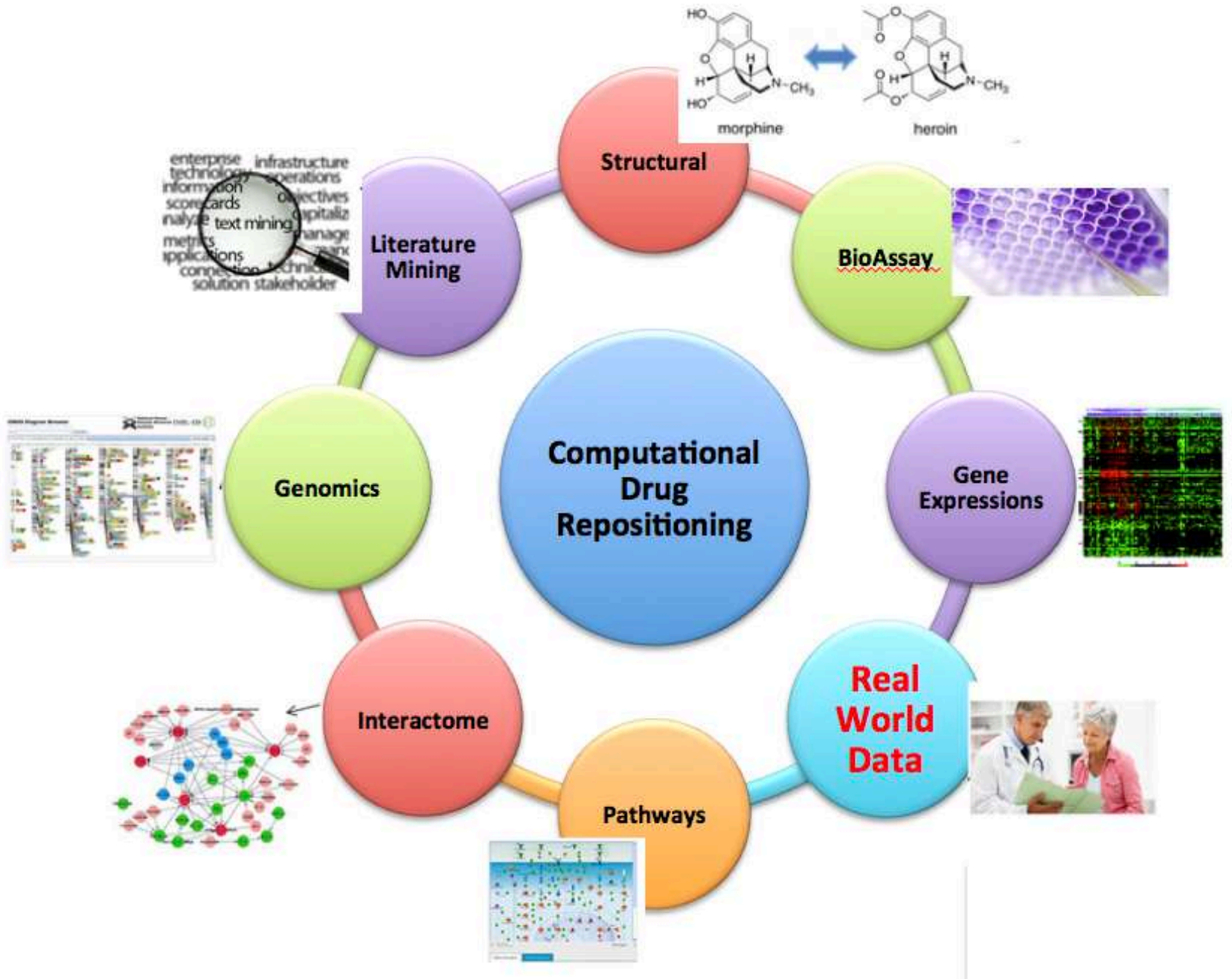
- The repositioned drug has already passed a significant number of toxicity and other tests, its safety is known and the risk of failure for reasons of adverse toxicology are reduced.

Shorter timelines & less risk



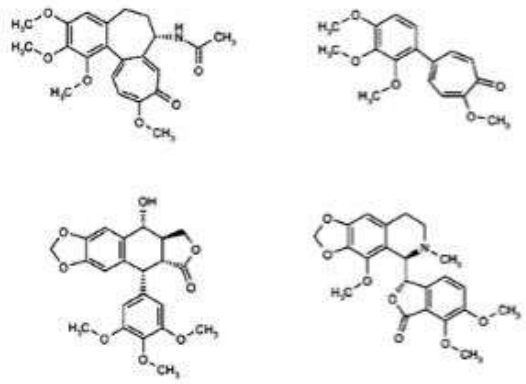
Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673-683, 2004.

Computational drug repositioning

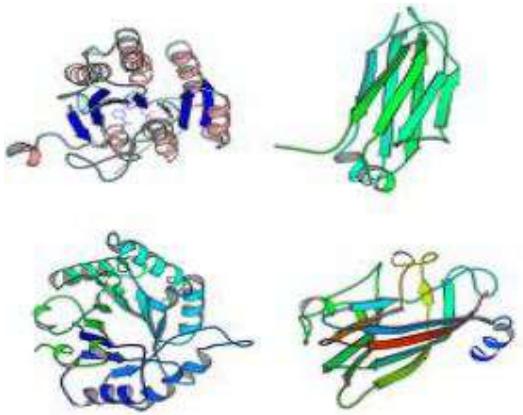


Drug Resources and Disease Resources

Drug



Chemical Structure



Target Proteins

- weight loss
- impotence
- dizziness
- blurred vision
- ……

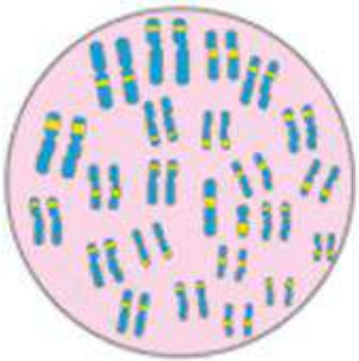
Side-effect Keywords

Calculate drug/disease similarities

Disease

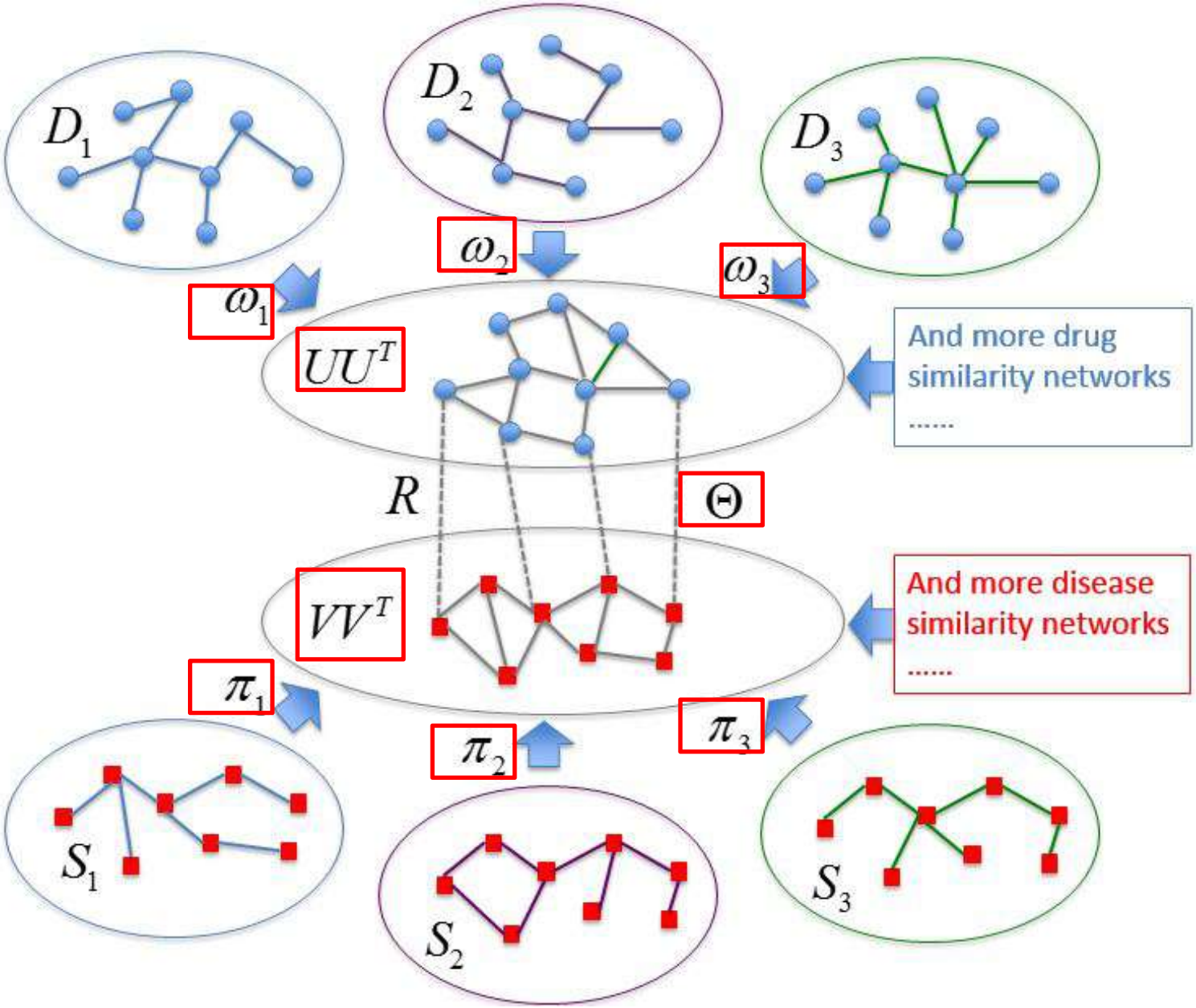
Phenotype/Symptom

Ontology



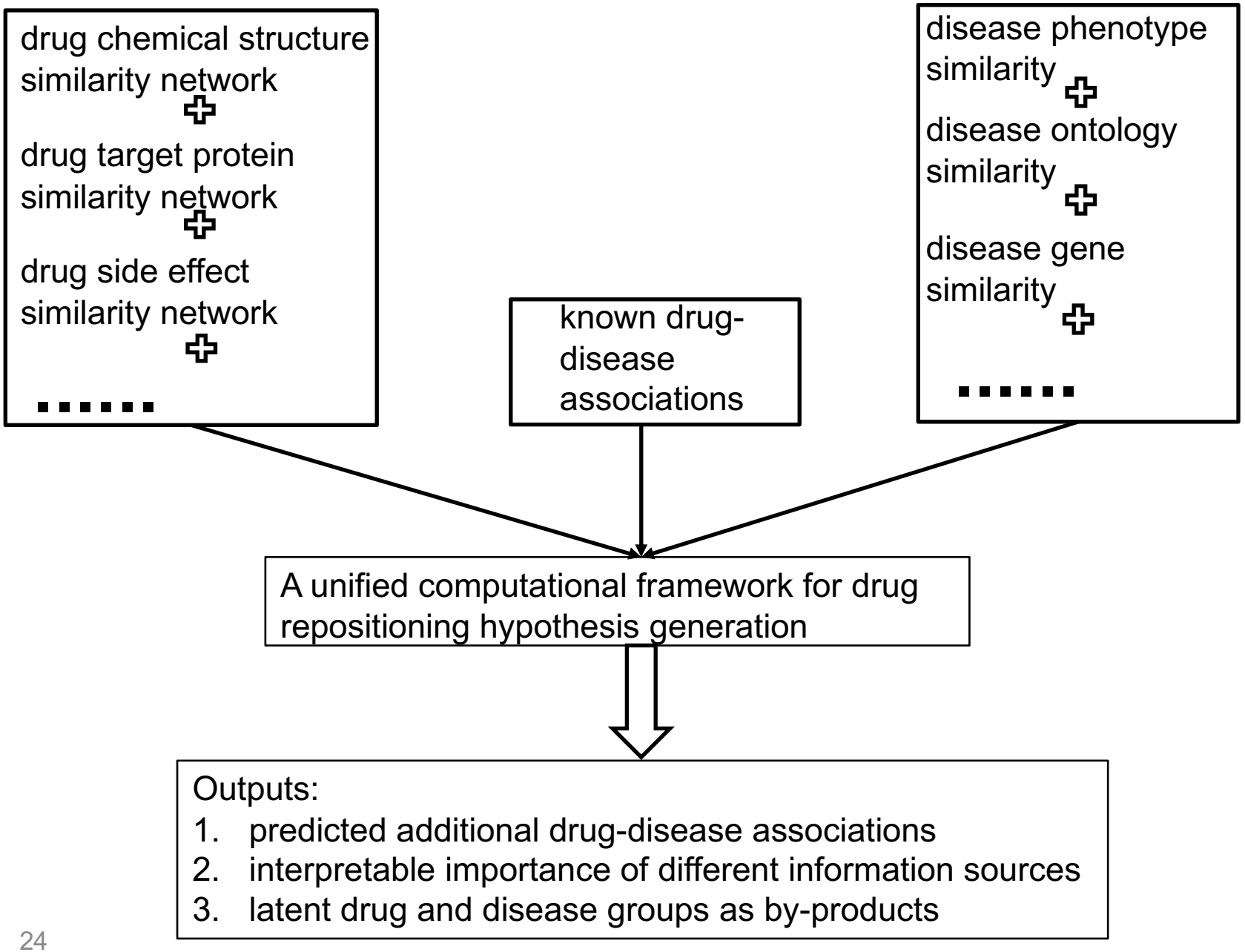
Disease Gene

Joint Matrix Factorization (JMF)



Zhang, P., Wang, F., Hu, J. Towards Drug Repositioning: A Unified Computational Framework for Integrating Multiple Aspects of Drug Similarity and Disease Similarity. *AMIA*, 2014.

Algorithm Flowchart of JMF



JMF as an optimization problem

Notations and symbols of the methodology

$\boldsymbol{D}_k$	n×n	The k-th drug similarity matrix
$\boldsymbol{S}_l$	m×m	The l-th disease similarity matrix
$\boldsymbol{U}$	n×C_D	Drug cluster assignment matrix
$\boldsymbol{V}$	m×C_S	Disease cluster assignment matrix
$\boldsymbol{\Lambda}$	C_D×C_S	Drug-disease cluster relationship matrix
$\boldsymbol{R}$	n×m	Observed drug-disease association matrix
$\boldsymbol{\Theta}$	n×m	Densified estimation of $\boldsymbol{R}$
ω	K_d×1	Drug similarity weight vector
π	K_s×1	Disease similarity weight vector

- We aim to analyze the drug-disease network by minimizing the following objective:

$$J = J_0 + \lambda_1 J_1 + \lambda_2 J_2$$

- The reconstruction loss of observed drug-disease associations:

$$J_0 = \| \Theta - U \Lambda V^T \|_F^2$$

Similar Drugs/diseases (latent groups) have similar behaviors

- The reconstruction loss of drug similarities:

$$J_1 = \sum_{k=1}^{K_d} \omega_k \| D_k - U U^T \|_F^2 + \delta_1 \| \omega \|_2^2$$

- The reconstruction loss of disease similarities:

$$J_2 = \sum_{l=1}^{K_s} \pi_l \| S_l - V V^T \|_F^2 + \delta_2 \| \pi \|_2^2$$

Reconstruct integrated drug/disease networks

- Putting everything together, we obtained the optimization problem to be resolved:

$\min_{U,V,\Lambda,\Theta,\omega,\pi} J$, subject to $U \geq 0$, $V \geq 0$, $\Lambda \geq 0$, $\omega \geq 0$, $\omega^T \mathbf{1} = 1$, $\pi \geq 0$, $\pi^T \mathbf{1} = 1$, $P_\Omega(\Theta) = P_\Omega(R)$

BCD approach for solving the problem

- **Block Coordinate Descent (BCD) strategy:** The BCD approach works by solving the different groups of variables alternatively until convergence. At each iteration, it solves the optimization problem with respect to one group of variables with all other groups of variables fixed.

Algorithm 1: A BCD Approach for Solving Problem (11)

Require: $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\delta_1 \geq 0$, $\delta_2 \geq 0$, $K_d > 0$, $K_s > 0$, $\{\boldsymbol{D}_k\}_{k=1}^{K_d}$, $\{\boldsymbol{S}_l\}_{l=1}^{K_s}$, $\boldsymbol{R}$

1: Initialize $\omega = (1/K_d)\mathbf{1} \in \mathbb{R}^{K_d \times 1}$, $\pi = (1/K_s)\mathbf{1} \in \mathbb{R}^{K_s \times 1}$

2: Initialize $\boldsymbol{U}$ and $\boldsymbol{V}$ by performing Symmetric Nonnegative Matrix Factorization on $\tilde{D} = \sum_{k=1}^{K_d} \omega_k D_k$ and $\tilde{S} = \sum_{l=1}^{K_s} \pi_l S_l$.

3: **while** Not Converge **do**

4: Solve $\boldsymbol{\Theta}$ as described in section 2 (as a constrained Euclidean projection)

5: Solve ω and π as described in section 3 (as a standard Euclidean projection onto a simplex)

Closed-form solution

6: Solve Λ as described in section 4 (as a nonnegative quadratic optimization problem)

7: Solve $\mathbf{U}$ as described in section 5 (as a nonnegative quadratic optimization problem)

8: Solve $\mathbf{V}$ as described in section 6 (as a nonnegative quadratic optimization problem)

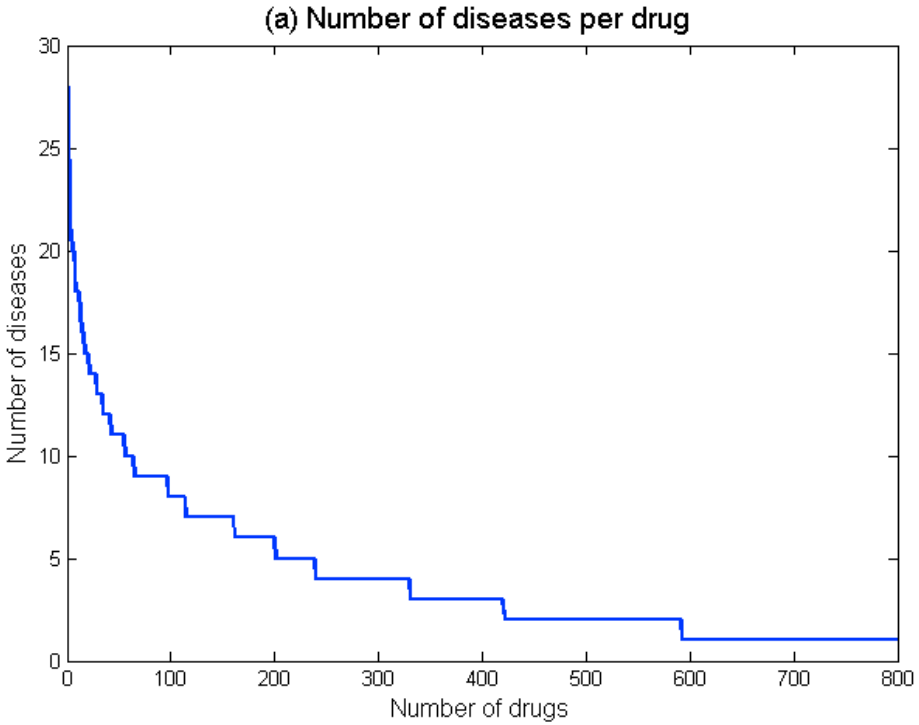
Solved by Projected Gradient Descent (PGD) method

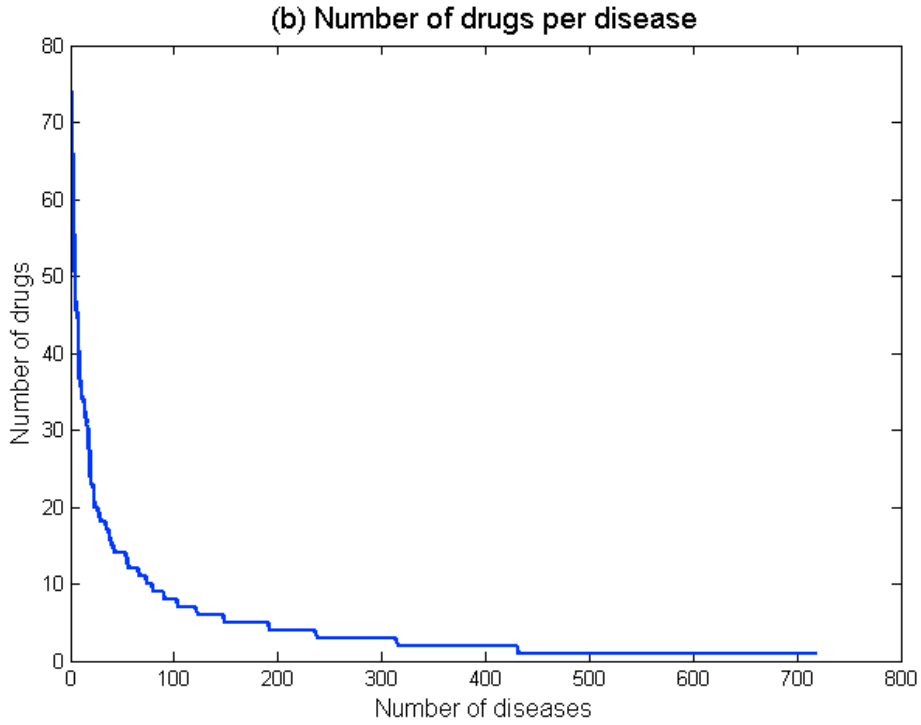
9: **end while**

Computational complexity is O(Rrmn) , where R is the number of BCD iterations, and r is the average PGD iterations when updating Λ, U, and V.

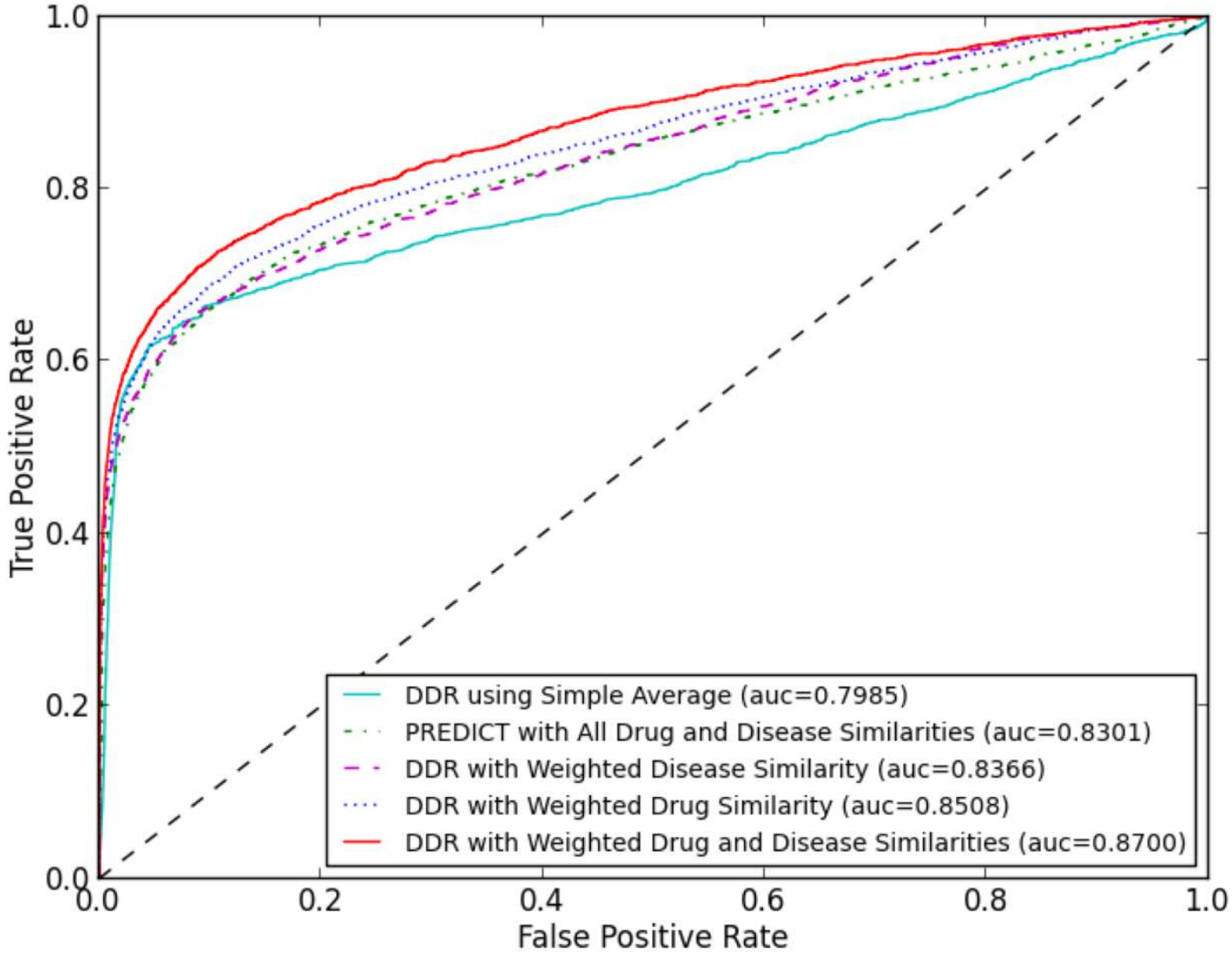
Data Description

- Benchmark dataset was extracted from NDF-RT, spanning 3,250 treatment associations between 799 drugs and 719 diseases
- **Three** 799×799 matrices were used to represent drug similarities between 799 drugs from different perspectives
- **Three** 719×719 matrices were used to represent disease similarities between 719 human diseases from different perspectives

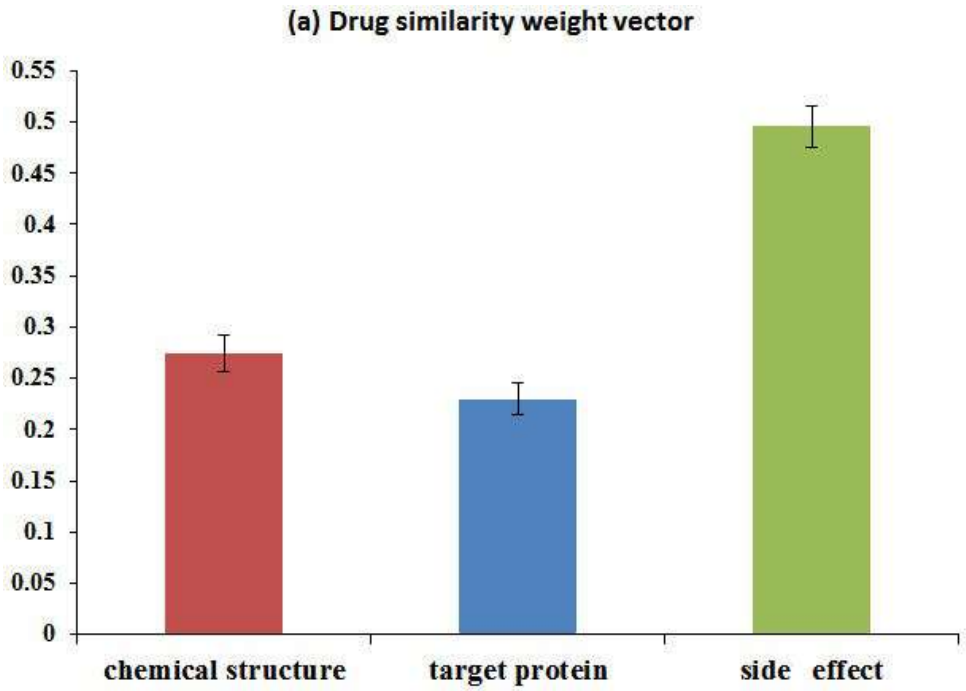


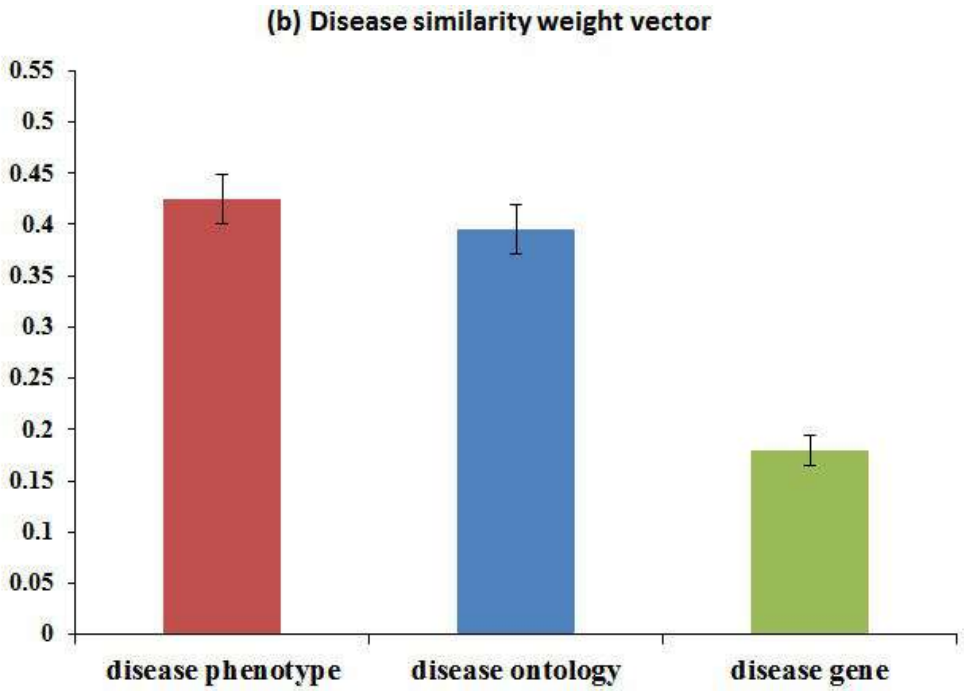


ROC comparisons of five drug repositioning approaches



Distribution of weights of the similarity weight vectors obtained by JMF





Top 10 drugs for diseases Alzheimer's Disease (AD) and Systemic Lupus Erythematosus (SLE)

(a) Top 10 drugs predicted for AD

Drug	Prediction Score	Clinical Evidence?
Selegiline*	0.7091	—
Carbidopa	0.6924	No
Amantadine	0.6897	No
Procyclidine	0.6826	No
Valproic Acid*	0.6745	—
Metformin	0.6543	Yes
Bexarotene	0.6426	Yes
Neostigmine	0.6385	No
Galantamine*	0.6348	—
Nilvadipine	0.6159	Yes

Repositioning candidates

(b) Top 10 drugs predicted for SLE

Drug	Prediction Score	Clinical Evidence?
Desoximetasone	0.7409	No
Azathioprine*	0.7269	—
Leflunomide	0.7078	Yes
Fluorometholone	0.7054	No
Triamcinolone*	0.6862	—
Beclomethasone	0.6522	No
Etodolac	0.6445	No
Hydroxychloroquine*	0.6374	—
Nelfinavir	0.6371	Yes
Mercaptopurine	0.6150	No

* denotes the drug is known and approved to treat the disease

Summary of joint matrix factorization framework

- We proposed a general computational framework, to explore drug-disease associations from multiple drug/disease sources
- Our method could help generate drug repositioning hypotheses, which will benefit patients by offering more effective and safer treatments
- The computational framework and its solution can be used in other applications (gene-disease, drug-patient, etc.)

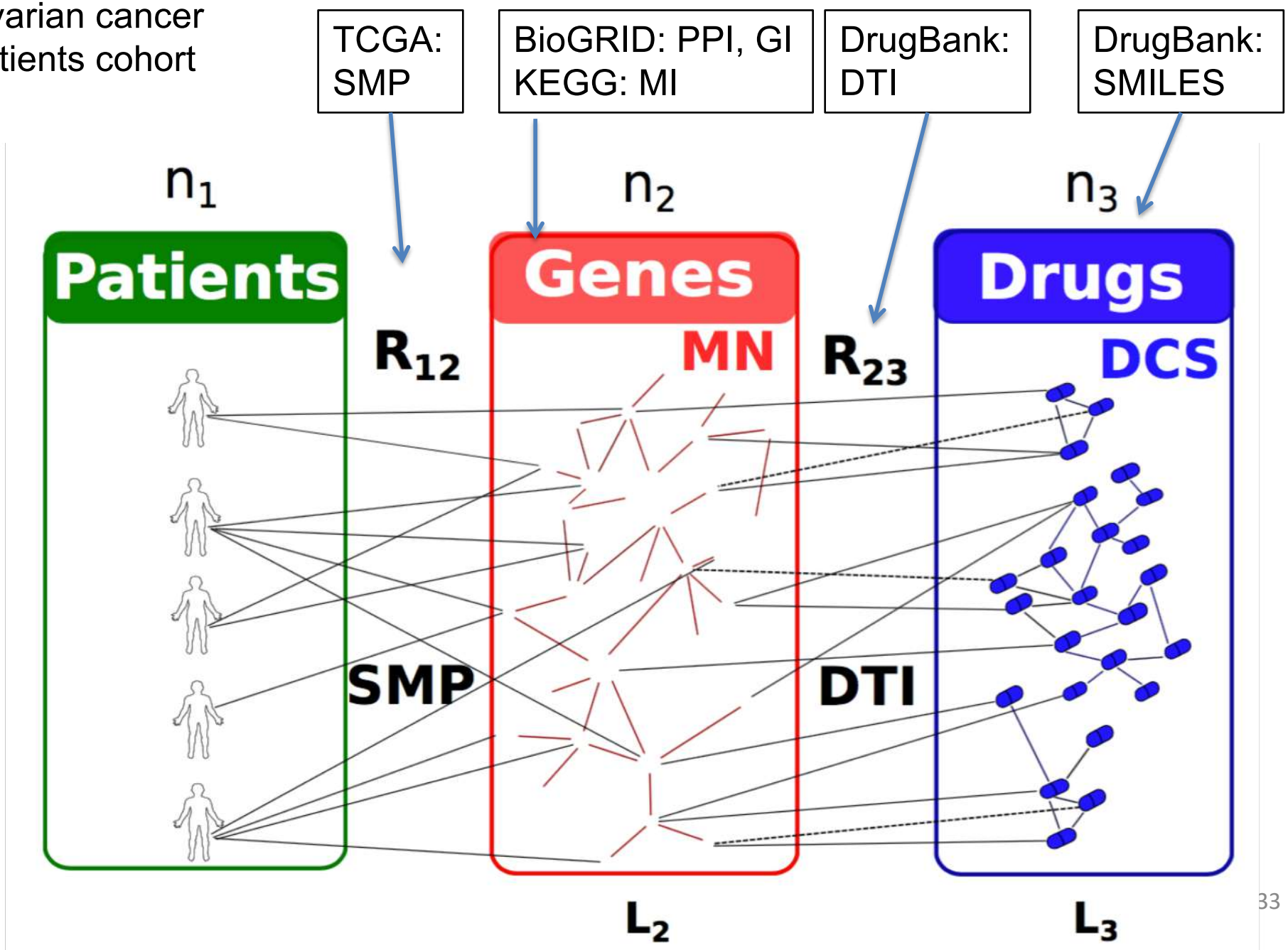
Zhang, P., Wang, F., Hu, J. Towards Drug Repositioning: A Unified Computational Framework for Integrating Multiple Aspects of Drug Similarity and Disease Similarity. *AMIA*, 2014.

Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Nonnegative Matrix Tri-Factorization and Patient-Specific Data Fusion
- Tensor Factorization and Patient Phenotyping

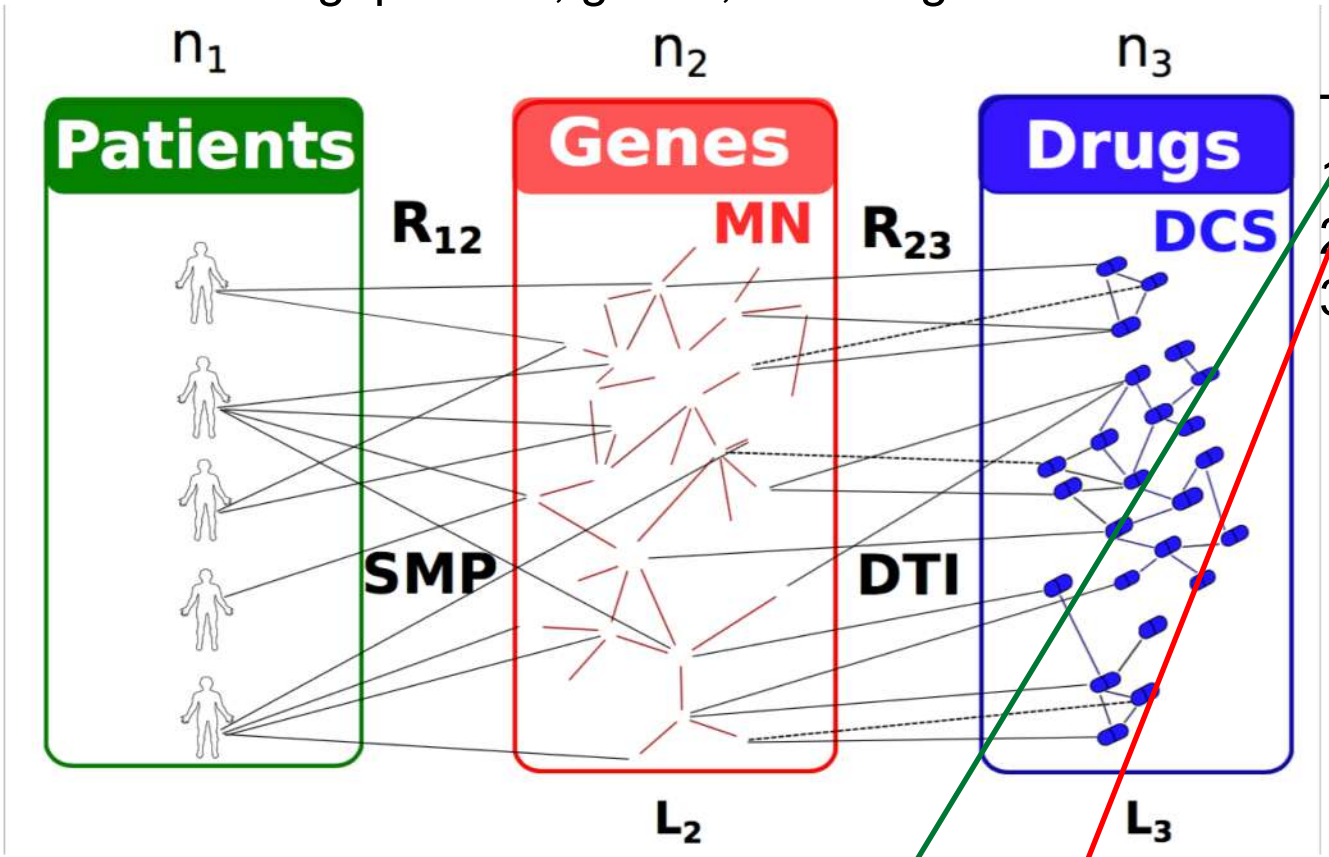
Network data integration

Ovarian cancer patients cohort



Patient-specific data fusion

Co-clustering: patients, genes, and drugs



Tasks:

1. patient stratification
2. driver gene prediction
3. drug-target interaction prediction

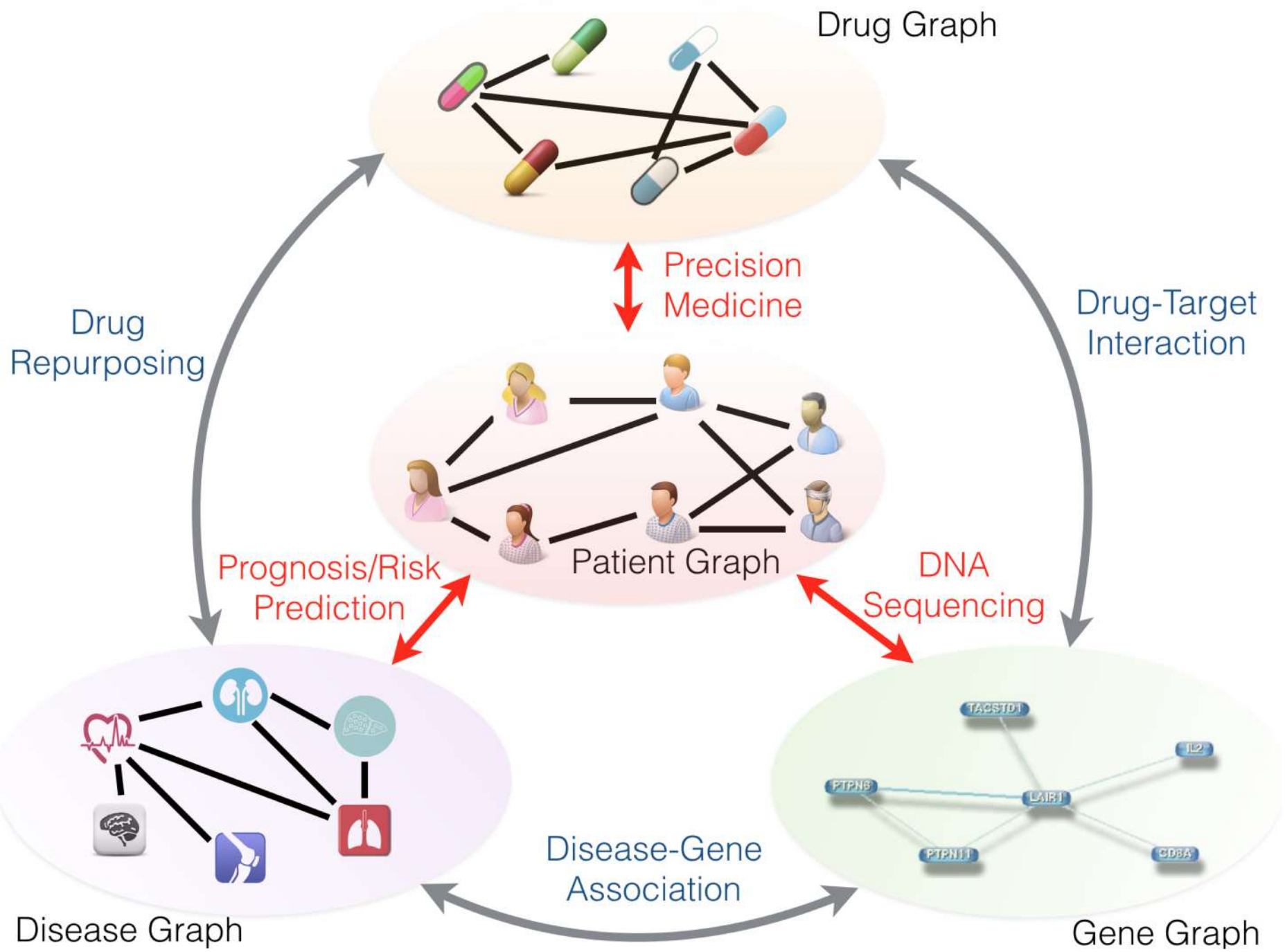
Solved by Graph-regularized Non-negative Matrix Tri-Factorization (GNMTF)

$$\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \left[\| \mathbf{R}_{12} - \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \|_F^2 + \| \mathbf{R}_{23} - \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \|_F^2 + tr(\mathbf{G}_2^T \mathbf{L}_2 \mathbf{G}_2) + tr(\mathbf{G}_3^T \mathbf{L}_3 \mathbf{G}_3) \right]$$

Gligorijevic V et al. Patient-specific data fusion for cancer stratification and personalized treatment. PSB, 2016.

Wang F, Li T, Zhang C. Semi-supervised clustering via matrix factorization. SDM, 2008.

Matrix models in biomedicine



Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Nonnegative Matrix Tri-Factorization and Patient-Specific Data Fusion
- Tensor Factorization and Patient Phenotyping

Phenotyping from Electronic Medical Records (EMR)

Phenotype (American Heritage Dictionary)

- The *observable* physical or biochemical *characteristics* of an organism, as determined by both genetic makeup and environmental influences.

Why phenotyping from EMR

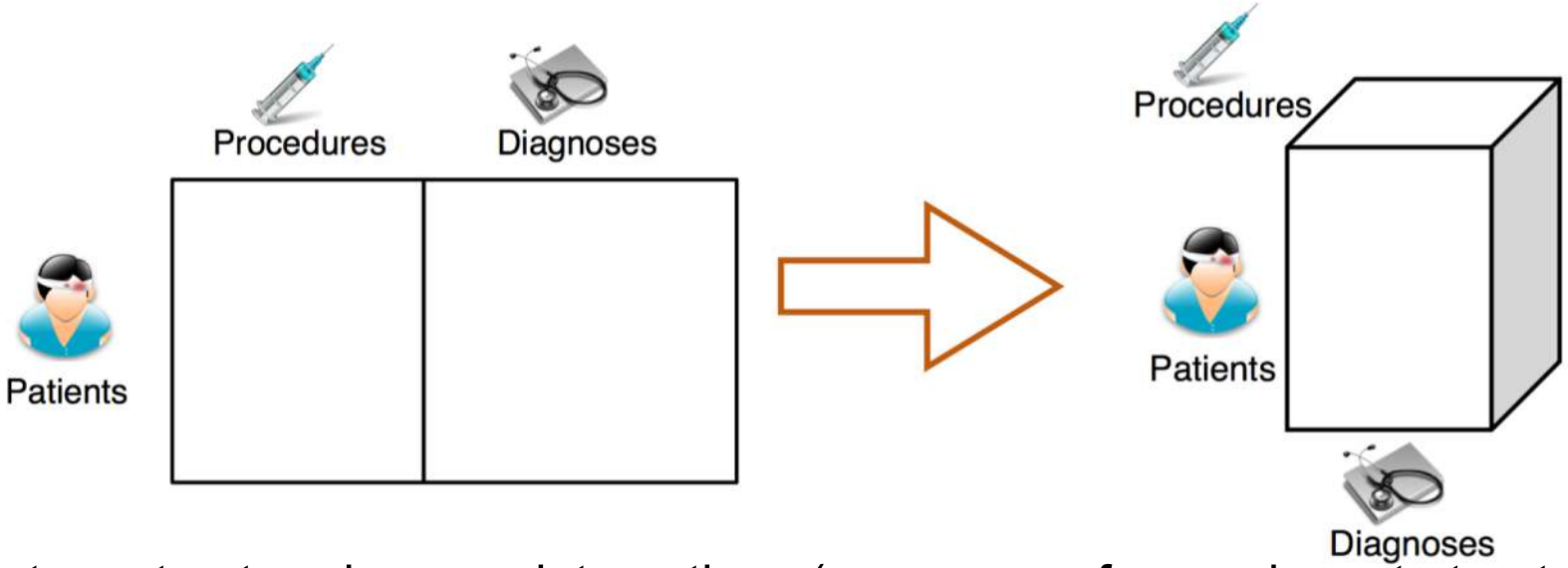
- Mapping *noisy*, *incomplete*, and potentially *inaccurate* patient representation from EMR to meaningful medical concepts Feature engineering
- Extracting clinical meaningful groups of patients from EMR Cohort generation

Diabetes Phenotype
Diseases of other endocrine glands
Complications of surgical and medical care
Chemistry Pathology and Laboratory Tests
Organ or Disease Oriented Panels
Hematology and Coagulation Procedures
Surgical Procedures on the Cardiovascular System

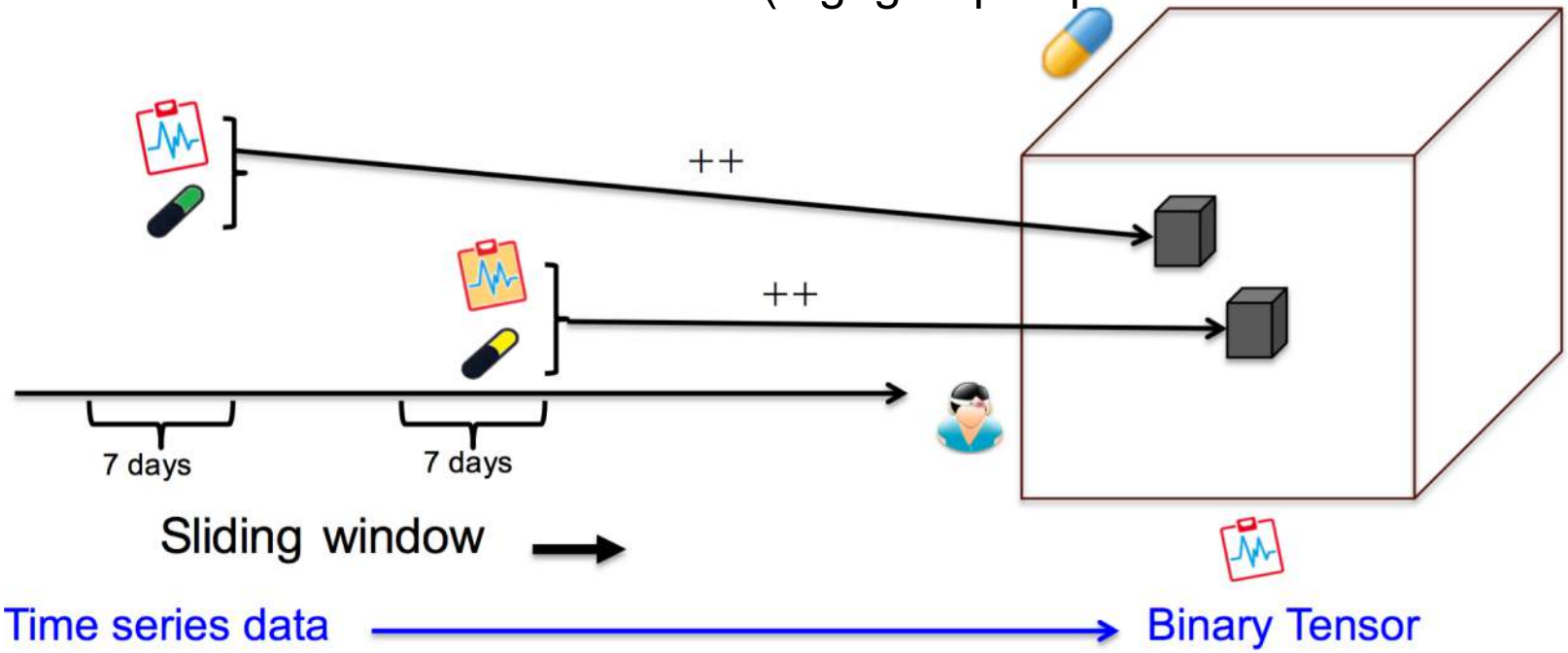
Heart Failure Phenotype
Other forms of heart disease
Complications of surgical and medical care
Symptoms
Cardiovascular Procedures
Hematology and Coagulation Procedures
Evaluation and Management of Other Outpatient Services
Surgical Procedures on the Cardiovascular System
Chemistry Pathology and Laboratory Tests

Ho J, Ghosh J, Sun J. Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization. KDD 2014.

Tensor representation for EMR

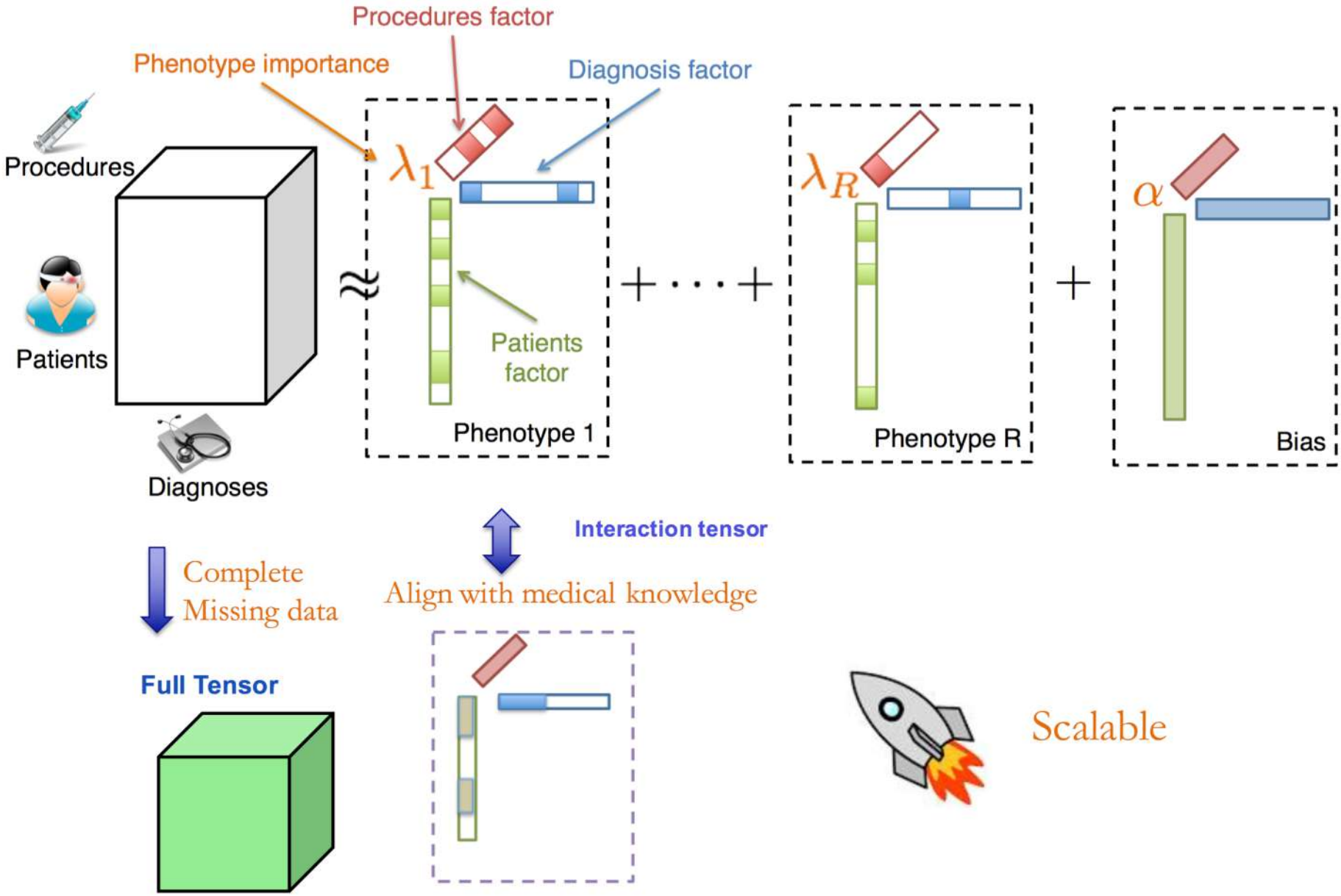


Capture structured source interactions (e.g. group of procedures to treat a disease)



Co-occurrences of events are captured in the tensor as binary values

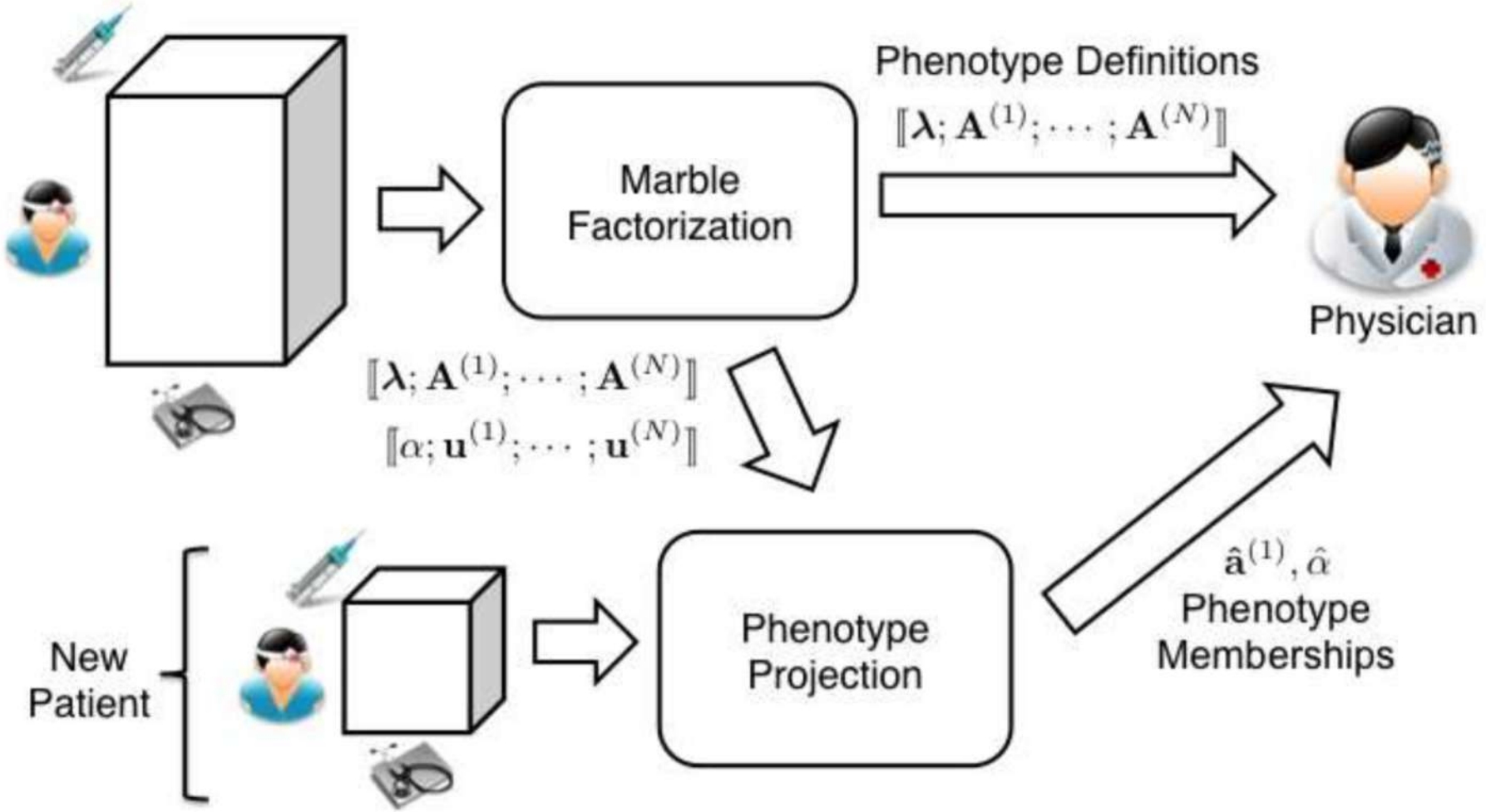
CP factorization for EMR



Ho J et al. Marble: High-throughput phenotyping from Electronic Health Records via sparse nonnegative tensor factorization. KDD 2014.

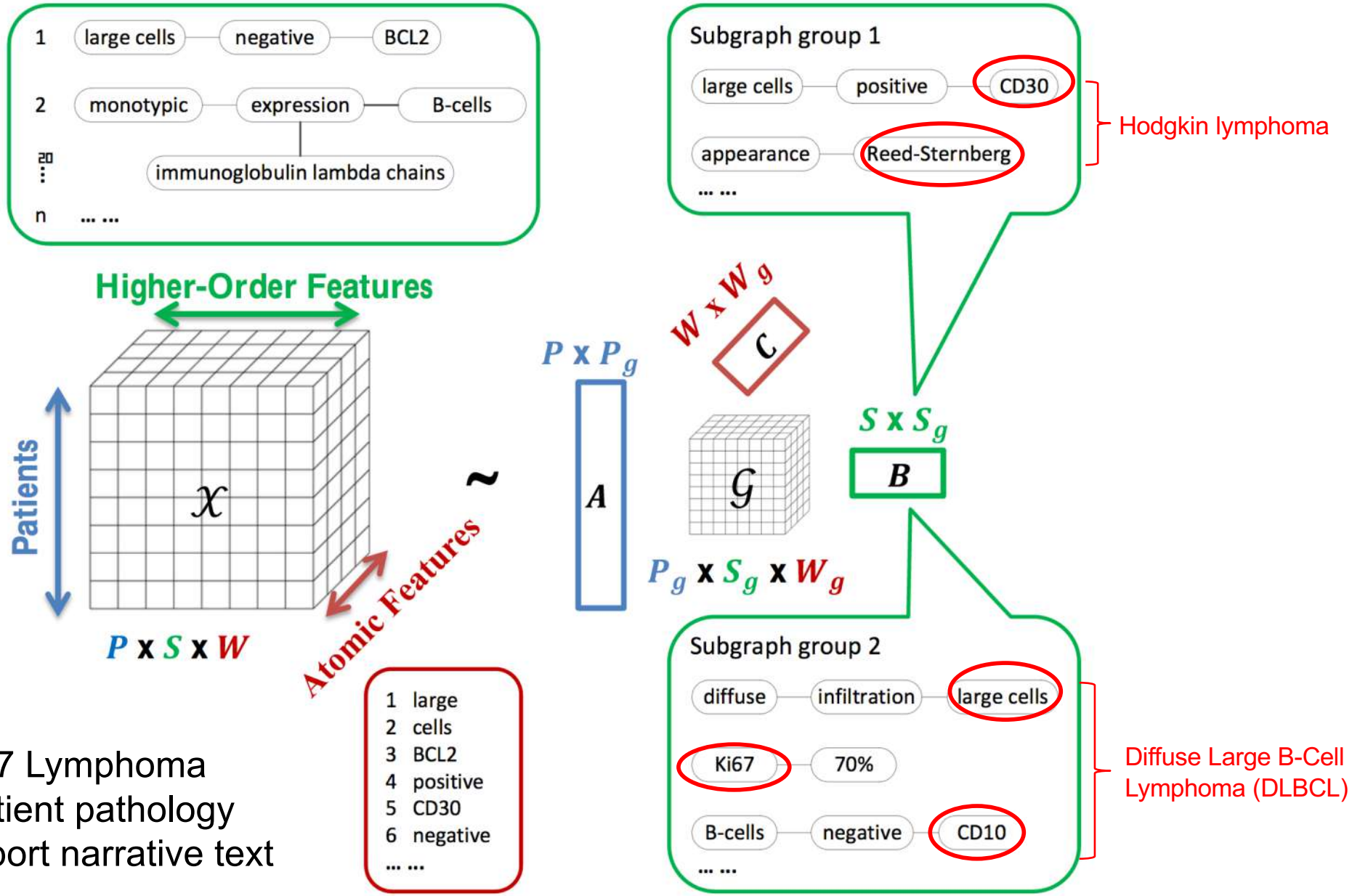
Wang Y et al. Rubik: Knowledge guided tensor factorization and completion for health data analytics. KDD 2015.

A possible application of EHR-phenotyping



Ho J, Ghosh J, Sun J. Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization. KDD 2014.

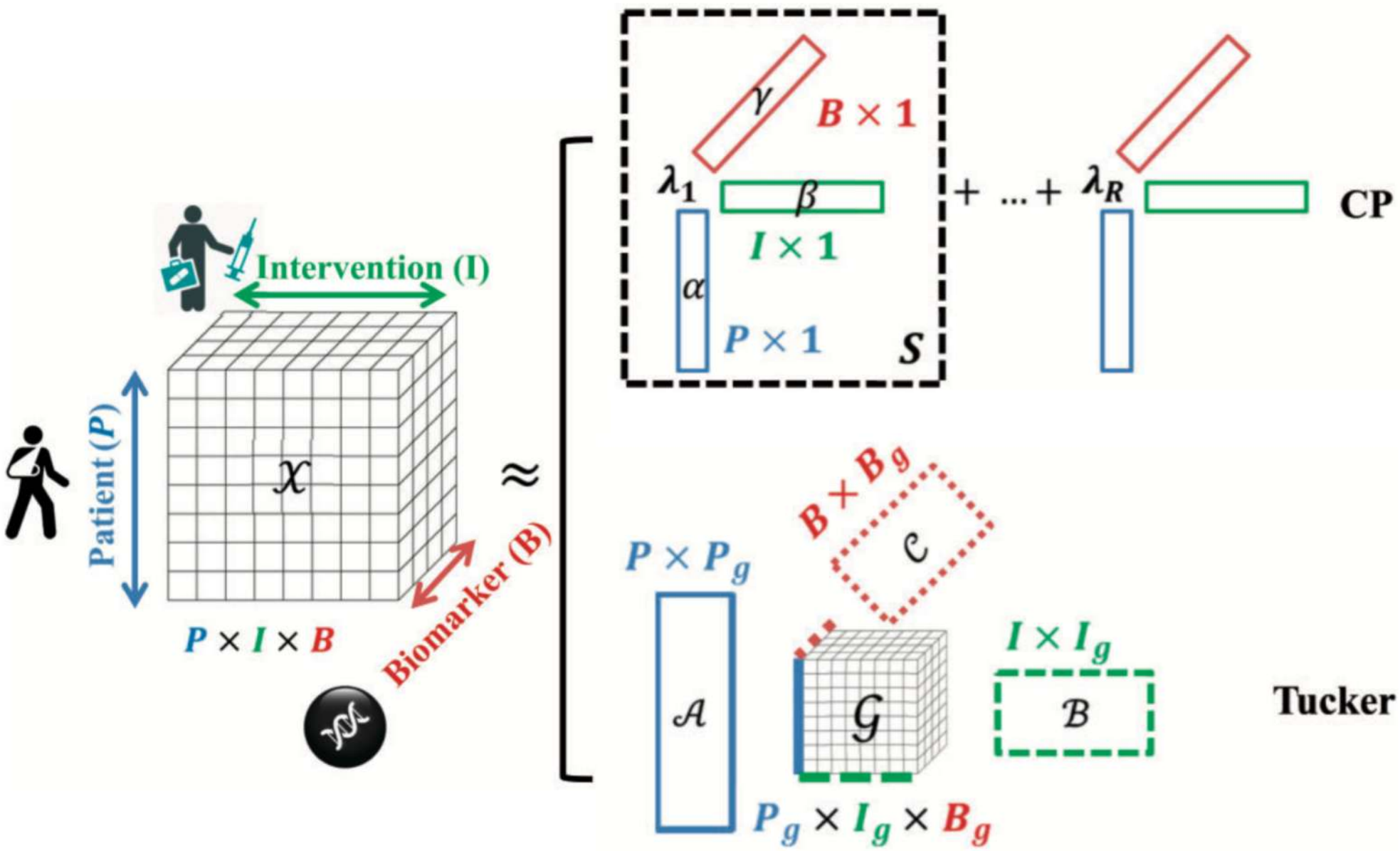
Tucker factorization for pathology reports



897 Lymphoma patient pathology report narrative text

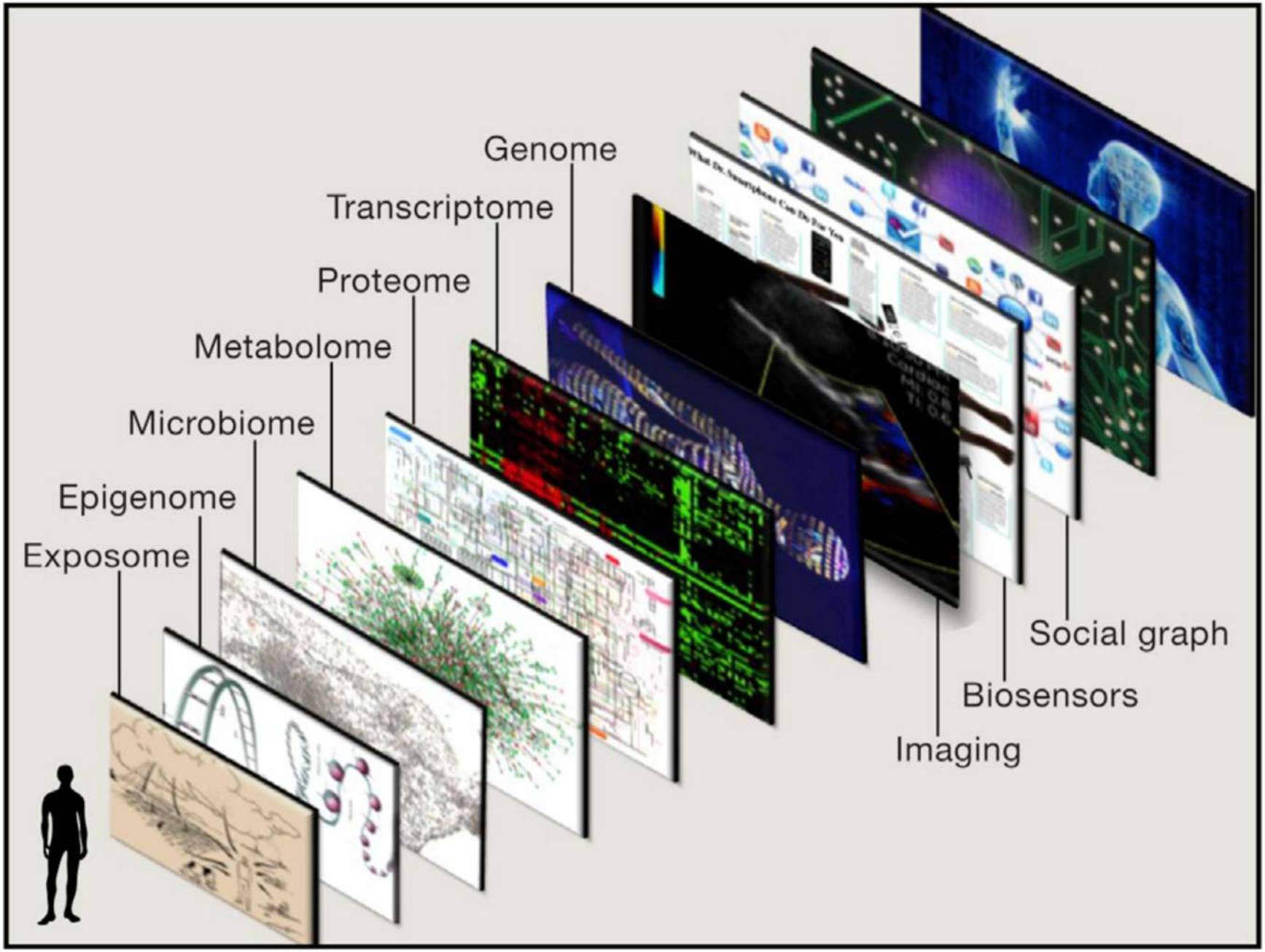
Luo Y et al. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. *JAMIA* 22:1009-1019, 2015.

Comparison of tensor modeling and factorization schemes



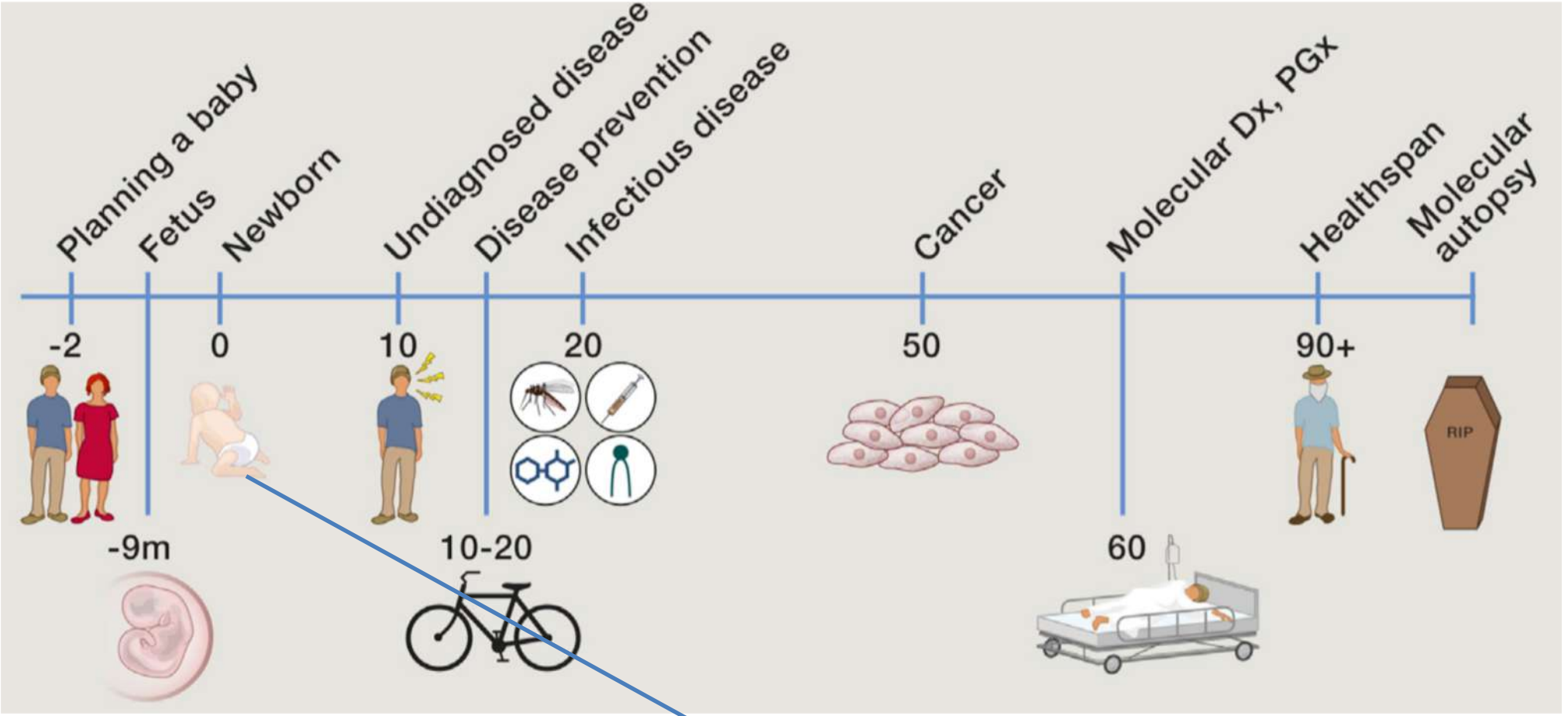
Luo Y, Wang F, Szolovits P. Tensor factorization toward precision medicine. *Brief Bioinform*, 2016

Challenges and opportunities: multiscale networks



Topol E. Individualized Medicine from Prewomb to Tomb. *Cell* 157, 2014.

Dynamic network: timeline of individualized genomic medicine

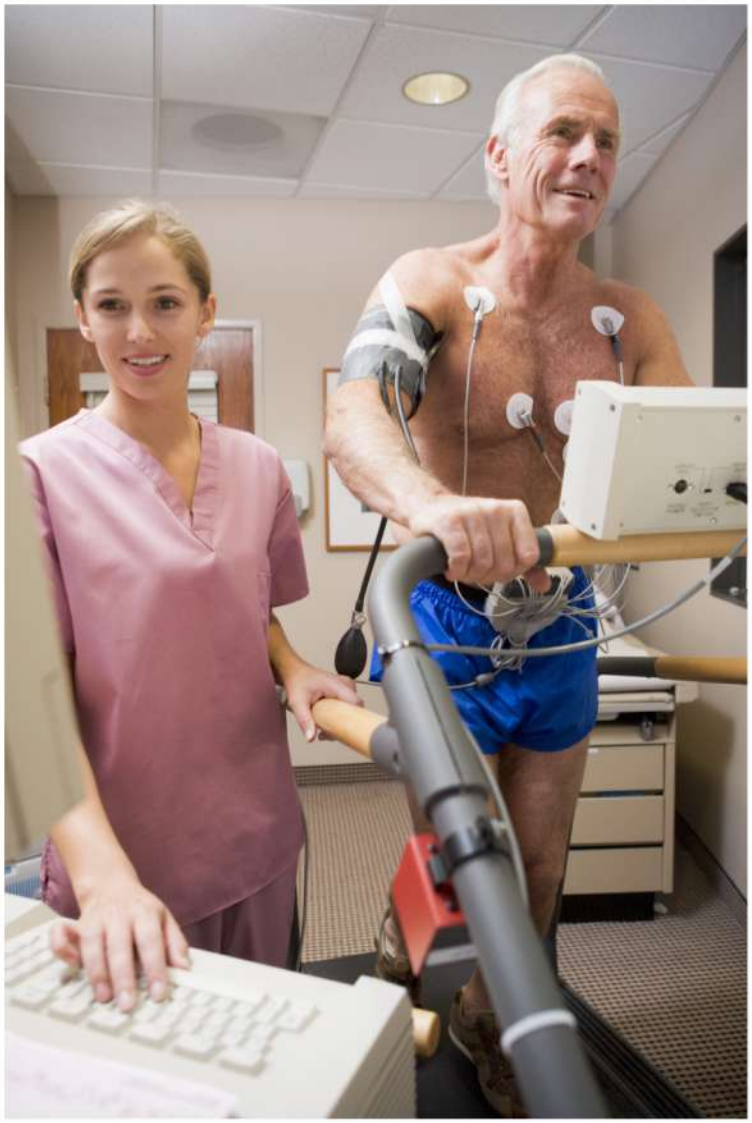


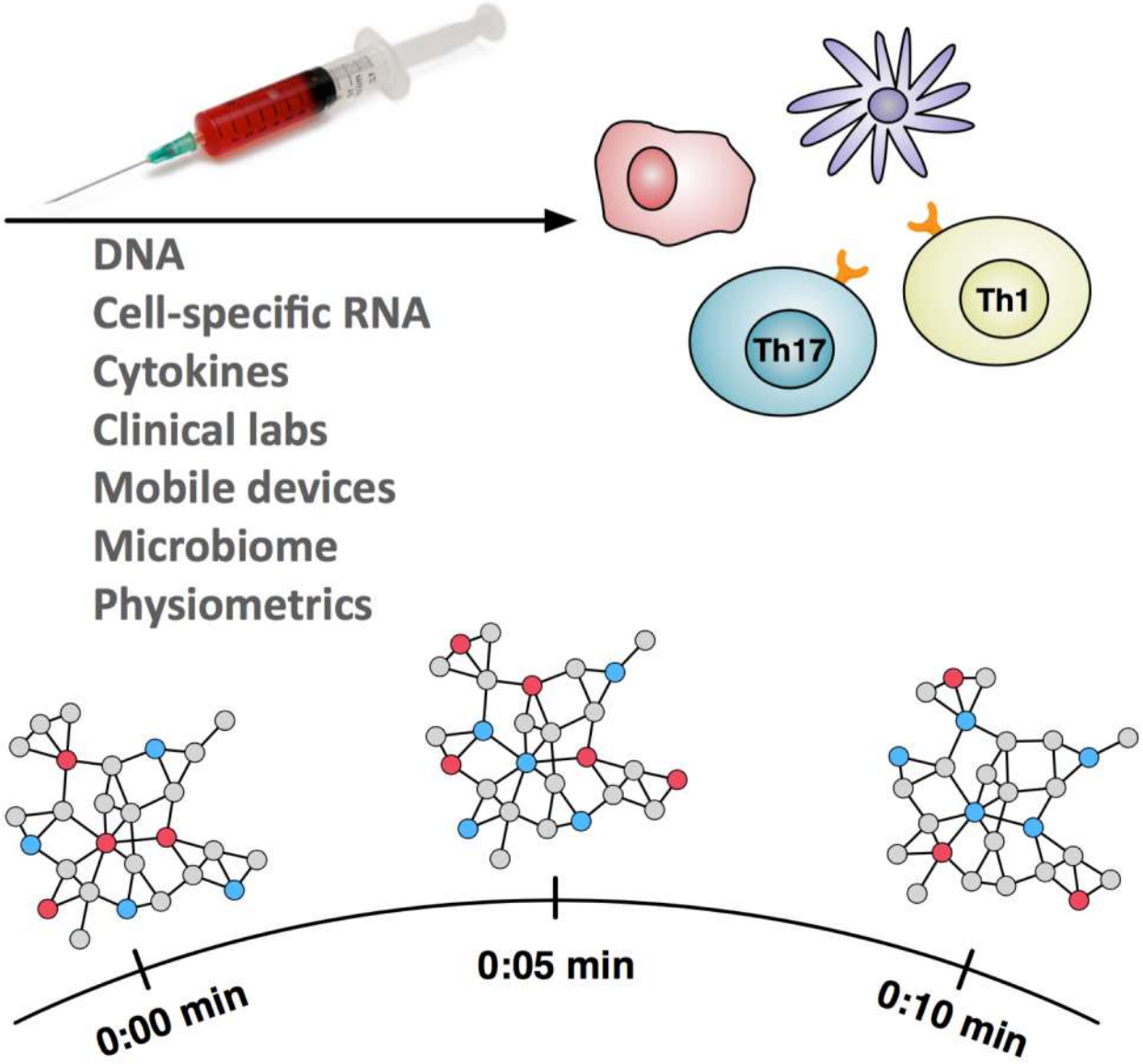
During an individual's lifespan: ***from prewomb to tomb***

Boland MR et al. Birth Month Affects Lifetime Disease Risk: A Phenome-Wide Method. JAMIA 2015.

Topol E. Individualized Medicine from Prewomb to Tomb. *Cell* 157, 2014.

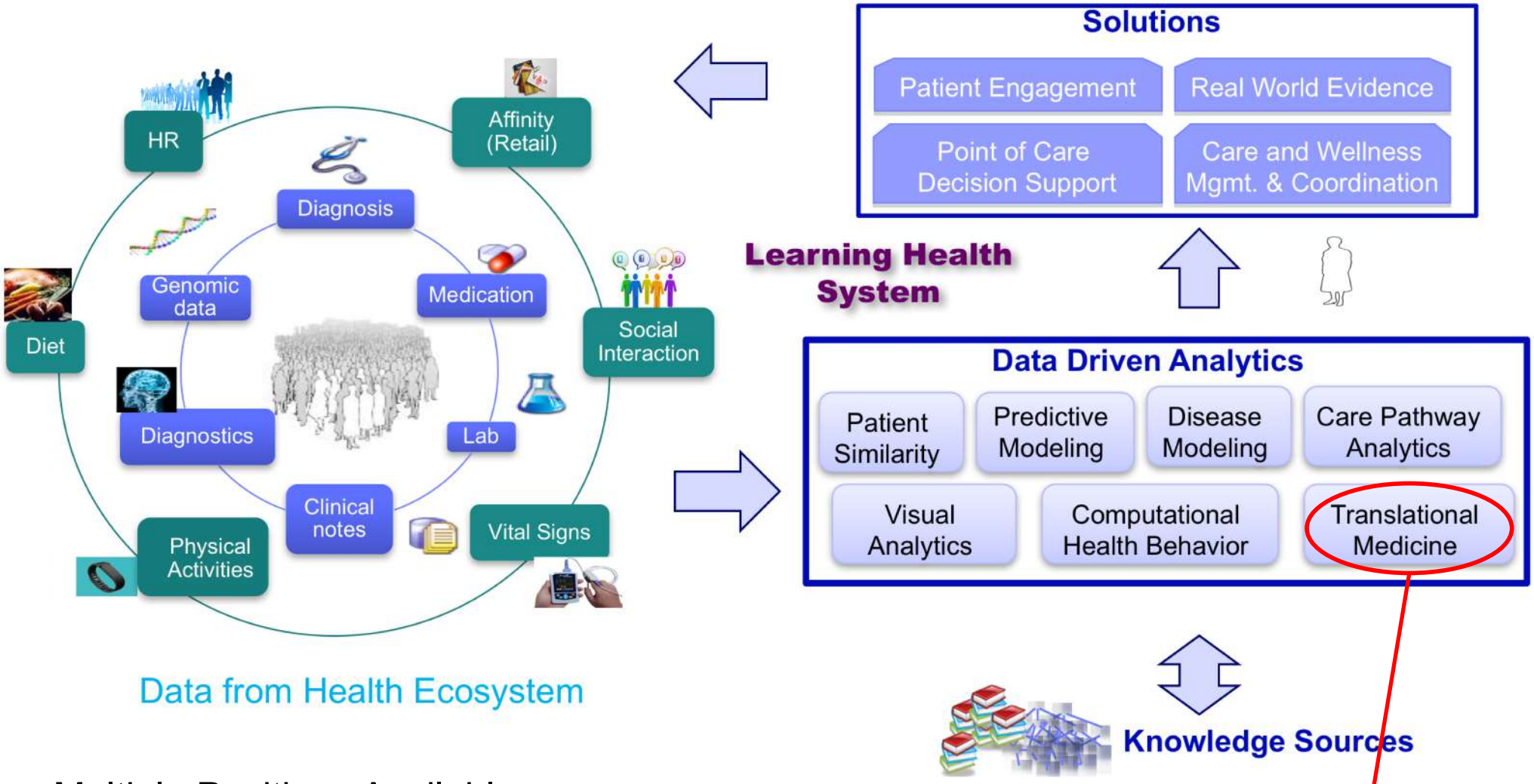
Personalized multiscale networks to model dynamics of complex disease





Dudley J. Big data in biology and medicine. Retreived at www.aaas.org

Center for Computational Health @ IBM



Translational Informatics

Multiple Positions Available:

- Interns
- Postdocs
- Research Engineers
- Research Staff Members

Contact:
pzhang@us.ibm.com

Thank you!!!