Healthcare Data Mining with Matrix Models

KDD 2016 Tutorial Part II

August 13th, 2016

Ping Zhang

Center for Computational Health

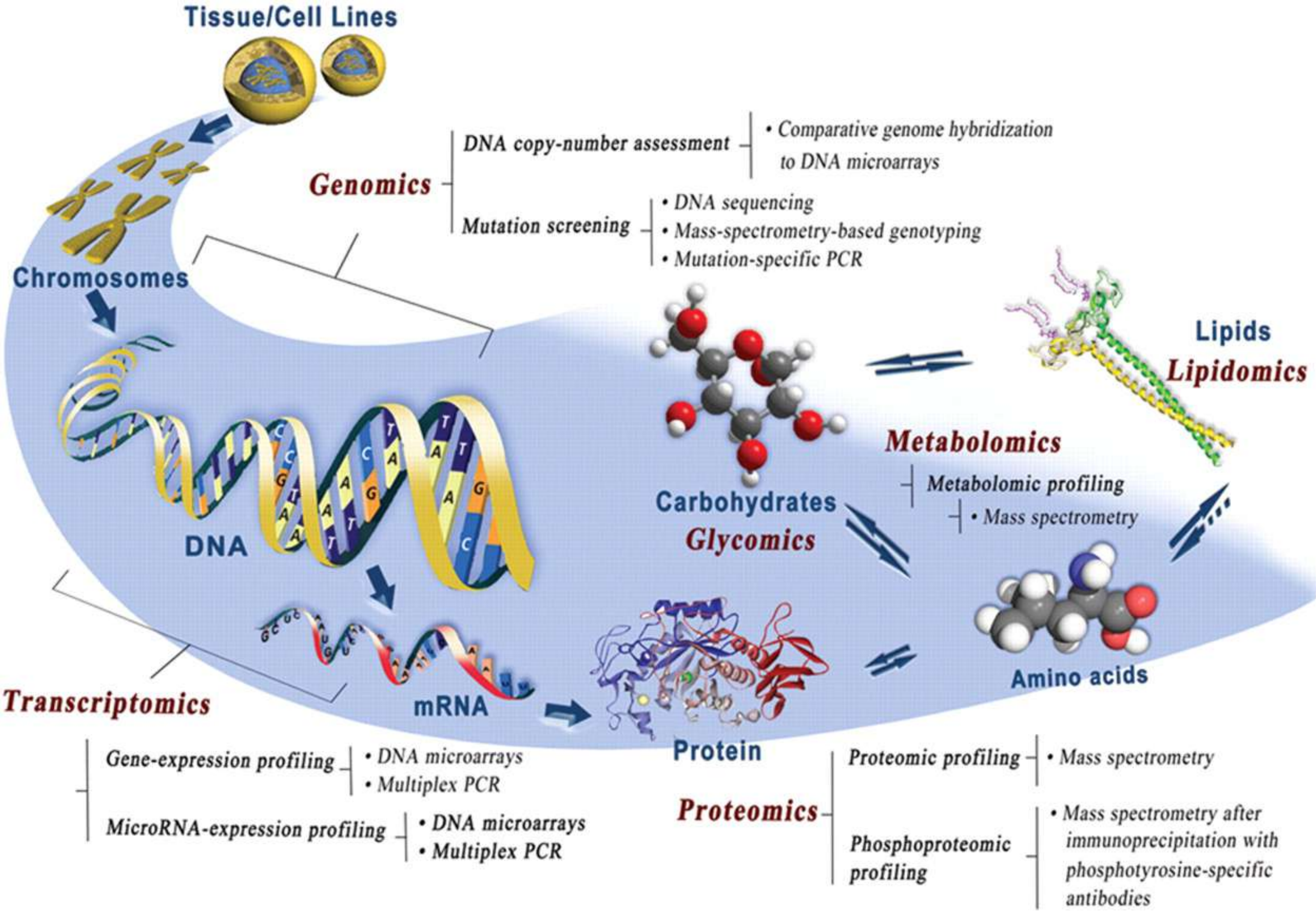
IBM T.J. Watson Research Center

pzhang@us.ibm.com

Recent Applications in Biomedicine

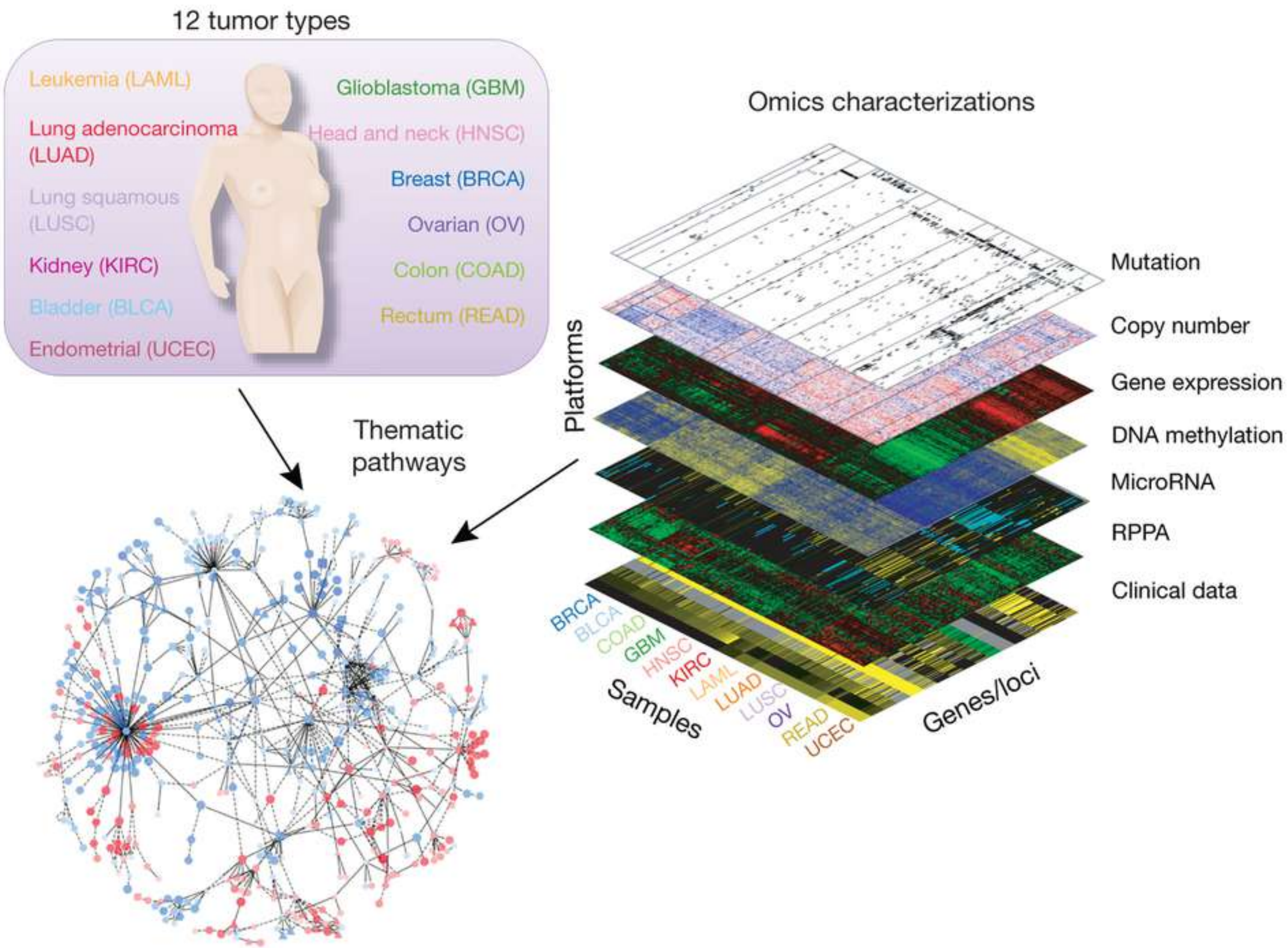
- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping

Omics technologies in biomedicine



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

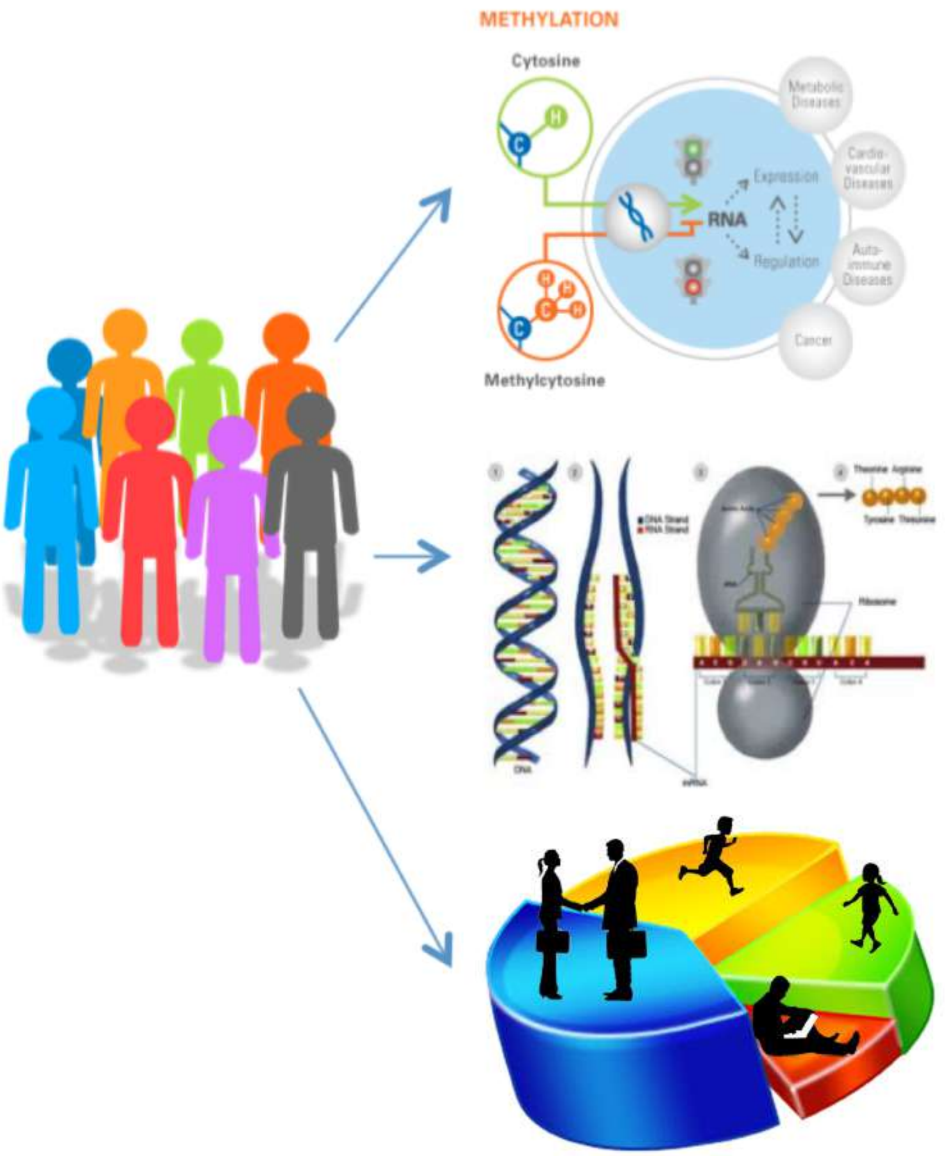
The Cancer Genome Atlas Pan-Cancer analysis project



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

Data integration from multiple heterogeneous sources

How to combine different measurements?



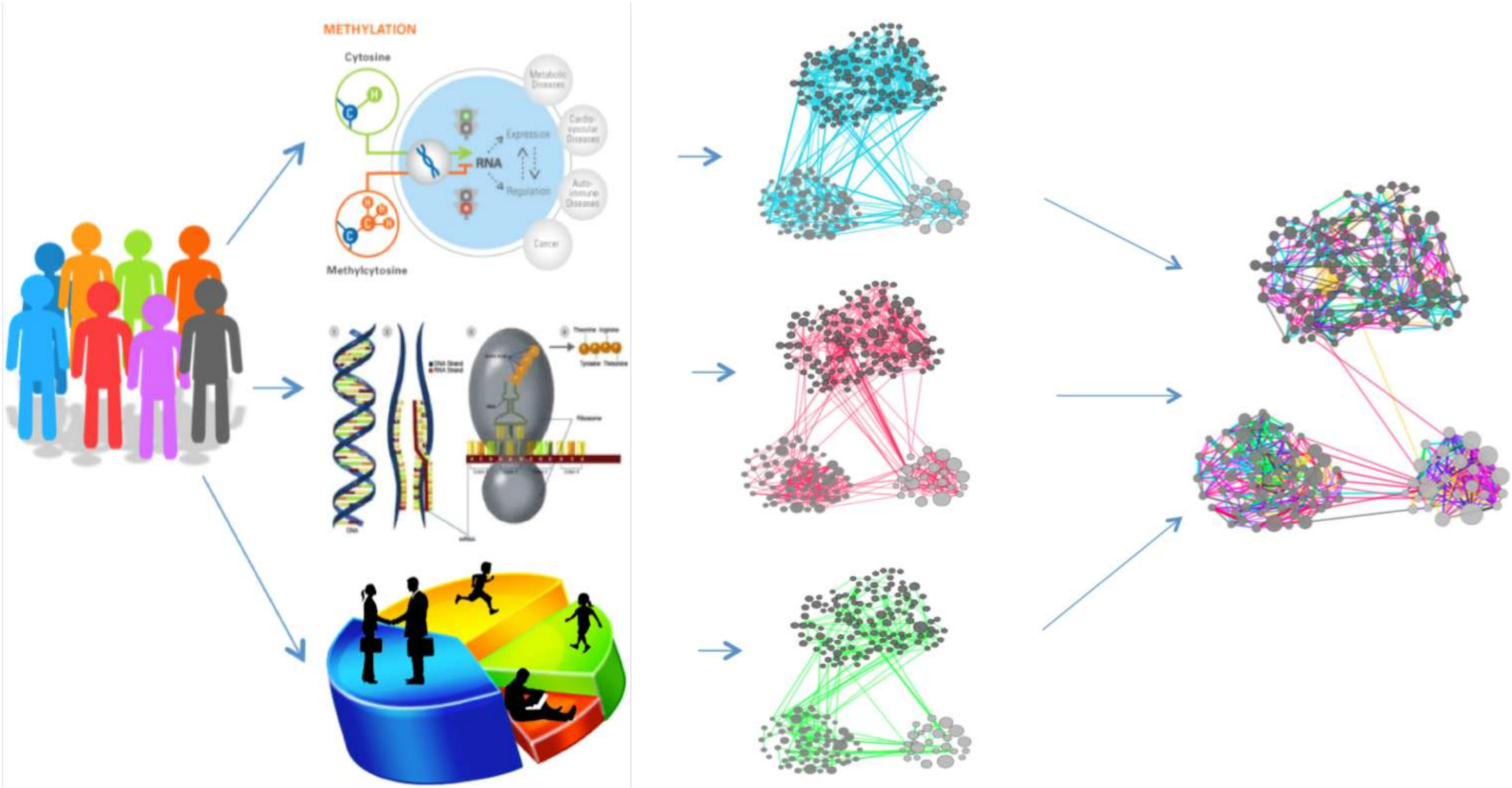
Issues:

- 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

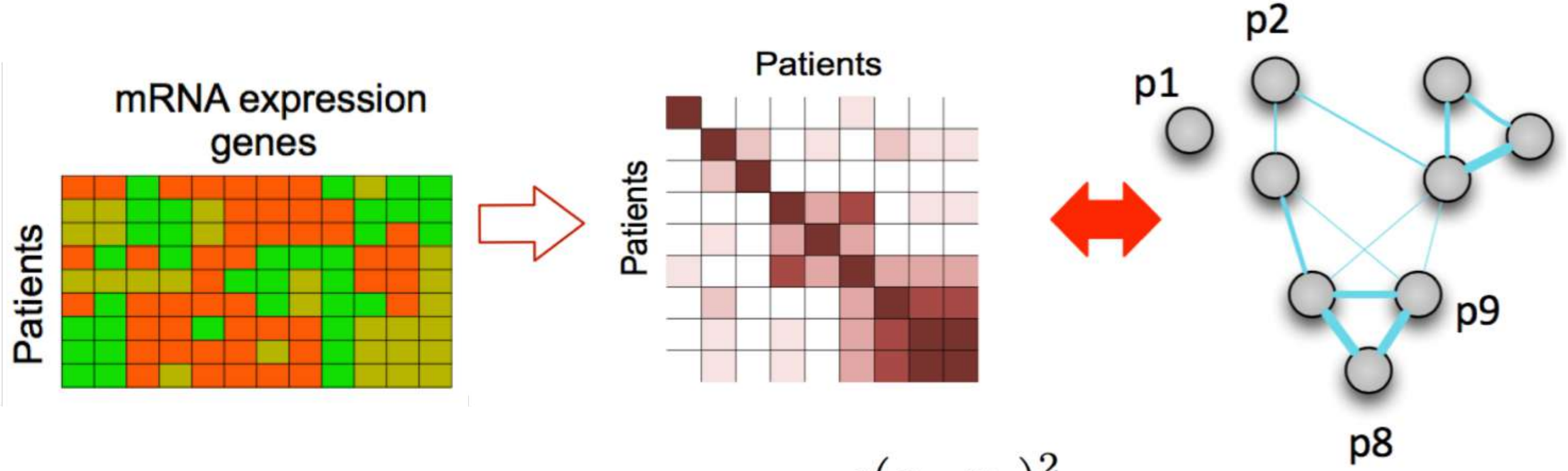
Step 1. Construct a similarity network for each data source

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

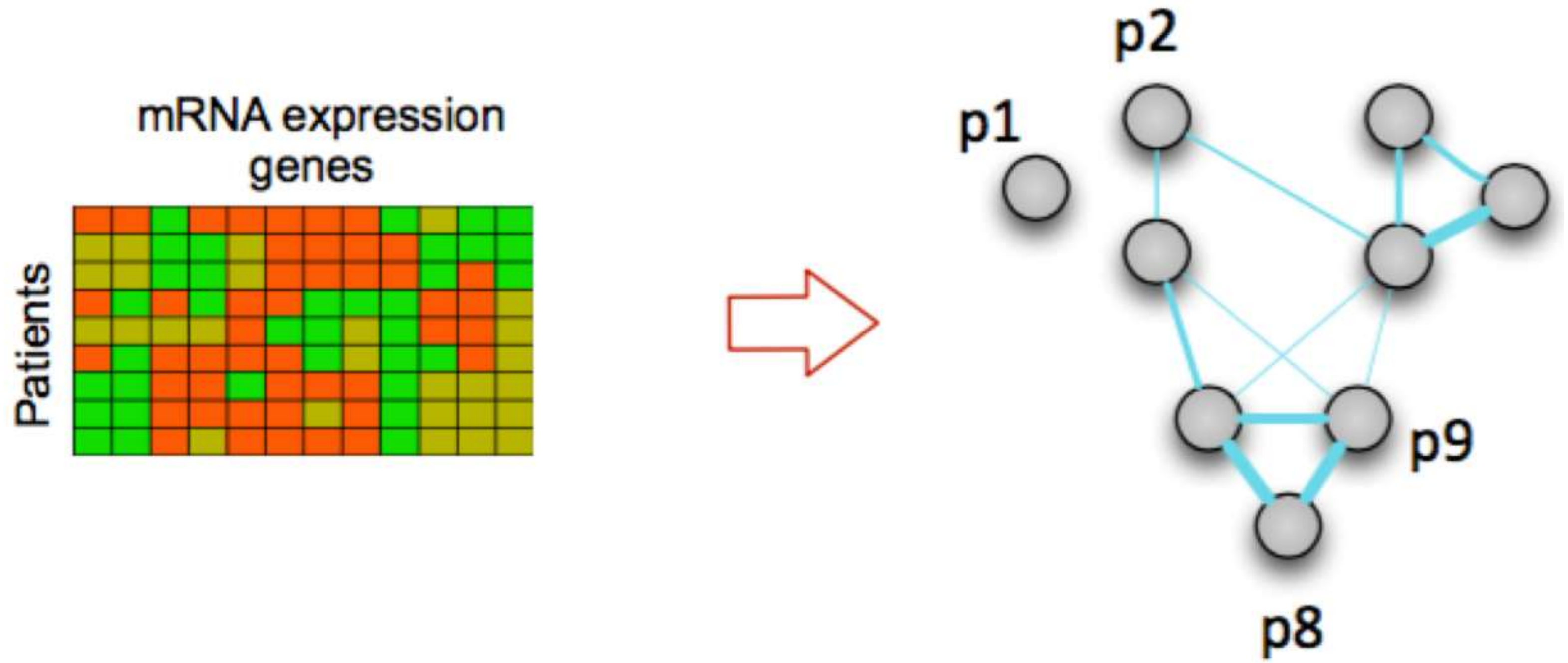
Adjacency matrix: $$P(i,j) = \frac{W(i,j)}{\sum_{k \in V} W(i,k)}$$

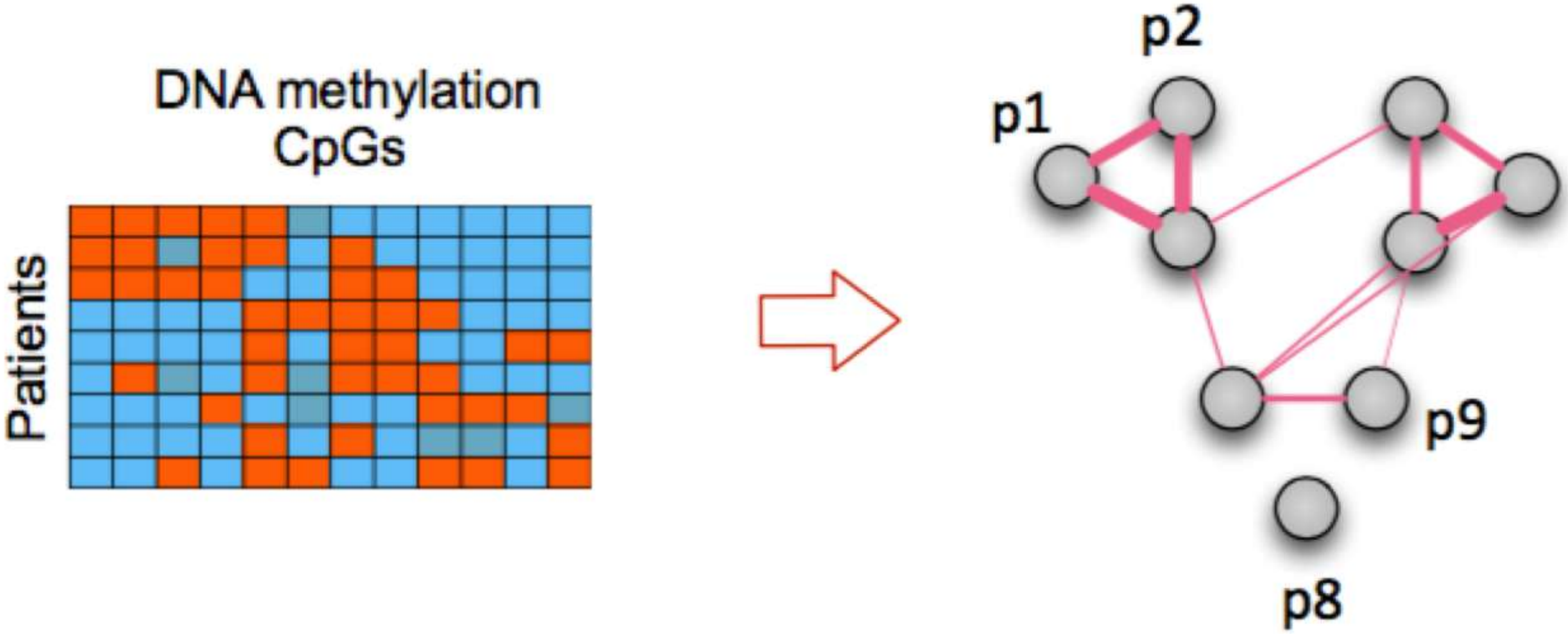
Sparsification

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

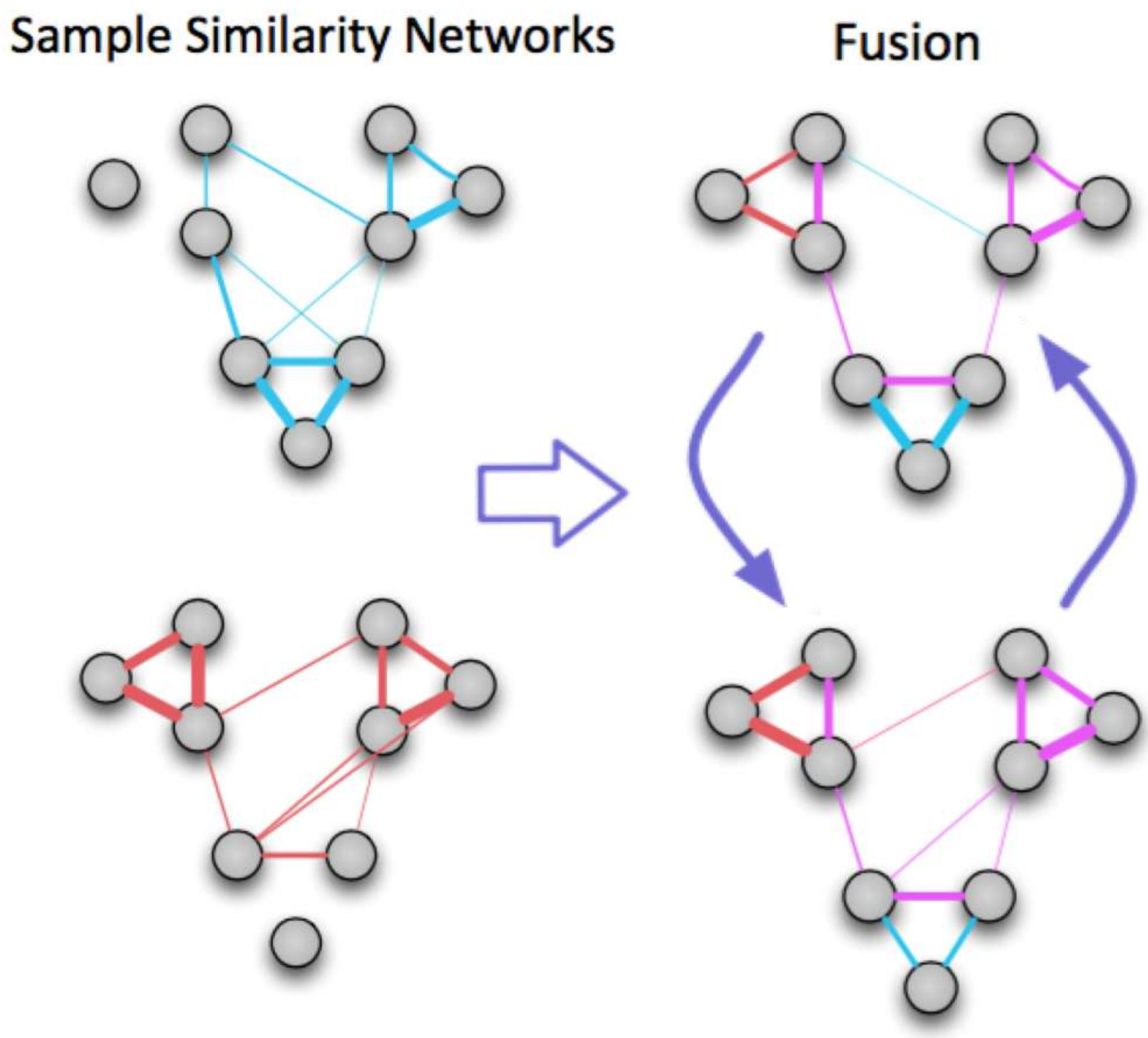
2) $$\mathbf{S}(i,j) = \frac{\mathcal{W}(i,j)}{\sum_{x_k \in KNN(x_i)} \mathcal{W}(i,k)}$$

Construct similarity networks (2)





Combine networks (1)



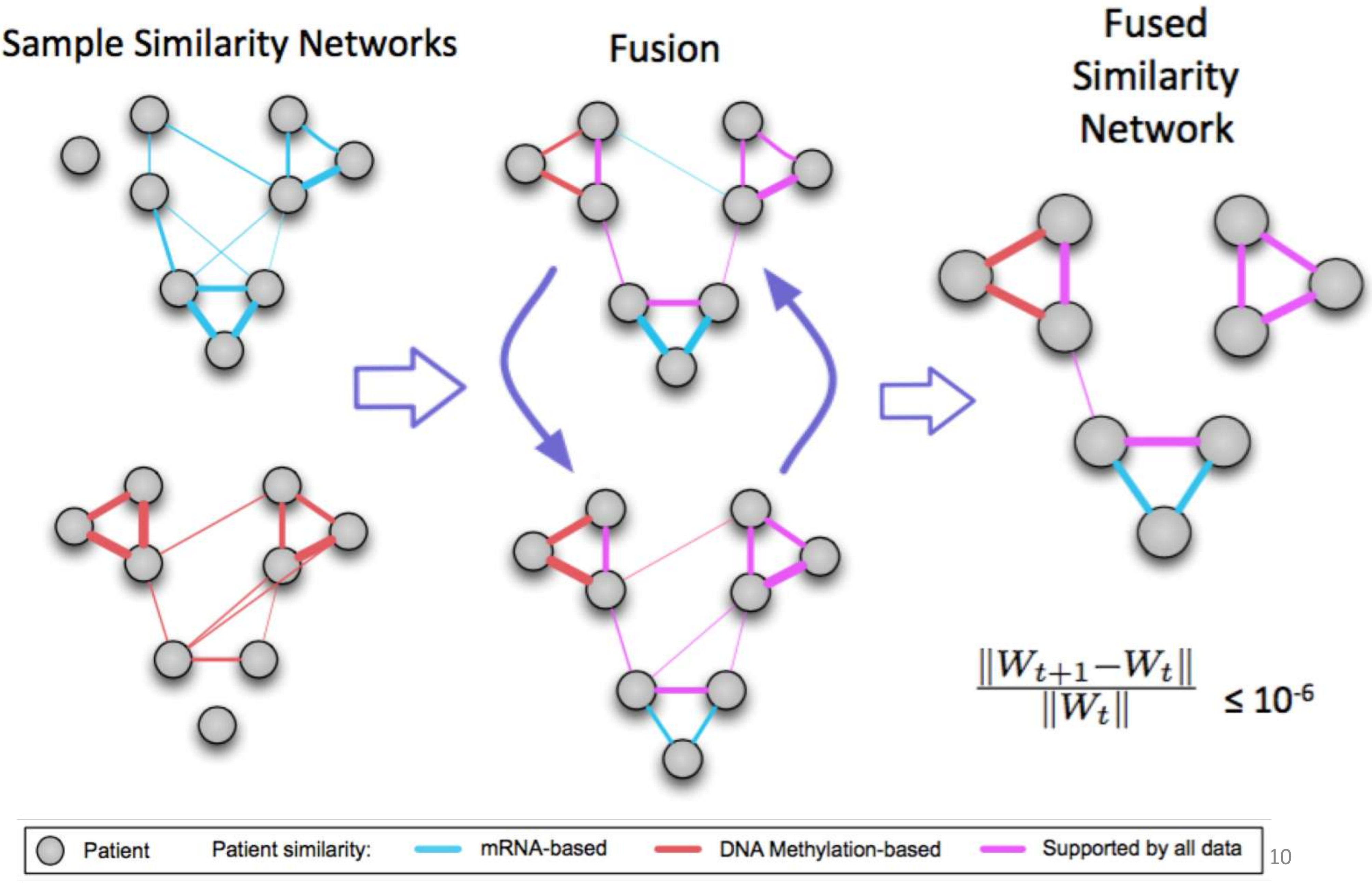
$$\mathbf{P}^{(1)}_{t+1} = \mathbf{S}^{(1)} \times \mathbf{P}^{(2)}_{t} \times (\mathbf{S}^{(1)})^{T}$$

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

Can also be extended to more than 2 data types

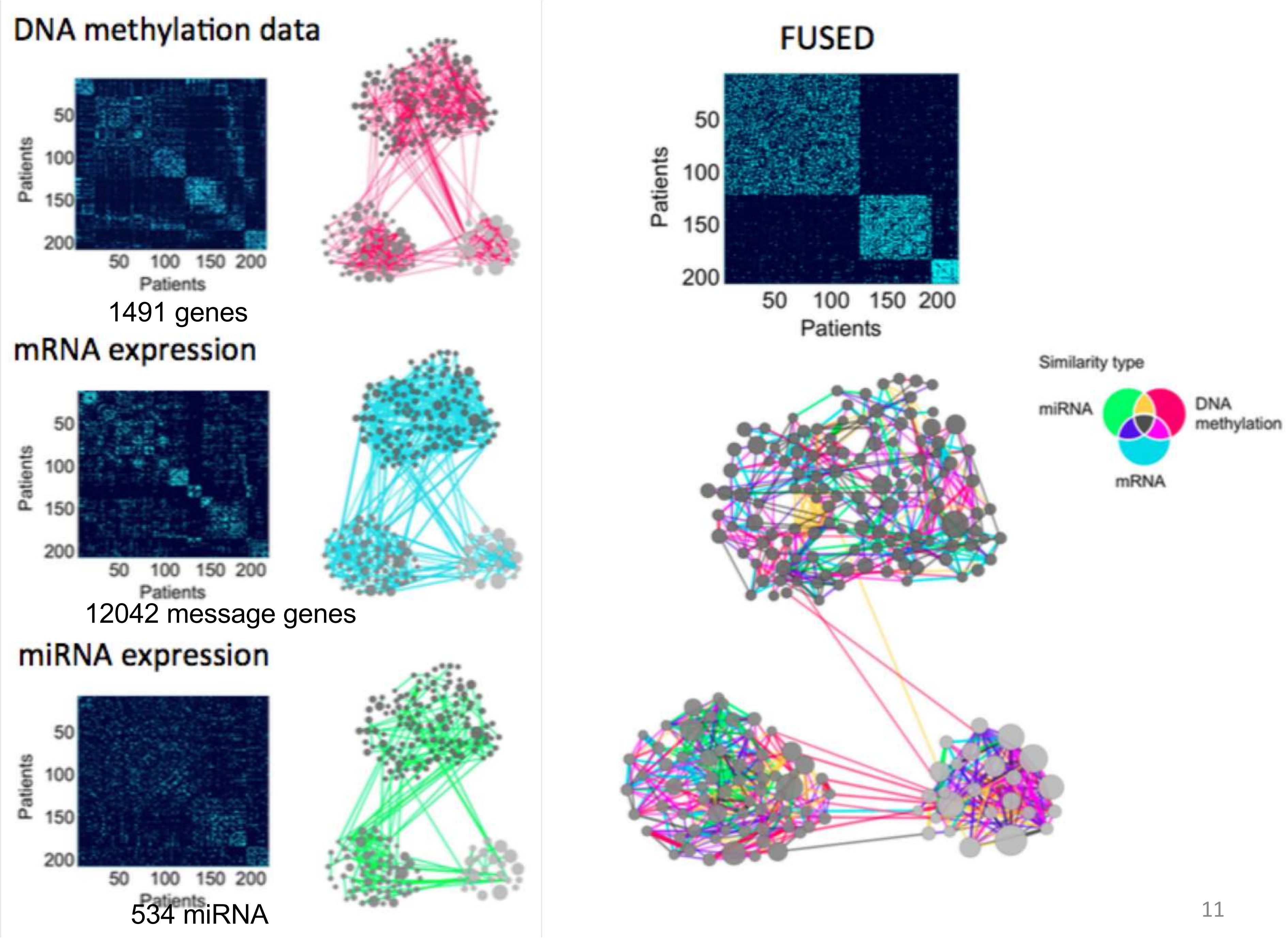
○ Patient Patient similarity: ▬ mRNA-based ▬ DNA Methylation-based ▬ Supported by all data

Combine networks (2)



$$\frac{\|W_{t+1}-W_t\|}{\|W_t\|} \leq 10^{-6}$$

Case study: glioblastoma multiforme (GBM)

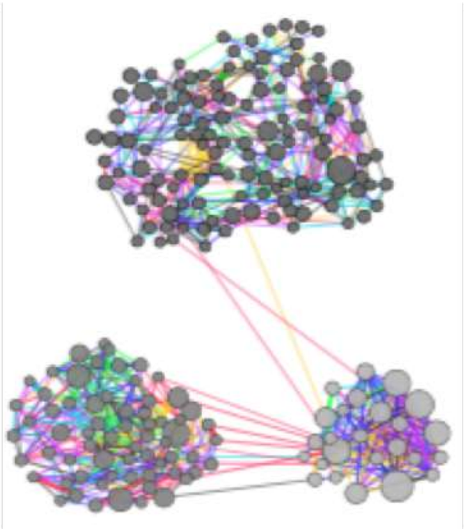


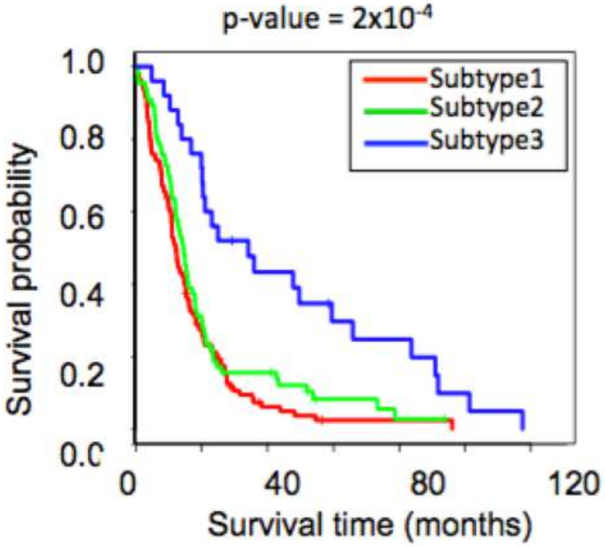
1491 genes

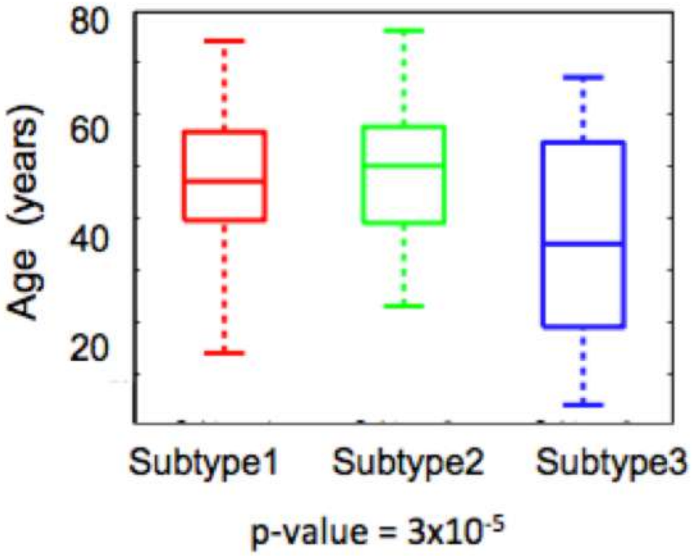
12042 message genes

534 miRNA

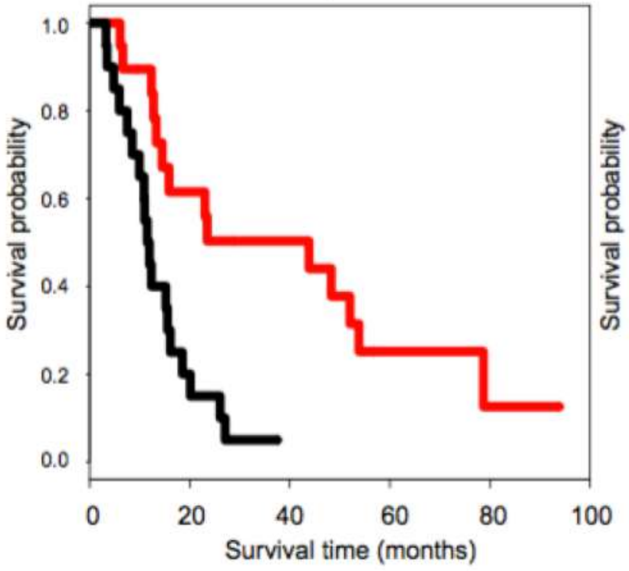
Clinical properties of the subtypes



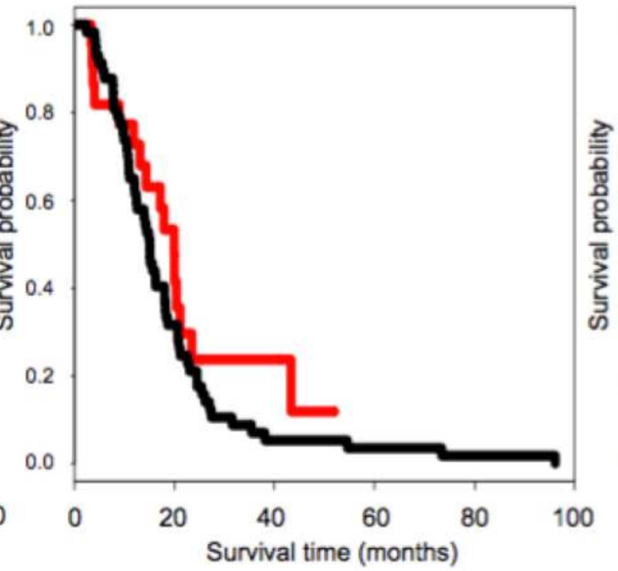




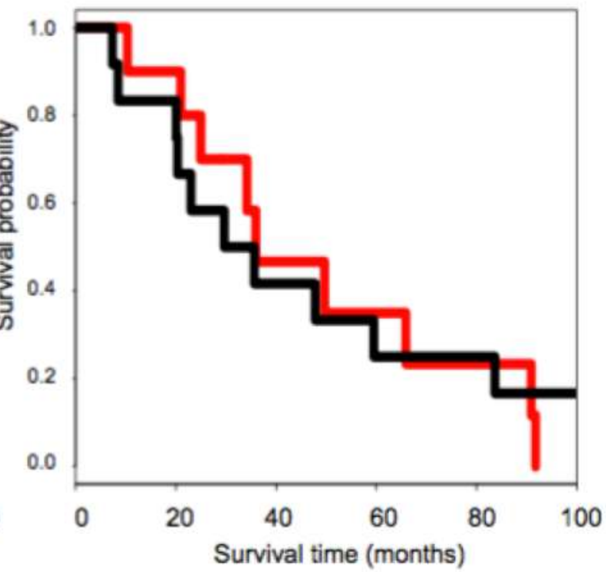
Subtype 1



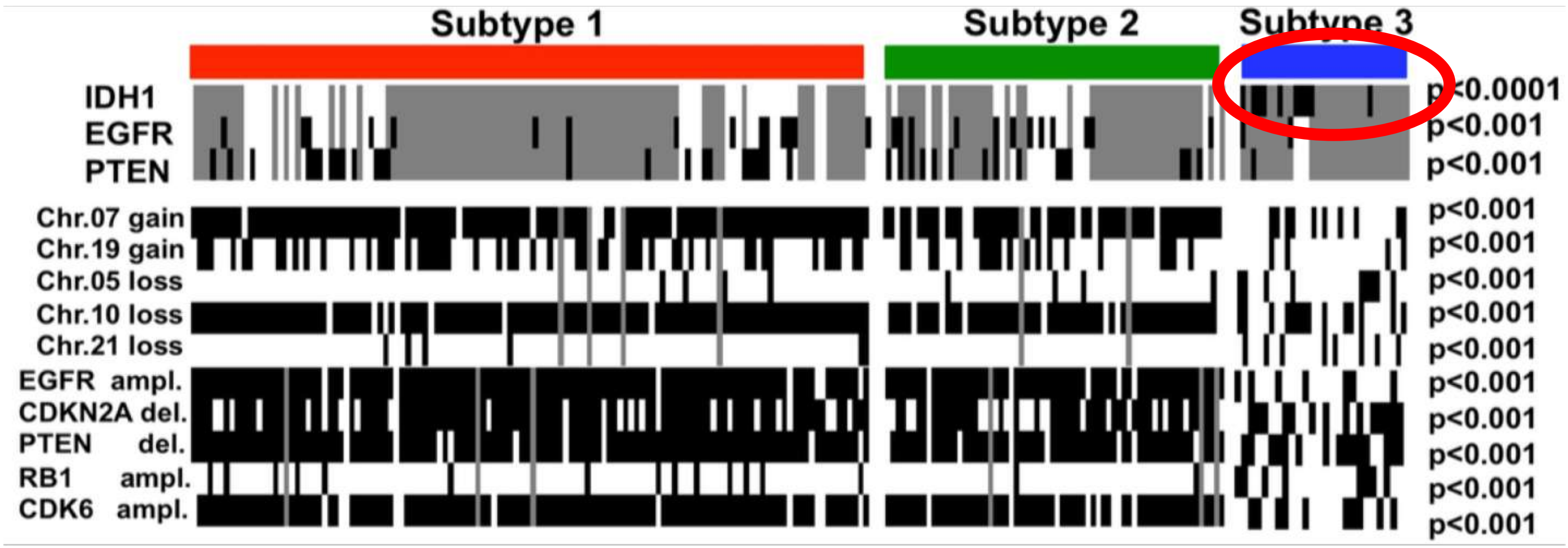
Subtype 2

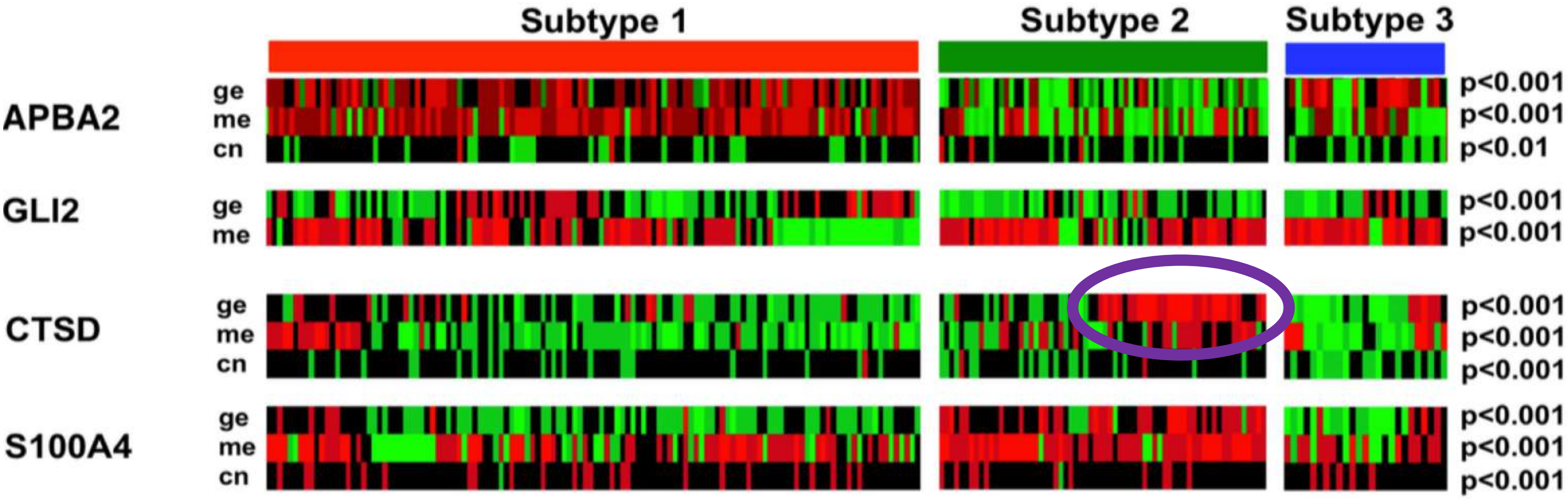


Subtype 3

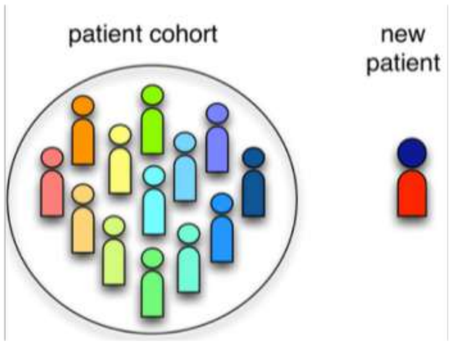


Biological characterization of the subtypes

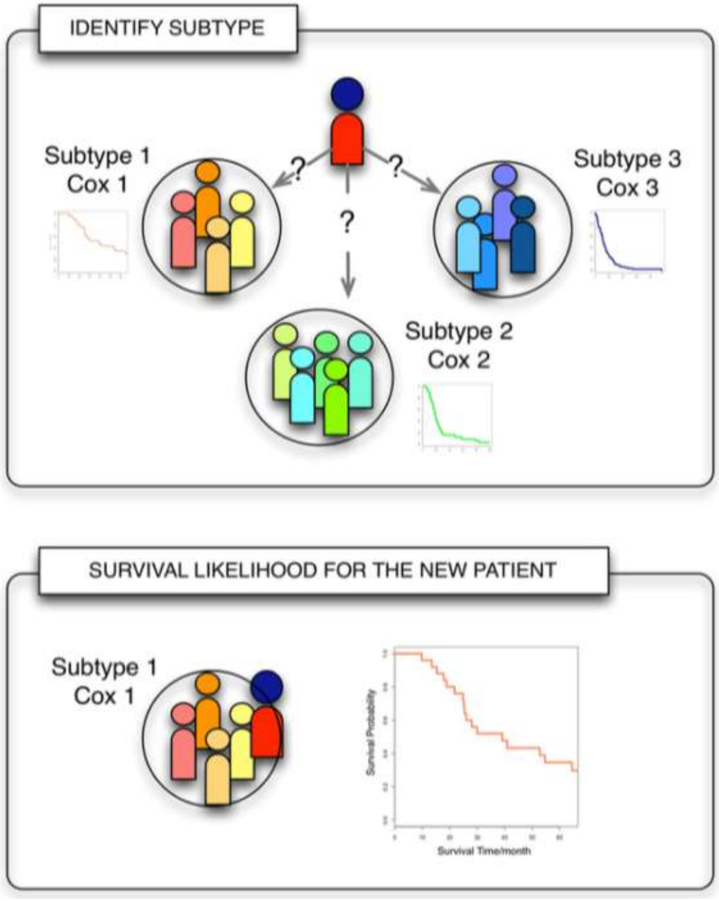




From subtype-based to network-based outcome prediction

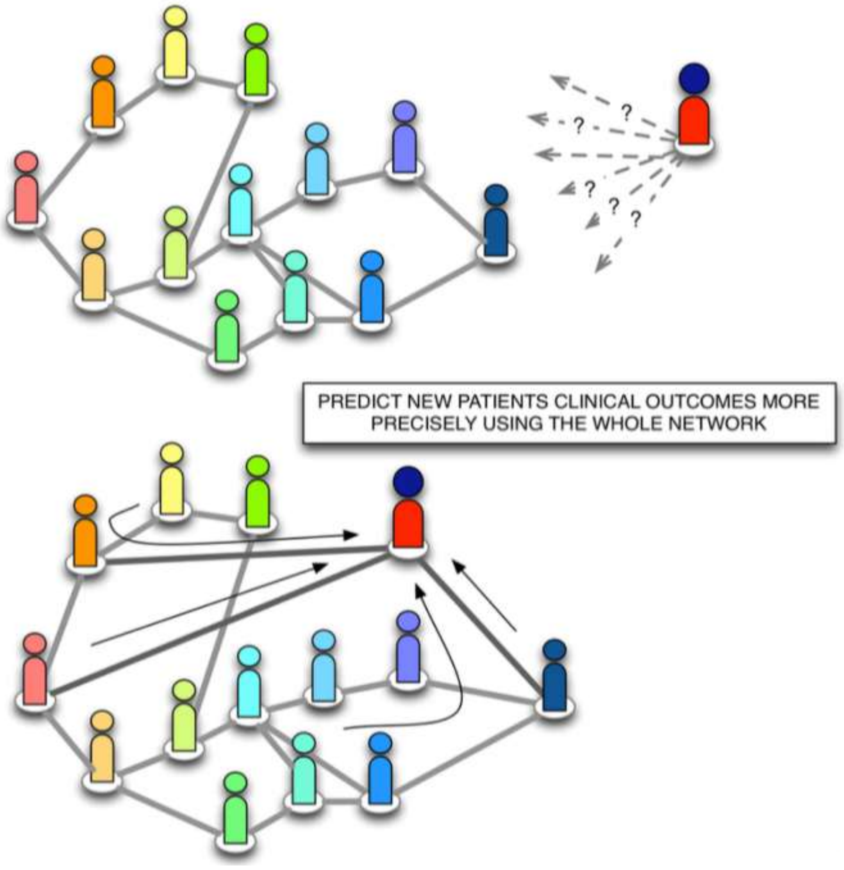


Current Analytics


IDENTIFY SUBTYPE

SURVIVAL LIKELIHOOD FOR THE NEW PATIENT

Future Analytics



PREDICT NEW PATIENTS CLINICAL OUTCOMES MORE PRECISELY USING THE WHOLE NETWORK

Comparisons on an METABRIC breast cancer data

Cox objective

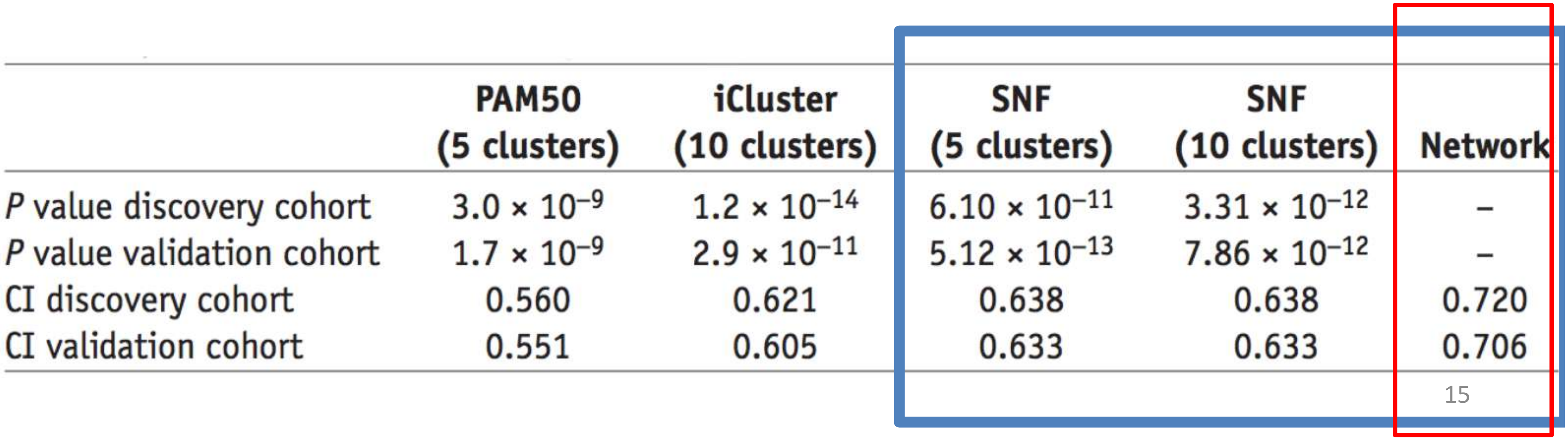
$$lp(z)=\sum_{i=1}^{n}\delta_i\left(\mathbf{X}_i^{T}z-\log\left(\sum_{j\in R(t_i)}\exp\left(\mathbf{X}_j^{T}z\right)\right)\right)$$

Network-regularized objective

Incorporate fused patient network structure

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

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

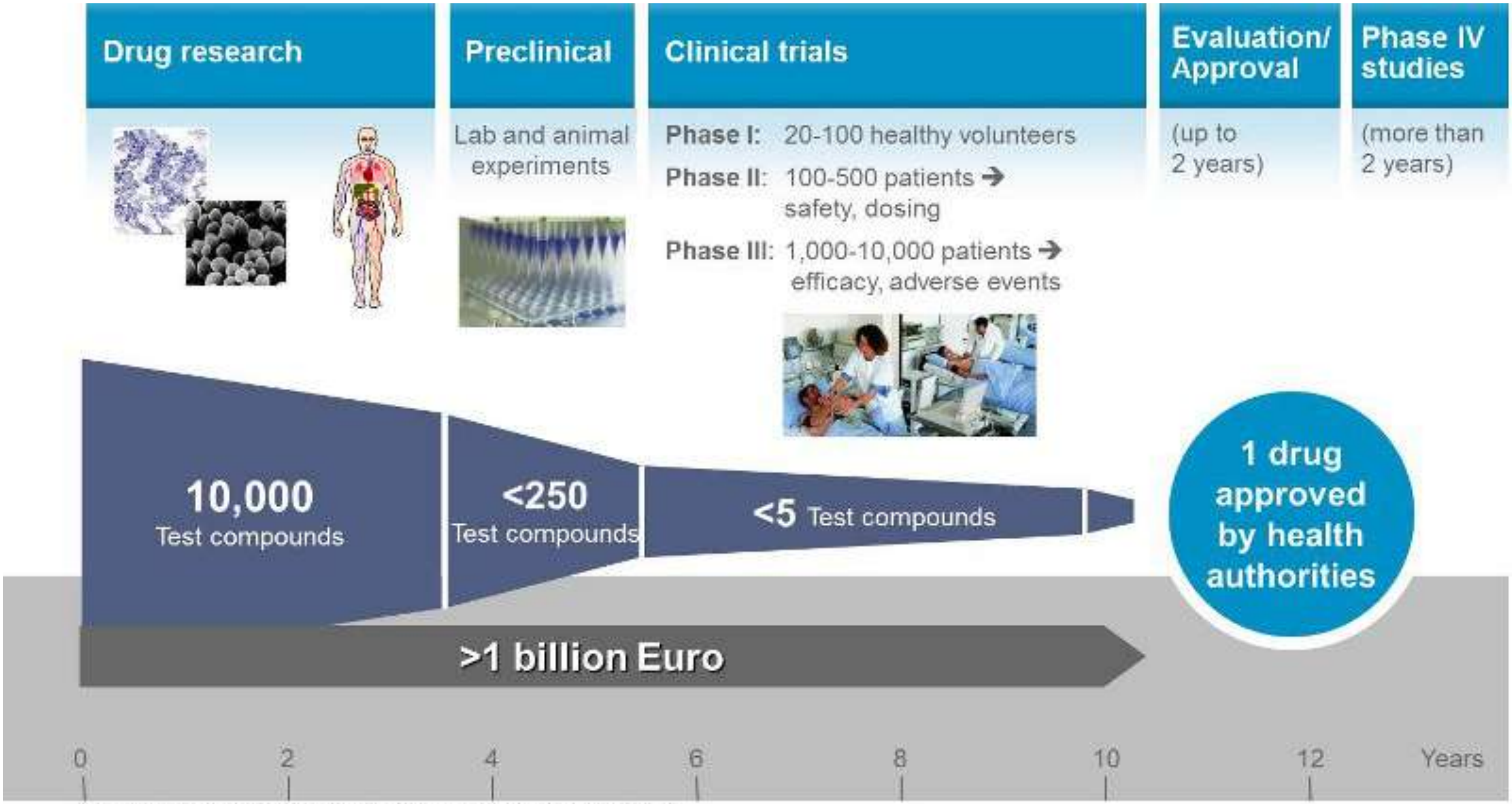
- Creates a unified view of patients based on multiple heterogeneous sources
- Integrates gene and non-gene based data
- Robust to different types of noise
- Obtain superior results on regular tasks such as subtyping and outcome prediction
- Scalable

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
- Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping

The Challenge of Drug Discovery



Drug research

Preclinical

Lab and animal experiments

Clinical trials

Phase I: 20-100 healthy volunteers

Phase II: 100-500 patients → safety, dosing

Phase III: 1,000-10,000 patients → efficacy, adverse events

Evaluation/ Approval

(up to 2 years)

Phase IV studies

(more than 2 years)

10,000 Test compounds

<250 Test compounds

<5 Test compounds

1 drug approved by health authorities

>1 billion Euro

Source: based on PhRMA Profile Pharmaceutical Industry 2010

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.

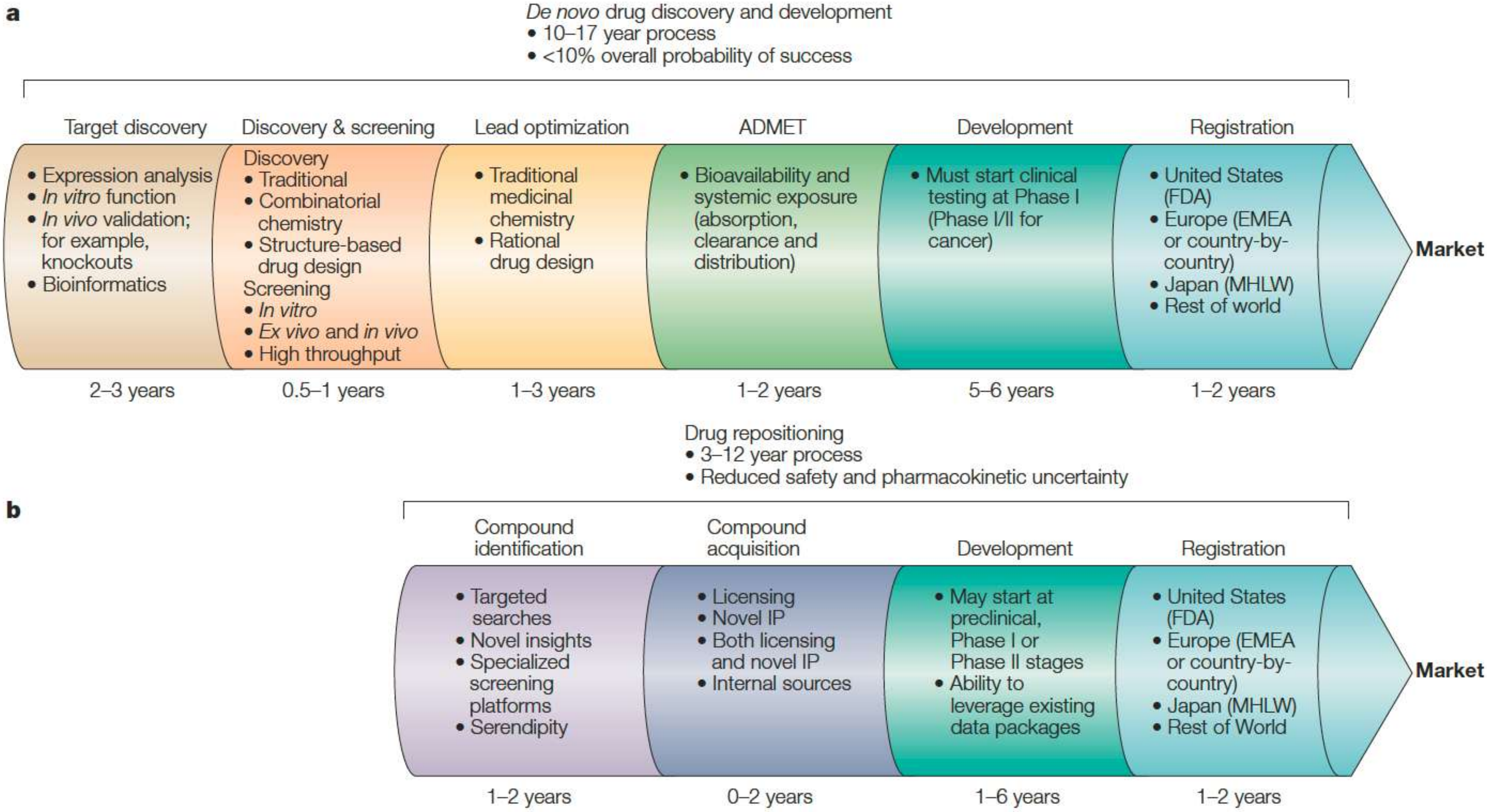
Drug repositioning

- **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



a

De novo drug discovery and development
- 10–17 year process
- <10% overall probability of success

Target discovery
- Expression analysis
- *In vitro* function
- *In vivo* validation; for example, knockouts
- Bioinformatics

2–3 years

Discovery & screening

Discovery
- Traditional
- Combinatorial chemistry
- Structure-based drug design

Screening
- *In vitro*
- *Ex vivo* and *in vivo*
- High throughput

0.5–1 years

Lead optimization
- Traditional medicinal chemistry
- Rational drug design

1–3 years

ADMET
- Bioavailability and systemic exposure (absorption, clearance and distribution)

1–2 years

Development
- Must start clinical testing at Phase I (Phase I/II for cancer)

5–6 years

Registration
- United States (FDA)
- Europe (EMEA or country-by-country)
- Japan (MHLW)
- Rest of world

1–2 years

Market

b

Drug repositioning
- 3–12 year process
- Reduced safety and pharmacokinetic uncertainty

Compound identification
- Targeted searches
- Novel insights
- Specialized screening platforms
- Serendipity

1–2 years

Compound acquisition
- Licensing
- Novel IP
- Both licensing and novel IP
- Internal sources

0–2 years

Development
- May start at preclinical, Phase I or Phase II stages
- Ability to leverage existing data packages

1–6 years

Registration
- United States (FDA)
- Europe (EMEA or country-by-country)
- Japan (MHLW)
- Rest of World

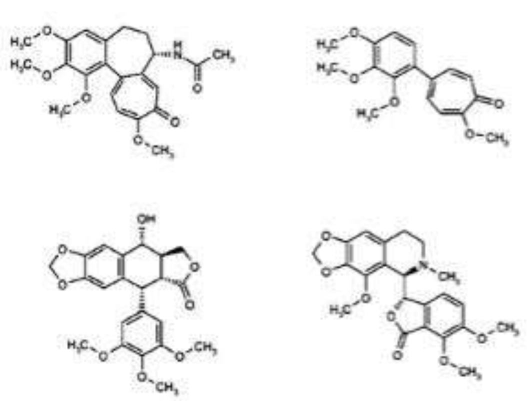
1–2 years

Market

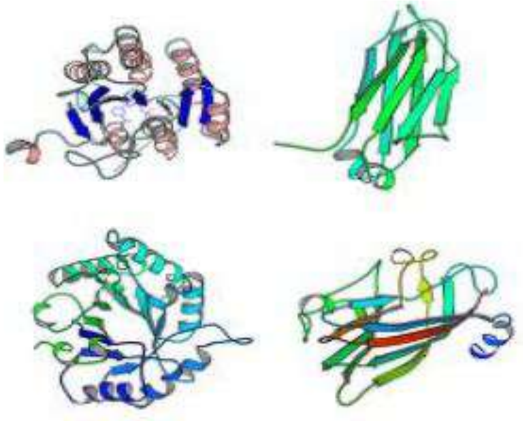
Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673-683, 2004.

Drug Resources and Disease Resources

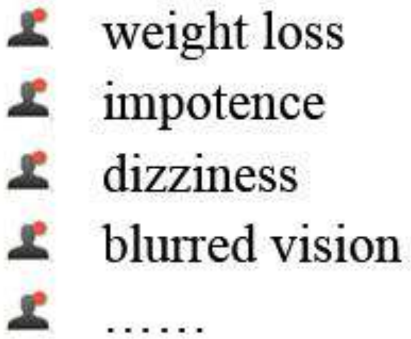
Drug



Chemical Structure



Target Proteins



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

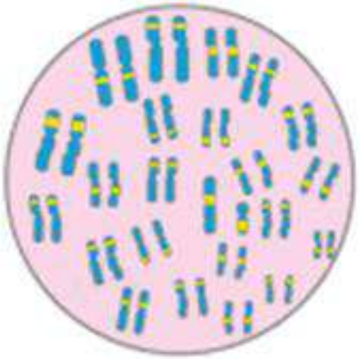
Side-effect Keywords

Calculate drug/disease similarities

Disease

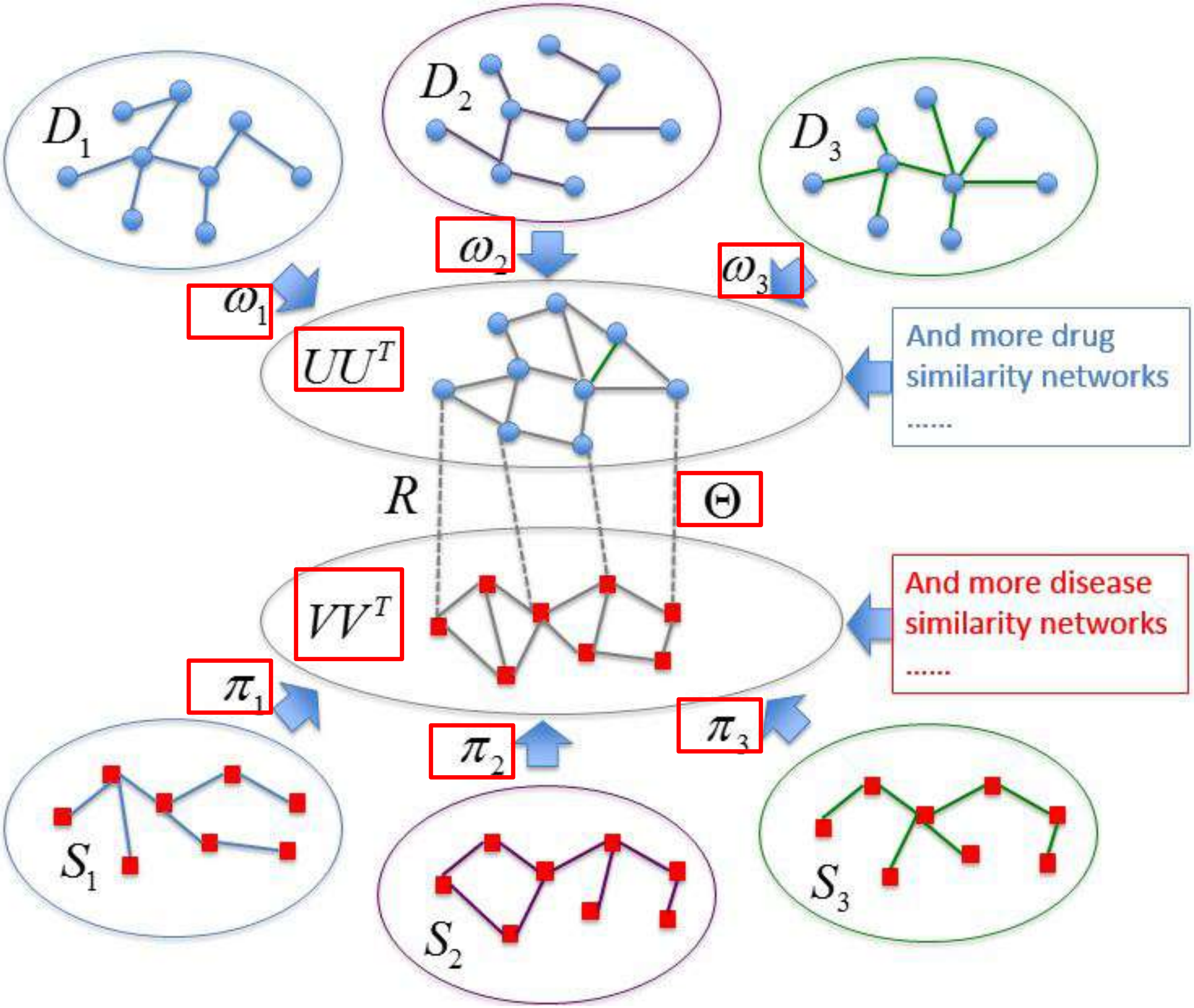
Phenotype/Symptom

Ontology



Disease Gene

Joint Matrix Factorization (JMF)

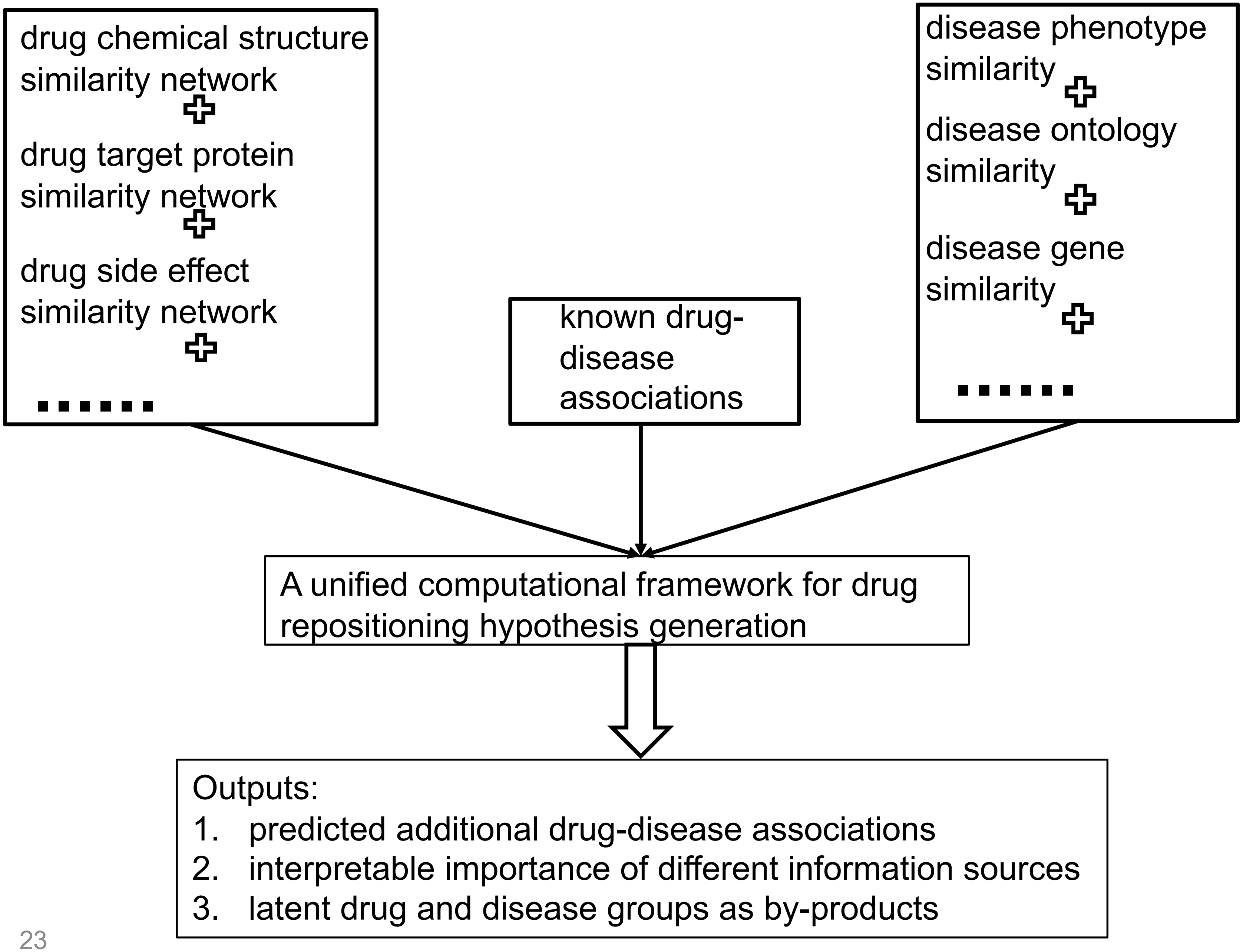


And more drug similarity networks
......

And more disease similarity networks
......

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



drug chemical structure similarity network
+
drug target protein similarity network
+
drug side effect similarity network
+
....

known drug-disease associations

disease phenotype similarity
+
disease ontology similarity
+
disease gene similarity
+
....

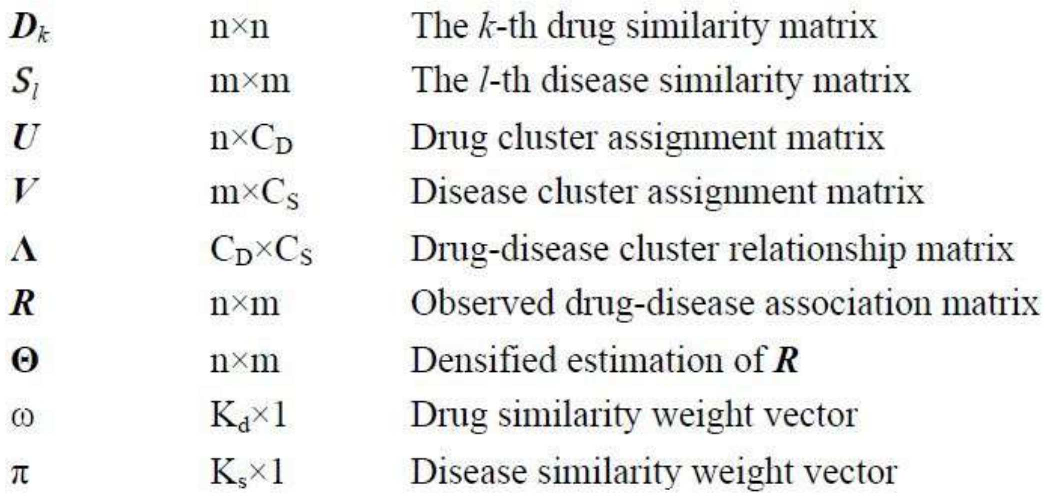
A unified computational framework for drug repositioning hypothesis generation

Outputs:

1. predicted additional drug-disease associations
2. interpretable importance of different information sources
3. latent drug and disease groups as by-products

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 ***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:

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

- The reconstruction loss of disease similarities:

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

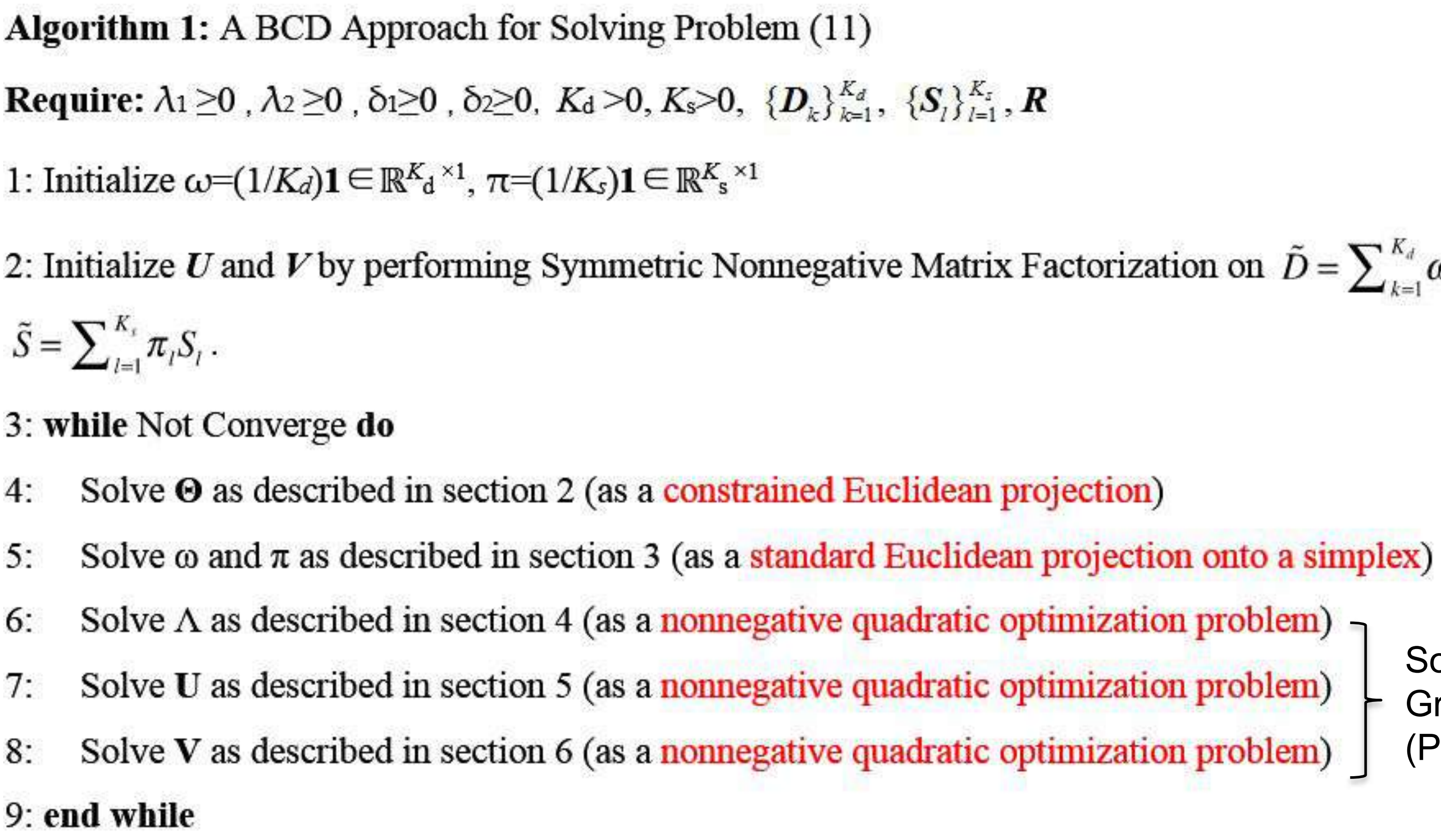
Reconstruct integrated drug/disease networks

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

```
Initialize $\omega = (1/K_d)\mathbf{1} \in \mathbb{R}^{K_d \times 1}$, $\pi = (1/K_s)\mathbf{1} \in \mathbb{R}^{K_s \times 1}$
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 $\tilde{S} = \sum_{l=1}^{K_s} \pi_l S_l$ .
while Not Converge do
    Solve $\boldsymbol{\Theta}$ as described in section 2 (as a constrained Euclidean projection)
    Solve $\omega$ and $\pi$ as described in section 3 (as a standard Euclidean projection onto a simplex)
    Solve $\Lambda$ as described in section 4 (as a nonnegative quadratic optimization problem)
    Solve $\mathbf{U}$ as described in section 5 (as a nonnegative quadratic optimization problem)
    Solve $\mathbf{V}$ as described in section 6 (as a nonnegative quadratic optimization problem)
end while
```

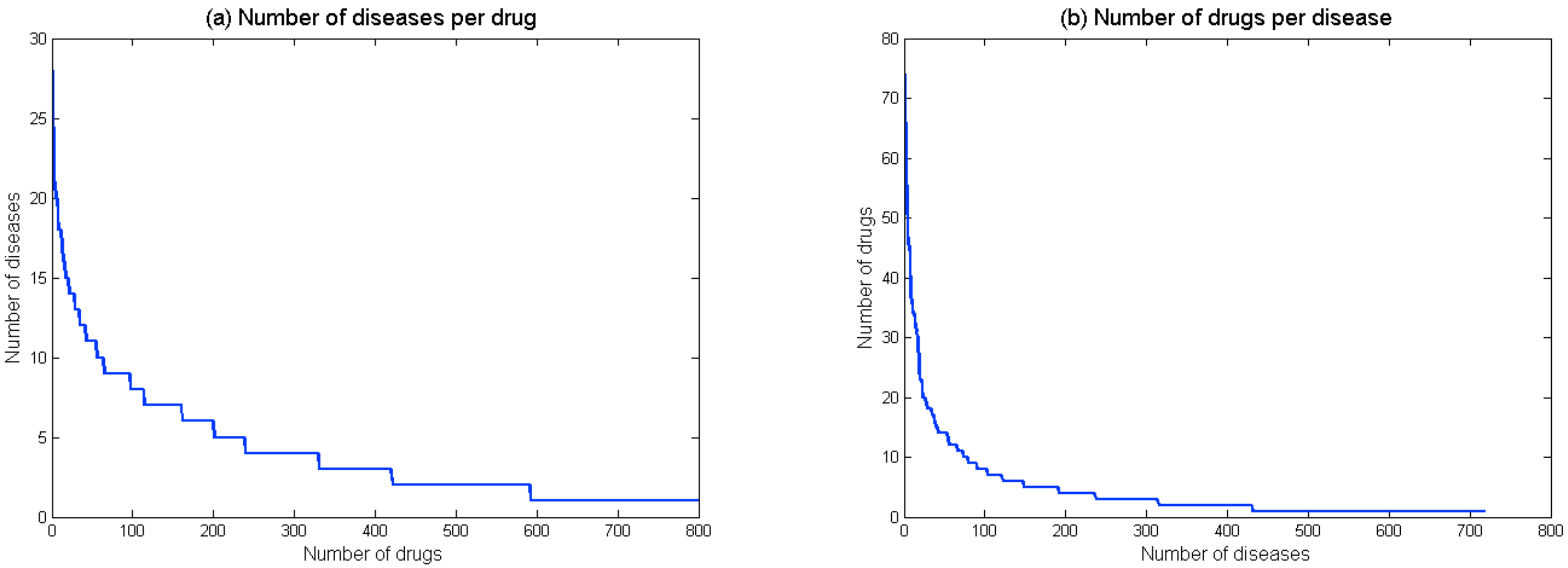
Closed-form solution

Solved by Projected Gradient Descent (PGD) method

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

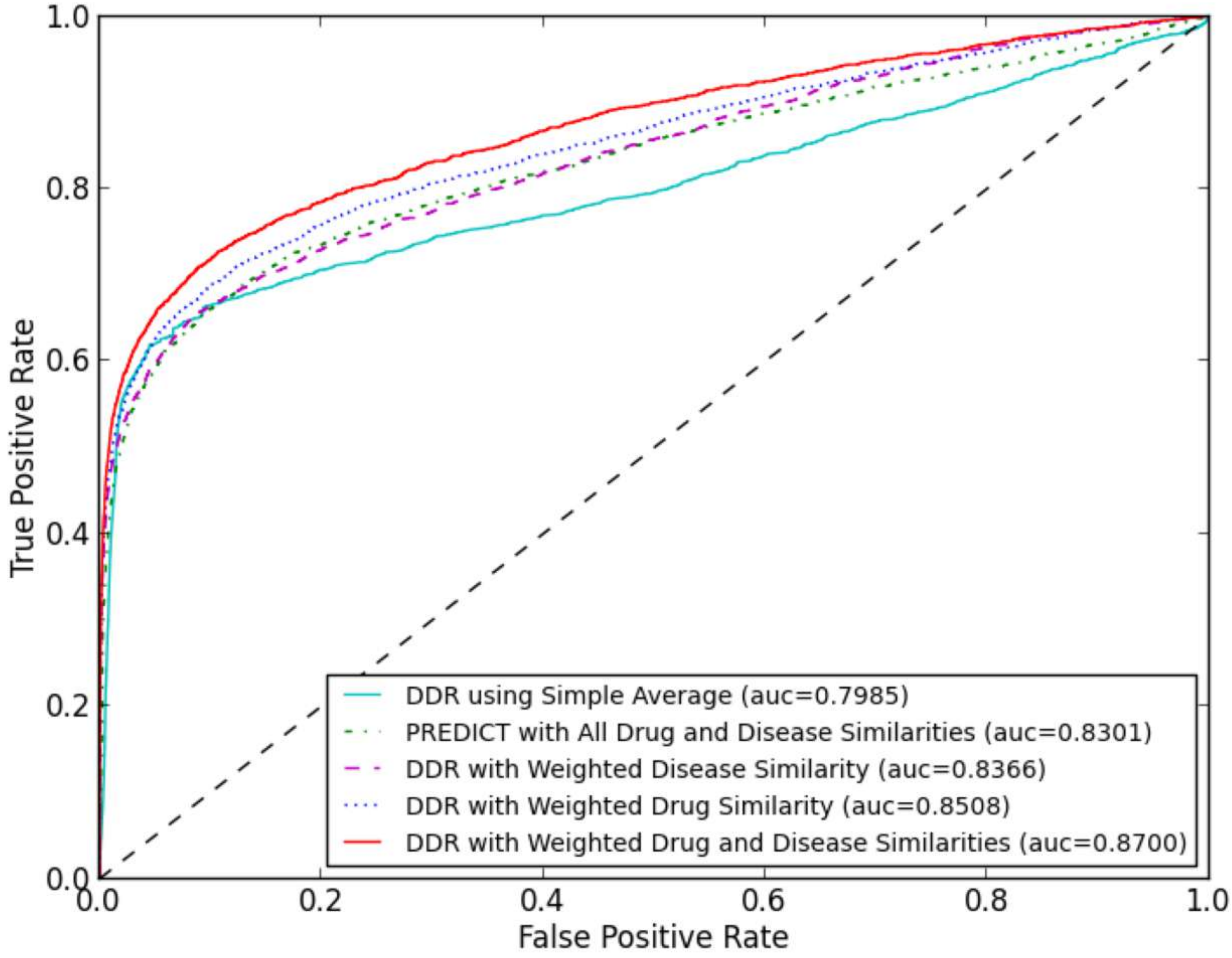
- Benchmark dataset was extracted from NDF-RT, spanning 3,250 treatment associations between 799 drugs and 719 diseases
- **Three** 799×799 matrices were used to represent drug similarities between 799 drugs from different perspectives
- **Three** 719×719 matrices were used to represent disease similarities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



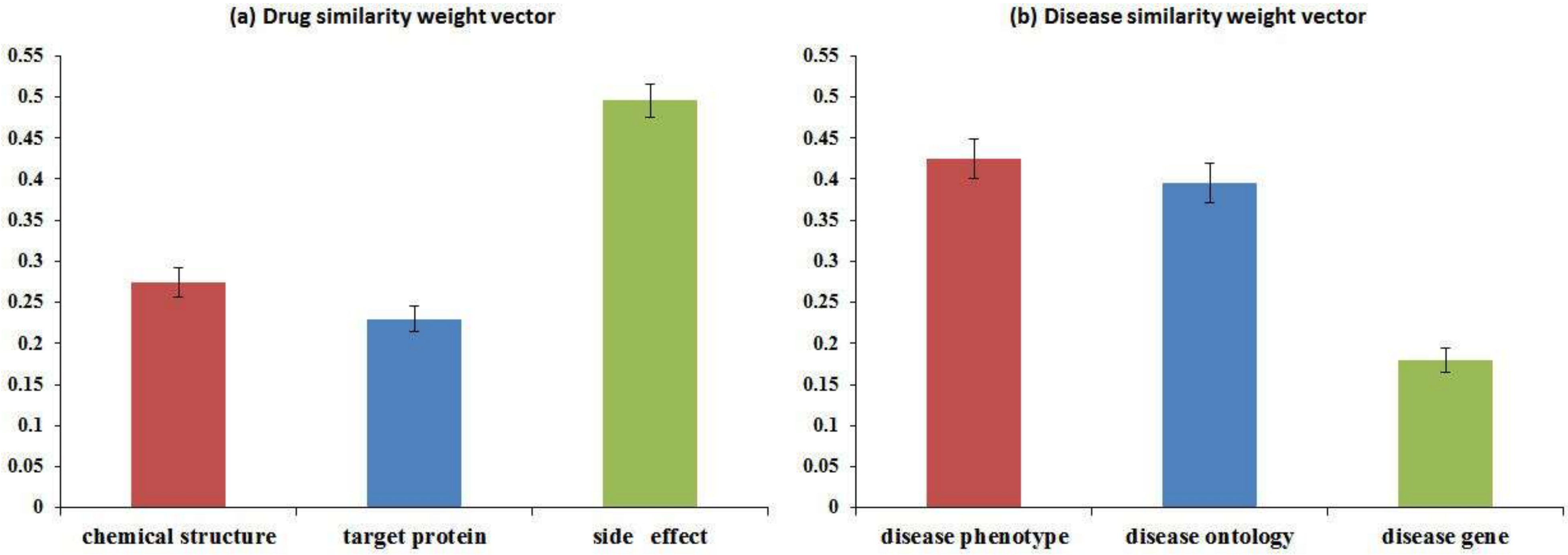
(a) Number of diseases per drug

(b) Number of drugs per disease

ROC comparisons of five drug repositioning approaches



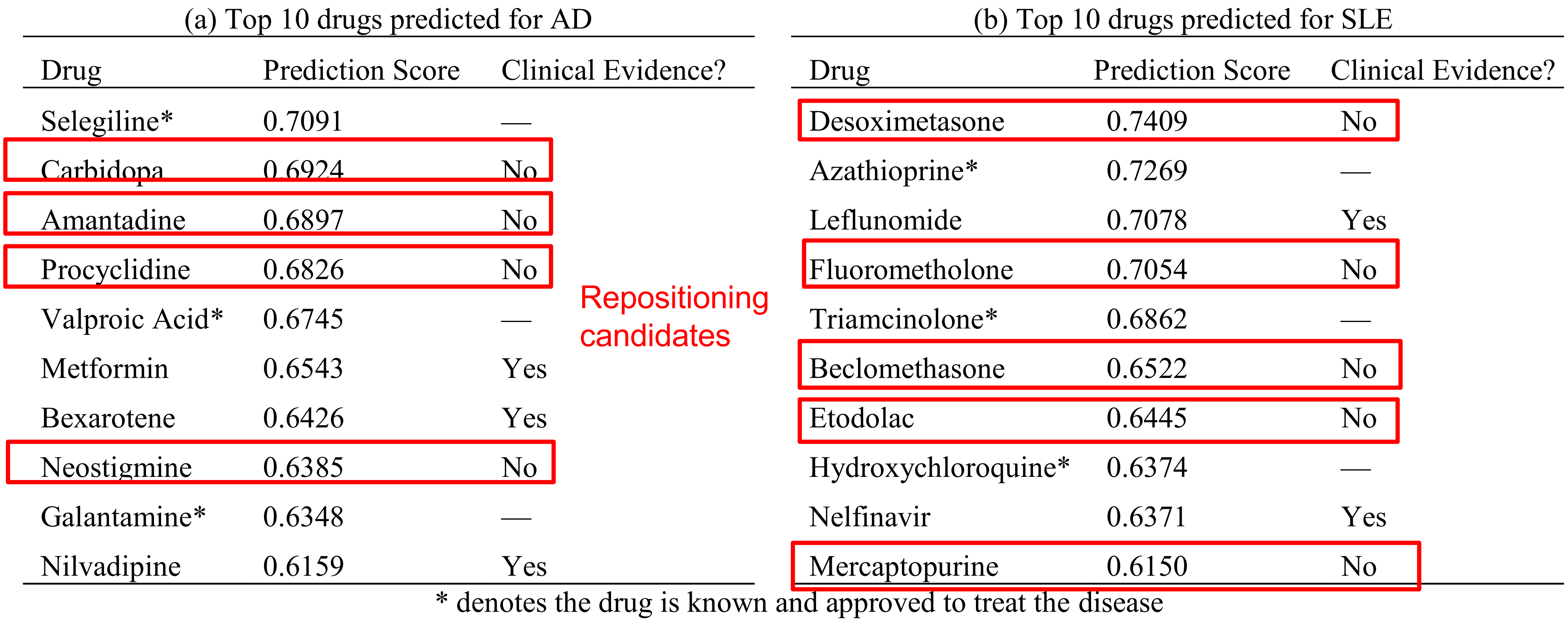
Distribution of weights of the similarity weight vectors obtained by JMF



(a) Drug similarity weight vector

(b) Disease similarity weight vector

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

(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

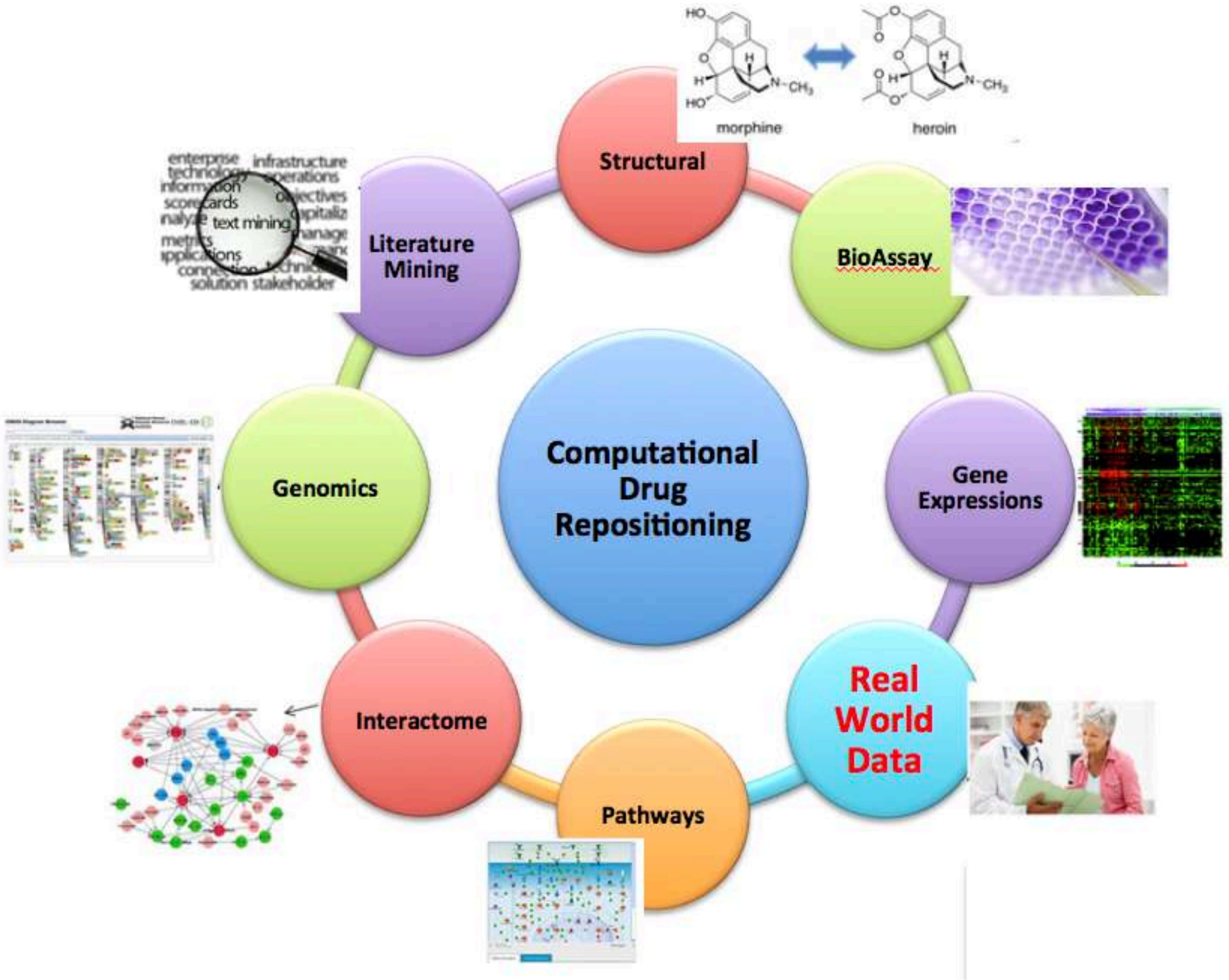
Repositioning candidates

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.

Next: Multi-channel detailed computational hypothesis generation



Computational Drug Repositioning

Structural

BioAssay

Gene Expressions

Real World Data

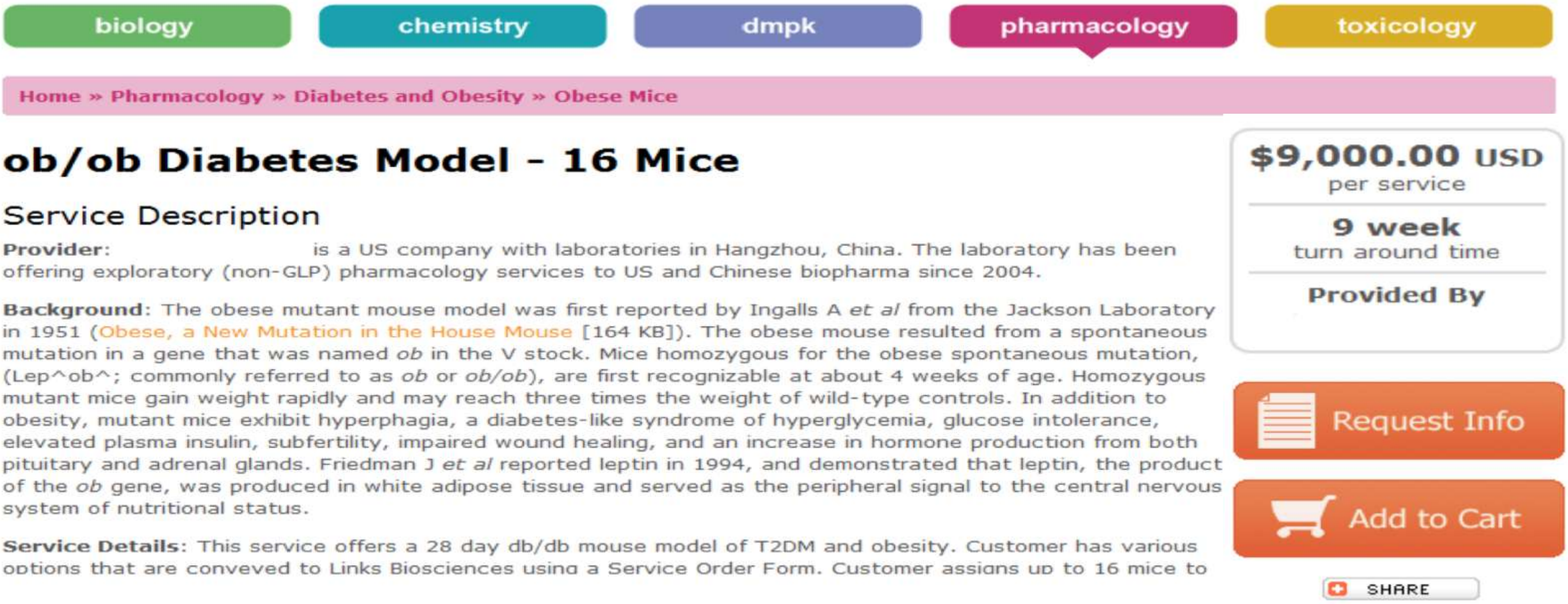
Pathways

Interactome

Genomics

Literature Mining

And even beyond the hypothesis generation…



Home » Pharmacology » Diabetes and Obesity » Obese Mice

ob/ob Diabetes Model - 16 Mice

$9,000.00 USD per service

9 week turn around time

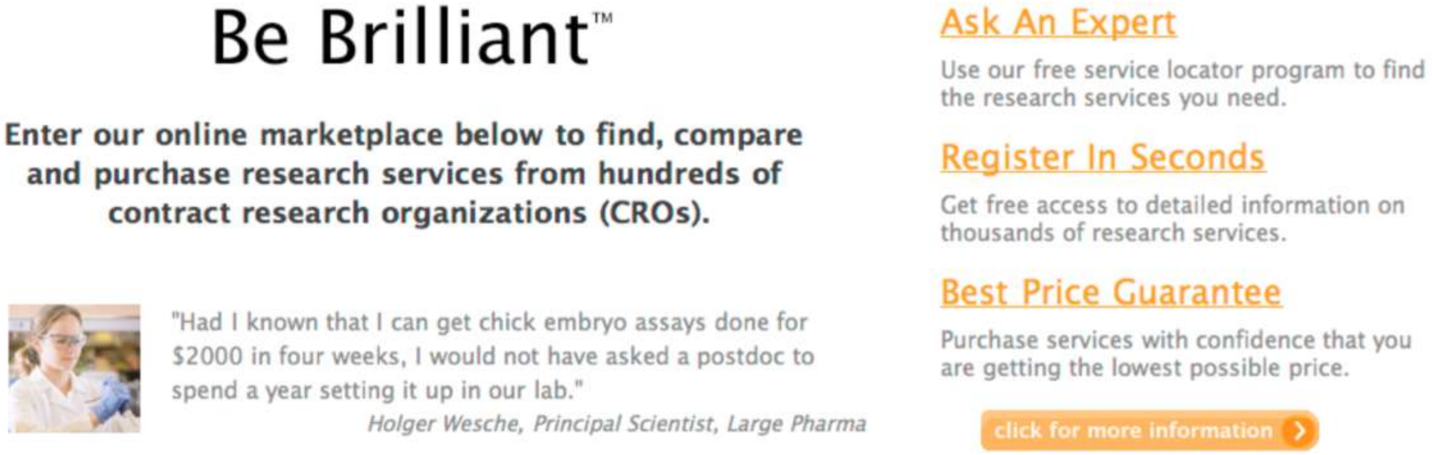
Provided By

Service Description

Provider: is a US company with laboratories in Hangzhou, China. The laboratory has been offering exploratory (non-GLP) pharmacology services to US and Chinese biopharma since 2004.

Background: The obese mutant mouse model was first reported by Ingalls A *et al* from the Jackson Laboratory in 1951 (Obese, a New Mutation in the House Mouse [164 KB]). The obese mouse resulted from a spontaneous mutation in a gene that was named *ob* in the V stock. Mice homozygous for the obese spontaneous mutation, (Lep^ob^; commonly referred to as *ob* or *ob/ob*), are first recognizable at about 4 weeks of age. Homozygous mutant mice gain weight rapidly and may reach three times the weight of wild-type controls. In addition to obesity, mutant mice exhibit hyperphagia, a diabetes-like syndrome of hyperglycemia, glucose intolerance, elevated plasma insulin, subfertility, impaired wound healing, and an increase in hormone production from both pituitary and adrenal glands. Friedman J *et al* reported leptin in 1994, and demonstrated that leptin, the product of the *ob* gene, was produced in white adipose tissue and served as the peripheral signal to the central nervous system of nutritional status.

Service Details: This service offers a 28 day db/db mouse model of T2DM and obesity. Customer has various options that are conveyed to Links Biosciences using a Service Order Form. Customer assigns up to 16 mice to



Be Brilliant™

Enter our online marketplace below to find, compare and purchase research services from hundreds of contract research organizations (CROs).

"Had I known that I can get chick embryo assays done for $2000 in four weeks, I would not have asked a postdoc to spend a year setting it up in our lab."

Holger Wesche, Principal Scientist, Large Pharma

Ask An Expert

Use our free service locator program to find the research services you need.

Register In Seconds

Get free access to detailed information on thousands of research services.

Best Price Guarantee

Purchase services with confidence that you are getting the lowest possible price.

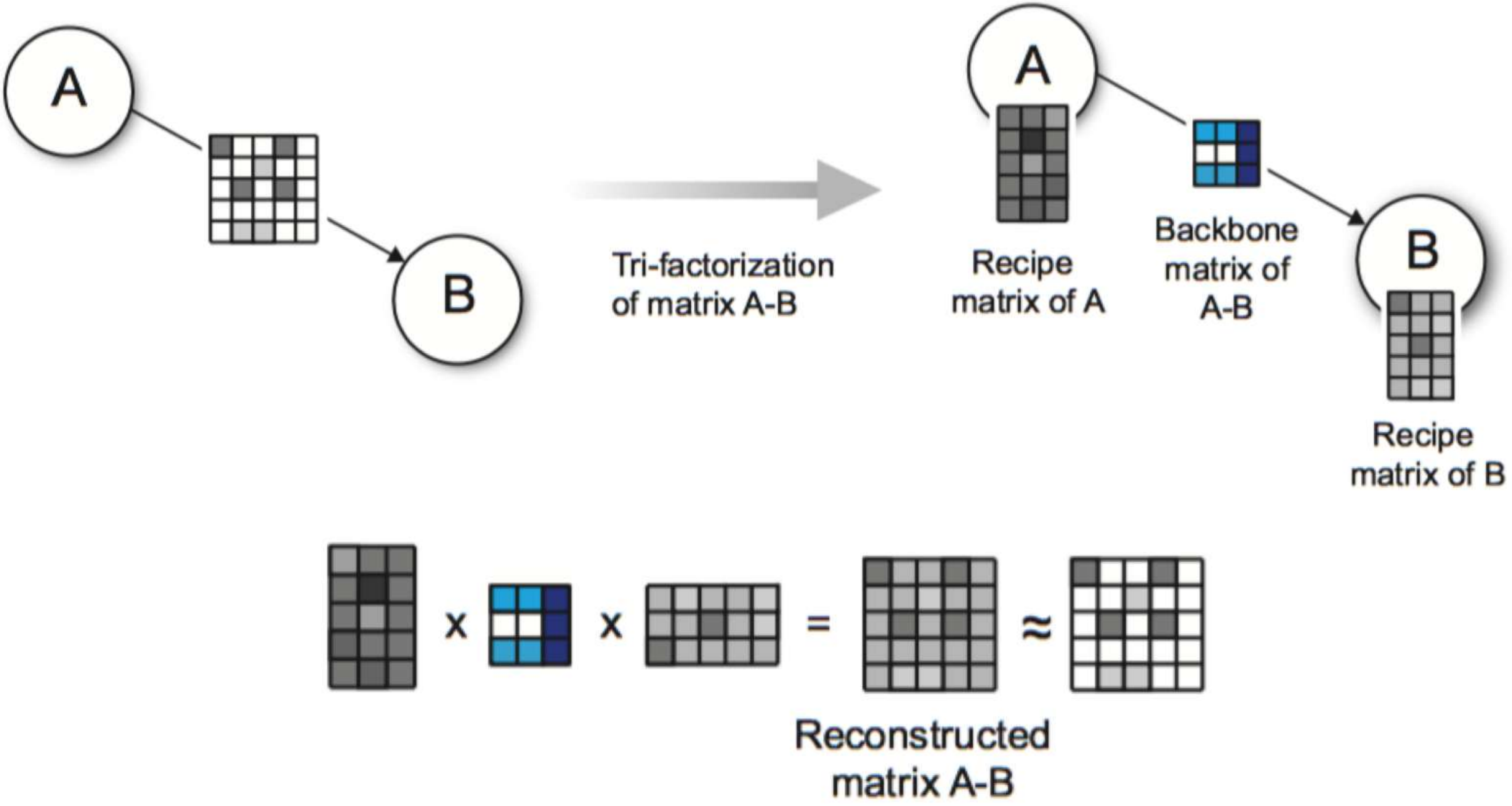
Validation methods are increasingly commoditized

Big data researchers will have a higher impact in biomedicine ☺

Recent Applications in Biomedicine

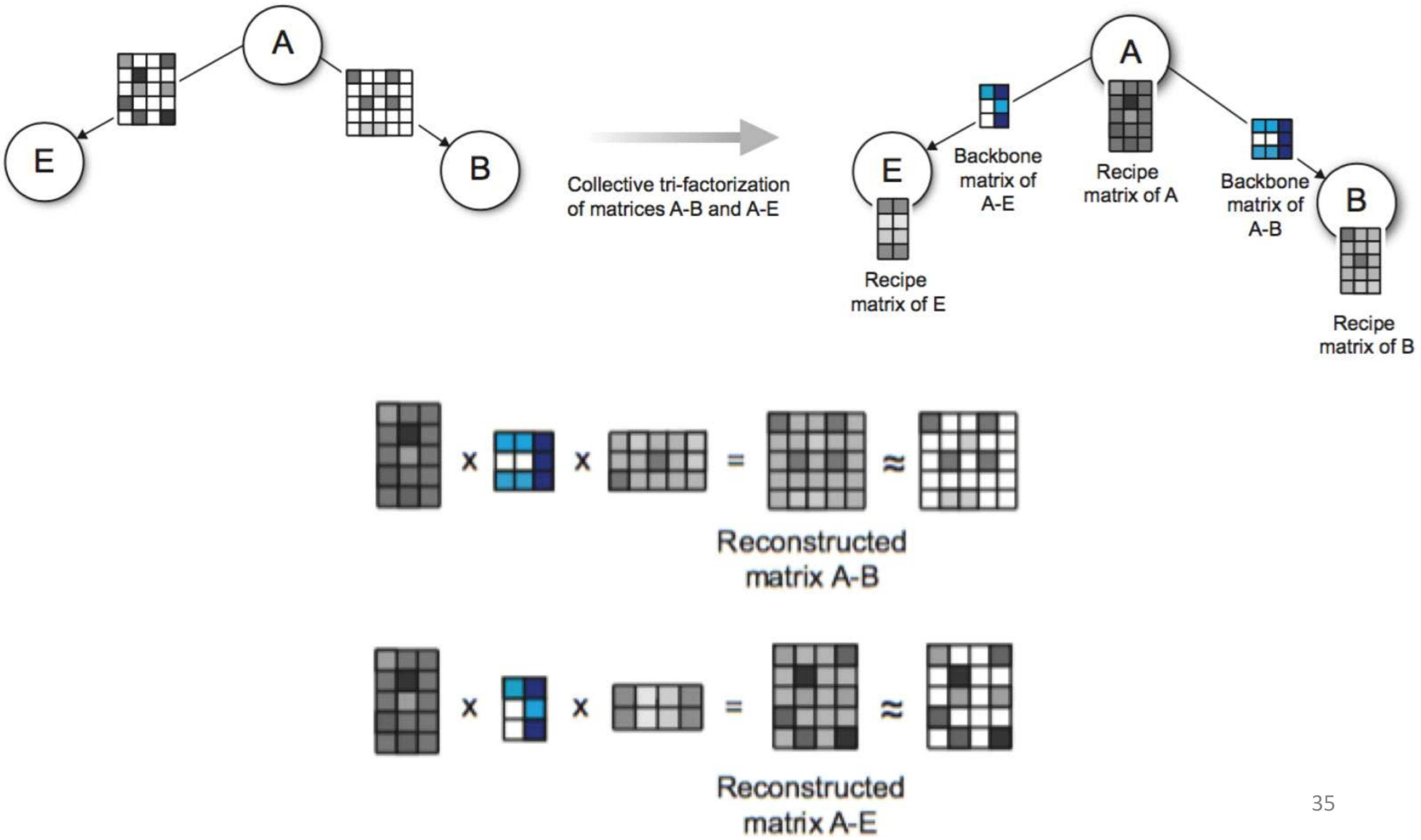
- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- ➡ Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping

Matrix Tri-Factorization



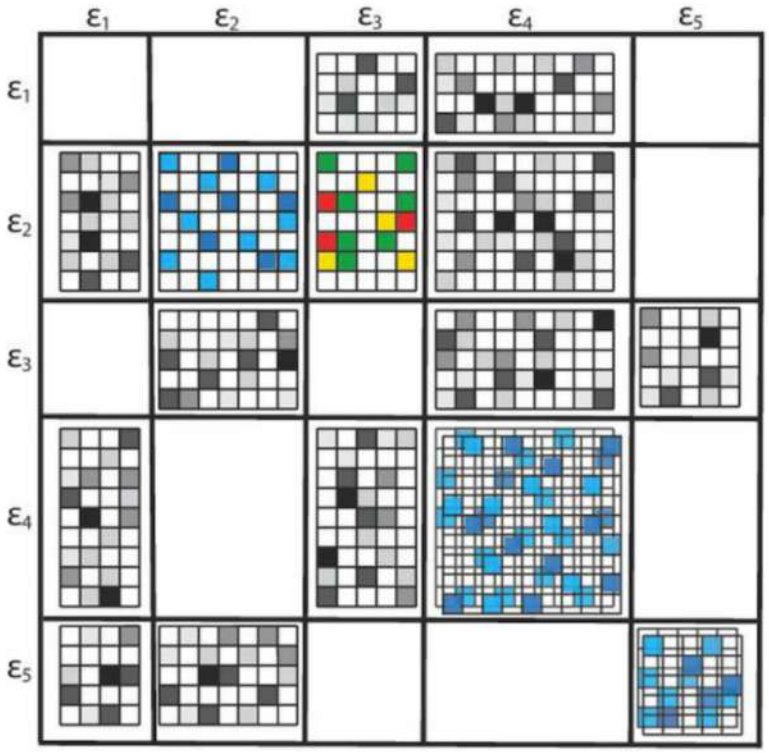
Ding C, Li T, Park H. Orthogonal Nonnegative Matrix Tri-factorizations for Clustering. KDD, 2006.
Wang F, Li T, Zhang C. Semi-supervised clustering via matrix factorization. SDM, 2008.

Simultaneous Matrix Tri-Factorization

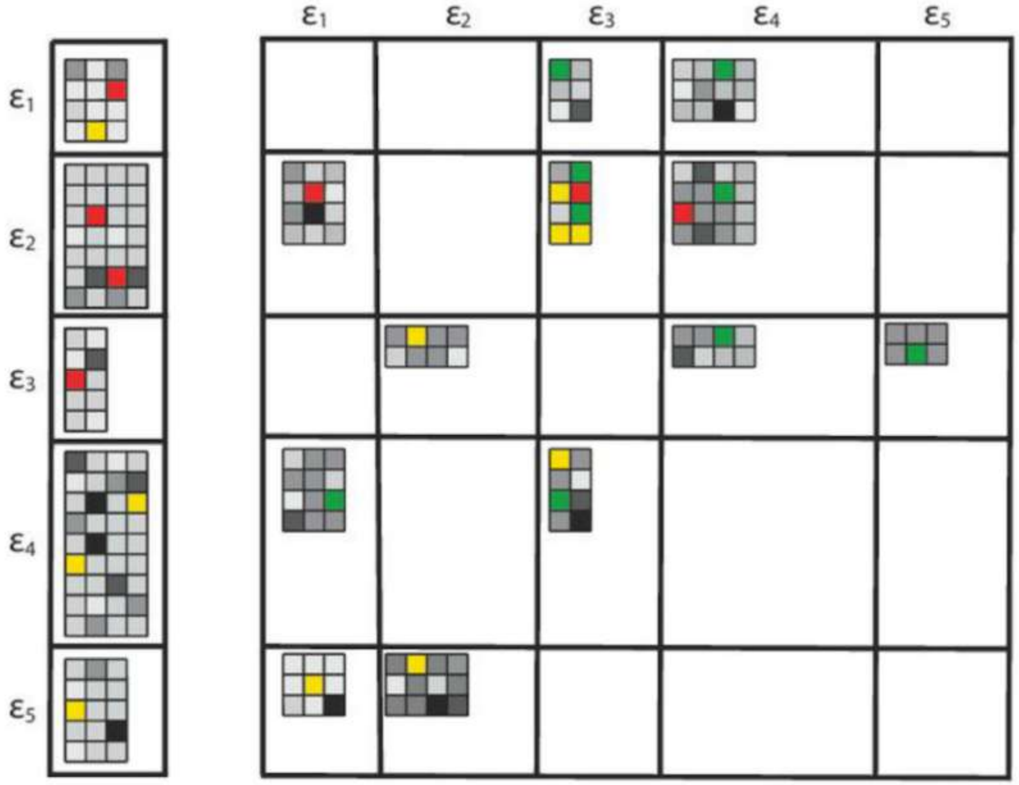


Data Fusion by Simultaneous Matrix Tri-Factorization

Input to data fusion



Simultaneous Constrained Decomposition



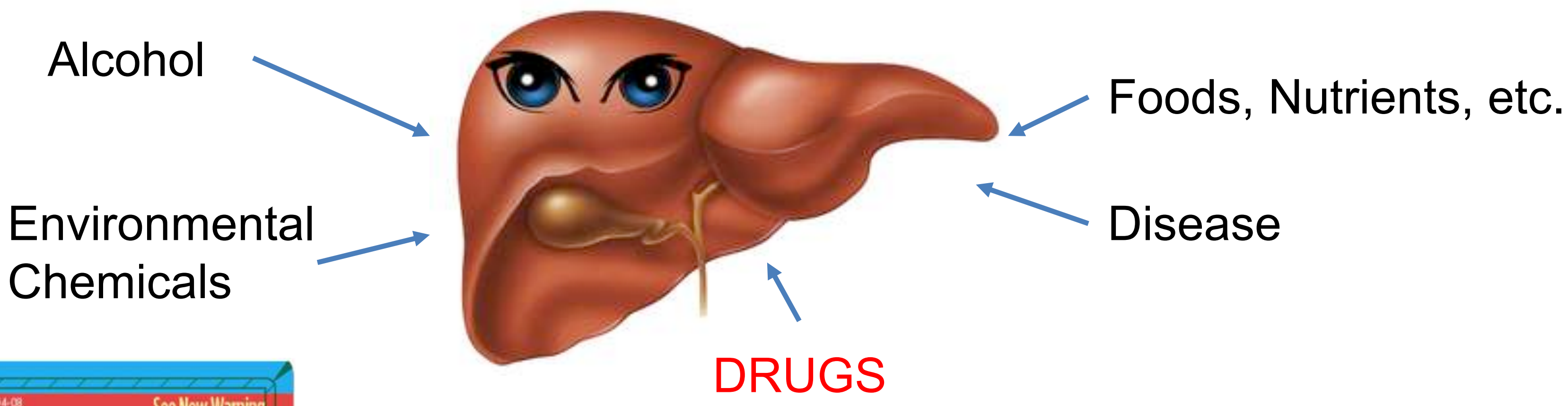
$$\min_{\mathbf{G}\geq 0} J(\mathbf{G};\mathbf{S}) = \sum_{\mathbf{R}_{ij}\in\mathcal{R}} ||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||^2 + \sum_{t=1}^{\max_i t_i} \mathrm{tr}(\mathbf{G}^T\mathbf{\Theta}^{(t)}\mathbf{G}),$$

Repeat until convergence:

- Fix G, update S
- Fix S, update G

Zitnik M, Zupan B. Data Fusion by Matrix Factorization. PAMI 2015.

Liver and Drug-Induced Liver Injury (DILI)



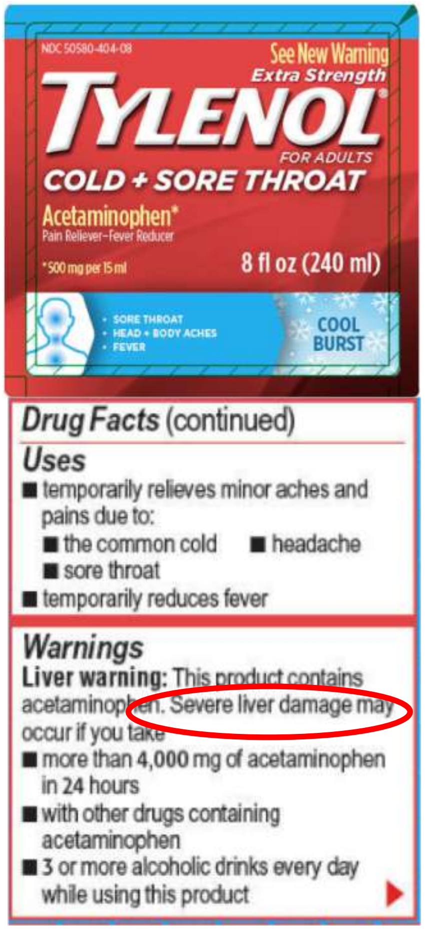
Alcohol

Environmental Chemicals

Foods, Nutrients, etc.

Disease

DRUGS



Drug Facts (continued)

Uses

■ temporarily relieves minor aches and pains due to:
■ the common cold ■ headache
■ sore throat
■ temporarily reduces fever

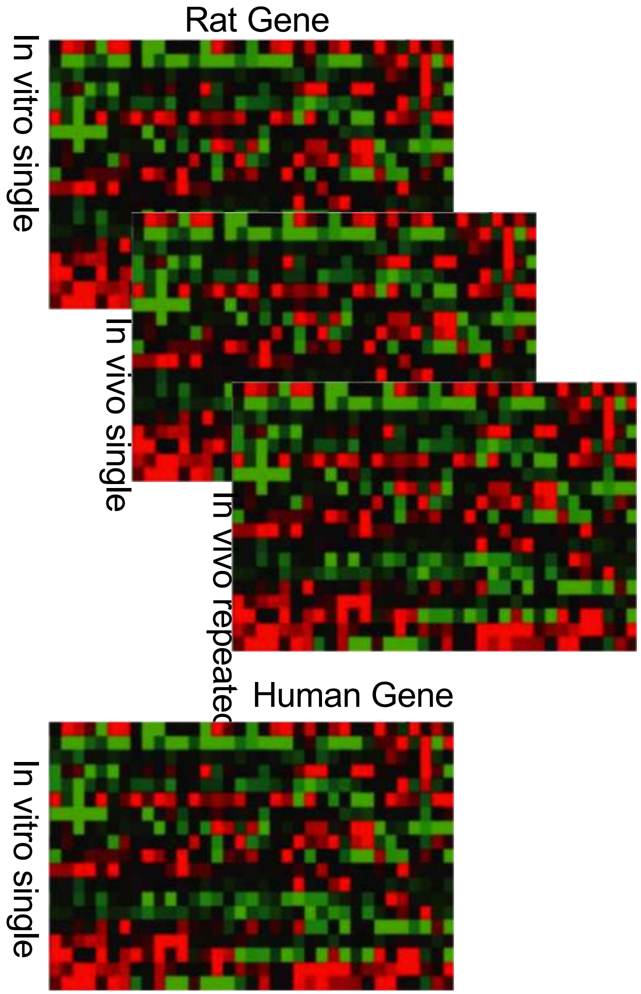
Warnings

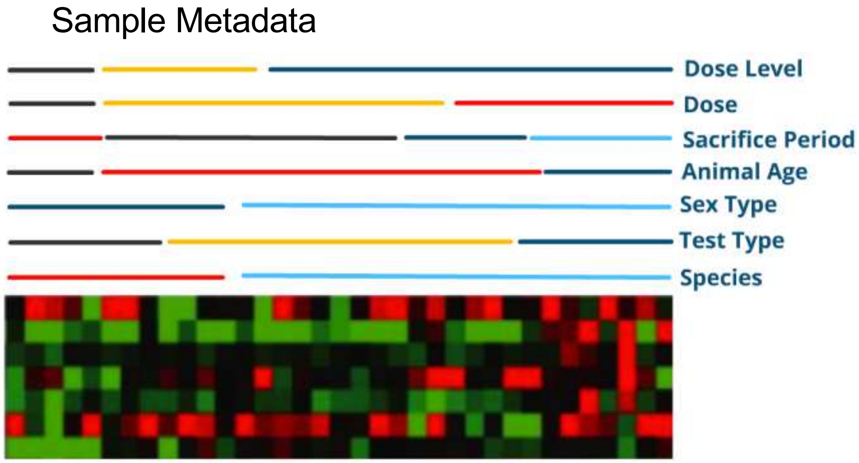
Liver warning: This product contains acetaminophen. Severe liver damage may occur if you take
■ more than 4,000 mg of acetaminophen in 24 hours
■ with other drugs containing acetaminophen
■ 3 or more alcoholic drinks every day while using this product

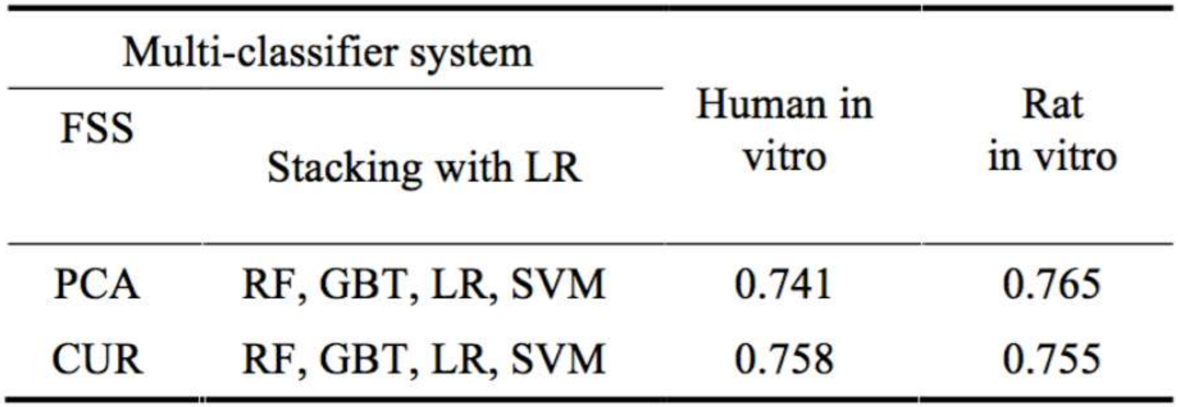
- "Approved drugs are the most common cause of acute liver failure in the USA" - FDA
- DILI is the MOST frequent reason for drug withdrawal during drug discovery, clinical trials, and after drugs are approved for the marketplace

CAMDA 2012 Task: DILI Prediction

- CAMDA: Critical Assessment of Massive Data Analysis
- The Japanese Toxicogenomics Project (TGP) creates a gene expression database using the Affymetrix GeneChip arrays to measure the effects of 131 chemicals, mainly medical drugs, on the liver.
- DILI potential has been categorized as severe, moderate, or mild.







Multi-classifier system		Human in vitro	Rat in vitro
FSS	Stacking with LR		
PCA	RF, GBT, LR, SVM	0.741	0.765
CUR	RF, GBT, LR, SVM	0.758	0.755

Data Fusion of Additional Sources

Histological and clinical chemistry data (Rat, in vivo)

Hematology

RBC, Neutrophil, Eosinophil, Basophil, Monocyte, Lymphocyte

Liver Weight

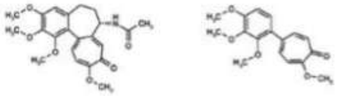
Terminal body weight
Liver weight,
Relative liver weight

Blood Chemistry

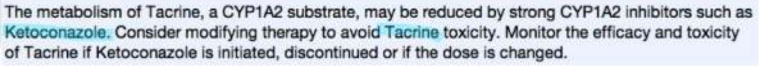
ALP, Cl, TC, Ca, TG, IP, PL, TP, TBIL, RALB, DBIL, A/G GLC, AST (GOT), BUN, ALT (GPT), CRE, LDH, Na, gamma-GTP, K

Drug information from DrugBank

Chemical Structure

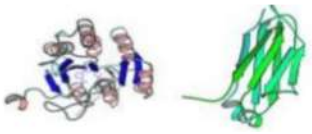


Drug Interactions

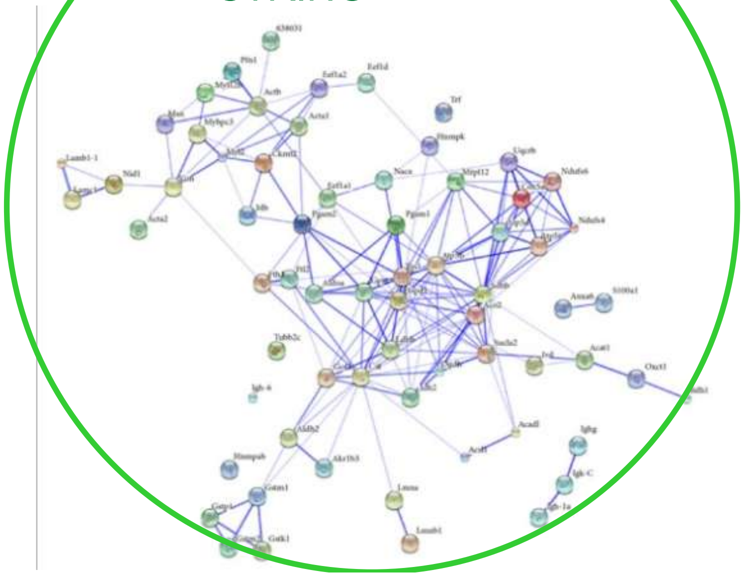


The metabolism of Tacrine, a CYP1A2 substrate, may be reduced by strong CYP1A2 inhibitors such as Ketoconazole. Consider modifying therapy to avoid Tacrine toxicity. Monitor the efficacy and toxicity of Tacrine if Ketoconazole is initiated, discontinued or if the dose is changed.

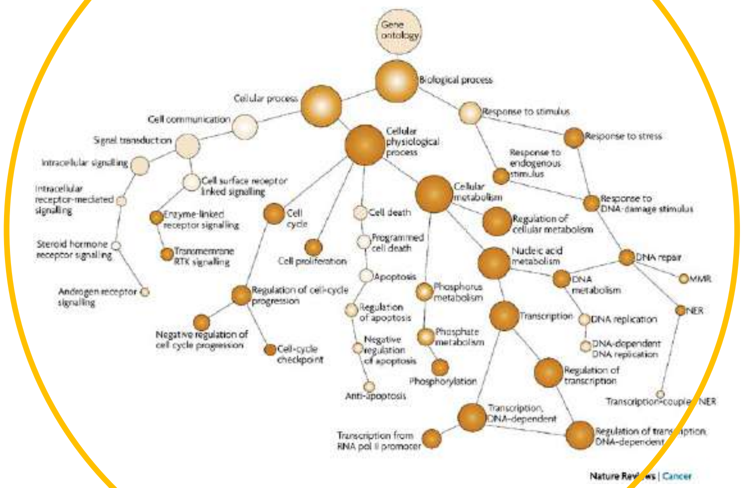
Drug Targets



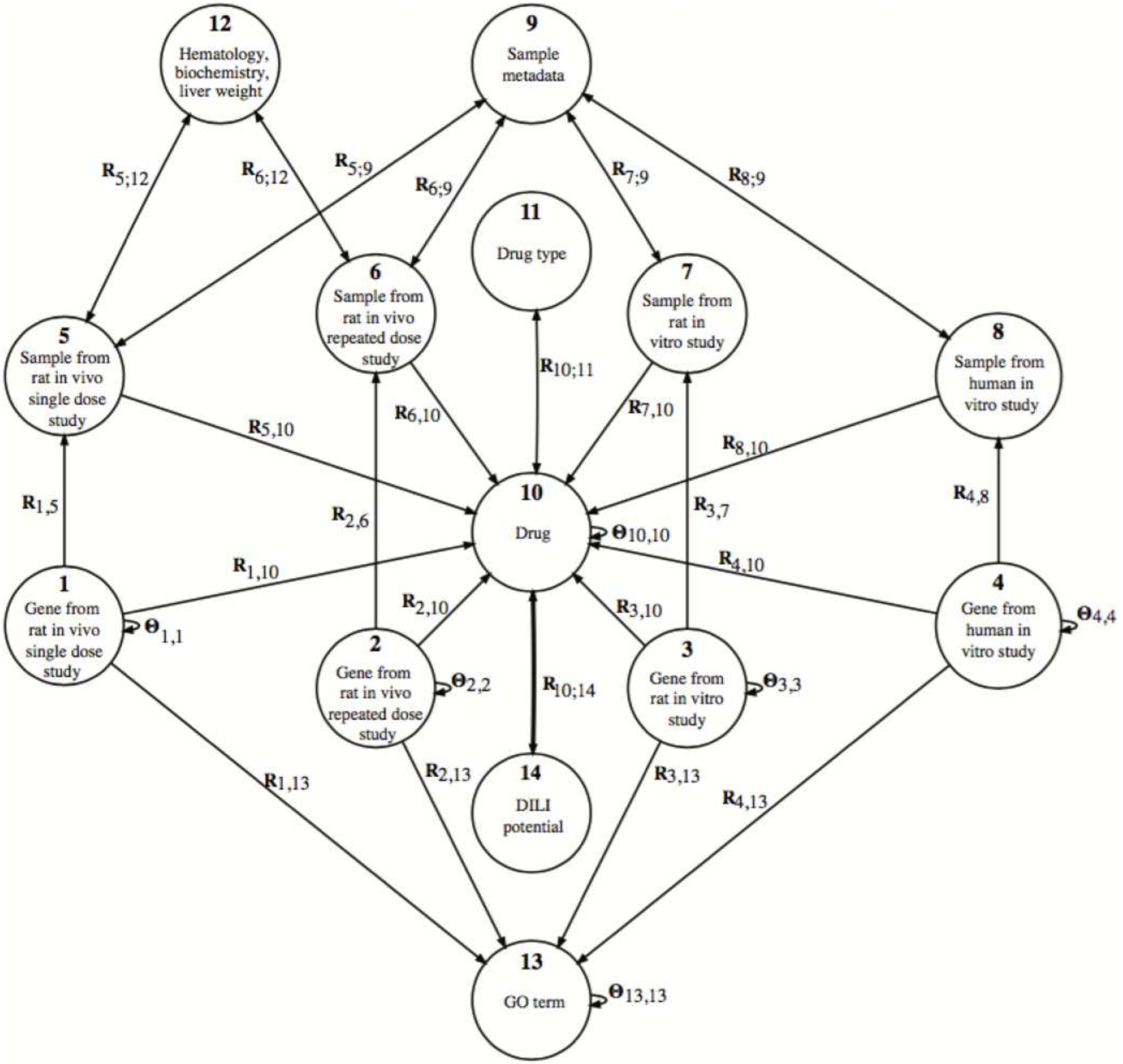
Protein-protein interactions (PPI) from STRING



Gene Ontology (GO)


Gene ontology
Biological process
Cellular process
Cell communication
Signal transduction
Intracellular signalling
Intracellular receptor-mediated signalling
Steroid hormone receptor signalling
Androgen receptor signalling
Cell surface receptor linked signalling
Enzyme-linked receptor signalling
Transmembrane RTK signalling
Cellular physiological process
Cell cycle
Cell proliferation
Regulation of cell-cycle progression
Negative regulation of cell cycle progression
Cell-cycle checkpoint
Cell death
Programmed cell death
Apoptosis
Regulation of apoptosis
Negative regulation of apoptosis
Anti-apoptosis
Cellular metabolism
Phosphorus metabolism
Phosphate metabolism
Phosphorylation
Regulation of cellular metabolism
Nucleic acid metabolism
Transcription
Transcription, DNA-dependent
Transcription from RNA pol II promoter
Regulation of transcription
Regulation of transcription, DNA-dependent
DNA metabolism
DNA replication
DNA-dependent DNA replication
Response to stimulus
Response to stress
Response to endogenous stimulus
Response to DNA-damage stimulus
DNA repair
MMR
NER
Transcription-coupled NER
Nature Reviews | Cancer

Matrix Factorization-Based DILI Prediction



Data fusion studies	AUC
In vivo studies	0.819
In vitro studies	0.790
Human in vitro study	0.793
Animal in vitro study	0.799
Animal studies	0.811
Human studies	0.792
All studies	0.810

Given the aim to predict DILI potential in humans:

- Animal studies may be replaced with in vitro assays (AUC = 0.799)
- Liver injury in humans can be predicted from animal data (AUC = 0.811)
- animal in vivo > animal in vitro ≈ human in vitro

Zitnik M, Zupan B. Matrix factorization-based data fusion for drug-induced liver injury prediction. Systems Biomedicine 2014. (First prize winner at CAMDA 2013 Conference)

Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping

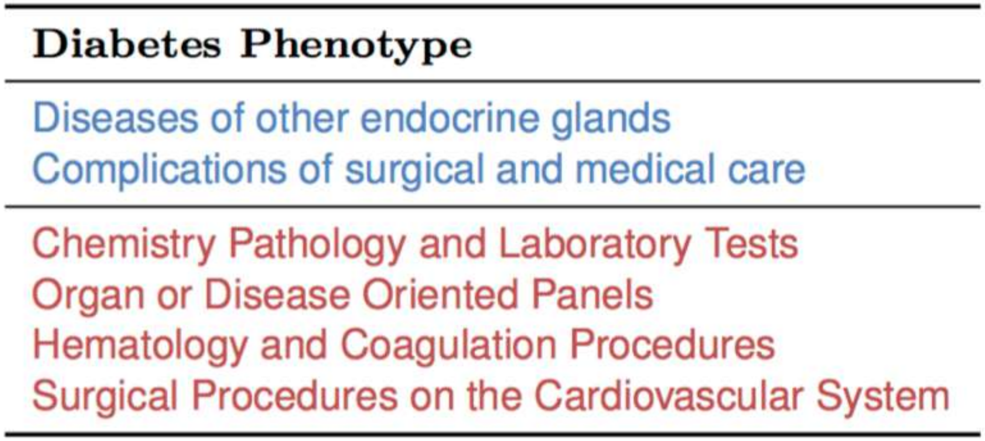
Phenotyping from Electronic Medical Records (EMR)

Phenotype (American Heritage Dictionary)

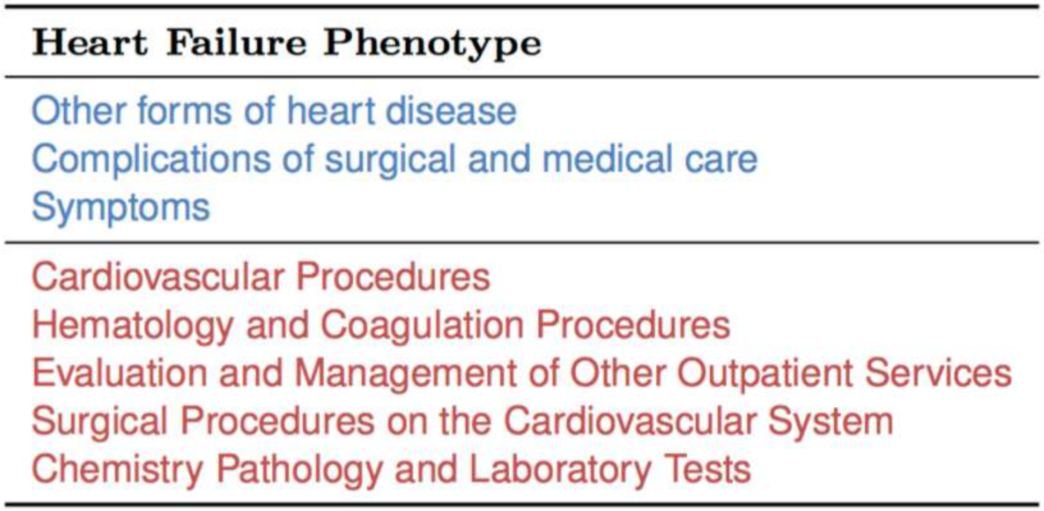
- The *observable* physical or biochemical *characteristics* of an organism, as determined by both genetic makeup and environmental influences.

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



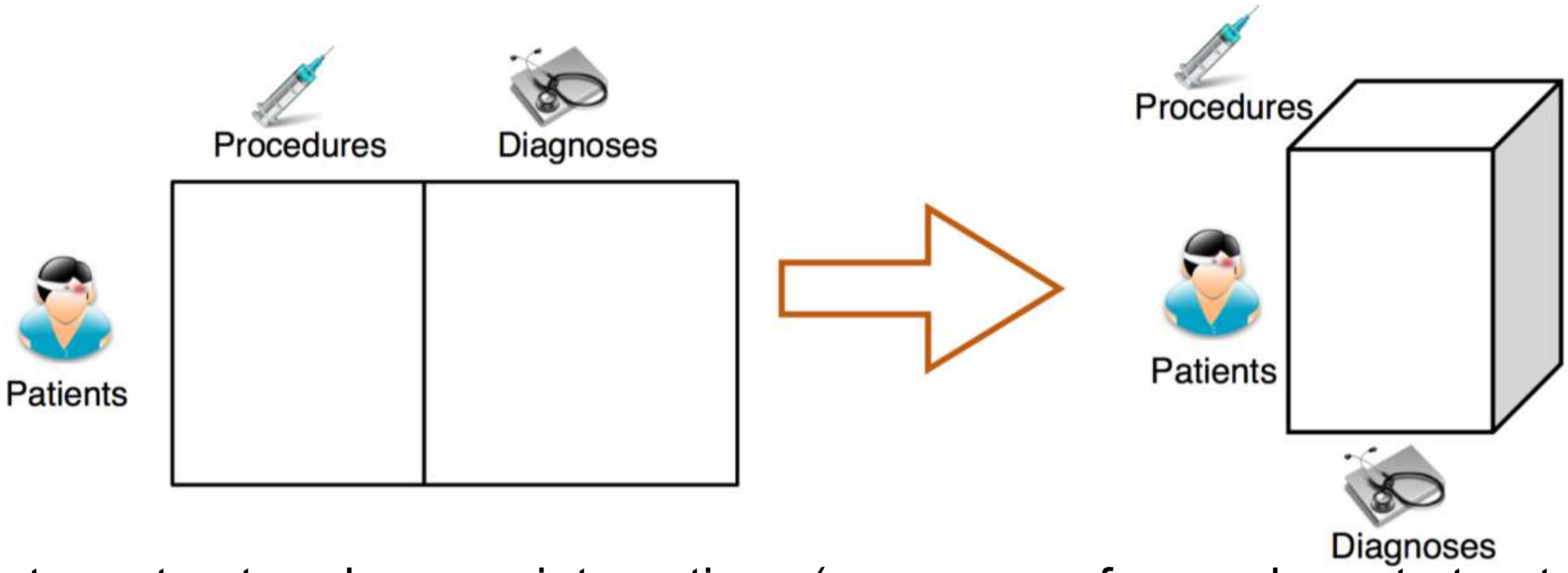
Diabetes Phenotype
Diseases of other endocrine glands
Complications of surgical and medical care
Chemistry Pathology and Laboratory Tests
Organ or Disease Oriented Panels
Hematology and Coagulation Procedures
Surgical Procedures on the Cardiovascular System



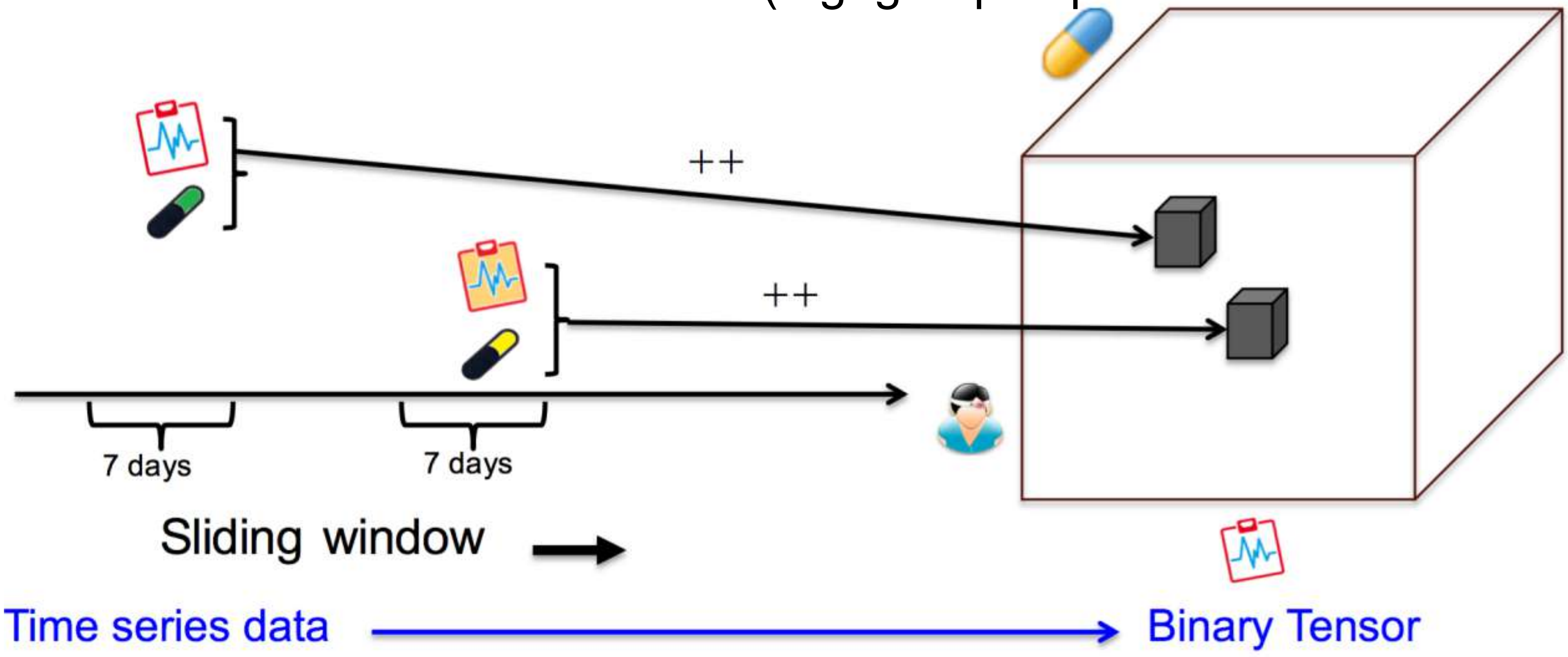
Heart Failure Phenotype
Other forms of heart disease
Complications of surgical and medical care
Symptoms
Cardiovascular Procedures
Hematology and Coagulation Procedures
Evaluation and Management of Other Outpatient Services
Surgical Procedures on the Cardiovascular System
Chemistry Pathology and Laboratory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

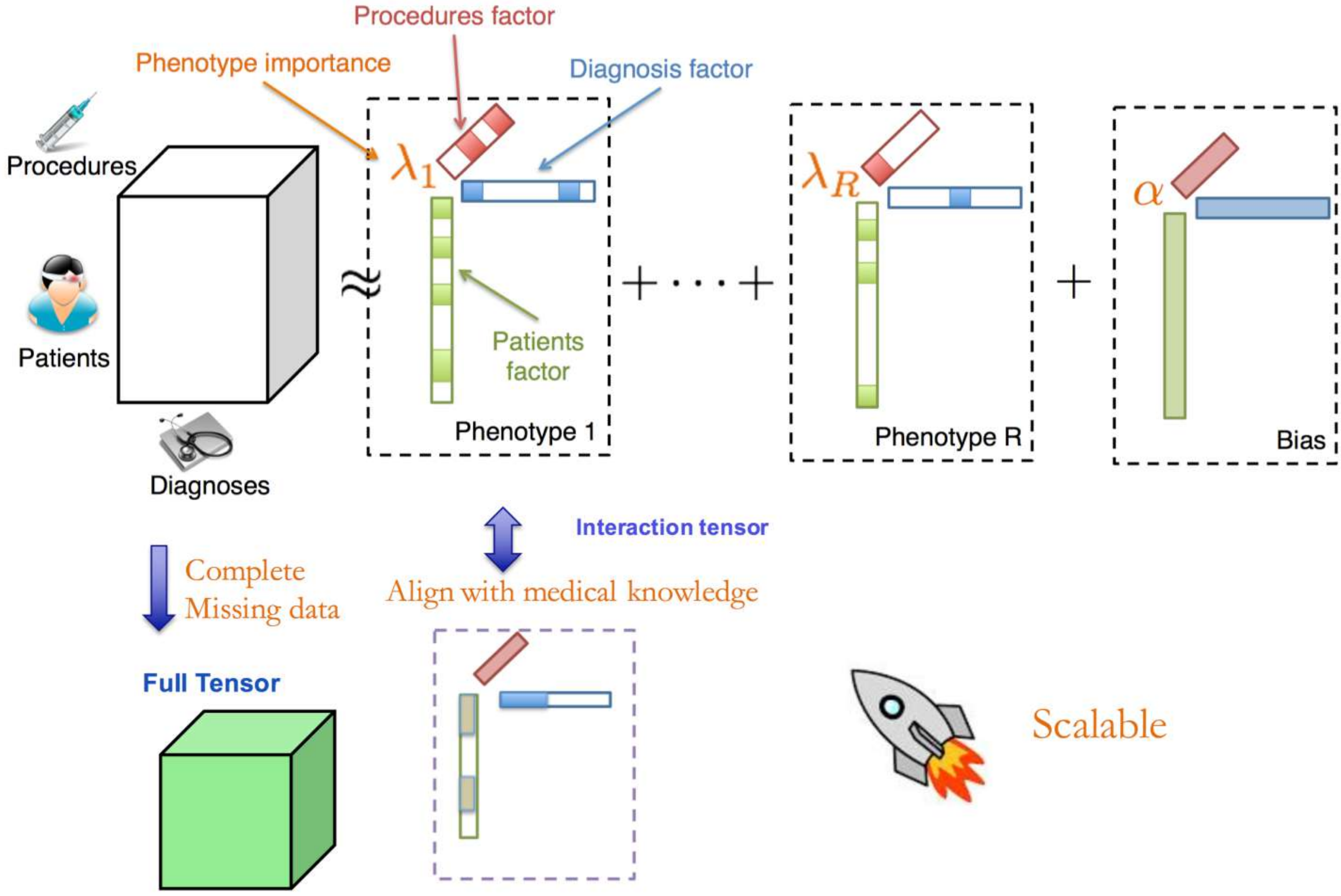


Capture structured source interactions (e.g. group of procedures to treat a disease)



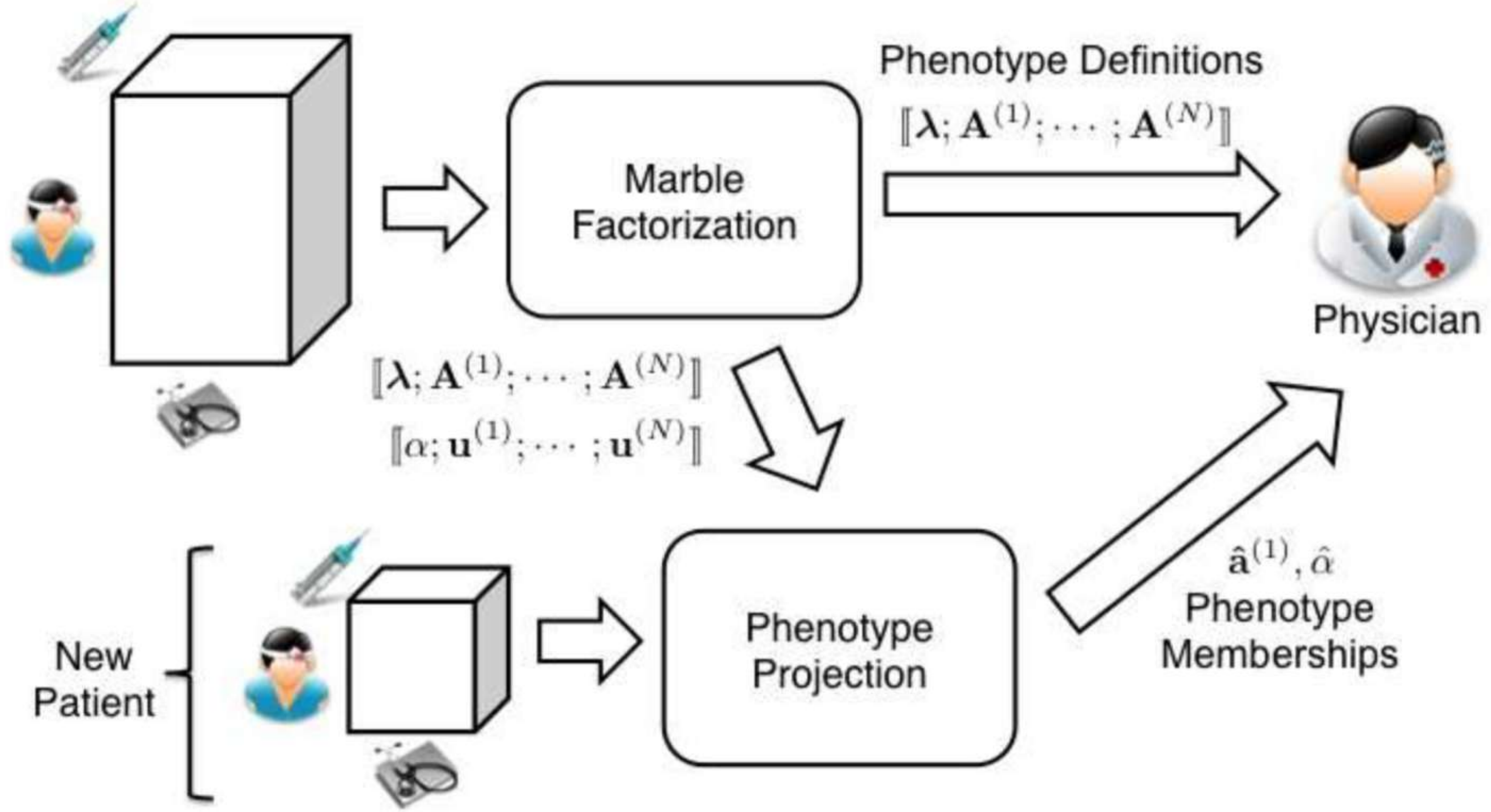
Co-occurrences of events are captured in the tensor as binary values

CP factorization for EMR



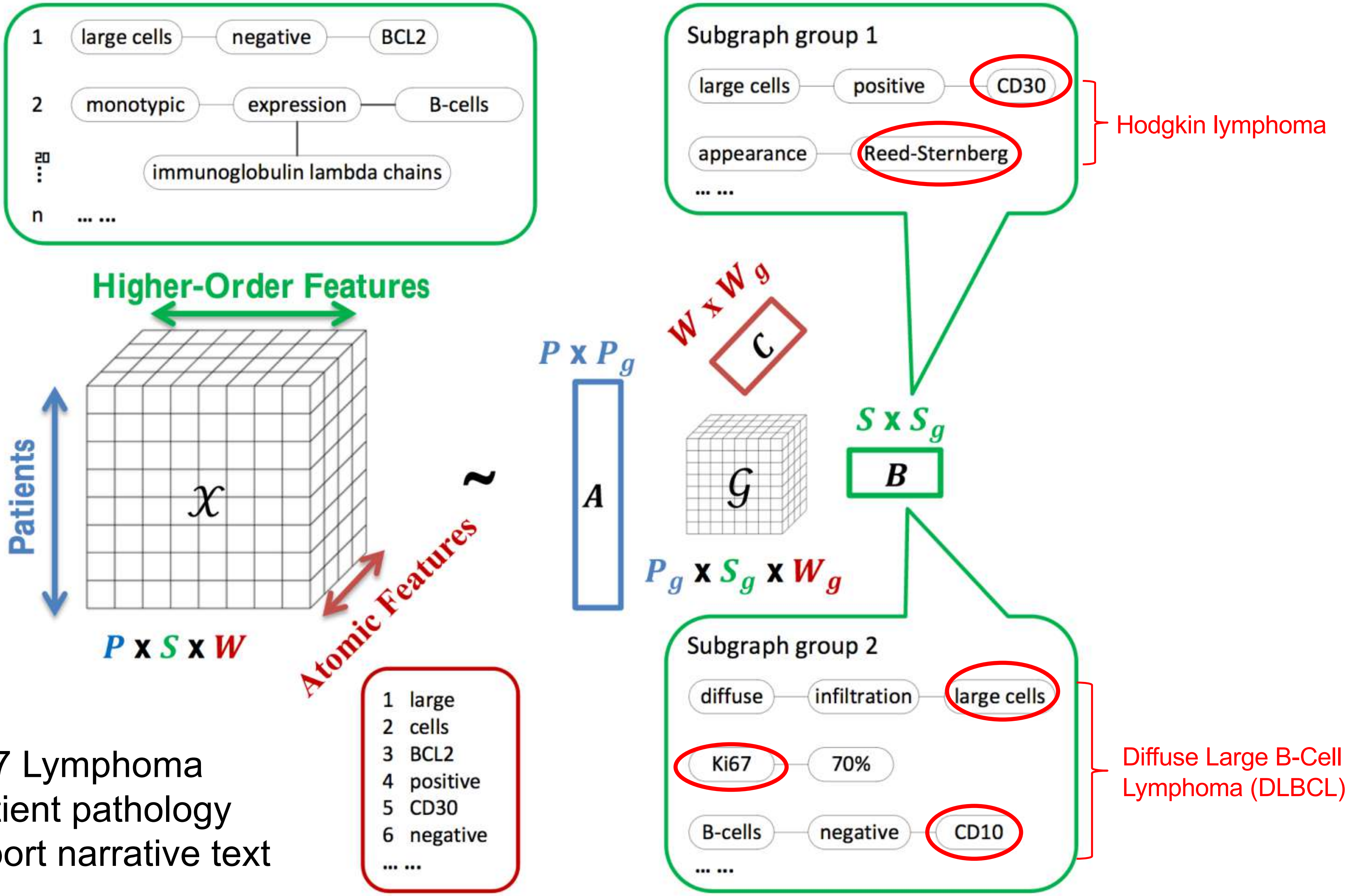
Ho J et al. Marble: High-throughput phenotyping from Electronic Health Records via sparse nonnegative tensor factorization. KDD 2014.
Wang Y et al. Rubik: Knowledge guided tensor factorization and completion for health data analytics. KDD 2015.

A possible application of EHR-phenotyping


Phenotype Definitions
$[\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!]$
Marble Factorization
Physician
$[\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!]$
$[\![\alpha; \mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!]$
$\hat{\mathbf{a}}^{(1)}, \hat{\alpha}$
Phenotype Memberships
New Patient
Phenotype Projection

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

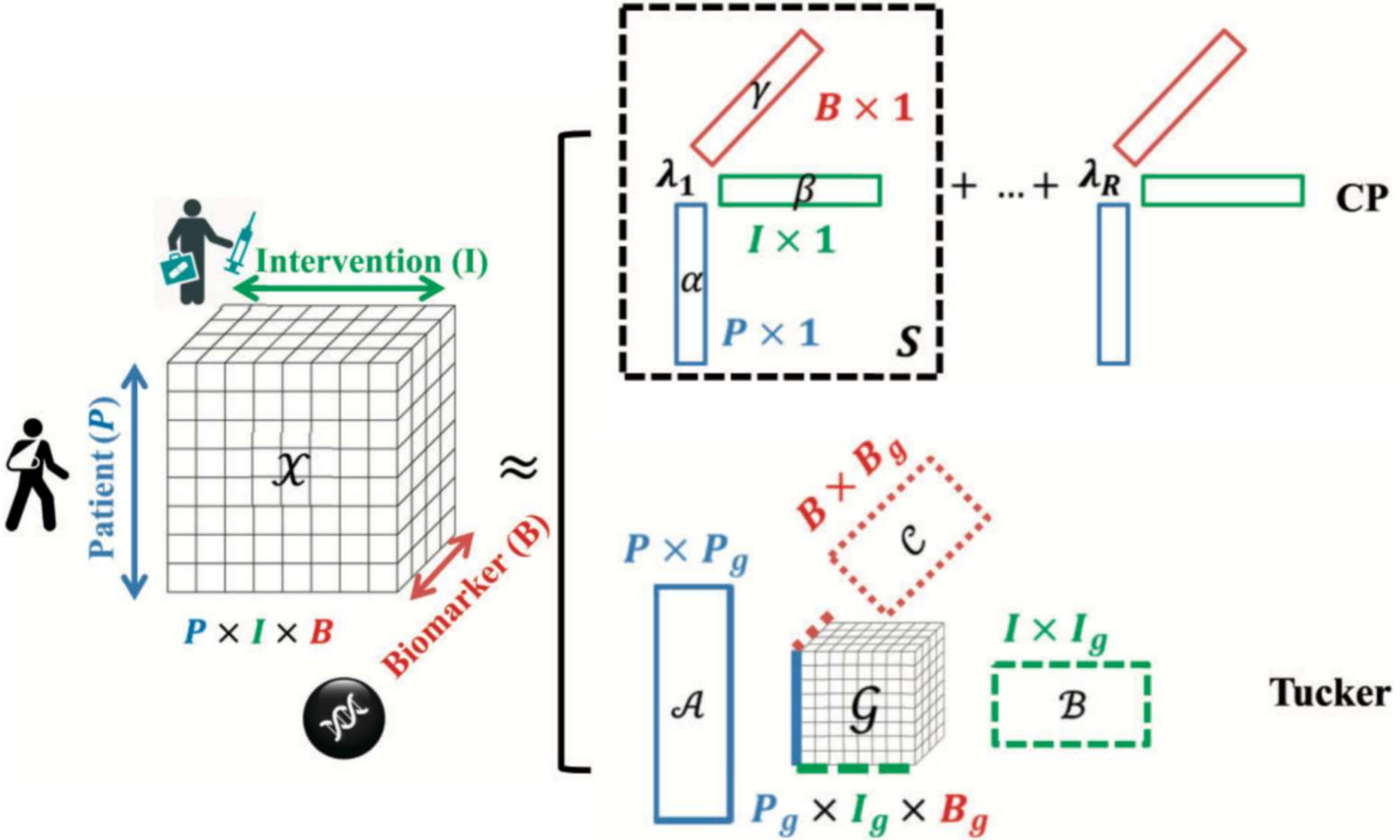
Tucker factorization for pathology reports



897 Lymphoma patient pathology report narrative text

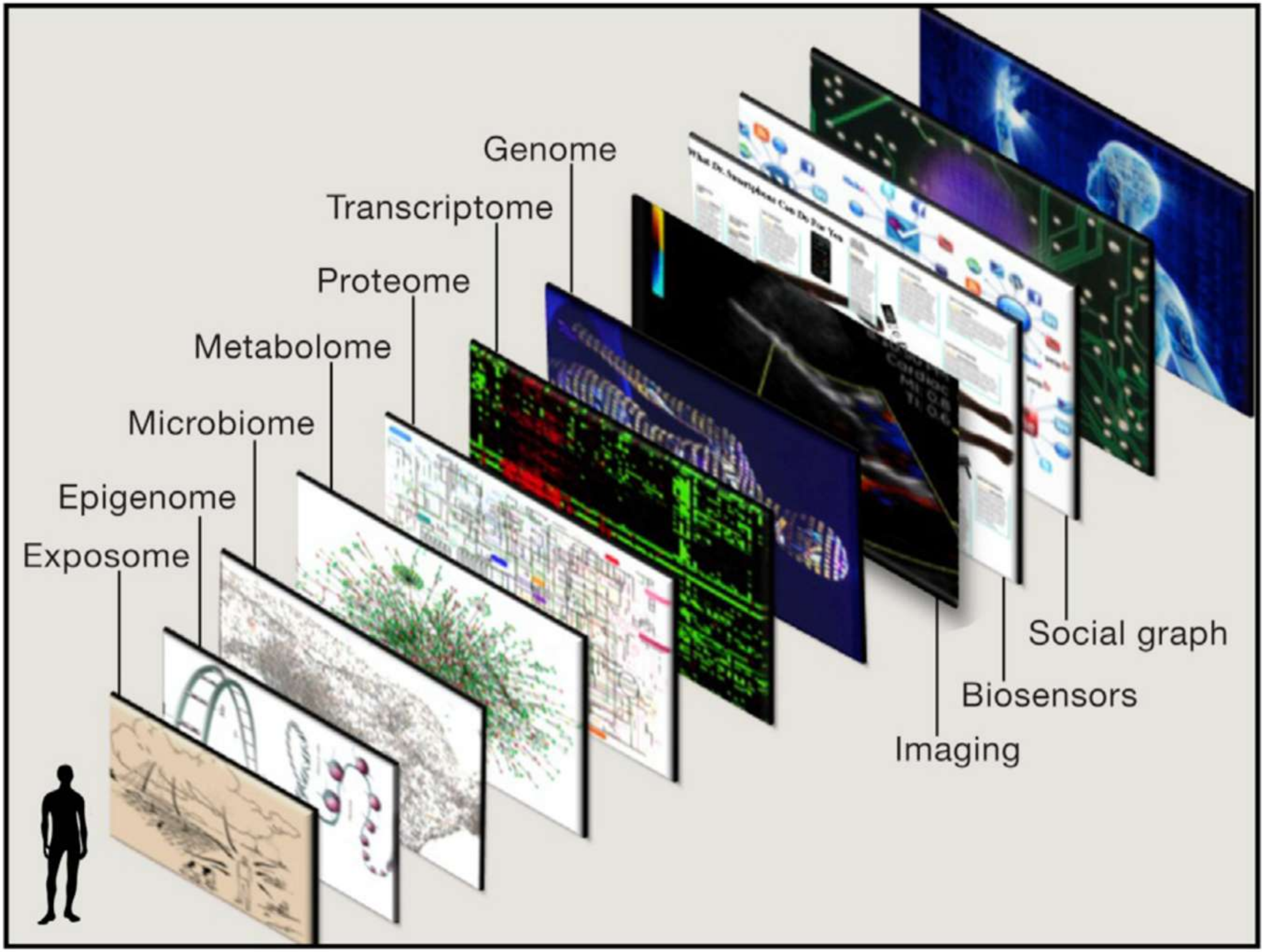
Luo Y et al. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. *JAMIA* 22:1009-1019, 2015.

Comparison of tensor modeling and factorization schemes



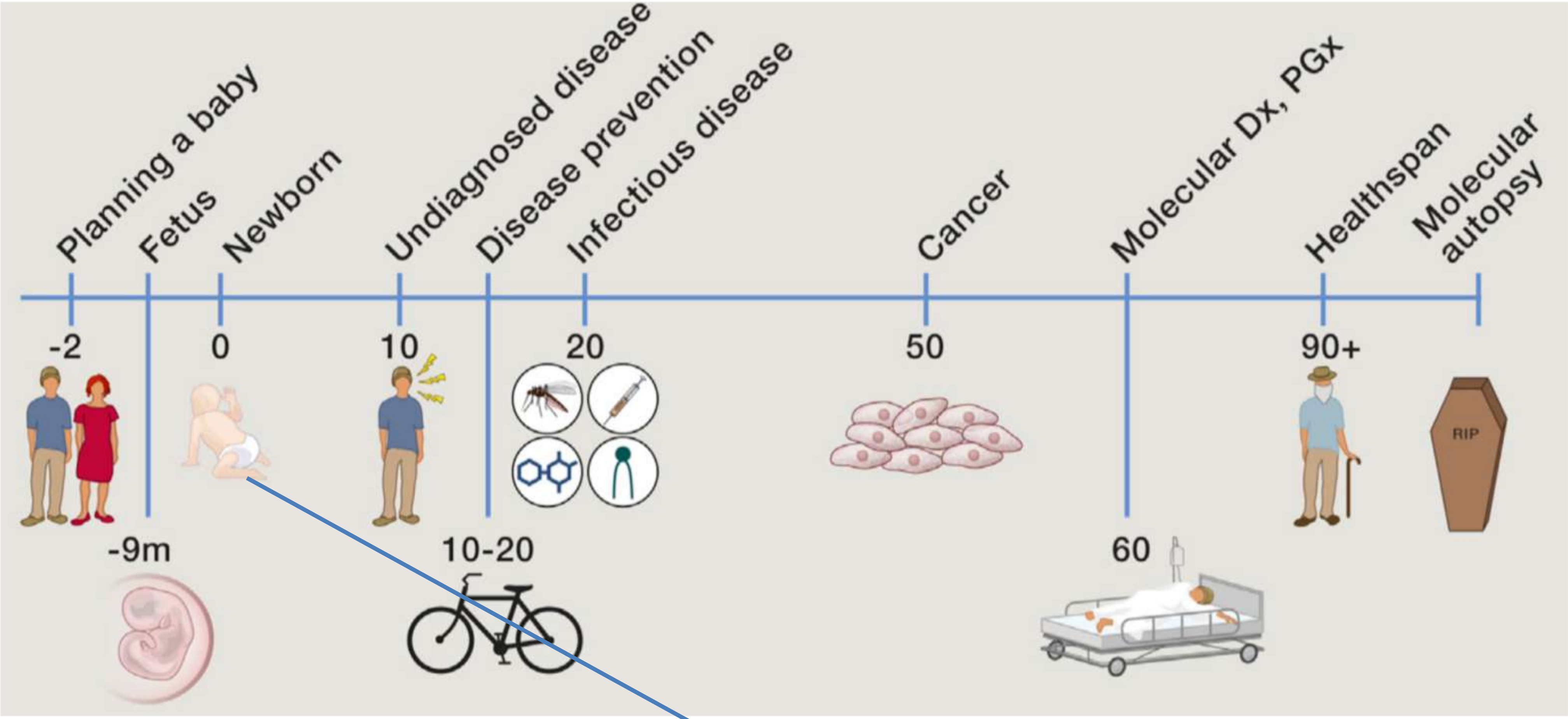
Luo Y, Wang F, Szolovits P. Tensor factorization toward precision medicine. *Brief Bioinform*, 2016

Challenges and opportunities: multiscale networks



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

Dynamic network: timeline of individualized genomic medicine

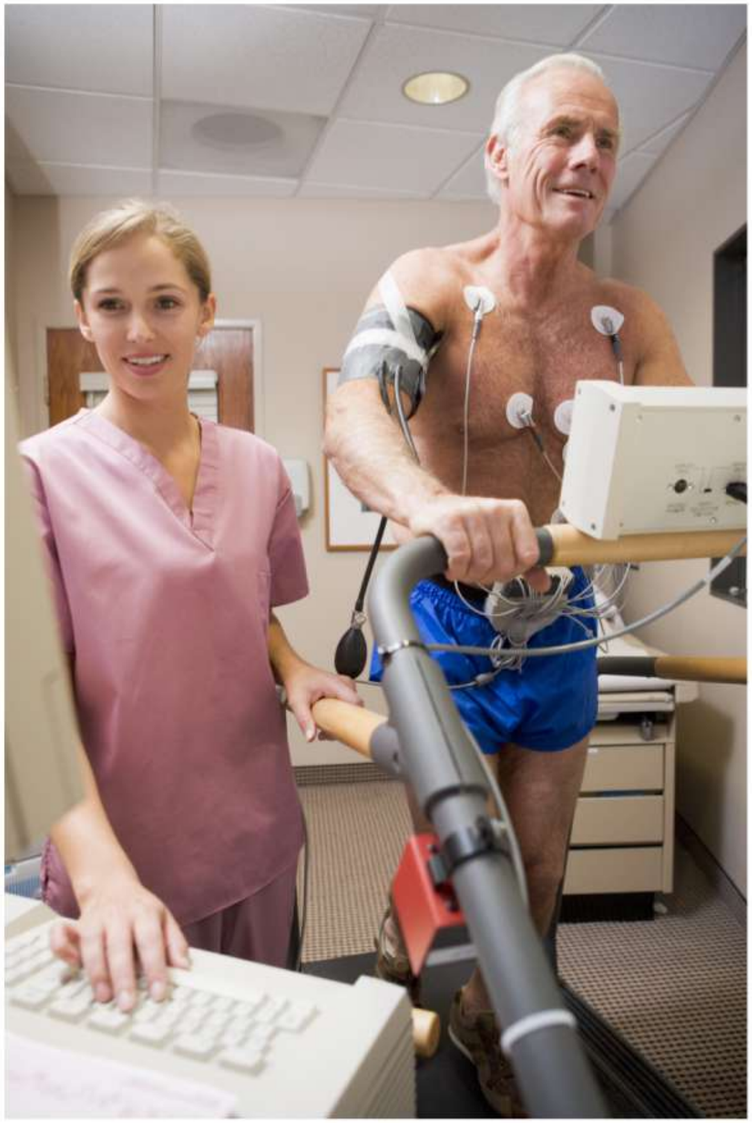


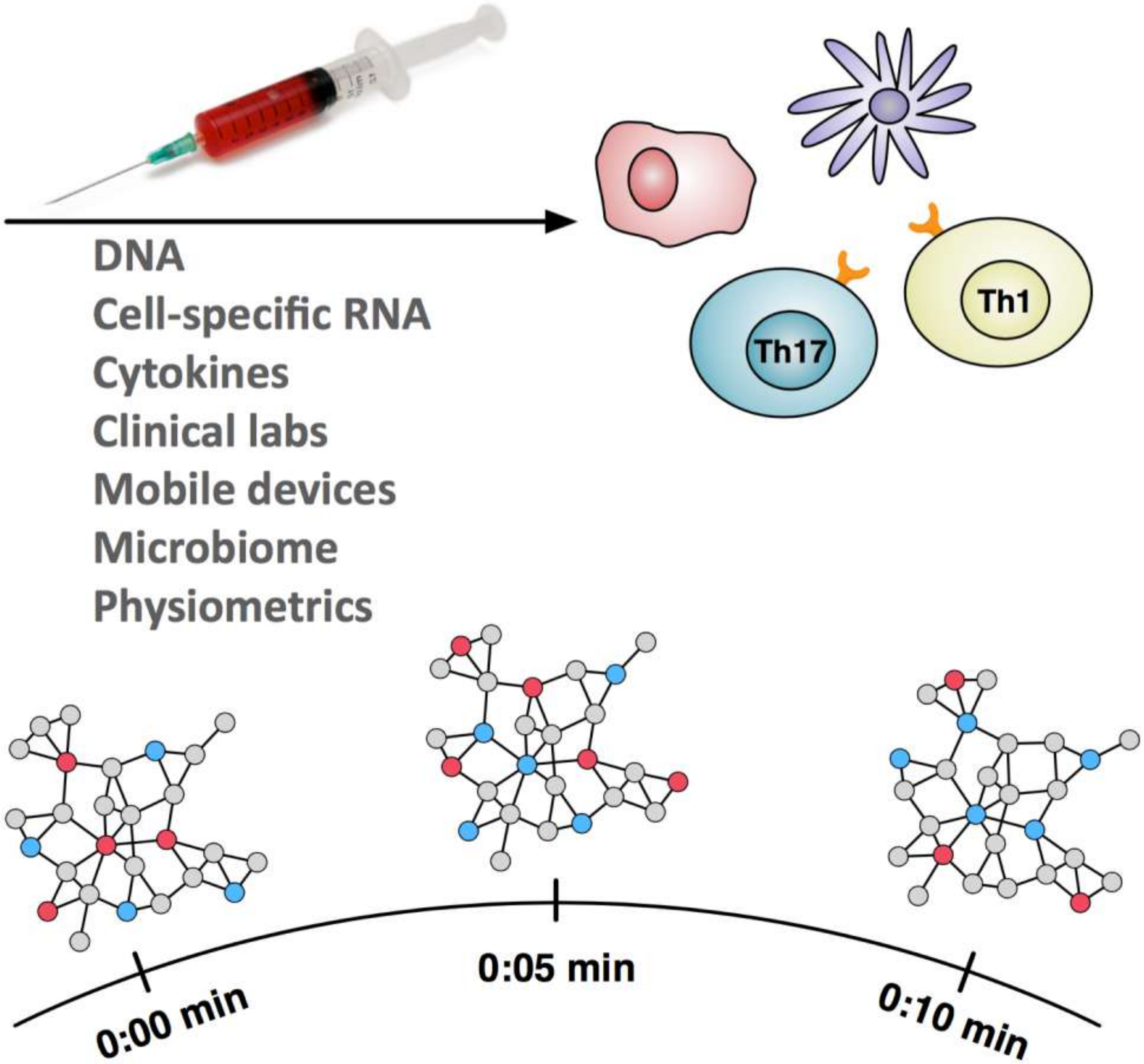
During an individual’s lifespan: ***from prewomb to tomb***

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

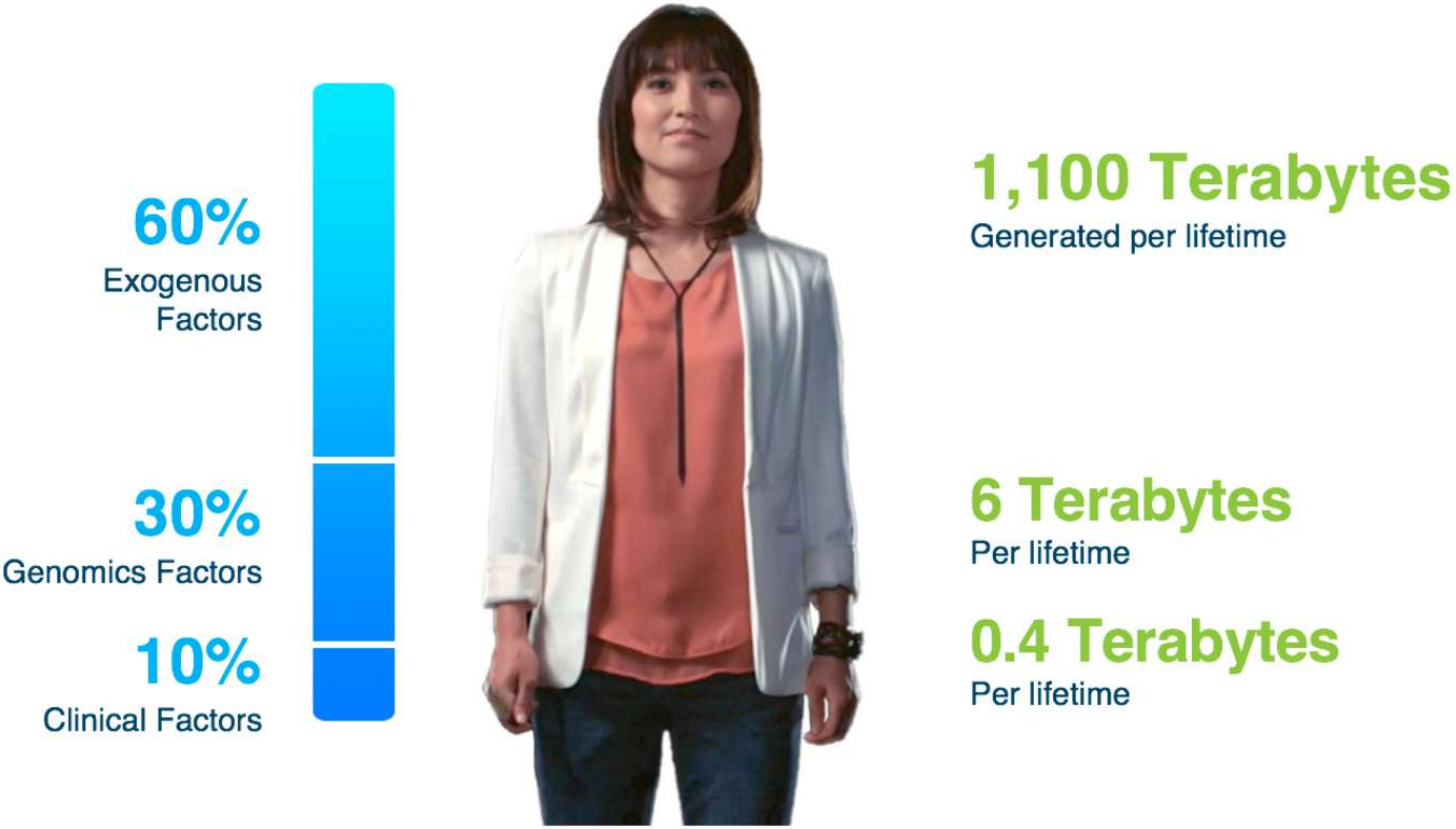




DNA
Cell-specific RNA
Cytokines
Clinical labs
Mobile devices
Microbiome
Physiometrics

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

Healthcare is really a big data industry

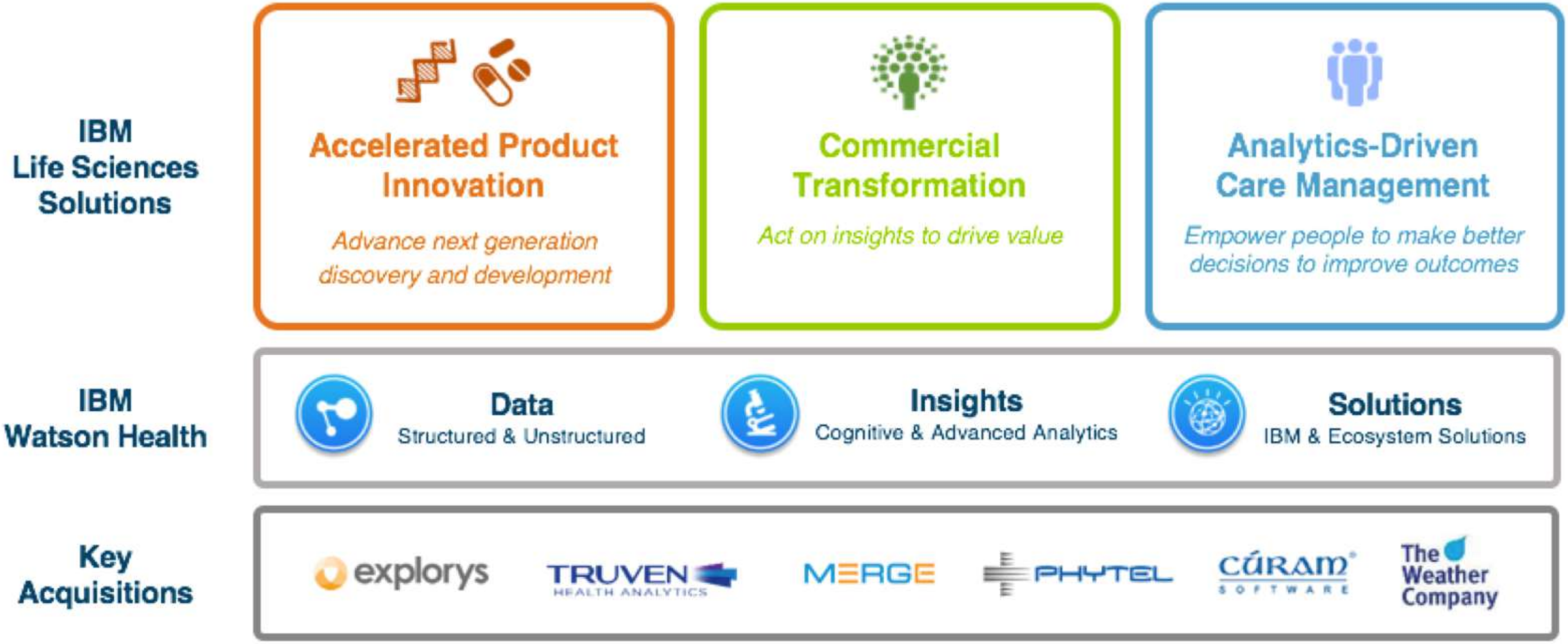


Help people live longer and feel better

Our commitment to Health – IBM Moonshot

> “I'm telling you, our moonshot will be the impact we will have on Healthcare. It has already started. We will change and do our part to change the face of Healthcare. I am absolutely positive about it. And that, to me, while we do many other things, that will be one of the most important.”

Ginni Rometty
IBM Chairman, President and CEO
April 16, 2015



IBM Life Sciences Solutions

Accelerated Product Innovation
Advance next generation discovery and development

Commercial Transformation
Act on insights to drive value

Analytics-Driven Care Management
Empower people to make better decisions to improve outcomes

IBM Watson Health

Data
Structured & Unstructured

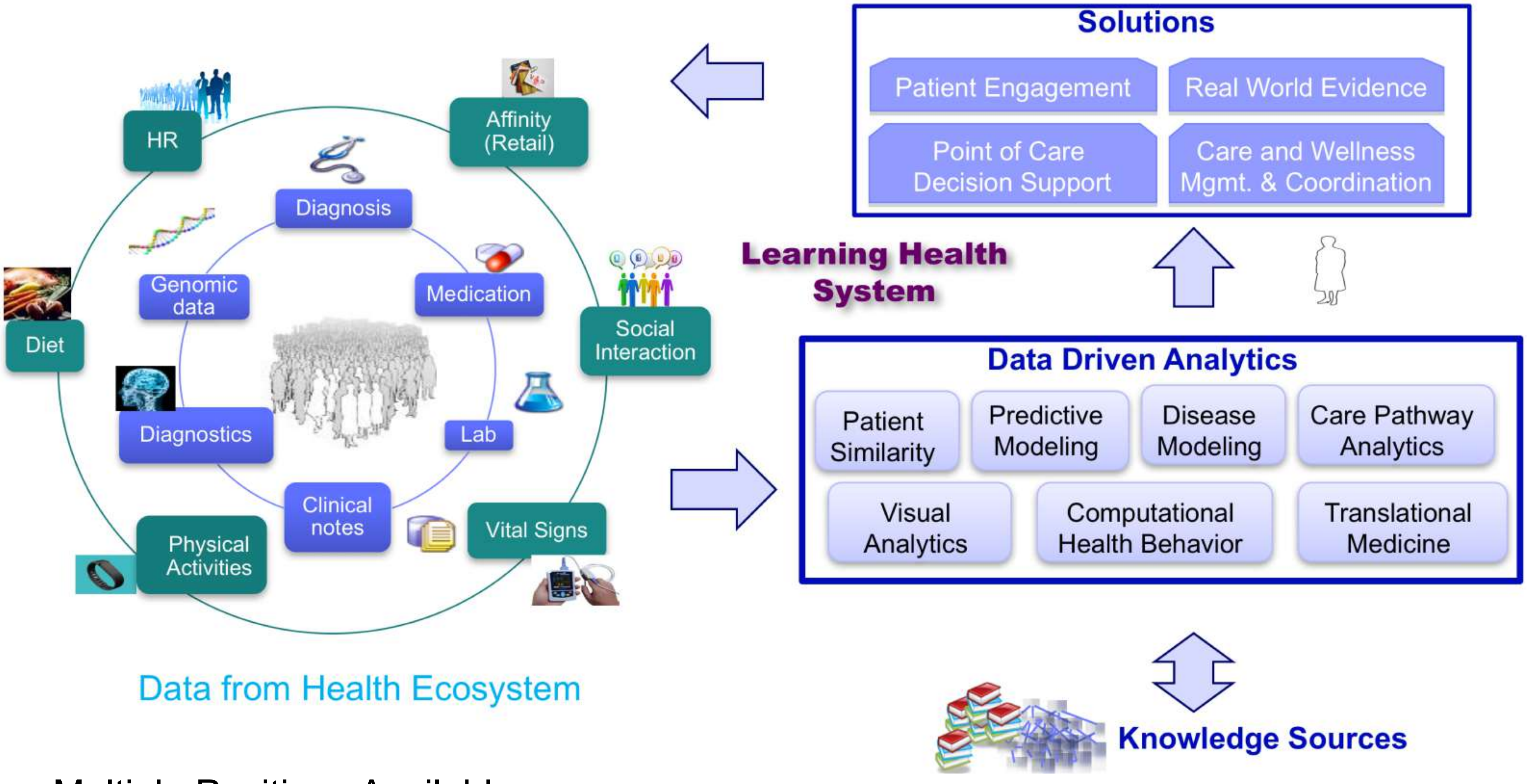
Insights
Cognitive & Advanced Analytics

Solutions
IBM & Ecosystem Solutions

Key Acquisitions

explorys · TRUVEN HEALTH ANALYTICS · MERGE · PHYTEL · CÚRAM SOFTWARE · The Weather Company

Center for Computational Health @ IBM



HR

Affinity (Retail)

Diagnosis

Genomic data

Medication

Diet

Social Interaction

Diagnostics

Lab

Clinical notes

Vital Signs

Physical Activities

Data from Health Ecosystem

Solutions

Patient Engagement

Real World Evidence

Point of Care Decision Support

Care and Wellness Mgmt. & Coordination

Learning Health System

Data Driven Analytics

Patient Similarity

Predictive Modeling

Disease Modeling

Care Pathway Analytics

Visual Analytics

Computational Health Behavior

Translational Medicine

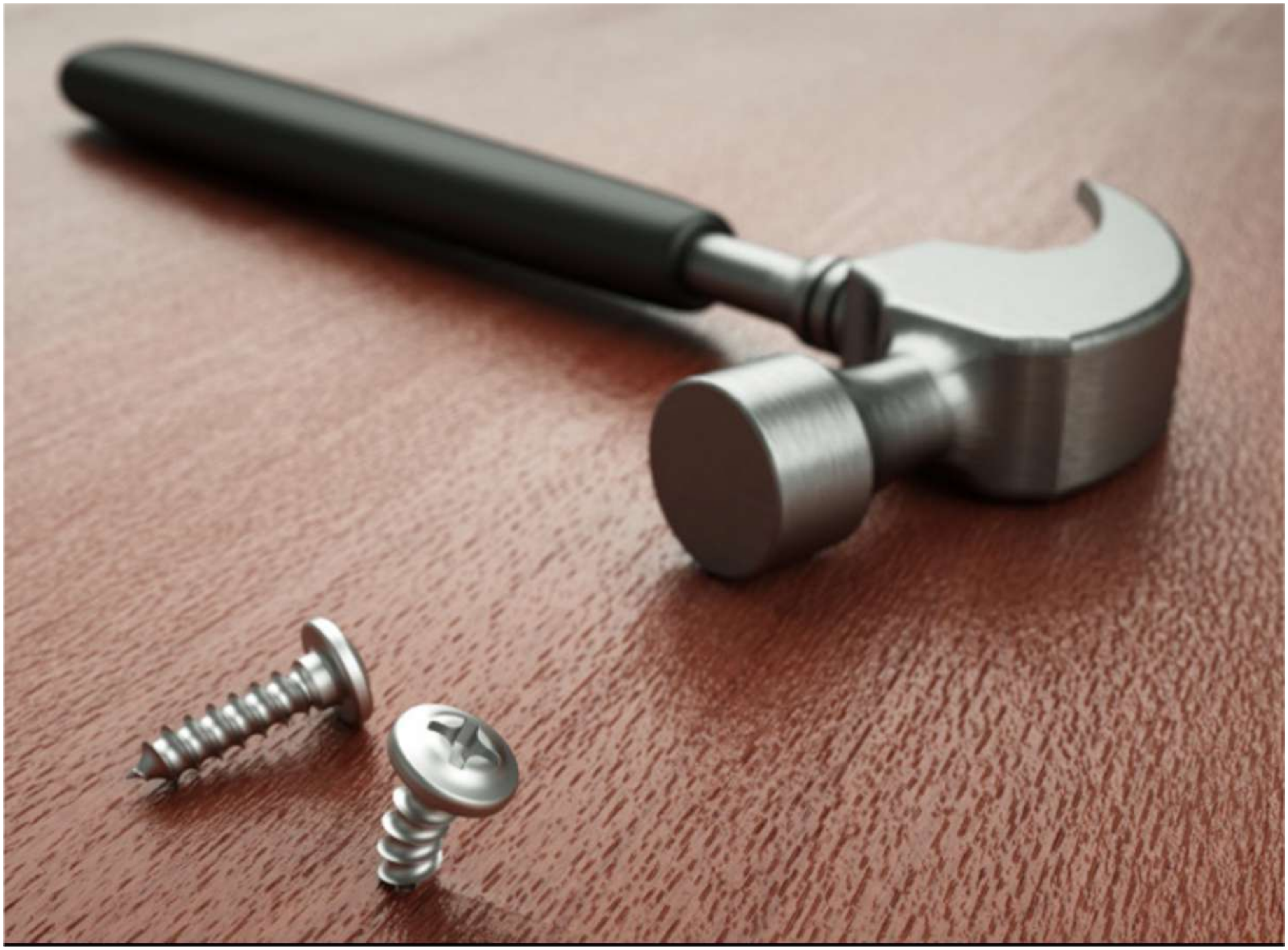
Knowledge Sources

Multiple Positions Available:

- Interns
- Postdocs
- Research Engineers
- Research Staff Members

Contact:
pzhang@us.ibm.com

Thank you!!!



"When you have a hammer, everything looks like a nail"