Introduction to Systems Biology II

Charles Blatti

NIH BD2K KnowEnG Center of Excellence in Big Data Computing
Carl R. Woese Institute for Genomic Biology
University of Illinois at Urbana-Champaign

June 13, 2019

Slides By Amin Emad

Systems Biology

- Systems biology is the computational and mathematical modeling of complex biological systems.

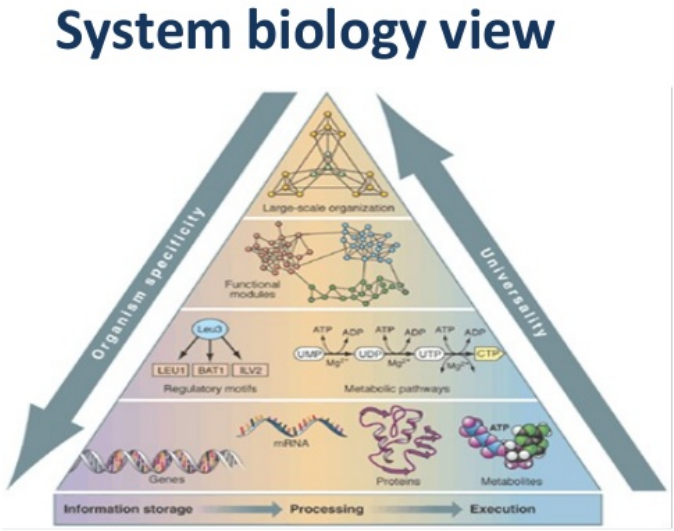


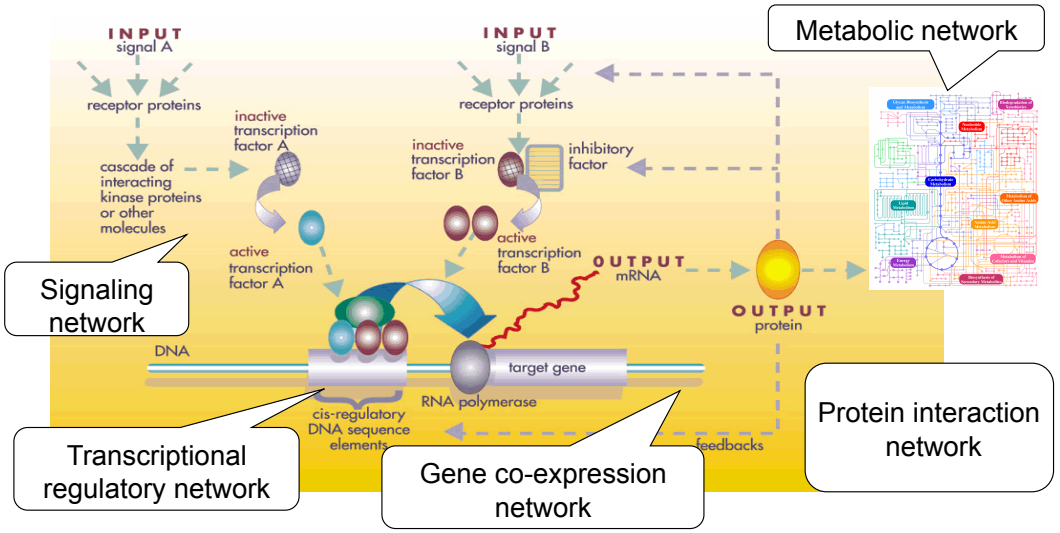
Figure from Oltvai , Z.N. and Barabasi
Life's complexity pyramid.

- Studies the interactions between the components of biological systems such as genes, proteins, metabolites, etc. (i.e. biological networks), and how these interactions give rise to the function and behavior of that system (phenotype)

Biological Networks

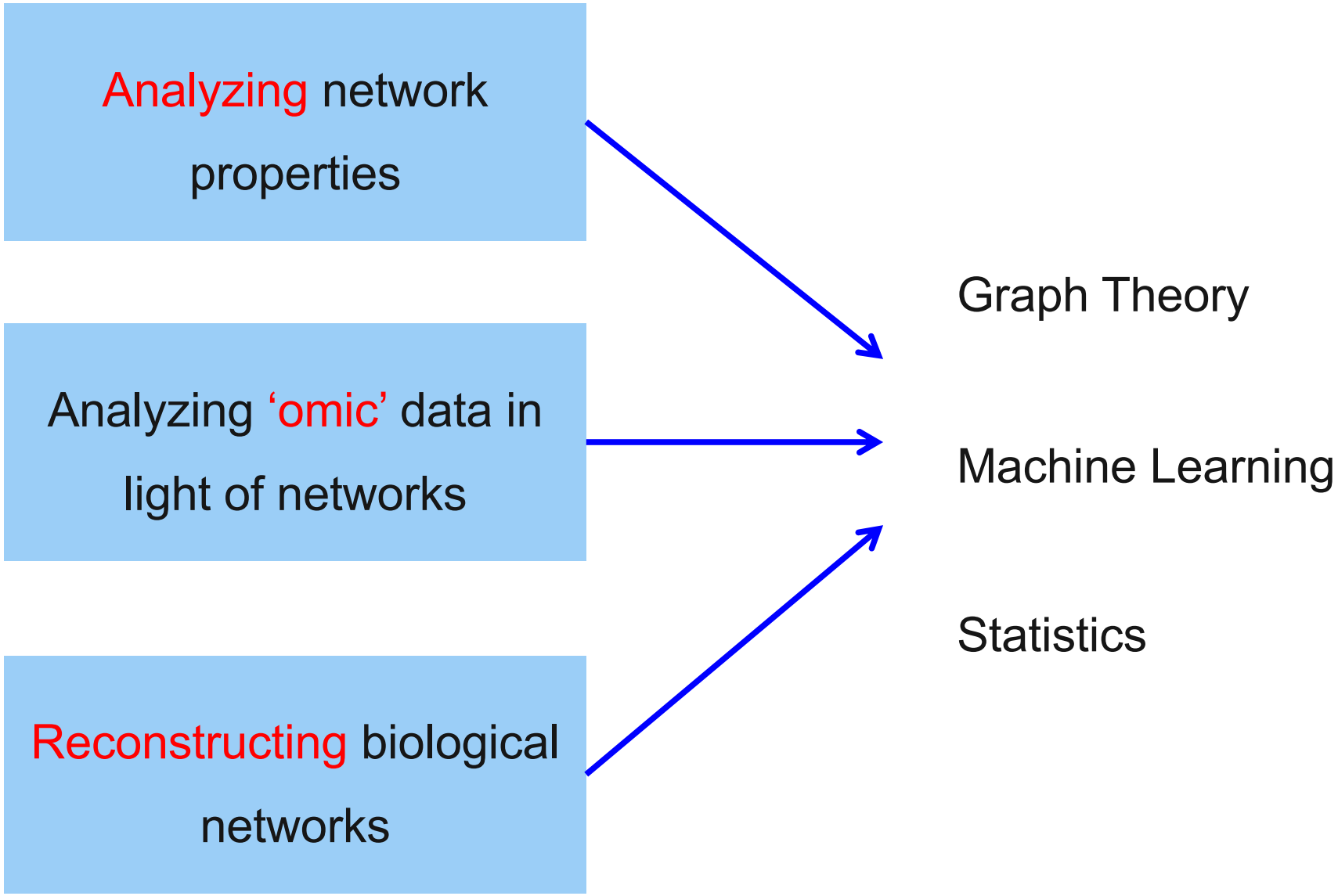
A graphical representation of the interactions of the components of a biological systems

- Cell signaling networks
- Gene regulatory networks
- Protein-protein interaction networks
- Gene co-expression networks
- Metabolic networks



Zhang (2009)

Biological Networks in Computational Biology

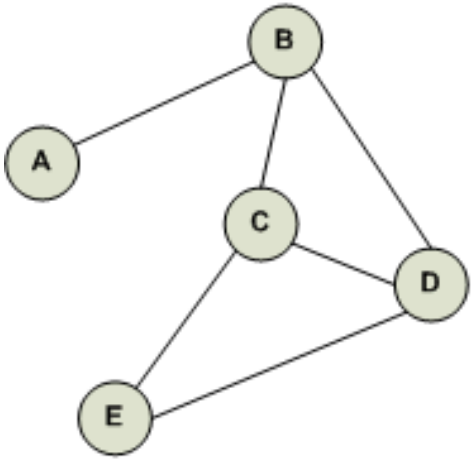


1) Analyzing network properties

What is a network/graph?

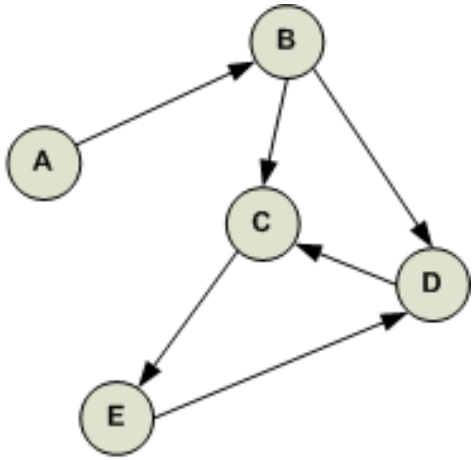
- Graph: A representation of relationship among objects
- A graph *G(V, E)* is a set of vertices (nodes) *V* and edges (links) *E*

Directed vs. Undirected:



Undirected graph

- Protein-protein interactions
- Co-expression network

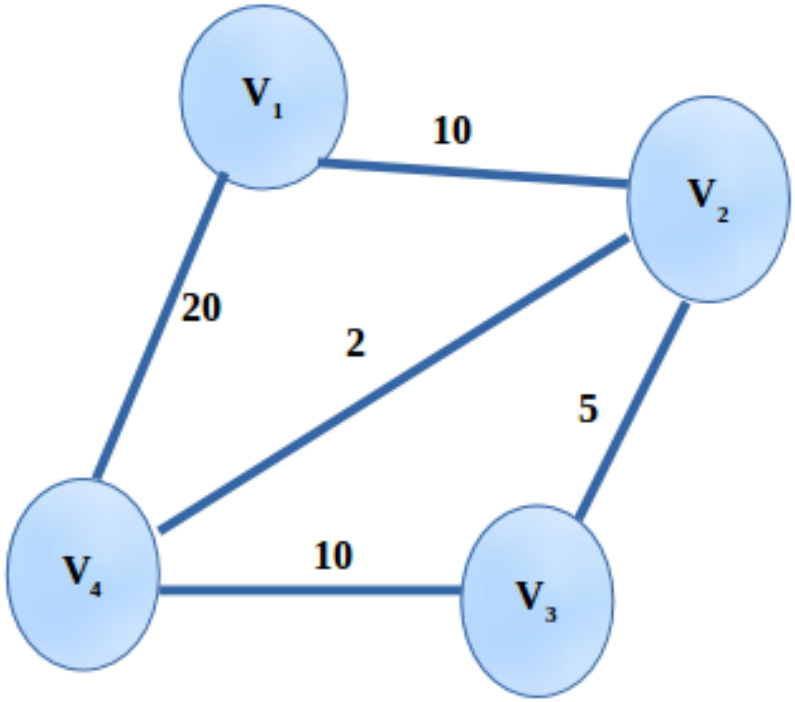


Directed graph

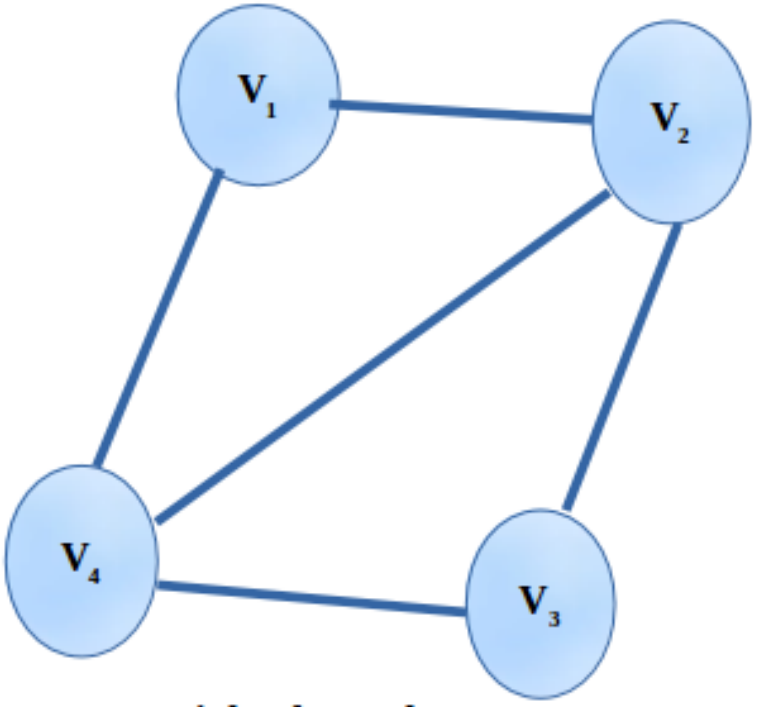
- Gene regulatory network
- Signaling pathways

Graph Properties

Weighted vs. Unweighted:



Weighted graph



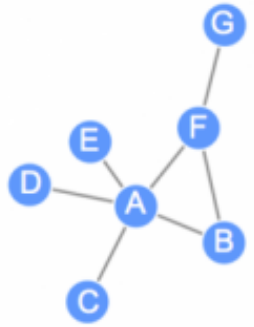
Unweighted graph

- Weights represent affinity in PPI, correlation coefficient in a co-expression network, confidence in a GRN, etc.

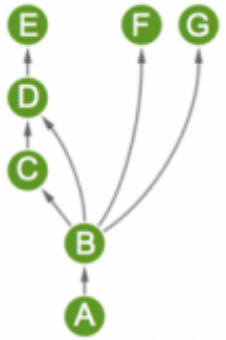
Graph Properties

Degree and degree distribution:

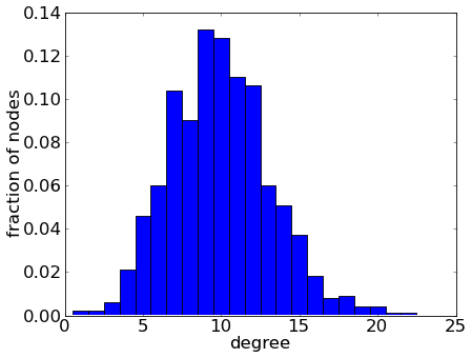
- Degree: Number of connections of a node to other nodes



- Indegree (outdegree) of a node in a directed graph is the number of edges entering (leaving) that node



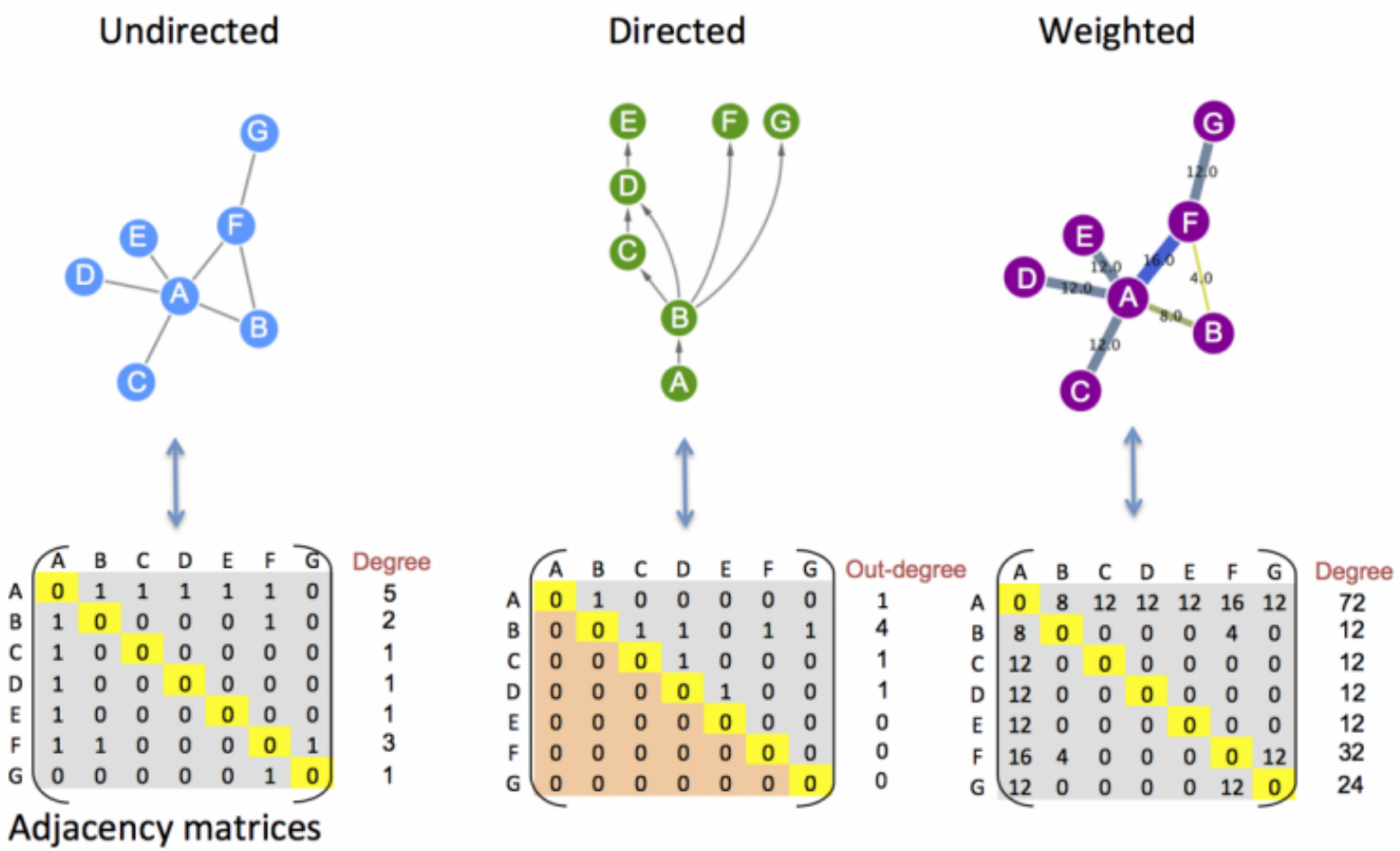
- Degree distribution of a network is the probability distribution of these degrees over the network:



Graph Properties

Adjacency matrix:

- A matrix representation of the graph

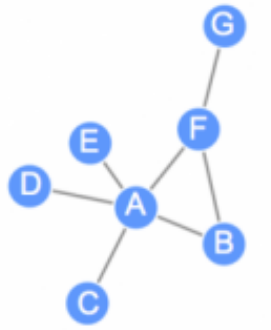


https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/introduction-graph-theory/graph-0

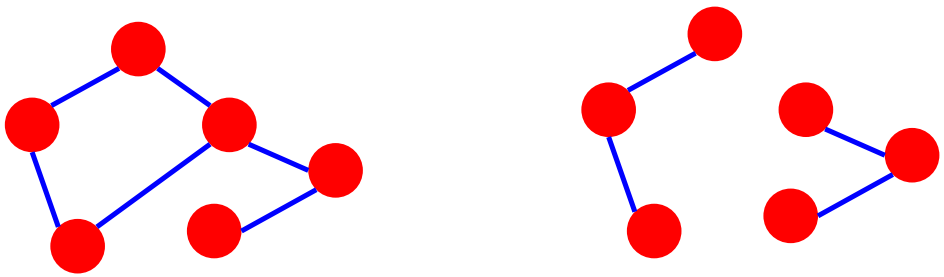
Graph Properties

Path and connectivity:

- Path: A sequence of distinct edges connecting a sequence of vertices: GFAB, EAC, etc.



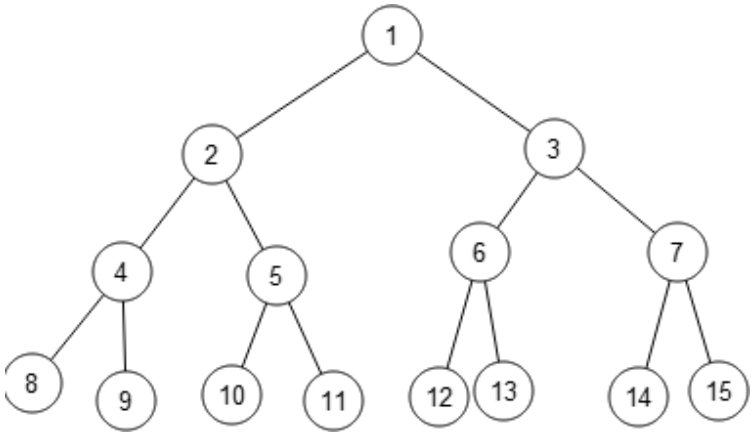
- Connectivity: A graph that in which a path exists between any two nodes



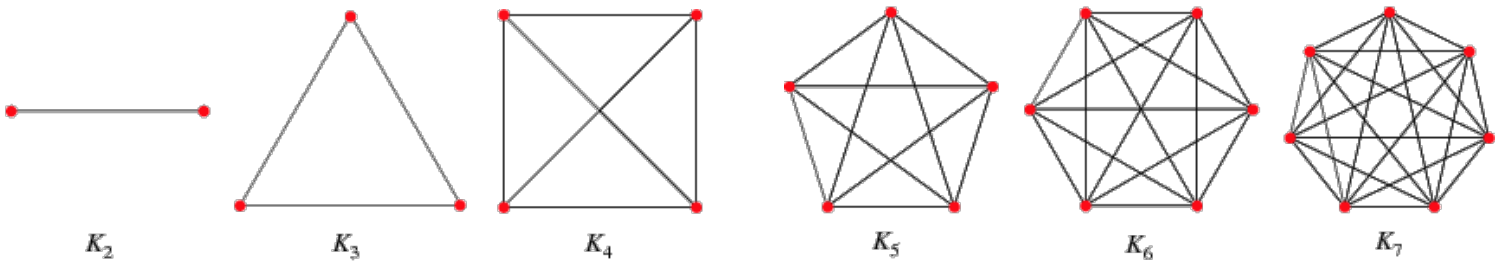
Graph Properties

Important classes of graphs:

- Tree: Any two vertices are connected by exactly one path (e.g. dendogram in hierarchical clustering)



- Complete graph: Each pair of vertices are connected by an edge



2) Analyzing ‘omic’ data in light of biological networks

Analyzing ‘omic’ data in light of networks

How to analyze large ‘omic’ datasets?

Statistics

Machine Learning

Analyzing ‘omic’ data in light of networks

How to analyze large ‘omic’ datasets?

Machine learning is concerned with utilizing statistical techniques to give computers the ability to “learn”.

Analyzing ‘omic’ data in light of networks

How to analyze large ‘omic’ datasets?

Machine learning is concerned with utilizing statistical techniques to give computers the ability to “learn”.

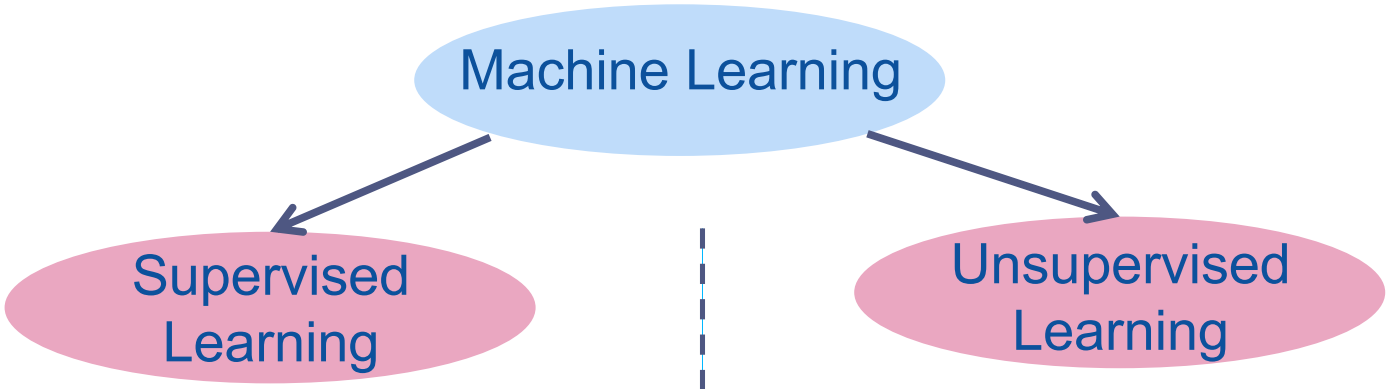
However, it can do much more!

Machine Learning in Computational Biology

Some examples:

- Predicting whether a patient is sensitive or resistant to a drug
- Predicting the survival probability of a cancer patient
- Identifying the subtypes of a disease
- Identifying genes associated with a disease
- etc.

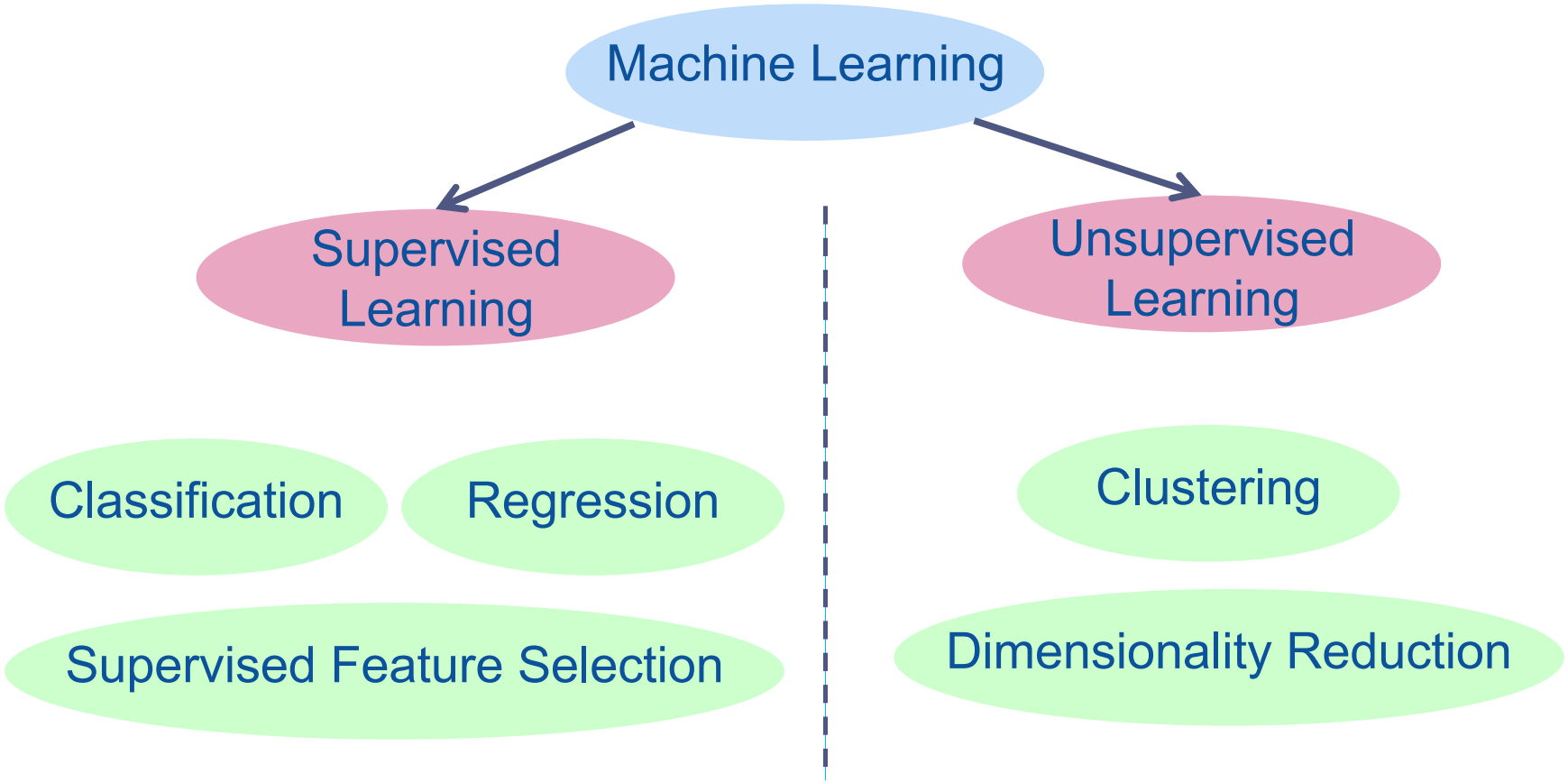
Machine Learning



Training examples are provided with desired inputs and outputs to help learning the desired rule

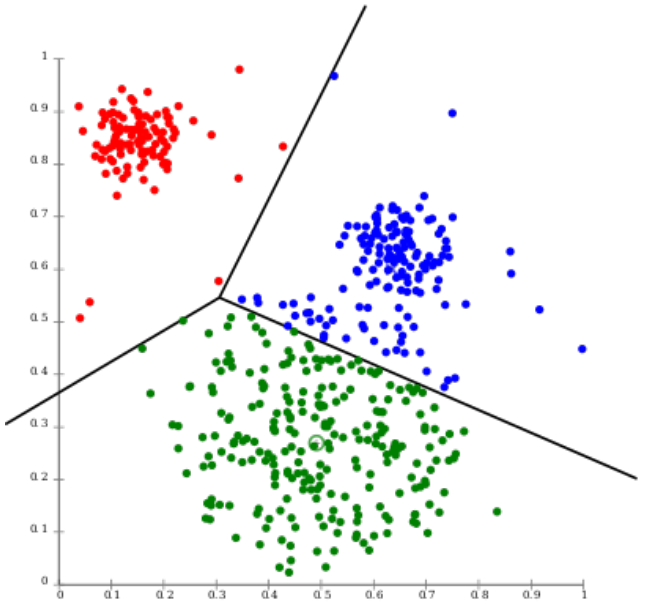
No training example exists and the goal is to learn structure in the data

Machine Learning



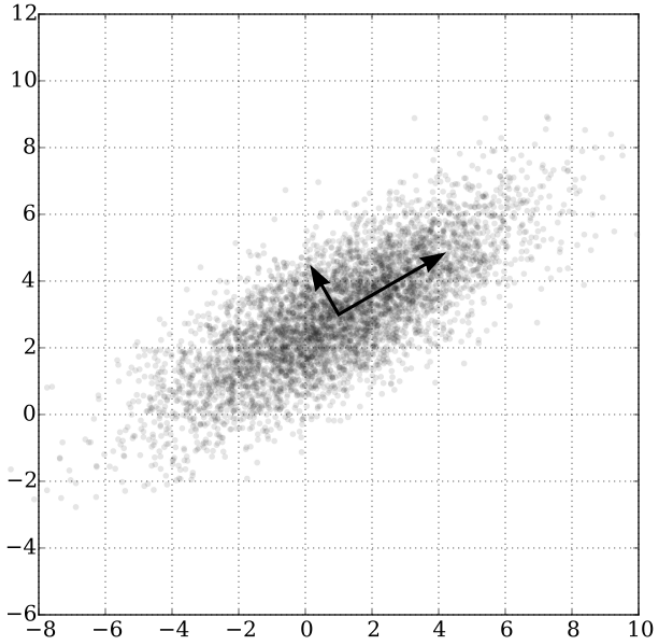
Unsupervised Machine Learning (Clustering)

- We have a set of samples characterized using several features (e.g. expression of thousands of genes for tumor samples)
- Goal: Group the sample such that those in the same group are more similar to each other than to those in other groups
- Many methods exist such as K-means, hierarchical clustering, matrix factorization, etc.
- Example: Identifying subtypes of breast cancer using transcriptomic data



Unsupervised ML (Dimensionality Reduction)

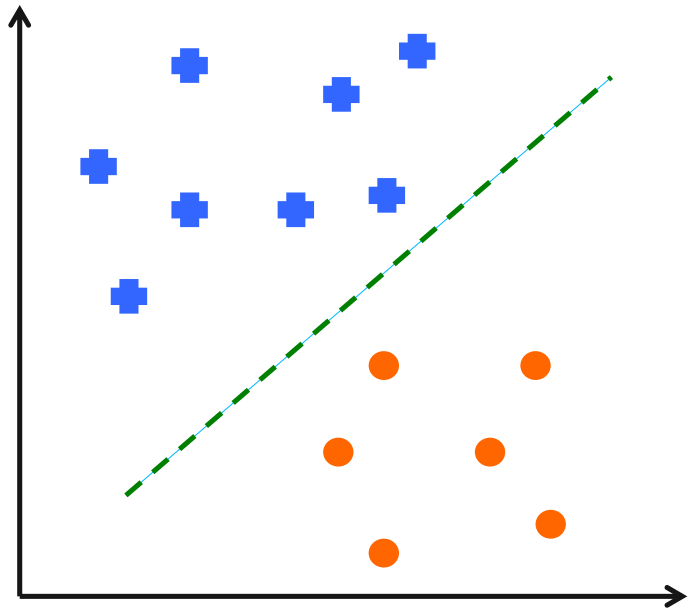
- We have a set of samples characterized using several features
- Goal: Reduce the number of features while preserving characteristics of the data
- Many methods exist such as principal component analysis, linear discriminative analysis, etc.
- Example: PCA identifies a few principal components, orthogonal to each other, such that they account for most of the variance in the data



Supervised Machine Learning (Classification)

Classification:

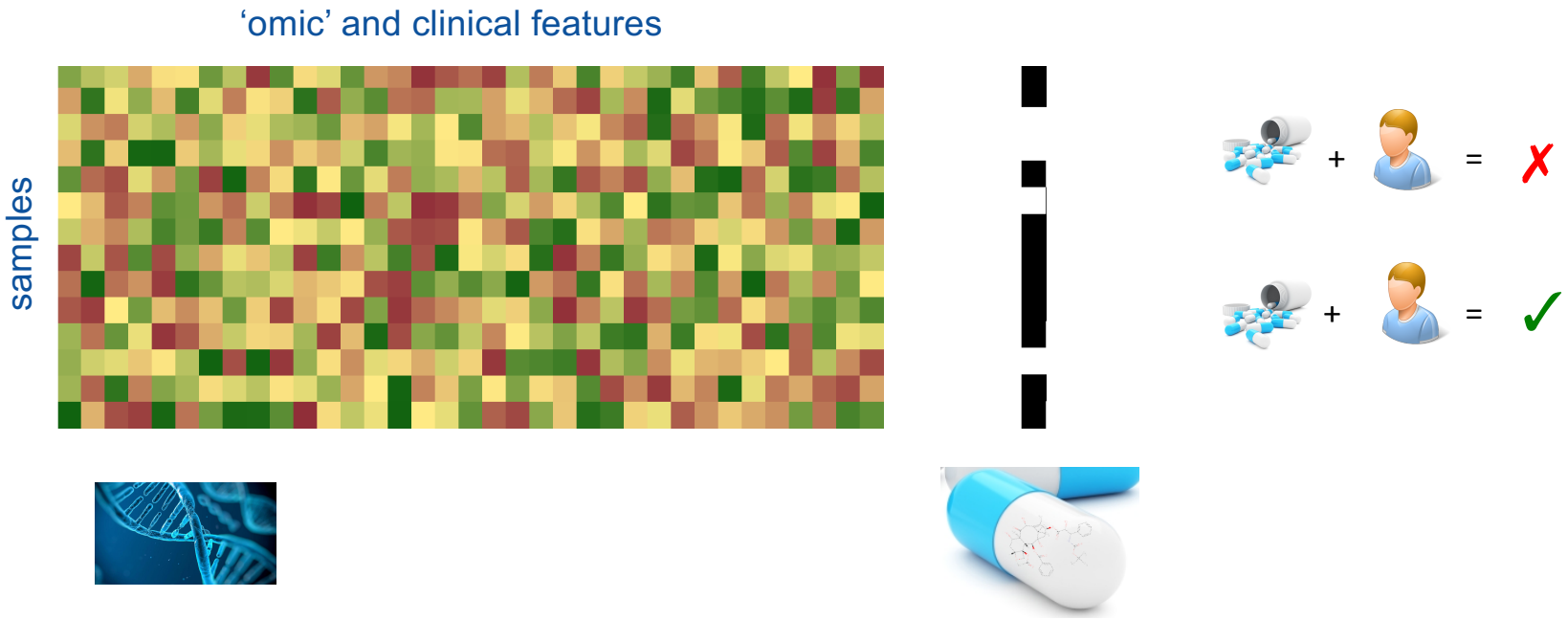
- We have a set of samples characterized using several features (e.g. expression of thousands of genes for tumor samples)
- The samples belong to set of known categories
- Goal: Given a new sample, to which category does it belong?
- Many methods exist such as KNN, SVM, logistic regression, decision trees, random forests, etc.



Supervised Machine Learning (Classification)

Example:

- We have ‘omic’ profiles and clinical information of breast cancer patients
- We also know which patients were resistant to a drug and which ones were not

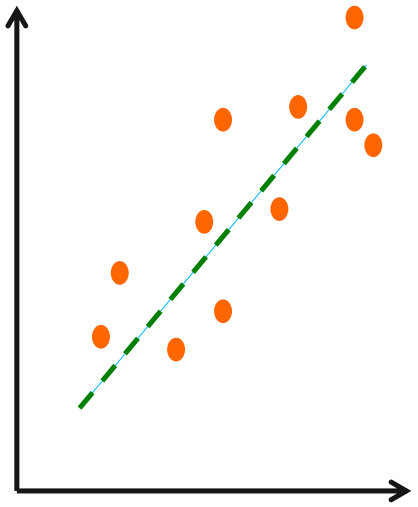


- Given the ‘omic’ profiles and clinical information of a new patient, will they be resistant to the drug or not?

Supervised Machine Learning (Regression)

- We have a set of samples characterized using several features (e.g. expression of thousands of genes for tumor samples)

- For each sample, we know a continuous-valued response (dependent variable) (e.g. number of years between diagnosis and occurrence of metastasis)



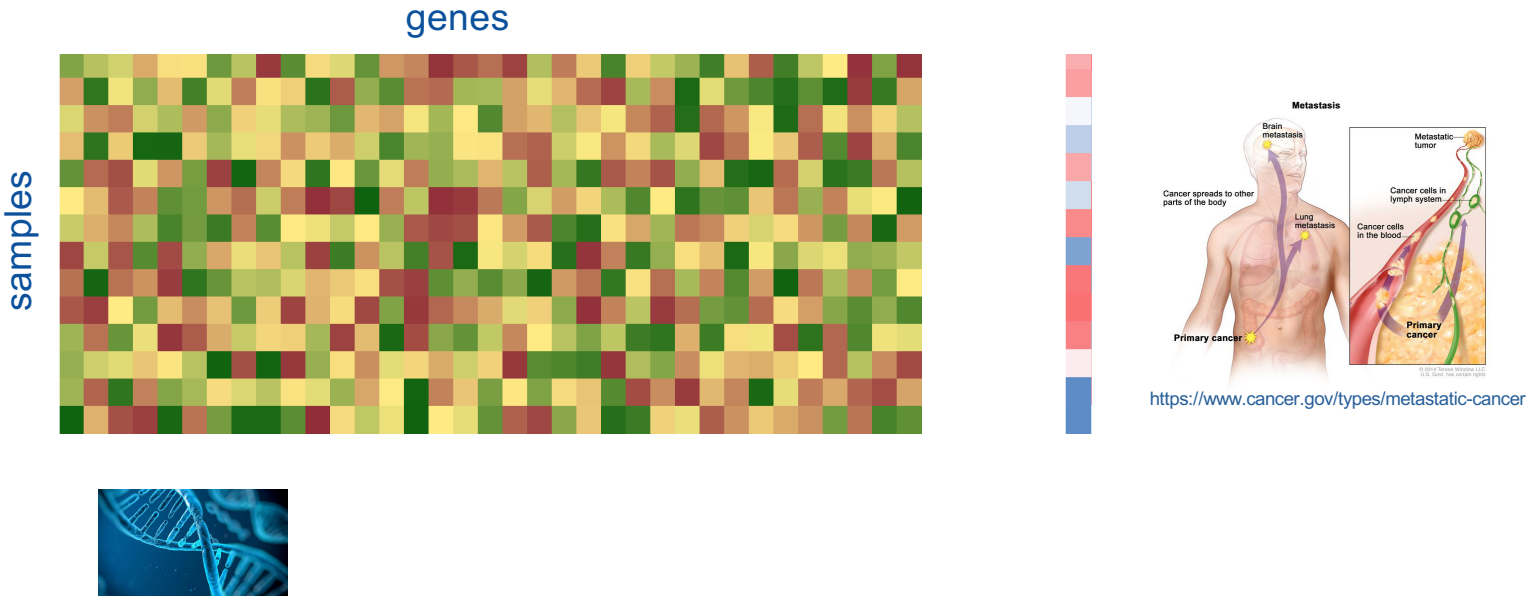
- Goal: Estimate the relationship between the response and features and predict the value of response for a new sample

- Many methods exist such as linear regression, LASSO, Elastic Net, Support vector regression, etc.

Supervised Machine Learning (Regression)

Example:

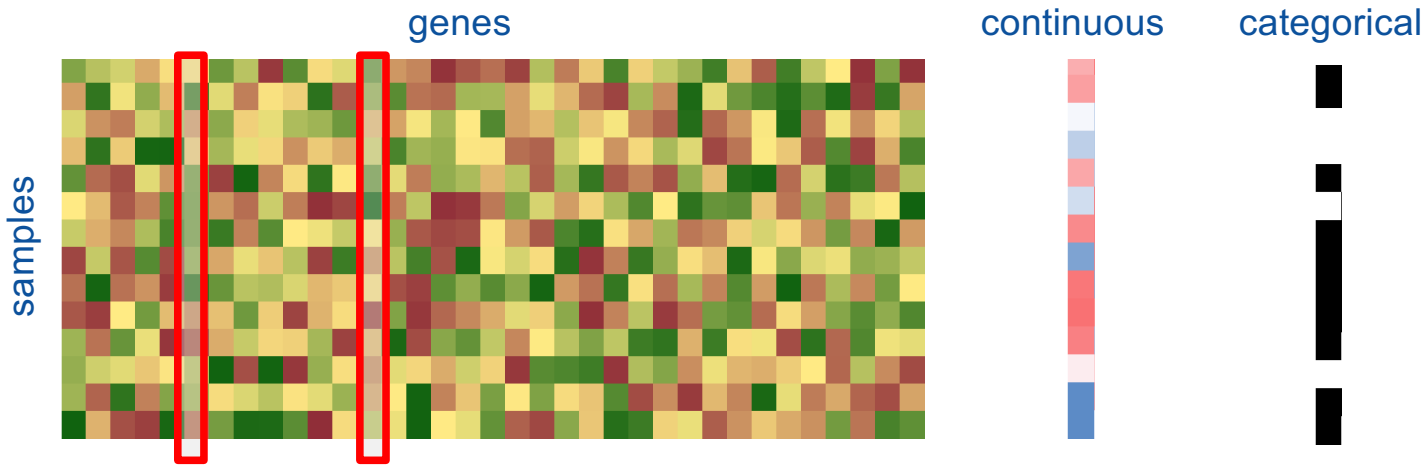
- We have transcriptomic profiles of breast cancer patients
- We also know number of months between diagnosis and occurrence of metastasis



- What is the relationship between gene expression and time of metastasis?

Supervised Machine Learning (Feature Selection)

- We have a set of samples characterized using several features
- We know a continuous-valued or categorical response for samples
- Goal: What are the features most predictive of the response?



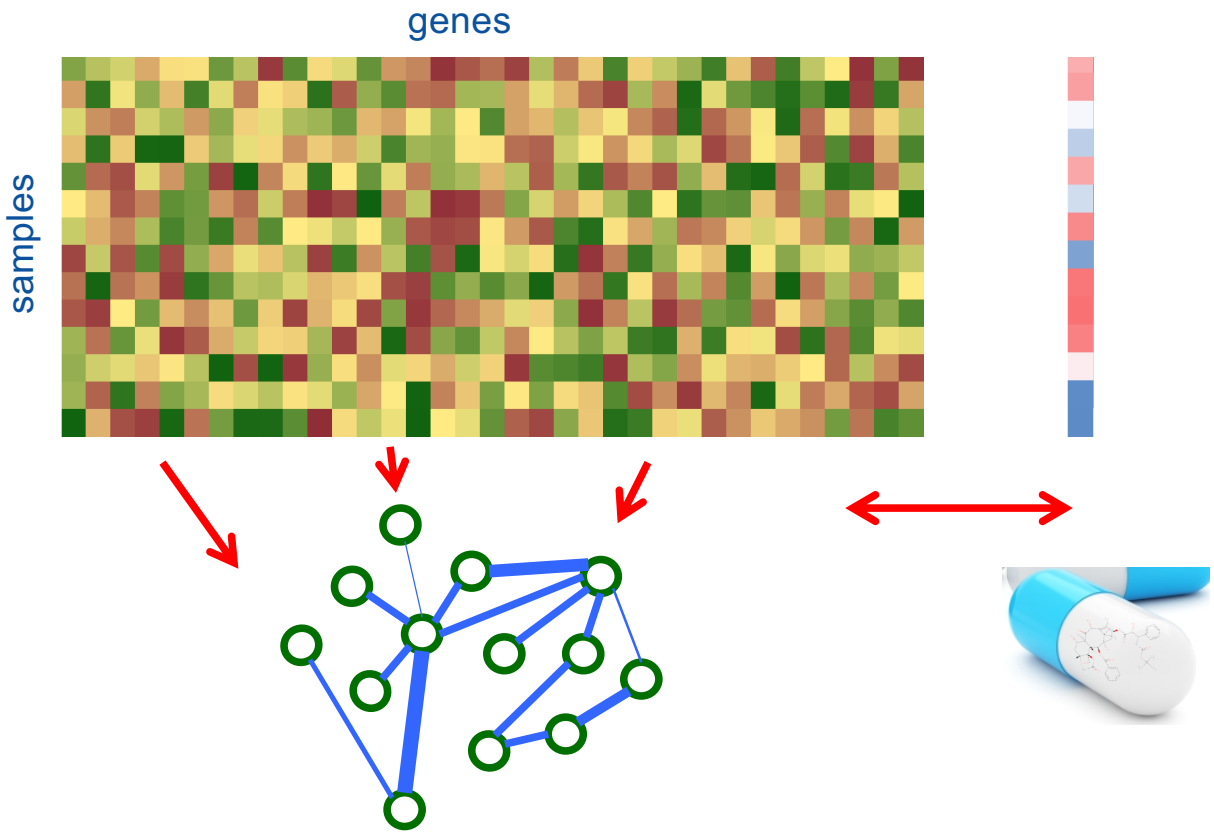
Examples:

- Differentially expressed genes (case vs. control)
- Correlation analysis (GWAS)
- etc.

Network guided analysis

How can biological networks help?

- When features correspond to genes or proteins (e.g. gene expression, mutation, etc.), these networks can provide information regarding the interactions and relationships of these features.

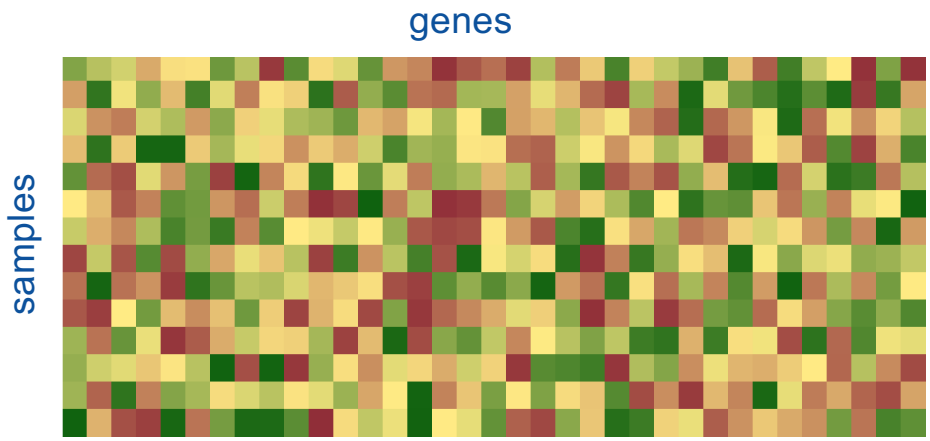


Network-guided gene prioritization using ProGENI

Background

- Phenotypic properties of a cell are determined (partially) by expression of its genes and proteins

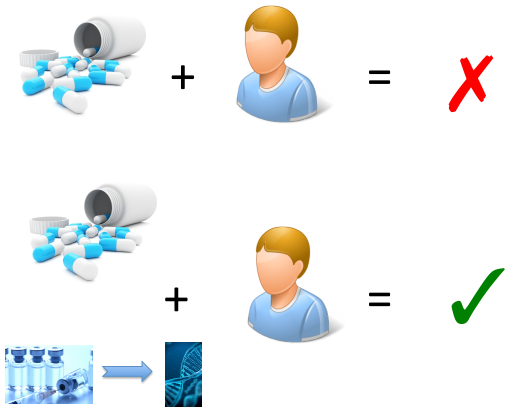
- Gene expression profiling measures the activity of thousands of genes to create a global picture of cellular function



Background

- Goal:
  - Identifying genes whose basal mRNA expression determines the drug sensitivity in different samples (supervised feature selection)

- Motivations:
  - Overcoming drug resistance
  - Revealing drug mechanism of action
  - Identifying novel drug targets
  - Predicting drug sensitivity of individuals



Gene prioritization

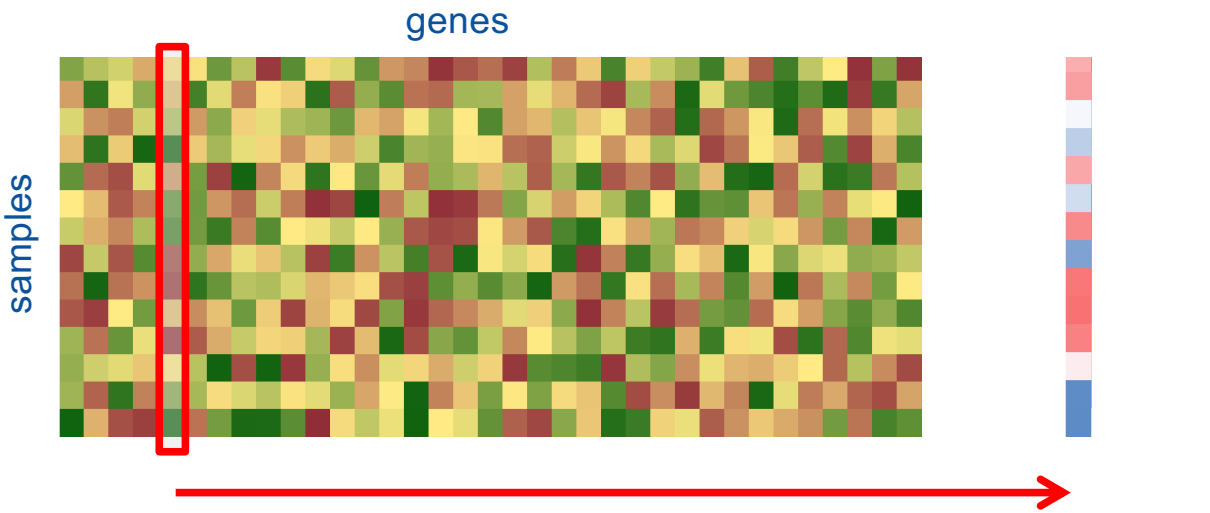
Examples of current methods:

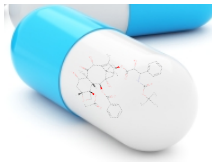
- Score each gene based on the correlation of its expression with drug response

Nat Chem Biol. 2016 Feb;12(2):109-16. doi: 10.1038/nchembio.1986. Epub 2015 Dec 14.

Correlating chemical sensitivity and basal gene expression reveals mechanism of action.

Rees MG[1], Seashore-Ludlow B[1,2], Cheah JH[1,2], Adams DJ[1,2], Price EV[1,2], Gill S[1], Javaid S[3], Coletti ME[1], Jones VL[1], Bodycombe NE[1,2], Soule CK[1,2], Alexander B[1], Li A[1], Montgomery P[1], Kotz JD[1], Hon CS[1], Munoz B[1], Liefeld T[1,2], Dančík V[1], Haber DA[3], Clish CB[1], Bittker JA[1], Palmer M[1,2], Wagner BK[1], Clemons PA[1], Shamji AF[1], Schreiber SL[1].





Gene prioritization

Examples of current methods:

- Score each gene based on the correlation of its expression with drug response

- Use multivariable regression algorithms such as Elastic Net to relate multiple genes' expression values to drug response

Nat Chem Biol. 2016 Feb;12(2):109-16. doi: 10.1038/nchembio.1986. Epub 2015 Dec 14.

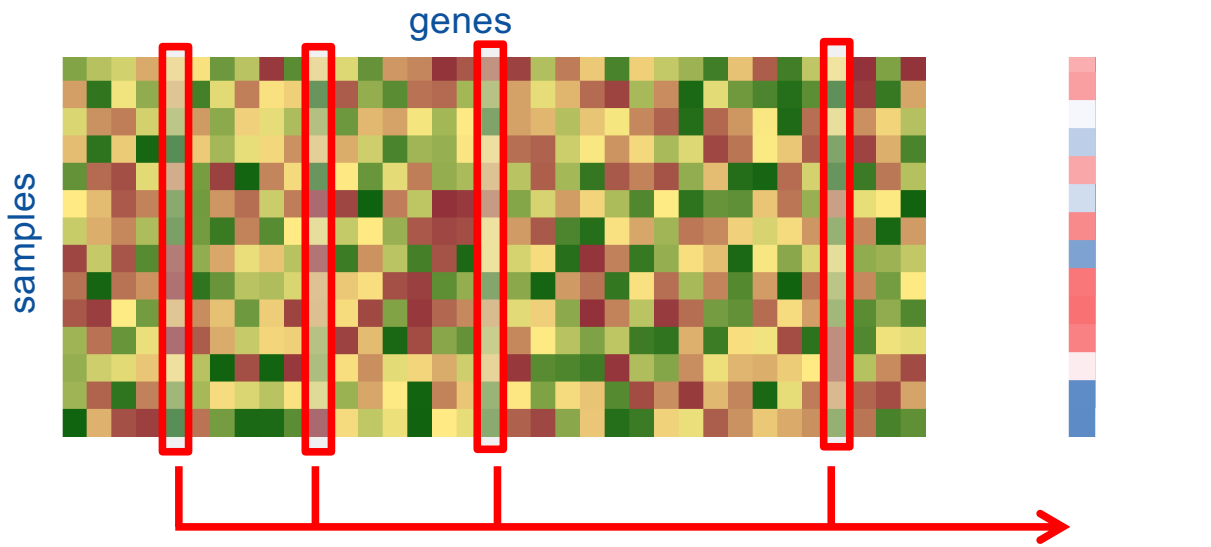
Correlating chemical sensitivity and basal gene expression reveals mechanism of action.

Rees MG[1], Seashore-Ludlow B[1,2], Cheah JH[1,2], Adams DJ[1,2], Price EV[1,2], Gill S[1], Javaid S[3], Coletti ME[1], Jones VL[1], Bodycombe NE[1,2], Soule CK[1,2], Alexander B[1], Li A[1], Montgomery P[1], Kotz JD[1], Hon CS[1], Munoz B[1], Liefeld T[1,2], Dančík V[1], Haber DA[3], Clish CB[1], Bittker JA[1], Palmer M[1,2], Wagner BK[1], Clemons PA[1], Shamji AF[1], Schreiber SL[1].

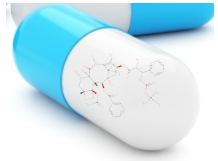
Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.

Barretina J[1], Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA.



$$\sum w_i \mathbf{x}_i$$



Gene prioritization

Examples of current methods:

- Score each gene based on the correlation of its expression with drug response

- Use multivariable regression algorithms such as Elastic Net to relate multiple genes' expression values to drug response

Nat Chem Biol. 2016 Feb;12(2):109-16. doi: 10.1038/nchembio.1986. Epub 2015 Dec 14.

Correlating chemical sensitivity and basal gene expression reveals mechanism of action.

Rees MG[1], Seashore-Ludlow B[1,2], Cheah JH[1,2], Adams DJ[1,2], Price EV[1,2], Gill S[1], Javaid S[3], Coletti ME[1], Jones VL[1], Bodycombe NE[1,2], Soule CK[1,2], Alexander B[1], Li A[1], Montgomery P[1], Kotz JD[1], Hon CS[1], Munoz B[1], Liefeld T[1,2], Dančík V[1], Haber DA[3], Clish CB[1], Bittker JA[1], Palmer M[1,2], Wagner BK[1], Clemons PA[1], Shamji AF[1], Schreiber SL[1].

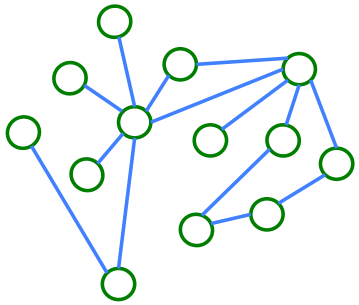
Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.

Barretina J[1], Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA.

Shortcoming:

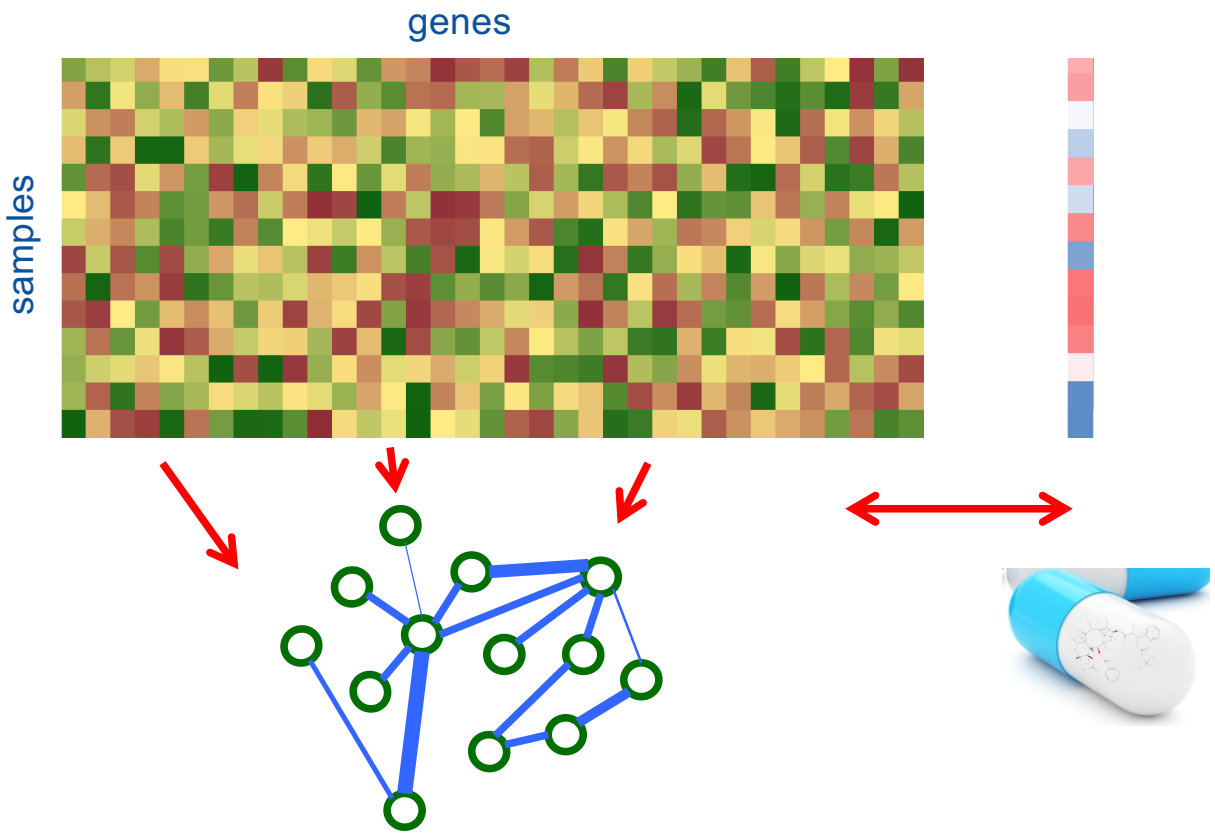
- These methods do not incorporate prior information about the interaction of the genes



ProGENI

Hypothesis:

- Since genes and proteins involved in drug MoA are functionally related, prior knowledge in the form of gene interaction network (e.g. PPI) can improve accuracy of the prioritization task



ProGENI: Network-guided gene prioritization

- An algorithm that incorporates gene network information to improve prioritization accuracy

Genome Biology

Featured article: new insights into mechanisms of chemoresistance

Emad *et al. Genome Biology* (2017) 18:153
DOI 10.1186/s13059-017-1282-3

Genome Biology

RESEARCH Open Access

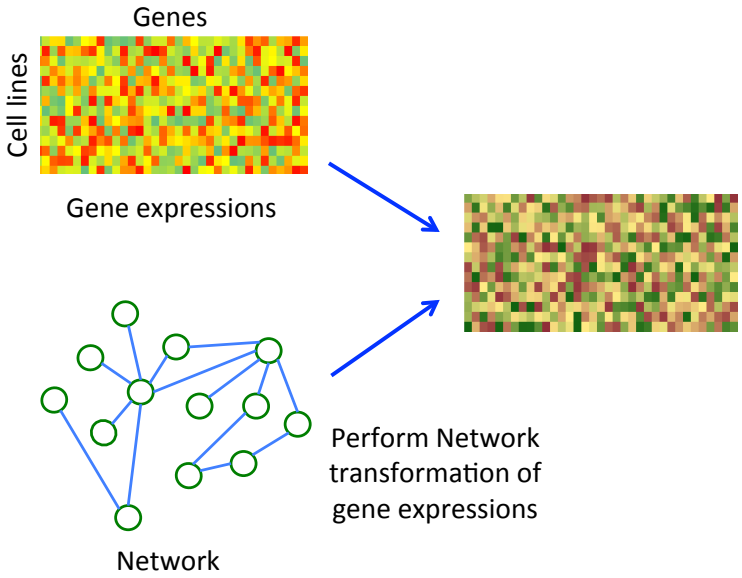
CrossMark

Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance

Amin Emad[1], Junmei Cairns[2], Krishna R. Kalari[3], Liewei Wang[2*] and Saurabh Sinha[4*]

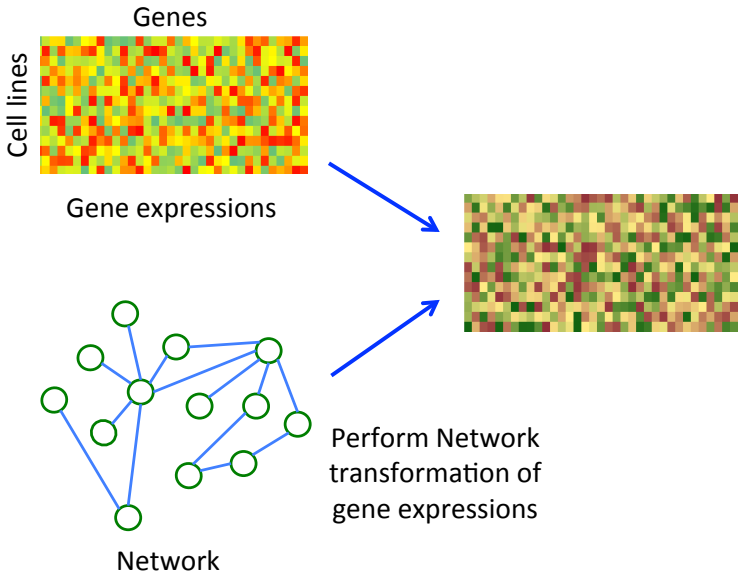
ProGENI

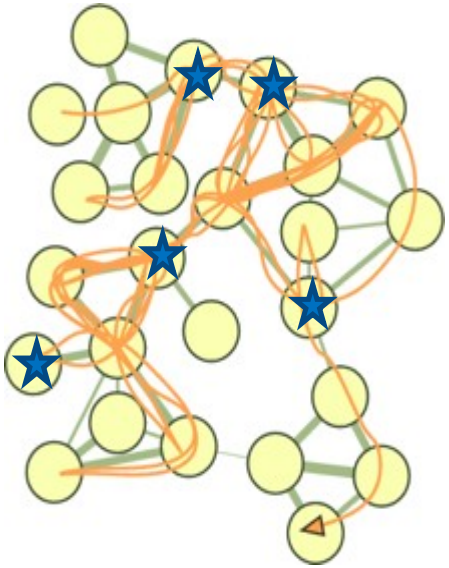
Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance



ProGENI

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance



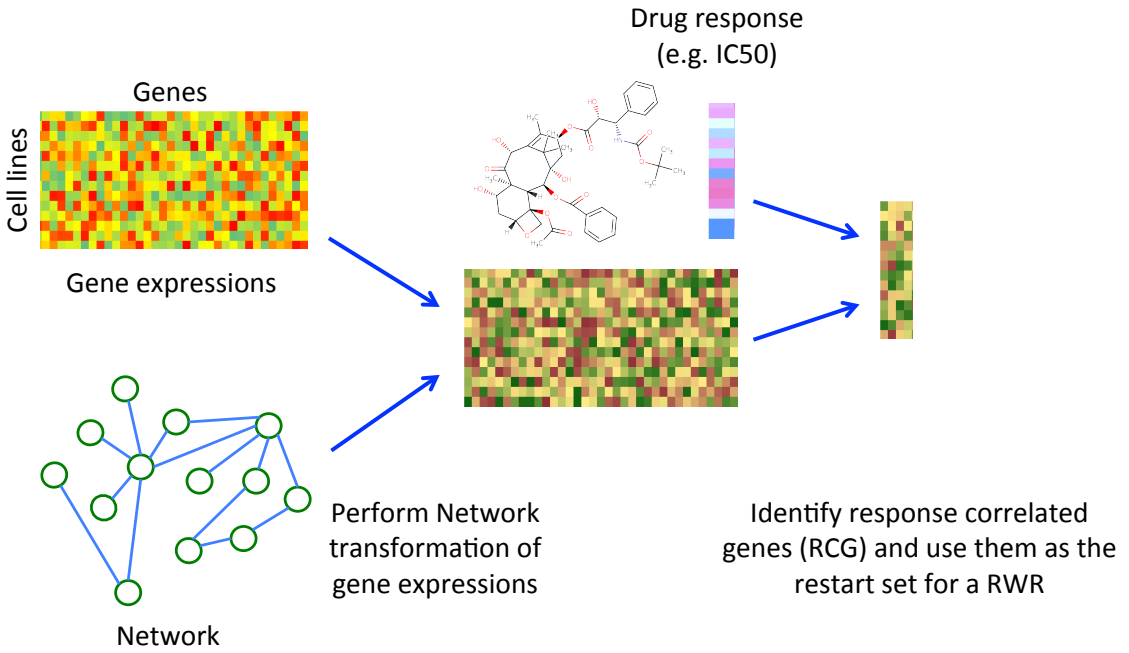


(Rosvall and Bergstrom, 2007)

ProGENI

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

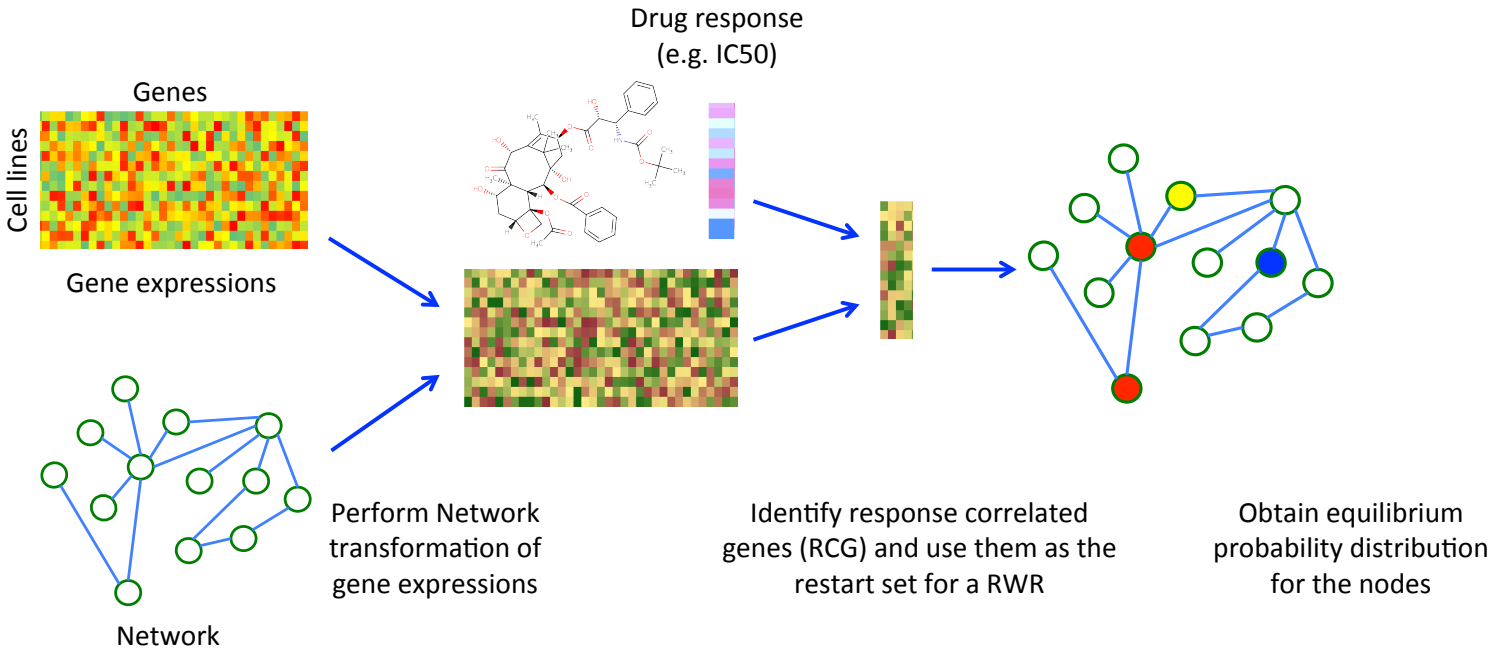


ProGENI

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

Step 3: Score genes based on their relevance to the RCG set



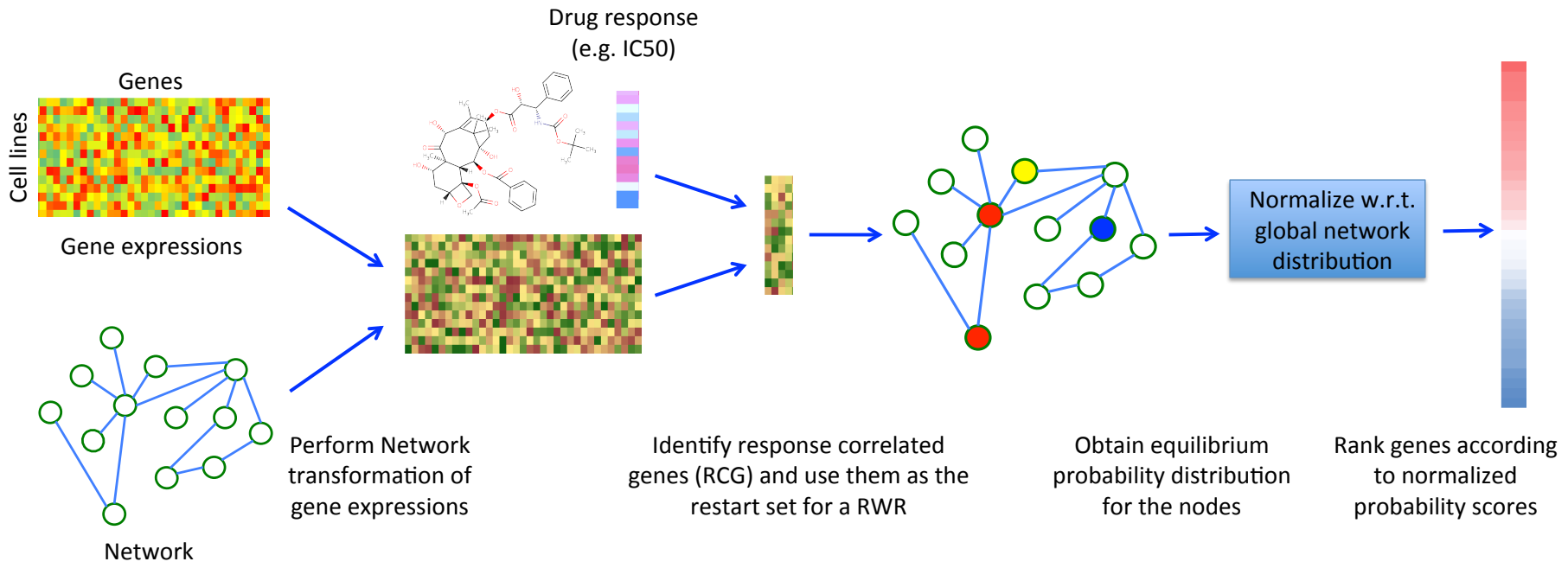
ProGENI

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

Step 3: Score genes based on their relevance to the RCG set

Step 4: Remove network bias by normalizing scores w.r.t. scores corresponding to global network topology



Datasets

- Human lymphoblastoid cell lines (LCL)
  - Gene expression (~17K genes of ~300 cell lines)
  - Drug response of 24 cytotoxic treatments

MAYO CLINIC

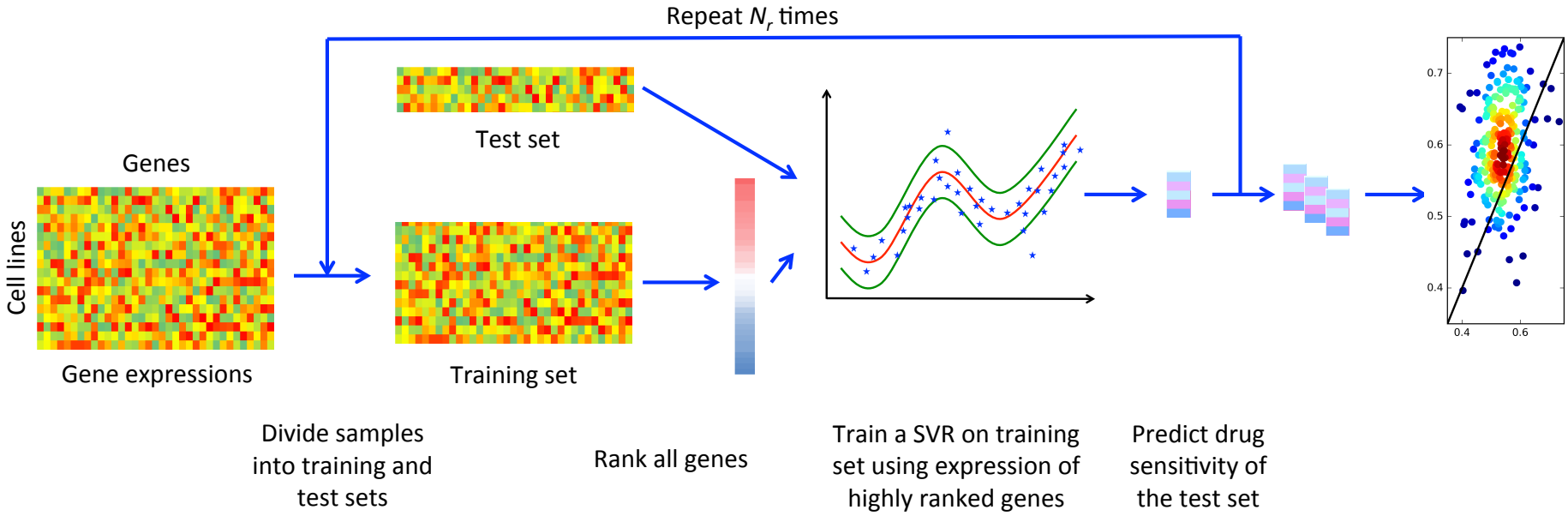
- Publicly available dataset from GDSC
  - Gene expression (~13K genes of ~600 cell lines from 13 tissues)
  - Drug response of 139 cytotoxic treatments

- Publicly available prior knowledge
  - Network of gene interactions (PPI and genetic interactions) from STRING (~1.5M edges, ~15.5K nodes)

STRING

Validation using drug response prediction

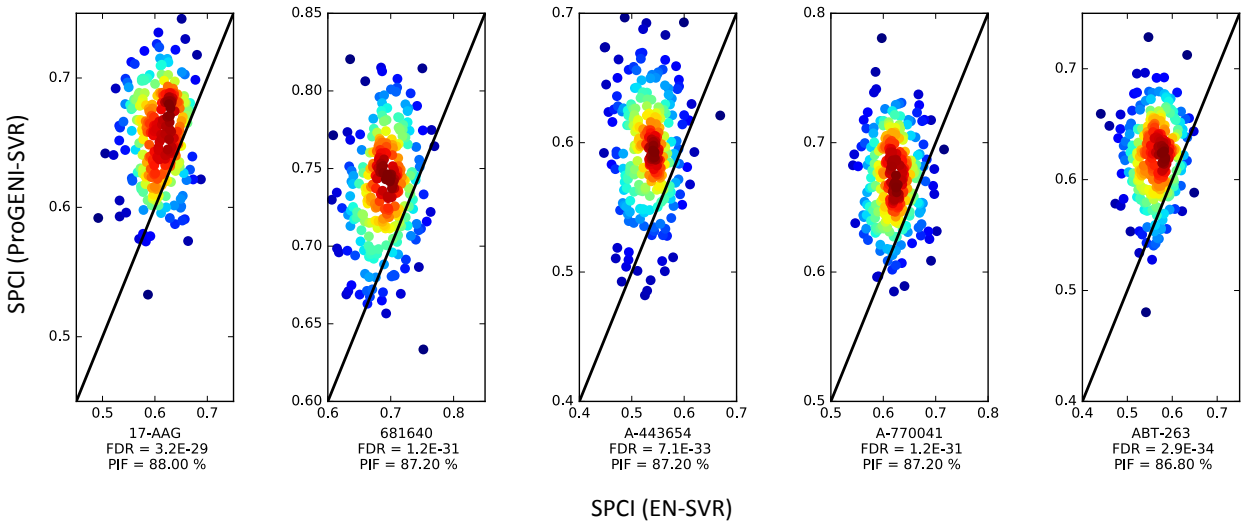
- Genes ranked highly using a good prioritization method are good predictors of drug sensitivity



Validation using drug response prediction

LCL Dataset	Pearson	Elastic Net
Num. Drugs (out of 24) ProGENI > Baseline	14	20
FDR (Wilcoxon signed-rank test)	6.5 E-3	9.6 E-5

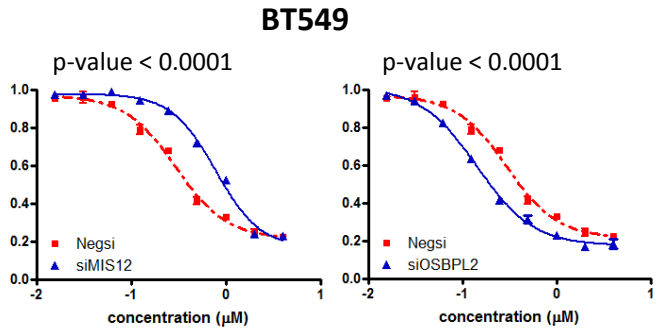
GDSC Dataset	Pearson	Elastic Net
Num. Drugs (out of 139) ProGENI > Baseline	66	110
FDR (Wilcoxon signed-rank test)	9.1 E-4	4.0 E-21

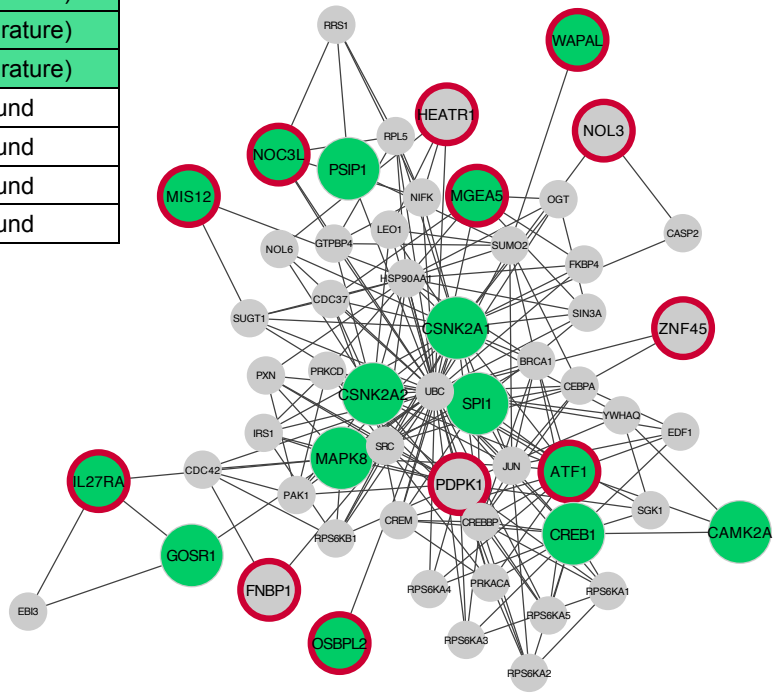


Functional validation

We validated role of 33 (out of 45) genes (73%) for three drugs.

Gene Symbol	Rank (ProGENI)	Rank (Pearson)	Absolute value of Pearson correlation coefficient	Evidence
ATF1	1	1	0.2000	Direct (this study)
MIS12	2	4	0.1887	Direct (this study)
OSBPL2	5	6	0.1865	Direct (this study)
CSNK2A1	7	1587	0.0752	Direct (literature)
PSIP1 (LEDGF)	8	46	0.1537	Direct (literature)
CAMK2A	9	6991	0.0157	Direct (literature)
CSNK2A2	10	4870	0.0347	Direct (literature)
GOSR1	11	6867	0.0167	Direct (this study)
MAPK8	13	7574	0.0112	Direct (literature)
SPI1	14	6287	0.0217	Direct (literature)
CREB1	15	665	0.1000	Direct (literature)
NOC3L	3	3	0.1893	Not found
IL27RA	4	2	0.1911	Not found
MGEA5	6	7	0.1814	Not found
WAPAL	12	8	0.1805	Not found





How about other ML tasks?

- Similar principles can be used for ML tasks other than feature selection/prioritization
- "Network-smoothing" of the features used in ProGENI can be used as a preprocessing step to regression and classification algorithms
- Network-smoothing can also be used for clustering and dimensionality reduction (e.g. Network-based stratification)

Nat Methods. 2013 Nov;10(11):1108-15. doi: 10.1038/nmeth.2651. Epub 2013 Sep 15.

Network-based stratification of tumor mutations.

Hofree M[1], Shen JP, Carter H, Gross A, Ideker T.

Network-based Stratification

Goal:

- Stratification (clustering) of tumor samples based on somatic mutation profiles

Main Issue:

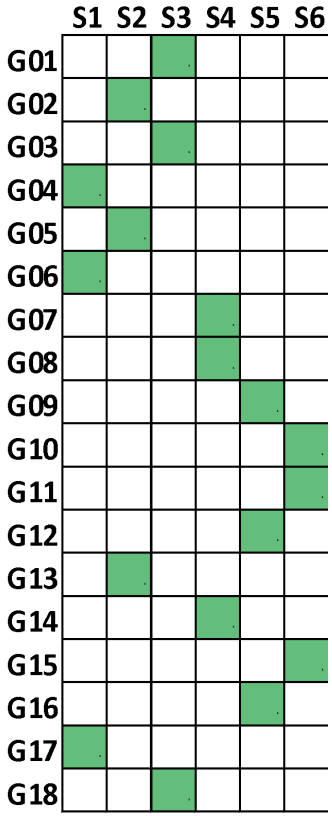
- The mutation data is very sparse and most conventional clustering techniques fail to identify reasonable patterns

- Although two tumors may not share the same somatic mutations, they may affect the same pathways and interaction networks

Value of network-guided analysis

Data sparsity:

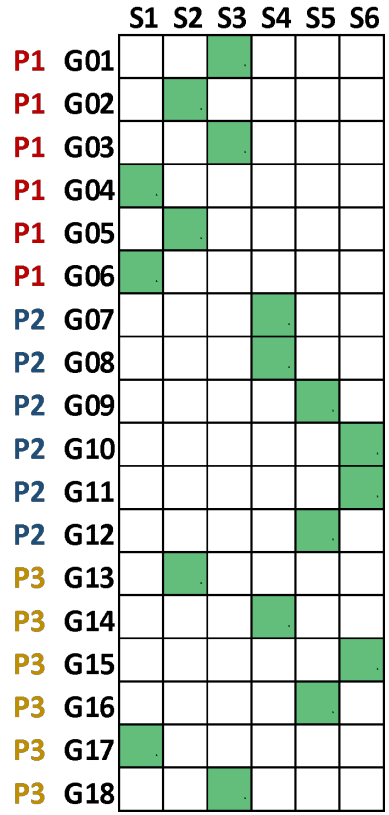
- Due to the sparsity of the data, all samples are at equal distance of each other



Value of network-guided analysis

Data sparsity:

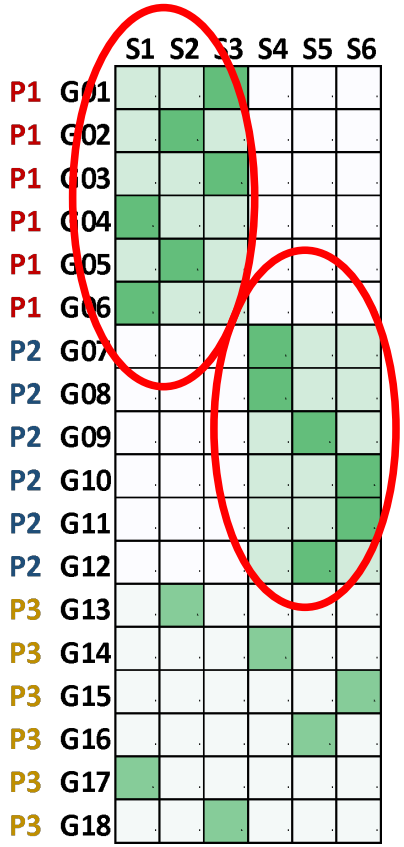
- Due to the sparsity of the data, all samples are at equal distance of each other
- Pathway information clarifies the similarity among some samples



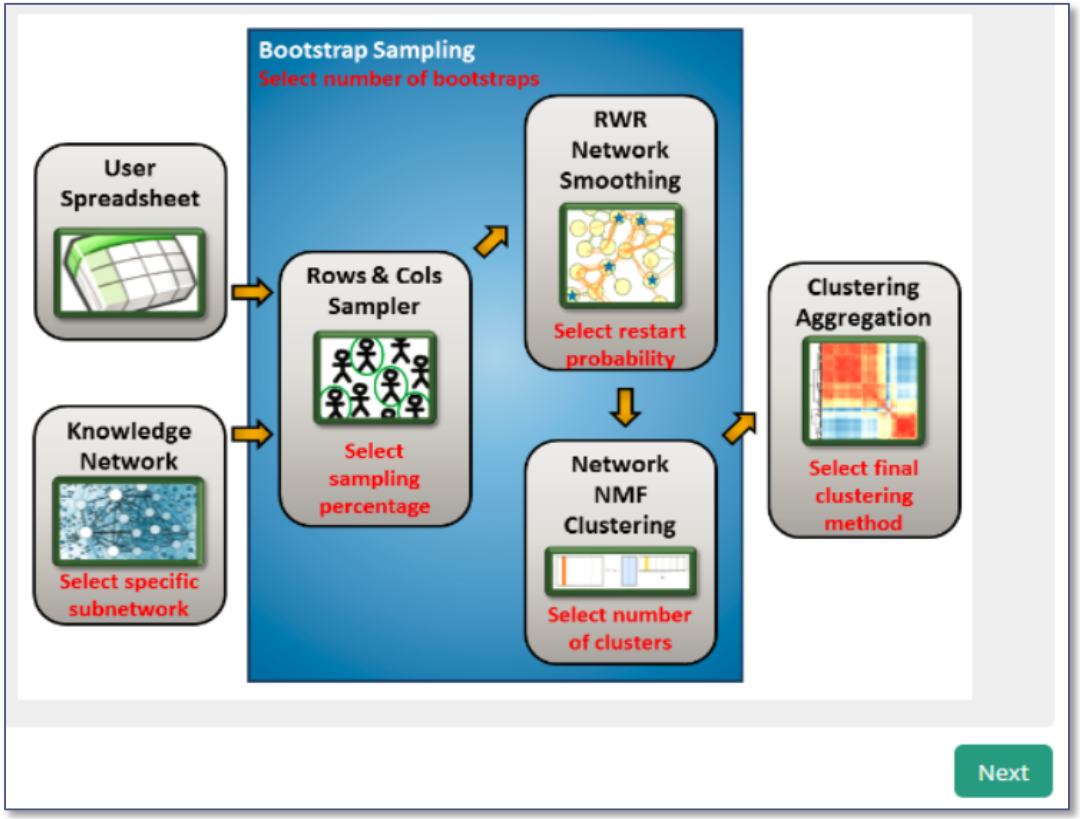
Value of network-guided analysis

Data sparsity:

- Due to the sparsity of the data, all samples are at equal distance of each other
- Pathway information clarifies the similarity among some samples
- Conventional clustering methods can identify clusters based on network-smoothed features



NBS (Algorithm Overview)



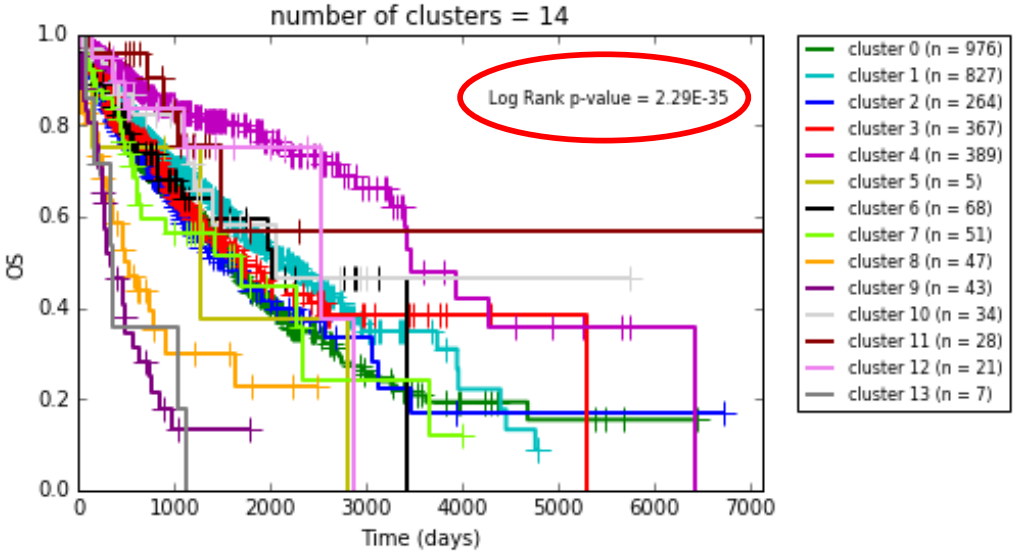
- Employs network smoothing to mitigate sparsity by transforming the binary gene-level somatic mutation vectors of patients into a continuous gene importance vector that captures the proximity of each gene in the network to all of the genes with somatic mutations in the patient sample
- Bootstrap sampling enables robust clustering

Stratification of TCGA Patients Mutation Data

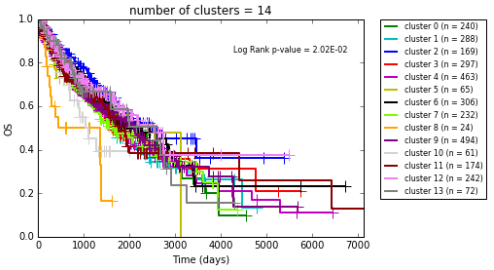
Knowledge-guided analysis of 'omics' data using the KnowEnG cloud platform

Charles Blatti III, Amin Emad, Matthew J. Berry, Lisa Gatzke, Milt Epstein, Daniel Lanier, Pramod Rizal, Jing Ge, Xiaoxia Liao, Omar Sobh, Mike Lambert, Corey S. Post, Jinfeng Xiao, Peter Groves, Aidan T. Epstein, Xi Chen, Subhashini Srinivasan, Erik Lehnert, Krishna R. Kalari, Liewei Wang, Richard M. Weinshilboum, Jun S. Song, C. Victor Jongeneel, Jiawei Han, Umberto Ravaioli, Nahil Sobh, Colleen B. Bushell, Saurabh Sinha

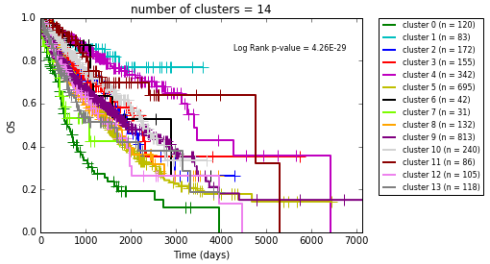
- 3276 tumor samples from TCGA from 12 cancer projects
  - Non-synonymous somatic mutation
- Network Based Stratification using the HumanNet Integrated Network
- Clusters significantly relate to survival outcome



- Much better than standard methods that do not incorporate prior knowledge



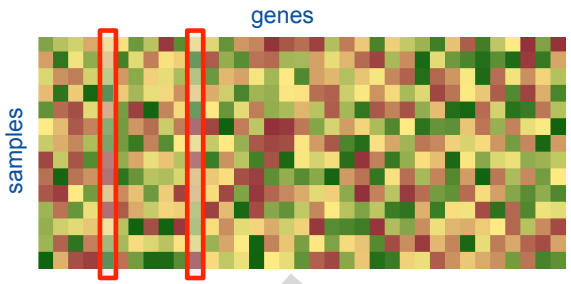
- In line with specialized method developed in TCGA paper that would be very difficult to reproduce



3) Reconstruction of Biological Networks

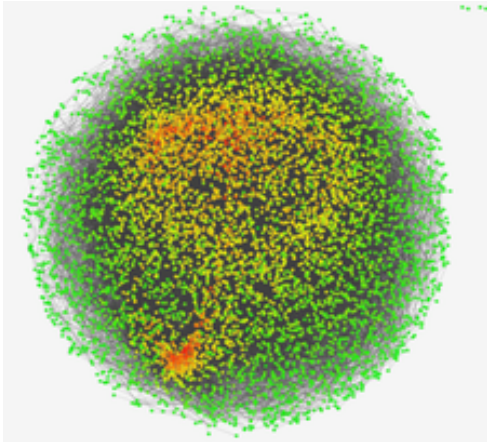
Gene Co-expression Networks

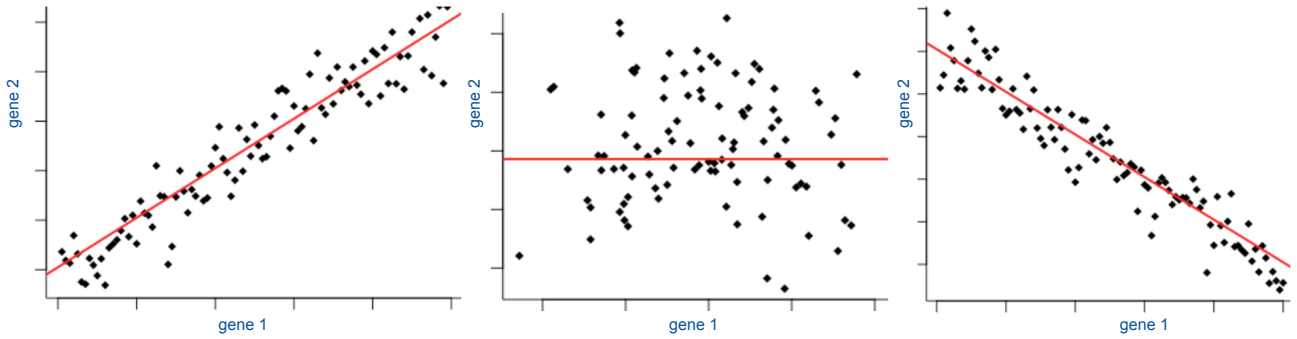
- Nodes represent genes
- An edge exists between two genes that are highly co-expressed across different samples



Calculate pair-wise correlations

Filter out small correlations





Gene Co-expression Networks

- Such networks provide a global view of co-expression patterns

BMC Bioinformatics. 2008; 9: 559. PMCID: PMC2631488
Published online 2008 Dec 29. doi: 10.1186/1471-2105-9-559

WGCNA: an R package for weighted correlation network analysis

Reviewed by Peter Langfelder[1] and Steve Horvath[2]

BMC Bioinformatics. 2006; 7(Suppl 1): S7. PMCID: PMC1810318
Published online 2006 Mar 20. doi: 10.1186/1471-2105-7-S1-S7

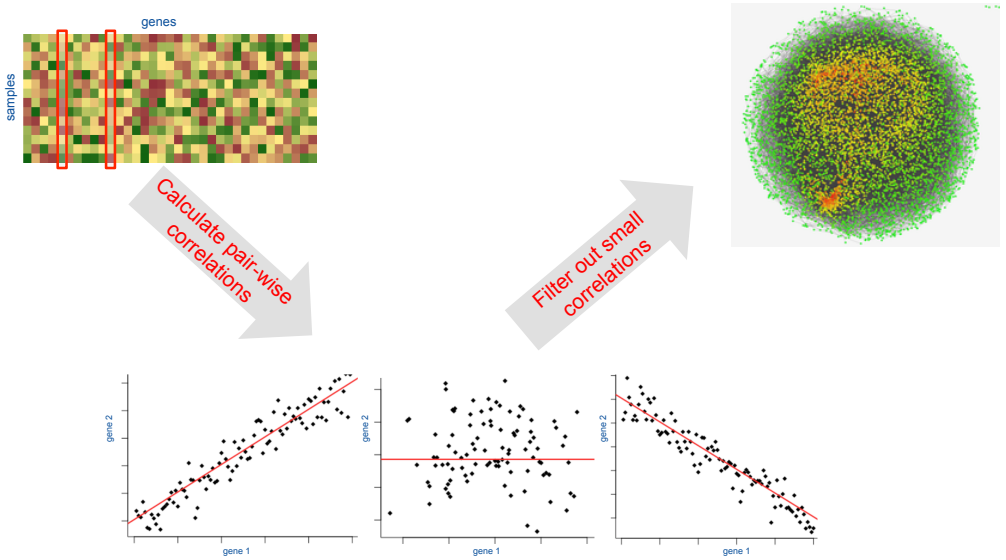
ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context

Adam A Margolin,[1,2] Ilya Nemenman,[2] Katia Basso,[3] Chris Wiggins,[2,4] Gustavo Stolovitzky,[5] Riccardo Dalla Favera,[3] and Andrea Califano[1,2]

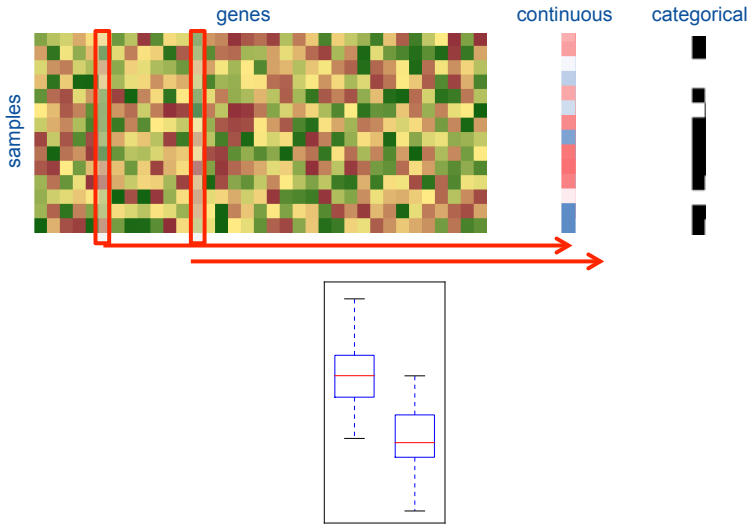
- But do not provide information on how these networks relate to the variation in a phenotypic outcome

Gene Co-expression Networks

How can we relate these networks to the phenotypic variation?



gene-gene correlation



gene-phenotype association

Gene Co-expression Networks

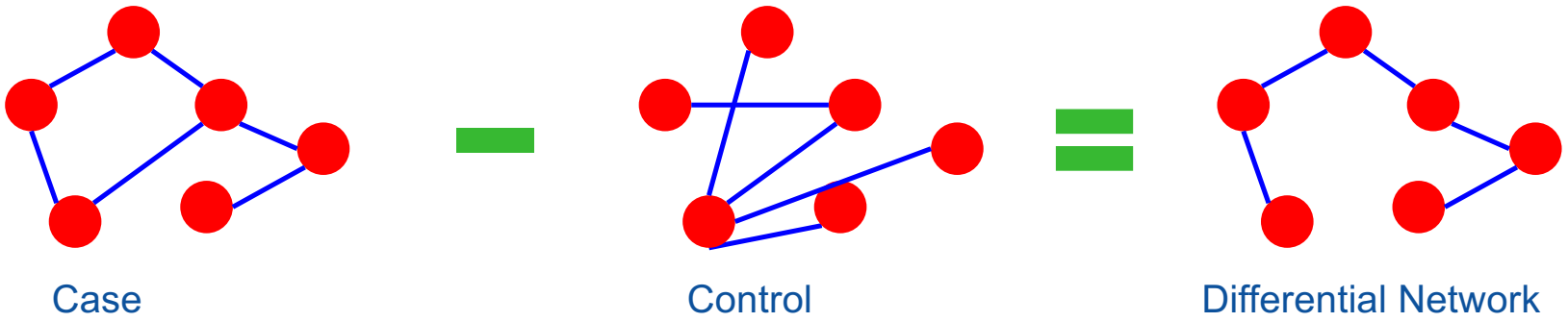
Approach 1: In reconstructing the network, we can limit our samples to one manifestation of the phenotypic outcome

- For example, build a Basal-like co-expression network by looking at the gene correlations across Basal breast cancer samples

- Issues:
  1. Only works if we have categorical phenotype
  2. Does not relate the network to the variation in the phenotypic outcome

Gene Co-expression Networks

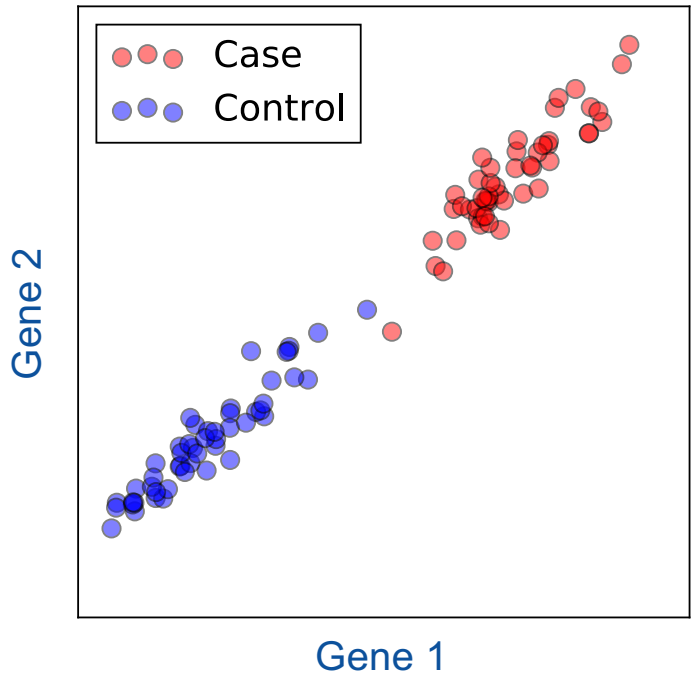
Approach 2: If the phenotype is binary, reconstruct two networks (one for each manifestation of the phenotype) and compare the two to build a differential network



- Shows changes in the co-expression pattern

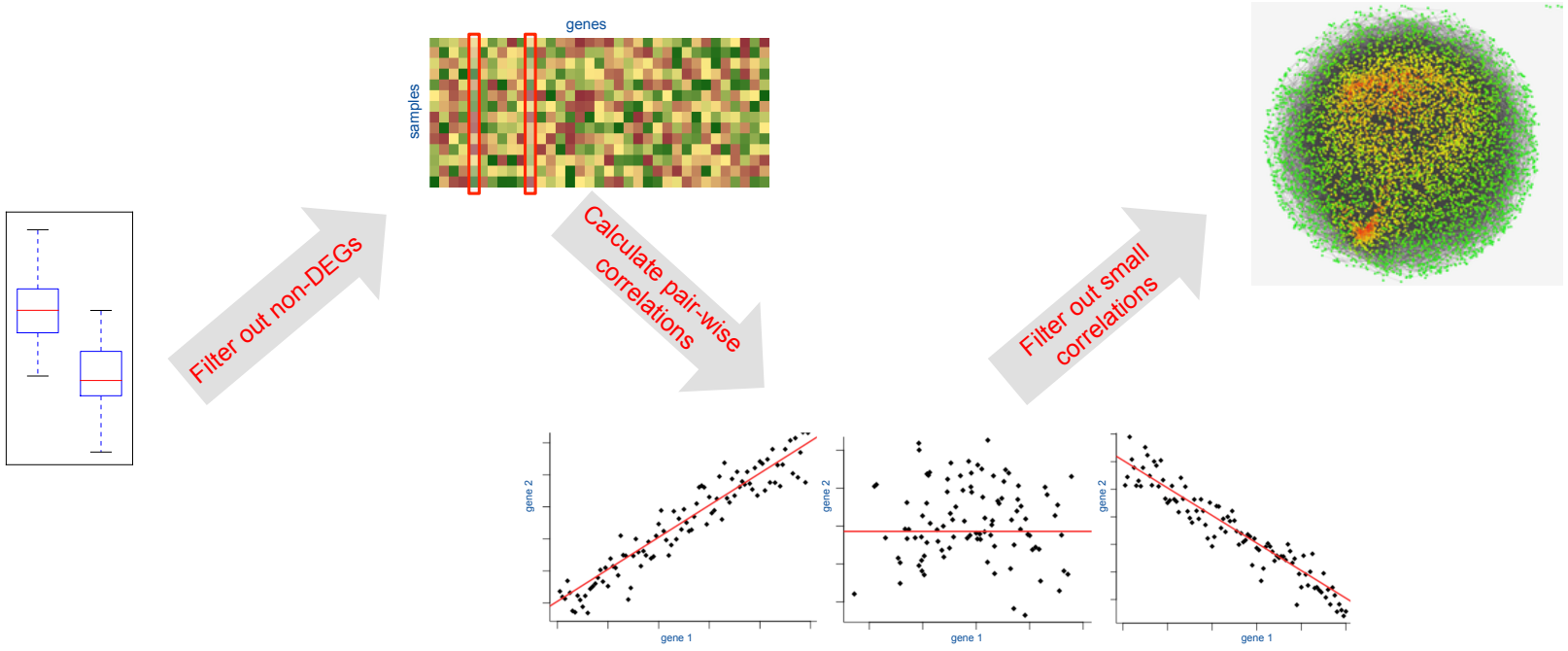
Gene Co-expression Networks

- Issues:
  1. Becomes very cumbersome if phenotype is not binary
  2. Does not work for continuous-valued phenotypes
  3. By dividing the samples into two groups, we will have less statistical power in identifying co-expression patterns
  4. Fails in a case shown below



Gene Co-expression Networks

Approach 3: First, find genes associated with the phenotype and then reconstruct a context-specific network only using those genes

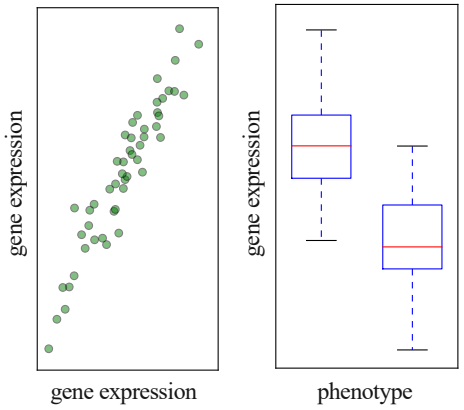


- Issue:

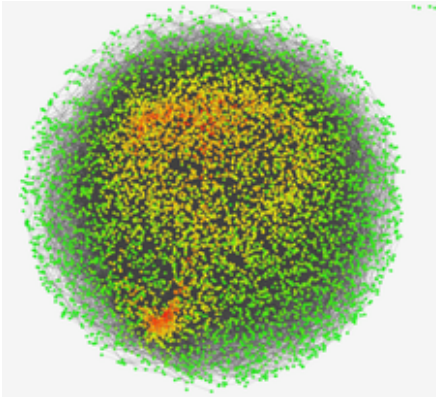
  Ignores the strength (p-value) of gene-phenotype association

Gene Co-expression Networks

Approach 4: Calculate p-values of gene-gene correlation and gene-phenotype associations separately and combine together using Fisher's method or Stouffer's method → Simplified-InPheRNo



Combine the two p-values



- Specifically useful to identify transcription factor-gene-phenotype associations

Inference of phenotype-relevant transcriptional regulatory networks elucidates cancer type-specific regulatory mechanisms in a pan-cancer study

Amin Emad, Saurabh Sinha

Questions?

Time for Course Photo!

Regression algorithms

- Lasso: learns a linear model from the training data using only a few features (sparse linear model)

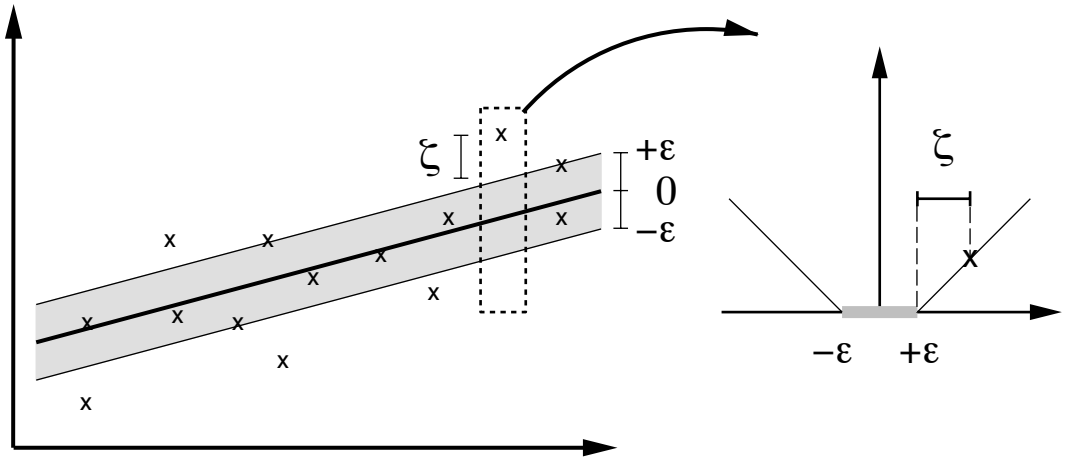
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right)$$

- Elastic Net: learns a linear model from the training data by linearly combining ridge and Lasso regression regularization terms (a generalization of both Lasso and ridge regression)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_2 ||\boldsymbol{\beta}||_2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right)$$

Regression algorithms

- Kernel-SVR:
  - Linear SVR learns a linear model such that it has at most ε-deviation from the response values and is as flat as possible



(Smola and Schölkopf, 1998)

  - Kernel-SVR generalizes the idea to nonlinear models by mapping the features to a high-dimensional kernel space