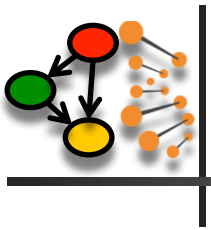


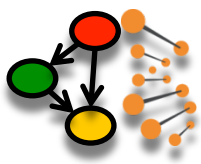
Cure cancer from your laptop

Su-In Lee

Computer Science & Engineering, Genome Sciences,

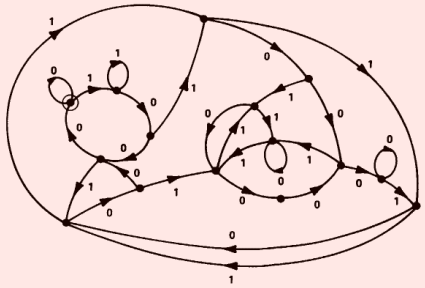
University of Washington

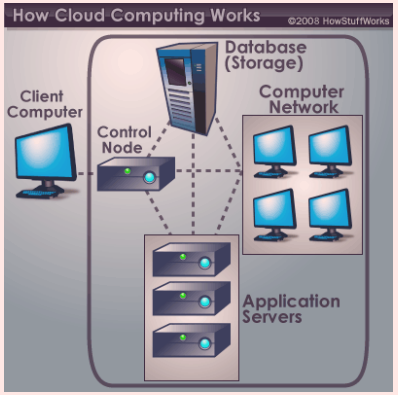
UW Direct Admits Seminar



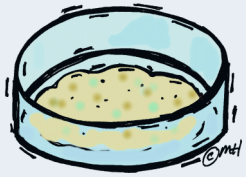
CS and biology seem to have very different goals and methodologies

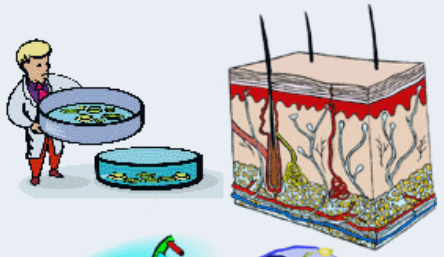
Computational



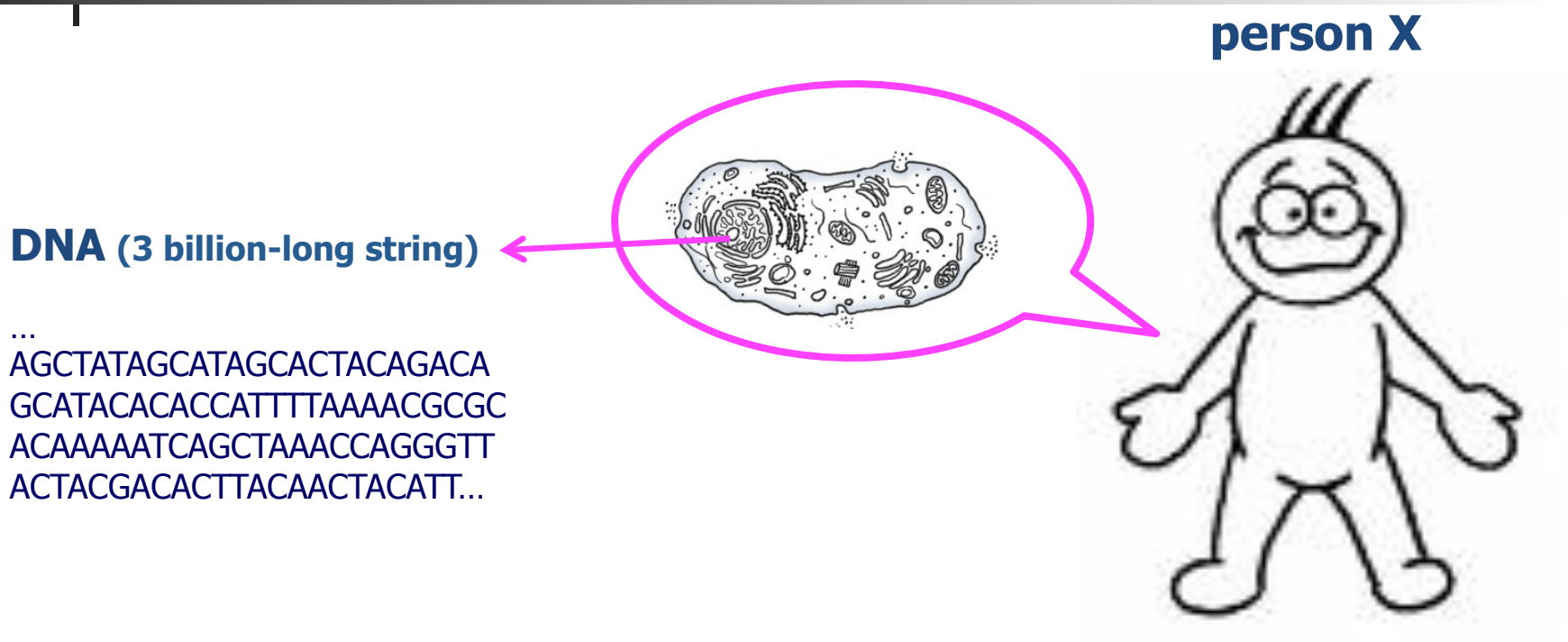


Biology



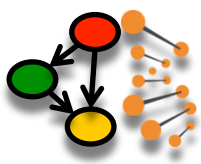


Biology is an information science



- DNA acts as the "brain" of the cell, telling the cell how to properly grow and work.

Each individual has a slightly different version of DNA sequence



DNA tells the cell how to perform various tasks in a cell

Gene (~20,000 in human) **Gene regulation♪**

AGATATGTGGATTGTTAGGATTTATGCGCGTCAGTGACTACGCATGTTACGCACCTACGACTAGGTAATGATTGATC

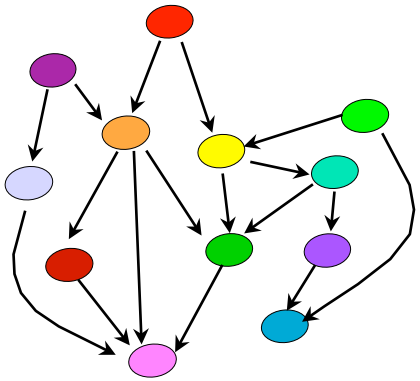
DNA

RNA AUGUGGAUUGUU AUGCGCGUC AUGAUUGAU

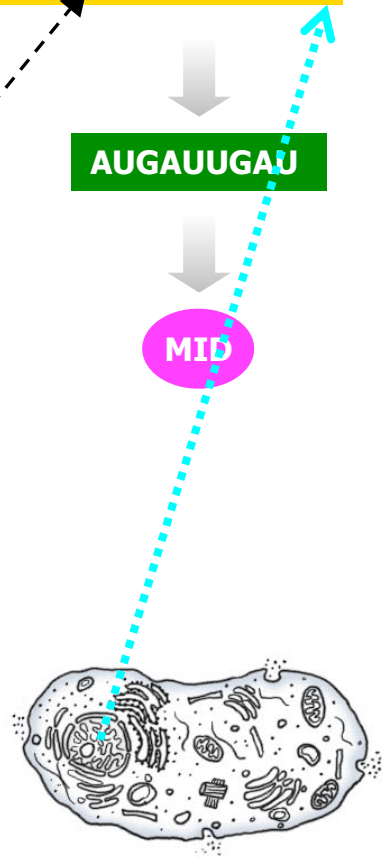
Protein MWIV MRV MRV MID

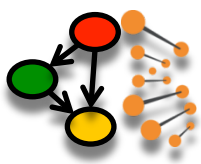
Gene expression

RNA degradation



Gene interaction map ("social network" of genes)

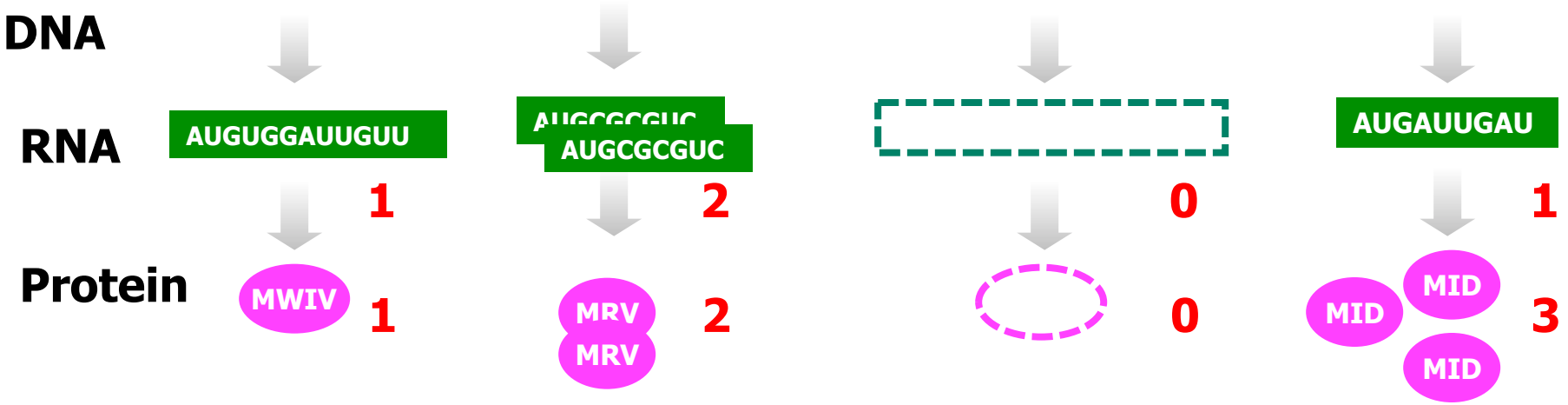




A cell's biological state can be described by millions of numbers

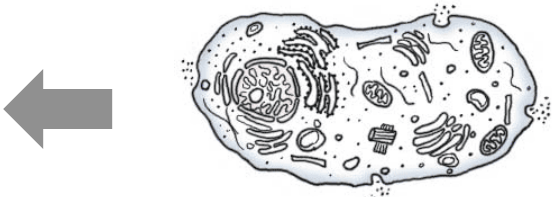
Gene (~20,000 in human)

AGATATGTGGATTGTTAGGATTTATGCGCGTCAGTGACTACGCATGTTACGCACCTACGACTAGGTAATGATTGATC



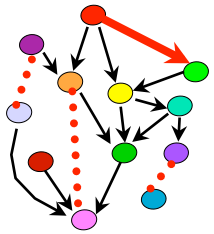
- **Biological information in the 21th century**
  - DNA sequence: **>1M** letters known to differ among individuals.
  - RNA expression levels of **20K** genes
  - Protein levels of **20K** genes

  :
- How many numbers? Definitely >1M !



Modern biology is about mining very large, complex data

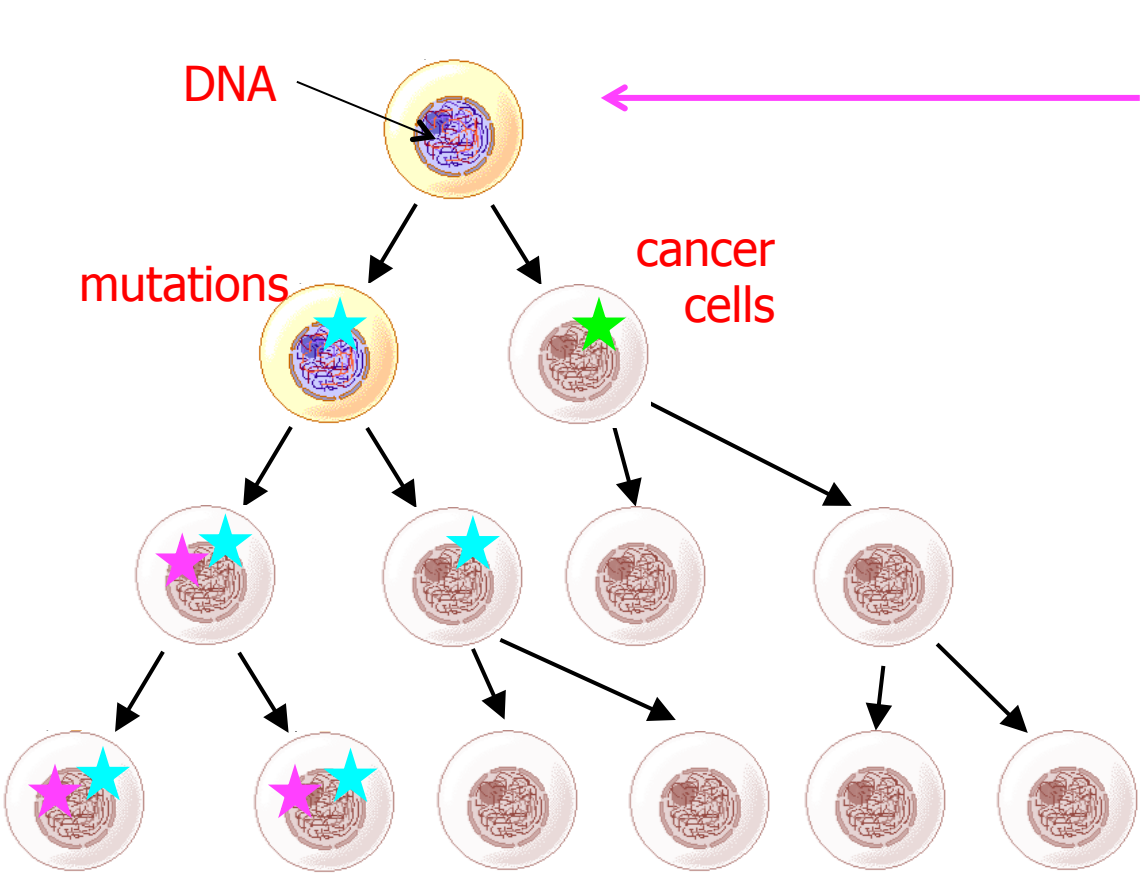
- Interesting questions arise…
  - Which parts in the DNA sequence determines susceptibility to Alzheimer's disease?
  - How the social network of genes are different between cancer and normal cells?
  - How the DNA sequences are different between different species?
- **CS and Statistics play a key role!**

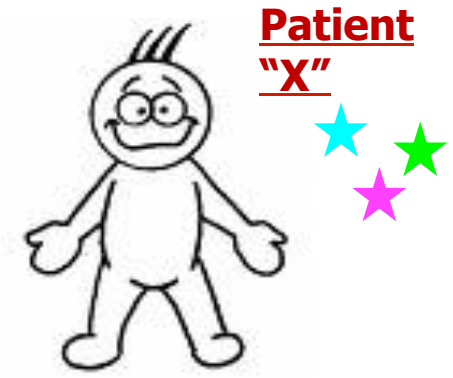


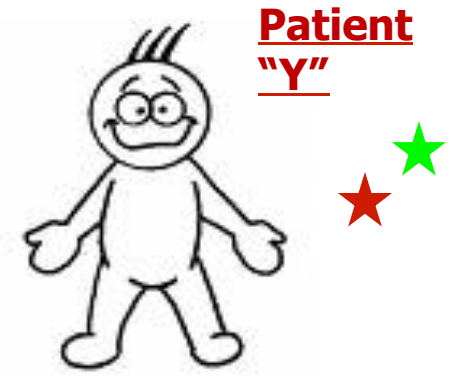
HOW CAN COMPUTER SCIENTISTS CURE CANCER?

Cancer is a disease of the genome

- Normal cells grow, divide and die in an orderly fasion.



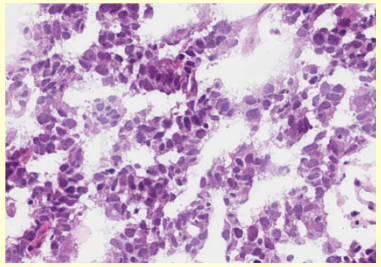




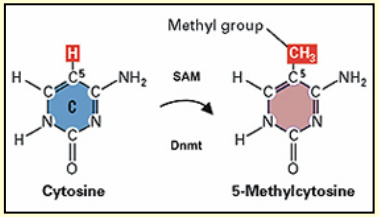
Using molecular profile to make treatment plan for patient X

Molecular snapshot

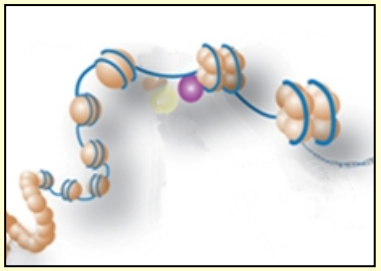
histopathology



DNA methylation



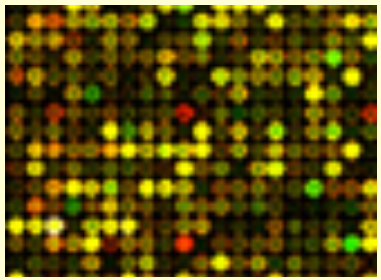
histone modification



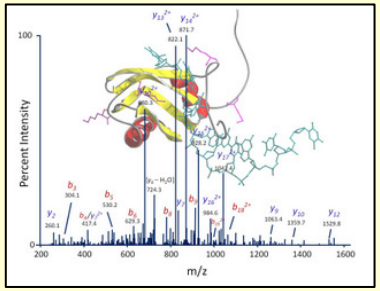
sequencing

```
...
ACGTAGCTAGCTAGCTAGCTGAT
GCTAGCTACGTGCTACATCTATC
TATCTATCTCCTCTCATCTATCT
ATCTATCATCTATCTATCTATCA
TTTCTATCTATCTTCTATCTTAC
ACCCCCAGGGCACCCCCCAAATC
TTCTATCTATCTTCTATCTAC...
```

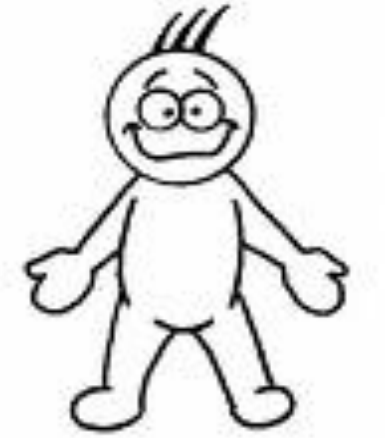
mRNA levels



protein levels



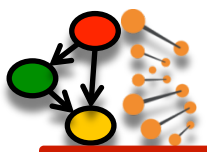
Patient X



Treatment plan

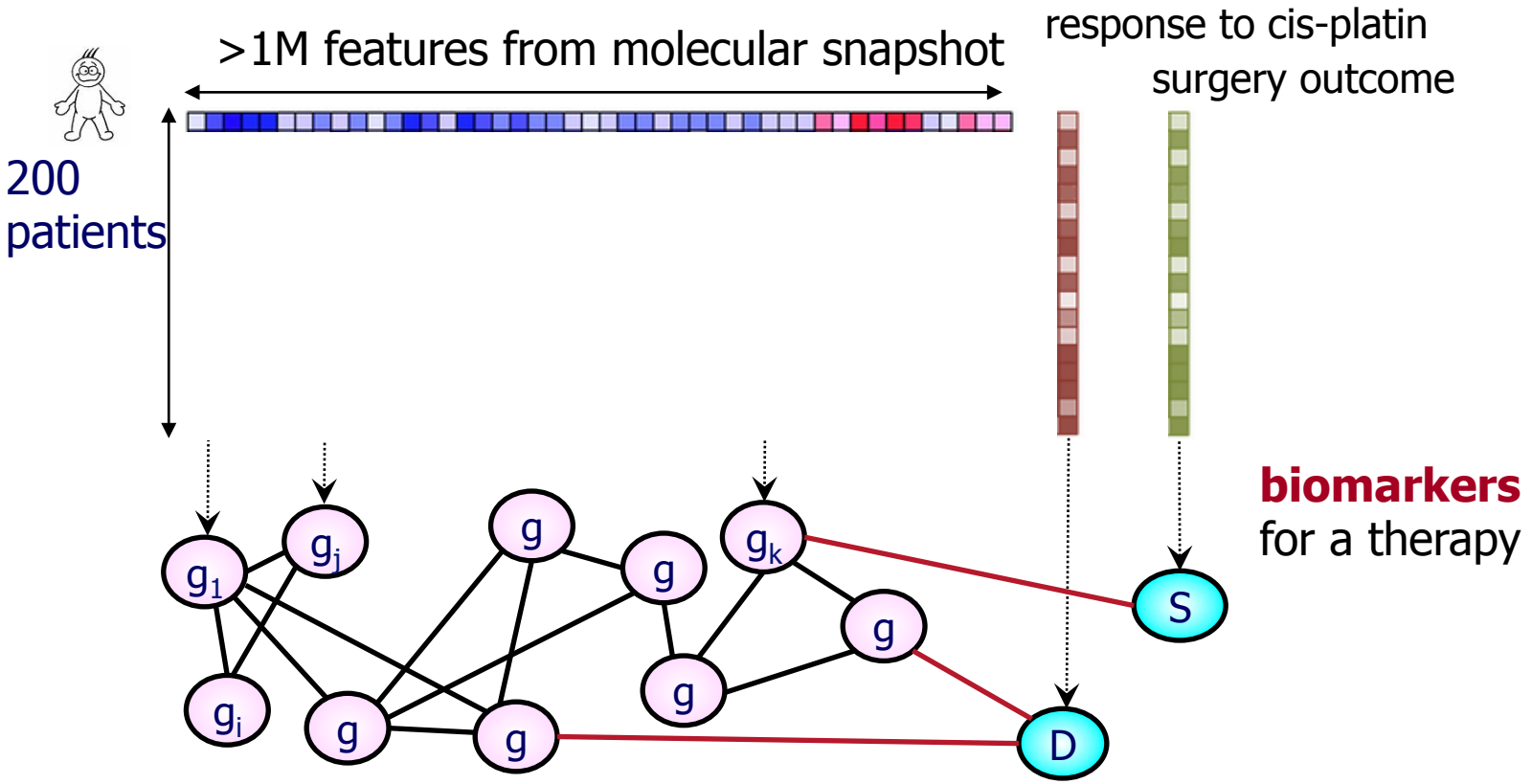
Would surgical resection be successful?
Need a neoadjuvant therapy?

Which of ~200 chemo drugs?
5-Iodotubercidin, Arsenic trioxide, Daunorubicin, Tipifarnib, ...
:

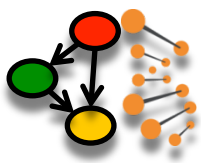


Learning relationships among variables from data

Learning relationships in high-dimensional data (1M » 200) is a very challenging statistical problem!

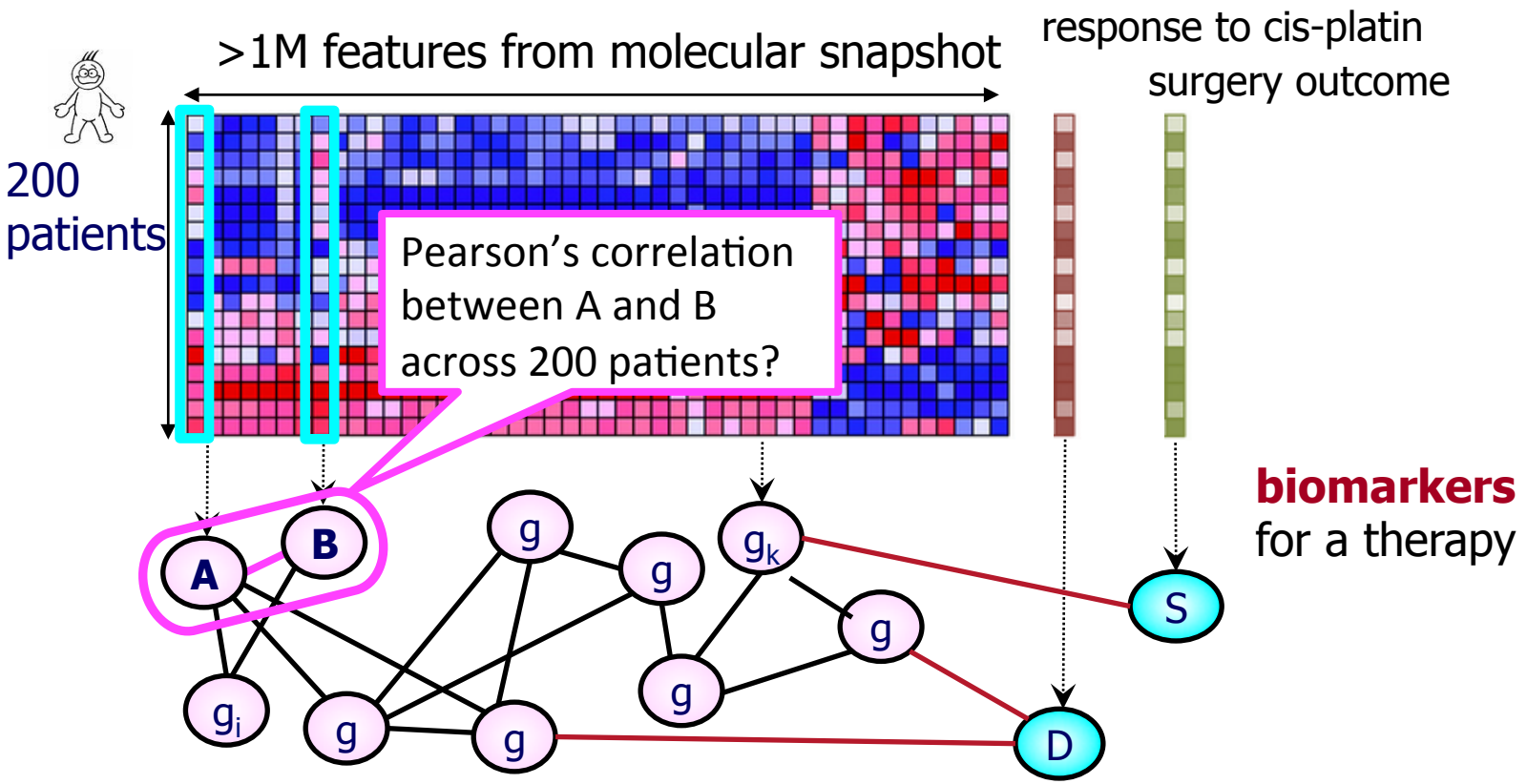


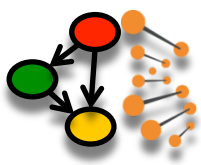
genes' social network in cancer



Really? How about we just compute pairwise correlations?

- **Pairwise correlations are not enough to reveal relationships among variables**





Coffeeshop Example

Like Coffeeshops?

Like Coffee?

Like Tea?

	Like Coffeeshops?	Like Coffee?	Like Tea?
Person 1	Y	Y	N
Person 2	N	N	N
.			
.			
.			
.			
.			
.			
.			
.			
Person 100	Y	N	Y

From Prof. Daniela Witten's slide

From Prof. Daniela Witten's slide

True network

From Prof. Daniela Witten's slide

Correlation Matrix

$$\Sigma = \begin{bmatrix} 1 & 0.58 & 0.58 \\ 0.58 & 1 & 0.5 \\ 0.58 & 0.5 & 1 \end{bmatrix}$$





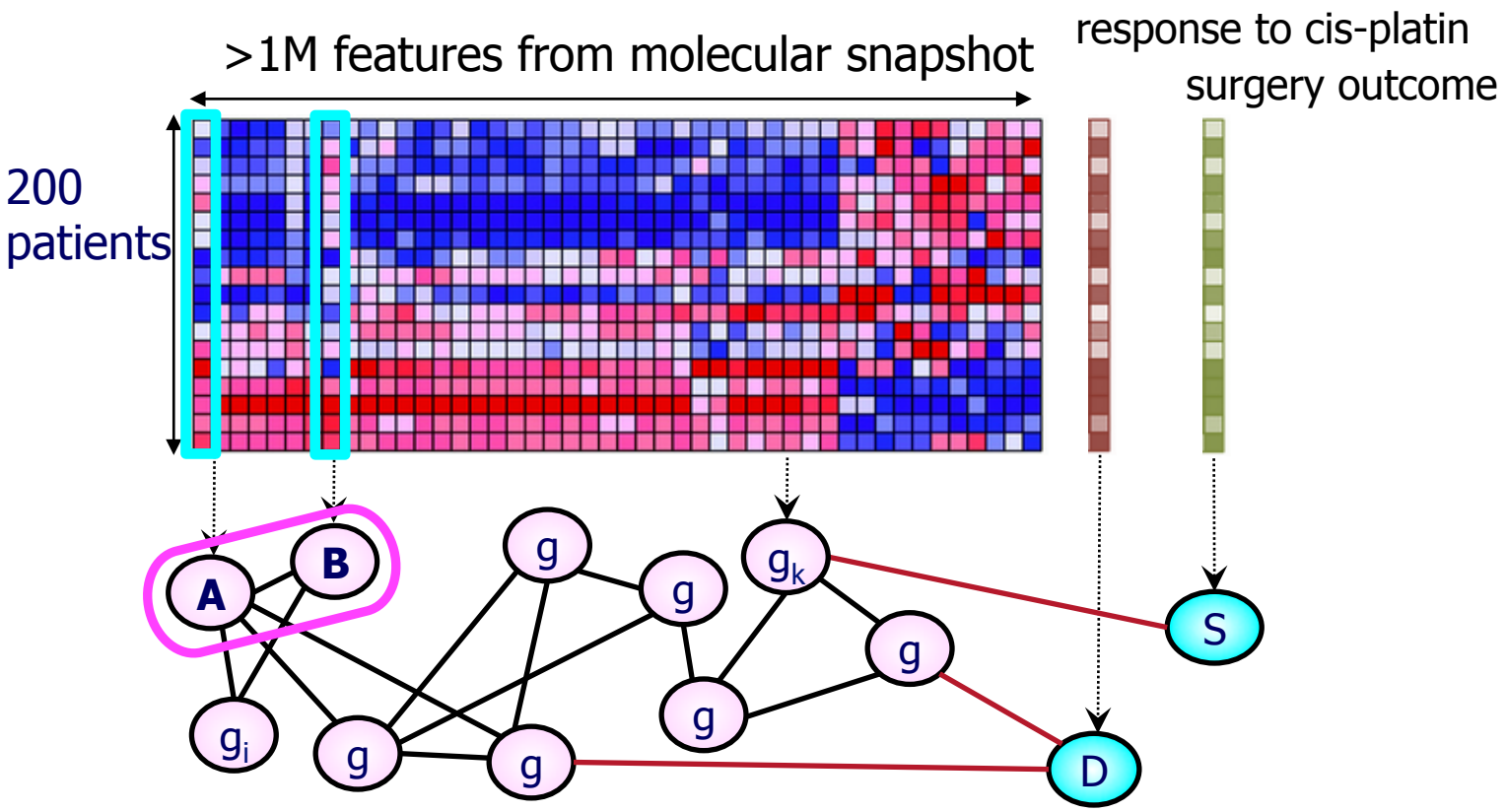

Variables that are indirectly connected can have high correlation

	Like Coffeeshops?	Like Coffee?	Like Tea?
Person 1	Y	Y	N
Person 2	N	N	N
.			
.			
.			
.			
.			
.			
.			
Person 100	Y	N	Y

From Prof. Daniela Witten's slide

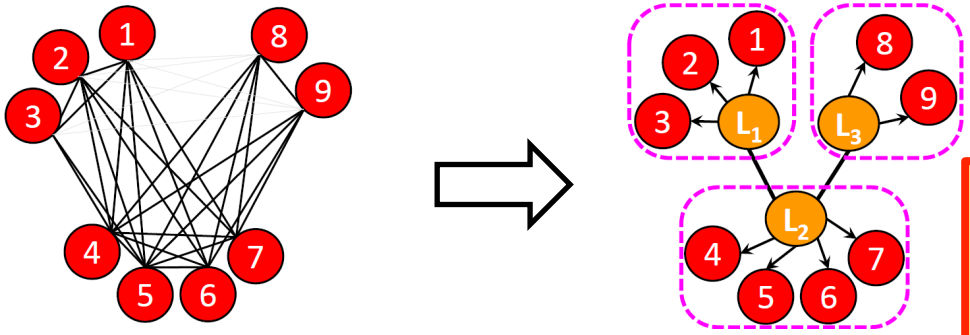
Relationship between A and B can be inferred by considering all variables

- **How many possible network structures?**



Reducing dimensionality in high-dimensional network inference

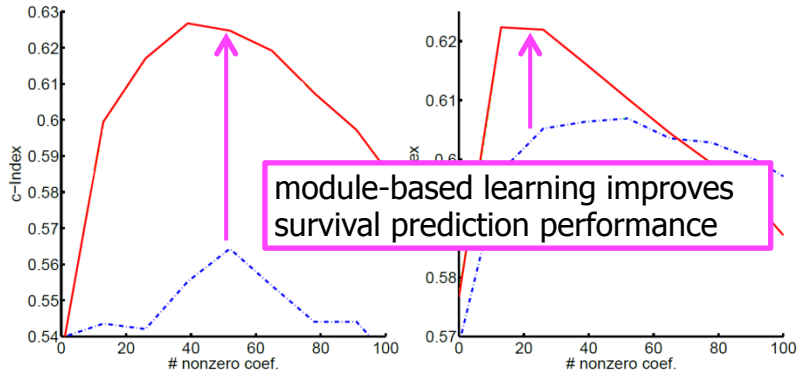
- Module graphical lasso*



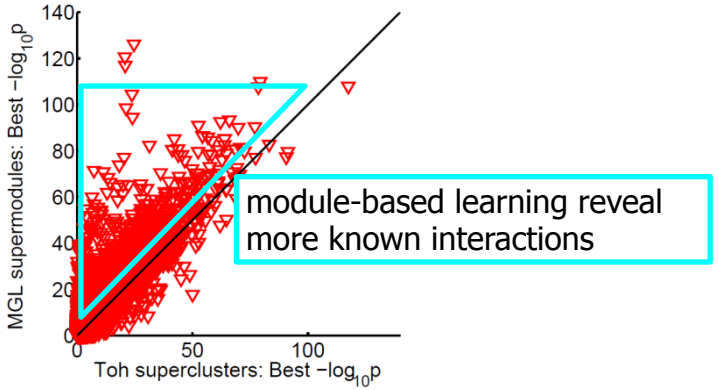
$$P(\mathbf{X}, \mathbf{L}, \mathbf{Z}, \Sigma_{\mathbf{L}})$$
$$= \prod_{i=1}^{p} P(X_i|L_{Z_i})P(\mathbf{L}|\Sigma_{\mathbf{L}})P(\Sigma_{\mathbf{L}}{}^{-1})P(\mathbf{Z})$$
$$= \prod^{p} \frac{1}{\sqrt{\quad}} \exp\left\{-\frac{(X_i - L_{Z_i})^2}{2\sigma^2}\right\}$$

Dimensionality reduction is important

survival time prediction

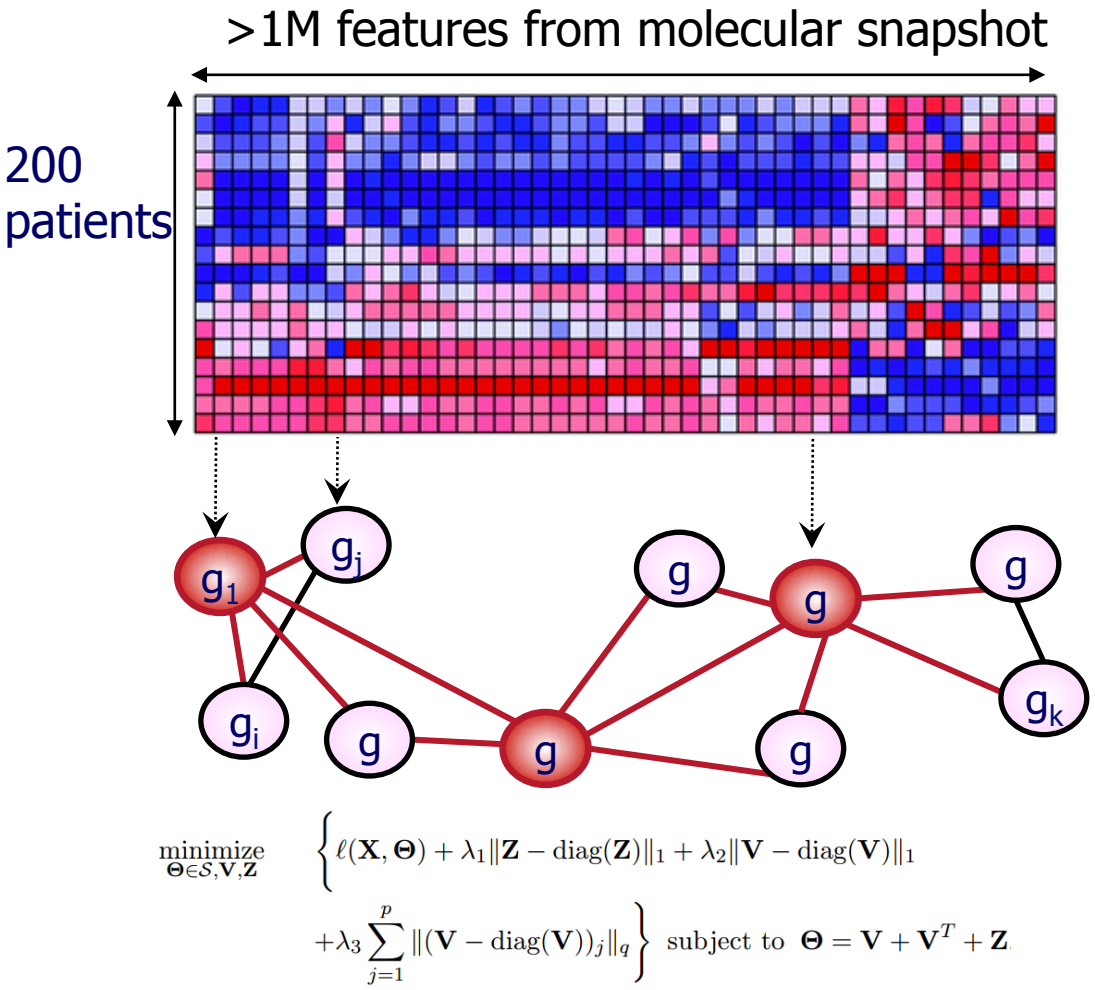


revealing known relationships



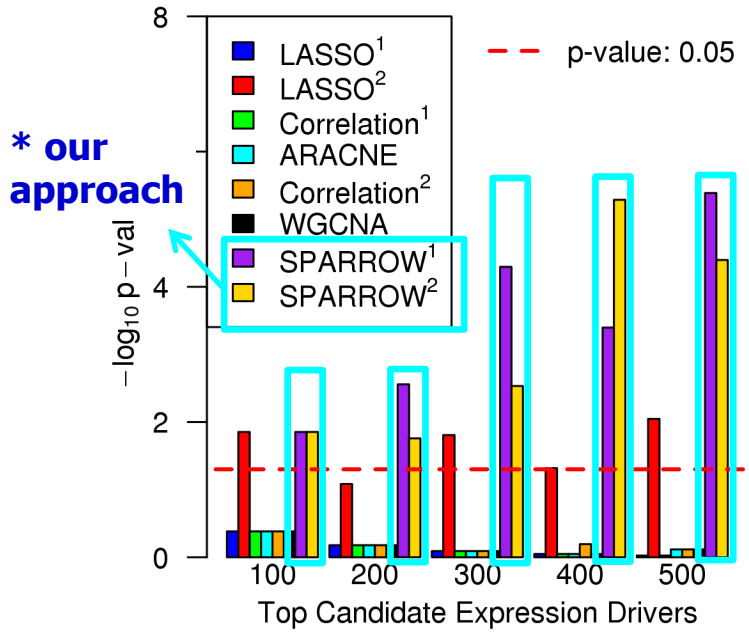
*Celik et al. Efficient Dimensionality Reduction for High-Dimensional Network Estimation (*ICML*'14, *NIPS MLCB*'13)

Identifying hub nodes in high-dimensional data



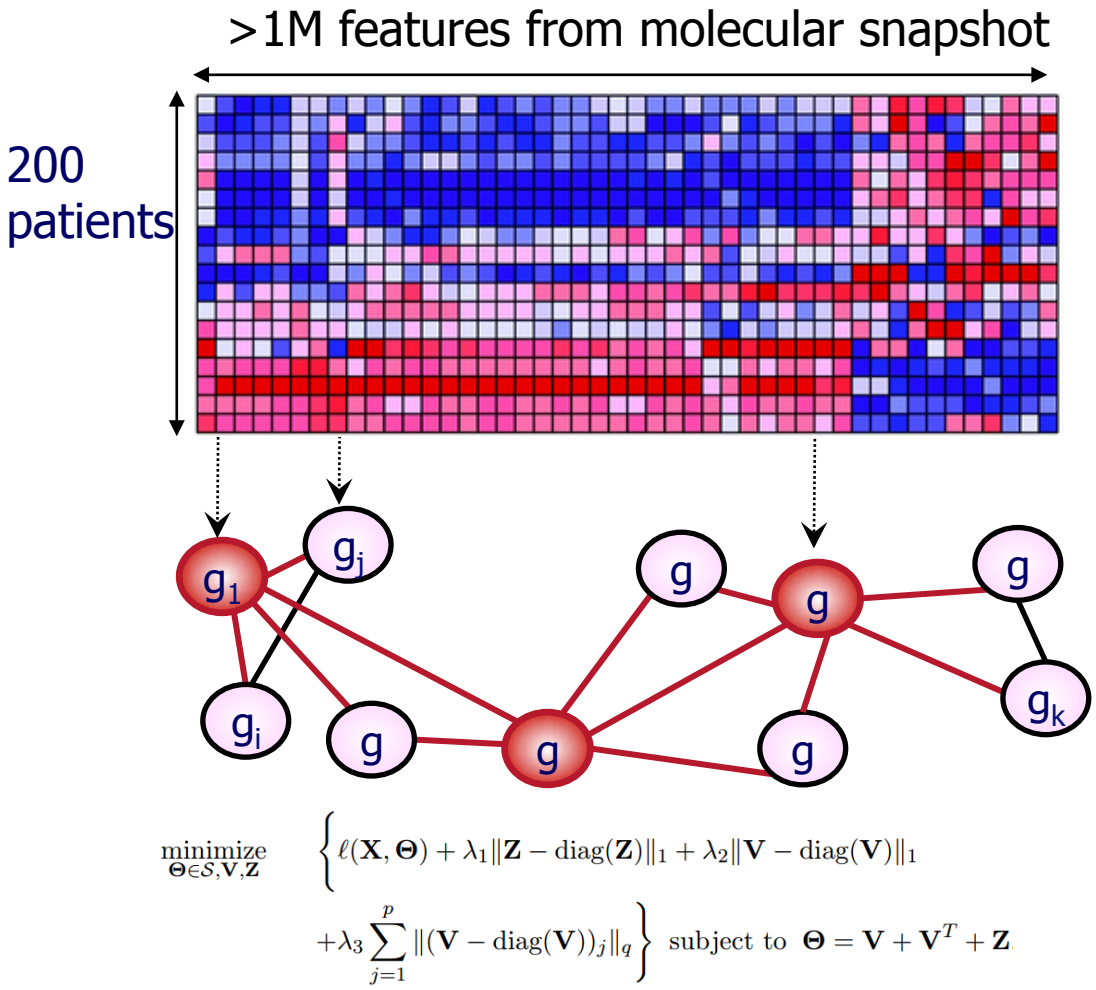
$$\underset{\Theta \in \mathcal{S}, \mathbf{V}, \mathbf{Z}}{\text{minimize}} \left\{ \ell(\mathbf{X}, \mathbf{\Theta}) + \lambda_1 \|\mathbf{Z} - \text{diag}(\mathbf{Z})\|_1 + \lambda_2 \|\mathbf{V} - \text{diag}(\mathbf{V})\|_1 + \lambda_3 \sum_{j=1}^{p} \|(\mathbf{V} - \text{diag}(\mathbf{V}))_j\|_q \right\} \text{ subject to } \mathbf{\Theta} = \mathbf{V} + \mathbf{V}^T + \mathbf{Z}$$

how well each method capture known cancer driving genes



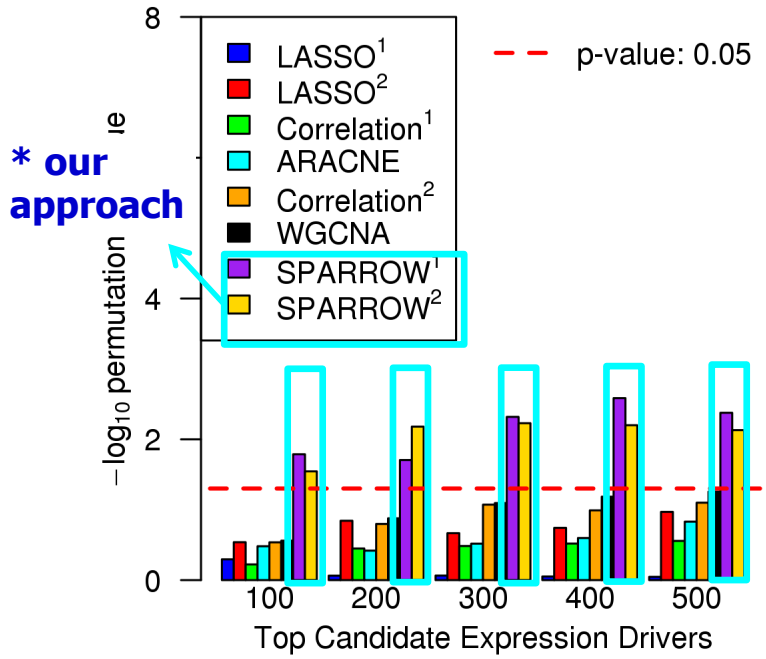
*Logsdon et al, Sparse expression bases in cancer reveal tumor drivers, Under review

Identifying hub nodes in high-dimensional data



$$\underset{\Theta \in \mathcal{S}, \mathbf{V}, \mathbf{Z}}{\text{minimize}} \left\{ \ell(\mathbf{X}, \mathbf{\Theta}) + \lambda_1 \|\mathbf{Z} - \text{diag}(\mathbf{Z})\|_1 + \lambda_2 \|\mathbf{V} - \text{diag}(\mathbf{V})\|_1 + \lambda_3 \sum_{j=1}^{p} \|(\mathbf{V} - \text{diag}(\mathbf{V}))_j\|_q \right\} \text{ subject to } \mathbf{\Theta} = \mathbf{V} + \mathbf{V}^T + \mathbf{Z}$$

how well each method predict survival time



*Logsdon et al, Sparse expression bases in cancer reveal tumor drivers, Under review

Hub nodes are effective biomarkers for chemo drugs

- Hubs in the gene social network are effective biomarkers for 160 chemo drugs*

Quellos High Throughput Screening Core

UW Center for Cancer Innovation

Dr. Pamela Becker
(oncologist)

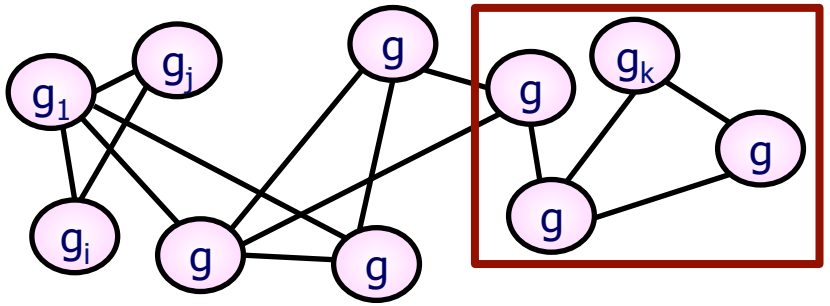
Dr. Tony Blau
(hematologist)

* Lee et al, Big data approach to identify molecular basis for drug sensitivity phenotypes in AML, Accepted for oral presentation (10%) at the *American Society of Hematology*

Patterns conserved across vastly different cancer types

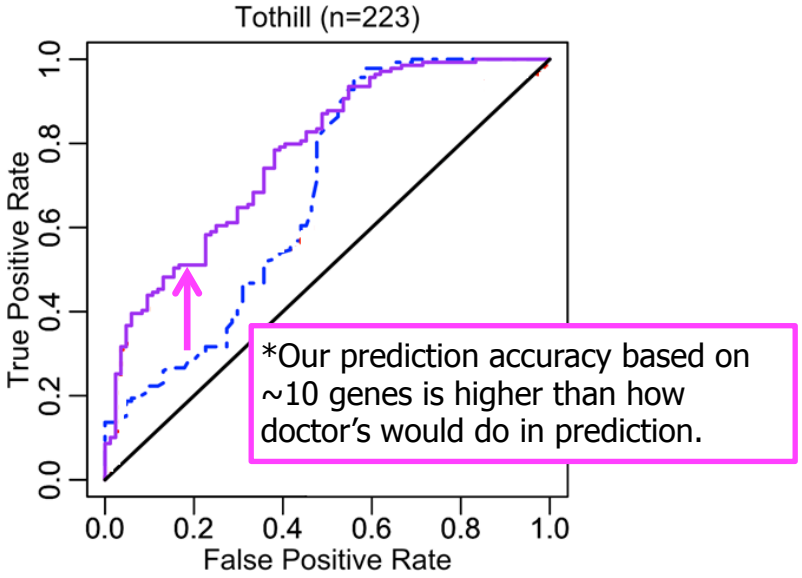
- Ovarian cancer resectability
  - Strongly associated with survival
  - Many ovarian cancers are difficult to remove from surrounding tissue
    - Molecular basis?

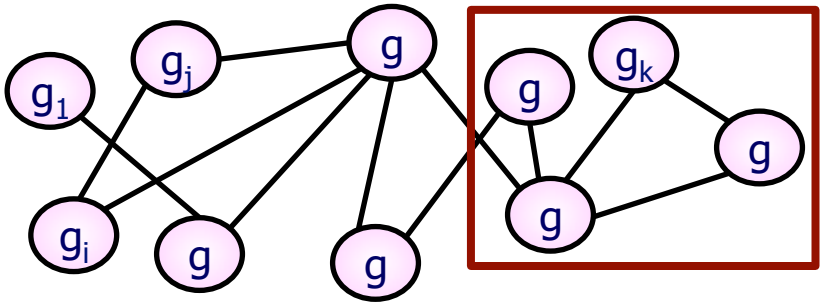
ovarian cancer



conserved gene social network

resectability prediction



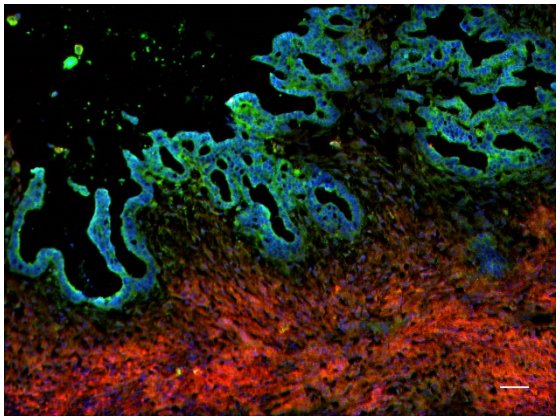


blood cancer

Revealing molecular basis for ovarian tumor resectability

- New findings
  - We identified genes associated with tumors that are harder to remove from the surrounding tissue
  - These genes can be therapeutic targets for neoadjuvant therapy (chemo before surgery)

Experimental validation
IHC staining



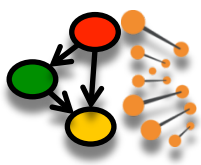
Dr. Charles Drescher
(surgeon)

Dr. Mara Rendi
(pathologist)

Cancer stem cell lab in UW GS

Stephanie Battle

David Hawkins



Conclusion

- Effective dimensionality reduction is the key to success in identifying relationships in high-dimensional data from cancer

- These are are challenging ML problems

Acknowledgement

Computational biologists
- Benjamin A. Logsdon (former postdoc), Sage Bionetworks
- Safiye Celik (3rd year CSE PhD student)

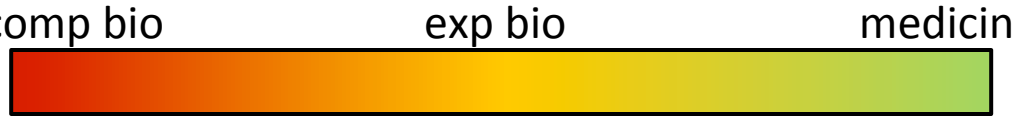
Clinicians
- Surgeon: Charles Descher (UW Medicine; FHCRC)
- Pathologist: Mara Rendi (UW Medicine)
- Leukemia oncologist: Pamela Becker (UW Medicine)

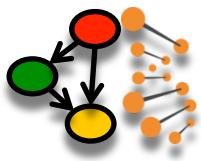
Systems biologists
- Hamid Bolouri (Fred Hutchinson Cancer Research Center)
- Muneesh Tewari (Univ of Michigan)
- Andrew Gentles (Stanford)

Hematologists
- Christopher Miller (UW Medicine)
- C. Anthony Blau (UW Medicine)

Cancer biologists in Hawkins Lab
- David Hawkins (UW Genome Sciences, Medical Genetics)
- Stephanie Battle (UW GS)

NSF, NIH, Royalty Research Funds, Fred Hutch Discovery funds, eScience/ ITHS





What is the coolest thing a computer scientist can do?

- Curing cancer.
- Predicting your disease susceptibilities based on your detailed biological information (>1M numbers!).
- Use big data to predict respiratory failure of patients under anesthesia during surgery in real time to more effectively save their lives.

:

- Visit suinlee.cs.washington.edu/research for more!