Ranking Methods in Machine Learning

A Tutorial Introduction

Shivani Agarwal

Computer Science & Artificial Intelligence Laboratory
Massachusetts Institute of Technology

In God we trust.
All others bring data.

W. Edwards Deming

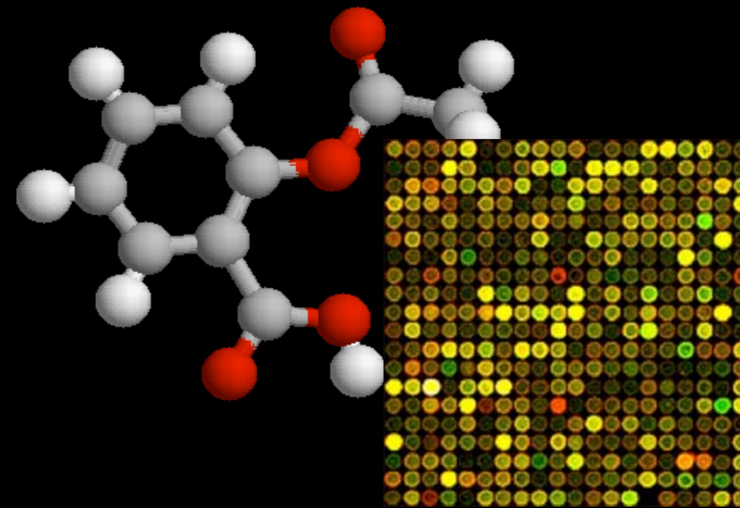
Machine Learning

Computer Science

Statistics

Machine Learning

Mathematical Optimization



Machine Learning

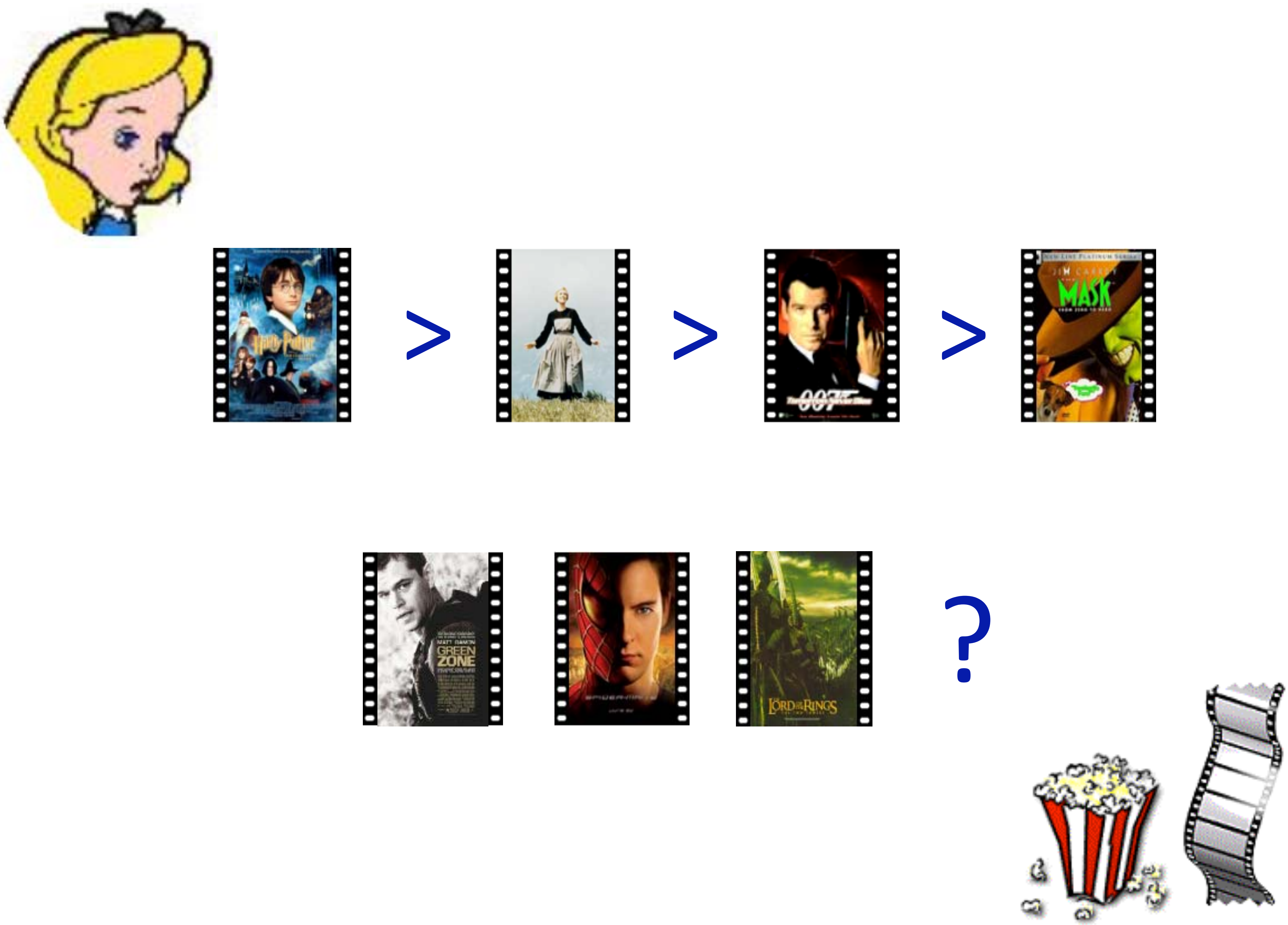
A > D > B > C

C > A > B > D

...

A > D > B > C

E F G ?



MASK
MATT DAMON
GREEN ZONE
>
>
>
?

Road Map

Theory

Algorithms

Applications

>

, 10

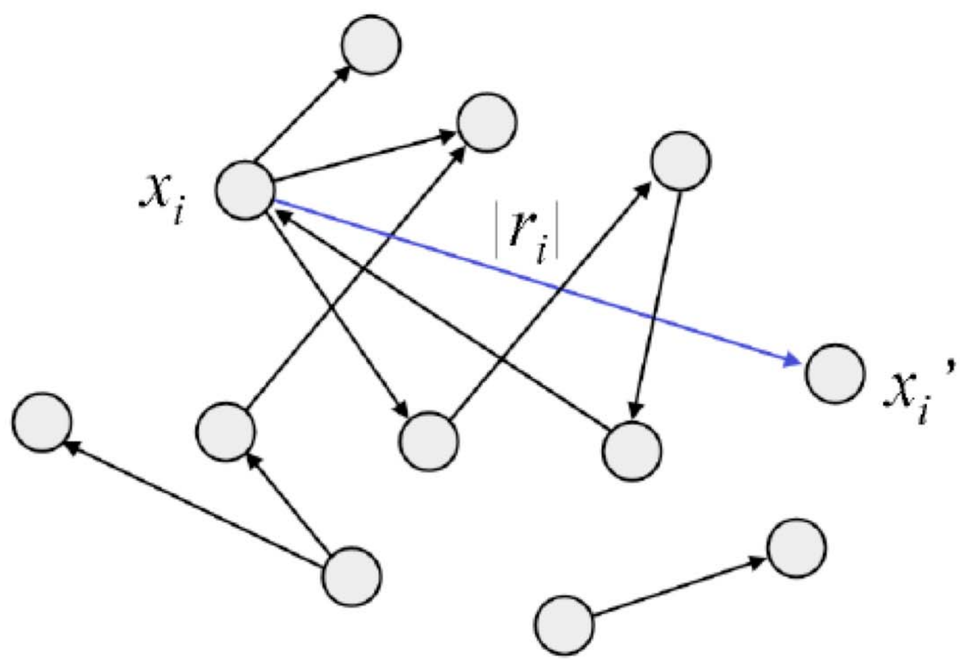
<

, 100

. . .

General Ranking Problem

- Instance space X
- Input: Training sample $S = ((x_1, x'_1, r_1), \ldots, (x_m, x'_m, r_m)) \in (X^2 \times \mathbb{R})^m$
- Output: Ranking function $f : X \to \mathbb{R}$



Likes   . . .

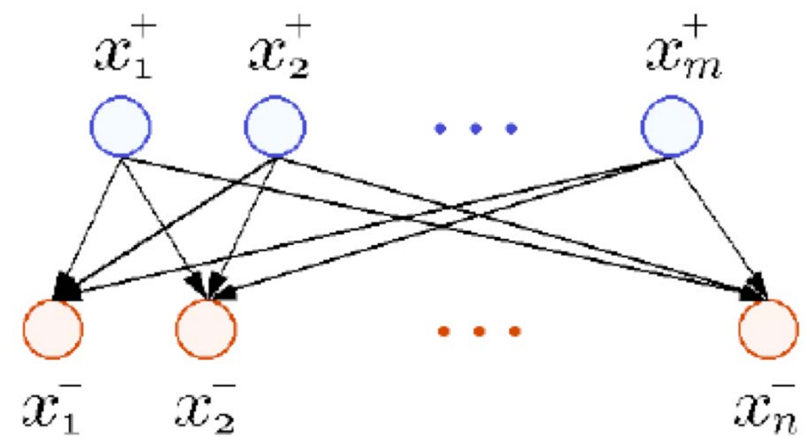
Dislikes    . . .

Bipartite Ranking Problem

- Instance space X
- Input: Training sample $S = (S_+, S_-)$:

$$S_+ = (x_1^+, \ldots, x_m^+) \in X^m \quad \text{(positive examples)}$$
$$S_- = (x_1^-, \ldots, x_n^-) \in X^n \quad \text{(negative examples)}$$

- Output: Ranking function $f : X \rightarrow \mathbb{R}$



Bipartite Ranking Problem

- Instance space X
- Input: Training sample $S = (S_+, S_-)$:

$$S_+ = (x_1^+, \ldots, x_m^+) \in X^m \quad \text{(positive examples)}$$
$$S_- = (x_1^-, \ldots, x_n^-) \in X^n \quad \text{(negative examples)}$$

- Output: Ranking function $f : X \rightarrow \mathbb{R}$

- Expected error: $\mathbf{er}(f) = \mathbf{P}_{(x,x')\sim\mathcal{D}_+\times\mathcal{D}_-}\left[f(x) < f(x')\right]$
- Empirical error: $\widehat{\mathbf{er}}_S(f) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{1}(f(x_i^+) < f(x_j^-))$

Generalization Bounds Review

Informally:

How does the empirical performance of a learned function generalize to its expected performance on future data?

Formally:

Let $f_S : X \rightarrow \mathbb{R}$ denote the ranking function learned from $S \in X^m \times X^n$.

- Expected error: $\mathbf{er}(f_S) = \mathbf{P}_{(x,x') \sim \mathcal{D}_+ \times \mathcal{D}_-} \left[f_S(x) < f_S(x') \right]$
- Empirical error: $\widehat{\mathbf{er}}_S(f_S) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}(f_S(x_i^+) < f_S(x_j^-))$

Assume $S \sim \mathcal{D}_+^m \times \mathcal{D}_-^n$. Can we bound $\mathbf{er}(f_S)$ in terms of $\widehat{\mathbf{er}}_S(f_S)$?

Generalization Bounds Based on Uniform Convergence

Let $f_S : X \to \mathbb{R}$ denote the ranking function learned from $S \in X^m \times X^n$.

Want to bound $\mathbf{er}(f_S)$ in terms of $\widehat{\mathbf{er}}_S(f_S)$.

Uniform convergence approach: If $f_S \in \mathcal{F}$, then

$$\mathbf{P}_S\left[\left|\mathbf{er}(f_S) - \widehat{\mathbf{er}}_S(f_S)\right| \geq \epsilon\right] \;\leq\; \mathbf{P}_S\left[\sup_{f \in \mathcal{F}}\left|\mathbf{er}(f) - \widehat{\mathbf{er}}_S(f)\right| \geq \epsilon\right].$$

Sufficient to bound this probability

[Vapnik & Chervonenkis, 1971]

Bounding $\mathbf{P}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbf{er}(f) - \widehat{\mathbf{er}}_S(f) \right| \geq \epsilon \right]$

Variance inequality due to Devroye (1991)

Step 1: Symmetrization

$$\mathbf{P}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbf{er}(f) - \widehat{\mathbf{er}}_S(f) \right| \geq \epsilon \right] \leq 2\,\mathbf{P}_{S,\tilde{S}} \left[\sup_{f \in \mathcal{F}} \left| \widehat{\mathbf{er}}_{\tilde{S}}(f) - \widehat{\mathbf{er}}_S(f) \right| \geq \frac{\epsilon}{2} \right]$$

Step 2: Permutations and reduction to a finite class

$$\mathbf{P}_{S,\tilde{S}} \left[\sup_{f \in \mathcal{F}} \left| \widehat{\mathbf{er}}_{\tilde{S}}(f) - \widehat{\mathbf{er}}_S(f) \right| \geq \frac{\epsilon}{2} \right] \leq \pi_{\mathcal{F}}(2m, 2n) \cdot 2 \exp\left(\frac{-mn\epsilon^2}{8(m+n)} \right)$$

Bipartite rank-shatter coefficients (New complexity measure for classes of ranking functions)

McDiarmid's inequality

Uniform Convergence Bound

Theorem. Let $\mathcal{F}$ be a class of real-valued functions on X. Then for any $0 < \delta < 1$, we have with probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}} \left| \mathbf{er}(f) - \widehat{\mathbf{er}}_S(f) \right| \;<\; \sqrt{\frac{8(m+n)}{mn} \left(\ln \pi_{\mathcal{F}}(2m, 2n) + \ln\left(\frac{4}{\delta}\right) \right)}\,.$$

[Agarwal et al, 2005]

Road Map

Theory

Algorithms

Applications

Bipartite Ranking: Basic Algorithmic Framework

Minimize a convex upper bound on the empirical ranking error, possibly with some regularization, over some class of ranking functions:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \ell(f, x_i^+, x_j^-) + \lambda N(f) \right]$$

where

$$\begin{aligned} \ell(f, x_i^+, x_j^-) &: \text{ convex upper bound on } \mathbf{1}(f(x_i^+) < f(x_j^-)) \\ N(f) &: \text{ regularizer} \\ \lambda > 0 &: \text{ regularization parameter} \\ \mathcal{F} &: \text{ class of ranking functions} \end{aligned}$$

Bipartite RankBoost Algorithm

$$\min_{f \in \mathcal{L}(\mathcal{F}_{\text{base}})} \left[\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \ell_{\text{exp}}(f, x_i^+, x_j^-) \right]$$

$$\ell_{\text{exp}}(f, x_i^+, x_j^-) = \exp\left(-\left(f(x_i^+) - f(x_j^-)\right)\right)$$

$\mathcal{L}(\mathcal{F}_{\text{base}})$ = linear combinations of functions in some base class $\mathcal{F}_{\text{base}}$

[Freund et al, 2003]

Bipartite RankSVM Algorithm

$$\min_{f \in \mathcal{F}_K} \left[\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \ell_{\text{hinge}}(f, x_i^+, x_j^-) + \frac{\lambda}{2} \|f\|_K^2 \right]$$

$$\ell_{\text{hinge}}(f, x_i^+, x_j^-) = \Big(1 - \big(f(x_i^+) - f(x_j^-)\big)\Big)_+ \quad \big[u_+ = \max(u, 0)\big]$$

$\mathcal{F}_K$ = reproducing kernel Hilbert space (RKHS) with kernel function K

$$N(f) = \frac{\|f\|_K^2}{2}$$

[Herbrich et al, 2000; Joachims, 2002; Rakotomamonjy, 2004]

Bipartite RankNet Algorithm

$$\min_{f \in \mathcal{F}_{\text{neural}}} \left[\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \ell_{\text{logistic}}(f, x_i^+, x_j^-) \right]$$

$$\ell_{\text{logistic}}(f, x_i^+, x_j^-) \; = \; \log\Big(1 + \exp\Big(-\Big(f(x_i^+) - f(x_j^-)\Big)\Big)\Big)$$

$\mathcal{F}_{\text{neural}}$ = functions represented by some class of neural networks

[Burges et al, 2005]

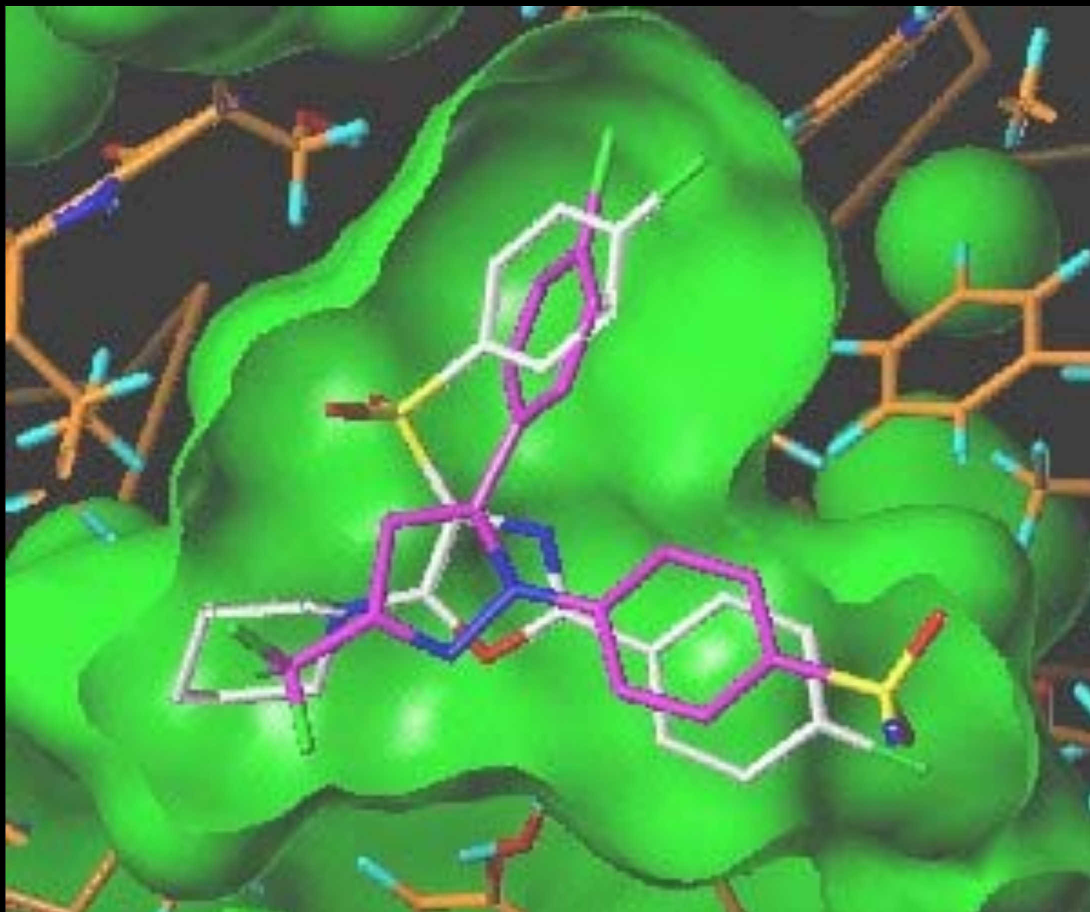
Road Map

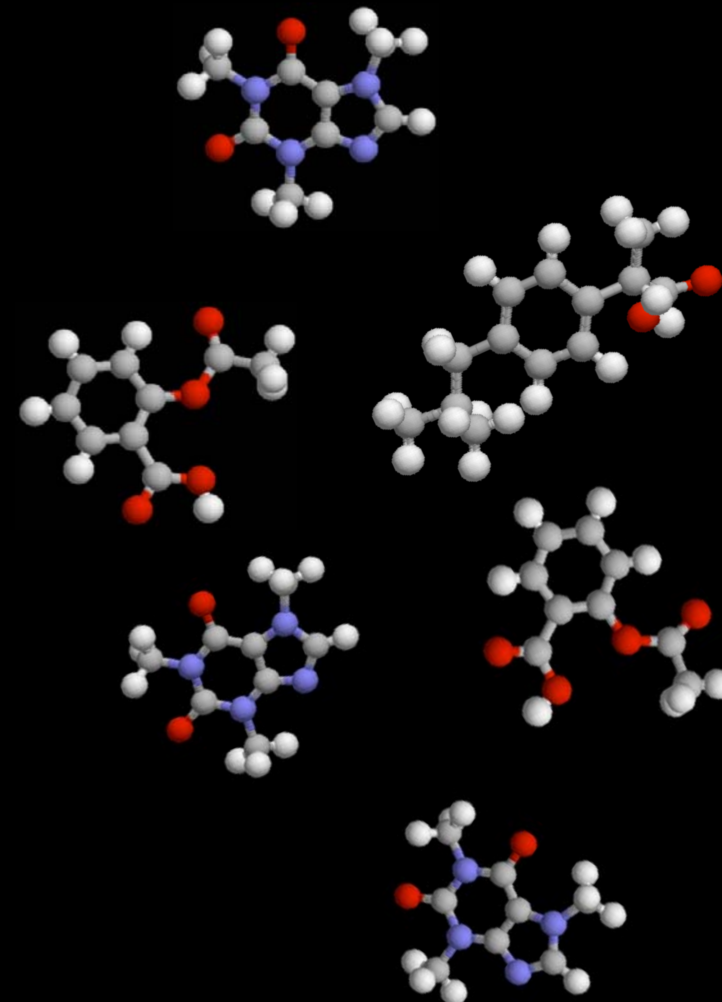
Theory

Algorithms

Applications

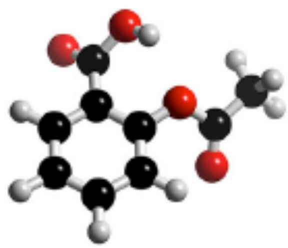
Application to Drug Discovery



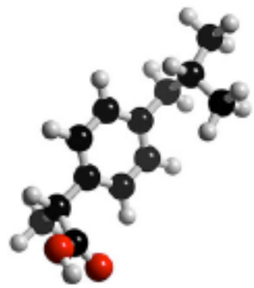


Problem: Millions of structures in a chemical library.
How do we identify the most promising ones?

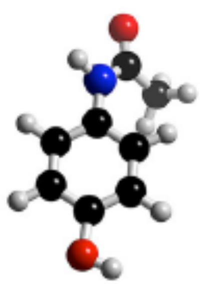
Formulation as a Ranking Problem with Real-Valued Labels



$\mathrm{pIC}_{50} = 5.6718$



$\mathrm{pIC}_{50} = 8.2991$



$\mathrm{pIC}_{50} = 4.1317$

$\cdots$

Ranking With Real-Valued Labels

- Instance space X
- Real-valued labels $Y = \mathbb{R}$
- Input: Training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times \mathbb{R})^m$
- Output: Ranking function $f : X \rightarrow \mathbb{R}$

- Expected error:

$$\mathbf{er}(f) = \mathbf{E}_{((x,y),(x',y'))\sim\mathcal{D}\times\mathcal{D}}\left[|y - y'|\, \mathbf{1}((y - y')(f(x) - f(x')) < 0)\right]$$

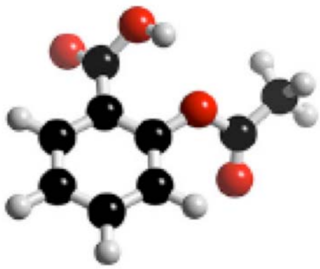
- Empirical error:

$$\widehat{\mathbf{er}}_S(f) = \frac{1}{\binom{m}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} |y - y'|\, \mathbf{1}((y_i - y_j)(f(x_i) - f(x_j)) < 0)$$

Cheminformatics Data Sets

[Sutherland et al, 2004]

Data Set	No. of Compounds	No. of Chemical (2.5D) Descriptors	pIC_{50} Values
DHFR inhibitors	361	70	3.3 – 9.8
COX2 inhibitors	292	74	4.0 – 9.0



DHFR Results Using RankSVM

2.5D chemical descriptors
Gaussian kernel

Training size	Ranking error	
	SVR	RankSVM
24	0.4755	**0.4601**
48	**0.3430**	0.3509
72	0.2840	**0.2726**
96	0.2483	**0.2351**
120	0.2171	**0.2121**
144	**0.2023**	0.2032
168	0.2019	**0.1817**
192	0.1808	**0.1749**
216	0.1816	**0.1722**
237	0.1714	**0.1681**

FP2 molecular fingerprints
Tanimoto kernel

Training size	Ranking error	
	SVR	RankSVM
24	0.3793	**0.3546**
48	0.2905	**0.2896**
72	0.2517	**0.2421**
96	0.2343	**0.2201**
120	0.2147	**0.2052**
144	0.2166	**0.1988**
168	0.2096	**0.1966**
192	0.2056	**0.1962**
216	0.1907	**0.1787**
237	0.1924	**0.1798**

[Agarwal et al, 2010]

about

visiting | maps | offices+services

admissions

undergrad | graduate | financial aid

education

schools+courses | OpenCourseWare

research

labs+centers | lincoln lab | libraries

community

students | faculty | staff | alumni

life@MIT

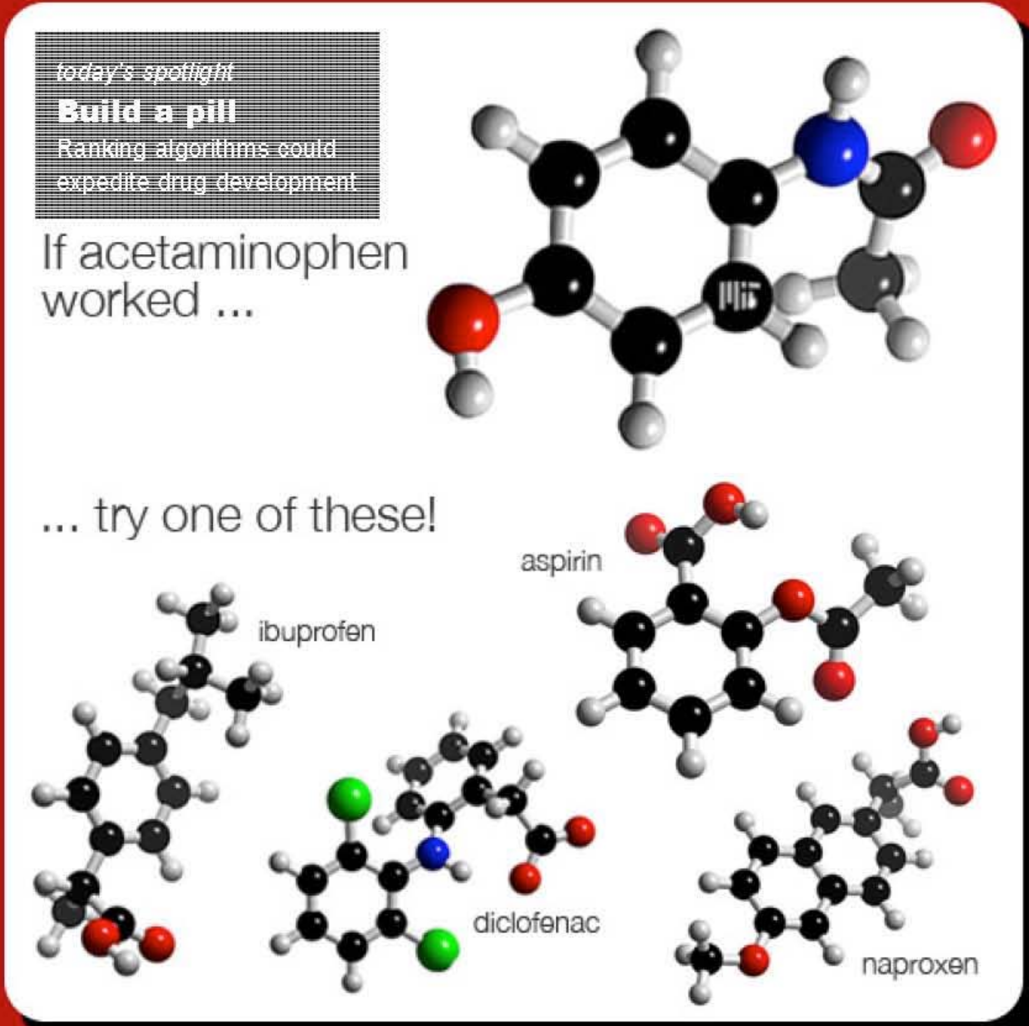
arts | athletics | video

initiatives

energy | cancer | diversity | global

impact

industry | public service



news

Letter to the community on MIT's financial condition

In The World: Better wound treatment for all

Toward more efficient wireless power delivery

research | campus | press

events

Memorial service for former MIT President Howard W. Johnson (today)

Artists Beyond the Desk concert (today)

Of Note: Celebrating SA+P's new Program in Art, Culture and Technology (tomorrow)

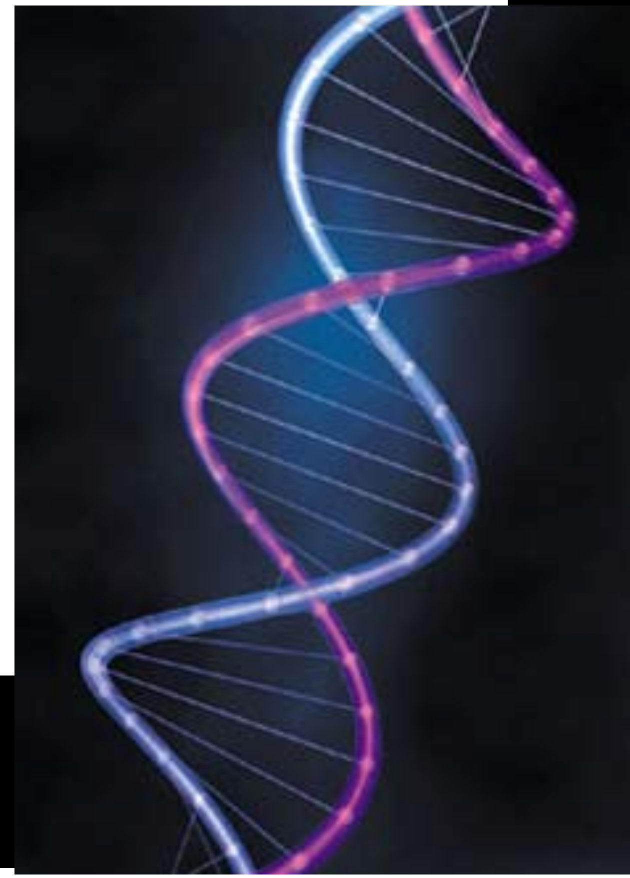
Legatum Lecture: Entrepreneur and investor Chuck Lacy (tomorrow)

Application to Bioinformatics

Searching for genetic determinants in the new millennium

N.J. Risch

Human genetics is now at a critical juncture. The molecular methods used successfully to identify the genes underlying rare mendelian syndromes are failing to find the numerous genes causing more common, familial, non-mendelian diseases . . .



Nature **405**:847–856, 2000

Application to Bioinformatics

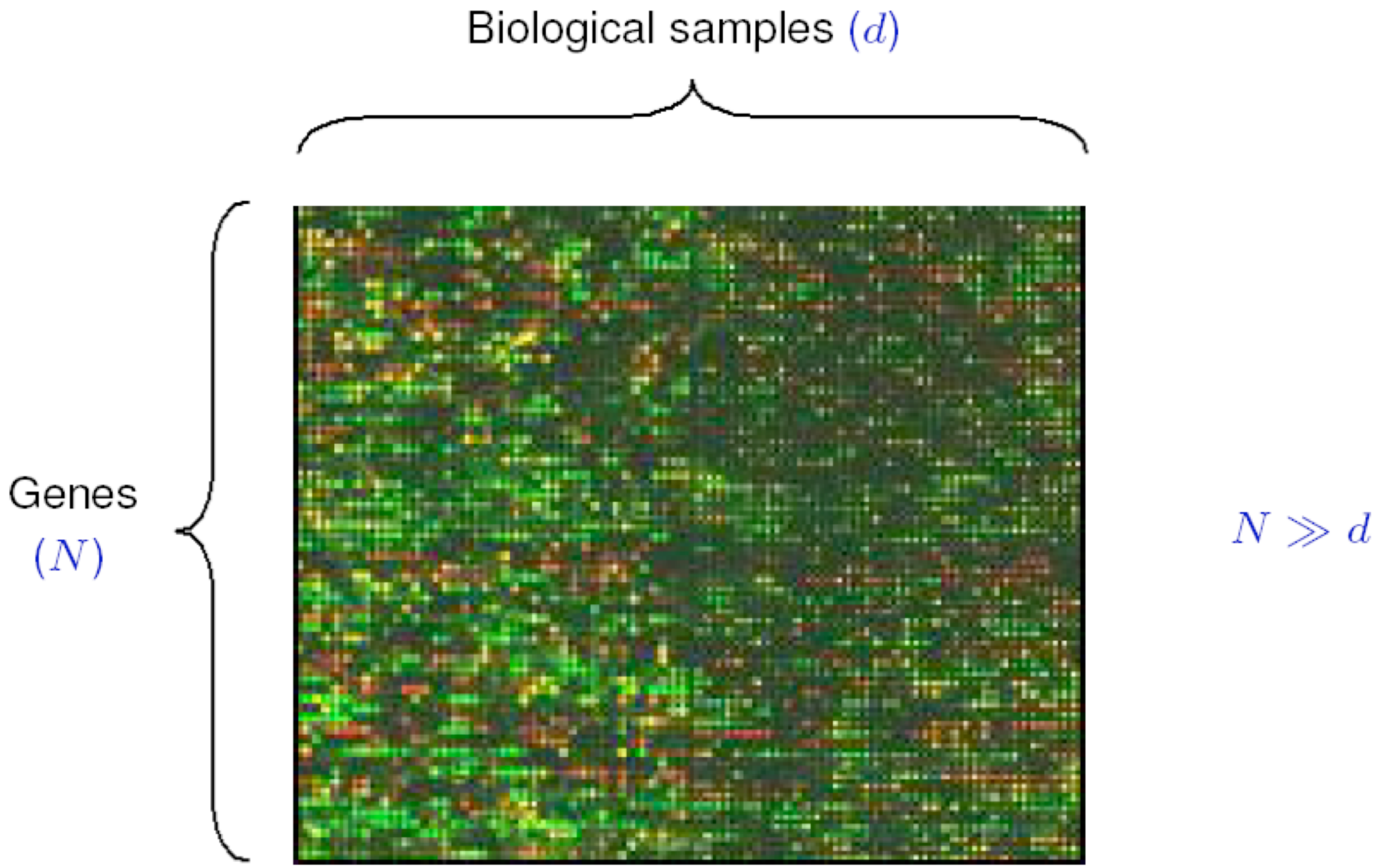
Searching for genetic determinants in the new millennium

N.J. Risch

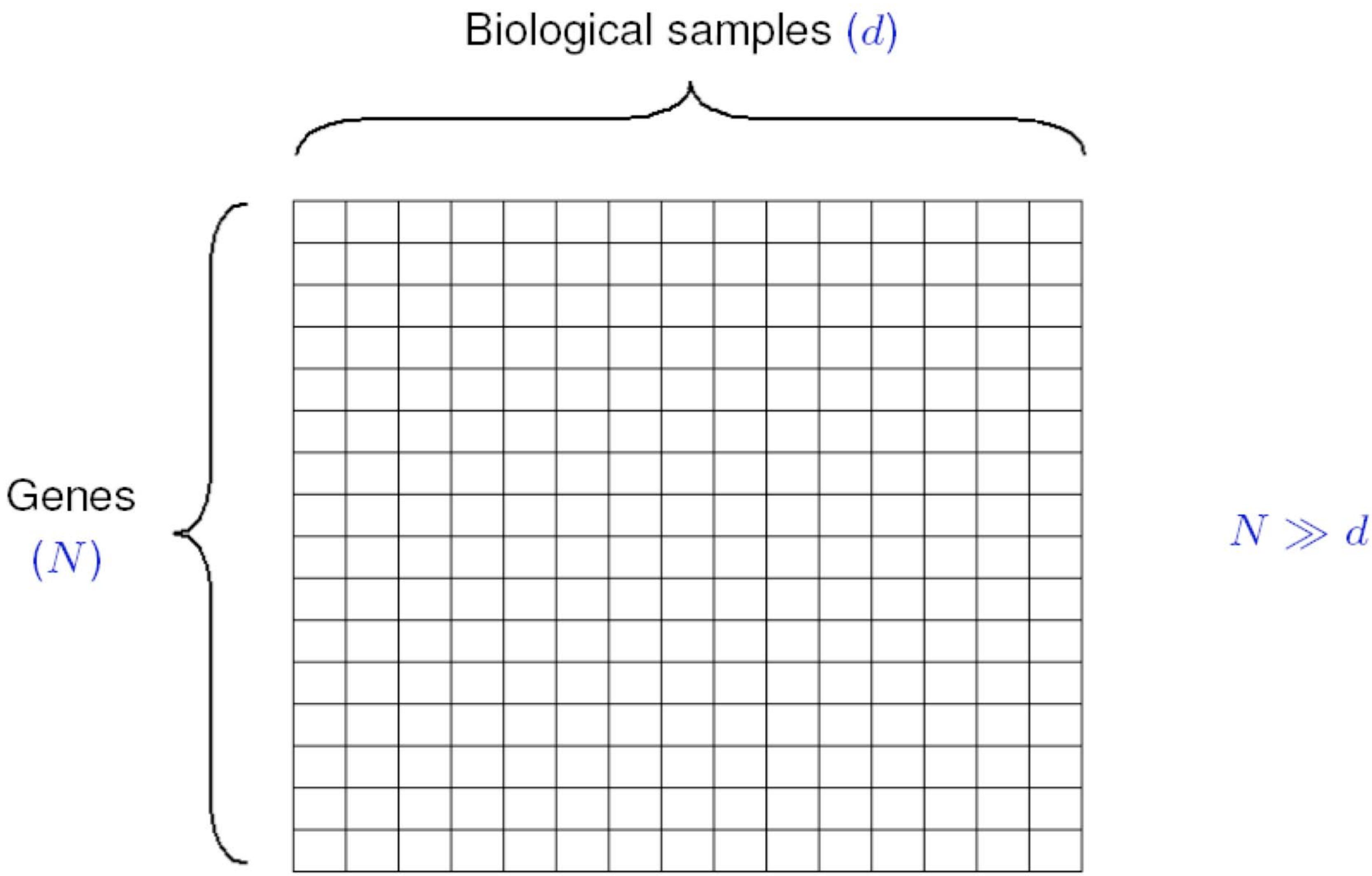
With the human genome sequence nearing completion, new opportunities are being presented for unravelling the complex genetic basis of nonmendelian disorders based on large-scale genomewide studies . . .

Nature **405**:847–856, 2000

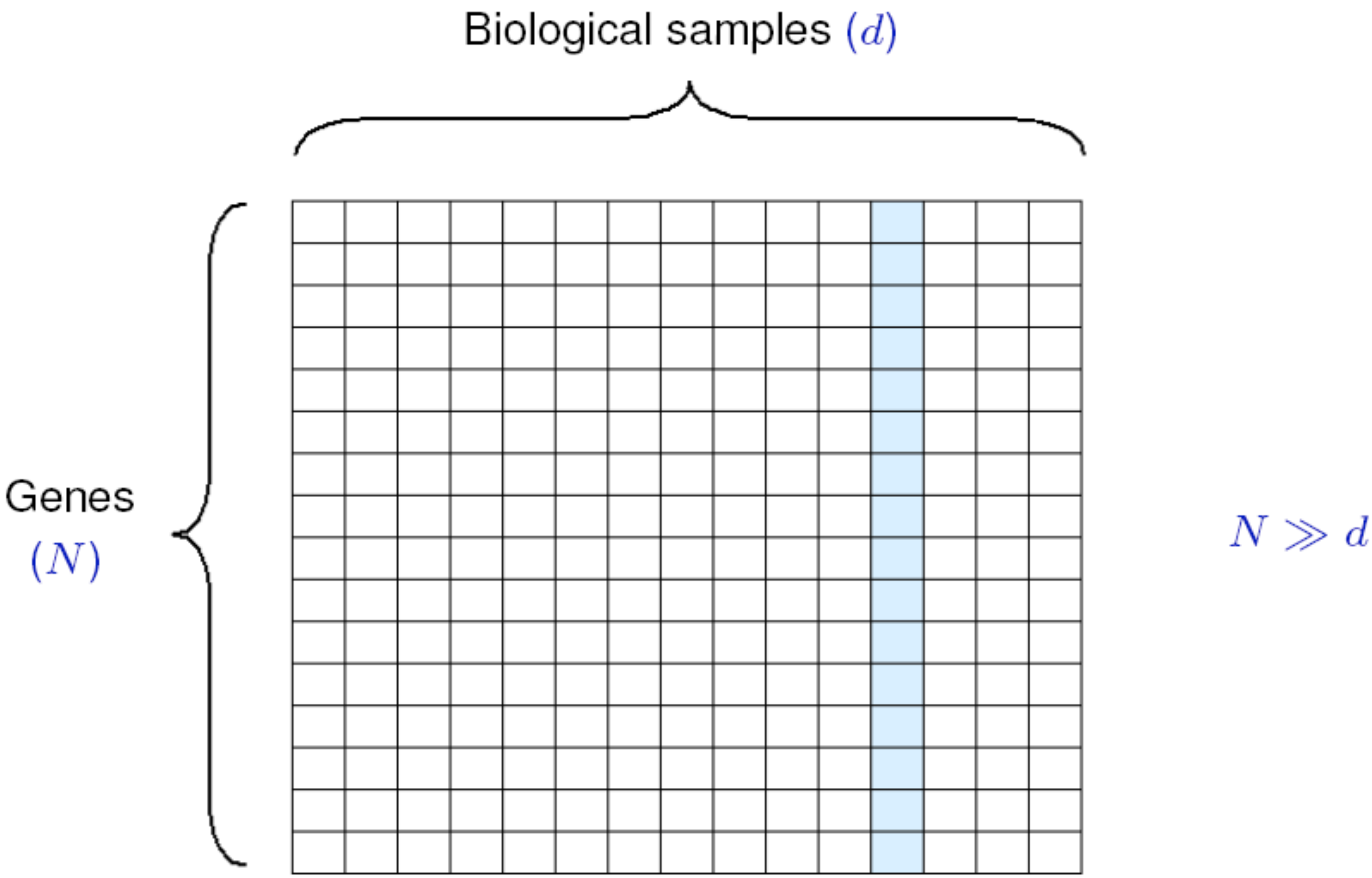
Identifying Genes Relevant to a Disease Using Microarrray Gene Expression Data



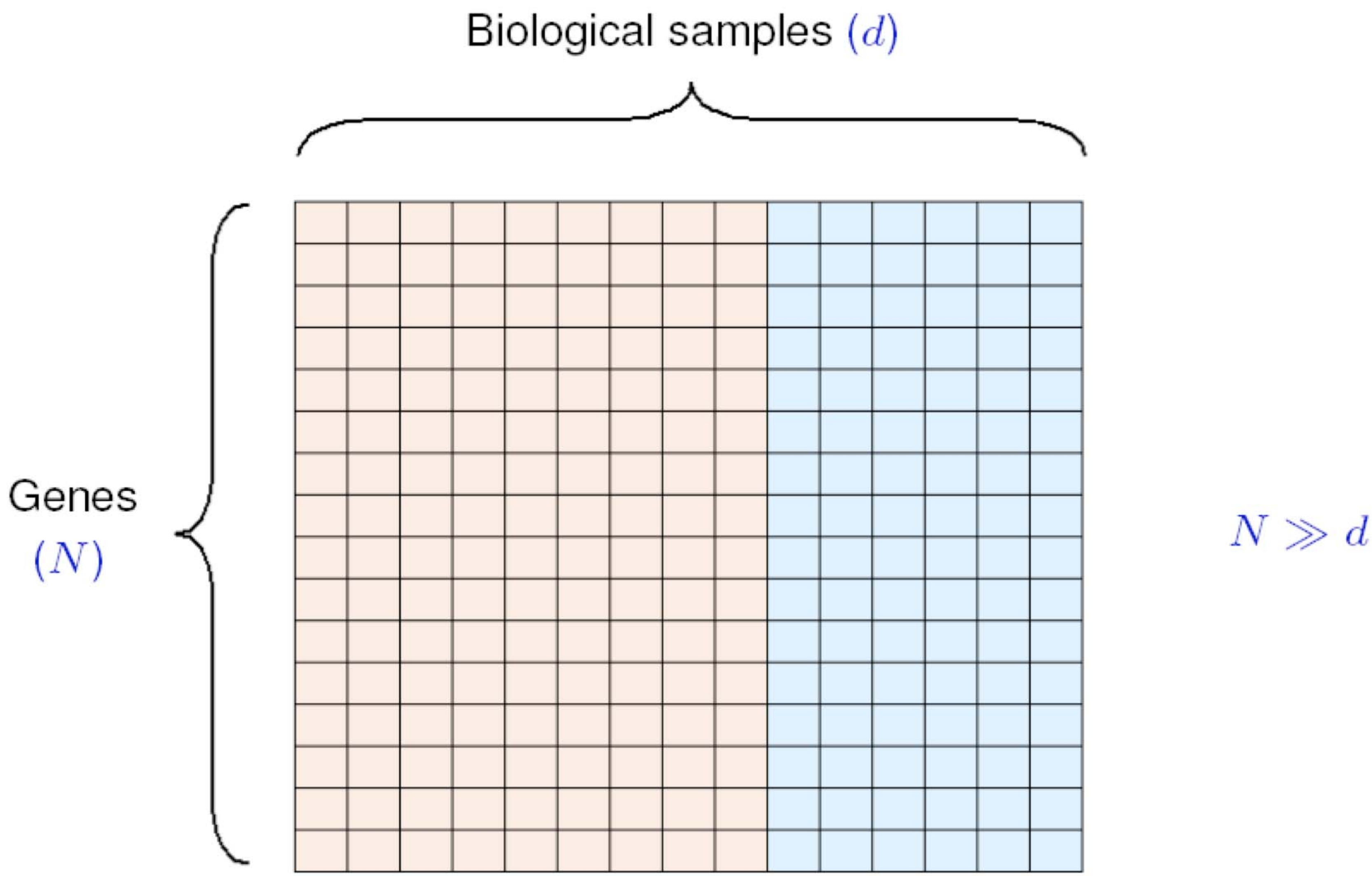
Identifying Genes Relevant to a Disease Using Microarrray Gene Expression Data



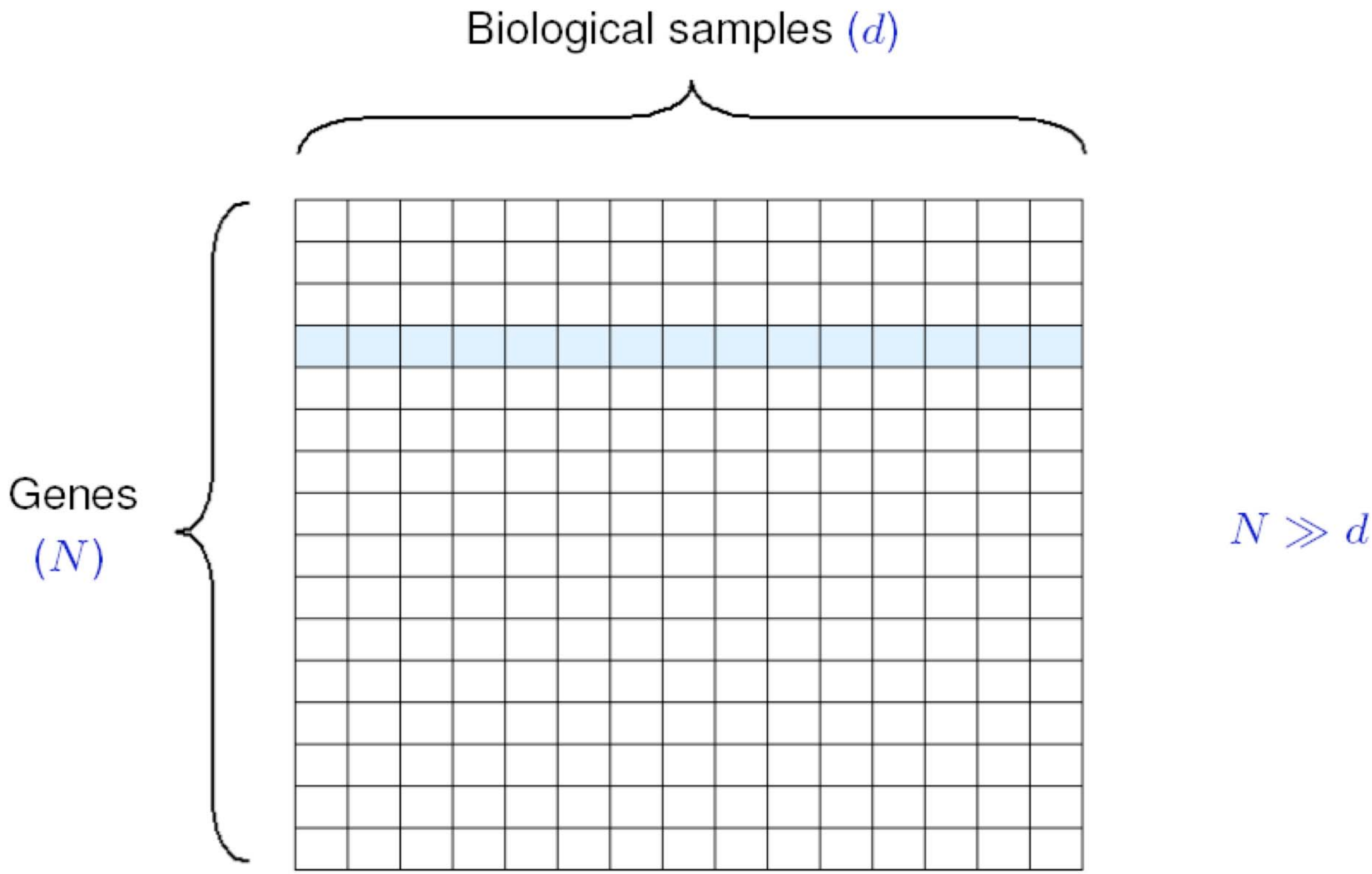
Identifying Genes Relevant to a Disease Using Microarrray Gene Expression Data



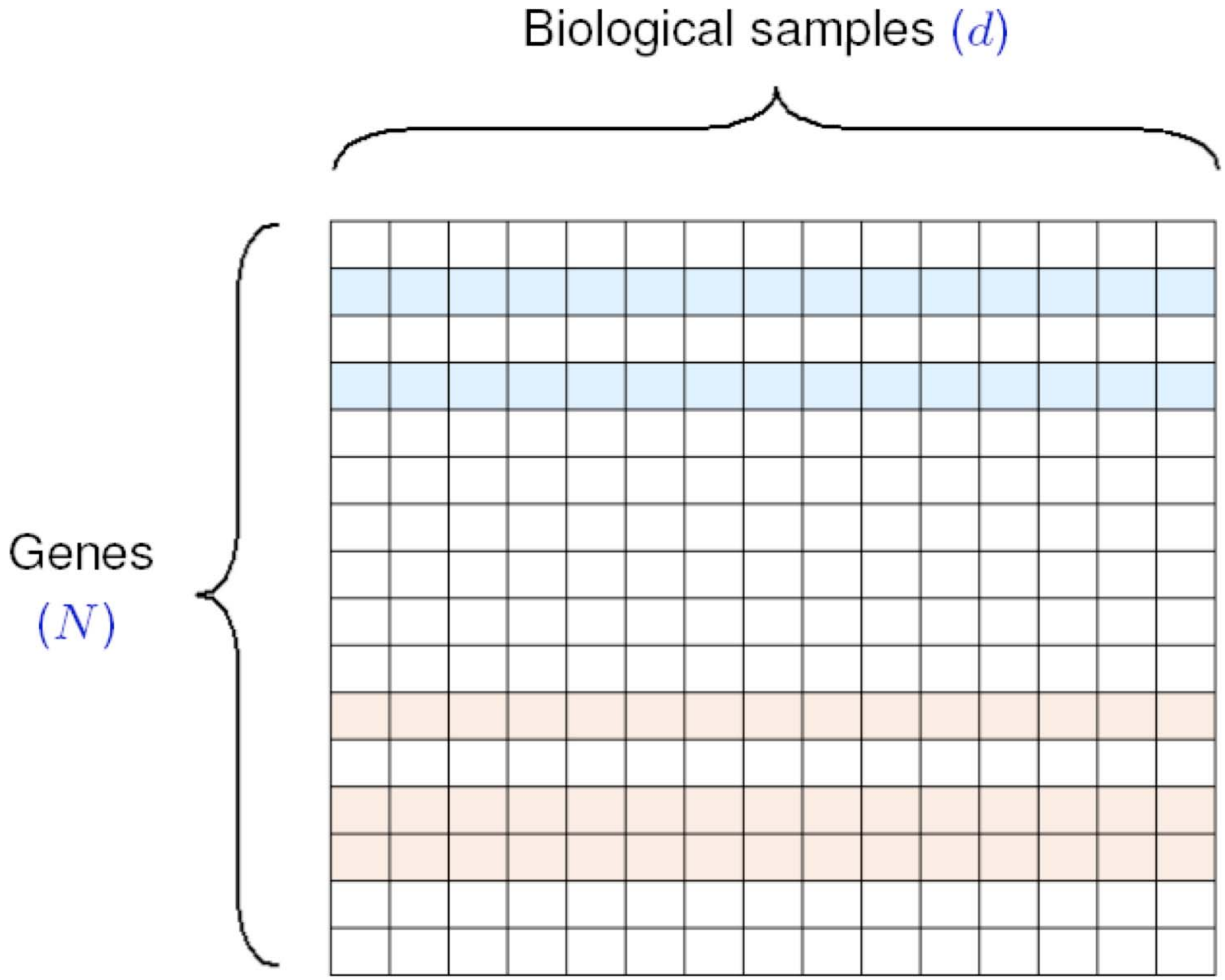
Identifying Genes Relevant to a Disease Using Microarrray Gene Expression Data



Identifying Genes Relevant to a Disease Using Microarrray Gene Expression Data



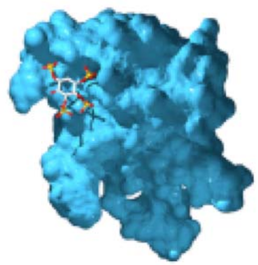
Identifying Genes Relevant to a Disease Using Microarrray Gene Expression Data

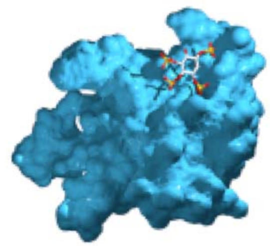


$N \gg d$

Formulation as a Bipartite Ranking Problem

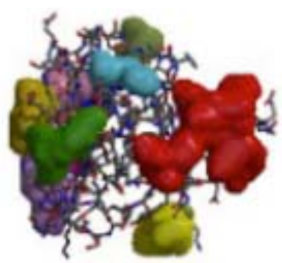
Relevant

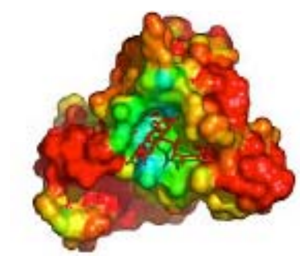


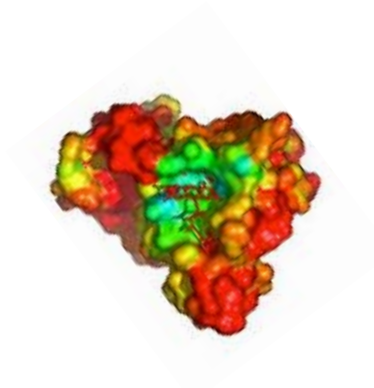


…

Not relevant





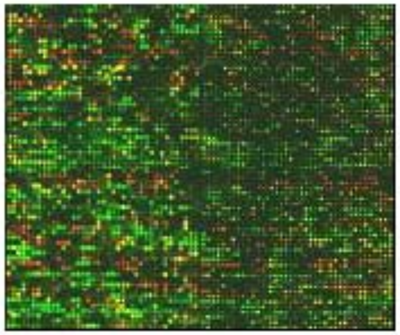


…

Microarray Gene Expression Data Sets

[Golub et al, 1999; Alon et al, 1999]

Data Set	No. of Genes	No. of Tissue Samples	Notes
Leukemia	7129	72	25 AML / 47 ALL
Colon cancer	2000	62	40 tumor / 22 normal



Selection of Training Genes

Leukemia	Colon cancer
Positive genes: Markers for AML/ALL	**Positive genes: Markers for colon cancer**
Myeloperoxidase CD13 CD33 HOXA9 Homeo box A9 V-myb CD19 CD10 (CALLA) TCL1 (T cell leukemia) C-myb Deoxyhypusine synthase	Phospholipase A2 Keratin 6 isoform PTP-H1 TF-IIIA V-raf oncogene MAPK kinase 1 CEA Oncoprotein 18 PEP carboxykinase ERK kinase 1
Negative genes	**Negative genes**
157 genes involved in physiological cellular functions	56 genes involved in physiological cellular functions

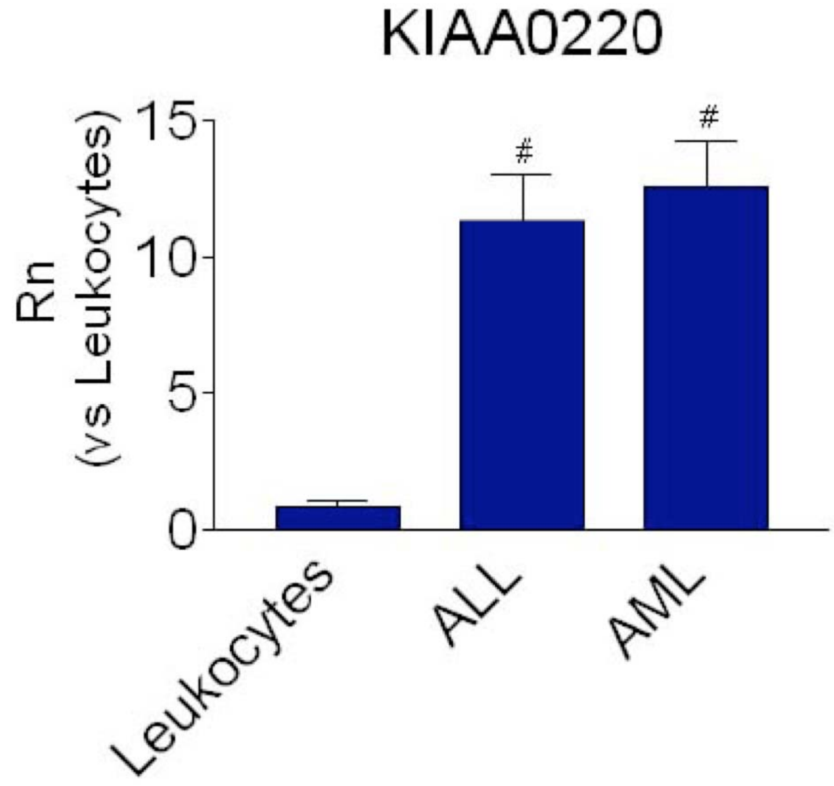
Top-Ranking Genes for Leukemia Returned by RankBoost

♦ Known marker; ♦ Potential marker;
■ Known therapeutic target; ■ Potential therapeutic target;
x No link found.

	Gene	Relevance Summary	t-Statistic Rank	Pearson Rank
1.	KIAA0220	■	6628	2461
2.	G-gamma globin	♦	3578	3567
3.	Delta-globin	♦	3663	3532
4.	Brain-expressed HHCPA78 homolog	■	6734	2390
5.	Myeloperoxidase	♦	139	6573
6.	Disulfide isomerase precursor	■	6650	575
7.	Nucleophosmin	♦	405	1115
8.	CD34	♦	6732	643
9.	Elongation factor-1β	x	4460	3413
10.	CD24	♦	81	1
11.	60S ribosomal protein L23	■	1950	73
12.	5-aminolevulinic acid synthase	■	4750	3351

[Agarwal & Sengupta, 2009]

Biological Validation



[Agarwal et al, 2010]

Further Topics & Some Pointers
[Incomplete!]

- Ranking performance measures that focus on accuracy at the top [Yue et al, 2007; Clemencon & Vayatis, 2007; Cossock & Zhang, 2008; Rudin, 2009; Agarwal, 2010; also see IR ranking algorithms below]

- Statistical consistency of ranking algorithms [Clemencon & Vayatis, 2007; Clemencon & Vayatis, 2008; Cossock & Zhang, 2008; Duchi et al, 2010]

- Other types of ranking problems, such as label ranking [Crammer & Singer, 2003; Shalev-Shwartz & Singer, 2006] and subset ranking [Cossock & Zhang, 2008]

- Ranking algorithms for information retrieval [many, many recent papers; see Liu, 2009 for a survey]

- Other applications of ranking, such as game move prediction [Stern et al, 2007], recommendation systems [Stern et al, 2009], manhole event prediction [Rudin et al, 2010]