statistical systems biology

department of applied physics and applied mathematics
+
center for computational biology and bioinformatics
columbia university
chris.wiggins@columbia.edu

statistical systems biology: agenda

1. challenges to keep in mind
2. microarrays / regulation
3. networks
4. final thoughts

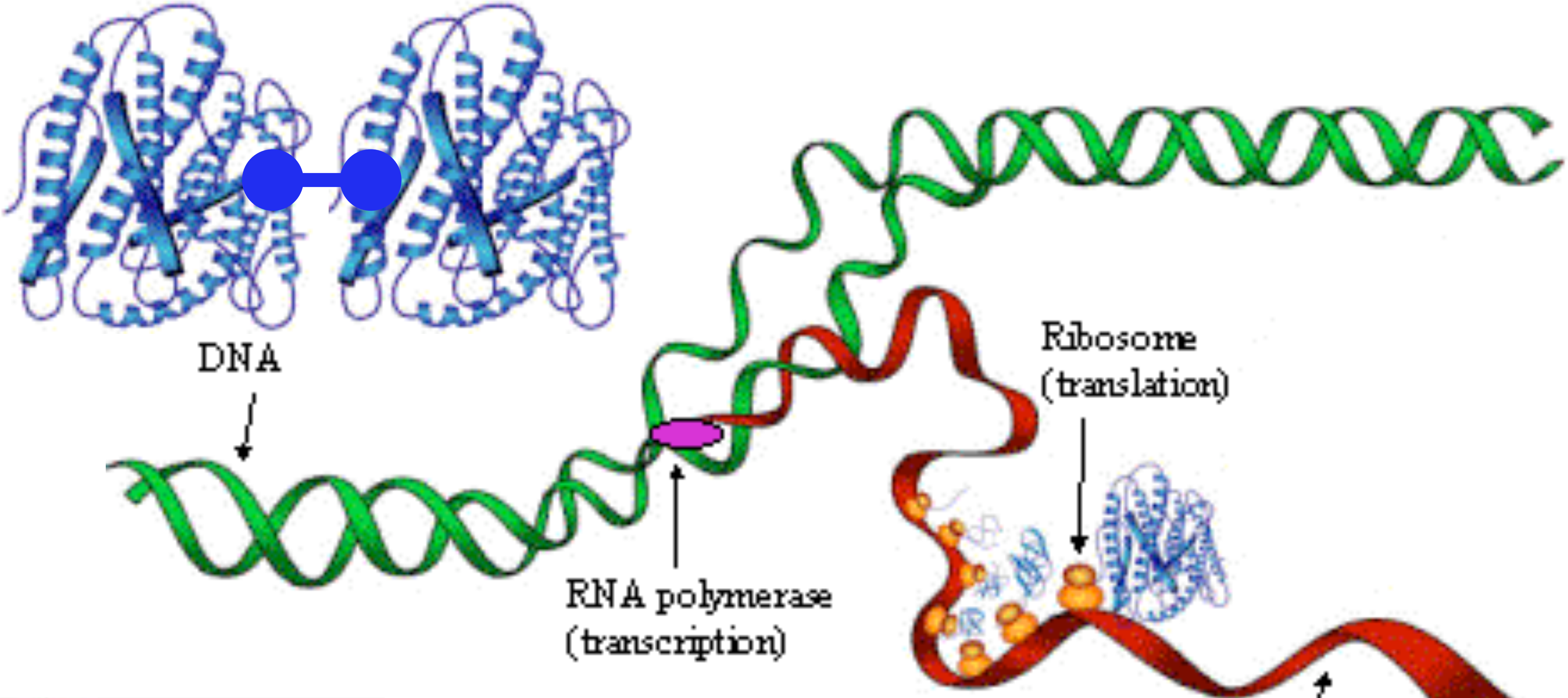
statistical systems biology: challenges

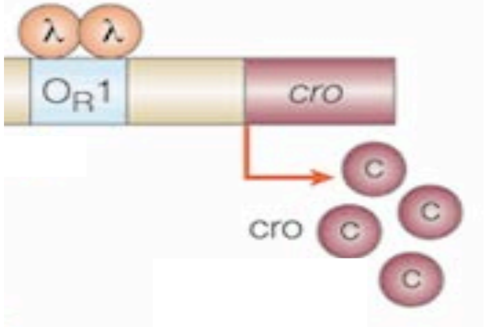
1. statistics
2. modeling
3. validation
4. interpretation

microarrays + transcriptional regulation

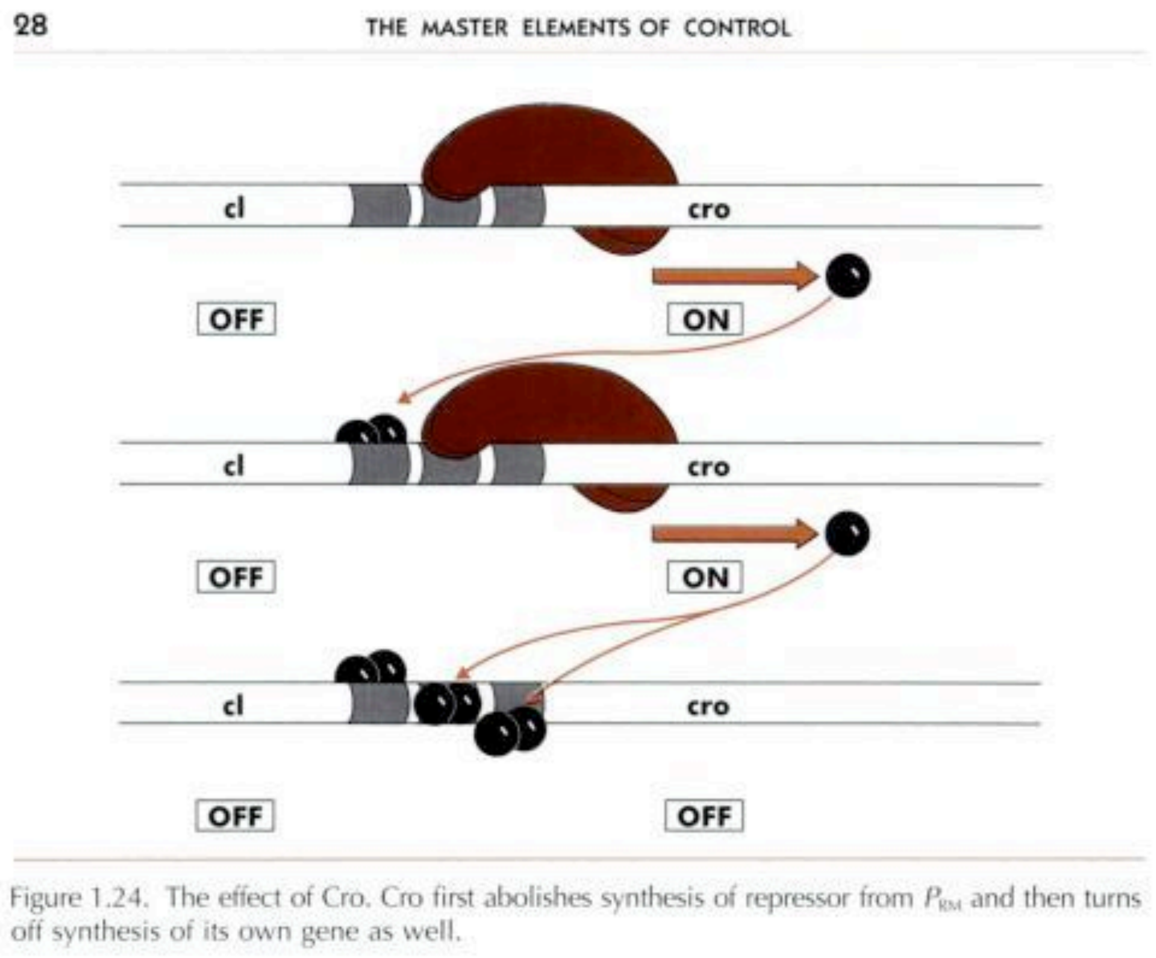
1. biological questions
2. history/context
3. methods
   - “unsupervised”: cluster first, ask questions later
   - “supervised”: predicting methods

biology as told by a theorist





biology as told by a biologist



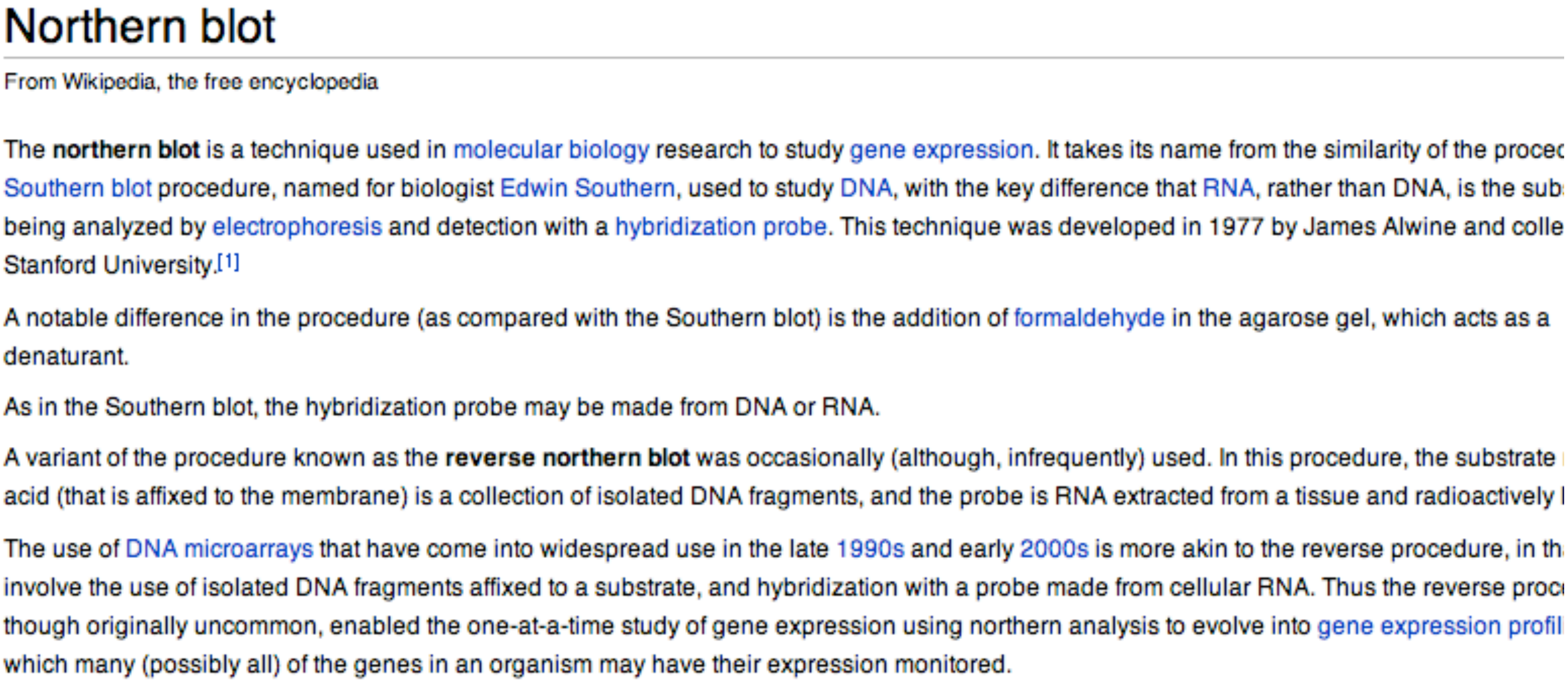
28 THE MASTER ELEMENTS OF CONTROL

Figure 1.24. The effect of Cro. Cro first abolishes synthesis of repressor from P_{RM} and then turns off synthesis of its own gene as well.

ptashne's “a genetic switch”

what is to be measured?

1. “expression” via RNA abundance



Northern blot

From Wikipedia, the free encyclopedia

The **northern blot** is a technique used in molecular biology research to study gene expression. It takes its name from the similarity of the proce Southern blot procedure, named for biologist Edwin Southern, used to study DNA, with the key difference that RNA, rather than DNA, is the sub being analyzed by electrophoresis and detection with a hybridization probe. This technique was developed in 1977 by James Alwine and colle Stanford University.[1]

A notable difference in the procedure (as compared with the Southern blot) is the addition of formaldehyde in the agarose gel, which acts as a denaturant.

As in the Southern blot, the hybridization probe may be made from DNA or RNA.

A variant of the procedure known as the **reverse northern blot** was occasionally (although, infrequently) used. In this procedure, the substrate acid (that is affixed to the membrane) is a collection of isolated DNA fragments, and the probe is RNA extracted from a tissue and radioactively

The use of DNA microarrays that have come into widespread use in the late 1990s and early 2000s is more akin to the reverse procedure, in th involve the use of isolated DNA fragments affixed to a substrate, and hybridization with a probe made from cellular RNA. Thus the reverse proc though originally uncommon, enabled the one-at-a-time study of gene expression using northern analysis to evolve into gene expression profil which many (possibly all) of the genes in an organism may have their expression monitored.

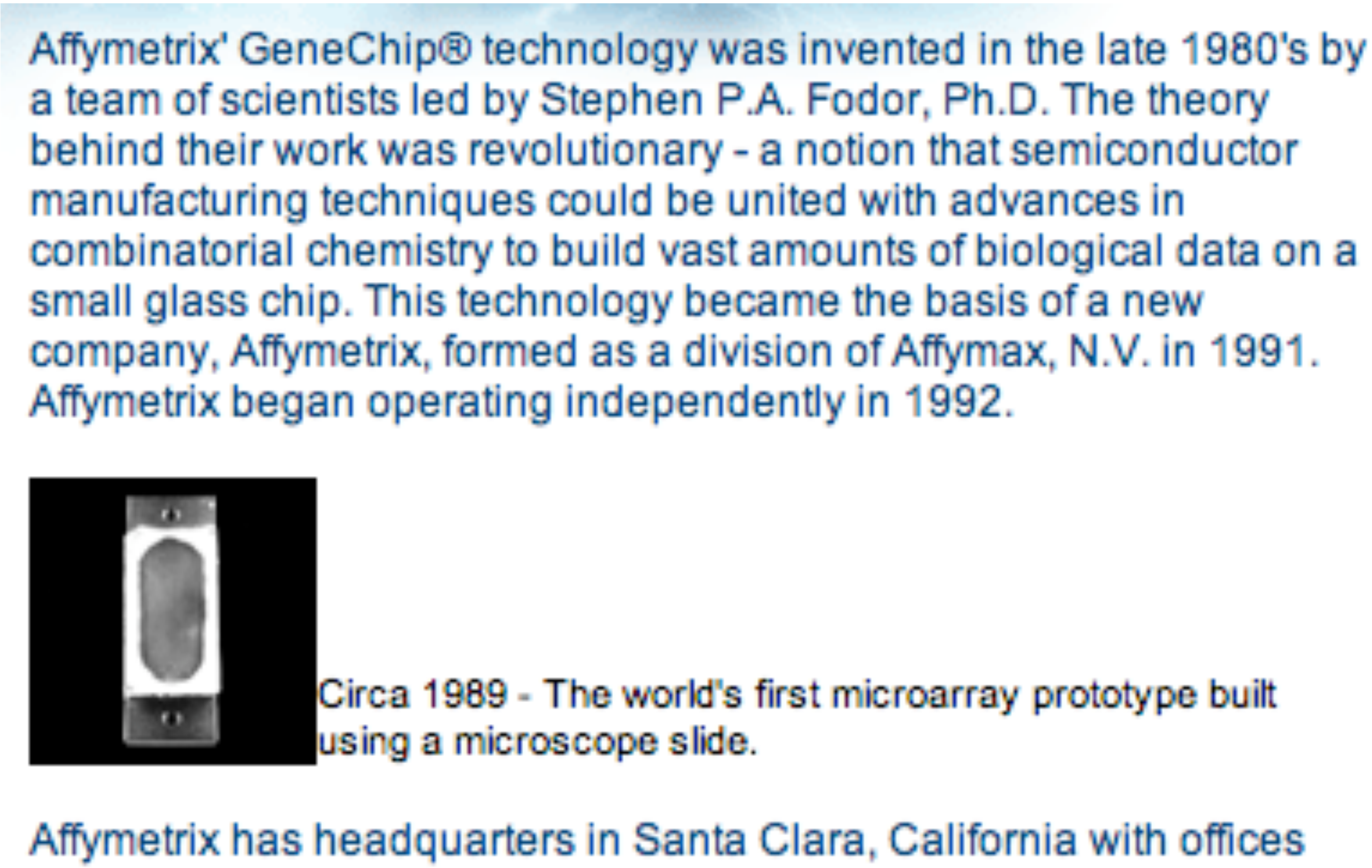
what is to be measured?

2. regulatory sequence

```
>YLR081W         GAL2
                                         CEN
AGGTTGCAATTTCTTTTTCTATTAGTAGCTAAAAATGGGTCACGTGATCT      -451
                                       GAL4
ATATTCGAAAGGGGCGGTTGCCTCAGGAAGGCACCGGCGGTCTTTCGTCC      -401

GTGCGGAGATATCTGCGCCGTTCAGGGGTCCATGTGCCTTGGACGATATT      -351
                       GAL4
AAGGCAGAAGGCAGTATCGGGGCGGATCACTCCGAACCGAGATTAGTTAA      -301
GCCCTTCCCATCTCAAGATGGGGAGCAAATGGCATTATACTCCTGCTAGA      -251
AAGTTAACTGTGCACATATTCTTAAATTATACAACATTCTGGAGAGCTAT      -201
TGTTCAAAAAACAAACATTTCGCAGGCTAAAATGTGGAGATAGGATAAGT      -151
TTTGTAGACATATATAAACAATCAGTAATTGGATTGAAAATTTGGTGTTG      -101
TGAATTGCTCTTCATTATGCACCTTATTCAATTATCATCAAGAATAGTAA      -51
TAGTTAAGTAAACACAAGATTAACATAATAAAAAAAATAATTCTTTCATA      -1
ATGGCAGTTGAGGAGAACAATATGCCTGTTGTTTCACAGCAACCCCAAGC      +50
```

GeneChip(R): “late 80’s”

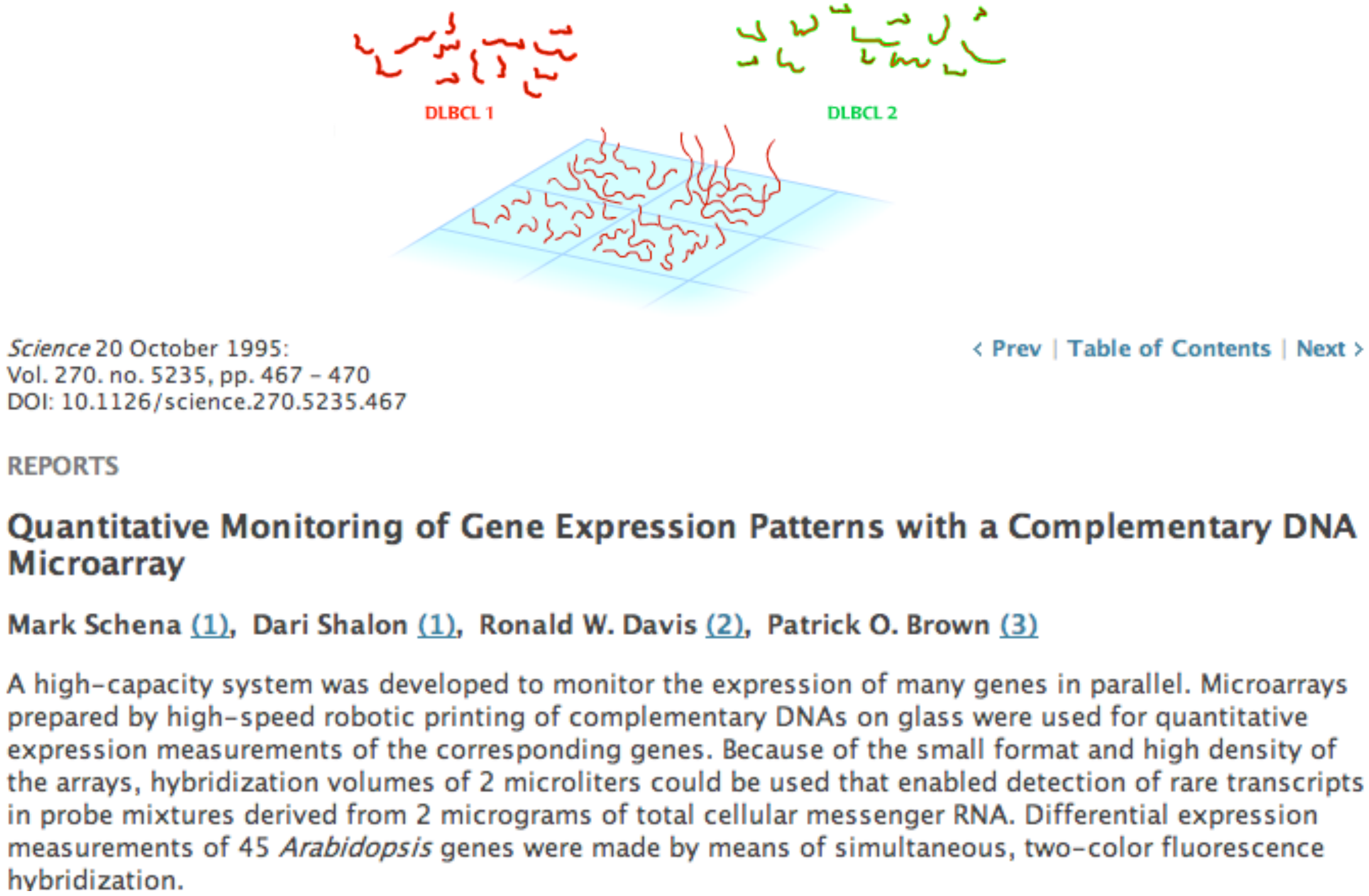


Affymetrix' GeneChip® technology was invented in the late 1980's by a team of scientists led by Stephen P.A. Fodor, Ph.D. The theory behind their work was revolutionary - a notion that semiconductor manufacturing techniques could be united with advances in combinatorial chemistry to build vast amounts of biological data on a small glass chip. This technology became the basis of a new company, Affymetrix, formed as a division of Affymax, N.V. in 1991. Affymetrix began operating independently in 1992.

Circa 1989 - The world's first microarray prototype built using a microscope slide.

Affymetrix has headquarters in Santa Clara, California with offices

cDNA “spot” arrays: 1995



Science 20 October 1995:
Vol. 270. no. 5235, pp. 467 - 470
DOI: 10.1126/science.270.5235.467

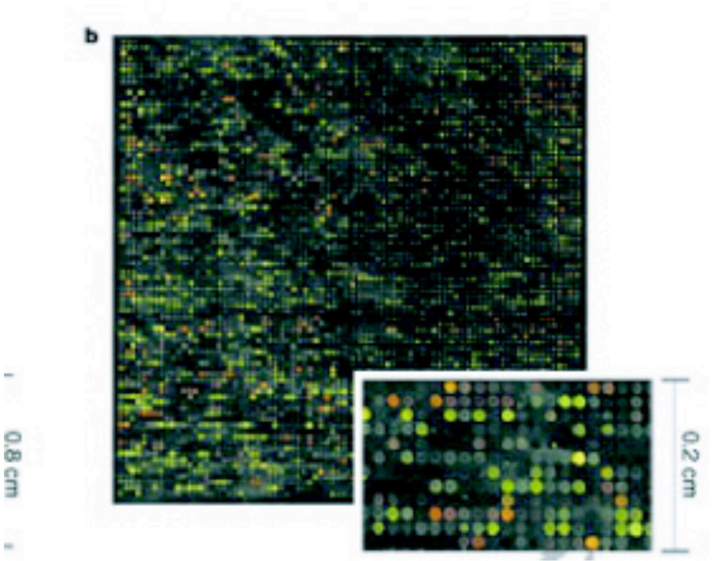
REPORTS

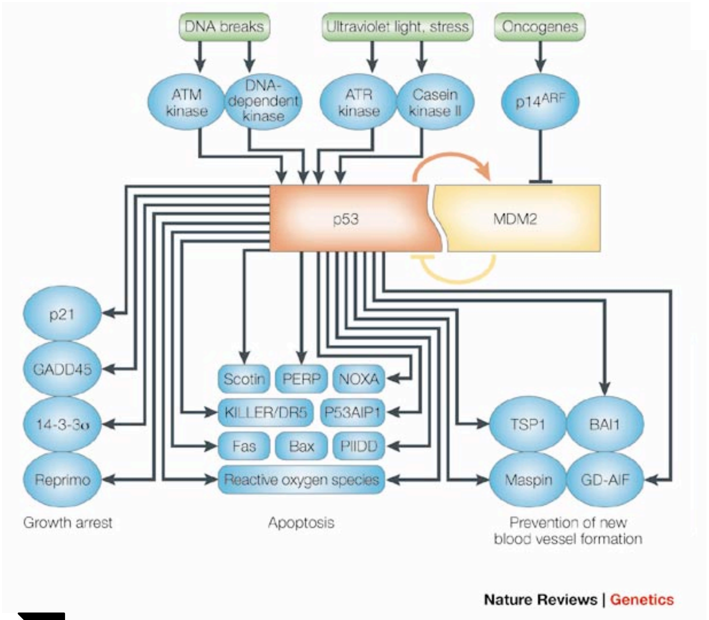
Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena (1), Dari Shalon (1), Ronald W. Davis (2), Patrick O. Brown (3)

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

the hope



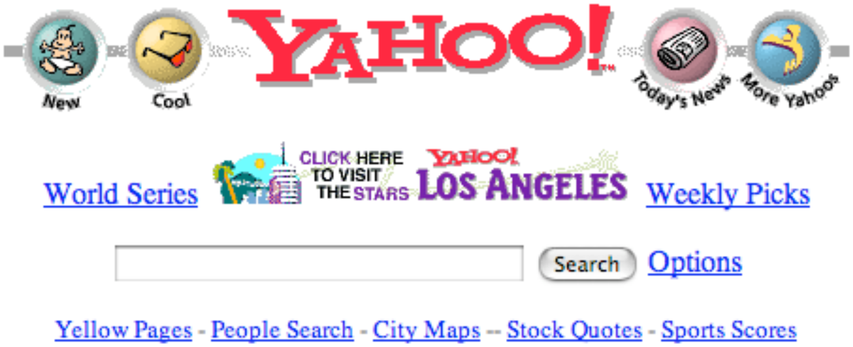


Nature Reviews | Genetics

Nature Reviews | Genetics

?

other relevant innovation:

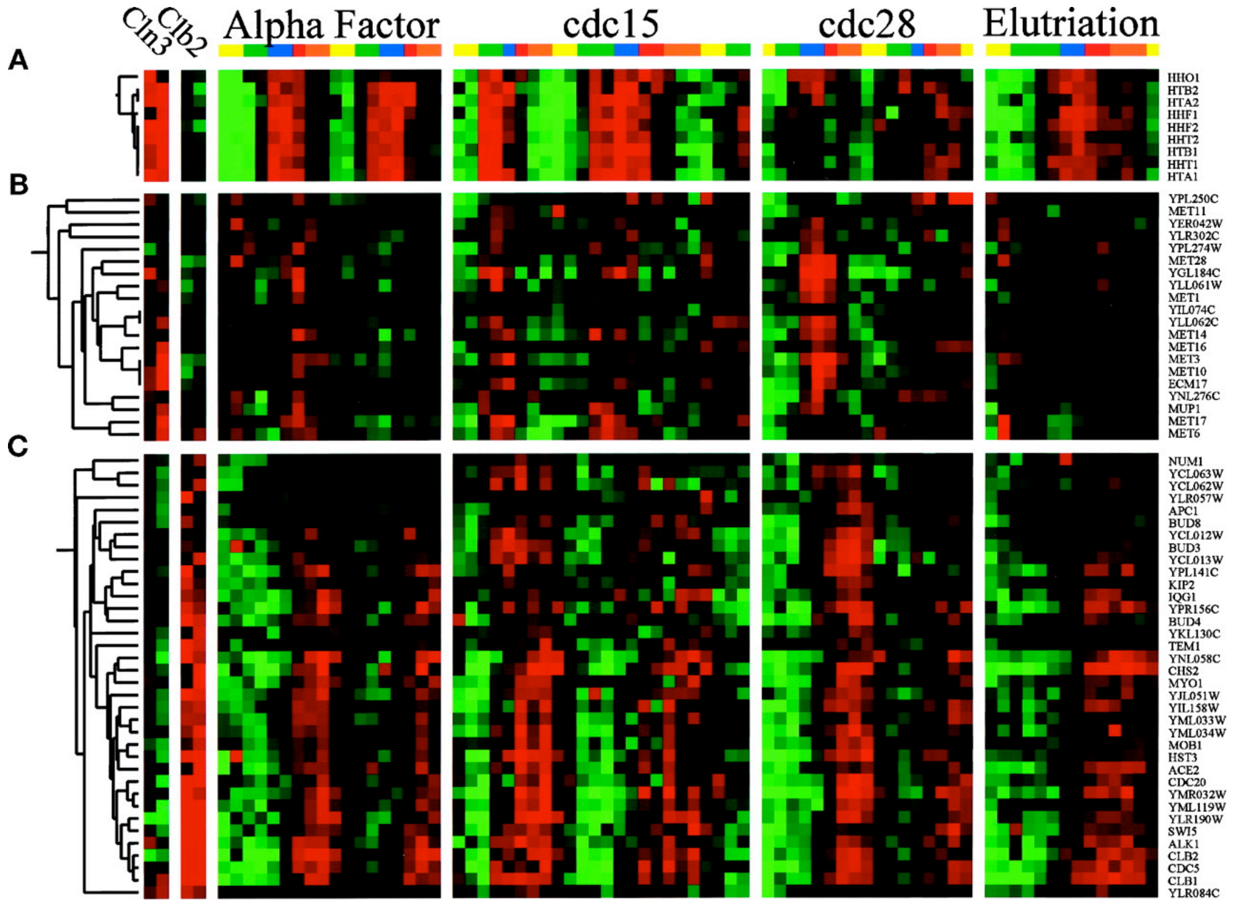

New
Cool
YAHOO!
Today's News
More Yahoos
World Series
CLICK HERE TO VISIT THE STARS
YAHOO! LOS ANGELES
Weekly Picks
Search
Options
Yellow Pages - People Search - City Maps -- Stock Quotes - Sports Scores

shared data.

microarrays + transcriptional regulation

1. biological questions
2. history/context
3. methods
   - “unsupervised”: cluster first, ask questions later
   - “supervised”: predicting methods

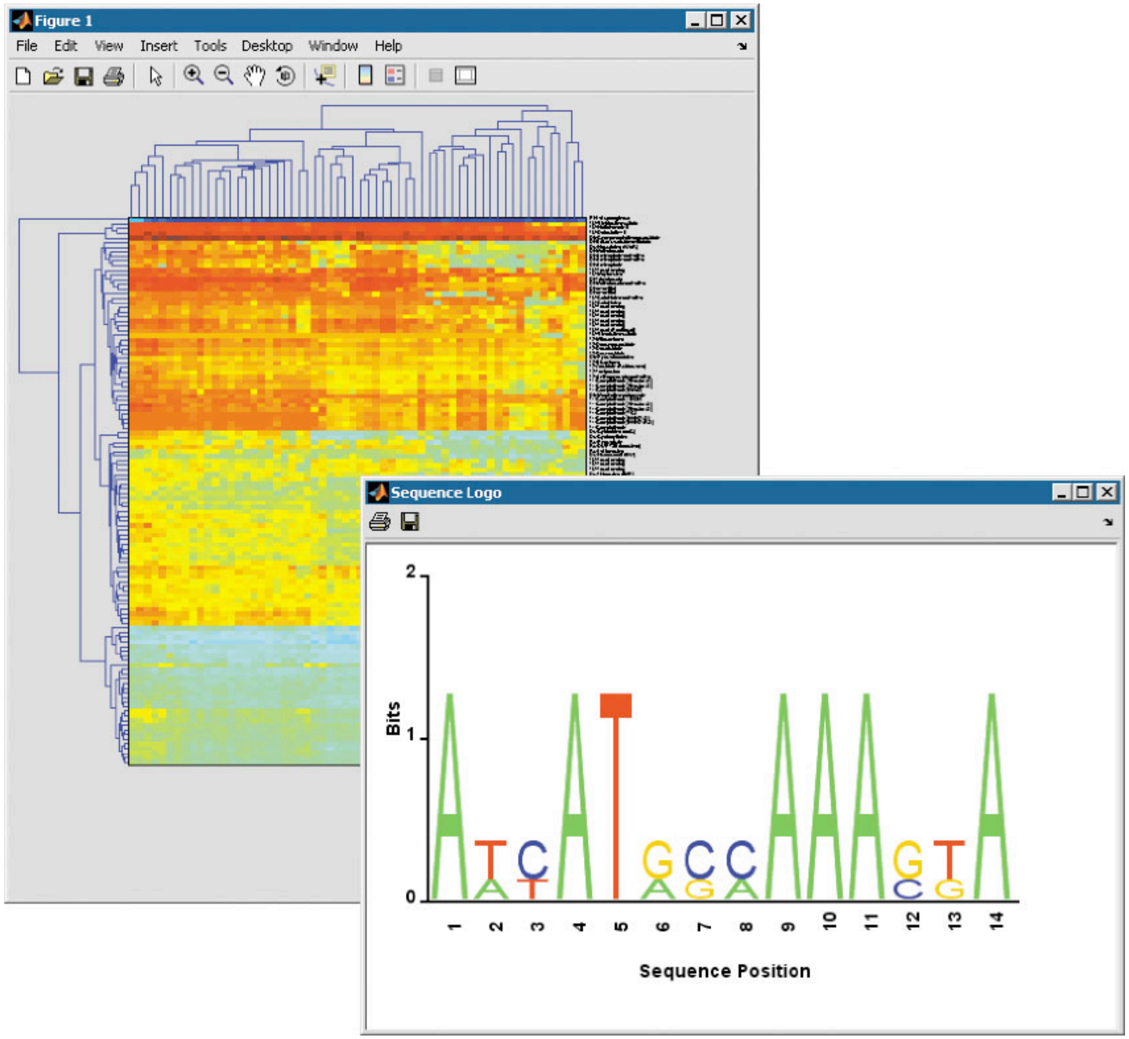
descriptive “models” of regulation:



Spellman et al., *Molecular Biology of the Cell* 1998 Dec;**9**(12):3273-97

- “unsupervised” (no input-output relation)

descriptive “models” of regulation:

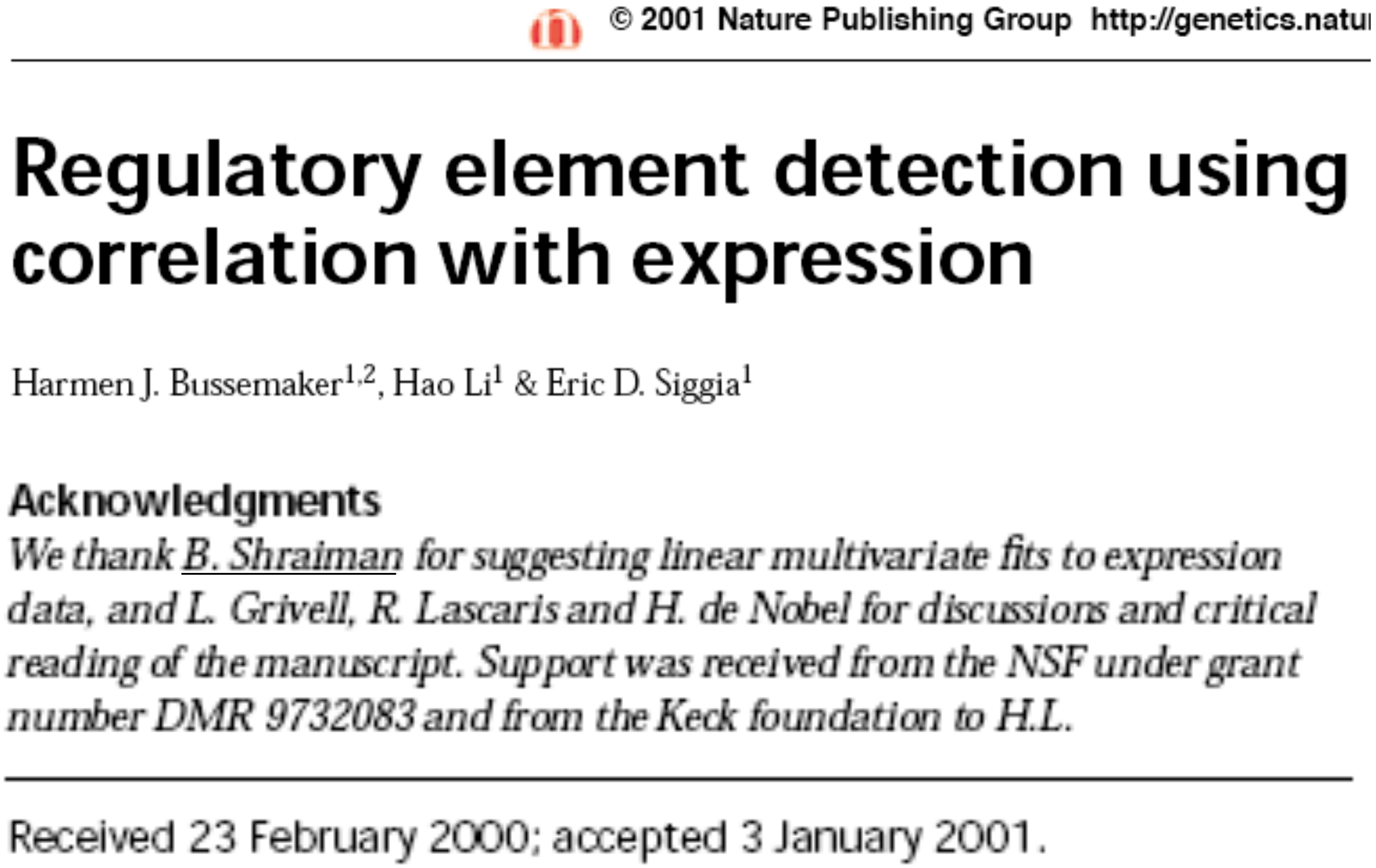


- “unsupervised” (no input-output relation)

microarrays + transcriptional regulation

1. biological questions
2. history/context
3. methods
    - “unsupervised”: cluster first, ask questions later
    - “supervised”: predicting methods

REDUCE: regression



Regulatory element detection using correlation with expression

Harmen J. Bussemaker[1,2], Hao Li[1] & Eric D. Siggia[1]

Acknowledgments

We thank B. Shraiman for suggesting linear multivariate fits to expression data, and L. Grivell, R. Lascaris and H. de Nobel for discussions and critical reading of the manuscript. Support was received from the NSF under grant number DMR 9732083 and from the Keck foundation to H.L.

Received 23 February 2000; accepted 3 January 2001.

REDUCE: why 7?

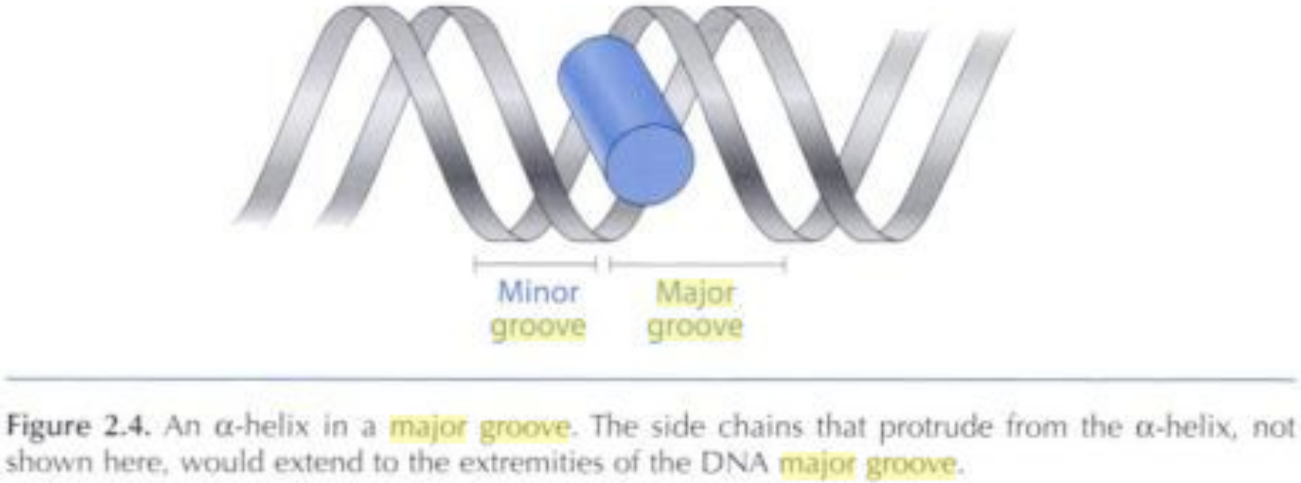
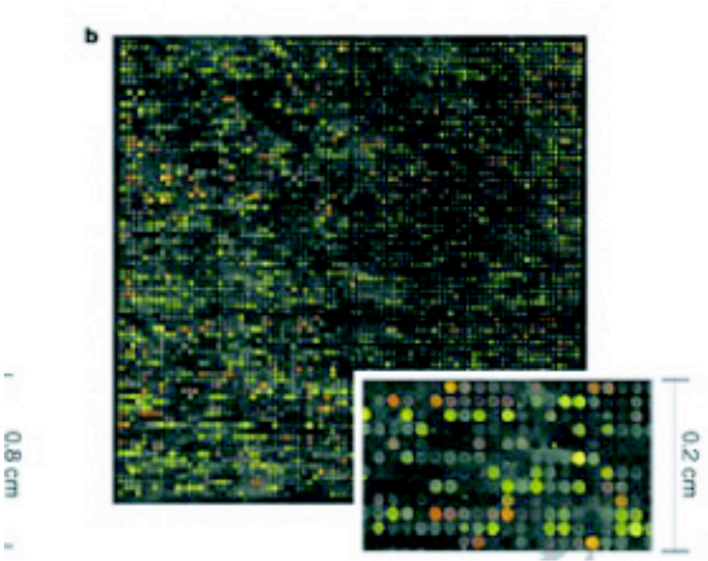
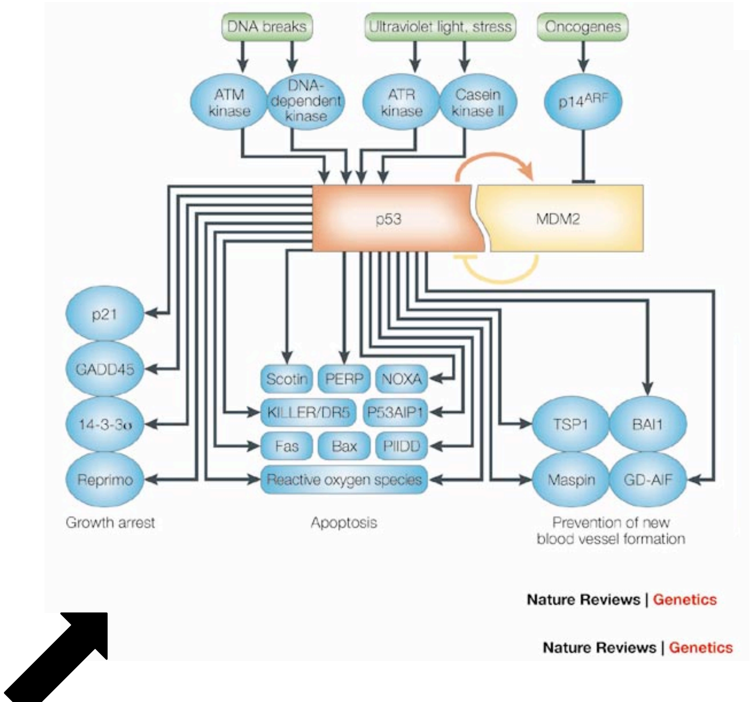


Figure 2.4. An α-helix in a major groove. The side chains that protrude from the α-helix, not shown here, would extend to the extremities of the DNA major groove.

ptashne’s “a genetic switch”

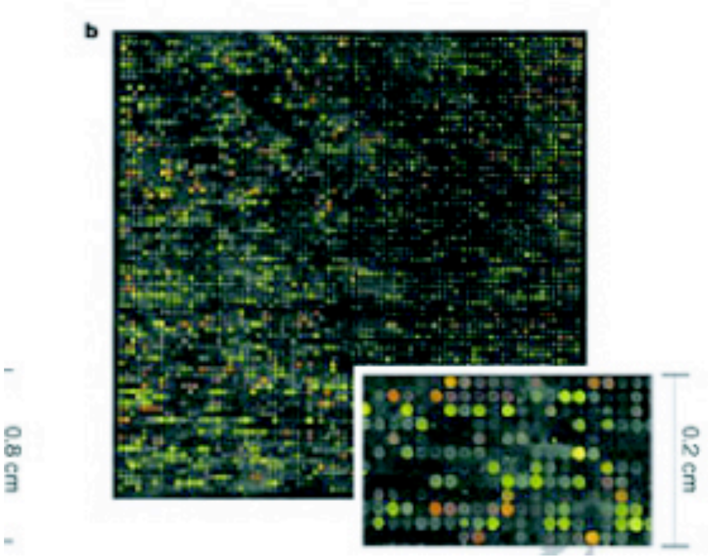
learning networks from biology

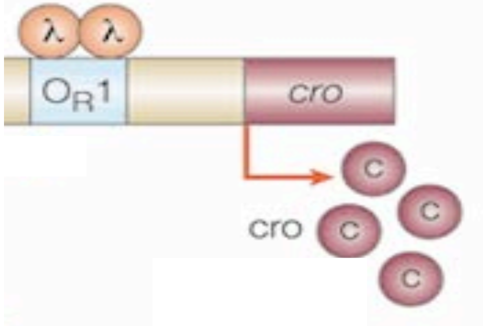




$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

“learning networks”: learn network-shaped f





$$A_g^t = f(\mu_g, \pi^t)$$

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

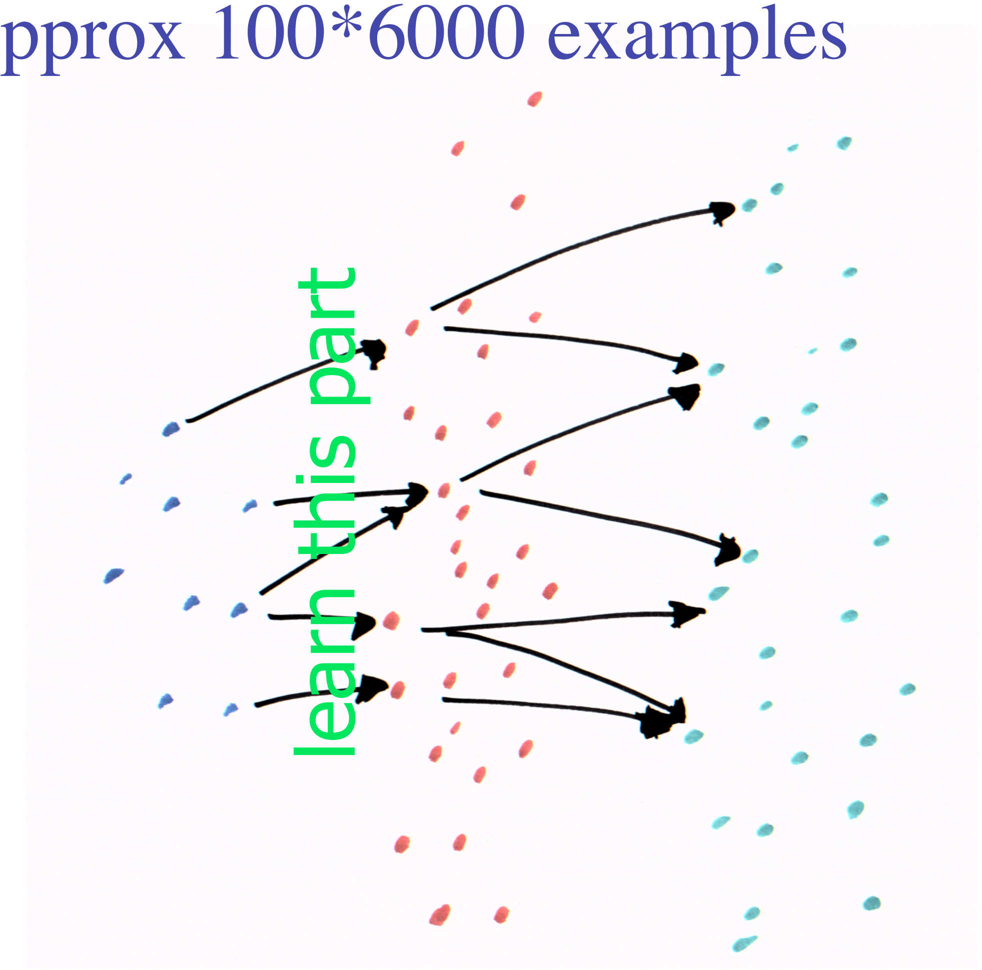
GeneClass: predict expression as class

- complex enough to learn from data
- simple enough
  - to generalize
    (predict on “held out” experiments)
  - and to be interpretable
    (based on biological rules)
- will exploit 3 tricks

trick #1: base on biological rules

parents - “motifs” - children

- 10M-dimensional feature space
- approx 100*6000 examples



trick #2: predict expression as class

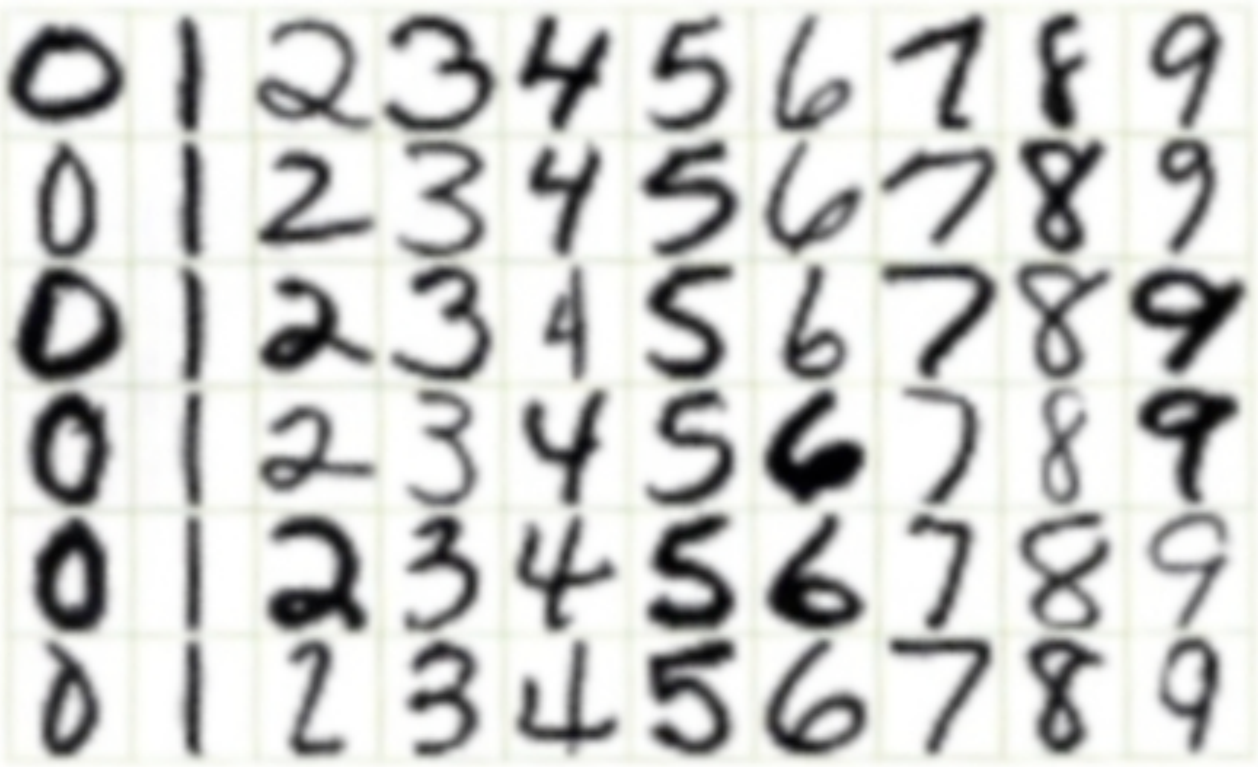
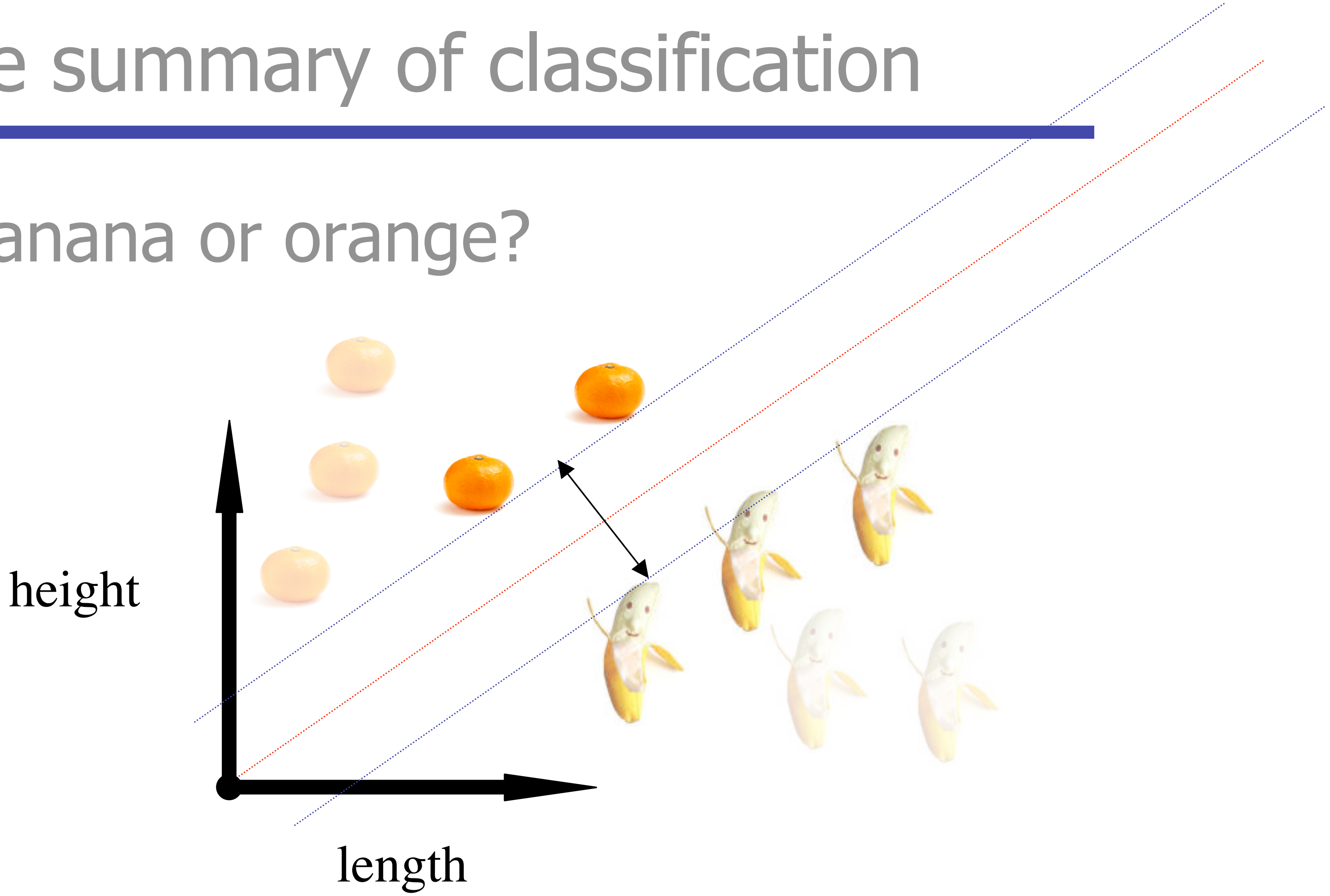


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

build a theory of 3’s?

1-slide summary of classification

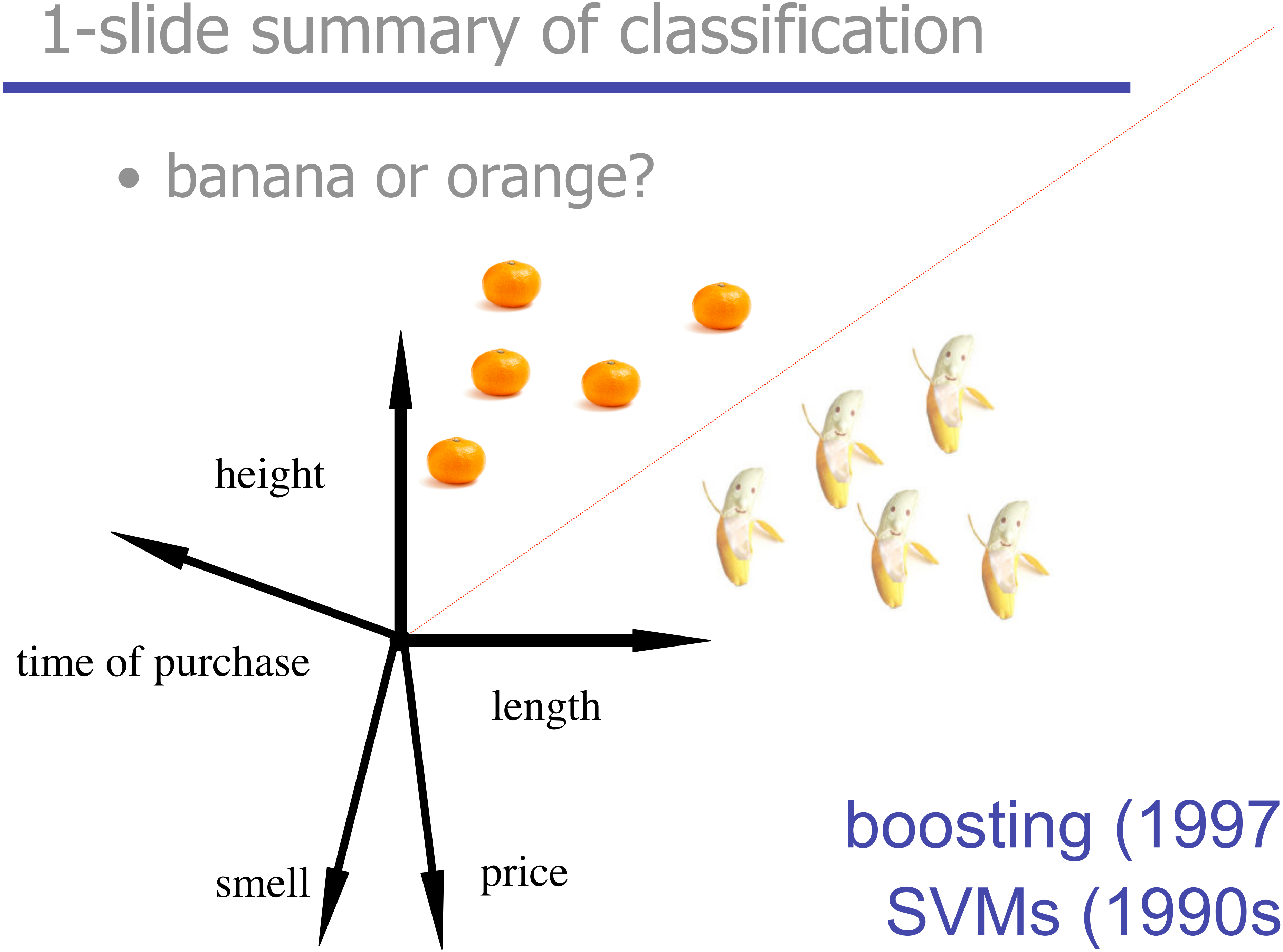
- banana or orange?



large deviation theory:
"maximum margin"

1-slide summary of classification

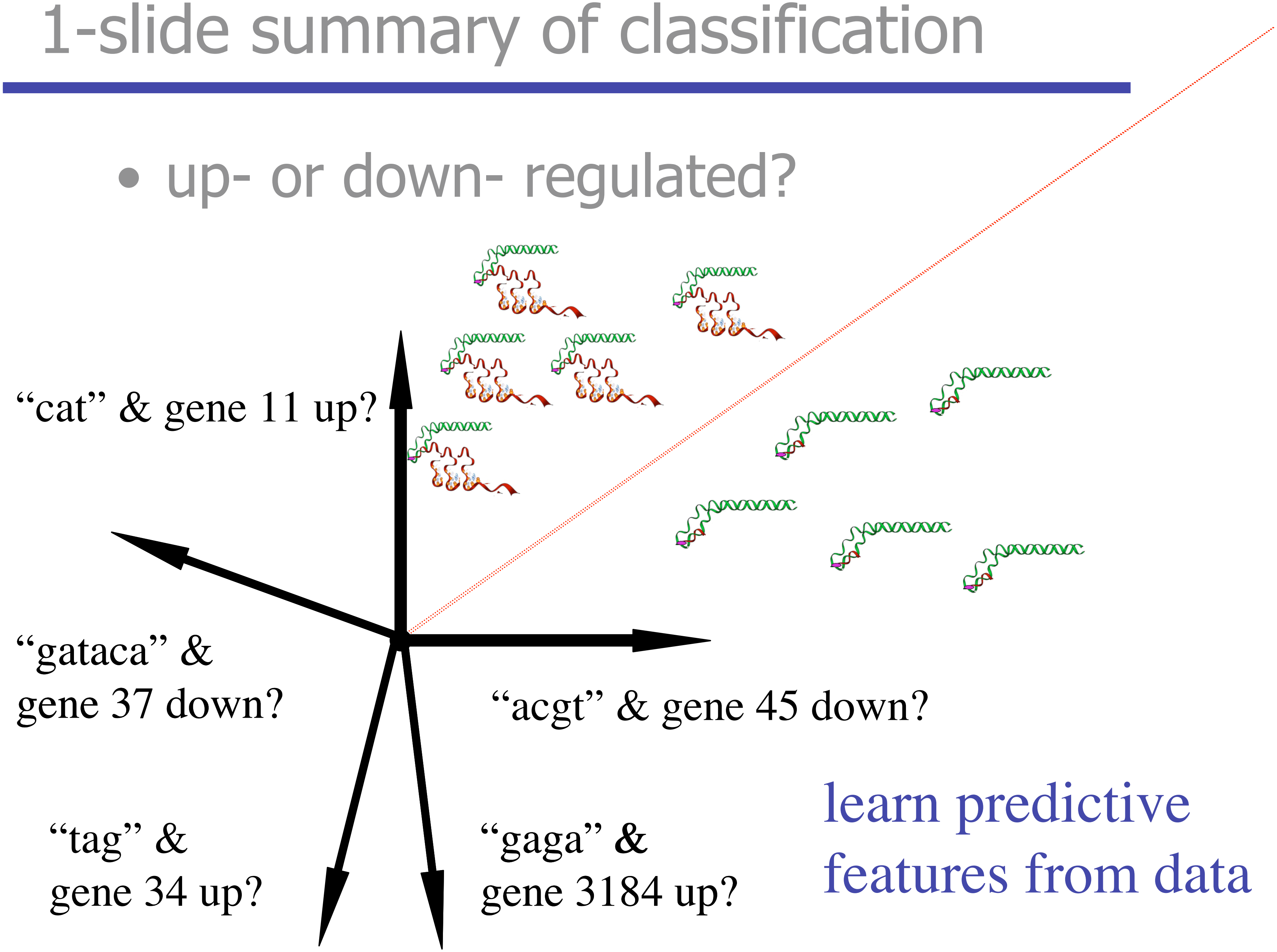
- banana or orange?



boosting (1997)
SVMs (1990s)

1-slide summary of classification

- up- or down- regulated?



“cat” & gene 11 up?

“gataca” & gene 37 down?

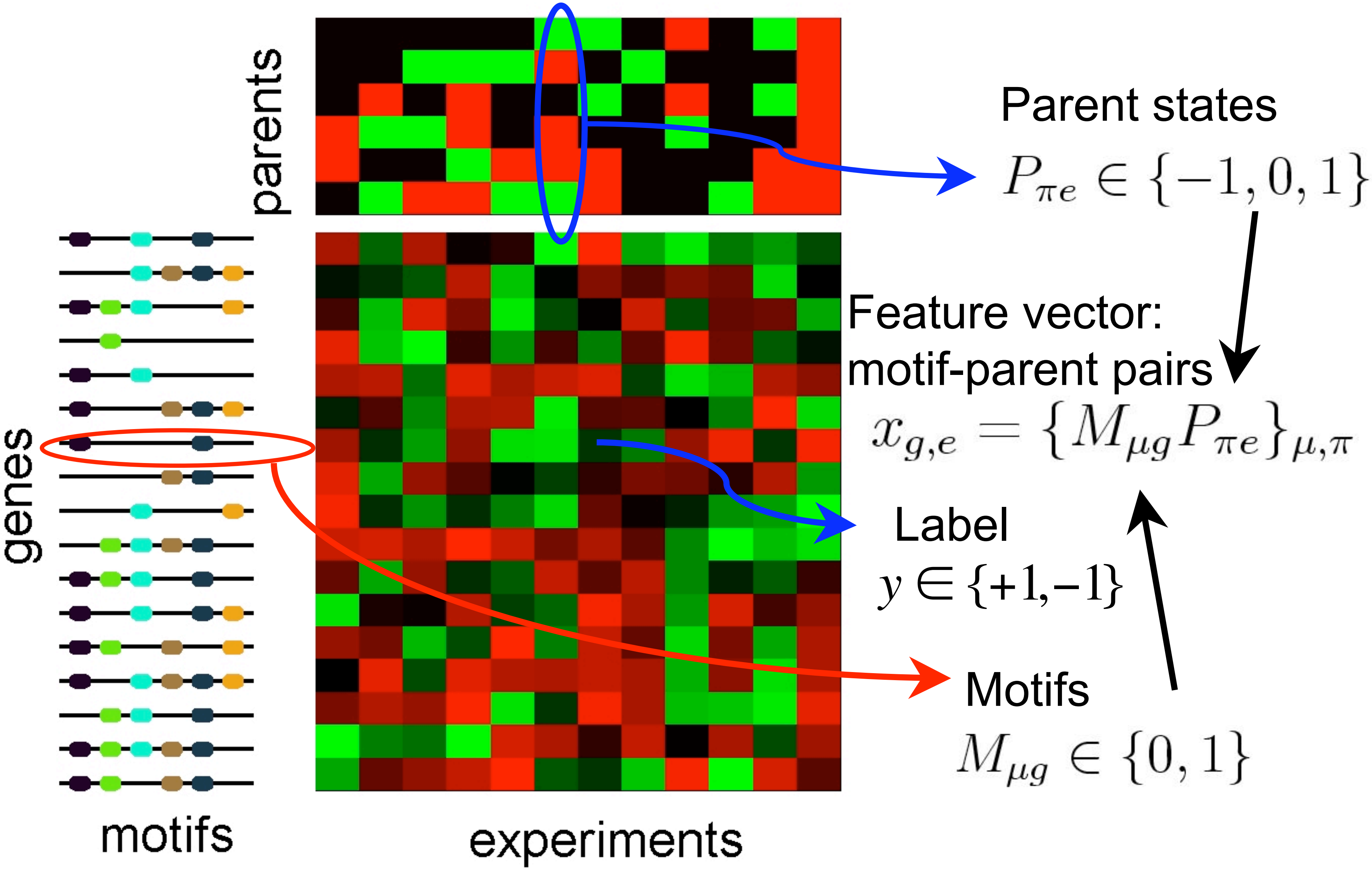
“acgt” & gene 45 down?

“tag” & gene 34 up?

“gaga” & gene 3184 up?

learn predictive features from data

model framework: $A_g^t = f(\mu_g, \pi^t)$



Parent states

$P_{\pi e} \in \{-1, 0, 1\}$

Feature vector:
motif-parent pairs

$x_{g,e} = \{M_{\mu g} P_{\pi e}\}_{\mu,\pi}$

Label

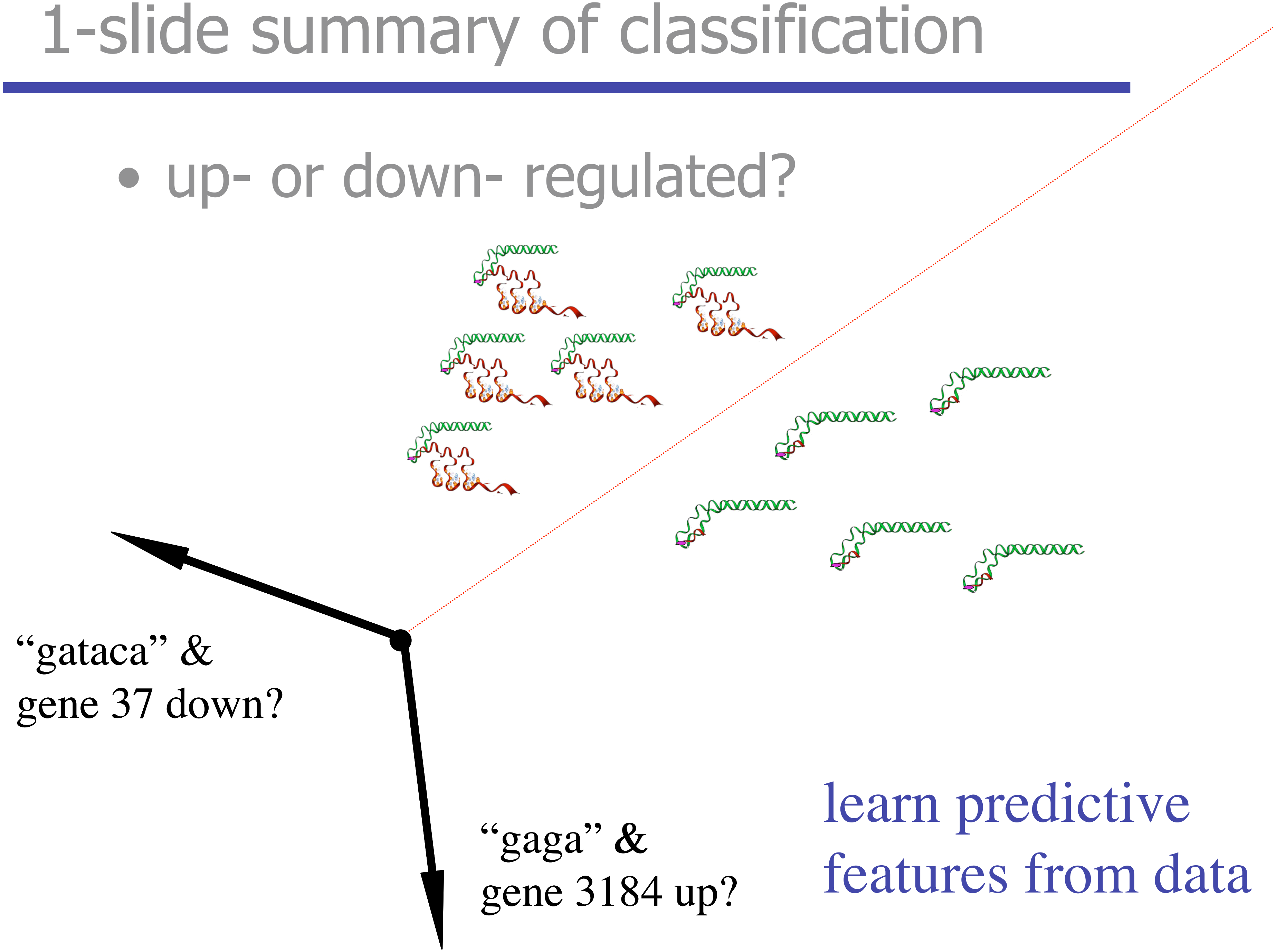
$y \in \{+1, -1\}$

Motifs

$M_{\mu g} \in \{0, 1\}$

1-slide summary of classification

- up- or down- regulated?



“gataca” &
gene 37 down?

“gaga” &
gene 3184 up?

learn predictive
features from data

"boosting"?

- Anachronistic observation:

$$\langle e^{-\sigma B(\vec{x})} \rangle \quad \text{minimized by} \quad B(\vec{x}) = \frac{1}{2} \ln \frac{p(+|\vec{x})}{p(-|\vec{x})}$$

- Therefore approximate

$$\langle e^{-\sigma B(\vec{x})} \rangle \approx Z \equiv \sum_s e^{-\sigma_s \sum_k c_k b_k(\vec{x}_s)}$$

- Coordinate descent

$$c_k \longrightarrow c_k + \alpha$$

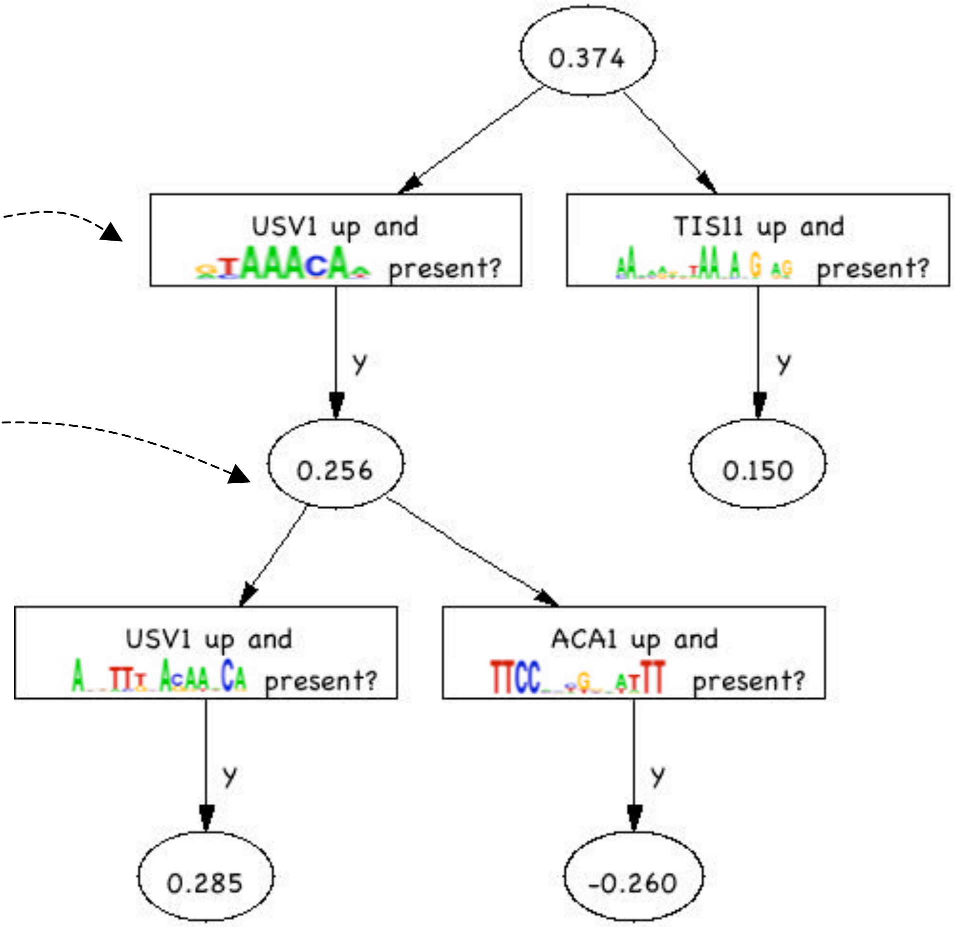
- Interpretations:
  - Add weight to hard examples
  - Greedily add 1 rule per iteration
  - learn predictive features from data.

trick #3: boosted alternating decision trees

- **One tree:** control logic all genes, all expts

1 interaction

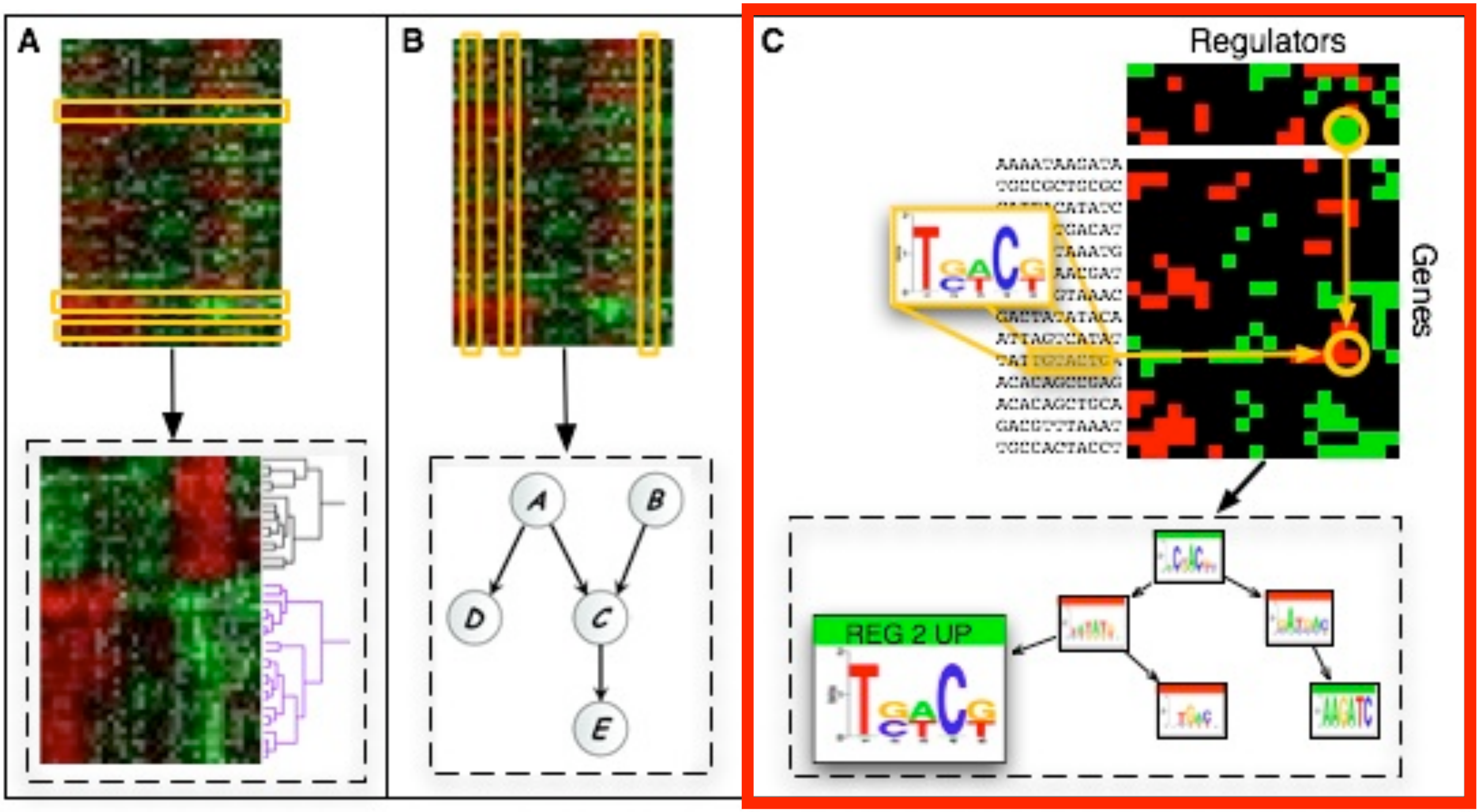
quantify regulation



- play “20 questions”
- output log(p(+)/p(-))
- highly interpretable

[ADTs: Freund & Mason 1999]

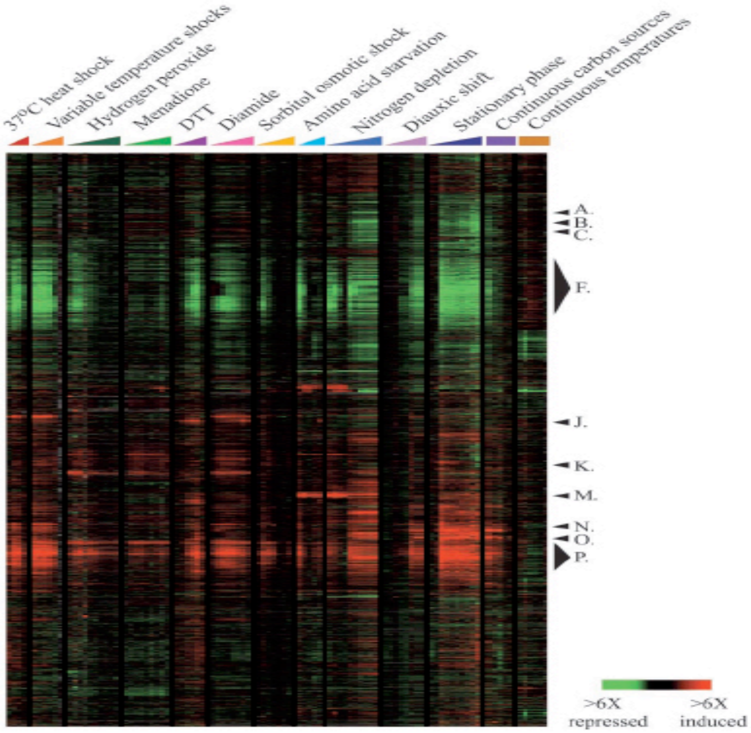
gene-centric vs. expt-centric vs. integrative

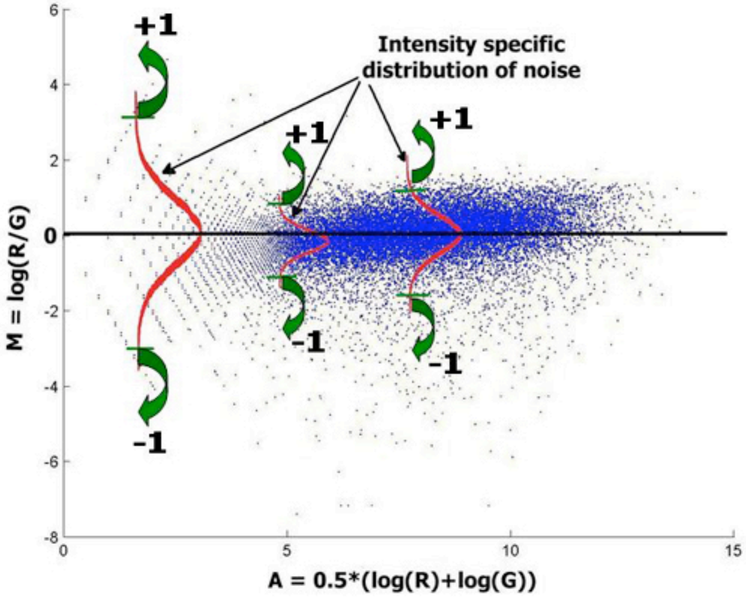


Learn *regulatory program* that makes genome-wide, context-specific predictions for differential (up/down) expression of target genes

yeast environmental stress response

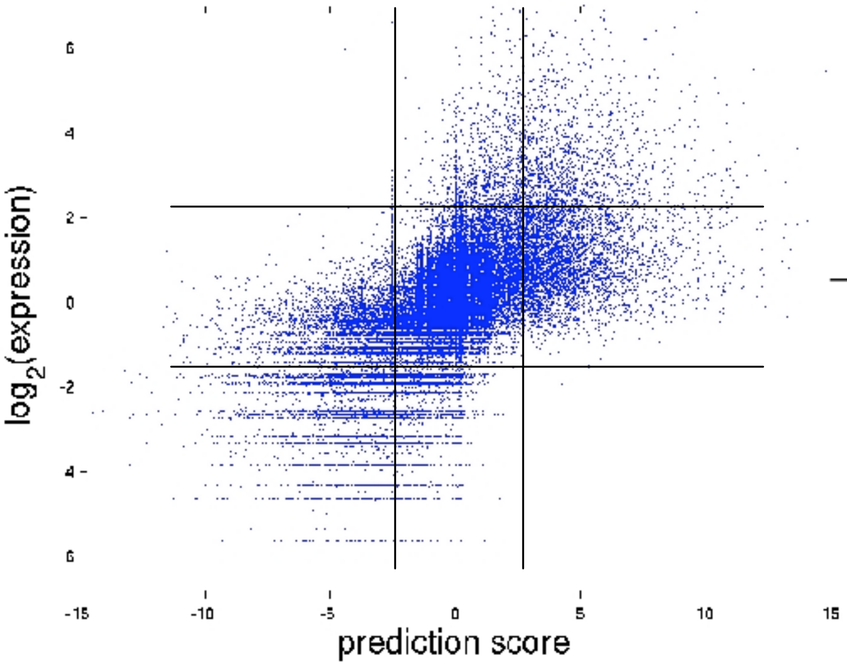
- Gasch et al. (2000) dataset, 173 microarrays, 13 environmental stresses
- *~5500 target genes, 475 regulators* (237 TF+ 250 SM)
- 500bp upstream promoter sequences
- Binning into +1/0/-1 expression levels based on wildtype vs.

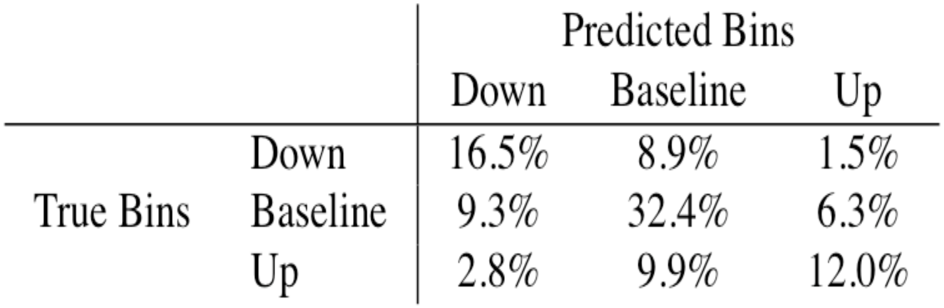




basic notions: fitting vs. overfitting

- “10-fold cross-validation” yields test loss of 13.6%



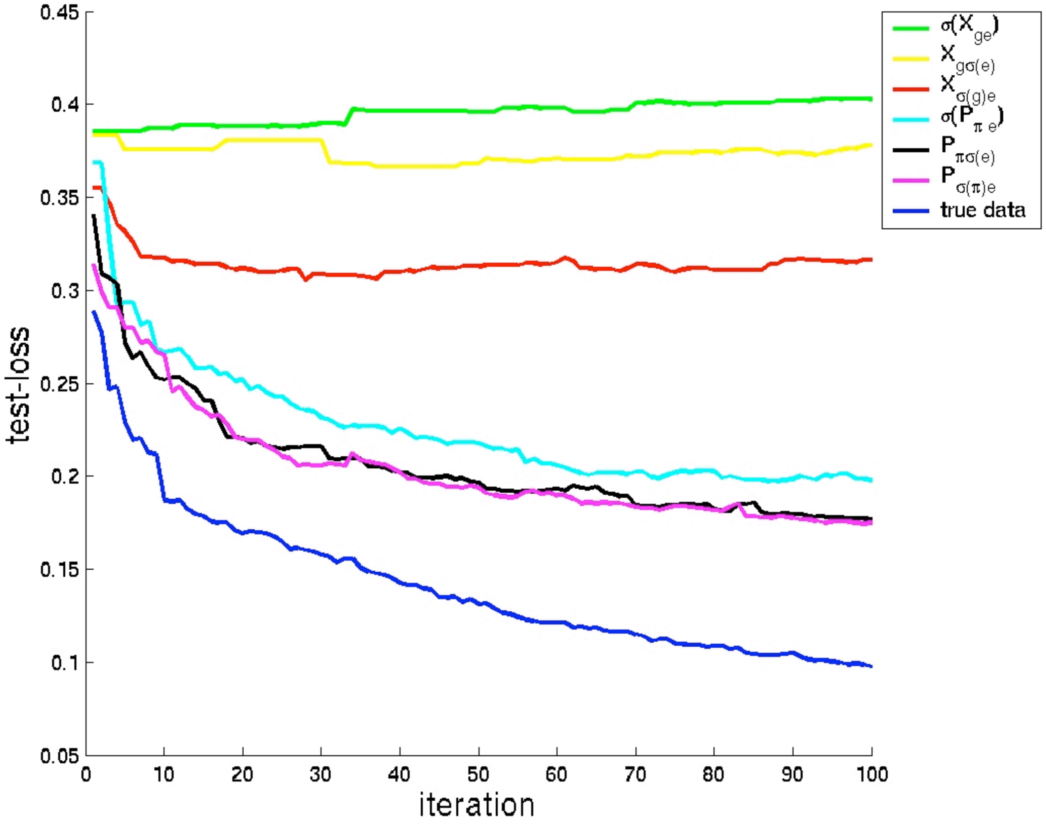


		Predicted Bins		
		Down	Baseline	Up
True Bins	Down	16.5%	8.9%	1.5%
	Baseline	9.3%	32.4%	6.3%
	Up	2.8%	9.9%	12.0%

- Empirical estimate of generalization error
- not chi squared (not training data, and not normal)

basic notions: mining vs. understanding

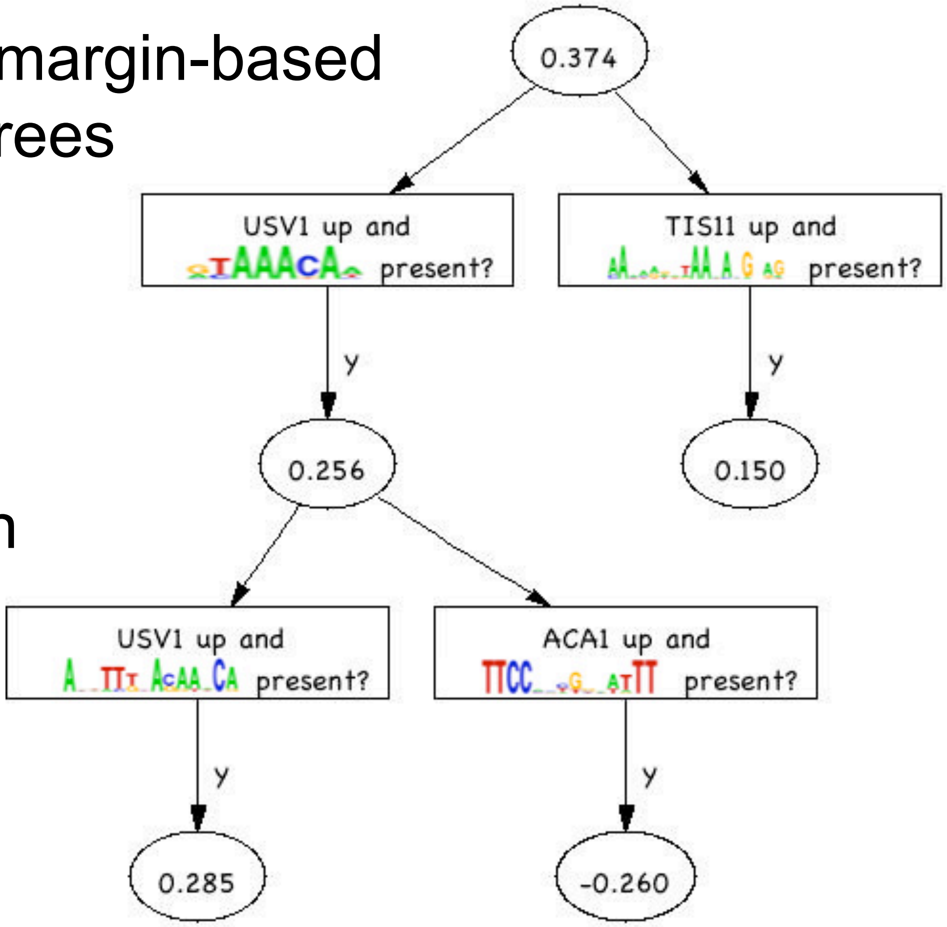
- Test Loss vs. “boosting iteration”=number of edges



- establish a baseline via randomizing

4th trick: learn predictive "f"+ motifs *ab initio*

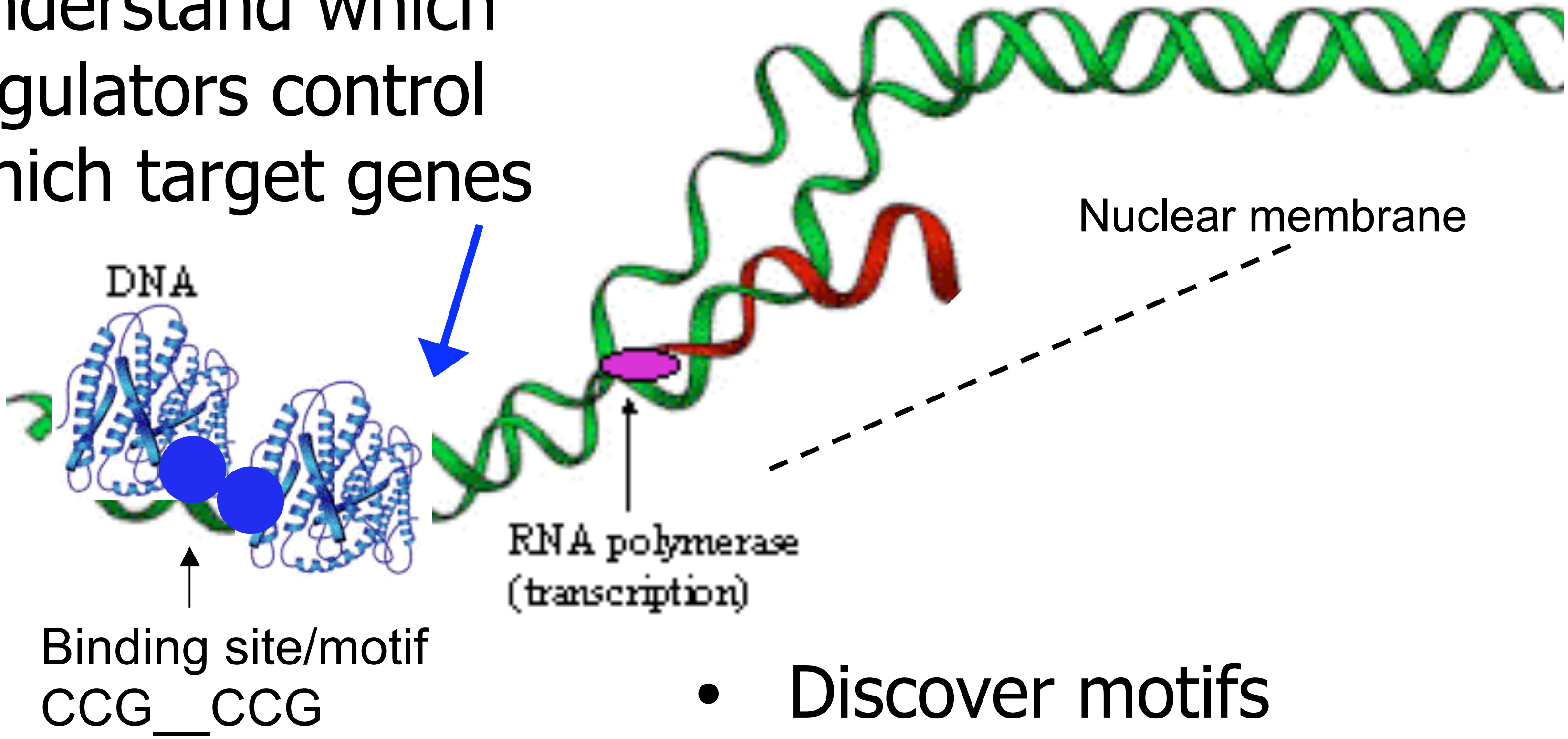
- Use *boosting* to iteratively combine predictive regulators and motifs into a tree-structure
- Alternating decision tree = margin-based generalization of decision trees
- Learn motifs ab initio from promoter sequences
- Lower nodes are conditionally dependent on higher nodes ⇒ can possibly reveal *combinatorial interactions*



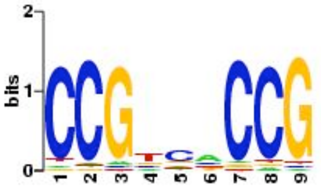
binding sites + “motif discovery”

Learning problems:

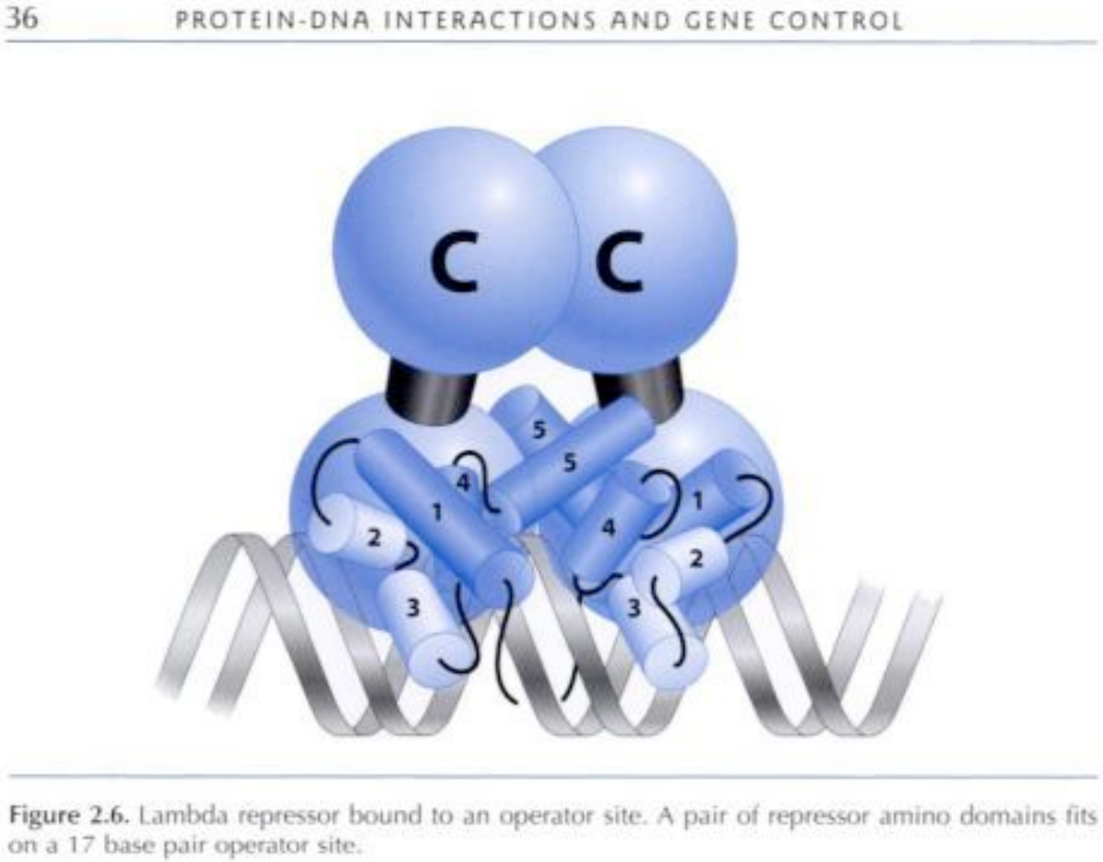
- Understand which regulators control which target genes



- Discover motifs representing regulatory elements



MEDUSA: why dimers?

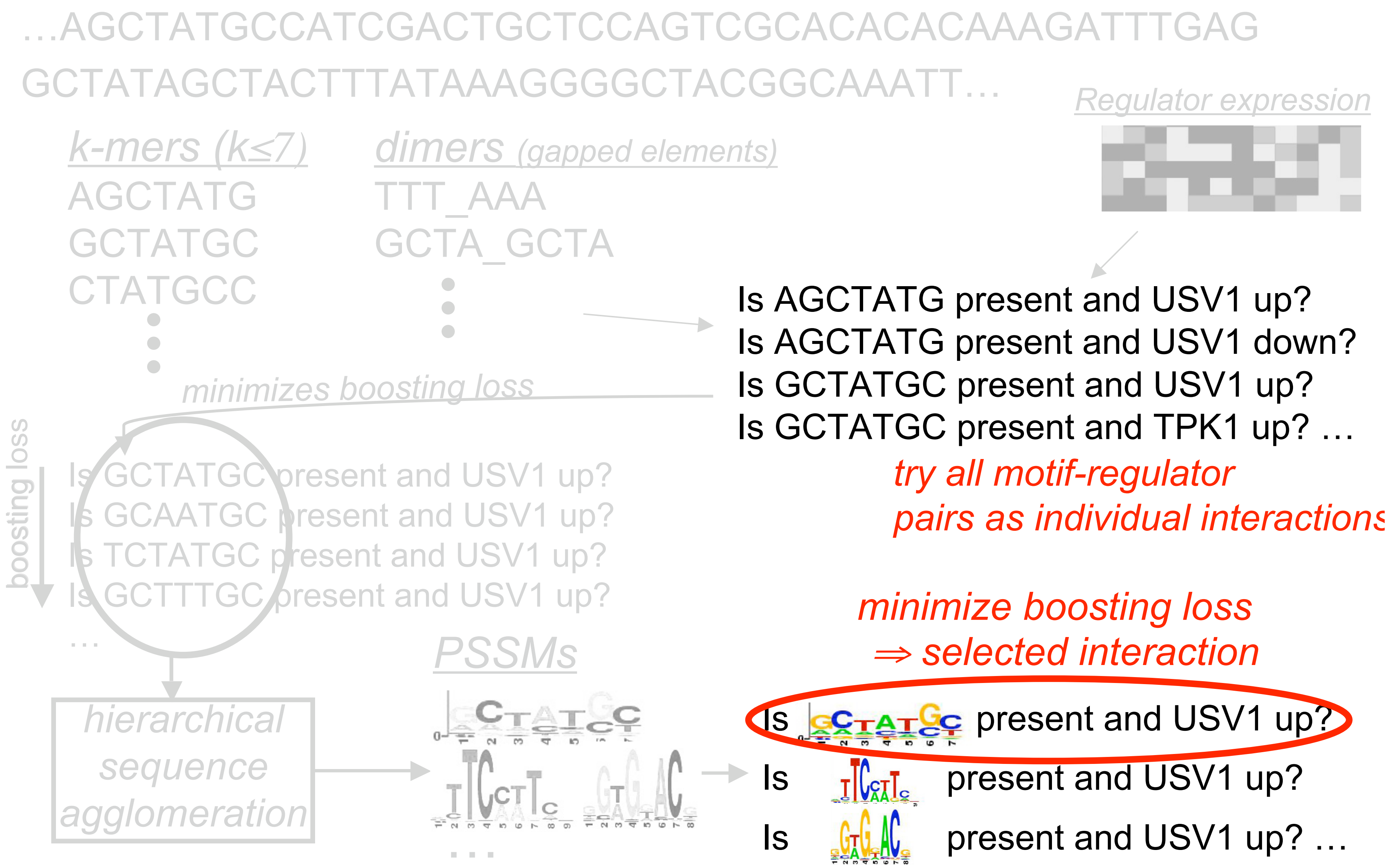


36 PROTEIN-DNA INTERACTIONS AND GENE CONTROL

Figure 2.6. Lambda repressor bound to an operator site. A pair of repressor amino domains fits on a 17 base pair operator site.

ptashne's "a genetic switch"

MEDUSA’s individual interactions



…AGCTATGCCATCGACTGCTCCAGTCGCACACACAAAGATTTGAG
GCTATAGCTACTTTATAAAGGGGCTACGGCAAATT…

Regulator expression

k-mers (k≤7)
AGCTATG
GCTATGC
CTATGCC

dimers (gapped elements)
TTT_AAA
GCTA_GCTA

Is AGCTATG present and USV1 up?
Is AGCTATG present and USV1 down?
Is GCTATGC present and USV1 up?
Is GCTATGC present and TPK1 up? …

try all motif-regulator pairs as individual interactions

minimizes boosting loss

boosting loss

Is GCTATGC present and USV1 up?
Is GCAATGC present and USV1 up?
Is TCTATGC present and USV1 up?
Is GCTTTGC present and USV1 up?
…

minimize boosting loss ⇒ selected interaction

hierarchical sequence agglomeration

PSSMs

…

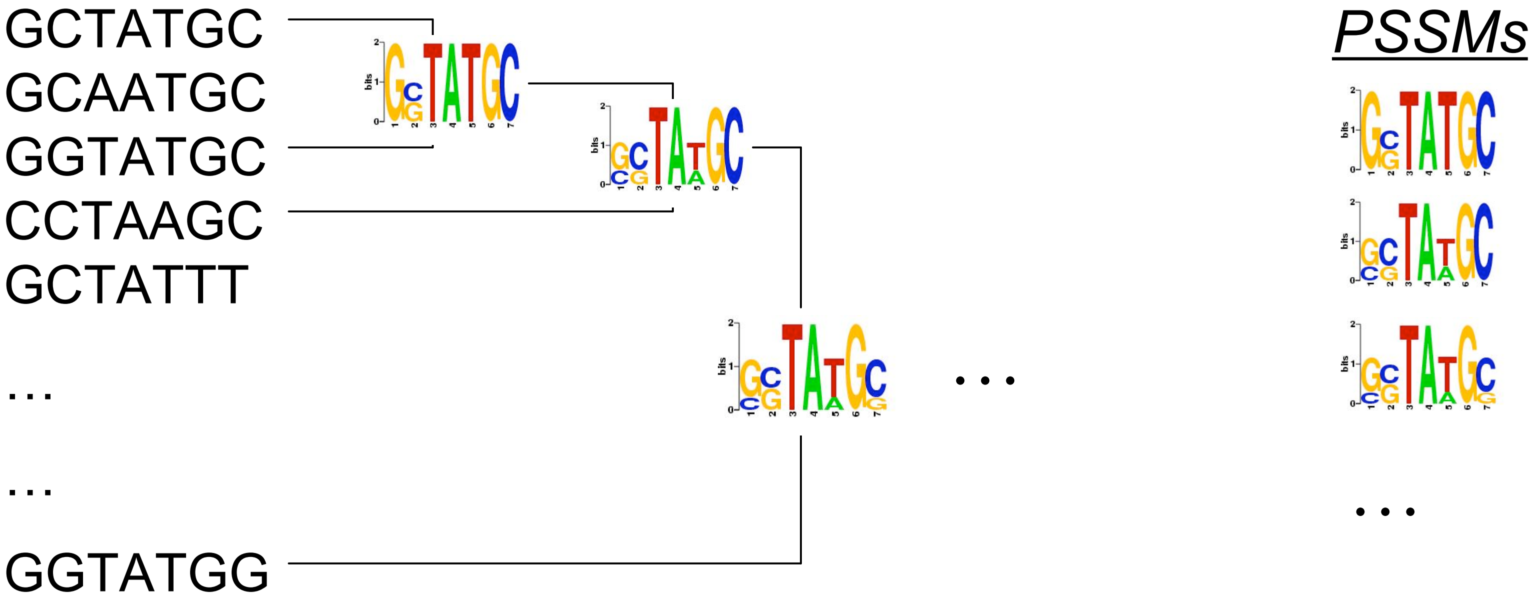
Is present and USV1 up?
Is present and USV1 up?
Is present and USV1 up? …

hierarchical sequence agglomeration

- Avoids masking of *correlated individual interactions*
- Improves prediction accuracy on test data

PSSM $p(x_1, \ldots, x_n) = \Pi_{i=1}^{n} p_i(x_i), x_i \in \{A, C, G, T\}$
score $S = \sum_{i=1}^{n} \ln(p_i(x_i)/p^{\mathrm{bg}}(x_i))$

2 PSSMs p and q
$d(p, q) \equiv \min_{\mathrm{offsets}} [w_1 D_{KL}(p||w_1 p + w_2 q) + w_2 D_{KL}(q||w_1 p + w_2 q)],$



GCTATGC
GCAATGC
GGTATGC
CCTAAGC
GCTATTT
…
…
GGTATGG

…

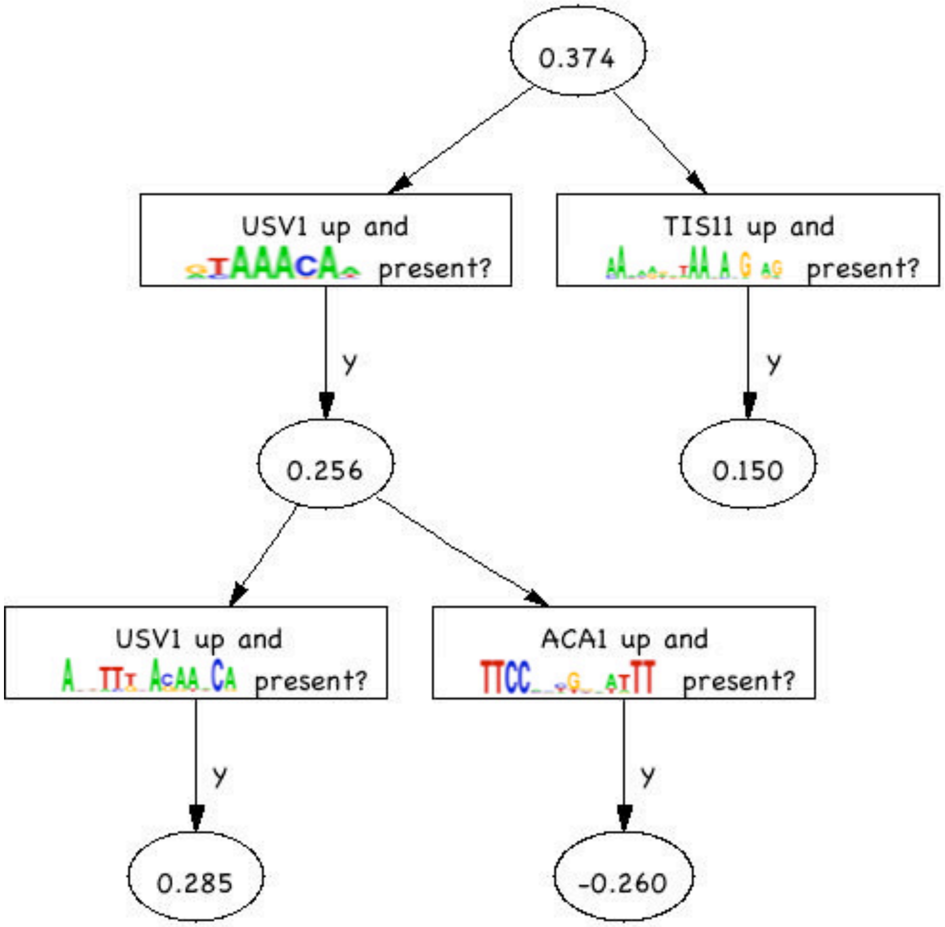
PSSMs

…

MEDUSA: summary

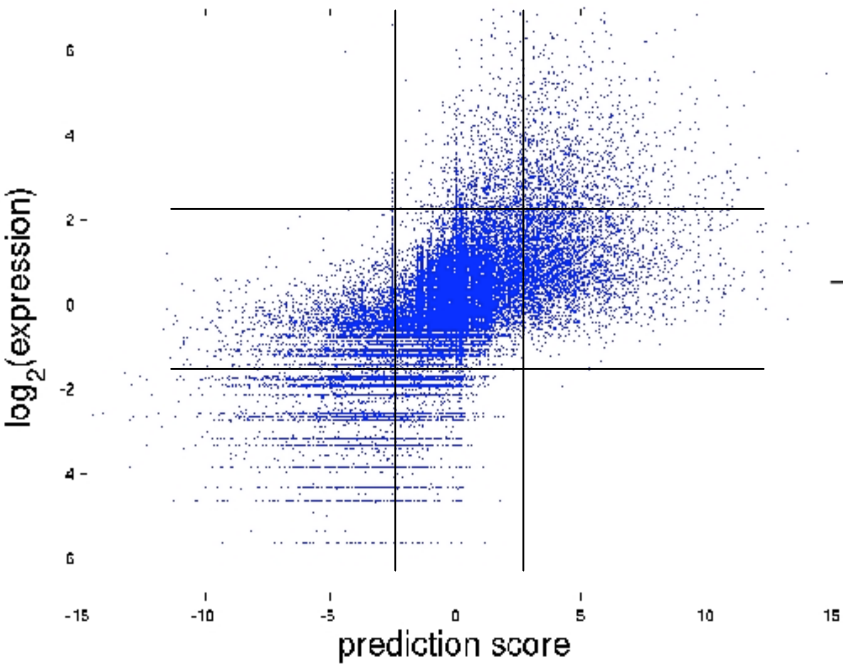
1. integrate sequence+ expression to learn a global regulatory program;
2. avoid overfitting
3. learn functional regulators-motif combos
4. learn binding site motifs, and thresholds, directly from sequence without seeding

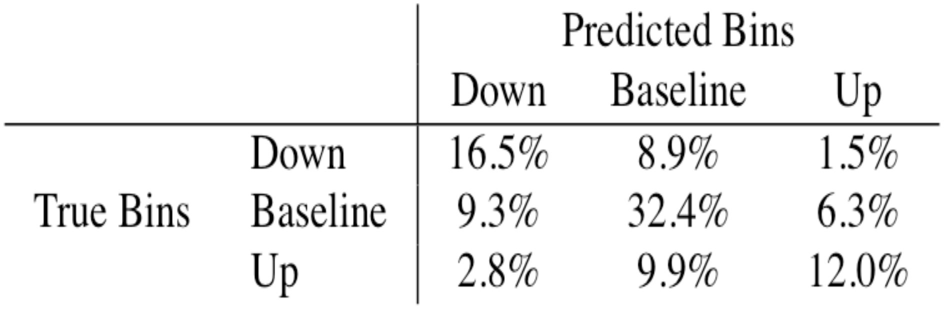
[Freund & Mason 1999]



reminder: fitting vs. overfitting

- “10-fold cross-validation” yields test loss of 13.6%





		Predicted Bins		
		Down	Baseline	Up
True Bins	Down	16.5%	8.9%	1.5%
	Baseline	9.3%	32.4%	6.3%
	Up	2.8%	9.9%	12.0%

- Empirical estimate of generalization error
- not chi squared (not training data, and not normal)

basic notions: fitting vs. overfitting

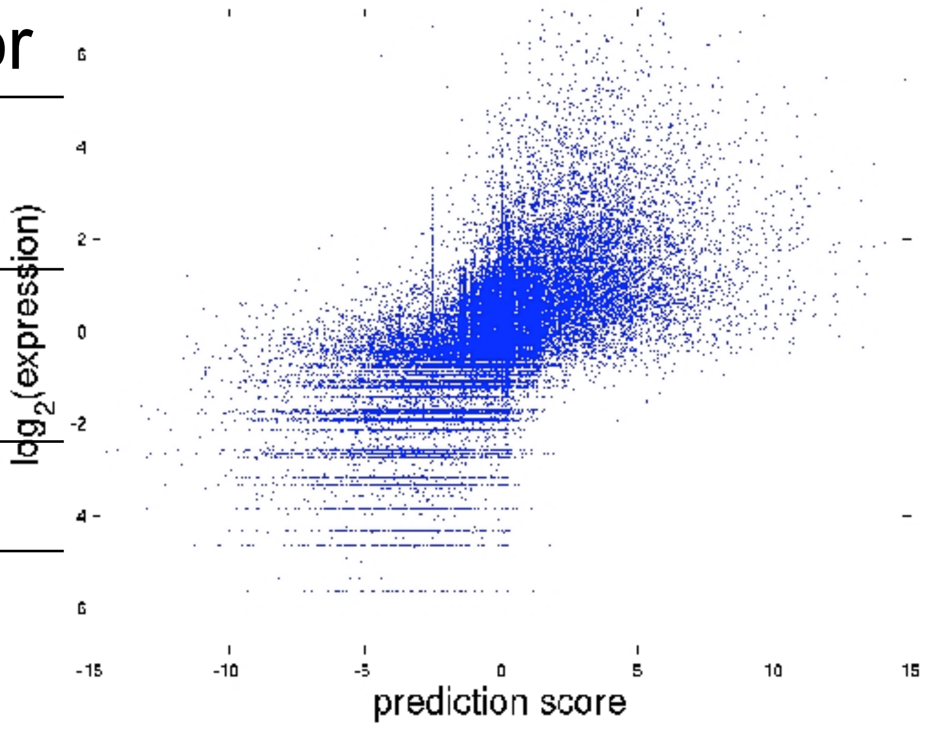
- 10-fold cross-validation (held-out experiments), ~60,000 (gene,experiment) training examples, 700 iterations
- $(N_{k\text{-mers}}+N_{dimers}+N_{PSSMs})*N_{reg}*2$ ~= 10^7 possible individual interactions at every node
- *MEDUSA*'s motifs give a *better prediction accuracy* on *held-out experiments* than database motifs

	test-loss
MEDUSA	13.4%
AlignACE (Pilpel et al. 2001)	16.1%
TRANSFAC	20.8%

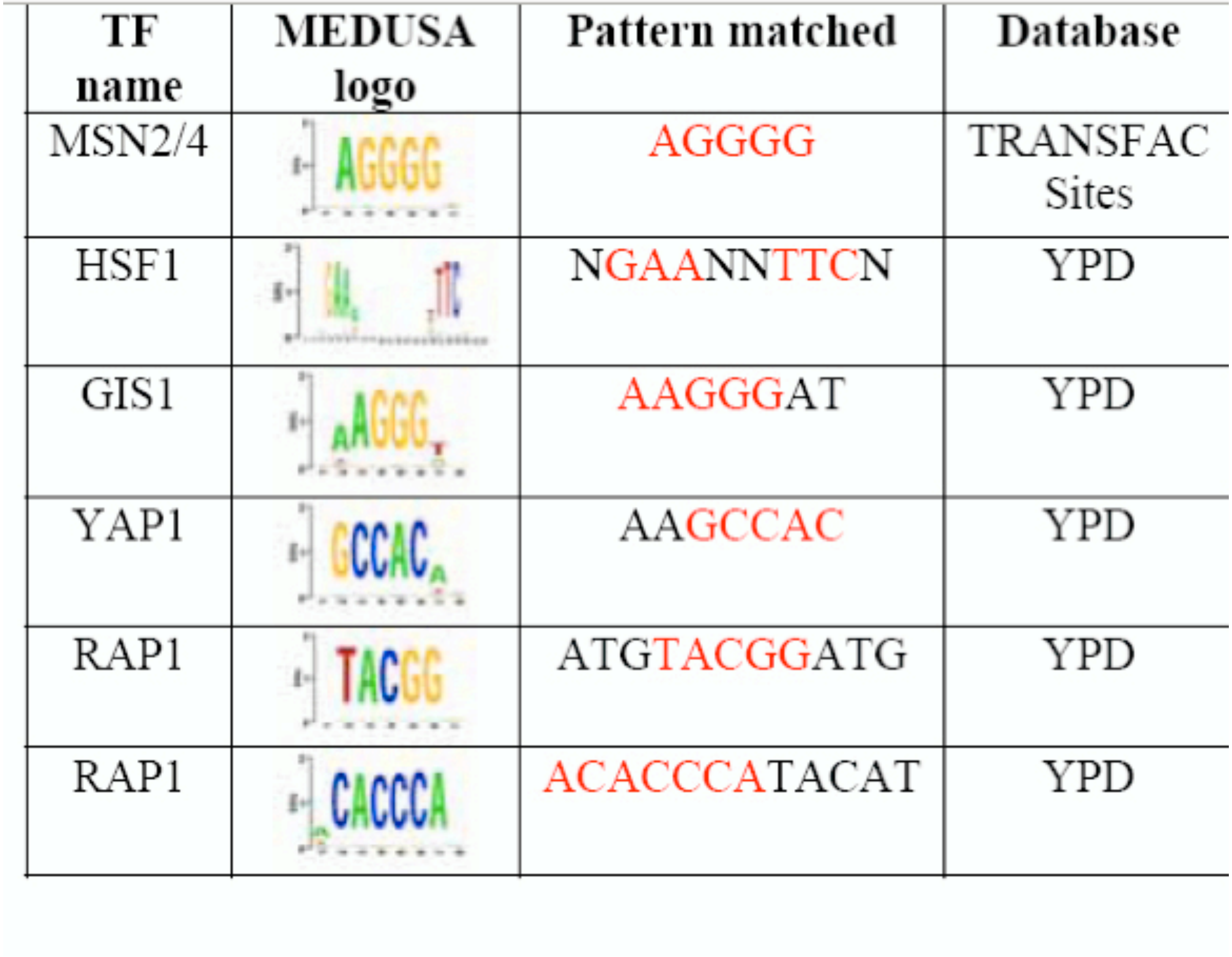
basic notions: fitting vs. overfitting

- Large-scale results: yeast ESR data set, ~170 microarrays, 5-fold cross-validation (held-out experiments), ~60,000 (gene,experiment) training examples, 700 iterations
- *MEDUSA*’s motifs give a *better prediction accuracy* on *held-out experiments* than database motifs

	test error
TRANSFAC motifs + nearest neighbor	31.3%
TRANFAC motifs + ADT	20.8%
AlignACE motifs + ADT	16.1%
MEDUSA	13.4%

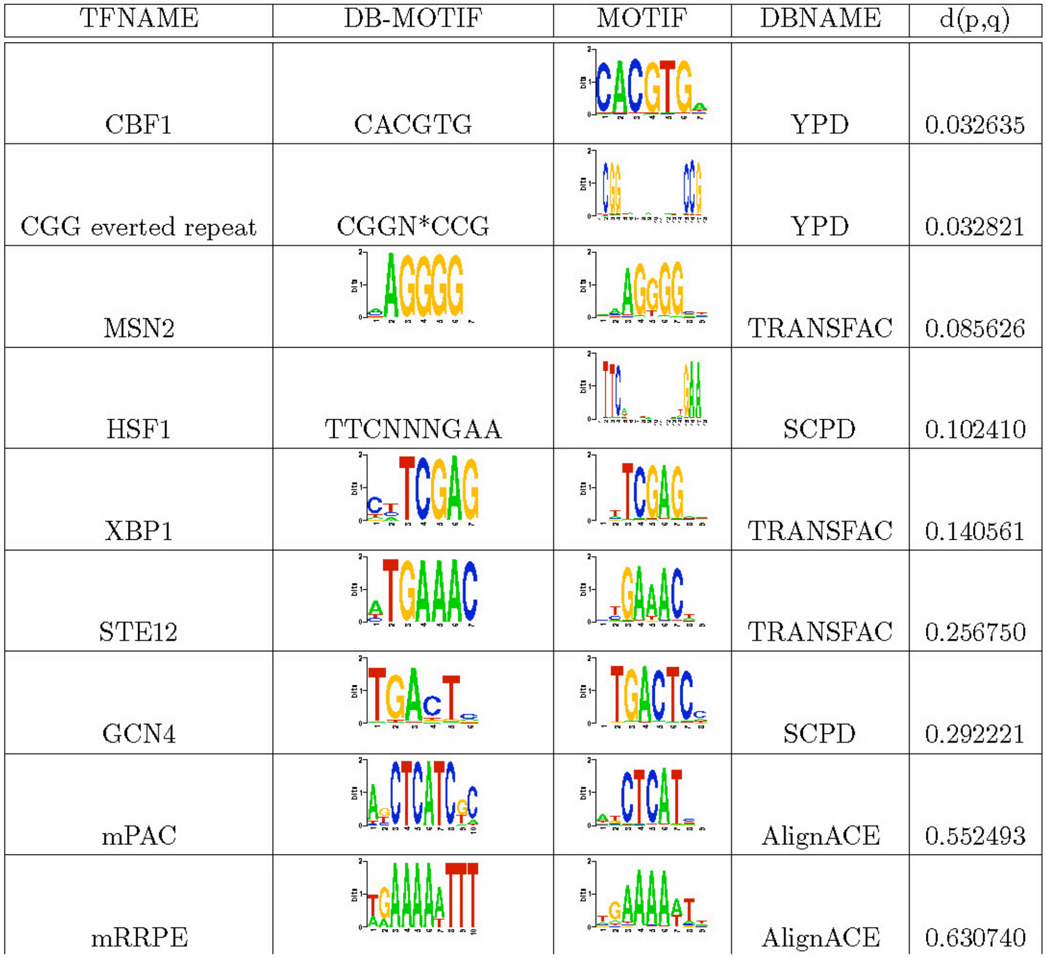


MEDUSA: ab initio PSSM discovery



TF name	MEDUSA logo	Pattern matched	Database
MSN2/4	AGGGG	AGGGG	TRANSFAC Sites
HSF1	GAA TTC	NGAANNTTCN	YPD
GIS1	AAGGGT	AAGGGAT	YPD
YAP1	GCCACA	AAGCCAC	YPD
RAP1	TACGG	ATGTACGGATG	YPD
RAP1	CACCCA	ACACCCATACAT	YPD

yeast ESR: biological validation



TFNAME	DB-MOTIF	MOTIF	DBNAME	d(p,q)
CBF1	CACGTG		YPD	0.032635
CGG everted repeat	CGGN*CCG		YPD	0.032821
MSN2			TRANSFAC	0.085626
HSF1	TTCNNNGAA		SCPD	0.102410
XBP1			TRANSFAC	0.140561
STE12			TRANSFAC	0.256750
GCN4			SCPD	0.292221
mPAC			AlignACE	0.552493
mRRPE			AlignACE	0.630740

← STRE element

← Heat shock factor

yeast ESR: biological validation

Important regulators identified by *MEDUSA*

# of weak rules	regulator
96	TPK1
64	USV1
57	AFR1
48	XBP1
19	ATG1
15	ETR1
15	SDS22
14	CIN5
12	PDR3
12	GPA2

Cellular localization of MSN2/4

Segal et al. 2003

Universal stress repressor

conclusions

- motif discovery + learning transcriptional regulation using *large-margin classification*
- learn binding sites *ab initio*
- PSSMs predictive on *test data*
- learn model of transcriptional regulation for *all genes* and *all experiments*
- simultaneous *discovery of important regulators*
- no gene clustering, no initialization
- open source:

http://www.cs.columbia.edu/compbio/medusa

agenda

- Theme: a predictive network model
  - predict expression
  - learn binding sites *ab initio*

- Breakdown: prediction? y=f(x)

- Variation: predicting evolution
  - validating models
  - letting the data decide

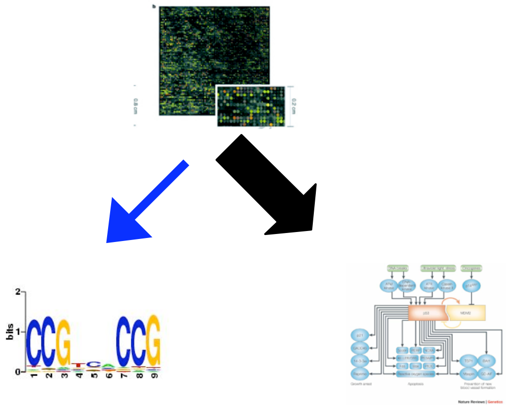
only important slide:

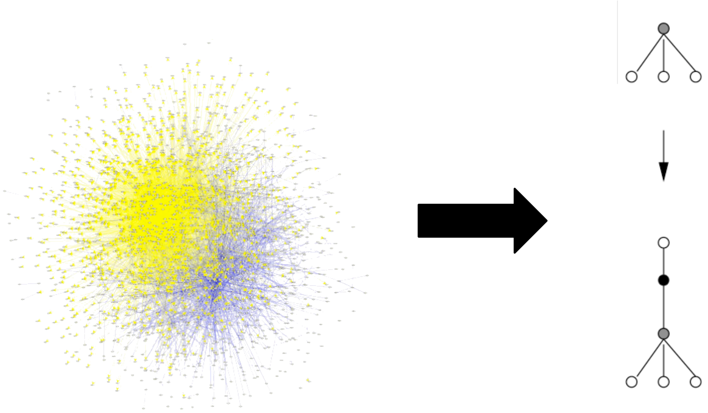
- task: learn *predictive* network from microarray and sequence data, w/o prior sequence annotation

- tool:
  y=f(x)

- task: predict evolution from topology

- tool:
  y=f(x)





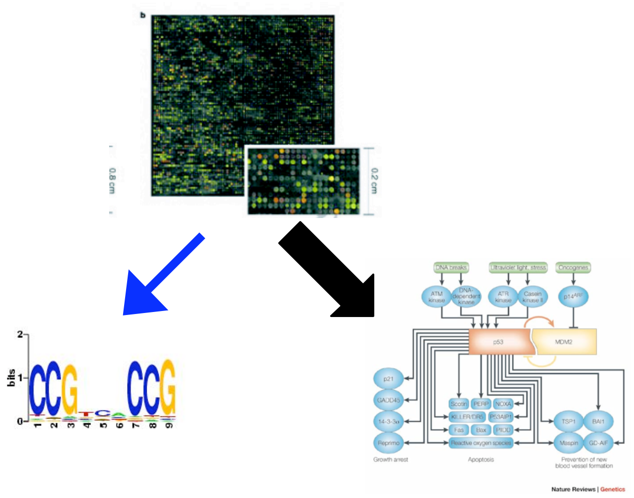
what's so great about y=f(x)???!?!?!??

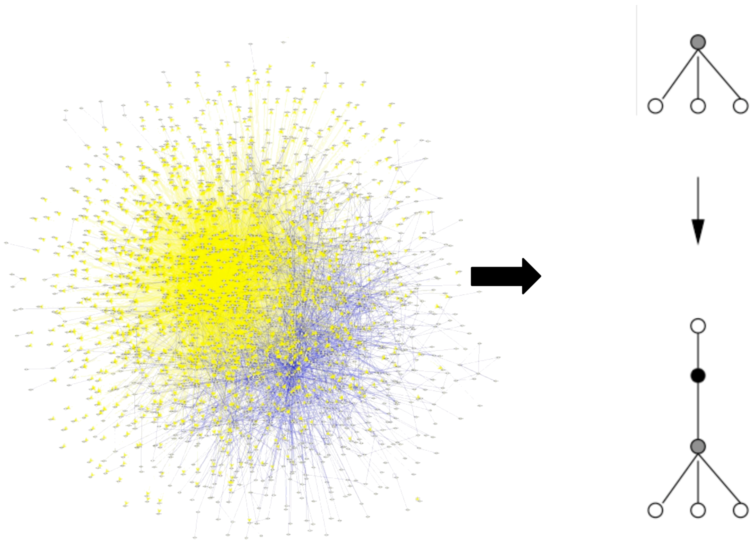
1. nothing up my sleeve:

CV: $y_V = f_L(x_V)?$

sig. $y_V = f_R(x_V)?$ $y_V = f_L(x_R)?$

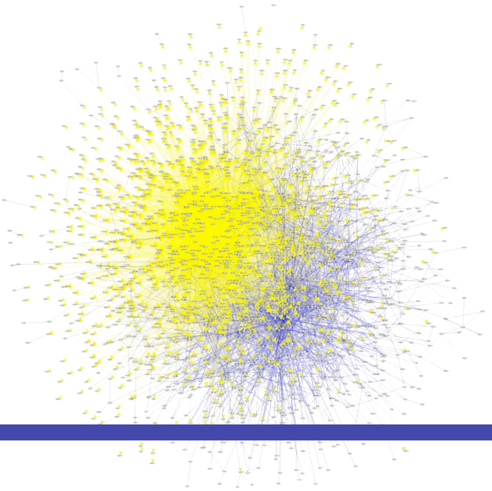
2. which x matter?





statistical network physics: definition

statistical analysis to reveal the mechanism responsible for an observed network topology

what information is in  ?

agenda

- statistical network physics
  - pseudohistory
  - the problem
- statistical learning
- biological networks

statistical network physics: pseudohistory

1999-2001:

1. measure p(k) for real networks
2. posit models/mechanisms:

   1. Erdos-Renyi $p(\omega) \sim 1$

   2. Yule/Simon/Barabasi-Albert $\dot{p}_k = f[k, p(k)]$

3. calculate $p(x) = \int_{\omega \in \Omega} d\omega p(x|\omega) p(\omega)$
4. select model which better agrees

statistical physics: cartoon

1800s-:

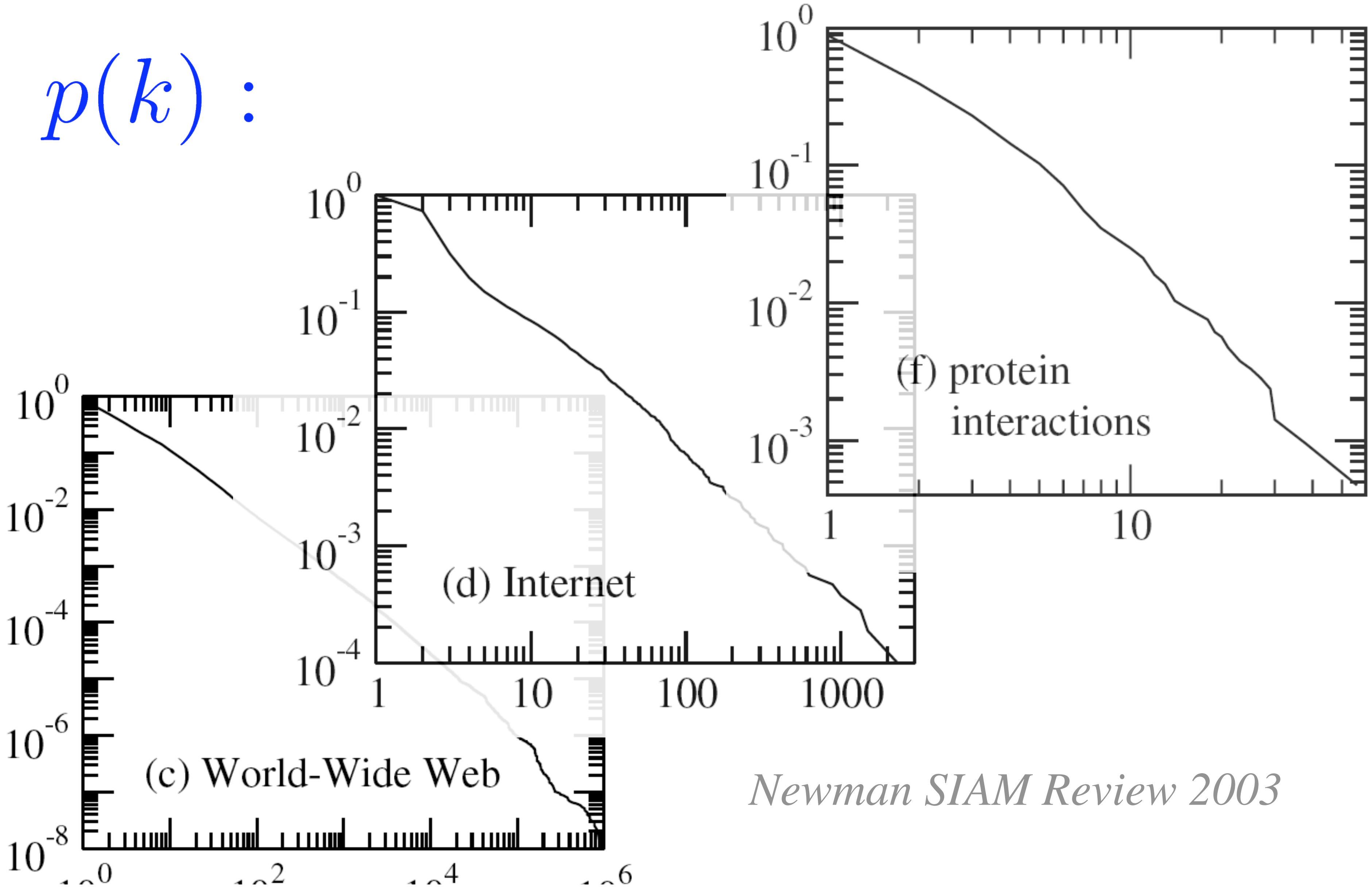
1. measure p(x) (or <x>)
2. posit models, e.g.:

$$p(\omega) \sim \mathrm{e}^{-E(\omega)/k_B T}$$

3. calculate $p(x) = \int_{\omega\in\Omega} d\omega p(x|\omega)p(\omega)$
4. select model which best agrees

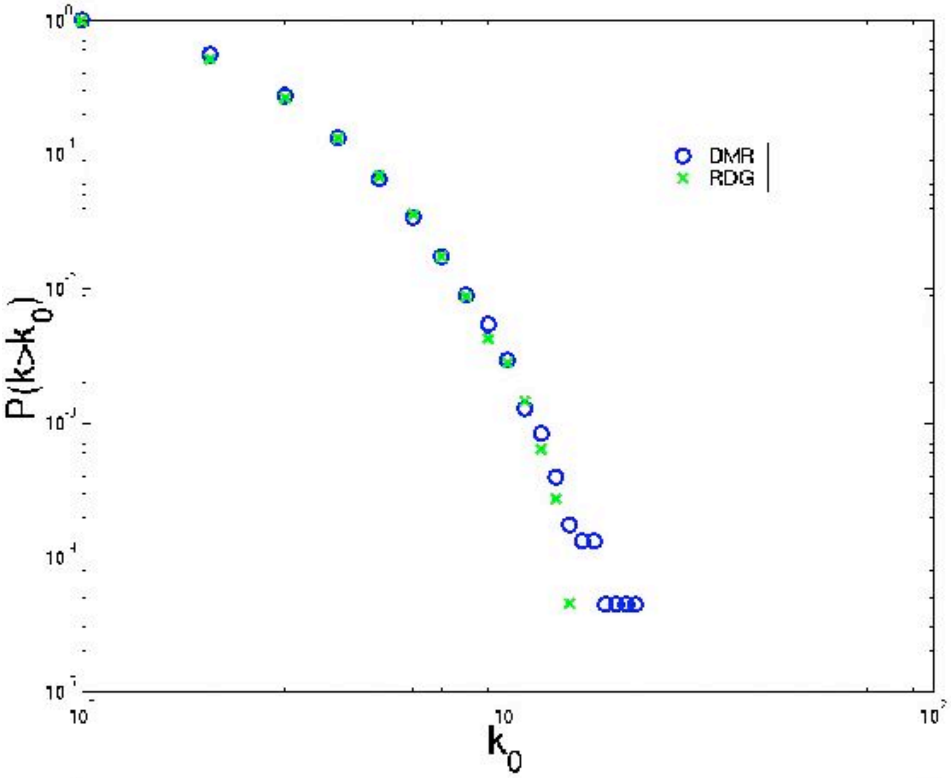
statistical network physics: measure

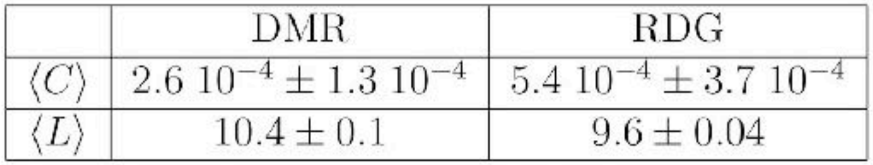
$p(k)$:



Newman SIAM Review 2003

the problem:





	DMR	RDG
$\langle C\rangle$	$2.6\ 10^{-4} \pm 1.3\ 10^{-4}$	$5.4\ 10^{-4} \pm 3.7\ 10^{-4}$
$\langle L\rangle$	10.4 ± 0.1	9.6 ± 0.04

informative statistics?

statistical network physics: history

1999-2001; 2001-2005

1. measure p(k) for real networks
2. posit models/mechanisms:
    1. Erdos-Renyi $p(\omega) \sim 1$
    2. Yule/Simon/Barabasi-Albert $\dot{p}_k = f[k, p(k)]$
3. calculate p $p(x) = \int_{\omega \in \Omega} d\omega p(x|\omega) p(\omega)$
4. mega)
5. select model which better agrees
6. overuniversality: almost all models can agree

proliferation of models (+metrics)

1. DMC
   (Vazquez, Flammini, Maritan, Vespignani,2003)
2. DMR
   (Sole, Pastor-Satorras, Smith, Kepler, 2002)
3. RDS
   (Erdos, Renyi,1959)
4. RDG
   (Callaway, Hopcroft, Kleinberg, Newman, Strogatz,2001)
5. LPA
   (Barabasi, Albert 1999)
6. AGV
   (Klemm, Eguiluz, 2002)
7. SMW
   (Watts, Strogatz 1998)

statistical network physics: problem

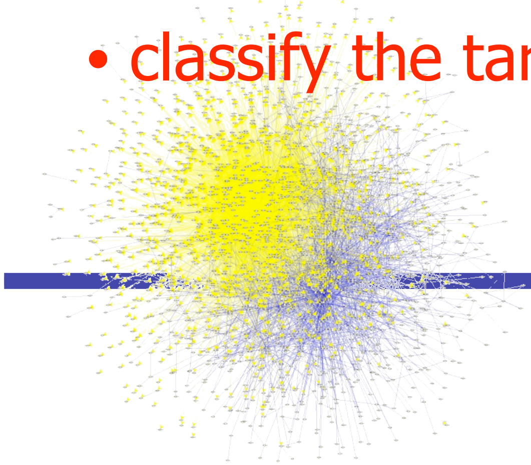
- " First, power law distributions are neither new nor rare;
- second, fitting available data to such distributions is suspiciously easy;
- third, even when the fit is robust, it adds little if anything to our knowledge of the actual architecture of the network (many different architectures can give rise to the same power laws)"

- **Revisiting "Scale-Free" Networks, E.F.Keller**

inferring design in the presence of overuniversality for a target network

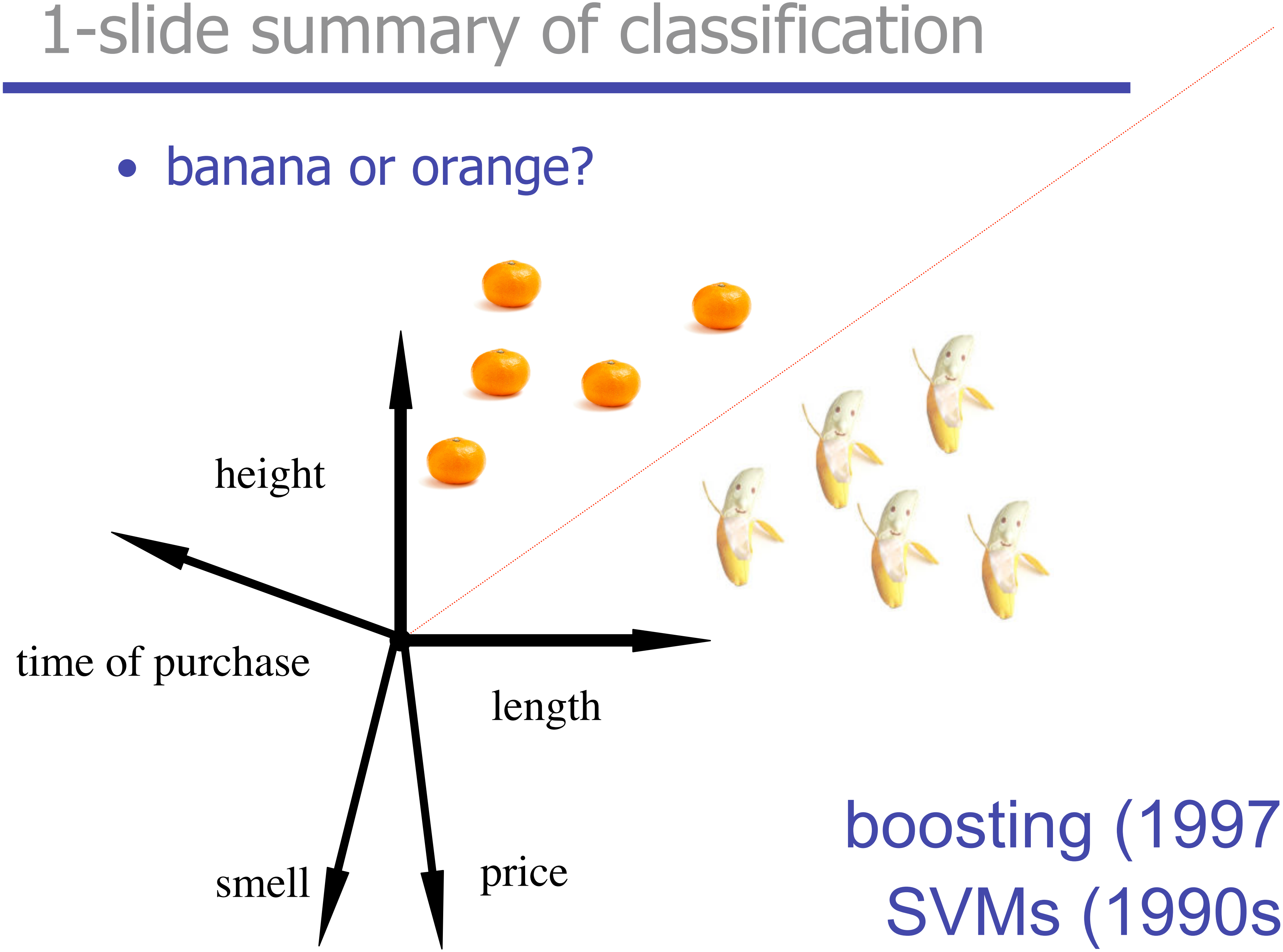
algorithm:

- forget your favorite design.
- forget your favorite feature.
- forget the target network.
- define a system for feature-generation.
- build a classifier to discriminate proposed designs.
- classify the target network.



1-slide summary of classification

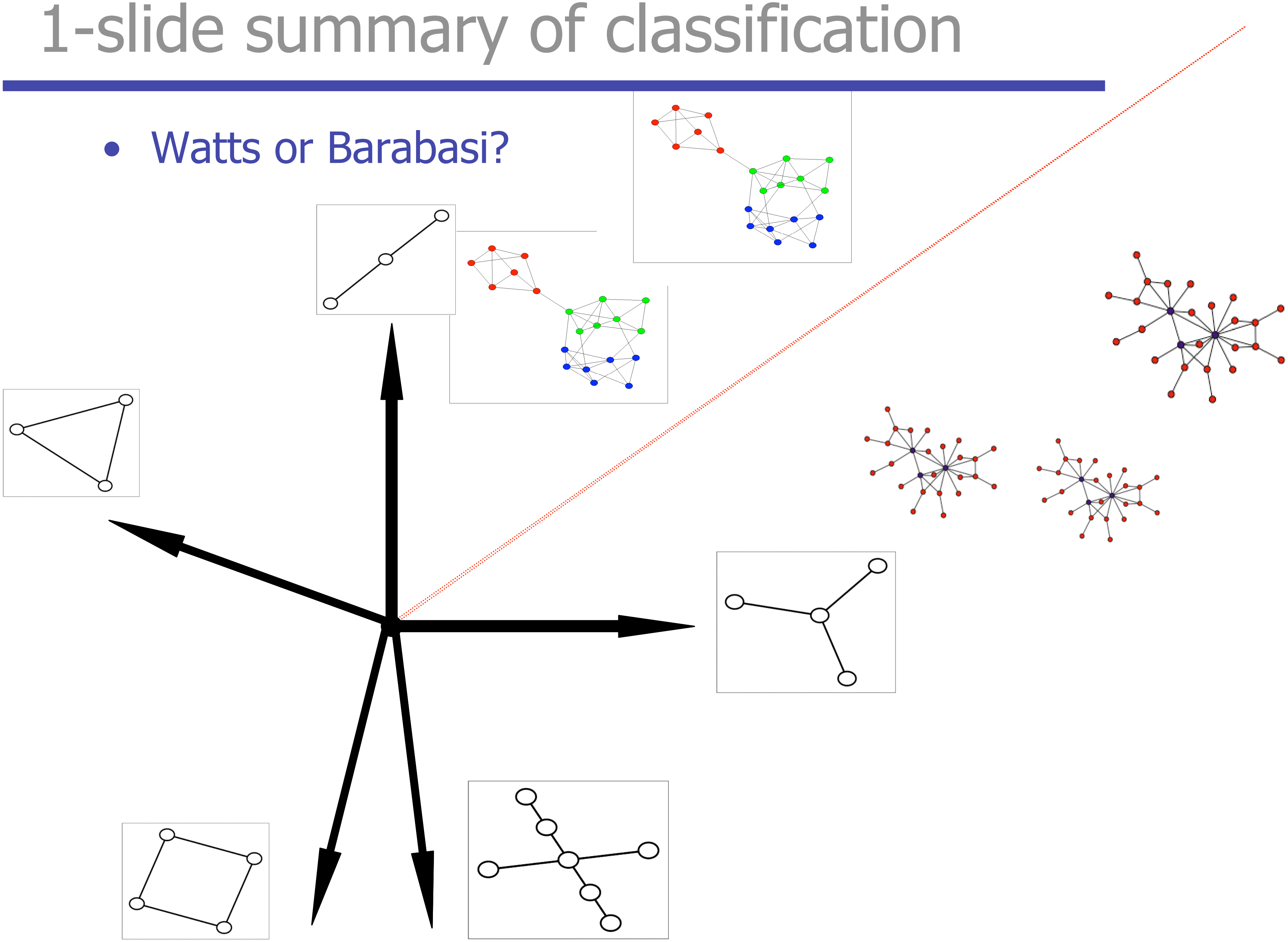
- banana or orange?



boosting (1997)
SVMs (1990s)

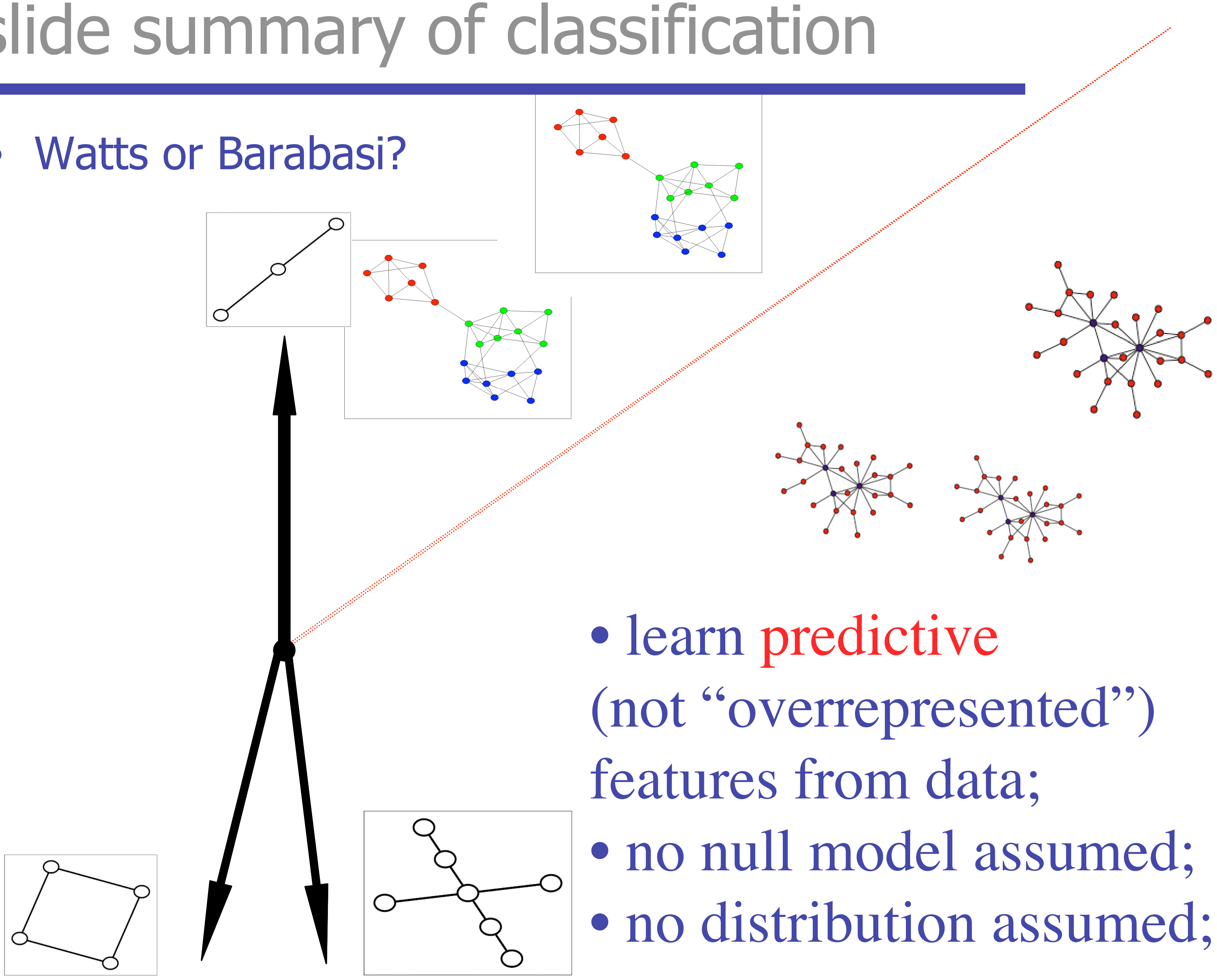
1-slide summary of classification

- Watts or Barabasi?



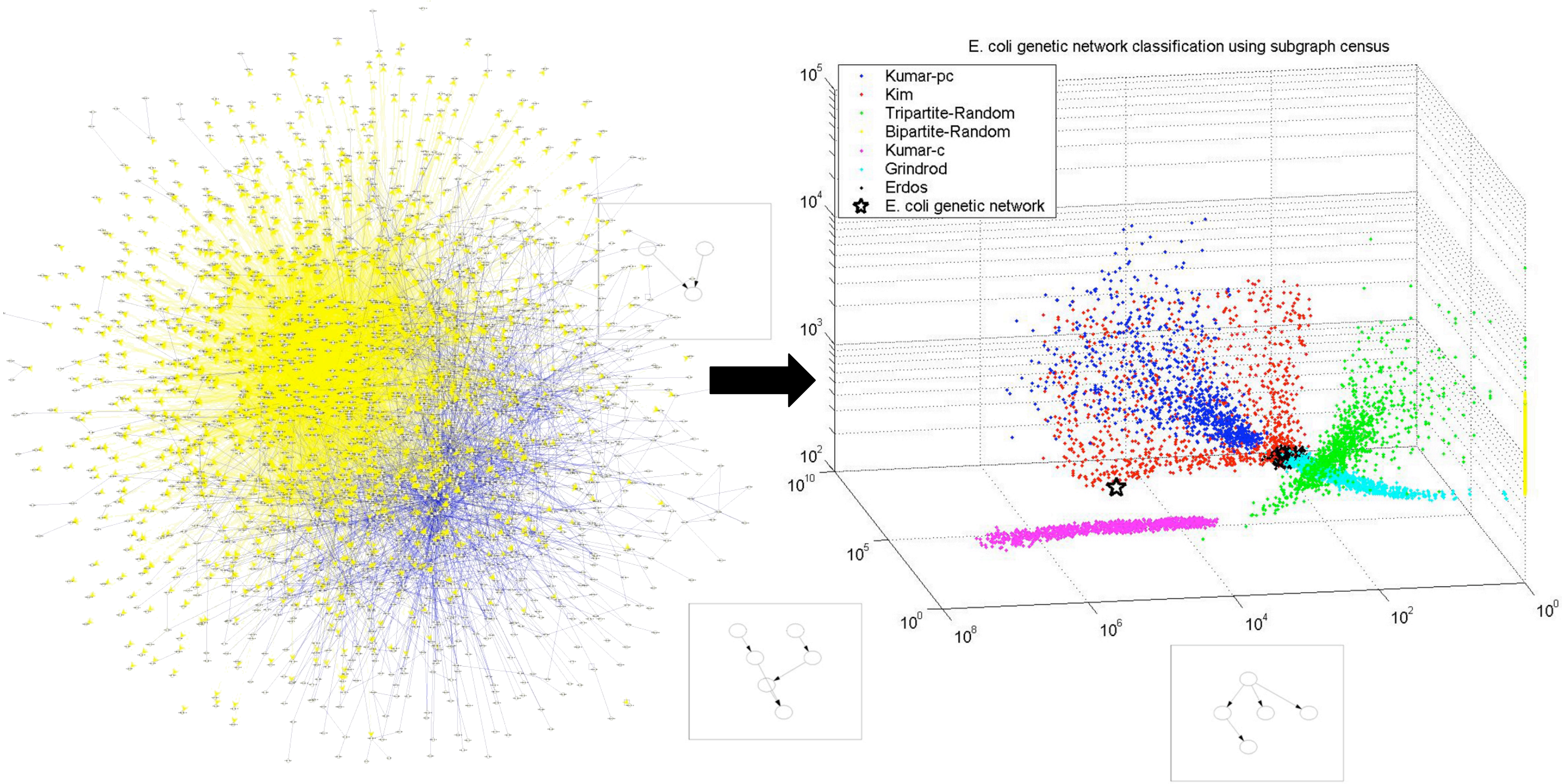
1-slide summary of classification

- Watts or Barabasi?



- learn predictive (not “overrepresented”) features from data;
- no null model assumed;
- no distribution assumed;

calculate discriminative features



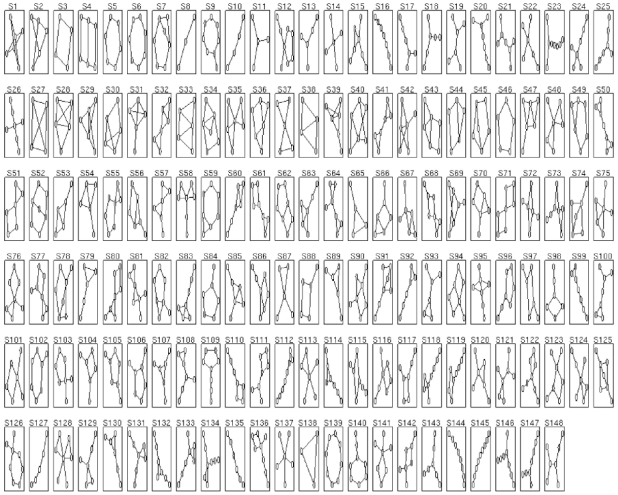
(and let the data decide which is best model)

agenda

- statistical network physics
  - the problem
  - probability
  - statistics
- statistical learning
- biological networks: predicting evolution
  - validating models
  - letting the data decide

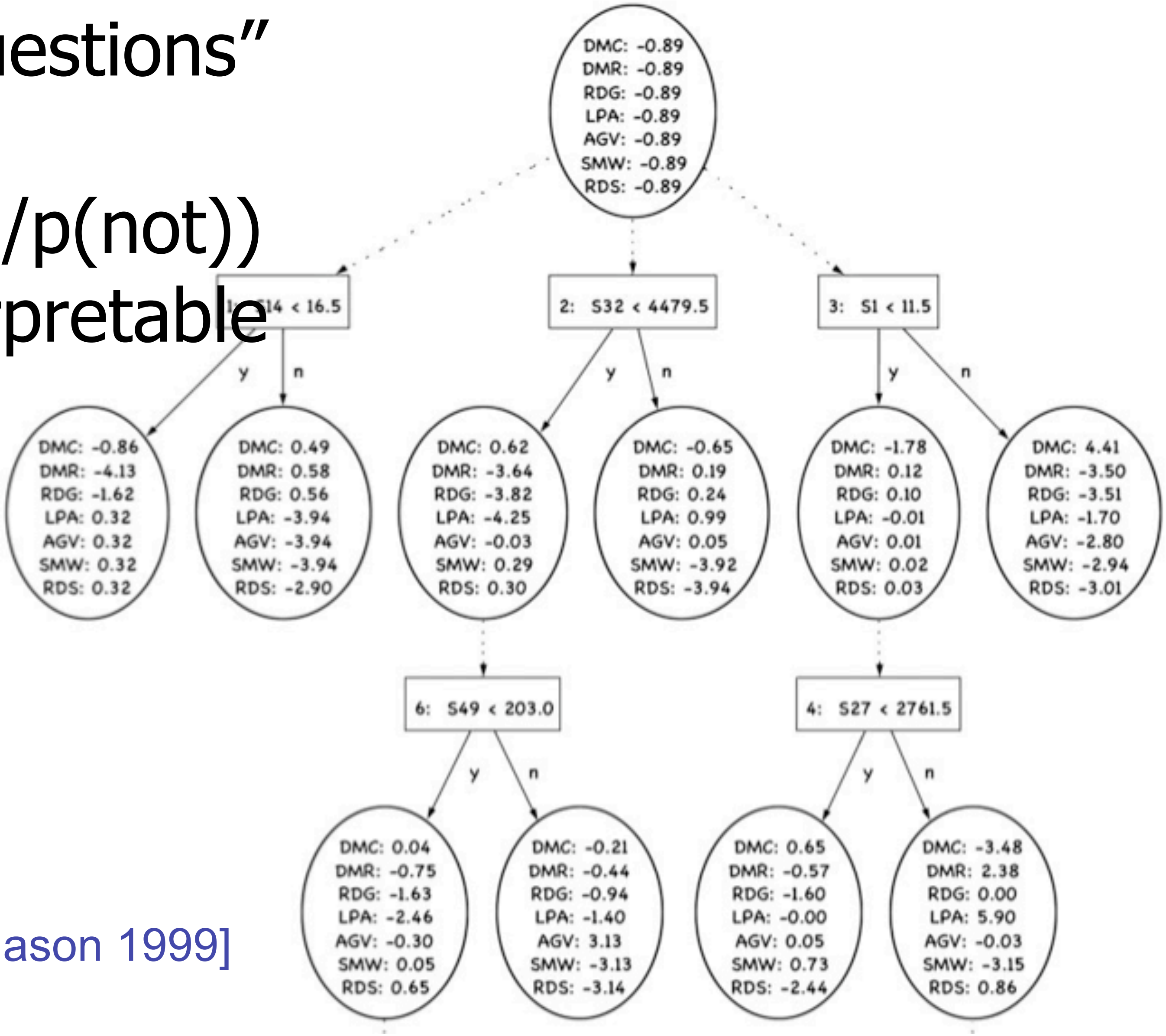
systematic enumeration of network features

- Subgraph census
  - exploit sparseness (“walks”)
  - use a pre-processed hash-table for subgraph isomorphisms
  - 148 subgraphs shown, can easily do 181 subgraphs



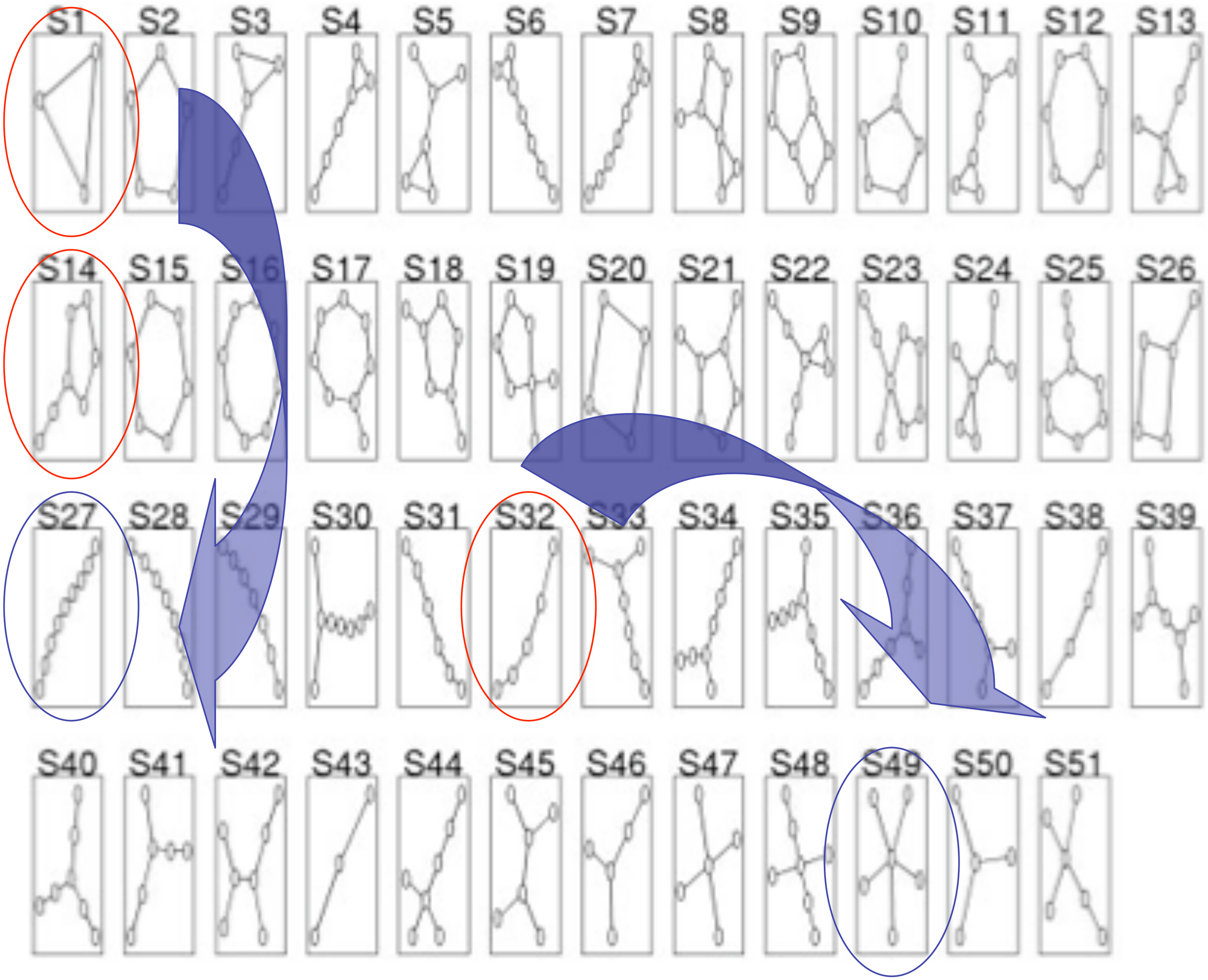
NetBoost: 20 questions

- play “20 questions”
- output log(p(model)/p(not))
- highly interpretable



[ADTs: Freund & Mason 1999]

conditionally important subgraphs



high accuracy (fit vs. overfit ; test-loss)

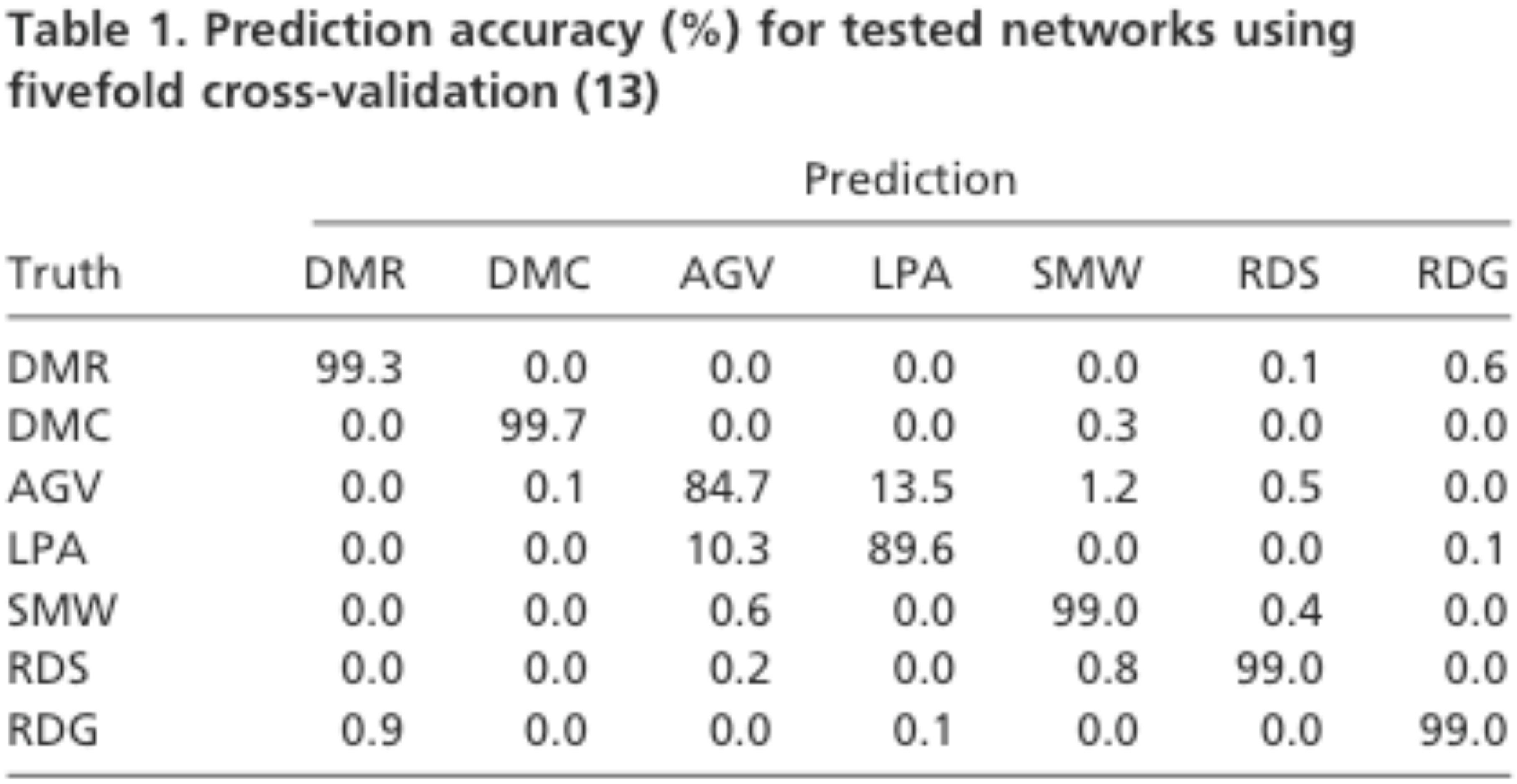


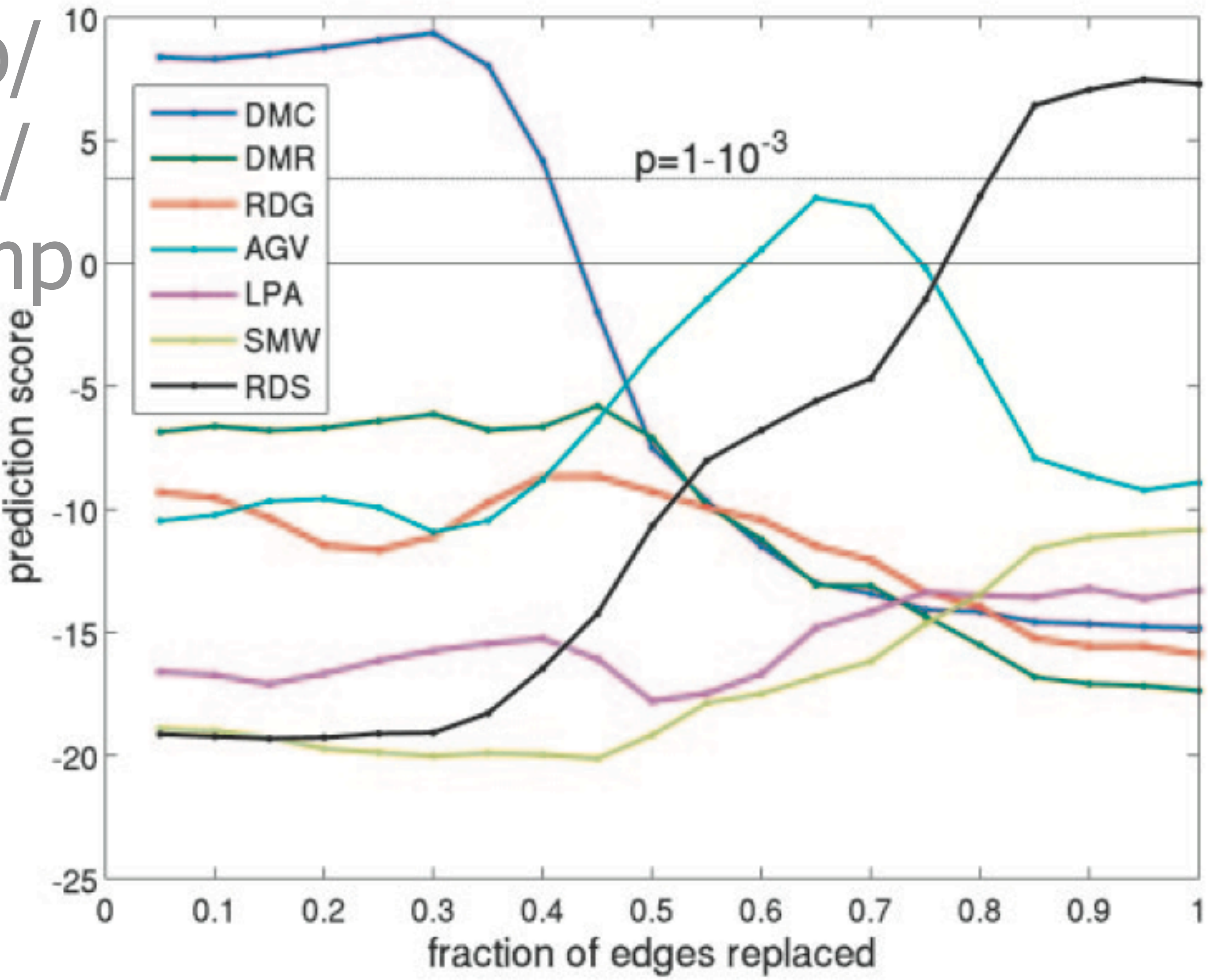
Table 1. Prediction accuracy (%) for tested networks using fivefold cross-validation (13)

	Prediction						
Truth	DMR	DMC	AGV	LPA	SMW	RDS	RDG
DMR	99.3	0.0	0.0	0.0	0.0	0.1	0.6
DMC	0.0	99.7	0.0	0.0	0.3	0.0	0.0
AGV	0.0	0.1	84.7	13.5	1.2	0.5	0.0
LPA	0.0	0.0	10.3	89.6	0.0	0.0	0.1
SMW	0.0	0.0	0.6	0.0	99.0	0.4	0.0
RDS	0.0	0.0	0.2	0.0	0.8	99.0	0.0
RDG	0.9	0.0	0.0	0.1	0.0	0.0	99.0

- Empirical estimate of generalization error
- not chi squared (not normal, too many parts=parameters)

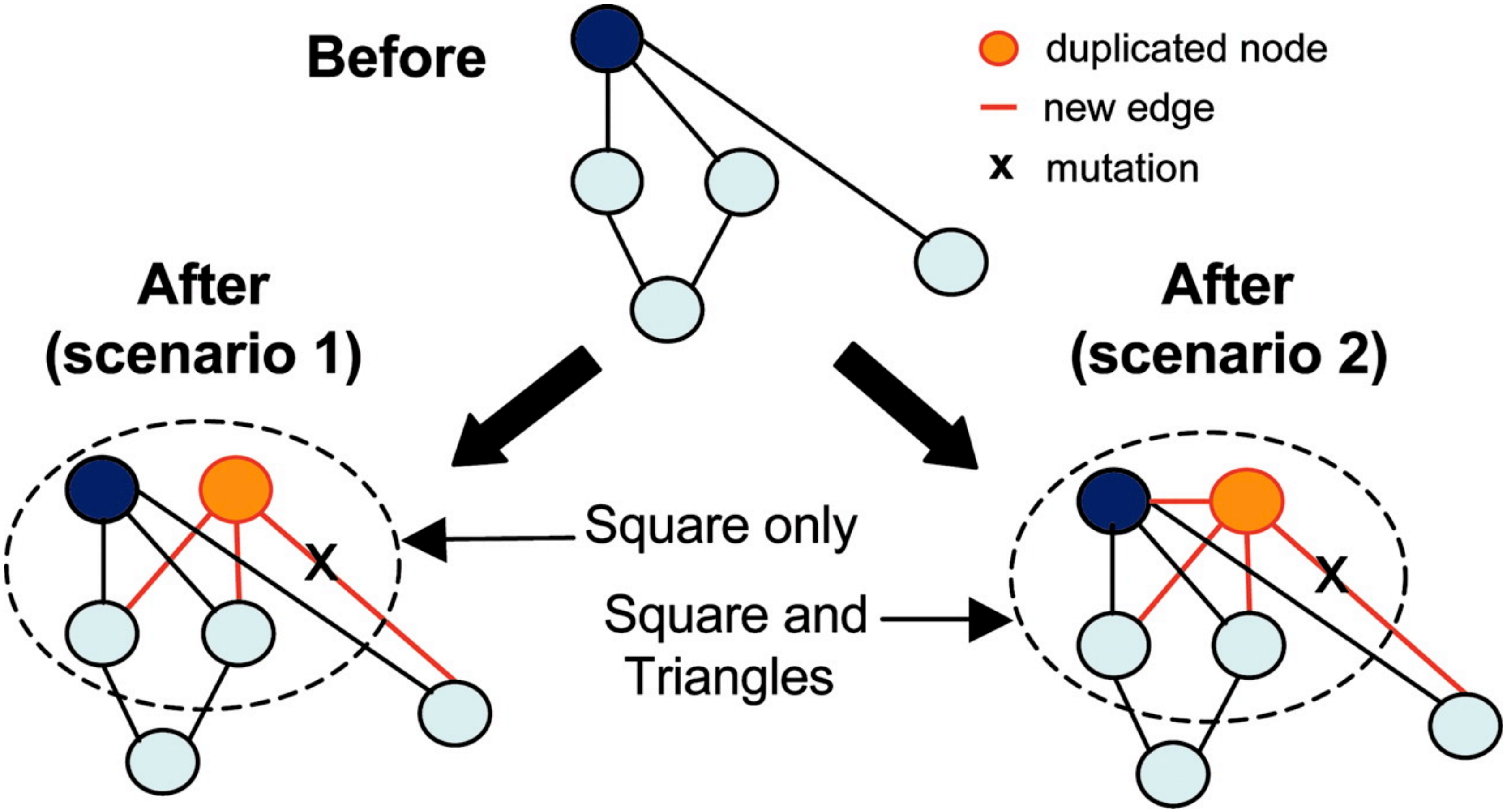
now look @ target: robust predictions

Dup/
Mut/
Comp



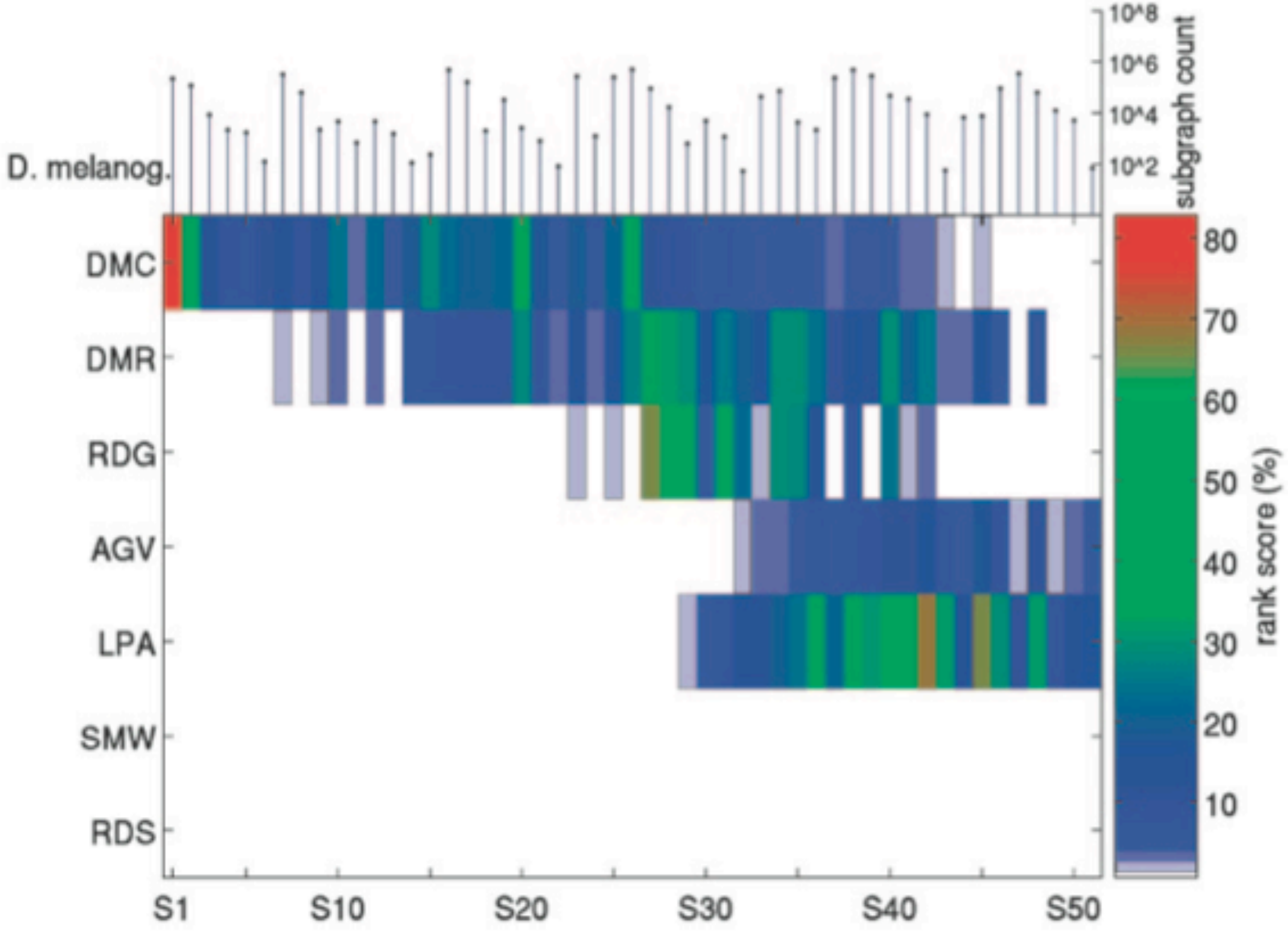
Erdös

DMC?

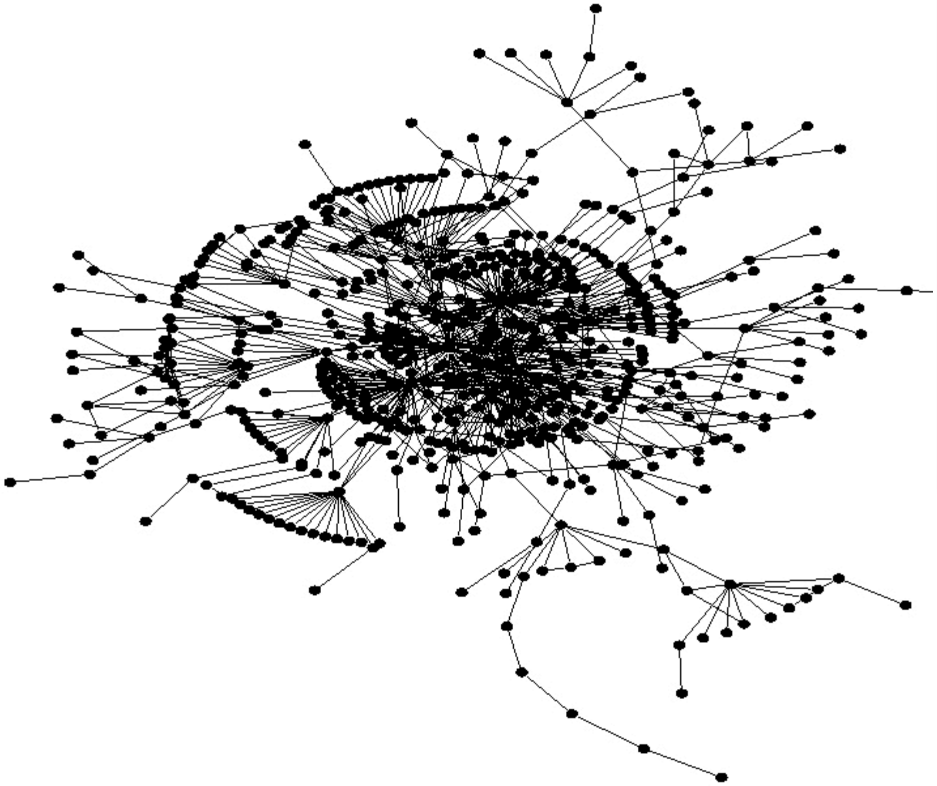


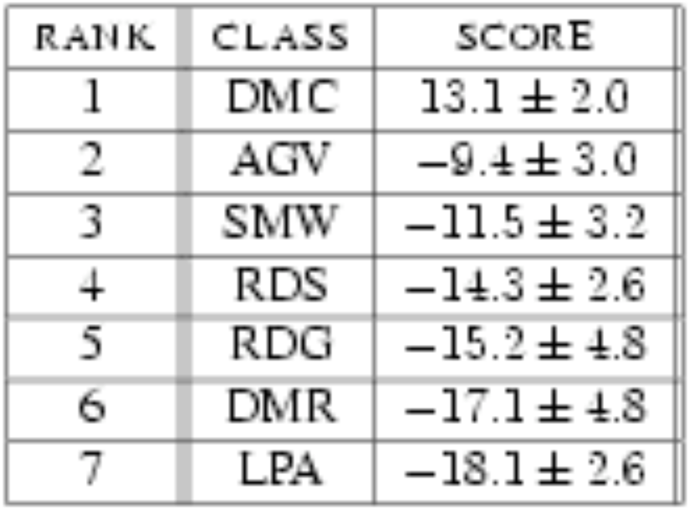
(from Rice, Kershenbaum, Stolovitzky's *Commentary*)

rank scores


D. melanog.
DMC
DMR
RDG
AGV
LPA
SMW
RDS
S1
S10
S20
S30
S40
S50
10^8
10^6
10^4
10^2
subgraph count
80
70
60
50
40
30
20
10
rank score (%)

not just for flies: yeast P-P network

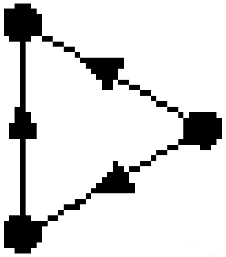




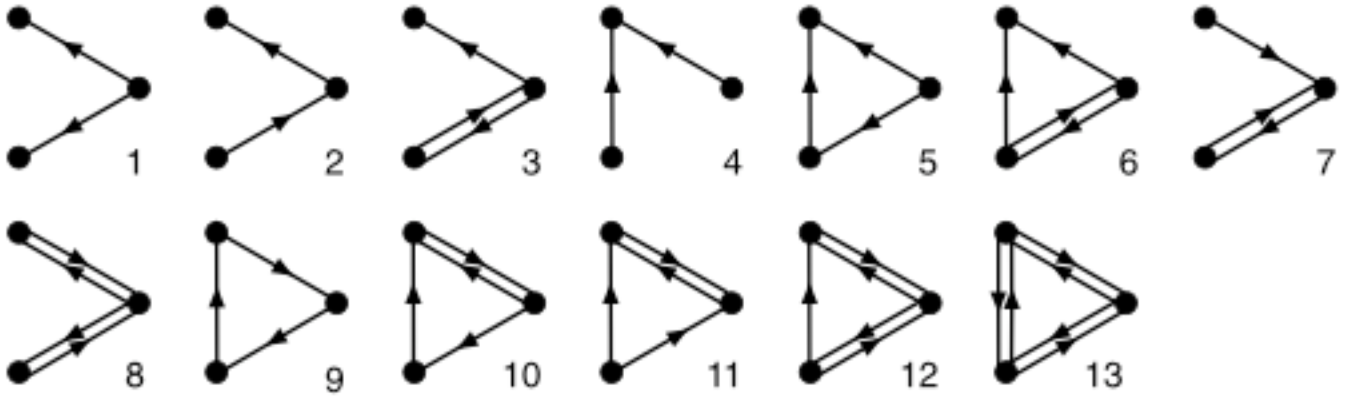
RANK	CLASS	SCORE
1	DMC	13.1 ± 2.0
2	AGV	−9.4 ± 3.0
3	SMW	−11.5 ± 3.2
4	RDS	−14.3 ± 2.6
5	RDG	−15.2 ± 4.8
6	DMR	−17.1 ± 4.8
7	LPA	−18.1 ± 2.6

data courtesy O. Troyanskaya

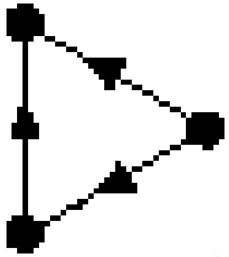
why subgraphs?



subgraph census: history



- Triad Census to test for **transitivity**,
  Holland and Leinhardt, 1970

‘FFL’ =  = ‘030T’ triad

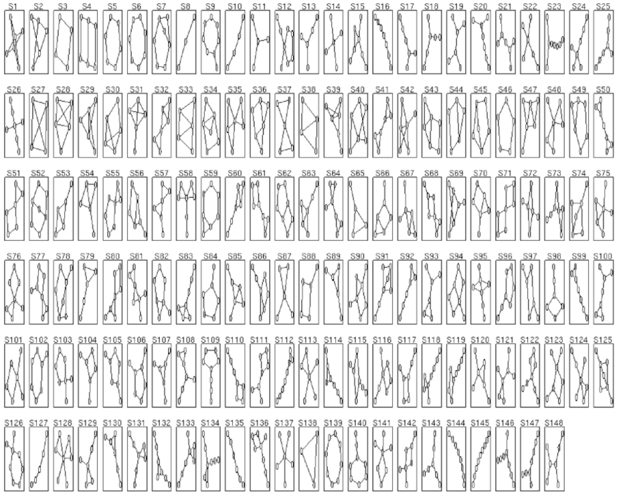
subgraph census: problems

- Number of isomorphism classes grows rapidly with graph size (Haray, 1955)

3	dyads
16	triads
218	tetrads
9608	pentads

- Census sensitive to density, clustering, degree distributions
- Traditional algorithms limited to n=3 or n=4
- Larger structures require tailored, parameterized algorithms

systematic enumeration of network features


S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12 S13 S14 S15 S16 S17 S18 S19 S20 S21 S22 S23 S24 S25
S26 S27 S28 S29 S30 S31 S32 S33 S34 S35 S36 S37 S38 S39 S40 S41 S42 S43 S44 S45 S46 S47 S48 S49 S50
S51 S52 S53 S54 S55 S56 S57 S58 S59 S60 S61 S62 S63 S64 S65 S66 S67 S68 S69 S70 S71 S72 S73 S74 S75
S76 S77 S78 S79 S80 S81 S82 S83 S84 S85 S86 S87 S88 S89 S90 S91 S92 S93 S94 S95 S96 S97 S98 S99 S100
S101 S102 S103 S104 S105 S106 S107 S108 S109 S110 S111 S112 S113 S114 S115 S116 S117 S118 S119 S120 S121 S122 S123 S124 S125
S126 S127 S128 S129 S130 S131 S132 S133 S134 S135 S136 S137 S138 S139 S140 S141 S142 S143 S144 S145 S146 S147 S148

- Subgraph census

  - exploit sparseness (“walks”)
  - use a pre-processed hash-table for subgraph isomorphisms
  - 148 subgraphs shown, can easily do 181 subgraphs

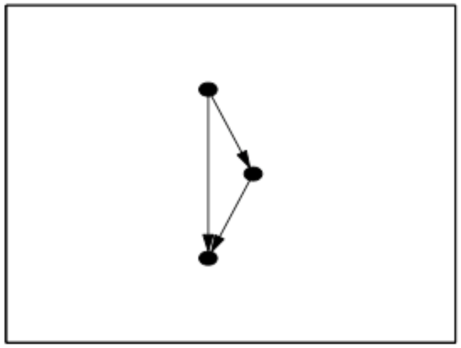
or

- Adjacency matrix functionals (“words”) (Ziv et al. cond-mat/0306610)

  - more efficient than subgraph census for denser networks
  - up to 4670 features tested

matrix functionals & graphs

- 030T (FFL) signature



$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} ; \mathrm{diag}(A^2 A^T) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Path operators

- – A = adjacency; (walking down the graph)
- – A^T = transpose; (walking up the graph)
- – D = diag; (restriction to closed walks)
- – U = I-D; (restriction to open walks)

sparse matrix functionals

In other words …

Number of FFLs =

“sum $D(A^2A^T)$”

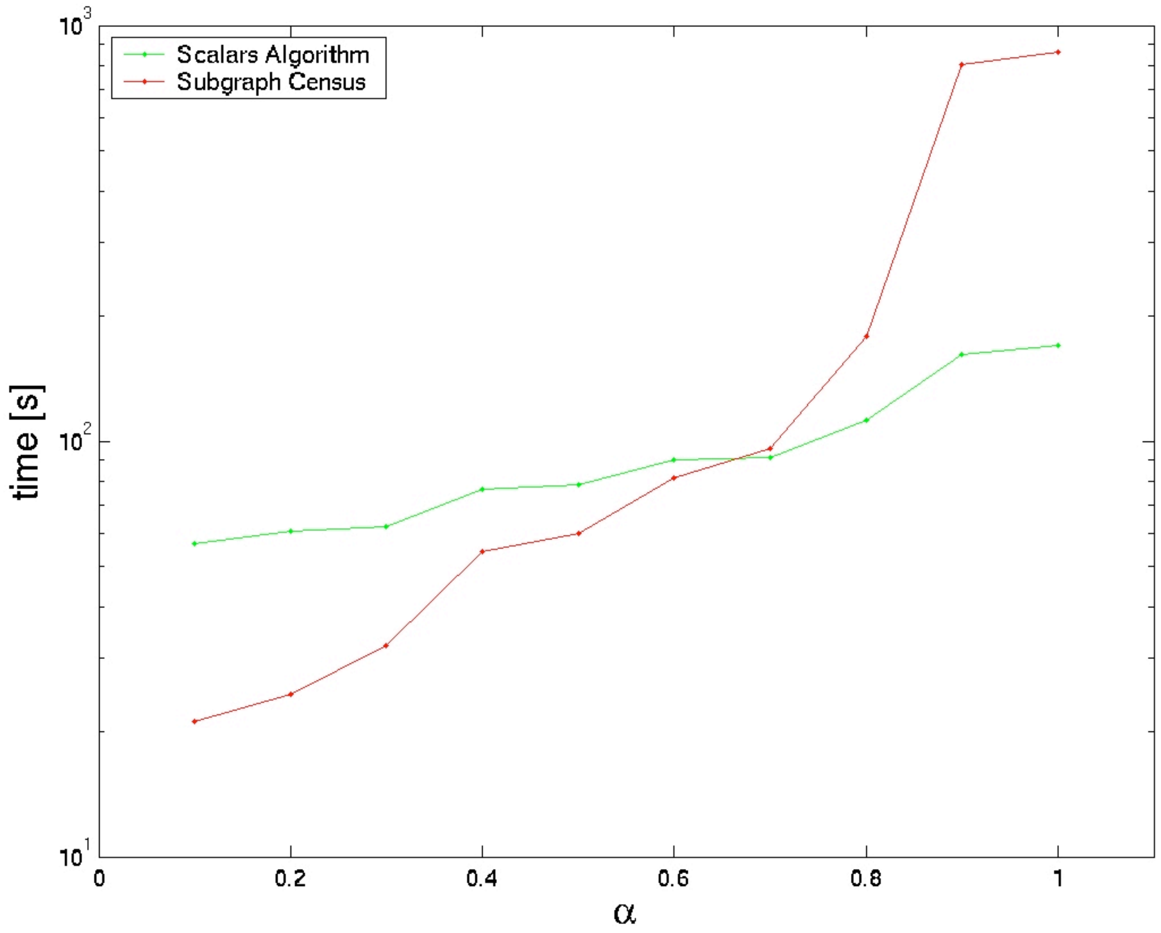
Example 1: $S(D(A^2A^T)) = 40$ = the number of FFLs in the E. coli network

Example 2: $nnz(D(A^2A^T)) = 10$

16 of 40 FFLs associated with gene csgBA

sum => number of distinct paths between all pairs of endpoints
nnz => number of distinct paths between unique pairs of endpoints

computational efficiency

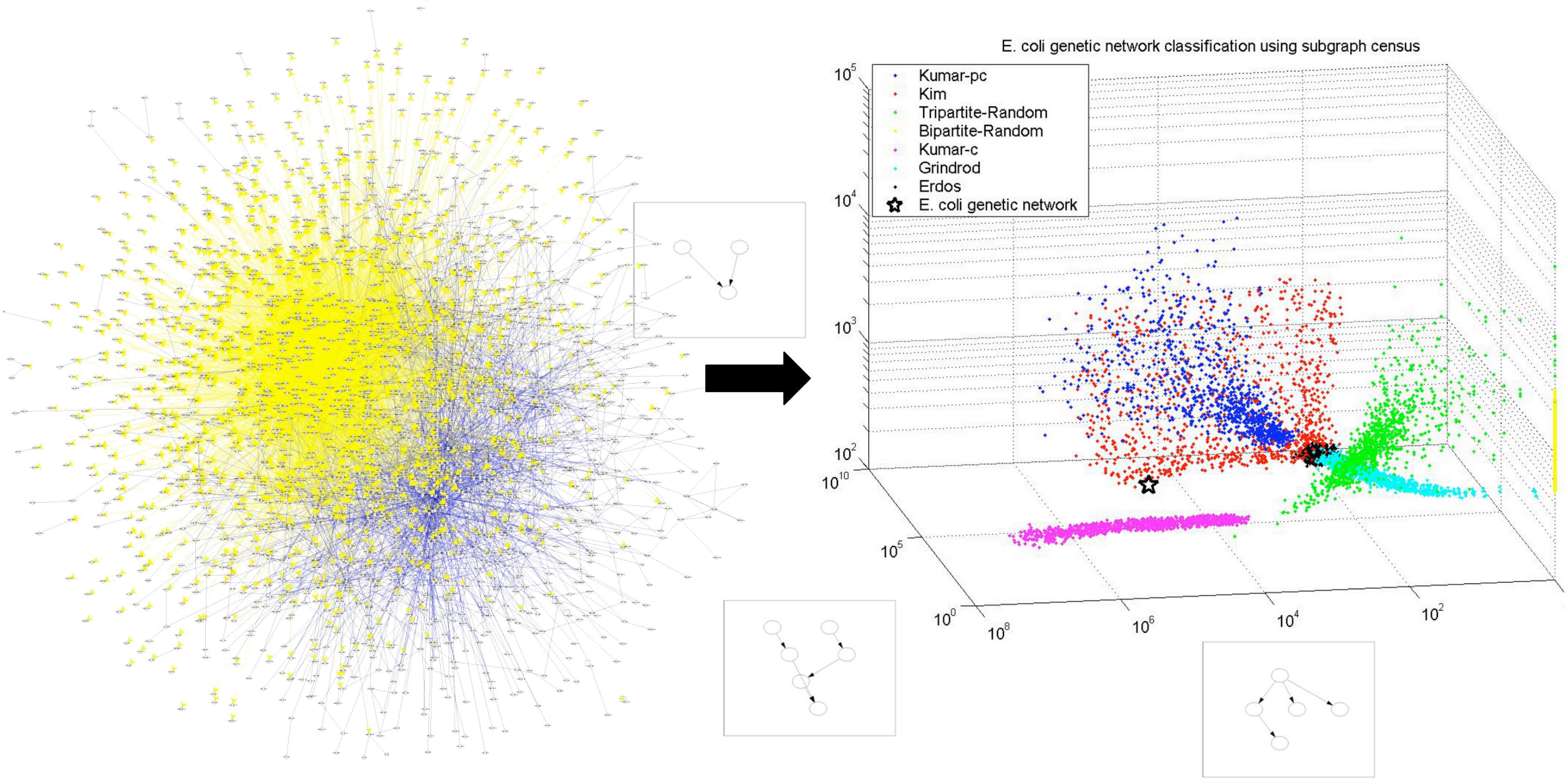


Tunable, preferential-attachment (PA) parameter

•Barabasi and Albert, Science '99

Scalars perform better for networks that are dense, clustered, or networks with long-tailed degree distributions

NetClass: predict mechanism as class

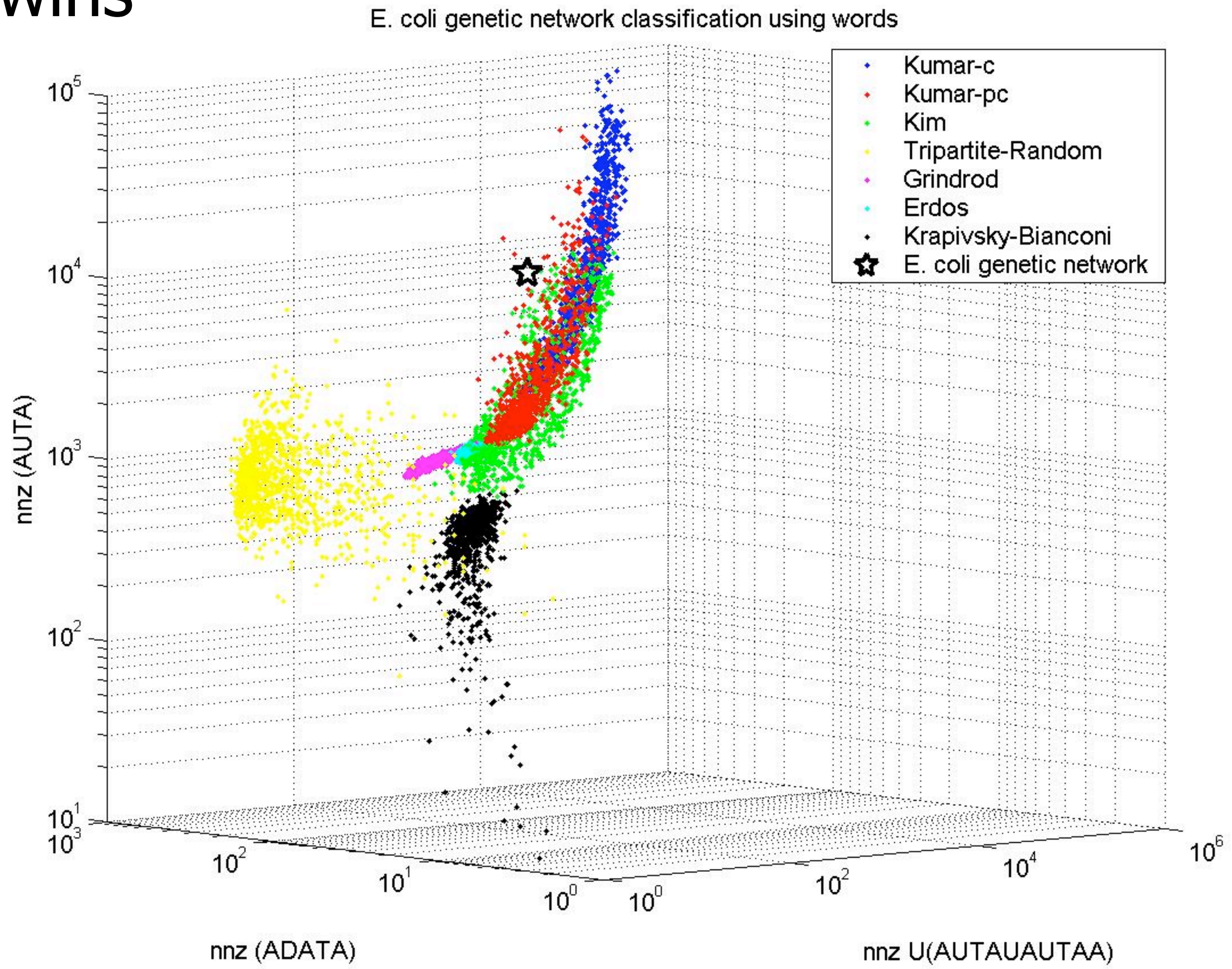


q-bio/**0402017;** BMC Bioinformatics 2004, 5:181

NetClass: *E. coli* Transcriptional Network

Kumar-C wins

(words)

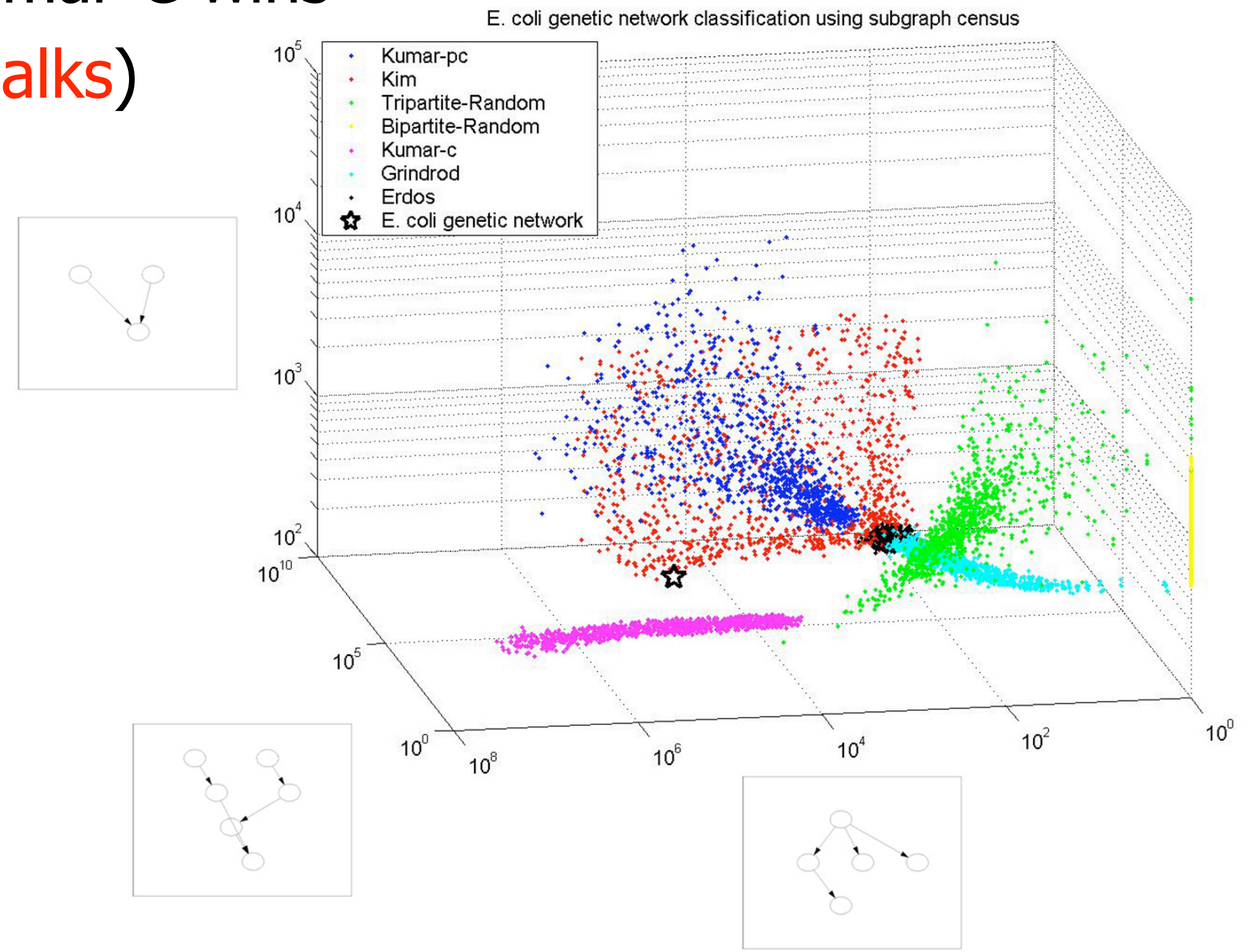


N=324; m=519; d=.3%; r=1.0

NetClass: *E. coli* Transcriptional Network

Kumar-C wins

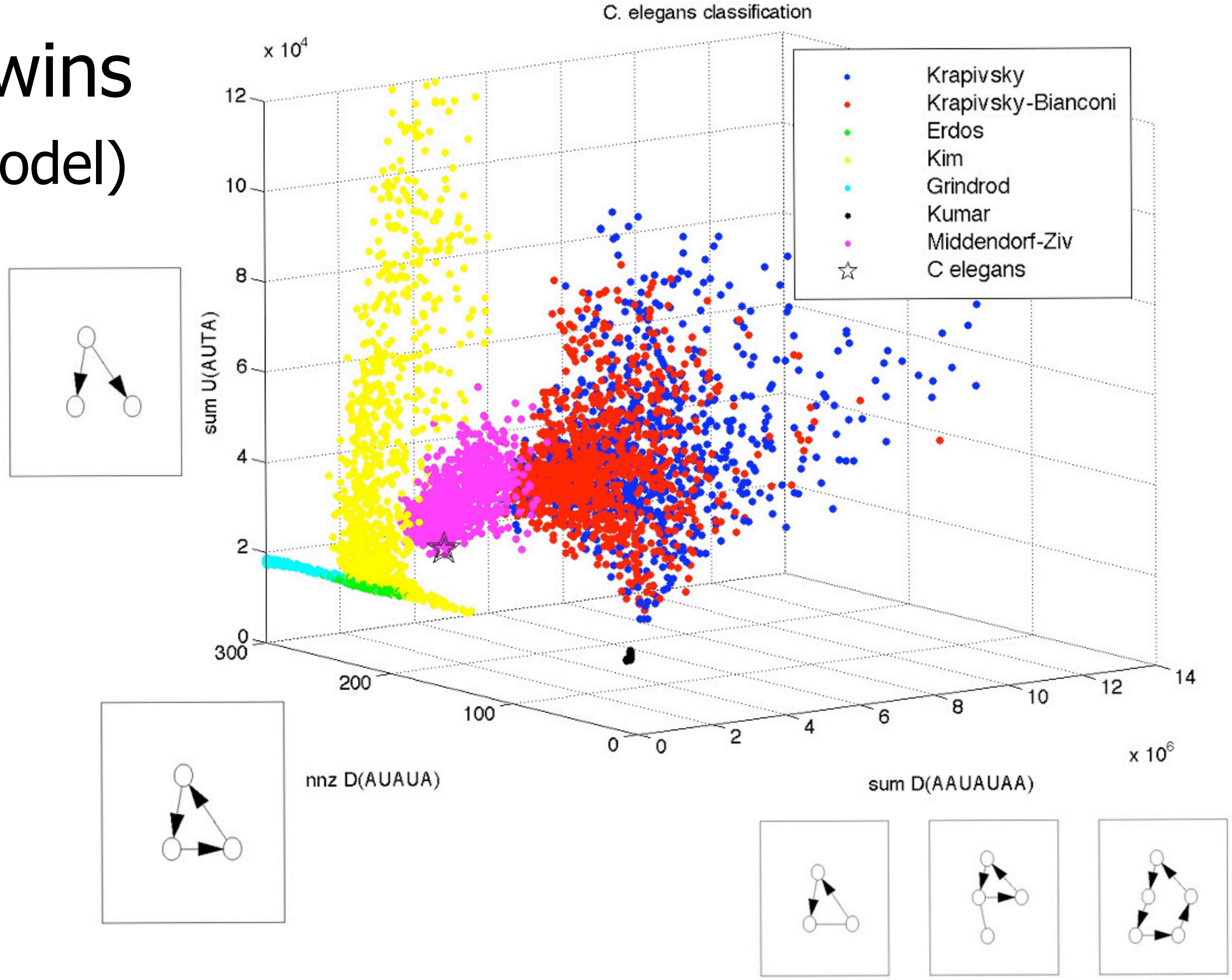
(walks)



NetClass: *C. elegans* Neural Network

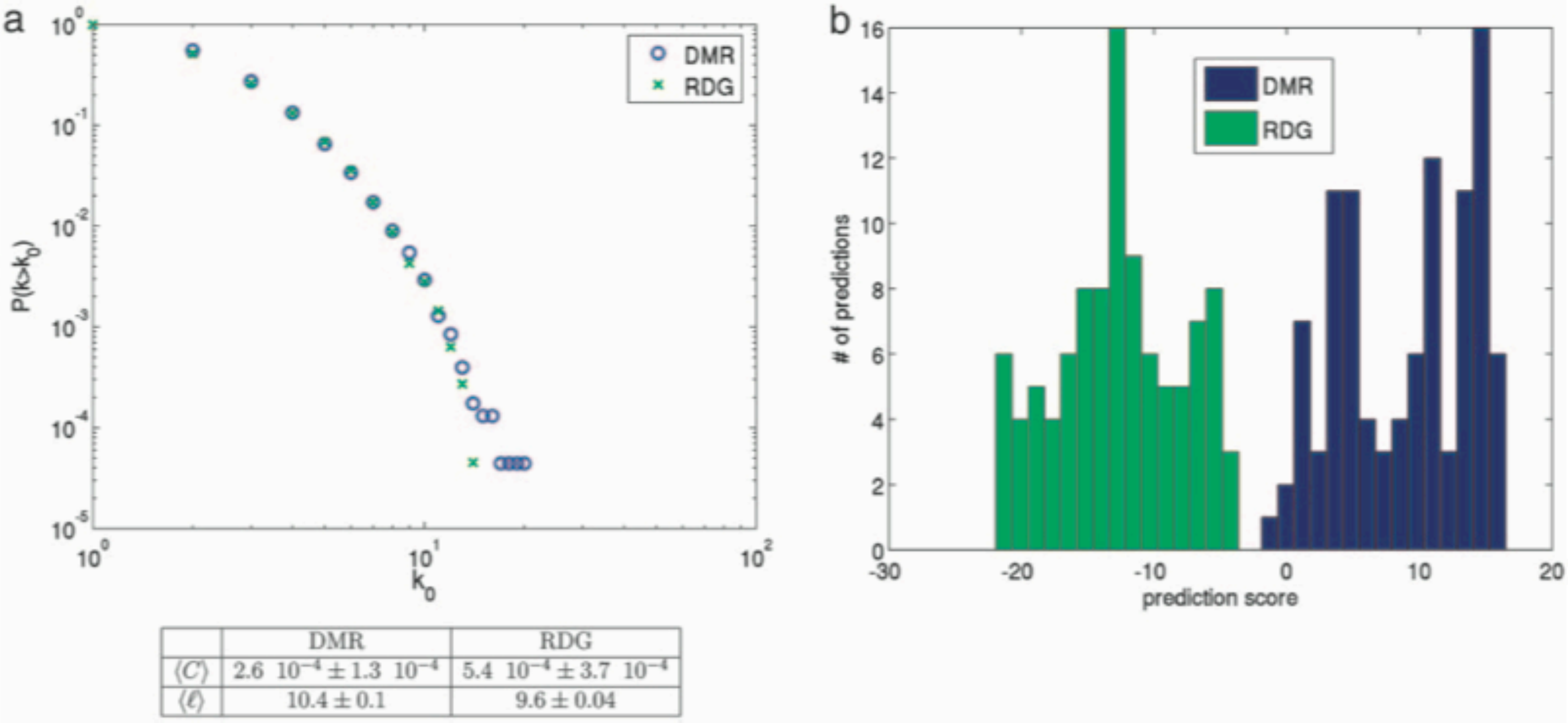
“MZ” wins

(new model)



N=306; m=2359; d=2.5%; r=.97

what is important? let the data decide



	DMR	RDG
$\langle C \rangle$	$2.6\ 10^{-4} \pm 1.3\ 10^{-4}$	$5.4\ 10^{-4} \pm 3.7\ 10^{-4}$
$\langle \ell \rangle$	10.4 ± 0.1	9.6 ± 0.04

Fig. 1. Discriminating similar networks. Ten graphs of two different mechanisms exhibit similar average geodesic lengths and almost identical degree distribution and clustering coefficients. (*a*) Cumulative degree distribution $p(k > k_0)$, average clustering coefficient $\langle C \rangle$ and average geodesic length $\langle \ell \rangle$, all quantities averaged over a set of 10 graphs. (*b*) Prediction scores for all 10 graphs and all five cross-validated (13) ADTs. The two sets of graphs can be perfectly separated by our classifier, even though none of these graphs is used in the classifier training.

statistical systems biology: agenda

1. challenges to keep in mind
2. microarrays / regulation
3. networks
4. final thoughts

things to watch out for:

1. methods / how to read
2. different data, same issues
3. “prediction”
4. validation

how to read/write a comp. sys. bio. paper:

1. background
2. intuition
3. question to be answered, in words
4. question to be answered, in math:
5. algorithm
6. validation

$$\boxed{\vartheta = \underset{\vartheta \in \Omega}{\operatorname{argmin}} \; \mathcal{L}(D, \vartheta; \lambda)}$$

“prediction”

1. overfitting
2. feature ranking / hypothesis generation (“qualitative predictions”)
3. predicting unseen data

validation, closely related to prediction

1. in literature / by friends
2. statistical validation (e.g., CV)
3. experiment

different data, same issues

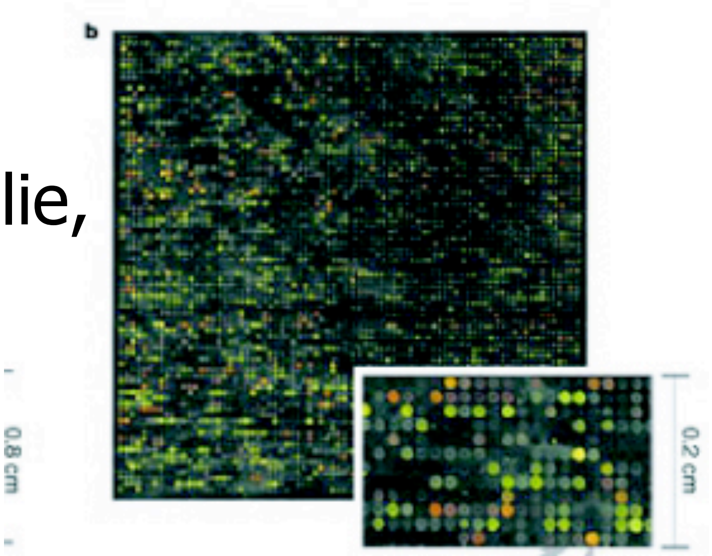
1. RNAi
2. ChIP-chip
3. PPi
4. image data
5. ...

learning networks from biology

- thanks:

  Freund, Kundaje, Leslie, Middendorf, + Shah

- for more info:
  - RECOMB, ISMB
- funding:
  - NIH NCBC
- open source.



$$A_g^t = f(\mu_g, \pi^t)$$

learning biology from networks

- thanks:

  Middendorf, Ziv

- for more info:
  - BMC Bioinfo, PNAS
- funding:
  - NSF/NIH/DOE
- open source:
  - sourceforge.net