On Learning Gene Regulatory Networks with Only Positive Examples

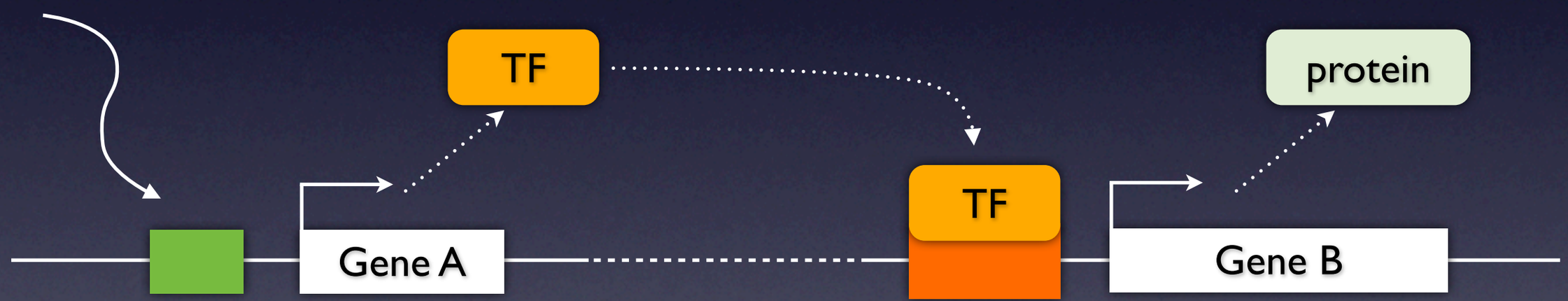
Luigi Cerulo, University of Sannio, Italy
Michele Ceccarelli, University of Sannio, Italy

Outline

- Supervised inference of gene regulatory networks
- The positive only problem
- Negative selection approaches
- Effect on prediction accuracy
- Conclusions and future directions

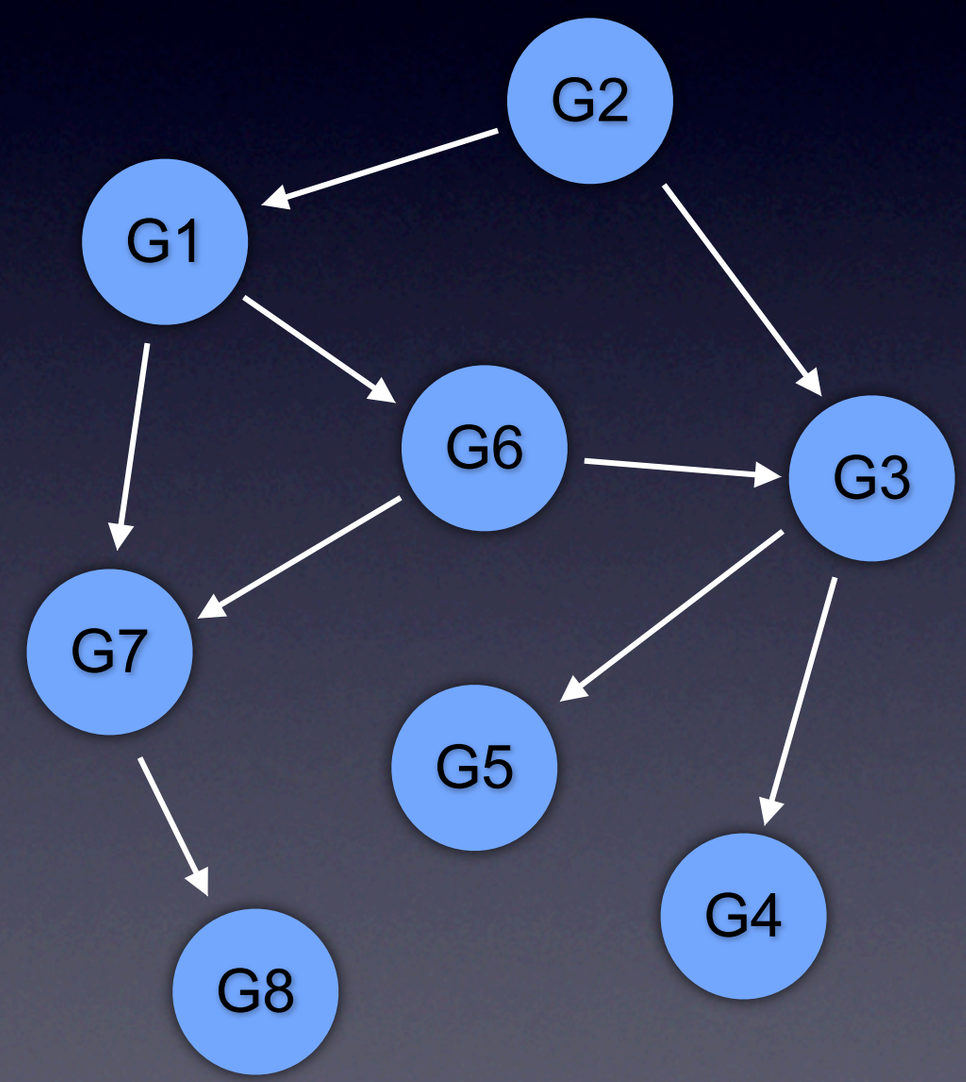
Gene Regulatory Network (GRN)

The network of transcription dependences among genes of an organism, known as transcription factors, and their binding sites.

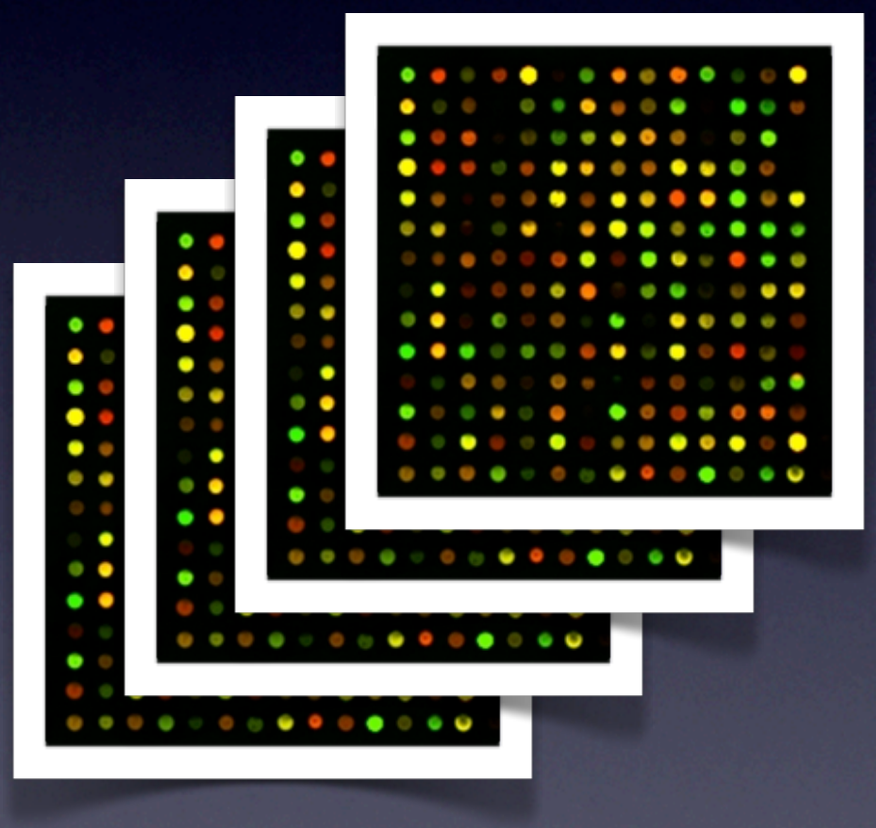


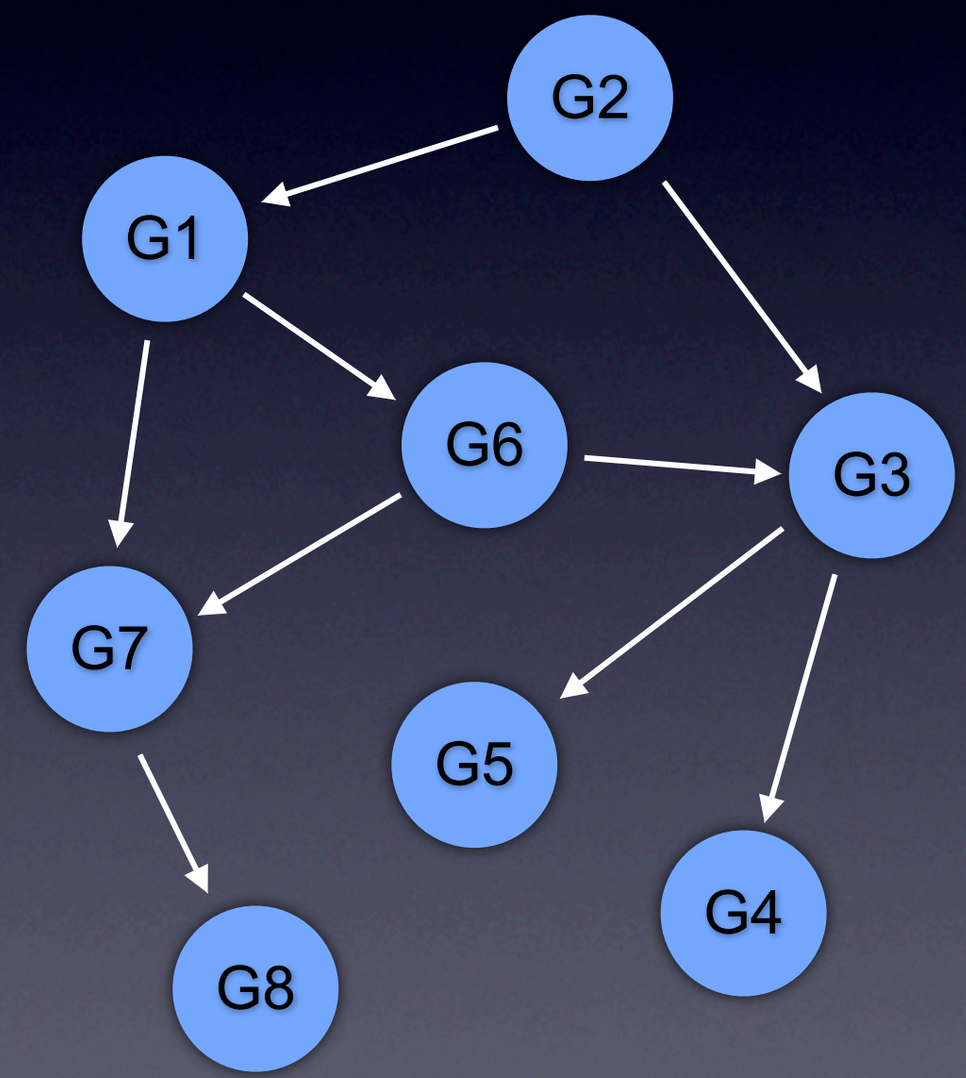
Gene Regulatory Network (GRN)

- A gene regulatory network can be represented as a graph G = (Vertices, Edges)
- Vertices = Genes
- Edges = Interactions



Inference of Gene regulatory networks





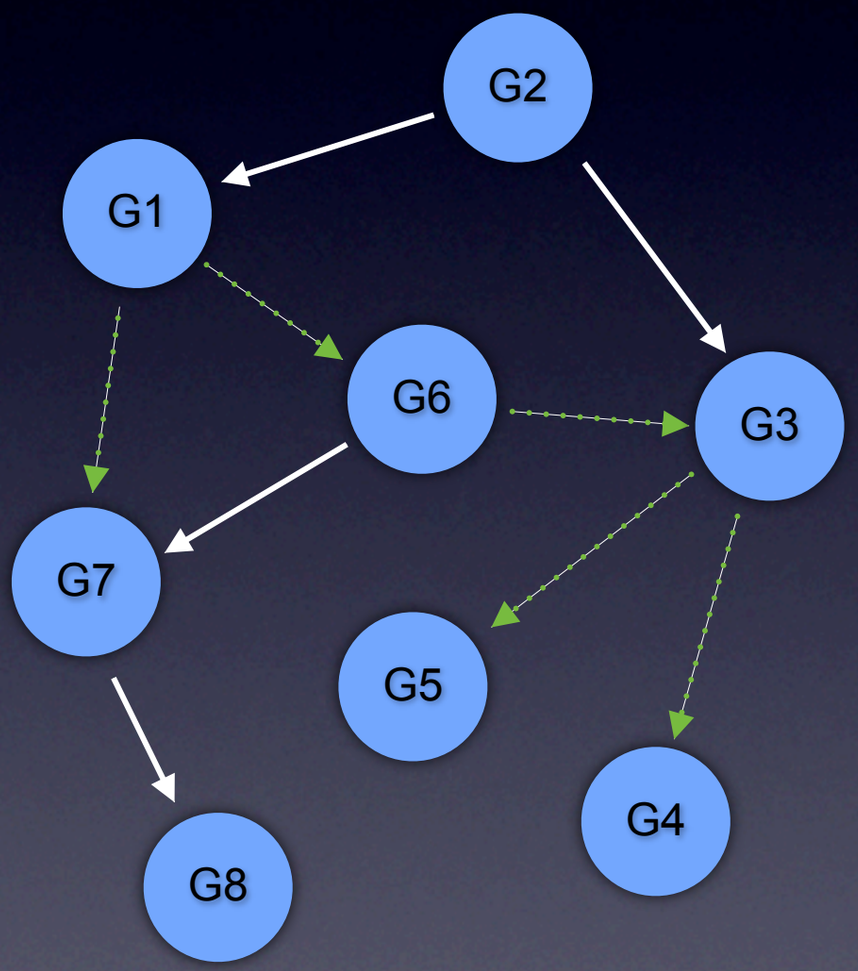
$$G_i = \{e_1, e_2, e_3, \ldots, e_n\}$$

GRN unsupervised inference

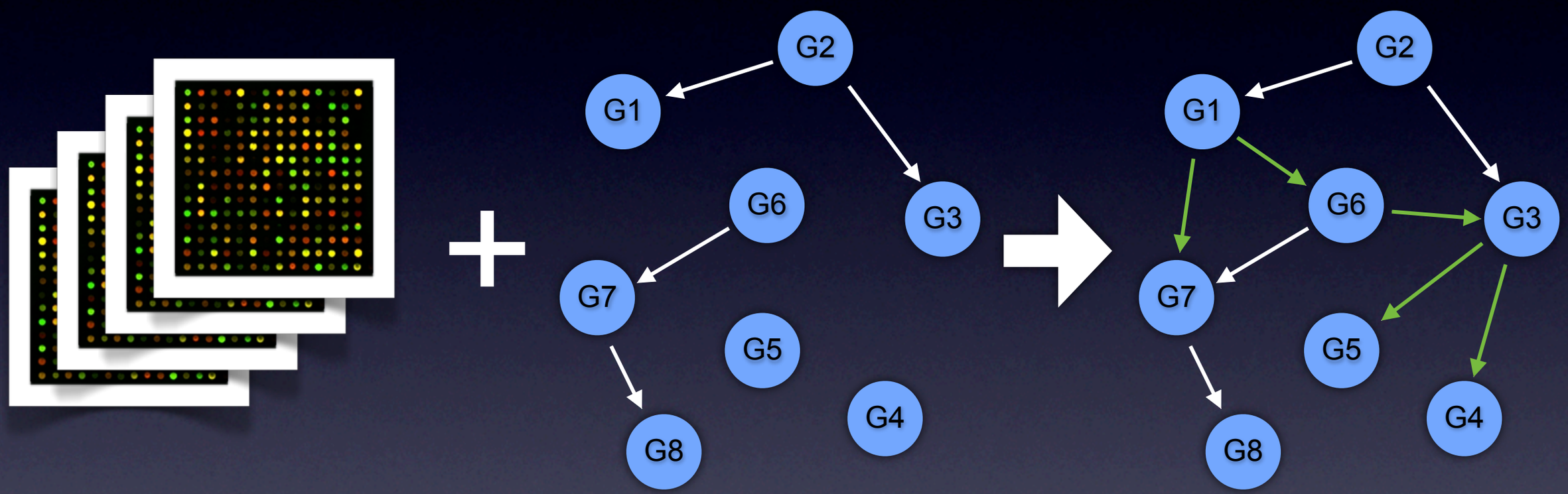
- Correlation models (eg. Mutual information)
- Bayesian Network
- Boolean networks
- ODE
- ...

GRN supervised Inference

- Part of the network is known in advance from public databases (Eg. RegulonDB)



GRN supervised Inference



$$G_i = \{e_1, e_2, e_3, \ldots, e_n\}$$

$$T = \{(G_1, G_2), (G_2, G_3), (G_6, G_7), (G_7, G_8)\}$$

Binary classifier (SVM, Decision Tree, Neural Networks,...)

Related work

BIOINFORMATICS

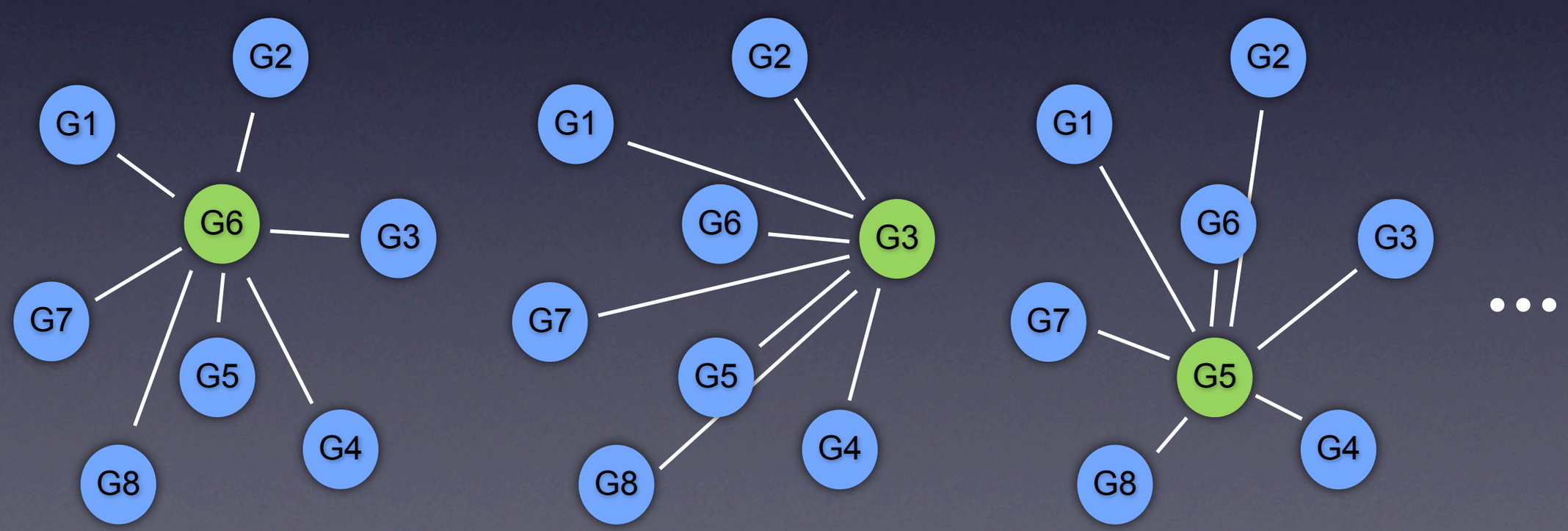
Vol. 24 ECCB 2008, pages i76–i82
doi:10.1093/bioinformatics/btn273

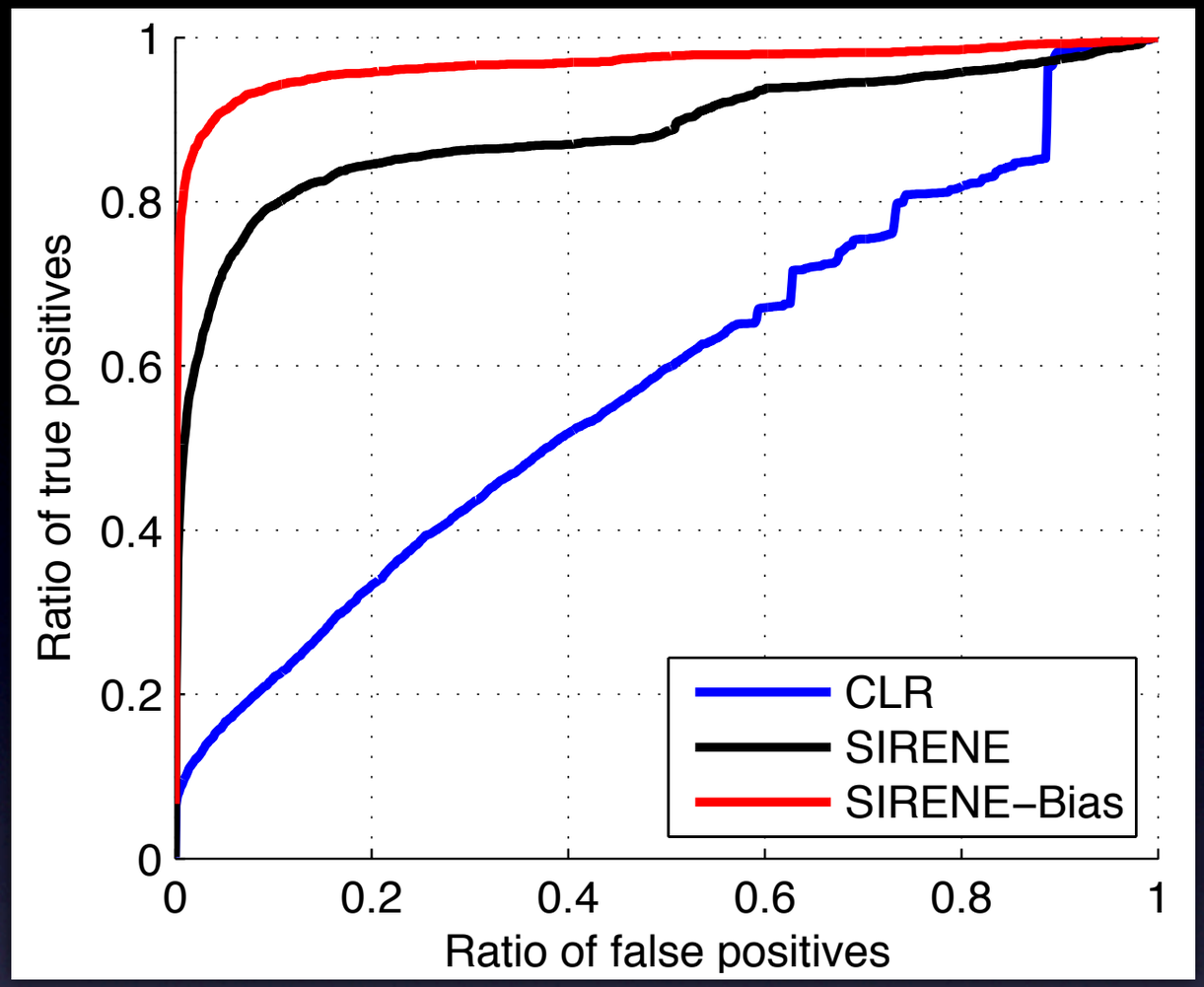
SIRENE: supervised inference of regulatory networks

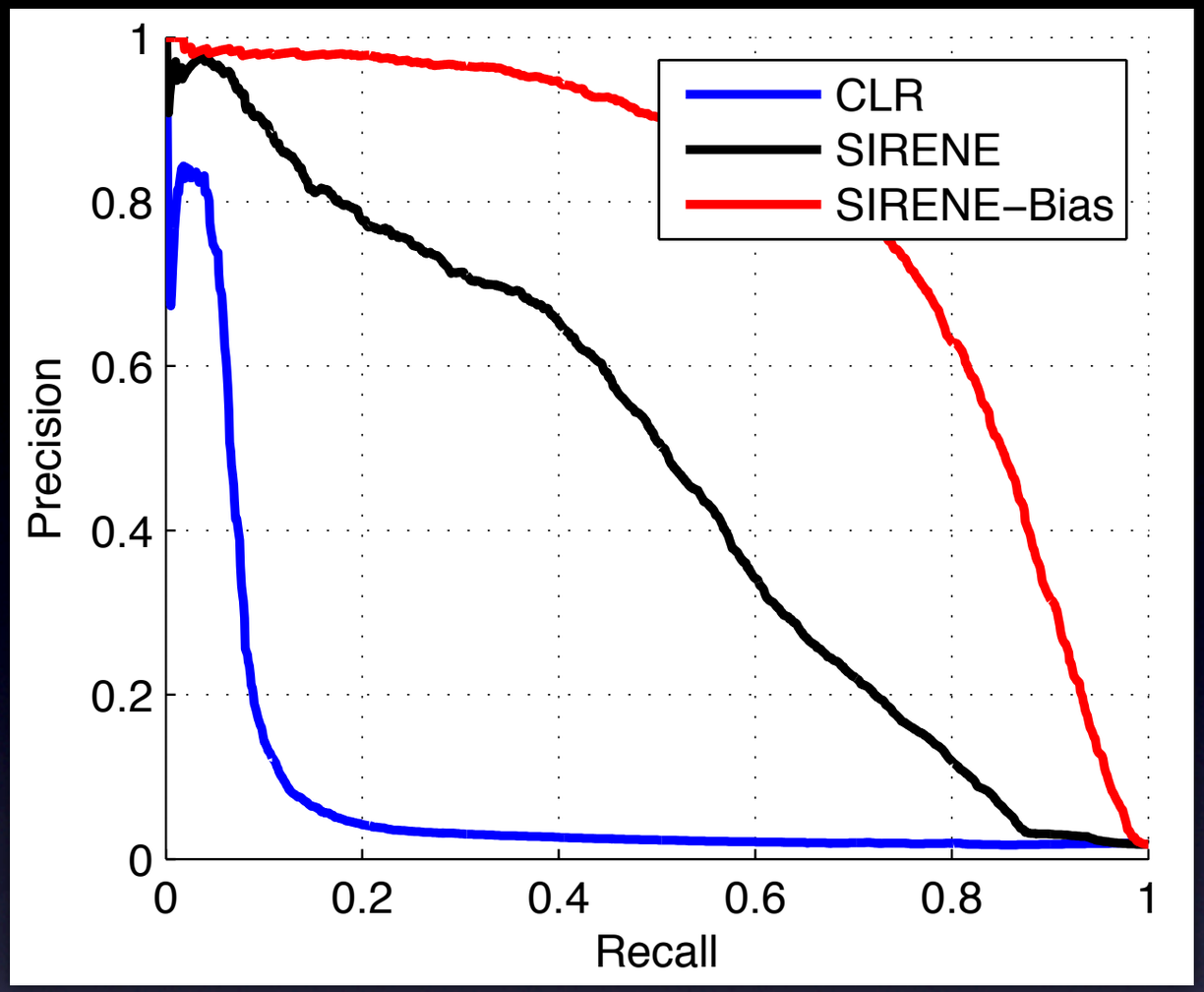
Fantine Mordelet[1,2,3,4,*] and Jean-Philippe Vert [1,2,3]

[1]Ecole des Mines de Paris, ParisTech, 35 rue Saint-Honoré, Fontainebleau F-77300, [2]Institut Curie, Paris F-75248, [3]INSERM, U900, Paris F-75248 and [4]CREST, INSEE, 3 av. Pierre Larousse, Malakoff, F-92240 France

- SIRENE approach
  - trains an SVM classifier for each gene and predicts which genes are regulated by that gene
  - combines all predicted regulations to obtain the full regulatory network



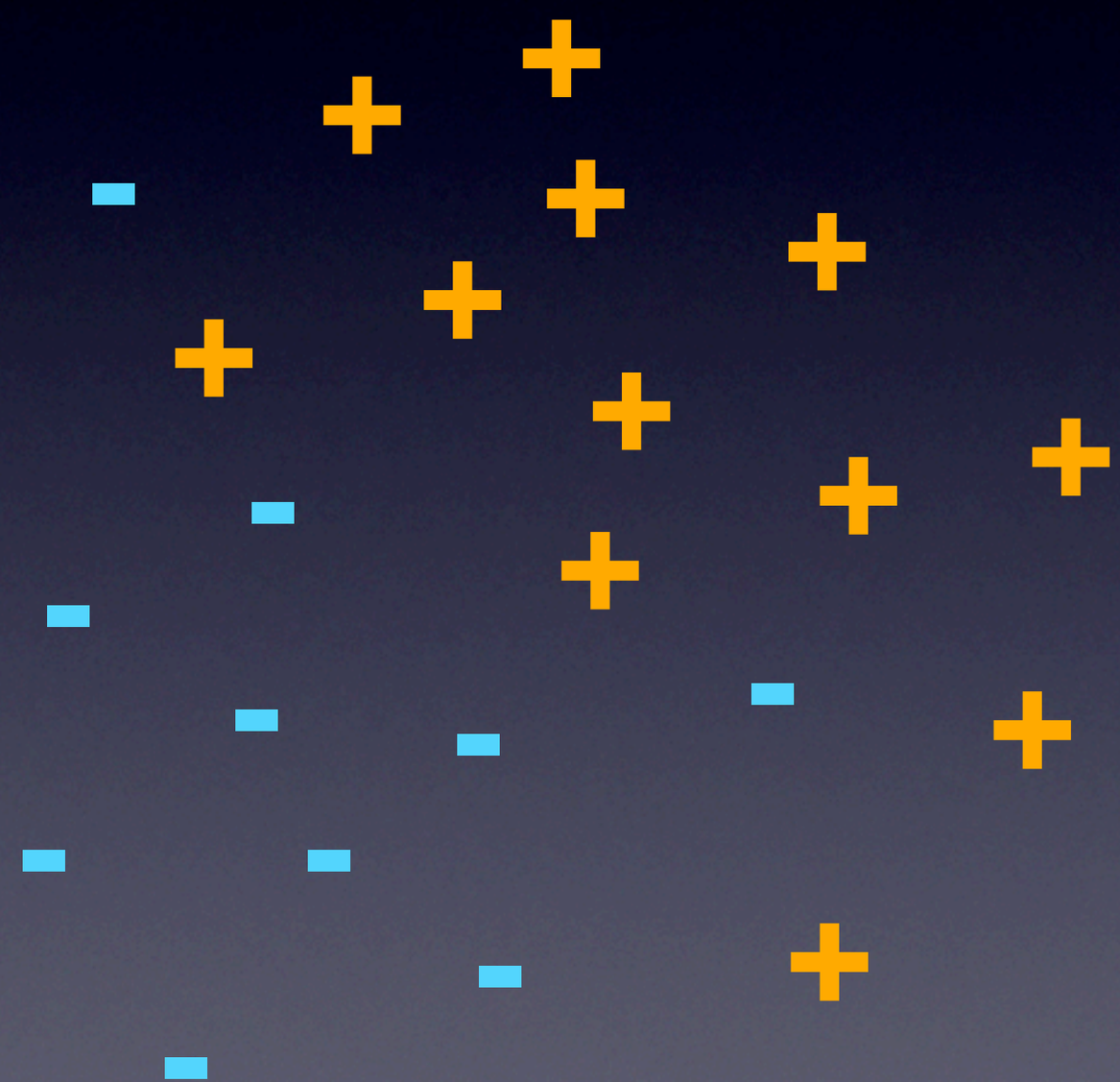




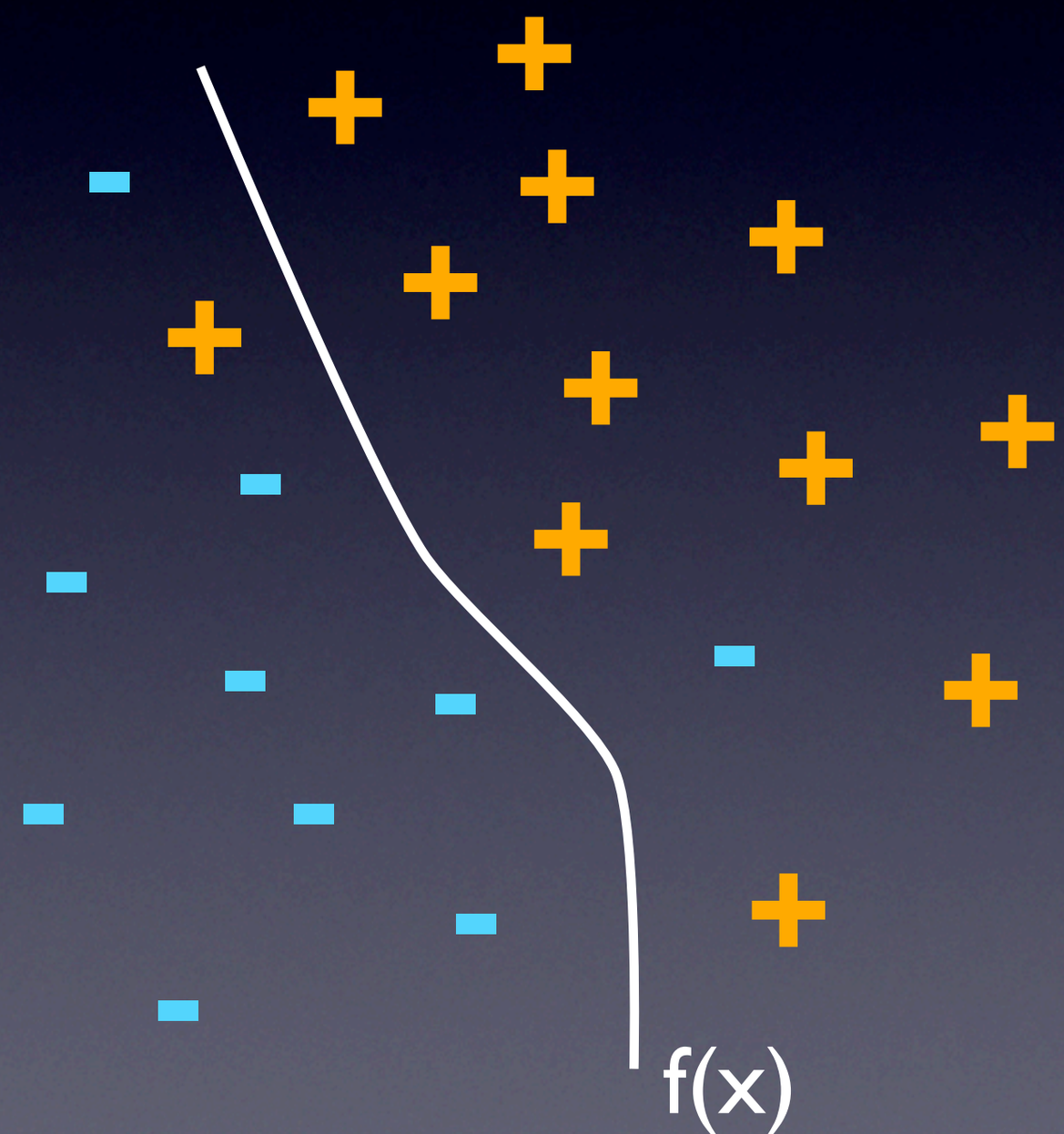
Method	Recall at 60% of Precision	Recall at 80% of Precision
SIRENE	44.5%	17.6%
CLR	7.5%	5.5%
Relevance networks	4.7%	3.3%
ARACNe	1%	0%
Bayesian network	1%	0%

Compared with unsupervised methods (Mordelet and Vert, 2008)

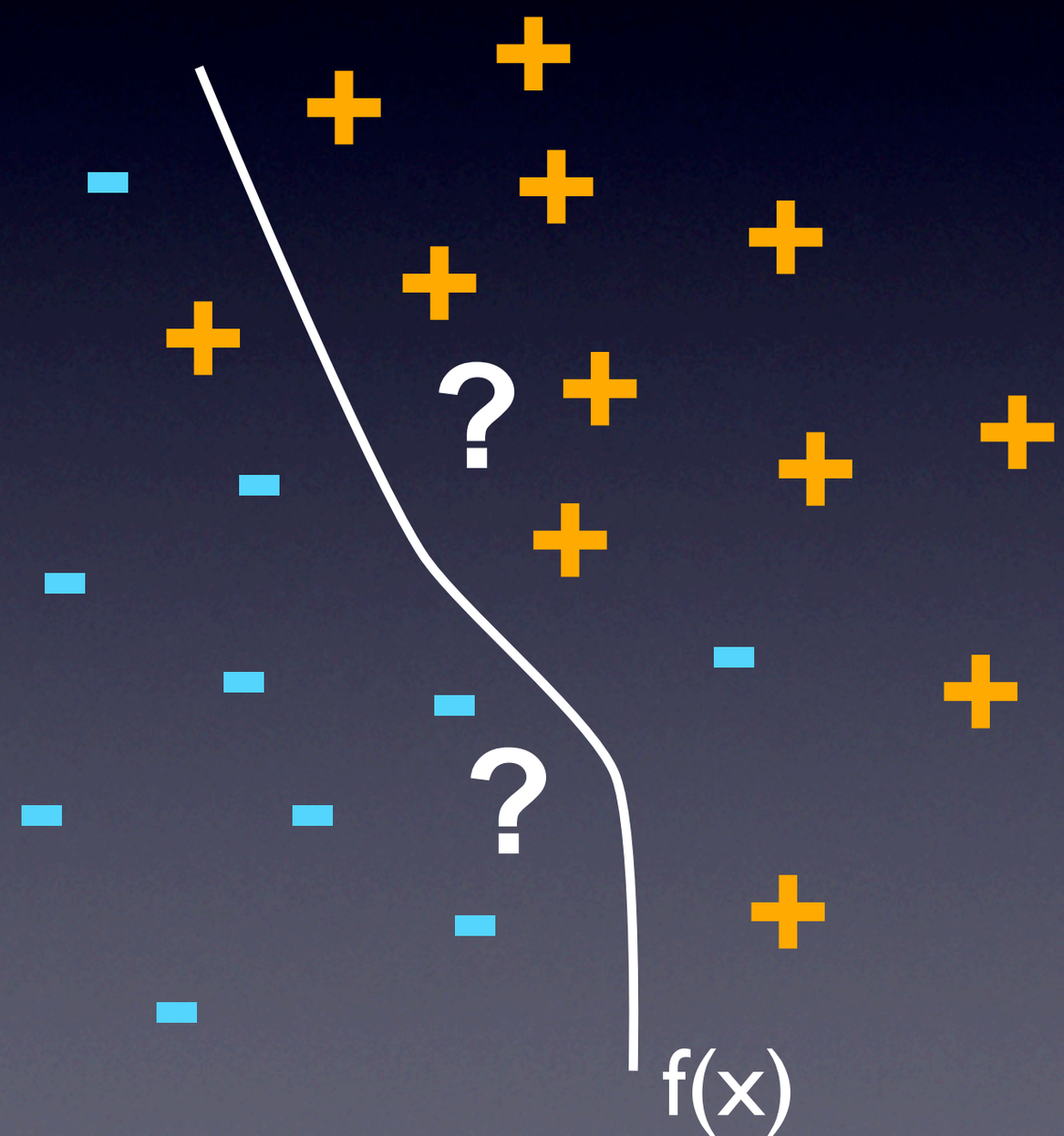
Supervised learning



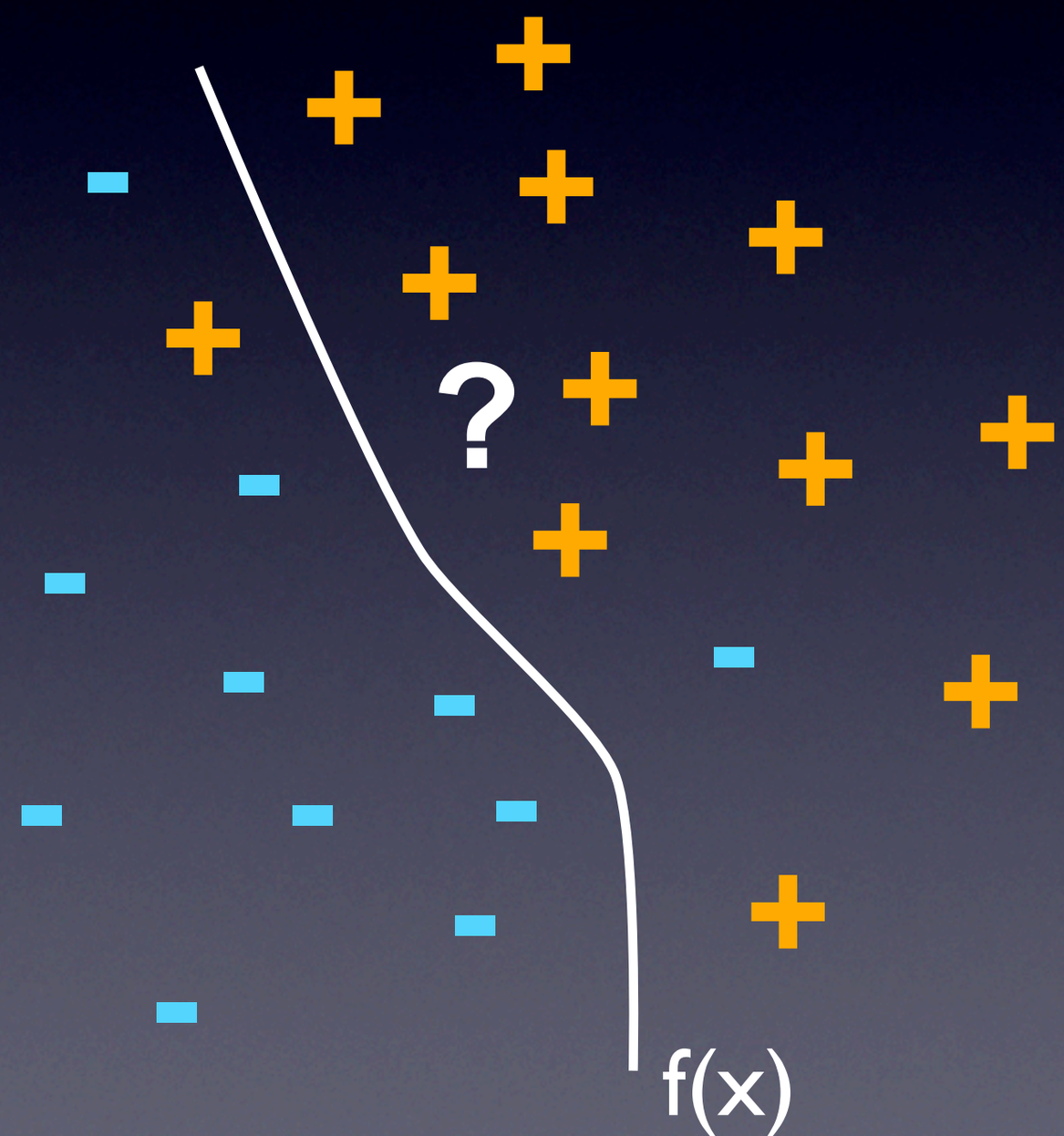
Supervised learning



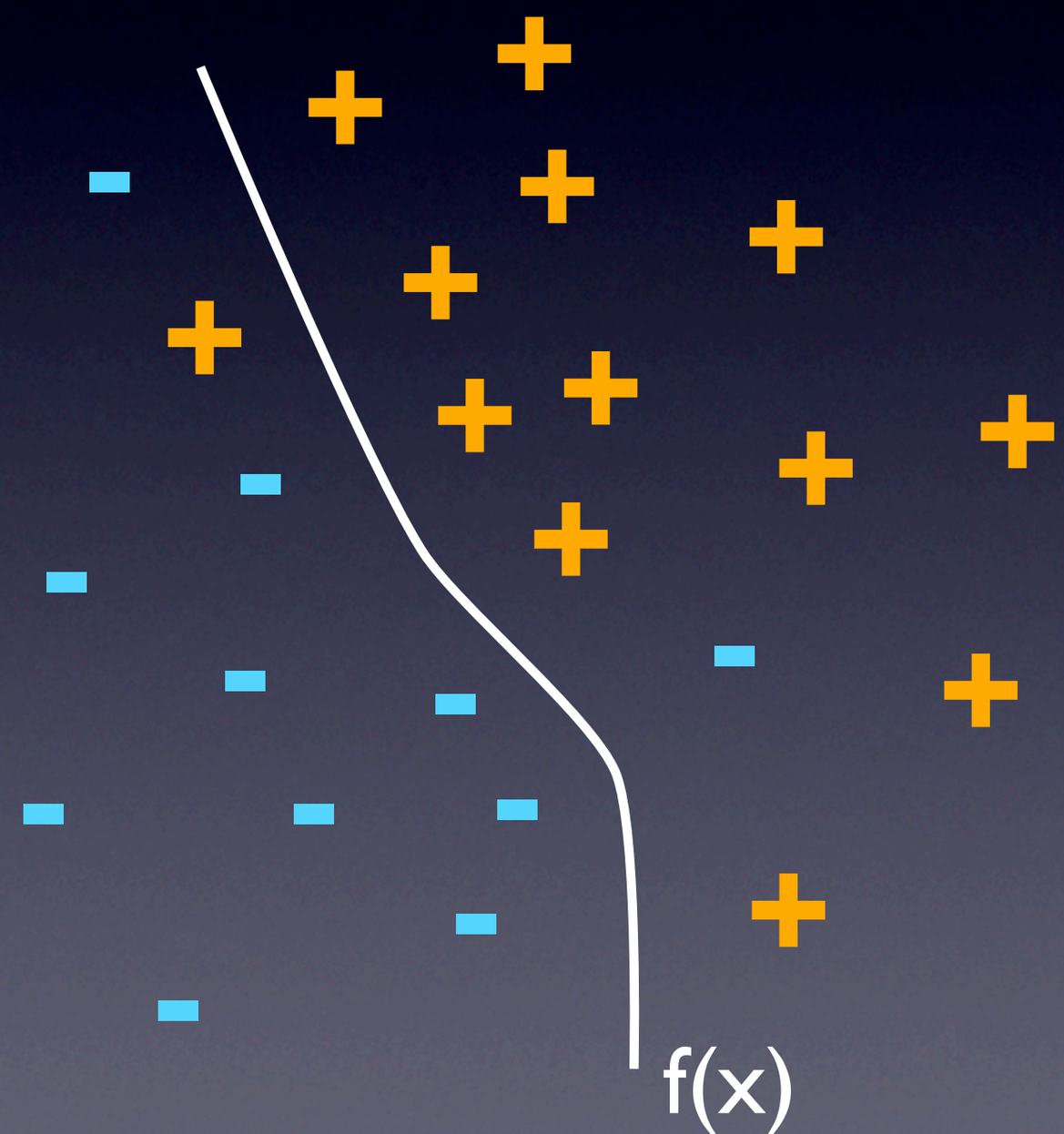
Supervised learning



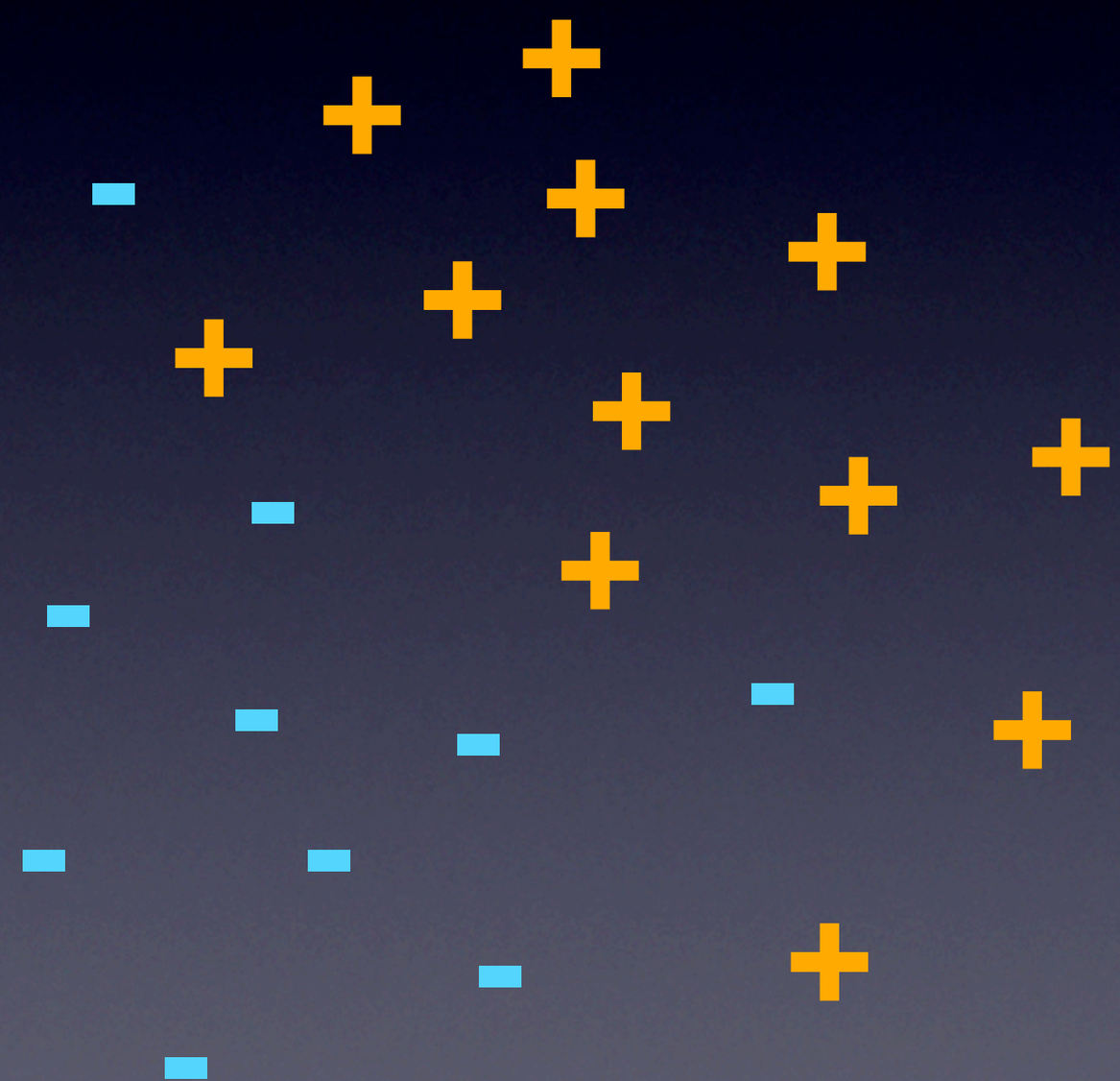
Supervised learning



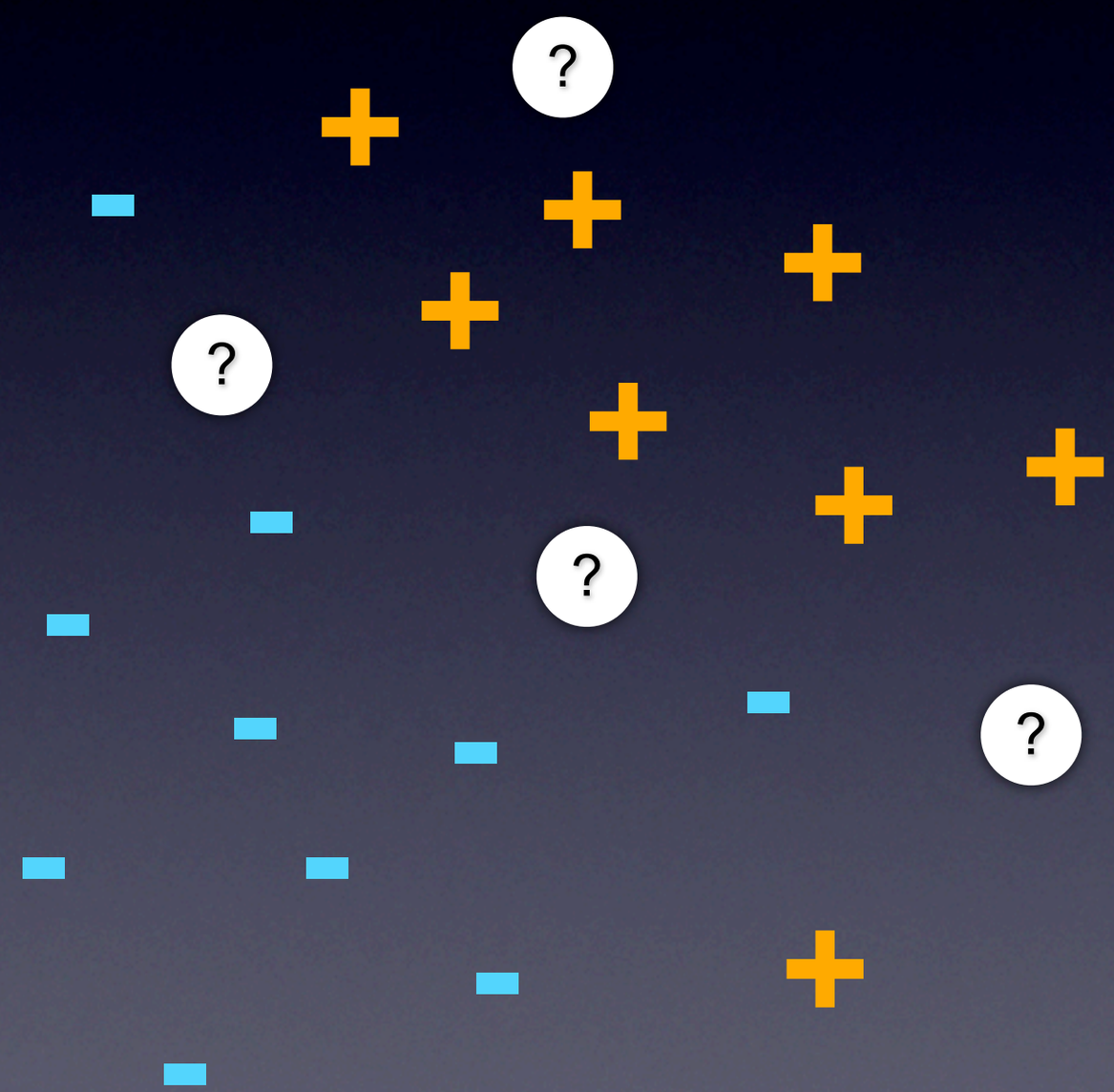
Supervised learning



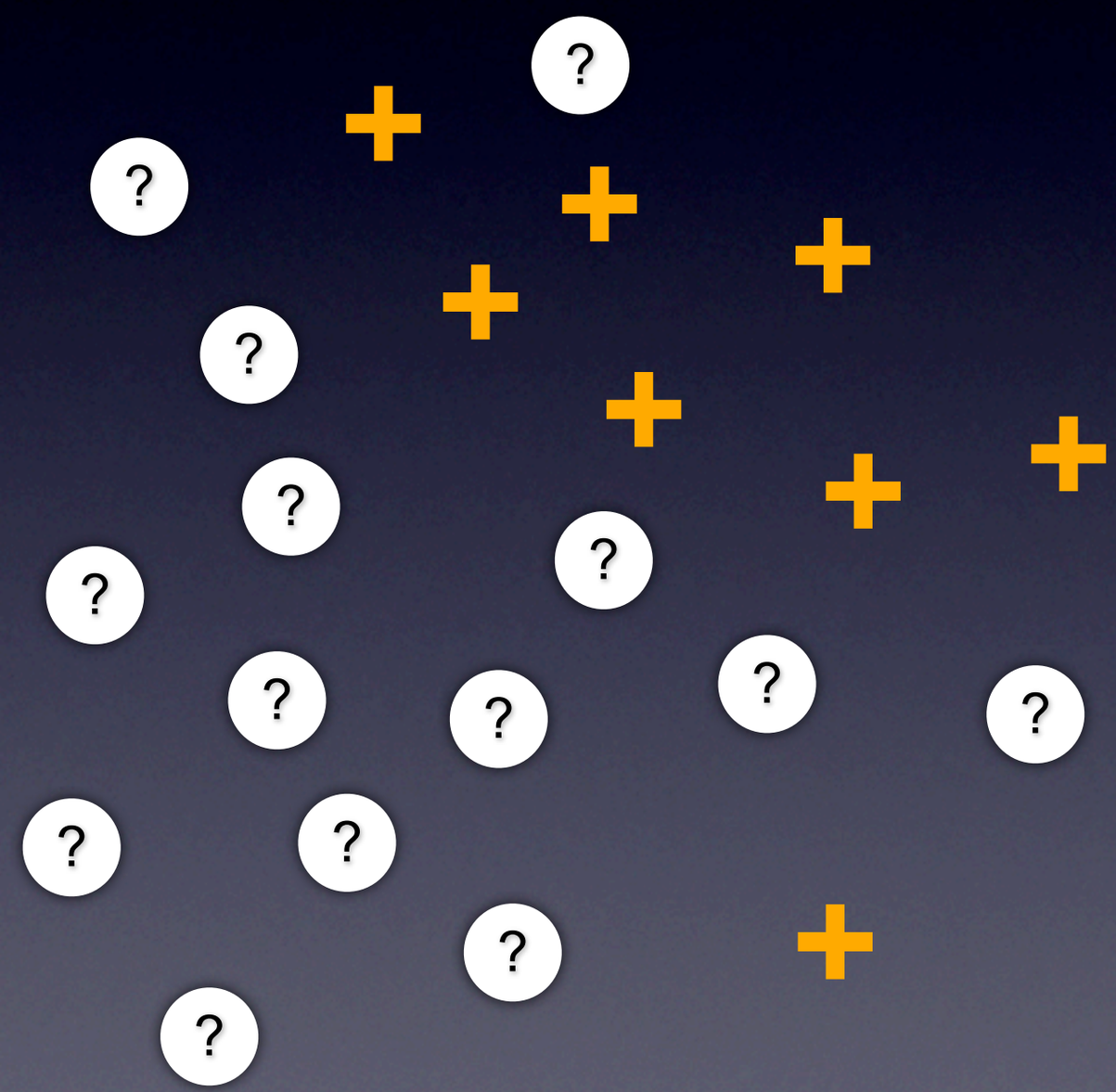
Supervised learning with unlabeled data



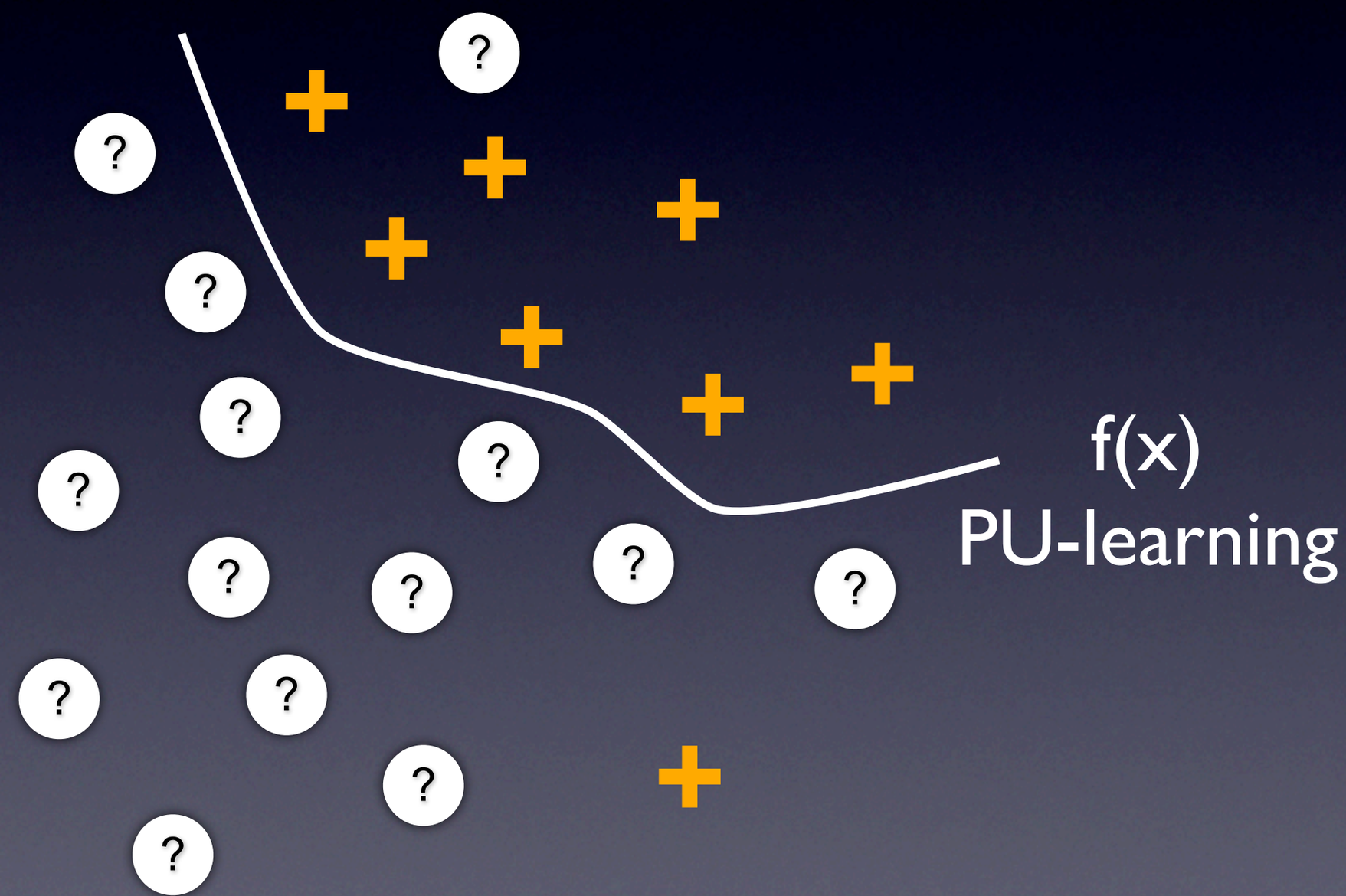
Supervised learning with unlabeled data



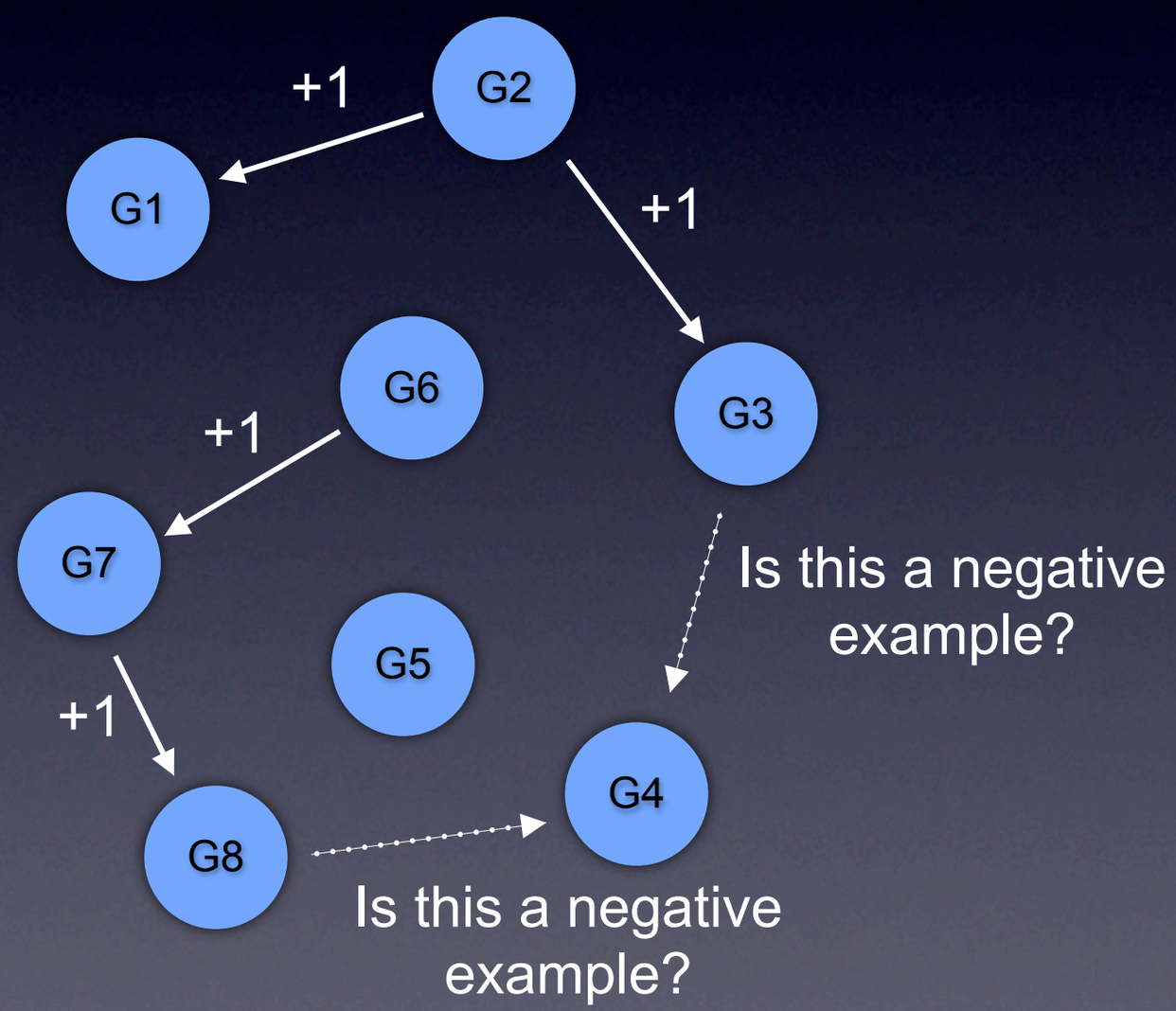
Supervised learning with unlabeled data



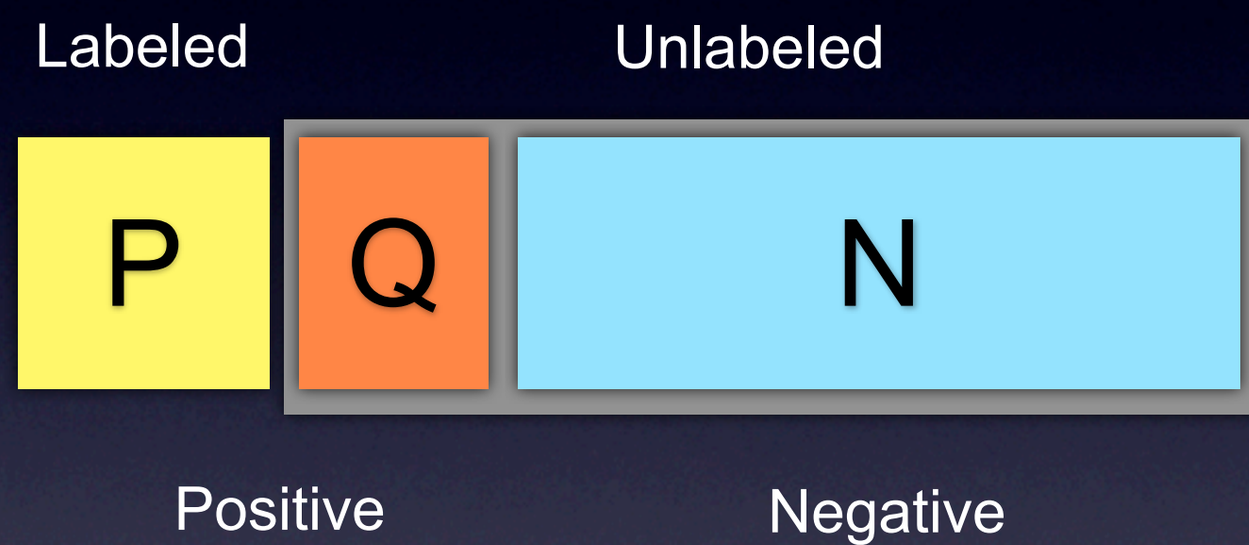
Supervised learning with unlabeled data



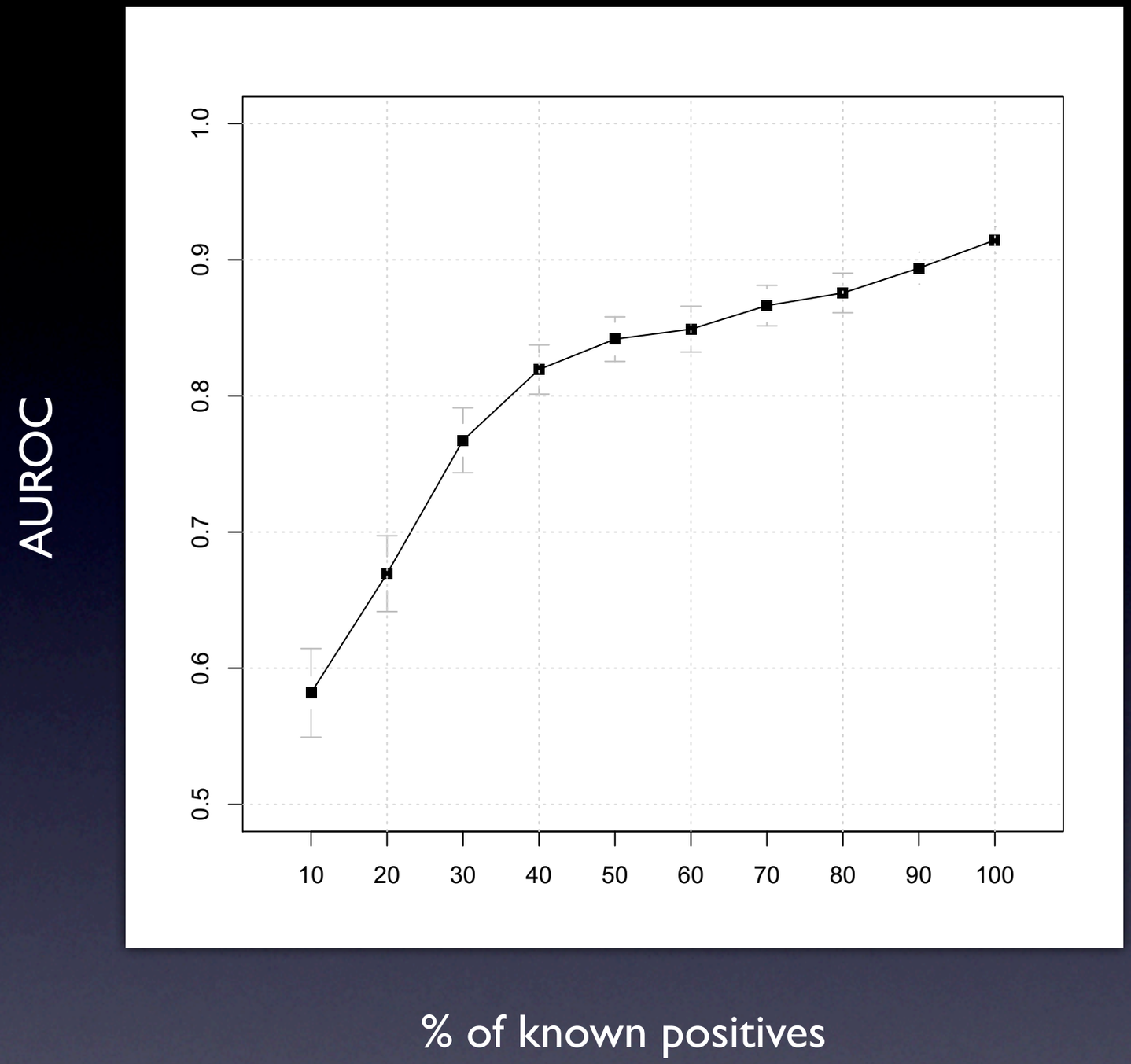
Supervised learning of gene regulatory networks



Training set



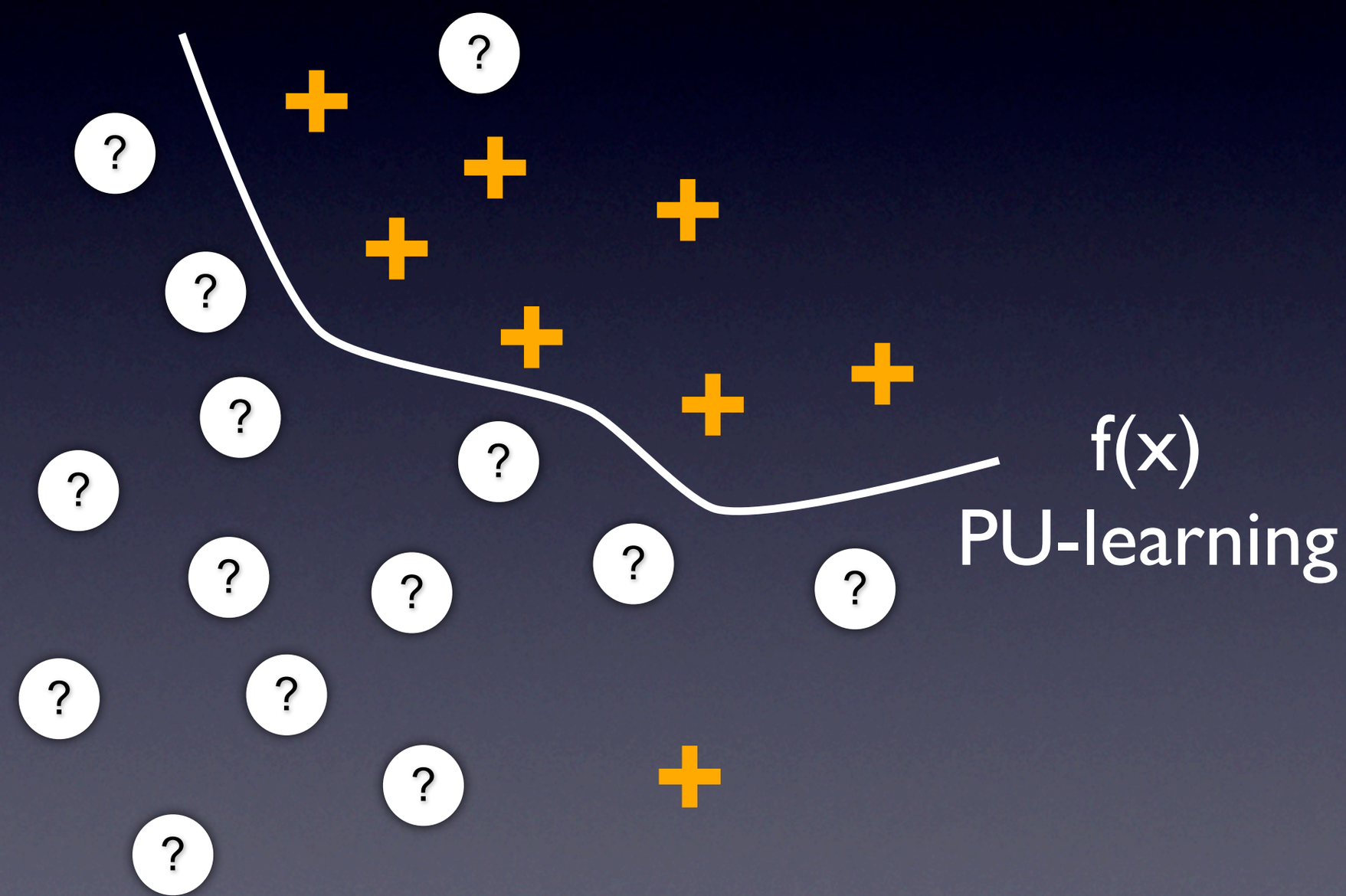
% of Known Positives $\frac{|P|}{|P \cup Q|}$



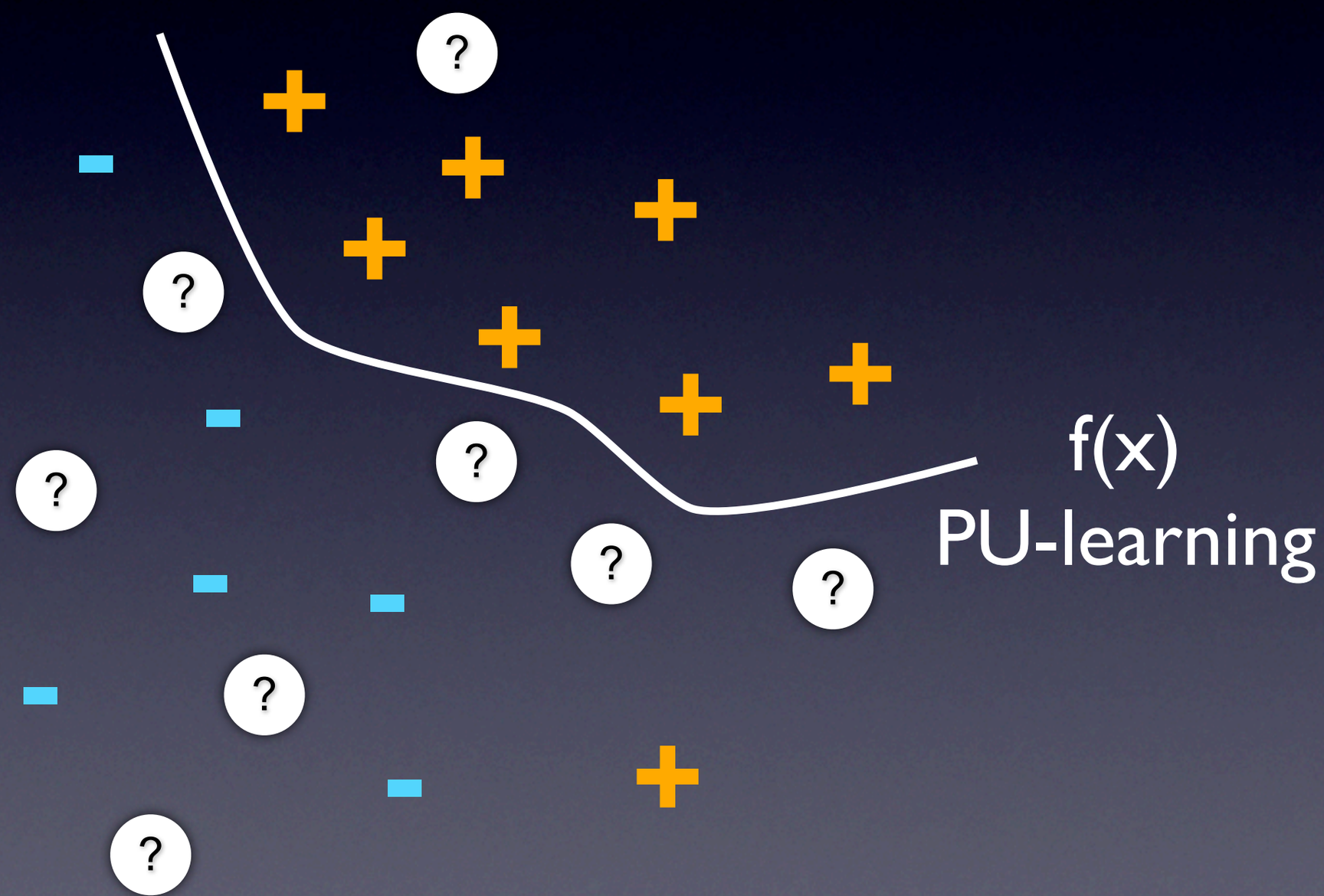
Effect of PU-learning

E.coli dataset [J.J. Faith *et al.*, 2007]

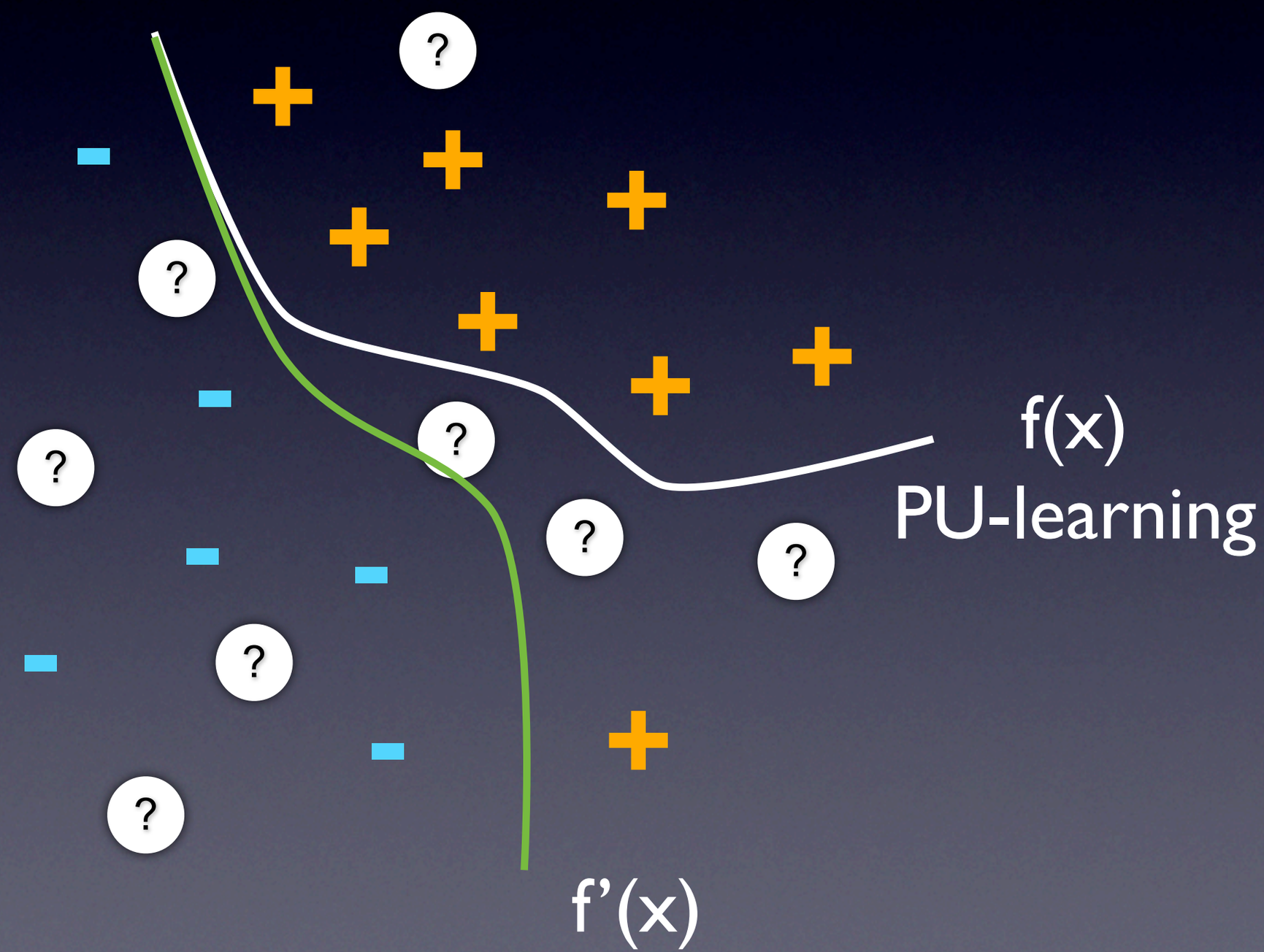
Reliable negative selection



Reliable negative selection



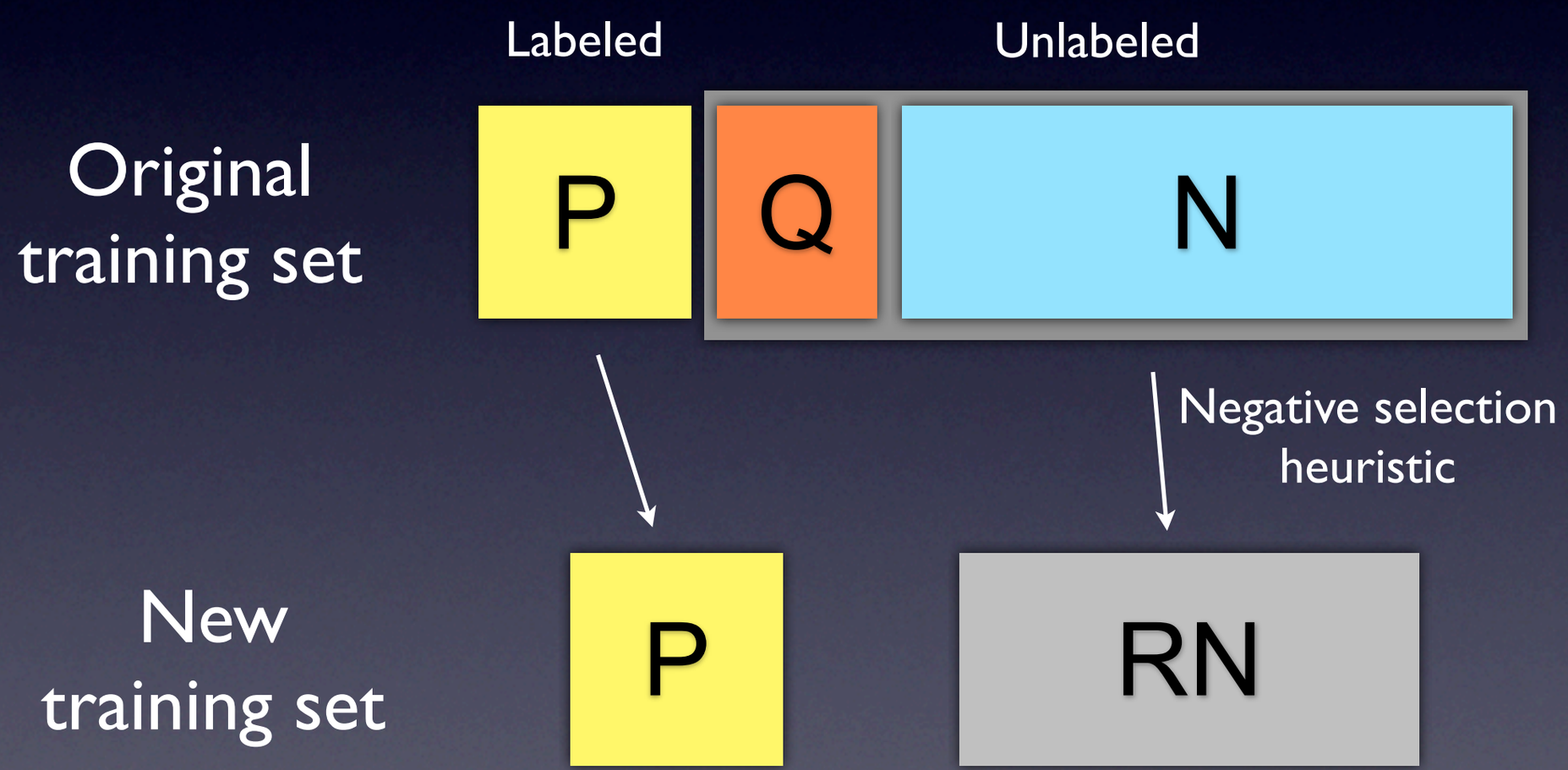
Reliable negative selection



Reliable negative selection in text mining

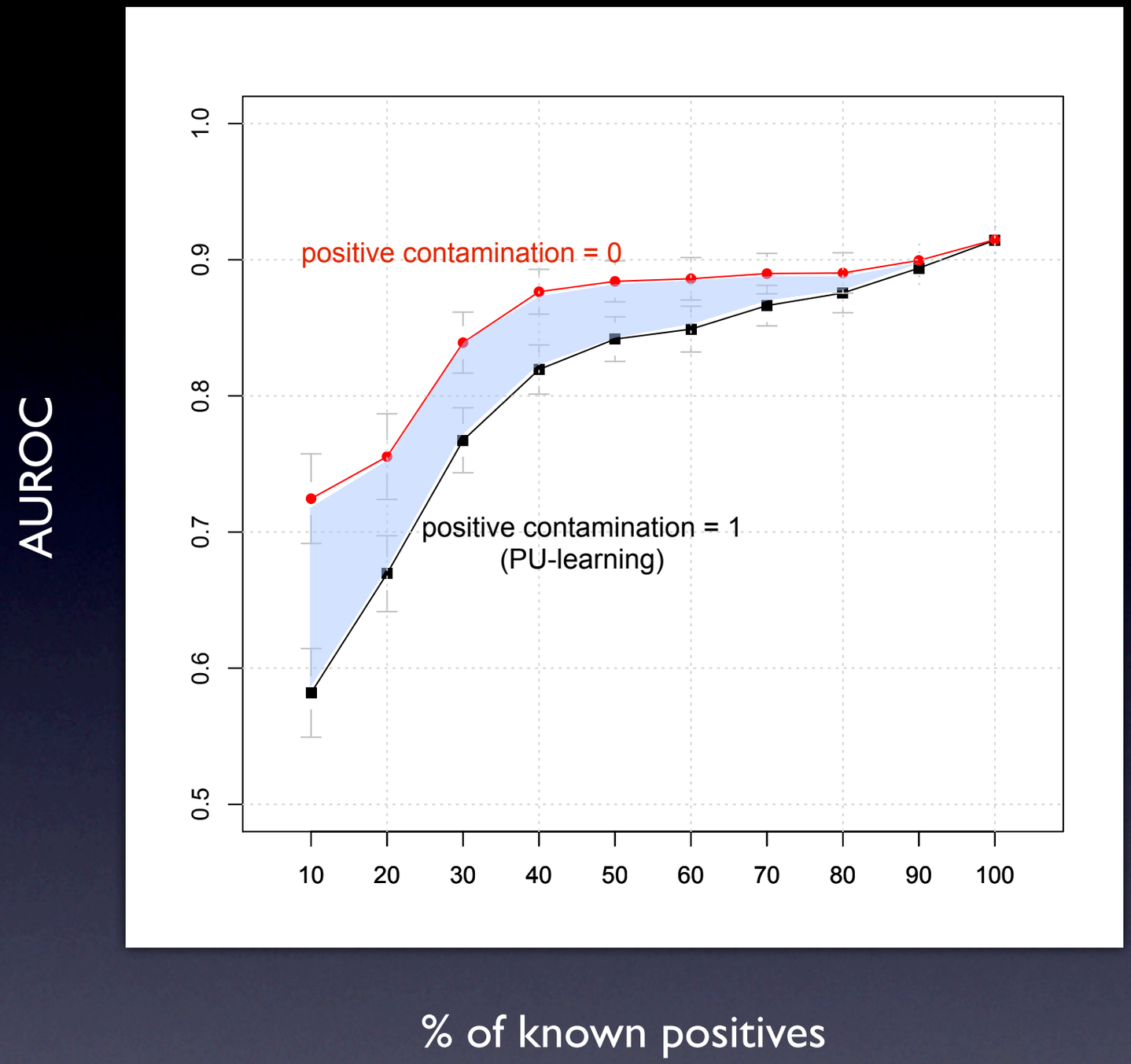
- B. Liu et al. Building Text Classifiers Using Positive and Unlabeled Examples, in ICDM 2003
- Yu et al. PEBL: Positive Example Based Learning for Web Page Classification Using SVM, in KDD 2002
- Denis et al. Text classification from positive and unlabeled Examples, in IPMU 2002

Methods based on reliable negative selection



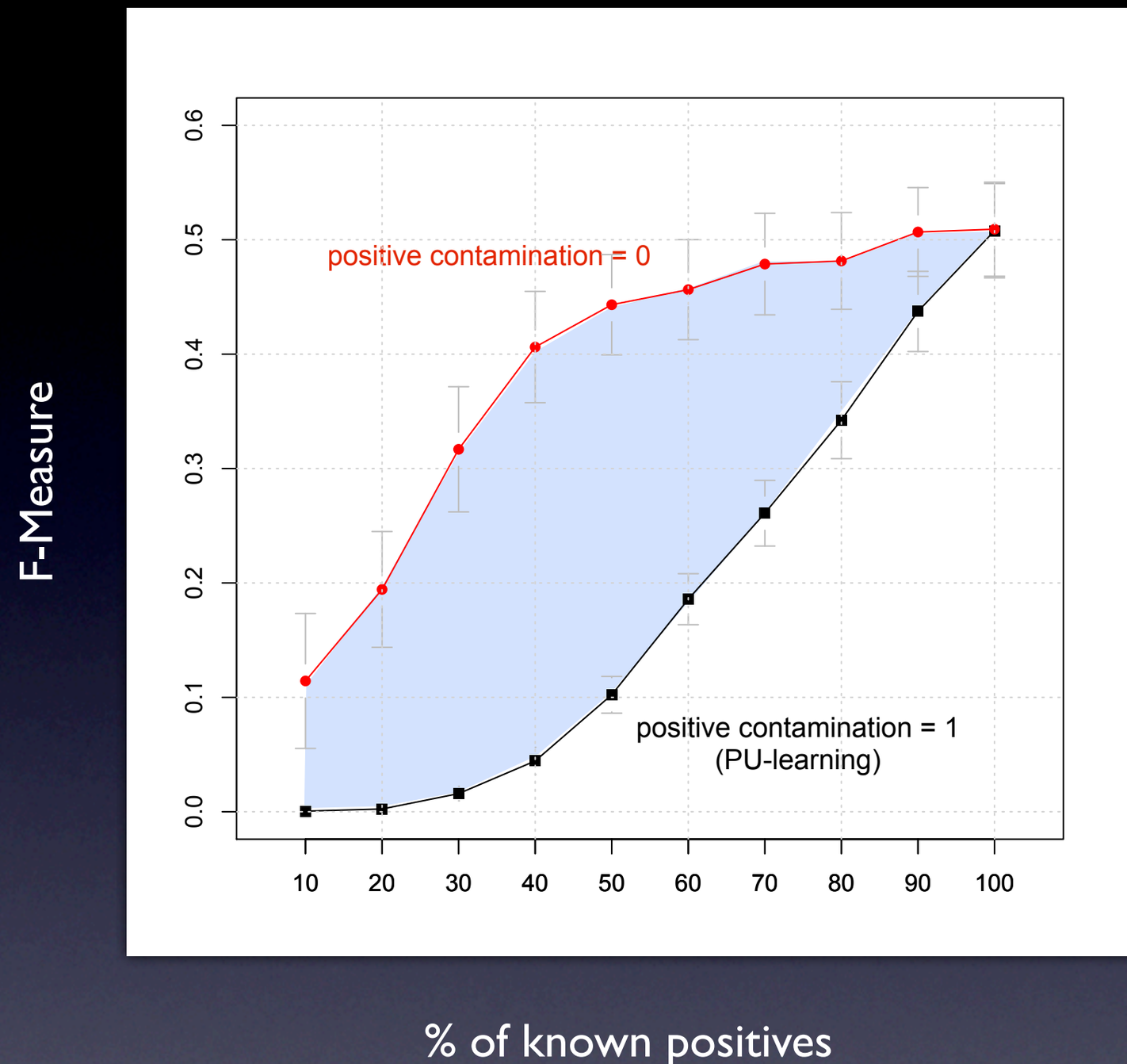
Quality of RN

- RN could be contaminated with positives embedded in unlabeled data
- The <u>fraction of positive contamination</u> is the ratio between the number of positives in RN and the total number of unknown positives |Q|



Effect of positive contamination

E.coli dataset [J.J. Faith *et al.*, 2007]



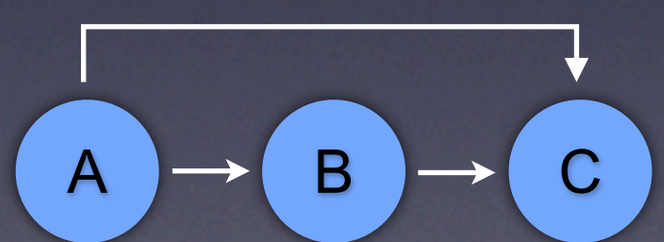
Effect of positive contamination

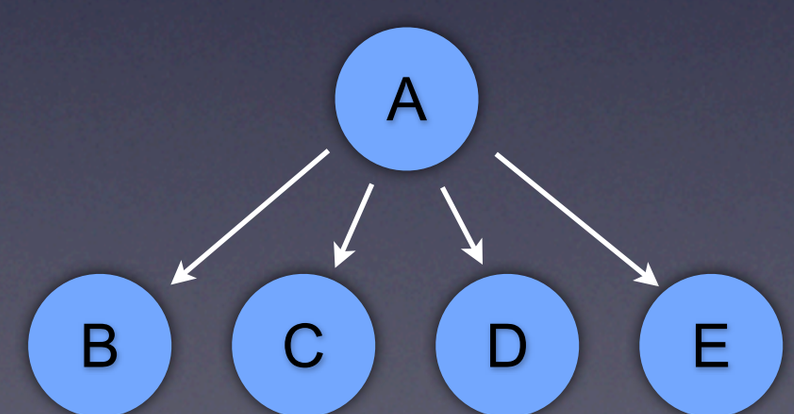
E.coli dataset [J.J. Faith *et al.*, 2007]

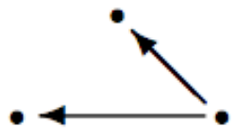
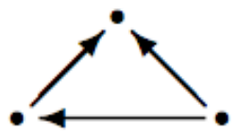
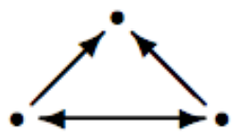
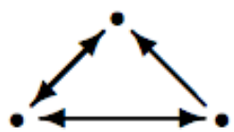
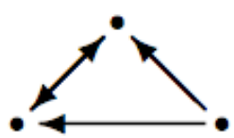
Network topology based heuristics

Network motifs

Network motifs are small connected subnetworks a network exhibits in a significant higher or lower occurrences than would be expected just by chance





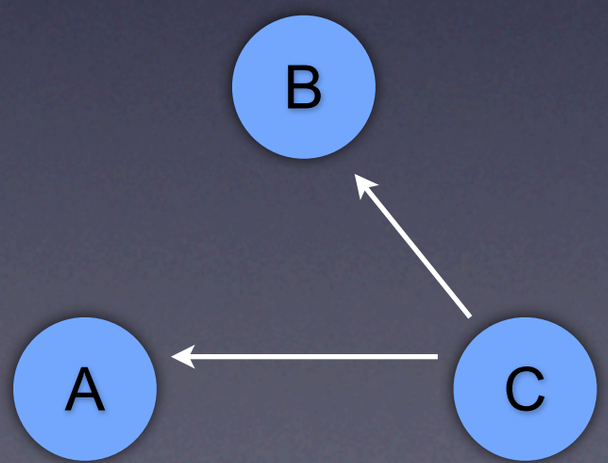
No.	Motif	E. coli Z-score	E. coli Freq.	S. cerevisiae Z-score	S. cerevisiae Freq.
1		20.343	97.467%	16.918	93.82%
2		13.295	0.318%	10.827	0.298%
3		14.401	0.105%	27.202	0.032%
4		2.058	<0.001%	4.233	<0.001%
5		4.533	0.004%	4.068	<0.001%

B. Goemann, E. Wingender, and A. P. Potapov, “An approach to evaluate the topological significance of motifs and other patterns in regulatory networks.” *BMC System Biology*, vol. 3, no. 53, May 2009.

S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, “Network motifs in the transcriptional regulation network of escherichia coli,” *Nature Genetics*, vol. 31, no. 1, pp. 64–68, May 2002.

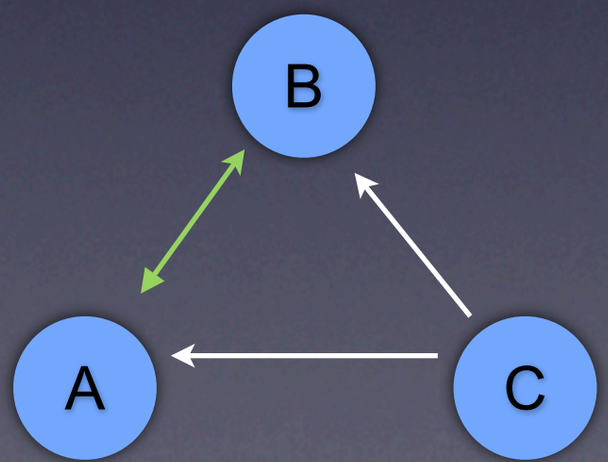
Network Motifs Heuristic

- For each three genes sub networks T:
- If matches a network motifs M then considers all connections not present in M as negatives



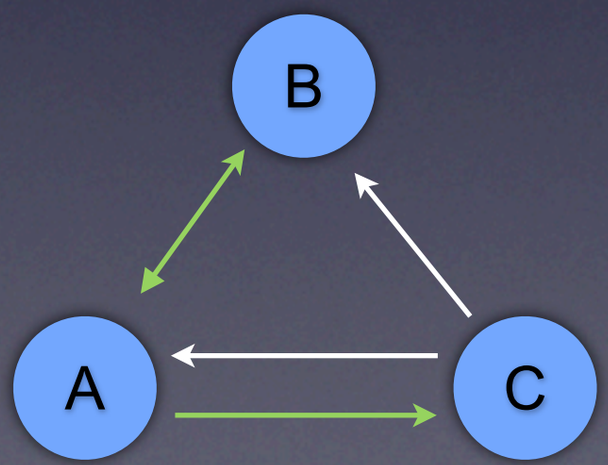
Network Motifs Heuristic

- For each three genes sub networks T:
- If matches a network motifs M then considers all connections not present in M as negatives



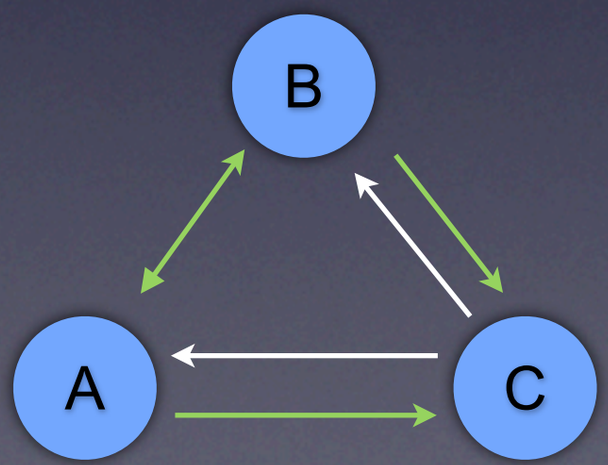
Network Motifs Heuristic

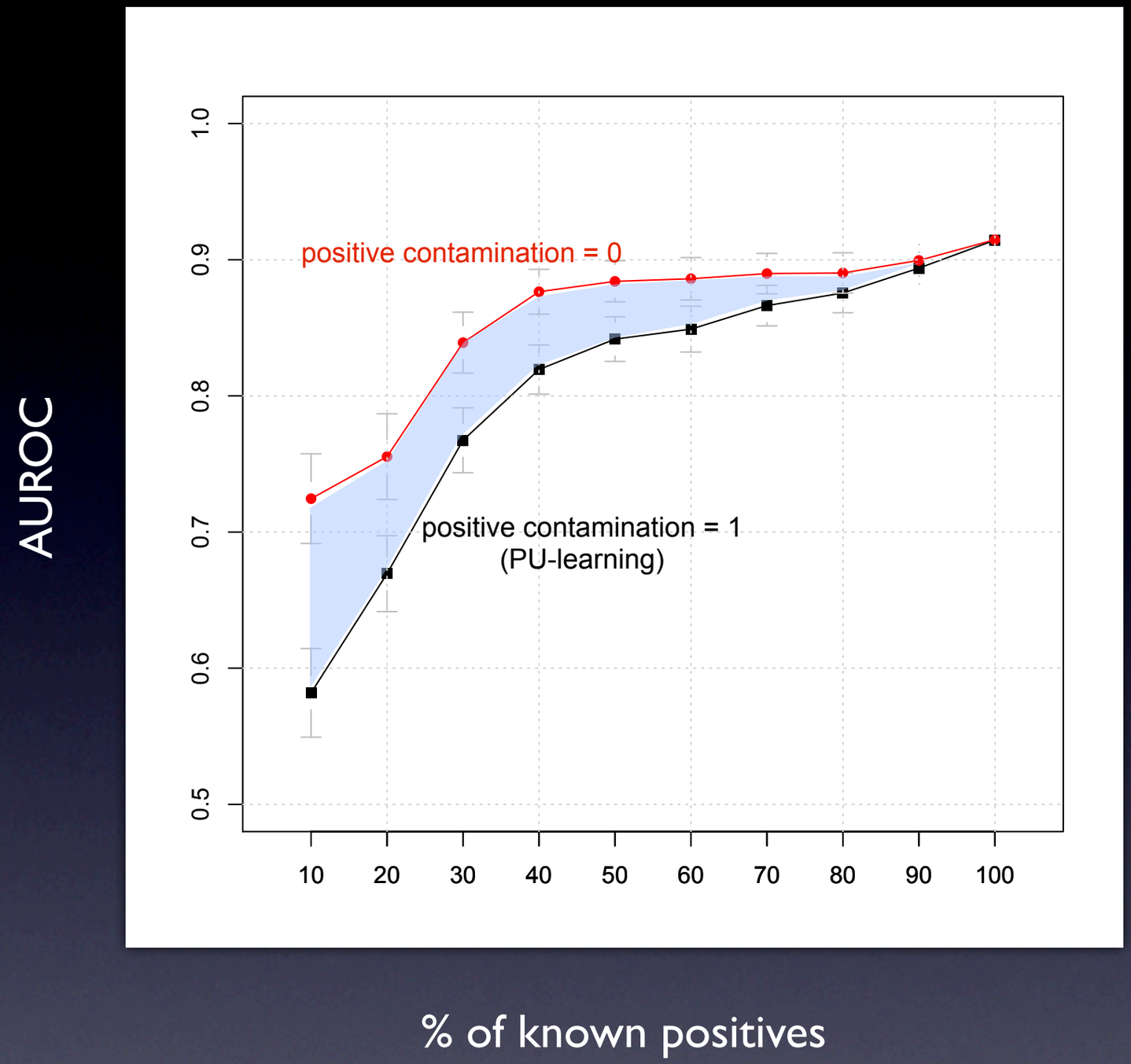
- For each three genes sub networks T:
- If matches a network motifs M then considers all connections not present in M as negatives



Network Motifs Heuristic

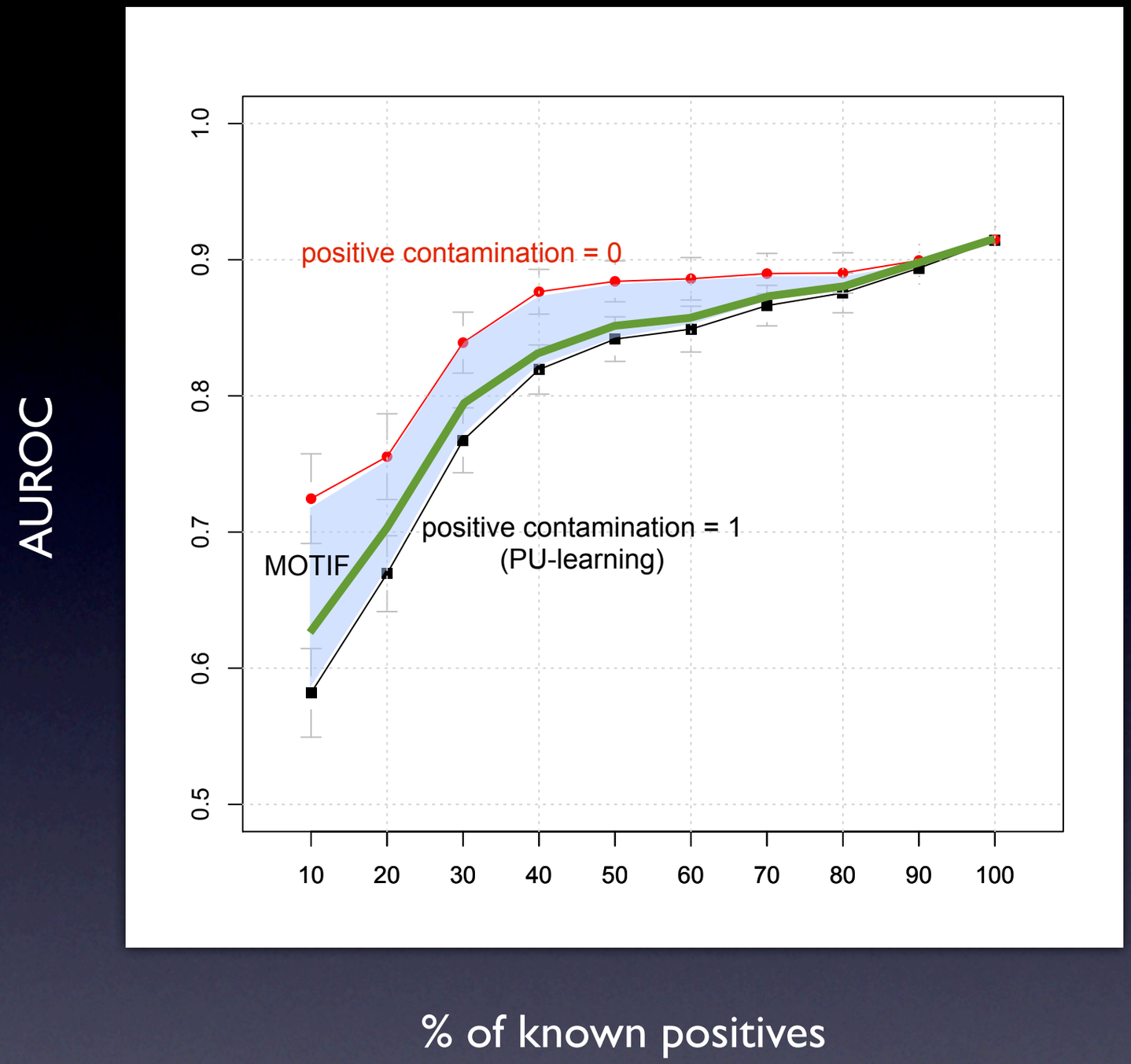
- For each three genes sub networks T:
- If matches a network motifs M then considers all connections not present in M as negatives





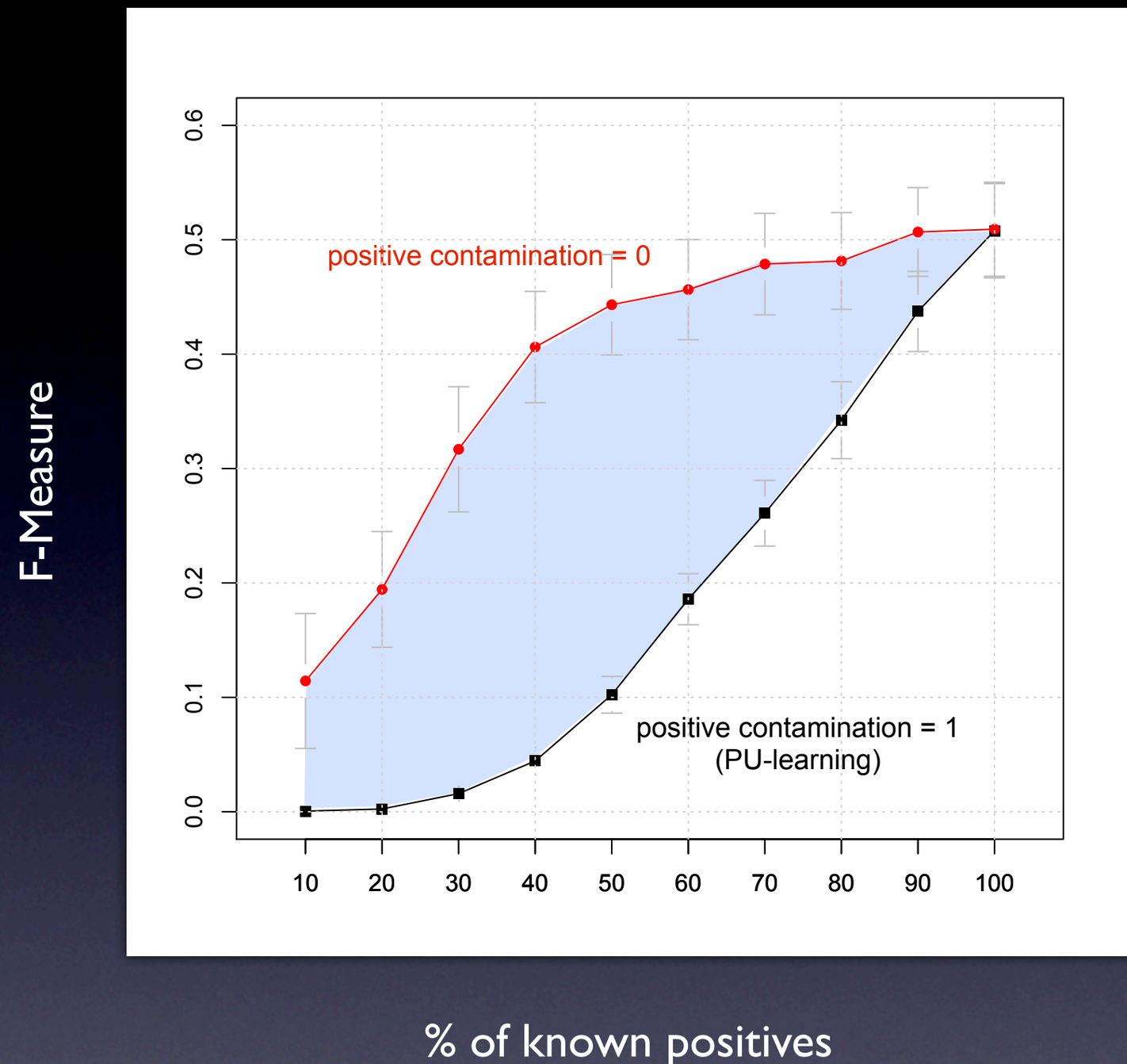
MOTIF selection performance

E.coli dataset [J.J. Faith *et al.*, 2007 and RegulonDB]



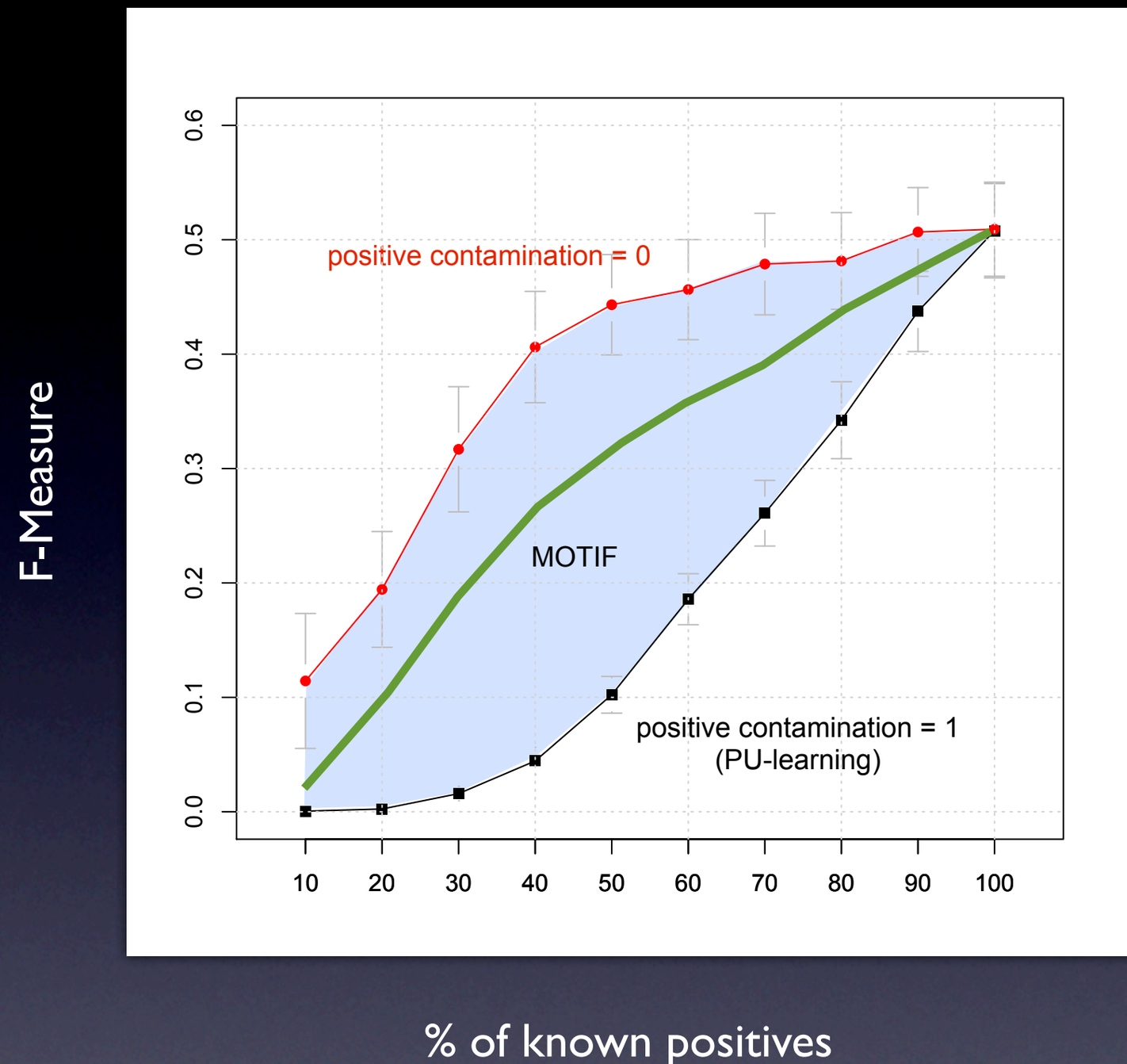
MOTIF selection performance

E.coli dataset [J.J. Faith *et al.*, 2007 and RegulonDB]



Effect of positive contamination

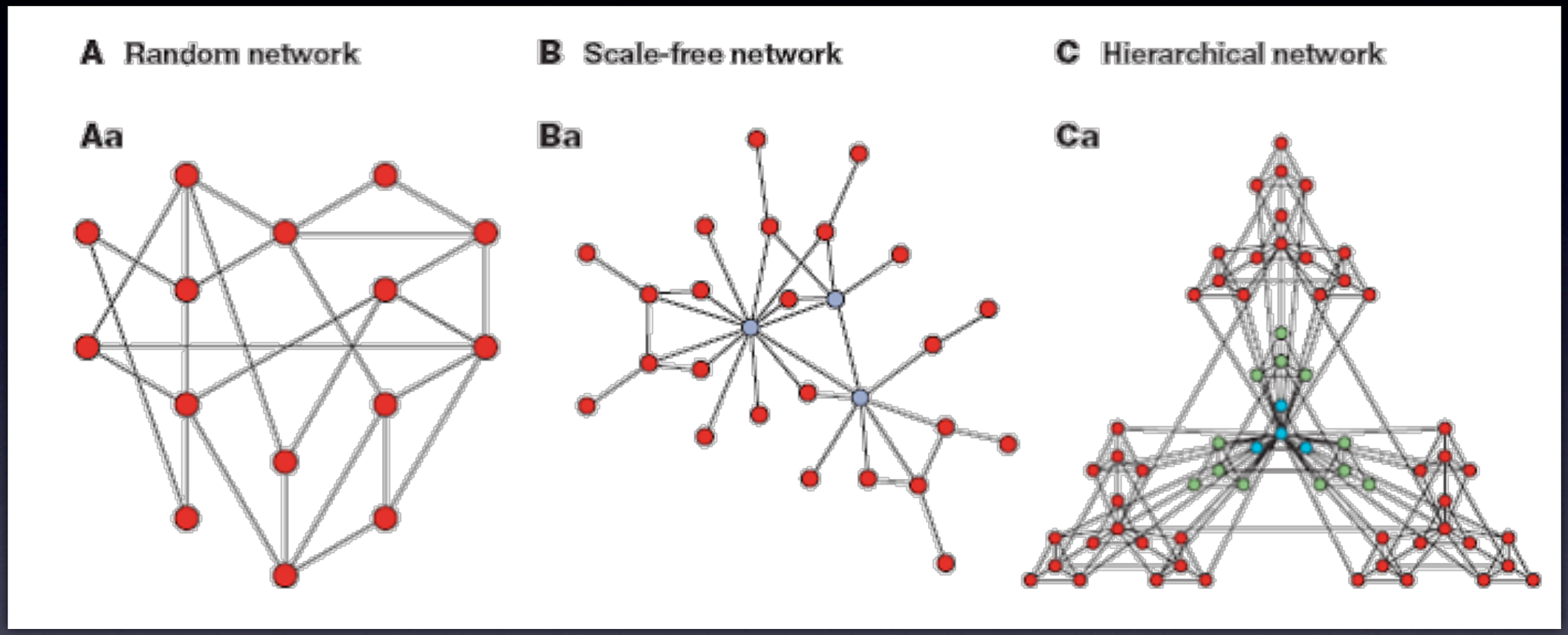
E.coli dataset [J.J. Faith *et al.*, 2007]



Effect of positive contamination

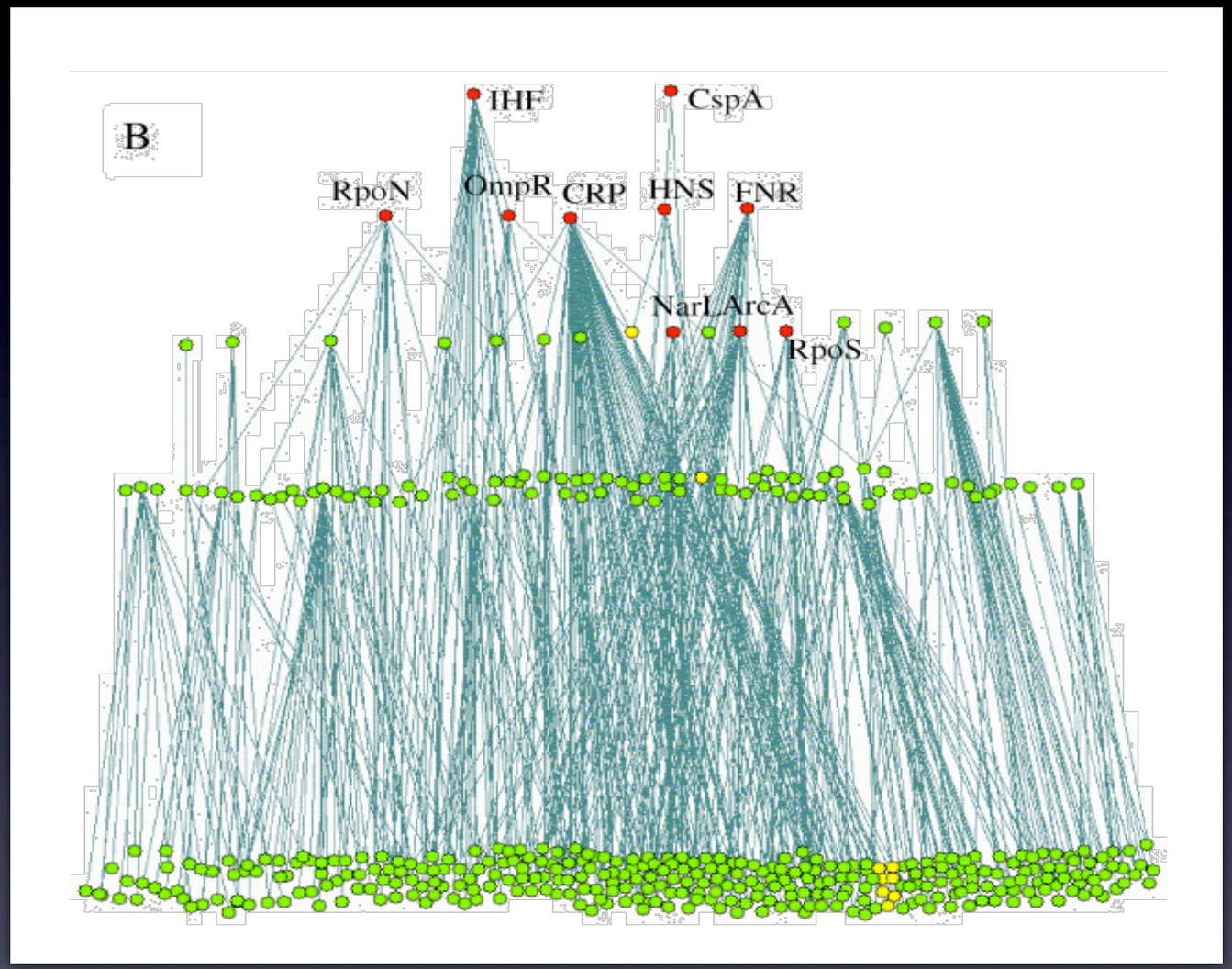
E.coli dataset [J.J. Faith *et al.*, 2007]

Scale free networks



Albert-László Barabási and Zoltán N. Oltvai
Network biology: Understanding the cell's functional organization
Nature Reviews Genetics 5, 101-113 (2004)

Hierarchical networks

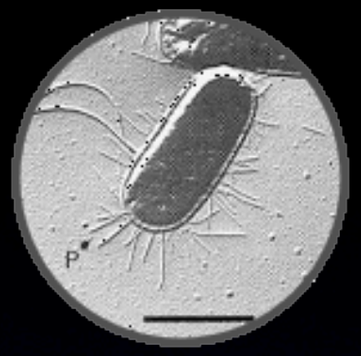


Hong-Wu Ma, Jan Buer, and An-Ping Zeng

Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach

BMC Bioinformatics 2004 5:199

Experimental data



- 445 Affymetrix Antisense2 microarray expression profiles for 4345 genes of E.coli [J.J. Faith *et al.*, 2007]
- Data were standardized (i.e. zero mean unit standard deviation)
- Regulations extracted from RegulonDB (v. 5) between 154 Transcription Factors and 1211 genes

Summary and conclusions

- Learning gene regulations is affected by the problem of learning from positive only data
- At least for E.coli
  - The study of positive contamination shows that there is room for new heuristics
  - Topology based heuristics (eg. motifs) have shown promising results.
- Open issues arise on higher level organisms where gene interactions are more complex