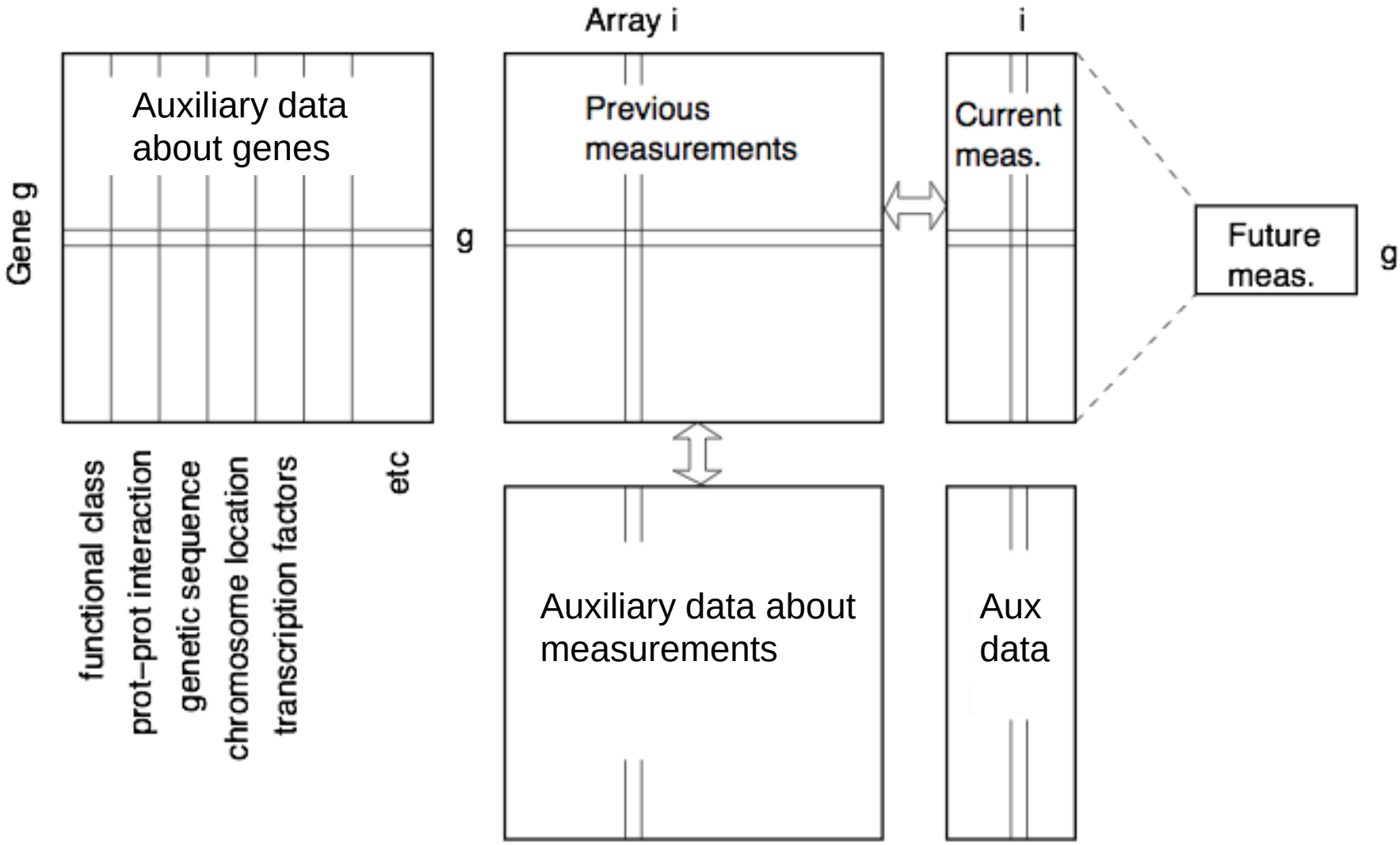
Learning and retrieval from multiple sources

Samuel Kaski

Learning from multiple sources

Learning from multiple sources



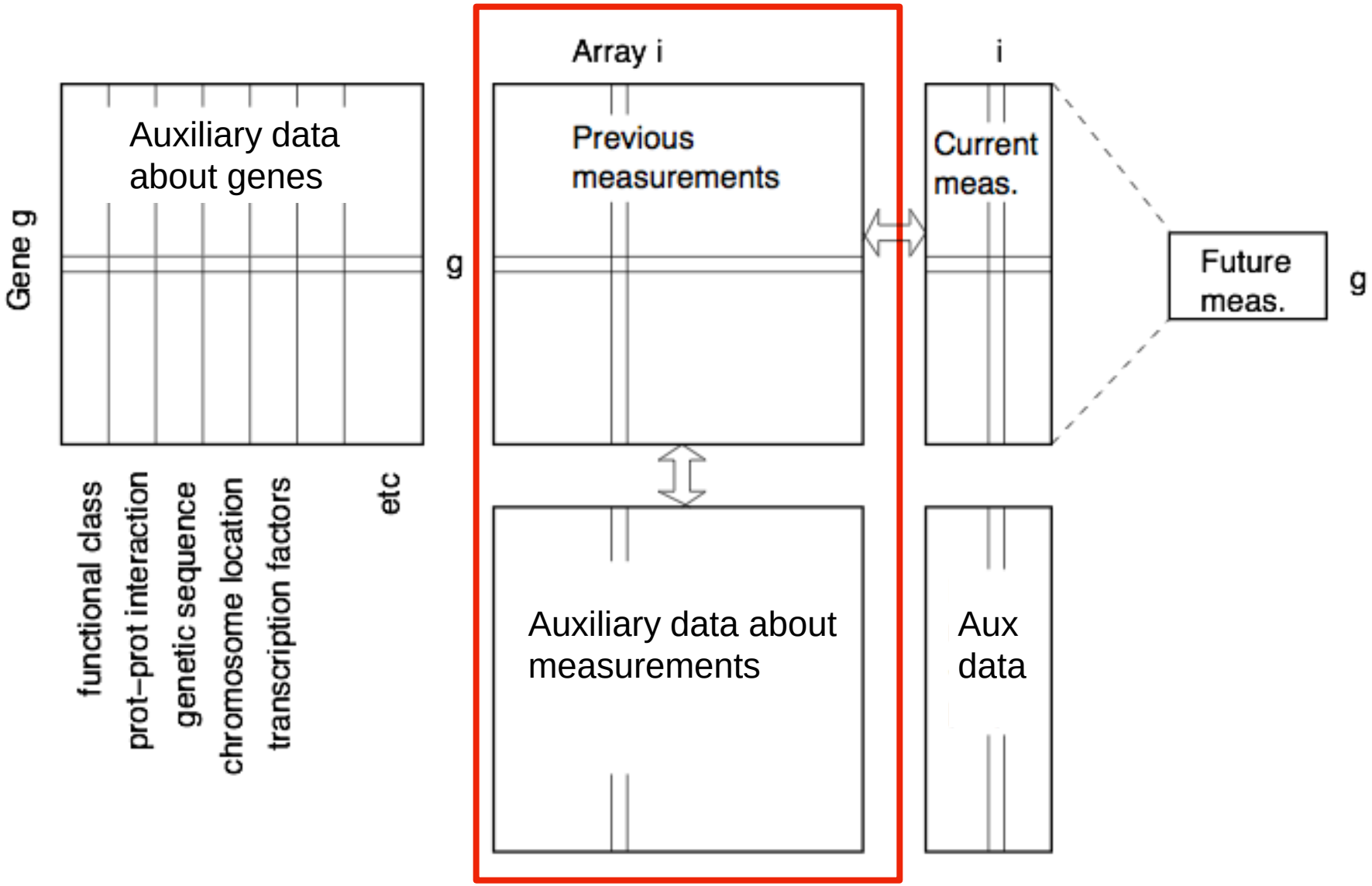
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

1. Focus on modeling relevant things:

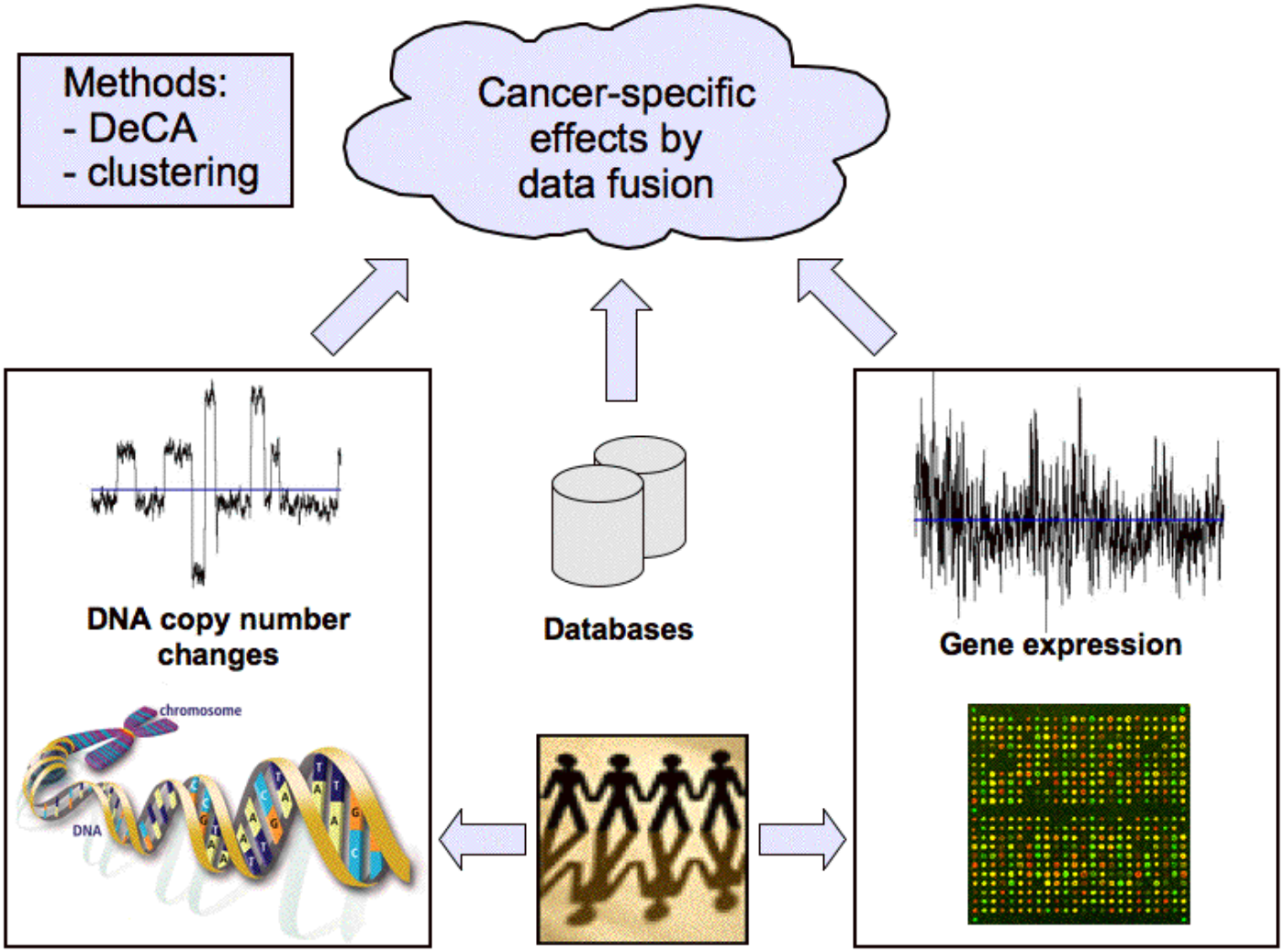
Unsupervised multi-view learning

Aalto University

UNIVERSITY OF HELSINKI

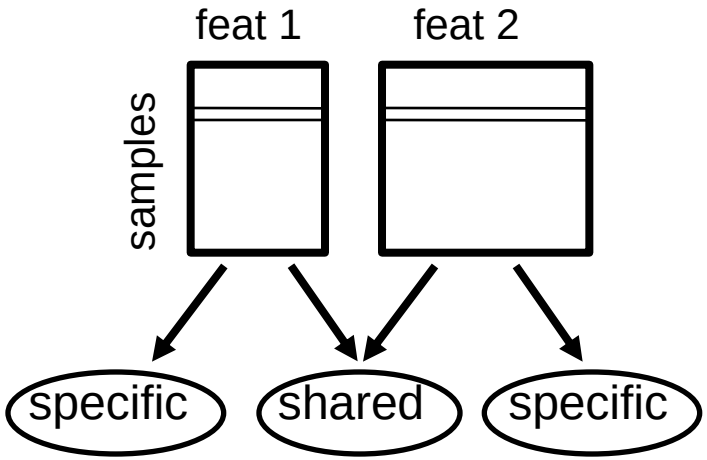

Array i
i
Auxiliary data about genes
Gene g
g
Previous measurements
Current meas.
Future meas.
g
functional class
prot–prot interaction
genetic sequence
chromosome location
transcription factors
etc
Auxiliary data about measurements
Aux data

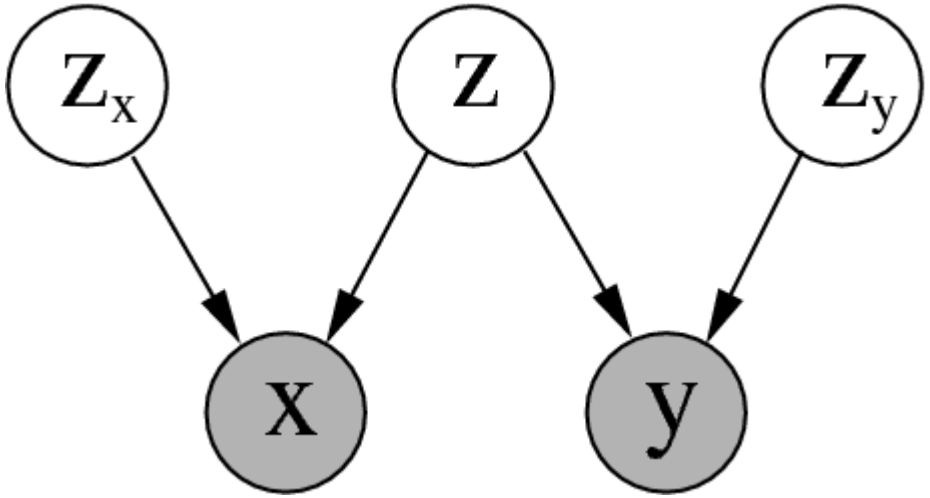
Relevance from co-occurring data: in search for cancer-related genes



Decompose multiple “views” into view-specific and shared components

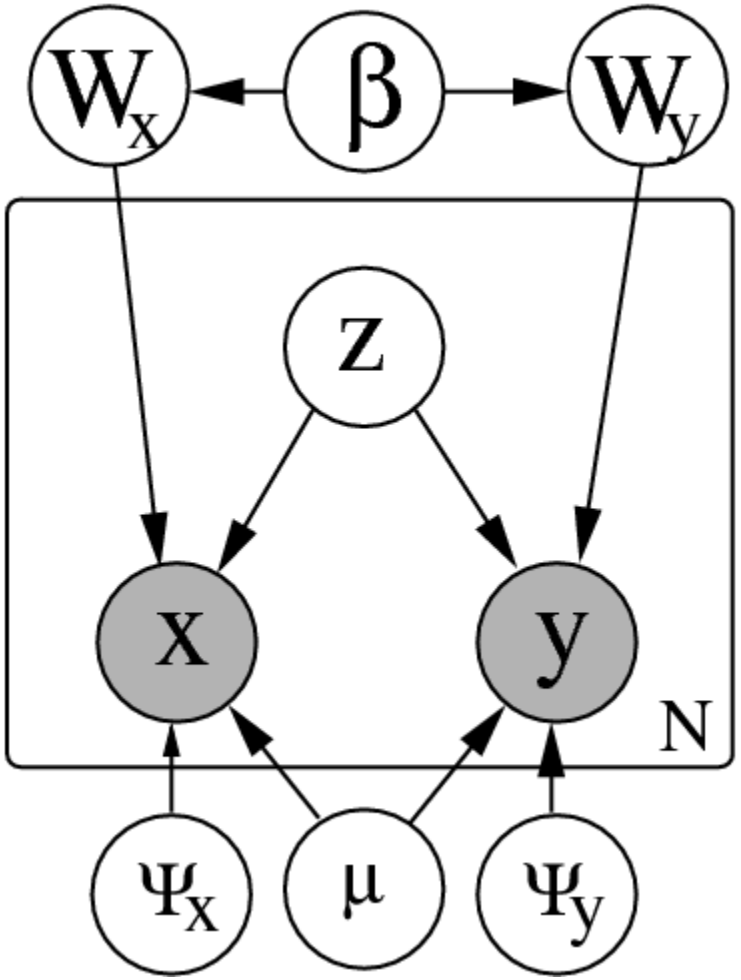
- Motivation #1: Shared or dependent components are *relevant* for both sources
- Motivation #2: Unknown type of noise (=source-specific signal) can be discarded
- Small samples => From dependency maximization to generative models





Local Dependent Components

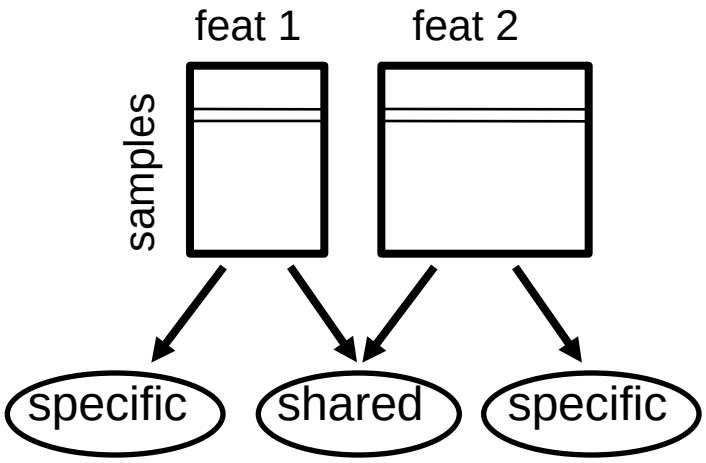
- Assume dependencies are linear only locally
- DP-mixture of Bayesian canonical correlation analyzers
- Marginalize out the specific latent sources

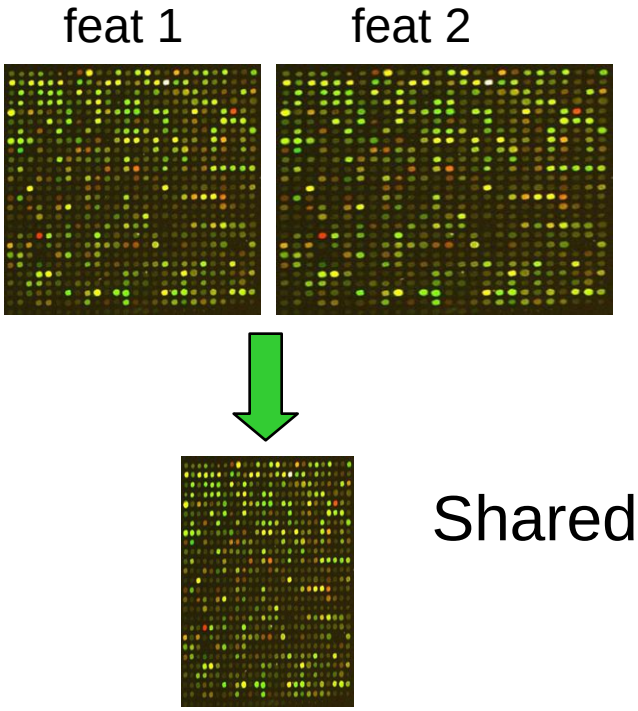


Klami and Kaski, ICML 2007

Preprocessing that preserves what is shared/dependent

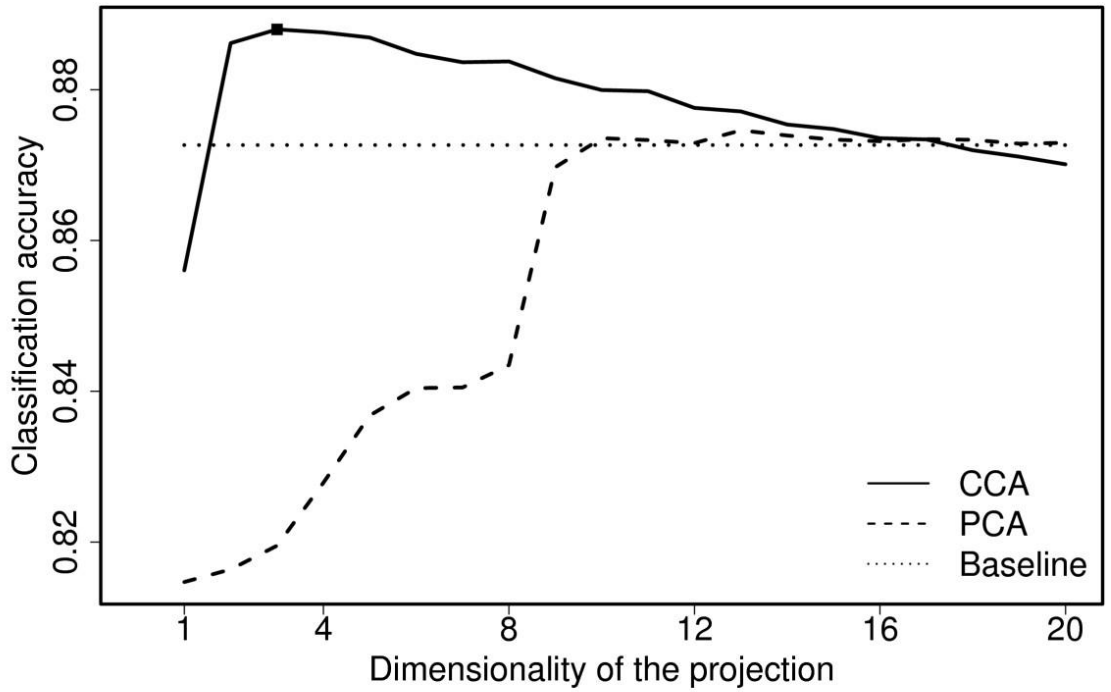
Under simplifying assumptions, the shared signal can be extracted by combining standard CCA components (fast!)





Tripathi et al, BMC Bioinformatics, 2008

Cell cycle regulation



Baseline : Simple column wise concatenation of all data matrices

PCA : PCA of column wise concatenation of all data matrices

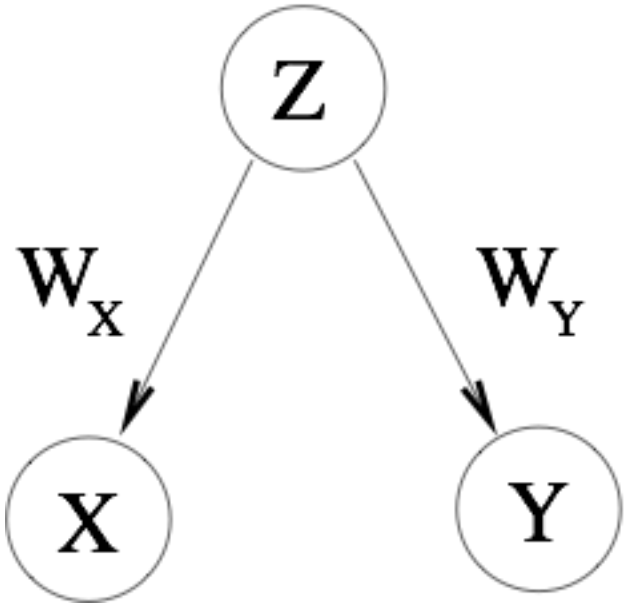
Problems with two-view learning

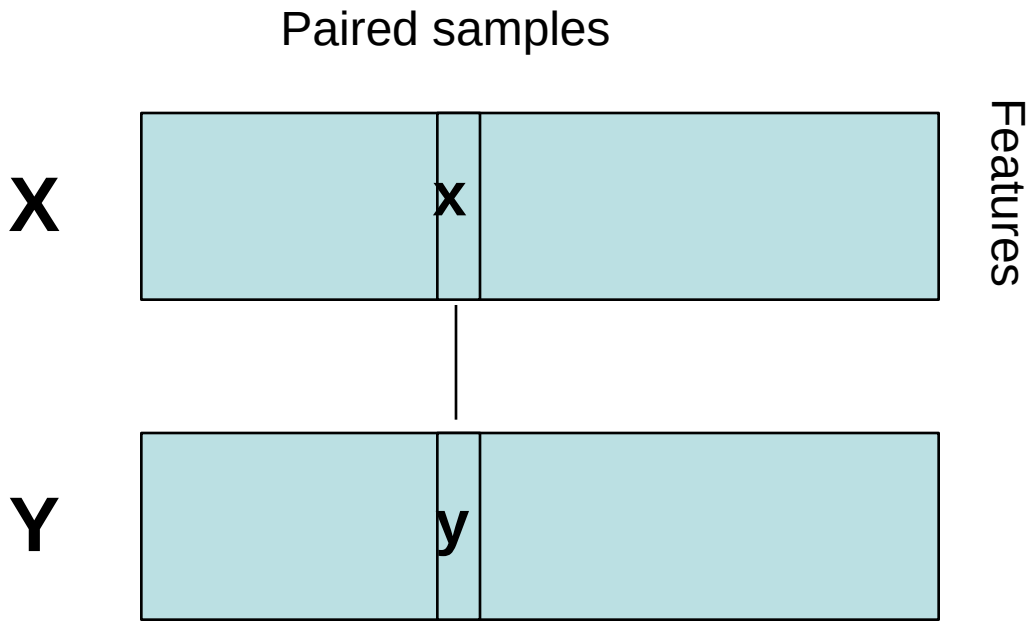
Strength of CCA-type approaches is *invariance* to transformations: CCA computes correlations in an optimized subspace

This turns into a weakness for small data sets: it is too flexible.

=> use prior knowledge to restrict the subspace

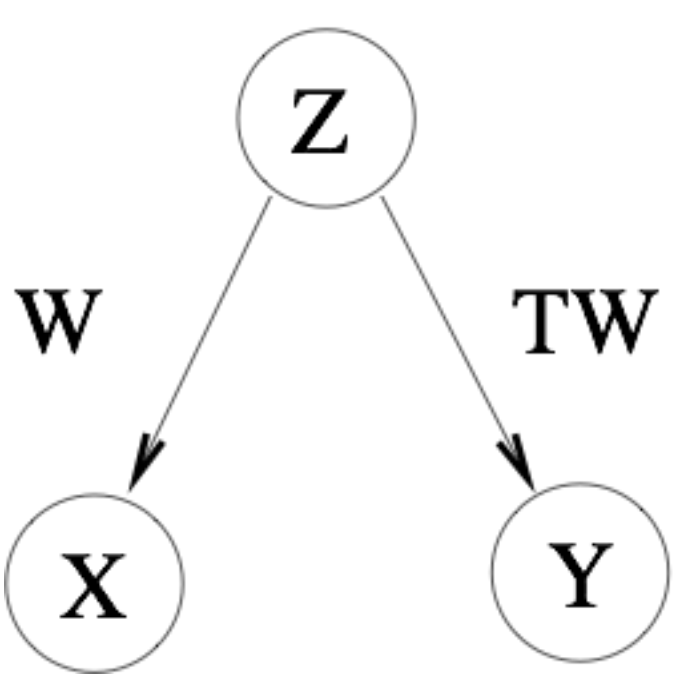
Standard probabilistic CCA

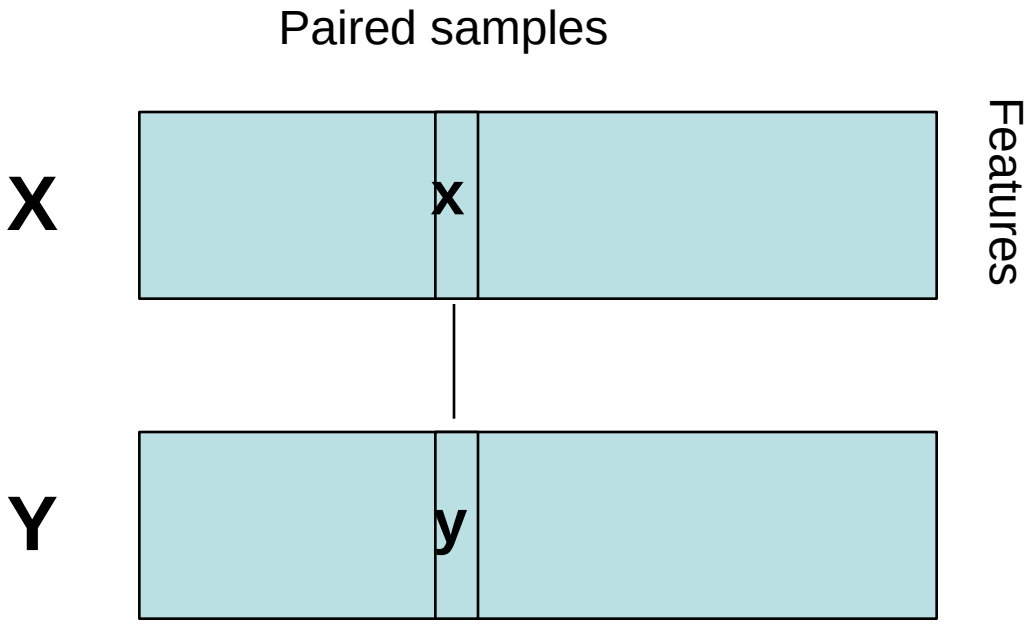




Dependency detection with similarity constraints

Similarity-constrained CCA:





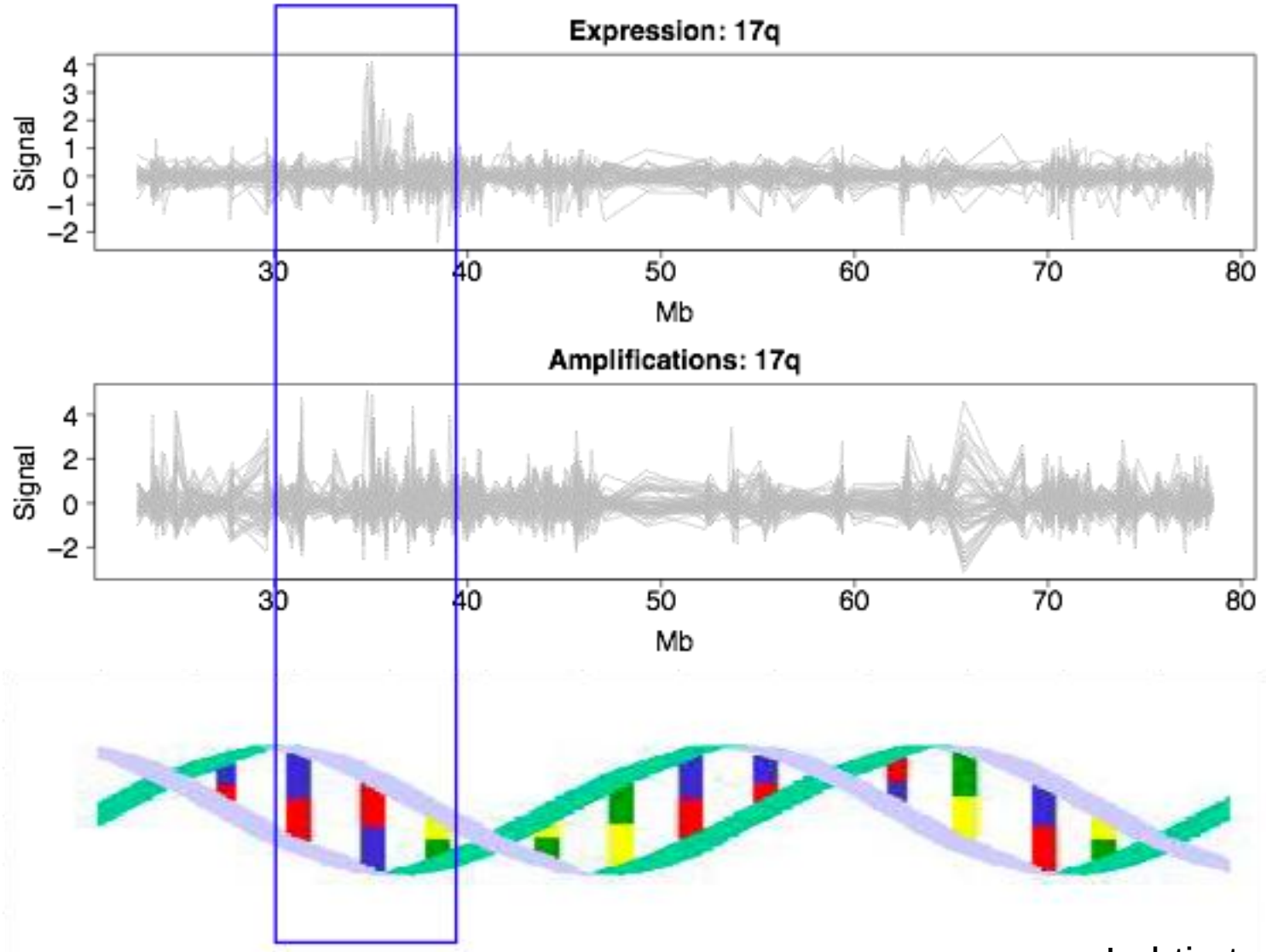
Setting suitable priors for T gives constraints

Lahti et al, MLSP2009

Dependencies between expression and structural variation

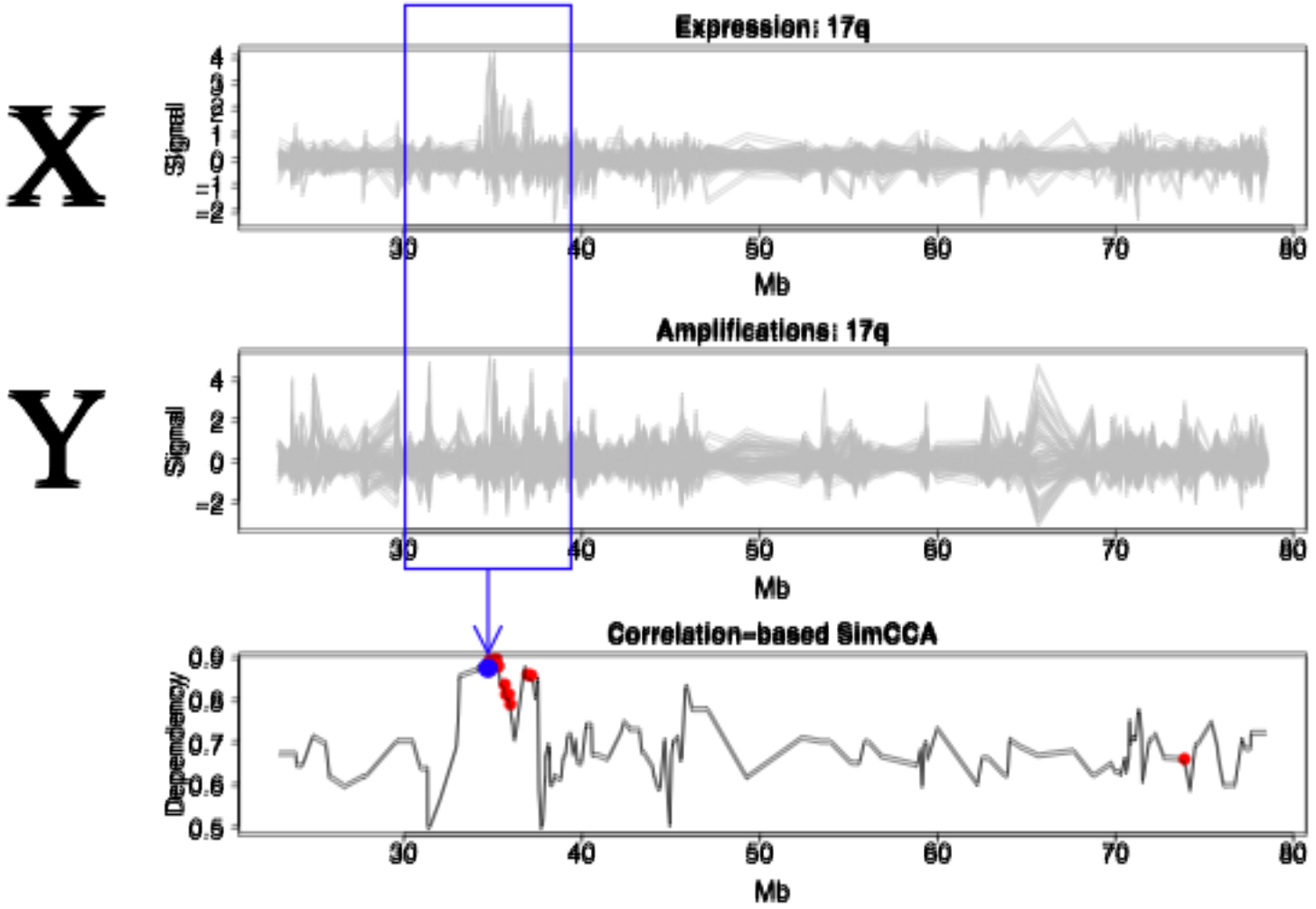
X

Y



Lahti et al, MLSP2009

Cancer study



Lahti et al, MLSP2009

Probabilistic Tools for Dependency Modelling

ICML/MLOSS 2010

Leo Lahti[1,2] Olli-Pekka Huovilainen[1] Tommi Suvitaival[1]
Ilkka Huopaniemi[1] Abhishek Tripathi[3] Samuel Kaski[1]

[1]Department of Information and Computer Science, Helsinki Institute for Information Technology HIIT and Adaptive Informatics Research Centre, Aalto University School of Science and Technology, Finland

[2]Haartman Institute and HUSLAB, Dpt of Pathology, University of Helsinki and Helsinki University Central Hospital, Finland

[3]Department of Computer Science, University of Helsinki, Finland

Development versions

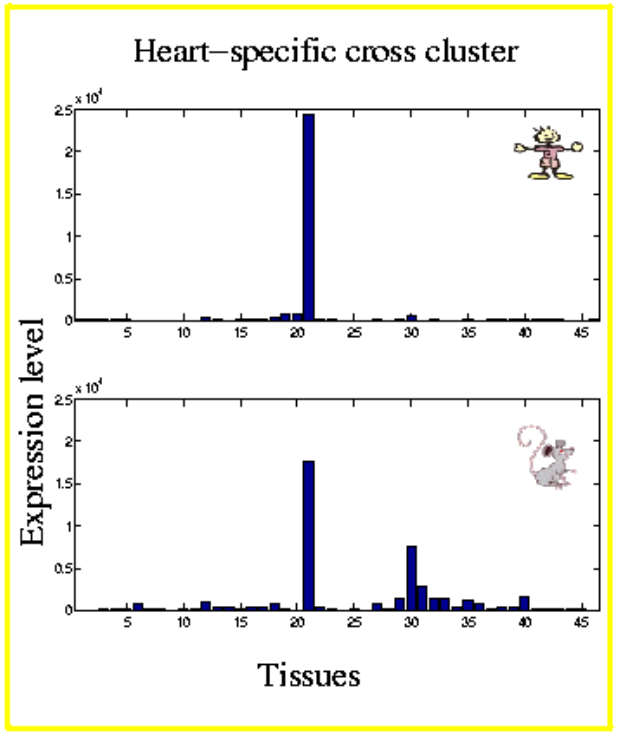
DNA copy number / gene expression / micro-RNA / methylation

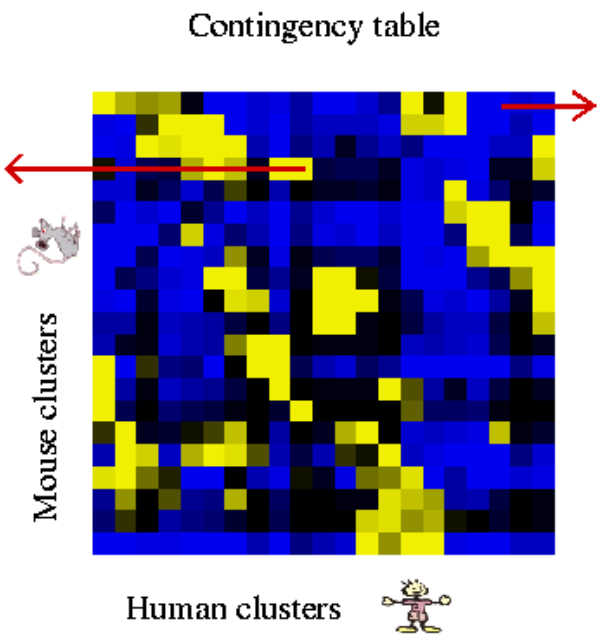
bioconductor.org/packages/devel/bioc/html/pint.html

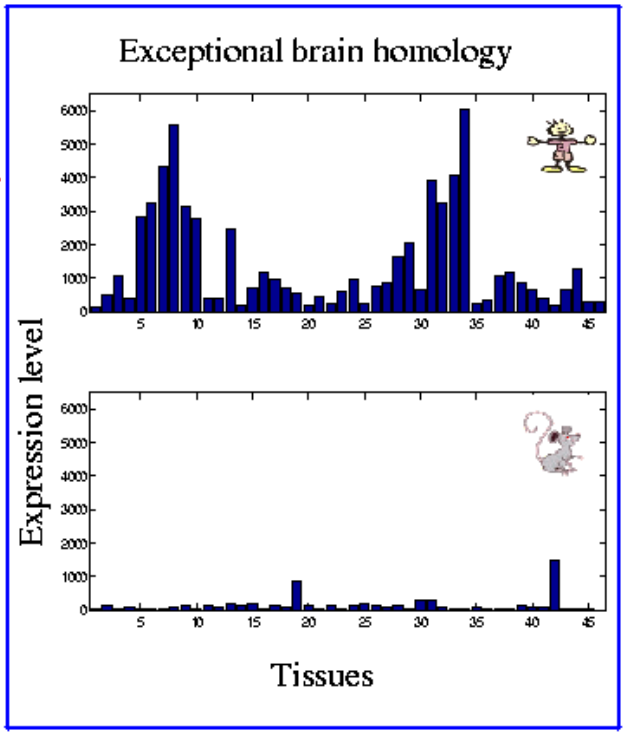
dmt.r-forge.r-project.org

(Leo Lahti, ICML/MLOSS'10)

From linear projections to clusterings. Associative clustering - of mice and men







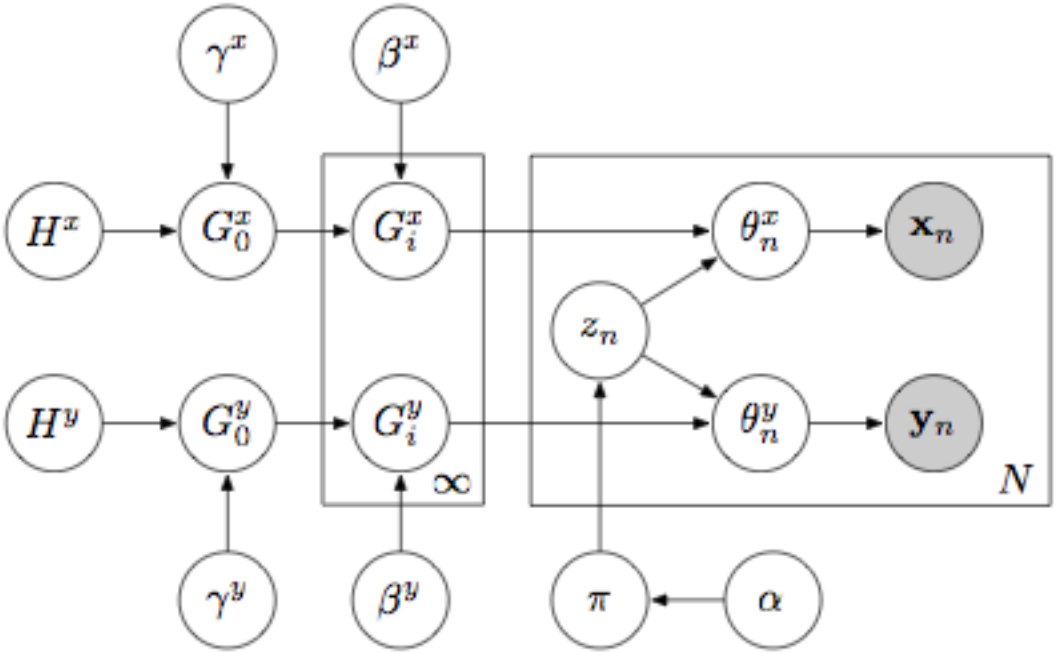
Dependency exploration through *associative clustering*: Search for regularities and exceptions in gene function between mouse and man

Kaski, et al, IEEE/ACM TrCBB 2005

Non-parametric dependencies between clusters

Triply infinite two-domain mixture model

Clusters x and y separately, and finds components that describe their dependencies



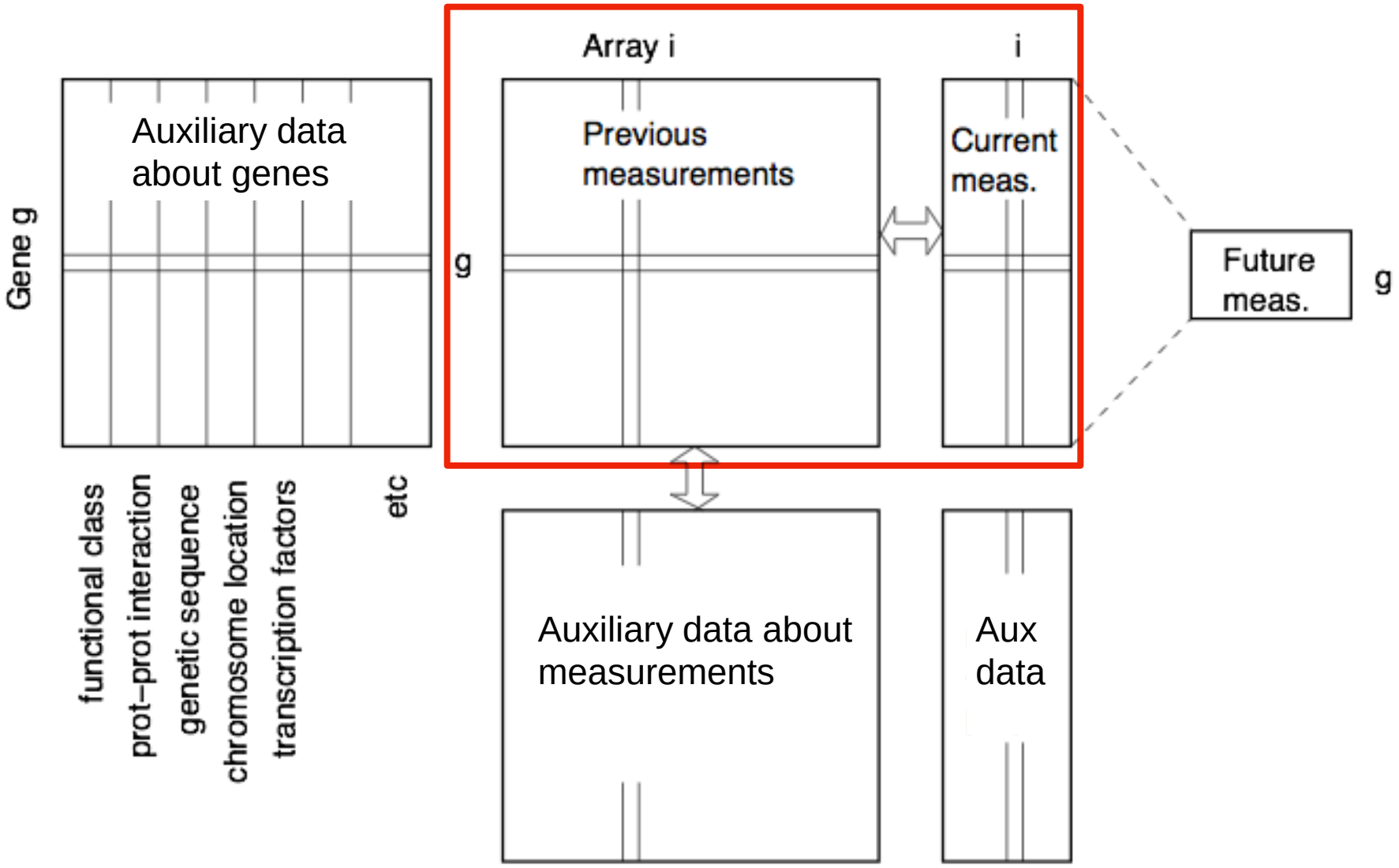
Rogers et al, Machine Learning 2010

Summary on multi-view learning

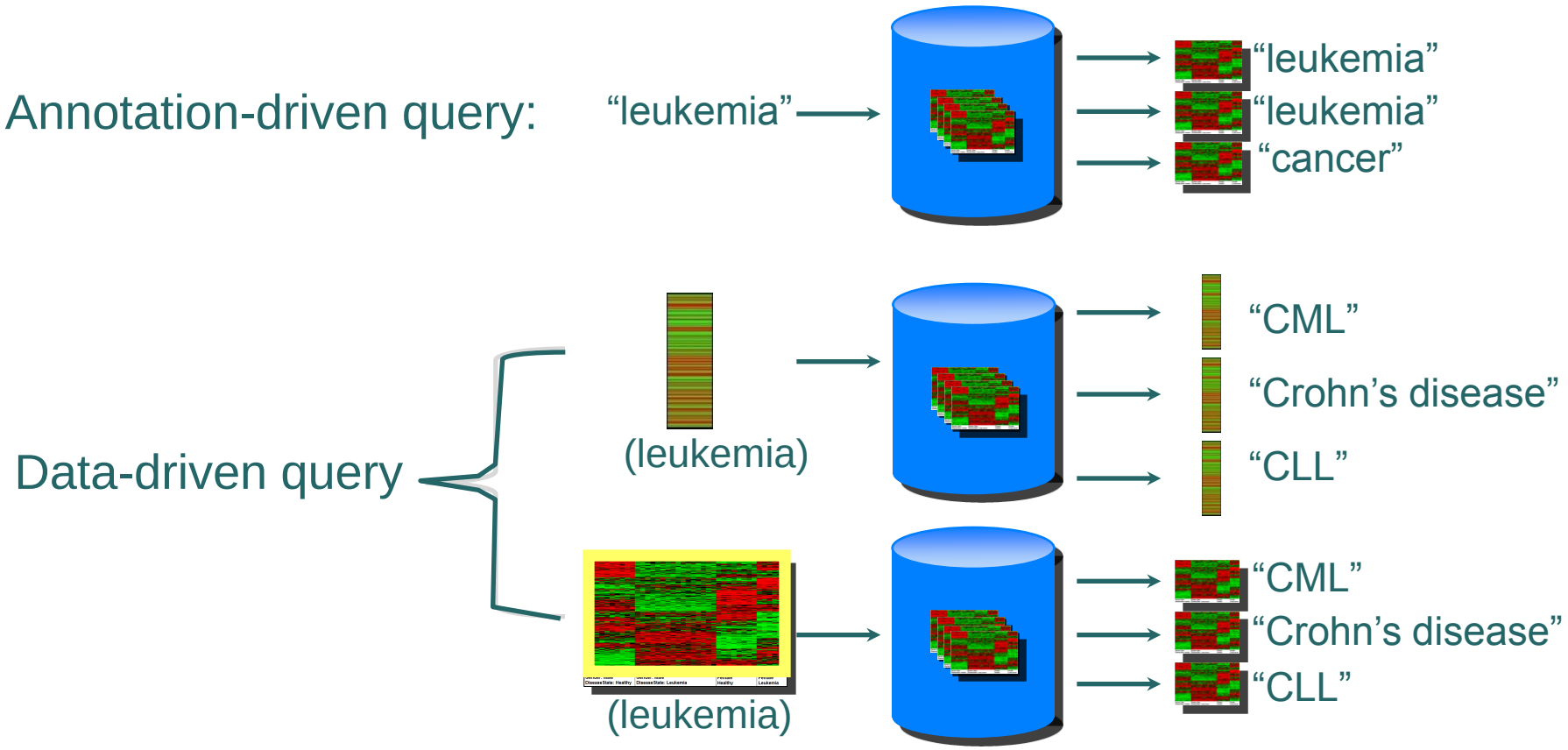
- Decomposition into shared and view-specific components
- Usable as a general-purpose preprocessing step
- Can be extended in several ways
  - Nonparametric methods
  - Associative clustering
  - Regularize
- Application to cancer studies

2. Get more data.

REx: Search for Relevant Experiments


Array i
i
Auxiliary data about genes
Previous measurements
Current meas.
Future meas.
Gene g
g
g
functional class
prot–prot interaction
genetic sequence
chromosome location
transcription factors
etc
Auxiliary data about measurements
Aux data

Querying collections



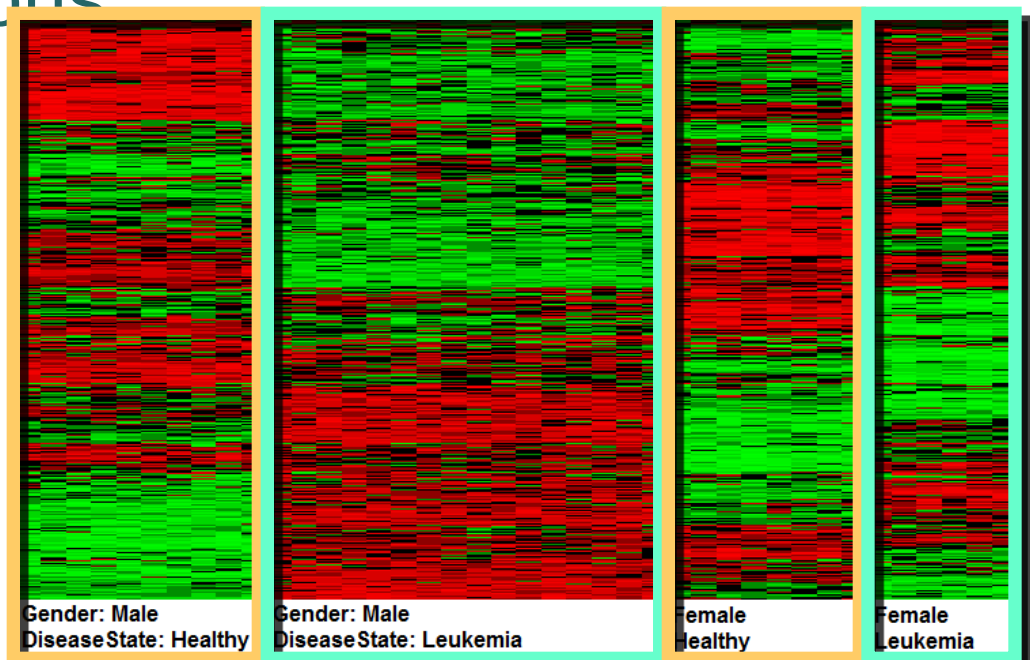
Caldas et al, Bioinformatics, 2009

What is interesting/relevant?

(i) Differential expression (Bring in covariates: treatment vs control). Why?

- The experimenter designed the controls to separate interesting variation
- The differences are more comparable across labs/situations

(ii) Bring in a model of biology



Modeling of an experiment collection

Task: Learn a decomposition of experiments into biological processes, given a database of experiments.

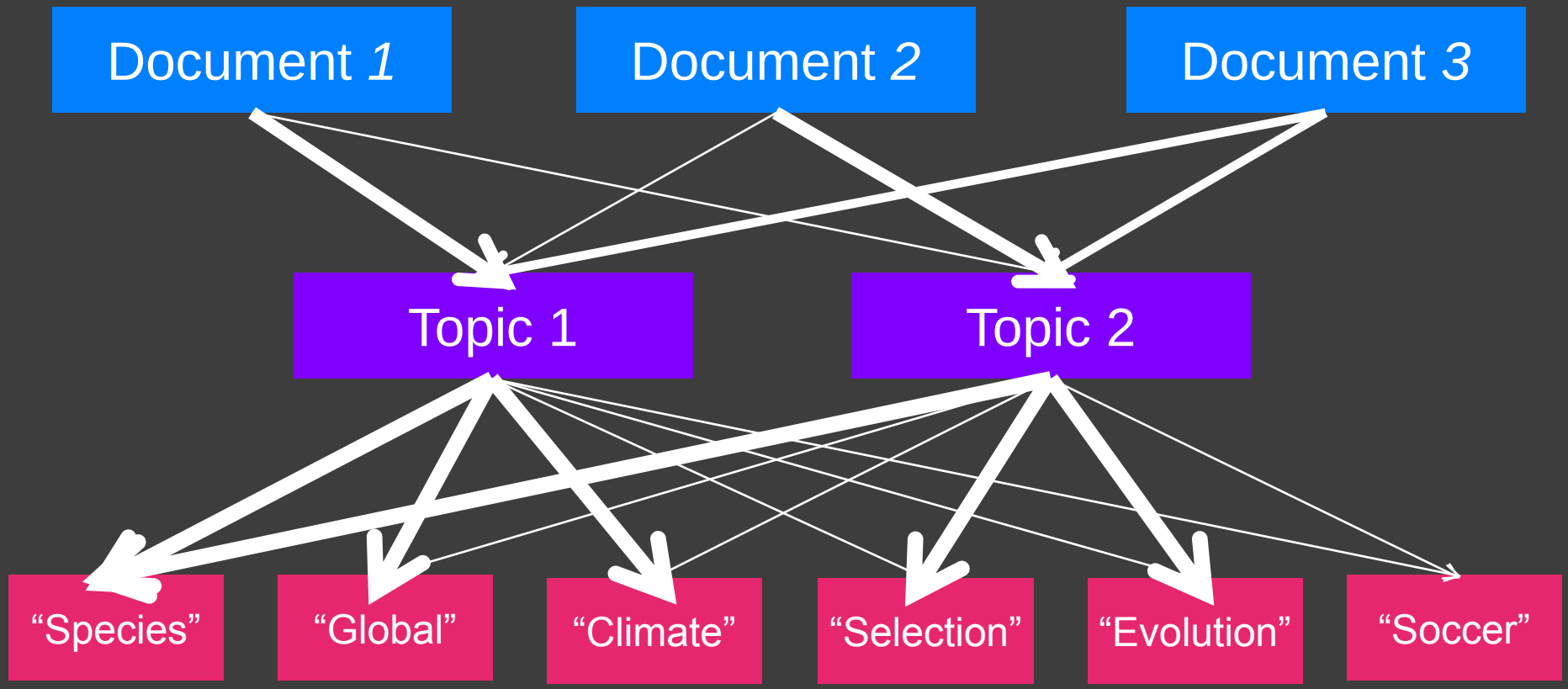
Solution in REx1.0:

- Assume experiments are bags of gene set activations (sets=biological constraints)
- Probabilistic overlapping components by topic models (data-driven modeling given the constraints)

Caldas et al, Bioinformatics, 2009

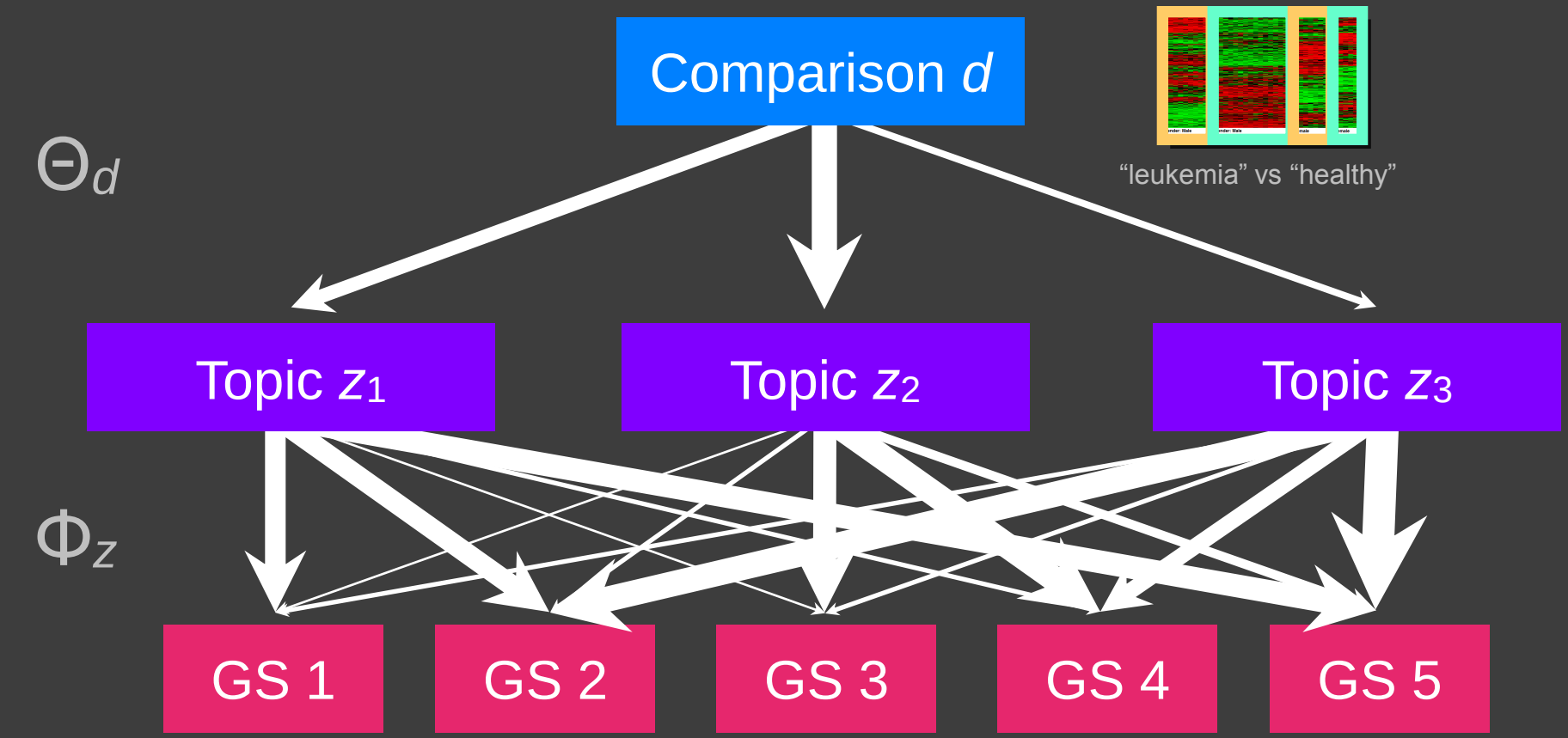
”Topic Model” / Latent variable model

- Extensively used in bag-of-words text data.
- Called Latent Dirichlet Allocation (LDA) or discrete PCA (dPCA)



Caldas et al, Bioinformatics, 2009

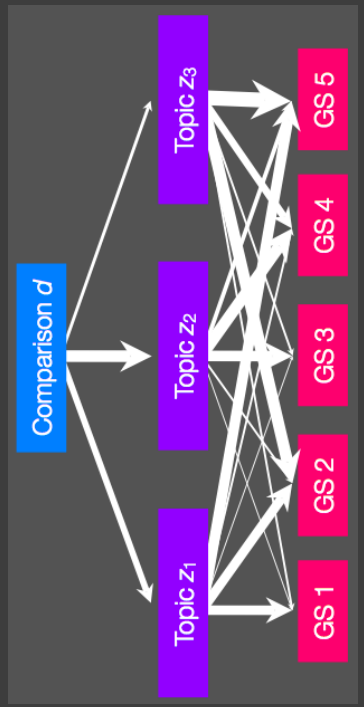
LDA and GSEA

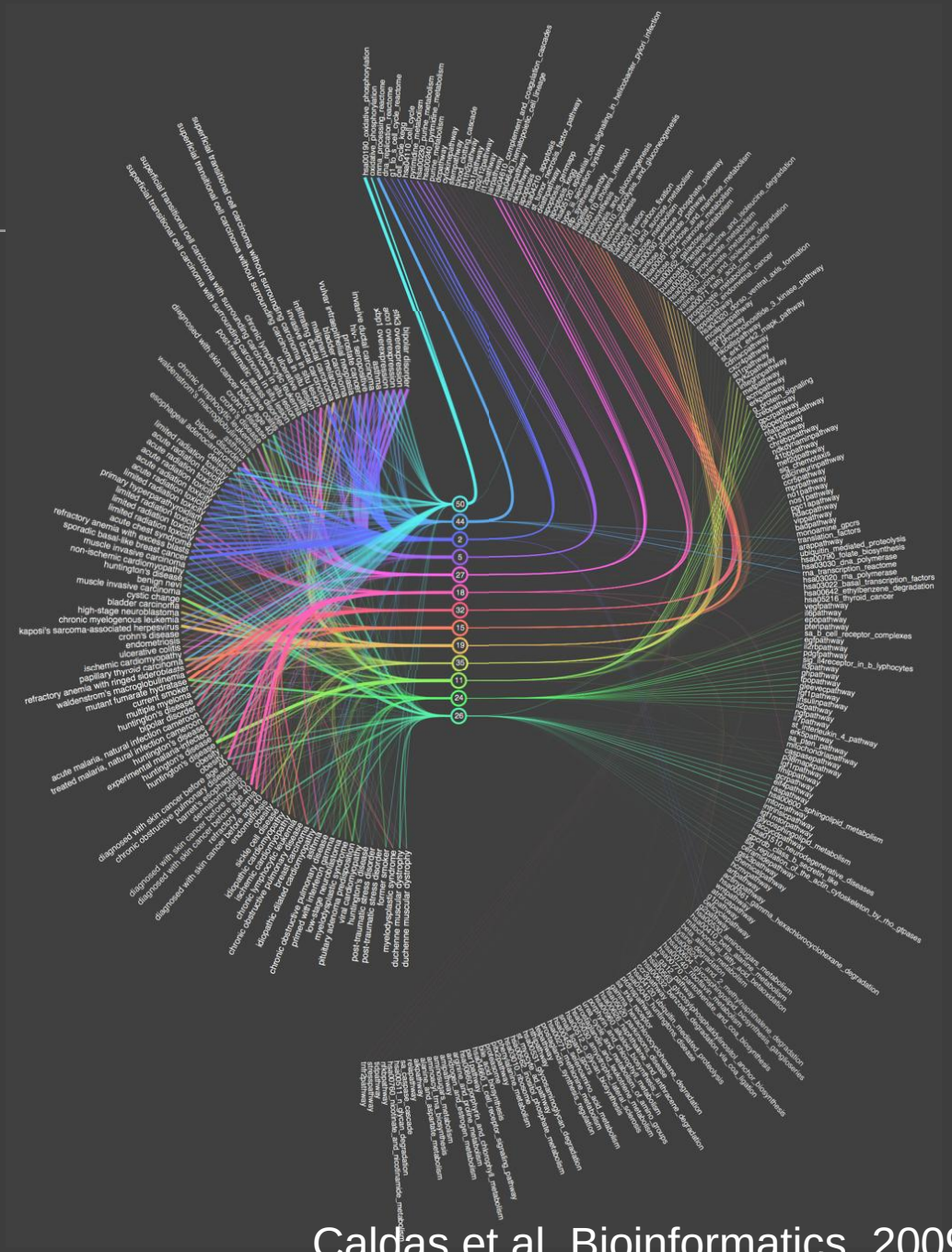


Estimate Θ and Φ with collapsed Gibbs sampler.

Caldas et al, Bioinformatics, 2009

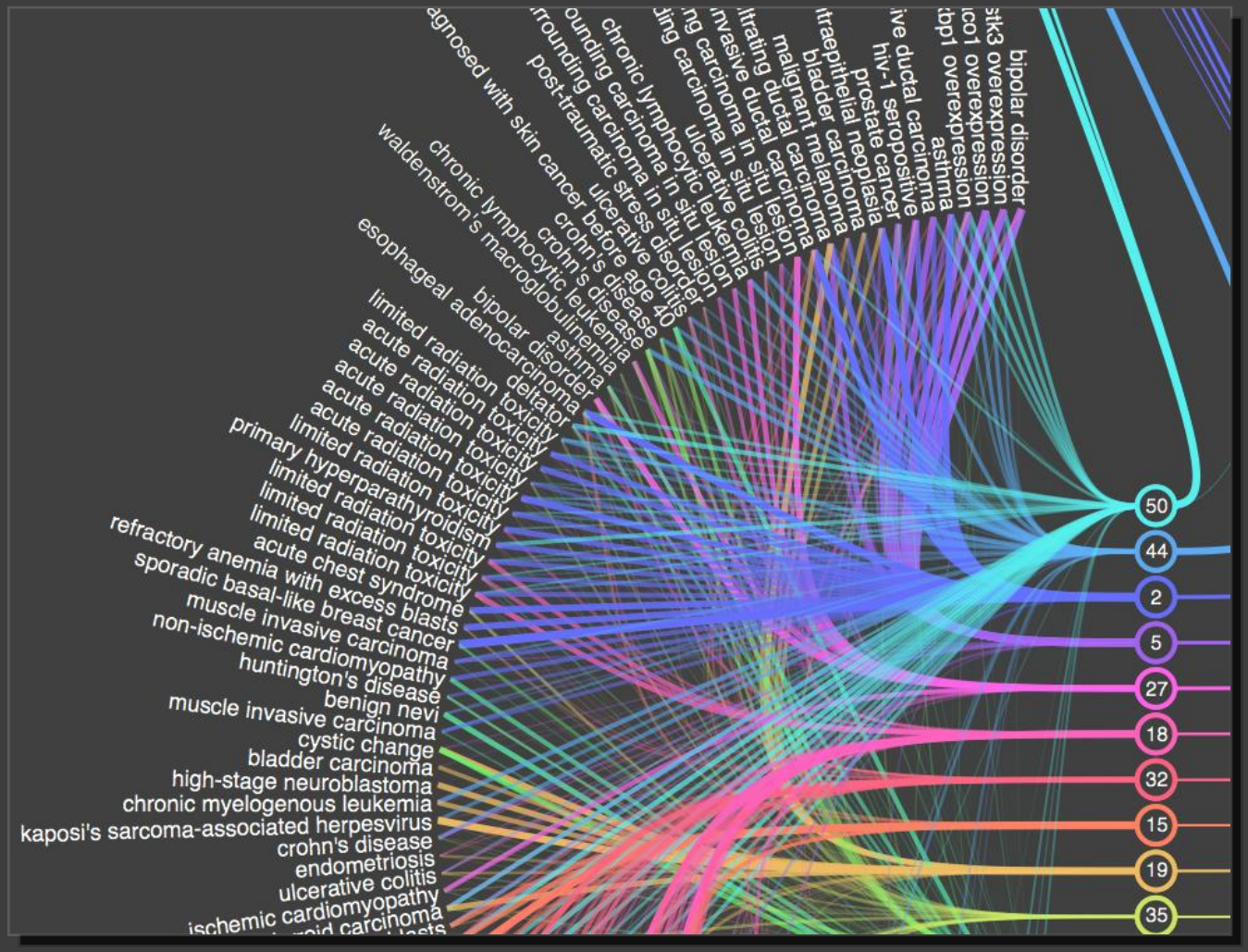
Components of experiments





Caldas et al, Bioinformatics, 2009

Components of experiments



Caldas et al, Bioinformatics, 2009

Retrieval of relevant experiments

Task: Find experiments in which the same biological processes are active.

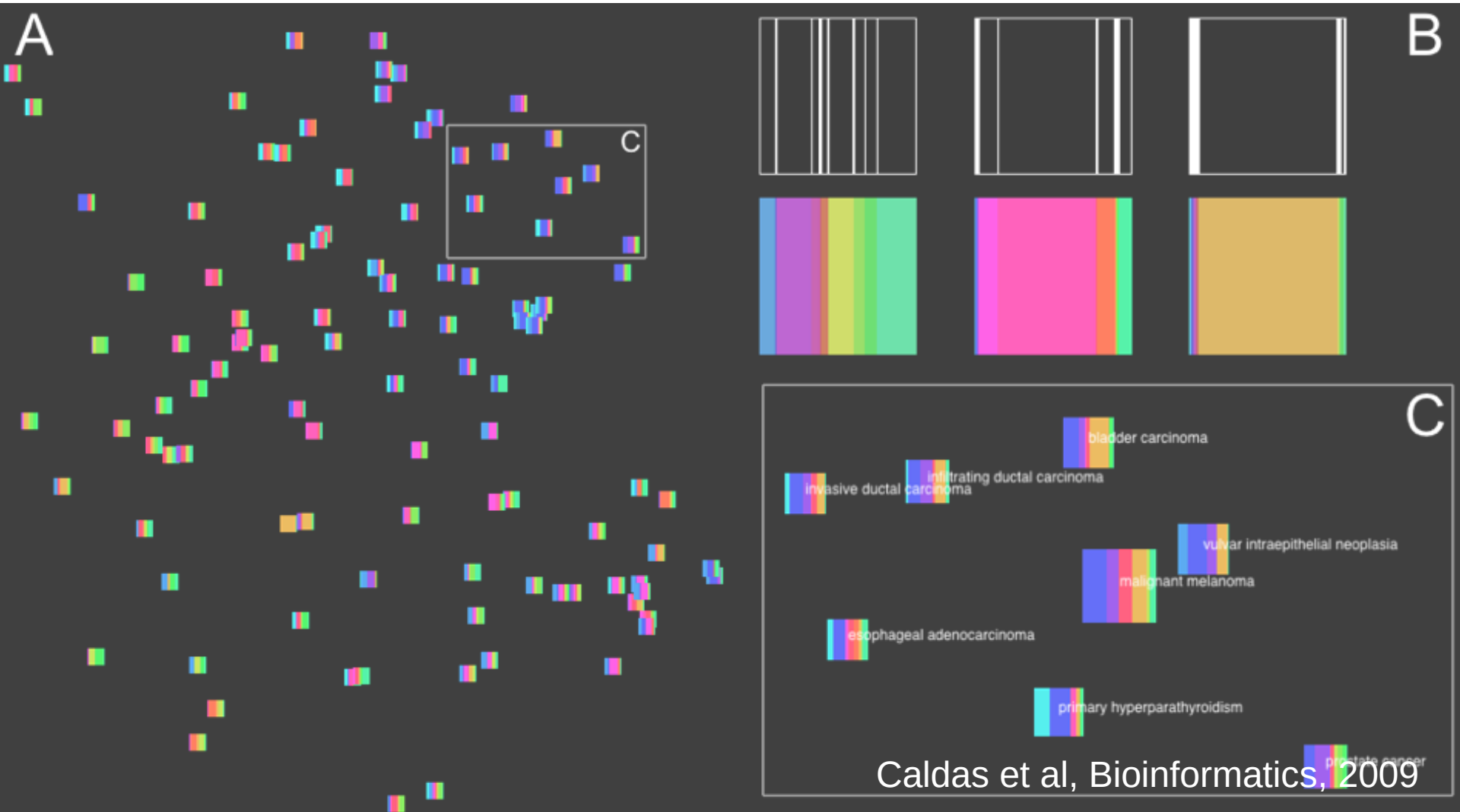
≈ find experiments where the same components are active

Convenient given the probabilistic model.
Rank the experiments by

p(query|experiment)

Caldas et al, Bioinformatics, 2009

Visualization of results: nonlinear projection



Caldas et al, Bioinformatics, 2009

Nonlinear projection

Task: Position each experiment on the plane such that relevant experiments are close to queries.

Solution:

Use *p(query|experiment)* to define relevance

Ask the relative cost of misses and false positives from the user

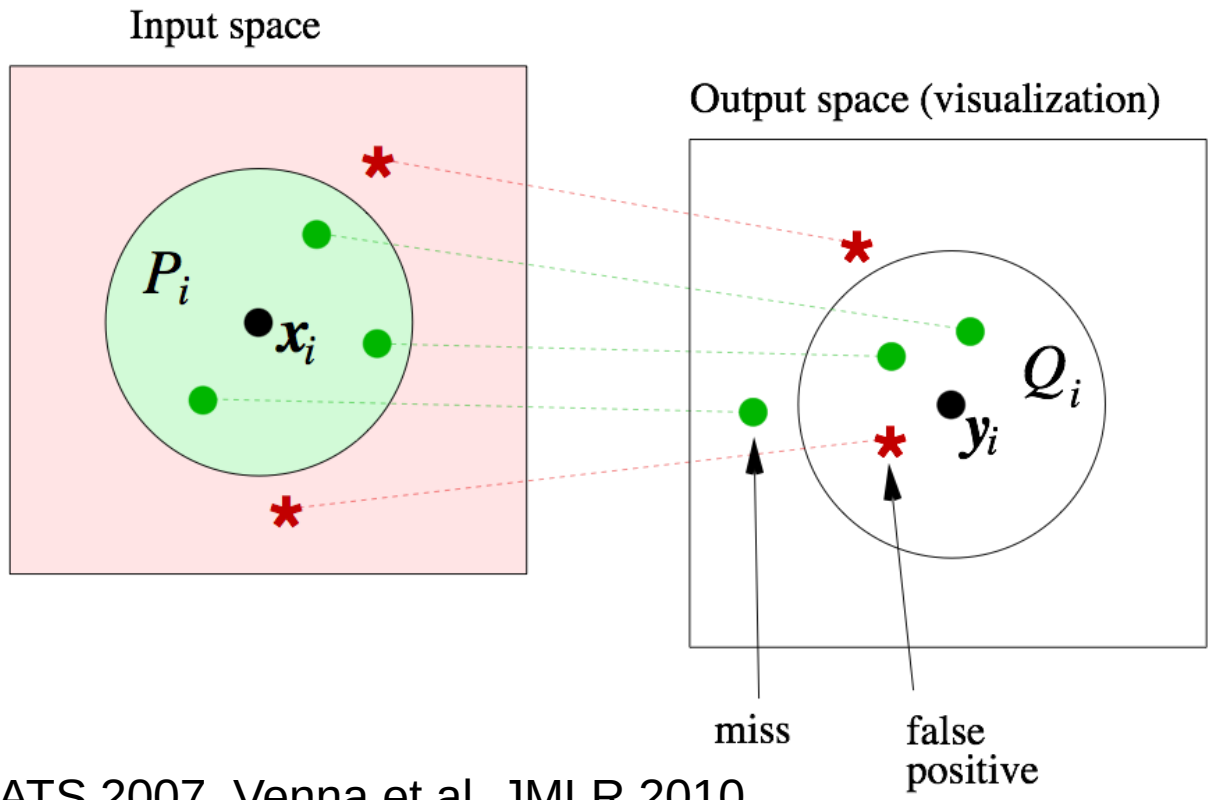
Minimize total cost by NeRV

Caldas et al, Bioinformatics, 2009

Neighbor retrieval visualizer NeRV

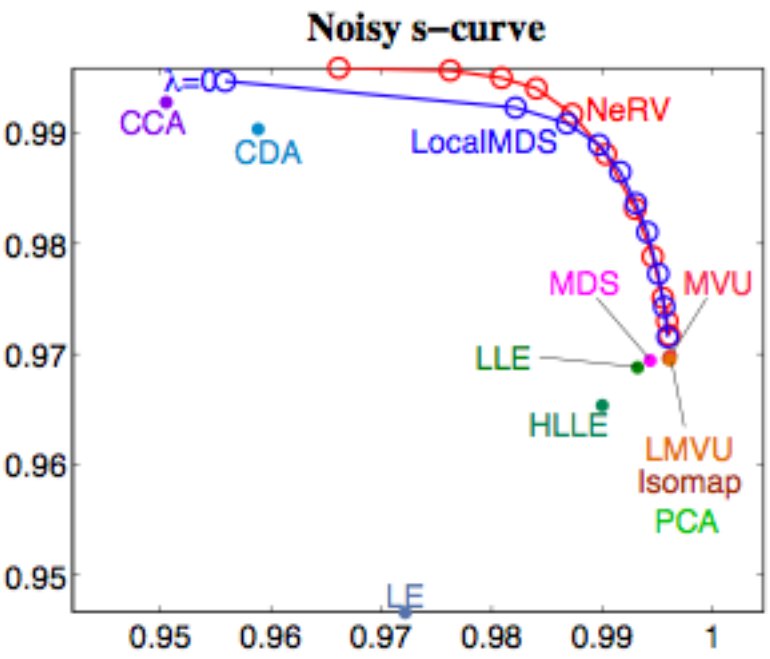
Optimizes a user-defined tradeoff between *precision* and *recall.*

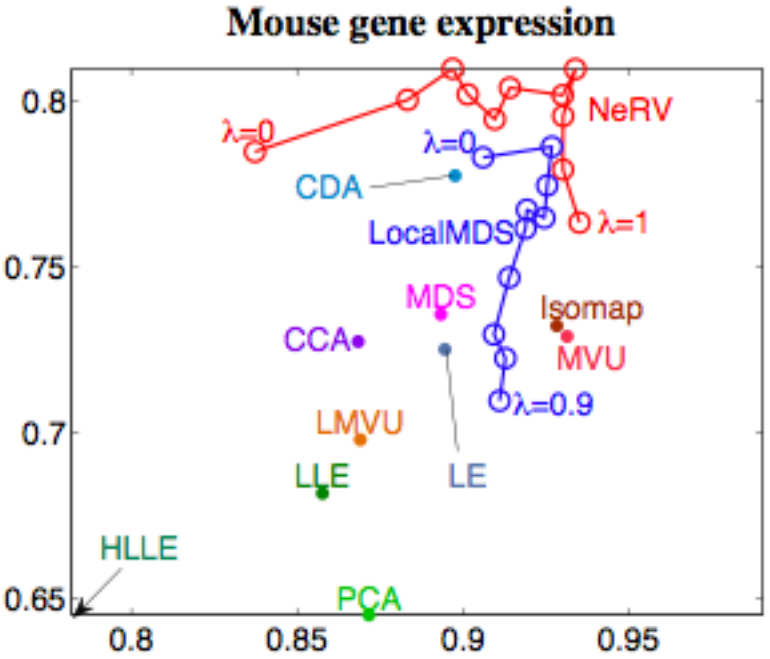
$$E_{\mathrm{NeRV}} = \lambda E_i[D(p_i, q_i)] + (1-\lambda) E_i[D(q_i, p_i)]$$



Venna and Kaski, AISTATS 2007, Venna et al, JMLR 2010

Does really work





Mean (rank-based, smoothed) precision

Mean (rank-based, smoothed) recall

http://www.cis.hut.fi/projects/mi/software/dredviz/

Venna and Kaski, AISTATS 2007; Venna et al, JMLR 2010

Querying the Model/Database

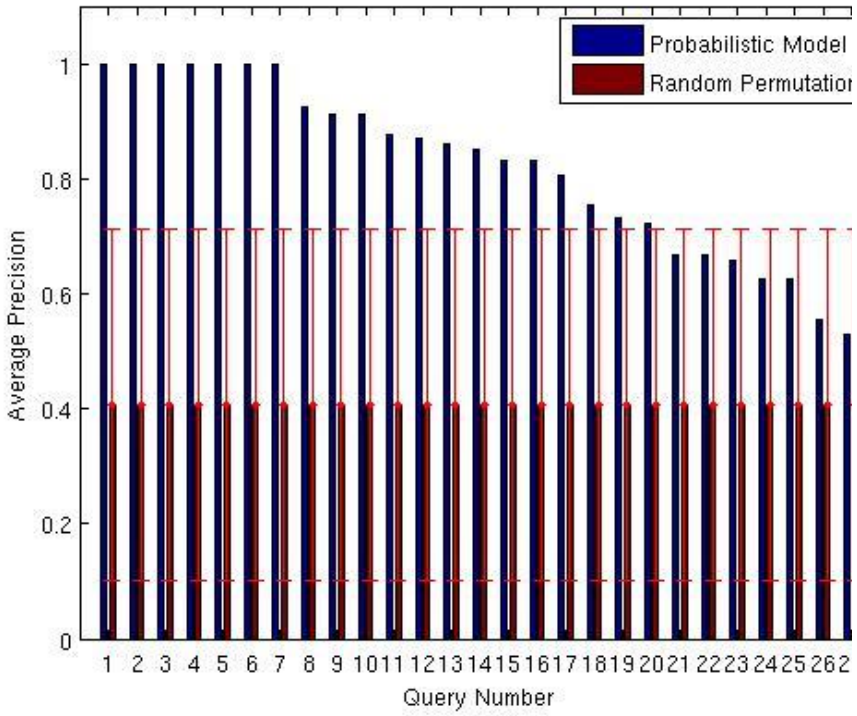
Query with “malignant melanoma” vs “normal” comparison.

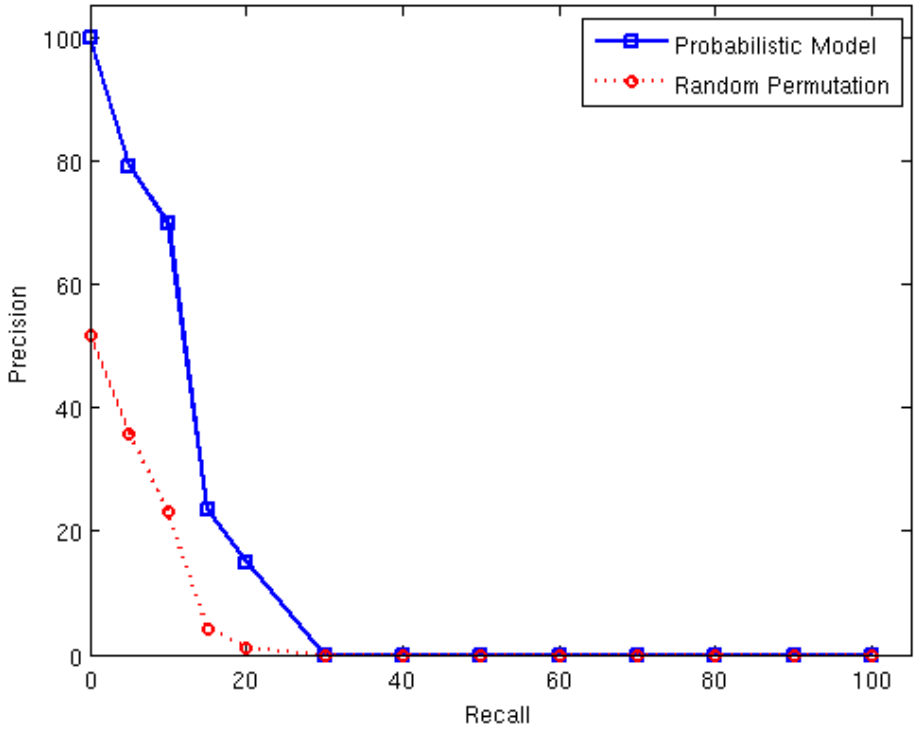
Rank	Comparison (... vs “normal”)
1	*Bladder Carcinoma*
2	*Vulvar Intraepithelial Neoplasia*
3	Hyperparathyroidism
4	Lung (smoker)
5	*Bladder Carcinoma*
6	*Bladder Carcinoma*
7	*Infiltrating Ductal Carcinoma*
8	*Prostate Cancer*
9	*Breast Carcinoma*
10	*Esophageal Adenocarcinoma*

Caldas et al, Bioinformatics, 2009

Retrieval results

- 105 normal vs. disease comparisons: ‘cancer’ (27) or ‘not cancer’ (78)
- Query with cancer comparisons
- Compare to random baseline





Caldas et al, Bioinformatics, 2009

Summary of REx: Retrieval of relevant Experiments

- Modeling of an experiment: Differential expression of biological processes (~gene sets)
- “Topic model” of bags of differentially expressed gene sets
- Probabilistic retrieval of relevant experiments, given the model
- Model-based visualization of results

Contributors from my group (current and former):

J. Caldas, A. Faisal, A. Klami, L. Lahti, J. Nikkilä, J. Parkkinen, J. Peltonen, J. Sinkkonen, A. Tripathi, J. Venna

Collaborators in these works (non-comprehensive list):

Laboratory of Cytomolecular Genetics, Univ. Helsinki: S. Knuutila, S. Myllykangas

VTT: M. Orešič

Univ of Turku: E. Savontaus

EBI, EMBL: A. Brazma, N. Gehlenborg

Medicel Oy: C. Roos + several companies

Univ. Glasgow: M. Girolami, S. Rogers

More information at:

http://www.cis.hut.fi/projects/mi