AIRC

TKK

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

Probabilistic retrieval and visualization of relevant experiments

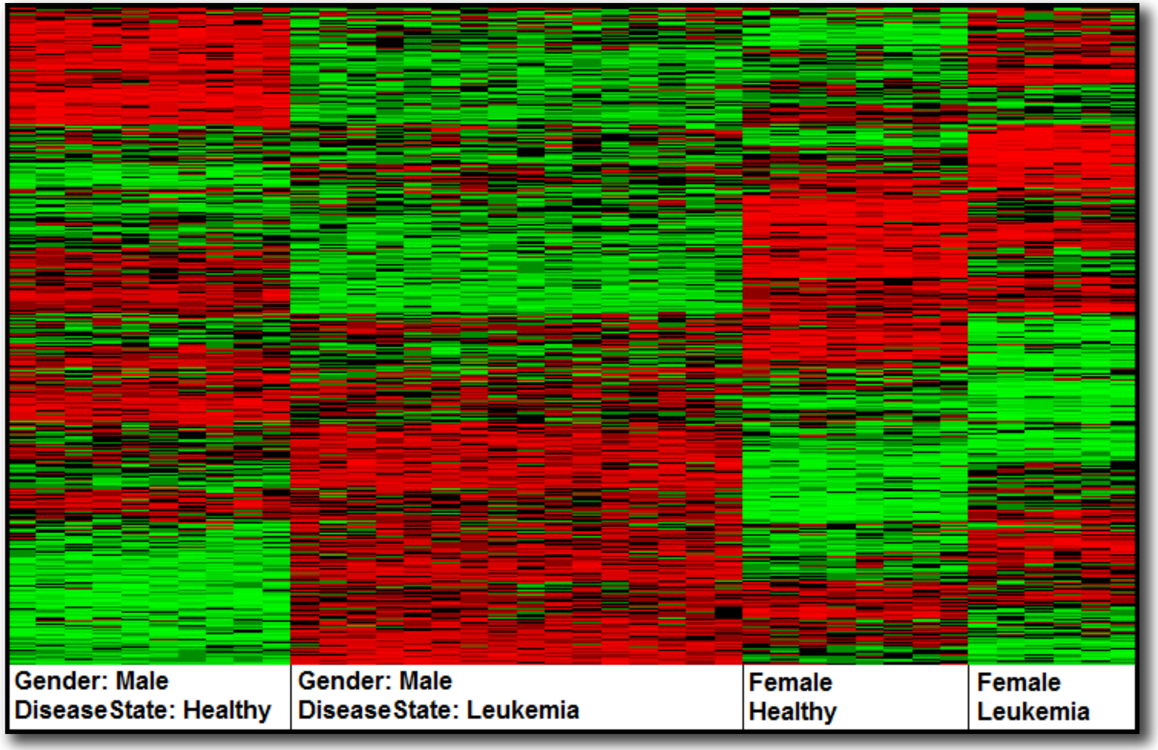
Samuel Kaski

Joint work with: José Caldas, Nils Gehlenborg, Ali Faisal, Alvis Brazma

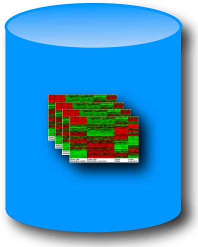
Motivation

How to best use collections of measurement data in biology?

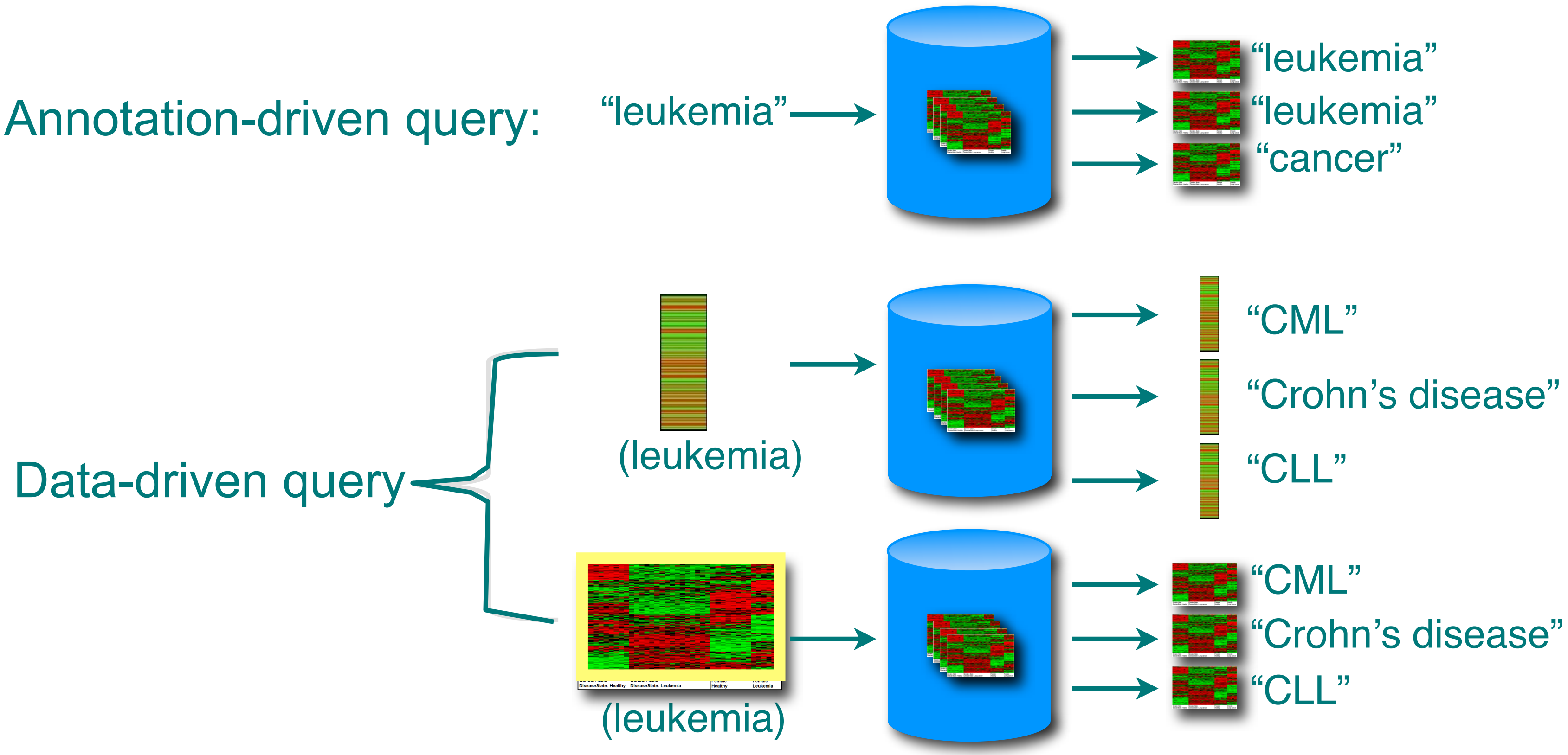
Example: Microarray experiments



Data and *experimental factors*



Querying collections



Annotation-driven query:

“leukemia”

“leukemia”

“leukemia”

“cancer”

Data-driven query

(leukemia)

“CML”

“Crohn’s disease”

“CLL”

(leukemia)

“CML”

“Crohn’s disease”

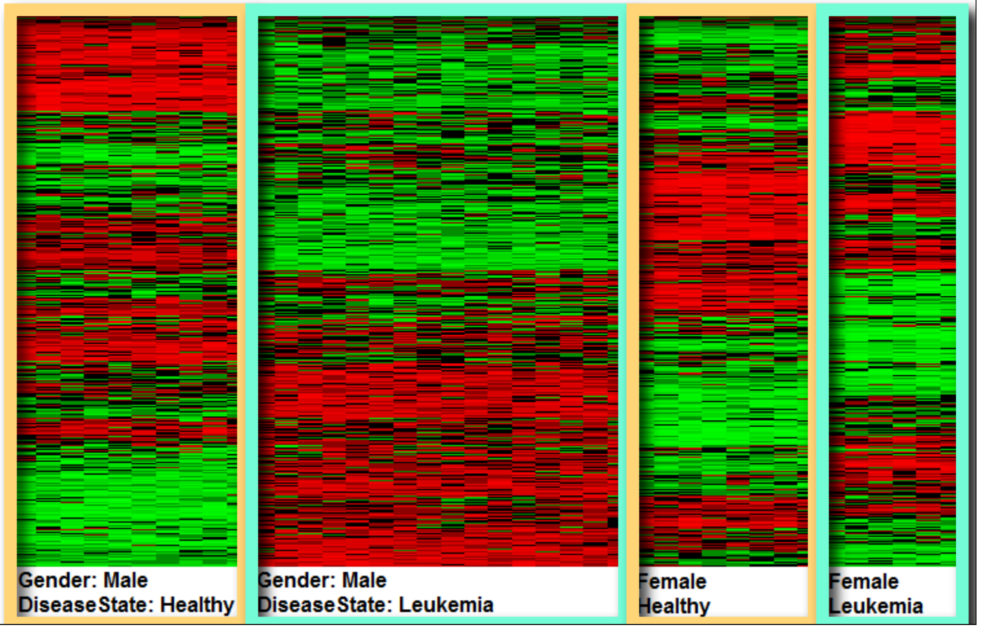
“CLL”

What is interesting/relevant?

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

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

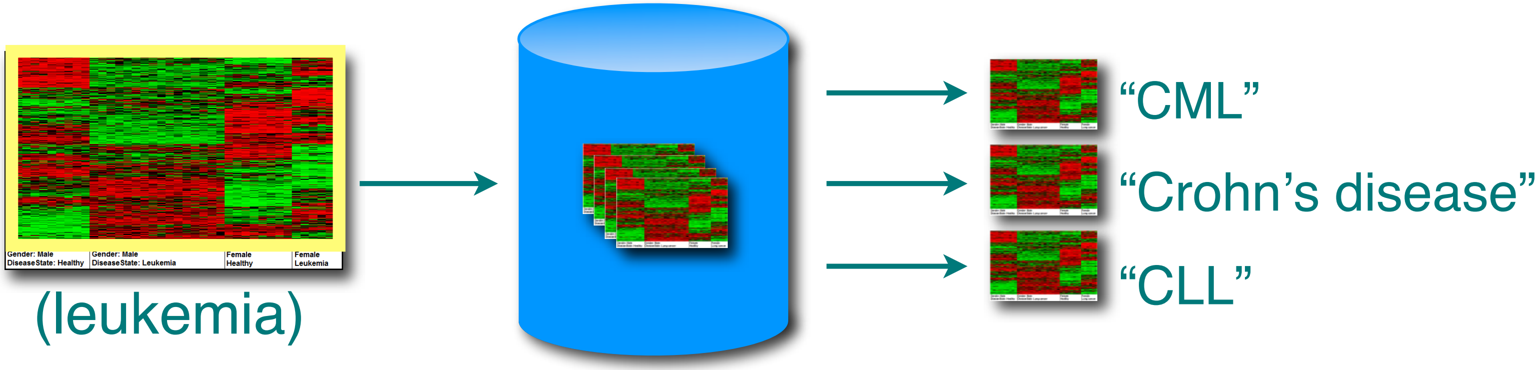
(ii) Bring in a model of biology



Content-based Retrieval of EXperiments

REX v1.0 (Caldas et al, ISMB 2009):

- Decompose data into binary comparisons
- Find similar comparisons (model-based)
- Include a biological model of the comparisons



(leukemia)

“CML”

“Crohn’s disease”

“CLL”

Methods needed

1. Modeling of an experiment
2. Modeling of experiment collection
3. Retrieval of relevant experiments
4. Visualization of results

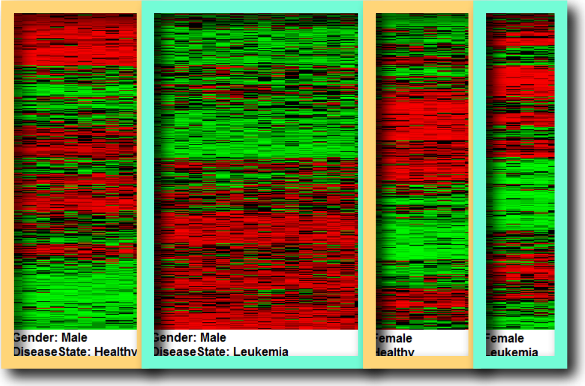
1. Modeling of an experiment

Question: How do case and control differ?

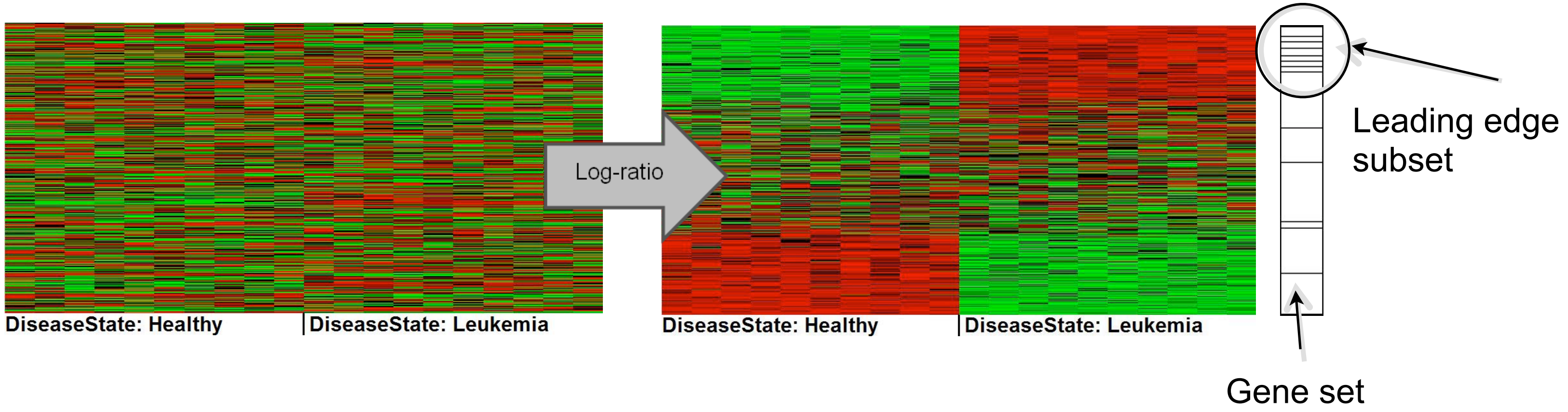
In biology: What biological processes are differentially active?

v0.1: Which genes are differentially expressed (t-test)

v1.0: Which *gene sets* are differentially expressed (and how much)



Gene Set Enrichment Analysis



Subramanian et al, PNAS 2005

Encode activity of the gene set by the count of active genes.

An experiment becomes encoded as a count vector. Dimension=gene set, count=number of active genes.

2. Modeling of an experiment collection

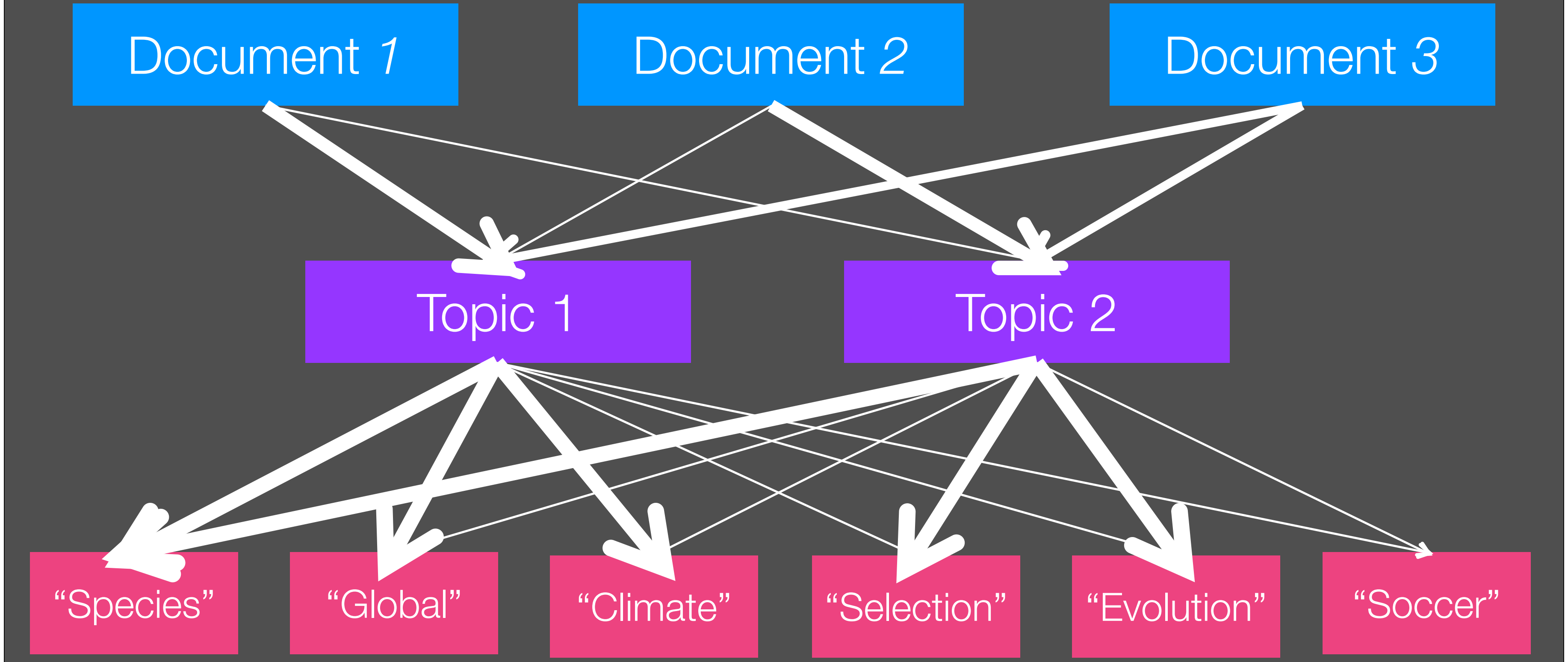
Task: Learn a decomposition of experiments into biological processes, given a database of experiments.

Solution v1.0:

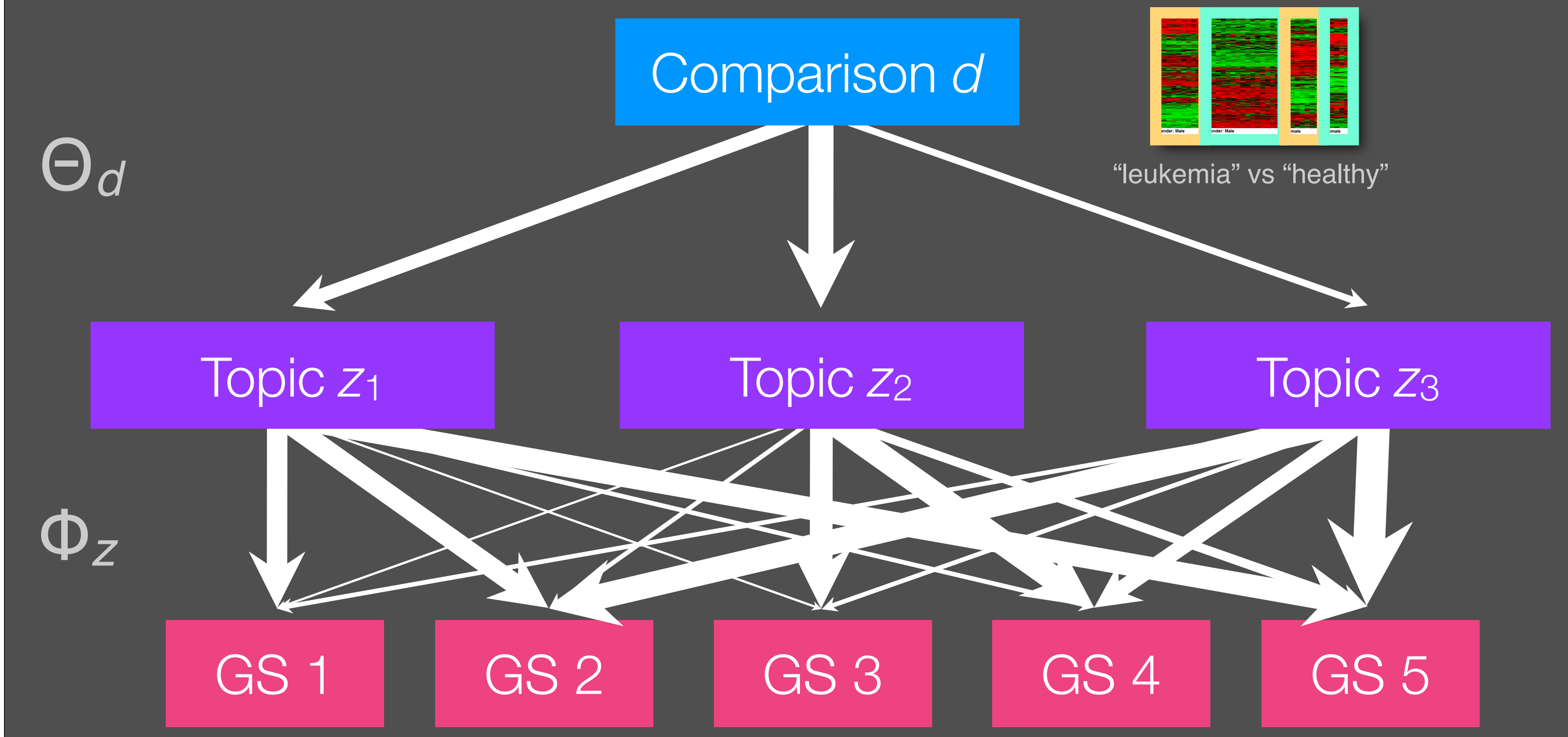
- Assume experiments are bags of gene set activations (sets=biological constraints)
- Probabilistic overlapping components by topic models (data-driven modeling given the constraints)

Topic Model

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



LDA and GSEA



Comparison *d*

“leukemia” vs “healthy”

Θ_d

Topic z_1

Topic z_2

Topic z_3

Φ_z

GS 1

GS 2

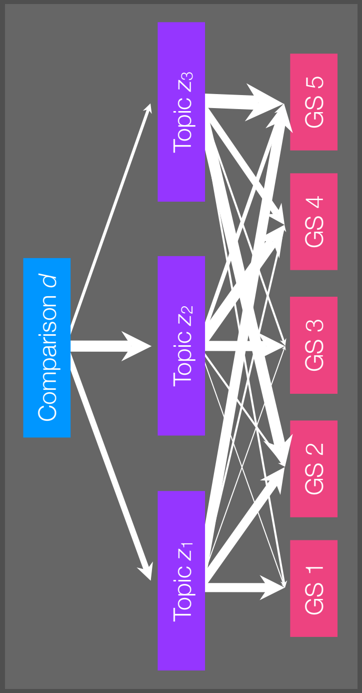
GS 3

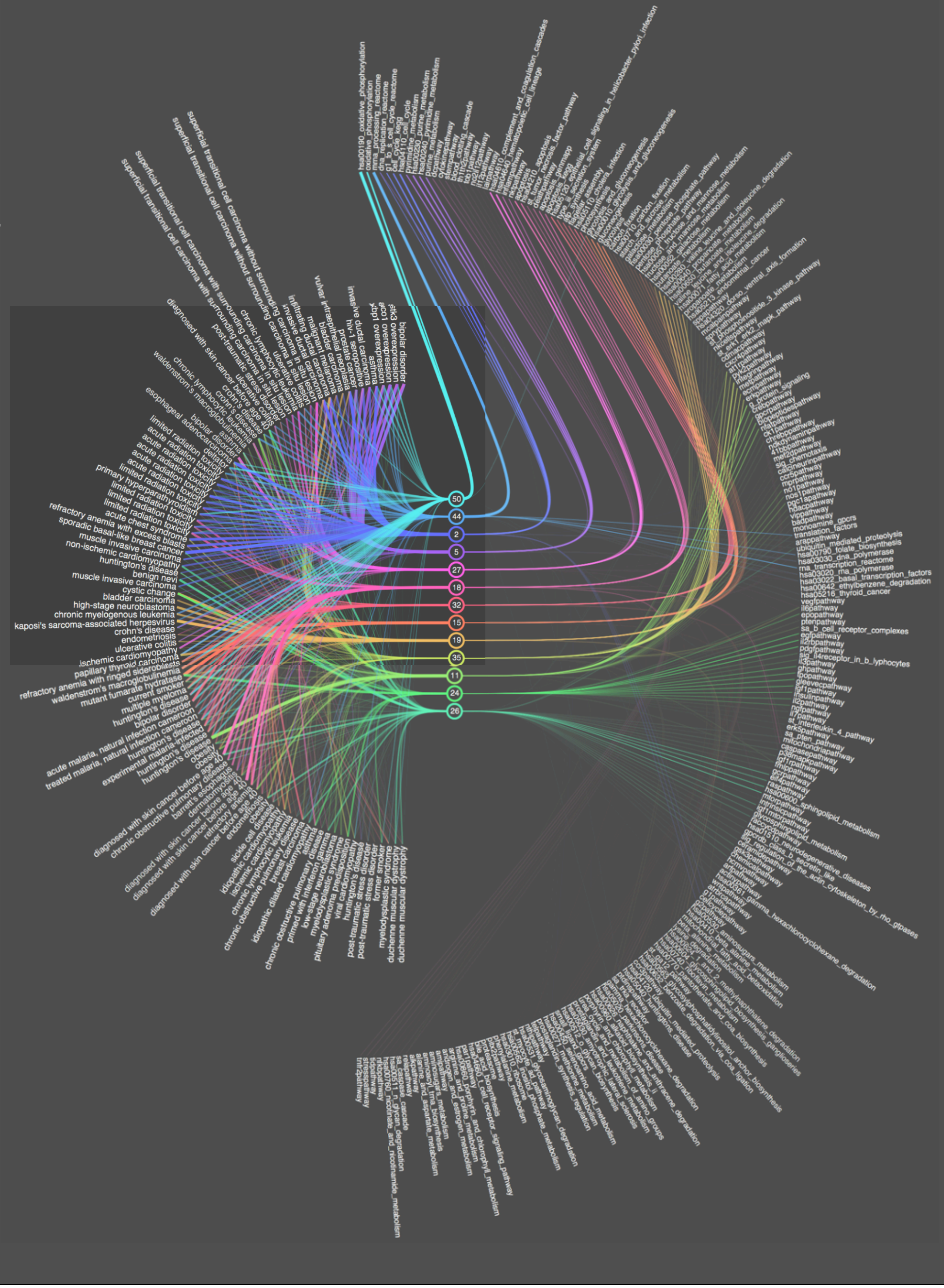
GS 4

GS 5

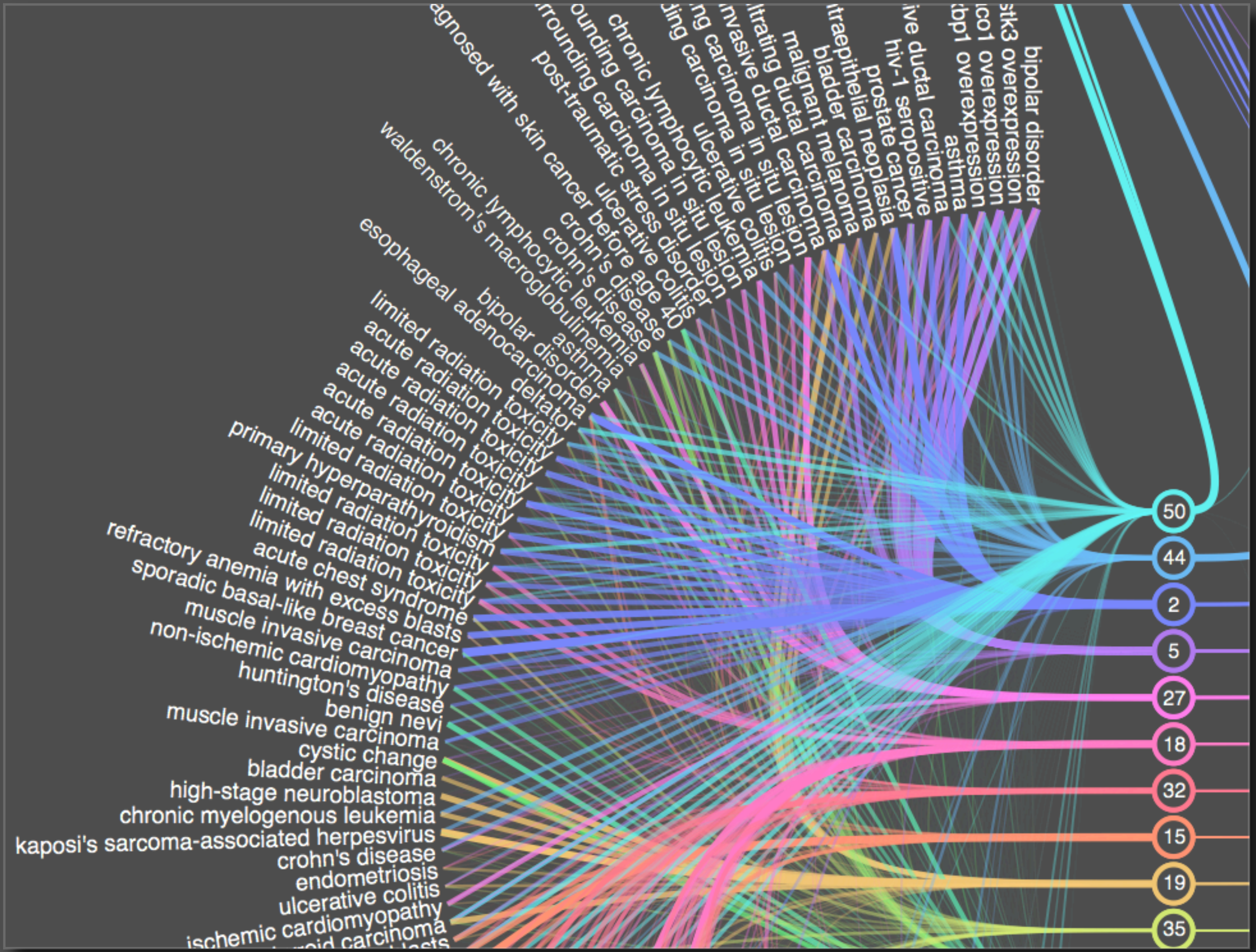
Estimate Θ and Φ with collapsed Gibbs sampler.

Components of experiments





Components of experiments



3. Retrieval of relevant experiments

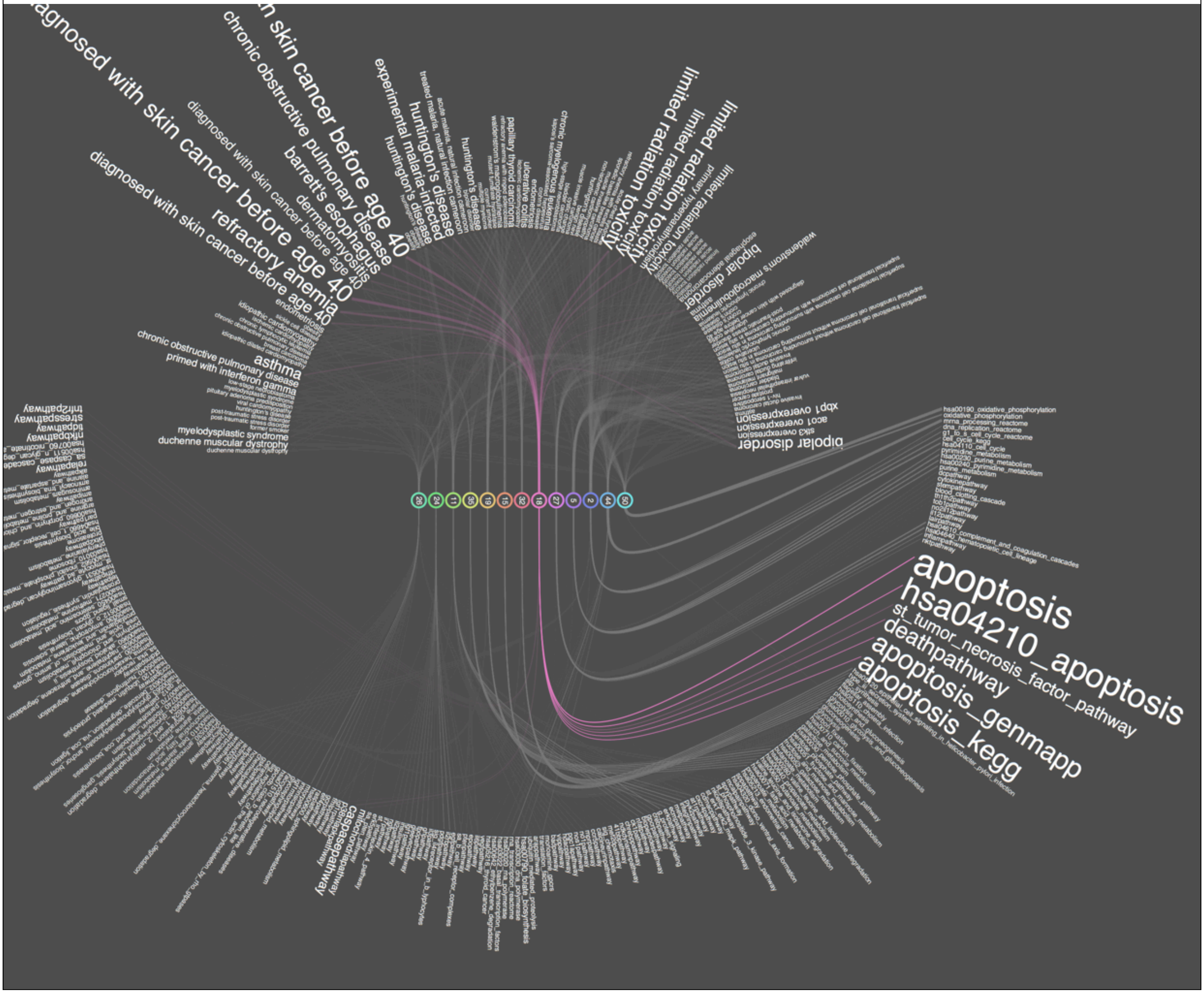
Task: Find experiments in which the same biological processes are active.

≈ find experiments where the same components are active

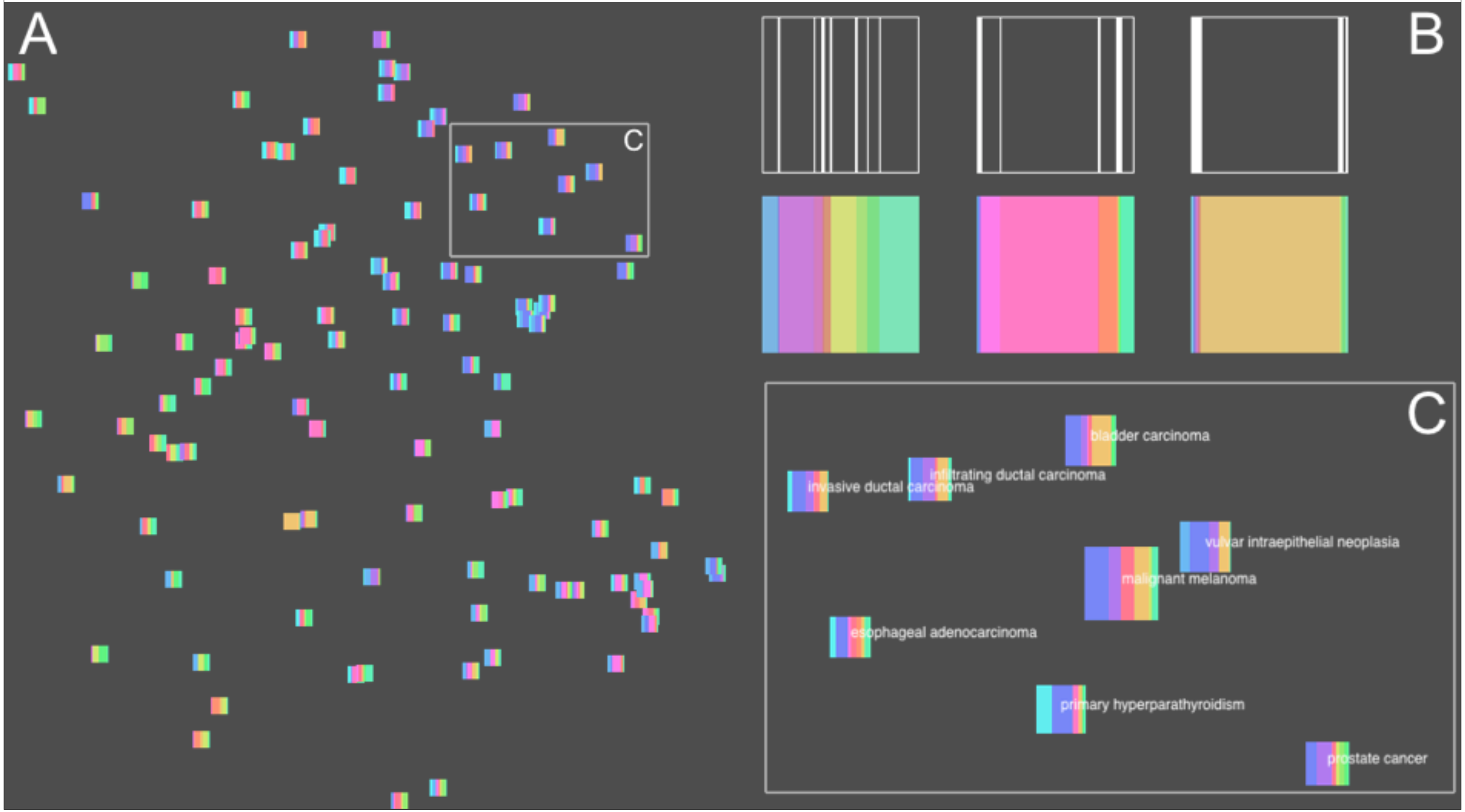
Convenient given the probabilistic model.
Rank the experiments by

$p(query|experiment)$

4. Visualization of results (a): interpretation of components


50
44
2
5
27
18
32
15
19
35
11
24
26
apoptosis
hsa04210_apoptosis
st_tumor_necrosis_factor_pathway
deathpathway
apoptosis_genmapp
apoptosis_kegg
bipolar disorder
xbp1 overexpression
limited radiation toxicity
limited radiation toxicity
limited radiation toxicity
limited radiation toxicity
bipolar disorder
waldenstrom's macroglobulinemia
chronic myelogenous leukemia
endometriosis
ulcerative colitis
papillary thyroid carcinoma
huntington's disease
huntington's disease
experimental malaria-infected
huntington's disease
n skin cancer before age 40
chronic obstructive pulmonary disease
barrett's esophagus
dermatomyositis
diagnosed with skin cancer before age 40
agnosed with skin cancer before age 40
refractory anemia
diagnosed with skin cancer before age 40
endometriosis
asthma
chronic obstructive pulmonary disease
primed with interferon gamma
myelodysplastic syndrome
duchenne muscular dystrophy
tnfr2pathway
stresspathway
tidpathway
nfkbpathway
relapathway
mitochondriapathway
caspasepathway
hsa00190_oxidative_phosphorylation
oxidative_phosphorylation

Visualization of results (b): nonlinear projection



Nonlinear projection

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

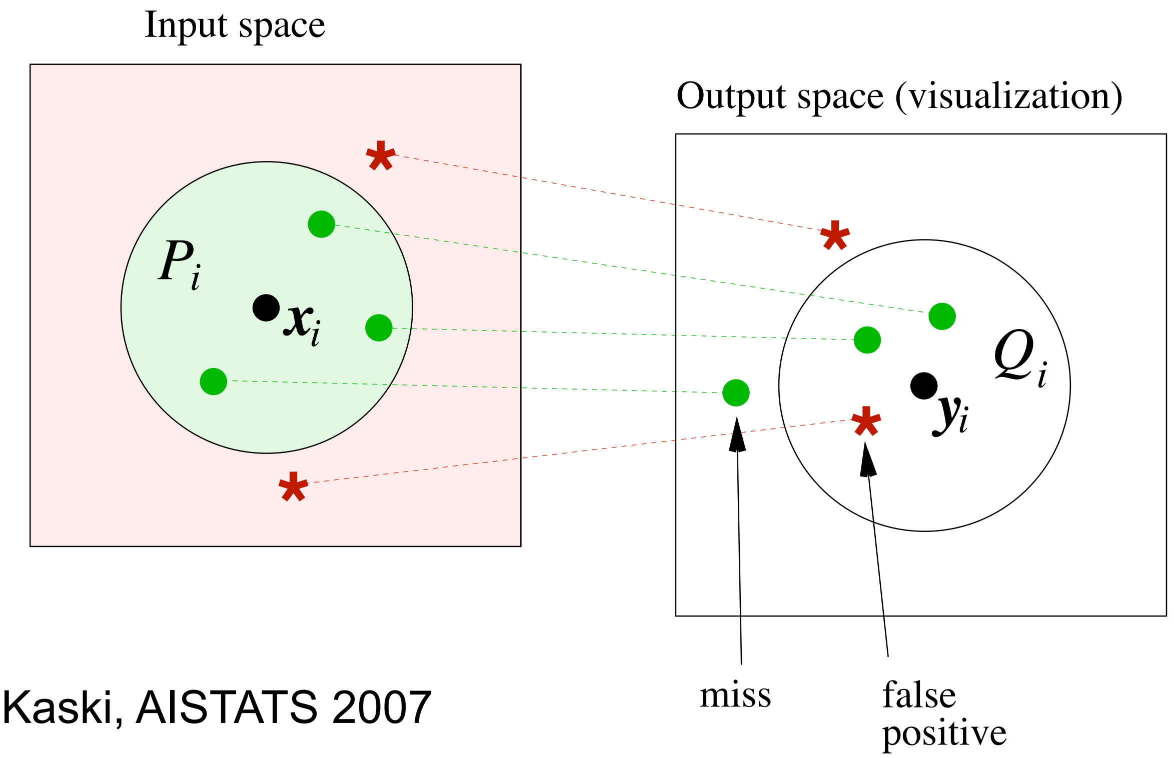
Solution:

1. Use *p(query|experiment)* to define relevance
2. Ask the relative cost of misses and false positives from the user
3. Minimize total cost by NeRV

Neighbor retrieval visualizer NeRV

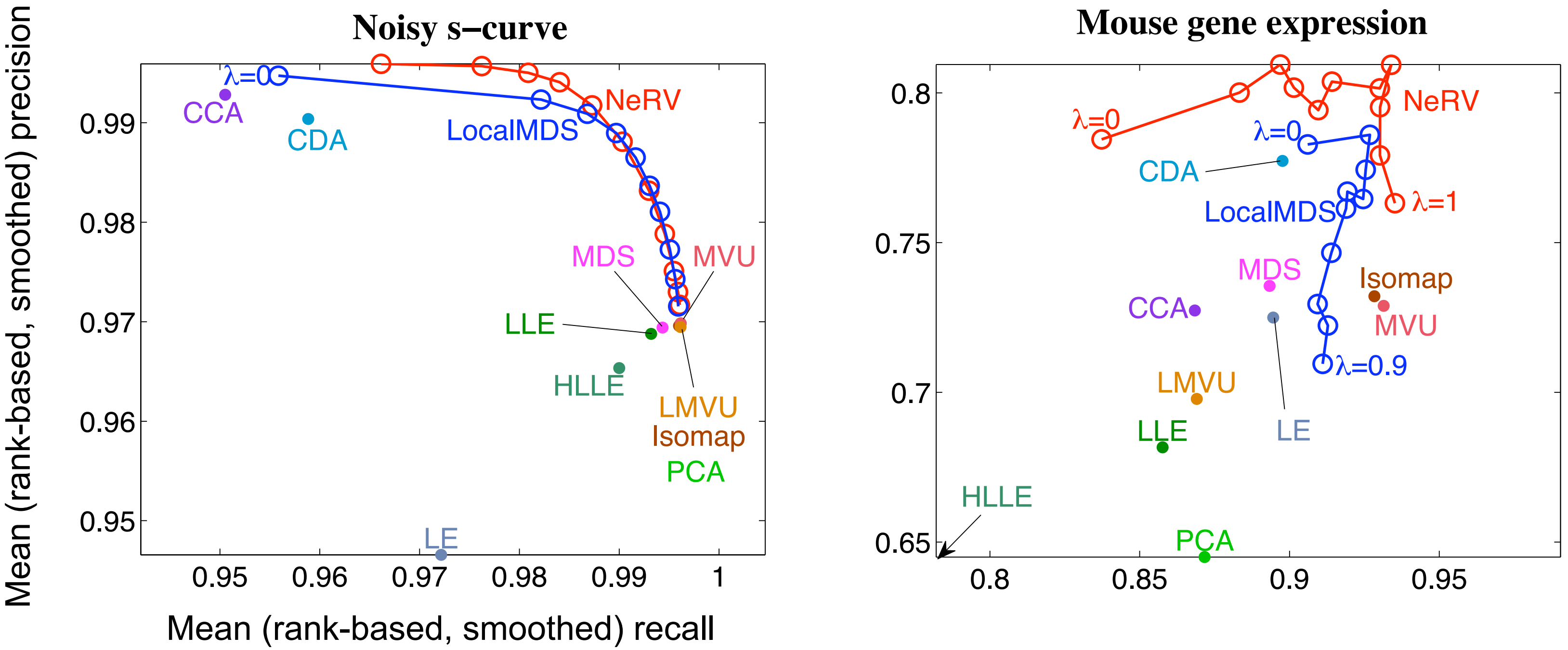
Optimizes a user-defined tradeoff between *precision* and *recall.*

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



Venna and Kaski, AISTATS 2007

Does really work



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

Results

Setup

- 288 preprocessed experiments from ArrayExpress: 768 comparisons.
- Focussed on 105 comparisons of healthy vs disease.
- Ran GSEA on collection of manually curated gene sets from MSigDB (KEGG, Biocarta,...).
- LDA with 50 topics.

Top Gene Sets for Representative Topics

Topic 2	Topic 24	Topic 44	Topic 50
Cell Cycle (BIOCARTA)	IL2RB Pathway	mRNA Processing (REACTOME)	Oxidative Phosphorylation (KEGG)
Cell Cycle (KEGG)	PDGF Pathway	RNA Transcription (REACTOME)	Oxidative Phosphorylation (GENMAPP)
G1 to S Cell Cycle (REACTOME)	EGF Pathway	Translation Factors	Glycolysis and Gluconeogenesis
DNA Replication (REACTOME)	Gleevec Pathway	Folate Biosynthesis	IL-7 Pathway
G2 Pathway	IGF-1 Pathway	Basal Transcription Factors	γ-Hexachlorocyclohexane Degradation

Topics are coherent and cover a wide range of biological processes.

Some of the top gene sets have significant gene overlap.

More Detailed Analysis of Topic 2

Topic 2 (cell cycle gene sets) has “ATR BRCA Pathway” gene set at 8th position.

BRCA is involved in breast cancer and DNA repair.

Comparisons with highest probability for Topic 2?

Rank	Comparison (... vs “normal”)
1	*Sporadic basal-like breast cancer*
2	Vulvar intraepithelial neoplasia
3	*Breast carcinoma*
4	Esophageal carcinoma

More Detailed Analysis of Topic 44

Topic 44 (transcription machinery) has “Folate Biosynthesis” gene set at 4th position.

Folate plays role in DNA and RNA synthesis.

Comparisons with highest probability for Topic 44?

Rank	Comparison (... vs “normal)
1	Crohn’s Disease
2	Chronic Lymphcytic Leukemia
3	Chronic Myelogenous Leukemia

Folate deficiency has been associated with both Crohn’s disease and leukemia.

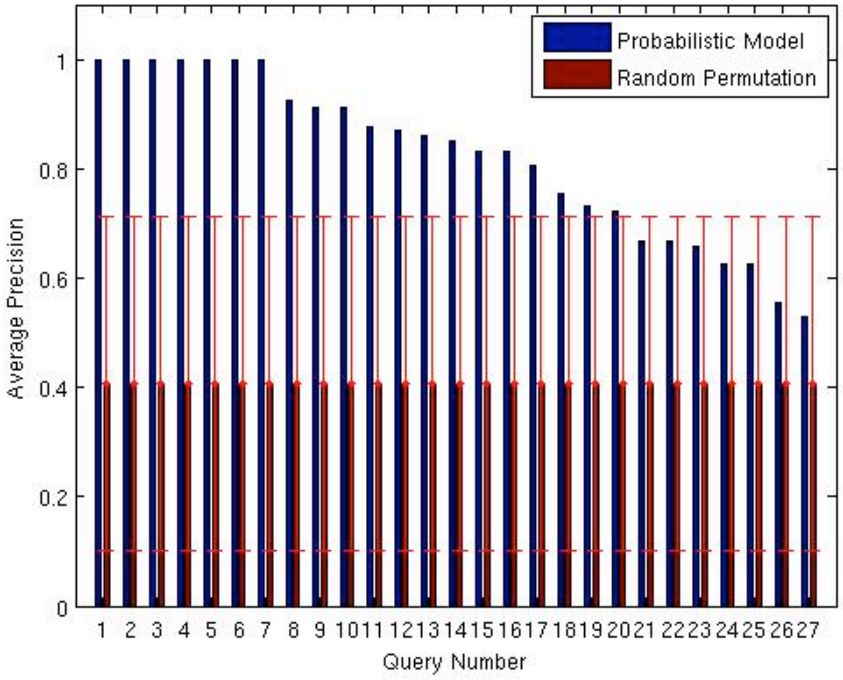
Querying the Model/Database

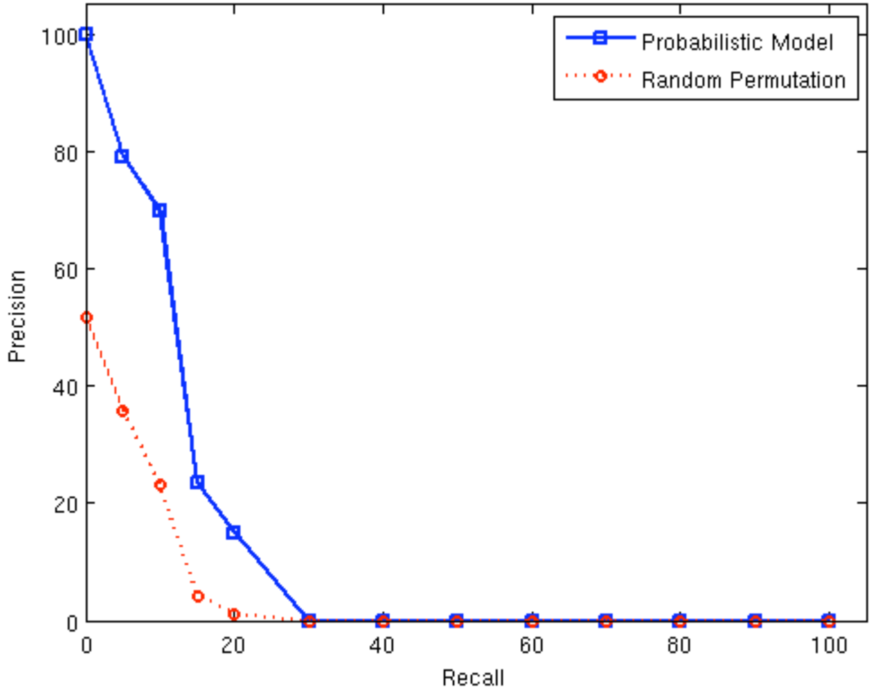
Query with “malignant melanoma” vs “normal” comparison.

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*

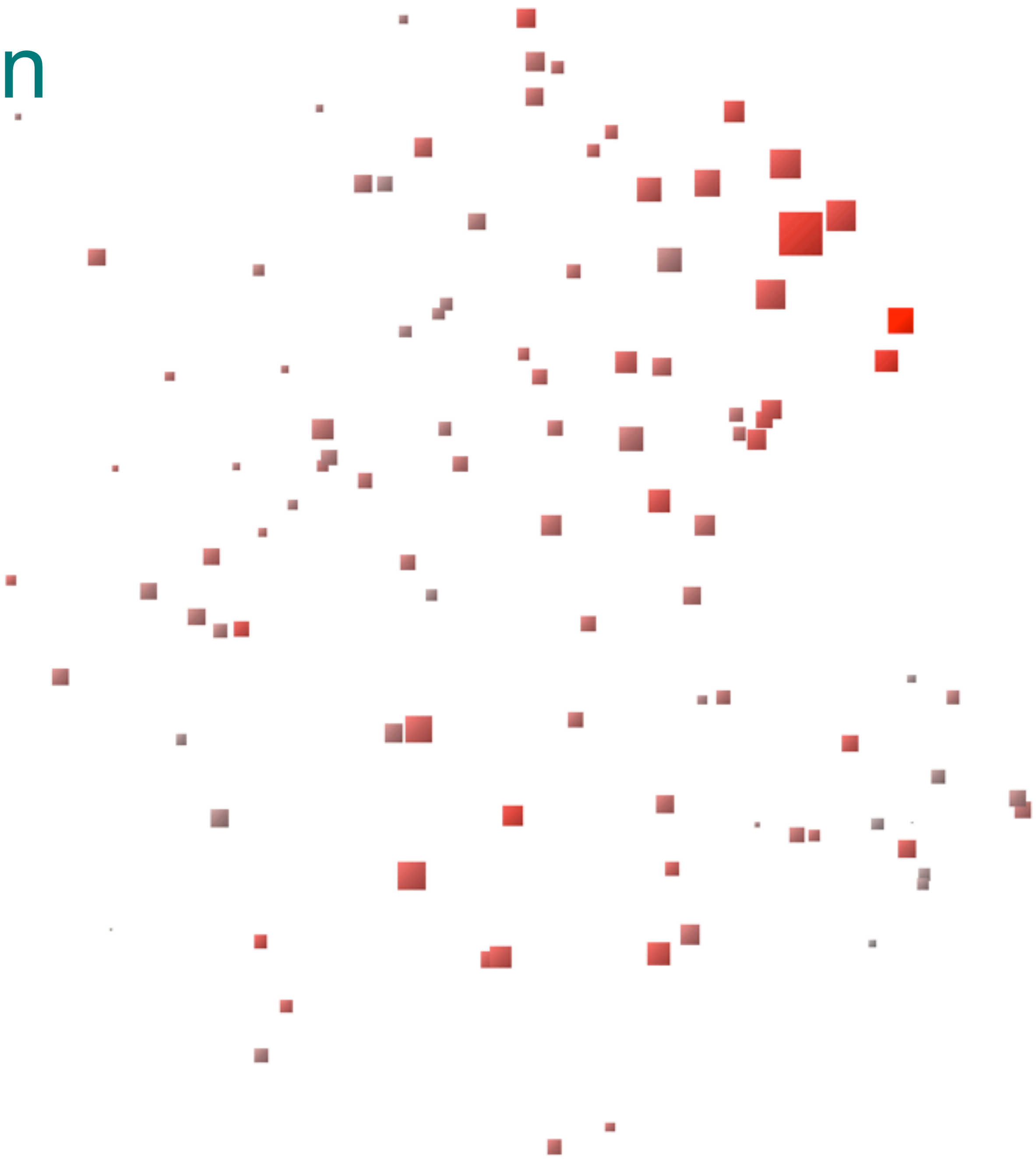
Retrieval results

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





Visualization of retrieval results



Conclusions

Summary: Content-based query for experiments with an experiment:

- Model of an experiment
  - bring in controls / experimenter’s intent
  - bring in biology/task-dependent prior knowledge
- Model of an experiment collection
  - data-driven machine learning
- Retrieval: *p(query|experiment)*
- Visualization of results: NeRV

More information at

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

MLSP 2010

2010 IEEE International Workshop on MACHINE LEARNING FOR SIGNAL PROCESSING

August 29 - September 1, 2010 Kittilä, Finland

Levi Summit conference and exhibition centre

`http://mlsp2010.conwiz.dk`

IEEE

CALL FOR PAPERS

The twentieth workshop in the series of workshops sponsored by IEEE Signal Processing Society will present the most recent and exciting contributions in machine learning for signal processing through keynote talks as well as special and regular single-track sessions. Papers are solicited that cover various aspects of machine learning for signal processing, as outlined in the following.

Machine Learning for Signal Processing

Machine learning in multi-dimensional and statistical signal processing is concerned with tasks such as detection, estimation, prediction, classification, and optimization. Typical approaches are modern implementations of supervised, unsupervised, reinforcement and semi-supervised learning, for instance using probabilistic modeling and kernel methods.

Machine learning has a wide range of applications: adaptive filtering, time-series analysis, pattern recognition, image processing, computer vision, data mining and visualization, information retrieval, robot control, data fusion, blind source separation, sparse and structured representations, context modeling, multimodal interfaces, neuroinformatics, bioinformatics, sensor networks, cognitive radio,

Organization

General chair: Erkki Oja
Program chairs:
Samuel Kaski, David Miller
Special session chairs:
Mikko Kurimo, Samy Bengio
Publicity chairs:
Marc Van Hulle, Jaakko Peltonen
Web and publication chairs:
Antti Honkela, Jan Larsen
Data competition: Kenneth Hild,
Vince Calhoun, Mikko Kurimo
Local arrangements chair:
Tapani Raiko

Schedule

Submission of full papers:	**April 1**
Notification of acceptance:	**May 28**
Camera-ready paper and author registration:	**June 18**
Advance registration before:	**June 23**