

Welcome to

MATH:7450 (22M:305) Topological Data Analysis

Office hours:

MWF 15:45 - 16:20 GMT (10:45 - 11:20 CDT),

M 2:00 - 3:00 am GMT (9pm - 10pm CDT)

and by appointment.

Office hours will be held in our online classroom
(same URL for entering class).

I am also available via google+, skype, and in
person at the University of Iowa.

Aug 30	<p>Download Mapper for Matlab Python Mapper Graphviz Web tool: Progression Analysis of Disease - PAD (includes Mapper) Ayasdi Iris, academic trial</p>
	<p>Additional readings: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition G. Singh , F. Memoli, G. Carlsson (2007) Topology and data, G Carlsson (2009) DNA MICROARRAY VIRTUAL LAB, youtube video Pearson Product-Moment Correlation</p>

Application 1: breast cancer gene expression

Data: microarray gene expression data from 2 data sets, NKI and GSE2034

Distance: pearson correlation distance

Filters: (1) L-infinity centrality:

$$f(x) = \max\{d(x, p) : p \text{ in data set}\}$$

captures the structure of the points far removed from the center or norm.

(2) NKI: survival vs. death

GSE2034: no relapse vs. relapse

Clustering: Single linkage.

Array CGH: The Complete Process

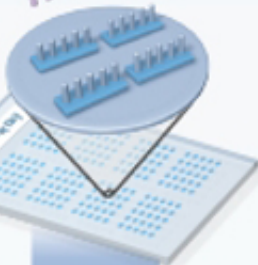
Step 1

Patient DNA Control DNA

Step 2

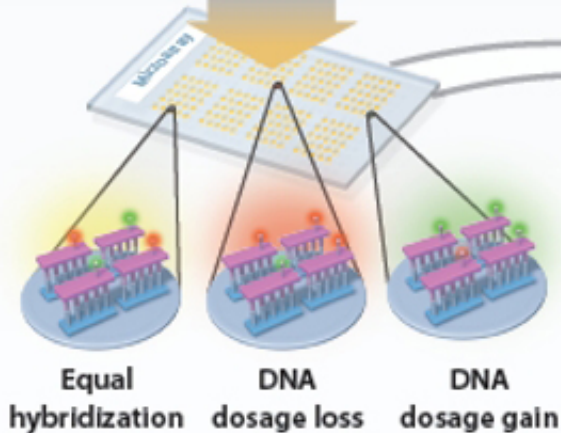


Step 3



Step 4

HYBRIDIZATION



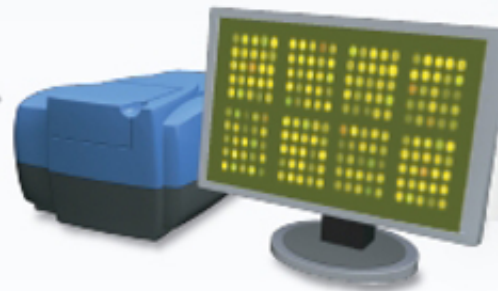
Steps 1-3 Patient and control DNA are labeled with fluorescent dyes and applied to the microarray.

Step 4 Patient and control DNA compete to attach, or hybridize, to the microarray.

Step 5 The microarray scanner measures the fluorescent signals.

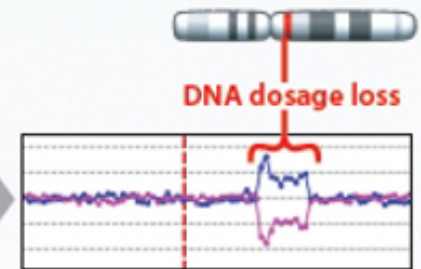
Step 6 Computer software analyzes the data and generates a plot.

Step 5



COMPUTER SOFTWARE

Step 6

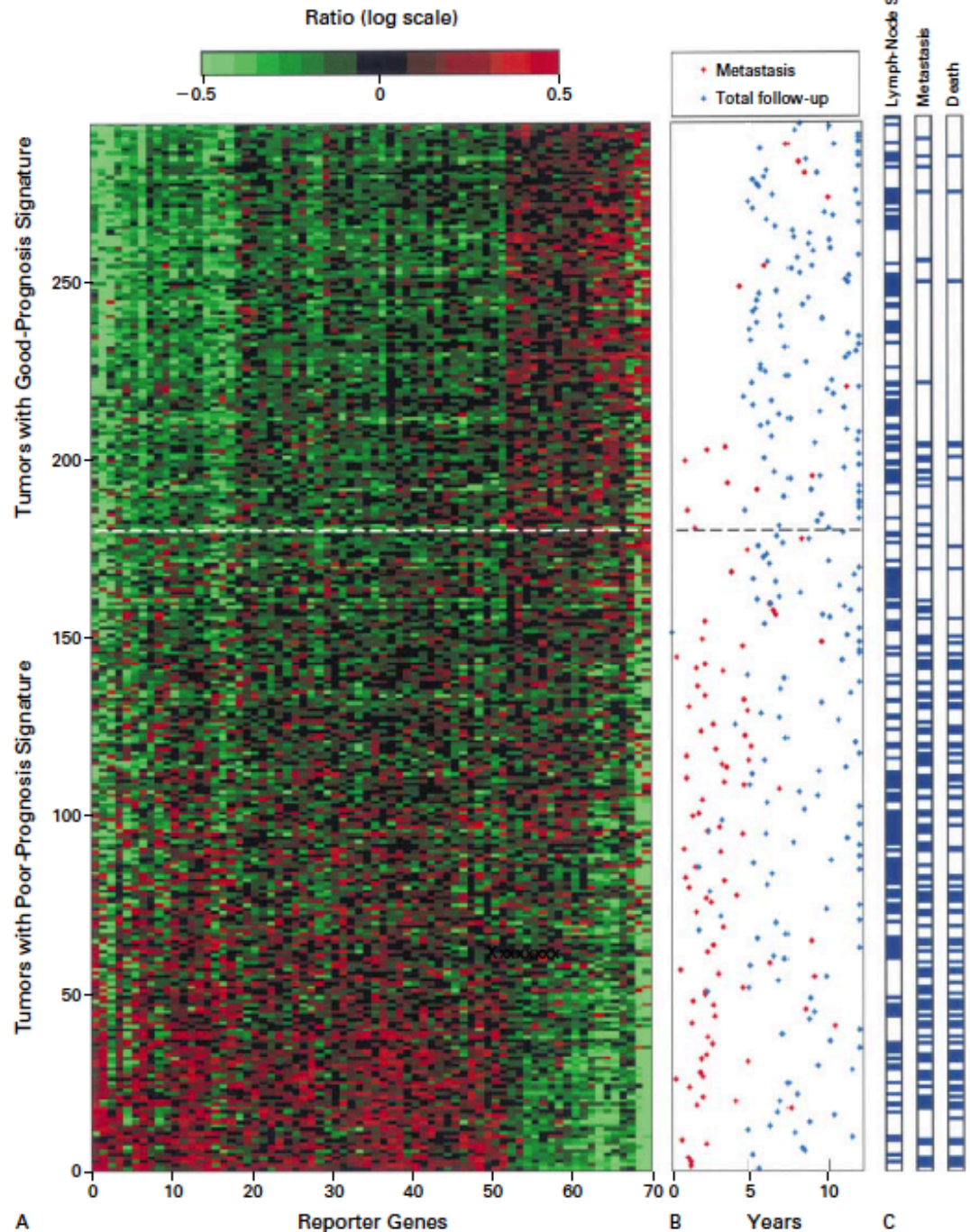


DATA PLOT
(Chromosome 7)

Gene expression profiling predicts clinical outcome of breast cancer

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernardis R, Friend SH

Nature. 2002 Jan 31;415(6871):530-6.



http://bioinformatics.nki.nl/data.php



Bioinformatics and Statistics

Division of Molecular Carcinogenesis, Netherlands Cancer Institute



About

People

Research

Publications

Software

Data

Vacancies

Student Projects

Courses

Meetings

Intranet

Contact

Data

- **Gene expression profiling predicts clinical outcome of breast cancer**
van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH
Nature 2002 Jan 31;415(6871):530-6.

Title: Software
Address: <http://bioinformatics.nki.nl/software.php>

Patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70-80% of patients receiving this treatment would have survived without it. None of the signatures of breast cancer gene expression reported to date allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients without tumour cells in local lymph nodes at diagnosis (lymph node negative). In addition, we established a signature that identifies tumours of BRCA1 carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and angiogenesis. This gene expression profile will outperform all currently used clinical parameters in predicting disease outcome. Our findings provide a strategy to select patients who would benefit from adjuvant therapy.

- Data can be downloaded [here](#).

2 breast cancer data sets:

1.) NKI (2002):

gene expression levels of 24,000 from 272 tumors. Includes node-negative and node-positive patients, who had or had not received adjuvant systemic therapy. Also includes survival information.

2.) GSE203414 (2005)

expression of 22,000 transcripts from total RNA of frozen tumour samples from 286 lymph-node-negative patients who had not received adjuvant systemic treatment. Also includes time to relapse information.

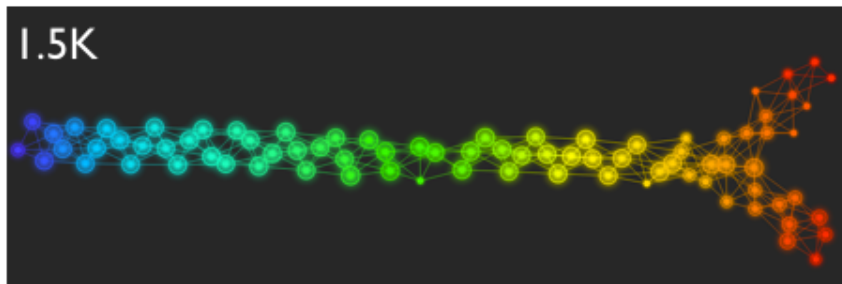
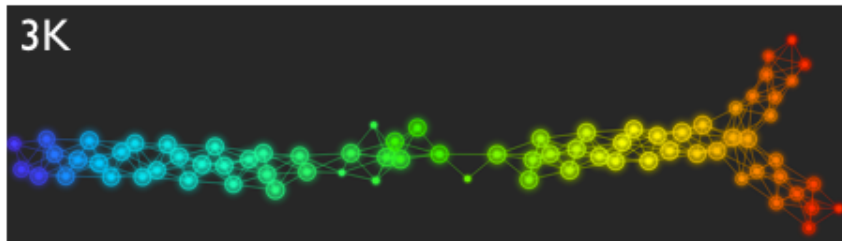
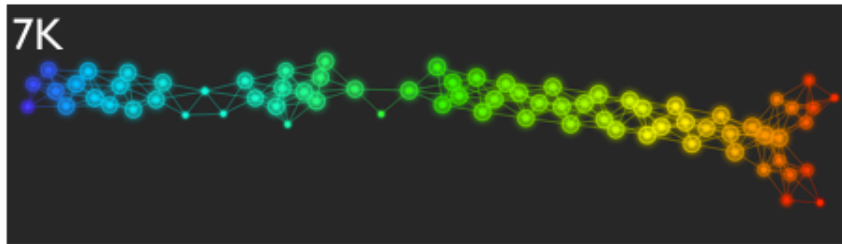
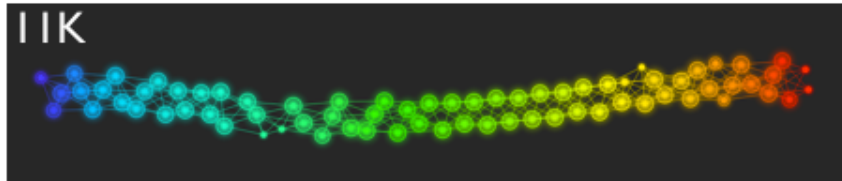
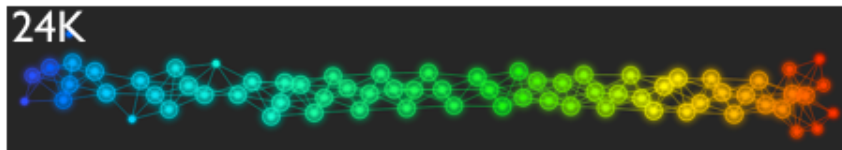


Fig. S1. Shape of the data becomes more distinct as the analysis columns are restricted to the top varying genes.

24K: all the genes on the microarray were used in the analysis;

11K: 10,731 top most varying genes were used in the analysis;

7K: 6,688 top most varying genes were used in the analysis;

3K: 3,212 top most varying genes were used in the analysis;

1.5K: 1,553 top most varying genes were used in the analysis.

Graphs colored by the L-infinity centrality values. Red: high; Blue: low

Comparison of our results with those of Van de Vijver and colleagues is difficult because of differences in patients, techniques, and materials used.

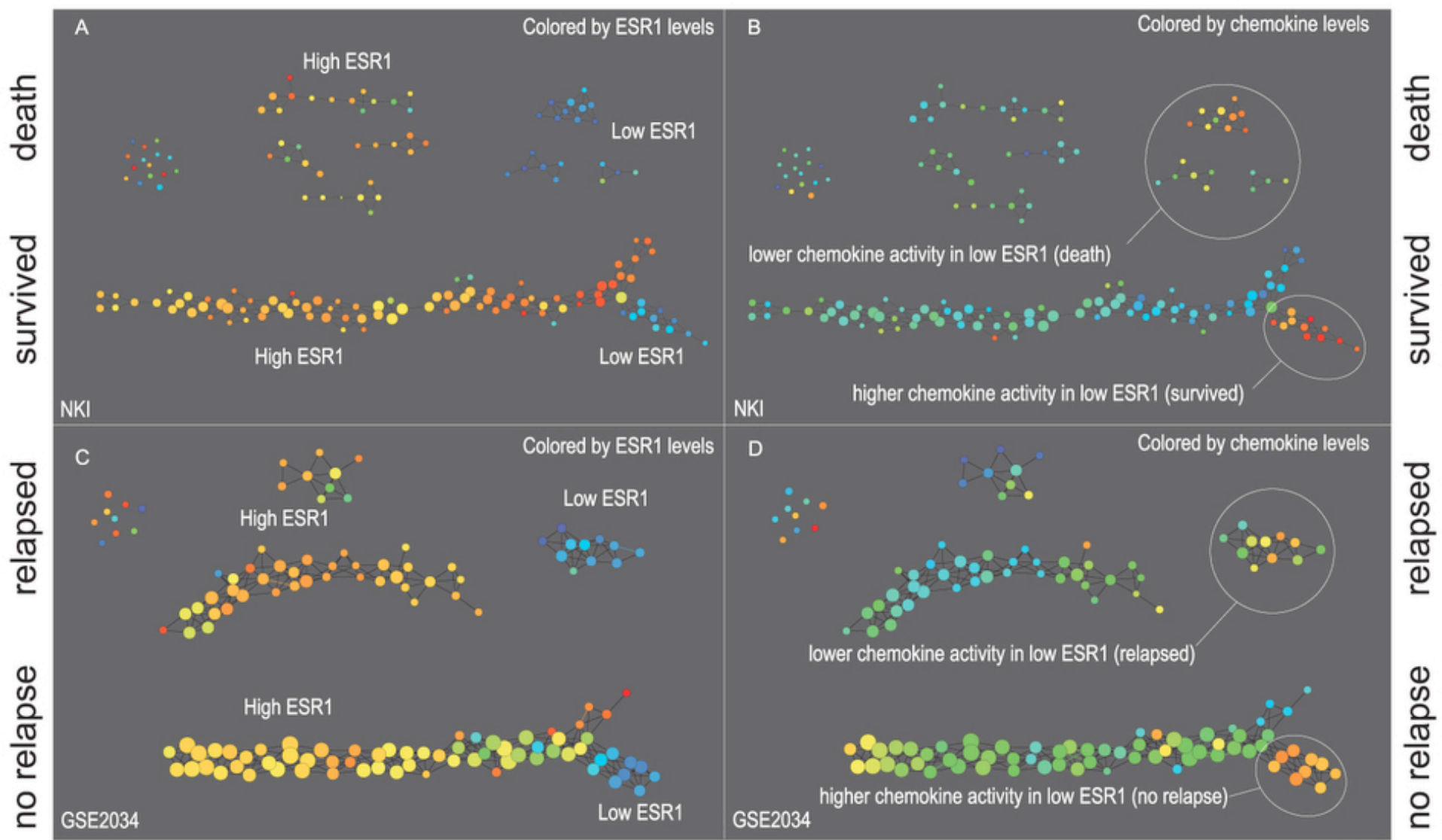
Their study included node-negative and node-positive patients, who had or had not received adjuvant systemic therapy, and only women younger than 53 years.

microarray platforms used in the studies differ—Affymetrix and Agilent.

Of the 70 genes in the study by van't Veer and co-workers, 48 are present on the Affymetrix U133a array, whereas only 38 of our 76 genes are present on the Agilent array. There is a three-gene overlap between the two signatures (cyclin E2, origin recognition complex, and TNF superfamily protein).

Despite the apparent difference, both signatures included genes that identified several common pathways that might be involved in tumour recurrence. This finding supports the idea that although there might be redundancy in gene members, effective signatures could be required to include representation of specific pathways.

From: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Yixin Wang et al, *The Lancet*, Volume 365, Issue 9460, 19–25 February 2005, Pages 671–679, <http://www.sciencedirect.com/science/article/pii/S0140673605179471>

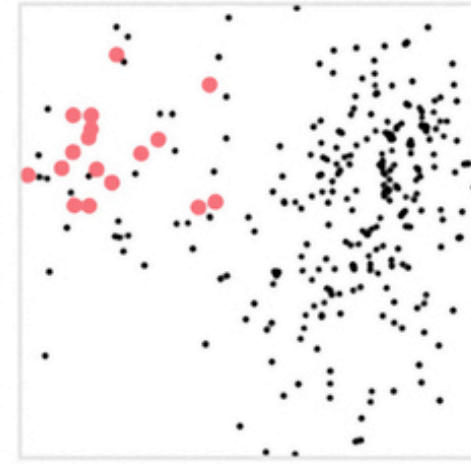
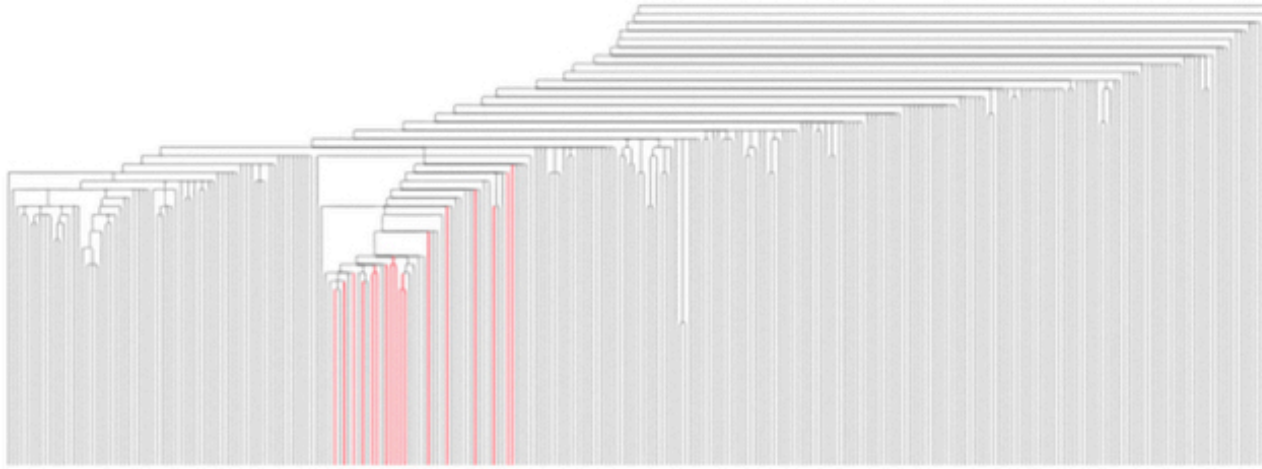


Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn't survive, the bottom half are the patients who survived. Panels C and D are similar to panels A and B except that one of the filters is relapse instead of survival. Panels A and C are colored by the average expression of the ESR1 gene. Panels B and D are colored by the average expression of the genes in the KEGG chemokine pathway. Metric: Correlation; Lens: L-Infinity Centrality (Resolution 70, Gain 3.0x, Equalized) and Event Death (Resolution 30, Gain 3.0x). Color bar: red: high values, blue: low values.

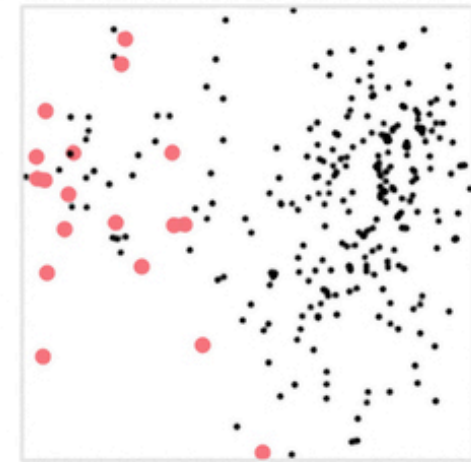
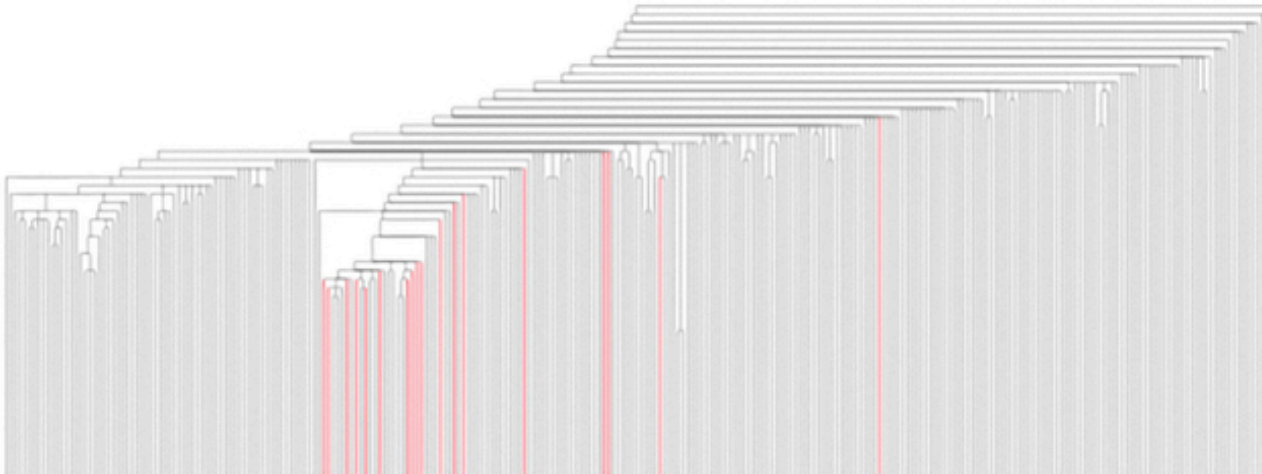
Clustering

PCA

ER- did not survive

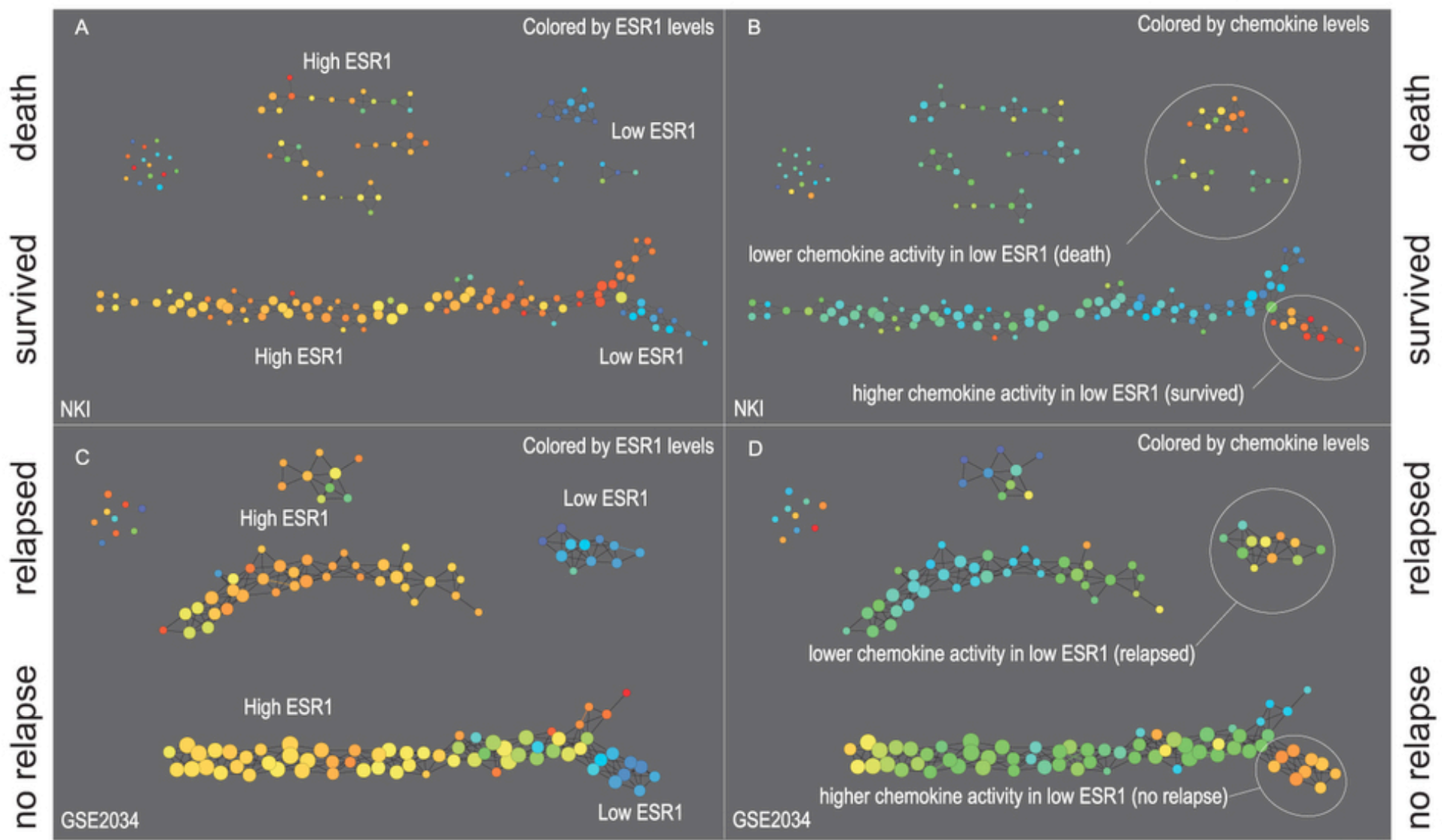


ER- survived



Highlighted in red are the lowERNS (top panel) and the lowERHS (bottom panel) patient subgroups.

<http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html>



Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn't survive, the bottom half are the patients who survived. Panels C and D are similar to panels A and B except that one of the filters is relapse instead of survival. Panels A and C are colored by the average expression of the ESR1 gene. Panels B and D are colored by the average expression of the genes in the KEGG chemokine pathway. Metric: Correlation; Lens: L-Infinity Centrality (Resolution 70, Gain 3.0x, Equalized) and Event Death (Resolution 30, Gain 3.0x). Color bar: red: high values, blue: low values.

Mapper Software

http://comptop.stanford.edu/pad/



Progression Analysis of Disease—*PAD*

A web tool for the data analysis method introduced in:

M. Nicolau, A. Levine, G. Carlsson: *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, Proc. Natl. Acad. Sci. USA (2011)

PAD is a data analysis method that integrates two methods:

Step 1: DSGA (*Disease Specific Genomic Analysis*) highlights the disease aspect of the data.

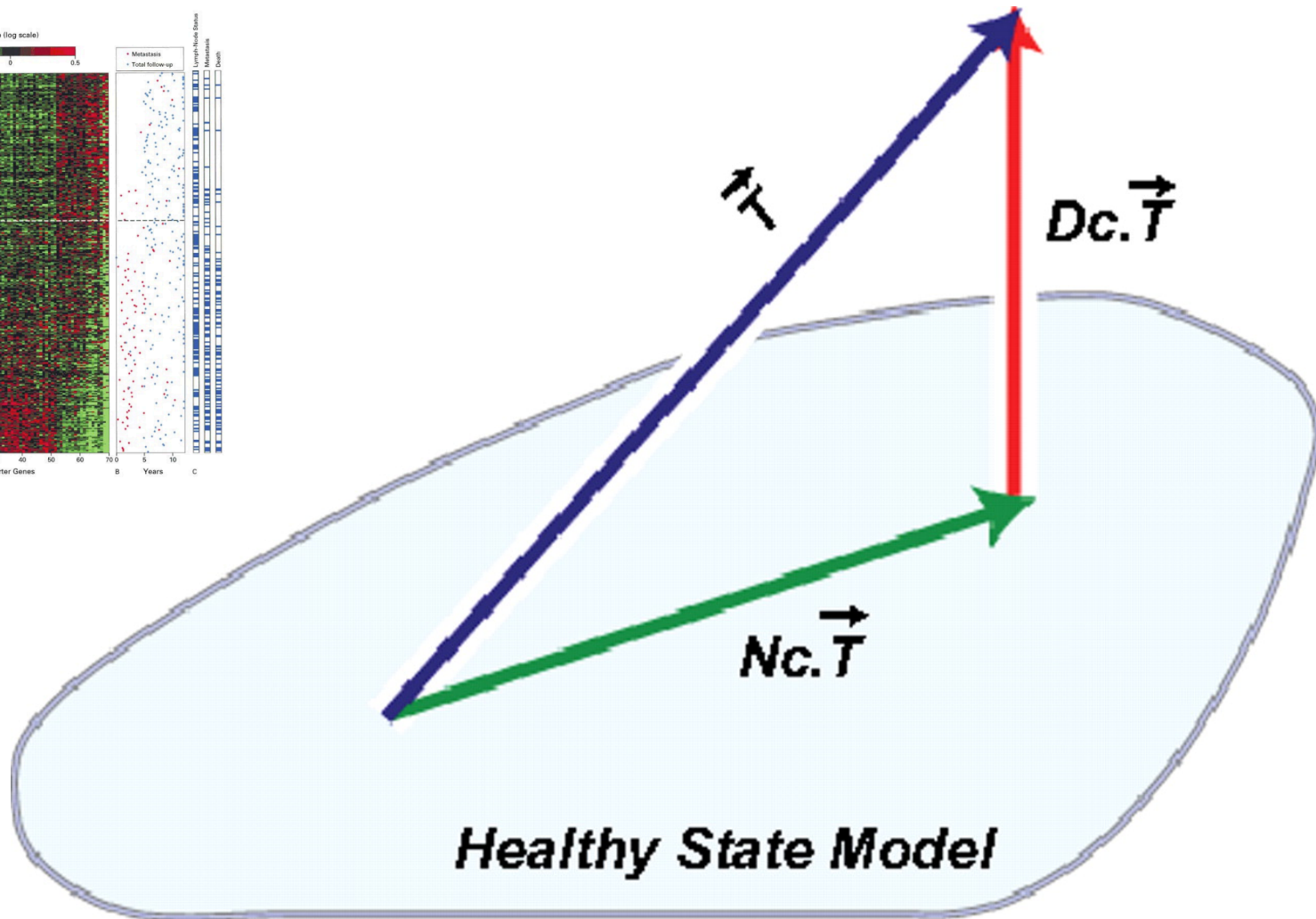
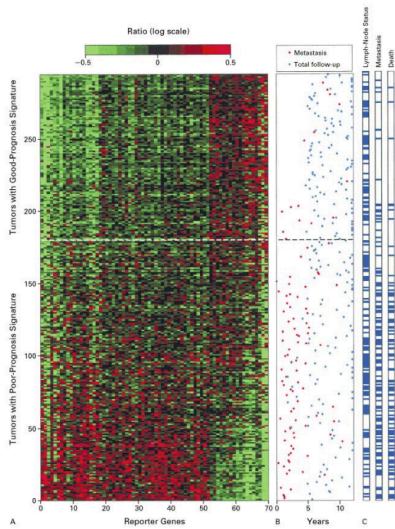
Step 2: Mapper identifies shape characteristics in the data.

A [tutorial](#) is available as a PDF document.

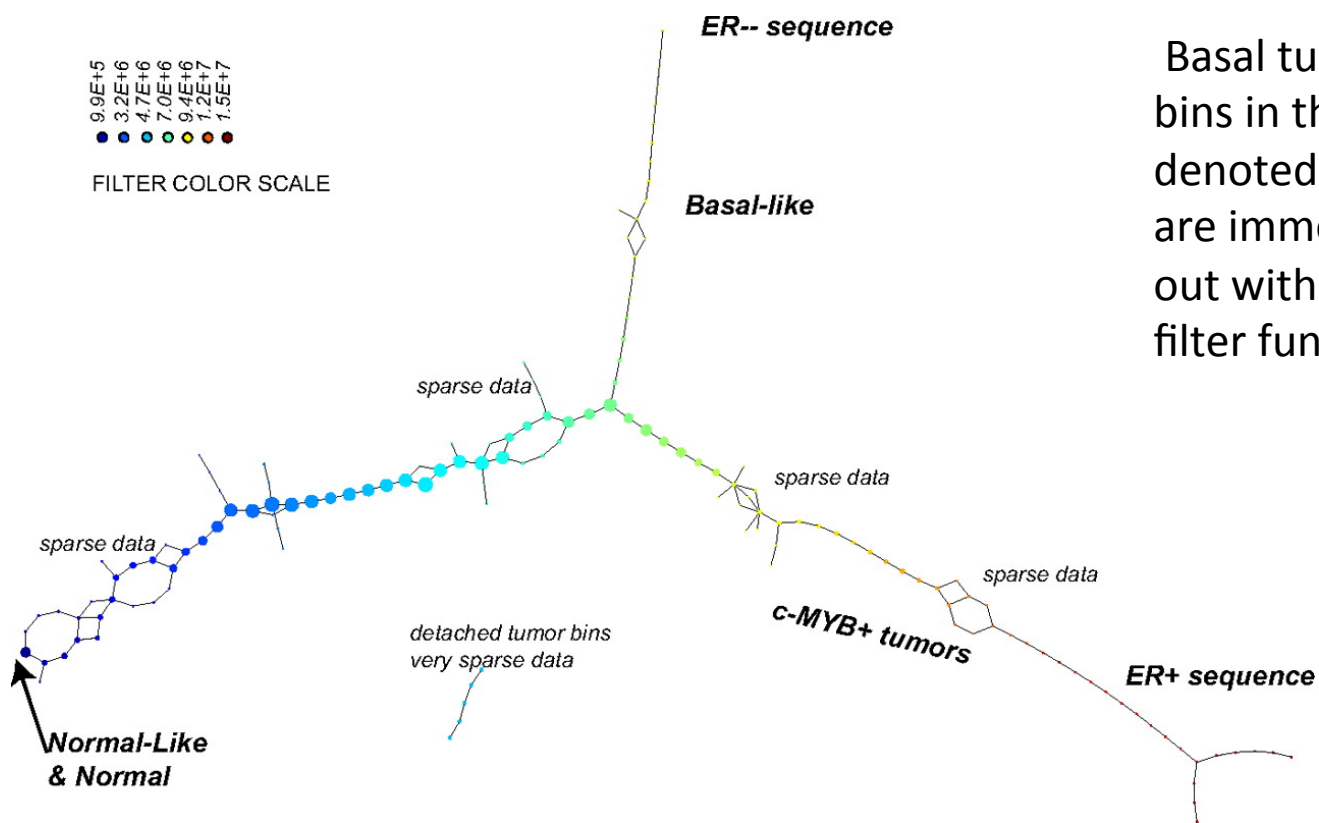
Upload normal data (max. 200 MB): No file chosen

Upload tumor data (max. 200 MB): No file chosen

DSGA decomposition of the original tumor vector into the Normal component its linear models fit onto the Healthy State Model and the Disease component vector of residuals.



Nicolau M et al. PNAS 2011;108:7265-7270



Basal tumors occupy most of the bins in the tumor sequence denoted as ER- sequence. They are immediately visible and stand out with large value (red) in the filter function

Normal tissue samples all fall in the same bin together with 15 additional ER+ tumors.

The known group of her2+ tumors is not yet visible, owing to the well-understood problem that only a small number of genes (on 17q) identify it, making them mathematically less visible, despite the fact that the small number of coordinates (17q genes) are biologically important.

A long tumor sequence on the graph, the ER+ sequence showing large deviation from normal, is visible, as defined by the filter. This tumor sequence also consists of ER+ tumors, but unlike the first (blue) group of tumors, these are distinct from normal tissue in that the value of the filter function—the L_p magnitudes of the tumor vectors in these bins—is very large.

http://math.stanford.edu/~muellner/mapper/



Welcome to the Python Mapper documentation!

Mapper is an algorithm for exploration, analysis and visualization of data.

- [What is Python Mapper?](#)
- [Installation](#)
 - [Requirements](#)
 - [Download](#)
 - [Installation](#)
 - [Troubleshooting](#)
 - [Mixed tips](#)

Table Of Contents

Welcome to the Python Mapper documentation!

- [Indices and tables](#)

Next topic

[What is Python Mapper?](#)

This Page

[Show Source](#)

Quick search

Go

Enter search terms or a module, class or function name.

- Compiling the documentation
- Quick start
- Input data
- Filter functions
 - Mathematical definition
 - Data structure
 - Filter functions in Python Mapper
- Cover methods
- Custom data processing in the GUI
 - Input data processing
 - Filter processing
- Copyright, references and citation info
 - Copyright
 - References
 - Citation info



Ayasdi Iris



Ayasdi Platform

API

Topological Data Analysis

Machine-learning algorithms

Scalable computing and distributed data store



Public or Proprietary Data

Ayasdi Iris

Ayasdi Iris is the world's first Insight Discovery solution that uses TDA to highlight the underlying geometric shapes of your data and allowing for real-time interaction to produce immediate insights.

Ayasdi Iris is offered as a multi-tenant cloud or as an on-premise solution, and is applicable for any organization needing to discover insights from complex data.

www.ayasdi.com/product/

<http://www.ayasdi.com/inquiry/academic-trial.html>

www.ayasdi.com/inquiry/academic-trial.html



AYASDI

Transform

Academic Free Trial Application

In order for your project to be considered for a 3-months free access to Iris, please fill in the form below and submit a one page proposal that highlights the project you are working on. Applications without this proposal cannot be considered.

GET STARTED

First Name*

Email Address - Address from Academic Required*

Organization*

Department

Matlab version demonstration