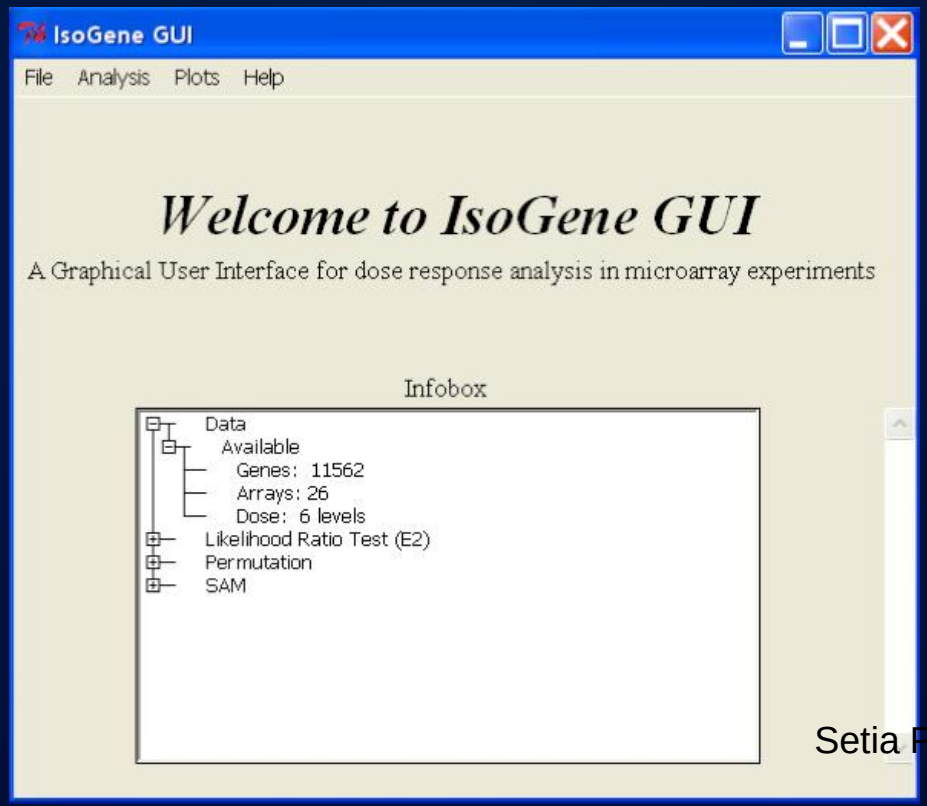
R for Bioinformatic

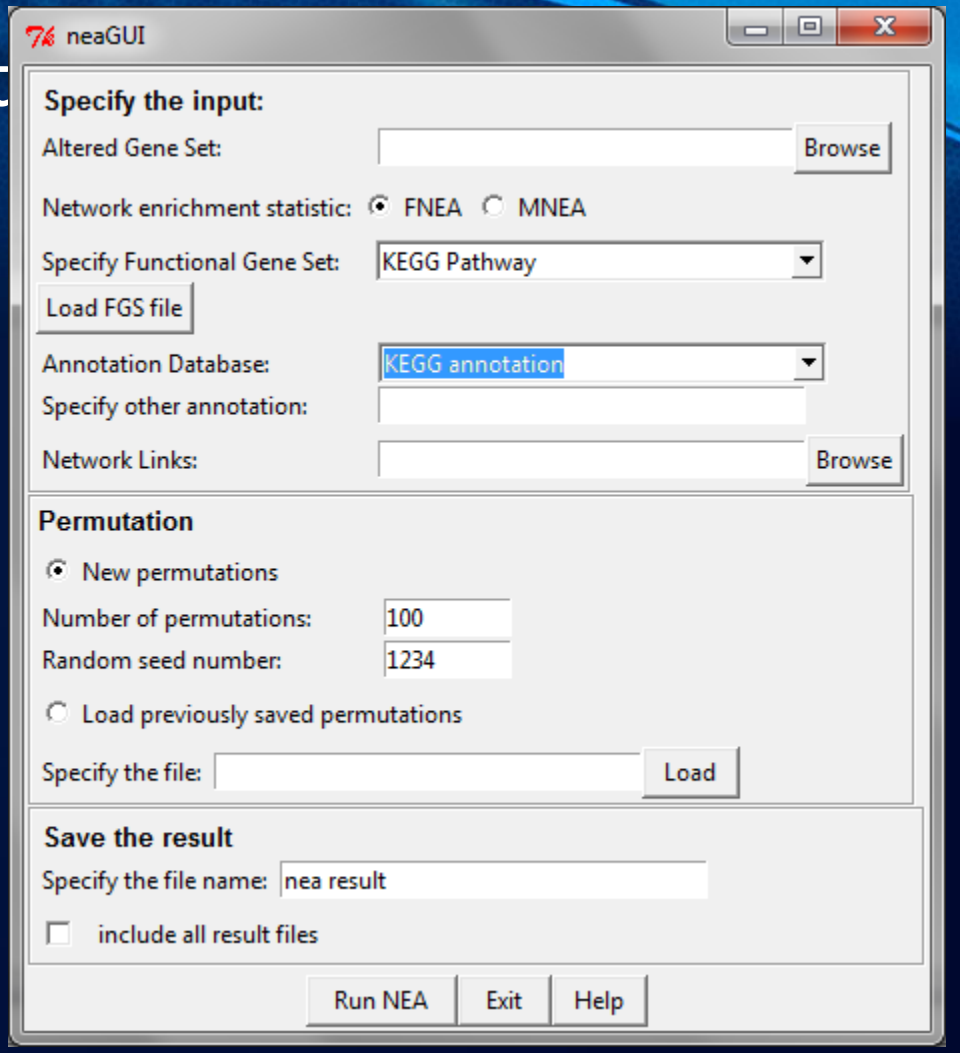
Setia Pramana

My R Packages

- IsoGene
- IsoGeneGUI
- nea
- neaGUI
- biclustGUI
- OCRME
- More detail: http://setiopramono.wordpress.com/r-programming/

RGUI Developed using t

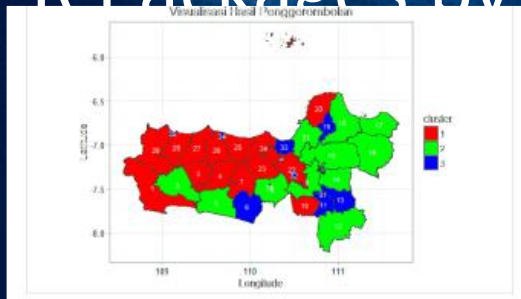


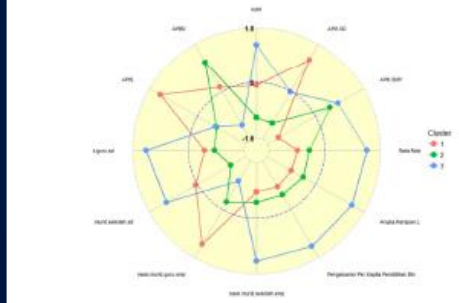


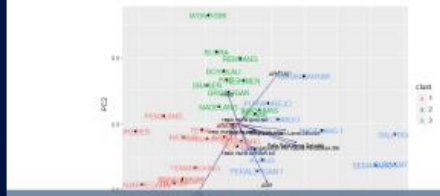
R Packages by STIS

Name	Title	Brief Description	Author	Repository
spatialClust	Spatial Clustering	Clustering analysis with pay attentation on membership via spatial effects	Imam Habib Pamungkas, Setia Pramana	CRAN
advclust	S4 Object Oriented for Advanced Clustering(Fuzzy Clustering and Cluster Ensemble)	Advance on clustering with fuzzy clustering for overlapping cluster and objects on gray area. Cluster Ensemble performs combining several result as one robust and stable result.	Achmad Fauzi Bagus F, Setia Pramana	CRAN
RcmdrPlugin.FuzzyClust	R commander plugin for fuzzy clustering	Graphical User interface via Rcmdr Plugin for fuzzy clustering analysis	Achmad Fauzi Bagus F, Setia pramana	CRAN
MetaheuristicFPA	Metaheuristic with Flower Pollinantion Algorithm	Optimization of function objectives to get global optimum of parameter by using Flower Pollination Algorithm	Amanda Pratama Putra, Margaretha Ari Anggorowati	CRAN
Multiplier	Social Accounting Matrix and Finansial Social Accounting Matrix	Graphical User Interface for performing SAM (Social Accounting Matrix) and FSAM (Financial Social Accounting Matrix)	Tiara Ratna Dewi, Aisyah Fitri Yuniarshi	R-Forge
RcmdrPlugin.PCARobust	Robust PCA plugin for Rcmdr	Graphical User Interface for Robust Principal Component Analysis (PCA) with Hubert Algorithm for Dimension Reduction	Monalisa Sipahutar, Setia Pramana	CRAN

R Packages by STIS







GEOGRAPHICALLY WEIGTHED CLUSTERING-GRAVITATIONAL

spatialClust

An R Package for Cluster Spatial Data

Available on CRAN:

September 3rd 2016

"spatialClust" - an R package

by Imam Habib Pamungkas, S.S.T and Setia Pramana, Ph.D

R Based Applications by STIS

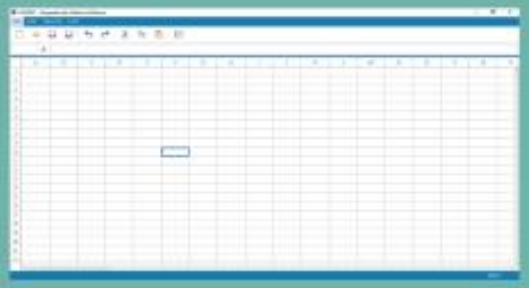
- Kalingga
- Muria

Asgard Alpha Version

ASGARD is a statistics software used to perform geographically weighted regression (GWR). This software was made in 2016 and currently contains some basic GWR functions like GWR, Geographically Weighted Poisson Regression (GWPR), Geographically Weighted Logistic Regression (GWLR), Geographically Weighted Negative-Binomial Regression (GWNBR) and some Assumption Test related to GWR. In addition, ASGARD is also integrated with the map that make it easier for users to performs analysis.

MAIN FEATURES

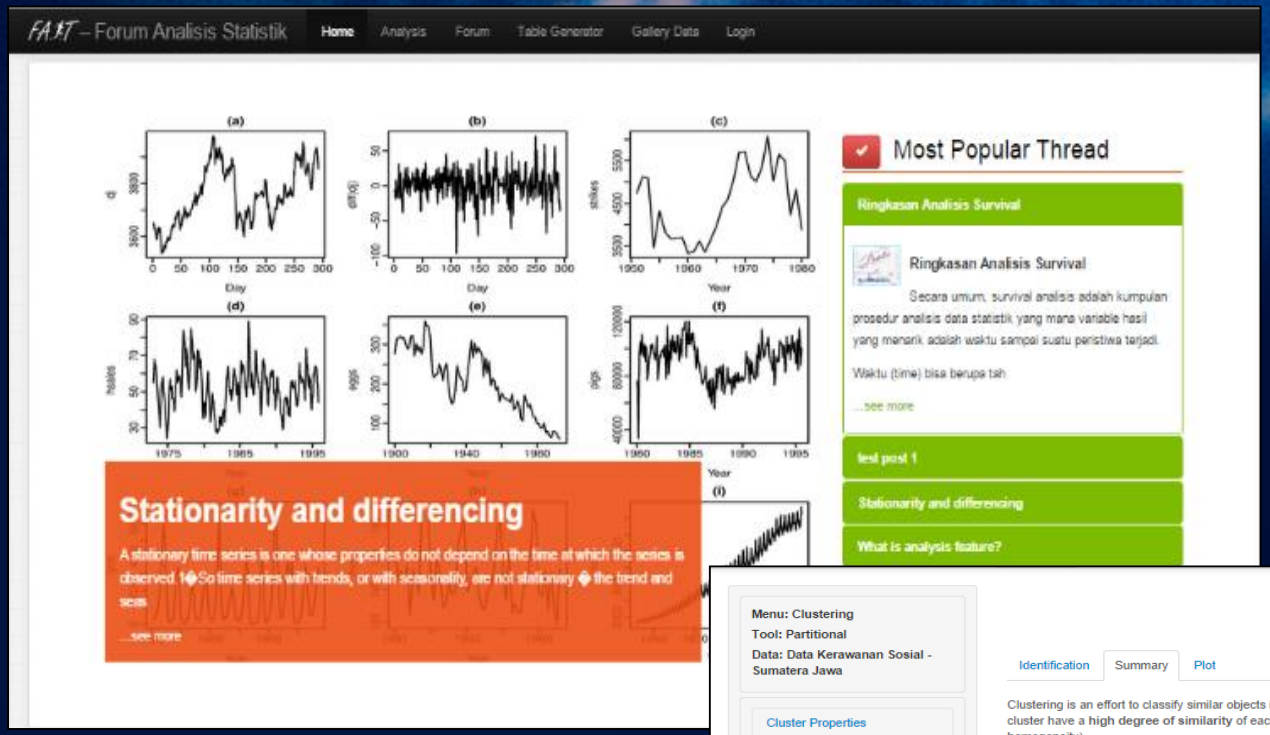
Spreadsheet

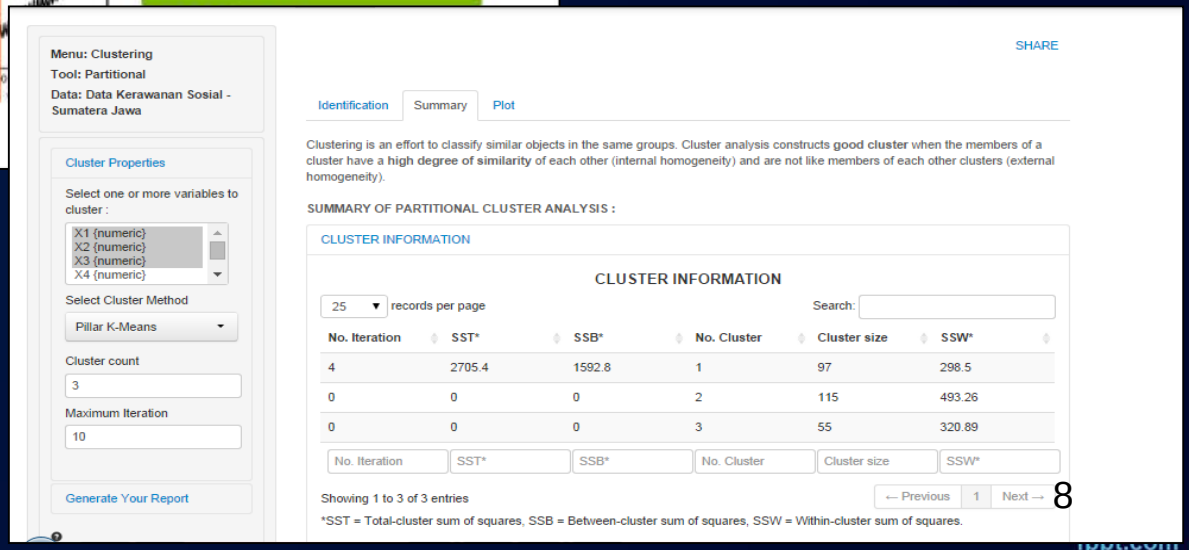


Fairly complete functions
- GWR
- GWPR
- GWLR
- GWNBR
- Variance Inflation Factor
- Breusch-Pagan Test

Map Visualization
Map Visualization can help users to understand the circumstances of the observation area.

FAST


FAST – Forum Analisis Statistik
Home
Analysis
Forum
Table Generator
Gallery Data
Login
Most Popular Thread
Ringkasan Analisis Survival
Ringkasan Analisis Survival
Secara umum, survival analisis adalah kumpulan prosedur analisis data statistik yang mana variable hasil yang menarik adalah waktu sampai suatu peristiwa terjadi.
Waktu (time) bisa berupa tah
...see more
test post 1
Stationarity and differencing
What is analysis feature?
Stationarity and differencing
A stationary time series is one whose properties do not depend on the time at which the series is observed. So time series with trends, or with seasonality, are not stationary — the trend and
...see more


Menu: Clustering
Tool: Partitional
Data: Data Kerawanan Sosial - Sumatera Jawa
Cluster Properties
Select one or more variables to cluster :
X1 {numeric}
X2 {numeric}
X3 {numeric}
X4 {numeric}
Select Cluster Method
Pillar K-Means
Cluster count
3
Maximum Iteration
10
Generate Your Report
SHARE
Identification
Summary
Plot
Clustering is an effort to classify similar objects in the same groups. Cluster analysis constructs good cluster when the members of a cluster have a high degree of similarity of each other (internal homogeneity) and are not like members of each other clusters (external homogeneity).
SUMMARY OF PARTITIONAL CLUSTER ANALYSIS :
CLUSTER INFORMATION
CLUSTER INFORMATION
25 records per page
Search:
No. Iteration
SST*
SSB*
No. Cluster
Cluster size
SSW*
4 2705.4 1592.8 1 97 298.5
0 0 0 2 115 493.26
0 0 0 3 55 320.89
Showing 1 to 3 of 3 entries
← Previous 1 Next →
*SST = Total-cluster sum of squares, SSB = Between-cluster sum of squares, SSW = Within-cluster sum of squares.

RGUI using C#: Wires

- Developed by STIS students
- For Spatial Data Analysis
- Still developing…

RGUI using C#: W

Features

Exploratory Spatial Data Analysis	Provide calculation of spatial autocorrelation based on Moran's I, Gearcy's C, Local Indicators of Spatial Association (LISA)
Spatial Weight Matrix	Spatial interactions among observations
Spatial Clustering	Clustering observation with spatial attributes
Spatial Regression	Regression analysis with spatial depedency
Regional Inequality	Inequality analysis especially on poverty subjects
Spatial Shift Share	Comparing growing rate of several sector based on spatial
Kriging	Imputation on missing data with spatial attributes

WIRES - [New Project : "Unsaved"]

File Thematic Analysis Tools

Data View | Variable View | Map View | Result View

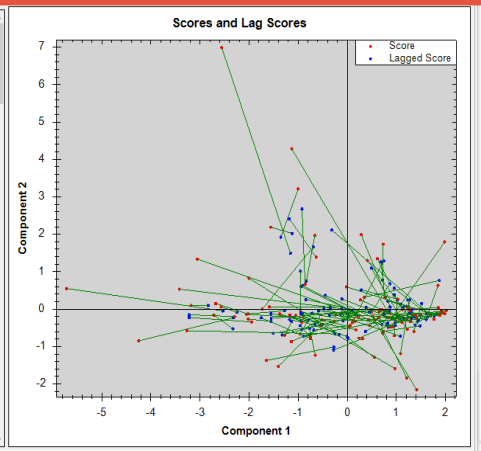
CLASSIFICATION MATRIX

before \ after	0	1
0	90	19
1	3	15

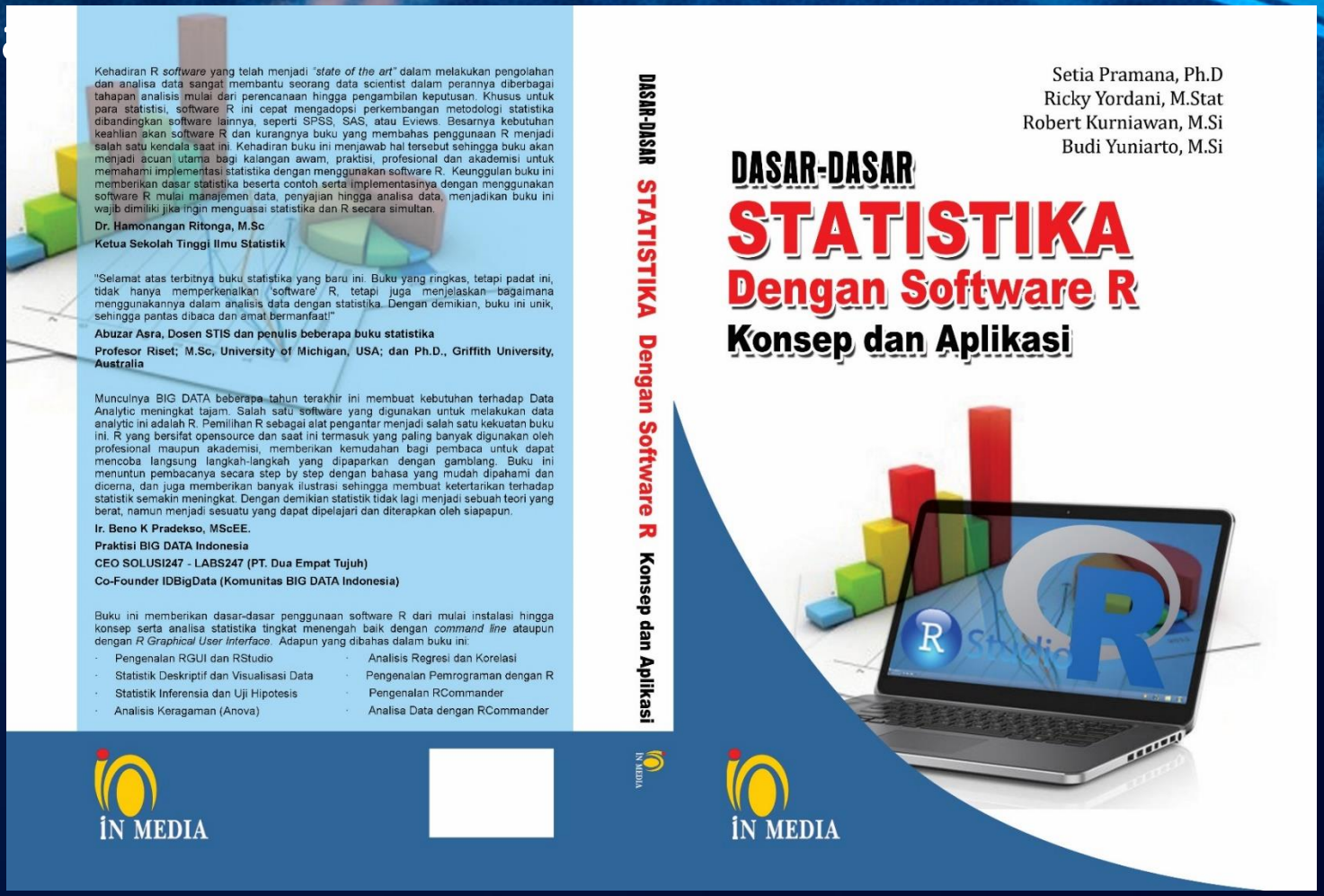
CCR: 82.6772 %

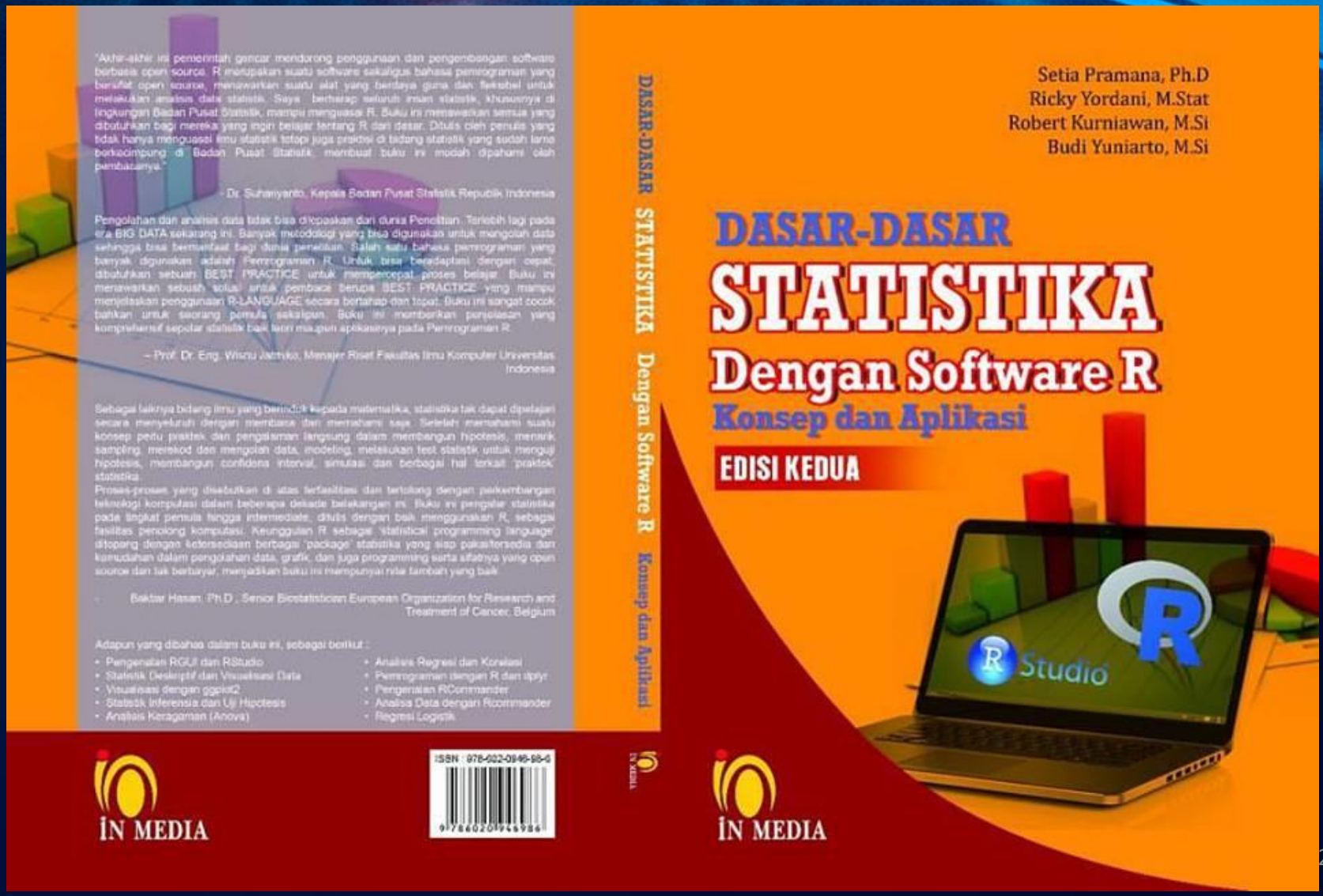
KABKOTA_NO	Before	After
01	0	0
03	0	0
06	0	0
07	0	0
08	0	0
10	0	0
12	0	0
14	0	0
74	0	0
79	0	0
01	0	0
02	0	0
03	0	0
04	0	0
05	0	0
06	0	0
07	0	0
08	0	0

	Score Component 1	Lag Score Component 1	Score Component 2	Lag Score Component 2	Inequality
SELAPARANG	-2.5570	-1.1572	6.9862	1.4920	5.6697
SUMBAWA	-1.1325	0.8284	4.2860	-0.0731	4.7798
DOMPU	-3.0534	0.1945	1.3361	-0.1566	3.5745
WOJA	-3.4202	-0.1159	0.5342	0.0115	3.3454
PRAYA	-5.7190	-2.5356	0.5513	-0.0462	3.2390
AIKMEL	-3.2690	-0.2523	-0.5785	-0.6618	3.0178
MOYO UTARA	1.6801	-0.3151	-0.1405	2.1151	3.0114
RHEE	1.9859	0.8254	1.7949	-0.3738	2.4597
SUKAMULIA	0.9716	-0.9446	-1.5866	-0.1341	2.4045
SELONG	-2.0031	-0.1675	0.8245	-0.6772	2.3716
SEMBALUN	1.3885	-0.8893	-0.1614	-0.7276	2.3471
TAMBORA	1.4137	0.4499	-2.1472	-0.0816	2.2794
KURIPAN	0.6826	-1.5507	-0.0123	-0.2446	2.2453
SOROMANDI	1.2874	-0.8441	-0.3371	0.2568	2.2127
PUJUT	-4.2471	-2.3220	-0.8469	-0.2324	2.0208
RABA	0.2869	0.8763	1.9864	0.0585	2.0160
MATARAM	-0.8592	-0.9215	0.6684	2.6770	2.0096
ALAS	0.4113	1.3838	1.3021	-0.4249	1.9820
GERUNG	-2.0509	-0.0880	-0.1390	-0.3074	1.9701
GUNUNG SARI	-1.7310	-0.6951	-0.0028	1.6604	1.9594
KAYANGAN	0.9773	-0.9436	-0.2941	-0.3153	1.9210
TARANO	1.2161	-0.5865	-0.2837	-0.0498	1.8177
TALIWANG	-0.0186	1.6467	0.5946	-0.0904	1.8007



A St


Kehadiran R *software* yang telah menjadi *"state of the art"* dalam melakukan pengolahan dan analisa data sangat membantu seorang data scientist dalam perannya diberbagai tahapan analisis mulai dari perencanaan hingga pengambilan keputusan. Khusus untuk para statistisi, software R ini cepat mengadopsi perkembangan metodologi statistika dibandingkan software lainnya, seperti SPSS, SAS, atau Eviews. Besarnya kebutuhan keahlian akan software R dan kurangnya buku yang membahas penggunaan R menjadi salah satu kendala saat ini. Kehadiran buku ini menjawab hal tersebut sehingga buku akan menjadi acuan utama bagi kalangan awam, praktisi, profesional dan akademisi untuk memahami implementasi statistika dengan menggunakan software R. Keunggulan buku ini memberikan dasar statistika beserta contoh serta implementasinya dengan menggunakan software R mulai manajemen data, penyajian hingga analisa data, menjadikan buku ini wajib dimiliki jika ingin menguasai statistika dan R secara simultan.
Dr. Hamonangan Ritonga, M.Sc
Ketua Sekolah Tinggi Ilmu Statistik
"Selamat atas terbitnya buku statistika yang baru ini. Buku yang ringkas, tetapi padat ini, tidak hanya memperkenalkan 'software' R, tetapi juga menjelaskan bagaimana menggunakannya dalam analisis data dengan statistika. Dengan demikian, buku ini unik, sehingga pantas dibaca dan amat bermanfaat!"
Abuzar Asra, Dosen STIS dan penulis beberapa buku statistika
Profesor Riset; M.Sc, University of Michigan, USA; dan Ph.D., Griffith University, Australia
Munculnya BIG DATA beberapa tahun terakhir ini membuat kebutuhan terhadap Data Analytic meningkat tajam. Salah satu software yang digunakan untuk melakukan data analytic ini adalah R. Pemilihan R sebagai alat pengantar menjadi salah satu kekuatan buku ini. R yang bersifat opensource dan saat ini termasuk yang paling banyak digunakan oleh profesional maupun akademisi, memberikan kemudahan bagi pembaca untuk dapat mencoba langsung langkah-langkah yang dipaparkan dengan gamblang. Buku ini menuntun pembacanya secara step by step dengan bahasa yang mudah dipahami dan dicerna, dan juga memberikan banyak ilustrasi sehingga membuat ketertarikan terhadap statistik semakin meningkat. Dengan demikian statistik tidak lagi menjadi sebuah teori yang berat, namun menjadi sesuatu yang dapat dipelajari dan diterapkan oleh siapapun.
Ir. Beno K Pradekso, MScEE.
Praktisi BIG DATA Indonesia
CEO SOLUSI247 - LABS247 (PT. Dua Empat Tujuh)
Co-Founder IDBigData (Komunitas BIG DATA Indonesia)
Buku ini memberikan dasar-dasar penggunaan software R dari mulai instalasi hingga konsep serta analisa statistika tingkat menengah baik dengan *command line* ataupun dengan *R Graphical User Interface*. Adapun yang dibahas dalam buku ini:
· Pengenalan RGUI dan RStudio
· Statistik Deskriptif dan Visualisasi Data
· Statistik Inferensia dan Uji Hipotesis
· Analisis Keragaman (Anova)
· Analisis Regresi dan Korelasi
· Pengenalan Pemrograman dengan R
· Pengenalan RCommander
· Analisa Data dengan RCommander
İN MEDIA
DASAR-DASAR STATISTIKA Dengan Software R Konsep dan Aplikasi
İN MEDIA
Setia Pramana, Ph.D
Ricky Yordani, M.Stat
Robert Kurniawan, M.Si
Budi Yuniarto, M.Si
DASAR-DASAR
STATISTIKA
Dengan Software R
Konsep dan Aplikasi


"Akhir-akhir ini pemerintah gencar mendorong penggunaan dan pengembangan software berbasis open source. R merupakan suatu software sekaligus bahasa pemrograman yang bersifat open source, menawarkan suatu alat yang berdaya guna dan fleksibel untuk melakukan analisis data statistik. Saya berharap seluruh insan statistik, khususnya di lingkungan Badan Pusat Statistik, mampu menguasai R. Buku ini menawarkan semua yang dibutuhkan bagi mereka yang ingin belajar tentang R dari dasar. Ditulis oleh penulis yang tidak hanya menguasai ilmu statistik tetapi juga praktisi di bidang statistik yang sudah lama berkecimpung di Badan Pusat Statistik, membuat buku ini mudah dipahami oleh pembacanya."
- Dr. Suhariyanto, Kepala Badan Pusat Statistik Republik Indonesia
Pengolahan dan analisis data tidak bisa dilepaskan dari dunia Penelitian. Terlebih lagi pada era BIG DATA sekarang ini. Banyak metodologi yang bisa digunakan untuk mengolah data sehingga bisa bermanfaat bagi dunia penelitian. Salah satu bahasa pemrograman yang banyak digunakan adalah Pemrograman R. Untuk bisa beradaptasi dengan cepat, dibutuhkan sebuah BEST PRACTICE untuk mempercepat proses belajar. Buku ini menawarkan sebuah solusi untuk pembaca berupa BEST PRACTICE yang mampu menjelaskan penggunaan R-LANGUAGE secara bertahap dan topat. Buku ini sangat cocok bahkan untuk seorang pemula sekalipun. Buku ini memberikan penjelasan yang komprehensif seputar statistik baik teori maupun aplikasinya pada Pemrograman R.
– Prof. Dr. Eng. Wisnu Jatmiko, Menajer Riset Fakultas Ilmu Komputer Universitas Indonesia
Sebagai layaknya bidang ilmu yang berbasis kepada matematika, statistika tak dapat dipelajari secara menyeluruh dengan membaca dan memahami saja. Setelah memahami suatu konsep perlu praktek dan pengalaman langsung dalam membangun hipotesis, menarik sampling, meriksod dan mengolah data, modeling, melakukan test statistik untuk menguji hipotesis, membangun confidona interval, simulasi dan berbagai hal terkait 'praktek' statistika.
Prosak-prosak yang disebutkan di atas terfasilitasi dan tertolong dengan perkembangan teknologi komputasi dalam beberapa dekade belakangan ini. Buku ini pengalar statistika pada tingkat pemula hingga intermediate, ditulis dengan baik menggunakan R, sebagai fasilitas penolong komputasi. Keunggulan R sebagai 'statistical programming language' ditopang dengan ketersediaan berbagai 'package' statistika yang siap pakai/tersedia dan kemudahan dalam pengolahan data, grafik, dan juga programming serta sifatnya yang open source dan tak berbayar, menjadikan buku ini mempunyai nilai tambah yang baik.
- Bakhtiar Hasan, Ph.D., Senior Biostatistician European Organization for Research and Treatment of Cancer, Belgium
Adapun yang dibahas dalam buku ini, sebagai berikut :
• Pengenalan RGUI dan RStudio
• Statistik Deskriptif dan Visualisasi Data
• Visualisasi dengan ggplot2
• Statistik Inferensia dan Uji Hipotesis
• Analisis Keragaman (Anova)
• Analisis Regresi dan Korelasi
• Pemrograman dengan R dan dplyr
• Pengenalan RCommander
• Analisa Data dengan Rcommander
• Regresi Logistik
İN MEDIA
ISBN : 978-602-0946-98-6
DASAR-DASAR STATISTIKA Dengan Software R Konsep dan Aplikasi
İN MEDIA
Setia Pramana, Ph.D
Ricky Yordani, M.Stat
Robert Kurniawan, M.Si
Budi Yuniarto, M.Si
DASAR-DASAR
STATISTIKA
Dengan Software R
Konsep dan Aplikasi
EDISI KEDUA
R
R Studio
İN MEDIA

DATA MINING dengan R

Konsep Serta Implementasi

Kebutuhan akan eksplorasi dan analisis data semakin meningkat beberapa tahun terakhir. Metode eskplorasi dan analisis data juga mulai bergeser ke arah penggunaan data mining dan beberapa algoritma machine learning. Hal ini mendorong perubahan kurikulum dan materi yang harus disampaikan dan dikuasai mahasiswa khususnya mahasiswa jurusan statistik. Buku ini sangat saya rekomendasikan baik kepada mahasiswa maupun para pengajar karena buku ini tidak hanya memberikan teori namun juga mengajarkan bagaimana mengaplikasikan teori tersebut dalam contoh-contoh praktis. Buku ini juga memberikan keberagaman aplikasi dari data mining dengan tipe data yang berbeda-beda yang dapat diaplikasikan dengan software R.

Dr. Erni Tri Astuti, M.Math - Direktur Politeknik Statistika STIS

R merupakan salah satu alat pengolahan data yang sangat ampuh. Dengan bahasa yang lugas dan "to-the-point", penulis berhasil menyajikan data mining dengan pendekatan praktis menggunakan R. Buku ini merupakan batu pijakan yang sangat berguna buat para aspiring data scientist yang ingin menggeluti bidang data science

Syafri Bahar S.Si., M.Sc., FRM - Vice President of Data Science GOJEK.

Bahasan buku ini mencakup:

1. Pengantar Data Mining
2. Eksplorasi dan Visualisasi Data
3. Regresi Linear dan Logistik
4. Analisis Komponen Utama
5. Multivariate Anova
6. Supervised Learning (KNN, Decision Tree, Random Forest, dll)
7. Unsupervised Learning (Cluster Analysis)
8. Text Mining
9. Analisis Sentimen
10. Data Mining dalam Bioinformatika

İN MEDIA

UMUM
ISBN 978-602-6469-

Harga P. Jawa Rp.

Setia Pramana, dkk

DATA MINING dengan R

İN MEDIA

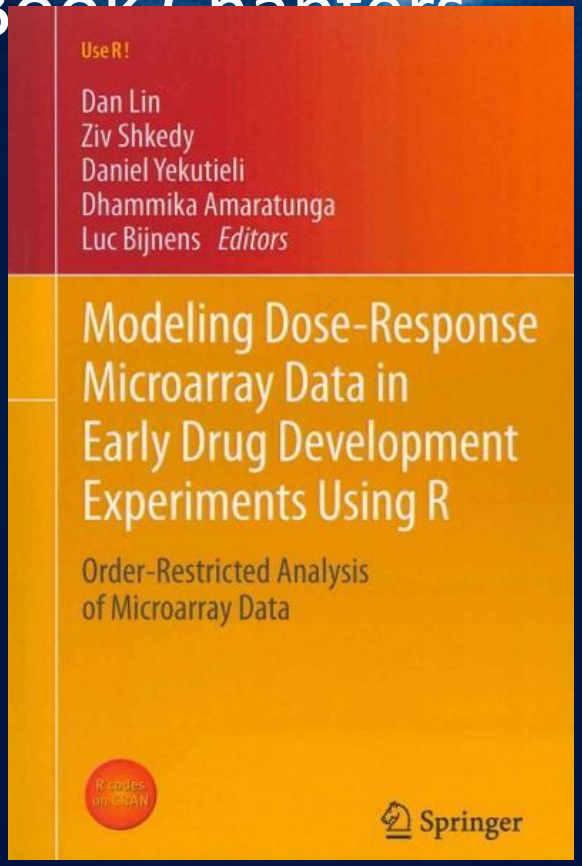
Setia Pramana
Budi Yuniarto
Siti Mariyah
Ibnu Santoso
Rani Nooraeni

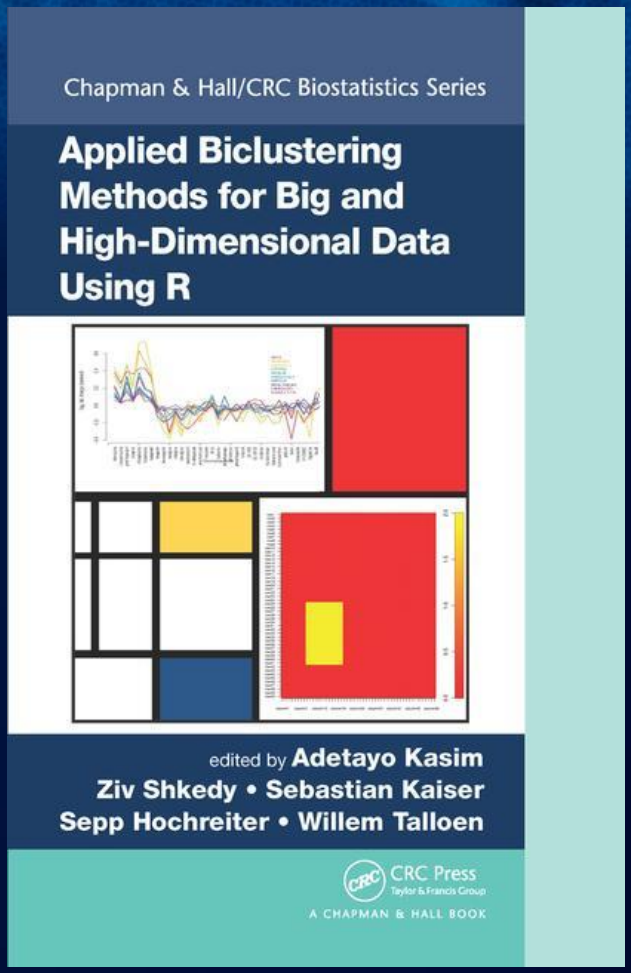
DATA MINING dengan R

Konsep Serta Implementasi

İN MEDIA

Book Chapters





Outline

- Intro Bioinformatics
- Bioinformatics pipeline
- R packages for Bioinformatics

What is Bioinformatics?

What is Bioinformatics ?

Bioinformatics - a definition[1]

(Molecular) **bio** – informatics: bioinformatics is conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying "***informatics techniques***" (derived from disciplines such as applied maths, computer science and statistics) to ***understand*** and ***organise*** the ***information*** associated with these molecules, on a ***large scale***. In short, bioinformatics is a management information system for molecular biology and has many ***practical applications***.

[1] As submitted to the Oxford English Dictionary

What is Bioinformatics ?

Bioinformatics is the use of computers for the acquisition, management, and analysis of biological information.

It incorporates elements of molecular biology, computational biology, database computing, and the Internet…

… bioinformatics is clearly a multi-disciplinary field including:
computer systems management
networking, database design, computer programming,
molecular biology

From Using Computers for Molecular Biology, Stuart M. Brown, PhD, RCR, NYU Medical Center

What is Bioinformatics ?

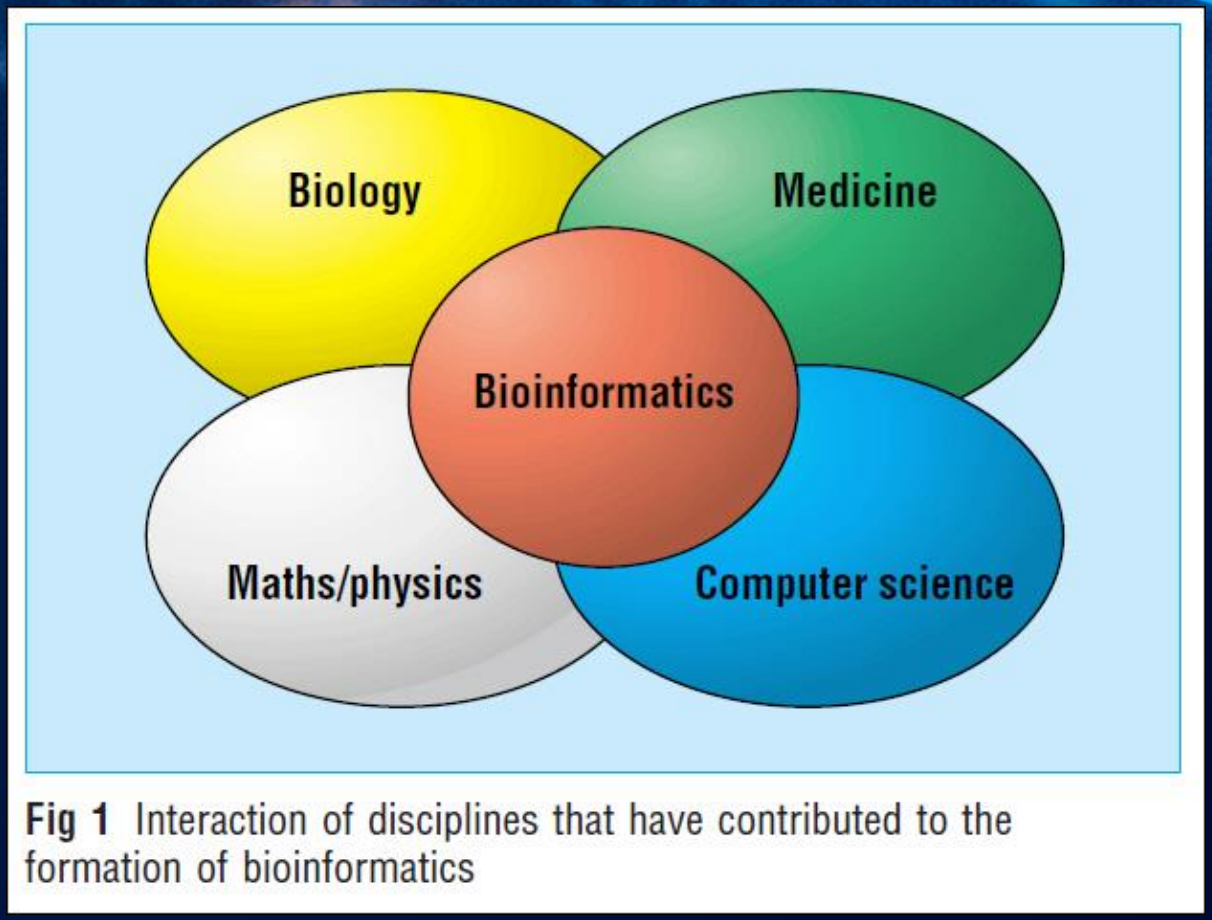


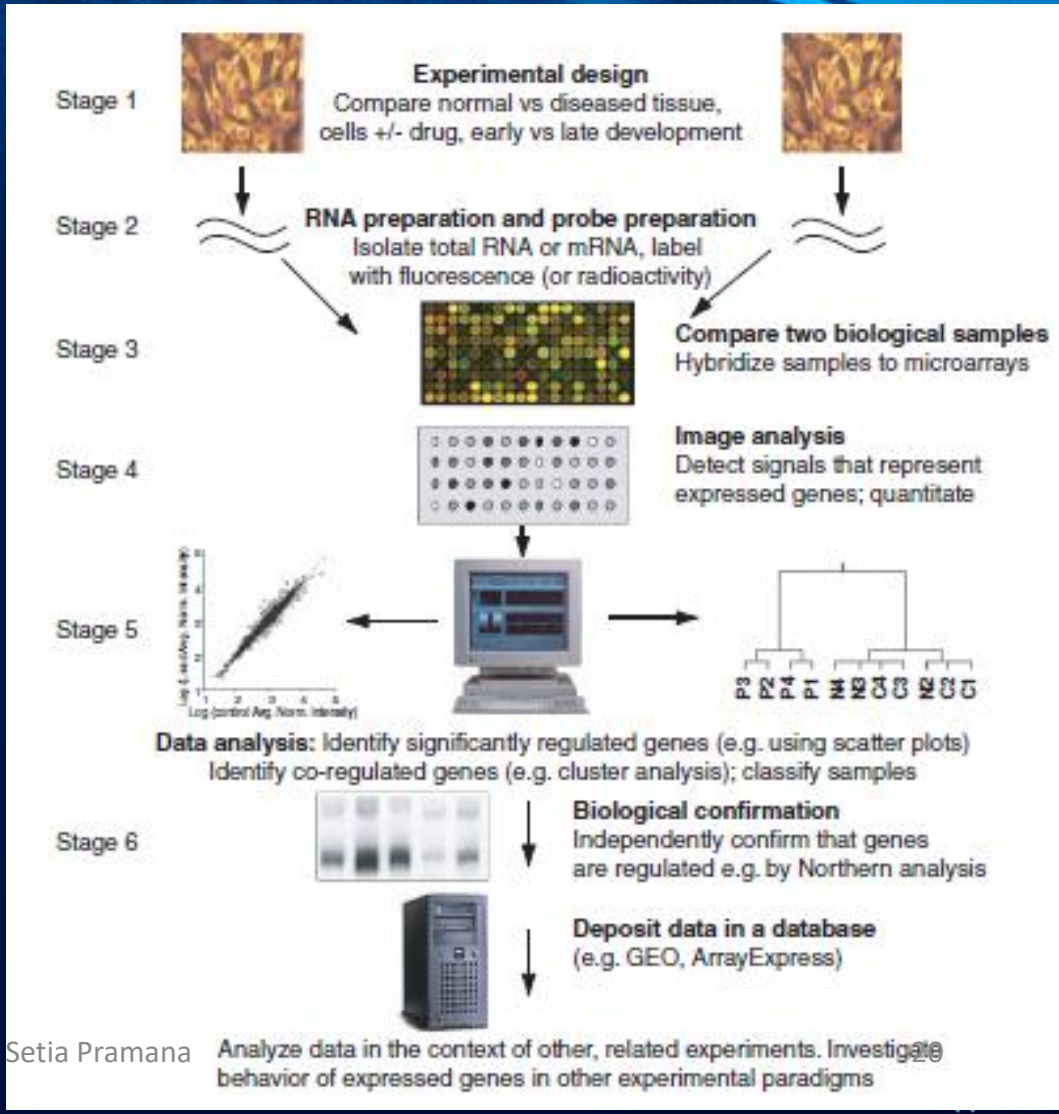
Fig 1 Interaction of disciplines that have contributed to the formation of bioinformatics

Bioinformatics is a multifaceted discipline combining many scientific fields including computational biology, statistics, mathematics, molecular biology and genetics (Fenstermacher, 2005, p. 440).

…from Bayat (2002), p 1018.

Microarray

Overview of the process of generating high throughput gene expression data using microarrays.



Pipeline

- Experiment design → Lab work → Image processing
- Signal summarization (RMA, GCRMA)
- Normalization
- Data Analysis:
  - Differentially Expressed genes
  - Clustering
  - Classification
  - Etc.

- Network / Pathways (GSEA etc..)
- Biological interpretations

Preprocessed data

Genes	C1	C2	C3	T1	T2	T3
G8521	6.89	7.18	6.60	7.40	7.15	7.40
G8522	6.78	6.55	6.37	6.89	6.78	6.92
G8523	6.52	6.61	6.72	6.51	6.59	6.46
G8524	5.67	5.69	5.88	7.43	7.16	7.31
G8525	5.64	5.91	5.61	7.41	7.49	7.41
G8526	4.63	4.85	5.72	5.71	5.47	5.79
G8527	8.28	7.88	7.84	8.12	7.99	7.97
G8528	7.81	7.58	7.24	7.79	7.38	8.60
G8529	4.26	4.20	4.82	3.11	4.94	3.08
G8530	7.36	7.45	7.31	7.46	7.53	7.35
G8531	5.30	5.36	5.70	5.41	5.73	5.77
G8532	5.84	5.48	5.93	5.84	5.73	5.75

Microarray Data Analysis Types

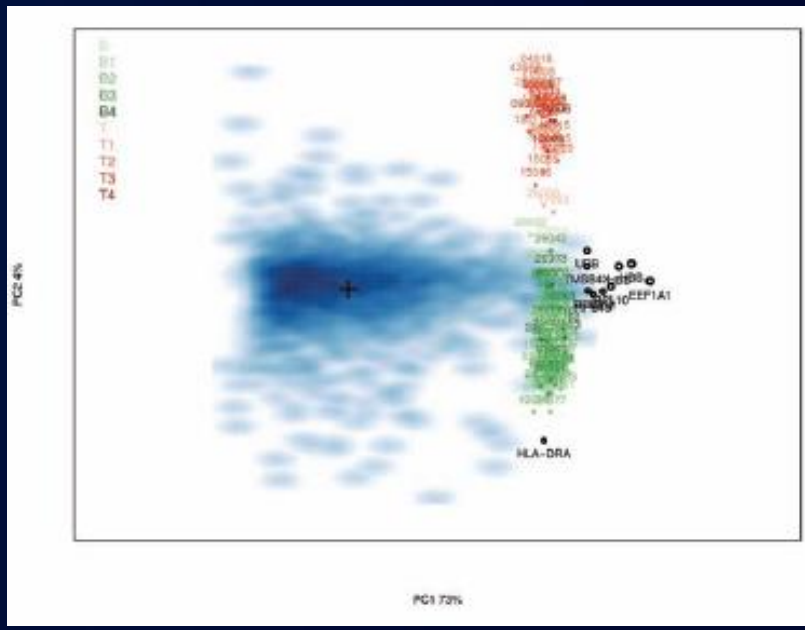
- Gene Selection
  - find genes for therapeutic targets
- Classification (Supervised)
  - identify disease (biomarker study)
  - predict outcome / select best treatment
- Clustering (Unsupervised)
  - find new biological classes / refine existing ones
  - Understanding regulatory relationship/pathway
  - exploration

Gene Selection

- Modified t-test
- Significance Analysis of Microarray (SAM)
- Limma (Linear model for microarrays)
- Random forest
- Lasso (least absolute selection and shrinkage operator)
- Linear Mixed model
- Elastic-net
- Etc,

Visualization

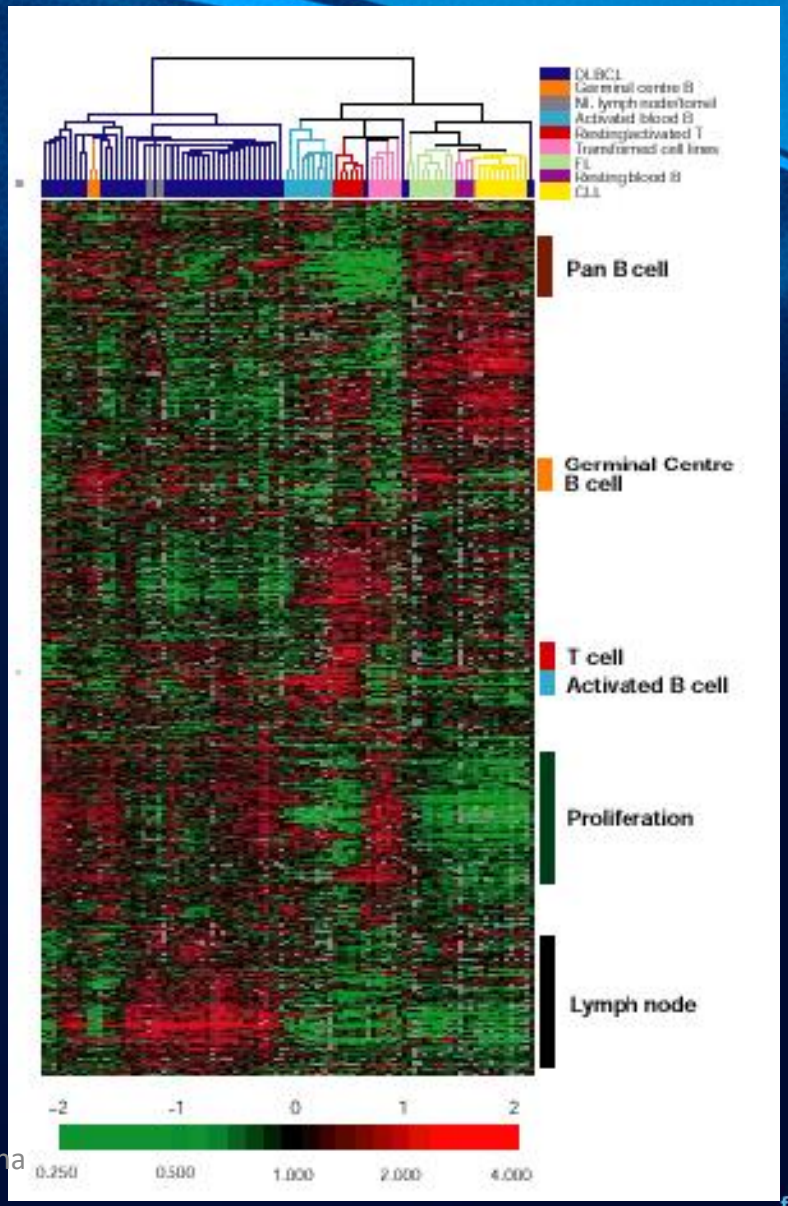
- Dimensionality reduction
- PCA (Principal Component Analysis)
- Biplot
- Multi dimensional scaling
- Etc



Clustering

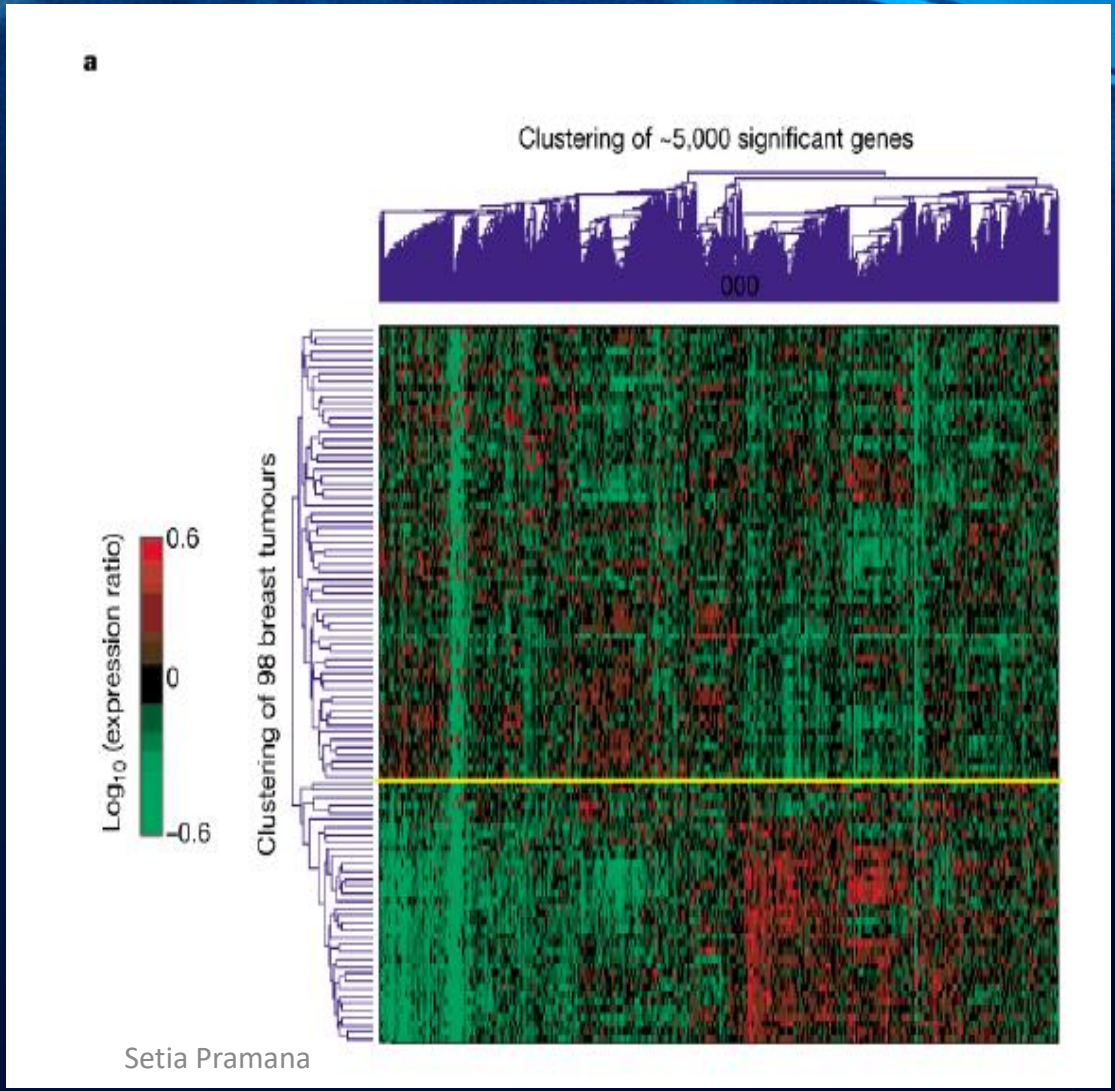
- Cluster the genes
- Cluster the arrays/conditions
- Cluster both simultaneously

- K-means
- Hierarchical
- Biclustering algorithms



Clustering

- Cluster or Classify genes according to tumors
- Cluster tumors according to genes

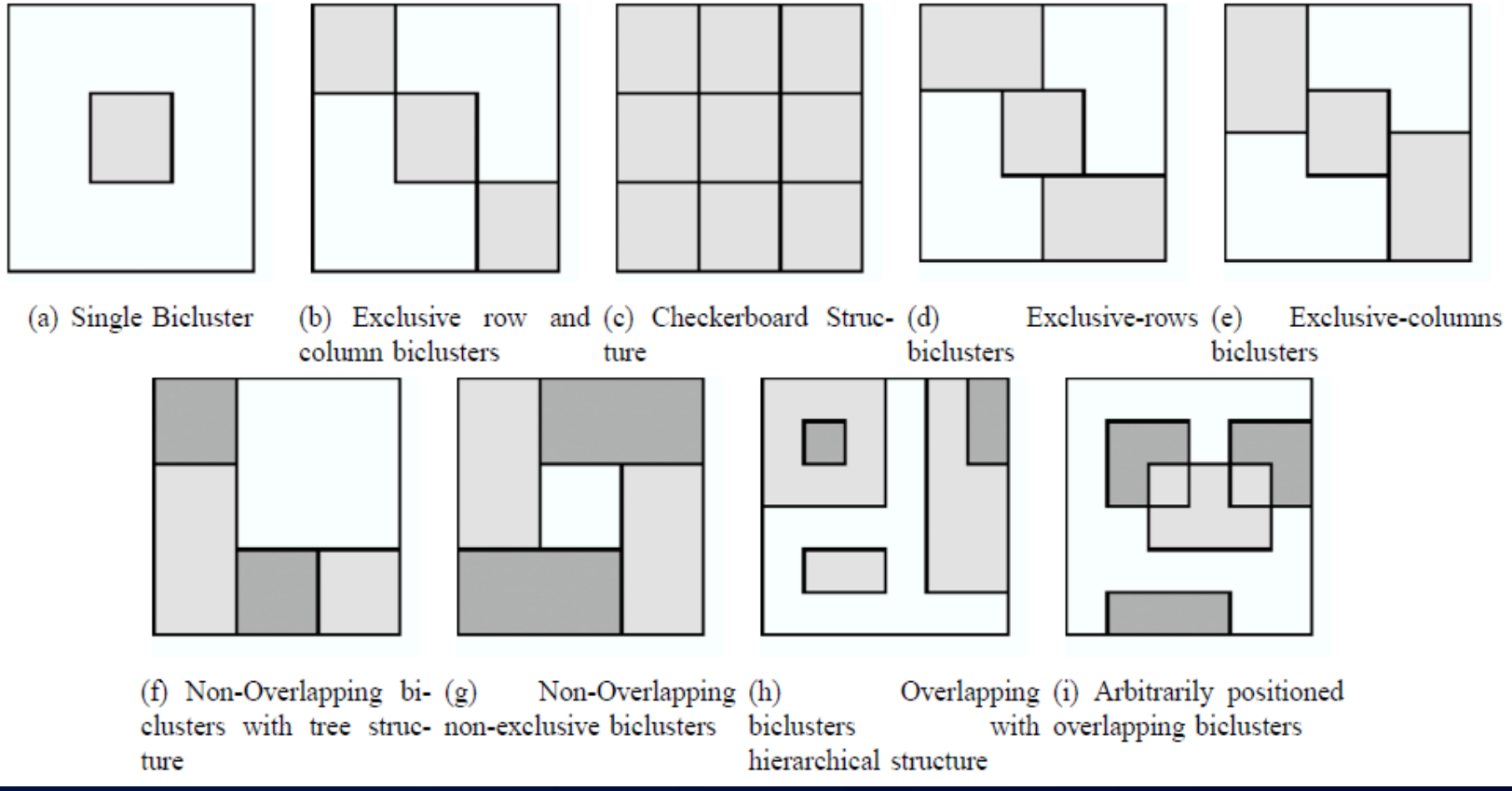


Setia Pramana

Biclustering

- A biclustering method is an unsupervised learning method which looks for sub-matrices in a data matrix with a high similarity of elements.
- Algorithms: Statistical based, AI, machine learning.
- BiclustGUI: A User Friendly Interface for Biclustering Analysis

Bicluster Structure



(a) Single Bicluster (b) Exclusive row and column biclusters (c) Checkerboard Structure (d) Exclusive-rows biclusters (e) Exclusive-columns biclusters

(f) Non-Overlapping biclusters with tree structure (g) Non-Overlapping non-exclusive biclusters (h) Overlapping biclusters with hierarchical structure (i) Arbitrarily positioned with overlapping biclusters

OPEN ACCESS Freely available online

PLOS | PATHOGENS

Affinity Proteomics Reveals Elevated Muscle Proteins in Plasma of Children with Cerebral Malaria

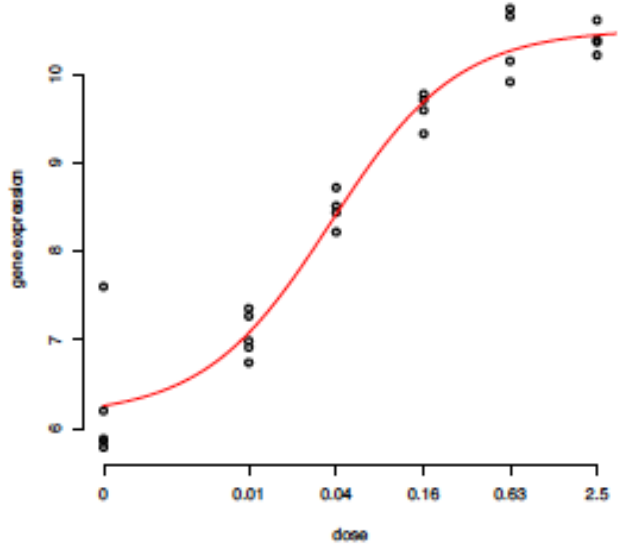
Julie Bachmann[1☯], Florence Burté[2☯], Setia Pramana[3], Ianina Conte[2], Biobele J. Brown[4,5,6], Adebola E. Orimadegun[4], Wasiu A. Ajetunmobi[4], Nathaniel K. Afolabi[4], Francis Akinkunmi[4], Samuel Omokhodion[4,6], Felix O. Akinbami[4,6], Wuraola A. Shokunbi[5,6], Caroline Kampf[7], Yudi Pawitan[3], Mathias Uhlén[1], Olugbemiro Sodeinde[2,4,5,6], Jochen M. Schwenk[1], Mats Wahlgren[8*¶], Delmiro Fernandez-Reyes[2,4,5,6,9*¶], Peter Nilsson[1*¶]

1 SciLifeLab Stockholm, School of Biotechnology, KTH-Royal Institute of Technology, Stockholm, Sweden, **2** Division of Parasitology, Medical Research Council National Institute for Medical Research, London, United Kingdom, **3** Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, **4** Department of Paediatrics, College of Medicine, University of Ibadan, University College Hospital, Ibadan, Nigeria, **5** Department of Haematology, College of Medicine, University of Ibadan, University College Hospital, Ibadan, Nigeria, **6** Childhood Malaria Research Group, University College Hospital, Ibadan, Nigeria, **7** Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden, **8** Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden, **9** Brighton & Sussex Medical School, Sussex University, Brighton, United Kingdom

Aim: To improve understanding of host protein profiles during disease progression especially in children.

Dose-response Microarray Studies

Monitoring of gene expression with respect to increasing dose of a compound.



No prior info about the dose-response shape.

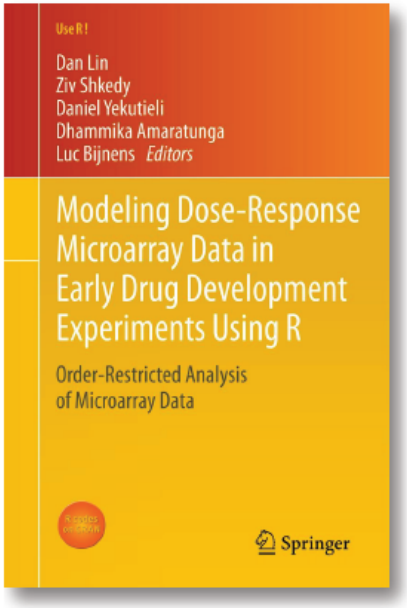
Genes have different shapes.

Many ”noisy” genes hence need for initial filtering.

Dose-response Microarray Studies

Springer

springer.com



2012, XV, 282 p. 96 illus., 4 illus. in color.

D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga, L. Bijnens (Eds.)

Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R

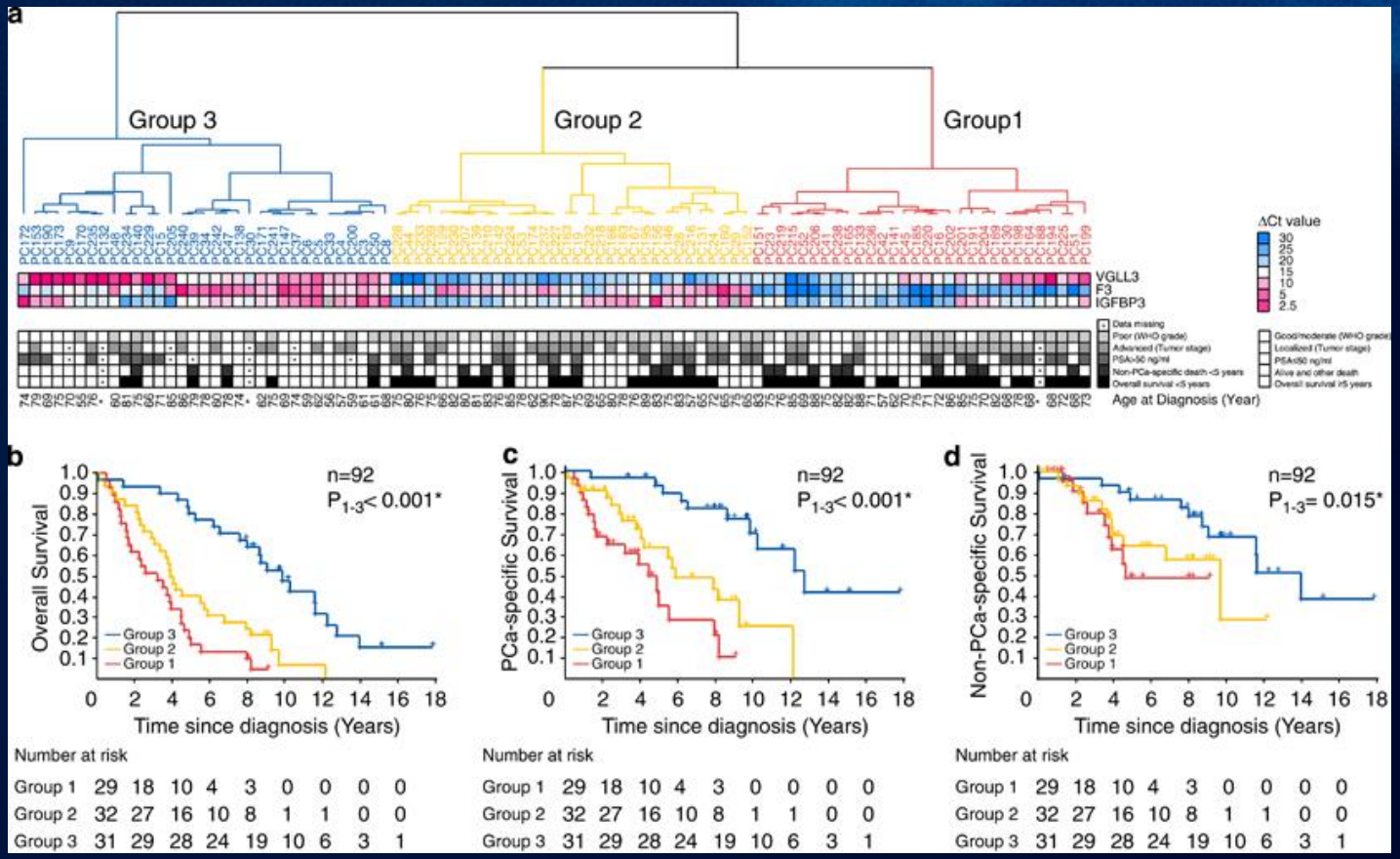
Order-Restricted Analysis of Microarray Data

Series: Use R!

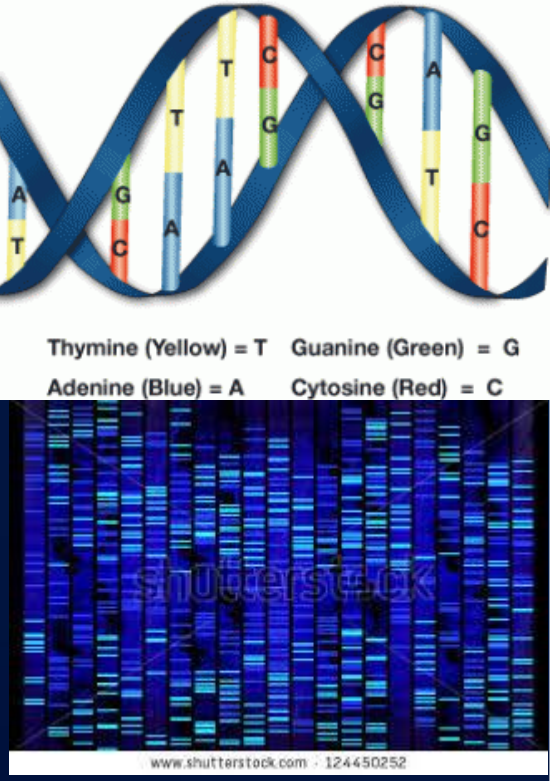
- This book focuses on the analysis of microarray data in the dose-response setting in early drug development experiments in the pharmaceutical industry
- Part I discusses the dose-response setting and the problem of estimation of normal means under order restrictions
- Part II demonstrates the use of the IsoGene R library and in particular its graphical capacity

This book focuses on the analysis of dose-response microarray data in pharmaceutical setting, the goal being to cover this important topic for early drug development and to provide user-friendly R packages that can be used to analyze dose-response microarray data. It is intended for biostatisticians and bioinformaticians in the pharmaceutical industry, biologists, and biostatistics/bioinformatics graduate students.

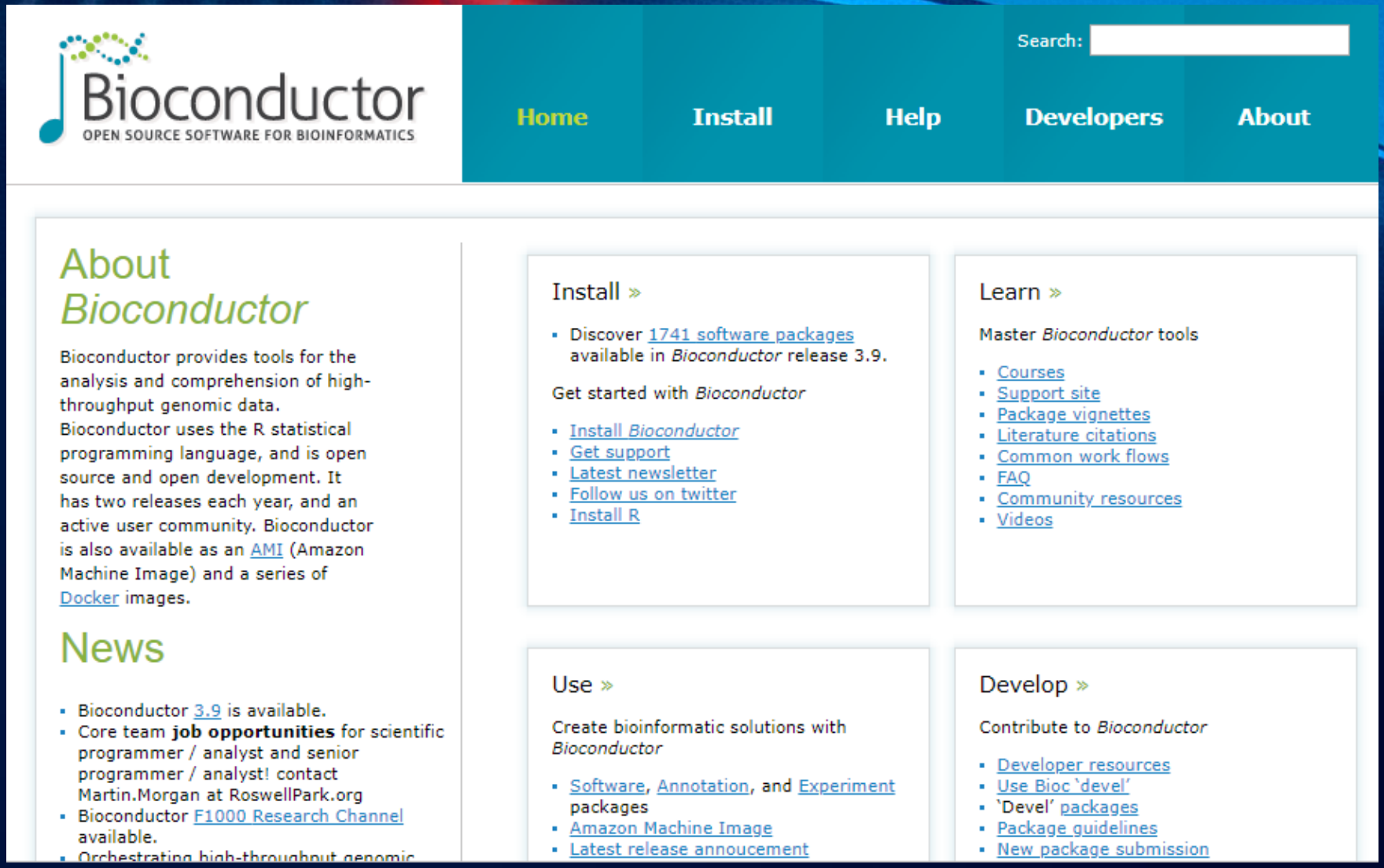
Gene Signature for Prostate Cancer



Bioinformatics R Packages: Bioconductor



- Microarray analysis: expression, copy number, SNPs, methylation, ...
- Sequencing: RNA-seq, ChIP-seq, called variants, ...
  - Especially *after* assembly / alignment
- Annotation: genes, pathways, gene models (exons, transcripts, etc.), ...
- Epigenetics
- Gene set enrichment analysis
- Network analysis
- Flow cytometry
- Proteomics and metabolomics
- Cheminformatics
- Images and high-content screens


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
About Bioconductor
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.
News
Bioconductor 3.9 is available.
Core team job opportunities for scientific programmer / analyst and senior programmer / analyst! contact Martin.Morgan at RoswellPark.org
Bioconductor F1000 Research Channel available.
Orchestrating high-throughput genomic
Install »
Discover 1741 software packages available in Bioconductor release 3.9.
Get started with Bioconductor
Install Bioconductor
Get support
Latest newsletter
Follow us on twitter
Install R
Learn »
Master Bioconductor tools
Courses
Support site
Package vignettes
Literature citations
Common work flows
FAQ
Community resources
Videos
Use »
Create bioinformatic solutions with Bioconductor
Software, Annotation, and Experiment packages
Amazon Machine Image
Latest release annoucement
Develop »
Contribute to Bioconductor
Developer resources
Use Bioc 'devel'
'Devel' packages
Package guidelines
New package submission

Great link to start

- https://www.bioconductor.org/help/course-materials/
- http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual
- http://www.cs.ukzn.ac.za/~hughm/bio/docs/a-little-book-of-r-for-bioinformatics.pdf

Bioconductor

Release coincides with R release.

Current: Bioconductor 2.10
(release coincide with R 2.15)

To install use script on Bioconductor Website

```
source("http://www.bioconductor.org/biocLite.R")
biocLite()
```

Packages Overview

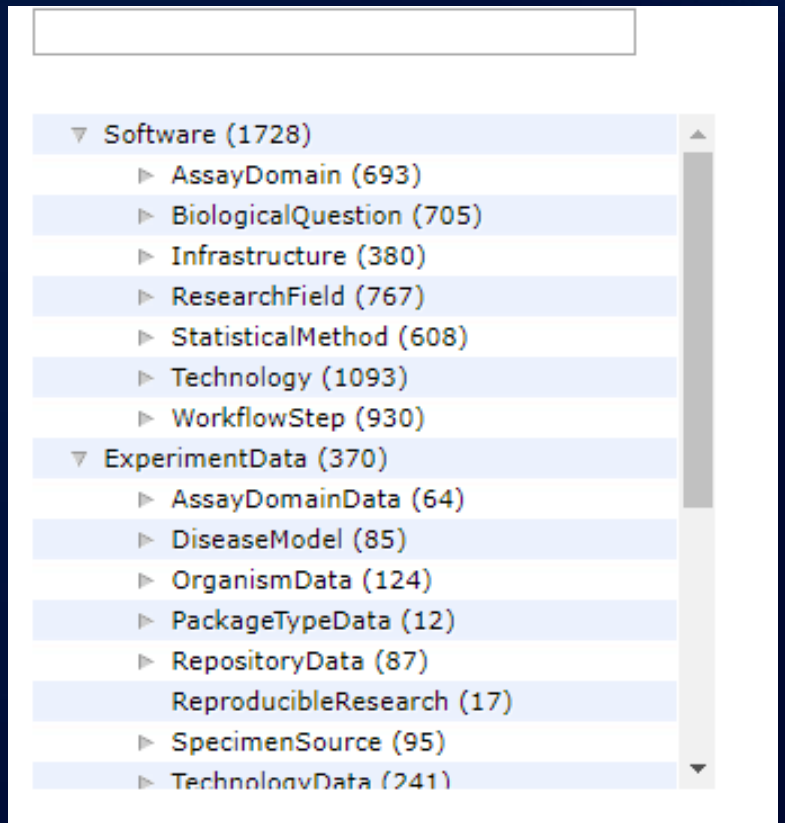
BioConductor web site

- Bioconductor BiocViews Task view

  Software

  Annotation Data

  Experimental Data



Main types of Annotation Packages

- Gene centric AnnotationDbi packages:
  - Organism: org.Mm.eg.db.
  - Technology/Platform: hgu133plus2.db.
  - GeneSets and Pathway (biology level): GO.db or KEGG.db
  - .db packages can be queried with sql or accessed using annotation package (totable, get, mget)
- Genome centric GenomicFeatures packages:
  - Transriptome level: TxDb.Hsapiens.UCSC.hg19.knownGene
  - Generic features: Can generate via GenomicFeatures
- biomaRt:
  - Query web-based `biomart' resource for genes, sequence, SNPs, and etc.
- See http://www.bioconductor.org/help/course-materials/2011/BioC2011/LabStuff/AnnotationSlidesBioc2011.pdf

onductor resources

- Mailing List (sign up for daily digest)
- Documentation, workshop/course material online
  - Slides from talks, pdf of tutorials, R code
- Help available for each software package
  - Each package MUST contain vignette (howto)
- Other resources ww.Rseek.org www.r-bloggers.com

Vignette

- Tutorials, provide worked example of package
- Required in Bioconductor packages
- Written in Sweave (Leisch, 2002).
  - $\LaTeX$ dynamic reports in which R code is embedded and executable
  - All R code in vignette is checked (and executed) by R CMD check
  - http://www.bioconductor.org/docs/vignettes.html

```
library("Biobase")
library("GOstats")    # Load package of interest
openVignette()
```

Some common data types

- Microarray
- SNP
- NGS

Reading Affymetrix Data

```
library(affy)
require(affy)   # Alternative

affybatch <- ReadAffy(celfile.path="[Location of your
    data]")

eSet<-justRMA()
```

Sample Workflow

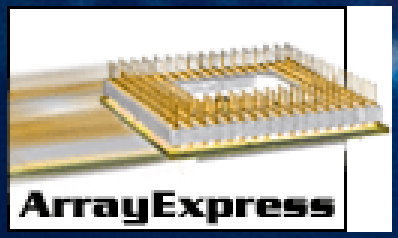
The following psuedo-code illustrates a typical R / Bioconductor session. It uses RMA from the affy package to pre-process Affymetrix arrays, and the limma package for assessing differential expression.

```
## Load packages
> library(affy)    # Affymetrix pre-processing
> library(limma)   # two-color pre-processing; differential
                   # expression

## import "phenotype" data, describing the experimental design
> phenoData <- read.AnnotatedDataFrame("sample-description.csv")

## RMA normalization
> eset <- justRMA("/celfile-directory", phenoData=phenoData)

## differential expression
> design <-                    # describe model to be fit
      model.matrix(~ Disease, pData(eset))
> fit <- lmFit(eset, design)  # fit each probeset to model
> efit <- eBayes(fit)        # empirical Bayes adjustment
> topTable(efit, coef=2)      # table of differentially expressed probesets
```

Other Arrays

- **Illumina**
  - Lumi package
- 2 color spotted arrays
  - Limma package
- Other arrays
  - http://www.bioconductor.org/help/workflows/oligo-arrays/

Public Microarray Data



ArrayExpress

- 21997 Studies (622,617 profiles,)

GEO Gene Expression Omnibus

GEO

- 22,735 Studies (558,074 profiles)

SMD

CIBEX

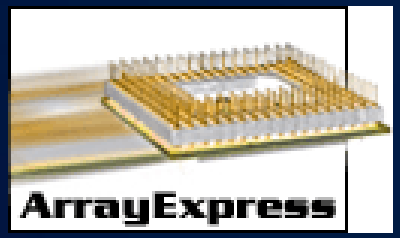
R Code

Gene Expression Omnibus

```
> library(GEOquery)
```

This loads the GEOquery library.

```
> gds <- getGEO("GDS1")
```




```
> library("ArrayExpress")
> sets = queryAE(keywords = "pneumonia", species = "homo
```

```
> mexp1422 = getAE("E-MEXP-1422", type = "full")
```

More on GEOquery

```
require(GEOquery)
```

Let's try to load the GDS810 dataset which contains data on Alzheimer's disease at various stages of severity.

```
GDS810<-getGEO("GDS810")
```

The *getGEO* function returns an object of class *GEOData*. You can get a description of this class like this:
help("GEOData-class")

```
Meta(GDS810)
Columns(GDS810)
head(Table(GDS810))
```

ExpressionSet Class in R

```
> library(ALL) # attach the ALL package to the search path
> data(ALL)    # load the ALL data into the global work space
> ALL          # view the ALL instance -- our first ExpressionSet!
```

`exprs(ALL)` returns the matrix of expression values (probe sets as rows, samples as columns). In the ALL data, the expression values are pre-processed and log-transformed.

`pData(ALL)` extracts a data frame describing the sample phenotype data.

`annotation(ALL)` reports the type of microarray chip used in this experiment.

DownStream Analysis

- Differentially expressed Genes
- Classification
- Clustering
- Pathway analysis
- Etc….

Thank you for your attention!!!

Setia.pramana@stis.ac.id