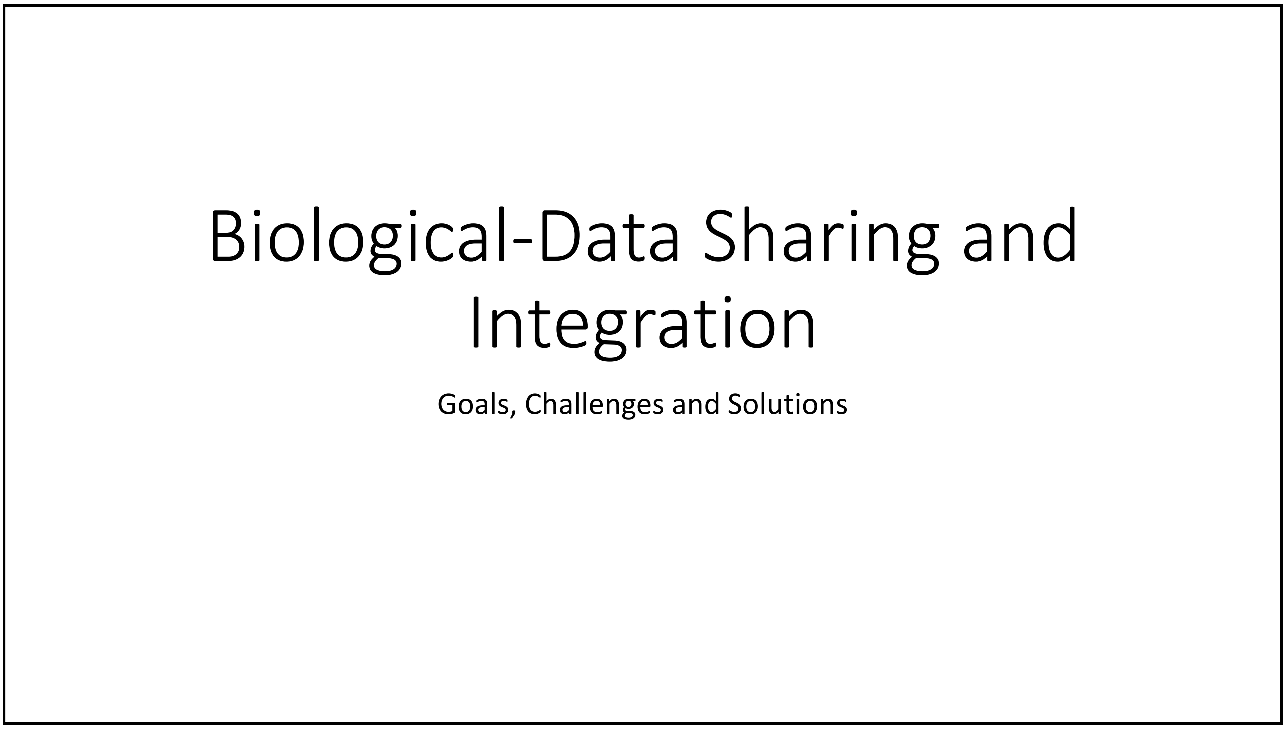
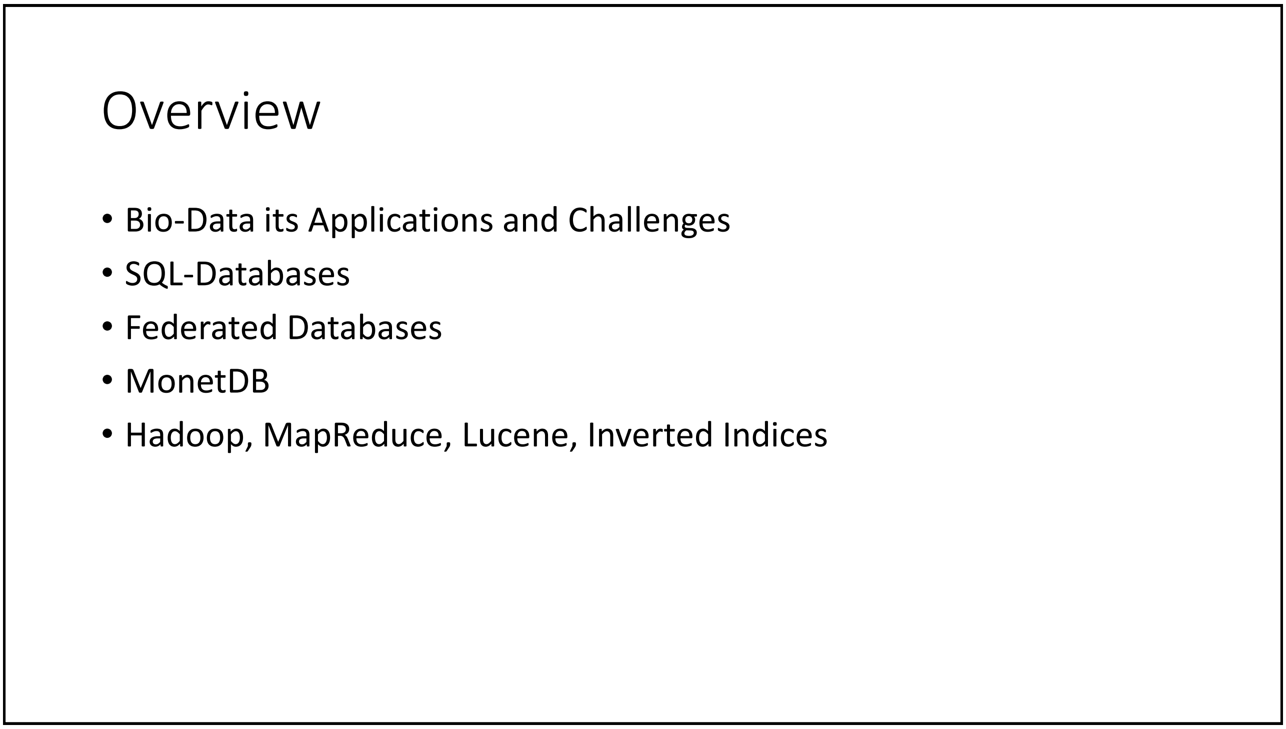


Biological-Data Sharing and Integration

Goals, Challenges and Solutions

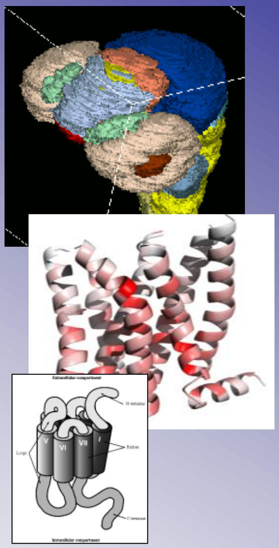


Overview

- Bio-Data its Applications and Challenges
- SQL-Databases
- Federated Databases
- MonetDB
- Hadoop, MapReduce, Lucene, Inverted Indices

Data Analyses and Data Modeling

- **Zebra Fish Atlas** (dr F. Verbeek)
- **Applied optimization techniques**: EA, GA, NN, etc. (prof T. Bäck)
- **Media Research: Content Based Indexing and Retrieval** (dr M.S. Lew)
- **Integrating Protein Databases:** Collecting and Analyzing Natural Variants in G Protein-Coupled Receptors (dr M.van Iterson,dr J. Kazius (LACDR))
- **Mining Phenotype Genotype Data** (dr F. Colas, LUMC)
- **Data Mining** (prof J. Kok)
  - VLe, sensor modeling, Hollandse brug, Cortana local pattern mining, Exception modeling, Complex pattern mining, …



Data Mining

Data Mining' and 'Knowledge Discovery in Databases' (KDD) are used interchangeably

- The process of **discovery of interesting, meaningful and actionable** patterns hidden in **large amounts** of data

- Multidisciplinary field originating from artificial intelligence, pattern recognition, statistics, machine learning, bioinformatics, econometrics

Data Mining in Bioinformatics

- Problem:
  - Leukemia (different types of Leukemia cells look very similar)
  - Given data for a number of samples (patients), can we
    - Accurately diagnose the disease?
    - Predict outcome for given treatment?
    - Recommend best treatment?
- Solutions (besides standard statistical analysis)
  - Data mining on micro-array data
  - Graph mining on co-expression networks

Centre for Medical Systems Biology

CMSB1

- **Epidemiology:** cohorts & genotyping
- **Systems Biology:** transcriptomics/arraying proteomics metabolomics
- **Technology:** magnetic resonance microscopy and others imaging molecular interactions
- **Model Systems:** animal models (mouse, zebra fish etc).
- **Clinical Applications:** translation (cells, vaccines, viral, pharmaceutical)
- **DIAL:** Data Integration, Analysis and Logistics

CMSB2

Alzheimer’s disease, Arthritis, Depression, Metabolic Syndrome, Migraine, Immunology, Social aspects of genomics

Phenotype Genotype Integration

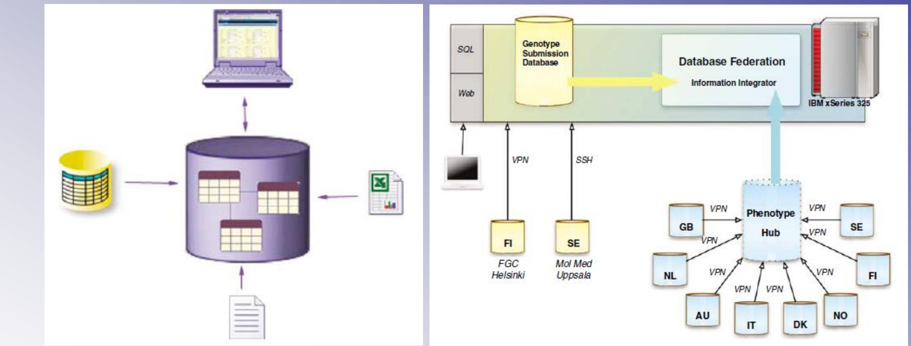
- Genotype data
  - Annotated genome databases
  - CGH-, SNP- databases
  - Expression databases
  - Etc.
- Phenotype data (Multimodal)
  - Blood samples
  - Weight, height, fat %, fat type, etc.
  - Echo, CT, MRI scans
  - Photographs
  - Etc.

DIAL Example Study Groups

- RotterdamStudy (ERGO: Hofman, van Duijn, a.o.)
  - population-based cohort study of 15,000 subjects aged 45+ years. Patients have been followed for over 20 years now.
- Grip Cohort Study (Rotterdam: van Duijn, Oostra)
  - population-based cohort study of 3 generation families (2500 subjects). They are screened for the presence of multiple diseases.
- Netherlands Twin Register (Boomsma VU Amsterdam)
  - number of twin-pairs ~87500, >18000 with DNA available

Nederlands Tweelingen Register
Vrije Universiteit
Nederlands Tweelingen Register

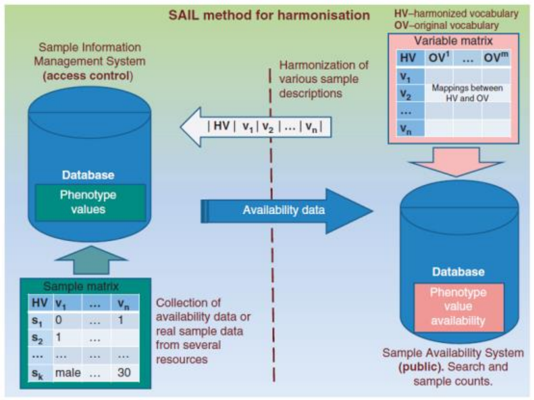
TWINE Database Architecture



A federated database implements an integrated transparent virtual view over several heterogeneous autonomous production databases.

From: Juha Muilu, Leena Peltonen and Jan-Eric Litton, *The federated database – a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Europe*, European Journal of Human Genetics (2007) 15, pp 718–723.

Data Harmonisation


SAIL method for harmonisation
HV–harmonized vocabulary
OV–original vocabulary
Sample Information Management System (access control)
Harmonization of various sample descriptions
| HV | v1 | v2 | ... | vn |
Database
Phenotype values
Availability data
Database
Phenotype value availability
Collection of availability data or real sample data from several resources
Sample Availability System (public). Search and sample counts.

Variable matrix

HV	OV^1	...	OV^m
v_1			
v_2	Mappings between HV and OV		
...			
v_n			

Sample matrix

HV	v_1	...	v_n
s_1	0	...	1
s_2	1	...	
...	...	...	...
s_k	male	...	30

From: O. Spjuth et al., Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. European Journal of Human Genetics EJHG Open, August 2015.

BBMRI-NL Rainbow project 12

D.I. Boomsma B. Penninx

Research on Major Depressive Disorder (MDD)

- Sample size should be huge as 13% of Dutch men and 24% of Dutch women have MDD at least once.
- High quality phenotypes
- DNA and GWAS (Genome wide association study) data

Solution:

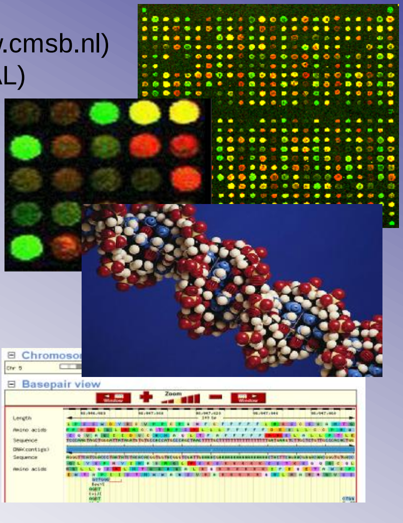
- Phenotype assessment tool for Biobanks

Recently (2015): A genome-wide significant SNP was found (3p14 in MAGI1 (rs35855737) related to Neuroticism. (Note: data harmonization was used.)

From: Meta-analysis of Genome-wide Association Studies for Neuroticism, and the Phylogenic Association With Major Depressive Disorder. Genetics Personality Consortium, 2015.

Data Integration, an example from the past: DIAL CMSB: CGH-DB

Center for Medical Systems Biology (www.cmsb.nl)
Data Integration and Logistics (DIAL)



CGH-DB a Microarray Database

- Consolidation of Experimental Data
- Integration of CGH/Oligo/SNP data with:
  - Other CGH/Oligo/SNP Experiments
  - Genome Databases
  - Expression Databases
  - Phenotype data
  - Etc.
- Publication, validation, repetition, etc.

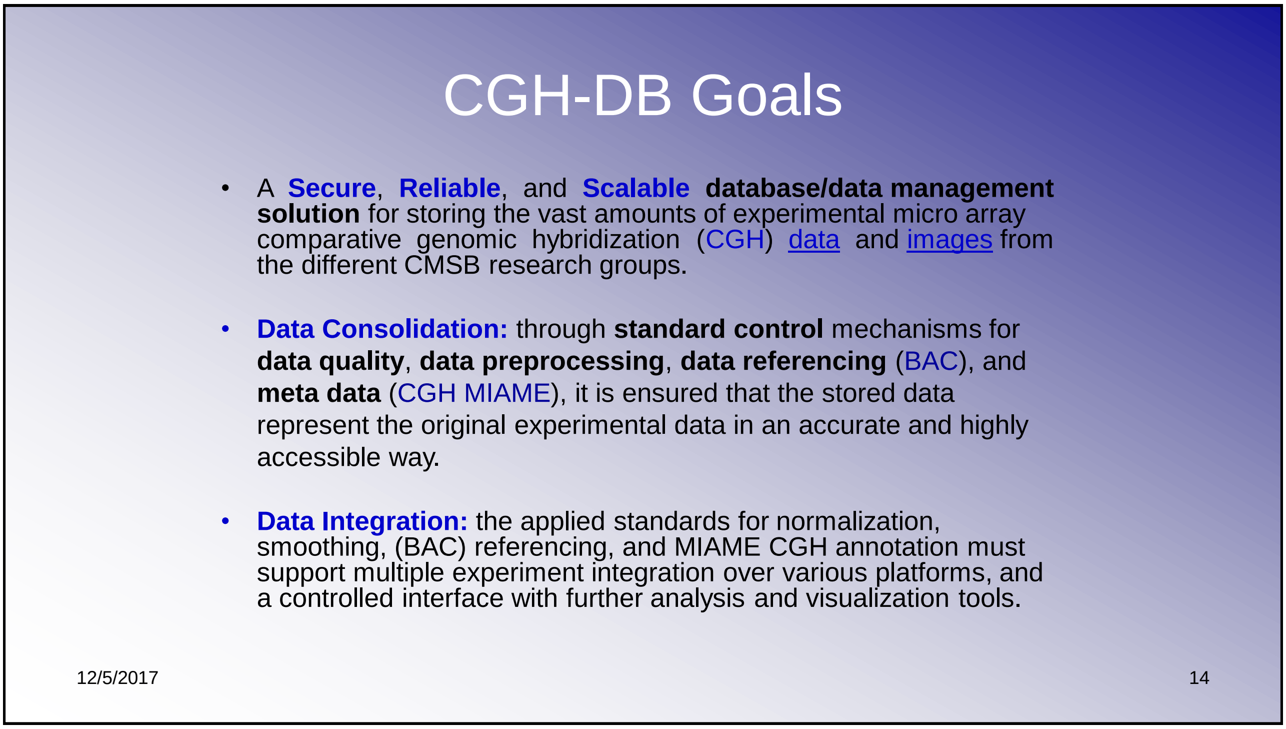
Centre for Medical Systems Biology

Groups Involved

- Micro Array Core Facility, VUMC: Bauke Ylstra, José Luis Costa, Anders Svensson, Paul vden IJssel, Mark van de Wiel, Sjoerd Vosse
- Center for Human and Clinical Genetics, LUMC: Judith Boer, Peter Taschner, and others
- Department of Molecular Cell Biology, Laboratory for Cytochemistry and Cytometry: Karoly Szuhai
- Leiden Institute of Advanced Computer Science, LIACS: Joost Kok, **Floris Sicking**, Erwin Bakker, Sven Groot, Michiel Ranshuysen, Harmen vder Spek, Antanas Kaziliūnas

Universiteit Leiden

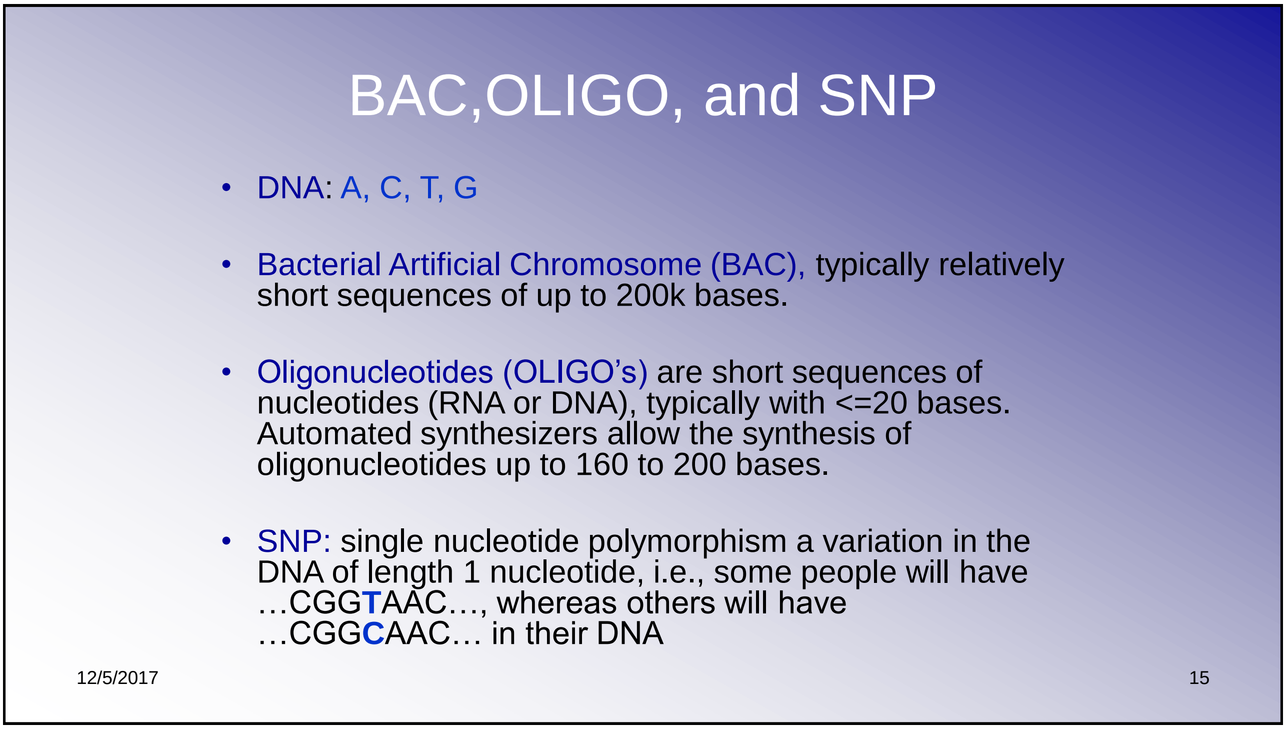
12/5/2017 13



CGH-DB Goals

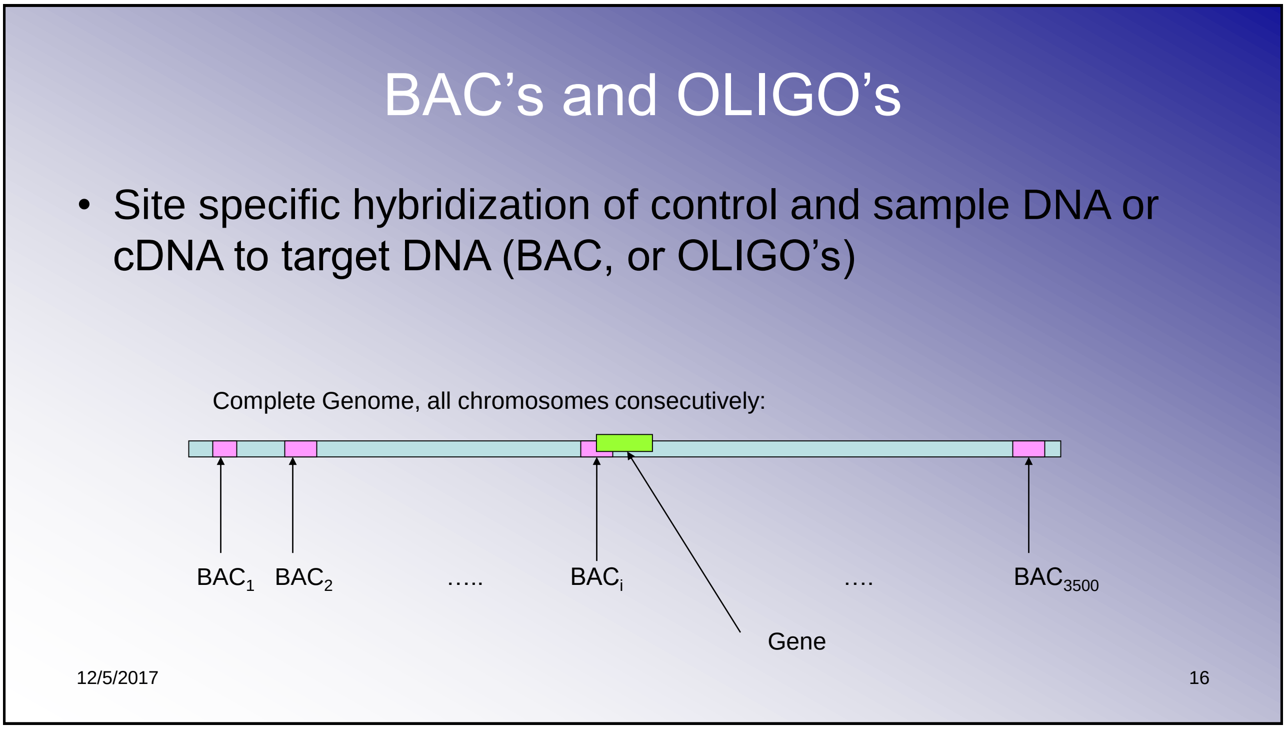
- A **Secure**, **Reliable**, and **Scalable** **database/data management solution** for storing the vast amounts of experimental micro array comparative genomic hybridization (CGH) data and images from the different CMSB research groups.
- **Data Consolidation:** through **standard control** mechanisms for **data quality**, **data preprocessing**, **data referencing** (BAC), and **meta data** (CGH MIAME), it is ensured that the stored data represent the original experimental data in an accurate and highly accessible way.
- **Data Integration:** the applied standards for normalization, smoothing, (BAC) referencing, and MIAME CGH annotation must support multiple experiment integration over various platforms, and a controlled interface with further analysis and visualization tools.

12/5/2017 14



BAC,OLIGO, and SNP

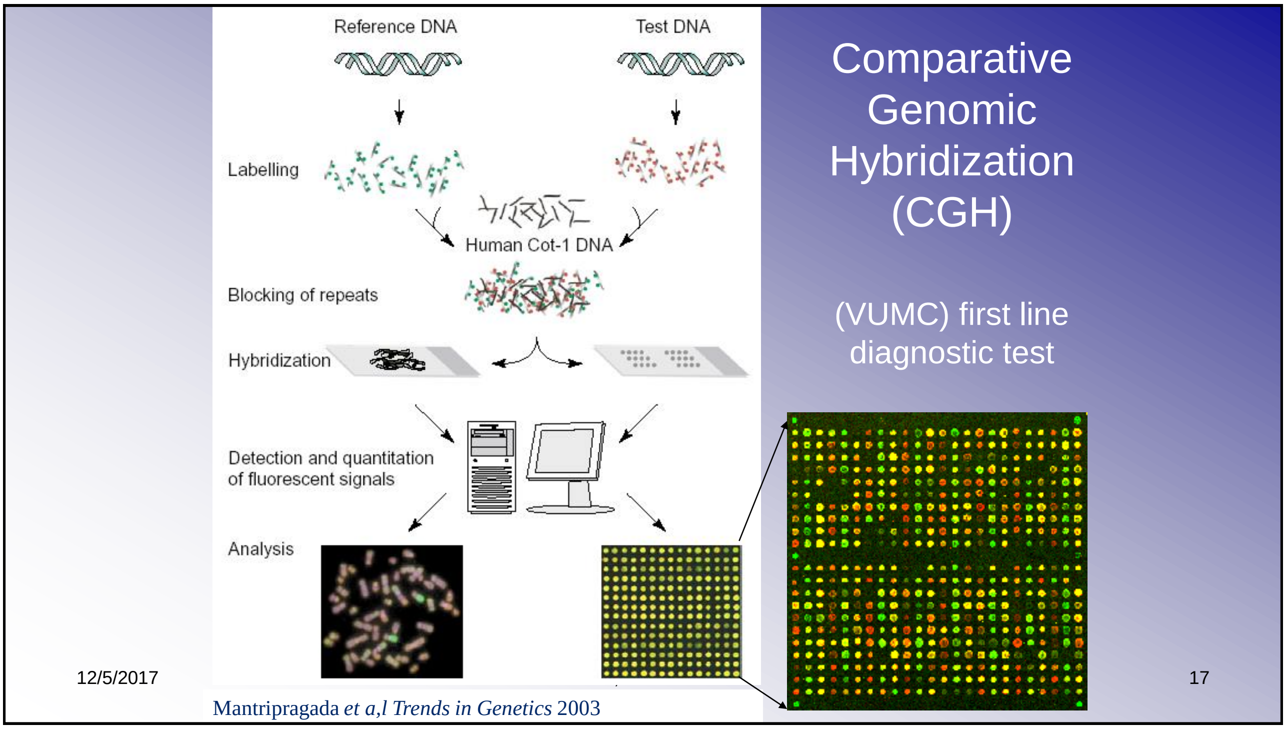
- DNA: A, C, T, G
- Bacterial Artificial Chromosome (BAC), typically relatively short sequences of up to 200k bases.
- Oligonucleotides (OLIGO's) are short sequences of nucleotides (RNA or DNA), typically with <=20 bases. Automated synthesizers allow the synthesis of oligonucleotides up to 160 to 200 bases.
- SNP: single nucleotide polymorphism a variation in the DNA of length 1 nucleotide, i.e., some people will have …CGG**T**AAC…, whereas others will have …CGG**C**AAC… in their DNA



BAC's and OLIGO's

- Site specific hybridization of control and sample DNA or cDNA to target DNA (BAC, or OLIGO's)

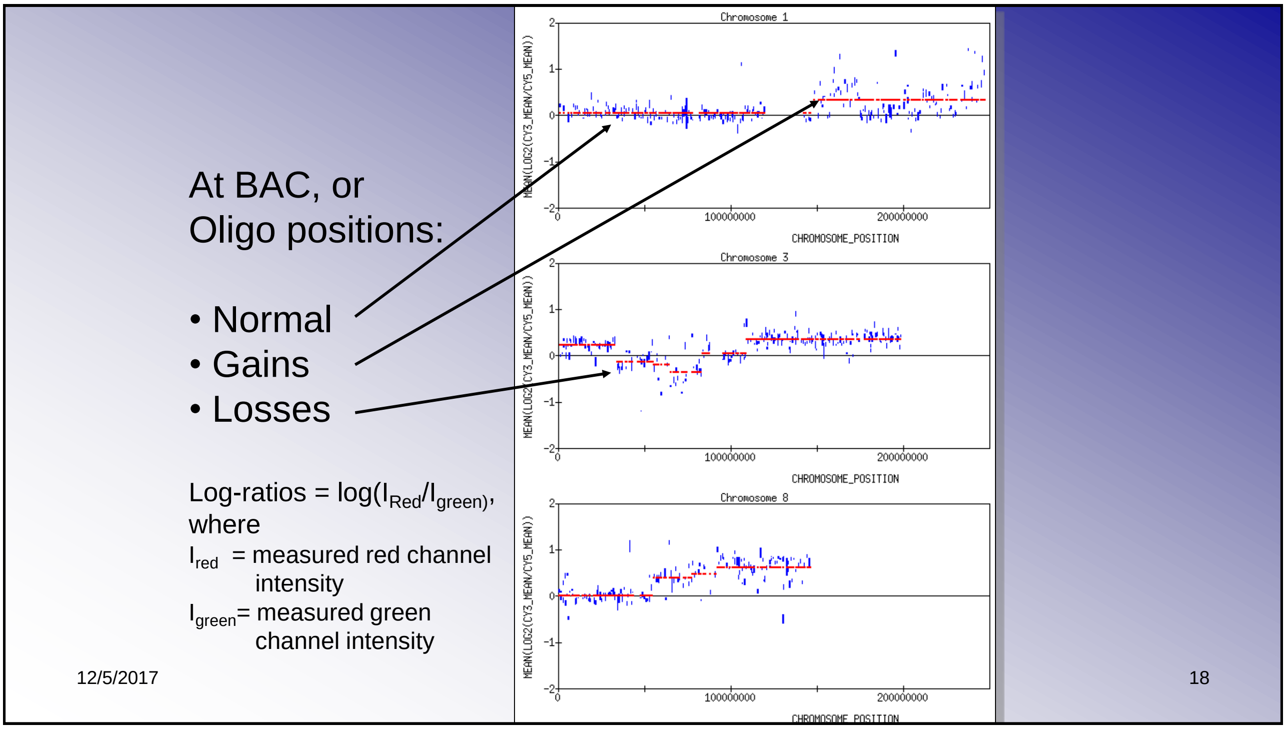
Complete Genome, all chromosomes consecutively:



Comparative Genomic Hybridization (CGH)

(VUMC) first line diagnostic test

Mantripragada *et a,l Trends in Genetics* 2003



At BAC, or Oligo positions:

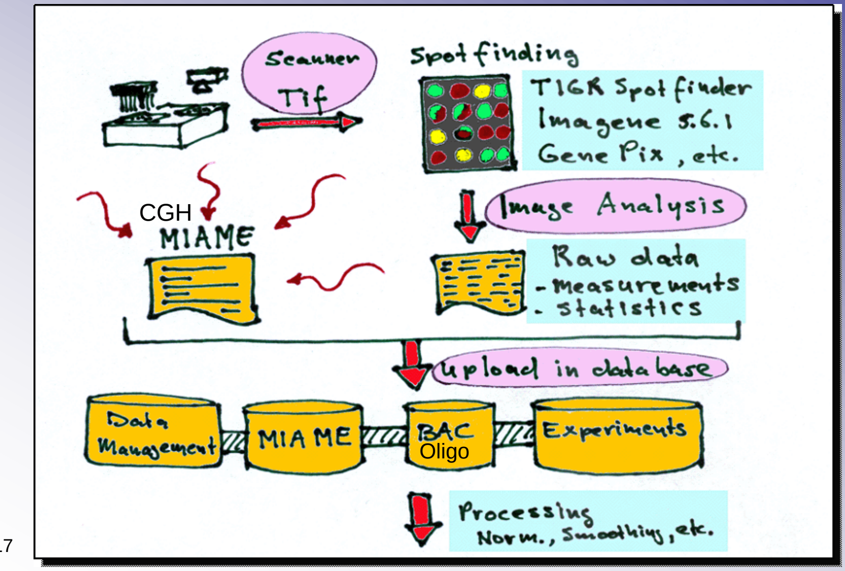
- Normal
- Gains
- Losses

Log-ratios = log(I_{Red}/I_{green}), where

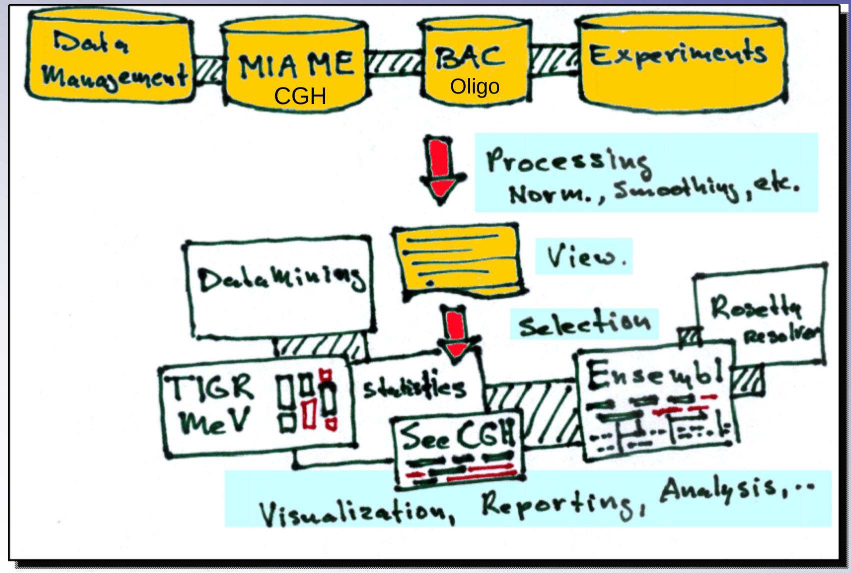
I_{red} = measured red channel intensity

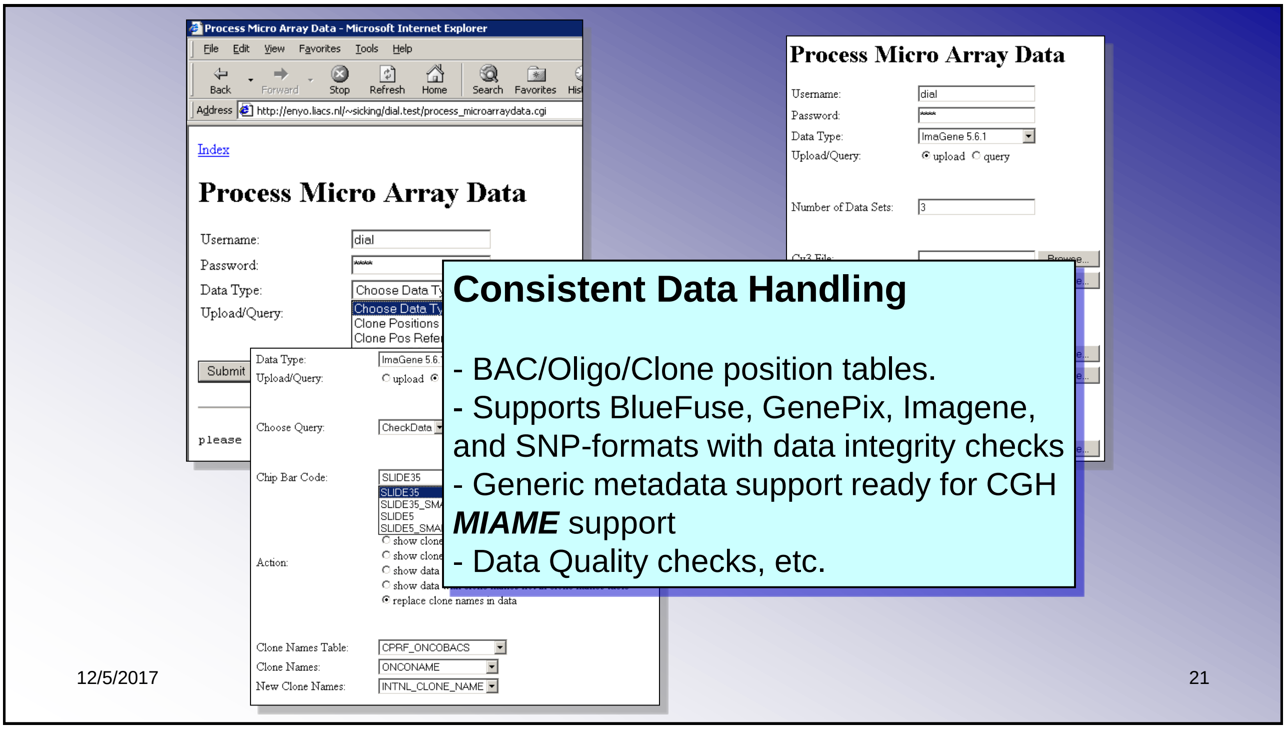
I_{green}= measured green channel intensity

Micro Array CGH Data Flow I



Micro Array CGH Data Flow II





Process Micro Array Data - Microsoft Internet Explorer

File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites His

Address http://enyo.liacs.nl/~sicking/dial.test/process_microarraydata.cgi

Index

Process Micro Array Data

Username: dial
Password: ****
Data Type: Choose Data Ty
Choose Data Ty
Clone Positions
Clone Pos Refer
Upload/Query:

Submit

please

Data Type: ImaGene 5.6.
Upload/Query: ○ upload ●
Choose Query: CheckData
Chip Bar Code: SLIDE35
SLIDE35
SLIDE35_SMA
SLIDE5
SLIDE5_SMA
Action: ○ show clone
○ show clone
○ show data
○ show data
● replace clone names in data
Clone Names Table: CPRF_ONCOBACS
Clone Names: ONCONAME
New Clone Names: INTNL_CLONE_NAME

Process Micro Array Data

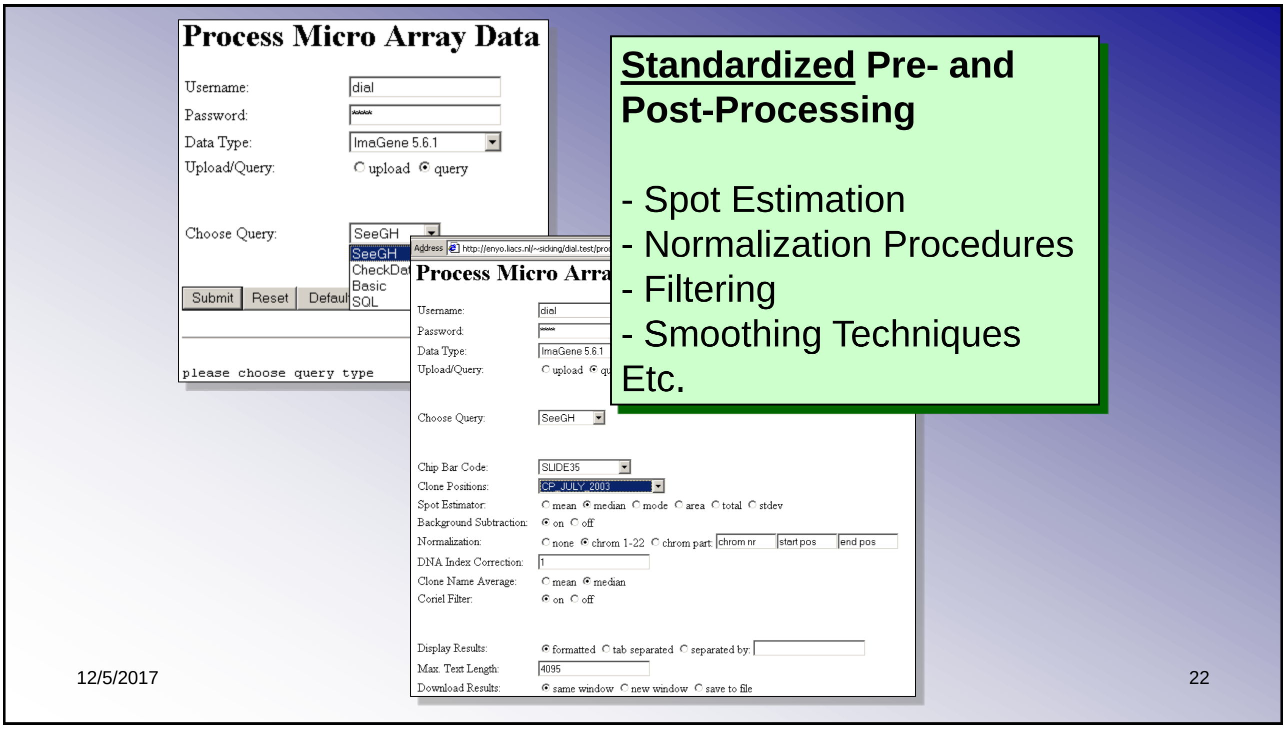
Username: dial
Password: ****
Data Type: ImaGene 5.6.1
Upload/Query: ● upload ○ query
Number of Data Sets: 3
Cy3 File: Browse...

Consistent Data Handling

- BAC/Oligo/Clone position tables.
- Supports BlueFuse, GenePix, Imagene, and SNP-formats with data integrity checks
- Generic metadata support ready for CGH ***MIAME*** support
- Data Quality checks, etc.

12/5/2017

21



Process Micro Array Data

Username: dial
Password: ****
Data Type: ImaGene 5.6.1
Upload/Query: ○ upload ● query
Choose Query: SeeGH
SeeGH
CheckDat
Basic
SQL

please choose query type

Address http://enyo.liacs.nl/~sicking/dial.test/proc

Process Micro Arra

Username: dial
Password: ****
Data Type: ImaGene 5.6.1
Upload/Query: ○ upload ● qu
Choose Query: SeeGH
Chip Bar Code: SLIDE35
Clone Positions: CP_JULY_2003
Spot Estimator: ○ mean ● median ○ mode ○ area ○ total ○ stdev
Background Subtraction: ● on ○ off
Normalization: ○ none ● chrom 1-22 ○ chrom part: chrom nr start pos end pos
DNA Index Correction: 1
Clone Name Average: ○ mean ● median
Coriel Filter: ● on ○ off
Display Results: ● formatted ○ tab separated ○ separated by:
Max. Text Length: 4095
Download Results: ● same window ○ new window ○ save to file

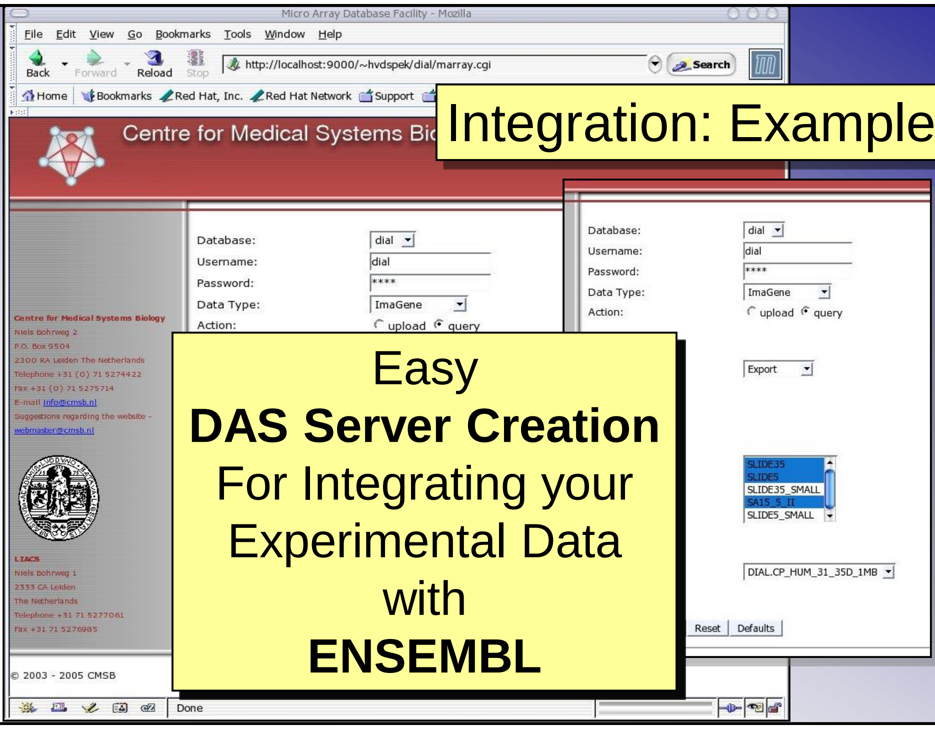
Standardized Pre- and Post-Processing

- Spot Estimation
- Normalization Procedures
- Filtering
- Smoothing Techniques

Etc.

12/5/2017

22



Integration: Example

Micro Array Database Facility - Mozilla

http://localhost:9000/~hvdspek/dial/marray.cgi

Centre for Medical Systems Biology

Centre for Medical Systems Biology
Niels Bohrweg 2
P.O. Box 9504
2300 RA Leiden The Netherlands
Telephone +31 (0) 71 5274422
Fax +31 (0) 71 5275714
E-mail info@cmsb.nl
Suggestions regarding the website - webmaster@cmsb.nl

LIACS
Niels Bohrweg 1
2333 CA Leiden
The Netherlands
Telephone +31 71 5277061
Fax +31 71 5276985

© 2003 - 2005 CMSB

Database: dial
Username: dial
Password: ****
Data Type: ImaGene
Action: ○ upload ● query

Export

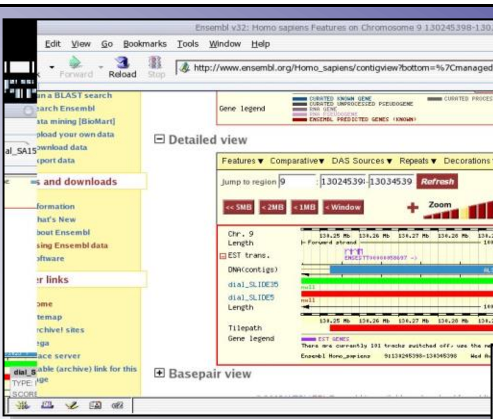
SLIDE35
SLIDE5
SLIDE35_SMALL
SA15_5_II
SLIDE5_SMALL

DIAL.CP_HUM_31_35D_1MB

Reset Defaults

Done

Easy
DAS Server Creation
For Integrating your Experimental Data with
ENSEMBL



Ensembl v32: Homo sapiens Features on Chromosome 9 130245398-130345398 - Mozilla

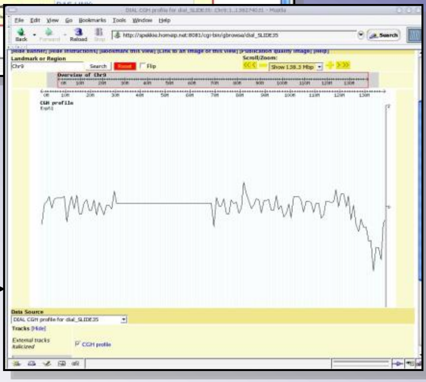
http://www.ensembl.org/Homo_sapiens/contigview?bottom=%7Cmanaged_extdas_dial_SA15_5_II%3Aoff&w=10000

Detailed view

Jump to region 9 130245398 13034539 Refresh

dial_SLIDE35
TYPE: loss
SCORE: -0.572024809140672
METHOD: loss
CATEGORY: loss

Basepair view

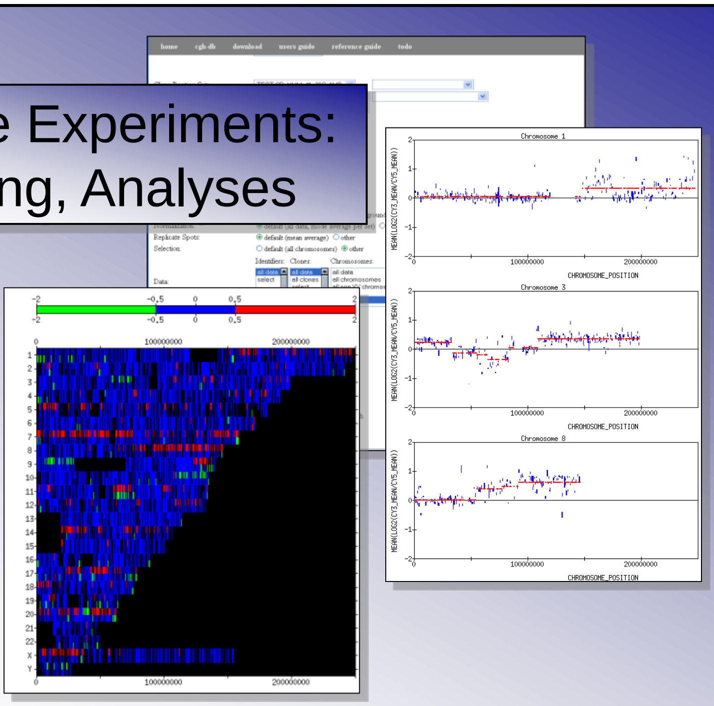


DIAL CGH profile for dial_SLIDE35

Get the original experimental Data in context.

Get a detailed GBrowse CGH Profile of your experiment.

Multiple Experiments: Viewing, Analyses


home cgh-db download users guide reference guide todo
Replicate Spots: ● default (mean average) ○ other
Selection: ○ default (all chromosomes) ● other
Identifiers: Clones: Chromosomes:
Data:
-2 -0,5 0 0,5 2
0 100000000 200000000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome 1
Chromosome 3
Chromosome 8
MEAN(LOG2(CY3_MEAN/CY5_MEAN))
CHROMOSOME_POSITION

DIAL Micro Array CGH

MIAME minimum information about a micro array experiment

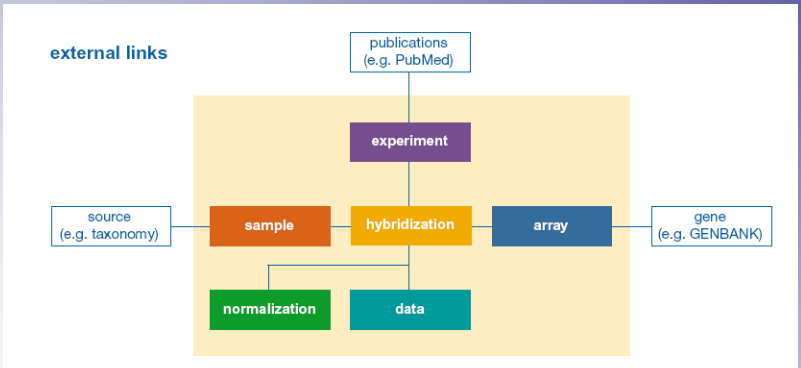
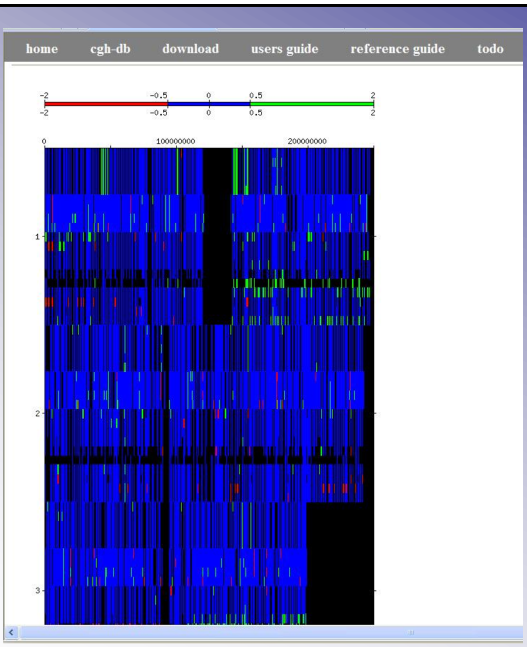


Fig. 3 A schematic representation of six components of a microarray experiment.



Integration of different experiments: BAC, Oligo, Bluefuse, Imagene, Affymetrix, Illumina, PacBio, etc.

DIAL CGH Database
Key Benefits

- **Consolidation** of the (Micro Array) Experiment
- **Converging data handling methods**
  Data Quality and Data Integration
- **Automatic BAC, Oligo and SNP referencing**
  with version management
- **Converging data annotation:** MIAME CGH
- **Straightforward Integration:** multi experiment; interfacing for further analyses; export to other databases; Ensembl; Data mining; Publication Export; Your Favorite Analysis Tool, etc.

Other Proof of Concepts and Projects

- Interfacing with MySQL data warehouse
- Clustering Module (Python, R)
- Data Mining Algorithms for Multiple CGH Experiments (C++)
- Experimentation with novel CGH Segmentation Methods (Matlab, R)
- Genotype Phenotype Integration using semantic wrappers (Postgres, JAVA)
- Processing pipeline: C#, R
- … cloud computing …

Genotype Databases

- Data Explosion
  - BAC 3500 data points
  - Oligo’s 20000 to 60000 data points 1000 experiments/year
  - 200k and 500k data points
  - Soon 5M data points for ‘routine’ diagnosis
  - 200MB - 1GB Images
  - De Novo Sequencing:
    - Complete genome scans => a multiple of 3Gbase scans
    - 1000 Genomes Project
  - Life Technologies’ **Benchtop Ion Proton^{TM} Sequencer** designed to decode a Human Genome in one day for $1000
  - Illumina, PacBIO, …
- Storage and Computational Requirements

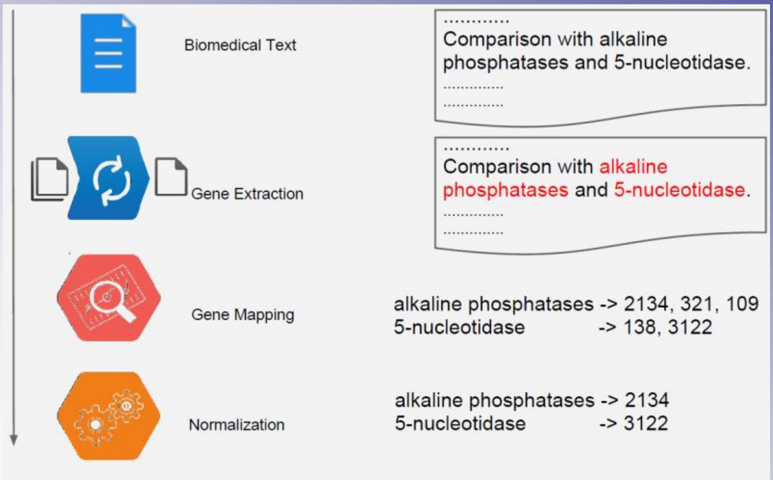
Challenges

Integration of Genomic Data

- Micro Array Expression Data mRNA levels, …
- Human Genome, Chimp, Rhesus, Mouse, etc.
- **Gene-name normalization**, dynamic ontologies, etc.
- Semantic integration

- Scale up of routine analysis
- Scale up of research analysis over integrated data sets
- Data mining for hidden relations
- …

A Gene Name Normalization Framework
Alex Wang (2013)



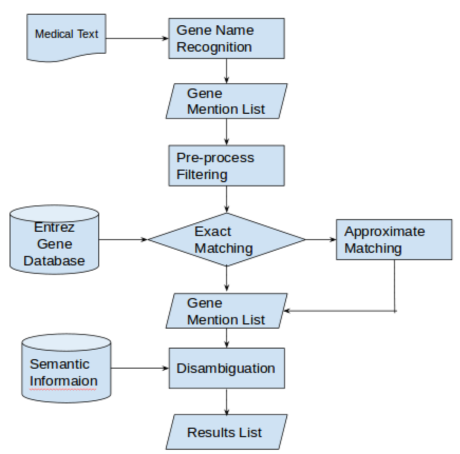
....
Comparison with alkaline phosphatases and 5-nucleotidase.
....
....

....
Comparison with alkaline phosphatases and 5-nucleotidase.
....
....

alkaline phosphatases -> 2134, 321, 109
5-nucleotidase -> 138, 3122

alkaline phosphatases -> 2134
5-nucleotidase -> 3122

A Gene Name Normalization Framework
Alex Wang (2013)



Longevity Studies at LUMC
Group headed by E. Slagboom (LUMC)

Data mining studies by Fabrice Colas (LIACS)

- Mining genetic data sets
- 1-, 2-, and 3-itemsets (frequent item sets)
- Solving the problems in reasonable time was only possible using parallel computing (DAS3)

Towards a Classification of Osteo Arthritis subtypes in Subjects with Symptomatic OA at Multiple Joint Sites. F. Colas et al NBIC-ISNB2007

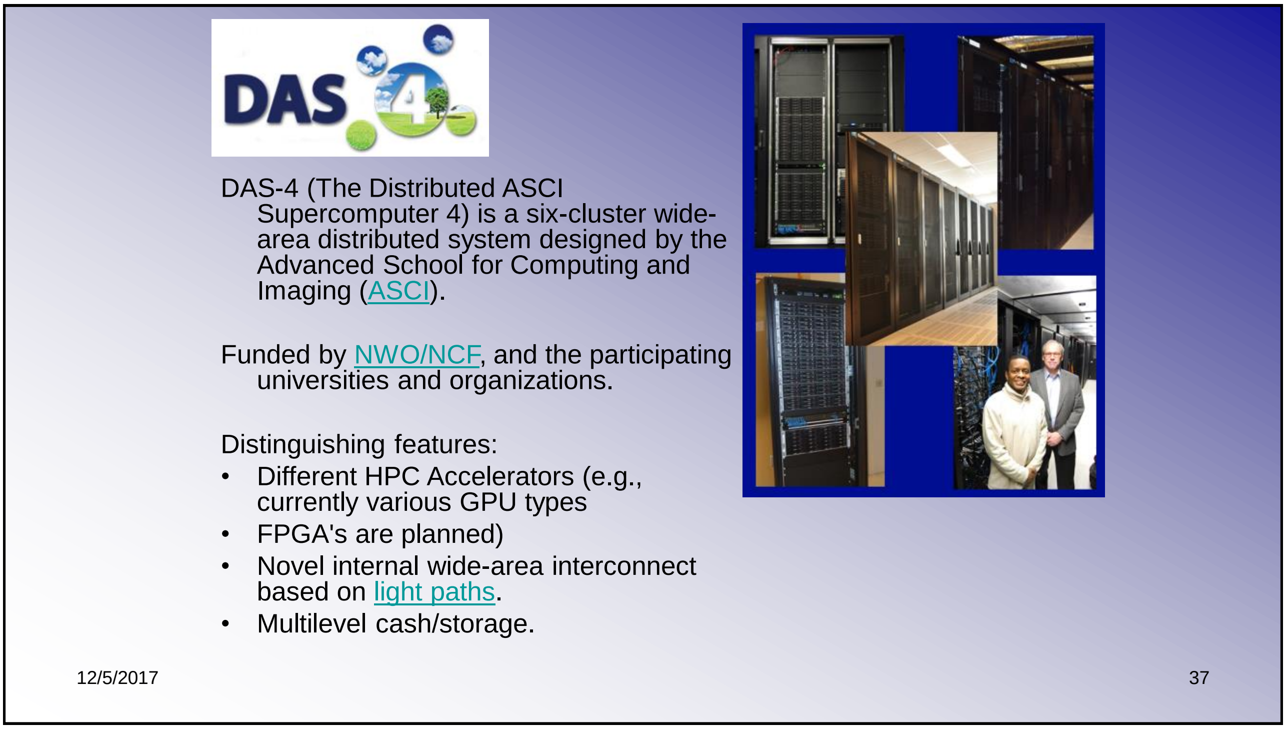
GARP study of OA (Osteo Arthritis) subtypes

- Identifying genetic factors
- Assist in development of new treatments
- Genetic causes of the disease are difficult to obtain because of the **clinical heterogeneity** of the disease
- Identification of homogeneous subgroups of OA
- Identify and characterize potentially new disease subtypes using machine learning techniques
- Parallel Computation (DAS3)

DAS3 GRID-Computing

- Data mining in Bioinformatics offer many challenging tasks in which DAS3 plays an essential role:
  - research on novel scalable high performance segmentation of high dimensional and high volume feature spaces.
  - Development and evaluation of novel high performance techniques for data mining
  - research on novel scalable data(base) structures for efficient data querying, analysis and mining of high volume data sets

DAS-3: The Next Generation Grid Infrastructure in The Netherlands
TU Delft
Leiden University
Vrije Universiteit
MultimediaN
University of Amsterdam
A Computer Science Grid with revolutionary Optical Interconnect

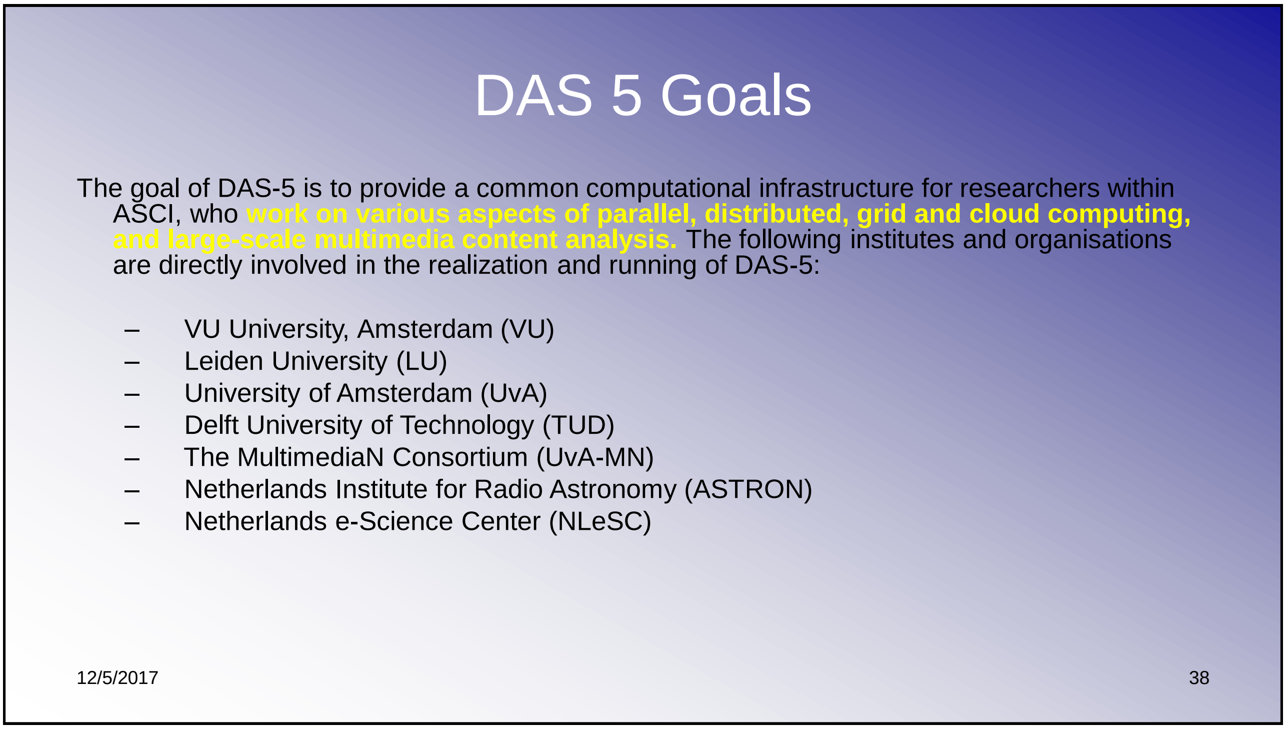


DAS-4 (The Distributed ASCI Supercomputer 4) is a six-cluster wide-area distributed system designed by the Advanced School for Computing and Imaging (ASCI).

Funded by NWO/NCF, and the participating universities and organizations.

Distinguishing features:

- Different HPC Accelerators (e.g., currently various GPU types
- FPGA's are planned)
- Novel internal wide-area interconnect based on light paths.
- Multilevel cash/storage.



DAS 5 Goals

The goal of DAS-5 is to provide a common computational infrastructure for researchers within ASCI, who **work on various aspects of parallel, distributed, grid and cloud computing, and large-scale multimedia content analysis.** The following institutes and organisations are directly involved in the realization and running of DAS-5:

- VU University, Amsterdam (VU)
- Leiden University (LU)
- University of Amsterdam (UvA)
- Delft University of Technology (TUD)
- The MultimediaN Consortium (UvA-MN)
- Netherlands Institute for Radio Astronomy (ASTRON)
- Netherlands e-Science Center (NLeSC)

DAS 5

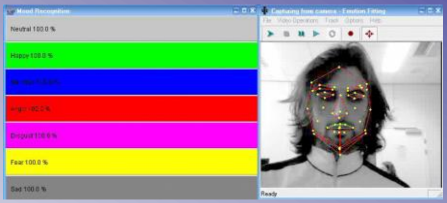
6 Jul 2015

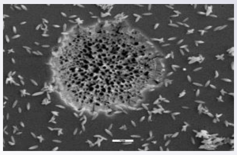
DAS-5 is fully operational!

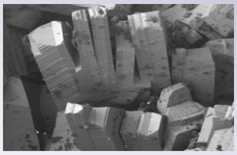
liacs medialab
LEIDEN UNIVERSITY

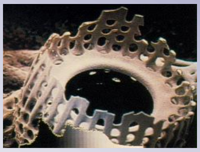
Content Based Indexing and Retrieval Techniques

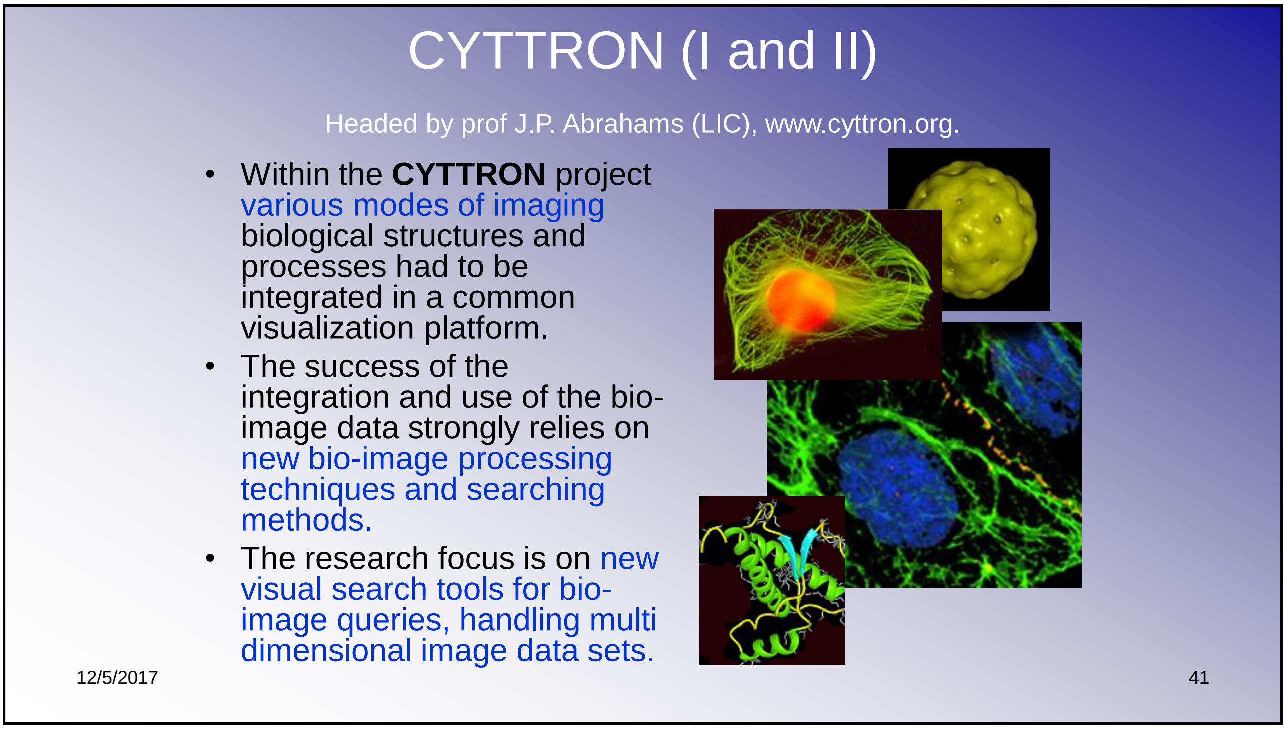
- Image Databases
- Speech Databases
- Video Databases
- Multimodal Databases
- Face recognition, bimodal emotion recognition (N. Sebe, UVA), Semantic Audio Indexing, etc.







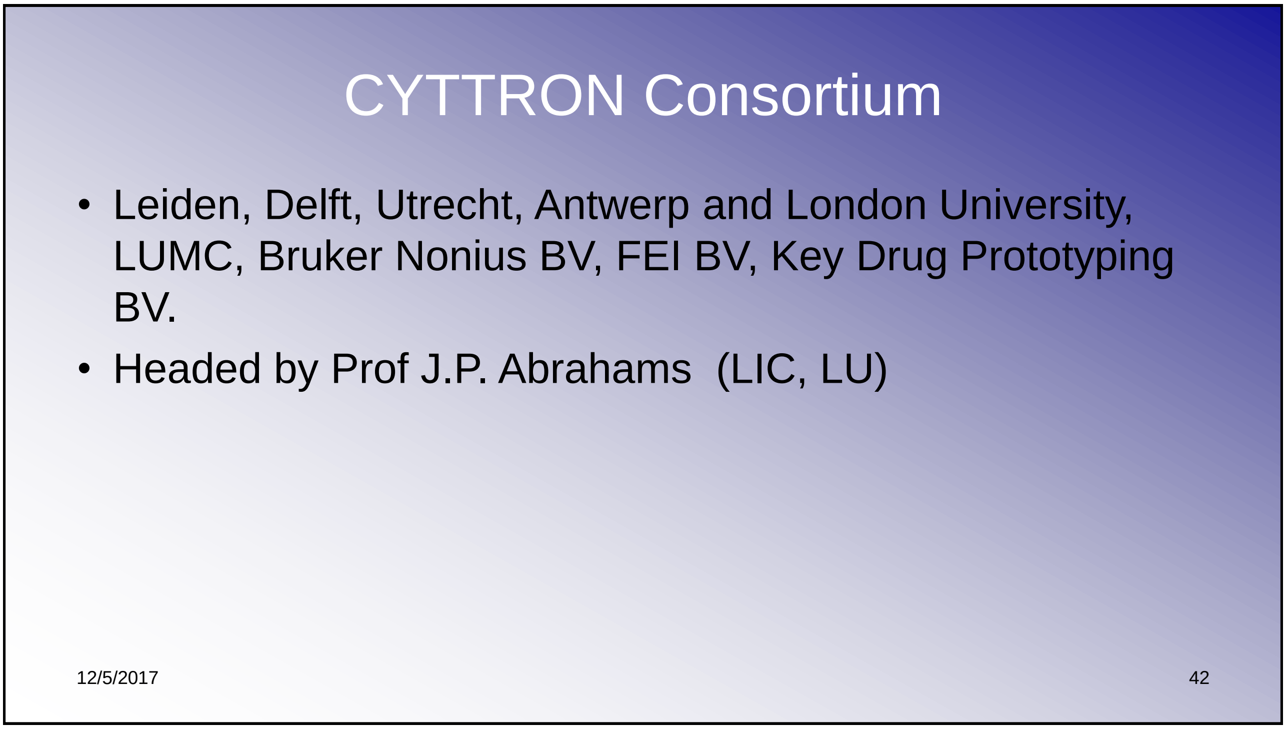




CYTTRON (I and II)

Headed by prof J.P. Abrahams (LIC), www.cyttron.org.

- Within the **CYTTRON** project various modes of imaging biological structures and processes had to be integrated in a common visualization platform.
- The success of the integration and use of the bio-image data strongly relies on new bio-image processing techniques and searching methods.
- The research focus is on new visual search tools for bio-image queries, handling multi dimensional image data sets.

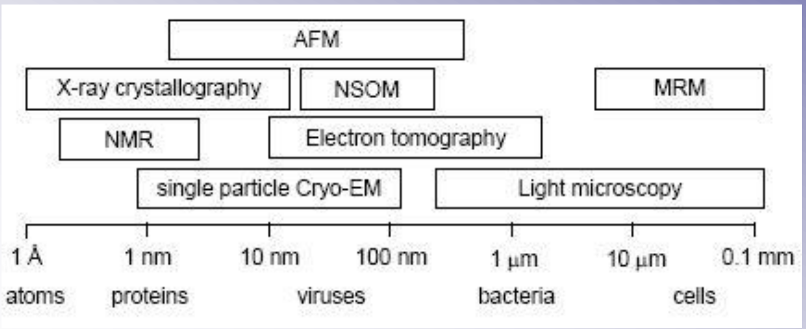


CYTTRON Consortium

- Leiden, Delft, Utrecht, Antwerp and London University, LUMC, Bruker Nonius BV, FEI BV, Key Drug Prototyping BV.
- Headed by Prof J.P. Abrahams (LIC, LU)

CYTTRON

- Different Bio-Imaging Techniques:
  - Light Microscopy
  - MRM
  - Confocal laser Microscopy
  - EM, Cryo, 3D EM
  - NMR
  - Crystallography
  - Etc.



Fluorescence Microscopy

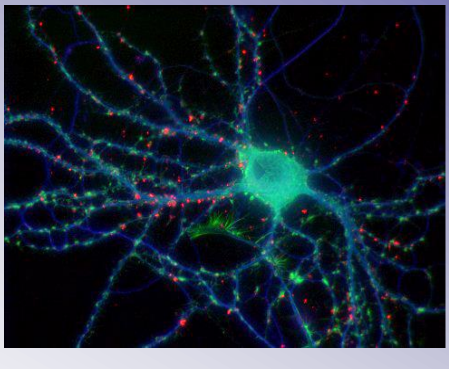


Figure from http://www.wadsworth.org/cores/alm/: A multi-wavelength, three dimensional, wide-field immunofluorescence image of a fixed neuron. The projection was generated using an extended depth of field algorithm. **Cell body** labeled for tubulin is shown in blue, **F-actin** in green, and **presynaptic protein** in Red. Specimen courtesy of Natalie Dowell-Mesfin BMS-PhD student

Fluorescence Microscopy

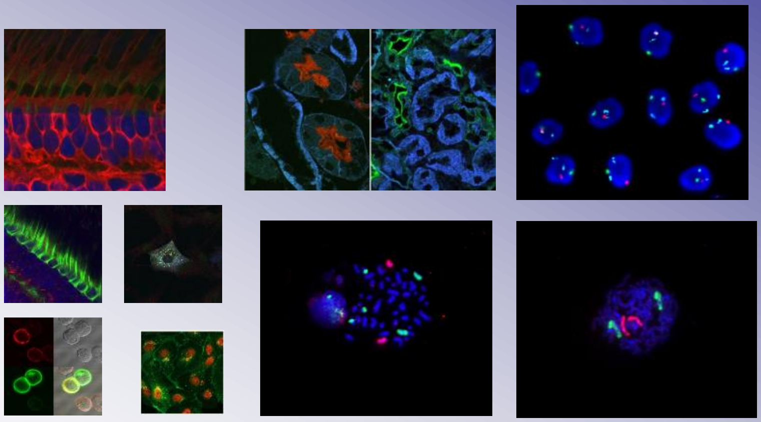


Figure from http://hsc.unm.edu/pathology/microscopy/instru.htm

Confocal Laser Scanning Microscope

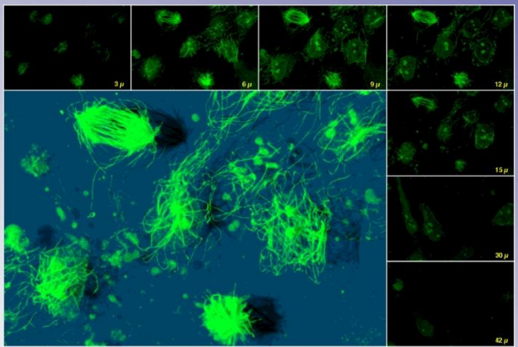
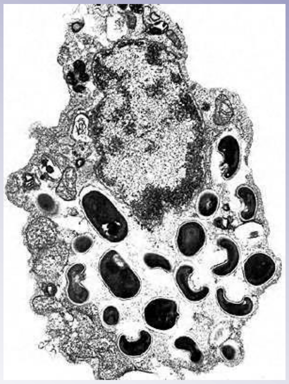
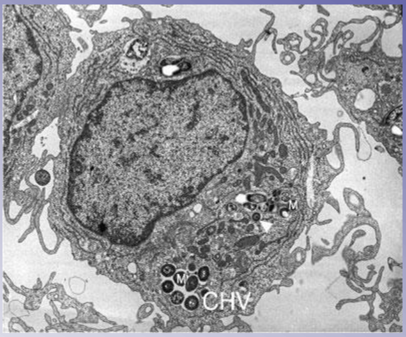


Figure (from http://www.mih.unibas.ch/Booklet/Booklet96/Chapter1/Chapter1.html). *Seven representative optical sections selected from 81 confocal planes (corresponding to a depth of 50 mm) "cut" through a collagen matrix containing growing fibroblasts labeled with fluorescent antibodies to tubulin. Inset, composite shadow-projection image of all 81 confocal sections revealing the spindle apparatus of dividing cells and the regular microtubular network of interphase (i.e., non-dividing) cells*

Electron Microscopy

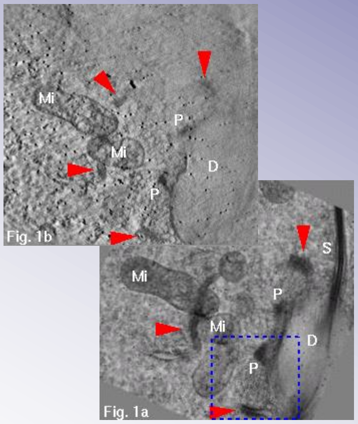


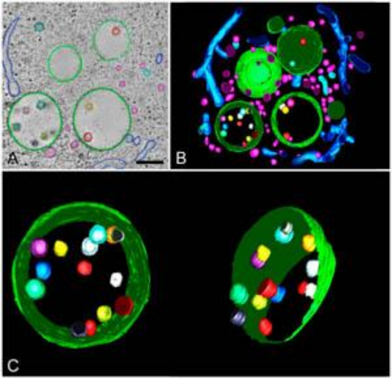


- Some standard (old technique) electron microscopic slides

12/5/2017 47

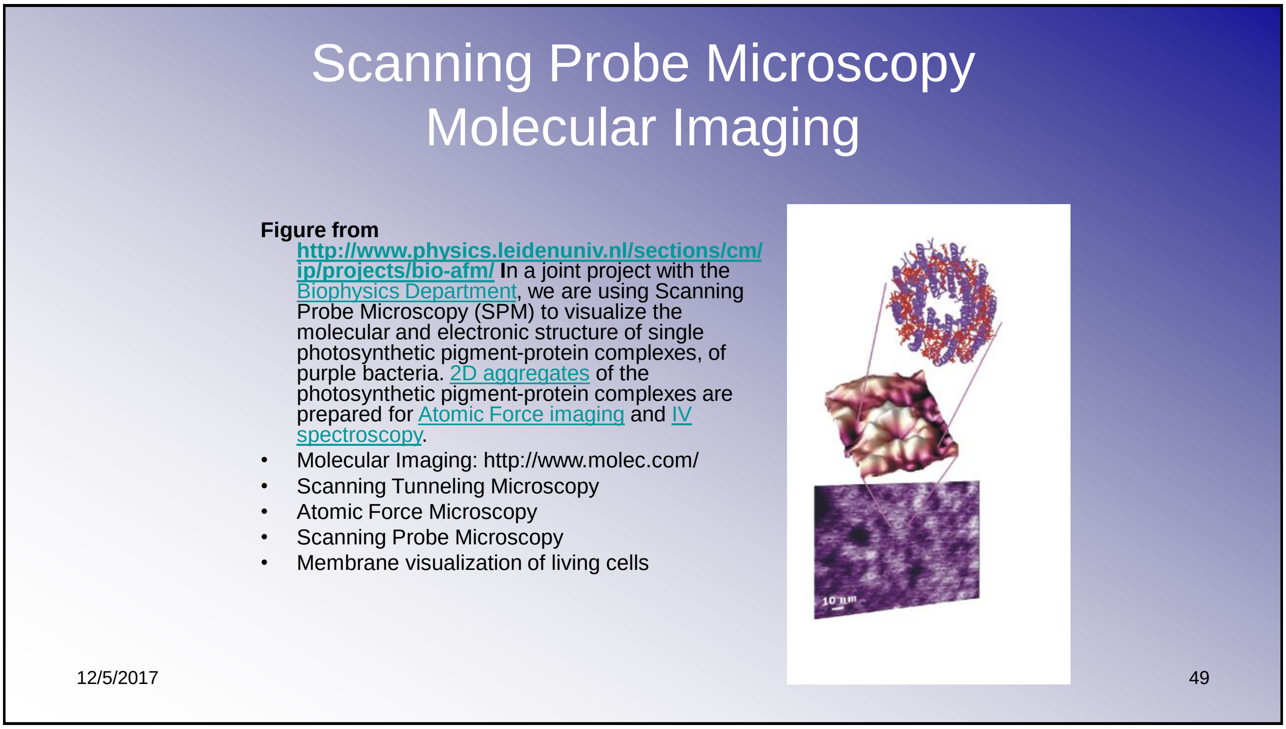
3D Electron Microscope Electron Tomography





Images from http://www.bio.uu.nl/mcb/3dem/

12/5/2017 48



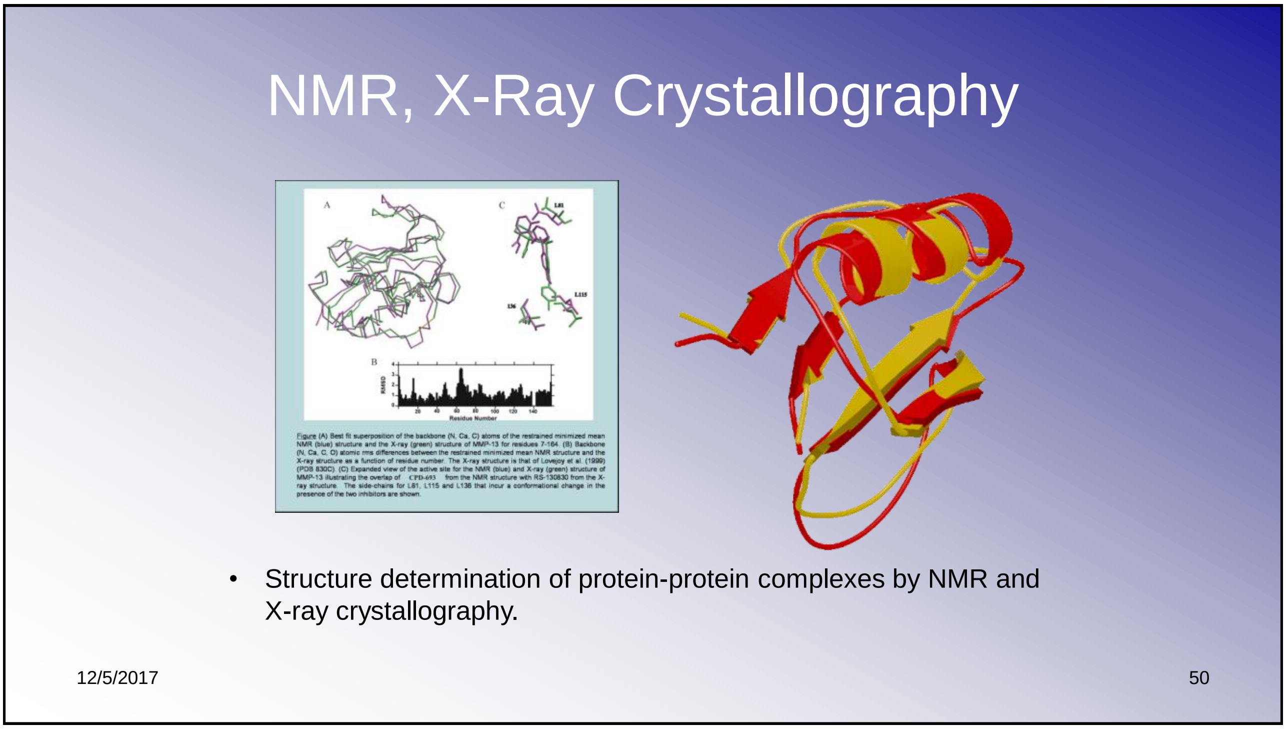
Scanning Probe Microscopy Molecular Imaging

Figure from **http://www.physics.leidenuniv.nl/sections/cm/ip/projects/bio-afm/** In a joint project with the Biophysics Department, we are using Scanning Probe Microscopy (SPM) to visualize the molecular and electronic structure of single photosynthetic pigment-protein complexes, of purple bacteria. 2D aggregates of the photosynthetic pigment-protein complexes are prepared for Atomic Force imaging and IV spectroscopy.

- Molecular Imaging: http://www.molec.com/
- Scanning Tunneling Microscopy
- Atomic Force Microscopy
- Scanning Probe Microscopy
- Membrane visualization of living cells

12/5/2017

49



NMR, X-Ray Crystallography

A
C
L81
L115
136
B
RMSD
0
1
2
3
4
20
40
60
80
100
120
140
Residue Number

Figure (A) Best fit superposition of the backbone (N, Ca, C) atoms of the restrained minimized mean NMR (blue) structure and the X-ray (green) structure of MMP-13 for residues 7-164. (B) Backbone (N, Ca, C, O) atomic rms differences between the restrained minimized mean NMR structure and the X-ray structure as a function of residue number. The X-ray structure is that of Lovejoy et al. (1999) (PDB 830C). (C) Expanded view of the active site for the NMR (blue) and X-ray (green) structure of MMP-13 illustrating the overlap of CPD-693 from the NMR structure with RS-130830 from the X-ray structure. The side-chains for L81, L115 and L136 that incur a conformational change in the presence of the two inhibitors are shown.

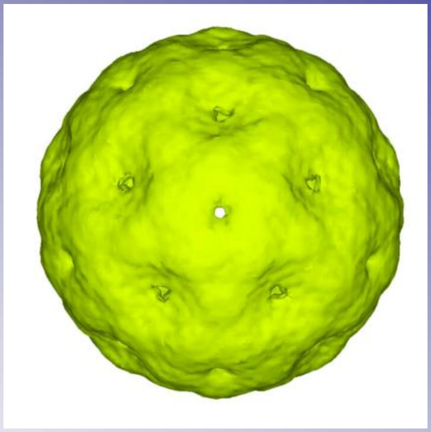
- Structure determination of protein-protein complexes by NMR and X-ray crystallography.

12/5/2017

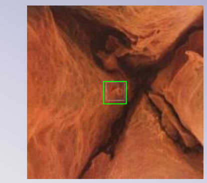
50

Single Particle Cryo Electron Microscopy

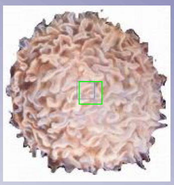
- Reconstruction made by Tyson (reconstruction package).



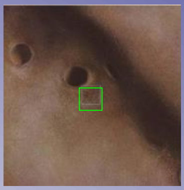
Example: White blood cell



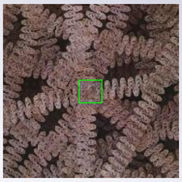
10^{-4}m



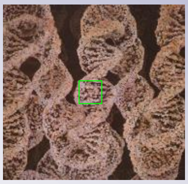
10^{-5}m



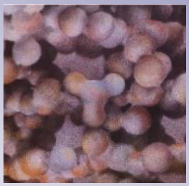
10^{-6}m



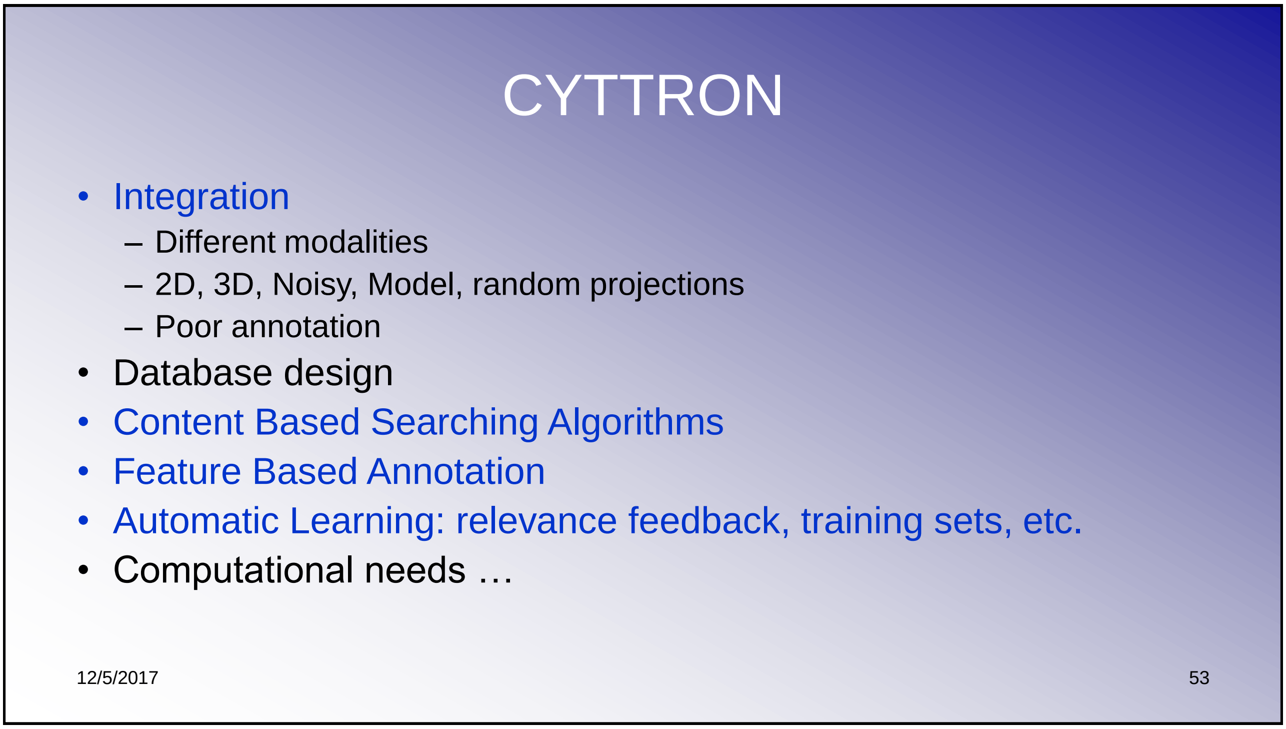
10^{-7}m



10^{-8}m

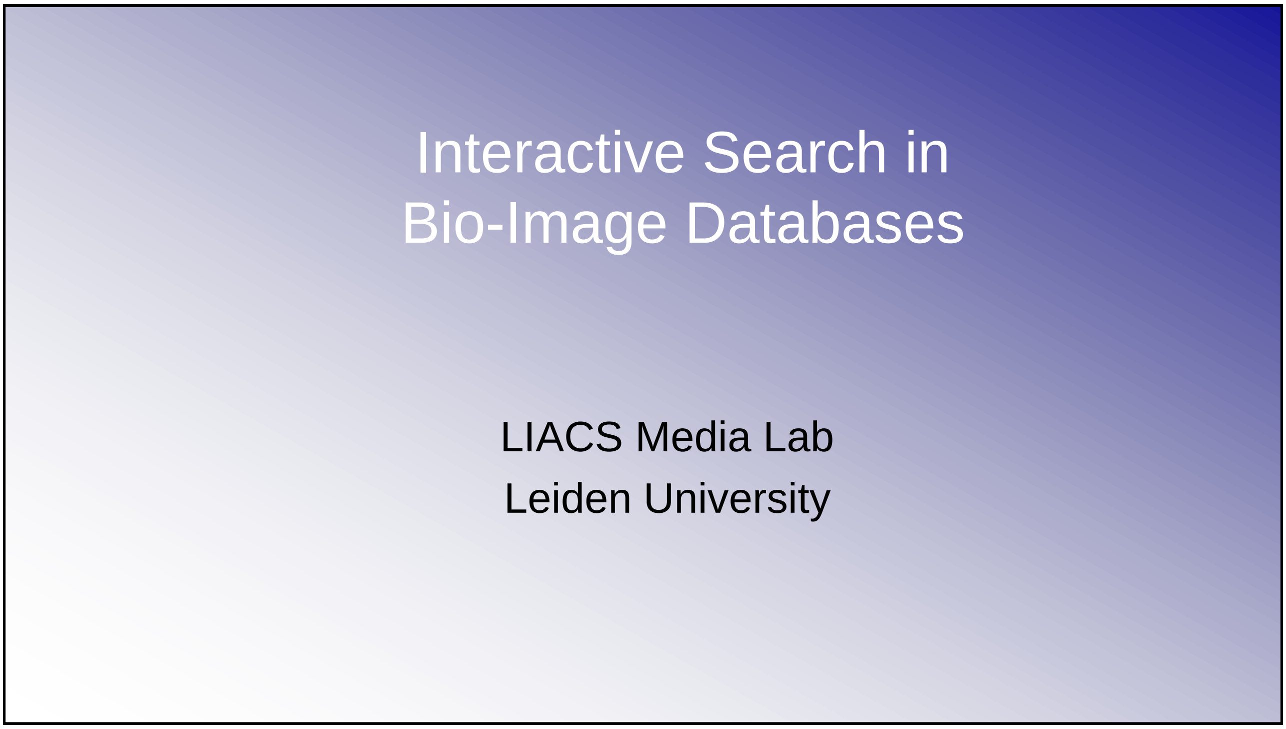


10^{-9}m



CYTTRON

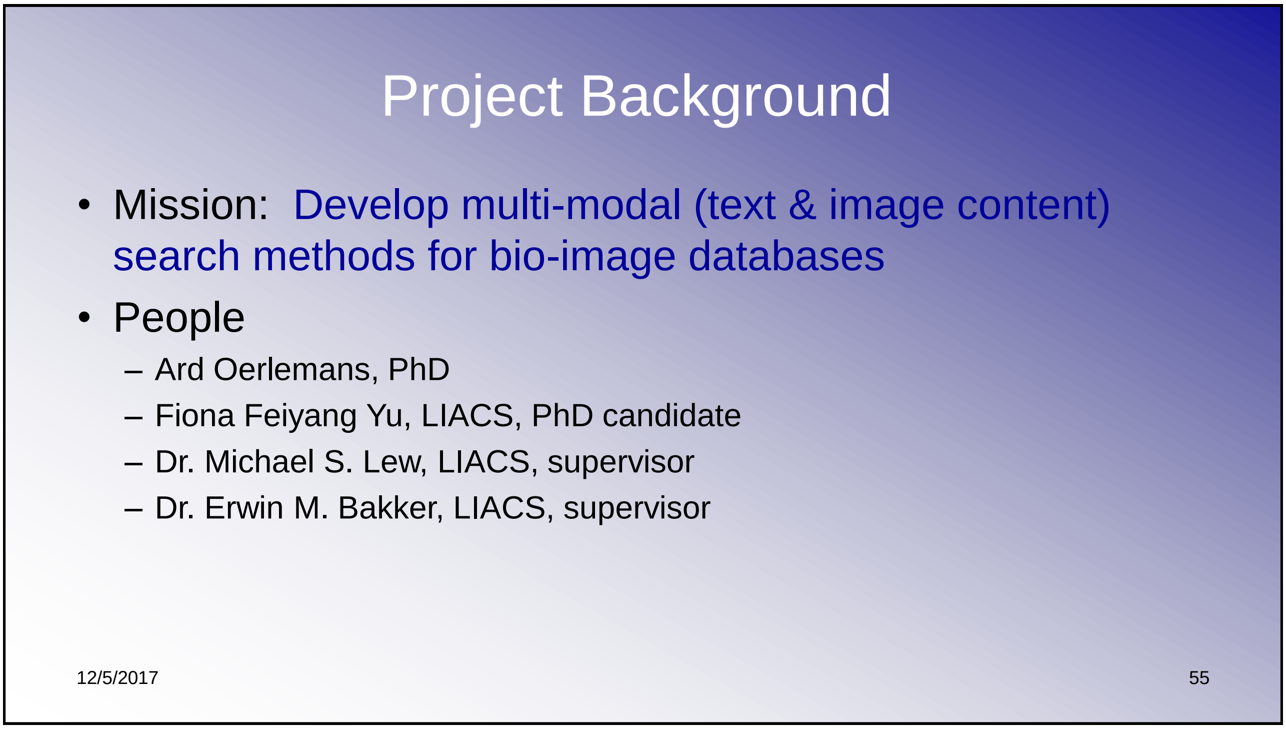
- Integration
  - Different modalities
  - 2D, 3D, Noisy, Model, random projections
  - Poor annotation
- Database design
- Content Based Searching Algorithms
- Feature Based Annotation
- Automatic Learning: relevance feedback, training sets, etc.
- Computational needs …



Interactive Search in Bio-Image Databases

LIACS Media Lab

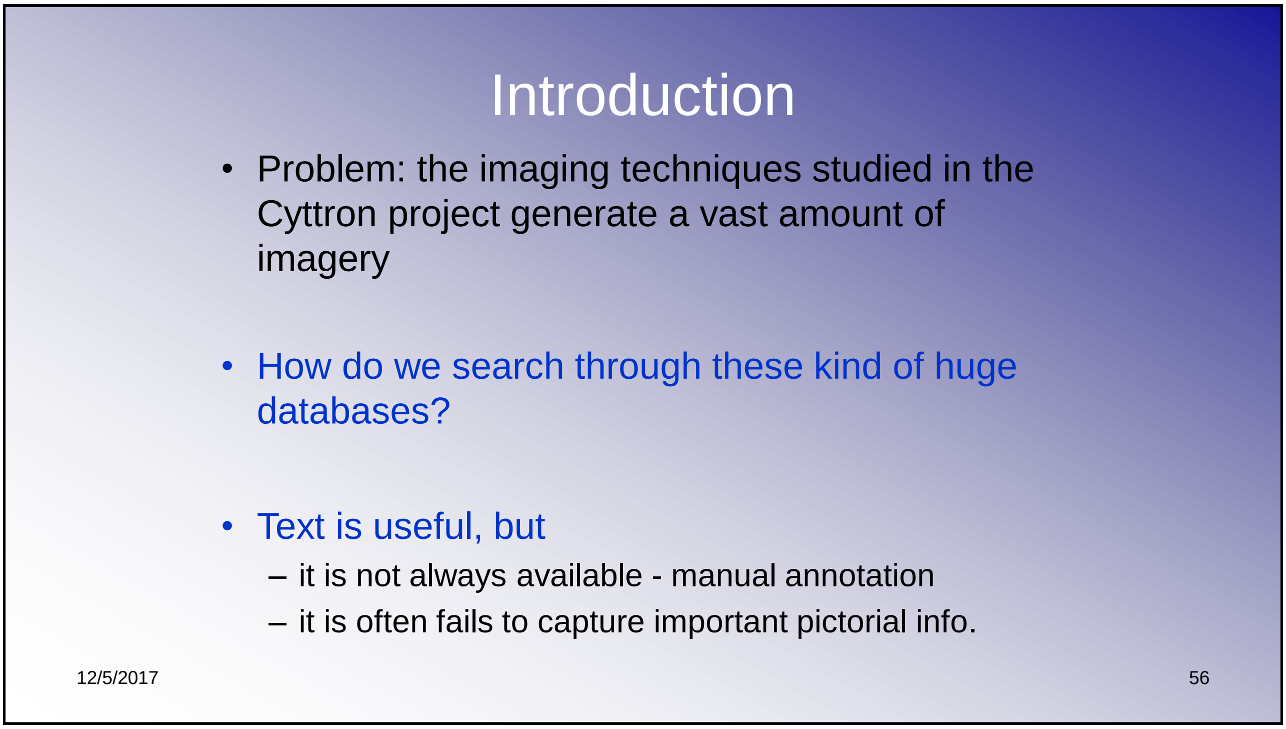
Leiden University



Project Background

- Mission: Develop multi-modal (text & image content) search methods for bio-image databases
- People
  - – Ard Oerlemans, PhD
  - – Fiona Feiyang Yu, LIACS, PhD candidate
  - – Dr. Michael S. Lew, LIACS, supervisor
  - – Dr. Erwin M. Bakker, LIACS, supervisor

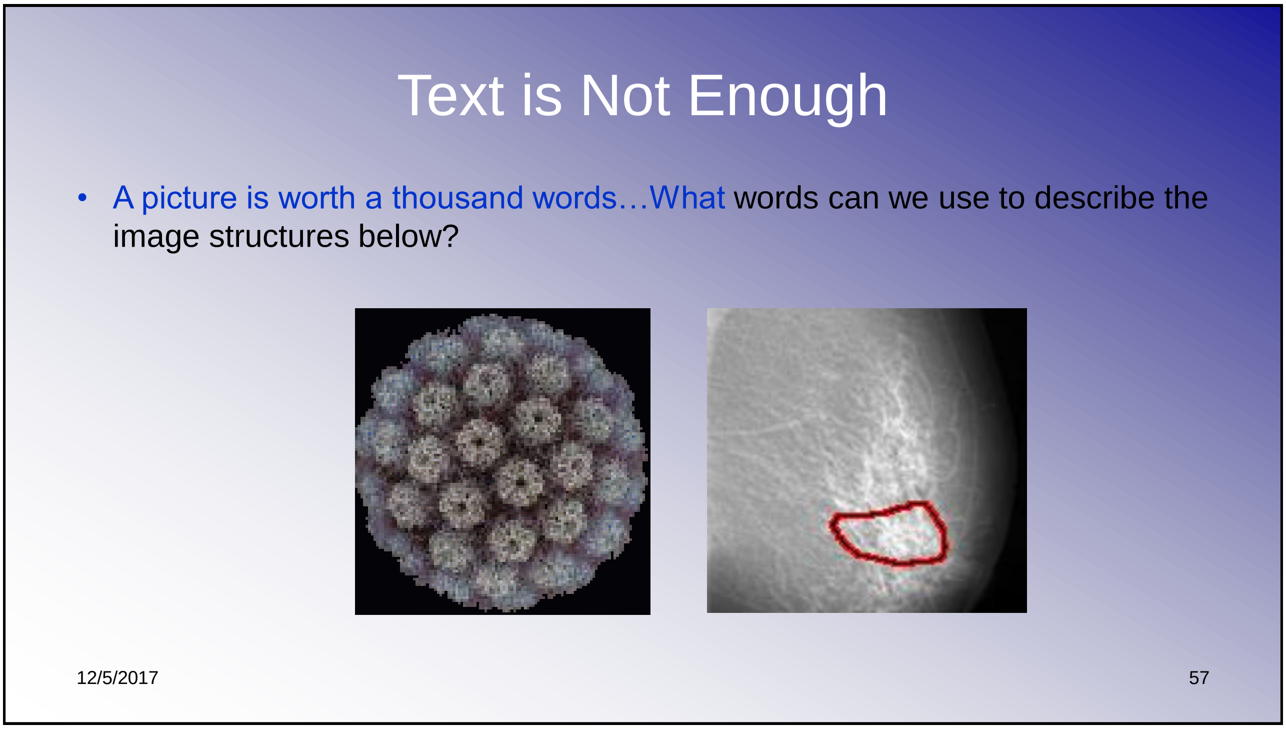
12/5/2017 55



Introduction

- Problem: the imaging techniques studied in the Cyttron project generate a vast amount of imagery
- How do we search through these kind of huge databases?
- Text is useful, but
  - – it is not always available - manual annotation
  - – it is often fails to capture important pictorial info.

12/5/2017 56



Text is Not Enough

- A picture is worth a thousand words…What words can we use to describe the image structures below?

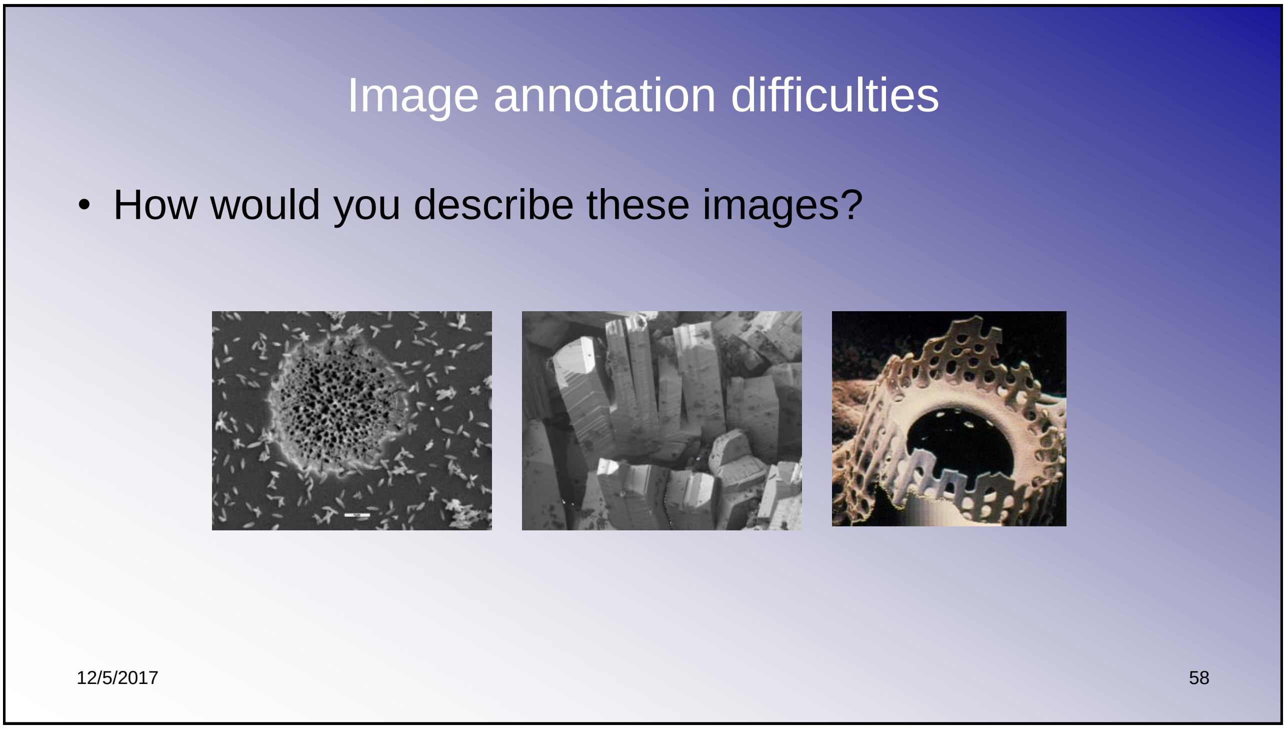


Image annotation difficulties

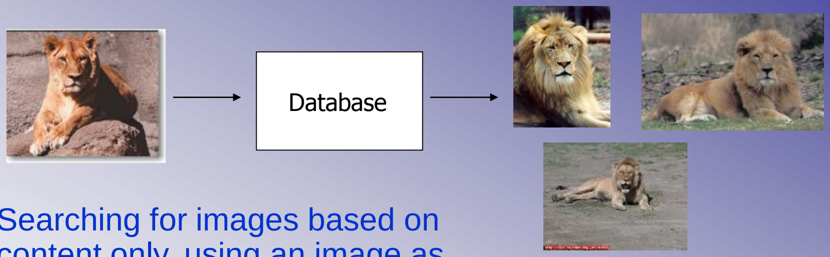
- How would you describe these images?

At that time: Going beyond Google

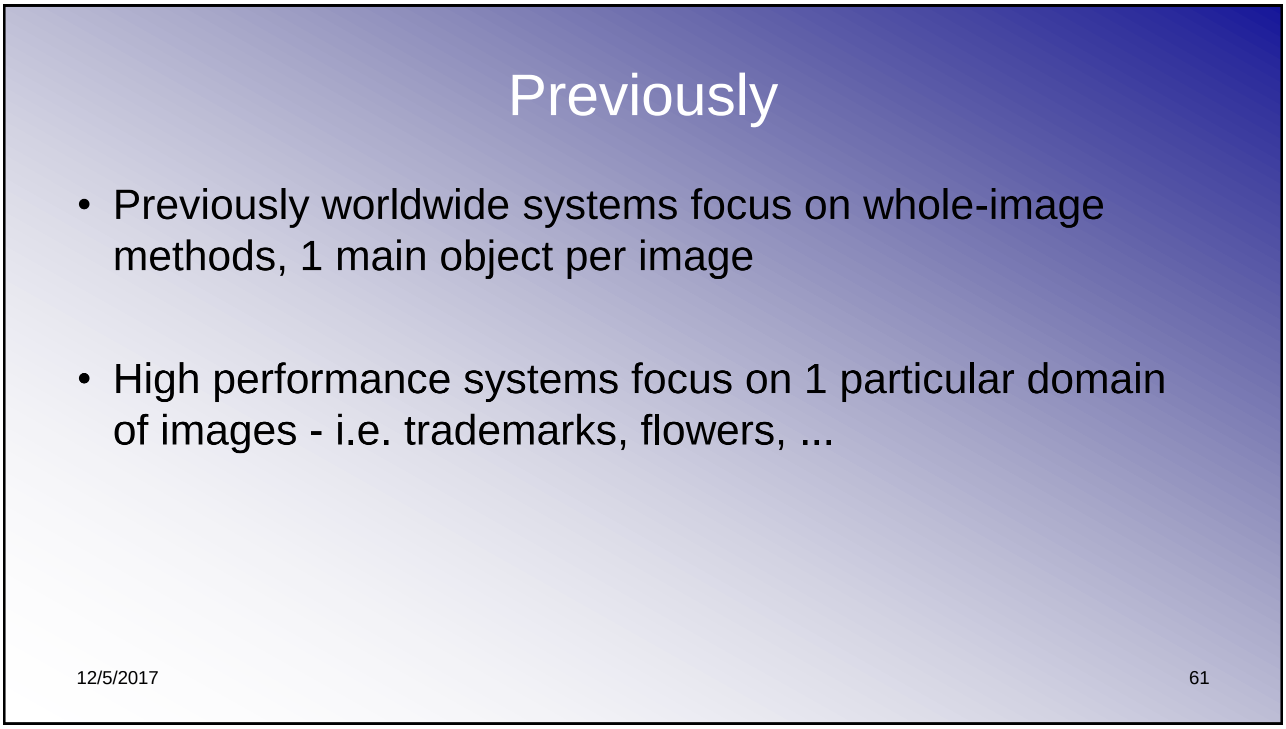
- Google used to search only using text annotation.
- Content-based retrieval techniques nowadays also incorporated at Google.

- We will be searching on both text and the pictorial content of the imagery.

Content-based image retrieval

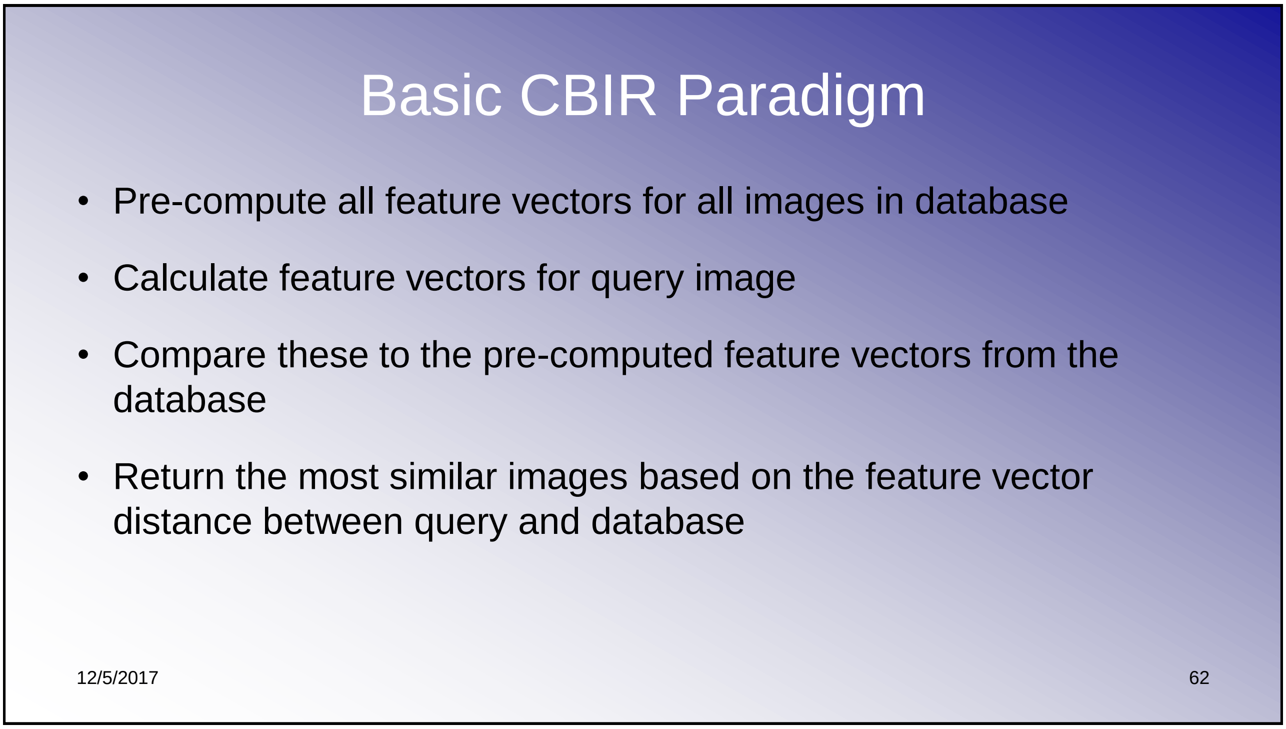


- Searching for images based on content only, using an image as a query
- Using text search for images requires every image to be annotated. Disadvantages:
  - Annotating images is time-consuming
  - Annotation can be incomplete
  - Annotation can be almost impossible



Previously

- Previously worldwide systems focus on whole-image methods, 1 main object per image
- High performance systems focus on 1 particular domain of images - i.e. trademarks, flowers, ...

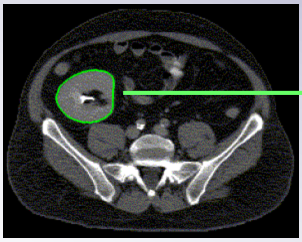


Basic CBIR Paradigm

- Pre-compute all feature vectors for all images in database
- Calculate feature vectors for query image
- Compare these to the pre-computed feature vectors from the database
- Return the most similar images based on the feature vector distance between query and database

Example

- Given the boundary, convert the interior region to a texture representation such as Linear Binary Patterns
- Quantize the information for efficient searching:



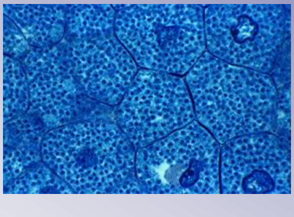
Texture representation:
Linear Binary Patterns

Feature Vector:
F[0…255]

Local Binary Patterns:

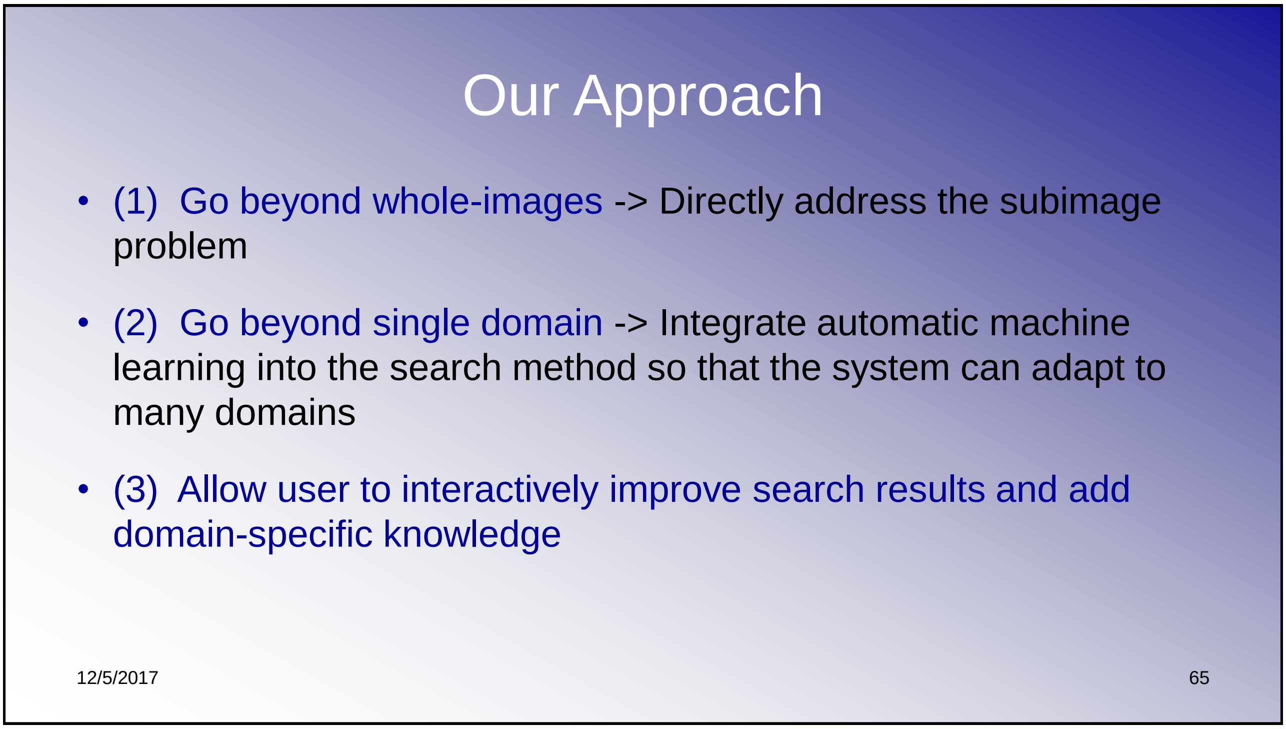
```
97    67  20        1 1 0
33    34   5    ->  0 0 0  ->  (110 01 111)
101 123 98          1 1 1
```

Basic CBIR paradigm



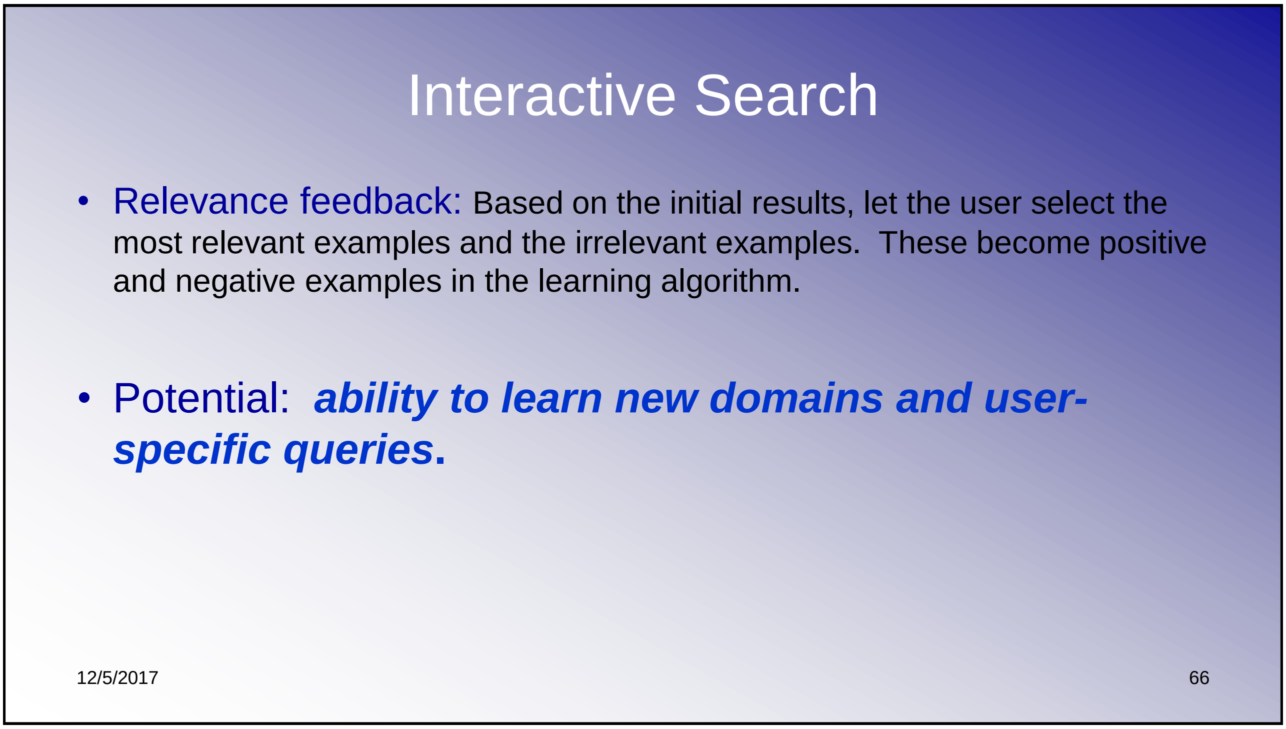
Average color → (23, 37, 241)

- Describe a specific visual property (feature) of an image as a vector
  - RGB Historgrams
  - Local Binary Patterns
  - Etc.
- Extract features for all database images
- Extract features from query image
- Calculate distance between query image and all database images
- Rank images by distance



Our Approach

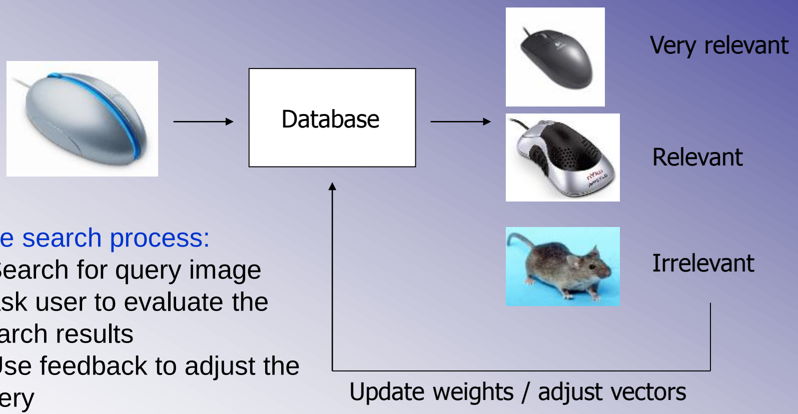
- (1) Go beyond whole-images -> Directly address the subimage problem
- (2) Go beyond single domain -> Integrate automatic machine learning into the search method so that the system can adapt to many domains
- (3) Allow user to interactively improve search results and add domain-specific knowledge



Interactive Search

- Relevance feedback: Based on the initial results, let the user select the most relevant examples and the irrelevant examples. These become positive and negative examples in the learning algorithm.
- Potential: ***ability to learn new domains and user-specific queries*****.**

Relevance Feedback



Database

Very relevant

Relevant

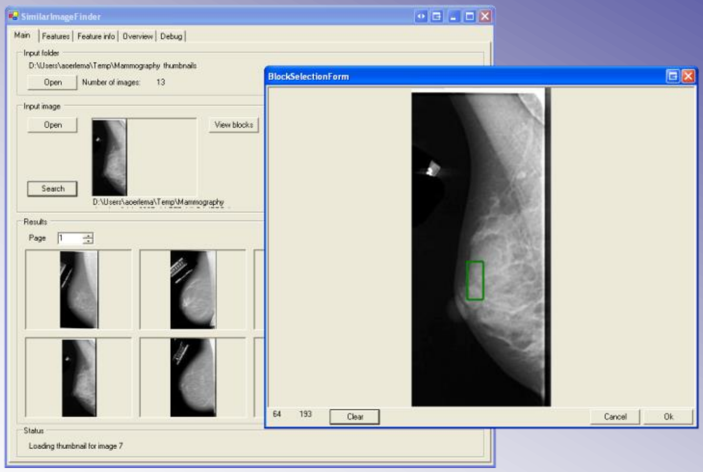
Irrelevant

Update weights / adjust vectors

Iterative search process:

- Search for query image
- Ask user to evaluate the search results
- Use feedback to adjust the query
- Repeat process until user is satisfied

Example Implementation



Input folder

D:\Users\aoerlema\Temp\Mammography thumbnails

Number of images: 13

Input image

D:\Users\aoerlema\Temp\Mammography

Results

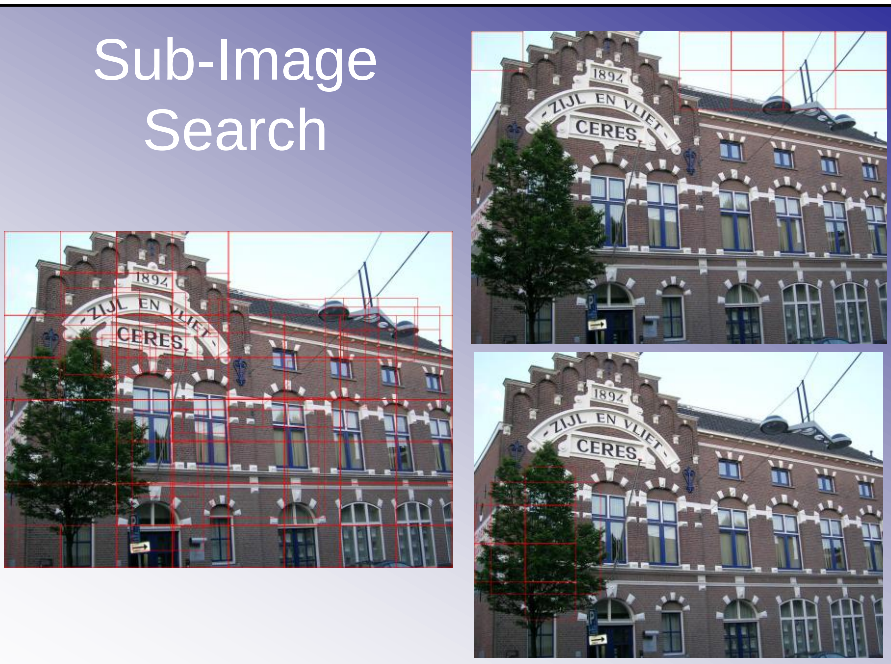
Page 1

64 193

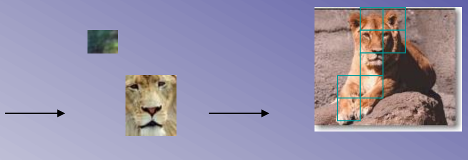
Status

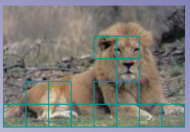
Loading thumbnail for image 7

Sub-Image Search



Sub-image search

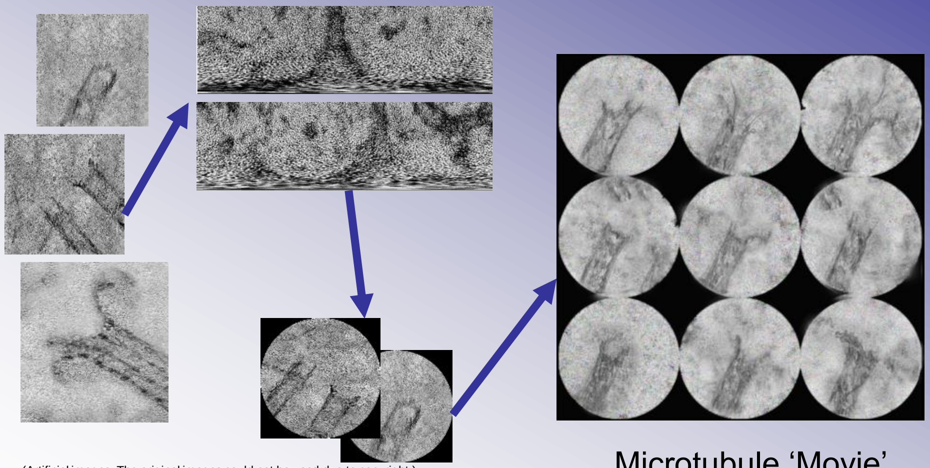




- Let the user select one or more parts of the query image
- For each database image, calculate the number of sub-images matching (are close to) the selected parts
- Rank results based on number of matching sub-images

Automatic Registration of Microtubule Images

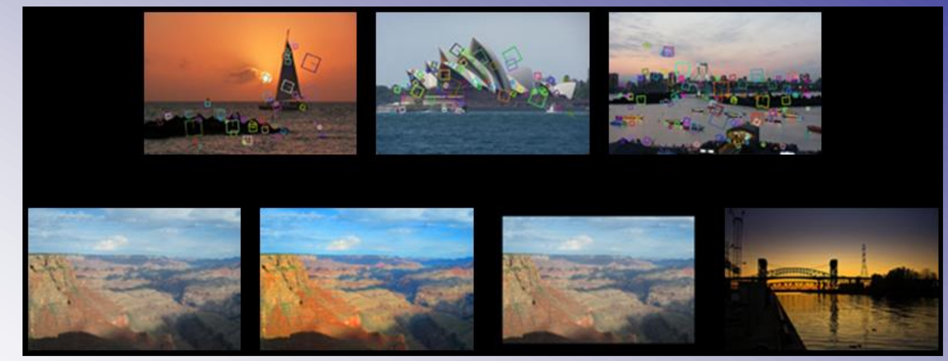
Feiyang Yu, Ard Oerlemans
Erwin M. Bakker and Michael S. Lew



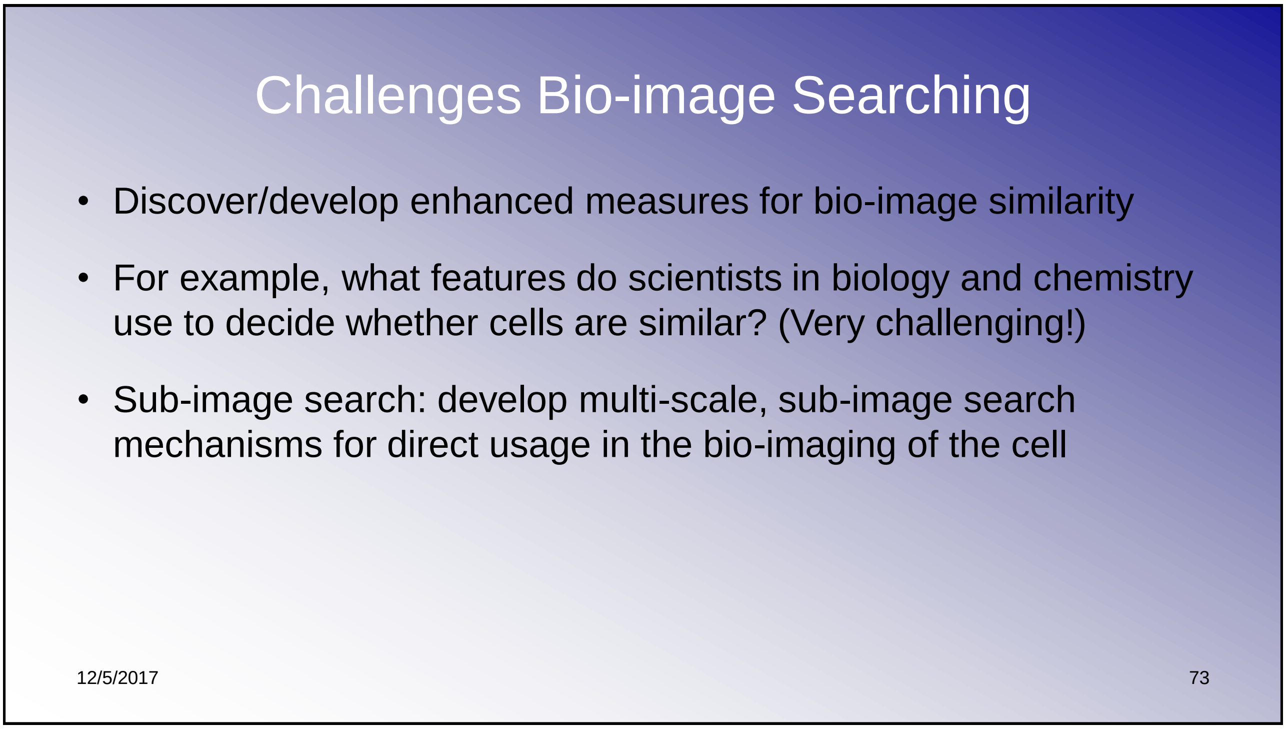
(Artificial images. The original images could not be used due to copyright.)

Microtubule 'Movie'

TOP-SURF: Visual Words

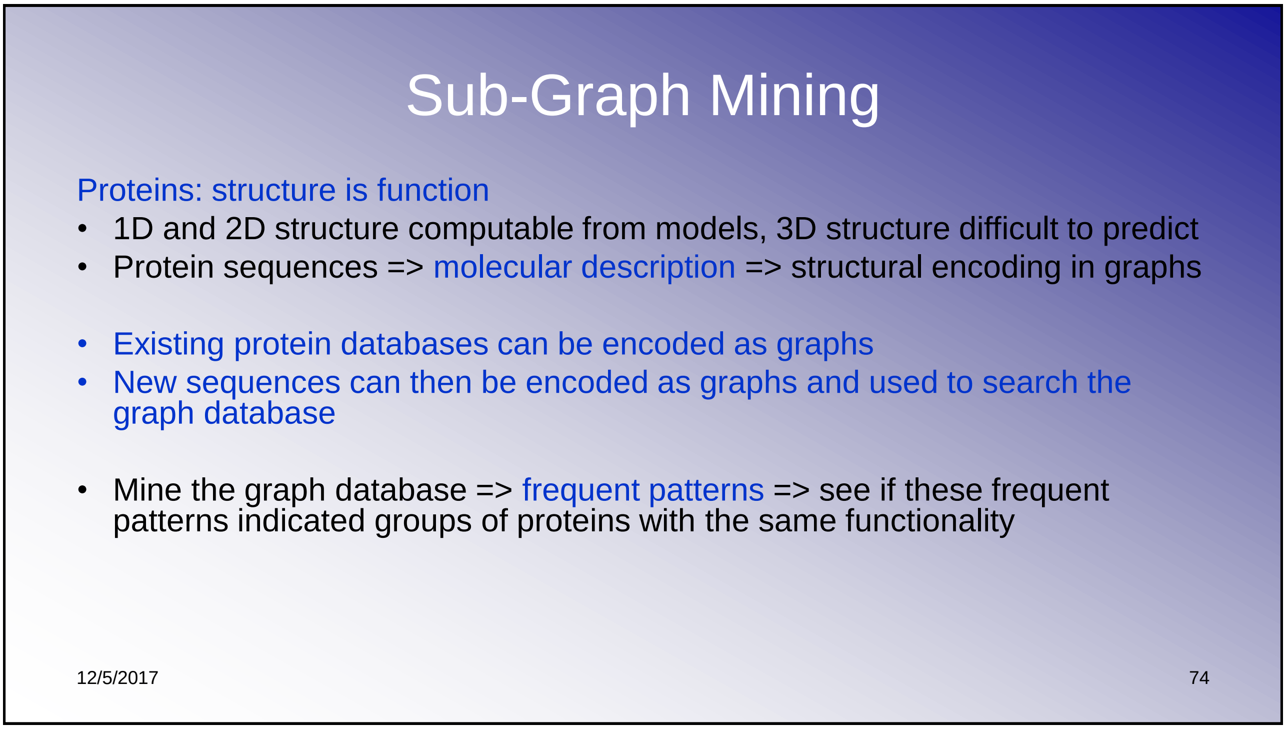


- TOP-SURF image descriptor for large scale image retrieval. By B. Thomee, EM Bakker, MS Lew.
  (Link: http://press.liacs.nl/researchdownloads/topsurf/)



Challenges Bio-image Searching

- Discover/develop enhanced measures for bio-image similarity
- For example, what features do scientists in biology and chemistry use to decide whether cells are similar? (Very challenging!)
- Sub-image search: develop multi-scale, sub-image search mechanisms for direct usage in the bio-imaging of the cell



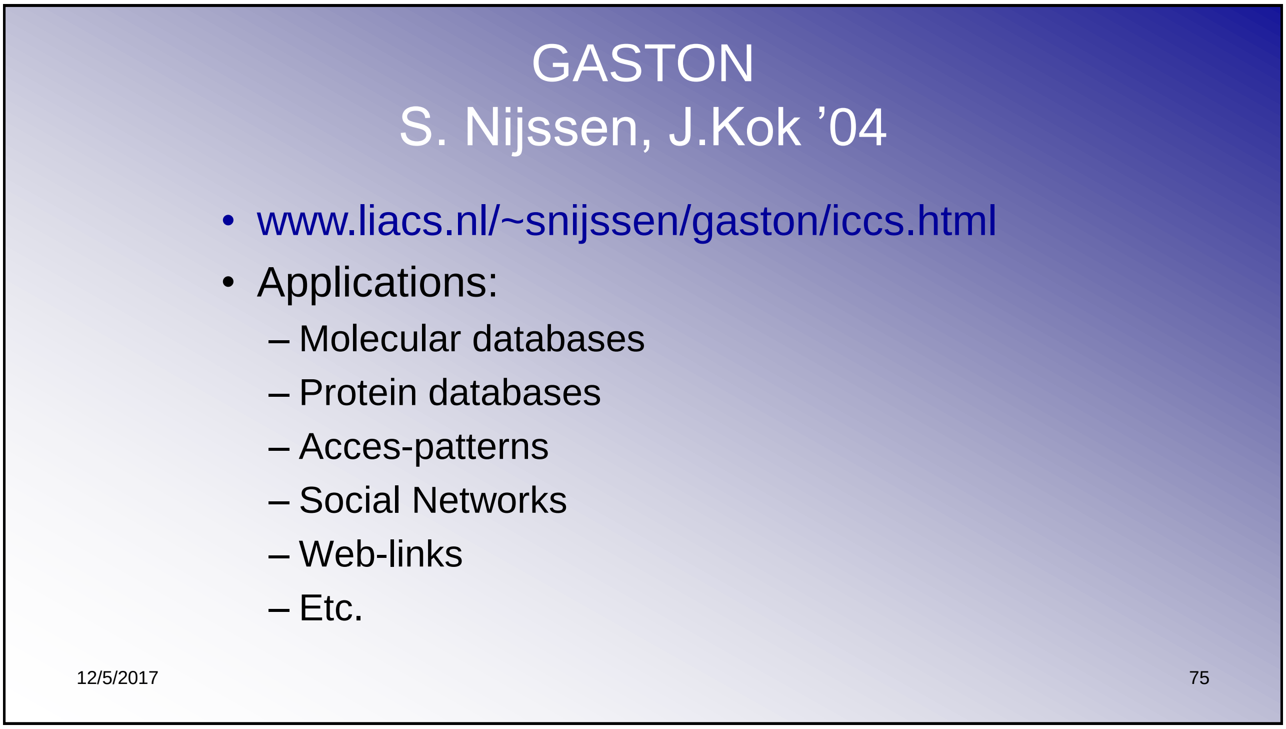
Sub-Graph Mining

Proteins: structure is function

- 1D and 2D structure computable from models, 3D structure difficult to predict
- Protein sequences => molecular description => structural encoding in graphs

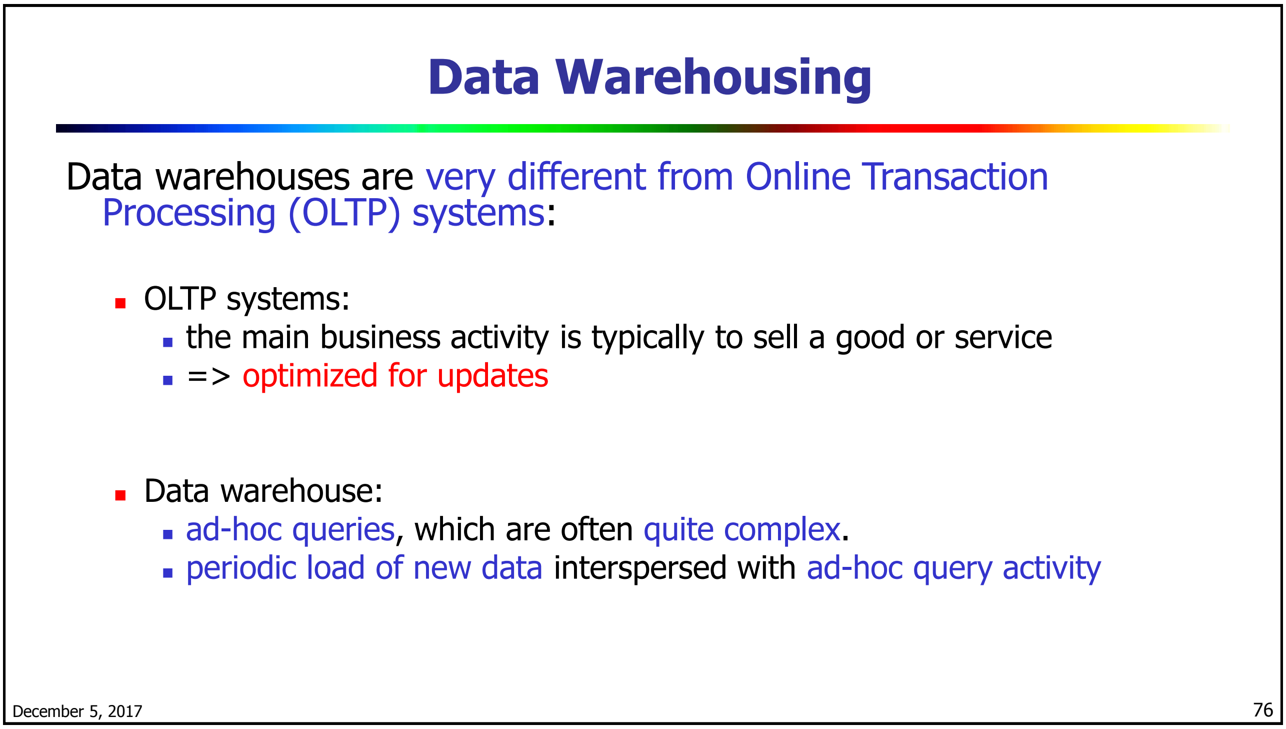
- Existing protein databases can be encoded as graphs
- New sequences can then be encoded as graphs and used to search the graph database

- Mine the graph database => frequent patterns => see if these frequent patterns indicated groups of proteins with the same functionality



GASTON
S. Nijssen, J.Kok ’04

- www.liacs.nl/~snijssen/gaston/iccs.html
- Applications:
  - Molecular databases
  - Protein databases
  - Acces-patterns
  - Social Networks
  - Web-links
  - Etc.



Data Warehousing

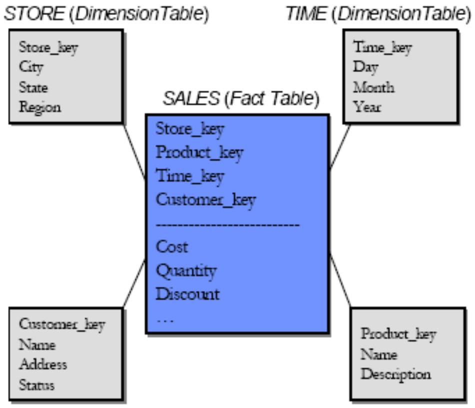
Data warehouses are very different from Online Transaction Processing (OLTP) systems:

- OLTP systems:
  - the main business activity is typically to sell a good or service
  - => optimized for updates
- Data warehouse:
  - ad-hoc queries, which are often quite complex.
  - periodic load of new data interspersed with ad-hoc query activity

Data Warehousing

The standard wisdom in data warehouse schemas is to create a fact table:

"who, what, when, where" about each operational transaction.



Data Warehousing

- Data warehouse applications run much better using bit-map indexes
- OLTP (Online Transaction Processing) applications prefer B-tree indexes.
- materialized views are a useful optimization tactic in data warehousing, but not in OLTP worlds.

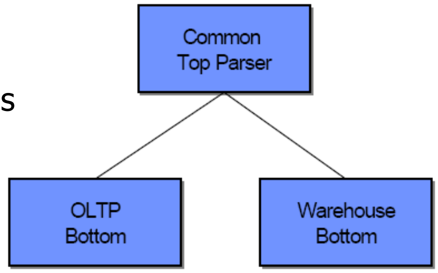
Data Warehousing

As a first approximation, most vendors have a

- warehouse DBMS (bit-map indexes, materialized views, star schemas and optimizer tactics for star schema queries) and

- OLTP DBMS (B-tree indexes and a standard cost-based optimizer), which are united by a common parser

		Bitmaps	
Index	Gender	F	M
1	Female	1	0
2	Female	1	0
3	Unspecified	0	0
4	Male	0	1
5	Male	0	1
6	Female	1	0



Example: An existing application: financial-feed processing

Detect Problems in Streaming stock ticks:

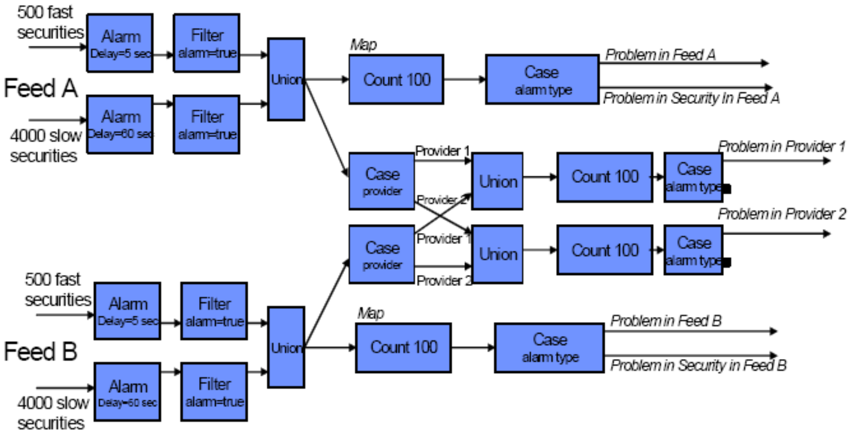
- Specifically, there are 4500 securities, 500 of which are "fast moving".

Defined by rules:

- A stock tick on one of the fast securities is late if it occurs more than 5 seconds after the previous tick from the same security.

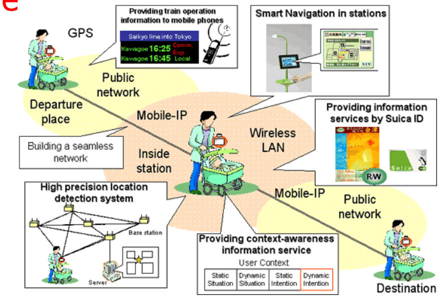
- The other 4000 symbols are slow moving, and a tick is late if 60 seconds have elapsed since the previous tick.

Stream Processing



Emerging Sensor Based Applications

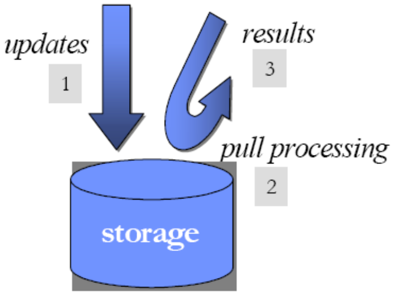
- Conventional DBMSs will not perform well on this new class of monitoring applications.
- For example: *Linear Road*, traditional solutions are nearly an order of magnitude slower than a special purpose **stream processing** engine



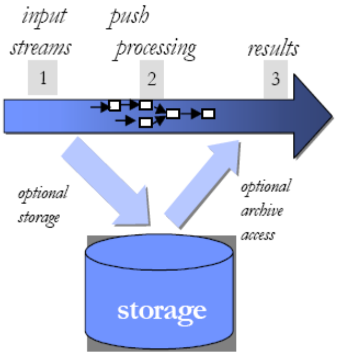
Performance

- Implemented in the StreamBase stream processing engine (SPE) [5], a commercial, industrial-strength version of Aurora [8, 13].
- On a 2.8Ghz Pentium processor with 512 Mbytes of memory and a single SCSI disk, the workflow in the previous figure can be executed at 160,000 messages per second, before CPU saturation is observed.
- In contrast, StreamBase engineers could only get 900 messages per second using a popular commercial relational DBMS.

Why?: Outbound vs Inbound Processing

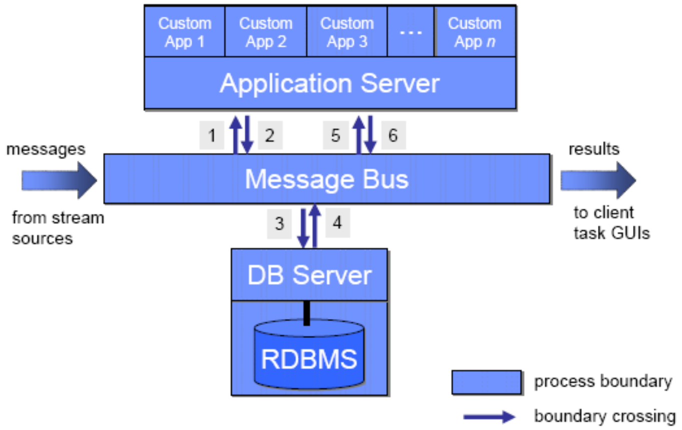


RDBMS
(Outbound Processing)



StreamBase
(Inbound Processing)

Inbound Processing



MapReduce

The problem:

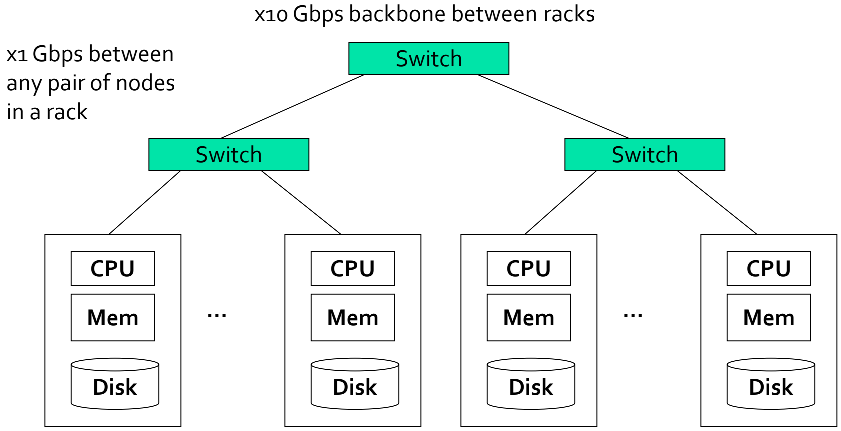
- 40+ billion web pages x 20KB ~ 1 petaB
- 1 computer reads 60-70 MB/sec from disk
  - ~4 months to read these pages
- ~1,000 hard drives to store 'the web'
- How can we do something useful with such amounts of data?

MapReduce

- Addresses distribution of computation
- Google's computational/data manipulation model

Slides adapted from: http://www.mmds.org

Cluster Architecture


x10 Gbps backbone between racks
x1 Gbps between any pair of nodes in a rack
Switch
Switch
Switch
CPU
Mem
Disk
…
CPU
Mem
Disk
CPU
Mem
Disk
…
CPU
Mem
Disk

Each rack contains 16-64 nodes

In 2011 it was guestimated that Google had 1M machines, http://bit.ly/Shh0RO
=> ~10sec to read 'the web'

Large-scale Computing

- **Large-scale computing** for **data mining** problems on **commodity hardware**
- **Challenges:**
  - **How can we make it easy to write distributed programs?**
    - One server may stay up 3 years (1,000 days)
    - If you have 1,000 servers, expect to loose 1/day
    - People estimated Google had ~1M machines in 2011
      - 1,000 machines fail every day!

Idea and Solution

- **Issue: Copying data over a network takes time**
- **Idea:**
  - **Bring computation close to the data**
  - Store files multiple times for reliability
- **Map-reduce** addresses these problems
  - Google’s computational/data manipulation model
  - Elegant way to work with big data
  - **Storage Infrastructure – File system**
    - Google: GFS. Hadoop: HDFS
  - **Programming model**
    - Map-Reduce

Storage Infrastructure

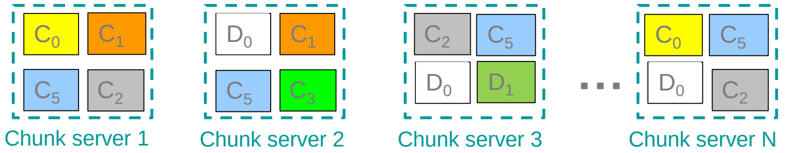
-
  - If nodes fail, how to store data persistently?
- **Answer:**
  - **Distributed File System:**
    - Provides global file namespace
    - Google GFS; Hadoop HDFS;
- **Typical data usage:**
  - Huge files (100s of GB to TB)
  - Data is rarely updated in place
  - Reads and appends are common

Distributed File System

-
  - File is split into contiguous chunks
  - Typically each chunk is 16-64MB
  - Each chunk replicated (usually 2x or 3x)
  - Try to keep replicas in different racks
- **Master node**
  - a.k.a. Name Node in Hadoop's HDFS
  - Stores metadata about where files are stored
  - Might be replicated
- **Client library for file access**
  - Talks to master to find chunk servers
  - Connects directly to chunk servers to access data

Distributed File System

- **Reliable distributed file system**
- Data kept in "chunks" spread across machines
- Each chunk on different machines
  - Seamless recovery from disk or machine failure



Bring computation directly to the data!

Chunk servers also serve as compute servers

Programming Model: MapReduce

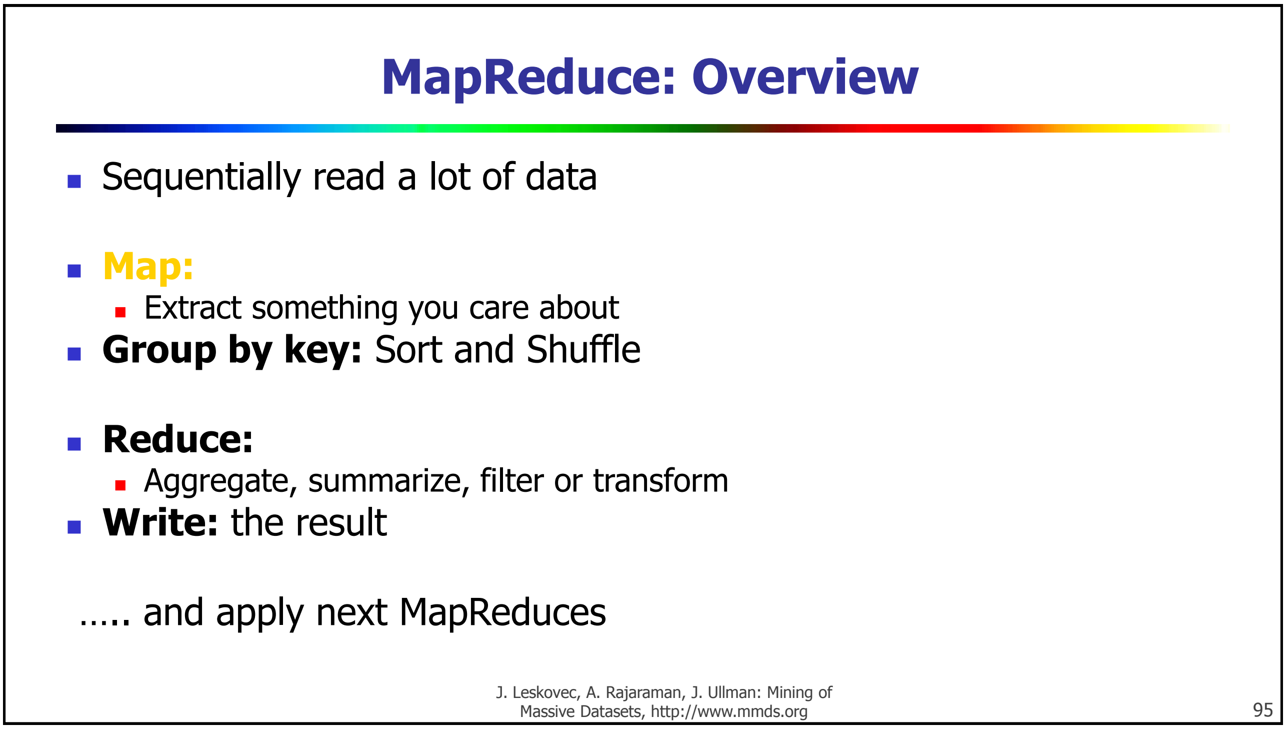
Example:

- We have a huge text document
- Count the number of times each distinct word appears in the file
- **Sample application:**
  - Analyze web server logs to find popular URLs

Task: Word Count

Case 1:
  - File too large for memory, but all <word, count> pairs fit in memory

Case 2:
- Count occurrences of words:
  - `words(doc.txt) | sort | uniq -c`
    - where `words` takes a file and outputs the words in it, one per a line
- Case 2 captures the essence of **MapReduce**
  - Great thing is that it is naturally parallelizable



MapReduce: Overview

- Sequentially read a lot of data
- **Map:**
  - Extract something you care about
- **Group by key:** Sort and Shuffle
- **Reduce:**
  - Aggregate, summarize, filter or transform
- **Write:** the result

….. and apply next MapReduces

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org