Creation of integrated data mining environment linking patient data to clinical, cellular and molecular information

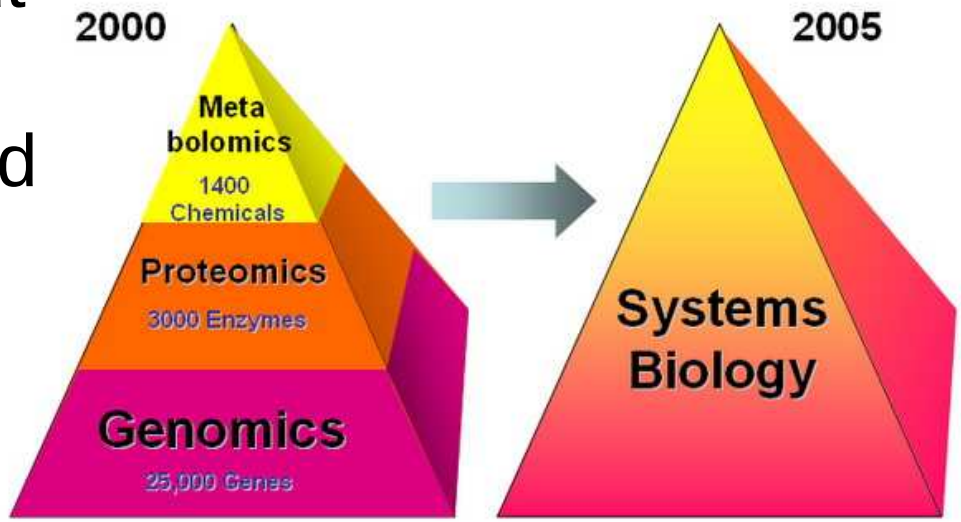
Juha Kononen

Problem in Biology and Medicine: Data overload

- too much
- too many sources
- hard to compare
- few tools to model
- constantly changing methodology

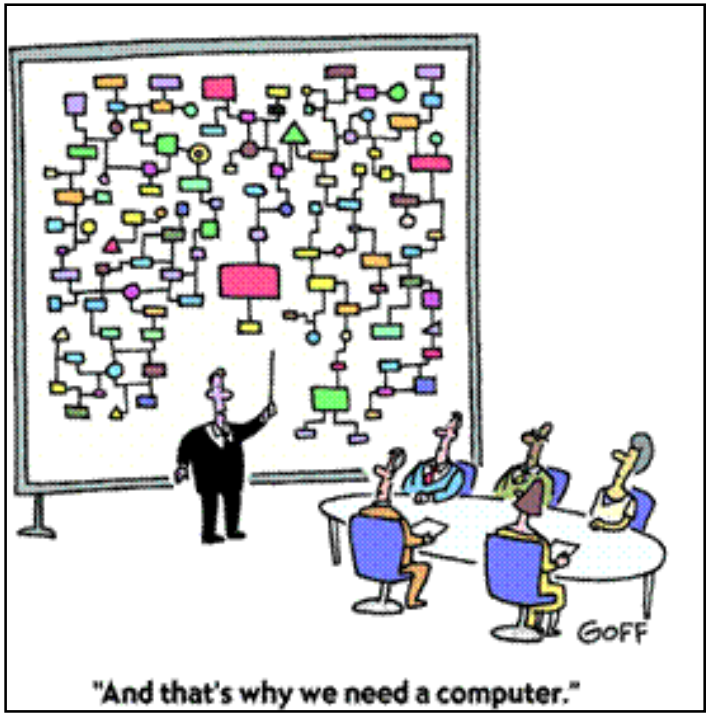
Merging of ‘omics - analyzing individual genes and proteins from few samples is no longer enough

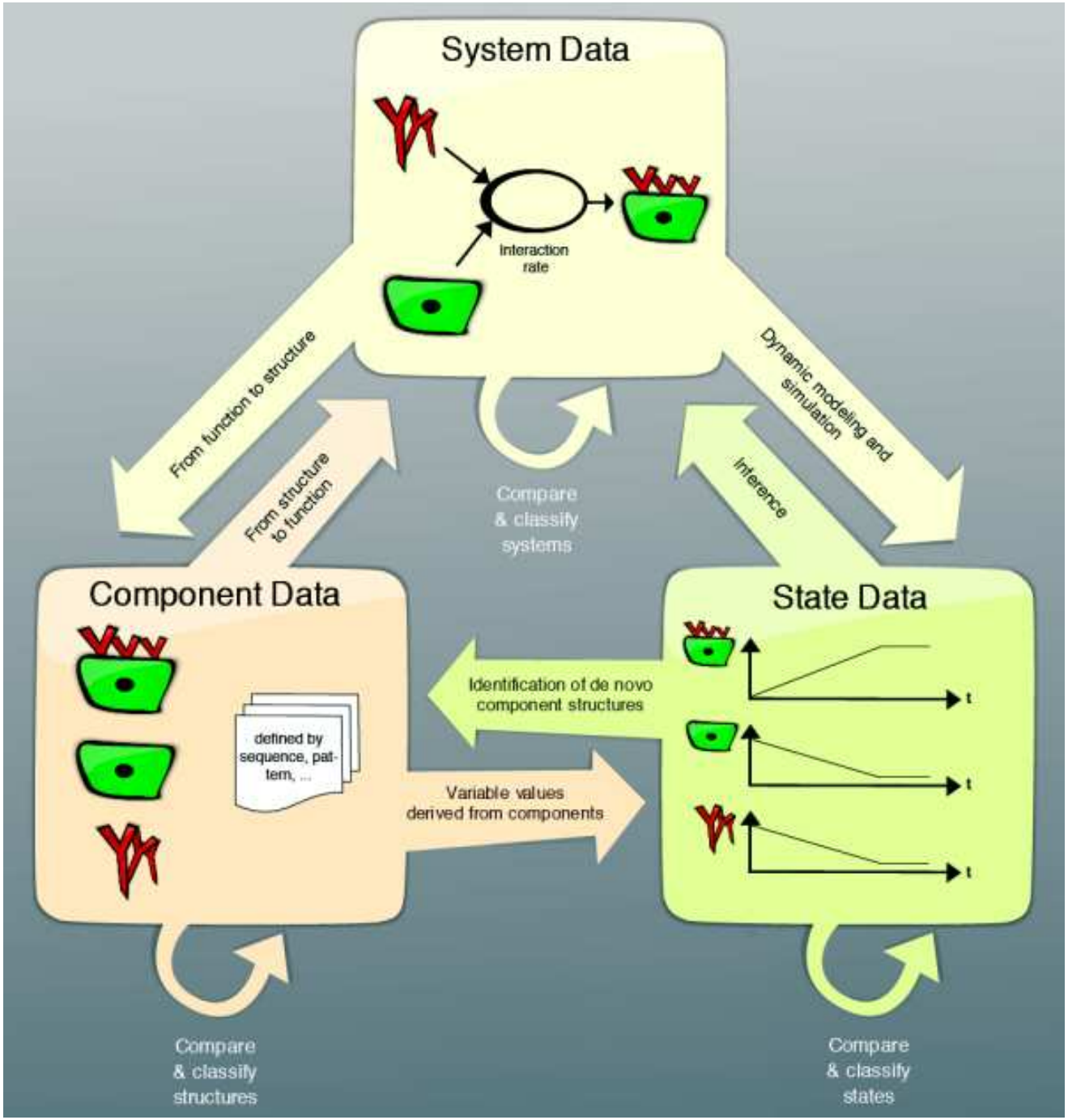
- New high-throughput methodologies constantly introduced
- Abundance of molecular targets
- Pathways and networks emerging



Data Management Challenge

- Experimental results are accumulating faster than ever
- Data types different and poorly defined, format conversions time-consuming
- How to manipulate data without extensive programming skills?
- How to best visualize and explore different data types?

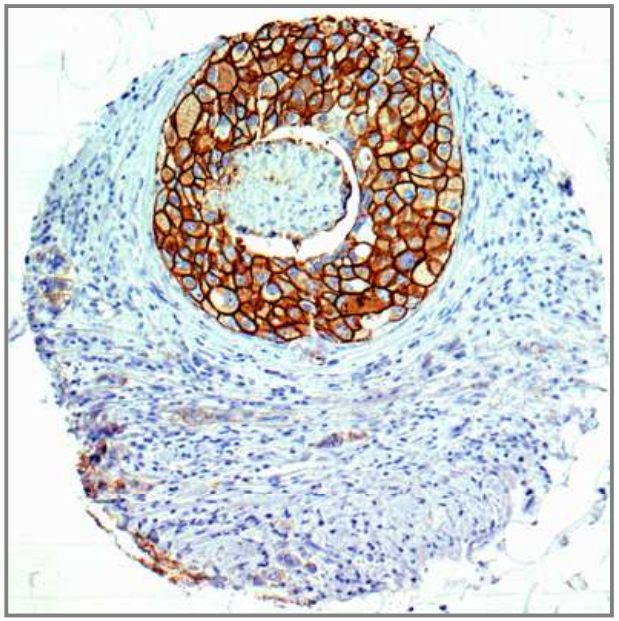




- Most of the biological databases contain component data (protein, gene, SNP, genome databases).
- State data is now accumulating (expression microarrays).
- Only few databases provide information of interactions between these objects (i.e. system data).
- However, ***all the different data domains must be exploited*** to be able to understand biology on a system level.

The goal is to create predictive models (system) from data

- Choose therapy type
- Predict responses
- Monitor responses



Develop better medical care

- individualized treatments

- preventive medicine

How to get there?

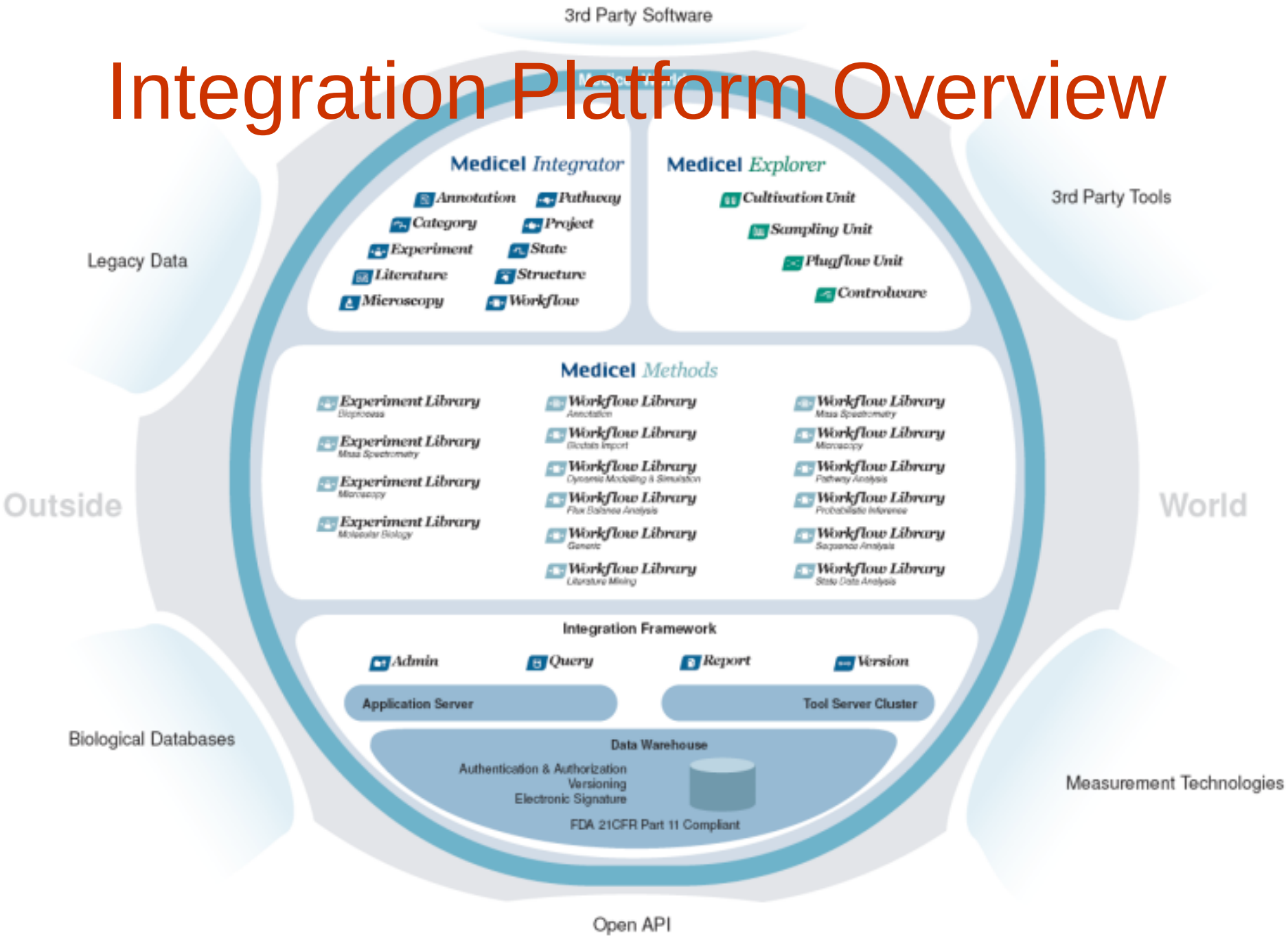
- Define flexible and comprehensive data model
- Create universal data manipulation language
- Automate laboratory
- Automate data management
- Create user interfaces for scientists

Integration platform

- Example: Medicel Integrator
  - Data warehouse built on unified information model capable of integrating legacy biomedical data from external databases and in-house research applications.

- Application-specific software running on top of the integration framework to manage experiment planning, monitor research workflows, interface with research instruments, execute in-silico computations based on data, simulate pathways, analyze images and annotate research literature.

Integration Platform Overview



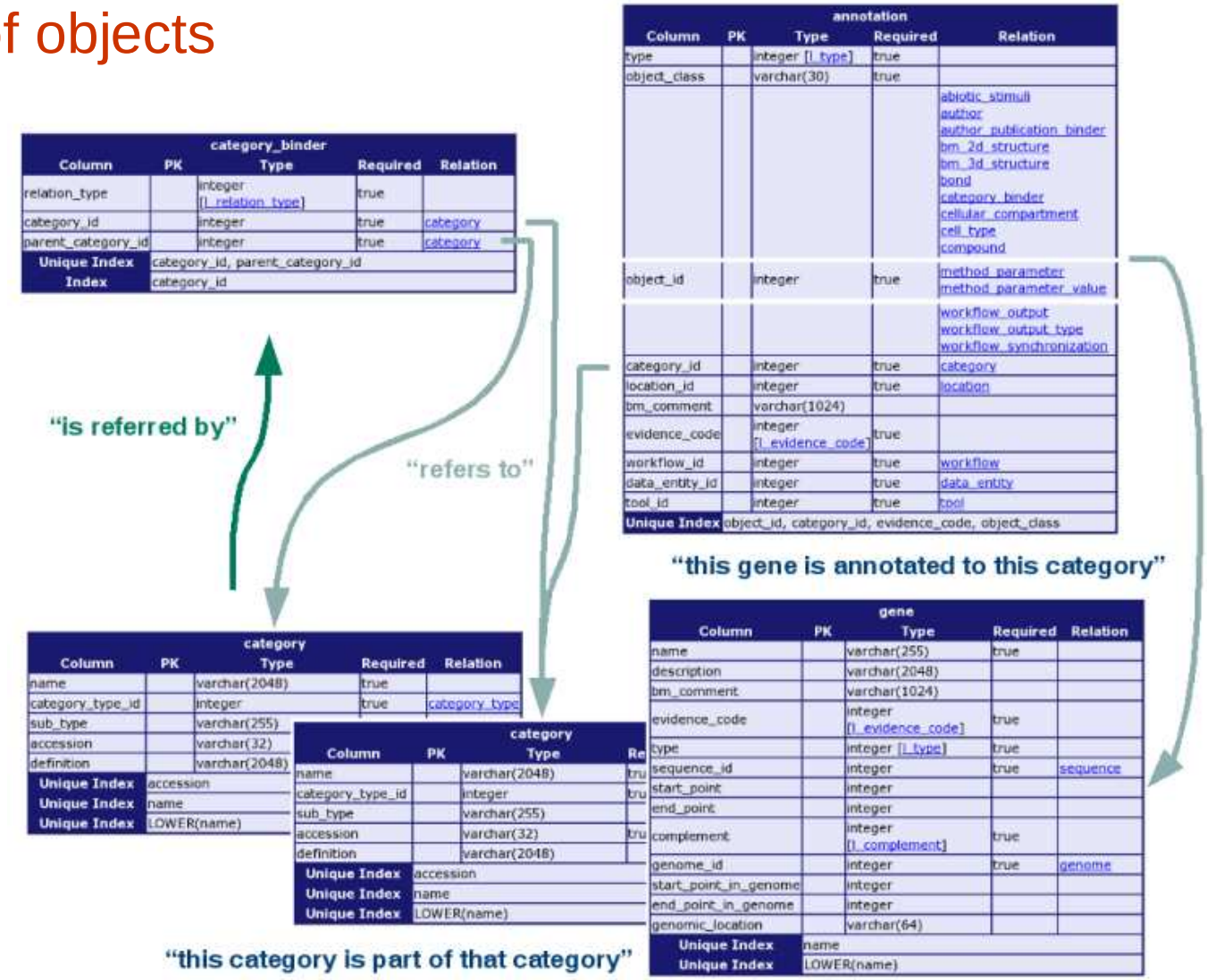
Unified Information Model – an optimized relational database with minimum number of tables to describe a wide variety of objects

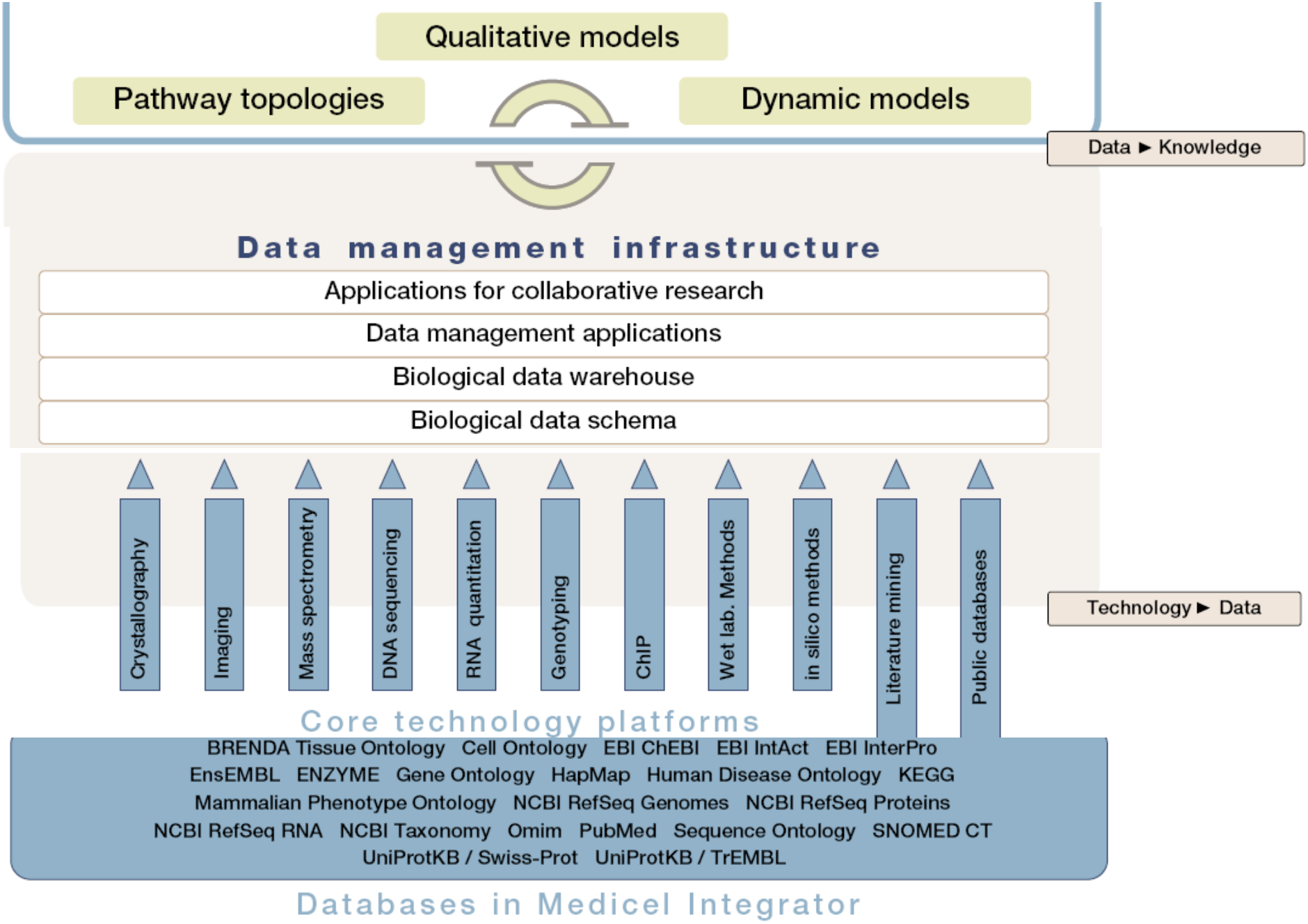
Data and meta-data is organised into tables and fields. Fields can point to other tables allowing the user to access data from different tables.

Abstraction is required to minimize number of tables needed.

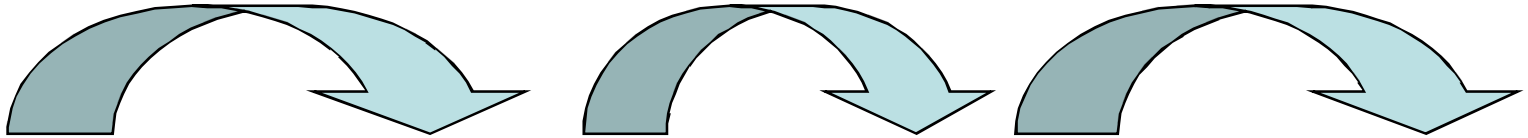
- A gene can be annotated to a category.
- Category can be defined as being part of another category.

- Farnesylation is a lipid modification of an aminoacid…
- …which is a post-translational modification of protein sequences.
- …and all three are categories of sequence modifications.




Qualitative models
Pathway topologies
Dynamic models
Data ► Knowledge
Data management infrastructure
Applications for collaborative research
Data management applications
Biological data warehouse
Biological data schema
Crystallography
Imaging
Mass spectrometry
DNA sequencing
RNA quantitation
Genotyping
ChIP
Wet lab. Methods
in silico methods
Literature mining
Public databases
Technology ► Data
Core technology platforms
BRENDA Tissue Ontology Cell Ontology EBI ChEBI EBI IntAct EBI InterPro
EnsEMBL ENZYME Gene Ontology HapMap Human Disease Ontology KEGG
Mammalian Phenotype Ontology NCBI RefSeq Genomes NCBI RefSeq Proteins
NCBI RefSeq RNA NCBI Taxonomy Omim PubMed Sequence Ontology SNOMED CT
UniProtKB / Swiss-Prot UniProtKB / TrEMBL
Databases in Medicel Integrator

Process management advantage: all research activities are accessible from one platform



Planning

- Literature
- DB Query

Wet-lab

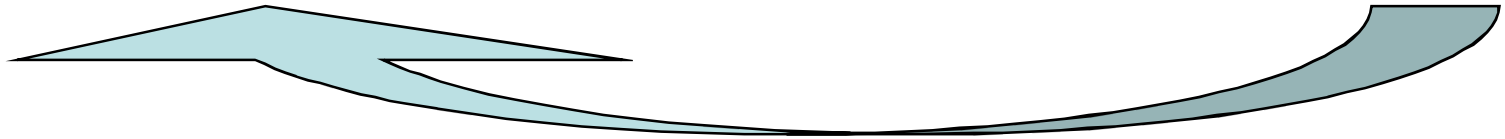
- Experiment management
- Interface to equipment
- Data entry

Analysis

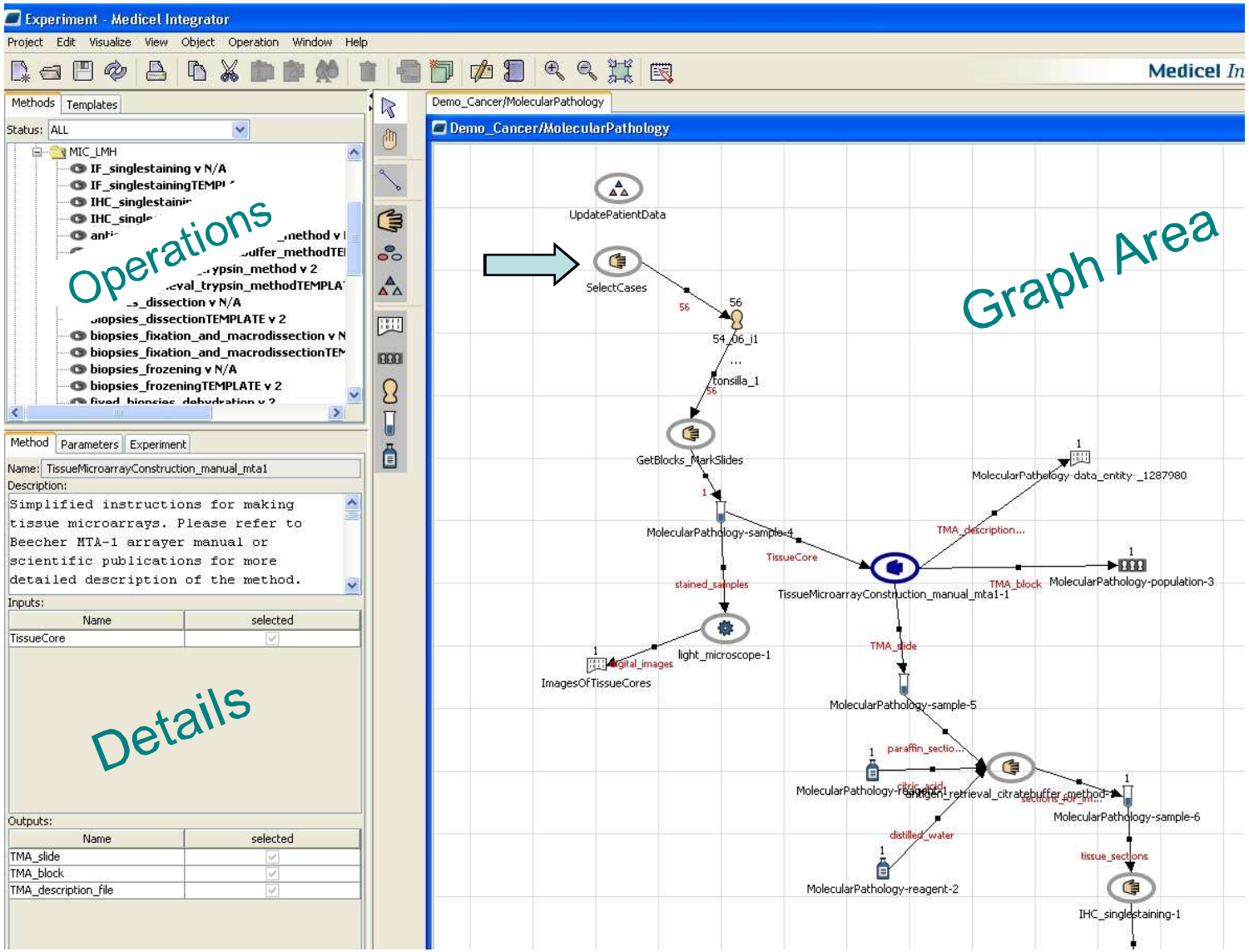
- In-silico workflow calculations
- External apps
- Data viewers

Review

- Visualization
- Annotation
- Pathway modeling



Virtual notebook use: experimental work easily visualized, planned, communicated and manipulated.

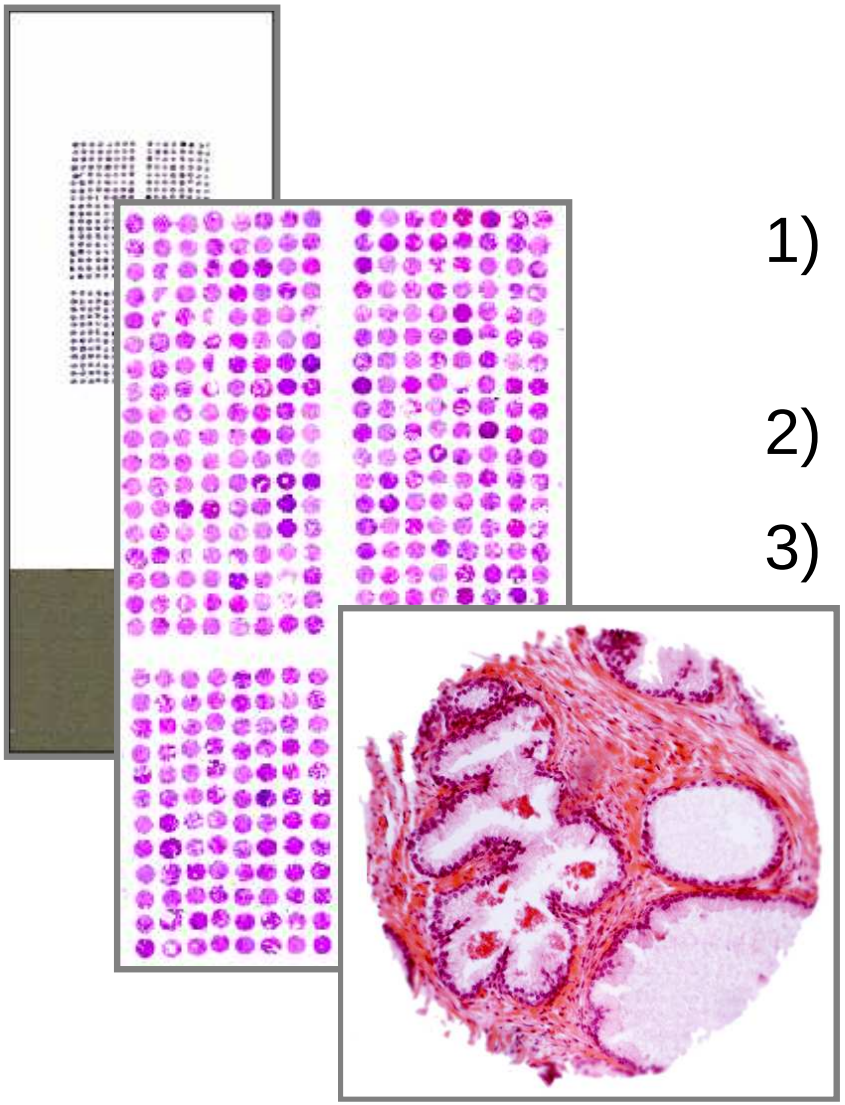


Pathology archives are treasure troves of clinical data

- Normal structures
- Early lesions
- Disease progression
- Disease variations
- Treatment responses
- Morphology
- Protein expression and structure

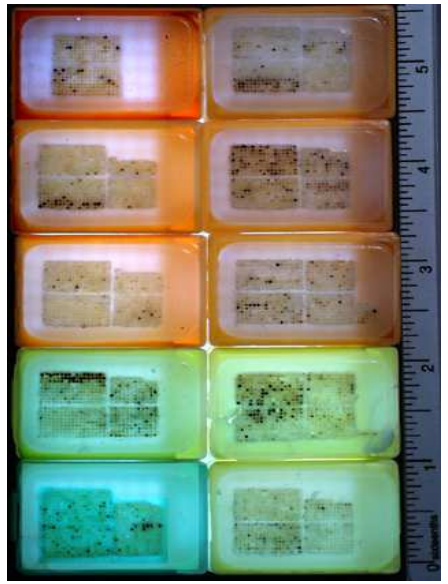
Pathology archive 1999…

Tissue Microarrays (TMAs)



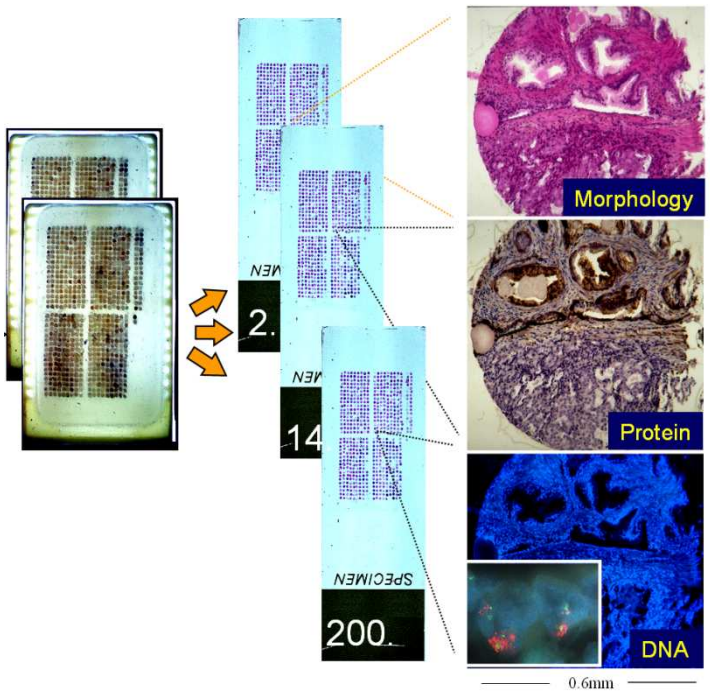
Platform to

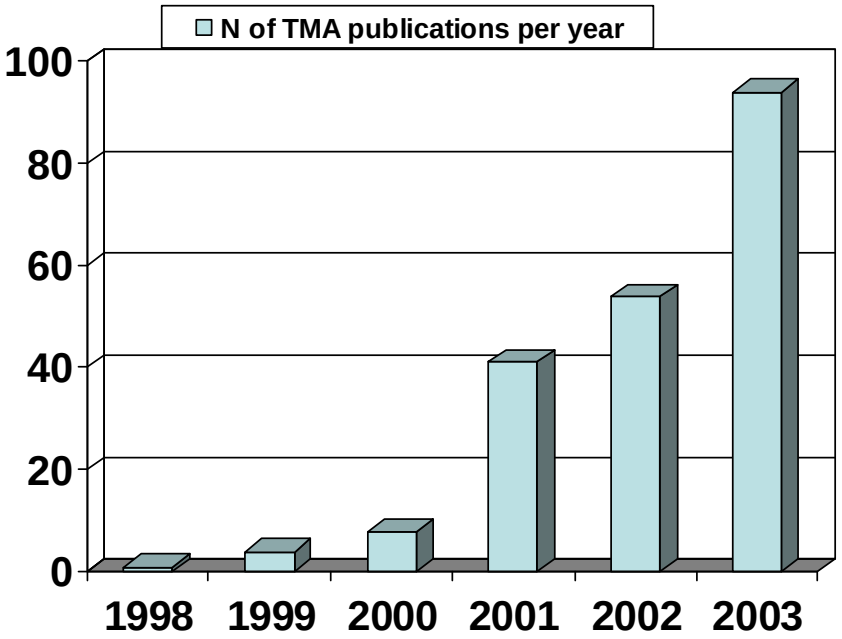
1) Manage tissue specimens
2) Collect data
3) Test new approaches



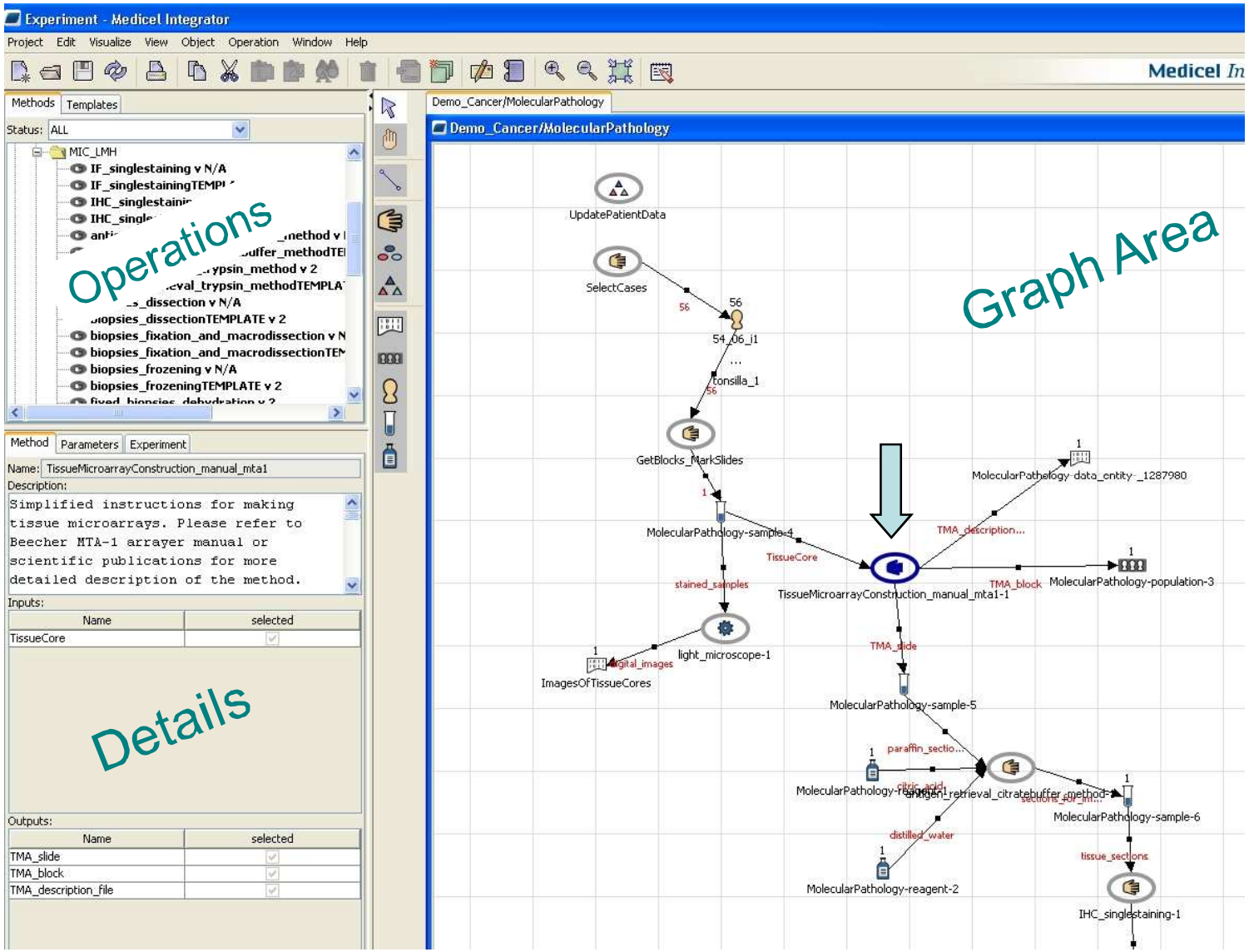
… TMA-based pathology archive today.

Tissue microarrays are a powerful platform for deriving new state data: linking genes to diseases

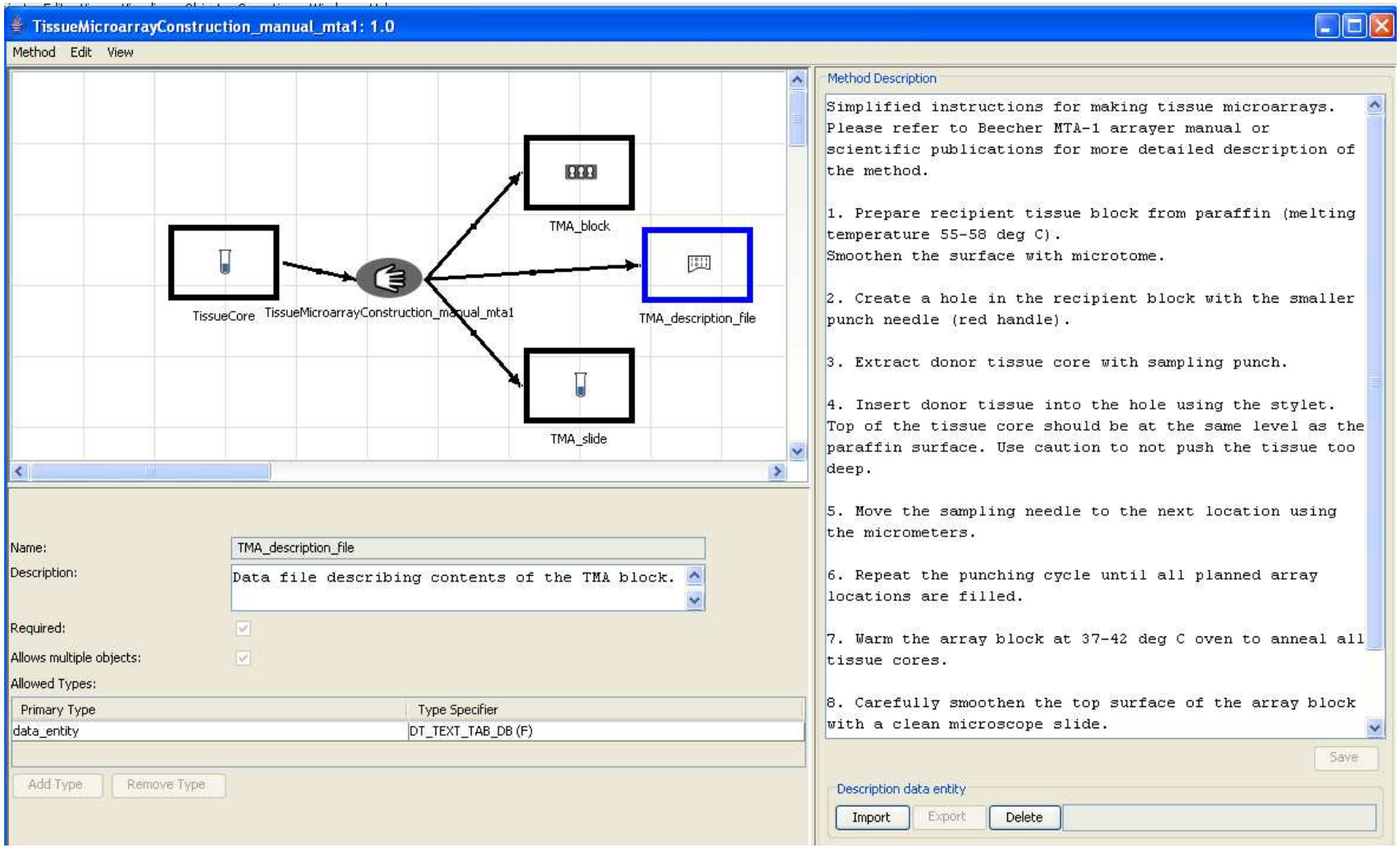




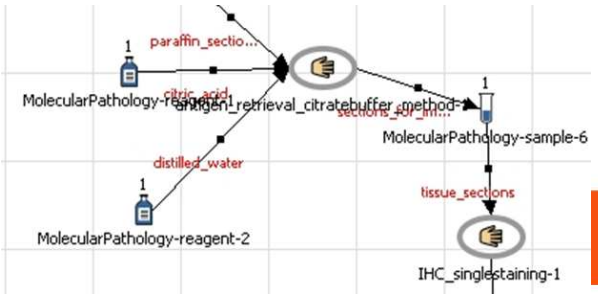
Virtual notebook use: experimental work easily visualized, planned, communicated and manipulated.

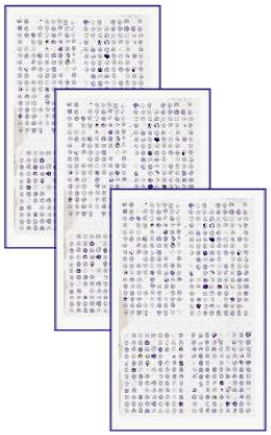


Operational benefit: Standard procedures, data formats, detailed instructions and references available for wet lab methods



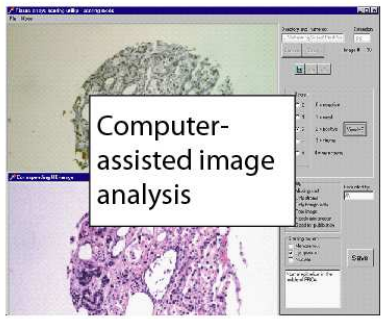
Bottleneck in TMA workflow: stained slides accumulate faster than can be manually analyzed





1 TMA -> 500 spots
20 TMAs -> 10,000 spots to score

1 array block -> 150 TMA slides
150 TMA slides -> 75,000 spots



- Image acquisition setup
- Region selection
- Manual measurements
- Import/export data

- Navigating the slide
- Setting scoring criteria
- Recording scores
- Manual data entry

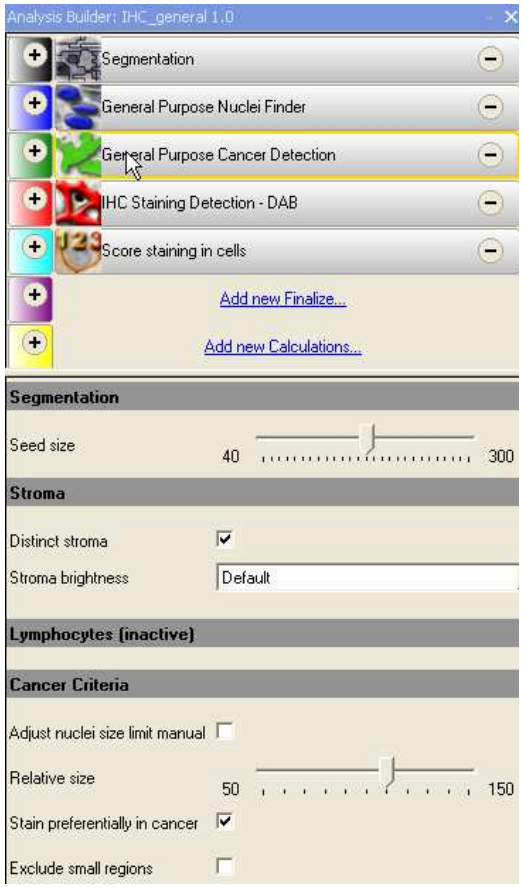
Image Data Domain

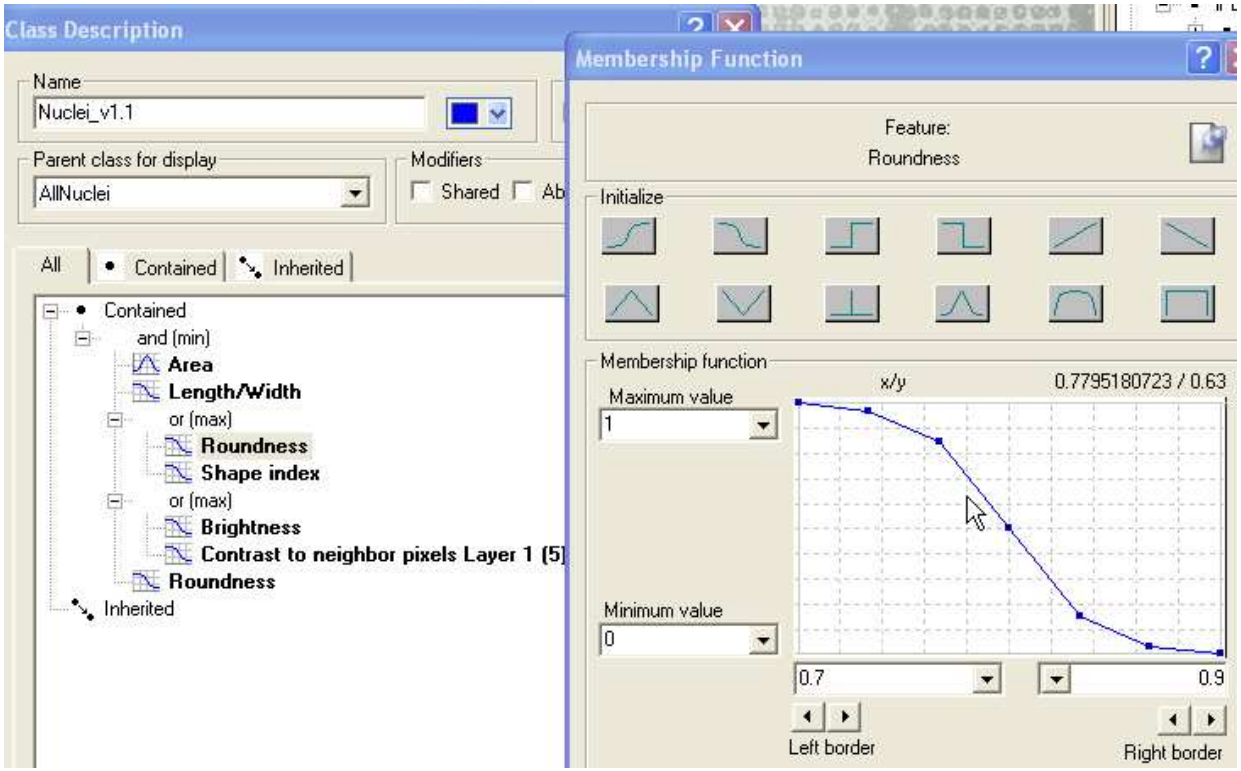
- Often the weakest link in the chain of collecting data linked to tissue samples
  - Specific, complex, unlinked and poorly documented image analysis algorithms exist
  - Image analysis too slow for clinical use
  - Image analysis too specific to easily adapt for different research use scenarios
- Unexplored
  - New (old) biomarkers to be found?
  - Could yield novel sources to mine data

Setting up TMA analysis

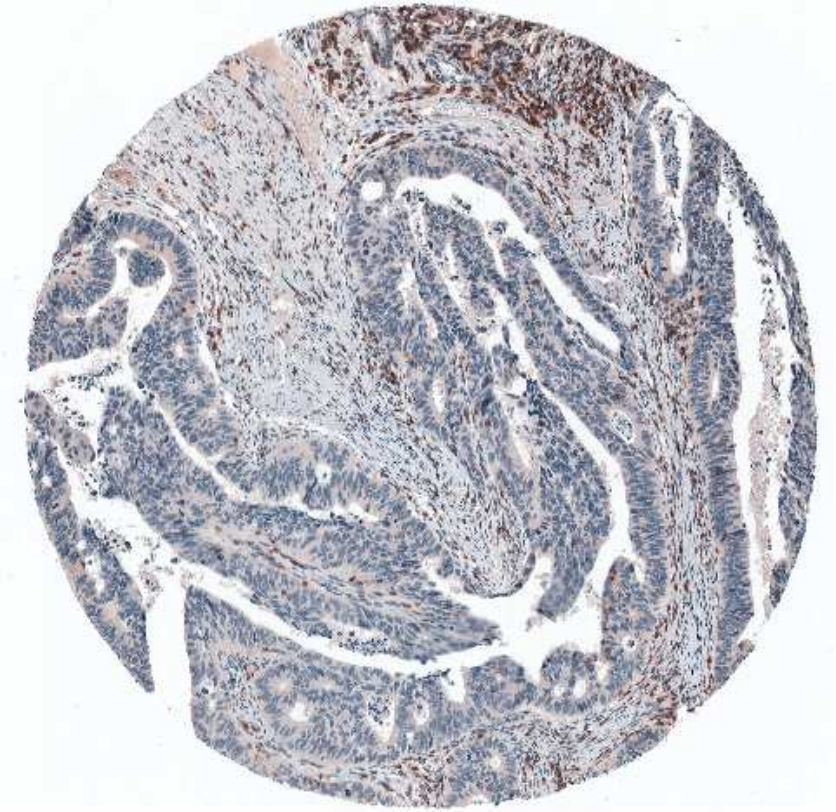
- Multiple analysis operations run for each TMA core image
- Layered analysis approach: tasks requiring different segmentation scales are done at specific levels.
  - Level 1: Determining and scoring IHC staining
  - Level 2: Nuclei segmentation
  - Level 3: Identifying cancer vs stromal regions
- Final output result combines data from all analysis levels.
  - Calculating number of cells in tumor region and stromal region
  - Assembling IHC staining scores for tumor and stroma
  - Determining cellular localization of IHC staining

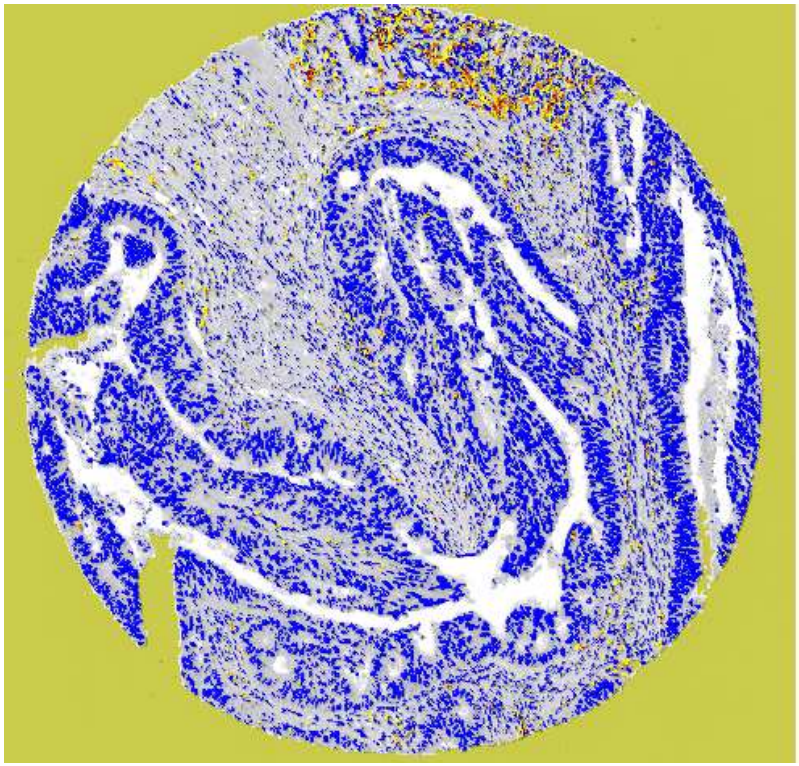

Microscope Image
Microscopy
Manual Image Modification, Visualization
Workflow
In Silico Analyses
Compound
Protein
Cellular Compartment
State
Visualization

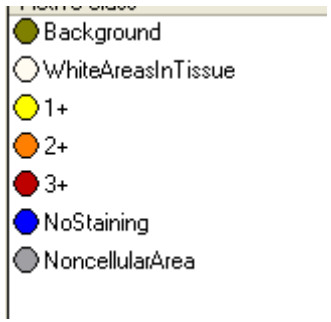

Analysis Builder: IHC_general 1.0
Segmentation
General Purpose Nuclei Finder
General Purpose Cancer Detection
IHC Staining Detection - DAB
Score staining in cells
Add new Finalize...
Add new Calculations...
Segmentation
Seed size
40
300
Stroma
Distinct stroma
Stroma brightness
Default
Lymphocytes (inactive)
Cancer Criteria
Adjust nuclei size limit manual
Relative size
50
150
Stain preferentially in cancer
Exclude small regions


Class Description
Name
Nuclei_v1.1
Parent class for display
AllNuclei
Modifiers
Shared
All
Contained
Inherited
Contained
and (min)
Area
Length/Width
or (max)
Roundness
Shape index
or (max)
Brightness
Contrast to neighbor pixels Layer 1 (5)
Roundness
Inherited
Membership Function
Feature:
Roundness
Initialize
Membership function
x/y
0.7795180723 / 0.63
Maximum value
1
Minimum value
0
0.7
0.9
Left border
Right border

Level 1: Staining detection and intensity evaluation





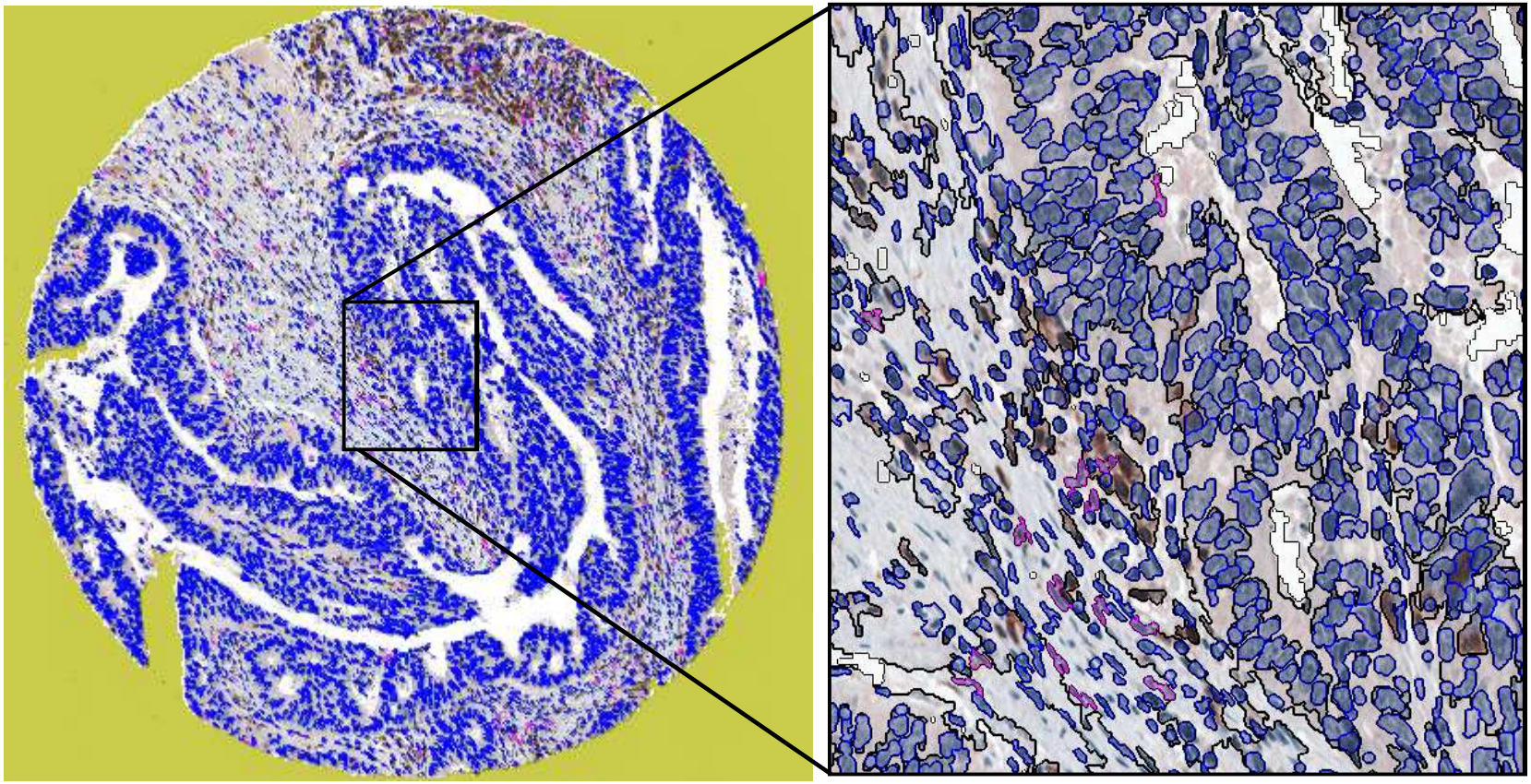


1+ = Weak IHC staining

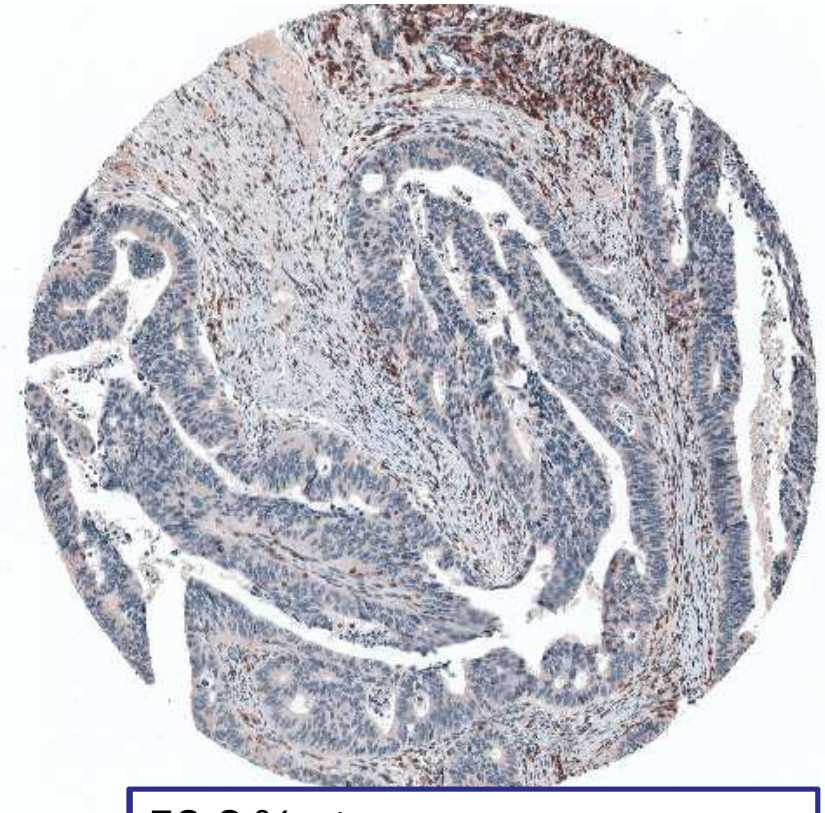
2+ = Moderate IHC staining

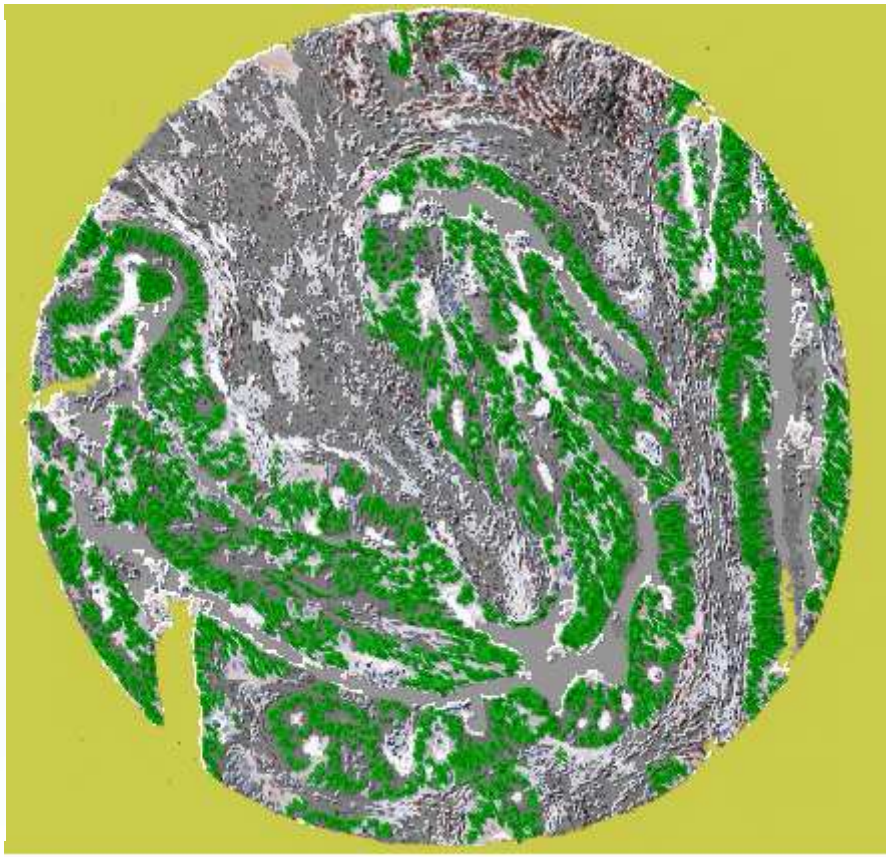
3+ = Strong IHC staining

Level 2: Cell nuclei segmentation

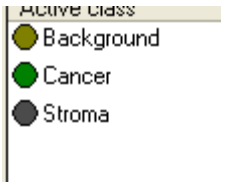


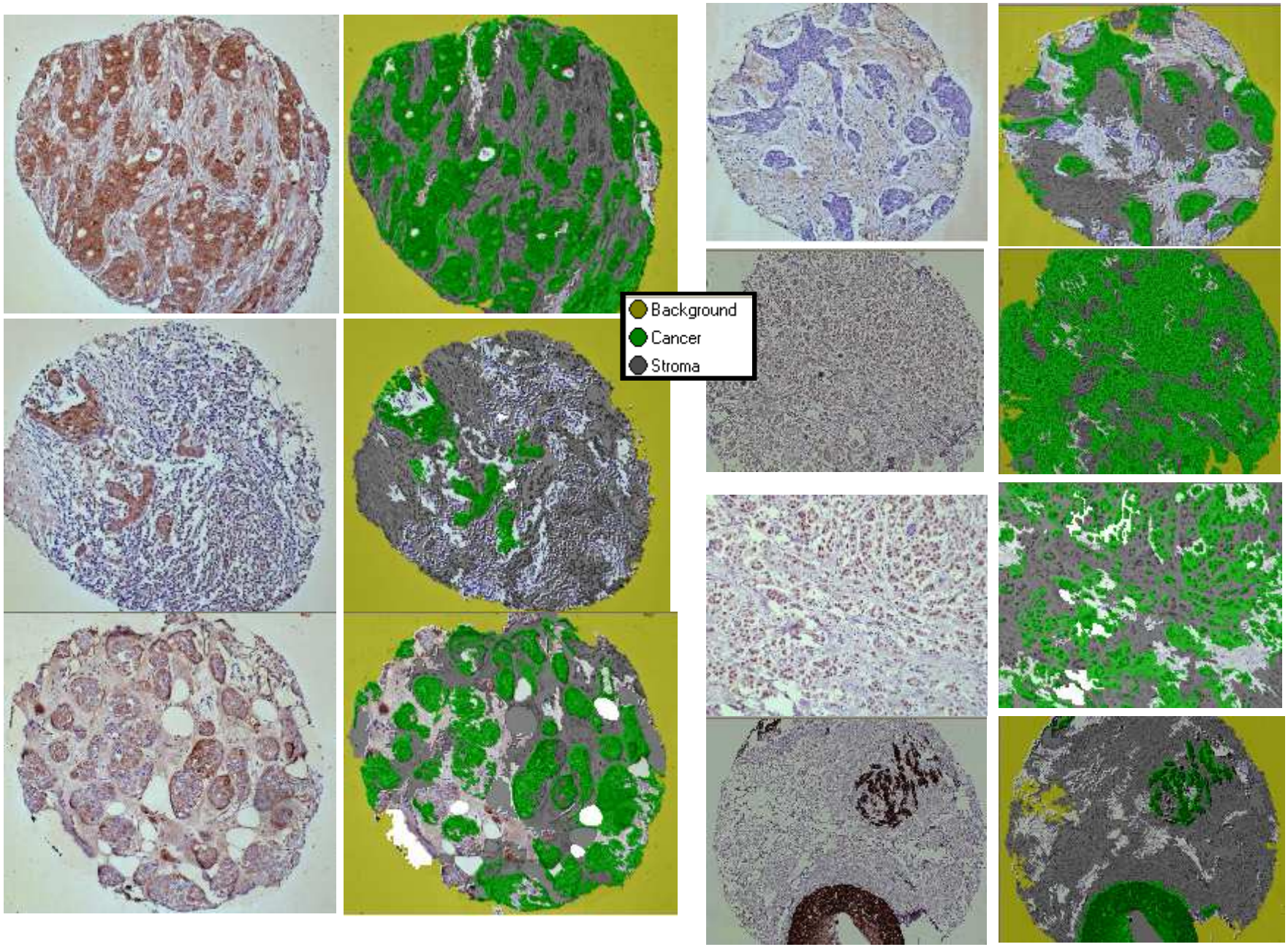
Level 3: Cancer vs stroma detection





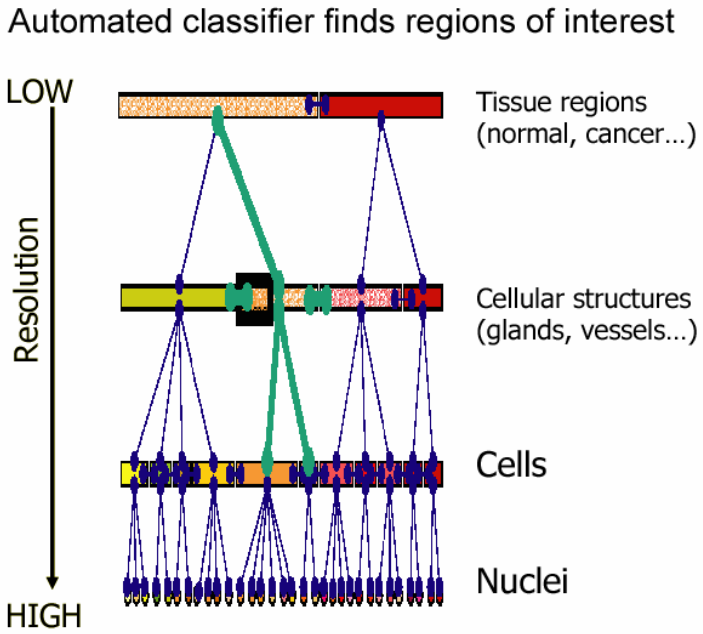
53.8 % stroma
10.9 % stained stroma
23.9 % stained cellular stroma
26.3 % cancer
2.7 % stained cancer




Background
Cancer
Stroma

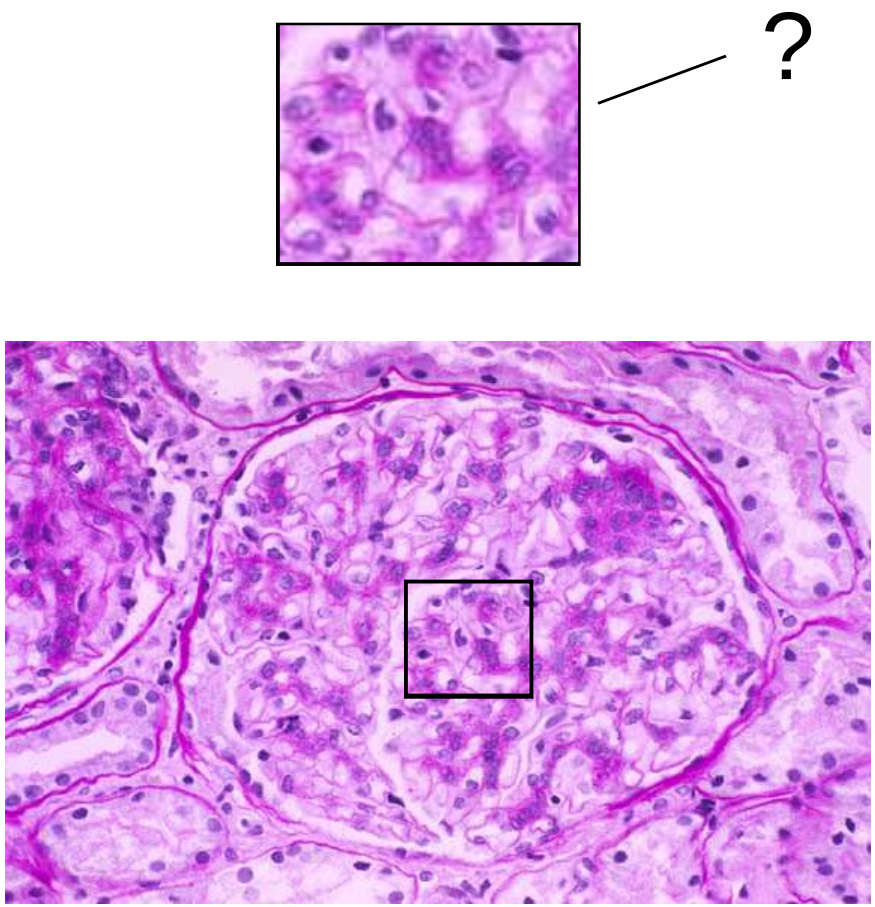
TMAx: Modeling tissue structures as a network of image objects

- The original pixel image is processed into a topological and semantical network of objects at multiple scales
- Hundreds of attributes (shape, color, texture, structure, relations) are available for fuzzy logic classification of each image object
- Object-based morphological operations and classification processes are alternating. As result of this iterative analysis, *objects of interest* are extracted in proper shape and proper labeling



Analysis is simultaneously performed at multiple scales

- Some objects need high resolution for detection, some can be best distinguished by overview and context analysis.

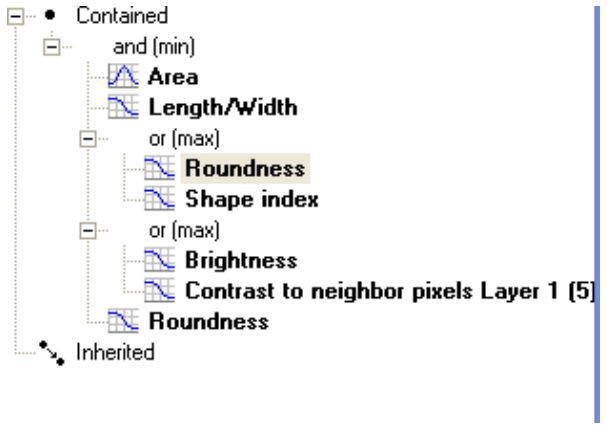


TMAx combines multiple classification criteria using fuzzy logic

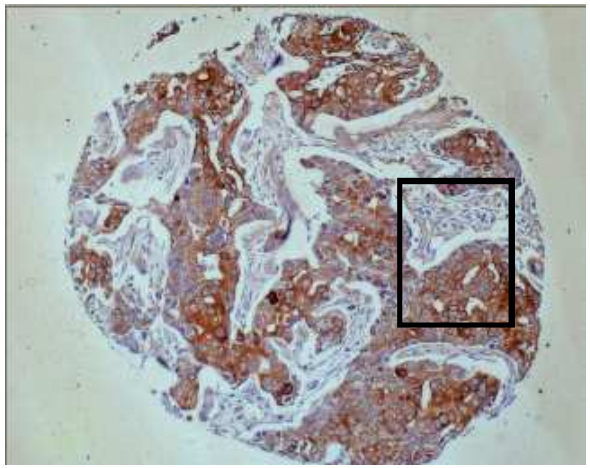
- **Instead of binomial black-or-white labeling, objects have a continuous classification value from 0…1 based on object features, such as brightness relative to neighbor objects**
- **Multiple features can be combined for each classifier**
- **Logical operations (And, Or, Min, Max) can be used to make a composite fuzzy classifier**
- **Logical operators can have a hierarchy**
- **Custom function curves can be applied for each component of the fuzzy classifier**

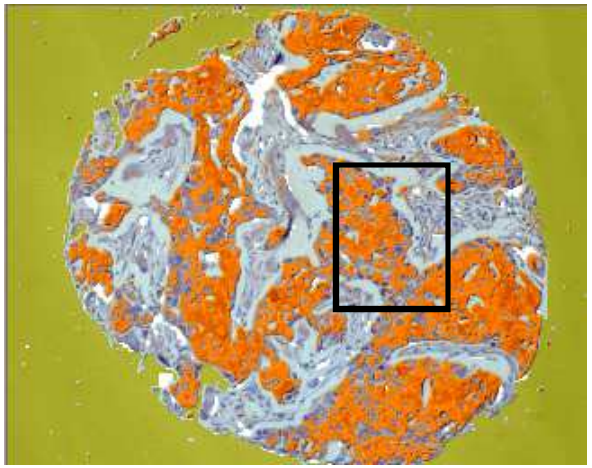
Plain language description of nuclei: "Nuclei are rather small, but not too small - there is a minimum size. They are round and have smooth borders. To be able to distinguish a nucleus it must have a certain contrast to the neighboring cytoplasm. Sometimes the nuclei in tumors are very pale and barely stain at all, sometimes they stain very intensely. They can be either blue if they are not stained, or brownish, almost black, if they are stained".

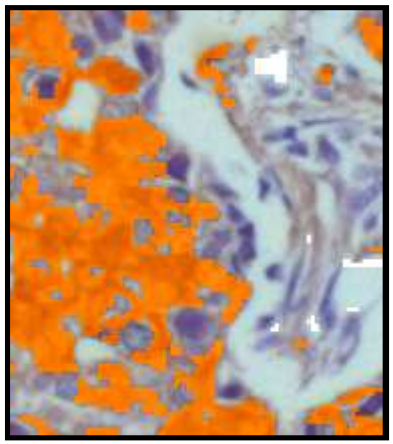
Translating the nuclei description to a fuzzy classifier :

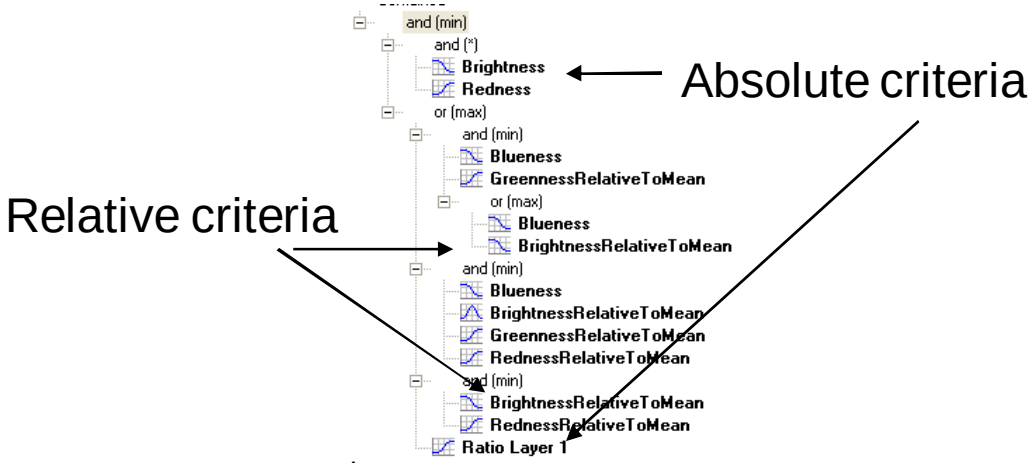


Fuzzy logic makes classifiers robust

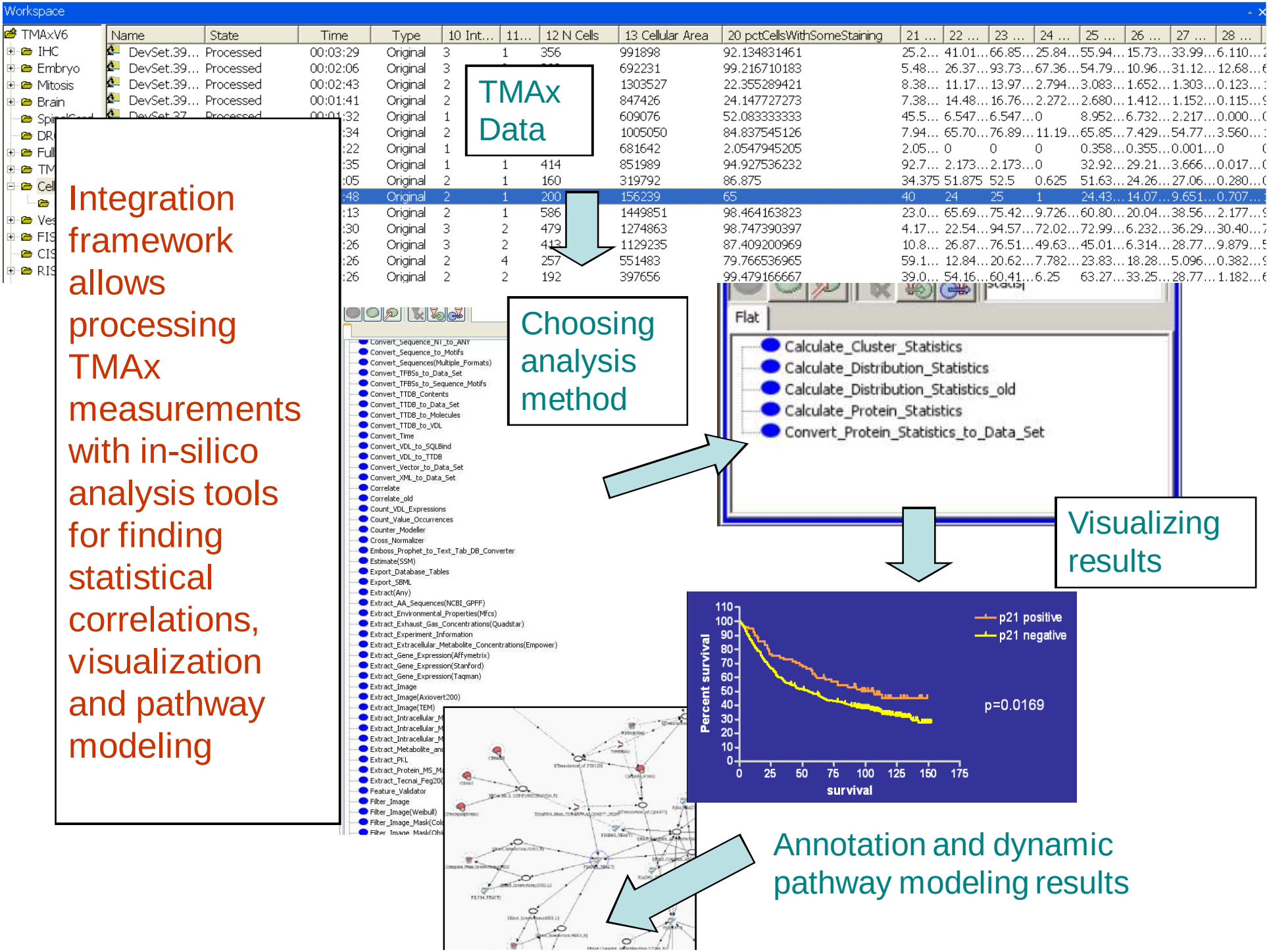








Integration framework allows processing TMAx measurements with in-silico analysis tools for finding statistical correlations, visualization and pathway modeling



Variable Description Language (VDL)

- Universal and exact way of defining data
- Import and export data between laboratories
- Import data from existing databases

Keywords make a limited vocabulary containing possible annotations for a data value.

Keyword	Name	Table	Column	Type	Example
V	Variable	variable	name	string	V[concentration]
U	Unit	Unit	Name	String	U[mol/L]
O	Organism	Organism	Latin_name	String	O[Homo sapiens]
Or	Organ	Organ	Name	String	Or[liver]

Example: V[concentration]U[mol/l]C[GDP-Mannose]Sa[1231]Ts[2003.06.12 12:00:00]

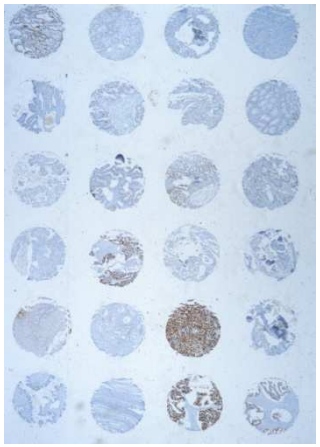
Example: P[her2]V[ihc_score]Or[Breast]Cg[ductal_inf]Cg[TMA34]Sa[br385.487]Sa[B2a]

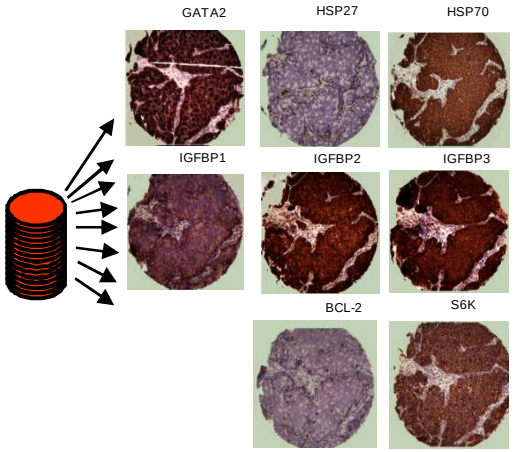
Practical challenges

- Data entry
- Infrastructure
- User interface

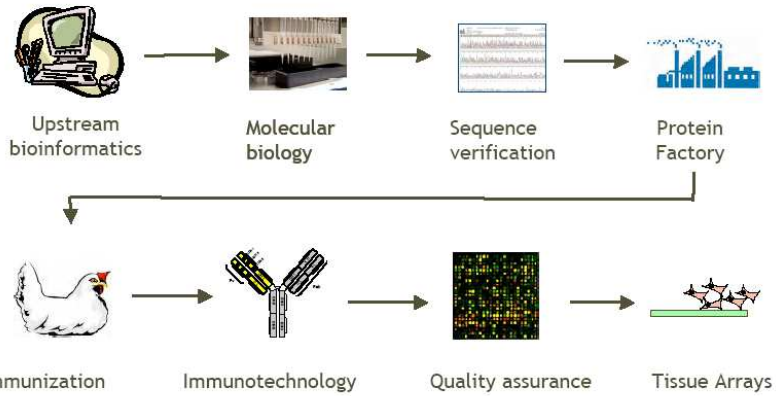
Tissue proteomics with TMAs

1999 idea: Generate a library of replica TMA blocks to make a resource of hundreds of thousands of sections – enough for proteome wide IHC profiling





2004 reality: TMAs now used in antibody-based large scale proteomic projects to generate protein expression atlases

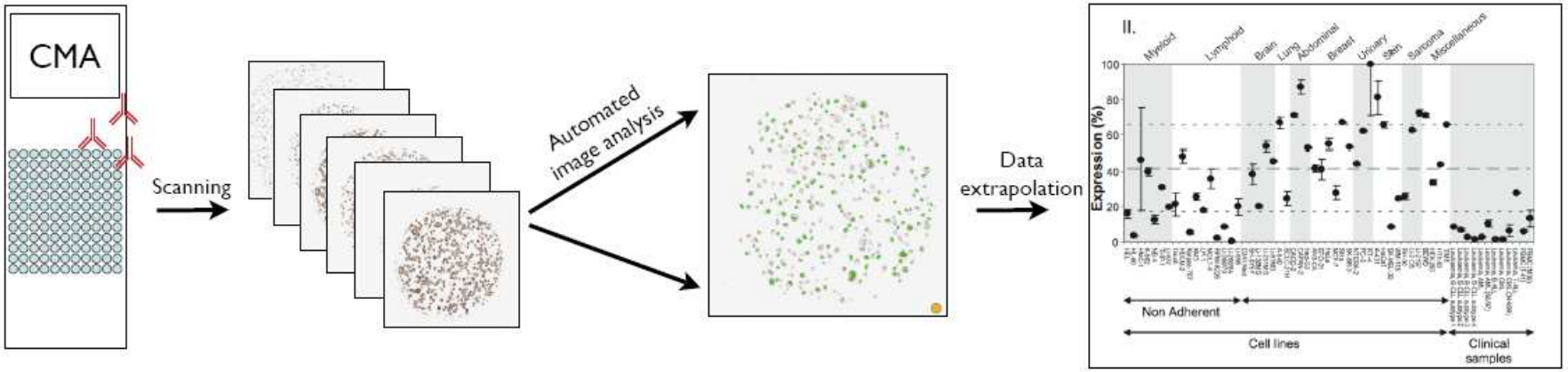


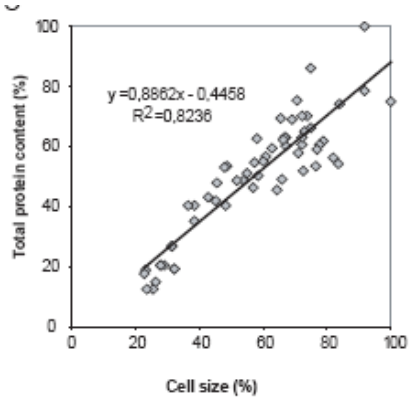
Swedish HPR project:

Version: **3.0** Atlas updated: **2007-10-09** (release history)
Atlas content: **3015** antibodies and **2,827,440** images.

http://www.proteinatlas.org/

Using TMAx to automate workflow allows a new level of throughput: a *pilot* study of 1862 proteins in 47 cell lines and 12 clinical samples





Correlation between cell size and global protein content -> reference standards and normalization approaches are required to reveal true expression changes

Summary

- Cellular-level gene and protein expression data from clinical samples remains valuable for translational research in post-genomic era.
- TMAs provide practical means to optimally manage biospecimen repositories.
- TMAs facilitate development and implementation of high-throughput research approaches.
- Information management is critical for tissue-based research. Automation and in silico analysis approaches are required to convert diverse wet-lab results to knowledge and predictive models.