

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

Juha Kononen

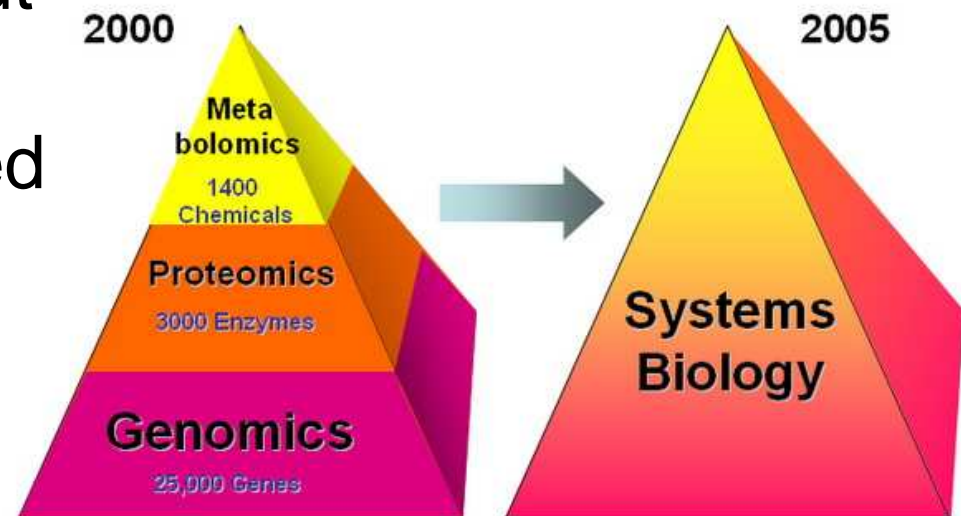
# Problem in Biology and Medicine: Data overload

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



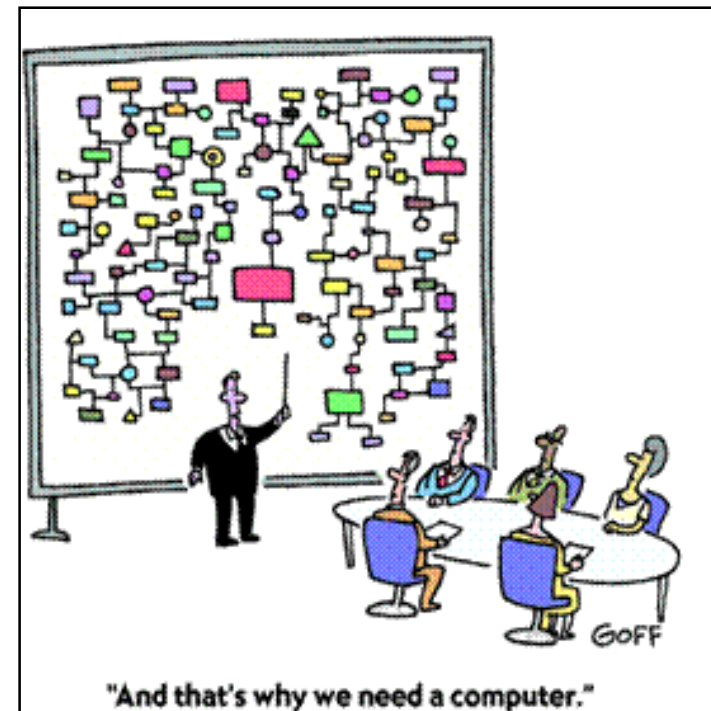
## Merging of 'omics - analyzing individual genes and proteins from few samples is no longer enough

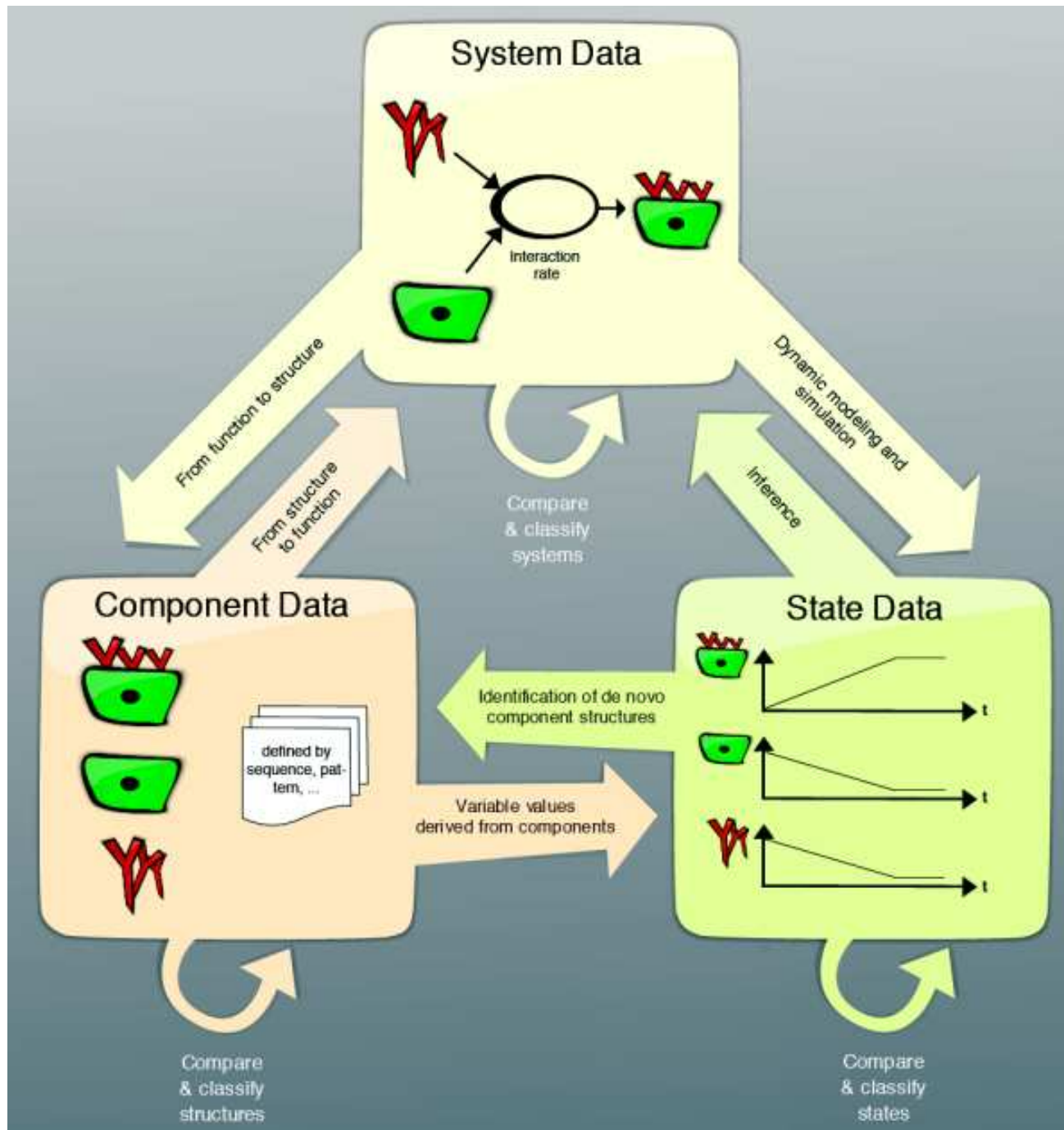
- New high-throughput methodologies constantly introduced
- Abundance of molecular targets
- Pathways and networks emerging



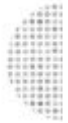
# Data Management Challenge

- 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.



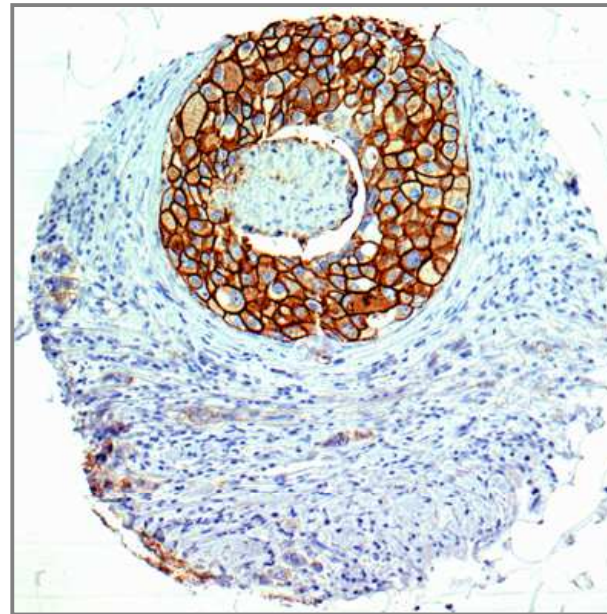
# The goal is to create predictive models (system) from data

- Choose therapy type
- Predict responses
- Monitor responses

*Develop better medical care*

*- individualized treatments*

*- preventive medicine*



# How to get there?

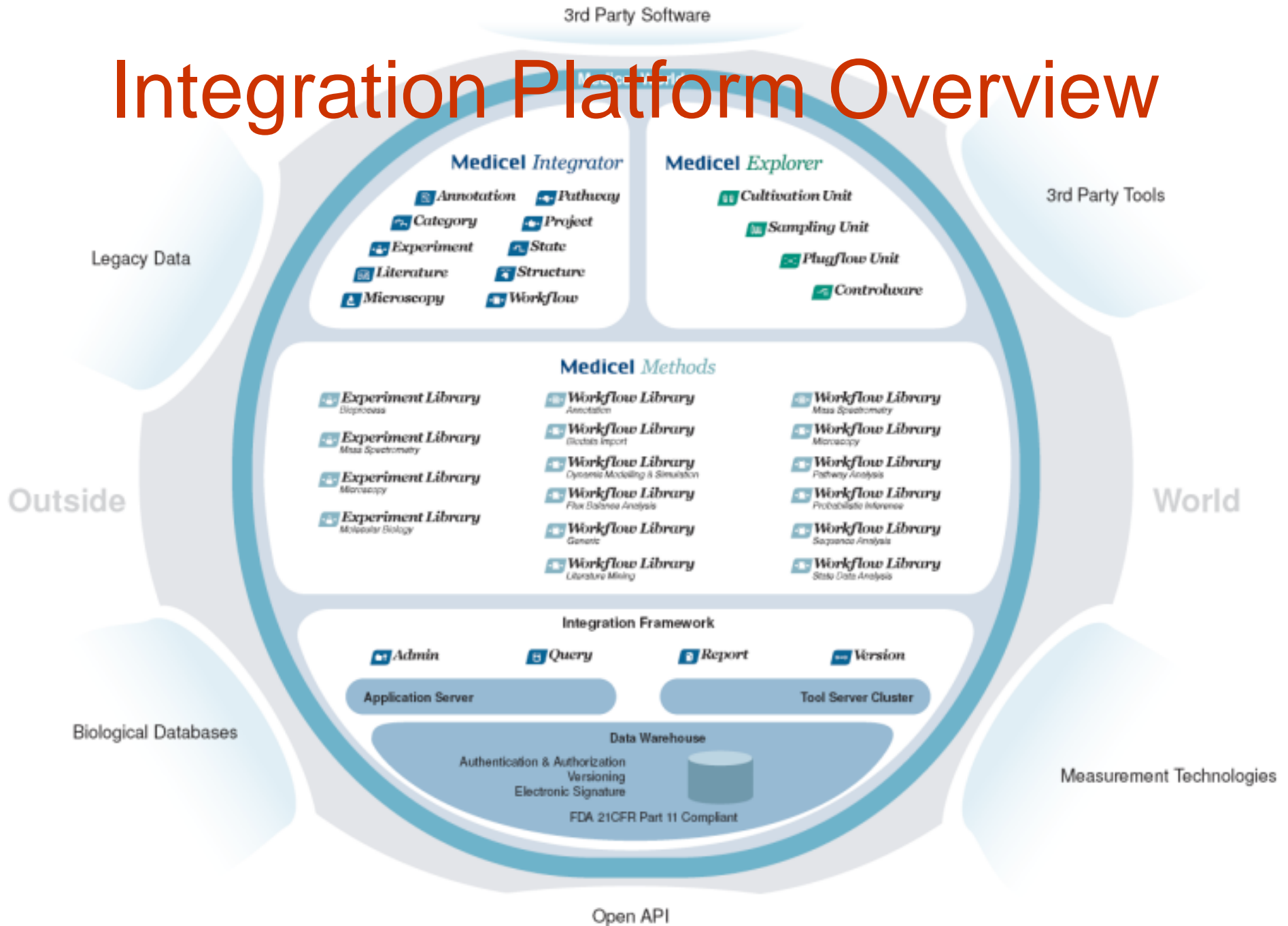
- Define flexible and comprehensive data model
- Create universal data manipulation language
- Automate laboratory
- Automate data management
- Create user interfaces for scientists

# Integration platform

- Example: Mediceal 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



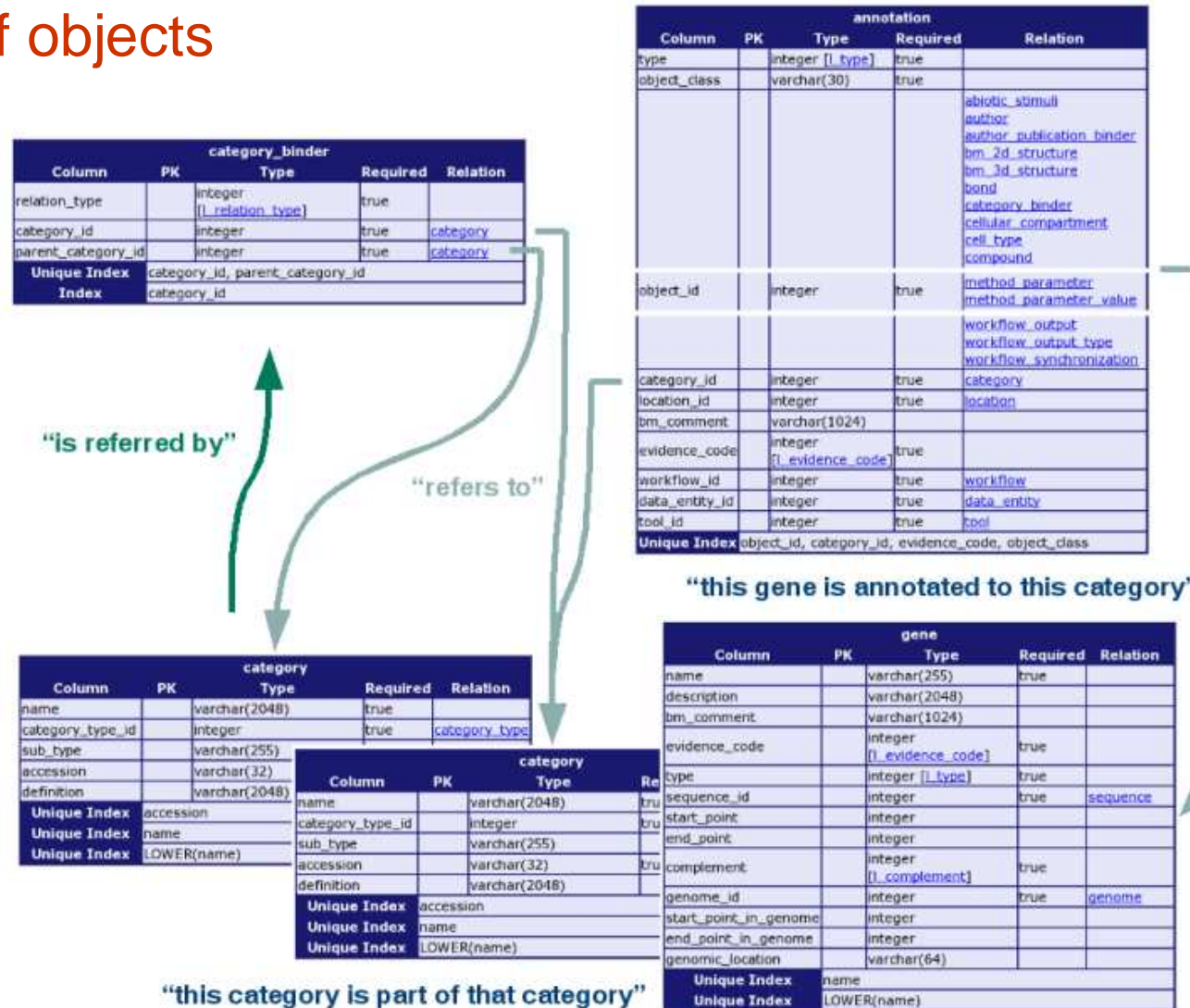
# Unified Information Model – an optimized relational database with minimum number of tables to describe a wide variety of objects

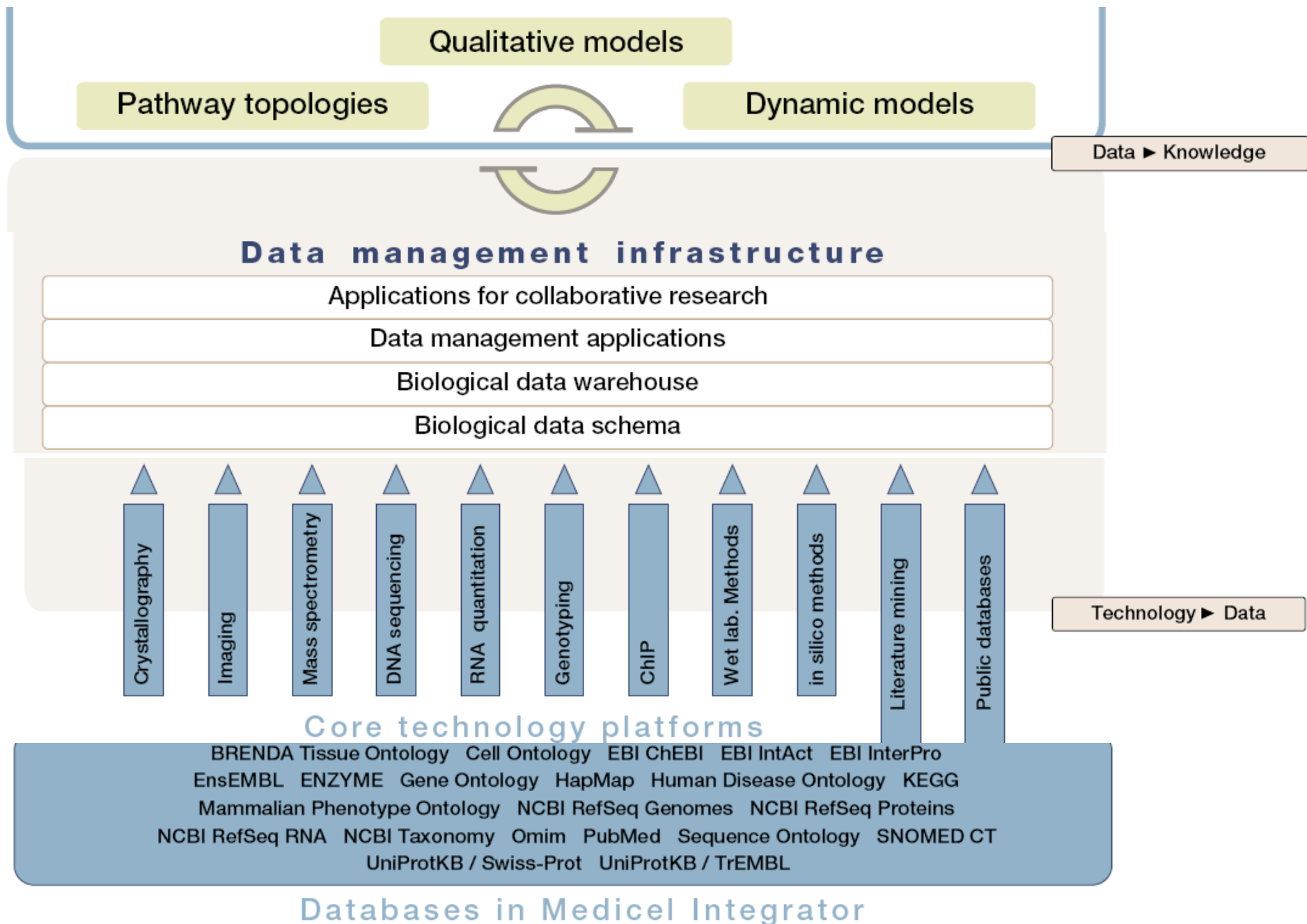
Data and meta-data is organised into tables and fields. Fields can point to other tables allowing the user to access data from different tables.

Abstraction is required to minimize number of tables needed.

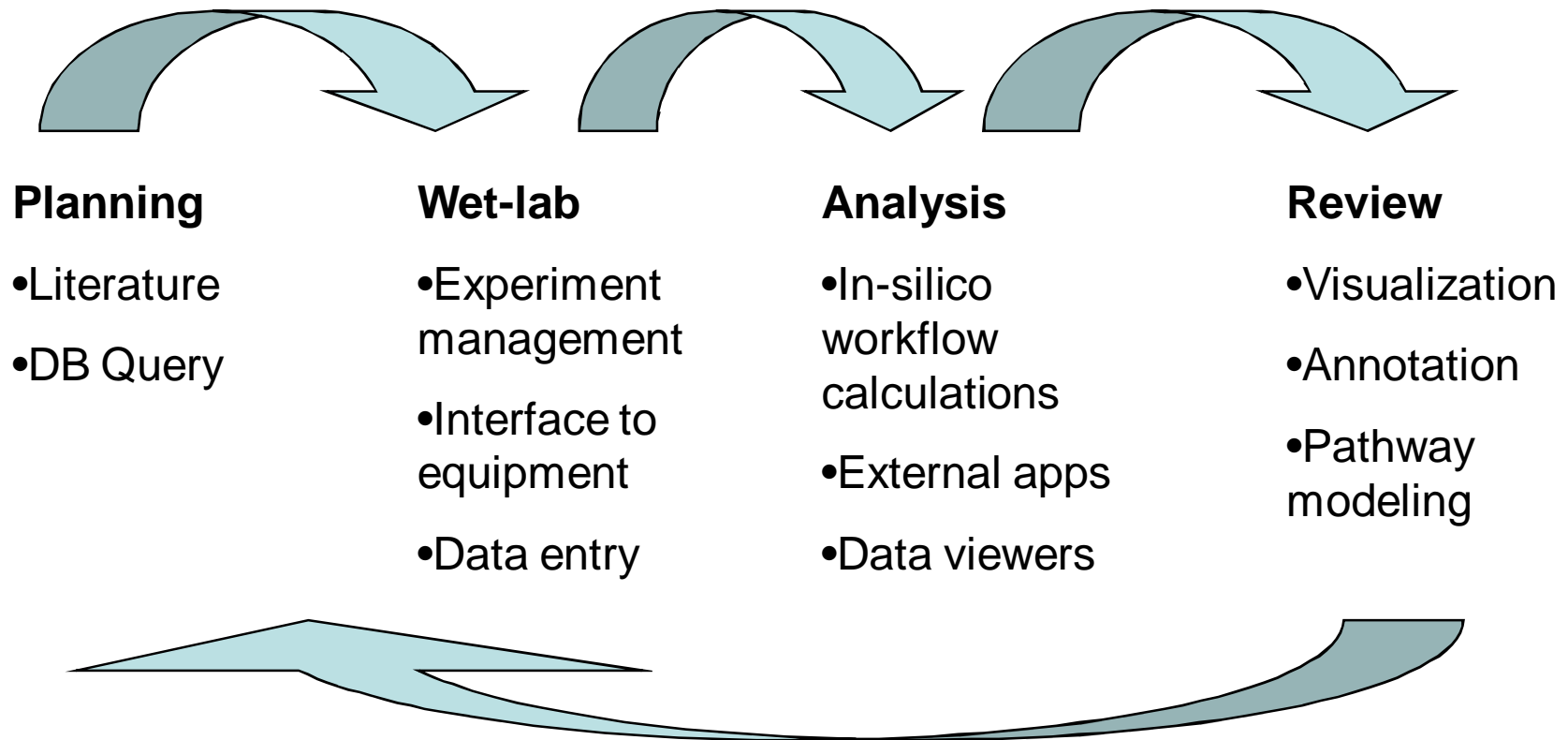
- A gene can be annotated to a category.
- Category can be defined as being part of another category.

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





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



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

The screenshot displays the Medical Integrator software interface, which is used for visualizing and managing experimental workflows. The interface is divided into several key sections:

- Methods/Operations Panel (Left):** A list of available methods under the 'MIC\_LMH' category. A large green arrow labeled 'Operations' points to this panel. The selected method is 'TissueMicroarrayConstruction\_manual\_mta1'.
- Method Details Panel (Bottom Left):** A section labeled 'Details' (indicated by a green arrow) showing the description and parameters of the selected method. The description states: 'Simplified instructions for making tissue microarrays. Please refer to Beecher MTA-1 arrayer manual or scientific publications for more detailed description of the method.' Below this are input and output tables.
- Graph Area (Right):** A large workspace labeled 'Graph Area' (indicated by a green arrow) containing a workflow graph. The graph starts with 'UpdatePatientData' and 'SelectCases', leading to 'GetBlocks\_MarkSlides'. This step produces 'MolecularPathology-sample-4' (TissueCore) and 'stained\_samples'. 'stained\_samples' leads to 'ImagesOfTissueCores' (Digital Images) and 'light\_microscope-1'. 'MolecularPathology-sample-4' leads to 'TissueMicroarrayConstruction\_manual\_mta1-1', which produces 'TMA\_slide'. 'TMA\_slide' leads to 'MolecularPathology-sample-5', which then leads to 'MolecularPathology-reagent-1' (paraffin\_section...). 'MolecularPathology-reagent-1' leads to 'MolecularPathology-reagent-2' (distilled\_water) and 'MolecularPathology-sample-6'. 'MolecularPathology-sample-6' leads to 'IHC\_singlestaining-1'. Other nodes in the graph include 'MolecularPathology-data\_entity\_1287980' and 'MolecularPathology-population-3'.

Name	selected
TissueCore	<input checked="" type="checkbox"/>

Name	selected
TMA_slide	<input checked="" type="checkbox"/>
TMA_block	<input checked="" type="checkbox"/>
TMA_description_file	<input checked="" type="checkbox"/>



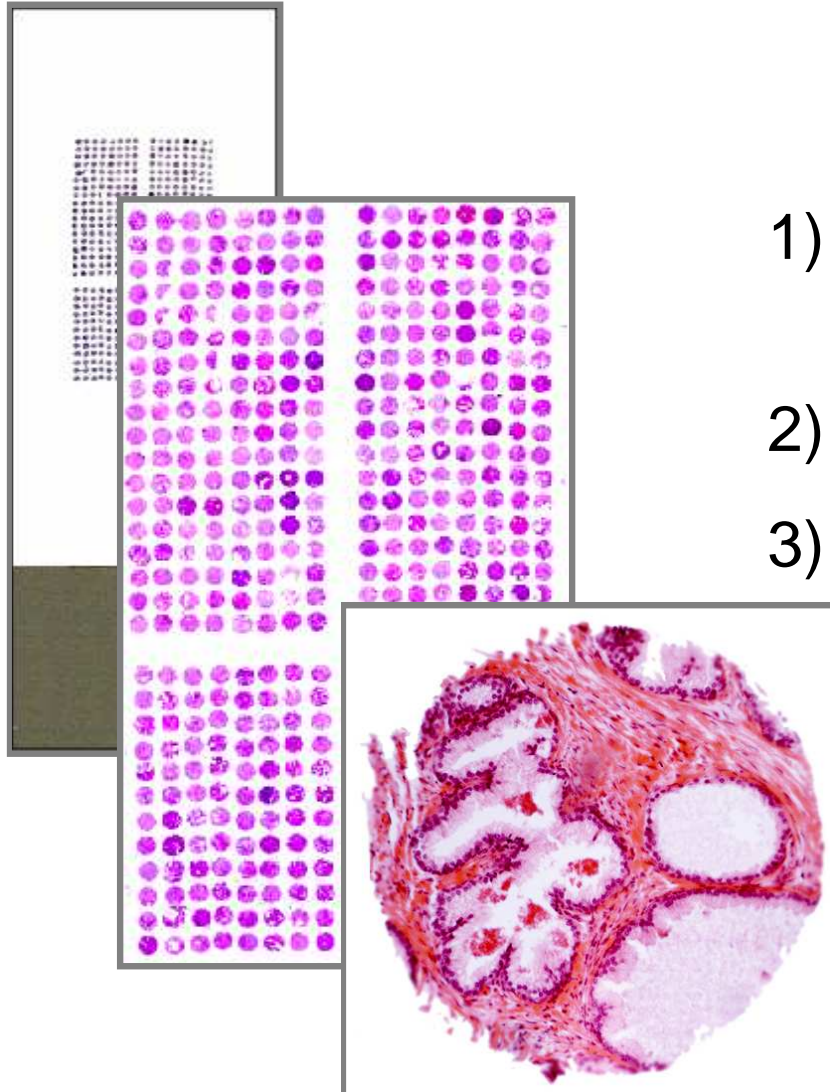
# Pathology archives are treasure troves of clinical data

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



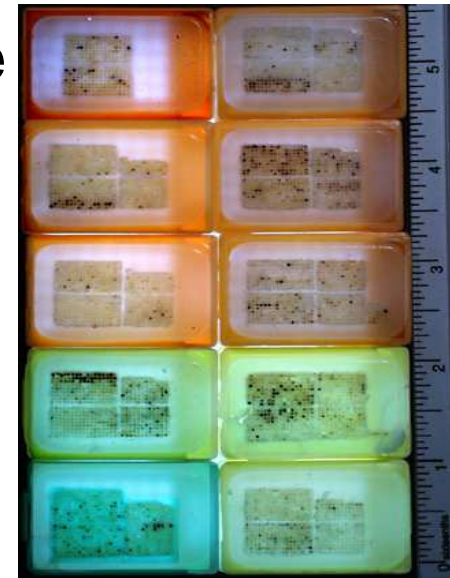
*Pathology  
archive 1999...*

# Tissue Microarrays (TMAs)



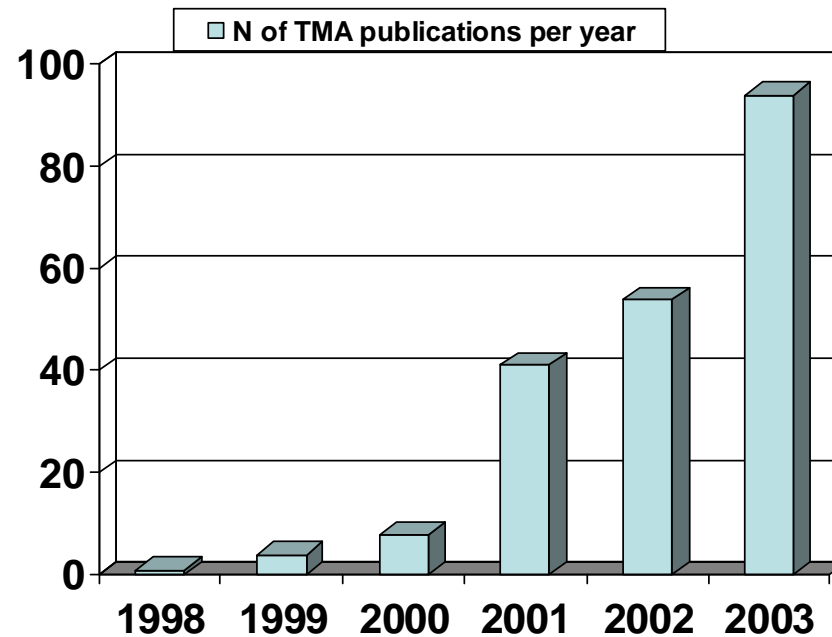
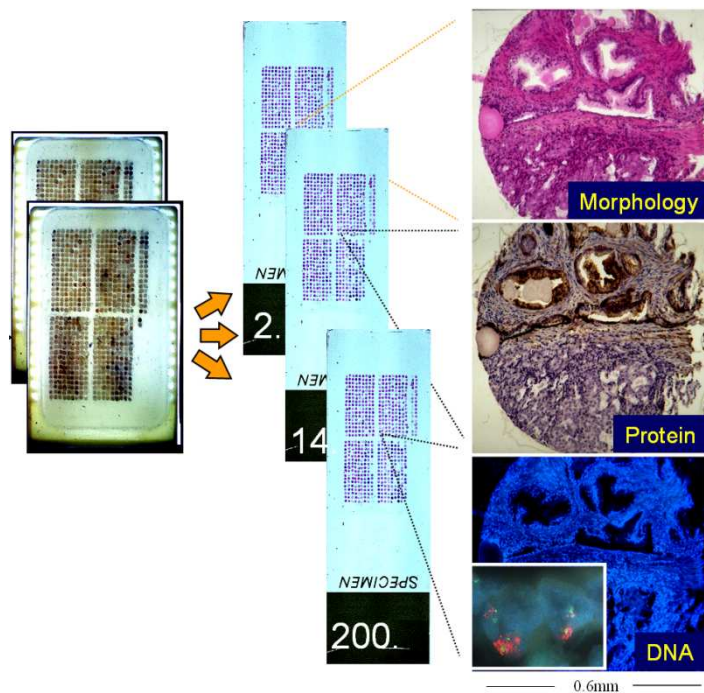
*Platform to*

- 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.

The screenshot displays the Medical Integrator software interface, which is used for visualizing and managing experimental workflows. The interface is divided into several key sections:

- Methods/Operations Panel (Left):** Lists various experimental methods such as `IF_singlestaining v N/A`, `IHC_singlestain`, and `biopsies_dissectionTEMPLATE v 2`. A large green text overlay labeled "Operations" is positioned over this list.
- Method Details Panel (Bottom Left):** Provides a detailed view of the selected method, `TissueMicroarrayConstruction_manual_mta1`. It includes a description, input parameters (e.g., `TissueCore`), and output parameters (e.g., `TMA_slide`, `TMA_block`, `TMA_description_file`). A large green text overlay labeled "Details" is positioned over this panel.
- Graph Area (Right):** A central workspace where a workflow graph is visualized. The graph shows a sequence of operations: `UpdatePatientData` leads to `SelectCases`, which then flows through `GetBlocks_MarkSlides`, `MolecularPathology-sample-4`, `TissueCore`, `ImagesOfTissueCores`, `light_microscope-1`, `MolecularPathology-sample-5`, `paraffin_section...`, `MolecularPathology-reagent-1`, `citric acid`, `distilled water`, `MolecularPathology-reagent-2`, `retrieval_citratebuffer_method...`, `MolecularPathology-sample-6`, and finally `IHC_singlestaining-1`. A large blue arrow points to the `TissueMicroarrayConstruction_manual_mta1-1` node in the graph, which is also labeled "Graph Area".

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

**TissueMicroarrayConstruction\_manual\_mta1: 1.0**

Method Edit View

**Diagram:**

```
graph LR; TissueCore[TissueCore] --> TMA_block[TMA_block]; TissueCore --> TMA_description_file[TMA_description_file]; TissueCore --> TMA_slide[TMA_slide];
```

**Method Description**

Simplified instructions for making tissue microarrays. Please refer to Beecher MTA-1 arrayer manual or scientific publications for more detailed description of the method.

1. Prepare recipient tissue block from paraffin (melting temperature 55-58 deg C). Smoothen the surface with microtome.
2. Create a hole in the recipient block with the smaller punch needle (red handle).
3. Extract donor tissue core with sampling punch.
4. Insert donor tissue into the hole using the stylet. Top of the tissue core should be at the same level as the paraffin surface. Use caution to not push the tissue too deep.
5. Move the sampling needle to the next location using the micrometers.
6. Repeat the punching cycle until all planned array locations are filled.
7. Warm the array block at 37-42 deg C oven to anneal all tissue cores.
8. Carefully smoothen the top surface of the array block with a clean microscope slide.

**Metadata Editor:**

Name: TMA\_description\_file

Description: Data file describing contents of the TMA block.

Required:

Allows multiple objects:

Primary Type	Type Specifier
data_entity	DT_TEXT_TAB_DB (F)

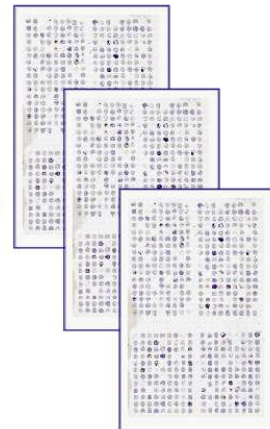
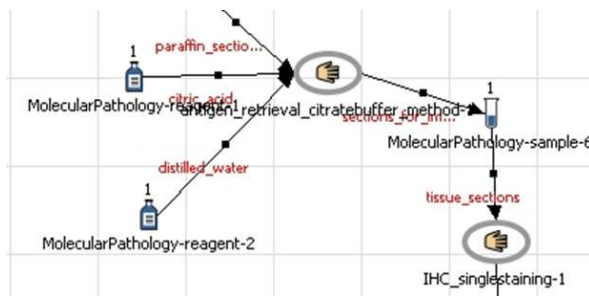
Add Type Remove Type

Save

Description data entity

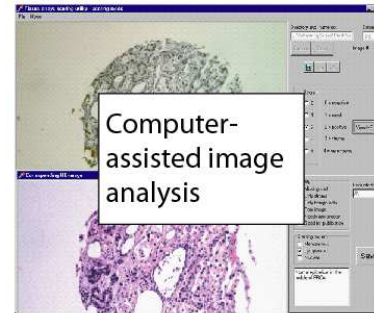
Import Export Delete

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

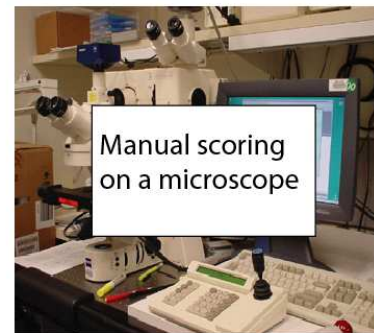


1 TMA -> 500 spots  
20 TMAs -> 10,000 spots to score

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

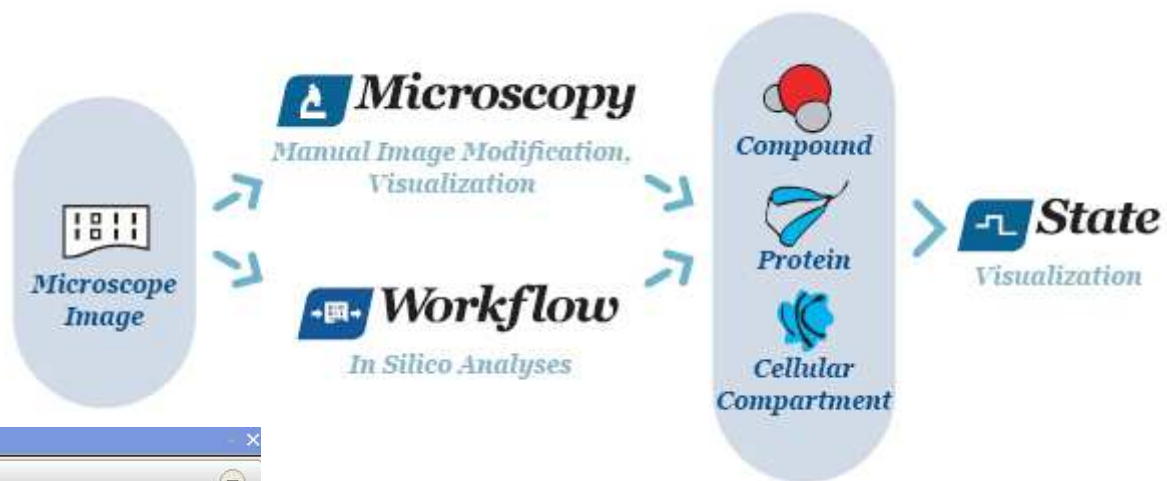


# Image Data Domain

- 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



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 (range 40 to 300)

---

**Stroma**

Distinct stroma:

Stroma brightness: Default

---

**Lymphocytes (inactive)**

---

**Cancer Criteria**

Adjust nuclei size limit manual:

Relative size: 50 (range 50 to 150)

Stain preferentially in cancer:

Exclude small regions:

**Class Description**

Name: Nuclei\_v1.1

Parent class for display: AllNuclei

Modifiers:  Shared  Ab

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: [Grid of function icons]

Membership function:  $x/y$  0.7795180723 / 0.63

Maximum value: 1

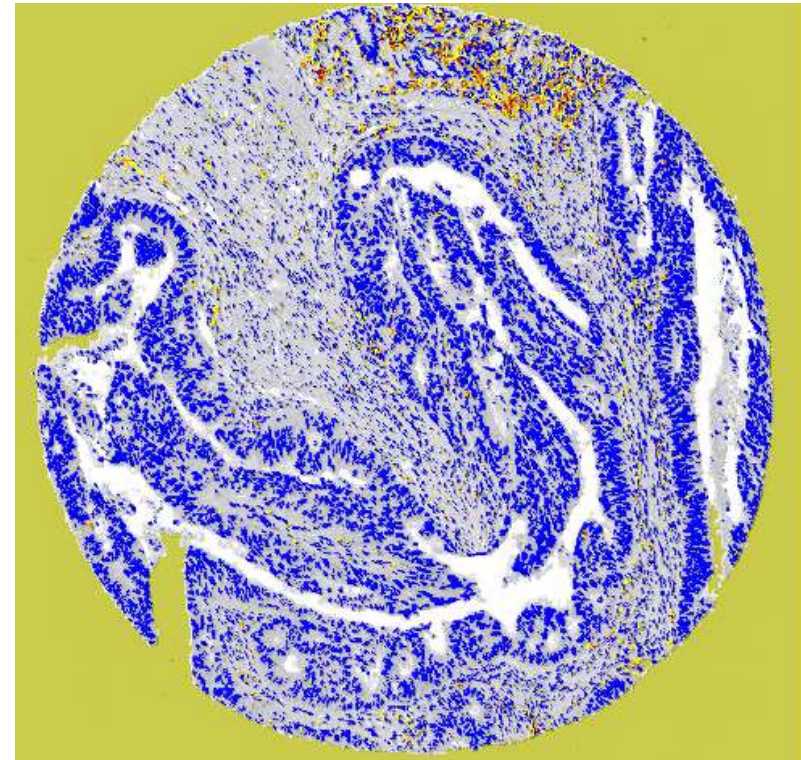
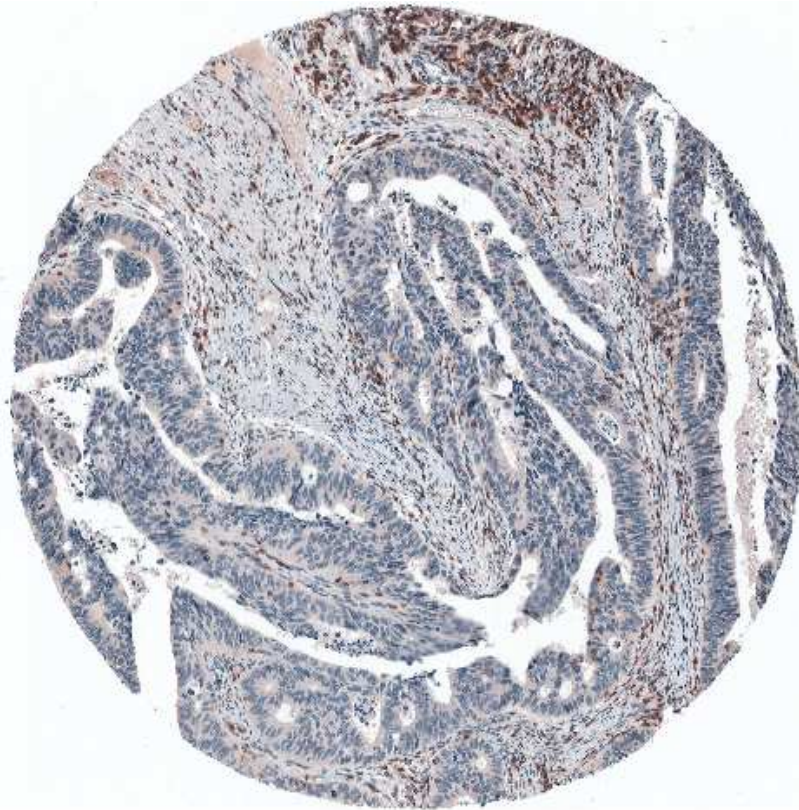
Minimum value: 0

Left border: 0.7 | Right border: 0.9

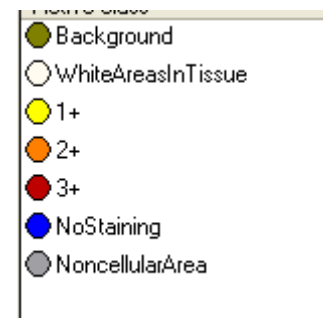
[Graph showing a sigmoidal membership function curve]



# Level 1: Staining detection and intensity evaluation

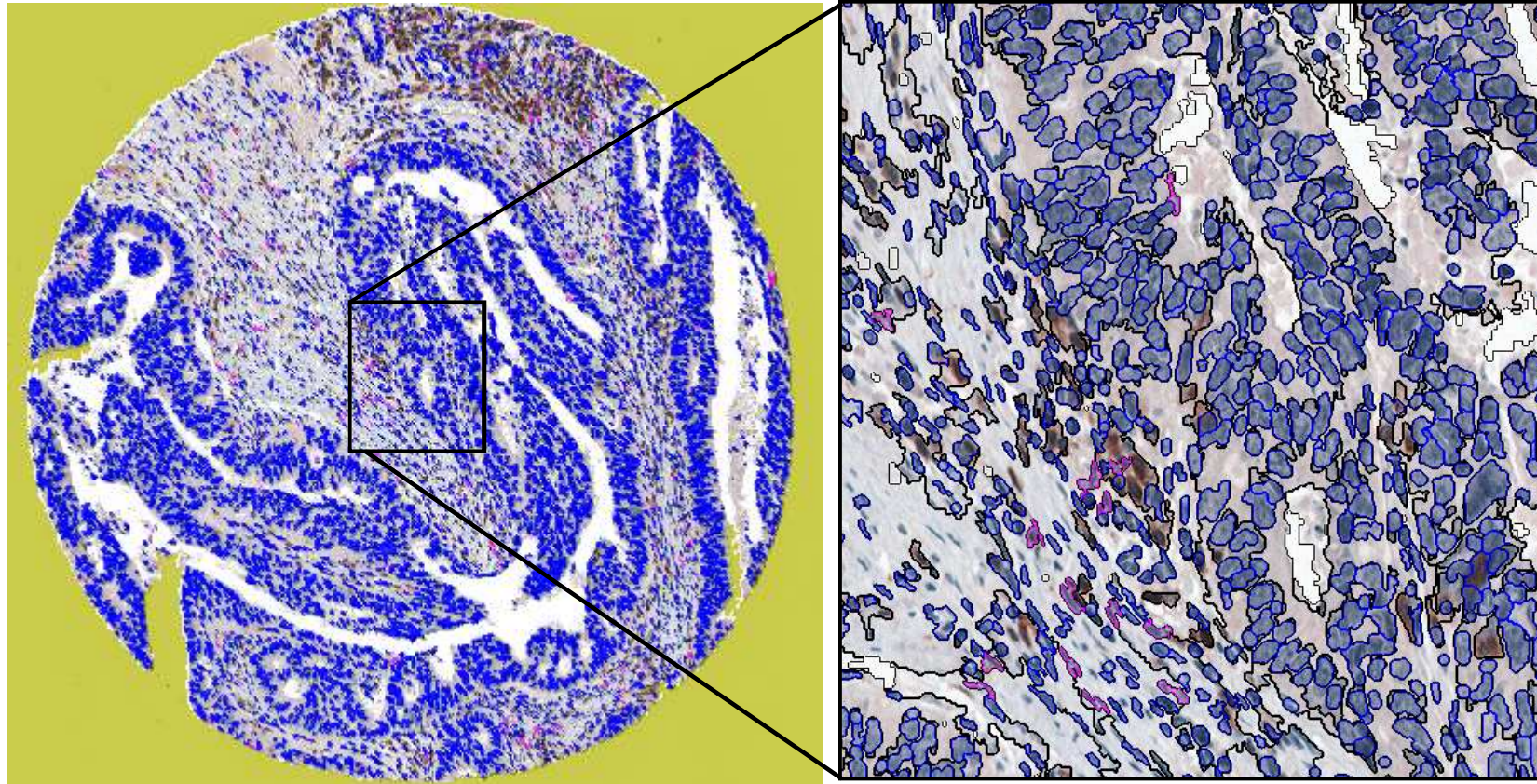


- 1+ = Weak IHC staining
- 2+ = Moderate IHC staining
- 3+ = Strong IHC staining



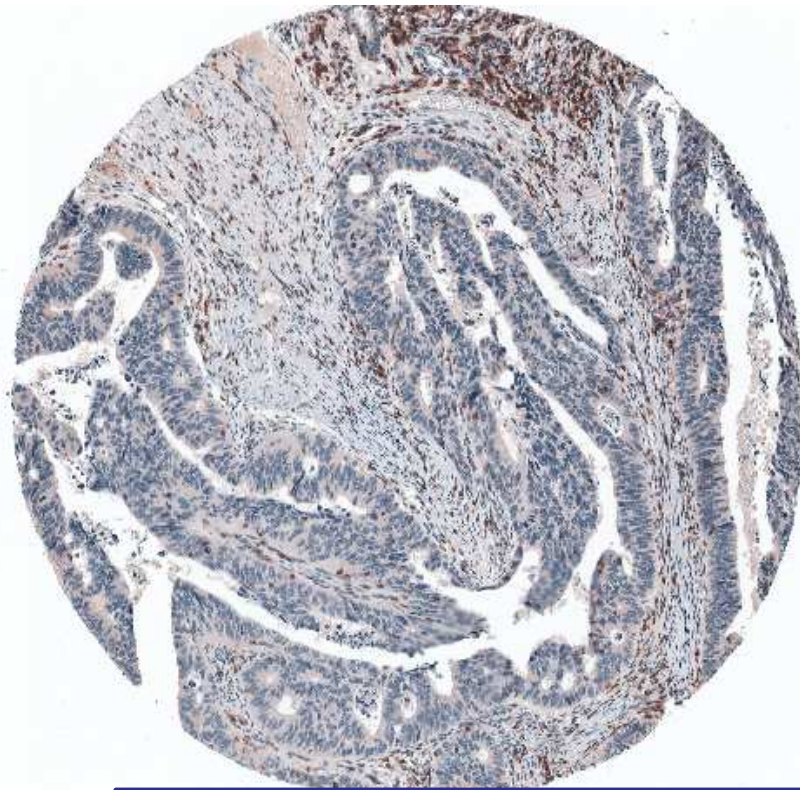


## Level 2: Cell nuclei segmentation

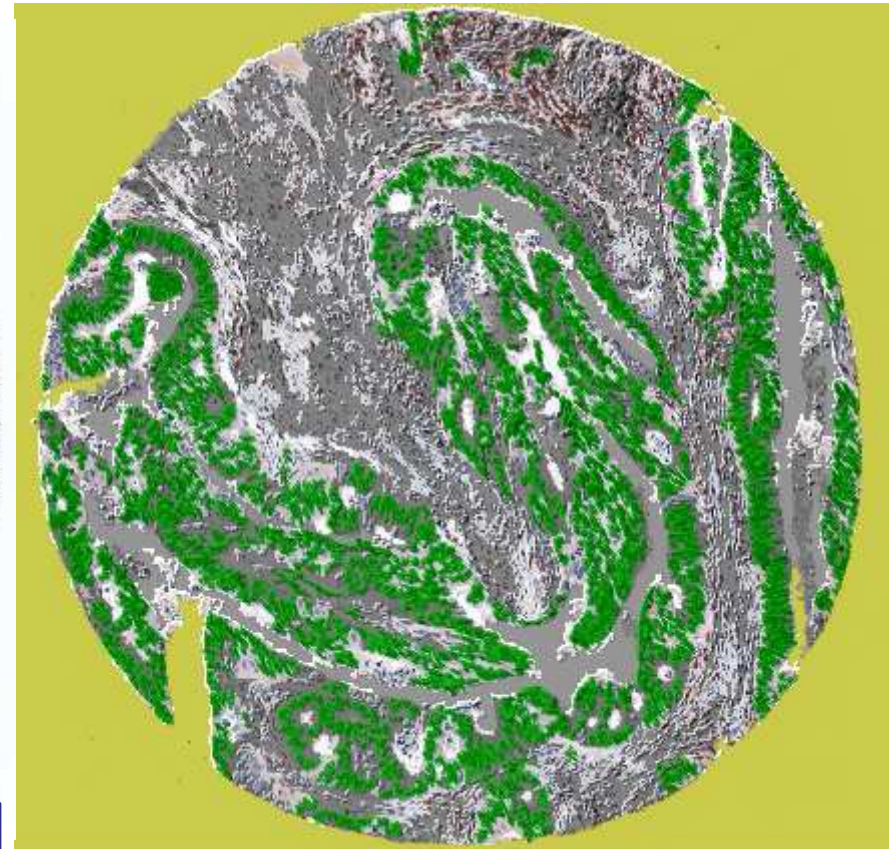




# Level 3: Cancer vs stroma detection

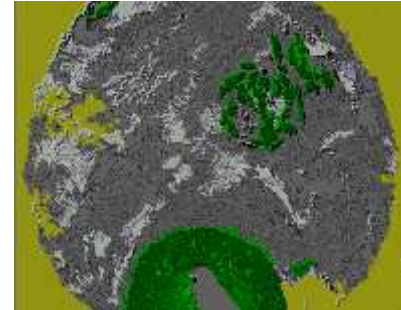
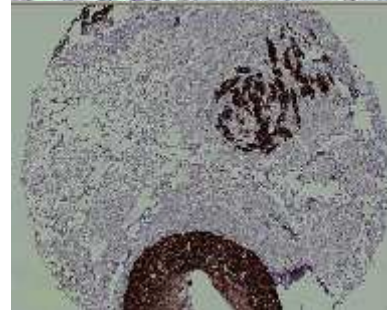
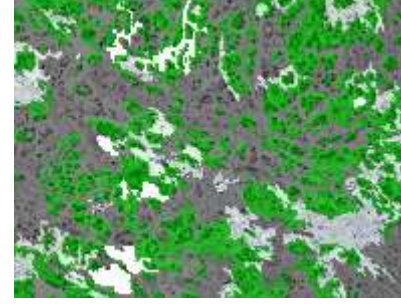
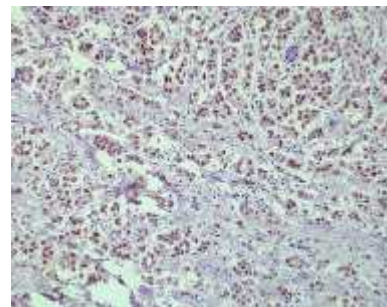
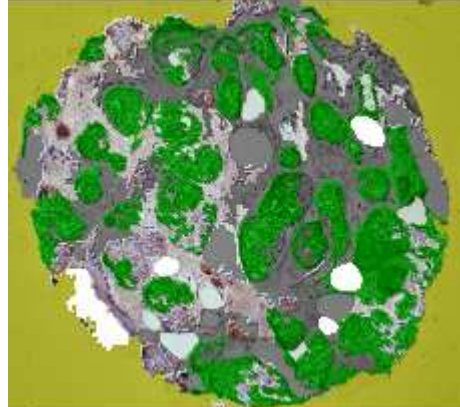
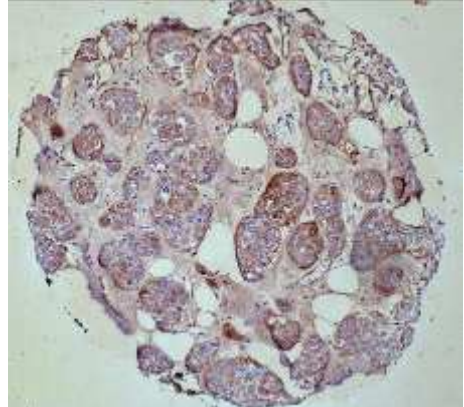
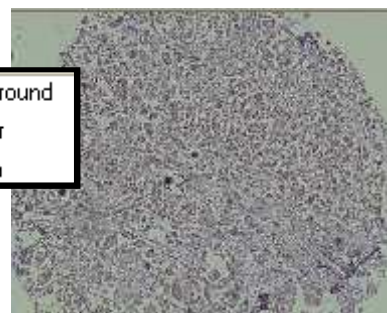
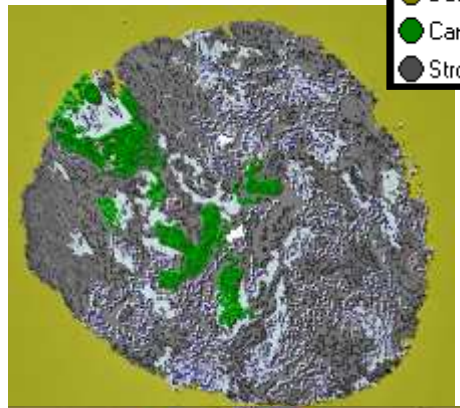
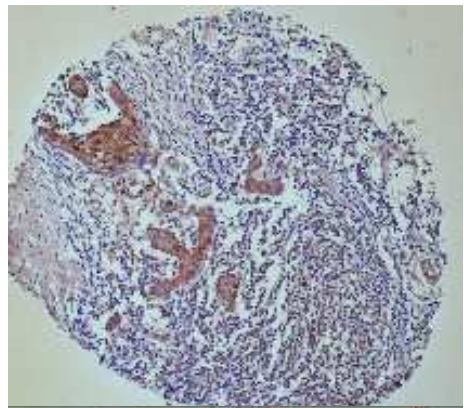
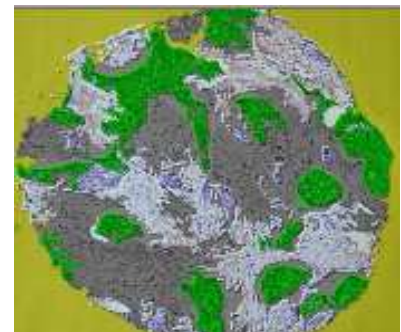
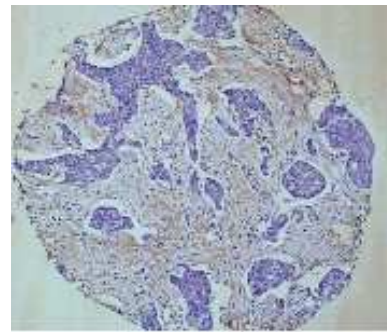
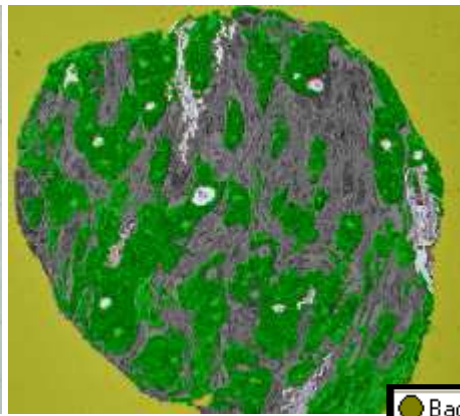
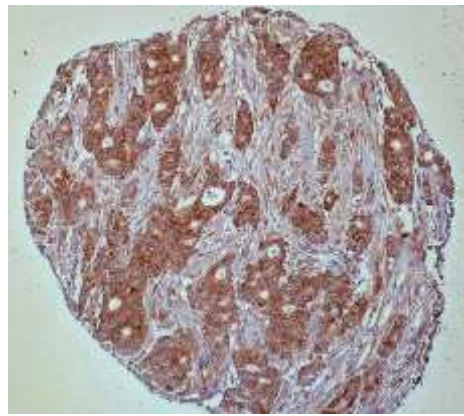


53.8 % stroma  
10.9 % stained stroma  
23.9 % stained cellular stroma  
26.3 % cancer  
2.7 % stained cancer



Active class  
● Background  
● Cancer  
● Stroma

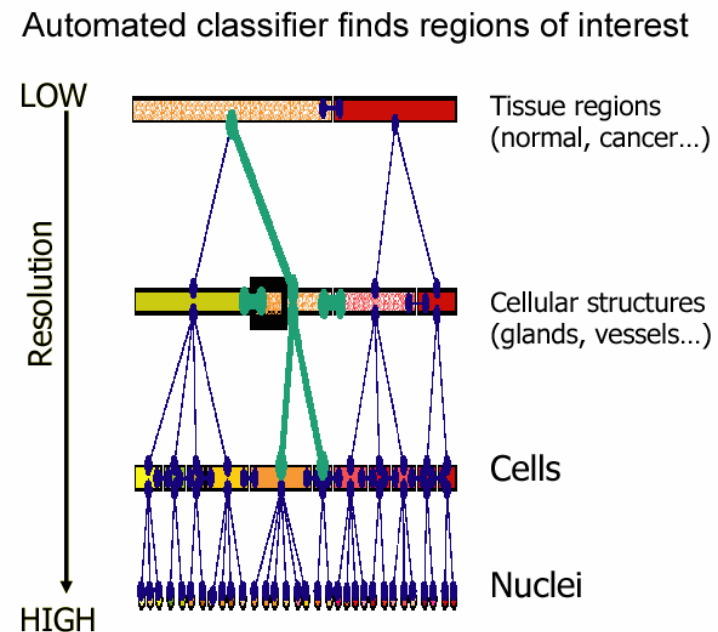




● Background  
● Cancer  
● Stroma

# TMAx: Modeling tissue structures as a network of image objects

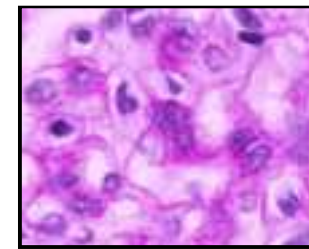
- The original pixel image is processed into a topological and semantical network of objects at multiple scales
- Hundreds of attributes (shape, color, texture, structure, relations) are available for fuzzy logic classification of each image object
- Object-based morphological operations and classification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



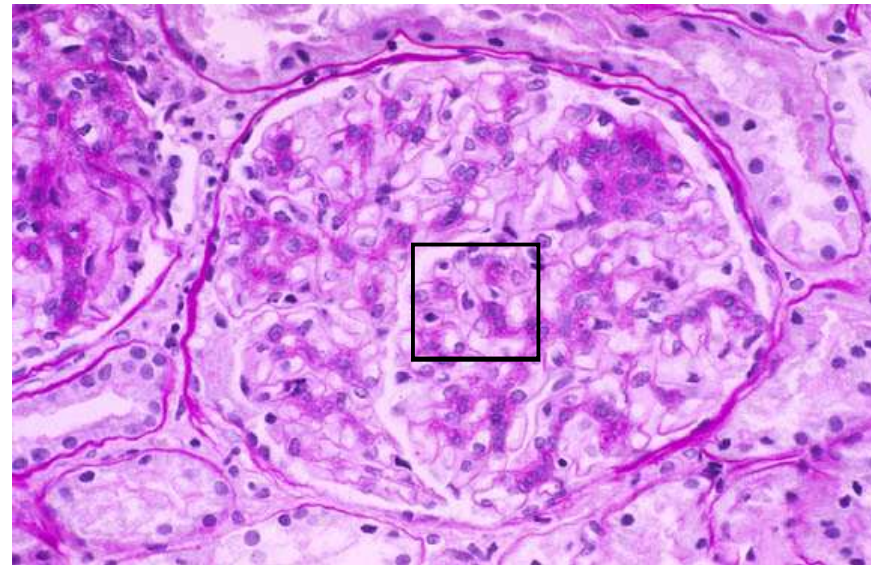


# Analysis is simultaneously performed at multiple scales

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



?

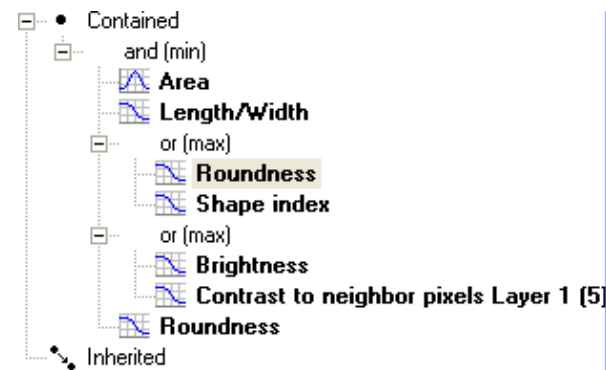


# TMax combines multiple classification criteria using fuzzy logic

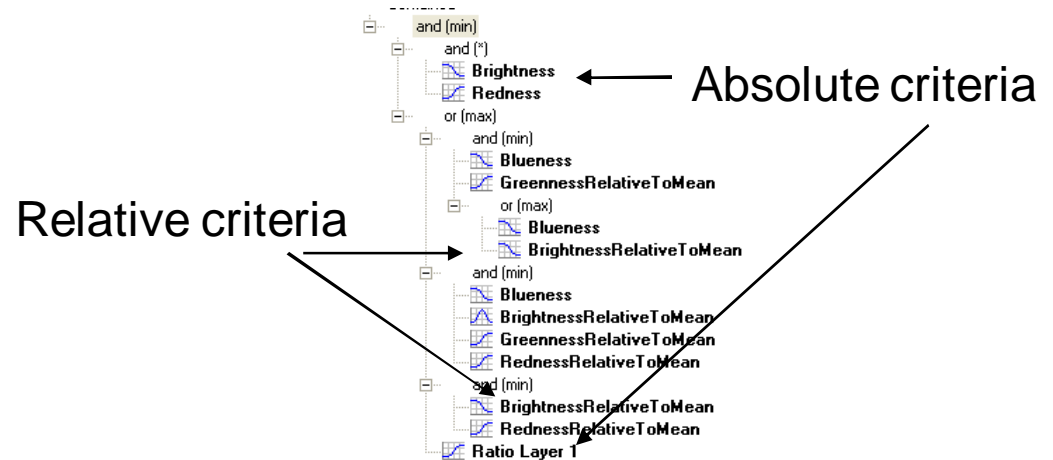
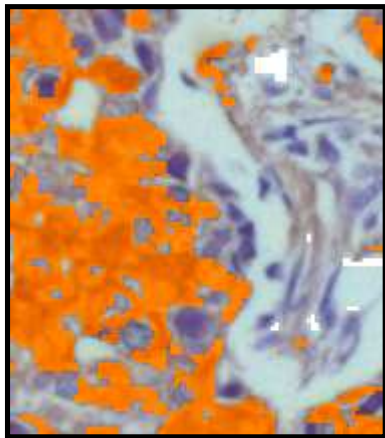
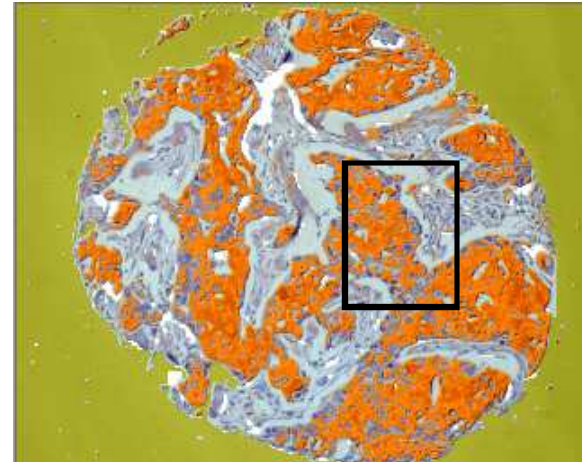
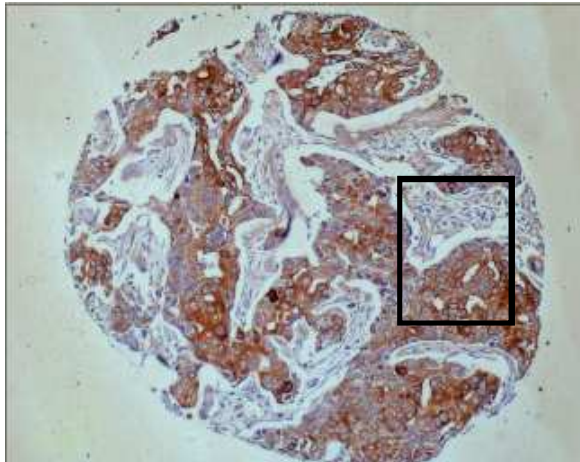
- **Instead of binomial black-or-white labeling, objects have a continuous classification value from 0...1 based on object features, such as brightness relative to neighbor objects**
- **Multiple features can be combined for each classifier**
- **Logical operations (And, Or, Min, Max) can be used to 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



Workspace

Name	State	Time	Type	10 Int...	11...	12 N Cells	13 Cellular Area	20 pctCellsWithSomeStaining	21 ...	22 ...	23 ...	24 ...	25 ...	26 ...	27 ...	28 ...
IHC	DevSet.39...	Processed	Original	3	1	356	991898	92.134831461	25.2...	41.01...	66.85...	25.84...	55.94...	15.73...	33.99...	6.110...
Embryo	DevSet.39...	Processed	Original	3			692231	99.216710183	5.48...	26.37...	93.73...	67.36...	54.79...	10.96...	31.12...	12.68...
Mitosis	DevSet.39...	Processed	Original	2			1303527	22.355289421	8.38...	11.17...	13.97...	2.794...	3.083...	1.652...	1.303...	0.123...
Brain	DevSet.39...	Processed	Original	2			847426	24.147727273	7.38...	14.48...	16.76...	2.272...	2.680...	1.412...	1.152...	0.115...
Spine	DevSet.37...	Processed	Original	1			609076	52.083333333	45.5...	6.547...	6.547...	0	8.952...	6.732...	2.217...	0.000...
DR			Original	2			1005050	84.837545126	7.94...	65.70...	76.89...	11.19...	65.85...	7.429...	54.77...	3.560...
Full			Original	1			681642	2.0547945205	2.05...	0	0	0	0.358...	0.355...	0.001...	0
TM			Original	1	1	414	851989	94.927536232	92.7...	2.173...	2.173...	0	32.92...	29.21...	3.666...	0.017...
Cell			Original	2	1	160	319792	86.875	34.375	51.875	52.5	0.625	51.63...	24.26...	27.06...	0.280...
			Original	2	1	200	156239	65	40	24	25	1	24.43...	14.07...	9.651...	0.707...
			Original	2	1	586	1449851	98.464163823	23.0...	65.69...	75.42...	9.726...	60.80...	20.04...	38.56...	2.177...
			Original	3	2	479	1274863	98.747390397	4.17...	22.54...	94.57...	72.02...	72.99...	6.232...	36.29...	30.40...
			Original	3	2	412	1129235	87.409200969	10.8...	26.87...	76.51...	49.63...	45.01...	6.314...	28.77...	9.879...
			Original	2	4	257	551483	79.766536965	59.1...	12.84...	20.62...	7.782...	23.83...	18.28...	5.096...	0.382...
			Original	2	2	192	397656	99.479166667	39.0...	54.16...	60.41...	6.25	63.27...	33.25...	28.77...	1.182...

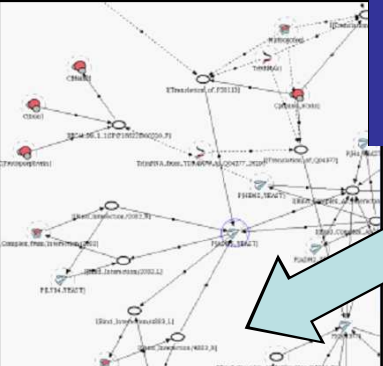
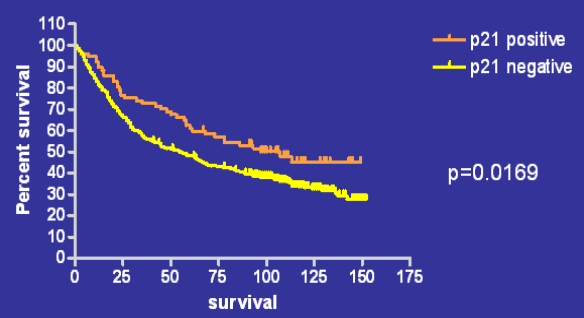
Integration framework allows processing TMax measurements with in-silico analysis tools for finding statistical correlations, visualization and pathway modeling

TMax Data

Choosing analysis method

- Calculate\_Cluster\_Statistics
- Calculate\_Distribution\_Statistics
- Calculate\_Distribution\_Statistics\_old
- Calculate\_Protein\_Statistics
- Convert\_Protein\_Statistics\_to\_Data\_Set

Visualizing results



Annotation and dynamic pathway modeling results

# 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]

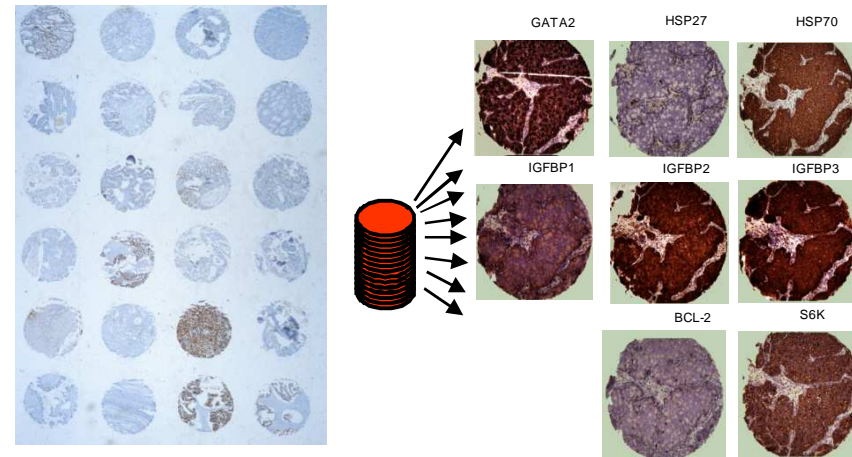


# Practical challenges

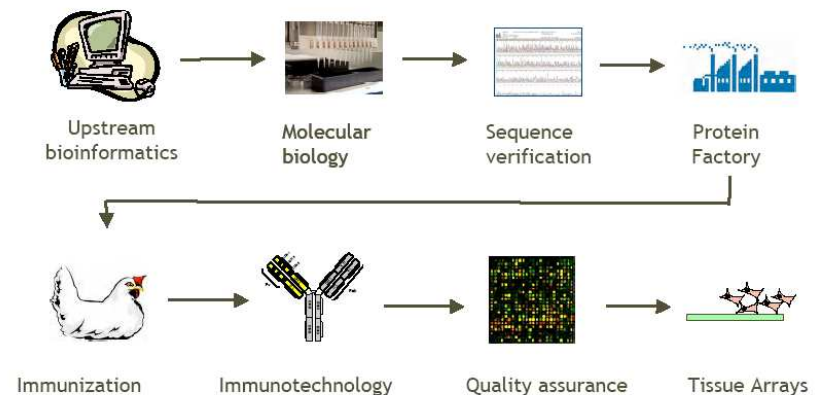
- Data entry
- Infrastructure
- User interface

# 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



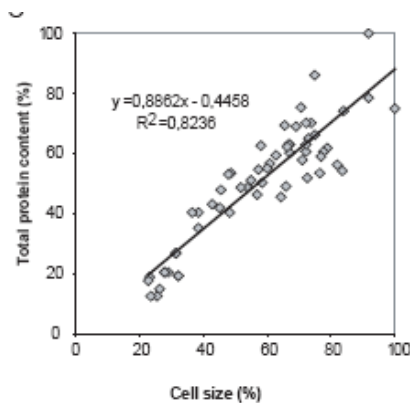
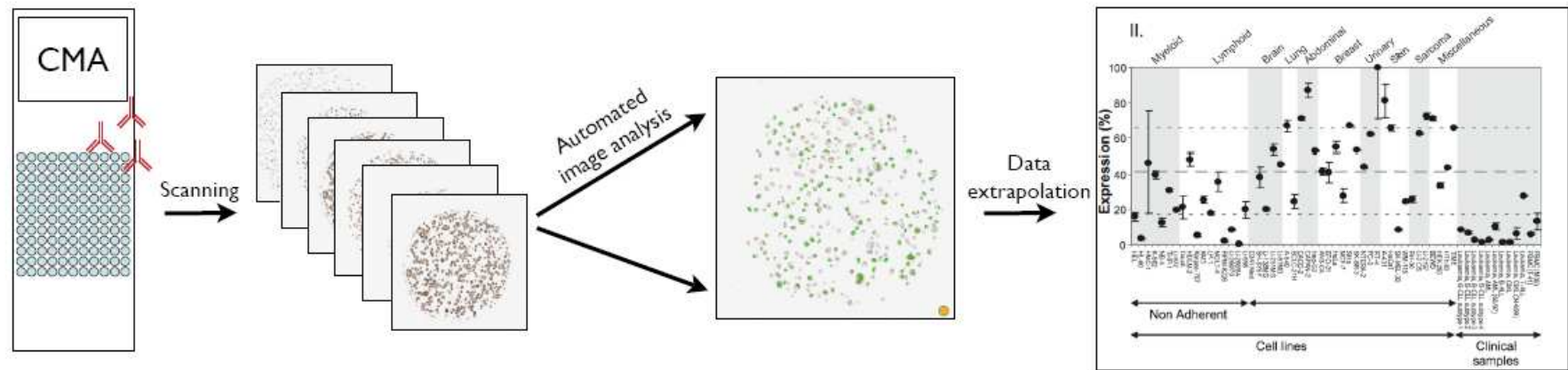
2004 reality: TMAs now used in antibody-based large scale proteomic projects to generate protein expression atlases



Swedish HPR project:

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

# 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.