Scientific Applications of Machine Learning

Eric Mjolsness

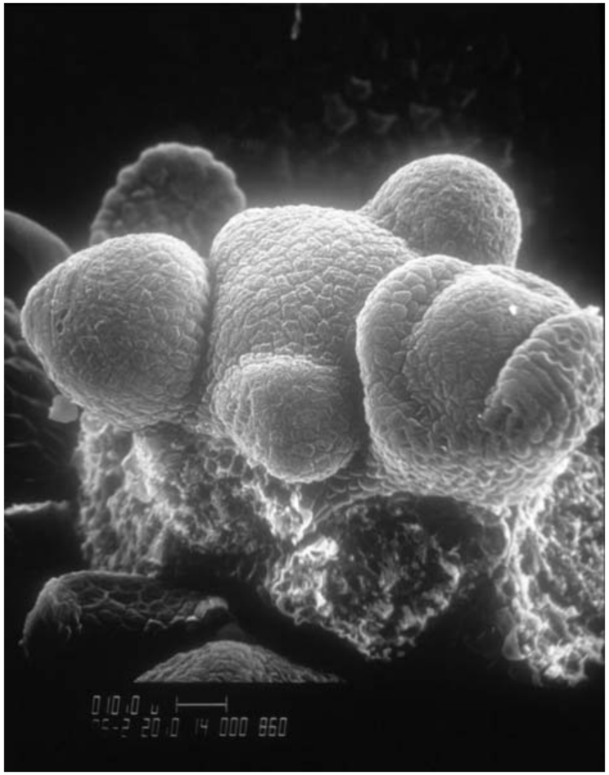
Scientific Inference Systems Laboratory

Donald Bren School of Information and Computer Sciences, *and*

Institute for Genomics and Bioinformatics

University of California, Irvine

Scientific Imagery Applications



Arabidopsis SAM - Meyerowitz Lab

NGC 7331 **-** http://photojournal.jpl.nasa.gov/catalog/PIA06322

Some Basic Machine Learning Distinctions

- Supervised vs. unsupervised learning
  - Supervised e.g. classification and regression
    - Feature selection
    - regression for phenomenological model fitting e.g. GRN’s
  - Unsupervised e.g. clustering; may be preprocessor
- Generative vs. Kernal methods
  - Generative (statistical inference) models
  - Kernal methods e.g Support Vector Machines
- Vector vs. Relationship data
  - Vector data: preprocessed image features Δlog I, Δx, …
  - Images, time series, shifted spectra - semigroup actions
  - Sparse graph/relationship data - permutation actions

Correspondence Problems

- Extended sources - map morphologies
  - Similar to biological imaging problems
  - Fewer sources but many pixels
- Moving or changing point sources
  - E.g. Ida and Dactyl / JPL MLS
- Dense point sources with instrument noise e.g. globular clusters (radial density function)
- Techniques:
  - soft permutations, geometric transformations via optimization & continuation
  - Embedding inside a graph clustering (optimization) algorithm
  - Multiscale acceleration of optimization

Mixture Models

- Mixture of Gaussians, t-distributions, …
  - Can do outlier detection
- Mixture of factor analyzers

$$f(\boldsymbol{X}|Y,\boldsymbol{Z},\Lambda,M,\Psi) = \prod_{i=1}^{n}\prod_{k=1}^{m}\{N(\boldsymbol{x}_i|\boldsymbol{\Lambda}_k\boldsymbol{y}_{ki}+\boldsymbol{\mu}_k,\boldsymbol{\Psi}_k)\}^{z_{ki}}$$

Utsugi and Kumagai 2000

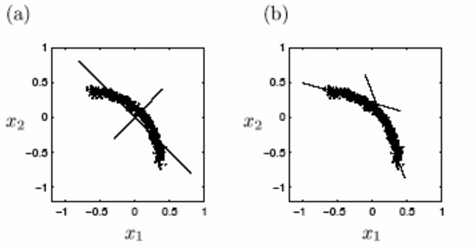
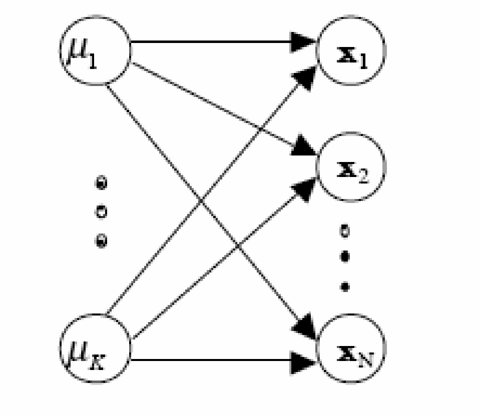


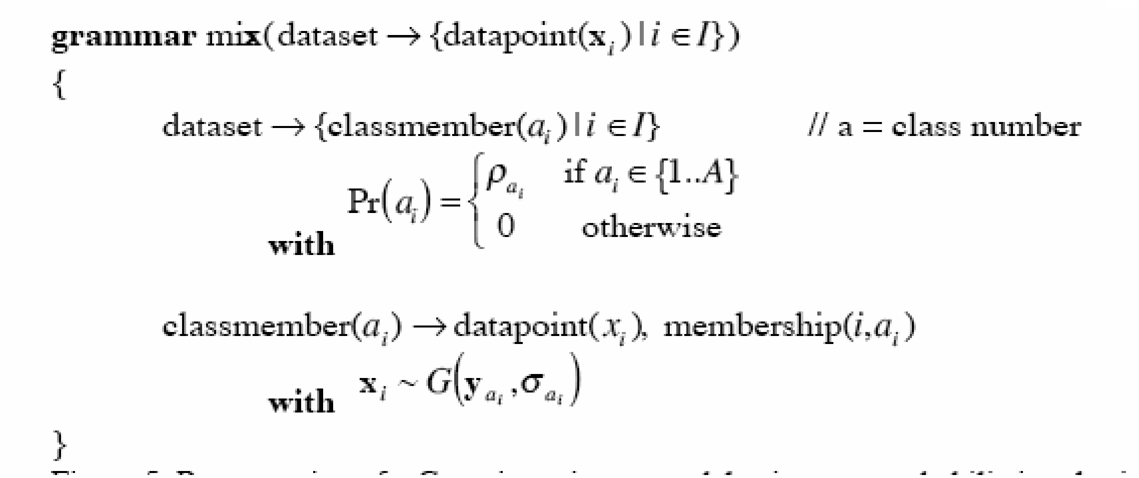
Figure 3: (a) A 2-dimensional scatter plot of some 3-dimensional data that lies on a 2-dimensional subspace. The data actually comes from a curved 1-dimensional manifold. The principal components fail to capture the curvature. (b) A mixture of 1-dimensional subspaces fits the data much better.

Frey et al. 1998

- Mixture of time series models
- * Problem-specific generative models
  - Can formulate with a Stochastic Parameterized Grammar
  - Clustering graphs

Stochastic Grammars for Data Modeling




grammar mix$(\text{dataset} \longrightarrow \{\text{datapoint}(\mathbf{x}_i) \mid i \in I\})$

{

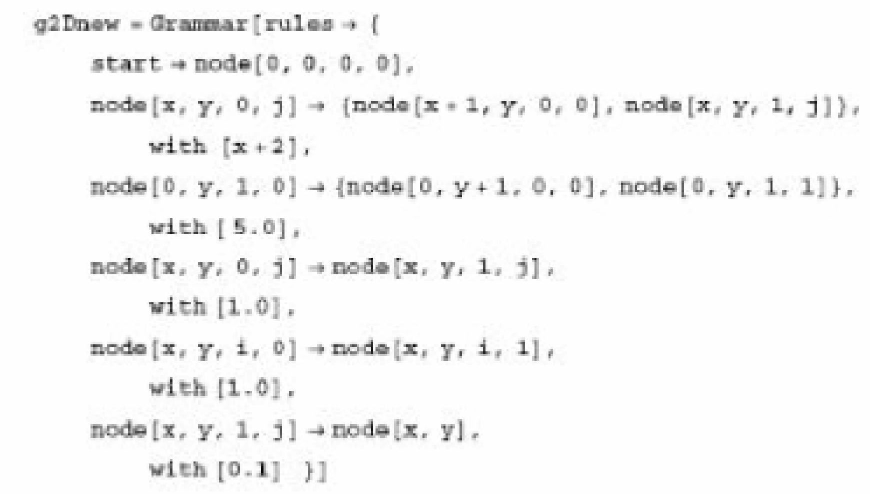
dataset $\longrightarrow \{\text{classmember}(a_i) \mid i \in I\}$ // a = class number

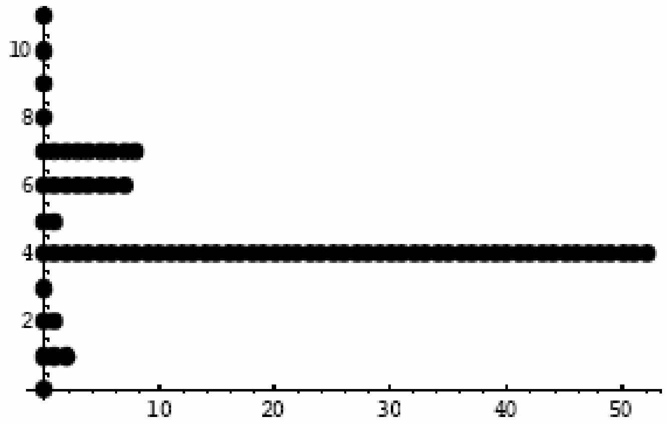
with $\Pr(a_i) = \begin{cases} \rho_{a_i} & \text{if } a_i \in \{1..A\} \\ 0 & \text{otherwise} \end{cases}$

classmember$(a_i) \longrightarrow \text{datapoint}(x_i),\ \text{membership}(i, a_i)$

with $\mathbf{x}_i \sim G(\mathbf{y}_{a_i}, \sigma_{a_i})$

}


```
g2Dnew = Grammar[rules → {
    start → node[0, 0, 0, 0],
    node[x, y, 0, j] → {node[x + 1, y, 0, 0], node[x, y, 1, j]},
        with [x + 2],
    node[0, y, 1, 0] → {node[0, y + 1, 0, 0], node[0, y, 1, 1]},
        with [ 5.0],
    node[x, y, 0, j] → node[x, y, 1, j],
        with [1.0],
    node[x, y, i, 0] → node[x, y, i, 1],
        with [1.0],
    node[x, y, 1, j] → node[x, y],
        with [0.1]  }]
```



Text & Biology Models

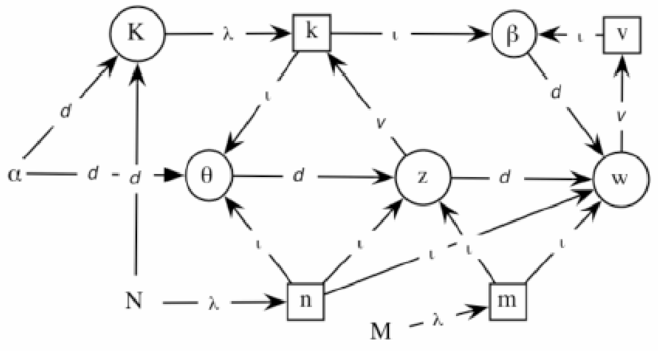


Figure 13. Document topic model.

In this diagram the notation is as follows. The index nodes are: $n \in \{1 \ldots N\}$ indexes the documents; $m \in \{1 \ldots M\}$ indexes the word positions in a document (padded out to maximal document length, or else subscripted as M_n); $k \in \{1 \ldots K\}$ indexes the topics a word or document can be "about"; $v \in \{1 \ldots V\}$ indexes the vocabulary of possible words.

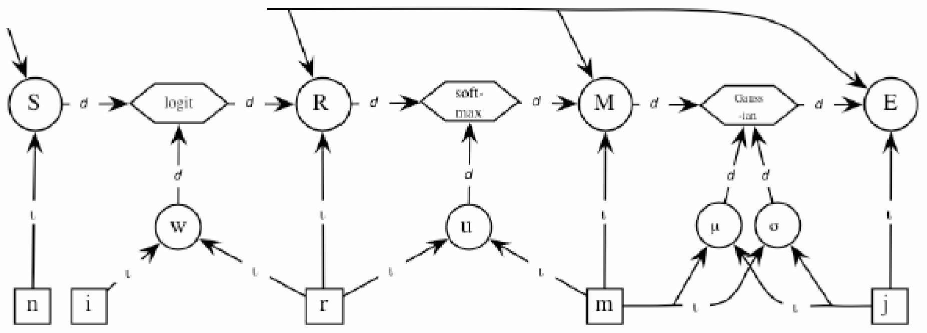
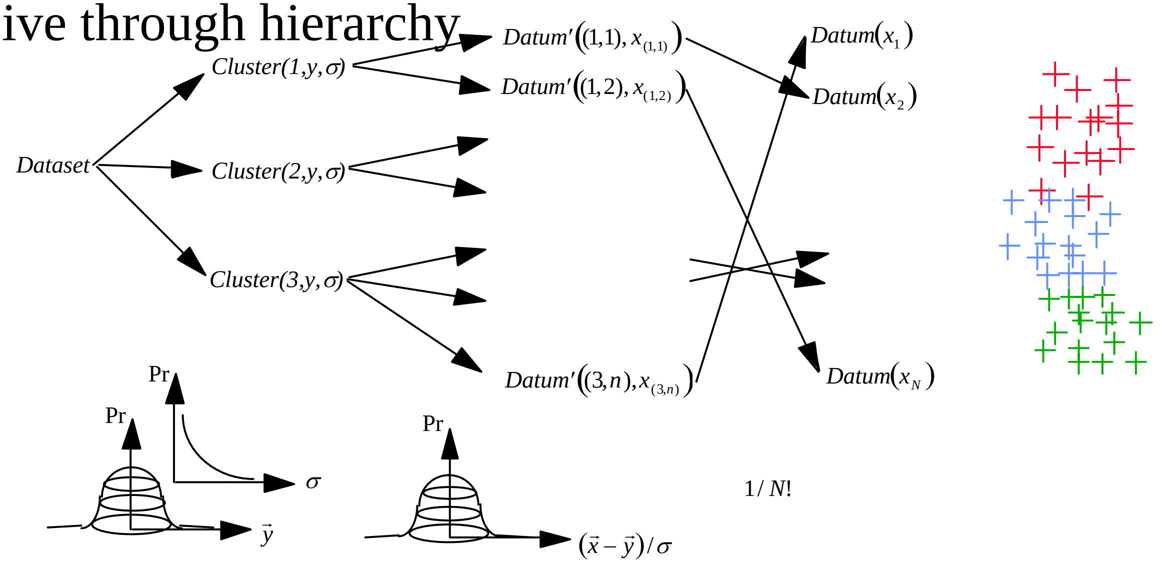


Figure 16. Model for regulation of mRNA expression level as a function of sequence information [Segal et al. 2003].

More Detailed Clustering Grammars

- Clusters generate data
- Priors on cluster centers & variances
- Iterative through levels in a hierarchy
- Recursive through hierarchy



Rock Field Grammar

grammar rockfield() start $\rightarrow \{\text{deposit}(a, \mathbf{y}_a, \mathbf{c}_a, \mathbf{p}_a) \mid a \in A\},$ distractors

{

deposit$(a, \mathbf{c}_a, \mathbf{p}_a) \rightarrow \{\text{patch}(a, b, \mathbf{x}_{ab}, \mathbf{c}_a, \mathbf{p}_{ab}) \mid a \in A, b \in B_a\}$

$$\sum_a \|\mathbf{c}_a\|^2 / 2\sigma_0^2 + \sum_a \|\mathbf{y}_a\|^2 / 2\sigma_0^2 + \sum_a \|\mathbf{p}_a - \ddot{\mathbf{p}}\|^2 / 2\sigma_1^2 + \log^2\left(\ddot{\sigma}_a / \sigma\right)$$

$$\sum_a \|\mathbf{y}_{ab} - \mathbf{y}_a\|^2 / 2\ddot{\sigma}_a^2 \qquad \mathbf{p}_{ab} \sim f\left(\mathbf{p}_a, \mathbf{x}_{ab} - \mathbf{x}_a\right)$$

patch$(a, b, \mathbf{x}_{ab}, \mathbf{c}_a, \mathbf{p}_a) \rightarrow \{\text{rock}\left(\mathbf{x}_{abc}, \mathbf{c}_{abc}, s_{abc}\right) \mid a \in A, b \in B_a, c \in C_{ab}\}$

$$\sum_{ab} \|\mathbf{c}_{abc} - \mathbf{c}_{ab}\|^2 / 2\sigma_4^2 + \sum_{ab} \|\mathbf{y}_{abc} - \mathbf{y}_{ab}\|^2 / 2\sigma_5^2$$

$$s_{abc} \sim \text{sizedistr}\left(\mathbf{p}_a\right)$$

distractors $\rightarrow \{\text{rock}\left(\mathbf{x}_{00d}, \mathbf{c}_{00d}, s_{00d}\right) \mid d \in D\}$

$$\sum_{ab} \|\mathbf{c}_{00d}\|^2 / 2\sigma_0^2 + \sum_{ab} \|\mathbf{x}_{00d}\|^2 / 2\sigma_3^2$$

$$s_{00d} \sim \text{sizedistr}\left(\ddot{\mathbf{p}}\right)$$

$\{\text{rock}\left(\mathbf{x}_{abc}, \mathbf{c}_{abc}, s_{abc}\right) \mid a \in A', b \in B', c \in C_{ab}\} \rightarrow \{\text{visible\_rock}\left(\mathbf{x}_i, \mathbf{c}_i, s_i\right) \mid i \in I\}$

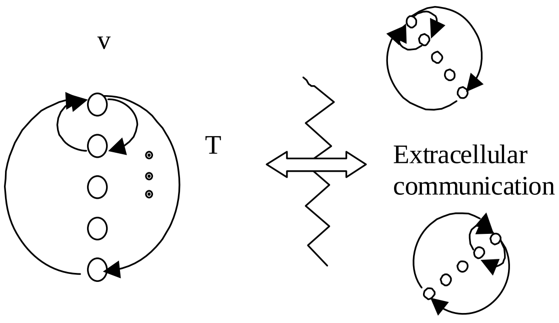
$$\{P_{i,abc}\} \sim \text{uniform\_permutations}\left(\sum_{ic} P_{i,abc} = n_{ab}\right)$$

$$\mathbf{x}_i = \sum_{ab} P_{i,abc} \mathbf{x}_{abc}$$

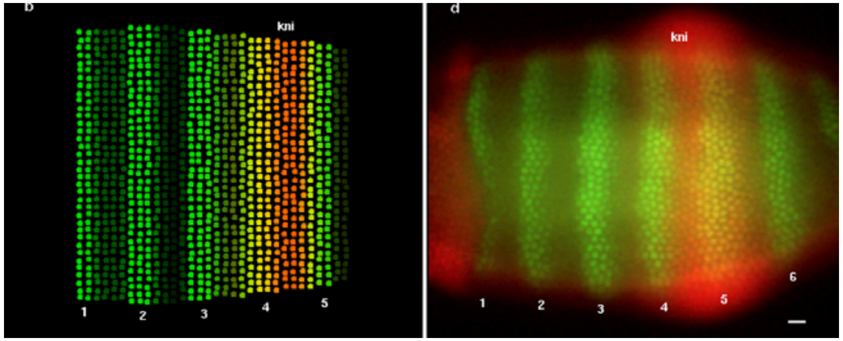
$$\underset{MFT}{\Rightarrow} E = \sum_a \|\mathbf{c}_a\|^2 / 2\sigma_0^2 + \sum_a \|\mathbf{y}_a\|^2 / 2\sigma_0^2 + \sum_a \|\mathbf{p}_a - \ddot{\mathbf{p}}\|^2 / 2\sigma_1^2 + \log^2\left(\ddot{\sigma}_a / \sigma\right)$$
$$+ \sum_a \|\mathbf{y}_{ab} - \mathbf{y}_a\|^2 / 2\ddot{\sigma}_a^2 + \sum_{iab} P_{iab}\left[\|\mathbf{c}_i - \mathbf{c}_a\|^2 / 2\sigma_4^2 + \|\mathbf{x}_i - \mathbf{y}_{ab}\|^2 / 2\sigma_5^2\right]$$

Transcriptional Gene Regulation Networks

- Gene Regulation Network (GRN) model



[Mjolsness et al. J. Theor. Biol. 152: 429-453, 1991]



Drosophila *eve* stripe expression in model (right) and data (left). Green: *eve* expression, red: *kni* expression.
From [Reinitz and Sharp, Mech. of Devel., 49:133-158,

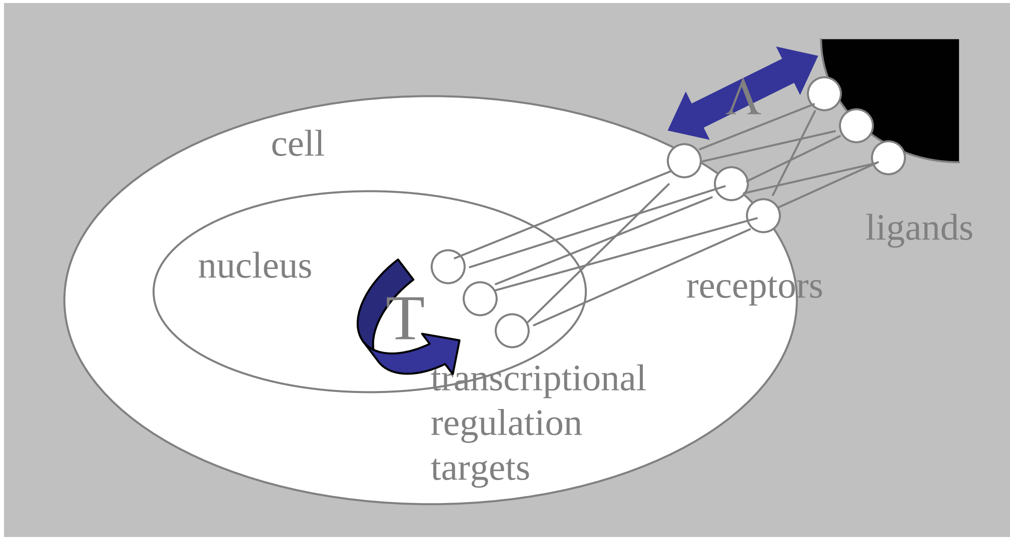
1995].

Gene Regulation + Signal Transduction Network

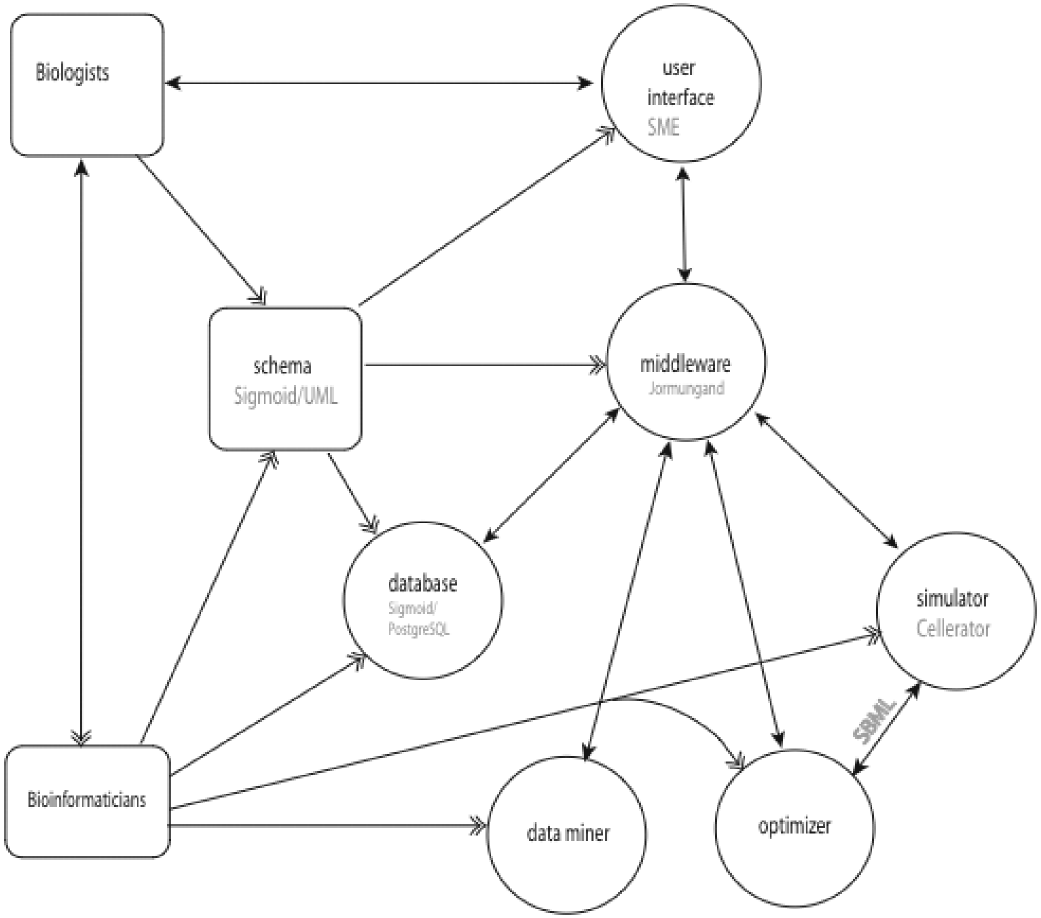
$$\frac{d}{dt} v_a(t) = \frac{1}{\tau_a}\left[g(u_a + h_a) - \lambda_a v_a \right],$$

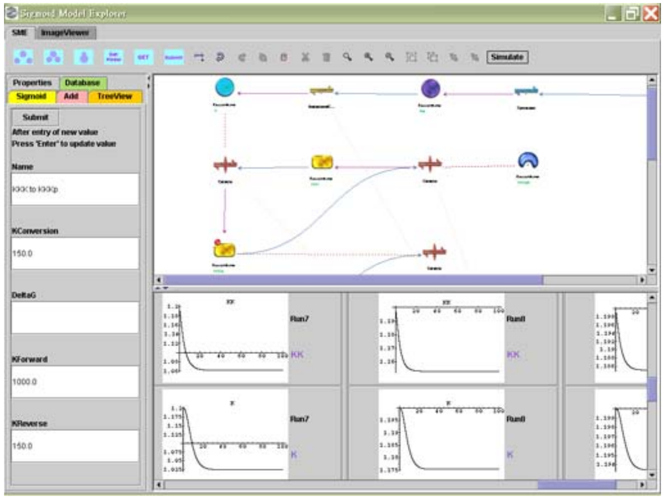
where

$$u_a(t) = \sum_b T_{ab} v_b(t) + \sum_{I \in Nbrs} \Lambda^I \sum_b \ddot{T}_{ab} v_b^I(t) + \sum_{I \in Nbrs} \Lambda^I \sum_b \sum_c \tilde{T}^1_{ac} \tilde{T}^2_{cb} v_c(t) v_b^I(t)$$



Software architectures for systems biology: Sigmoid & Cellerator





After entry of new value
Press 'Enter' to update value

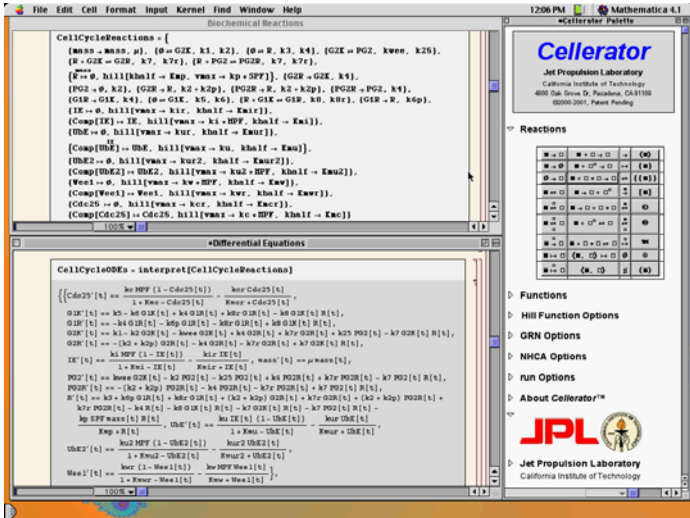
Name
KKK to KKKp

KConversion
150.0

DeltaG

KForward
1000.0

KReverse
150.0



```
CellCycleReactions = {
  {mass → mass, μ}, {∅ ⇌ G2K, k1, k2}, {∅ ⇌ R, k3, k4}, {G2K ⇌ PG2, kwee, k25},
  {R + G2K ⇌ G2R, k7, k7r}, {R + PG2 ⇌ PG2R, k7, k7r},
  {R ↦ ∅, hill[khalf → Kmp, vmax → kp * SPF]}, {G2R → G2K, k4},
  {PG2 → ∅, k2}, {G2R → R, k2 + k2p}, {PG2R → R, k2 + k2p}, {PG2R → PG2, k4},
  {G1R → G1K, k4}, {∅ ⇌ G1K, k5, k6}, {R + G1K ⇌ G1R, k8, k8r}, {G1R → R, k6p},
  {IE ↦ ∅, hill[vmax → kir, khalf → Kmir]},
  {Comp[IE] ↦ IE, hill[vmax → ki * MPF, khalf → Kmi]},
  {UbE ↦ ∅, hill[vmax → kur, khalf → Kmur]},
  {Comp[UbE] ↦ UbE, hill[vmax → ku, khalf → Kmu]},
  {UbE2 ↦ ∅, hill[vmax → kur2, khalf → Kmur2]},
  {Comp[UbE2] ↦ UbE2, hill[vmax → ku2 * MPF, khalf → Kmu2]},
  {Wee1 ↦ ∅, hill[vmax → kw * MPF, khalf → Kmw]},
  {Comp[Wee1] ↦ Wee1, hill[vmax → kwr, khalf → Kmwr]},
  {Cdc25 ↦ ∅, hill[vmax → kcr, khalf → Kmcr]},
  {Comp[Cdc25] ↦ Cdc25, hill[vmax → kc * MPF, khalf → Kmc]}
```

```
CellCycleODEs = interpret[CellCycleReactions]
```

$$\Big\{\Big\{\mathrm{Cdc25}'[t] == \frac{\mathrm{kc\,MPF}\,(1-\mathrm{Cdc25}[t])}{1+\mathrm{Kmc}-\mathrm{Cdc25}[t]} - \frac{\mathrm{kcr\,Cdc25}[t]}{\mathrm{Kmcr}+\mathrm{Cdc25}[t]},$$

$$\mathrm{G1K}'[t] == \mathrm{k5} - \mathrm{k6\,G1K}[t] + \mathrm{k4\,G1R}[t] + \mathrm{k8r\,G1R}[t] - \mathrm{k8\,G1K}[t]\,\mathrm{R}[t],$$

$$\mathrm{G1R}'[t] == -\mathrm{k4\,G1R}[t] - \mathrm{k6p\,G1R}[t] - \mathrm{k8r\,G1R}[t] + \mathrm{k8\,G1K}[t]\,\mathrm{R}[t],$$

$$\mathrm{G2K}'[t] == \mathrm{k1} - \mathrm{k2\,G2K}[t] - \mathrm{kwee\,G2K}[t] + \mathrm{k4\,G2R}[t] + \mathrm{k7r\,G2R}[t] + \mathrm{k25\,PG2}[t] - \mathrm{k7\,G2K}[t]\,\mathrm{R}[t],$$

$$\mathrm{G2R}'[t] == -(\mathrm{k2}+\mathrm{k2p})\,\mathrm{G2R}[t] - \mathrm{k4\,G2R}[t] - \mathrm{k7r\,G2R}[t] + \mathrm{k7\,G2K}[t]\,\mathrm{R}[t],$$

$$\mathrm{IE}'[t] == \frac{\mathrm{ki\,MPF}\,(1-\mathrm{IE}[t])}{1+\mathrm{Kmi}-\mathrm{IE}[t]} - \frac{\mathrm{kir\,IE}[t]}{\mathrm{Kmir}+\mathrm{IE}[t]},\ \mathrm{mass}'[t] == \mu\,\mathrm{mass}[t],$$

$$\mathrm{PG2}'[t] == \mathrm{kwee\,G2K}[t] - \mathrm{k2\,PG2}[t] - \mathrm{k25\,PG2}[t] + \mathrm{k4\,PG2R}[t] + \mathrm{k7r\,PG2R}[t] - \mathrm{k7\,PG2}[t]\,\mathrm{R}[t],$$

$$\mathrm{PG2R}'[t] == -(\mathrm{k2}+\mathrm{k2p})\,\mathrm{PG2R}[t] - \mathrm{k4\,PG2R}[t] - \mathrm{k7r\,PG2R}[t] + \mathrm{k7\,PG2}[t]\,\mathrm{R}[t],$$

$$\mathrm{R}'[t] == \mathrm{k3} + \mathrm{k6p\,G1R}[t] + \mathrm{k8r\,G1R}[t] + (\mathrm{k2}+\mathrm{k2p})\,\mathrm{G2R}[t] + \mathrm{k7r\,G2R}[t] + (\mathrm{k2}+\mathrm{k2p})\,\mathrm{PG2R}[t] + \mathrm{k7r\,PG2R}[t] - \mathrm{k4\,R}[t] - \mathrm{k8\,G1K}[t]\,\mathrm{R}[t] - \mathrm{k7\,G2K}[t]\,\mathrm{R}[t] - \mathrm{k7\,PG2}[t]\,\mathrm{R}[t] - \frac{\mathrm{kp\,SPF\,mass}[t]\,\mathrm{R}[t]}{\mathrm{Kmp}+\mathrm{R}[t]},\ \mathrm{UbE}'[t] == \frac{\mathrm{ku\,IE}[t]\,(1-\mathrm{UbE}[t])}{1+\mathrm{Kmu}-\mathrm{UbE}[t]} - \frac{\mathrm{kur\,UbE}[t]}{\mathrm{Kmur}+\mathrm{UbE}[t]},$$

$$\mathrm{UbE2}'[t] == \frac{\mathrm{ku2\,MPF}\,(1-\mathrm{UbE2}[t])}{1+\mathrm{Kmu2}-\mathrm{UbE2}[t]} - \frac{\mathrm{kur2\,UbE2}[t]}{\mathrm{Kmur2}+\mathrm{UbE2}[t]},$$

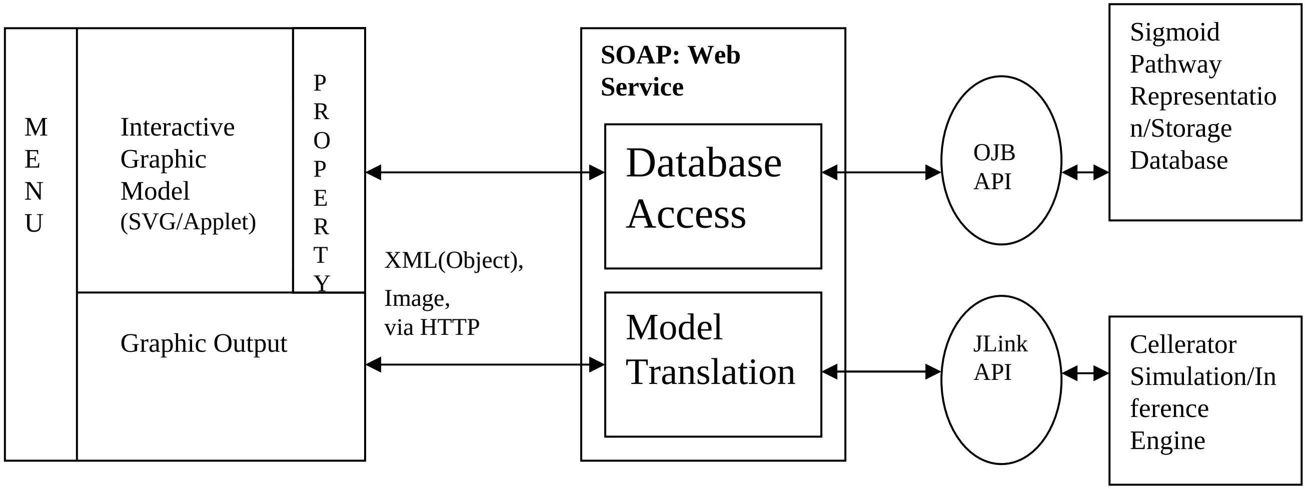
$$\mathrm{Wee1}'[t] == \frac{\mathrm{kwr}\,(1-\mathrm{Wee1}[t])}{1+\mathrm{Kmwr}-\mathrm{Wee1}[t]} - \frac{\mathrm{kw\,MPF\,Wee1}[t]}{\mathrm{Kmw}+\mathrm{Wee1}[t]}\Big\},$$

Cellerator

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Dr, Pasadena, CA 91109
©2000-2001, Patent Pending

- Reactions
- Functions
- Hill Function Options
- GRN Options
- NHCA Options
- run Options
- About Cellerator™
- Jet Propulsion Laboratory
  California Institute of Technology

3-tier architecture


M
E
N
U
Interactive Graphic Model (SVG/Applet)
P
R
O
P
E
R
T
Y
Graphic Output
XML(Object),
Image,
via HTTP
SOAP: Web Service
Database Access
Model Translation
OJB API
JLink API
Sigmoid Pathway Representation/Storage Database
Cellerator Simulation/Inference Engine

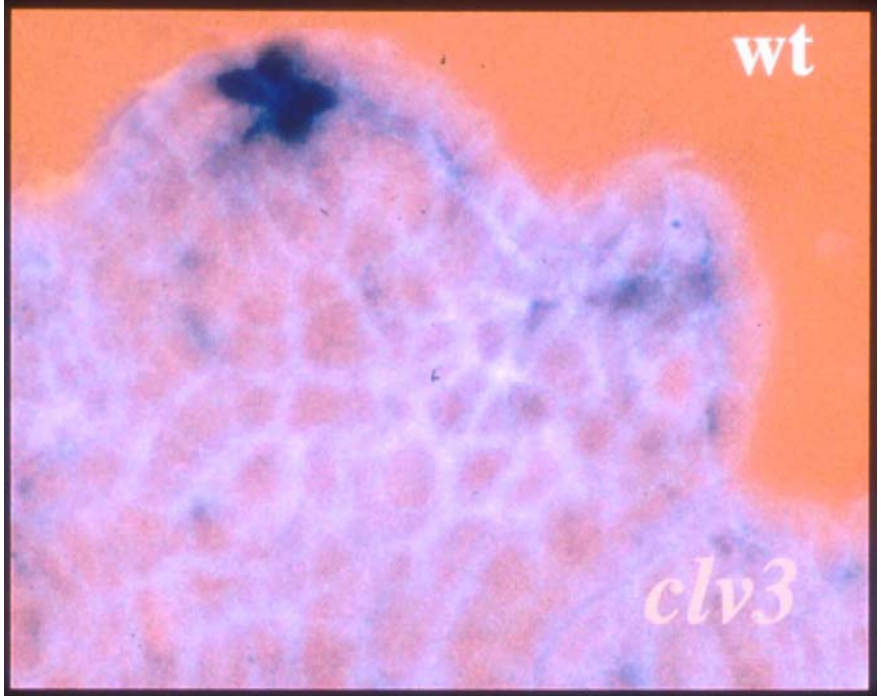
Possible software support

- Machine learning (open source/academic)
  - CompClust (CIT/JPL):
    - Scripting/GUI dichotomy data point;
    - dataset views
  - WEKA data mining
  - Intel: PNL Probabilistic Networks Library
  - Future: stochastic grammar modeler
    - + autogeneration (as in Cellerator)
- Image processing, data environments
  - Matlab, IDL, Mathematica, Khoros/VisiQuest, …
  - NIHImage/ImageJ, …

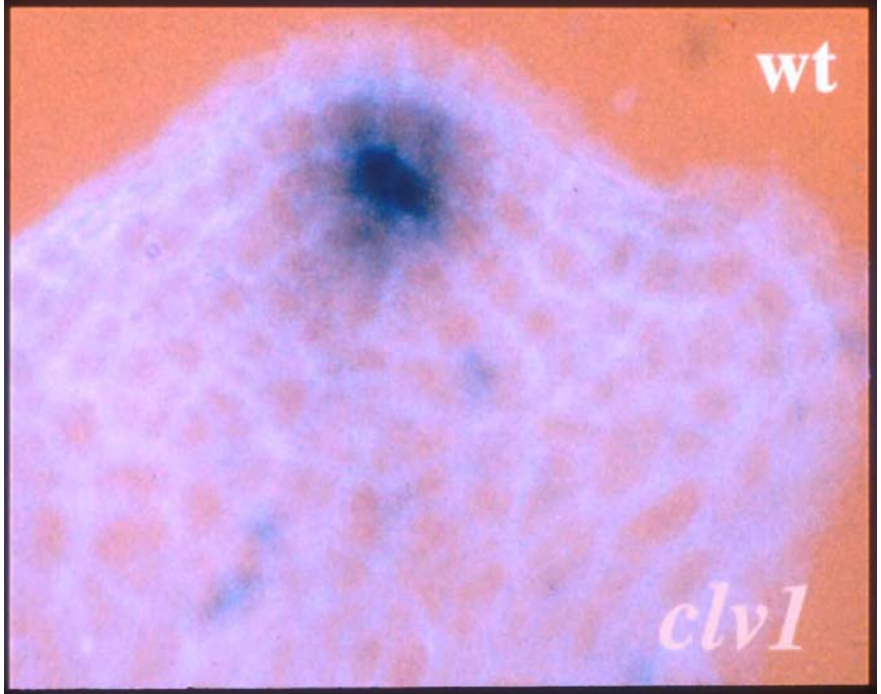
Metadata in Systems Biology

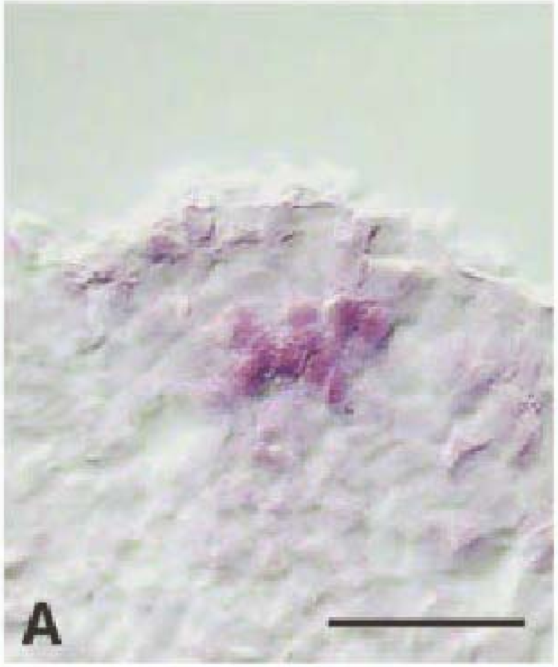
- SBML
- Sigmoid UML

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.



Fletcher et al., Science v. 283, 1999

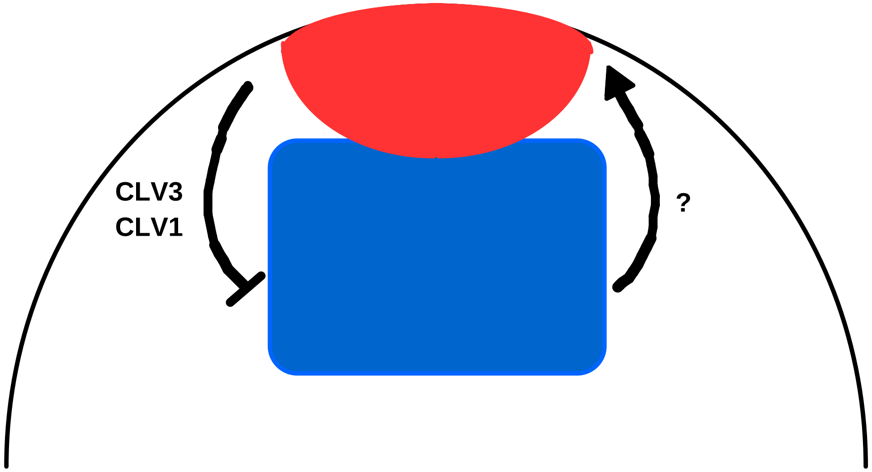




SISL

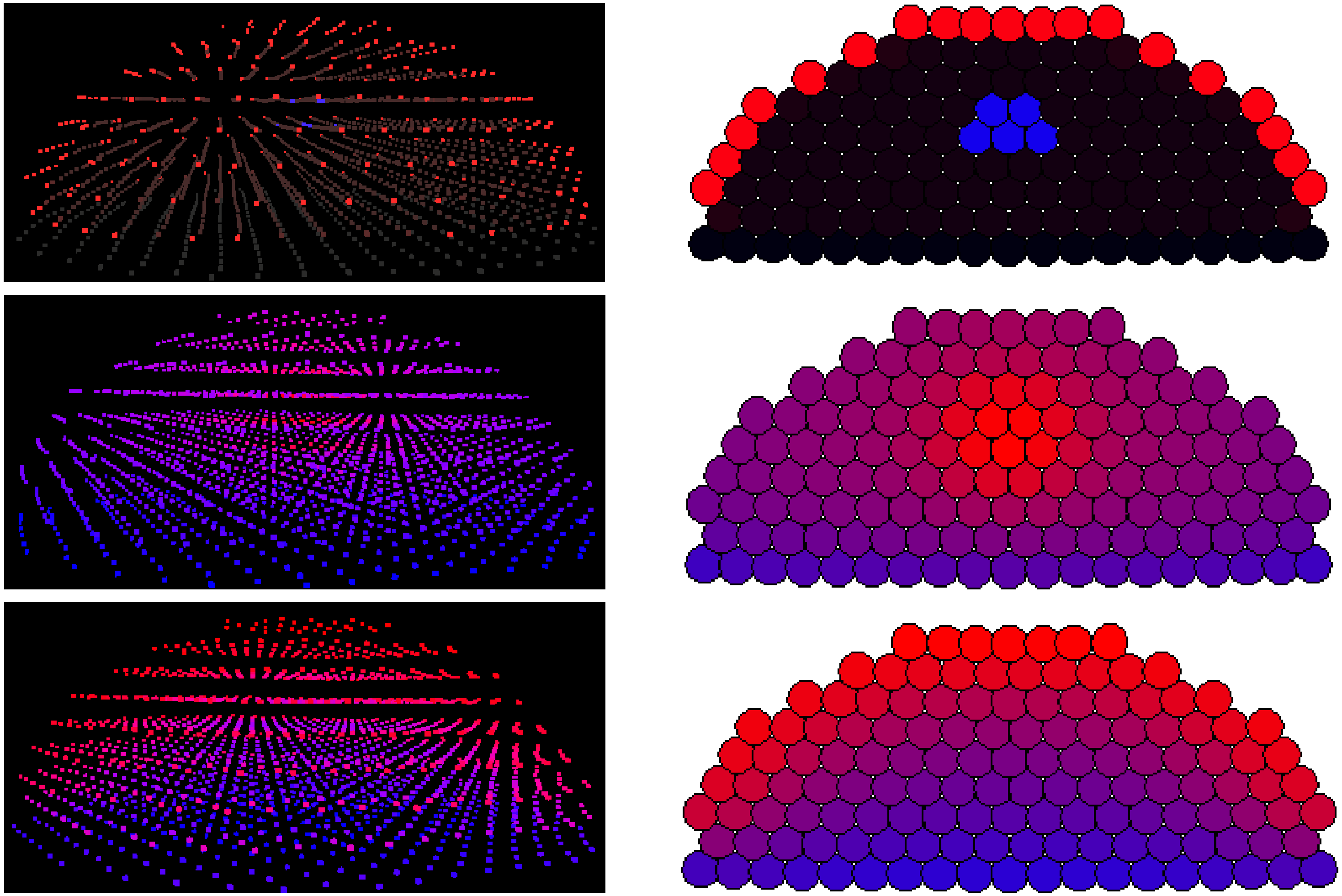
WUS

Brand et. al., Science **289**, 617-619, (2000)



SAM gene network: Results

protein concentrations

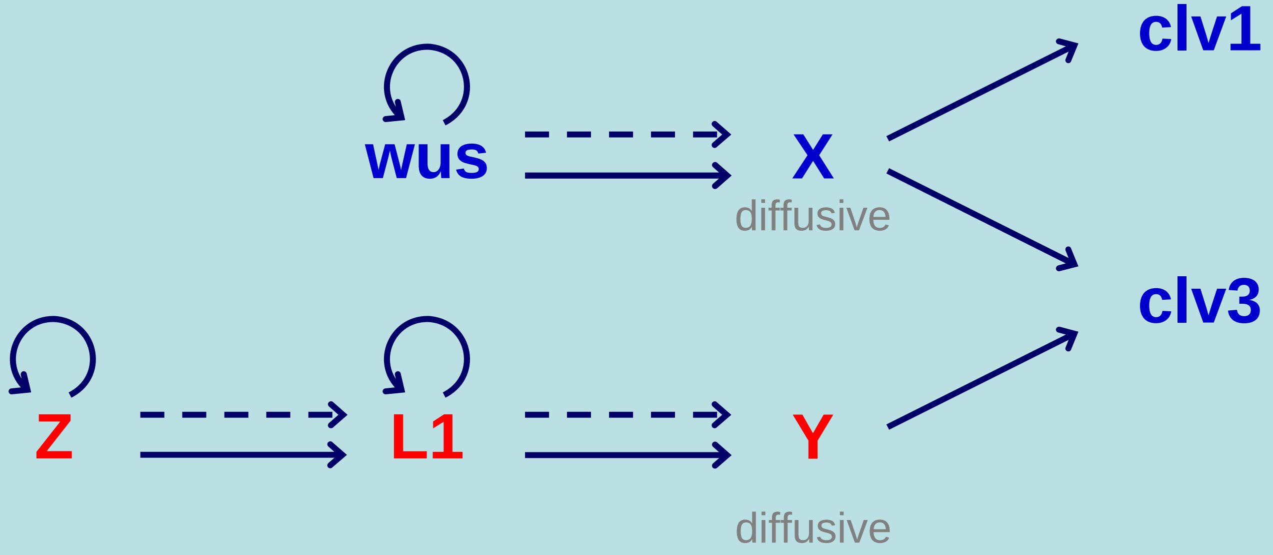


wus(init) and **L1**

X

Y

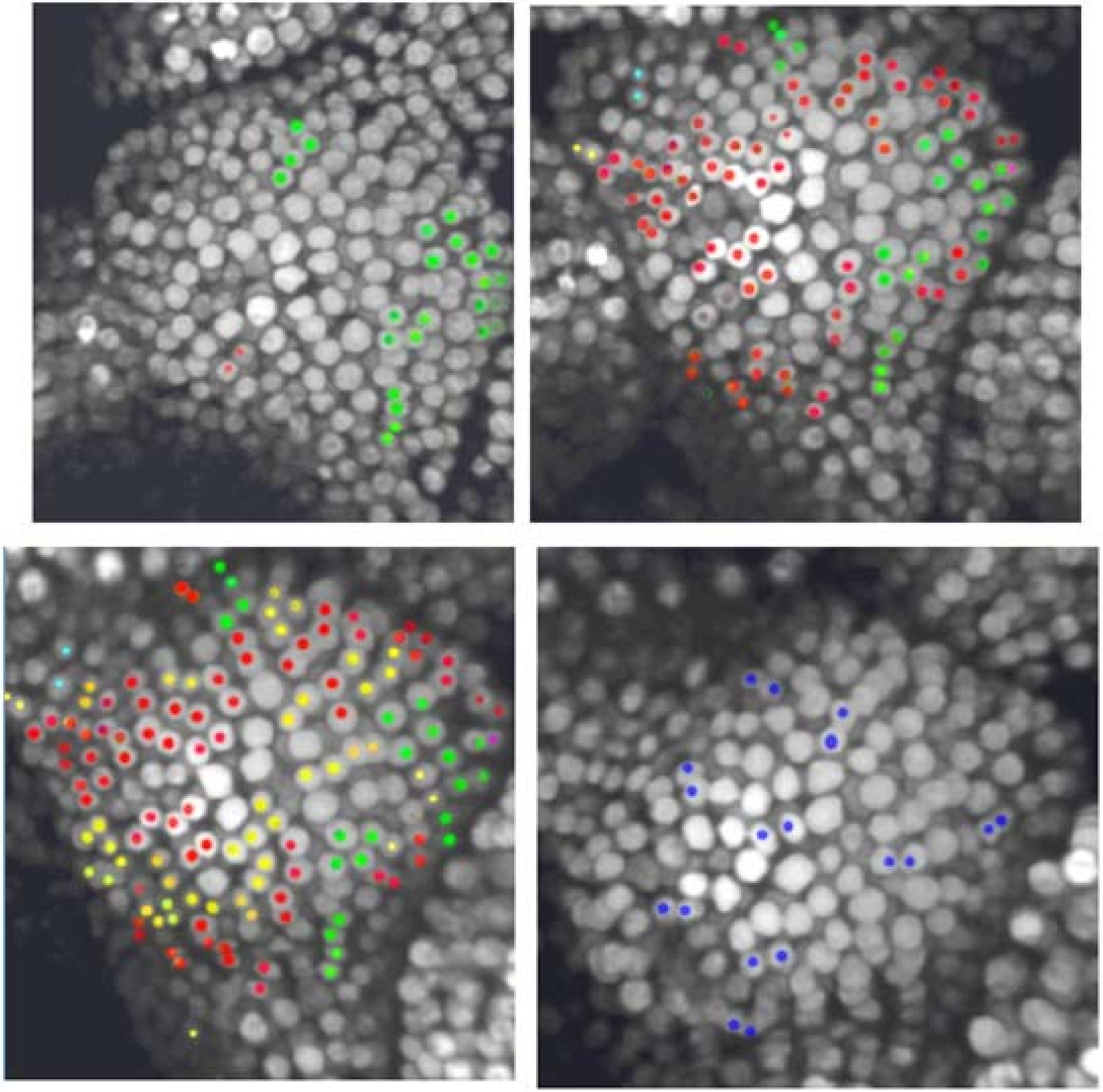
SAM: Gene Network Model



SAM
growth
imagery
PIN1
cell
walls

QuickTime™ and a
TIFF decompressor
are needed to see this picture.

IGB *SISL*



Venu Gonehal

Basic Machine Learning Distinctions

- Supervised vs. unsupervised learning
    - Supervised e.g. classification and regression
        - Feature selection
        - regression for phenomenological model fitting e.g. GRN’s
    - Unsupervised e.g. clustering; may be preprocessor
- Generative vs. Kernal methods
    - Generative (statistical inference) models
    - Kernal methods e.g Support Vector Machines
- Vector vs. Relationship data
    - Vector data: preprocessed image features Δlog I, Δx, …
    - Images, time series, shifted spectra - semigroup actions
    - Sparse graph/relationship data - permutation actions

Contacts

- Wayne Hayes, UCI ICS faculty
  - scientific computing
- UCI ICS Maching Learning
  - Padhraic Smyth
  - Pierre Baldi
- Chris Hart, Caltech Biology grad student