Chemical Space: Modeling Exploration & Understanding

Rajarshi Guha

School of Informatics
Indiana University

16th August, 2006

Outline

1. Overview
2. Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information
3. Tool Development
   - CDK
   - R

Outline

1. Overview

2. Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information

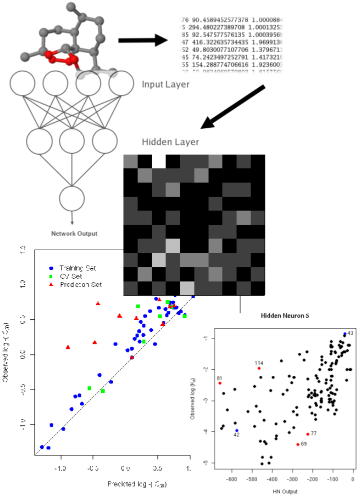
3. Tool Development
   - CDK
   - R

Goals of Cheminformatics Research at IU

- Extend the state of the art in cheminformatics
- Extract chemistry, not just R^2, q^2 et al.
- Provide intuitive and efficient ways to handle chemical information
- Supply expertise to bench chemists and non-computational chemists

Statistical Modeling of Chemical Information

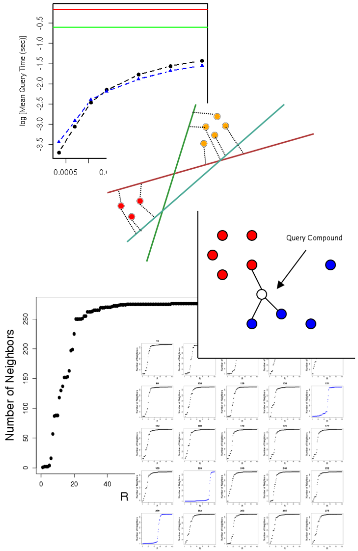
- QSAR model development
  - Lazy regression
  - Ensemble descriptor selection
  - Wavelet-based spectral descriptors
- Interpretation of QSAR models
- Measuring model applicability



```
76 90.4589452577378 1.000088
5 294.480227389708 1.00013253
85 92.547577576135 1.00039569
47 416.322635734435 1.9699136
52 49.8030077107706 1.3796711
45 74.2423497252791 1.4173219
55 154.288774706616 1.9236003
```

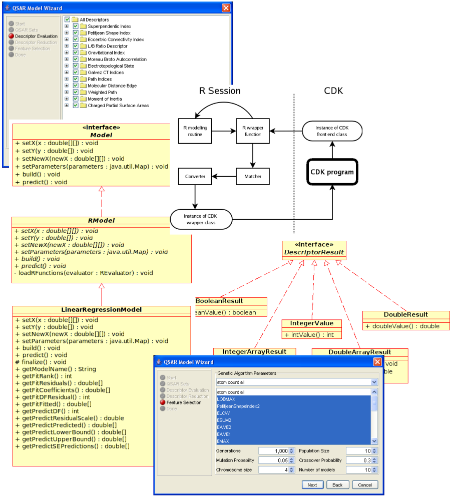
Cheminformatics Algorithms

- R-NN curves
  - Outlier detection
  - Cluster cardinality
- LSH based applications
- Ensemble descriptor selection
- Interpretation techniques



Tools and Pipelines for Cheminformatics

- Contributions to the CDK
- Packages for R
  - rcdk
  - spe
  - fingerprint
  - spectral clustering
  - rpubchem (to come)
- Automated QSAR pipeline (collaboration with P&G)
- Development of workflows



QSAR Model Wizard

- Start
- QSAR Sets
- ● Descriptor Evaluation
- Descriptor Reduction
- Feature Selection
- Done

☑ All Descriptors
- ☑ Superpendentic Index
- ☑ Petitjean Shape Index
- ☑ Eccentric Connectivity Index
- ☑ L/B Ratio Descriptor
- ☑ Gravitational Index
- ☑ Moreau Broto Autocorrelation
- ☑ Electrotopological State
- ☑ Galvez CT Indices
- ☑ Path Indices
- ☑ Molecular Distance Edge
- ☑ Weighted Path
- ☑ Moment of Inertia
- ☑ Charged Partial Surface Areas

```
«interface»
Model
+ setX(x : double[][]) : void
+ setY(y : double[]) : void
+ setNewX(newX : double[][]) : void
+ setParameters(parameters : java.util.Map) : void
+ build() : void
+ predict() : void
```

```
RModel
+ setX(x : double[][]) : void
+ setY(y : double[]) : void
+ setNewX(newX : double[][]) : void
+ setParameters(parameters : java.util.Map) : void
+ build() : void
+ predict() : void
- loadRFunctions(evaluator : REvaluator) : void
```

```
LinearRegressionModel
+ setX(x : double[][]) : void
+ setY(y : double[]) : void
+ setNewX(newX : double[][]) : void
+ setParameters(parameters : java.util.Map) : void
+ build() : void
+ predict() : void
# finalize() : void
+ getModelName() : String
+ getFitRank() : int
+ getFitResiduals() : double[]
+ getFitCoefficients() : double[]
+ getFitDFResidual() : int
+ getFitFitted() : double[]
+ getPredictDF() : int
+ getPredictResidualScale() : double
+ getPredictPredicted() : double[]
+ getPredictLowerBound() : double[]
+ getPredictUpperBound() : double[]
+ getPredictSEPredictions() : double[]
```

```
«interface»
DescriptorResult

BooleanResult
eanValue() : boolean

IntegerValue
+ intValue() : int

DoubleResult
+ doubleValue() : double

IntegerArrayResult

DoubleArrayResult
void
ble
```

QSAR Model Wizard

- Start
- QSAR Sets
- Descriptor Evaluation
- Descriptor Reduction
- ● Feature Selection
- Done

Genetic Algorithm Parameters

atom count all

atom count all
LOBMAX
PetitjeanShapeIndex2
ELOW
ESUM2
EAVE2
EAVE1
EMAX

Generations	1,000	Population Size	10
Mutation Probability	0.05	Crossover Probability	0.3
Chromosome size	4	Number of models	10

Outline

1. Overview

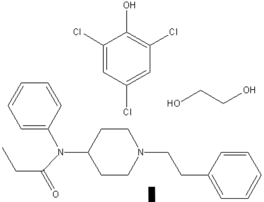
2. Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information

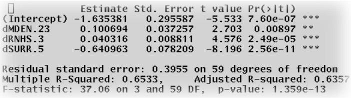
3. Tool Development
   - CDK
   - R

QSAR Model Interpretation

Why interpret?

- Predictive ability is useful for screening
- Interpretation provides extra value
- Interpretability might be more useful than predictive ability
- Interpretability depends on modeling technique and descriptors involved





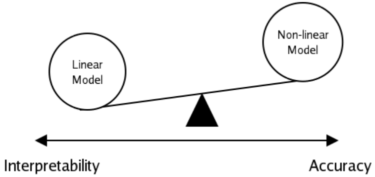
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.635381   0.295587  -5.533 7.60e-07 ***
dMDEN.23     0.100694   0.037257   2.703  0.00897 **
dRNHS.3      0.040316   0.008811   4.576 2.49e-05 ***
dSURR.5     -0.640963   0.078209  -8.196 2.56e-11 ***

Residual standard error: 0.3955 on 59 degrees of freedom
Multiple R-Squared: 0.6533,     Adjusted R-squared: 0.6357
F-statistic: 37.06 on 3 and 59 DF,  p-value: 1.359e-13
```

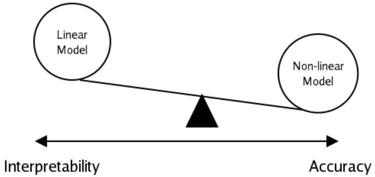
?

QSAR Model Interpretation

The trade-offs in interpretation

- Interpretability is usually a trade-off with accuracy
- OLS models are easily interpretable, not always accurate
- CNN models are usually black boxes, more accurate
- Some methods (RF) lie in between





Detailed Interpretation Methods

OLS Models

- Build a *good* model
- Run it through PLS
- The X-loadings indicate which descriptors are important
    - magnitude lets us rank them
    - sign lets us indicate the nature of their effects
- Allows us to identify effects of descriptors on a *molecule-wise* basis

CNN Models

- Analogous to PLS based interpretations
- *Linearizes* the CNN
- Information is lost
- Resultant interpretations match the corresponding interpretations for an OLS model quite well

Guha, R. et al., *J. Chem. Inf. Model.*, **2005**, *45*, 321–333

Guha, R. et al., *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1440–1449

Extensions of Interpretation Methods

- The CNN interpretation uses significant approximations and does not make full use of biases
- Develop interpretation techniques for other methods, RF in particular
- Ensemble descriptor selection to choose a subset that is *simultaneously* good for multiple model types
- Use these methods on real datasets

QSAR Model Applicability

Model Validation

- Goal is to test the reliability of the model
- Ensures that the model is not due to chance factors
- Based on dataset used to develop the model

Model Applicability

- Goal is to test the applicability of the model to new compounds
- Tells us: The model will predict the activity well (or not)
- Similar to confidence measures

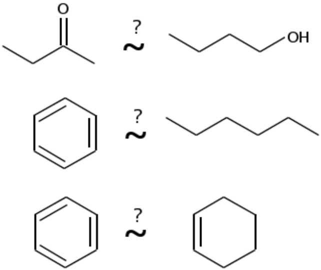
What is Model Applicability

Question?

How will a model perform when faced with molecules that it has not been trained on or validated with?

Aspects

- Similarity to the TSET?
- Can we consider a *global* chemistry space?
- Structural or statistical similarity?
- Quantitative or qualitative?



How to Assess Model Applicability

Define Model Performance

Performance is measured by prediction residuals. The model performs well on a new molecule if it predicts its activity with low residual error.

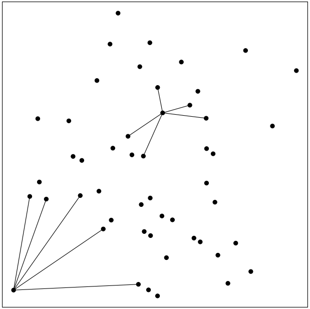
Correlate *'X'* With Performance

- 'X' could be similarity between a query molecule and the original training set
- 'X' could be derived from a cluster membership approach
- Alternatively, *predict performance* itself

Guha, R.; Jurs, P.C; *J. Chem. Inf. Model.*, **2005**, *45*, 65–73

Nearest Neighbor Methods

- Traditional *k*NN methods are simple, fast, intuitive
- Applications in
  - regression & classification
  - diversity analysis
- Can be misleading if the *nearest* neighbor is far away
- *R*-NN methods may be more suitable



R-NN Curves for Diversity Analysis

The question . . .

Does the variation of nearest neighbor count with radius allow us to characterize the location of a query point in a dataset?

The answer . . . *R*-NN curves

```
$D_{max} \leftarrow$ max pairwise distance
for molecule in dataset do
    R $\leftarrow 0.01 \times D_{max}$
    while $R \leq D_{max}$ do
        Find NN's within radius $R$
        Increment $R$
    end while
end for
Plot NN count vs. $R$
```

Guha, R., et al., *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1772

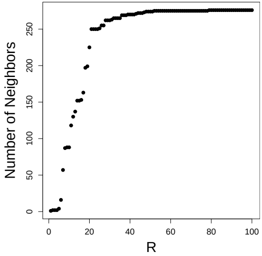
R-NN Curves for Diversity Analysis

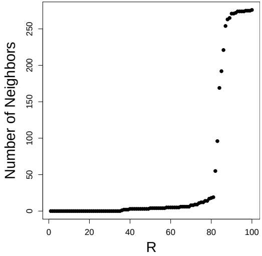
The question . . .

Does the variation of nearest neighbor count with radius allow us to characterize the location of a query point in a dataset?

The answer . . . *R*-NN curves

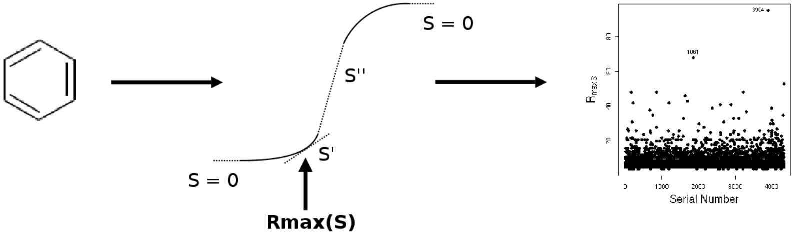
```
$D_{max} \leftarrow$ max pairwise distance
for molecule in dataset do
    R $\leftarrow 0.01 \times D_{max}$
    while $R \leq D_{max}$ do
        Find NN's within radius $R$
        Increment $R$
    end while
end for
Plot NN count vs. $R$
```



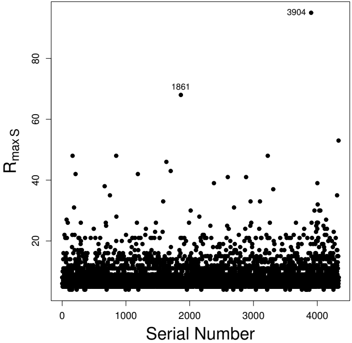


Guha, R., et al., *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1772

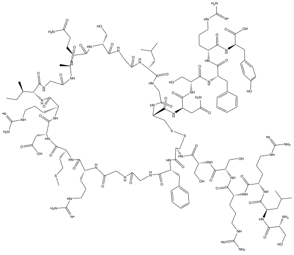
Characterizing *R*-NN Curves



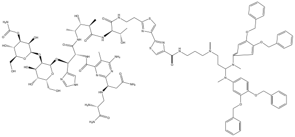
R-NN Curves and Outliers



A single plot identifies the location characteristics of *all* the molecules



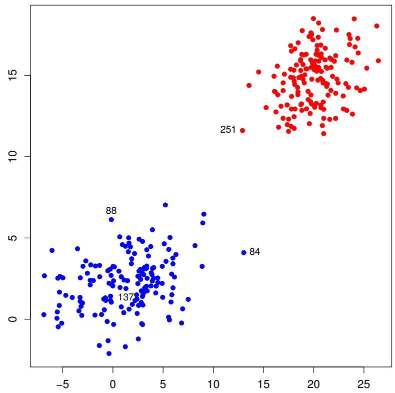
3904

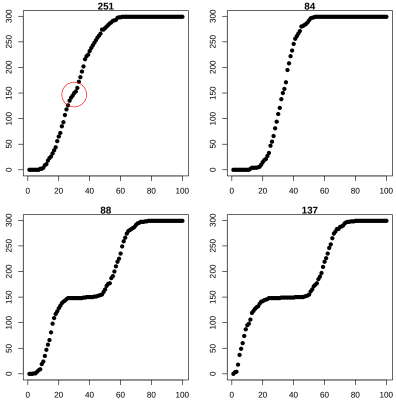


1861

Kazius, J.; McGuire, R.; Bursi, R.; *J. Med. Chem.* **2005,** *48,* 312–320

R-NN Curves and Clusters





Smoothed *R*-NN Curves

R-NN curves are indicative of the number of clusters

R-NN Curves and Clusters

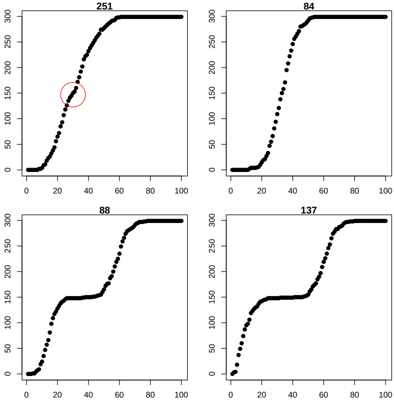
Counting the steps

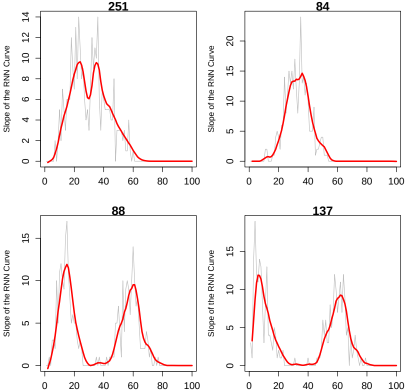
- Essentially a curve matching problem
- All points will not be indicative of the number of clusters
- Not applicable for concentric clusters

Approaches

- Hausdorff / Fréchet distance - requires *canonical* curves
- RMSE from distance matrix
- Slope analysis

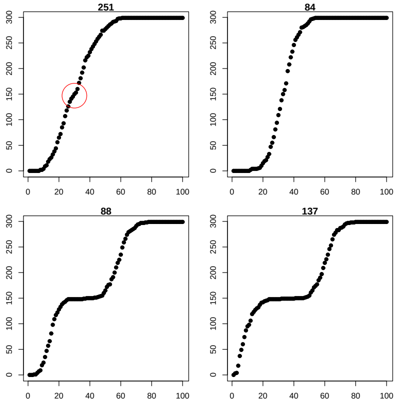
R-NN Curves and Their Slopes

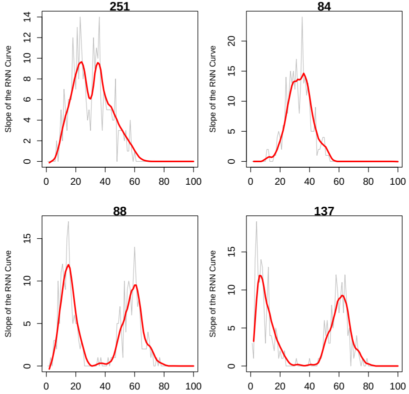




Smoothed first derivative of the *R*-NN Curves

R-NN Curves and Their Slopes





Smoothed first derivative of the *R*-NN Curves

- Identifying peaks identifies the number of clusters
- Automated picking can identify spurious peaks

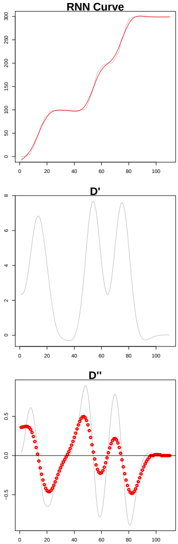
Slope Analysis of *R*-NN Curves

Procedure

```
for i in molecules do
    Evaluate $R$-NN curve
    $F \leftarrow$ smoothed $R$-NN curve
    Evaluate $F''$
    Smooth $F''$
    $N_{root,i} \leftarrow$ no. of roots of $F''$
end for
$N_{cluster} = [\max(N_{root}) + 1]/2$
```

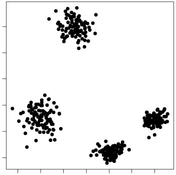
Possible improvements

- Sample from the collection of *R*-NN curves
- Improve handling of concentric clusters

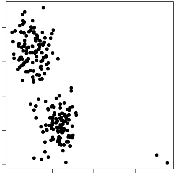


Preliminary Results

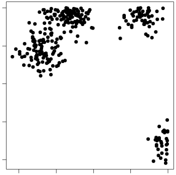
- Simulated 2D data
- Predicted *k*, followed by kmeans clustering using *k*
- Investigated similar values of *k*



k	ASW
4	0.61
3	0.74
5	0.70



k	ASW
3	0.44
2	0.65
4	0.47



k	ASW
4	0.64
3	0.48
5	0.56

ASW - average silhouette width, higher is better; *k* - number of clusters

Ontologies

What is an ontology?

- Defines a controlled vocabulary (keywords)
- Defines a set of relationships between them
- Allows for
  - meaning to be added to data and algorithms
  - automated inference of relationships
- An example is the *Gene Ontology*

Ontologies in Chemistry

What's available?

- No really comprehensive ontology available
- Some work is on to add chemistry semantics to biology-related ontologies

What can we do?

- Start small - focus on one area of chemistry, descriptors
- The CDK currently provides an ontology for implemented descriptors
- Vocabulary includes
  - author
  - literature reference
  - class (molecular, atomic, bond)
  - type (geometric, electronic, . . .)

How Can We Use Ontologies?

Allow interoperability of software

- Different program use different naming schemes
- Programs may be open- or close-source
- Annotation via dictionaries allows us to make conclusions regarding descriptors, suggest similar descriptors etc.

Utilize expertise from chemists

- Some descriptors are useful for certain properties
- Chemists who use models can *tag* descriptors as useful for a property
- Over time the tagging can be indicative of the utility of certain descriptors

Outline

1 Overview

2 Modeling & Algorithms
- Aspects of QSAR Modeling
- Exploring Chemical Space
- Adding Meaning to Chemical Information

3 Tool Development
- CDK
- R

The Chemistry Development Kit

What does it do?

- Reads multiple file formats
- Calculates molecular descriptors (in progress)
  - Topologicals (Chi, Kappa, . . .)
  - Geometric (Gravitational indices, MI, . . .)
  - Hybrid (CPSA)
  - Global (BCUT, WHIM, RDF, . . .)
- Provides access to statistical engines (R & Weka)
- Fingerprint calculation, 2D structure diagrams
- Kabsch alignment
- 3D coordinates and force fields (in progress)

Steinbeck, C. et al., *Curr. Pharm. Des.*, **2006**, *12*, 2111–2120

Steinbeck, C. et al., *J. Chem. Inf. Sci.*, **2003**, *43*, 493–500

Contributions

Main areas of contributions

- Descriptor framework
- QSAR modeling framework (interface to R)
- Web service functionality
- Other areas include
    - Numerical surface areas
    - Rigid alignment
    - Descriptor ontology
    - Build system
    - QA, debugging, support

`http://almost.cubic.uni-koeln.de/cdk/cdk_top`

Cheminformatics and R

What is R?

- An open-source statistical *environment* for
  - model development
  - algorithm prototyping
- Open source version of Splus
  - Splus code runs (mostly) unchanged on R
- Provides a wide array of mathematical and statistical functionality
  - Linear models (OLS, robust regression, GLM, PLS)
  - Neural networks, random forests, SOM, SVM
  - Clustering methods (kmeans, agnes, pam, . . .)
  - Optimization routines
  - Database interfaces
- When working with chemical data it would be nice to have access to cheminformatics functionality inside R

Cheminformatics and R

What is R?

- An open-source statistical *environment* for
  - model development
  - algorithm prototyping
- Open source version of Splus
  - Splus code runs (mostly) unchanged on R
- Provides a wide array of mathematical and statistical functionality
  - Linear models (OLS, robust regression, GLM, PLS)
  - Neural networks, random forests, SOM, SVM
  - Clustering methods (kmeans, agnes, pam, . . .)
  - Optimization routines
  - Database interfaces
- When working with chemical data it would be nice to have access to cheminformatics functionality inside R

Some Cheminformatics Related Packages

Connecting R and the CDK

- R can access Java code via the SJava package
- This allows us to use CDK functionality within R
- The `rcdk` package provides user friendly wrappers to CDK classes and methods
- The result is that we can stay inside R and handle molecules directly

Fingerprints

- The `fingerprint` package handles binary fingerprint data from CDK and MOE
- Calculates similarity matrices using the Tanimoto metric
- Converts binary fingerprints to Euclidean vectors

Guha, R., *CDK News*, **2005**, *2*, 2–6

R Miscellanea

R as a Web Service

- A number of packages are available to access R via the web
- An effort is also underway to provide an explicit SOAP interface
- Very simple to access a remote R process via RServe
  - Currently under investigation as our statistical backend for workflows

Summary

Broadly focused on 2 areas . . .

- Modeling
  - Predictive model development
  - Model interpreation and applicability
- Algorithm development
  - Exploring nearest neighbor methods
  - Descriptor selection and interpretation
  - Dictionaries and ontologies
- Underlying motiviation is the extraction of *chemistry* from the numbers and making it available

Collaborations . . .

- More brains are useful
- Always on the lookout for data

Summary

Broadly focused on 2 areas . . .

- Modeling
  - Predictive model development
  - Model interpreation and applicability
- Algorithm development
  - Exploring nearest neighbor methods
  - Descriptor selection and interpretation
  - Dictionaries and ontologies
- Underlying motiviation is the extraction of *chemistry* from the numbers and making it available

Collaborations . . .

- More brains are useful
- Always on the lookout for data

Summary

Broadly focused on 2 areas . . .

- Modeling
  - Predictive model development
  - Model interpreation and applicability
- Algorithm development
  - Exploring nearest neighbor methods
  - Descriptor selection and interpretation
  - Dictionaries and ontologies
- Underlying motiviation is the extraction of *chemistry* from the numbers and making it available

Collaborations . . .

- More brains are useful
- Always on the lookout for data