

What is analytical informatics?

*Randy Julian
Lilly Research Laboratories*

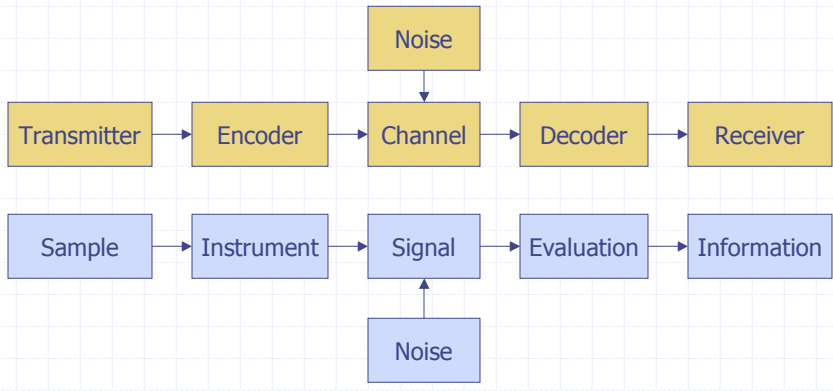
1

Information Theory and Analytical Chemistry

- ◆ The analytical process attempts to obtain information on chemical systems
- ◆ Information is encoded in nature itself
- ◆ Some information can be accessed directly
- ◆ Measuring is a sampling and translation step
- ◆ Measured values are still encoded signals
- ◆ Decoding is required to interpret information
- ◆ Prediction is based on understanding relationships between units of information

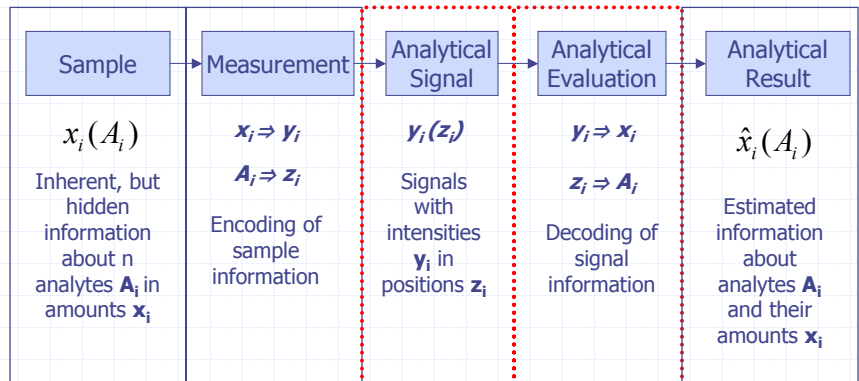
[1]

Comparison



[1]

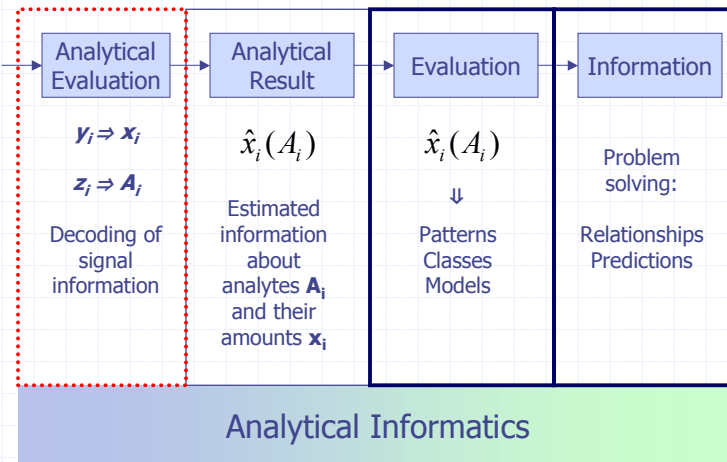
Encoding/Decoding: Part 1



Physics and Analytical Chemistry

[1]

Encoding/Decoding: Part 2



[1]

Machine evaluation of analytical results *(biased, but useful definitions)*

- ◆ Patterns, form, order are the way we "understand" anything – they allow prediction
- ◆ Pattern Recognition: Automatic grouping, description, and classification of patterns.
- ◆ Machine Learning: A method for getting computers to do pattern recognition

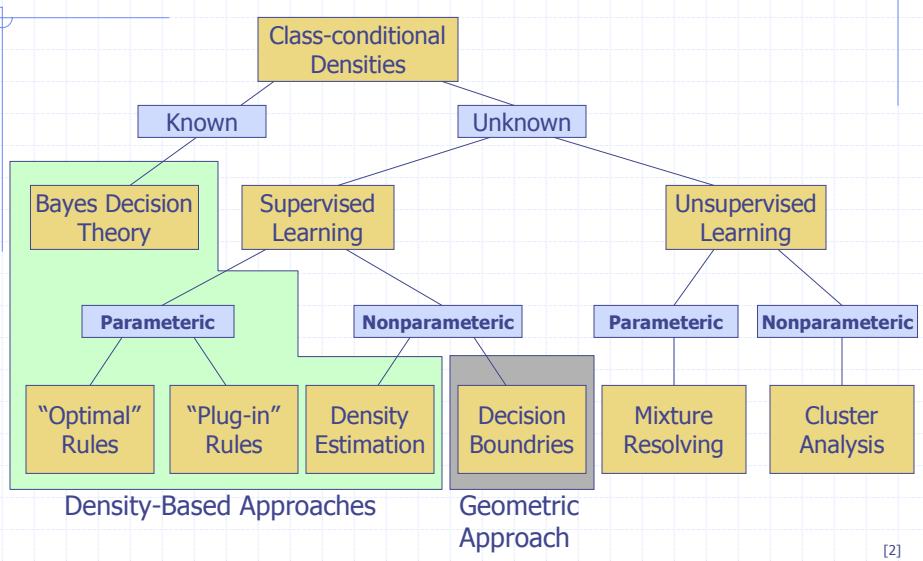
[2]

Examples of Machine Learning

Problem Domain	Application	Input Pattern	Pattern Classes
Bioinformatics	Sequence analysis	DNA/Protein sequence	Known types of genes/patterns
Data mining	Marketing & sales	Sales data	"Hot" product & services
Document classification	E-mail SPAM filtering	Text documents	Semantic categories
Multimedia database search	Search for biochemical pathway data	Images in electronic documents	Figure types, pathway types
Speech recognition	Directory assistance	Speech waveform	Spoken words

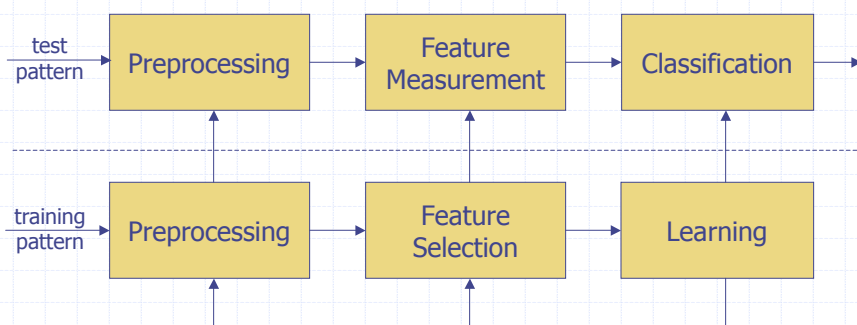
[2]

Various approaches



[2]

Model for machine learning



Machine Learning In Analytical Chemistry - Examples

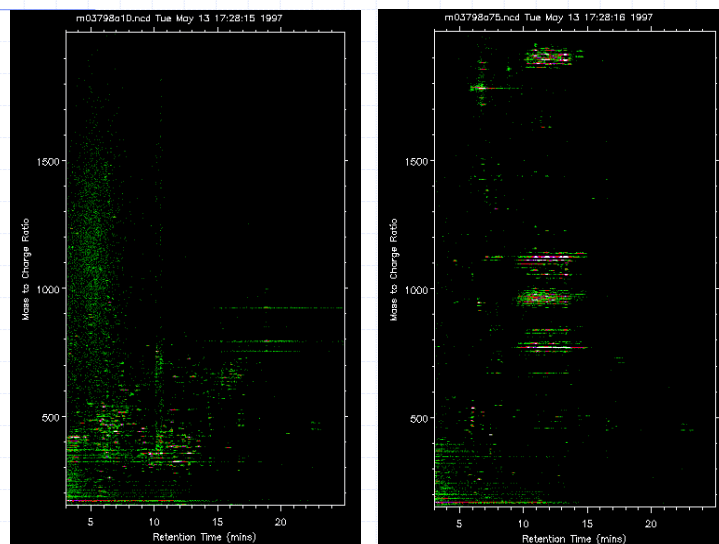
- ◆ Improve efficiency of experts
 - Power-tools allow more from less
- ◆ Embed expertise for non-experts to use
 - Automate application of expert knowledge
- ◆ Elucidate non-trivial relationships
 - Complex relationships in large volumes of data
- ◆ Create a framework for Understanding
 - Create new knowledge in testable models

Example: High Volume LC-MS

- ◆ Look for new natural products in extracts
 - >300,000 extracts
- ◆ Extracts analyzed with 5-30 minute LC run
 - 150-2000 spectra/sample
- ◆ Too much 3D data to look at by hand...

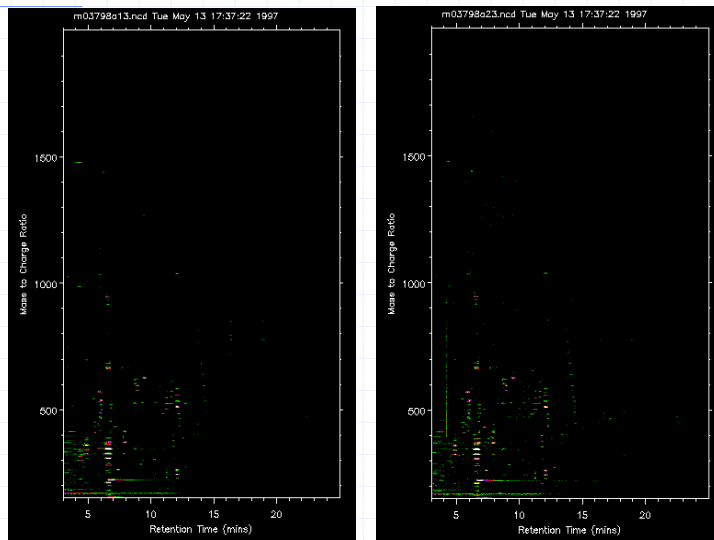
[3]

Fungal Extracts: Samples not similar (Similarity = 0.11)



[3]

Fungal Extracts: Similar samples (Similarity = 0.78)

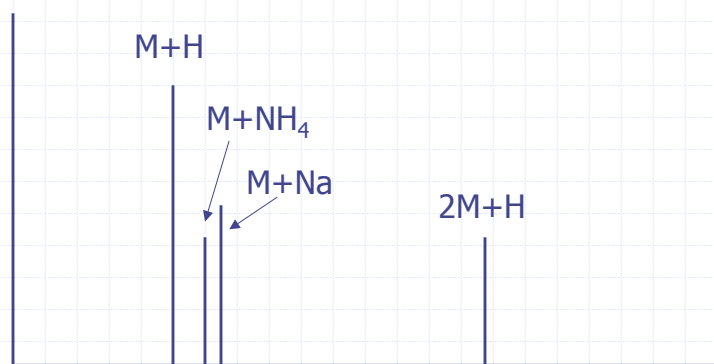


[3]

Example: Automated MS

- ◆ Problem: High volume of combinatorial chemistry samples with unknown purity
- ◆ Solution: High speed FIA-MS
- ◆ ML application: Does the measured data match the proposed structure?

Simple ESI-MS Interpretation



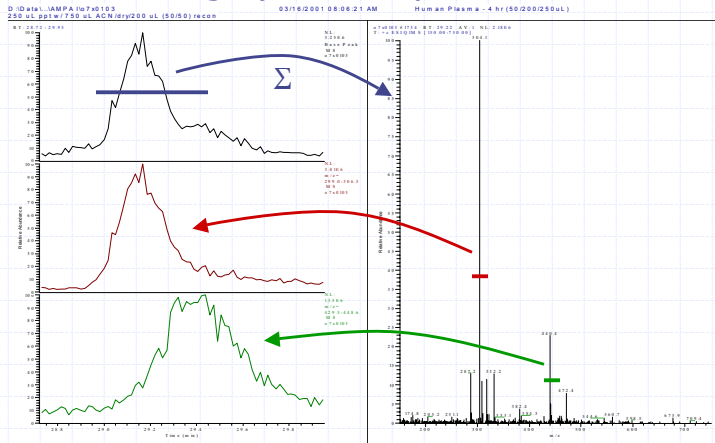
- What is the MW?
- What if some peaks are missing?
- What if there is noise?

Simplistic Approach to MW/Purity

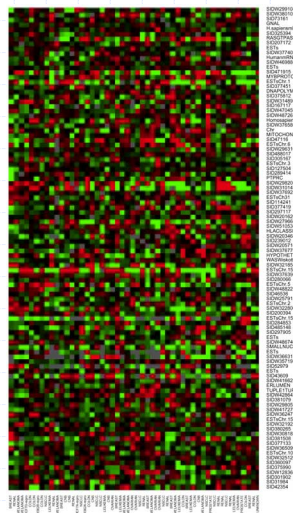
- ◆ Make a list of all possible adducts
- ◆ Compute all possible m/z values
- ◆ Determine if m/z values are present
- ◆ If yes, then call it **OK**
- ◆ If no, call it **BAD**
- ◆ If can't tell, call it **MAYBE**

Problems with simplistic approach

'chromatographic' components:



Complex Relationships: Microarrays



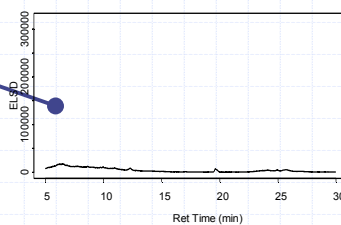
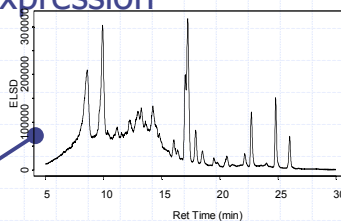
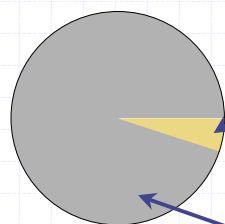
6830 Genes (Rows)
64 samples (cols)

What do you see?

Figure 1.3 From the textbook (p6)

Framework for understanding

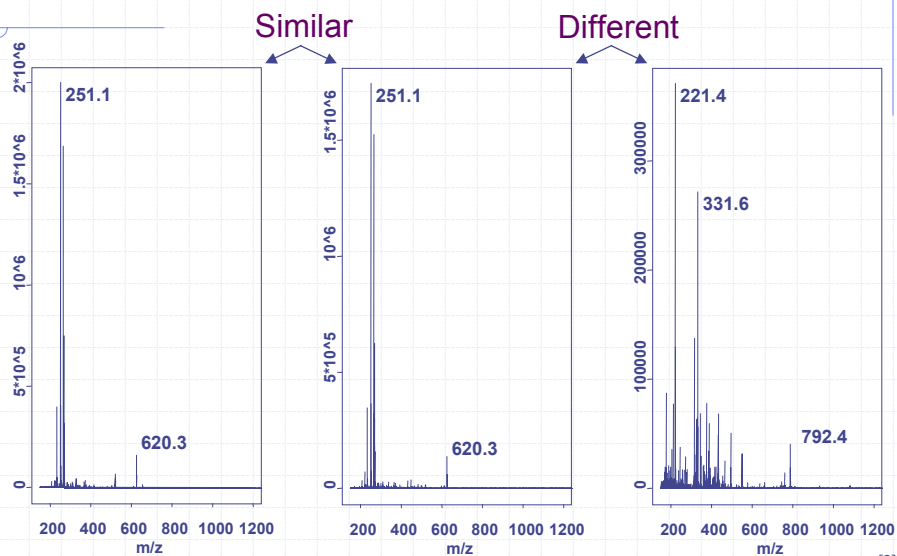
◆ Fungal metabolite expression



[3]

Rapid Chemical Screening – Diversity

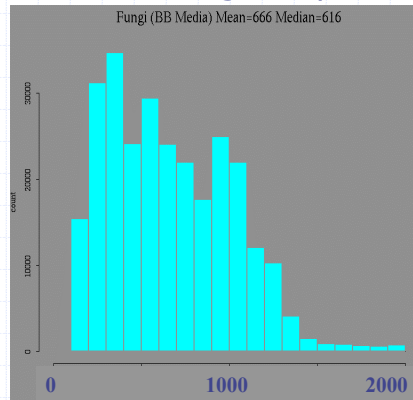
Media subtracted average spectra from 5 minute analysis



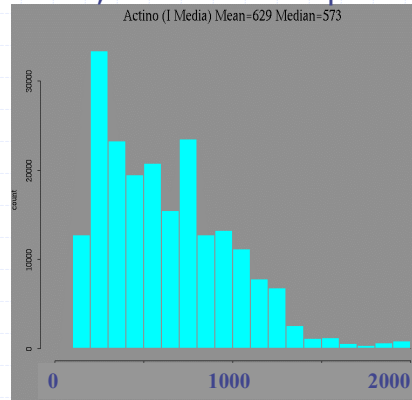
[3]

Mass of Secondary Metabolite Ions

3,337 Fungi Samples



1,700 Actino Samples



[3]

Patterns and Models

- ◆ "All models are wrong, but some are useful"
- G.E.P. Box
- ◆ "If you look for patterns, you will find them" – Anon
- ◆ "First, you must not fool yourself – and you are the easiest one of all to fool..." - Feynman
- ◆ "Prediction is very difficult, especially if it's about the future" - Bohr

Parsimony, Numerology and Genes

- ◆ Many genomic sequences contain simple repetitive patterns
- ◆ Patterns can be 'found' by a number of methods
- ◆ What is the significance of the pattern?

Discovering Simple DNA Sequences

- ◆ Look for repetitive patterns:
 - Use the "The name game" to generate supposed sequence...
 - Sum(ASCII("RandallKeithJulian")) = 1814
 - Julian Family: 11437; $x^2 = 130804969$
 - Binary: 111110010111110110011101001
 - DNA Code: 00=A; 01=B; 10=G; 11=T
 - Code: CGGTATGTTGATTC
 - Exact BLAST Hits at NCBI: 6
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
 - GenBank Human EST entries 4,889,573 sequences;
2,536,808,622 total letters

Alternative:

◆ Uniform Random Number Generator

- 14 digits 0-3, converted to DNA Base letters
- (1): CGACACGGGTGGGC 8 hits
- (2): CTGCCCCCAATTCG 3 hits
- (3): TTGAAGTGTGACAG 1 hit
- (4): AGTGTGTCGTCGTA 2 hits

- ◆ Is the numerology approach different than a random number generator?

[4]

Parsimony

- ◆ Simpler answers are more likely
- ◆ Complex methods carry a "Minimum Description Length".
- ◆ The uniform random number generator is "shorter" than the Name Game.
- ◆ For a method to be relevant it has to be different than random.

[4]

Conclusions:

- ◆ Informatics can play a significant role in analytical chemistry.
- ◆ Machine learning represents a next step in using computers to analyze data.
- ◆ Machine learning and pattern recognition share a great deal with “curve fitting”.
- ◆ To keep from being fooled, we need to know the best and worst we can do.

What's next:

- ◆ Ideas from statistics
 - What is needed to do pattern recognition?
 - What are the limits of performance?
- ◆ Tools to make this reasonable
 - What software to use?
- ◆ Working examples to try things out
 - Least squares
 - Nearest neighbors

References

- ◆ [1] "Information Theory in Analytical Chemistry", Eckschlager, Danzer
- ◆ [2] "Statistical Pattern Recognition: A Review", A.K. Jain, R.P.W. Duin, and J. Mao, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 1, JANUARY 2000
- ◆ [3] "Advances in the Application of LC-MS to Mixture Analysis in Natural Products Drug Discovery", R.K. Julian, R.E. Higgs, M.D. Hilton, 4th Conference on Molecular Diversity, Lake Tahoe, CA, March 2000
- ◆ [4] "Pattern Discovery in Biological Data", J.T.L. Wang, B.A. Shapiro, D.Shasha, Eds. Oxford Univ. Press 1999