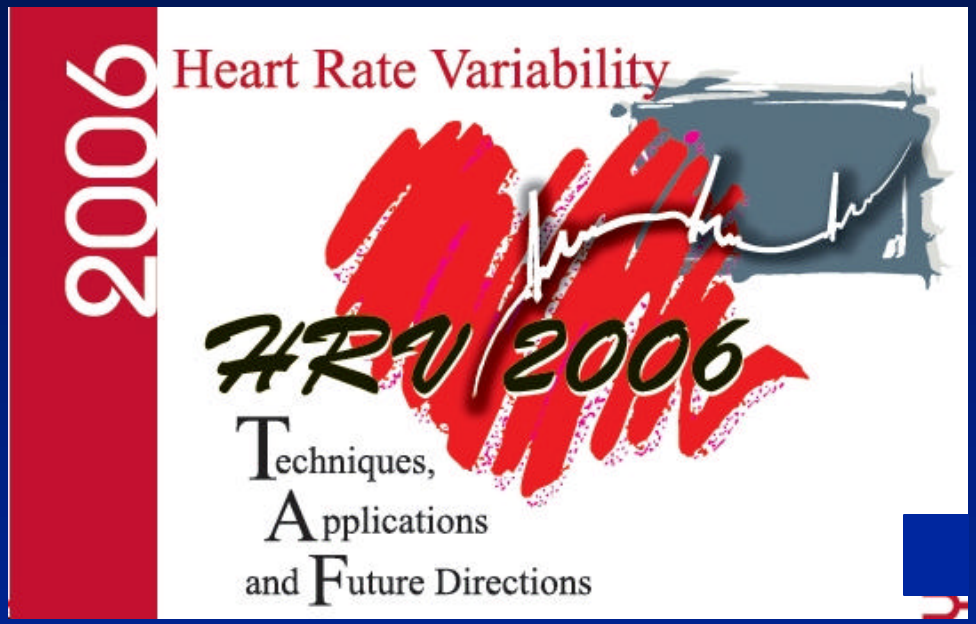
Complexity Measures: Linguistic and Symbolic Dynamical Approaches



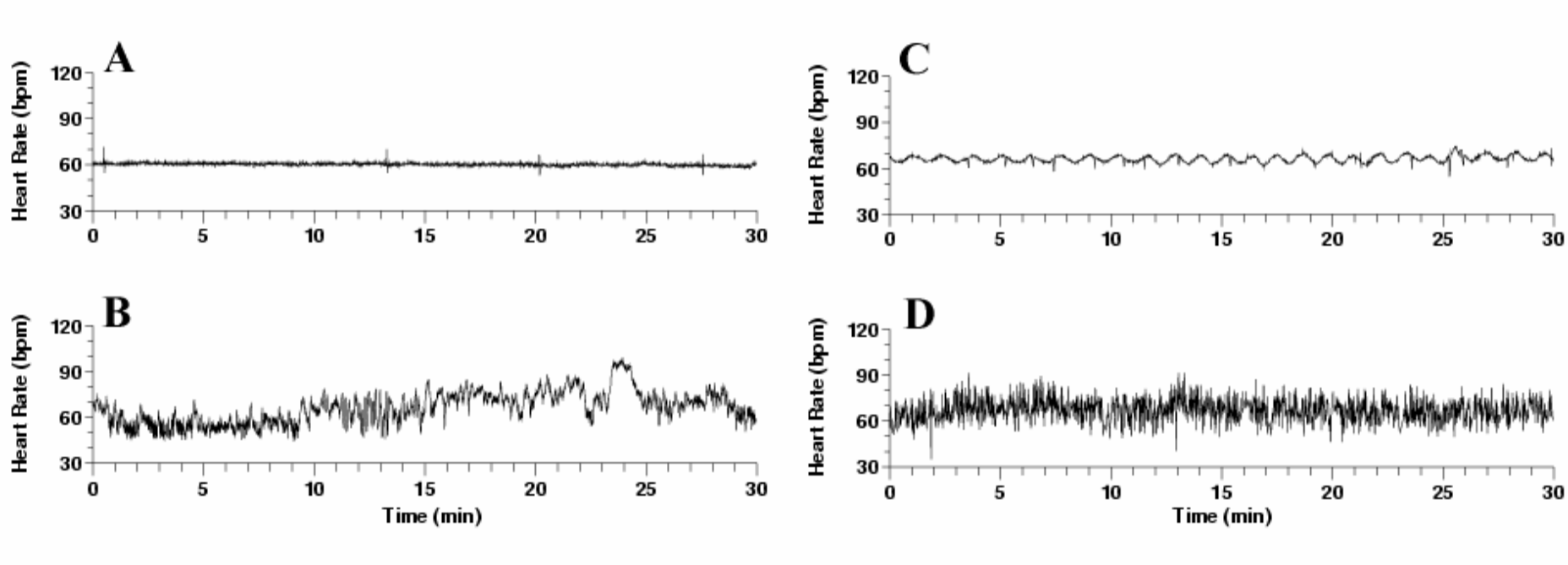
Chung-Kang Peng, PhD

Beth Israel Deaconess Medical Center
Harvard Medical School

Objectives

- Provide an analogy between heart rate time series and other symbolic sequences (e.g., language)
- Introduce an intuitive measurement of similarity based on repetitive patterns
- Discuss clinical applications of this linguistic approach

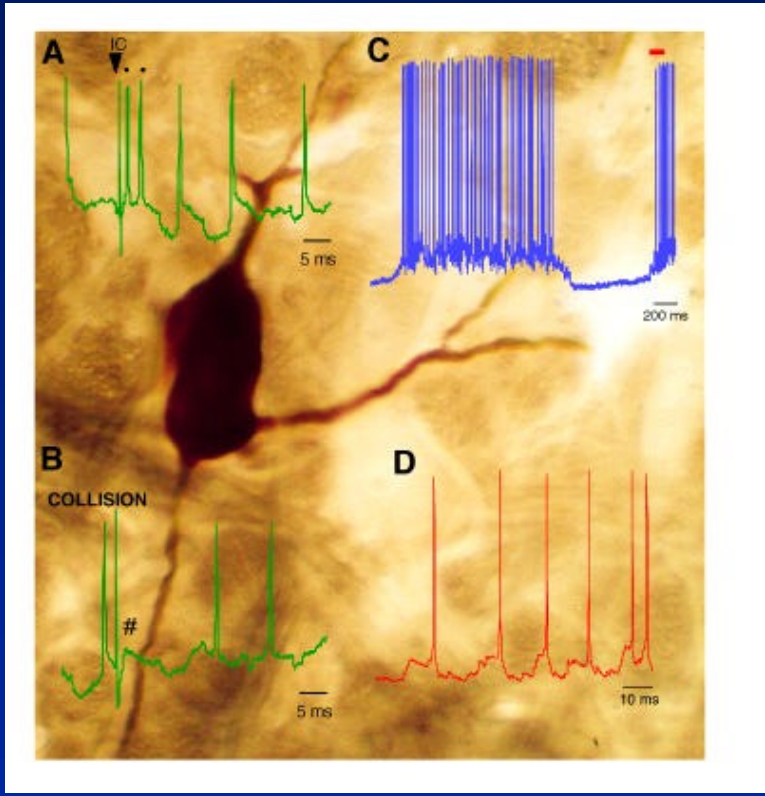
Heart Rate Dynamics in Health and Disease

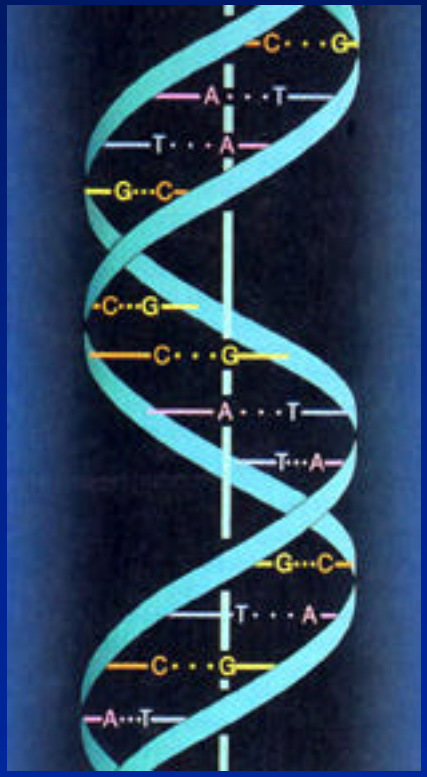


These time series contain information that reflects their underlying control mechanisms

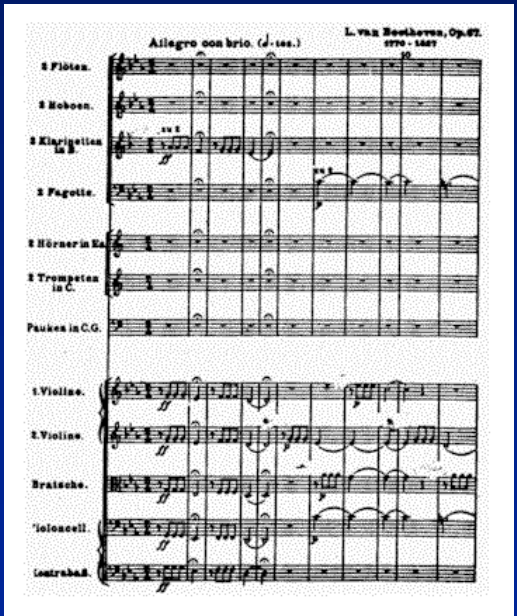
Question: Can we usefully classify different types of information even *without* initially understanding their content?

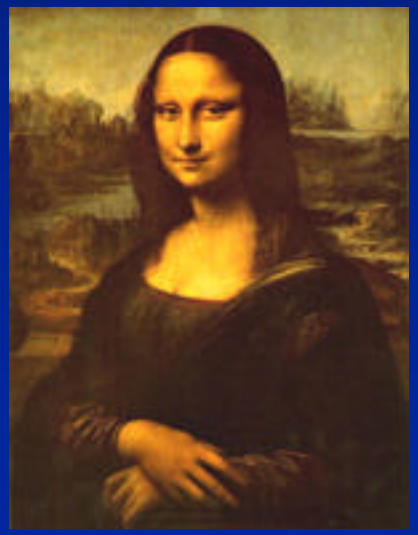
Information Created by Biological Systems

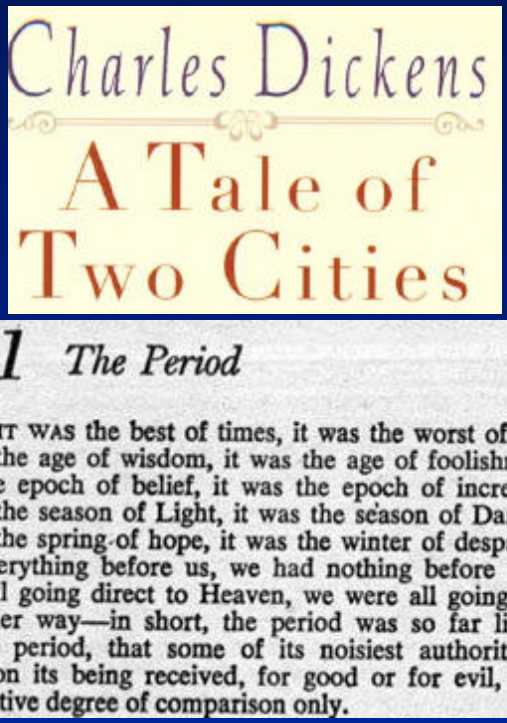




Information Created by Humans







Our approach:
Categorize complex signals based on the occurrences of ***repetitive patterns***

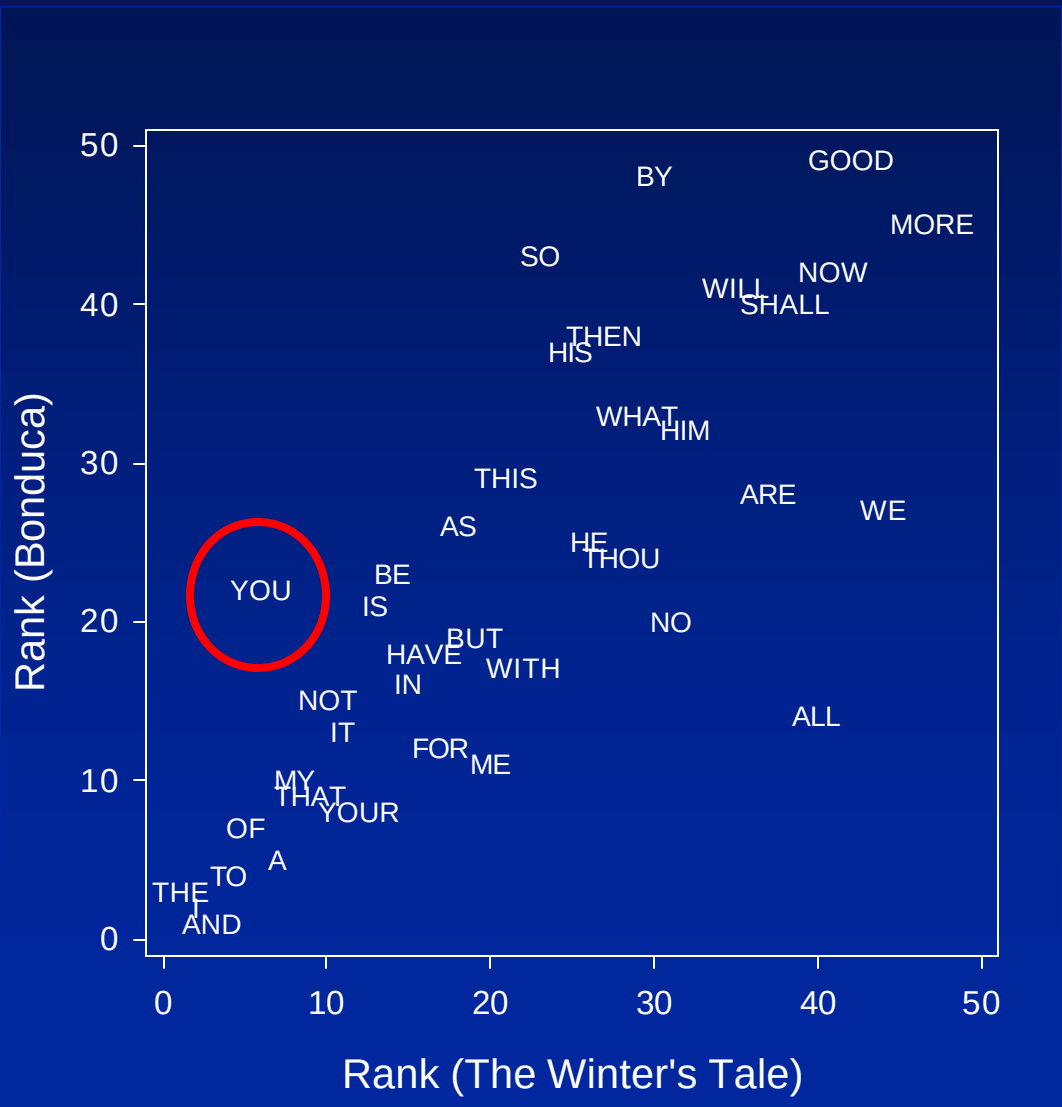
Illustration of our algorithm:
Comparison of human literary texts

Repetitive patterns: words

Refs: Phys Rev Lett 2003; 90:108103; Physica A 2003:329:473-83

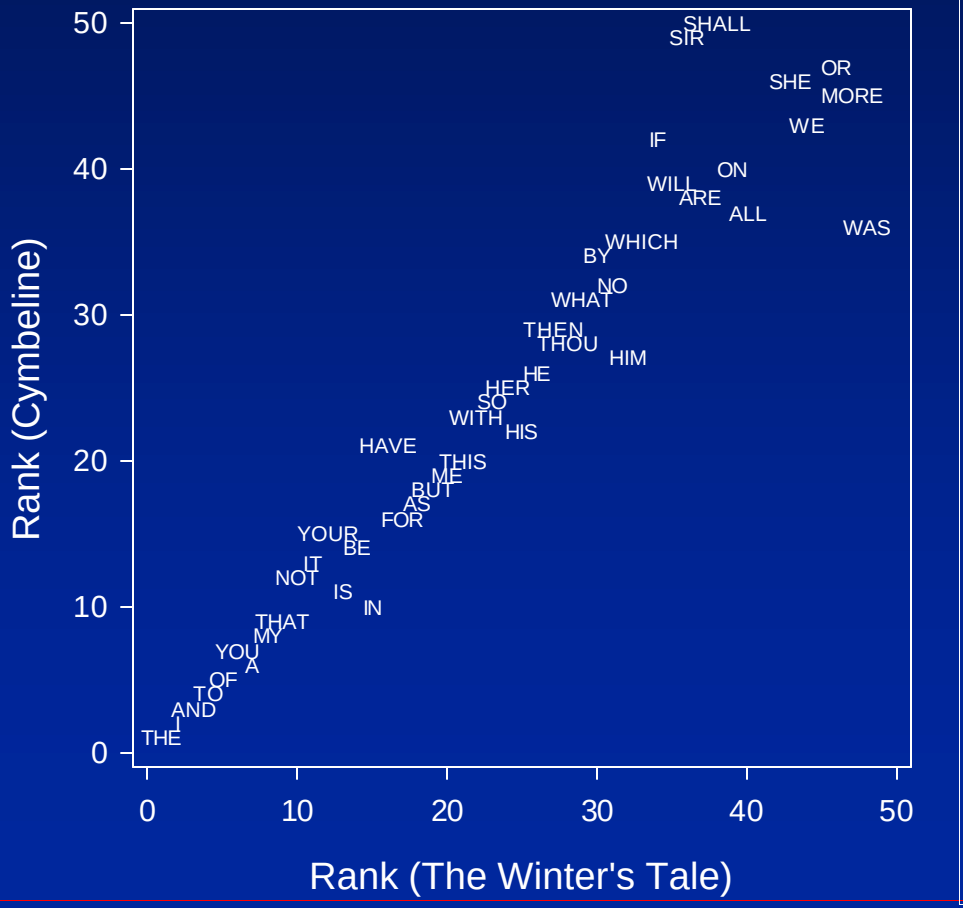
Comparison of Human Texts

Word	Rank (The Winter's Tale)	Rank (Bonduca)
THE	1	3
I	2	2
AND	3	1
TO	4	4
OF	5	7
YOU	6	22
A	7	5
MY	8	10
THAT	9	9
NOT	10	15

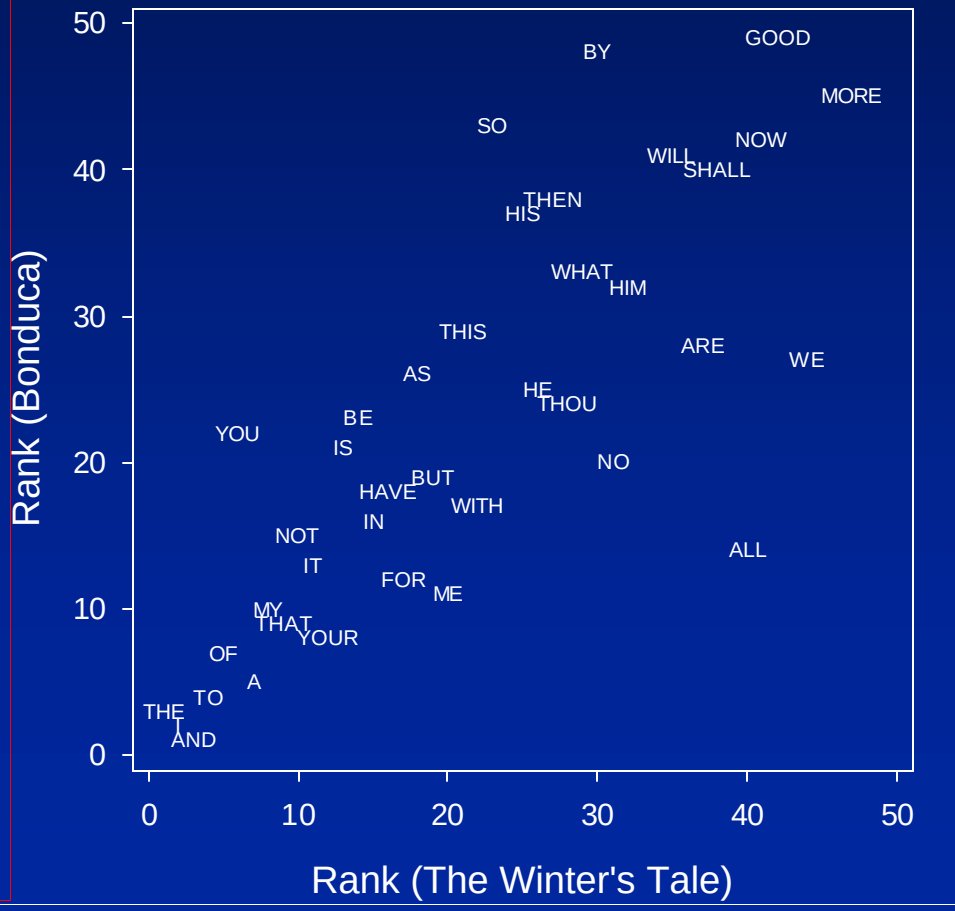


Rank Comparison Maps

Shakespeare vs. Shakespeare

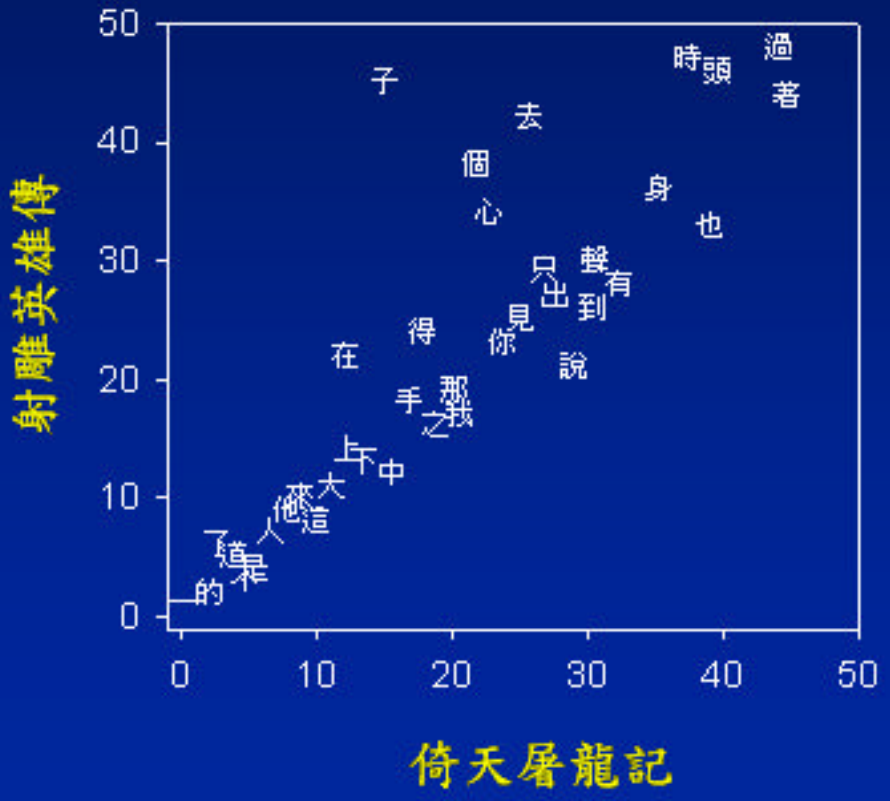


Shakespeare vs. Fletcher

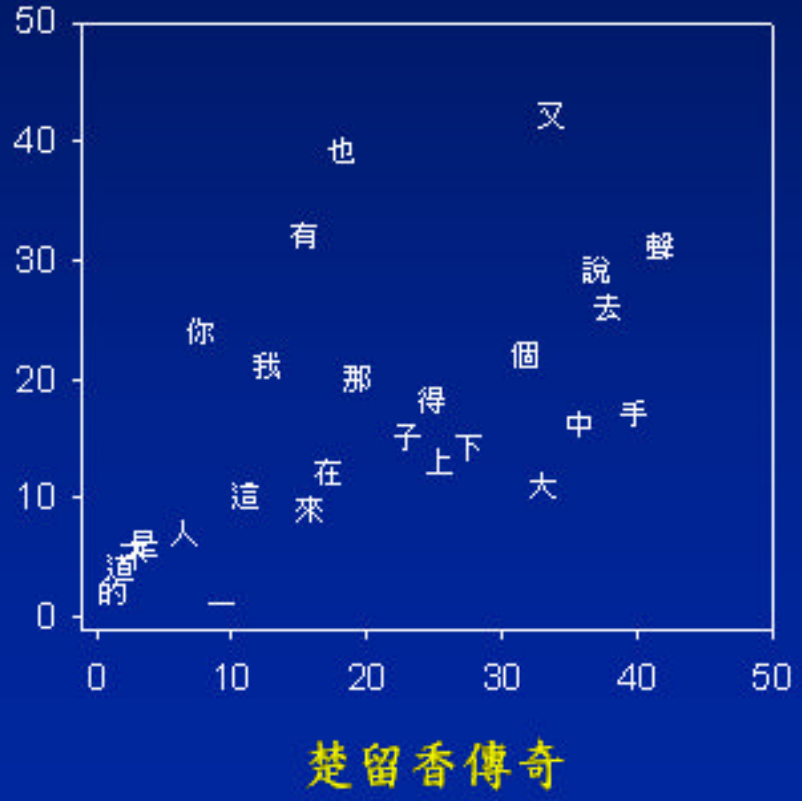


Rank Comparison Maps

Same Author



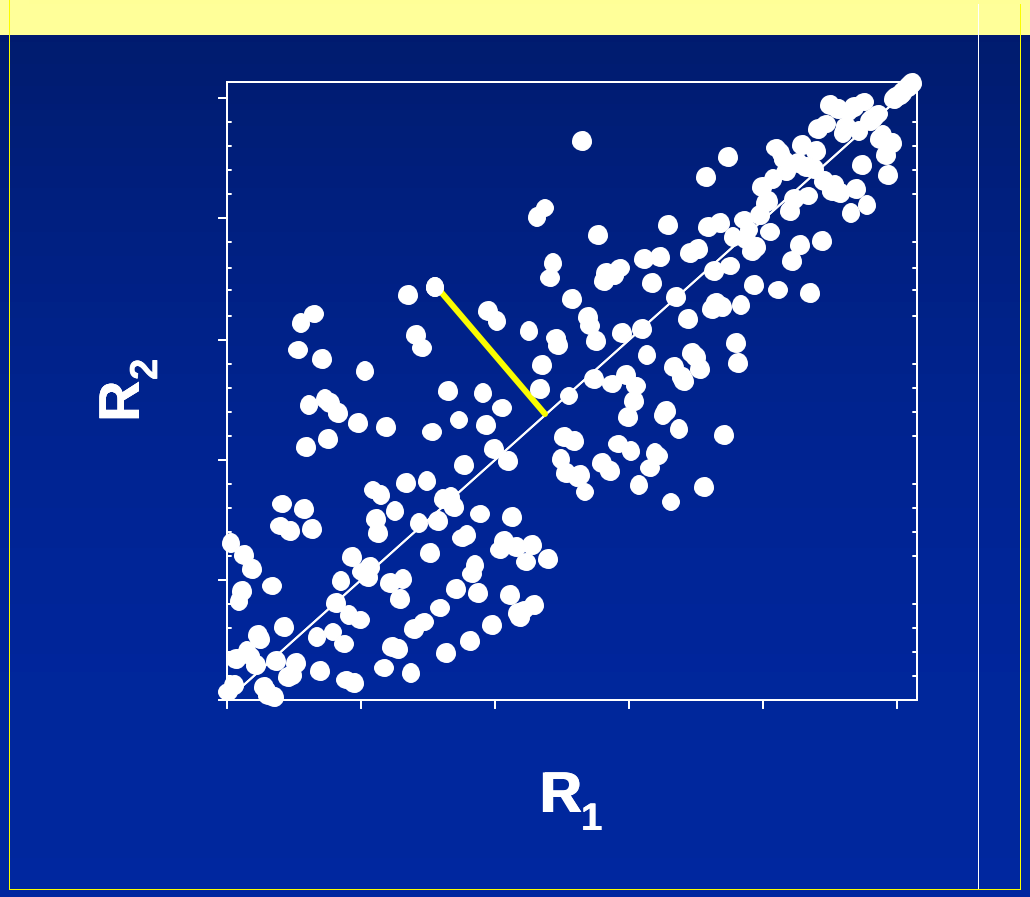
Different Authors



Dis-Similarity Index

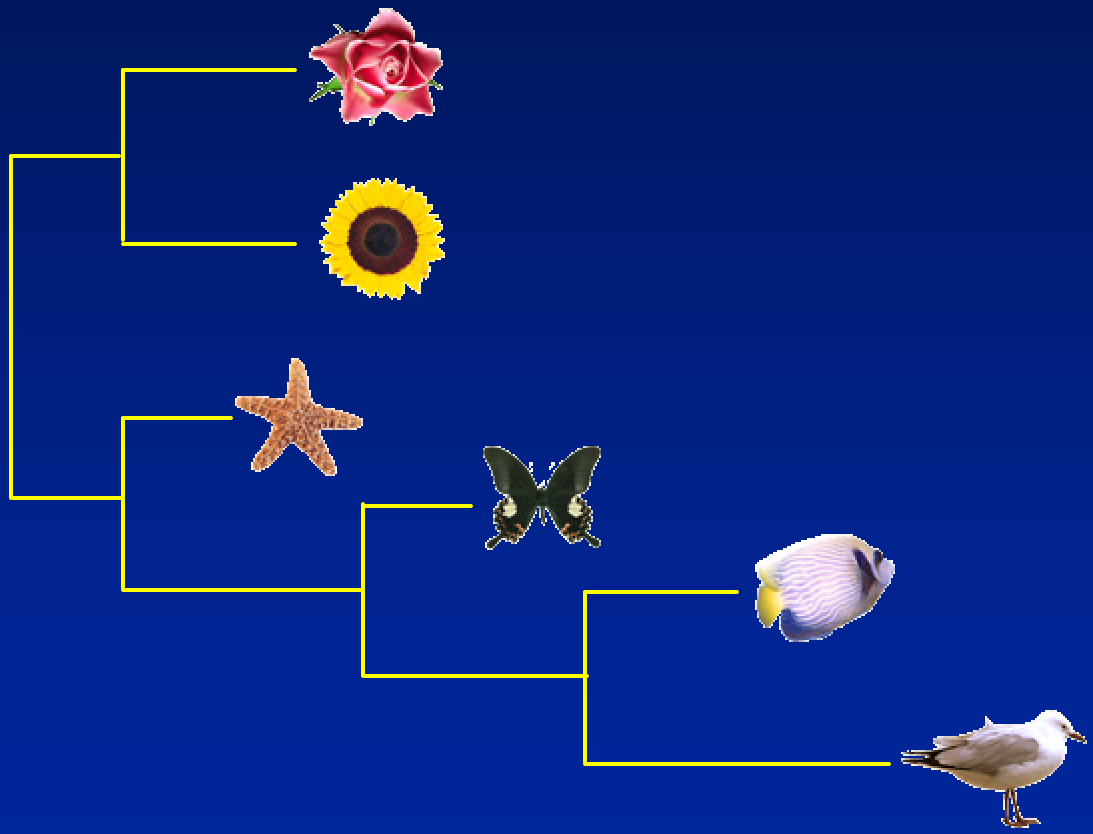
$$D(S_1, S_2) = \sum_{k=1}^{N} \left| R_1(w_k) - R_2(w_k) \right| \; F(w_k)$$

$$F(w_k) = \frac{1}{Z}\left[- p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k) \right]$$

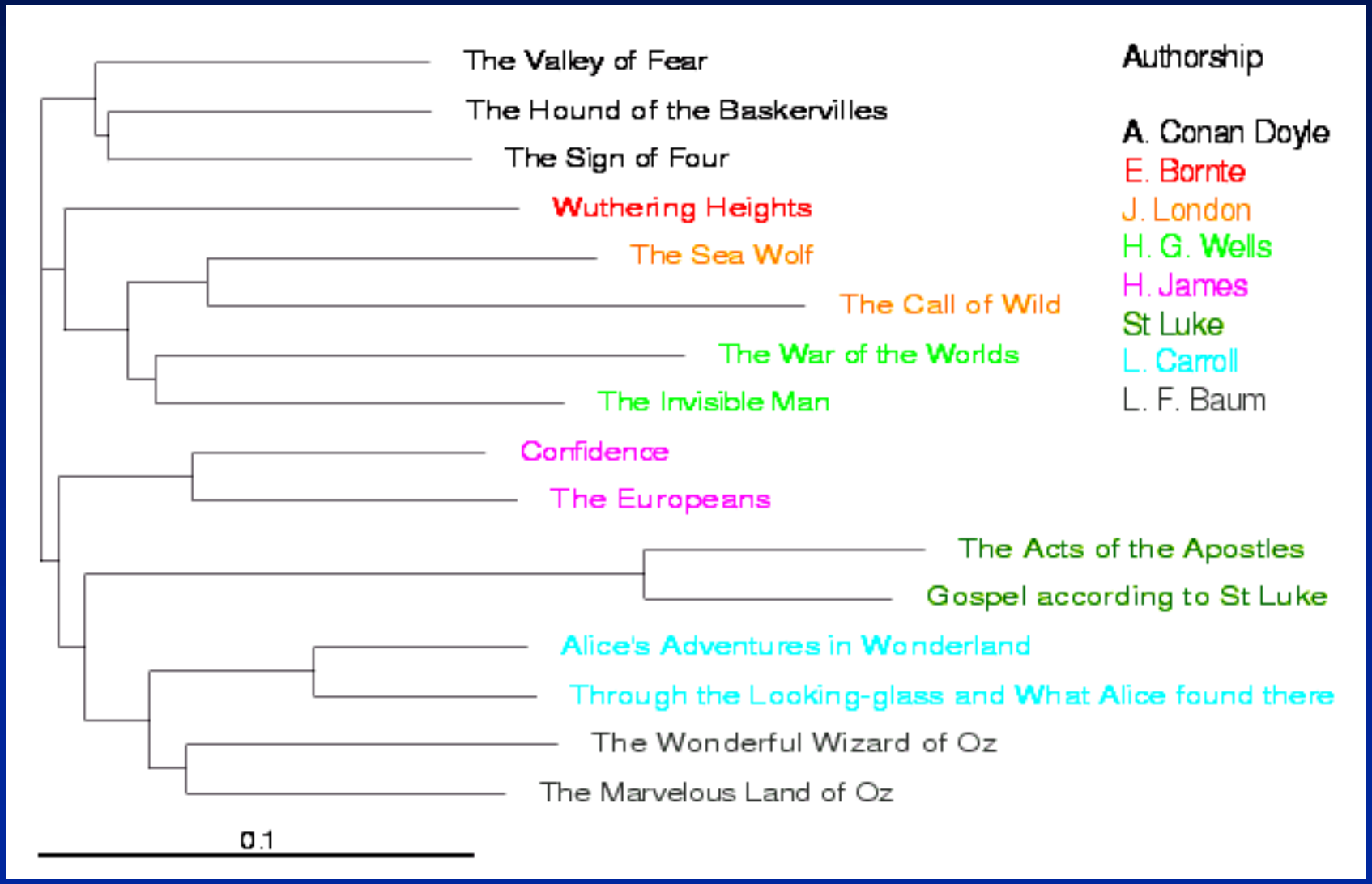


Phylogenetic Tree Distance matrix method

Phylogenetic Tree



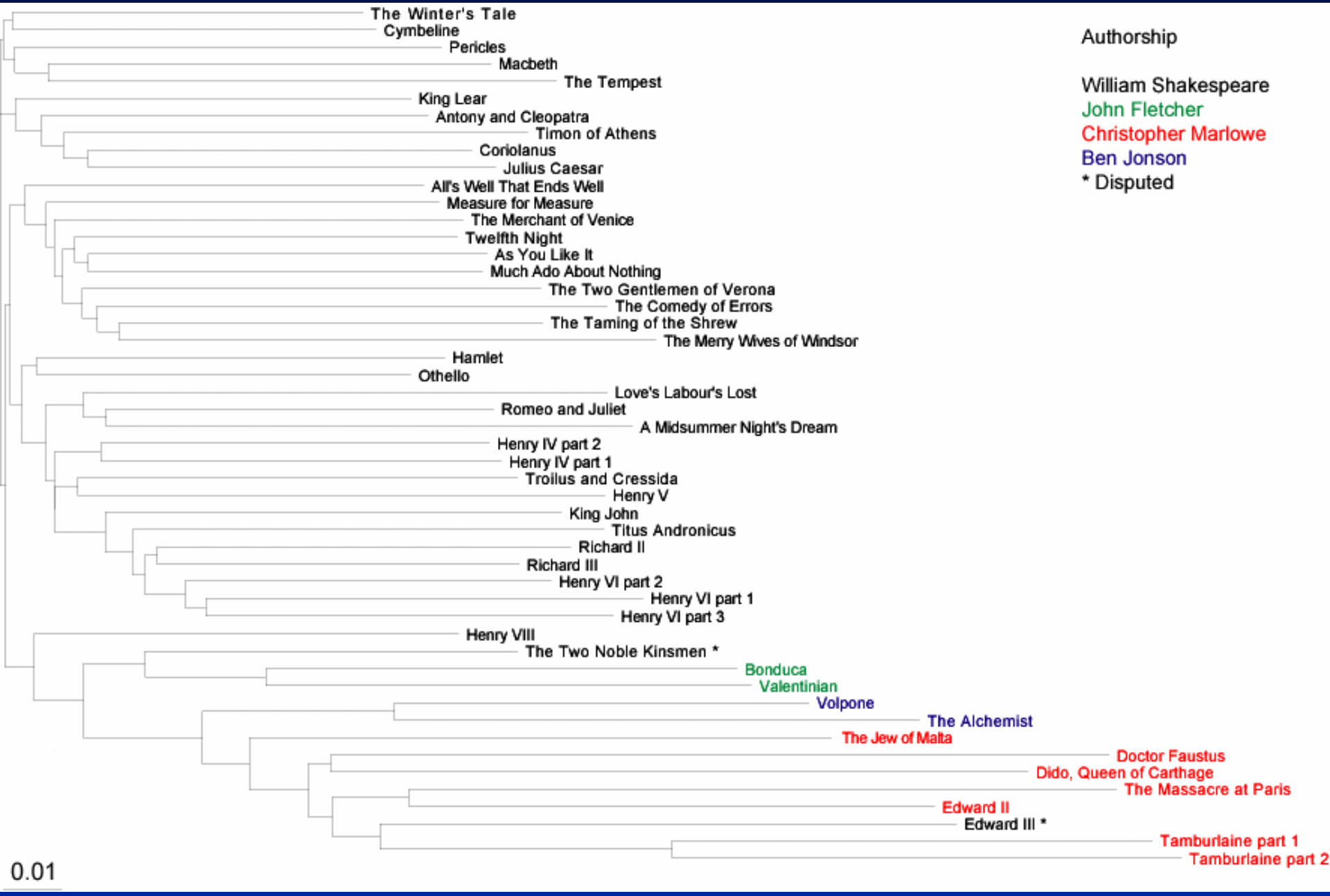
Authorship Problem

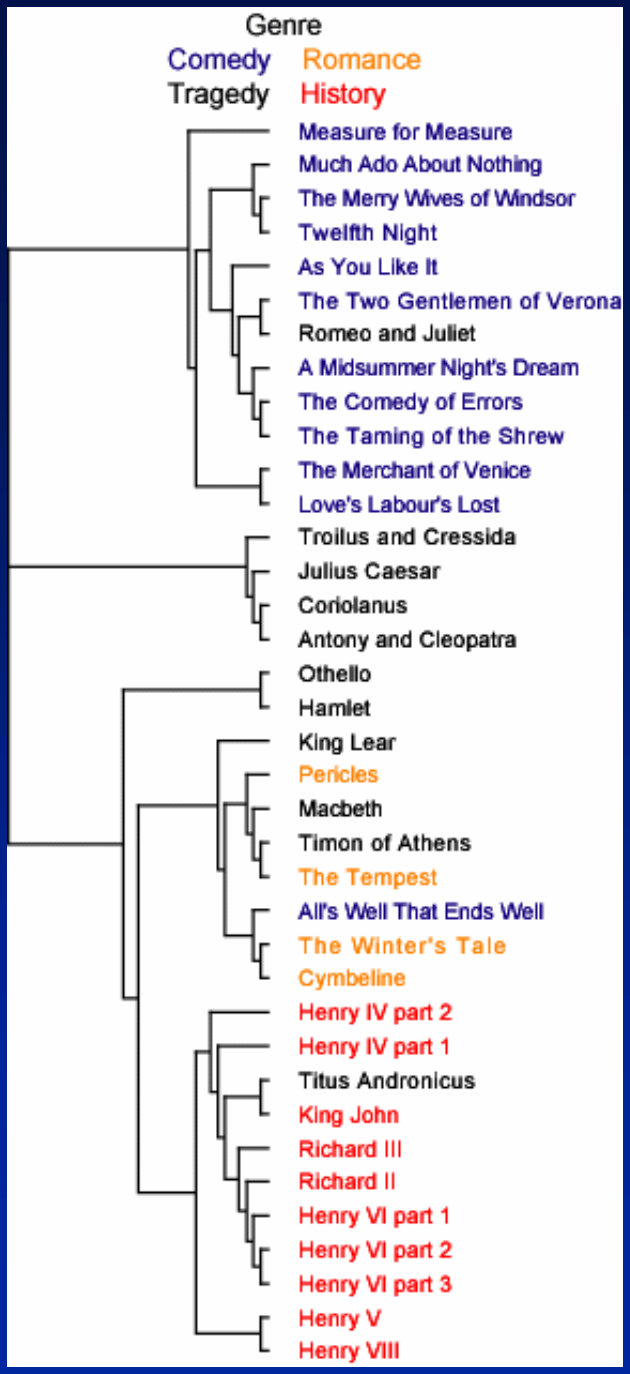


Who Wrote Shakespeare's Plays?

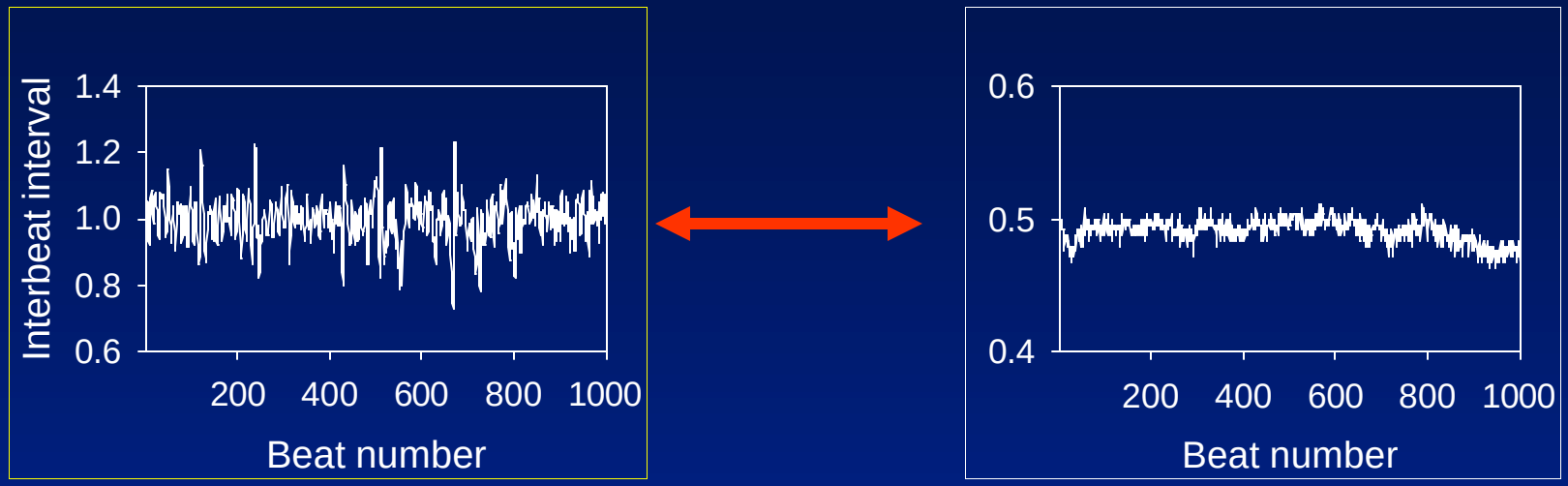
Shakespeare

Marlowe


The Winter's Tale
Cymbeline
Pericles
Macbeth
The Tempest
King Lear
Antony and Cleopatra
Timon of Athens
Coriolanus
Julius Caesar
All's Well That Ends Well
Measure for Measure
The Merchant of Venice
Twelfth Night
As You Like It
Much Ado About Nothing
The Two Gentlemen of Verona
The Comedy of Errors
The Taming of the Shrew
The Merry Wives of Windsor
Hamlet
Othello
Love's Labour's Lost
Romeo and Juliet
A Midsummer Night's Dream
Henry IV part 2
Henry IV part 1
Troilus and Cressida
Henry V
King John
Titus Andronicus
Richard II
Richard III
Henry VI part 2
Henry VI part 1
Henry VI part 3
Henry VIII
The Two Noble Kinsmen *
Bonduca
Valentinian
Volpone
The Alchemist
The Jew of Malta
Doctor Faustus
Dido, Queen of Carthage
The Massacre at Paris
Edward II
Edward III *
Tamburlaine part 1
Tamburlaine part 2
Authorship
William Shakespeare
John Fletcher
Christopher Marlowe
Ben Jonson
* Disputed
0.01

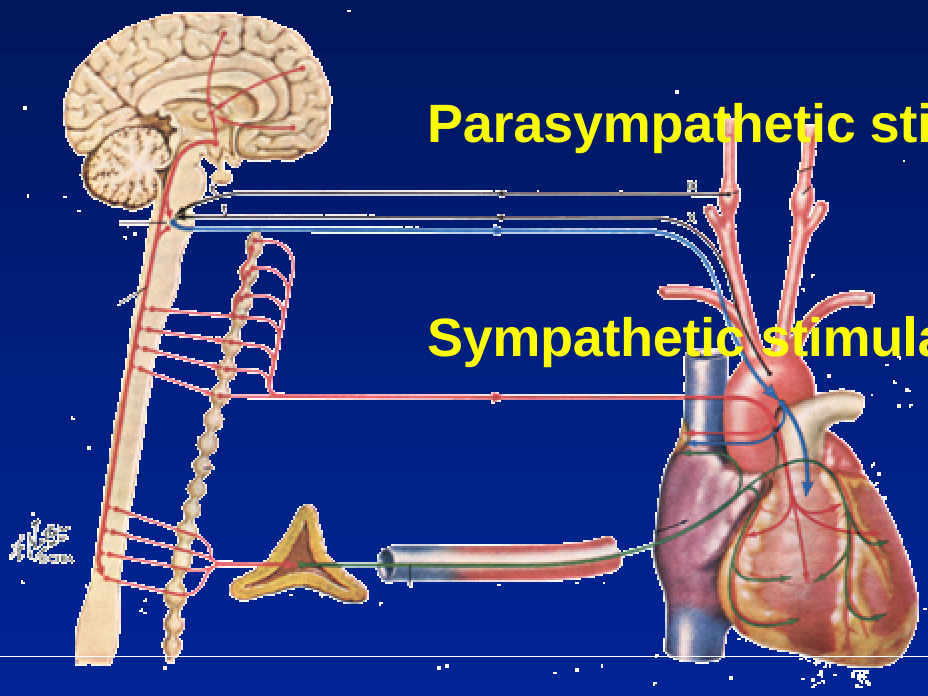

Genre
Comedy
Romance
Tragedy
History
Measure for Measure
Much Ado About Nothing
The Merry Wives of Windsor
Twelfth Night
As You Like It
The Two Gentlemen of Verona
Romeo and Juliet
A Midsummer Night's Dream
The Comedy of Errors
The Taming of the Shrew
The Merchant of Venice
Love's Labour's Lost
Troilus and Cressida
Julius Caesar
Coriolanus
Antony and Cleopatra
Othello
Hamlet
King Lear
Pericles
Macbeth
Timon of Athens
The Tempest
All's Well That Ends Well
The Winter's Tale
Cymbeline
Henry IV part 2
Henry IV part 1
Titus Andronicus
King John
Richard III
Richard II
Henry VI part 1
Henry VI part 2
Henry VI part 3
Henry V
Henry VIII

Application to Heartbeat Time Series



- How to map a time series to a symbolic sequence?
- How to define ***words*** in symbolic sequences?

Heart rate dynamics



Symbolic Mapping

Step 1: Binary code representation

Increment of successive RR intervals can be mapped to 0 or 1

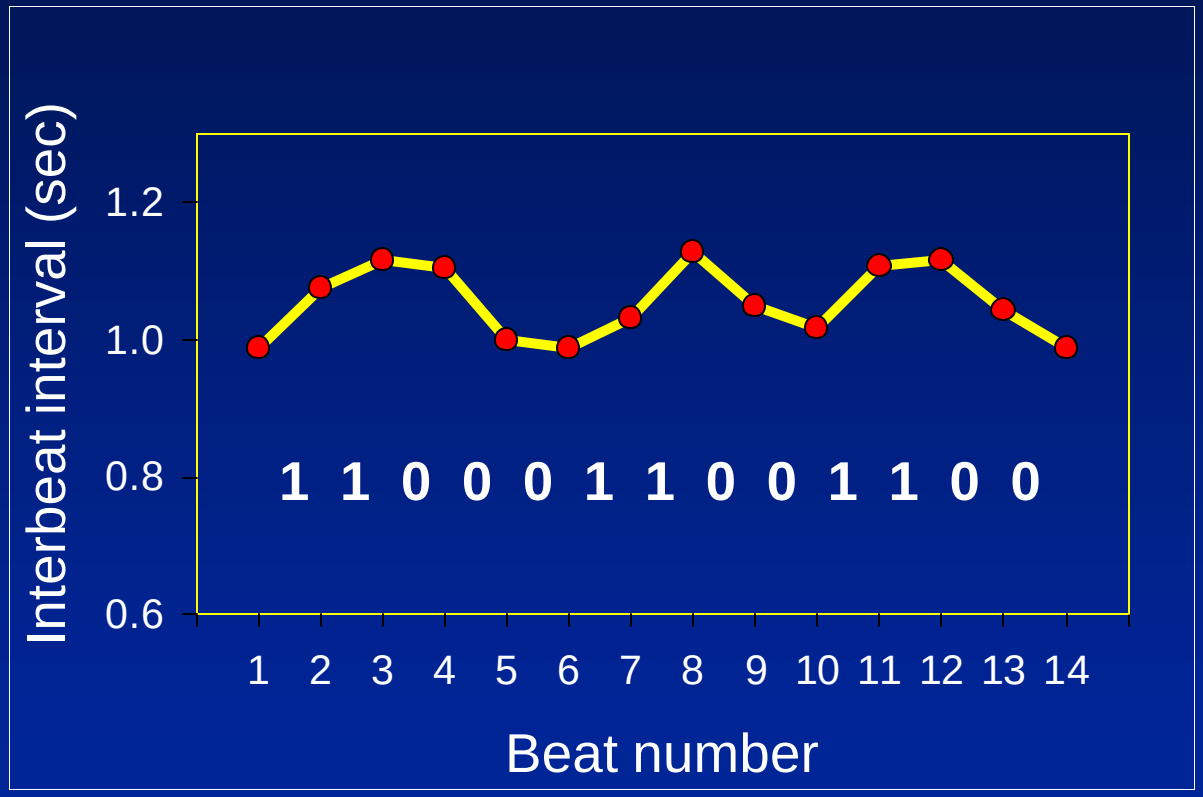
$$I_n = \begin{cases} 0, & \text{if } x_n \le x_{n-1} \\ 1, & \text{if } x_n > x_{n-1} \end{cases}$$

Ashkenazy et al., Phys. Rev. Lett. 86, 1900 (2001).

Step 2: “Word” partitioning

m+1 successive RR intervals are mapped to a binary sequence of length m, called an **m-bit word**

Symbolic Mapping

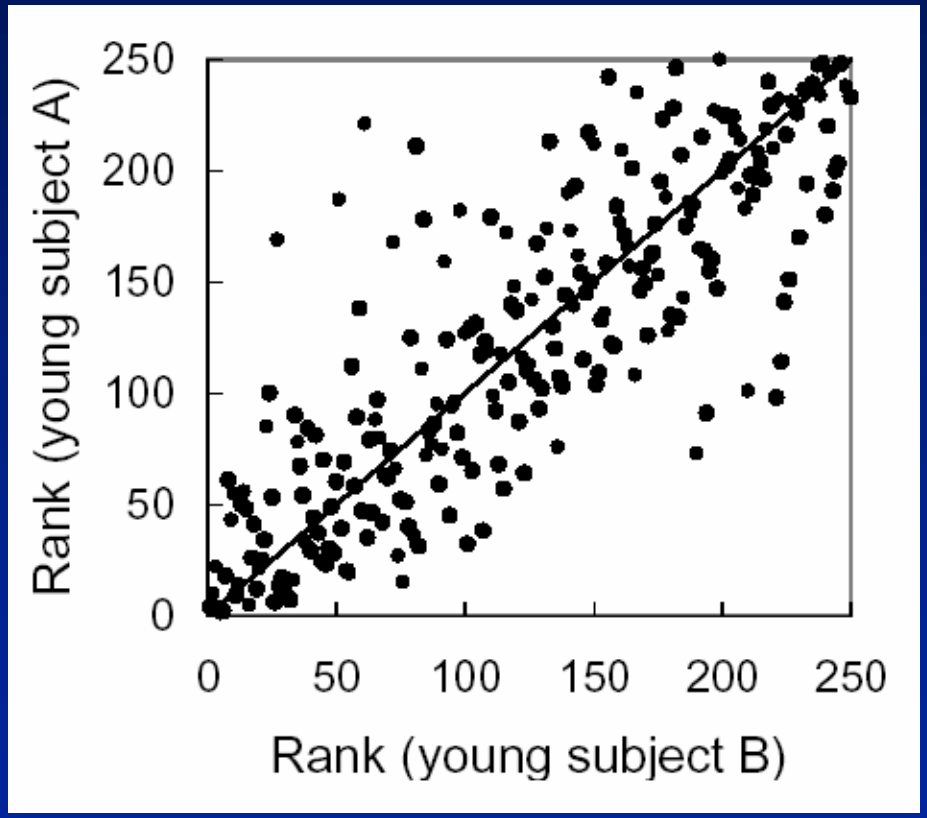


8-bit word: 11000110, 10001100, 00011001

Scientific Questions

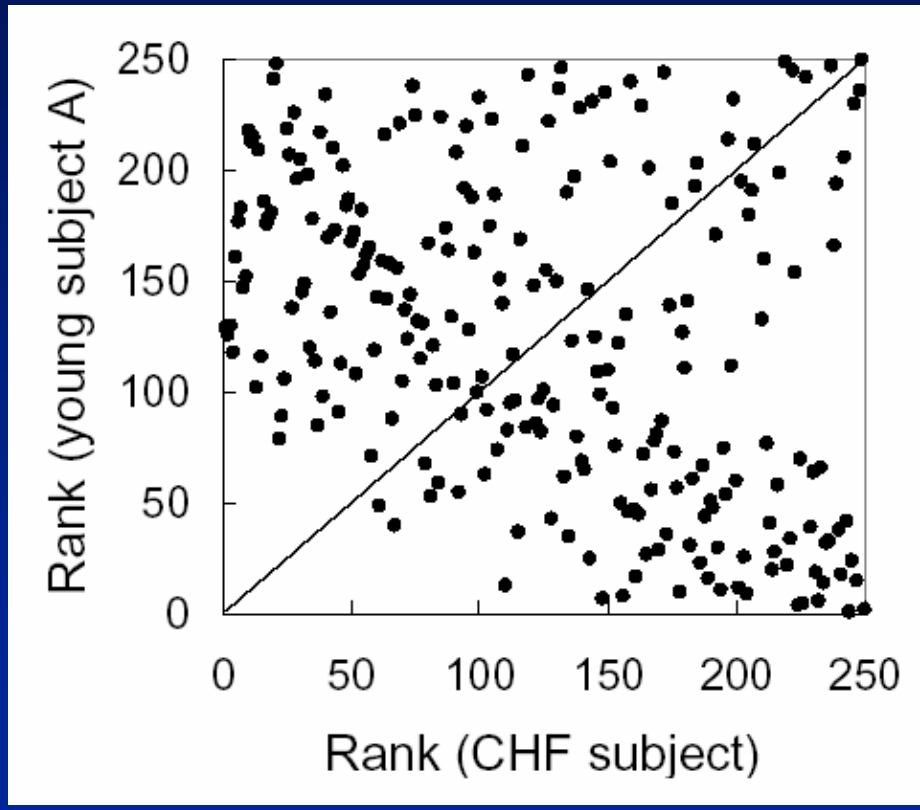
- As a healthy system is altered by disease and aging, can we quantify changes in the dynamical patterns?

Health vs. Health



D = 0.10

Health vs. Disease



D = 0.25

PhysioNet: www.physionet.org

NIH Research Resource for Complex Physiologic Signals

- Healthy subjects **2 hours**

  Young (10 male, 10 female, average 25.9 years)

  Elderly (10 male, 10 female, average 74.5 years)

- Congestive heart failure database **16-24 hours**

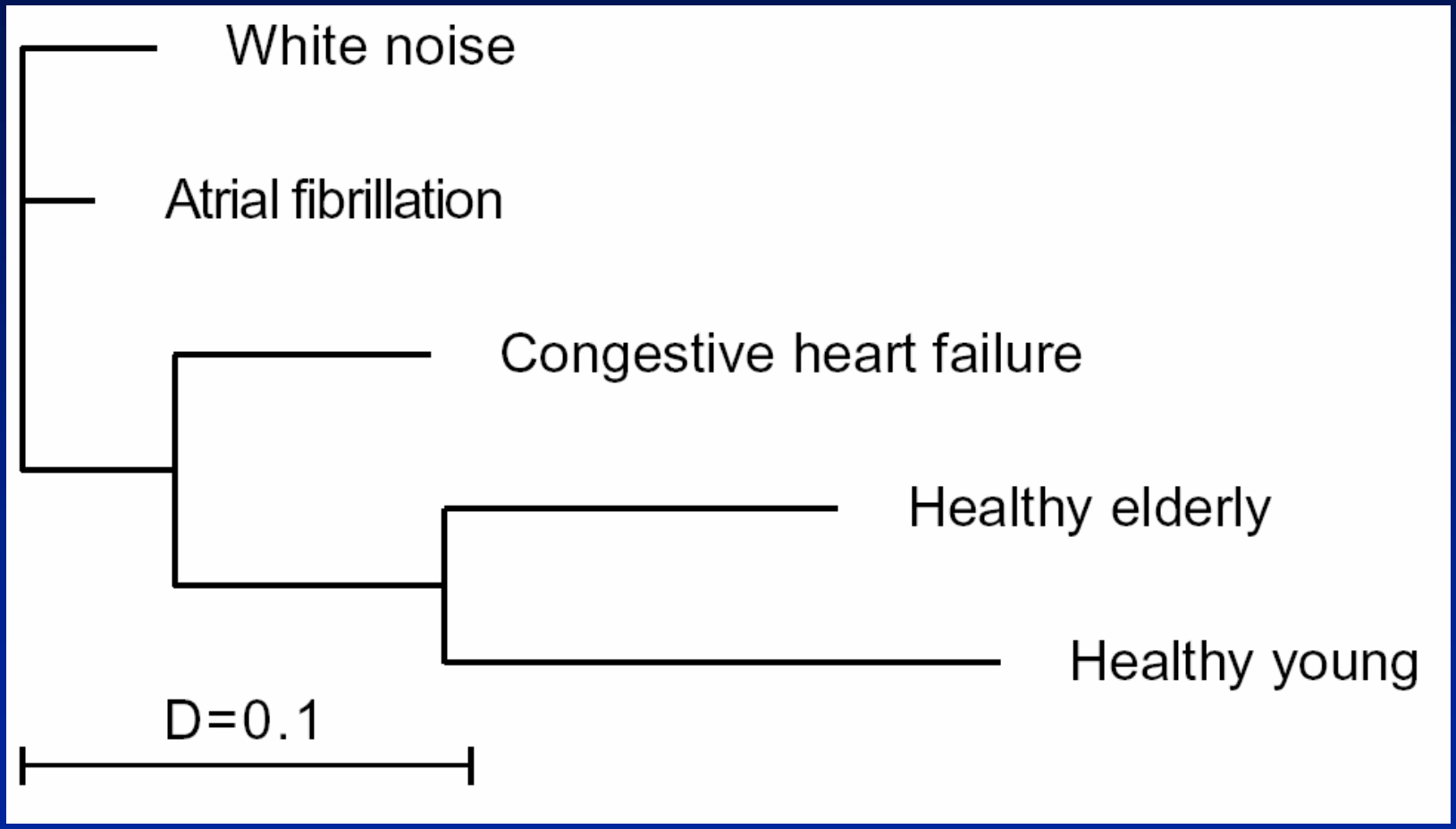
  15 female, 28 males, average 55.5 years

- Atrial fibrillation database **2 hours**

  9 subjects

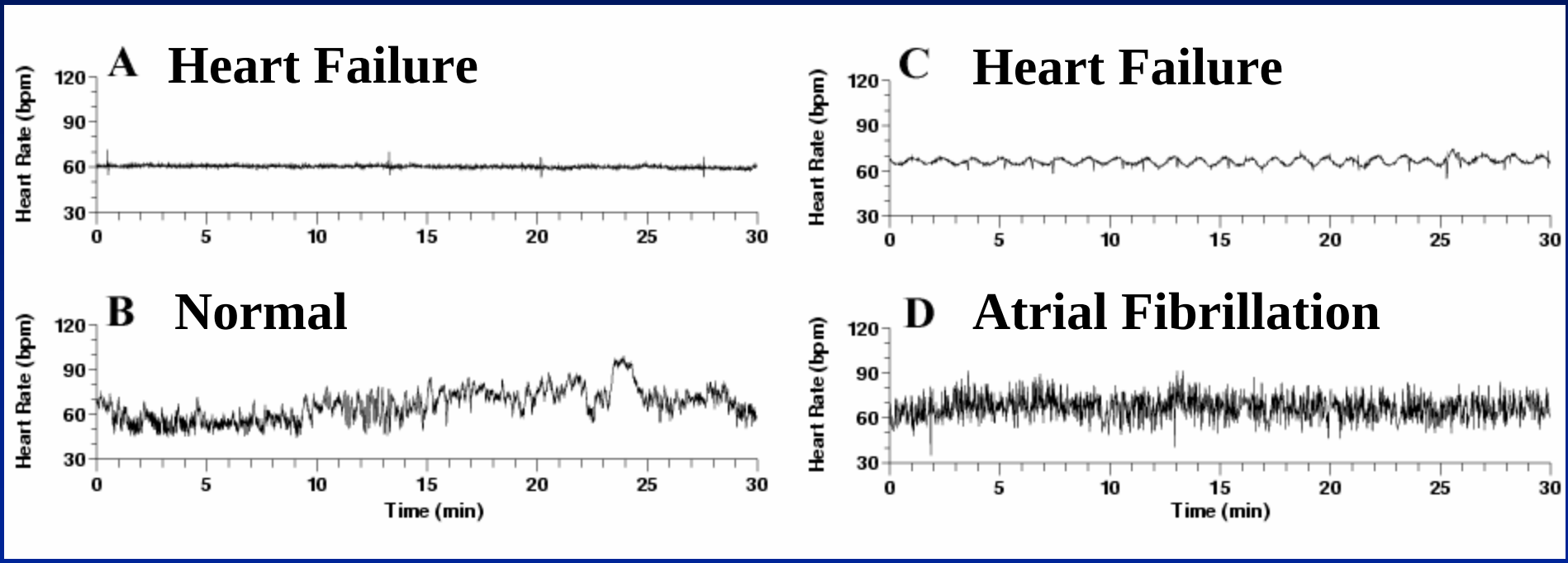
- White noise (random Gaussian noise)

Phylogenetic Tree



Yang *et al.* Phys Rev Lett, 2003; **90**: 108103

Heart Rate Dynamics in Health and Disease

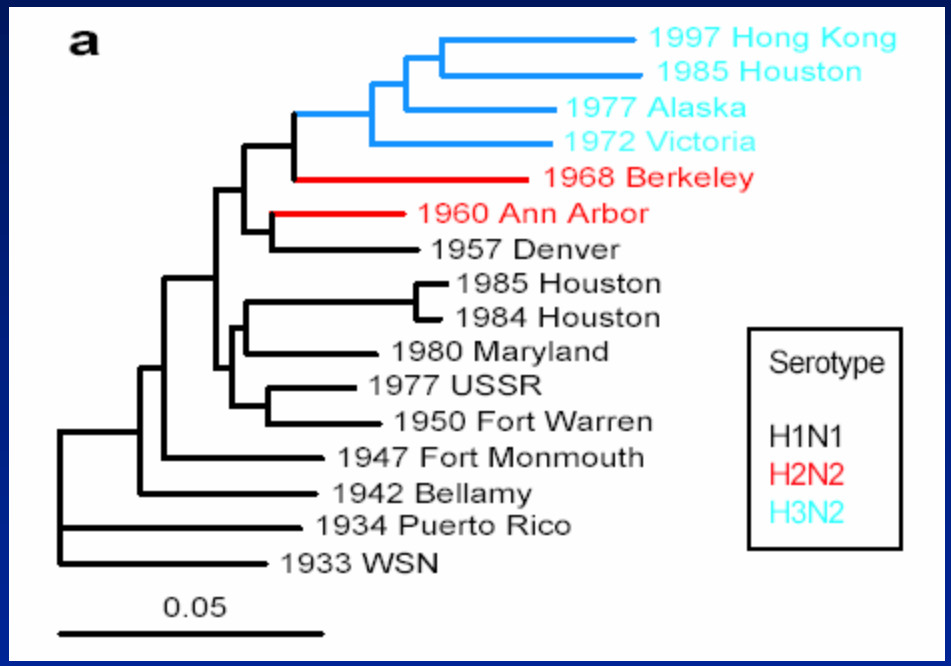


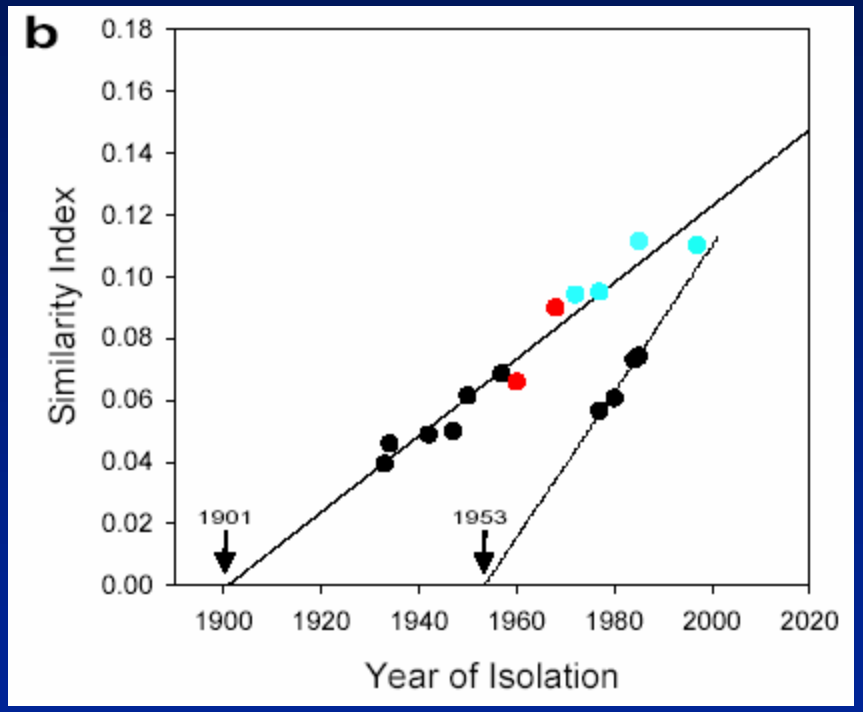
It is possible to categorize these time series!

Other Applications

- Gene chip (micro-array) analysis
- Dynamics of sleep stage transitions
- EEG classification
- Image analysis (pathology)
- Psychiatric evaluation

Application to DNA Sequences: Human Influenza Virus

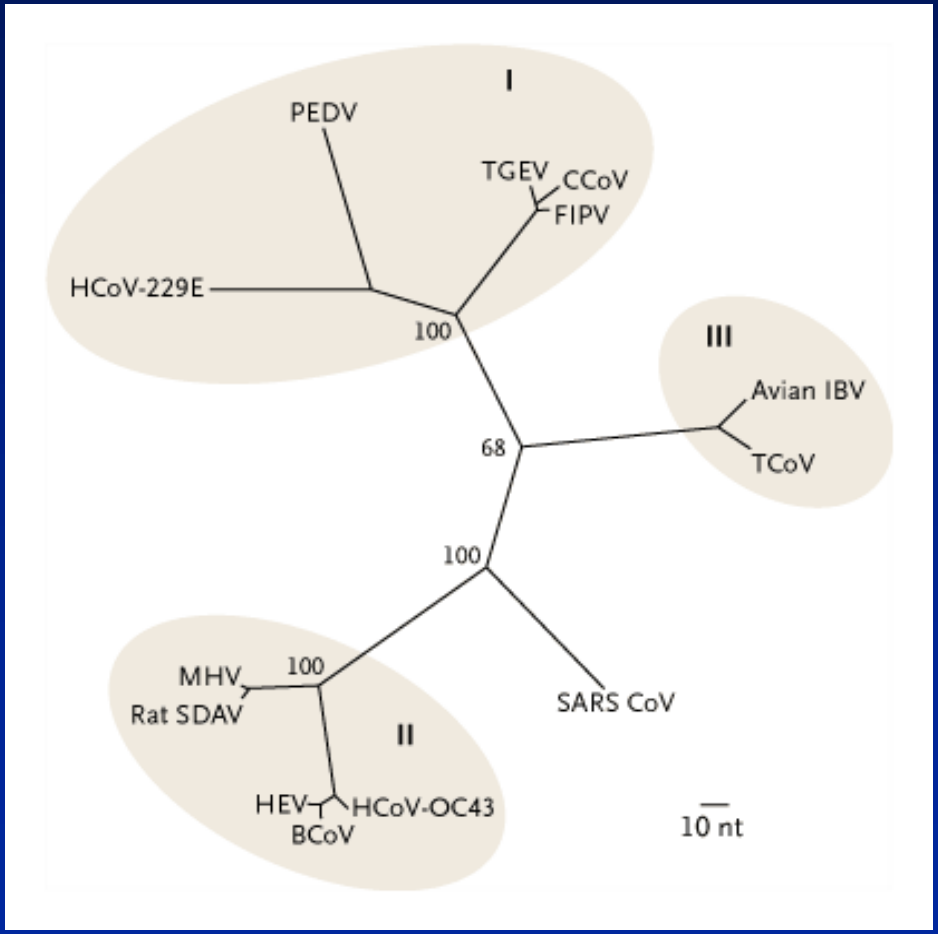




Our result is consistent with previous finding based on sequence alignment technique (*Science* 1986; **232**: 980)

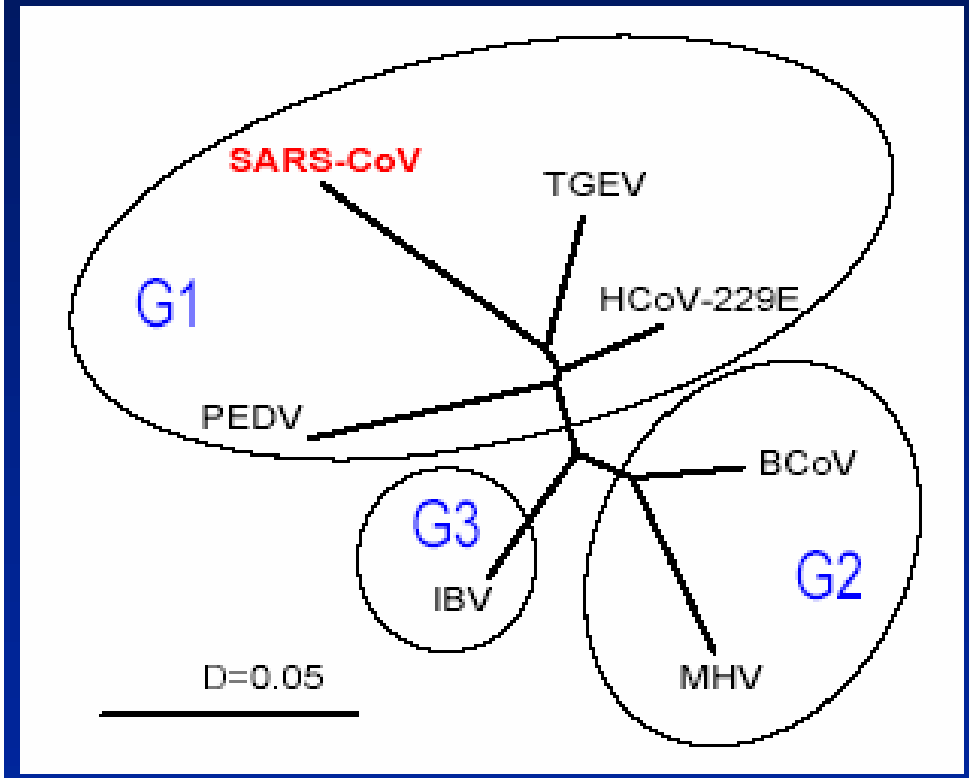
Classification of SARS Virus (~29751 bps)

WHO Result (based on **405 bps)**



N Engl J Med 2003; **348**: 1947

Our result (complete genome)



Yang, Goldberger, Peng. J Comput Biol 2005;12:1103-1116

Conclusions

- Biological signals are complex and contain hidden information
- Information categorization analysis based on repetitive elements may uncover certain interesting features relevant to modeling and diagnostics
- Applications range from literary texts, DNA sequences, to physiologic time series