Information Theoretic Clustering
of Astrobiology Documents

L. Miller & S. Still
Information & Computer Sciences, University of Hawaii at Mānoa

Information Bottleneck Clustering Method

Information Theory (C. Shannon, 1948)

- Reliable communication over noisy channels

Rate Distortion Theory

$$R(D) \equiv \min_{\{p(c|x):\langle d(x,c)\rangle \leq D\}} I(C;X)$$

- Gives theoretical limits on the rate x can be compressed to and successfully reconstructed if you know c.
- R, rate: bits per data sample transmitted •D, average distortion: "difference" between transmitted and received signal.

Mutual Information

$$I(C;X) = H(X) - H(X|C)$$

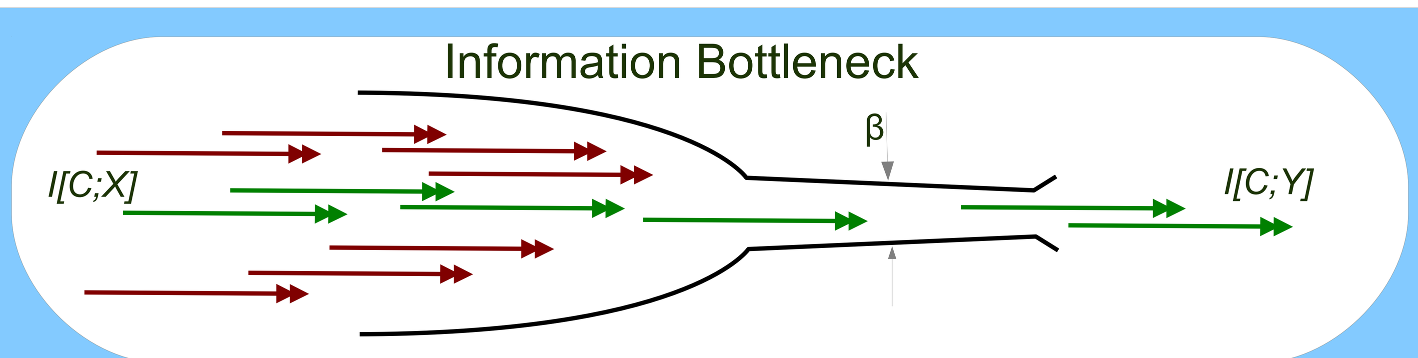
- Measures how much more certain you are about x if you know c.
- H is the entropy (measures uncertainty.)

Information Bottleneck (Tishby, Pereira, & Bialek, 1999)

- Define Y as the relevant variable you want to keep that is contained in X.
  - In our case, Y is the words in documents X.
- Instead of minimizing distortion, maximize the information kept about Y, while minimizing the information kept about X.
- β controls the trade-off between information preservation and compression.

$$\min_{p(c|x)} I(C;X) - \beta I(C;Y)$$

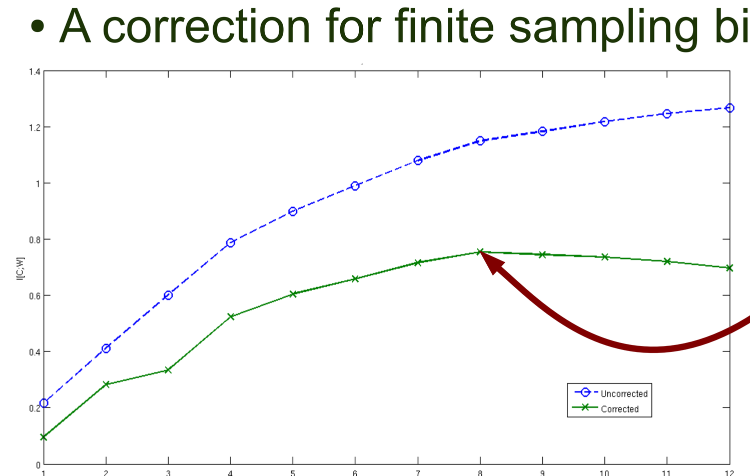
- ***Lossy of compression of documents into clusters of words.***



How Many Clusters to Make?

The maximum number of meaningful clusters in a dataset can be determined using:

- The mutual information between clusters and the relevant variable (words), $I[C;W]$
- A correction for finite sampling bias. (Still & Bialek 2006.)



The maximum of the corrected curve indicates the largest number of clusters that can be resolved in the data.

Preliminary Results

Some article titles from clusters created

- *Confocal laser scanning microscopy and Raman imagery of ancient microscopic fossils*
- **Late Archean rise of aerobic microbial ecosystems**
- *Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age*
- Evolution of uranium and thorium minerals
- **The worm turned, and the ocean followed**

- **Time–Evolution of Viscous Circumstellar Disks due to Photoevaporation by FUV, EUV and X-ray Radiation from the Central Star**
- **Diagnostic Line Emission From Extreme Ultraviolet And X-Ray-Illuminated Disks And Shocks Around Low-Mass Stars**
- RESOLVING THE CHEMISTRY IN THE DISK OF TW HYDRAE
- Titan's Methane as a Primordial Chemical Species

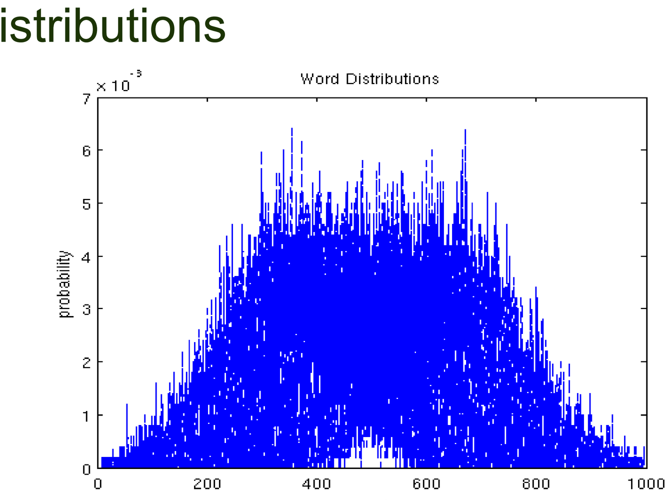
NAI 2009 Annual Report and attached publications

- **Extensive carbon isotopic heterogeneity among methane seep microbiota**
- Lipid biomarker and phylogenetic analyses to reveal archaeal biodiversity and distribution in hypersaline microbial mat and underlying sediment
- **Diversity of hopanoids and squalene-hopene cyclases across a tropical land-sea gradient**
- *Processes of carbonate precipitation in modern microbial mats*

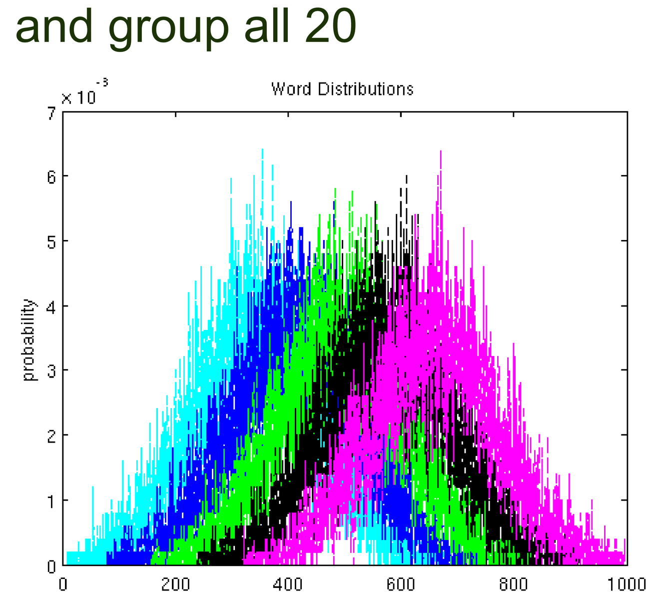
- ***Mars Bulk Composition***
- High-Precision Isotopic Studies of Meteorites
- ***High-precision SIMS oxygen, sulfur and iron stable isotope analyses of geological materials: accuracy, surface topography and crystal orientation***
- New frontiers in micro-analysis of isotopic compositions of natural materials: Development of Fe isotopes

Testing and Evaluation

- On synthetic data
  - 20 "document" datapoints with 5000 elements each
  - Drawn from 1 of 5 Gaussian distributions
  - Ranging over 1000 "words":



- We can recover all 5 distributions and group all 20
  - 100% precision and accuracy



Acknowledgments: This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science.