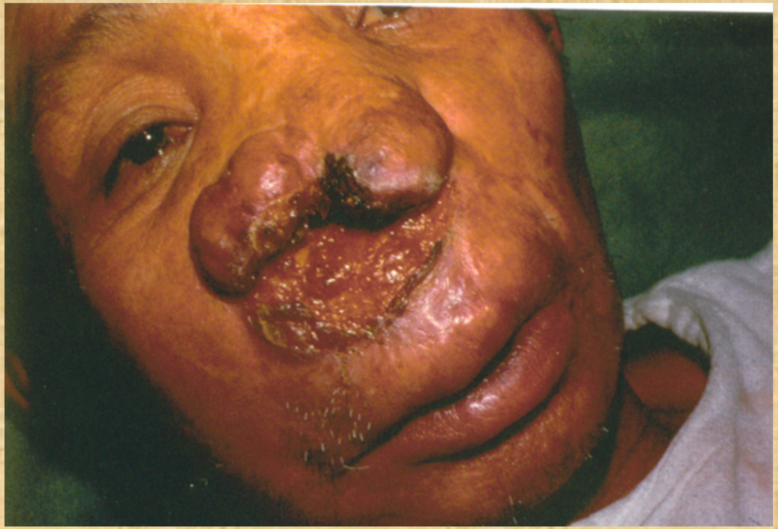
Topic Models with Relational Features for Drug Design

Ashwin Srinivasan
Tanveer Faruquie
Indrajit Bhattacharya
Ross D. King

Parasitic Tropical Diseases

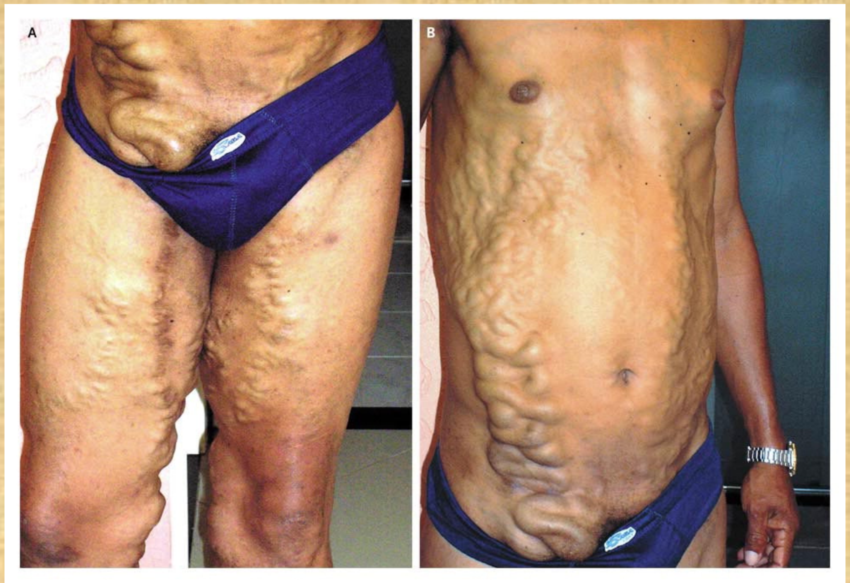
Malaria

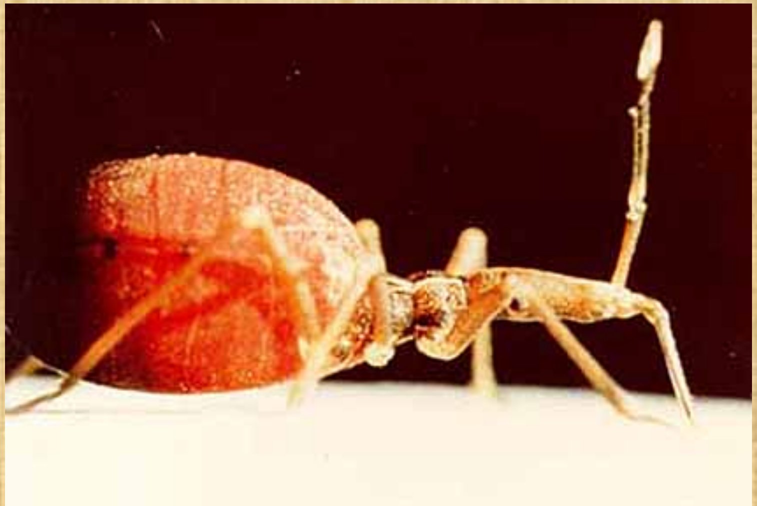


Shistosomaisis

Leishmania

Chagas



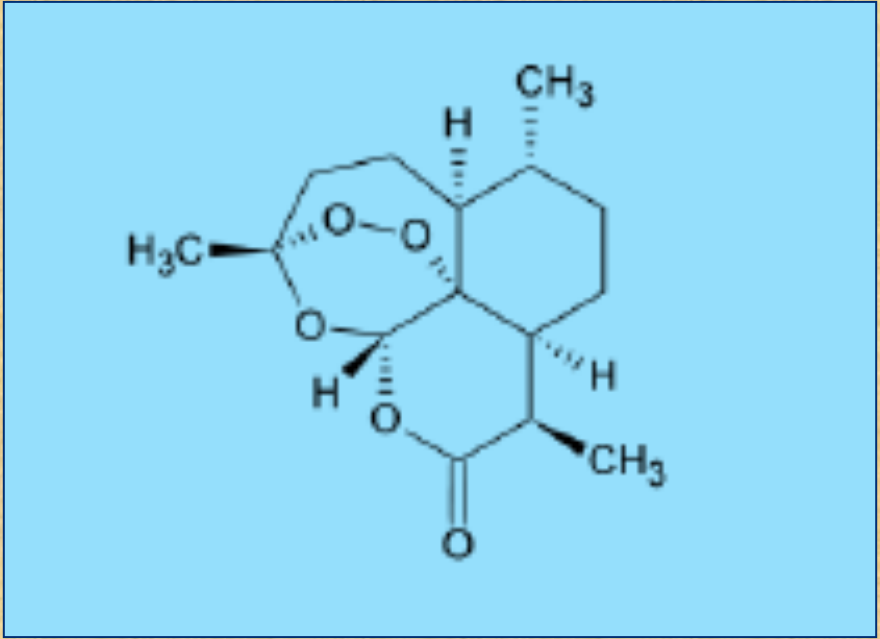


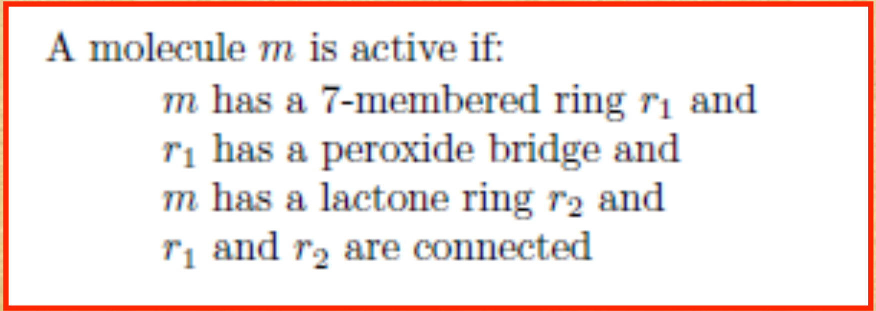
Why Tropical Diseases?

- Millions of people die of these diseases, and hundreds of millions of people suffer infection.
- It is clear how to cure these diseases – kill the parasites.
- They are “neglected”, so avoid competition from the Pharmaceutical industry.

Eve

Artemisinin





A molecule m is active if:

m has a 7-membered ring r_1 and
r_1 has a peroxide bridge and
m has a lactone ring r_2 and
r_1 and r_2 are connected

“Business-as-Usual”: ILP-based discovery

- An ILP system can discover a rule to describe structures like this
  - Background knowledge of ring structures, peroxide bridges, connected rings etc.
  - Data of artemisinin-like compounds and their efficacy
- But, there are limitations
  - Cannot (easily) discover clusters of useful sub-concepts
  - Cannot (easily) weight different sub-concepts to generate new molecules stochastically
  - Cannot (easily) account for uncertainty arising from noisy biology

Topic Models for Molecules

- Originally used in the analysis of text documents, is concerned with three principal entities: documents, topics, and words. Documents consist of one or more topics, which in turn consist of one or more words.
- Molecules ("documents") will be taken to consist of one or more concepts ("topics")
  - For example, concepts may be like: "activity" and "toxicity".
- Concepts will be taken to consist of one or more features ("words", although more like phrases)
  - For example, an active molecule may consist of the following features: a 7-membered ring, a lactone ring connected to a 7-membered ring, and a peroxide bridge in a 7-membered ring.

A Probabilistic Model for Molecules

- **Given**: A set of (Boolean) features, and a set of molecules:
  1. Associated with each molecule *m* is a multinomial distribution over the entire set of K concepts. The parameter θ_m of this distribution gives the probability of observing the K concepts for the molecule *m*. It is assumed that the molecule-specific concept-distributions are drawn using some prior distribution over multinomial parameters

  2. Associated with each concept *k* is a multinomial distribution over the entire set of V features. The parameter ϕ_k of this gives the probability of observing the features for concept *k*. It is assumed that these feature-specific distributions are in fact drawn using some prior distribution over multinomial parameters

- Once priors and multinomials are known (estimated) we can compute posterior probabilities, or “generate” molecules randomly

What is the Result?

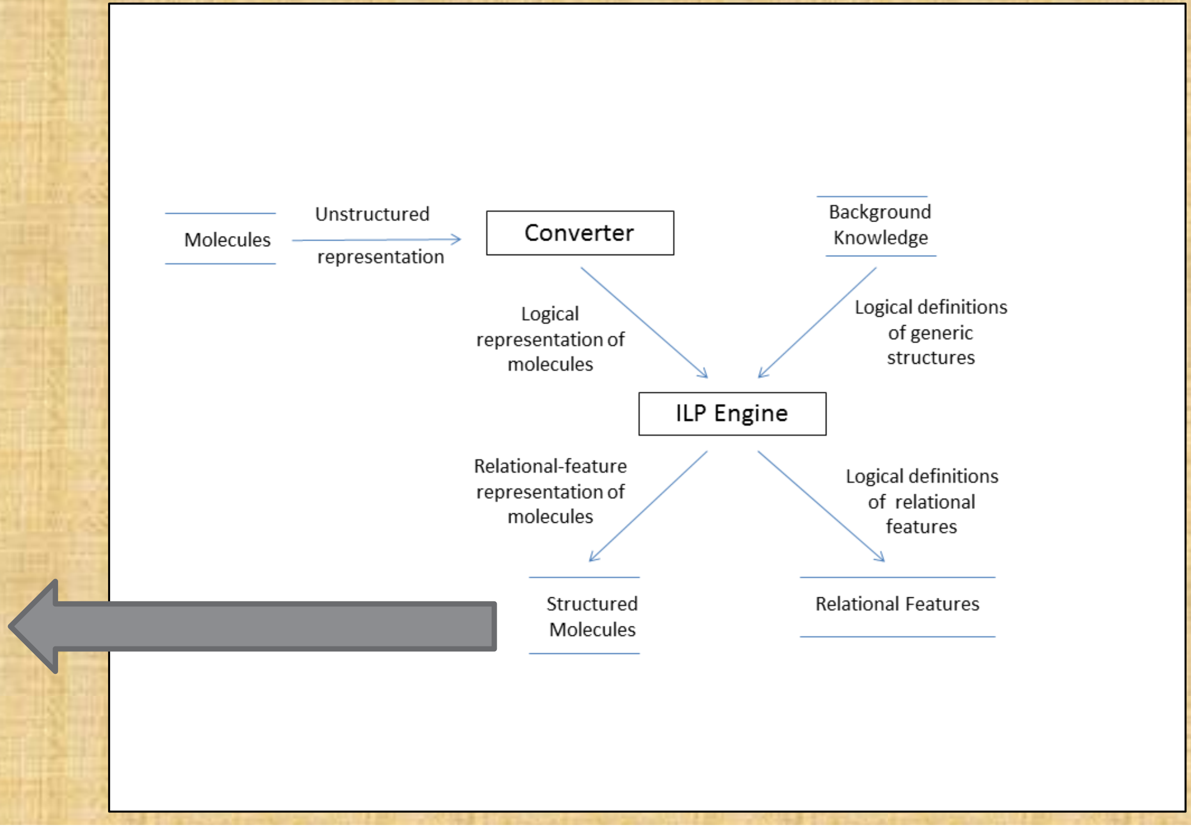
Molecules	Features							
	F_1	F_2						F_V
m1	1	0	0	0	1	…	…	…	1
m2	0	0	1	1	…	…	…	…	0

Extract Automatically

Molecules	Topics (subsets of features)			
	T_1	T_2		T_K
m1	0.1	0.3	…	0.4
m2	0.3	0.1	…	0.1

Where do the Features Come From?

- Since the early 1990s, a very effective use of ILP systems has been as engines for discovering relational features



Molecules

Unstructured representation

Converter

Background Knowledge

Logical representation of molecules

Logical definitions of generic structures

ILP Engine

Relational-feature representation of molecules

Logical definitions of relational features

Structured Molecules

Relational Features

Topic Model

Topic Models with Relational Features for Drug Design

- Three kinds of problems in drug design:
    1. Discrimination of active molecules (classification)
    2. Finding molecules similar to a specific kinds of active molecules with a known target (retrieval and ranking)
    3. Generating molecules that share structural properties with known active molecules (synthesis)

- Today, we will report on (1), and indicate what we are doing (have done) towards (2) and (3)

Problem 1: Discrimination

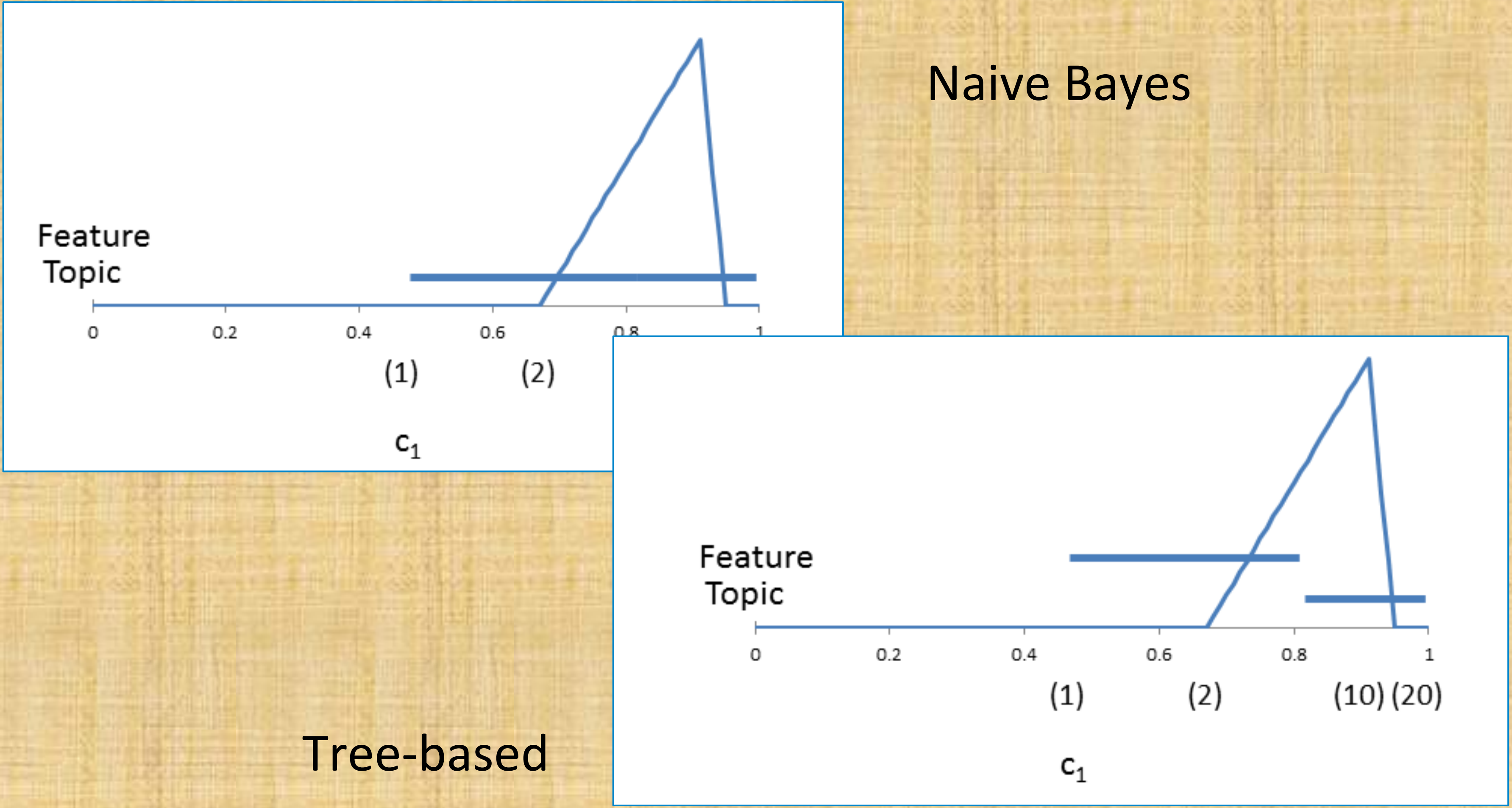
- **Data and Problem:**
  - The Tres Cantos Antimalarial TCAMS dataset (freely available)
  - Screening GlaxoSmithKline's library of approximately 2 million compounds. The database consists of 13,000 of chemicals that were found, on screening, to inhibit significantly the growth of the 3D7 strain of *P. falciparum* in human erythrocytes
  - Task: identify molecules in the top 15-percentile of activity
- **Approximate Differential Costs:**

	Predicted		
Actual		Less Active	Very Active
	Less Active	0	1
	Very Active	10	-10

Problem 1: Materials & Method

- **Background Knowledge and Algorithms:**
  - Standard definitions of cyclic structures and groups used in the past
  - Conversion programs from SMILES representation to Prolog
  - ILP engine for constructing features (Aleph)
  - Programs for constructing topic models (R)
  - Programs for using cost-sensitive classification using topic-vectors (WEKA)
- **Method:**
  1. Partition data into “training” and “test”
  2. Learn features using the ILP engine and training data
  3. For a distribution of costs around the values specified: learn topic models
  4. Compare classification models (on test data) constructed in “topic-space” against those constructed in the original “feature-space”

Problem 1: Results

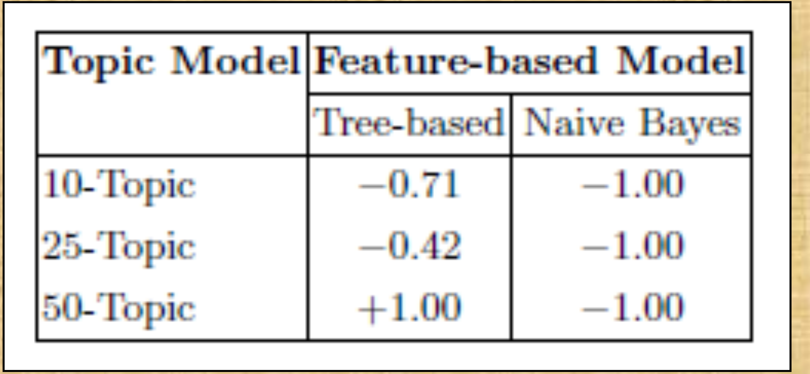


Naive Bayes

Tree-based

Problem 1: Results

- **Quantitative**
  - See Adams and Hand, “Comparing classifiers when misallocation costs are uncertain”, *Patt. Recog.* 32, 1139-1147 (1999)
  - LC-index between -1 and +1 (-ve values means topic model is better, averaged across costs)



Topic Model	**Feature-based Model**	
	Tree-based	Naive Bayes
10-Topic	-0.71	-1.00
25-Topic	-0.42	-1.00
50-Topic	$+1.00$	-1.00

Problem 1: Sanity Checks

- **Simple feature-selection?**
  - Are topic-models doing anything other than simple feature selection?
  - Yes. A 25-topic model does better than simply selecting the best 25 features from the data etc.

- **Conditional independence?**
  - The topic modelling technique used requires conditional independence of the features, given the molecule-specific multinomial. Does this hold?
  - Probably. The features are generated using a random strategy, and are largely dissimilar (using a simple Jaccard index calculation)

Problem 2: Retrieval

- **Data and Problem:**
  - Subset of molecules known to inhibit Dihydrofolate Reductase in *P. falciparum* (freely available: few hundreds)
  - The Tres Cantos Antimalarial TCAMS dataset (freely available, several thousands), which contains molecules active against *P. falciparum*, but with targets mostly unclear.
  - Task: identify molecules in the TCAMS dataset that are most likely to be DHFR inhibitors
- **Materials and Method:**
  - Background knowledge as before
  - Construct ILP-features for known DHFR inhibitors
  - Construct topic models using these features
  - Use topic distribution vectors to rank molecules in TCAMS
  - Compare against using feature vectors to rank molecules in TCAMS

Problem 2: Snapshot of Results

- **Topic-Based and Feature-Based Rankings**
  - We have constructed topic models for the DHFR inhibitors and used the model to compute representations for the TCAMS dataset in "topic-space"
  - We have ranked molecules in TCAMS based on aggregate similarity to the DHFR inhibitors, based on the original feature-values and based on topic-distribution values
  - At this point, we are able to state that the ranking of the TCAMS based on the original Boolean features is significantly different to the ranking based on topics
  - But, which is better? For example, how many of the top-k ranked molecules in each ordering really are DHFR inhibitors?
  - We are investigating this.

Problem 3: Synthesis

- **Data and Problem:**
  - Task: generate molecule substructures that are shared with molecules in this class, using their distribution of occurrence in the class

- **Materials and Method:**
  - Background knowledge as before
  - Construct ILP-features for molecules in the specified class
  - Construct topic models using these features
  - Use the distributions to generate molecule fragments
  - Validate using wet-lab experiments and expertise

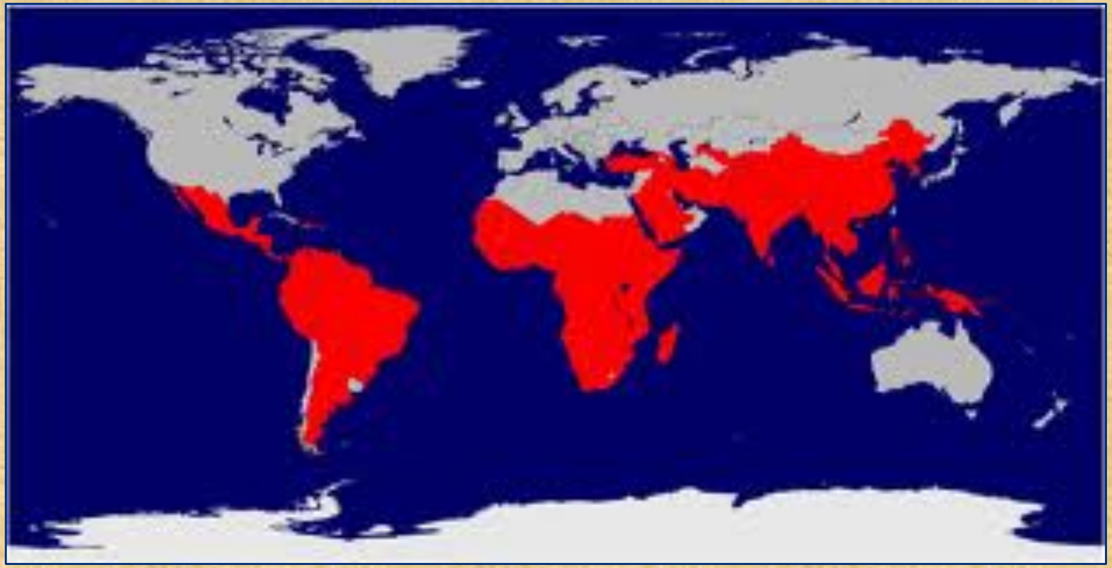
Concluding Remarks

- **Hierarchical Bayesian Models with Relational Features**
  - Combine the advantages of ILP (explicit use of background knowledge, discovery of relational features) with the power of a parametric model. Result: a "poor man's SRL": ILP engine for features + Probabilistic Model + Standard Methods for Classification/Retrieval
  - First-time for ILP-based drug design
  - Results obtained are good: better to operate in "topic-space" than in the original high-dimension feature-space that results from "propositionalisation". Complete version of the paper will contain findings for Problems 2 and 3 as well.
  - Some further advantages: discovery of topics (sub-concepts), handling uncertainty, and a well-specified probabilistic model for generating molecular substructures
  - A key limitation: need to pre-determine the number of topics. This can be overcome by using a different kind of topic model

Acknowledgements

- Dept. of Science & Technology, Govt. of India
- Royal Academy of Engineering, UK
- Dept. of Computer Science, Oxford
- School of Computer Science & Eng., UNSW
- School of Computational & Integrative Science, JNU
- Dept. of Computer Science, SAU

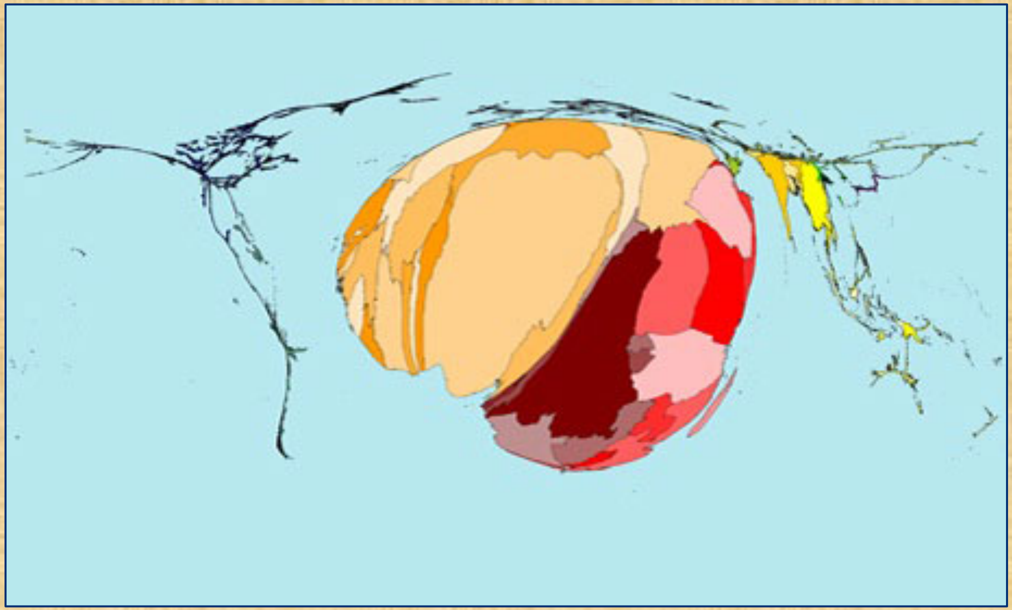
Malaria: 216m cases, 3.3b affected



Most parts of Asia, sub-Saharan Africa and South America

Source: Am. Jnl. Emergency Med. (2012), 30(6), 972-980.

Malaria: 655,000 deaths, mostly children



A country's size is proportional to the deaths that occur there. (Africa and Asia dominate)

Source: Benjamin D Hennig, University of Sheffield / UNICEF

A Growing Problem: Drug Resistance

- The WHO has two substantial reports on the emergence of drug-resistance across the world. Large parts are already resistant to Aminoquinolines (chloroquine etc.)

- Evidence is now emerging of resistance to the most effective drug family (Sesquiterpene lactones: artemisinin etc.)

- Key recommendation: New anti-malarial drugs as alternatives to artemisinin