Young H. Cho, Ph.D.
University of Southern California

EE 542
Lecture 19: Machine Learning

USC Viterbi School of Engineering

Final Project Proposal

- Requirement
  - xDot + Gateway + Node Red Data collection
  - Amazon Web Services + Thingsboard
  - Data Analytics/Machine Learning on Data
  - Inferential or Model-based Result
  - Real world problem that needs solution
  - Novel solution
- Submissions
  - Final Project Proposal: Summary Outline, Oct 25
  - Final Project Progress Videos
  - Final Project: Final Report, Software Source Package, Slides, Final video Due Dec 13

Final Project Presentation

- Audience
  - Investor/Board of Directors/CEO
  - Grand parents/People without expertise
  - Technology Expert
- Composition
  - Attention Grab/Relevance/Problem to Investor+People without expertise
  - Most important/novel aspect using a detailed example
  - Summary of Result showing Supriority
  - Summary of what that means: Money? Safety? Good?
- Methodology
  - Use of Animations and Pictures
  - Minimum Words
  - Practiced Talk

Common Myths of Machine Learning

- Myth#1: Machines can learn autonomously
- Reality: Machine learning is carefully architected by a programmer and trained with the necessary training data. Most of the machine learning algorithms require large amounts of structured data that are often manually filtered and fed into the algorithm.

Common Myths of Machine Learning

- Myth#2: Machines can learn like humans
- Reality: If we compare the learning process of a machine with that of a child, it becomes evident that machine learning is still in its infancy. For example, a baby doesn't need to listen to millions of other humans before it learns how to talk. Machines on the other hand requires guidance and support at each step of learning.

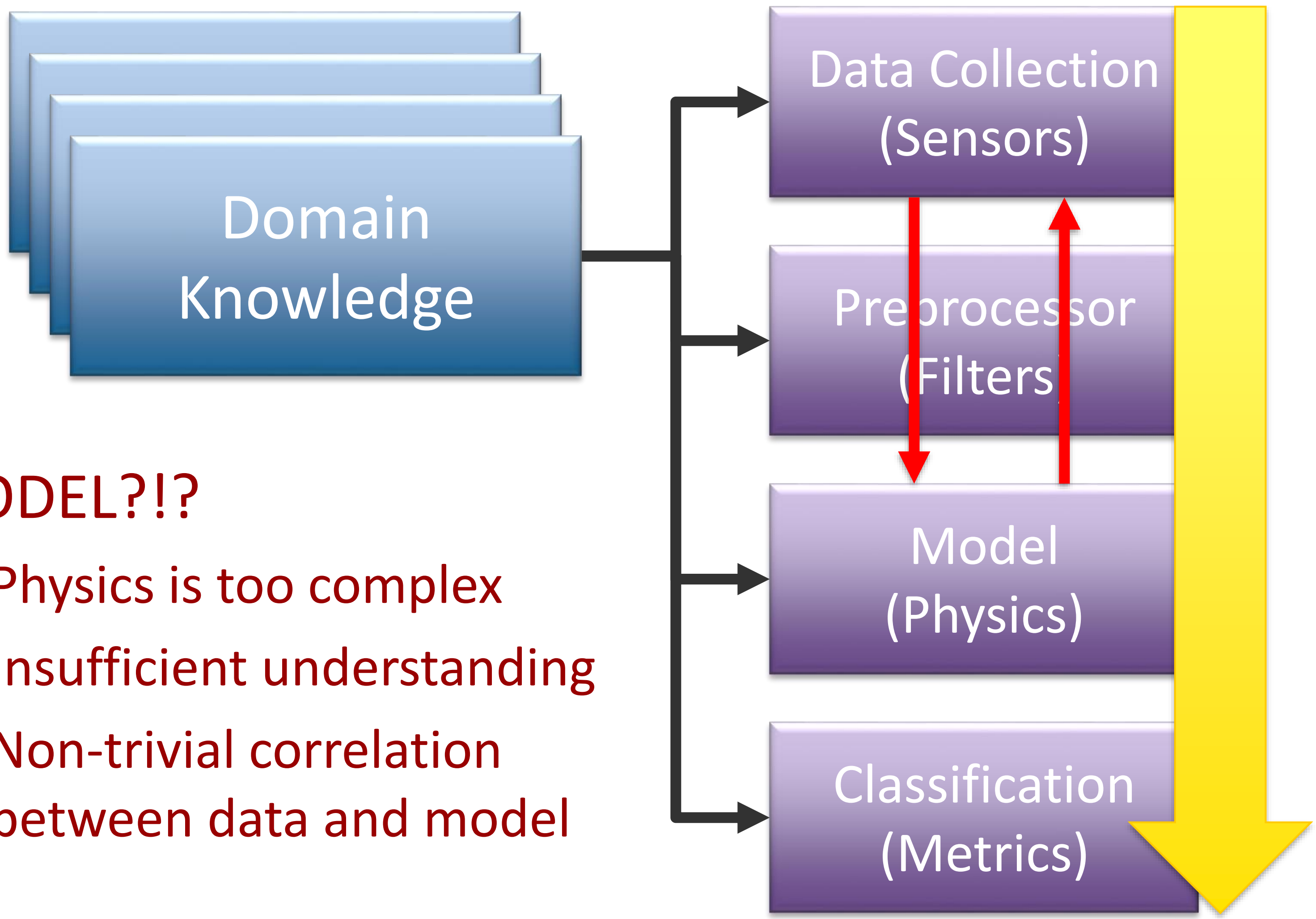
Common Myths of Machine Learning

- Myth#3: Machine learning can be applied to any task
- Reality: Currently, machine learning can only be applied to tasks where large and sufficient number of input data sets exist or can potentially be captured. Most of the successes in AI have come in the applications where companies like Google and Facebook have access to enormous data sets (texts, voices or images) coming from a variety of sources.

USC Viterbi
School of Engineering

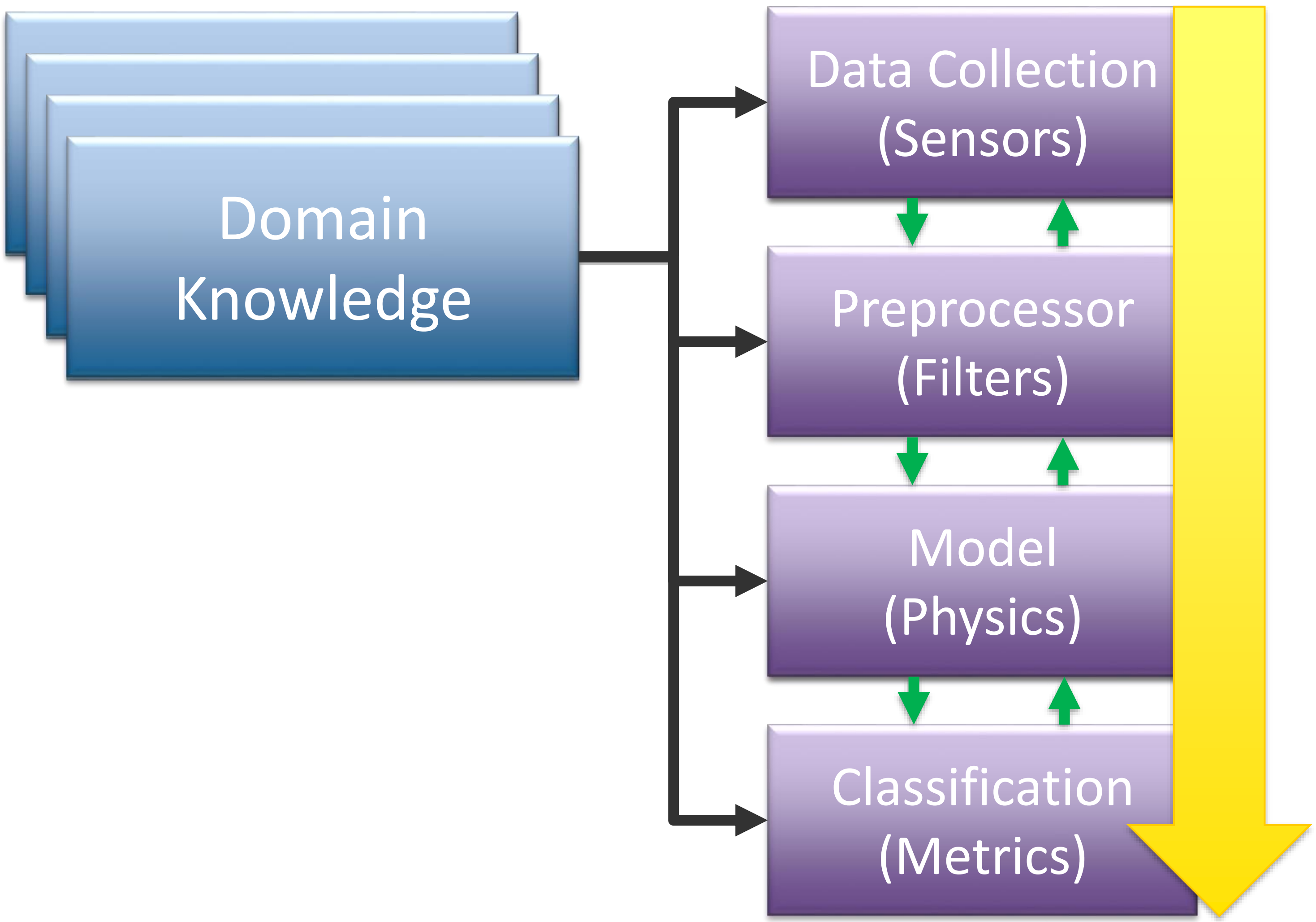
Traditional Data Analytics

Current ML Approach



- MODEL?!?
  - Physics is too complex
  - Insufficient understanding
  - Non-trivial correlation between data and model

My Thoughts


Domain Knowledge
Data Collection (Sensors)
Preprocessor (Filters)
Model (Physics)
Classification (Metrics)

Machine Learning Approaches

	Unsupervised Learning	Supervised Learning
Discrete	Clustering	Classification
Continuous	Dimensionality Reduction	Regression

Unsupervised Learning Experience

- Goal: Determine Meaning of Internet Document Content then Hierarchically and Dynamically Cluster and Discover New Topics using Hardware Accelerator

- Application: Discovering and Filter Topics of Interest on the Internet in Real-time

- Funded by US Department of Defense 2004-2007

Processing Raw Documents

- N-Gram Analysis
  - Samples taken at every byte offset
  - Multiple lengths of n-grams sampled

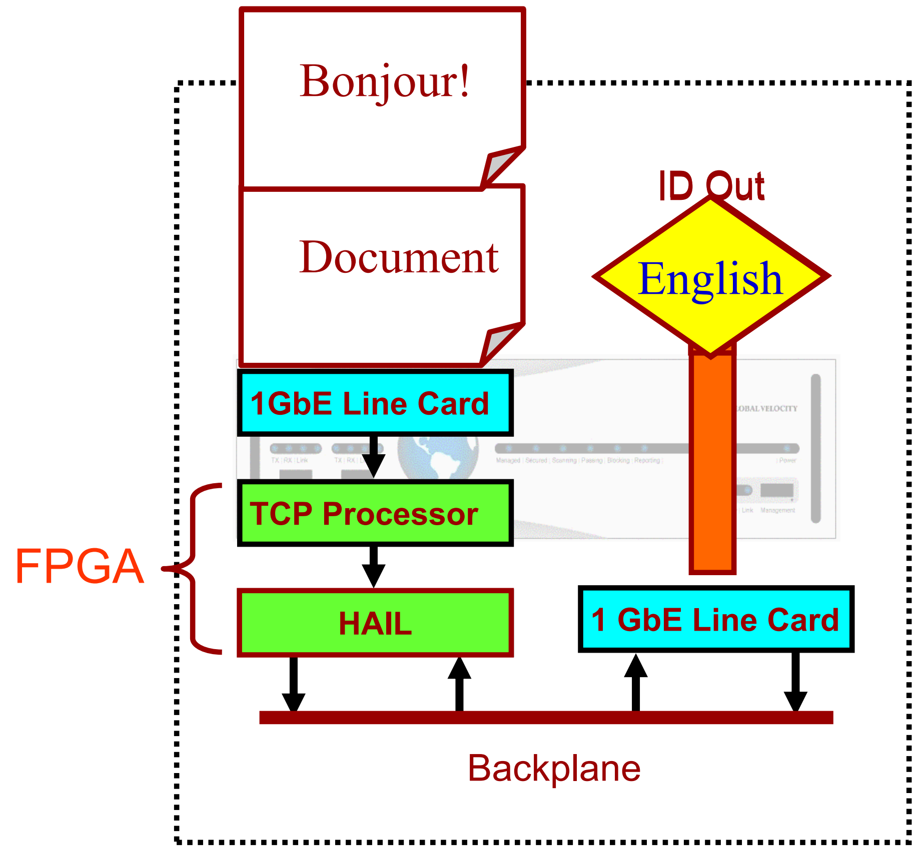
FPGAs are

Language Identification

- Keep Track of Word Count
  - Compare only the relevant languages
  - Count subsequent appearance of words
- Performance
  - Significant reductions for comparisons
  - Little or no impact on accuracy

Hardware Implementation

- Implementation
  - GVS-1000 Platform
  - Stackable FPGA Cards
- System Features
  - Highly Customizable
    - Reconfigurable
    - Modular Unit
  - High Accuracy
    - 99.8%+ *raw* text docs
  - Low Latency
    - Able to ID single packet
  - High Performance
    - 2.4+ Gigabits/second



Identifying Document Structure

TAGBOT

HTML Source Document

hdr2 strg hdr2
<h2>Company Overview</h2>
comm para strg link href
<!-- Corporate Fact Sheet --> <p>Founded by <a href=
quot strg link
"/about/profile.html">Dr. J. Robert Beyster and a
small group of scientists in 1969, SAIC, a Fortune 500
strg
company, now ranks … and have more than 43,000 <!- Also
comm
update employee number on: saic.com/news/0722.html -->
strg para
employees with offices in over 150 cities. </p>

Token List

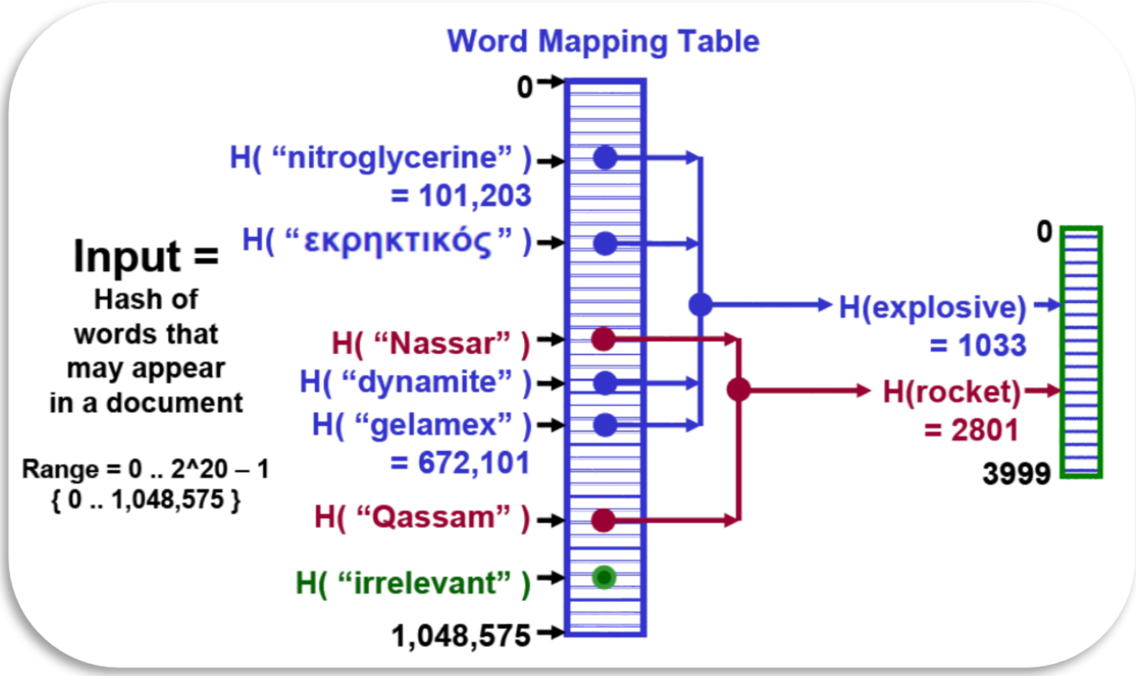
```
(1) hdr2 : 'h2'
(2) para : 'p'
(3) link : 'a'
(4) href : 'href='
(5) quot : '"'.alpnum*.'"'
(6) comm : alpnum*
(7) strg : alpnum*
```

Simple Grammar

```
(1) Tag_Name → hdr2 | para | link
(2) Comment  → '<!-'.comm.'-->'
(3) Attrib   → href.quot | ε
(4) Tag_Head → '<'.Tag_Name.Attrib.'>'
(5) Tag_Tail → '</'.Tag_Name.'>'
(6) Expr     → Comment | strg | ε
(7) Line     → Tag_Head.Line.Tag_Tail
              | Expr.Line.Expr | Expr
(8) Content  → Line.Content
```

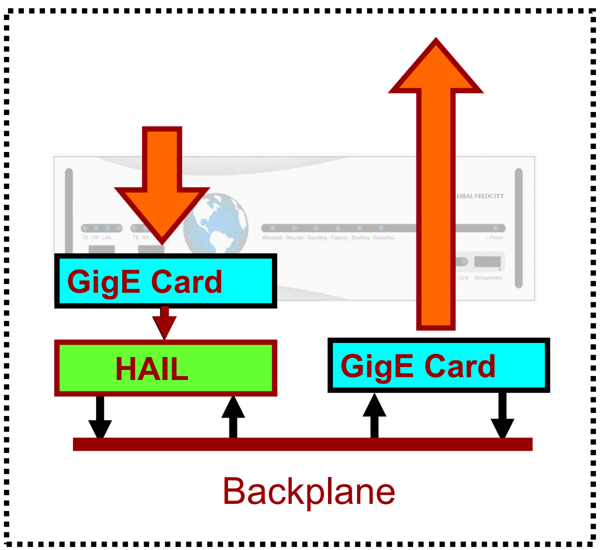
Document Classification

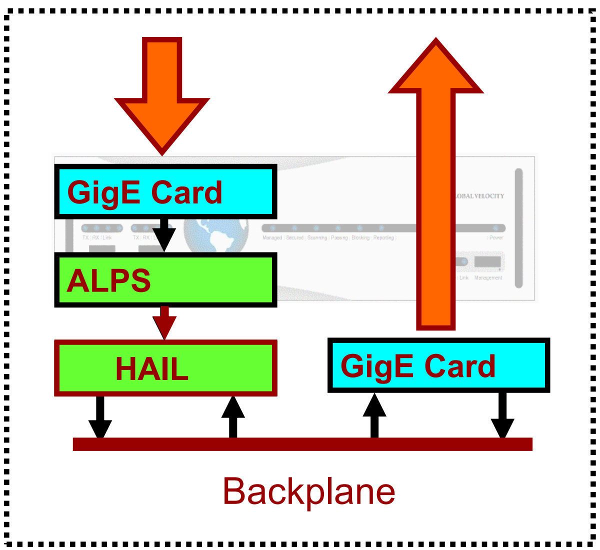
- Documents transformed to 4000-wide numerical vectors with 4-bit dynamic range
- Document similarity computed based on vector similarity



Application Level Processor

- Experiments
  - Five different email data sets were created
    - 75-bytes, 150-bytes, 300-bytes, 600-bytes, 1200-bytes
  - 10,816 email messages per data set
  - 14 different language documents in each data set

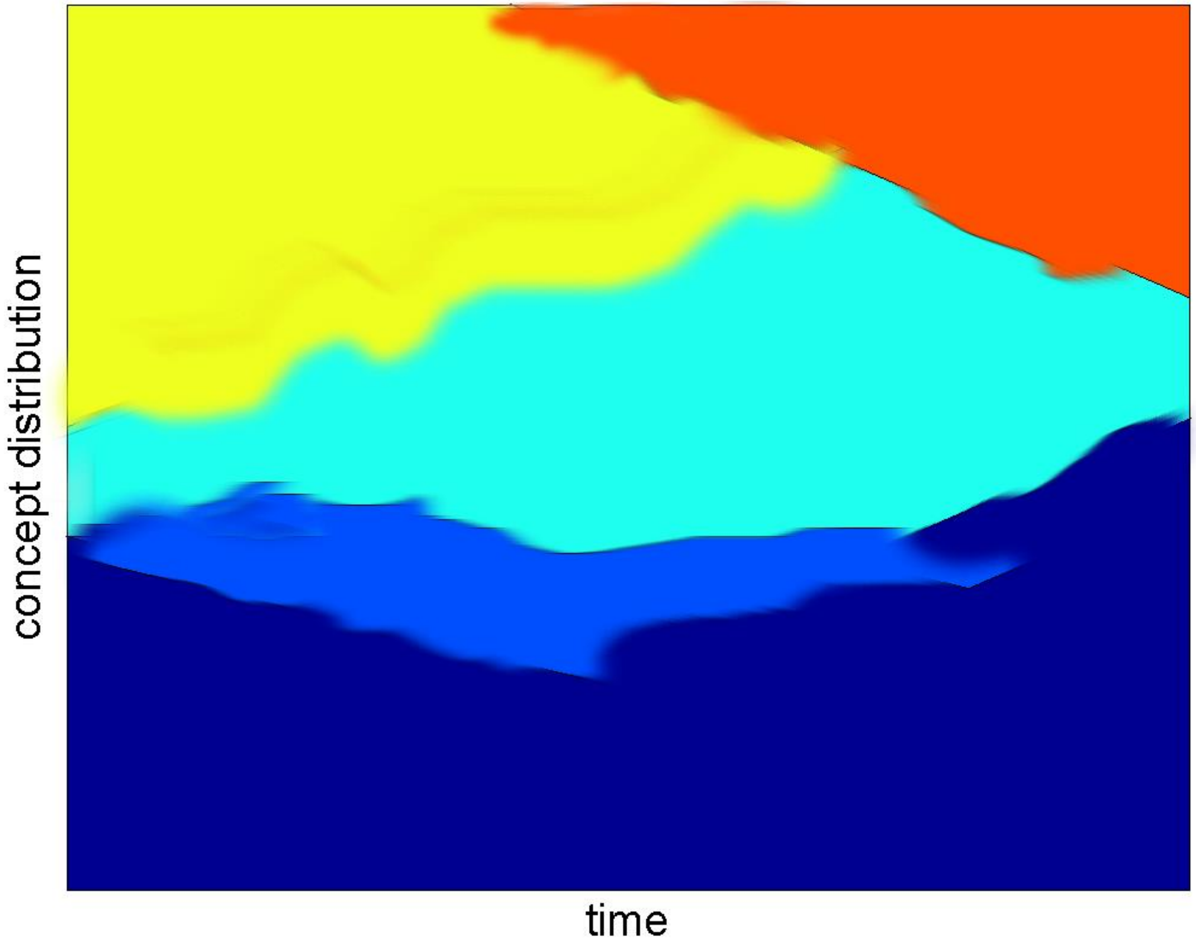




Hierarchical Clustering

- Document Insertion
  - Adds a single document to an existing tree
  - Top-down descent, greedy matching
- Batch Clustering
  - Original top-down (global) hierarchical clustering
  - Necessary to avoid getting stuck in local optima
- Document Removal
  1. Dead topics
     - Least Recently Used (LRU) caching (or an approximation)
  2. Over-representation
     - Don’t store multiple copies of (nearly) identical documents

Concept Drift



USC Viterbi School of Engineering

Results

- Algorithms trained and tested on CMU 20-newsgroups
  - Standard benchmark (13,000 messages)
  - Added “noise” from talk.origins (11,000 messages)
  - Used K=60, flattened hierarchies to be comparable
- K-means results
  - Vast majority of documents in only two clusters
  - Few concepts discovered
  - Discovered many meaningful concepts
  - However, ~50% of all concepts dominated by noise data
- Streaming Hierarchical Partitioning results
  - Discovered many meaningful concepts
  - Noise data effectively isolated to ~10% of concepts
- **Discovered Concept Vectors leading to Supervised Learning**

Lessons Learned

- Need Enough High Quality Data to Train
  - It is very difficult to find good data
  - Corpus of data needed a lot of manual prefiltering
- Erroneous Data Leads to Poor Result
  - Better to have less data than bad data