# Challenges in Data-driven Agriculture

Anne Denton

# Data-driven Agriculture

- What do we mean by that?
- Big data
  - Geospatial data tend to be large
  - Time adds a dimension
  - Many data relate to grower practices
- Treasure trove for data miners?
- I didn't actually encounter any clear data miner at the International Conference for Precision Agriculture

# Conventional Topic: Precision Agriculture

- Originally: Soil-specific agriculture
- Development of zone maps for managing nutrients at sub-field level
- Big business: Precision application requires special equipment
- Crop consultants help in developing zone maps

# Existing Data Mining Work

- Classic problem: Predict sub-field-level yield based on satellite imagery
- Some special problems:  Predict specific pests
- In practice, yield prediction is often done based on specific crop models
- Data mining of many fields?
  - Data are not centrally collected
- Some work is done with data aggregated at county level
- Problem:  Generally no central data collection

# Special Opportunity: Work With American Crystal Sugar

- Origins of collaboration were Capstone course
- Scientific problem, around which consortium was created
  - Prediction of annual sugarbeet yield
- Important additional potential
  - American Crystal Sugar is a cooperative
  - Collects far more information than is commonly assembled

# Funding

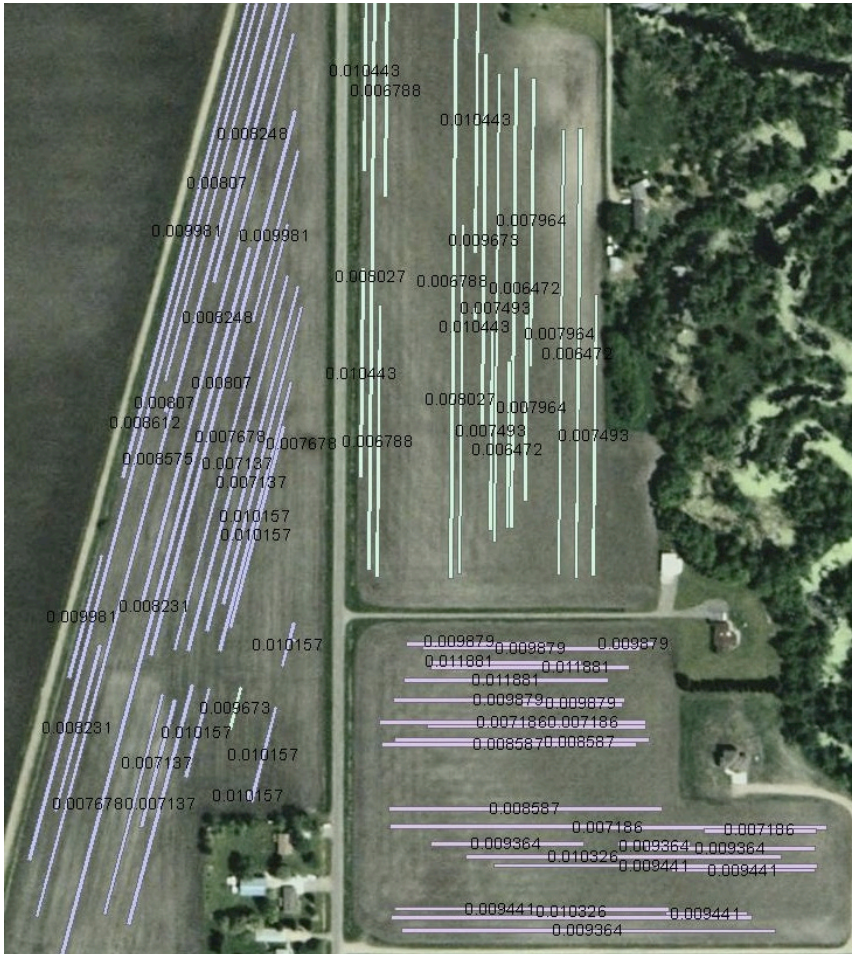
- Consortium
  - American Crystal Sugar
  - Phoenix International (John Deere)
  - RDO Equipment
- NSF-PFI Grant
  - All of the above plus three small companies
    - Holland Scientific
    - Meridian
    - Agri ImaGIS
  - Co-PIs from multiple related disciplines
    - Dave Franzen
    - John Nowatzki
    - Kambiz Fahramand
    - Phil Boudjouk

# Specific Problems

- Consortium data
  - Year 1: Sub-field-level yield from truck information
  - Year 2: Field-level data
- Data collection and processing
  - Interesting task for distributed processing
- Annual yield problem
  - Very different from regular yield prediction
- Vector-item pattern mining problem
  - What my group has been doing in other contexts
- Computational infrastructure
  - Cloud issues

# Sub-field-level Yield (Consort. Y1)

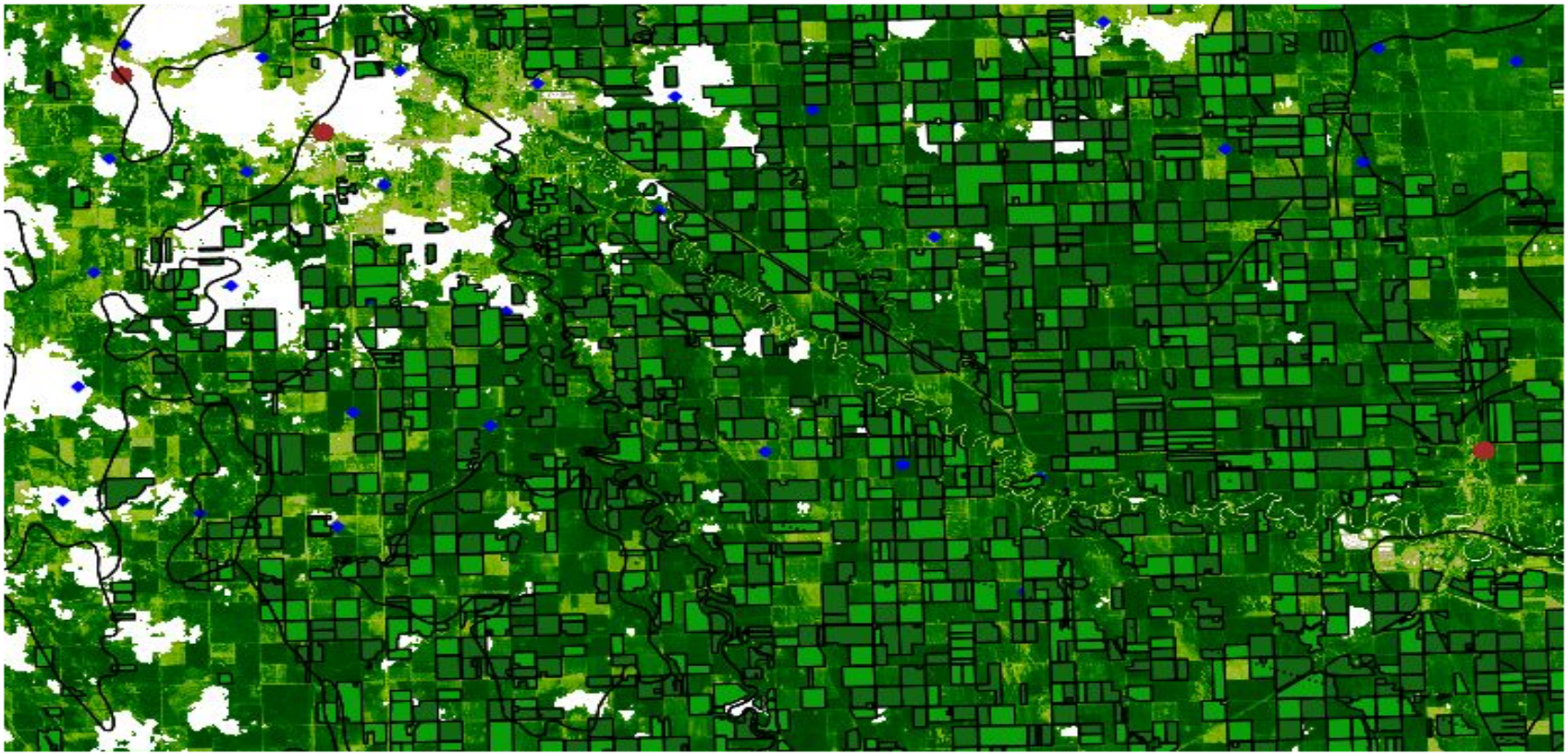
- Undergraduates did a fabulous job
  - Zech Andersen
  - Ben Bechtold
  - Sean Gerhard
- Many practical problems
- Unfortunately coverage was not sufficient



# Consortium Year 2

- Tabular data from Crystal Sugar for > 4000 fields
  - Planting date
  - Soil type
  - Preceding crop
  - Beet variety
- Meteorological data available from NWS
- Satellite image data from Landsat satellites
- Growing degree days calculated

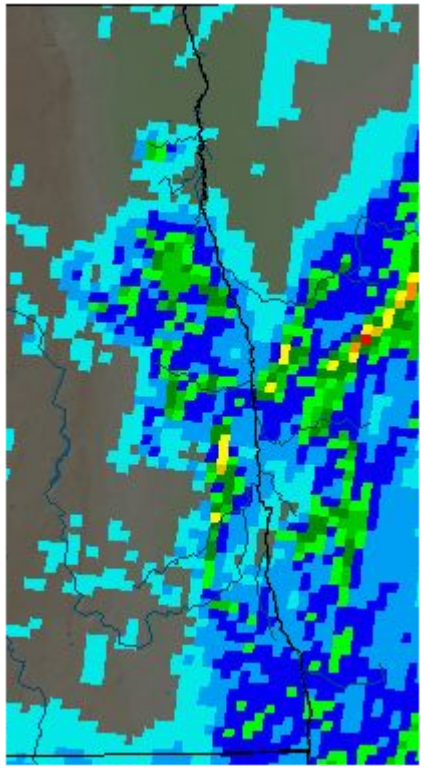
# Data Sources and their Scale



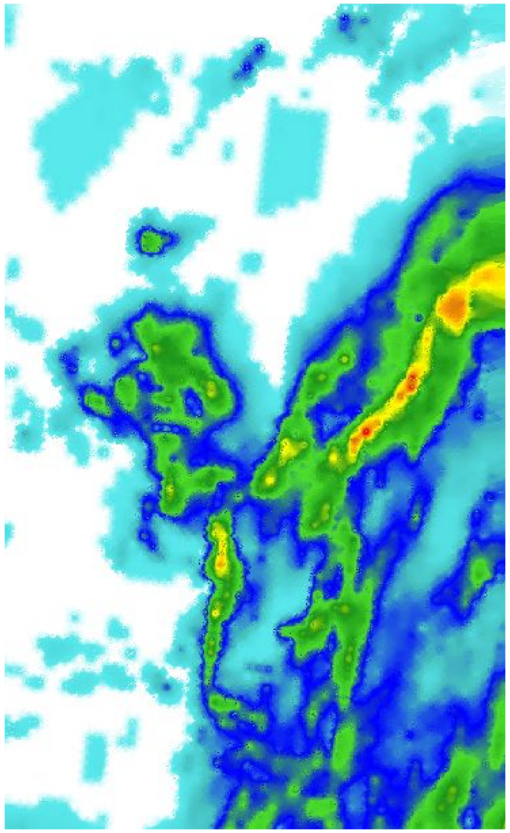
- NDVI, good detail for field level, but clouds over 50.1% of fields in July/Aug/Sept
- Red circles for weather stations (temperature)
- Blue diamonds for precipitation estimates by NWS - HRAP grid

# Data Processing Substantial

NWS website



GRASS – available to query



# Parts Amenable to MapReduce

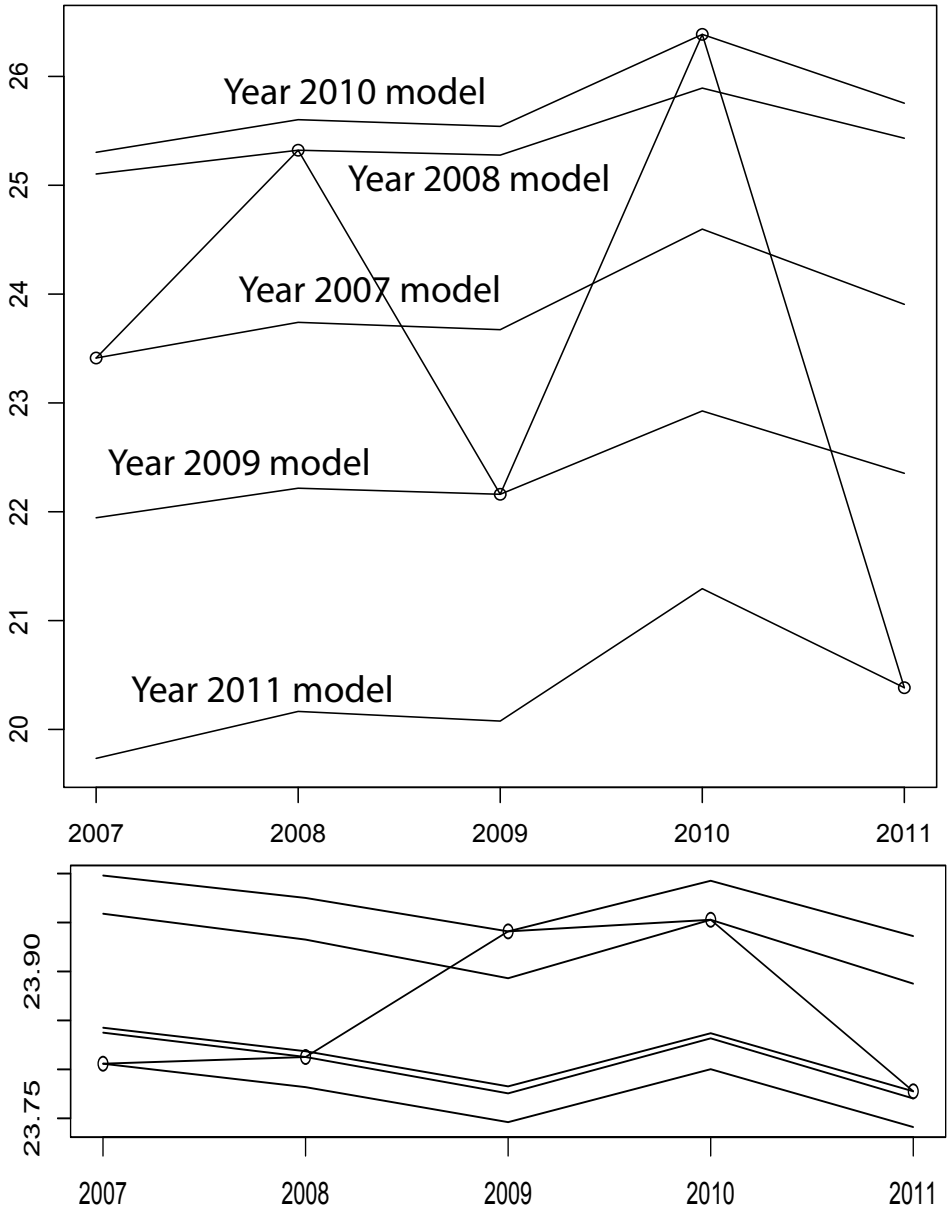
- Hadoop MapReduce suitable for some tasks in processing raster images
- MapReduce concept introduced by Google and now available as part of open source project Hadoop
- Works well when majority of task can be parallelized (map function)
- Aggregation small part of task (reduce task)

# Annual Yield Problem

- Only 5 years of relevant data
- Differs from normal regression problem
  - Normal evaluation procedures assume that goal is accuracy at record level
  - Differences between years much higher than would be expected

# How does this affect prediction?

- Example of trivial model: BT100FT
- Very different models built each year
- Bottom: Random selection of points



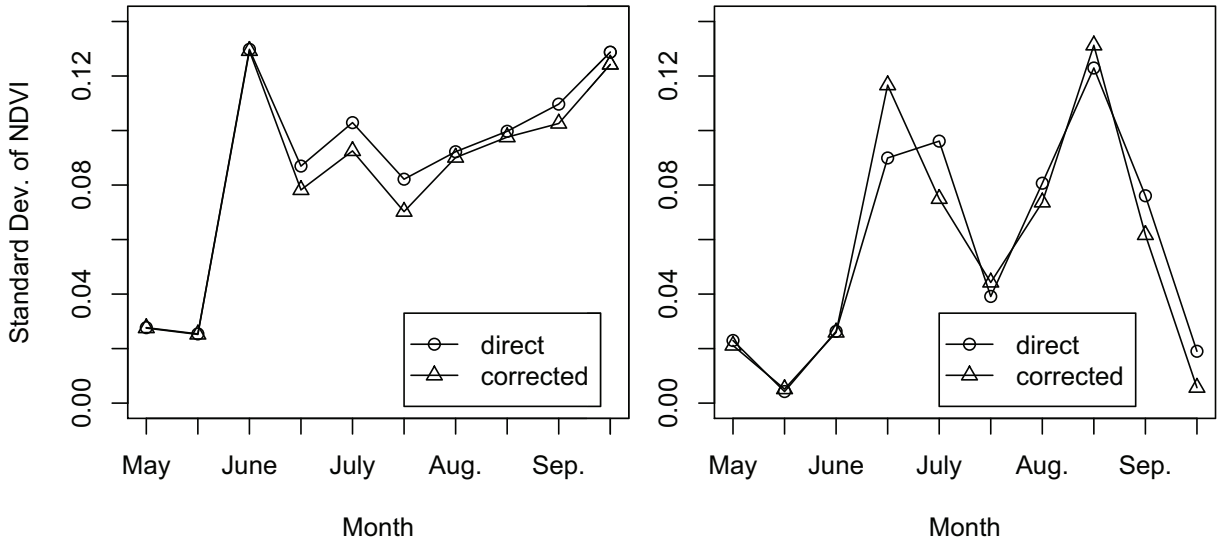
# Economics Analogy

- Predicting gross national product is a macro-economics problem
- Predicting company performance is a micro-economics problem
- Could one predict GNP by aggregating public company information?
  - Not done. Why not?
  - Really relevant predictors affect all companies in the same way
- GNP prediction is notorious for overuse of data
  - "Data mining" used to be swear word for this

# Systematic Errors vs. Random Errors

- Does it matter if we get soil type of one field wrong?
  - Matters for prediction of that field
  - Law of large numbers suggests that's irrelevant for grand total
- Does it matter if satellite image was taken on a hazy day?
  - Affects all the fields in the area

Left Fig.: Standard deviation on corrected data at the field level is consistently reduced as expected
Right Fig.: Average values do not show a consistently reduced standard deviation across years

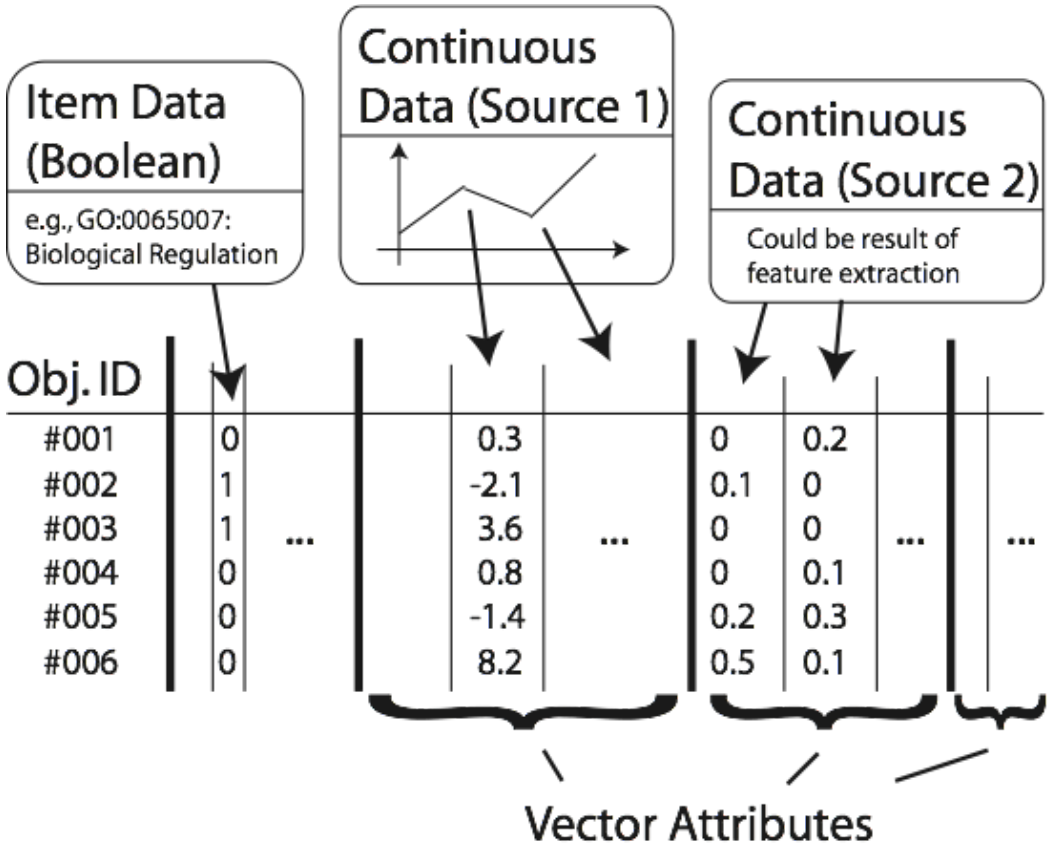


# Vector-item Pattern Mining

- Look for attributes that predict combination of attributes well
- Less likely to create problems with over-fitting
- Sometimes multiple attributes relevant
  - Weight, sugar content and sugar lost to molasses in sugarbeets
  - Weight and protein in wheat

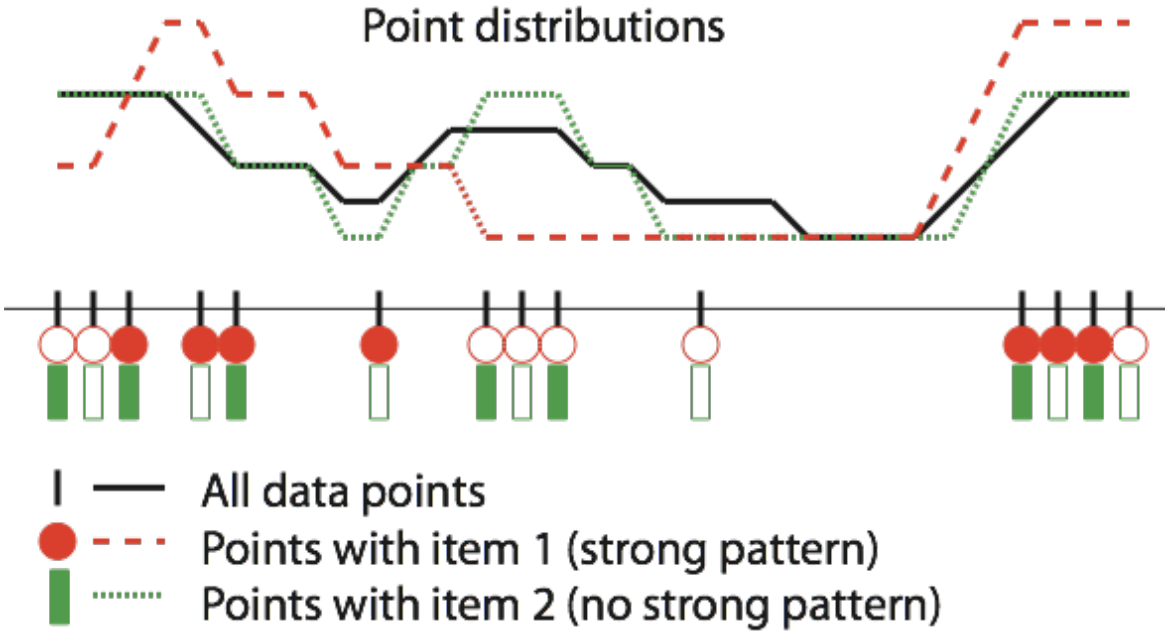
# Vector-item Pattern Mining

- Vector data
  - Groups of continuous data from multiple sources
  - Could be result of feature extraction
- Item data
  - Binary with presence less frequent then absence
  - Could be item sets



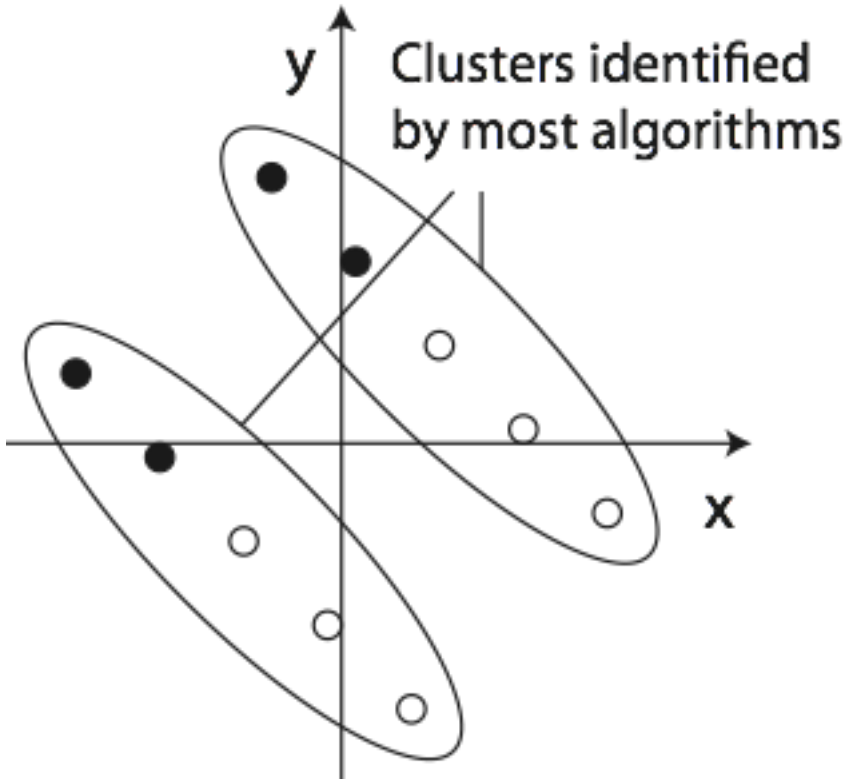
# Problem Statement

- Identify items for which distribution of points with item differs significantly from overall distribution
- Related work: Is classification significant?



# Common Approach in Bioinformatics

- Cluster, then look for enrichment of clusters
- Can miss significant relationships



# Approach

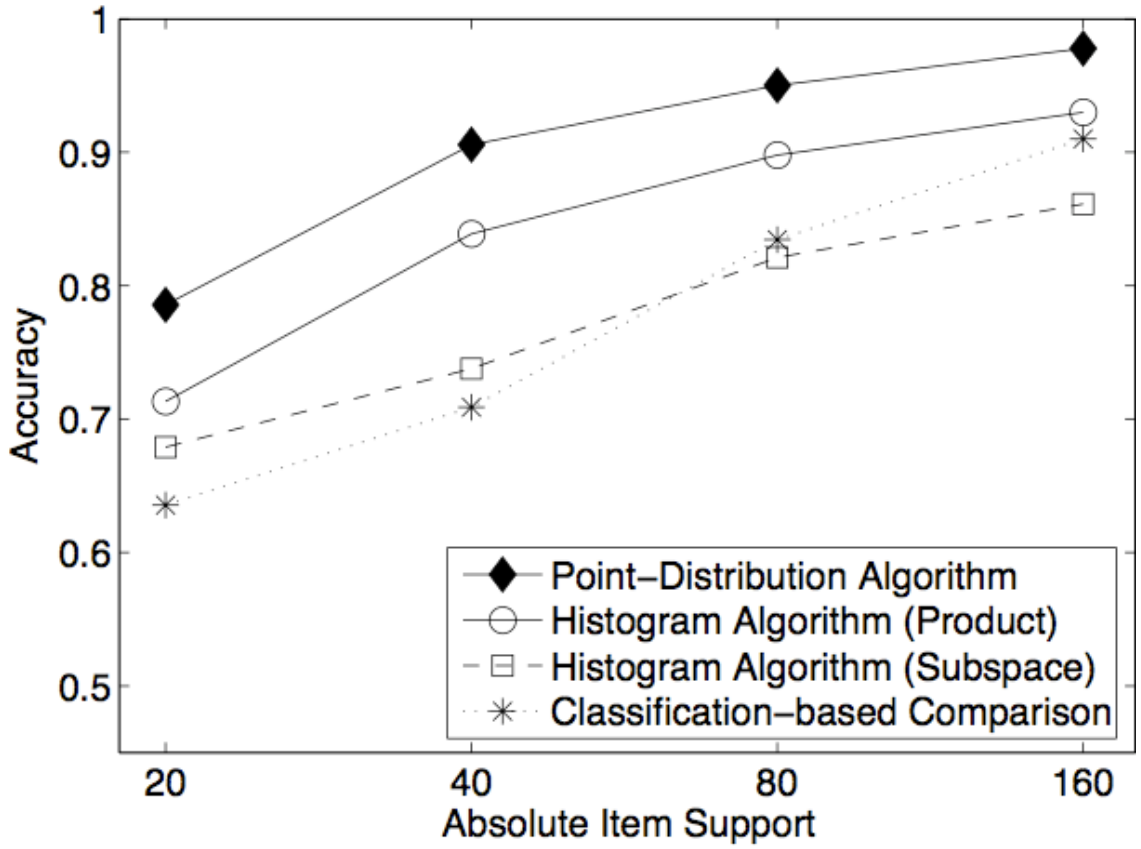
- Define density, using kernel function (uniform kernel)
- Compare densities of points with item to densities of all points
- Previous approach used histograms
- Kullback-Leibler divergence quantifies difference between distributions directly

$$D_{\mathrm{KL}}^{(d)}(P||Q) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \, \mathrm{d}x_1, \ldots, \mathrm{d}x_d$$

$$D_{\mathrm{KL}}^{(d)}(P||Q) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$

# Results on Time Series Data

- Clearly superior to comparison algorithms
- For very large item support classification may become competitive



# Big Data Problems

- Satellite data going from 30 m (Landsat) spatial resolution to 5 m (RapidEye)
- For meteorological data there may be a benefit in going to multiple time points per day
- As in any field
  - More is measured
  - More is stored
- But: Geospatial data starts out large already

# Dealing with the Infrastructure Problem

- At data mining conferences
  - Companies like Google use orders of magnitude larger data sets
  - Researchers often look at scaling of algorithms but that potentially misses processing issues
- Geospatial data
  - Not really a choice:  Data are large
  - To get high-quality results requires high resolution

# Options

- Desktop computers
  - Clearly too limited eventually
- Virtualized departmental infrastructure
  - Has any of the growth potential associated with "the cloud"
  - Difficulties growing it
- CCAST
  - Traditionally too inflexible
  - Negotiations are long and frustrating

# CCAST

- Awareness of importance of data-intensive applications has been rising
- Recent MRI (Major Research Instrumentation) grant was funded that emphasizes storage
- Still primarily driven by
  - Chemistry
  - Engineering
  - Pharmacy
- I've often been "excuse computer scientist" but
  - I use CCAST heaviest for collaborative parts of my bioinformatics work

# Problems with CCAST

- Showing off results
  - Currently no web servers at CCAST
- Installing software
  - It's a nuisance when any installation has to be done by somebody at CCAST
  - They may say "can't be done" without additional information
- Running jobs
  - Quota and limitations can be unclear

# Industry Solutions

- How do you get privileges on shared resources?
  - You create a sandbox
- How do you create a sandbox that's a whole system?
  - Virtualization
- How do you let users create their own sandbox?
  - Cloud

# CCAST and Cloud Computing

- At supercomputer conferences
  - "Is the cloud ready for supercomputing?"
  - Always >5% overhead
  - Answer may always stay "no"
- Not really the issue for CS
  - If a platform is too inflexible to use 5% don't matter!
- Why should cloud be at CCAST?
  - Farm out some jobs to conventional infrastructure
  - Access to storage
  - Large-scale funding may preclude "closet solutions"

# Case Study: iPlant

- NSF-funded project
  - http://www.iplantcollaborative.org/
- Plant-genomics community
  - Large genomes
  - Well-organized community
  - Visionary grant, but has had problems
- Embodies all three major cloud design concepts
  - Platform as a System (PaaS)
  - Software as a System (SaaS)
  - Infrastructure as a System (IaaS)

# Cloud Computing Services

- Platform as a Service, PaaS
  - Idea of Google App Engine
  - Doesn't require virtualization
  - Apps are developed to use large infrastructure
  - Security issues addressed by app development environment
- Software as a Service, SaaS
  - Application and data centrally hosted
  - Distinction to web interfaces? Sometimes vague
- Infrastructure as a Service, IaaS
  - Amazon cloud offerings and others
  - Virtual machine can be created by user (usually cloning)
  - Parallelism typically through (Elastic) MapReduce types of environments
  - Typically work with storage clouds

# Comparison with Departmental Infrastructure

- Virtualized system using VMware
  - Requires System Administrator to create VMs
  - VMs can be created by installing operating system or by cloning
  - Storage configured as RAID arrays in a SAN (not distributed cloud storage)
- Newer platform: OpenStack
  - Full cloud platform
  - Amazon Web Services compatible
  - User can create VMs
  - VMs normally created by cloning
  - Specifically configures systems created as “Boot from volume”
  - The project started as IBM Capstone
  - Storage:  Currently same SAN as for VMware

# Future of Cloud Efforts

- We clearly need to keep working on funding for infrastructure
  - Funding agencies will also want to see our understanding of this!
- It seems OpenStack is the way to go
  - Not absolutely clear that CCAST agrees with that
- Educational side
  - This sort of stuff is what fascinates our undergrads

# Student ACM

- Has in recent years put major effort in mobile development “SIG-MOBI”
- Several students have been involved in cloud
  - Logan Paschke
  - Davin Loegering
  - Matt Odden (now at IBM)
- Students know they need this in their jobs
- I think there is a great opportunity for working with student ACM
- There actually is no SIG-CLOUD even at ACM level!?