The 4th Unit Project

Required if you are taking class for 4 credits

Offered for extra credit (5%) if you are taking class for 3 credits and *cannot* take it for 4 credits
Project Idea: Reliable Real-time Information Distillation from the Physical World

Physical World
- Civil Unrest
- Hurricanes
- Man-made disasters

People

Sensors

Information
There exists a unique "ground truth" state (vector) is being estimated

As opposed to: opinion mining, sentiment analysis, statistical correlation mining, ...
Reconstructing Event Timelines

The Apollo Fact-finder

- Credibility of sources
- Correctness of claims
- Confidence intervals

Clean Event Summary

Maximum Likelihood Estimation

- Formulate the fact-finding problem as one of maximum likelihood estimation
- Solve it using the Expectation Maximization (EM) algorithm
- Compute a bound on estimation accuracy (using the Cramer Rao Bound)

Events

- Civil Unrest
- Hurricanes
- Man-made disasters

Sources

Clubs

Claims

Attribute: Credibility

Attribute: True/False
Social Channel “Decoding”
A Maximum Likelihood Estimation Problem

Joint estimation of
- Source reliability
- True/false value of each observation

Given
- Who said what

Events
- Civil Unrest
- Hurricanes
- Man-made disasters

Sources

Claims

Attribute: Reliability

Attribute: True/False

$$P(SC|\theta) = \sum_z P(SC, z|\theta)$$
Apollo: A Social Sensing System with a Twitter Front-end
Humans as (Noisy) Sensors

- Example of tweets collected in the aftermath of the Syrian chemical weapons attack in August 2013.
- Tweets were crawled for ten days after the event using the keywords “Syria”, “attack”, “dead”
- Table shows results of maximum likelihood estimation, automatically separating tweets into “socially corroborated” and “not corroborated”.

<table>
<thead>
<tr>
<th>Triage Result: Recommended for Viewing</th>
<th>Triage Result: Dismissed/Unimportant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medecins Sans Frontieres says it treated about 3,600 patients with ‘neurotoxic symptoms’ in Syria, of whom 555 died</td>
<td>So sad. All but one of the activists who filmed the chemical attack in Syria died of toxins:</td>
</tr>
<tr>
<td>Weapons expert says #Syria footage of alleged chemical attack “difficult to fake”</td>
<td>Saudis offer Russia secret oil deal if it drops Syria via @Telegraph <a href="http://t.co/oOutxSiaRs">http://t.co/oOutxSiaRs</a></td>
</tr>
<tr>
<td>U.N. experts in Syria to visit site of poison gas attack</td>
<td>Putin Orders Massive Strike Against Saudi Arabia If West Attacks Syria <a href="http://t.co/SFLJ39hwbh">http://t.co/SFLJ39hwbh</a></td>
</tr>
<tr>
<td>Syria Gas Attack: ‘My Eyes Were On Fire’</td>
<td>Miley Cyrus twerks meanwhile in other news the U.S.A. might declare war on Syria...</td>
</tr>
<tr>
<td>Long-term nerve damage feared after Syria chemical attack</td>
<td>I posted a new photo to Facebook <a href="http://t.co/FRWBFC0vKh">http://t.co/FRWBFC0vKh</a></td>
</tr>
<tr>
<td>Syrian official blames rebels for deadly attack</td>
<td>Two Minds on Syria <a href="http://t.co/ogDjKPH7Rs">http://t.co/ogDjKPH7Rs</a> via @NewYorker</td>
</tr>
<tr>
<td>Assad regime responsible for Syrian chemical attack, says UK government</td>
<td>We may be going to war in Syria, and somehow Miley Cyrus is trending on twitter</td>
</tr>
<tr>
<td>US forces move closer to Syria as options weighed: WASHINGTON (AP) — U.S. naval forces are moving closer to Sy...</td>
<td>Syrian Chemical Weapons Attack Carried Out by Rebels, Says UN (UPDATE) <a href="http://t.co/lN4CkUePlj">http://t.co/lN4CkUePlj</a> #Syria <a href="http://t.co/TorVEUuZF">http://t.co/TorVEUuZF</a></td>
</tr>
<tr>
<td>400 tonnes of arms sent into #Syria through Turkey to boost Syria rebels after CW attack in Damascus --&gt;</td>
<td>For those in the US, please text SYRIA to 864233 to donate $10 via @unicefusa <a href="http://t.co/YMXnru1jeb">http://t.co/YMXnru1jeb</a> #childrenofsyria</td>
</tr>
<tr>
<td>UN Syria team departs hotel as Assad denies attack</td>
<td>Attack! <a href="http://t.co/vY5KKm7R3s">http://t.co/vY5KKm7R3s</a></td>
</tr>
<tr>
<td>Vehicle of @UN #Syria #ChemicalWeapons team hit by sniper fire. Team replacing vehicle &amp; then returning to area.</td>
<td>A fathers last words to his dead daughters killed by Bashar al-Assad &amp; his supporter army with chemical weapon attack <a href="http://t.co/DN25pJwCq8">http://t.co/DN25pJwCq8</a></td>
</tr>
<tr>
<td>International weapons experts leave Syria. U.S. prepares attack. More @ <a href="http://t.co/pZ6Z2R9QOIE">http://t.co/pZ6Z2R9QOIE</a></td>
<td>What the media isn’t telling you about the Syrian chemical attack <a href="http://t.co/LQ479S1Tw">http://t.co/LQ479S1Tw</a></td>
</tr>
<tr>
<td>Military strike on Syria would cause retaliatory attack on Israel, Iran declares <a href="http://t.co/M950sEvGwW">http://t.co/M950sEvGwW</a></td>
<td>France on the phone. Apparently they surrendered to #Syria weeks ago.</td>
</tr>
<tr>
<td>Asia markets fall on Syria concerns: Asian stocks fall, extending a global market sell-off sparked by growing ... <a href="http://t.co/06a9h2x4nJ">http://t.co/06a9h2x4nJ</a></td>
<td>Poll: Do you think the chemical attack in #Syria could have been a false flag attack to push for war? RT for yes. Favourite for no</td>
</tr>
<tr>
<td>UK Prime Minister Cameron loses Syria war vote (from @AP) <a href="http://t.co/UIFF1wY9gx">http://t.co/UIFF1wY9gx</a></td>
<td>Lebanon was once part of Syria and will forever be with Syria. #PrayForSyria #PrayForLebanon</td>
</tr>
</tbody>
</table>

For more details, refer to the table and the accompanying text.
Extensions:

- The current estimation framework makes simplifying assumptions on sources and observations (e.g., independence)
  - How to detect copying/influence?
  - How to account for source non-independence due to information dissemination?
  - How to account for physical relations between observations?
  - How to include inference and other logical relations when some observations imply others?
  - How to separate “opinions” from ground-truthable facts?
  - How to de-bias observations?
  - How to detect degree of “polarization” among sources?
  - How to compute fundamental error bounds?
  - How to influence sources such as error bound is reduced?
The Social Signal: An Analogy

Physical target

Response of physical propagation medium (e.g., acoustic, vibration, optical, ...)

Received signature (energy in multiple signal frequency bands)
An Analogy

Physical target
Response of physical propagation medium (e.g., acoustic, vibration, optical, ...)

Received signature (energy in multiple signal frequency bands)

Physical event
Response of social propagation medium (e.g., tweets)

Received signature (energy in multiple keyword frequency bands)
Demultiplexing

- A world of “protest” – this morning:
  - Angry French farmers and 1,000 tractors head for Paris protest. Photo @MartinBureau1 #AFP [http://t.co/j5DdveSHZh](http://t.co/j5DdveSHZh)
  - VIDEO: Tractor protest descends on Paris: French farmers protesting about high taxes have taken a convoy of tr... [http://t.co/hKievMFpq3](http://t.co/hKievMFpq3)
  - WATCH LIVE: Farmers on tractors gather in Paris streets [https://t.co/peTOvKrIAF http://t.co/3vDK6qc060](https://t.co/peTOvKrIAF)
  - MORE: Police detained refugees who lay on train tracks in protest at being taken to a camp, This is 2015 not 1940's [http://t.co/TbQrwWBWrH](http://t.co/TbQrwWBWrH)
  - RIGHT NOW: Activists & giant polar bear protest Arctic oil outside Shell London HQ [http://t.co/1Ae9mgc1ZF #ArcticRoar](http://t.co/1Ae9mgc1ZF)
  - Underwater sculptures emerge from Thames in climate change protest [http://t.co/mg6RiURn6t](http://t.co/mg6RiURn6t)
Events and Signal Processing: The Lexical Frequency Domain

- Observation: Targets can be recognized using frequency domain signatures
- Question: Can we detect and track events using “frequency domain” signatures only?
  - At first glance: text has complex semantics, so the ordering of keywords has great impact on meaning
    - “John killed Mary” versus “Mary killed John”
  - Do we need natural language processing to identify and track distinct events?
Events and Signals: A Data Association Problem

Easy to associate data with events

Densely populated feature space

Hard to associate data with events

Sparsely populated feature space
Most languages have about 10,000 frequent words.

Consider a 2-word event signature
- There are at least 100,000,000 possible signatures

Number of “events” in a Twitter data trace may be in the 100s or 1000s

The space of keyword signatures is vastly sparse:
- Different events → Different signatures (assuming independent keywords)
Event Detection, Consolidation, and Tracking: Signal Processing Questions

- How to detect new event signatures?
  - Find high-information-gain signatures (new spikes in the frequency spectrum)
  - Bin tweets that contain a new signature into a cluster
  - Determine if this cluster is of a new event or not using frequency domain distance (note: some events will have more than one signature)
Event Detection, Consolidation and Tracking

Three key ideas:
1. Use information gain to detect new keyword pairs (event signatures)
2. Each pair gives rise to a cluster of tweets (that contain the pair)
3. Merge clusters with similar keyword distributions

Automatically detected high-information-gain keyword pairs
Clusters of tweets containing keyword pairs

Time
Distance Metrics (For Merging Event Data Clusters)

- Cosine similarity
- Term frequency difference
- Jaccard distance
- KL divergence

Clusters of tweets containing keyword pairs

Event track

Tweet cluster

(keyword pair)

(keyword pair)

(keyword pair)

Lexical frequency domain signal

Distance?
Event Tracks
Recognizing Distinct Event Tracks

- **Project contribution:** Efficient algorithms that “demultiplex” Twitter feed into sub-streams associated with different events in a class (e.g., different concurrent flashmobs or different concurrent protests)

<table>
<thead>
<tr>
<th>Protest Name</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh protests</td>
<td>Religion Bangladesh braces for protests after Islamists’ execution: senior official of the largest Islamis… [link]</td>
</tr>
<tr>
<td></td>
<td>World News: Bangladesh braced for protests after Islamist leader’s execution: Bangladesh security personnel s… [link]</td>
</tr>
<tr>
<td></td>
<td>Bangladesh braces for protests after Jamaat leaders execution: Bangladesh braced for protests and fresh violen… [link]</td>
</tr>
<tr>
<td>Brazil protests</td>
<td>Protests across Brazil seek ouster of president [link]</td>
</tr>
<tr>
<td></td>
<td>FollowMePlease Brazil braces for nationwide protests, as groups seeking impeachment of president… [link]</td>
</tr>
<tr>
<td></td>
<td>Fresh anti-government protests in Brazil: Brazil on Sunday braced for more huge demonstrations against government… [link]</td>
</tr>
<tr>
<td>Turkey protests</td>
<td>DTN Turkey: Turkey protests to Pope Francis after he brands Armenian killings ‘genocide’: Pontif’s run-in wit… [link]</td>
</tr>
<tr>
<td></td>
<td>Pope refers to Armenian genocide; Turkey protests. 24 April is 100th anniversary of start of the Armenian genocide [link]</td>
</tr>
</tbody>
</table>
|                        | Telegraph: Turkey protests to Pope Francis after he brands Armenian killings ‘genocide’ [link] [link] [link]
The Social Signal Layer

Keywords
Frequency counts
Event Detection and Tracking
Social Sensing Signal, $Signal (k)$
Event Map
Events and Trajectories
Event Data (Tweets Associated with the Event)
Event Medium
Observers of Physical Events

Event: Marathon, Protest, Flash mob
Event Localization with Instagram

- Taking a picture requires being on location
- There is a substantial overlap between Twitter users and Instagram users
  - Implication: Many shared hashtags/labels
- “Demultiplex” events on Twitter, identify relevant keywords/hashtags, search Instagram, find location!
Instagram Localization

- Tracking “LA Marathon”
Instagram

- Tracking “LA Marathon”
Instagram Tracking

- Tracking “LA Marathon”: Early Stage
Instagram Tracking

- Tracking “LA Marathon”: Middle
Instagram Tracking

- Tracking “LA Marathon”: Late Stage
Challenge: Extractive Summarization

Build a data service that allows applications to retrieve (extractive) data summaries at arbitrary levels of granularity in accordance with an application-specific redundancy metric.
Customizability: The Distance Metric

Data Object → Application Callback → Feature Vector

Data Object → Application Callback → Feature Vector

Difference Function → Distance Metric

(Must obey triangle inequality)
Customizability: The Distance Metric

Opaque type (not interpreted by service)

Distance Metric
(Must obey triangle inequality)

Application specific functions
(customization API)
Customizability: The Distance Metric

**Examples**: Scalars, vectors, pictures, text, etc.

Distance Metric (Must obey triangle inequality)

Data Object

- Application Callback

Data Object

- Application Callback

Feature Vector

Difference Function

---

Customizability: The Distance Metric

**Examples**: Scalars, vectors, pictures, text, etc.

Distance Metric (Must obey triangle inequality)

Data Object

- Application Callback

Data Object

- Application Callback

Feature Vector

Difference Function

---
Customizability: The Distance Metric

Data Object

Application Callback

Data Object

Application Callback

Feature Vector

Difference Function

Distance Metric
(Must obey triangle inequality)
Customizability: The Distance Metric

Data Object → Application Callback → Feature Vector → Difference Function → Distance Metric

(Distance Metric must obey triangle inequality)

Hierarchical Clustering
Summarization

Data Object

Application Callback

Feature Vector

Difference Function

Distance Metric
(Must obey triangle inequality)

Hierarchical Clustering

Data Object

Application Callback

Feature Vector
Summarization

Hierarchical Clustering

Data Object

Application Callback

Feature Vector

Distance Metric
(Must obey triangle inequality)

Difference Function

Summarization
Hierarchical Clustering

Data Object

Application Callback

Feature Vector

Difference Function

Distance Metric
(Must obey triangle inequality)

Summarization
Summarization

Data Object

Application Callback

Feature Vector

Difference Function

Distance Metric

(Must obey triangle inequality)

Hierarchical Clustering

Representative sampling versus noise reduction?

Data Object

Application Callback

Feature Vector
The data fire-hose effect

A Network Paradigm Shift
Communication → Information Distillation

Present Networks

**Goal:** Communication
- Maximizes bit throughput between end-points
- Most data is “logical”
- Protocols geared primarily for point-to-point communication
- Data loss may be a problem

Future Distillation Networks

**Goal:** Information Distillation
- Maximizes *information flow*
- Much data is “physical”
- Protocols geared for data filtering, and aggregation
- Data loss may be a feature intended to reduce less informative bits
A Primary Network Design Challenge

How to build networks that maximize useful information flow from the physical world?
Information-maximizing Prioritization

- Determine transmission order?
Information-maximizing Prioritization

- Determine transmission order?
Information-maximizing Prioritization

- Determine transmission order?
Information-maximizing Prioritization

- Determine transmission order?
Information-maximizing Prioritization

- Determine transmission order?
Information-maximizing Prioritization

- Determine transmission order?
Information-maximizing Prioritization

- Determine transmission order?

Coverage-monotonic scheduling
Information-maximizing Prioritization

Note: Coverage can be defined in an abstract feature space

Coverage-monotonic scheduling
Example: Data Forwarding in Disruption-tolerant Networks

A big disaster strikes a city...

- Volunteers are recruited
- They scout the area, capture pictures and send them to a rescue center
- Network constraints prevent sending all pictures

Images are collected from the Internet

Hurricane Katrina 2005
Nepal earthquake 2015
Thailand flood 2011
Challenge: Data Selection to Maximize Coverage

- Fire on 6th and Main.
- Collapse on Park Ave.
- Flooding on State St.
- Structural damage on Pier Square
Example of Bad Coverage

Fire on 6th and Main.

Collapse on Park Ave.

An Example of Poor Data Selection (Low Coverage)
Example of Good Coverage

Fire on 6th and Main.  
Collapse on Park Ave.  
An Example of Good Data Selection (High Coverage)

Flooding on State St.  
Structural damage on Pier Square
A Scheduling Approach: Coverage-maximizing Priorities

- Implement coverage-maximizing in-network prioritization for forwarding and storage
  - Objects are forwarded/dropped in a priority order aimed to maximize coverage of delivered content
    - Objects similar to previously forwarded ones get lower priority
  - Challenge: Forwarding and dropping must be made aware of the degree of semantic redundancy (i.e., similarity) between objects
US is 5% of world’s population but 21% of GHG emissions.
The transportation sector is one of the largest sources of GHG emissions in the US.
How to reduce energy & emissions?
Trends

- Carbon emissions by sector
- Transportation emissions by vehicle type
Trends

- Modes of transportation (to work)
Class Project Idea

- Improve fuel-efficiency of transportation via “green” navigation
  - Measure fuel-efficiency of vehicles
  - Model fuel-consumption as a function of driver characteristics, road characteristics (average speed, speed variability, waiting time, slope, etc), and vehicle characteristics
  - Compute least-energy routes for a given vehicle and driver
Green GPS
Saves 6% over shortest path and 13% over fastest path

Subscribers
OBDII-WiFi Adaptor ($50) + GPS Phone

Subscribers:
Premium service
High savings

Open access:
Standard service
Average savings

Fuel Data + Physical Models

Server

Shortest and fastest

Most fuel-efficient

Fuel Error (%) vs. Trip Length

Green GPS

Physical Models:

\[ F_{\text{engine}} = \frac{\Gamma(\omega)GG_k}{r} \]

\[ F_{\text{air}} = \frac{1}{2} \rho A v^2 \]

\[ F_{\text{friction}} = c_{\text{rr}} m g \cos(\theta) \]

\[ F_g = m g \sin(\theta) \]

\[ F_{\text{car}} = F_{\text{engine}} - F_{\text{friction}} - F_{\text{air}} - F_g \]
Faster? Shorter? Try Cheaper, Greener

Program Gives Drivers the Most Fuel-Efficient Route

Tracy Cozzens

Most GPS devices in cars today give the driver two choices: shortest route or fastest route. GreenGPS provides a third option: most fuel-efficient route.

With gas prices skyrocketing, many drivers would be happy to spend a few more minutes on the road, or take the engine’s fuel efficiency and customizes navigation advice to the particular vehicle, Abdelzaher explained.

The best route computed by GreenGPS may not be the shortest or fastest, but it will use less fuel, he said. For example, if one route had a steep hill that would burn fuel, GreenGPS would avoid it.

Most GPS devices in cars today give the driver two choices: shortest route or fastest route. GreenGPS provides a third option: most fuel-efficient route.

With gas prices skyrocketing, many drivers would be happy to spend a few more minutes on the road, or take the engine’s fuel efficiency into account. GreenGPS customizes navigation advice to the particular vehicle, said Tarek Abdelzaher.

The best route computed by GreenGPS may not be the shortest or fastest, but it will use less fuel, he said. For example, if one route had a steep hill that would burn fuel, GreenGPS would avoid it.
A Modeling Challenge

Fuel consumption of a few cars driven on a few roads by a few driver

Predict fuel consumption of any car on any road by any driver
## Fuel Savings Evaluation

- **How efficient is the fuel-efficient route?**

<table>
<thead>
<tr>
<th>Car Details</th>
<th>Landmarks</th>
<th>Route</th>
<th>Savings %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honda Accord 2001</td>
<td>H1 to Mall</td>
<td>Shortest</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td>H1 to Gym</td>
<td>Shortest</td>
<td>19.7</td>
</tr>
<tr>
<td>Ford Taurus 2001</td>
<td>H2 to Restaurant</td>
<td>Shortest</td>
<td>26</td>
</tr>
<tr>
<td>Toyota Celica 2001</td>
<td>H2 to Work</td>
<td>Fastest</td>
<td>10.1</td>
</tr>
<tr>
<td>Nissan Sentra 2009</td>
<td>H3 to CUPHD</td>
<td>Fastest</td>
<td>8.4</td>
</tr>
<tr>
<td>Honda Civic 2002</td>
<td>Grad to Work</td>
<td>Fastest</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Average fuel savings across 5 cars
Buildings and Smart Spaces

- On average, Americans spend about 90 percent or more of their time indoors.
- Buildings accounted for 38.9% of total U.S. energy consumption in 2005.
- Buildings accounted for 72% of total U.S. electricity consumption in 2006.
- The average household spends at least $2,000 a year on energy.
- Out of the total energy consumption in an average household, 50% goes to space heating, 27% to run appliances, 19% to heat water and 4% goes to air conditioning.
Related Class Projects

- Build smart services that improve residential energy consumption
  - Energy consumption modeling
  - Smart lighting
  - Smart door/window control
## Testbed

<table>
<thead>
<tr>
<th>Knob</th>
<th>Sensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Frequency, Utilization,</td>
</tr>
<tr>
<td></td>
<td>Frequency, Temperature</td>
</tr>
<tr>
<td>MEM</td>
<td>Utilization,</td>
</tr>
<tr>
<td>NIC</td>
<td>Received / Sent packets/bytes</td>
</tr>
<tr>
<td>PDU</td>
<td>Power consumption of each</td>
</tr>
<tr>
<td></td>
<td>individual machine</td>
</tr>
<tr>
<td>CRAC</td>
<td>set point*</td>
</tr>
<tr>
<td></td>
<td>Input and outlet temperature,</td>
</tr>
<tr>
<td></td>
<td>Set point</td>
</tr>
</tbody>
</table>
Failures in Complex Systems

When systems fail, a common goal is: *Localize and fix the root cause!*
Failures in Complex Systems

Another Thought

Individual software components are easy to “debug”
- Therefore, they are typically built reliably

Systems do not fail because of “bugs” localized to single components
- Systems fail because of unexpected interactions between many *individually well-behaved* components
- No single component is to blame
- No predicate over current state explains failure
- Unexpected *sequences of events* lead to problems