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

The Real-time Information Distillation Problem

Physical World

Civil Unrest

Hurricanes

Man-made disasters

People
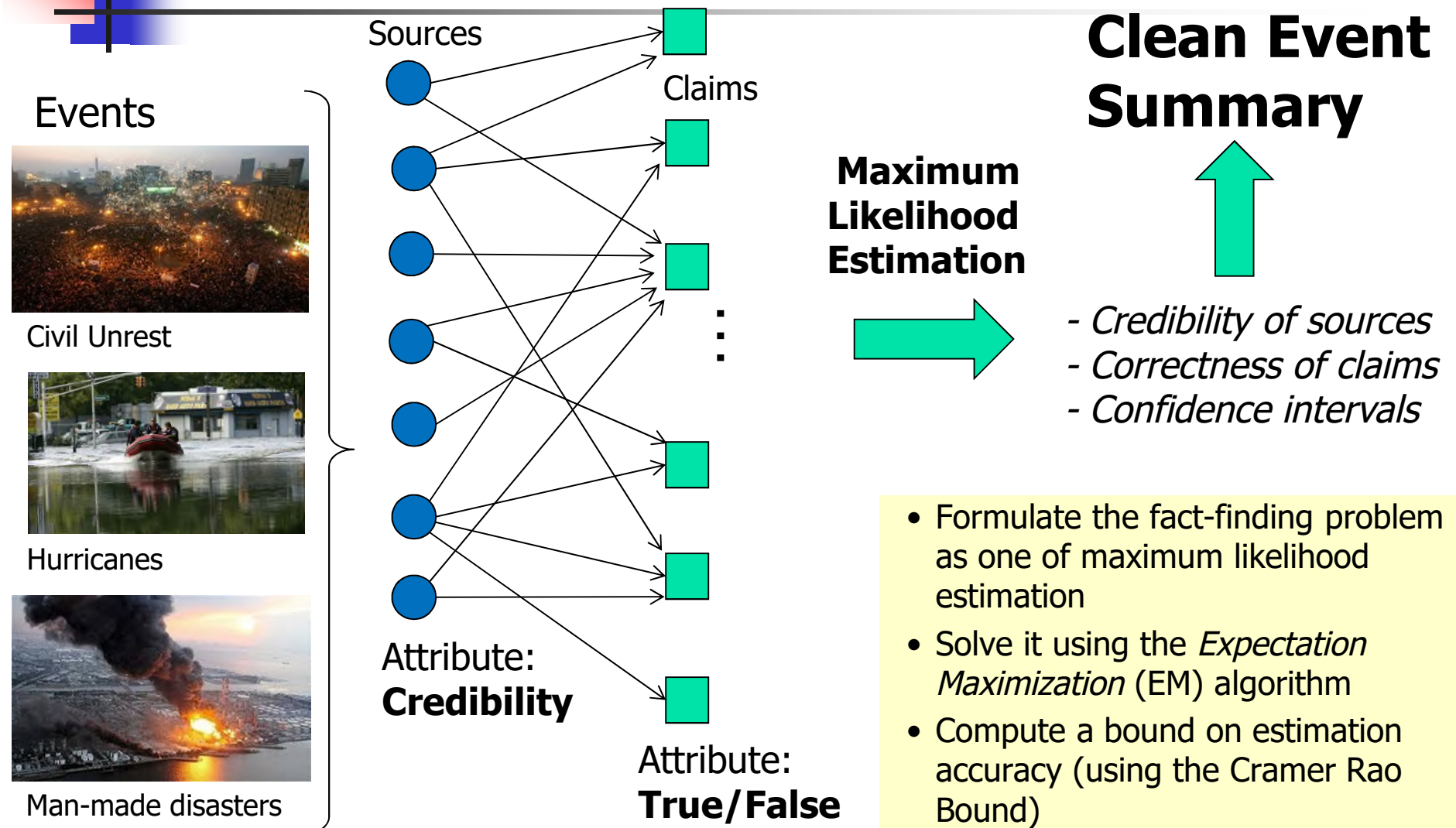
Sensors

Data Mining/Machine Learning/ Estimation

Estimated State

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

Events

Civil Unrest

Hurricanes

Man-made disasters

Sources

Claims

Attribute: **Credibility**

Attribute: **True/False**

Maximum Likelihood Estimation

Clean Event Summary

- Credibility of sources
- Correctness of claims
- Confidence intervals

- 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)

Social Channel "Decoding"
A Maximum Likelihood Estimation Problem

Events

Civil Unrest

Hurricanes

Man-made disasters

Sources

Claims

Attribute: **Reliability**

Attribute: **True/False**
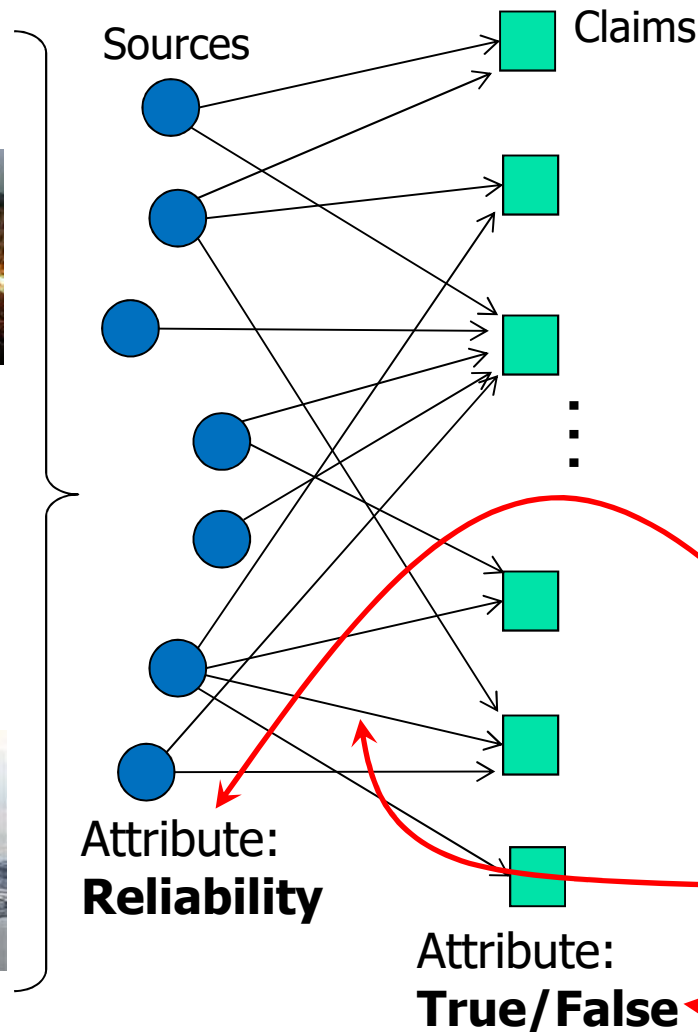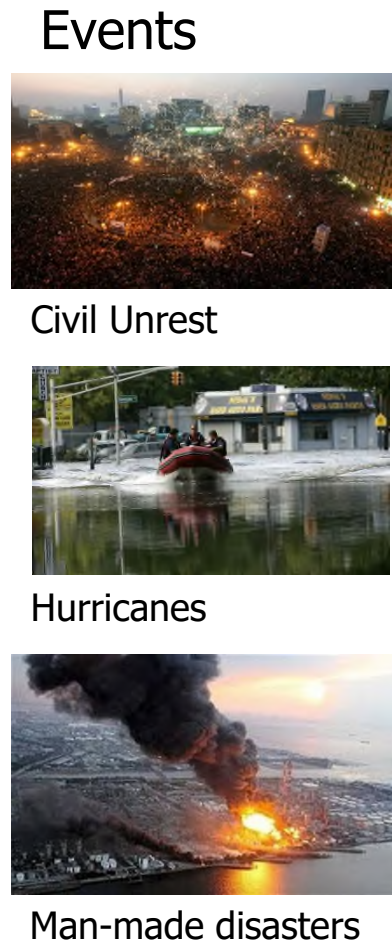
- Joint estimation of
  - Source reliability
  - True/false value of each observation
- Given
  - Who said what

$$P(SC|\theta) = \sum_{z} P(SC, z|\theta)$$

Apollo: A Social Sensing System with a Twitter Front-end
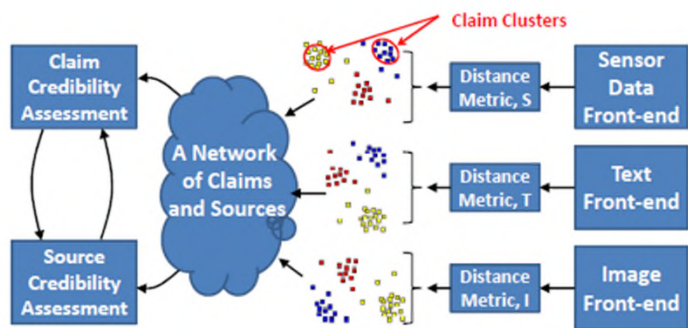
Humans as (Noisy) Sensors

- 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".

Triage Result: Recommended for Viewing	Triage Result: Dismissed/Unimportant
Medecins Sans Frontieres says it treated about 3,600 patients with 'neurotoxic symptoms' in Syria, of whom 355 died http://t.co/eHWY77jdS0	So sad. All but one of the activists who filmed the chemical attack in Syria died of toxins: http://t.co/7Xc9u8achL
Weapons expert says #Syria footage of alleged chemical attack "difficult to fake" http://t.co/zfDMujaCTV	Saudis offer Russia secret oil deal if it drops Syria via @Telegraph http://t.co/iOutxSiaRs
U.N. experts in Syria to visit site of poison gas attack http://t.co/jol8OlFxnf via @reuters #PJNET	Putin Orders Massive Strike Against Saudi Arabia If West Attacks Syria http://t.co/SFLJ9ghwbt
Syria Gas Attack: 'My Eyes Were On Fire' http://t.co/z76MiHj0Em	Miley Cyrus twerks meanwhile in other news the U.S.A. might declare war on Syria....
Long-term nerve damage feared after Syria chemical attack http://t.co/8vw7BiOxQR	I posted a new photo to Facebook http://t.co/FRWBFC0vKb
Syrian official blames rebels for deadly attack http://t.co/76ncmy4eqb	Two Minds on Syria http://t.co/ogDjKFH7Rs via @NewYorker
Assad regime responsible for Syrian chemical attack, says UK government http://t.co/pMZ5z7CsNZ	We may be going to war in Syria, and somehow Miley Cyrus Is trending on twitter
US forces move closer to Syria as options weighed: WASHINGTON (AP) — U.S. naval forces are moving closer to Sy... http://t.co/F6UAAXLa2M	Syrian Chemical Weapons Attack Carried Out by Rebels, Says UN (UPDATE) http://t.co/lN4CkUePUj #Syria http://t.co/tTorVFUfZF
400 tonnes of arms sent into #Syria through Turkey to boost Syria rebels after CW attack in Damascus --> http://t.co/KLwESYChCc	For those in the US, please text SYRIA to 864233 to donate $10 via @unicefusa http://t.co/YMXnrk1jcb #childrenofsyria
UN Syria team departs hotel as Assad denies attack http://t.co/O3SqPoiq0x	Attack! http://t.co/wY5KKm7R3s
Vehicle of @UN #Syria #ChemicalWeapons team hit by sniper fire. Team replacing vehicle & then returning to area.	A fathers last words to his dead daughters killed by Bashar al-Assad & his supporter army with chemical weapon attack http://t.co/DN25pLfCq8
International weapons experts leave Syria, U.S. prepares attack. More @ http://t.co/4Z62RhQKOE	What the media isn't telling you about the Syrian chemical attack http://t.co/LQ479S1Tiv
Military strike on Syria would cause retaliatory attack on Israel, Iran declares http://t.co/M950o5VcgW	France on the phone. Apparently they surrendered to #Syria weeks ago.
Asia markets fall on Syria concerns: Asian stocks fall, extending a global market sell-off sparked by growing ... http://t.co/06A9h2xCnJ	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
UK Prime Minister Cameron loses Syria war vote (from @AP) http://t.co/UlFF1wY9gx	Lebanon was once part of Syria and will forever be with Syria. #PrayForSyria #PrayForLebanon

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
  - VIDEO: Tractor protest descends on Paris: French farmers protesting about high taxes have taken a convoy of tr... http://t.co/hKievMFpq3
  - WATCH LIVE: Farmers on tractors gather in Paris streets https://t.co/peTOvKrIAF http://t.co/3vDK6qc060
  - 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
  - RIGHT NOW: Activists & giant polar bear protest Arctic oil outside Shell London HQ http://t.co/1Ae9mgc1ZF #ArcticRoar http://t.co/5tJaKv0mHZ
  - Underwater sculptures emerge from Thames in climate change protest 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"
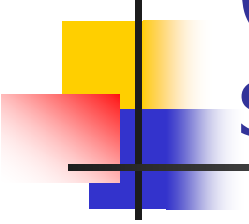 - Do we need natural language processing to identify and track distinct events?

Events and Signals: A Data Association Problem

Sparsely populated feature space

Easy to associate data with events

Feature axis

Densely populated feature space

Hard to associate data with events
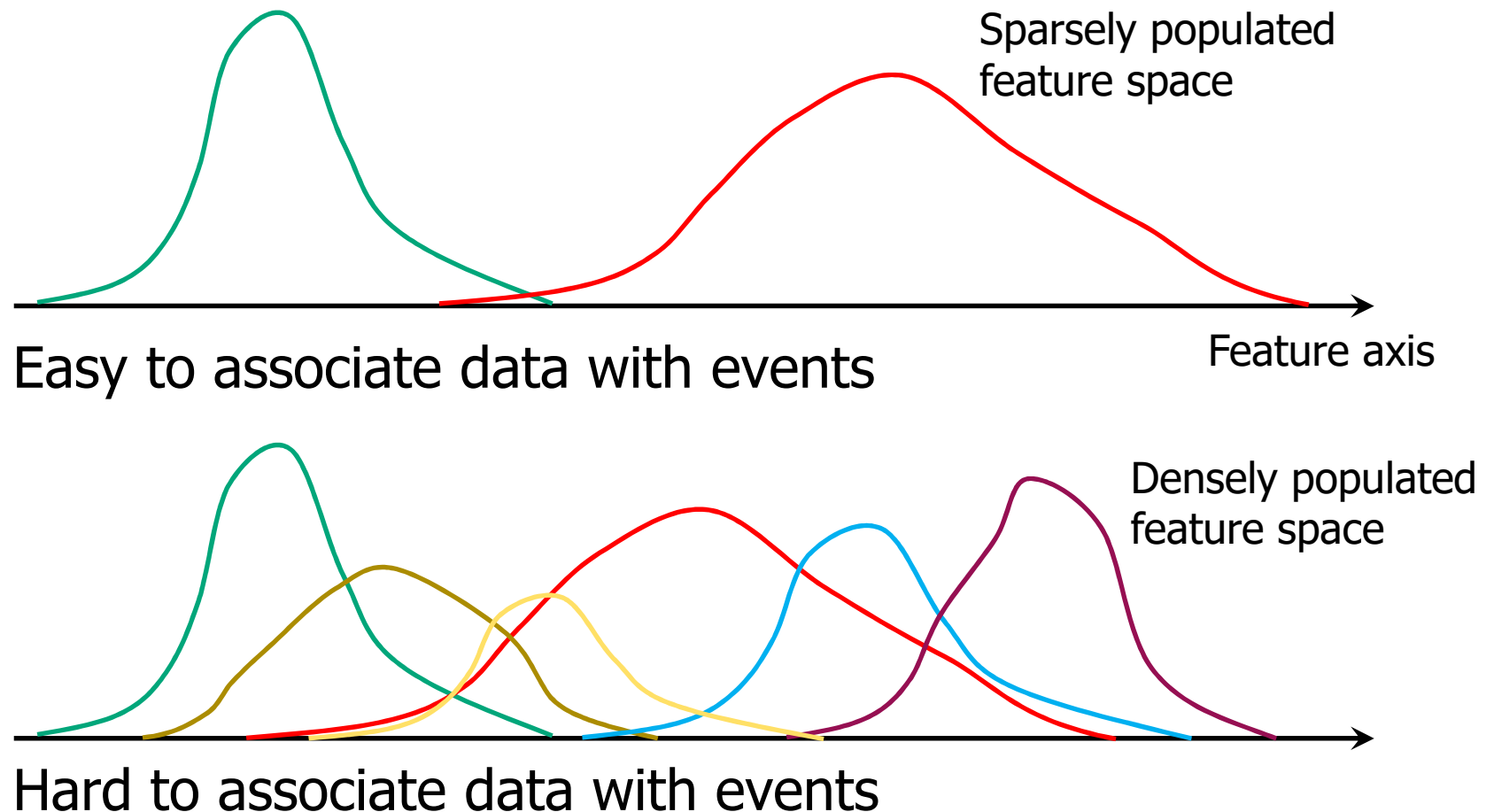
A Signal Sparsity Observation

- 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

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

Event track

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Event track

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

Time

Distance Metrics
(For Merging Event Data Clusters)
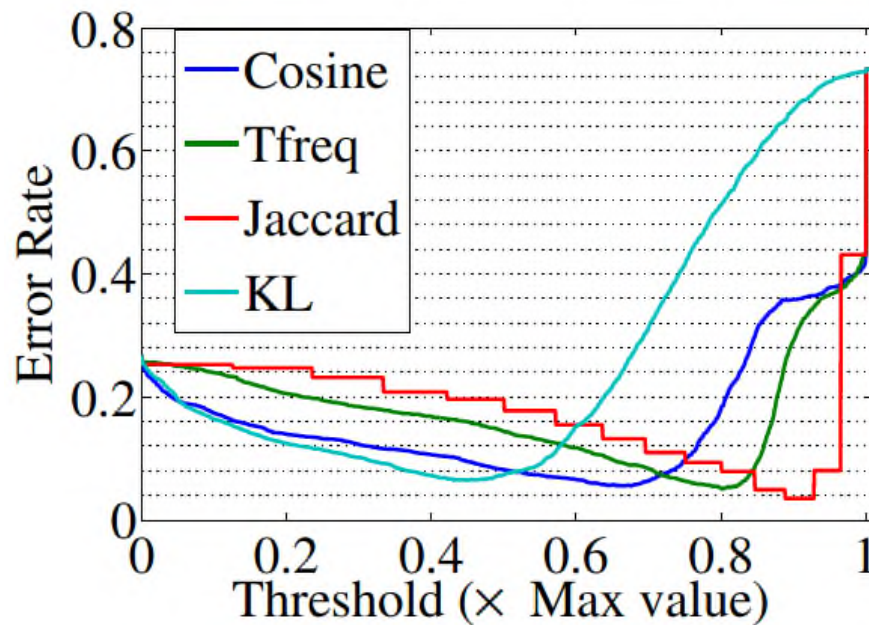
- Cosine similarity
- Term frequency difference
- Jaccard distance
- KL divergence

Lexical frequency domain signal

Distance?

Lexical frequency domain signal

(keyword pair)

(keyword pair)

(keyword pair)

(keyword pair)

Event track

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Tweet cluster

Clusters of tweets containing keyword pairs
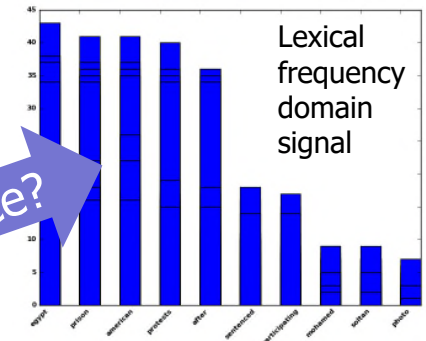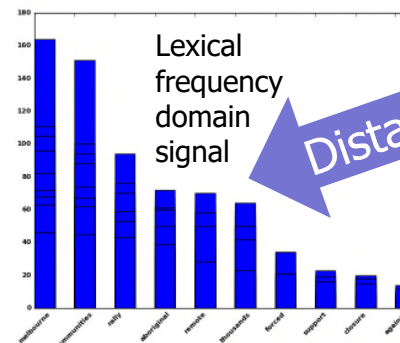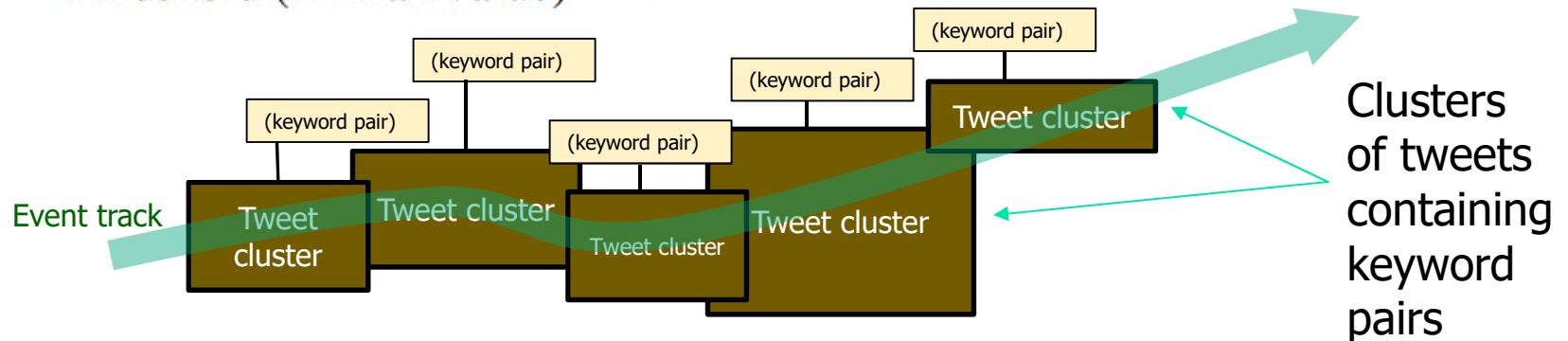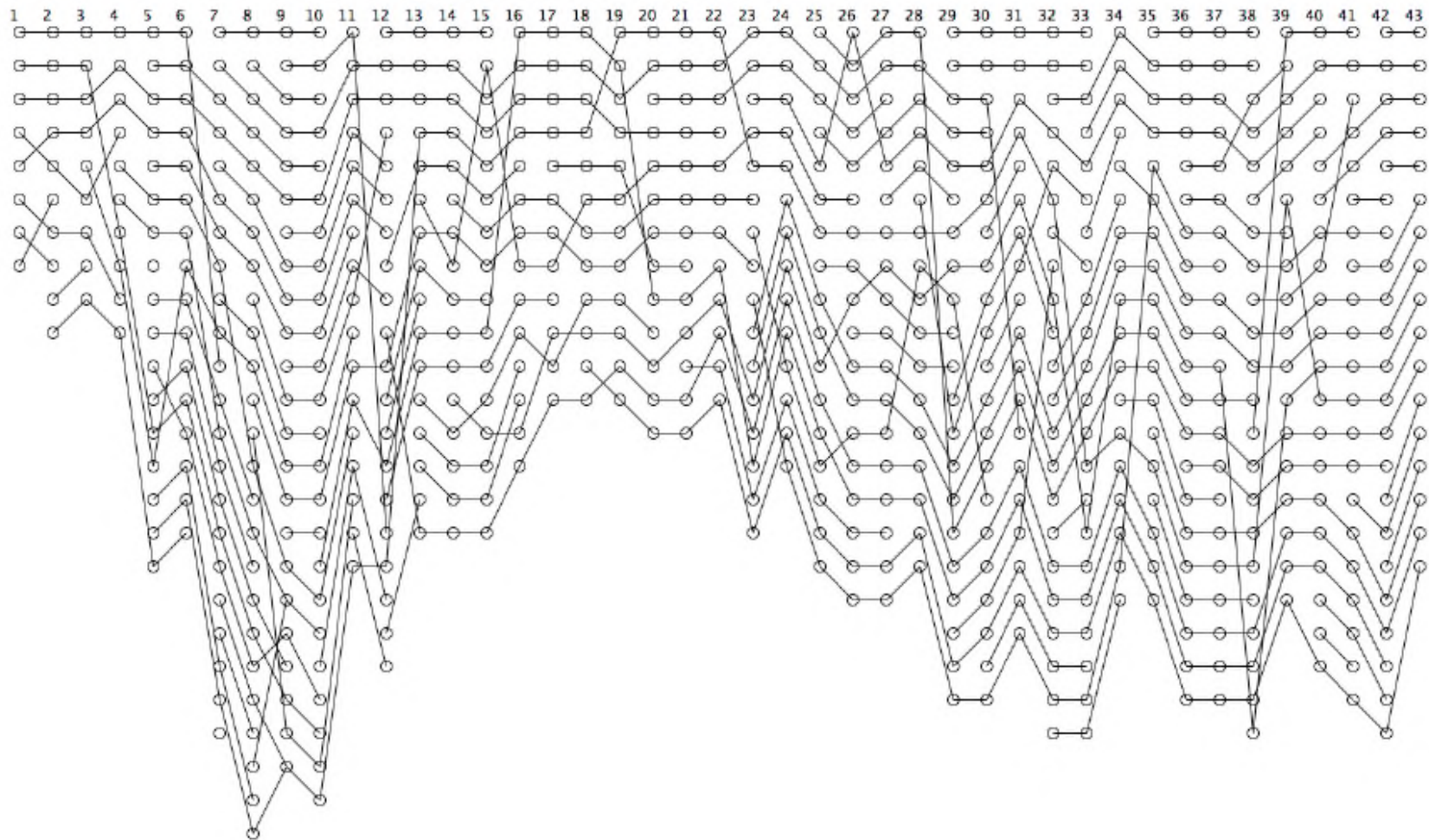
Event Tracks

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)

Protest Name	Tweets
Bangladesh protests	Religion Bangladesh Braces for Protests After Islamist's Execution: senior official of the largest Islamis... http://t.co/NTBmIWsTme
	World News: Bangladesh braced for protests after Islamist leader's execution: Bangladeshi security personnel s... http://t.co/UPKiaFFHtW
	Bangladesh braces for protests after Jamaat leaders execution: Bangladesh braced for protests and fresh violen... http://t.co/3dcPFqKAQE
Brazil protests	Protests across Brazil seek ouster of president http://t.co/YmXZnsxbAQ
	FollowMePlease Brazil braces for nationwide protests, as groups seeking impeachment of presiden... http://t.co/5T150D0zlL BrinaldyHere
	Fresh anti-government protests in Brazil: Brazil on Sunday braced for more huge demonstrations against governm... http://t.co/8nJmX56MUm
Turkey protests	DTN Turkey: Turkey protests to Pope Francis after he brands Armenian killings 'genocide': Pontiff's run-in wit... http://t.co/dKx5iCwP9w
	Pope refers to Armenian genocide; Turkey protests. 24 April is 100th anniversary of start of the Armenian genocide http://t.co/ZFOCZnC411
	Telegraph: Turkey protests to Pope Francis after he brands Armenian killings 'genocide' http://t.co/YnJew4foN1 http://t.co/O6d5oYwG2p

The Social Signal Layer

Frequency counts

Event Detection and Tracking

Event Data
(Tweets Associated with the Event)

Keywords

Social Sensing Signal, *Signal (k)*

Event
Marathon

Event
Protest

Event
Flash mob

Event Map
Events and Trajectories

Social Medium
Observers of
Physical Events
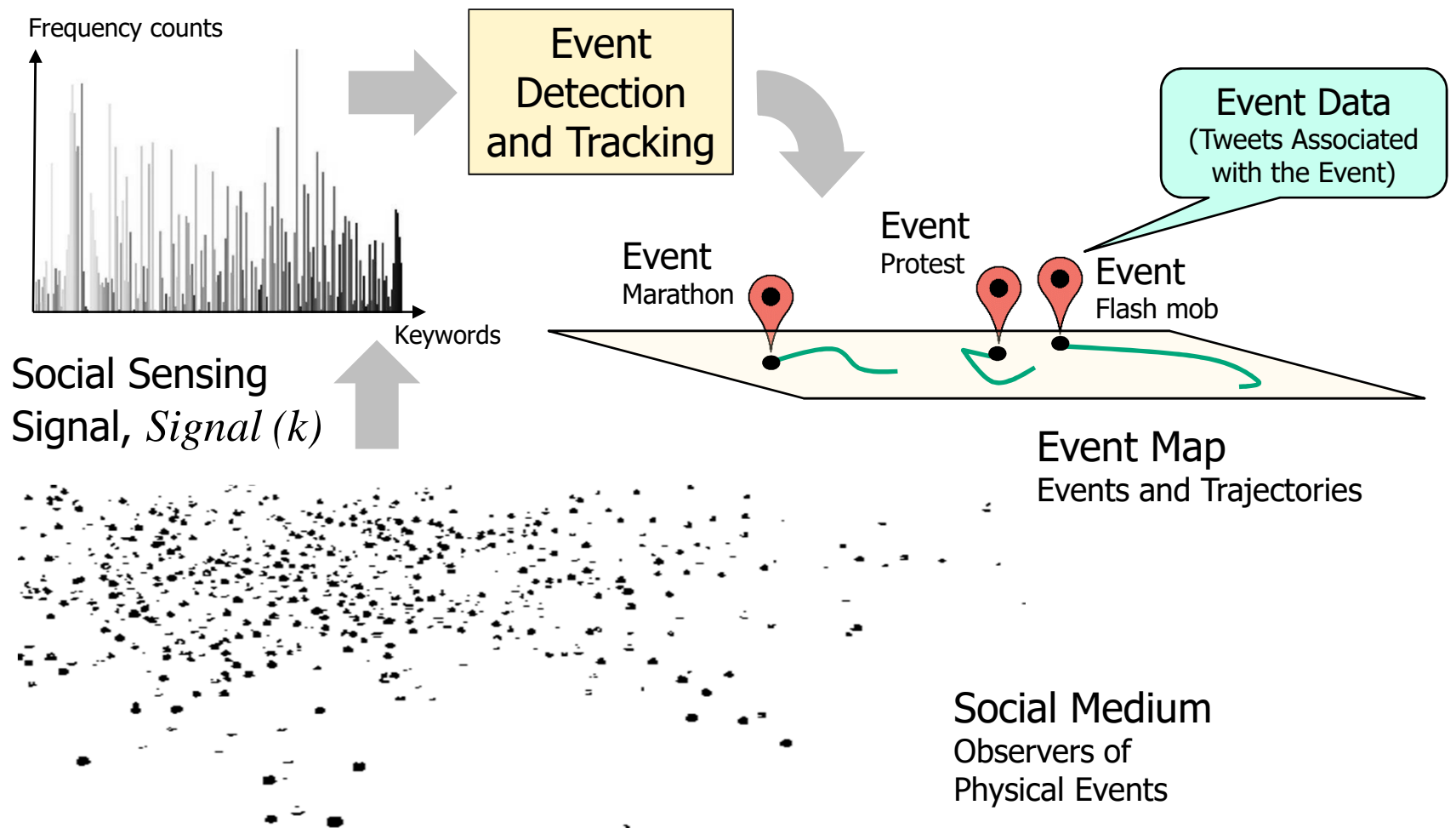
Event Localization with Instagram

- 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"
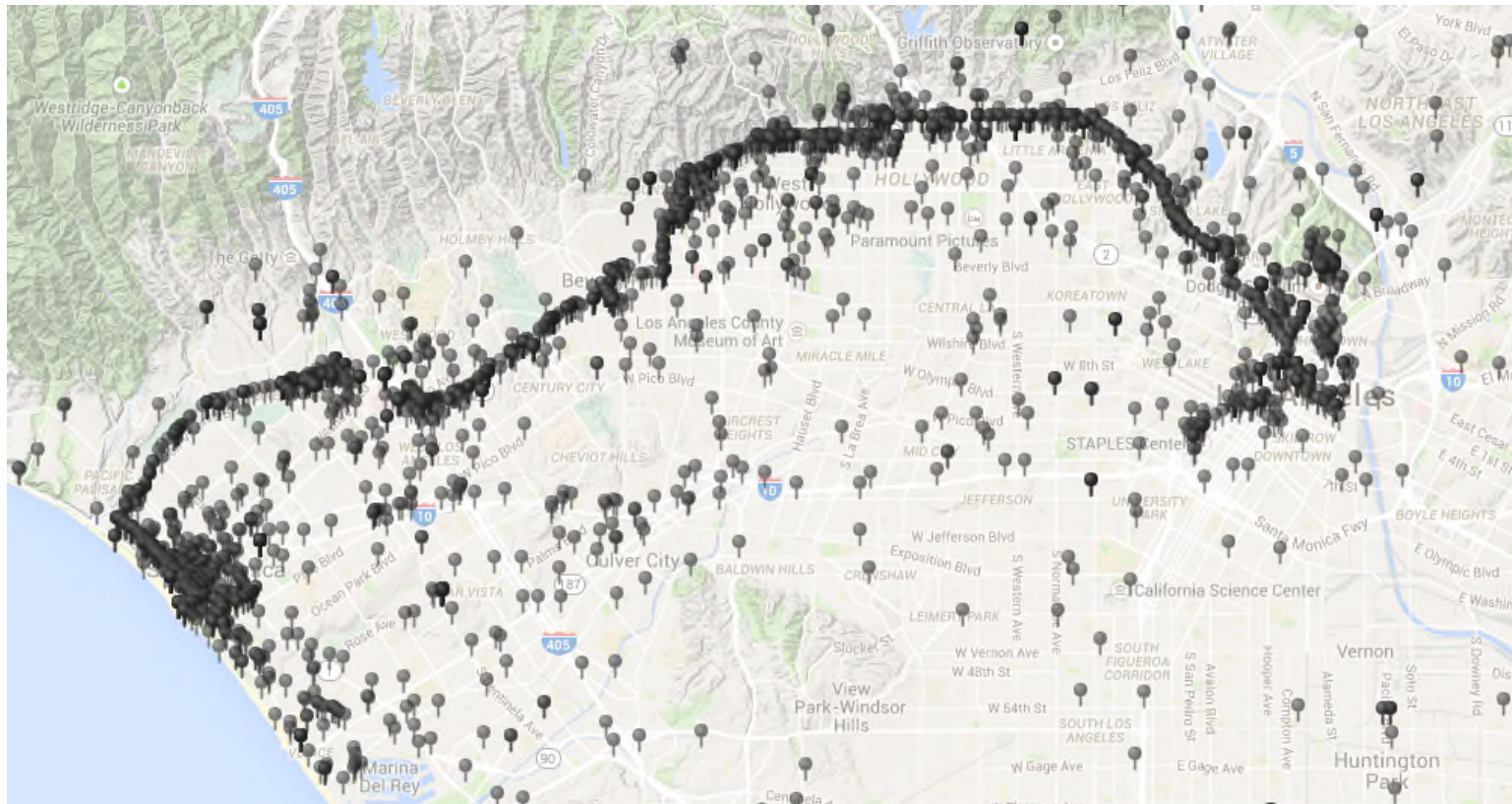
Instagram

- Tracking "LA Marathon"
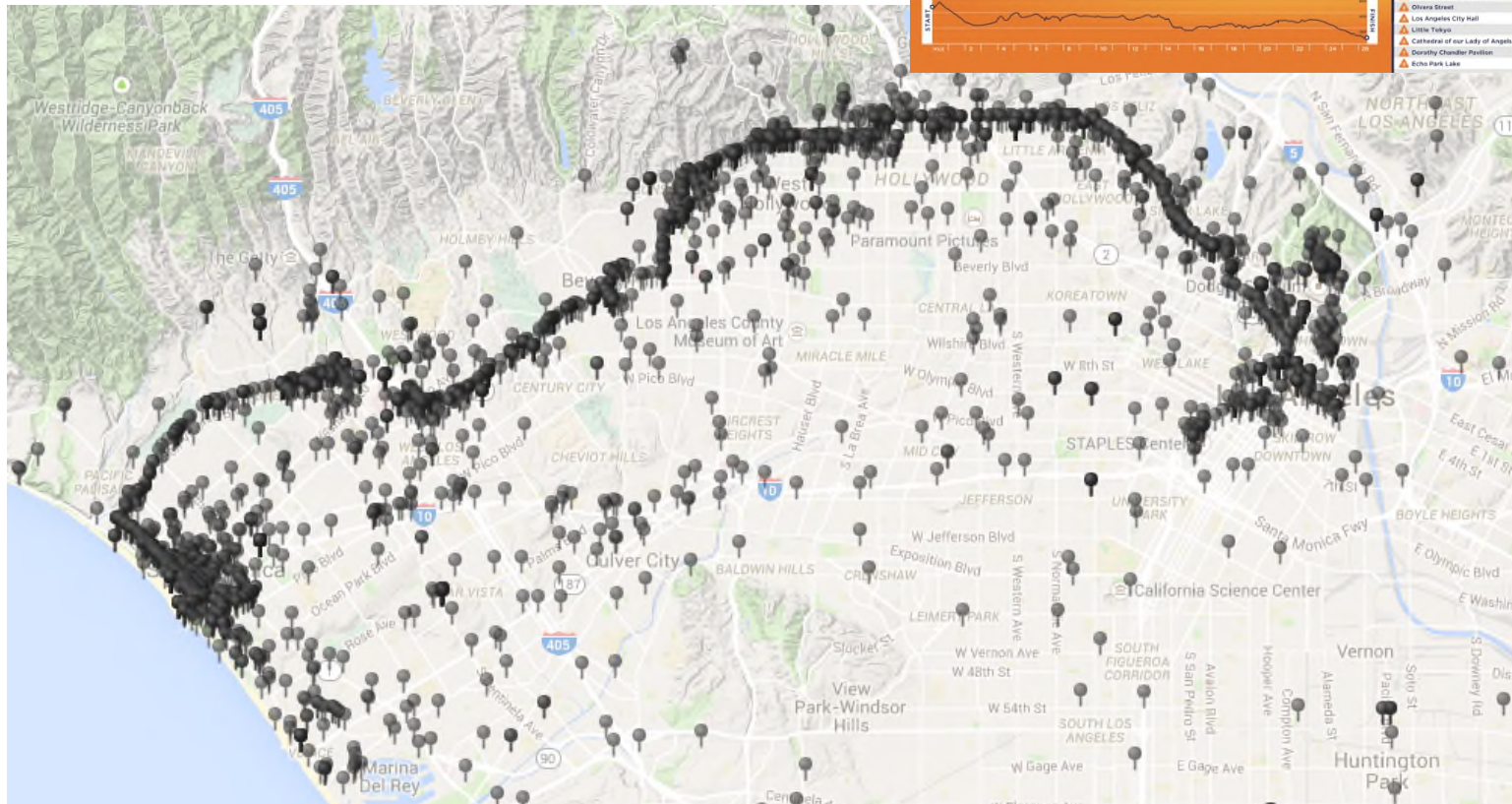
Instagram Tracking

- Tracking "LA Marathon": Early Stage
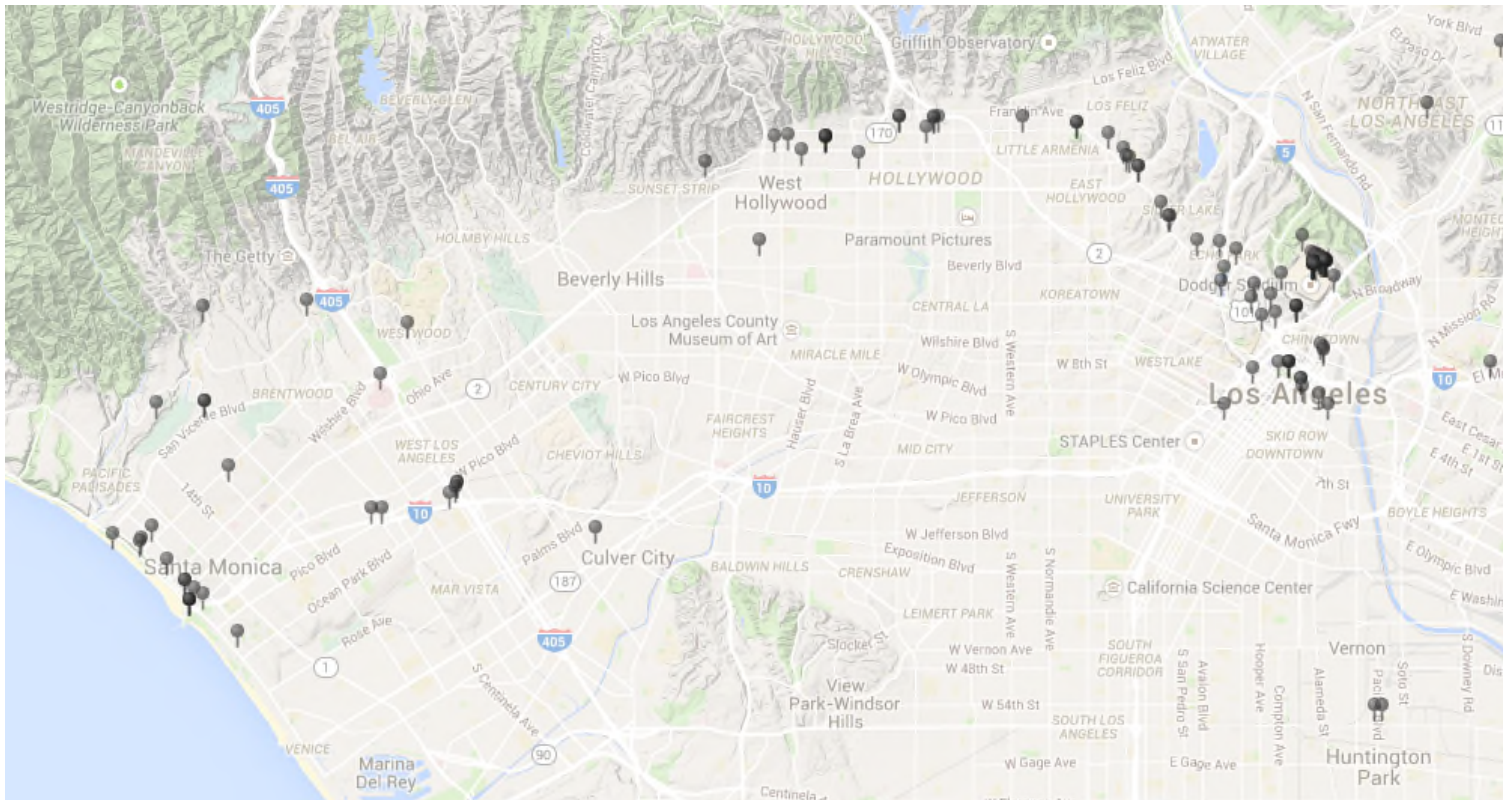
Instagram Tracking

- Tracking "LA Marathon": Middle
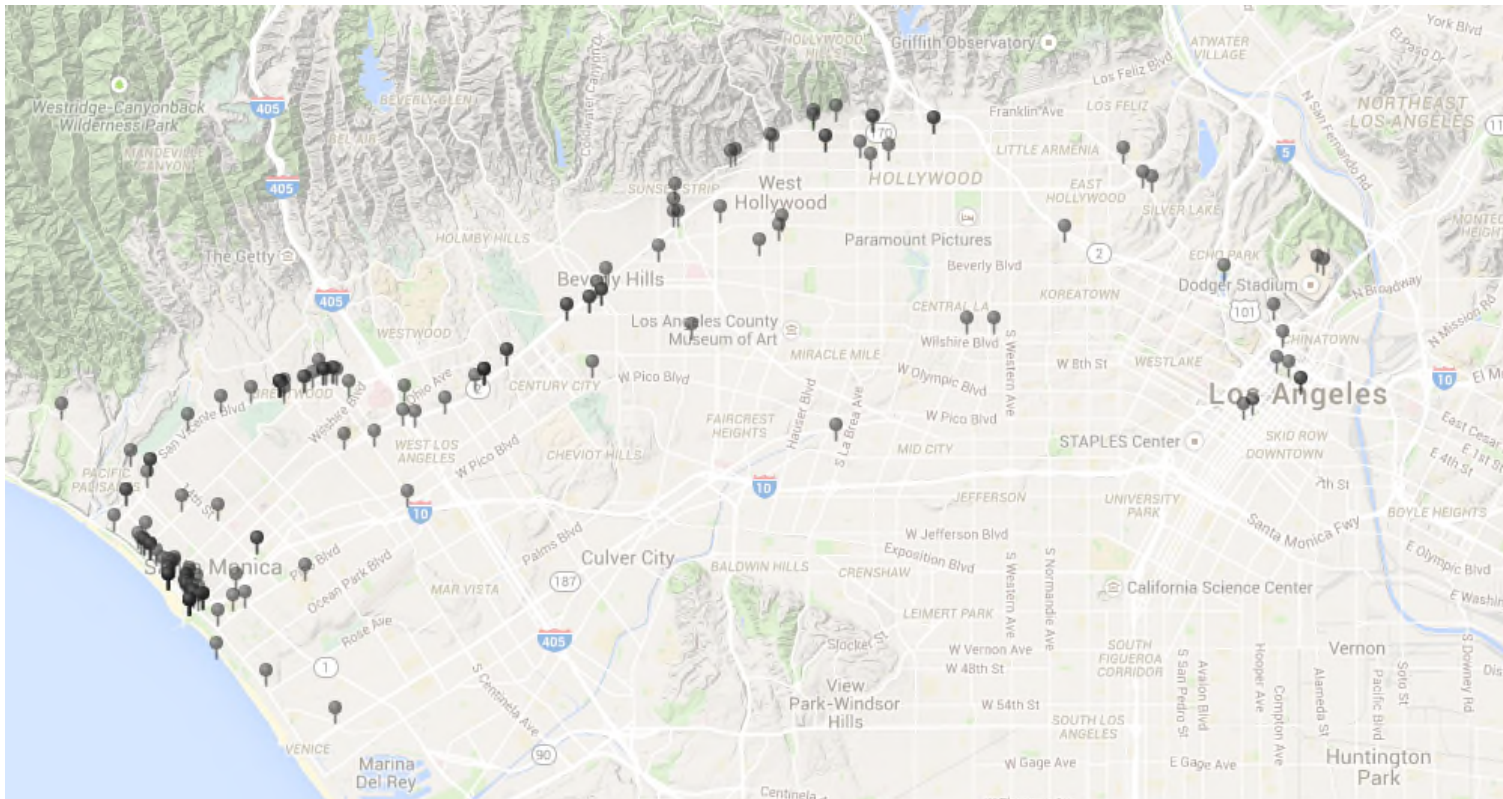
Instagram Tracking

- Tracking "LA Marathon": Late Stage
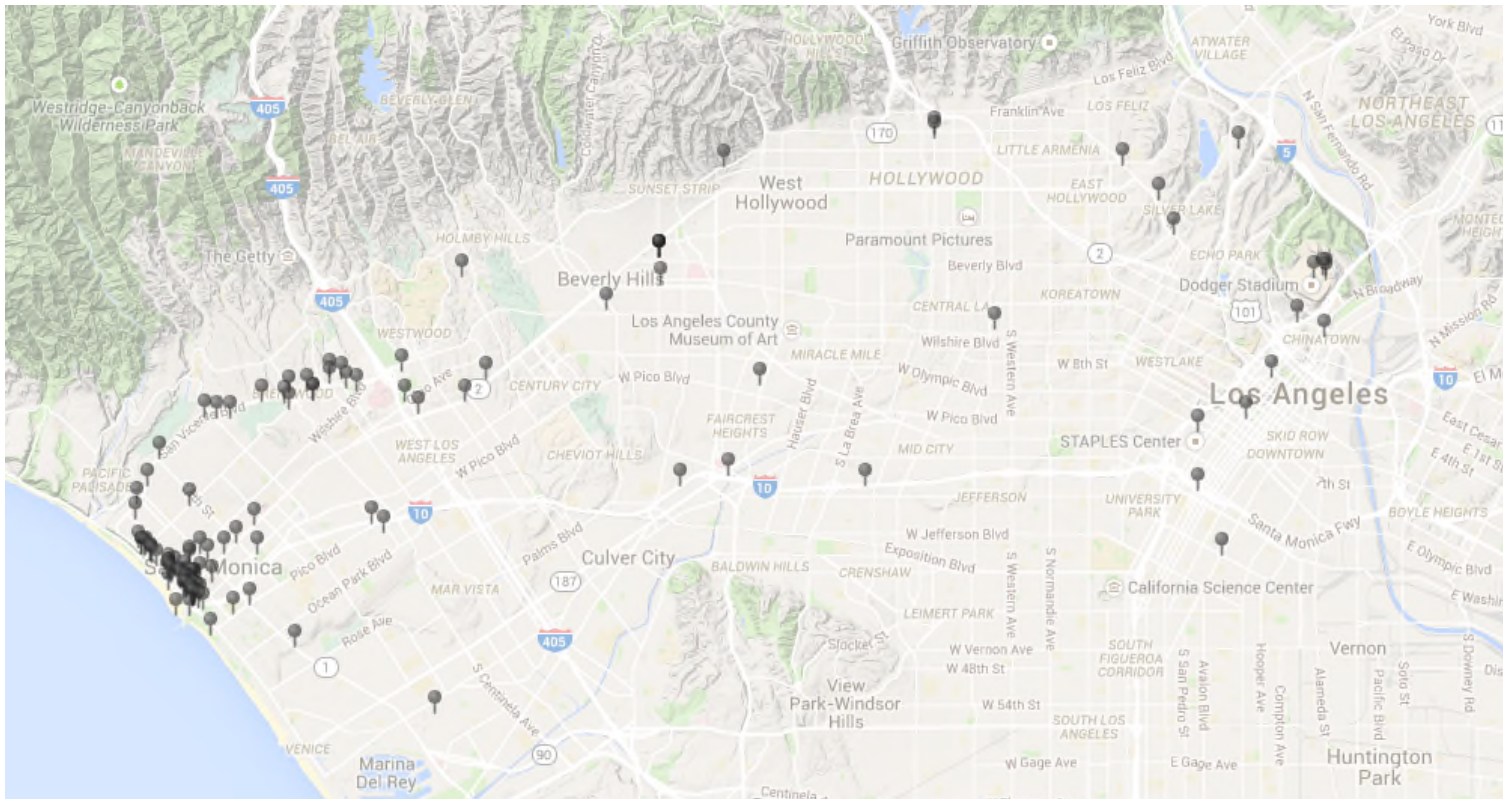
Challenge: Extractive Summarization

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
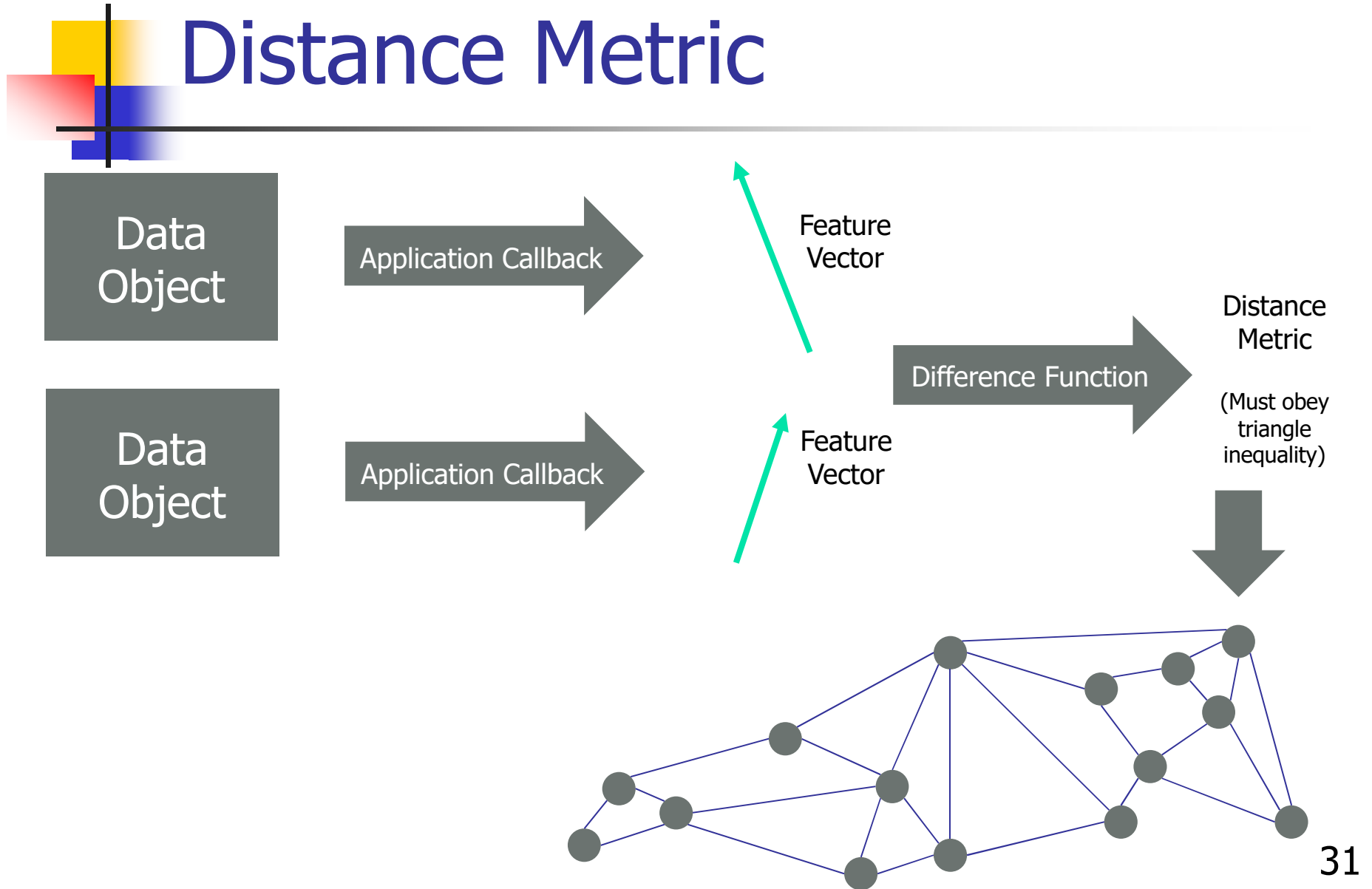
Customizability: The Distance Metric

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)

Data Object

Application Callback

Feature Vector

Data Object

Application Callback

Feature Vector

Difference Function

Distance Metric

(Must obey triangle inequality)

Application specific functions (customization API)

29

Customizability: The Distance Metric

Data Object → Application Callback → Feature Vector

Data Object → Application Callback → Feature Vector

Feature Vector + Feature Vector → Difference Function → Distance Metric

(Must obey triangle inequality)

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

Customizability: The Distance Metric

Data Object → Application Callback → Feature Vector

Data Object → Application Callback → Feature Vector

Difference Function → Distance Metric

(Must obey triangle inequality)

31

Customizability: The Distance Metric

Data Object	Application Callback →	Feature Vector
Data Object	Application Callback →	Feature Vector

Difference Function →

Distance Metric

(Must obey triangle inequality)

Hierarchical Clustering

32

Summarization

Data Object

Application Callback

Feature Vector

Data Object

Application Callback

Feature Vector

Difference Function

Distance Metric

(Must obey triangle inequality)

Hierarchical Clustering

33

Summarization

Data Object → Application Callback → Feature Vector

Data Object → Application Callback → Feature Vector

Difference Function → Distance Metric

(Must obey triangle inequality)

Hierarchical Clustering

34

Summarization

Data Object → Application Callback → Feature Vector

Data Object → Application Callback → Feature Vector

Difference Function → Distance Metric

(Must obey triangle inequality)

Hierarchical Clustering

35

Summarization

Data Object → Application Callback → Feature Vector

Data Object → Application Callback → Feature Vector

Feature Vector, Feature Vector → Difference Function → Distance Metric

(Must obey triangle inequality)

Hierarchical Clustering

Representative sampling versus noise reduction?

36

A Network Paradigm Shift
Communication → Information Distillation

The data fire-hose effect

- 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*?
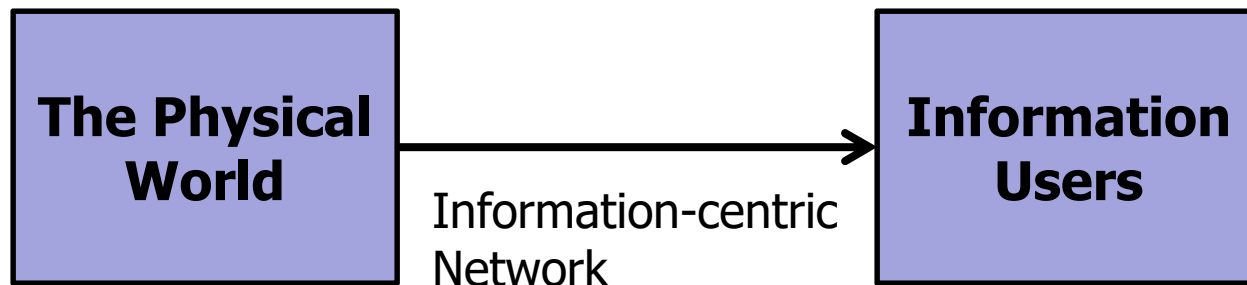
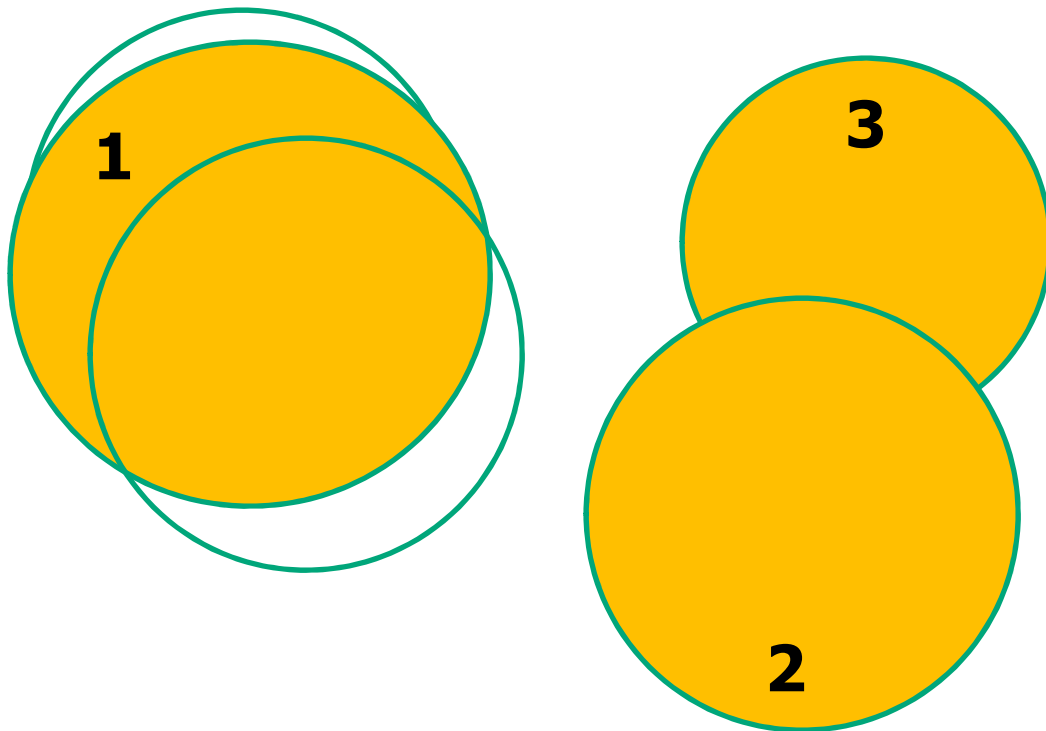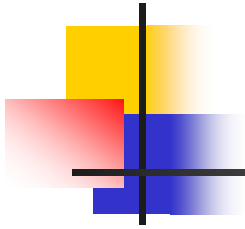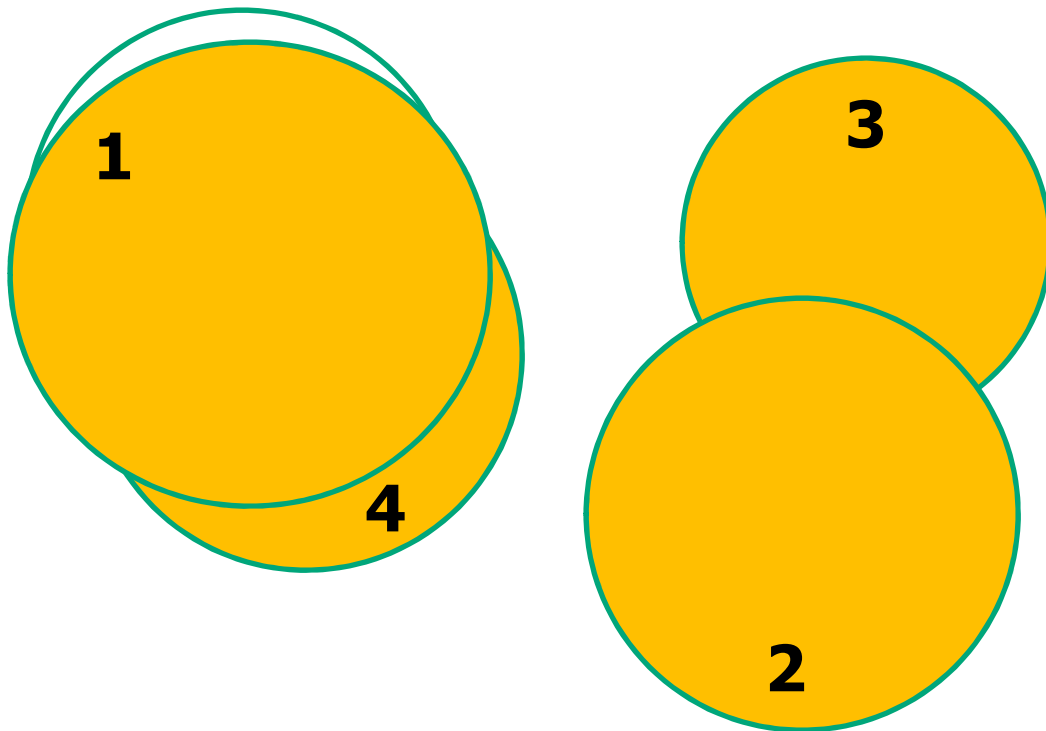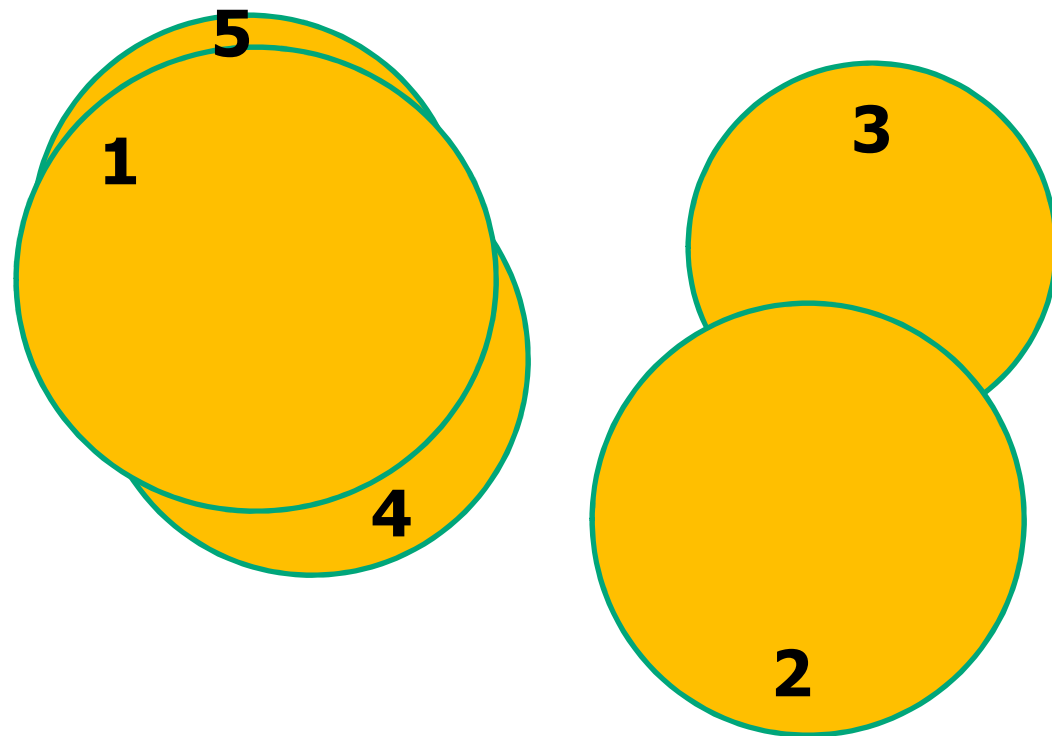|		
The Physical World	→ Information-centric Network	**Information Users**
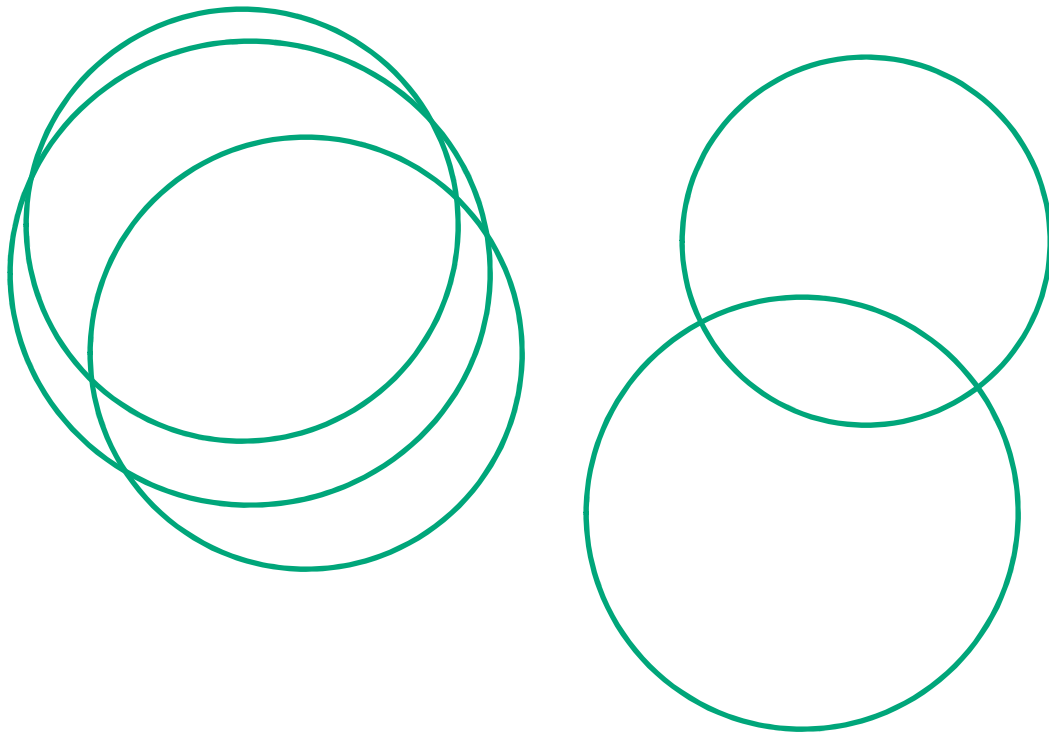
Information-maximizing Prioritization

- Determine transmission order?

Information-maximizing Prioritization

1

- Determine transmission order?

Information-maximizing Prioritization
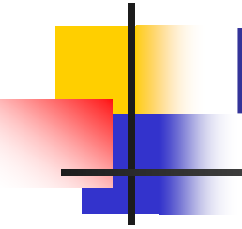
- Determine transmission order?
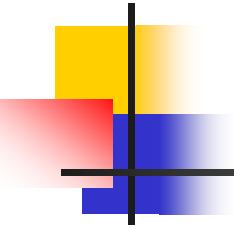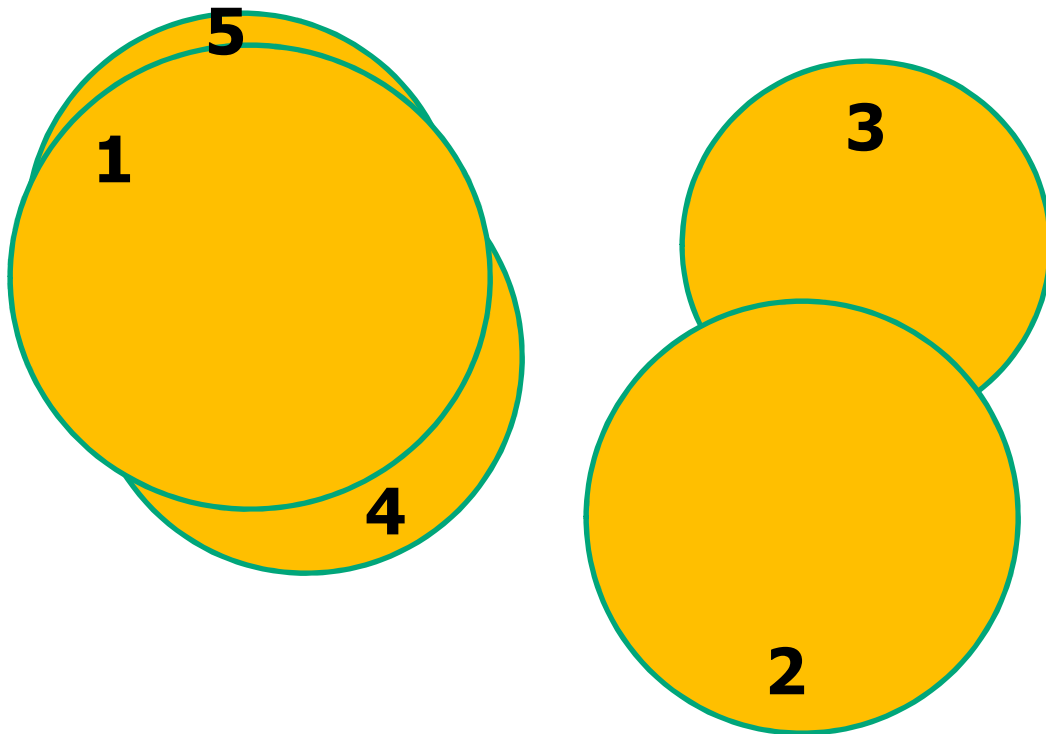
1

2

Information-maximizing Prioritization

- Determine transmission order?

Information-maximizing Prioritization

- Determine transmission order?
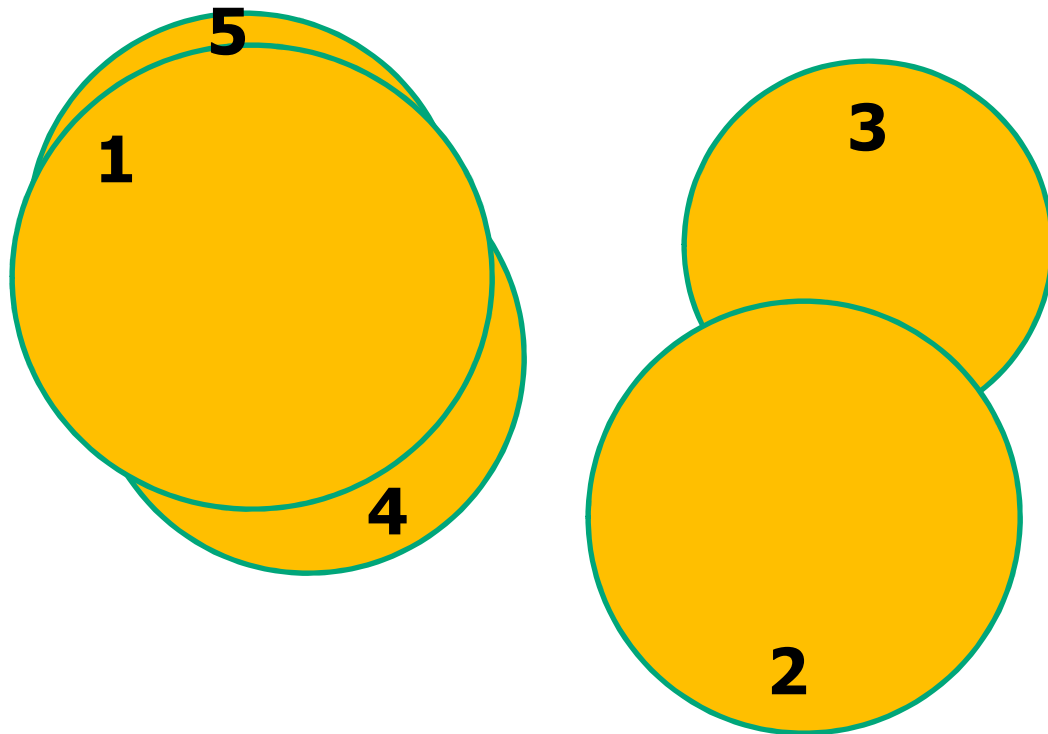
Information-maximizing Prioritization

- Determine transmission order?

Information-maximizing Prioritization

- Determine transmission order?

5

1

3

4

2

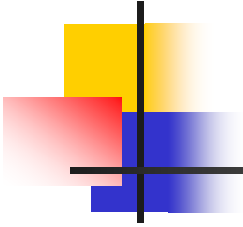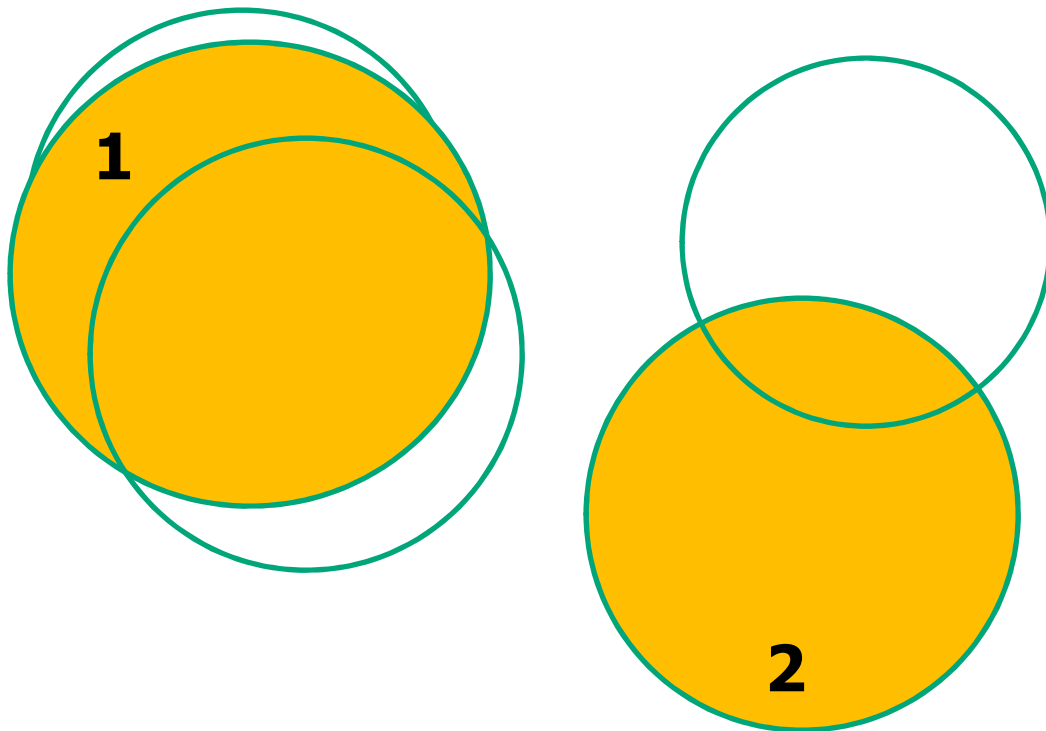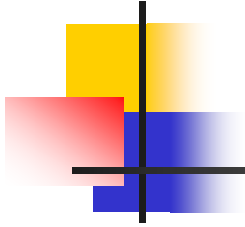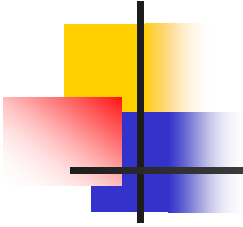**Coverage-monotonic scheduling**

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…

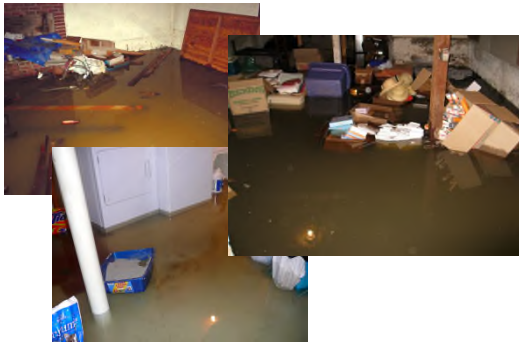Images are collected from the Internet

Hurricane Katrina 2005

Nepal earthquake 2015

Thailand flood 2011

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

Challenge: Data Selection to Maximize Coverage

Collapse on Park Ave.

Fire on 6th and Main.

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
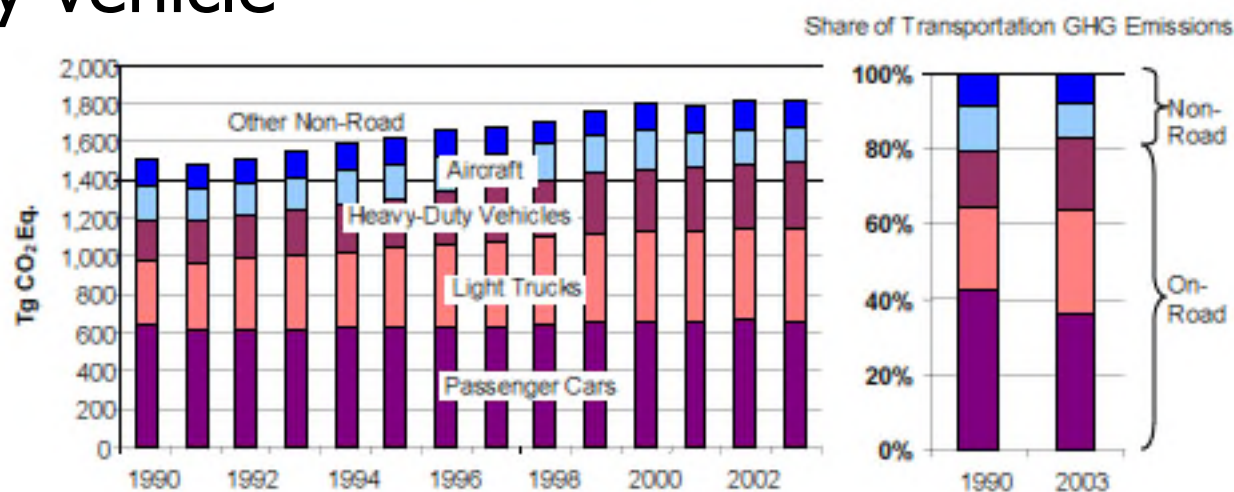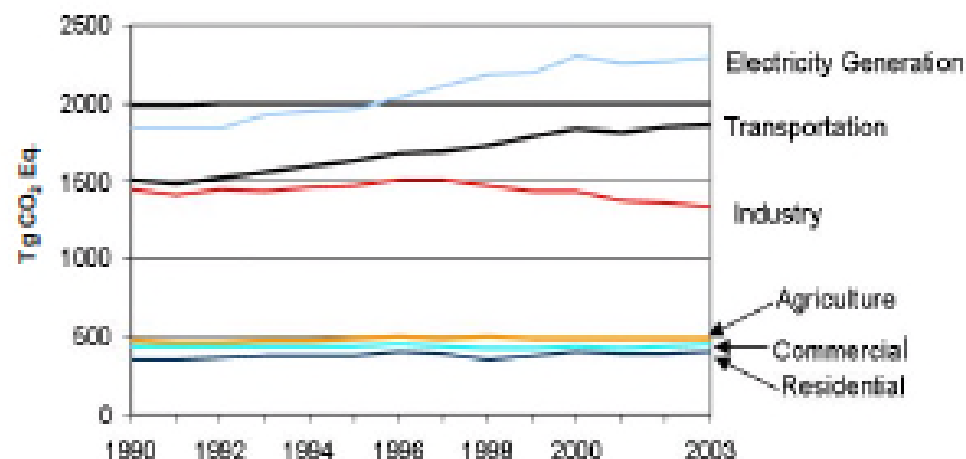
Class Projects – Energy
Transportation Energy Minimization

- 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
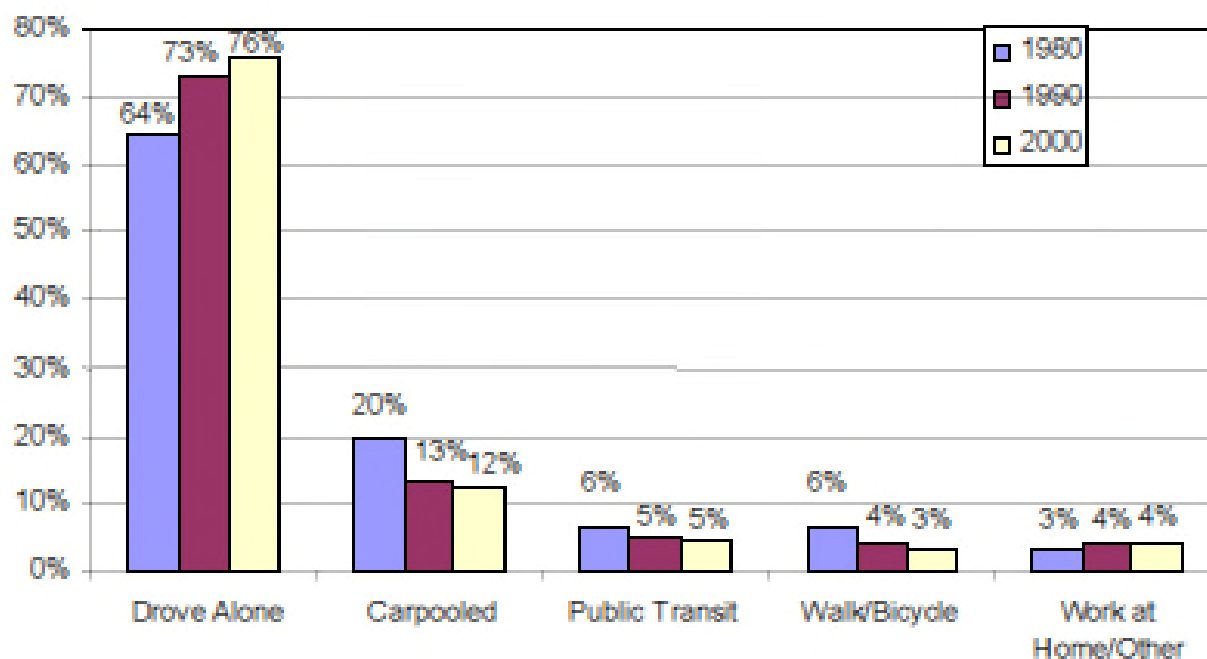
- How to reduce energy & emissions?

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

Shortest and fastest
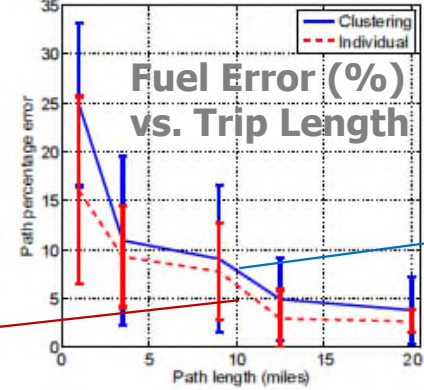
Green GPS

Most fuel-efficient

Subscribers

Subscribers:
Premium service
High savings

Open access:
Standard service
Average savings

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

Fuel Error (%) vs. Trip Length

Clustering
Individual

Path percentage error

Path length (miles)

Fuel Data   **+**   Physical Models

$$F_{engine} = \frac{\Gamma(\omega) G g_k}{r}$$

F_{engine}

F_{air}

$$F_{air} = \frac{1}{2} c_d A \rho v^2$$

$$F_{friction} = c_{rr} mg \cos(\theta)$$

F_g

$F_{friction}$

$$F_g^g = mg \sin(\theta)$$

θ
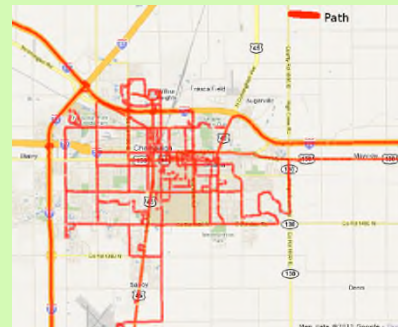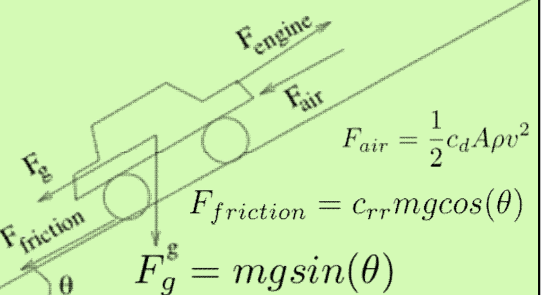
$$F_{car} = F_{engine} - F_{friction} - F_{air} - F_g$$

Server

Path

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 Green... may ... other. ... about ...

loped by University

ponent

GPS GOING GREEN

Green GPS

Tarek Abdelzaher
U of I Professor

17 HD
66° 5:04 PM

☑ shortest
☑ fastest

URBANA fuelOptim

WCIA

TOYOTA BLE NEARLY 140 POINTS ON WORD FIRST· 63° 5:06
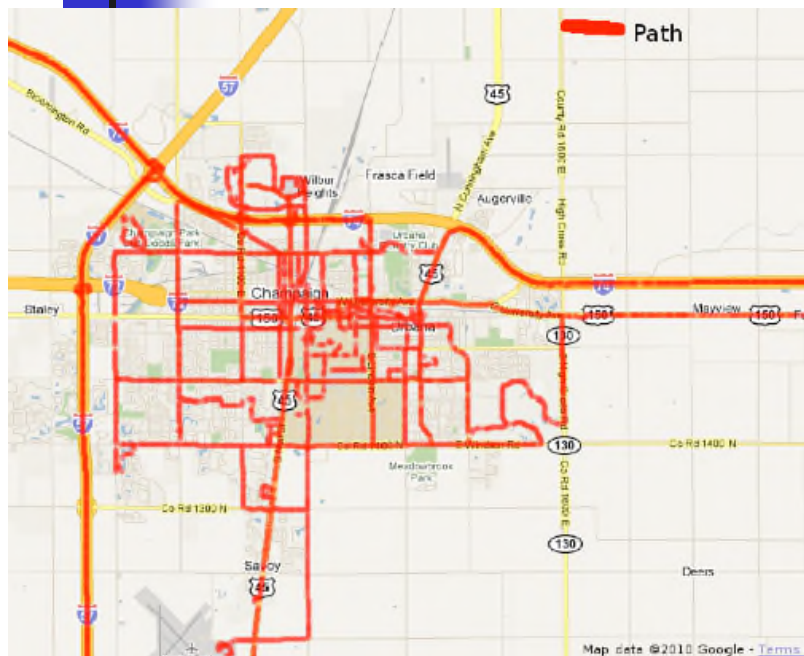
A Modeling Challenge

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?

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

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

	Knob	Sensor
CPU	Frequency	Utilization, Frequency, Temperature
MEM		Utilization,
NIC		Received / Sent packets/bytes
PDU		Power consumption of each individual machine
CRAC	set point*	Input and outlet temperature, Set point

Failures in Complex Systems
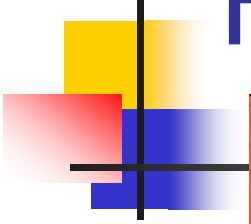
When systems fail, a common goal is:

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