



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



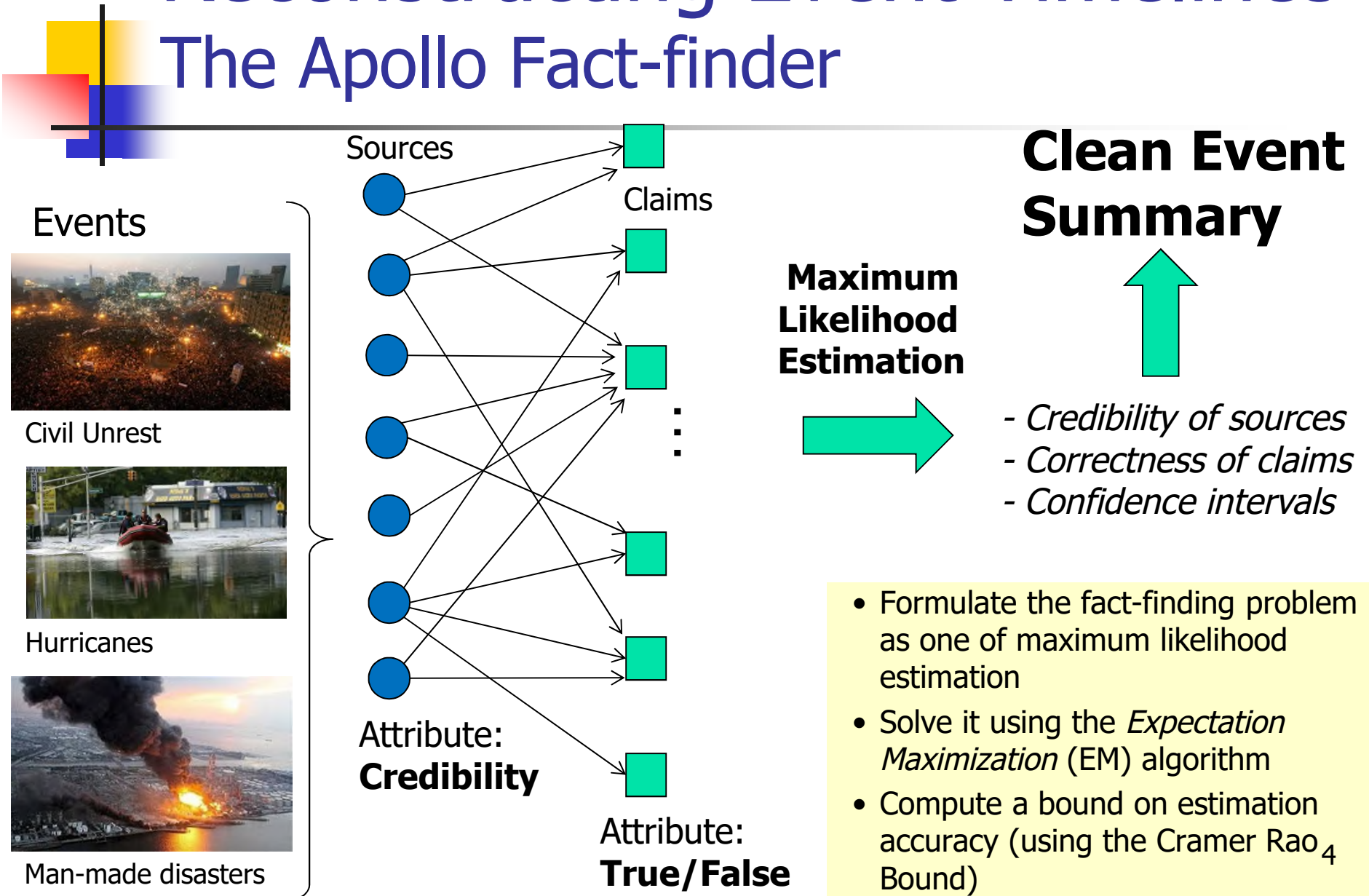
**Estimated
State**

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



Attribute:
Credibility



Claims

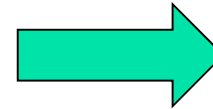


⋮



Attribute:
True/False

Maximum
Likelihood
Estimation



**Clean Event
Summary**

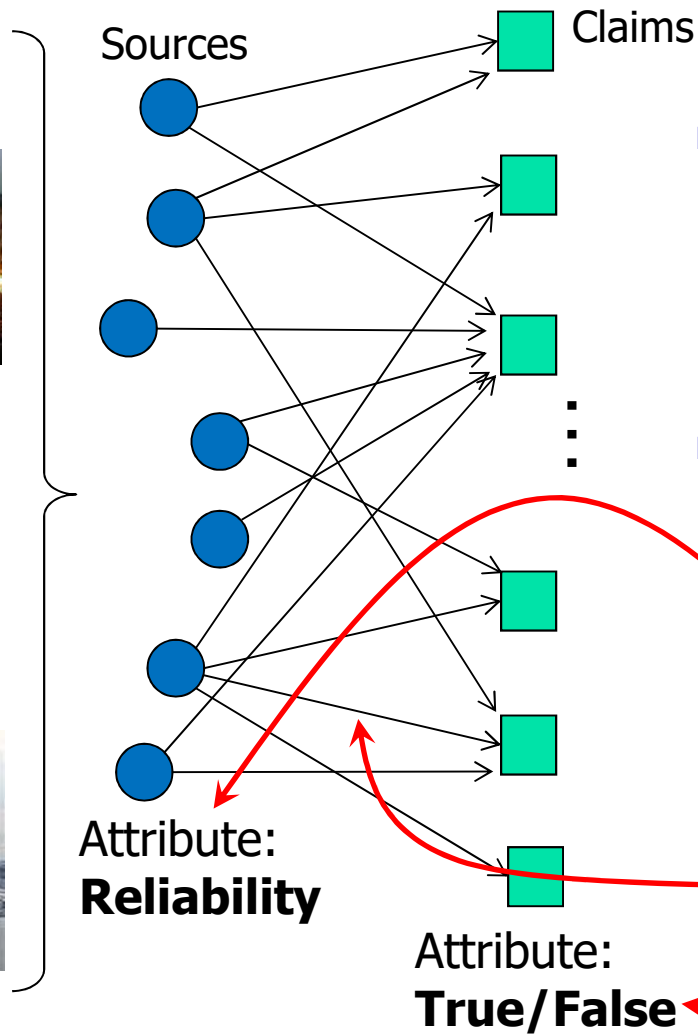


- *Credibility of sources*
- *Correctness of claims*
- *Confidence intervals*

- 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 Rao₄ Bound)

Social Channel "Decoding"

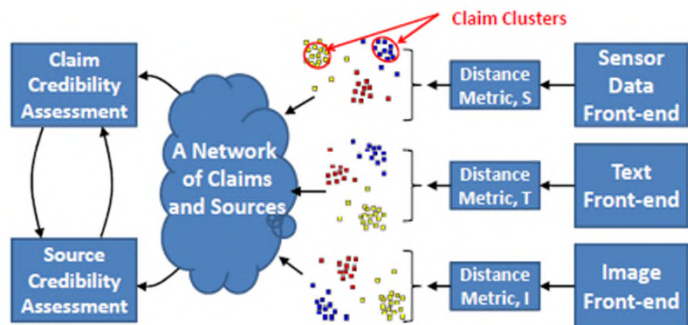
A Maximum Likelihood Estimation Problem



- 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



Create new task

Keyword 1 or
Keyword 2 or
Keyword 3 or from
Latitude Longitude Radius (miles) [Hide Map](#)



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/jol8OlFxfn via @reuters #PJNET	Putin Orders Massive Strike Against Saudi Arabia If West Attacks Syria http://t.co/SFLJ9ghwb
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/IN4CkUePUj #Syria http://t.co/TorVFUfZF
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/UIFF1wY9gx	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, ...)

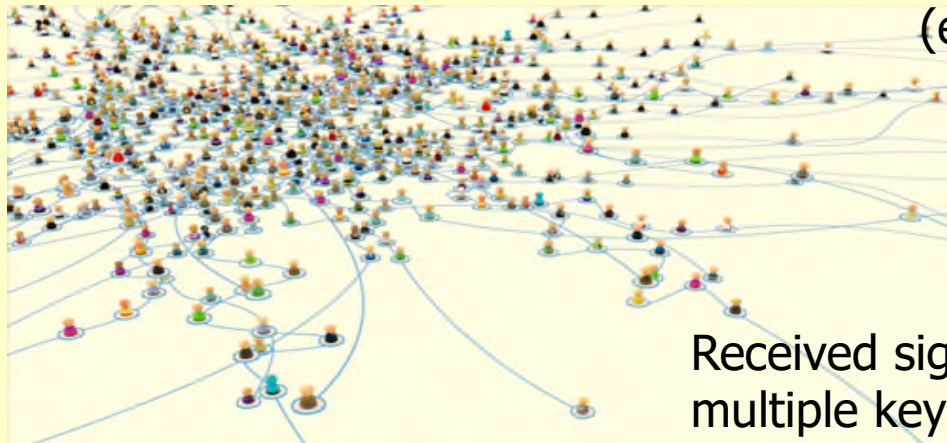


Received signature (energy in
multiple signal frequency bands)



Physical event

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



Received signature (energy in
multiple keyword frequency bands)



Demultiplexing

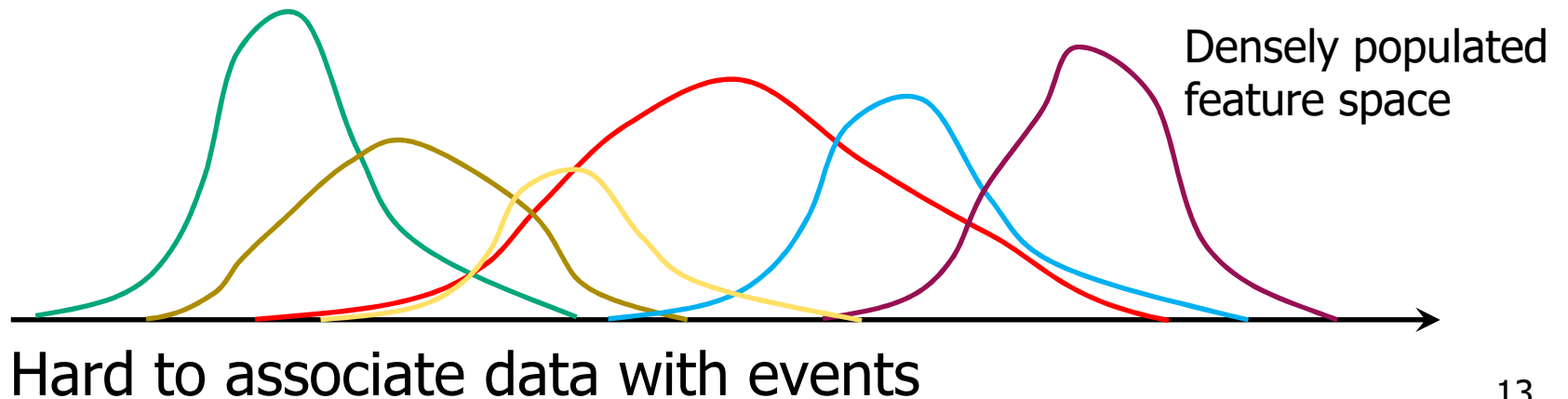
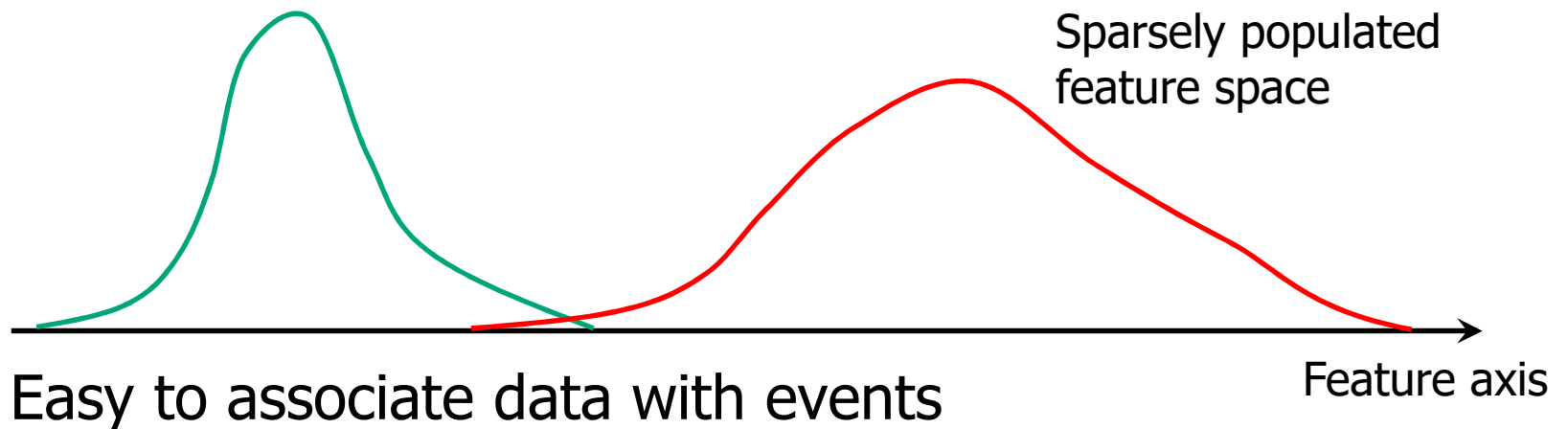
- A world of “protest” – this morning:
 - Angry French farmers and 1,000 tractors head for Paris protest. Photo @MartinBureau1 #AFP <http://t.co/j5DdveSHZh>
 - VIDEO: Tractor protest descends on Paris: French farmers protesting about high taxes have taken a convoy of tr... <http://t.co/hKievMFpq3>
 - 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?

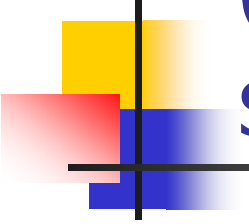
Events and Signals: A Data Association Problem





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

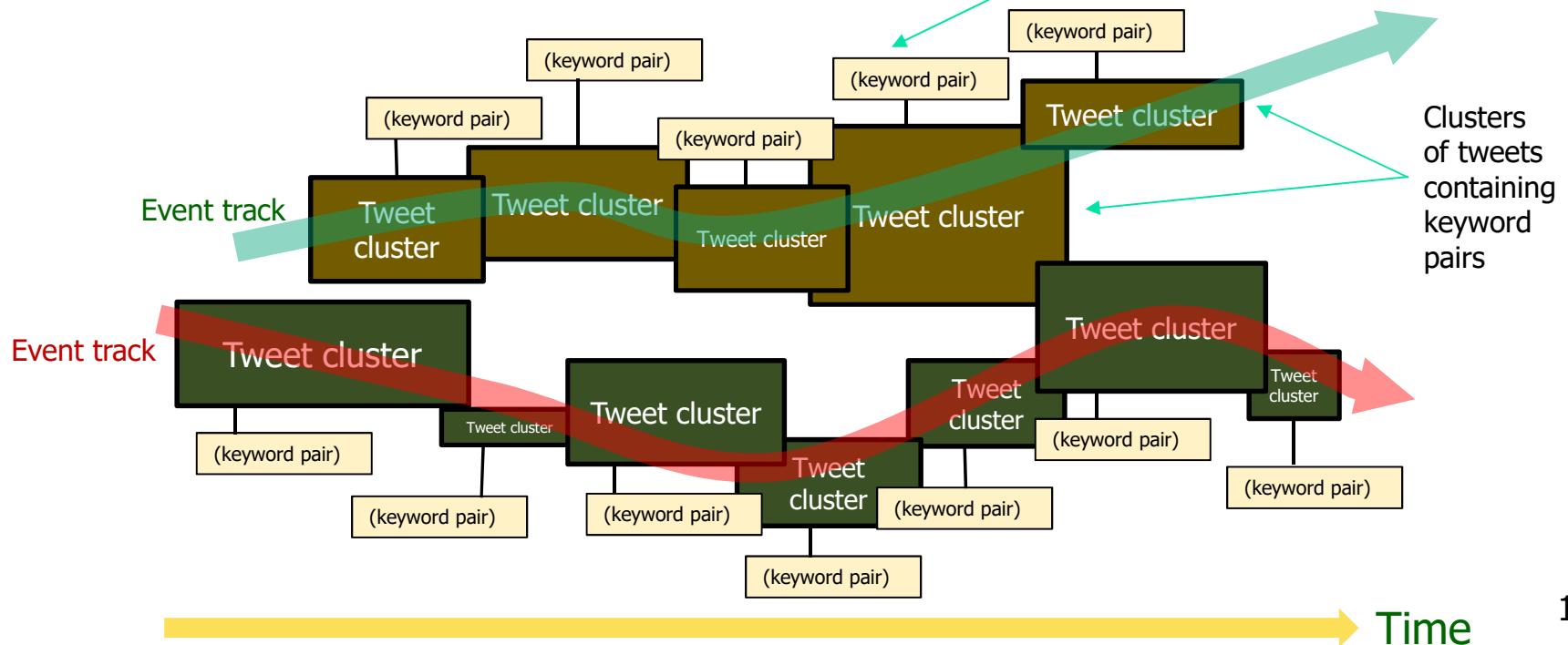
- How to detect new event signatures?
 - Find high-information-gain signatures (new spikes in the frequency spectrum)
 - Bin tweets that contain a new signature into a cluster
 - Determine if this cluster is of a new event or not using frequency domain distance (note: some events will have more than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

Event Detection, Consolidation and Tracking

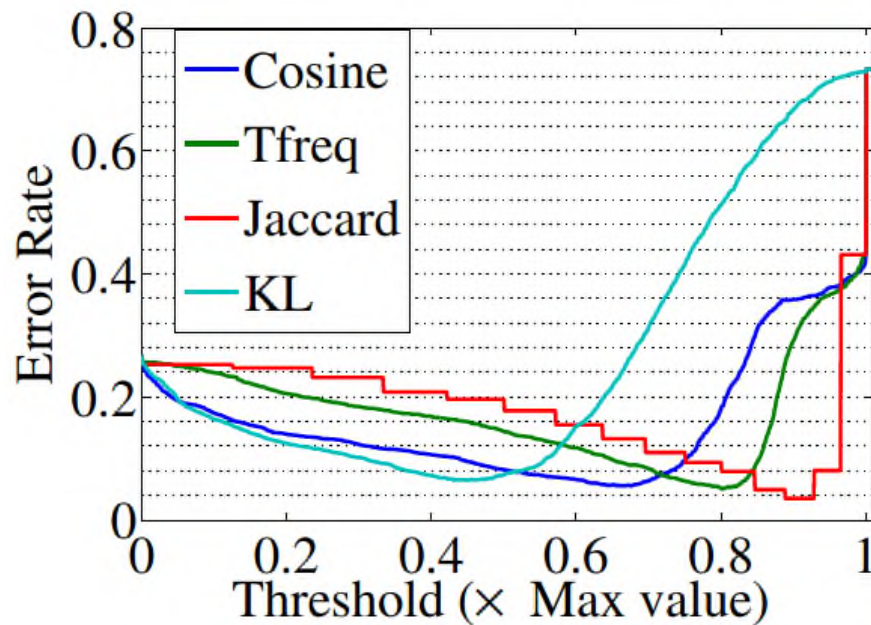
Three key ideas:

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

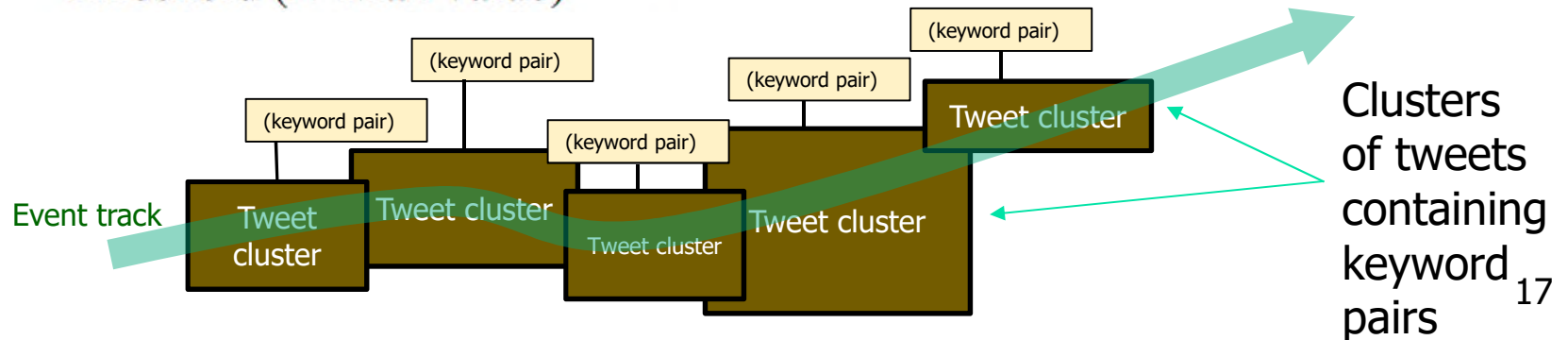
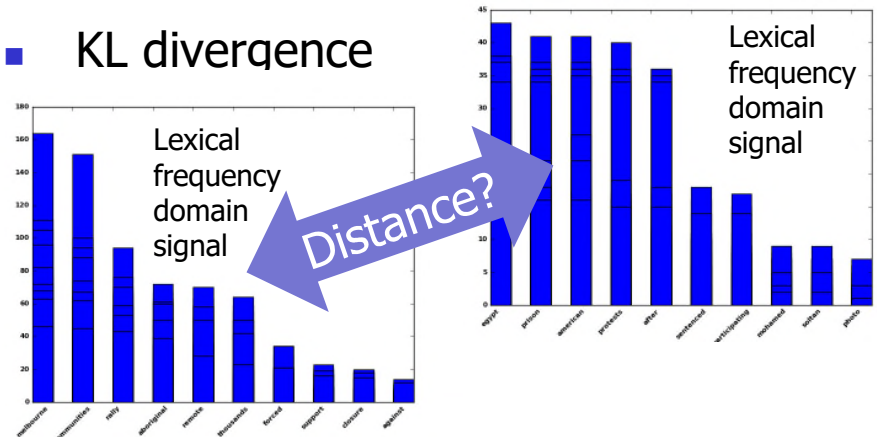
Automatically detected high-information-gain keyword pairs



Distance Metrics (For Merging Event Data Clusters)

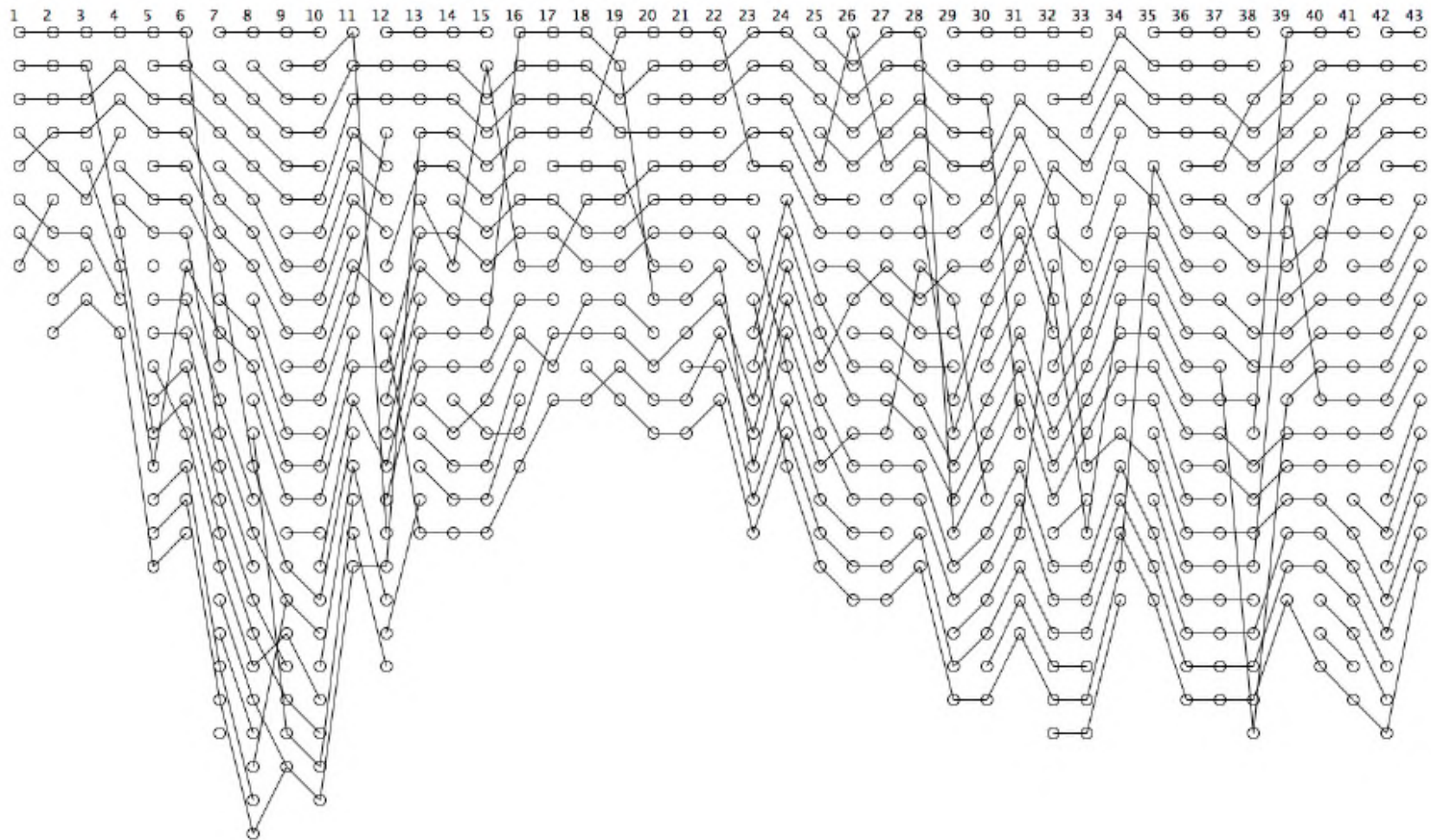


- Cosine similarity
- Term frequency difference
- Jaccard distance
- KL divergence





Event Tracks

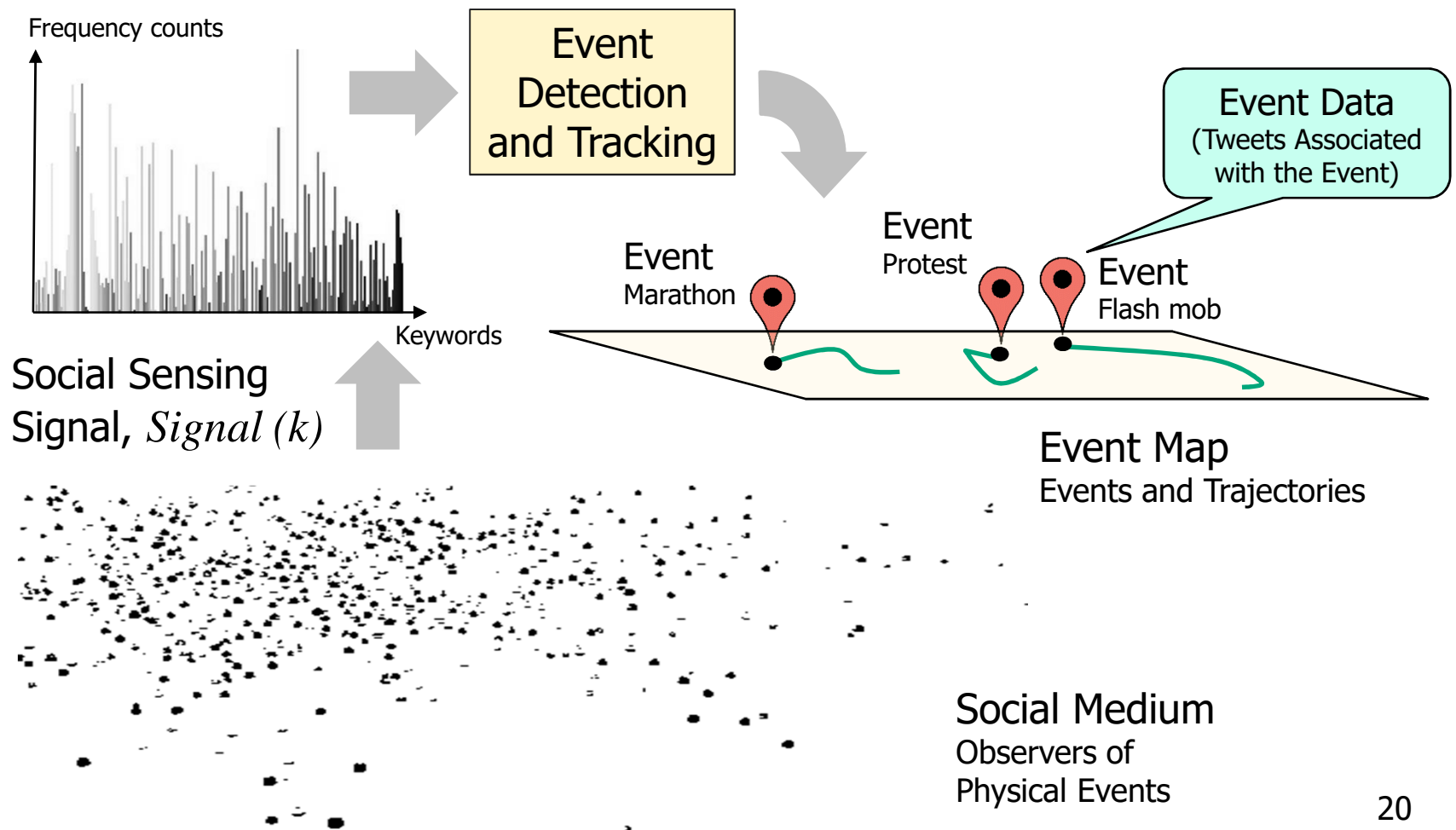


Recognizing Distinct Event Tracks

- 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	<p>Religion Bangladesh Braces for Protests After Islamist's Execution: senior official of the largest Islamis... http://t.co/NTBmIWSTme</p> <p>World News: Bangladesh braced for protests after Islamist leader's execution: Bangladeshi security personnel s... http://t.co/UPKiaFFHtW</p> <p>Bangladesh braces for protests after Jamaat leaders execution: Bangladesh braced for protests and fresh violen... http://t.co/3dcPFqKAQE</p>
Brazil protests	<p>Protests across Brazil seek ouster of president http://t.co/YmXZnsxbAQ</p> <p>FollowMePlease Brazil braces for nationwide protests, as groups seeking impeachment of presiden... http://t.co/5T150D0zIL BrinaldyHere</p> <p>Fresh anti-government protests in Brazil: Brazil on Sunday braced for more huge demonstrations against governm... http://t.co/8nJmX56MUM</p>
Turkey protests	<p>DTN Turkey: Turkey protests to Pope Francis after he brands Armenian killings 'genocide': Pontiff's run-in wit... http://t.co/dKx5iCwP9w</p> <p>Pope refers to Armenian genocide; Turkey protests. 24 April is 100th anniversary of start of the Armenian genocide http://t.co/ZFOCZnC411</p> <p>Telegraph: Turkey protests to Pope Francis after he brands Armenian killings 'genocide' http://t.co/YnJew4foN1 http://t.co/O6d5oYwG2p</p>

The Social Signal Layer



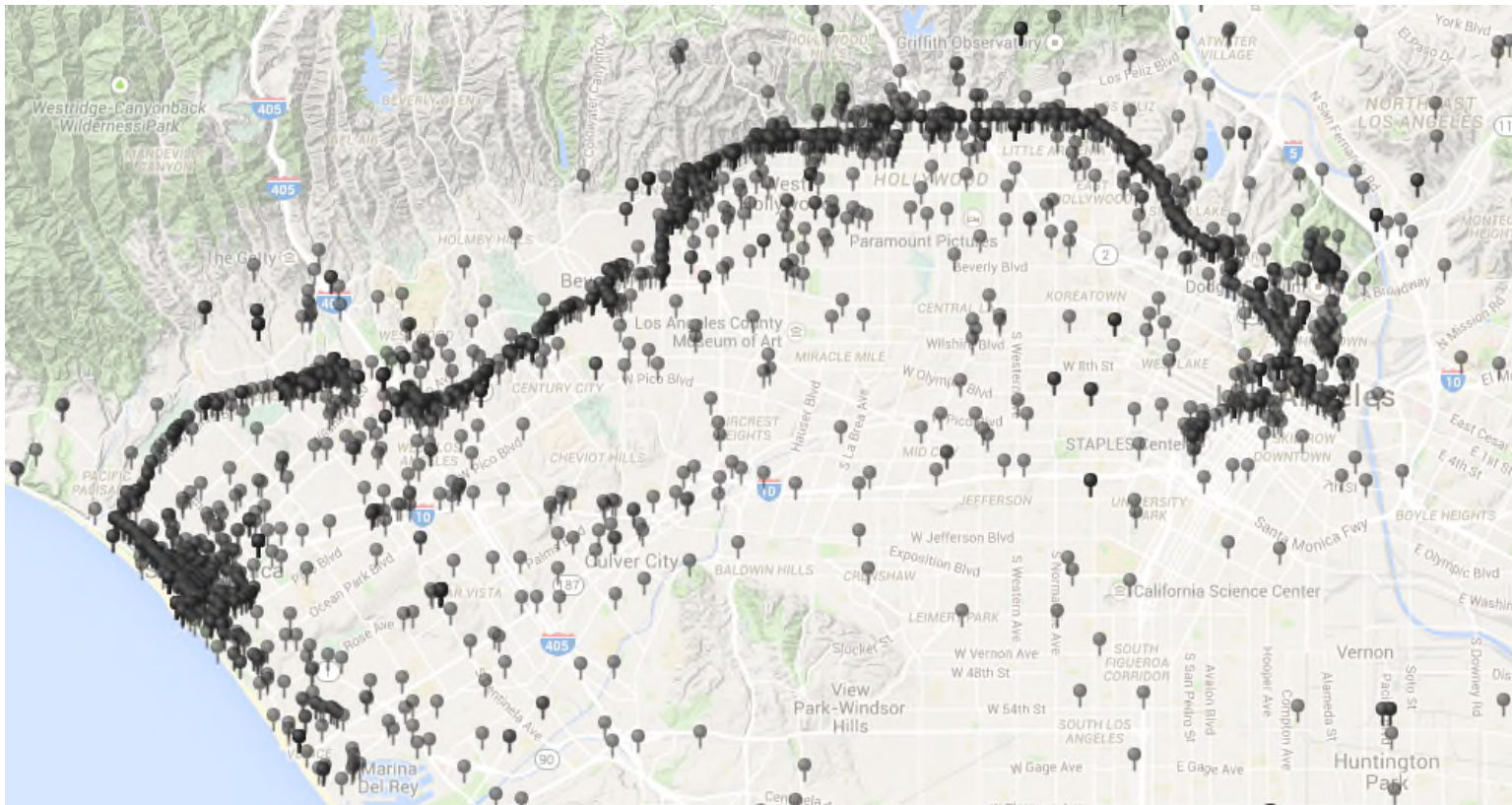


Event Localization with Instagram

- 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!

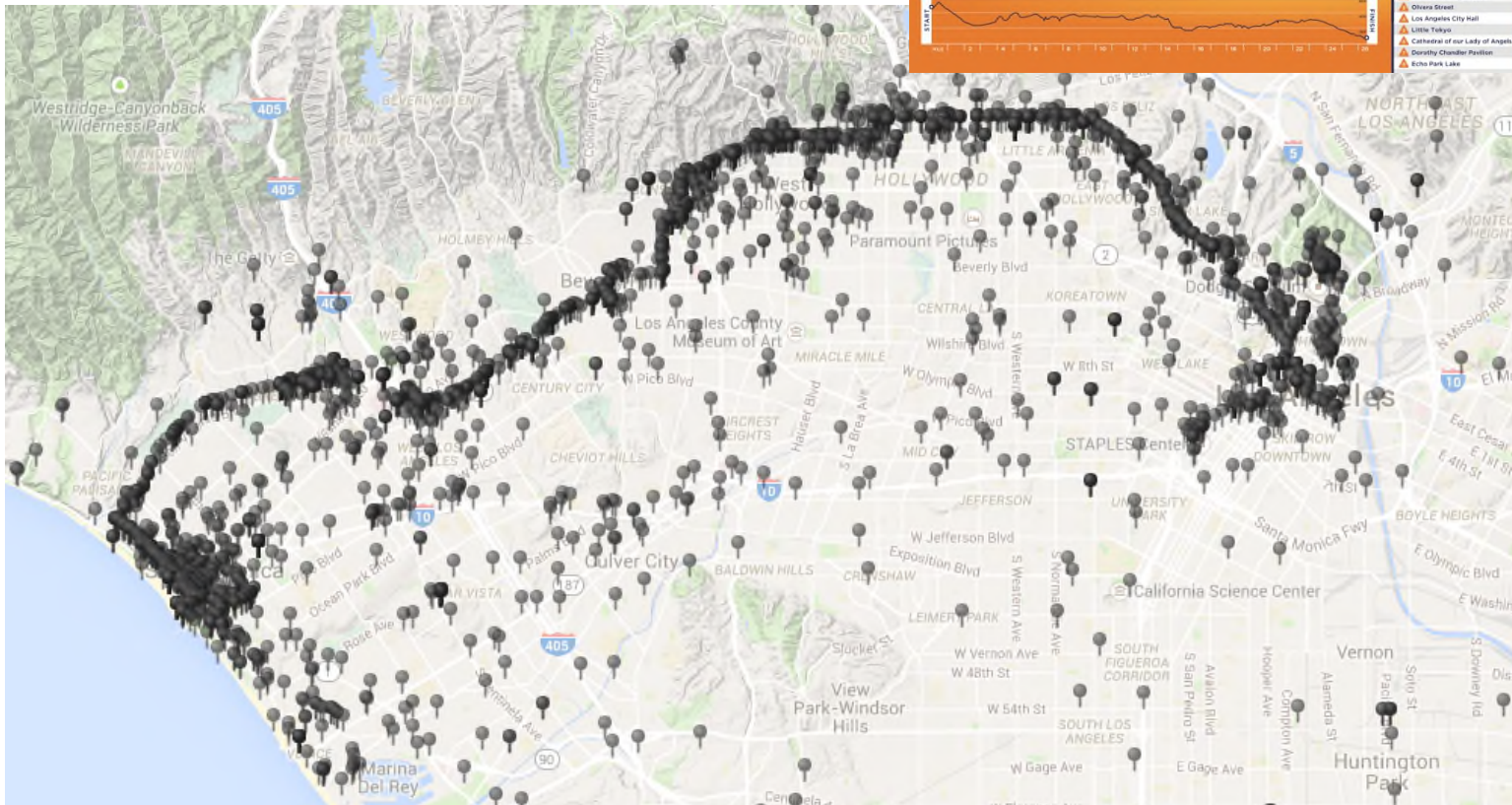
Instagram Localization

- Tracking “LA Marathon”



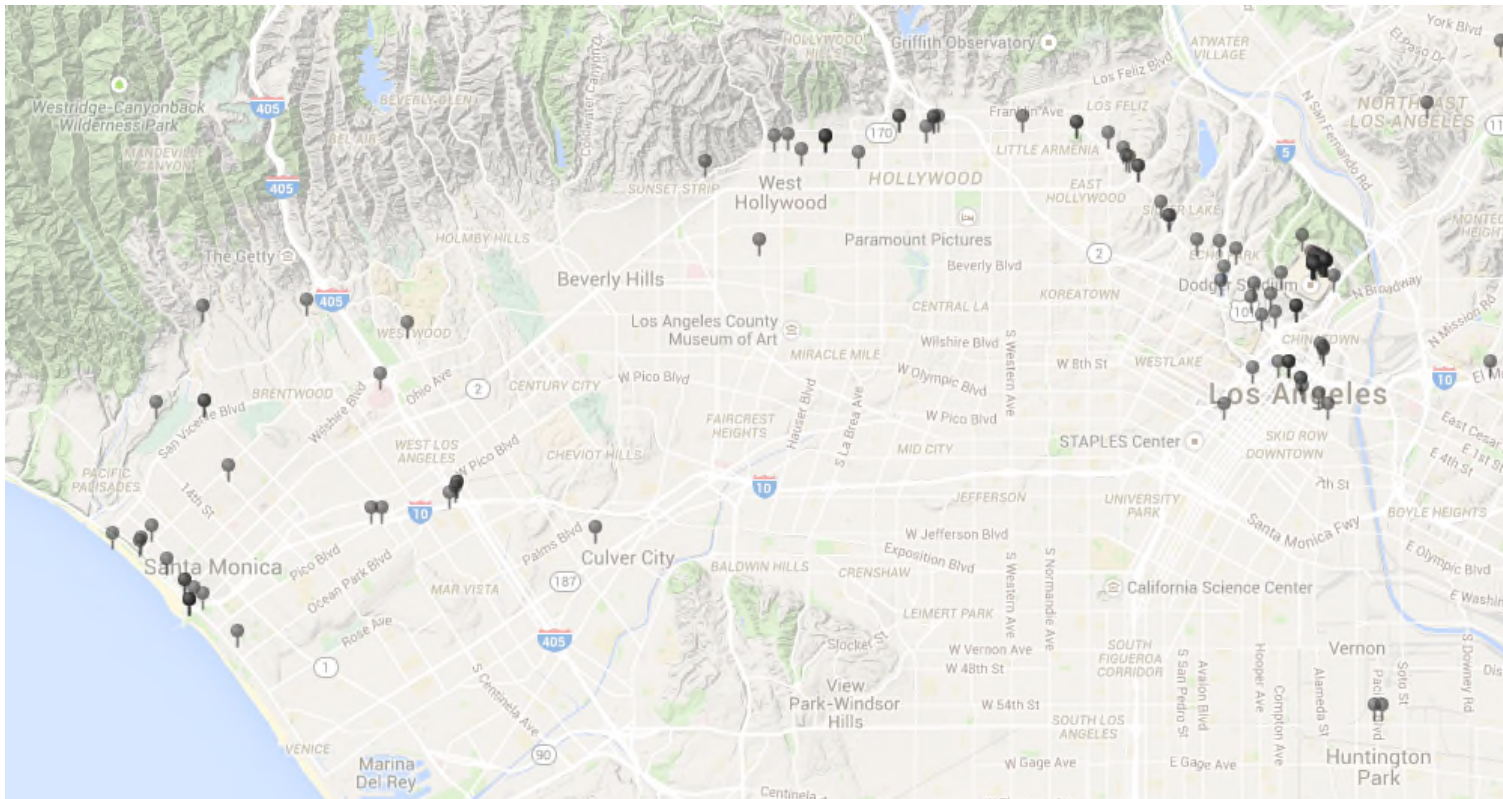
Instagram

Tracking "LA Marathon"



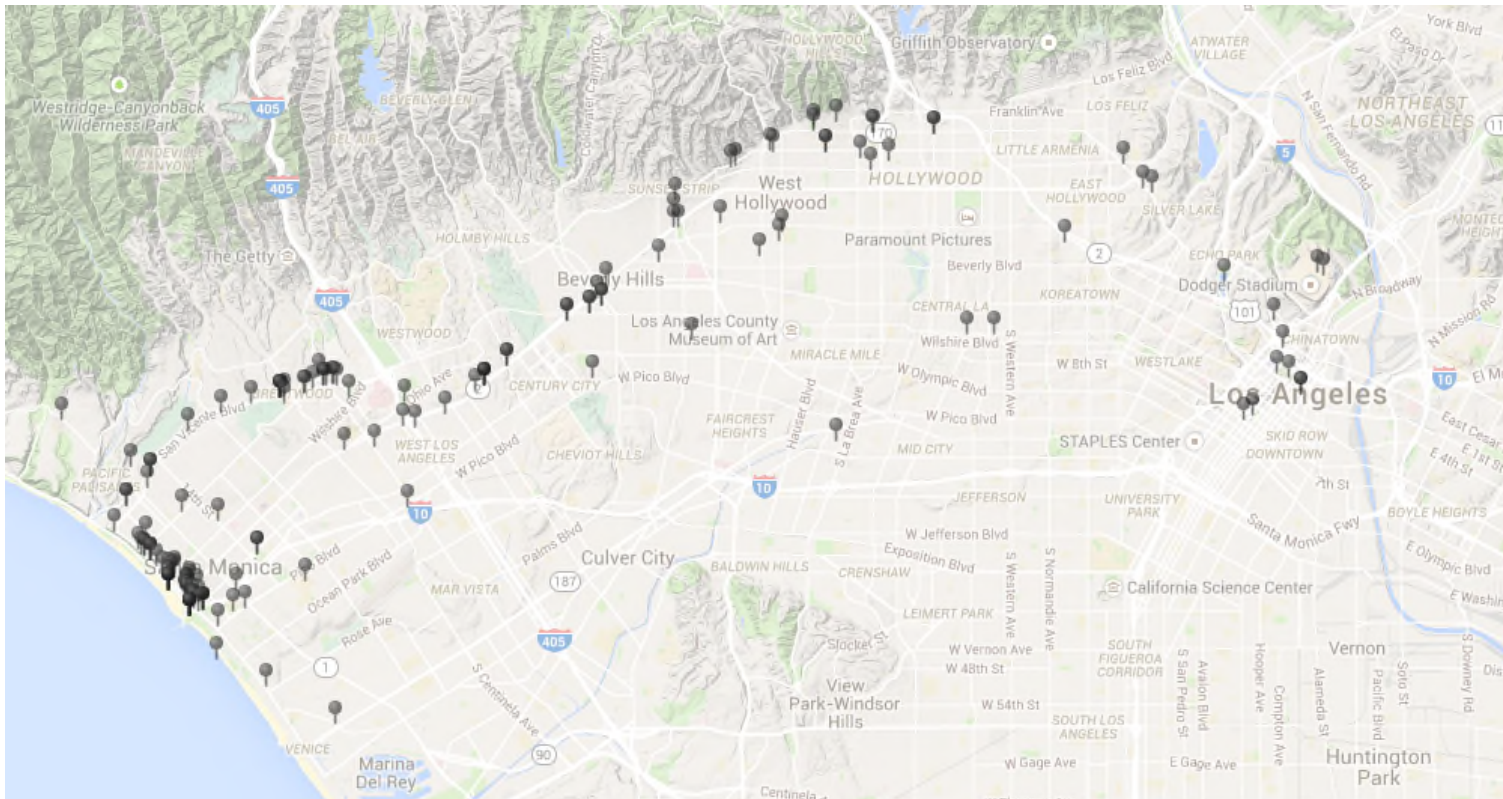
Instagram Tracking

- Tracking “LA Marathon”: Early Stage



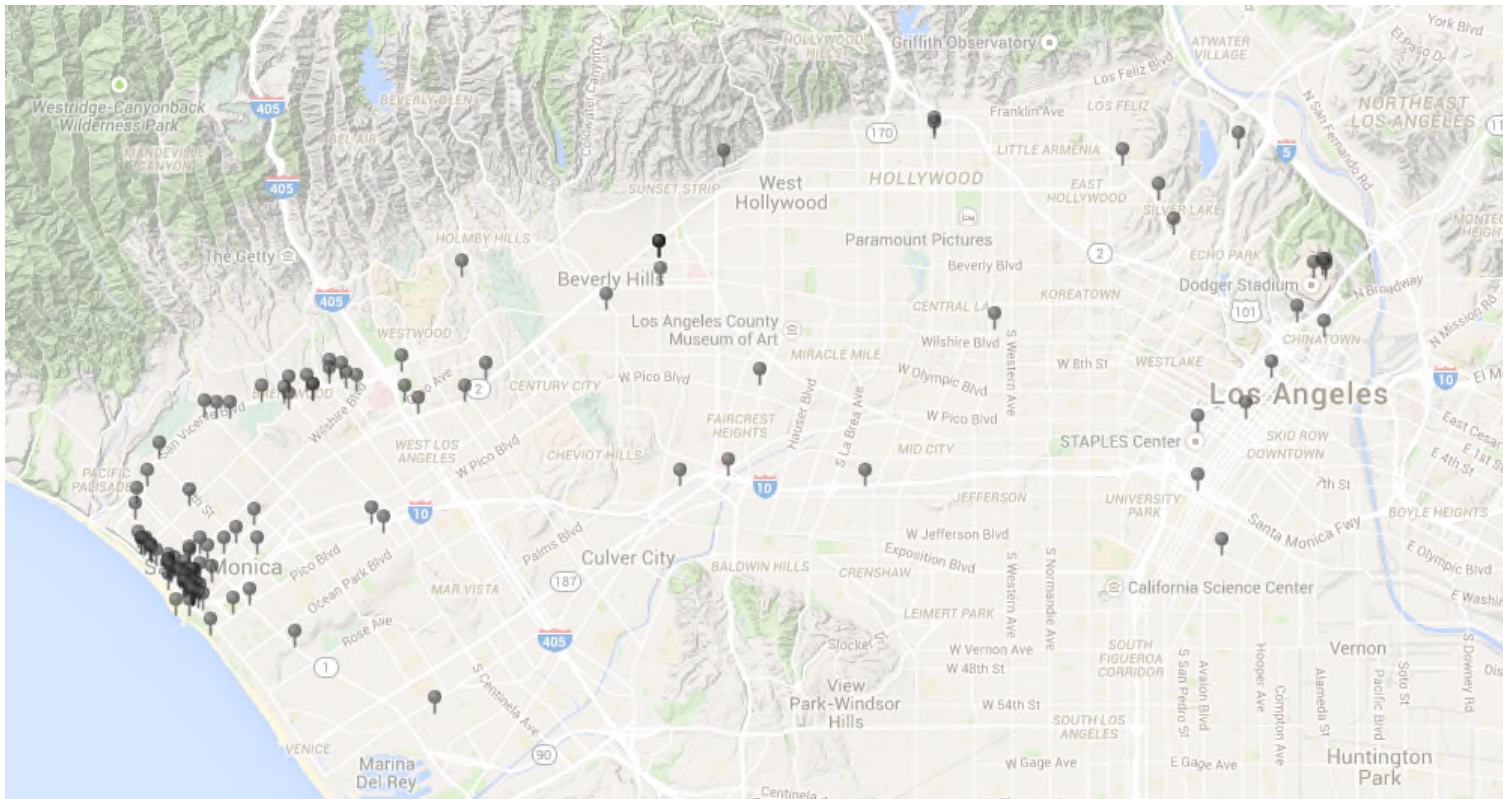
Instagram Tracking

■ Tracking “LA Marathon”: Middle



Instagram Tracking

- Tracking "LA Marathon": Late Stage

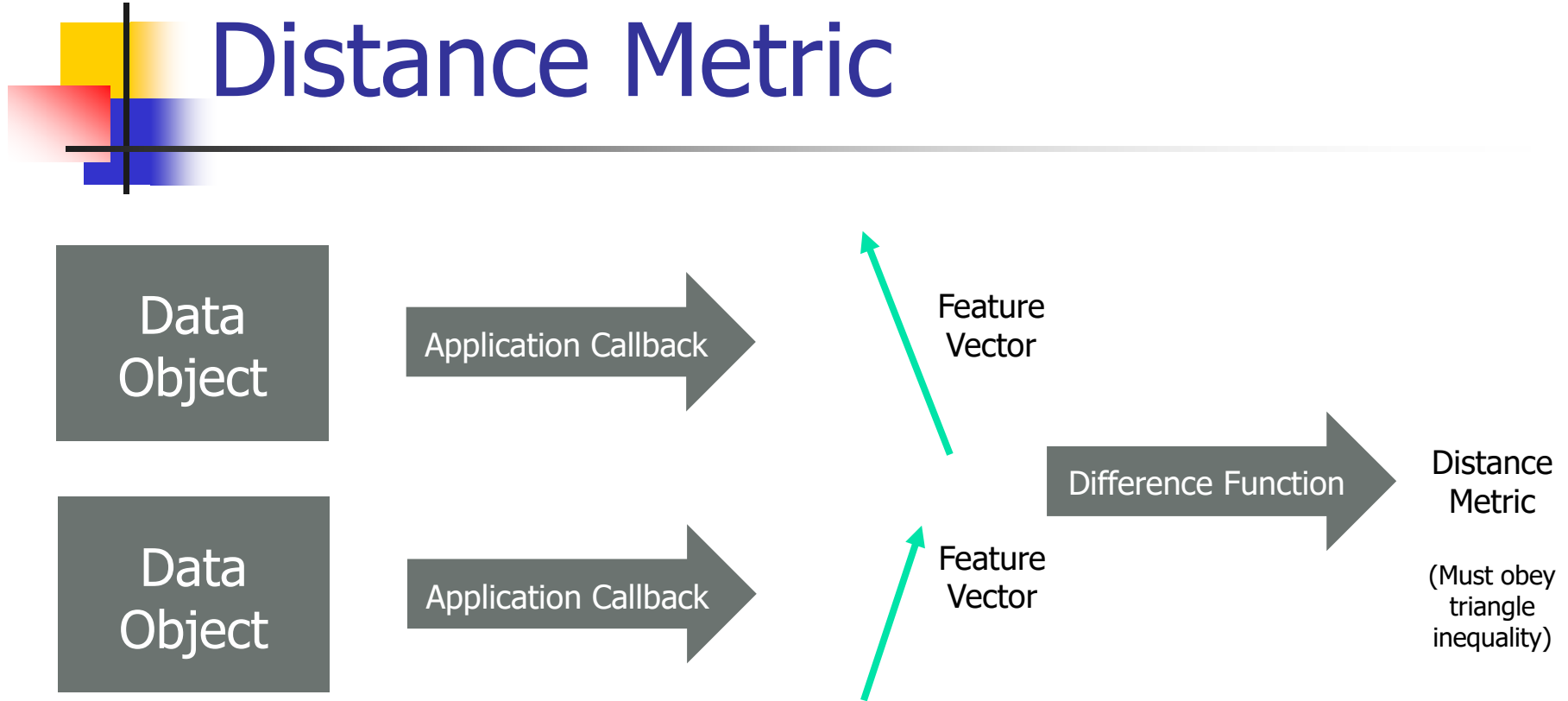


Challenge: Extractive Summarization

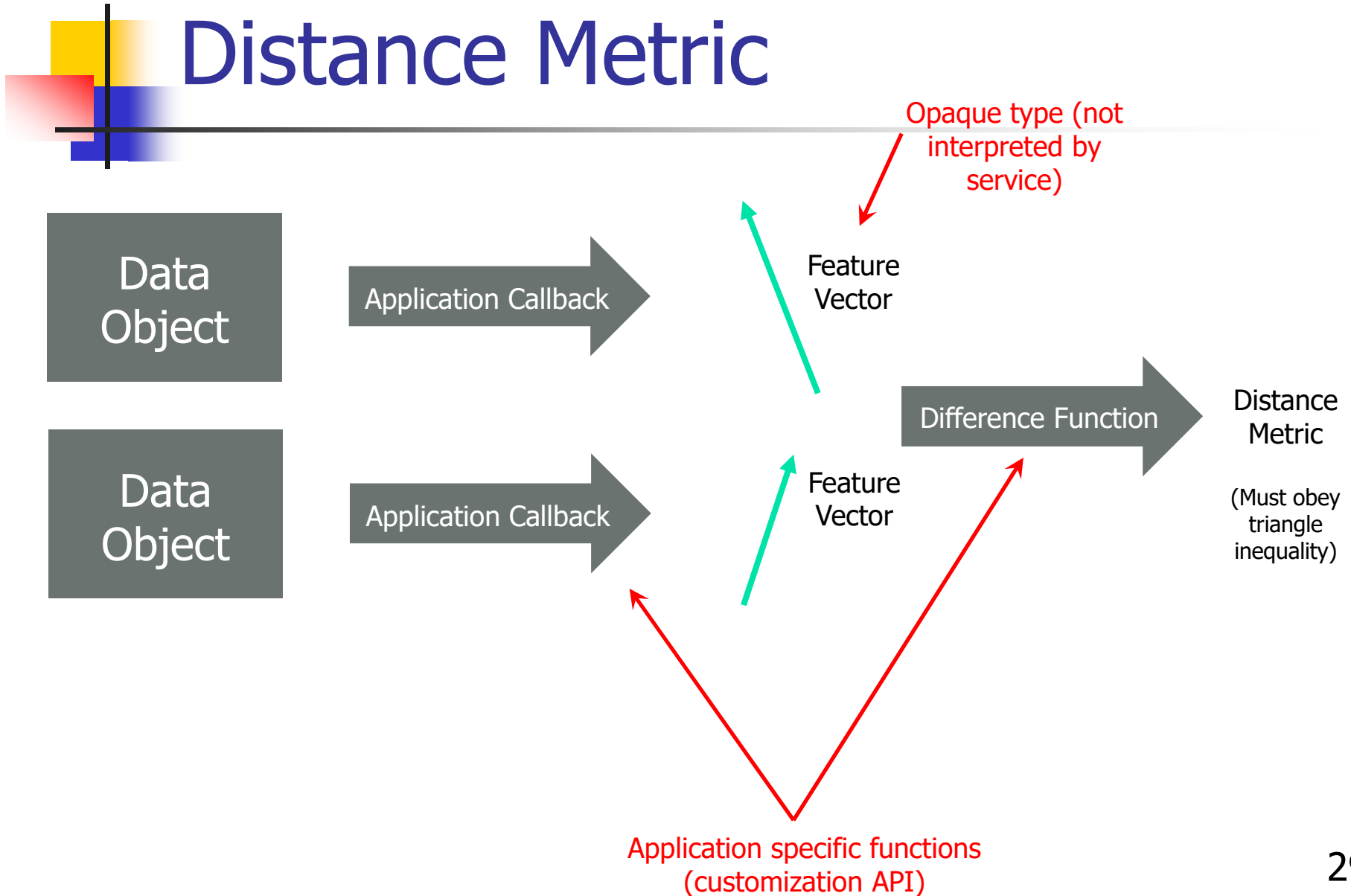


Build a data service that allows applications to retrieve (extractive) data summaries at arbitrary levels of granularity in accordance with an application-specific redundancy metric

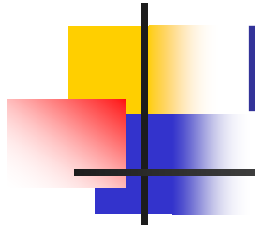
Customizability: The Distance Metric



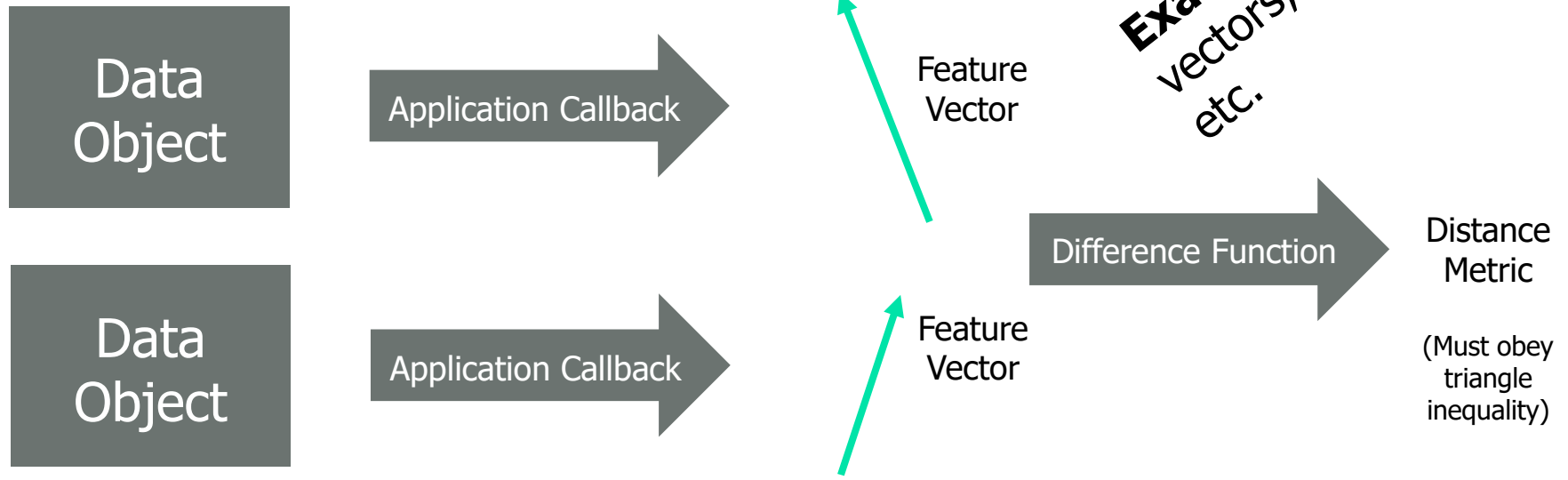
Customizability: The Distance Metric



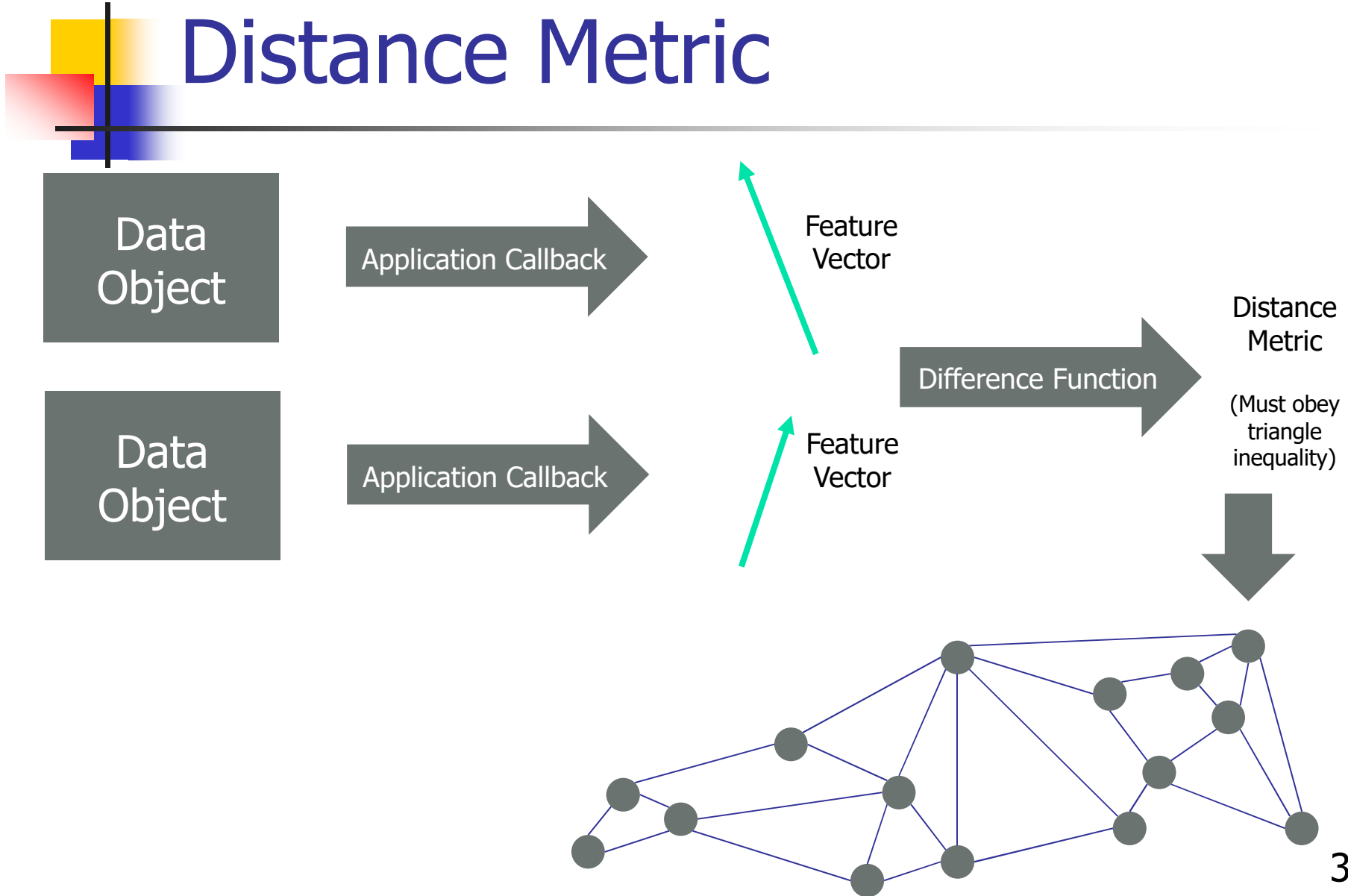
Customizability: The Distance Metric



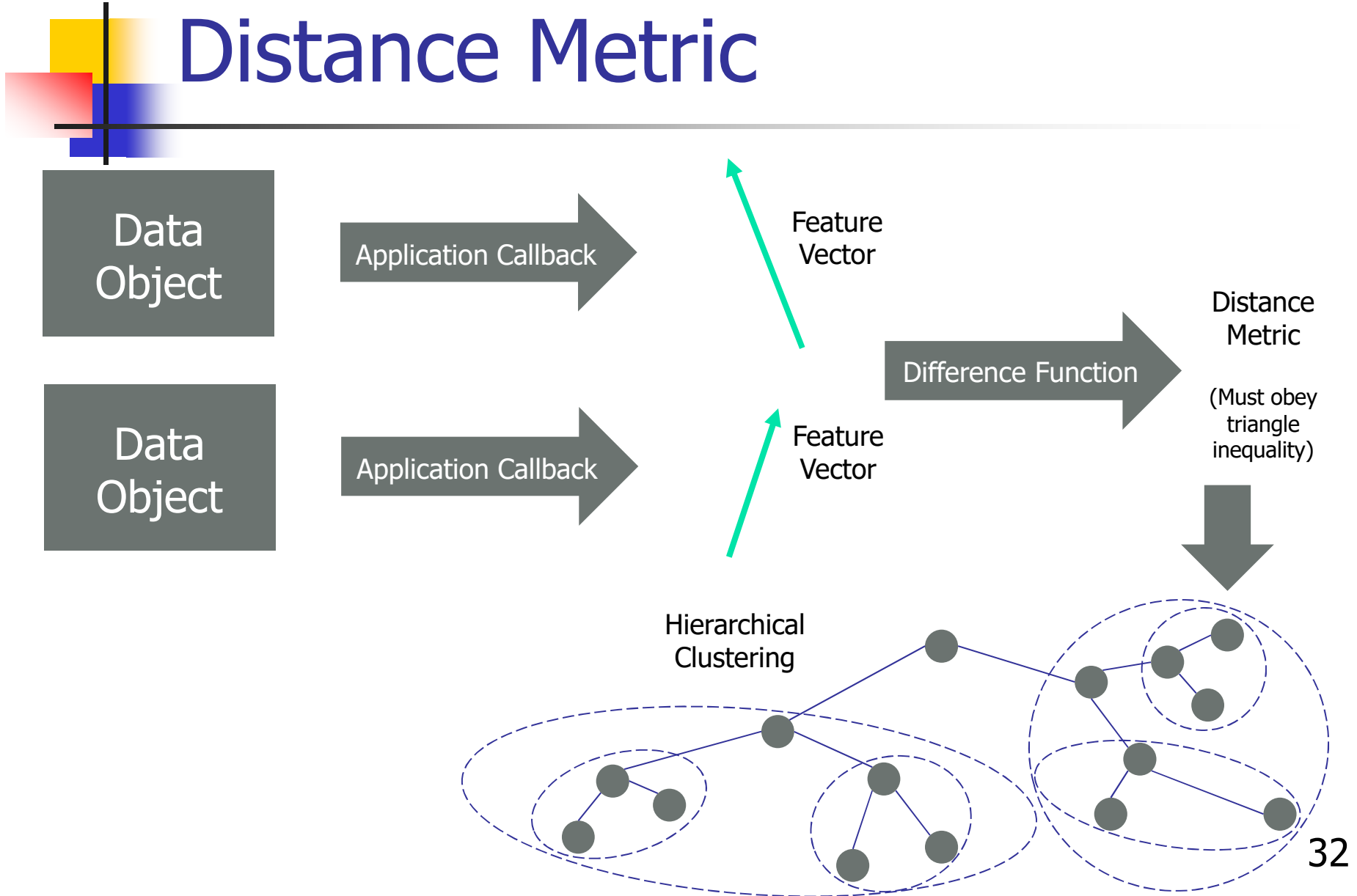
Examples: Scalars, vectors, pictures, text, etc.



Customizability: The Distance Metric



Customizability: The Distance Metric



Summarization

Data Object



Feature Vector

Data Object

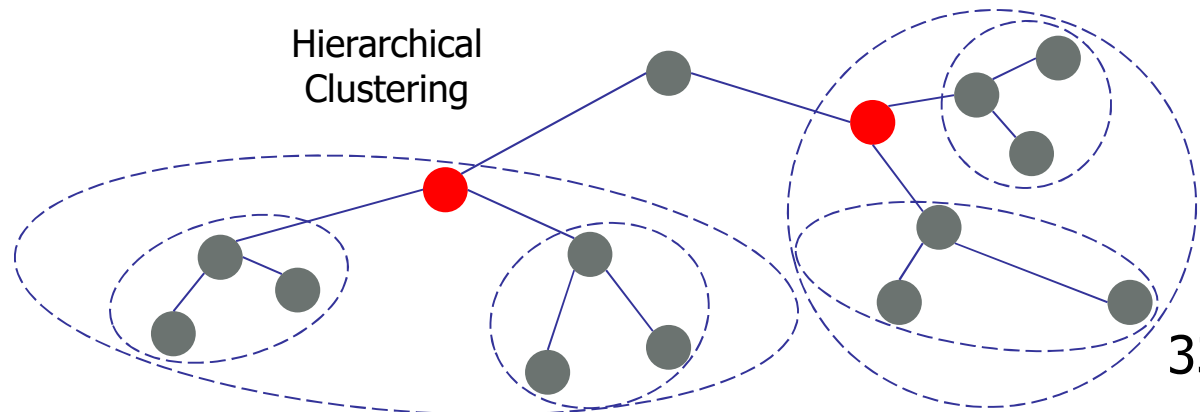


Feature Vector



Distance Metric

(Must obey triangle inequality)



Summarization

Data Object



Feature Vector

Data Object

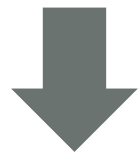


Feature Vector

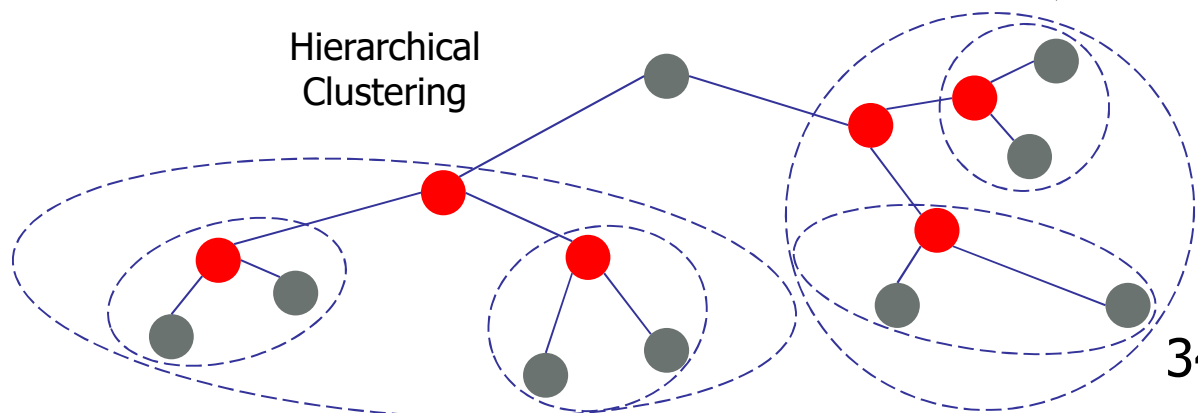


Distance Metric

(Must obey triangle inequality)

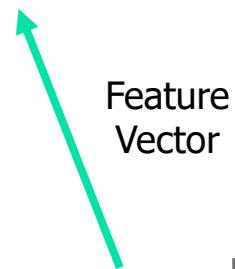


Hierarchical Clustering

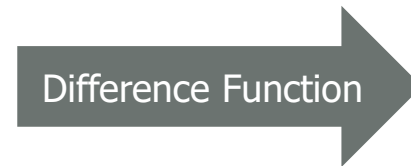
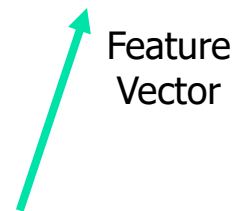


Summarization

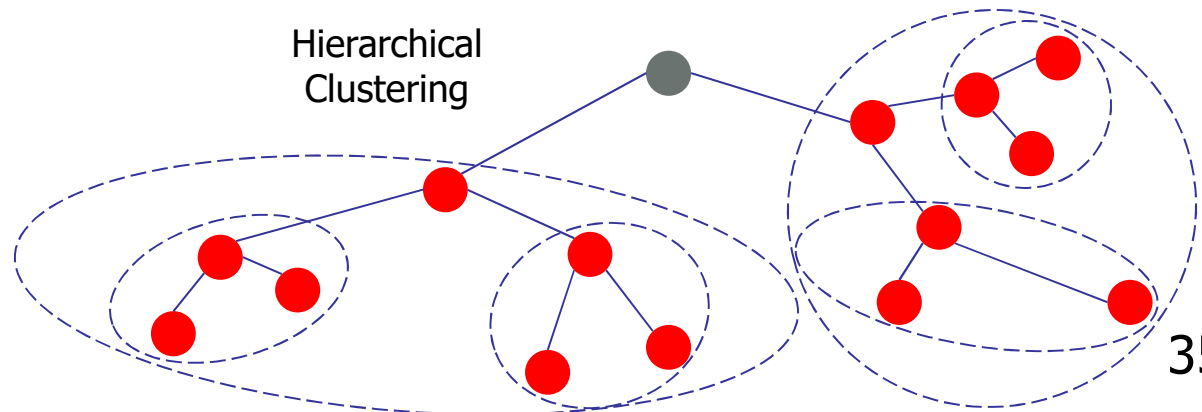
Data Object



Data Object



Distance Metric
(Must obey triangle inequality)



Summarization

Data Object



Feature Vector

Data Object



Feature Vector



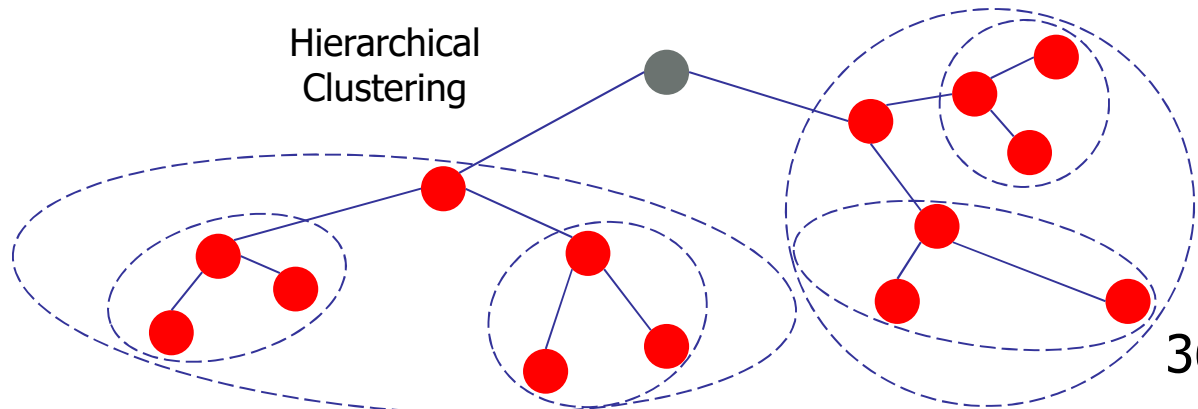
Distance Metric

(Must obey triangle inequality)



Representative sampling versus noise reduction?

Hierarchical Clustering



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

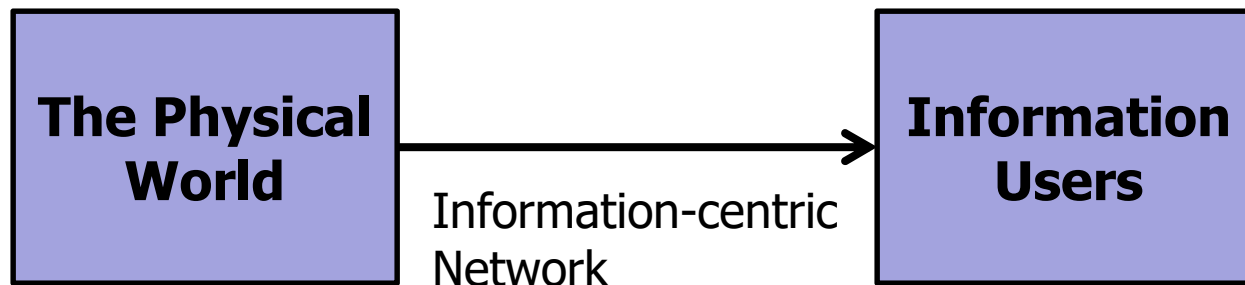
Goal:

Information Distillation

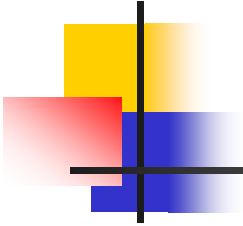
- Maximizes *information flow*
- Much data is "physical"
- Protocols geared for data filtering, and aggregation
- Data loss may be a feature intended to reduce less informative bits

A Primary Network Design Challenge

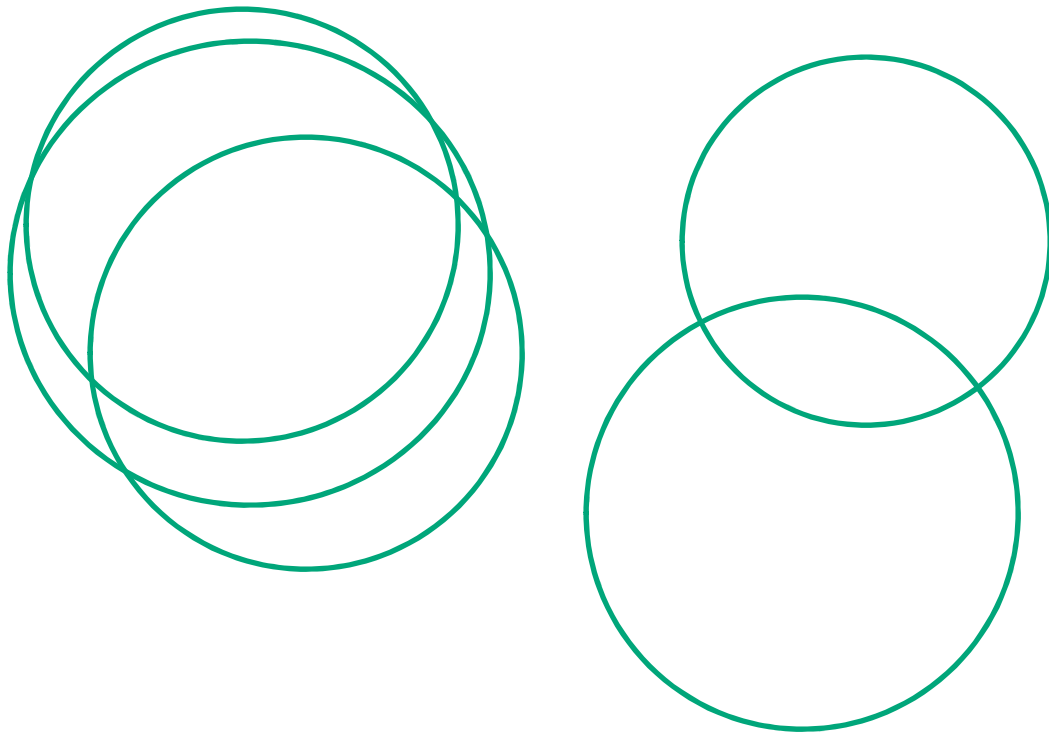
- How to build networks that *maximize useful information flow from the physical world?*



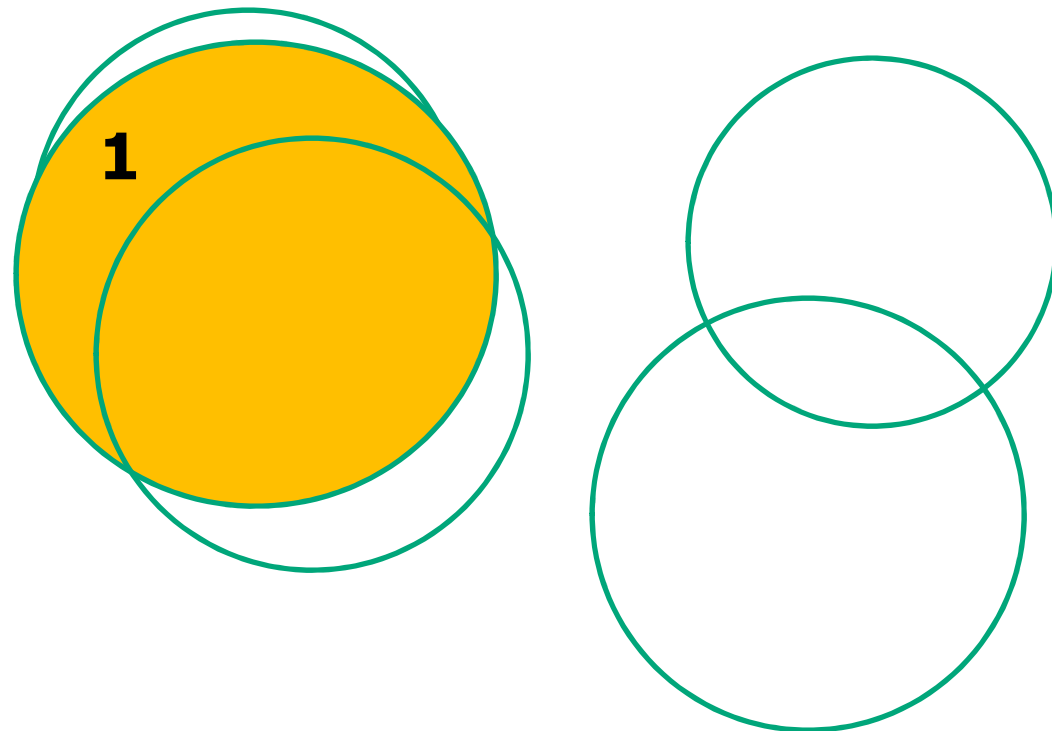
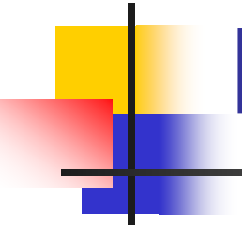
Information-maximizing Prioritization



- Determine transmission order?

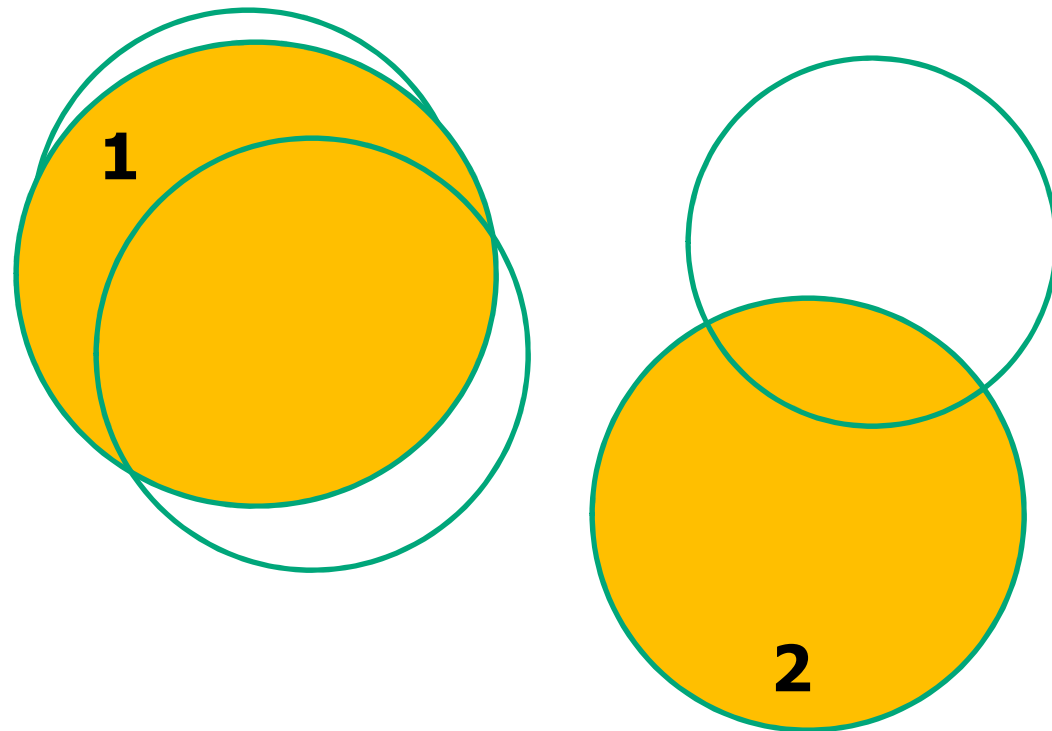
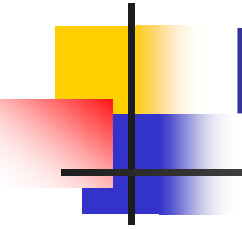


Information-maximizing Prioritization



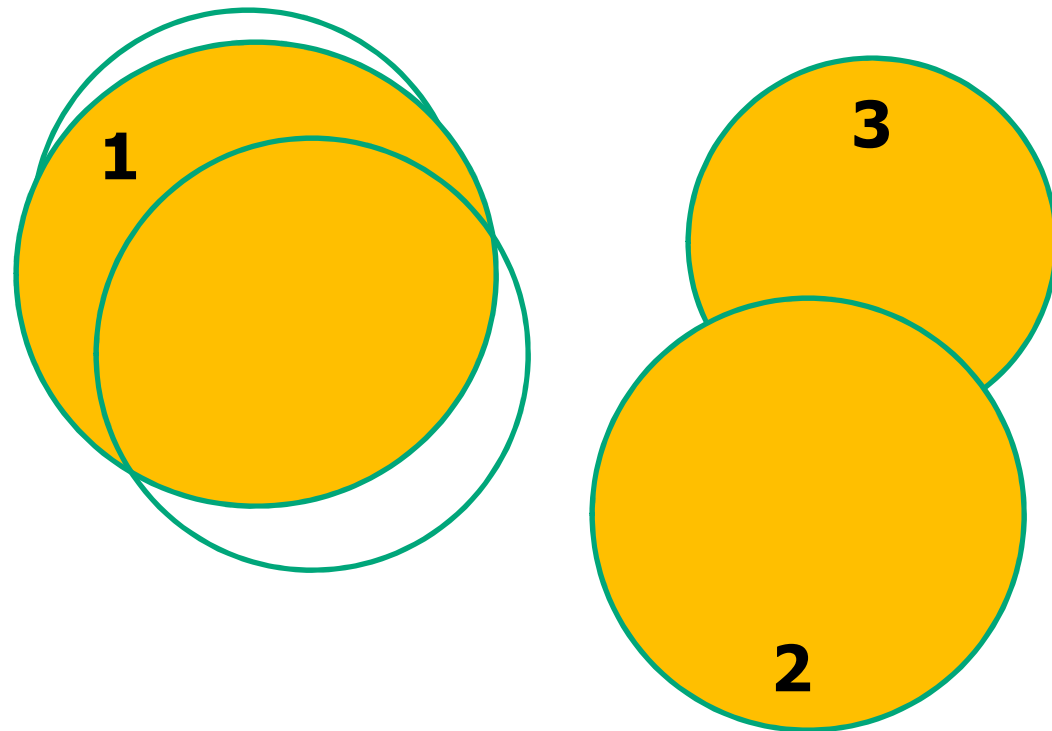
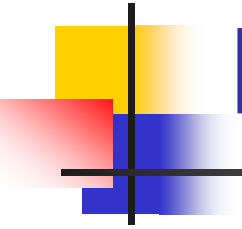
- Determine transmission order?

Information-maximizing Prioritization



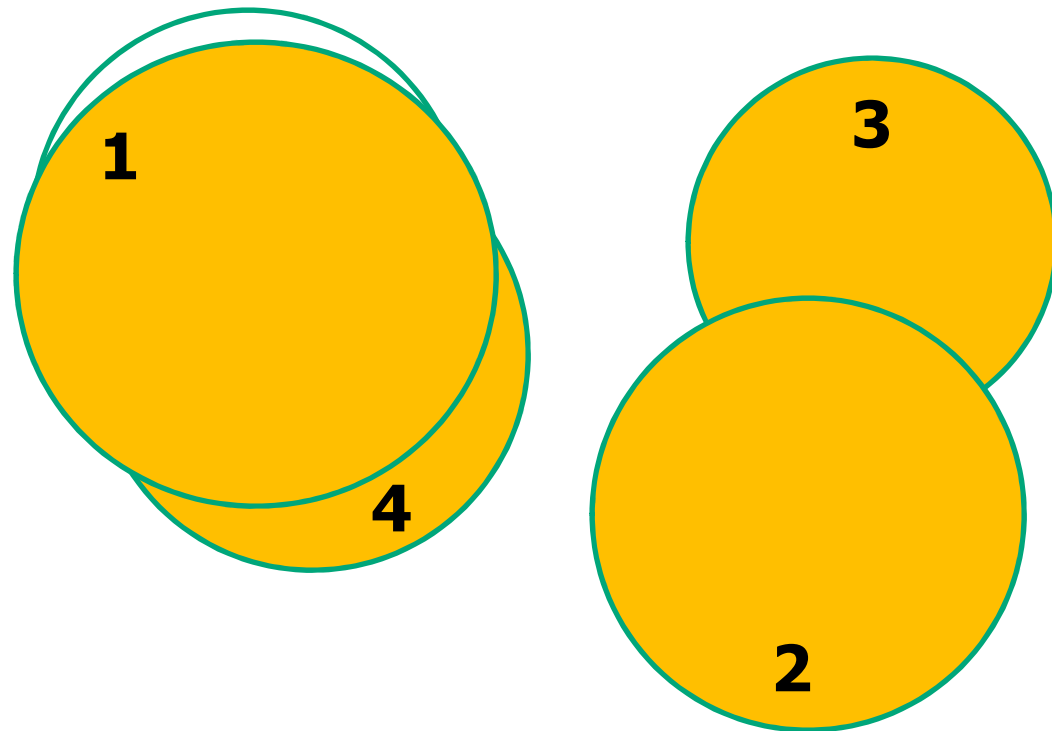
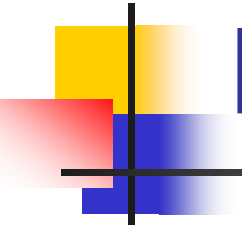
- Determine transmission order?

Information-maximizing Prioritization



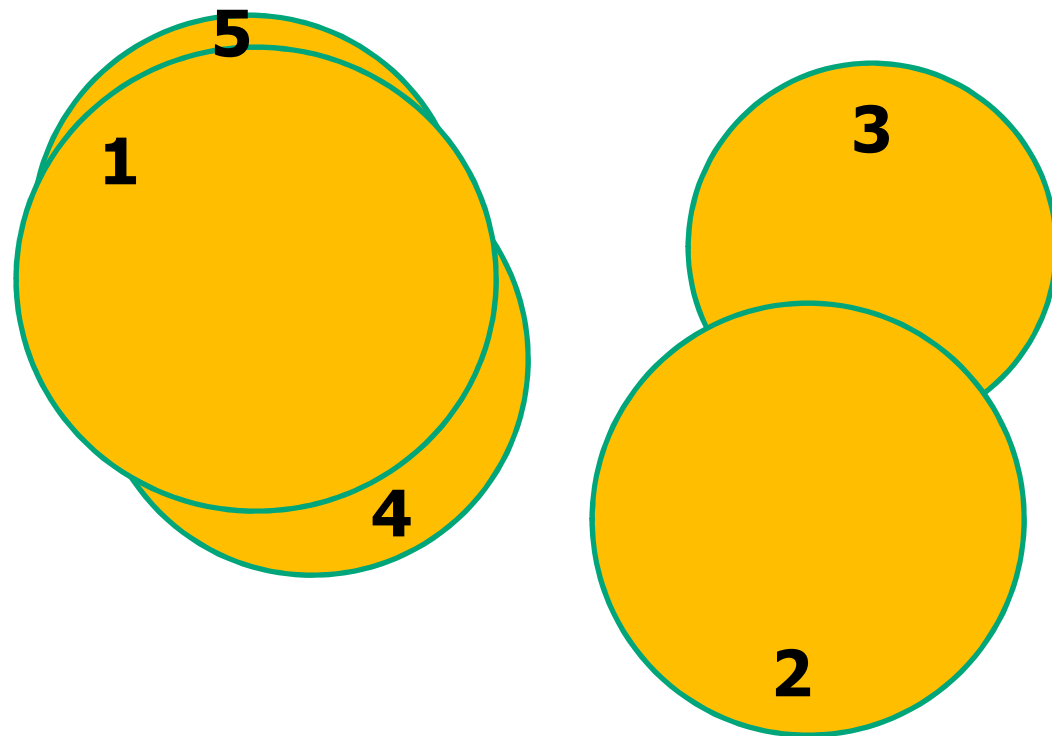
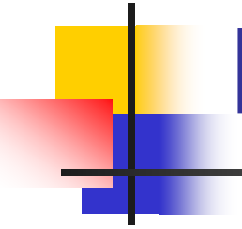
- Determine transmission order?

Information-maximizing Prioritization



- Determine transmission order?

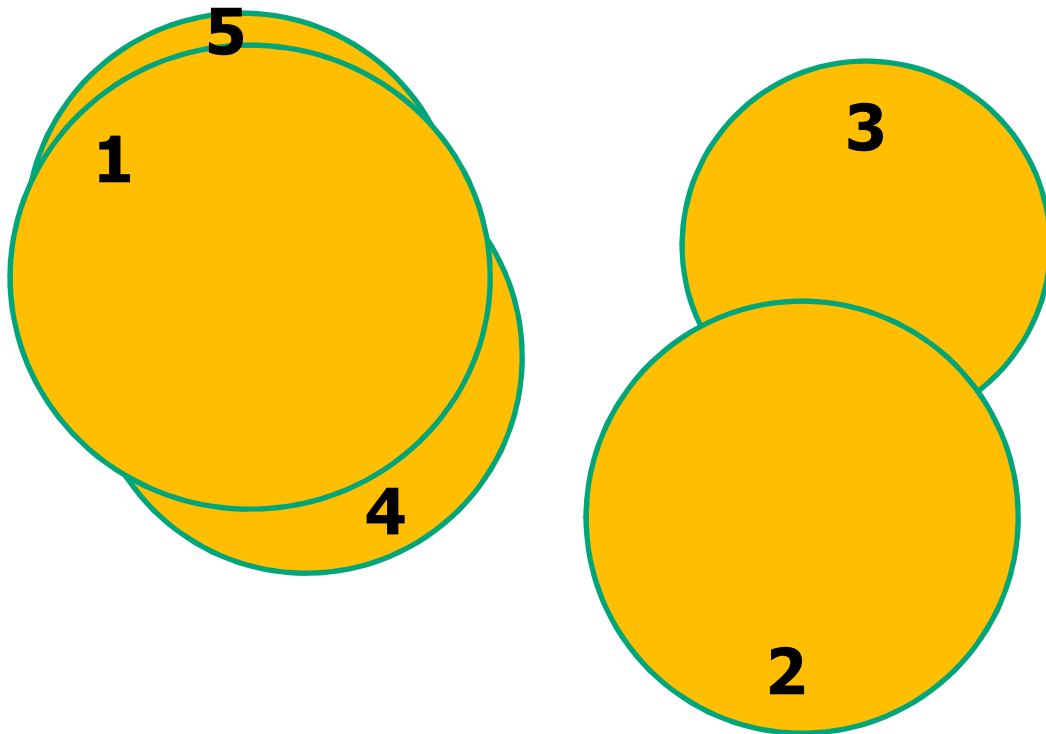
Information-maximizing Prioritization



- Determine transmission order?

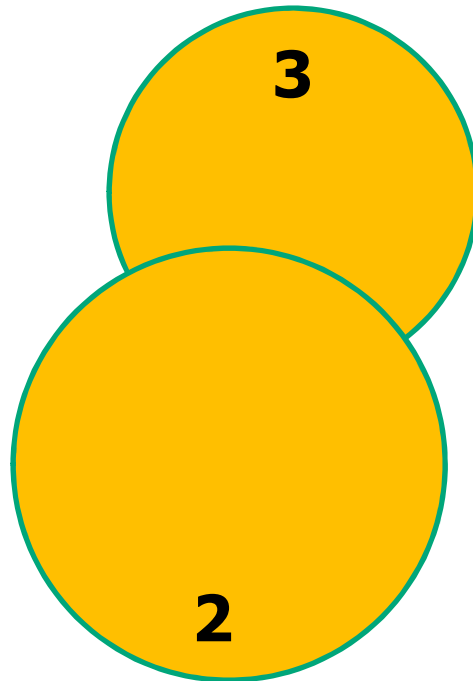
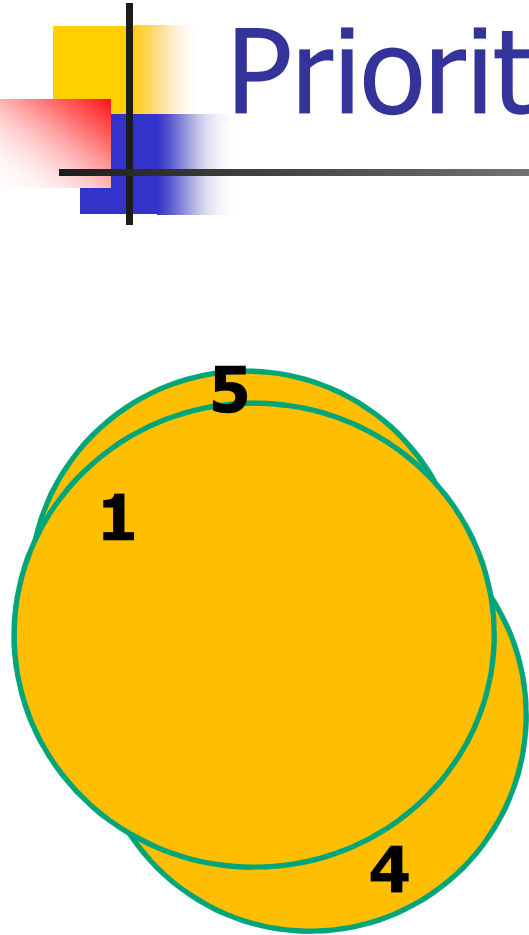
Information-maximizing Prioritization

- 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...

Images are collected from the Internet



Hurricane Katrina 2005



Nepal earthquake 2015



Thailand flood 2011

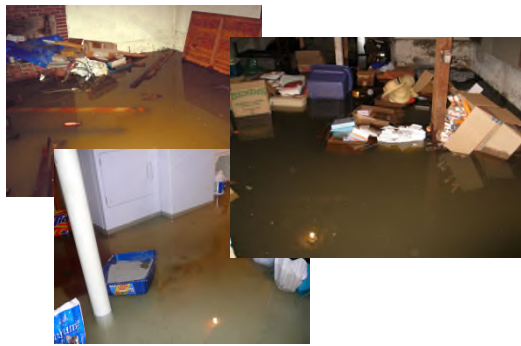
- Volunteers are recruited
- They scout the area, capture pictures and send them to a rescue center
- Network constraints prevent sending all pictures 47

Challenge: Data Selection to Maximize Coverage

Fire on 6th and Main.

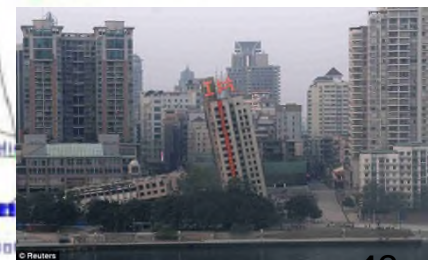


Collapse on Park Ave.



Flooding on State St.

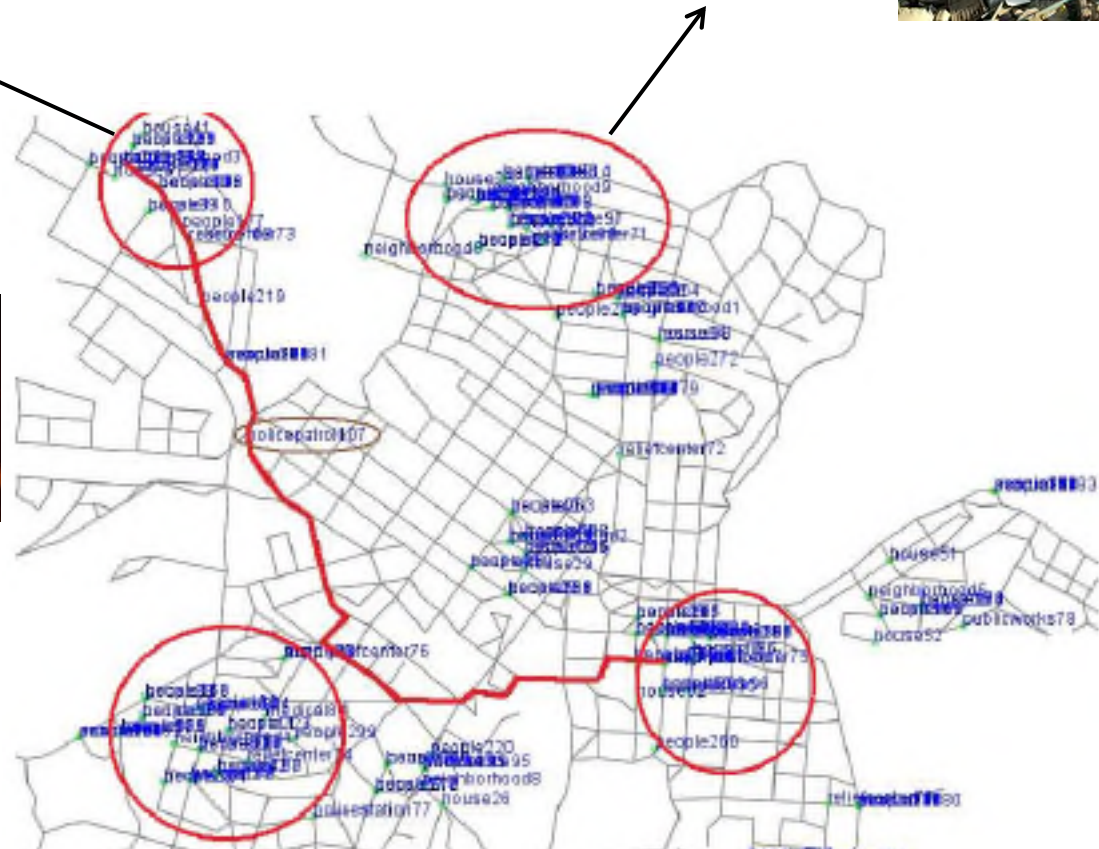
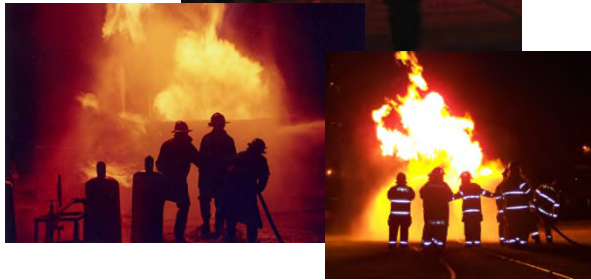
Structural damage on Pier Square



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