Big Data in Climate:
Opportunities and Challenges for Machine Learning and Data Mining

Vipin Kumar

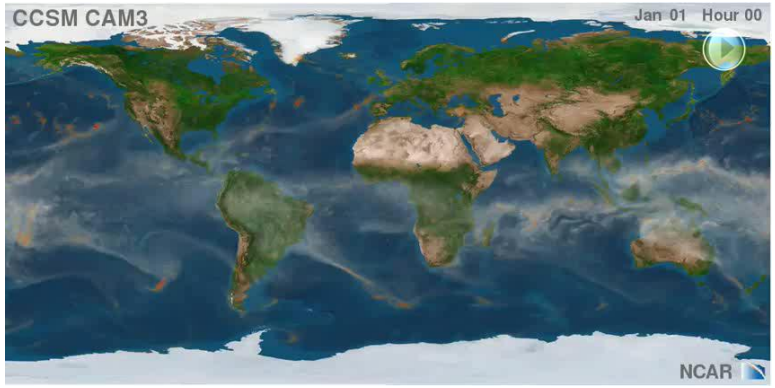
University of Minnesota

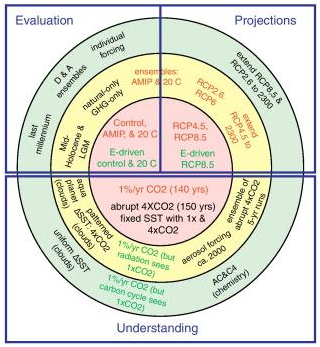
kumar001@umn.edu
www.cs.umn.edu/~kumar

Big Data in Climate

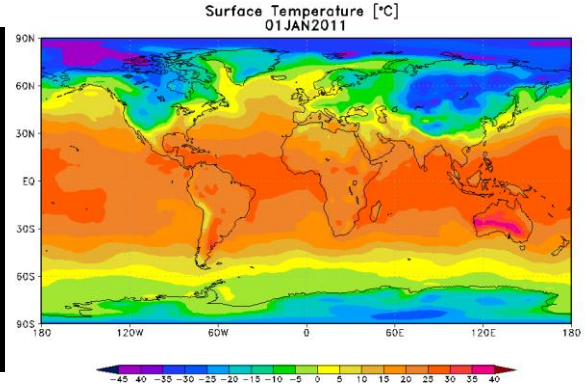
- Satellite Data
  - Spectral Reflectance
  - Elevation Models
  - Nighttime Lights
  - Aerosols
- Oceanographic Data
  - Temperature
  - Salinity
  - Circulation
- Climate Models
- Reanalysis Data
- River Discharge
- Agricultural Statistics
- Population Data
- Air Quality
- …

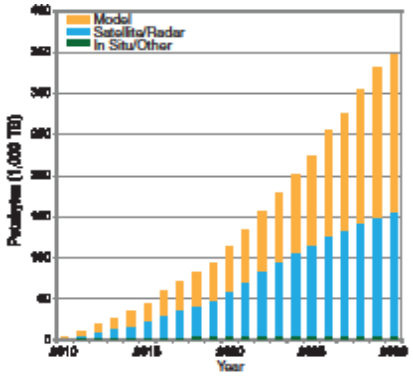
Source: NCAR





Source: NASA





Big Data in Climate

- Satellite Data
  - Spectral Reflectance
  - Elevation Models
  - Nighttime Lights
  - Aerosols
- Oceanographic Data
  - Temperature
  - Salinity
  - Circulation
- Climate Models
- Reanalysis Data

> "Climate change research is now 'big science,' comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics."
> **(Nature Climate Change, Oct 2012)**

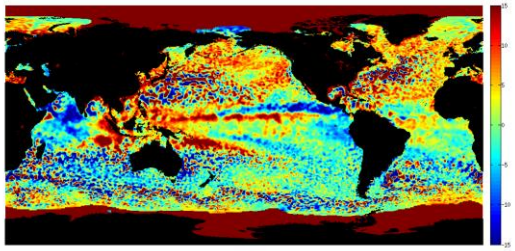
Source: NCAR

Source: NASA

Five Year, $ 10m NSF Expeditions in Computing Project (1029711, PI: Vipin Kumar, U. Minnesota)

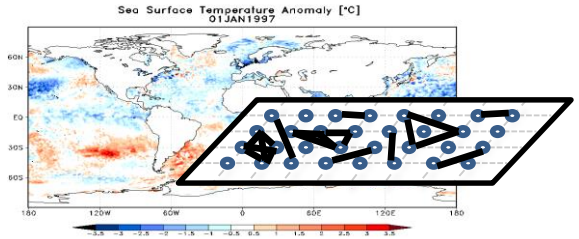
Understanding Climate Change: A Data-driven Approach

Research Highlights



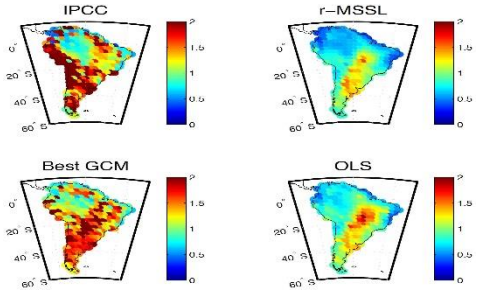
Pattern Mining:
Monitoring Ocean Eddies

- Spatio-temporal pattern mining using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



Network Analysis:
Climate Teleconnections

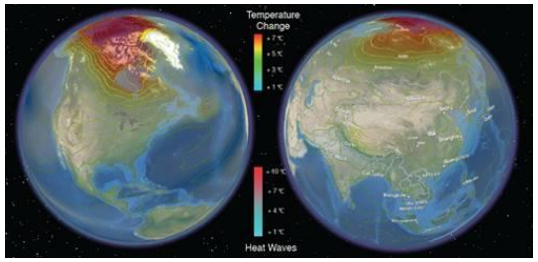
- Scalable method for discovering related graph regions
- Discovery of novel climate teleconnections
- Also applicable in analyzing brain fMRI data



Sparse Predictive Modeling:
Precipitation Downscaling

- Hierarchical sparse regression and multi-task learning with spatial smoothing
- Regional climate predictions from global observations

[Arindam Banerjee]



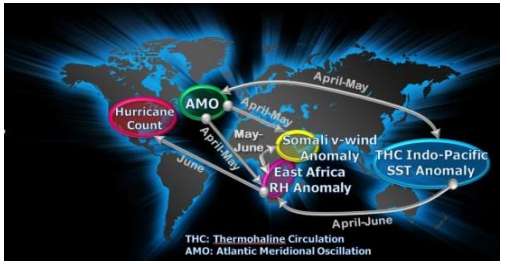
Extremes and Uncertainty:
Heat waves, heavy rainfall

- Extreme value theory in space-time and dependence of extremes on covariates
- Spatiotemporal trends in extremes and physics-guided uncertainty quantification

[Auroop Ganguly]

Change Detection:
Monitoring Ecosystem Distrubances

- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in surface water and vegetation, e.g. fires and deforestation.



Relationship mining:
Seasonal hurricane activity

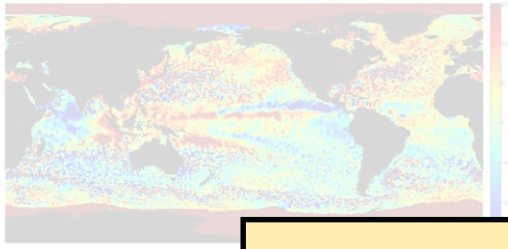
- Statistical method for automatic inference of modulating networks
- Discovery of key factors and mechanisms modulating hurricane variability

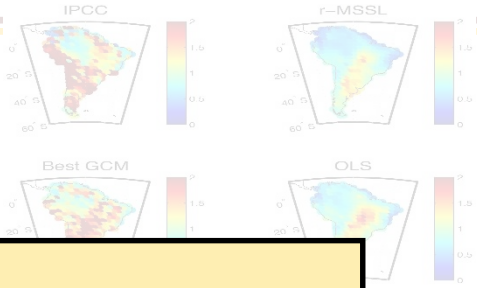
[Nagiza Samatova]

http://climatechange.cs.umn.edu/

Understanding Climate Change: A Data-driven Approach

Research Highlights





Pattern Mining:
Monitoring Ocean

- Spatio-temporal patte
  multiple object trackir
- Created an open sour
  eddies and eddy track

Challenges

- Multi-resolution, multi-scale data
- High temporal variability
- Spatio-temporal auto-correlation
- Spatial and temporal heterogeneity
- Large amount of noise and missing values
- Lack of representative ground truth
- Class imbalance (changes are rare events)

ng:
g

nd multi-task

m global

Extremes and U
Heat waves, heavy rainfall

- Extreme value theory in space-time and dependence of extremes on covariates
- Spatiotemporal trends in extremes and physics-guided uncertainty quantification

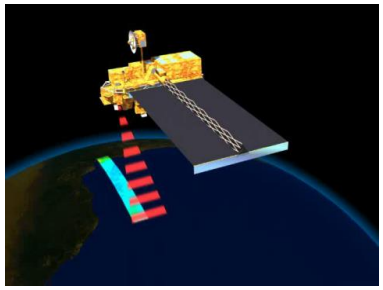
Monitoring Ecosystem Distrubances

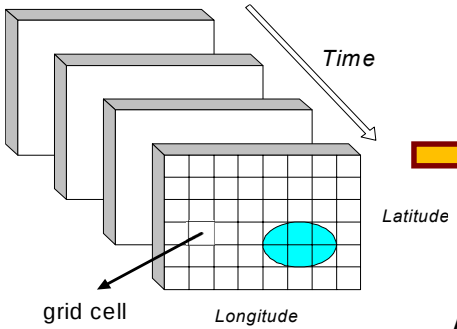
- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in surface water and vegetation, e.g. fires and deforestation.

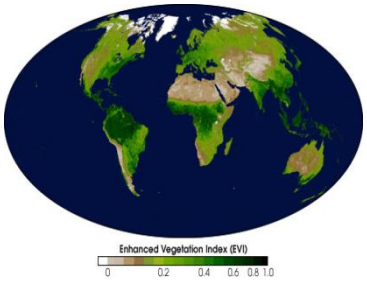
Seasonal hurricane activity

- Statistical method for automatic inference of modulating networks
- Discovery of key factors and mechanisms modulating hurricane variability

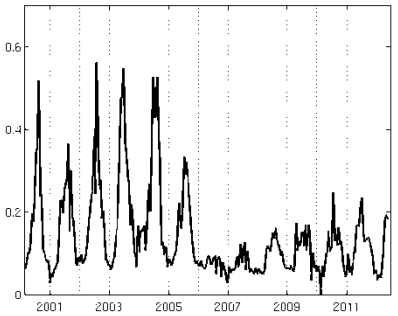
Big Data in Earth System Monitoring







A **vegetation index** measures th surface "greenness" – proxy for total biomass



This vegetation **time series** captures temporal dynamics around the site of the China National Convention Center

MODIS covers ~ 5 billion locations globally at 250m resolution daily since Feb 2000.

Data	Type	Coverage	Spatial Resolution	Temporal Resolution	Spectral Resolution	Duration	Availability
MODIS	Multispectral	Global	250 m	Daily	7	2000 - present	Public
LANDSAT	Multispectral	Global	30 m	16 days	7	1972 - present	Public
Hyperion	Hyperspectral	Regional	30 m	16 days	220	2001 - present	Private
Sentinal - 1	Radar	Global	5 m	12 days	-	2014 - present	Public
Quickbird	Multispectral	Global	2.16 m	2 to 12 days	4	2001 - 2014	Private
WorldView - 1	Panchromatic	Global	50 cm	6 days	1	2007 - present	Private

Monitoring Global Change: Case Studies

1. Global mapping of forest fires:
   - RAPT: Rare Class Prediction in Absence of Ground Truth

2. Mapping of plantation dynamics in tropical forests:
   - Handling Heterogeneity in Space, Time and Target Class

3. Global mapping of inland surface water dynamics
   - Heterogeneous Ensemble Learning and Physics-guided Labeling

Lake Oroville near the Bidwell Marina in 2011 and 2014

Case Study 1: Global Forest Fires Mapping

Monitoring fires is important for climate change impact

A record number of more than 150 countries signed the landmark agreement to tackle climate change at a ceremony at UN headquarters on 22 April, 2016.

"the best chance to save the one planet we have"


SEARCH

The New York Times

ENVIRONMENT

Delegates at Climate Talks Focus on Saving the World's Forests

By JUSTIN GILLIS DEC. 10, 2015

The canopy of the forest in Puerto Viejo, Costa Rica, in October 2014. Climate change negotiations in Paris could lead to a sweeping effort to save the world's forests. Adriana Zehbrauskas for The New York Times

State-of-the-art: NASA MCD64A1

- Most extensively used global fire monitoring product
- Uses MODIS surface reflectance and Active Fire data in a predictive model
- Performance varies considerably across different geographical regions
- Known to have very low recall in tropical forests that play a critical role in regulating the Earth's climate, maintaining biodiversity, and serving as carbon sinks

Predictive Modeling:
Traditional Paradigm

Given a feature vector $\boldsymbol{x} \in \mathbf{R}^d$
predict the class label $y \in \{0, 1\}$

Learn a classification function

$$f : \mathbf{R}^d \rightarrow \mathcal{Y}$$

which generalizes well on unseen data that comes from the same distribution as training data.

8000 sq.Km scene in SE Asia(2005)

Burned area mapping

Predicts whether a given pixel is burned or not?

Explanatory Variable $\boldsymbol{x}_i \in \mathbf{R}^d$	**Target Label** $y_i \in \mathcal{Y} = \{0, 1\}$
$\boldsymbol{x}_1$	1
$\boldsymbol{x}_2$	0
$\boldsymbol{x}_3$	0
$\boldsymbol{x}_4$	1
.	.
$\boldsymbol{x}_N$	1

Predictive Modeling for Global Monitoring of Forest Fires

Challenges:

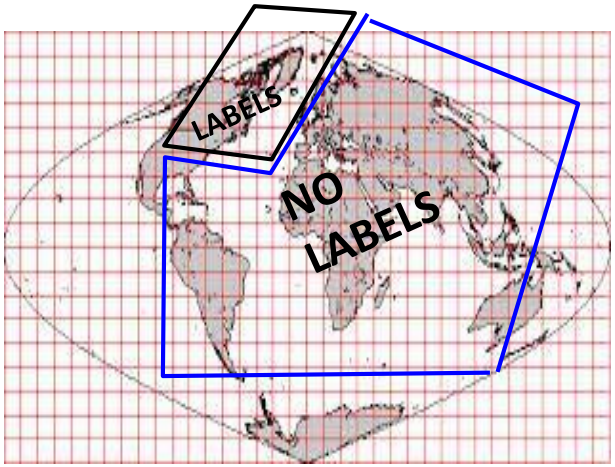
(1) Complete absence of target labels for supervision

(however, imperfect annotations of poor quality labels are available for every sample)

Variations in the relationship between the explanatory and target variable

- Geographical heterogeneity
- Seasonal heterogeneity
- Land class heterogeneity
- Temporal heterogeneity

$\boldsymbol{x}_i \in \mathbf{R}^d$	$y_i \in \mathcal{Y} = \{0, 1\}$
$\boldsymbol{x}_1$	?
$\boldsymbol{x}_2$	?
$\boldsymbol{x}_3$	?



Global availability of labeled samples for burned area classification

Predictive Modeling for Global Monitoring of Forest Fires

Challenges:

(1) Complete absence of target labels for supervision

(however, imperfect annotations of poor quality labels are available for every sample)

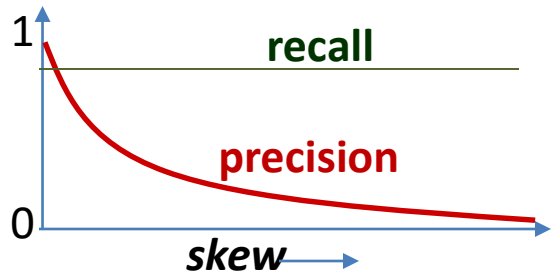
(2) Highly imbalanced classes

For eg. **California State**

Year 2008 (experienced maximum fire activity in last decade)

2,296 sq. km. of forests burned
out of a total
73,702 sq. km. forested area

True Positive Rate = 0.9
False Positive Rate = 0.01



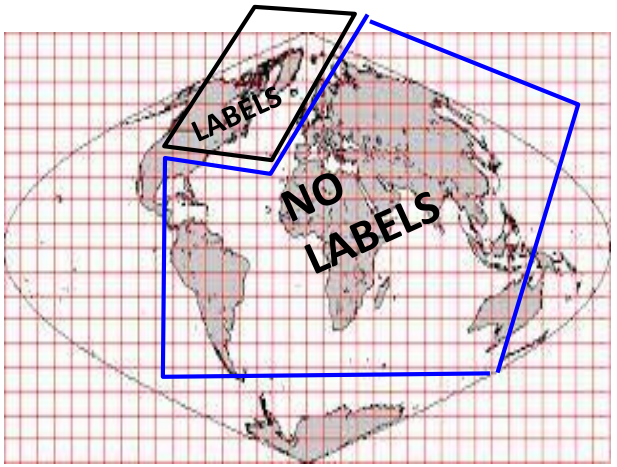
Predictive Modeling for Global Monitoring of Forest Fires

Challenges:

(1) Complete absence of target labels for supervision
(however, imperfect annotations of poor quality labels are available for every sample)

(2) Highly imbalanced classes

(3) How to evaluate performance of a model using imperfect labels?

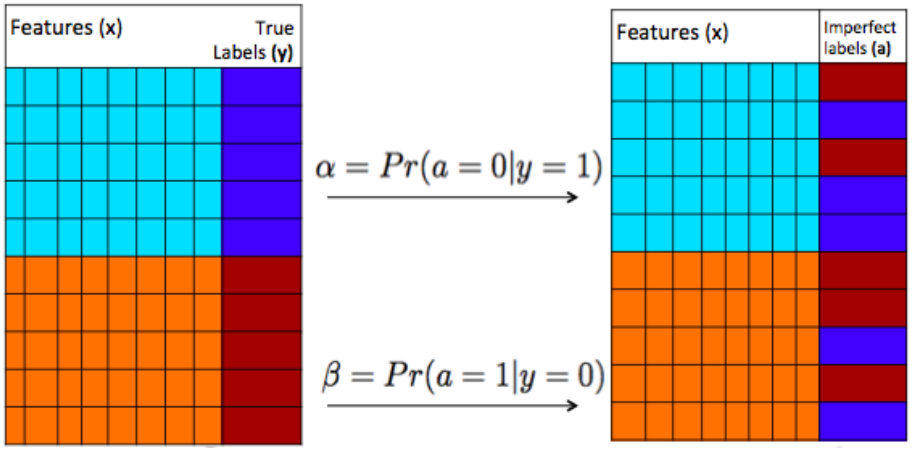


Global availability of labeled samples for burned area classification

Predictive Modeling for a Rare Target Class using Imperfect Labels

What are imperfect labels ?

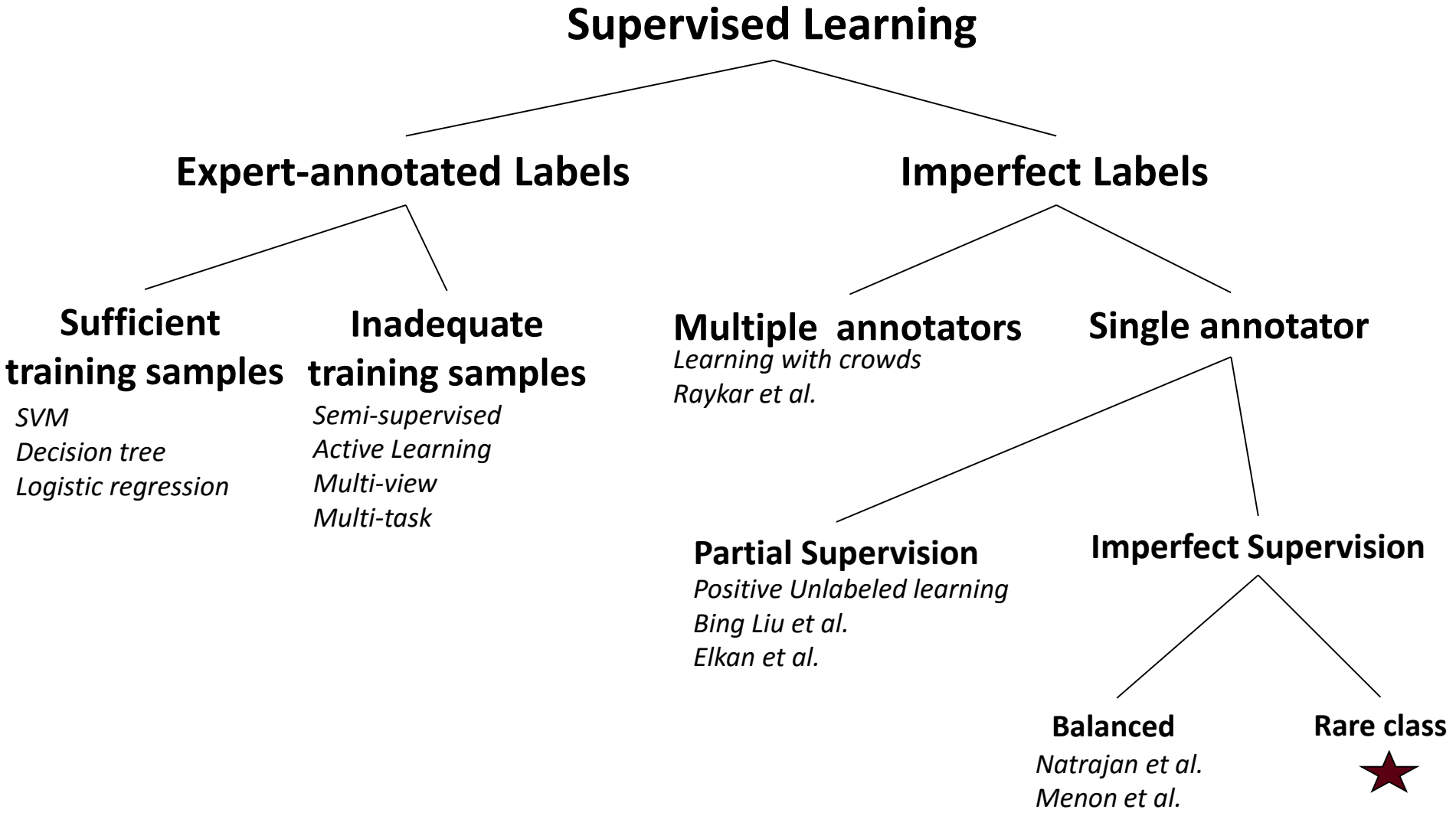
- Noisy/perturbed true labels
- Inexpensive to obtain
  - Raw feature
  - Heuristics (given by expert)
- Available for all test instances



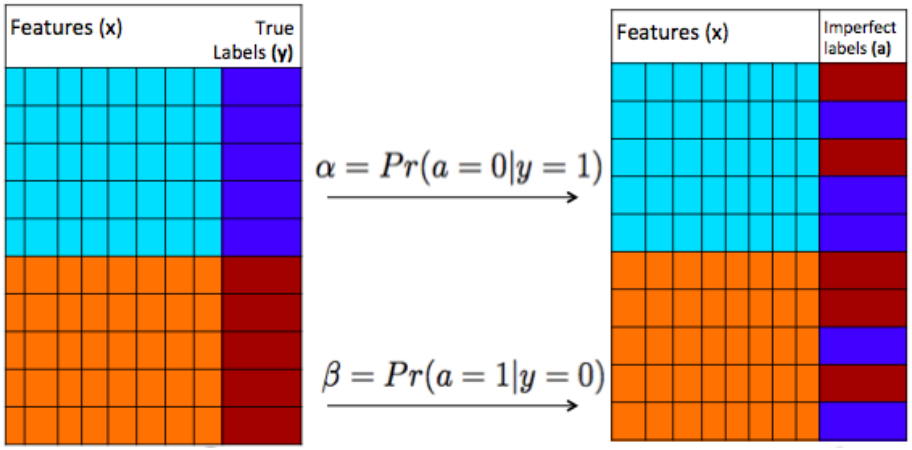
Examples of imperfect labels

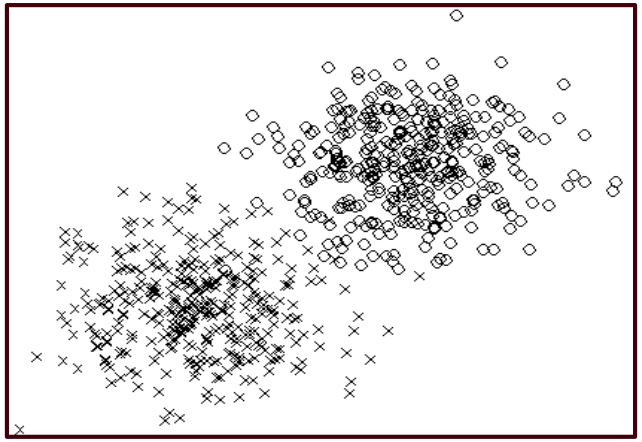
Application	Target label	Imperfect label
Burned Area	Fire/No Fire	Thermal anomaly
Urban settlement	Urban/No urban	Night time light
Recommending items to a new user	Interested/ No interested	Friends interest

Learning with imperfect labels



Learning with imperfect labels





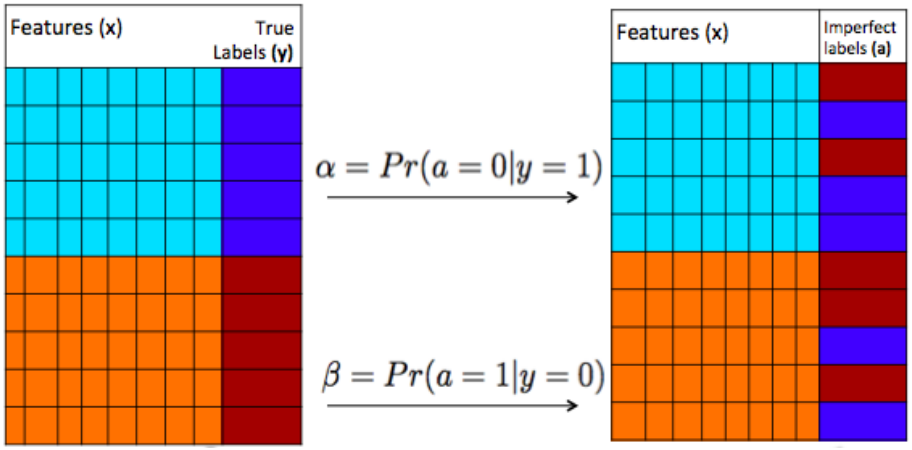
True Labels

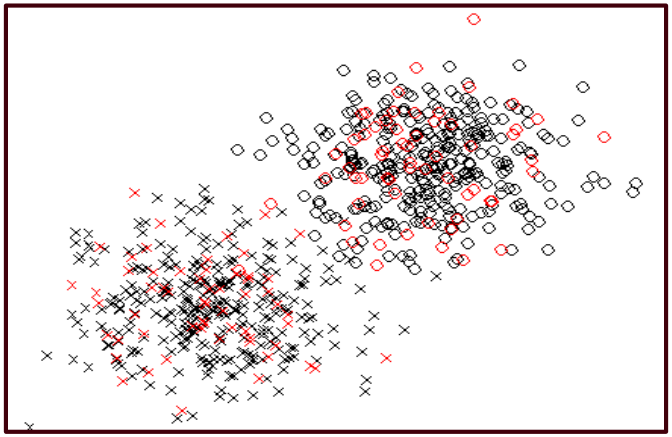
Assumptions

(1) $\alpha + \beta < 1$

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

Learning with imperfect labels





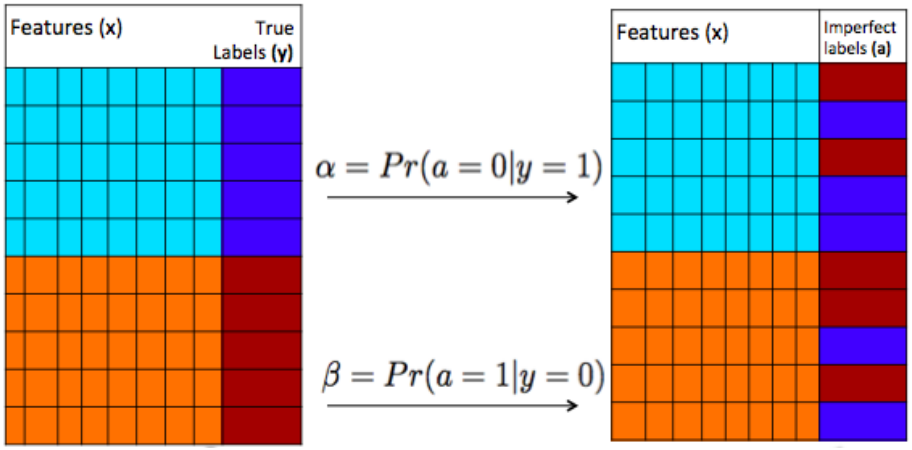
Imperfect Labels that are **conditionally independent** (Errors shown in Red)

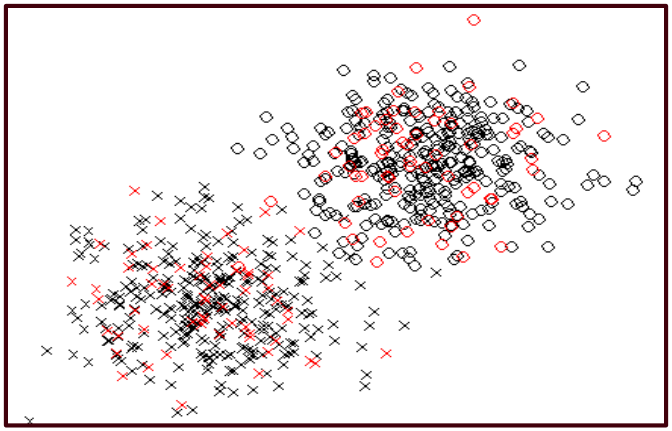
Assumptions

(1) $\alpha + \beta < 1$

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

Learning with imperfect labels



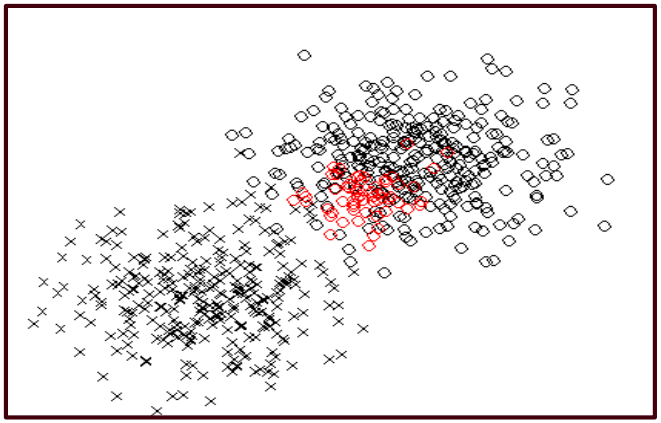


Imperfect Labels that are **conditionally independent** (Errors shown in Red)

Assumptions

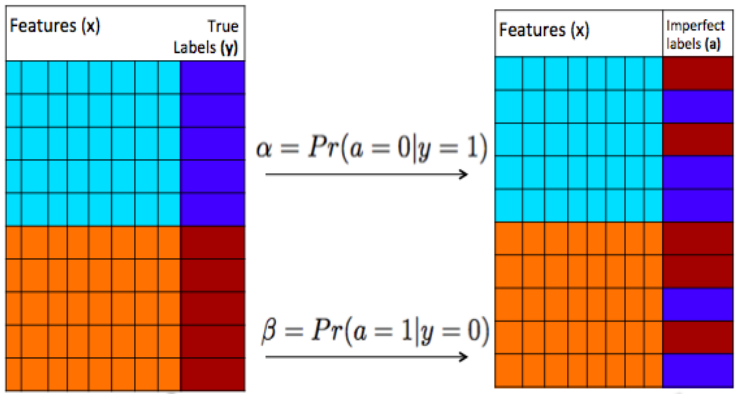
(1) $\alpha + \beta < 1$

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)



Imperfect Labels that are **not conditionally independent** (Errors shown in Red)

Learning with imperfect labels



Assumptions

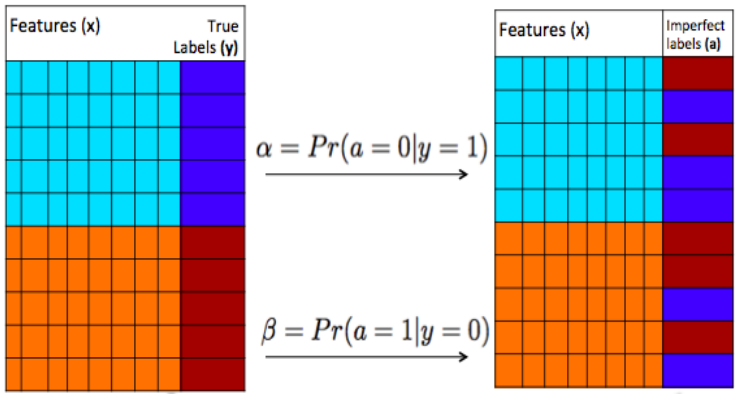
(1 $\alpha + \beta < 1$

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

Blum et.al, COLT 1998

Given enough samples, it is possible to learn a classifier in CCN label noise scenario that is as good as one learned using perfectly labeled samples.

Learning with imperfect labels



Assumptions

(1 $\alpha + \beta$ < 1

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

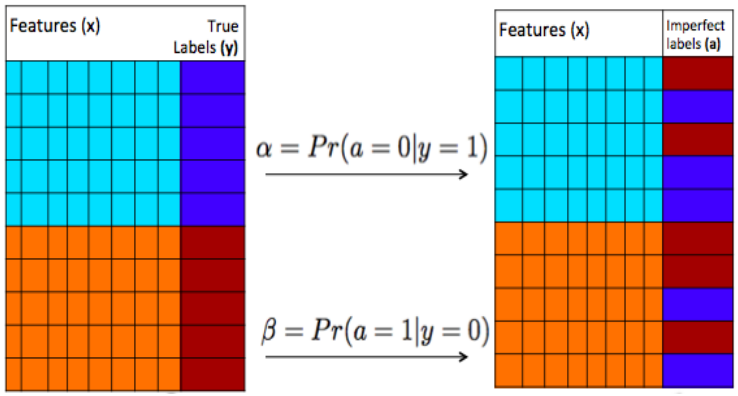
Blum et.al, COLT 1998

Given enough samples, it is possible to learn a classifier in CCN label noise scenario that is as good as one learned using perfectly labeled samples.

Natarajan et.al, NIPS 2013

For some performance measures like accuracy, a classifier can be learned using CCN labels treating them as perfect labels.

Learning with imperfect labels

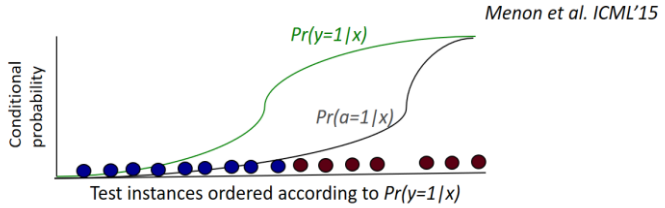


Assumptions

(1 α + β : < 1

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

Ranking according to *Pr(a=1|x)* and *Pr(y=1|x)* is identical



Blum et.al, COLT 1998

Given enough samples, it is possible to learn a classifier in CCN label noise scenario that is as good as one learned using perfectly labeled samples.

Natarajan et.al, NIPS 2013

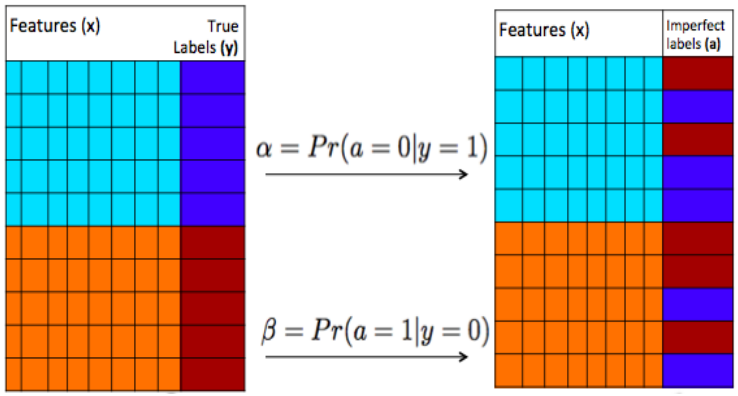
For some performance measures like accuracy, a classifier can be learned using CCN labels treating them as perfect labels.

Menon et.al, ICML 2015

-For general CCN scenario showed the linear relationship between P(**y**=1|**x**) and P(***a***=1|**x**).

- Presented a method to optimize balanced error rate, AUC

Learning with imperfect labels



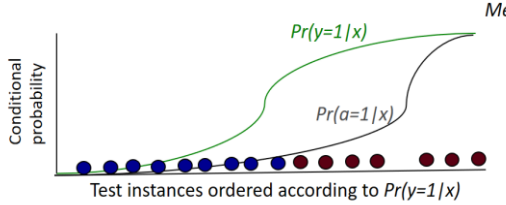
Assumptions

(1 α + β : < 1

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

Ranking according to *Pr(a=1|x)* and *Pr(y=1|x)* is identical

Menon et al. ICML'15



Blum et.al, COLT 1998

Given enough samples, it is possible to learn a classifier in CCN label noise scenario that is as good as one learned using perfectly labeled samples.

Natarajan et.al, NIPS 2013

For some performance measures like accuracy, a classifier can be learned using CCN labels treating them as perfect labels.

Menon et.al, ICML 2015

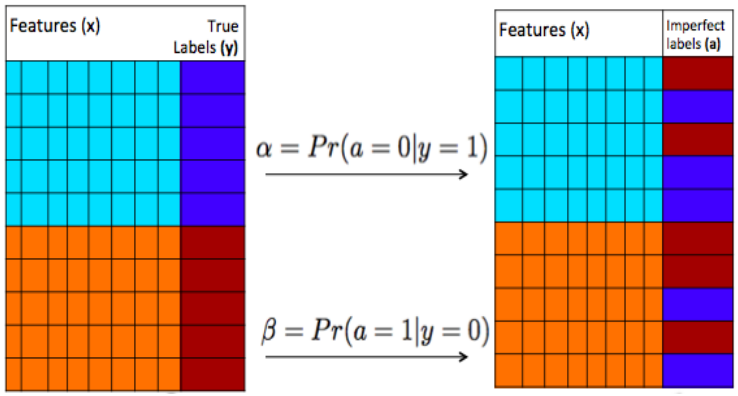
-For general CCN scenario showed the linear relationship between P(**y**=1|**x**) and P(***α***=1|**x**).

- Presented a method to optimize balanced error rate, AUC

Imperfect supervision under rare class scenario

Models built using imperfect labels ***α*** to optimize combinations of **precision** and **recall** can perform very poorly compared to models built using true labels **y**

Learning with imperfect labels



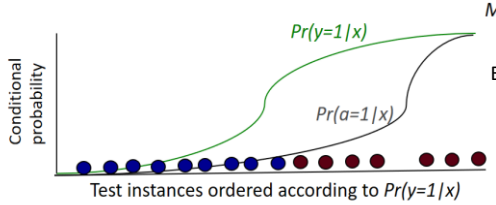
Assumptions

(1 α + β < 1

(2) Imperfect label is conditionally independent of feature space given the true label (CCN)

Ranking according to *Pr(a=1|x)* and *Pr(y=1|x)* is identical

Menon et al. ICML'15



Elkan et.al, KDD 2008

Blum et.al, COLT 1998

Given enough samples, it is possible to learn a classifier in CCN label noise scenario that is as good as one learned using perfectly labeled samples.

Natarajan et.al, NIPS 2013

For some performance measures like accuracy, a classifier can be learned using CCN labels treating them as perfect labels.

Menon et.al, ICML 2015

True class probability P(**y**=1|**x**) and corrupted class probability P(***a***=1|**x**) are monotonically related under CCN assumption and showed that balanced error rate, AUC can be optimized using corrupted labels.

Liu et.al, ICML 2003

For Positive and Unlabeled (PU) learning setting (CCN with β =0) presented an algorithm to optimize **precision*recall** without using any perfect labels.

RAPT: RAre class Prediction in absence of ground Truth

- ***Step 1***: Learn classification models using imperfect labels
  - We provide a new method to optimize precision*recall using imperfect labels and prove that the selected threshold maximizes the true precision*recall

- ***Step 2***: Combine predictions from classification model and the imperfect label
  - For ultra-rare class scenarios, the gain in precision after aggregation is significantly higher compared to the loss in recall
  - The improvement in G-measure after step 2 is by a factor of $\frac{(1-\alpha)}{\sqrt{P(a=1|g(x)>\gamma_\delta^{g,y})}}$.

- ***Step 3***: Collective classification to use spatial context
  - Exploits the relationship structure such as spatial neighborhood, biological network or social network, to improve the coverage (recall) of the rare class instances.

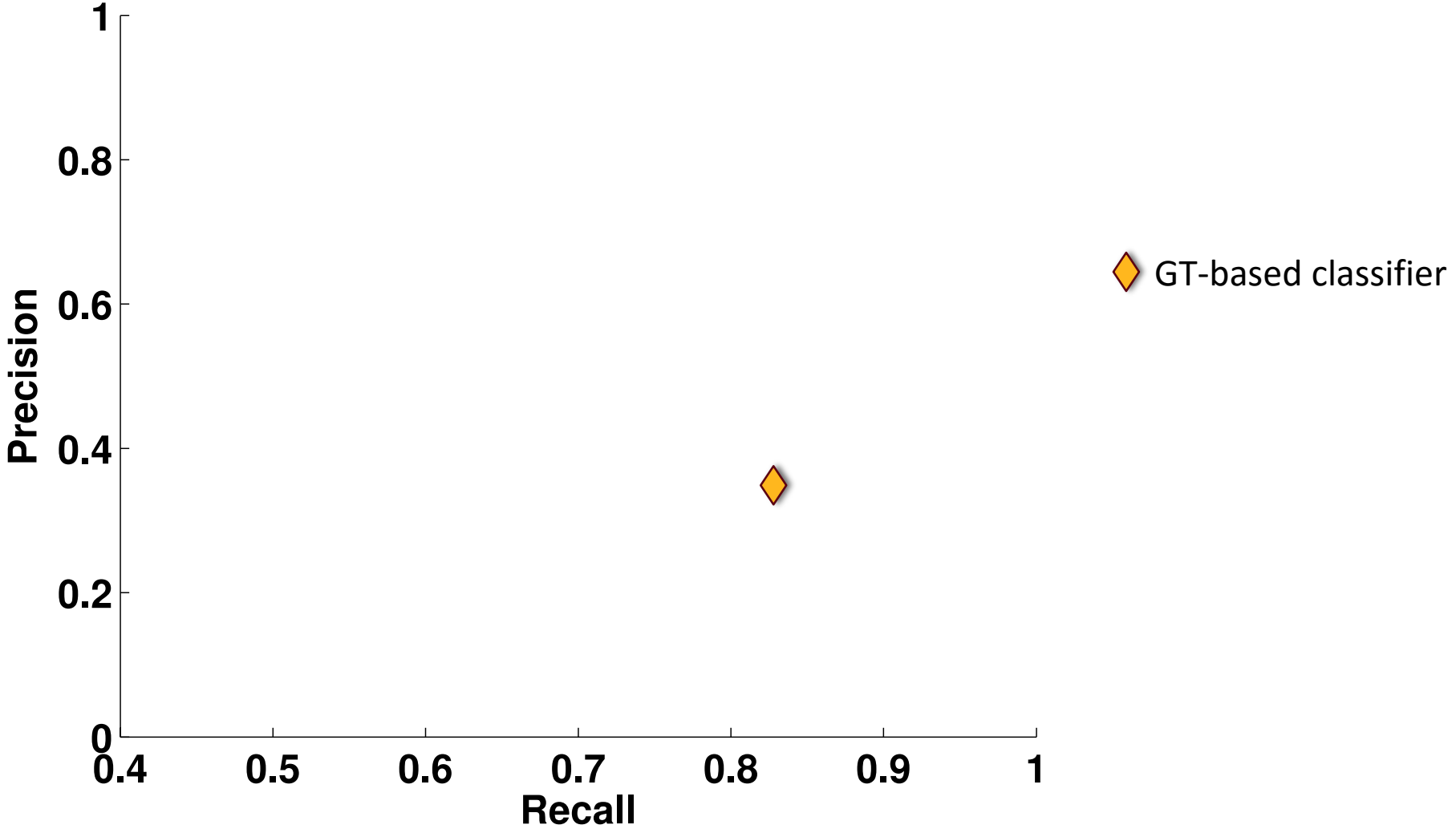
RAPT: RAre class Prediction in absence of ground Truth

- ***Step 1***: Learn classification models using imperfect labels that is as good as the one that could be built using perfect labels
  - We provide a new method to optimize precision*recall using imperfect labels and prove that the selected threshold maximizes the true precision*recall

- ***Step 2***: Combine predictions from classification model and the imperfect label to improve precision at some loss in recall
  - For ultra-rare class scenarios, the gain in precision after aggregation is significantly higher compared to the loss in recall
  - The improvement in G-measure after step 2 is by a factor of $\frac{(1-\alpha)}{\sqrt{P(a=1|g(x)>\gamma_o^{g,y})}}$.

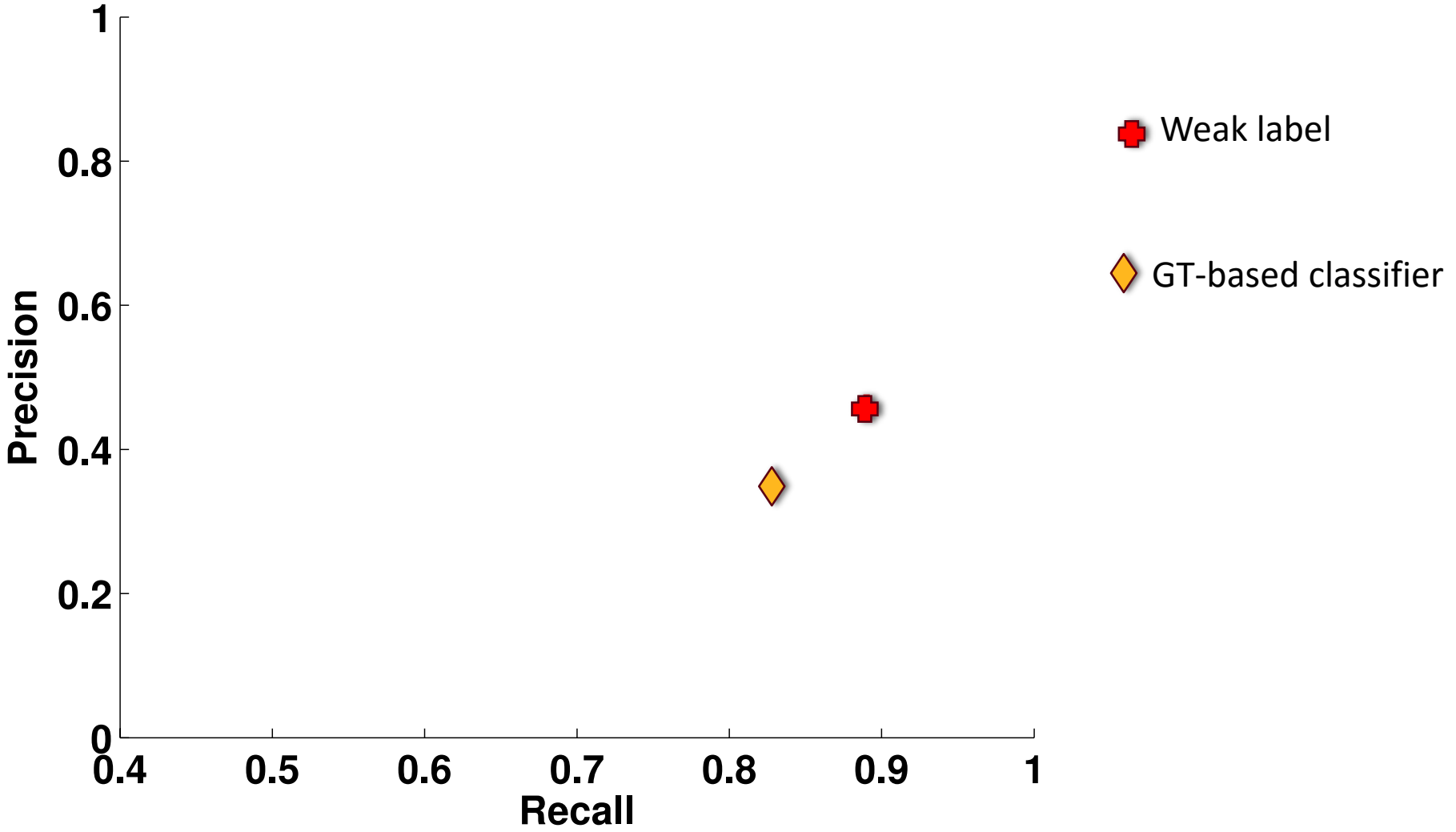
- ***Step 3***: Use spatial context to improve recall
  - Exploits the relationship structure such as spatial neighborhood, biological network or social network, to improve the coverage (recall) of the rare class instances.

Results for Burned Area Mapping



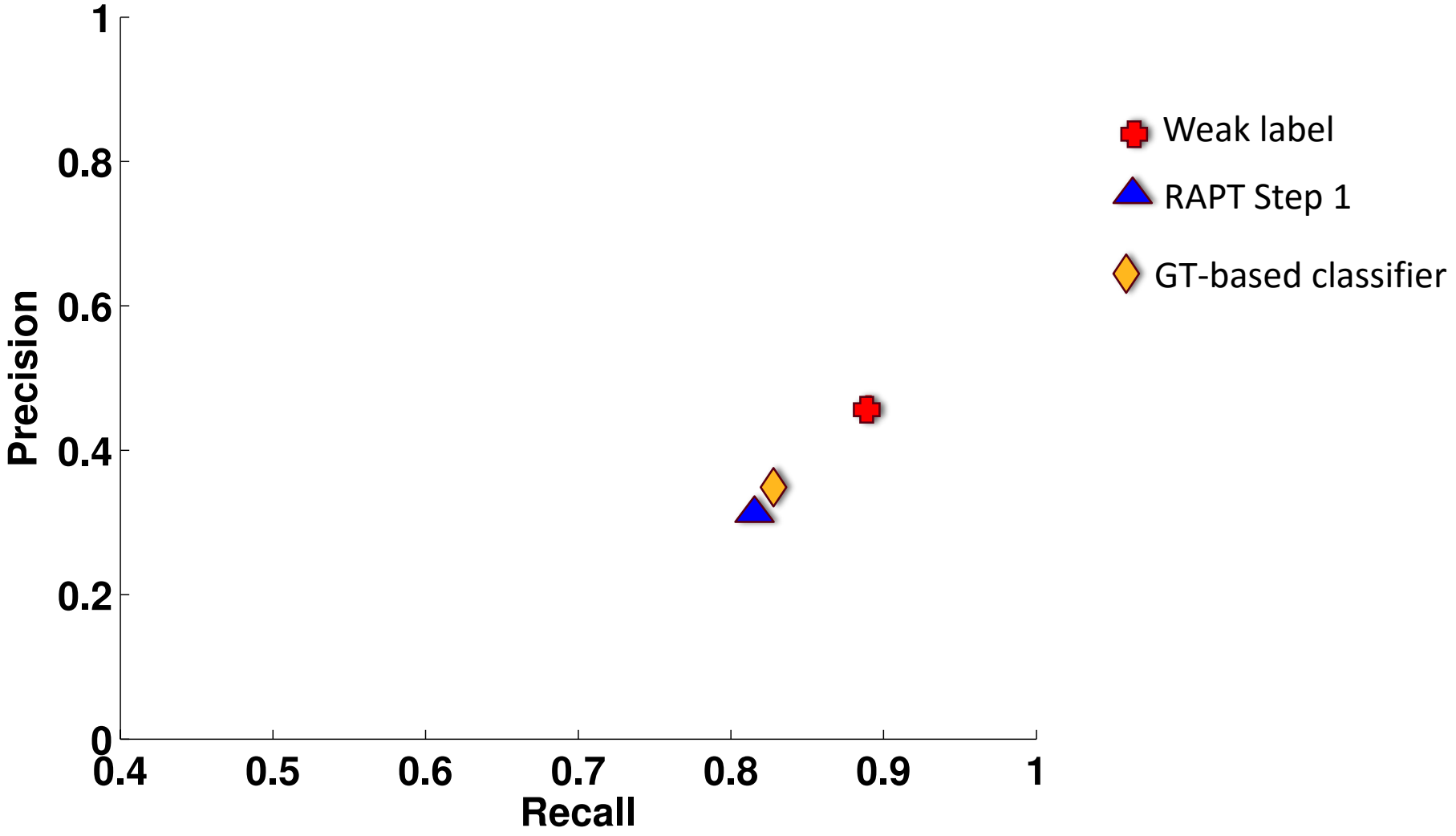
California State

Results for Burned Area Mapping



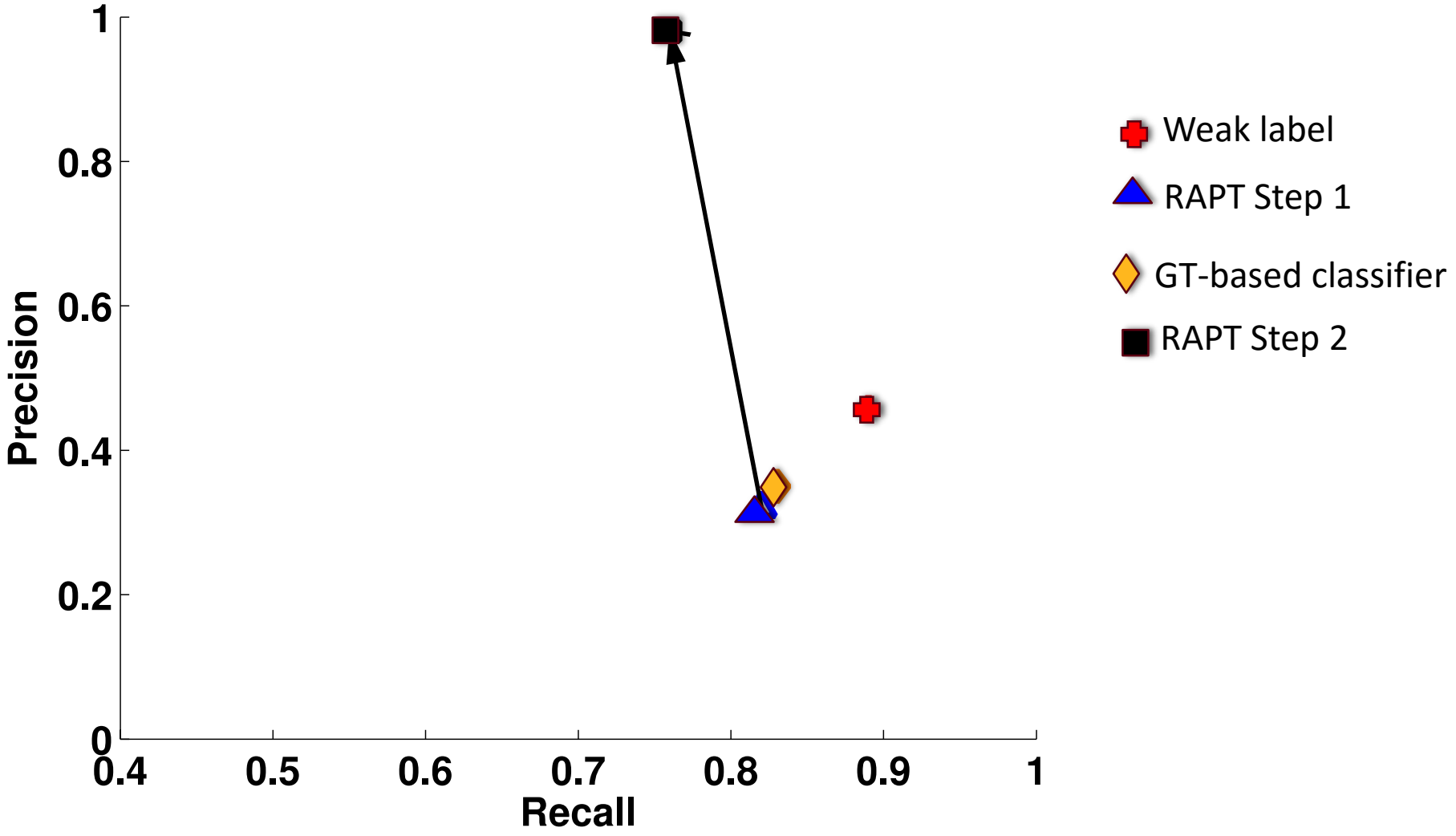
California State

Results for Burned Area Mapping



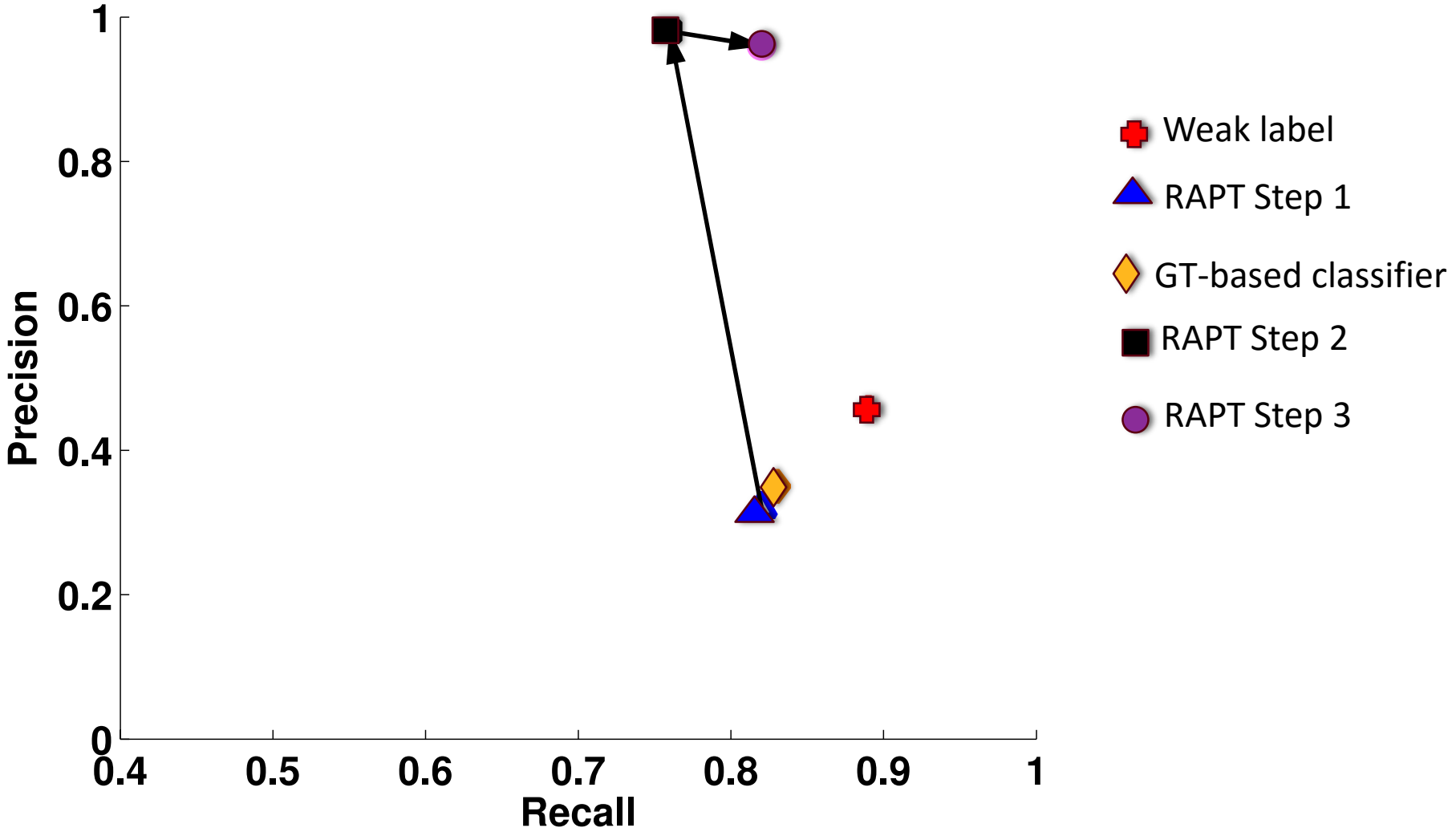
California State

Results for Burned Area Mapping



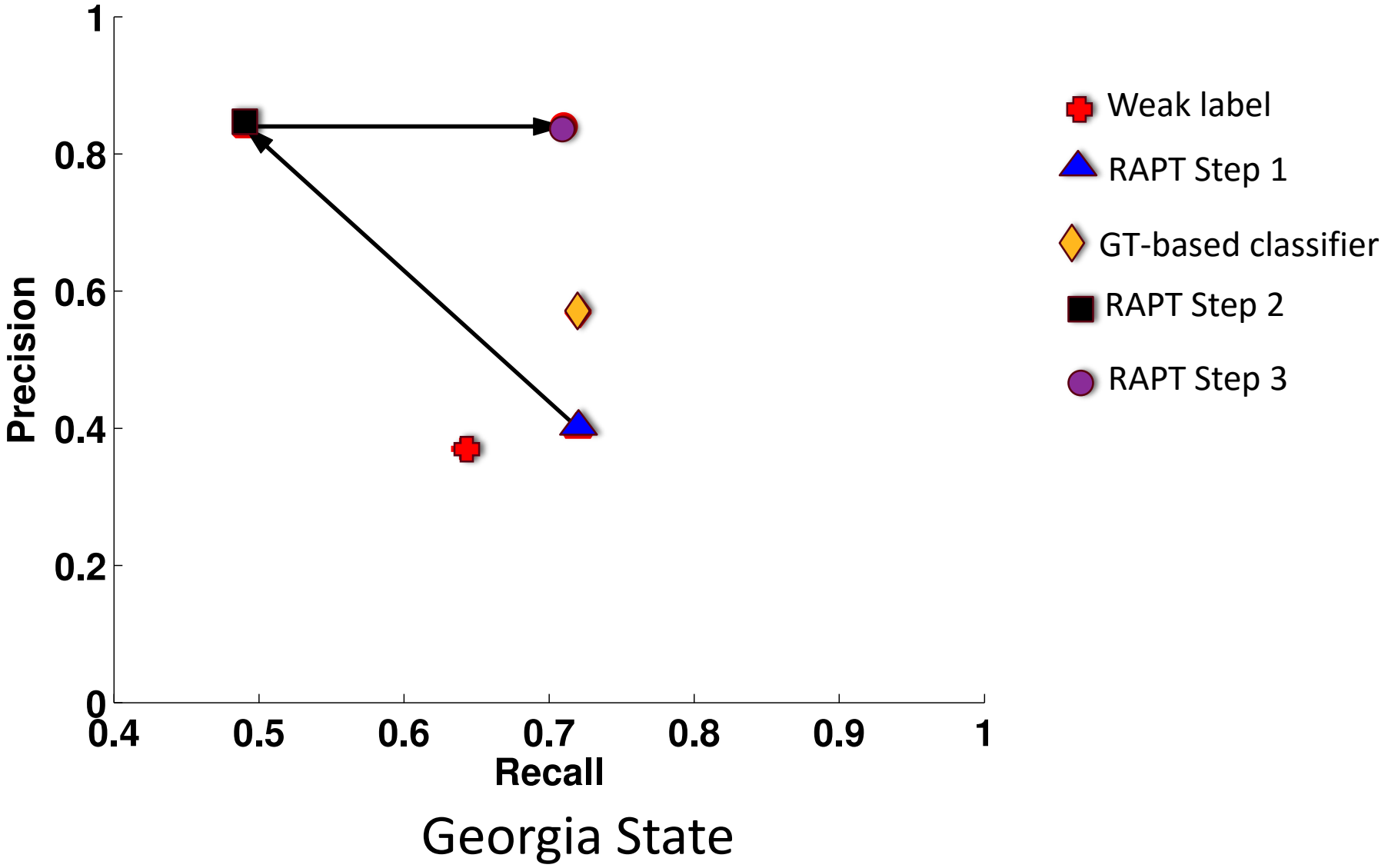
California State

Results for Burned Area Mapping



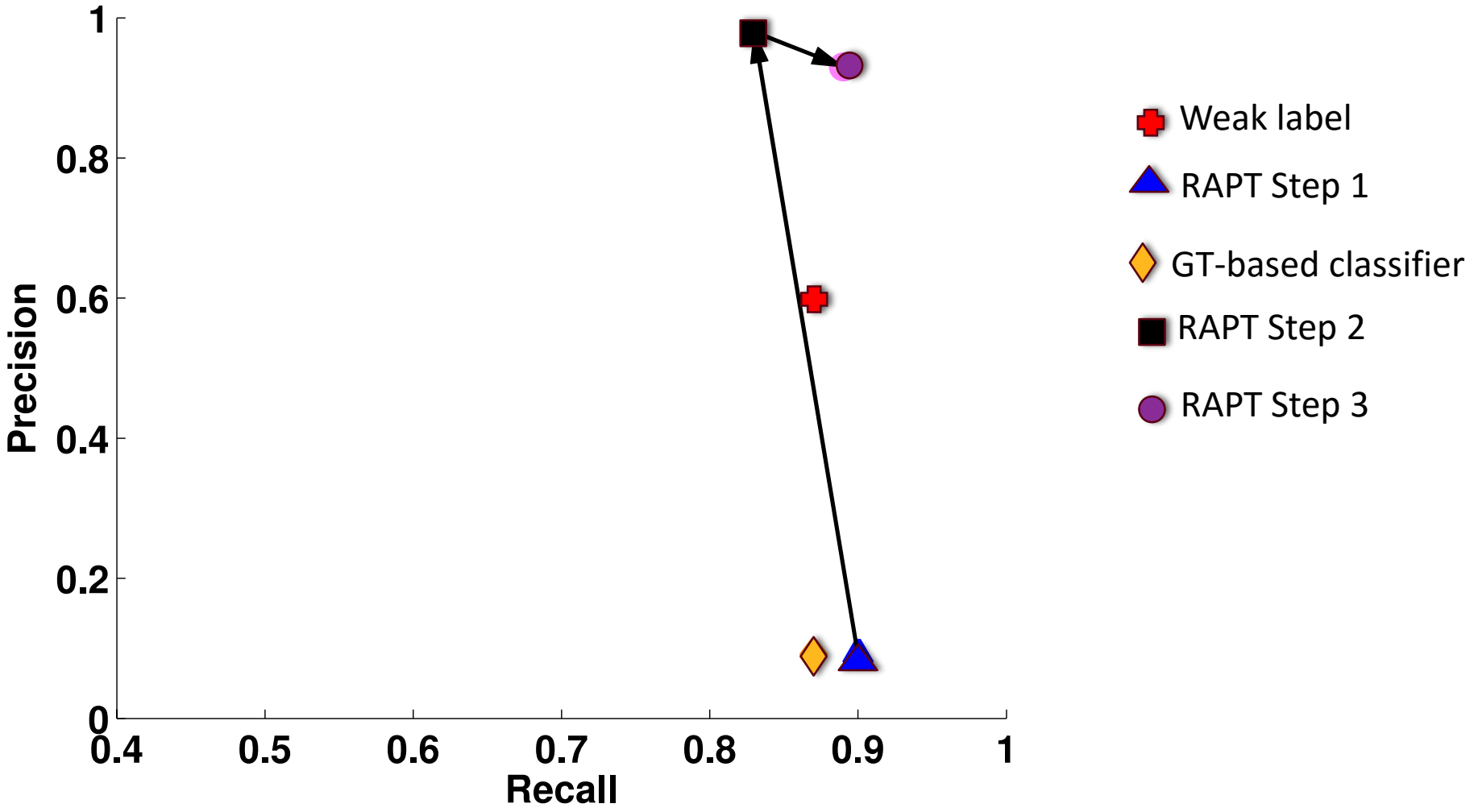
California State

Results for Burned Area Mapping



Georgia State

Results for Burned Area Mapping



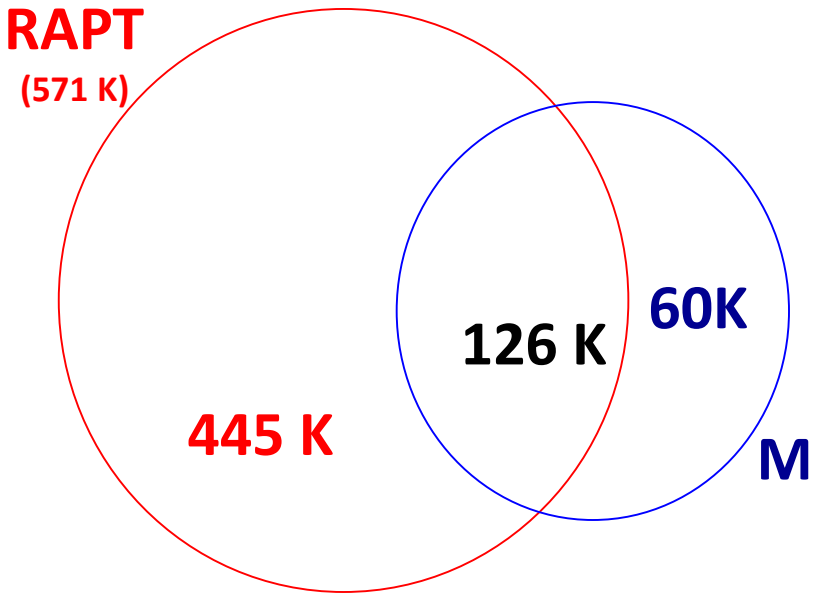
Montana State

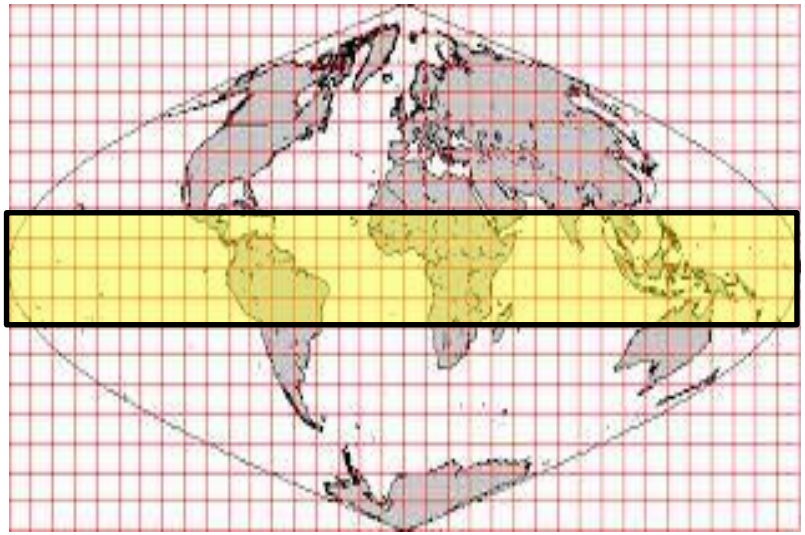
Global Monitoring of Fires in Tropical Forests

Fires in tropical forests during 2001-2014

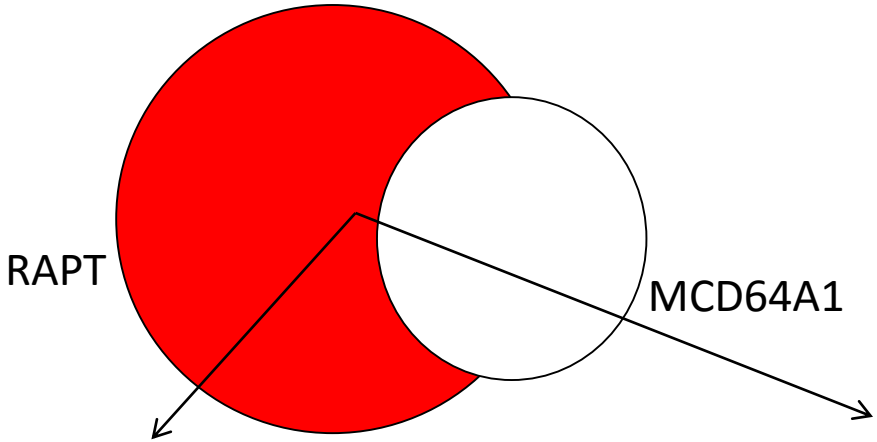
571 K sq. km. burned area found in tropical forests

three times the area reported by state-of-art NASA product: MCD64A1





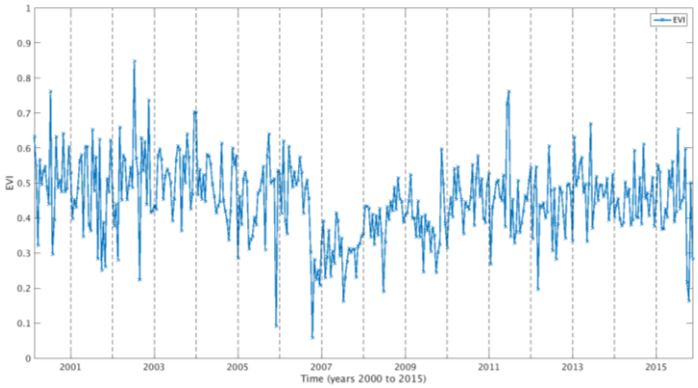
Validation

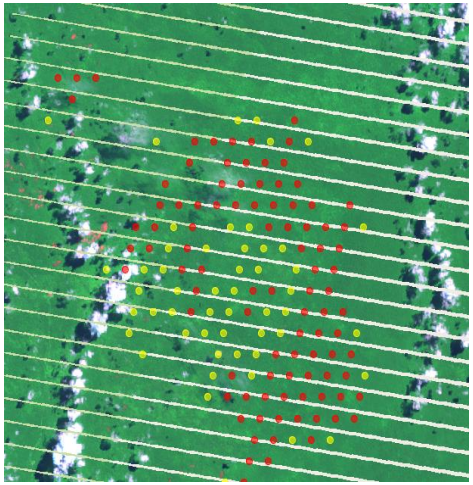


Multiple lines of evidence indicate that RAPT-only points are actual forest fires

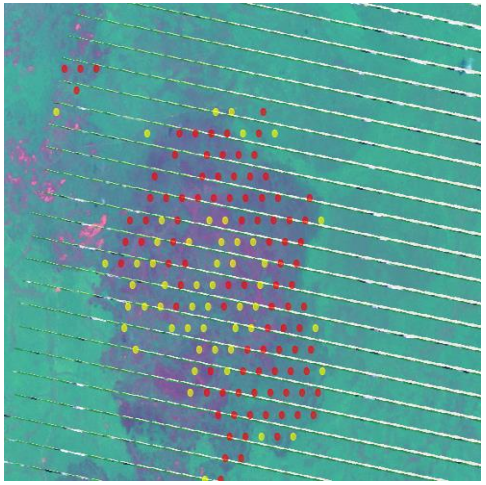
Burn scar in Landsat composite

Change in Vegetation series





Before Fire Event

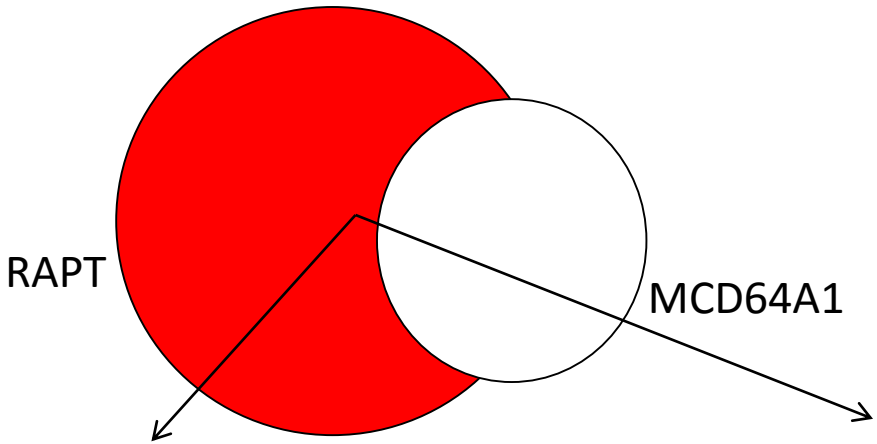


After Fire Event

Sudden drop followed by recovery is a key signature of forest fires

Landsat false-color composite shows the scar after the fire event identified by RAPT

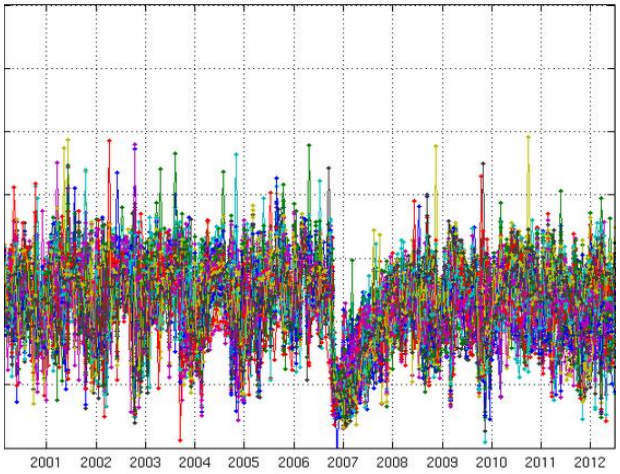
Validation



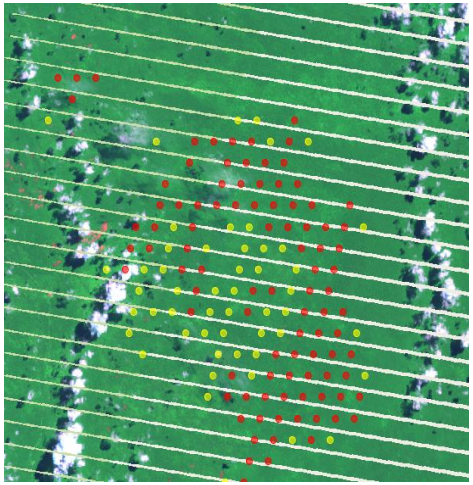
Multiple lines of evidence indicate that RAPT-only points are actual forest fires

Burn scar in Landsat composite

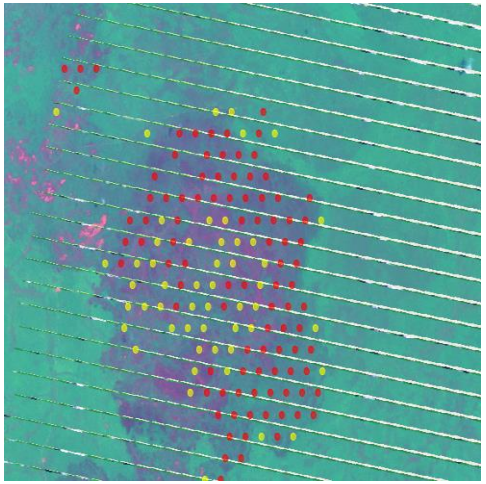
Change in Vegetation series



Synchronized drop followed by recovery is a key signature of forest fires



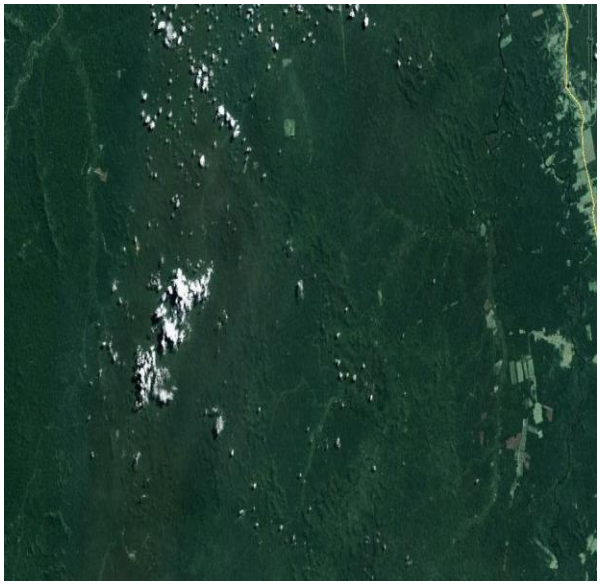
Before Fire Event



After Fire Event

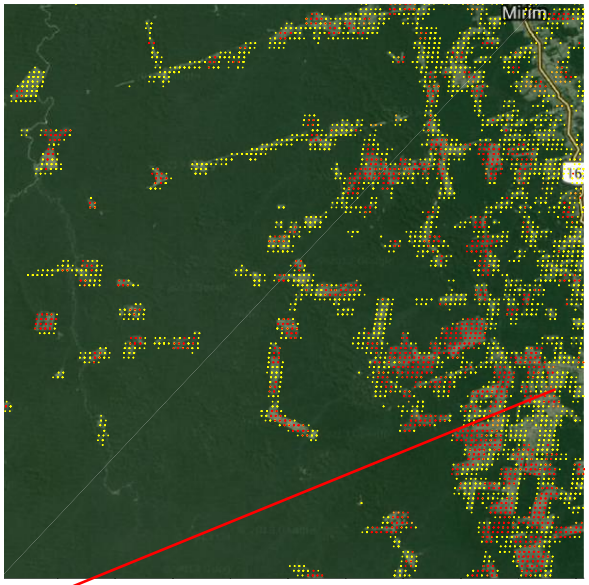
Landsat false-color composite shows the scar after the fire event identified by RAPT

Deforestation via Burning in Amazon



Google Earth Image: Year 2002

Google Earth Image: Year 2015



RAPT detection 2002-2014
(RAPT only Common)

Burn Detection					B	B	B						
Land cover	F	F	F	F	F	F	F	N	N	N	N	N	N
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014

Case Study 2: Mapping of Plantation Dynamics

SCIENTIFIC AMERICAN™ SUSTAINABILITY

Palm Oil Set to Grow Indonesia's Climate Changing Emissions

Draining peatlands and replacing forests with palm oil plantations may cause Indonesian pollution to soar, despite pledges

By Nathanael Massey, ClimateWire on October 10, 2012

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

NATURE | NEWS

Fibre production drives deforestation in Indonesia

Study debunks belief that palm-oil plantations are main culprit.

Natasha Gilbert 21 July 2014

SCIENTIFIC AMERICAN™ SUSTAINABILITY December 1, 2012

Stop Burning Rain Forests for Palm Oil

The world's growing appetite for cheap palm oil is destroying rain forests and amplifying climate change

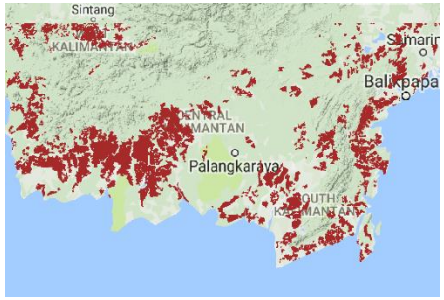
Interplay between food, energy and water:

Production of edible oils and biofuels.
High carbon emissions.
Degradation of water quality.

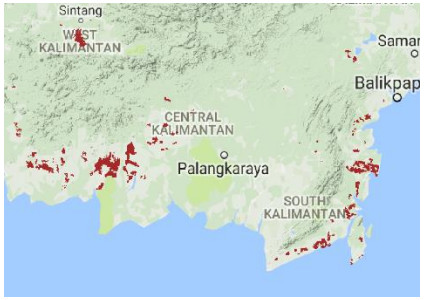
State-of-the-art and Challenges

•**Tree Plantation (TP):** This data set is created by Transparent World, with the support of Global Forest Watch. Plantations are manually annotated on 2014. *TP has high recall and low precision.*

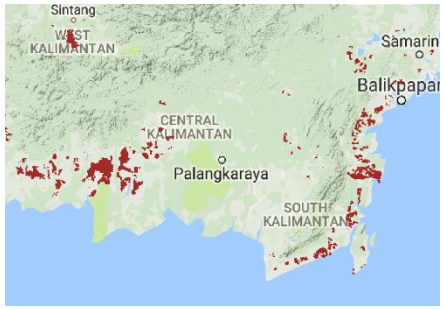
•**Roundtable on Sustainable Palm Oil (RSPO):** This dataset is available across Indonesia in 2000, 2005, and 2009. In addition, the study digitized all the locations into 19 land cover types in these eras. *RSPO has high precision and low recall.*



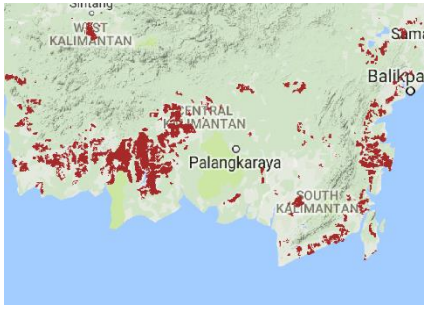
TP, 2014



RSPO, 2001



RSPO, 2005



RSPO, 2009

➢ Challenges

•Imperfect annotators
- •*Tree Plantation (TP):* high recall and low precision.
- •*Roundtable on Sustainable Palm Oil (RSPO):* high precision and low recall.

•Data heterogeneity
- •Land cover heterogeneity
  - ❑ differentiate plantation from a variety of land covers, e.g. forest, are highly confused with plantations.
- •Spatial heterogeneity
- •Temporal heterogeneity
  - ❑ Seasonal variation - e.g., a crop land after harvest looks very similar to a barren land.
  - ❑ Yearly variation – the spectral features of a land cover change across years.

•Noisy and high-dimensional feature space

Our Contribution

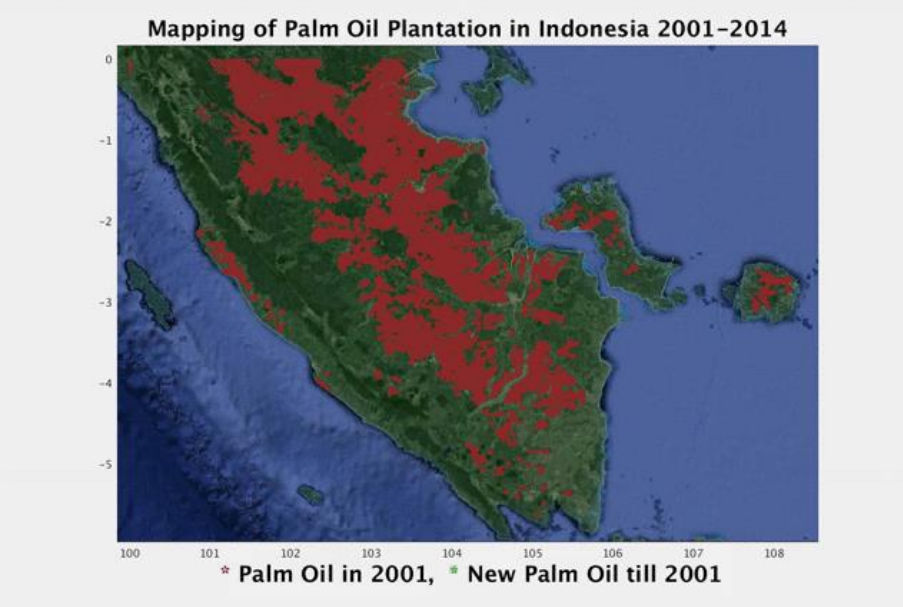
- Learning from multiple imperfect annotators
  (Jia et al. BigData 2016)
  - Each annotator has different expertise level on different plantation types.
  - We recursively update the expertise of each annotator and estimate true labels.

- Handling temporal heterogeneity in prediction
  (Jia et al. SDM 2017)
  - We model temporal and spatial dependencies across years in an LSTM model.
  - We propose an incremental learning strategy to update the LSTM model.

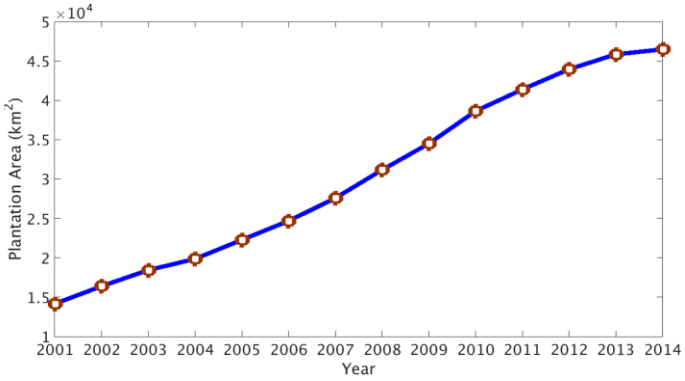
- Aggregating classes, collecting samples and validating results
  (Jia et al. Technical Report, 2017)
  - We aggregate similar classes according to domain expertise.
  - For each aggregated class, we sample equal amount of samples from each sub-class across multiple years.
  - We validate the generated plantation maps by comparing random sampled locations to high-resolution images.

- 1. Jia, X., Khandelwal, A., Gerber, J., Carlson, K., West, P., and Kumar, V. Learning Large-scale Plantation Mapping from Imperfect Annotators. In IEEE Big Data (Big Data), 2016.
- 2. Jia, X., Khandelwal, A., Nayak, G., Gerber, J., Carlson, K., West, P., and Kumar, V. Predict Land Covers with Transition Modeling and Incremental Learning. In SDM, 2017.
- 3. Jia, X., Khandelwal, A., Gerber, J., Carlson, K., Samberg, L., West, P., and Kumar, V. Automated Plantation Mapping in Southeast Asia Using Remote Sensing Data. In Department of Computer Science and Engineering-Technical Reports.

Annual Plantation Maps

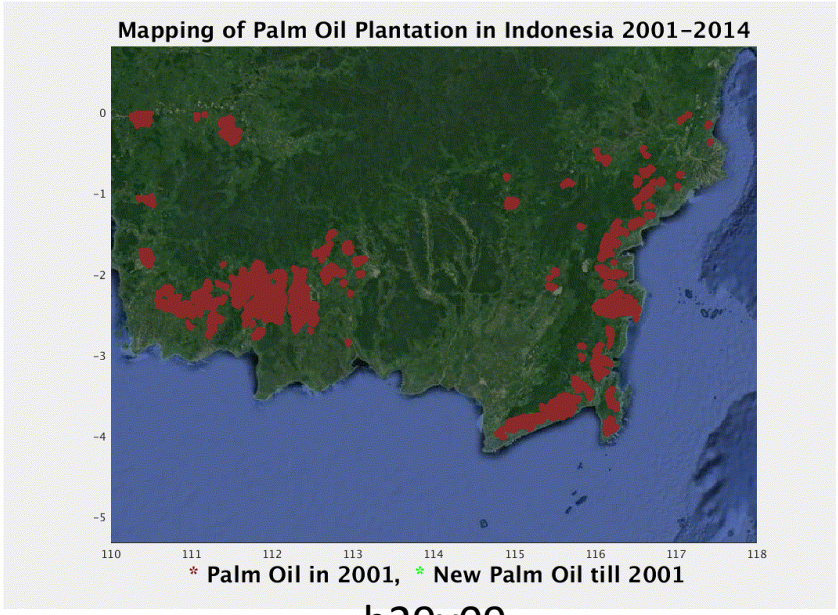


h28v09



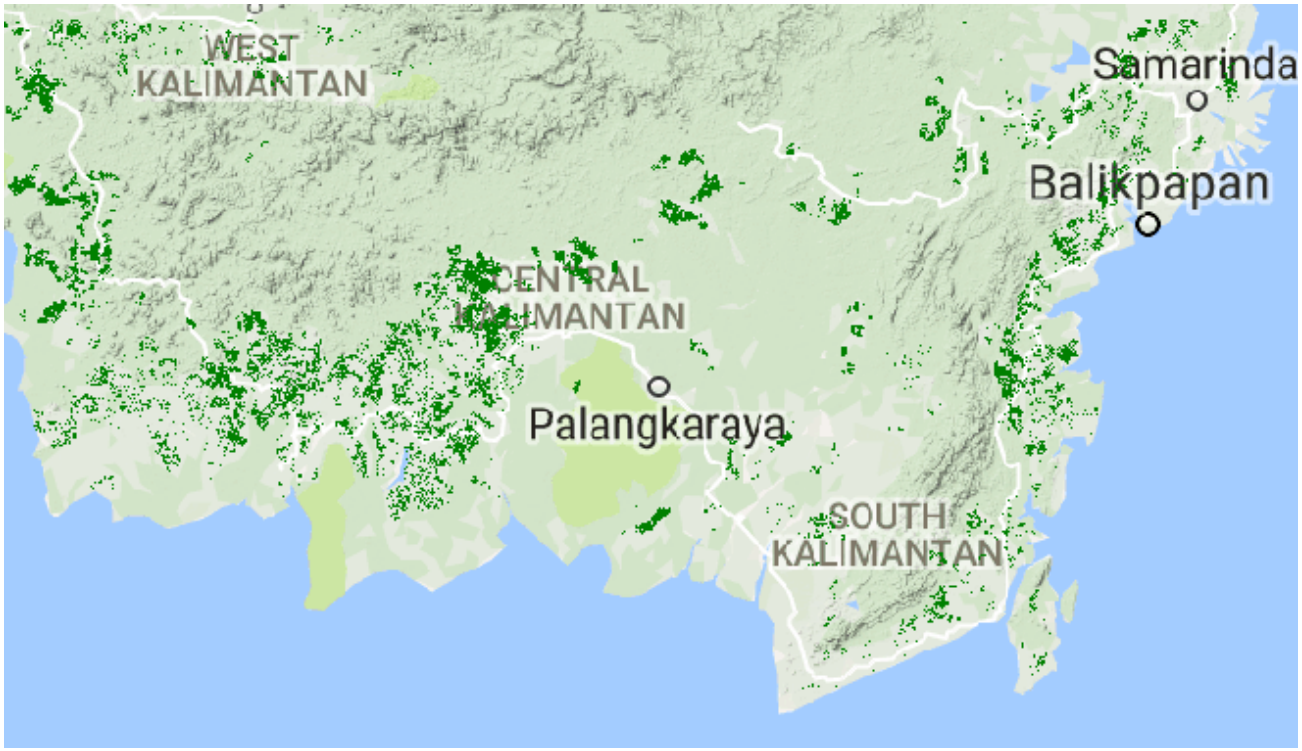
- Annual growth rate ≈ 9.57%

h29v08



h29v09

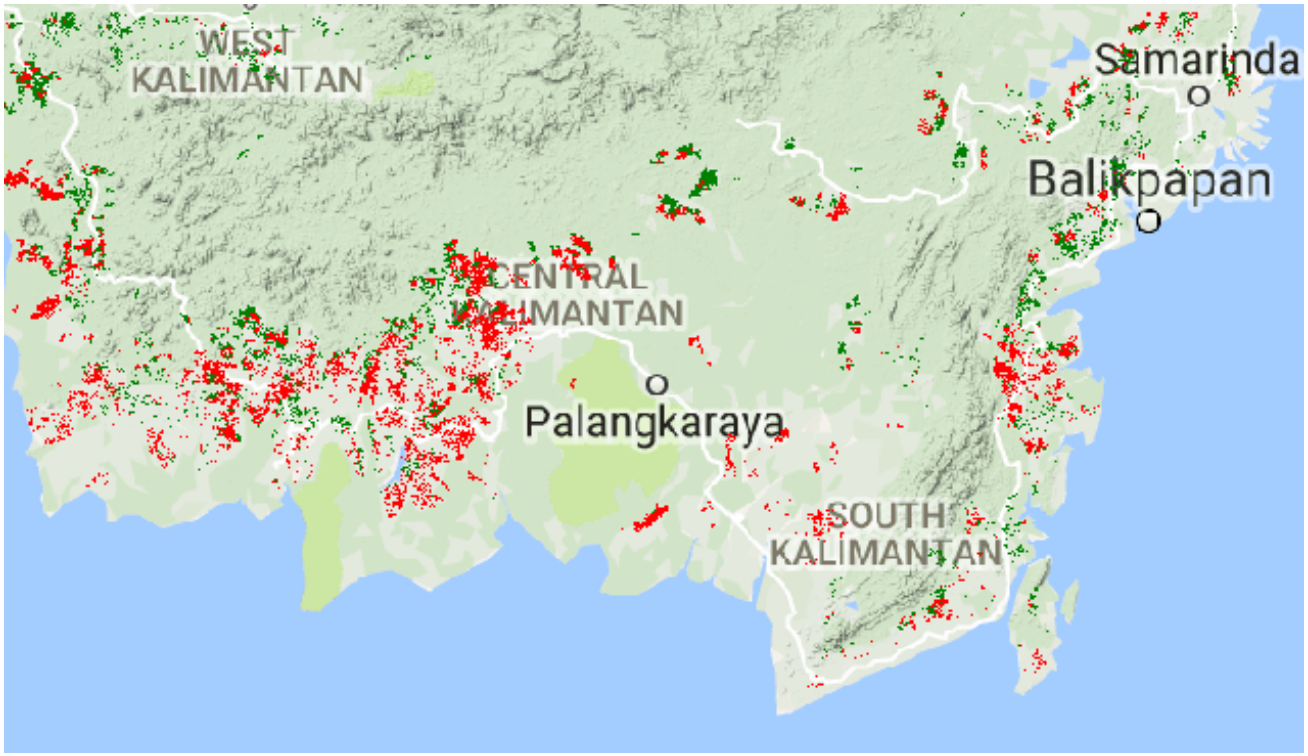
Interaction between Fires and Palm Oil Plantation



All plantations

This and all following figures show only confident forest pixels.

Interaction between Fires and Palm Oil Plantation

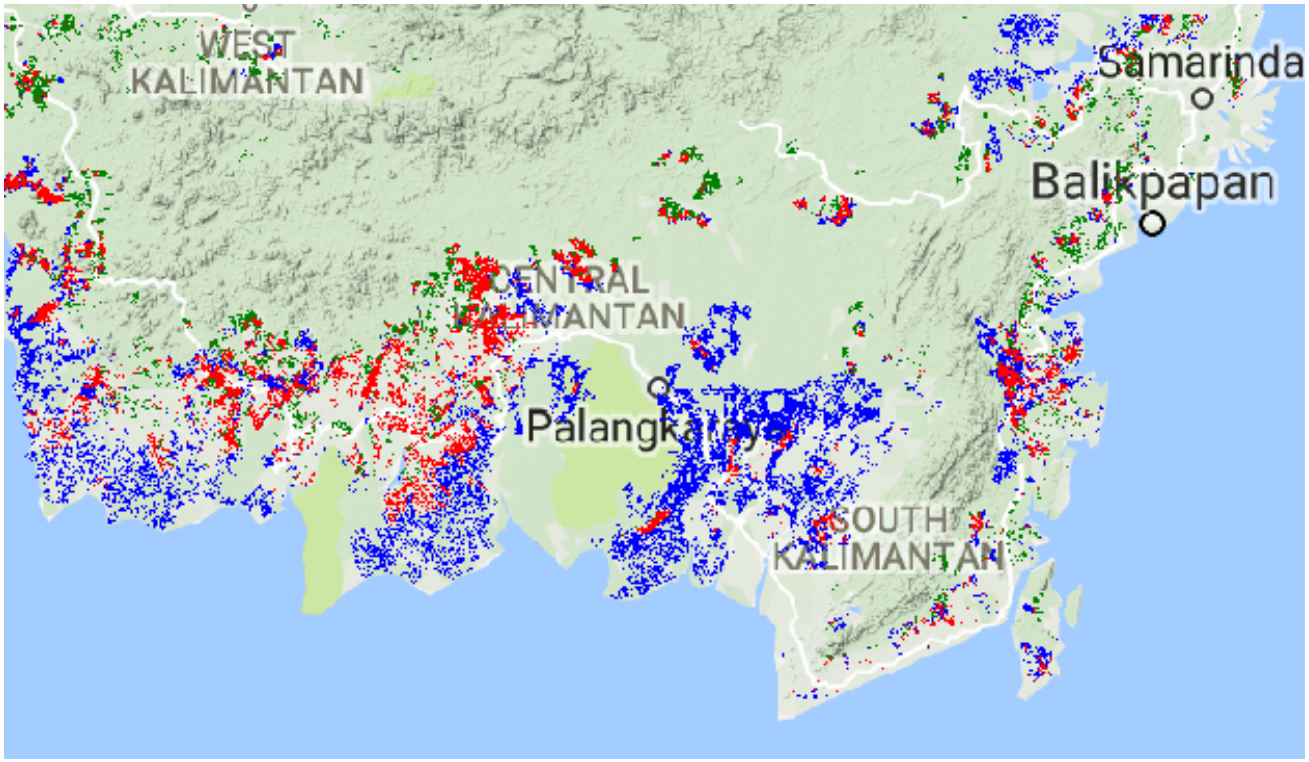


Plantations with no burn scars

Plantations with burn scar during 2001-2014

This and all following figures show only confident forest pixels.

Interaction between Fires and Palm Oil Plantation



Plantations with no burn scars

Plantations with burn scar during 2001-2014

Burned pixels around plantations (same burn date as nearby red pixels)

This and all following figures show only confident forest pixels.

Case Study 3: Global Mapping of Surface Water Dynamics

Brazil's Severe Drought Dries Up Reservoirs

California is not alone: São Paulo is also facing severe water restrictions.

Oil-Rich Persian Gulf Looks to Renewables to Avert Water Crisis BloombergBusiness January 19, 2016

Kariba Dam Water Levels 'Dire,' Zambian Energy Minister Says January 8, 2016

nature International weekly journal of science

Published online 12 August 2009 | *Nature* **460**, 789 (2009) | doi:10.1038/460789a

News

Satellite data show Indian water stocks shrinking

Groundwater depletion raises spectre of shortages.

Effect Of Climate Change On Agriculture: Droughts, Heat Waves Cut Global Cereal Harvests By 10 Percent In 50 Years

TECH TIMES January 7,

Smithsonian.com

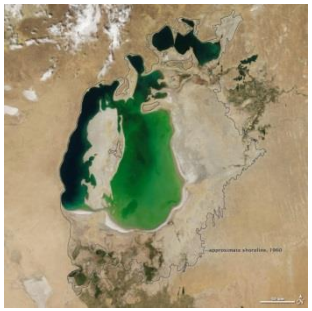
The Colorado River Runs Dry

Dams, irrigation and now climate change have drastically reduced the once-mighty river. Is it a sign of things to come?

Cedo Caka Lake in Tibet, 1984

Cedo Caka Lake in Tibet, 2011

Melting of glacial lakes in Tibet



Aral Sea in 2000

Aral Sea in 2014

Shrinking of Aral Sea since 1960s

Importance of Monitoring Global Surface Water Dynamics

Brazil's Severe Drought Dries Up Reservoirs

California is not alone: São Paulo is also facing severe water restrictions.

Oil-Rich Persian Gulf Looks to Renewables to Avert … Crisis BloombergBusiness …

Kariba Dam Water Levels

nature International weekly journal of science

Published online 12 August 2009 | *Nature* **460**, 789 (2009) | doi:10.1038/460789a

News

…w Indian water stocks

…es spectre of shortages.

Effect Of Climate C… Droughts, Heat Wav… Harvests By 10 Per…

TECH TIMES Jan…

…ian.com

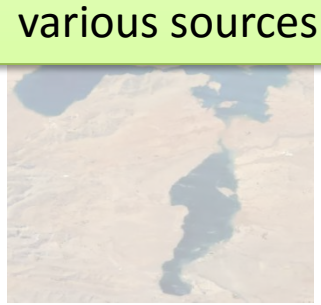
…ver Runs Dry

…limate change have drastically reduced …a sign of things to come?

Opportunity in using Remote Sensing Data

- Multi-spectral data
  - MODIS (at 500m, from 2000)
  - Landsat (at 30m, from 1970s)
- Can be used to classify every location at a given time as water or land (binary classes)
- Ground truth on specific dates available from various sources: SRTM, GLWD

Cedo Caka Lake in Tibet, 1984



Cedo Caka Lake in Tibet, 2011

Melting of glacial lakes in Tibet

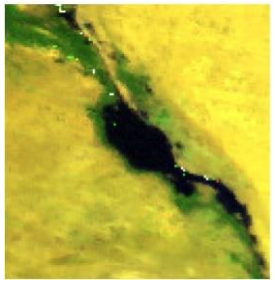
Aral Sea in 2000

Aral Sea in 2014

Shrinking of Aral Sea since 1960s

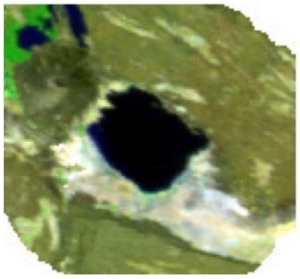
Challenges for Traditional Big Data Methods in Monitoring Water

- **Challenge 1: Heterogeneity in space and time**
  - Water and land bodies look different in different regions of the world
  - Same water body can look different at different time-instances

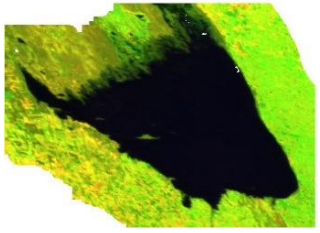


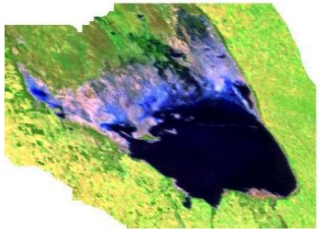
Great Bitter Lake, Egypt

Lake Tana, Ethiopia



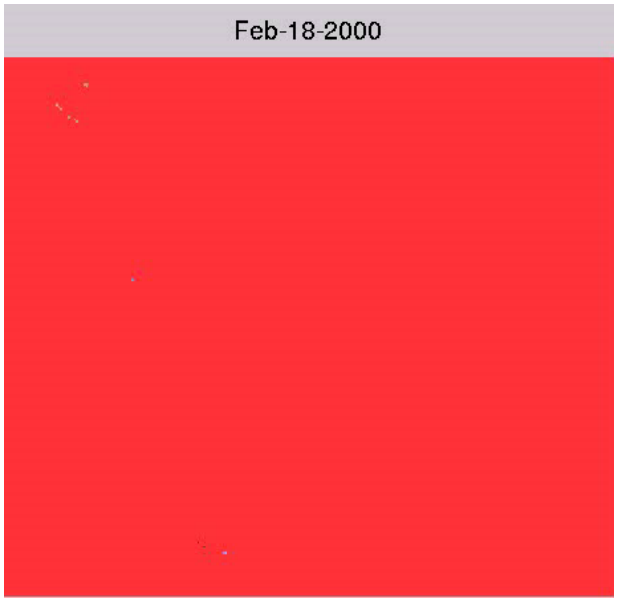
Lake Abbe, Africa





Mar Chiquita Lake, Argentina in 2000 (left) and 2012 (right)

- **Challenge 2: Data Quality**
  - Noise: clouds, shadows, atmospheric disturbances
  - Missing data

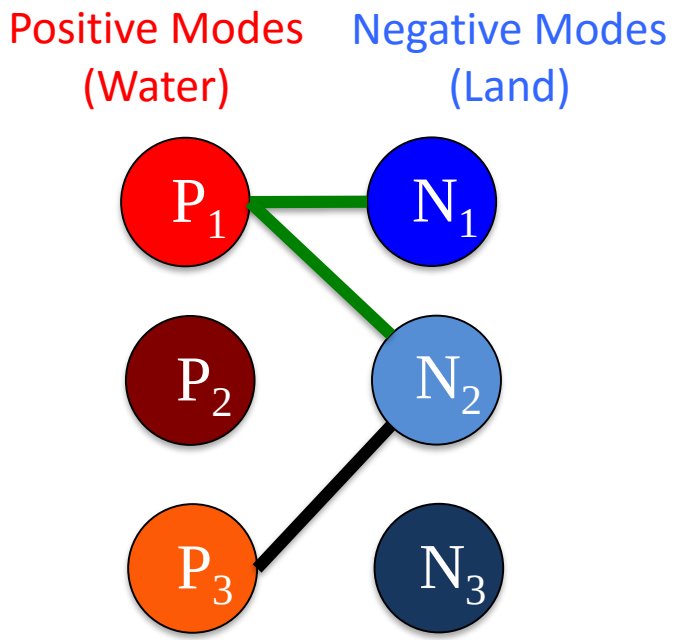


Poyang Lake, China
(Pink color shows missing data)

Method Innovations for Monitoring Water

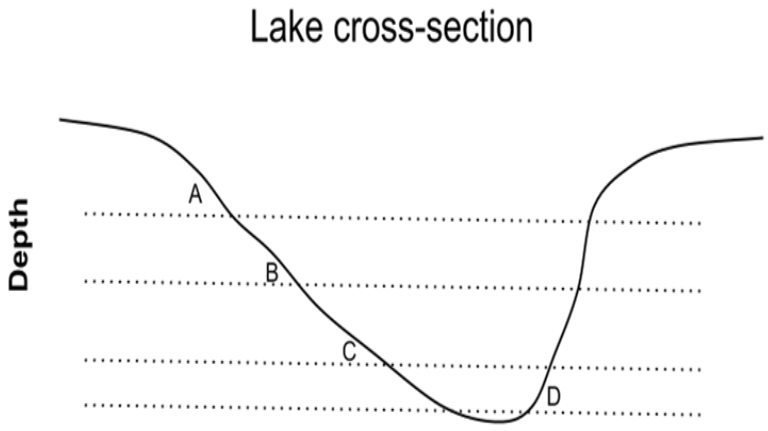
- **Ensemble Learning Methods for Handling Heterogeneity in Data** [1,2]

Learn an ensemble of classifiers to distinguish b/w different pairs of positive and negative modes



- **Using Physics Guided Labeling to Handle Poor Data Quality**[3,4]

Use elevation information to constrain physically-consistent labels



Elevation A > B > C > D

[1] Karpatne et al. SDM 2015
[2] Karpatne et al. ICDM 2015
[3] Khandelwal et al. ICDM 2015
[4] Mithal et al. (PhD Dissertation)

A Global Surface Water Monitoring System

http://z.umn.edu/monitoringwater

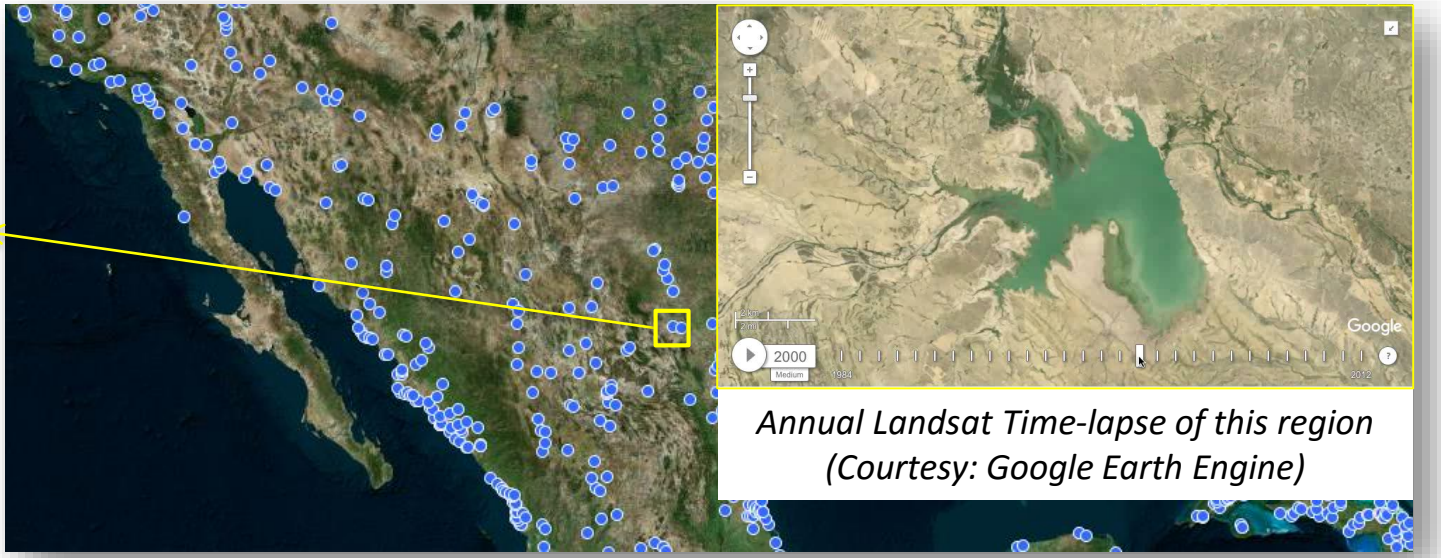
- Maps the dynamics of all major surface water bodies (surface area > 2.5 km^2) shown as *blue dots*

Key Highlights:

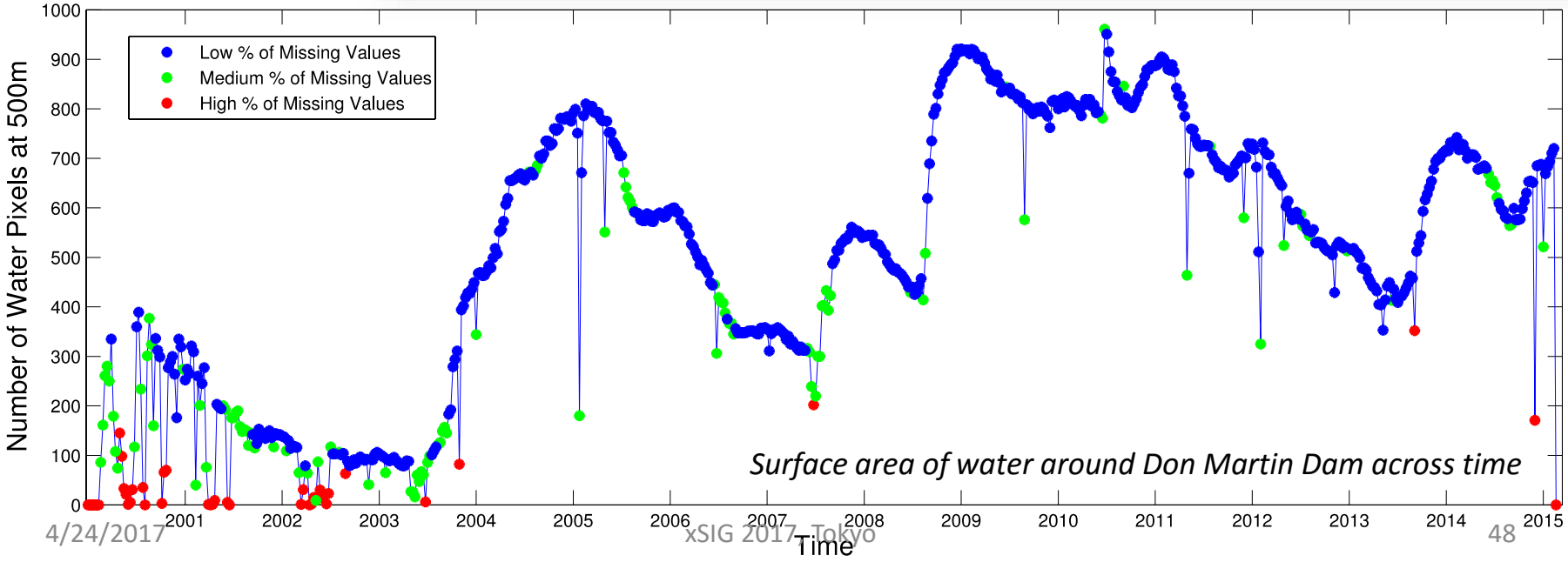
- Detects melting of glacial lakes
- Maps changes in river morphology
- Identifies reservoir constructions
- Finds relationships b/w surface water and precipitation/groundwater

Showing Surface Water Dynamics

Don Martin Dam, Mexico



Annual Landsat Time-lapse of this region (Courtesy: Google Earth Engine)

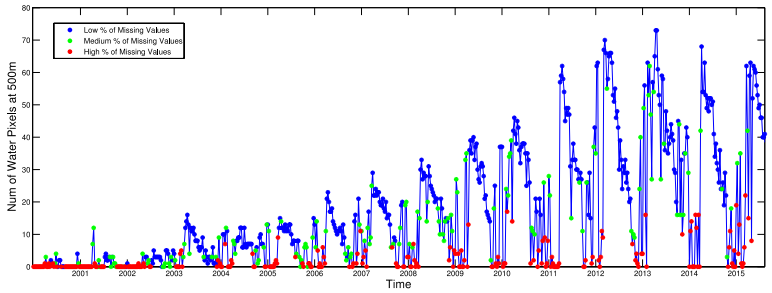


Surface area of water around Don Martin Dam across time

Regions of Change in South America

Red Dots *(Water Gain):*

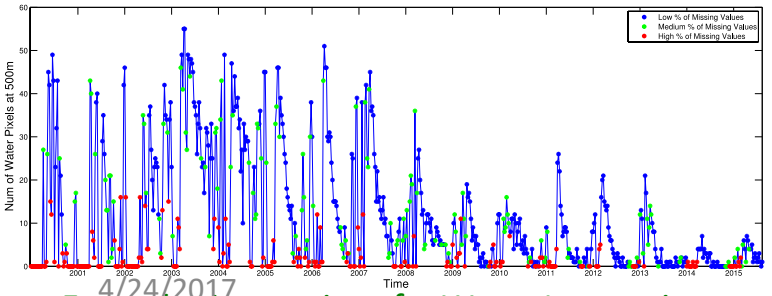
Region of size > 2.5 km^2 that have changed from land to water in the last 15 years



Example time series of a *Water Gain* region

Green Dots (Water Loss):

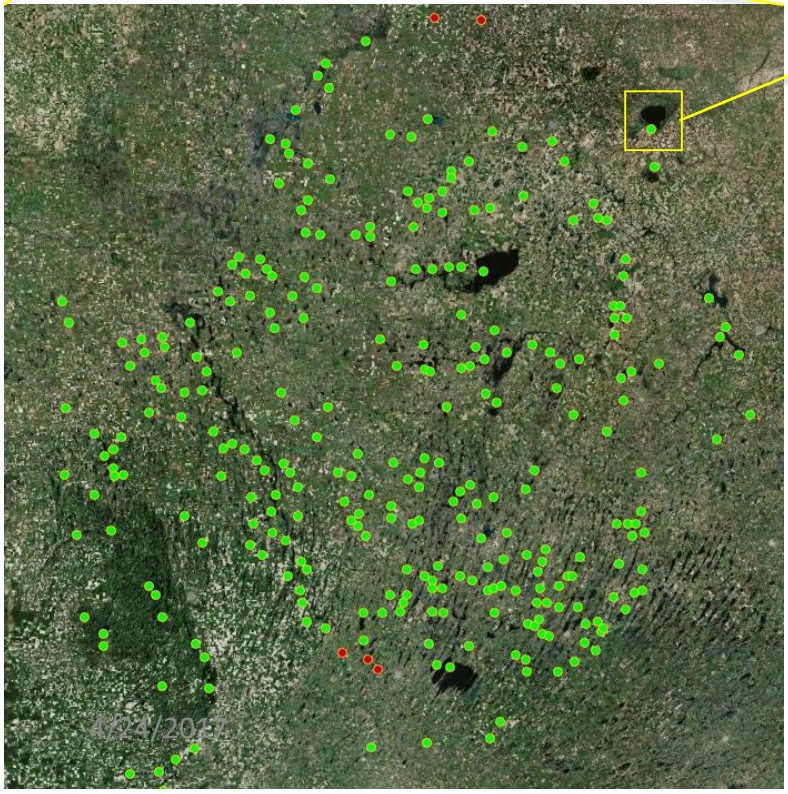
Region of size > 2.5 km^2 that have changed from water to land in the last 15 years



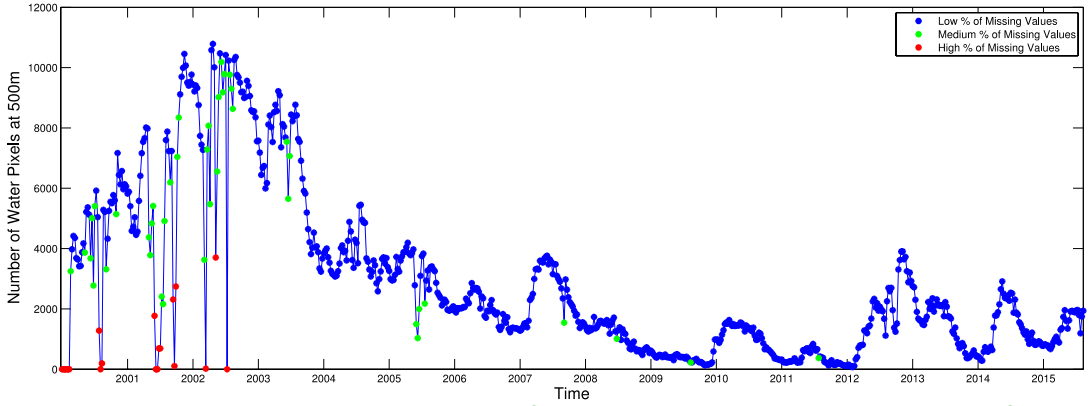
Example time series of a *Water Loss* region

Examples of Change: Shrinking Water Bodies

(Green dots show regions changing from water to land in last 15 years)



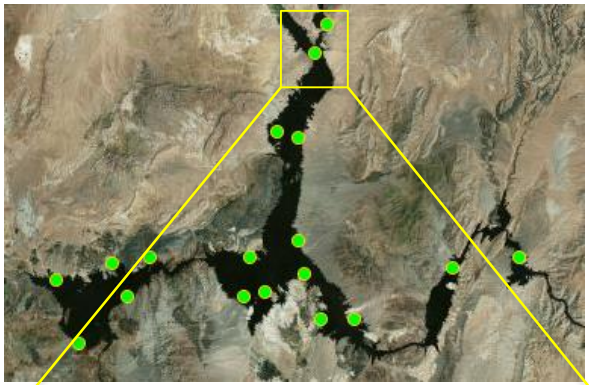
Annual Time-lapse of an example green dot



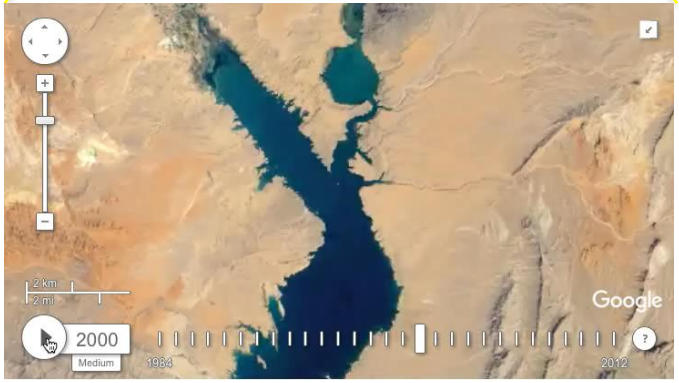
Aggregate dynamics of all green dots shown on left

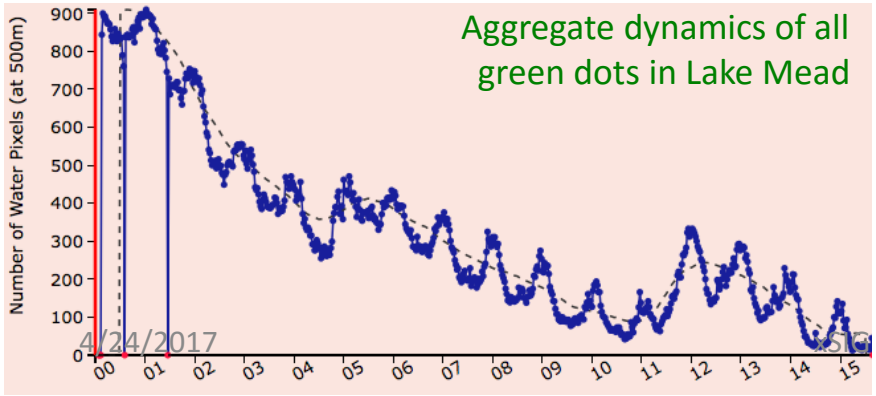
Examples of Change: Colorado River

Shrinking of Lake Mead

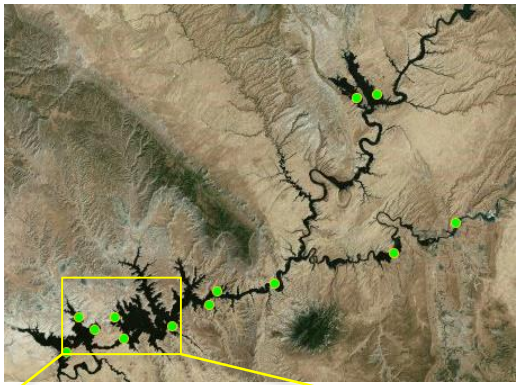


Annual time-lapse of selected region

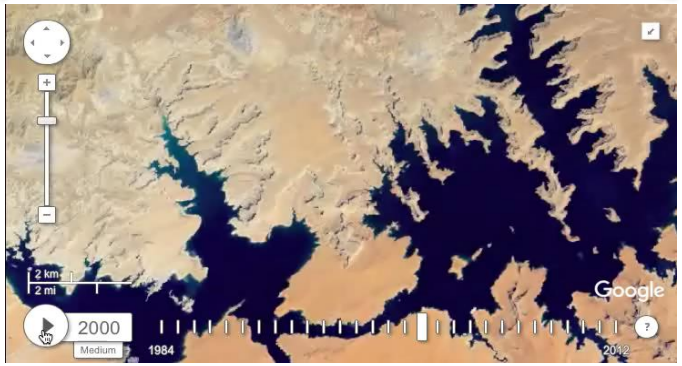


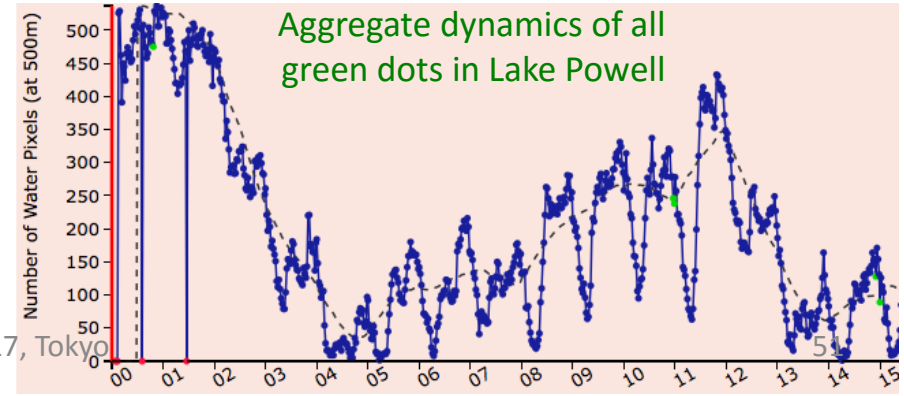


Shrinking of Lake Powell

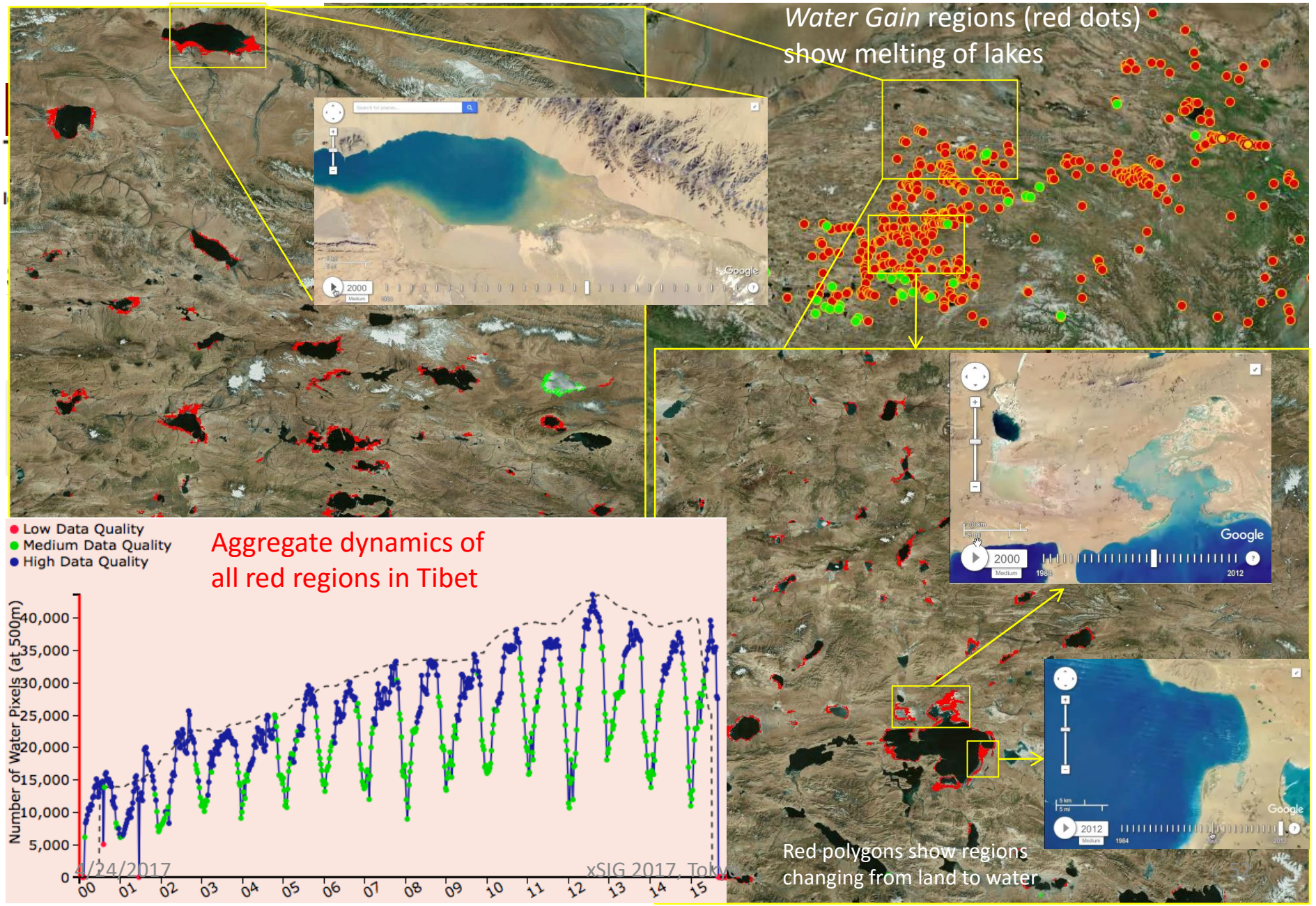


Annual time-lapse of selected region



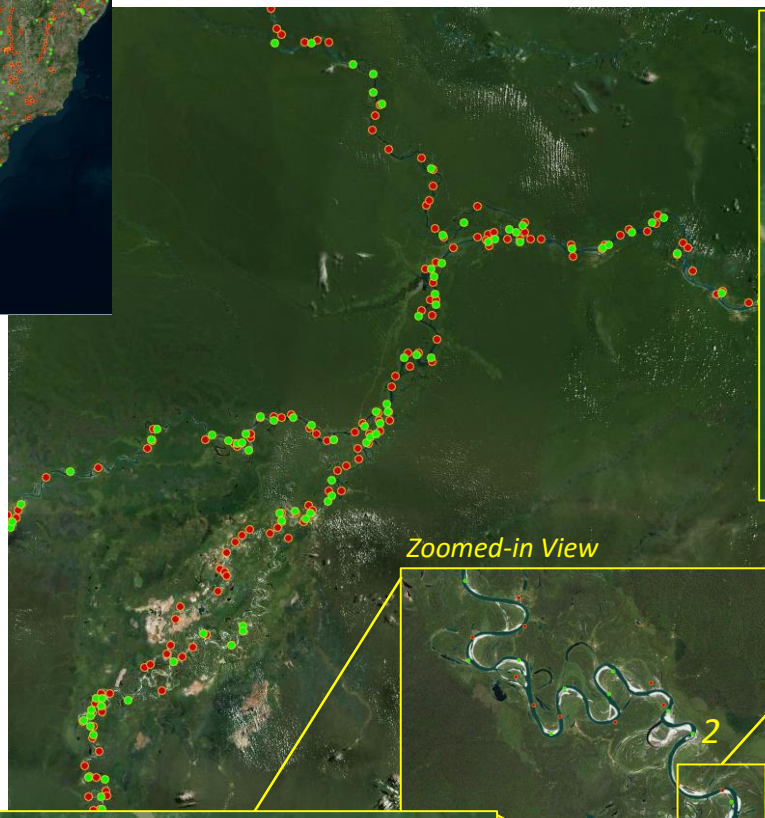


Examples of Change: Melting Glacial Lakes in Tibet

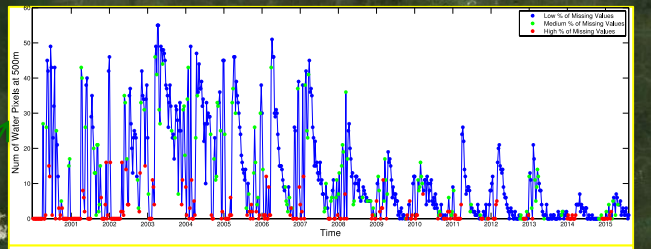


Examples of Change: River Meandering

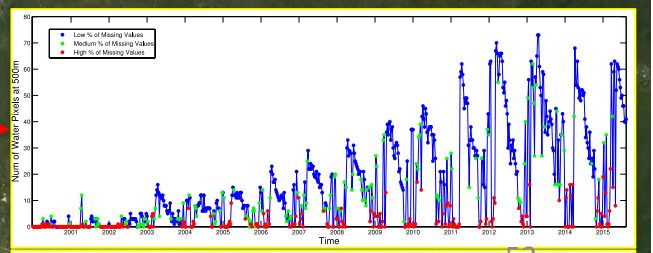
(Adjacent occurrence of *Water Gain (red)* and *Water Loss (green)* regions all along the river indicate the displacement of water from the green dots to the red dots)



Zoomed-in View



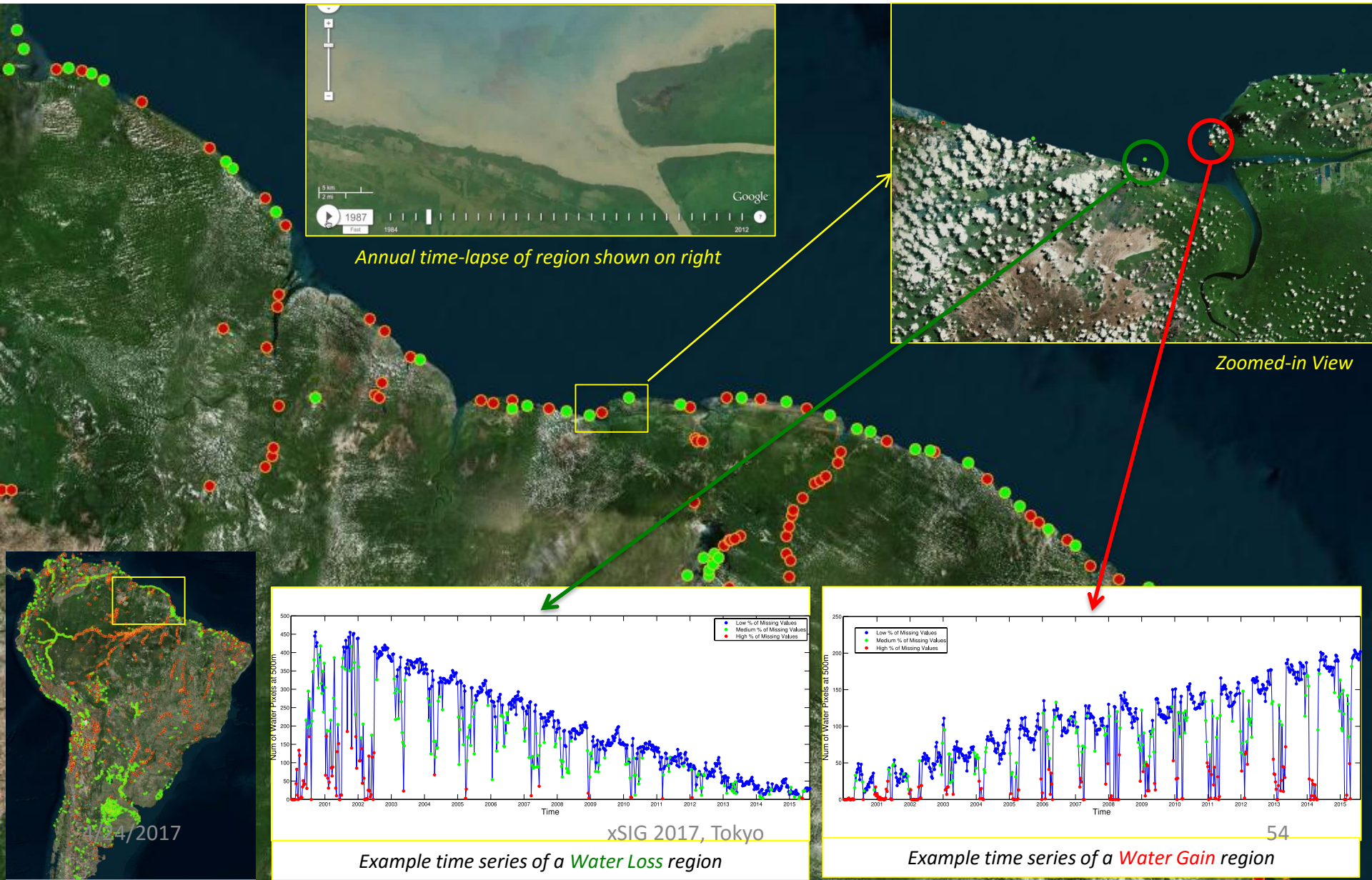
Example time series of a Water Loss region



Example time series of a Water Gain region

Examples of Change: Delta Erosion

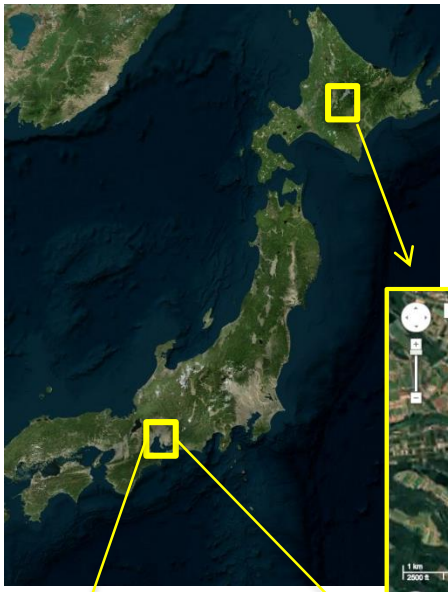
(*Water Gain* and *Water Loss* regions appear on the coastline, due to displacement of sediments around river deltas)



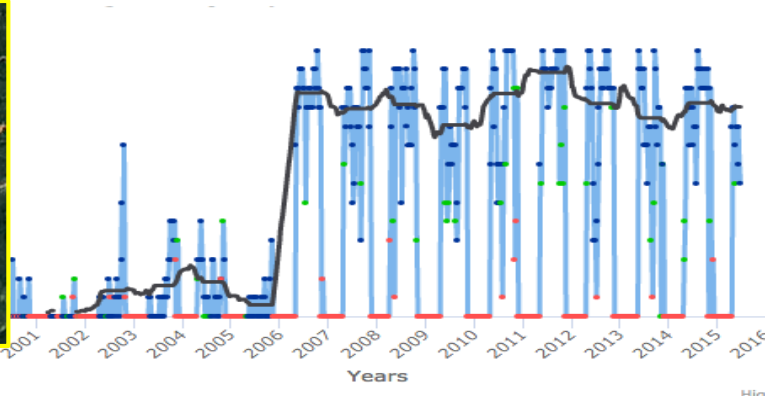
Example time series of a Water Loss region

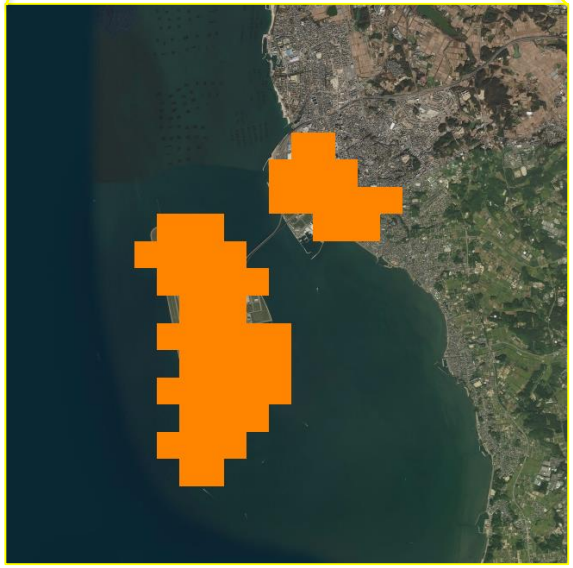
Example time series of a Water Gain region

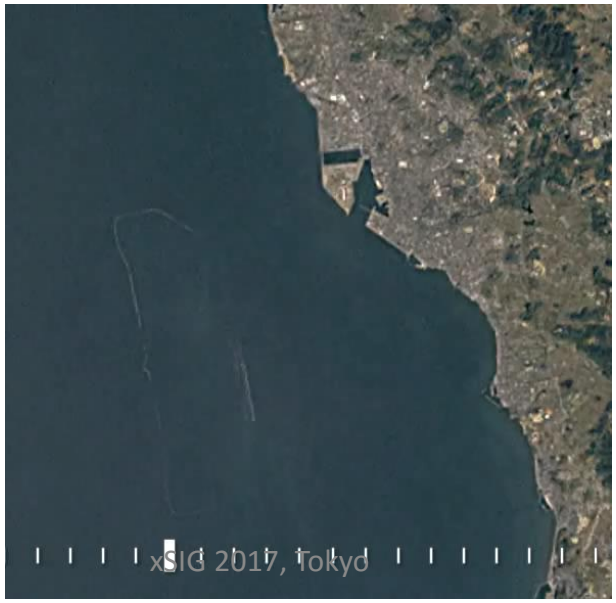
Examples of Changes in Japan



1. Construction of Chubetsu Dam, Hokkaido







2. Construction of Chubu Centrair International Airport

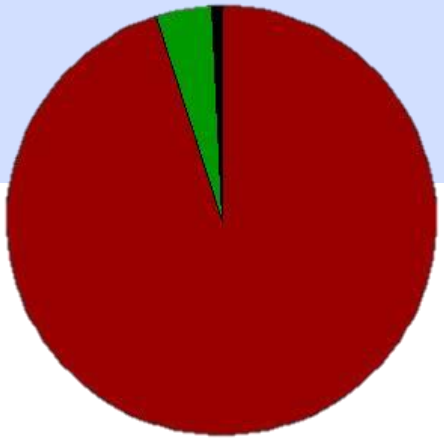
(Orange polygons show regions changing from land to water)

Global Reservoir and Dam (GRanD) Database

A data curation initiative by Global Water System Project (GWSP)

Dams reported by GRanD since 2001: 35

Comparison of Dam Detections with GRanD



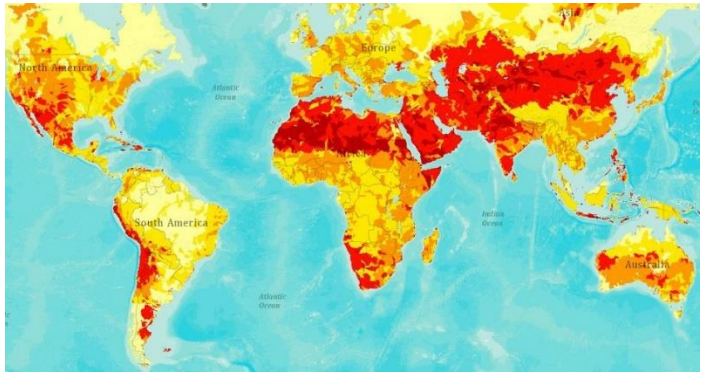
Dams only reported by GRanD: 5

Dams reported by both UMN and GRanD: 30

Dams only reported by UMN: 671

Potential Use Cases of a Water Monitoring System

Quantifying water stocks and risks

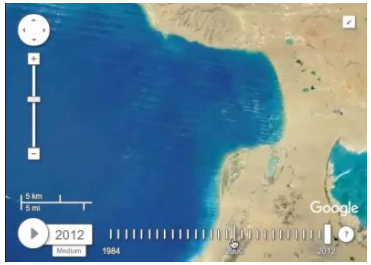


Global projections of water risks (red)

Global mapping of river discharge

Gleason et al. PNAS 2014

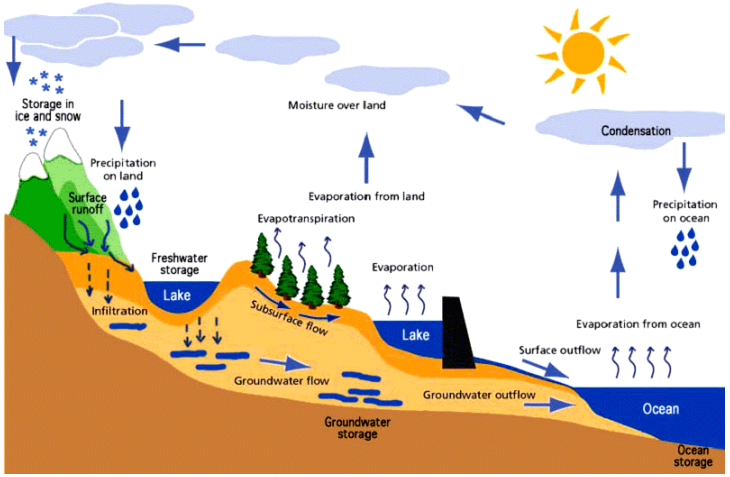
Assessing the impact of climate change and human actions



Melting Glacial Lakes

Constructions of Dams

Integrating with hydrological models



Concluding Remarks

- Big data techniques hold great promise for increasing our understanding of the Earth's climate and environment.
- Domain theory can be used to guide the process of knowledge discovery in scientific data
  - "Theory-guided Data Science"
- Methods have applicability across diverse domains:
  - Ecosystem management
  - Epidemiology
  - Geospatial Intelligence
  - Neuroscience

Thank You! Questions?

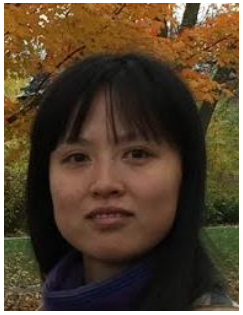
UMN Graduate Students


Anuj Karpatne


Ankush Khandelwal


Varun Mithal


Guruprasad Nayak


Xi Chen


Xiaowei Jia


Saurabh Agrawal

UMN Team Members
Arindam Banerjee, Snigdhansu Chatterjee, Stefan Liess, Shashi Shekhar, Michael Steinbach

NSF Expeditions Collaborators
NCSU: Nagiza Samatova, Fredrick Semazzi; Northeastern: Auroop Ganguly; North Carolina A&T: Abdollah Homaifar, Fred Semazzi

External Collaborators
NASA Ames: Rama Nemani, Nikunj Oza; IonE, UMN: Kate Braumann; UCLA: Dennis Lettenmaier, Miriam Marlier

Work supported by NASA and NSF Expeditions in Computing project on Understanding Climate Change using Data-driven Approaches