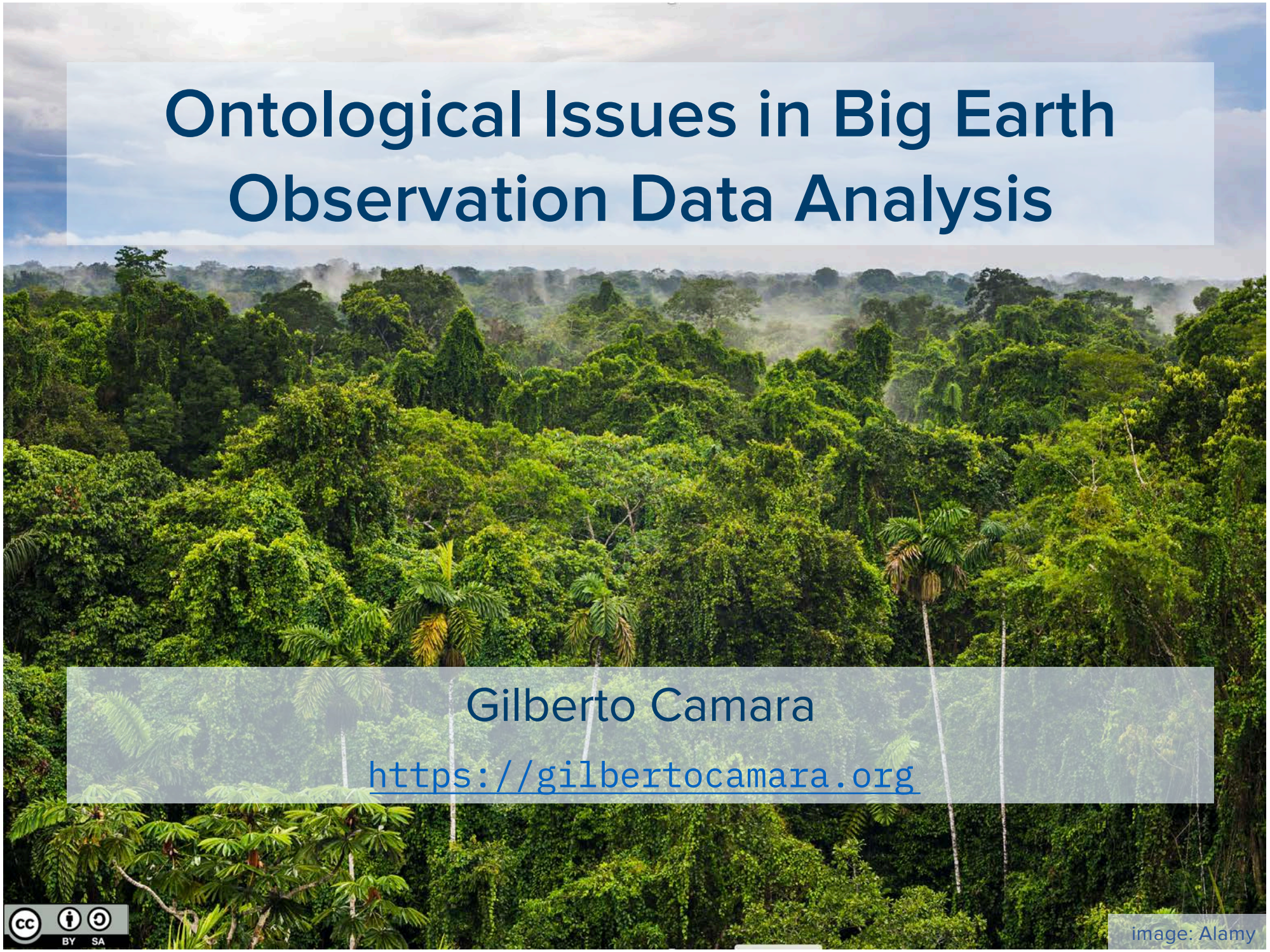
# Ontological Issues in Big Earth Observation Data Analysis



Gilberto Camara

https://gilbertocamara.org



image: Alamy

# Thanks to the INPE team!

Karine

Lúbia

Adeline

Lorena

Gilberto

Isabel

Michelle

Alber

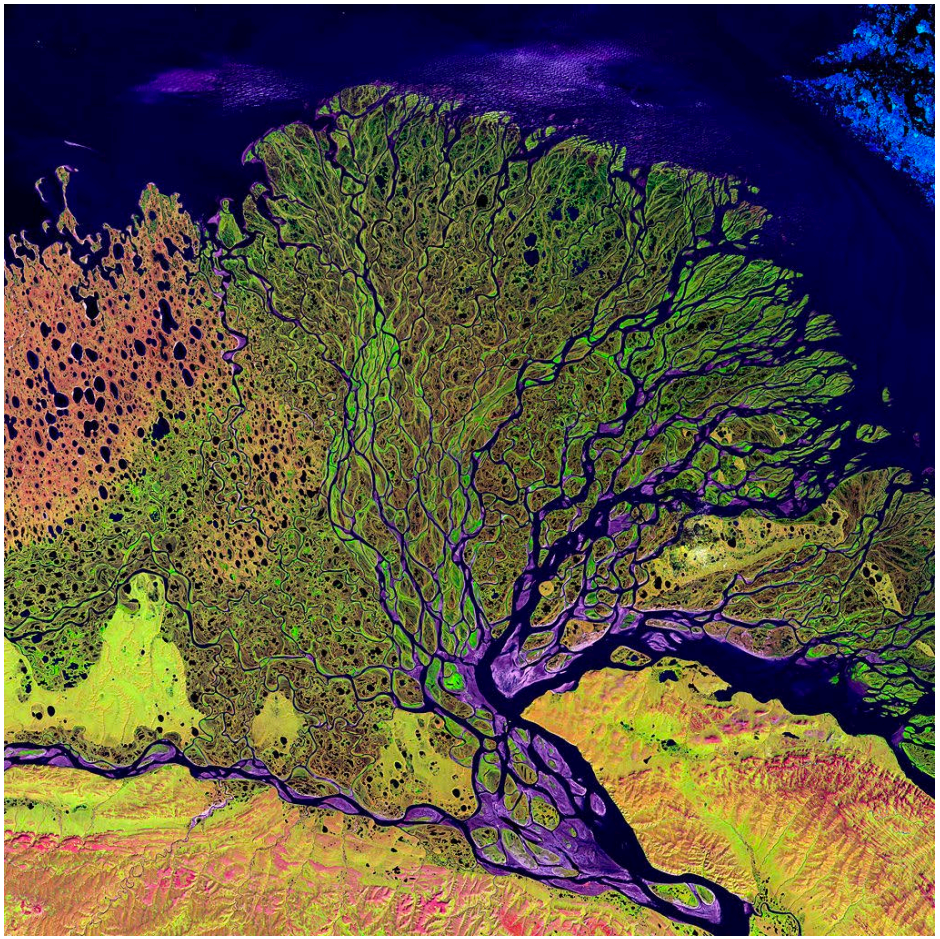
Pedro

Cartaxo

Victor

Rolf

# Sustainable development: knowledge and action



Knowledge: informs us about the limits of our planet

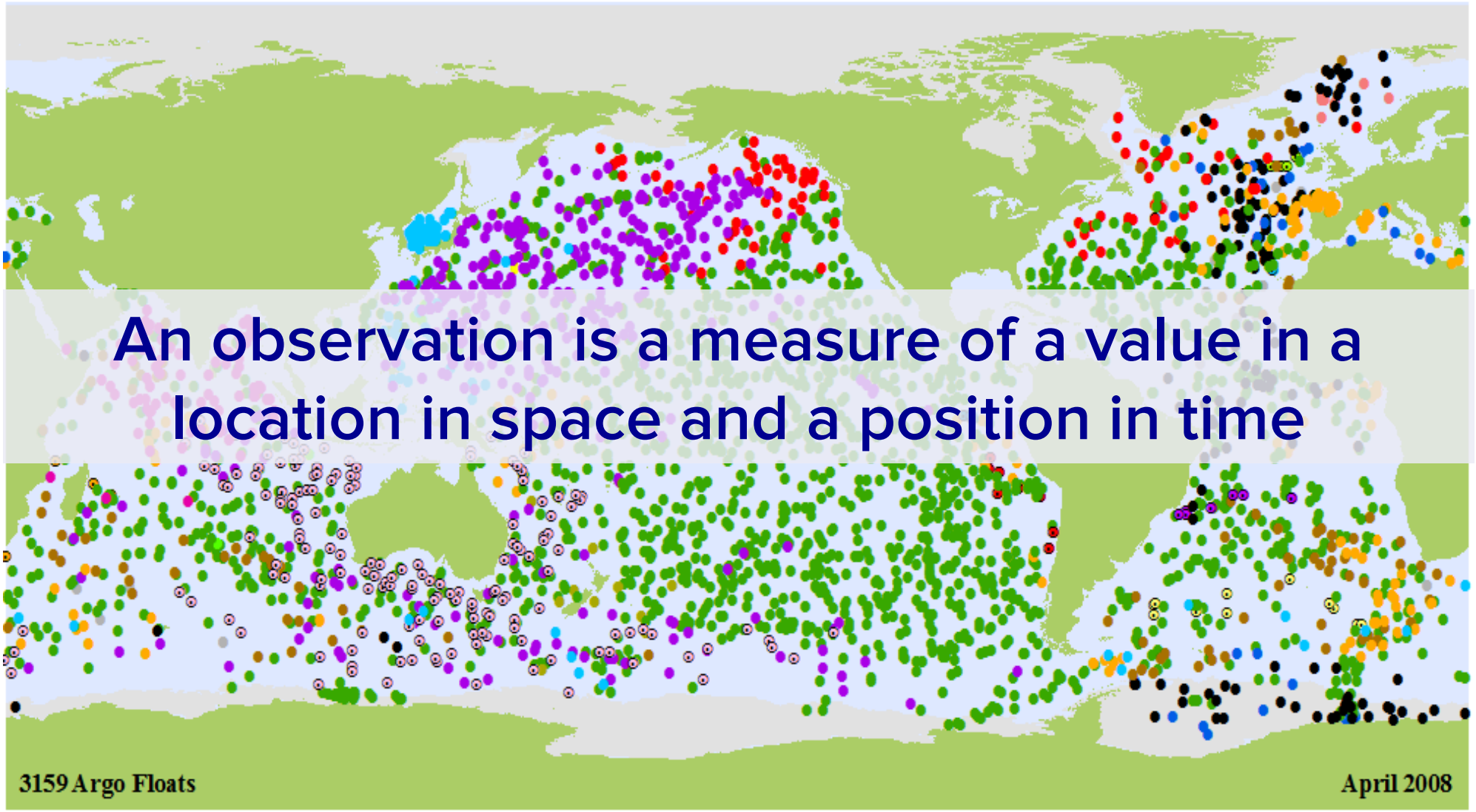
Action: societies decide how to use our planet's resources

images: USGS, Alamy

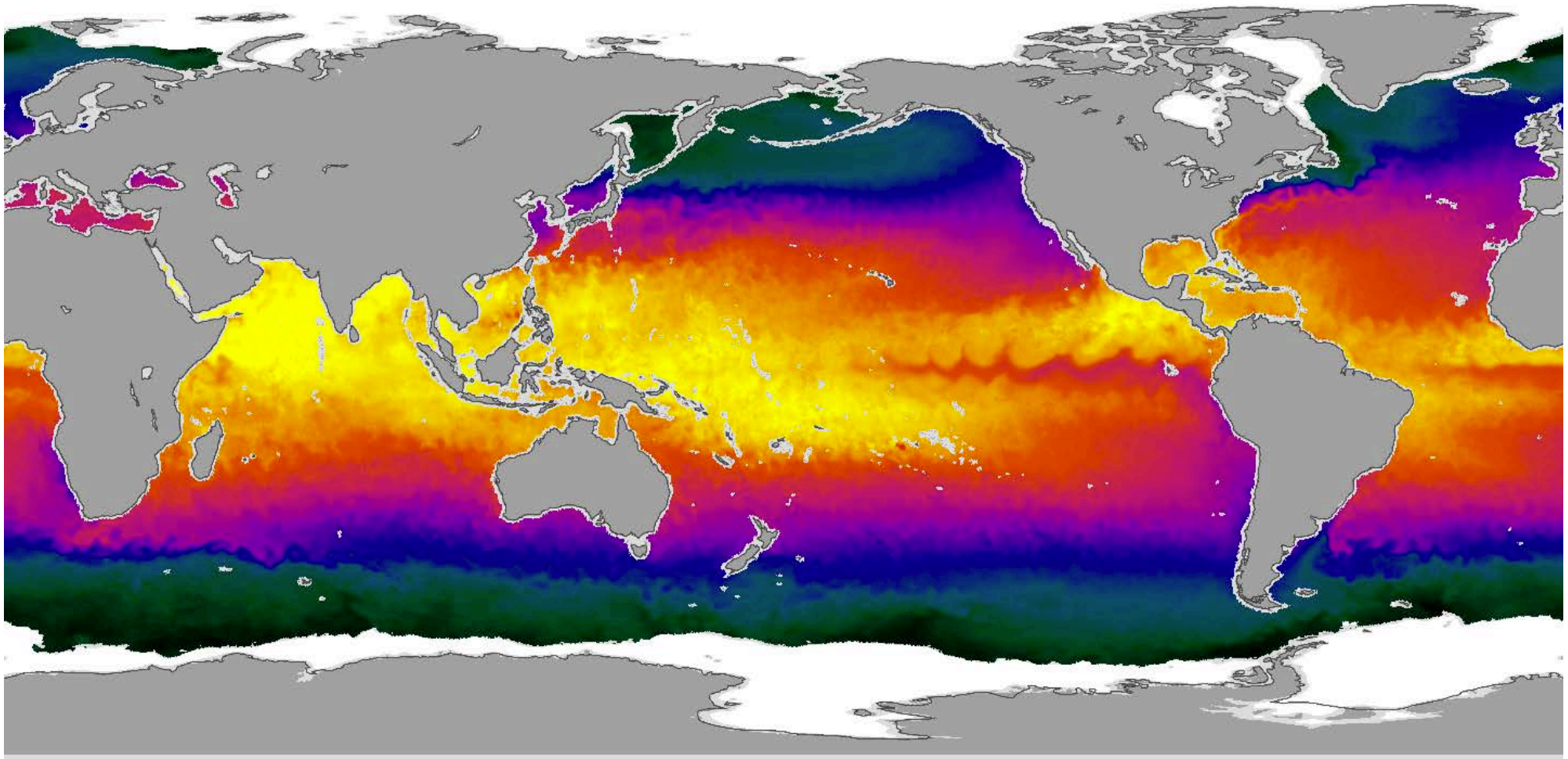
"Ontology aims to reduce the complexity of Nature so our limited intellect can try to understand the Earth's environment" (Ian Jarvis, GEOGLAM)

# It all begins with observations…


We use words in our language to describe the world


An observation is a measure of a value in a location in space and a position in time
3159 Argo Floats
April 2008
ARGENTINA (11)
AUSTRALIA (163)
BRAZIL (7)
CANADA (97)
CHILE (8)
CHINA (11)
COSTA RICA (0)
ECUADOR (3)
EUROPEAN UNION (28)
FRANCE (152)
GERMANY (153)
INDIA (88)
IRELAND (4)
JAPAN (381)
SOUTH KOREA (99)
MAURITIUS (4)
MEXICO (0)
NETHERLANDS (16)
NEW ZEALAND (10)
NORWAY (7)
RUSSIAN FEDERATION (1)
SPAIN (2)
UNITED KINGDOM (101)
UNITED STATES (1813)
jcommops

A field is a space-time continuous function that assigns a unique value to a location in spacetime



Sea-surface temperature

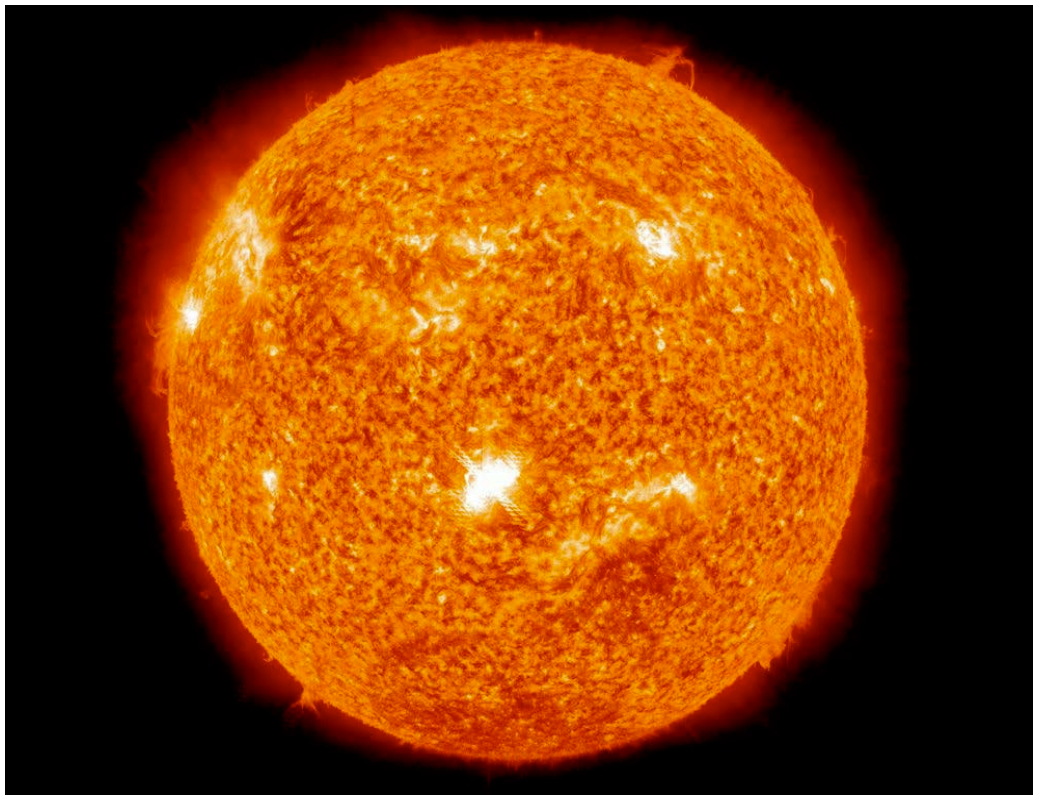
# What about objects?


Objects are language constructs, built upon observations

They require both an external reality and a conscious act to identify their existence

# Geospatial ontology: Barry Smith

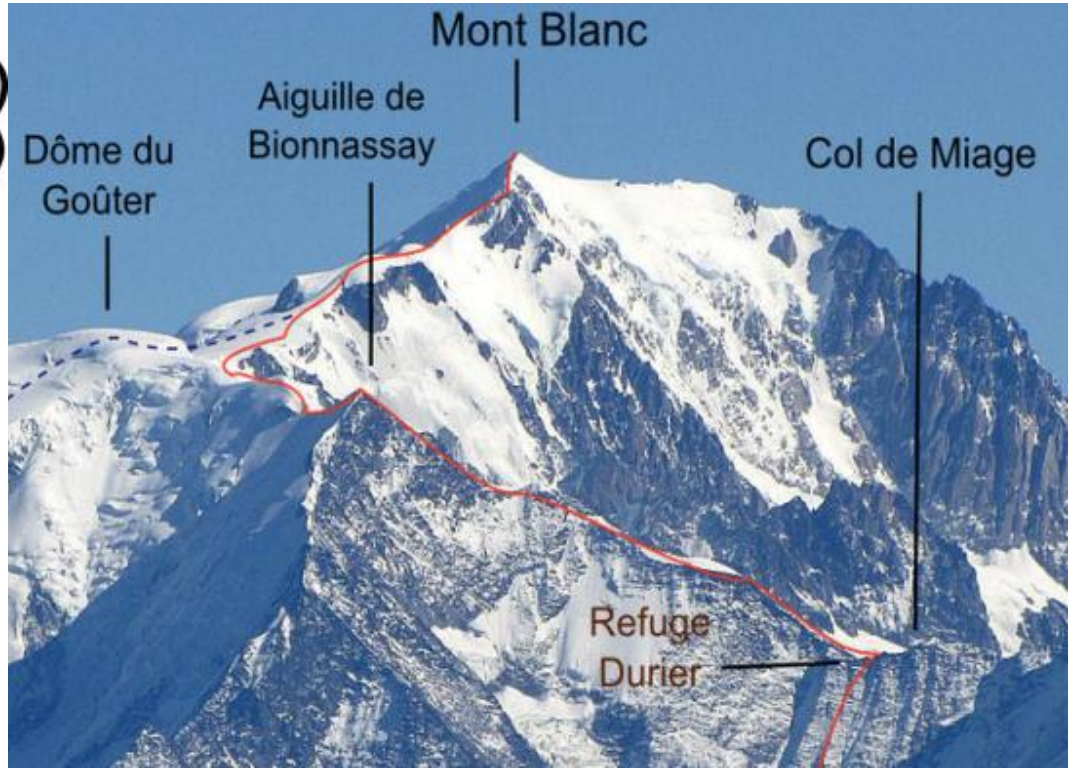
Fiat and bona fide objects



A bona fide object
(the Sun)


A fiat object
(Germany)

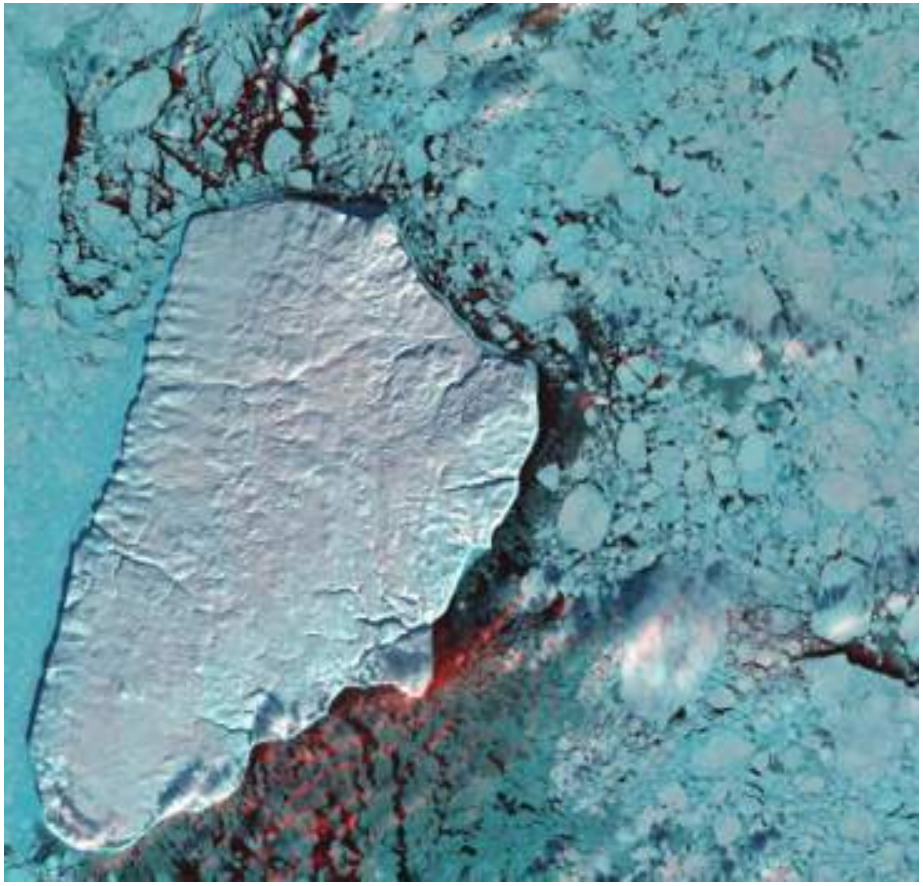
# Objects as mental constructions derived from physical reality


"Mont Blanc" is a socially-accepted name for a specific topographic feature

Where does "Mont Blanc" start and "Dôme du Goûter" end?

# Natural objects are bona fide



What is the size of the coast of England?


CAUTION
DUST STORMS
MAY EXIST

# objects exist, events occur

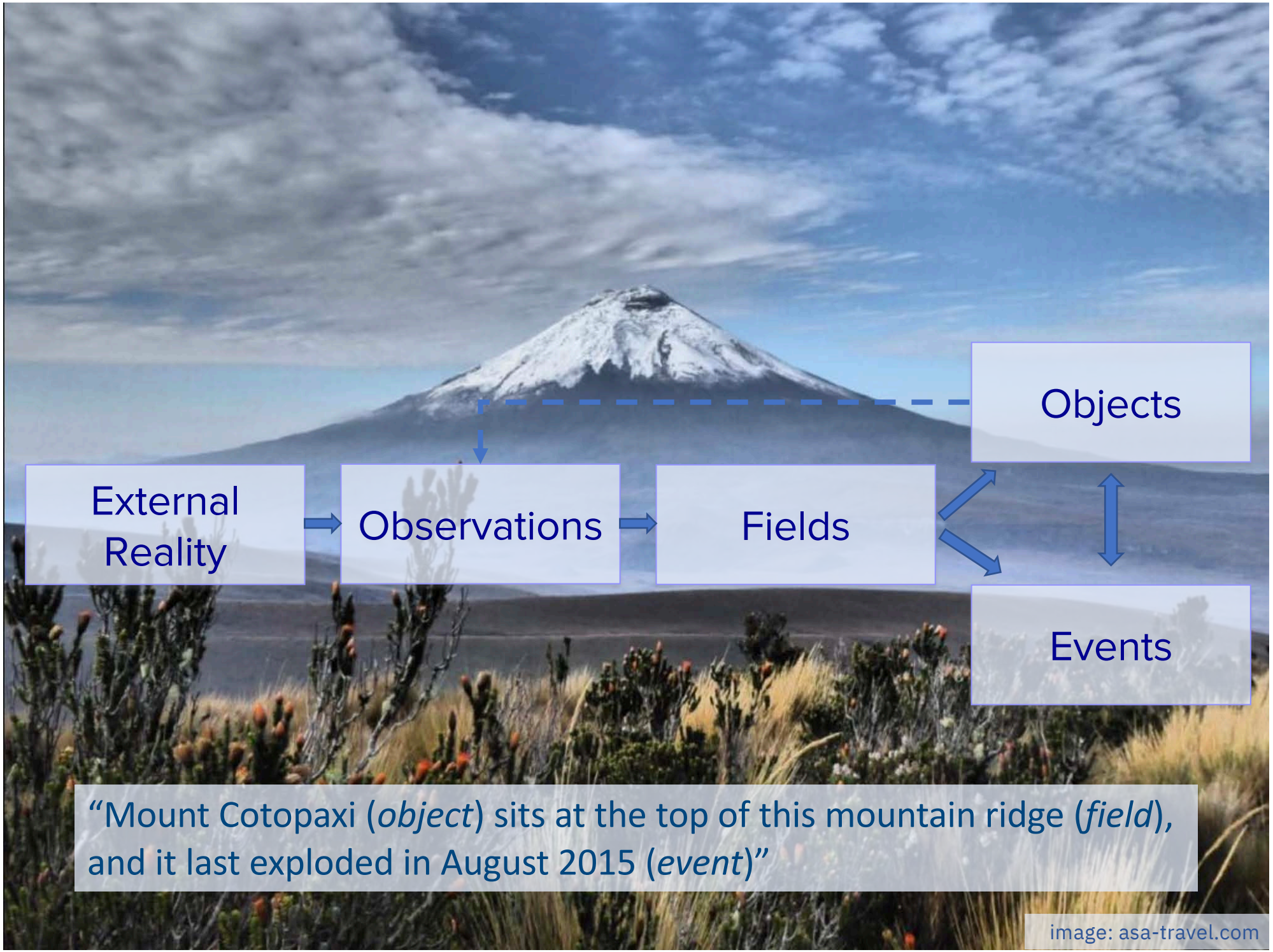
Mount Etna is an object

Etna's  2002 eruption was an event

# Objects exist, events occur


Japan is an object

The 2011 Tohoku tsunami was an event

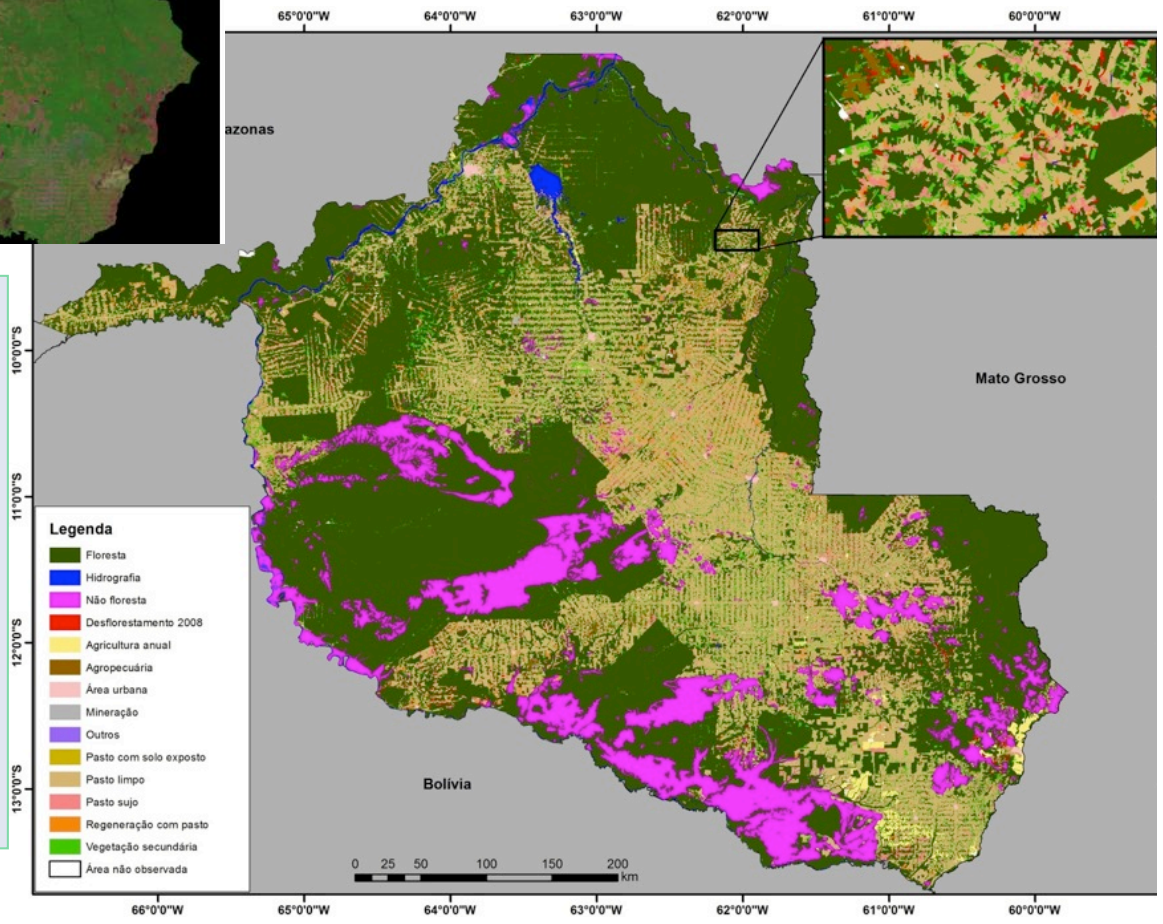

Objects
External Reality
Observations
Fields
Events
"Mount Cotopaxi (*object*) sits at the top of this mountain ridge (*field*), and it last exploded in August 2015 (*event*)"
image: asa-travel.com

# The quest for the perfect map

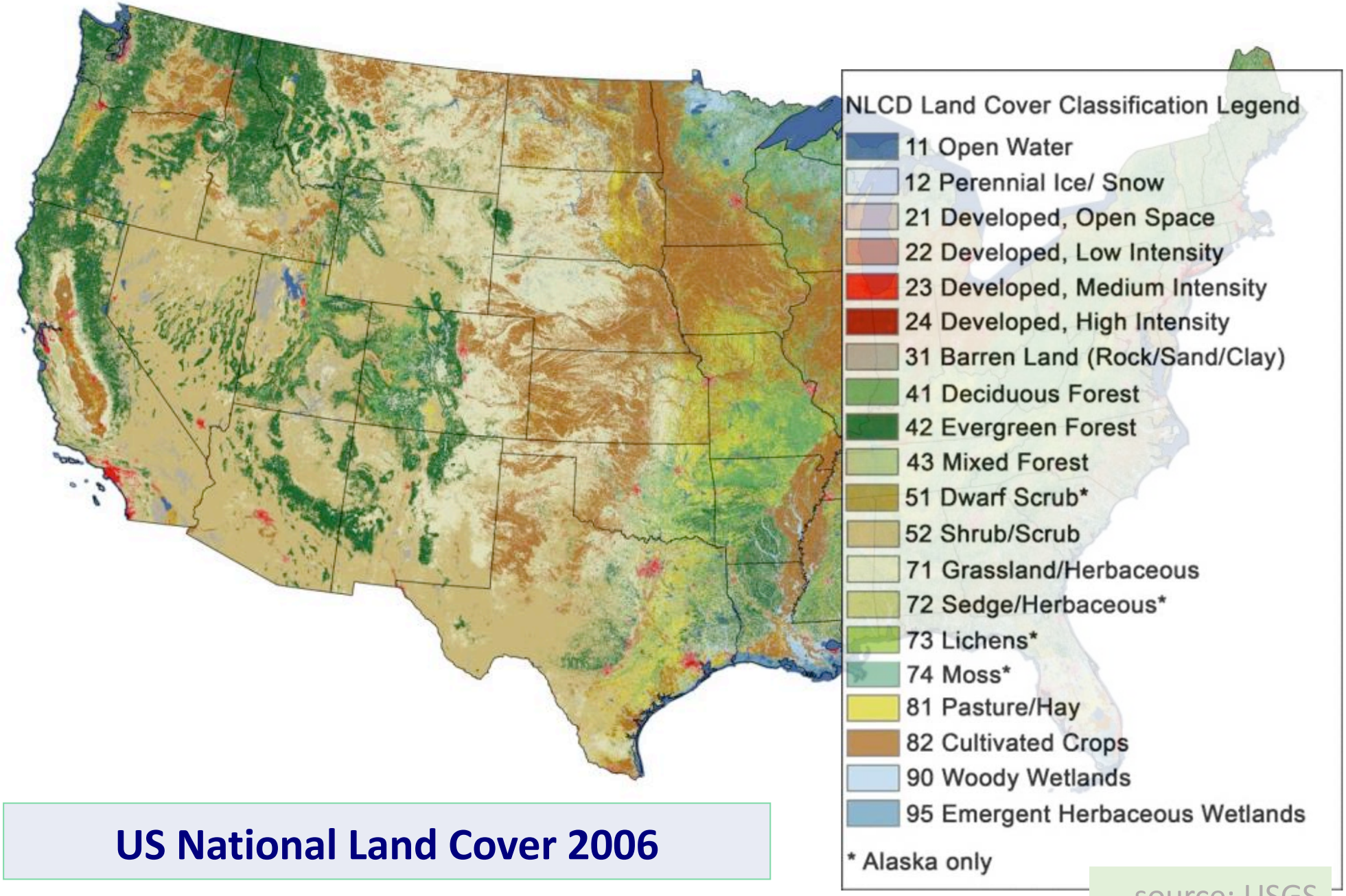
The world is divided in cells.
Each cell has a single class.

There is a correct classification.

The more our classification approaches the ideal, the better.



# In search of the perfect map




**US National Land Cover 2006**

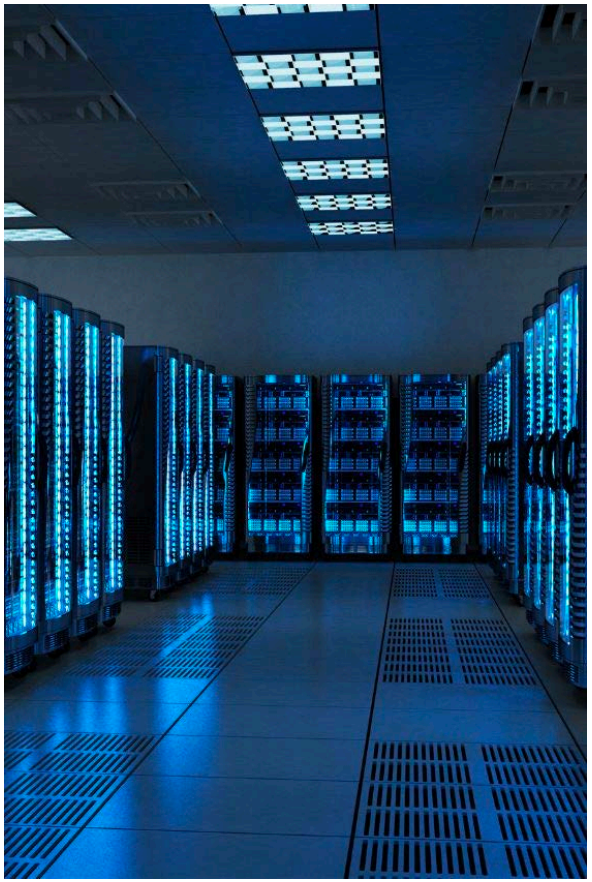
source: USGS

**Grassland/Herbaceous** - areas dominated by gramanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.

## Land cover or land use?

**Pasture/Hay** – areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20% of total vegetation.

# What has changed in the 21st century?



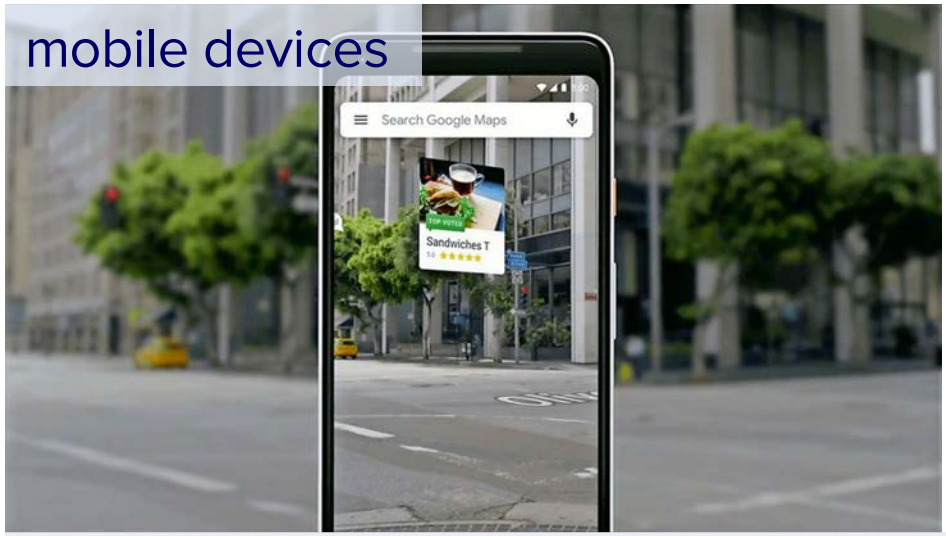
Low access cost

data services

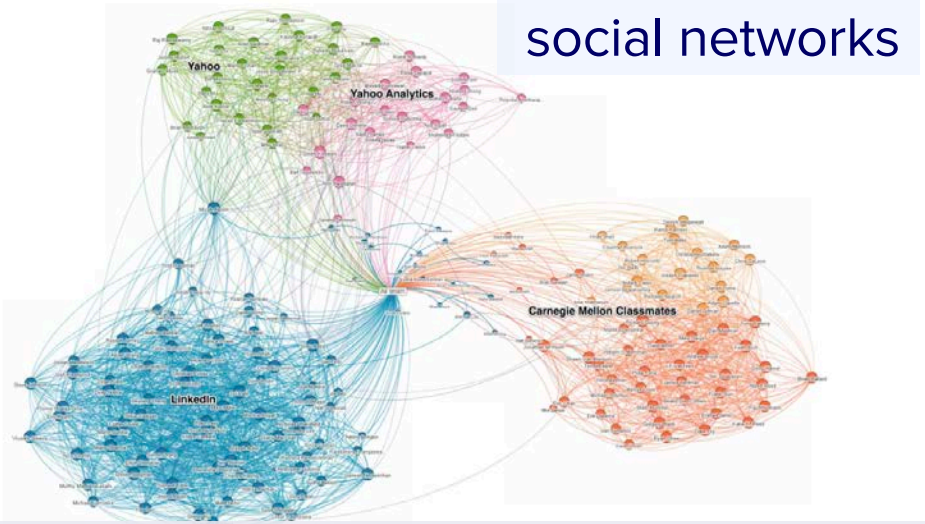
massive use

The new digital economy: data services, social networks, cloud computing, machine learning

images: shutterstock




mobile devices




social networks

# Mobile devices, social network, big Earth observation sets: new technologies, new challenges


sensors everywhere

images: Economist, USGS, LinkedIn, tagesanzeiger.ch

lots of images

WIRED

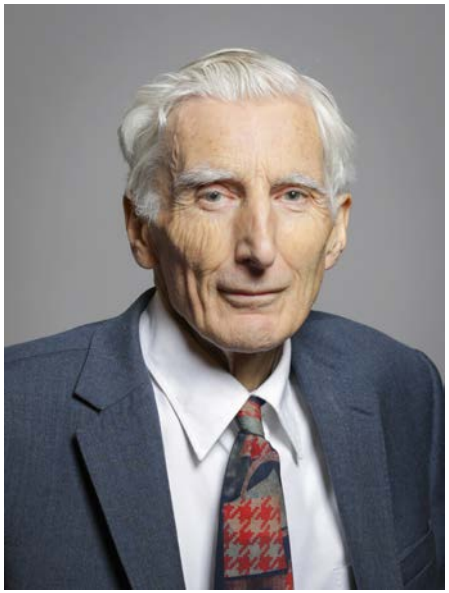
# The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

George Box (Univ Wisconsin): “All models are wrong, but some are useful”

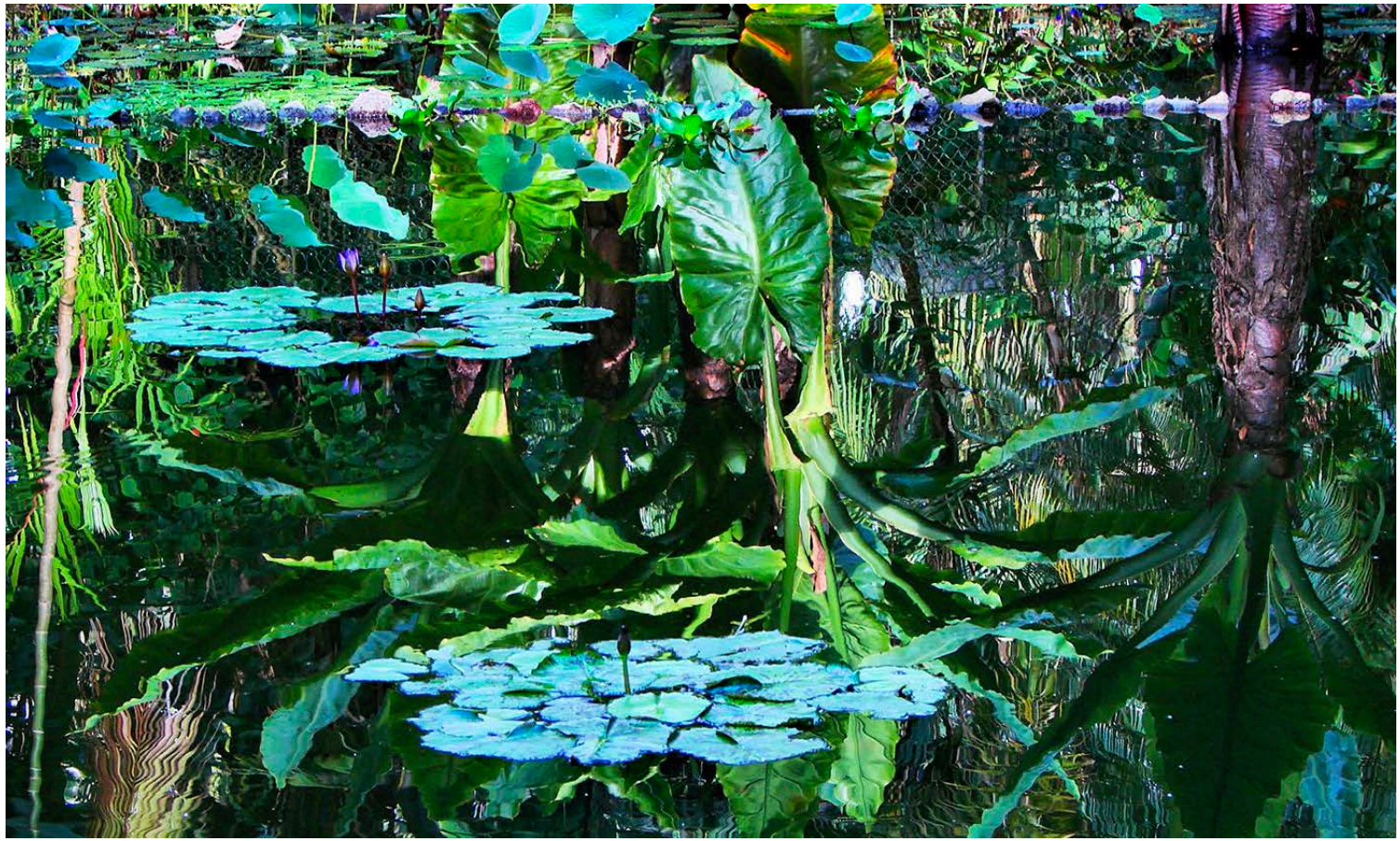
Chris Anderson (Wired): “What can Science learn from Google?”

Peter Norvig (Google): “All models are wrong, and increasingly you can succeed without them”.

# “Black holes are simpler than forests and Science has its limits”

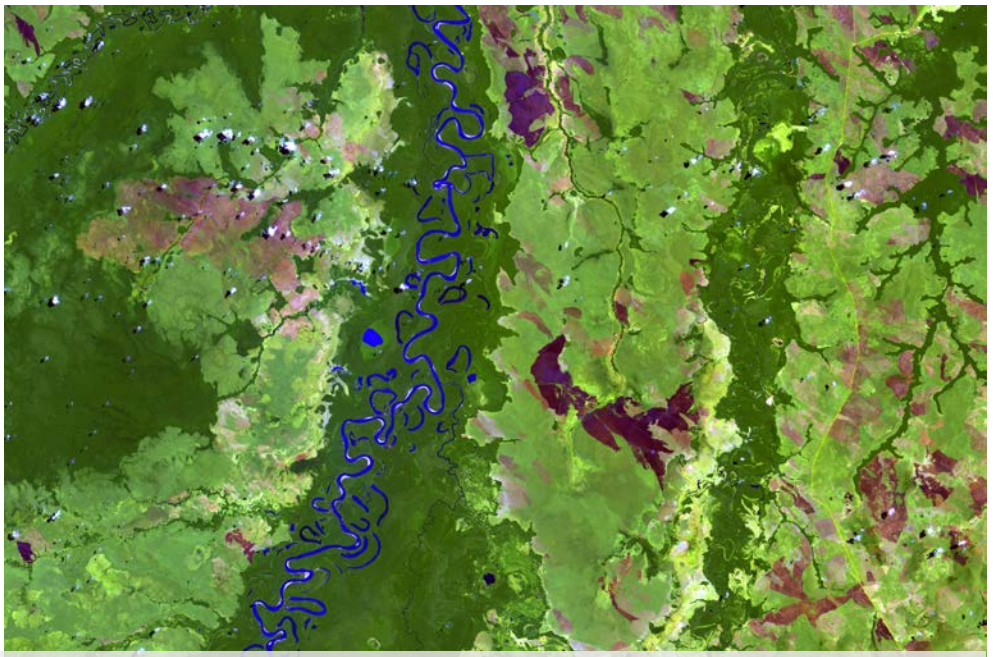


Sir Martin Rees
(Astronomer Royal)

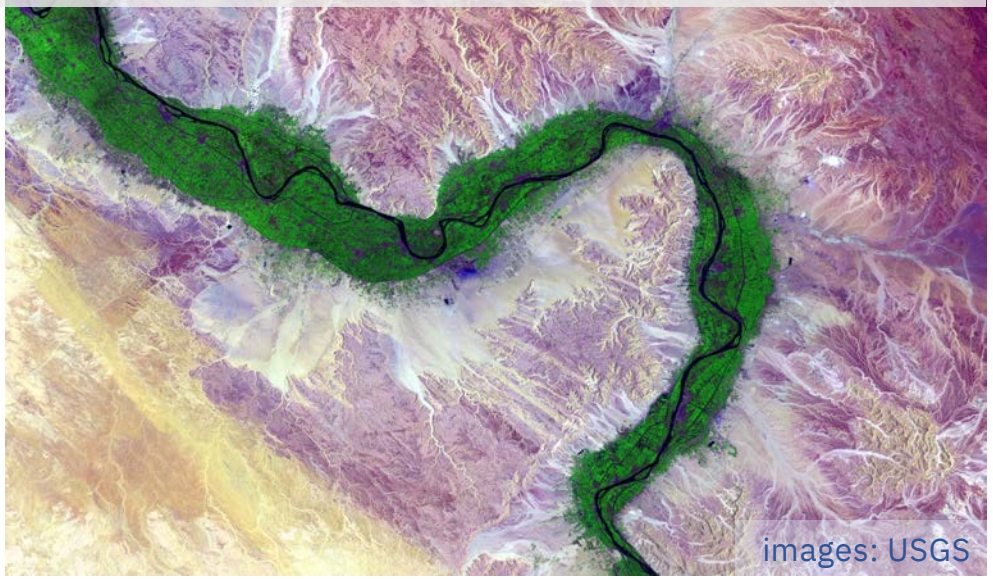


“More is different” (Anderson, Science 1972)

aeon



# What is the semantic content of information measured by big Earth observation data?

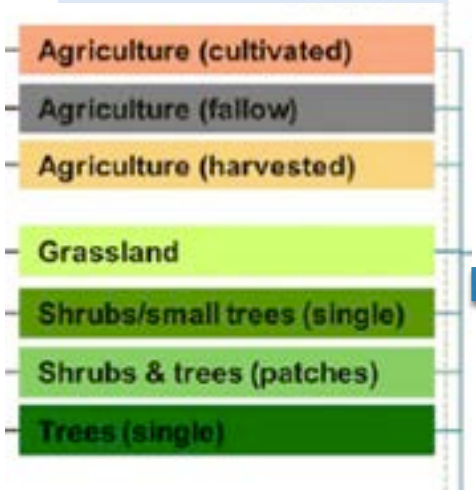


images: USGS

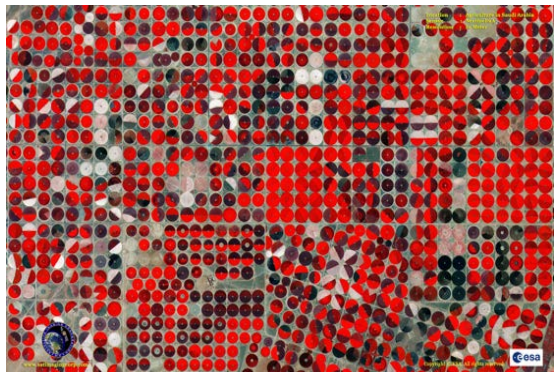
# From reality to representation

reality

ontologies

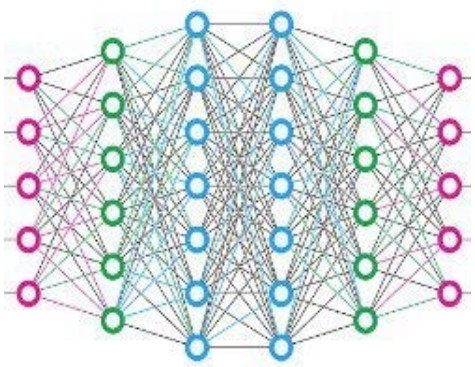



classification



measurements

Data organisation



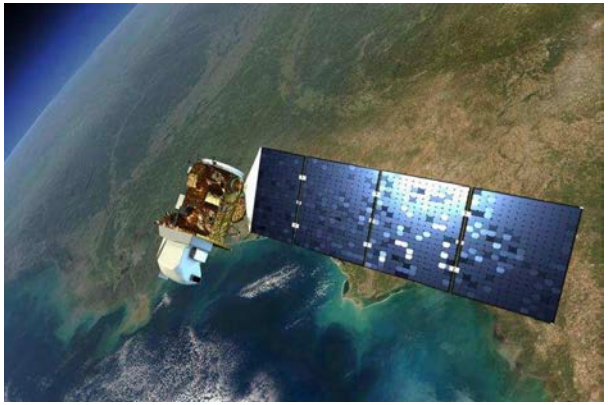
Data analysis

# What are the limits of machine learning?

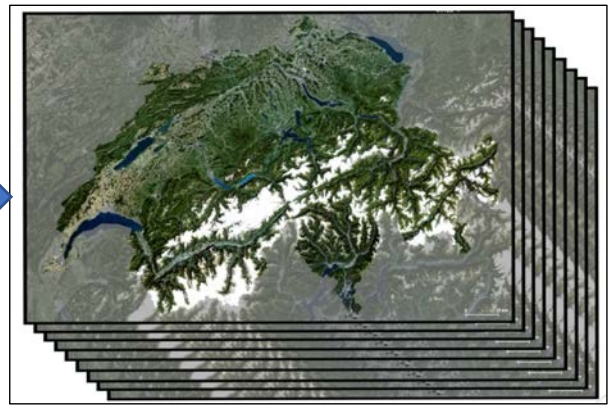
What works for face recognition, automatic translation, and Chess/Go games does also work for big spatial data?

image: manmax31.github.io

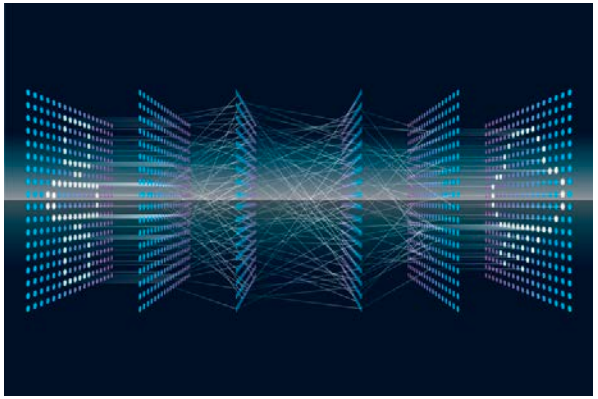
# Big spatial data for sustainable development: Earth observation images



Big satellite data
(2PB per day)



Data cubes
(space-time fields)

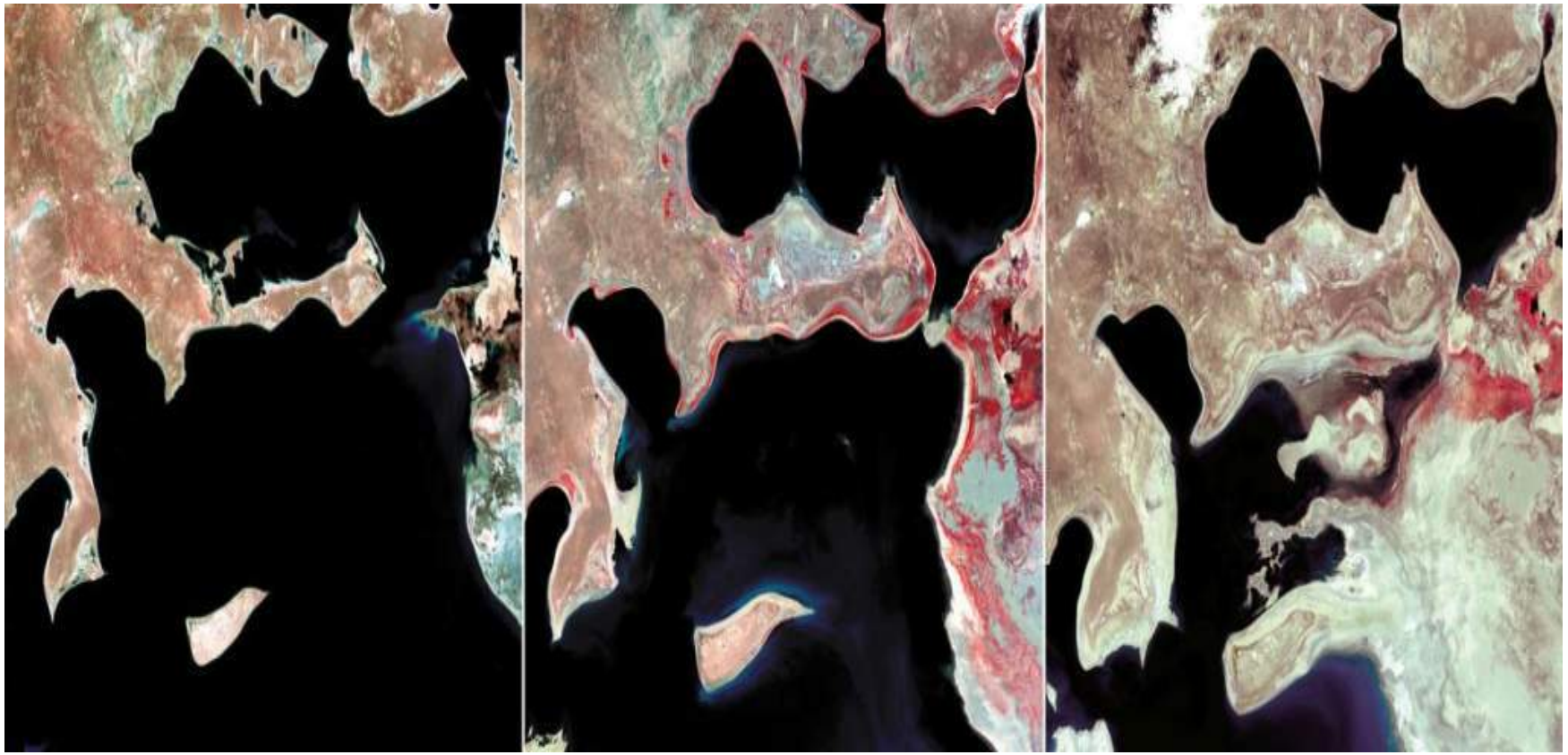


Machine learning
(classification)

In what ways is big spatial data special?

images: NASA, Swiss Data Cube, ECMWF

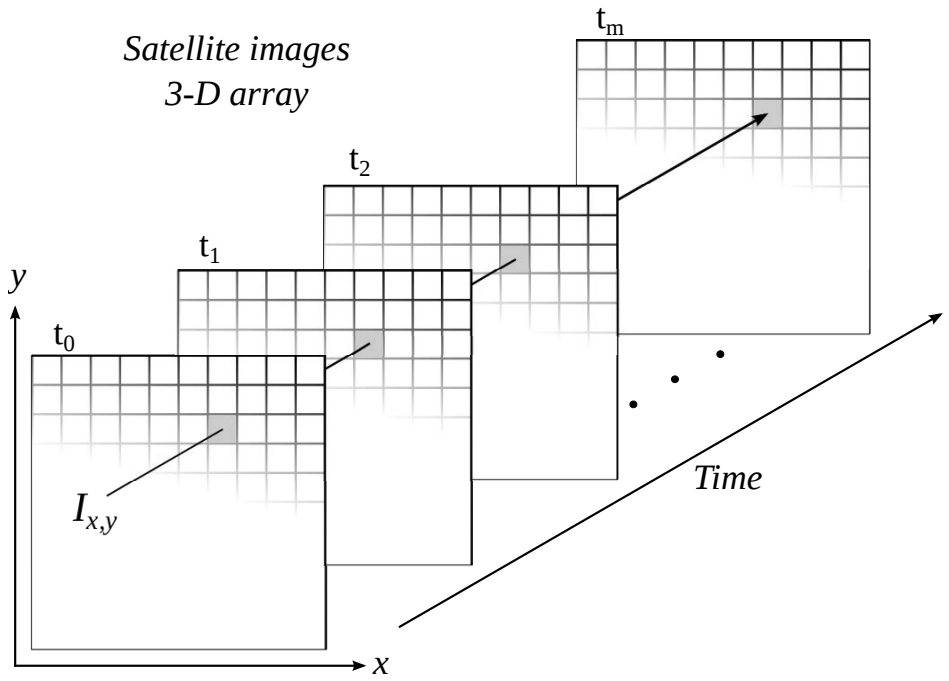
# What's in an image?



"A remote sensing image is a measurement that captures snapshots of change trajectories. The focus of the ontological characterization of images should be on searching for changes instead of searching for content." (Camara et al, COSIT 2001)

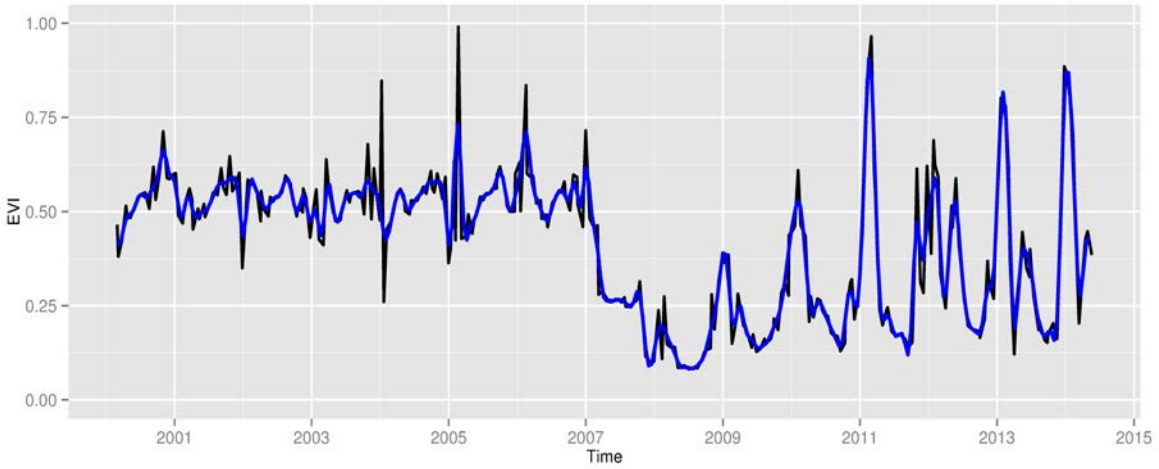
images: USGS

# Space first, time later or time first, space later?



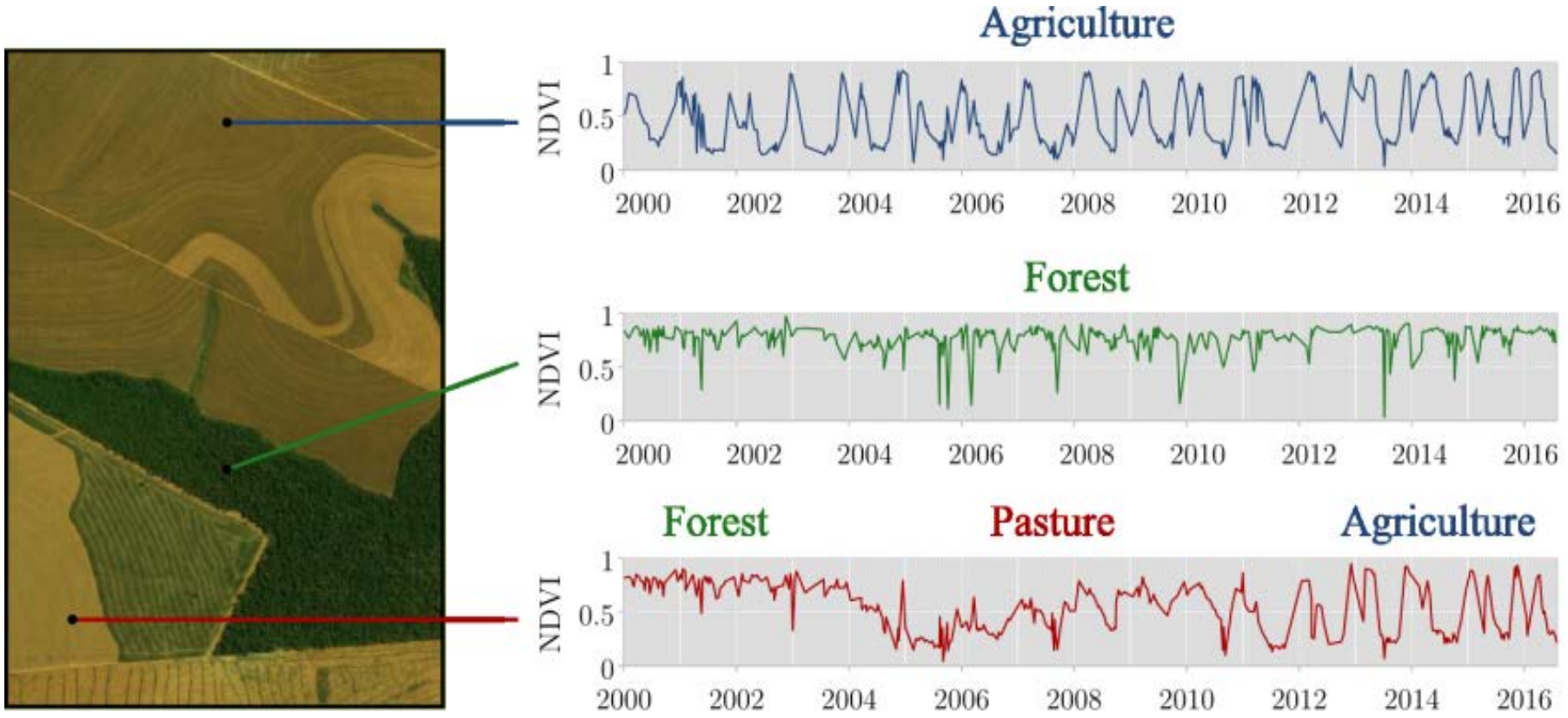

Space first: classify images; compare results in time

Time first: classify time series; join results to get maps




graphics: V Maus, author

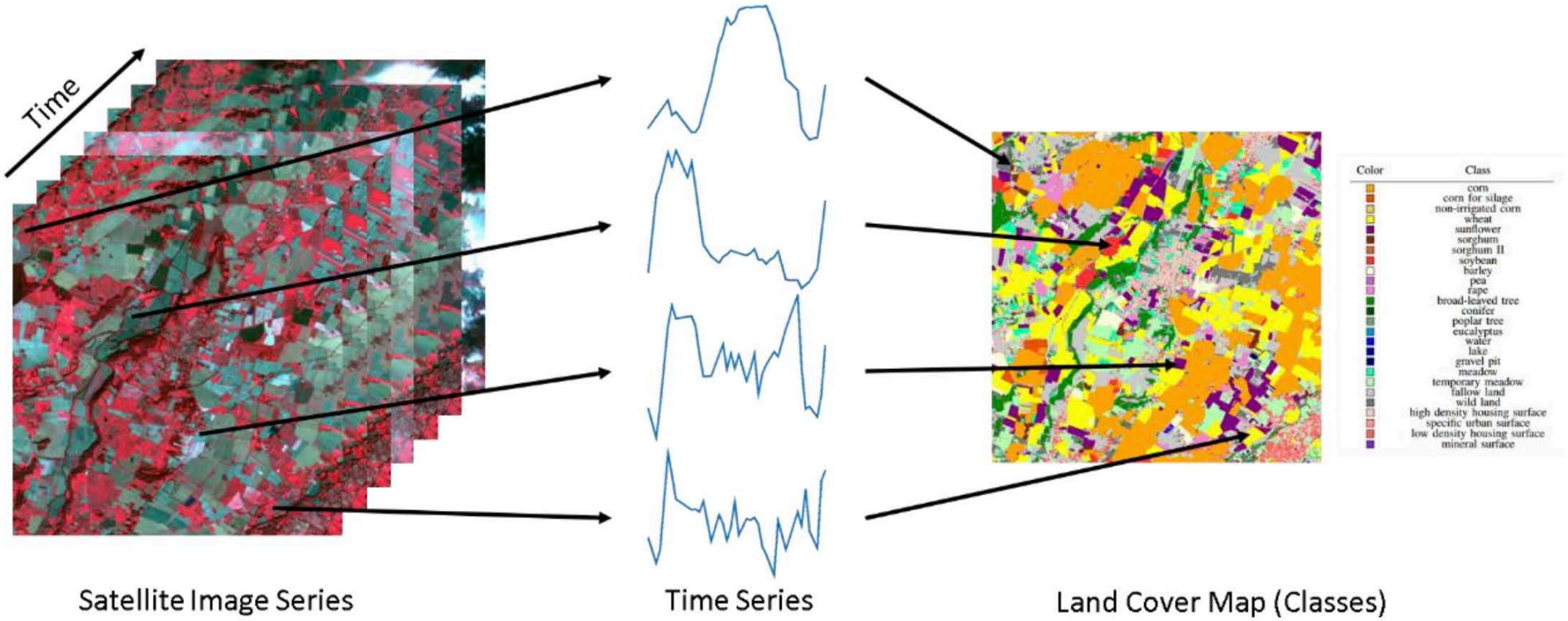
# Land-use change trajectories measured by time series




"Transformations of land cover due to actions of land use"
(continuous monitoring of landscapes)

graphics: Rolf Simoes (INPE)

# Using time series to estimate land use and land cover change




Satellite Image Series

Time Series

Land Cover Map (Classes)

1. Extract samples from a data cube
2. Train a machine learning model
3. Classify all time series for a chosen period

image: F Petitjean, Monash Univ

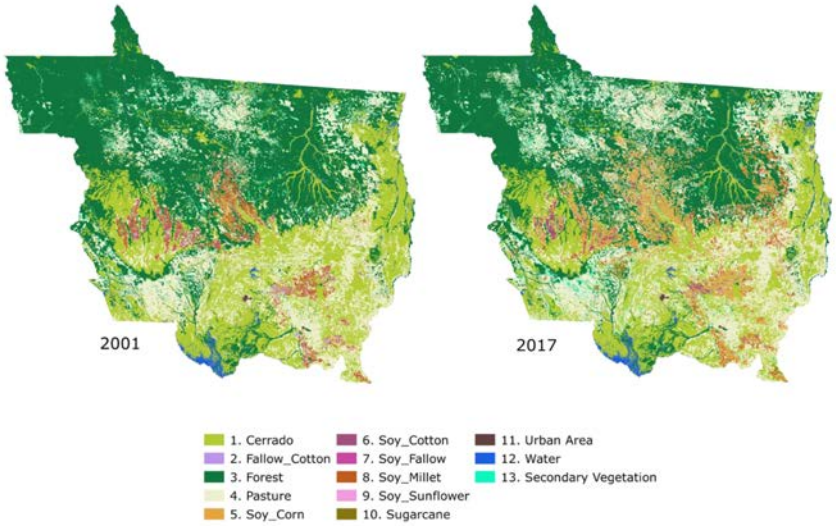
# Finding: Machine learning is useful for big EO data, but...

SCIENTIFIC DATA

Data Descriptor | Open Access | Published: 27 January 2020

## Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017

Rolf Simoes, Michelle C. A. Picoli, Gilberto Camara, Adeline Maciel, Lorena Santos, Pedro R. Andrade, Alber Sánchez, Karine Ferreira & Alexandre Carvalho



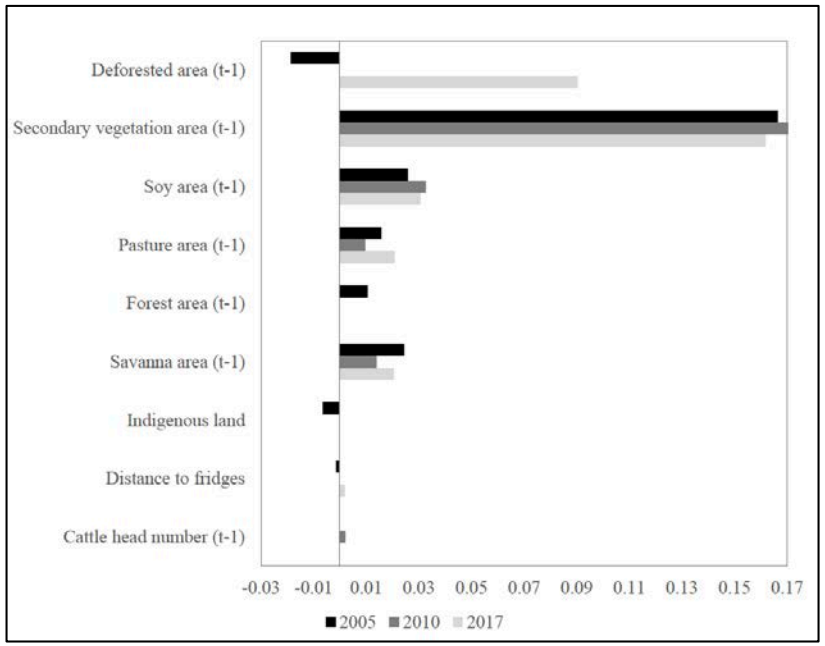

land

MDPI

*Article*

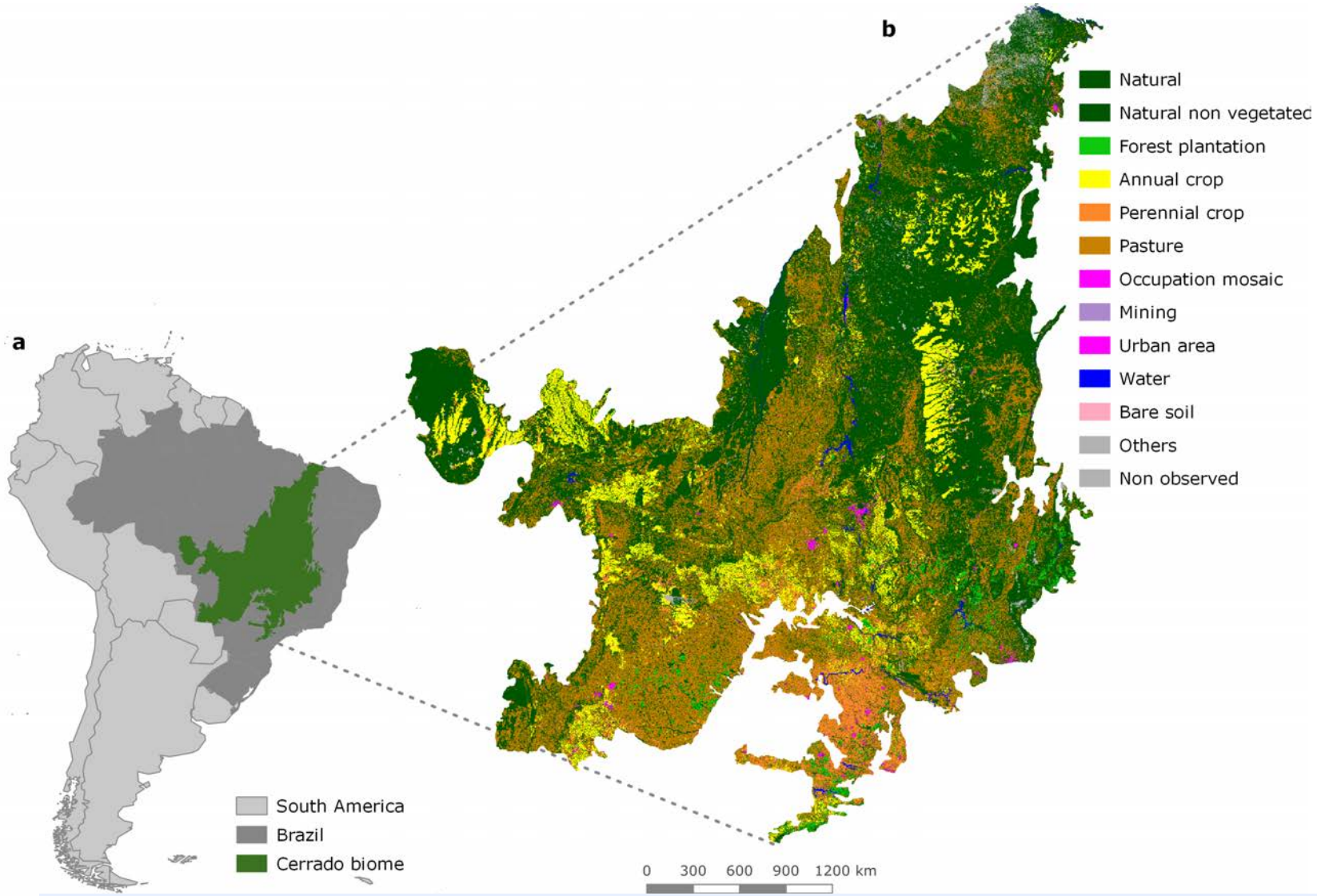
## Impacts of Public and Private Sector Policies on Soybean and Pasture Expansion in Mato Grosso—Brazil from 2001 to 2017

Michelle C. A. Picoli [1,*,†], Ana Rorato [1,†], Pedro Leitão [2,3,†], Gilberto Camara [1,4,†], Adeline Maciel [1,†], Patrick Hostert [3,5,†] and Ieda Del'Arco Sanches [1]




Pasture expansion

# Cerrado: Brazilian Savanna

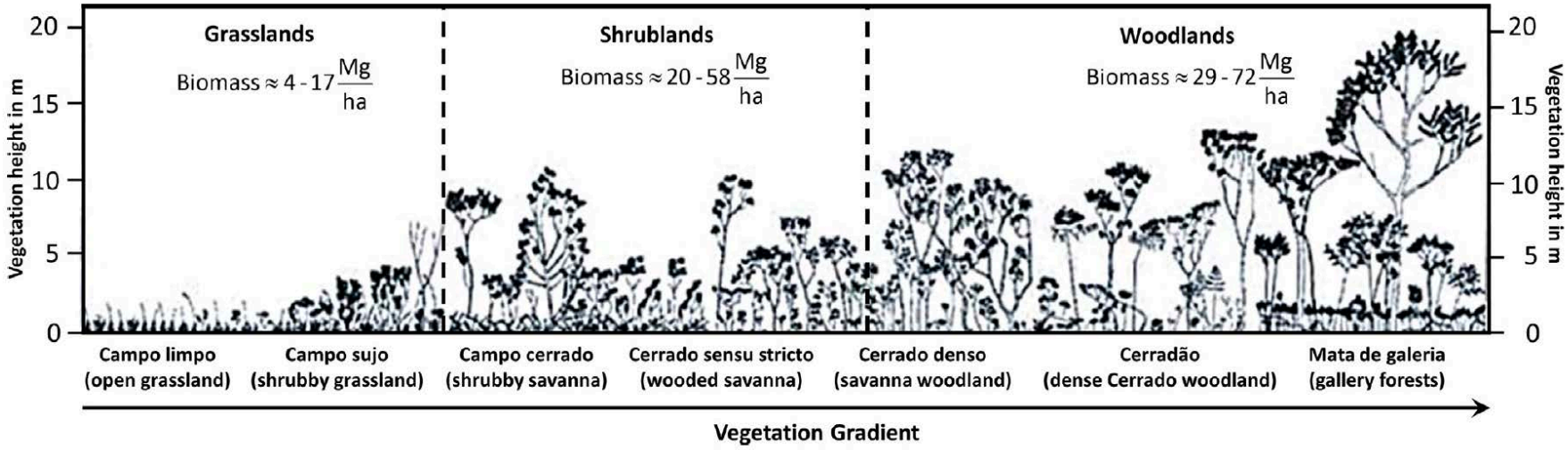



covers 2 million km$^2$, latitude ranges from 5$^0$S to 25$^0$S

image: TerraClass, INPE/EMBRAPA

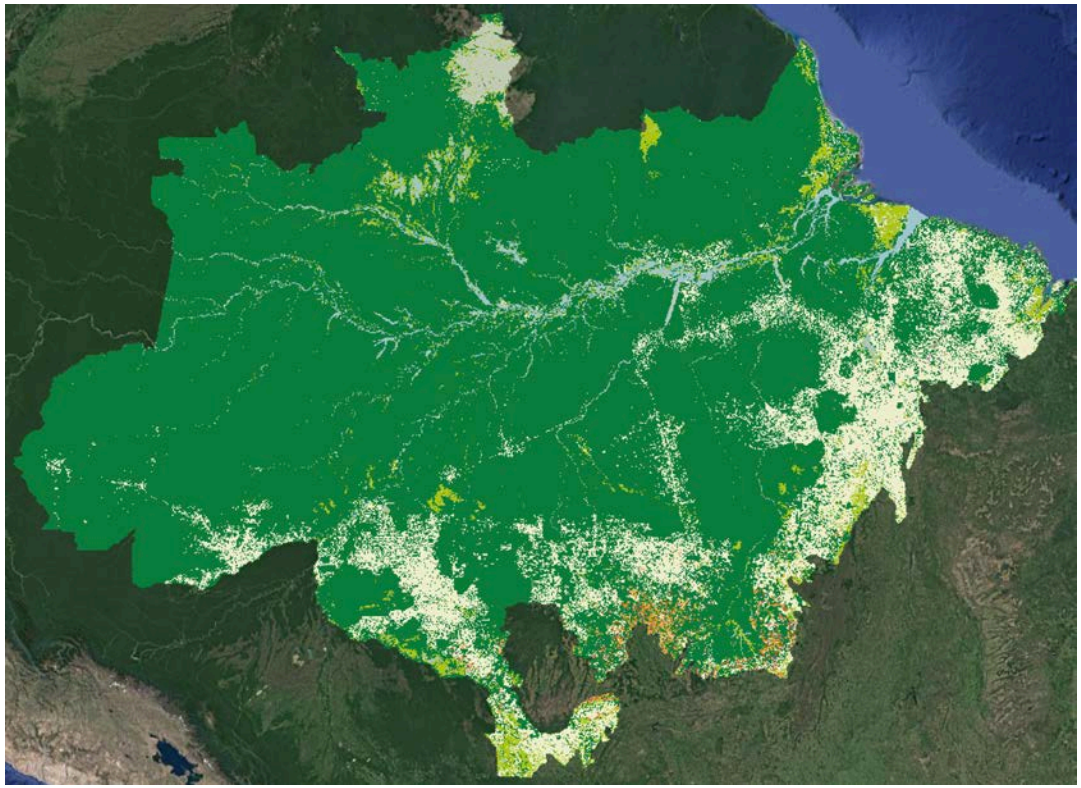
# Cerrado: Brazilian savanna




images: Del-Claro and Silingardi (Annals Brazilian Academy Sci, 2019 ), Schwieder et al. (IJAEOG, 2016)

# Too good to be true?



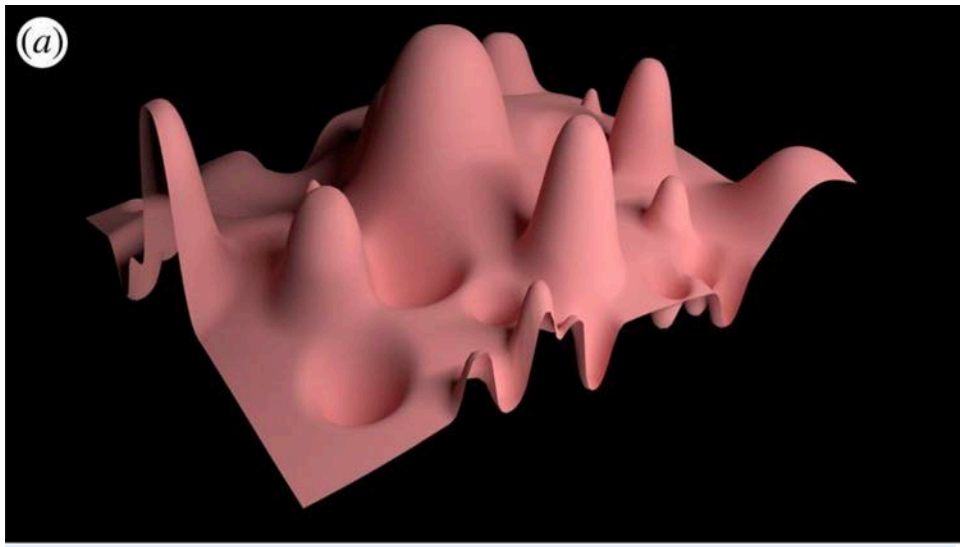
Brazilian Amazonia – MODIS data
33,000 samples

| Model | 5-fold validation accuracy |
|---|---|
| SVM | 97.6% |
| Random Forest | 98.5% |
| Perceptron | 99.2% |
| FCNN | 98.9% |
| tempCNN | 99.1 % |
| ResNet | 99.0% |

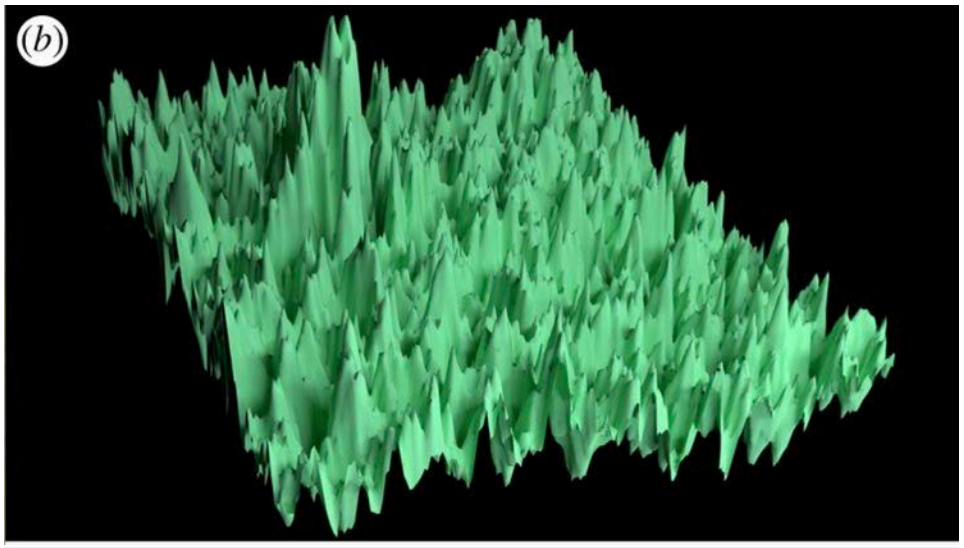
Models fit samples well, but do they represent reality?

data: Simoes and Camara (2020)

# Machine learning in simulated landscapes




Small variability: easy for ML

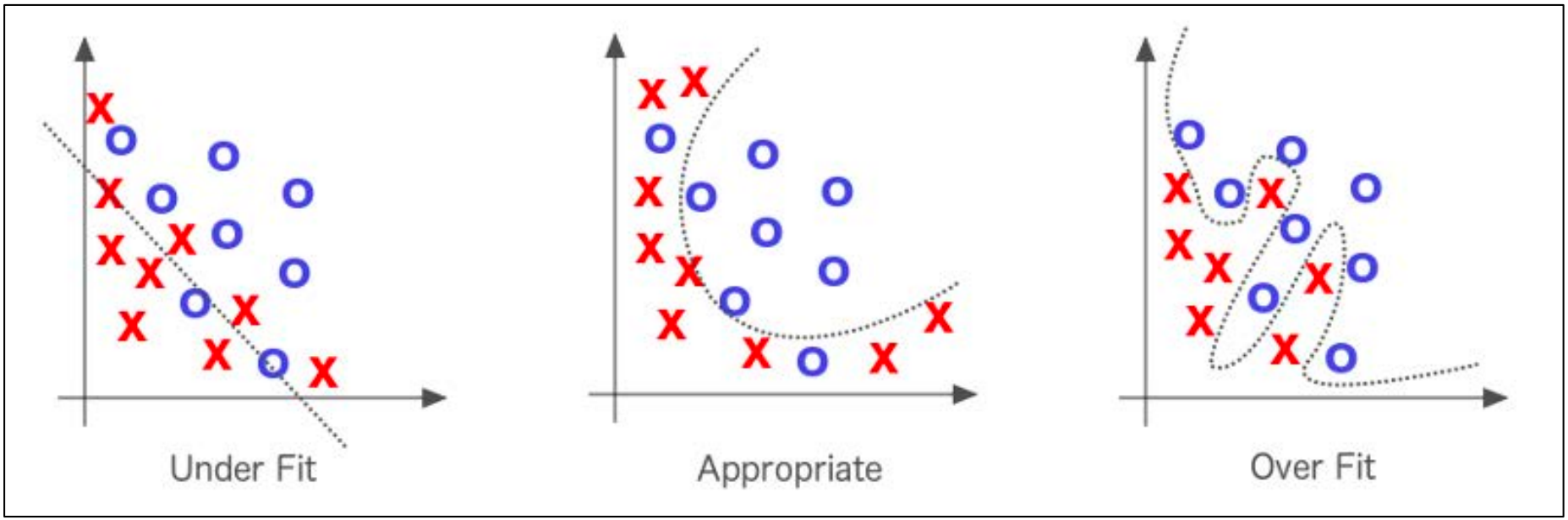



Large variability: hard for ML

Machine learning relies on optimization and works best on problems with low variability

source: Coveney et al., “Big data need big theory too” (2017)

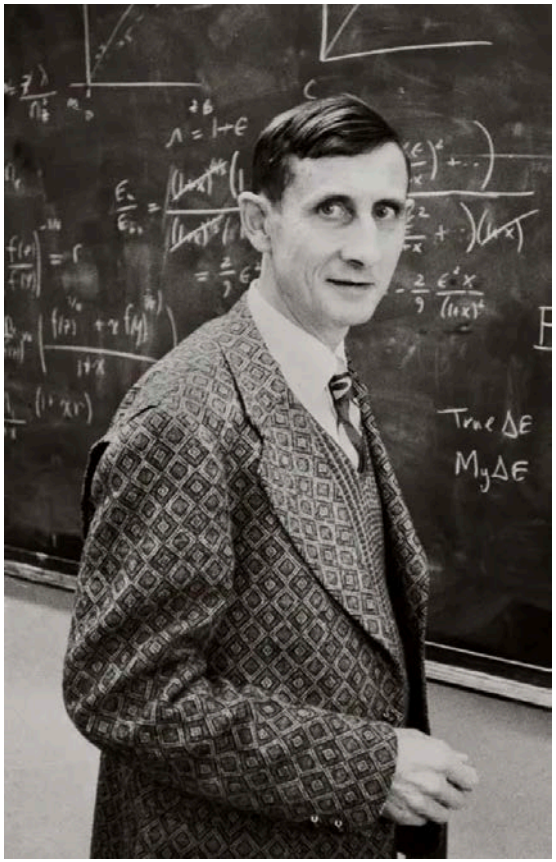
# The ghost of overfitting




SVM (support vector machines) have seven (7) tuning parameters. Other machine learning methods are similar. Overfitting is easy to occur.

graphics: Patni (medium.com)

# Data-rich, theory poor



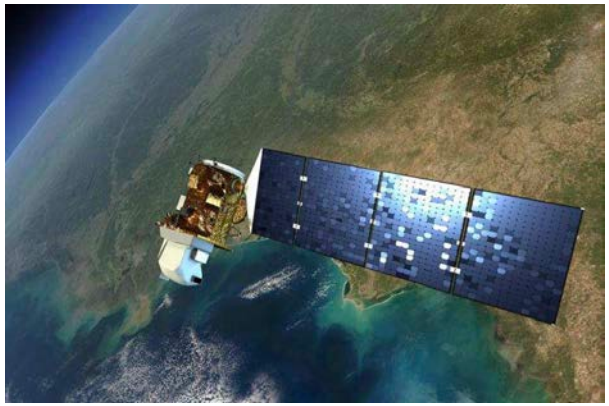

Freeman Dyson

Enrico Fermi

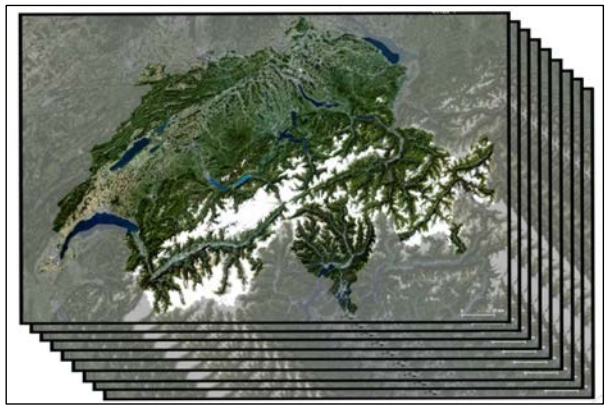
Fermi: How many parameters did you use in your calculations?
Dyson: Four.
Fermi: My friend John von Neumann used to say that with four parameters, I can fit an elephant and with five, I can make him wiggle his trunk.

images: NYT, Wikipedia

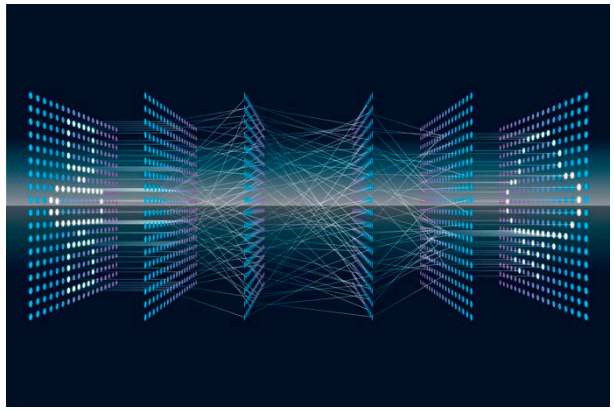
# The elephant in the room



Big satellite data
(2PB per day)



Data cubes
(space-time fields)

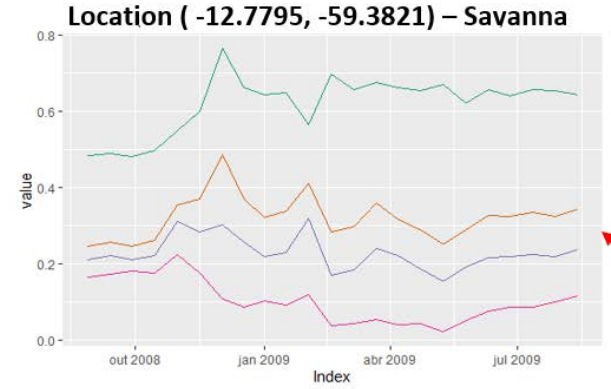


Machine learning
(classification)

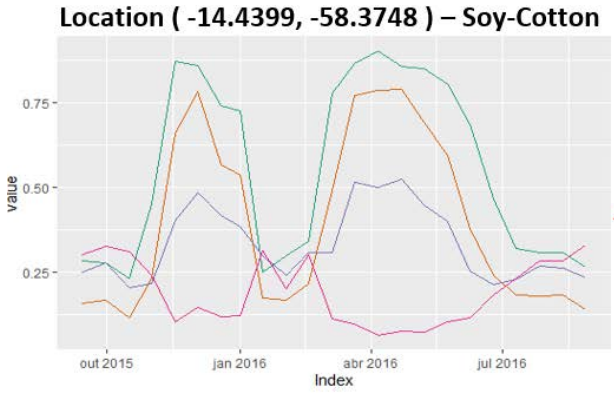
Ecosystems are highly variable
Local knowledge is essential

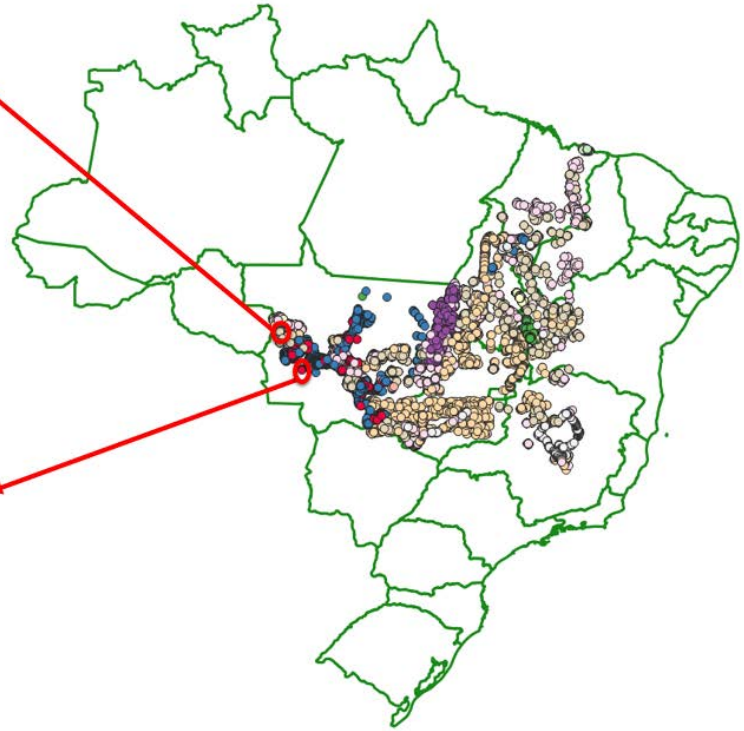
images: Alamy, NASA, Swiss Data Cube, ECMWF

# How to find out if our samples are good?




| longitude | latitude | start_date | end_date | class |
|---|---|---|---|---|
| -58.3748 | -14.4399 | 2015-08-29 | 2016-08-13 | Soy-Cotton |
| -59.3821 | -12.7795 | 2008-08-29 | 2009-08-13 | Savanna |

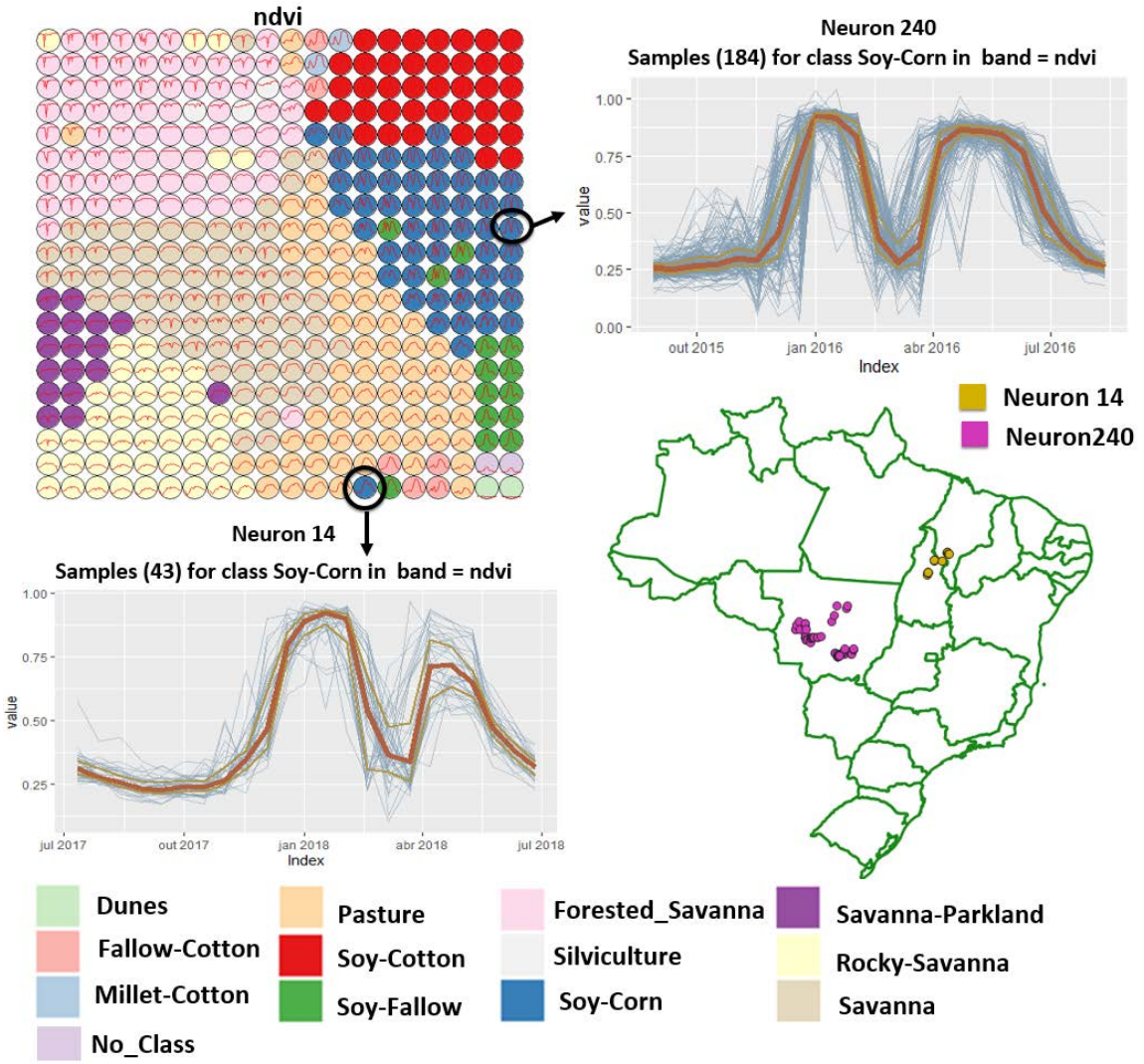


Dunes
Fallow-Cotton
Millet-Cotton
Pasture
Soy-Cotton
Soy-Fallow
Forested_Savanna
Silviculture
Soy-Corn
Savanna-Parkland
Rocky-Savanna
Savanna

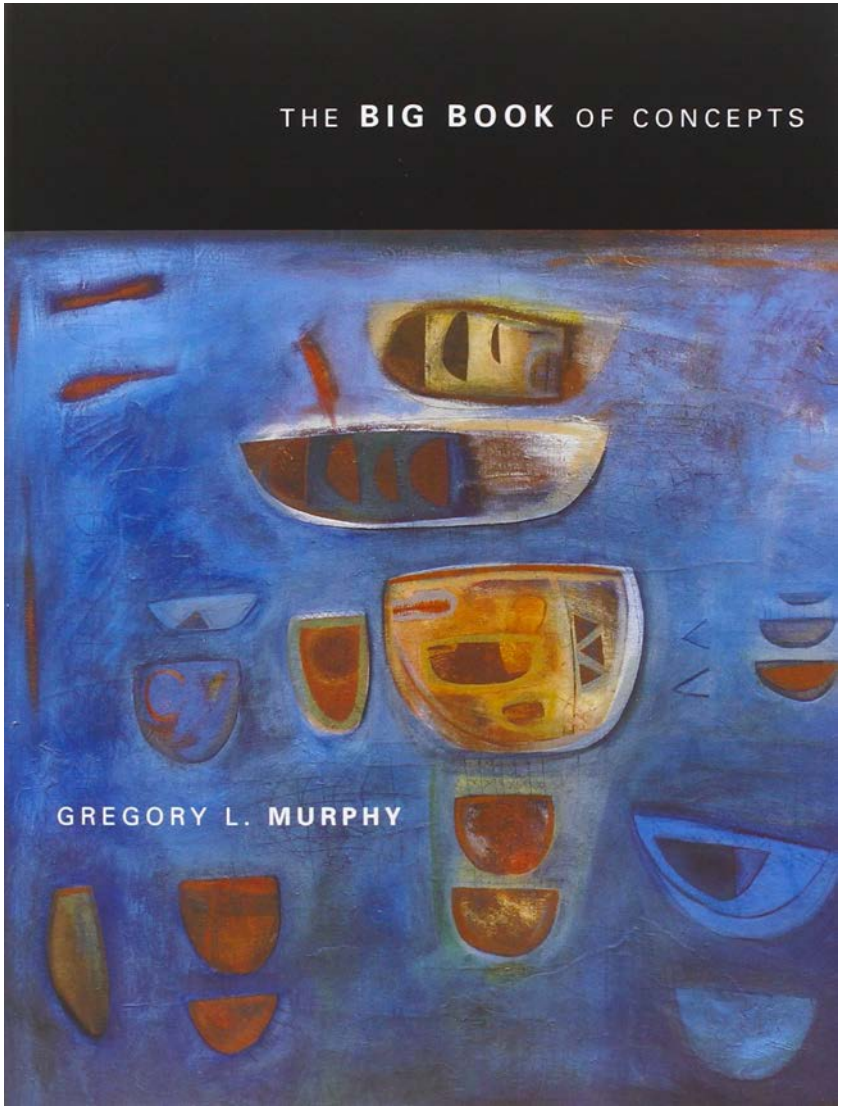
Alves et al. (in preparation), “Quality control and class noise reduction of satellite image time series”

# Using SOM to find out ecological differences




Alves et al. (in preparation), “Quality control and class noise reduction of satellite image time series”

# The limits of our concepts




"The gradation of properties in the world means that our smallish number of categories will never map perfectly onto all objects" (Murphy, *The Big Book of Concepts*, 2004)

image: Amazon, referral: O Ahlqvist

images: Alamy

# The ambiguity of “forest”




# A scientific question linked to public policy

When is a forest not a forest?

images: Alamy

images: Alamy, INPE

# Deforestation as an event




# Event composition

me

Event 1

Event 2

Event 3

Event 4

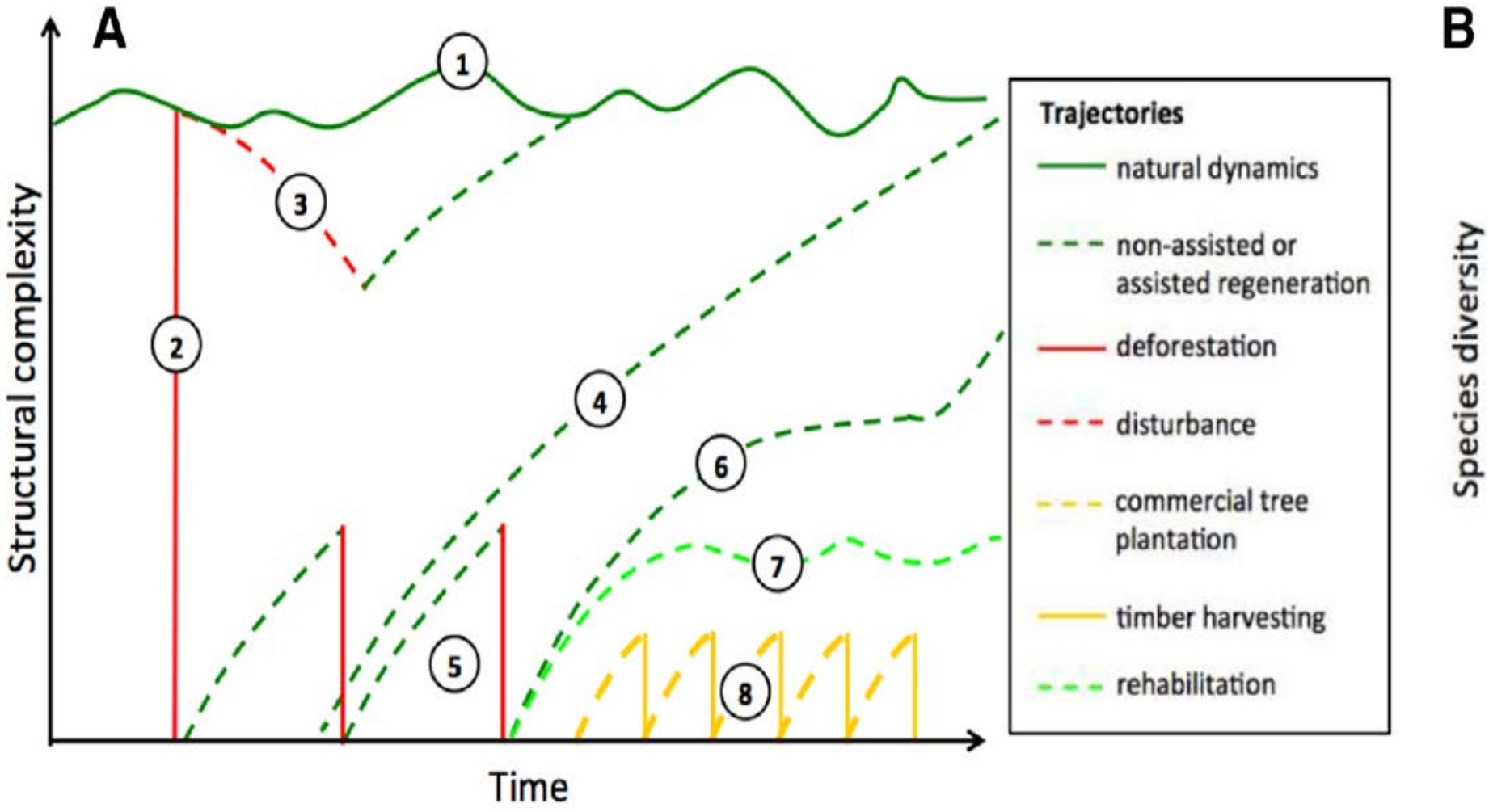
Forest loss > 20%

Loss > 50%

Loss > 90%

Clear cut

# Distinguishing forests by temporal evolution



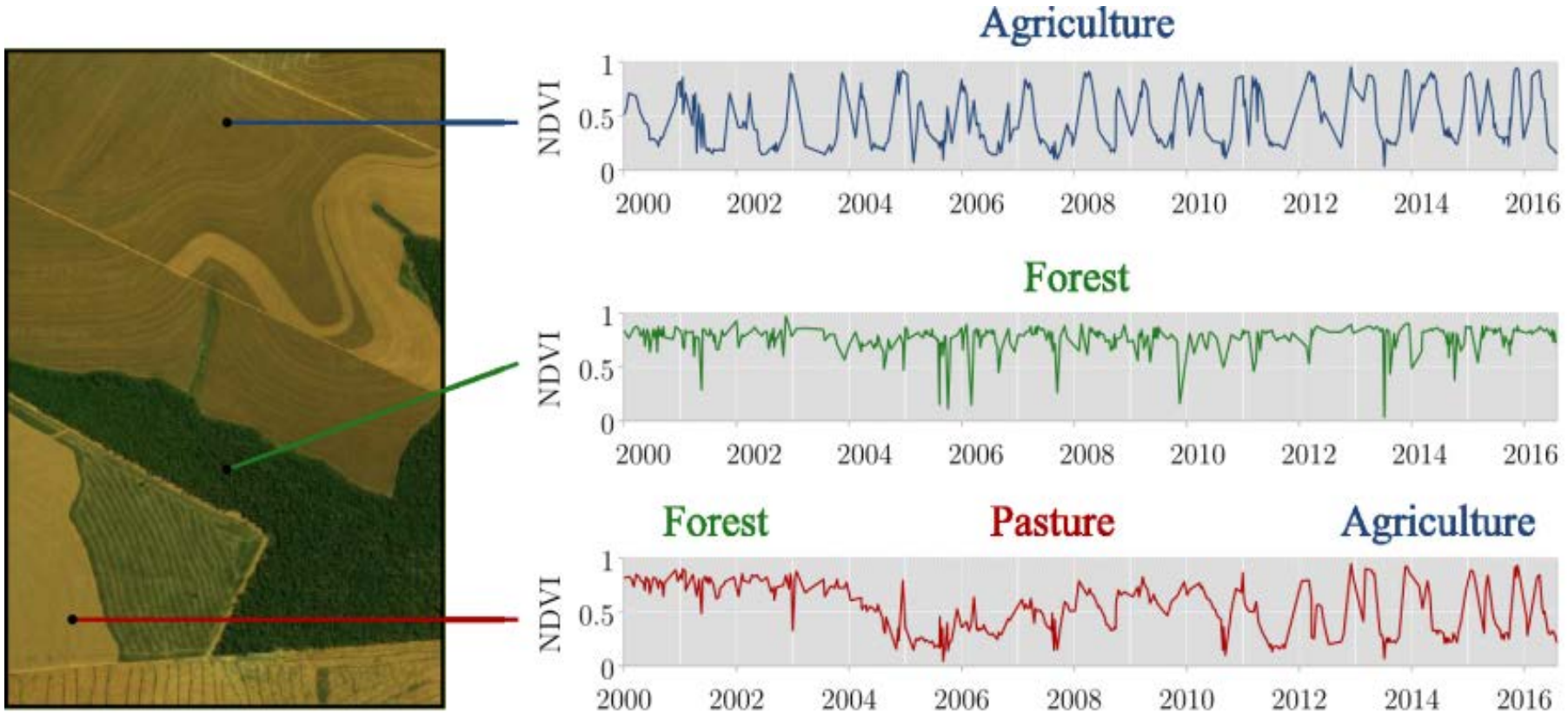

source: Chazdon et al (2016)

# Objects and events in land change

| Space | Time |
| --- | --- |
| Continuants | Occurrents |
| Indigenous area | Deforestation |

images: ISA, Greenpeace

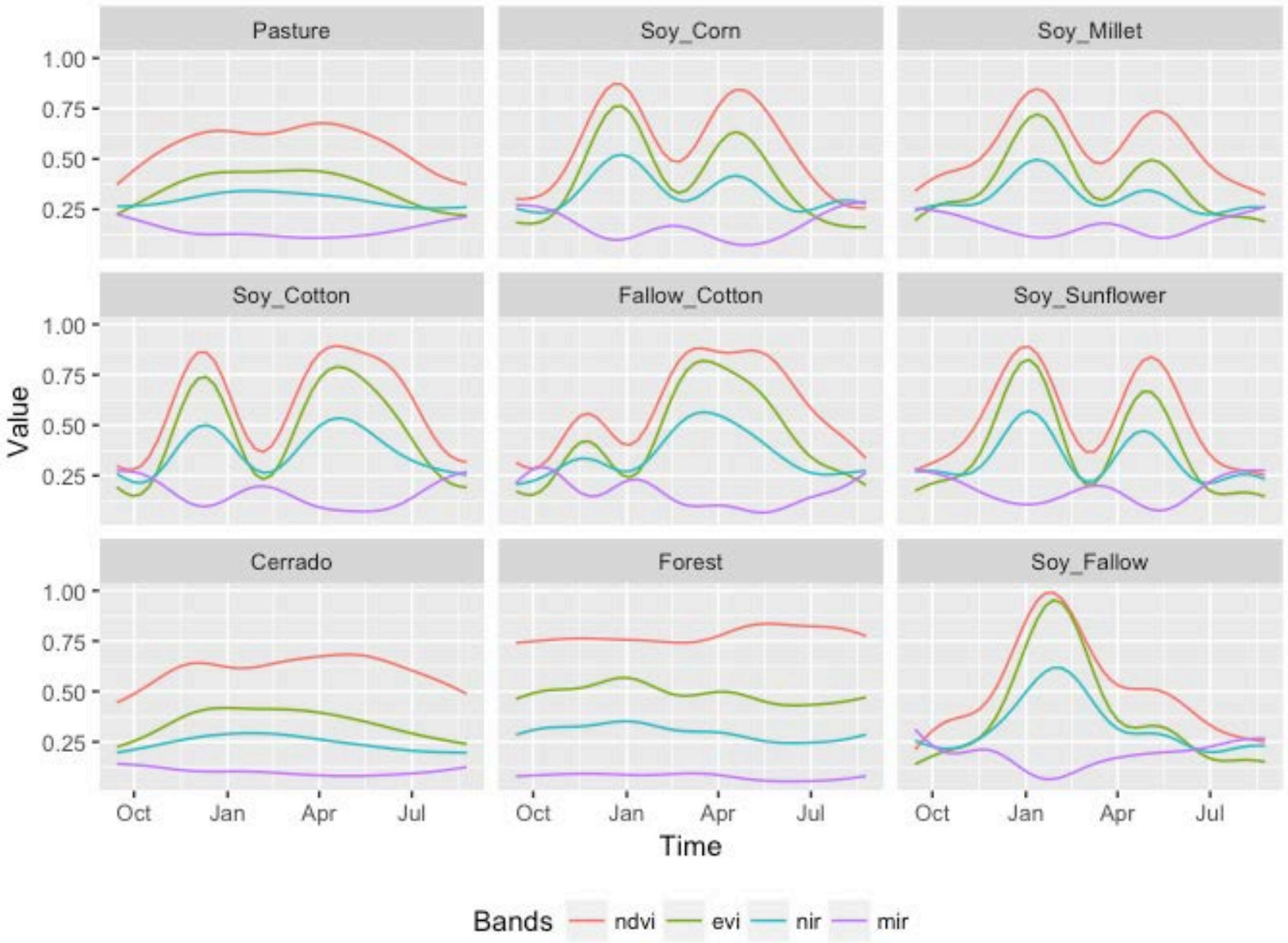
# Land-use change trajectories are measurements of events




How to extract events from time series?

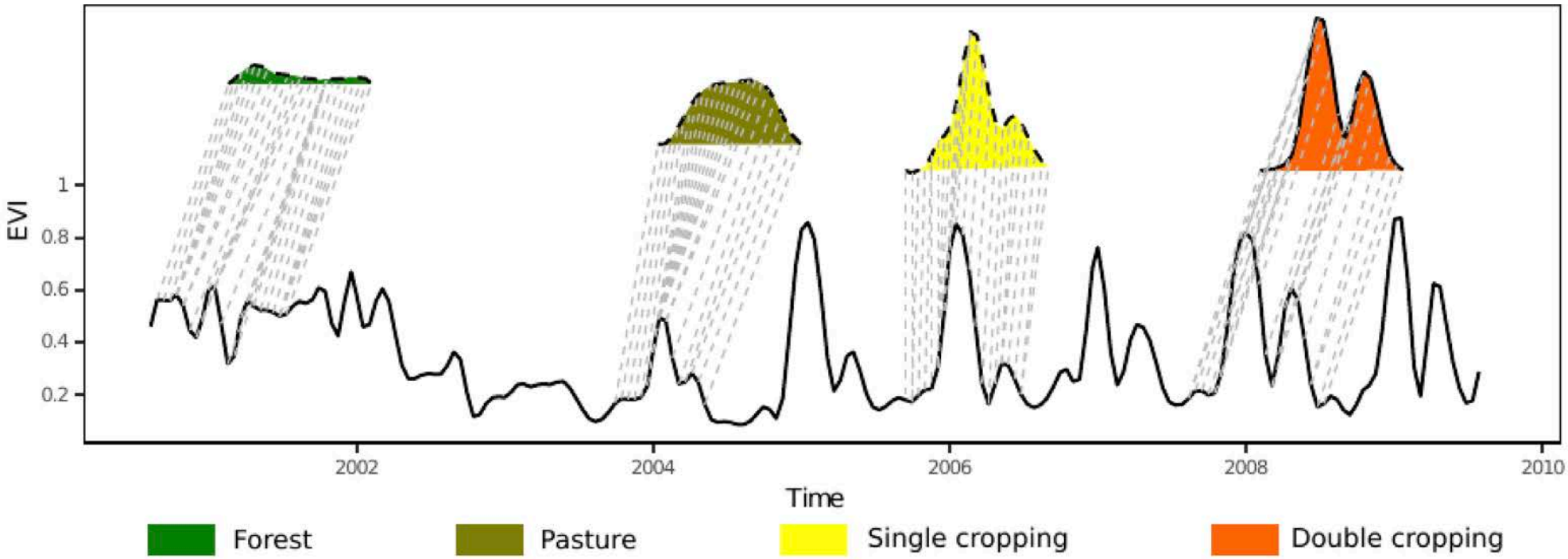
graphics: Rolf Simoes (INPE)

# Temporal patterns of different classes




graphics: author

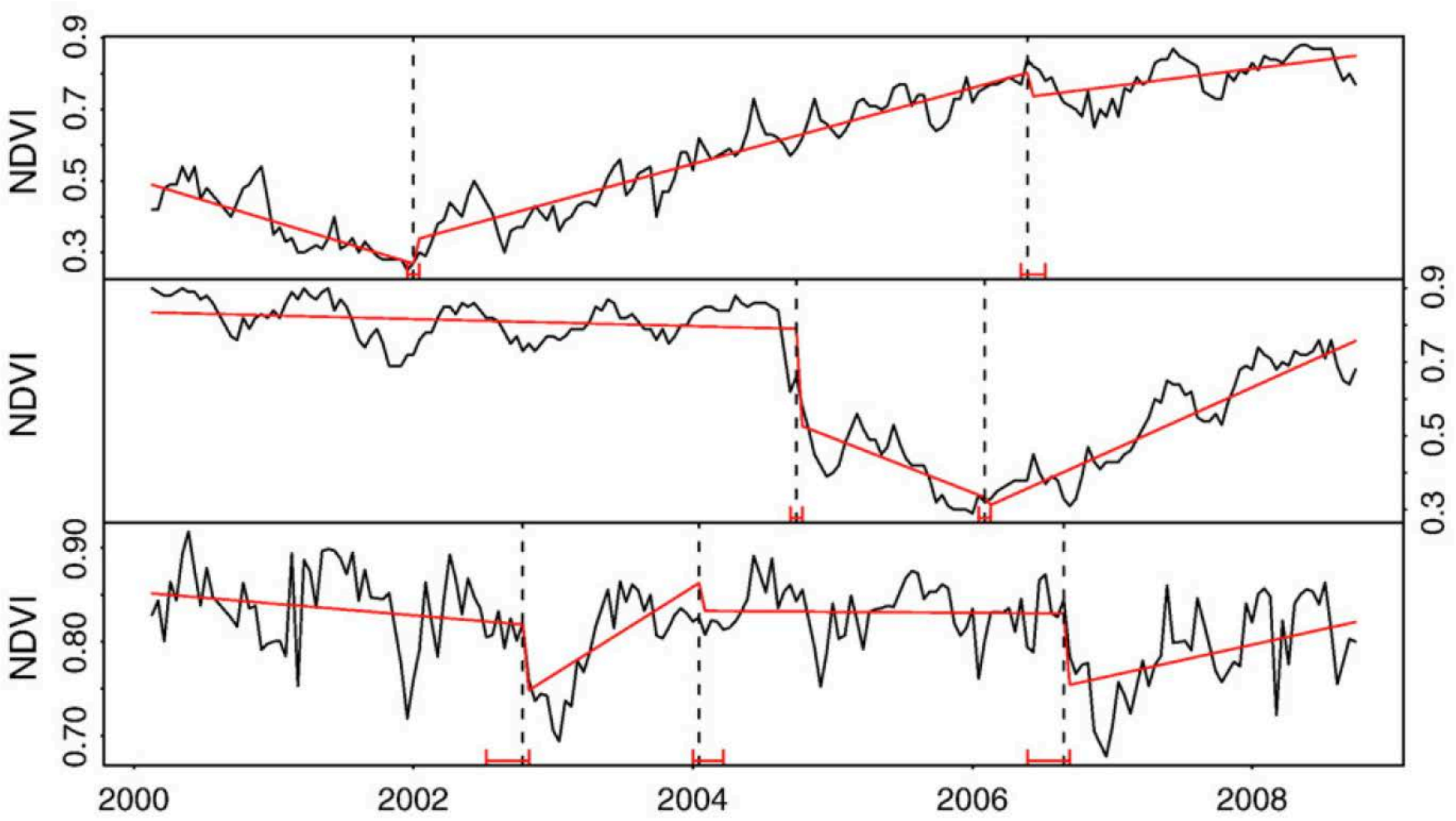
# Finding events in a remote sensing time series




"A Time-Weighted Dynamic Time Warping Method for Land-Use and Land-Cover Mapping" (Maus, Câmara et al, 2016)
"dtwSat R Package" (Maus et al., 2019)

graphics: V Maus (Univ of Vienna)

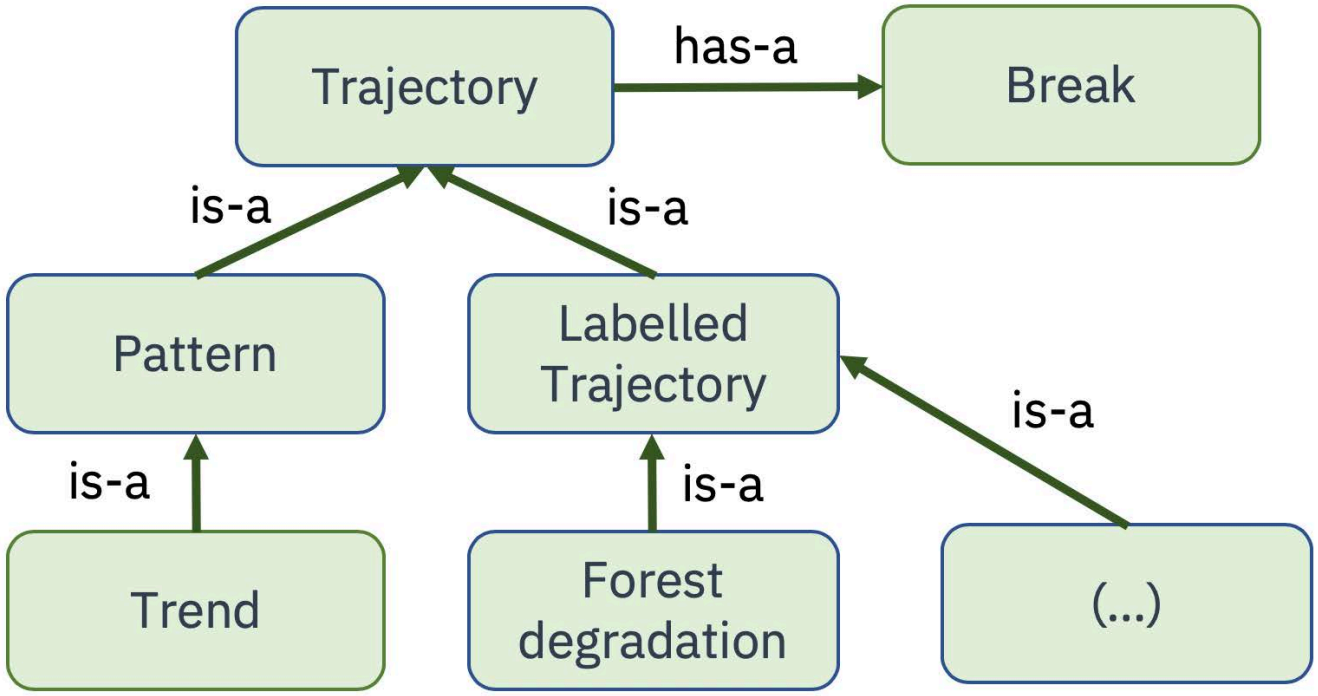
# Finding breaks in time series (events)




BFAST: Detected changes in an 16-day NDVI time series for pine plantations (Verbesselt et al., 2010)

graphics: J Verbesselt

# Proposed hierarchy of events linked to land use change




Caveat: names of high-level event types in the hierarchy are conflated with their measures

source: Camara, JOSIS 2020 (submitted)

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2018.1520235

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

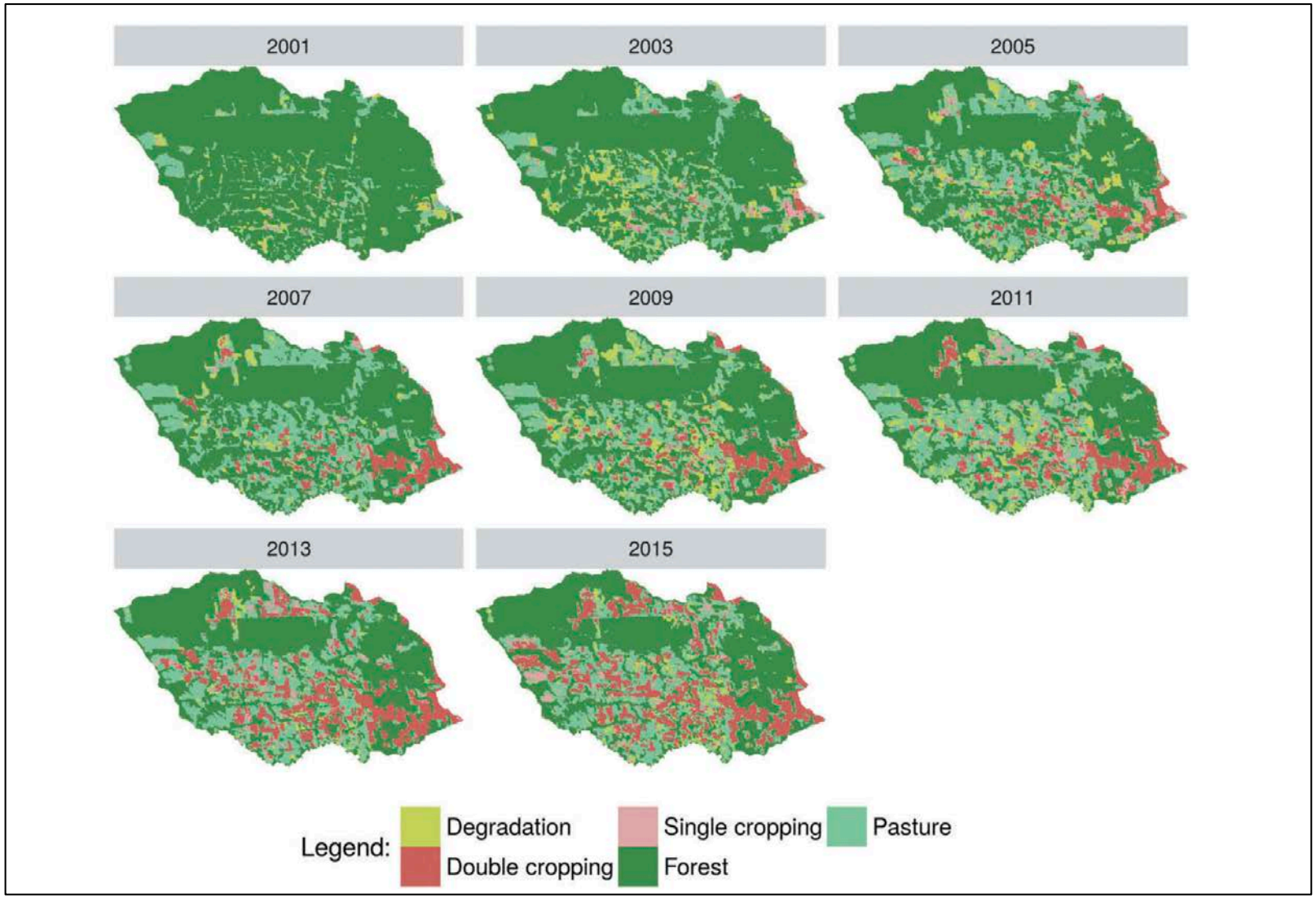
# A spatiotemporal calculus for reasoning about land-use trajectories

**Table 4.** Examples of LUC Calculus with multiple transitions.

| |
|---|
| *Search for all 'forest" locations where forest regrowth has occurred* |
| $\forall\, l_1 \in L, \forall\, t_i, t_j \in T,\ \text{RECUR}(l_1, Forest, t_i, t_j)$ |
| *Search for all 'forest" locations that have been converted into 'pasture"* |
| $\forall\, l_2 \in L, \forall\, t_i, t_j \in T,\ \text{CONVERT}(l_2, Forest, t_i, Pasture, t_j)$ |
| *Search for all 'forest" locations that have evolved into 'double cropping"* |
| $\forall\, l_2 \in L, \forall\, t_i, t_j \in T,\ \text{EVOLVE}(l_2, Forest, t_i, Double\_Cropping, t_j)$ |

source: Maciel et al, IJGIS (2019)

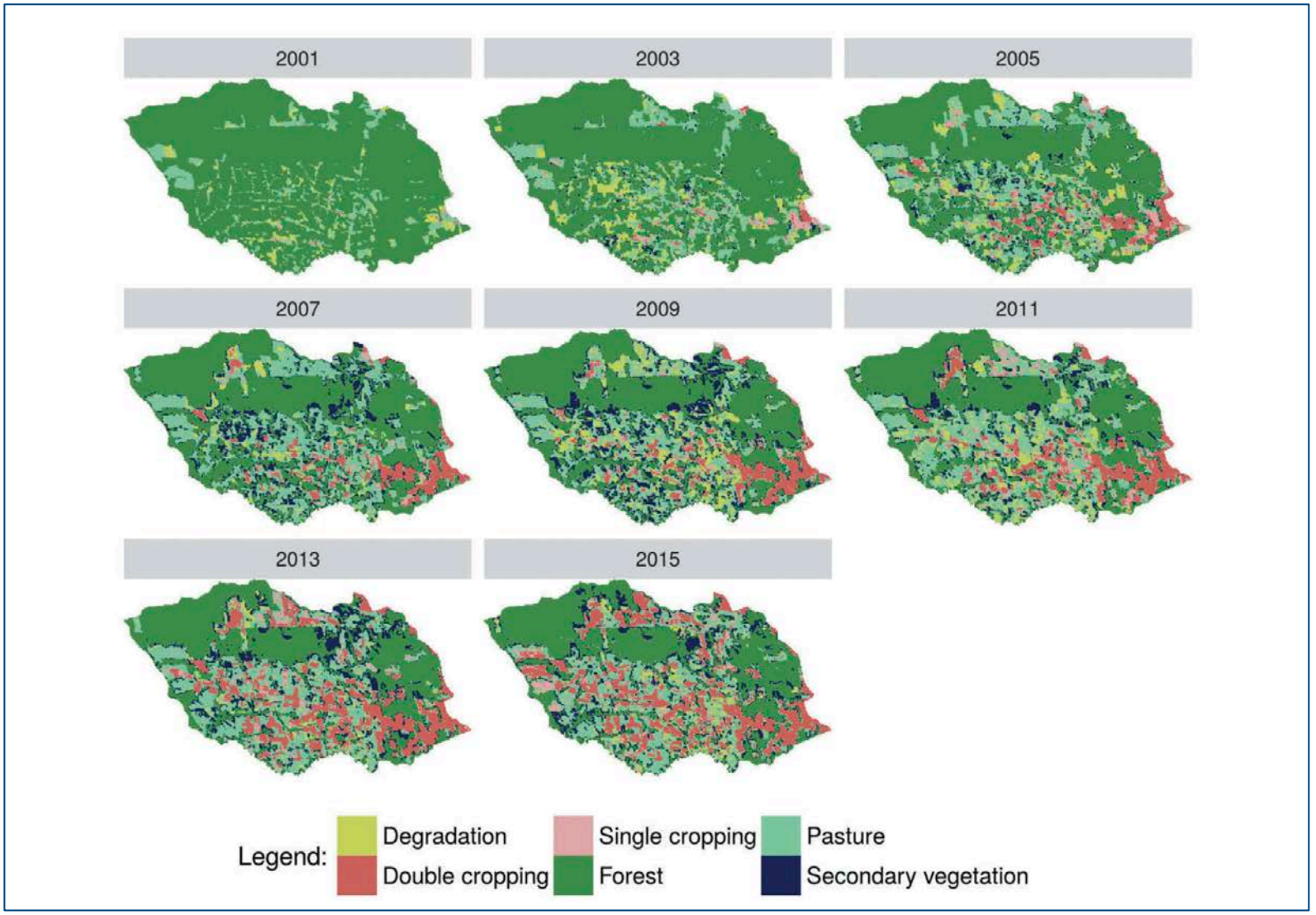
# Year-by-year land classification




Classification does not consider recurring events

source: Maciel et al, IJGIS (2019)

# Year-by-year land classification

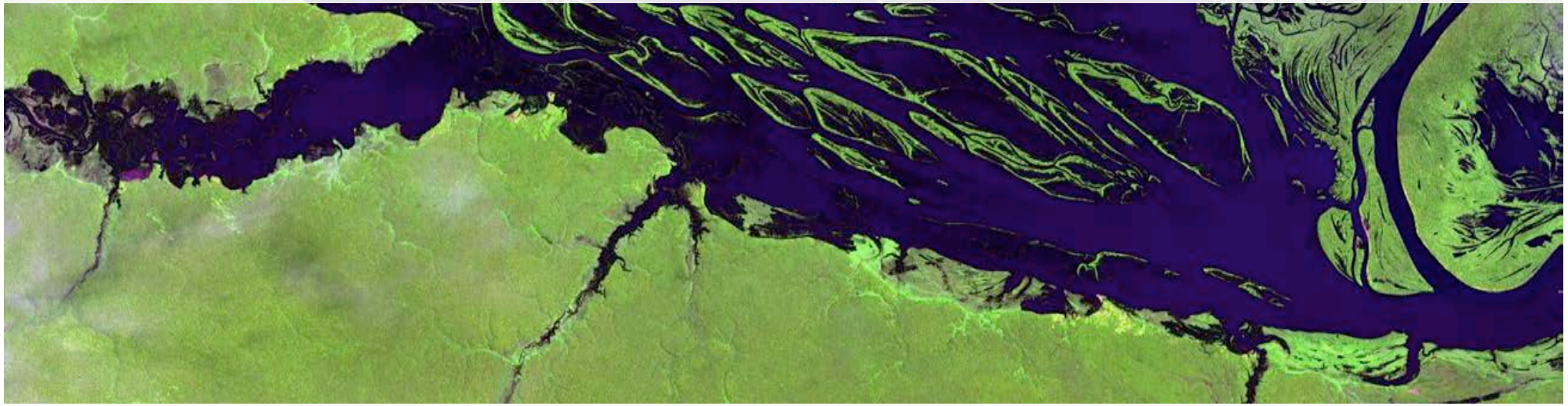



After spatiotemporal calculus reasoning

source: Maciel et al, IJGIS (2019)

# Findings

1. Nature's diversity limits machine learning and big data apps
2. Understanding human-enviromental interactions is essential
3. Big data enables continuous monitoring (trajectories)
4. Trajectories are measurements of events
5. Event recognition replaces object identification in big data



JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 20 (2020), pp. 21–34

doi:10.5311/JOSIS.2020.20.645

INVITED ARTICLE

# On the semantics of big Earth observation data for land classification

Gilberto Camara

National Institute for Space Research, Brazil

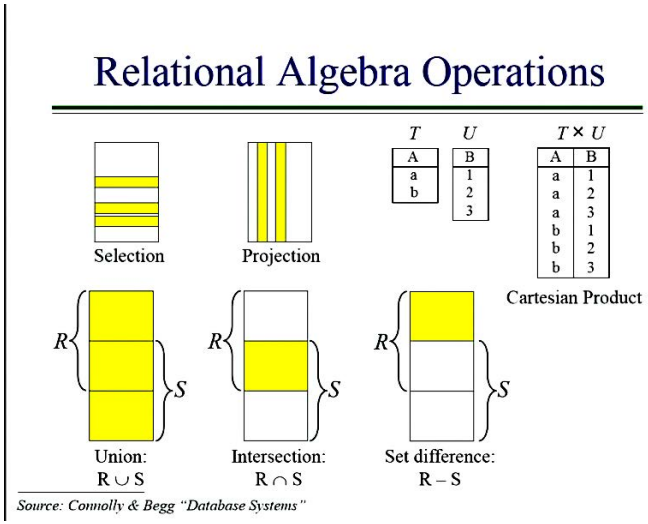
*Received: May 11, 2020; accepted: June 13, 2020*

# Research agenda

1. How to identify ground samples that account for natural and social variation?
2. How to capture local knowledge in big EO data?
3. How to recognise events in big EO data?
4. How best to reason with events?
5. How to use conceptual models to propose sound APIs to deal with big EO data?
6. How to develop and share robust open source tools to work with machine learning and big EO data?
7. How to make these technologies useful to support sound environmental policies?

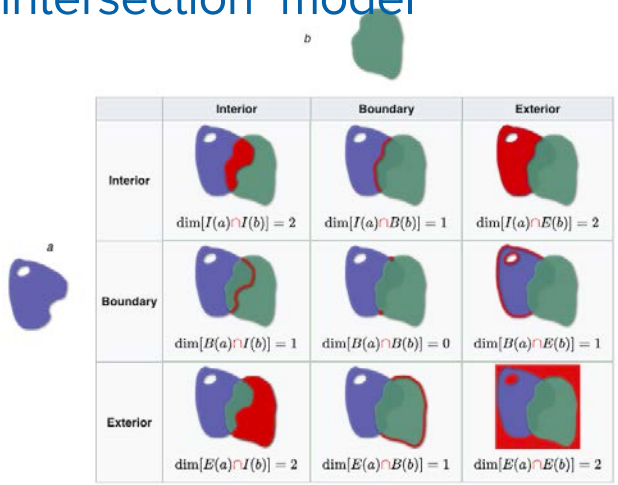
# Backup slides

# Standards need theory
# (and theory can come from good practices)




SQL

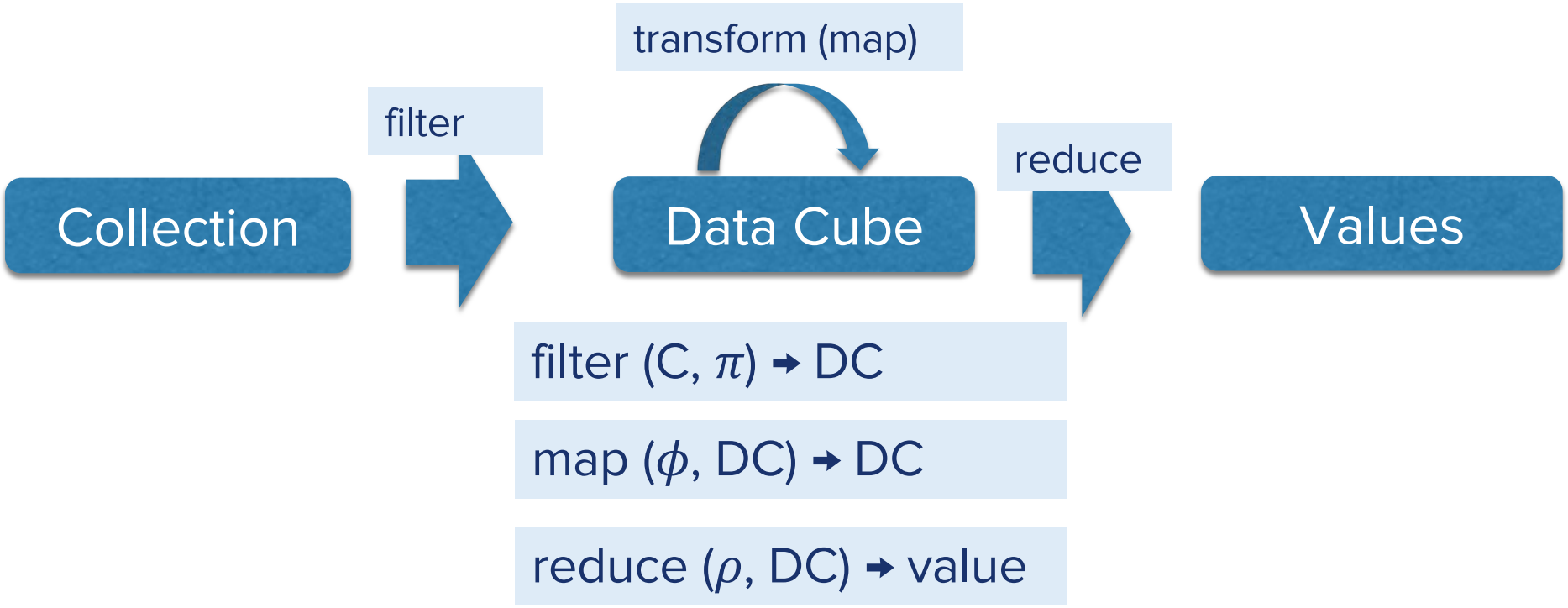
Dimensionally extended
9-intersection model



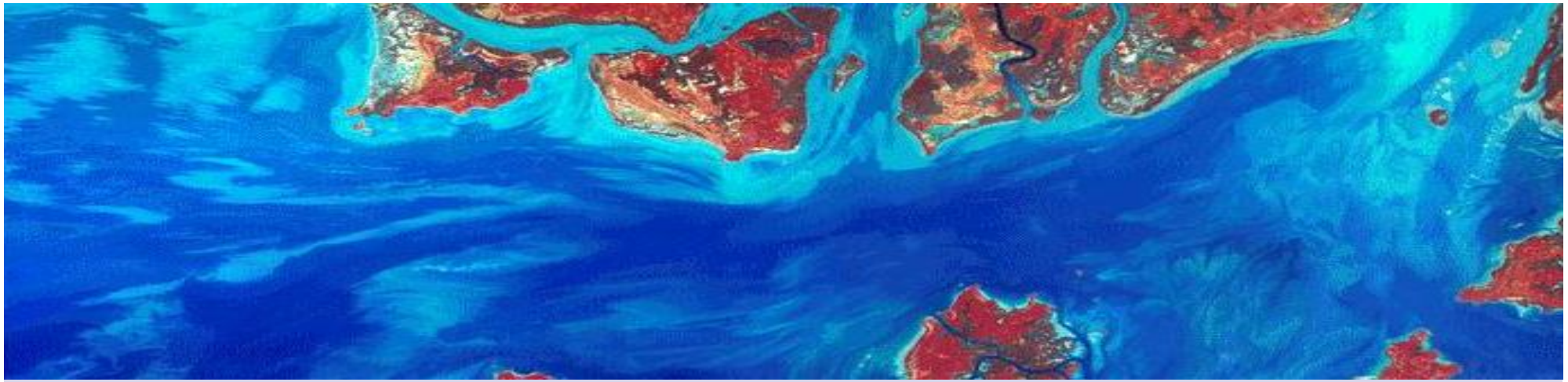

SF-SQL, GeoSPARQL, WFS, ...

OGC®
Making location count.

# Functional programming in big EO data: foundation for a sound API




Is big EO data analysis an extended case of map-reduce?

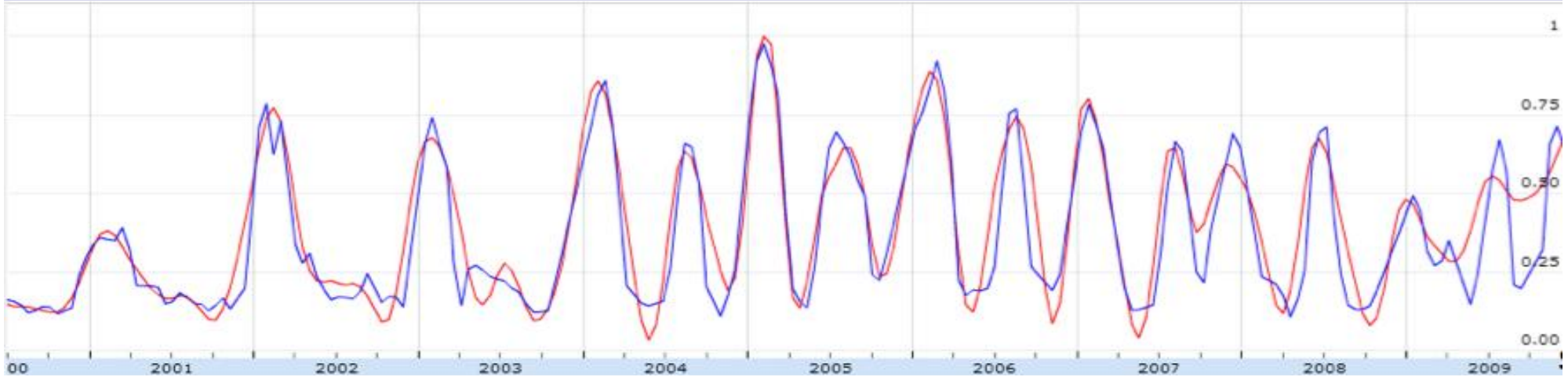


# Fields as a Generic Data Type

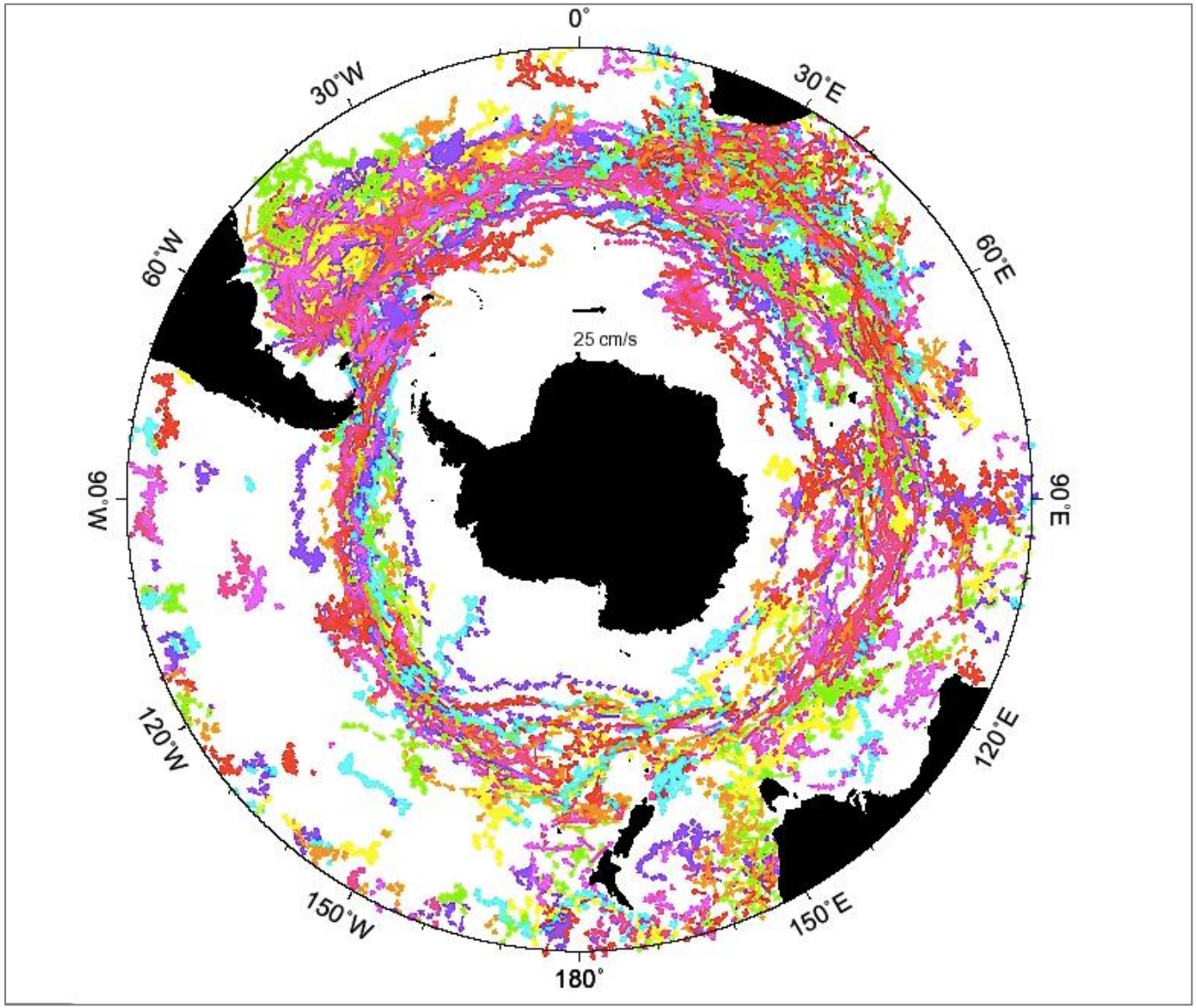
$$\texttt{estimate: Position} \rightarrow \texttt{Value}$$

Instances of Position: space, time, and space-time

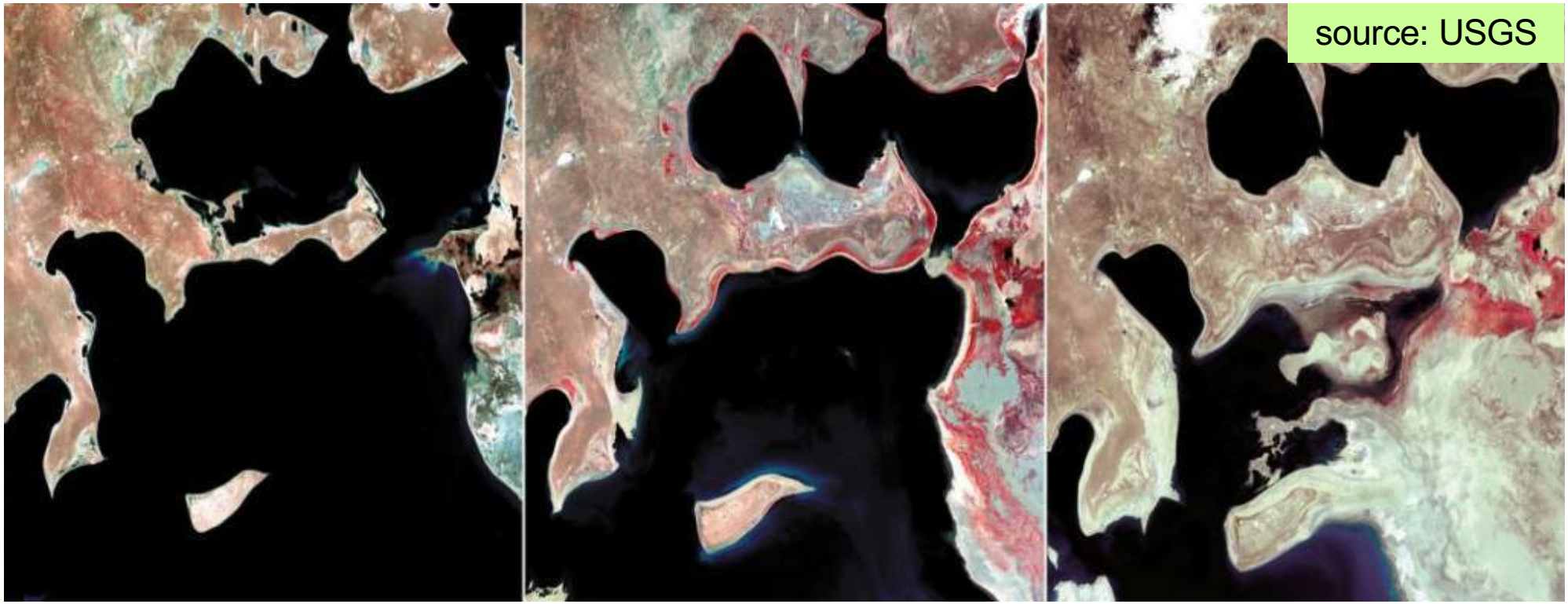
Instances of Value: numbers, strings, space-time



# A field of fields (Argo floats in Southern Ocean)




Positions: space  Values: trajectories (time ⍰ space)

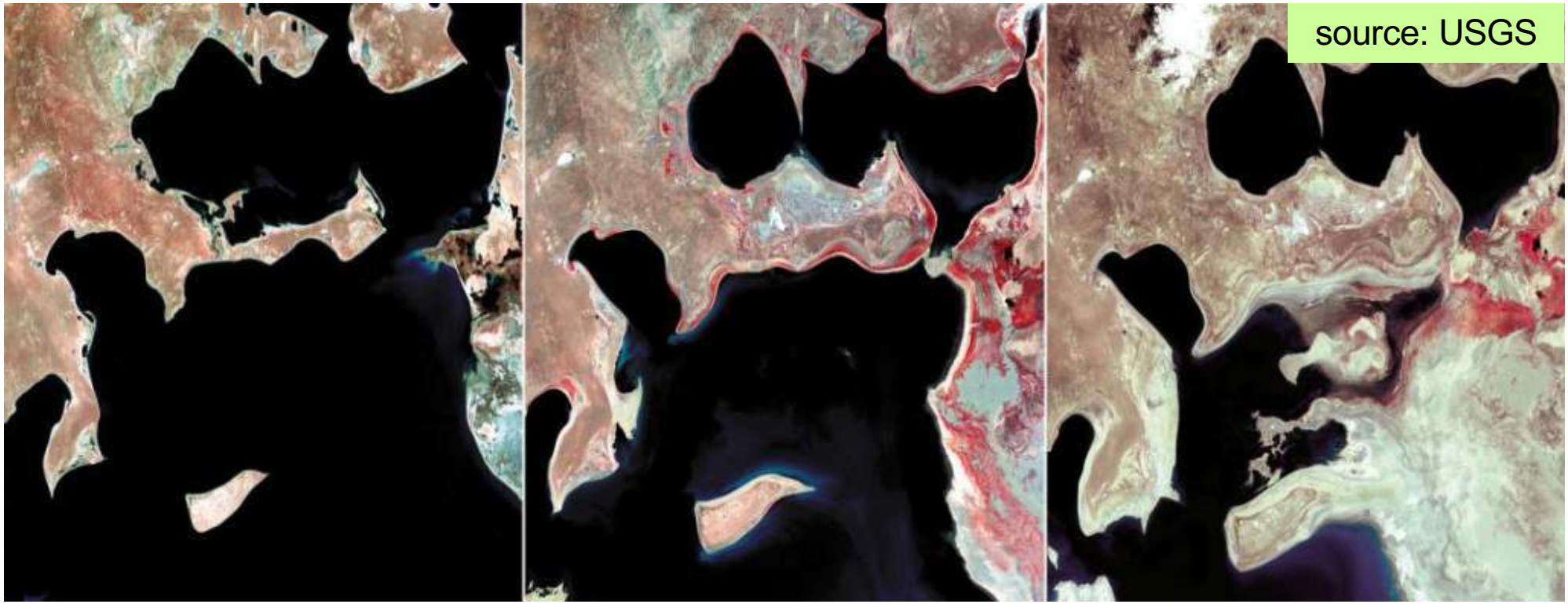



## Events are categories (Frank, Galton)

identity : $id \cdot a = a$

composition : $\forall a, \forall b, \exists c, c = a.b$

associativity : $a \cdot (b \cdot c) = (a \cdot b) \cdot c$




## Events are categories (Frank, Galton)

cause (a,b) = effect (b,a)

before (a,b) = after (b,a)

on-commutativity : a · (b · c) /= (b.a) · c

# Event composition

me

Event 1

Event 2

Event 3

Event 4

Forest loss > 20%

Loss > 50%

Loss > 90%

Clear cut


nature
Vol 452 | Issue no. 7184 | 13 March 2008
“A few satellites can cover the entire globe, but there needs to be a system in place to ensure their images are readily available to everyone who needs them. Brazil has set an important precedent by making its Earth-observation data available, and the rest of the world should follow suit.”

# Humans and machines on remote sensing

| | | Humans | |
|---|---|---|---|
| | | Easy | Hard |
| Machines | Easy | Water mask, fires | Time series |
| | Hard | Space-time objects | Cause-effect |

# Forest (FAO definition)

Land spanning more than 0.5 hectares with trees higher than 5 meters and a canopy cover of more than 10 percent, or trees able to reach these thresholds in situ. It does not include land that is predominantly under agricultural or urban land use.

# A working definition of big data

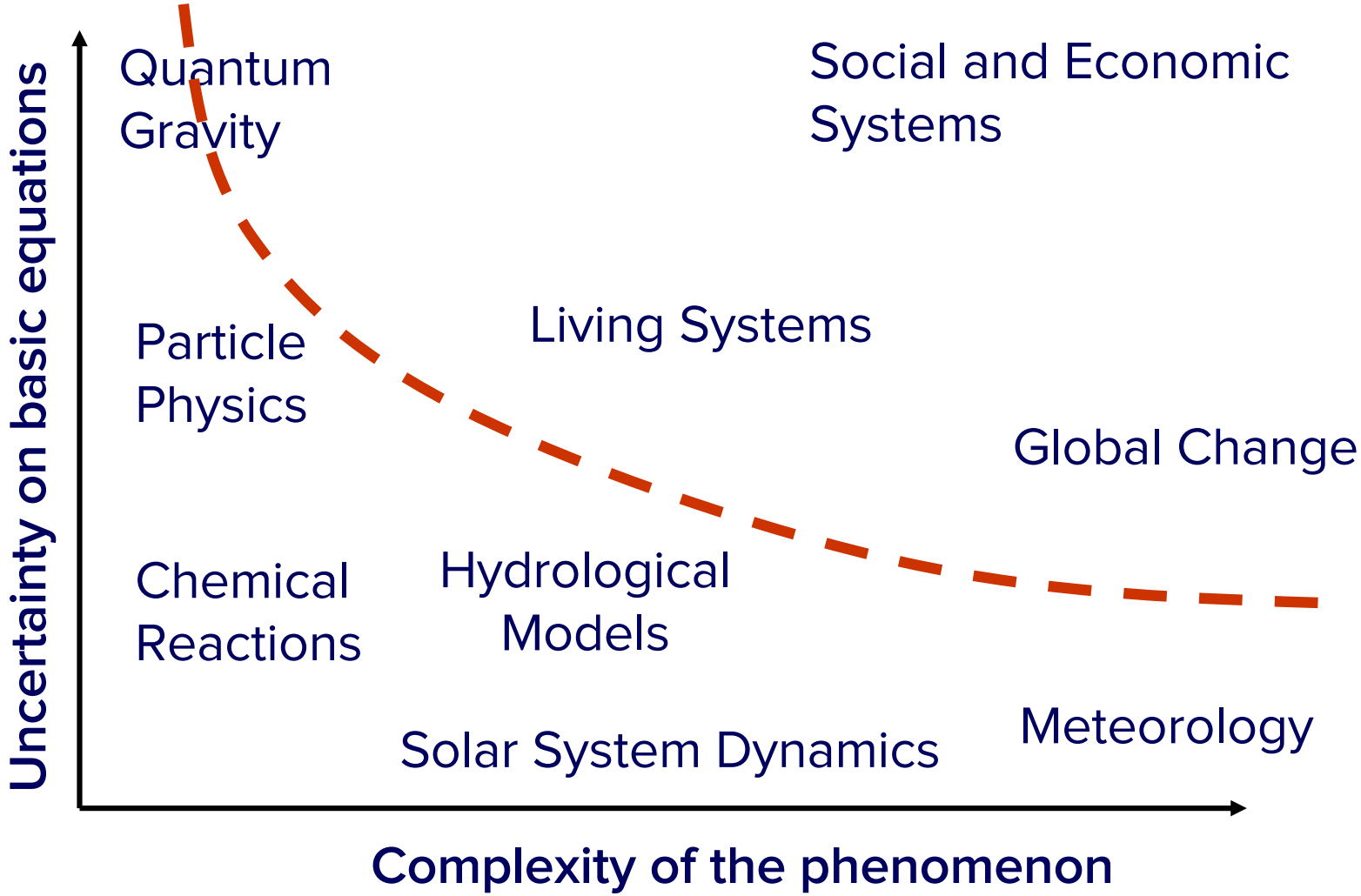
≃

n << all  Statistics: we have a small part of the data

n == all  Big brother: we have all the data (do we?)

n ≃≃ all  Big data: we have data close to problem size

# Limits of our models




source: John Barrow
(after David Ruelle)

# From research to decision-making

## Research

problem-based

innovative

objective

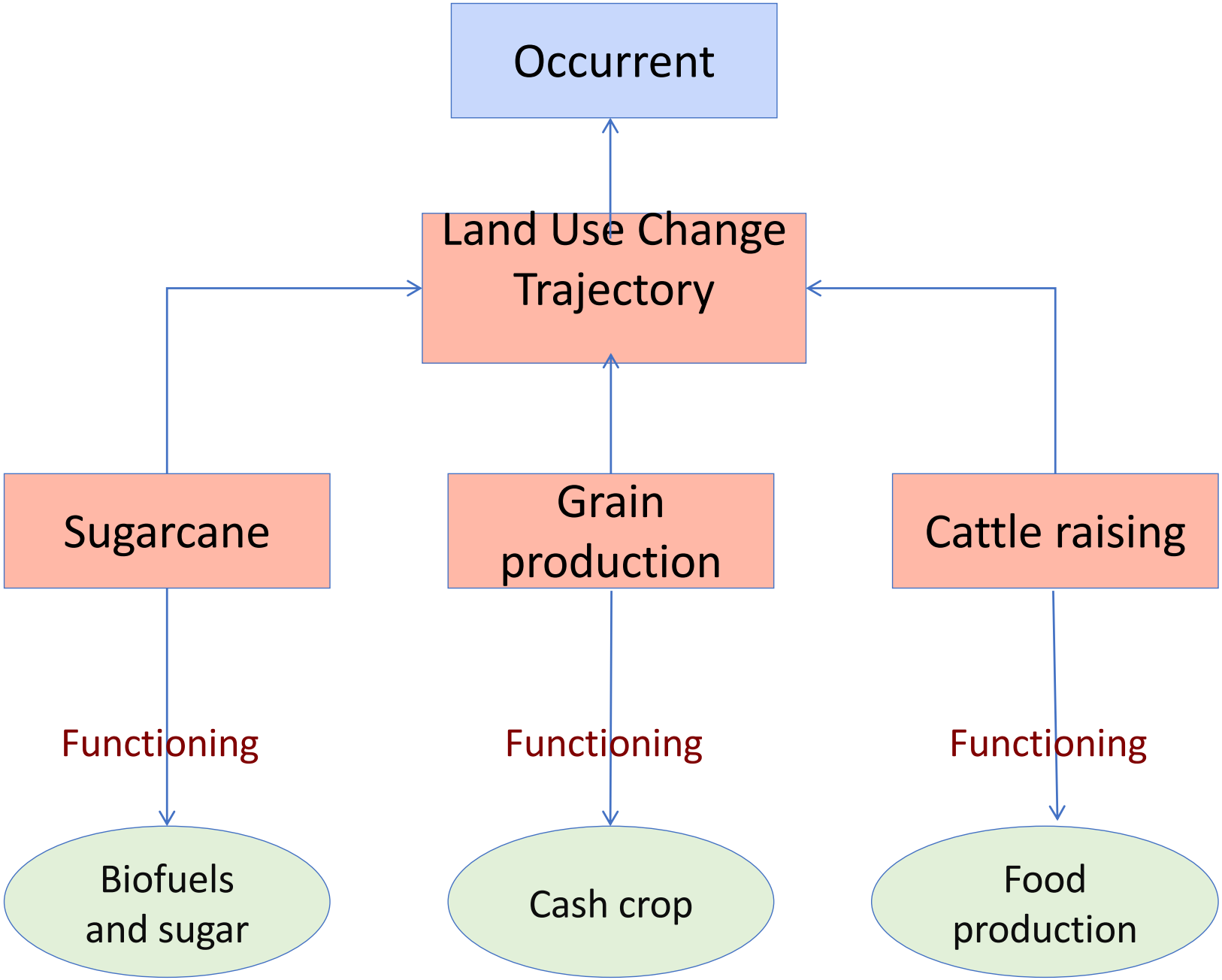
## Valley of Death

## Decision-making

outcome-based

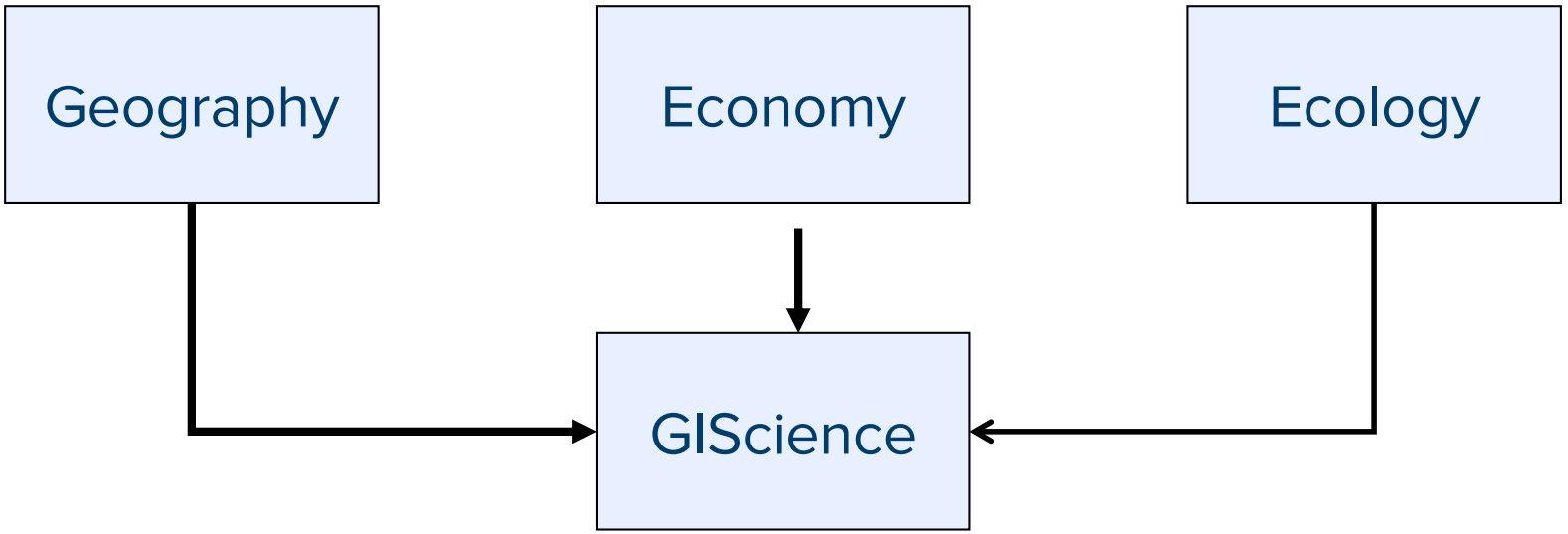
compromise

best guess

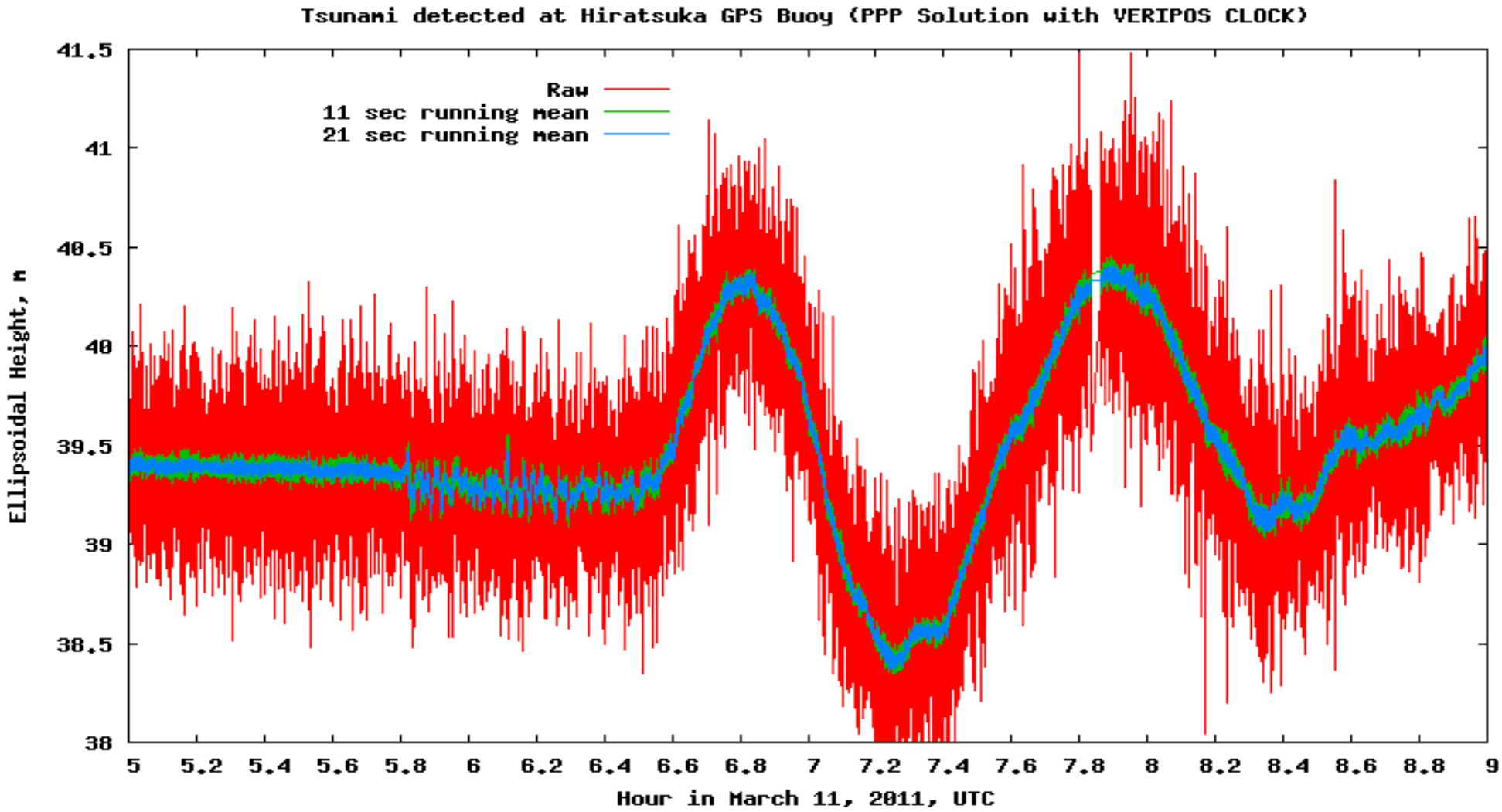

Occurrent
Land Use Change Trajectory
Sugarcane
Grain production
Cattle raising
Functioning
Functioning
Functioning
Biofuels and sugar
Cash crop
Food production

# Collaborative e-science

Connect expertise from different fields
Make the different conceptions explicit in space and time



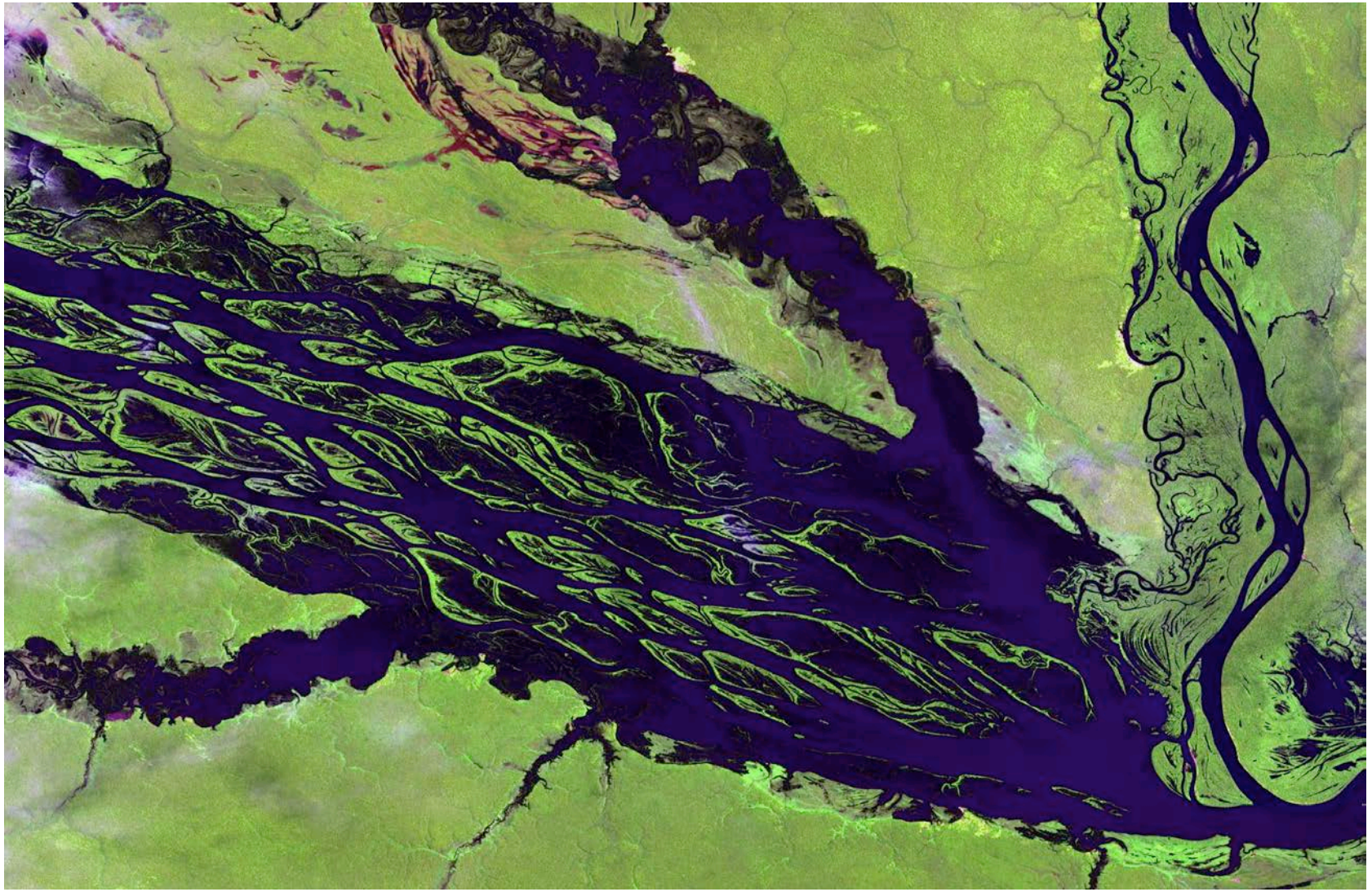
# A time series field (tsunami buoy)




positions: time  values: R

(Câmara et al., 2014)

# A remote sensing image



positions: 2D  values: $R^n$

(Câmara et al., 2014)

image: USGS

# Mato Grosso, Brasil,  May 8 – Jun 9, 1984

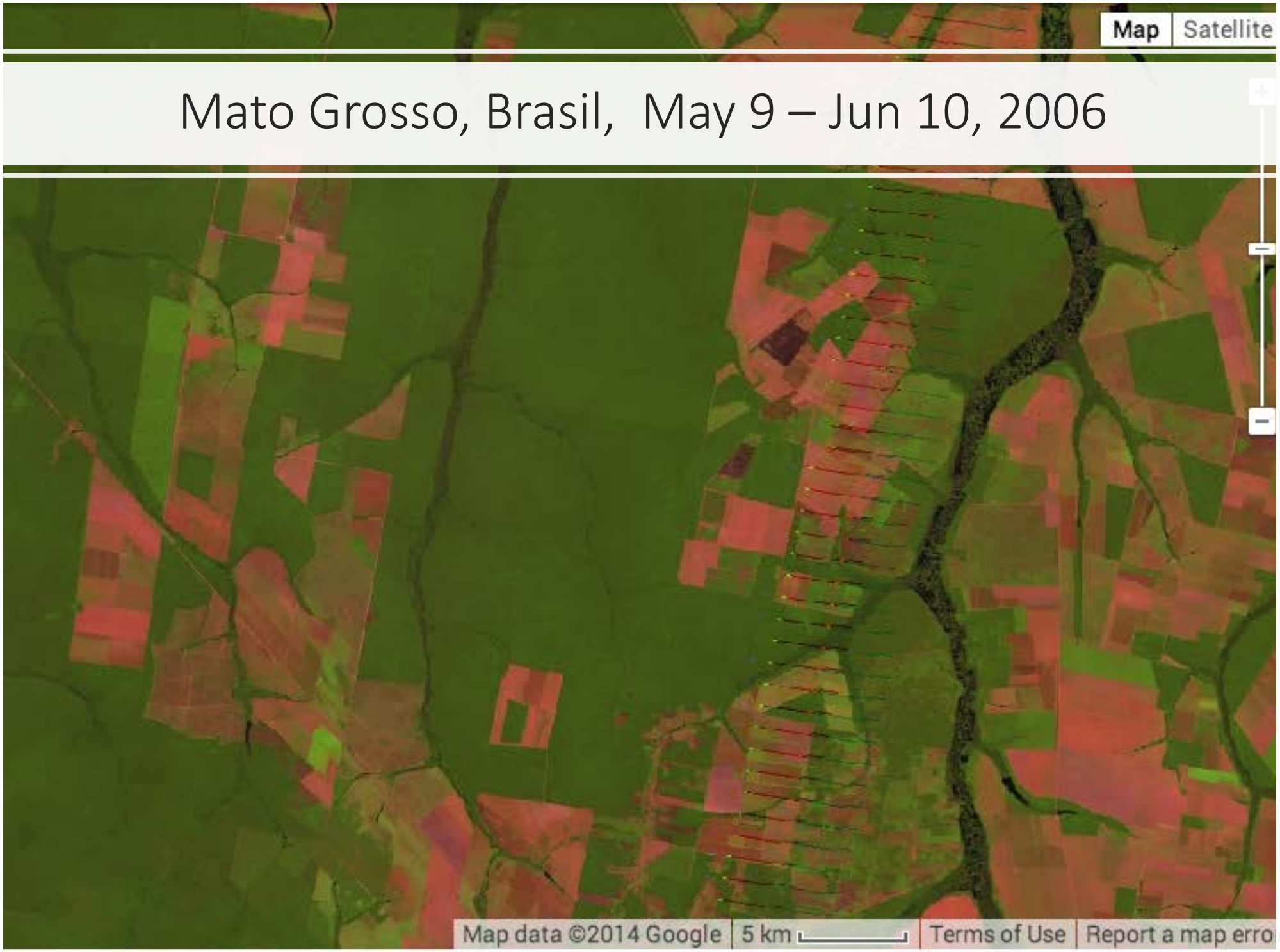
# Mato Grosso, Brasil,  Aug 13 – Sep 14, 1989

# Mato Grosso, Brasil,  Jul 12 – Aug 13, 1994

# Mato Grosso, Brasil,  Jun 9 – Jul 11, 2000

# Mato Grosso, Brasil,  Jun 9 – Jul 11, 2004

# Mato Grosso, Brasil,  May 9 – Jun 10, 2006



# Mato Grosso, Brasil,  Jun 9 – Jul 11, 2008

# Mato Grosso, Brasil,  Jun 10 – Jul 12, 2010