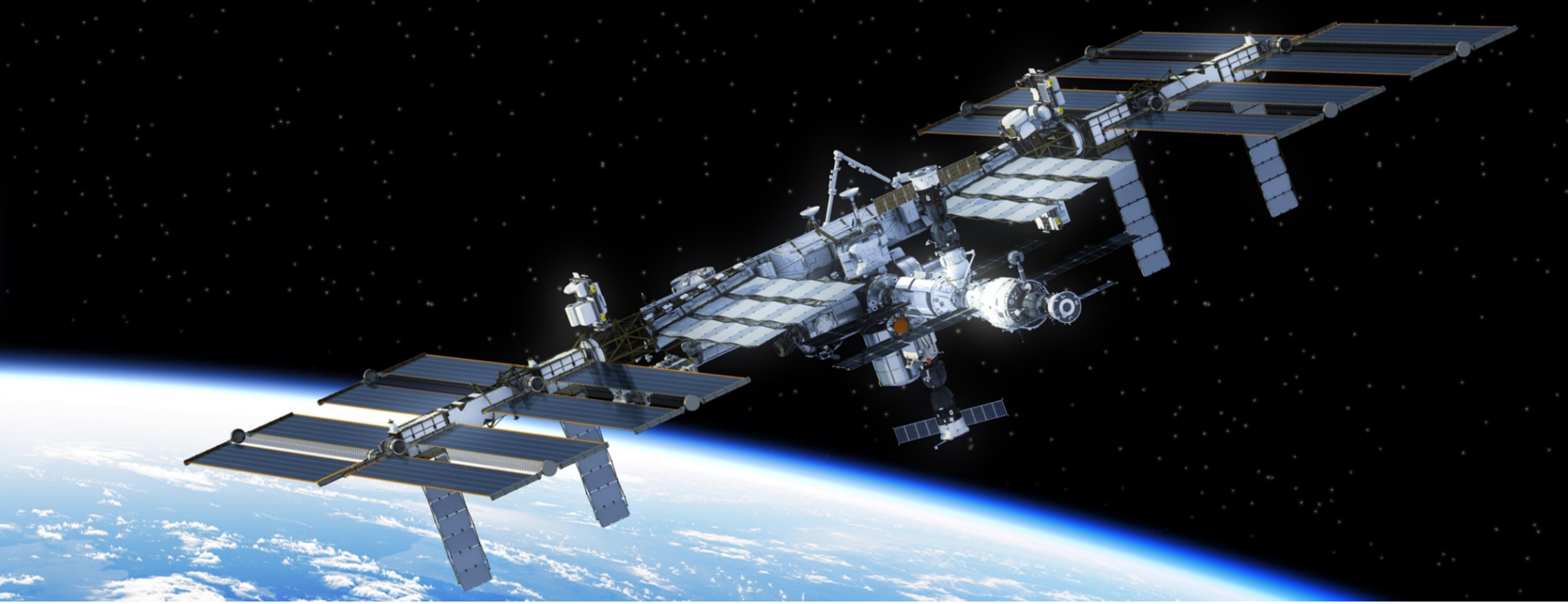


# Big data in remote sensing

Heather Reese
Department of Earth Sciences
University of Gothenburg

# Big data….

- …data sets whose size or type is beyond the ability of traditional systems to capture, manage, and process the data

- … has 1 or more of the 3 V's:
  - high volume,
  - high velocity,
  - high variety

Source: www.IBM.com; Ma et al., 2015

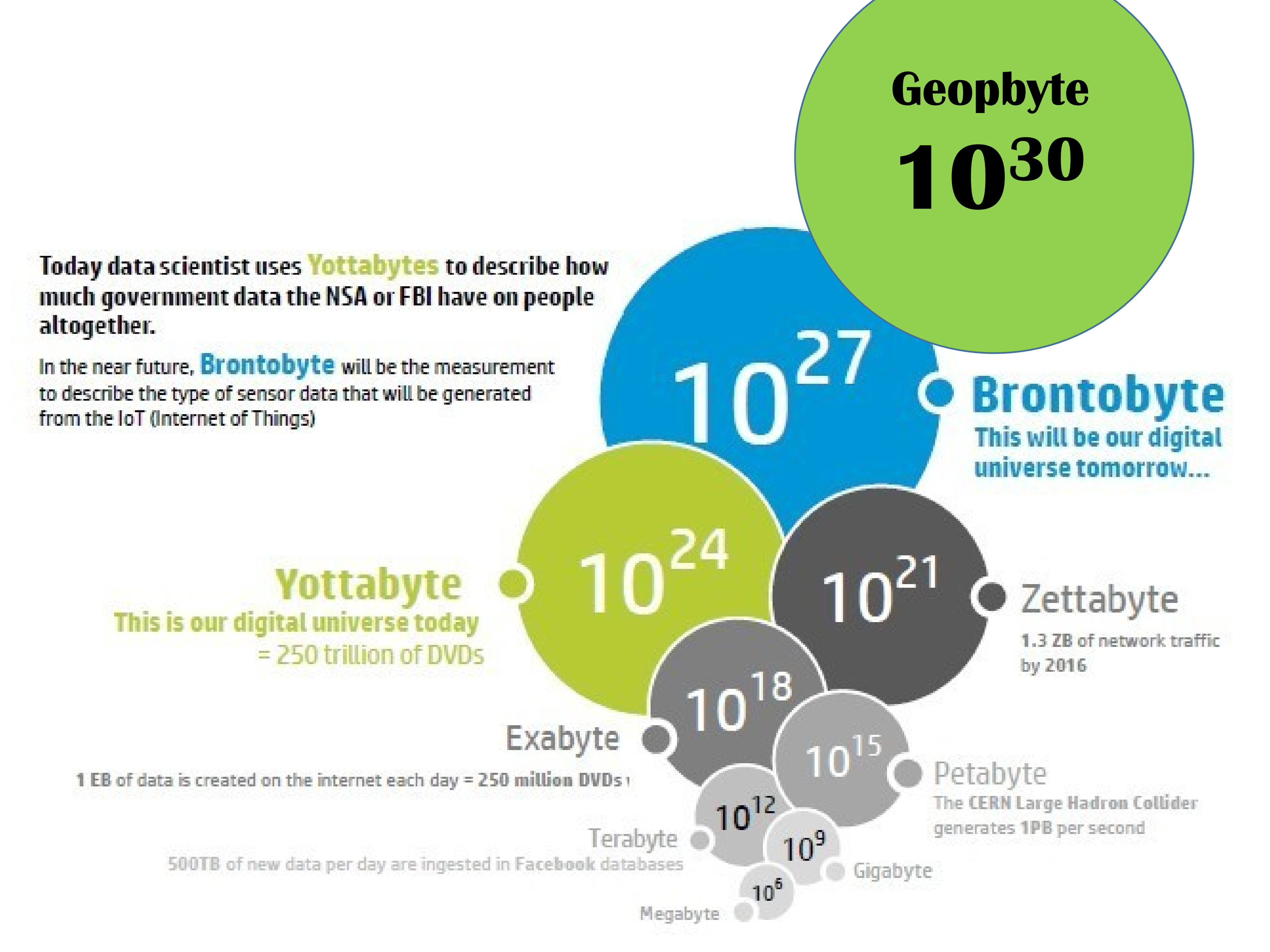
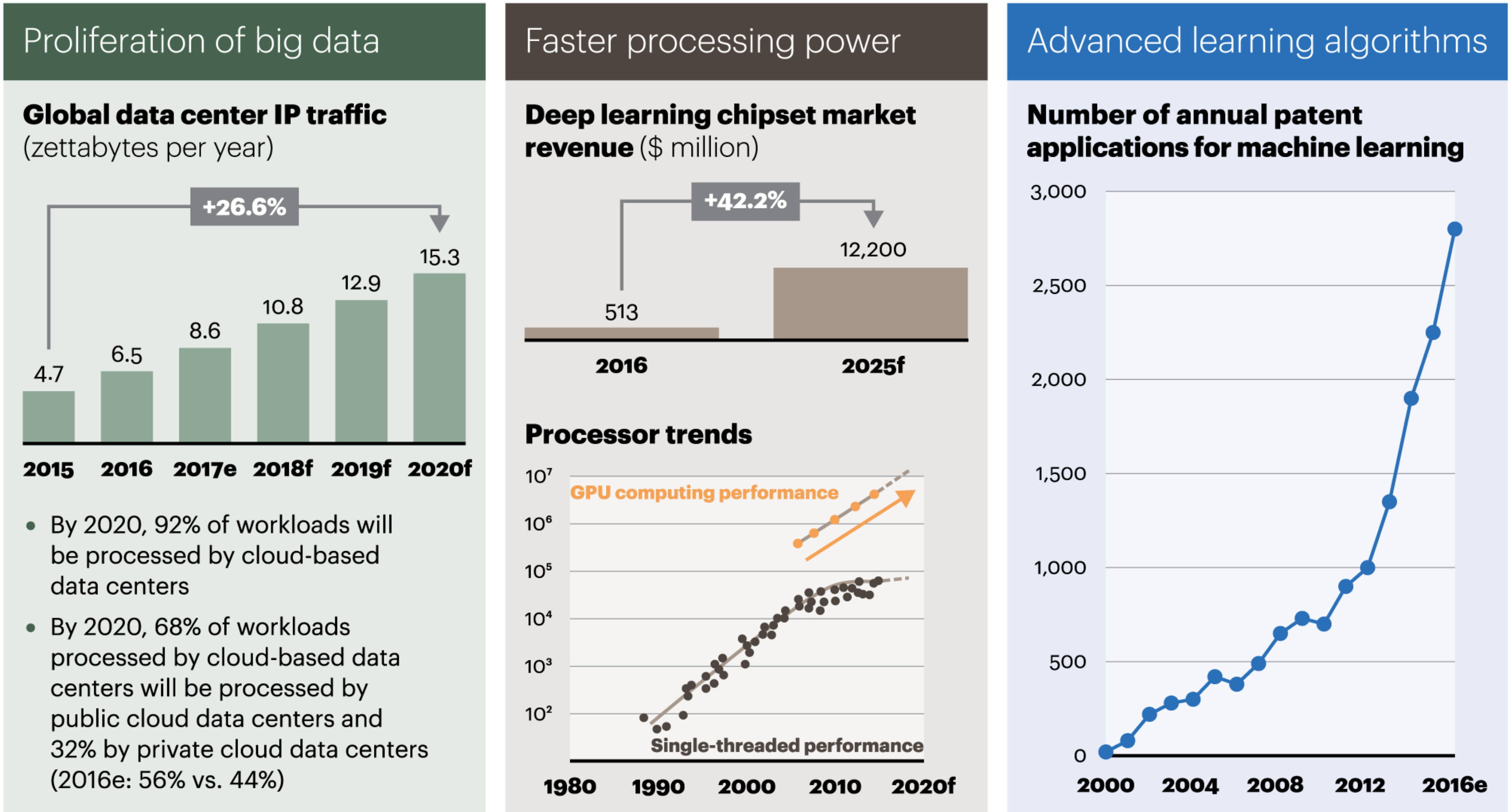

Geopbyte
10^30
Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.
In the near future, Brontobyte will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)
10^27
Brontobyte
This will be our digital universe tomorrow...
10^24
Yottabyte
This is our digital universe today
= 250 trillion of DVDs
10^21
Zettabyte
1.3 ZB of network traffic by 2016
10^18
Exabyte
1 EB of data is created on the internet each day = 250 million DVDs
10^15
Petabyte
The CERN Large Hadron Collider generates 1PB per second
10^12
Terabyte
500TB of new data per day are ingested in Facebook databases
10^9
Gigabyte
10^6
Megabyte

Figure 1

**Three trends are fueling the artificial intelligence boom**




Notes: GPU is graphics processing unit. Patents are those filed with the US Patent and Trademark Office.

Sources: Cisco, Tractica, Nextbigfuture, Teqmine Analytics; A.T. Kearney analysis

https://www.atkearney.com/operations-performance-transformation/artificial-intelligence/article/?/a/4-facts-every-business-leader-should-know-about-ai-and-enterprise-automation

# Some data sources we use

- Sentinel 1,2,3,4,5, Landsat, Planet, Hyperspectral, GEDI
- Airborne LiDAR (discrete return and full-waveform and multispectral and single photon)
- Aerial photos and point clouds
- Terrestrial Laser Scanning
- Harvester data
- Field-based reference data
- Crowd-sourced data
- Personally collected data from UAVs and other sensors

HIGH VARIETY !

# Some figures

- ESA collects 1.6 Tb/day from Sentinel-2A and 2B
- ESA publishes 10 Tb/day of Sentinel data
- Sentinel archive back to 2014 (S1) is now > 5 Petabytes

- USGS currently publishes 1.5 Tb/day of Landsat 7 & 8
- NASA's EOSDIS archives 4Tb/day

- Collective public archives of RS data are ca 1 Exabyte

*Figures from ESA, Menti et al (2018) and Ma et al. (2015)*

# Some figures

A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring (Li and Roy, 2017)

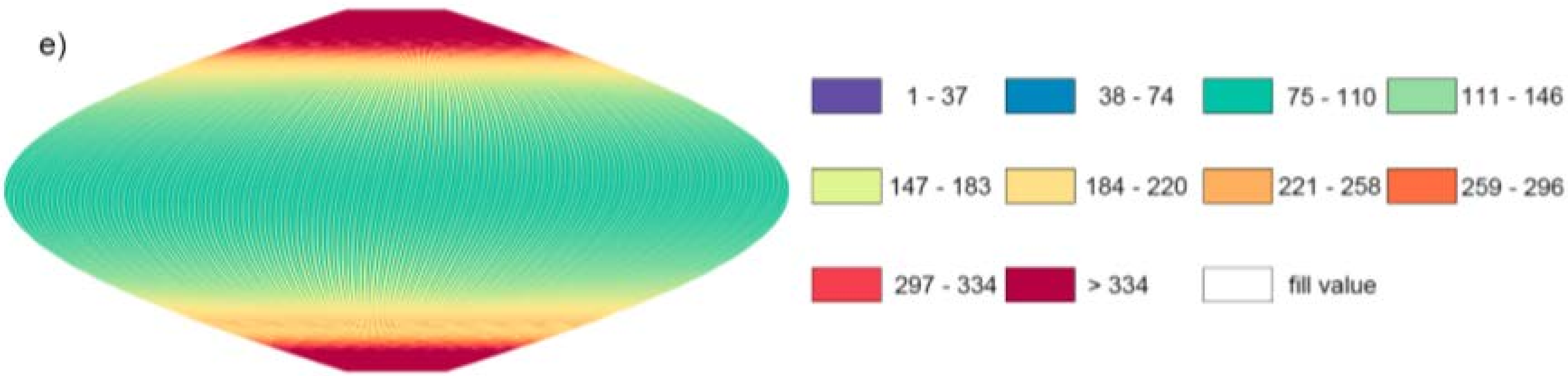



**Figure 3.** The total annual number of satellite observations (1 January to 31 December 2016) for: **(a)** Sentinel-2A; **(b)** Landsat-8; **(c)** Landsat-8 and Sentinel-2A; **(d)** Sentinel-2A and Sentinel-2B; and **(e)** Landsat-8, Sentinel-2A and Sentinel-2B. Global results derived at 7201 × 3601 points spaced every 5.559752 km, equivalent to 0.05° at the Equator, in the equal area sinusoidal projection.

# Big Data Challenges for RS

- Multi-source
- Variable spatial resolution
- Variable noise
- Missing data
- Relevance of geographic location
- Need of reference data for model building

*Requires new ways to process remote sensing data*

# Current big data processing solutions

- Parallel processing on own system
- Google Earth Engine (Cloud-based solutions)
  - Gorelick et al., 2018.
- Digital Globe’s GBDX platform using Amazon cloud with 100 Pb image data from 16 year archive
- … more to come in the future

# Some current Big Data algorithms

- Machine learning
- Active learning
- Deep ensemble learning
- Deep fusion learning
- Deep reinforcement learning
- Deep and shallow fusion
- Representation learning
- Semi-supervised deep learning
- Supervised deep learning
- Transfer deep learning
- Unsupervised deep learning

# Machine learning

…is the term for algorithms with a learning phase, and re-processing the data based on the gained knowledge to improve the final result.

We ”train” or ”teach” the program how to interpret the data with very little involvement from the analyst.

e.g.,

- Random Forest
- Support Vector Machines

# Deep learning

…is a subset of Machine Learning, but is an advanced level where the algorithm determines the input variables needed without involvement from the analyst.

The algorithm operates as a human brain might, coming to solutions without always directly being provided the answers.

”Deep” refers to the number of layers in the neural network

e.g., Neural Networks

# Deep learning

Advantages:

- Good results

Disadvantages:

- Requires large amount of training data
- Complex classifications (such as EO data) are still difficult
- Potential for overfitting

e.g.,

Convolutional Neural Networks (CNN)

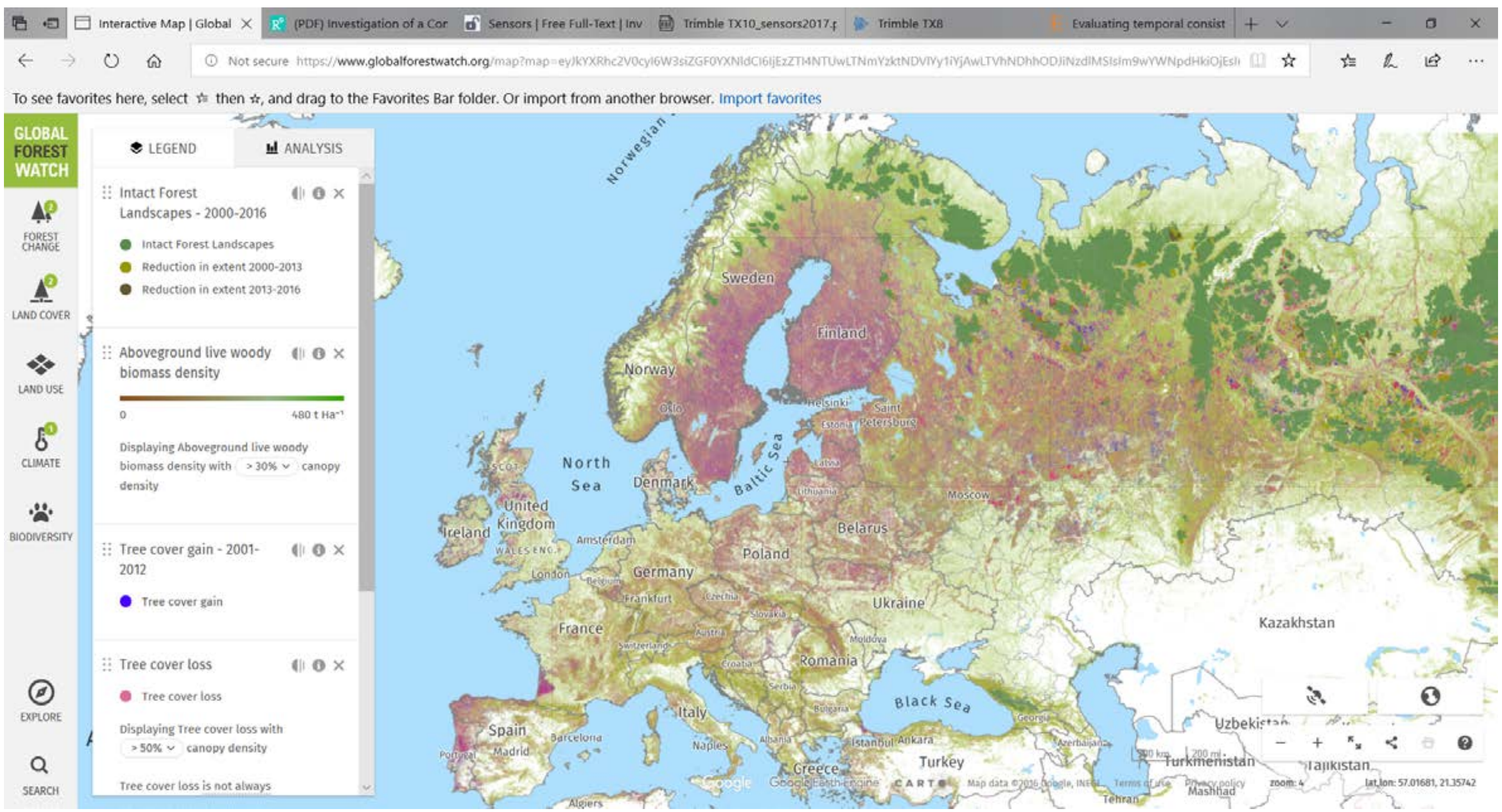
Long Short-Term Memory (LSTM)

Deep Convolutional Network

# Published case studies

Besides widely known ones, such as

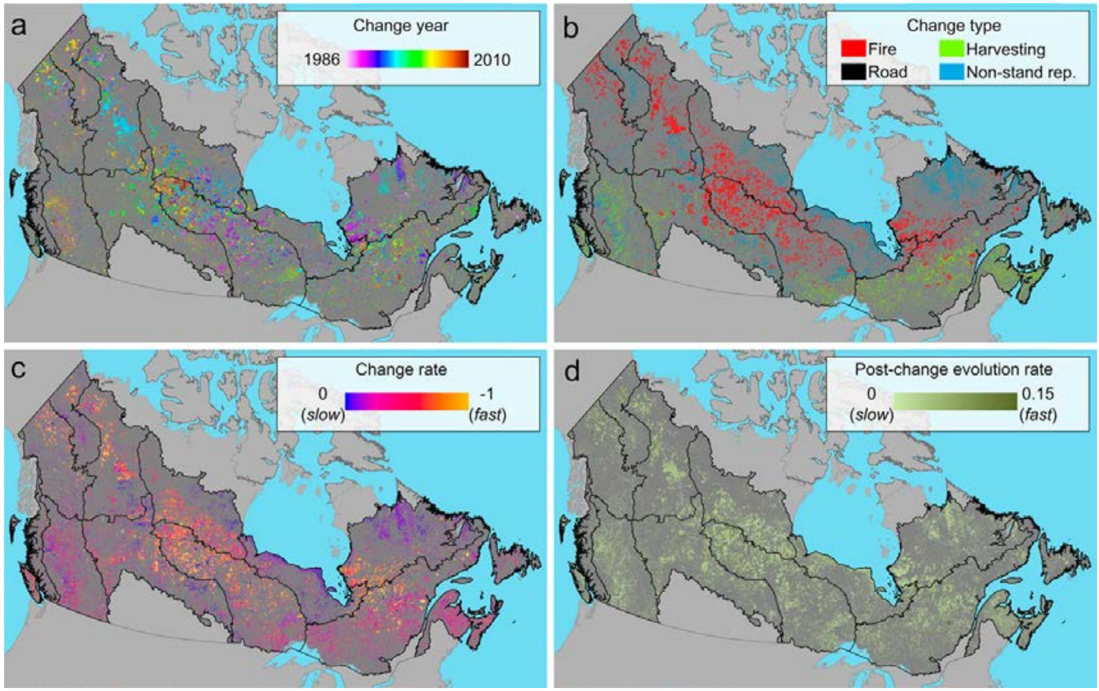
- Global forest change (Hansen et al, 2013)
- Global NDVI trends from 1984 to present (e.g., GIMMS3g)



# Published case studies

<u>Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring (Hermosilla et al., 2016)</u>

- Identified occurence and type of forest from 1984-2012
- 73,544 Landsat images
- Used spectral trend analysis per year using best-available-pixel
- 400 Tb in data products




Conclusions

- High accuracy (86% overall) for identifying that a change occurred
- Little lower accuracy when identifying change type (fire, road, harvest, non-stand replacing)

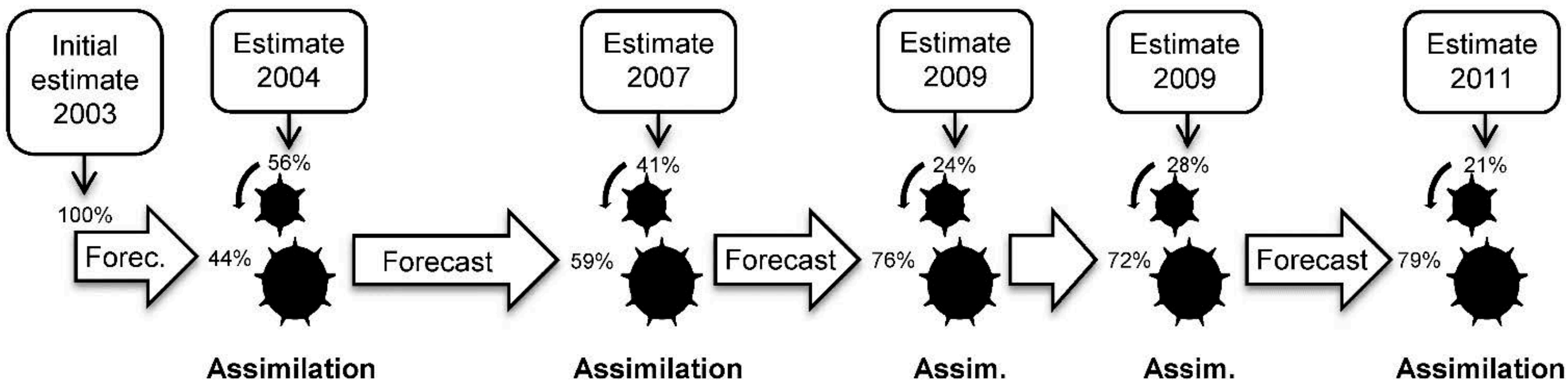
# Published case studies

<u>Data assimilation in Forest Inventory: First empirical results</u>

(Nyström et al., 2015; SLU)

- Used six dates of aerial photo point clouds (2003-2011)
- Estimated forest parameters
- Data assimilation combines information sources using weights that are inversely proportional to their uncertainty

Time 1 + Time 2 ⟶ New estimate for Time 2 ⟶ …. Time *i*



# Published case studies

<u>Data assimilation in Forest Inventory: First empirical results</u>

(Nyström et al., 2015; SLU)

**Table 4.** Root mean squared error (*RMSE*) of the deviation from the field measurement 2011 for the nine assimilated plots. In parentheses is relative *RMSE*.

| Target Variable | Assimilated | Most Recent Estimate | Forecasted |
|---|---|---|---|
| V | 40.1 (13.5%) | 44.7 (15.0%) | 58.5 (19.7%) |
| BA | 3.80 (12.0%) | 4.05 (12.8%) | 4.49 (14.2%) |
| $H_L$ | 1.81 (9.3%) | 1.86 (9.6%) | 1.86 (9.5%) |

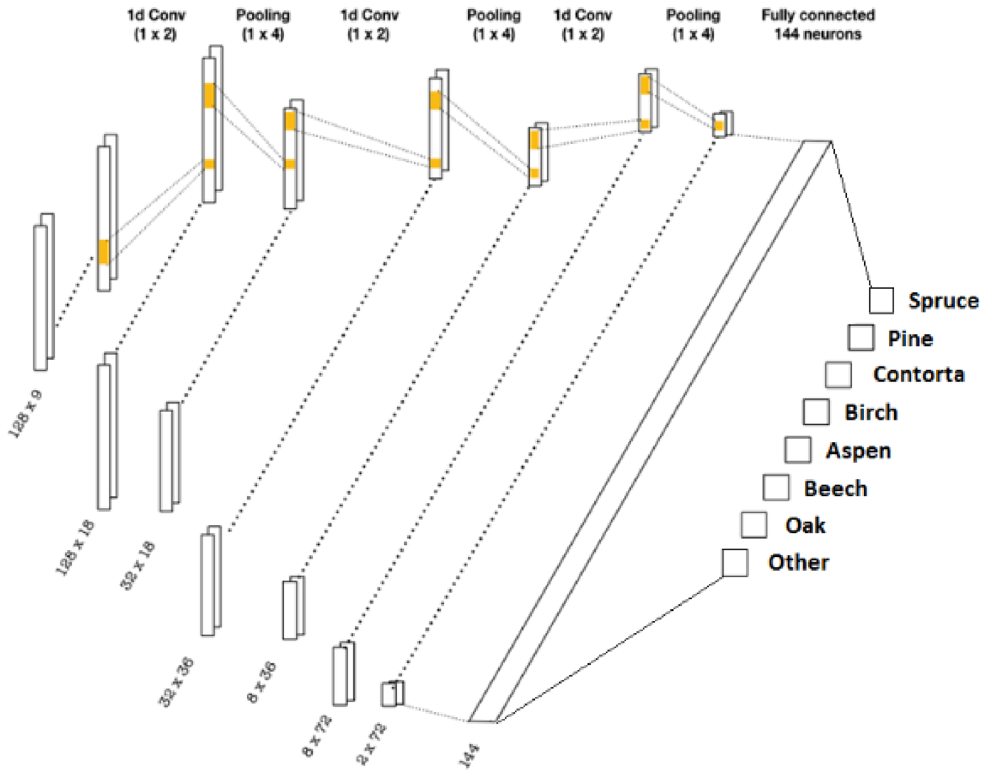
Conclusions:

- Data assimilation provided better forest parameter estimates
- High quality field data and growth models are needed for this approach

# Coming case studies

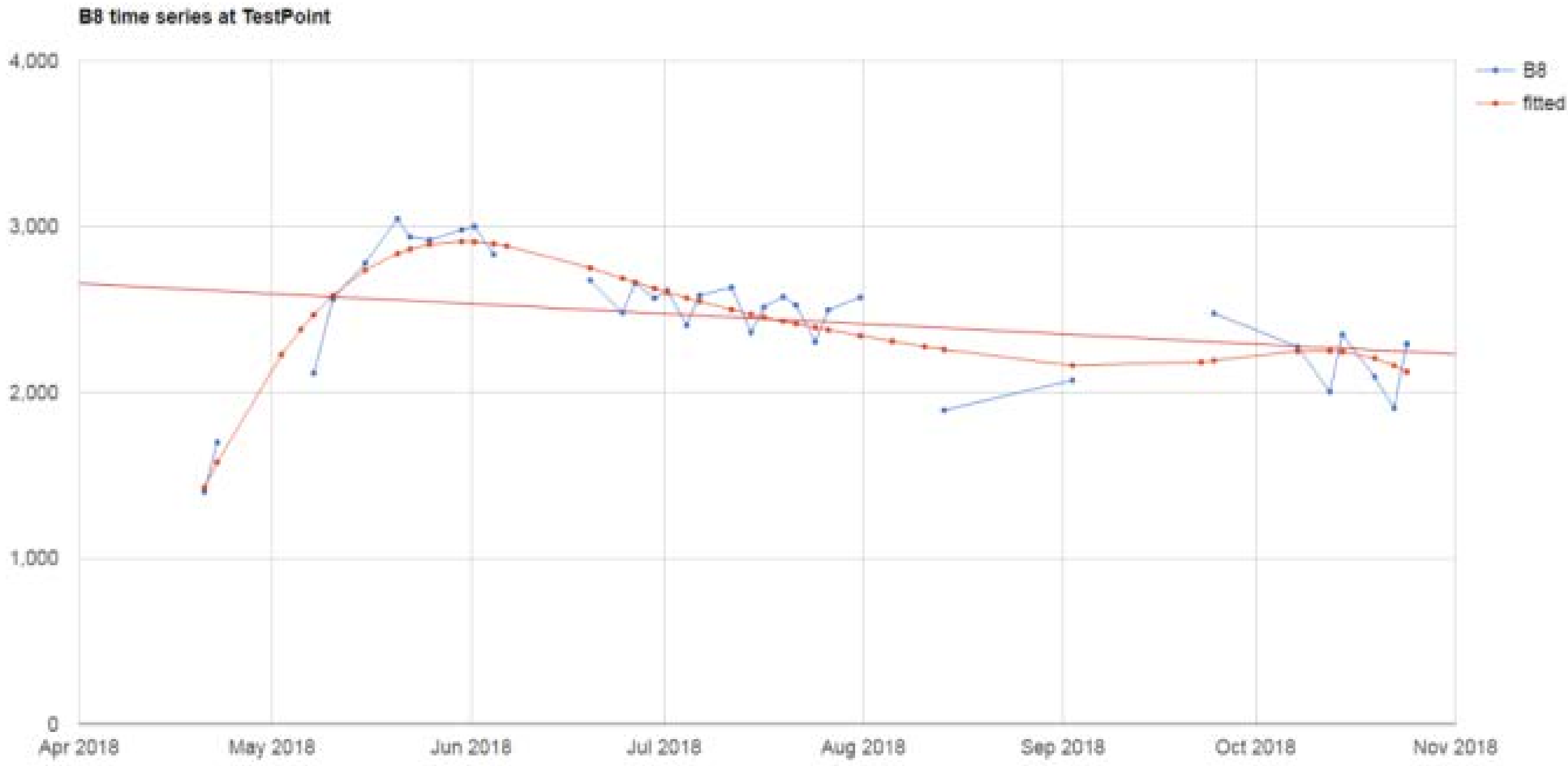
<u>Neural Networks for tree species classification using Sentinel-2, NFI and Google Earth Engine (André Wästlund, SLU)</u>

- Thirty-five Sentinel-2 images from 2018 to predict tree species
- All of Sweden + National Forest Inventory data
- Currently tested Convolutional Neural Network



# Coming case studies

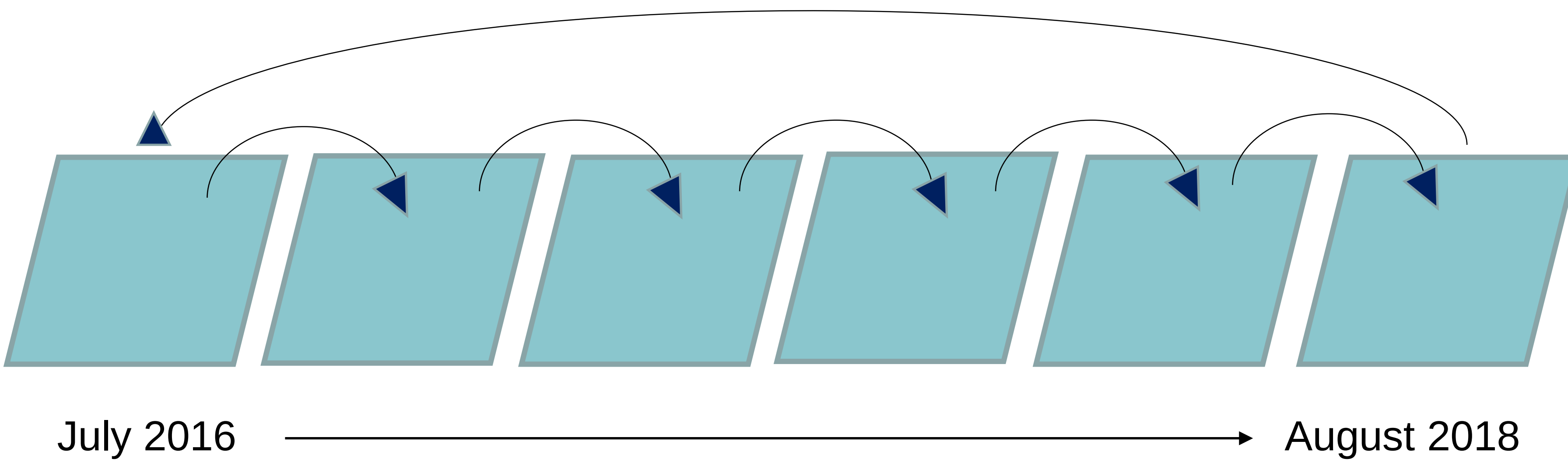
Neural Networks for tree species classification using Sentinel-2, NFI and Google Earth Engine (André Wästlund, SLU)



# Coming case studies

<u>Tree Species Classification with Sentinel-2 Time Series and Bayesian Inference </u>(Arvid Axelsson, SLU)

- Bayesian inference uses the probability of belonging to a class
- Updates the probability given the images that have come before
- Result should be a continuous improvement of the classification
- Preliminary result gave 80% overall accuracy for four tree species

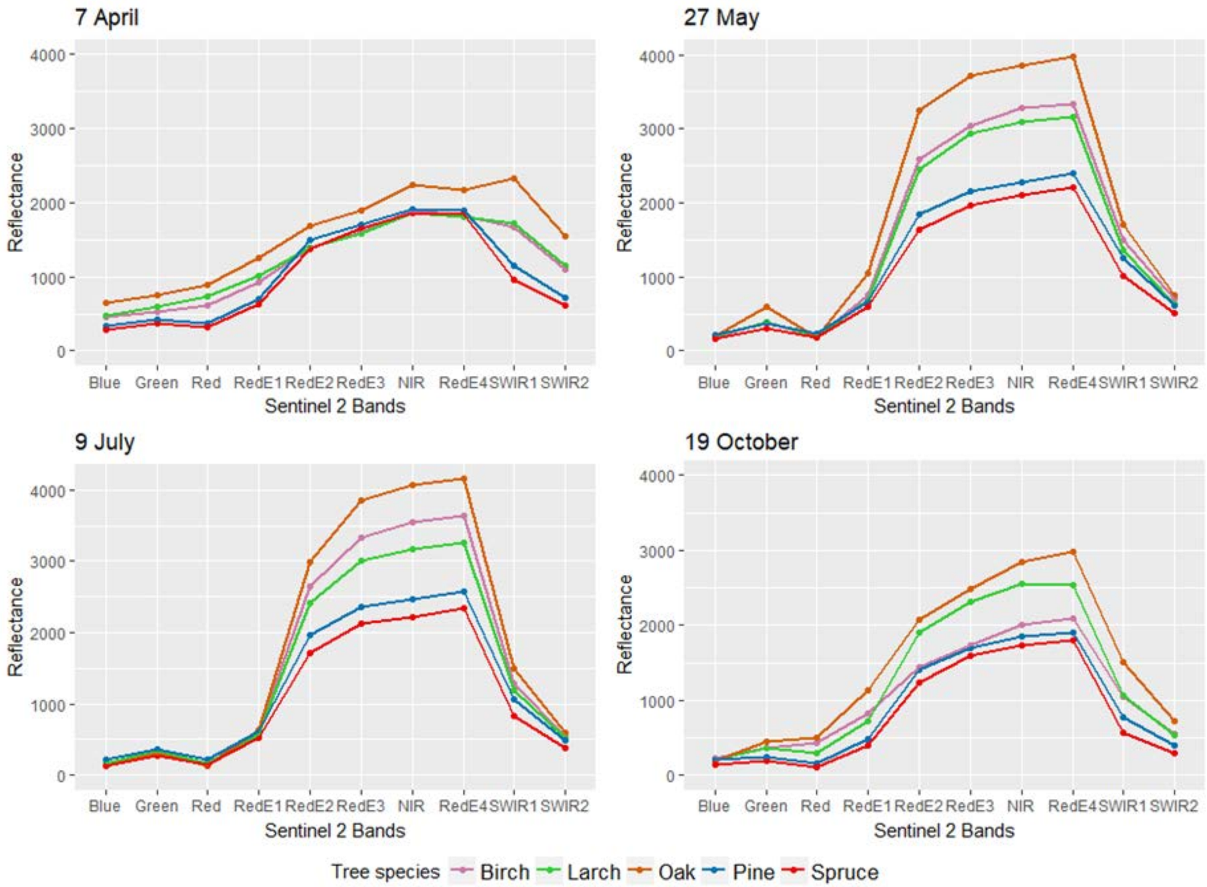


# Published case studies

<u>Tree species classification with multi-temporal Sentinel-2 data</u>

(Persson et al., 2018)

Goal: Exploratory analysis to see which Sentinel-2 dates and bands were most useful for classifying (using Random Forest) tree species



# Published case studies

<u>Tree species classification with multi-temporal Sentinel-2 data</u>

(Persson et al., 2018)

Conclusion: Two well-timed Sentinel-2 images gave 85% overall accuracy

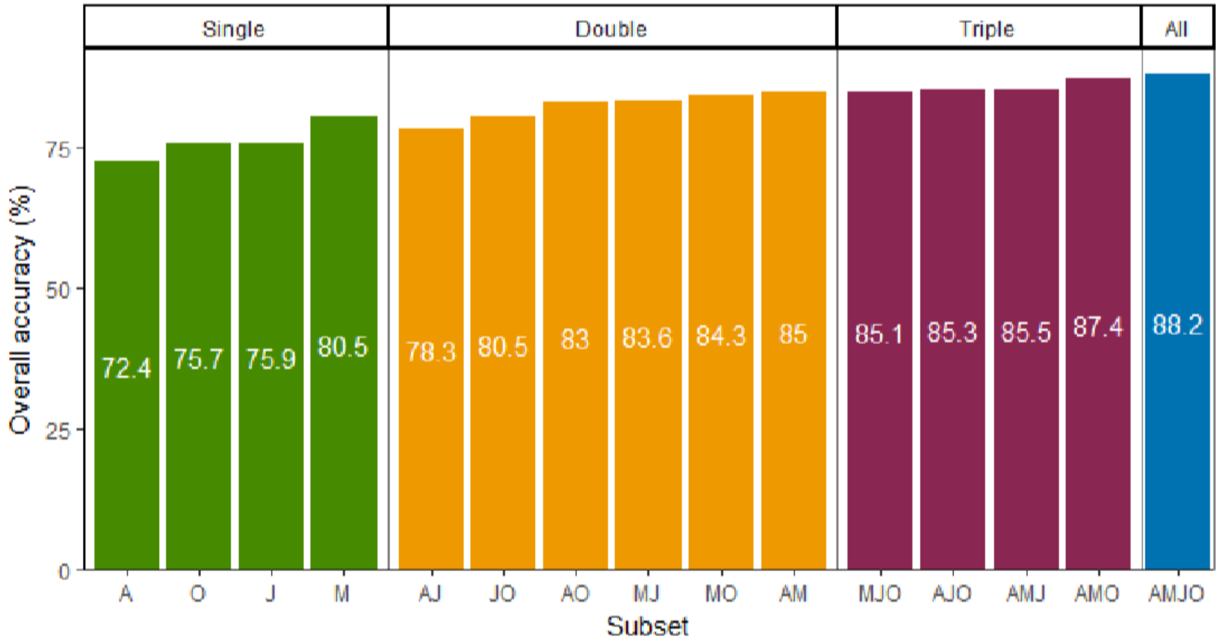



**Figure 2.** The overall accuracy (%) for individual RF models produced using all bands for single image, and multiple image combinations. The colors represent the number of satellite images used and the single letters indicate the image date: A = April 7; M = May 27; J= July 9; O = October 19.

# When do we need Big RS Data?

- Identify appropriate applications

    - Tree species at this point in time *This depends ...*

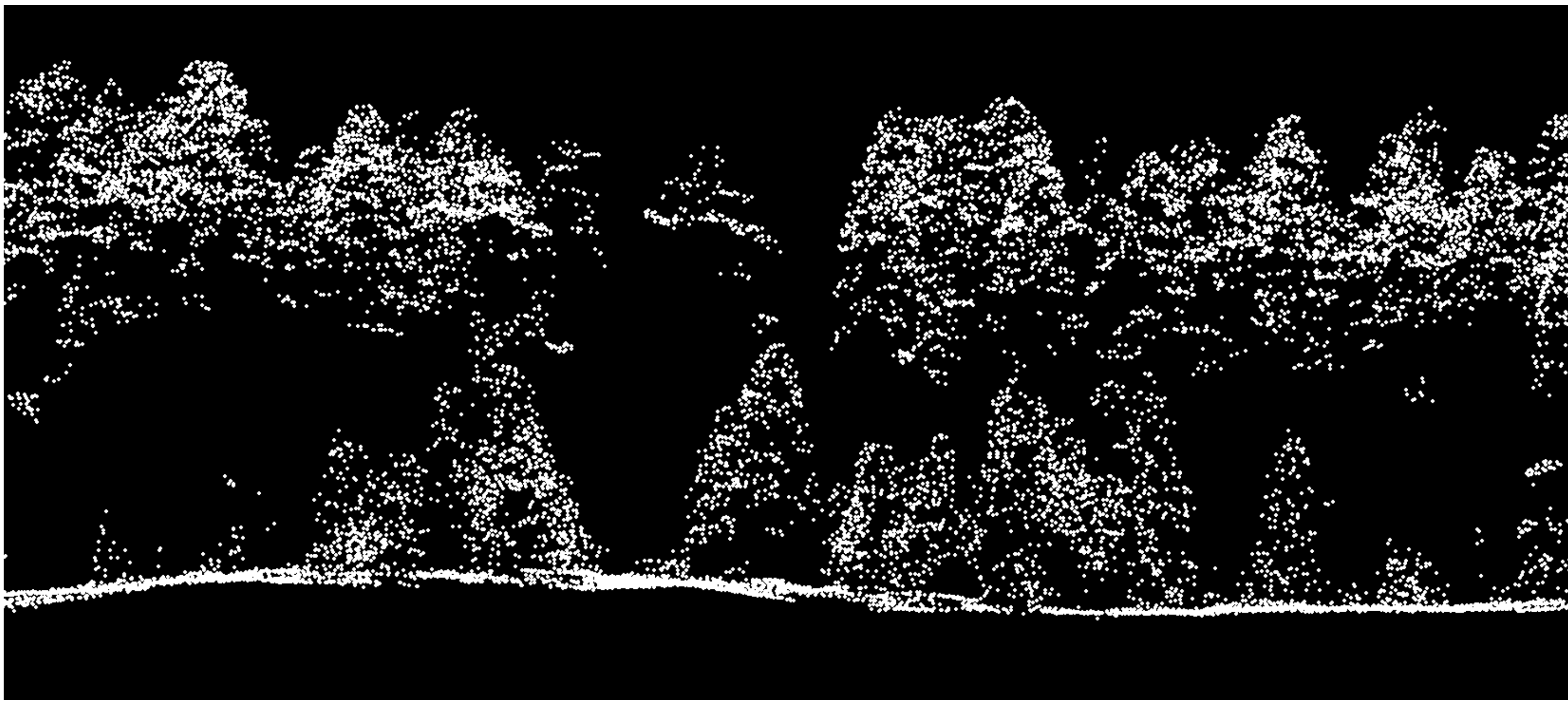
# When do we need Big RS Data?

- Identify appropriate applications

    - Tree species at this point in time
    - Long-term data studies
        - Site index (e.g., Wallerman et al., Persson et al., SLU)
        - Continuously updated forest biomass over large areas
        - Change in tree species composition over large areas
        - Long term effects of climate on forest health
        - Combining data sources (e.g., harvester data, TLS, ALS)
        - ….

Singh, K.K., et al., 2016. When Big Data are Too Much: Effects of LiDAR Returns and Point Density on Estimation of Forest Biomass. IEEE Vol 9 (7).

# More data?!
# Single Photon LiDAR

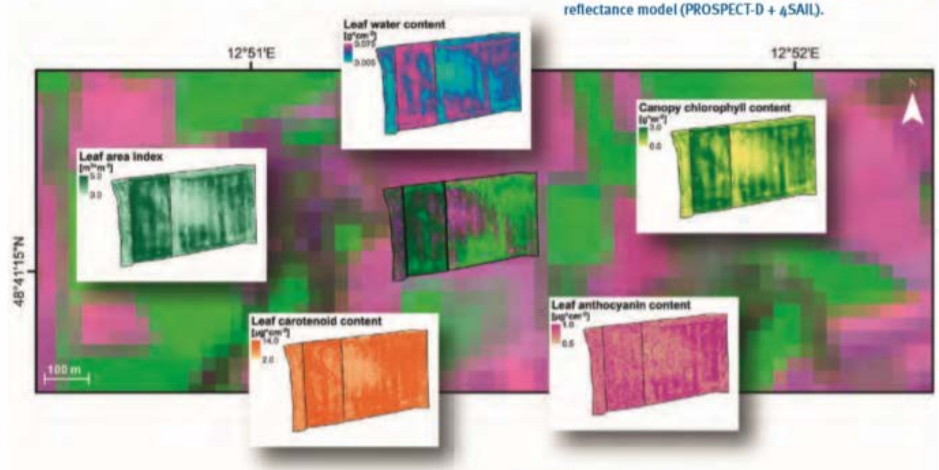


*Wästlund, A., et al., 2018. Forest Variable Estimation Using a High Altitude Single Photon Lidar System. Remote Sensing 10, 1422*

# More data?!
# EnMAP

- Hyperspectral sensor from DLR
- 88 bands i 420 – 1000 nm (Vis – NIR)
- 154 bands i 900 – 2450 nm (NIR-SWIR)
- Swath width 30 km
- Pixel size 30 x 30 m
- One image every 4 days (pointable sensor)
- Launch 2020

Fig. 8: Result of the spatial determination of various agriculturally relevant variables for a cropped field in the area near Neusling, Southern Germany, based on simulated EnMAP-Data. The estimations were achieved by inverting a canopy reflectance model (PROSPECT-D + 4SAIL).



# More data?!
# Landsat in the future

- Landsat-9
  - A copy of Landsat-8
  - Launch December 2020

- Landsat-10 (or Landsat-X)
  - Launch ca 2028
  - Will take consideration to new developments, such as
    - increased temporal resolution
    - synching with Sentinel-2,
    - creating information and not just images

# Free data = Big data

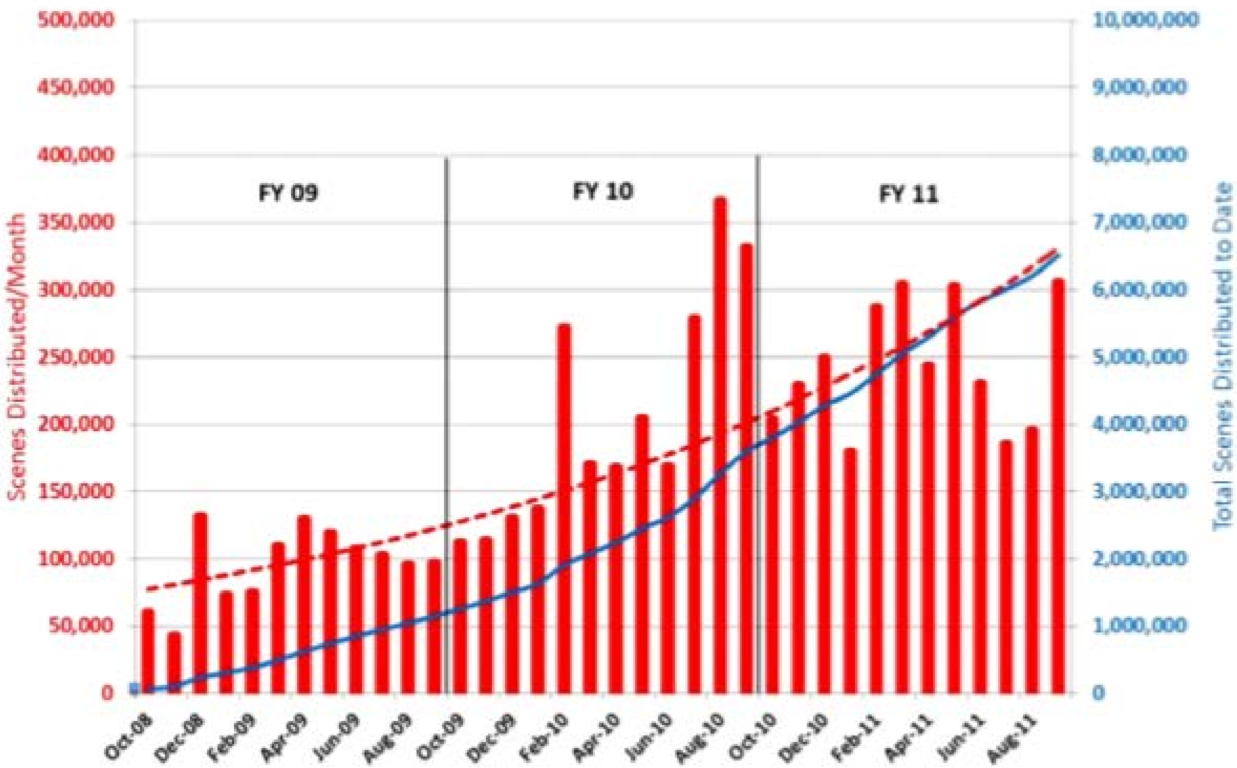



**Fig. 2.** Monthly summary of scene downloads from the EROS Data Center, covering the period from October 2008 to September 2011, further delineated by US Government fiscal year.

*Wulder et al., 2012. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. Rem Sens of Env. 2-10.*

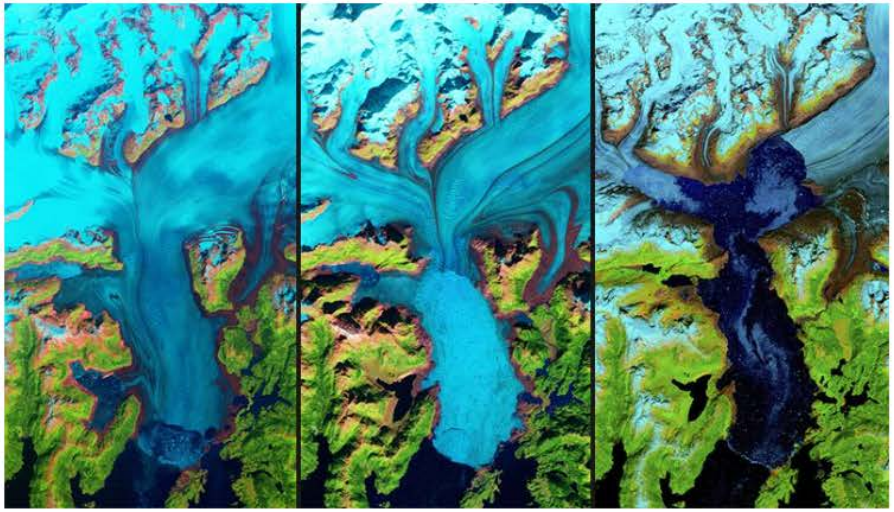
# Free data = Big data

MENU nature

# US government considers charging for popular Earth-observing data

*Images from Landsat satellites and agricultural-survey programme are freely available to scientists — but for how long?*

Gabriel Popkin



The ongoing melt of Alaska's Columbia glacier is revealed in these images captured by the US government's Landsat satellites in 1986, 1999 and 2017. Credit: Landsat/EO/NASA

PDF version

RELATED ARTICLES

Earth-observing companies push for more-advanced science satellites

Landsat 8 to the rescue

Satellite alerts track deforestation in real time

Satellite images reveal gaps in global population data

# References


- Gorelick, N. et al. 2017. Google Earth Engine: Planetary-scale geospatial data analysis for everyone. Rem. Sens. Of Env. 202:18-27.
- Hansen, M. C., et al.,. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850–53.
- Hermosilla, et al., 2016. Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring. Intl J of Dig Earth, Vol 9 (11):1035-1054.
- IBM (https://www.ibm.com/analytics/hadoop/big-data-analytics)
- Kussul, N. et al., 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. IEEE Geosci and RSL, Vol 14, No 5.
- Li and Roy, D. 2017. A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring
- Ma et al ,2015. Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems. 51:47-60
- Mirt, A. et al., 2018. Big Data Analytics on combining radar and optical satellite data. Presentation.
- Nyström, M., et al., 2015. Data Assimilation in Forest Inventory: First Empirical Results. Forests 2015, 6(12), 4540-4557
- Persson, M., Lindberg, E., and Reese, H. 2018. Tree species classification with multi-temporal Sentinel-2 data. Remote Sensing, 10(11):1794.
- Singh, K.K., et al., 2016. When Big Data are Too Much: Effects of LiDAR Returns and Point Density on Estimation of Forest Biomass. IEEE Selected topics in App. Earth Obs. and Rem. Sens. Vol 9 (7).
- Wulder et al., 2012. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. Rem Sens of Env. 2-10.
- Wästlund, A., et al., 2018. Forest Variable Estimation Using a High Altitude Single Photon Lidar System. Remote Sensing 10, 1422


Then watch for publications from André Wästlund and Arvid Axelsson