The National Academies of SCIENCES ENGINEERING MEDICINE

Exploring a Dynamic Soil Information System: A Workshop

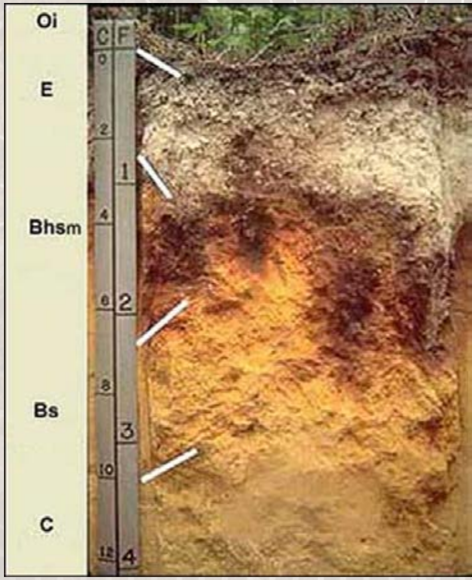
A Virtual Workshop

Breakout Topic A: Measurement and Sampling and Archiving

Charles Rice and Marcos Sarto

What should be measured in soils?

- **What is the question to solve?**
  - **Scientific, policy, land management, soil resources**
  - **Soil Science theories**
- **Every agency or project has its individual goals at temporal and spatial scales.**
- **At what temporal and spatial scale. Depends on objectives**
  - **Erosion at decadal scale**
  - **Gas flux at minutes**
- **Soil Information system needed for models**
- **The challenge is to integrate different projects to have a more robust soil information system.**

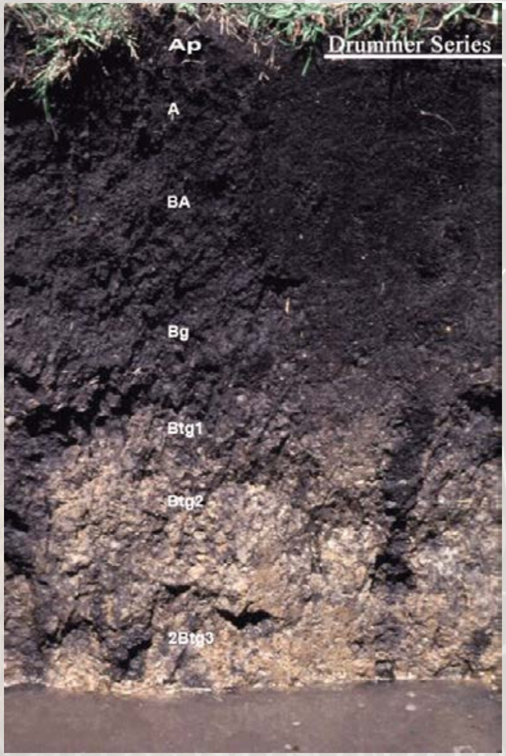


What should be measured in soils?

- **Physical: Texture; Aggregate Stability, Bulk density**
  - *Bulk density cannot be archived*
  - Soil water content
  - Color, aeration status
  - soil pore distributions
  - Mineralogy
  - Horizons
- **Chemical: Soil organic carbon; pH**
  - Nutrients
  - SOC fractions
- **Biological: Methods are evolving**
  - Diversity genomics
  - Phospholipid Fatty Acids (PLFA)
  - Do not forget about simpler measure chloroform fumigation (Microbial biomass but also active C fraction)
  - Enzymes
- **At Depth**

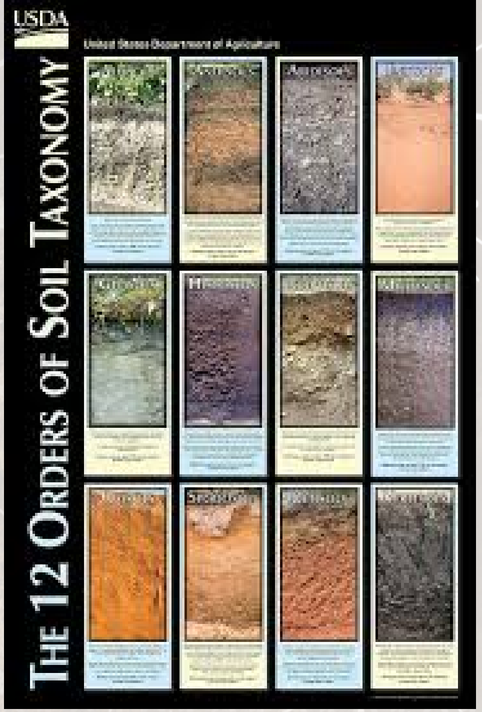
- We did not discuss at what temporal and spatial scale. Depends on objectives
    - Erosion at decadal scale
    - Gas flux at minutes
- Meta data is the biggest need/deficiency
- Location
- Weather
- Land use history
    - Vegetation type and biomass, ROOTS
    - If ag tillage, fertility, yield, crop rotation
    - If urban contamination
- Sensing
    - Aerial: elevation, productivity, tillage, vegetation type
    - Belowground: measure dynamic properties

    - *Need new techniques to avoid physical extraction of the soil*



- Standardization
  - Hard to standardized based on soil chemistry
    - Bray P or Olsen P or Mehlick P
  - Need to be able to compare method and data
- Archiving samples
  - Dry
    - Storage conditions (temperature)
  - Biological
    - -80°C
    - Extract DNA and freeze
    - Enzymes can be from dried samples
  - Issues
    - Space
    - Other examples
      - Svalbard Global Seed Vault in Norway

Reference Sites



- Need to have a network of reference sites
  - Calibrate new methods
  - Calibrate models
- Use USDA-ARS, NEON, LTER, Land Grant University
  - Have each land grant dedicate a reference site
  - Would allow soil and climate variability

  - Submit proposal to NSF for RCN

Breakout Room B
Collection and Curation

- Harmonization of data and methodology
- Semantic and ontology structures
- Data models
- Avoiding the proliferation of standards
- Internet of Things
- Cloud infrastructure for data storage and data processing
- Data repository federations
- Security/privacy issues
- Machine learning tools in databases
- Data platforms

Breakout Room B: *Collection and Curation*

- How do you make the process FAIR (findable, accessible, interoperable, reproducible)?
- Incentivizing sharing
  - Journals
  - Younger gen more open to sharing, also older people
  - Need to be consistent with industry
  - Buy-in from stakeholder: academia, government, private sector, farmers
  - New business models, e.g. Carbon markets? Crop insurance?
- How to find data?
  - Semantic web, link data, knowledge graphs
  - DOI citation to link datasets

Breakout Room B: *Collection and Curation*

- How to train people to use data?
  - New NLP AI Schemes, e.g. conversational AI, machine teaching
  - Wolfram’s Natural language for non-coders
  - Visualization is an important problem
- Data annotation
  - Who annotates the metadata

Breakout Room B: *Collection and Curation*

- Data fidelity/veracity?
  - Tradeoff:
    - Share the entire data
      - Capture meta data, record methods
    - Have strict cleaning procedures
      - Callibration of sites and data
- What level of data is shared?
  - Temporal, spatial – maybe through ontologies
  - Need level of privacy

Breakout Room B: *Collection and Curation*

- Data storage
  - Centralized vs distributed
  - Cloud has benefits, but pricey for downloads
  - Regulatory concerns prevent centralization
- AI on soils data (can we predict the state of soil)
  - Bias
  - Explainability
  - Ethics
  - Need to communicate uncertainty

The National Academies of SCIENCES ENGINEERING MEDICINE

Exploring a Dynamic Soil Information System: A Workshop

A Virtual Workshop

Breakout Topic C Data Analysis and Models

Bruno Basso, Allison Marklein, Katherine Todd-Brown, Rafael Martinez-Feria

Exploring a Dynamic Soil Information System: A Workshop

The National Academies of SCIENCES ENGINEERING MEDICINE

What is the promise of the current ML/AI methods?

- It all depends on what we are trying to model, and any modeling approach is objective driven.
- ML can help generate new hypotheses to be verified experimentally or with process-based modeling.
- Ensemble modeling of ML and process-based model to build metamodels
- Triangulation of different methods (ML, process-based model, Remote Sensing, measured data on top of ML

- ML can be used to predict management practices based on detailed farmer surveys.
- ML can help find missing data/fill in data gaps.
- ML can help determine when/where we need take measurements.
- Optimization can be augmented with ML – helps explore the parameter space more efficiently.
- ARPA-E SMARTFARM has new sensors/ways of measuring things; AI and RS can be used to model observed data (N2O)

What are the concerns in AI?

- The ML/AI are black boxes. It is hard to fully understand how ML makes decisions.
- We need to match the methods with the right questions.
- Cannot extrapolating beyond the training data
- Frequent overfitting
- How do we use ML/AI to derive insights (ML to Machine Teaching/Mknowing)?
- Data availability for ML is limited, validated process-based model can help overcome this issue because they can generate data/ fill gaps. We are data starved.

- We should not use ML alone for sparse data streams. It behaves very unpredictably. I see more potential for a hybrid PM-process-based approach. It is hard to decouple the noise from data with ML.
- ML is another tool, doesn't replace other tools.
- Modeling predictions without knowing the background: legacies, site history, previous management, water table, in-season heterogeneity
- How do you harness the power of ML while preventing abuses?

What are the challenges with integrating models with data? How can it be improved? Parameterization, validation, and benchmarking? Matching model in/out with measurements?

- Bayesian methods for parameter estimation

- Process-based models should be better at predicting the future, while AI might not have all the data (for training).

- You have areas where yields are high, others where they are low consistently, this allows to identify key uncertainties and use proxies for improving inputs to the models.

- Modeling SOC, nutrient and water dynamics require proper simulation of crop yield, biomass, roots partitioning, uptake, water balance, etc.
- We capture the variability of using crop history from yield maps and remote sensing (through yield stability maps and thermal stability maps).
- Learned from AgMIP (ensemble and median of models, overcalibrating)
- Scales for measurement may not be the same as those needed by most models to make predictions (microbial respiration flux towers). How do we link those?
- We are generating new types of data: Metagenomics, QTL

- Dynamic soil information system needs to have site history metadata to help modelers.
- Ideally would like to see a nested hierarchy of models to upscale predictions from different models.
- Information systems needs to have repeated measurements to reduce the uncertainty.
- What about underground sensors? To measure N20, VOCs. Where are we in the development of these sensors? Can they be integrated too?
- There is untapped potential to linking remote-sensing with process-based models, RS to inform/correct processes and input within the model on the fly.

- We need something like LUCAS!
- Organized soils databases to make it easy, usable to modelers
- We need a system that helps us synthesize from multiple different sources.
- Information system needs to be able to link multiple datasets to help us contextualize the data.
- Works with global assessments and it is hard to cobble data together, different sources, resolutions, etc. Every time is a custom data stream.
- How do you select important covariates from different sources?
- We need better soil ontologies.
- Communicating the uncertainties and measure of risks

- Small sample, large uncertainty. We need to be more transparent
- Increasing the sample data for ML doesn’t make it more accurate if your population has a bias.

Conclusion

- Setting scale and the objective of any modeling approach is critical to choose right type of model system
- ML great promises as well as concerns
- Tradeoffs between model complexity vs simplicity – right balance in modeling processes
- Integration and fusion of domain knowledge to make sound assumption, ML, remote sensing, observed data and process-based model is key
- We need soil ontologies and FAIR standard and file format
- We need to improve the resolution of the data

The National Academies of SCIENCES ENGINEERING MEDICINE

(Some) opportunities to stay engaged in soil informatics…

Exploring a dynamic soil information system: A workshop

Kathe Todd-Brown (University of Florida)

4 March 2020

Data resources from group

- **USDA-NRCS - Soil Database Interface**
  - http://ncss-tech.github.io/soilDB/docs/
- **National Microbiome Data Collaborative**
  - https://microbiomedata.org/
- **Agricultural Collaborative Research Outcomes System, AgCROS**
  - https://agcros-usdaars.opendata.arcgis.com/
- **The Long-Term Agroecosystem Research Network**
  - https://ltar.ars.usda.gov/
- **NASA's Soil Moisture Active Passive**
  - https://smap.jpl.nasa.gov/
- **National Geochemical Survey database**
  - https://mrdata.usgs.gov/geochem/
- **The Long-Term Ecological Research Network**
  - https://lternet.edu/
- **Consortium of Universities for the Advancement of Hydrologic Science, Inc.**
  - https://criticalzone.org/learn-more.html
- **National Ecological Observatory Network**
  - https://www.neonscience.org/
  - https://www.neonscience.org/resources/learning-hub/teaching-modules

Data resources from group

- IIASA/FAO - Soil Revealed
  - https://soilsrevealed.org/
- ISRIC Soil Data Hub
  - https://www.isric.org/explore
- SoilGrids web portal
  - **https://soilgrids.org/**
- ESA WORLDSOILS
  - https://www.isric.org/projects/esa-worldsoils
- Soils 4 Africa - Horizon 2020 programme of the European Union
  - https://www.soils4africa-h2020.eu/
- OpenGeoHub - OpenLandMap
  - https://gitlab.com/openlandmap/compiled-ess-point-data-sets
  - http://www.openlandmap.org/

Data resources from group

- **MESONET**
  - https://climate.umt.edu/mesonet/default.php
- **Soil Health Kit created by Colorado State researchers**
  - https://smallholder-sha.org/protocol-1/
- **Open Data Science**
  - https://opendatascience.eu/
- **Soil Spectroscopy 4 Global Good**
  - https://soilspectroscopy.org/
- **Global Soil Biodiversity Initiative**
  - https://www.globalsoilbiodiversity.org/objectives
- **International Soil Moisture Network**
  - https://ismn.geo.tuwien.ac.at/en/
- **International Soil Carbon Network**
  - https://iscn.fluxdata.org/

Examples of a non-soil data repository / information system?

- World Resources Institute
  - https://www.wri.org/
- Resource Watch
  - https://resourcewatch.org/
- Global Biodiversity Information Facility
  - https://www.gbif.org/
- Avian Knowledge Network
  - http://avianknowledge.net/index.php/about-akn/
- CUAHSI's Hydrologic Information System
  - http://his.cuahsi.org/

Resources from Breakout Room B

- **DOE Joint Genome Institute-A Functional Genomics Database for Plant Microbiome Studies**
  - **https://jgi.doe.gov/functional-genomics-database-for-plant-microbiome-studies/**
- **DOE Joint Genome Institute-National Microbiome Data Collaborative Now on Trellis**
  - https://jgi.doe.gov/join-national-microbiome-data-collaborative-trellis/
- **Zooniverse, citizen science**
  - http://zooniverse.org/

Data resources from group – highlighted funding

- **Signals in the Soil | NSF - National Science Foundation**
  - https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505752
- **Center for Advancement and Synthesis of Open Environmental Data and Sciences | NSF - National Science Foundation**
  - https://nsf.gov/funding/pgm_summ.jsp?pims_id=505829

Other soil-data organizations

- **Coastal Carbon Research Coordination Network**
  - https://serc.si.edu/coastalcarbon
- **Ecological Forecasting Initiative**
  - https://ecoforecast.org/
- **International Soil Modeling Consortium, DO-Link science panel**
  - https://soil-modeling.org/science-panels/DO-Link
- **International Soil Radiocarbon Database**
  - https://soilradiocarbon.org/

Talking about soil informatics: Conferences

- **Big society annual and specialty meetings**
  - American Geophysical Union, European Geophysical Union
  - Soil Science Society of America, Ecological Society of America, Soil Ecology Society, and Geological Society of America
- **Bi-annual meetings**
  - Earth Science Information Partnership
  - Research Data Alliance
- **Every other year**
  - International Data Week (https://codata.org/events/conferences/international-data-week-2021/)

Where to do collaborative science: Working groups

- **Earth Science Information Partnership**
  - Clusters: “Soil Ontology and Informatics” and others https://www.esipfed.org/get-involved/collaborate
- **Research Data Alliance**
  - Agriculture working groups: https://www.rd-alliance.org/
- **Global Soil Partnership - GLOSOLAN (Global Soil Laboratory Network)**
  - Lab method harmonization and lab quality improvement http://www.fao.org/global-soil-partnership/glosolan/en/
  - Soil spectra library development http://www.fao.org/global-soil-partnership/glosolan/soil-analysis/dry-chemistry-spectroscopy/en/

The National Academies of SCIENCES ENGINEERING MEDICINE

Exploring a Dynamic Soil Information System: A Workshop

A Virtual Workshop

Closing Synthesis

Jim Tiedje

March 4 2021

What are the Values of a DySIS?

- **Agricultural**
- Increased productivity
- Increased profitability
- Sustainable or even enhanced land quality
- **Environmenta**l
- Improved ecosystem services
- Improved soil & water quality
- Contributes to global goal of GHG mitigation
- **Fundamental Planetary Science**
- Hypotheses about scale, about trajectories, lithosophere
- About interactions between bio, chem, phy
- Ecological, evolutionary change, mechaisms

What Are the Functions?

- Support: crops, cities, roads, forests, cactus, tundra….
  - we don't know the land use in the future
- Store water, recycle water, distribute water, safe water
- Supply nutrients: recycle, bank
- Reliable biogeochemical cycles, C,N, P. S, Fe and more
- Planetary clean-up, at least of the terrestrial

- Jerry Hatfield
"It's about Water & Carbon"
"Understand dynamics and Interdependences"

What Can (Should) We Measure?

- Standard, historical proven core chemical & physical measures
- Plants as sensors, roots as the window below
- **Scaling:** in time & space, aggregate to ecosystem level modeling
- Microsensor to drone to satellite-based technologies
- Multidisciplinary needs
- **Breakout A summary**

-Joe Cornelius

Soils at What Scale? Global of Course

- Management, change over time, projections
- C mining, erosion
- Importance of historical data; ***time series*** an extremely valuable resource
- ***Soil archives***, valuable
- Connectivity and integration of data & efforts
- Value in diverse approaches
  Systematic continental & grass roots, boutique

-Alison Hoyt

What About the Big Data?

Breakout B

- FAIR principles
- Sharing incentives
- Finding, tracking data
- Key, metadata, annotation
- Data fidelity
- Data privacy
- Data storage
- Training

Breakout C

- Date use
- Models
- Uncertainty
- Tradeoffs
- Integration & fusion
- Ontologies

Data Science Era, probably 1/3 of the “effort”

What has Been Done?

- A lot
- Existing data
  - Archived, various status, different country interests
  - Data at risk – that between eras of science & data mgt
- Archived soils
  - Air dried, corrections for
  - Time series
  - Example: Dutch stored samples and antibiotic resistance

What is Being Done?

- Here: In USA, in North America
- Elsewhere
- Lessons Learned
    - Values
    - Scales of effort
    - Resourcing
    - Users
    - Nimbleness

Our Listening Sessions: Valuable Resource

US AGENCIES

- NRCS
- NOAA
- NEON
- US Forest Service
- Soil Health Partnership
- University of Minnesota
- NCEAS

INTERNATIONAL AGENCIES

- FAO: Global Soil Partnership
- ISRIC
- CSIRO
- ISRaD
- Prairie Soil Carbon Project (Canada)
- Chinese Academy of Sciences
- JRC European Soil Data Centre
- iSDA
- Rothamsted

PRIVATE SECTOR

- Gates Foundation
- METER Group
- dataONE
- Cyverse
- Viresco

From Alison Marklein's Day One Summary

What we learned: Challenges

- Continuous funding - monitoring
- Understaffed for soil science and/or data analytics
- Data privacy and security
- Data naming conventions, harmonization
- Common methodology including sampling and analysis procedures
- Error associated with methodology, analysis, facilities
- Capturing spatial and temporal data at varying scales
- Destructive samples

From Alison Marklein's Day One Summary

Recommendations

- Funding monitoring for long-term understanding
- More repeated measurements
- More communication between agencies
- Increased clarity on nomenclature, etc.
- More metadata on sampling methodology and processing
- Data at scale relevant to farmers
- Archived soils for future analysis

Coales

A coalescence

From Alison Marklein's Day One Summary

How Are We Interacting, Our Connectivity?

Is This - “DySIS” - Big Science?

If so, Plan Like It Is:

- Organize at several levels & components
  - Leadership & community building
  - Develop major guiding science questions
  - Build support base in several sectors
  - Heavily multidisciplinary - engage

If a Big Science Approach is Taken?

- What are the downsides, dangers of a BiG Science effort?
  - Fragmentation of community
  - Over designed, sinks on its own massive weight
  - Sustainable resources, long-term commitment is difficult
  - Keep dynamic, visionary, but don't lose focus
- What is the plan fails, is all lost, effort wasted?
  - No, the planning will provide for better pieces that can be pursued
  - Likely larger effort funded than if bottom-up

Learning from Other "Models"

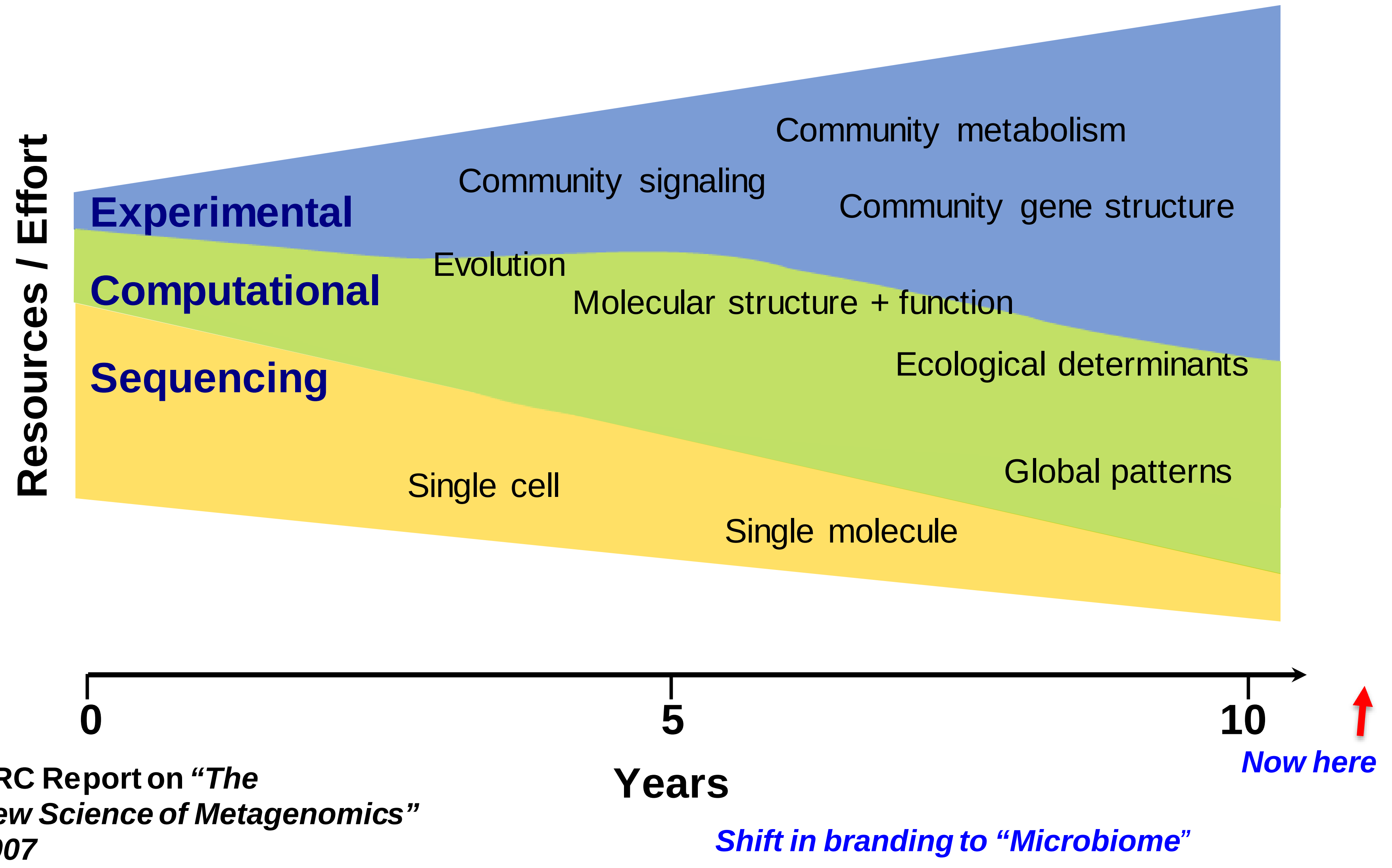
- Marine science, oceanography
  - the other 2/3 of Earth's surface
  - parallel in many aspects: chemical, biology, physical
- Atmospheric
- Astronomy, decadal plans
- Particle, Neutrino Physics, decadal plans
- Human, road map initiatives
- Geosciences

  - Infrastructure needs
  - Science drivers
  - Organizational strategies

Learning from Others

- Lessons from the restroom
- GSC (Genomic Standards Consortium)
- Lessons from the physicists, Deep Science NSF/DOE
- NSF-MREFC
    - NEON, LIGO, DKIST
- GenBank, INSDC (Intl nucl seq db collaboration)

Possible Shifts in Emphasis as the Field of Metagenomics Develops



NRC Report on *"The New Science of Metagenomics"* *2007*

Thank *you** & to our previous presenters

- David Lindbo, U.S. Department of Agriculture–Natural Resources Conservation Service
- Ronald Vargas and colleagues, FAO's Global Soil Partnership
- Marina Skumanich, NOAA's National Integrated Drought Information System
- Maria Bowman, Soil Health Partnership
- Rik van den Bosch and colleagues, ISRIC
- Andrew Bissett, Mark Farrell, and V. Gupta, CSIRO
- Linda Kinkel, University of Minnesota
- Brian McConkey and Ben Ellert, Prairie Soil Carbon Project
- Mark Schildhauer, National Center for Ecological Analysis and Synthesis
- Cory Lawrence, Susan Trumbore, Alison Hoyt, and colleagues, ISRaD
- Luca Montanarella and Panos Panagos, JRC European Soil Data Centre
- Zhongjun Jia, Xian-Zhang Pan, and Ganlin Zhang, Chinese Academy of Sciences
- Lee Stanish and Samantha Weintraub, NEON
- Christian Witt and colleagues, Gates Foundation
- Colin Campbell, METER Group
- Andy Macdonald and Richard Ostler, Rothamsted Research
- Parker Antin, Cyverse

* >280

Exploring a Dynamic Soil Information System: A Workshop

The National Academies of SCIENCES ENGINEERING MEDICINE

Concluding remarks

- **Dynamic soil information system (DSIS)** requires a coordinated interdisciplinary system approach with multiple objectives: from monitoring minimum datasets at benchmark sites using different techniques (sampling to sensing) to modeling (ML to process based) to insight to guide decisions; and where scale is clearly identified in each sub-objectives (e.g. soil biology) and resulting data and knowledge is available to continue to have healthy soil support our lives

- **Funding** from public and private parternship to make DSIS real and sooner than later