



Data-Intensive Research in Education: NSF Initiatives in “Big Data” and Data Science

Chris Dede

Harvard University

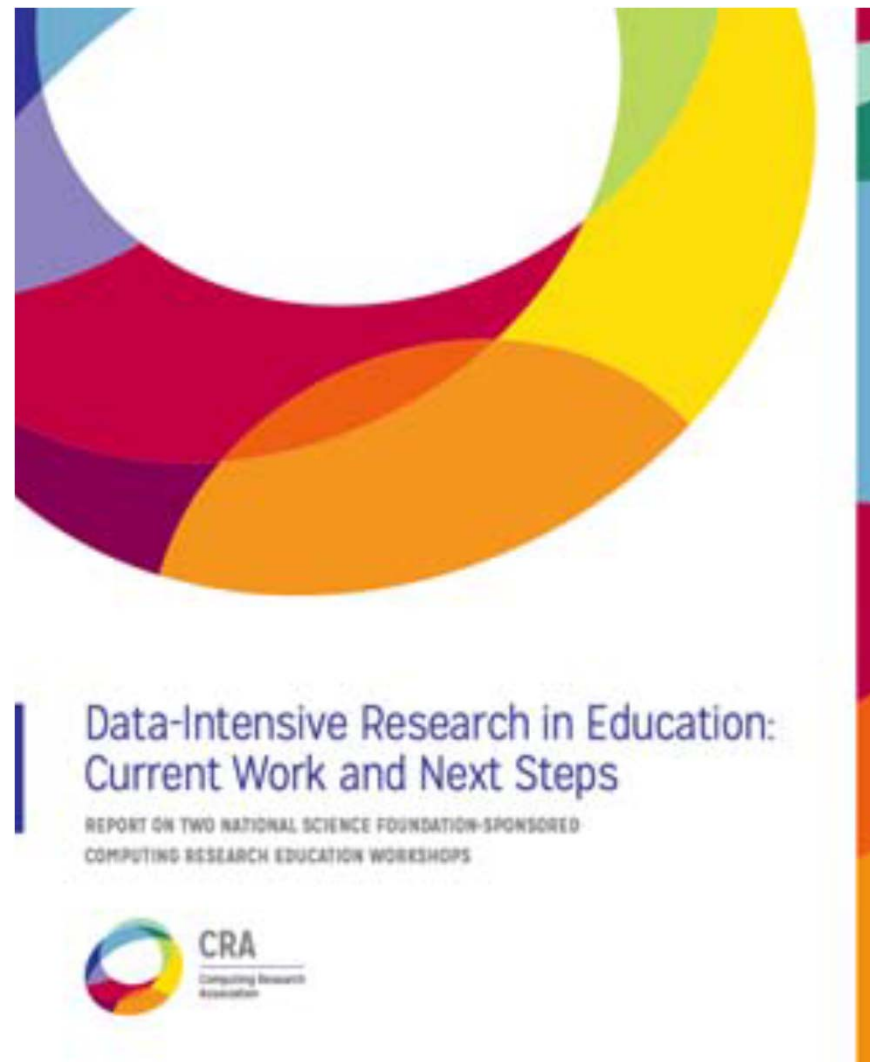
Chris_Dede@harvard.edu

www.gse.harvard.edu/faculty/christopher-dede

My Current Role in Data-Intensive Research in Education

- Confront “big data” issues in my design-based research in ecosystems science education
- Organized a two workshop sequence on data-intensive research for NSF and the field: insights from relatively mature data-intensive research initiatives in the sciences and engineering were applied to nascent data-intensive research efforts in education

<http://cra.org/cra-releases-report-on-data-intensive-research-in-education/>



Definitions

- Big Data is characterized by the ways in which it allows researchers to do things not possible before (i.e., Big data enables the discovery of new information, facts, relationships, indicators, and pointers that could not have been realized previously).
- Data-intensive research involves data resources that are beyond the storage requirements, computational intensiveness, or complexity that is currently typical of the research field.
- Data science is the large-scale capture of data and the transformation of those data into insights and recommendations in support of decisions.

Tools for Transformational Insights



Illustrative Types of Big Data in Education

- *Micro-behavioral data about students' activities in learning ecosystem science*
- *Micro-behavioral data about diagnostic performance assessments formative for learning and instruction*
- *Micro-descriptive data about activities in MOOCs*
- *Macro- and meso-level data about attributes and outcomes for teachers and schools*
- *Macro-behavioral data related to students' dropping out or staying in college*

Tools, Infrastructures, Repositories;

Privacy, Security, Safety;

Models from the Sciences and Engineering



BIG
DATA

→ VOLUME

→ VELOCITY

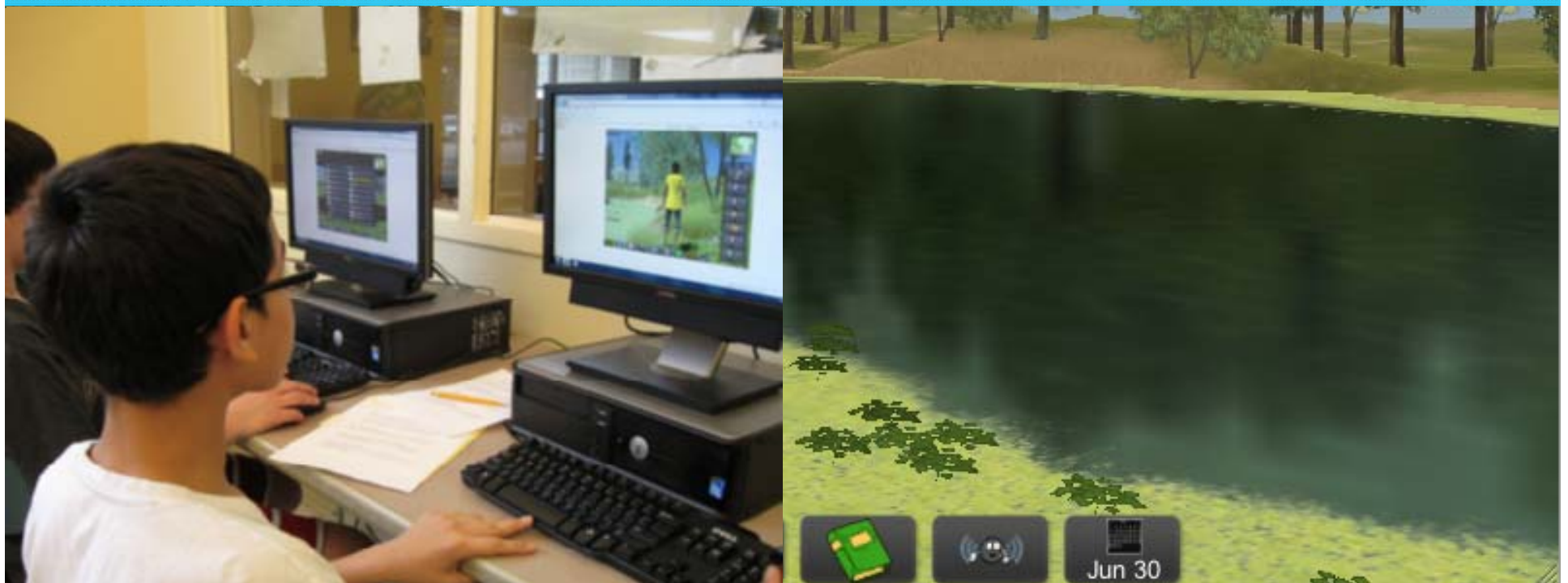
→ VARIETY

→ VERACITY

Settings

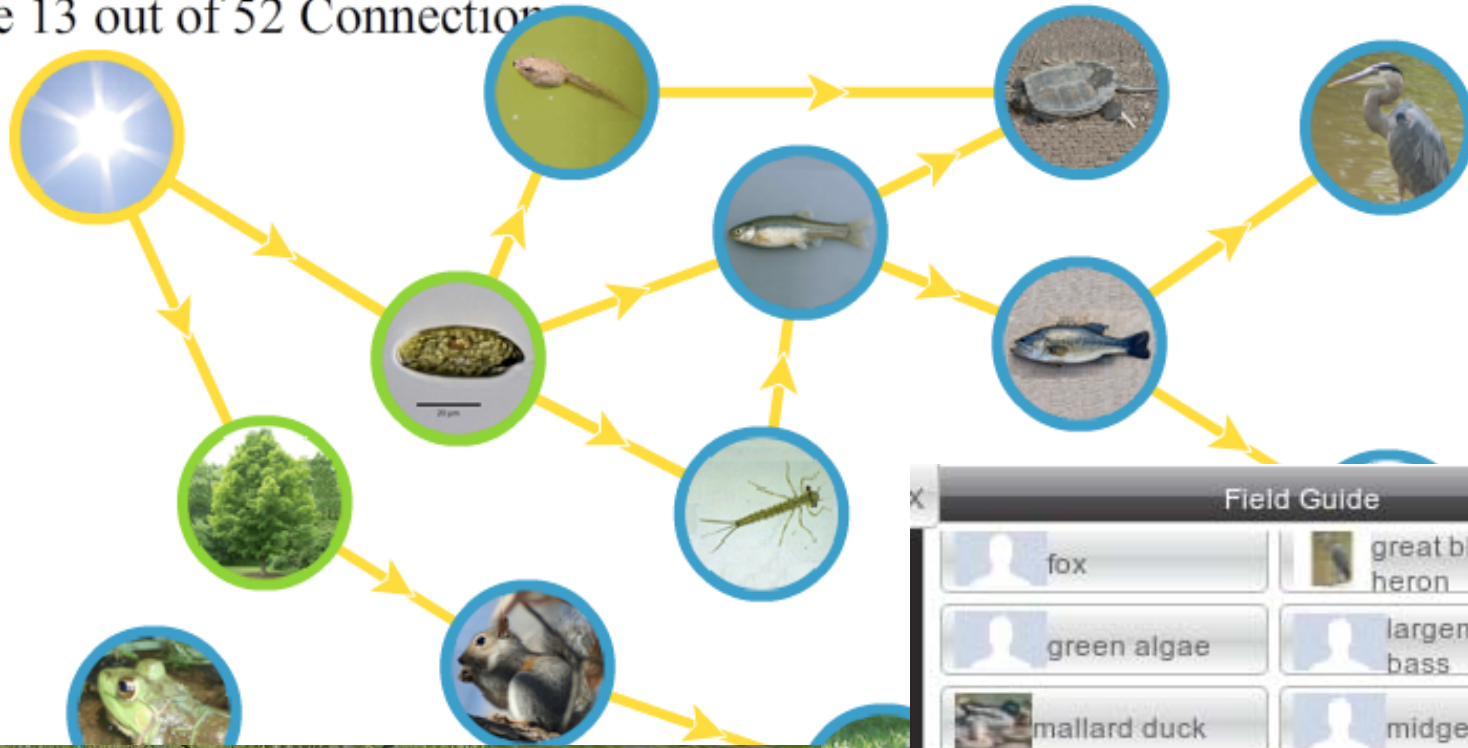


EcoMUVE – Multi-User Virtual Environment



⏏ Drag
↕ Connect
↕ Disconnect
🌿 Check
🖨 Print
? Field Guide

You have 13 out of 52 Connections



snapping turtle

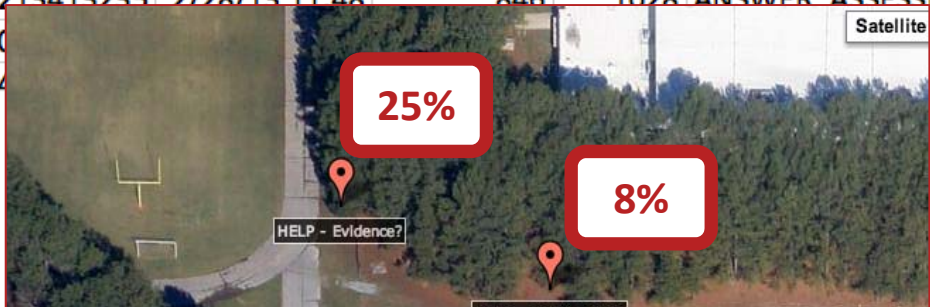
Save Cancel

Field Guide

fox	great blue heron
green algae	largemouth bass
mallard duck	midge larvae
mosquito larvae	predaceous diving beetle
protist	red-tailed hawk
rotifer	snail
snapping turtle	squirrel
sugar maple	tadpole
water chestnut	white pine

Log File Data

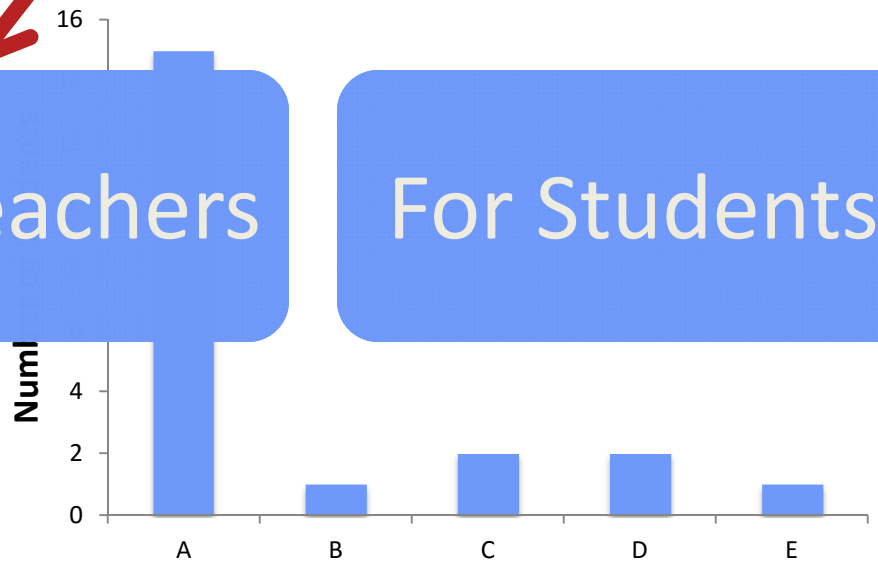
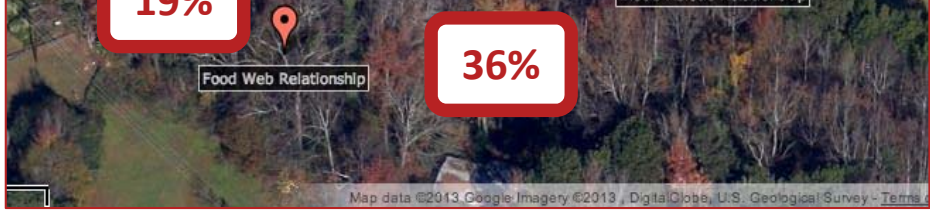
id	timestamp	lat	lon	eventType	latitude	longitude	event	details
8680767965	2/28/13 15:30	515	1028	ANSWER_ASSESSMENT	33.8902313	-84.272389	Before you r	any bugs
3467479015	2/28/13 15:38	464	1028	ANSWER_ASSESSMENT	33.8900381	-84.272309	Before you r	does it rain alot.
1950194462	2/28/13 15:37	855	1028	ANSWER_ASSESSMENT	33.8902205	-84.272443	Before you r	are there any animals
3323505825	2/28/13 11:52	950	1028	ANSWER_ASSESSMENT	33.8899791	-84.271837	Before you r	do animals live there?
9885642951	2/28/13 11:48	812	1028	ANSWER_ASSESSMENT	33.8902634	-84.272266	Before you r	do bugs live there ?
1911444640	2/28/13 15:25	655	1028	ANSWER_ASSESSMENT	33.8902795	-84.272427	Before you r	do they die fast
1541099324	2/28/13 11:38	273	1028	ANSWER_ASSESSMENT	33.8902473	-84.272534	Before you r	how many
1574700448	2/28/13 15:35	112	1028	ANSWER_ASSESSMENT	33.890225	-84.272438	Before you r	how many plants does
4213413235	2/28/13 11:48	846	1028	ANSWER_ASSESSMENT	33.8902713	-84.272373	Before you r	how old is the baby pi
10				MENT	33.8902205	-84.272481	Before you r	i see a lot of dead plan
84				MENT	33.8903868	-84.272711	Before you r	i want to know if their



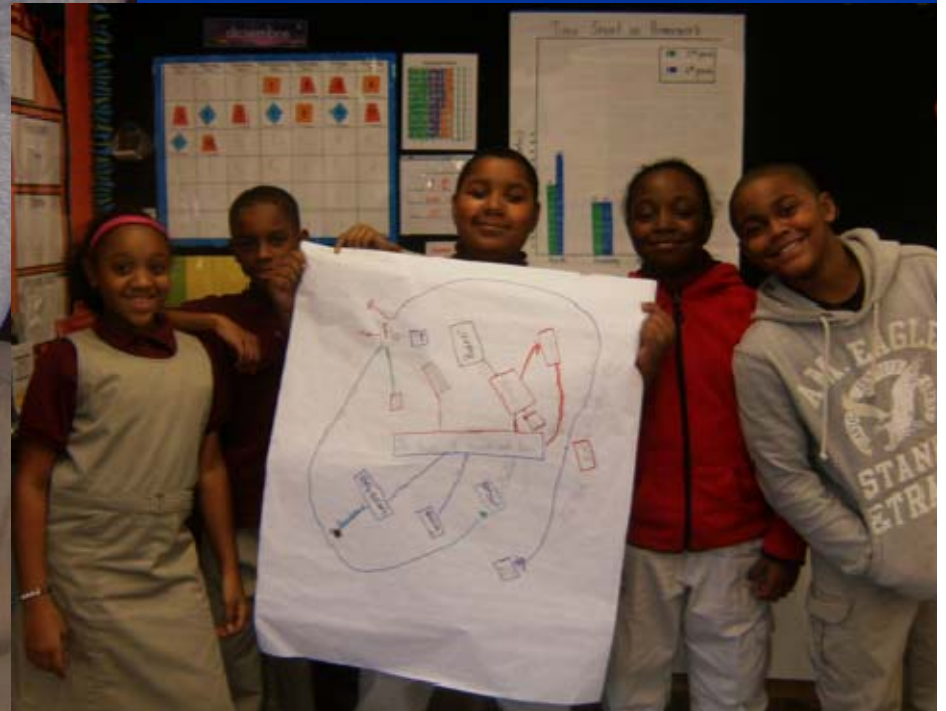
For Researchers

For Teachers

For Students



Collaborative construction of concept maps





(Conner Flynn)

Augmenting Real World Ecosystems

<http://ecomobile.gse.harvard.edu>

GoPro Cameras Capture EcoMOBILE Experience



P2: Now we need to write two things we see that could affect dissolved oxygen.

P1: Plants.

S: Plants.

P2: Plants. And...you guys would rather say...rain, or the dead matter?

S: Mm, dead matter, maybe. 'Cause—

P1: (Why the) dead matter is bacteria?

S: —Yeah. The bacteria. And we don't know, you know, how long...this has been...

P2: You got plants already? Plants, 'cause they release dissolved oxygen into the water.

P1: This could

P2: ... Provide food for bacteria, increasing their population and increasing their need for dissolved oxygen.

P1: The bacteria and —

P2: And um —

P1: And causing an increase in population.

P2: Yeah, increasing their population and their need for dissolved oxygen.

S: [Student talking to other student] Quinn.

P1: Um, provide food for bacteria, increasing population?

P2: Mhm. [Partner 1 continues typing in Evernote]

Evernote: Plants could release dissolved oxygen into the water and dead matter could provide food for bacteria, increasing the bacteria population and their need for dissolved oxygen.



EcoMUVE

- MUVES promote self-efficacy in science
- Simulate experiences otherwise impossible in school settings.
- Explore time and scale
- Opportunities to take on roles, work in teams
- Shared immersive experience that contextualizes learning and supports inquiry

(Ketelhut et al. 2010, Metcalf et al. 2011)

EcoMOBILE

- Greater fidelity and sensory richness, physical interactions with organisms and environments.
- Self-directed collection of real-world data and artifacts.
- Facilitated use of cameras, recording devices, probes, GPS, mapping, graphing, augmented reality.



What Can We Inculcate and Assess?

- Inquiry skills?
- Collaboration?
- Leadership?
- Self-efficacy?
- Metacognition?

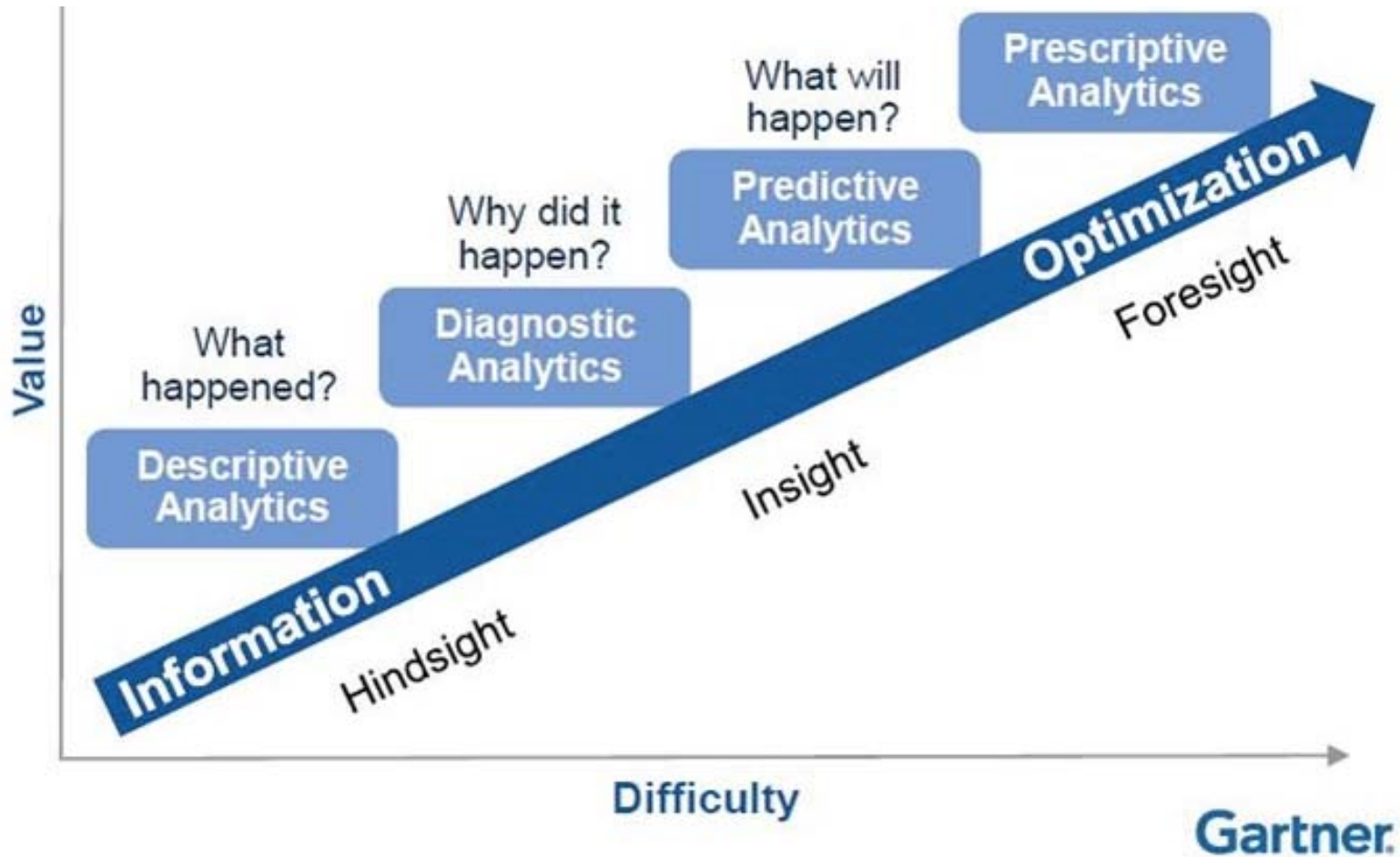
Key Research Questions

- Can we detect problems that students are having *as they are happening*, through automated analysis?
- Can we provide real-time feedback to students and educators in response to the problem detection?
- Is the feedback effective in helping students attain more sophisticated behaviors? Does it *make sense* to the students and educators? Is it *actionable* in that they are able to do something useful with it?

From Description to Prescription

- Determine students' probabilities of failure (*predictions*)
- Determine which students respond to which interventions (*uplift modeling*)
- Determine which interventions are most effective (*explanatory modeling*)
- Allocate resources accordingly (*cost benefit analysis*)

From Hindsight to Foresight



Questions for Field

- To what types of behavioral data could we now apply these methods?
 - *Micro-level* data (e.g., each student's second-by-second behaviors as they learn)
 - *Neso-level* data (e.g., teachers' patterns in instruction; students' patterns in retention)
 - *Macro-level* data (e.g., aggregated student outcomes for accountability purposes) *Gummer's work with EdWise*
- What are the barriers to collecting, storing, sharing and analyzing these data?
- How can we build human and organizational capacity to use evidence-based findings effectively?

3 E's of Immersive Learning

- **Engagement**

Students are motivated to do well, see the relevance of their learning, and increase in self-efficacy

- **Evocation**

Immersive interfaces can evoke a wide spectrum of authentic performances with embedded support

- **Evidence**

Log files, chat logs, shared notebooks, and similar artifacts provide a rich evidentiary trail

Key Next Steps

- Mobilize Communities around Opportunities based on New Forms of Evidence
- Develop New Forms of Educational Assessment
- Develop New Types of Analytic Methods
- Build Human Capacity to Do and to Understand Data Science
- Develop Advances in Privacy, Security, and Ethics
- Infuse Evidence-based Decision-Making throughout Organizations and Systems

NSF Initiatives in Data-Intensive Research

- Christopher Hoadley
- John Cherniavsky
- Anthony Kelly
- Susan Singer
- Finbarr Sloane

Cyberlearning and Future Learning
Technologies and Big Data
Chris Hoadley choadley@nsf.gov

AERA April 2016



Cyberlearning and Future Learning Technologies Description

WHAT IS THE CYBERLEARNING PROGRAM?

Vision of the Cyberlearning Program

- New technologies change what and how people learn
- The best of these will be informed by research on how people learn, how to foster learning, how to assess learning, and how to design environments for learning.
- New technologies give us new opportunities to learn more about learning

Cyberlearning Program Purpose and Goals

The purpose of the Cyberlearning program is to

1. advance design and effective use of the next generation of learning technologies, especially to address pressing learning goals, and
2. increase understanding of how people learn and how to better foster and assess learning, especially in technology-rich environments

A Cross-Directorate Effort

- CISE – Computer and Information Science and Engineering
- EHR – Education and Human Resources
- ENG – Engineering
- SBE – Social, Behavioral, and Economic Sciences

Cyberlearning & Future Learning Technologies project “recipe”

Need

- Pressing societal need or technological opportunity
- *Any domain of learning* (not just STEM)

Innovation

- Design and iteration of new cyberlearning system that could spawn a new genre of learning environments
- Imagining/inventing the future of learning

Learning

- Builds on what we know about how people learn
- Contributes back to the learning sciences

Genre

- Advances design knowledge for a whole category of learning environments
- Research to inform development of the genre

Ways Cyberlearning supports big data research

- Big data as a way to support assessment and feedback to learners (e.g., Aleveln)
- Big data as a way to support research in support of cyberlearning R&D (e.g., Resnick, Ito, Graesser)
- Big data as a tool for learners (e.g. Finzer)



Building Capacity in Data Intensive Education Research

John C Cherniavsky

National Science Foundation

Division of Research on Learning in Formal and Informal Environments

ichernia@nsf.gov

703-292-5136



Education Research Data

- Traditional Data – data bases of local, state, national student and/or school performances
- Interactive Data – data collected from learners interacting with systems – e.g. intelligent tutoring systems, MOOCs
- Sensor Data – e.g. data collected from instrumented learning environments such as video, sound, eye trackers, gps, EEG data, etc.
- Exogenous Data – e.g. data collected for other purposes that can usefully be combined with data collected for education or learning use
- Velocity, Volume, Variety?



Some Problems with Education Research Data

- Restricted access
- Limited standardization
- Scattered data

Resulting in

- Inability to replicate research
- Inability to build on other researcher's results
- Limited trustworthiness of research built upon individual research data



NSF Programs directly addressing some of these Issues with EHR participation

Big Data http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767

Software Infrastructure for Sustained Innovation (SI2)

http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf16532&org=NSF

Data Infrastructure Building Blocks (DIBBS)

<http://www.nsf.gov/pubs/2016/nsf16530/nsf16530.htm>

Building Community and Capacity for Data Intensive Research (BCC)

https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505161

Smart and Connected Communities (S&CC)

<http://www.nsf.gov/pubs/2015/nsf15120/nsf15120.jsp>



Big Data in Education– Some sample possibilities

- Research using large federated traditional data sets (millions of students) and interventions to
 - Effectively use HLM to infer effects on groups or clusters of groups
 - Identify groups that benefit and those that don't from interventions
- Research using interactive data and sensor data collected from learning environments to begin to address affect and/or physiological effects on learning



Big Data in Education– some sample possibilities

- Research using exogenous data such as socioeconomic data on poverty, criminal records, financial records, family records, etc. and combine with other education data to better model factors outside of schools and learning environments that affect learning
- Research addressing questions that arise in using, in particular, large education data sets for research (FERPA, IRBs related to privacy, informed consent, data use and ownership, etc.)
- Developing infrastructure (software and data) for sharing large education data sets



DIBBS addresses data infrastructure

SI2 addresses software infrastructure

Big Data addresses data intensive research issues

BCC addresses both infrastructure and community development for education researchers

S&CC addresses research issues surrounding real community needs mediated by connections – mostly networking



Smart and Connected Communities

- Living Laboratory Model
- Demands demonstration of marked community improvement
- Demands community engagement – e.g. government, industry, technology developers, and end users
- Demands a strong sociotechnical component
- Demands involvement of K-16 education institutions and informal learning institutions (museums, etc.)
- Research should inform and be informed by Complex Systems methodologies



Some Possible S&CC Projects in EHR

- Using location aware software to get data, then analyze the data to develop a more efficient transportation network addressing the needs of all citizens (including both K-12 and university students that will be involved in data collection and analysis).
- Using water analysis software involve citizen scientists – especially K-12 students – in analyzing the community water supply for contamination. Incorporate into the science classroom
- Address adult team learning in development teams that can respond to community emergency response situations



Future STEM Education: The
Potential Value of *Smart and
Connected Communities*
AERA 2016

Anthony E. Kelly
Senior Advisor
Directorate for Education and Human Resources
National Science Foundation
akelly@nsf.gov

“Smart and Connected Communities”

- White House “Smart Cities” initiative September 2015
- NIST Global Cities challenge + NSF DCL (expired)
- USIGNITE- seeking partners
- <https://www.usignite.org/globalcityteams/actioncluster/needs-partner/>
- National Science Foundation Dear Colleague Letter on “smart and connected communities,” September 2015 (expired)

What is a “smart and connected community” problem?

- The problem is complex (and perhaps wicked [3]) and motivates some community (e.g., tribe, region, town, rural group, city, megacity) to work with researchers and other professionals to design, deploy and evaluate an intervention that has potential to ameliorate the identified problem [4].
- An intervention is “smart and connected” when it takes advantage of **emerging nested systems of cyber physical sensors, context-aware computing, Internet of Things, wearable technologies, mobile systems, augmented reality, etc.**
- An intervention is “smart and connected” when it involves **the creative engagement of one or more communities and their distributed human and social capital** (e.g., tribal representatives, city planners, formal and informal education participants, including teachers, students, citizen scientists, or the maker movement).
- A compelling case needs to be made that the intervention is likely to lead to outcomes such as powerful and resilient models and solutions, efficiencies in resources, **advances in science and engineering knowledge and practices**, sociotechnical systems, and STEM education practices and research.

Methodology development for smart and connected research should:

- Account for complex contexts
- Account for complex system interactions at multiple grain sizes
- Design community interventions for resilience
- Support evidence-based claims

Research is necessary on smart and connected communities is necessary for...

Hypothesis generation and testing

Outcomes and metrics

Data management, sharing, and analysis

Enhancing community and capacity

Summary

Consistent with the goals of broadening participation, advancing scientific knowledge and educational practices, and promoting scientific workforce development, NSF seeks ideas on how:

- the wide range of resources of formal and informal education
- research on teaching and learning
- knowledge of curricular design and development
- research on graduate and postdoctoral education
- effective cyberlearning strategies
- workforce development strategies
- research and evaluation innovations
- indicator and assessment innovations
- and related resources . . .

may maximize the many opportunities provided by smart and connected technological and social ecosystems to enable more livable, workable, sustainable, and connected communities.

Some sources that may be valuable in guiding proposal writing

- NSF December meeting: <http://www.bu.edu/systems/nsf-conference-december-3-4-2015/nsf-agenda/>
- NSF Seattle Meeting: <http://cps-vo.org/group/NSF-SmartCities2016/program-agenda>
- EnvisionAmerica: <http://envisionamerica.org/proposed-agenda/>
- White House S&CC: <https://www.whitehouse.gov/the-press-office/2015/09/14/fact-sheet-administration-announces-new-smart-cities-initiative-help>
- NIST Global Cities: http://www.nist.gov/public_affairs/releases/nist-global-city-teams-challenge-aims-to-create-smart-cities.cfm
- NIST and NSF EAGER on Global City Teams Challenge: <http://www.nsf.gov/pubs/2016/nsf16036/nsf16036.jsp>
- CIRCL Ideas Lab: <http://circlcenter.org/events/innovation-lab/>
- European Open Living Labs: <http://openlivinglabs.eu/node/923>
- Living Lab Handbook: <http://www.ltu.se/centres/cdt/Resultat/2.59039/Metoder-och-handbocker/Living-Labs-1.101555?l=en>

References

[1] Dear Colleague Letter

<http://www.nsf.gov/pubs/2015/nsf15120/nsf15120.jsp>

[2] EAGER guidelines:

http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#IID2

[3] Akamani, K., Holzmüller, E. J., & Groninger, J. W. (2016). Managing Wicked Environmental Problems as Complex Social-Ecological Systems: The Promise of Adaptive Governance. In *Landscape Dynamics, Soils and Hydrological Processes in Varied Climates* (pp. 741-762). Springer International Publishing.

[4] Michelucci, P., & Dickinson, J. L. (2016). The power of crowds. *Science*, 351(6268), 32-33.

[5] Bannan, B. (2015).

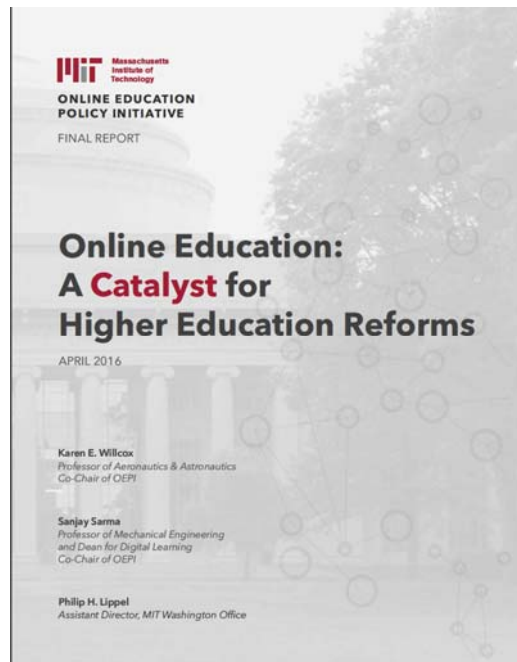
https://www.nitrd.gov/nitrdgroups/index.php?title=SmartCities_CaseExample_Bannan].

[6] Research on privacy:

<https://www.nitrd.gov/cybersecurity/nationalprivacyresearchstrategy.aspx>, and <https://www.nitrd.gov/>).

Improving STEM Education through Data-intensive Research

Susan Rundell Singer
Division Director
Undergraduate Education
National Science Foundation



The Effects of Education and Professional Development on Beginning STEM Teacher Persistence: A Longitudinal Study

Richard Ingersoll U. Penn
(153517500)



MAGAZINE STORY in *The Atlantic*, "Why Do Teachers Quit?"

Analyze data from the newly released nationally representative large-scale, longitudinal survey - The Beginning Teacher Longitudinal Study (BTLS) - conducted by NCES.

- 1) What are the levels of job persistence and job transition among beginning STEM school teachers over their first 5 years after entering teaching;
- 2) What are the types and amounts of preservice education and preparation that beginning STEM school teachers receive and what impact do these have on their job persistence and transitions?
- 3) What are the types and amounts of inservice induction and professional development that beginning STEM school teachers receive in their first 5 years and what impact do these have on their job persistence and transitions?

The project uses Event History Analysis and other advanced statistical methods.

Using data-mining to enable early interventions in introductory engineering courses

UC Riverside - 1432820

- Develop technology-based techniques to directly capture and analyze student learning steps and pathways, and then provide interventions based on that analysis to promote success in engineering courses
- Students will use smartpens and tablet computers to carry out learning activities in undergraduate engineering courses.
- Data mining techniques are used to examine the correlation between these learning activities and academic achievement
- Create an early warning system that identifies students at risk of poor academic performance and recommends suitable learning strategies.



“REBUILD: Changing the Culture of Introductory STEM Instruction at the University of Michigan.” (DUE 1347697)

- Researching Evidence-Based Undergraduate Instructional and Learning Developments (REBUILD)
- The goal of REBUILD is to advance a culture of evidence-based teaching through the work of a team of ten leading faculty members from physics, chemistry, biology, math, and education.
- REBUILD is leveraging the efforts of the existing Learning Analytics Task Force and the Center for Research on Learning and Teaching.

<https://rebuild.lsa.umich.edu/home-3/about/>

Mining MOOCS to Build Instruments

Developing Community & Capacity to Measure Noncognitive Factors in Digital Learning Environments, SRI International (NSF 1338487, Andrew Krumm)

Building a collaborative research community to support the measurement of noncognitive factors associated with learning in science, technology, engineering and mathematics using data from digital learning environments

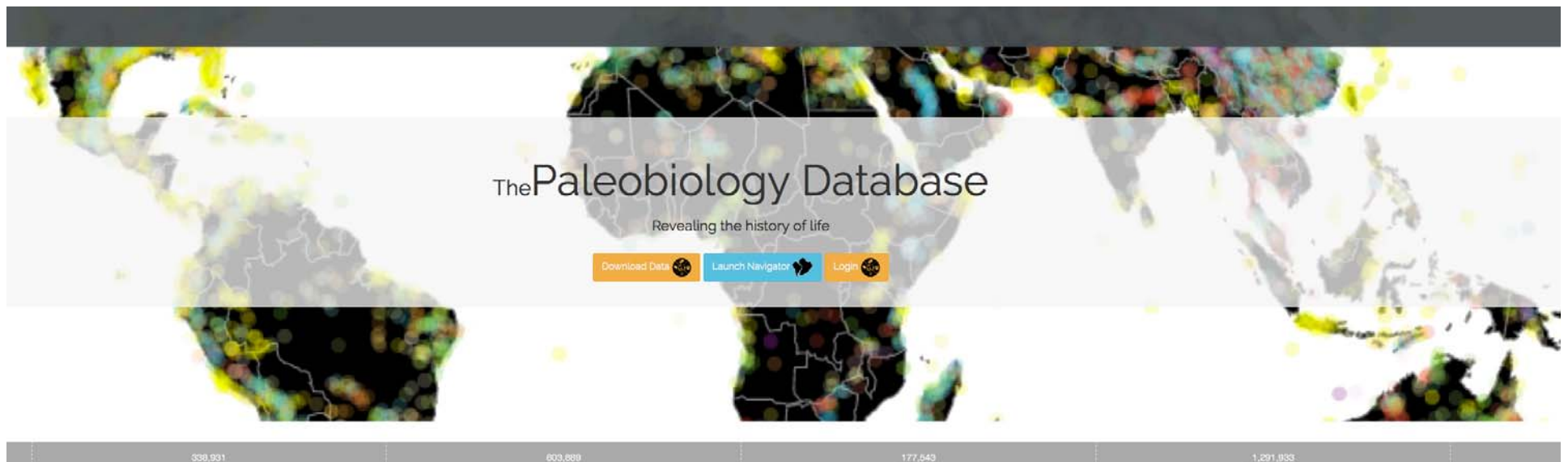
- Competencies related to academic success: Engagement, grit, tenacity, perseverance
- Workshops
- Building on Learning Registry platform
- Framework and shared worked examples that can be used to build common measurement approaches

Leveraging "Big Data" to Explore Big Ideas : Utilizing the Paleobiology Database to Provide Hands-on Research Opportunities for Undergraduates

George Mason University & College of William and Mary

Preparing data scientists within their discipline:


Determine how research experiences using the PBDB compare to field or lab-based research experiences



Ocean Tracks for K-16 Learners

Education Development Center, Inc. (EDC), the Scripps Institution of Oceanography, and Stanford University have been conducting research that has led to the development of a unique Web interface called "Ocean Tracks"

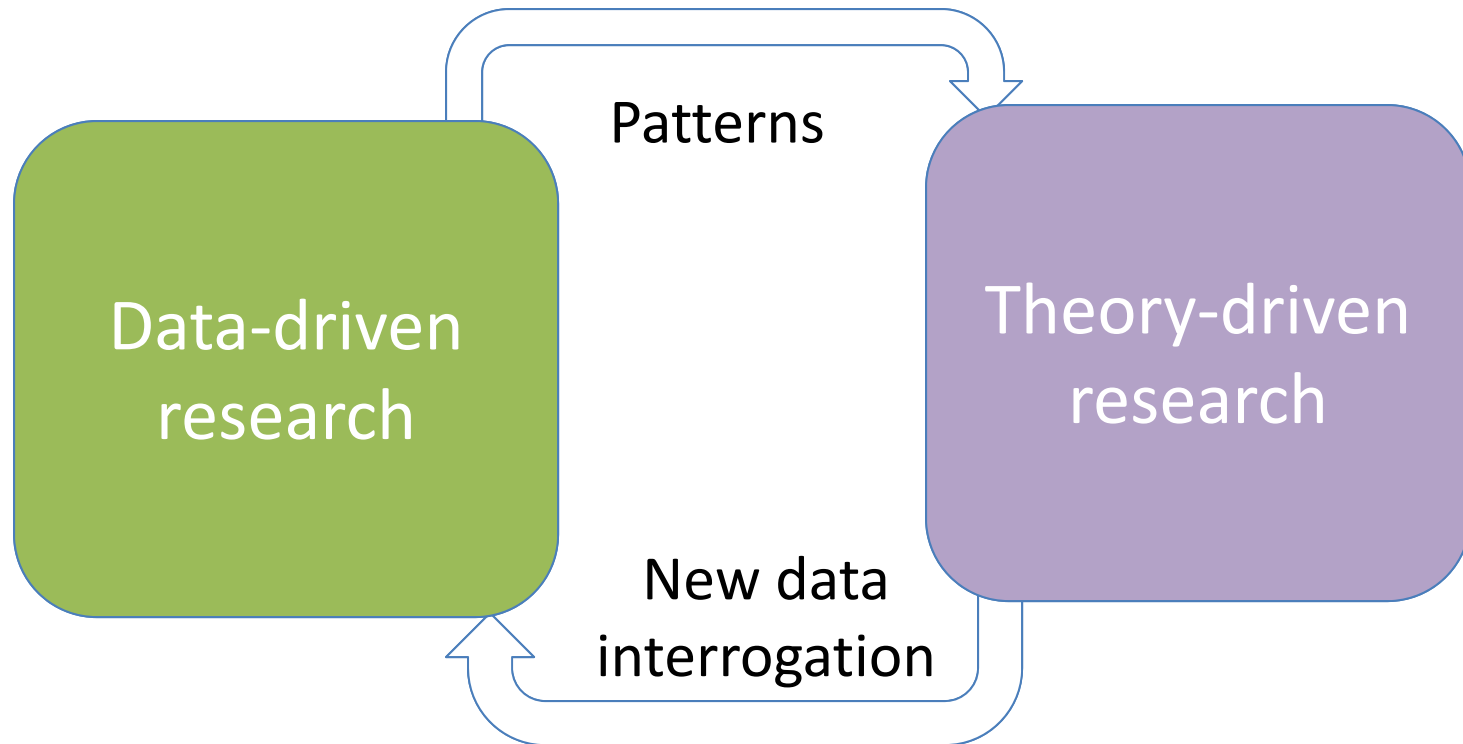
"The Ocean Tracks College Edition" (OT-CE), builds and expands on the prior work to understand how to engage students in scientific inquiry with large-scale datasets.



WELCOME TO THE OCEAN TRACKS DEVELOPMENT SITE!
CONNECTING STUDENTS WITH SCIENTIFIC DATA

[Home](#) | [The Map](#) | [The Library](#) | [Getting Started](#) | [Curriculum](#) | [About](#) | [Contact Us](#)

Data-intensive research in education



Needed: Data-intensive Research Infrastructure

- Interoperable data (standards)
- Community workspace
- Shared tools
- Fixed and flexible workflows
- Growing the next generation researchers



The Future of Educational Data Science

Finbarr Sloane

NSF

Towards a Greater Data Science and its Implications for Education Research

- Our point of departure for *50 years* is the current squabble in the data industry as to whether data science is really the same as traditional statistics.
- His starting point is the simple but elegant depiction of data science as the science of learning from data.

Definition

- Most definitions today focus on skills – the “industrial” – rather than the basic academic or “intellectual” foundations, which are independent of particular technologies and algorithms.

John Tukey: Ever Prescient

Tukey (1962) depicts “data analysis” as the combination of:

- 1. The formal theories of statistics;
- 2. Accelerating developments in computers and display devices;
- 3. The challenge, in many fields, of more and ever larger bodies of data;
- 4. The emphasis on quantification in an ever-widening variety of disciplines.

The Divide

- The divide between mathematical statistics and “data analysis” persisted with Tukey's younger colleagues at Bell labs: John Chambers and William Cleveland.
 - John Chambers
 - William Cleveland

The Divide

- The schism between mathematical statistics and learning from data more prominent than in the seminal 2001 paper *Statistical modeling: The two cultures* by the late UC Berkeley statistician, Leo Breiman.
 - Generative models
 - Predictive models

COMMON TASK FRAMEWORK (CTF)

- Breiman's emphasis on predictive in contrast to generative models led to the “secret” sauce methodology for developing predictive models now called the Common Task Framework.
 - A publically available data set
 - A set of enrolled competitors
 - A scoring referee

Donoho's Vision: A Greater Data Science

- Data Exploration and Preparation;
- Data Representation and Transformation;
- Computing with Data;
- Data Modeling;
- Data Visualization and Presentation;
- Science about Data Science.

A Future Science of Data Science

- The Data Science Foundry for MOOCS
 - Kaylan Veeramamachaneni (MIT)
 - Estimate the time spent on on data cleaning versus data analysis for six COSERA courses?
- EDS at the intersection of Stats and CS: A new science?
 - A need for novel foundational experiences that blend the statistical and the computational;
 - Intellectual traction;

Training in the New Science

- Combinations across knowledge spaces:
 - Learning;
 - A solid knowledge of the content area the learning model is being applied to;
 - CS (programming, algorithms, machine learning??)
 - Statistical Modeling (???)
 - New Model Development
- What is the appropriate mix?
- What might be a Grand Challenge for this new science?