Data and Society
Data science – Lecture 13

3/18/21

Today (3/18/21)

- IDEA would like to announce the RPI StudySafe app. Students working with the Rensselaer IDEA's Data INCITE program during the Summer and Fall 2020 terms created an app to help students easily identify crowded and open spots around campus. Would you please help us to share this information with faculty, staff and students within your school? https://idea.rpi.edu/media/rensselaer-students-create-studysafe-app-safe-distancing-campus.
- For questions related to StudySafe, please contact: Dr. Kristin P. Bennett, bennek@rpi.edu or Dr. John S. Erickson, erickj4@rpi.edu.

- **Personal Essay #2 instructions**
- Lecture / Discussion – Data Science
- Presentations

Personal Essay 2: Facial Recognition

- 500-600 words / 11 point font / 10 points
- Send .docx to bermaf@rpi.edu before/by **Sunday, April 1, 11:59 p.m..**
- **Read:** https://www.nytimes.com/interactive/2021/03/18/magazine/facial-recognition-clearview-ai.html?referringSource=articleShare

- TOPIC: **Should the use of facial recognition be limited? What do you think the rights of subjects included in facial recognition databases should be?**

Facial recognition is a key topic of our time. Arguments abound about whether it is an effective tool that protects us or a tool that intrudes on our privacy and rights. What do you think? What is the best use of facial recognition and how should we promote that through policy, legislation and practice?

Writing Personal Narratives / Storytelling

- **GOAL:** Tell your audience (Fran/general public) an interesting (true) story on the assigned topic
- **PURPOSE:** Personal essays explore a **specific experience** and tell the story from **your point of view**. They may illustrate how a personal conflict, event, or experience left a lasting impression or how it changed your views or perspective.
- **TONE:** Can be more conversational than formal writing but should **establish you as an articulate and credible individual**.
- **FORMAT:**
  - **Introduction** -- Grab the reader and summarize your points
  - **Body** – main text that tells the story / provides information / explains and supports your points
  - **Conclusion** – may include a lesson, message, moral, take-away

Tips and Grading Rubric for Personal Narratives

TIPS

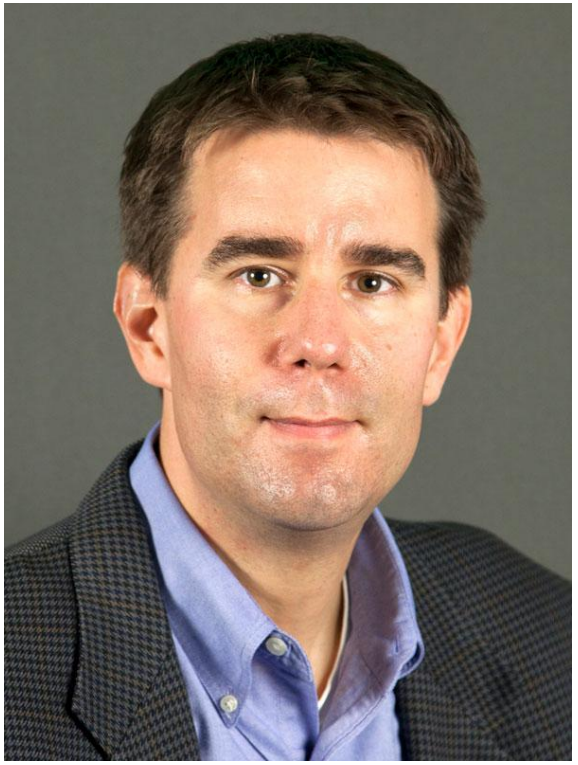
- Create an **outline** of the piece (don't turn this in) before you write with the main points.
- Do **more than one draft** before turning your piece in.
- **Spell and grammar check** your piece
- Relevant statistics or facts should be cited and included as endnotes.
- Resources for writing personal essays:
  - https://www.thoughtco.com/write-the-perfect-personal-essay-3858745
  - https://www.indeed.com/career-advice/career-development/how-to-write-a-personal-essay

ciety, C

- **GRADING RUBRICK** (10 points total)
  - **5 points – content**
    - Is the story compelling?
    - Does the content comply to the personal essay format?
  - **5 points – writing**
    - Is there a clear tone and narrative?
    - Is it well-written (English, grammar, spelling, flow)?

"Watching" for March 22

- Guest Speaker: **Brett Bobley,** CIO, National Endowment for the Humanities, Director of the Office of Digital Humanities



- **Watch this video** of hypertext and the early Digital Humanities! https://triproftri.wordpress.com/2016/05/09/claims-of-early-dh-dp/

Date	Topic	Speaker	Date	Topic	Speaker
1-25	Introduction	Fran	1-28	The Data-driven World	Fran
2-1	Data and COVID-19	Fran	2-4	Data and Privacy -- Intro	Fran
2-8	Data and Privacy – Differential Privacy	Fran	2-11	Data and Privacy – Anonymity / Briefing Instructions	Fran
2-15	NO CLASS / PRESIDENT'S DAY		2-18	NO CLASS	
2-22	Legal Protections	Ben Wizner	2-25	Data and Discrimination 1	Fran
3-1	Data and Discrimination 2	Fran	3-4	Data and Elections 1	Fran
3-8	Data and Elections 2	Fran	3-11	NO CLASS / WRITING DAY	
3-15	Data and Astronomy (Op-Ed due)	Alyssa Goodman	3-18	Data Science	Fran
3-22	Digital Humanities	Brett Bobley	3-25	Data Stewardship and Preservation	Fran
3-29	Data and the IoT	Fran	4-1	Data and Smart Farms	Rich Wolski
4-5	Data and Self-Driving Cars	Fran	4-8	Data and Ethics 1	Fran
4-12	Data and Ethics 2	Fran	4-15	Cybersecurity	Bruce Schneier
4-19	Data and Dating	Fran	4-22	Data and Social Media	Fran
4-26	Tech in the News	Fran	4-29	Wrap-up / Discussion	Fran
5-3	NO CLASS				

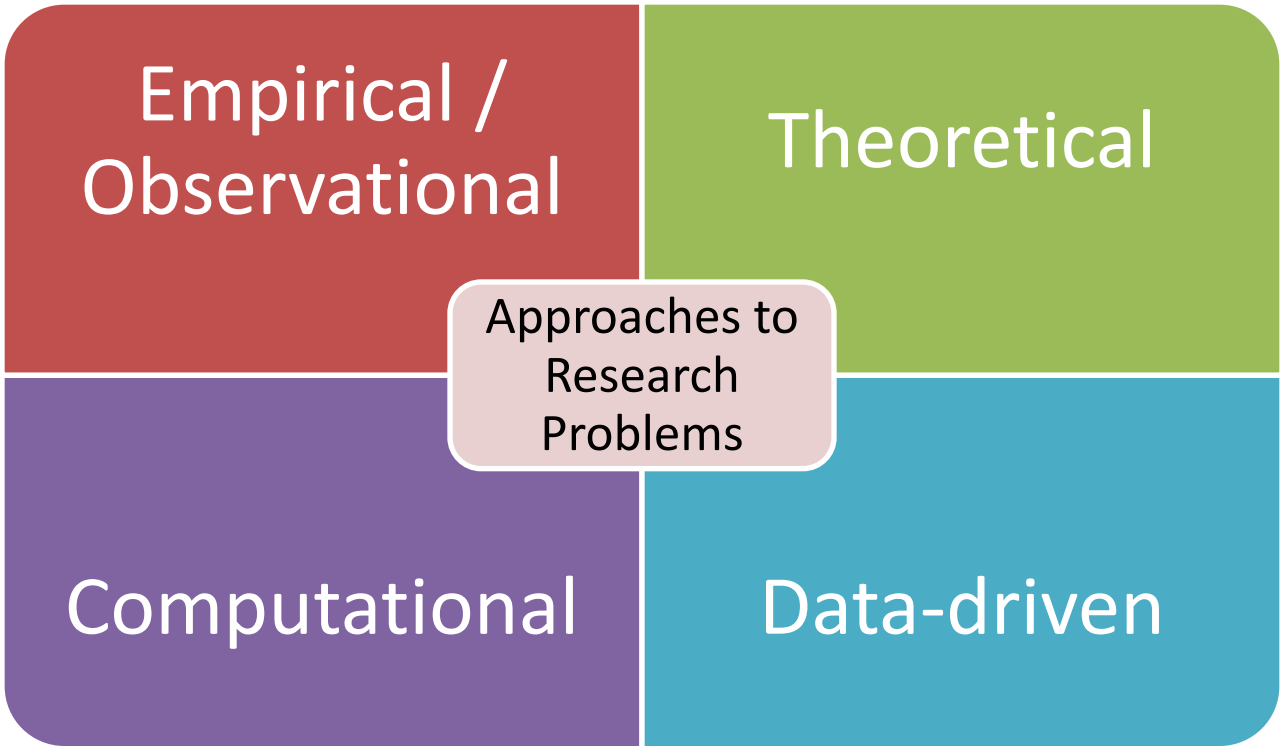
Lecture – 3 Faces of Data science

- Data science
- Big data
- Data-driven science

What is data science?

- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.
- **Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.** It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science. [Wikipedia, Data Science]

Data science as a 4th paradigm for scientific research [Jim Gray]



Data Science as a discipline vs. Data-driven Science

- **Data Science**
  questions advance the ways in which data is organized, analyzed, visualized, etc.

- Publications often found in computer science venues

- **Data-driven science**
  questions focus on the use of data to drive solutions to domain questions.

- Publications often found in domain science venues

Developing Data Science as a discipline

- **Research**
  - How to advance data science? How to advance data-driven science? What are the programmatic needs?
- **Curriculum & pedagogy**
  - Who teaches data science? What classes? Where in the university? To whom?
- **Infrastructure**
  - What is needed to support data science?
- **Applications**
  - How is data science used to provide services?
- **Social and ethical impacts**
  - What are the consequences of data science?

Data science research

Important research problems in optimizing the potential of the data analysis pipeline

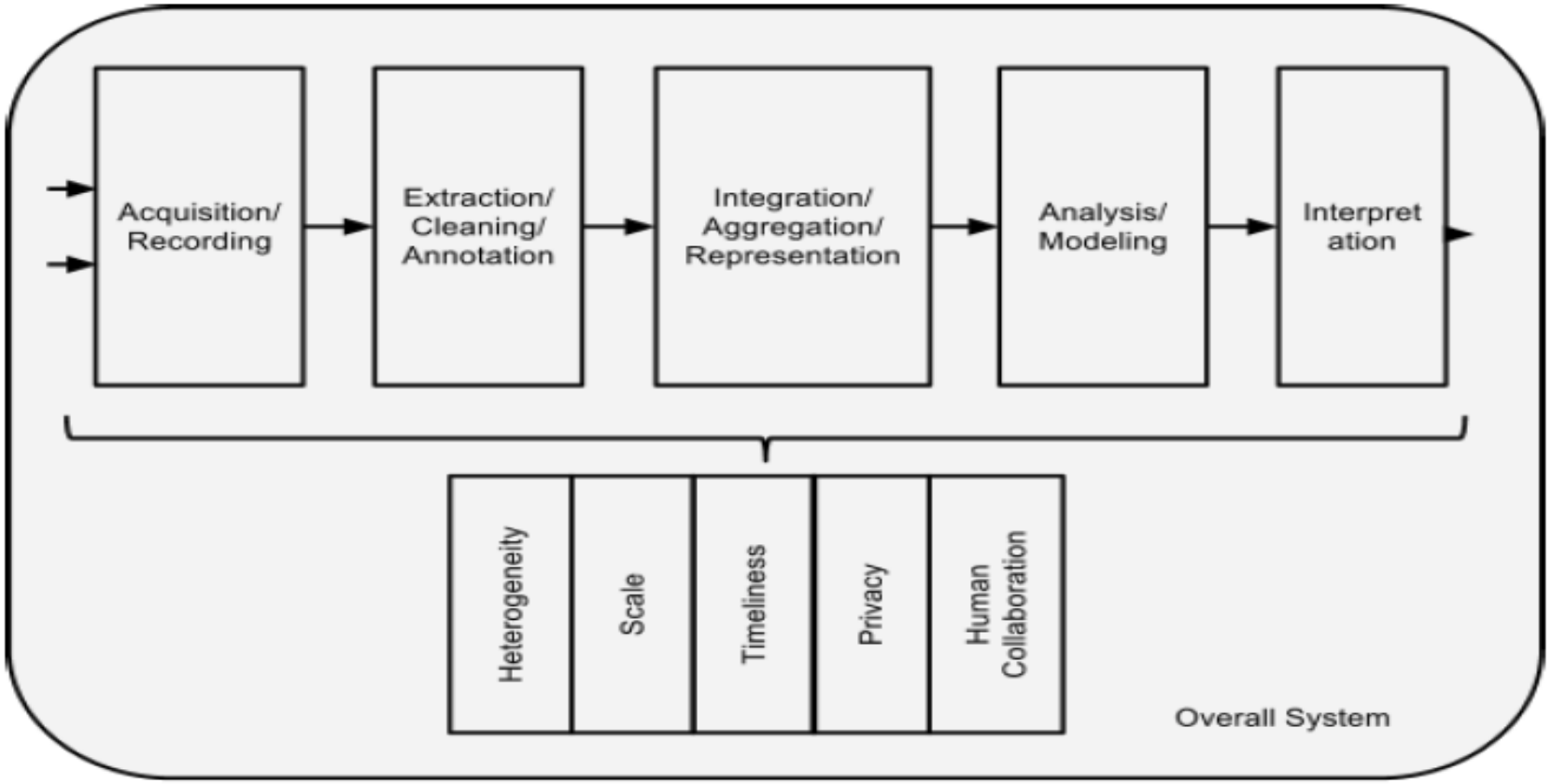
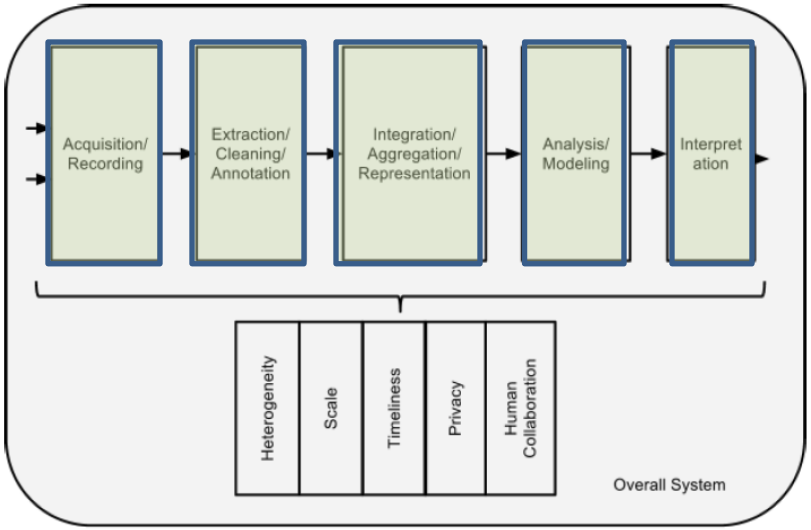


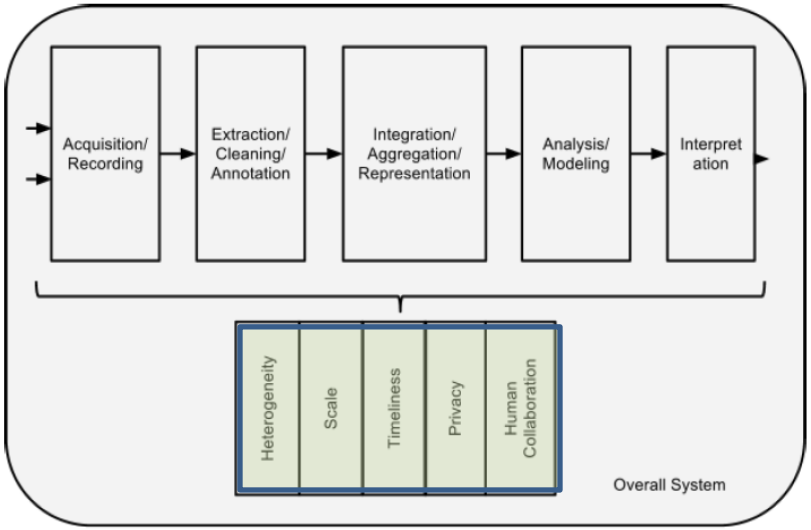
Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Many areas for investigation throughout the data pipeline



- **Data acquisition and recording challenges:**
  - Collection, data integrity, curation, metadata
- **Information extraction and cleaning challenges:**
  - How to extract and structure, how to clean, how to deal with error
- **Data integration, aggregation and representation challenges:**
  - Machine and human understandable, interoperability
- **Query processing, data modeling and analysis challenges:**
  - Meaningful mining, modeling, analysis. statistics
  - Scaling, error estimation, prediction accuracy
- **Interpretation challenges:**
  - Visualization, reproducibility, provenance, etc.

Many areas for investigation throughout the data pipeline – cross-cutting issues



- **Heterogeneity challenges:**
  - Interoperability, harmonization of error, effective representation, flexible DB design
- **Scale challenges:**
  - Extensible storage and cloud technologies; I/O and query performance,
  - Data security, replication
- **Timeliness challenges:**
  - Real-time analysis , common vs. custom searches
- **Privacy/ policy challenges:**
  - Support for privacy, differential privacy and sharing mechanisms, policy
  - Clarification of rights, ownership, access
- **Human collaboration challenges:**
  - Harmonization of input from multiple human experts and shared exploration of results
  - Crowd-sourcing approaches that facilitate correction of errors.

What is big data?

- *Wikipedia*: "Broad term for data sets so large or complex that traditional data processing applications are inadequate."
- *McKinsey*: "Datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze"
- *O'Reilly Radar*: "Data that exceeds the processing capacity of conventional database systems. The data that is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it."

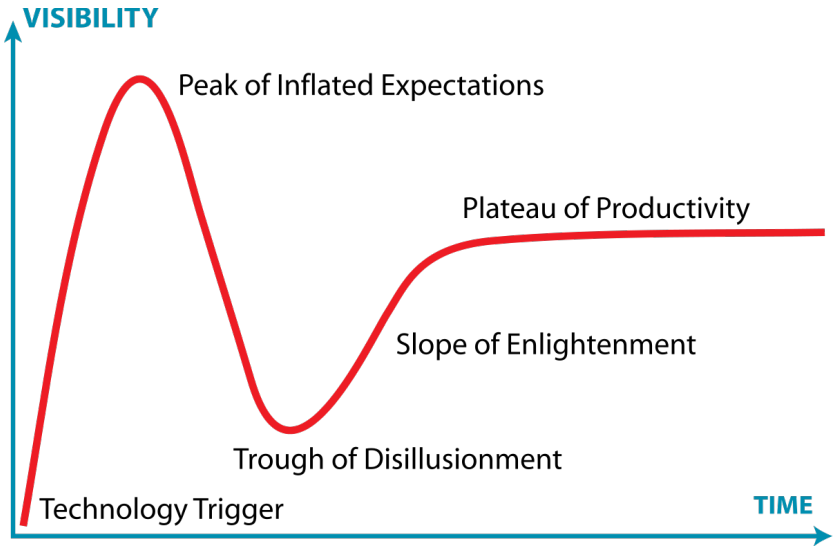
What does big data tell us?

- **Big data is often noisy, dynamic, heterogeneous. Inter-related and untrustworthy**.
- Why do we find it useful?
  - General statistics obtained from frequent patterns and correlation analysis can disclose more **reliable hidden patterns and knowledge**
  - Interconnected big data forms large heterogeneous information networks, with which **information redundancy can be explored** to compensate for missing data, cross check conflicting cases, validate trustworthy relationships, disclose inherent clusters, and uncover hidden relationships and models.

Not a magic bullet: big data has limits

- *(From "Eight (No, Nine!) Problems with Big Data", NY Times).*
  Limitations of big data:
  - "… although big data is very good at detecting correlations, …, it **never tells us which correlations are meaningful**"
  - " … big data can work well as an adjunct to scientific inquiry but **rarely succeeds as a wholesale replacement**."
  - " … many tools that are based on big data **can be easily gamed**."
  - " … even when the results of a big data analysis aren't intentionally gamed, they often turn out to be **less robust than they initially seem**."
  - " … whenever the source of information for a big data analysis is itself a product of big data, opportunities for vicious cycles abound [**echo chamber effect**]."
  - " … risk of **too many correlations**."
  - " … big data is prone to giving **scientific-sounding solutions to hopelessly imprecise questions**."
  - " …big data is at its best when analyzing things that are extremely common, but often **falls short when analyzing things that are less common**."
  - " … **the hype**."

Gartner's 2020 predictions on Data Science and Machine Learning

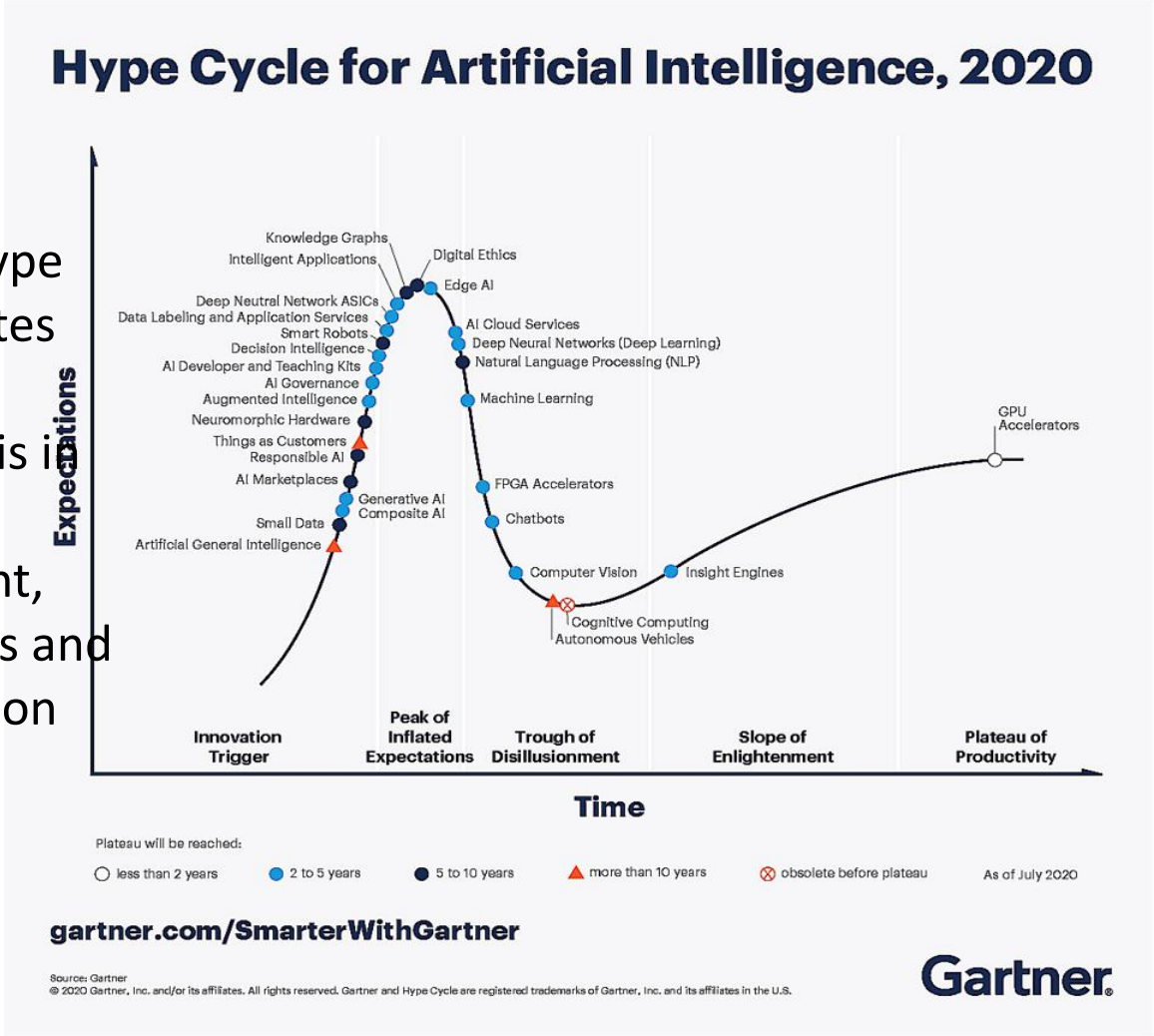


- **On the Rise**
  - Quantum ML
  - Self-Supervised Learning
  - Generative Adversarial Networks
  - Differential Privacy
  - Federated Machine Learning
  - Adaptive ML
  - Kubeflow
  - Reinforcement Learning
  - Transfer Learning
  - Synthetic Data
- **At the Peak**
  - Decision Intelligence
  - Large-Scale Pretrained Language Model
  - AI-Related C&SI Services
  - Data Labeling and Annotation Services
  - Explainable AI
  - MLOps
  - Augmented DSML
  - AutoML
  - Citizen Data Science
  - Deep Neural Networks (Deep Learning)
  - Prescriptive Analytics
- **Sliding Into the Trough**
  - Graph Analytics
  - Advanced Video/Image Analytics
  - Event Stream Processing
- **Climbing the Slope**
  - Predictive Analytics
  - Text Analytics
- **Entering the Plateau**
  - Apache Spark
  - Notebooks

https://www.gartner.com/en/documents/3988118

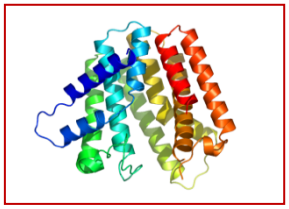
Gartner hype cycle: Data Science and Machine Learning

- Gartner's hype cycle indicates where a technology is in terms of development, expectations and productization



https://www.forbes.com/sites/louiscolumbus/2020/10/04/whats-new-in-gartners-hype-cycle-for-ai-2020/?sh=2acea770335c, 70/6370

Data-Driven Research (other technologies involved too)



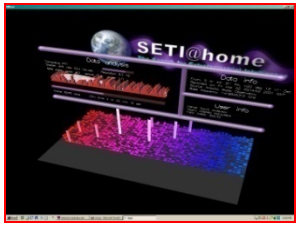
Protein analysis and modeling of function and structures

Storage and Analysis of Data from the CERN Large Hadron Collider

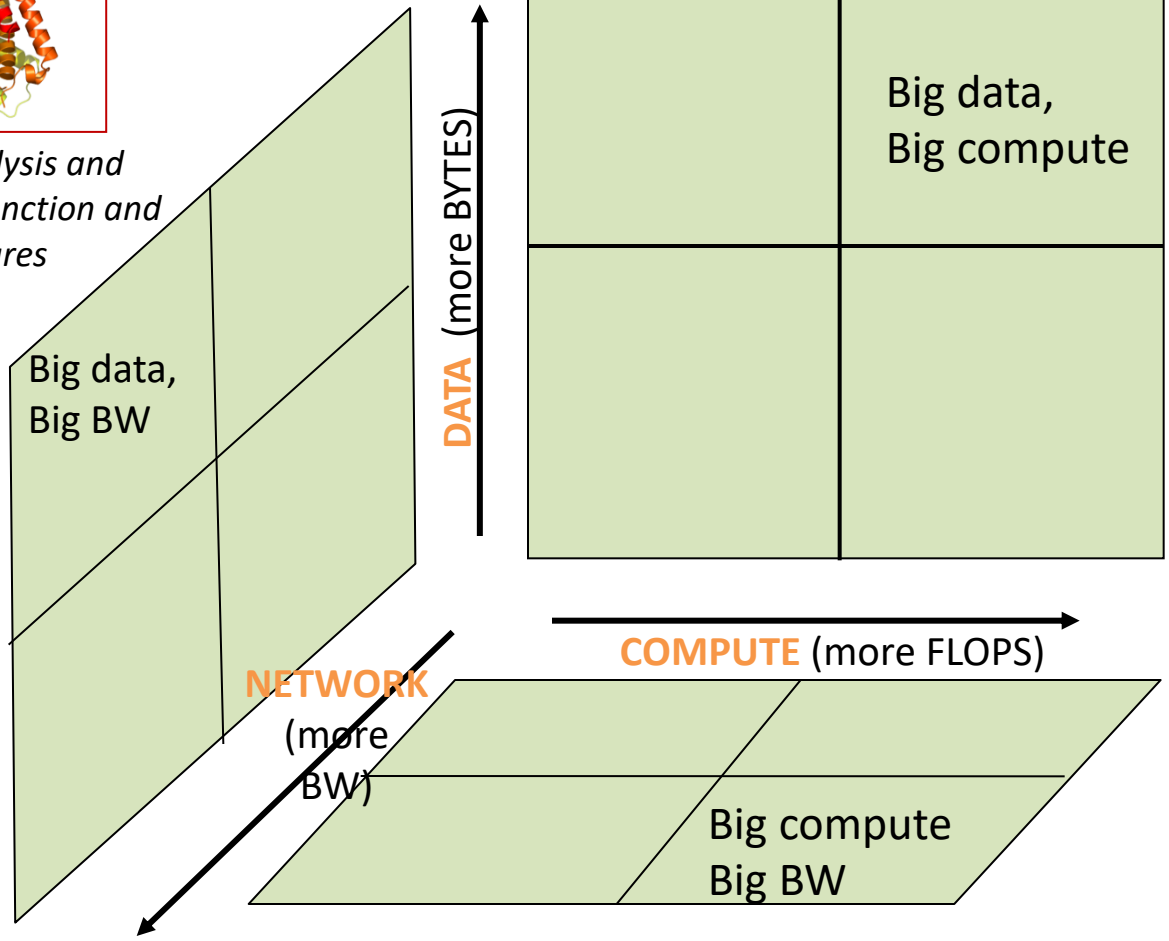
Real-time disaster response

Cosmology

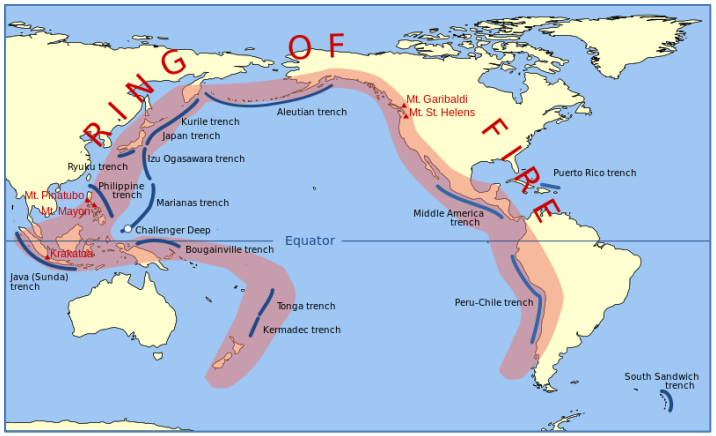
Development of biofuels



Seti@home, MilkyWay@Home, BOINC



Earthquake Simulation



Background:

- Earth constantly evolving through the movement of “plates”
- Using plate tectonics, the Earth's outer shell (lithosphere) is posited to consist of seven large and many smaller moving plates.
- As the plates move, their boundaries collide, spread apart or slide past one another, resulting in geological processes such as earthquakes and tsunamis, volcanoes and the development of mountains, typically at plate boundaries.

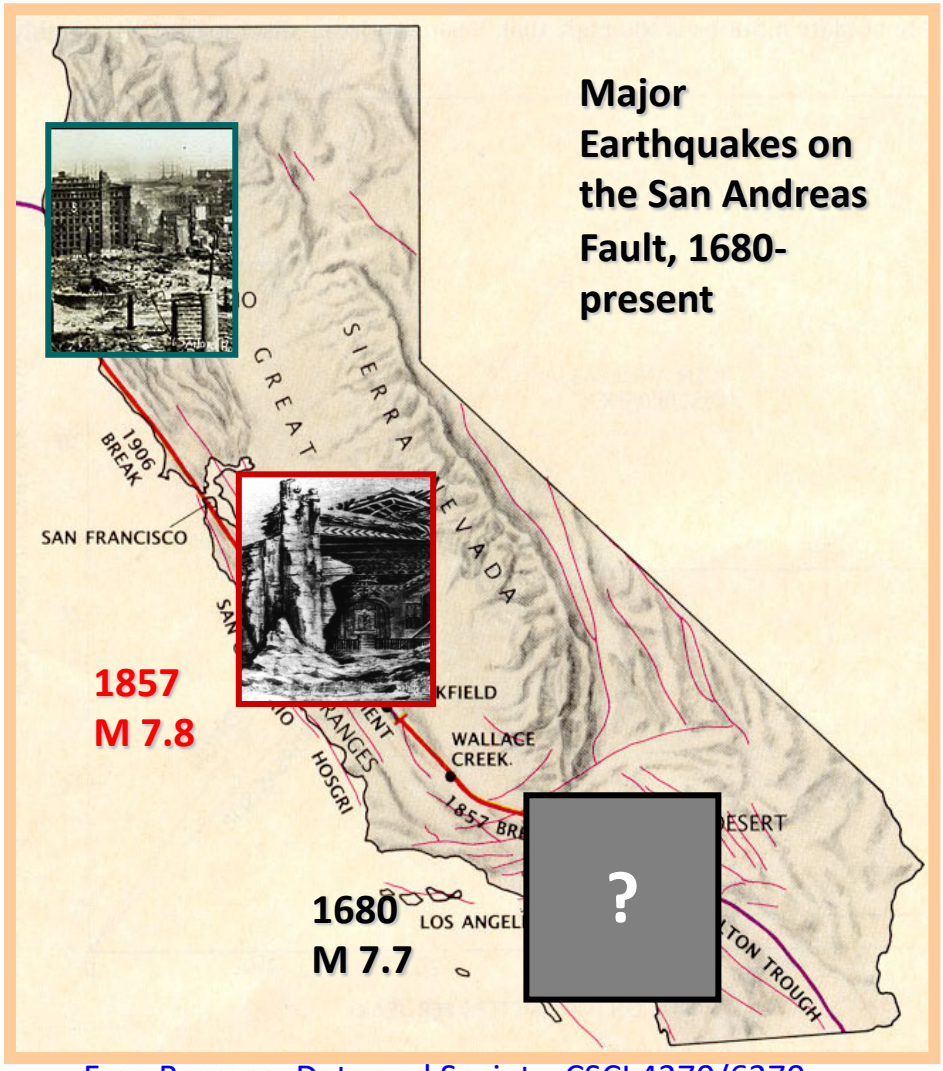
Why Earthquake Simulations are Important

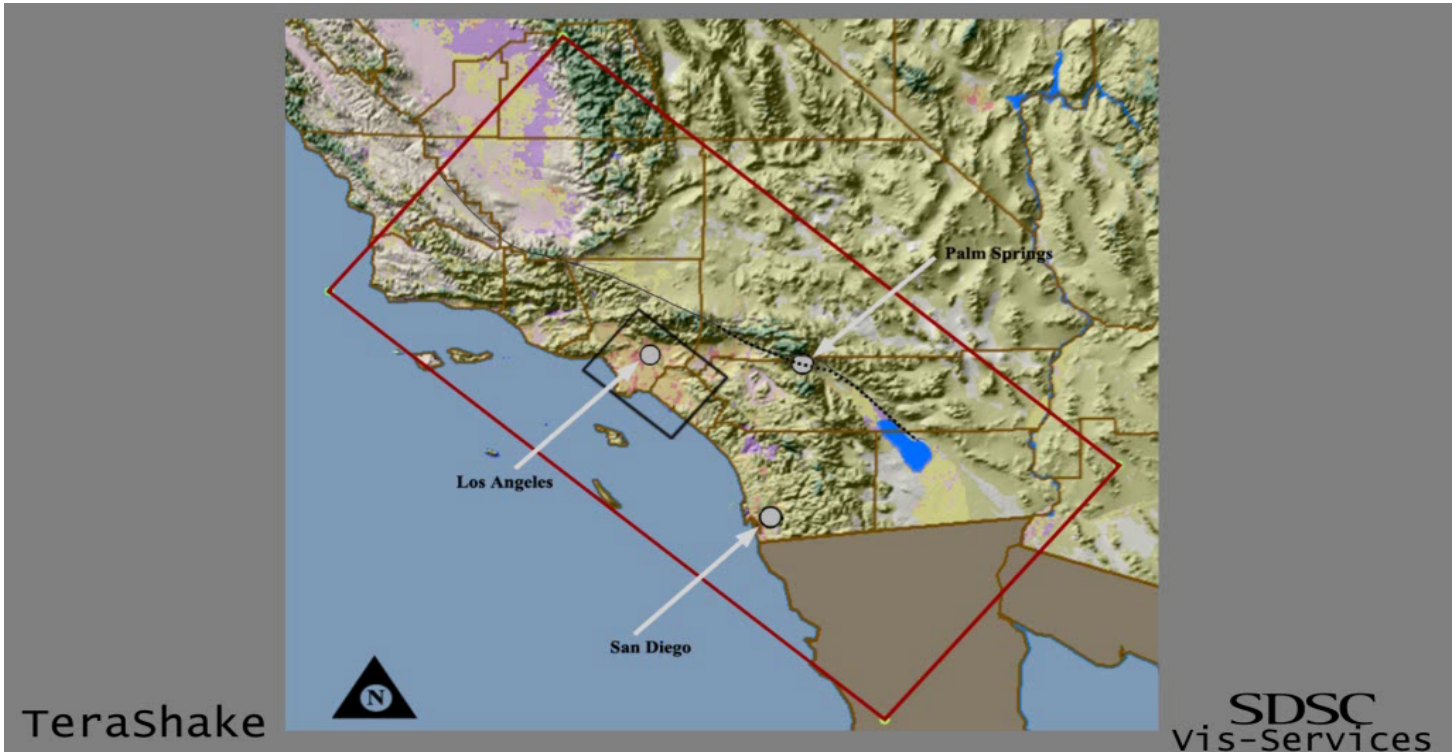
- **If we understand how earthquakes can happen, we can**
  - Predict which places might be hardest hit
  - Reinforce bridges and buildings to increase safety
  - Prepare police, fire fighters and doctors in high-risk areas to increase their effectiveness
- Information technologies drive more accurate earthquake simulation

Tsunamis come from earthquakes in the ocean

Terrestrial earthquakes damage homes, buildings, bridges, highways

What would be the impact of an earthquake on the lower San Andreas fault?





Simulation decomposition strategy leverages parallel high performance computers

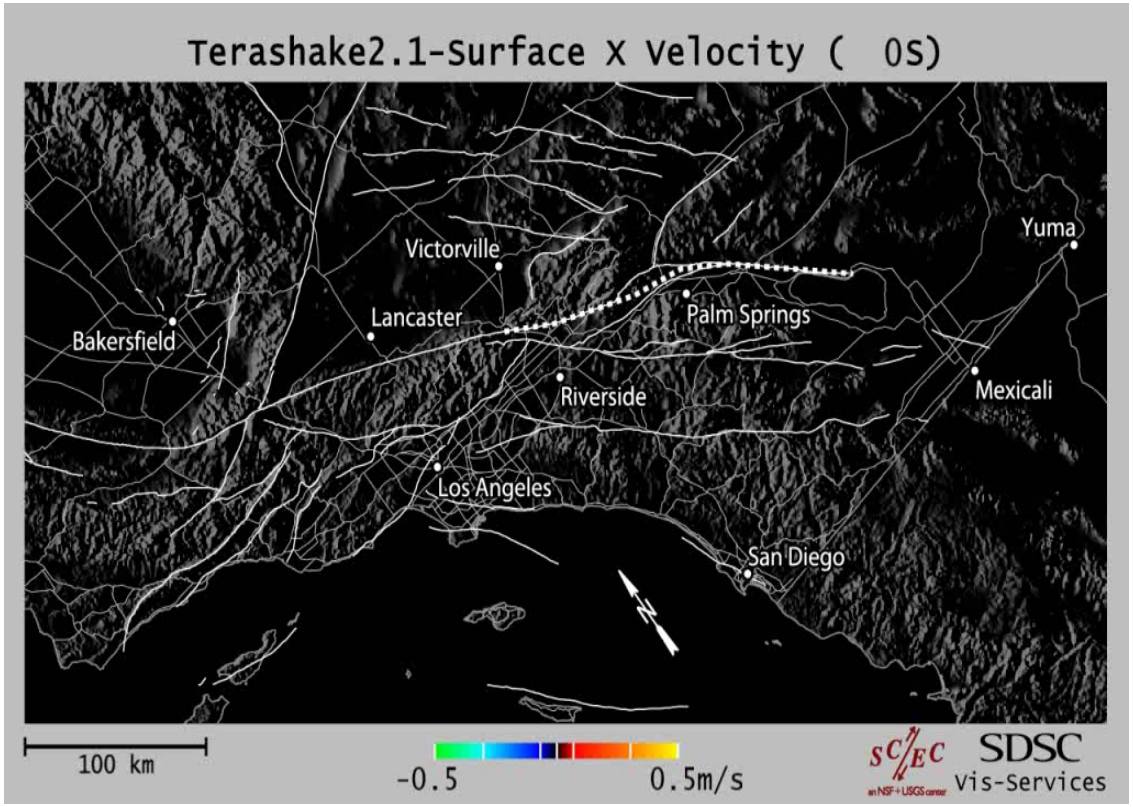
- *Southern California partitioned into "cubes" then mapped onto processors of high performance computer*
- *Data choreography used to move data in and out of memory during processing*

Builds on data and models from the Southern California Earthquake Center, Kinematic source (from Denali) focuses on Cajon Creek to Bombay Beach

TeraShake Simulation (2000's)

Simulation of Southern of 7.7 earthquake on lower San Andreas Fault

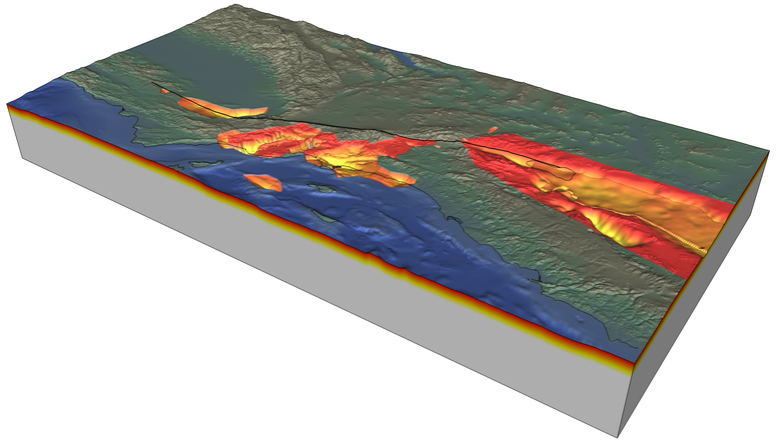
- Physics-based dynamic source model – Anelastic Wave Propagation Model code -- simulation of mesh of 1.8 billion cubes with spatial resolution of 200 m
- Simulated first 3 minutes of a magnitude 7.7 earthquake, 22,728 time steps of 0.011 second each
- Simulation for TeraShake 1 and 2 simulations generated 45+ TB data



Under the surface

Application Evolution

- TeraShake → PetaSHA, PetaShake, CyberShake, M8, etc. at SCEC and SDSC
- Evolving applications improving
  - Resolution
  - Models and algorithms
  - Simulation accuracy
  - Features
  - Workflow and simulation efficiency and performance, etc.



- **M8:** Regional scale wave propagation simulation using realistic 3D earth model, dynamic earthquake source, simulating ground motions at frequencies of interest to building engineers

Image: https://scec.usc.edu/scecpedia/M8

More compute and data → more science
SCEC 2017-2021 Science Priorities

- The strategic framework for the SCEC5 Science Plan is cast in the form of five basic questions of earthquake science:

    1. **How are faults loaded on different temporal and spatial scales?**
    2. **What is the role of off-fault inelastic deformation on strain accumulation, dynamic rupture, and radiated seismic energy?**
    3. **How do the evolving structure, composition and physical properties of fault zones and surrounding rock affect shear resistance to seismic and aseismic slip?**
    4. **How do strong ground motions depend on the complexities and nonlinearities of dynamic earthquake systems?**
    5. **In what ways can system-specific studies enhance the general understanding of earthquake predictability?**

- These questions cover the key issues driving earthquake research in California, and they provide a basis for gauging the intellectual merit of proposed SCEC5 research activities.
- (From https://www.scec.org/research/scec3_objectives.html)

Lecture Materials (not already on slides)

- "**Challenges and Opportunities with Big Data**" a community white paper developed by leading researchers across the U.S., http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf
- **NSF CISE Data Science Report**, https://www.nsf.gov/cise/ac-data-science-report/
- "

Presentations

Upcoming Presentations

March 22

- "**New open source database tracks data on slaves, slavers, and allies**", Harvard Gazette, https://news.harvard.edu/gazette/story/2021/03/new-open-source-database-tracks-data-on-slaves-slavers-allies/ (Eric X.)
- "**Robo-writers: the rise and risks of language-generating AI**", Nature, https://www.nature.com/articles/d41586-021-00530-0 (Josh M.)

March 25

- "**How Scientists scrambled to stop Donald Trump's EPA from wiping out climate data**", The Verge, https://www.theverge.com/22313763/scientists-climate-change-data-rescue-donald-trump
- "**More than 100 scientific journals have disappeared from the Internet**", Nature, https://www.nature.com/articles/d41586-020-02610-z

Need Volunteers – Presentations for March 29

- "**Animal Planet**", New York Times, https://www.nytimes.com/interactive/2021/01/12/magazine/animal-tracking-icarus.html?referringSource=articleShare (Julian C.)

- "**Ring and Nest helped normalize American surveillance and turned us into a nation of voyeurs**", Washington Post, (Hannah L.) https://www.washingtonpost.com/technology/2020/02/18/ring-nest-surveillance-doorbell-camera/?utm_campaign=wp_post_most&utm_medium=email&utm_source=newsletter&wpisrc=nl_most

Presentations for March 18

- "**Why so many data science projects fail to deliver**", MIT Sloan Management Review, (Sola) https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/?og=Home+Editors+Picks

- "**Using big data to measure environmental inclusivity in cities**," EOS, https://eos.org/articles/using-big-data-to-measure-environmental-inclusivity-in-cities (Nate)