Hearing #7 on Competition and Consumer Protection in the 21st Century

Howard University

School of Law

November 13, 2018

Welcome

We Will Be Starting Shortly

Welcome and Introductory Remarks

Andrew I. Gavil
Howard University School of Law

Opening Address

Michael Kearns

University of Pennsylvania

Introduction to Algorithms, Artificial Intelligence, and Predictive Analytics

John P. Dickerson
Assistant Professor of Computer Science
University of Maryland, College Park

The path to "AI"

… although machines can perform certain things as well or perhaps better than any of us can, they infallibly fall short in others …

… by which means we may deduce that they did not act from knowledge, but only from the disposition of their organs.

[Descartes 1600s]

The path to "AI"

Reasoning is nothing but reckoning.

[Hobbes 1600s]

- 1900s: Breakthroughs in the formalization of mathematical reasoning
  - Some hard limits on what can be done
  - Subject to those limits, a Turing machine can do it!

The path to “AI”

If intelligence can be simulated by mathematical reasoning …

… and mathematical reasoning can be simulated by a machine …

… then can a machine simulate intelligence?

The path to "AI"

- Artificial Intelligence coined by John McCarthy ('55/'56)
- Dartmouth Summer Research Project on Artificial Intelligence aka "Dartmouth Conference" ('56)

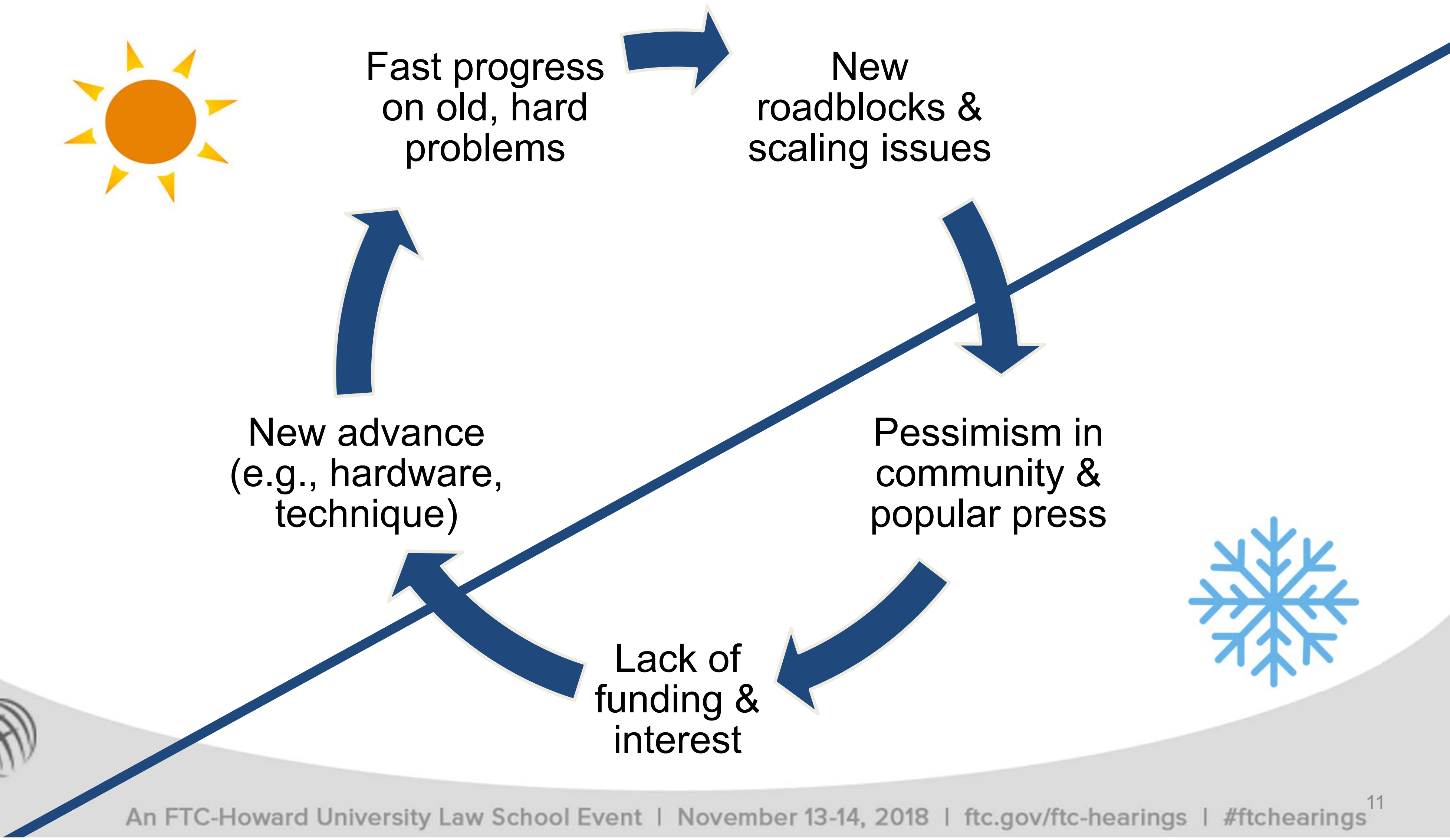
> … every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.

[McCarthy et al. 1955]

(Spoiler)

This hasn't happened yet.

But progress has been made.


Fast progress on old, hard problems
New roadblocks & scaling issues
Pessimism in community & popular press
Lack of funding & interest
New advance (e.g., hardware, technique)

But progress has been made.

So, what is AI?

- Artificial intelligence is the ability to process and act based on information via automation

Perceive	Learn
Abstract & Generalize	Reason & Act

First-Wave AI

- Search
  - Brute force search through solution space
  - Domain-specific and/or general heuristics
- Expert systems
  - A knowledge database (rules, facts)
  - Inference engine
  - I/O system to interact with a human

First-Wave AI: Drawbacks

- No real learning capability
- Huge overhead to encoding knowledge
- Brittle systems:
  - In-depth specific reasoning
  - Difficult to generalize

Perceive	Learn
Abstract & Generalize	Reason & Act

Transition Point(s)

- Natural Language Processing (NLP):
  - Before: hand-written syntax/semantics rules
  - 1980s+: probabilistic models based on large text corpora
- Autonomous vehicles (CV):
  - First DARPA Grand Challenge
  - 2005: in one year, **five** completions based on statistical models

[Vehicles] were scared of their own shadow, hallucinating obstacles when they weren't there.

[Strat 2004]

Transition Point(s)

- Similar transition points throughout core AI areas.

Computational power increases

Storage costs decrease

Reliance on statistical models increases

Second-Wave AI

- Encoding all knowledge explicitly does not work
  - Does not scale, very brittle, difficult to handle uncertainty, …
- New idea:
  - Create a general statistical model for a problem domain
  - Train that model on real-world data until it “looks right”
- Characterized by statistical learning
  - Give a different dataset, learn a different model

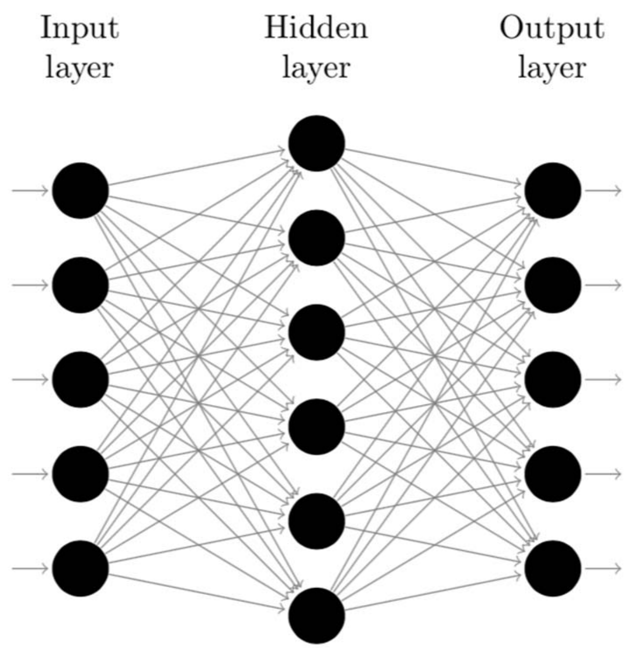
Second-Wave AI

- Machine translation:
  - – Multilingual text corpora → learn relationships between languages
- Autonomous vehicles:
  - – Videos/Tests of successful driving → learn what scenarios are "safe"
- Face detection and recognition:
  - – Many labeled faces of many people → learn what a face "looks like"

Perceive	Learn
Abstract & Generalize	Reason & Act

Example Model: Neural Networks

- Not a new idea!
- Information passed into an input layer
- Cascades through network to output layer
- Adjust strength of connections based on observed output

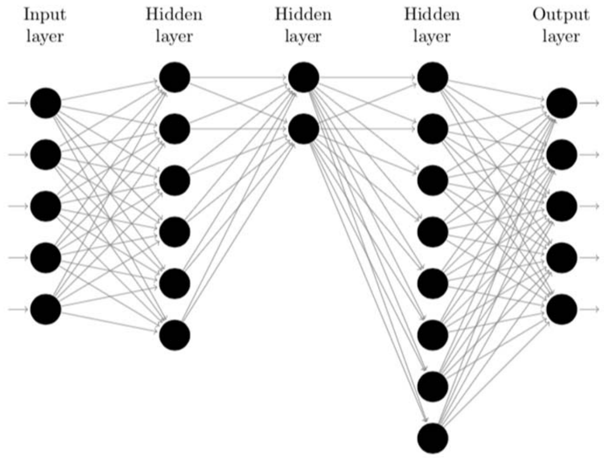


Feed data through network, observe computed outputs

Adjust weights on edges based on observed errors

Deep Neural Networks Work Very Well

- Deep networks are just neural networks with more "hidden layers"
  – Sometimes many more hidden layers
- Idea for, and exploration of, deep networks has existed since 1980s
  – Advances in hardware
  – Huge increase in training data
  – Better methods developed for training



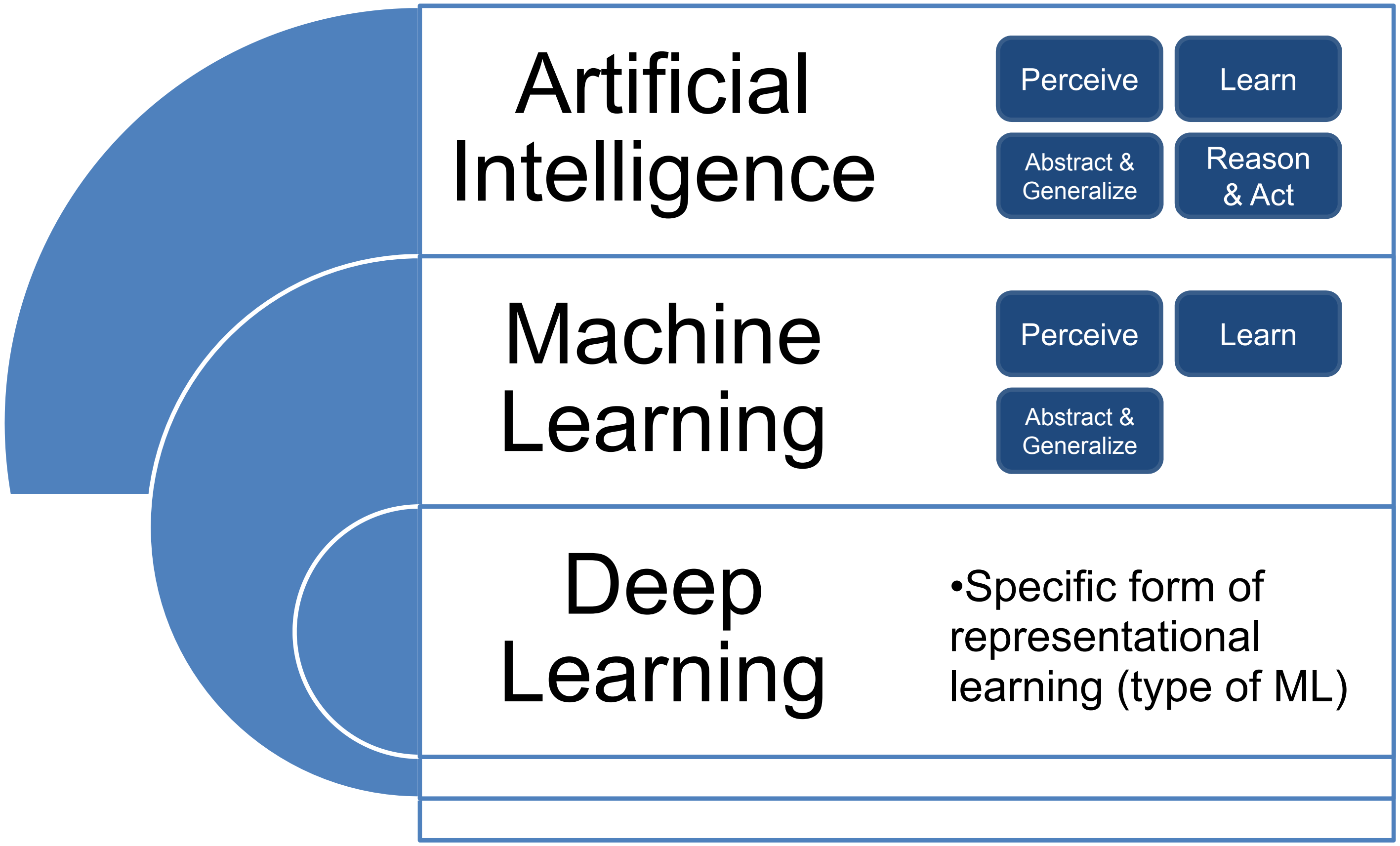
Deep Neural Networks Work Very Well

- Hugely successful
  - Anomaly detection
  - Voice recognition (Alexa, Siri, Assistant)
  - Machine translation, language generation
  - Game playing (AlphaGo, DeepStack)
  - Self-driving cars
  - Video search, audio search, finance, …

Nobody Understands Why Deep Neural Networks Work Very Well

- Humans design the network structure
  - – Encodes domain expertise, known heuristics
  - – Trial & error process – some automation here
- Nobody knows when or why they do not work*
  - – Work well in expectation
  - – Individual failure cases can be confusing & hard to explain
  - – Behavior can be exploited by adversaries

* except in special cases


Artificial Intelligence
Perceive
Learn
Abstract & Generalize
Reason & Act
Machine Learning
Perceive
Learn
Abstract & Generalize
Deep Learning
•Specific form of representational learning (type of ML)

Present-Day Movements in AI

- Understanding bias & methods for debiasing
  – Skewed training data produces skewed ML-based systems
- Adversarial reasoning & multi-agent systems
  – Learning to act with cooperative and/or adversarial actors
- Robustness to noise / adversarial attacks
  – Designing automated systems that fail less / more predictably
- Explainable AI (e.g., DARPA XAI)
  – Produce human-understandable models that also work well

Present-Day Movements in AI

- Reinforcement learning is a type of machine learning
  - Agent (physical or virtual) acts in an environment
  - Receives a reward signal, wants to maximize total reward
- Deep networks used extensively to learn, e.g., to reduce complexity of representing environment, value of actions

Perceive	Learn
Abstract & Generalize	Reason & Act

AI & Market Design

- Markets provide agents the opportunity to gain from trade
  - Many markets require structure to operate efficiently
  - Market design tackles this problem via “economic engineering”
- AI increasingly helps with the design of markets:
  - Automated methods use data to help designers characterize families of market structures
  - Predictive methods anticipate future supply and demand

Example: AI in Online Advertising

- Online advertising markets match advertisers with consumers
  - Many billions of USD, a driving force in technology sector
- Machine learning models:
  - Divide customers into fine-grained and automatically-generated segments
  - Set reserve prices in auctions based on user modeling and bidder behavior
  - Automatically generate creatives fit to customers' predicted wants
- Reinforcement-learning-based tools help advertisers bid better on fine-grained segments

Example: AI in Electricity Markets

- Matching supply and demand in electricity markets relies heavily on demand forecasting
- Machine-learning-based techniques:
  - Provide accurate demand forecasting, leading to stable market prices and more efficient power usage
- Reinforcement-learning-based techniques:
  - (De)activate heterogeneous power sources to maintain stability

Example: AI in Kidney Allocation

- Kidney exchanges are organized markets where patients with end-stage renal disease swap willing donors
  - 10%+ of all US living donations, 100s of transplant centers
- AI-based tools:
  - Automatically and optimally* match donors to patients (UNOS, UK NHS, Netherlands, …)
  - Provide sensitivity analysis for basic dynamic matching policies
  - Learn from data the quality of potential matches

* With respect to a human-defined model

Open Questions & Current Pushes

- How and why does deep learning work?
- How can we handle incentives of competing agents?
- Fairness, Accountability, and Transparency (FAT*)
  - How to define?
  - How to implement?
- Ethical AI
  - How is labor divided between ethically-minded policymakers and technically-trained AI/ML experts?
  - Close ties to privacy and social norms

The End Goal

Perceive

Learn

Abstract & Generalize

Reason & Act

Break

10:15-10:30 am

Understanding Algorithms, Artificial Intelligence, and Predictive Analytics Through Real World Applications

Session moderated by:

Karen A. Goldman
Federal Trade Commission
Office of Policy Planning

Harry Keeling
Howard University
Department of Computer Science

AI and Creativity

Dana Rao

General Counsel,

Executive Vice-President

Adobe

Why am I on this panel?

ACADEMIC JOURNAL ARTICLE

The George Washington Journal of International Law and Economics

Neural Networks: Here, There, and Everywhere-An Examination of Available Intellectual Property Protection for Neural Networks in Europe and the United States

By Rao, Dana S

Read preview

Article excerpt

1. Introduction

The development of neural networks, engineering constructs that simulate human neural interconnections,l has expanded rapidly in recent years.2 A neural network's structure allows it to "learn" information while training for a particular application.3 The network then can generalize the information to solve new problems outside the scope of its initial training.4 Neural networks differ from other forms of "artificial intelligence," such as expert systems and fuzzy logic, in that those technologies use a rules-based decision-making process and have no ability to learn.5 The U.S. Patent and Trademark Office (PTO) recognizes this difference and places "artificial intelligence" in a separate category.6 The particular characteristics of a neural network distinguish it from all other existing technologies and, thus, present unique intellectual property issues.

“The chief enemy of creativity is good sense.”

- Pablo Picasso

“Learn the rules like a pro,
so you can break them like an artist.”

- Pablo Picasso

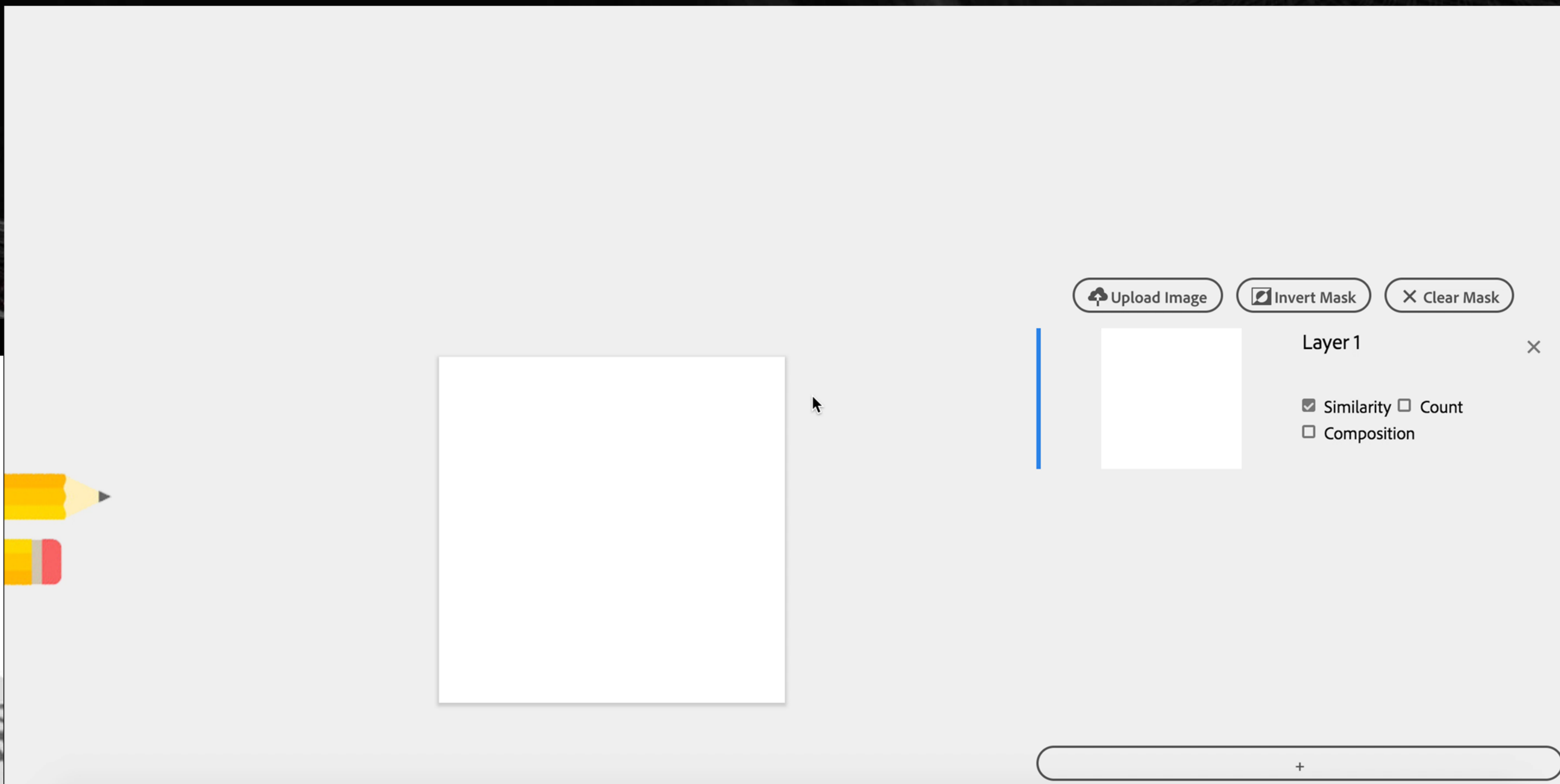
The Role of Today's AI in Digital Creativity

- Augment the creativity in all of us
- Minimize the mundane
- Meet the new demands for high volume content creation

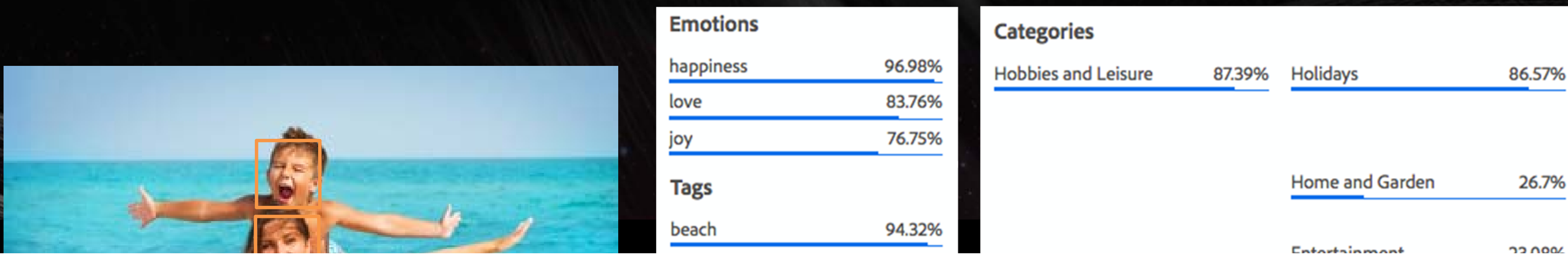

Content Understanding
Computational Creativity
Experience Intelligence


Actions
Content
Understanding
Objects
Concepts
Style
Aesthetics
Sentiment

Selective Similarity



Deep Learning for Content Understanding



Auto-Phrasing

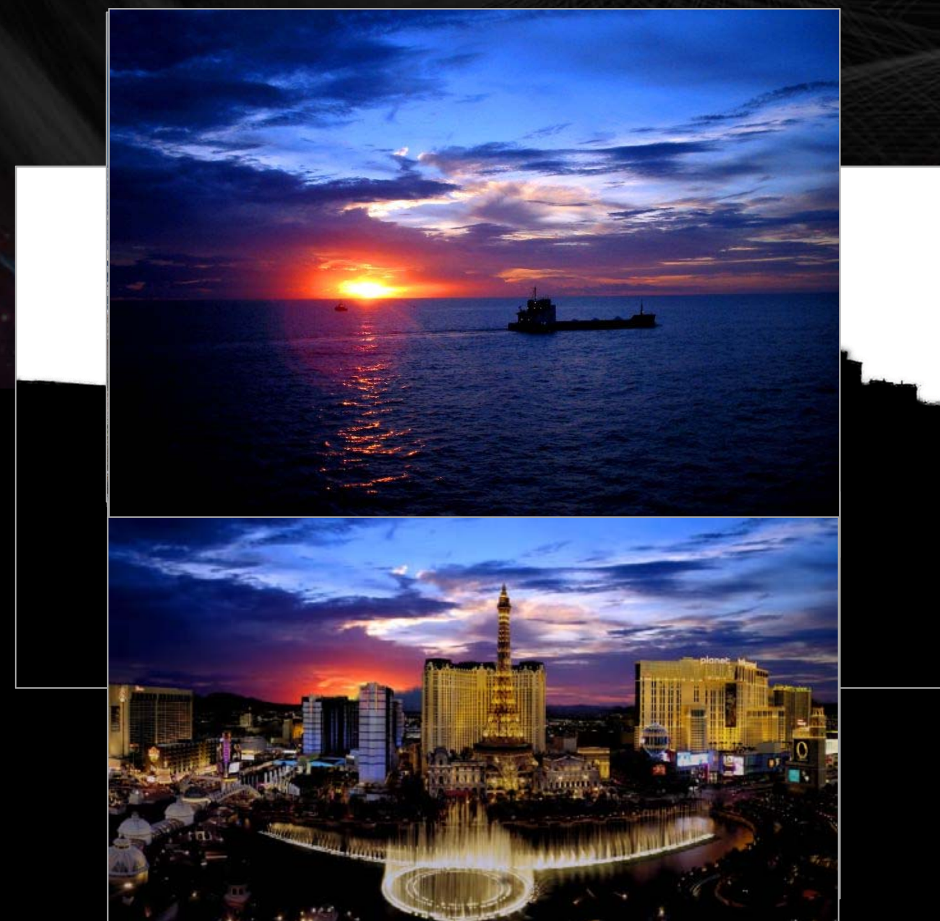
AI Techniques In Content Understanding

- Traditional ML and Deep Learning
- Behavioral feedback
- Image similarity
- Aesthetics, style
- Colors, foreground/background

Computational Creativity

- Augmentation of creators' skills and capabilities
- Workflow optimization

Semantic-Aware Sky Replacement



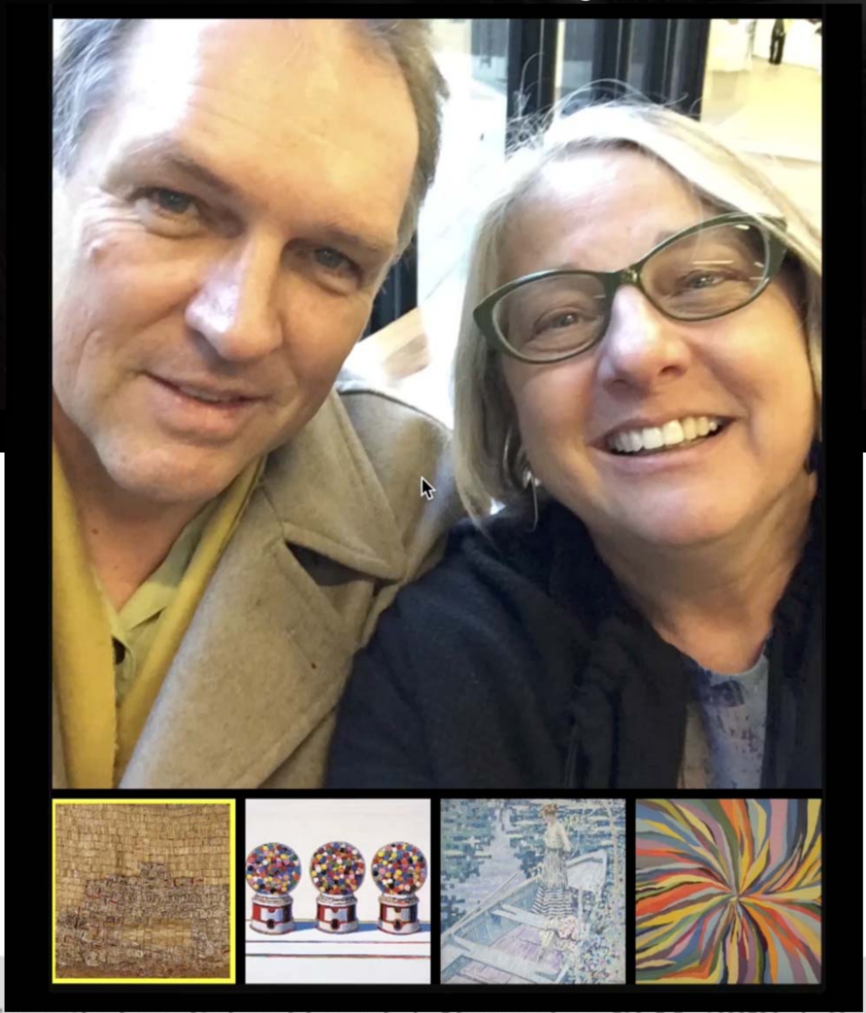
Tsai, Y., Shen, X., Lin, Z., Sunkavalli, K., Yang, M.

Neural Stylization

Eli Schechtman with Leon Gatys, Matthias Bethge (Tubingen), Aaron Hertzmann

Adobe

Artistic Eye



Project Cloak

Experience Intelligence

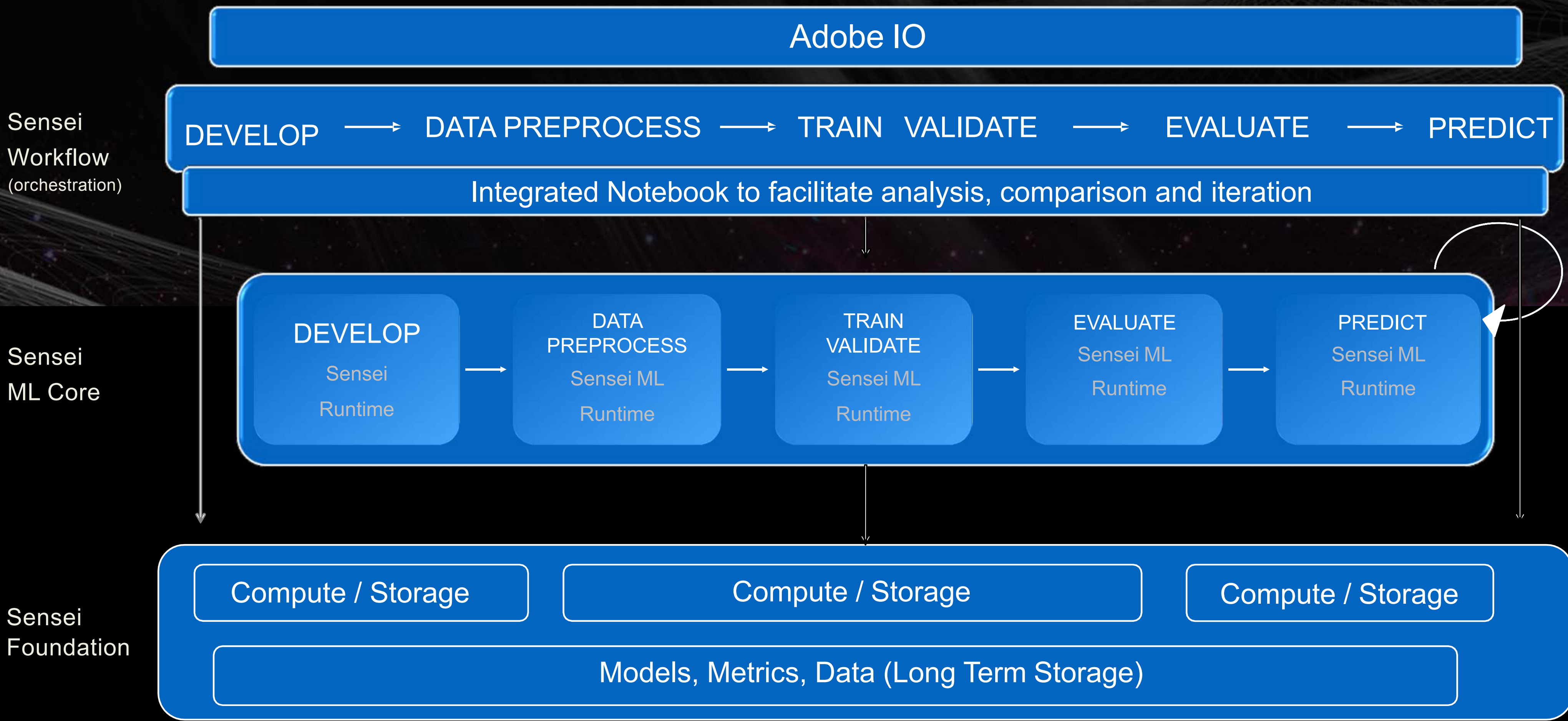
The right content for the right people at the right time

- Prediction and forecasting
- Personalization
- Recommendations
- Audience segmentation and clustering
- Optimization

A boom in content creation demand

How do we get there?

Sensei AI Training Lifecycle



Key principles for high quality AI training

- Millions of data points to train NN across images, videos, documents
- The more variety of data, the more accurate your NNs will be
- Bias is real

Adobe Sensei

Creativity in AI

- AI will become the Creative's assistant
- AI will help creatives be more creative
- AI will help bring creativity to all

Adobe Sensei

Adobe

NSF Support for AI Applications for Social Good

Henry Kautz

Director, Information & Intelligent Systems

Computer & Information Science & Engineering

National Science Foundation

NSF Award Criteria

- Advances Science or Engineering
- Potential for Broader Positive Impacts on Society
- Traditional broader impacts
  - Training graduate students
  - ***Potential future*** applications of the results to socially beneficial applications
- Increasingly:
  - Science and broader impacts are entwined
  - Work on beneficial applications leads to new scientific questions

AI and Broader Impacts

- AI methods are being used / advanced by researchers in every discipline funded by NSF – and many other agencies
- Crosscutting programs fund collaborations between AI researchers and application domain researchers
  - Smart and Connected Health
  - Smart and Connected Communities
  - Big Data

Future of Work at the Human-Technology Frontier

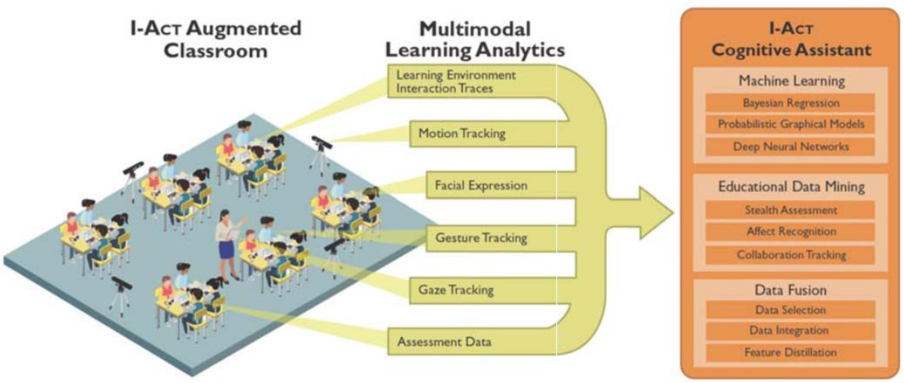
- Directorates: Computer Science, Engineering, Education, Social, Behavioral & Economic Sciences
- **Opportunities and (mitigating) risks** of the changing work landscape – many driven by advances in AI

Whole-body Exoskeletons for Advanced Vocational Enhancement

- **Work Domain:** Factory Work
- **Tech Innovation:** Full-body exoskeletons with embedded AI for factory workers.
  - Provides physical *and* cognitive enhancement.
  - Enables workers to seamlessly accomplish rapidly evolving, physically demanding tasks in networked, information-dense factory environment.
- **Work Impact:** Improve worker productivity, safety, comfort, & longevity; expand job opportunities by increasing employment & retention of diverse populations in physically demanding jobs.
- **SE Impact:** Understand the effects of augmentation on worker productivity & work-life satisfaction and labor market outcomes.

Virginia Polytech and State University

Transforming Teacher Work with Intelligent Cognitive Assistants



- **Work Domain:** Classroom teaching
- **Tech Innovation:** Intelligent Augmented Cognition for Teaching (I-ACT)
  - AI; motion tracking; eye tracking; facial expression reading; interaction logging; multimodal data fusion
  - Provides teachers with prospective, concurrent and retrospective pedagogical guidance
- **Work Impact:** improve teacher performance and quality of teacher work-life.
- **SE Impact:** Increase retention of K-12 STEM teachers, create a stronger STEM pipeline and a larger and better skilled STEM workforce. Improve US economic and societal well-being.

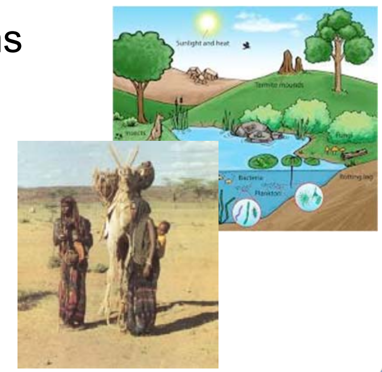
North Carolina State University

Expeditions in Computing

- NSF’s largest grants for computer science research
- Research of the highest intellectual merit
- Broader impacts address nation’s greatest needs
- Case Study:
  Institute for Computational Sustainability
  Cornell, Stanford, University of Southern California

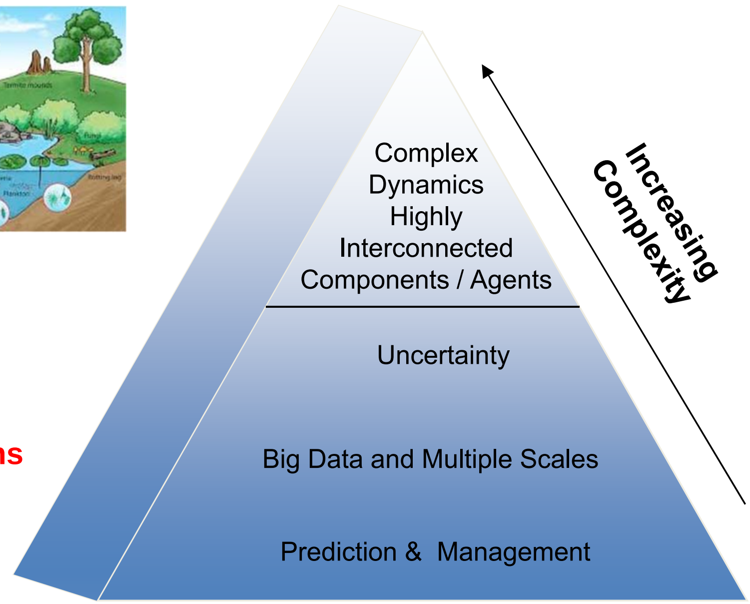
Sustainability Problems: Complex Systems

Sustainability problems
unique in scale
and complexity



Natural Ecosystems

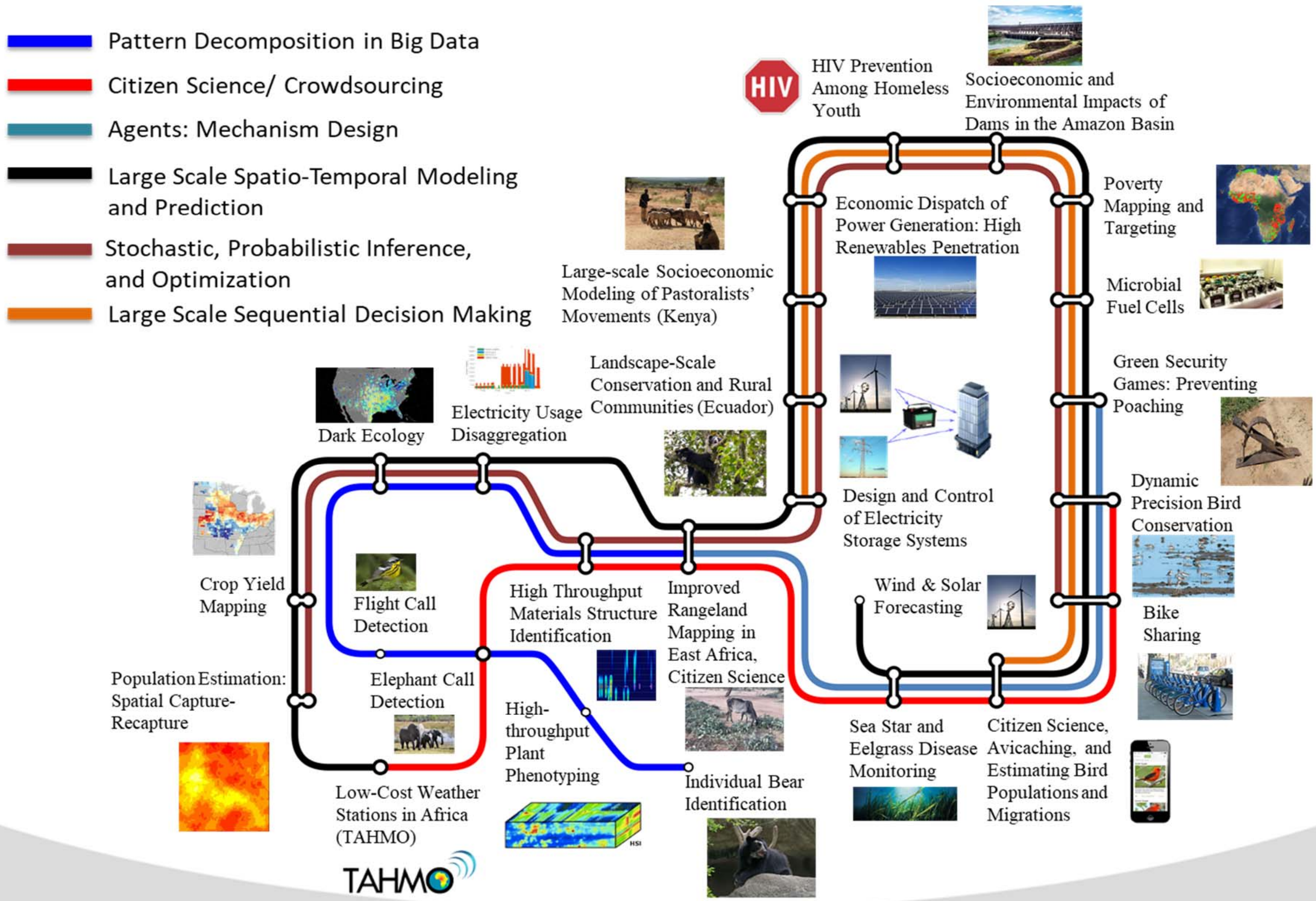
Smart Power Grid: Complex Digital Ecosystem



Complexity levels in Computational Sustainability Problems

Significant computational challenges:
Clear need for AI technology

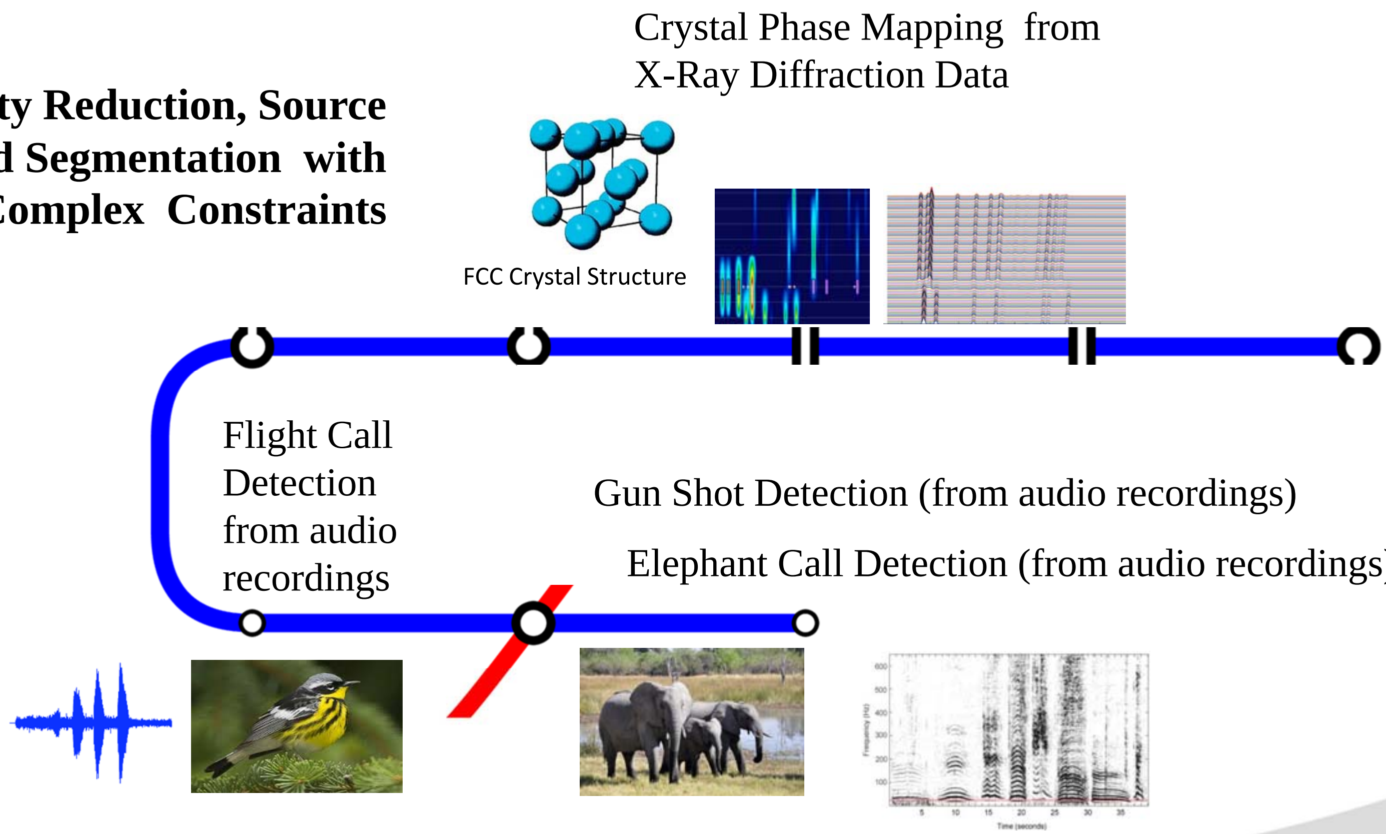
SUBWAY MAP



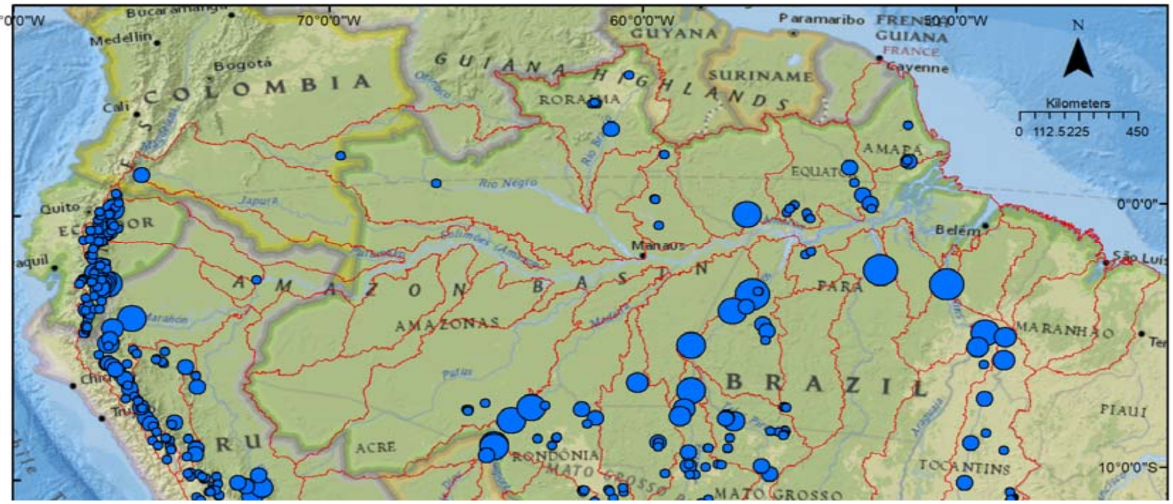
Pattern Decomposition in Big Data
Citizen Science/ Crowdsourcing
Agents: Mechanism Design
Large Scale Spatio-Temporal Modeling and Prediction
Stochastic, Probabilistic Inference, and Optimization
Large Scale Sequential Decision Making
HIV
HIV Prevention Among Homeless Youth
Socioeconomic and Environmental Impacts of Dams in the Amazon Basin
Poverty Mapping and Targeting
Economic Dispatch of Power Generation: High Renewables Penetration
Large-scale Socioeconomic Modeling of Pastoralists' Movements (Kenya)
Microbial Fuel Cells
Green Security Games: Preventing Poaching
Landscape-Scale Conservation and Rural Communities (Ecuador)
Electricity Usage Disaggregation
Dark Ecology
Dynamic Precision Bird Conservation
Design and Control of Electricity Storage Systems
Crop Yield Mapping
Flight Call Detection
High Throughput Materials Structure Identification
Improved Rangeland Mapping in East Africa, Citizen Science
Wind & Solar Forecasting
Bike Sharing
Population Estimation: Spatial Capture-Recapture
Elephant Call Detection
High-throughput Plant Phenotyping
Sea Star and Eelgrass Disease Monitoring
Citizen Science, Avicaching, and Estimating Bird Populations and Migrations
Individual Bear Identification
Low-Cost Weather Stations in Africa (TAHMO)
TAHMO

Pattern Decomposition in Big Data

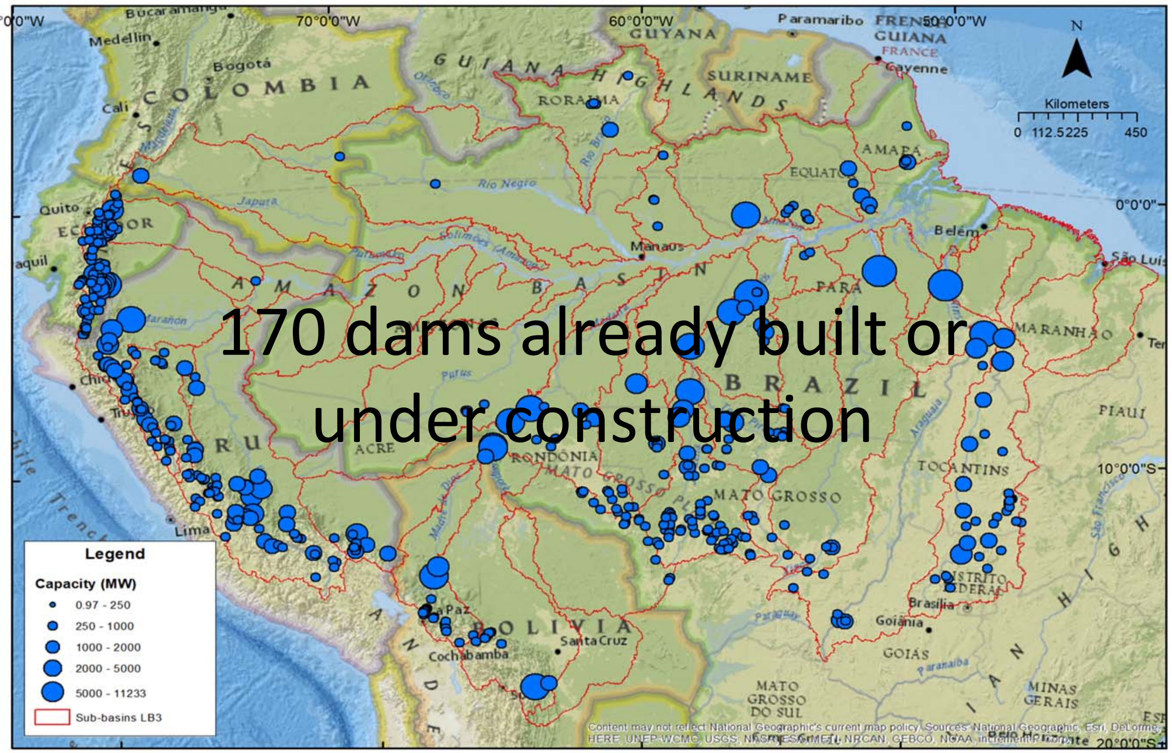
Dimensionality Reduction, Source Separation, and Segmentation with Complex Constraints



Hydropower Dam Proliferation in the Amazon Basin

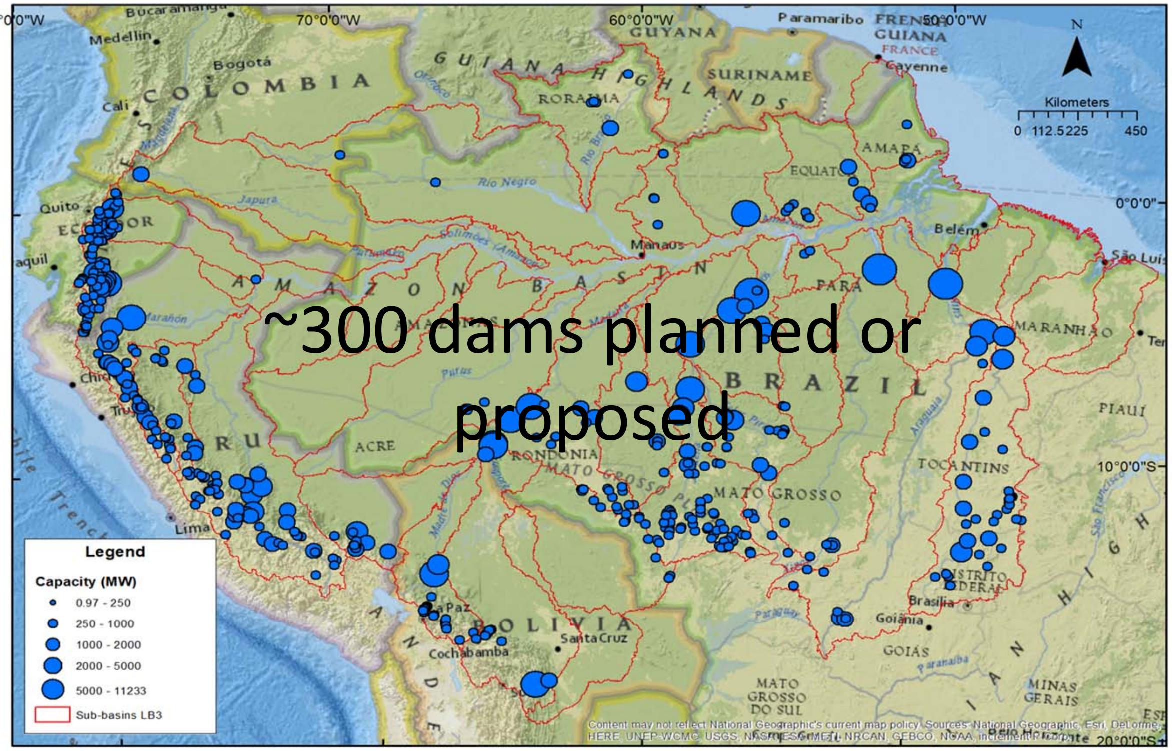


Hydropower Dam Proliferation in the Amazon Basin



170 dams already built or under construction

Hydropower Dam Proliferation in the Amazon Basin

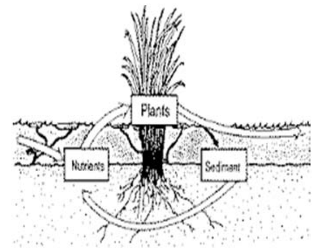


Ecosystem Services of River Networks

Energy

Fisheries

Transportation and navigation

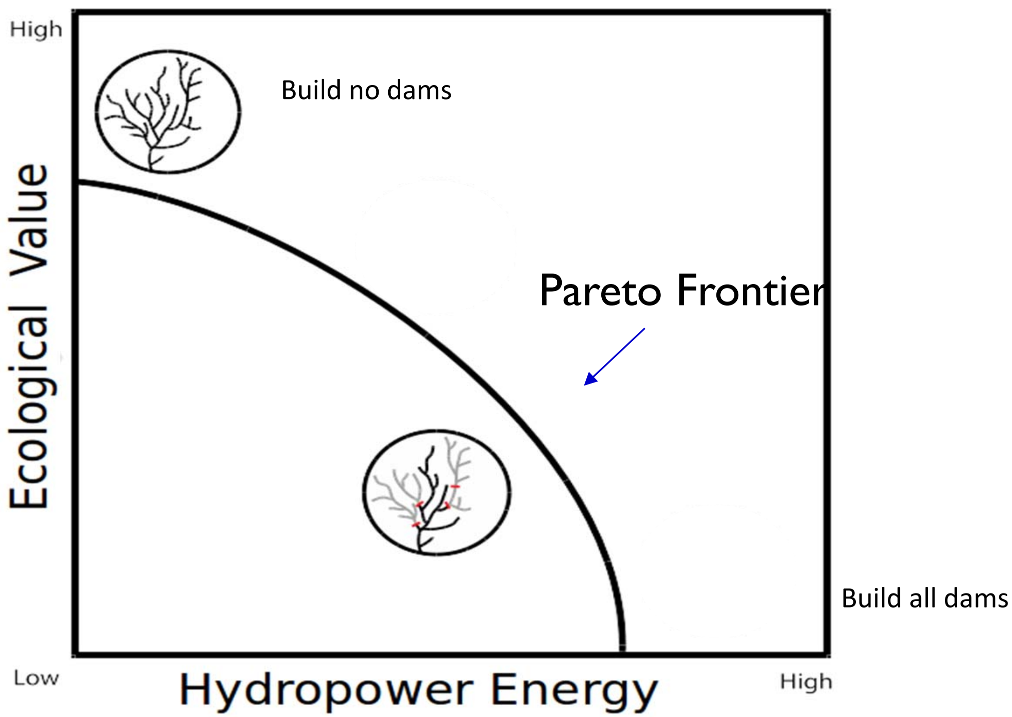


Sediments and Nutrients

Computational Perspective:
Multi-objective Optimization Problem

Pareto frontier:
the **trade-offs** wrt to the different objectives **of different non-dominated solutions of dam portfolios**

Goal: Find Optimal Portfolios of Dams to Build



AtlasAI

Startup (out of Stanford)

Funded by:

THE ROCKEFELLER FOUNDATION

Atlas AI uses AI cutting-edge techniques to build an accessible analytics platform to analyze and predict crop yields, economic well-being, and other **sustainable development indicators** at fine resolution across the developing world.

CompSustNet
Beyond Research:
Community Building, Education & Outreach

Research

- Coordinating transformative synthesis collaborations
- Interdisciplinary Research Projects (IRPs)

Community Building

- Web Portal
- Cross-Cutting Problems
- Panels & tutorials at major conferences
- Annual Conference
- Host visiting Scientists

CompSust.Net

Education

- Postdocs
- Doctoral students
- Undergrad Courses and Projects
- Research Virtual seminar series
- CompSust/DC Series
- Tutorials
- Cogs: ComputSust Graduate Seminars
- Sharing Code

Outreach

- Citizen Science Projects
- Engaging Gov., NGOs Institutions and Companies
- K-12 Outreach
- General Public Outreach
- Diversity
- Broad Dissemination of scientific results
- Shaping National Research Agenda on Sustainability and Societal Issues

AI in Credit Scoring

Angela Granger

VP Analytics, Experian

Why AI in Credit Scoring

- Credit scoring context
  - scores used to assess eligibility for credit where adverse action may be taken

- Benefits Lenders and Consumers
  - Better lending decisions: greater insights and more accurate scores
  - Financial inclusion: ability to include more data to broaden access to credit

Data Used in Credit Scoring

TRADITIONAL CREDIT DATA	**ALTERNATIVE** CREDIT DATA
Data assembled and managed in the core credit files of the nationwide consumer reporting agencies, which includes: • tradeline information (including certain loan or credit limit information, debt repayment history, and account status) • credit inquiries • public records relating to bankruptcies. Data customarily provided by consumers as part of applications for credit, such as income or length of time in residence and employment.	Data that are not “traditional.” We use “alternative” in a descriptive rather than normative sense and recognize there may not be an easily definable line between traditional and alternative data. Examples include: • Alternative Financial Service data (Short term/ Payday Loan, Title, Loan, Rent to Own) • Rental payments • Asset ownership • Utility payments • Full File Public Records • Consumer permissioned data

All data used in credit scores are FCRA compliant

1. Accurate
2. Disclosable
3. Disputable/correctable
4. Data furnishers play a role in the dispute process
5. Data is blind to ECOA factors: age, gender, marital status, ethnicity, race, religion.

Data Used in Credit Scoring

Generally acceptable	Generally not acceptable	Under consideration
• Data that complies with FCRA • Proven payment data like telephone and utilities • Rental data • DDA account transactions	• Unverified social media data • Data that could result in discriminatory decisions	• Education level

Developing Credit Scores

- Regulatory guidelines around accuracy and fairness in practice
  - Model governance (OCC guidelines) – documentation of build process, uses, monitoring
  - Controls around discrimination (ECOA) – need for transparency
  - Adverse action and Consumer Disclosures (required by FCRA) – need to be able to explain and provide consumer options to dispute and remedy

Key Considerations When Developing Credit Scores

- Model objective
- Data integrity
- Techniques
- Overfitting
- Parsimony
- Transparency
- Variable Selection
- Inference
- Adverse Action
- In and out of time validation
- Benchmarking
- Documentation to support model governance
- Deployment constraints
- Monitoring

Credit Scoring Modeling Methods

Sample parameters	**Techniques (single model)**

• Generic bureau data samples
  • Auto
  • Bankcard
• 90+ DPD performance flag
• 24-month outcome period
	• Logistic regression (LR)
	• Neural network (NN)
	• Random forest (RF)
	• Support vector machines (SVM)
	• Extreme gradient boosting (XGB)

Built preliminary unrefined models, evaluated performance on hold-out sample

Validation Gini					
	LR	**NN**	**RF**	**SVM**	**XGB**
Auto	71.34	73.87	73.21	73.98	**74.80**
Bankcard	69.30	72.11	72.31	72.22	**73.18**

Addressing Overfitting through model refinement

Logistic regression (LR)

- Built separate niche models
  - Thick clean
  - Thick dirty
  - Thin
- Variable count: 45 (15 per model)

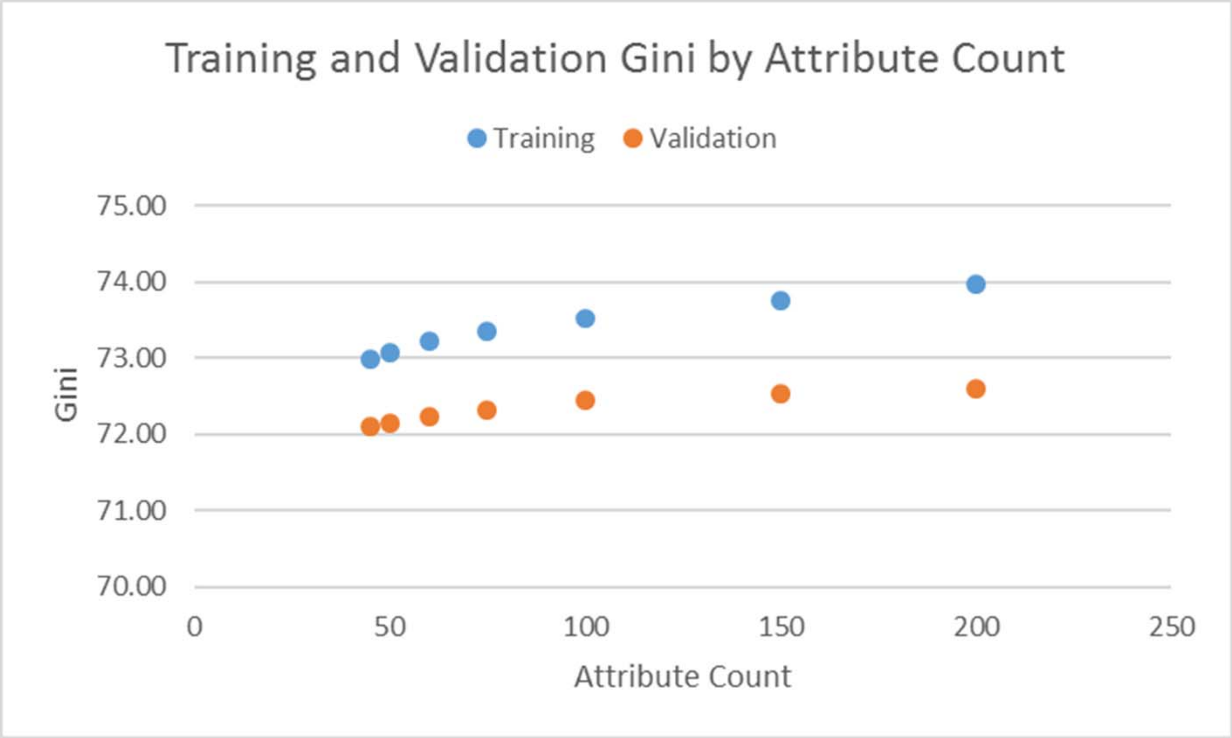
Extreme gradient boosting (XGB)

- Single model with conservative parameters to minimize overfitting and ease implementation
  - Forced attribute monotonicity
  - Imposed maximum variable count of 45

Validation Gini (after refinement)

	LR	XGB
Auto	72.51	**73.74**
Bankcard	70.62	**72.10**

Trade-off between attribute count and performance



Developing Credit Scores

- Advantages of AI in Credit Scoring
  - Quicker answers leading to faster time to market
  - Inclusion of new data sources, less time between updates

- AI uses beyond modeling methodology
  - Variable creation
  - Variable selection
  - Inference models
  - Benchmarking

Production Credit Scoring

- Credit scores are *static models*
  - Use real-time and batch data, no real-time model updates, refreshed regularly, replicable, documented changes for governance
  - Changes generally introduced through retro-studies and then Champion/Challenger approach

PROSPECTING	ACQUISITIONS	ACCOUNT MANAGEMENT	COLLECTIONS
• Decision on who to target for credit offers	• Decision on who to approve and at what credit terms	• Decision on authorizations and changes to credit terms	• Decision on collection strategy based on likelihood to collect

Future Policy Regarding Credit Scoring

- Financial inclusion
  - CFPB estimates 45million consumers are “credit invisible”
  - Regulatory incentives to include more rental, utility and telecom data in credit files – H.R. 435
    - Removes barriers to reporting but still voluntary
    - Positive payments known to be predictive of credit worthiness
      - PERC and Brookings Institute studies
    - Ease of implementation, ease of consumer understanding

Thanks!

Understanding Algorithms, Artificial Intelligence, and Predictive Analytics Through Real World Applications

Melissa McSherry

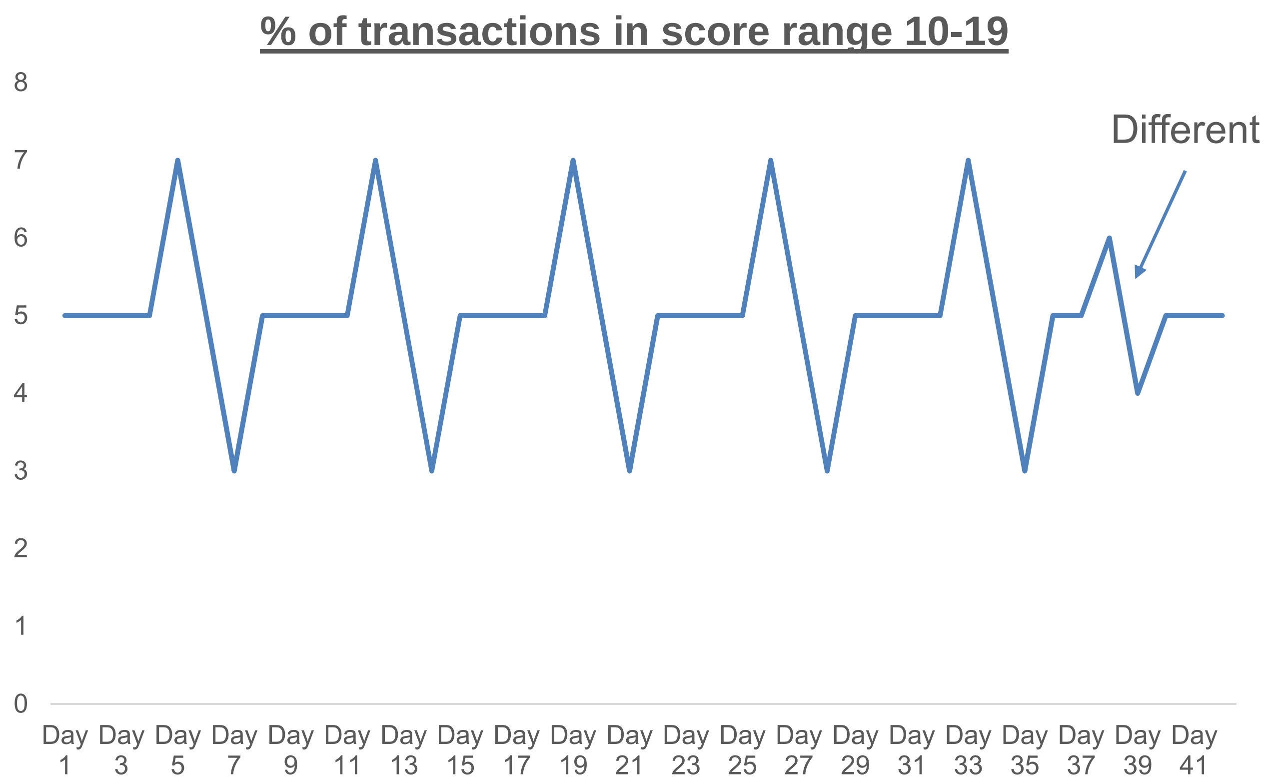
SVP, Global Head of Data Products

VISA

What is a fraud score? A fraud score distribution?

- A fraud score is the predicted probability that a transaction request is from someone other than the card holder. Visa provides scores from 0 to 99.
- Visa calculates a fraud score for every transaction going through VisaNet.
- Across all fraud scores, Visa can calculate the percentage that are in a particular range, for example 10-19. The percentages in each range are the fraud score distribution.
- We can monitor the fraud score distribution over time. This gives us a point of view on the stability of the whole system.

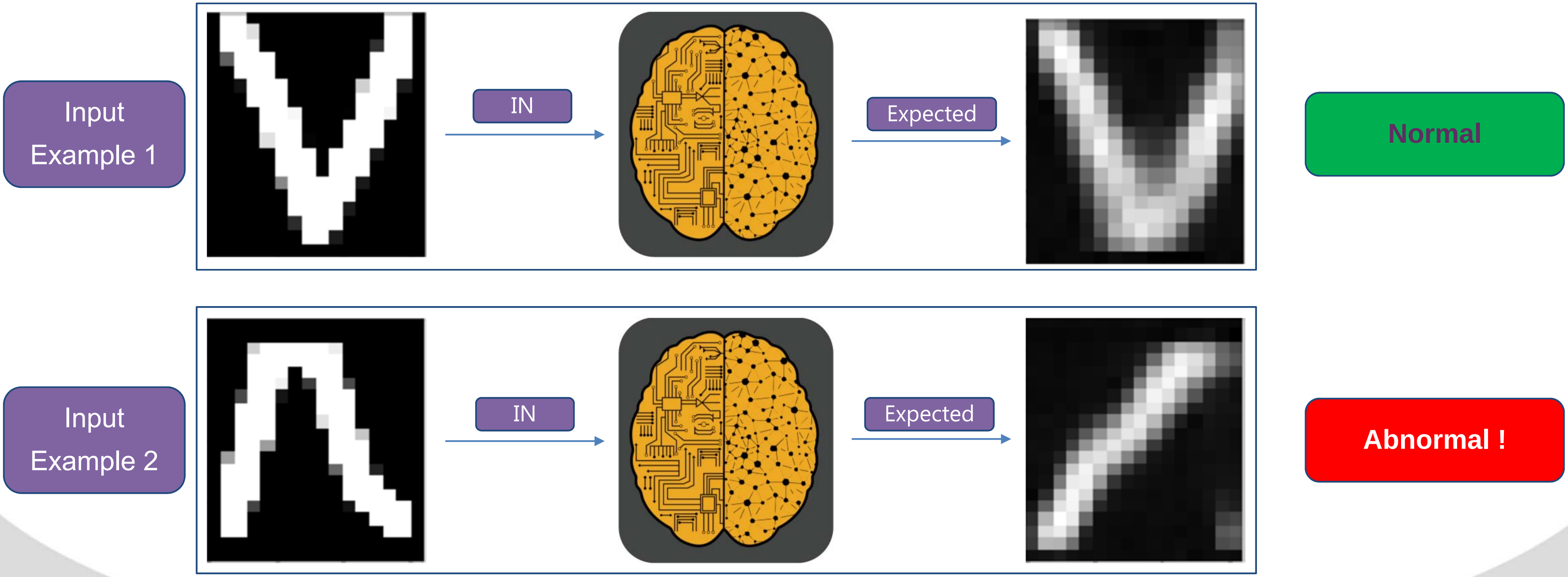
AI is a highly effective tool for identifying pattern variations



Imagine

- Every hour, every day
- Hundreds of metrics, not just one
- Hundreds of dimensions at once

We are using computer vision to monitor fraud score distributions today



Autonomous AI in Healthcare

Michael D. Abramoff, MD, PhD
Founder and CEO, IDx
The Robert C. Watzke Professor
University of Iowa

IDx
EyeDiagnosis.net

The why

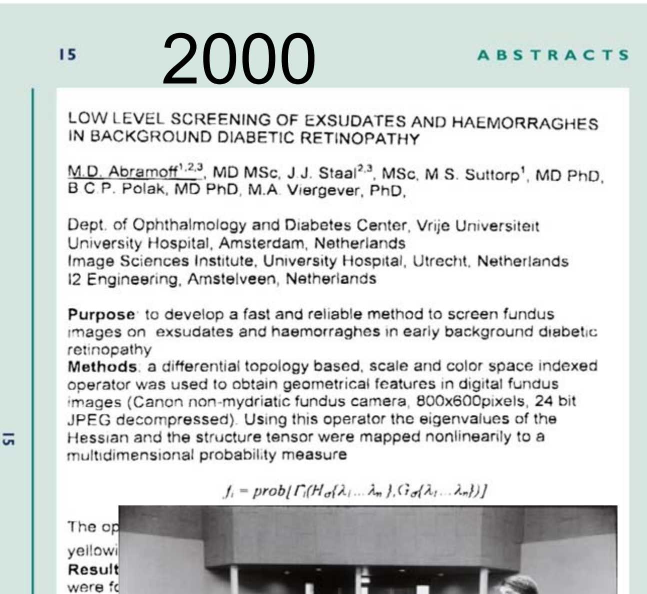
- Diabetes is primary cause of blindness: US ~24,000 p.a.
- Blindness is most feared complication of diabetes
- Vision loss preventable if detected early
- But we are not catching the patients early enough
  - <50% get retinal exam (CDC)
  - Many ACO's ~10-15%

The what

- Autonomous diagnostic AI system
  - Point of care result with minutes
  - No human review or oversight (IDx carries malpractice insurance)
  - Shifts specialty diagnostics from academic to primary care
- Robotic Camera
- Assistive AI for operator
- High school graduation level operator training
- Aligns with Clinical Standards

The How

2000


15 ABSTRACTS

LOW LEVEL SCREENING OF EXSUDATES AND HAEMORRAGHES IN BACKGROUND DIABETIC RETINOPATHY

M.D. Abramoff[1,2,3], MD MSc, J.J. Staal[2,3], MSc, M.S. Suttorp[1], MD PhD, B.C.P. Polak, MD PhD, M.A. Viergever, PhD,

Dept. of Ophthalmology and Diabetes Center, Vrije Universiteit University Hospital, Amsterdam, Netherlands
Image Sciences Institute, University Hospital, Utrecht, Netherlands
I2 Engineering, Amstelveen, Netherlands

Purpose: to develop a fast and reliable method to screen fundus images on exsudates and haemorraghes in early background diabetic retinopathy
Methods: a differential topology based, scale and color space indexed operator was used to obtain geometrical features in digital fundus images (Canon non-mydriatic fundus camera, 800x600pixels, 24 bit JPEG decompressed). Using this operator the eigenvalues of the Hessian and the structure tensor were mapped nonlinearly to a multidimensional probability measure

$$f_i = prob[\Gamma_i(H_\sigma\{\lambda_1 \ldots \lambda_m\}, G_\sigma\{\lambda_1 \ldots \lambda_m\})]$$

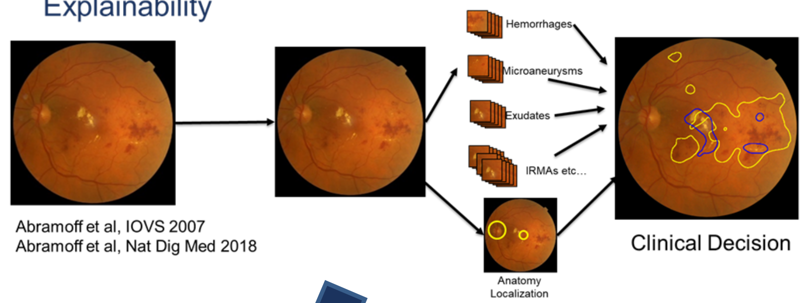
The op
yellowi
Result
were fo

15

2014

Explainability

Biomarker Detection (mostly CNN)



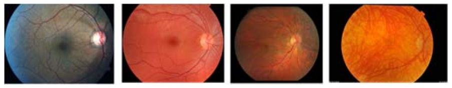
Abramoff et al, IOVS 2007
Abramoff et al, Nat Dig Med 2018

Clinical Decision

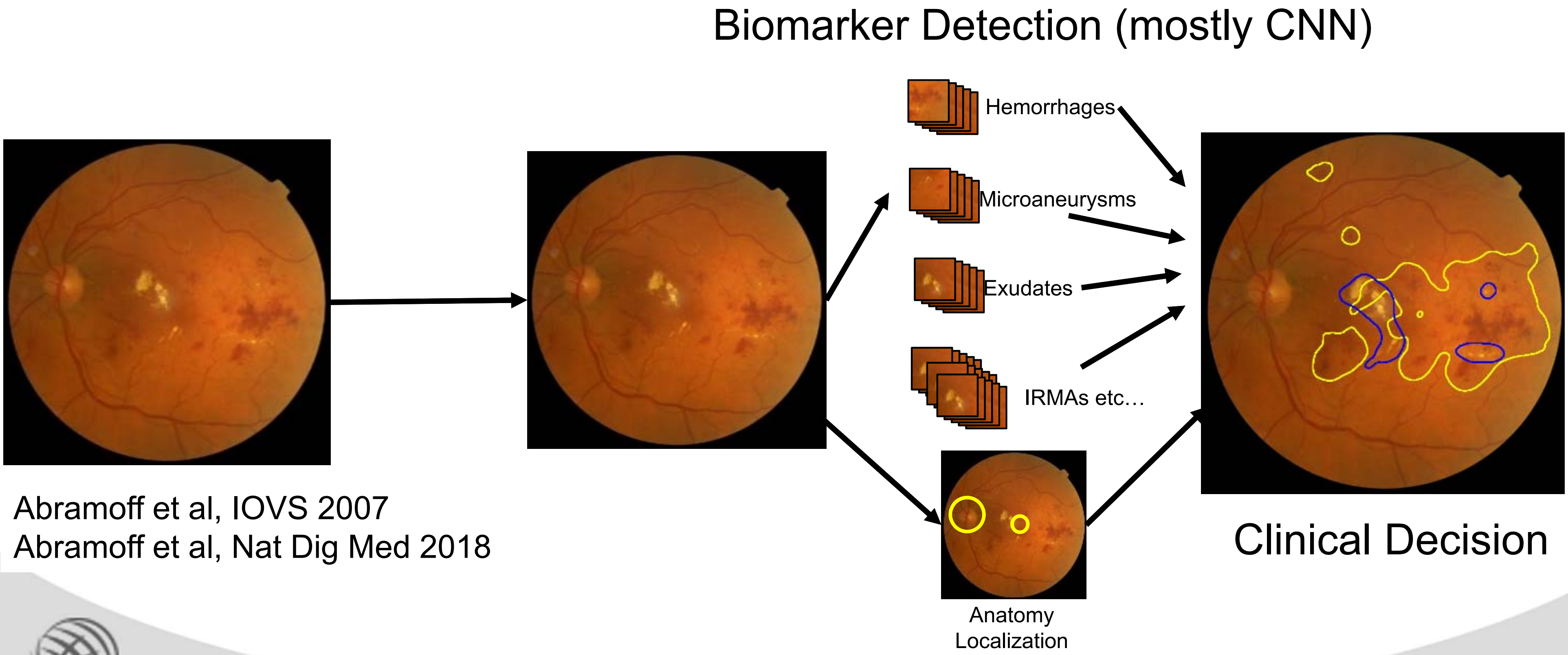
2018

2000 – 2014: Scientific / ‘Device’ Stage

- Insights from neuroscience and evolution of mammalian vision
  - -> explainable AI and how to avoid racial and ethnic bias
- Insights from clinical evidence
  - -> dealing with diagnostic drift and measurable performance
  - -> how do you validate AI when expert sensitivity < 40%
- Insights from implementation:
  - -> Importance of image quality



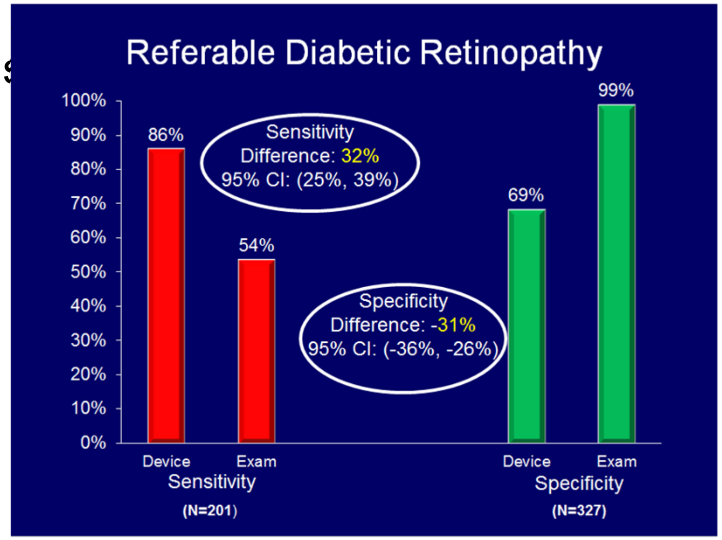
Explainable AI based on evolution of mammalian vision



Abramoff et al, IOVS 2007
Abramoff et al, Nat Dig Med 2018

2014: FDA ‘rejects’ clinical trial that met endpoint

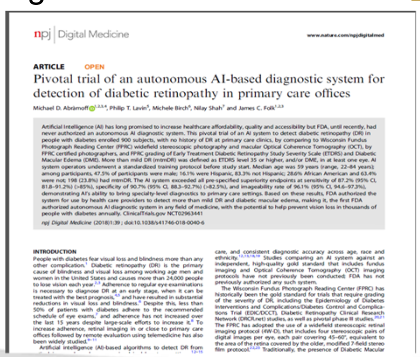
- Now, we agree with them!
  - Even though Sensitivity significantly better than clinicians
- Evaluated AI algorithm on image reading
  - No widefield stereo imaging, 3D OCT imaging
  - AI and Reading Center use same images
  - Not widely accepted reading center
- Clinical trial not in real world setting
  - In ophthalmology clinics, not primary care clinics
  - Not primary care diabetes population
  - Excluded patients with insufficient image quality
  - Highly experience ophthalmic photographers



Maguire et al, ARVO 2015

2018: Clinical Trial and FDA De Novo ‘Authorization’

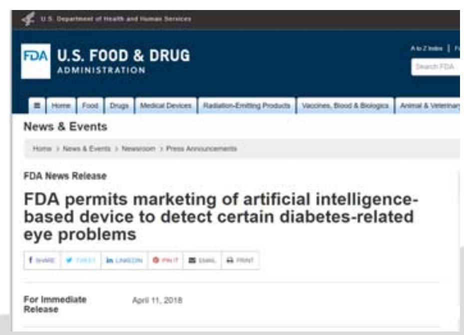
- System validation
  - Primary care clinics
  - Primary care patient sample
  - Primary care existing staff
- Highest level truth
  - Leading Reading center
  - Experienced retinal imagers
  - Both 2D and 3D imaging
    - More than 2x retinal area
- Autonomous AI system consists of
  - Robotic camera
  - Operator 4 hour standardized training
  - Assistive operator AI
  - Diagnostic AI


npj | Digital Medicine

ARTICLE OPEN

Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices

- Preregistered clinical trial
- Repeatability and reproducibility
- Human Factors Validation
- ‘Endpoints’:
  - Diagnostic accuracy
    - Sensitivity
    - Specificity
  - Image-ability



AI System design/deployment standards

- IDx Quality Management System, audited by Underwriter Laboratory
  - complaint, feedback, corrective action, regulatory reporting, post market monitoring, etc.
- 21 CFR 820 FDA Current Good Manufacturing Practice
- ISO 13485 – Medical Device Quality Management Systems
- IEC 14971 – Applications of Risk Management to Medical Devices
- IEC 62366 - Application of Usability Engineering to Medical Devices
- ISO 62304 – Medical Device Software Life Cycle Process
- HIPAA & EU General Data Protection Regulation (GDPR)
- SOC 2 Auditing - Cybersecurity

Cleared the way for Autonomous AI

- IDx established FDA pathway
  - Product code: PIB
  - Regulation Number: 21 CFR 886.1100
  - Special Controls for Autonomous AI
- <u>Reducing time to market for future products</u>

U.S. FOOD & DRUG
ADMINISTRATION

April 11, 2018

IDx, LLC
% Janice Hogan
Regulatory Counsel
Hogan Lovells US LLP
1735 Market Street, Suite 2300
Philadelphia, Pennsylvania 19103

Re: DEN180001
Trade/Device Name: IDx-DR
Regulation Number: 21 CFR 886.1100
Regulation Name: Retinal diagnostic software device
Regulatory Class: Class II
Product Code: PIB
Dated: January 12, 2018
Received: January 12, 2018

Implications for Autonomous AI

- Biomarker based diagnostic algorithms
  - Explainable
  - Avoid catastrophic failure
  - Avoid racial and ethnic bias in diagnostic accuracy
- Highest achievable Reference standard
  - Reference imaging protocol
  - Reading protocol – avoid clinicians
  - Alignment of reference standard with PPP
- System level validation
  - Operator
  - Camera
  - Diagnostic AI
  - Operator assistive AI
- Preregistered clinical trial
  - AI locked down – deterministic AI
  - Algorithm Integrity Provider
  - Prospectively defined se/sp, imageability
  - Avoid replication crisis
- Human Factors Validation
  - Training, image capture, patient workflow
  - Regulatory oversight of medical labeling/output
  - Indicated for only clinically validated camera
- Specialty specific requirements
  - 7-field stereo equivalent field of view
  - 3D imaging: OCT
  - Individual disease biomarker validation
- Training data stewardship
  - Full traceability and transparency of training data
  - Source, truth, number of images
- Implementation
  - QMS, Cybersecurity, Privacy
  - Medical malpractice Insurance

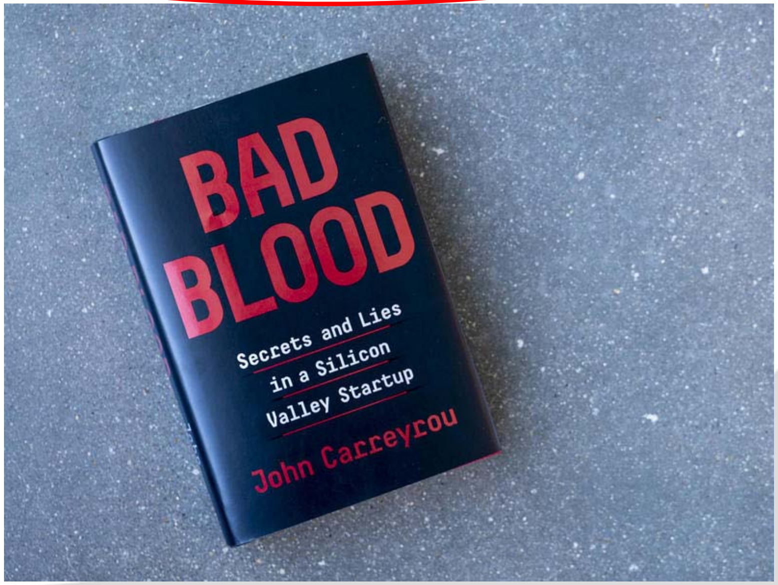
Safe Implementation of Autonomous AI in Medicine

- Agreement on definitions and nomenclature
- Tech-world software design and development paradigms do not directly transfer
  - cannot use a “fail fast and learn’ approach
  - cannot bypass regulatory / medical-scientific standards
- New AI algorithms do not directly transfer
  - Explainable AI instead of Black box
  - Full stewardship and transparency of training data
  - Design addresses bias & catastrophic failure
- Preregistered clinical trials paramount
  - Best reference standard: frequently not clinicians
  - Incorporating the intended context and workflow
- Oversight and claims enforcement
  - Reliable framework to understand and trust autonomy levels
- AI autonomy and company liability vs physician liability


THE NEW YORKER

A REPORTER AT LARGE OCTOBER 22, 2018 ISSUE

“If it is your job to advance technology, safety *cannot* be your No. 1 concern,” Levandowski told me. “If it is, you’ll never do anything. It



Artificial Intelligence for Health and Health Care

Teresa Zayas Cabán, PhD
Chief Scientist
Office of the National Coordinator for Health Information Technology

Agenda

- Study Goals and Questions
- Areas of Focus
- Is AI Ripe for Health and Health Care?
- Challenges
- Current Efforts

Recent Clinical Applications

Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection

Insights Into the Black Box for Pathologists

Yun Liu, PhD; Timo Kohlberger, PhD; Mohammad Norouzi, PhD; George E. Dahl, PhD; Jenny L. Smith, MD; Arash Mohtashamian, MD; Niels Olson, MD; Lily H. Peng, MD, PhD; Jason D. Hipp, MD, PhD; Martin C. Stumpe, PhD

• *Context.*—Nodal metastasis of a primary tumor influences therapy decisions for a variety of cancers. Histologic identification of tumor cells in lymph nodes can be laborious and error-prone, especially for small tumor foci.

Objective.—To evaluate the application and clinical implementation of a state-of-the-art deep learning–based artificial intelligence algorithm (LYmph Node Assistant or LYNA) for detection of metastatic breast cancer in sentinel lymph node biopsies.

Design.—Whole slide images were obtained from hematoxylin-eosin–stained lymph nodes from 399 patients (publicly available Camelyon16 challenge dataset). LYNA was developed by using 270 slides and evaluated on the remaining 129 slides. We compared the findings to those obtained from an independent laboratory (108 slides from 20 patients/86 blocks) using a different scanner to measure reproducibility.

Results.—LYNA achieved a slide-level area under the receiver operating characteristic (AUC) of 99% and a tumor-level sensitivity of 91% at 1 false positive per patient on the Camelyon16 evaluation dataset. We also identified 2 "normal" slides that contained micrometastases. When applied to our second dataset, LYNA achieved an AUC of 99.6%. LYNA was not affected by common histology artifacts such as overfixation, poor staining, and air bubbles.

Conclusions.—Artificial intelligence algorithms can exhaustively evaluate every tissue patch on a slide, achieving higher tumor-level sensitivity than, and comparable slide-level performance to, pathologists. These techniques may improve the pathologist's productivity and reduce the number of false negatives associated with morphologic detection of tumor cells. We provide a framework to aid practicing pathologists in assessing such algorithms for adoption into their workflow (akin to how a pathologist assesses immunohistochemistry results).

(*Arch Pathol Lab Med.* doi: 10.5858/arpa.2018-0147-OA)

Study Goals and Questions

- Understand the full impact that AI can have on health and health care
  - How can AI shape the future of public health, community health, and health care delivery from a personal level to a system level?
  - Understand the opportunities and considerations that can better prepare and inform developers and policy makers and promote the general welfare of health care consumers

Areas of Focus

- Opportunities
- Considerations
- Implementation

Areas of Focus

- **Opportunities**
- Considerations
- Implementation

Areas of Focus

- Opportunities
- **Considerations**
- Implementation

Areas of Focus

- Opportunities
- Considerations
- **Implementation**

Is AI Ripe for Health and Health Care?

- **Broad advances in AI are significant and real**
  1. Frustration with the existing – or legacy – medical systems among patients and health professionals
  2. Ubiquity of networked smart devices in society
  3. Comfort with at-home services like those provided through Amazon and other technology companies

Blog Post: www.healthit.gov/buzz-blog/Jason

Report: www.healthit.gov/jason

Six Domains of Significant Challenges

1. Acceptance of AI applications in clinical practice will require immense validation
2. Ability to leverage the confluence of personal networked devices and AI tools
3. Availability of and access to high quality training data from which to build and maintain AI applications in health
4. Executing large-scale data collection to include missing data streams
5. Building on the success in other domains, creating relevant AI competitions
6. Understanding the limitations of AI methods in health and health care applications

Six Domains of Significant Challenges

1. **Acceptance of AI applications in clinical practice will require immense validation**
2. Ability to leverage the confluence of personal networked devices and AI tools
3. Availability of and access to high quality training data from which to build and maintain AI applications in health
4. Executing large-scale data collection to include missing data streams
5. Building on the success in other domains, creating relevant AI competitions
6. Understanding the limitations of AI methods in health and health care applications

Six Domains of Significant Challenges

1. Acceptance of AI applications in clinical practice will require immense validation
2. **Ability to leverage the confluence of personal networked devices and AI tools**
3. Availability of and access to high quality training data from which to build and maintain AI applications in health
4. Executing large-scale data collection to include missing data streams
5. Building on the success in other domains, creating relevant AI competitions
6. Understanding the limitations of AI methods in health and health care applications

Six Domains of Significant Challenges

1. Acceptance of AI applications in clinical practice will require immense validation
2. Ability to leverage the confluence of personal networked devices and AI tools
3. **Availability of and access to high quality training data from which to build and maintain AI applications in health**
4. Executing large-scale data collection to include missing data streams
5. Building on the success in other domains, creating relevant AI competitions
6. Understanding the limitations of AI methods in health and health care applications

Six Domains of Significant Challenges

1. Acceptance of AI applications in clinical practice will require immense validation
2. Ability to leverage the confluence of personal networked devices and AI tools
3. Availability of and access to high quality training data from which to build and maintain AI applications in health
4. **Executing large-scale data collection to include missing data streams**
5. Building on the success in other domains, creating relevant AI competitions
6. Understanding the limitations of AI methods in health and health care applications

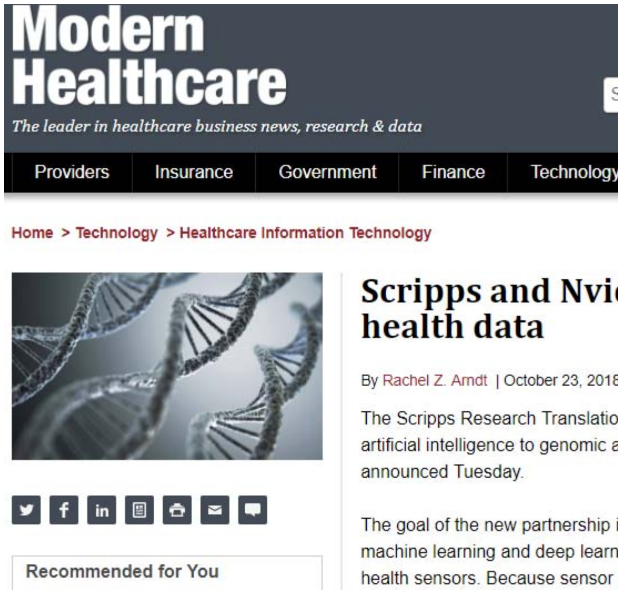
Six Domains of Significant Challenges

1. Acceptance of AI applications in clinical practice will require immense validation
2. Ability to leverage the confluence of personal networked devices and AI tools
3. Availability of and access to high quality training data from which to build and maintain AI applications in health
4. Executing large-scale data collection to include missing data streams
5. **Building on the success in other domains, creating relevant AI competitions**
6. Understanding the limitations of AI methods in health and health care applications

Six Domains of Significant Challenges

1. Acceptance of AI applications in clinical practice will require immense validation
2. Ability to leverage the confluence of personal networked devices and AI tools
3. Availability of and access to high quality training data from which to build and maintain AI applications in health
4. Executing large-scale data collection to include missing data streams
5. Building on the success in other domains, creating relevant AI competitions
6. **Understanding the limitations of AI methods in health and health care applications**

The Potential of AI for Health and Health Care


Modern Healthcare

The leader in healthcare business news, research & data

Providers | Insurance | Government | Finance | Technology

Home > Technology > Healthcare Information Technology

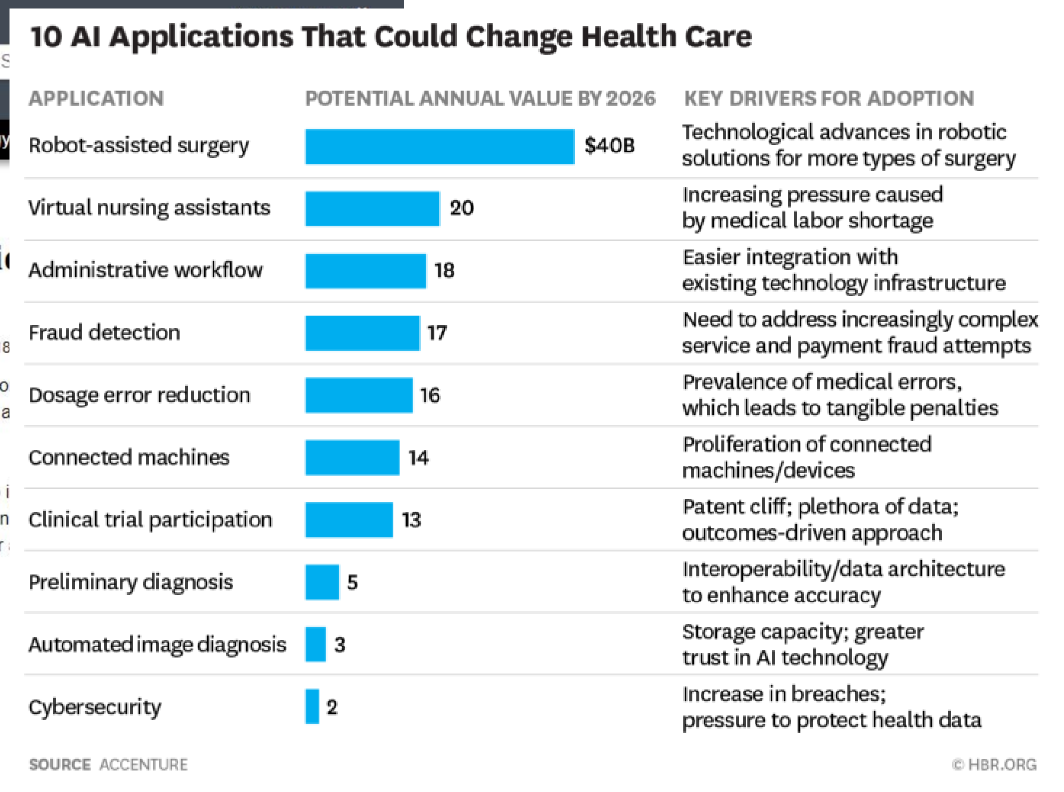
Recommended for You

Scripps and Nvi
health data

By Rachel Z. Arndt | October 23, 2018

The Scripps Research Translatio
artificial intelligence to genomic a
announced Tuesday.

The goal of the new partnership i
machine learning and deep learn
health sensors. Because sensor



ONC's Role Moving Forward

- Work with other agencies to define and identify possible opportunities

- Work towards interoperable and standardized health data

Current Efforts

- National Cancer Institute – Department of Energy (DoE)
  - CANDLE (CANcer Distributed Learning Environment)
- Veterans Affairs – DoE
  - Big Data Science Initiative

Understanding Algorithms, Artificial Intelligence, and Predictive Analytics Through Real World Applications

Panel Discussion:

Michael D. Abràmoff, Angela Granger,
Henry Kautz, Melissa McSherry,
Dana Rao, Teresa Zayas Cabán

Moderators: Karen A. Goldman & Harry Keeling

Lunch

12:15-1:15 pm

Perspectives on Ethics and Common Principles in Algorithms, Artificial Intelligence, and Predictive Analytics

Session moderated by:

Karen A. Goldman
Federal Trade Commission
Office of Policy Planning

James Trilling
Federal Trade Commission
Division of Privacy and Identity Protection

Fairness and Bias in Machine Learning and Artificial Intelligence Systems

James Foulds
Department of Information Systems
University of Maryland,
Baltimore County

Work sponsored in part by the
National Institute of Standards and Technology (NIST)

Machine Learning

- **Machine learning algorithms**, which make predictions based on data, are having an increasing impact on our daily lives.

- Example: **credit scoring**
  - Predicting whether you will repay or default on a loan

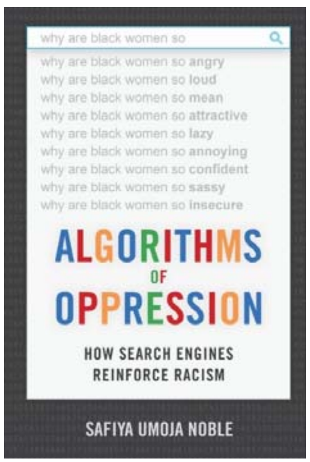
# Late Payments	% of available credit used	Previous defaults?	Employed?	…	**Repay Loan?**

Feature vector X **Class label Y**

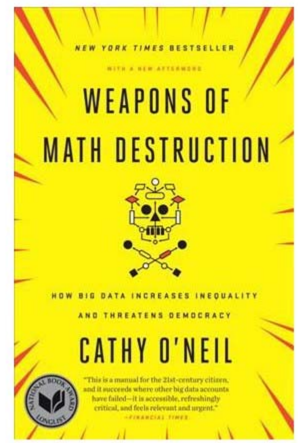
- The models are "trained" on many labeled feature vectors
- This is called **classification**, an instance of **supervised machine learning**

Fairness in Machine Learning

- There is growing awareness that **biases inherent in data** can lead the behavior of machine learning algorithms to **discriminate against certain populations**



Big Data: A Report on



Bias in Predicting Future Criminals

- Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**)
  - An algorithmic system for **predicting risk of re-offending** in criminal justice, by Northpointe company
  - Used for sentencing decisions across the U.S.

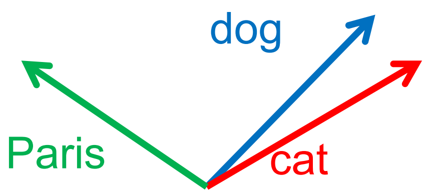
- ProPublica study (Angwin et al., 2016):
  - **COMPAS almost twice as likely to incorrectly predict re-offending for African Americans than for white people**.
    Similarly much more likely to incorrectly predict that white people would not re-offend. Northpointe disputes the findings

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Illustrative Example: Sentiment Analysis

- An example from “How to make a racist AI without really trying” blog post by Rob Speer

- Application: sentiment analysis
  - Predict whether the sentiment expressed in a text is positive or negative



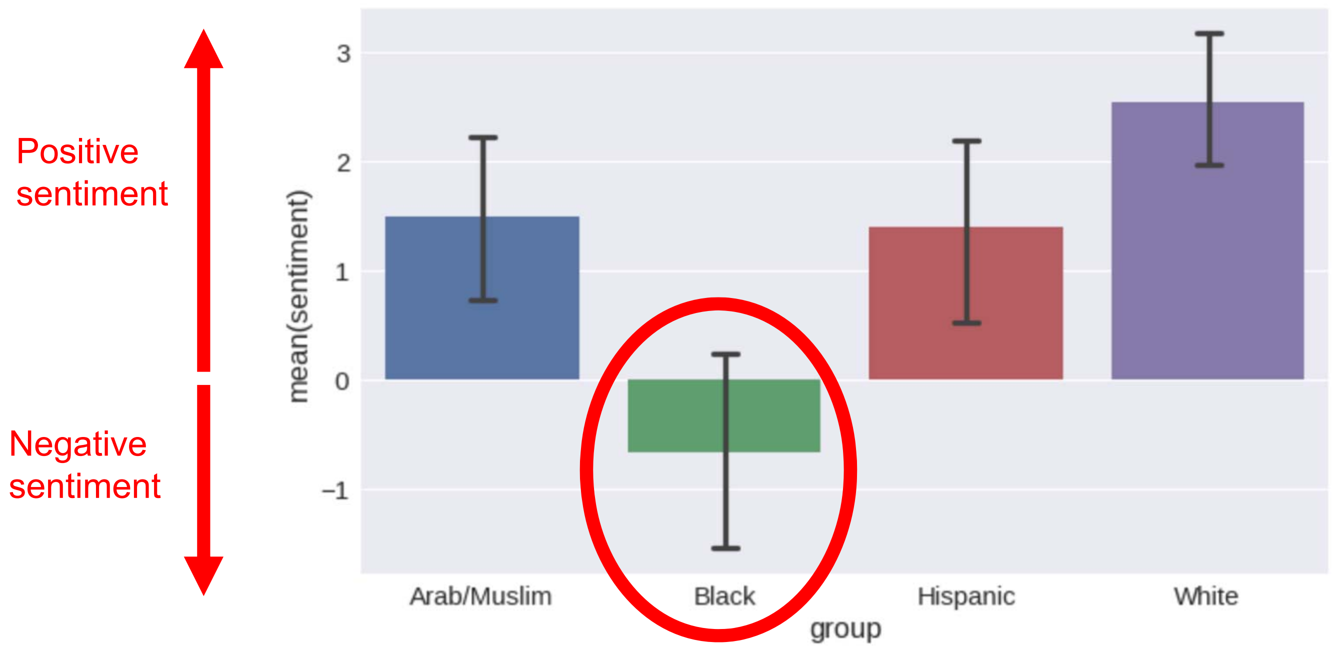
dog: (0.11, -1.5, 2.7, …)
cat: (0.15, -1.2, 3.2, …)
Paris: (4.5, 0.3, -2.1, …)

Feature vector X

Class label Y

Illustrative Example: Sentiment Analysis

- **Sentiment of stereotypical names for different race groups**
  (bar plot with 95% confidence interval of means shown)



Amazon scraps secret AI recruiting tool that showed bias against women

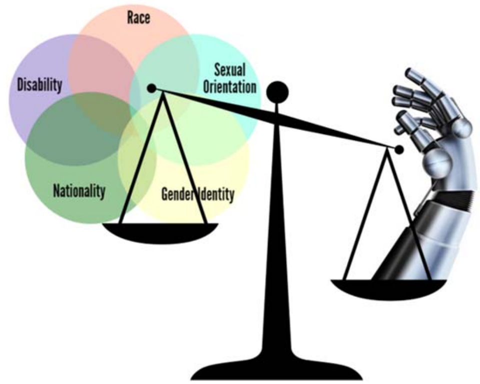
Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, retrieved 10.24.2018.

Sources of Bias in Data (cf. *Barocas and Selbst (2016)*)

- Data encodes **societal prejudices**
  - e.g. racism/sexism in social media data
- Data encodes **societal (dis)advantages**
  - college admissions, criminal justice
- **Less data** for minorities
- **Collection bias**
  - data from smartphones, automobiles,…
- **Intentional** prejudice. **Digital redlining**, **masking**
  - St. George's Hospital Med School encoded its existing race/gender-biased decision-making for admissions interviews in an algorithm (Lowry & McPherson, 1988)
- **Proxy variables**
  - (e.g. zip code highly correlated with race, leading classifier to unintentionally consider race)



Considerations

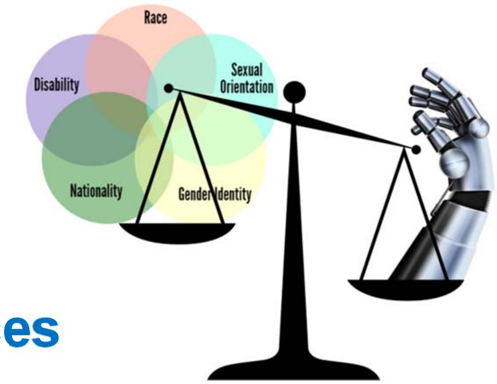
- **Fairness is a highly complicated socio-technical-political-legal construct**

- **Harms of representation** vs **harms of outcome**
  (cf. Kate Crawford, Bolukbasi et al. (2016))

- Differences between **equality** and **fairness**
  (Starmans and Sheskin, 2017) How to balance these?

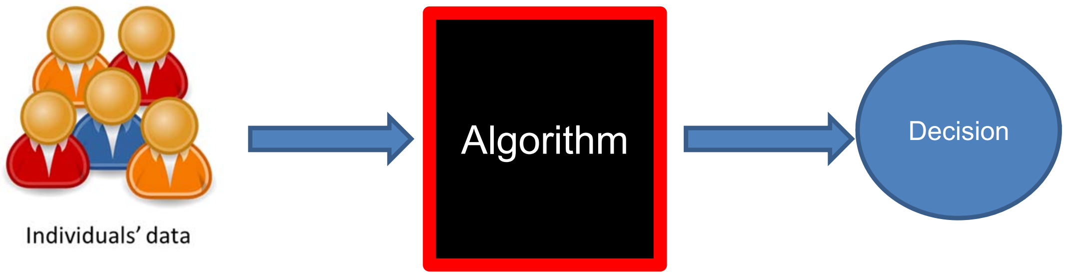
- Whether (and how) to model underlying **differences between populations** *(Simoiu et al., 2017)*

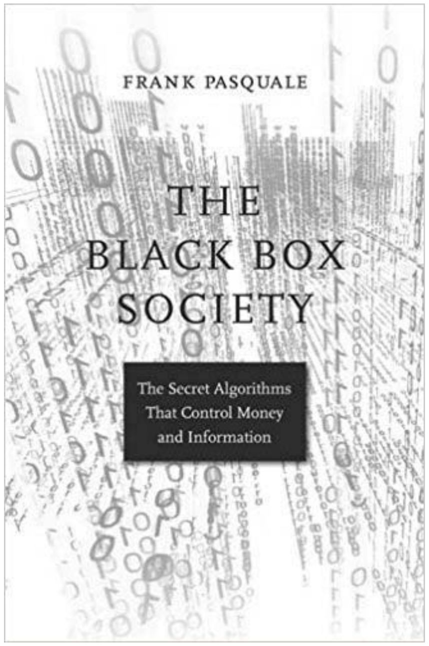
- Whether to aim to correct **biases in society** as well as biases in data (**fair affirmative action**) *(Dwork et al., 2012)*



Explainability and Transparency

- The algorithms making decisions that affect our lives are often inscrutable black boxes





- If an algorithm harms us, often we have no knowledge or recourse
- Legal "right to explanation" in certain cases
  - U.S.: Credit scores, Equal Credit Opportunity Act (1974)
  - European Union: General Data Protection Regulation (2018)

"the existence of automated decision-making … and … meaningful information about the logic involved"

Fairness and the Law: Adverse Impact Analysis

- Title VII, other anti-discrimination laws prohibit employers from intentional discrimination against employees with respect to protected characteristics
  - gender, race, color, national origin, religion

- Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission)

The "four-fifths rule" (a.k.a. 80% rule)

*"A **selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate** will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact."*

-Code of Federal Regulations 29 Part 1607 (1978)

The Machine Learning / AI Community's Response to Fairness

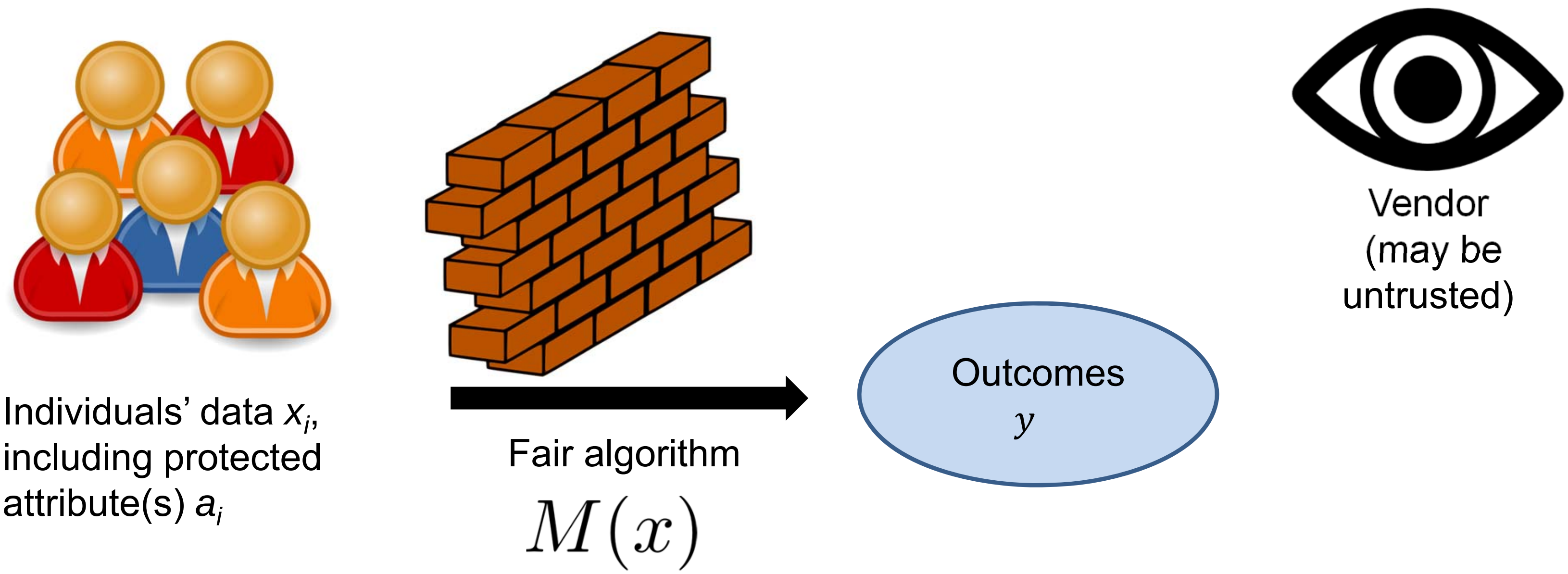
- A recent explosion of research (since circa 2016)

- **Publication venues** dedicated to fairness and related issues
  - Fairness, Accountability and Transparency in ML (FAT/ML) Workshop
  - ACM FAT*
  - AAAI/ACM Conference on AI, Ethics & Society

- **Mathematical definitions**, **algorithms** for enforcing and measuring fairness

AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY

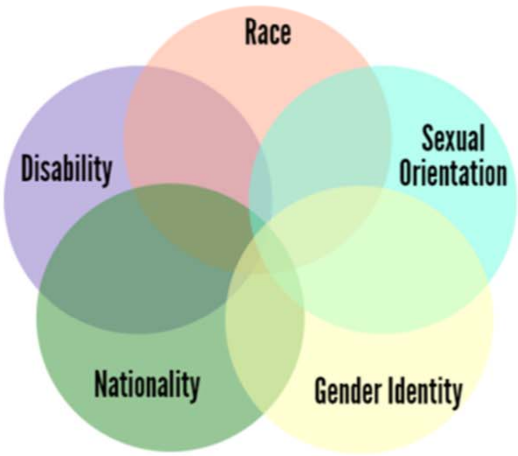
Fairness and Privacy: the Untrusted Vendor (Dwork et al., 2012)



The user of the algorithm's outputs (the *vendor*) may discriminate, e.g. in retaliation for a fairness correction (Dwork et al., 2012)

Fairness and Intersectionality

- **Intersectionality:**
  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  - gender, race, nationality, sexual orientation, disability status, socioeconomic class, …



versus

- **Infra-marginality:**
  attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.

My research: Differential Fairness (DF)

We propose a fairness definition with the following properties:

- **Measures** the **fairness cost** of **algorithms** and **data**
  - Can measure difference in fairness between algorithms and data: **bias amplification**
- **Privacy** and **economic guarantees**
  - Privacy perspective provides an **interpretation** of definition, based on **differential privacy**
- Implements **intersectionality**, e.g. fairness for (gender, race) probably ensures fairness for gender and for race separately

Essentially, differential fairness generalizes the 80% rule. Multiple protected attributes and outcomes, provides a privacy interpretation

Paper preprint: J. R. Foulds and S. Pan. **An Intersectional Definition of Fairness.** arXiv:1807.08362 [CS.LG], 2018. https://arxiv.org/pdf/*1807.08362*

Thank you!

- Contact information:
  - James Foulds
    Assistant professor
    Department of Information Systems
    UMBC
    Email: jfoulds@umbc.edu
    Webpage: http://jfoulds.informationsystems.umbc.edu

- A pre-print of our work is online at arxiv.org:
  J. R. Foulds and S. Pan. ***An Intersectional Definition of Fairness.*** *ArXiv preprint arXiv:1807.08362 [CS.LG], 2018.*

https://arxiv.org/pdf/1807.08362

SIIA’s Ethical Principles for Artificial Intelligence

Mark MacCarthy
Senior Vice President for Public Policy
Software & Information Industry Association

Software & Information Industry Association

- The Software & Information Industry Association is the principal trade association for the software and digital content industry. SIIA provides global services in government relations, business development, corporate education and intellectual property protection to the leading companies that are setting the pace for the digital age.

Ethical Principles for AI

- Belmont Principles
- FAT/ML Principles
- ACM Principles
- SIIA Principles

SIIA’s Ethical Principles for AI

- Rights
- Justice
- Welfare
- Virtue

Rights

- Engage in data practices that respect internationally recognized principles of human rights.
- The framework of human rights requires organizations to respect the **equal dignity and autonomy of individuals**

Which Rights?

- These rights include the right to **life, privacy, religion, property, freedom of thought, and due process before the law.**
- Organizations should validate these universal aspects of human nature by engaging only in data practices that respect fundamental human rights.

Justice

- Individuals have rights based on justice to a **fair share of the benefits and burdens of social life.**
- Aim for an equitable distribution of the benefits of data practices and **avoid data practices that disproportionately disadvantage vulnerable groups.**

Distribution of Benefits

- The benefits of advanced analytical services should be available to all and **not restricted** based on arbitrary and irrelevant characteristics such as **race, ethnicity, gender, or religion**

Responsibility for the Use of Models

- Organizations share responsibility for how the models they develop **are used and by whom and how the benefits of their new analytical services are distributed**

Welfare

- Aim to create the greatest possible benefit from the use of data and advanced modeling techniques
- **Increase human welfare** through improvements in the provision of public services and low-cost, high-quality goods and services.

Virtue

- Engage in data practices that encourage the practice of virtues that contribute to human flourishing.
- Data and advanced modeling techniques should be designed and implemented to enable people, individually and collectively, to **further their efforts to become people capable of living genuinely good lives in their communities.**

Which Virtues?

- Data practices should allow affected people to develop and maintain moral virtues
- Such as **honesty, courage, moderation, self-control, humility, empathy, civility, care, and patience**

Do it All!

- Organizations need not choose one of these principles to the exclusion of the others.
- Use them jointly as general guides to the development of ethical data practices.

Domain-Specific Principles

- These general principles need to be supplemented with specific principles appropriate to the context or domain of use.

Disparate Impact Analysis

- A key part of assessing compliance with statutory and constitutional prohibitions on discrimination.
- Should also be used to assess AI decision-making algorithms as designed and as they evolve and adjust themselves in use.

Stages of Disparate Impact Analysis

- Evidence of a disproportionate adverse impact
- Legitimate purpose served
- Alternatives that achieve the legitimate objective with less

Which Groups to Assess?

- Protected classes include race, gender, religion, ethnicity.
- Consider expanding to vulnerable groups also at risk but not explicitly protected by law.

Which Purposes to Assess?

- Law protects eligibility decisions in employment, housing, insurance, credit.
- Consider expanding to include consequential decisions that affect a person’s life chance.

SIIA Issue Brief:
Ethical Principles for Artificial Intelligence

Ethical Principles for Artificial Intelligence:

http://bit.ly/2zkNmp5

Understanding Algorithmic Bias: Primary and Secondary Consumer Harms

Dr. Rumman Chowdhury
Global Lead, Responsible AI,
Accenture

@ruchowdh
www.rummanchowdhury.com

The **Responsible AI team at Accenture** is dedicated to creating human-centric Artificial Intelligence. Our goal is to understand and address the social, regulatory and economic impact of this technology from development to deployment and beyond. The team also serves as the starting point for governance internally at Accenture.

The Responsible Artificial Intelligence Team led by Dr. Rumman Chowdhury, data scientist and social scientist, and Deborah Santiago, senior leadership in Accenture Legal.

Why does technology need ethics?

2017:
AWARENESS

Evangelizing and educating on Responsible AI as a global imperative.

2018:
ACTION

Moving from virtue signaling to positive action. Providing concrete tools and assets for clients.

2019:
AGENCY & ACCOUNTABILITY

Democratizing RAI with a focus on enterprise applications of Responsible AI.

What is bias?

Technologists mean: Experimental Bias

DATA BIAS

- Selection or sampling bias: Is your data representative of the population the model will be used on?
- Measurement bias: Both measurement instrument and operationalization can be faulty.

RESPONSE /REPORTING BIAS

- How is the data being picked up, and might that introduce bias?
- Is the data sensitive in nature; is there reason to misrepresent the truth? Will people have the same metrics of reporting (e.g., yelp effect)?

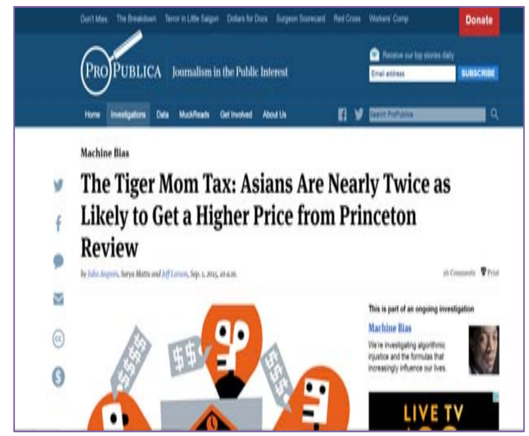
DESIGN BIAS

- What assumptions are you making about your model and its applicability to the question?
- Are you engineering a feedback loop?

Non-technologists mean: Societal Bias

Data is not an objective truth.
It is reflective of pre-existing institutional, cultural, and social biases.

- Loss of Opportunity
- Economic Loss
- Social Detriment
- Loss of Liberty

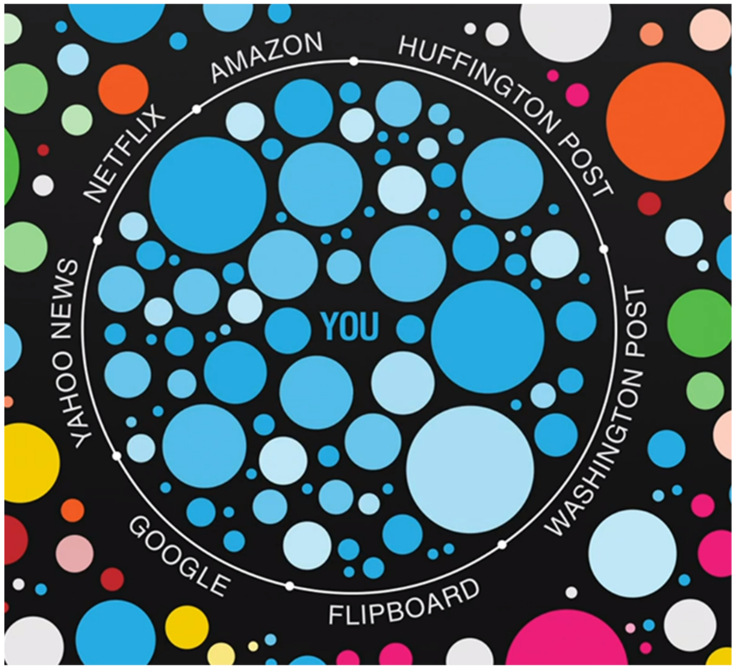


Our language around ‘bias’ tends to mean ‘primary harms’.

Algorithmic Determinism and Secondary Harms

Is social media radicalizing us?

- Does the filter bubble lead to ideological polarization?
- Personalization paradox and confirmation bias



TELEVISION

The New York Times

SUBSCRIBE NOW LOG IN

Some Viewers Think Netflix Is Targeting Them by Race. Here's What to Know.

Promotional images taken from four different Netflix accounts for the movie "Set It Up." Clockwise, from top left: Zoey Deutch and Glen Powell; Deutch and Powell; Taye Diggs and Lucy Liu; and Pete Davidson. Netflix

algorithmic determinism

= measurement bias + feedback loop

measurement bias – what you think you are measuring is not what you are actually measuring feedback loop – structure that causes output to eventually influence it’s own input

Conclusions

- Bias means different things to different groups
- Our language of ‘harms’ needs to evolve to embrace algorithmic determinism and the effects of secondary harms

Tools for understanding machine learning

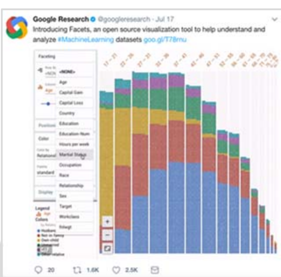
Martin Wattenberg
Senior Staff Research Scientist,
Google

PAIR | People + AI Research Initiative

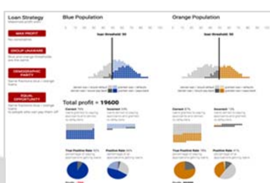
Bringing Design Thinking and HCI to Machine Learning

google.ai/pair

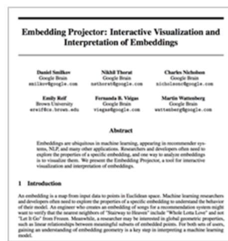
Open Source tools and platforms

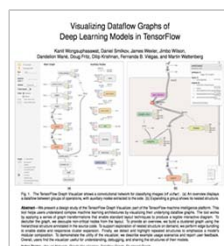


Educational Materials

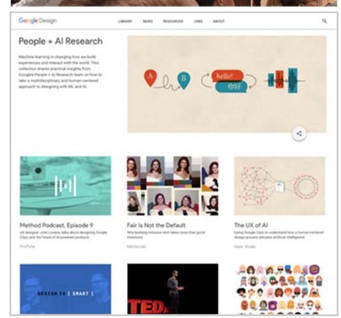


Academic Publications



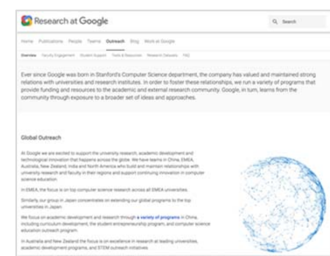


Public presentations, sharing best practices



Public Symposia & meetings

Visiting Faculty, Faculty Grants



Google AI Principles

AI should:

1. be socially beneficial
2. avoid creating or reinforcing unfair bias
3. be built and tested for safety
4. be accountable to people
5. incorporate privacy design principles
6. uphold high standards of scientific excellence
7. be made available for uses that accord with these principles

applications we will not pursue:

1. likely to cause overall harm
2. principal purpose to direct injury
3. surveillance violating internationally accepted norms
4. purpose contravenes international law and human rights

Tools that help humans understand AI

- Understanding of AI is a critical task
  - o Engineering: key for design & debugging
  - o Ethical usage: are systems doing the right thing?

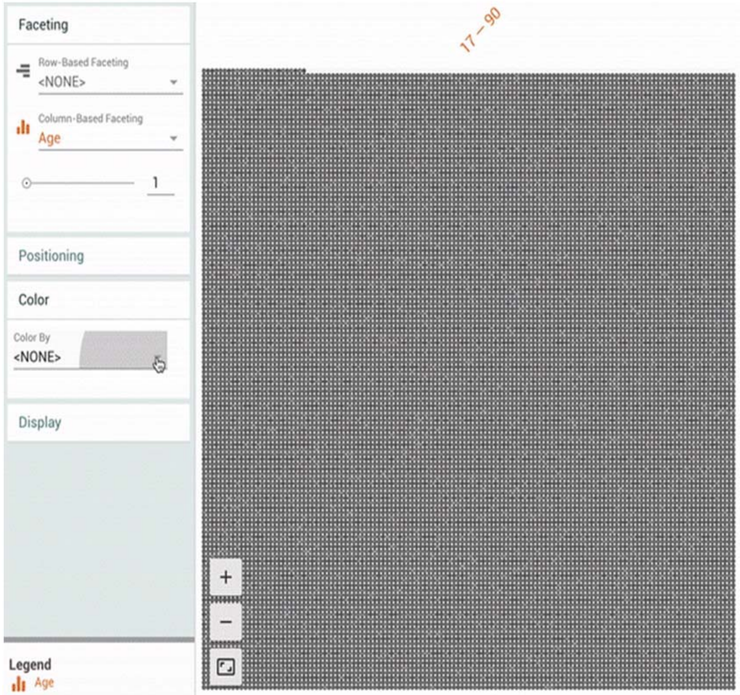
- Machine learning systems: not "black boxes"
  - o They can in some cases be more explainable than human systems
  - o Unlike traditional software, they have explicit goals, or "objective functions"
  - o Biggest problem: there is too *much* data about what they're doing, not too little

- Today's presentation: three tools that Google has created (and open-sourced) to analyze systems machine learning / AI

1. Facets: Visualizing training data

- Machine learning is driven by training data
- To understand a machine learning system, we therefore need to understand the data it's been trained on
- Huge tables of numbers are difficult to analyze - but data visualization is an efficient method of communicating a lot of information at once
- PAIR has open-sourced a visualization tool, "Facets," to help people inspect and analyze their training data.

Facets: debug data, not just programs



Users can group and filter data, visually

Instant, animated responses to user queries

No coding is required; interface can be used by non-programmers, to bring as many stakeholders into the process as possible.

Open source: https://pair-code.github.io/facets/

2. "What-If Tool": Probing an ML model

- Example questions people ask of ML systems:
  - **What if** the system sees data that doesn't look like the training data?
  - **What would happen** if a particular field changed from "true" to "false" on a data point?
- Answering hypotheticals usually means writing code
  - Requires time, money
  - Restricts the set of stakeholders who can easily ask these questions
- PAIR has open-sourced the "What-If Tool," which lets users ask these and many other questions, no coding required.

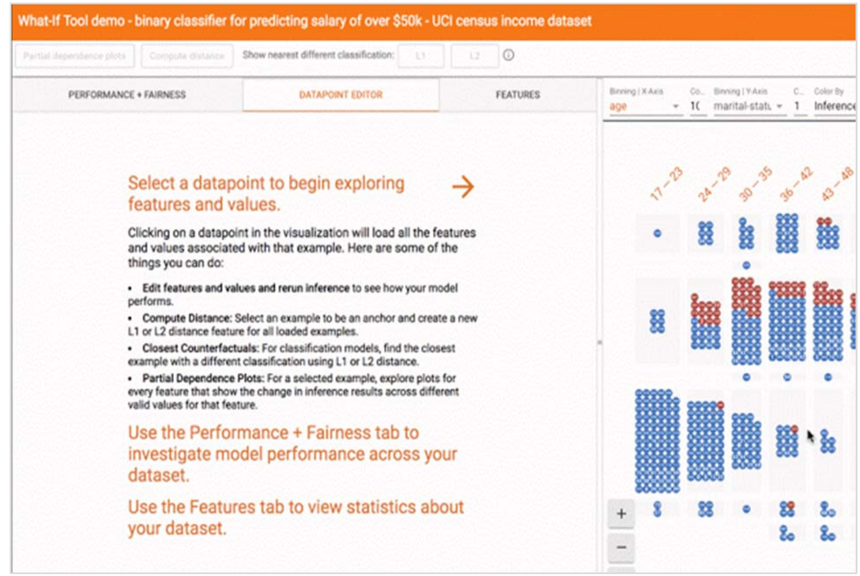
What-If Tool

Edit data point values; instantly see results

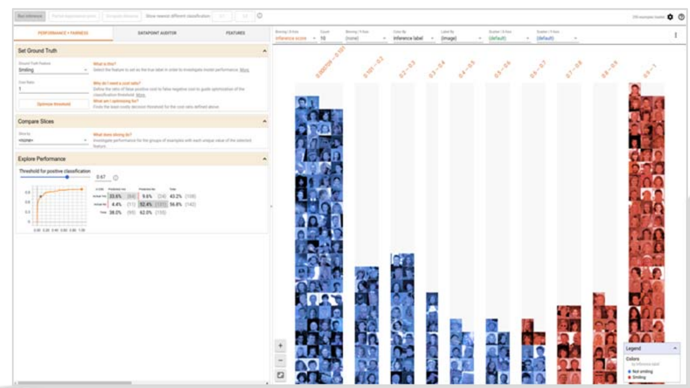
Ask, "what's the nearest point that was classified differently?"

See how strongly different data fields affect results

Global measures of how different groups are treated



pair-code.github.io/what-if-tool



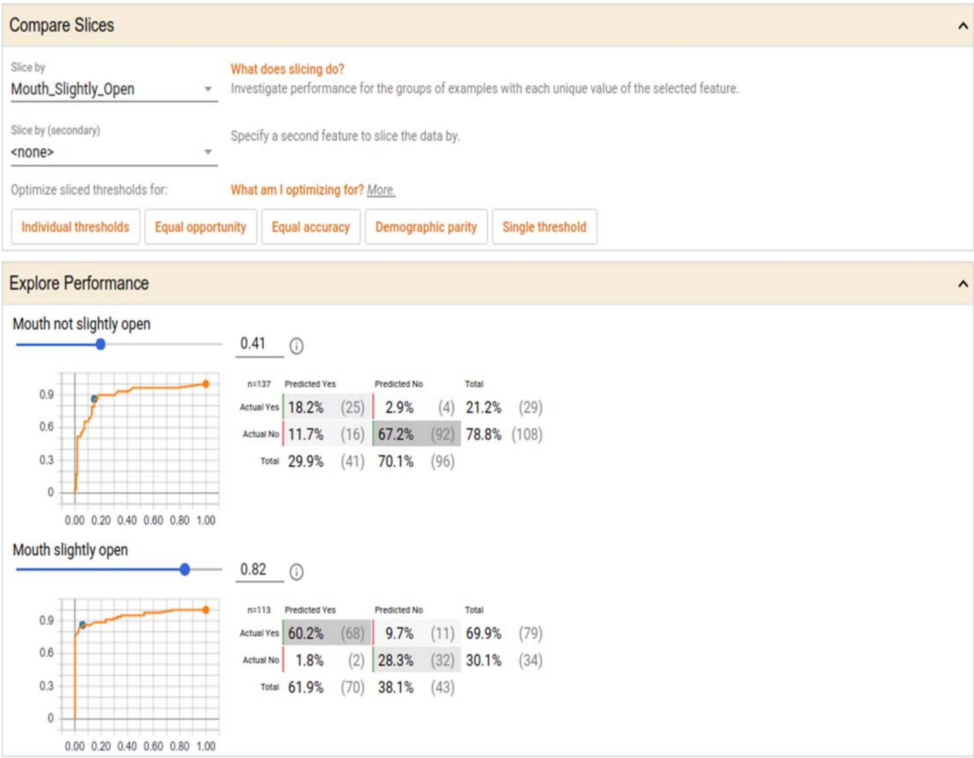
What-If Tool

Fairness metrics

Define groups of interest and calculate "fairness metrics"

Tool suggests threshold changes to achieve different types of equity.

Hardt et al. "Equality of Opportunity in Supervised Learning," NIPS 2016.
See also https://research.google.com/bigpicture/attacking-discrimination-in-ml/



Understanding neural networks

- Neural networks are often called "black boxes"
- But we have vast data on their internals
- PAIR created a new technique, "TCAV," to help people understand this data in human terms
    - Instead of: "Was this photo classified as a zebra because of pixel (17, 255)"
    - Users can ask, "Was this classified as a zebra because of the stripes?"
- You can ask this for **any** high-level concept, as long as you can provide a few dozen examples.

Technical details: see Kim et al., ICML 2018
Open source: https://github.com/tensorflow/tcav

Perspectives on Ethics and Common Principles in AI

Erika Brown Lee
Senior Vice President
Assistant General Counsel
Mastercard

Responsible AI Principles

- Transparency

- Accountability

- Privacy by Design

Transparency

- Build consumer trust and confidence
  - Explainability
  - Disclosures

Accountability

- Practices
- Governance
- Documentation

Privacy by Design

- Minimization
- Data quality
- Anonymization
- Security

AI and NIST Privacy

Naomi Lefkovitz
Senior Privacy Policy Advisor
Research and Standards Development
Information Technology Lab, NIST

NIST Research

- More than 50 projects contemplated or underway in artificial intelligence and machine learning

- Exploring fundamental questions related to measurement and quantification

IEEE Standards Association – AI Policy Standards

- AI-Related Standards Projects:
  - IEEE P7000™ - Model Process for Addressing Ethical Concerns During System Design
  - IEEE P7001™ - Transparency of Autonomous Systems
  - IEEE P7002™ - Data Privacy Process
  - IEEE P7003™ - Algorithmic Bias Considerations
  - IEEE P7004™ - Standard for Child and Student Data Governance
  - IEEE P7005™ - Standard for Transparent Employer Data Governance
  - IEEE P7006™ - Standard for Personal Data Artificial Intelligence (AI) Agent
  - IEEE P7007™ - Ontological Standard for Ethically Driven Robotics and Automation Systems
  - IEEE P7008™ - Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
  - IEEE P7009™ - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
  - IEEE P7010™ - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
  - IEEE P7011™ - Standard for the Process of Identifying and Rating the Trustworthiness of News Sources
  - IEEE P7012™ - Standard for Machine Readable Personal Privacy Terms
  - IEEE P7013™ - Inclusion and Application Standards for Automated Facial Analysis Technology

ISO/IEC - AI Technical Standards

ISO/IEC JTC 1/SC 42 - AI

WG 1 – Foundational standards

WG 2 – Big data

WG 3 – Trustworthiness

WG 4 – Use cases and applications

SG 1 - Computational approaches and characteristics of artificial intelligence systems

JWG with SC40* - AWI 38507 Information technology – Governance of IT - Governance implications of the use of Artificial Intelligence by organizations

** Upon conditional agreement*

ISO/IEC - AI Technical Standards

WG 1 – Foundational standards

- 22989 (IS) – Artificial Intelligence Concepts and Terminology, WD
- 23053 (IS) – Framework for Artificial Intelligence (AI) Systems
  Using Machine Learning (ML), WD
- Consideration of AI Lifecycle

ISO/IEC - AI Technical Standards

WG 2 – Big data

- 20546 (IS) – Big data overview and vocabulary, FDIS
- 20547 – Big data reference architecture
  - o 20547-1 (TR) Part 1: Framework and application, WD
  - o 20547-2 (TR) Part 2: Use cases and derived requirements, Published, April 2018
  - o 20547-3 (IS) Part 3: Reference architecture, DIS
  - o 20547-4 (IS) Part 4: Security and Privacy, CD (under SC 27)
  - o 20547-5 (TR) Part 5: Standards roadmap, Published, April 2018
- Potential new projects include:
  - o Business process management for data analytics
  - o Big data reference architecture interfaces
    - Part 1: Characteristics and capabilities
    - Part 2: Best practices

ISO/IEC - AI Technical Standards

WG 3 – Trustworthiness

- Bias in AI systems and AI aided decision making (TR)
- Overview of Trustworthiness in Artificial Intelligence (TR)
- Assessment of the robustness of neural networks
  Part 1: Overview (TR)
- NP Ballot on Artificial Intelligence – Risk Management (IS)

ISO/IEC - AI Technical Standards

WG 4 – Use cases and applications

- Artificial Intelligence (AI) – Use cases (TR)

SG 1 - Computational approaches and characteristics of artificial intelligence systems

- Study - Computational approaches, processes and methods for applications of AI systems
- Study - Assessment of classification performance for machine learning models

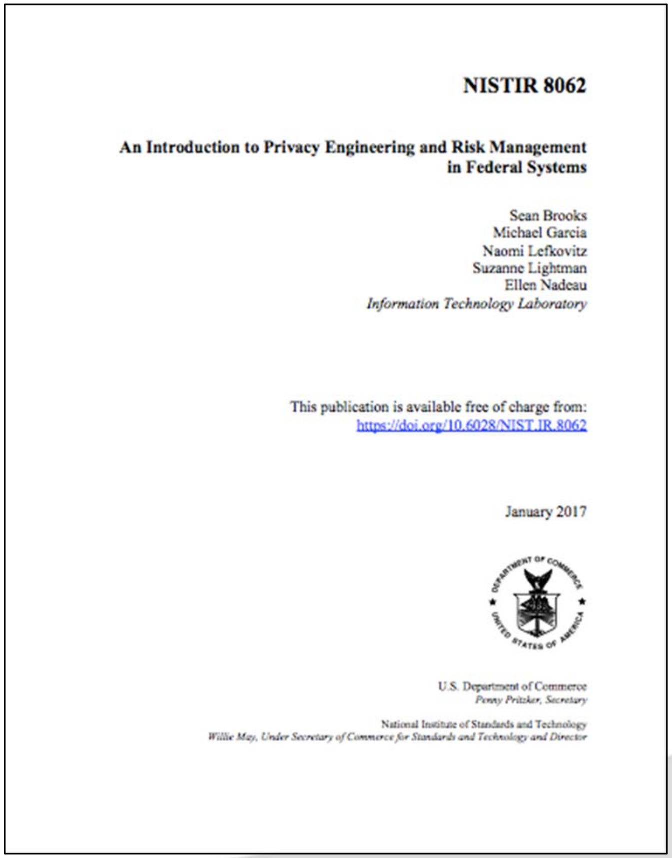
JWG with SC40*

- AWI 38507 Information technology – Governance of IT – Governance implications of the use of Artificial Intelligence by organizations

* *Upon conditional agreement*

NIST Internal Report 8062

An Introduction to Privacy Engineering and Risk Management in Federal Systems



NISTIR 8062

An Introduction to Privacy Engineering and Risk Management in Federal Systems

Sean Brooks
Michael Garcia
Naomi Lefkovitz
Suzanne Lightman
Ellen Nadeau
Information Technology Laboratory

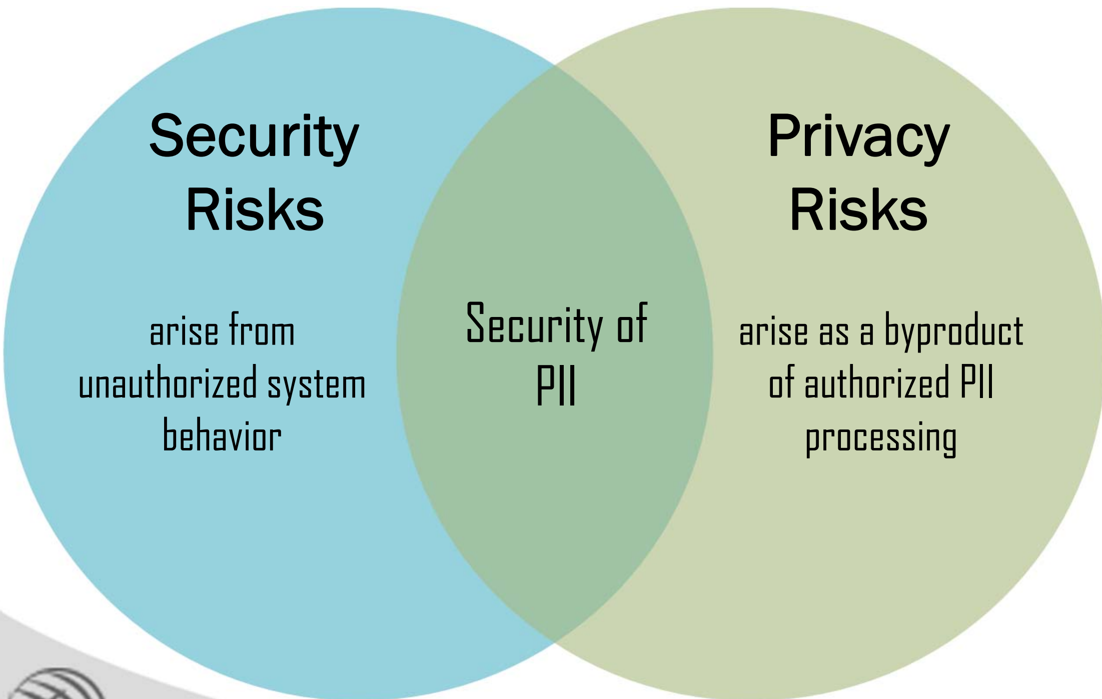
This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8062

January 2017

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

Information Security and Privacy Relationship



There is a clear recognition that security of PII plays an important role in the protection of privacy

Individual privacy cannot be achieved solely by securing PII

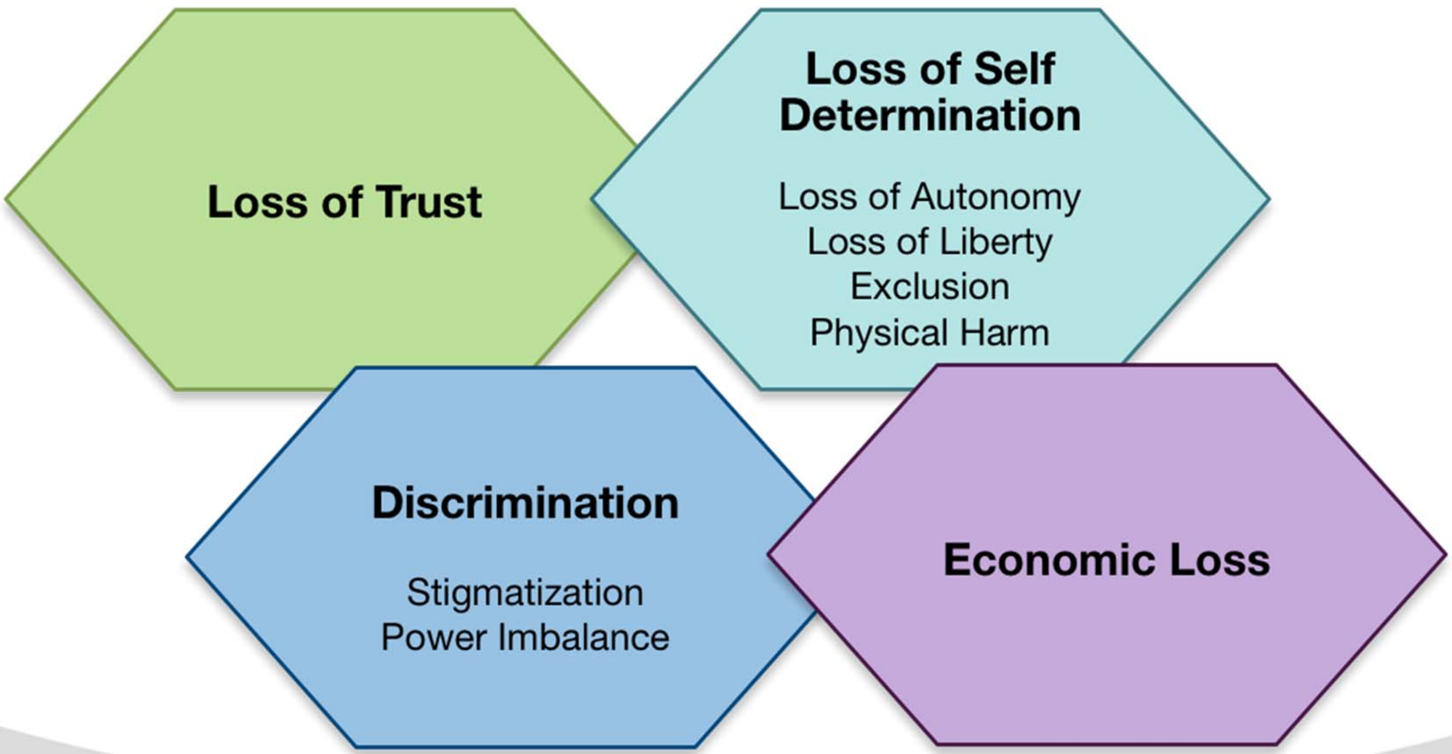
Authorized processing: system operations that handle PII (collection - disposal) to enable the system to achieve mission/business objectives

Security Risk Model

Risk factors:

Likelihood | Vulnerability | Threat | Impact

Processing PII Can Create Problems for Individuals



NIST Working Model for System Privacy Risk

Privacy Risk Factors:

Likelihood | Problematic Data Action | Impact

Likelihood is a contextual analysis that a data action is likely to create a problem for a representative set of individuals

Impact is an analysis of the costs should the problem occur

NIST Privacy Engineering Objectives

System properties that support

Predictability	enabling reliable assumptions by individuals, owners, and operators about PII and its processing by an information system
Manageability	providing the capability for granular administration of PII including alteration, deletion, and selective disclosure
Disassociability	enabling the processing of PII or events without association to individuals or devices beyond the operational requirements of the system

- How can reliable assumptions about AI and data processing be enabled?
- How much manageability of AI and intervention in data processing/behavior is needed?
- How can data be dissociated from individuals or devices while still permitting functionality?

Resources

Naomi Lefkovitz
naomi.lefkovitz@nist.gov

NIST Internal Report 8062
https://doi.org/10.6028/NIST.IR.8062

NIST Privacy Engineering Website
https://www.nist.gov/programs-projects/privacy-engineering

Perspectives on Ethics and Common Principles in Algorithms, Artificial Intelligence, and Predictive Analytics

Panel Discussion:

Erika Brown Lee, Rumman Chowdhury,
James Foulds, Naomi Lefkovitz,
Mark MacCarthy, Martin Wattenberg

Moderators: Karen A. Goldman & James Trilling

Break

3:00-3:15 pm

Consumer Protection Implications of Algorithms, Artificial Intelligence, and Predictive Analytics

Session moderated by:

Tiffany George
Federal Trade Commission
Division of Privacy and Identity Protection

Katherine Worthman
Federal Trade Commission
Division of Financial Practices

Consumer Protection Implications of Algorithms, Artificial Intelligence, and Predictive Analytics

Panel Discussion:

Ryan Calo, Fred H. Cate,
Jeremy Gillula, Irene Liu,
Marianela López-Galdos

Moderators: Tiffany George & Katherine Worthman

Thank You

Join us tomorrow for more on this exciting topic!