National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

March 27, 2025

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

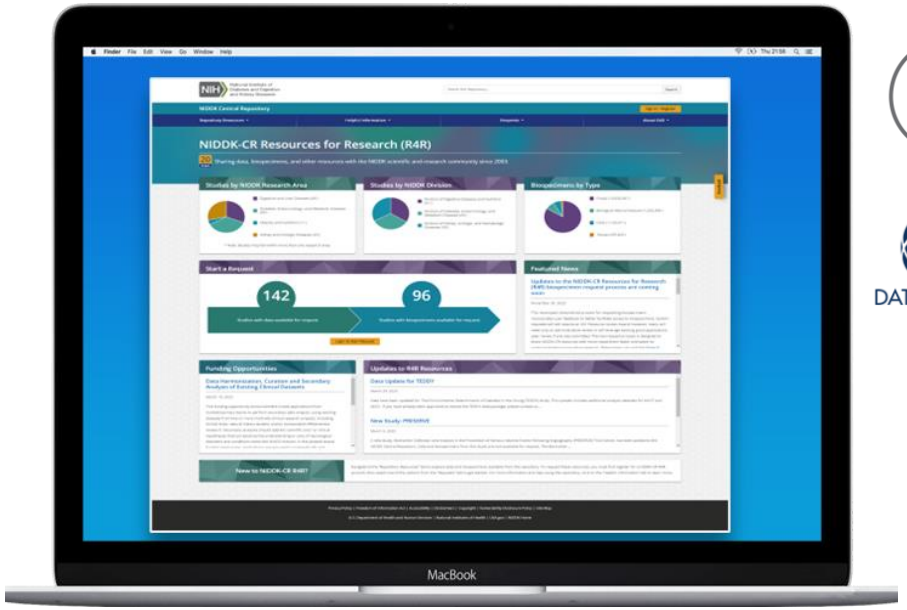
NIDDK Central Repository Overview

Mission

Established in 2003 to **facilitate sharing of data, biospecimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.

- Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles

Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website



CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files
15.8 M

Clinical Datasets
>8,400
from 189 clinical studies

Biospecimens
>16 M

Registered Users
6,827

Weekly Users
>5,000

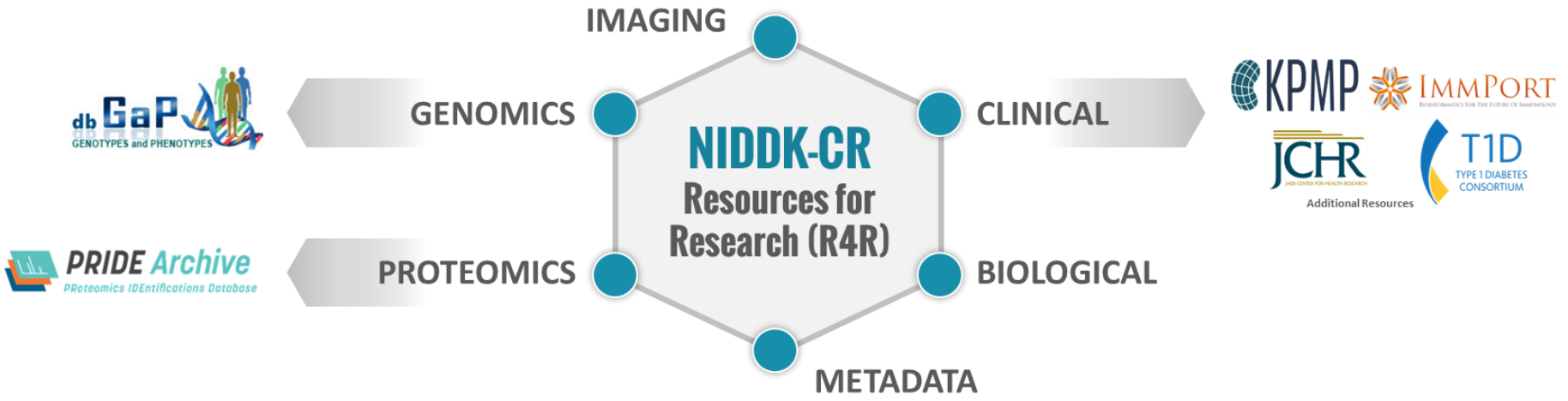
Public Releases
>800

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK Data Sharing Ecosystem

The NIDDK-CR is a part of the broader NIH-funded biomedical data ecosystem and plays a key role in NIH's FAIRness and TRUSTworthiness goals. The NIDDK-CR houses a broad range of data types for secondary research, provides access to biospecimens, and direct links to other repositories with additional resources such as genomics data.



FAIRsharing.org standards, databases, policies

DataCite FIND, ACCESS, AND REUSE DATA

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

DK net

Google Dataset Search / Schema.org

NIH U.S. National Library of Medicine ClinicalTrials.gov

Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA

PLANNING PHASE

figshare

NIH HEAL INITIATIVE

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Future Functionality: Analytics Workbench

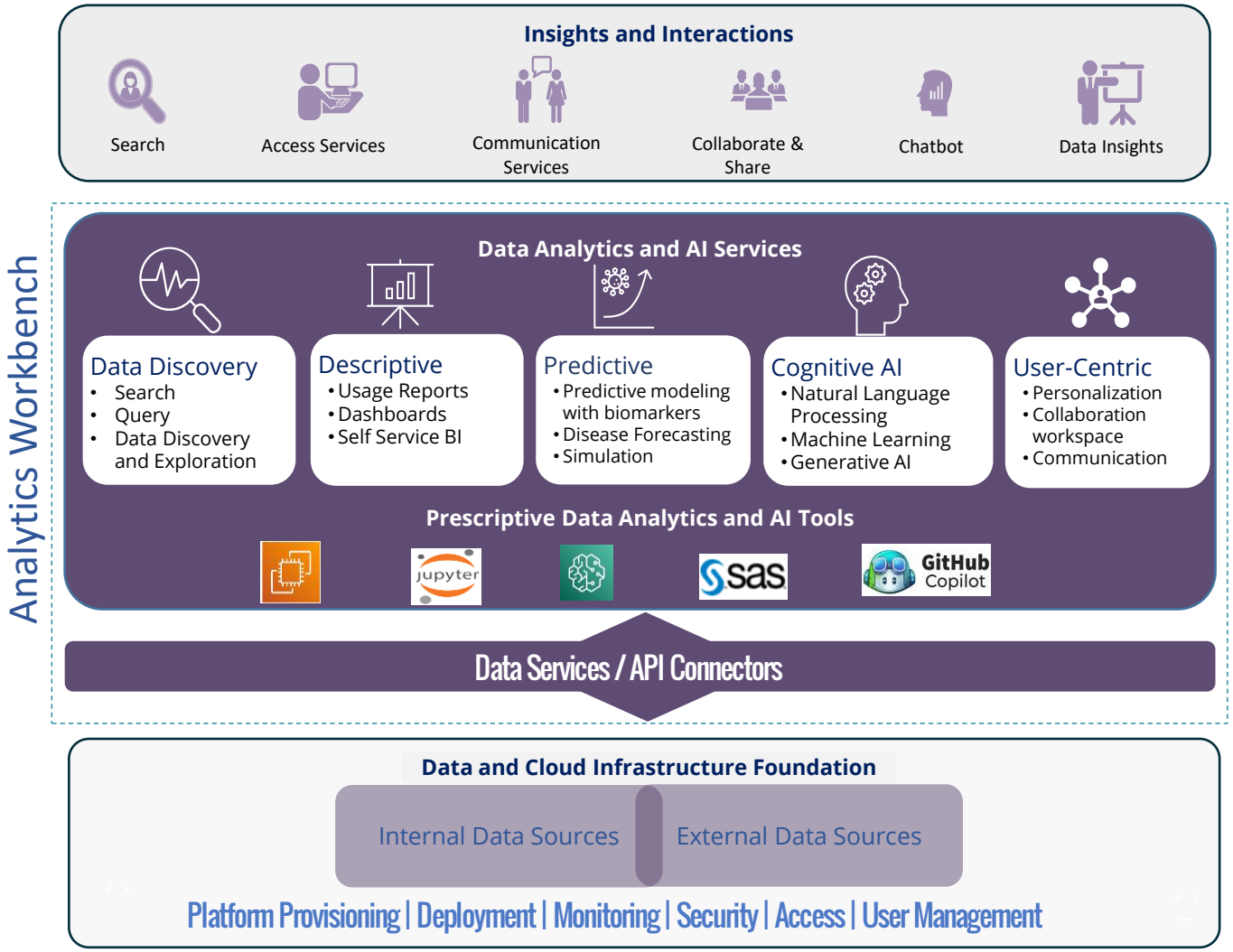
Streamlining *end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

Innovation and ease of use

A cloud-based analytics environment where researchers and data scientists can access a suite of integrated analytics tools and cloud computing resources to participate in data challenges and AI innovation.

Expected Benefits of Analytics Workbench:

- **Promote Collaboration**
- **Support AI Innovation**
- **Minimize Data Movement**
- **Improve User Experience**
- **Discover Data Insights**
- **Advance NIDDK Research Mission**



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications

Visit our website for more information on our data-centric movement and to learn more about our past data-challenges

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Secondary Data Science and Meet the Expert Webinar Series

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month**

Upcoming Webinars

- Today – Artificial Intelligence fundamentals
- April 24 – Artificial Intelligence fundamentals applications
- May 29 – FAIR data sharing
- June 26 – Different privacy preserving techniques and implications for researchers
- July 31 – Challenges, opportunities, and considerations for secondary researchers using electronic health records and real-world data sources
- August 28 – Impact and innovations realized

Learn more about the webinar series, register for future webinars, and access past webinars materials and recordings

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Gordon Aiello is a Lead Scientist at Booz Allen Hamilton with a PhD in Applied Mathematical and Computational Sciences. He works full-time developing and delivering specialized data science, artificial intelligence, machine learning, and Python trainings for clients in the Navy and Intelligence Community. Prior to joining Booz Allen Hamilton, Dr. Aiello worked in the Office of Macroeconomic Affairs at the U.S. Department of State, using machine learning techniques to analyze developing and emerging market economies. Additionally, he has taught courses on data science and the R programming language for the Foundation for Advanced Education in the Sciences (FAES) at the NIH. He is passionate about working with others to expand their understanding of data science techniques and their applications.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI Fundamentals

NIDDK-CR Data Science

Meet the Experts Webinar Series

March 27th, 2025

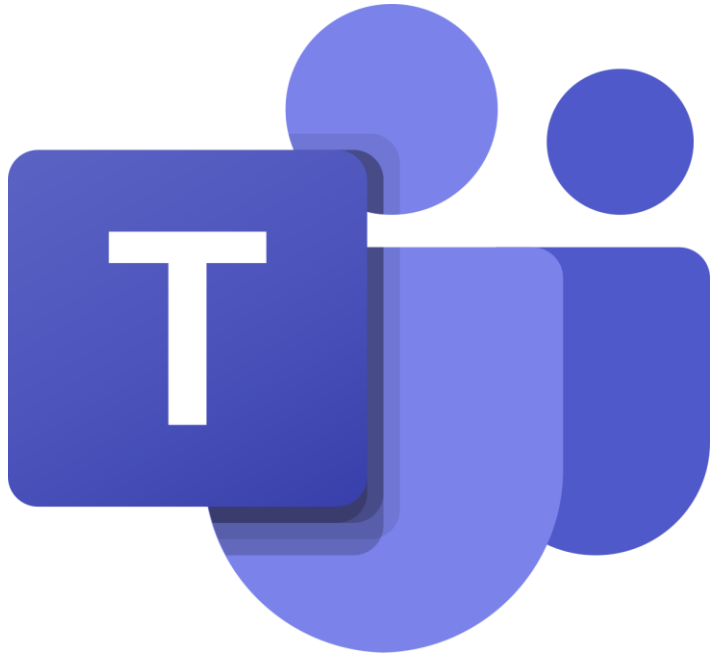
Presented by: Booz Allen Hamilton

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Training Guidance

- Avoid CUI/PII/PHI conversations
- Questions in Teams Chat are encouraged
- Due to size of class, stay on mute until end of class



Data Science Learning Program

If you're new to data science, start your learning journey with the **Foundations** courses. A more in-depth learning track starts with the **Data Science Fundamentals** course and continues to the **Data Science Labs**. Those interested in more specialized topics can explore courses in the **Select Topics** track.

INTRODUCTION

Foundations for Data Citizens

- Data Citizen best practices
- Data governance
- Data-driven organization

udemy

Foundations of Data Analytics

- For NAVWAR supervisors
- Data Science Overview
- Machine Learning and Artificial Intelligence

udemy

THEORY-TO-PRACTICE

Data Science Fundamentals

- Comprehensive intro to Data Science
- Python programming
- Statistics, Probability and Linear Algebra refresher
- Machine learning and Artificial Intelligence

Live Training

udemy

10.5 hours (3 sessions)

Data Science Project Lab*

- Theory-to-practice
- Case study format
- Hands-on exercises
- Tabular data cleansing and processing techniques
- Full-cycle analytics process

Live Training JUPITER

12 hours (3 sessions)

Data Science NLP Lab*

- Theory-to-practice
- Case study format
- Hands-on exercises
- Natural Language Processing Techniques
- Large Language Models

Live Training JUPITER

12 hours (3 sessions)

*Completion of the Introduction to Python course is recommended for those without programming experience.

SELECT TOPICS

Introduction to Data Visualization

- Telling a story with your data
- How to create more impactful briefings
- Not product specific

Live Training

udemy

3 hours

Python Fundamentals for Data Science

- Foundational Python syntax
- Develop essential analytic skills
- Machine Learning and Artificial Intelligence

Live Training JUPITER

7 hours (2 sessions)

Artificial Intelligence Fundamentals

- AI initiatives and foundational AI
- AI ecosystems and AI operations
- Responsible and Ethical AI
- Neural Networks

Live Training

7 hours (2 sessions)

Data Science for Managers

Developed in partnership with NGA

- Management responsibilities in Data Science Projects
- Ethical considerations in Data Science
- Data Science and AI Opportunities

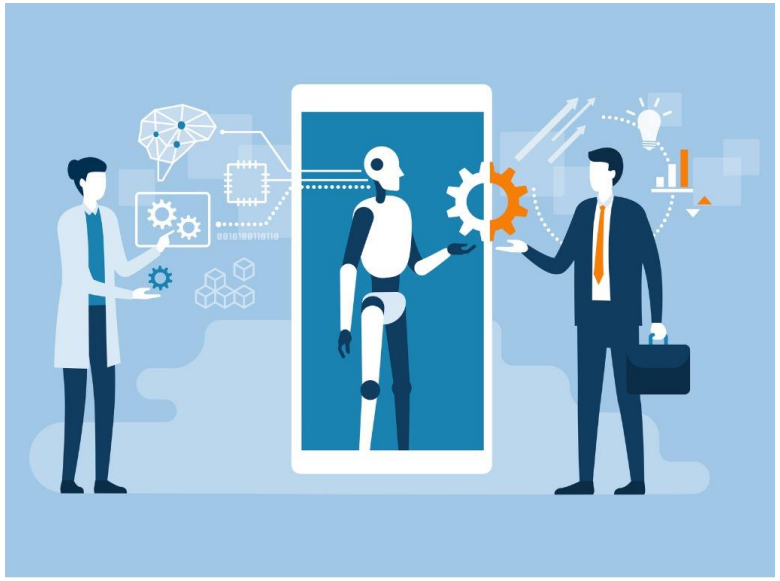
In Person Training

8 hours

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Agenda



1. AI Demystified
2. Responsible AI
3. AI Operations
4. AI Implementations

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI Demystified

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Defined

The goal of data science is to extract meaningful insights from data.

Data – any kind of qualitative or quantitative set of values

- Common examples in data science today:
  - Natural text: "I'm cold," "I'm not very cold"
  - Categories: "yellow," "green," "red"
  - Numbers: 1, 2.53, -4
  - Images:  
- Sometimes you have the data, sometimes you need to procure the data

Science – a systematic approach to building knowledge by testing hypotheses

- Think Scientific Method:

  Define a hypothesis → Collect the data → Analyze results → Draw conclusions
- Hypotheses must be testable, and experiments must be reproducible

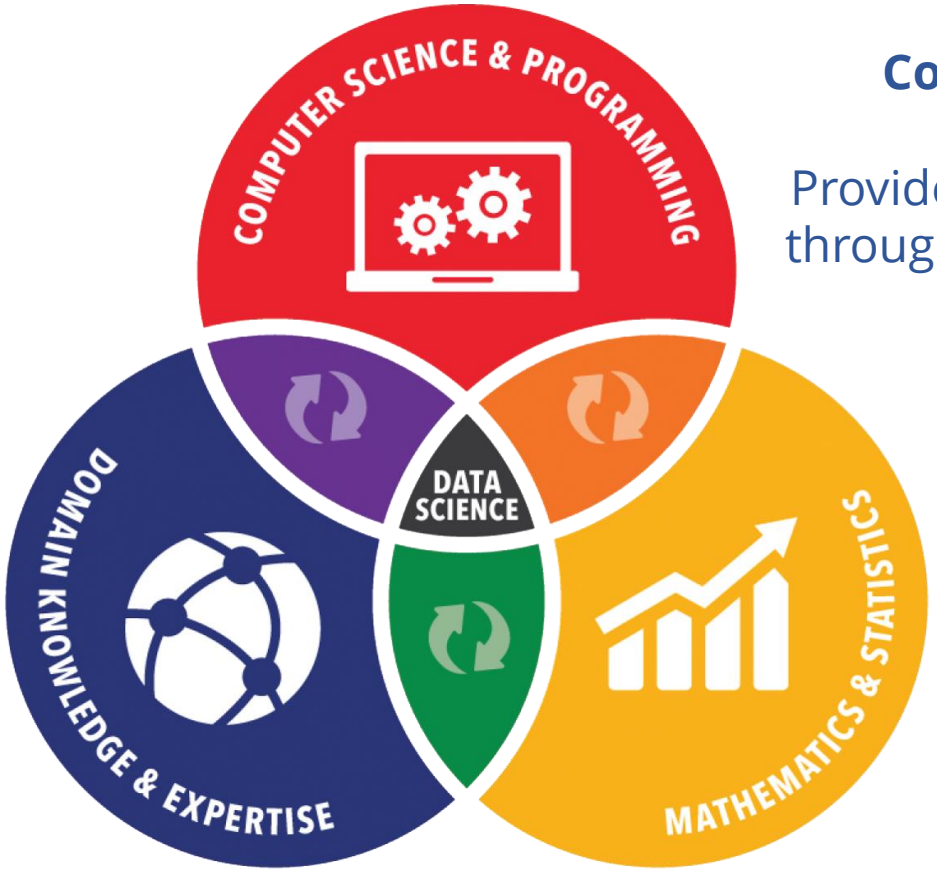
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Is an Interdisciplinary Field

Data science lies at the intersection of mathematics and statistics, computer science and programming, and domain knowledge and expertise.

Computer Science and Programming
Provides the technical resources through which data products are created



Domain Knowledge and Expertise
Provides context and specialized understanding of the application in which the problem exists

Mathematics and Statistics
Provides the theoretical tools and structures used to analyze data science problems

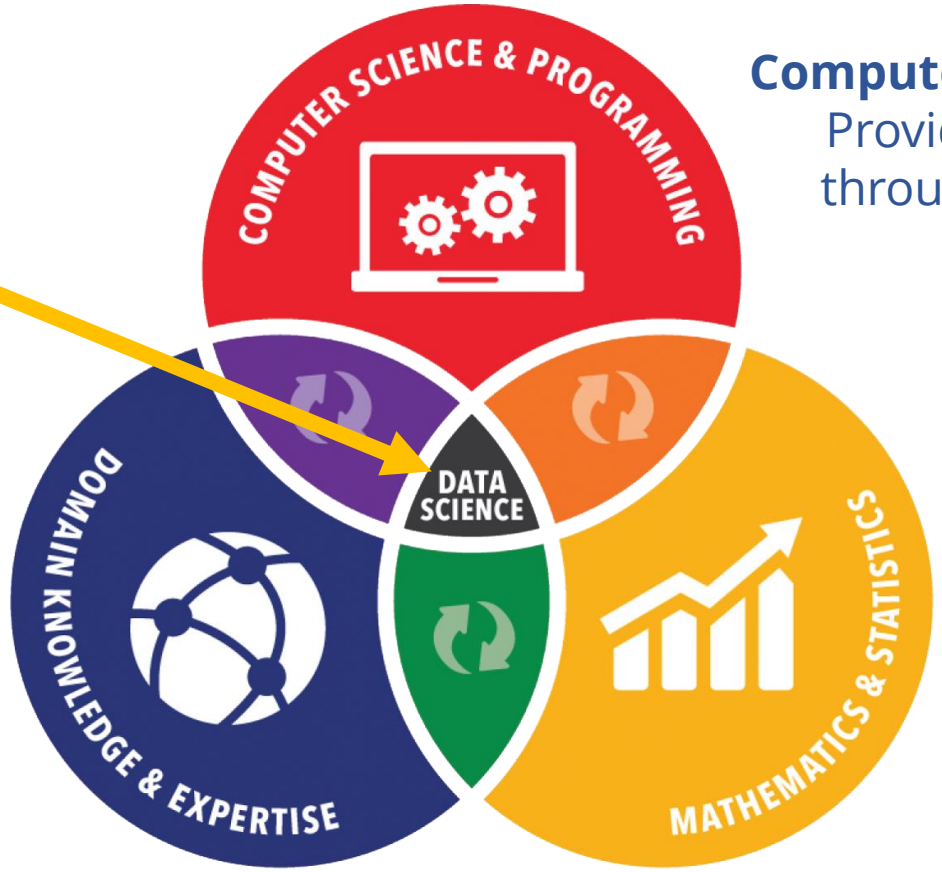
Source: Booz Allen Hamilton

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Is an Interdisciplinary Field

Data science lies at the intersection of mathematics and statistics, computer science and programming, and domain knowledge and expertise.



Computer Science and Programming
Provides the technical resources through which data products are created

Domain Knowledge and Expertise
Provides context and specialized understanding of the application in which the problem exists

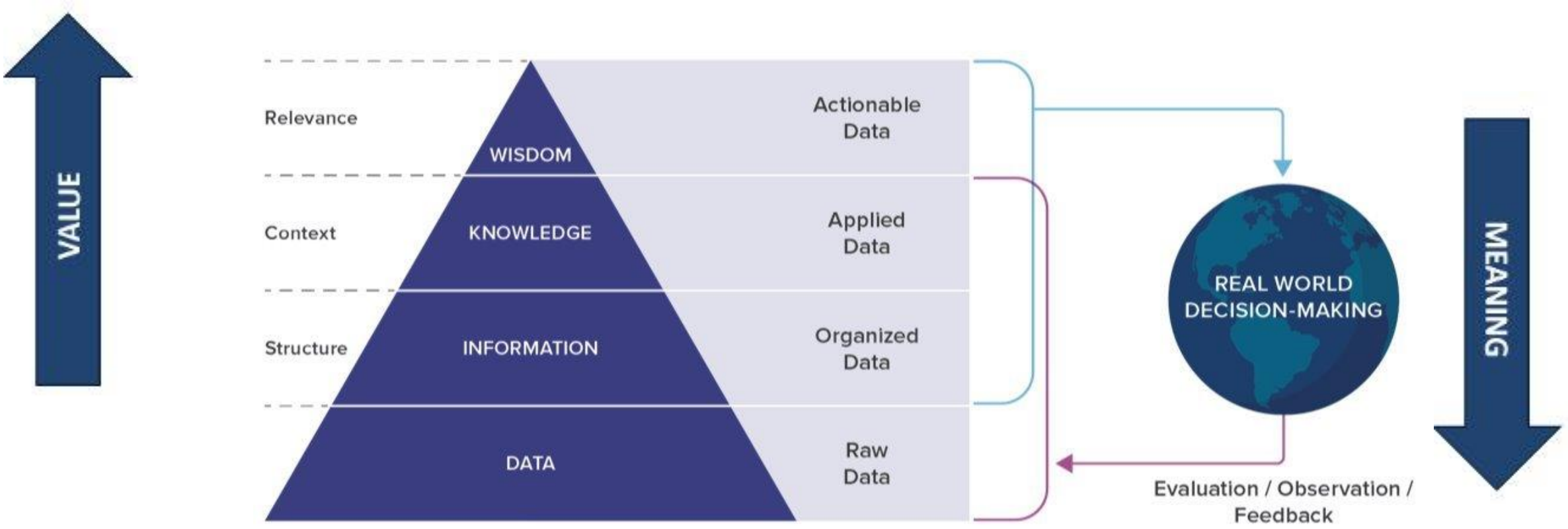
Mathematics and Statistics
Provides the theoretical tools and structures used to analyze data science problems

Source: Booz Allen Hamilton

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data-Information-Knowledge-Wisdom Model



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

How DS Intersects with AI & ML

The various subfields of data science intersect with one another:

Artificial Intelligence:

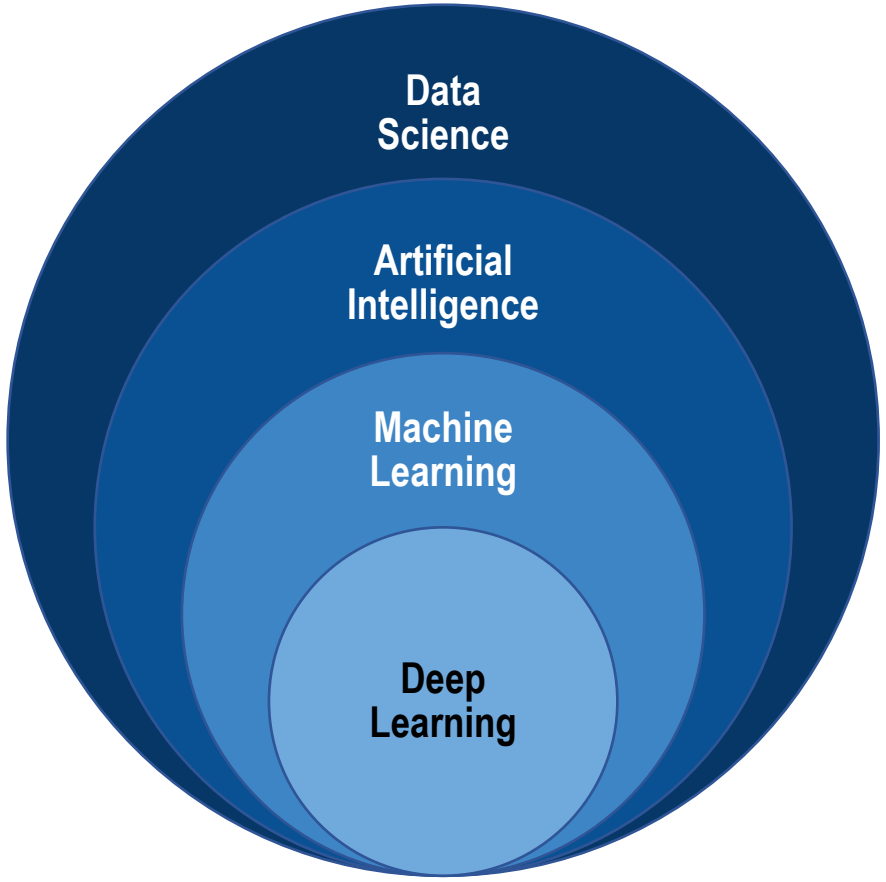
- The theory and development of computer systems able to perform tasks normally requiring human intelligence

Machine Learning:

- A field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed

Deep Learning:

- Is part of a broad family of machine learning methods based on learning data representations, as opposed to task-specific algorithms (predominately based on artificial neural networks)




REPOSITORY
TECHNOLOGY ADVANCEMENT
BIOSTATISTICS
CLINICAL STUDIES
REGULATORY
Office of Clinical Research Support
NIH
National Institute of Diabetes and Digestive and Kidney Diseases

Why Now?

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Why Now? Big Data

- Businesses, governments, and societies are only starting to tap data's vast potential
- Calls for individuals to “strengthen data and statistics for accountability and decision-making purposes”

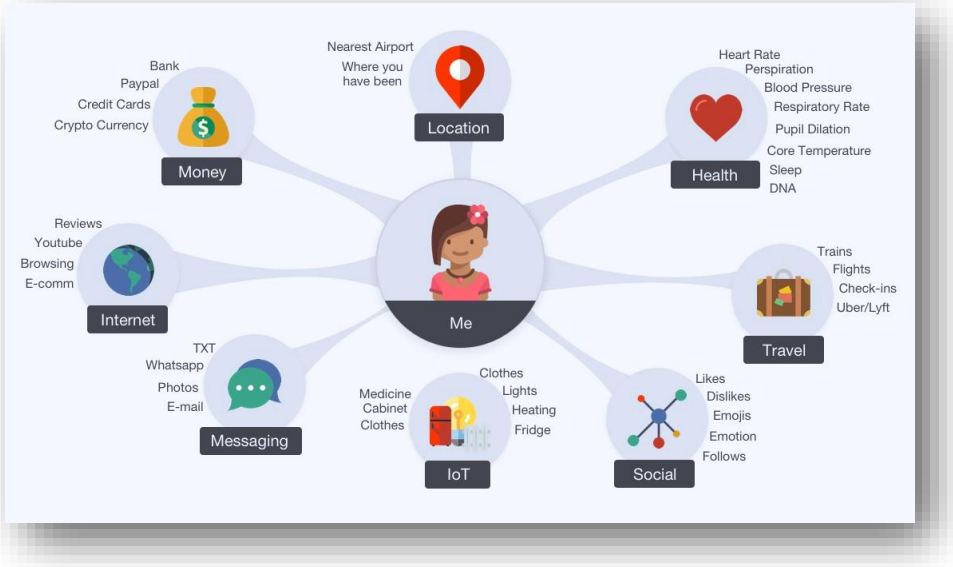
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Why Now? Data Is Universal

- Data has replaced oil as the world's most valuable resource, according to the Economist
- In 2000, 5% of the world was connected to the internet
- In 2024, 67.5% of the world is connected to the internet, causing massive increases in the volume of data being produced daily
- In 2024, approximately 400 quintillion* bytes of data are created every day

*400 quintillion = 400,000,000,000,000,000,000

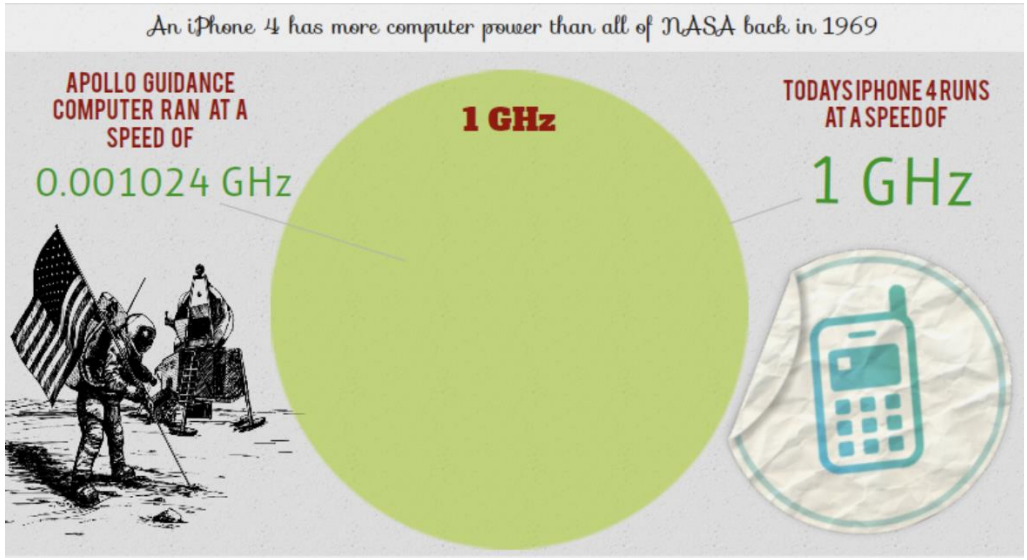


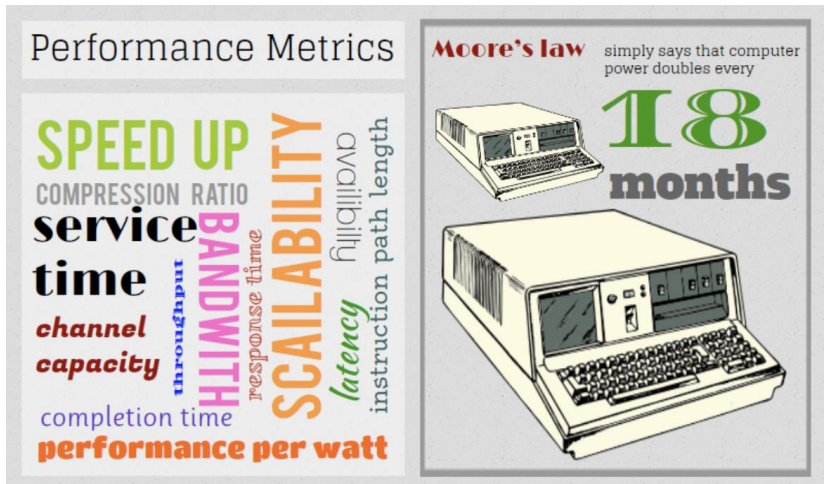
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Why Now? Computer Capabilities

Computer capabilities have greatly improved over the years, allowing us to perform computations that would take longer than the age of the universe on previous models.

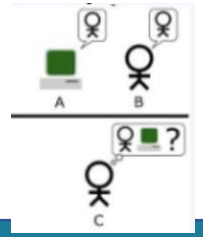




National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Brief History of AI & Its Rapid Ascension

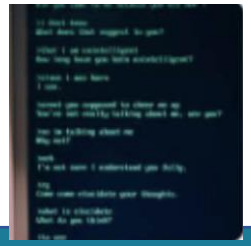


1950

- Alan Turing famous for breaking the Nazi's ENIGMA code during World War II, publishes Computing Machinery and Intelligence. In the paper Turing proposes to answer the question "can machines think?" and introduces the Turing Test to determine if a computer can demonstrate the same intelligence as a human

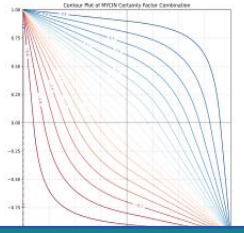
1956

- John McCarthy coins the term "Artificial Intelligence" at the first ever AI conference at Dartmouth College.



1964-1966

- Joseph Weizenbaum creates ELIZA, an early natural language processing computer, at the MIT AI laboratory to demonstrate the superficiality of communication between humans and machines.



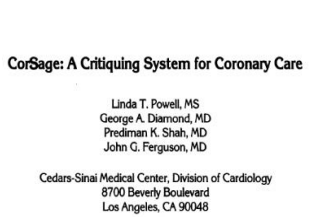
1970s

- AI System MYCIN creates suggested antibiotic treatments for potential bacterial pathogens.

National Institute of Diabetes and Digestive and Kidney Diseases

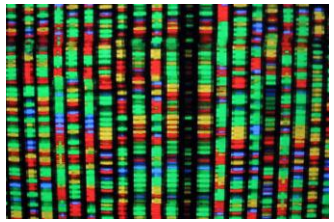
Central Repository

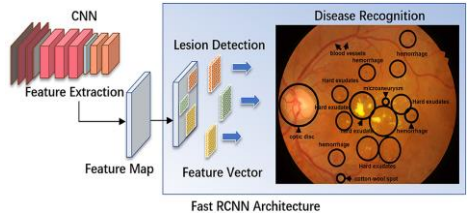
Brief History of AI & Its Rapid Ascension


CorSage: A Critiquing System for Coronary Care

Linda T. Powell, MS
George A. Diamond, MD
Prediman K. Shah, MD
John G. Ferguson, MD

Cedars-Sinai Medical Center, Division of Cardiology
8700 Beverly Boulevard
Los Angeles, CA 90048





Fast RCNN Architecture

1986

- UMass releases DXplain, which uses inputted symptoms to generate diagnoses for 500 diseases.

1989

- CorSage, a clinical tool that combines AI and statistical techniques to help physicians identify heart patients who are most likely to suffer another coronary event, is debuted.

2003

- The Human Genome Project provides a wealth of data on the genetic basis of disease.

2017

- Deep-learning applications screen for diseases ranging from diabetic retinopathy to skin cancer with astonishing accuracy. The FDA approves the first AI-powered device for operating-room use.

2020

- Google DeepMind uses AI to predict a protein's 3D structure from its amino-acid sequence, solving one of biology's greatest challenges.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Limitations of Data Science & AI

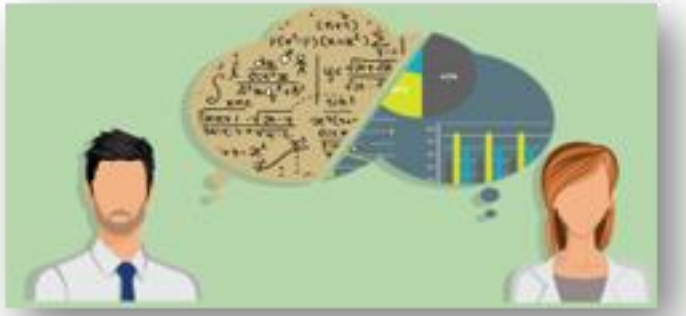
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Myths

Myth 1:
It's as simple as pushing a button

Myth 2:
Data science will make analysts obsolete



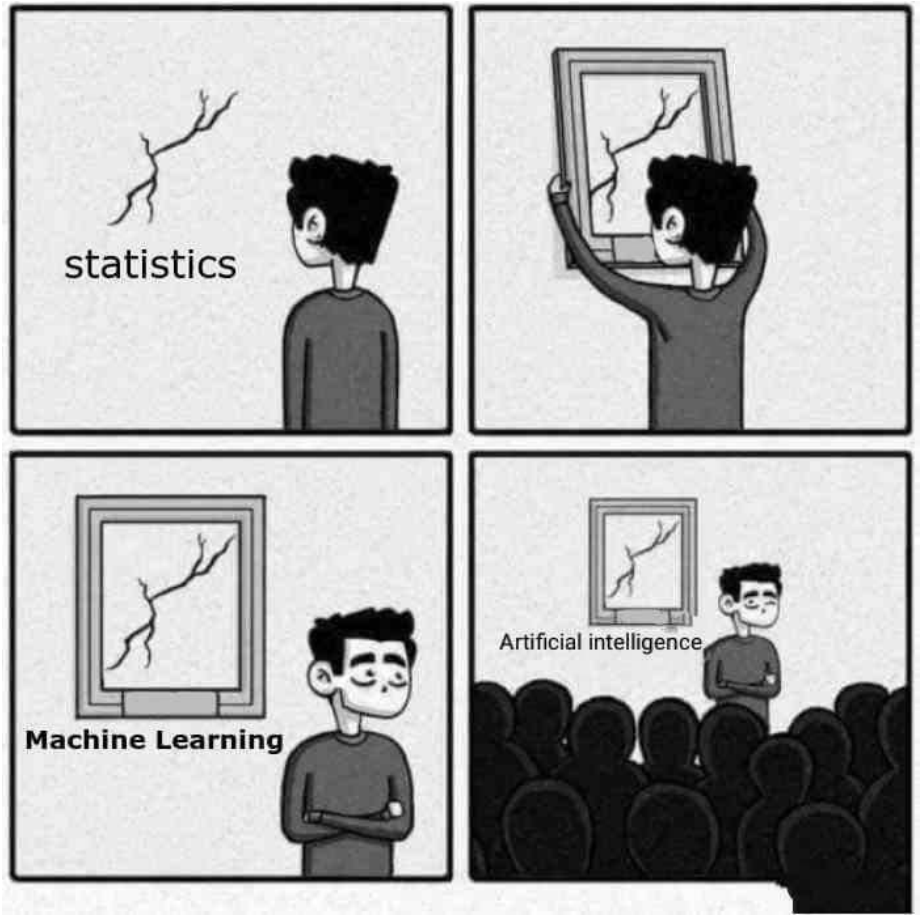
Myth 3:
It's too soon to talk

Myth 4:
Everybody needs to learn coding

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI Is Statistics Over Larger Datasets



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

What Is AI?

AI is a field concerned with producing machines able to autonomously perform tasks that would normally require human intelligence by giving them the ability to perceive, learn from, abstract (classify, conceptualize, and generate rules), and act using data.

What is AI?

- Interchangeable with the term machine intelligence or "MI"
- Able to perform certain narrowly defined tasks as well or better than humans
- A substitute for human intelligence in certain tasks within jobs

What isn't AI?

- Interchangeable with the term "data science"
- Able to perform wide-ranging tasks as well as or better than humans
- A substitute for any human's entire job

Tasks Well Suited to AI

- Simple, requires little to no context
- Involves finding patterns in data
- Characterized by large amounts of data preferably labeled
- Occurs in a statis environment or one with little uncertainty

Tasks Poorly Suited to AI

- Complex, requires contextual understanding
- Requires explaining patterns in data
- Little or no data to characterize the problem
- Occurs in a dynamic environment with a lot of uncertainty

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

How Is AI Used in Medical Care and Biomedical Research?

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI Uses in Healthcare

- **Radiology**: The ability of AI to interpret imaging results may aid in detecting a minute change in an image that a clinician might accidentally miss.
- **Early Detection of Kidney Disease**: AI can help in early detection by analyzing patterns in patient data (e.g., blood tests, medical history, and imaging scans). For instance, AI systems are used to predict the onset of chronic kidney disease (CKD) by analyzing blood markers like creatinine levels or urine albumin-to-creatinine ratio.
- **Diabetes Prediction and Risk Assessment**: AI can analyze patient data, including family history, lifestyle factors, and clinical records, to predict the risk of developing Type 2 diabetes. By identifying high-risk individuals, early interventions like lifestyle changes or medications can be recommended.
- **Continuous Glucose Monitoring (CGM) and AI-Driven Insulin Dosing**: For individuals with Type 1 diabetes, continuous glucose monitors (CGM) provide real-time glucose level data. AI models can predict glucose fluctuations and automatically adjust insulin doses to maintain optimal glucose levels.

How Machines Learn

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

How Machines Learn

- We can get an intuition for how machines learn by experimenting with an interactive applet demonstrated on the upcoming slides.

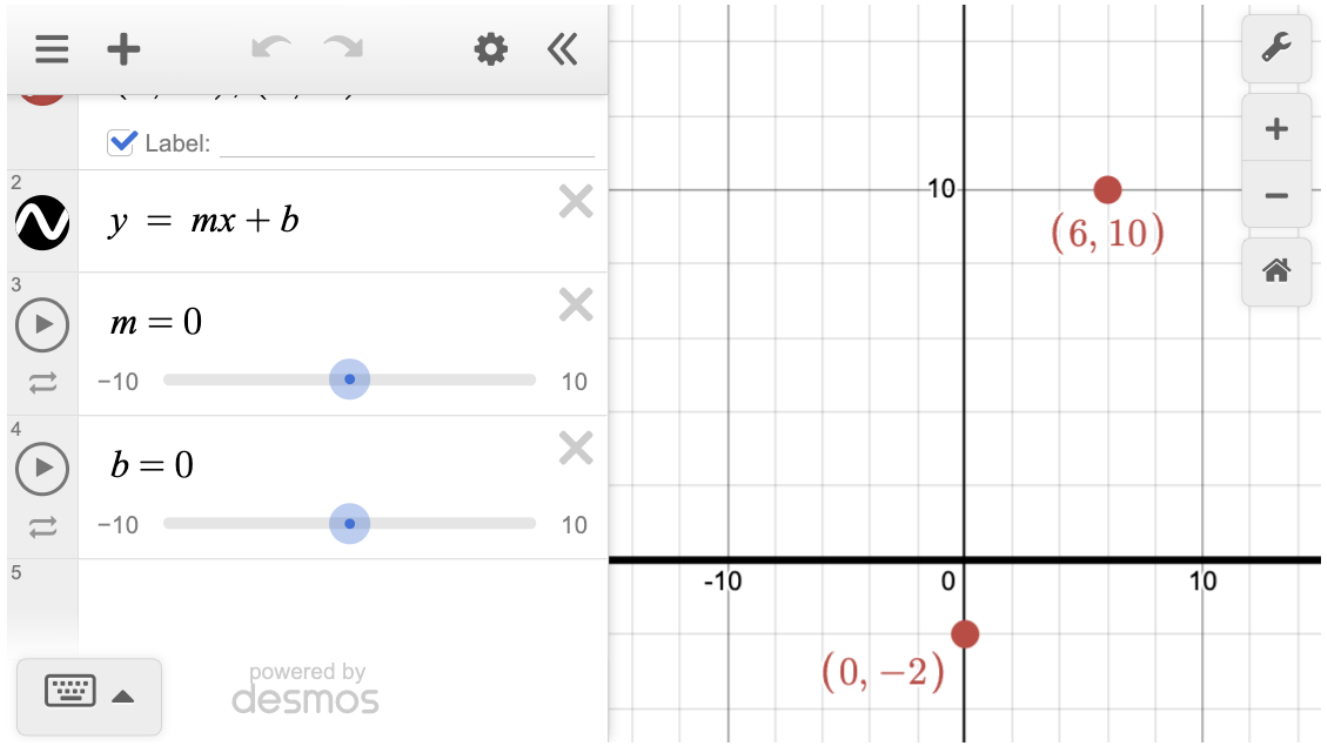
- A helpful analogy for thinking of this process is to relate machine learning with baking, as illustrated in the correspondence table below.

Machine Learning	Baking
Data – Human Provided	Ingredients to be mixed together (e.g., flour, sugar, butter, cream, etc.)
Desired Model Output – Human Provided	Tasty treat (e.g., cake, cookie, biscuit, etc.)
Model Parameters – Computer Determines These	Quantities of ingredients used in recipe (e.g., 3 cups sugar, 4 tbsp butter, etc.)

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

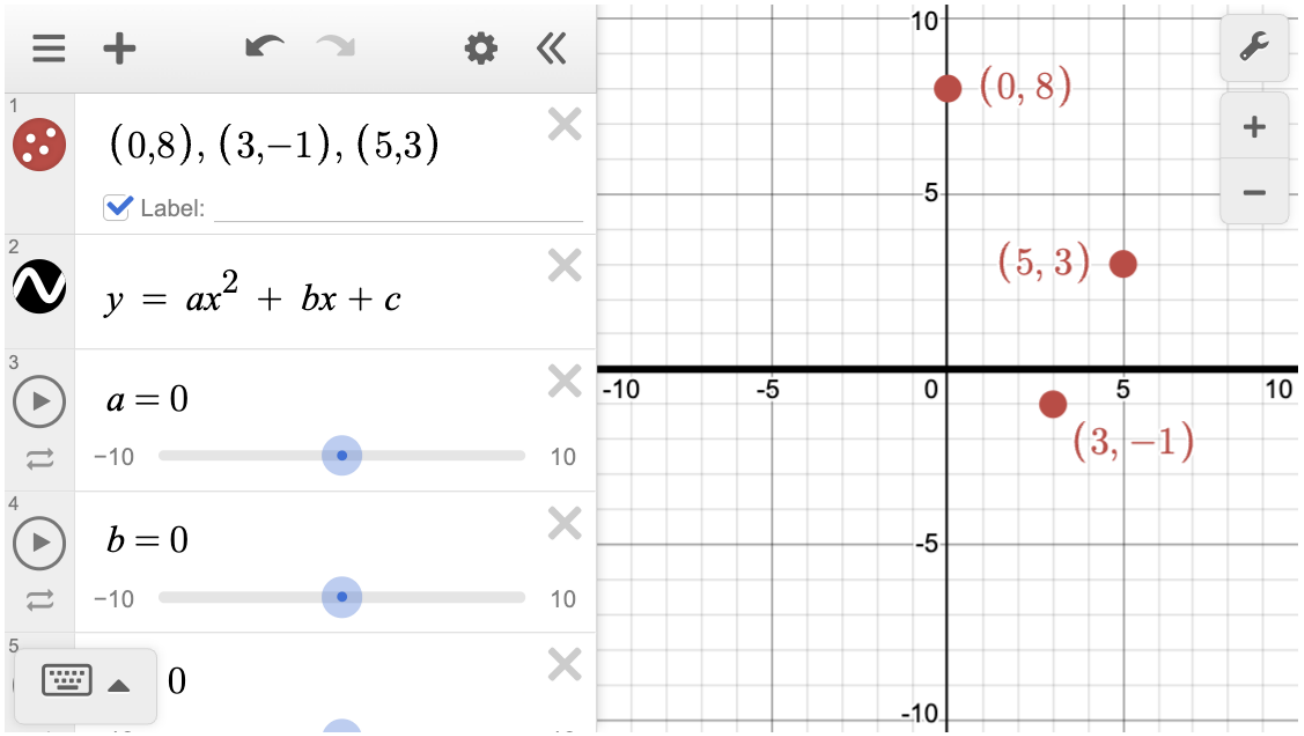
Fitting a Linear Model



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

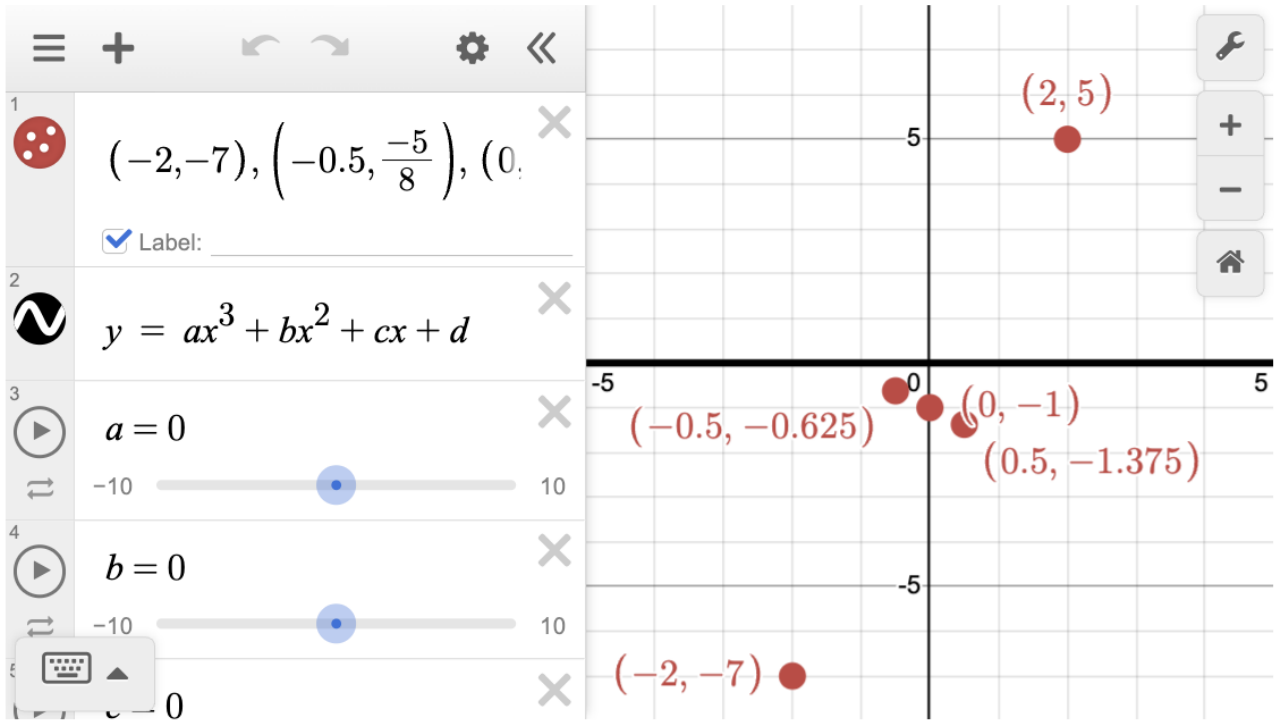
Fitting a Quadratic Model



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

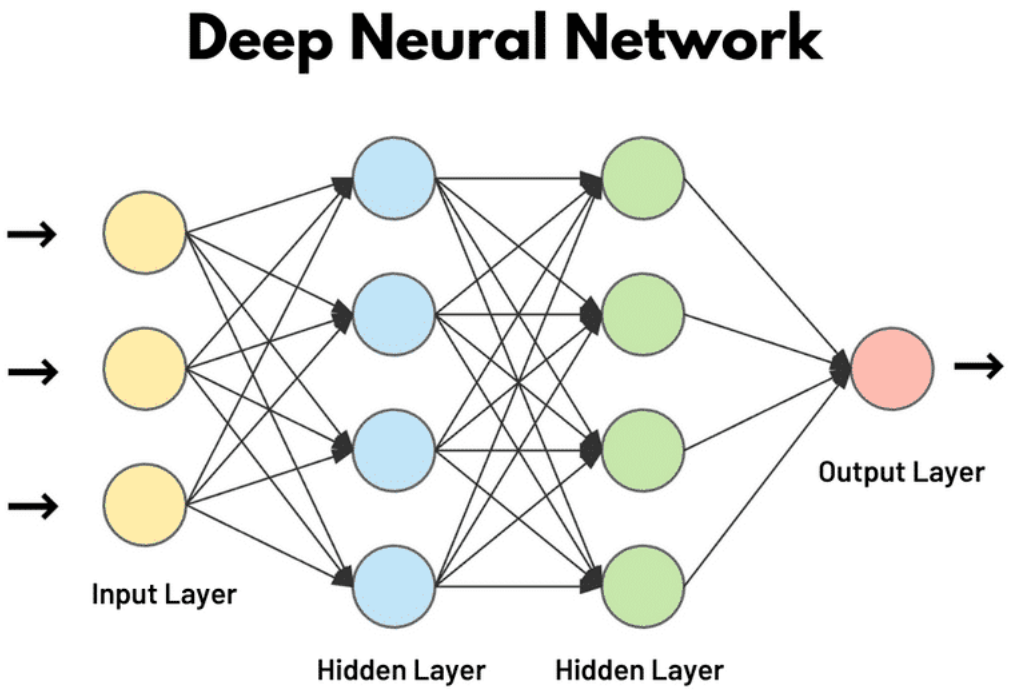
Fitting a Cubic Model



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Deep Learning

- **Neural Networks**: A machine learning approach modeled on the human brain in which algorithms process signals via interconnected layers of artificial neurons.
    - Mimicking biological nervous systems, artificial neural networks have been used to recognize and predict patterns of neural signals involved in brain function, especially in differentiating typographic characters and in recognizing facial features.
- **Deep Learning**: A form of machine learning that uses many layers of computation to form a more complex neural network — often called a deep neural network — that is capable of learning from large amounts of complex, unstructured data.
    - Deep neural networks enable such devices as voice-controlled virtual assistants; self-driving vehicles, which learn to recognize traffic signs; and computer-vision systems used by doctors to more quickly and accurately assess X-rays and other medical images.



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

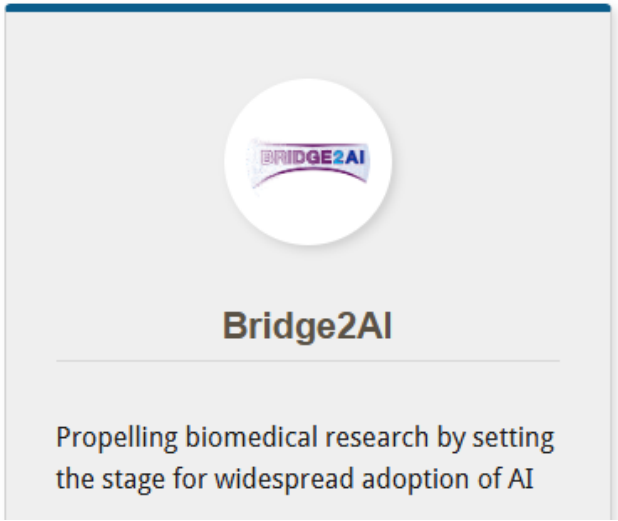
Responsible AI

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI at the NIH

- The NIH makes a wealth of biomedical data and AI opportunities available to research communities and aims to make these data findable, accessible, interoperable, and reusable—or FAIR.



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIH AI Policy Guidance

Protection of Human Subjects (45 CFR 46): Outlines basic provisions for Institutional Review Boards, informed consent, and assurance of compliance for NIH-supported research involving human participants and their data, including considerations of risks & benefits.

Establishes the requirement to submit a DMS Plan and comply with NIH-approved plans. In addition, NIH Institutes, Centers, and Offices can request additional or specific information be included within the plan to support programmatic priorities.

Research Participant Protections

Data Management and Sharing

Health Insurance Portability and Accountability Act (HIPAA) helps protect the privacy and security of health data used in research, including research involving AI, thereby fostering trust in healthcare research activities.

Health Information Privacy

Peer Review

NOT-OD-23-149: Informs the extramural community that the NIH prohibits NIH scientific peer reviewers from using natural language processors, large language models, or other generative AI technologies for analyzing and formulating peer review critiques for grant applications and R&D contract proposals.

Biosecurity and Biosafety

Research funded by NIH, including research using the tools and technologies enabled or informed by AI, fall under this oversight framework.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

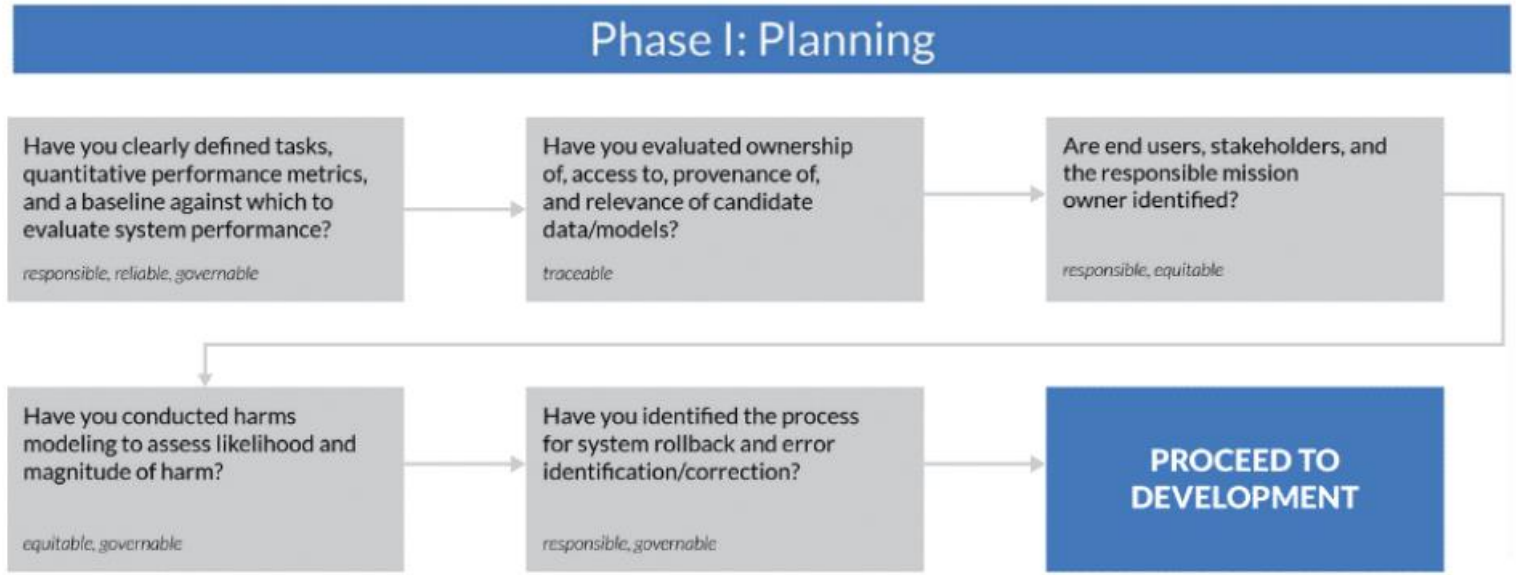
NIH AI Policy Resources

- Use of Generative AI in Peer Review FAQs (NIH Office of Extramural Research)
- NIH Office of Data Science Strategy
- US Department of Health and Human Services Artificial Intelligence Use Cases Inventory
- Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence
- PCAST Report to the President – Supercharging Research: Harnessing Artificial Intelligence to Meet Global Challenges
- NIH STRIDES Initiative | NIH STRIDES

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

RAI Lifecycle: Planning



The planning phase is the process to map out, identify the objective, and define the functionality of the AI system to solve a given problem.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

RAI Lifecycle: Planning

Phase I: Planning

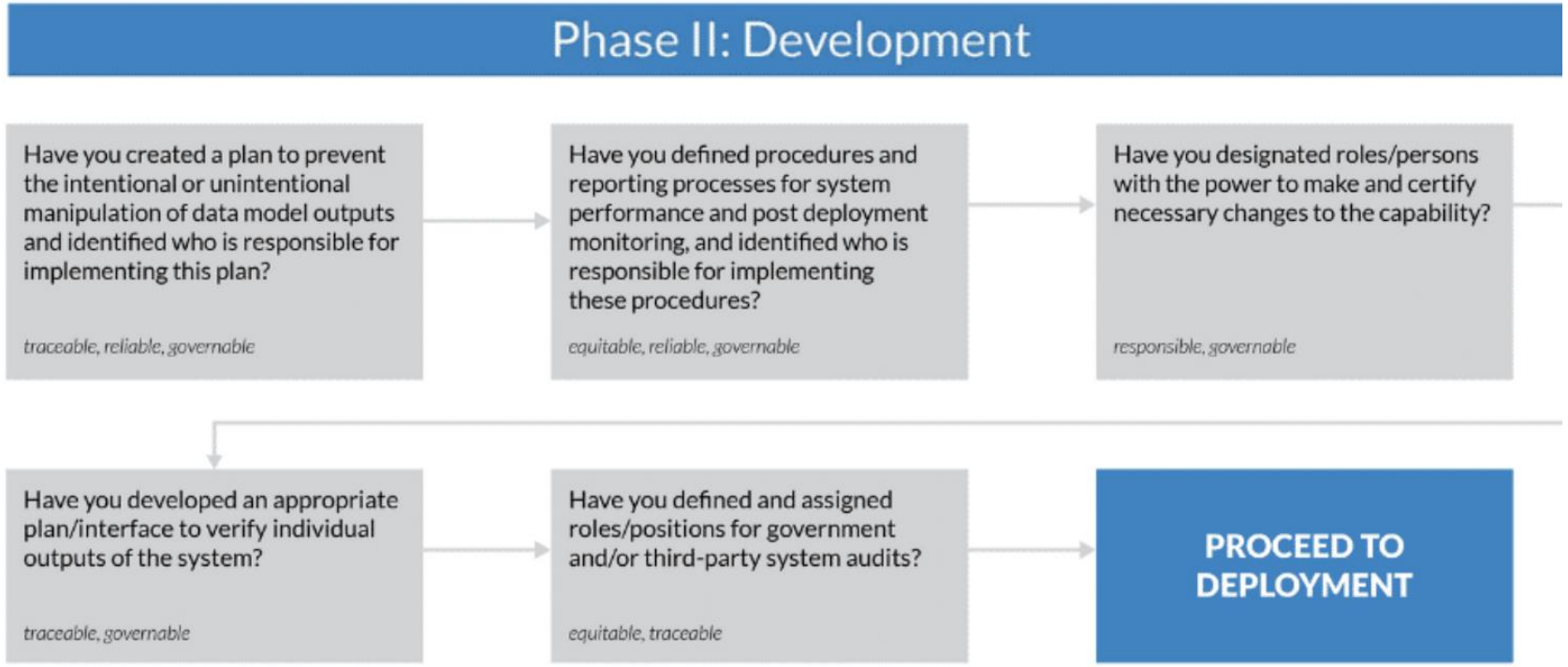
Questions to answer throughout the Planning phase are:

- Have you evaluated other methods of implementing a similar objective?
- What do the success metrics look like?
- Do you have the right and appropriate data for the problem?
- What are the resources required to create the solution?
- Is ownership of the AI solution clearly defined?
- Who are the stakeholders and are they informed of the process and solutions?
- Are the risks identified and the process laid out for a system malfunction?

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

RAI Lifecycle: Development



In the development phase your team will iterate on the process to build the planned-out AI system through data modeling, output verification, and data governance.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

RAI Lifecycle: Development

Phase II: Development

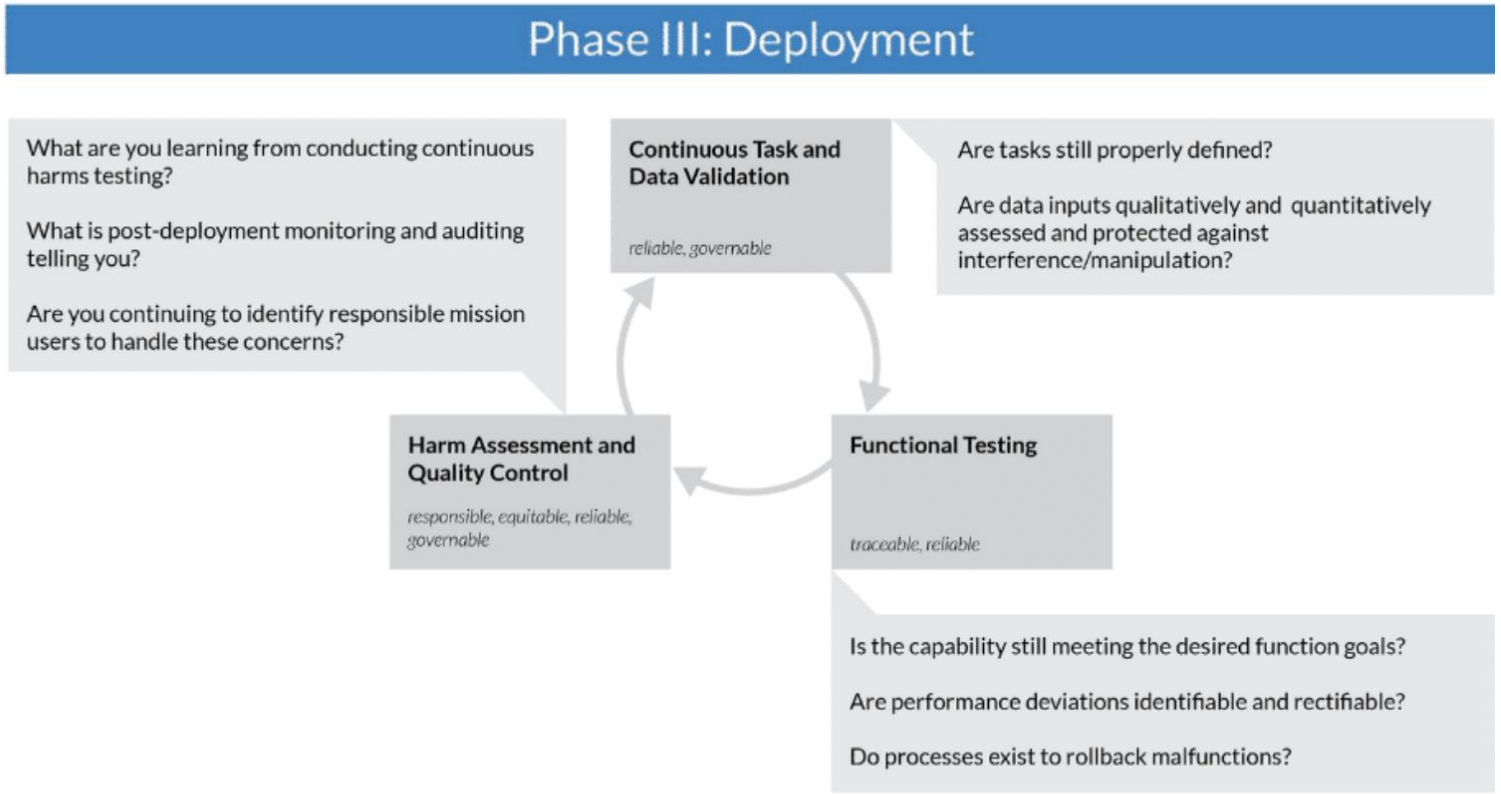
Questions to answer throughout the Development phase are:

- What are the impacts of data or model biases and how can you mitigate them?
- What are the metrics needed for post deployment monitoring?
- Who is the owner for maintaining the AI system?
- Are the AI system and outputs explainable?
- What is the process for routine maintenance?

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

RAI Lifecycle: Deployment



In the Deployment phase, the AI system is put to use in an operational setting to solve the problem initially posed.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

RAI Lifecycle: Deployment

Phase III: Deployment

Questions to answer throughout the Deployment phase are:

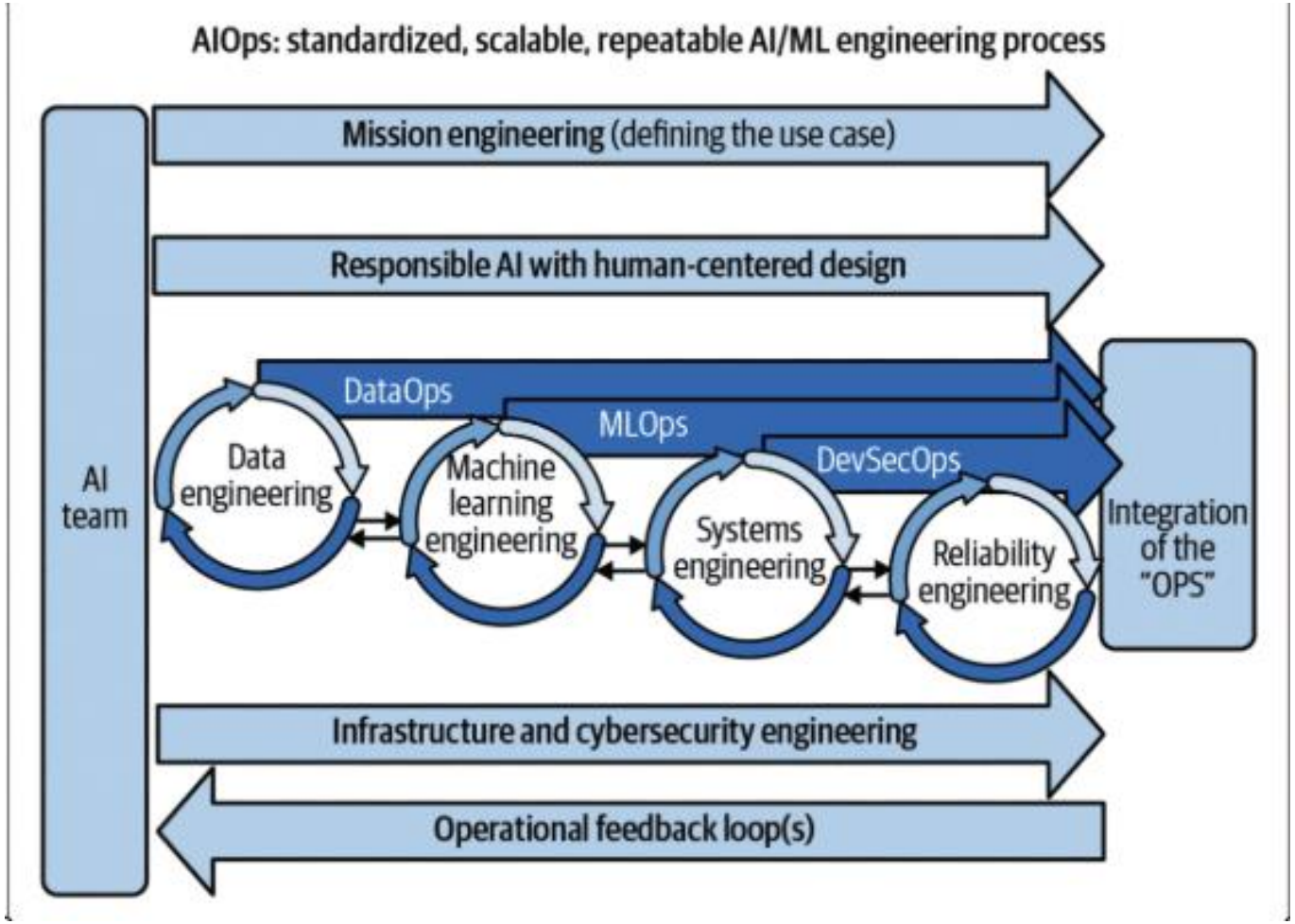
- Is there a process in place to continuously monitor and validate the outputs from the AI system?
- Does the capability still meet the needs of the desired solution and tasking?
- Are the potential biases and negative impacts of the outputs considered and mitigated?

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI Operations

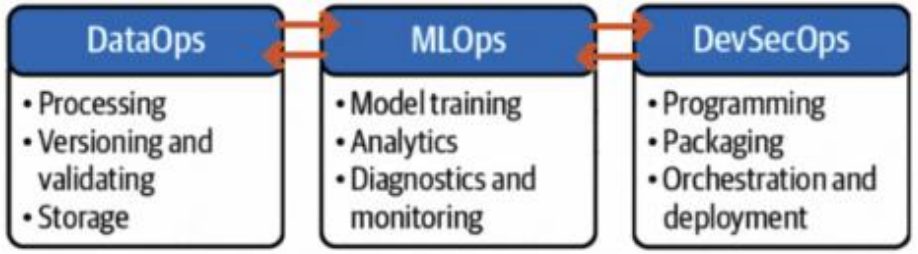
AI Ops


AIOps: standardized, scalable, repeatable AI/ML engineering process
Mission engineering (defining the use case)
Responsible AI with human-centered design
DataOps
MLOps
DevSecOps
AI team
Data engineering
Machine learning engineering
Systems engineering
Reliability engineering
Integration of the "OPS"
Infrastructure and cybersecurity engineering
Operational feedback loop(s)

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AIOps Processes – Data Ops



Data Ops – a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and data consumers across an organization

Data readiness can be measured through several key drivers:

- **Accuracy:** The degree to which data correctly describes the "real world" object or event being described.
- **Usability**: A measure of data's reliability. (Data that is messy, is missing substantial records, is difficult to ingest, or is otherwise prohibitive would be less usable.)
- **Provenance:** A record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.
- **Security:** How safe the data is from loss or unwanted manipulation.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

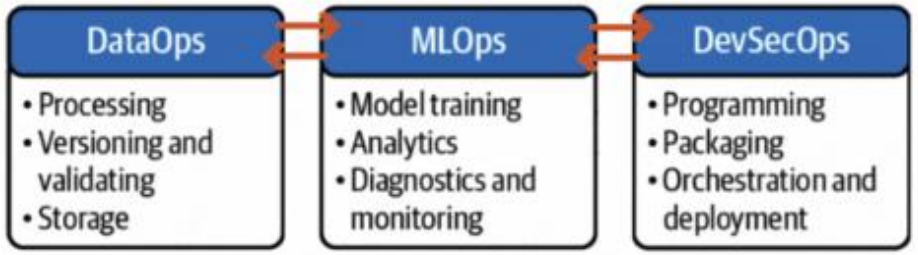
Data Governance Priorities

- Creating policies and procedures tailored to support data needs for AI at enterprise scale.
- Addressing data acquisition policies and procedures to include legal processes, agreements, and data brokers/data services.
- Empowering key stakeholders to facilitate and lead your organization's data governance.
- Developing data security plans and data privacy policies in accordance with appropriate regulatory frameworks (e.g., Health Insurance Portability and Accountability Act [HIPAA], Federal Information Security Modernization Act [FISMA], European Union's General Data Protection Regulation [GDPR], etc.).
- Prioritizing metadata management to ensure compliance and troubleshooting support.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Machine Learning Operations



Machine Learning Ops – focused primarily on the governance and life cycle management of a wide range of operationalized artificial intelligence (AI) and decision models

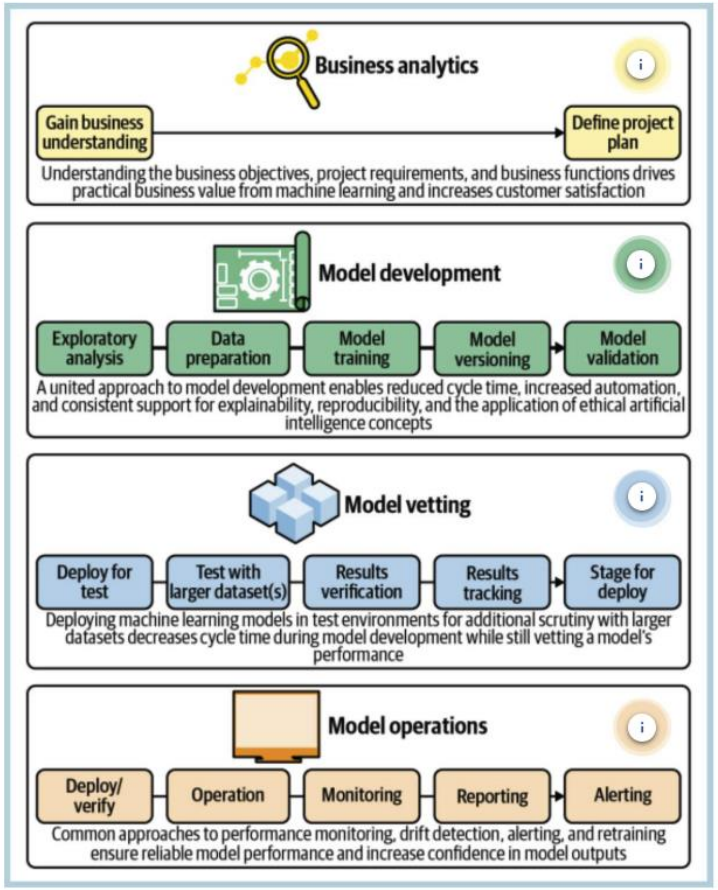
MLOps governs the development and sustainment of an AI solution's analytic backbone throughout its operational lifecycle. Not all ML solutions have the same requirements in terms of:

- Required data
- Development techniques
- Deployment considerations.

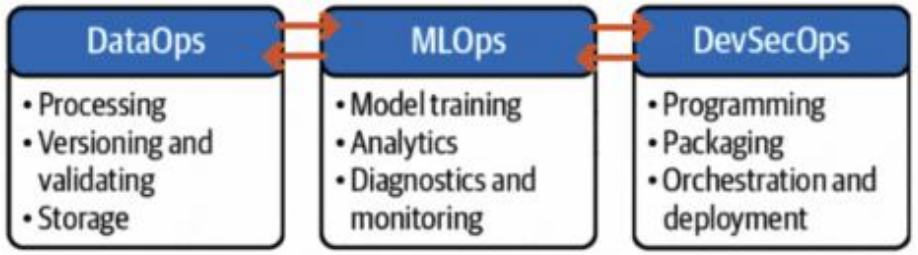
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

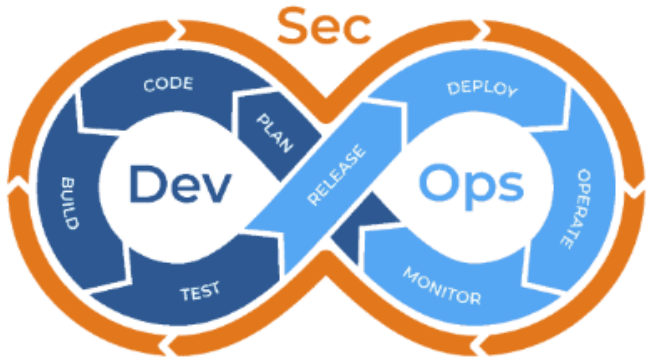
Machine Learning Lifecycle



DevSecOps



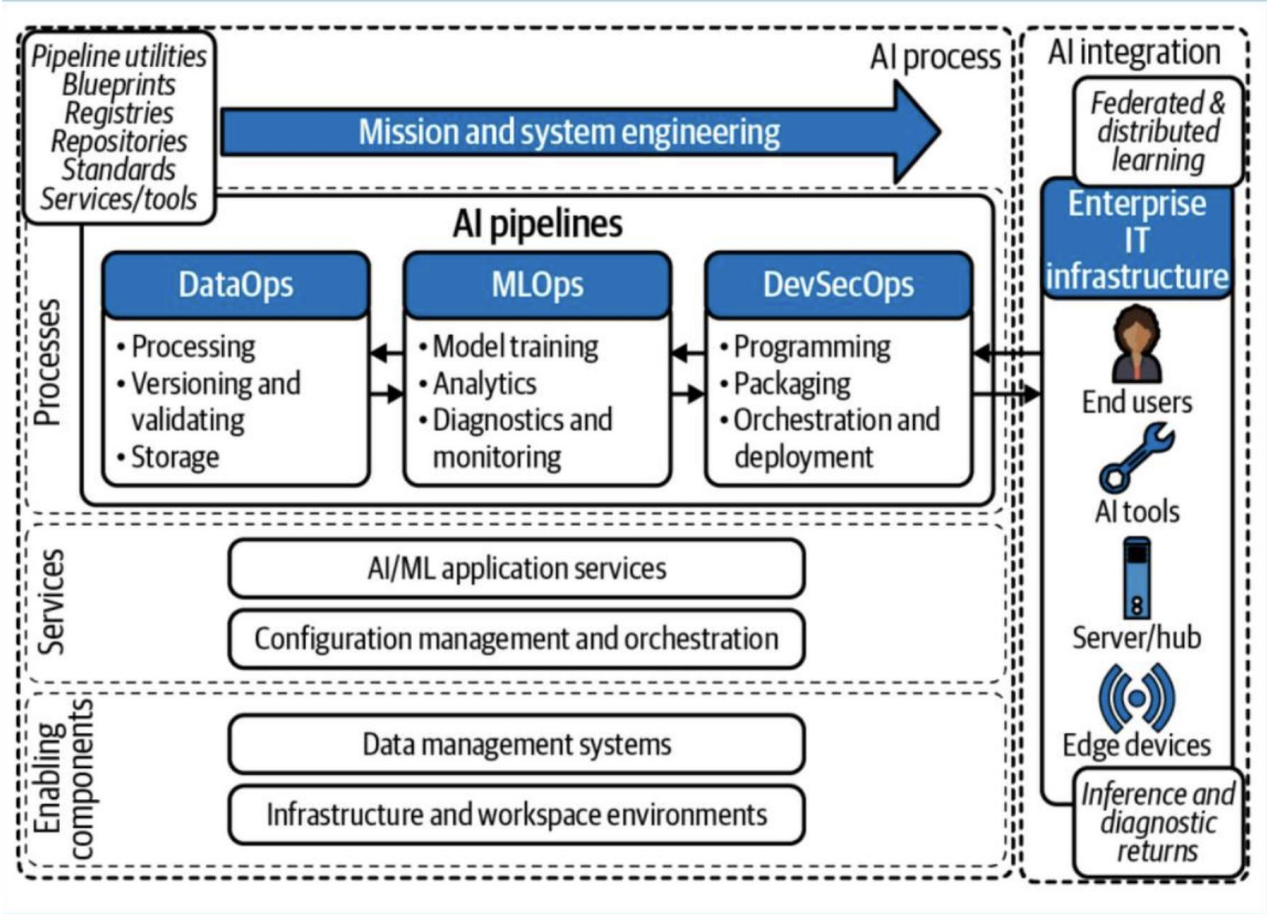
- **DevOps** – a set of processes and techniques devoted to increasing the efficiency of communication and coordination between developers and operations staff, **DevSecOps** expands this relationship further to include security and the need to make systems robust against intentional meddling by a motivated adversary



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

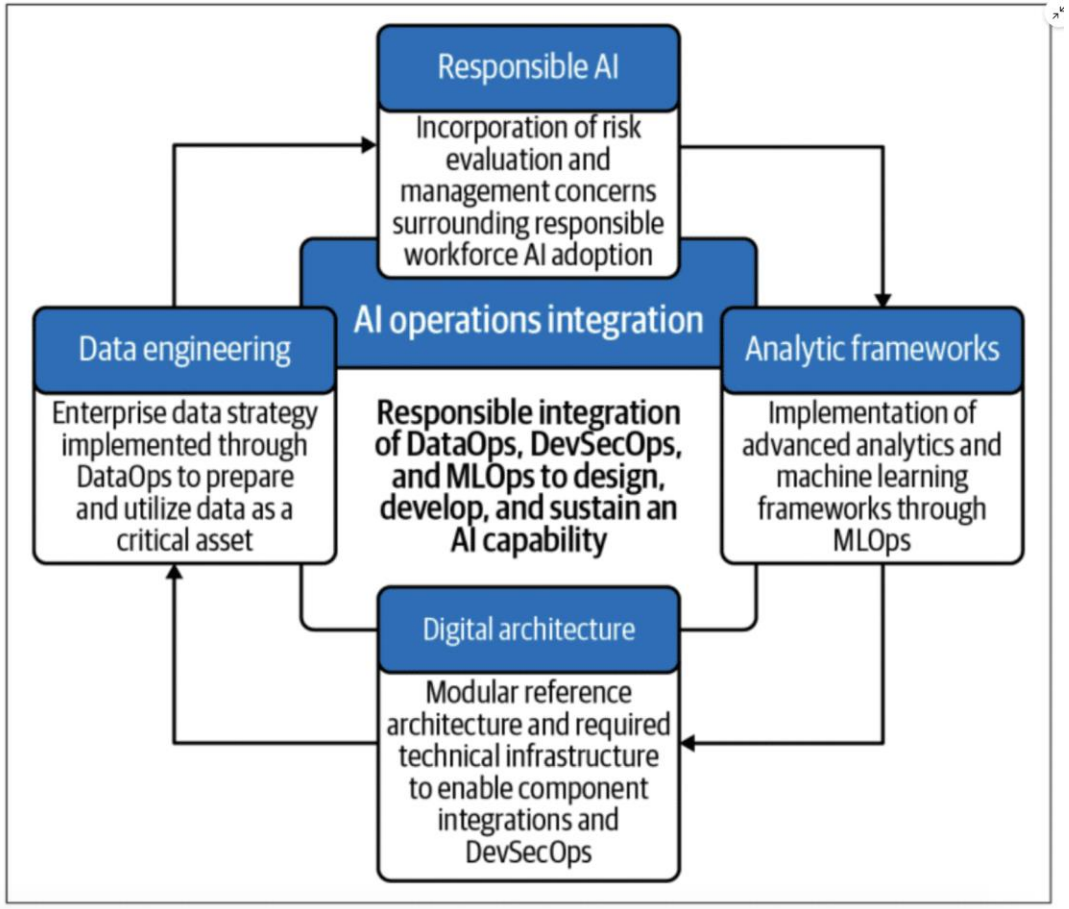
AIOps Integration



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

AI Adoption Blueprint



- **Responsible AI**: addresses risk evaluation and management concerns to promote responsible workforce adoption of AI solutions
- **Analytic frameworks**: leveraging advanced modeling and MLOps frameworks to implement analytics.
- **Digital architecture**: outlines reference architectures and infrastructure to enable component integration and DevSecOps.
- **Data engineering**: realizes enterprise data strategy through implementation of DataOps, utilizing data as a strategic asset.

National Institute of Diabetes and Digestive and Kidney Diseases

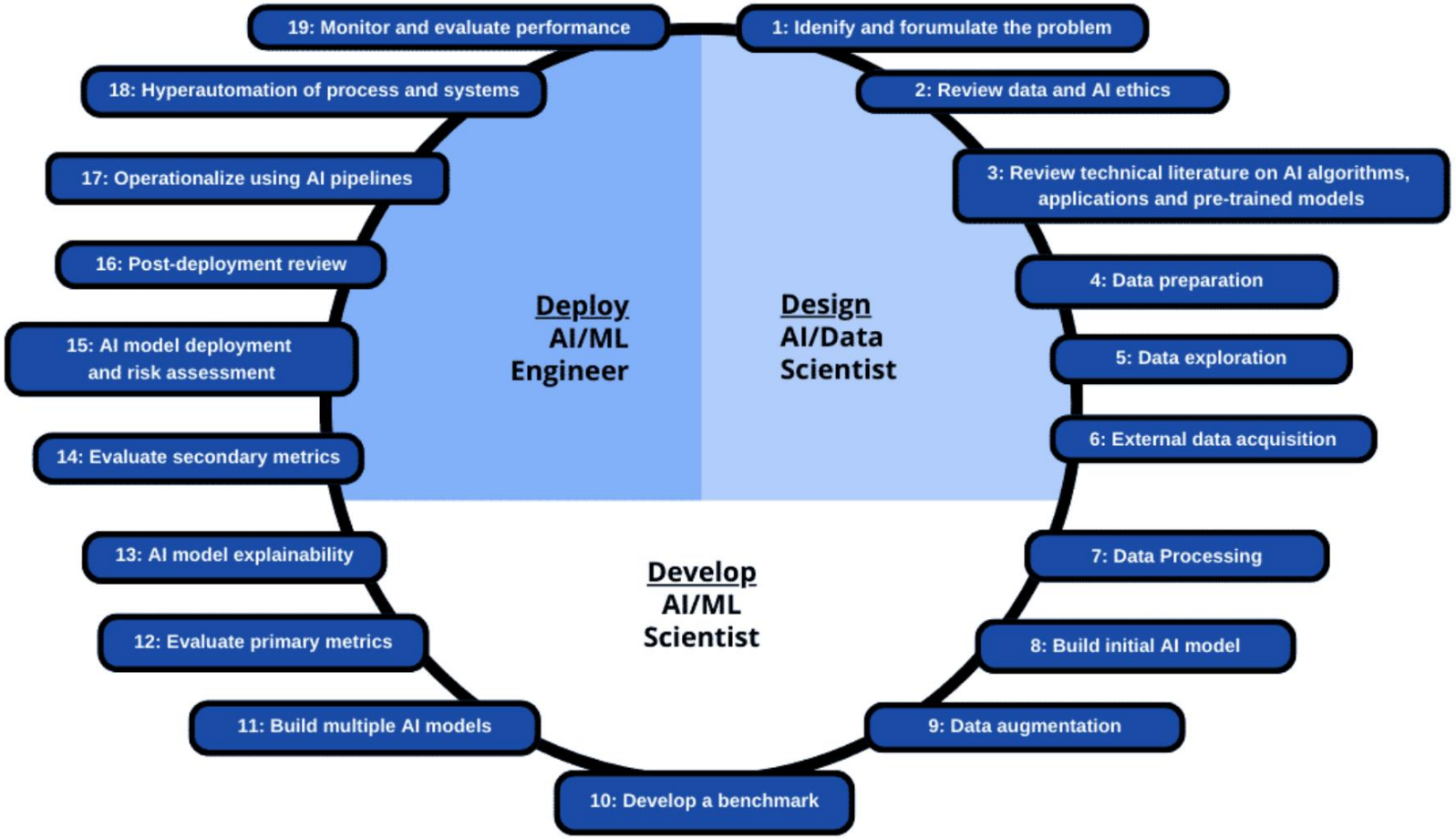
Central Repository

AI Solutions

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Guidelines for Approaching AI Projects

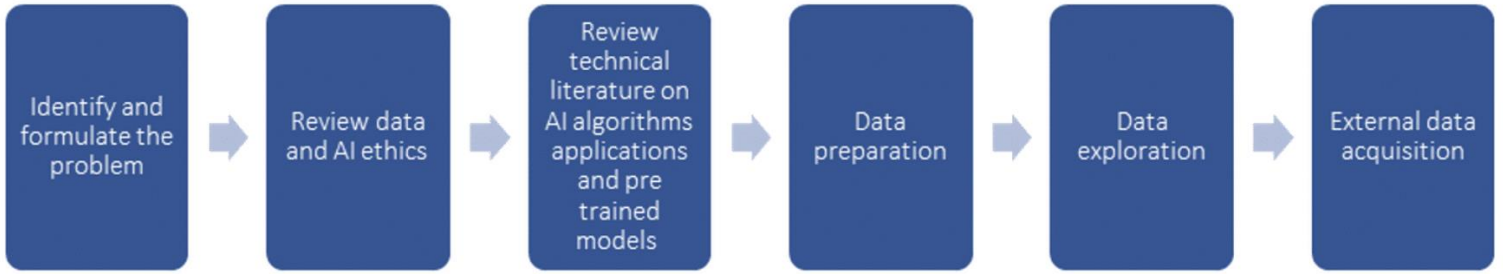


National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Design

- The design phase represents the beginning of the AI lifecycle. The inputs to this phase are typically either a business or mission **need** or an innovative **idea**.
- Critical Activities:
  - Requirements Analysis – decomposition of end-user needs
  - Reference Architecture – infrastructure elements that the AI solution requires
- **Useful Questions to ask:**
  - *What is the problem the solution must address?*
  - *What relevant data is available to the team?*
  - *Is there necessary data that the team lacks access to?*
  - *Do any relevant pre-trained models exist?*
  - *How should available data be explored, understood, and prepared?*



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Develop

- The Develop phase picks up where design ends and seeks to turn a detailed idea into a functioning prototype.
- In practice, these steps are not sequential but rather **cyclical** and **iterative**.
- Use Agile development - w**here cross-functional teams focus on continuous improvement, in which planning, design, and development all run in parallel**
- **Focus on a Minimum Viable Product**
  - An MVP is a product with **just enough features** to attract early-adopter users and **validate a product idea** early in the development cycle.
    - Focus on what matters most.
    - Get small wins early in the development process.
    - Generate user feedback early and often.

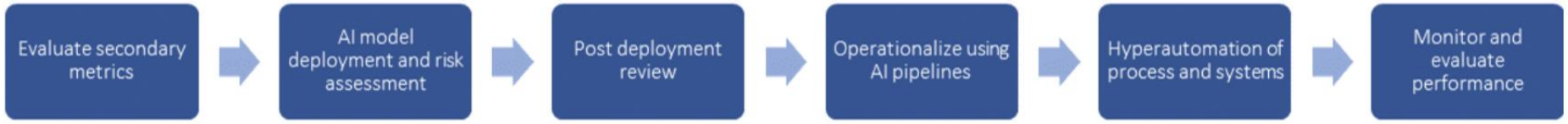
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Deploy

- The Deploy phase is all about putting your AI models to use. Often, this means deploying models to users who will employ them in their day-to-day work streams.
- AIOps for Deployment
  - Activities needed to manage the AI Lifecycle
  - Seeks to take the repetitive, low-value work of deploying and maintaining models off the plate of AI experts, allowing them to focus on innovation tasks that cannot be automated

- If you don’t have a solid operations process in place on your AI project, your team will have the following challenges:
  - Reproducing experimental results will require manual record-keeping and data archiving.
  - Updating an AI model to a new version will be time consuming, and users might experience downtime.
  - All of your critical processes will be prone to human error.
  - Retraining models will be cumbersome and time-consuming.
  - Your team will spend a lot of time on repetitive, low-value tasks, detracting from innovation opportunities.

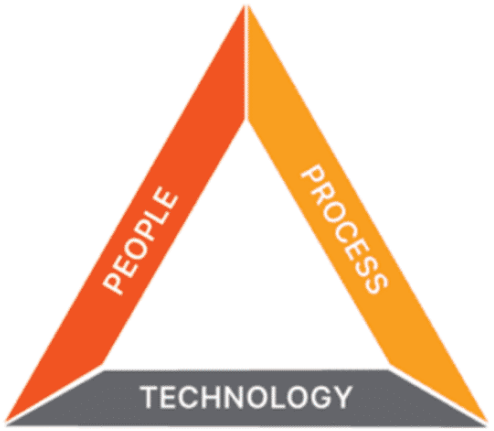


National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

People-Process-Technology (PPT) Framework

Golden Triangle



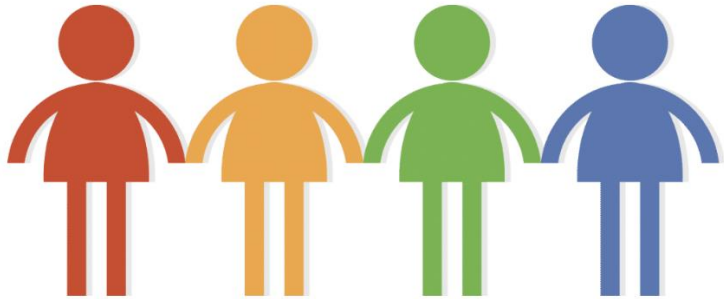
Golden Triangle

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

People

- **People*:*** Workforce knowledge, skills, and abilities to perform required tasks across the AI solution lifecycle.
  - Technical
  - Non-technical
- **Leadership and project/resource management**: Who is driving the overarching strategic AI objective, product, or service forward?
- **Domain and business expertise**: Who is analyzing investments and risks? Who is familiar with the data?
- **Data engineering and stewardship**: Who is defining and maintaining processes for reliable and clean data?
- **Data science and applied mathematics**: Who is analyzing the data in detail to identify the optimal models/algorithms?
- **Machine learning engineering**: Who is researching, selecting, optimizing models and the software/hardware configurations to run them?
- **Software development**: Who is building and managing the software infrastructure?
- **Security and deployment engineering**: Who is ensuring secure implementation of the product or service?
- **Visualization and reporting expertise**: Who is facilitating efficient content digestion for project stakeholders?

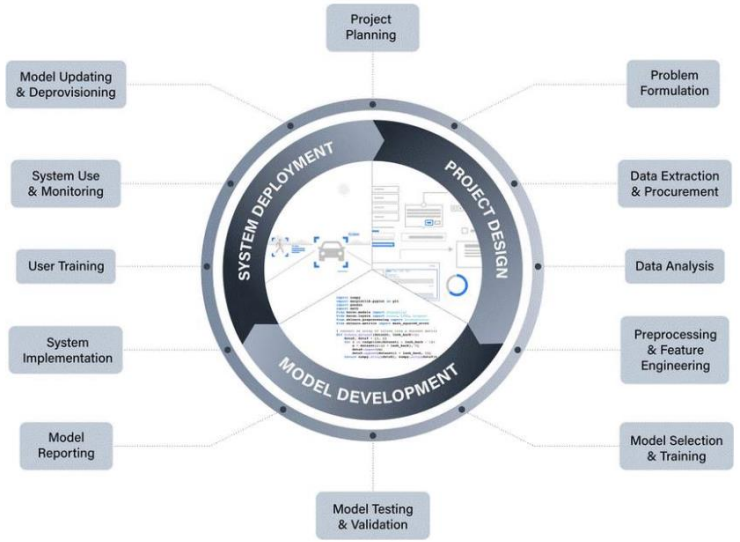


National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Processes

- AI processes seek to orchestrate the interactions between the workforce (people) and the platform or tools (technology) required to engage AI solutions.
- From an operational perspective, you can consider AI Operations in terms of its technical component processes:
  - **DataOps**
  - **MLOps**
  - **DevSecOps**
- You might also leverage AI lifecycle frameworks to inform processes around high-level phases: design, develop, and deploy.



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Technology

- Technical components, in contrast to people and processes, are the fairly tangible components of AI-enabling infrastructure. However, these technologies are extremely diverse in their intended use, capabilities, and application to your unique AI challenge.
- **Platforms**: A digital environment (hardware or software) that houses centralized technology including development tools, applications, services, and user interfaces.
  - EX: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), Advana/Jupiter, etc.
- **Development Tools**: A tool that supports the developer to create, test, and share repositories to enable collaborative AI solutions.
  - EX: Integrated Development Environments (IDEs), GitHub, Databricks, Jupyter Notebooks, Google Colab, etc.
- **Infrastructure as Code**: Software that manages and provisions data centers through machine-readable definition files, rather than physical hardware configuration.
  - EX: Terraform, Ansible, Puppet, Chef, Otter, etc.
- **Physical Infrastructure**: Hardware used to house and provide secure access to platforms.
  - EX: servers, storage media, personal devices, edge devices, sensors, communications technologies, etc

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Technology Demo

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Day 1 Summary

- AI Demystified
- Responsible AI at the NIH
- AI Operations
- AI Implementations

Questions?


REPOSITORY
TECHNOLOGY ADVANCEMENT
BIOSTATISTICS
CLINICAL STUDIES
REGULATORY
Office of Clinical Research Support
NIH
National Institute of Diabetes and Digestive and Kidney Diseases

Arica Christensen – Christensen_arica@bah.com
Dr. Gordon Aiello – Aiello_Gordon@bah.com

April 24 – AI Fundamentals Applications

Thank You!