AI-Ready Data on FASRC Clusters

Manasvita Joshi

Harvard - FAS Research Computing

AI in the News

ICT News
For Indigenous communities, AI brings peril and promise

Artificial Intelligence News
Balancing AI cost efficiency with data sovereignty

Essence
AI Data Centers Are the New Environmental Burden Black Communities Didn't Ask For

PwC
CEO confidence in revenue outlook hits five-year low – as AI becomes a defining divide between leaders and laggards: PwC 2026 Global CEO Survey

Mass.gov
Governor Healey Announces Massachusetts AI Hub to Make State Global Leader in Applied AI Innovation

AI in the News


Mass.gov

Search Mass.gov SEARCH

Governor Maura Healey and Lt. Governor Kim Driscoll

OFFERED BY Governor Maura Healey and Lt. Governor Kim Driscoll | Show 2 more

PRESS RELEASE

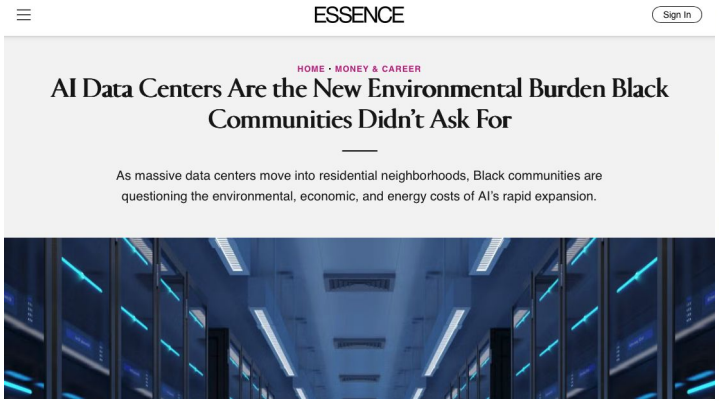
Governor Healey Announces Massachusetts AI Hub to Make State Global Leader in Applied AI Innovation

Leveraging the Mass Leads Act, Initiative Will Support Game-Changing AI Research, Drive Business Growth, Train a Workforce of Tomorrow, and Harness AI to Advance Solutions to World's Greatest Challenges

FOR IMMEDIATE RELEASE:
12/19/2024

Governor Maura Healey and Lt. Governor Kim Driscoll
Executive Office of Technology Services and Security
Executive Office of Economic Development

https://www.mass.gov/news/governor-healey-announces-massachusetts-ai-hub-to-make-state-global-leader-in-applied-ai-innovation


ESSENCE

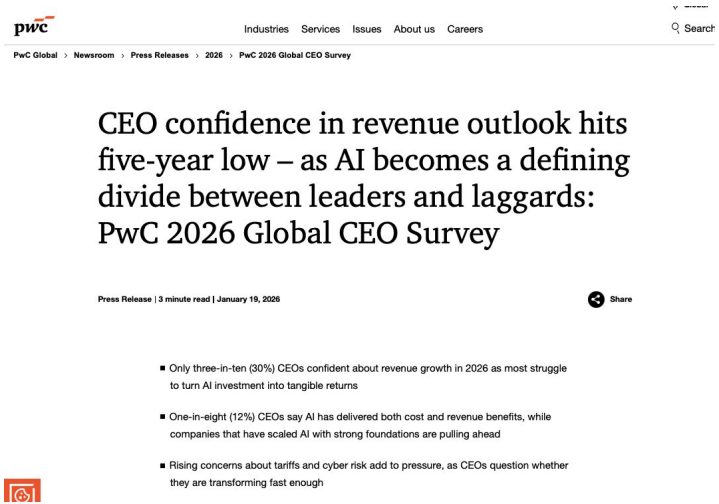
Sign In

HOME · MONEY & CAREER

AI Data Centers Are the New Environmental Burden Black Communities Didn't Ask For

As massive data centers move into residential neighborhoods, Black communities are questioning the environmental, economic, and energy costs of AI's rapid expansion.

https://www.essence.com/news/money-career/ai-data-centers-black-communities/


pwc

Industries Services Issues About us Careers Search

PwC Global › Newsroom › Press Releases › 2026 › PwC 2026 Global CEO Survey

CEO confidence in revenue outlook hits five-year low – as AI becomes a defining divide between leaders and laggards: PwC 2026 Global CEO Survey

Press Release | 3 minute read | January 19, 2026

Share

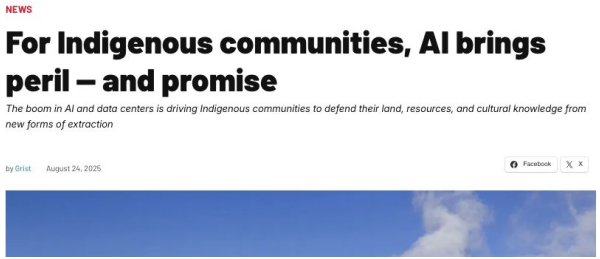
- Only three-in-ten (30%) CEOs confident about revenue growth in 2026 as most struggle to turn AI investment into tangible returns
- One-in-eight (12%) CEOs say AI has delivered both cost and revenue benefits, while companies that have scaled AI with strong foundations are pulling ahead
- Rising concerns about tariffs and cyber risk add to pressure, as CEOs question whether they are transforming fast enough

https://www.pwc.com/gx/en/news-room/press-releases/2026/pwc-2026-global-ceo-survey.html


Balancing AI cost efficiency with data sovereignty

Ryan Daws January 21, 2026

https://www.artificialintelligence-news.com/news/balancing-ai-cost-efficiency-with-data-sovereignty/


NEWS

For Indigenous communities, AI brings peril – and promise

The boom in AI and data centers is driving Indigenous communities to defend their land, resources, and cultural knowledge from new forms of extraction

by Grist August 24, 2025

Facebook X

https://ictnews.org/news/for-indigenous-communities-ai-brings-peril-and-promise/

Benefits of AI for Research

- Data analysis, increasing speed & efficiency of repetitive tasks
- Streamlining data curation through systematic review
- Academic research & writing with ideas, topic organization, & editing capabilities
- Prediction of long-term models or trends using Machine Learning (ML) techniques
- Decreasing time2science (T2S) - research at accelerated pace
  - AlphaFold for protein structure predictions & their benchmark tests
  - AlphaGeometry for solving hard Euclidean geometry problems
  - TxAgent - AI agent for therapeutic reasoning capable of performing multi-step reasoning, real-time knowledge retrieval, and tool-assisted decision-making to analyze drug interactions, contraindications, & patient-specific treatment strategies

Potential Risks of AI for Research

- Growing dependency on AI tools without full comprehension of how the tools operate or whether the results are accurate
- Lack of awareness and understanding around data privacy and security requirements, as well as ownership of intellectual property
- Potential for inadvertent or overt plagiarism in research products
- Unconscious political, cultural or other types of bias injected into research
  - How to ensure that no bias is introduced while using an AI scientist?
  - How to account for bias when an AI agent interacts with a human on a research topic?
- Can produce false information (hallucinations) which may appear to be real
  - https://www.reddit.com/r/singularity/comments/1enqk04/how_many_rs_in_strawberry_why_is_this_a_very/
  - Video: https://www.instagram.com/reel/DTfk0bqiJ52/?hl=en

Impact of AI on data

- Streamlining data curation by automating processing & cleaning
- Analyzing vast amounts of data to accelerate predictive analytics in areas, such as finance, healthcare, marketing, weather forecasting, etc.
- Enabling real-time insights for your research, lab, or an organization
- Enhancing data management by improving data quality, security, & compliance
    - Detecting errors,
    - Flagging inconsistencies,
    - Managing complex data lifecycles
- See https://rivery.io/data-learning-center/ai-data-management/

What is AI-ready data?

- Curated data with AI use in mind
  - High-quality (astronomy phase - “no-data better than bad-data”)
  - Cleaned & Structured (beyond tabular)
  - Data governance compliant
  - Easily consumable by AI systems for training & inference
  - Includes essential context & metadata
  - Stored in a standardised format suitable for research or the organization
  - Is aligned with AI use case at hand, for accuracy & efficiency
- *Almost ML-ready data (complete, consistent, unbiased - featurized)*
- *Additional layers of relevant context & what good looks like for that AI app*
- A necessity if organizations/researchers need the most value from AI efforts
- AI-readiness evaluation intuition-based with: “How fast can I go from raw data to my ML pipeline?” (Interesting report) - *No standardized definition or evaluation framework*

Challenges for Researchers

- No standardized definition or evaluation framework to validate data AI-readiness => inefficiencies in data sharing & reuse
- Publishing data that can support scientific claims while acting as a source for AI/ML applications becomes difficult
- Working with datasets with varying degrees of documentation, formatting, structural consistency => time investment in data preparation & validation
- Theoretical frameworks for AI-ready data exist but critical gap between theory & practice.
- Domain experts & AI practitioners could have different ideas on what AI-ready data is for “them”

Credit: AI-Ready Data Pilot Project Report

How to Achieve AI-Ready Data

- Data Story
  - What's the problem at hand and where the data has originated from
  - Review datasets to understand how to prepare them for AI
- Data Curation & Cleaning
  - Garbage in garbage out - ensure high quality datasets with little inconsistency
  - Quality of output dependent on quality of input
- Metadata
  - Ensure contextual information is retained, so results can be compared for accuracy
  - Enriched with metadata (documentation, schemas, sources) for machine understanding
  - Identify gaps or language issues (abbreviations, acronyms) that could affect AI models' interpretation of the data

Credit: https://youtu.be/95NrwIIRo5o

How to Achieve AI-Ready Data

- Data Accessibility & Discoverability
  - Data lives in a centralized, shared location, accessible by group members
- Formatting & Organization
  - Structured data
    - Easier to parse than unstructured data
    - Consistent templates, sections; or clean & complete tables
  - Consistency
    - Formatting - uniform fonts & layouts or tabular designs
    - Consistent schema across tables & files
    - Language - similar terminology across documents
  - Organized in formats (like txt, CSV, JSON, XML, etc.) - interoperable than proprietary format
  - Machine-readable text, clear visuals, & consistent formatting, naming conventions, etc.

How to Achieve AI-Ready Data

- Data Governance
  - Dataset is ethical & compliant; Follows privacy standards, & anonymized (if needed)
  - DO NOT include an PII (personally identifiable information) in the datasets
- Review
  - Continually monitor model's inputs & outputs to verify results are as expected
  - Establish data quality evaluation metrics
    - completeness, accuracy, consistency, timeliness, & uniqueness
- Establish Best Practices
  - Build pipeline - automate document processing or leverage tools to transform data
  - Standardize templates & maintain single source of truth for tabular data
  - Automate data inspection & quality and set up alerts
  - Build knowledge graphs & semantic layer for documents; data catalog for tabular data
  - Follows FAIR principles - Findable, Accessible, Interoperable, & Reusable

Key Characteristics

- High Quality & Clean
- Structured, Labelled, & Contextualized
- Governed & Secure
- Use-Case Aligned
- Faster AI-deployment - Easily digestible into processes & data pipelines to accelerate development of AI solutions
- Accessible & Scalable - Easily integrated into AI workflows, not just for human analysis but for automated systems

Traditional vs AI Quality

- Analytics quality ≈ clean, de-duplicated, consistent, outliers removed.
  - Predicting wait times would need clean and accurate training and testing data sets for their output - no context needed

- AI quality & extra needs for AI
  - Represents real patterns with noise & edge cases. (gartner.com)
  - Labels / annotations (e.g., intents, categories, outcomes)
  - Rich metadata (source, time, permissions, lineage) (alteryx.com)

Traditional vs AI Quality

- Gen AI cases include quality of prompts & definition of “good” output - having good training sets is not enough
- RAG-based AI agent needs clear & up to date documentation, labels & metadata for additional context to output accurate and relevant responses
- Prompts for AI agents - clear & well structured prompts to provide right context & instructions to AI agents
- **“High quality” alone ≠ “AI-ready.”**

Why prepare data for AI?

- **Enhance Reliability**: accurate patterns => better performance & predictions by deployed models.
- **Accelerate AI Projects**: less time on data prep => faster development to production.
- **Failure Prevention**: access to AI-ready data => AI projects less prone to failure
- **Ongoing Process**: forces to have effective CI/CD pipelines involving monitoring, triage, resolution, enhancement, & timeliness (fresh enough for decisions)
- **Reduces Data Sprawl**: no fragmentation, no data silos => centralized, federated, & standardized access
    a. Copying data (duplication) everywhere increases governance & cost problems

Credits: denodo.com; ibm.com

Why prepare data for AI?

- Enables consistent data quality management practices
- Maintains its standards, quality, governance, security, etc throughout the AI lifecycle.
- Secured from data leakage or prompt injection attacks (for LLM-based applications)
- Introduces best practices:
  a. Profiling, automated checks, anomaly detection, feedback loops from model outputs
  b. Policies for PII, PHI, and regulated data (GDPR, HIPAA, etc.) - **Security**
  c. Strong governance is essential for responsible, compliant AI - **Governance** & **Compliance**
  d. “Zero-trust” data governance as AI-generated content enters data pipelines: don’t assume any data is safe or human-generated; verify and control
- Pilot with one or two AI projects, refine patterns, then scale
  a. Start small but end-to-end (not just a data cleanup exercise in isolation)

Credits: itpro.com

Data Preparation

Task Area	Key Tasks	Tools
Data Cleaning	Remove duplicates, handle outliers, missing values, formatting issues, errors	OpenRefine, Pandas, Tidyverse, Scikit-learn, Jupyter
Quality Assurance	Ensure data quality, remove bias, check for completeness, consistency, & outliers	Custom scripts (Python/R), DVC (Data Version Control), FAIR principles
Privacy	Remove, mask, or swap identifiers, aggregate sensitive fields, document methods to support transparency and privacy compliance	OpenRefine, R (sdcMicro, anonymizer, custom scripts with tidyverse), Python (Faker, custom scripts with pandas), ARX
Transformation	Normalize, scale, augment (audio, image, video), index data, text tokenization, optimize storage & query performance	Scikit-learn, Pandas, spaCy, R (dplyr, stringr, spaCy, tidytext), Hadoop/Spark
Exploratory Data Analysis (EDA)	Visualize distributions, spot anomalies, patterns, and correlations	Matplotlib, Seaborn, Plotly, Bokeh, Excel, R (dplyr, ggplot2)
Feature Engineering	Create/select/dimension reduce/encode/transform features, address multicollinearity	FeatureTools, Scikit-learn, Pandas, R (tidyverse, tidymodels, data.table)
Reproducibility & Versioning	Track data transformations and versions, ensure transparency	DVC, git

Credit: Borrowed from AI Facilitation Materials WG

AI-Ready Data Preparation Use Cases

- Cleaning & curating raw global record of 4.3B tweets spanning time, geography, & language (geotweets) using Jupyter notebook for sentiment analysis with natural language processing (NLP) models, such as BERT
- Converting raw sequencing data into table of gene-level counts (or a count matrix) with rows representing genes & columns for samples
- Introducing a labeling scheme for behavioral sensor data & parallelizing annotation workflow for downstream ML using Jupyter notebook, Label Studio, or Apache Spark
- Enrich large datasets with metadata, such as subject- and keyword-enrichment of social sciences, humanities, and medical data

Credit: Borrowed from AI Facilitation Materials WG

The Metadata Problem in AI Ready Data

Credit: AI Metadata Enrichment Project

AI Is Only As Good As The Data Behind It- What can we do about it?

185,000+

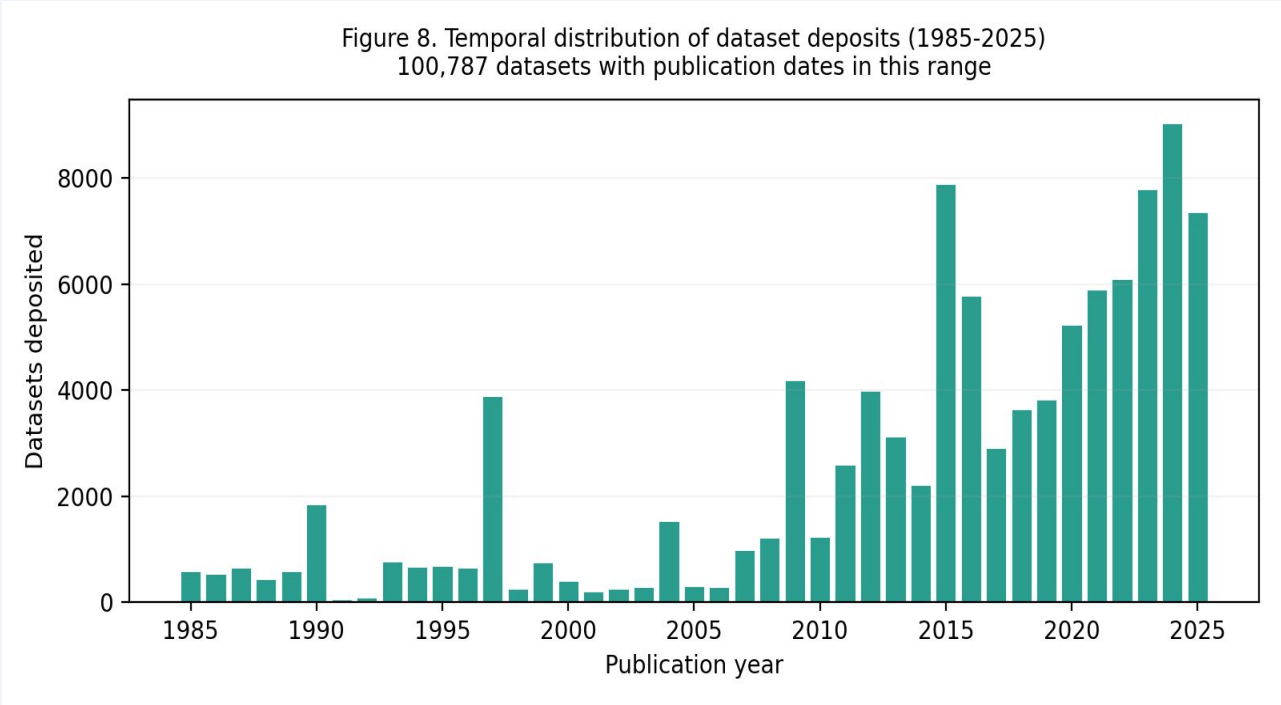
datasets in Harvard Dataverse

5–10%

have rich, curator-ready metadata

43,991

carry any geospatial metadata



"If a book needs great footnotes, a dataset needs great metadata"- Dr. O'Neill, Harvard University

Towards AI-Ready (Meta) Data for GeoAI

Credit: AI Metadata Enrichment Project

Two-way street

- AI can enhance metadata enrichment.
- Enriched metadata improves data discoverability and usability.
- AI is only as reliable as the data it learns from!
- High-quality metadata, in turn, can train more accurate AI models.

Future Work

- Geospatial Enrichment: Enrich with spatial fields, standardize place name
- LLM enrichment: facts about the metadata from codebooks & replication packages
- Collaborations: Trusted Data Collaboration, NIH GREI, DataCite
- Curator interviews: validate metadata measures against expert judgment!

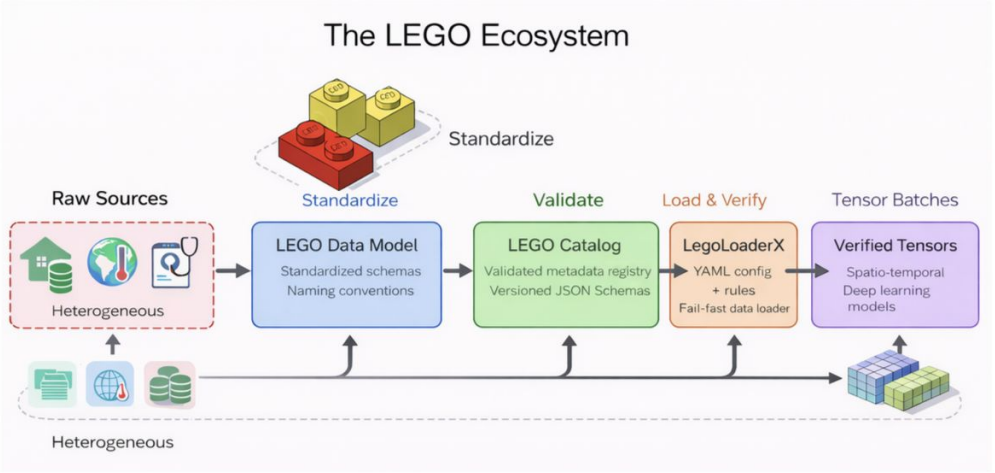
Batch script on gpu_test with checkpointing

GeoAI is only as intelligent as the (meta)data it learns from!

AI-Ready Environmental-Health Data

- HDSI building an open-source data management stack, LEGO Ecosystem, to streamline the path from fragmented multi-domain sources to AI-ready spatio-temporal inputs, specifically for tabular data
- Data & setup for multi-source data fusion for predictive modeling on Cannon

- Standardized schemas & naming conventions to signal cross-domain patterns
- Validated metadata registry catalog
- Specialized PyTorch Dataset/ Dataloader for structured tensor outputs to provide standardized interface for downstream AI/DL architectures



AI Tools for Research

Versatile Chatbots	Google Gemini (Bard); Microsoft Copilot; OpenAI ChatGPT, Claude
Data Visualization	Adobe Firefly, Canva, Prezi
Research Tools	Research Rabbit, Elicit, Perplexity, Consensus, Scholarcy, Scite, Semantic Scholar
Code	GitHub Copilot

Generative AI tool comparison, Harvard Information Technology: https://www.huit.harvard.edu/ai/tools

Storing AI-Generated Data

Storage needs to be scalable, efficient, fast, and reliable

- **Adaptable**
  - Ability to integrate with changing research requirements
  - Keep up with new AI pipelines and ML workflows
- **Scalable**
  - Exponential growth in dataset size and the rate of acquisition
  - Accommodating a variety of data types
  - Storage availability in existing data centers
- **Efficient**
  - Connectivity to HPC environments with immediate access required
  - Parallel processing; read and write speeds
- **Reliable**
  - Continuously running with little to no downtime; updated to newest technology and security requirements
- **Integration**
  - Integrate with AI tools for Research (as listed in previous slide)

Roles and Responsibilities

Preparing data to be AI-ready is a **team effort** (not just IT!)

- Research Administration - Compliance, regulations, data policies
- Research Computing and IT - AI tools, storage, networking, data science
- Security and Privacy Office - safety, risk reduction, data agreements, secure data
- Library - training, resources, courses, sharing
- Labs and researchers - data curation, metadata, stewardship

Summary

- AI offers many benefits to academic research, streamlining data curation, executing repetitive tasks, predicting trends, and the development of research products
- It can also increase risks such as plagiarism, unconscious bias, hallucinations, and intellectual property concerns
- AI-ready data is achievable by creating high-quality, clean, and contextualized datasets
- Preparing data for AI can enhance reliability, improve data quality, accelerate projects, streamline pipelines, and reduce potential security risks

Summary

- Many AI tools already exist, helping with data preparation as well as the generation of other research materials (papers, code, presentations etc.)
- Storage for AI-ready data differs from standard data storage, requiring further consideration when selecting an appropriate option, as it must be scalable, reliable, and adaptable
- Preparing data for AI is a collaborative effort between entities across research institutions

FASRC documentation

- User Docs: FASRC DOCS
- Training
  - Calendar: Training Calendar | FAS Research Computing
  - Material: Training Materials – FASRC DOCS
- Getting help
  - Office hours: Virtual Office Hours | FAS Research Computing
  - Ticket:
    - HUIT Service Portal -> Submit Ticket: Submit Ticket - IT Help
    - Email: rchelp@rc.fas.harvard.edu

Training session evaluation

Please, fill out our training session evaluation. Your feedback is essential for us to improve our trainings!!

https://tinyurl.com/FASRC-training

Thank You!

Questions? Comments?

Supplemental Content

Example: Making Support Data AI-Ready

Use Case – Support AI Assistant

- Target AI use case: agent assist / customer self-service.
- Needed data:
  - Historical tickets with resolution and categories
  - Knowledge base articles, FAQs
  - Customer profile and product configuration (carefully scoped)
- Steps to AI-ready:
  - Standardize ticket fields, deduplicate customers
  - Label tickets with intents / topics
  - Govern PII exposure; restrict free-text fields as needed. (uniphore.com)

Glossary

1. Data Governance -
    a. System of policies, processes, roles & standards that ensures an organization's data is secure, accurate, consistent, & usable throughout its lifecycle
2. Data Literacy -
    a. Ability to read, understand and utilize data for making better decisions (data-driven decisions)
3. Data Democratization -
    a. Making relevant data or digital information accessible to the average non-technical user, within an organization, so that they can analyze & utilize it for decisions without relying on IT
4. Data Accuracy -
    a. Data that is correct, precise, relevant, & free from errors
5. Data Consistency -
    a. State of data with instances that are same across all systems & databases
6. Data Strategy -
    a. Detailed plan for using data to enhance decision making, and optimize research/business outcome

contd.

1. Data Quality -
    a. Measurement of the relevance, accuracy, completeness, validity, consistency, uniqueness, & timeliness of a dataset while complying with the data governance & security policies, if any
2. Data Management -
    a. Collecting, processing, and using data securely & efficiently for better outcomes throughout a project's lifecycle
3. Data Security -
    a. Protecting data from unauthorized access, corruption, & theft throughout its lifecycle
4. Data Privacy -
    a. Principle that a person should be in control of their personal data with the freedom to decide how it should be shared & handled by organizations collecting it
5. Data Provenance & Data Lineage -
    a. Documented, chronological record of a dataset's origins, transformations, & movements throughout its lifecycle to provide its authenticity by capturing its metadata and keeping that intact

contd..

1. Data Silos -
    a. Isolated collections of data that prevent data sharing between different groups, departments, systems, or business units
2. Data Integration -
    a. Combining & harmonizing data from multiple sources into unified, coherent formats for AI use cases
3. Data Fabrics -
    a. Data architecture designed to democratize data access across an organization
4. Interoperability -
    a. Standards based approach for sharing data & functionality between different IT systems with minimal end user intervention
5. EU AI Act
    a. Law to govern the use &/or development of AI in the European Union
6. AI Ethics
    a. Multidisciplinary field to understand the benefits of AI while reducing risks & adverse outcomes