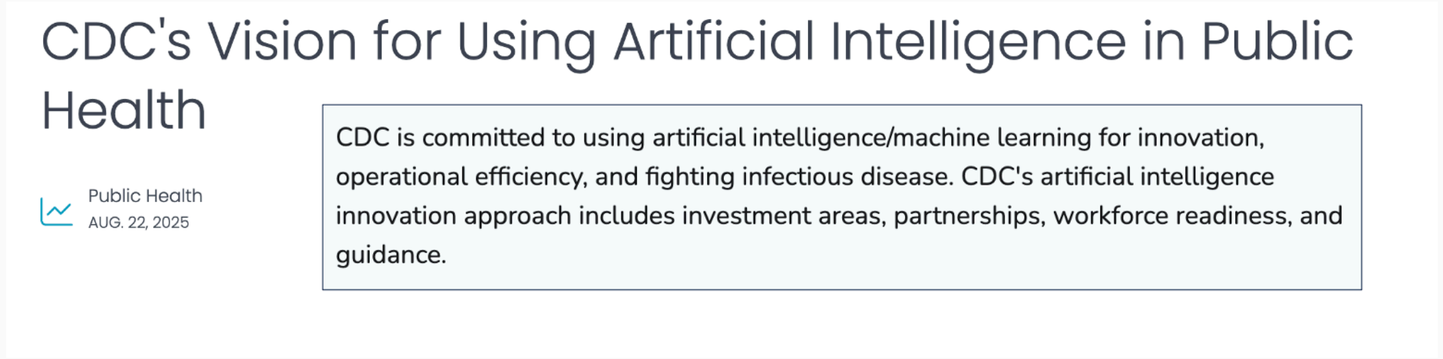
Data Science, AI, and Machine Learning in Public Health using R

Part 2

December 2025

Presented By: Wronski Associates

Introduction: Directing AI Use in 2025



- **M-25-21:** "Accelerating Federal Use of AI through Innovation, Governance, and Public Trust" (April 3, 2025)
- **M-25-22:** "Driving Efficient Acquisition of Artificial Intelligence in Government" (April 3, 2025)
- **Executive Order 14319** - "Preventing Woke AI in the Federal Government" (July 23, 2025)
- **America’s AI Action Plan** (July 2025)

  Source: The White House

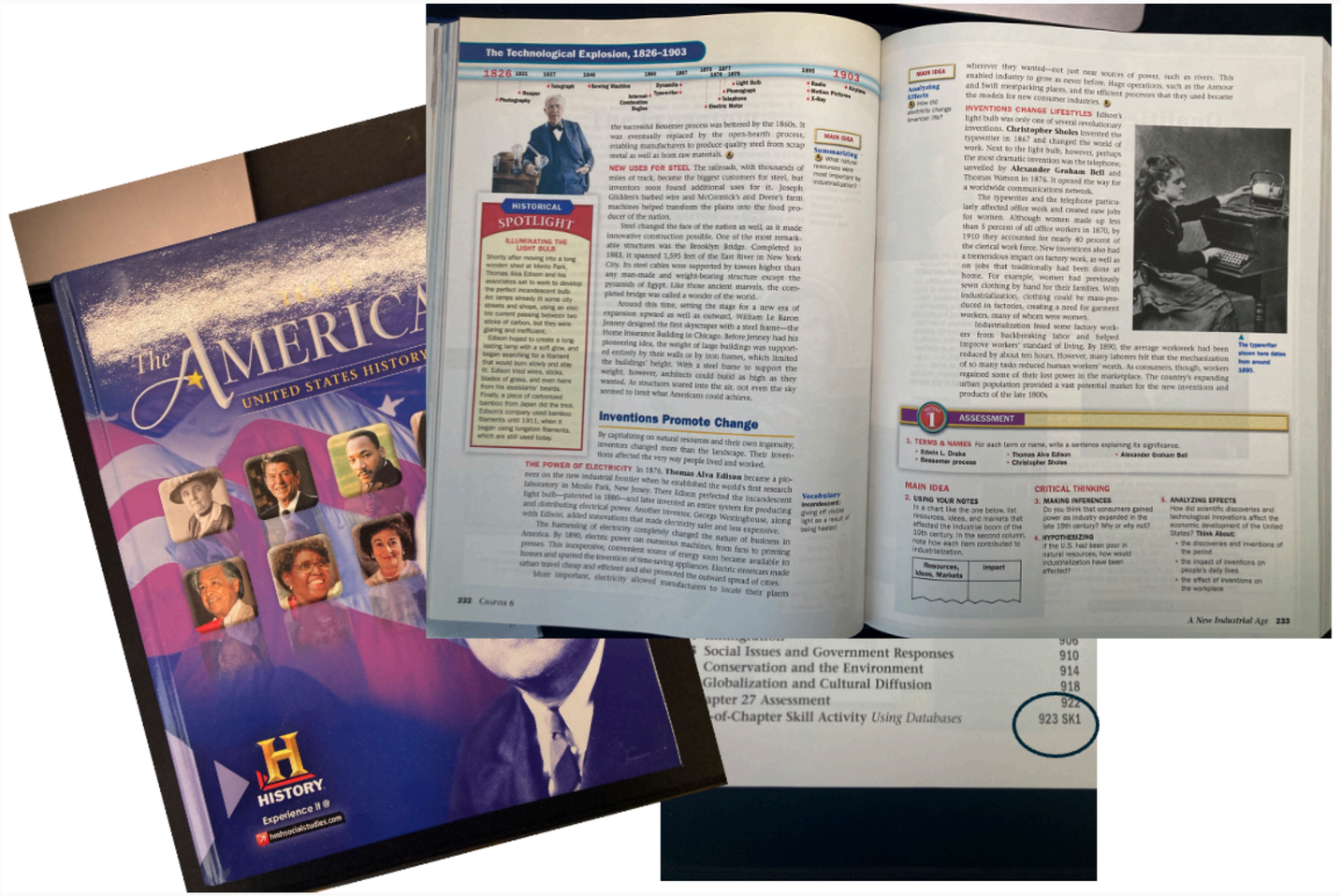
To Do What, Exactly?

That is What We Are Here to Figure Out

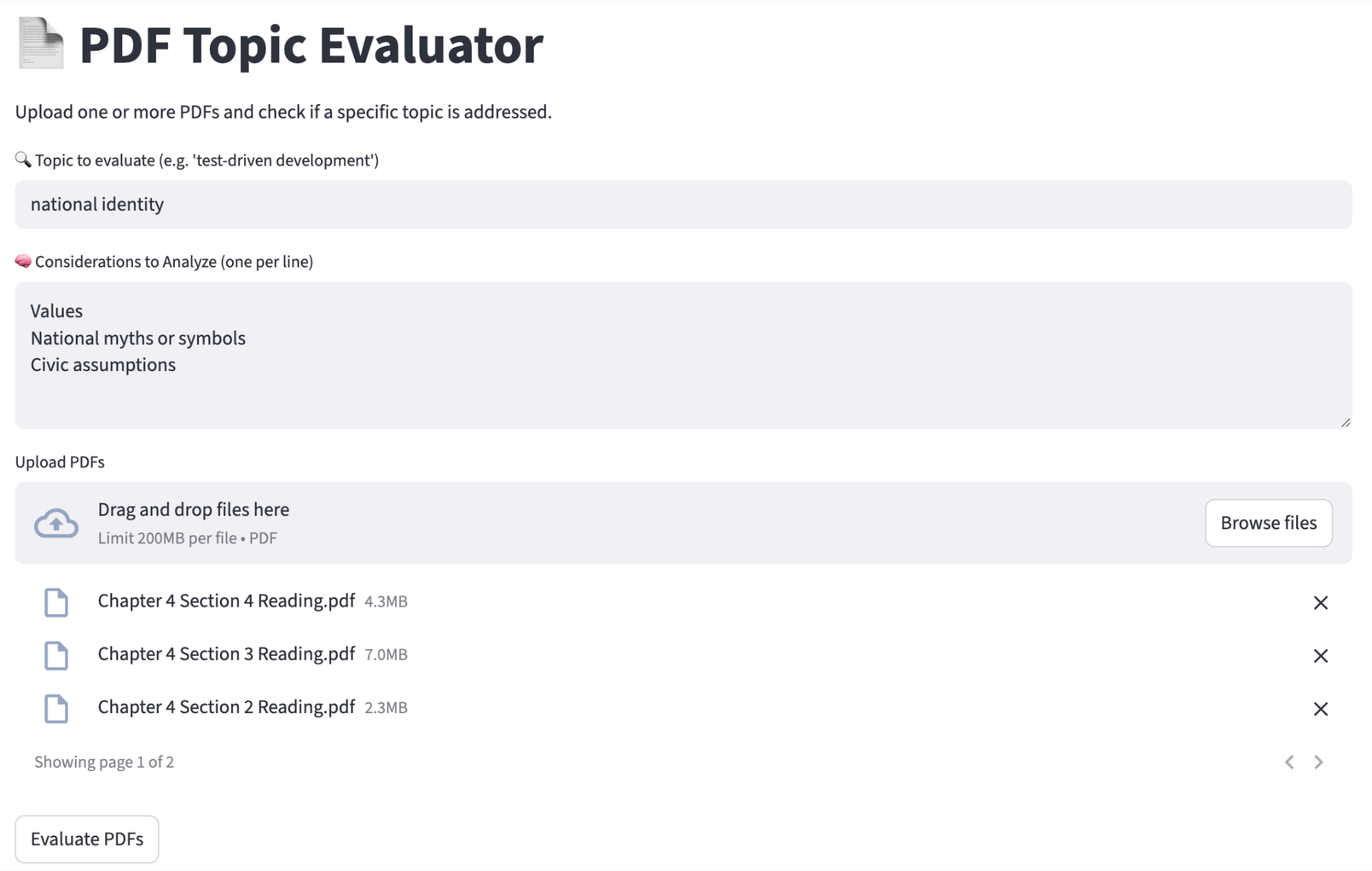
Prompt: Create an image related to AI. The image presents the learning objectives for a course. The learning objectives are: Learning AI Foundational Concepts, Problem Triage - Can AI be useful?, and Technologies and Terminology.

A Quick Working Example

I have 20 secondary school history textbooks. Each book is broken down into chapters. Each chapter has multiple sections. Each book has about 90 sections totaling around 1000 pages. I only want to read the sections that are pertinent to my study.



Creating a Computational Filter - The Input



Creating a Computational Filter - The Output

Chapter 4 Section 3 Reading.pdf

1. **Determination**: The document does not directly address the topic of "national identity."
2. **Explanation**: The excerpts provided focus primarily on historical events related to the American Revolutionary War, including military strategies, battles, and key figures involved in the war effort. The content is centered around the practical aspects of achieving independence, such as military campaigns, alliances, and the challenges faced by the Patriots. While these events are foundational to the development of American national identity, the document does not explicitly discuss the concept of national identity itself.

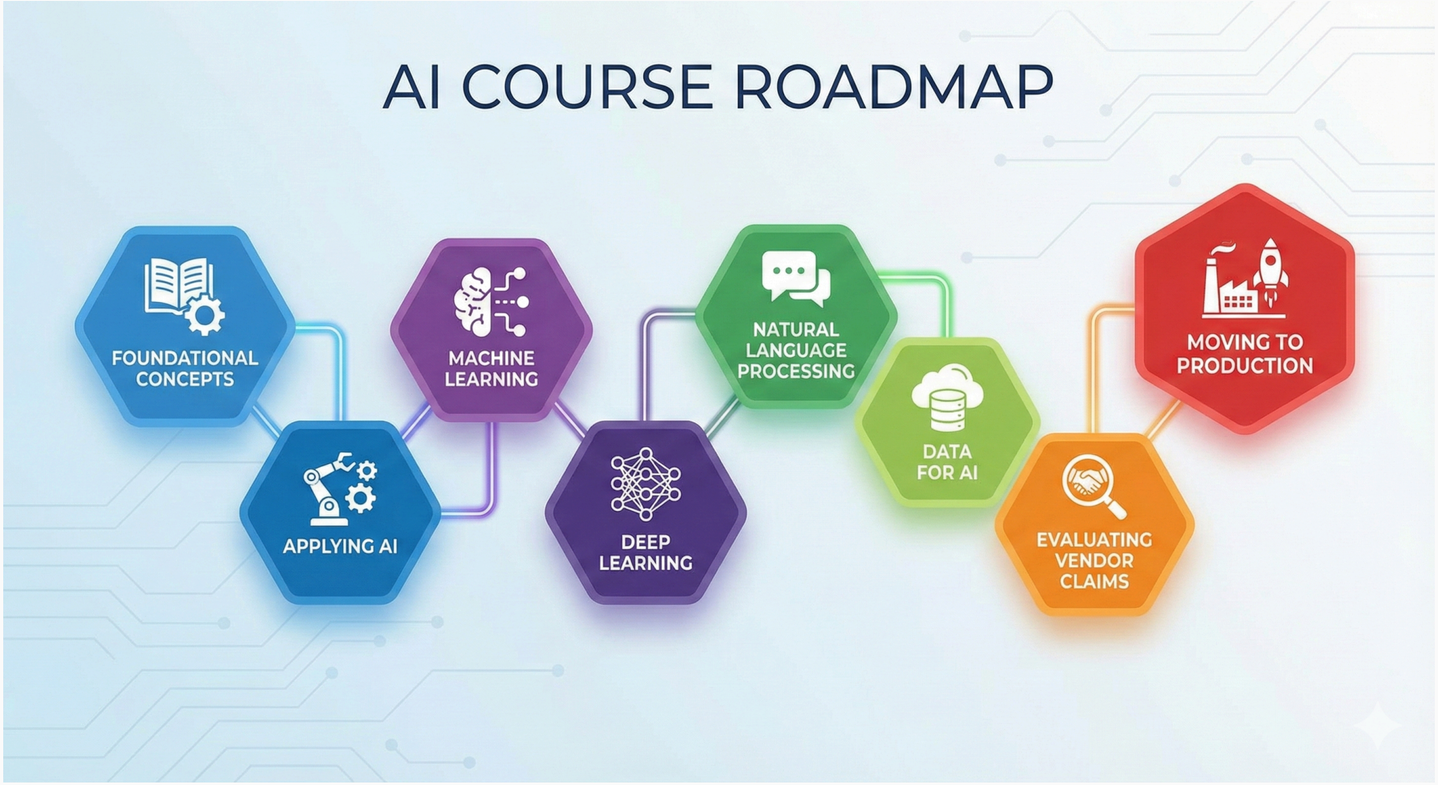
- The excerpts do not delve into values, national myths or symbols, or civic assumptions that are typically associated with discussions of national identity. Instead, they focus on the tactical and strategic elements of the war, such as the use of mercenaries, the role of foreign allies, and specific military engagements.
- There are no direct quotes or phrases in the excerpts that explicitly address the formation or expression of a national identity, such as discussions of shared values, symbols like the flag, or the creation of a national narrative.

In summary, while the events described in the document are part of the historical context that contributed to the formation of American national identity, the document itself does not explicitly address the topic of national identity.

Chapter 4 Section 4 Reading.pdf

1. **Determination**: Yes, the document addresses the topic of "national identity."
2. **Analysis**:
    - **Values**: The document reflects values such as liberty, courage, and patriotism. These are evident in the passages that discuss the contributions of women and soldiers during the war, emphasizing their dedication to the public good and the fight for freedom. The emphasis on liberty and resistance against tyranny are core values that contribute to the formation of a national identity.
    - **National Myths or Symbols**: The document alludes to the mythic narrative of the American Revolution as a struggle for freedom and independence. The mention of the Treaty of Paris and the return of soldiers to their homes symbolizes the birth of a new nation and the establishment of its borders, which are key elements in the national identity of the United States.
    - **Civic Assumptions**: The document assumes a shared commitment to the ideals of independence and self-governance. The references to the formation of a new government and the collective efforts of soldiers and civilians highlight the civic responsibility and unity necessary for nation-building.
    - **Direct Quotes**:
        - "Born for liberty, disdaining to bear the irons of a tyrannic Government, we associate ourselves..."
        - "The courage of soldiers and civilians had made America's victory possible."

Course Roadmap



Prompt: Create an image related to AI. The image presents the topics, or roadmap, for a course. The topics to be covered in order are: Foundational concepts, applying AI, Machine Learning, Deep Learning, Natural Language Processing, Data for AI Evaluating Vendor Claims, Moving to Production.

Foundational Concepts: What is AI?

- **What is AI?** — The landscape: AI, Machine Learning, Deep Learning, and NLP
- **Machine Learning Overview** — How machines learn from data; supervised, unsupervised, and reinforcement learning
- **How AI Differs from Classical Statistics** — Prediction vs. inference; different goals, different workflows
- **Data Types** — What your data needs to look like; labels, structure, volume, quality, and access
- **Putting It Together** — A framework to connect it all

By the end of this section you will be able to:

- Understand the vocabulary of AI and ML
- Recognize what kind of AI is being proposed
- Ask the right questions about data
- Distinguish when ML is appropriate vs. classical statistics
- Build on this foundation for the rest of the course

Defining AI?

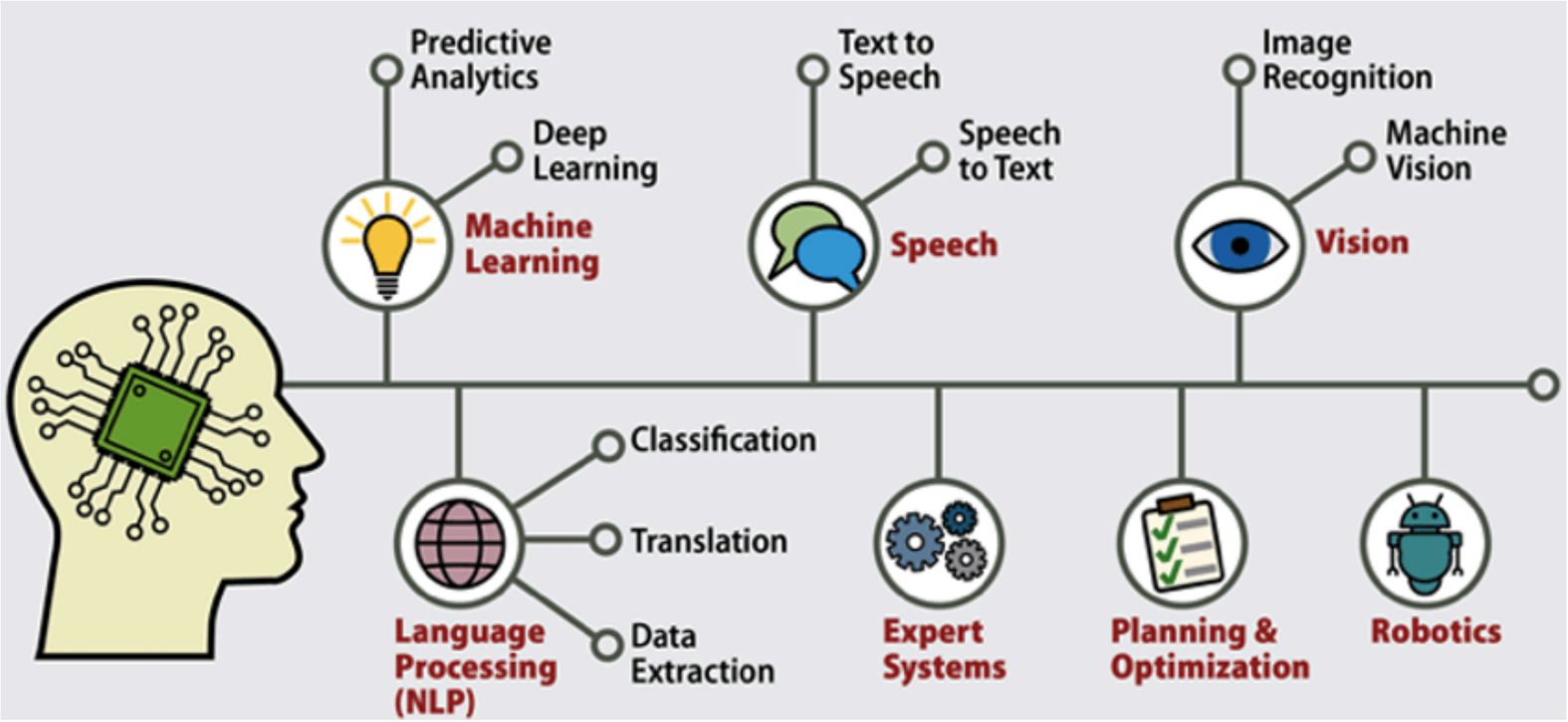
> ***Artificial intelligence (AI)*** *is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action. (15 U.S.C. 9401(3)).*

https://www.cdc.gov/data-modernization/php/ai/cdcs-vision-for-use-of-artificial-intelligence-in-public-health.html

Artificial Intelligence is **a set of computer techniques that let machines perform tasks that normally require human thinking**, such as recognizing patterns, making decisions, summarizing information, or predicting what might happen next.

Instead of following a strict list of rules, AI systems learn from examples, adjust themselves as they go, and improve when given more data.

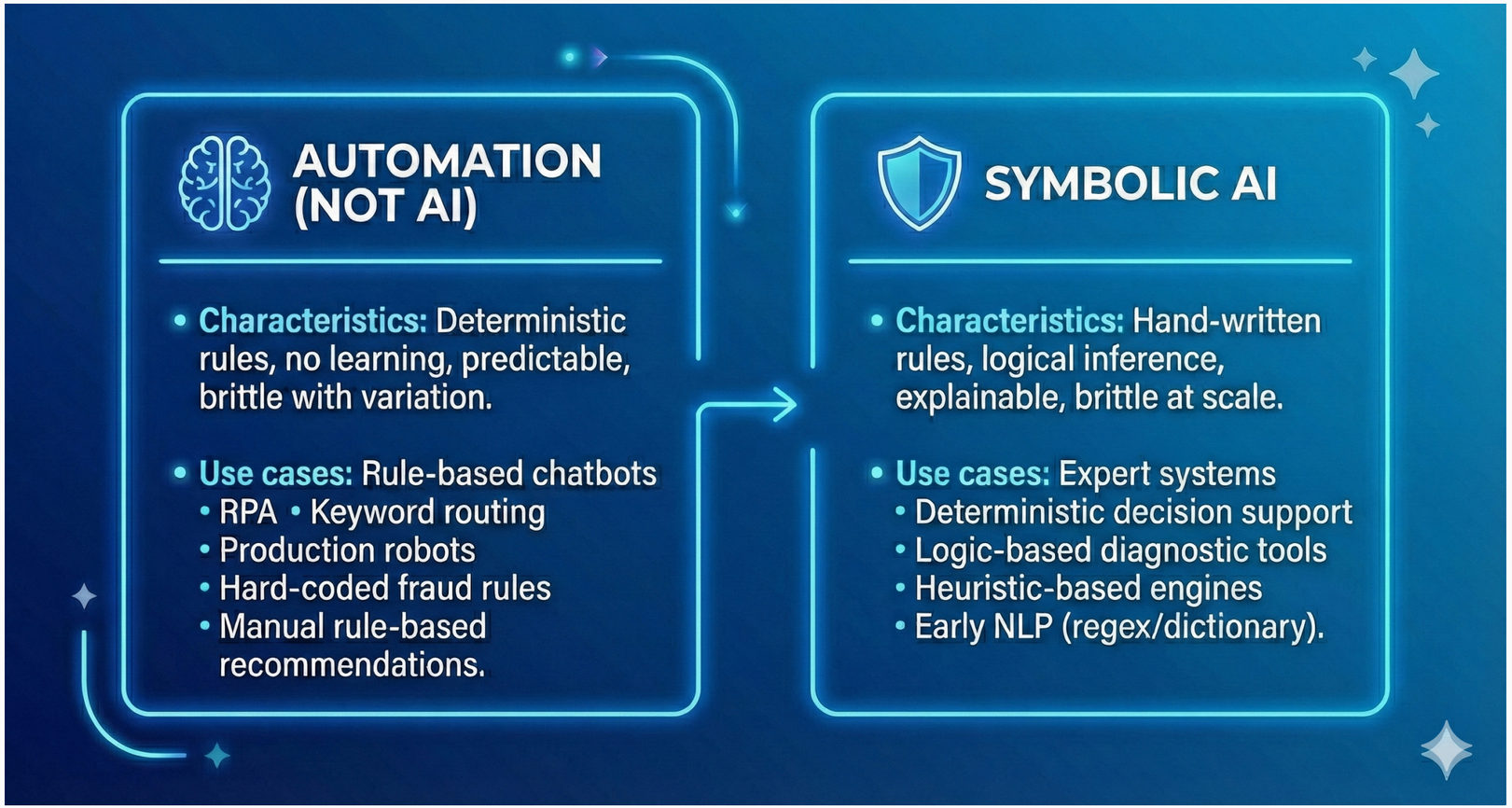
A Very Broad View of AI



Note: This slide is a confusing mixture of AI techniques, possible use cases and use cases that might not be modern AI.

It Probably Isn’t AI

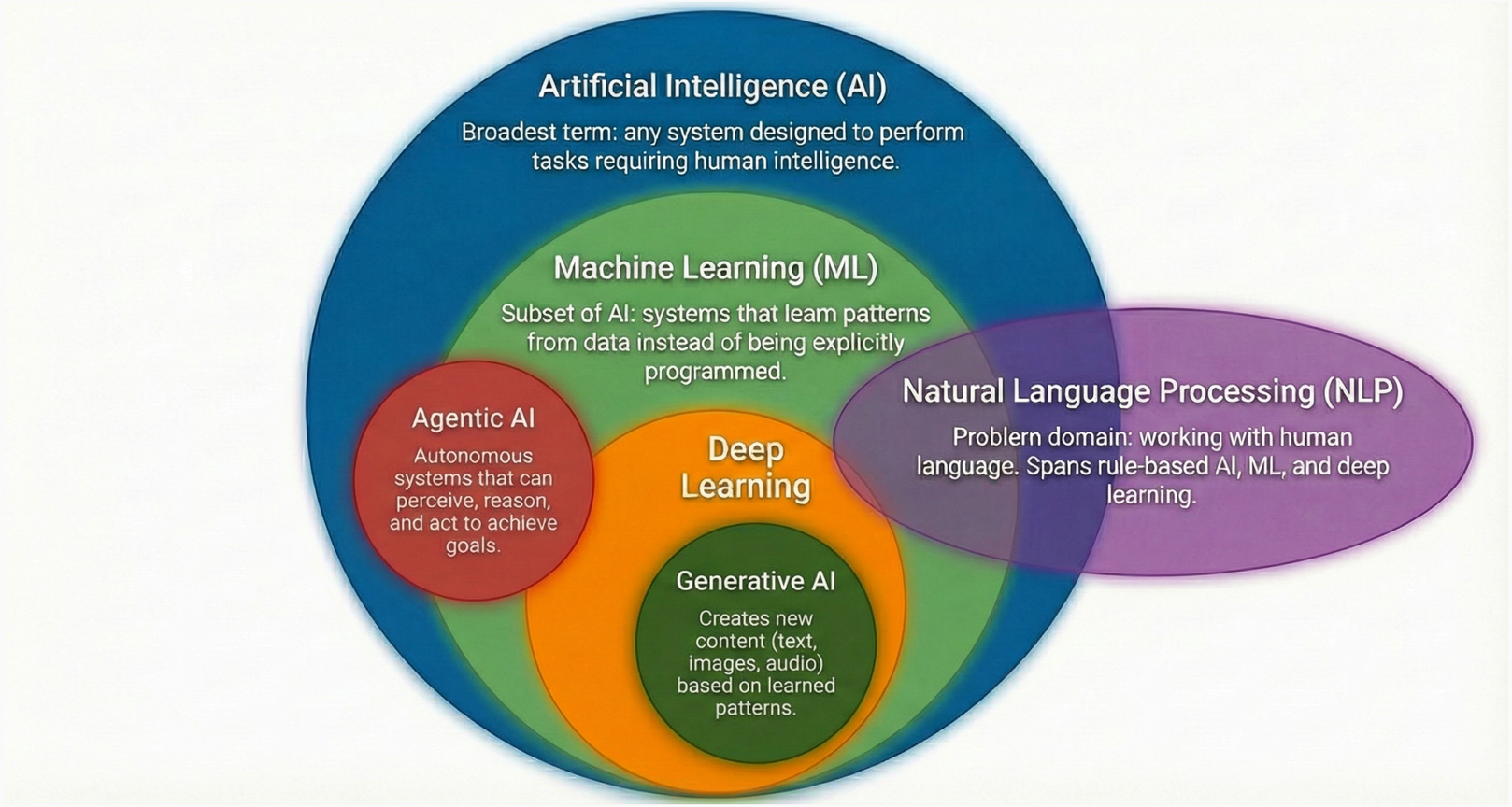
Some automation and expert systems use human-written if/then rules to imitate expert reasoning, while AI discovers its rules from data.



Prompt: Make the first item after 'Use Cases:' a bullet point. Rules-basedd chatbots is a bullet point. Expert systems is a bullet point. Make sure each bullet point is on a separate line.

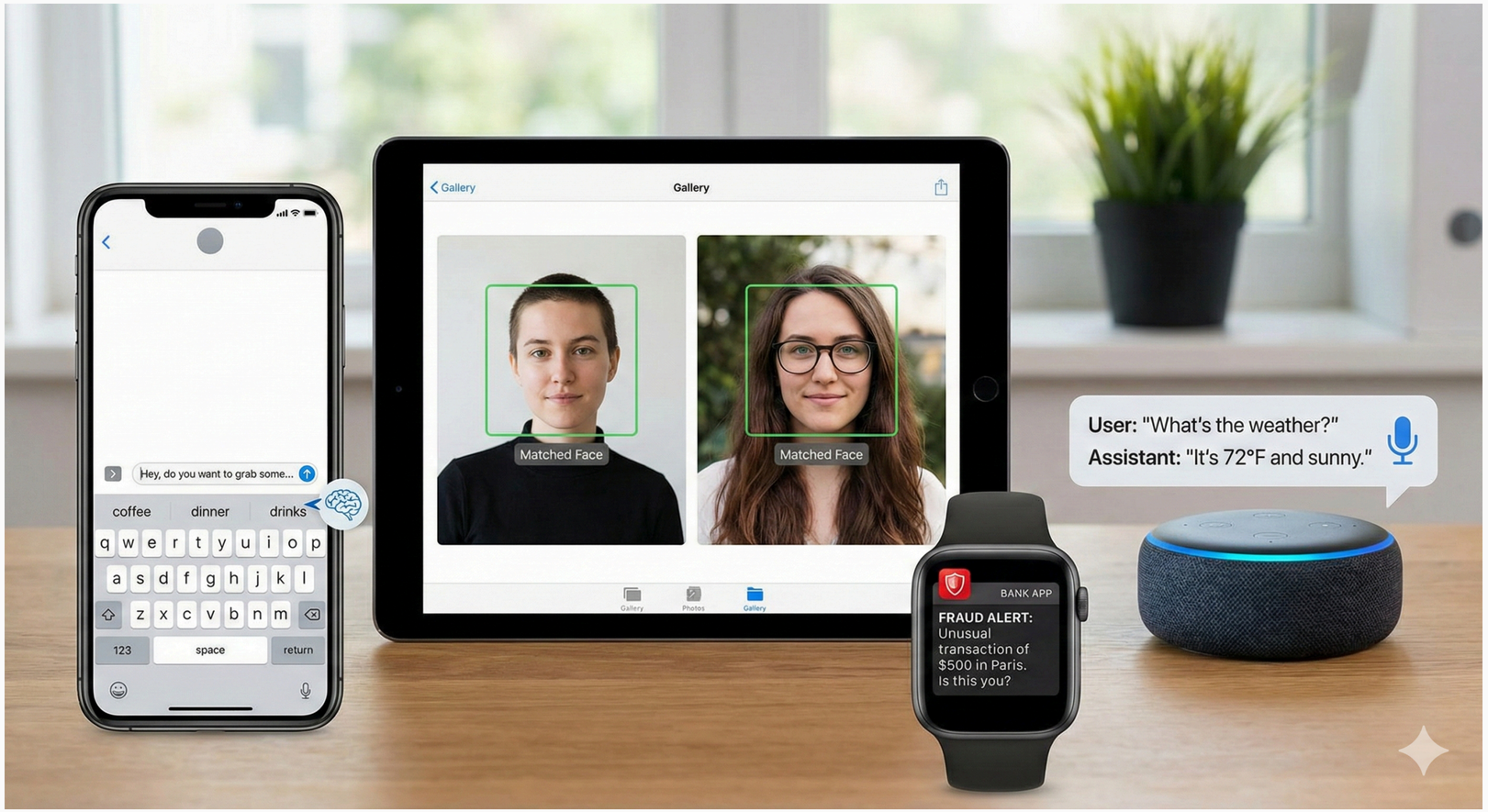
AI Subfields

AI is a field that includes multiple techniques (ML, deep learning, etc.).



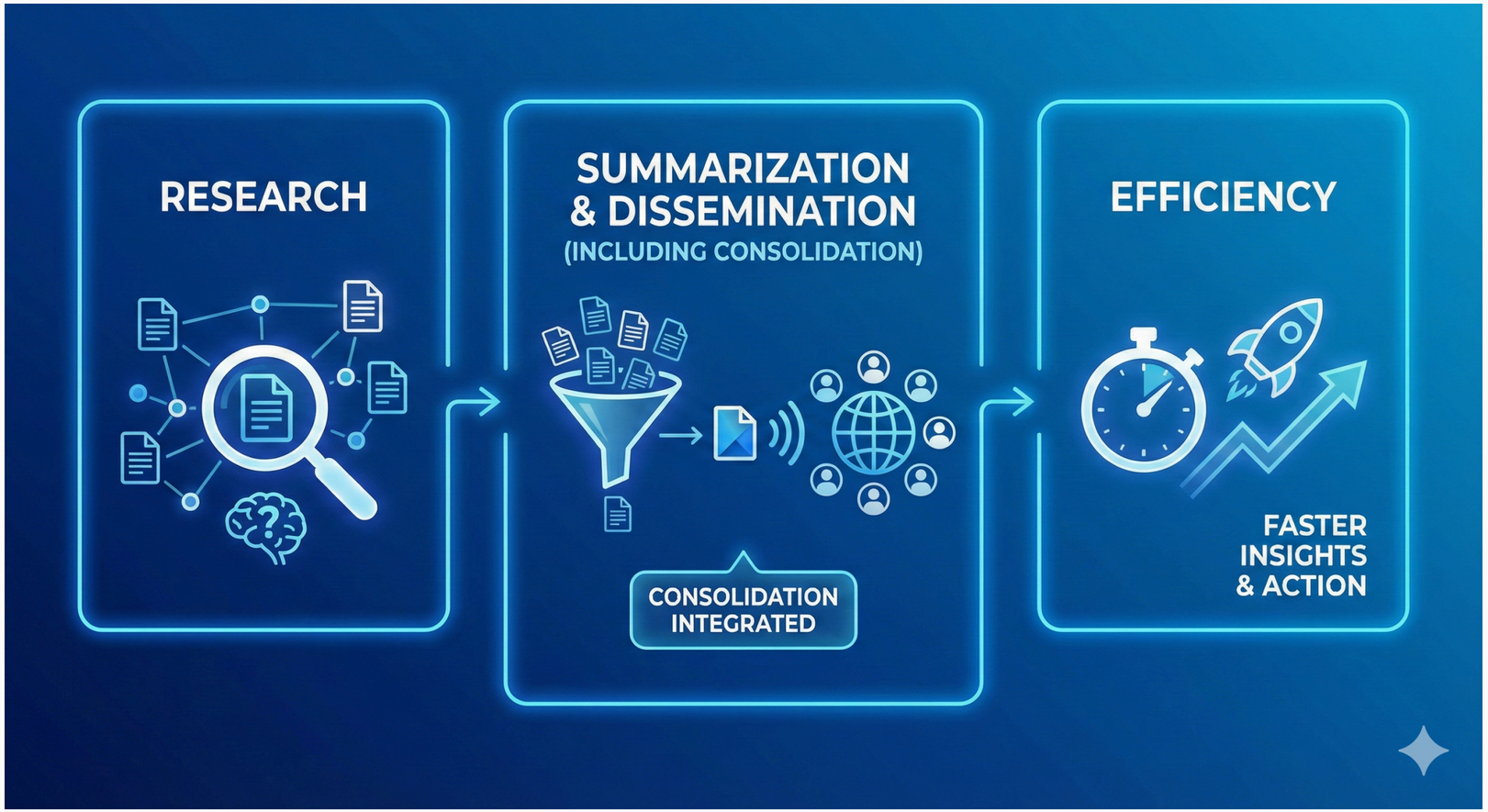
Note: NLP is a **problem domain** that leverages AI. Generative AI and Agentic AI are 'technologies'. They are systems built using the techniques.

Everyday Examples of AI



Prompt: Create an image that has images representing these ideas. Typing suggestions on your phone; the phone learns from previous words to predict the next one. Photo apps recognizing faces, even if the person changes their hair or angle. Fraud alerts from your bank, which notice unusual patterns in your spending. Chatbots or voice assistants, which understand your question and generate a helpful response.

Organizing AI by Use



Prompt: Create an image that has images representing these aspects of AI - Research, Consolidation,Summarization and Dissemination, Efficiency.

Examples: AI for Research in Public Health

Projects focused on developing predictive models, identifying risk factors, and generating new insights

#	Title/Project	Institution/Organization	Year	Description
1	US Diabetes Risk Prediction Using BRFSS Data	Multiple institutions	2024-2025	Analysis of 253,680 adult respondents using machine learning models (Extra Trees Classifier, Random Forest, XGBoost) to identify influential predictors of diabetes likelihood
2	Social Media Analysis for Mental Health	Research study	2022	Research showing social media discussions can predict mental health consultations on US college campuses, enabling earlier intervention

Examples: AI for Consolidation, Summarization & Dissemination in Public Health

Projects focused on extracting insights from large datasets, analyzing narratives, and synthesizing information for public health decision-making

#	Title/Project	Institution/Organization	Year	Description
3	BlueDot and HealthMap Surveillance Systems	BlueDot, HealthMap	Ongoing	AI-based epidemiological surveillance systems providing early warnings and real-time disease outbreak monitoring and visualization
4	Transforming Public Health Practice with Generative AI	US Health Departments	2024	Exploring how AI supports core public health functions including communications, organizational performance, and novel insights for decision-making

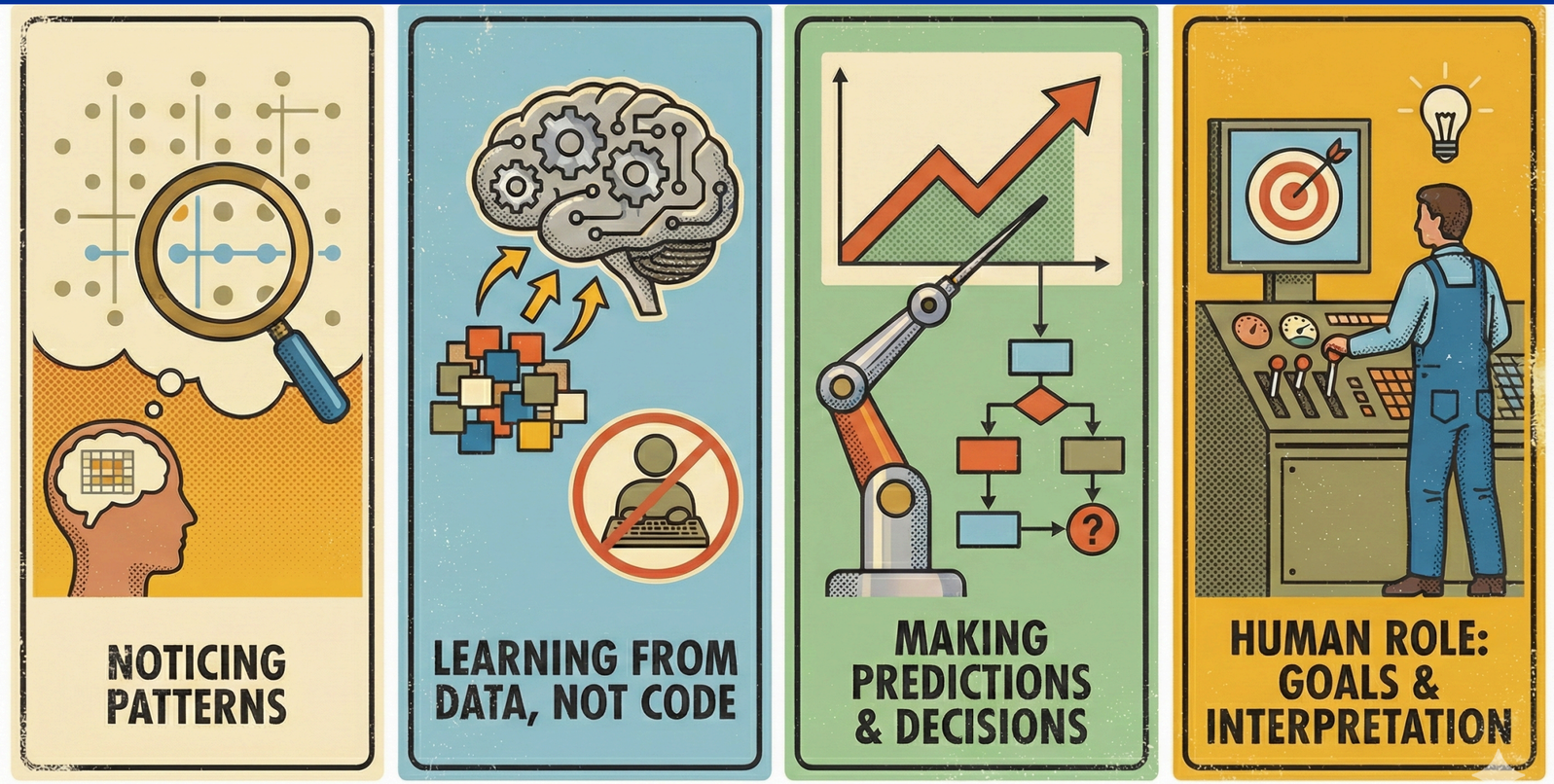
Examples: AI for Efficiency & Automation in Public Health

Projects focused on automating routine tasks, speeding up processes, and improving operational efficiency

#	Title/Project	Institution/Organization	Year	Description
5	Conversational AI for Vaccine Communication	Research review	2023	Systematic review showing chatbot studies measuring influence on vaccine attitudes found evidence of positive effects with no "backfire effects"
6	Motivational Interviewing-Oriented AI Digital Assistant	Research study (Hong Kong & US)	2022-2024	RCT with 177 participants testing an AI-driven chatbot with motivational interviewing techniques

Note Some projects could fit into multiple categories. Classification is based on the primary application or objective of each initiative.

The key idea: AI is not one thing; it is a collection of methods that let computers:



Prompt: Create an image containing images that represent these statements. AI is not one thing; it is a collection of methods that let computers: Notice patterns in data (like humans noticing trends). Learn from those patterns (without a programmer giving step-by-step instructions). Use that learning to make predictions or decisions. Humans still define the goal, set limits, and interpret the results.

A helpful analogy

- Think of AI like teaching a new employee:
  - You don’t give them a rulebook for every situation.
  - Instead, they watch examples, practice, and get feedback.
  - Over time, they get better and faster at handling similar tasks.
  - AI learns the same way, but with data instead of experience.

What AI is not

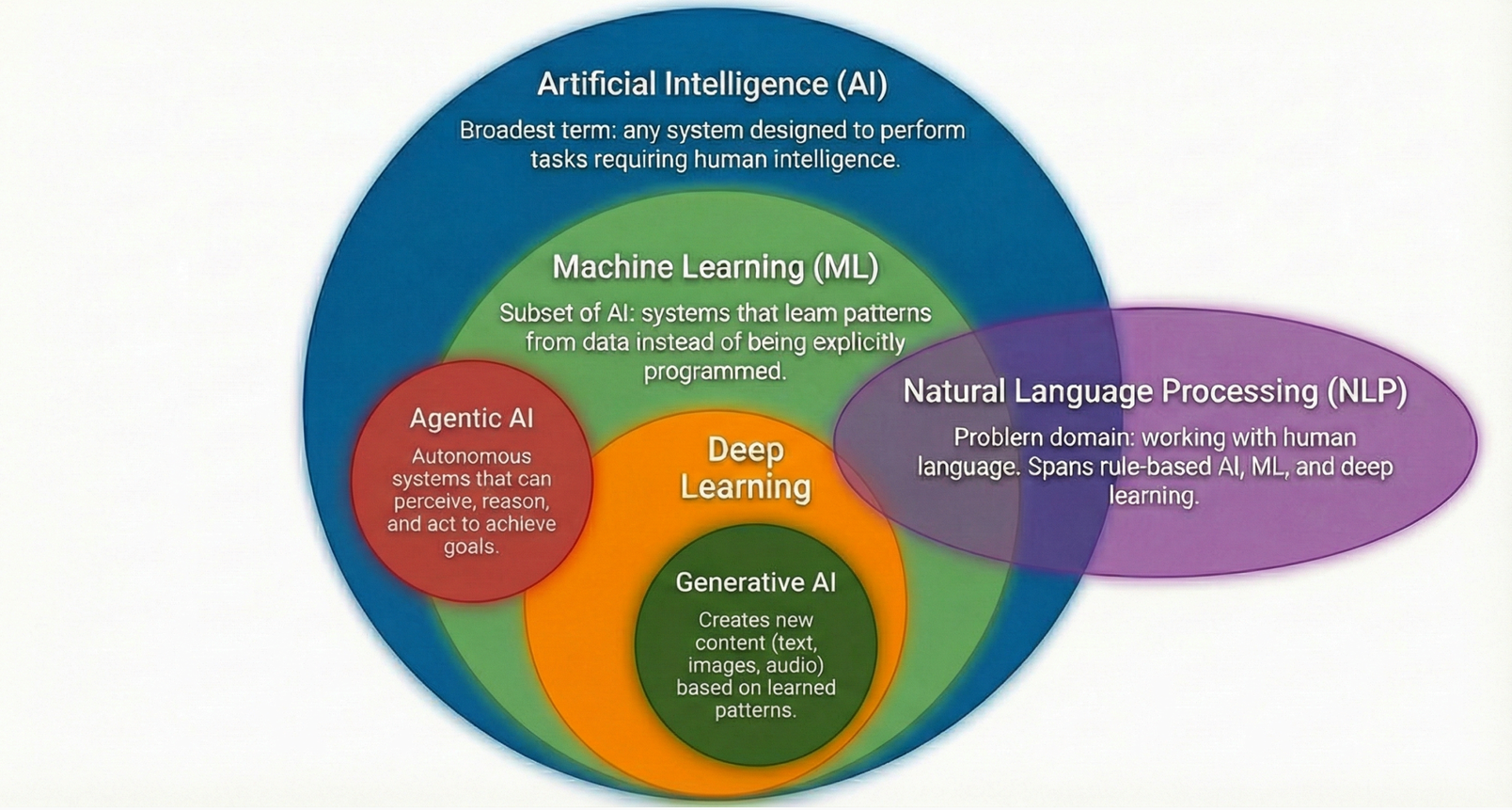
- AI is not magic; it is math and pattern recognition.
- AI is not a full replacement for human judgment; it needs oversight.
- AI does not understand the world the way humans do, even when it generates fluent language.

Major AI Subfields

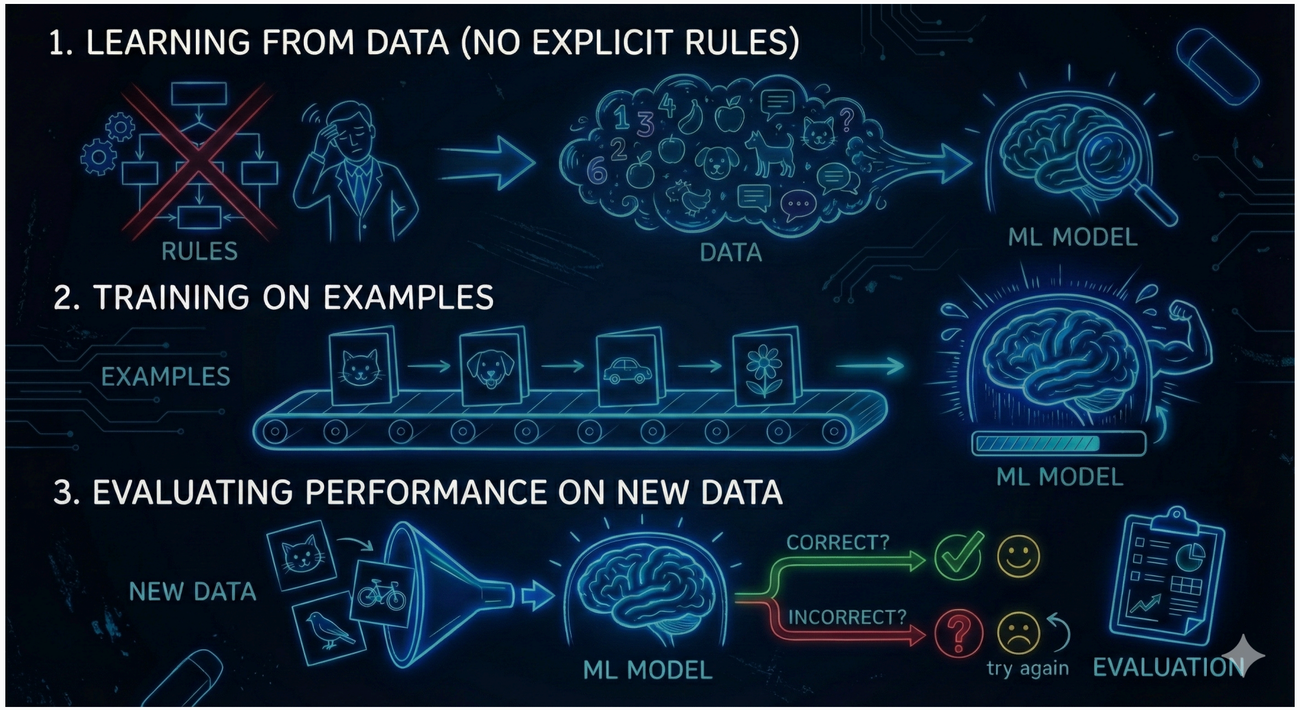
Concept	Description
Artificial Intelligence (AI)	**Broadest term**: any system designed to perform tasks requiring human intelligence; includes both rule-based systems and learning-based systems; the umbrella category.
Machine Learning (ML)	**Subset of AI**: systems that learn patterns from data instead of being explicitly programmed. Requires structured data and engineered features; often interpretable. Examples include decision trees, random forests, logistic regression, and support vector machines.
Deep Learning	**Subset of ML**: neural networks with many layers that learn representations automatically. Works with unstructured data, needs large datasets and computation, and is often a “black box.” Examples include CNNs, RNNs, and transformers.
Generative AI	**Subset of Deep Learning**: systems that create new content (text, images, code, audio, etc.) by learning patterns from data. Typically built using deep learning, especially transformer models, but not inherently autonomous; produces outputs in response to prompts.
Agentic AI	**System behavior**: AI that can take actions toward a goal. Uses planning, decision-making, tool use, memory, and feedback loops. May incorporate generative AI as one component, but adds autonomy (deciding what to do next) and the ability to trigger actions, workflows, or operations.
Natural Language Processing (NLP)	**Problem domain**: working with human language. Spans rule-based AI, ML, and deep learning. The method used depends on task, data, and compute resources.

Foundational Concepts: Machine Learning Overview

Machine learning (ML) is the engine that powers modern AI. Deep learning, NLP, and other advanced techniques all build on ML's core ideas. Understanding ML is the foundation for understanding all of modern AI.



Machine Learning Core Concepts

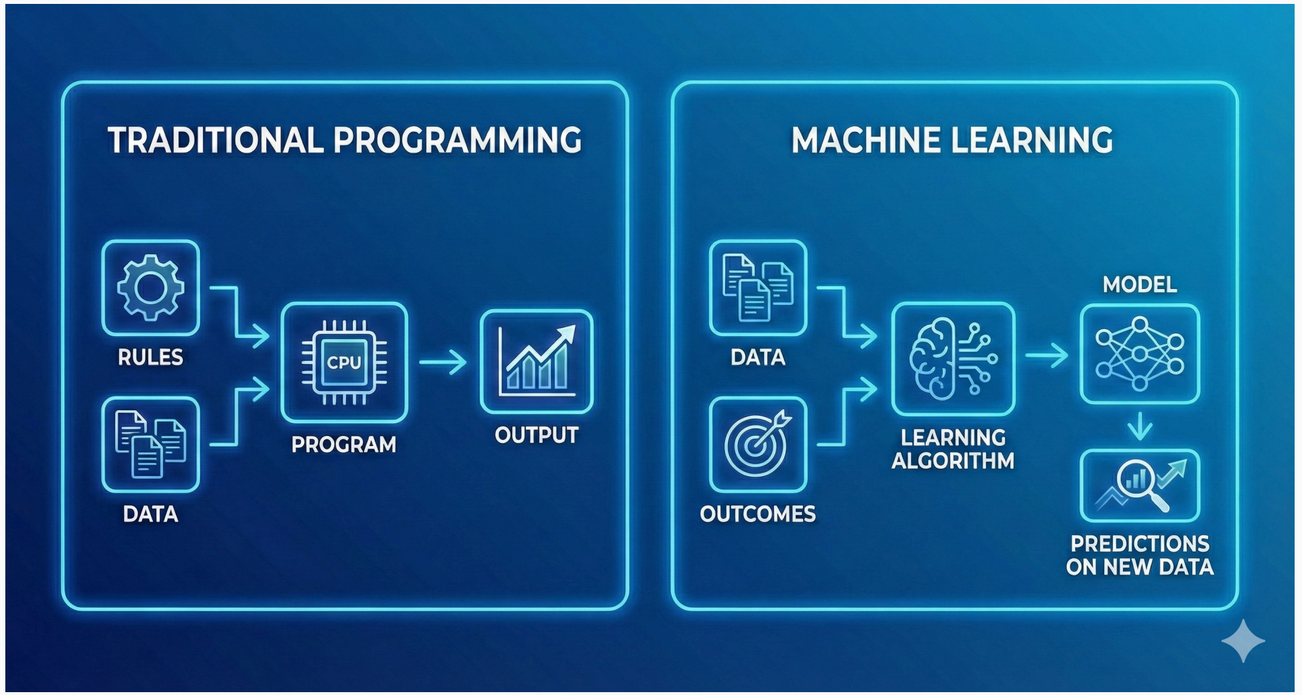


Prompt: Create an image that portrays the three core concepts of Machine learning. Don’t be word heavy. Use images. Create the image in a 'hand drawing' style. Use this as a basis for the image.ML provides the three core concepts that everything else builds on: Learning from data (not from explicit rules), Training on examples, Evaluating performance on new data. Follow-up: Can you make the images a little opaque so that the 1., 2. 3. text stands out more

---

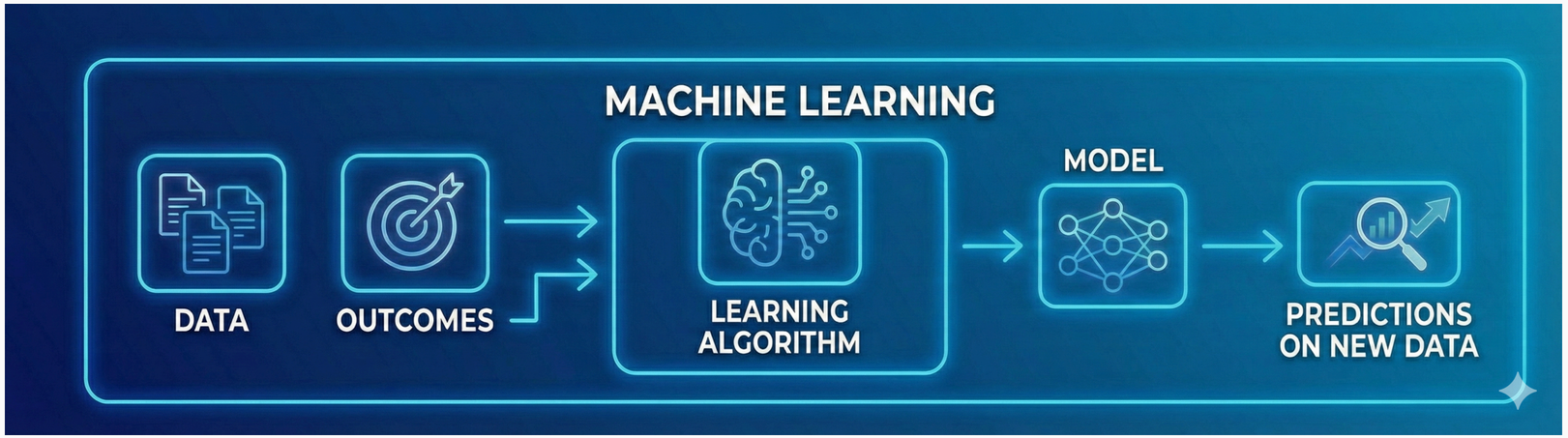
The Core Idea: Learning from Data

Machine learning is a way for computers to **learn patterns from data**, rather than being explicitly programmed with rules.



Prompt: I am discussing machine learning. create a Two-panel diagram: Left: Traditional programming (Rules + Data → Output) Right: Machine learning (Data + Outcomes → Learning Algorithm → Model → Predictions on new data). Use the attached file as a style guide. Follow-up: Leave the title of each box but remove the subtitle that describes the flow. Leave all else as is.

The Learning Process



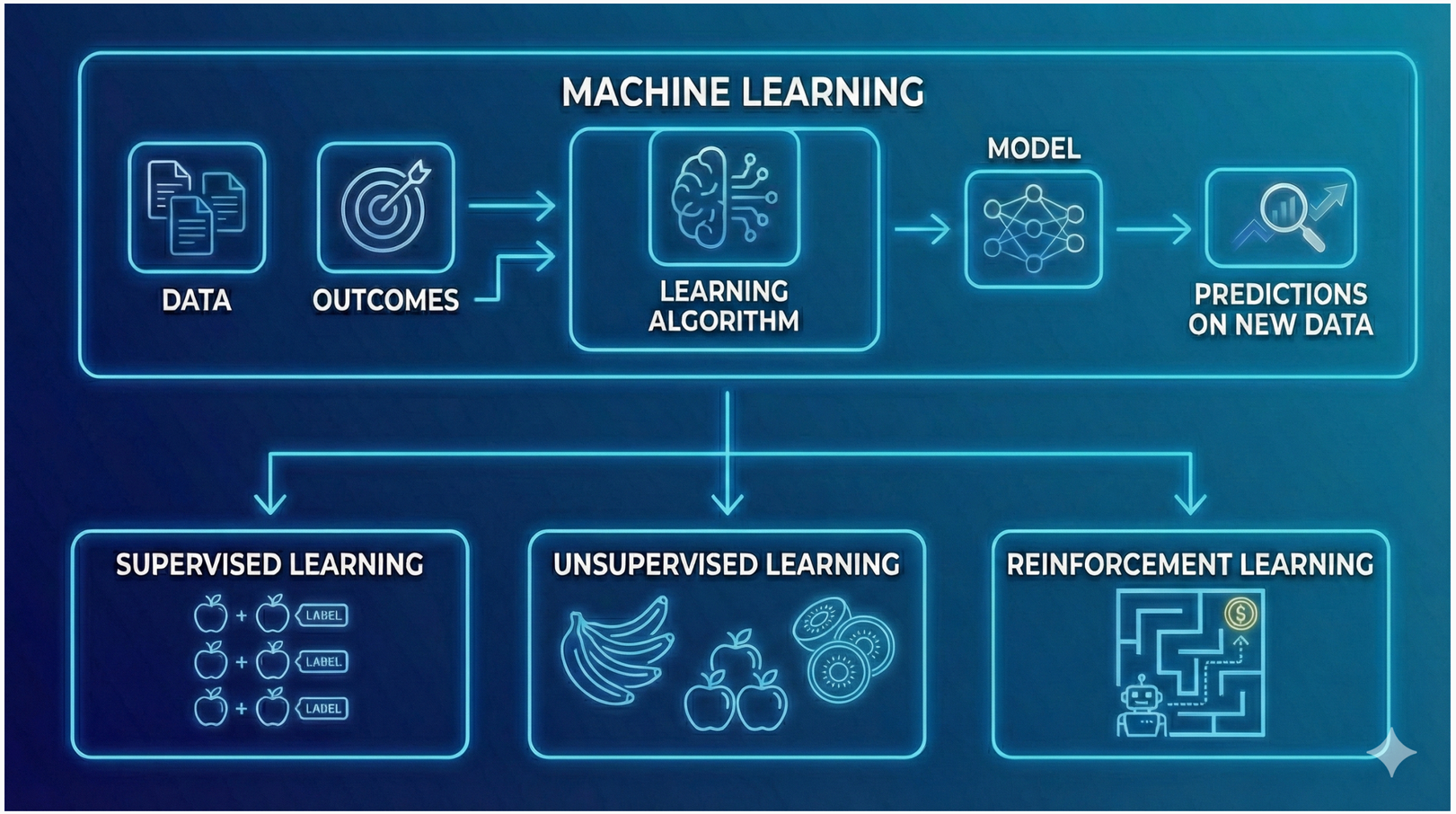
How learning works (intuition, not math):

- The algorithm makes a guess, checks how wrong it is, and adjusts.
- It repeats this process—thousands or millions of times—until it minimizes errors.
- Think of it like learning to throw darts: we throw, see where it lands, adjust our aim, throw again.

The train/test split:

- We can’t test learning on the same examples we learned from (that’s just memorization).
- So we split the data: learn from some (training data), test on the rest (test data).
- Performance on test data tells us if the model actually learned generalizable patterns.

ML Learning Paradigms



Supervised Learning

In supervised learning, the algorithm learns from **labeled** examples - data where we know the right answer.

encounter_id	patient_nbr	race	gender	age	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	A1Cresult	insulin	change	diabetesMed	readmitted
2278392	8222157	Caucasian	Female	[0-10)	1	41	0	1	0	0	0	None	No	No	No	NO
149190	55629189	Caucasian	Female	[10-20)	3	59	0	18	0	0	0	None	Up	Ch	Yes	>30
64410	86047875	AfricanAmeric	Female	[20-30)	2	11	5	13	2	0	1	None	No	No	Yes	NO
500364	82442376	Caucasian	Male	[30-40)	2	44	1	16	0	0	0	None	Up	Ch	Yes	NO
16680	42519267	Caucasian	Male	[40-50)	1	51	0	8	0	0	0	None	Steady	Ch	Yes	NO
35754	82637451	Caucasian	Male	[50-60)	3	31	6	16	0	0	0	None	Steady	No	Yes	>30
55842	84259809	Caucasian	Male	[60-70)	4	70	1	21	0	0	0	None	Steady	Ch	Yes	NO

- "Supervised" because the learning is being supervised with correct answers.
- The data includes both inputs (features) and the outcome (label) we want to predict.
- The algorithm learns the relationship between inputs and outcomes.
- Once trained, the model can predict outcomes for new cases when the answer is unknown.

Classification examples:

- Will this patient be readmitted? (Yes / No)
- Is this email spam? (Spam / Not spam)
- What type of request is this? (Complaint / Question / Compliment)

Regression examples:

- How many people will enroll next month? (a number)
- How long will this patient stay in the hospital? (days)
- What will the temperature be tomorrow? (degrees)

Unsupervised Learning

In unsupervised learning, there are no labels. The algorithm finds structure or patterns in the data on its own.

- "Unsupervised" because there’s no right answer to learn from.
- The algorithm explores the data to find natural groupings, patterns, or simplifications.
- Useful when we don’t know what we’re looking for, or when we want to understand the structure of your data.

Table 1. Two Main Applications

Application	**What It Does**	**Example**
Clustering	Groups similar cases together	Which patients have similar profiles? What types of complaints do we receive?
Dimensionality Reduction	Simplifies data with many variables	We have 200 variables—which combinations capture most of the information?

Clustering & Dimensionality Reduction

Clustering:

- Finds natural groupings in data without being told what the groups are.
- You might discover that your population falls into 4–5 distinct segments you didn’t know existed.
- Useful for: customer segmentation, identifying outbreak clusters, grouping similar facilities.

Dimensionality reduction:

- When you have too many variables to work with, this technique reduces them to a smaller set.
- It finds combinations of variables that capture most of the important variation.
- Think of it as data simplification: 50 variables become 5 composite variables that retain most of the signal.
- Useful for: simplifying complex data before analysis, identifying which factors matter most.

Reinforcement Learning

Reinforcement learning is a third paradigm where an agent learns by taking actions and receiving feedback (rewards or penalties).

- Different from supervised (no labeled examples) and unsupervised (not just finding structure).
- An agent interacts with an environment, takes actions, and learns from the results.
- Good outcomes = rewards; bad outcomes = penalties.
- Over time, the agent learns a strategy (policy) that maximizes rewards.

Classic examples:

- Game-playing AI (learns to win by playing millions of games)
- Robotics (learns to walk by trial and error)
- Recommendation systems (learns what to show you based on your clicks)

Why it’s not our focus:

- Less common in public health and government contexts (so far).
- Requires an environment where the agent can take actions and observe results.
- Most agency AI applications are supervised or unsupervised.

Why This Matters for Everything Else

The learning paradigms are core concepts. **Learning from data**, supervised vs. unsupervised, training and testing—appear in every AI application you'll encounter.

- Deep learning uses the same paradigms:
  - Supervised deep learning: image classification, speech recognition
  - Unsupervised deep learning: finding patterns in text or images
  - Reinforcement learning: game-playing AI, robotics
- NLP uses the same paradigms:
  - Supervised: classify emails as spam/not spam, sentiment analysis
  - Unsupervised: topic modeling, clustering documents by similarity
- This vocabulary transfers everywhere:
  - Training data, test data
  - Labels, features
  - Classification, regression, clustering
  - Overfitting (learning too much from training data)

The Three Paradigms at a Glance

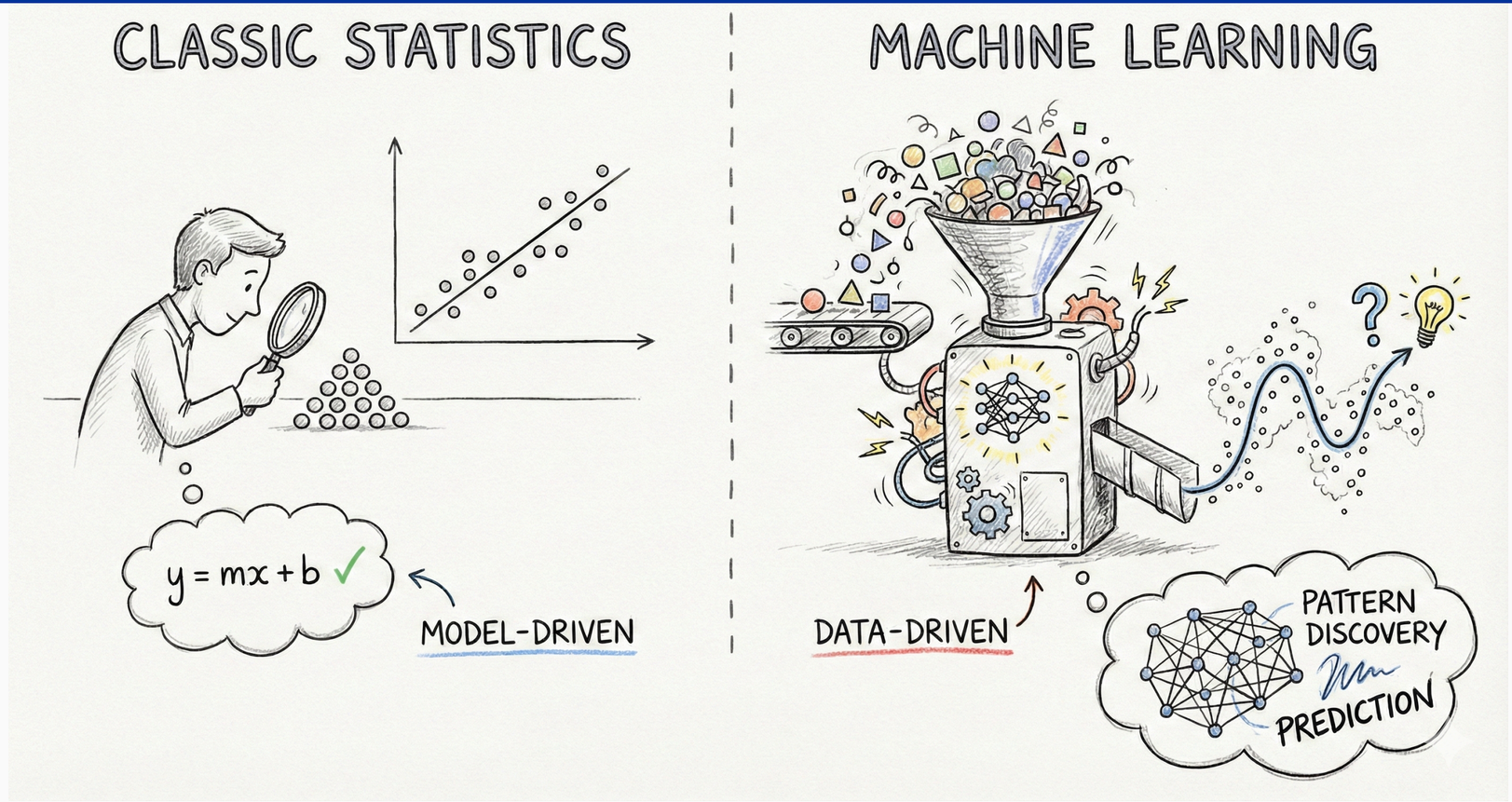
A summary reference for the three learning paradigms.

Paradigm	What It Needs	What It Does	Common Tasks	Public Health Example
Supervised	Labeled data (inputs + outcomes)	Learns to predict outcomes	Classification, Regression	Predict readmission risk
Unsupervised	Unlabeled data (inputs only)	Finds structure or patterns	Clustering, Dimensionality reduction	Identify patient segments
Reinforcement	Environment + feedback	Learns strategy through trial and error	Game playing, Robotics	(Less common in public health)

Machine Learning Summary

Component	Key Takeaway
ML is the foundation	Everything else in AI builds on ML’s core concepts
Learning from data	ML finds patterns through iteration, not explicit rules
Supervised learning	Learns from labeled examples; classification and regression
Unsupervised learning	Finds structure without labels; clustering and dimensionality reduction
Reinforcement learning	Learns from feedback; less common in public health
Why this matters	These concepts recur throughout deep learning and NLP
Three paradigms at a glance	Summary reference for quick recall

Foundational Concepts: How Does AI Differ from Classic Inferential Statistics



The Core Distinction

Classical Statistics	Machine Learning
Asks "WHY?"	Asks "WHAT NEXT?"
Inference & Understanding	Prediction & Generalization

About the Dataset

This walkthrough uses a real-world clinical dataset derived from the Health Facts database (Cerner Corporation), containing approximately 5,000 hospital encounters for patients with diabetes. The dataset was originally used in a published study examining whether HbA1c measurement affects hospital readmission rates.

encounter_ic	patient_nbr	race	gender	age	time_in_hosp	num_lab_pro	num_proced	num_medica	number_outp	number_eme	number_inpa	diag_1	number_diag	max_glu_ser	A1Cresult	insulin	readmitted
2278392	8222157	Caucasian	Female	[0-10)	1	41	0	1	0	0	0	250.83	1	None	None	No	NO
149190	55629189	Caucasian	Female	[10-20)	3	59	0	18	0	0	0	276	9	None	None	Up	>30
64410	86047875	AfricanAmeri	Female	[20-30)	2	11	5	13	2	0	1	648	6	None	None	No	NO
500364	82442376	Caucasian	Male	[30-40)	2	44	1	16	0	0	0	8	7	None	None	Up	NO
16680	42519267	Caucasian	Male	[40-50)	1	51	0	8	0	0	0	197	5	None	None	Steady	NO
35754	82637451	Caucasian	Male	[50-60)	3	31	6	16	0	0	0	414	9	None	None	Steady	>30
55842	84259809	Caucasian	Male	[60-70)	4	70	1	21	0	0	0	414	7	None	None	Steady	NO
63768	114882984	Caucasian	Male	[70-80)	5	73	0	12	0	0	0	428	8	None	None	No	>30
12522	48330783	Caucasian	Female	[80-90)	13	68	2	28	0	0	0	398	8	None	None	Steady	NO
15738	63555939	Caucasian	Female	[90-100)	12	33	3	18	0	0	0	434	8	None	None	Steady	NO
28236	89869032	AfricanAmeri	Female	[40-50)	9	47	2	17	0	0	0	250.7	9	None	None	Steady	>30
36900	77391171	AfricanAmeri	Male	[60-70)	7	62	0	11	0	0	0	157	7	None	None	Steady	<30
40926	85504905	Caucasian	Female	[40-50)	7	60	0	15	0	1	0	428	8	None	None	Down	<30
42570	77586282	Caucasian	Male	[80-90)	10	55	1	31	0	0	0	428	8	None	None	Steady	NO
62256	49726791	AfricanAmeri	Female	[60-70)	1	49	5	2	0	0	0	518	8	None	None	Steady	>30
73578	86328819	AfricanAmeri	Male	[60-70)	12	75	5	13	0	0	0	999	9	None	None	Up	NO
77076	92519352	AfricanAmeri	Male	[50-60)	4	45	4	17	0	0	0	410	8	None	None	Steady	<30
84222	108662661	Caucasian	Female	[50-60)	3	29	0	11	0	0	0	682	3	None	None	No	NO
89682	107389323	AfricanAmeri	Male	[70-80)	5	35	5	23	0	0	0	402	9	None	None	Steady	>30
148530	69422211	?	Male	[70-80)	6	42	2	23	0	0	0	737	8	None	None	Steady	NO

Asking Questions of Our Diabetes Data

'Does measuring HbA1c significantly reduce the probability of 30-day readmission, controlling for patient age, length of stay, and number of diagnoses?' The goal is to understand the **mechanism**. Does paying attention to glucose control during hospitalization actually matter?
'Given a patient’s age, length of stay, lab results, medications, and diagnoses, can readmission within 30 days be predicted accurately?' The goal is a model that **works** on the next patient who walks in the door."

Logistic Regression as a Bridge

Same algorithm, different goals → different workflows

Element	Traditional Statistical Approach	Machine Learning Approach
Primary Goal	**Inference & Understanding** Understand relationships between variables	**Prediction & Generalization** Accurate predictions on unseen data
Key Question	**"Does X affect Y?"** Does HbA1c measurement reduce readmission?	**"Can I predict Y from X?"** Which patients will be readmitted?
Data Usage	**Full Dataset** Use all data to estimate parameters and test hypotheses	**Train / Validation / Test Split** 70% train, 15% validation, 15% test; honest evaluation on held-out data
Variable Selection	**Theory-Driven** Select variables based on clinical hypotheses; each must be interpretable	**Data-Driven** Include anything that improves prediction; interpretation optional
Validation & Output	**p-values, Confidence Intervals, Odds Ratios** "OR = 0.85, 95% CI: 0.74–0.97, p = 0.02" Is the effect real? How large?	**AUC, Precision, Recall on Test Set** "AUC = 0.72, Precision = 0.34, Recall = 0.58" How well does it predict new cases?

Why the Workflow Differs - Splitting the data

- In statistics:
  - Typically use all data to estimate parameters and test hypotheses—like using all 5,000 encounters to estimate the odds ratio for HbA1c measurement.
- In ML:
  - Intentionally hide some patients from the model during training.
  - Why? Because the workflow requires an honest test of whether the model generalizes. If the model is validated on the same patients it was trained on, it might just be measuring how well the model memorized those specific patients—not how well it learned the underlying pattern.

Why the Workflow Differs - Including more variables in ML

- In the original study (inferential statistics):
  - Researchers carefully selected variables based on clinical relevance—age, race, admission type, primary diagnosis, HbA1c.
- In ML:
  - The workflow might utilize everything: number of lab procedures, number of medications, all three diagnosis codes, admission source.
  - Why? Because the effort is not trying to interpret each coefficient. It is trying to capture any signal that helps predict the outcome. A variable that’s uninterpretable might still improve predictions.

Why the Workflow Differs - Using different metrics

- The original study:
  - Reports an odds ratio of 0.85 with a p-value. This indicates the association is likely real and its magnitude.
  - It doesn’t tell you how well you can actually predict readmission for a specific patient.
- The ML process provides:
  - AUC, precision, and recall as indicators of how well the predictions match reality.
  - This is what matters if your goal is to flag high-risk patients before discharge.

The Bias-Variance Tradeoff

- **Bias:** Error from oversimplifying. If your model is too simple, it misses real patterns. (Underfitting)
  - Just using age—might miss important patterns related to medications or diagnoses. That’s high bias; it underfits.
- **Variance:** Error from overcomplicating. If your model is too complex, it fits noise and doesn’t generalize. (Overfitting)
  - Using every variable and their interactions, memorizing the specific combination of features for each of the 5,000 patients. The model might perfectly 'predict' readmission for patients it’s seen, but fail completely on new patients. That’s high variance; it overfits.

The ML workflow balances bias and variance by using held-out data to detect overfitting, adjusting model complexity and regularization through hyperparameter tuning, and evaluating generalization via validation and cross-validation, often with ensembles and early stopping to stabilize performance.

Balancing Bias and Variance Summary Table

Concept	What It Does	Diabetes Example
Regularization + Hyperparameter Tuning	Penalizes complexity; tune penalty strength on validation data	Test λ = 0.001, 0.01, 0.1; pick λ = 0.1 because it gives best validation AUC
Validation	Tests if patterns generalize to unseen data	Train on 3,500 patients, test on 750; validation AUC = 0.68 is our honest estimate
Cross-Validation	Rotates train/validation splits for stable estimates	5-fold CV gives AUC = 0.71, 0.73, 0.70, 0.72, 0.74 → average 0.72
Ensembles	Combines multiple models for better accuracy	Random Forest (100 trees) achieves AUC = 0.72 vs. single tree at 0.65
Early Stopping	Stops training when validation performance peaks	Stop at round 150 (validation AUC = 0.72) before overfitting sets in

Techniques Working Together

These techniques work as a system to prevent overfitting:

1. **Split your data** so you have an honest test (validation)
2. **Use cross-validation** to get stable performance estimates
3. **Add regularization** to prevent the model from getting too complex
4. **Tune hyperparameters** to find the right amount of regularization
5. **Use ensembles** to combine multiple models and reduce variance
6. **Apply early stopping** to halt training at the right moment

The goal is always the same: a model that performs well on **new patients**, not just the ones it trained on.

Sometimes the Boundaries Blur

- The original HbA1c study used logistic regression
- The researchers did care about prediction (readmission rates) but framed it as inference (does HbA1c measurement matter?)
- Some ML practitioners care about interpretability and coefficient meaning
- Modern practice often blends both traditions

The distinction should not be about which is better. It’s needs to be about **which approach is appropriate for the question**. 'Should we implement a policy requiring HbA1c measurement?', needs the statistical approach. We need to know if it actually helps. 'Can you flag patients at high risk of readmission so we can intervene?', can use the ML approach for accurate predictions.

Optional: Code Walkthrough Statistical approach:

```
# Load data
diabetes <- read.csv("diabetes_for_ml.csv")

# Create binary readmission variable
diabetes$readmit_30 <- ifelse(diabetes$readmitted == "<30", 1, 0)

# Fit logistic regression for inference
model_stats <- glm(readmit_30 ~ A1Cresult + age + time_in_hospital + number_diagnoses,
                   data = diabetes, family = binomial)

# Examine coefficients, p-values, odds ratios
summary(model_stats)
exp(coef(model_stats))  # Odds ratios
confint(model_stats)     # Confidence intervals
```

Optional: Code Walkthrough ML approach:

```
library(caret)

# Split data
set.seed(42)
train_index <- createDataPartition(diabetes$readmit_30, p = 0.7, list = FALSE)
train_data <- diabetes[train_index, ]
test_data <- diabetes[-train_index, ]

# Train model with cross-validation
model_ml <- train(as.factor(readmit_30) ~ A1Cresult + age + time_in_hospital +
                  number_diagnoses + num_lab_procedures + num_medications,
                  data = train_data, method = "glm", family = "binomial",
                  trControl = trainControl(method = "cv", number = 5))

# Evaluate on test set
predictions <- predict(model_ml, test_data)
confusionMatrix(predictions, as.factor(test_data$readmit_30))
```

The Confusion Matrix: Where It All Starts

Accuracy alone isn’t enough—we need to understand the **types** of errors the model makes.

When our model predicts "will be readmitted" or "will not be readmitted" for each patient, there are four possible outcomes:

	Actually Readmitted (Yes)	**Actually Not Readmitted (No)**
Predicted Readmitted (Yes)	✅ **True Positive (TP)** Model said yes, patient was readmitted	❌ **False Positive (FP)** Model said yes, but patient was NOT readmitted
Predicted Not Readmitted (No)	❌ **False Negative (FN)** Model said no, but patient WAS readmitted	✅ **True Negative (TN)** Model said no, patient was not readmitted

Example with our diabetes data:

Suppose we test our model on 1,000 patients. The results:

	Actually Readmitted	**Actually Not Readmitted**
Predicted Readmitted	80 (TP)	120 (FP)
Predicted Not Readmitted	70 (FN)	730 (TN)

- 150 patients were actually readmitted (80 + 70)
- 850 patients were not readmitted (120 + 730)
- The model flagged 200 patients as high-risk (80 + 120)

Putting It All Together

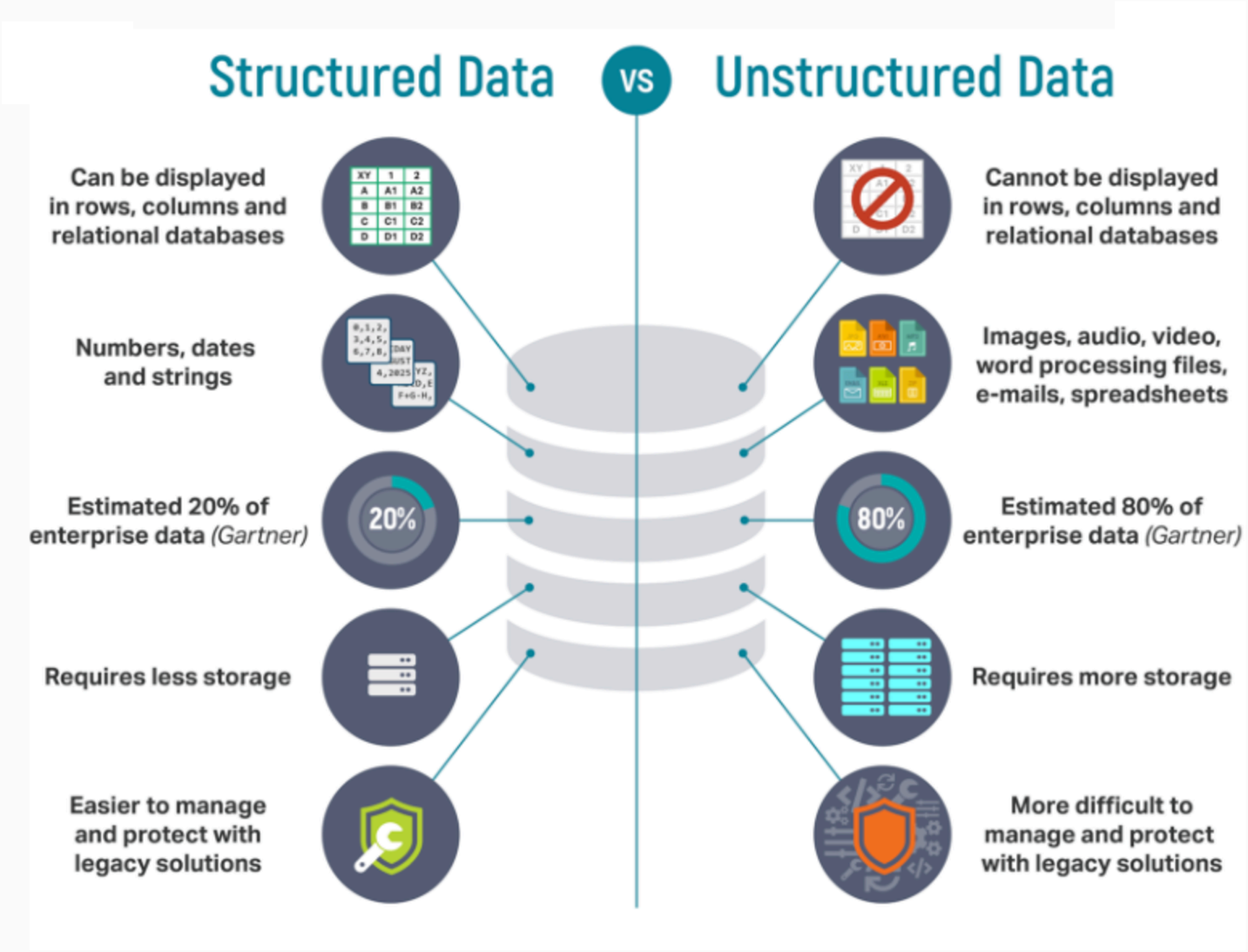
Metric	Question It Answers	Formula	Our Diabetes Model
Precision	Of those we flagged, how many were actually readmitted?	TP ÷ (TP + FP)	40%
Recall	Of those actually readmitted, how many did we catch?	TP ÷ (TP + FN)	53%
AUC	Overall, how well does the model rank patients by risk?	Area under ROC curve	0.72

Questions to Ask About Any Model

When someone presents model performance, ask:

1. **What’s the baseline?** (If 15% of patients are readmitted, a model that always predicts "no" gets 85% accuracy—but catches no one)
2. **What’s the AUC?** (Is the model actually learning something, or just guessing?)
3. **What threshold was used?** (Performance depends on where you draw the line)
4. **What’s the precision-recall tradeoff?** (What are you willing to accept in terms of false positives vs. missed cases?)
5. **Do we have the capacity to act?** (If we can only intervene with 50 patients, a model that flags 500 isn’t useful without prioritization)

Foundational Concepts: Data Types



Structured vs. Unstructured Data

Key point: Data comes in different forms, and the form affects how easily ML can work with it.

- **Structured data:** Organized in rows and columns, where each column is a defined field with a consistent format.
  - Examples: Spreadsheets, databases, claims data, enrollment records
  - Each row is a case/record; each column is a variable
  - ML algorithms can work with this directly
- **Unstructured data:** No predefined format; the information is embedded in content that requires processing to extract.
  - Examples: Free-text notes, emails, PDFs, images, audio recordings
  - The meaning isn’t in neat columns—it’s in the words, pixels, or sounds

Implication for AI

Data Type	What It Means for ML
Structured	Most ML algorithms work directly; relatively straightforward
Unstructured	Requires extra processing before ML can use it

Extra processing for unstructured data:

- Text → Natural Language Processing (NLP) to extract meaning, categorize, or convert to structured features
- Images → Computer vision to identify objects, patterns, or classifications
- Audio → Speech recognition to transcribe, then NLP to analyze

Unstructured data isn’t unusable—but it adds complexity, time, and specialized techniques. If someone proposes an AI project using free-text case notes or scanned documents, that’s a signal the project will require more than basic ML.

Labeled vs. Unlabeled Data

Key point: A "label" is the outcome you're trying to predict—not the column headers. Whether you have labels determines what kind of ML you can do.

Common confusion: People hear "labeled data" and think it means the data has variable names or column headers. That's not what it means.

What "label" actually means:

- The label is the outcome, target, or answer you want the model to learn to predict.
- In a dataset of patients, the label might be: "Was this patient readmitted within 30 days? Yes/No"
- In a dataset of emails, the label might be: "Is this spam? Yes/No"
- The label is what you're trying to predict for future cases where you don't know the answer yet.

Labeled data:

- You have both the inputs (features) AND the outcome (label) recorded for historical cases.
- Example: Patient records where you know who was readmitted and who wasn't.
- This enables supervised learning—the algorithm can learn the relationship between inputs and outcomes.

Unlabeled data:

- You have inputs but NOT the outcome.
- Example: Patient records, but readmission status was never tracked.
- This limits you to unsupervised learning—finding patterns or structure, but not predicting a specific outcome.

Where Labels Come From

Source	Example	Consideration
Recorded in normal operations	Claims data includes whether patient was readmitted	Best case—labels already exist
Created from historical decisions	"Fraud" label based on cases that were investigated	Labels reflect past decisions, which may be biased
Manually created for the project	Staff review 1,000 cases and label them	Time-consuming and expensive; need enough labeled examples
Doesn’t exist	"Success" was never defined or tracked	Can’t do supervised learning without creating labels first

Key insight:
Labels don’t appear magically—someone or something recorded them. And the way labels were created matters. If "fraud" labels come from who got investigated (not who actually committed fraud), the model learns to predict who gets investigated, not who commits fraud.

Connection to Learning Paradigms

Data Situation	Learning Paradigm	What You Can Do
Have labels	Supervised	Predict outcomes (classification, regression)
No labels	Unsupervised	Find patterns, groups, structure (clustering, dimensionality reduction)

Volume: How Much Data Do You Have?

Key point: ML learns from examples. More examples generally means better learning—but "enough" depends on the problem.

- ML algorithms find patterns by seeing many examples. Too few examples, and the algorithm can’t learn reliably.
- There’s no magic number, but rough guidelines help set expectations.

Situation	Rough Minimum	Notes
Simple problem, structured data	Hundreds to low thousands	Logistic regression, simple decision trees
Moderate complexity	Thousands to tens of thousands	Random forests, gradient boosting
Complex patterns, many variables	Tens of thousands or more	May need more sophisticated approaches
Deep learning (images, text)	Often tens of thousands to millions	Data-hungry techniques

Why volume matters

- **Too little data:** Model may memorize the training examples rather than learning generalizable patterns (overfitting).
- **Rare outcomes:** If you’re predicting something that happens 1% of the time, you need enough data to have sufficient examples of that rare event.
  - Example: If readmission happens 10% of the time and you have 500 records, you only have ~50 readmission cases to learn from.

Questions to ask about volume:

- How many records do we have?
- How many examples of the outcome we’re trying to predict?
- Is the outcome common or rare?

Key insight:

"We have a lot of data" isn’t enough. You need enough examples of the thing you’re trying to predict.

Quality: Is the Data Any Good?

Key point: Data quality issues can undermine or derail an AI project. Garbage in, garbage out.

- Real-world data is messy. It was usually collected for operational purposes, not for ML.
- Quality issues are the norm, not the exception. The question is how severe they are and whether they can be addressed.

Common Quality Issues

Issue	Example	Why It Matters
Missing values	30% of records have no income data	Model may learn wrong patterns or exclude too many cases
Inconsistent coding	"Diabetes," "DM," "Type 2 DM," "diabetic" all mean the same thing	Model treats them as different; patterns get diluted
Data entry errors	Birthdate of 1/1/1900; weight of 5,000 lbs	Outliers distort learning
Duplicates	Same patient appears multiple times with different IDs	Inflates apparent data volume; may leak information
Outdated information	Address from 10 years ago	May not reflect current reality
Inconsistent definitions	"Enrollment date" means different things in different systems	Apples-to-oranges comparisons

Questions to ask about quality:

- How complete is the data? What's missing?
- Are fields coded consistently?
- Are there known data quality issues?
- When was the data last validated or cleaned?

Accessibility: Can You Actually Get to It?

Key point: Data that exists but can’t be accessed is the same as data that doesn’t exist. Accessibility is often the hidden blocker.

- Just because data exists somewhere doesn’t mean you can use it for an AI project.
- Accessibility barriers are common and often underestimated.

Common Accessibility Barriers

Barrier	Example	What It Means
Different systems	Data is in three different databases that don’t talk to each other	Need to extract and link—may require technical work and approvals
Permissions	You don’t have access to the system where data lives	Need to request access; may take weeks or months
Data sharing agreements	Data belongs to another agency or partner	Need legal agreements (DUAs, MOUs); can take months
Privacy and legal restrictions	Data contains PHI, PII, or is subject to consent limitations	May need IRB approval, legal review, or de-identification
Format barriers	Data is in paper files, scanned PDFs, or legacy systems	Need extraction or digitization before ML is possible
Political barriers	Data owner is uncooperative or protective	May be insurmountable without executive intervention

The accessibility reality check:

- "We have that data" often means "that data exists somewhere."
- The real question is: can you get it in a usable form, in a reasonable timeframe, with the approvals you need?

Questions to ask about accessibility:

- Where does this data live? Who controls it?
- What approvals are needed to access it?
- Is there a data sharing agreement in place, or would we need to create one?
- How long would it take to get access?
- What format is it in? Would we need to extract or transform it?

Putting It Together: What Data Questions to Ask

Key point: Before proposing or evaluating an AI project, know what questions to ask about data.

This topic has introduced five dimensions of data that affect AI feasibility:

Table 2. Five Dimensions of Data

Dimension	Key Question
Structure	Is the data structured (rows and columns) or unstructured (text, images, etc.)?
Labels	Do we have the outcome we want to predict recorded? Where did it come from?
Volume	How many records? How many examples of the outcome?
Quality	How complete and consistent is the data? What are the known issues?
Accessibility	Can we actually get to the data? What approvals are needed?

The data conversation:

- These questions don’t require technical expertise—they require curiosity and persistence.
- Asking them early surfaces blockers before time and resources are invested.
- The answers shape what’s possible: supervised vs. unsupervised, simple vs. complex approaches, quick start vs. long runway.

Summary

Component	Key Takeaway
4.1 Structured vs. unstructured	Unstructured data requires extra processing; adds complexity
4.2 Labeled vs. unlabeled	Label = outcome to predict; no labels means no supervised learning
4.3 Volume	Need enough examples, especially of the outcome you’re predicting
4.4 Quality	Garbage in, garbage out; assume issues exist and ask about them
4.5 Accessibility	Data existing ≠ data you can use; ask about access early
4.6 Putting it together	Know what questions to ask before proposing a project

---

Foundational Concepts: Putting It Together

- **What AI is** — A broad field; ML is the foundation; deep learning and NLP build on ML
- **How ML works** — Learning from data, not explicit rules; supervised, unsupervised, and reinforcement learning; classification, regression, and clustering
- **How ML differs from classical statistics** — Prediction vs. inference; different workflows and validation approaches
- **What data you need** — Structured vs. unstructured; labeled vs. unlabeled; volume, quality, and accessibility

The Reference Table: How It All Connects

Concept	What It Means	Key Distinctions	Questions to Ask
AI / ML / Deep Learning / NLP	Nested relationship: AI contains ML, ML contains Deep Learning, NLP spans multiple levels	ML is the foundation; DL and NLP build on ML concepts	Is this really AI, or is it something simpler?
Learning Paradigm	How the model learns	Supervised (has labels), Unsupervised (no labels), Reinforcement (feedback/rewards)	Do we have labeled data? What outcome are we predicting?
Task Type	What the model produces	Classification (category), Regression (number), Clustering (groups), Generation (content)	What output do we need? A prediction? A grouping?
Data Structure	How the data is organized	Structured (rows/columns) vs. Unstructured (text, images)	Is the data in a database, or is it free text/documents?
Labels	The outcome you're trying to predict	Labeled (outcome recorded) vs. Unlabeled (outcome not recorded)	Is the outcome we care about actually in the data?

The Reference Table: How It All Connects

Concept	What It Means	Key Distinctions	Questions to Ask
Data Volume	How much data you have	More is generally better; rare outcomes need more data	How many examples? How many of the outcome we're predicting?
Data Quality	How good the data is	Completeness, consistency, accuracy	What's missing? What's messy?
Data Accessibility	Whether you can get to the data	Exists vs. accessible vs. usable	Where does it live? What approvals do we need?
Prediction vs. Inference	The goal of the analysis	ML optimizes for prediction; statistics focuses on inference (understanding relationships)	Do we need to predict, or do we need to explain?

What You Now Know

- **Understand the vocabulary** — When someone says "supervised learning," "classification," or "training data," you know what they mean
- **Ask basic data questions** — Is there labeled data? Is it accessible? Is there enough?
- **Distinguish prediction from inference** — Know when ML is appropriate vs. when classical statistics might be better
- **Spot when something isn’t really AI** — A rule-based system or a simple report isn’t ML, even if someone calls it "AI"

What you don’t yet know (and that’s okay):

- How to evaluate whether a specific problem is a good fit for AI
- How to assess whether AI is the right solution vs. a simpler approach
- How to recognize what kind of AI is being proposed — Is it ML? Deep learning? NLP? What task type?
- How to frame a problem clearly for technical staff
- How to spot bias and ethical concerns
- How to evaluate vendor claims

Transition to Applying AI

Next up: Applying AI

In the next section, you'll learn how to:

- Take a vague directive ("use AI for this") and turn it into a clear problem
- Evaluate whether AI is the right solution
- Identify where AI opportunities exist in your work
- Flag bias and ethical concerns
- Communicate problems clearly to technical staff
