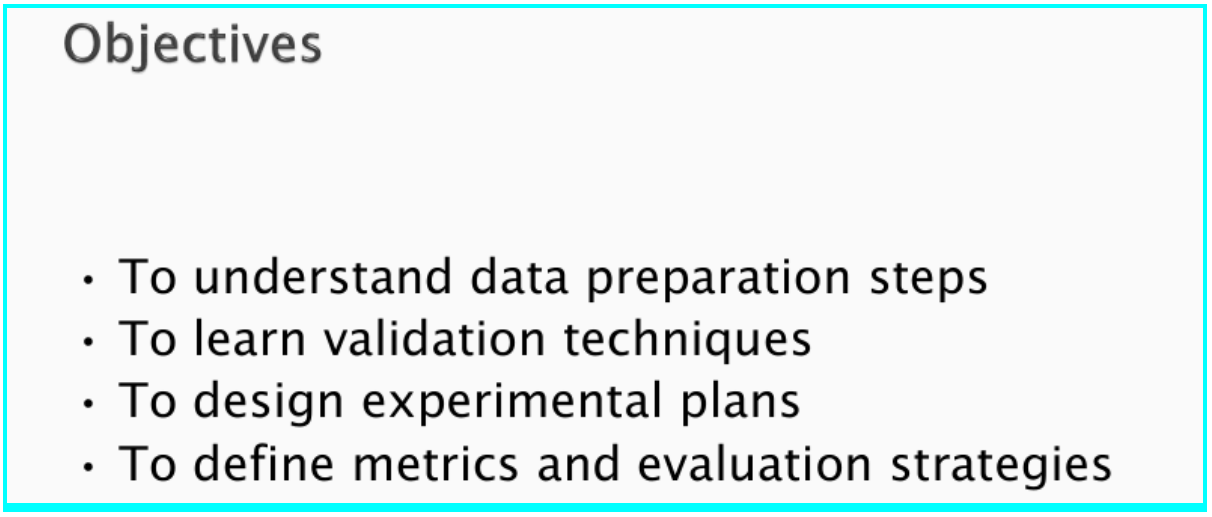
📘 CSE414 — Lecture 6: Research Methodology — Data Preparation, Validation & Experimental Planning

---



🔹 Slide 1: Title Slide

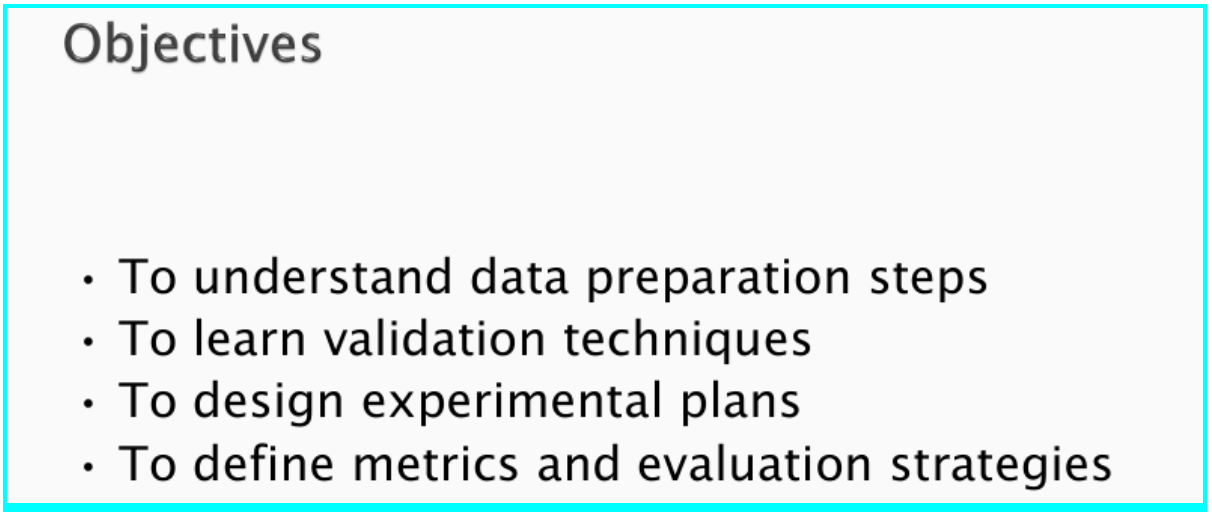
Topic: **Research Methodology — Data Preparation, Validation & Experimental Planning**

এই lecture টা মূলত Lecture 5 এর continuation। Lecture 5 এ শিখেছিলে data কীভাবে collect করতে হয় (qualitative/quantitative) এবং ethically কীভাবে handle করতে হয়। এখন এই lecture এ শেখানো হবে — সেই data নিয়ে collect করার পর কী করতে হয়, অর্থাৎ:

1. Data কে প্রস্তুত **(prepare)** করা

2. Data কে যাচাই **(validate)** করা
3. Experiment টাকে সঠিকভাবে পরিকল্পনা **(plan)** করা

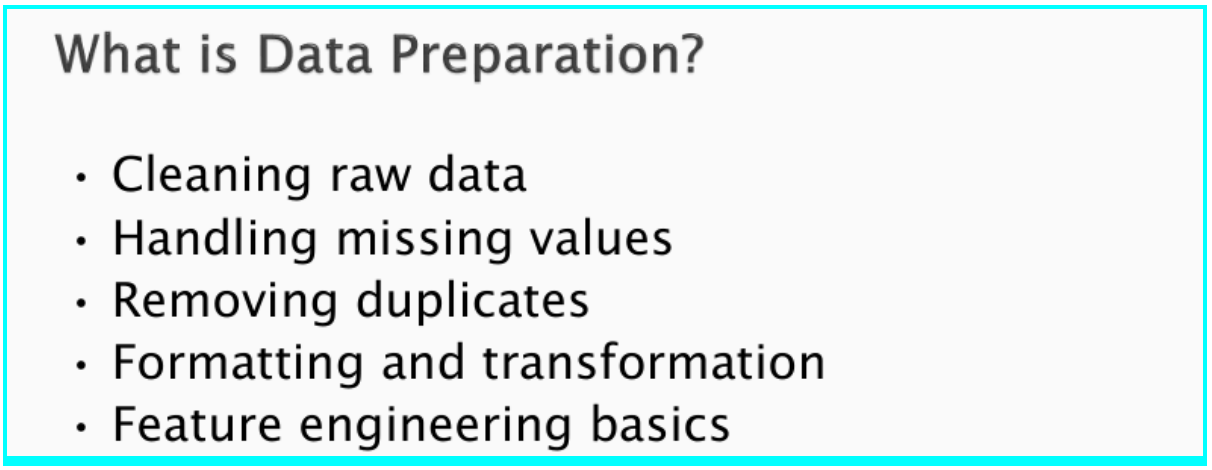
এইটা একটা technical/practical lecture — machine learning research এর workflow বোঝায়।



🔹 Slide 2: Objectives

চারটা লক্ষ্য:

- **To understand data preparation steps** — raw data কে কীভাবে ধাপে ধাপে ব্যবহারযোগ্য বানানো হয়।
- **To learn validation techniques** — data ঠিক আছে কিনা তা যাচাই করার পদ্ধতি।
- **To design experimental plans** — একটা গবেষণা experiment কীভাবে design করতে হয়।
- **To define metrics and evaluation strategies** — ফলাফল measure করার জন্য কোন metric ব্যবহার করবো তা ঠিক করা।



🔹 Slide 3: What is Data Preparation? (Diagram সহ)

Definition: Data Preparation হলো raw (কাঁচা) data কে পরিষ্কার করে বিশ্লেষণ বা model training এর জন্য প্রস্তুত করার প্রক্রিয়া।

পাঁচটা মূল ধাপ:

1. **Cleaning raw data (**কাঁচা তথ্য পরিষ্কার করা**)** — ভুল, অপ্রয়োজনীয় বা noise যুক্ত data বাদ দেওয়া বা ঠিক করা। *Example:* একটা age column এ "-5" বা "200" এর মতো অবাস্তব value থাকলে সেটা fix করা।

2. **Handling missing values (**অনুপস্থিত মান সামলানো**)** — কিছু data point এ value না থাকলে (blank/NULL) সেটা কীভাবে handle করবে। সাধারণত ৩টা উপায়:

    - **Deletion (**মুছে ফেলা**)** — যে row এ value missing, সেটা পুরোপুরি বাদ দেওয়া।
    - **Imputation (**অনুমানভিত্তিক পূরণ**)** — missing value এর জায়গায় mean/median/mode বসিয়ে দেওয়া।
    - **Flagging (**চিহ্নিত করা**)** — missing কে আলাদা category হিসেবে রাখা।

3. **Removing duplicates (**পুনরাবৃত্ত তথ্য মুছে ফেলা**)** — একই data একাধিকবার থাকলে (duplicate rows) সেটা বাদ দেওয়া, নাহলে result skew (বিকৃত) হয়ে যেতে পারে।

4. **Formatting and transformation (**বিন্যাস ও রূপান্তর**)** — Data কে standard format এ আনা। *Example:* Date কে "DD/MM/YYYY" থেকে "YYYY-MM-DD" এ convert করা, অথবা text কে সব lowercase করা।

5. **Feature engineering basics (**ফিচার প্রকৌশলের প্রাথমিক ধারণা**)** — নতুন meaningful feature (variable) তৈরি করা existing data থেকে, যাতে model ভালো শিখতে পারে। *Example:* "Date of Birth" থেকে "Age" নামে নতুন একটা column তৈরি করা।

১. Noisy Dataset (কোলাহলযুক্ত বা বিকৃত ডেটা)

নয়েজি ডেটা বলতে বোঝায় ডেটার মধ্যে থাকা ভুল, অবাস্তব, বা অপ্রত্যাশিত মান **(Errors, Outliers, or Random Variations)**। সাধারণত ডেটা সংগ্রহ করার সময় সেন্সরের ত্রুটি, স্পেলিং মিস্টেক বা টাইপিং ভুলের কারণে এটি ঘটে।

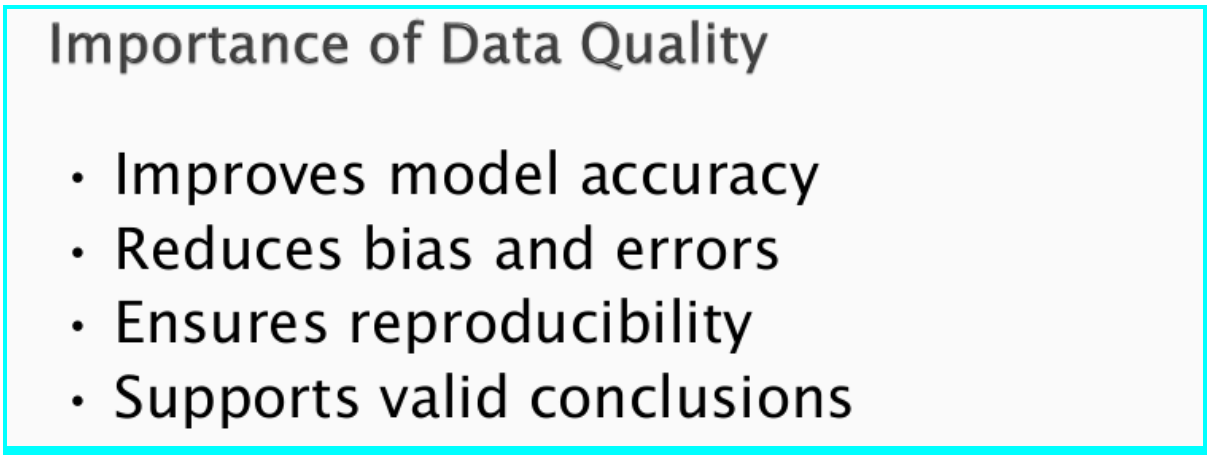
- বৈশিষ্ট্য**:** ডেটাটি ভুল বা বিকৃত, যা খালি চোখেই অসঙ্গতিপূর্ণ বলে মনে হয়।
- বাস্তব উদাহরণ**:**
    - বয়স: -২৫ বা ১৫০০ (বয়স কখনো মাইনাস বা ১৫০০ হতে পারে না)।
    - তাপমাত্রা: সেন্সর এররের কারণে ঘরের তাপমাত্রা দেখাচ্ছে ৫০০°C (যা অস্বাভাবিক/নয়েজ)।
- সমাধানের উপায়**:** Outlier Removal, Smoothing Techniques (যেমন- Binning, Regression), বা Mean/Median ইম্পুটেশন।

২. Inconsistent Dataset (অসামঞ্জস্যপূর্ণ বা সাংঘর্ষিক ডেটা)

ইনকনসিস্টেন্ট ডেটা বলতে বোঝায় এমন ডেটা, যেখানে একই বিষয় নিয়ে দুটি ভিন্ন জায়গায় দুই ধরণের তথ্য বা সাংঘর্ষিক দাবি **(Contradictory/Conflicting Data)** রয়েছে। ডেটার মানগুলো হয়তো নিজ নিজ জায়গায় সঠিক, কিন্তু অন্য ডেটার সাথে মিলালে বৈপরীত্য দেখা দেয়।

- বৈশিষ্ট্য: ভিন্ন ভিন্ন সোর্স থেকে ডেটা একত্রিত (Data Integration) করার সময় এই সমস্যা তৈরি হয়।
- বাস্তব উদাহরণ:
    - একই ব্যক্তির জন্মতারিখ দেওয়া আছে: `01/01/2000`, কিন্তু তার বয়স লেখা আছে: `10` বছর (অথচ ২০০০ সালে জন্ম হলে বয়স আরও বেশি হওয়ার কথা)।
    - পাসপোর্ট ও জাতীয় পরিচয়পত্রে একই ব্যক্তির নাম দুইভাবে লেখা (যেমন: এক জায়গায় `Ahmad`, অন্য জায়গায় `Ahmed`) বা জেন্ডার দুই জায়গায় দুই রকম থাকা।
- সমাধানের উপায়: Data Integration, Redundancy Elimination, এবং Business Rules তৈরি করে বৈপরীত্য দূর করা।

📌 এই ৫টা step পুরো data preparation pipeline এর মূল কাঠামো — exam এ "explain data preparation steps" ধরনের প্রশ্নে এই ৫টাই লিখতে হবে।



🔹 Slide 4: Importance of Data Quality (কেন Data Quality গুরুত্বপূর্ণ)

চারটা কারণ:

1. **Improves model accuracy** (মডেলের নির্ভুলতা বৃদ্ধি করে) — ভালো মানের data দিলে model ভালো ফলাফল দেয়।
2. **Reduces bias and errors** (পক্ষপাত ও ভুল কমায়) — ভুল/অসম্পূর্ণ data দিলে model ভুল সিদ্ধান্ত নিতে শেখে।
3. **Ensures reproducibility** (পুনরুৎপাদনযোগ্যতা নিশ্চিত করে) — অন্য কেউ একই data ও পদ্ধতি ব্যবহার করলে একই ফলাফল পাবে।
4. **Supports valid conclusions** (বৈধ সিদ্ধান্তে সাহায্য করে) — সঠিক data থেকেই সঠিক গবেষণা সিদ্ধান্ত আসতে পারে।

সহজ **concept:** এইটাকে বলে **"Garbage In, Garbage Out" (GIGO) principle** — তুমি যদি খারাপ মানের data দাও, তাহলে model থেকেও খারাপ ফলাফল আসবে, যতই ভালো algorithm ব্যবহার করো না কেন।

Data Validation

- Checking data consistency
- Range and type validation
- Outlier detection
- Cross-checking sources

🔹 Slide 5: Data Validation

Definition: Data Validation হলো collect করা data সঠিক, সামঞ্জস্যপূর্ণ (consistent) এবং ব্যবহারযোগ্য কিনা তা যাচাই করার প্রক্রিয়া — এটা Data Preparation এর পরের ধাপ।

চারটা মূল কাজ:

1. **Checking data consistency (**তথ্যের সামঞ্জস্য যাচাই করা**)** — বিভিন্ন জায়গায় একই তথ্য একইরকম আছে কিনা দেখা। *Example:* একই patient এর age দুই জায়গায় ২৫ আর ৩০ লেখা থাকলে সেটা inconsistency।

2. **Range and type validation (**সীমা ও ধরন যাচাই**)** — Value গুলো তাদের নির্ধারিত সীমার (range) মধ্যে আছে কিনা এবং সঠিক data type (number/text) এ আছে কিনা check করা। *Example:* "percentage" column এ value 0-100 এর মধ্যে হওয়া উচিত, ১৫০ হলে সেটা ভুল।

3. **Outlier detection (**ব্যতিক্রমী মান শনাক্তকরণ**)** — যে data point অন্য সব data থেকে অনেক আলাদা/অস্বাভাবিক, সেটা খুঁজে বের করা। *Example:* বেশিরভাগ salary 20k-50k এর মধ্যে, কিন্তু একজনের salary 5 million দেখাচ্ছে — এটা outlier, হয়তো data entry error।

4. **Cross-checking sources (**একাধিক উৎসের সাথে মিলিয়ে দেখা**)** — একই তথ্য অন্য কোনো নির্ভরযোগ্য উৎসের সাথে মিলিয়ে verify করা।

Common Validation Techniques

- Manual inspection
- Automated scripts
- Statistical checks
- Visualization methods

🔹 Slide 6: Common Validation Techniques

চারটা পদ্ধতি data validate করার জন্য:

1. **Manual inspection (**হাতে-কলমে পরিদর্শন**)** — মানুষ নিজে চোখে দেখে data check করা। ছোট dataset এর জন্য উপযোগী।

2. **Automated scripts (**স্বয়ংক্রিয় স্ক্রিপ্ট**)** — Programming (Python ইত্যাদি) দিয়ে script লিখে automatic ভাবে ভুল খুঁজে বের করা। বড় dataset এর জন্য দরকার।

3. **Statistical checks (**পরিসংখ্যানগত যাচাই**)** — Mean, standard deviation, correlation এর মতো statistical method দিয়ে data এর pattern বোঝা এবং অস্বাভাবিকতা ধরা।

4. **Visualization methods (**দৃশ্যমান পদ্ধতি**)** — Graph, chart, boxplot ব্যবহার করে চোখে দেখে data এর ভুল/outlier ধরা। *Example:* একটা boxplot বানালে outlier গুলো সহজেই চোখে পড়ে।

Experimental Planning Overview

- Define research question
- Select methodology
- Choose dataset and variables
- Plan workflow

🔹 Slide 7: Experimental Planning Overview

Definition: Experimental Planning হলো একটা গবেষণা experiment কীভাবে পরিচালনা করা হবে তার আগে থেকেই সুসংগঠিত পরিকল্পনা তৈরি করা।

চারটা ধাপ:

1. **Define research question (**গবেষণা প্রশ্ন নির্ধারণ করা**)** — কী জানতে চাও তা স্পষ্টভাবে ঠিক করা। *Example: "*কোন *CNN model white blood cell classification* এ সবচেয়ে ভালো *accuracy* দেয়*?"*

2. **Select methodology (**পদ্ধতি নির্বাচন করা**)** — Qualitative/Quantitative/Mixed method এর মধ্যে কোনটা ব্যবহার করবে তা ঠিক করা।

3. **Choose dataset and variables (Dataset** ও চলক বাছাই করা**)** — কোন dataset ব্যবহার করবে, কোন variable (input/output feature) দিয়ে কাজ করবে তা নির্ধারণ করা।

4. **Plan workflow (**কার্যপ্রবাহ পরিকল্পনা করা**)** — পুরো experiment টা কোন ধারাবাহিকতায় (sequence) হবে তার একটা map তৈরি করা (data collection → cleaning → training → evaluation)।

Train/Test Split

- Training set: model learning
- Testing set: performance evaluation
- Typical split: 70/30 or 80/20
- Avoid data leakage

🔹 Slide 8: Train/Test Split (অত্যন্ত গুরুত্বপূর্ণ — Machine Learning Concept)

Definition: এটি হলো একটা dataset কে দুই ভাগে ভাগ করার প্রক্রিয়া — একভাগ দিয়ে model train (শেখানো) করা হয়, আরেক ভাগ দিয়ে সেটার performance test (পরীক্ষা) করা হয়।

চারটা মূল point:

1. **Training set: model learning (**প্রশিক্ষণ সেট**)** — Dataset এর যে অংশ দিয়ে model শেখে (pattern চিনতে শেখে)।

2. **Testing set: performance evaluation (**পরীক্ষার সেট**)** — যে অংশ দিয়ে model কতটা ভালো শিখেছে তা যাচাই করা হয়। এই data model আগে দেখেনি।

3. **Typical split: 70/30 or 80/20** — সাধারণত dataset এর ৭০% বা ৮০% training এর জন্য, বাকি ৩০% বা ২০% testing এর জন্য রাখা হয়।

4. **Avoid data leakage (**তথ্য ফাঁস এড়ানো**)** — Testing data যেন কোনোভাবেই training এর সময় ব্যবহার না হয়, নাহলে model ভুয়া ভালো result দেখাবে (কারণ সে "পরীক্ষার প্রশ্ন আগেই দেখে ফেলেছে")।

সহজ **Analogy (**উপমা**):** এটা অনেকটা student exam এর মতো — Training set হলো practice question (যেগুলো দিয়ে পড়াশোনা করো), Testing set হলো আসল exam question (যেগুলো তুমি আগে দেখোনি)। যদি exam এর প্রশ্ন আগে থেকে জেনে যাও (data leakage), তাহলে তোমার result সত্যিকারের জ্ঞান প্রকাশ করবে না।

Cross Validation

- K-fold validation
- Improves reliability
- Reduces overfitting risk
- Used when data is limited

🔹 Slide 9: Cross Validation

Definition: Cross Validation হলো একটা advanced technique, যেখানে dataset কে বারবার বিভিন্নভাবে ভাগ করে model কে train ও test করা হয়, যাতে ফলাফল বেশি নির্ভরযোগ্য হয়।

চারটা point:

1. **K-fold validation** — Dataset কে "K" সংখ্যক ভাগে (fold) ভাগ করা হয়। প্রতিবার একটা fold কে test set হিসেবে রেখে বাকি সবগুলো দিয়ে train করা হয় — এভাবে K বার repeat করা হয়। *Example:* যদি K=5 হয়, তাহলে dataset ৫ ভাগে ভাগ হবে, এবং ৫ বার আলাদা আলাদা combination দিয়ে train/test হবে, তারপর সবগুলোর average result নেওয়া হয়।

2. **Improves reliability (**নির্ভরযোগ্যতা বৃদ্ধি করে**)** — একবার split করার চেয়ে বারবার split করে average নিলে result বেশি বিশ্বাসযোগ্য হয়।

3. **Reduces overfitting risk (Overfitting** এর ঝুঁকি কমায়**)** — Model যেন শুধু একটা নির্দিষ্ট split এর জন্যই ভালো কাজ না করে, বরং সার্বিকভাবে ভালো কাজ করে তা নিশ্চিত করে।

4. **Used when data is limited (**ডেটা কম থাকলে ব্যবহৃত হয়**)** — যখন dataset ছোট থাকে, তখন একবার split করলে test set এ পর্যাপ্ত data থাকে না, তাই cross-validation ব্যবহার করা হয়।

Overfitting এর **simple definition:** যখন একটা model training data একদম মুখস্থ করে ফেলে (memorize করে), কিন্তু নতুন/unseen data তে ভালো performance দেখাতে পারে না — একে overfitting বলে।

Evaluation Metrics

- Accuracy, Precision, Recall
- F1-score
- Mean Squared Error
- Confusion matrix

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

🔹 Slide 10: Evaluation Metrics (খুব গুরুত্বপূর্ণ — মুখস্থ রাখা দরকার)

চারটা প্রধান metric:

1. **Accuracy** — মোট prediction এর মধ্যে কতগুলো সঠিক হয়েছে তার শতকরা হার।

   Formula (সহজভাবে): Accuracy = (সঠিক Prediction সংখ্যা) / (মোট Prediction সংখ্যা)

2. **Precision (**নির্ভুলতা**)** — Model যতগুলো "Positive" বলেছে তার মধ্যে কতগুলো আসলেই Positive ছিল। *Example:* Model ১০০ জনকে "cancer আছে" বলল, তার মধ্যে ৯০ জনের আসলেই cancer

ছিল → Precision = 90%।

3. **Recall (স্মরণশক্তি/সংবেদনশীলতা)** — যত জনের আসলেই disease ছিল, তার মধ্যে model কতজনকে সঠিকভাবে ধরতে পারল। *Example:* মোট ১২০ জনের cancer ছিল, model তার মধ্যে ৯০ জনকে ধরতে পারলো → Recall = 75%।

4. **F1-score** — Precision এবং Recall এর মধ্যে একটা balance/average (harmonic mean)। যখন Precision আর Recall দুটোই গুরুত্বপূর্ণ, তখন F1-score ব্যবহার করা হয়।

5. **Mean Squared Error (MSE)** — মূলত regression problem এ ব্যবহৃত হয়। Predicted value আর Actual value এর মধ্যে পার্থক্যের বর্গের গড় (average)। যত কম MSE, তত ভালো model।

6. **Confusion Matrix (বিভ্রান্তি ম্যাট্রিক্স)** — একটা table যেখানে True Positive, True Negative, False Positive, False Negative দেখানো হয়, যা থেকে Accuracy, Precision, Recall সবকিছু বের করা যায়।

সহজ মনে রাখার **Trick:**

- **Precision** = "যা বলেছি তার কতটা ঠিক" (Quality of positive predictions)
- **Recall** = "যা ঠিক ছিল তার কতটা ধরতে পেরেছি" (Coverage of actual positives)

📌 তোমার প্রশ্নপত্রের Q2 এ দেখলাম "precision of 99%, F1-score of 99%" উল্লেখ আছে — এই metric গুলো ঠিক এখান থেকেই এসেছে।

Precision vs Recall: সহজ ব্যাখ্যা ও বাস্তব উদাহরণ

মেশিন লার্নিং (Machine Learning) বা ডেটা সায়েন্সে কোনো মডেলের পারফরম্যান্স পরিমাপ করার জন্য **Precision** এবং **Recall** দুটি অত্যন্ত গুরুত্বপূর্ণ পরিমাপক (Metrics)। তুমি যেন এটি সরাসরি ফাইলে বা ডকুমেন্টে পেস্ট করতে পারো, সেজন্য বড় সাইজের টাইটেল, সাবটাইটেল এবং সহজ ব্যাখ্যাসহ নিচে বুঝিয়ে দেওয়া হলো:

1. Precision (নির্ভুলতা বা নির্ভুলতার হার)

বাংলা সংজ্ঞা**:** প্রিসিশন (Precision) হলো— মডেলটি যতগুলো ফলাফলকে "পজিটিভ" (যেমন: রোগ আছে) বলে দাবি করেছে, তার মধ্যে আসলেই কতগুলো ফলাফল প্রকৃতপক্ষে "পজিটিভ" ছিল। অর্থাৎ, মডেলের নিজের করা পূর্বাভাসের নির্ভুলতার হারই হলো প্রিসিশন।

English Definition: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It answers the question: "Of all patients the model predicted as having cancer, how many actually have it?"

- সহজ বাস্তব উদাহরণ**:** * একটি মেডিকেল এআই (AI) মডেল ১০০ জন রোগীর রিপোর্ট দেখে বলল, **"**আপনাদের সবার ক্যান্সার **(Cancer)** আছে**"** (Predicted Positive = ১০০)।
    - পরবর্তীতে ডাক্তার নিখুঁতভাবে পরীক্ষা করে দেখলেন, ওই ১০০ জনের মধ্যে ৯০ জনের আসলেই ক্যান্সার ছিল (True Positive = ৯০), আর বাকি ১০ জন সম্পূর্ণ সুস্থ ছিল।
    - এখানে মডেলের প্রিসিশন হবে: frac{90}{100} = 90%।

- এর অর্থ হলো: মডেলটি যখন কাউকে রোগী বলে, তখন তার কথাটি ৯০% সত্যি হওয়ার সম্ভাবনা থাকে।

2. Recall (সংবেদনশীলতা বা ধরার ক্ষমতা)

বাংলা সংজ্ঞা: রিকল (Recall) হলো— পুরো ডেটাসেটে বা বাস্তবে যতজন মানুষ আসলেই "পজিটিভ" (যেমন: মোট যতজন সত্যিকারের রোগী আছে), তাদের মধ্য থেকে মডেলটি কতজনকে সঠিকভাবে খুঁজে বের করতে বা "ধরতে" পেরেছে।

English Definition: Recall (also known as Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. It answers the question: "Of all the people who truly have cancer, how many did the model manage to detect?"

- সহজ বাস্তব উদাহরণ:
  - মনে করো, একটি হাসপাতালে বা ডেটাসেটে মোট ১২০ জন সত্যিকারের ক্যান্সার রোগী আছেন (Actual Positive = ১২০)।
  - এবার ওই এআই মডেলটিকে সবার পরীক্ষা করতে দেওয়া হলো। মডেলটি ওই ১২০ জন রোগীর মধ্যে থেকে ৯০ জনকে সঠিকভাবে সনাক্ত করতে পারল (True Positive = 90)। বাকি ৩০ জন রোগীকে সে সুস্থ বলে ভুল করল (ছেড়ে দিল)।
  - এখানে মডেলের রিকল হবে: frac{90}{120} = 75%।
  - এর অর্থ হলো: বাস্তবে যত রোগীই থাক না কেন, এই মডেলটির তাদেরকে খুঁজে বের করার ক্ষমতা বা সাকসেস রেট হলো ৭৫%।

৩. দ্রুত মনে রাখার শর্টকাট ট্রিক (The Core Difference)

- **Precision-**এর ফোকাস হলো মডেলের নিজের ওপরে: মডেলটি পজিটিভ বলার পর নিজের কথায় কতটুকু নির্ভুল **(Precise)** ছিল। (ভুল পজিটিভ বা False Positive কমানো এর লক্ষ্য)।
- **Recall-**এর ফোকাস হলো পুরো বাস্তবের ওপরে: বাস্তবে যত রোগী ছিল, তাদের সবাইকে মডেলটি স্মরণ করে বা খুঁজে **(Recall)** বের করতে পারল কি না। (রোগী যেন বাদ না পড়ে বা False Negative কমানো এর লক্ষ্য)।

৪. কখন কোনটি বেশি গুরুত্বপূর্ণ? (Exam Scenario)

- মেডিকেল টেস্টে **Recall** বেশি গুরুত্বপূর্ণ: কারণ ক্যান্সার বা হার্ট অ্যাটাকের মতো মারাত্মক রোগে কোনো রোগী যেন বাদ না পড়ে। মডেলের রিকল ১০০% হওয়া দরকার, যেন সে সব রোগীকে ধরতে পারে (ভলে দুই একজন সুস্থ মানুষকে রোগী বললে ক্ষতি কম, কিন্তু আসল রোগীকে সুস্থ বলে ছেড়ে দিলে সে মারা যেতে পারে)।
- স্প্যাম ইমেইল ফিল্টারে **Precision** বেশি গুরুত্বপূর্ণ: কারণ মডেল যদি কোনো একটি জরুরি ইমেইলকে ভুল করে স্প্যাম ফোল্ডারে পাঠিয়ে দেয় (False Positive), তবে ব্যবহারকারীর বড় ক্ষতি হতে পারে। তাই স্প্যাম বলার আগে মডেলকে ১০০% নিশ্চিত বা প্রিসাইজ হতে হয়।

Choosing Metrics

- Depends on problem type
- Classification vs regression
- Business or research goal
- Sensitivity vs specificity trade-off

🔹 Slide 11: Choosing Metrics (কোন Metric কখন বাছাই করবে)

চারটা বিবেচনা:

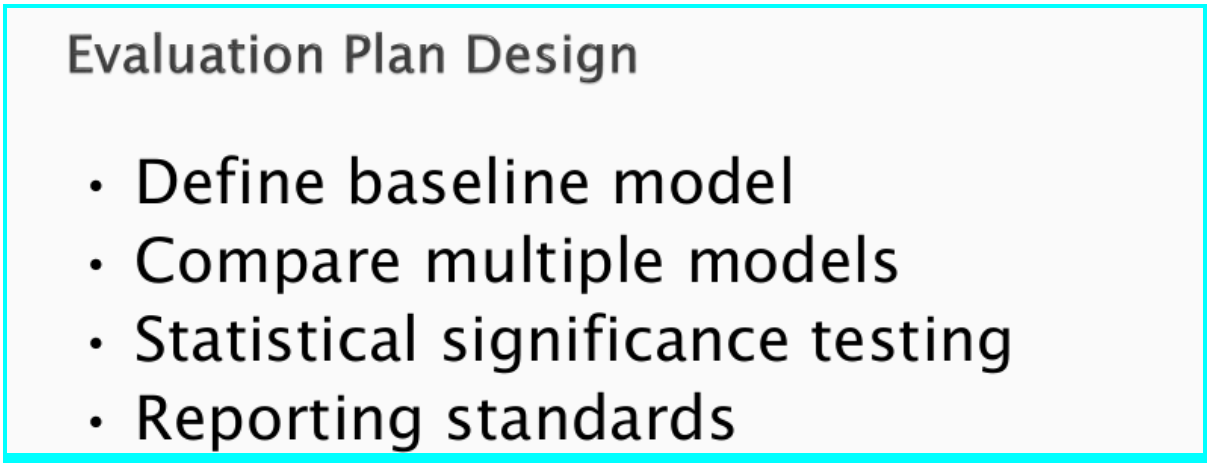
1. **Depends on problem type (**সমস্যার ধরনের উপর নির্ভর করে**)** — Classification নাকি Regression সমস্যা তার উপর metric নির্বাচন নির্ভর করে।

2. **Classification vs regression** —

   - **Classification** (যেমন: রোগ আছে/নেই) → Accuracy, Precision, Recall, F1-score ব্যবহার হয়।
   - **Regression** (যেমন: বাড়ির দাম predict করা - একটা continuous number) → MSE, RMSE ব্যবহার হয়।
3. **Business or research goal (**ব্যবসায়িক বা গবেষণার লক্ষ্য**)** — তোমার আসল উদ্দেশ্য কী তার উপর নির্ভর করে metric ঠিক করতে হয়।

4. **Sensitivity vs specificity trade-off (**সংবেদনশীলতা বনাম নির্দিষ্টতার ভারসাম্য**)** — কিছু ক্ষেত্রে false negative (রোগ থাকা সত্ত্বেও ধরতে না পারা) এড়ানো বেশি জরুরি (তখন Recall বেশি গুরুত্বপূর্ণ), আবার কিছু ক্ষেত্রে false positive এড়ানো জরুরি (তখন Precision বেশি গুরুত্বপূর্ণ)।

Real-life Example: Cancer detection এ **Recall** কে বেশি গুরুত্ব দেওয়া হয় — কারণ একজন প্রকৃত cancer রোগীকে miss করা (False Negative) অনেক বেশি বিপজ্জনক, বরং কিছু healthy মানুষকে ভুলবশত positive বলে ফেলা (False Positive) তুলনামূলক কম ক্ষতিকর (কারণ পরে আবার test করে নিশ্চিত করা যায়)।



🔹 Slide 12: Evaluation Plan Design

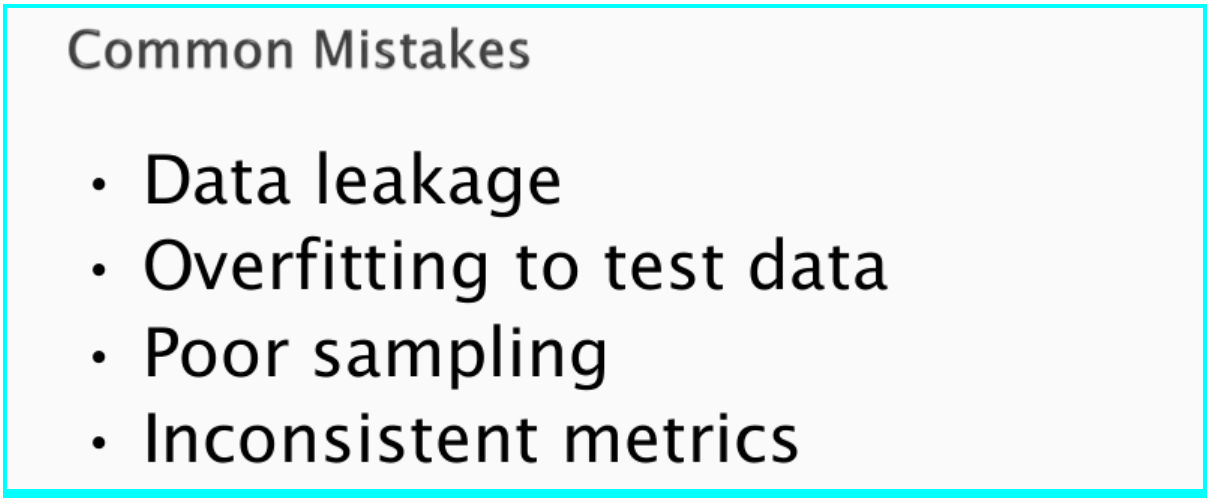
চারটা ধাপ যেভাবে একটা সঠিক evaluation plan তৈরি করতে হয়:

1. **Define baseline model (**ভিত্তি মডেল নির্ধারণ করা**)** — একটা সাধারণ/সহজ model দিয়ে শুরু করা, যার সাথে পরবর্তী উন্নত model গুলো তুলনা করা হবে। *Example:* একটা simple Logistic Regression কে baseline ধরে, তার সাথে CNN, MobileNetV2 এর result compare করা (ঠিক যেমন তোমার প্রশ্নপত্রের abstract এ VGG19, MobileNetV2, DenseNet201 compare করা হয়েছে)।

2. **Compare multiple models (**একাধিক মডেল তুলনা করা**)** — বিভিন্ন model/algorithm ব্যবহার করে কোনটা সবচেয়ে ভালো ফলাফল দেয় তা যাচাই করা।

3. **Statistical significance testing (**পরিসংখ্যানগত তাৎপর্য পরীক্ষা**)** — দুইটা model এর মধ্যে যে পার্থক্য দেখা যাচ্ছে, সেটা কি সত্যিকারের পার্থক্য নাকি শুধু কাকতালীয় (by chance) — তা পরিসংখ্যানগতভাবে পরীক্ষা করা।

4. **Reporting standards (**প্রতিবেদনের মান**)** — ফলাফল কীভাবে সঠিকভাবে, স্বচ্ছভাবে এবং সবার বোঝার মতো করে উপস্থাপন করতে হবে তার নিয়ম মেনে চলা।



🔹 Slide 13: Common Mistakes (গবেষণায় সাধারণ ভুল)

চারটা common mistake যা এড়ানো উচিত:

1. **Data leakage (**তথ্য ফাঁস**)** — Testing data ভুলবশত training এর সময় ব্যবহার হয়ে যাওয়া, যার ফলে model কে ভুয়া ভালো মনে হয়।

2. **Overfitting to test data (Test data** তে অতি-মানানসই হয়ে যাওয়া**)** — বারবার একই test set এ model tune করতে করতে model সেই নির্দিষ্ট test set এর জন্যই ভালো হয়ে যায়, কিন্তু নতুন unseen data তে খারাপ করে।

3. **Poor sampling (**দুর্বল নমুনা নির্বাচন**)** — এমনভাবে sample বাছাই করা যা পুরো population কে সঠিকভাবে প্রতিনিধিত্ব করে না (representative না হওয়া), ফলে ফলাফল biased হয়ে যায়।

4. **Inconsistent metrics (**অসামঞ্জস্যপূর্ণ মেট্রিক্স**)** — বিভিন্ন model compare করার সময় একই metric ব্যবহার না করা, যার ফলে তুলনা ভুল হয়ে যায়।

📌 **Exam Tip:** এই ৪টা mistake যদি কোনো scenario তে দেখা যায় (যেমন কেউ test data দিয়েই model tune করছে), তাহলে সেটা identify করে বলতে পারতে হবে "এটা data leakage/overfitting এর সমস্যা"।

Summary

- Data quality is critical
- Validation ensures reliability
- Proper planning improves outcomes
- Metrics guide interpretation

🔹 Slide 14: Summary

চারটা মূল takeaway:

- **Data quality is critical** (ডেটার গুণমান অত্যন্ত গুরুত্বপূর্ণ)
- **Validation ensures reliability** (যাচাইকরণ নির্ভরযোগ্যতা নিশ্চিত করে)
- **Proper planning improves outcomes** (সঠিক পরিকল্পনা ফলাফল উন্নত করে)
- **Metrics guide interpretation** (মেট্রিক্স ফলাফল ব্যাখ্যায় সাহায্য করে)

✅ Quick Recap Table — Lecture 6 (Exam এর জন্য সবচেয়ে গুরুত্বপূর্ণ)

Concept	মনে রাখো
Data Preparation 5 steps	Cleaning → Handling missing values → Removing duplicates → Formatting/transformation → Feature engineering
Data Validation 4 checks	Consistency, Range/type, Outlier detection, Cross-checking
Train/Test Split	সাধারণত 70/30 বা 80/20; data leakage এড়াতে হবে
Cross Validation	K-fold; data কম থাকলে ব্যবহার হয়; overfitting কমায়
Metrics	Accuracy, Precision, Recall, F1-score, MSE, Confusion Matrix
Precision vs Recall	Precision = "যা বলেছি ঠিক কিনা"; Recall = "যা ঠিক ছিল তার কতটা ধরেছি"
Common Mistakes	Data leakage, Overfitting, Poor sampling, Inconsistent metrics

---

🔗 দুই Lecture এর মধ্যে সংযোগ (Connection between Lecture 5 & 6)

এটা বোঝা জরুরি, কারণ exam এ দুই lecture মিলিয়ে scenario question আসতে পারে:

- **Lecture 5** শেখায় → data কীভাবে সংগ্রহ **(collect)** করবে (qualitative/quantitative) এবং **ethically** কীভাবে সংগ্রহ করবে (privacy, consent, integrity)।
- **Lecture 6** শেখায় → সেই সংগৃহীত data নিয়ে কী করবে — প্রস্তুত (prepare), যাচাই (validate) করবে, এবং experiment plan করে সঠিক metric দিয়ে ফলাফল মূল্যায়ন (evaluate) করবে।

পুরো **Research Workflow (**একসাথে**):** Research Question → Data Collection (Lec 5: Qualitative/Quantitative + Ethics) → Data Preparation → Data Validation → Experimental Planning → Train/Test Split → Model Building → Evaluation (Metrics) → Conclusion

---