خلاصه جلسه دوم مرور کتاب:

برای مدل‌های یادگیری ماشین، به داده‌های حجیم (big data) نیاز است. اما حجم زیاد داده‌ها به معنای کیفیت بالای آن‌ها نیست.

در ابتدا، دانشمندی به نام لنین مفهوم داده‌های بزرگ را با بیان ۳ ویژگی (یا ۳ V) مطرح کرد که ویژگی‌های اساسی داده‌های بزرگ را نشان می‌داد. امروزه این مفهوم گسترش یافته و ۵۶ V برای داده‌های بزرگ تعریف شده است.

مهم‌ترین V ها به شرح زیر هستند:

۱. حجم (Volume): داده‌های بزرگ حجم زیادی دارند که نیاز به منابع مناسب برای ذخیره‌سازی و دسترسی جهت تحلیل و تفسیر آن‌ها دارد.
۲. تنوع (Variety): داده‌ها از منابع، محل‌ها و انواع مختلفی تأمین می‌شوند.
۳. سرعت (Velocity): یکی از ویژگی‌های کلیدی داده‌های بزرگ، سرعت بالای تولید داده و وجود داده‌ها به‌صورت لحظه‌ای (real-time) است.
۴. ارزش (Value): برای استفاده از داده‌ها، دقت و ارزش آن‌ها بسیار مهم است. ارزش داده‌ها خود به‌طور مستقیم وجود ندارد، بلکه باید از بین داده‌های موجود استخراج و ایجاد شود.
۵. صحت (Veracity): بررسی هرگونه بایاس، نویز و ناهماهنگی در داده‌ها بسیار اهمیت دارد. داده‌ها باید کامل، تمیز، منسجم و سازگار باشند.
۶. اعتبار (Validity): داده‌ها باید با هدف موردنظر مرتبط و مناسب باشند.
۷. تغییرپذیری (Variability): این مفهوم تفاوت داخلی داده‌ها را نشان می‌دهد؛ به‌عنوان مثال، هر نوع داده (متن، تصویر، عدد) ویژگی‌های خاص خود را دارد.
۸. تصویری‌سازی (Visualization): یکی از مهم‌ترین بخش‌ها در داده‌های بزرگ، توانایی ارائه نتایج به‌صورت تصویری واضح، شفاف و قابل‌فهم برای دیگران است (مانند بیماران یا سیاست‌گذاران).

متادیتا (Metadata) به معنای داده‌هایی است که اطلاعاتی درباره داده‌های دیگر ارائه می‌دهند. این داده‌ها به توصیف محتوا، ساختار و ویژگی‌های داده اصلی کمک می‌کنند تا مدیریت، درک و بازیابی آن‌ها راحت‌تر شود. متادیتا

مانند یک برچسب یا خلاصه‌ای از داده اصلی عمل می‌کند و جزئیاتی مانند منبع، تاریخ ایجاد، نویسنده، اندازه فایل، فرمت و نحوه سازماندهی داده‌ها را ارائه می‌دهد.

به عنوان مثال، در یک عکس دیجیتال، تصویر خود داده اصلی است و متادیتا می‌تواند شامل اطلاعاتی مانند تنظیمات دوربین، مکان، زمان عکسبرداری و فرمت فایل باشد. این اطلاعات اضافی (متادیتا) برای سازماندهی، جستجو و پردازش داده‌ها بسیار مفید است.