مقایسه حفاری داده‌ها با تحلیل داده‌ها

Data Analysis and Data Mining Comparison

عادل محمدپور
دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)
AdelM.ir
adel@aut.ac.ir

مفهوم داده‌كاوي

عبارت داده‌كاوي مترادف با يكي از عبارت‌هاي
استخراج دانش،
برداشت اطلاعات،
وارسي داده‌ها
و حتي لايروبي كردن داده‌ها است
كه در حقيقت كشف دانش در پايگاه داده‌ها را توصيف مي‌كند

- Knowledge Discovery of Database (KDD)

کشف دانش در پایگاه داده‌ها

Knowledge Discovery in Data is the
non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

From *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

تعاريف داده‌كاوي

- داده‌كاوي در حقيقت كشف ساختارهاي جالب توجه، غيرمنتظره و با ارزش از داخل مجموعه وسيعي از داده‌ها مي‌باشد و فعاليتي است كه اساساً با آمار و تحليل دقيق داده‌ها منطبق است
- داده‌كاوي فرايند كشف رابطه‌ها، الگوها و روندهاي جديد معني‌داري است كه به بررسي حجم وسيعي از اطلاعات ذخيره شده در انبارها با فناوري‌هاي تشخيص الگو (مانند رياضي و آمار) مي‌پردازد
- داده‌كاوي يا به تعبير ديگر كشف دانش در پايگاه داده‌ها، استخراج غير بديهي اطلاعات بالقوه مفيد از روي داده‌هايي است كه قبلاً ناشناخته مانده‌اند. اين مطلب برخي از روش‌هاي فني مانند خوشه‌بندي، خلاصه‌سازي داده‌ها، فراگيري قاعده‌هاي رده‌بندي، يافتن ارتباط شبكه‌ها، تحليل تغييرات و كشف بي‌قاعدگي‌ها را شامل مي‌شود

چرا حفاري داده‌ها بجاي تحليل داده‌ها

داده‌هاي بيشتري توليد مي‌شوند

Bank, telecom, other business transactions ...

Scientific Data: astronomy, biology, etc

Web, text, and e-commerce

داده‌هاي بيشتري ذخيره مي‌شوند

Storage technology faster and cheaper

DBMS capable of handling bigger DB

مثال

Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session

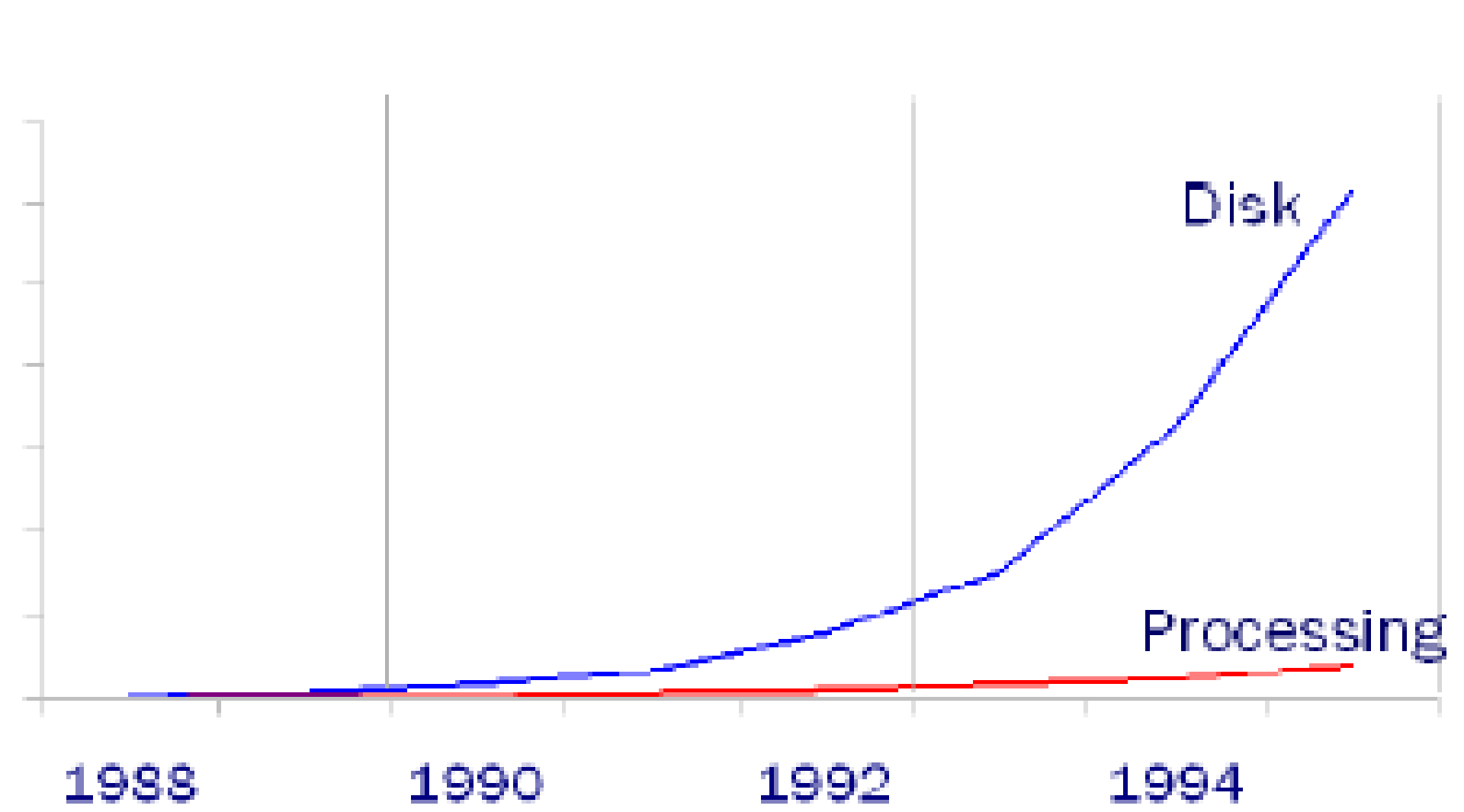
storage and analysis a big problem

Walmart reported to have 24 Tera-byte DB

AT&T handles billions of calls per day

data cannot be stored -- analysis is done on the fly

چرا حفاري داده‌ها بجاي تحليل داده‌ها

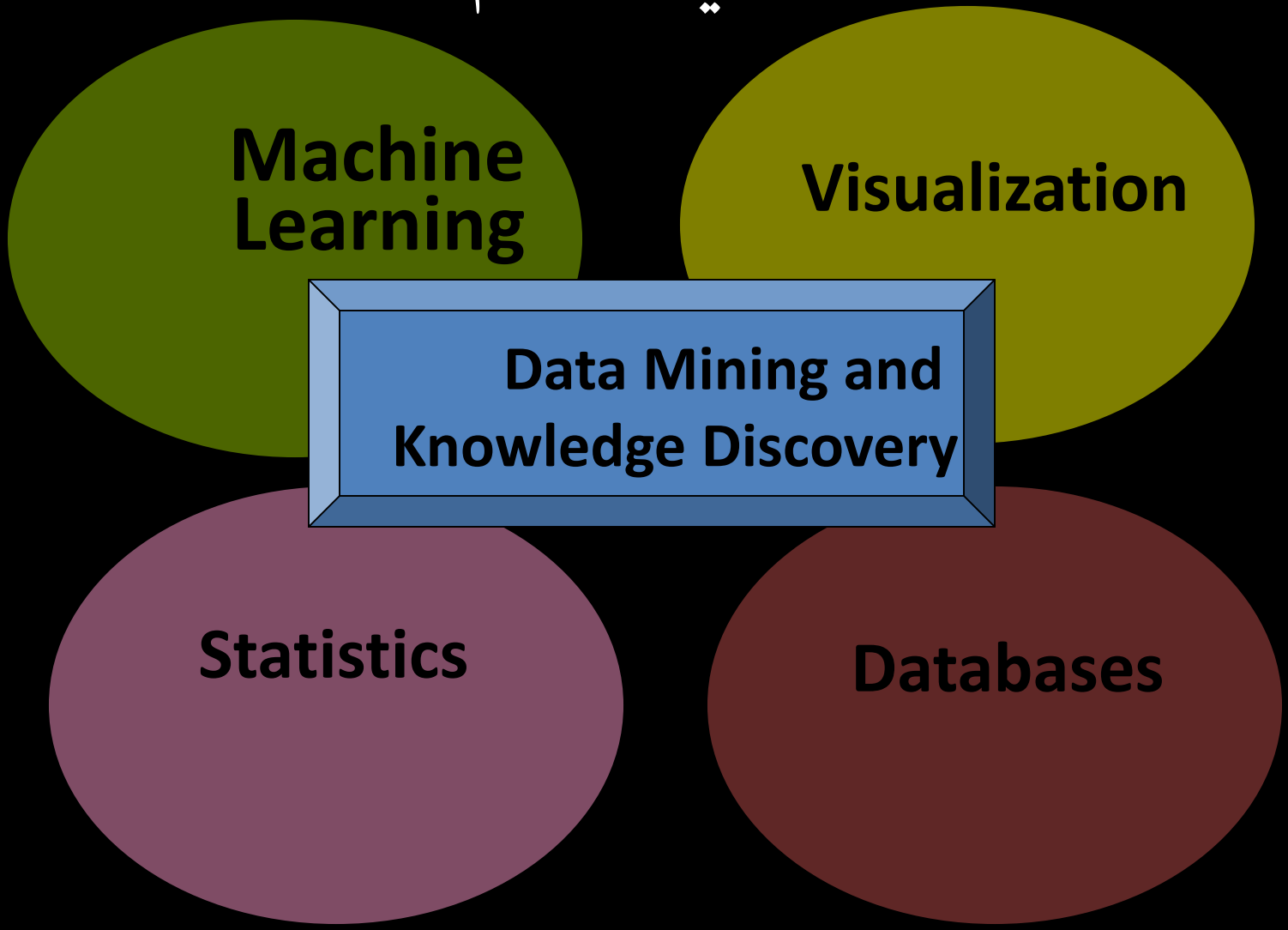


Knowledge Discovery in Data is the

non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

ارتباط داده‌کاوي با علوم مختلف



- هوش مصنوعي در زمينة تحليل آماري، هوش مصنوعي عبارت از شيوه‌هاي خودكار براي به كارگيري روش‌هاي آماري در تحليل داده‌ها است
- الگوريتم يادگيري ماشين، ابزاري است كه مي‌تواند آزمايش‌ها (مثال‌هاي مشاهده شده) را با درنظرگرفتن رده بعضي از توابع و معيارهاي اندازه‌گيري، ياد بگيرد

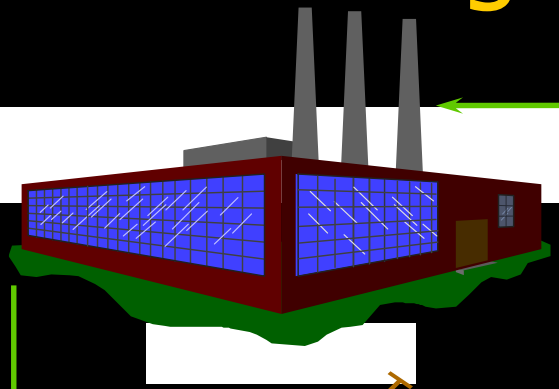
گام‌هاي داده‌كاوي

1) درك قلمرو
2) آماده كردن مجموعه داده‌ها
3) كشف الگوها (داده‌كاوي)
4) پردازش بعد از كشف الگو
5) استفاده از نتايج

گام‌هاي داده‌كاوي

1) بيان مسئله
2) انتخاب داده
3) آماده‌سازي داده‌ها
4) مدل‌سازي
5) ارزيابي مدل
6) توسعه مدل

Knowledge Discovery Process



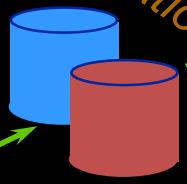
Integration

Interpretation
& Evaluation

Data Mining

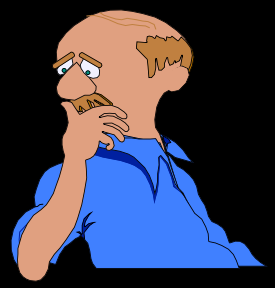
Transformation

Selection
& Cleaning



DATA
Ware
house

Understanding



آمار و داده‌كاوي

- **نوع داده‌ها** در آمار كمي است ولي داده‌كاوي داده‌هاي كمي، كيفي و متني را پوشش مي‌دهد.
- **نوع متغيرهاي ورودي** در آمار عددي ولي در داده‌كاوي متغيرها از نوع عددي، طبقه‌اي و متني مي‌باشند.
- **نوع تمركز** در آمار روي مدل ولي در داده‌كاوي تمركز روي الگو است.
- **زيربخش‌هاي اصلي آمار** برآورد، توزيع‌هاي احتمال، آزمون فرضيه، امتيازبندي مدل و پيشگويي است، ولي زيربخش‌هاي داده‌كاوي مدل‌بندي پيشگويي‌ها، بخش‌بندي پايگاه داده‌ها و ساخت فرضيه است.

آمار و داده‌كاوي

- در آمار جستجو براي دستيابي به نتايج محدود به جستجوهاي جهت‌دار بوده و با نتايج نيز آشنا هستيم. ولي در داده‌كاوي جستجوها اكثراً از طريق كاربر تعيين شده و ممكن است جهت‌دار باشند، اما اصولاً روش خودكار است و نوع نتايج نامعلوم است.
- در مقايسه با آمار، داده‌كاوي توجه كمتري به ويژگيهاي مجانبي استنباط‌هاي بزرگ نمونه‌اي دارد و فلسفه كلي يادگيري، بيشتر شامل ملاحظه پيچيدگي مدل‌ها و محاسباتي كه آنها نياز دارند، است.

آمار و داده‌كاوي

- **هدف اصلي** از به كارگيري روش‌ها در آمار استفاده از برآوردگرها و توزيع‌ها براي ادغام اطلاعات، ولي در داده‌كاوي هدف اصلي كشف دانش مورد علاقه است.
- **شكل نتايج** در آمار به صورت مدل‌هاي كلي برآورد مي‌شود ولي در داده‌كاوي مدل‌هاي موضعي محاسبه مي‌شوند.
- **ارزش اطلاعاتي** نتايج در آمار معلوم و محدود است در حالي كه در داده‌كاوي نامعلوم و نامحدود مي‌باشد.

آمار و داده‌كاوي

- با يك فرضيه شروع مي‌شود
- آمارشناسان بايد رابطه‌هايي را ايجاد كنند كه به فرضيه مربوط شود
- داده‌ها عددي هستند

- به فرضيه احتياجي ندارد
- الگوريتم‌هاي داده‌كاوي خودكار روابط را ايجاد مي‌كند
- از هر نوع داده‌اي مي‌توانند استفاده كنند

آمار و داده‌کاوي

- داده‌هاي نابجا و نادرست ضمن تحليل تشخيص داده مي‌شوند
- داده‌کاوي به داده‌هاي صحيح و درست طبقه‌بندي شده بستگي دارد
- نتايج کار آمارشناسان براي مديران نياز به تفسير ندارد
- تفسير نتايج داده‌کاوي آسان نسيت و براي تحليل و بيان آن‌ها به متخصصان آمار نياز است

آمار و داده‌کاوي

- نمونه‌گيري
- کم کردن هزينه
- مخرب بودن
- زمان‌بر بودن سرشماري
- هزينه محاسباتي
- محدوديت سخت افزار و نرم افزار
- زمان‌بر بودن تحليل

- نمونه‌گيري
- فروکاهي داده‌ها
- اندازه زياد داده‌ها
- ضروري نبودن ذخيره آن‌ها
- هزينه محاسباتي
- محدوديت سخت افزار و نرم افزار
- زمان‌بر بودن تحليل