
بسم الله الرحمن الرحيم

# مبانی داده‌کاوی

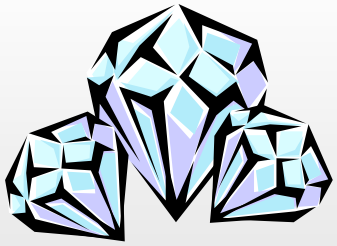
علی شکیبا
ali.shakiba@vru.ac.ir
دانشگاه حضرت ولی‌عصر (عج) رفسنجان

# داده‌کاوی
# Data Mining

# داده‌کاوی
Data Mining

- استخراج الگوها یا دانش جالب

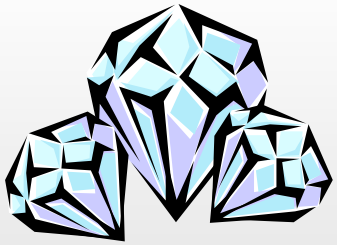
از حجم زیادی از داده



# داده‌کاوی
Data Mining

- استخراج الگوها یا دانش جالب

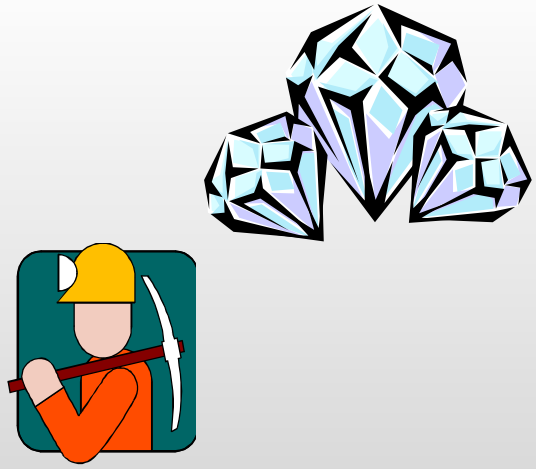
از حجم زیادی از داده



# داده‌کاوی
Data Mining

- استخراج الگوها یا دانش جالب
  - غیر بدیهی (non-trivial)
  - سودمند (useful)
  - ناشناخته (unknown)

از حجم زیادی از داده



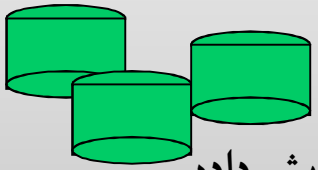
# کشف دانش از داده‌ها

Knowledge Discovery from Data (KDD)

# کشف دانش از داده‌ها
## Knowledge Discovery from Data (KDD)

**1. پالایش داده**

- حذف داده نویز و ناسازگار



پالایش داده

# کشف دانش از داده‌ها

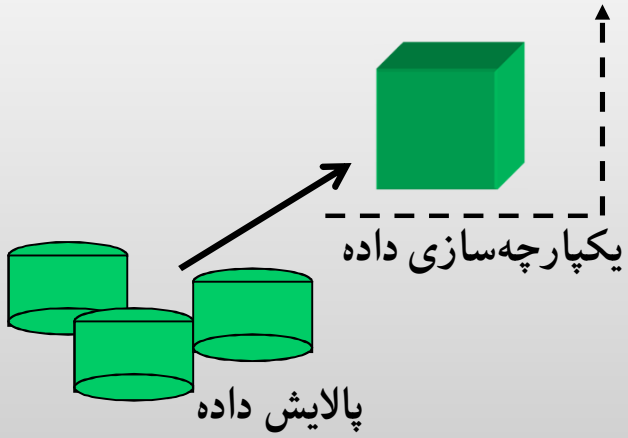
## Knowledge Discovery from Data (KDD)

**1. پالایش داده**

- حذف داده نویز و ناسازگار

**2. یکپارچه‌سازی داده**

- ترکیب از چندین منبع



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

**1. پالایش داده**

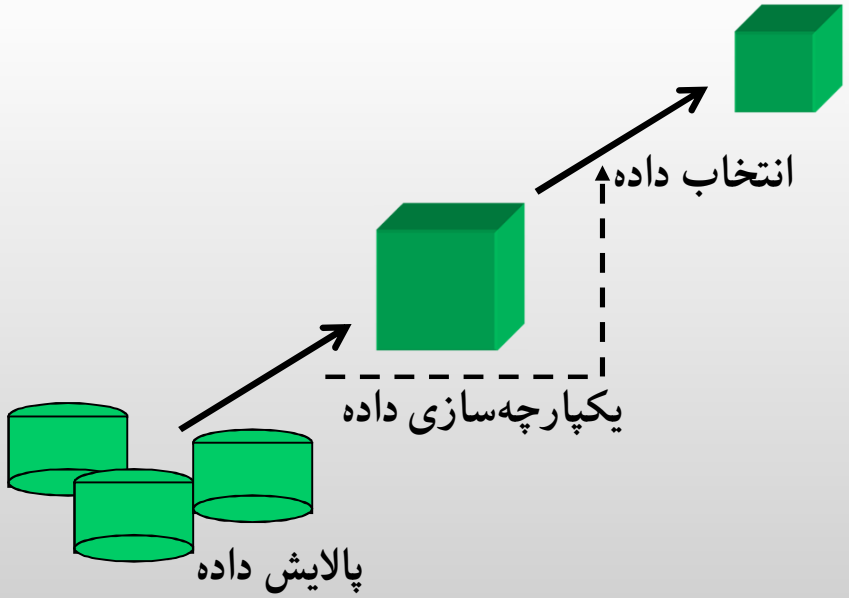
- حذف داده نویز و ناسازگار

**2. یکپارچه‌سازی داده**

- ترکیب از چندین منبع

**3. انتخاب داده**

- بازیابی داده مرتبط با تحلیل



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

**1. پالایش داده**

- حذف داده نویز و ناسازگار

**2. یکپارچه‌سازی داده**

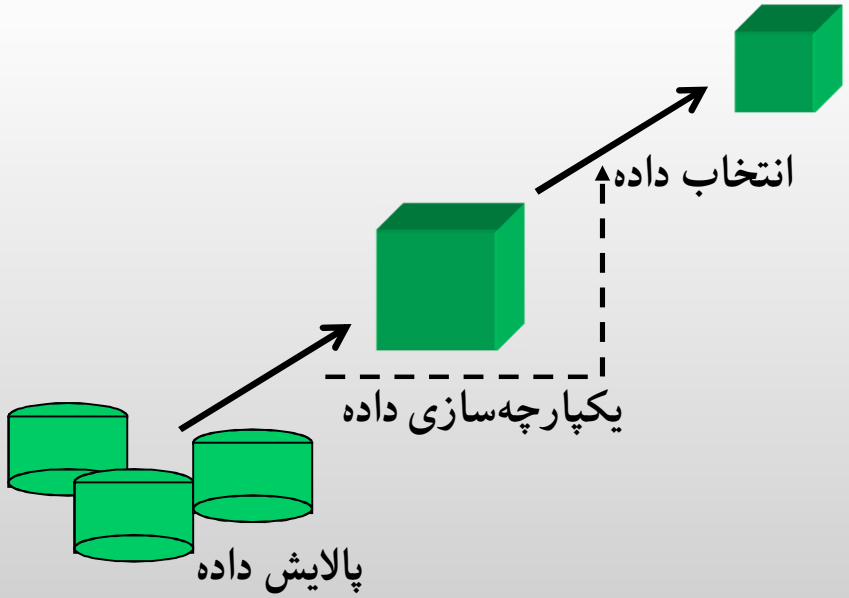
- ترکیب از چندین منبع

**3. انتخاب داده**

- بازیابی داده مرتبط با تحلیل

**4. تبدیل داده**

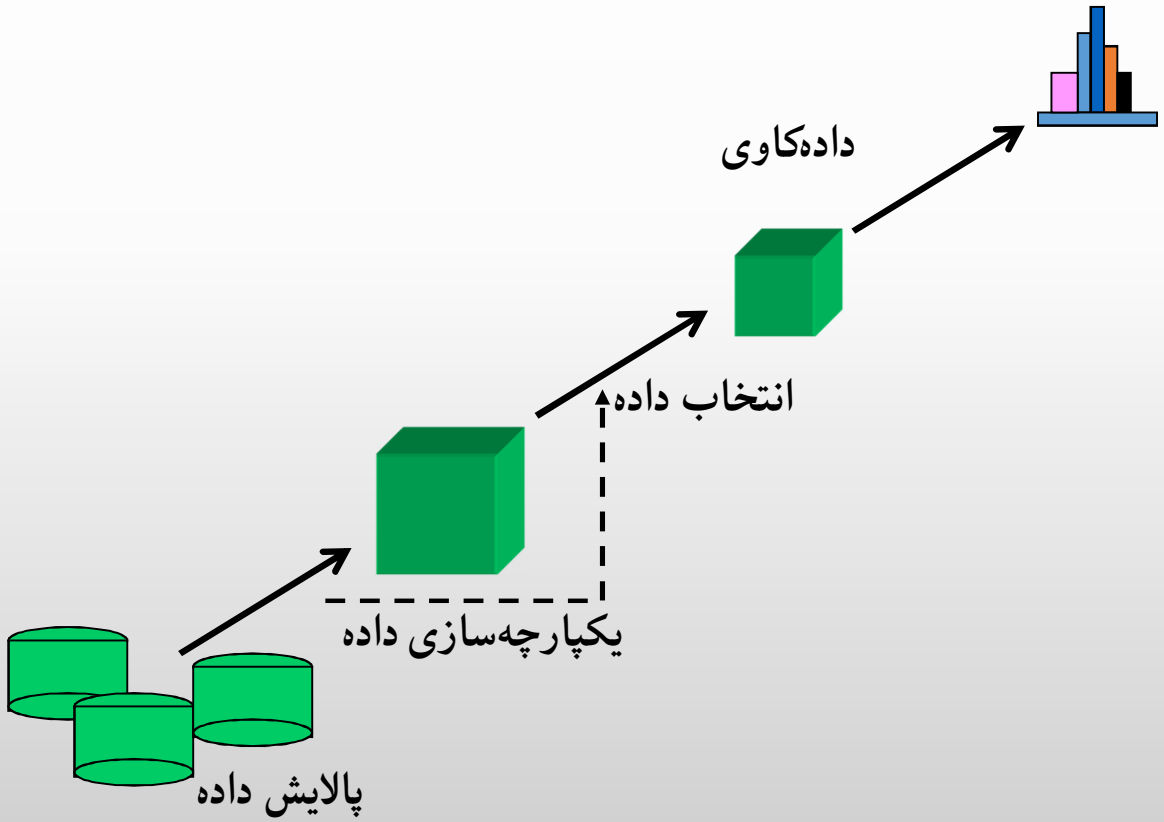
- تبدیل داده به شکل مناسب با تلخیص و تجمیع



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

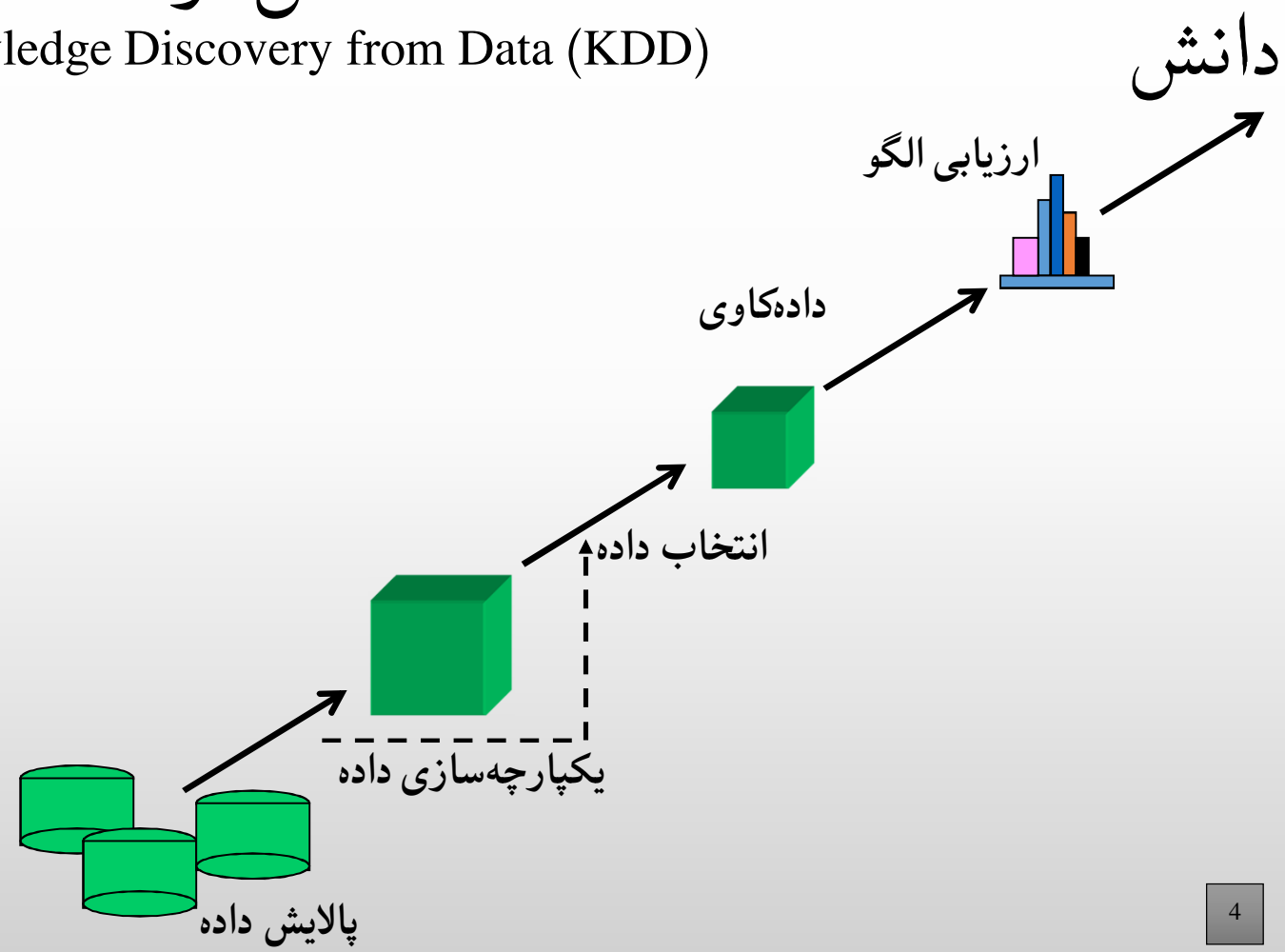
1. **پالایش داده**
   - حذف داده نویز و ناسازگار
2. **یکپارچه‌سازی داده**
   - ترکیب از چندین منبع
3. **انتخاب داده**
   - بازیابی داده مرتبط با تحلیل
4. **تبدیل داده**
   - تبدیل داده به شکل مناسب با تلخیص و تجمیع
5. داده کاوی
   - استخراج الگو از داده با روش‌های هوشمند



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

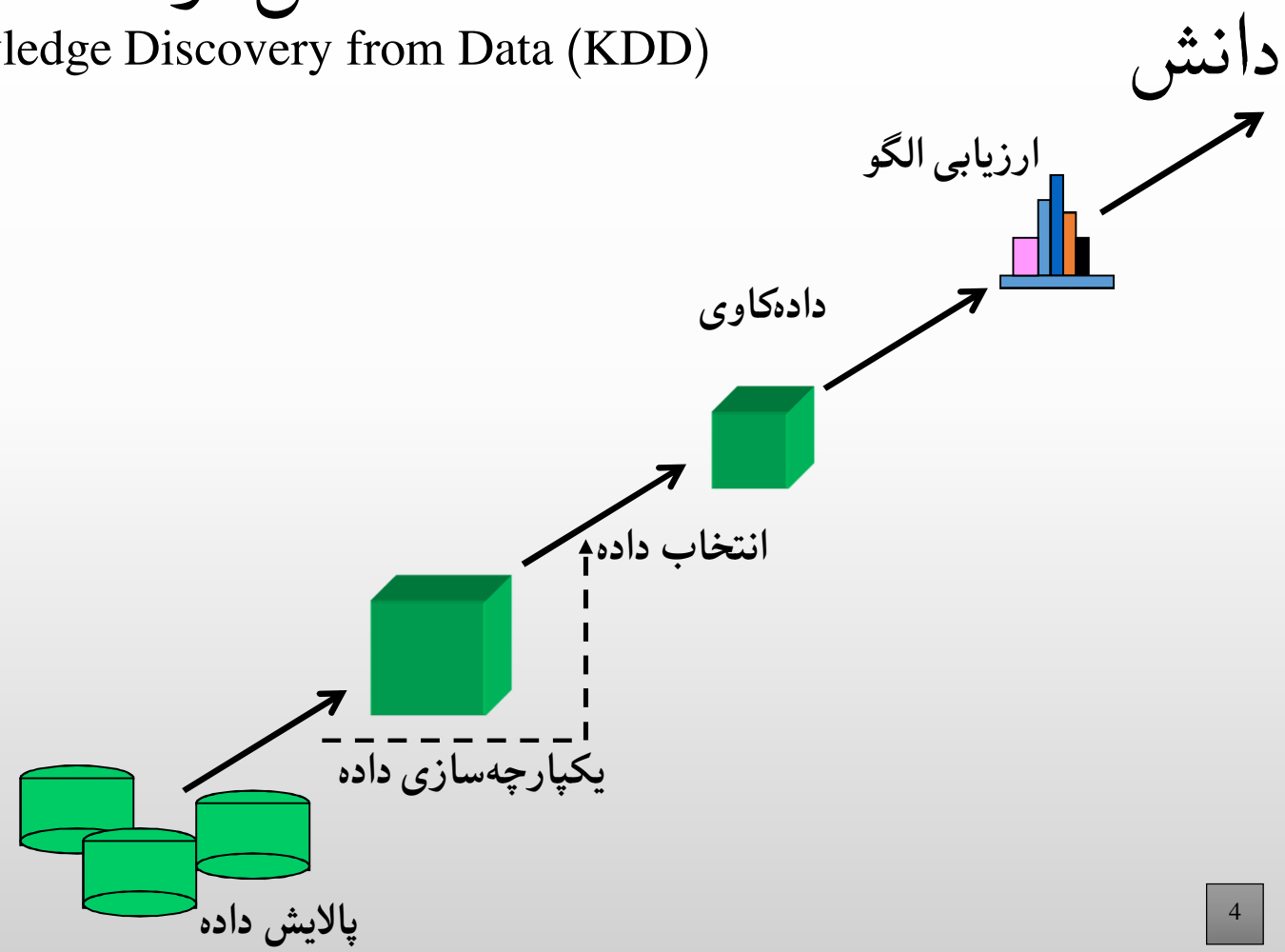
1. **پالایش داده**
   - حذف داده نویز و ناسازگار
2. **یکپارچه‌سازی داده**
   - ترکیب از چندین منبع
3. **انتخاب داده**
   - بازیابی داده مرتبط با تحلیل
4. **تبدیل داده**
   - تبدیل داده به شکل مناسب با تلخیص و تجمیع
5. داده کاوی
   - استخراج الگو از داده با روش‌های هوشمند
6. **ارزیابی الگو**
   - انتخاب بهترین الگوها از میان الگوهای استخراج شده



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

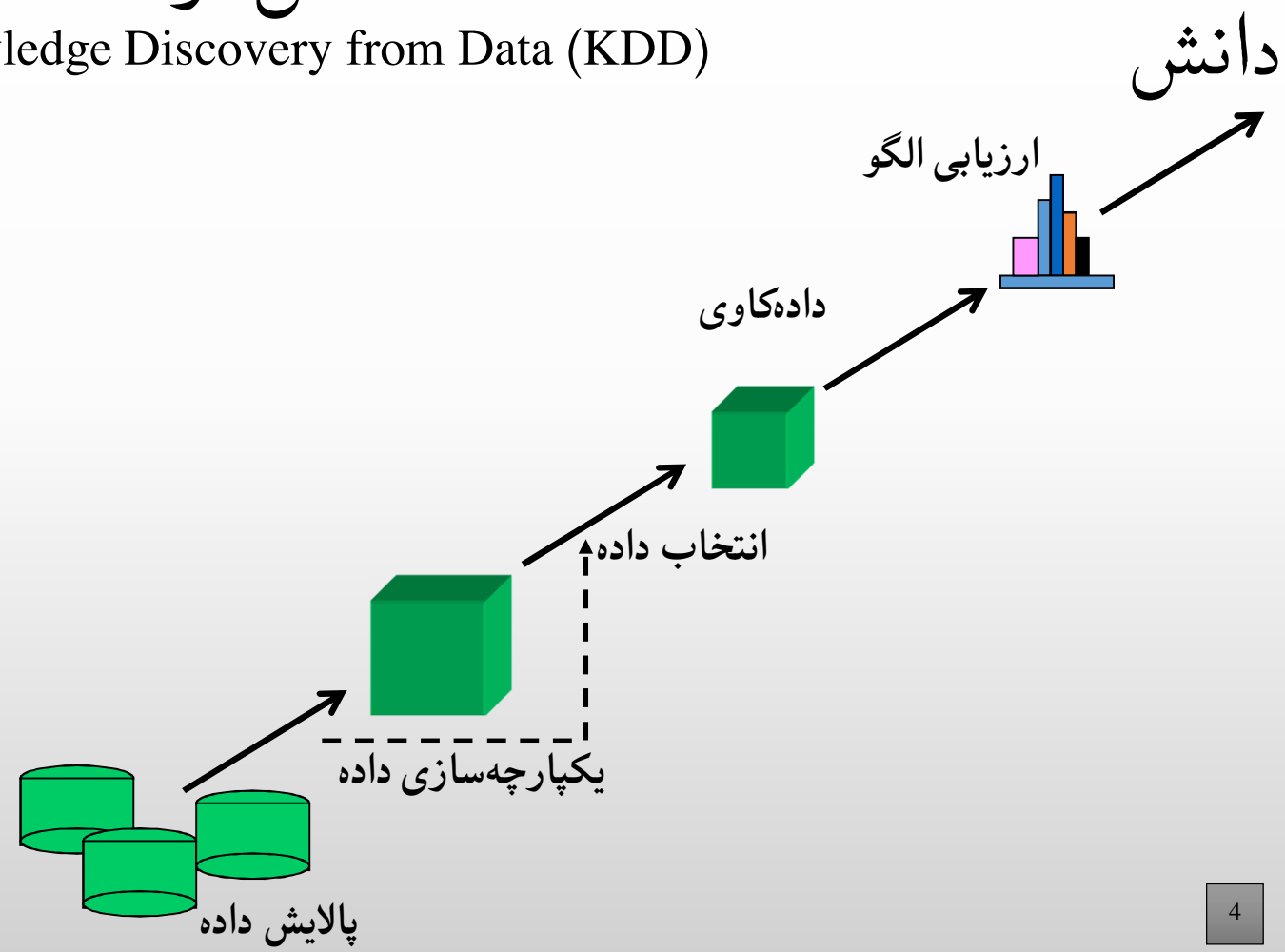
1. **پالایش داده**
   - حذف داده نویز و ناسازگار
2. **یکپارچه‌سازی داده**
   - ترکیب از چندین منبع
3. **انتخاب داده**
   - بازیابی داده مرتبط با تحلیل
4. **تبدیل داده**
   - تبدیل داده به شکل مناسب با تلخیص و تجمیع
5. د
   - استخراج الگو از داده با روش‌های هوشمند
6. **ارزیابی الگو**
   - انتخاب بهترین الگوها از میان الگوهای استخراج شده
7. **ارائه دانش**
   - ارائه دانش کسب‌شده با استفاده از تکنیک‌های دیداری متفاوت



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

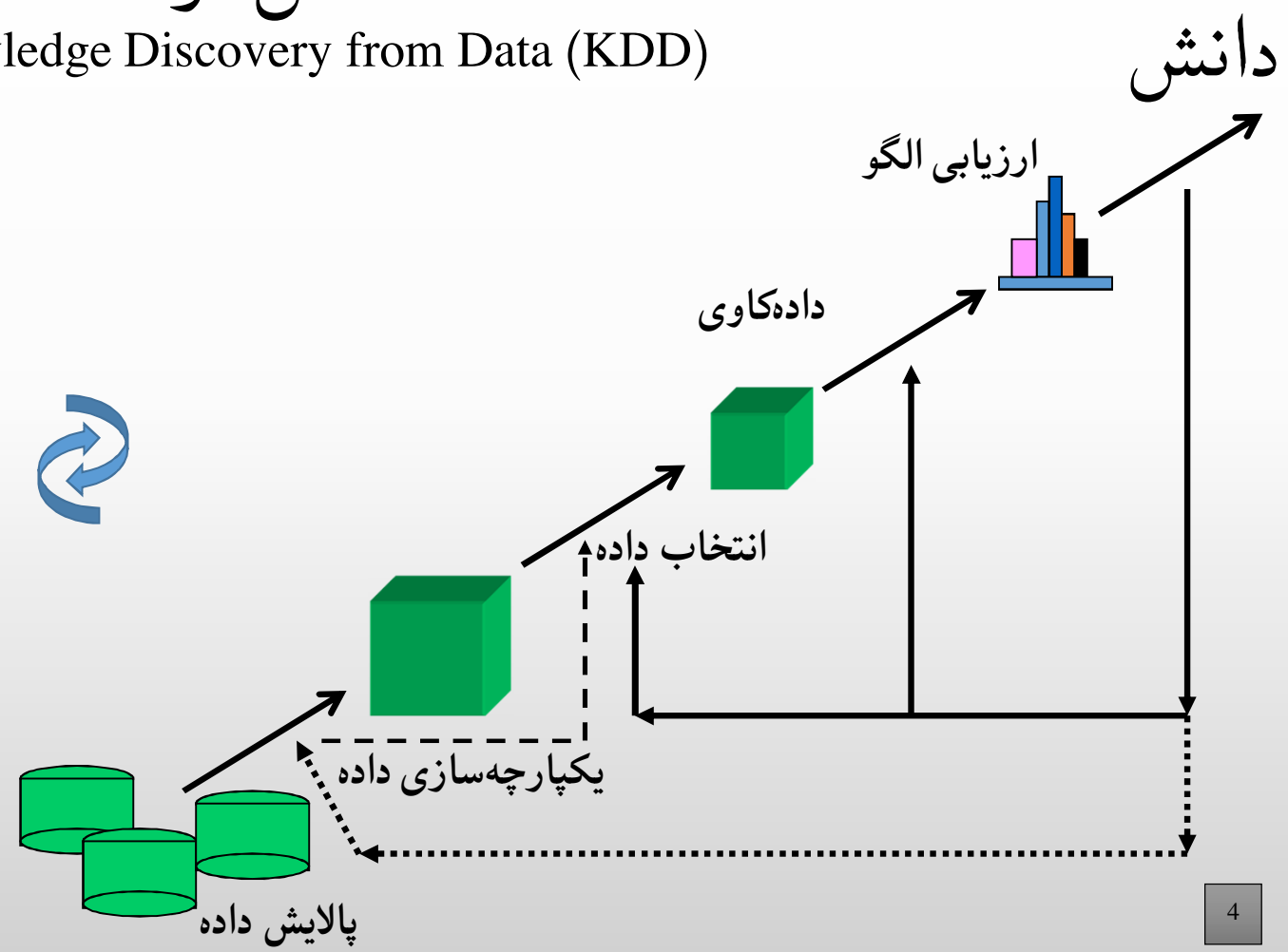
1. **پالایش داده**
   - حذف داده نویز و ناسازگار
2. **یکپارچه‌سازی داده**
   - ترکیب از چندین منبع
3. **انتخاب داده**
   - بازیابی داده مرتبط با تحلیل
4. **تبدیل داده**
   - تبدیل داده به شکل مناسب با تلخیص و تجمیع
5. د
   - استخراج الگو از داده با روش‌های هوشمند
6. **ارزیابی الگو**
   - انتخاب بهترین الگوها از میان الگوهای استخراج شده
7. **ارائه دانش**
   - ارائه دانش کسب‌شده با استفاده از تکنیک‌های دیداری متفاوت



# کشف دانش از داده‌ها

## Knowledge Discovery from Data (KDD)

1. **پالایش داده**
   - حذف داده نویز و ناسازگار
2. **یکپارچه‌سازی داده**
   - ترکیب از چندین منبع
3. **انتخاب داده**
   - بازیابی داده مرتبط با تحلیل
4. **تبدیل داده**
   - تبدیل داده به شکل مناسب با تلخیص و تجمیع
5. د
   - استخراج الگو از داده با روش‌های هوشمند
6. **ارزیابی الگو**
   - انتخاب بهترین الگوها از میان الگوهای استخراج شده
7. **ارائه دانش**
   - ارائه دانش کسب‌شده با استفاده از تکنیک‌های دیداری متفاوت



# چرا داده‌کاوی؟

# چرا داده‌کاوی؟

- تولید حجم وسیعی از داده در طول روز
  - ده‌ها و صدها پتابایت و حتی بیشتر

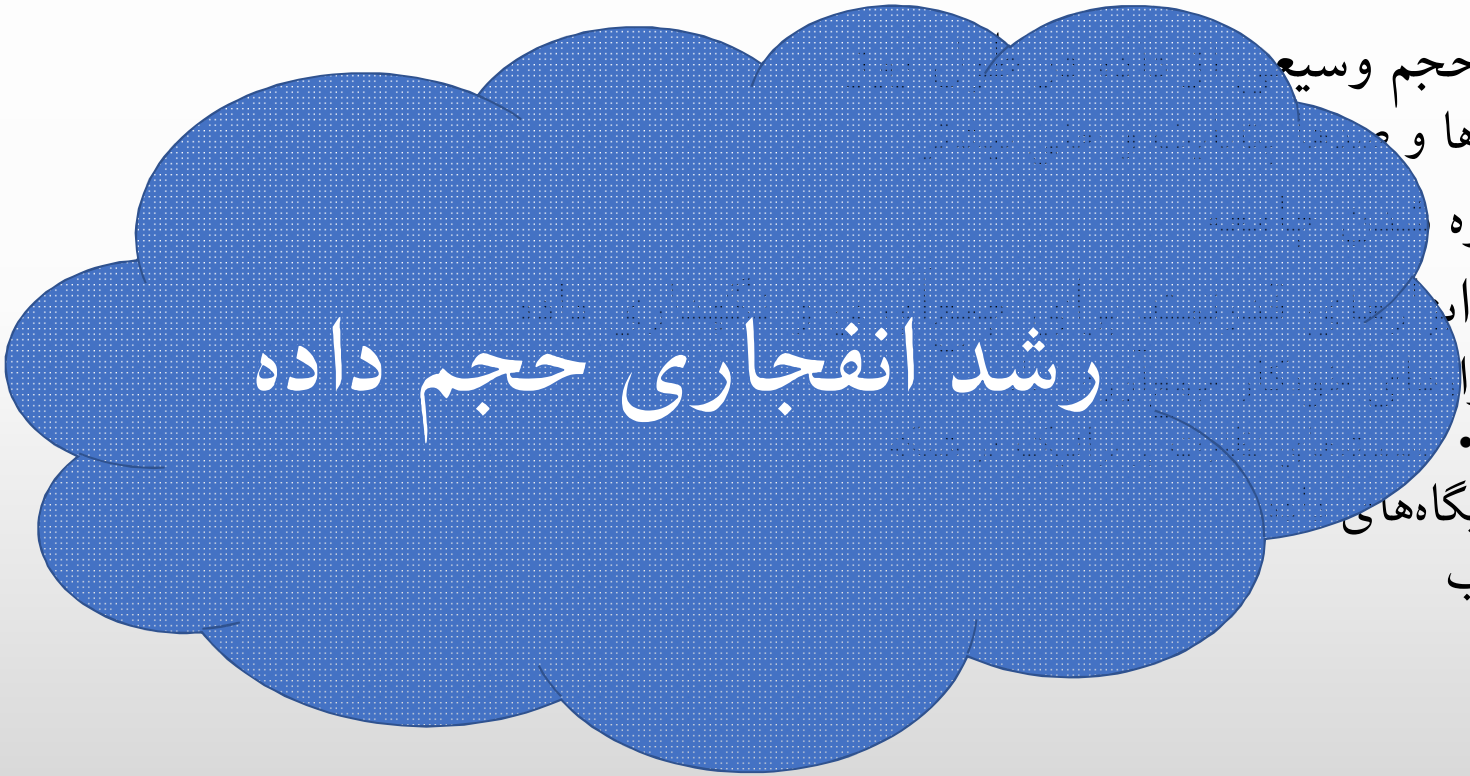
# چرا داده‌کاوی؟

- تولید حجم وسیعی از داده در طول روز
  - ده‌ها و صدها پتابایت و حتی بیشتر
- مکانیزه شدن جامعه
  - کسب‌وکارهای الکترونیکی متعدد
    - ثبت صدها میلیون تراکنش در هفته در فروشگاه‌های بزرگی مانند وال-مارت
  - ثبت و ذخیره‌ی وسیع تجربیات علمی و مهندسی
    - ثبت داده‌ها توسط حسگرهای محیطی مختلف
  - ثبت و ذخیره‌ی حجم وسیعی از داده‌های سلامت
    - داده‌های مربوط به بیماران در بیمارستان‌ها و بیمه‌ها

# چرا داده‌کاوی؟

- تولید حجم وسیعی از داده در طول روز
  - ده‌ها و صدها پتابایت و حتی بیشتر
- مکانیزه شدن جامعه
- وجود ابزارهای قدرتمند برای جمع‌آوری و نگهداری داده
  - ابزارهای خودکار جمع‌آوری داده
    - سیستم‌های نظارت بر ترافیک در شبکه
  - پایگاه‌های داده
  - وب

# چرا داده‌کاوی؟




- تولید حجم وسیع
  - ده‌ها و
- مکانیزه
- وجود ابز
  - ابزا
  -
  - پایگاه‌های
  - وب

# چرا داده کاوی؟ (ادامه)

در داده غرق شده‌ایم؛ اما در قحطی دانش به سر می‌بریم.

# چرا داده کاوی؟ (ادامه)

در داده غرق شده‌ایم؛ اما در قحطی دانش به سر می‌بریم.

نیاز؛ مادر اختراعات است.

**داده‌کاوی: تحلیل خودکار مجموعه دادگان حجیم**

# چه چیزهایی کاوش می‌شوند؟

# چه چیزهایی کاوش می‌شوند؟

- هر نوع داده‌ی معنا دار

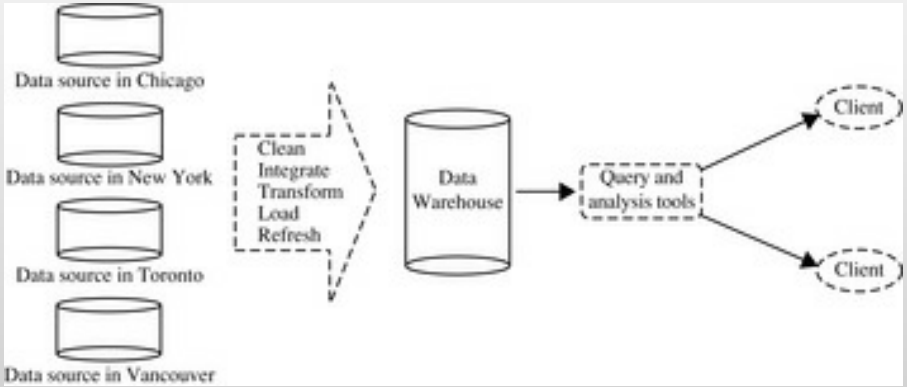
# چه چیزهایی کاوش می‌شوند؟

- هر نوع داده‌ی معنا دار
- شکل‌های معمول
  - پایگاه‌های داده
  - پرس‌وجوی رابطه‌ای

| | |
|---|---|
| *customer* | (*cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .*) |
| *item* | (*item_ID, brand, category, type, price, place_made, supplier, cost, . . .*) |
| *employee* | (*empl_ID, name, category, group, salary, commission, . . .*) |
| *branch* | (*branch_ID, name, address, . . .*) |
| *purchases* | (*trans_ID, cust_ID, empl_ID, date, time, method_paid, amount*) |
| *items_sold* | (*trans_ID, item_ID, qty*) |
| *works_at* | (*empl_ID, branch_ID*) |

# چه چیزهایی کاوش می‌شوند؟ (ادامه)

- هر نوع داده‌ی معنا دار
- شکل‌های معمول
  - پایگاه‌های داده
  - انباره‌های داده
    - مخزنی از اطلاعات که از چندین منبع (مختلف) جمع‌آوری شده و تحت شمای یکسانی نگهداری می‌شوند.
    - فرایندی شامل (۱) پالایش؛ (۲) یکپارچه‌سازی؛ (۳) تبدیل؛ (۴) بارگذاری داده
    - مکعب داده (Data Cube)



# چه چیزهایی کاوش می‌شوند؟ (ادامه)

- هر نوع داده‌ی معنا دار
- شکل‌های معمول
  - پایگاه‌های داده
  - انباره‌های داده
  - داده‌ی تراکنشی
    - هر رکورد حاوی اطلاعاتی درباره‌ی یک تراکنش است.

| *trans_ID* | *list_of_item_IDs* |
|---|---|
| T100 | I1, I3, I8, I16 |
| T200 | I2, I8 |
| ... | ... |

# چه چیزهایی کاوش می‌شوند؟ (ادامه)

- هر نوع داده‌ی معنا دار
- شکل‌های معمول
  - پایگاه‌های داده
  - انباره‌های داده
  - داده‌ی تراکنشی
- گونه‌های دیگر داده
  - داده‌ی زمانی
  - داده‌ی توالی
  - داده‌ی جریانی
  - داده‌ی فضایی
  - داده‌ی چندرسانه‌ای
  - گراف‌ها
  - ...

# چه نوع الگوهایی کاوش می‌شوند؟

- ...

# چه نوع الگوهایی کاوش می‌شوند؟

- توصیف و تفکیک
- ...

# چه نوع الگوهایی کاوش می‌شوند؟

- توصیف و تفکیک
- الگوهای مکرر

- ...

# چه نوع الگوهایی کاوش می‌شوند؟

- توصیف و تفکیک
- الگوهای مکرر
- مشارکت‌ها و وابستگی‌ها

- ...

# چه نوع الگوهایی کاوش می‌شوند؟

- توصیف و تفکیک
- الگوهای مکرر
- مشارکت‌ها و وابستگی‌ها
- دسته‌بندی

- ...

# چه نوع الگوهایی کاوش می‌شوند؟

- توصیف و تفکیک
- الگوهای مکرر
- مشارکت‌ها و وابستگی‌ها
- دسته‌بندی
- رگرسیون

- ...

# چه نوع الگوهایی کاوش می‌شوند؟

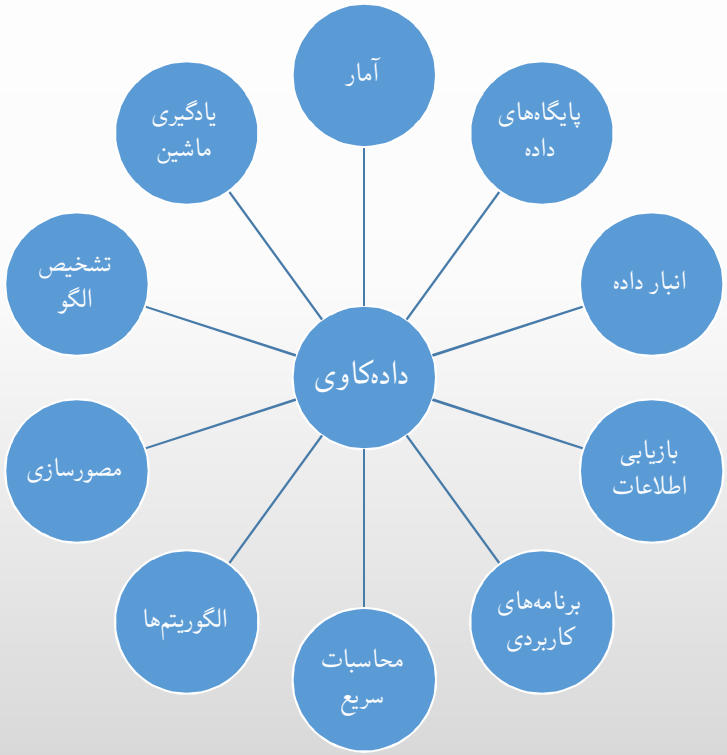
- توصیف و تفکیک
- الگوهای مکرر
- مشارکت‌ها و وابستگی‌ها
- دسته‌بندی
- رگرسیون
- خوشه‌بندی

- ...

# چه نوع الگوهایی کاوش می‌شوند؟

- توصیف و تفکیک
- الگوهای مکرر
- مشارکت‌ها و وابستگی‌ها
- دسته‌بندی
- رگرسیون
- خوشه‌بندی
- تحلیل داده‌های پرت
- ...

# کدام فناوری‌ها و تکنیک‌ها؟



# چه کاربردهایی؟

- هوش تجاری
- موتورهای جستجوی وب
- سیستم‌های پیشنهاددهنده
- تحلیل سبد خرید
- تحلیل داده زیستی و پزشکی
- ...

# موضوعات عمده در داده‌کاوی

- متدولوژی کاوش
  - کاوش گونه‌های جدید و مختلف دانش
  - کاوش دانش در فضای چند بعدی
  - داده‌کاوی؛ کوششی میان رشته‌ای
  - ارتقا قدرت کشف دانش در محیط شبکه‌بندی شده
  - کنترل عدم‌قطعیت، نویز و ناقص‌بودن داده
  - ارزیابی الگوها و هدایت فرایند کاوش

# موضوعات عمده در داده‌کاوی

- متدولوژی کاوش
  - کاوش گونه‌های جدید و مختلف دانش
  - کاوش دانش در فضای چند بعدی
  - داده‌کاوی؛ کوششی میان رشته‌ای
  - ارتقا قدرت کشف دانش در محیط شبکه‌بندی شده
  - کنترل عدم‌قطعیت، نویز و ناقص‌بودن داده
  - ارزیابی الگوها و هدایت فرایند کاوش
- تعامل کاربر
  - کاوش تعاملی
  - درآمیختن دانش پس زمینه
  - ارائه و مصورسازی نتایج داده‌کاوی

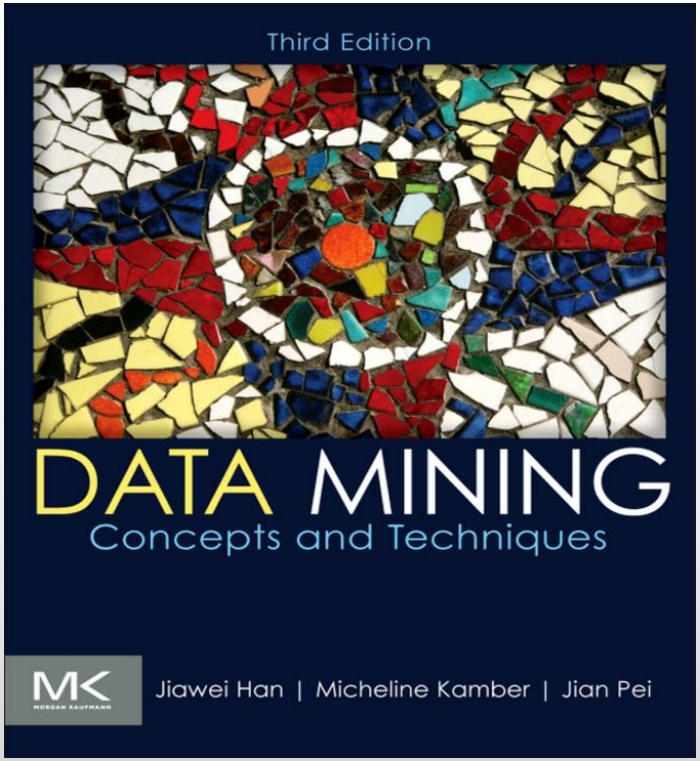
# موضوعات عمده در داده‌کاوی (ادامه)

- کارآمدبودن و قابلیت مقیاس‌پذیری
- کارآیی و مقیاس‌پذیری الگوریتم‌های داده‌کاوی
- الگوریتم‌های کاوش موازی؛ توزیع شده؛ جریانی و افزایشی (Incremental)

# موضوعات عمده در داده‌کاوی (ادامه)

- کارآمدبودن و قابلیت مقیاس‌پذیری
  - کارآیی و مقیاس‌پذیری الگوریتم‌های داده‌کاوی
  - الگوریتم‌های کاوش موازی؛ توزیع شده؛ جریانی و افزایشی (Incremental)
- تنوع گونه‌های داده
  - داده‌ی پیچیده
  - کاوش از مخازن داده‌ی پویا؛ شبکه‌بندی شده و سراسری (Global)
  - داده‌کاوی و جامعه
  - تاثیر اجتماعی داده‌کاوی
  - حفظ محرمانگی و داده‌کاوی
  - داده‌کاوی غیرمحسوس و غیرقابل مشاهده

# مرجع اصلی



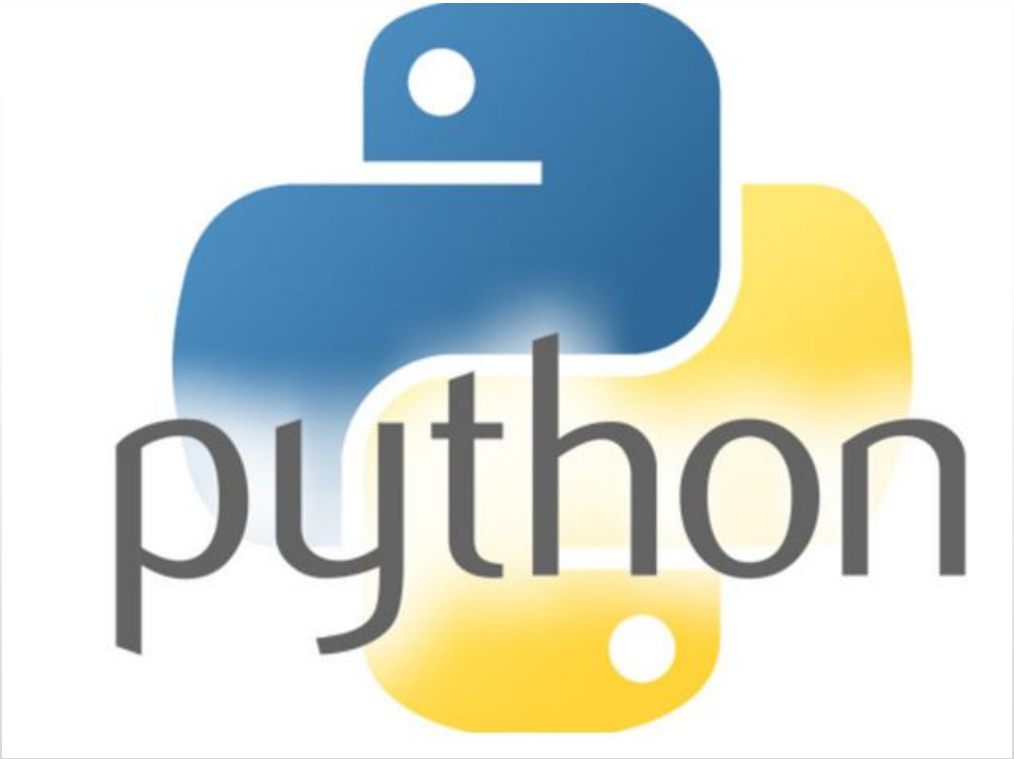

*Jiawei Han, Micheline Kamber, and Jian Pei. 2011.* ***Data Mining: Concepts and Techniques (3rd ed.)****. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.*

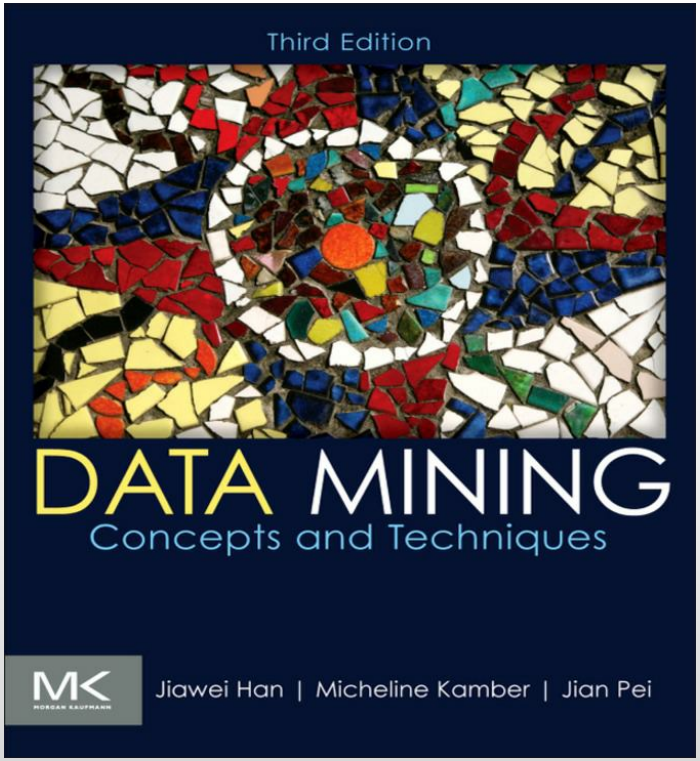
هان، ژ.، کمبر، م. و پی، ژ.، اسماعیلی، مهدی (م.) ۱۳۹۳. داده‌کاوی: مفاهیم و کاربردها (ویراست سوم). نیاز دانش، تهران، ج.ا.ایران.

# اطلاعات مهم

| عنوان | بارم | توضیحات |
|---|---|---|
| میان‌ترم | ۴ | دوشنبه؛ ۱۱ اردیبهشت ۱۳۹۶ – ساعت کلاسی |
| پایان‌ترم | ۵ | طبق سیستم گلستان |
| تمرین | ۵ | تقریبا؛ هر هفته یک سری (محاسباتی؛ برنامه‌نویسی و ...) |
| پروژه | ۶ | چند فاز؛ انفرادی و تیمی |
| تلاش بیشتر | +۲ | بسته به عملکرد فرد |
| جمع | ۲۰+۲ | |

# ابزار مورد استفاده






- آنچه آموختیم:
  - فصل ۱: مقدمه
- جلسه‌ی آینده:
  - فصل ۲: شناخت داده‌ها