Data Science Overview

Data Analytics, Data Mining

Decision Science, Business Intelligence

Artificial Intelligence, Machine Learning

Statistical Learning, Data Science

Sexiest Job of the twenty-first century

- by Harvard Rev

Data Science is new electricity

- by Andrew Ng

Data Science Overview

Why it becomes so popular now?

What are the problems does Data Analytics try to solve?

How does it work? When it will be useful and relevant?

What are the criteria that determine if it succeed in delivering its goal?

What are all these Data Warehouse, Data Mart, ETL Process, Business Intelligence, OLAP, Dashboard, Data Mining, Machine Learning terms mean?

How to become a successful Data Scientist? What are the required skillset?

Why Data Science becomes so popular

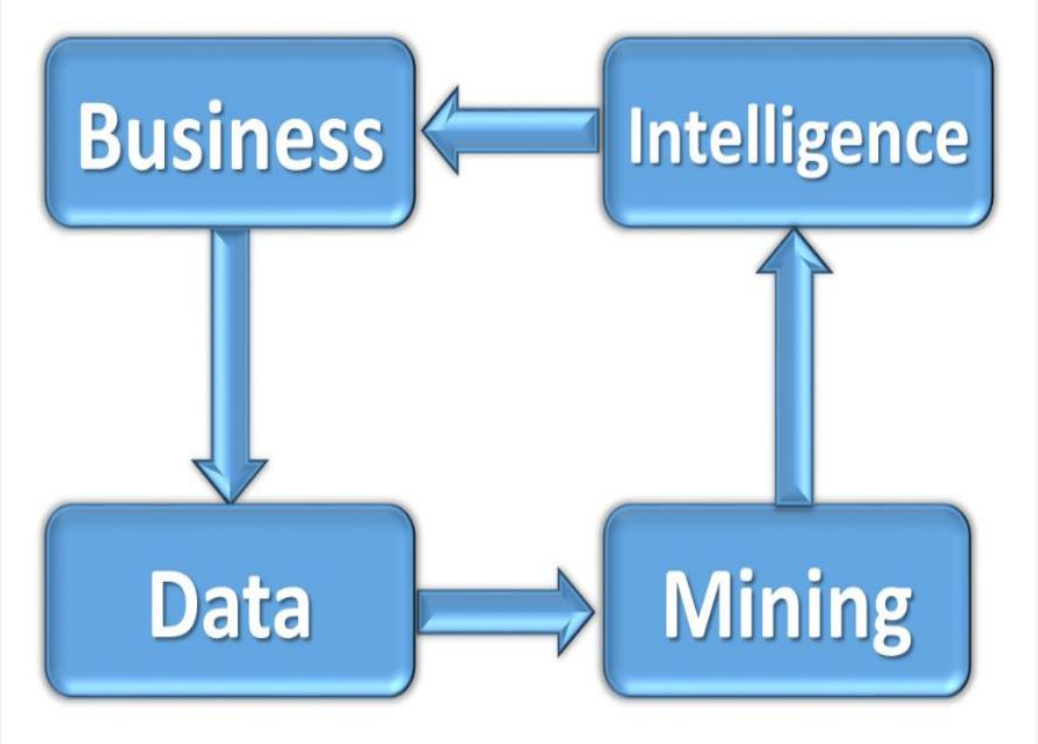
- More and more data available b/c web 2.0
- Faster and Faster Hardware
- Social Media, Availability of unstructured data/alternative data
- Deep Neural Network
- Real-life Proof: Alpha-Go, Alexa, Driverless car ….

It answers questions and solve business problems

Goal of Data Analytics

- *Answer basic strategic business questions* based on data a company has collected.
  - What is the best-selling beer brand.
  - Does reducing classroom size (teacher to students ratio) really improve student performance?

- Discover not-so-obvious *patterns* from the data



BIDM Cycle

- Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place.
- Business activities are recorded on paper or using electronic media, and then these records become data. There is more data from customers' responses and on the industry as a whole.
- All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning.
- These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs. And the cycle continues on

Real-life example of Data Analytics Solutions

- Example from the textbook: MoneyBall

- Another real-life Example:

Most Expensive State to buy a home

The Most Expensive State to Buy a Home in Isn’t New York or California

https://www.barrons.com/articles/most-expensive-state-to-buy-a-home-51567108698

https://smartasset.com/mortgage/cheapest-states-to-buy-a-home-2019?mod=article_inline

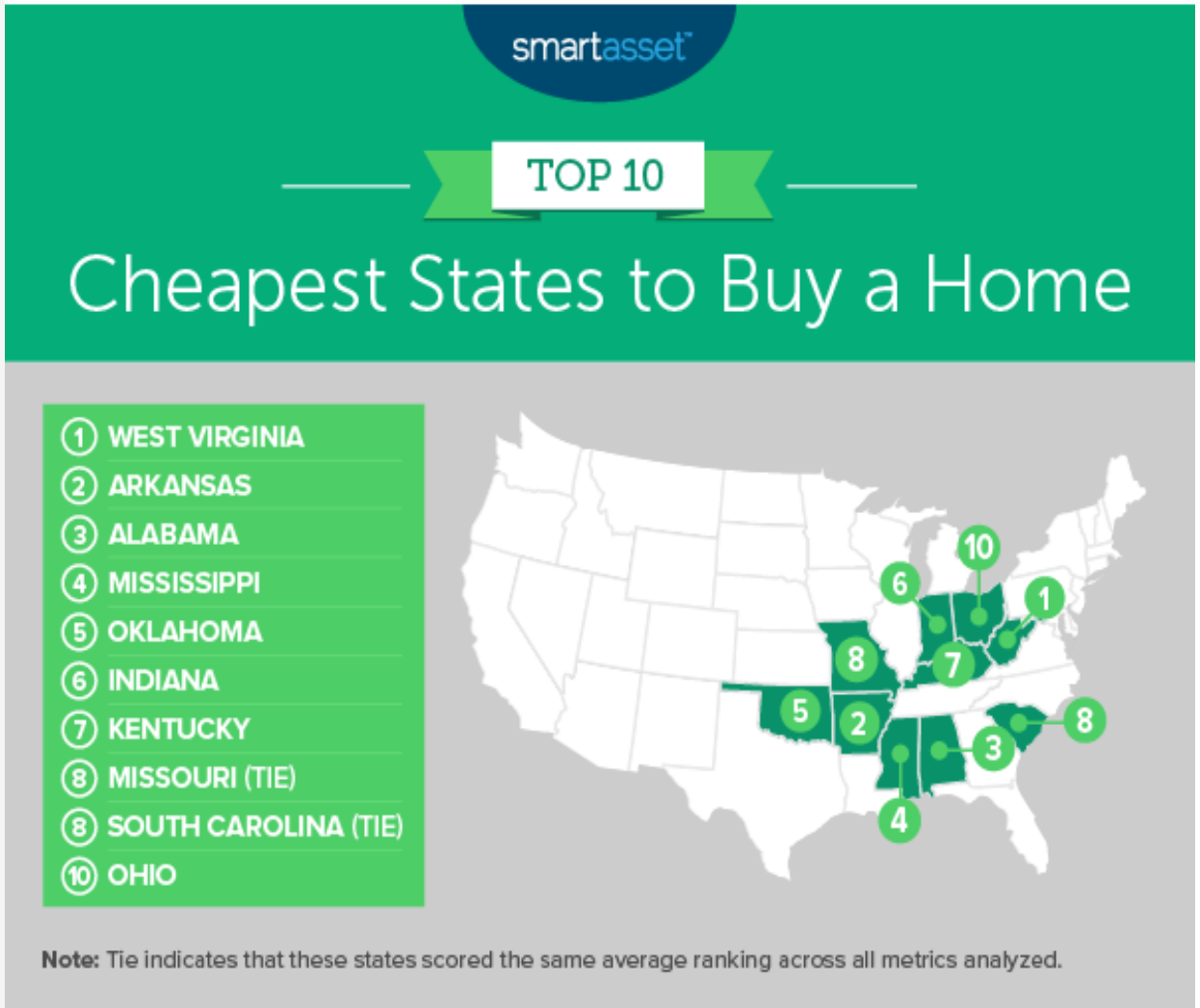
Data Science Real Life Example

Photograph by Breno Assis

Massachusetts ranked as the most expensive state to buy a home in, according to a report from personal-finance website SmartAsset. The analysis assessed 48 states and Washington, D.C., based on the following metrics (Delaware and Louisiana were not included due to insufficient data):

- Effective property tax rate, based on U.S. Census Bureau data
- Median listing price and price per square foot, according to Zillow ZG
- Median value for homes in the bottom third of the market
- Average closing costs, according to SmartAsset's own closing cost calculator

All the states were ranked for each of these metrics, and then researchers calculated their average ranking. This then determined the index value they were assigned on a scale from 0 to 100, with 100 being the cheapest. Massachusetts' index came in at zero, while West Virginia was the most affordable state with a value of 100.



Extra Credits (Class Participation)

Find a similar news article where it clearly showcases how Data Analytics can solve or answer interesting questions.

I will start a discussion thread on this. No explanation is needed. But you are more than welcome to describe how the article does help you realize how data analytics is realized in real-life

"Data Science" consists of two words:

Data and Science

Let's start with Data

Different Type of Data

Structured Data

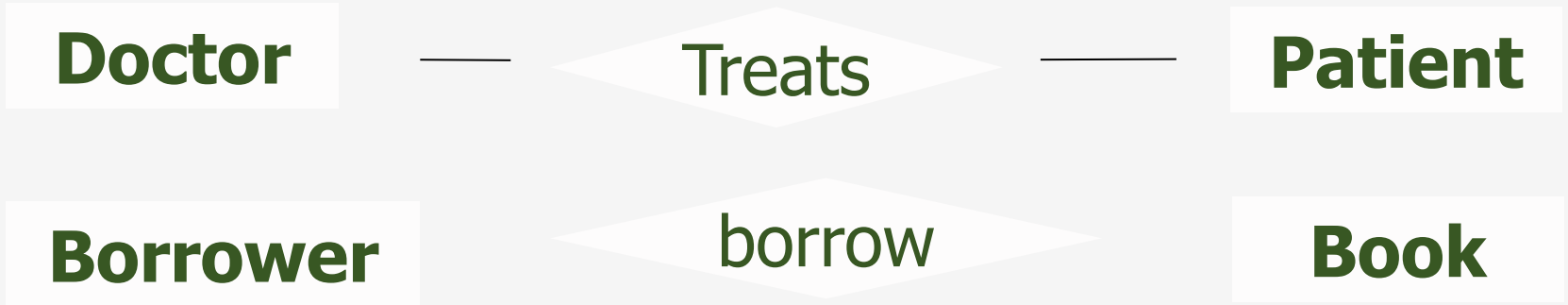
- Tabular Data, Typically stored in a relational database table
  - Example: Think in terms of managing computer system in keeping track of books in a library
    - A Borrower Table with borrower ID, name, birth-date, address as columns
    - A Book table with author, name of the book, published date, publisher as columns

- Storage format
  - CSV file, an excel file or a JSON file, but more commonly in a SQL database

- Type of data: Numerical fields, categorical field, date-time

Structured Data (database and data modeling)

- Data is organized in tables
  - Tables relate to entities and relationships among entities
  - Entities are nouns like Doctor, Patient
  - Relationships are verbs like ‘Treats’
- Data tables can be managed using SQL language in simple declarative mode



Different Type of Data

Unstructured Data

- Blob (binary large objects)
- Satellite Data
- Graph to links friends on Facebook.
- Image Data
- Video with annotation
- Twitter feeds
- Time-Series data

Storage format:

- XML, JSON

Type of Databases

- Relational databases
  - MSSQL
  - MySQL
  - Postgres
- NoSQL databases
  - MongoDB
- Time Series Databases

Common Databases for storing data

SQL Database:

- Microsoft SQL Server (https://www.microsoft.com/en-us/sql-server/sql-server-editions-express)
- MySQL (https://www.mysql.com/)
- Postgres SQL (https://www.postgresql.org/)

NoSQL Database:

- MongoDB (https://www.mongodb.com/)

Data Science is the new electricity

Data is a new set of asset class

How to collect and manage Data ?

Expectation vs Reality

Expectation:

Your CSCI-381 /780 instructor

Expectation vs Reality

Reality:

Your CSCI-381 /780 instructor

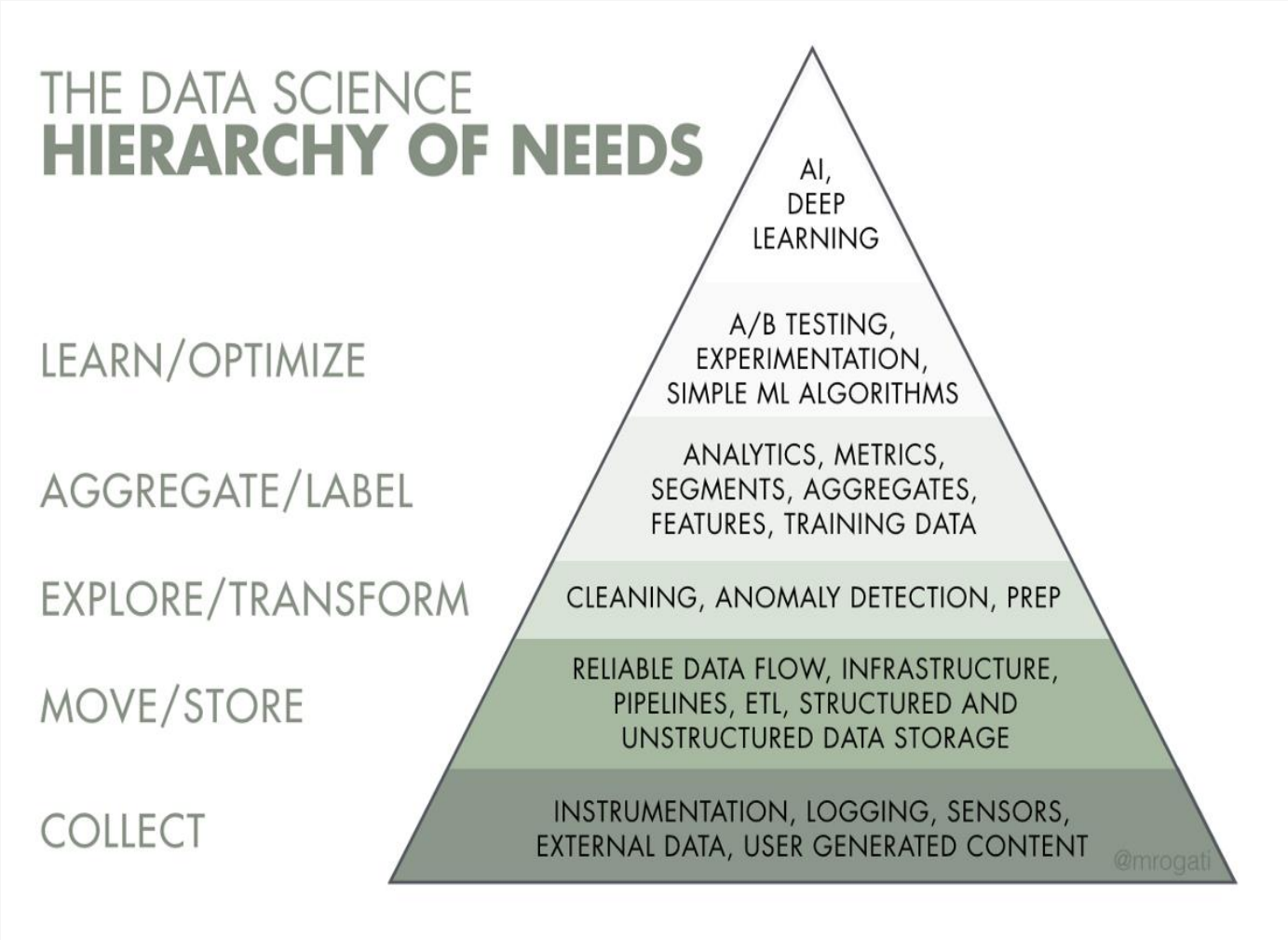
Three type of Data Science jobs

Data Science: Expectation and Reality

https://www.youtube.com/watch?v=8LucP1wiX1g&t=214s

3 common job classifications:

- Data Engineers
- Data Analytics
- Data Scientists



Data as a new natural resource or a new asset class

Data Processing Chain, Data Pipeline, Data Flow

Datafication, Internet of Things

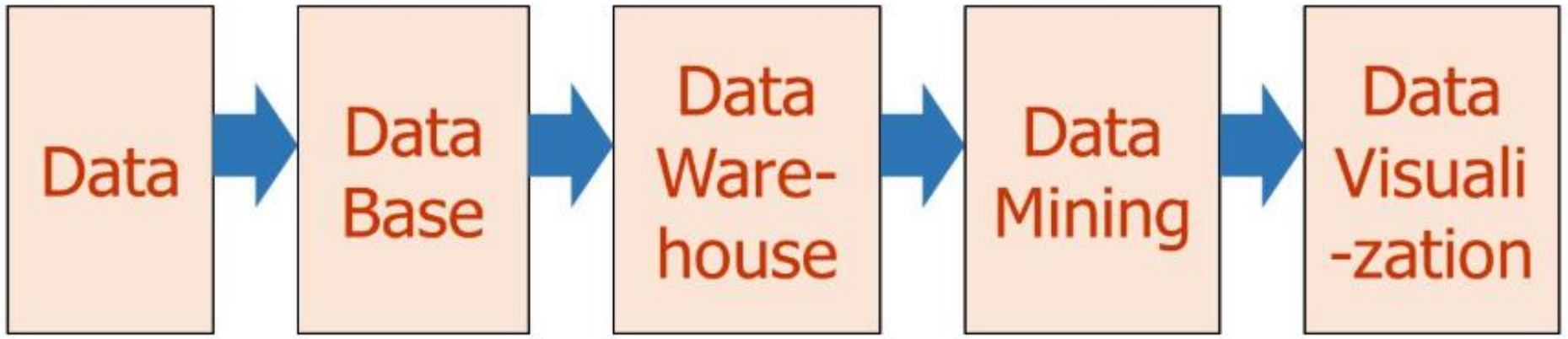
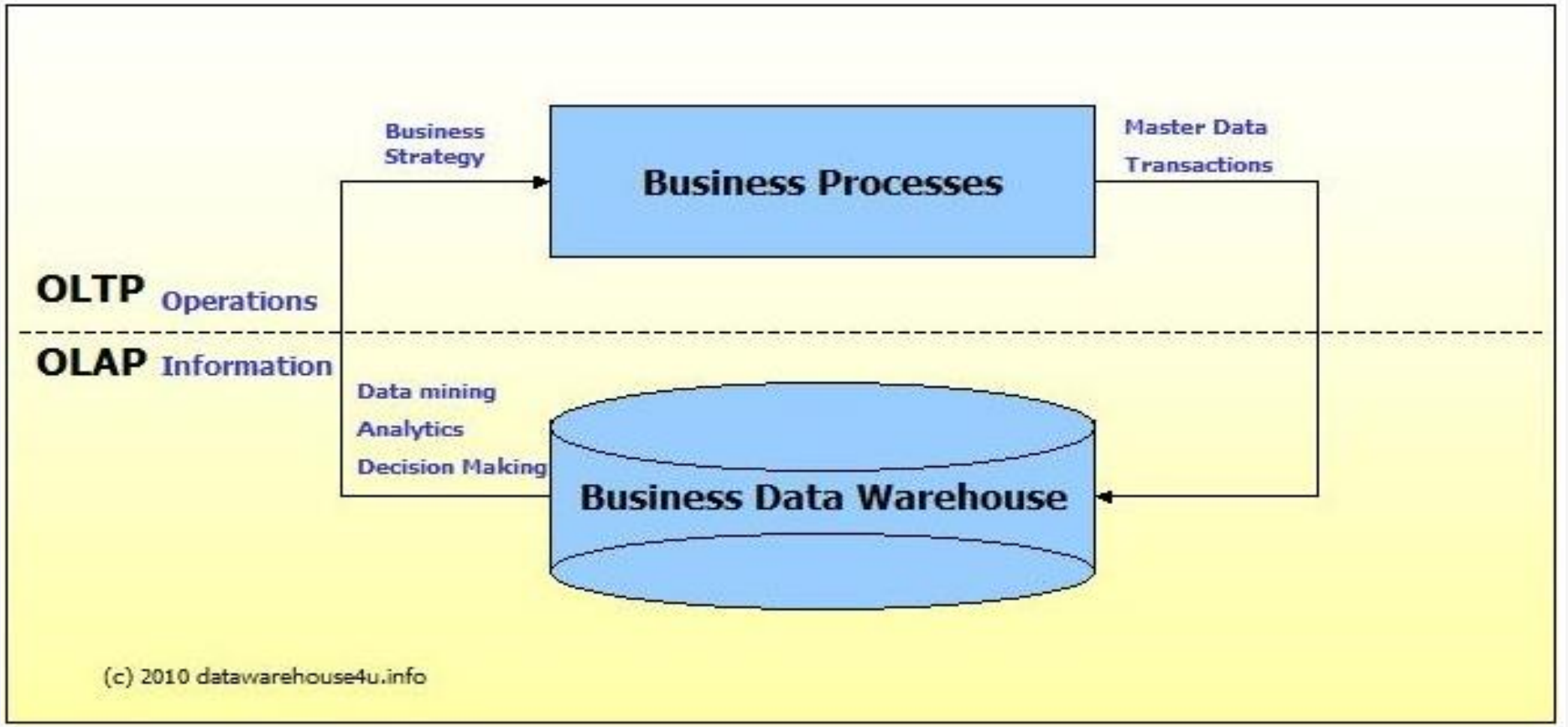


Figure 1.2: Data Processing Chain

Database vs Data Warehouse

OLTP (On-Line Transaction Processing) vs OLAP (On-Line Analytical Processing)



Data Warehouse Definition

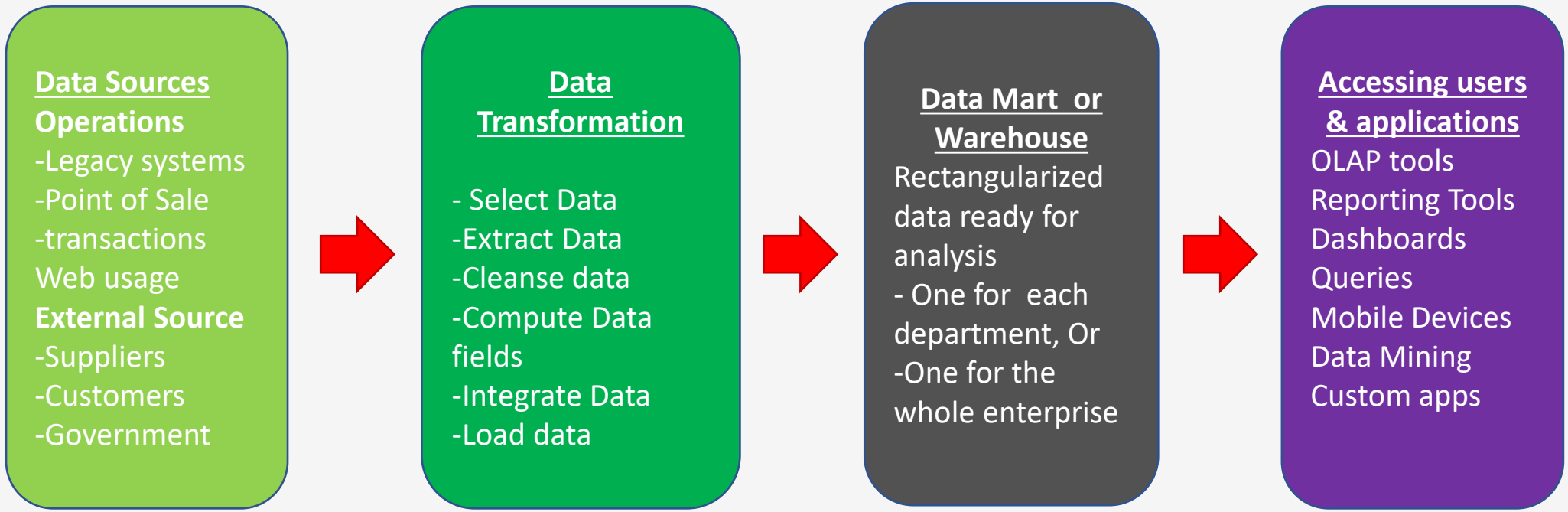
Data warehouse (DW) is an organized collection of integrated, subject-oriented databases designed to support business decision functions.

- organized at the right level of granularity (usually some unit of time)
- provides clean enterprise-wide data in a standardized format for reports, queries and analysis.
- has to be constantly kept up-to-date to be useful
- Has clear Metadata (Data about data)

A DW is physically and functionally separate from an operational transactional database.

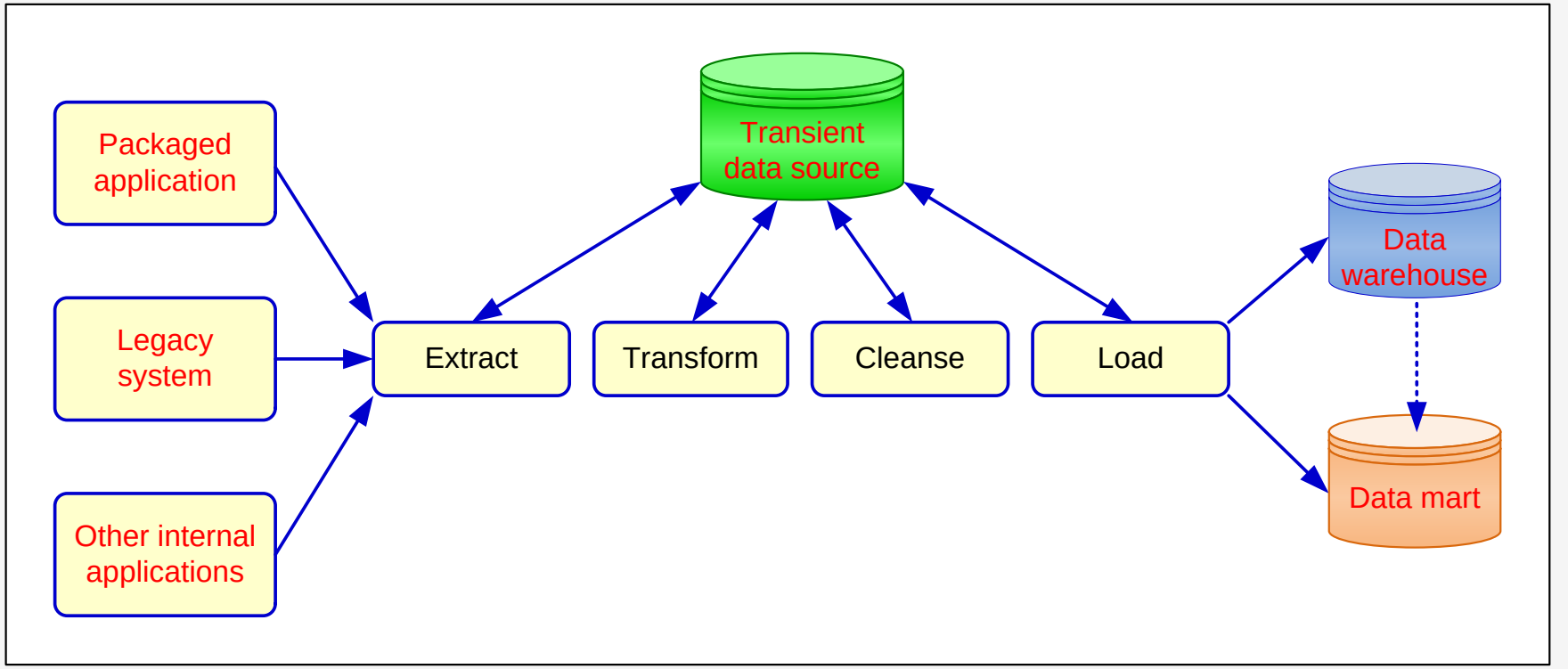
- Creating a DW for analysis and queries represents significant investment in time and effort.

A Conceptual Framework for Data Warehouse



ETL Process for creating a Datawarehouse

Extraction, Transformation, and Load (ETL) process



DataWarehouse Concepts

Watch this 8 min video presentation on Data Warehousing concepts and approaches

https://www.youtube.com/watch?v=zTs5zjSXnvs

Quick reminder of what is 1-2-3 Normal form
https://www.dummies.com/programming/sql/sql-first-second-and-third-normal-forms/

Database vs Data Warehouse

Function	Database	Data Warehouse
Purpose	Data stored in databases can be used for many purposes including day-to-day operations	Data in DW is cleansed data useful for reporting and analysis
Granularity	Highly granular data including all activity and transaction details	Lower granularity data; rolled up to certain key dimensions of interest
Complexity	Highly complex with dozens or hundreds of data files, linked through common data fields	Typically organized around a large fact tables, and many lookup tables
Size	Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size.	Grows as data from operational databases is rolled-up and appended every day. Data is retained for long-term trend analyses
Architectural choices	Relational, and object-oriented, databases	Star schema, or Snowflake schema
Data Access mechanisms	Primarily through high level languages such as SQL. Traditional programming access DB through Open DataBase Connectivity (ODBC) interfaces	Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools

What's next

Now we have the data stored in Data Warehouse

What's next?

Data Mining
Data Visualization
…