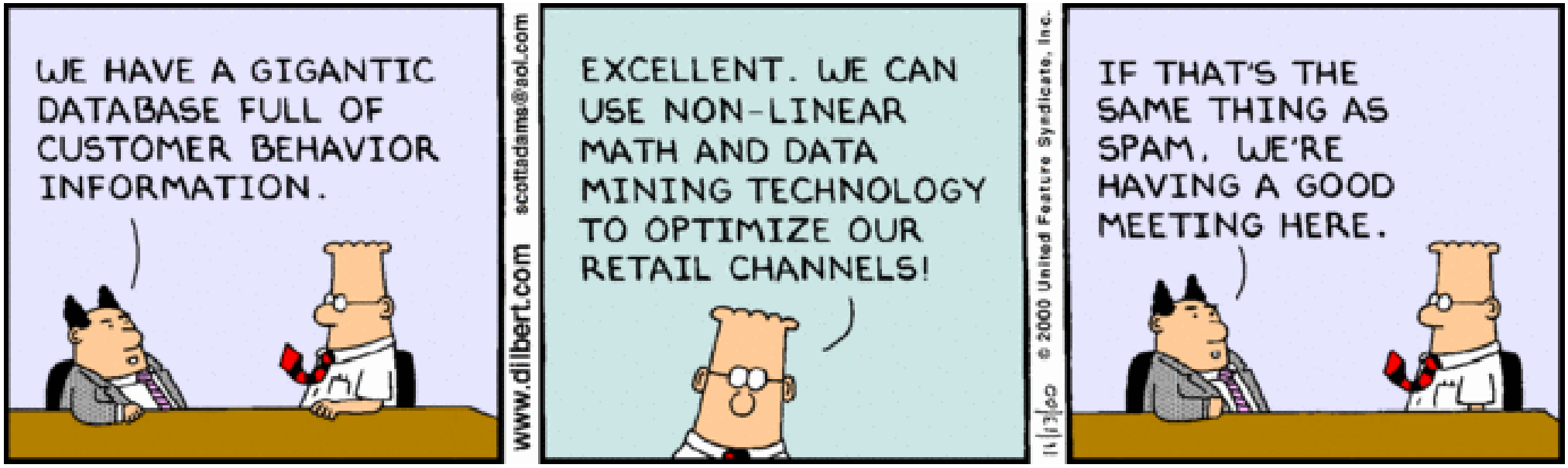
OREM 3309: Information Engineering

(Including a Short Introduction to Analytics)



Slides by Michael Hahsler

What is Information Engineering?

*"Information engineering (IE) or information engineering methodology (IEM) is a software engineering approach to designing and **developing information systems**. It can also be considered as the generation, distribution, analysis and use of information in systems."*

[Wikipedia]

*"Information Engineering is the incorporation of an **engineering approach** and discipline to the generation of information and the promotion of the **better use of information** and resources."*

[Steven A. Demurjian, CSE, UConn]

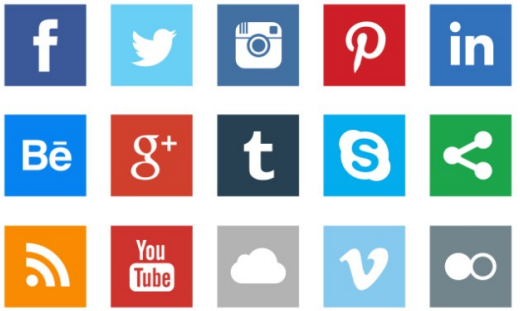
What is Analytics?

- Analytics is the discovery and communication of **meaningful patterns** in data.
- Analytics relies on the simultaneous application of **statistics, computer programming and operations research** to quantify performance.
- Analytics often favors data **visualization** to communicate insight.

Why do companies care?

Businesses collect and warehouse lots of **data**.

- Bank/credit card transactions
- Web data, e-commerce
- Social media
- Internet of things (IOT)



Computers are cheaper and more powerful.

- SaaS/IaaS/PaaS

Competition to provide better services.

- Mass customization and recommendation systems
- Targeted advertising
- Improved logistics

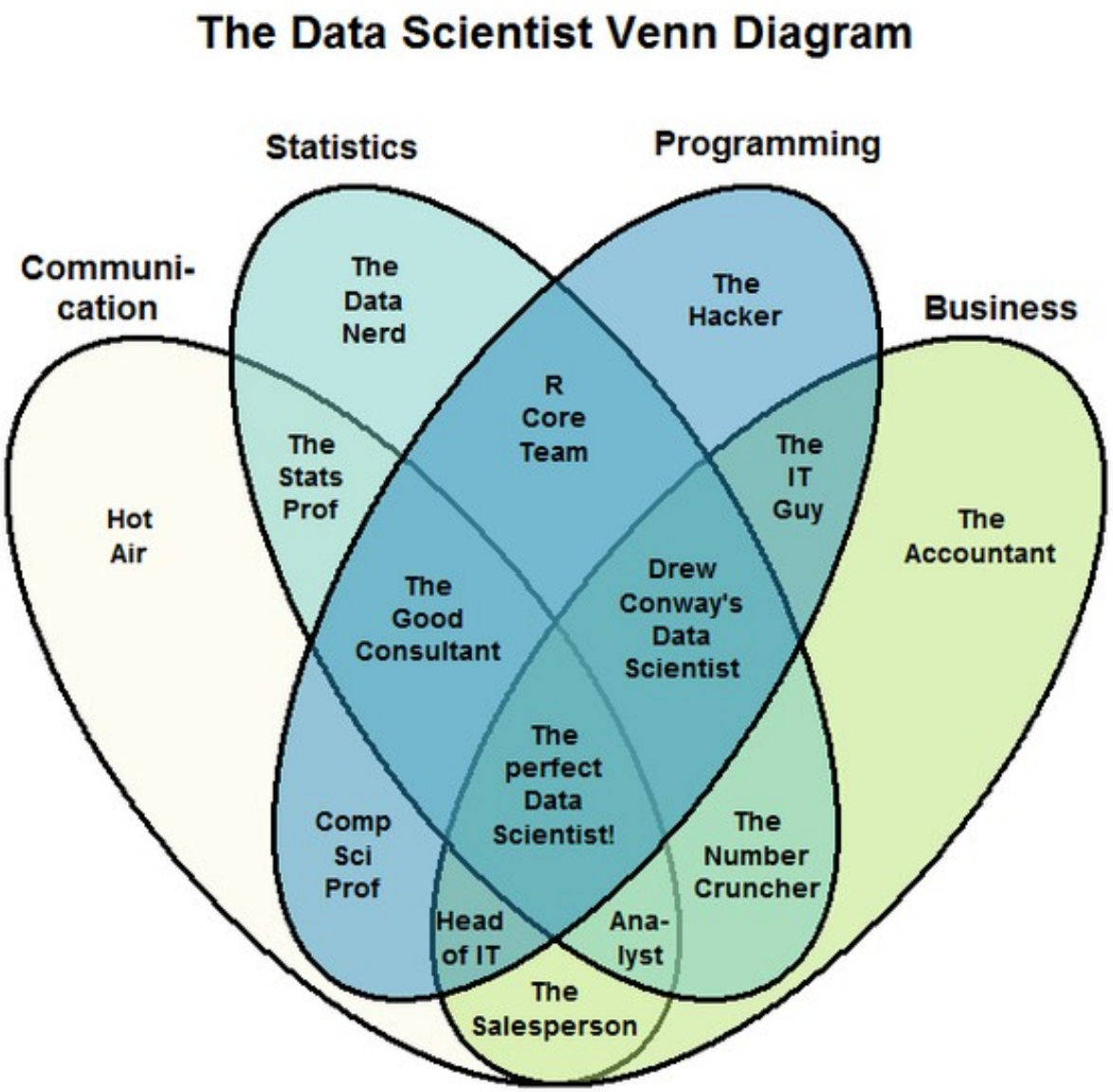
Who does all this?

And who gets the big paycheck?

Who does all this?

And who gets the big paycheck?

Of course! That weird DATA SCIENTIST living in an overpriced house in Silicon Valley!



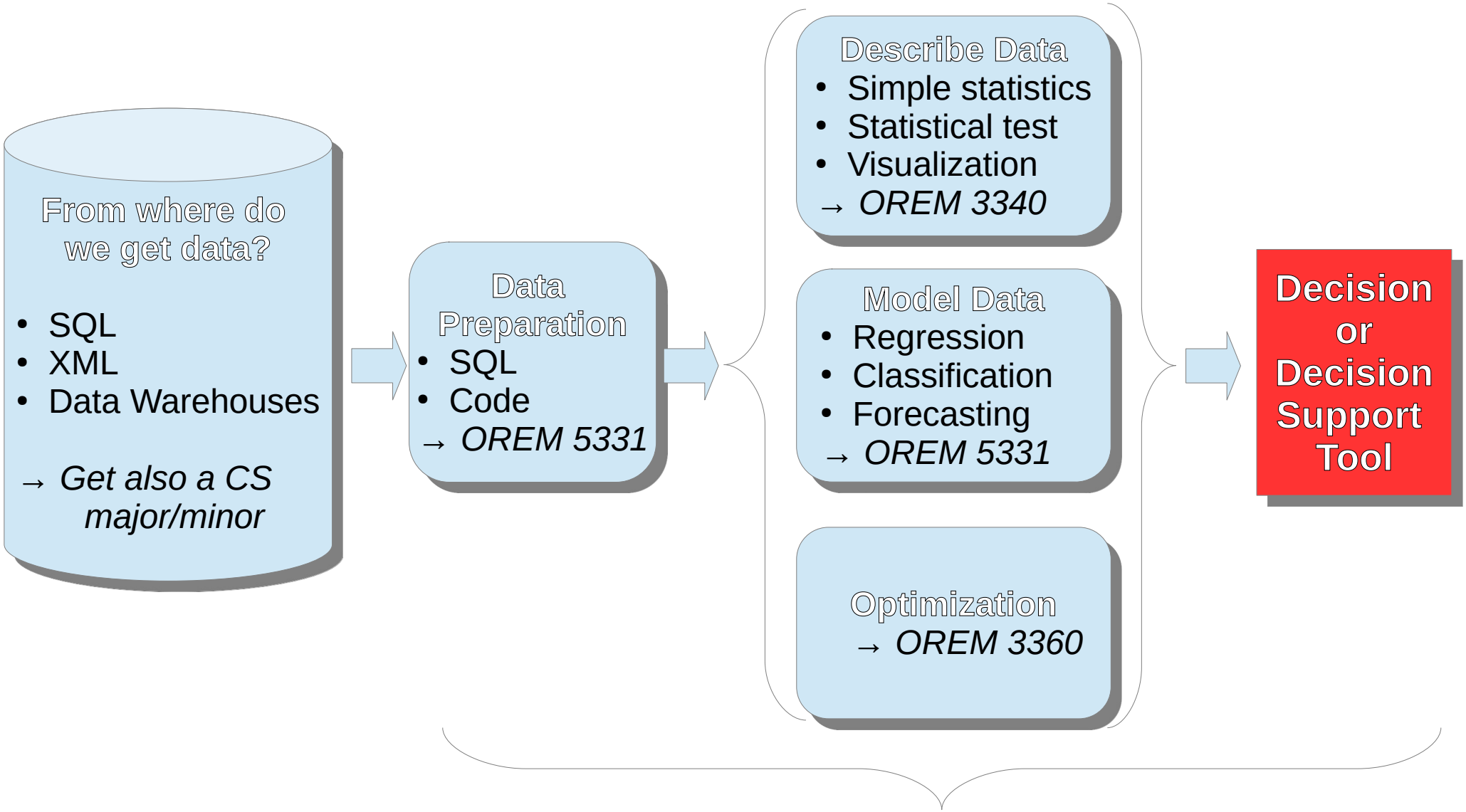
Who is a data scientist?

- The perfect data scientist from Kolassa's Venn diagram is a mythical sexy unicorn ninja rockstar who can transform a business just by thinking about its problems.
- A person who is better at statistics than any software engineer and better at software engineering than any statistician.
- Data scientist is now widely used for people working with data.

https://yanirseroussi.com/2016/08/04/is-data-scientist-a-useless-job-title/

What will we learn in this course?

And where can you learn more?



Havard Business Review, 2006

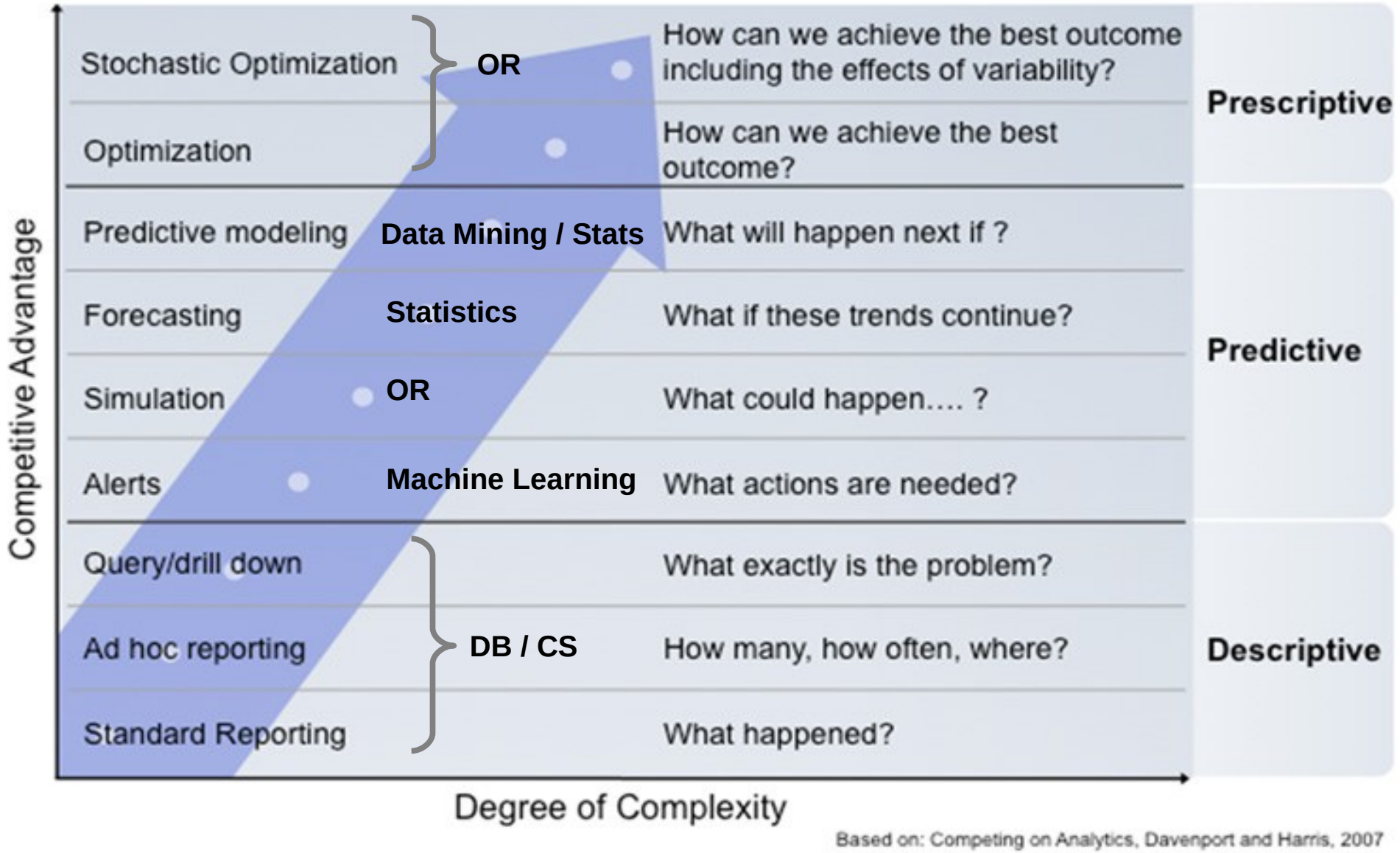
THINGS YOU CAN COUNT ON

Analytics competitors make expert use of statistics and modeling to improve a wide variety of functions. Here are some common applications:

FUNCTION	DESCRIPTION	EXEMPLARS
Supply chain	Simulate and optimize supply chain flows; reduce inventory and stock-outs.	Dell, Wal-Mart, Amazon
Customer selection, loyalty, and service	Identify customers with the greatest profit potential; increase likelihood that they will want the product or service offering; retain their loyalty.	Harrah's, Capital One, Barclays
Pricing	Identify the price that will maximize yield, or profit.	Progressive, Marriott
Human capital	Select the best employees for particular tasks or jobs, at particular compensation levels.	New England Patriots, Oakland A's, Boston Red Sox
Product and service quality	Detect quality problems early and minimize them.	Honda, Intel
Financial performance	Better understand the drivers of financial performance and the effects of nonfinancial factors.	MCI, Verizon
Research and development	Improve quality, efficacy, and, where applicable, safety of products and services.	Novartis, Amazon, Yahoo

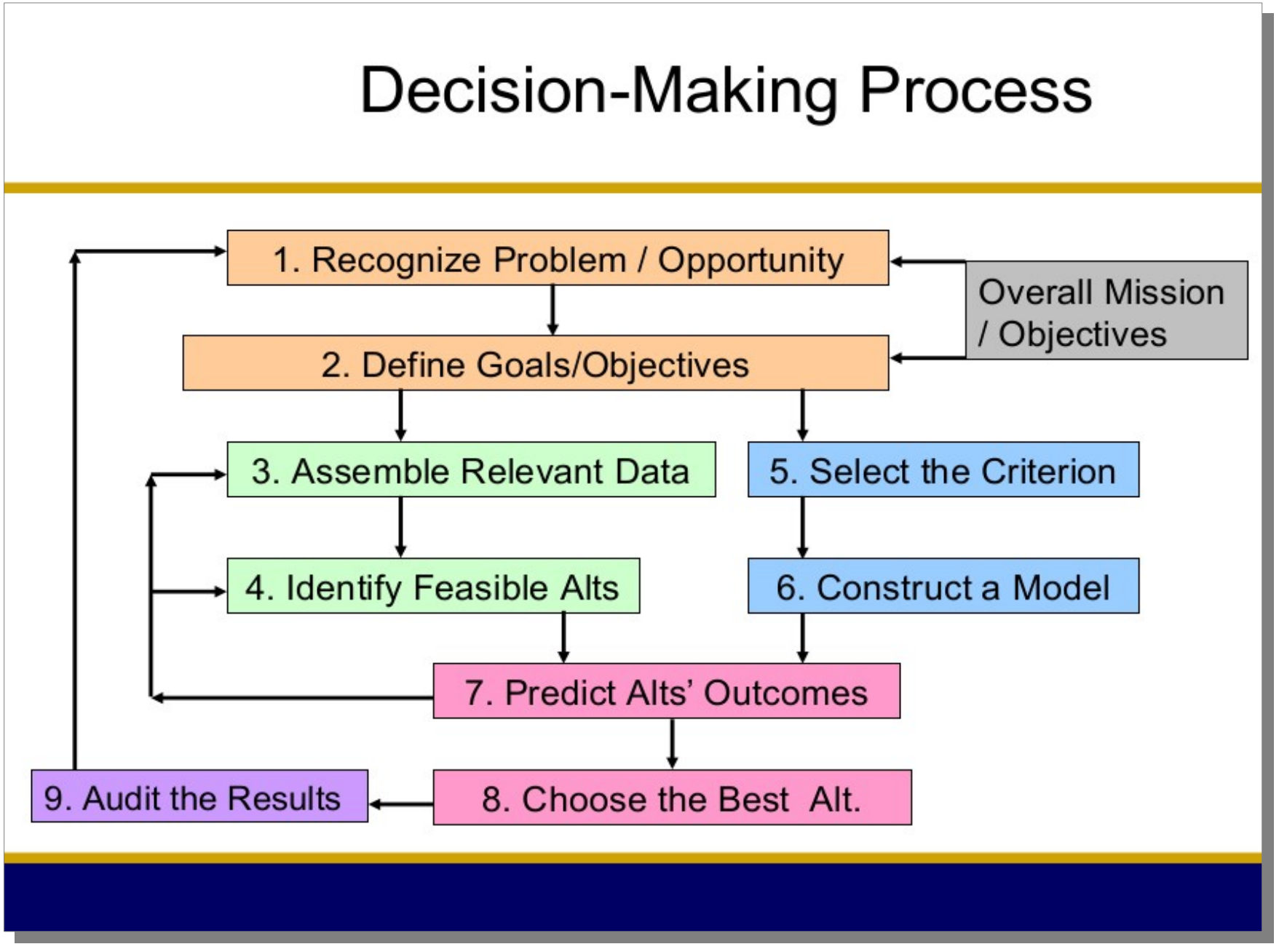
Havard Business Review, 2006

Types of Analytics



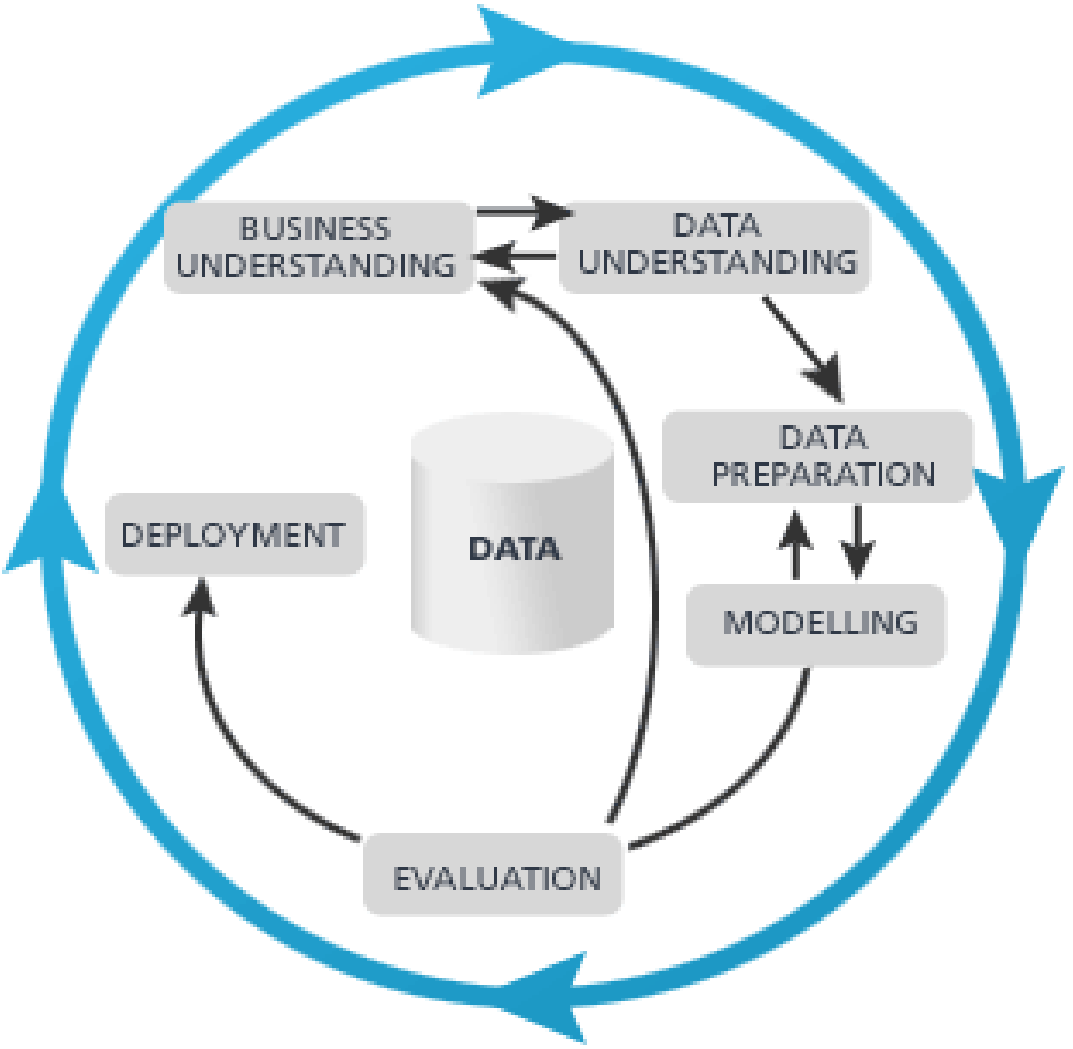
Based on: Competing on Analytics, Davenport and Harris, 2007

How to do an analytics project? Remember this from EMIS 2360?



How to do an analytics project? CRISP-DM Reference Model

- **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- De facto standard for conducting data mining and knowledge discovery projects.
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own process.



Tasks in the CRISP-DM Model

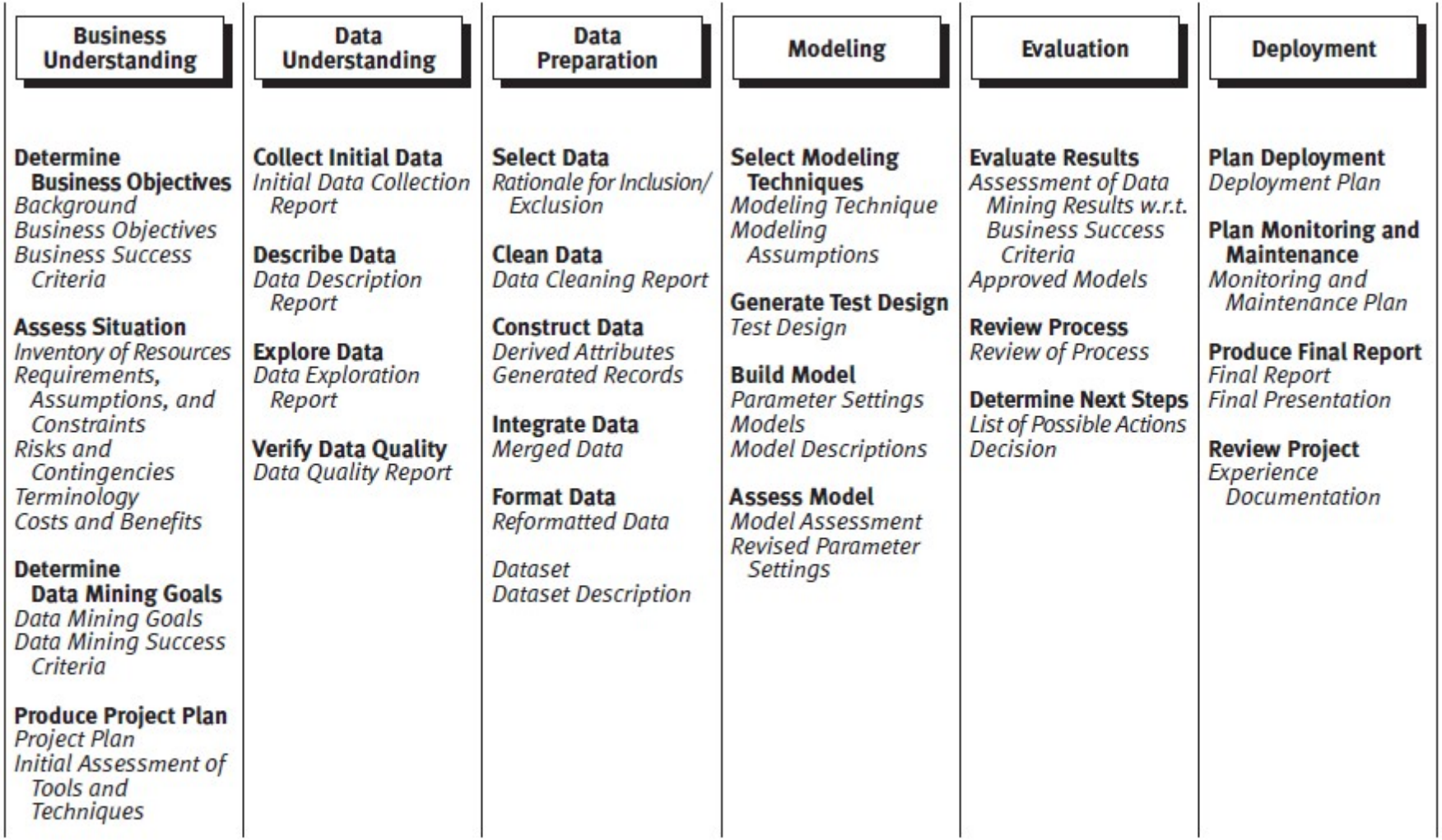


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

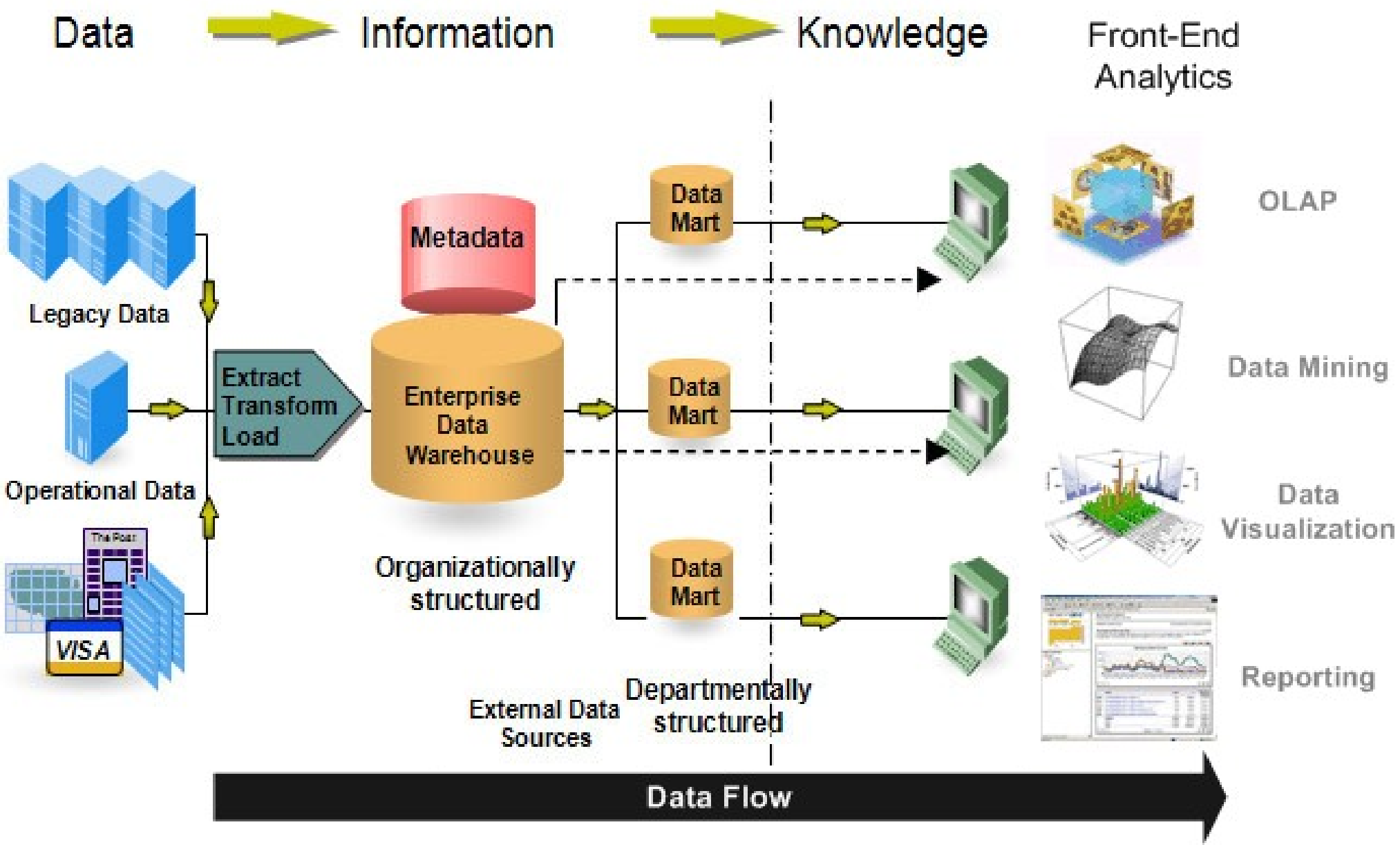
Example: How is POS data stored?

- Relational data base?
- How do the tables look like?
  → On Line Transaction Processing
- Has every store/region its own data base?

- What if I want to know how many units of product A were sold in the last three month in Texas?

- There must be an easier way!

Data Warehouse



Data Warehouse
ELT: Extract, Transform and Load

- **Extracting** data from outside sources
- **Transforming** it to fit analytical needs. E.g.,
  - Clean (missing data, wrong data)
  - Translate (1 → "female")
  - Join (from several sources)
  - Calculate and aggregate data
- **Loading** it into the end target (data warehouse)

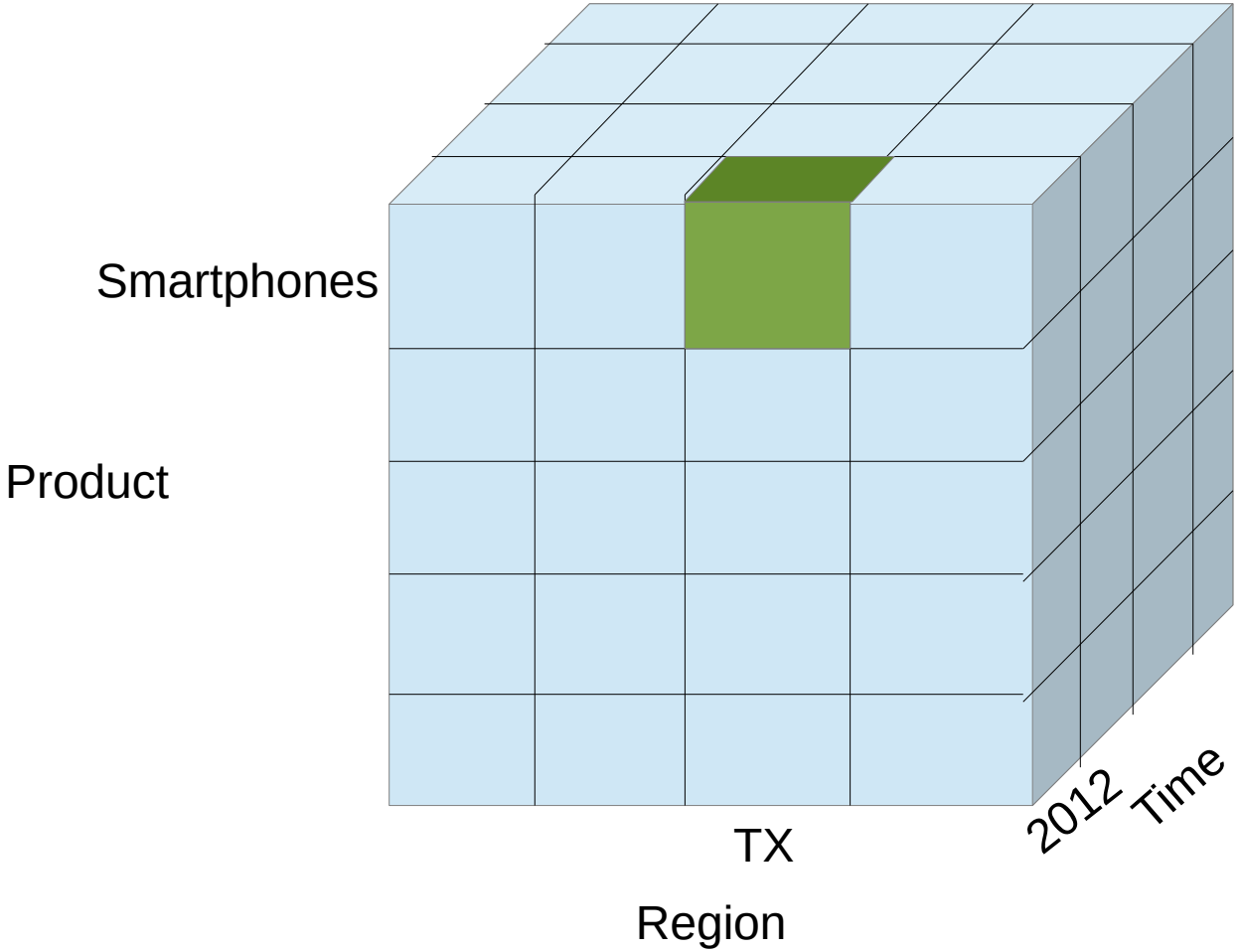
Data Warehouse

Properties

- **Subject Oriented:** Data warehouses are designed to help you analyze data in a certain area (e.g., sales).
- **Integrated:** Integrates data from disparate sources into a consistent format.
- **Nonvolatile**: Data in the data warehouse are never overwritten or deleted.
- **Time Variant**: they maintain both historical and (nearly) current data.

OnLine Analytical Processing (OLAP)

- Stores data in "data cubes" for fast OLAP operations.
- Requires a special database structure (Snow-flake scheme)



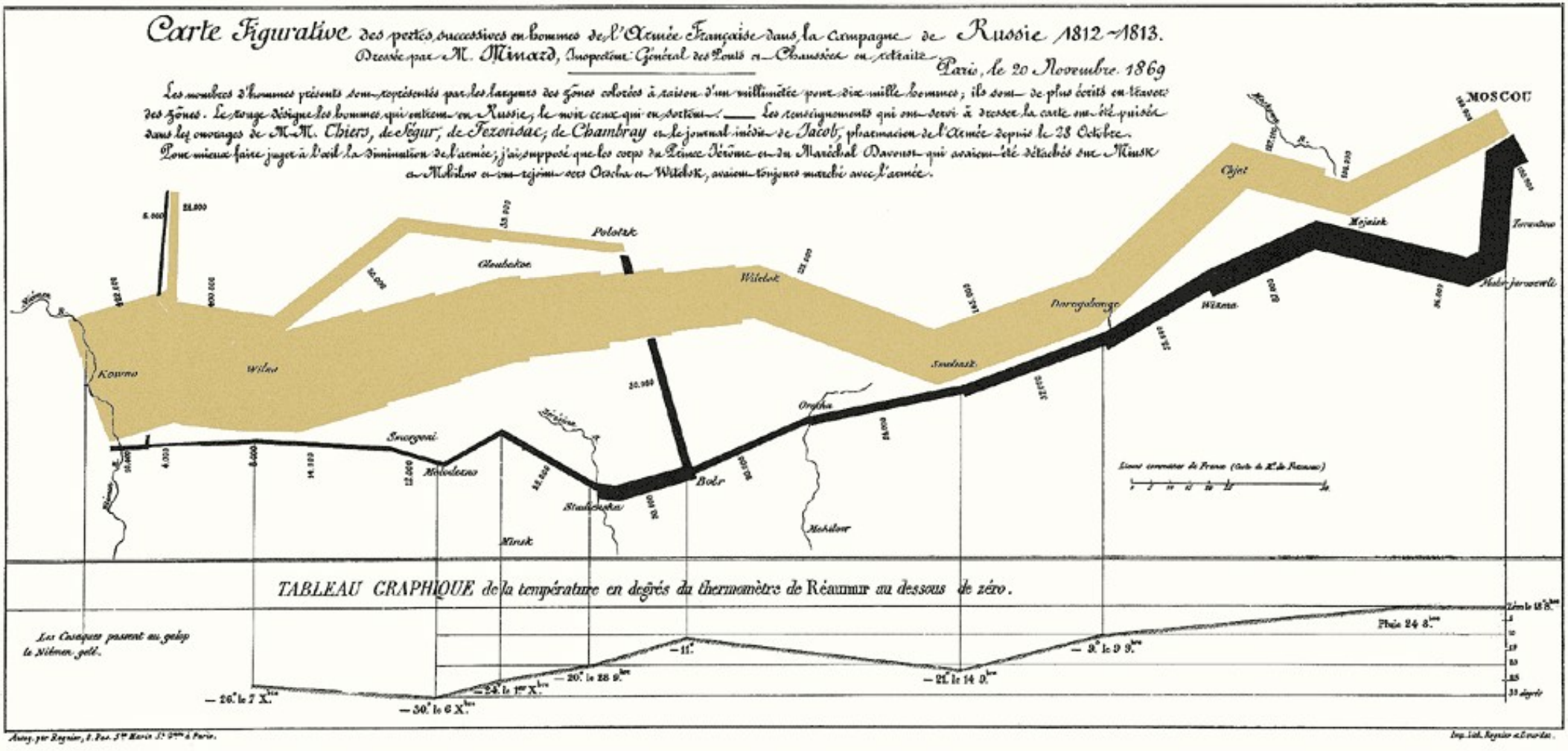
Operations:

- Slice
- Dice
- Drill-down
- Roll-up
- Pivot

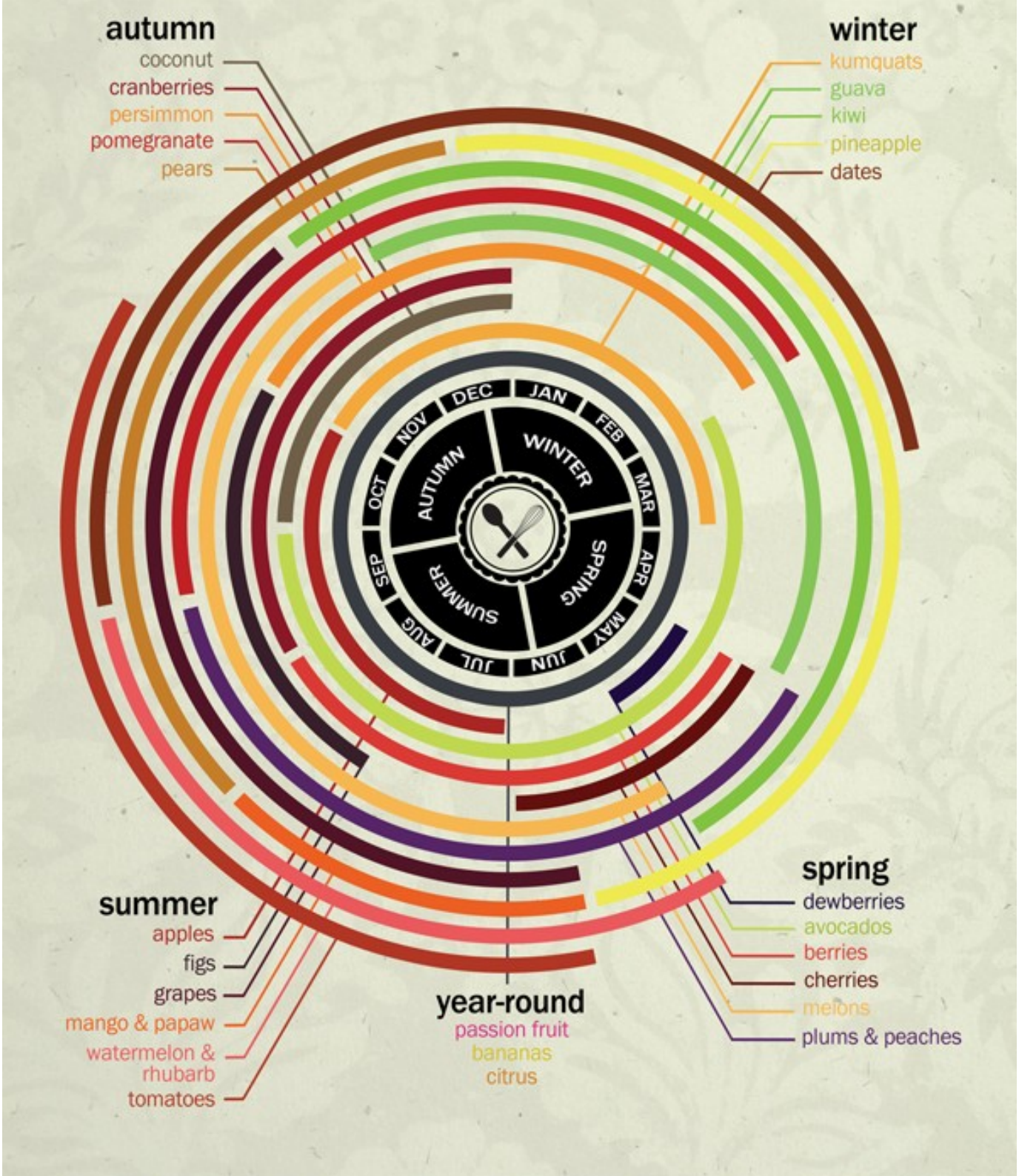
→ Similar to Pivot Tables

Data Visualization

- Infoviz is a field by its own.
- Napoleon's Army in Russia by Charles Minard (around 1850)

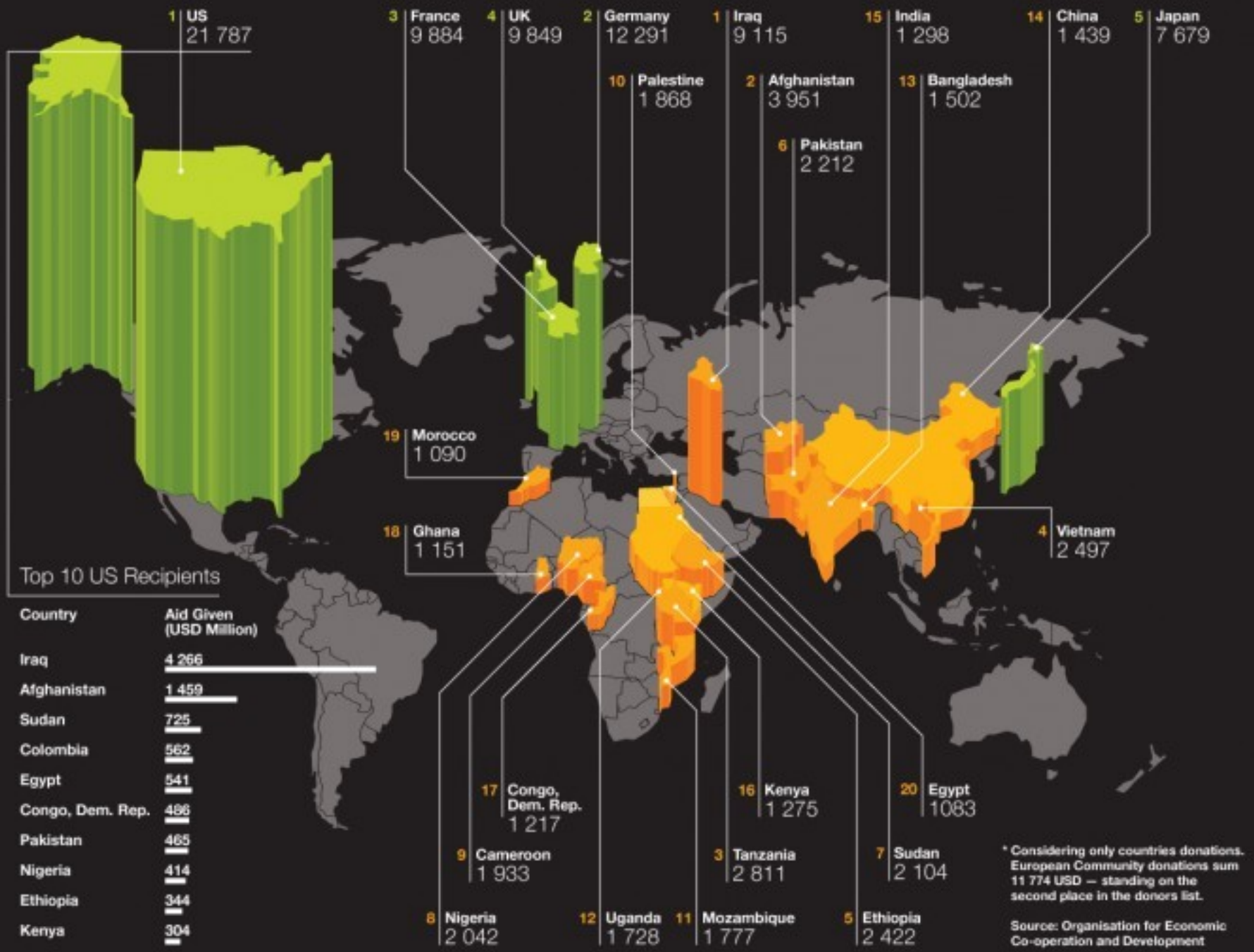


Eat **fruits** when they are in **season!!!**



Developmental Aid Flows around the World

■ Top 20 recipients ■ Top 5 donors* • Values in USD million



Top 10 US Recipients

Country	Aid Given (USD Million)
Iraq	4 266
Afghanistan	1 459
Sudan	725
Colombia	562
Egypt	541
Congo, Dem. Rep.	486
Pakistan	465
Nigeria	414
Ethiopia	344
Kenya	304

* Considering only countries donations. European Community donations sum 11 774 USD — standing on the second place in the donors list.

Source: Organisation for Economic Co-operation and Development

Do you notice the slight flaw?

Legal, Privacy and Security Issues

Legal, Privacy and Security Issues

1)Are we allowed to collect the data?

2)Are we allowed to use the data?

3)Is privacy preserved in the process?

4)Is it ethical to use and act on the data?

Problem

Internet is global but legislation is local!

Legal, Privacy and Security Issues

The New York Times

Data-Gathering via Apps Presents a Gray Legal Area

By KEVIN J. O'BRIEN

Published: October 28, 2012

BERLIN — Angry Birds, the top-selling paid mobile app for the iPhone in the United States and Europe, has been downloaded more than a billion times by devoted game players around the world, who often spend hours slinging squawking fowl at groups of egg-stealing pigs.

When Jason Hong, an associate professor at the Human-Computer Interaction Institute at Carnegie Mellon University, surveyed 40 users, all but two were *unaware that the game was storing their locations so that they could later be the targets of ads....*


Pokémon
GO

Here is what the small print says…

USA Today Network Josh Hafner, USA TODAY 2:38 p.m. EDT July 13, 2016

Pokémon Go's constant location tracking and camera access required for gameplay, paired with its skyrocketing popularity, could provide data like no app before it.

***"Their privacy policy is vague,"** Hong said. "I'd say deliberately vague, because of the lack of clarity on the business model."*

...

*The agreement says Pokémon Go collects data about its users as a **"business asset."** This includes data used to personally identify players such as email addresses and other information pulled from Google and Facebook accounts players use to sign up for the game.*

If Niantic is ever sold, the agreement states, all that data can go to another company.