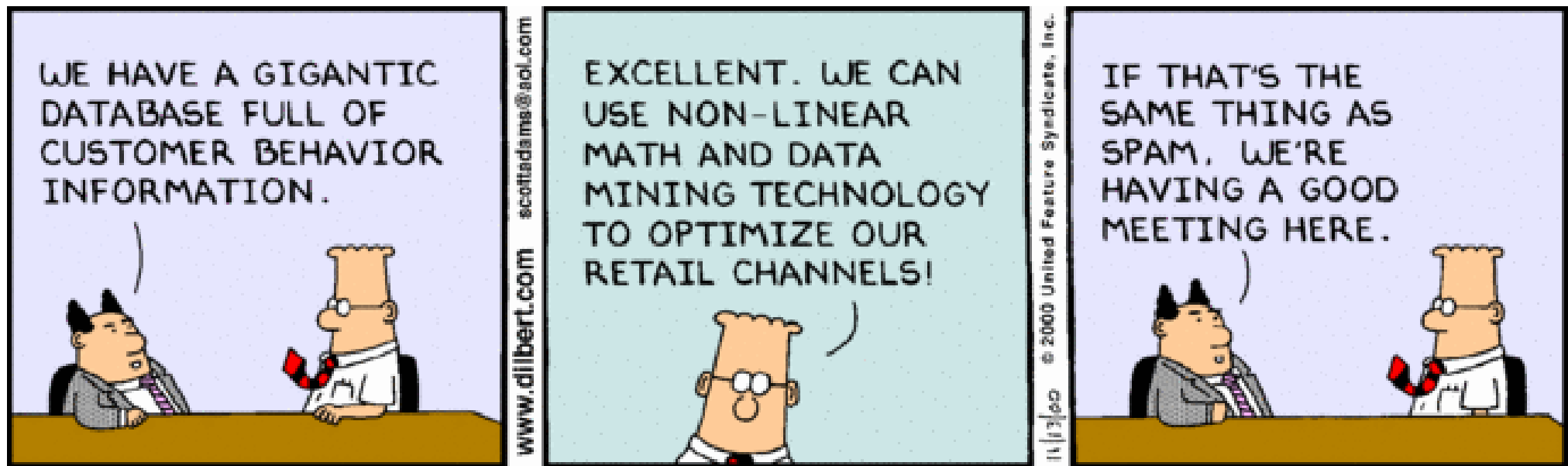


EMIS/DS 1300: A Practical Introduction to Data Science



Slides by Michael Hahsler

Agenda



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools

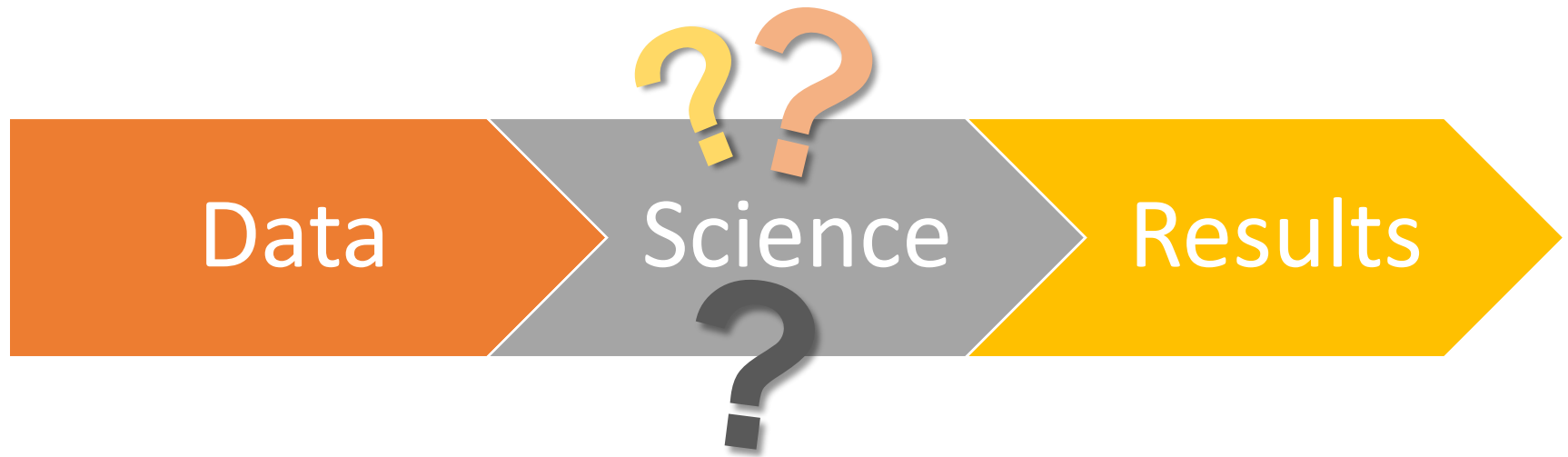


Visualization



Ethics, Privacy and Security Issues

Data + Science = Results?

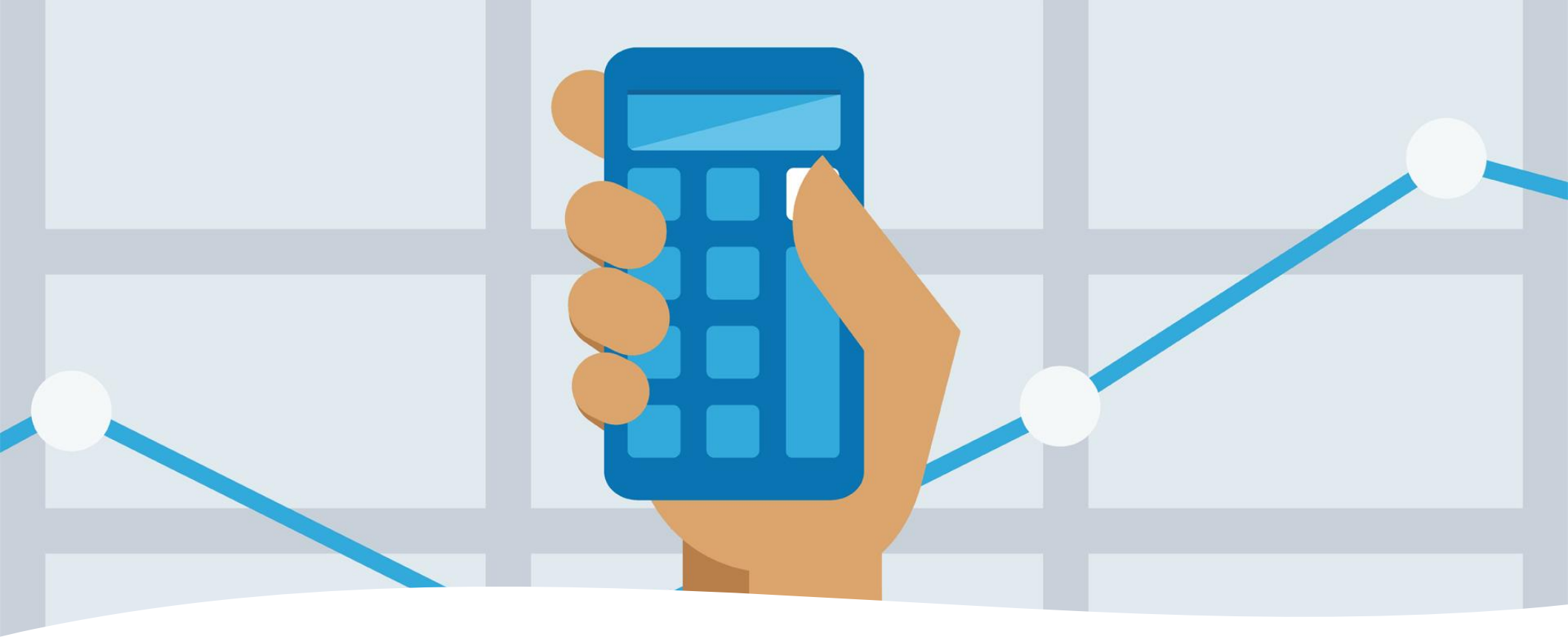




What is Data Science?

*“Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to **understand** and analyze **actual phenomena with data.**”*

[Hayashi, Chikio "What is Data Science?"]



What is Statistics?

- “Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.”

[Wikipedia]

- Techniques:
 - Design of experiments (sampling)
 - Descriptive statistics
 - Statistical inference (estimation, testing)

What are Analytics and Data Mining?

- Analytics and Data Mining focus on the discovery and the communication of **meaningful patterns** in data.
- Analytics relies on the simultaneous application of **statistics, computer programming and optimization** to quantify performance.
- Analytics often favors data **visualization** to communicate insight.
- Data Mining focuses on **predictive models**.

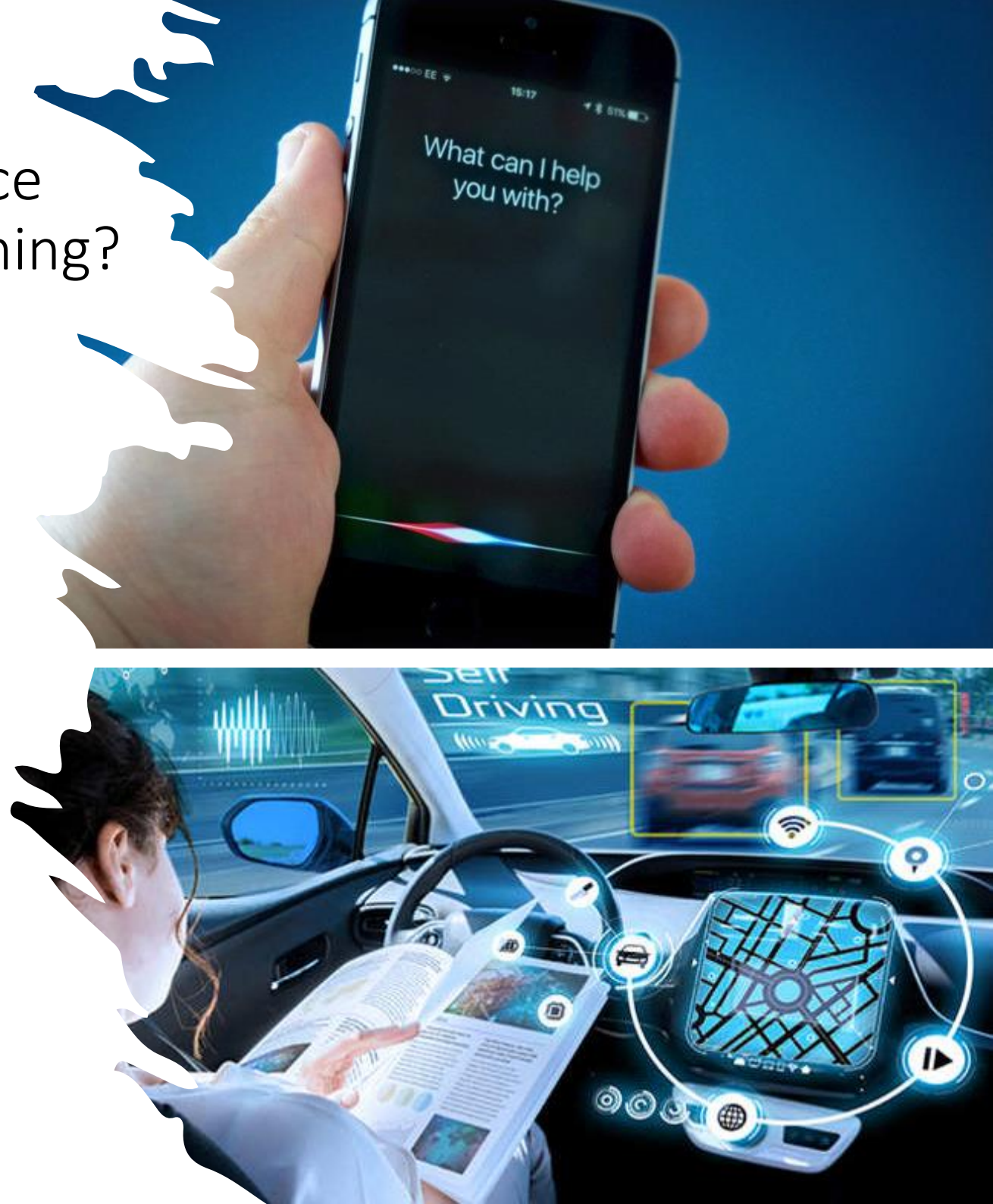
[Wikipedia]



What are Artificial Intelligence and Machine Learning?

- **AI** is the study of **intelligent agents**, devices that perceives its environment and takes actions that maximize its chance of successfully **achieving its goals**.
- **Machine learning (ML)** is the study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task. The goal is to **make accurate predictions** or decisions without being explicitly programmed to perform the task.

[Wikipedia]



Why do companies care about Data

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

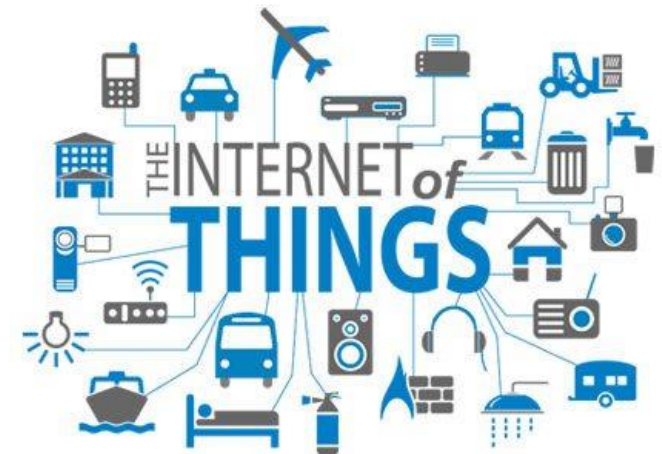
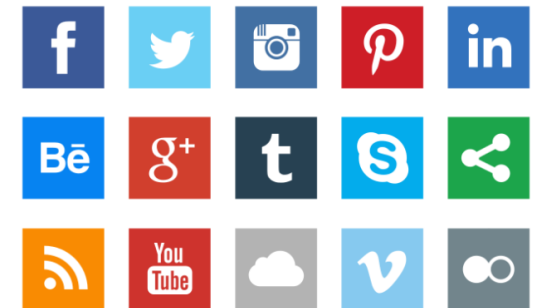
- Bank/credit card transactions
- Web data, e-commerce
- Social media
- Internet of things (IOT)

Computers are cheaper and more powerful.

- SaaS/IaaS/PaaS

Competition to provide better services.

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



Assignment:
Why should
you care?

Answer the
following
questions.



Examples of information
that is collected about you
in your daily life.



Who do you think collects
the information?



Who do you think has
access to the information?



How may this information
be used/misused?

Some Applications of Data Science

- Uber
- Airbnb
- Netflix
- Amazon

- Logistics
- Banking, loans, insurance
- Pharmaceutical industry
- Healthcare
- Sports



Agenda



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools



Visualization

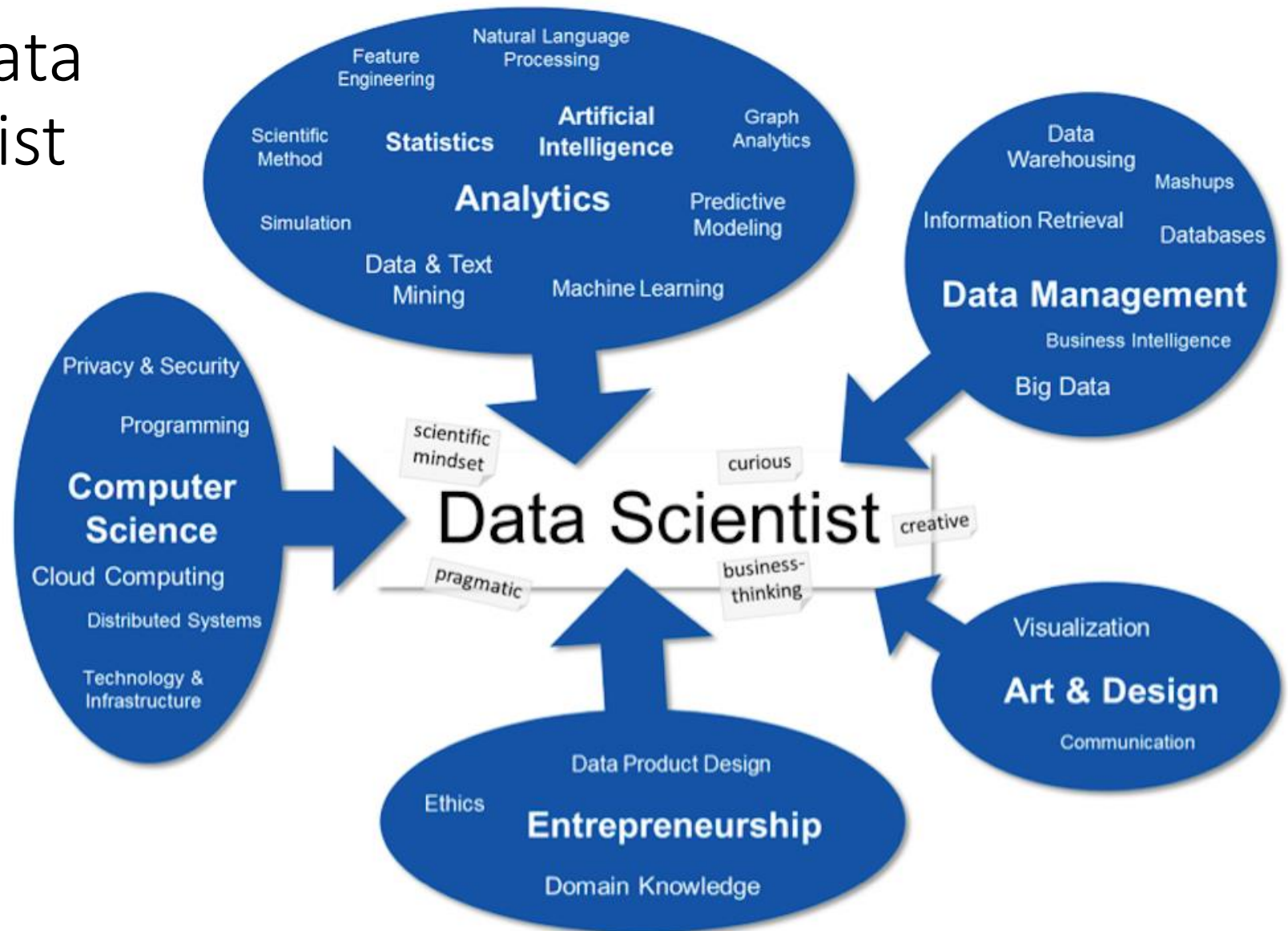


Ethics, Privacy and Security Issues

Who does all this?
And who gets the big paycheck?



The Data Scientist



Source: T. Stadelmann, et al., Applied Data Science in Europe

**Good luck finding this person!
Probably a team effort!**

What Does a Data Scientist Do?



Identifying data analytics **opportunities**.



Find/collect the correct **data** sets and variables.



Clean the data and ensure accuracy and completeness.



Decide on appropriate **models** and algorithms to mine the data. Identify patterns and trends.



Interpret the results to data to discover solutions and opportunities.



Communicate findings to stakeholders using visualization and prototypes.

Agenda



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools



Visualization

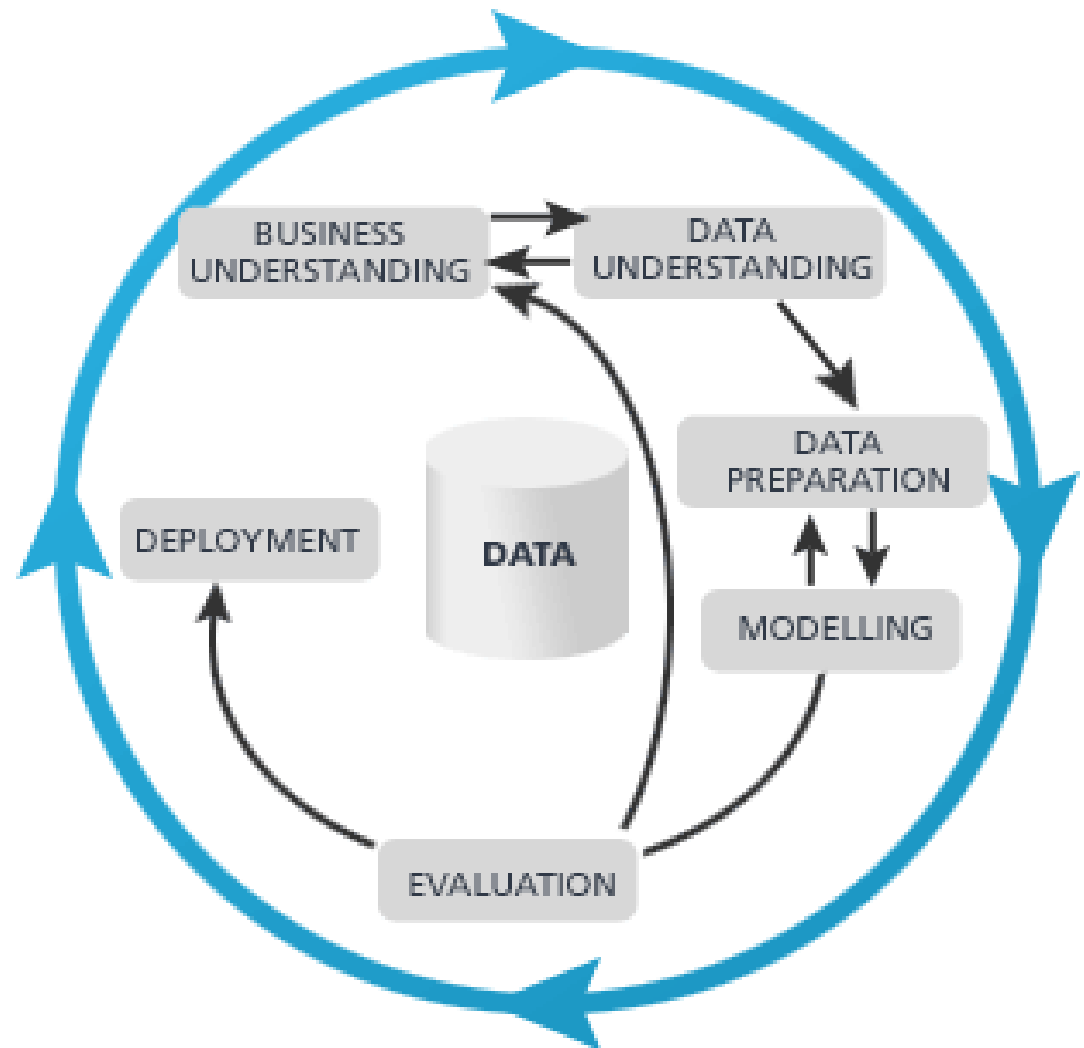


Ethics, Privacy and Security Issues

How to do a Data Science project?

CRISP-DM Reference Model

- **Cross Industry Standard Process for Data Mining**
- 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 Analytic (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own process.

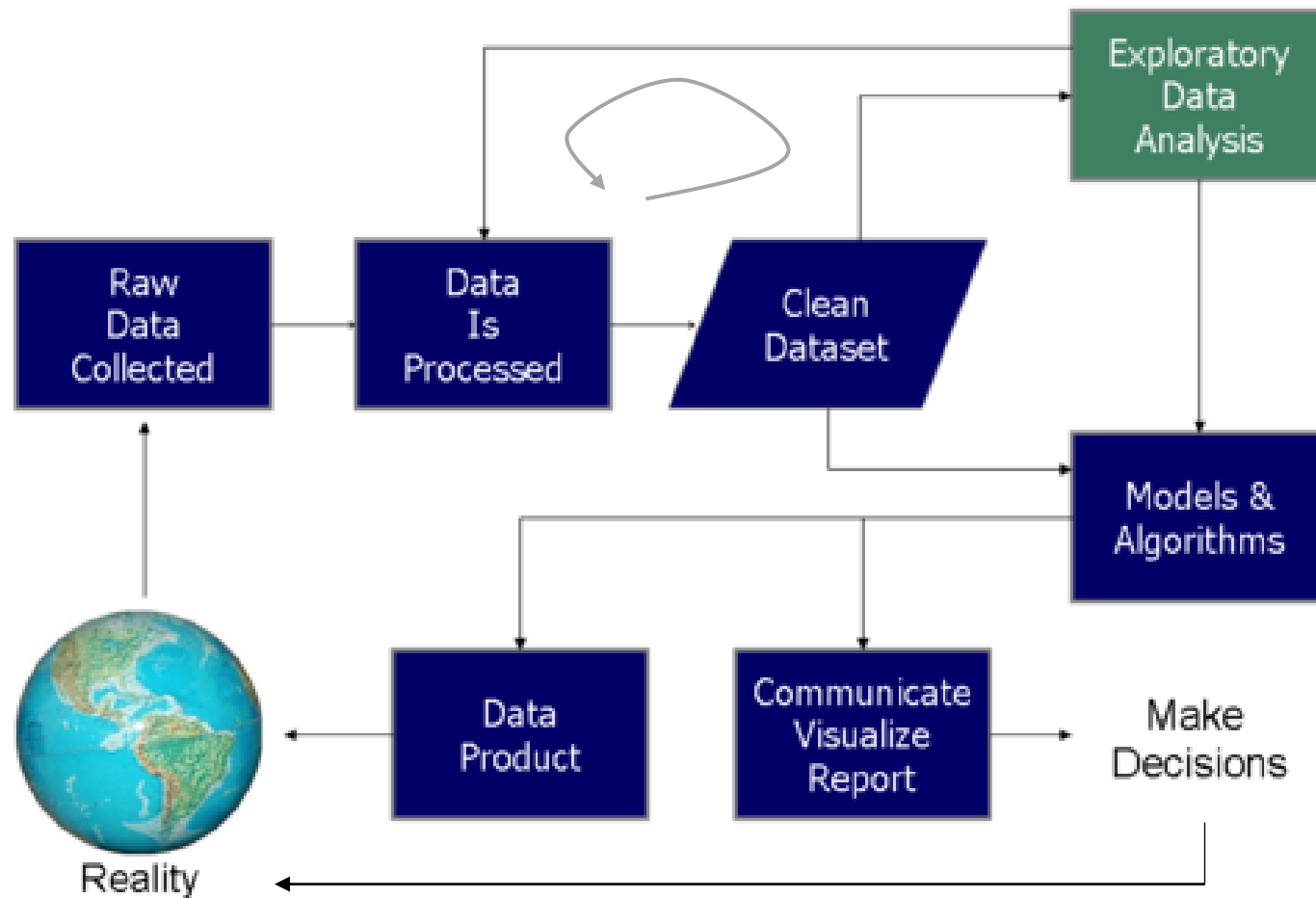


Tasks in the CRISP-DM Model

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i></p> <p>Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i></p> <p>Determine Data Mining Goals <i>Data Mining Goals Data Mining Success Criteria</i></p> <p>Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i></p>	<p>Collect Initial Data <i>Initial Data Collection Report</i></p> <p>Describe Data <i>Data Description Report</i></p> <p>Explore Data <i>Data Exploration Report</i></p> <p>Verify Data Quality <i>Data Quality Report</i></p>	<p>Select Data <i>Rationale for Inclusion/ Exclusion</i></p> <p>Clean Data <i>Data Cleaning Report</i></p> <p>Construct Data <i>Derived Attributes Generated Records</i></p> <p>Integrate Data <i>Merged Data</i></p> <p>Format Data <i>Reformatted Data</i></p> <p><i>Dataset Dataset Description</i></p>	<p>Select Modeling Techniques <i>Modeling Technique Modeling Assumptions</i></p> <p>Generate Test Design <i>Test Design</i></p> <p>Build Model <i>Parameter Settings Models Model Descriptions</i></p> <p>Assess Model <i>Model Assessment Revised Parameter Settings</i></p>	<p>Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i></p> <p>Review Process <i>Review of Process</i></p> <p>Determine Next Steps <i>List of Possible Actions Decision</i></p>	<p>Plan Deployment <i>Deployment Plan</i></p> <p>Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i></p> <p>Produce Final Report <i>Final Report Final Presentation</i></p> <p>Review Project <i>Experience Documentation</i></p>

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

The Data Science Process



Source: *The Data Science Process*, Springboard
<https://www.kdnuggets.com/2016/03/data-science-process.html>

Agenda



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools

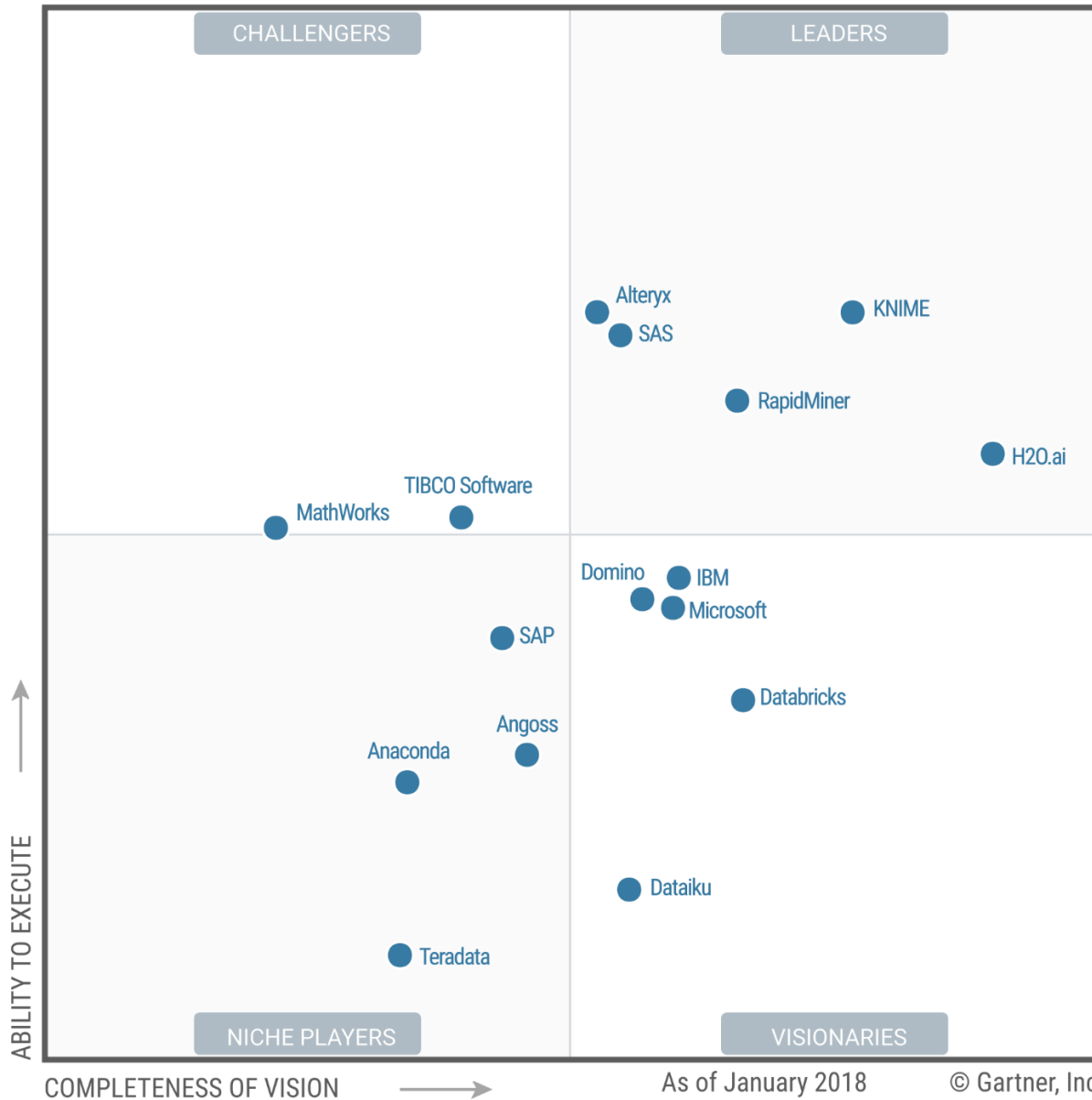


Visualization



Ethics, Privacy and Security Issues

Tools



Gartner

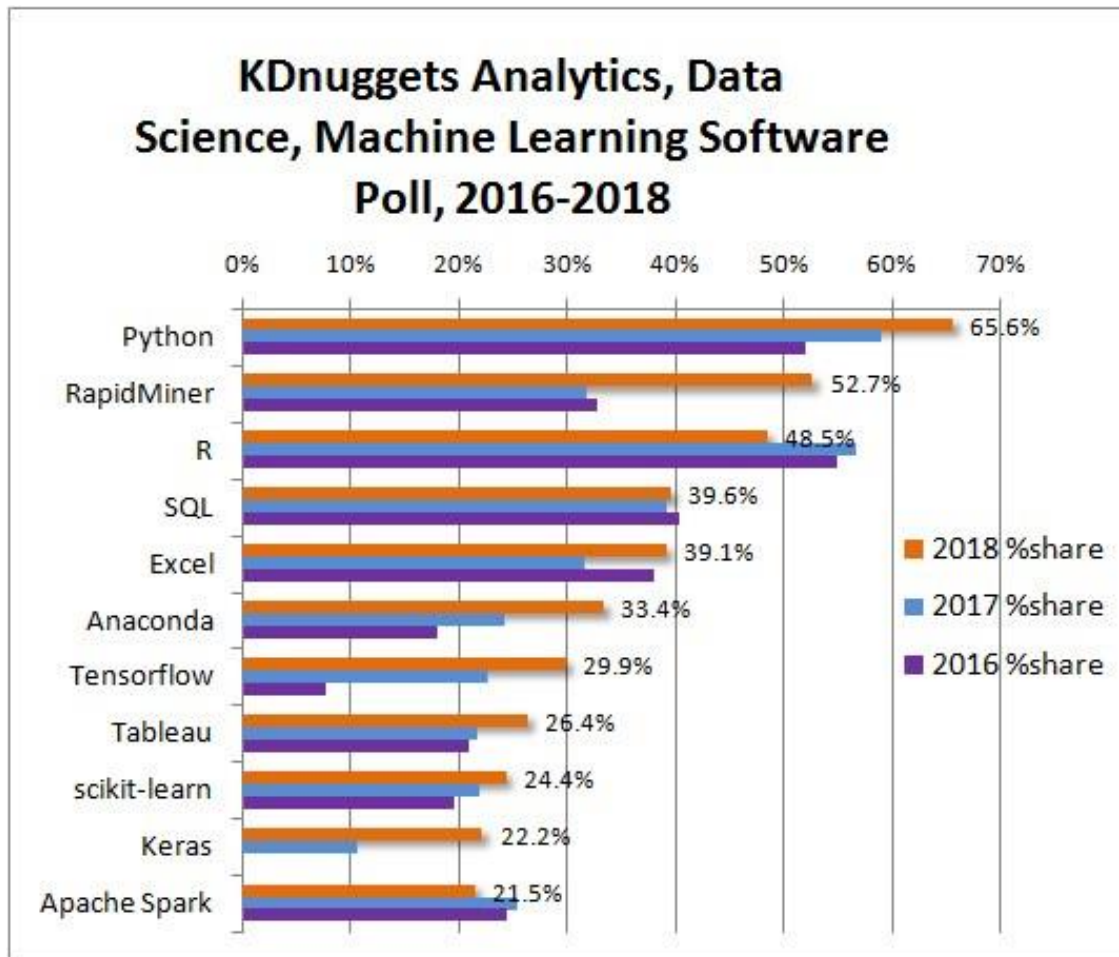
2018 Magic Quadrant for
Data Science and Machine
Learning Platforms

COMPLETENESS OF VISION

As of January 2018

© Gartner, Inc

Tools - Popularity



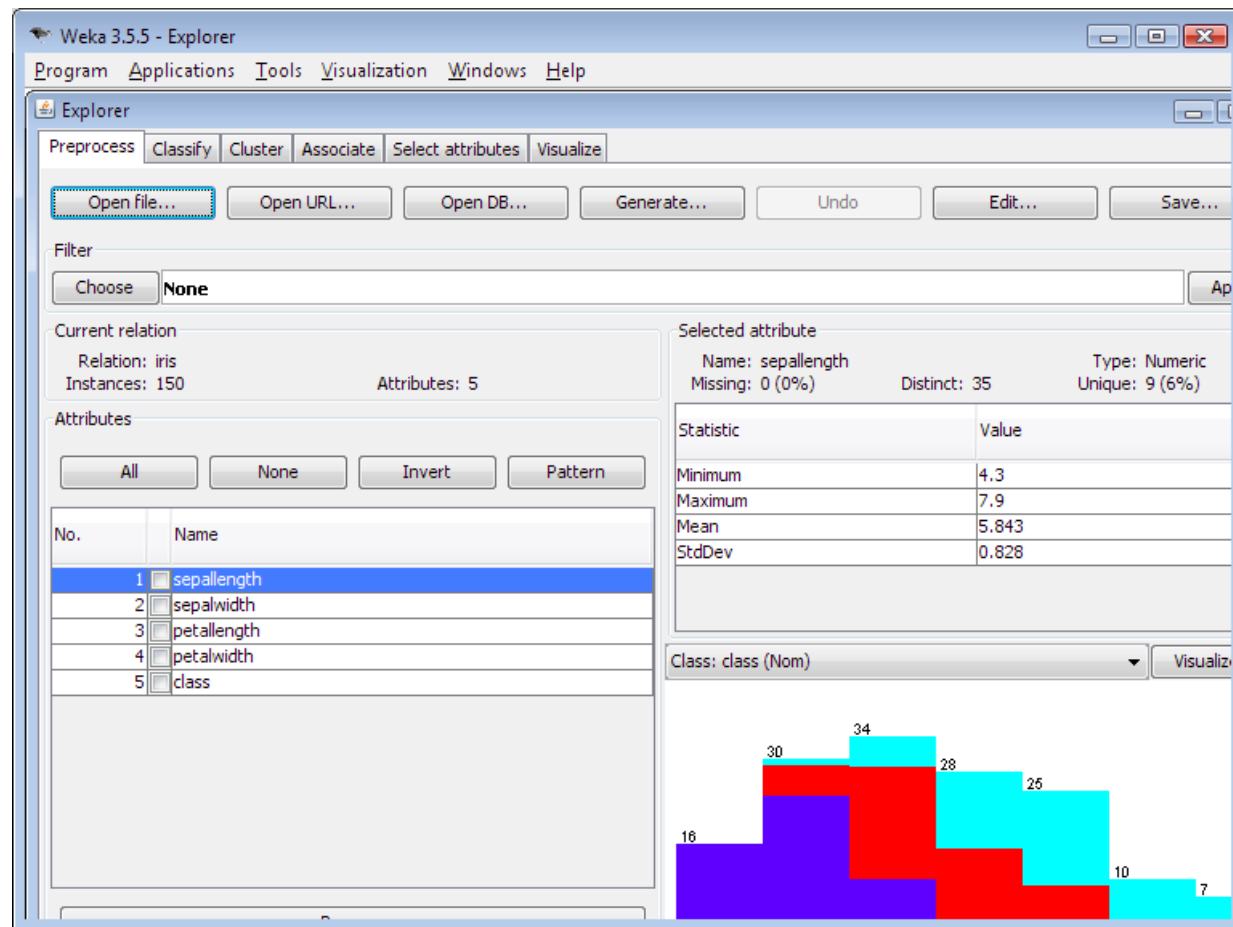
<https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html>

Tools - Types

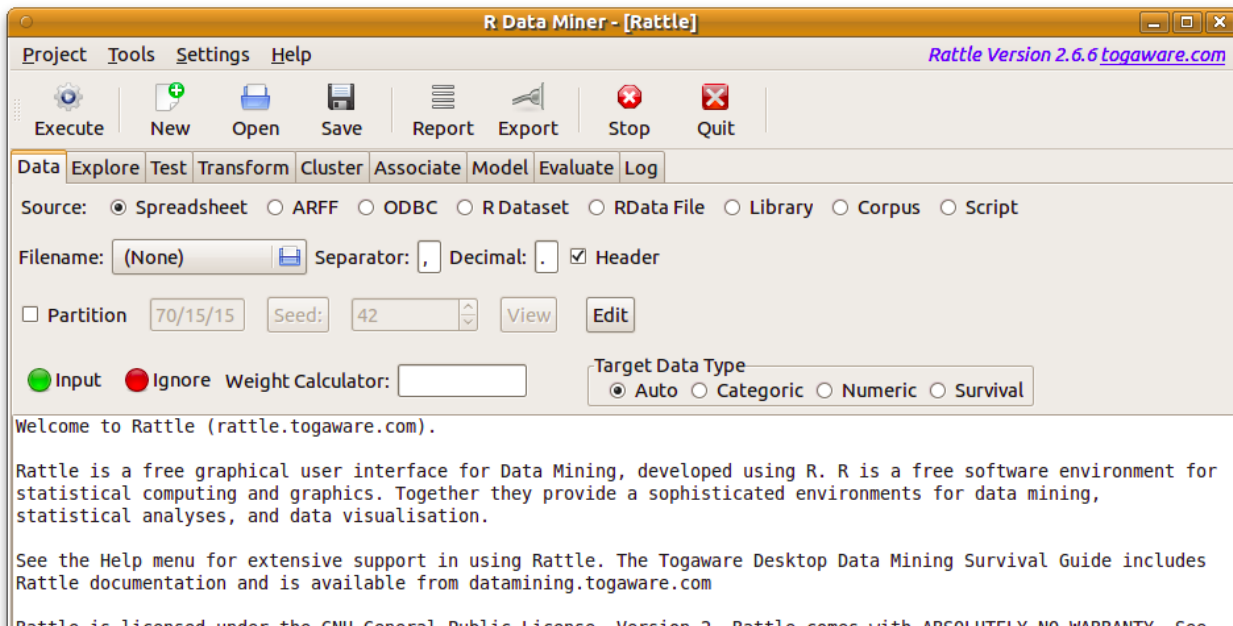
- **Data:** Relational databases (**SQLite**), NoSQL databases
- **Spreadsheet:** **Excel**, Google Sheets
- **Visualization:** Tableau, Microsoft Power BI, SAS jmp
- **Data Science Platforms**
 - Simple graphical user interface
 - Process oriented
 - Programming oriented

Tools Simple GUI

- **Weka:** Waikato Environment for Knowledge Analysis (Java API)

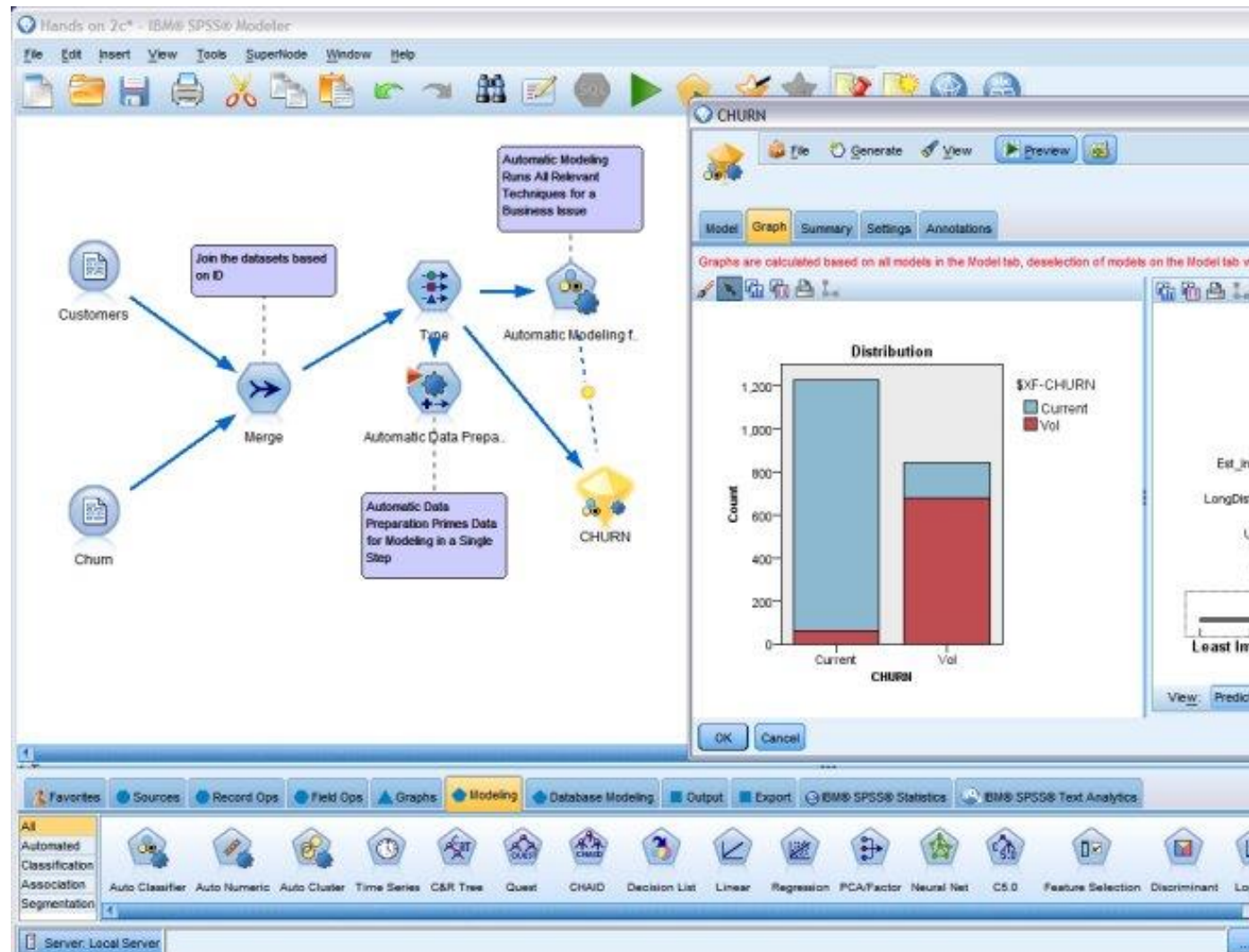


- **Rattle:** GUI for Data Mining using R



Tools -Process oriented

- SAS Enterprise Miner
- IBM SPSS Modeler
- **RapidMiner**
- Knime
- Orange



Tools -Programming oriented

- Python

- Scikit-learn, pandas
- IPython, notebooks



- R

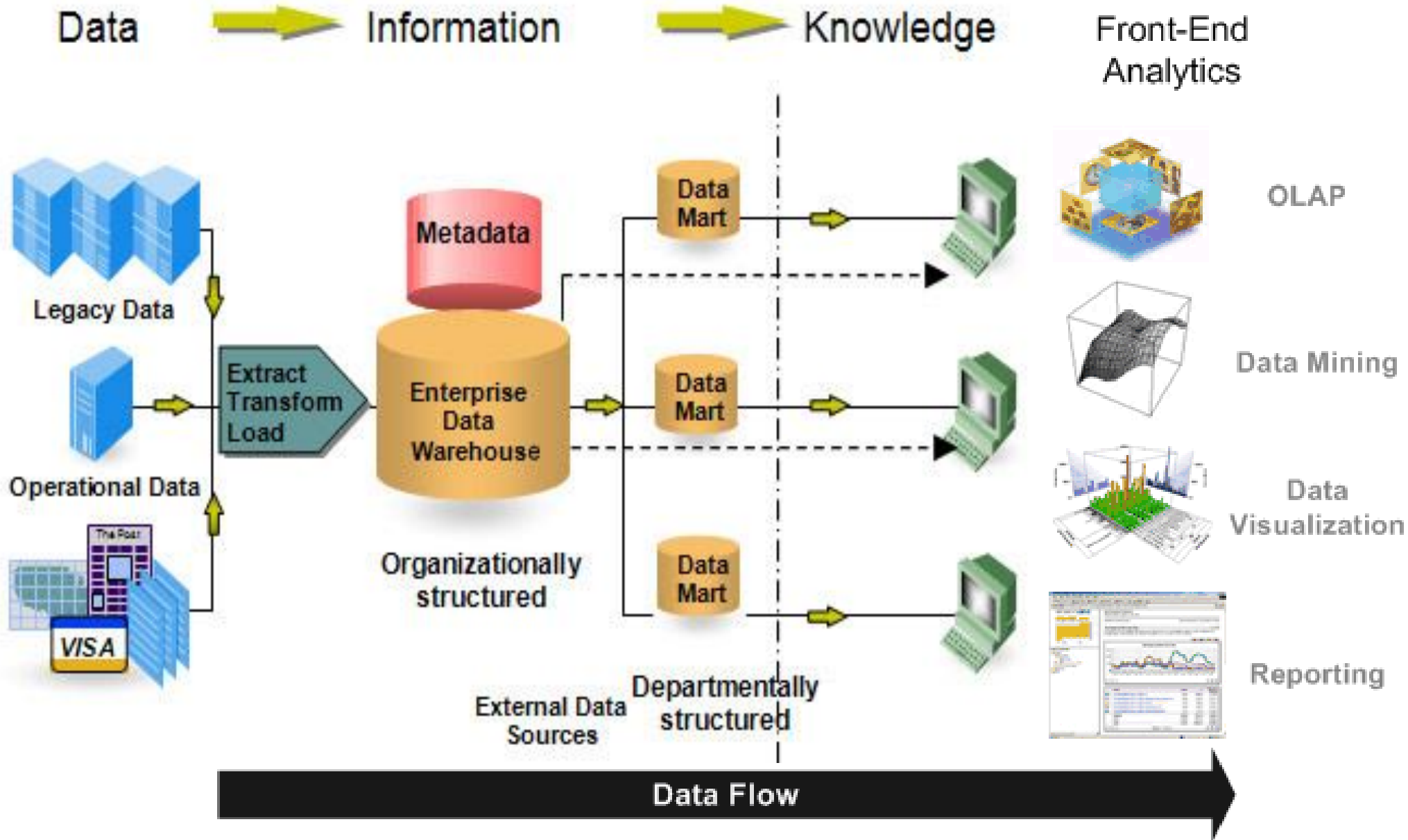
- Rattle for beginners
- RStudio IDE, markdown, shiny
- Microsoft Open R



- → Both have similar capabilities. Slightly different focus:

- R: Statistical computing and visualization
- Python: Machine learning and big data

Data Warehouse



Agenda



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools



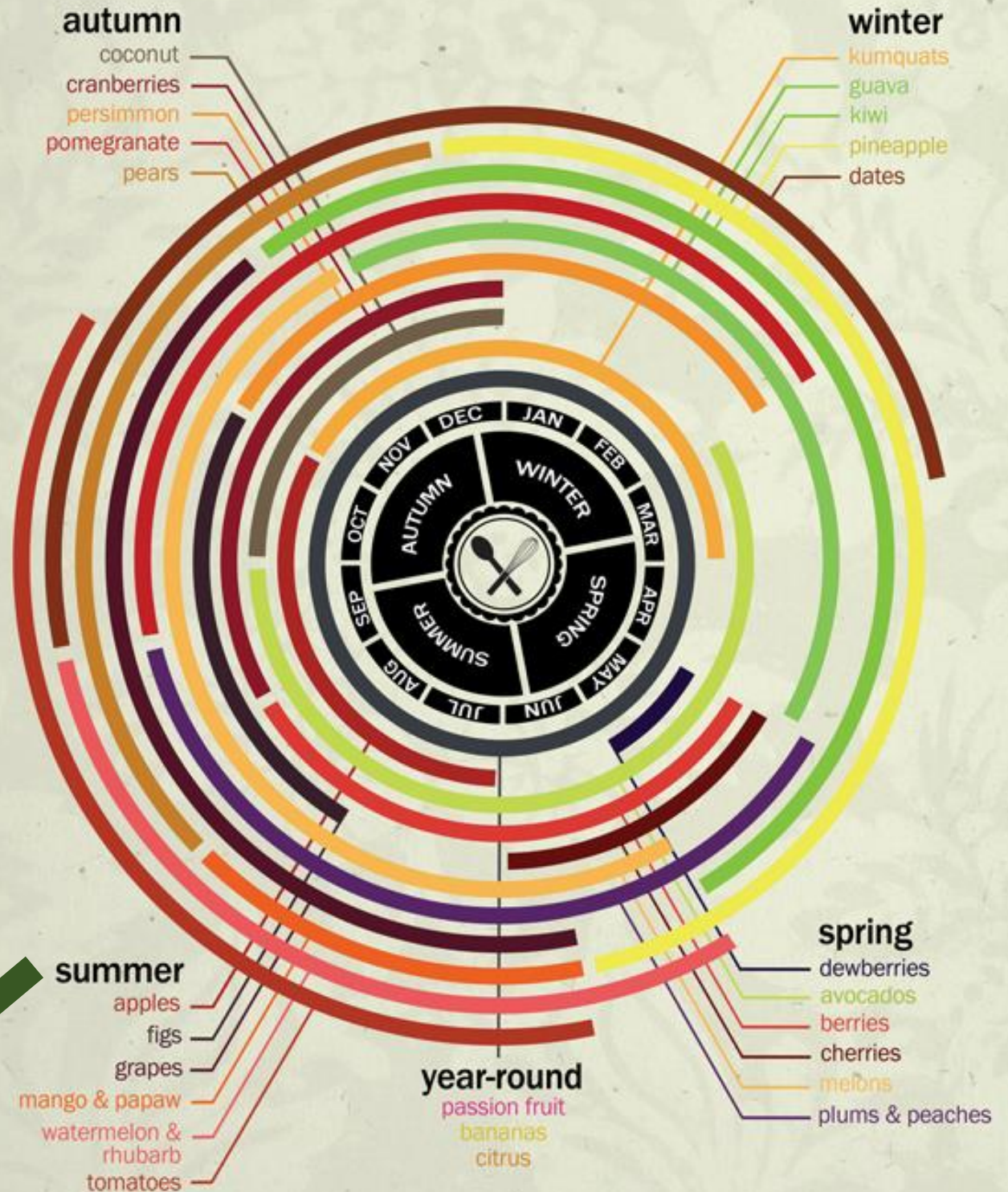
Visualization



Ethics, Privacy and Security Issues

Data Visualization

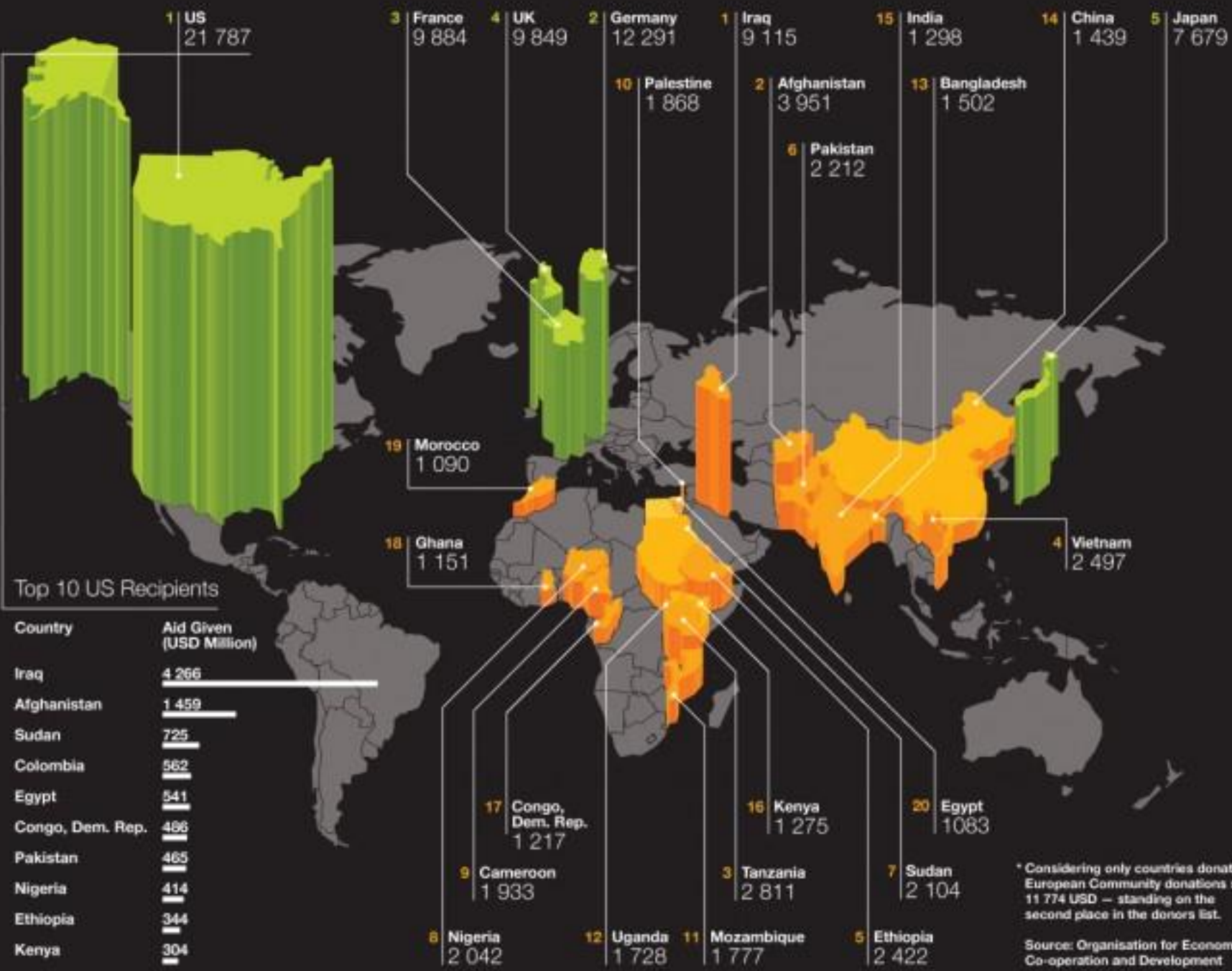
Infoviz is a field of its own.



Eat fruits when they are in season!!!

Developmental Aid Flows around the World

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



* 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?

Agenda



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools

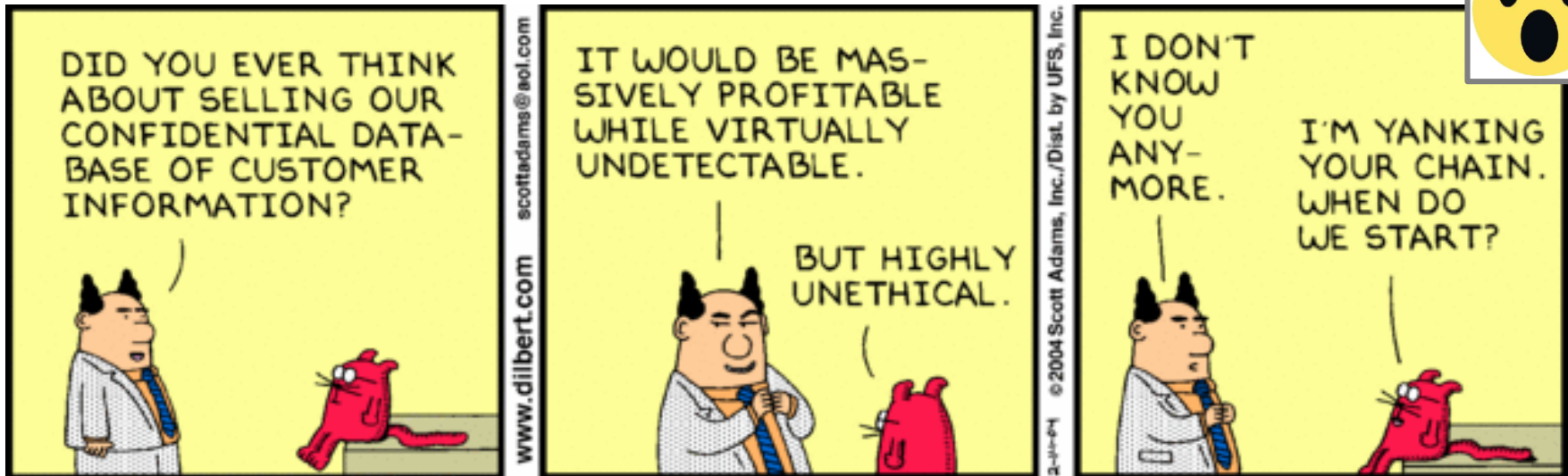


Visualization



Ethics, Privacy and Security Issues

Legal, Privacy and Security Issues



Questions:

- Are we allowed to **collect** the data?
- Are we allowed to **use** the data?
- Is it ethical to use and **act** on the data?
- Is **privacy** preserved in the process?

Problem: Internet is global, but legislation is local!

GDPR



EU law on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA)

Implementation: 25 May 2018


Personal data may not be processed unless there is at least one legal basis to do so. Lawful purposes are:

- Consent by the individual (Opt-in)
 - Legal obligations of the data controller
 - Protect the vital interests of a data subject or another individual
 - To perform a task in the public interest or in official authority
 - For the legitimate interests of a data controller
-
- Applies to US companies doing business in the EU.
 - California passed a similar bill called The California Consumer Privacy Act of 2018.

<https://www.informs.org/About-INFORMS/Privacy-Policy>

← → ↻ 🏠 <https://online.informs.org/info>

INFORMS.org INFORMS Connect PubsOnLine Certified Analyt
2019 Analytics Conference



[Return to the INFORMS Self Service Men](#)

DR MICHAEL HAHLER

Membership and subscription renewal for January 01, 2019 through D
Membership Type: Regular Member

Change Membership Class/Type
If you want to change your INFORMS membership type (regular, retired, or student) yo
need to reselect **ALL** your items. To **cancel all your current items** and reselect, click h
NEW! Multi-Year Memberships!
Pay now for 2 or 3 years at the current rate. Click [HERE](#) to proceed.


PLEASE NOTE: By selecting the multi-year option, you **cancel** your membership order
appears below. You must re-select any community memberships and subscriptions.

Current Items

MEMBERSHIPS

INFORMS site uses cookies to store information on your computer. Some are
essential to make our site work; Others help us improve the user experience.
By using this site, you consent to the placement of these cookies. Please read
our [Privacy Statement](#) to learn more. \$10.00

← → ↻ 🏠 <https://www.informs.org/About>



[Home](#) > [About INFORMS](#) > [Privacy Policy](#)

☰ IN THIS SECTION

Privacy Policy

SHARE: [f](#) [in](#) [t](#) [✉](#)

Effective: May 25, 2018

With over 12,500 members from around the globe, Institute for
Operations Research and the Management Sciences (INFORMS) is the
leading international association for professionals in operations research
and analytics.

[Agree](#)

Privacy

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**

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.

Security

A screenshot of a ZDNet article header. The ZDNet logo is in the top left. Navigation links include 'BEST VPNS', 'CLOUD', 'SECURITY', 'AI', 'MORE', 'NEWSLETTERS', and 'ALL WRITERS'. The date and time are 'February 13, 2014 -- 08:24 GMT (00:24 PST)'. The main headline is 'How hackers stole millions of credit card records from Target'. Below it is a sub-headline: 'How did the cyberattack on Target, which resulted in the theft of millions of records, take place?'

ZDNet

BEST VPNS CLOUD SECURITY AI MORE NEWSLETTERS ALL WRITERS

February 13, 2014 -- 08:24 GMT (00:24 PST)

How hackers stole millions of credit card records from Target

How did the cyberattack on Target, which resulted in the theft of millions of records, take place?

<https://www.zdnet.com/article/how-hackers-stole-millions-of-credit-card-records-from-target/>

A screenshot of a Seattle Times article header. The Seattle Times logo is at the top right. A hamburger menu icon is on the left. The category is 'Local Politics'. Navigation links include 'Business', 'Data', 'Economy', 'Local News', 'Local Politics', and 'Technology'. The main headline is 'Banking, Social Security info of more than 1.4 million people exposed in hack involving Washington state auditor'. Below it is the date and time: 'Feb. 1, 2021 at 10:39 am | Updated Feb. 3, 2021 at 4:57 pm'.

The Seattle Times

Local Politics

Business | Data | Economy | Local News | Local Politics | Technology

Banking, Social Security info of more than 1.4 million people exposed in hack involving Washington state auditor

Feb. 1, 2021 at 10:39 am | Updated Feb. 3, 2021 at 4:57 pm

<https://www.seattletimes.com/seattle-news/politics/personal-data-of-1-6-million-washington-unemployment-claimants-exposed-in-hack-of-state-auditor/>

What you should know now...



What is Data Science?



Who is a Data Scientist?



The Data Science Process



Data Science Tools



Visualization



Ethics, Privacy and Security Issues