Introduction to data-driven decision making

Data science, Analytics, Business Intelligence & Market Research

Prof. Dr. Jan Kirenz
HdM Stuttgart

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

“The ability to —

take data,

be able to **understand** it,

process it,

extract value from it,

visualize it,

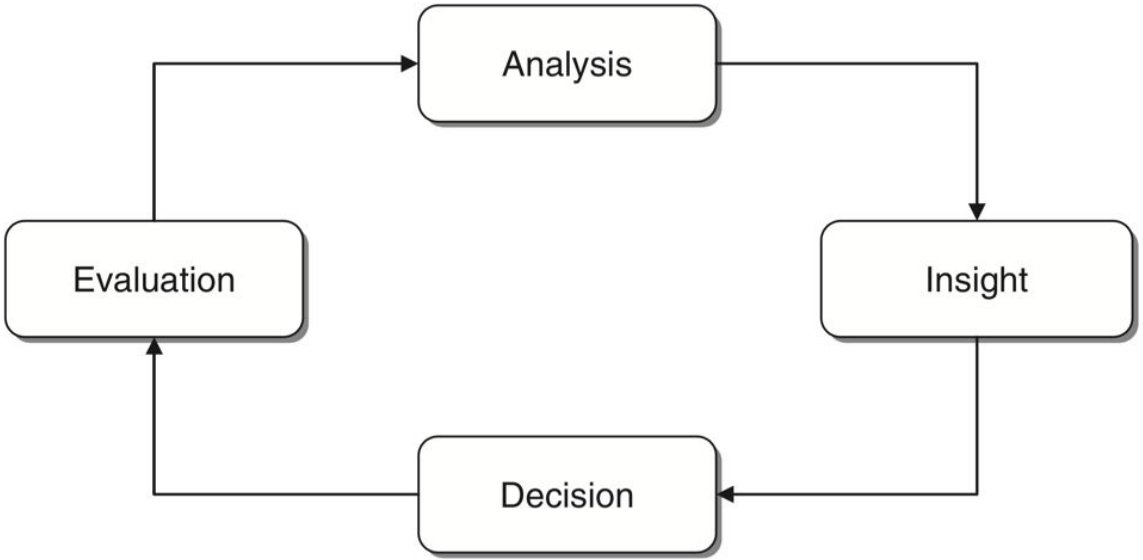
communicate it —

that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

Data-driven decision making is the process of **making decisions** based on actual **data** rather than intuition or personal experience alone.

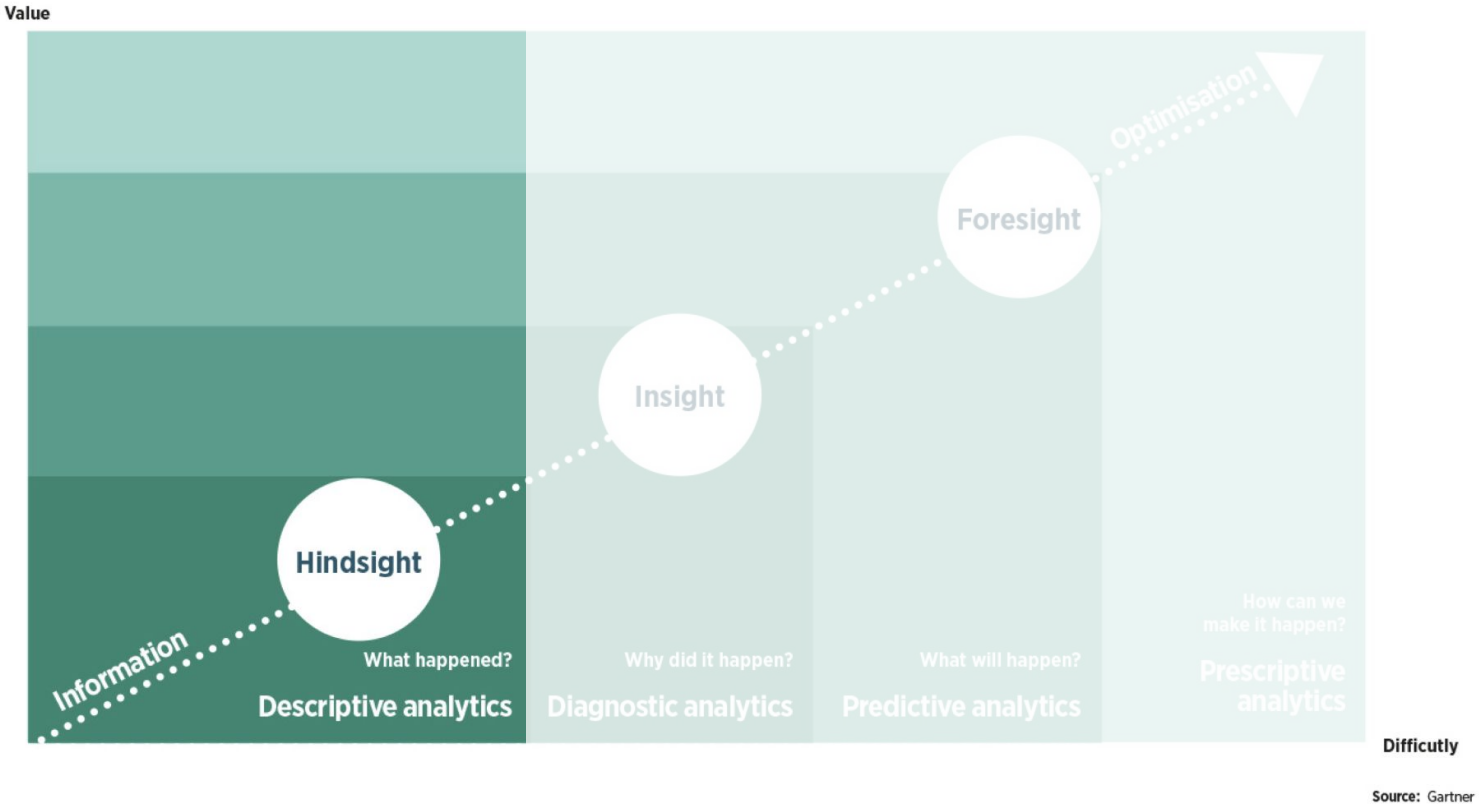
Cycle of data driven **decision making**



Source: Vercellis (2009)

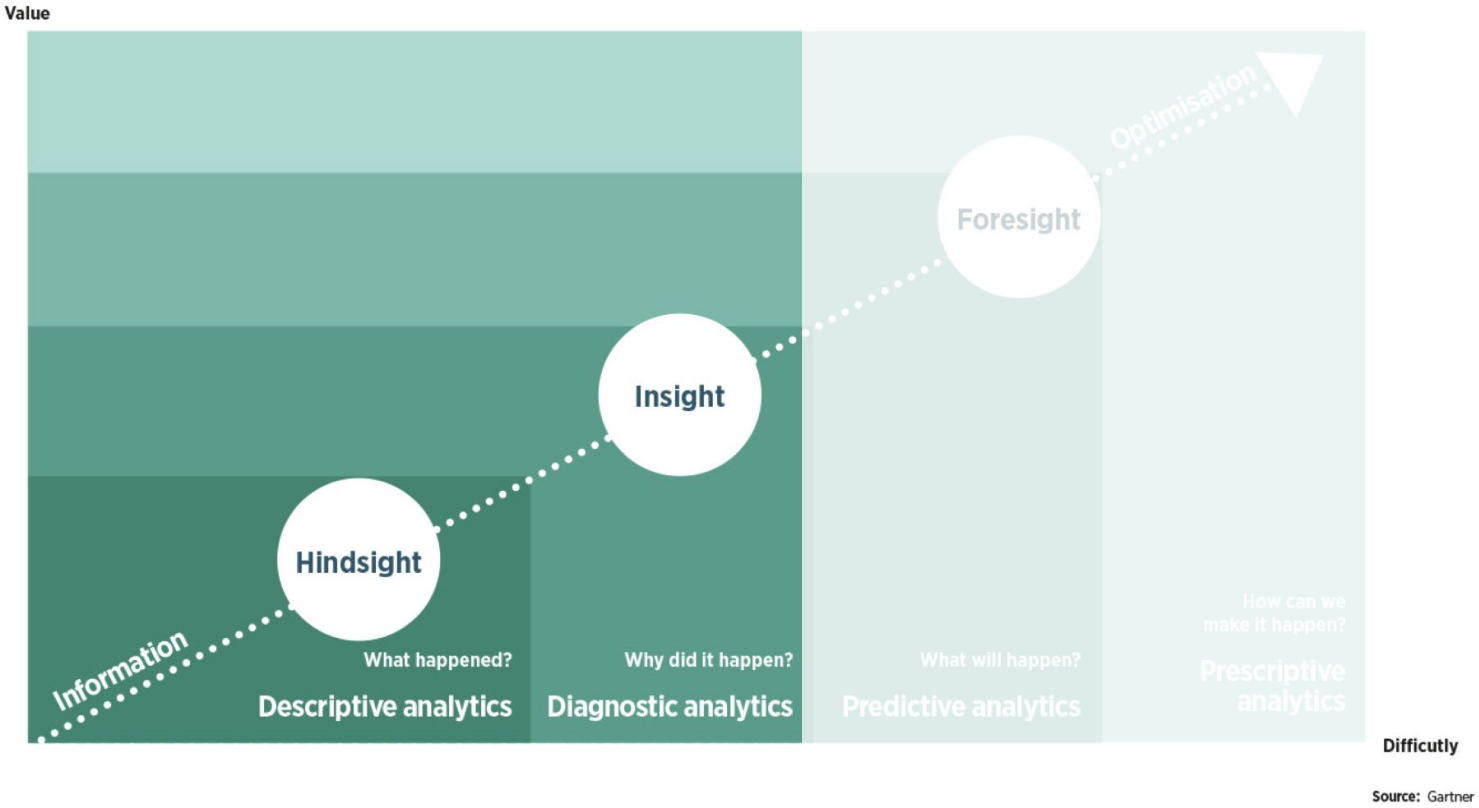
Types of analytics

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS



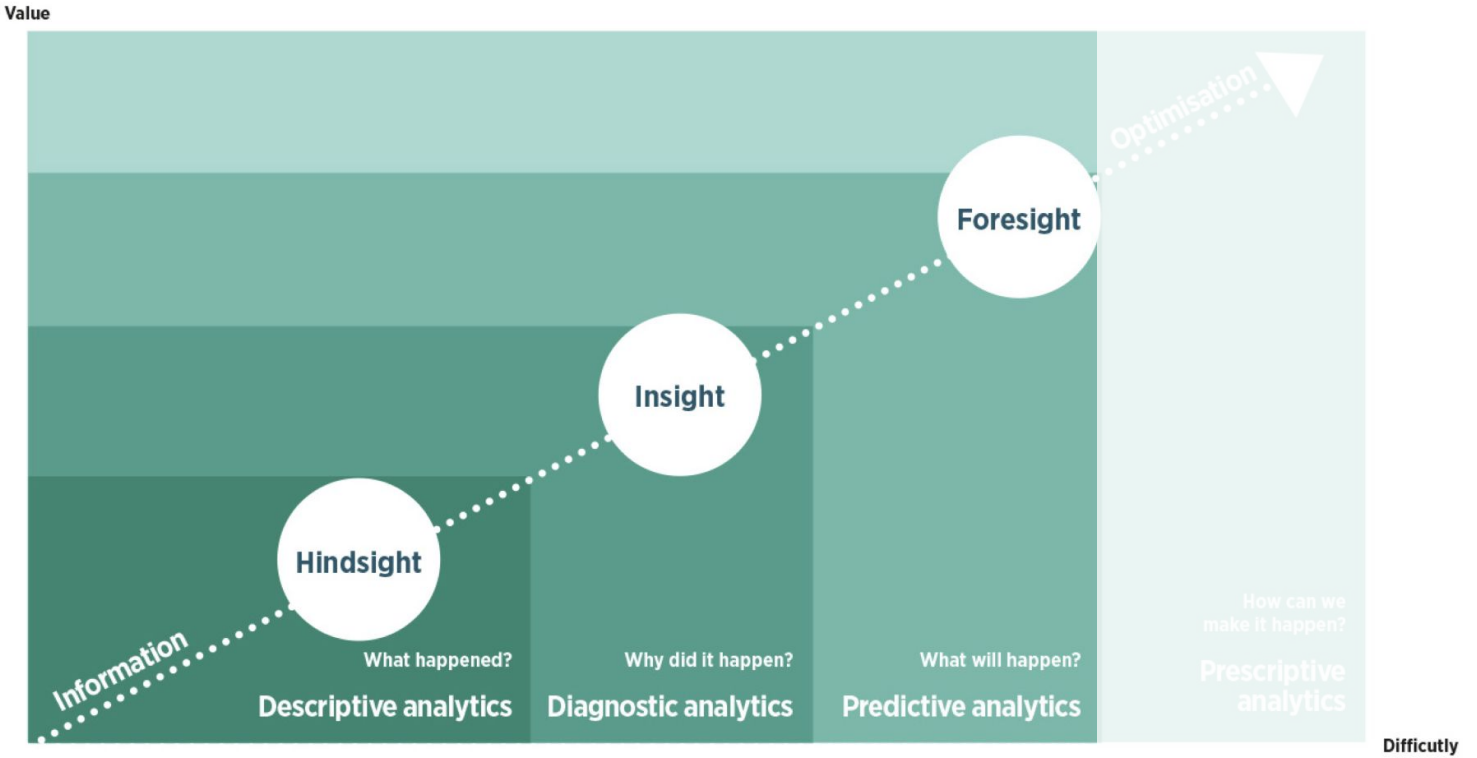
Source: Gartner

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS



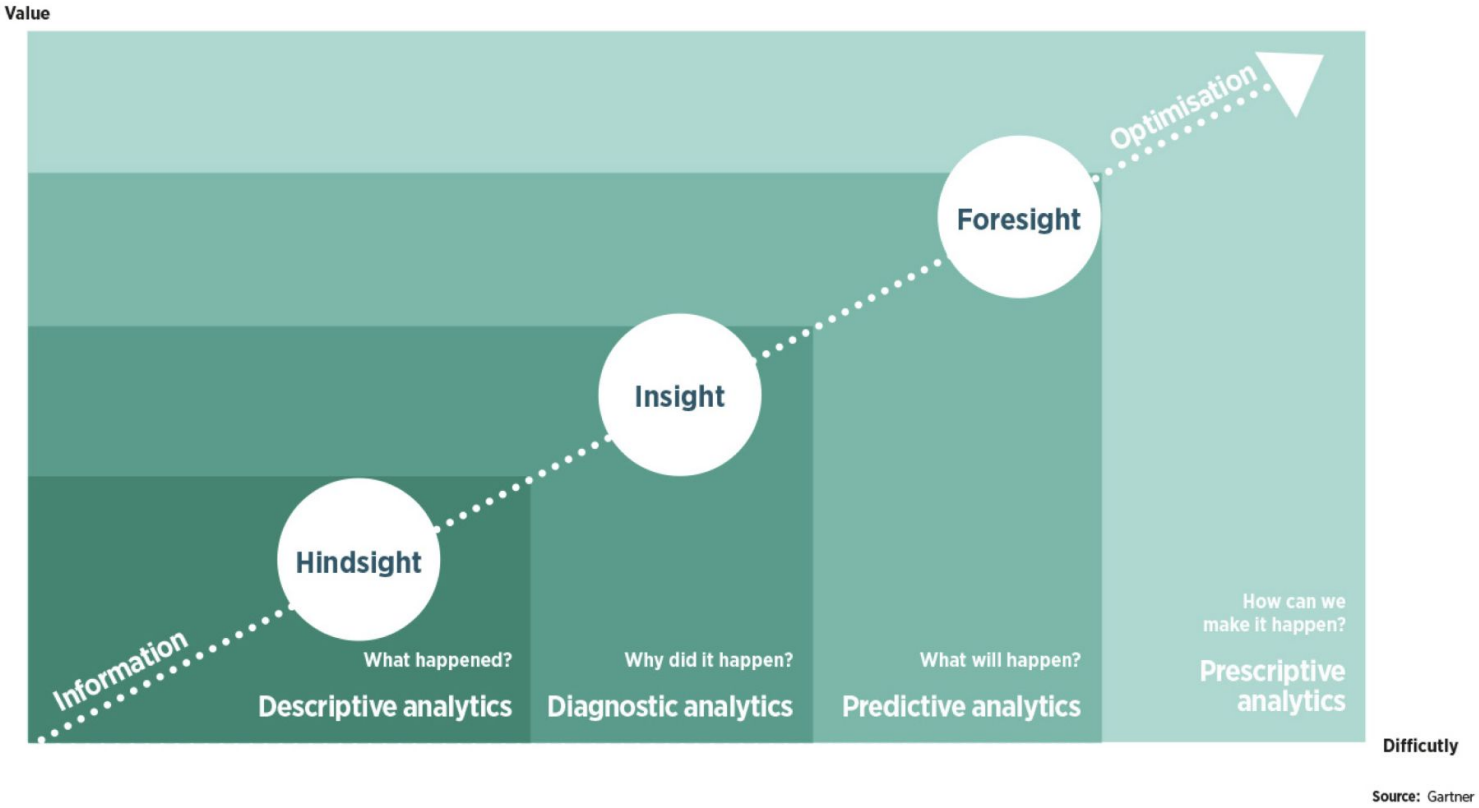
Source: Gartner

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS



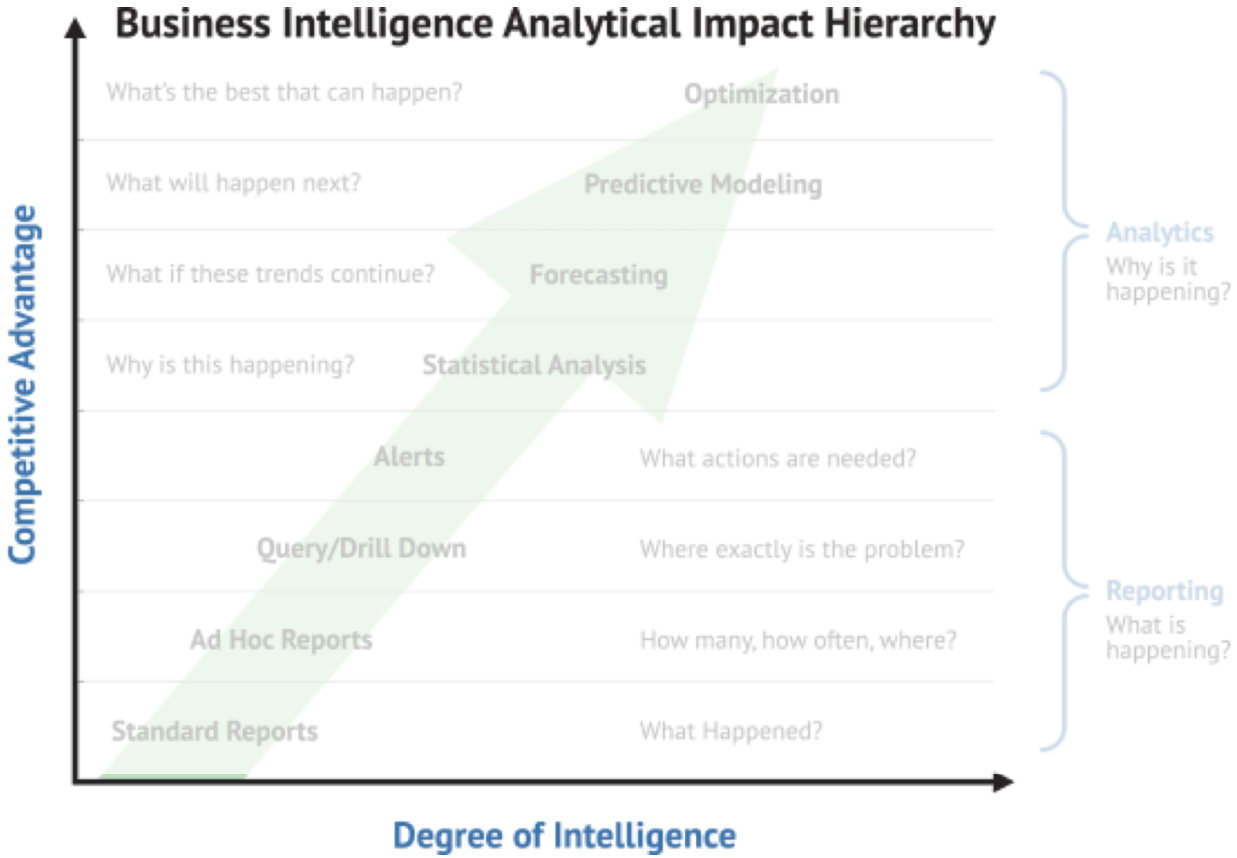
Source: Gartner

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS

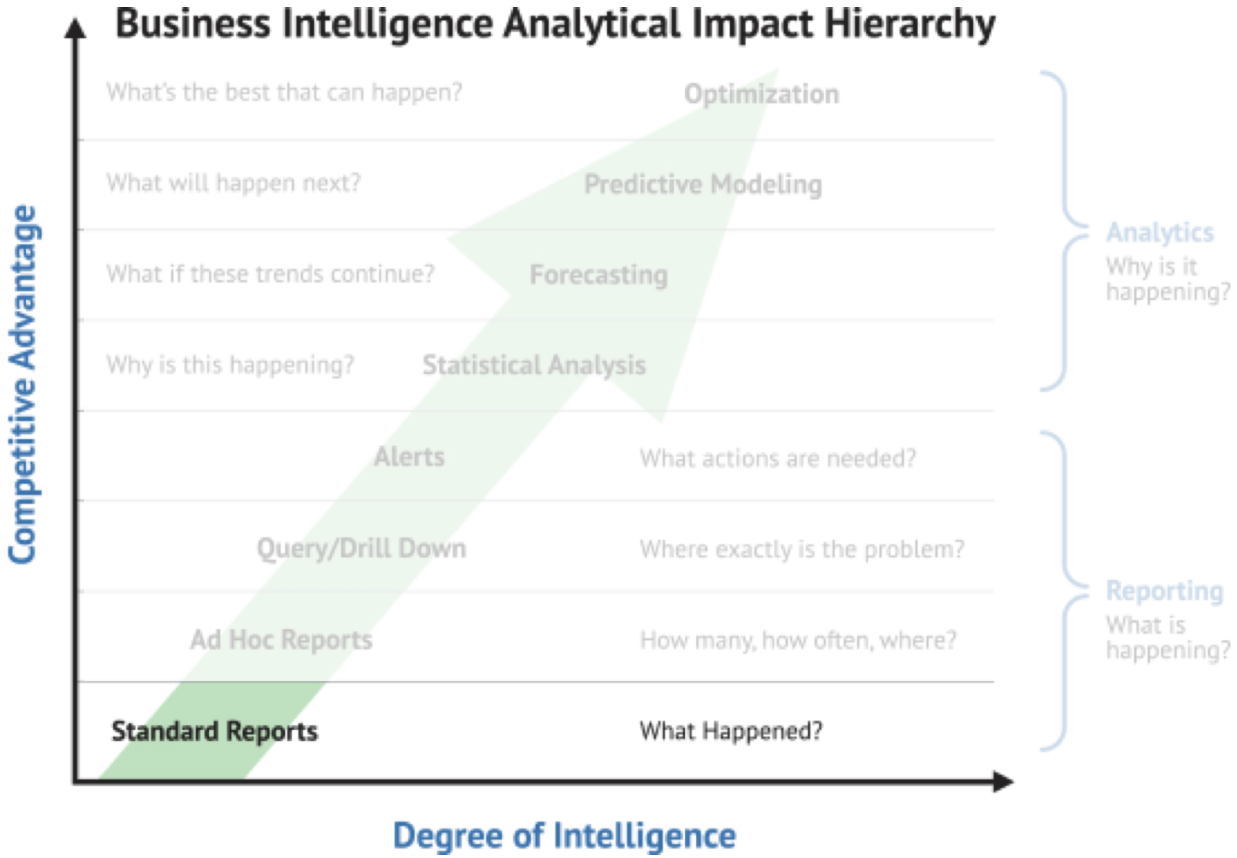


Source: Gartner

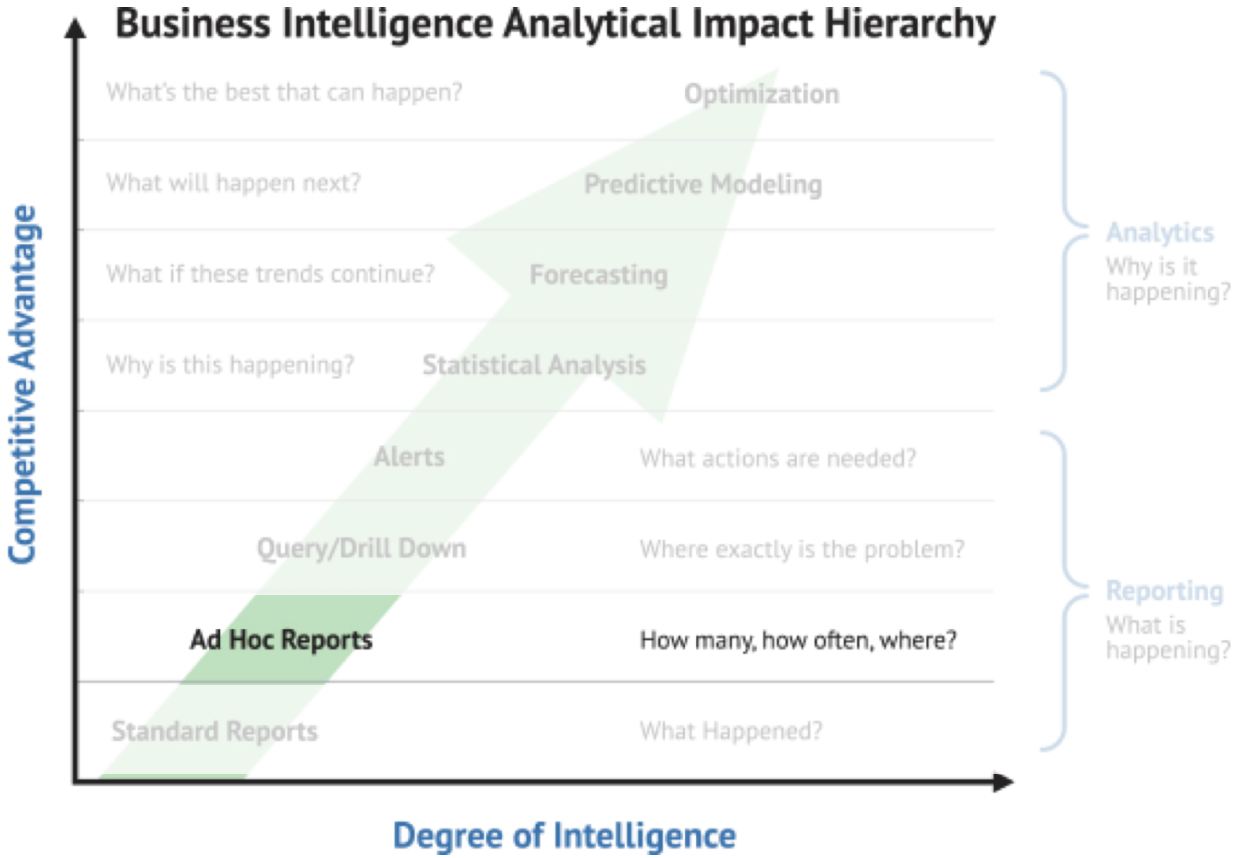
Business intelligence analytical impact hierarchy



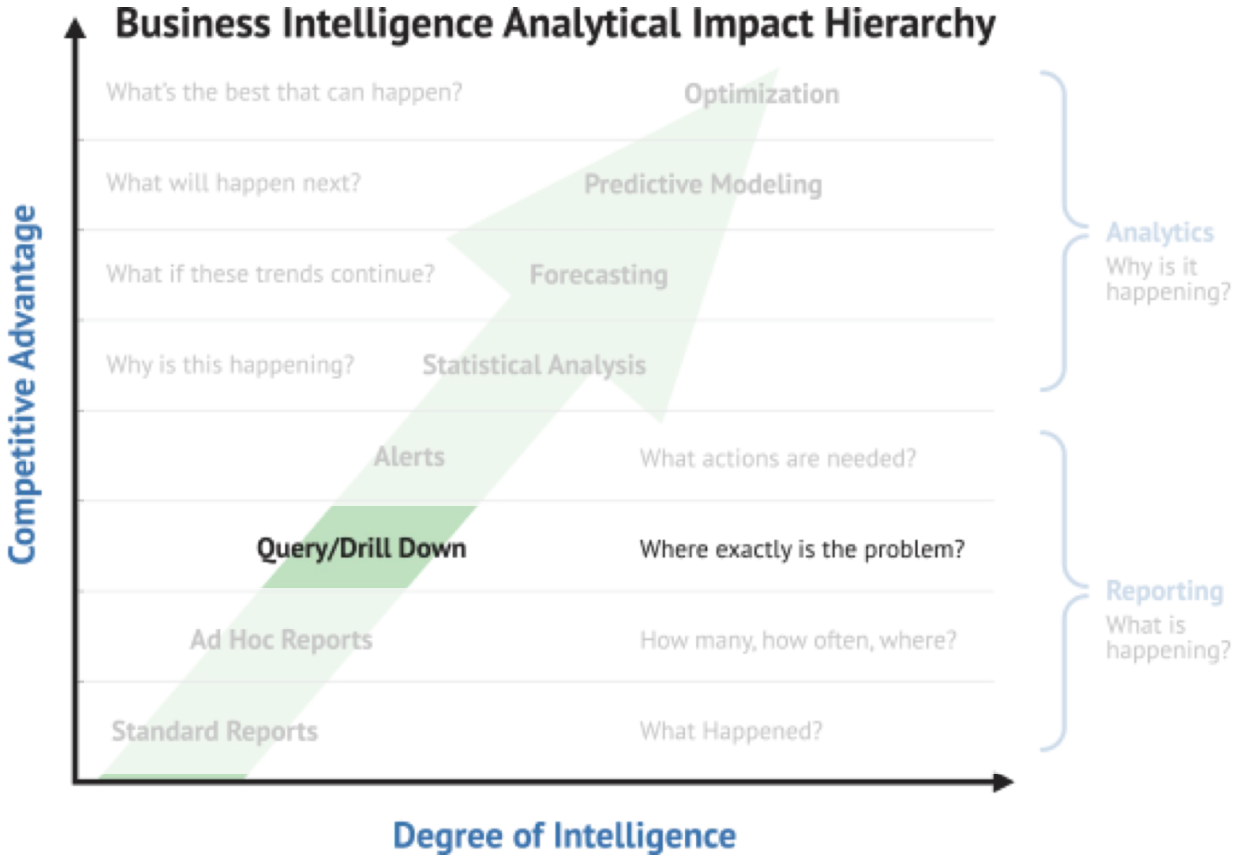
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



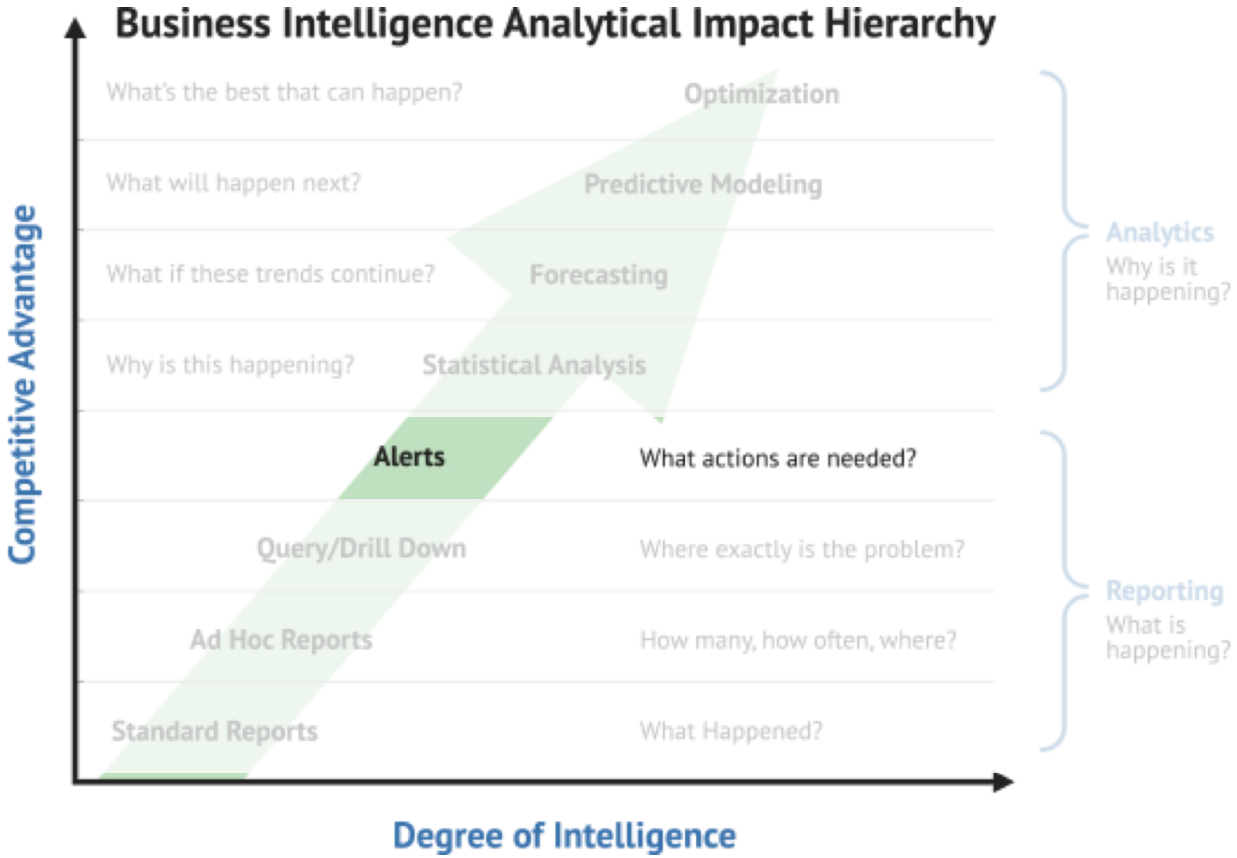
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



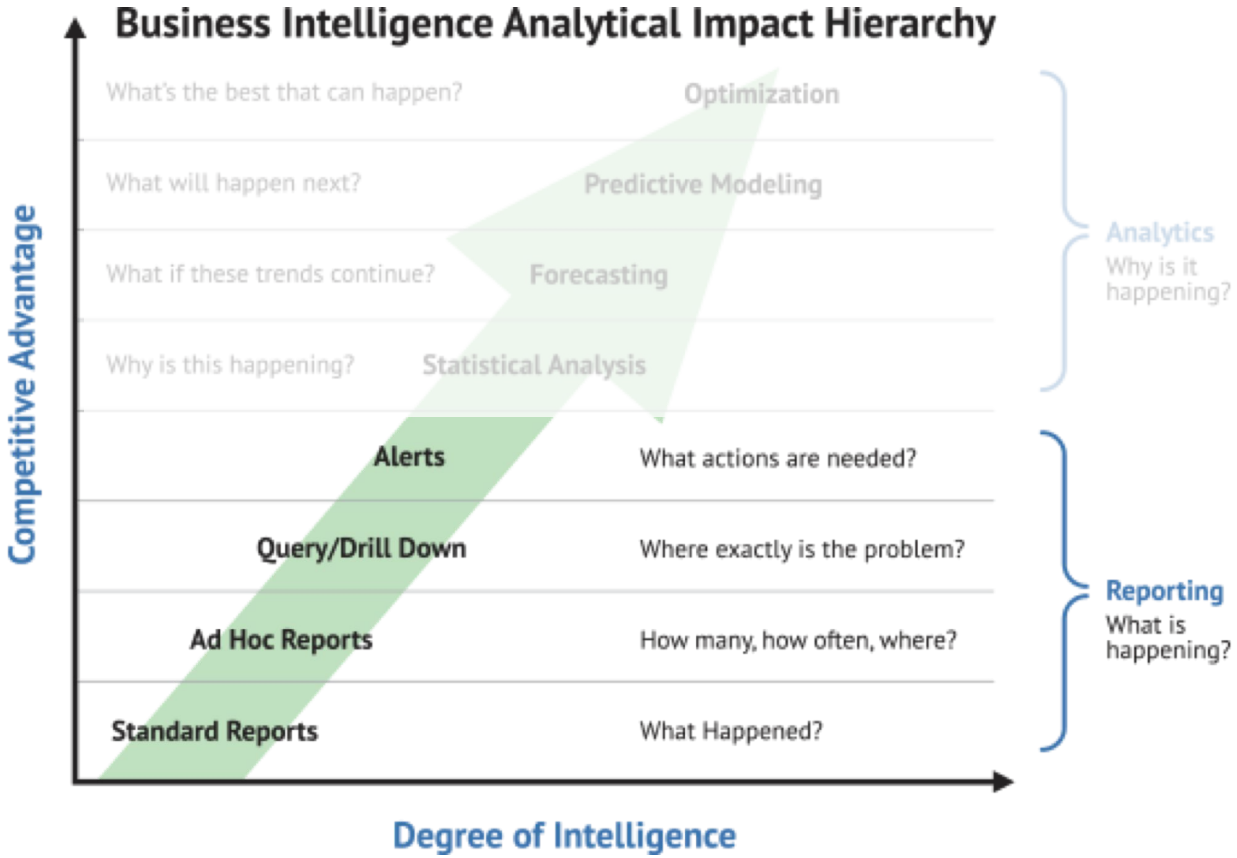
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



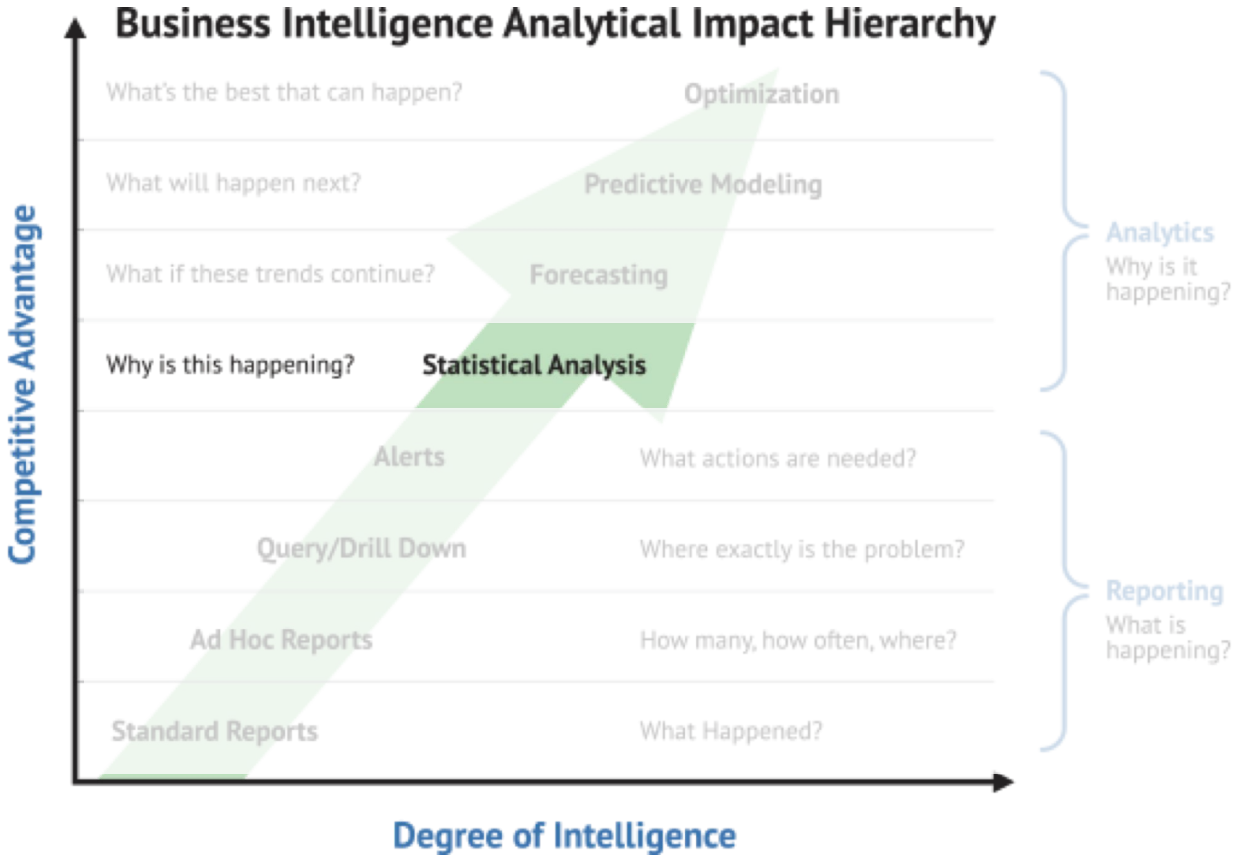
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



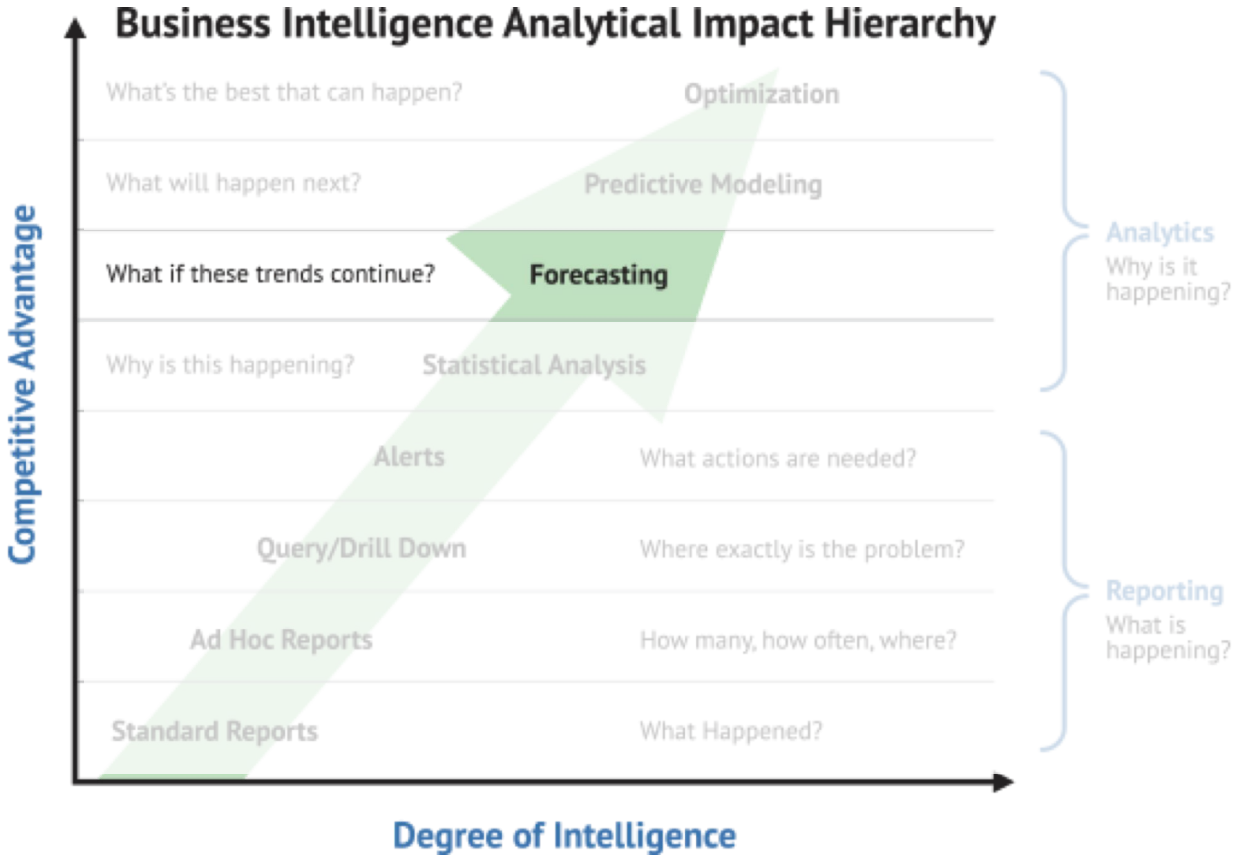
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



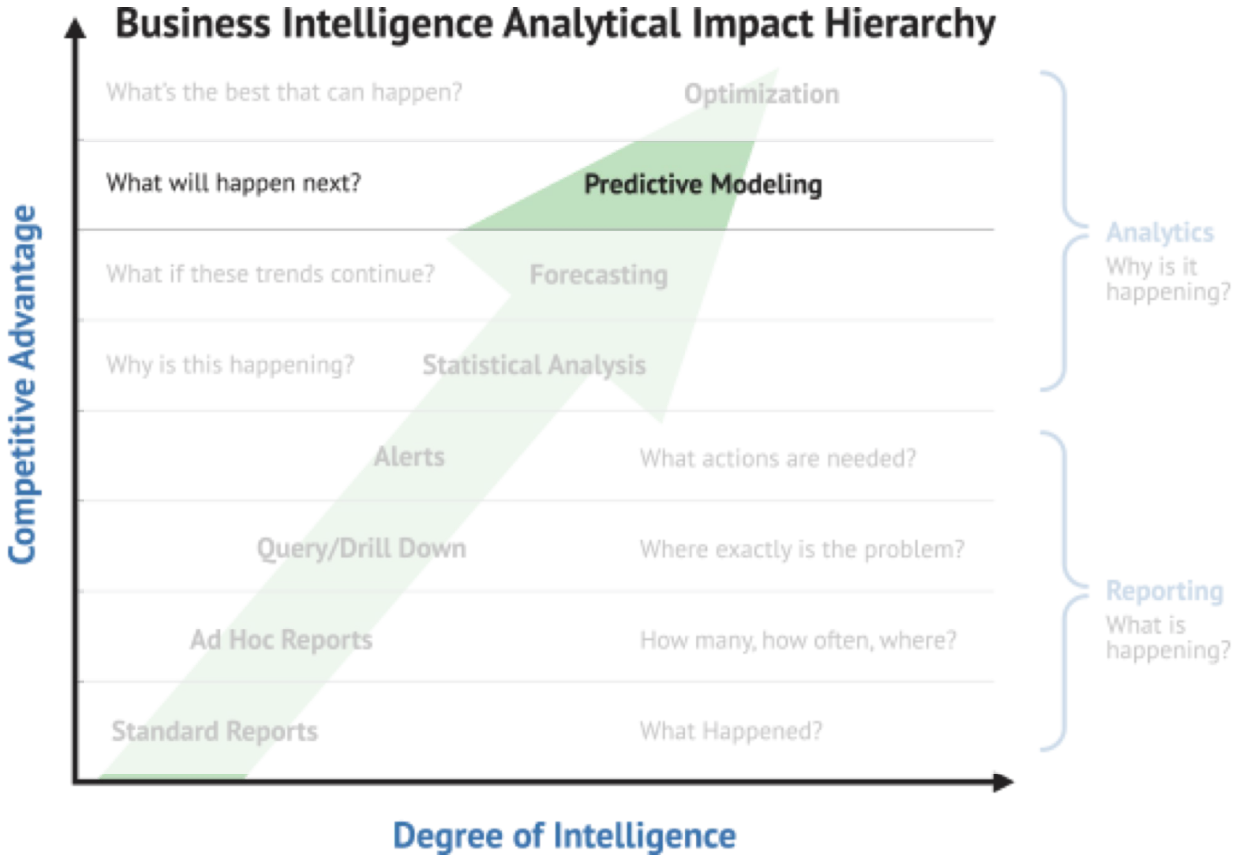
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



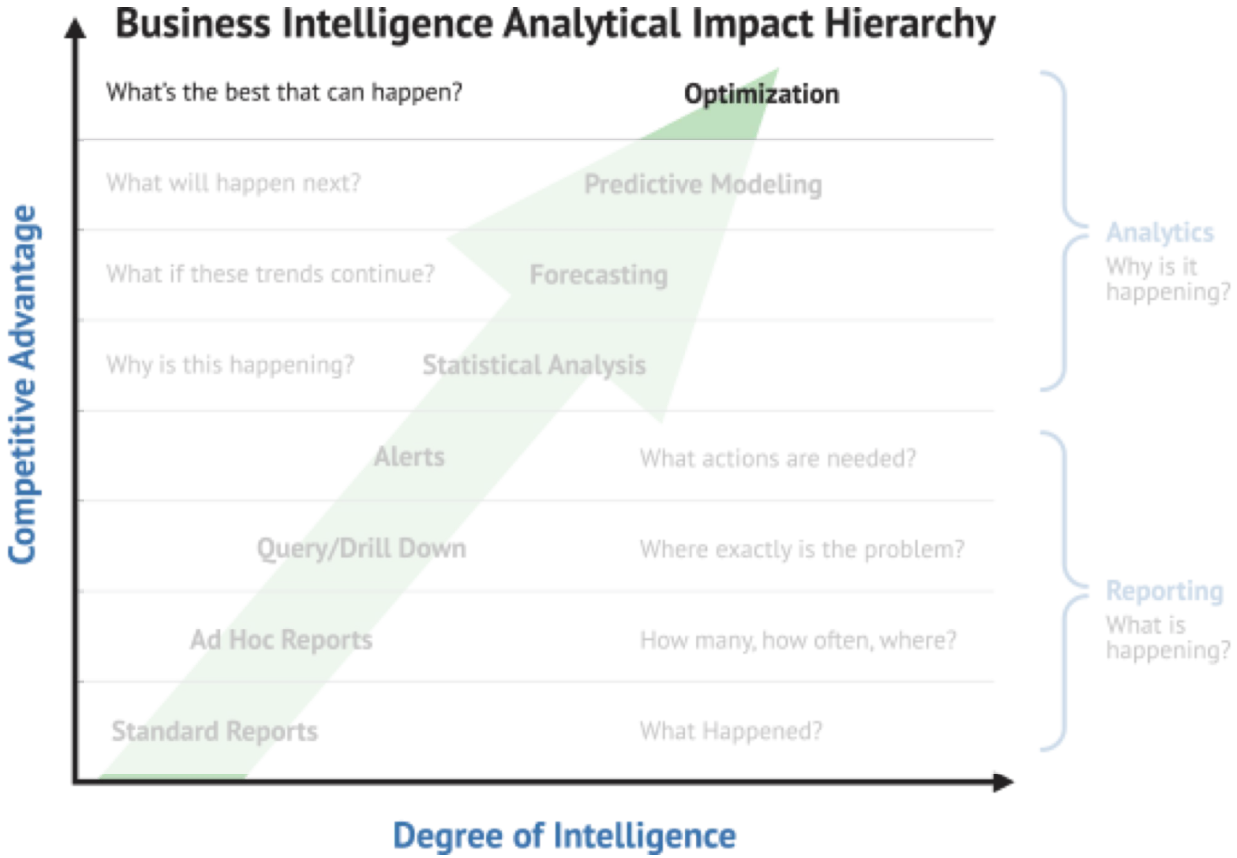
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



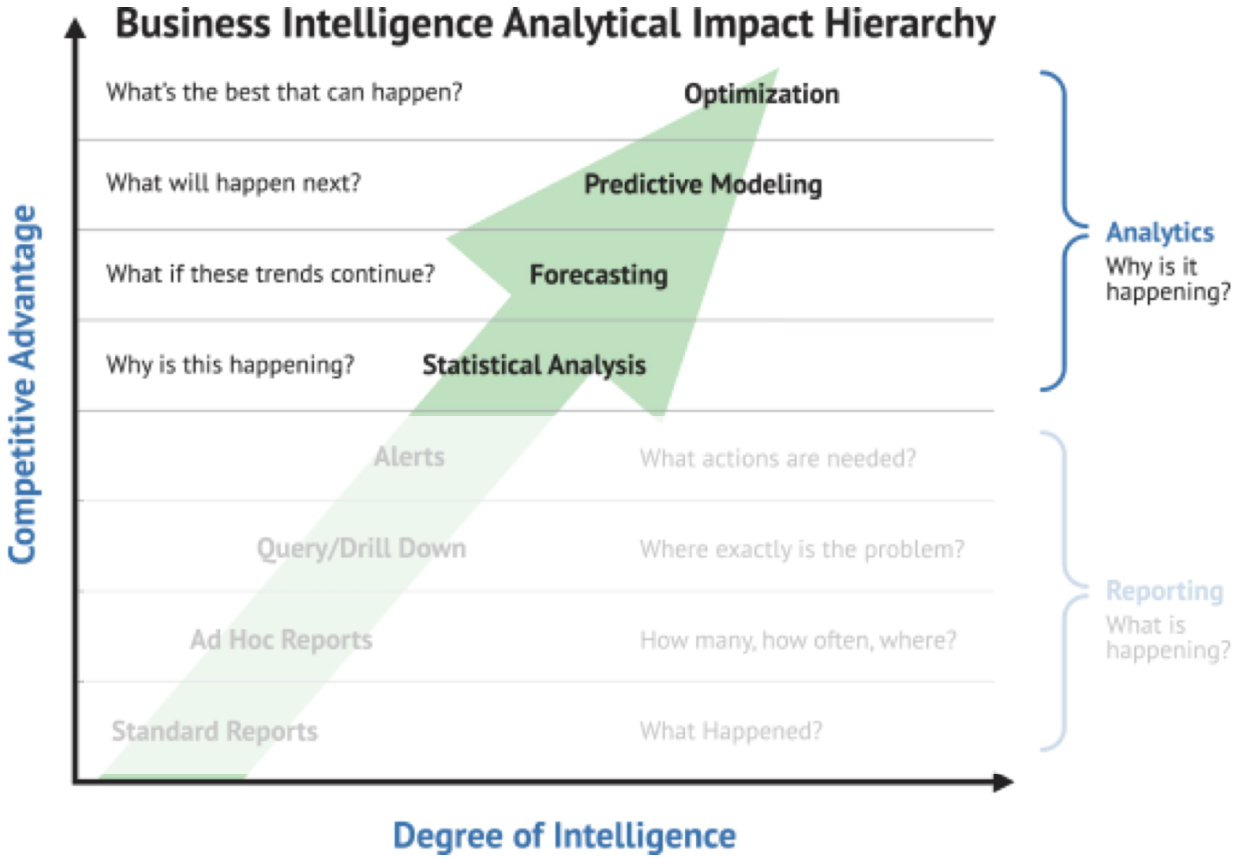
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



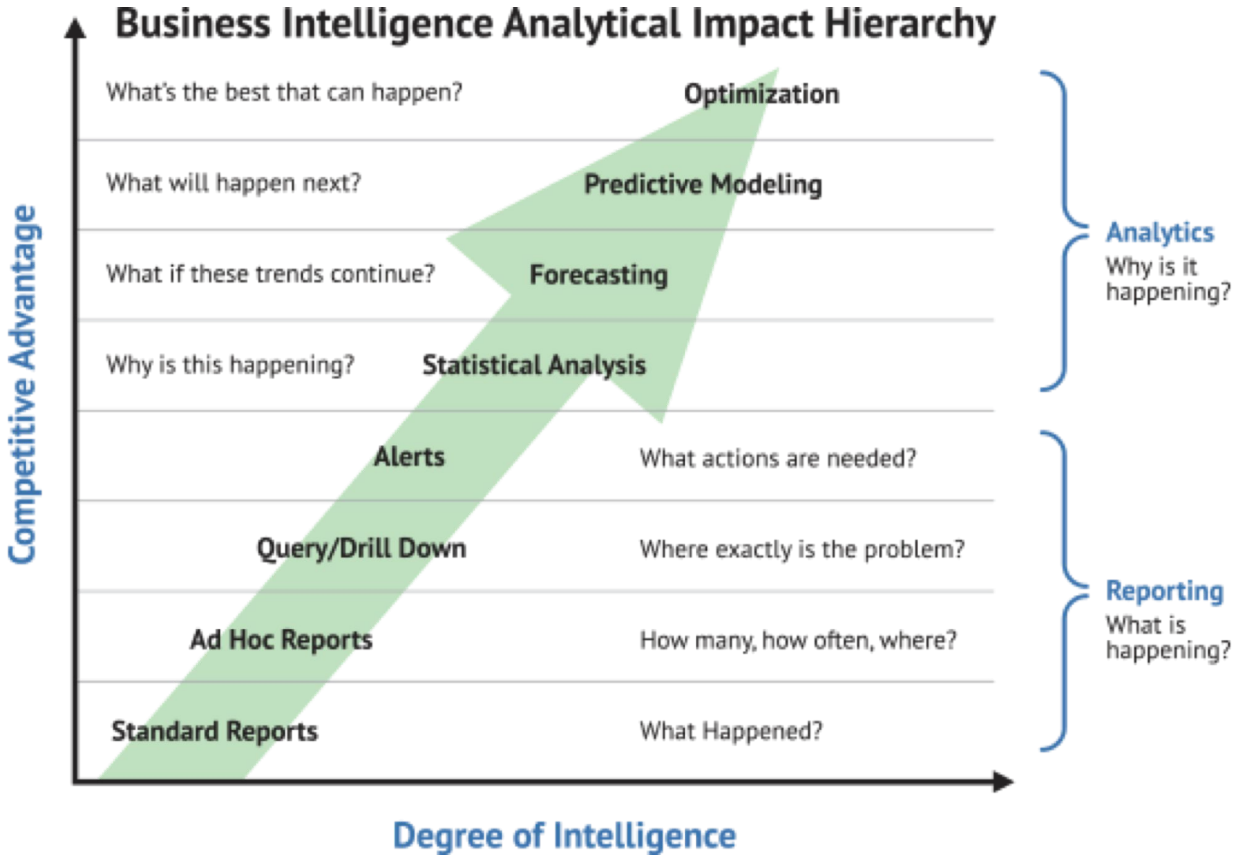
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



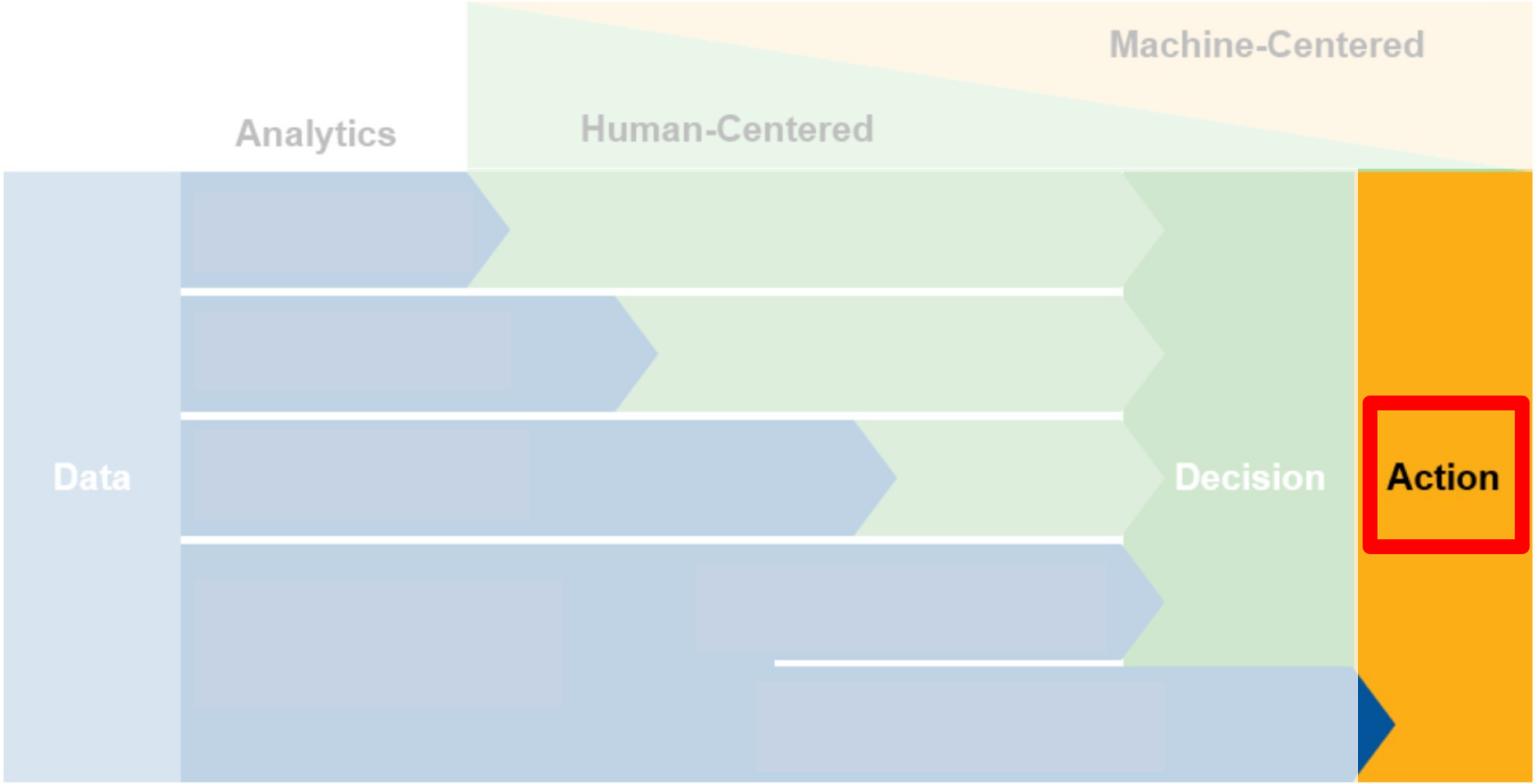
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)



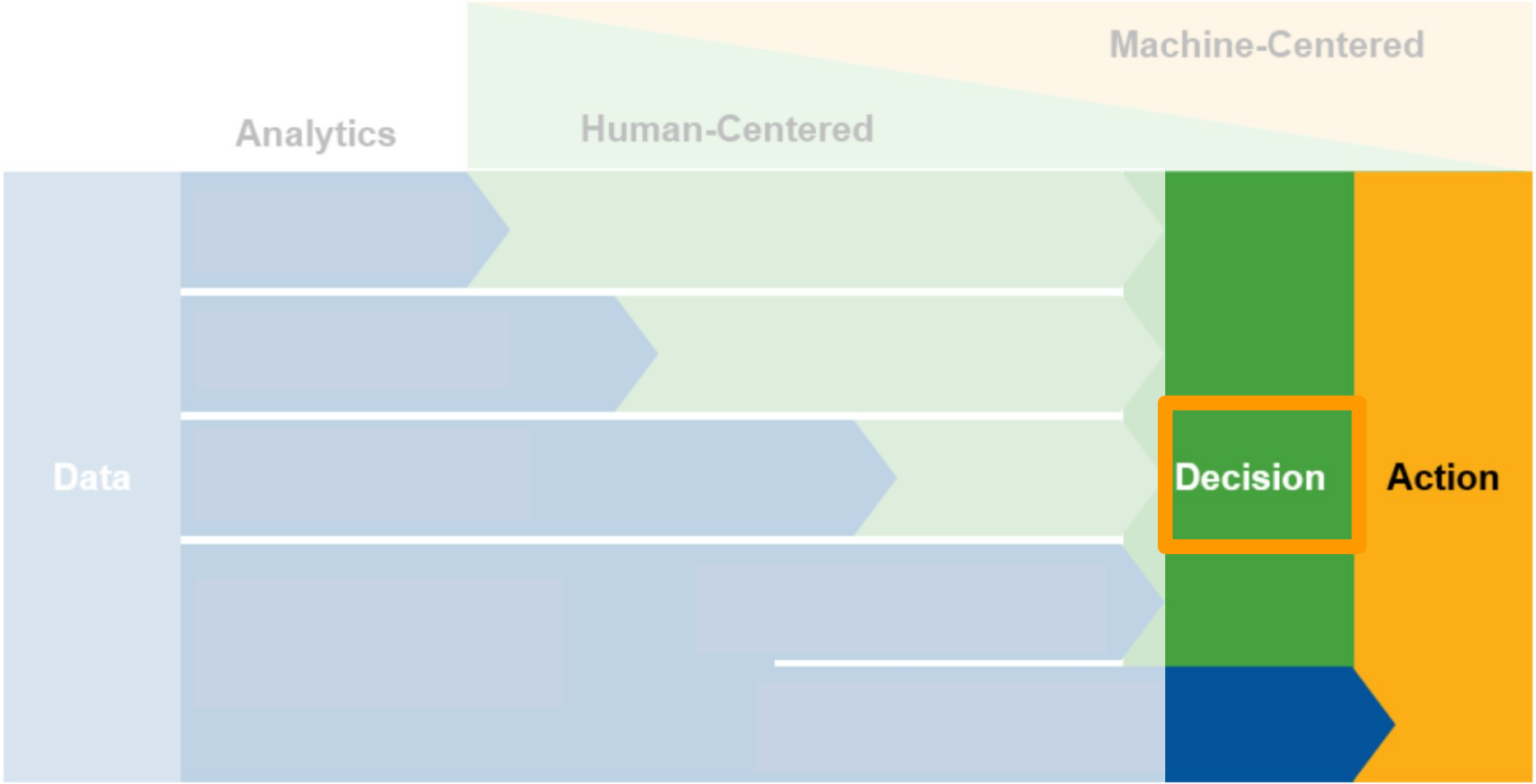
Business Intelligence Analytical Impact Hierarchy (Davenport & Harris, 2007)

Analytics in decision support systems

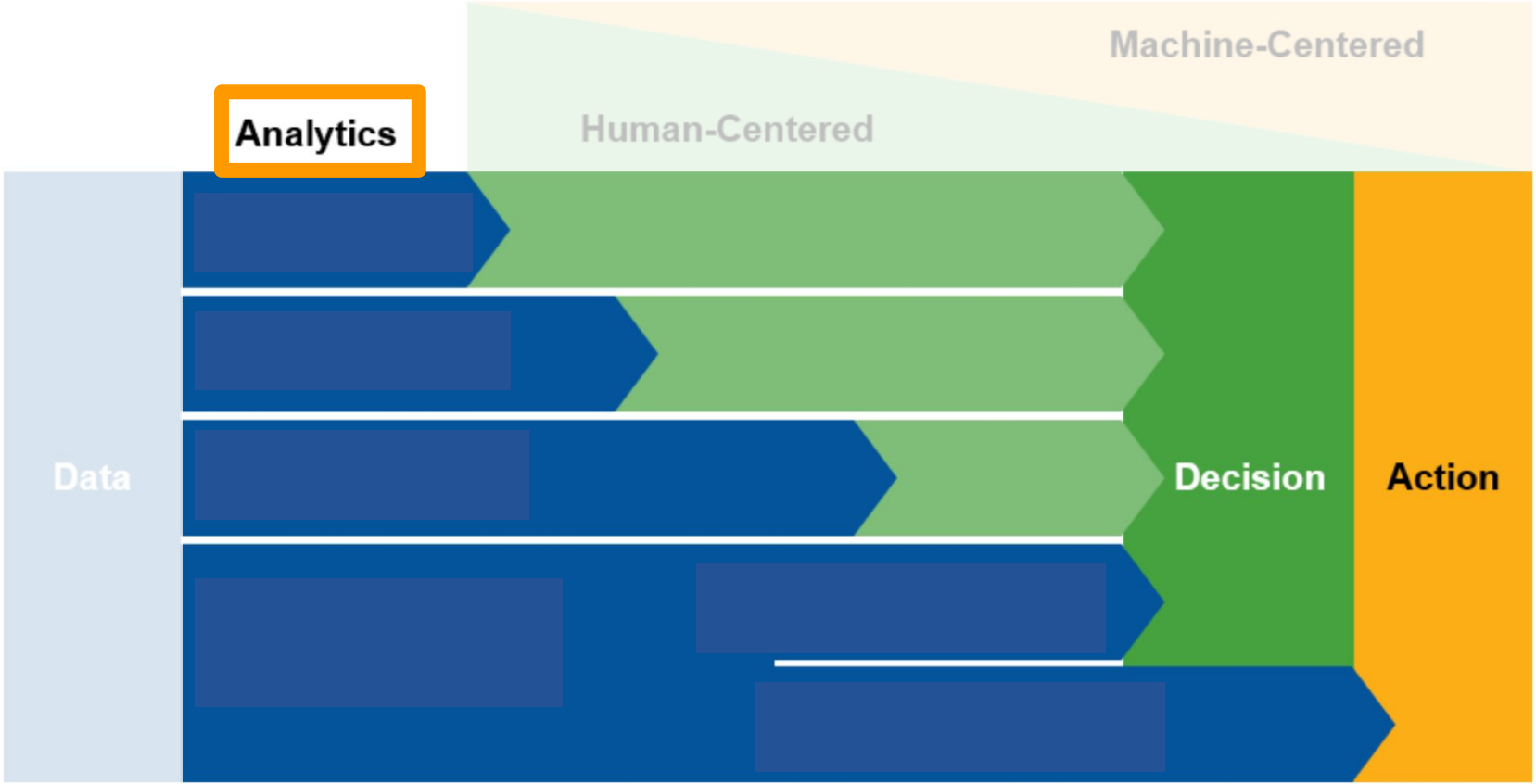
Prof. Dr. Jan Kirenz



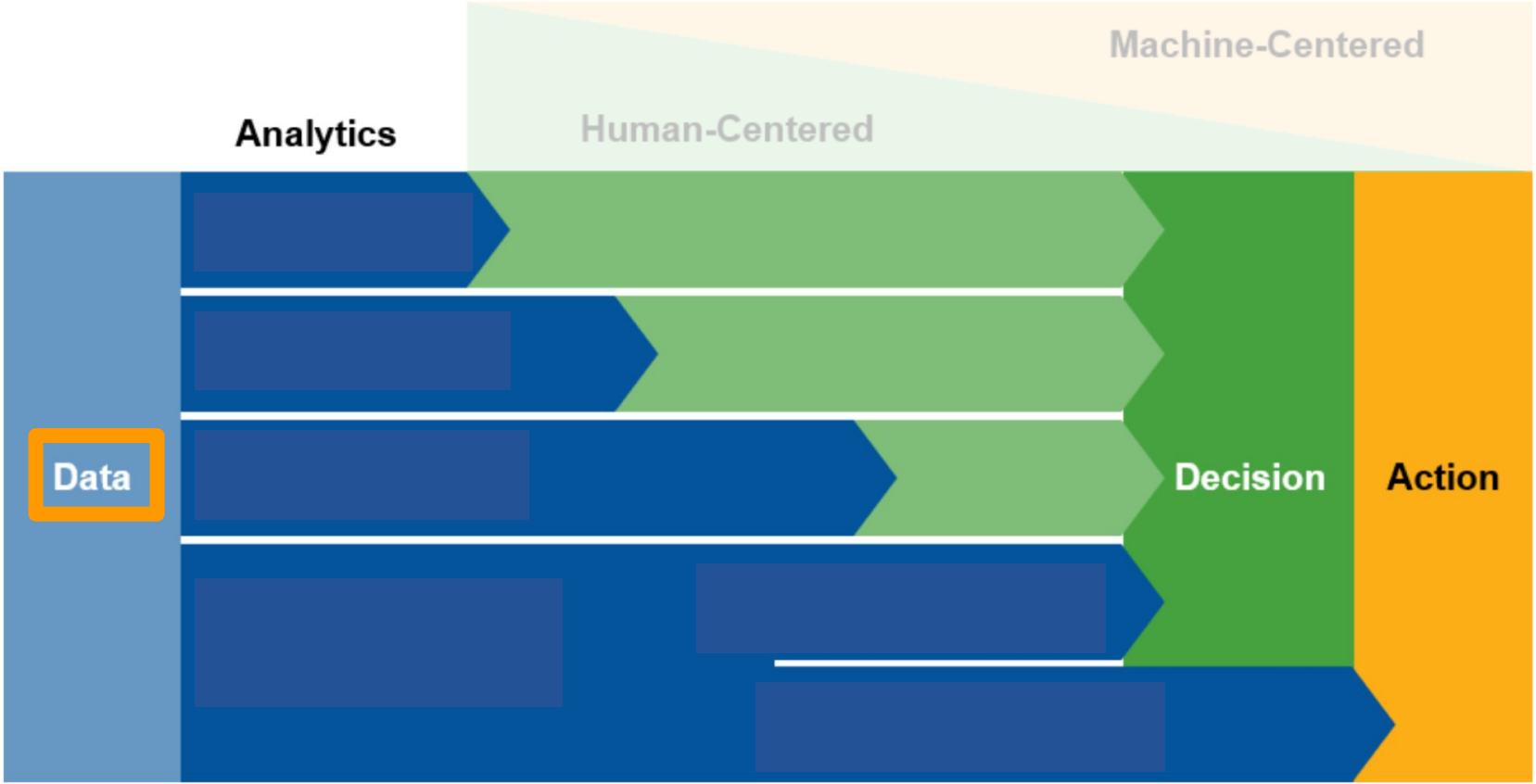
Source: Gartner (October 2016)



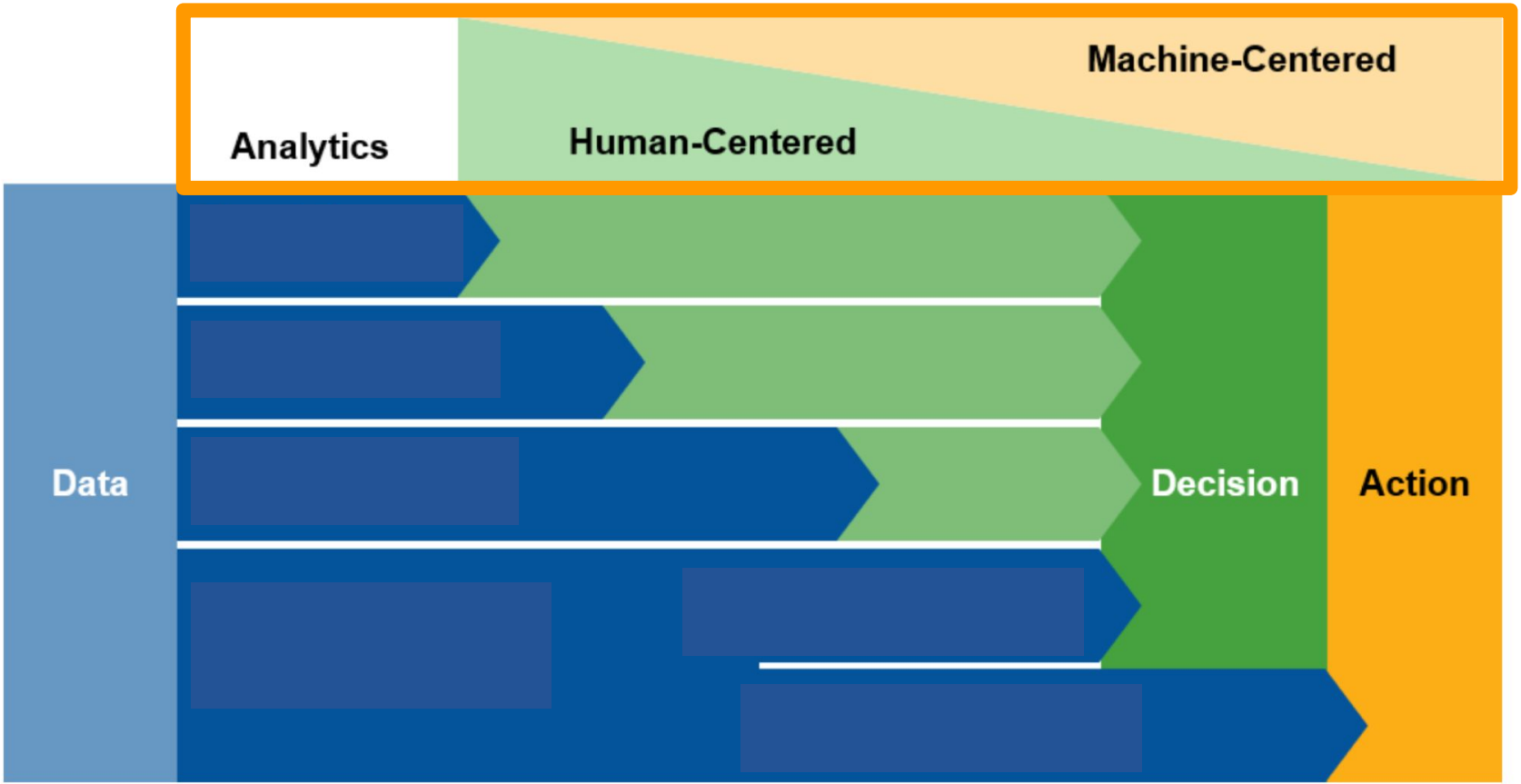
Source: Gartner (October 2016)



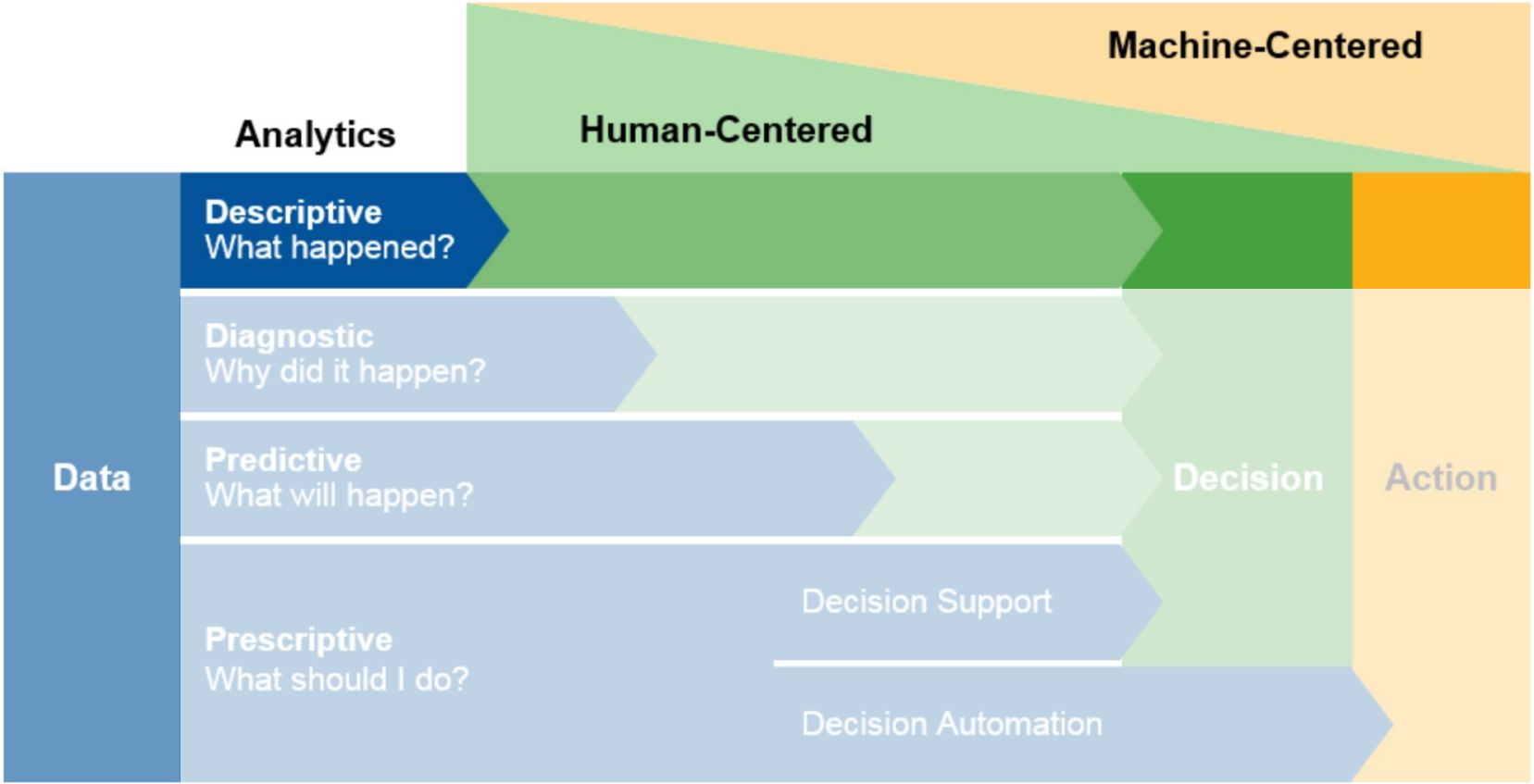
Source: Gartner (October 2016)



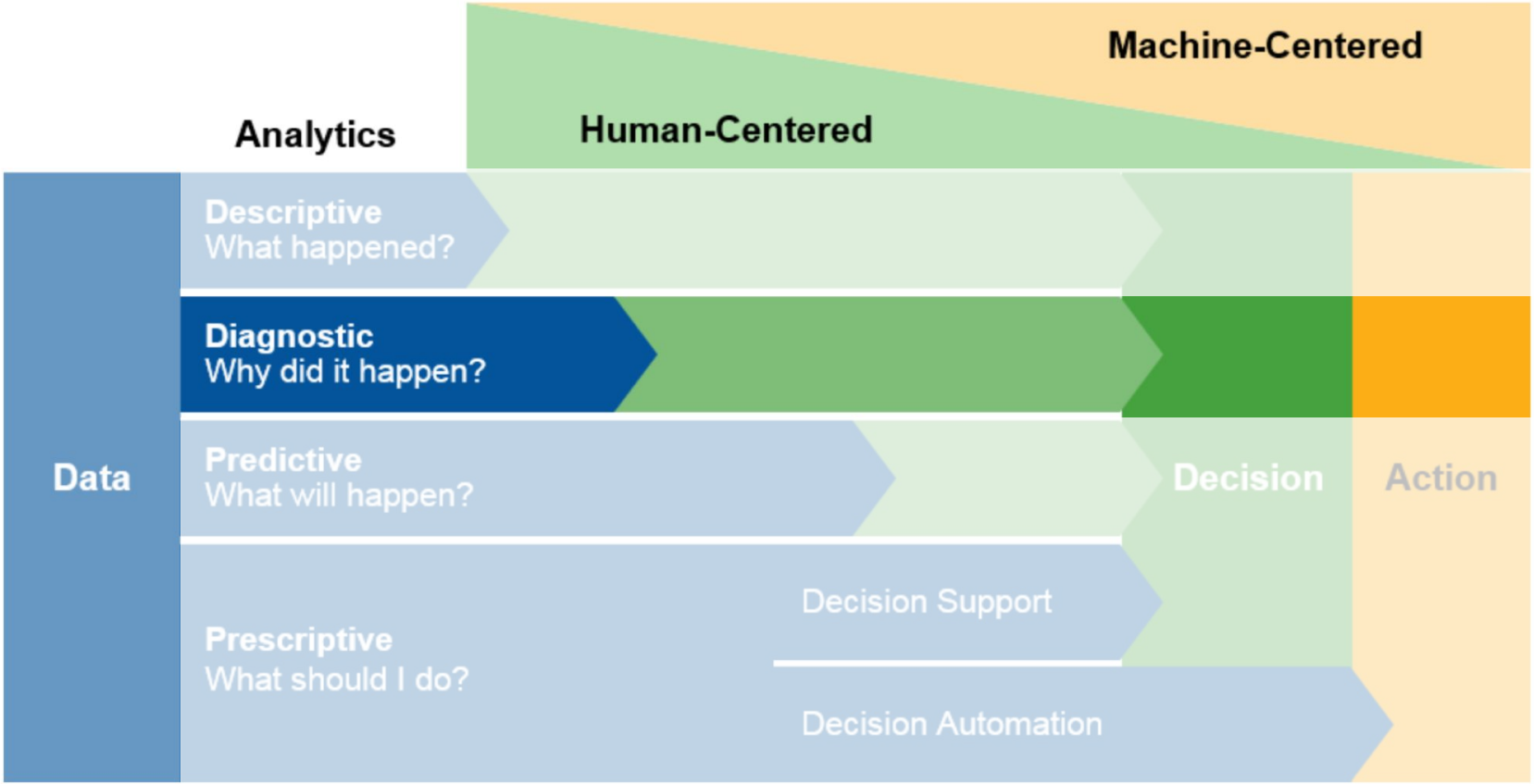
Source: Gartner (October 2016)



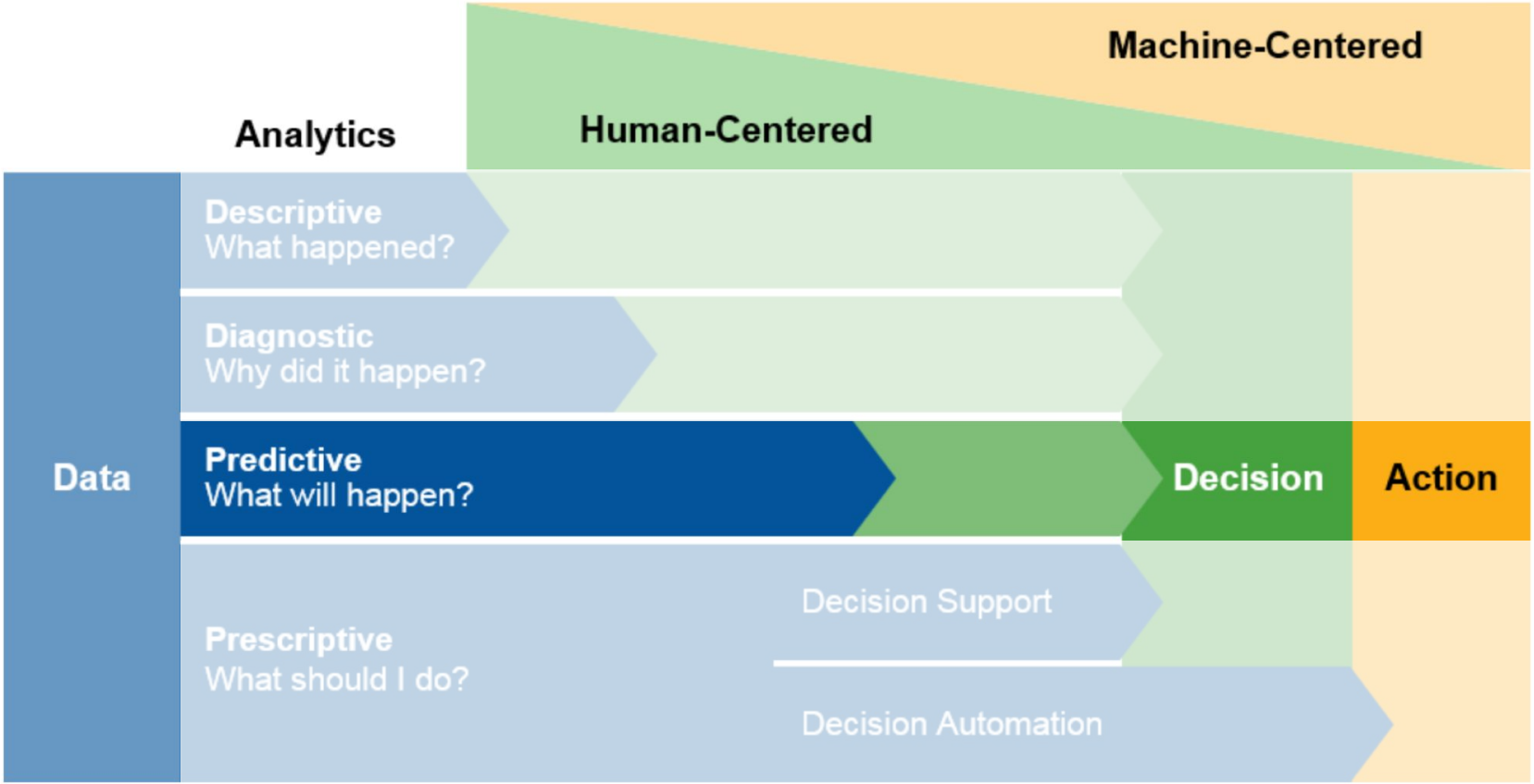
Source: Gartner (October 2016)



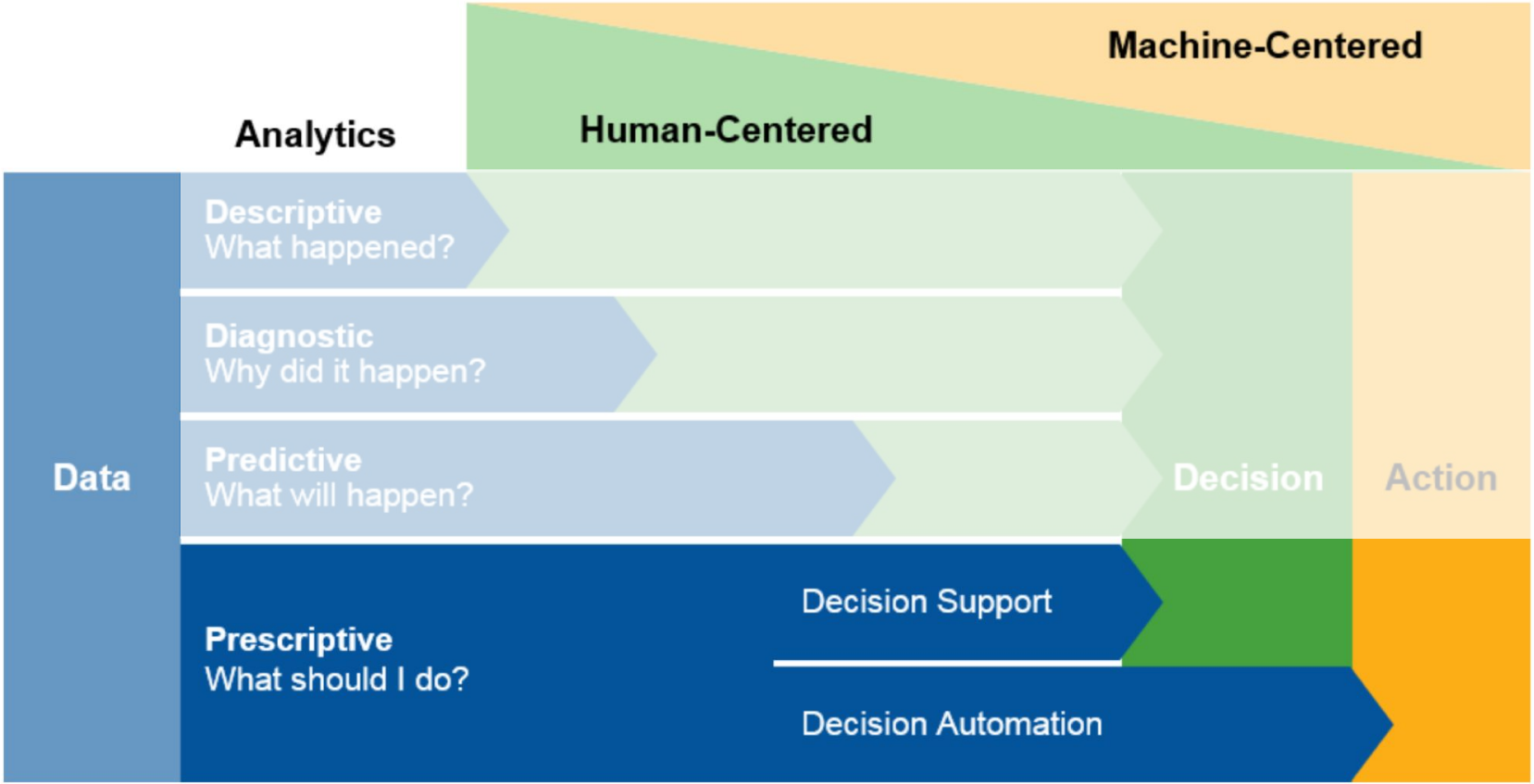
Source: Gartner (October 2016)



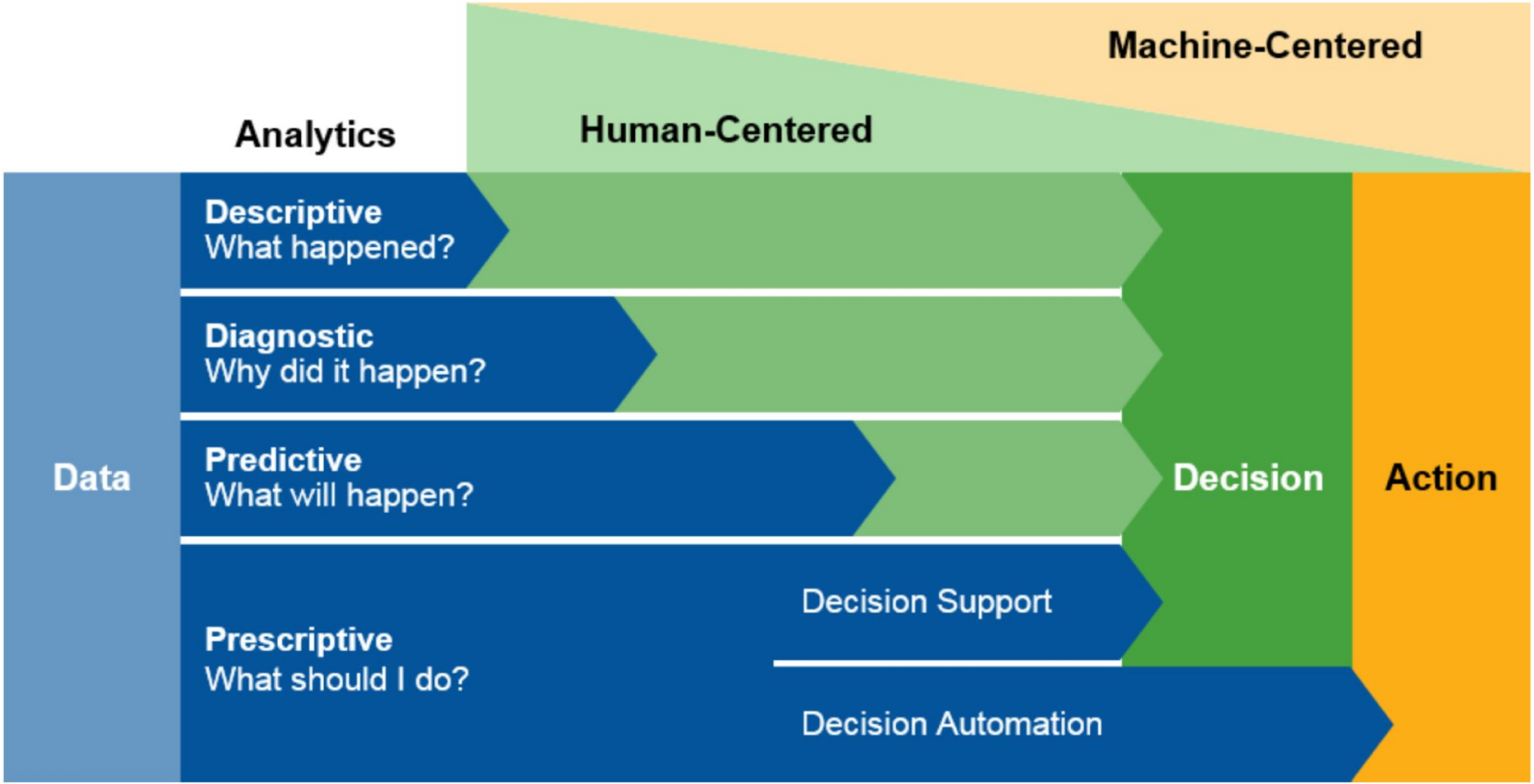
Source: Gartner (October 2016)



Source: Gartner (October 2016)

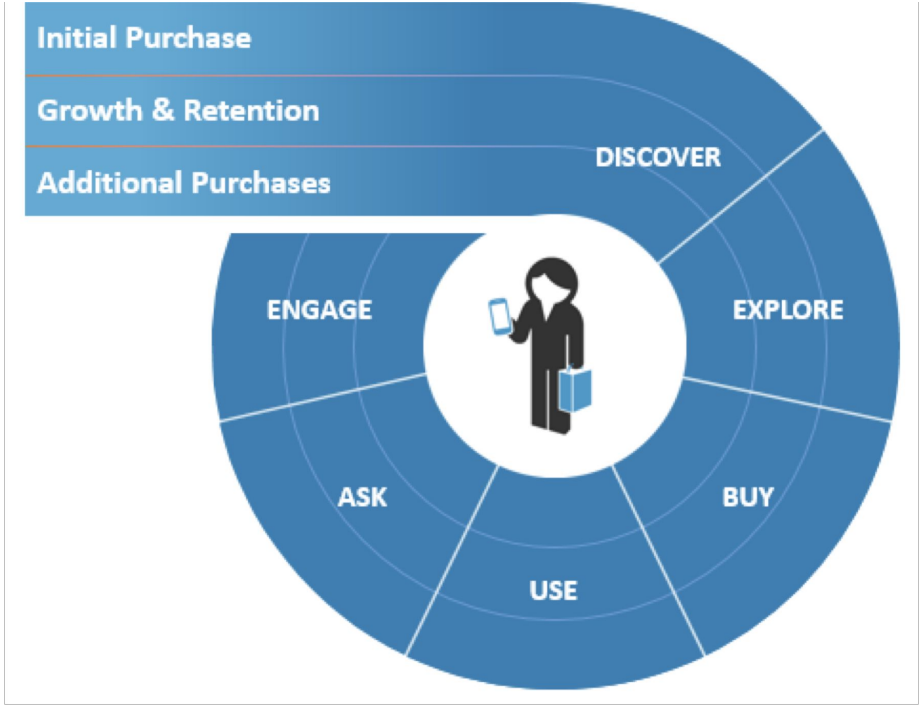


Source: Gartner (October 2016)

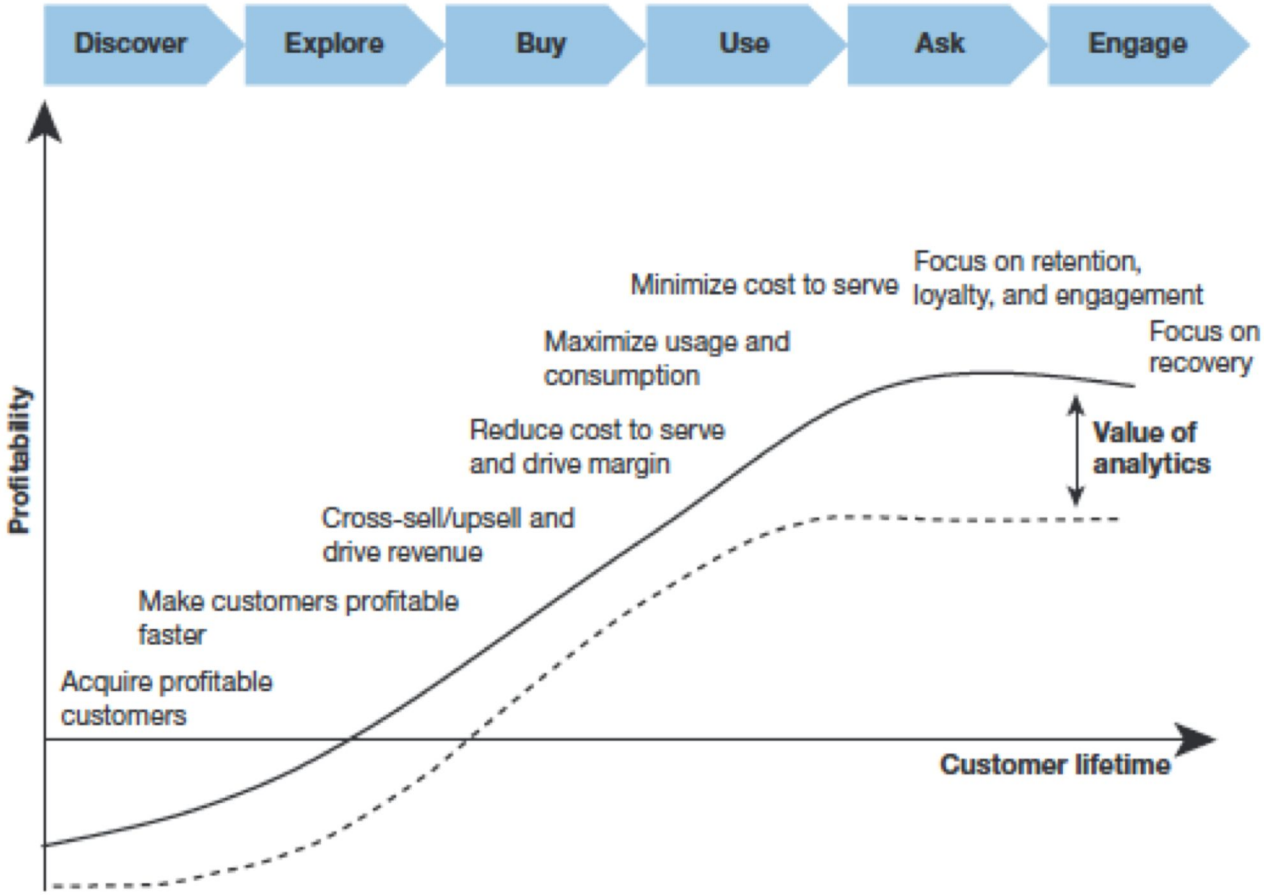


Source: Gartner (October 2016)

Analytics use cases



Source: Forrester (2016)



Source: Forrester (2016)

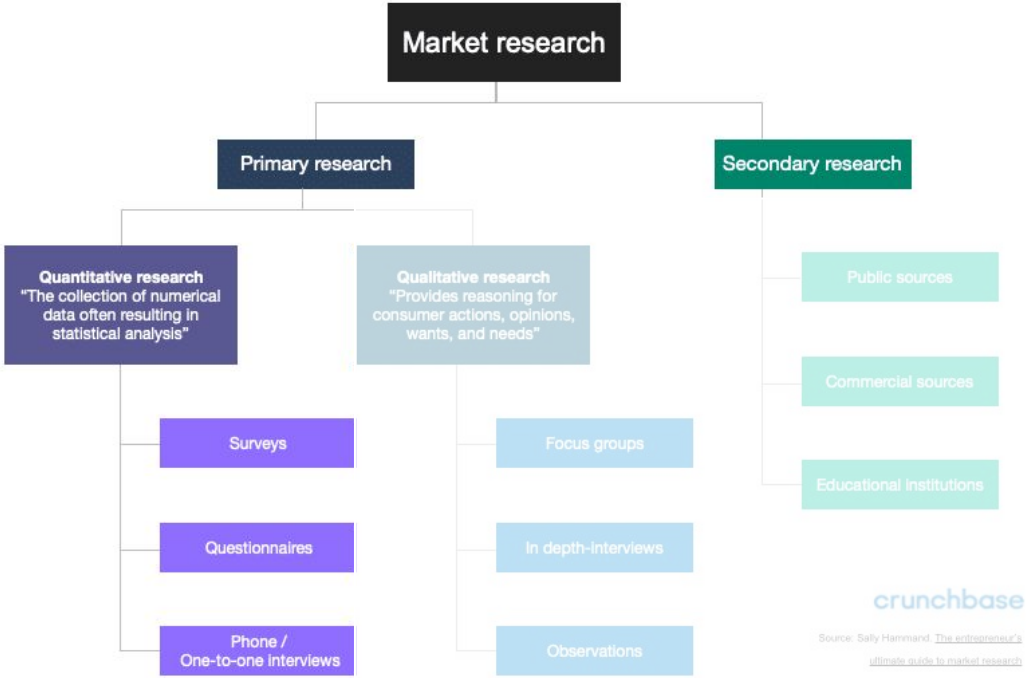
Life-cycle stage	Business objective	Analytical method
Discover	Profile customers	Segmentation
	Evaluate prospects	Lead scoring
	Reach right prospects	Customer look-alike targeting
Explore	Analyze customer response	Offer/contact optimization
	Optimize marketing mix	Marketing mix modeling
	Test marketing inputs	A/B and multivariate testing
Buy	Predict future events	Propensity models
	Expand wallet share	Cross-sell/upsell
	Target accurately	In-market timing models
Use	Drive deeper product use	Product and recommendation analysis
	Understand context behind usage	Sentiment analysis
Ask	Learn about drivers of engagement	Engagement analysis
	Understand customer satisfaction	Voice of the customer analysis
Engage	Manage defection of customers	Churn models
	Personalize marketing efforts	Next-best-action models
	Maximize customer value	Lifetime value models
	Add context to behavior	Customer location analysis
	Increase depth of relationship	Loyalty models

Source: Forrester (2016)

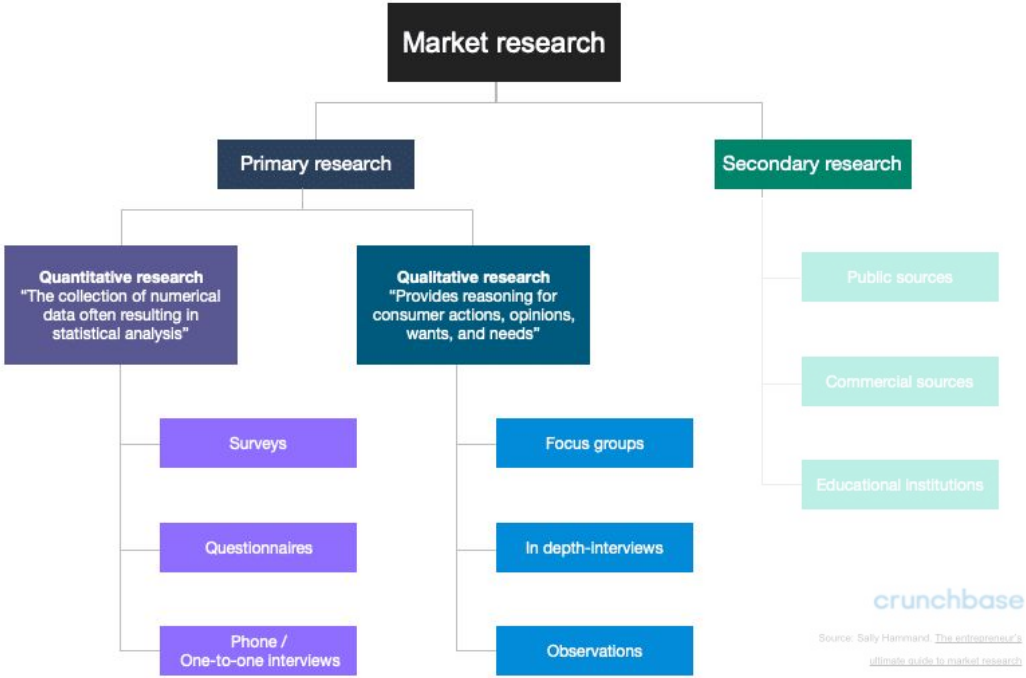
How do we obtain data?

Primary vs secondary research

Quantitative research



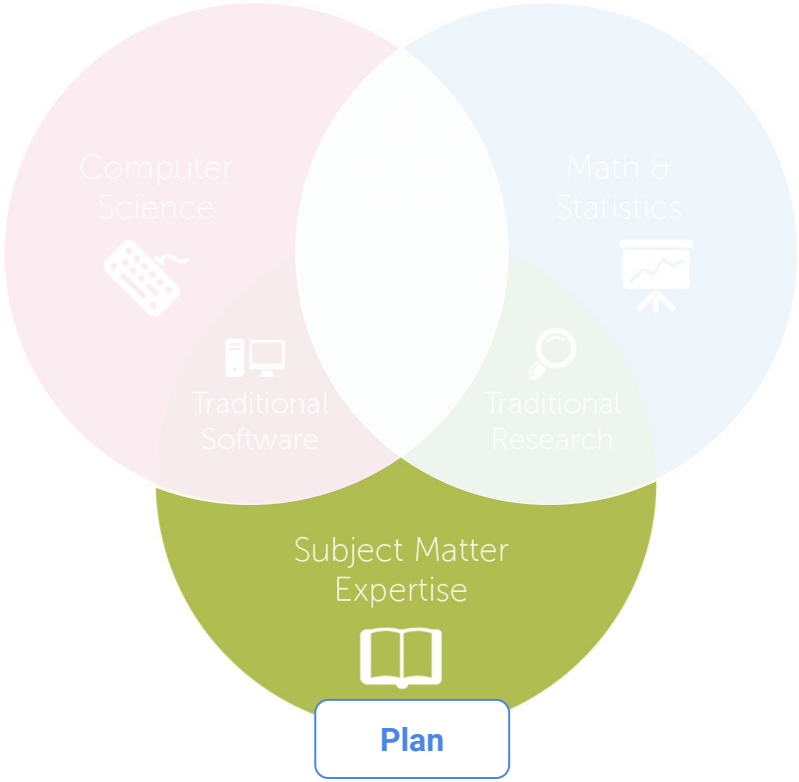
Qualitative research



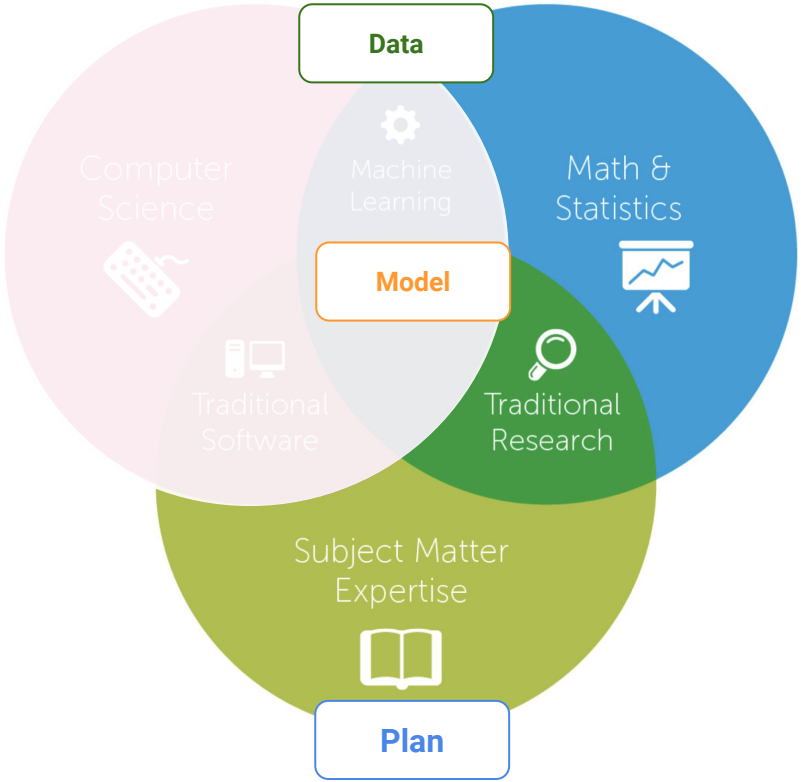
Types of market research

Data science

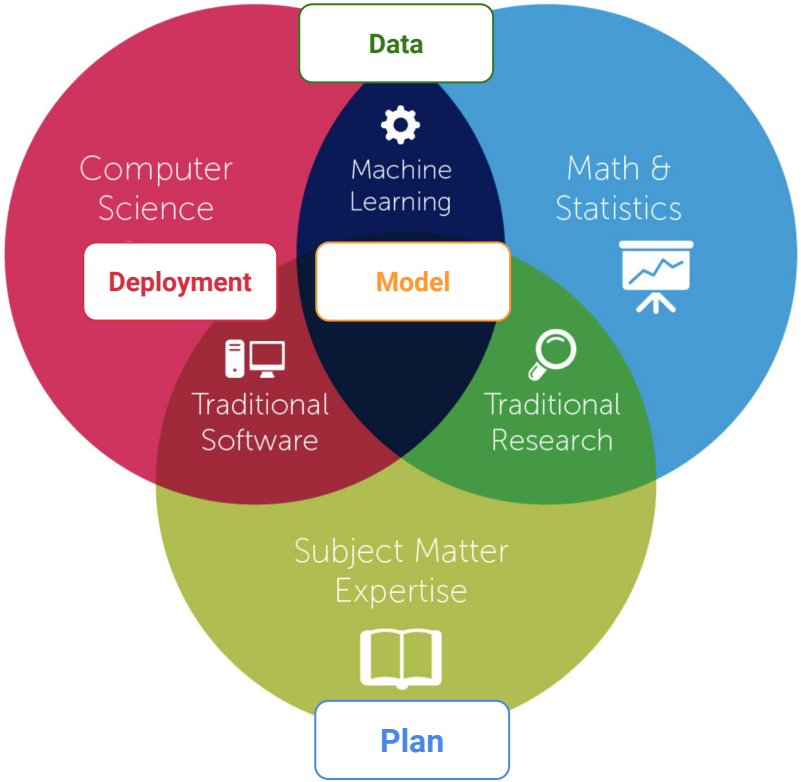
Data Science



Data Science

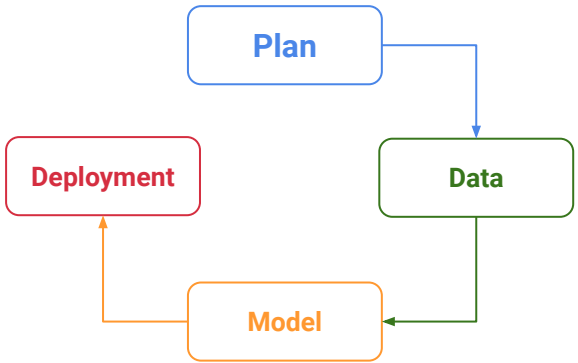


Data Science



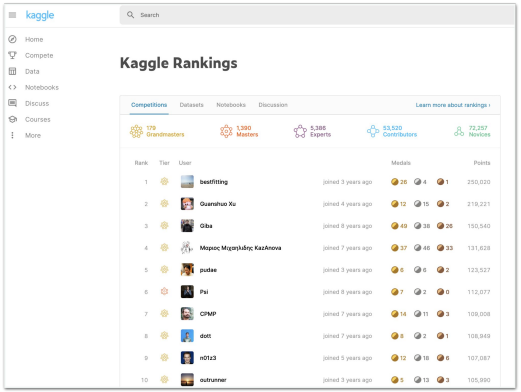
Data science is about ...

- Identifying promising use cases and creation of **plans**
- **Data** abilities
- Creation of **models**
- Understanding of **deployment** options

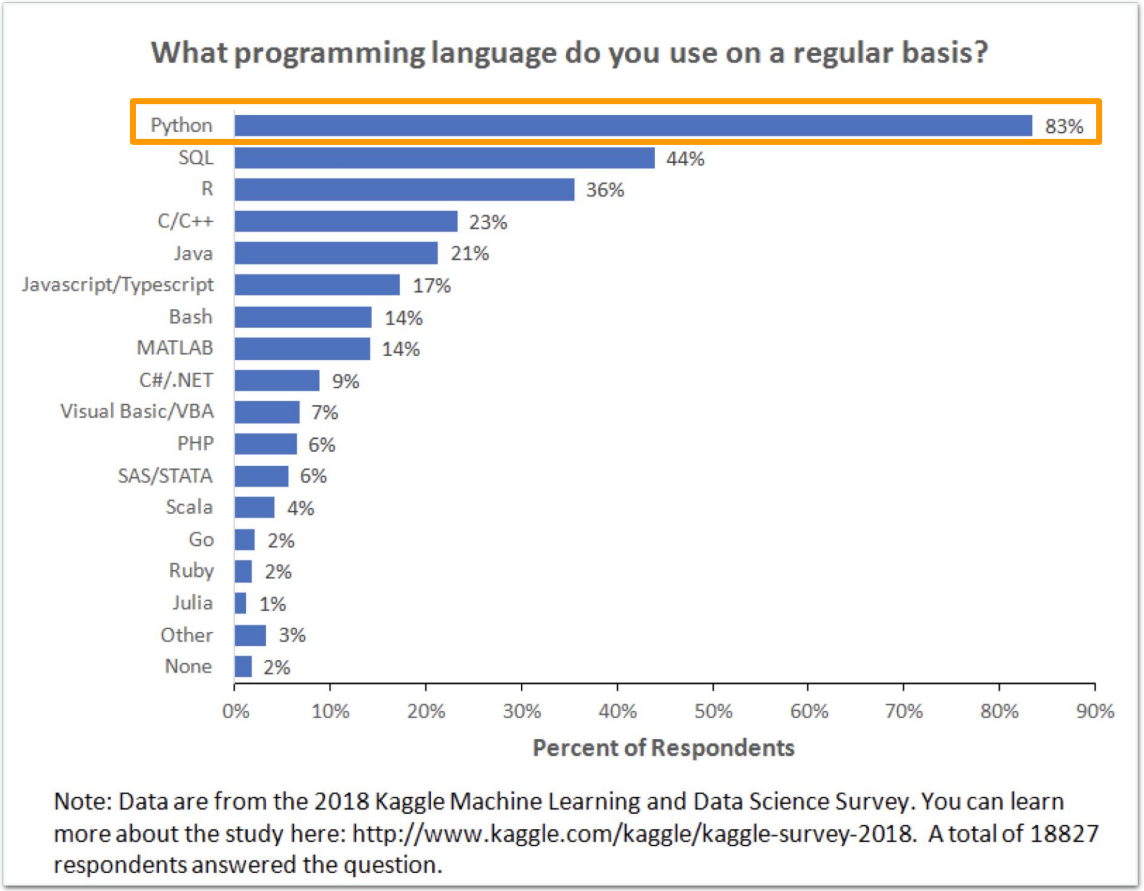


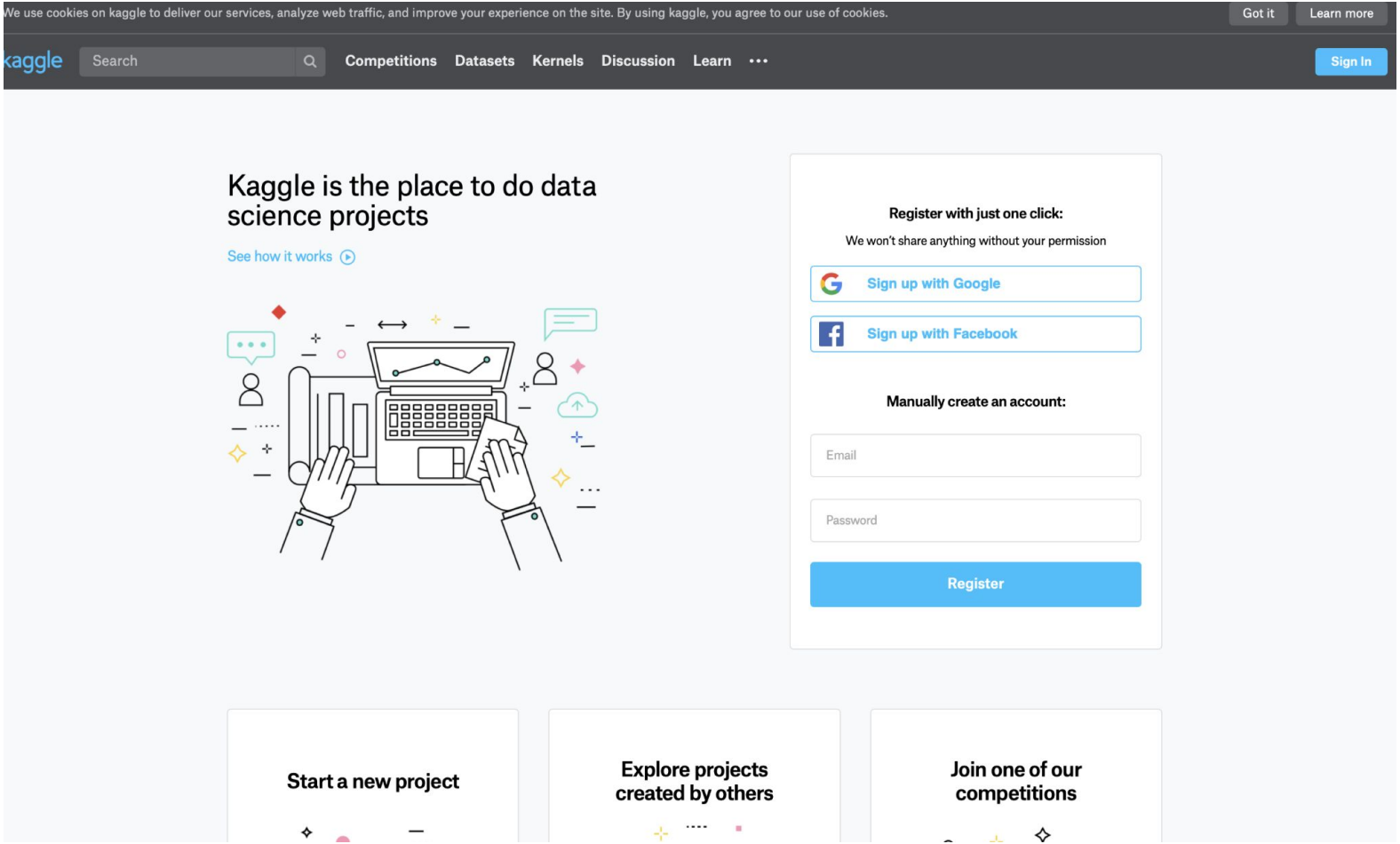
Computer science: Python

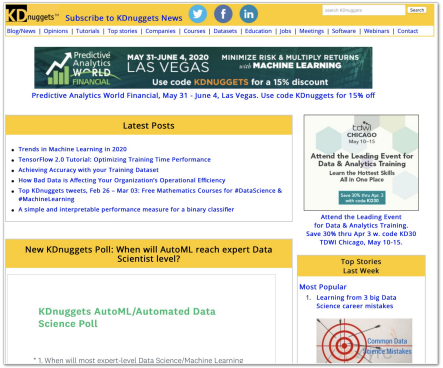
Prof. Dr. Jan Kirenz



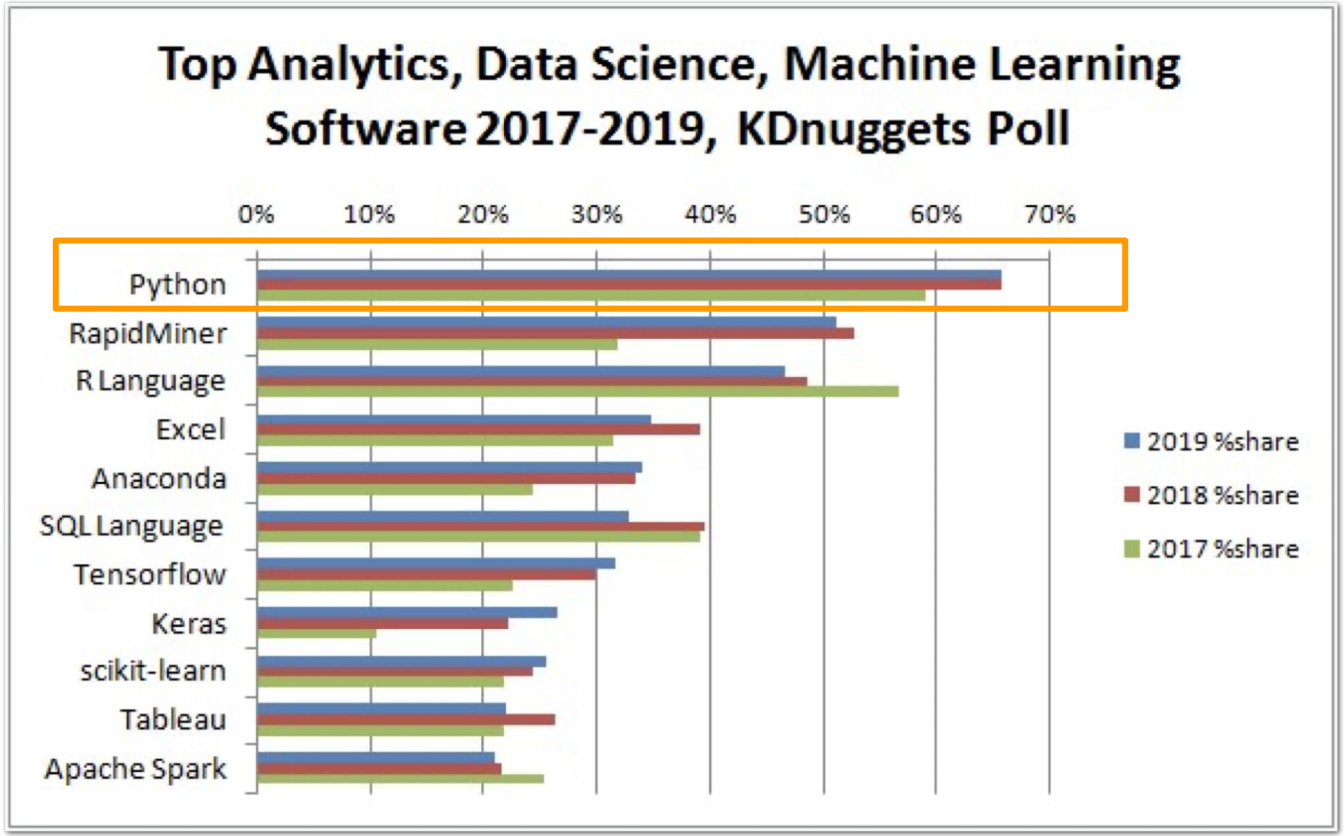
Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.




We use cookies on kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using kaggle, you agree to our use of cookies.
Got it
Learn more
kaggle
Search
Competitions
Datasets
Kernels
Discussion
Learn
Sign In
Kaggle is the place to do data science projects
See how it works
Register with just one click:
We won't share anything without your permission
Sign up with Google
Sign up with Facebook
Manually create an account:
Email
Password
Register
Start a new project
Explore projects created by others
Join one of our competitions



KDnuggets™ is a leading site on AI, Analytics, Big Data, Data Mining, Data Science, and Machine Learning and is edited by Gregory Piatetsky-Shapiro



Python 

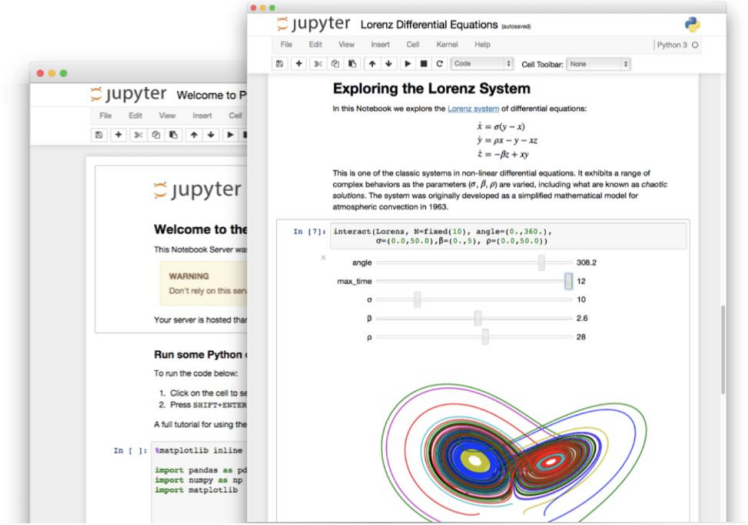
- Python is an **object-oriented** language (an object is an entity that contains data along with associated metadata and/or functionality).
- One thing that distinguishes Python from other programming languages is that it is **interpreted** rather than compiled.
- This means that it is executed **line by line** which is particular useful for data analysis, as well as the creation of interactive, executable documents (VanderPlas, 2016).
- On top of this, there is a broad ecosystem of **third-party tools and modules** that offer more specialized data science functionality (like Scikit-Learn, which provides a toolkit for applying machine learning algorithms to data).

Python 

- Python is an **object-oriented** language (an object is an entity that contains data along with associated metadata and/or functionality).
- One thing that distinguishes Python from other programming languages is that it is **interpreted** rather than compiled.
- This means that it is executed **line by line** which is particular useful for data analysis, as well as the creation of interactive, executable documents (VanderPlas, 2016).
- On top of this, there is a broad ecosystem of **third-party tools and modules** that offer more specialized data science functionality (like Scikit-Learn, which provides a toolkit for applying machine learning algorithms to data).

Python 

- Python is an **object-oriented** language (an object is an entity that contains data along with associated metadata and/or functionality).
- One thing that distinguishes Python from other programming languages is that it is **interpreted** rather than compiled.
- This means that it is executed **line by line** which is particular useful for data analysis, as well as the creation of interactive, executable documents (VanderPlas, 2016).
- On top of this, there is a broad ecosystem of **third-party tools and modules** that offer more specialized data science functionality (like Scikit-Learn, which provides a toolkit for applying machine learning algorithms to data).



The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Try it in your browser | Install the Notebook

Language of choice

The Notebook has support for over 40 programming languages, including Python, R, Julia, and Scala.

Share notebooks

Notebooks can be shared with others using email, Dropbox, GitHub and the Jupyter Notebook Viewer.

Interactive output

Your code can produce rich, interactive output: HTML, images, videos, LaTeX, and custom MIME types.

Big data integration

Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, TensorFlow.

Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Share

Table of contents

Introducing Colaboratory
Getting Started
More Resources
Machine Learning Examples: Seedbank
Section

+ Code + Text Copy to Drive

Connect Editing

Welcome to Colaboratory!

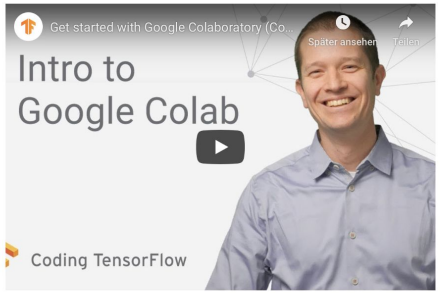
Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud.

With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.

[ ]

Introducing Colaboratory

This 3-minute video gives an overview of the key features of Colaboratory:



Getting Started

The document you are reading is a Jupyter notebook, hosted in Colaboratory. It is not a static page, but an interactive environment that lets you write and execute code in Python and other languages.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ] seconds_in_a_day = 24 * 60 * 60
    seconds_in_a_day
```

```
86400
```

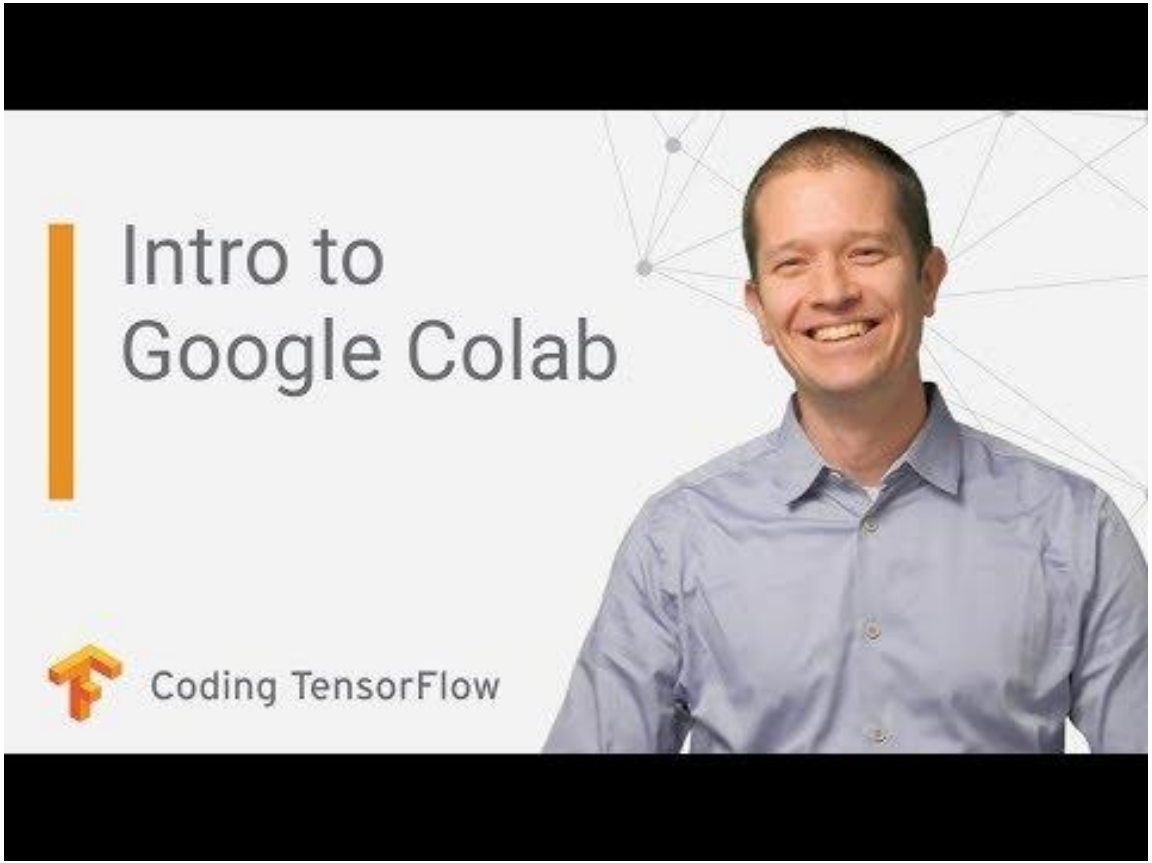
To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter".

All cells modify the same global state, so variables that you define by executing a cell can be used in other cells:

```
[ ] seconds_in_a_week = 7 * seconds_in_a_day
    seconds_in_a_week
```

```
604800
```

For more information about working with Colaboratory notebooks, see Overview of Colaboratory.


Intro to
Google Colab
Coding TensorFlow

Python 

- Python is an **object-oriented** language (an object is an entity that contains data along with associated metadata and/or functionality).
- One thing that distinguishes Python from other programming languages is that it is **interpreted** rather than compiled.
- This means that it is executed **line by line** which is particular useful for data analysis, as well as the creation of interactive, executable documents (VanderPlas, 2016).
- On top of this, there is a broad ecosystem of **third-party tools and modules** that offer more specialized data science functionality (like Scikit-Learn, which provides a toolkit for applying machine learning algorithms to data).

Install User Guide API Examples More Go

scikit-learn

Machine Learning in Python

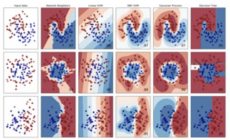
Getting Started | What's New in 0.22.2 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

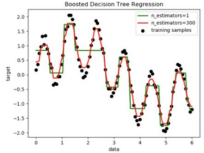


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

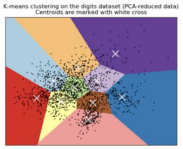


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

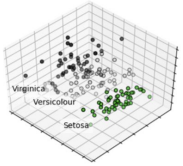


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

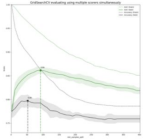
Applications: Visualization, Increased efficiency
Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...



Model selection

Comparing, validating and choosing parameters and models.

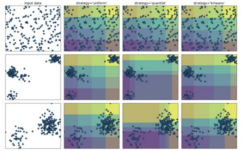
Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...

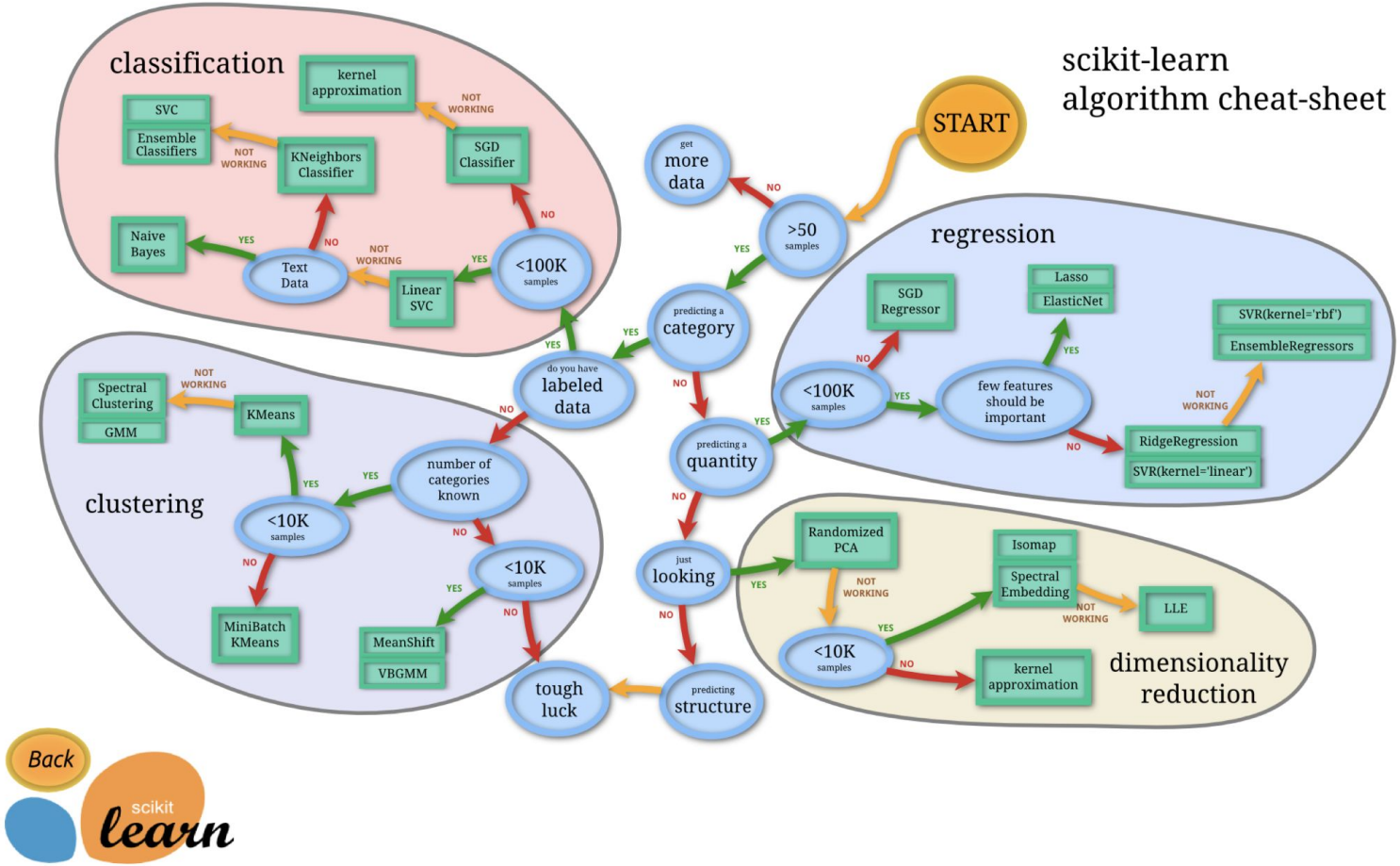


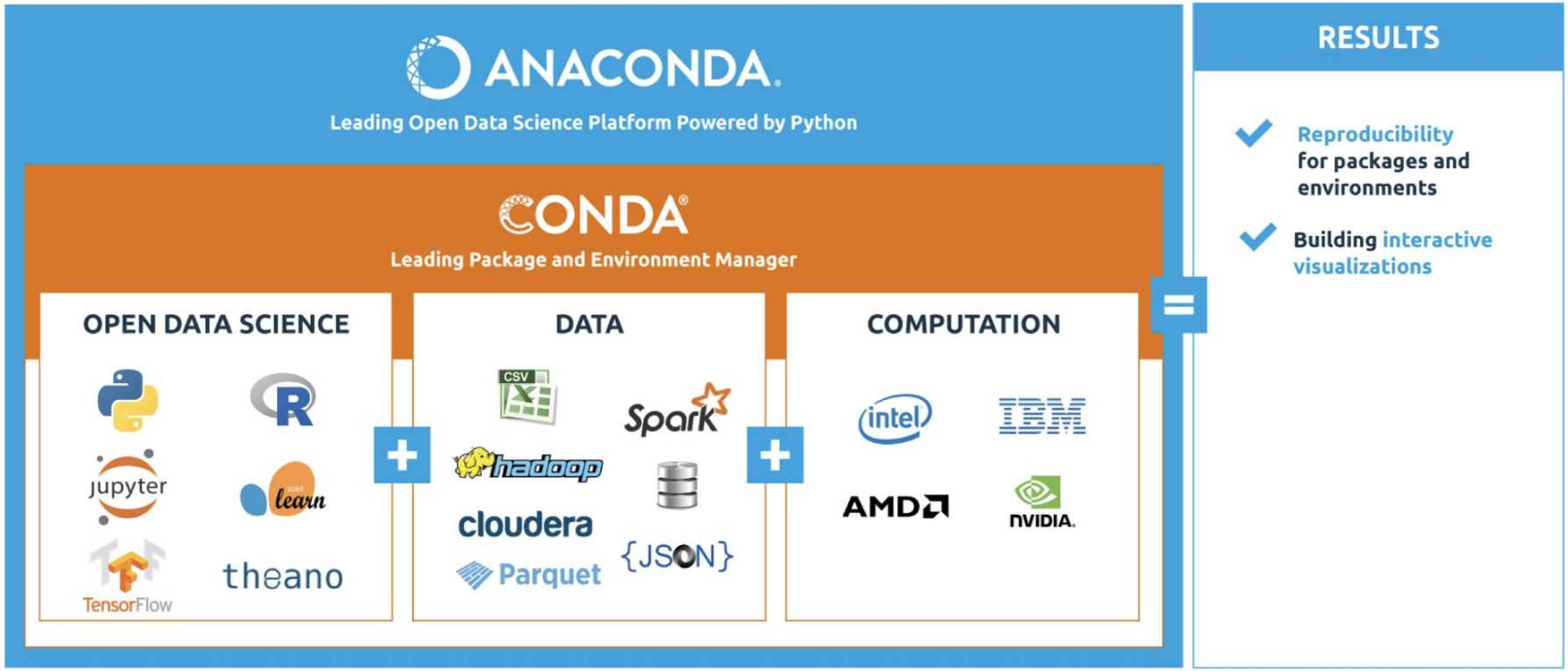
Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...




scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
NO
do you have labeled data
YES
<100K samples
NO
YES
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
NOT WORKING
Linear SVC
YES
classification
predicting a quantity
YES
NO
<100K samples
NO
YES
SGD Regressor
Lasso
ElasticNet
YES
few features should be important
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
<10K samples
YES
NO
kernel approximation
dimensionality reduction
predicting structure
NO
tough luck
number of categories known
YES
NO
<10K samples
YES
NO
KMeans
YES
Spectral Clustering
GMM
NOT WORKING
MiniBatch KMeans
<10K samples
NO
YES
MeanShift
VBGMM
clustering
Back
scikit learn


ANACONDA.
Leading Open Data Science Platform Powered by Python
CONDA
Leading Package and Environment Manager
OPEN DATA SCIENCE
DATA
COMPUTATION
jupyter
learn
theano
TensorFlow
CSV
Spark
hadoop
cloudera
Parquet
{JSON}
intel
IBM
AMD
NVIDIA.
RESULTS
Reproducibility for packages and environments
Building interactive visualizations

Python 

- Python is an **object-oriented** language (an object is an entity that contains data along with associated metadata and/or functionality).
- One thing that distinguishes Python from other programming languages is that it is **interpreted** rather than compiled.
- This means that it is executed **line by line** which is particular useful for data analysis, as well as the creation of interactive, executable documents (VanderPlas, 2016).
- On top of this, there is a broad ecosystem of **third-party tools and modules** that offer more specialized data science functionality (like Scikit-Learn, which provides a toolkit for applying machine learning algorithms to data).

Git & GitHub

Prof. Dr. Jan Kirenz

Git is an open source distributed version control system

- Git can be used to **store content**
- Code can be changed and other developers can **add code** in **parallel**.
- Git has a **remote repository** which is stored in a server and a **local repository** which is stored in the computer of each developer.

Git is an open source distributed version control system

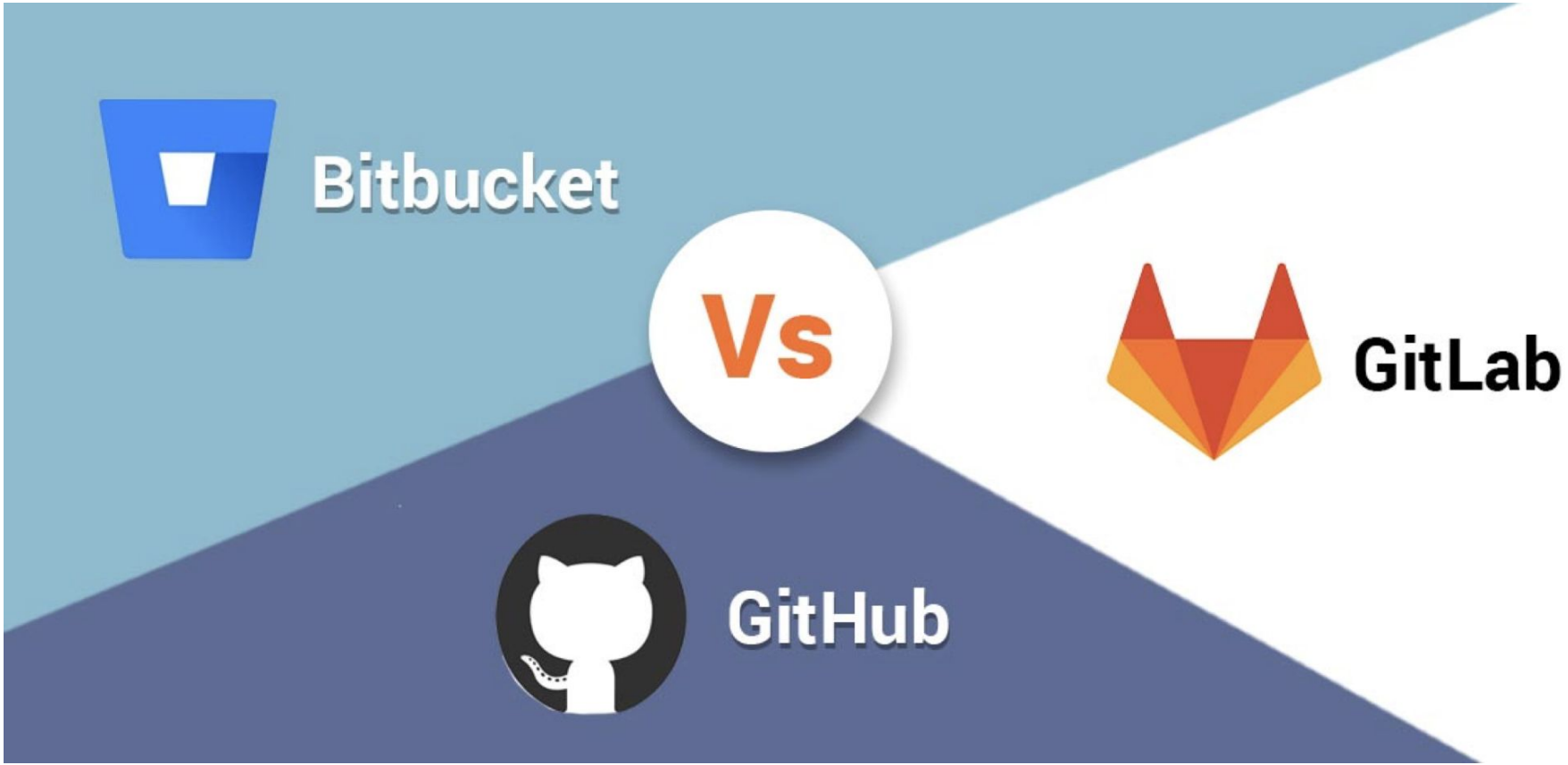
- Git can be used to **store content**
- Code can be changed and other developers can **add code** in **parallel**.
- Git has a **remote repository** which is stored in a server and a **local repository** which is stored in the computer of each developer.

Git is an open source distributed version control system

- Git can be used to **store content**
- Code can be changed and other developers can **add code** in **parallel**.
- Git has a **remote repository** which is stored in a server and a **local repository** which is stored in the computer of each developer.

Git is an open source distributed version control system

- Git can be used to **store content**
- Code can be changed and other developers can **add code** in **parallel**.
- Git has a **remote repository** which is stored in a server and a **local repository** which is stored in the computer of each developer.


Bitbucket
Vs
GitLab
GitHub

kirenz / first_steps_in_python

Watch 2 | Star 1 | Fork 6

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: master ▾ first_steps_in_python / 5_data_science_programming_process.ipynb

Find file | Copy path

kirenz Add files via upload 4759298 on 16 Mar

1 contributor

2025 lines (2024 sloc) | 372 KB

Raw | Blame | History

Introduction to Data Science with Python

Prof. Dr. Jan Kirenz
Hochschule der Medien Stuttgart

In [1]:
```
import numpy as np
import pandas as pd
from pandas.api.types import CategoricalDtype
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('ggplot') # define plot style
import seaborn as sns
sns.set() # seaborne standard settings
from IPython.display import Image # display image in the frontend
```

Module overview:

- **NumPy** provides efficient storage and computation for multidimensional data arrays.
- **Pandas** provides a DataFrame object along with a powerful set of methods to manipulate, filter, group, and transform data. statistical models, as well as for conducting statistical tests, and statistical data exploration. Furthermore, you can use R-style formulas together with pandas data frames to fit your models.
- **Matplotlib** provides a useful interface for creation of publication-quality plots and figures.
- **Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Data Science