

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."



"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."



"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."



"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."



"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."



"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."

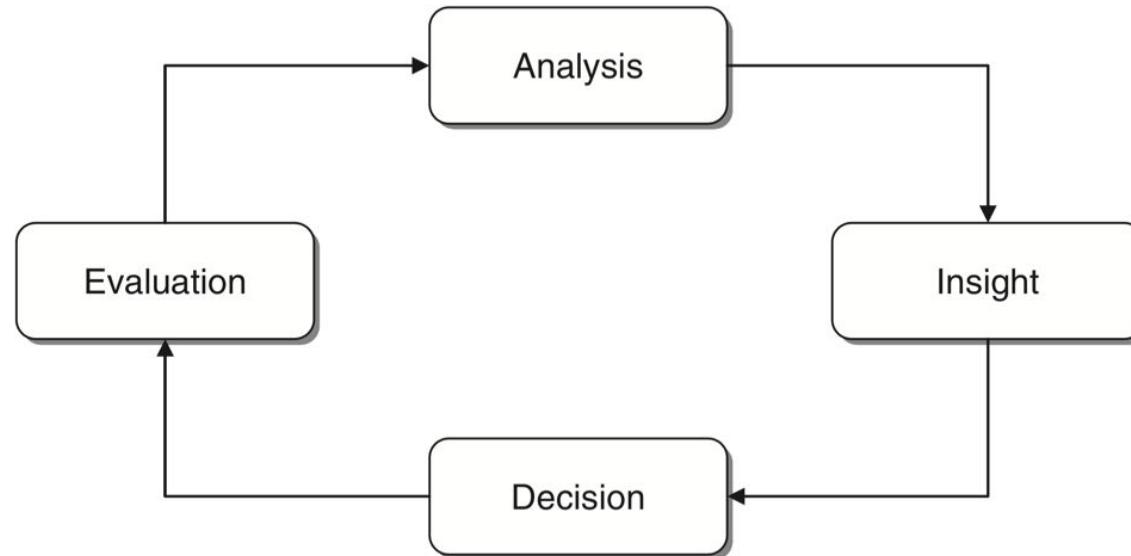


"The ability to –
take data,
be able to **understand** it,
process it,
extract value from it,
visualize it,
communicate it –
that's going to be a hugely important
skill in the next decades."



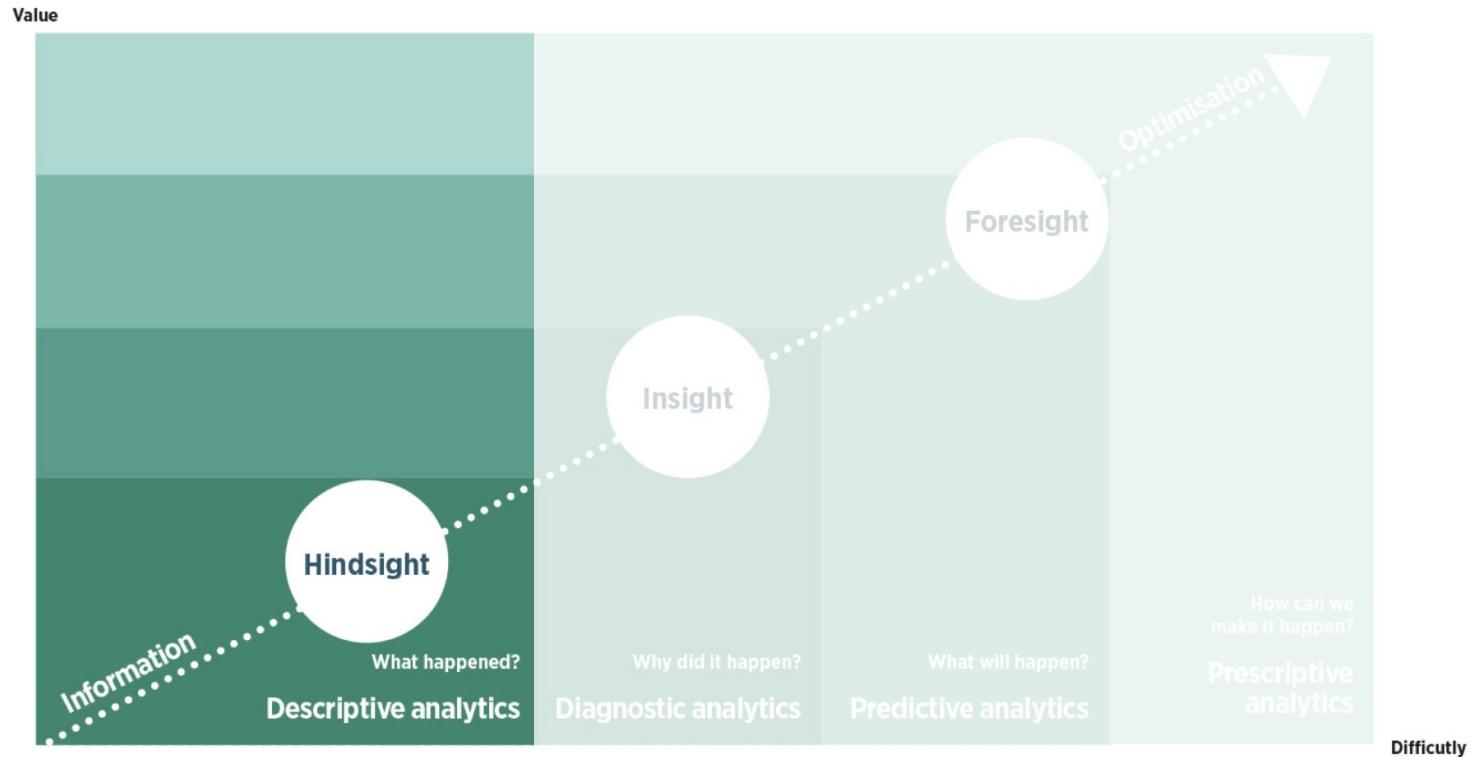
Data-driven decision making is the process of **making decisions** based on actual **data** rather than intuition or personal experience alone.

Cycle of data driven **decision making**



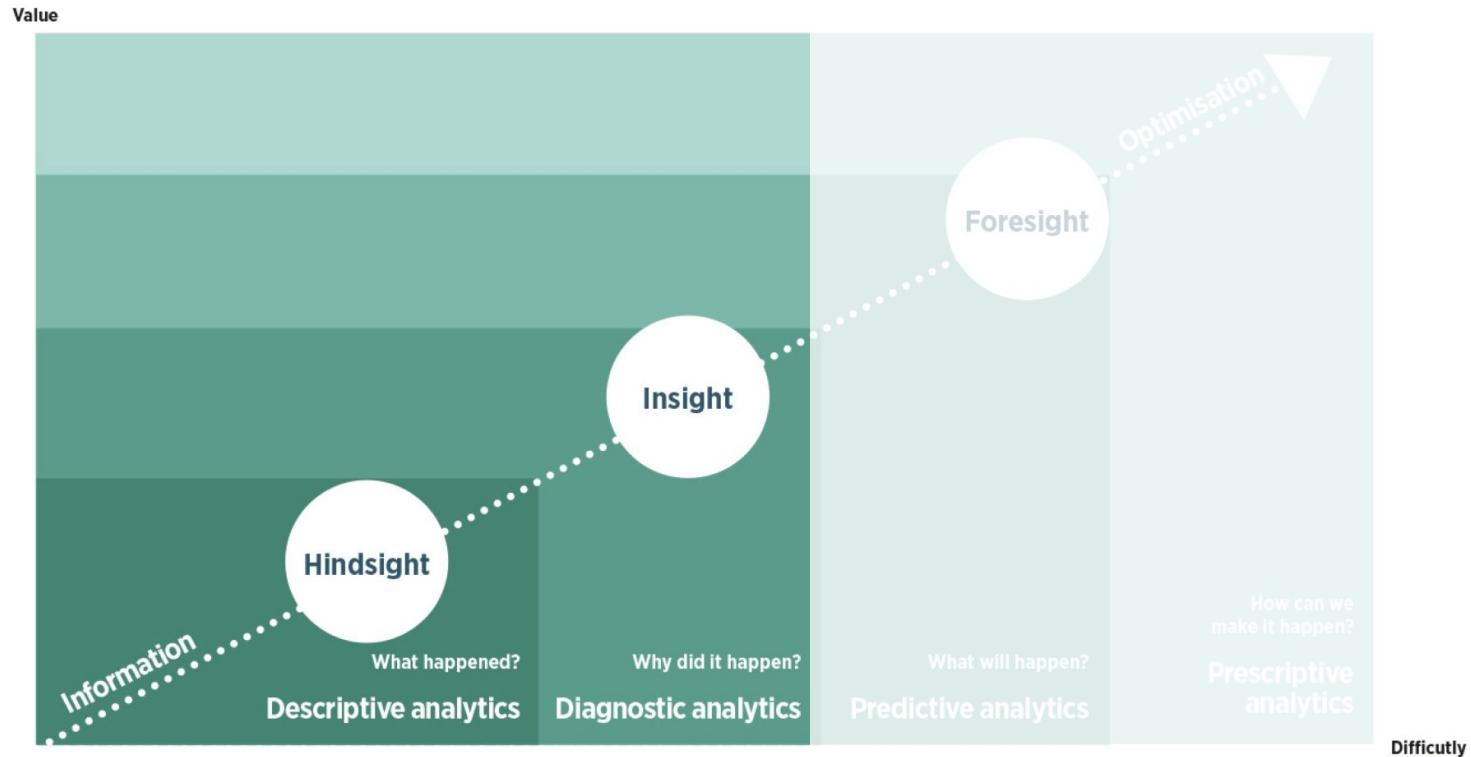
Types of analytics

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS



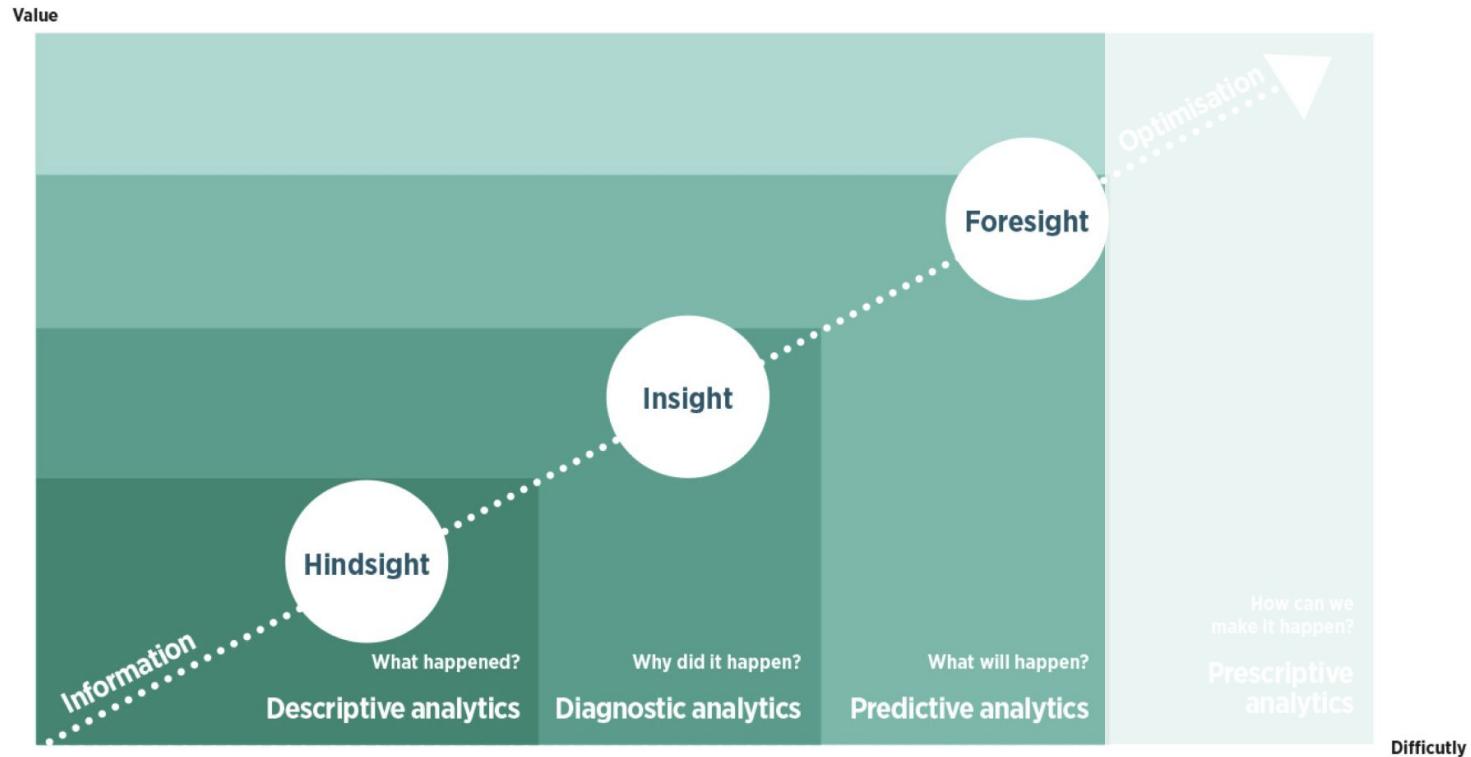
Source: Gartner

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS



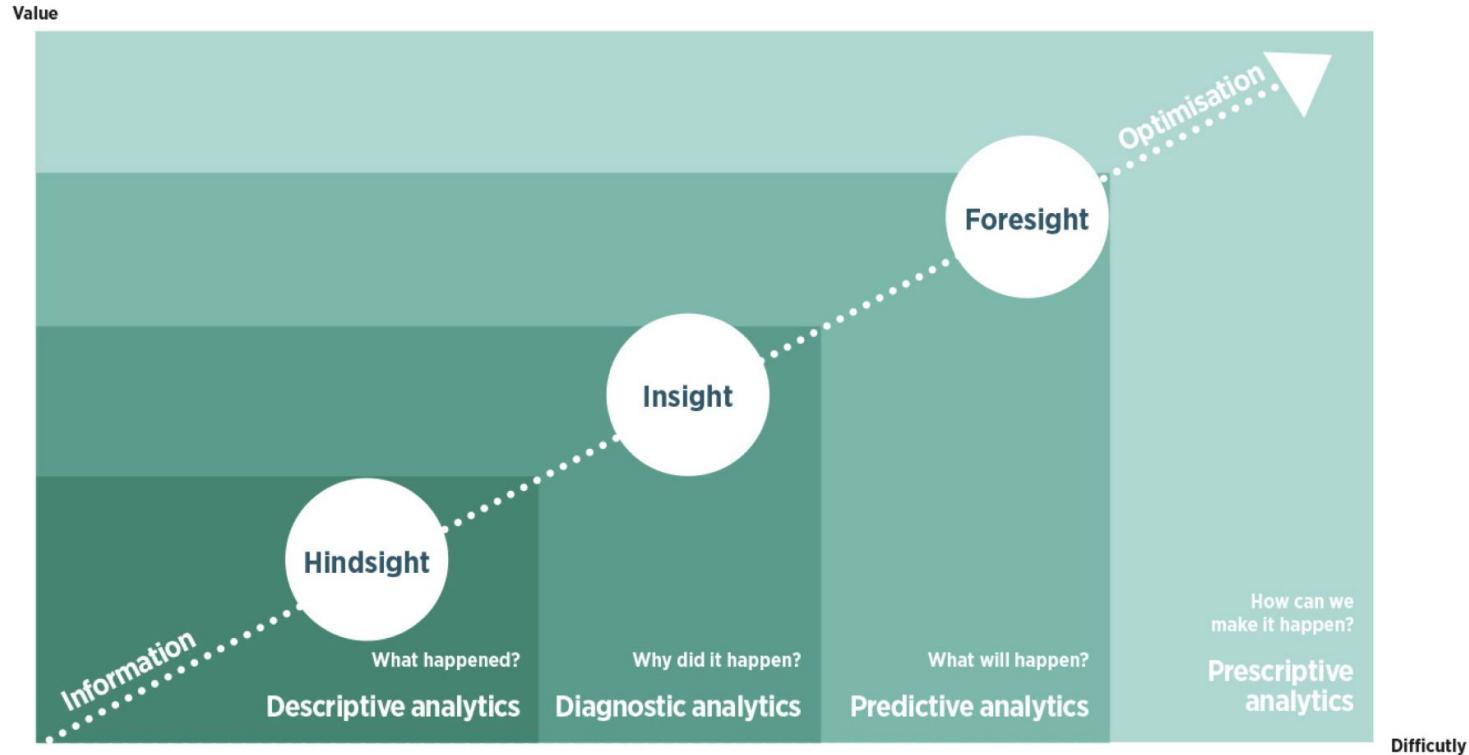
Source: Gartner

MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS



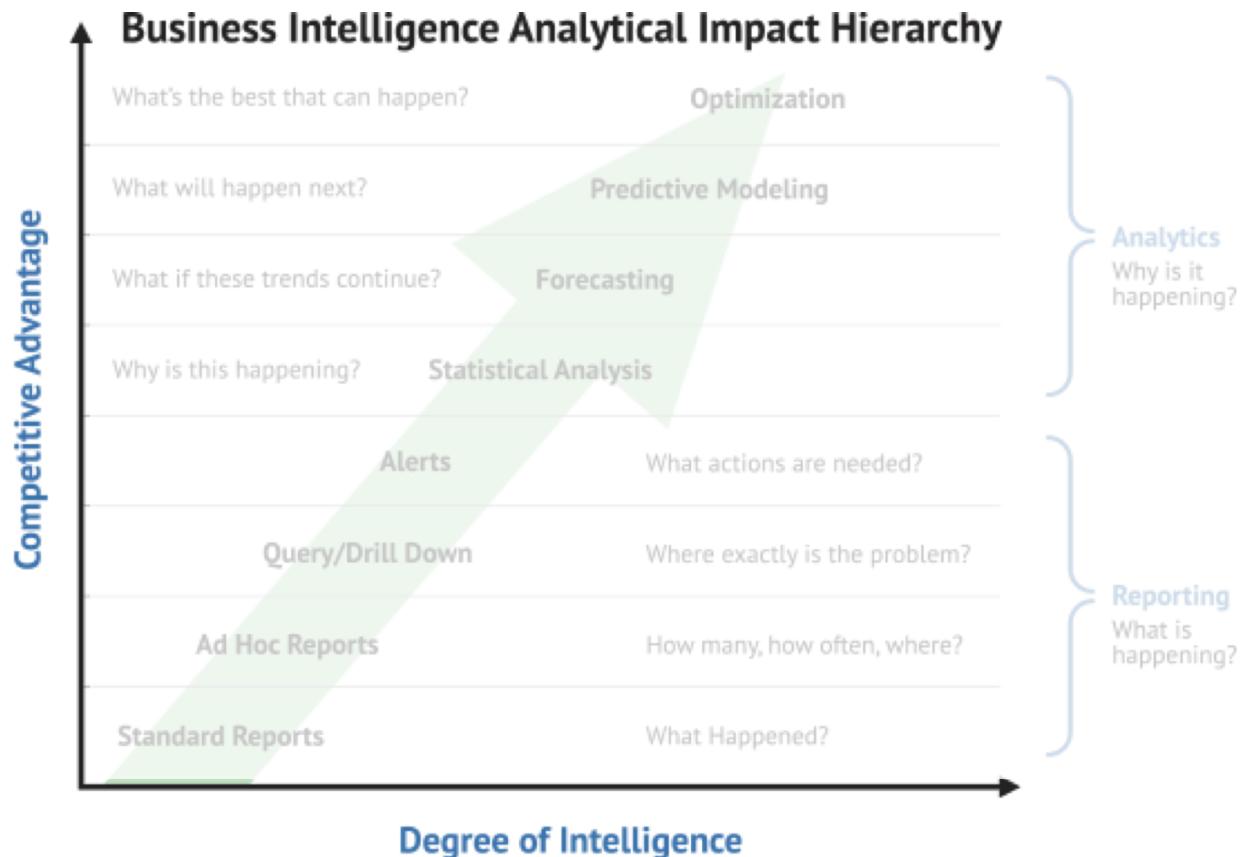
Source: Gartner

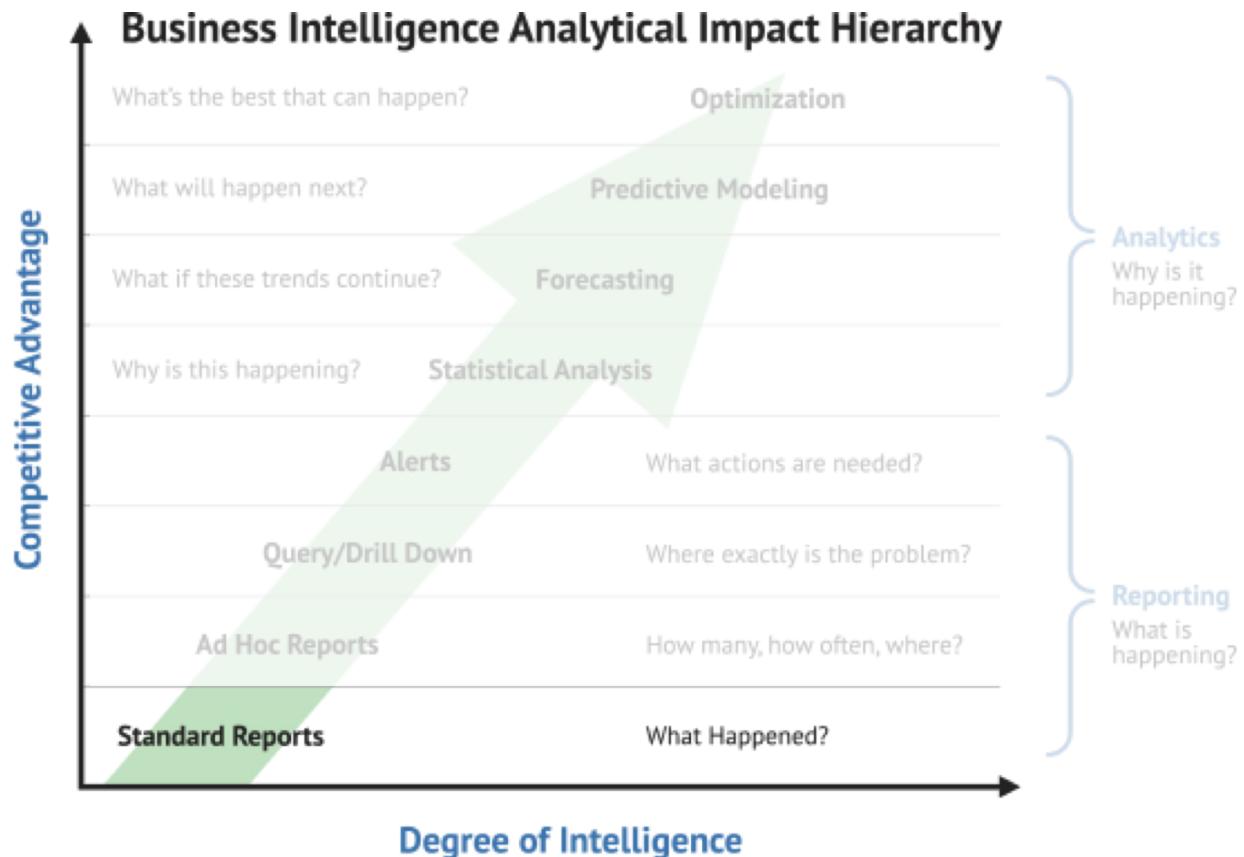
MEASURING THE DIFFICULTY AND VALUE OF ANALYTICS

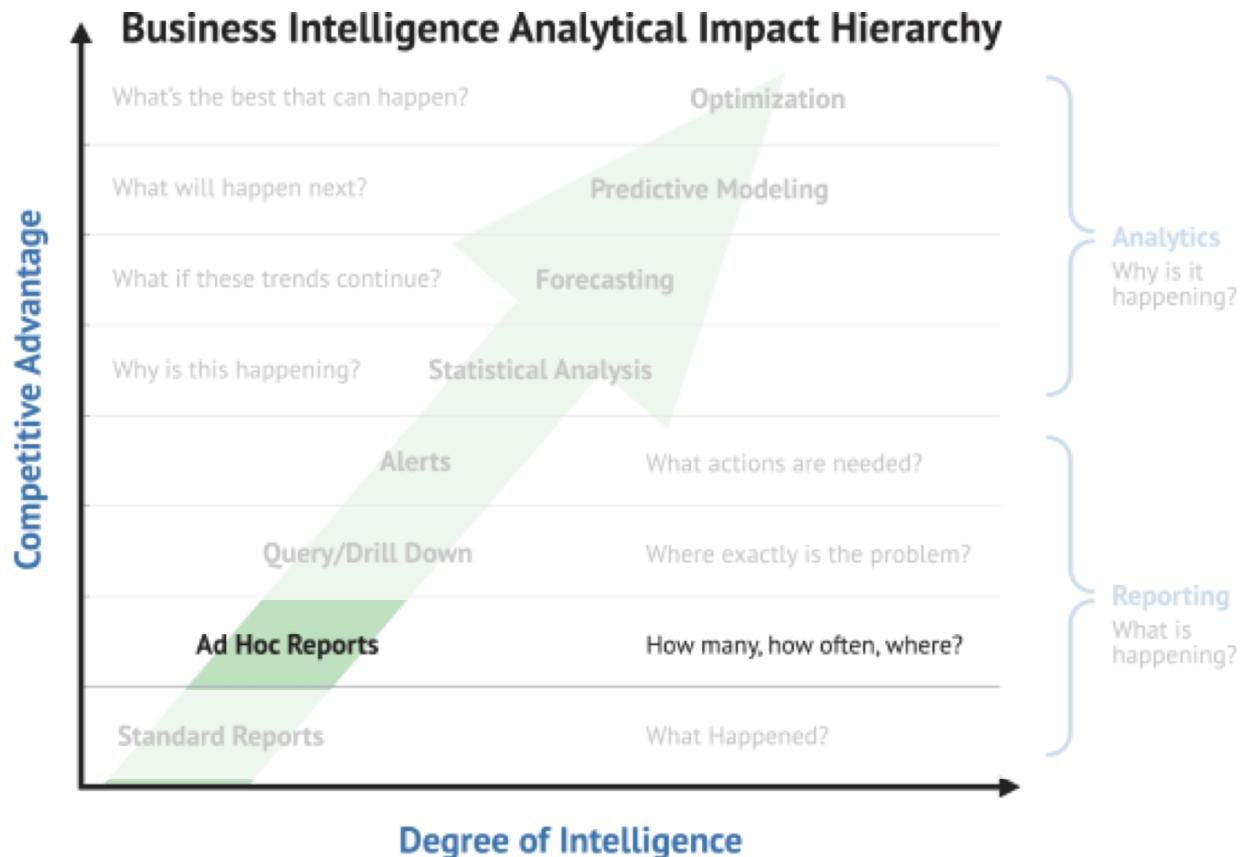


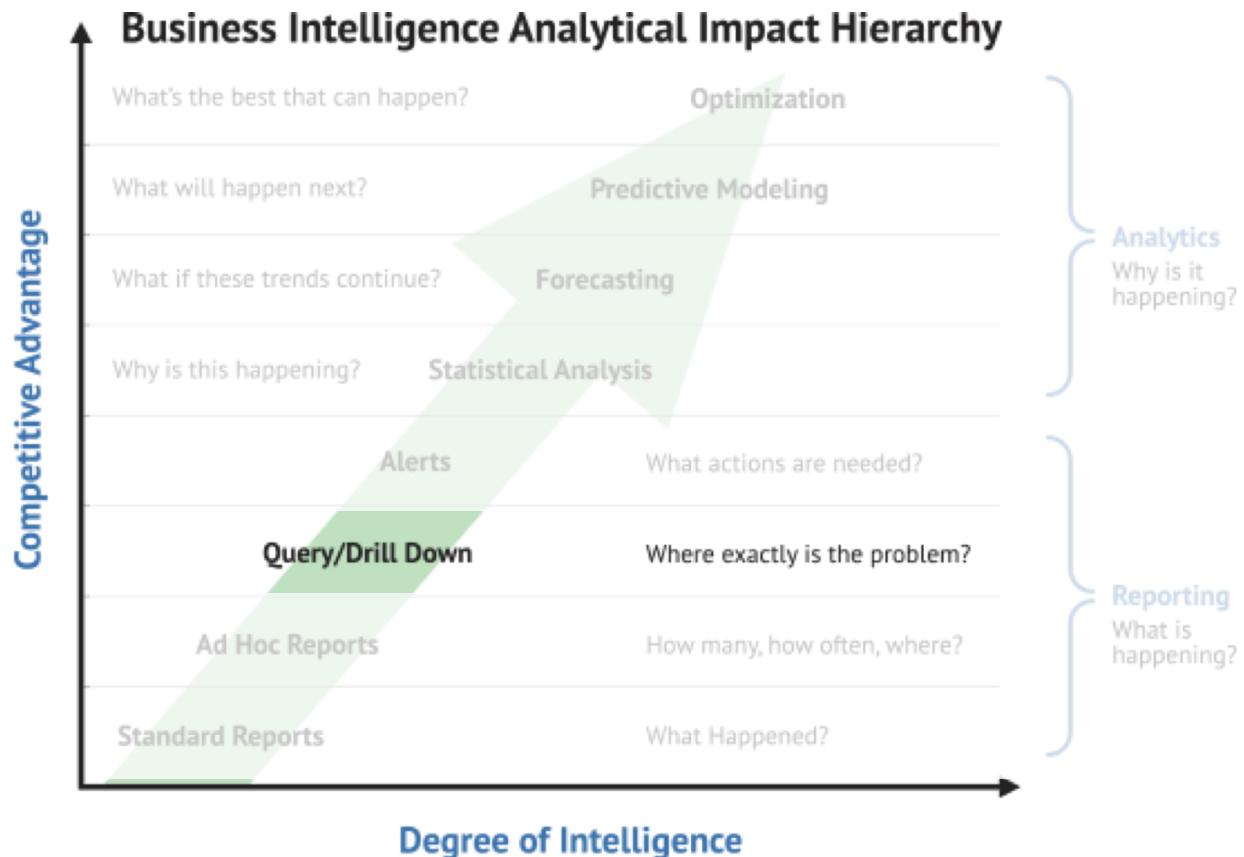
Source: Gartner

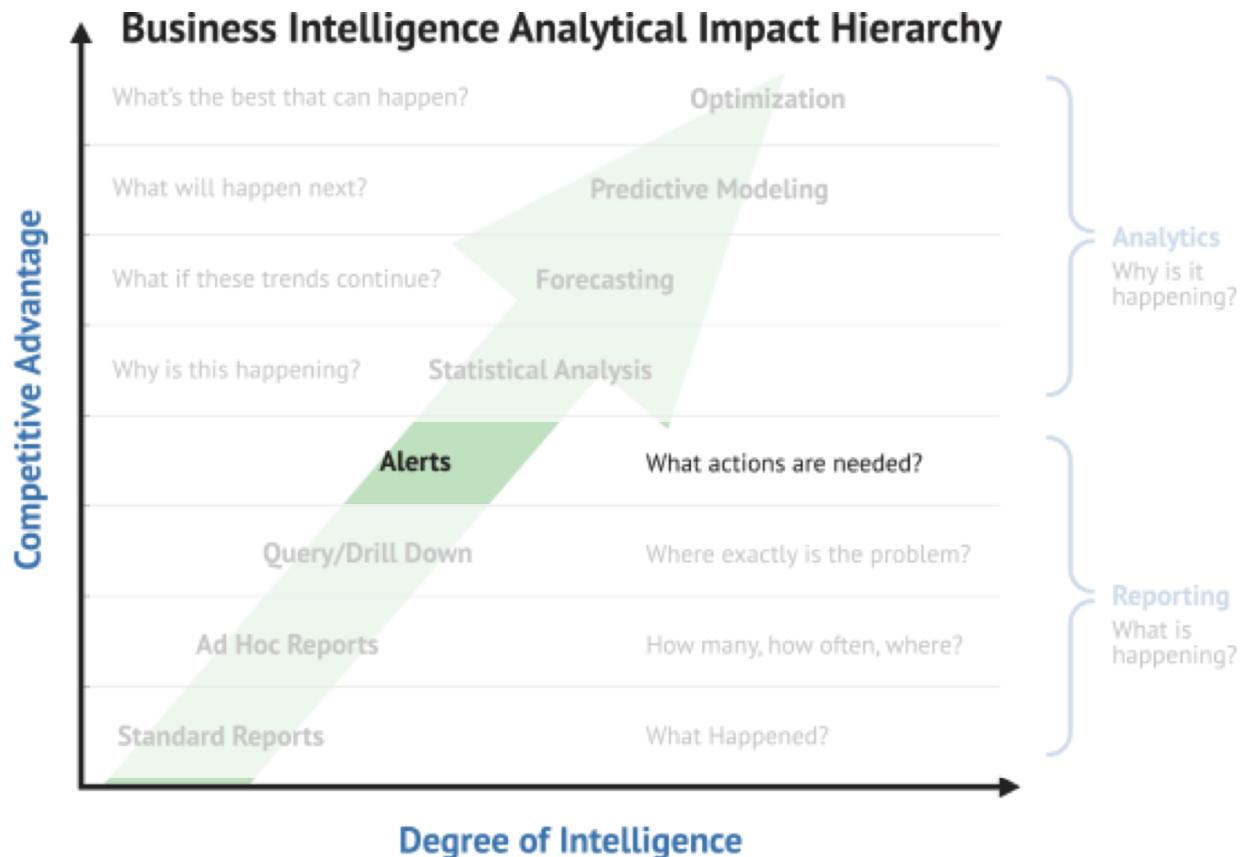
Business
intelligence
analytical impact
hierarchy

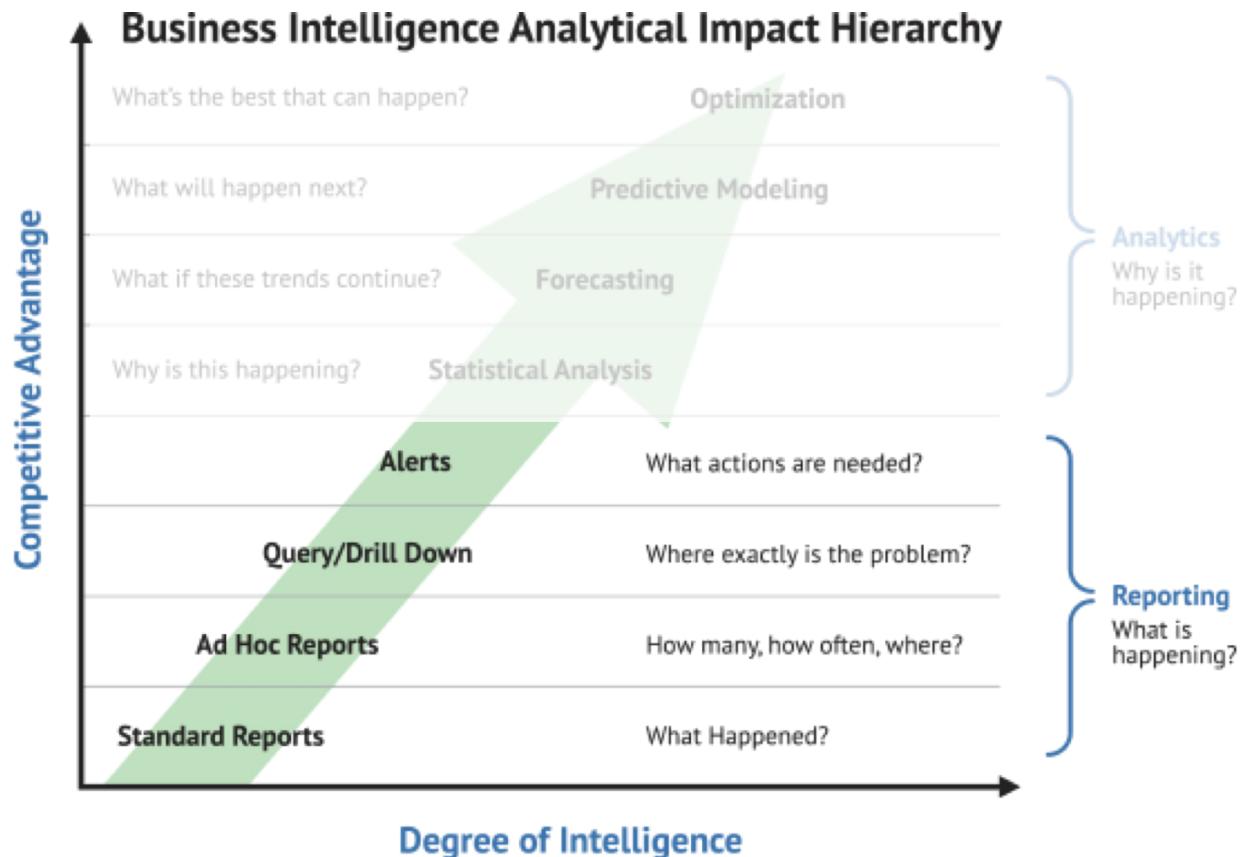


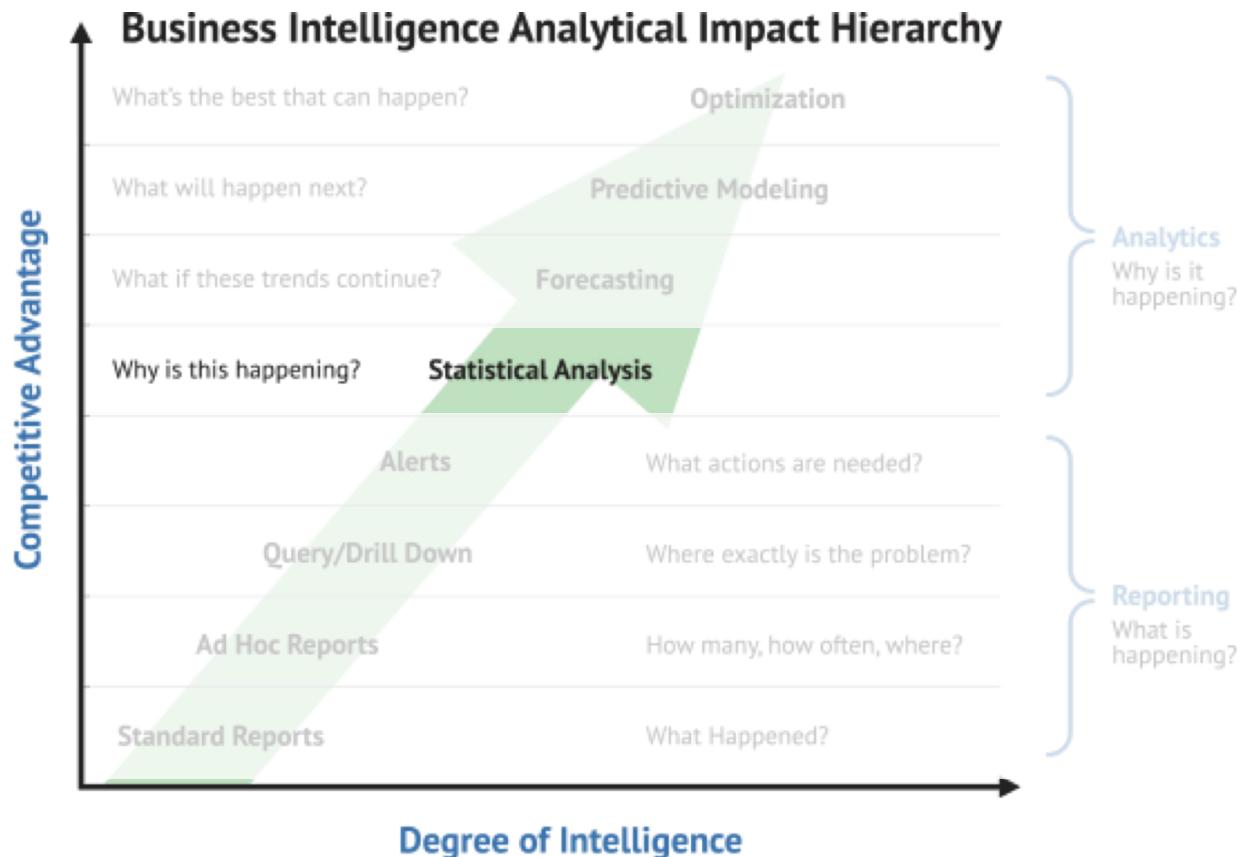


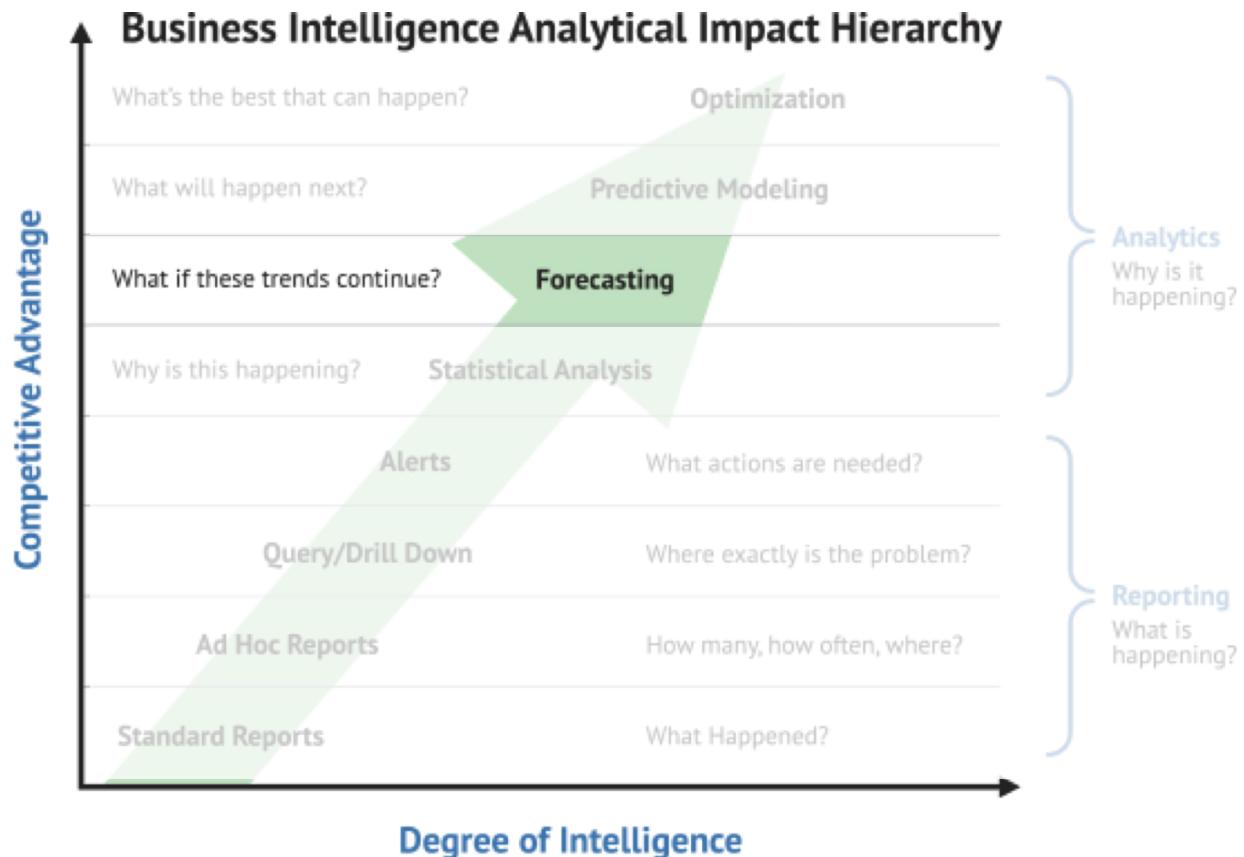


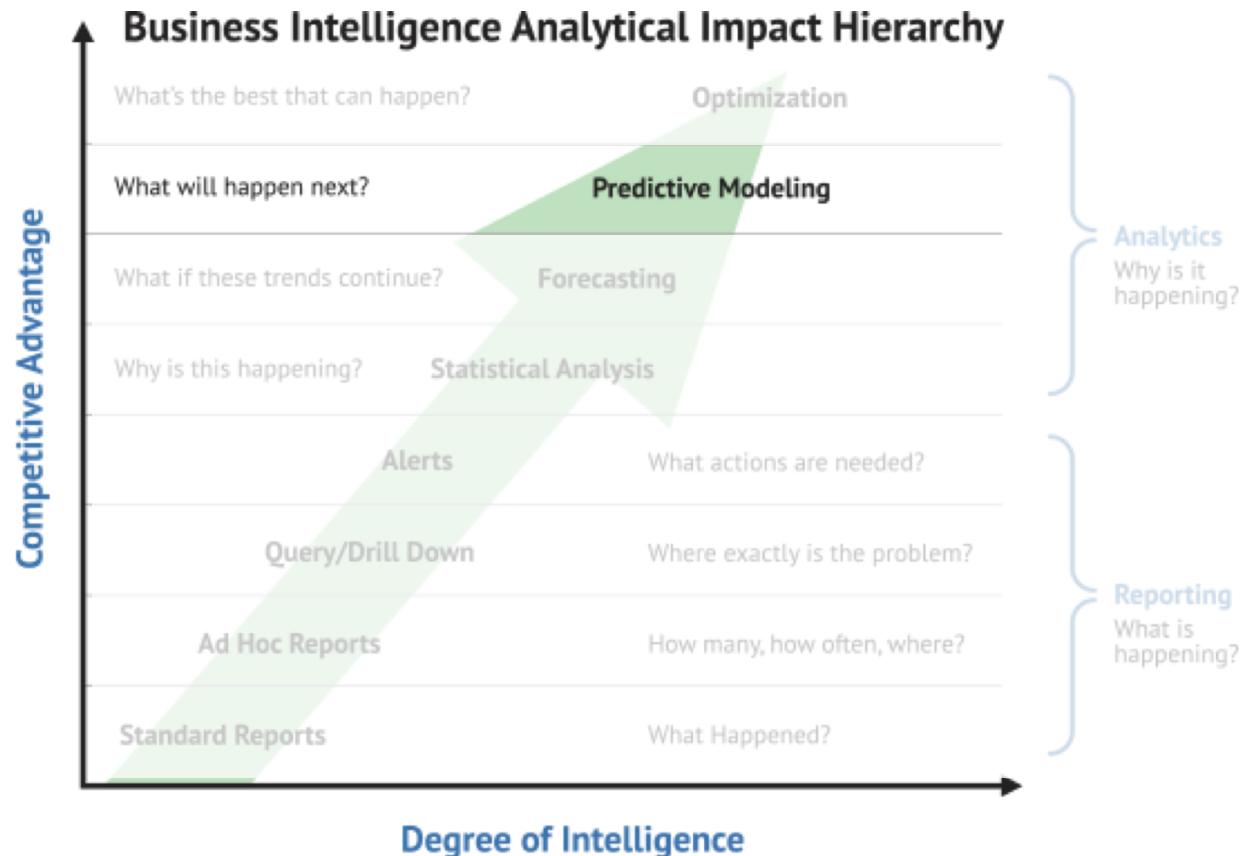


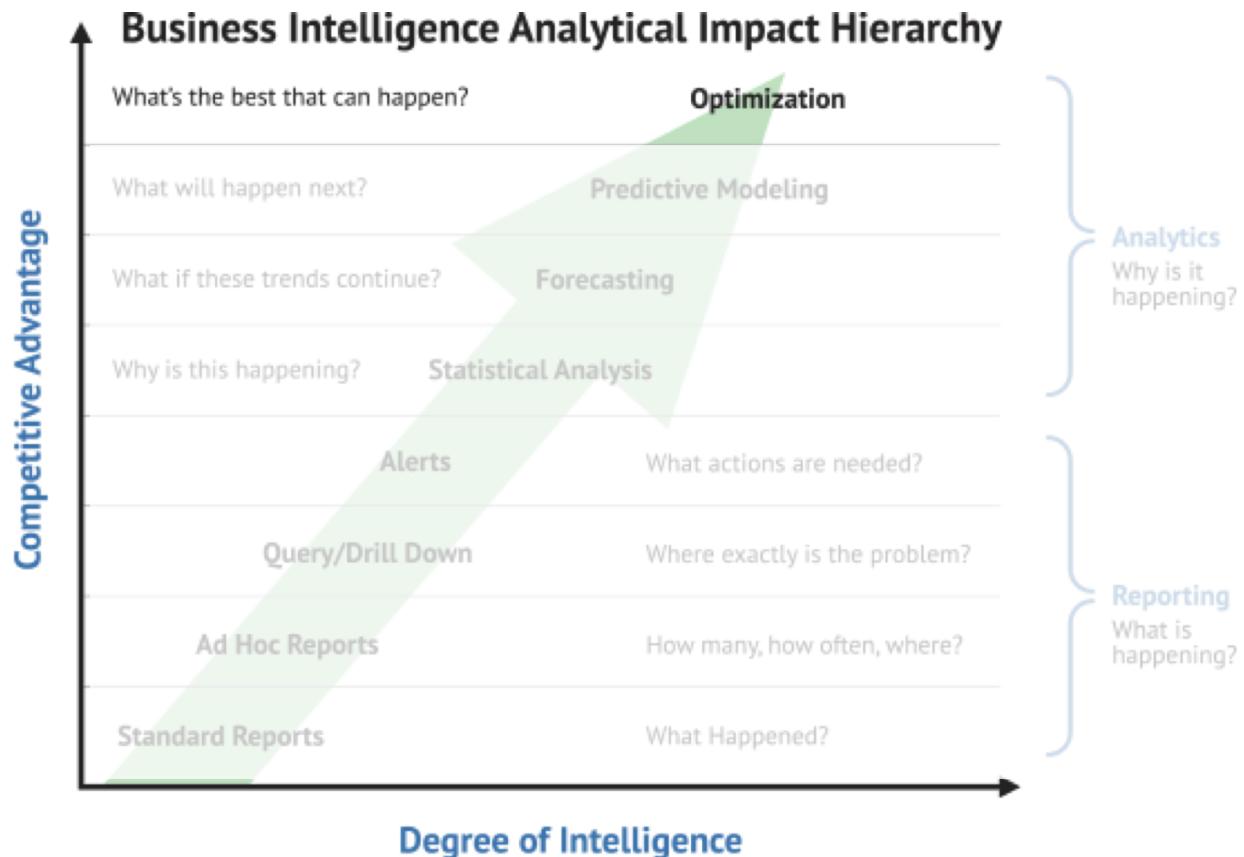


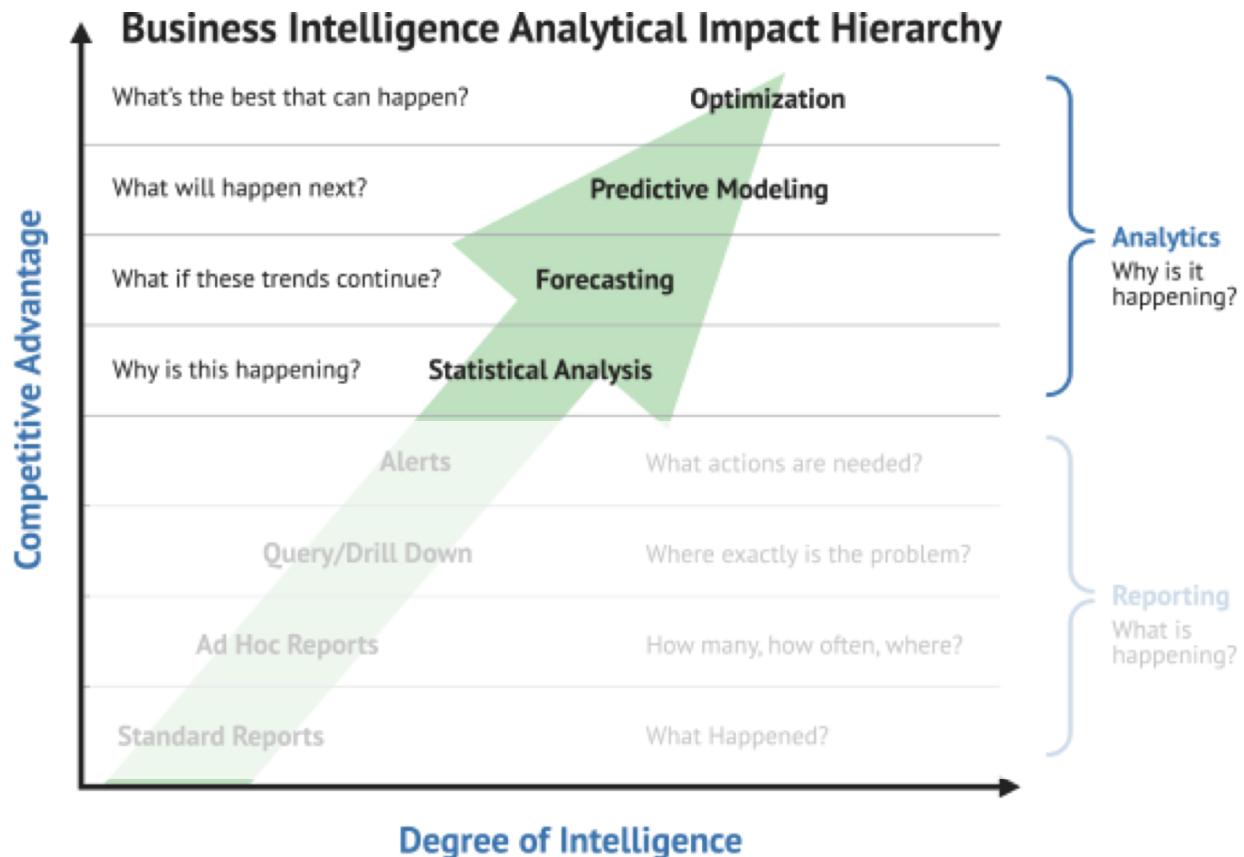


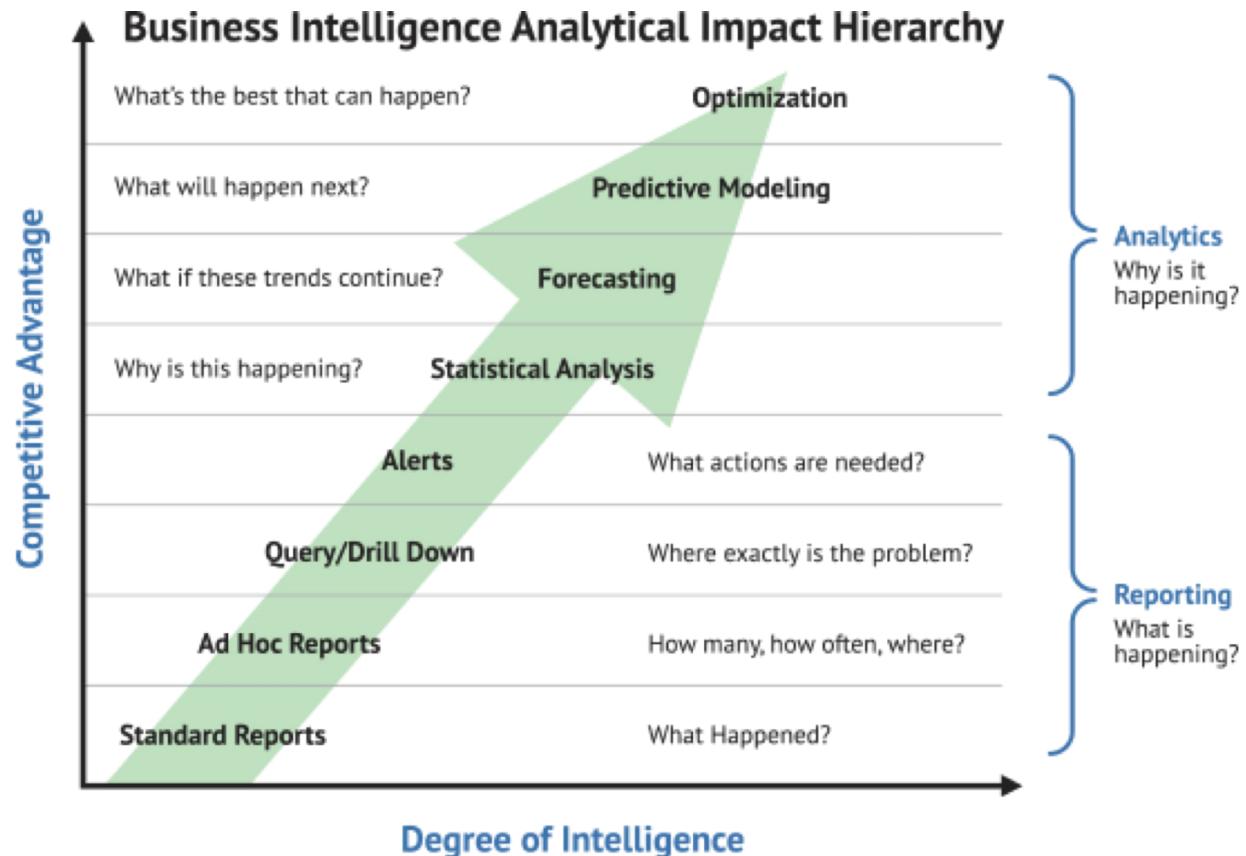




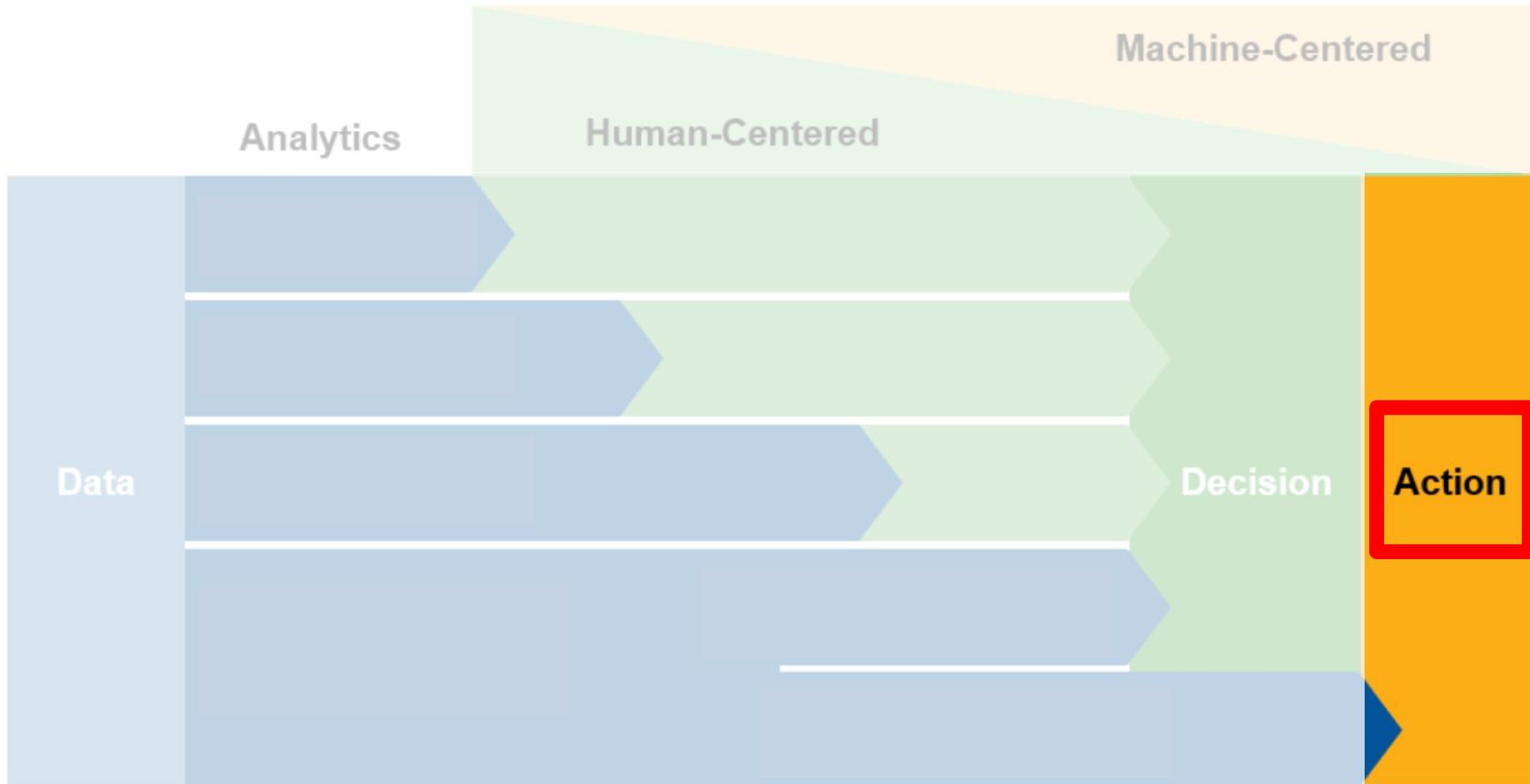




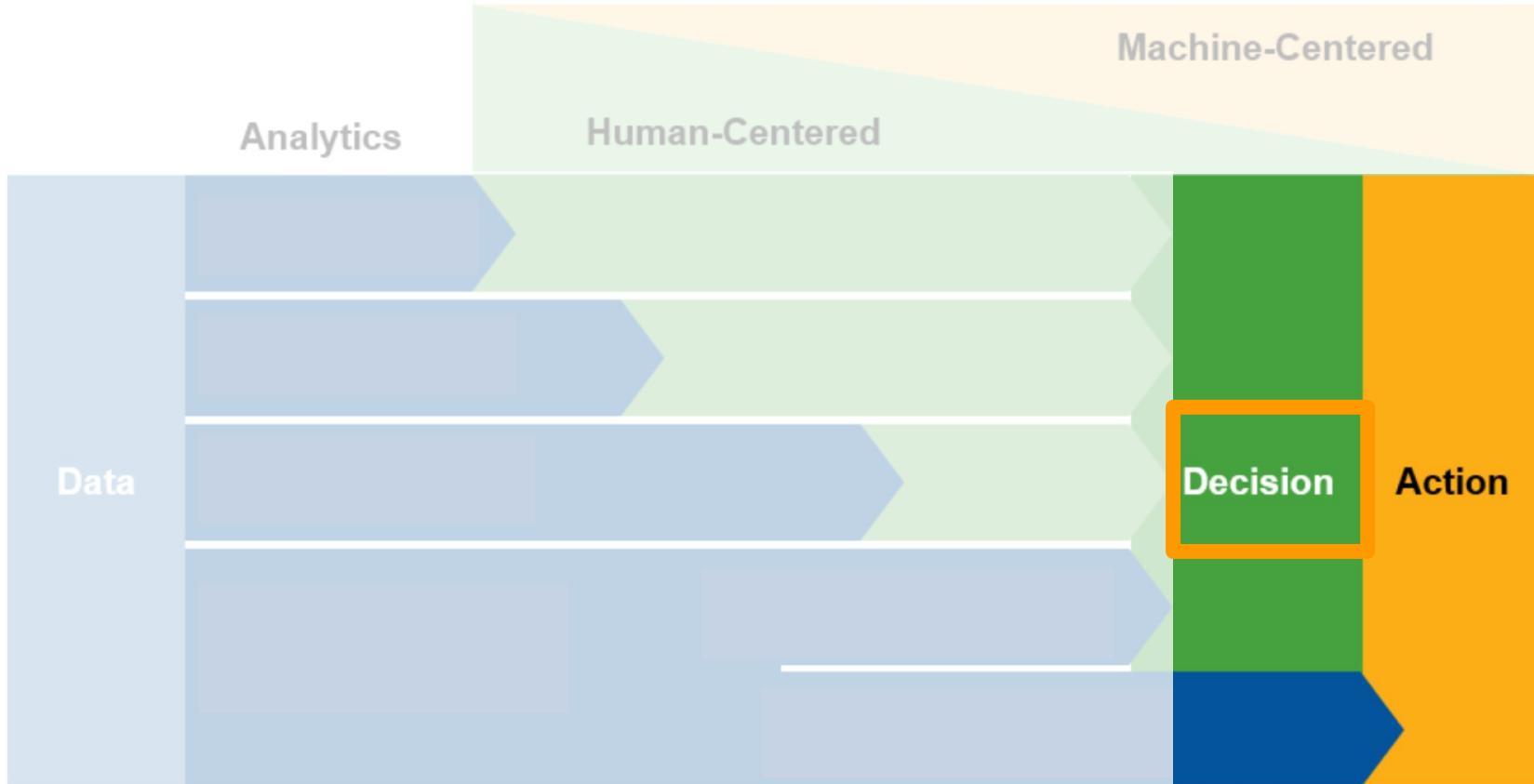




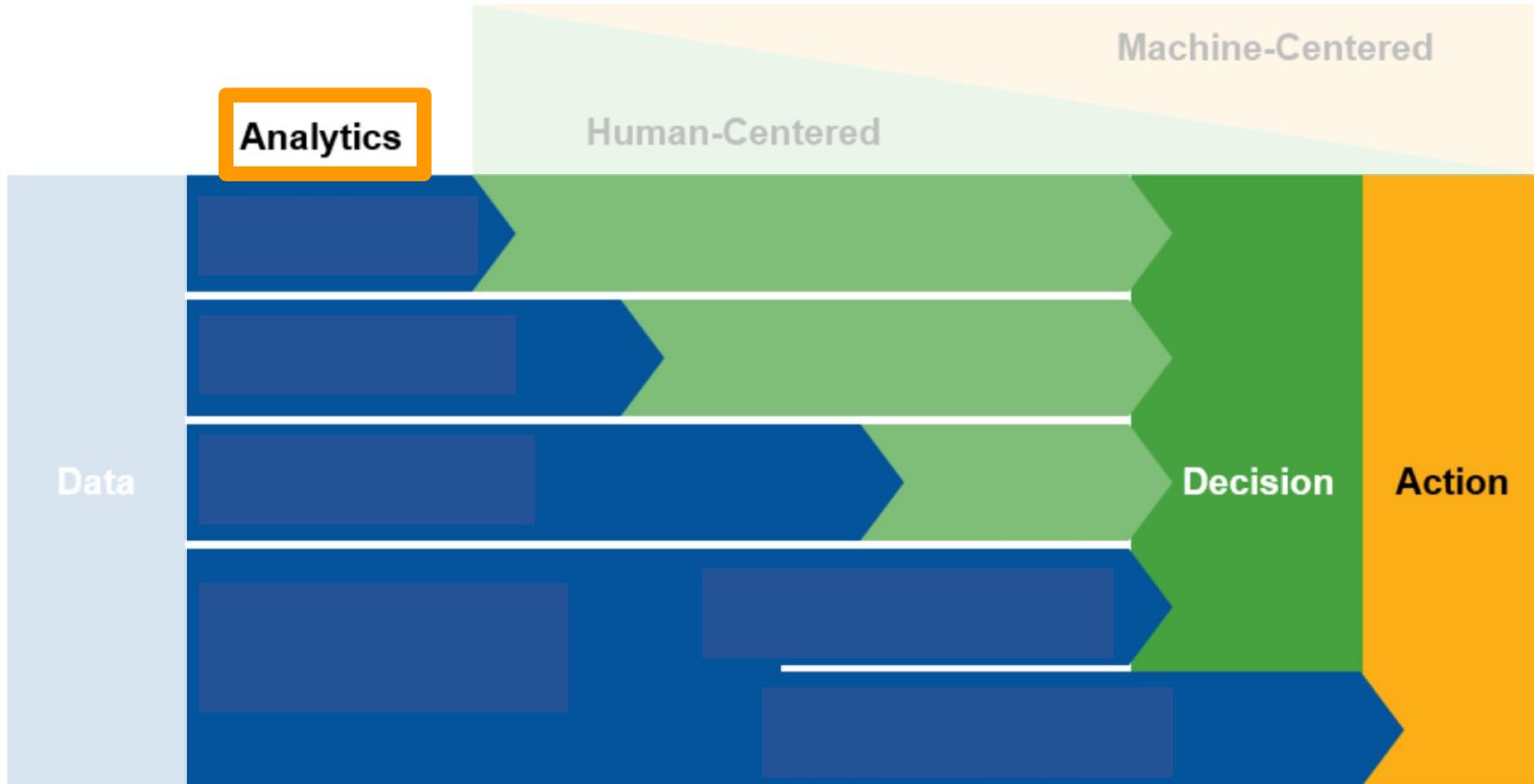
Analytics in decision support systems



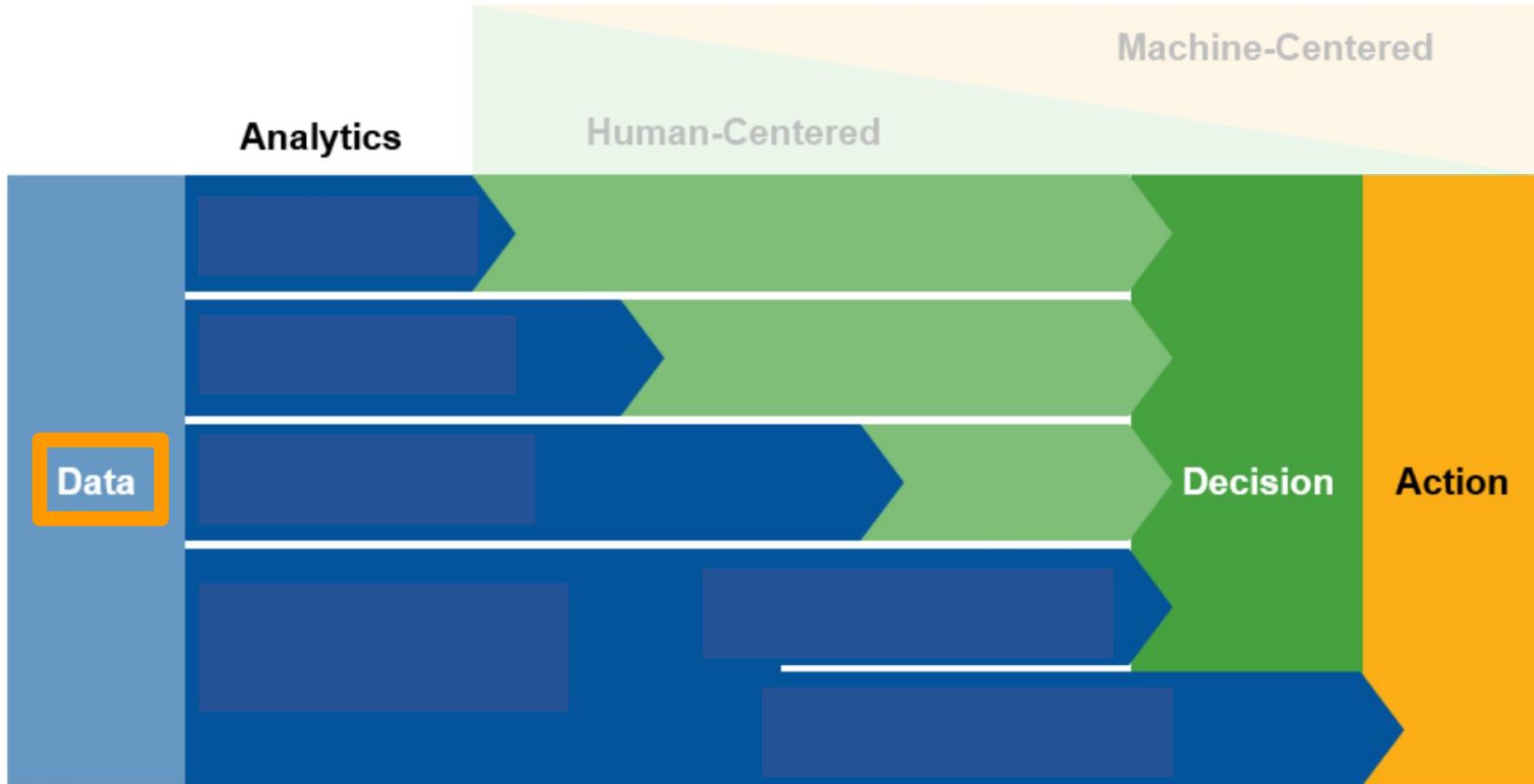
Source: Gartner (October 2016)



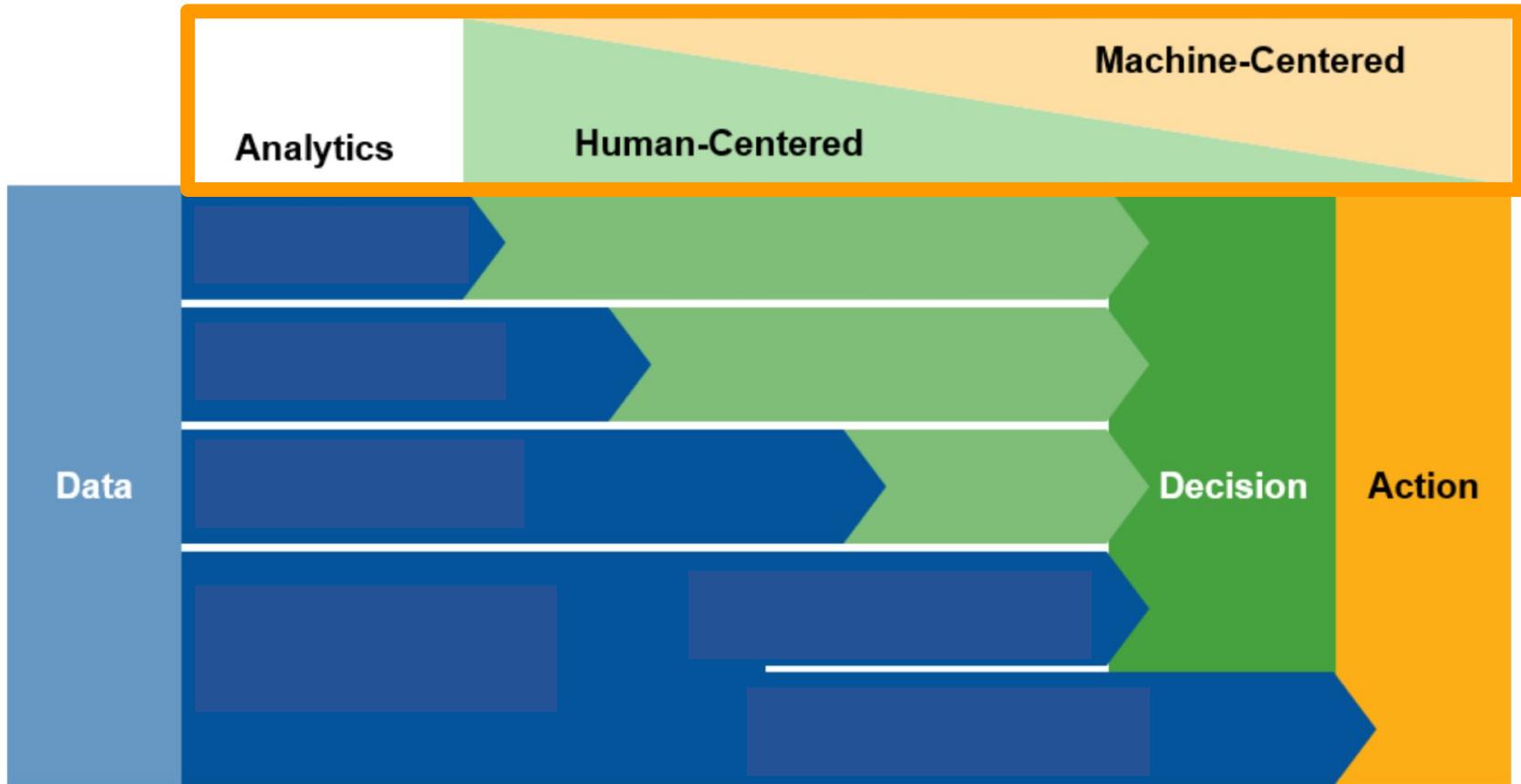
Source: Gartner (October 2016)



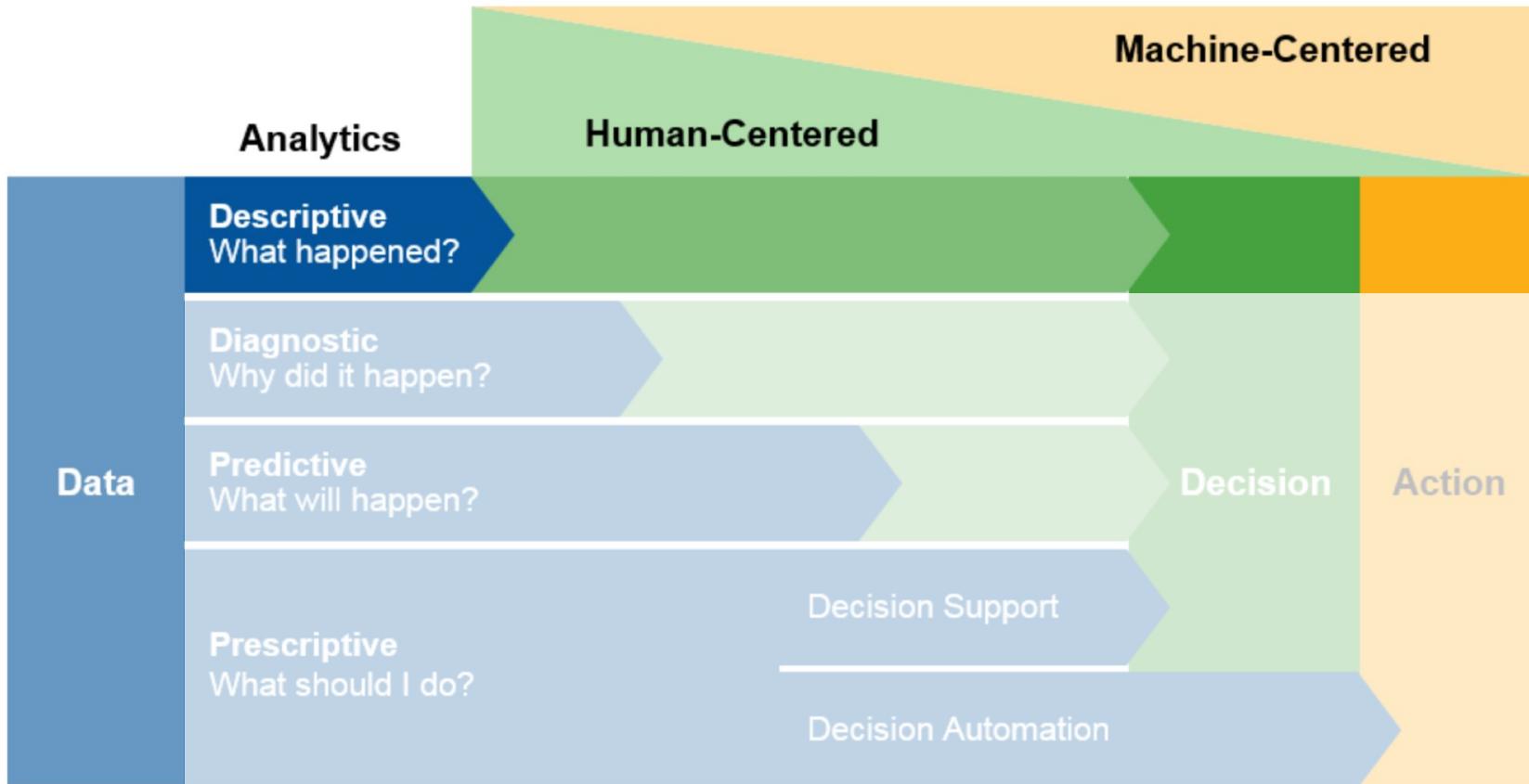
Source: Gartner (October 2016)



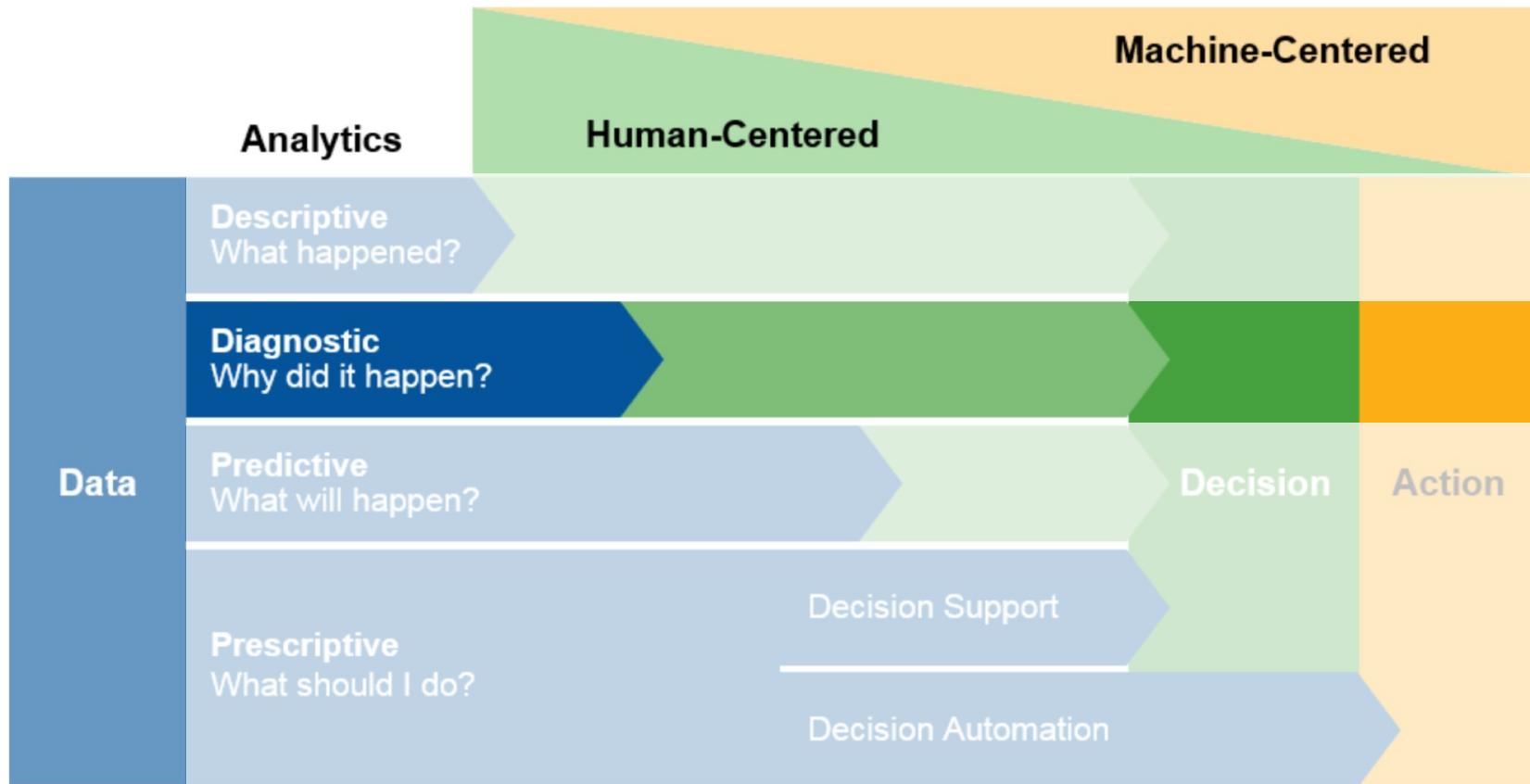
Source: Gartner (October 2016)



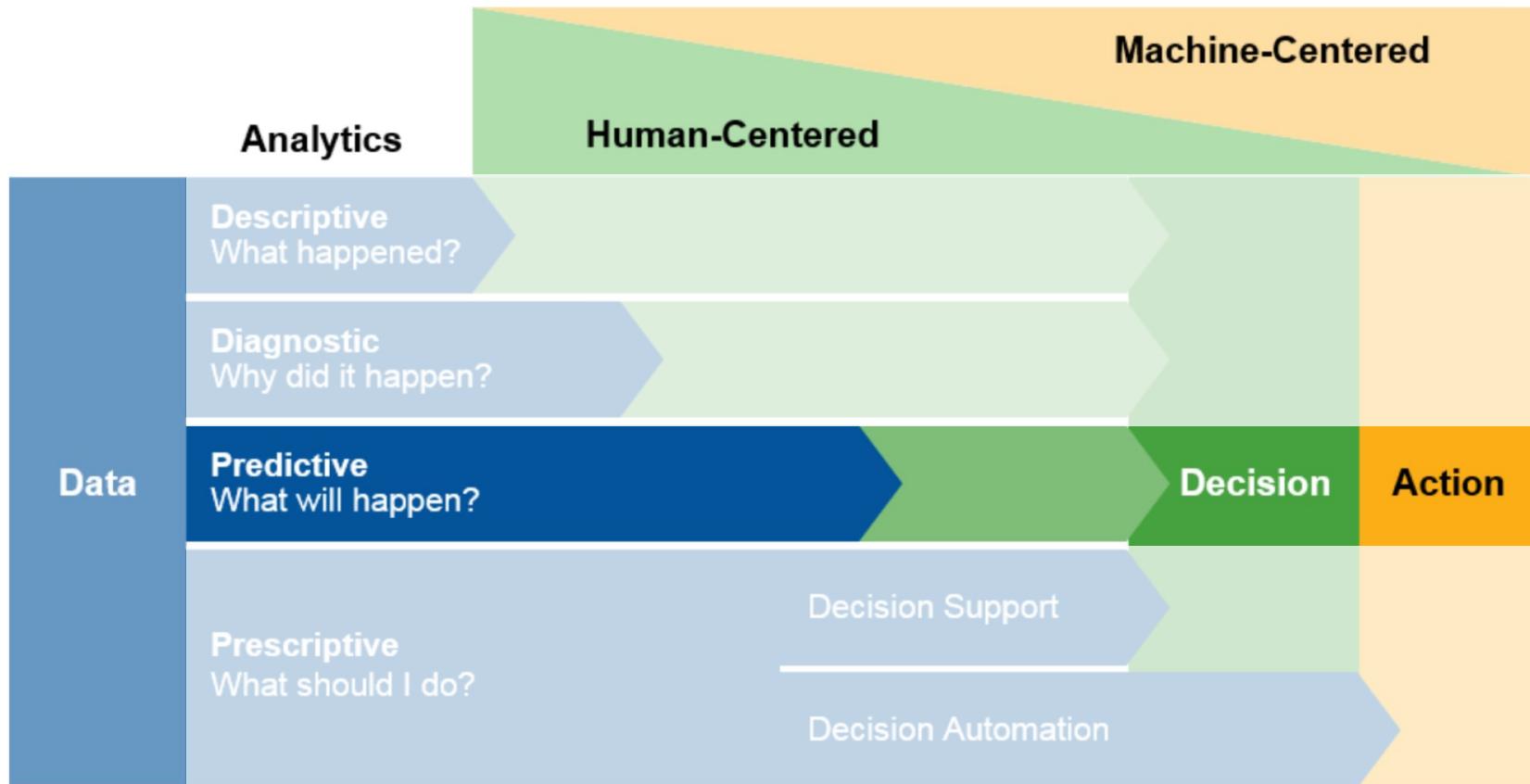
Source: Gartner (October 2016)



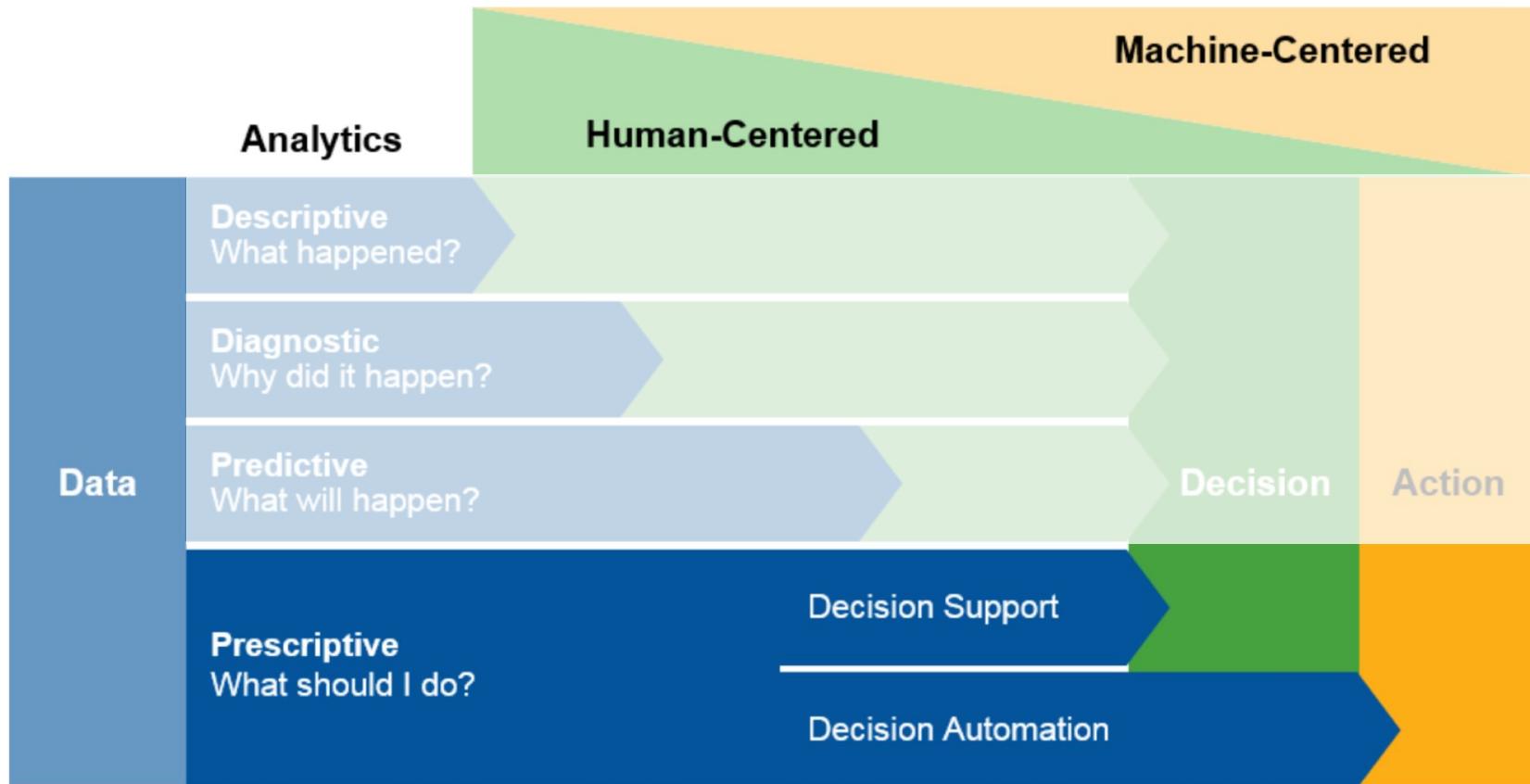
Source: Gartner (October 2016)



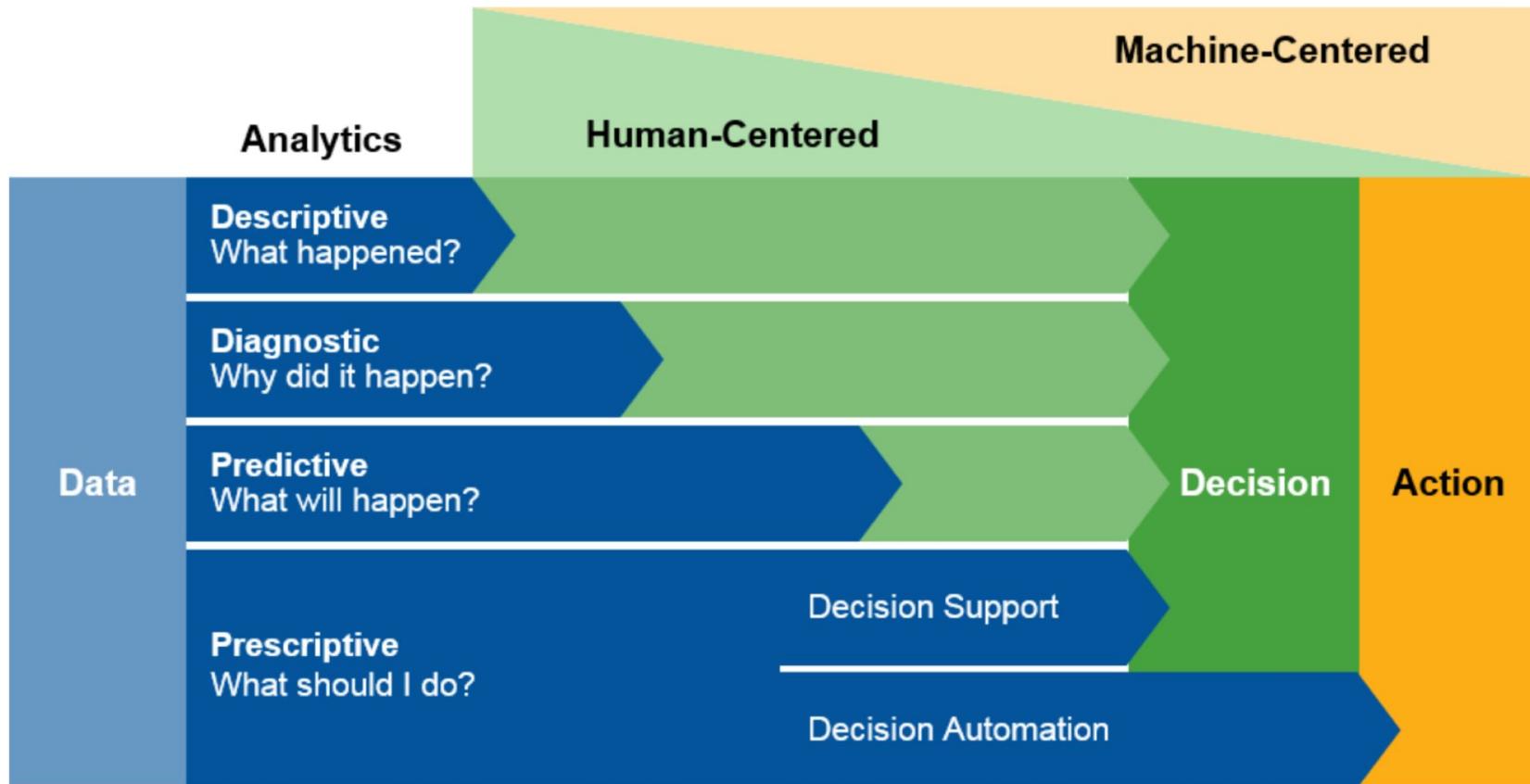
Source: Gartner (October 2016)



Source: Gartner (October 2016)

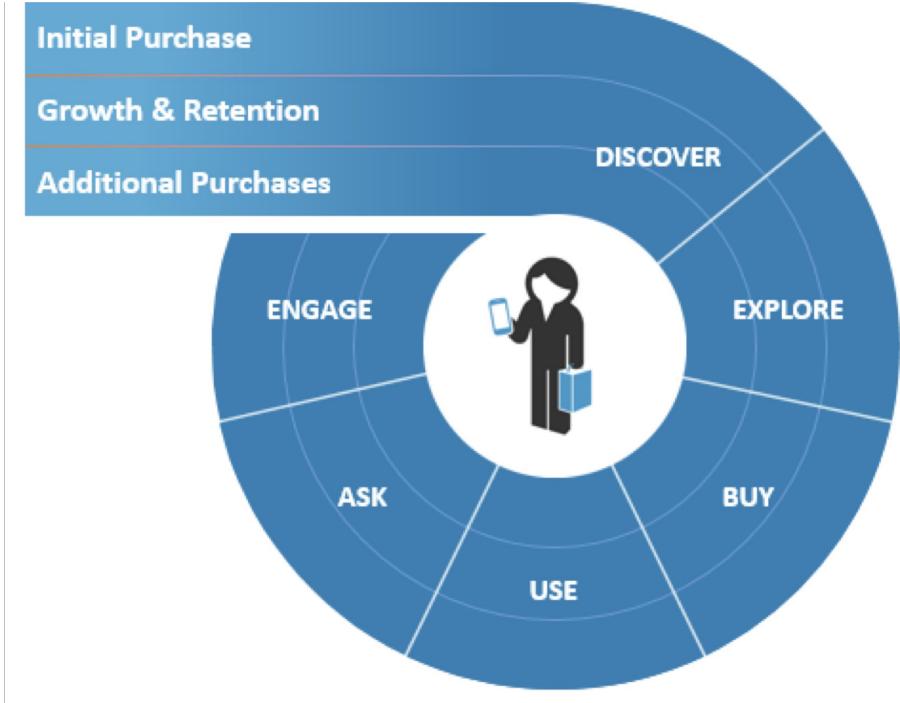


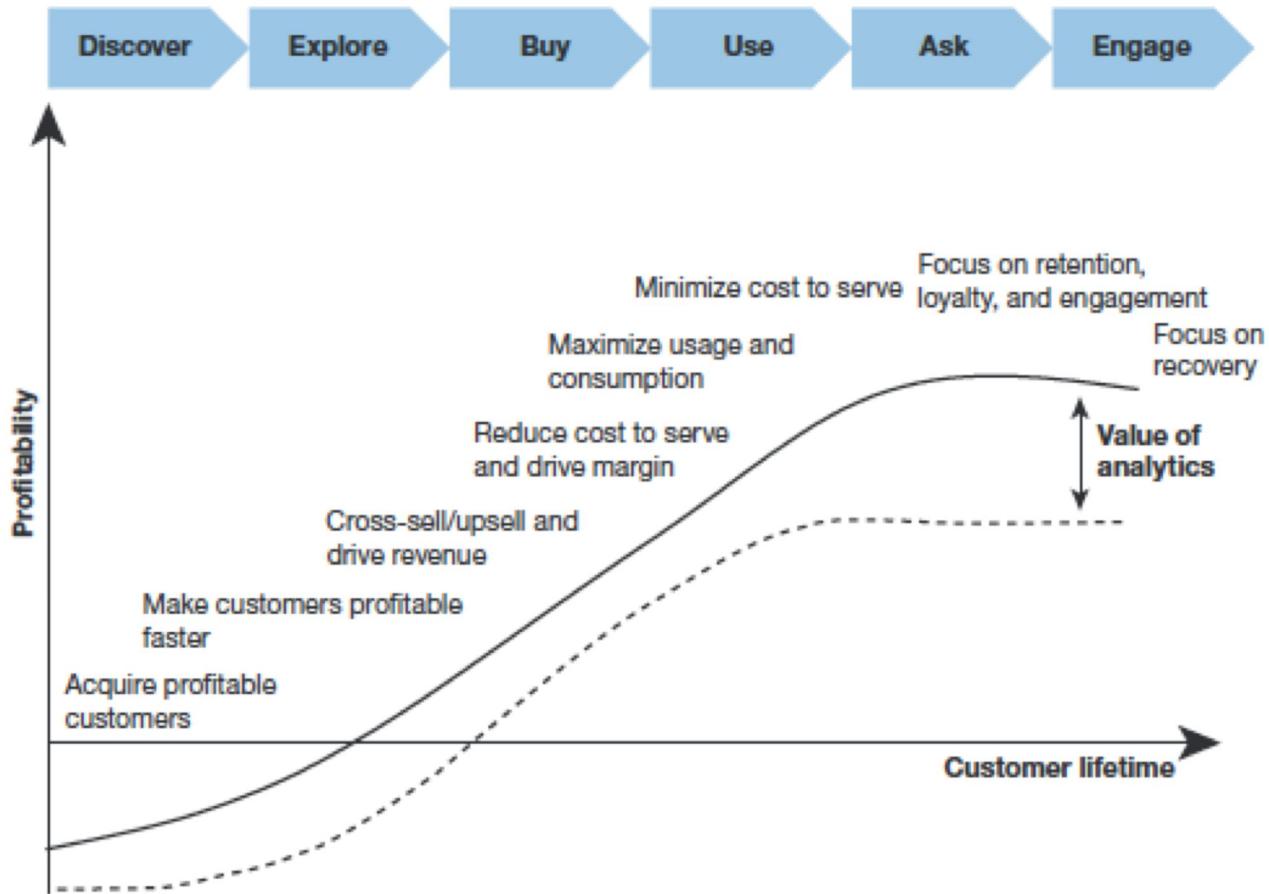
Source: Gartner (October 2016)



Source: Gartner (October 2016)

Analytics use cases

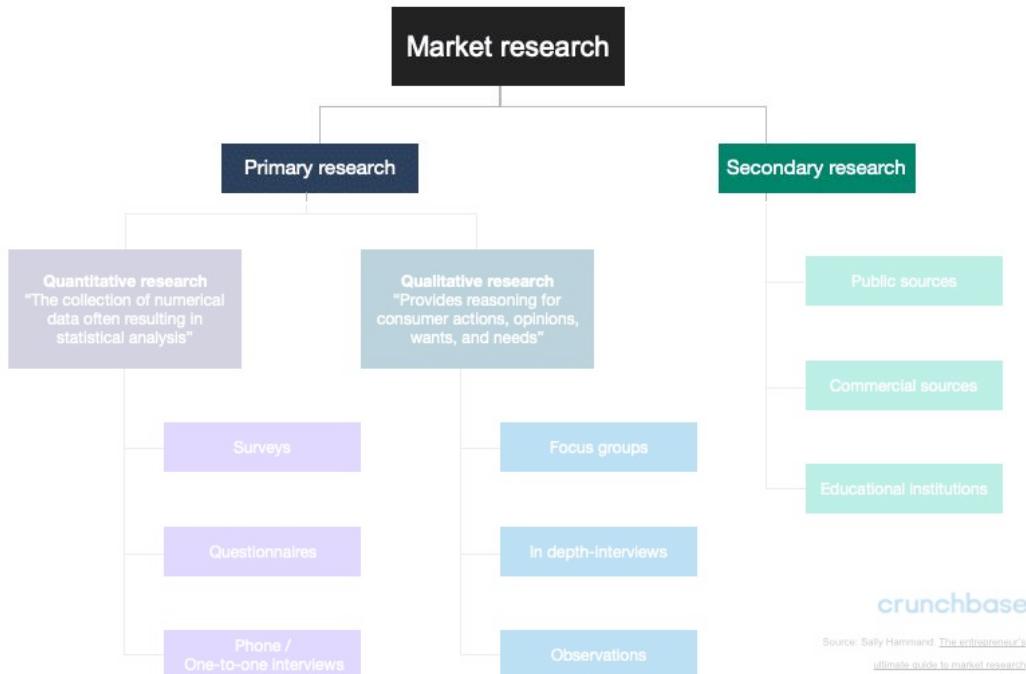




Life-cycle stage	Business objective	Analytical method
Discover	Profile customers	Segmentation
	Evaluate prospects	Lead scoring
	Reach right prospects	Customer look-alike targeting
Explore	Analyze customer response	Offer/contact optimization
	Optimize marketing mix	Marketing mix modeling
	Test marketing 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

How do we obtain
data?

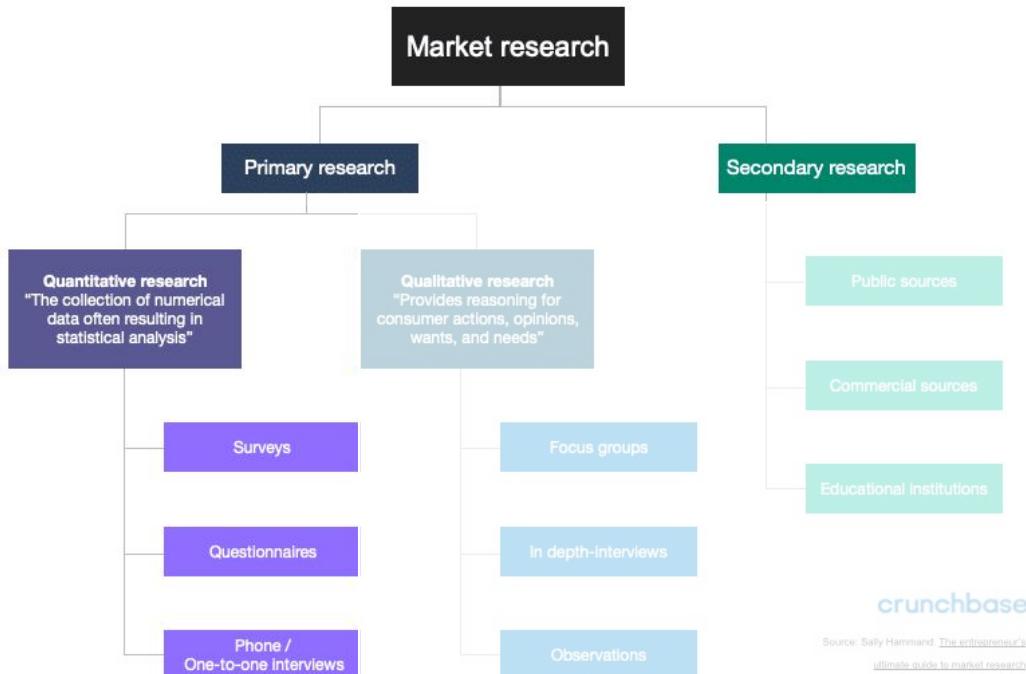
Primary vs secondary research



crunchbase

Source: Sally Hammand. [The entrepreneur's ultimate guide to market research](#)

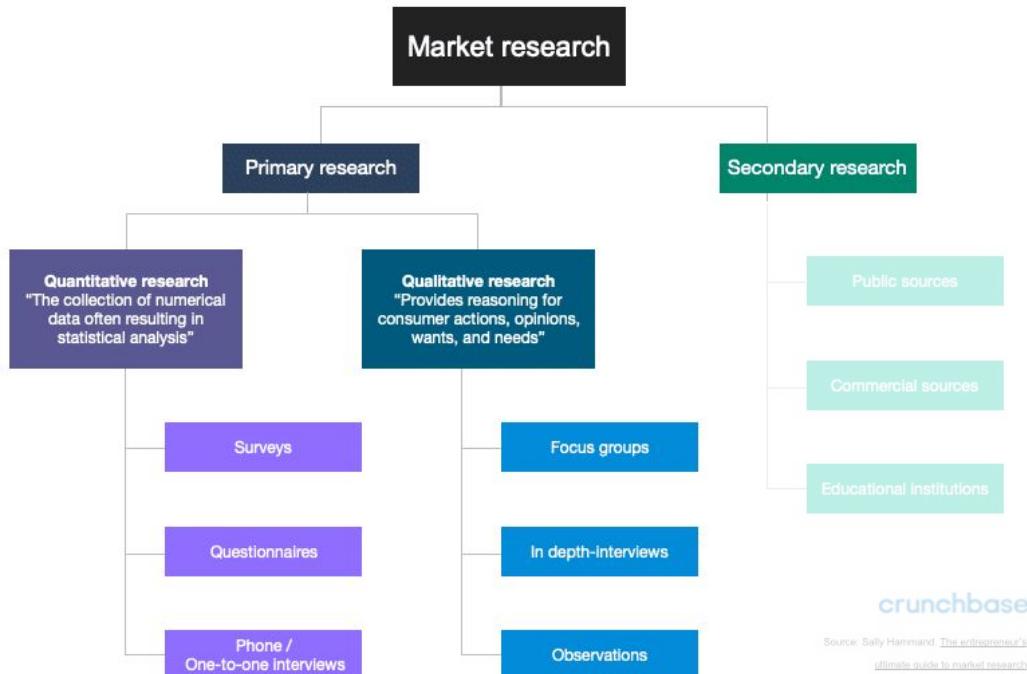
Quantitative research



crunchbase

Source: Sally Hammand. [The entrepreneur's ultimate guide to market research](#)

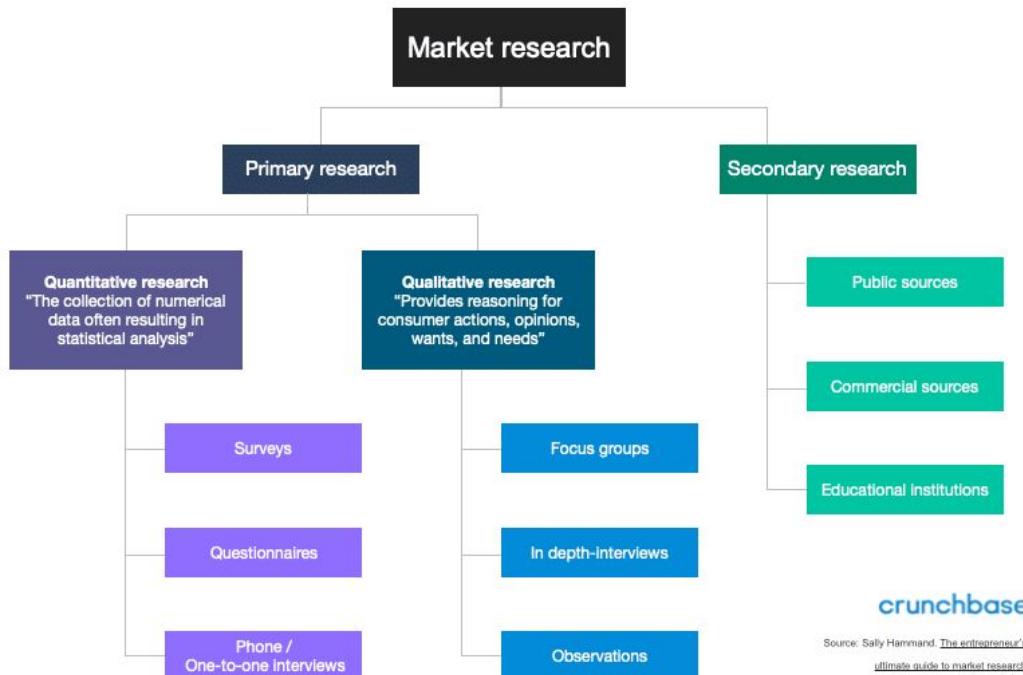
Qualitative research



crunchbase

Source: Sally Hamman. [The entrepreneur's ultimate guide to market research](#)

Types of market research

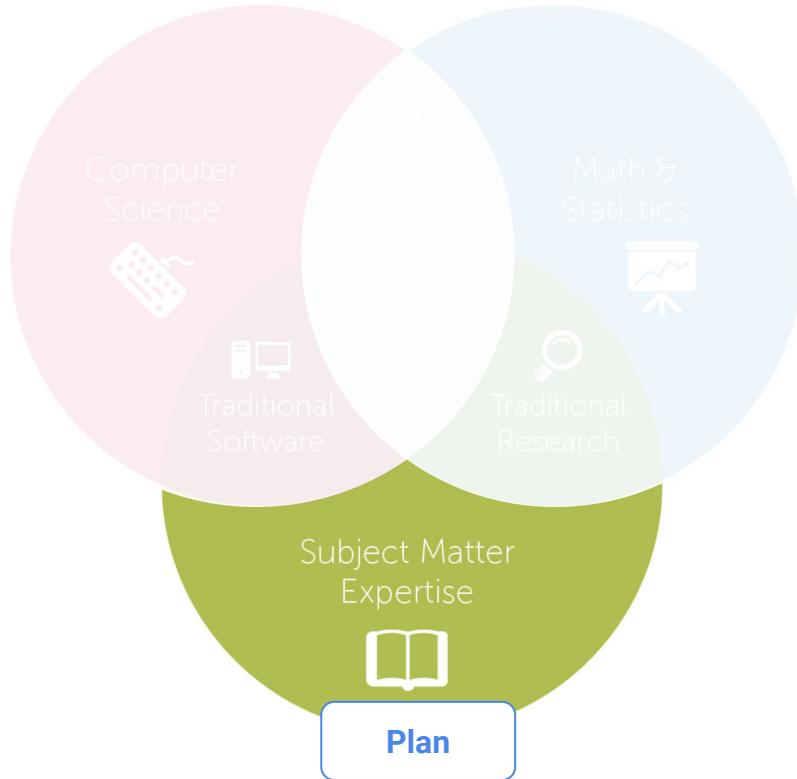


crunchbase

Source: Sally Hammand. [The entrepreneur's ultimate guide to market research](#)

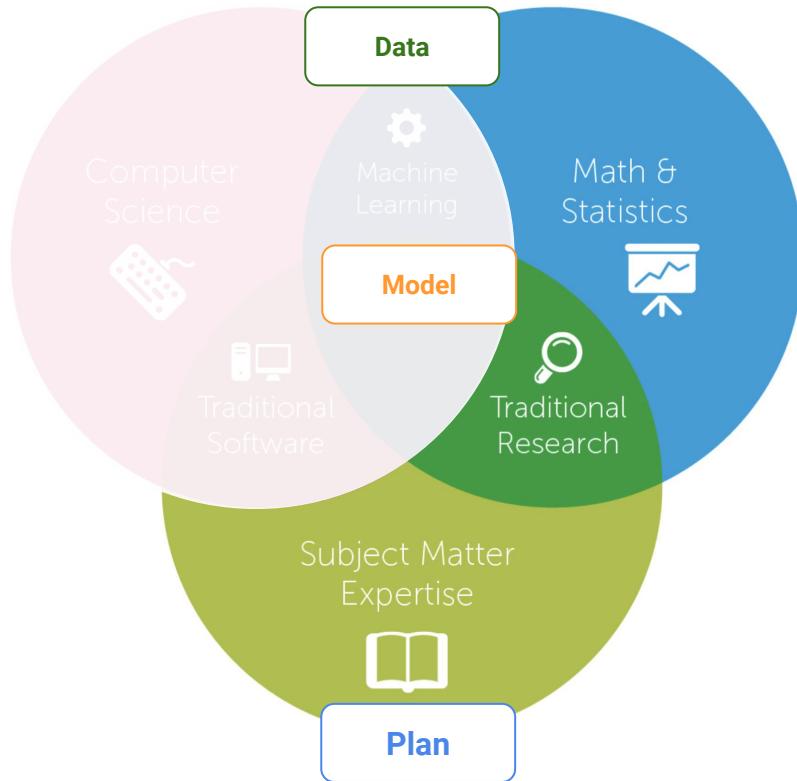
Data science

Data Science



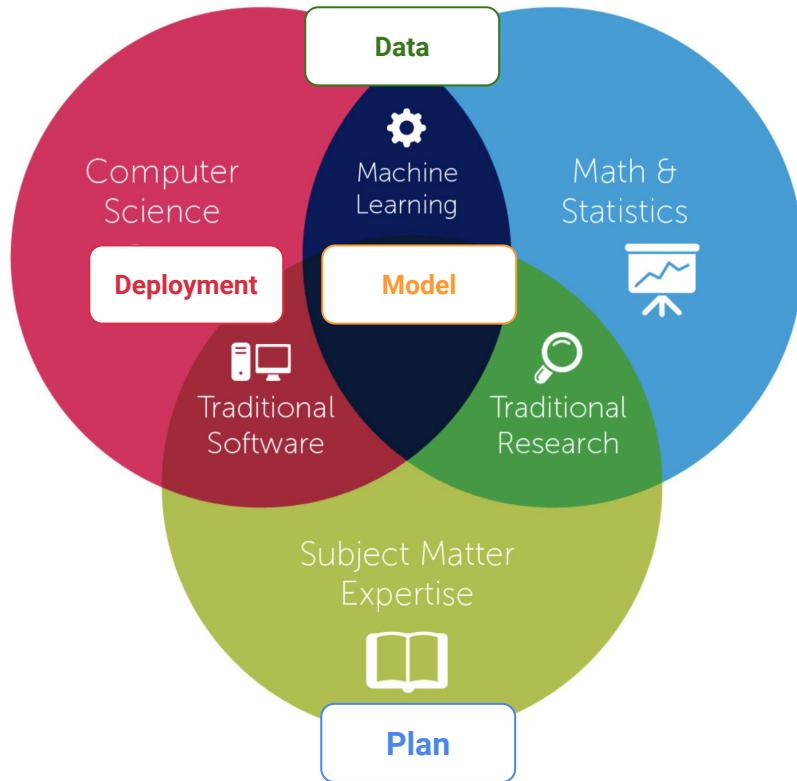
Copyright © 2014 by Steven Geringer Raleigh, NC.

Data Science



Copyright © 2014 by Steven Geringer Raleigh, NC.

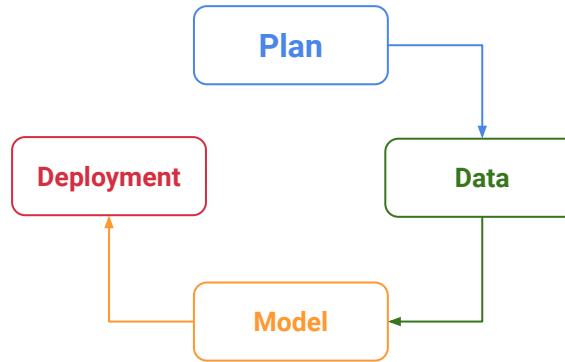
Data Science



Copyright © 2014 by Steven Geringer Raleigh, NC.

Data science is about ...

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



Computer science: Python

kaggle

Search

Kaggle Rankings

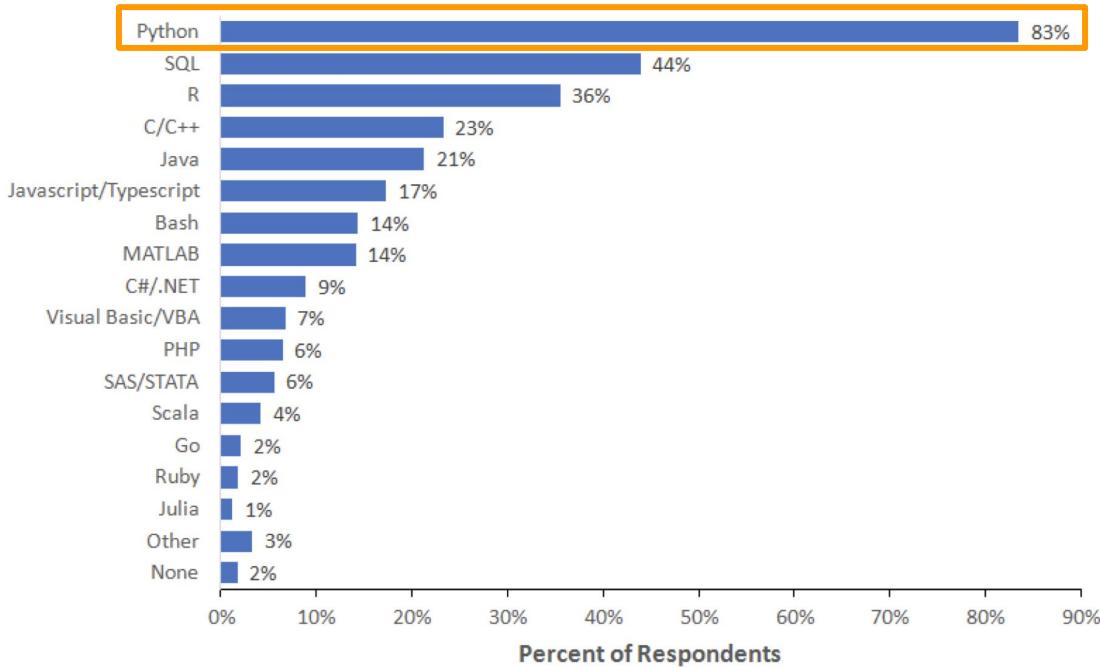
Competitions Datasets Notebooks Discussion Learn more about rankings

179 Grandmasters 1280 Masters 3,286 Experts 53,520 Contributors 71,297 Novices

Rank	Tier	User	Joined	Medals	Points
1	Gold	beefitting	joined 3 years ago	26 4 1	250,020
2	Gold	Guanshuo Xu	joined 4 years ago	12 2 2	219,221
3	Gold	Oba	joined 9 years ago	40 28 20	150,540
4	Gold	Moscow_Myanya_dg Kazanova	joined 7 years ago	37 16 23	131,628
5	Gold	stude	joined 3 years ago	8 2 2	123,527
6	Gold	Pai	joined 8 years ago	7 2 0	112,077
7	Gold	CIPMP	joined 7 years ago	14 2 3	109,008
8	Gold	zot	joined 7 years ago	8 2 1	108,949
9	Gold	w0t12	joined 5 years ago	12 18 6	107,087
10	Gold	schuhner	joined 3 years ago	9 13 3	105,990

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

What programming language do you use on a regular basis?



Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: <http://www.kaggle.com/kaggle/kaggle-survey-2018>. A total of 18827 respondents answered the question.



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](#)

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll

KDnuggets™ Subscribe to KDnuggets News [Twitter](#) [Facebook](#) [LinkedIn](#) [Search](#)

Blog/News | Opinions | Tutorials | Top Stories | Companies | Courses | Databases | Education | Jobs | Meetings | Software | Webinars | Contact

Predictive Analytics World Financial, May 31 - June 4, 2020 **LAS VEGAS** **MAY 31-JUNE 4, 2020** **MINIMIZE RISK & MULTIPLY RETURNS WITH MACHINE LEARNING** Use code **KDNUGGETS** for a 15% discount

Latest Posts

- Trends in Machine Learning In 2020
- TensorFlow 2.0 Tutorial: Optimizing Training Time Performance
- Achieving Accuracy with your Training Dataset
- How Bad Data is Affecting Your Organization's Operational Efficiency
- Top 10 Data Science Projects, Feb 26 - Mar 03: Free Mathematics Courses for #DataScience & MachineLearning
- A simple and interpretable performance measure for a binary classifier

New KDnuggets Poll: When will AutoML reach expert Data Scientist level?

Top Stories Last Week

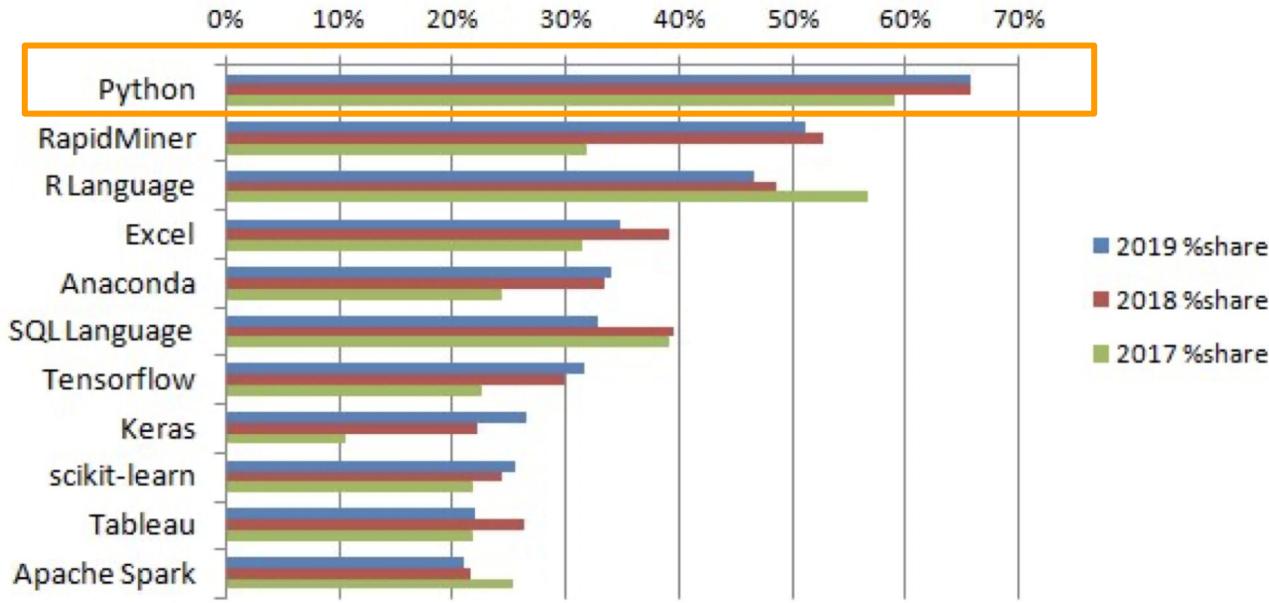
Most Popular

- Learning from 3 big Data Science career mistakes

KDnuggets AutoML/Automated Data Science Poll

* 1. When will most expert-level Data Science/Machine Learning

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

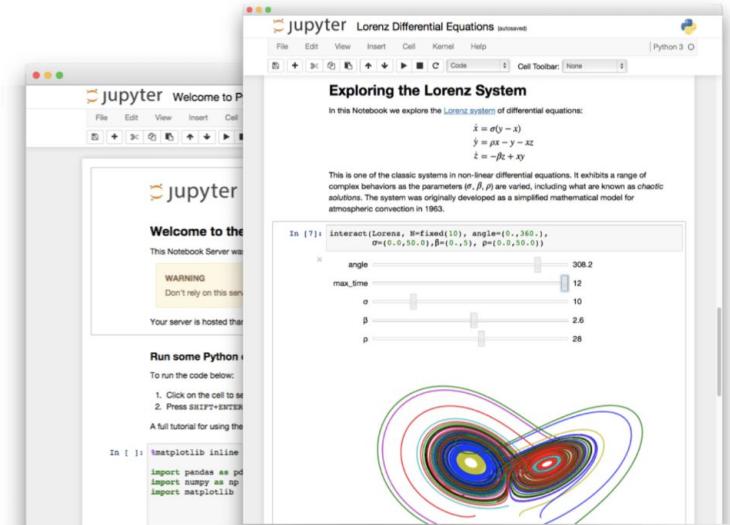


- Python is an **object-oriented** language (an object is an entity that contains data along with associated metadata and/or functionality).
- One thing that distinguishes Python from other programming languages is that it is **interpreted** rather than compiled.
- This means that it is executed **line by line** which is particularly 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.

Table of contents X

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

>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).



scikit-learn

Machine Learning in Python

[Getting Started](#)[What's New in 0.22.2](#)[GitHub](#)

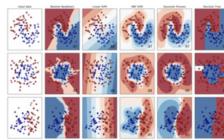
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...

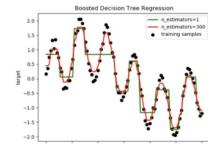
[Examples](#)

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...

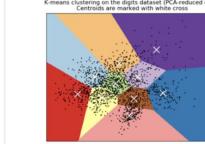
[Examples](#)

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...

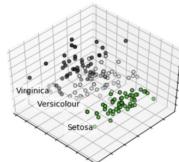
[Examples](#)

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...

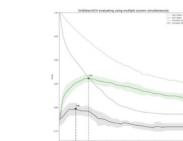


Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning

Algorithms: grid search, cross validation, metrics, and more...

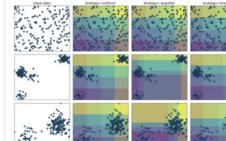


Preprocessing

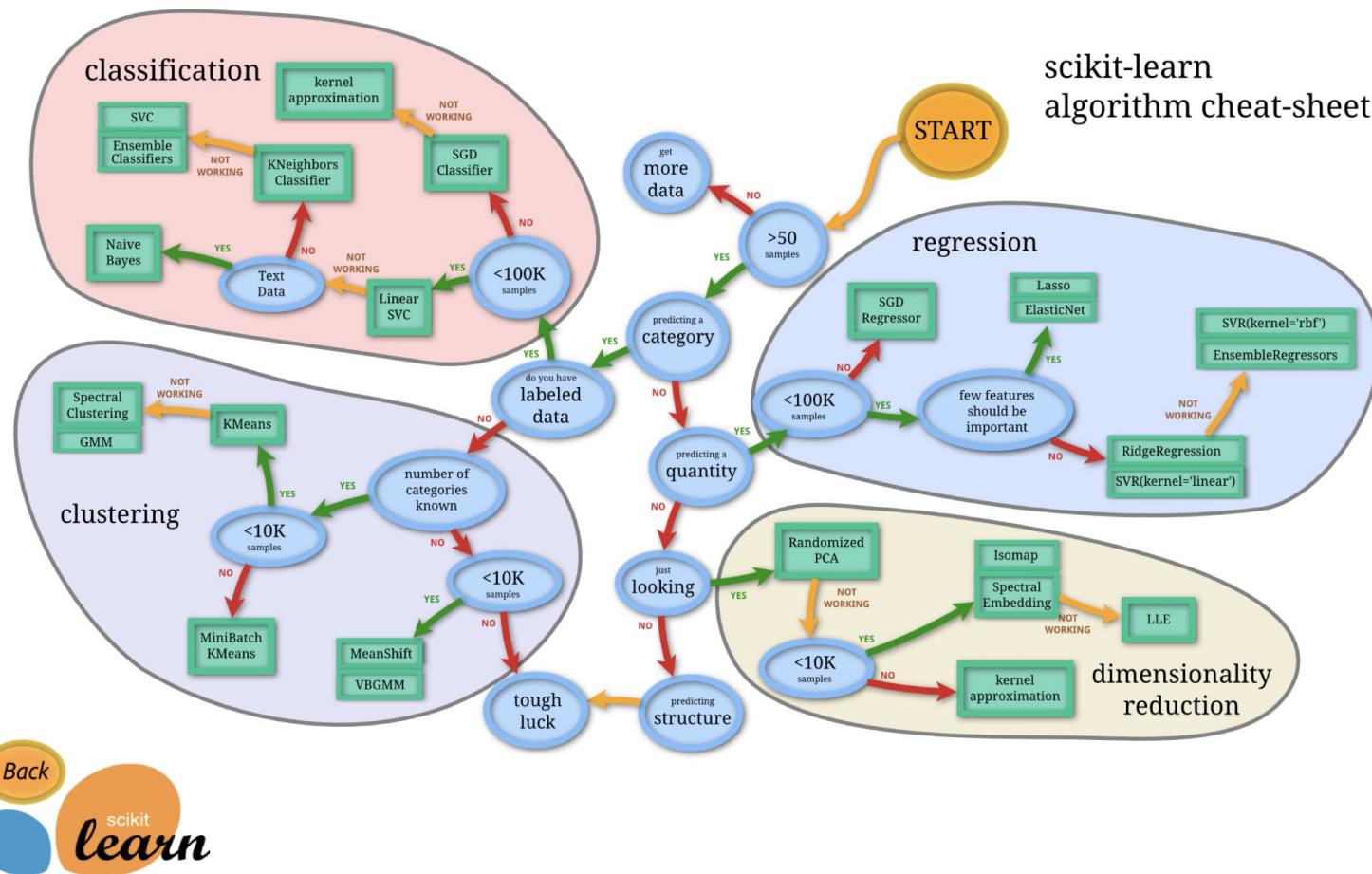
Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.

Algorithms: preprocessing, feature extraction, and more...



scikit-learn algorithm cheat-sheet





Leading Open Data Science Platform Powered by Python

CONDA[®]

Leading Package and Environment Manager

OPEN DATA SCIENCE



theano

DATA



+



cloudera



COMPUTATION



AMD



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

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



GitLab



GitHub

[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

[Copy](#) [Raw](#) [Blame](#) [History](#)

Introduction to Data Science with Python

Prof. Dr. Jan Kirenz

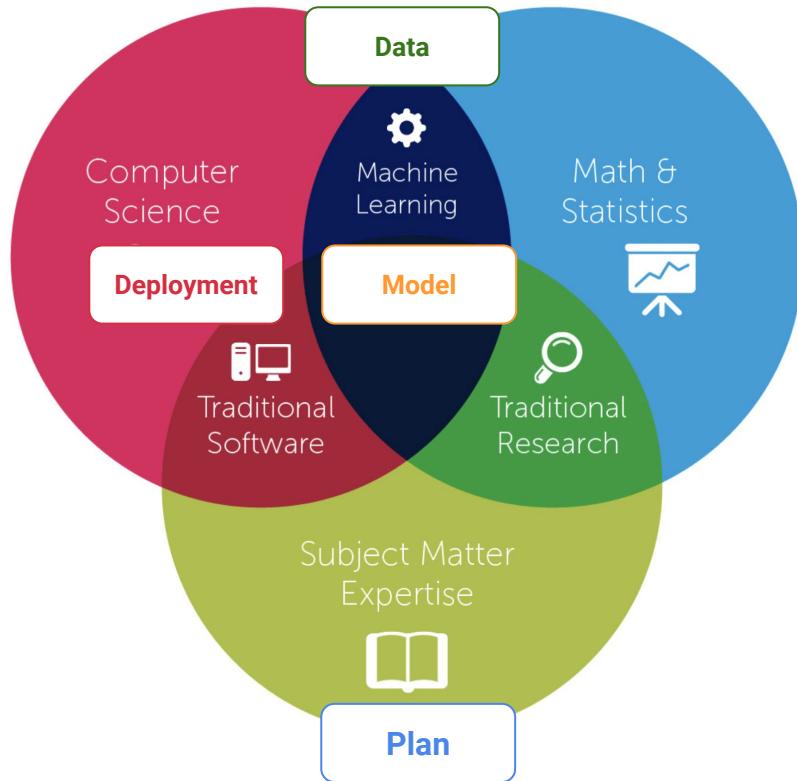
Hochschule der Medien Stuttgart

```
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() # seaborn 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



Copyright © 2014 by Steven Geringer Raleigh, NC.