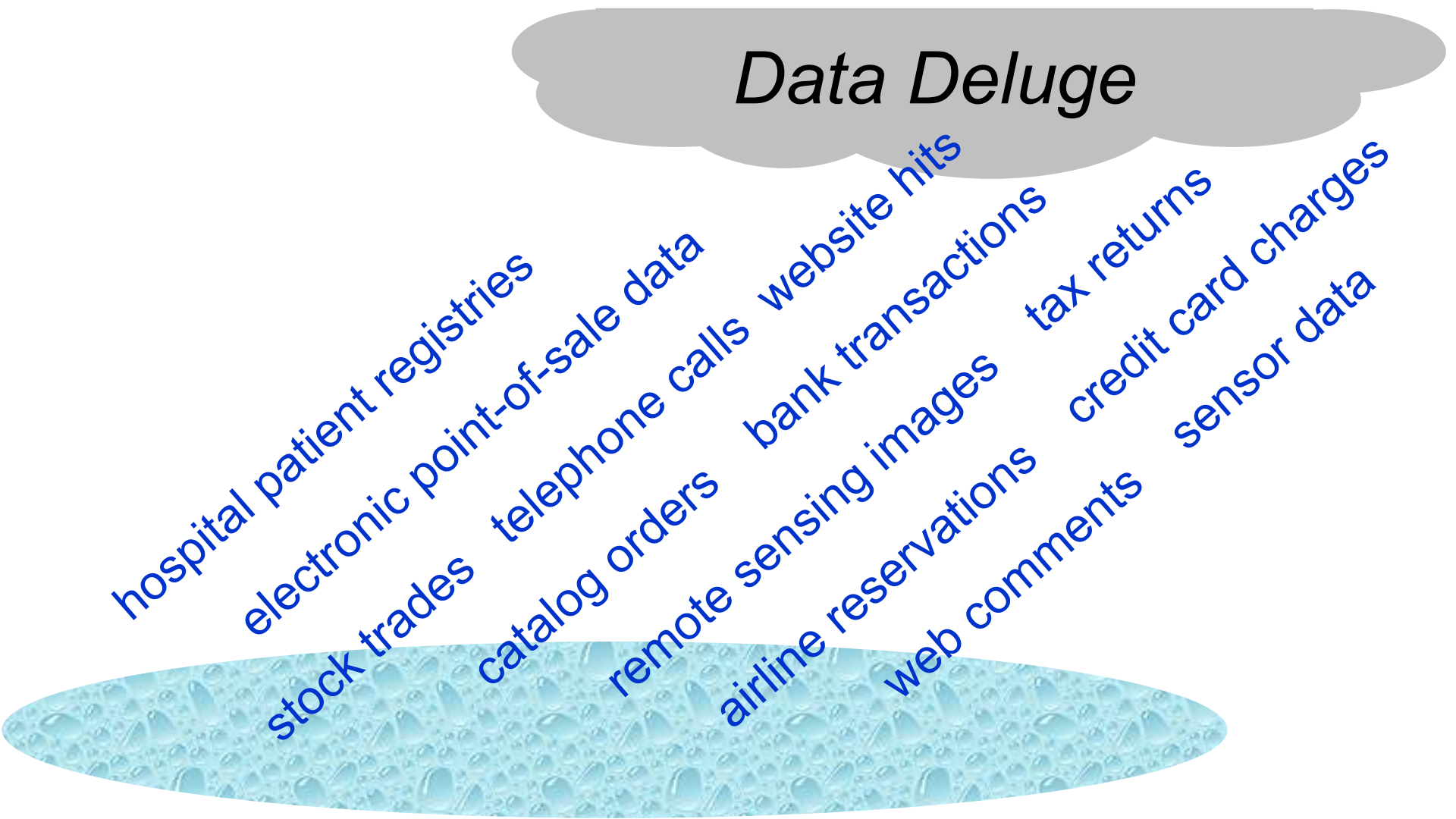
# Lesson 1:
# Big Data, Analytics, and Data Science

| |
|---|
| 1.1 Introduction |
| 1.2 The (Citizen) Data Scientist |
| 1.3 Skills for the Data Scientist |

# Lesson 1:
# Big Data, Analytics, and Data Science

| |
|---|
| **1.1 Introduction** |
| 1.2 The (Citizen) Data Scientist |
| 1.3 Skills for the Data Scientist |

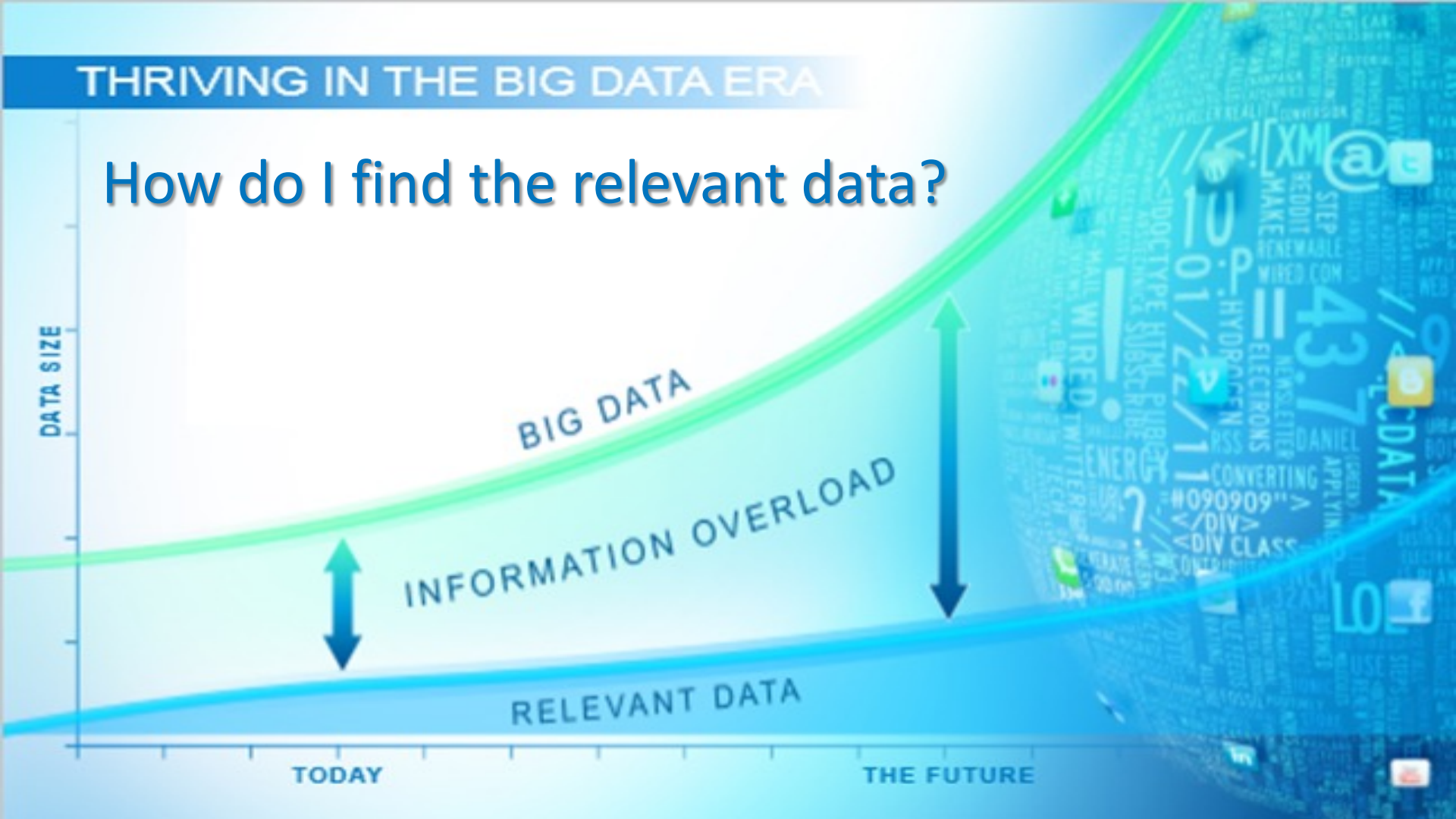

Data Deluge
hospital patient registries
electronic point-of-sale data
stock trades
telephone calls
catalog orders
website hits
bank transactions
remote sensing images
airline reservations
tax returns
credit card charges
web comments
sensor data

# Consequences of the Data Deluge

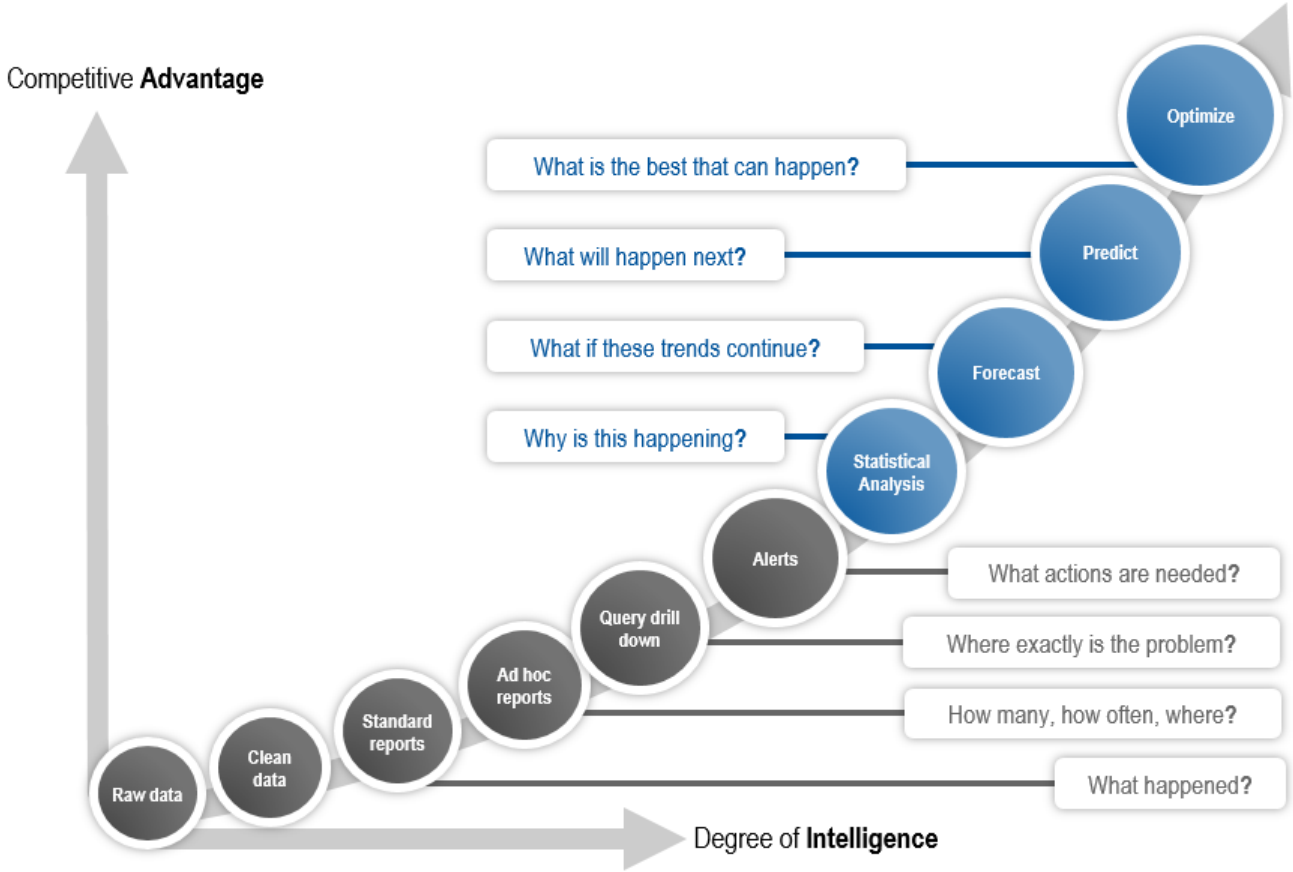
- **Every problem generates data eventually.**
  Proactively defining a data collection protocol results in more useful information. This leads to more useful analytics.

- **Every company needs analytics eventually.**
  Proactively analytical companies compete more effectively.

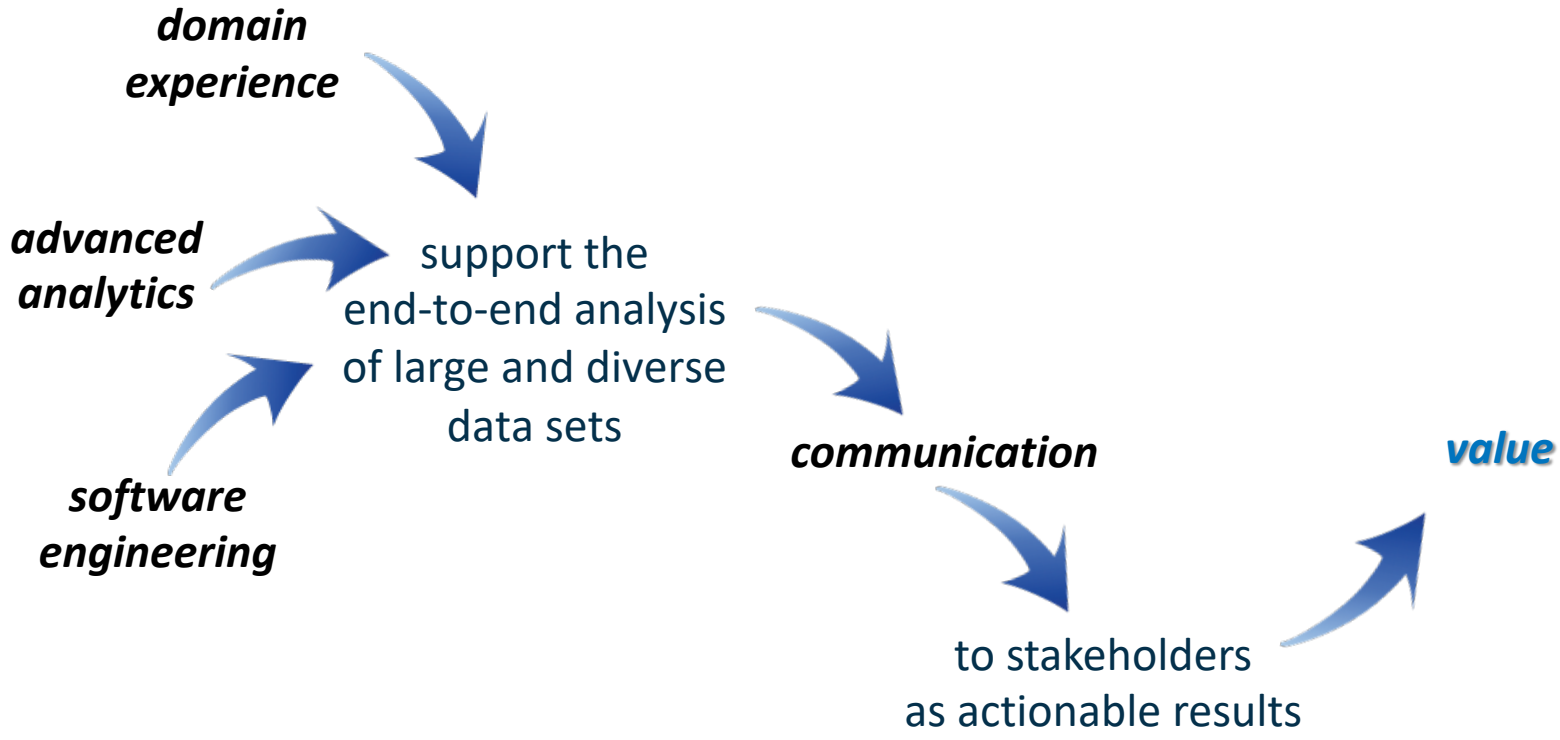
- **Everyone needs analytics eventually.**
  Proactively analytical people are more marketable and more successful in their work.


THRIVING IN THE BIG DATA ERA
How do I find the relevant data?
DATA SIZE
BIG DATA
INFORMATION OVERLOAD
RELEVANT DATA
TODAY
THE FUTURE

# Levels of Analytics



# Data Science: A Definition According to SAS



# Analytic Methods

what to do by providing information about optimal decisions based on the predicted future scenarios

## Prescriptive model

the use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data

## Predictive model

### types

classification -> predict class membership
regression -> predict a number

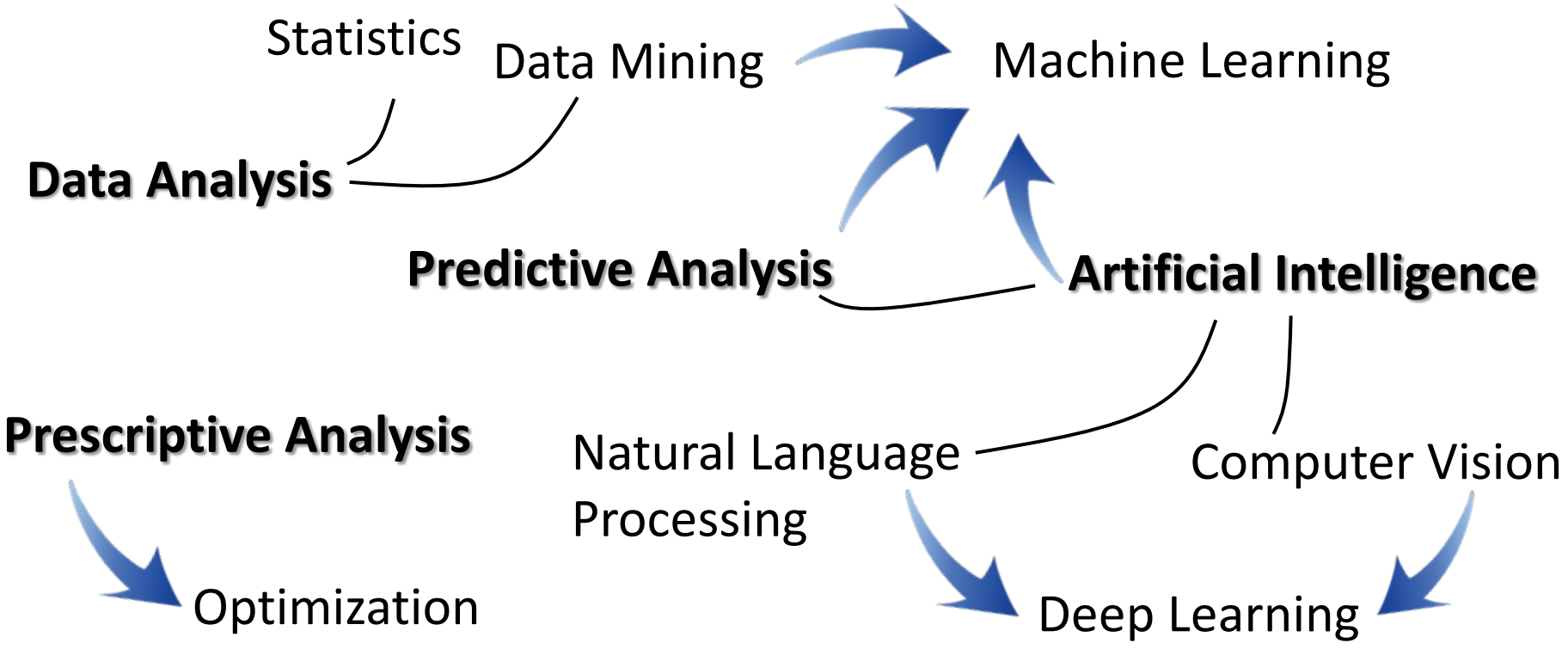
### techniques

decision trees | linear/logistic regression
neural networks
gradient boosting | random forests
support vector machines

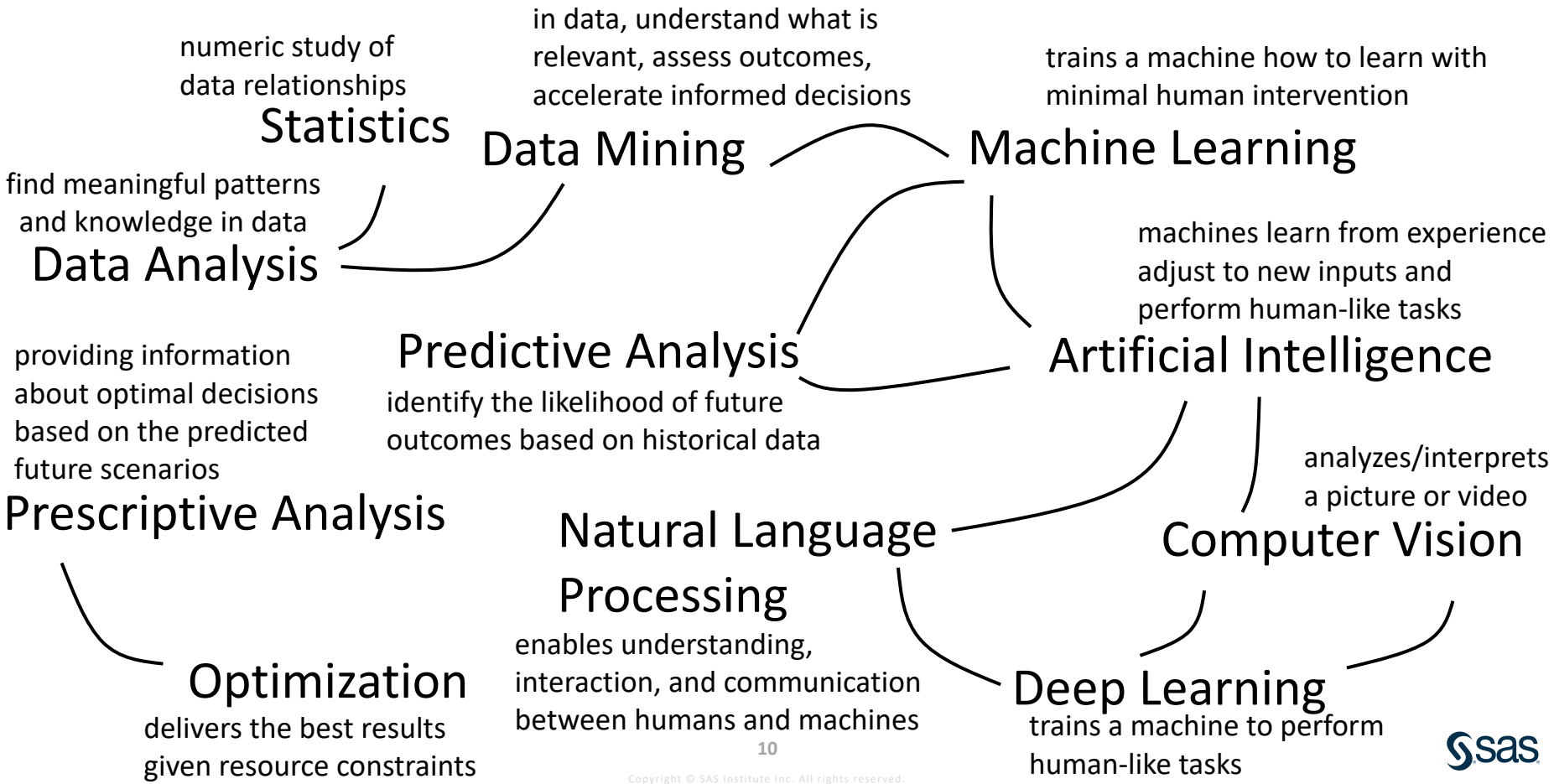
helps you understand what happened, or diagnostic models that help you understand key relationships and determine why something happened

## Descriptive model

# Glossary of Terms



# Glossary of Terms



# Reasons for the Big Data Explosion

- increasing “data velocity” due to the following:
  - streaming data feeds
  - point-of-sale (POS) transactional systems
  - radio-frequency identification (RFID) tags
  - smart metering
- bigger and cheaper data storage capabilities
- social media
- improved and automated business processes
- mergers and acquisitions, leading to the merge of multiple data sources
- more online self-service applications being used

# Factors Driving Demand for Big Data Solutions

In addition to rapidly increasing data growth rates, consider these factors:

- availability of data from social media sources
- in-memory technology
- demand for mobile business intelligence
- increasing requirements around real-time reporting
- desire to mine data from social media sources (sentiment analysis)
- …

# Big Data Explained

"Big data is what happened when the **cost of storing information became less than the cost of making the decision to throw it away**."

*- George Dyson*
*Science Historian and TED Speaker*

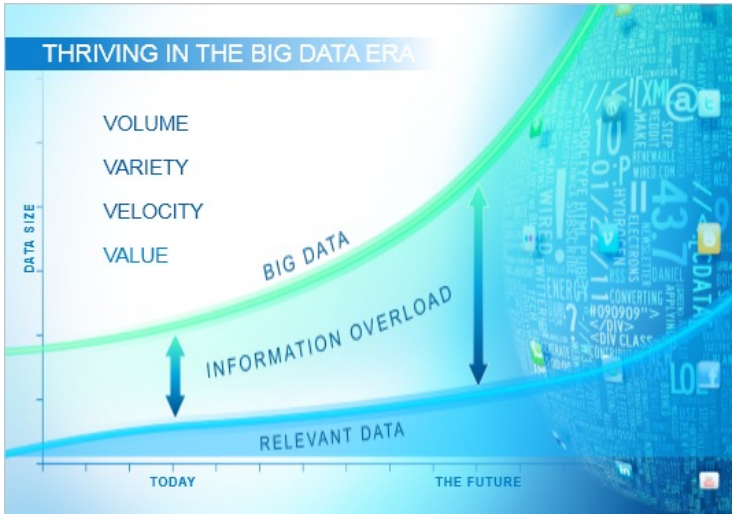
# Big Data: What Is It?

The SAS definition of *big data*:

> The point at which the volume, velocity, and variety of data exceed an organization's storage or computation capacity for accurate and timely decision making

Here are some factors associated with big data:

- data volume
- data velocity
- data variety
- data variability
- data complexity



# Data Volume

Data volumes are increasing due to use of the following:

- social media (Facebook, Twitter, Instagram)
- machines talking to machines
- improvements in the manufacturing process (quality control)
- automated tracking devices
- streaming data feeds

# Data Velocity

- business processes that are more automated
- mergers and acquisitions
- more use of social media
- more use of self-service applications
- integration of business applications

facebook

twitter

# Data Variety

- structured data
- unstructured data
  - business applications
  - unstructured text documents (articles, blogs, and so on)
  - emails
  - digital images
  - video and audio clips
- streaming data
  - stock ticker data
  - RFID tag data
  - sensor data

# Data Variability

- The flow of data changes over time (seasonality, peak response, social media trends, and so on).
- Data values change over time. How much history do you keep?
- Data values are different across data sources.
- Data is stored in different formats.
- Data standards change across time. What was “valid” five years ago might not be “valid” today.

# Data Complexity

Data comes from a variety of systems in a variety of formats. This can make it difficult to merge, cleanse, and transform data in a uniform manner.

> “Analytics is core to success in the digital economy. Data and analytics driven organizations will thrive.”

Chandana Gopal, IDC, December 2017

# Artificial Intelligence

is the science of training systems to emulate human tasks through **learning** and **automation**

Understand Context

Learn Patterns

Recognize Objects

| Learning | Automation | Benefit |
| --- | --- | --- |
| Images | Is this you? | More Secure |
| Transactions | Is it fraud? | Lower Risk |
| Users | Will they buy? | Higher Return |
| Medical Images | Is this healthy? | Better Outcome |
| Languages | Translate? | Reduced Cost |
| Emails | Is it spam? | Better Experience |

# 61%

of organizations identified machine learning and AI as the most significant data initiative for next year

Source: Machine Learning and AI survey, O'Reilly Media and MemSQL, 2018

Tomorrow, AI will impact your industry

Tomorrow, AI will impact your industry

# 3x

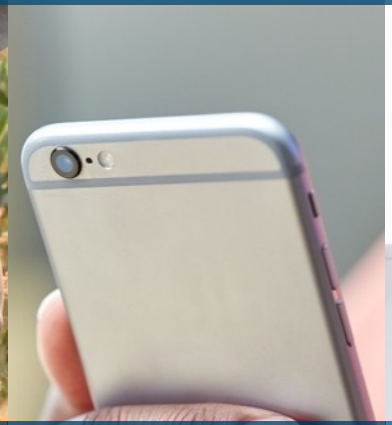
Increased investment in 2017 for AI technology compared to investment in 2016

Source: Business Tech Predictions: 10 Ways AI, Big Data, and Cloud evolved in 2017

# Organizations That Are Using SAS AI

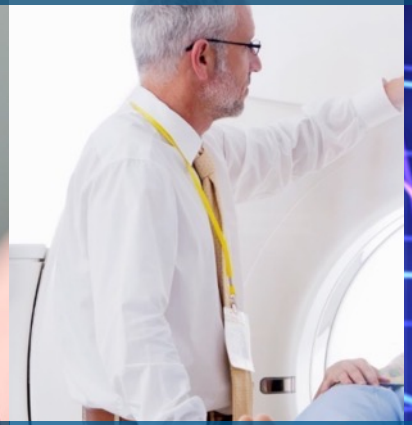
## WildTrack
Data for Good

**90%**
accuracy for ID of wildlife using tracks[5]



## Rogers
Telecom

**53%**
fewer customer complaints[1]



## Amsterdam UMC
Health Care

Improved liver and brain tumor diagnosis with AI and analytics


## Daiwa
Financial

**2.7x**
increase in client purchase rates[4]

## Honda
Manufacturing

Continuous learning and Insight from clients to improve design and quality[3]

# Lesson 1:
# Big Data, Analytics, and Data Science

1.1 Introduction

**1.2 The (Citizen) Data Scientist**

1.3 Skills for the Data Scientist

# What Is a Data Scientist?

Data scientists are a new breed of analytical data expert
who have the technical skills to solve complex problems
and the curiosity to explore what problems need to be solved.

They are part **mathematician**, part **computer scientist**, part **trend spotter**.
They are a **sign of the times**. Their popularity reflects how businesses now think about **big data.**

That unwieldy mass of unstructured information can no longer be ignored and forgotten.
It is a **virtual gold mine** that helps **boost revenue** – as long as there is **someone** who digs in and unearths **business insights** that no one thought to look for before.

## Enter the data scientist

# Typical Job Duties for a Data Scientist

## It is not definitive, but think of …

- Collecting large amounts of unruly data and transforming it into a more usable format
- Solving business-related problems using data-driven techniques
- Working with a variety of programming languages, including SAS, R and Python
- Having a solid grasp of statistics, including statistical tests and distributions
- Staying on top of analytical techniques such as machine learning, deep learning, and text analytics
- Communicating and collaborating with both IT and business
- Looking for order and patterns in data, as well as spotting trends that can help a business's bottom line

# Typical Job Responsibilities for a Data Scientist

- collect large amounts of unruly data and transform it into a more usable format
- solve business-related problems using data-driven techniques
- work with a variety of programming languages (for example, SAS, R, and Python)
- have a solid grasp of statistics, such as statistical tests and distributions
- stay on top of analytical techniques such as social network analysis, text analytics, and new methodologies for predictive modeling
- communicate and collaborate with both IT and business
- look for order and patterns in data

# But …

- There just are not enough data scientists in the workforce.
- it is important to realize one data scientist might not have all the necessary skills.
- it is important to develop a team of data scientists that are “scattered across the business.”
- There is a rise of easier-to-use analytics tools.
- Analytics is so important to society that it cannot be something that is only the domain of experts.

➢ So companies rely on **Citizen Data Scientists**.

(Gartner research director, Alexander Linden, April 2015)

# How to Find Citizen Data Scientists?

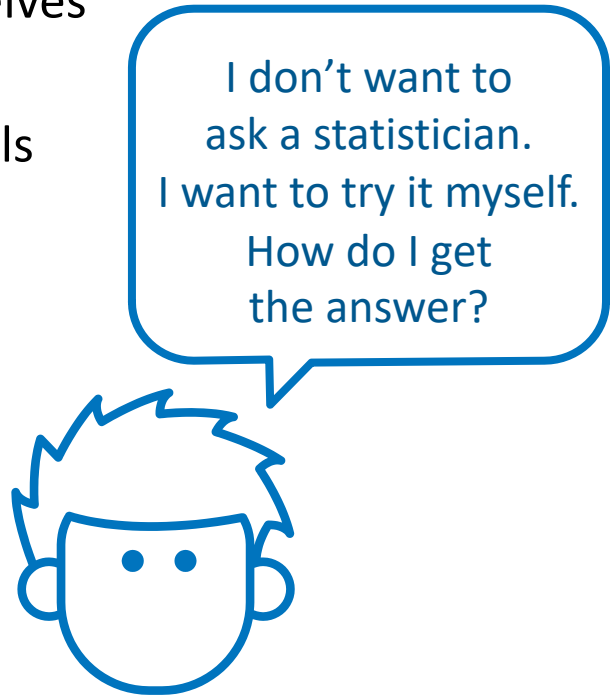
In most organizations, they're already there, working in many different roles and departments throughout the organization. They are citizen data scientists, businesspeople with the right attitude – curious, adventurous, determined – to research and improve things in your organization.

The demand for citizen data scientists will increase **five times more quickly** than the demand for "traditional," highly skilled data scientists.

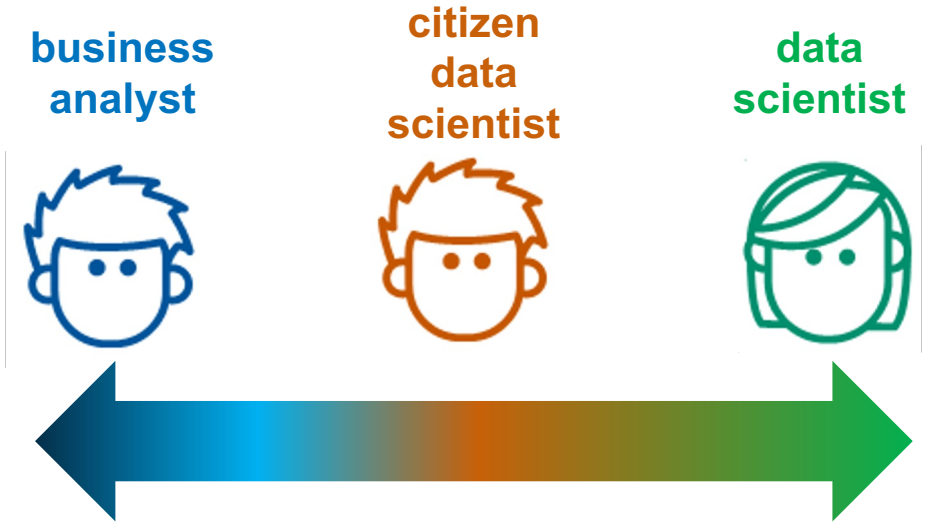
http://www.sas.com/en_us/insights/articles/analytics/how-to-find-and-equip-citizen-data-scientists.html

# Characteristics of Citizen Data Scientists

- tired of looking at the same reports
- want to get their hands on all the data themselves and find new ways to get answers
- willing to learn new methods and use new tools
- analytically minded



# Three Roles Working Together …

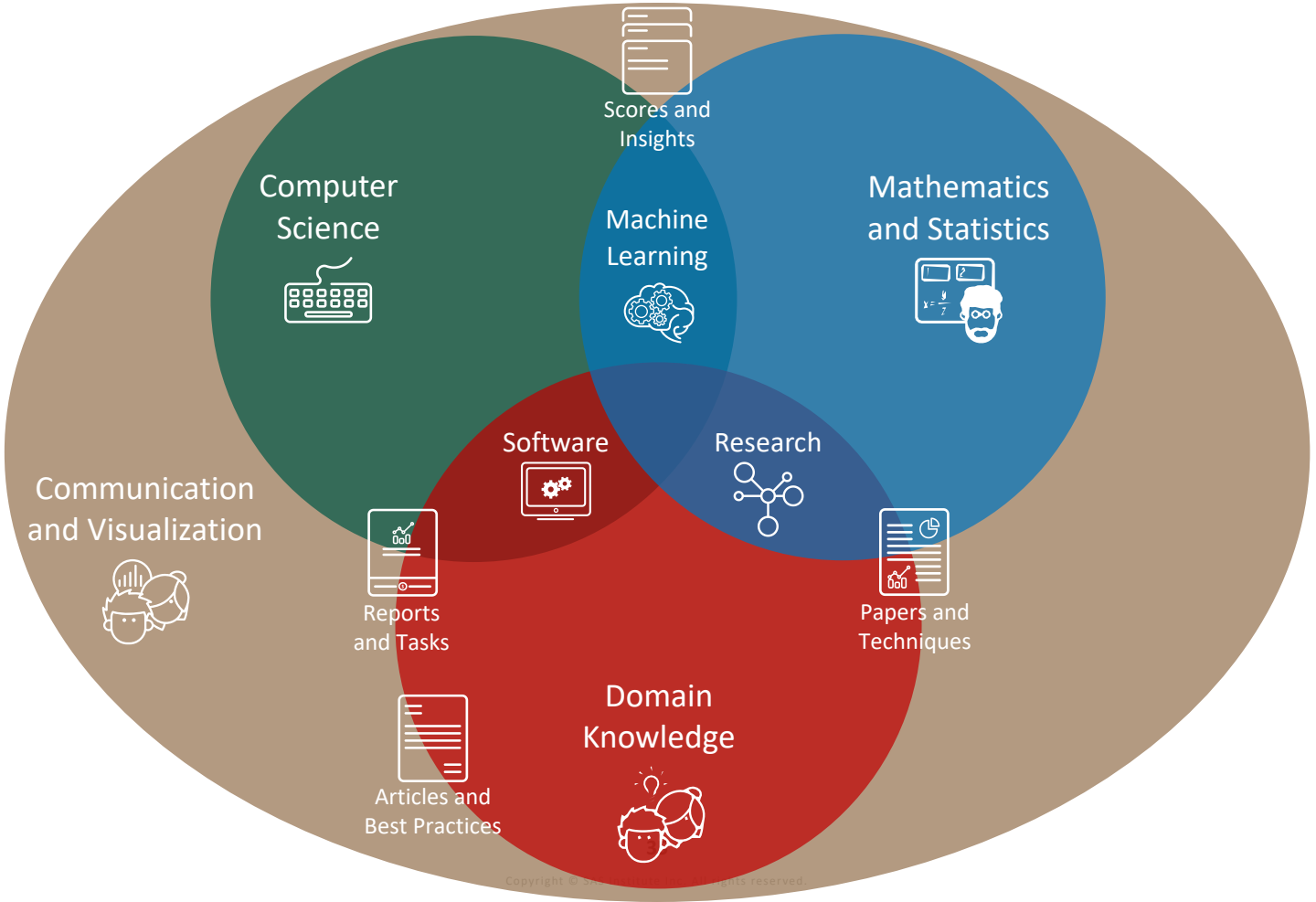



**… from basic discovery to data science …**

# Lesson 1:
# Big Data, Analytics, and Data Science

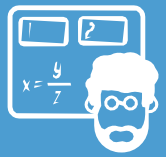
| |
|---|
| 1.1 Introduction |
| 1.2 The (Citizen) Data Scientist |
| **1.3 Skills for the Data Scientist** |

# Data Scientist Skills



# Data Scientist Skills

## Mathematics and Statistics



- Design of Experiments
- Descriptive Statistics
- Statistical Inference
- Supervised Modeling (*Regression, Decision Tree, Forest, Gradient Boosting, Neural Networks, Support Vector Machine, Factorization Machine, Ensemble Models, Two-Stage Models*)
- Unsupervised Modeling (*K-Means, Self-Organizing Maps, Variable Clustering, Principal Components, Association Rules, Sequence, Association, Path Analysis, Link Analysis*)
- Optimization
- Forecasting
- Econometrics
- Text Mining

## Computer Science

- Programming Language
- Statistical Package
- Scripting Language
- Mathematical Package
- Machine Learning Package
- Deep Learning Package
- Data Cleansing
- Data Preparation
- Visualization Tools
- Databases (SQL, NoSQL, Graph)
- Parallel Database and Parallel Query
- Distributed Computing
- Hadoop and Hive
- MapReduce
- Cloud Computing
- Graphical Processing

## Domain Knowledge

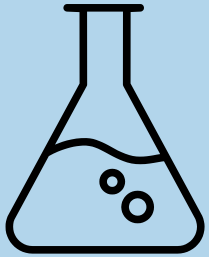
- Business Knowledge
- Data Curiosity
- Analytical Approach
- Problem Solver
- Proactive
- Strategic
- Creative
- Innovative
- Collaborative

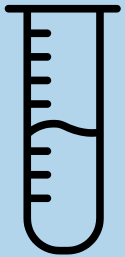
## Communication and Visualization

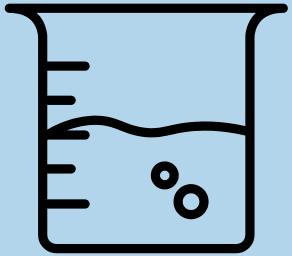
- Engagement with Business and Management Levels
- Translation Insights into Business Decisions and Actions
- Visual Presentation Expertise
- Data Visualization Tools Skills
- Storytelling Capabilities

# Data Scientist Approach

## Science







Math

Statistics

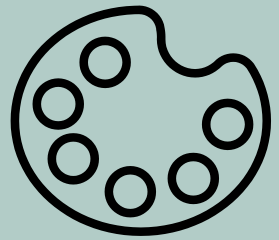
Computer Science

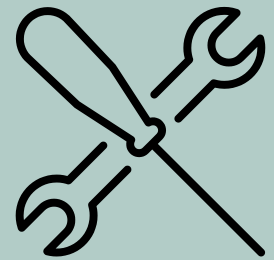
## Art

Creativity

Trial and Error

Invention





# Data Scientist

Harvard Business Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 19"

DATA

## Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT

Big Data: The Management Revolution

# Applied Data Science