Data and Measurement

Kirk Bansak

January 24th, 2023

Logistics

- Be sure to install `R` and `RStudio` this week!

To R!

Data Foundations

Observations

- **The rows in our data**
- These are the units of analysis, or objects of interest, that we care about.
- Could be comprised of individual entities, such as individual people, animals, etc.
- Could be more aggregate units of analysis: schools, districts, etc.
- Defines the scope and nature of the analysis that will be performed

Observations

- **The rows in our data**
- These are the units of analysis, or objects of interest, that we care about.
- Could be comprised of individual entities, such as individual people, animals, etc.
- Could be more aggregate units of analysis: schools, districts, etc.
- Defines the scope and nature of the analysis that will be performed
- Terminology (often used interchangeably, though sometimes context-dependent):
  - Observation
  - Case
  - Instance
  - Individual
  - Record
  - ...

Variables

- **The columns in our data**
- These are attributes, qualities, or characteristics of objects
- Variables are what describe the observations in one's data
- Etymology: "Variable" from Latin word for "capable of changing"

Variables

- **The columns in our data**
- These are attributes, qualities, or characteristics of objects
- Variables are what describe the observations in one's data
- Etymology: "Variable" from Latin word for "capable of changing"
- Terminology
  1. For primary variables of interest (that we want to explain or predict):
     - Outcomes
     - Labels
     - Dependent Variables
     - Target Variables
     - Output Variables

Variables

- **The columns in our data**
- These are attributes, qualities, or characteristics of objects
- Variables are what describe the observations in one's data
- Etymology: "Variable" from Latin word for "capable of changing"
- Terminology
  1. For primary variables of interest (that we want to explain or predict):
     - Outcomes
     - Labels
     - Dependent Variables
     - Target Variables
     - Output Variables
  2. For other variables used to make predictions (or for other analyses):
     - Predictors
     - Features
     - Independent Variables
     - Input Variables
     - Covariates

Data Set

- A collection of data is referred to as a data set
- "Clean" or structured data sets typically arranged in tables (a.k.a. data frames)

A Data Set of Classic Books

ID	Title	Author	Year	Cover	Edition	Price
1	Emma	Austen	1815	Paperback	20	5.75
2	Dracula	Stoker	1897	Hardback	15	12.00
3	Ivanhoe	Scott	1820	Hardback	8	25.00
4	Kidnapped	Stevenson	1886	Paperback	11	5.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Variable Types (Mathematical/Conceptual View)

Variable Types (Mathematical/Conceptual View)

1. Quantitative/Numeric
   - For characteristics that can naturally be represented by numbers
   - Can be integers, continuous numbers (e.g. real numbers), percentage or proportion (i.e. measurements with natural zero)
   - Usual practice is to treat quantitative variables purely as numbers, without units being given explicitly in the data values (See Codebooks!)

Variable Types (Mathematical/Conceptual View)

1. Quantitative/Numeric
   - For characteristics that can naturally be represented by numbers
   - Can be integers, continuous numbers (e.g. real numbers), percentage or proportion (i.e. measurements with natural zero)
   - Usual practice is to treat quantitative variables purely as numbers, without units being given explicitly in the data values (See Codebooks!)
2. Categorical/Nominal
   - For characteristics comprised of classes, categories, etc.
   - When used for rigorous analysis, should be a fixed set of possibilities, known as levels (possible categories/values)
   - Sometimes represented as numbers for convenience but should not be treated as such (See Codebooks!)

Variable Types (Mathematical/Conceptual View)

Variable Types (Mathematical/Conceptual View)

- Ordinal
  - Categorical values, but ranked/ordered
  - e.g. Likert scale
    - Strongly disagree
    - Disagree
    - Neither agree nor disagree
    - Agree
    - Strongly agree

Variable Types (Mathematical/Conceptual View)

- Ordinal
  - Categorical values, but ranked/ordered
  - e.g. Likert scale
    - Strongly disagree
    - Disagree
    - Neither agree nor disagree
    - Agree
    - Strongly agree
- Indicator/Binary
  - Variable measuring/indicating whether a case belongs to a particular category or not
  - Often coded as 1/0
  - Multi-class categorical variables often transformed into a set of indicator ("dummy") variables for analysis

Variable Types: Conceptual View vs. Computer's View

- Sometimes need to be careful to distinguish between a variable's conceptual/mathematical "type" and its representation/encoding on the computer
- Representation/encoding is how it appears in a data set and is viewed by the computer (i.e. its `class` according to `R`)...
- But its conceptual type determines how it should be analyzed

Variable Types: Conceptual View vs. Computer's View

- For instance, consider a categorical variable that is encoded as numeric/integers:

```
> favorite_color
[1] 1 5 2 6 3 2 1 3 4 2

> class(favorite_color)
[1] "numeric"

> mean(favorite_color)
[1] 2.9
```

Variable Types: Conceptual View vs. Computer's View

- Or a binary indicator that is encoded as character:

```
> won_election
[1] "Yes" "No"  "Yes" "Yes" "Yes" "No"  "No"  "Yes"

> class(won_election)
[1] "character"

> mean(won_election)
[1] NA
Warning message: In mean.default(won_election) :
  argument is not numeric or logical: returning NA

> mean(won_election == "Yes")
[1] 0.625
```

Type vs. Representation: Examples of Possible Confusion

Type vs. Representation: Examples of Possible Confusion

- A categorical variable might be coded as integers for convenience (e.g. 1 : Alabama, 2 : Alaska, 3 : Arizona, ...)

Type vs. Representation: Examples of Possible Confusion

- A categorical variable might be coded as integers for convenience (e.g. 1 : Alabama, 2 : Alaska, 3 : Arizona, ...)
- An ordinal variable is often mapped onto quantitative scale and analyzed accordingly (e.g. Strongly disagree $\rightarrow$ -2, Disagree $\rightarrow$ -1, Neither agree nor disagree $\rightarrow$ 0, Agree $\rightarrow$ 1, Strongly agree $\rightarrow$ 2)

Type vs. Representation: Examples of Possible Confusion

- A categorical variable might be coded as integers for convenience (e.g. 1 : Alabama, 2 : Alaska, 3 : Arizona, ...)
- An ordinal variable is often mapped onto quantitative scale and analyzed accordingly (e.g. Strongly disagree $\rightarrow$ -2, Disagree $\rightarrow$ -1, Neither agree nor disagree $\rightarrow$ 0, Agree $\rightarrow$ 1, Strongly agree $\rightarrow$ 2)
  - Caution: This is somewhat reasonable given the ordering, but there is no natural or necessarily uniform "distance" between the values

Type vs. Representation: Examples of Possible Confusion

- A categorical variable might be coded as integers for convenience (e.g. 1 : Alabama, 2 : Alaska, 3 : Arizona, ...)
- An ordinal variable is often mapped onto quantitative scale and analyzed accordingly (e.g. Strongly disagree $\rightarrow$ -2, Disagree $\rightarrow$ -1, Neither agree nor disagree $\rightarrow$ 0, Agree $\rightarrow$ 1, Strongly agree $\rightarrow$ 2)
  - Caution: This is somewhat reasonable given the ordering, but there is no natural or necessarily uniform "distance" between the values
- Numeric variables are often dichotomized into binary variables (e.g. above average income vs. below average income) or converted into set of indicator variables (e.g. several age groups)

Type vs. Representation: Examples of Possible Confusion

- A categorical variable might be coded as integers for convenience (e.g. 1 : Alabama, 2 : Alaska, 3 : Arizona, ...)
- An ordinal variable is often mapped onto quantitative scale and analyzed accordingly (e.g. Strongly disagree $\rightarrow$ -2, Disagree $\rightarrow$ -1, Neither agree nor disagree $\rightarrow$ 0, Agree $\rightarrow$ 1, Strongly agree $\rightarrow$ 2)
  - Caution: This is somewhat reasonable given the ordering, but there is no natural or necessarily uniform "distance" between the values
- Numeric variables are often dichotomized into binary variables (e.g. above average income vs. below average income) or converted into set of indicator variables (e.g. several age groups)
- Binary variables could be treated as numeric (1/0) or categorical without changing results

Type vs. Representation: Examples of Possible Confusion

- A categorical variable might be coded as integers for convenience (e.g. 1 : Alabama, 2 : Alaska, 3 : Arizona, ...)
- An ordinal variable is often mapped onto quantitative scale and analyzed accordingly (e.g. Strongly disagree $\rightarrow$ -2, Disagree $\rightarrow$ -1, Neither agree nor disagree $\rightarrow$ 0, Agree $\rightarrow$ 1, Strongly agree $\rightarrow$ 2)
  - Caution: This is somewhat reasonable given the ordering, but there is no natural or necessarily uniform "distance" between the values
- Numeric variables are often dichotomized into binary variables (e.g. above average income vs. below average income) or converted into set of indicator variables (e.g. several age groups)
- Binary variables could be treated as numeric (1/0) or categorical without changing results
- A variable comprised of numbers might have such a limited number of values that it makes more sense to treat it as a categorical variable (e.g. election year for a sample of Senators)

Character vs. Factor Variables in R

- Both "character" and "factor" variable types can be used to represent categorical data comprised of strings in R.
- While factors look (and often behave) like character vectors, they are actually stored as integers under the hood (with character labels associated with the unique integers).
- Once created, factors can only contain a pre-defined set of values, known as levels.
- Strings are converted to factor variables by default in many standard functions in R, including `data.frame()`, `read.csv()`, and `read.table()`.
- Behavior of character and factor variables can vary in important ways, so always be aware of your variable class in R!

A Data Set of Classic Books

ID	Title	Author	Year	Cover	Edition	Price
1	Emma	Austen	1815	Paperback	20	5.75
2	Dracula	Stoker	1897	Hardback	15	12.00
3	Ivanhoe	Scott	1820	Hardback	8	25.00
4	Kidnapped	Stevenson	1886	Paperback	11	5.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮

“Tidy Data”

Hadley Wickham: “Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).”

"Tidy Data"

Hadley Wickham: "Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)."

Criteria for "tidy data":

1. Each variable forms a column.
2. Each observation (case) forms a row.
3. Each type of observational unit forms a table.
   - i.e. rows of the table should correspond to a constant unit of analysis

"Tidy Data"

Hadley Wickham: "Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)."

Criteria for "tidy data":

1. Each variable forms a column.
2. Each observation (case) forms a row.
3. Each type of observational unit forms a table.
   - i.e. rows of the table should correspond to a constant unit of analysis

I would also add:

- ⇝ Each variable has a constant type and representation

"Tidy Data"

Hadley Wickham: "Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)."

Criteria for "tidy data":

1. Each variable forms a column.
2. Each observation (case) forms a row.
3. Each type of observational unit forms a table.
   - i.e. rows of the table should correspond to a constant unit of analysis

I would also add:

- ⇝ Each variable has a constant type and representation

"Messy data" is any other other arrangement of the data. "Tidy datasets are all alike, but every messy dataset is messy in its own way."

Key Considerations and Common Challenges

Population vs. Sample

Population vs. Sample

- Population
  - A population is the set of all the possible objects or units which might have been included in the collection
  - Does not only refer to people; can refer to anything

Population vs. Sample

- Population
  - A population is the set of all the possible objects or units which might have been included in the collection
  - Does not only refer to people; can refer to anything
- Sample
  - Selection/subset of cases from the population that you actually have data on
  - "Sample size" refers to the number cases in the sample
  - Usually a relatively small proportion of the population
  - If the sample is the full population, then it is a "census"

Sampling

- Goal: Usually to understand or find something out about a broad population
- Problem: It can be expensive, impossible, or destructive to collect information on an entire population
- Solution: Sampling, or the act of collecting a sample
- Challenge: External validity (is what you can learn about the sample reflective of the truth about the population?)
- Types of samples (not exhaustive)
  - Simple random sample
  - Convenience samples

Size

- Generally, the more data (cases and variables) the better

Size

- Generally, the more data (cases and variables) the better
- More cases means having access to a larger proportion of the population, which means ability to estimate things about that population more precisely

Size

- Generally, the more data (cases and variables) the better
- More cases means having access to a larger proportion of the population, which means ability to estimate things about that population more precisely
- In the realm of prediction, the amount of data has implications for the signal vs. noise problem
  - Relationships between variables are governed by signal (systematic interconnection) and noise (randomness, incidental correlation)
  - More cases increases ability to detect more signal (i.e. distinguish it from noise)
  - More variables creates more signal to detect

Size

- Generally, the more data (cases and variables) the better
- More cases means having access to a larger proportion of the population, which means ability to estimate things about that population more precisely
- In the realm of prediction, the amount of data has implications for the signal vs. noise problem
  - Relationships between variables are governed by signal (systematic interconnection) and noise (randomness, incidental correlation)
  - More cases increases ability to detect more signal (i.e. distinguish it from noise)
  - More variables creates more signal to detect
- But downsides to collecting more cases and measuring more variables
  - Costs time and money
  - Adding nonsense variables could make things worse through (a) distraction, (b) computational tractability, (c) adding noise
  - And adding more cases in a nonrepresentative manner will damage external validity

Common Problems

- Data Missingness (variable values missing for some cases)
  - Failure to measure
  - Non-response bias
- Mismeasurement
  - Entry errors or typos
  - Mechanical/natural mismeasurement
- Atypical Values
  - Outliers (quantitative)
  - Sparsity (categorical)
- Inconsistent variable coding
  - Variable unit of measurement
  - Variable type

Always Consider Representativeness of Data in Sample

- Analysis of the data can always be useful for learning about the sample itself
- But must think carefully about sample representativeness to consider what can be learned about the population

Always Consider Representativeness of Data in Sample

- Analysis of the data can always be useful for learning about the sample itself
- But must think carefully about sample representativeness to consider what can be learned about the population
- Nonrepresentativeness of bias in the sample could be the result of:
  - Imperfect sampling strategy by the researcher
  - Non-response bias / non-random data missingness
  - Self-selection bias into the sample
- These issues are notoriously thorny in the case of surveys of people
  - e.g. Political polls or surveys

Bonus Slides: Human Discretion with Data

The Three Stages of a Variable

Three stages or versions of a variable to distinguish between:

The Three Stages of a Variable

Three stages or versions of a variable to distinguish between:

1. Conceptual
   - Ideal, theoretical quantity or phenomenon the researcher is interested in
   - Often impossible to directly or perfectly measure
   - It is a choice what to care about conceptually

The Three Stages of a Variable

Three stages or versions of a variable to distinguish between:

1. Conceptual
   - Ideal, theoretical quantity or phenomenon the researcher is interested in
   - Often impossible to directly or perfectly measure
   - It is a choice what to care about conceptually
2. Operational
   - More specific, concrete quantity or phenomenon used to serve as a proxy for ("operationalize") the conceptual variable of interest
   - Must be measurable in some practical manner
   - Must determine how to best capture the conceptual variable of interest with a measurable phenomenon

The Three Stages of a Variable

Three stages or versions of a variable to distinguish between:

1. Conceptual
   - Ideal, theoretical quantity or phenomenon the researcher is interested in
   - Often impossible to directly or perfectly measure
   - It is a choice what to care about conceptually
2. Operational
   - More specific, concrete quantity or phenomenon used to serve as a proxy for ("operationalize") the conceptual variable of interest
   - Must be measurable in some practical manner
   - Must determine how to best capture the conceptual variable of interest with a measurable phenomenon
3. Actualized
   - Precise quantity that is actually measured such that the operational variable can be systematically encoded into data
   - This is what is actually present in the dataset
   - Must decide how to measure the operational variable, in the real world, subject to practical requirements and constraints

Measuring Variables

1. Conceptual
   - Ex. 1 Size
   - Ex. 2 Health
   - Ex. 3 Happiness
2. Operational
   - Ex. 1 ...
   - Ex. 2 ...
   - Ex. 3 ...
3. Actualized
   - Ex. 1 ...
   - Ex. 2 ...
   - Ex. 3 ...

Measuring Variables

1. Conceptual
   - Ex. 1 Size
   - Ex. 2 Health
   - Ex. 3 Happiness
2. Operational
   - Ex. 1 Height
   - Ex. 2 ...
   - Ex. 3 ...
3. Actualized
   - Ex. 1 Height measured in feet
   - Ex. 2 ...
   - Ex. 3 ...

Measuring Variables

1. Conceptual
   - Ex. 1 Size
   - Ex. 2 Health
   - Ex. 3 Happiness
2. Operational
   - Ex. 1 Height
   - Ex. 2 Comorbidity score (number of active chronic conditions)
   - Ex. 3 ...
3. Actualized
   - Ex. 1 Height measured in feet
   - Ex. 2 Sum of the number of a patient's conditions that are classified as "active" and belong to the reference dictionary of "chronic conditions"
   - Ex. 3 ...

Measuring Variables

1. Conceptual
   - Ex. 1 Size
   - Ex. 2 Health
   - Ex. 3 Happiness
2. Operational
   - Ex. 1 Height
   - Ex. 2 Comorbidity score (number of active chronic conditions)
   - Ex. 3 Self-reported happiness score
3. Actualized
   - Ex. 1 Height measured in feet
   - Ex. 2 Sum of the number of a patient's conditions that are classified as "active" and belong to the reference dictionary of "chronic conditions"
   - Ex. 3 Self-reported level of happiness on a scale of 1 to 10, where 1 denotes "extremely unhappy" and 10 denotes "extremely happy"

The differences are not simply an issue of objectivity vs. subjectivity. Actualized variables can be quite subjective.

Measurement

- Going from the conceptual to the operational and finally to the actual involves decisions on measurement!

Measurement

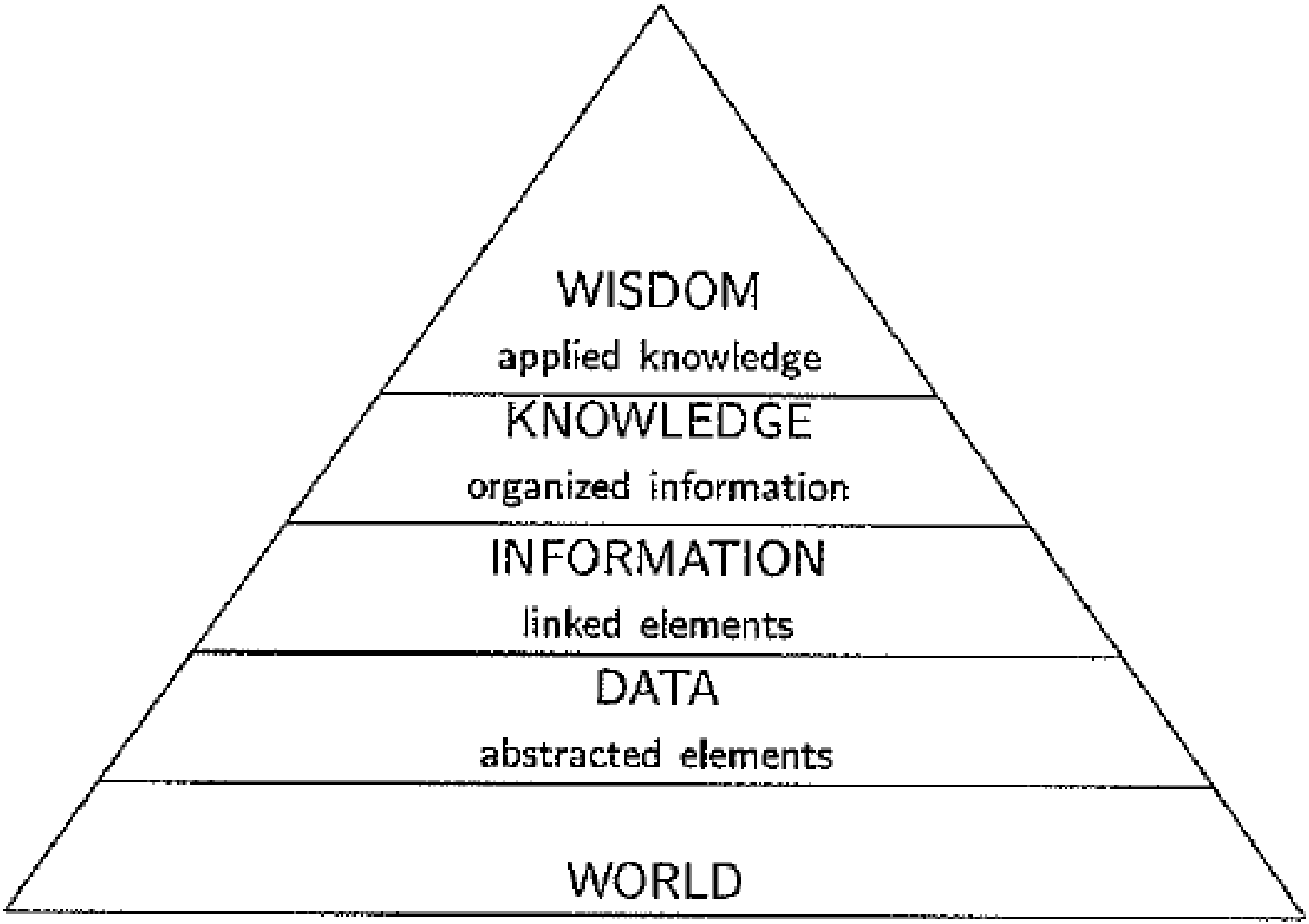
- Going from the conceptual to the operational and finally to the actual involves decisions on measurement!
- Sometimes easily justifiable choices:
  - You know it when you see it
  - Limitations on what is possible
  - The conceptual is the operational is the actual

Measurement

- Going from the conceptual to the operational and finally to the actual involves decisions on measurement!
- Sometimes easily justifiable choices:
  - You know it when you see it
  - Limitations on what is possible
  - The conceptual is the operational is the actual
- Other times, you may not know if what you are measuring is actually achieving the goal
- Establishing validity of measurement is often subjective

Measurement

- Going from the conceptual to the operational and finally to the actual involves decisions on measurement!
- Sometimes easily justifiable choices:
  - You know it when you see it
  - Limitations on what is possible
  - The conceptual is the operational is the actual
- Other times, you may not know if what you are measuring is actually achieving the goal
- Establishing validity of measurement is often subjective
- Andrew Gelman: "the #1 neglected topic in statistics is measurement."


WISDOM
applied knowledge
KNOWLEDGE
organized information
INFORMATION
linked elements
DATA
abstracted elements
WORLD

A Day in the Life of a Data Scientist

According to 2016 survey of data scientists, this is the breakdown of time spent on different tasks:

- Data Collection: 19%
- Cleaning and Organizing Data: 60%
- Building Training Sets: 3%
- Mining Data for Patterns: 9%
- Refining Algorithms: 4%
- Performing Other Tasks: 5%

Install R and RStudio!!!