Defining Data Science

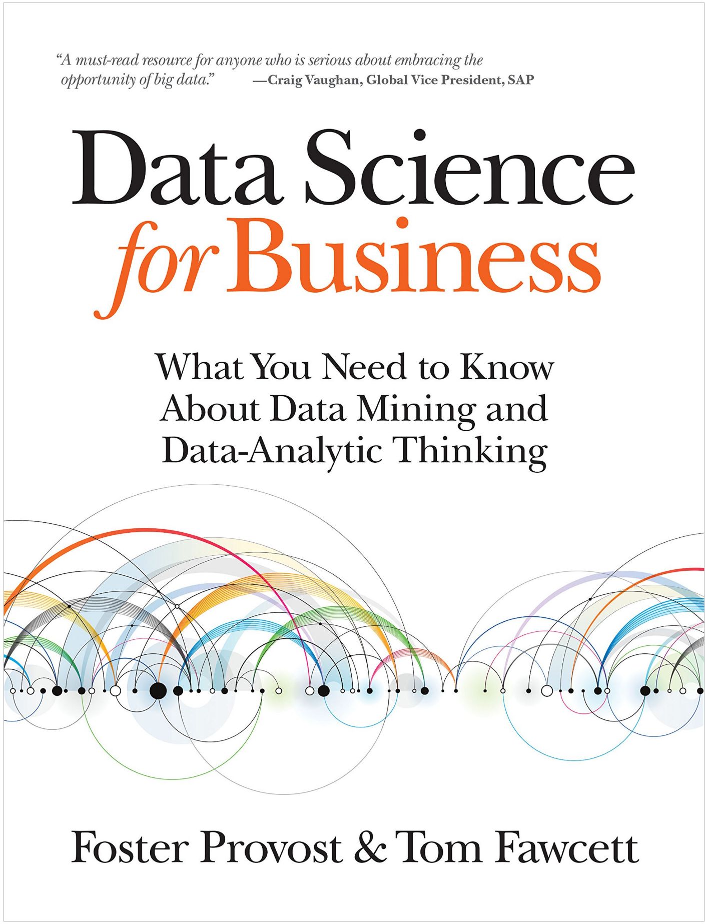
⚗️

1201 Data Science: Tim Raiswell

2018/10/15

Part 1: Data Science and Statistics

The Source of A Lot of Tonight's Content



"A must-read resource for anyone who is serious about embracing the opportunity of big data." —Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking

Foster Provost & Tom Fawcett

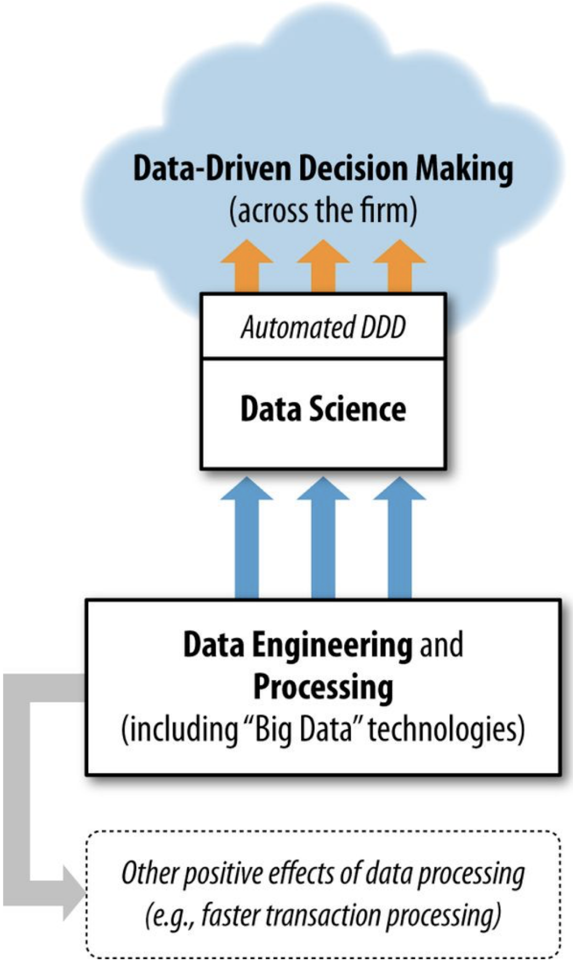
Caveat Analyticum

It impossible to know everything about data science. Learn the basics, then find the things that are useful and the things that interest you.

A Practical Definition

> Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.
>
> — Josh Wills (@josh_wills) May 3, 2012

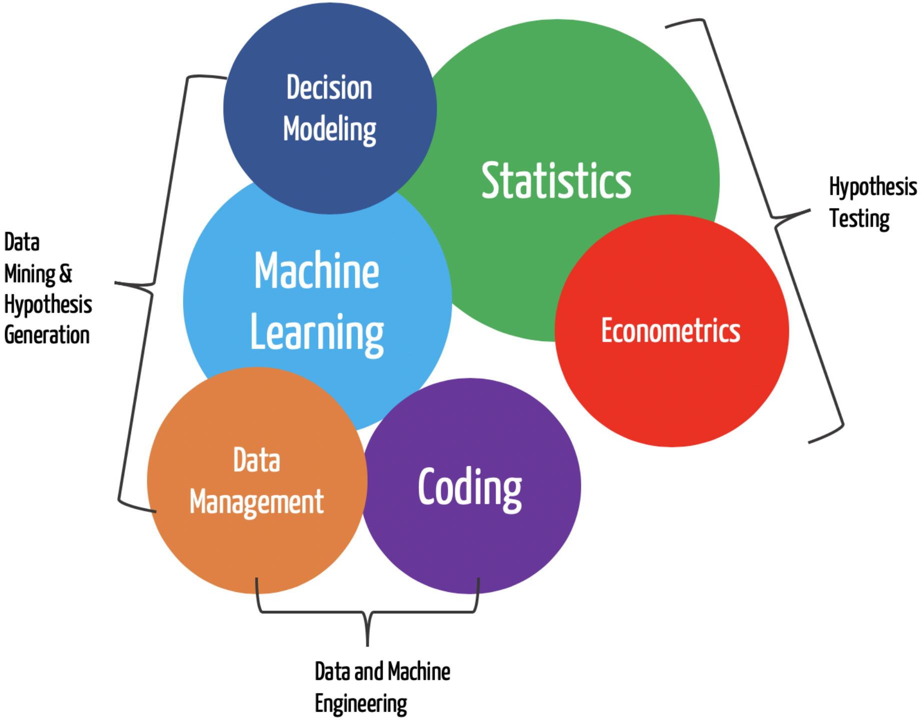
The Data Science Process


Data-Driven Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing
(including "Big Data" technologies)
Other positive effects of data processing
(e.g., faster transaction processing)

The CRISP Data Mining Process

> "Extracting useful knowledge from data to solve business problems can be treated systematically by following a process with reasonably well-defined stages." Foster Provost, *Data Science for Business*

Data Science



Putting it Together

Data Mining

1. Relies on data mining to source and analyze sometimes very large amounts of data;
2. Often concerned with prediction to improve a decision outcome;
3. In a business context, data science implies a mature infrastrucure to procure, analyze and deploy data analytics;
4. Should be 100% scientific in approach but is often criticised by statisticians for playing fast and loose with machine learning approaches;

Applied Statistics

1. Relies on traditional data gathering tools like experimentation and surveying;
2. Often concerned with causality to improve a medical or policy outcome;
3. In a business context, statistician is often an advisorial and *ad hoc* role related to bringing a scientific approach to decision-making;
4. Should be 100% scientific in approach but, like data science, is only as good as the person running the experiment.

The Data Science Skillset

> "Although data mining involves software, it also requires skills that may not be common among programmers. In software engineering, the ability to write efficient, high-quality code from requirements may be paramount. Team members may be evaluated using software metrics such as the amount of code written or number of bug tickets closed. In analytics, it's more important for individuals to be able to formulate problems well, to prototype solutions quickly, to make reasonable assumptions in the face of ill-structured problems, to design experiments that represent good investments, and to analyze results. In building a data science team, these qualities, rather than traditional software engineering expertise, are skills that should be sought."[1]

[1] Provost & Fawcett

Five Reasons I ❤️ Data Science

Five Reasons I ❤️ Data Science

1. It is an incredibly diverse global community;

Five Reasons I ❤️ Data Science

1. It is an incredibly diverse global community;
2. Most of the useful tools and frameworks are open source and free;

Five Reasons I ❤️ Data Science

1. It is an incredibly diverse global community;
2. Most of the useful tools and frameworks are open source and free;
3. It is evolving quickly and changes every day and week;

Five Reasons I ❤️ Data Science

1. It is an incredibly diverse global community;
2. Most of the useful tools and frameworks are open source and free;
3. It is evolving quickly and changes every day and week;
4. It is a discovery-driven process;

Five Reasons I ❤️ Data Science

1. It is an incredibly diverse global community;
2. Most of the useful tools and frameworks are open source and free;
3. It is evolving quickly and changes every day and week;
4. It is a discovery-driven process;
5. Everything is data. 🎧🎮🚗$⚽🏈⏰📡🔮

   There is something for everyone.

Some Important Foundations

Machine Learning

Supervised Learning: Predicting a target. For example, "What is the probability that customer will default on their loan in the next 12 months?" The target is `loan-default-in-12-months`.

Unsupervised Learning: Targetless exploration. For example, "Are there natural groups of similar customers in our CRM system?" There is no target. In fact by running a clustering algorithm we may not find anything useful.

Data and Model Types

Categorical Variables: A finite number of discrete values. The type nominal denotes that there is no ordering between the values, such as last names and colors. The type ordinal denotes that there is an ordering, such as in an attribute taking on the values `low, medium, or high`, e.g.

`{man, woman}, {low income, middle income, high income}, {dead, alive}, {cat, dog}, {good, bad}`.

If the target of a machine learning model is categorical, it is a *classification* model.

Continuous Variables: Commonly, subset of real numbers, where there is a measurable difference between the possible values. Integers are usually treated as continuous in practical problems, e.g.

Prices in dollars `{25, 34.2, 100.75}`, height of adults in inches: `{74, 70, 81.5}.`

If the target of a machine learning model is continuous, it is a *regression* model.

Model Types: Classification

Will something happen?

Classification is named as such because it is an attempt to predict class membership.The most common business example is churn prediction.

Example
What is the probability that a given customer will churn within the next 12 months?

A machine model like a classification tree would be trained on data of customers who

- Class 1: Did churn within 12 months
- Class 2: Did *not* churn

Model Types: Regression

How much will something happen?

Regression is a quick way of saying "value estimation".

Example
How much bandwidth will a customer use in 6 months?

A machine model like a regression tree would be trained on data of customers with varying bandwidth use. The model could be fed the details of a new customer and predict how much bandwidth they will consume in the next 12 months.

Regression 1 of 4

```
# we will start with a little regression because most of us
# have dealt with it in the past.
head(cars, 10) # type this code into your R console
```

```
#    speed dist
# 1      4    2
# 2      4   10
# 3      7    4
# 4      7   22
# 5      8   16
# 6      9   10
# 7     10   18
# 8     10   26
# 9     10   34
# 10    11   17
```

```
# Now this code, and hit enter
# ?cars don't include the hash in front of yours
```

Regression 2 of 4

```r
# type this code into your R console
fit <- lm(dist ~ speed, data = cars) # this first, hit enter
coef(summary(fit)) # then this, hit enter
```

```
#                Estimate Std. Error   t value             Pr(>|t|)
# (Intercept) -17.579095  6.7584402 -2.601058 0.012318816153809090
# speed         3.932409  0.4155128  9.463990 0.000000000001489836
```

Regression 2 of 4

```
# type this code into your R console
fit <- lm(dist ~ speed, data = cars) # this first, hit enter
coef(summary(fit)) # then this, hit enter
```

```
#               Estimate Std. Error   t value             Pr(>|t|)
# (Intercept) -17.579095  6.7584402 -2.601058 0.012318816153809090
# speed         3.932409  0.4155128  9.463990 0.000000000001489836
```

Congratulations! You just built a regression model in R.

Regression 3 of 4

Now let's actually predict something with our model.

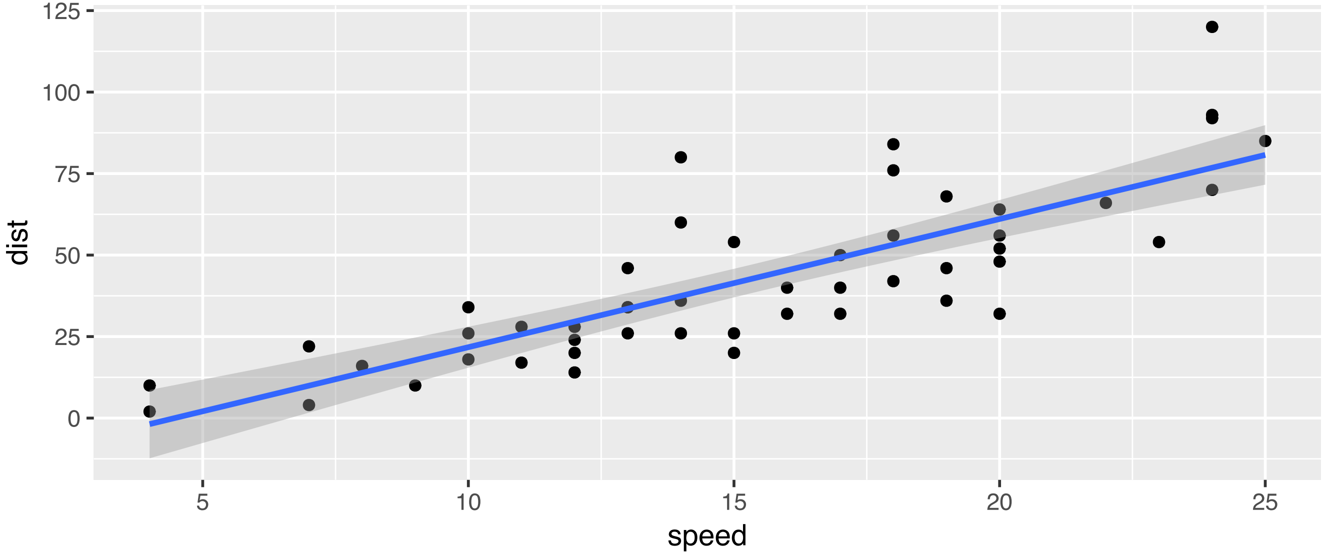
```
new.speeds <- data.frame(
  speed = c(12, 19, 24, 50)
)

predict(fit, newdata = new.speeds)
```

```
##         1         2         3         4 
##  29.60981  57.13667  76.79872 179.04134
```

Regression 4 of 4

```
#This is a scatterplot of the data visualizing the relationships
# in the model. The target variable always goes on the y-axis.
library(ggplot2)
ggplot(cars, aes(speed, dist)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Classification 1 of 4

```
head(iris, 5) # type this code into your R console
```

```
#   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
# 1          5.1         3.5          1.4         0.2  setosa
# 2          4.9         3.0          1.4         0.2  setosa
# 3          4.7         3.2          1.3         0.2  setosa
# 4          4.6         3.1          1.5         0.2  setosa
# 5          5.0         3.6          1.4         0.2  setosa
```

```
# what classes are we predicting?
levels(iris$Species) # enter this code and hit enter
```

```
## [1] "setosa"     "versicolor" "virginica"
```

Classification 2 of 4

We'll shrink the problem to two classes for demo purposes.

```
# enter these lines one after the other. Hit 'enter' after each.
iris.small <- dplyr::filter(iris,
                            Species %in%
                              c("virginica", "versicolor"))

# logistic regression
glm.out <- glm(Species ~ Sepal.Width + Petal.Width + Petal.Length,
               data = iris.small,
               family = "binomial")

coef(summary(glm.out)) # take a look at the coefficients.
```

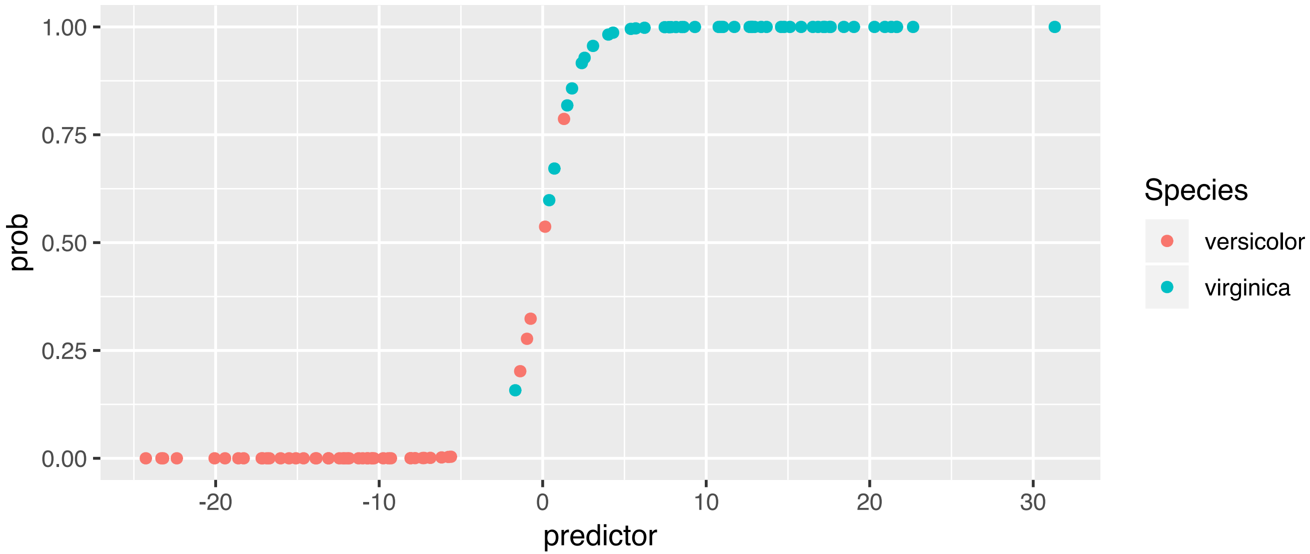
```
##                 Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)   -50.526834  23.994601 -2.105758 0.03522534
## Sepal.Width    -8.376070   4.761147 -1.759255 0.07853428
## Petal.Width    21.429592  10.707268  2.001406 0.04534864
## Petal.Length    7.874547   3.840637  2.050323 0.04033292
```

Classification 3 of 4

Now we'll look at how well the model seaparates the two classes

```
lr_data <- data.frame(predictor = glm.out$linear.predictors,
                      prob = glm.out$fitted.values,
                      Species = iris.small$Species)

ggplot(lr_data, aes(x = predictor, y = prob, color=Species)) +
  geom_point()
```



Classification 4 of 4

A Quick Explanatory Note on Logistic Regression

- Logistic regression models a binary response variable: `0, 1`
- It is called regression but it's really a classifiction algorithm;
- It is referred to as a regression because the model uses a link function (a logit function) to constrain the fitted values of the regression to always lie between 0 and 1.
- Our logistic model answers this question: **Given some linear combination of the predictors (petal widths etc.), what are the odds of the dependent variable being a `virginica iris`? (and conversely a `versicolor iris`?)

Part 2: Data Exploration with R