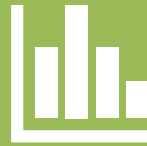


Tonight's Agenda



Discussion



Data Science in
Context
Presentations



Assignment 2
student solutions

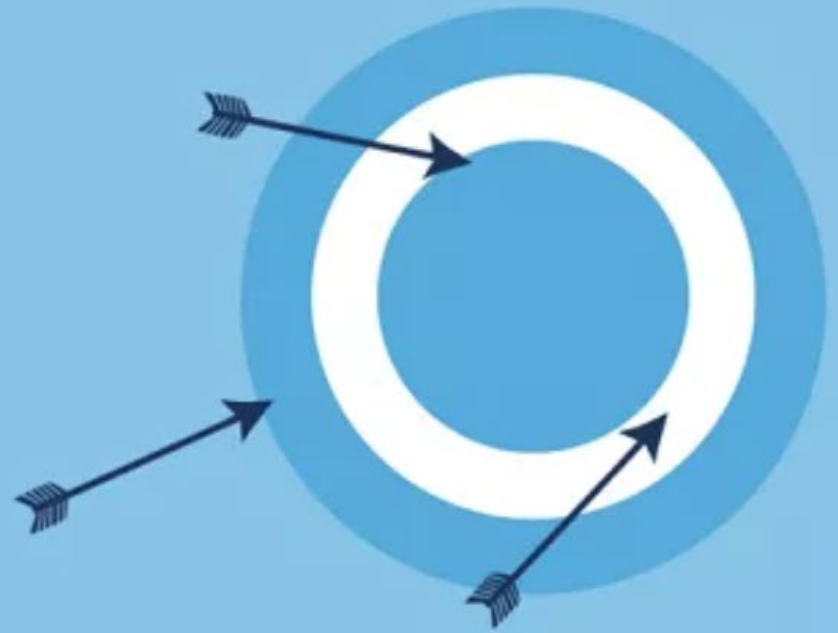
Discussion

Is it better to be a generalist or a specialist?

What are traits of the best data scientists?

R and SQL

Resources for SQL and Regular Expressions



Is it better to be a
generalist or a
specialist?

- Source: cleverism.com

Is it better
to be a
generalist
or a
specialist?



Suppose you work on a team of hundreds of data scientists at American Express or UPS?



Suppose you work in a small start-up, where you need to assume several roles?

Analyzing the Analyzers

An Introspective Survey
of Data Scientists and Their Work



Barlan D. Harris, Sean Patrick Murphy
& Marck Vaisman

Is it better to be a generalist or a specialist?

“The most successful data scientists are those with substantial, deep expertise in at least one aspect of data science, be it statistics, big data, or business communication.

In many ways, this pattern matches the “T-shaped skills” idea that has been promoted since at least the early 1990s... The “T” represents breadth of skills, across the top, with depth in one area represented by the vertical bar. T-shaped professionals can more easily work in interdisciplinary teams than those with less breadth and can be more effective than those without depth. Data science is an inherently collaborative and creative field, where the successful professional can work with database administrators, business people, and others with overlapping skill sets to get data projects completed in innovative ways.”

Source: <https://www.oreilly.com/data/free/analyzing-the-analyzers.csp>

#1 NEW YORK TIMES BESTSELLER

RANGE

WHY GENERALISTS TRIUMPH
IN A SPECIALIZED WORLD



"I loved RANGE."
—Malcolm Gladwell

DAVID EPSTEIN

AUTHOR OF THE SPORTS GENE

Is it better for everyone on a data science team to have similar backgrounds?

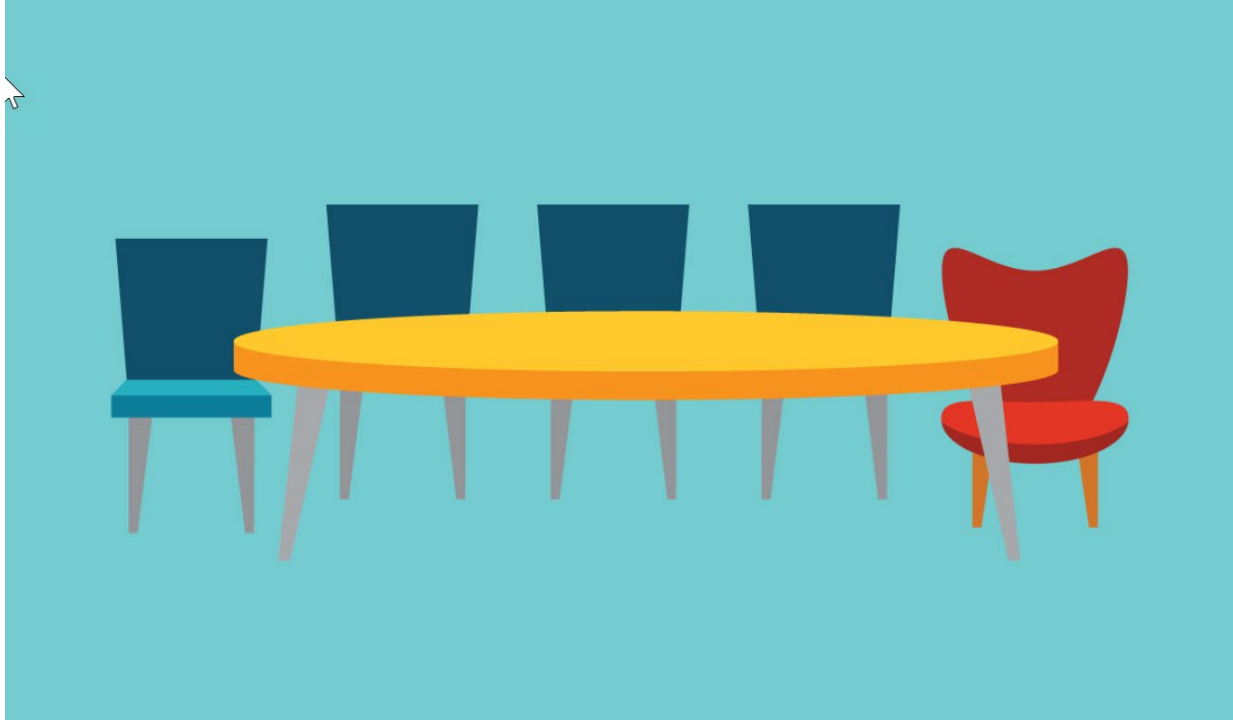
"Lateral thinking is a term coined in the 1960's for the re-imagining of information in new contexts, including the drawing together of seemingly disparate concepts or domains that can give old ideas new uses."

What separates the best data scientists from ordinary data scientists?



The best data scientists have sufficient **mastery of the tools** (e.g. R, Python, SQL) and depth of **knowledge about the data** that they can **focus on the business questions/problems** at hand.

These data scientists earn a seat at the table alongside the business stakeholders, which makes them much more effective at using data to help organizations make better decisions and improve their products, processes, and services.





Another “superpower” for data scientists is to have strong **data engineering skills**, so that you are able to help operationalize data science models.

“If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes.”


– attributed to Albert Einstein

“Experts in their fields tend to be motivated by criticism, and to see it as a sign of how well they’re progressing toward their goals, according to a 2011 study co-wrote by Dr. Finkelstein. Novices are more likely to seek praise, and to interpret it as a sign of whether to remain committed to the goals they’ve set, the study shows..”

– Sue Shellenberg, Wall Street Journal

R and SQL

- Structured (“rectangular”) data operations
- Where should the work be done?
- How should I connect my (R) client to the database (native drivers or ODBC or ORM)

	Python pandas	PySpark RDD	PySpark DF	R dplyr 	Revo R dplyrXdf
subset columns	<code>df.colname , df['colname']</code>	<code>rdd.map()</code>	<code>df.select('col1', 'col2', ...)</code>	<code>select(df, col1, col2, ...)</code>	
new columns	<code>df['newcolumn']=...</code>	<code>rdd.map(function)</code>	<code>df.withColumn("newcol", content)</code>	<code>mutate(df, col1=col2+col3, col4=col5^2,...)</code>	
subset rows	<code>df[1:10] , df.loc['rowname':]</code>	<code>rdd.filter(function or boolean vector) , rdd.subtract()</code>		<code>filter</code>	
sample rows		<code>rdd.sample()</code>			
order rows			<code>df.sort('col1')</code>	<code>arrange</code>	
group & aggregate	<code>df.sum(axis=0) , df.groupby(['A', 'B']).agg([np.mean, np.std])</code>	<code>rdd.count() , rdd.countByValue() , rdd.reduce() , rdd.reduceByKey() , rdd.aggregate()</code>	<code>df.groupBy('col1', 'col2').count().show()</code>	<code>group_by(df, var1, var2,...) %>% summarise(col=func(var3), col2=func(var4), ...)</code>	<code>rxSummary(formula, df) or group_by() %>% summarise()</code>
peek at data	<code>df.head()</code>	<code>rdd.take(5)</code>	<code>df.show(5)</code>	<code>first() , last()</code>	
quick statistics	<code>df.describe()</code>		<code>df.describe()</code>	<code>summary()</code>	<code>rxGetVarInfo()</code>
schema or structure			<code>df.printSchema()</code>		

Where should the data processing work be done?

Can it / should it be done all in SQL?

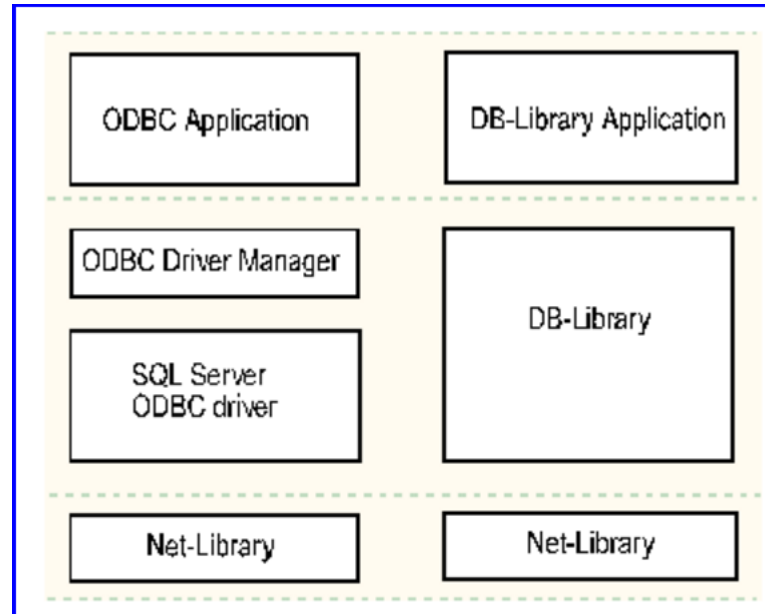
Can it / should it be done all in R?

If you are going to use both SQL and R...
how should you connect your SQL to your R?

- File handoff
- Programmatic connection

Often, the data scientist will not have direct access to the SQL data for security reasons.

Native Drivers vs. ODBC



The DBI package defines a common interface between the R and database management systems (DBMS). The interface defines a small set of classes and methods similar in spirit to Perl's [DBI](#), Java's [JDBC](#), Python's [DB-API](#), and Microsoft's [ODBC](#). It defines a set of classes and methods defines what operations are possible and how they are performed:

- connect/disconnect to the DBMS
- create and execute statements in the DBMS
- extract results/output from statements
- error/exception handling
- information (meta-data) from database objects
- transaction management (optional)

DBI separates the connectivity to the DBMS into a "front-end" and a "back-end". Applications use only the exposed "front-end" API. The facilities that communicate with specific DBMSs (SQLite, MySQL, PostgreSQL, MonetDB, etc.) are provided by "drivers" (other packages) that get invoked automatically through S4 methods.

```
# Connect to a database driver
library("RSQLite")
con = dbConnect(SQLite(), dbname = ghg_db) # Also username & password arguments
dbListTables(con)
rs = dbSendQuery(con, "SELECT * FROM `ghg_ems` WHERE (`Country` != 'World')")
df_head = dbFetch(rs, n = 6) # extract first 6 row
```

sources: top: <https://github.com/rstats-db/DBI>; bottom: Colin Gillespie and Robin Lovelace, *Efficient R Programming*, O'Reilly, Dec 8, 2016.

RMySQL Beachhead

```
library(RMySQL)
drv <- dbDriver("MySQL")
con <- dbConnect(drv, user="root", password =
  "Kasparov", dbname = "flights",
                  host = "localhost")
dbListTables(con)
dbListFields(con, "weather")
query <- "SELECT year, month, MAX(wind_speed)
  AS maxwind FROM weather GROUP BY year, month;"
max_wind <- dbGetQuery(con, query)
```

dplyr 0.7 The newest version of dplyr has “improved tools for connecting to databases.” We cover dplyr in weeks 6 and 7.



See: <https://www.rstudio.com/resources/webinars/whats-new-in-dplyr-0-7-0/>

dplyr can generate SQL against remote database

source: <https://github.com/rstudio/webinars/blob/master/39-dplyr-0.7.0/dbplyr.R>

```
library(dplyr)
```

```
library(RSQLite)
```

```
con <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")
```

```
DBI::dbWriteTable(con, "mtcars", mtcars)
```

```
mtcars2 <- tbl(con, "mtcars")
```

```
mtcars2
```

```
mtcars2 %>%
```

```
  filter(cyl > 2) %>%
```

```
  select(mpg:hp) %>%
```

```
  head(10) %>%
```

```
  show_query()
```

Resources for learning more about SQL and Regular Expressions

SQL Resources – Foundational

- For foundational SQL knowledge, the four courses in DataCamp's PostgreSQL-based SQL Fundamentals skill track:

–<https://www.datacamp.com/tracks/sql-fundamentals>

- Larry Rockoff's book



Regex resources

In addition to the content in R for Data Science and the optional content in Automated Data Collection in R...

- DataCamp courses

- String Manipulation in R with stringr. The back half of the course covers regex
- Regular Expressions in Python. Python-based, but excellent coverage of concepts

- <https://regexr.com/>

- Thanks to Salma Elshahawy for her recommendation on our Slack channel!

Data Science in Context Presentations

Your Assignment 2 solutions

Data Interrogation

- In many of the best assignment 2 solutions...
- Students asking questions of their data (which sometimes led to collecting additional information), such as
 - "Which movie had the highest average rating?"
 - "Which movie genre was rated the highest?" or
 - "How did my friend's ratings compare with those on IMBD?"

They then analyzed the data and wrote up their conclusions in the form of findings and possible recommendations on how to extend their work going forward.