

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com

# the art and science of teaching data science mine çetinkaya-rundel

# bit.ly/introds-ecots2020

Image credit: Thomas Pedersen, data-imaginist.com/art



## **1. Teach statistical thinking.**

- Teach statistics as an investigative process of problem-solving and decision making. Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision making process that is fundamental to scientific inquiry and essential for making sound decisions.
- Give students experience with multivariable thinking. We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

## 2. Focus on conceptual understanding.

- 3. Integrate real data with a context and purpose.
- 4. Foster active learning.

# 5. Use technology to explore concepts and analyse data. 6. Use assessments to improve and evaluate student learning.

amstat.org/asa/files/pdfs/GAISE/GaiseCollege\_Full.pdf

## **1. Teach statistical thinking.**

- Teach statistics as an investigative process of problem-solving and decision making. Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision making process that is fundamental to scientific inquiry and essential for making sound decisions.
- Give students experience with multivariable thinking. We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

## 2. Focus on conceptual understanding.

- 3. Integrate real data with a context and purpose.
- 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data. 6. Use assessments to improve and evaluate student learning.

amstat.org/asa/files/pdfs/GAISE/GaiseCollege\_Full.pdf

(1)NOTacommonly used subset of tests and intervals and produce them with hand calculations





## **1. Teach statistical thinking.**

- Teach statistics as an investigative process of problem-solving and decision making. Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision making process that is fundamental to scientific inquiry and essential for making sound decisions.
- Give students experience with multivariable thinking. We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

## 2. Focus on conceptual understanding.

- 3. Integrate real data with a context and purpose.
- 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data. 6. Use assessments to improve and evaluate student learning.

amstat.org/asa/files/pdfs/GAISE/GaiseCollege\_Full.pdf

(2) Multívariate analysis requires the use of computing





## **1. Teach statistical thinking.**

- Teach statistics as an investigative process of problem-solving and decision making. Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision making process that is fundamental to scientific inquiry and essential for making sound decisions.
- Give students experience with multivariable thinking. We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

## 2. Focus on conceptual understanding.

- 3. Integrate real data with a context and purpose.
- 4. Foster active learning.
- 5. Use technology to explore concepts and analyse data.

## 6. Use assessments to improve and evaluate student learning.

amstat.org/asa/files/pdfs/GAISE/GaiseCollege\_Full.pdf

(3) NOT use technology that is only applicable in the intro course or that doesn't follow good science principles

## **1. Teach statistical thinking.**

- Teach statistics as an investigative process of problem-solving and decision making. Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision making process that is fundamental to scientific inquiry and essential for making sound decisions.
- Give students experience with multivariable thinking. We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

## 2. Focus on conceptual understanding.

- 3. Integrate real data with a context and purpose.
- 4. Foster active learning.

## 5. Use technology to explore concepts and analyse data. 6. Use assessments to improve and evaluate student learning.

amstat.org/asa/files/pdfs/GAISE/GaiseCollege\_Full.pdf

(4) Data analysis ísn't just inference and modelling, it's also data importing, cleaning, preparation, exploration, and visualisation







a course that satisfies these four points is looking more like today's intro data science courses than (most) intro stats courses

but this is not because intro stats is inherently "bad for you"

instead it is because it's time to visit intro stats in light of emergence of data science









tidy data, data frames vs. summary tables, recoding & transforming, web scraping & iteration + collaboration on GitHub





tidy data, data frames vs. summary tables, recoding & transforming, web scraping & iteration + collaboration on GitHub

building & selecting models, visualising interactions, prediction & validation, inference via simulation





tidy data, data frames vs. summary tables, recoding & transforming, web scraping & iteration + collaboration on GitHub building & selecting models, visualising interactions, prediction & validation, inference via simulation

data science ethics, text analysis, Bayesian inference + communication & dissemination





tidy data, data frames vs. summary tables, recoding & transforming, web scraping & iteratic + collaboration on GitHu building & selecting models, visualising interactions, prediction & validation, inference via simulation

data science ethics, text analysis, Bayesian inference + communication & dissemination

# Go to bit.ly/rscloud-ecots2020 Start the project titled UN Votes



# Go to bit.ly/rscloud-ecots2020

- Start the project titled UN Votes
- Open the R Markdown document called unvotes.Rmd



		0 \$				
	<b>\$</b> 😔	) 🎁 Mine's (				
Help						
		R				
6	Environment History Connections					
8	😅 📊 📑 Import Dataset 👻 🍯					
	💼 Global Environment 🖌	С,				
	Environment is empty					
	Files Plots Packages Help Viewer					
	Vew Folder V Opload V Delete • Ken	ame i 👷 More 🗸				
	▲ Name Size	Modified				
	<b>1</b>					
	C .Rhistory 0 B	May 20,				
	01-unvotes.html 1.6	MB May 20,				
	01-unvotes.Rmd 2.9	KB May 20,				
	project.Rproj     205	B May 20,				
	Help	Help     Environment History Connections     Import Dataset -     Global Environment -     Environment Is em     Files Plots Packages Help Viewer     New Folder     Upload        Files Plots Packages Help Viewer     New Folder   Upload   Delete   Ren     Cloud   project   Name   Size   Size   Size   Import Cloud   Delete   Ren   Import Cloud   Delete   Ren   Import Cloud   Import Cloud <tr< td=""></tr<>				



- Go to bit.ly/rscloud-ecots2020
- Start the project titled UN Votes
- Open the R Markdown document called unvotes.Rmd



# Knit the document and review the data visualisation you just produced

- Go to bit.ly/rscloud-ecots2020
- Start the project titled UN Votes
- > Open the R Markdown document called unvotes.Rmd
- another country of your choice

Knit the document and review the data visualisation you just produced Then, look for the character string "Turkey" in the code and replace it with

## Knit again, and review how the voting patterns of the country you picked compares to the United States and United Kingdom & Northern Ireland

Percentage of 'Yes' votes in the UN General Assembly 1946 to 2015





# three questions that keep me up at night...



what should students learn? how will students learn best? what tools will enhance student learning?

# three questions that keep me up at night...

# content pedagogy infrastructure

what should students learn? how will students learn best? what tools will enhance student learning?

## content

# ex. 1 money in politics





)	Country of Origin/Parent Company	0	Total	¢	Dems	0	Repubs	0
up)	Switzerland/Asea Brown Boveri		\$1,000		\$1,000		\$0	
)	Ireland/Accenture plc		\$73,500		\$45,000		\$28,500	
h	Mexico/Grupo Salinas		\$3,000		\$1,000		\$2,000	
	125		72		- 25		22	

## SELECT A CYCLE

## 2020

PAC Name (Affiliate)	Country of Origin/Parent Company	0 Total	0 Dems	<b>ତ Repubs</b> ି
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	\$1,000	\$1,000	\$0
Accenture (Accenture)	Ireland/Accenture plc	\$73,500	\$45,000	\$28,500
Advance America Cash				
Advance Centers (Grupo Salinas)	Mexico/Grupo Salinas	\$3,000	\$1,000	\$2,000
Air Liquide America	France/L'Air Liquide SA	\$11,500	\$5,000	\$6,500
Airbus Group	Netherlands/Airbus Group	\$81,500	\$26,500	\$55,000
Alkermes Inc	Ireland/Alkermes Plc	\$40,250	\$11,250	\$29,000
Allergan PLC (Allergan PLC)	Ireland/Allergan PLC	\$111,000	\$6,000	\$105,000
Allianz of America (Allianz)	Germany/Allianz AG Holding	\$35,500	\$17,100	\$18,400

	•	

# \* web scraping \* text parsing \* data types \* regular expressions

SELECT A CYCLE					
<ul> <li>✓ 2020</li> <li>2018</li> <li>2016</li> <li>2014</li> <li>2012</li> <li>2010</li> <li>2008</li> <li>2006</li> <li>2004</li> <li>2002</li> <li>2000</li> <li>1998</li> </ul>				Repubs 0 \$0 \$28,500	<ul> <li>* web scraping</li> <li>* text parsing</li> <li>* data types</li> <li>* regular expression</li> </ul>
Advance Centers (Grupo Salinas)	Mexico/Grupo Salinas	\$3,000	\$1,000	\$2,000	<pre>* iteration</pre>
Air Liquide America	France/L'Air Liquide SA	\$11,500	\$5,000	\$6,500	
Airbus Group	Netherlands/Airbus Group	\$81,500	\$26,500	\$55,000	
Alkermes Inc	Ireland/Alkermes Plc	\$40,250	\$11,250	\$29,000	
Allergan PLC (Allergan PLC)	Ireland/Allergan PLC	\$111,000	\$6,000	\$105,000	
Allianz of America (Allianz)	Germany/Allianz AG Holding	\$35,500	\$17,100	\$18,400	

ns



\* web scraping \* text parsing \* data types \* regular expressions \* iteration \* data visualisation \* interpretation

- Party
- Democrat
- Republican

paths\_allowed("https://www.opensecrets.org")

## [1] TRUE

\* web scraping \* text parsing \* data types \* regular expressions  $\star$  iteration \* data visualisation \* interpretation \* data science ethics



## **Project:** The North South Divide: University Edition **Question:** Does the geographical location of a UK university affect its university score? Association between overall score and the latitude of a university 1000

## Team: Fried Egg Jelly Fish

## University League Tables 2020

How to use | Methodology | Full Table | Print

Order by

Overall Score

Our League Tables rank UK universities both nationally and in 70 subjects. You can sort each table on the measure important to you and compare universities. Read university profiles for more information about the institutions or search for courses. We also rank 12 specialist colleges and conservatoires separately in the Arts, Drama & Music League Table.

Filter by	Subjects 😔	Year	~	Region	~	Group	~
				C Hel		-	

	Rank		University Name	Entry Standards	Student Satisfaction	Research Quality	Graduate Prospects	Overall Score	Next Steps	
*	1st	• 0	Cambridge	274	4.09	5.65	86.7	1000	PROFILE	COURSES
*	2nd	• 0	Oxford	215	4.10	3.34	83.4	989	PROFILE	
*	3rd	• 2	St Andrews	207	4.26	3.13	79.7	944	PROFILE	COURSES
Y	4th	•1	London School of Economics	189	3.67	3.35	86.1	934	PROFILE	COURSES
~	Sth	•1	Imperial College London	205	4.02	3.36	90.4	933	PROFILE	
*	5th	• 0	Durham	194	4.01	3.14	84.8	915	PROFILE	COURSES





# Resources

- Sample assignment: introds.org/hw/hw-06/hw-06-money-inpolitics.html
- Code: Go to bit.ly/rscloud-ecots2020, start the project titled 02 -Money in politics
- Paper: Web Scraping in the Statistics and Data Science Curriculum:

Challenges and Opportunities (Dogucu & Cetinkaya-Rundel, 2020) github.com/mdogucu/web-scrape (conditionally accepted to JSE)

# **ex. 2** fisheries of the world

100

15

Charles States Mark

-

1000

10





fi	she	eries %>% select(country)	CC	onti	inents
##	# /	A tibble: 82 x 1	##	# /	A tibble
##		country	##		country
##		<chr></chr>	##		<chr></chr>
##	1	Angola	##	1	Afghani
##	2	Argentina	##	2	Åland 1
##	3	Australia	##	3	Albania
##	4	Bangladesh	##	4	Algeria
##	5	Brazil	##	5	America
##	6	Cambodia	##	6	Andorra
##	7	Cameroon	##	7	Angola
##	8	Canada	##	8	Anguill
##	9	Chad	##	9	Antigua
##	10	Chile	##	10	Argenti

fisheries <- left\_join(fisheries, continents)</pre>

## Joining, by = "country"

## \* data joins

e: 245 x 2 continent <chr> Asia istan Islands Europe Europe Africa an Samoa Oceania Europe 1 Africa Americas la a & Barbuda Americas Americas ina

```
fisheries %>%
  filter(is.na(continent))
```

```
## # A tibble: 3 x 5
##
    country
    <chr>
                                     <dbl>
##
## 1 Democratic Republic of the Congo
                                    237372
## 2 Hong Kong
## 3 Myanmar
```

```
fisheries <- fisheries %>%
 mutate(continent = case_when(
    country == "Democratic Republic of the Congo" ~ "Africa",
    country == "Hong Kong"
                                                   ~ "Asia",
    country == "Myanmar"
                                                   ~ "Asia",
    TRUE
                                                   ~ continent
```

# \* data joins \* data science ethics

capture aquaculture total continent <dbl> <dbl> <chr> 3161 240533 <NA> 142775 4258 147033 <NA> 2072390 1017644 3090034 <NA>





Average share of aquaculture by continent out of total fisheries harvest, 2016



Source: bit.ly/2VrawTt

\* data joins \* data science ethics \* critique \* improving data visualisations



## Average share of aquaculture by country



20%

0%

40%

60%

80%

Aquaculture %

\* data joins \* data science ethics critique \* improving data visualisations \* mapping

Source: bit.ly/2VrawTt



## **Project:** 2016 US Election Redux **Question:** Would the outcome of the 2016 US Presidential Elections been different had Bernie Sanders been the Democrat candidate? **Team:** 4 Squared







# Resources

- Sample lab: introds.org/labs/lab-04/lab-04-ugly-charts.html
- Code: Go to bit.ly/rscloud-ecots2020, start the project titled 03 -**Fisheries of the world**
- effective-dataviz.html

Sample lecture: introds.org/slides/w4\_d1-effective-dataviz/w4\_d1-

# ex. 3 spam filters





# \* logistic regression \* prediction



	Email is spam	Email i
Email labelled spam	True positive	False positi error)
Email labelled not spam	False negative (Type 2 error)	True negati

## s not spam

ive (Type 1

ive

\* logistic regression \* prediction \* decision errors \* sensitivity / specificity \* intuition around loss functions



**Project:** Spotify Top 100 Tracks of 2017/18 **Question:** Is it possible to predict the year a song made the Top Tracks playlist based on its metadata? Team: weR20

year ~ danceability + energy + key + loudness + mode + speechiness + acousticness + instrumentalness + liveness + valence + tempo + duration\_s

## 2017

artists name I'm the One DJ Khaled Childish Gambino Redbone Sign of the Times Harry Styles

## 2018

artists name **Everybody Dies In Their Nightmares XXXTENTACION** Jocelyn Flores XXXTENTACION Rich The Kid Plug Walk Moonlight **XXXTENTACION** Nevermind **Dennis Lloyd** In My Mind Dynoro changes XXXTENTACION



# Resources

- w10\_d1-logistic-regression.html
- filter

# Sample lecture: introds.org/slides/w10\_d1-logistic-regression/

Code: Go to bit.ly/rscloud-ecots2020, start the project titled 04 - Spam

Book chapter: OpenIntro Statistics, 4th Edition (Diez, Cetinkaya-Rundel, and Barr, 2019), Chapter 9.5 with randomised controlled trial data on discrimination on job application evaluation openintro.org/book/os

# pedagogy

# **teams:** weekly labs in teams + periodic team evaluations + term project in teams

"minute paper": weekly online quizzes ending with a brief reflection of the week's material **peer feedback:** used minimally so far, but positive experience

• • •	S Week 07 - Simple linear regres × +	
$\leftarrow \   \rightarrow $	C minecr.shinyapps.io/07-modeling-s	simple/#section-
١	Nook 07 - Simpla	Fini
2 1	veer or - oimple	1 11 11
ļ	inear regression	To finis
Ĩ		1
	Teacher salaries	
	Single numerical predictor	
	Prediction	
	A second set 40	
2	Assessing model fit	
	Working backwards	
	Single categorical predictor	
	Finish up	
	Start Over	

-up

## ९ ☆ 🕐 🌄 🐁 😁 🔉 🕈 🛆 🚺 🌍 :

## up

the quiz go to the form linked below and answer a few simple questions.

Hi Mine, when you submit this form, the owner will be able to see your name and email address.	
* Required	
1. What is your name? *	
Enter your answer	
2. What is your student ID? *	
This is the number that starts with s.	
Enter your answer	
3. Write about one or two questions you didn't get right initially but were able to solve after a few	

tries. What was difficult about them? What did you ultimately learn?

OR

If you got every single question correct on the first try, write one question you would still like clarified on the topics covered in this quiz. \*

Your answers can be brief / in bullet point form. The goal isn't to make you write too much, but instead to make you quickly reflect on your learning.

Enter your answer



## teams: weekly labs in teams + periodic team evaluations + term project in teams

"minute paper": weekly online quizzes ending with a brief reflection of the week's material

peer feedback: used minimally so far, but positive experience

creativity: assignments that make room for creativity



• • •			×	+		
$\leftarrow \   \rightarrow $	C 🔒 i	introds.org				
	IDS	Timetable	Schedule	Syllabus	Project	Help
	Pro	oject	t			TL;C
	Show	case your	inner da	ta scien	tist	Pick a
						and o below.
						May
						The fin choosi survey interes project this cla meanin
						The go
						statisti
						that yo
						results

Q (

People Resources 🕥 😨 P 🔘

## ;DR

a dataset, any dataset...

d do something with it. That is your final project in a nutshell. More details w.

## ay be too long, but please do read

final project for this class will consist of analysis on a dataset of your own osing. The dataset may already exist, or you may collect your own data using a ey or by conducting an experiment. You can choose the data based on your rests or based on work in other courses or research projects. The goal of this ect is for you to demonstrate proficiency in the techniques we have covered in class (and beyond, if you like) and apply them to a novel dataset in a ningful way.

goal is not to do an exhaustive data analysis i.e., do not calculate every stic and procedure you have learned for every variable, but rather let me know you are proficient at asking meaningful questions and answering them with Its of data analysis, that you are proficient in using R, and that you are

at internation and muse outing the near star Peace on methods that halp





e

 $\leftarrow$ 

 $\rightarrow$ 



## infrastructure







) + pghclass +		
Code Dissues 0 Pull requests 1 Projects 0 EWiki Decurity	L⊥ Insights © Settings	
Draft mine-cetinkaya wants to merge 1 commit into master from styler 🕃		+16 -1
Results of running styler: Styling 1 files: hw-03-ncbikecrash.Rmd i		
StatusCountLegend✓0File unchanged.i1File changed.×0Styling threw an error.		
Please review the changes carefully!		
mine-cetinkaya-rundel committed 16 minutes ago	commit 721c39cf8b	6c0e6b3bc50f29187474159
<ul> <li>✓ 34 ■ ■ ■ hw-03-ncbikecrash.Rmd </li> <li>28 @@ -7,7 +7,7 @@ output: github_document</li> </ul>		
7 8 ```{r load-packages, message=FALSE} 9 # load packages	<pre>7 8 ```{r load-packages, message=FALSE} 9 # load packages 18 tlib ramu(tiduus coo)</pre>	











## openness



0 ρ R

## Introduction to Data Science

Learn to explore, visualize, and analyze data to understand natural phenomena, investigate patterns, model outcomes, and make predictions, and do so in a reproducible and shareable manner. Gain experience in data collection, wrangling, and visualization, exploratory data analysis, predictive modeling, and effective communication of results while working on problems and case studies inspired by and based on real-world questions. The course will focus on the R statistical computing language. No statistical or computing background is necessary.

🚓 Incognito

☆

Q (

## 🔮 Data Science in a Box :: Data S 🗙 🛛 🕂

datasciencebox.org



Q Search...

Hello #dsbox

Course content

Infrastructure

Pedagogy

# Data Science in a Box

How can we effectively and efficiently teach data science to students with little to no background in computing and statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and leave them wanting to learn more? This introductory data science course is our (working) answer to this question.

The core content of the course focuses on data acquisition and wrangling, exploratory data analysis, data visualization, inference, modeling, and effective communication of results. Time permitting, the course also introduces additional concepts and tools like interactive visualization and reporting, text analysis, and Bayesian inference. A heavy emphasis is placed on a consitent syntax (with tools from the tidyverse), reproducibility (with R Markdown), and version control and collaboration (with Git and GitHub). In addition, out-of-class learning is supplemented with interactive tutorials. The goal of the course is to bring students from zero to being able to work in a team on a fully reproducible data science project analyzing a dataset of their choice and answering questions they care about.

Data Science in a Box contains the materials required to teach (or learn from) the course described above, all of which are freely-available and open-source. They include course materials such as slide decks, homework assignments, guided labs, sample exams, a final project assignment, as well as materials for instructors such as pedagogical tips, information on computing infrastructure, technology stack, and course logistics.

Built with 💙 and blogdown, logo by muuuuge.

Q 🕁

🚓 Incognito



## Day 1 - Monday, Jan 27

ne	Activity
00 - 10:30	Welcome + Curriculum design
30 - 11:00	Coffee break
00 - 12:30	Teaching the tidyverse
30 - 13:30	Lunch break
30 - 15:00	Computing infrastructure with RStudio Cloud
00 - 15:30	a Coffee break
30 - 17:00	Case study: Design your assignment on RStudio Cloud

0 🔤 🐁 🞯 🔬 🔘 🥩 🗄

## Day 2 - Tuesday, Jan 28

## Materials for Day 2 will be posted in the morning of Day 2

ne	Activity
:00 - 10:30	Reproducible workflows: R Markdown, Git, GitHub
:30 - 11:00	Coffee break
:00 - 12:30	Getting more out of GitHub
:30 - 13:30	Lunch break
:30 - 15:00	Interactivity and immediate feedback
:00 - 15:30	🥌 Coffee break
:30 - 17:00	#rstats lifehacks for instructors + Wrap up

# the art and science of teaching data science

mine çetinkaya-rundel

## bit.ly/introds-ecots2020 Code for all case studies: bit.ly/introds-ecots2020-cases or bit.ly/rscloud-ecots2020

aminebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com



Image credit: Thomas Pedersen, <u>data-imaginist.com/art</u>

