出國報告（出國類別：研究）

參加 R 語言暨開發平台技術研討會及工作坊

服務機關：疾病管制署
姓名職稱：陳必芳 技正
派赴國家/地區：美國/舊金山
出國期間：109 年 1 月 26 日至 109 年 2 月 6 日
報告日期：109 年 3 月 31 日

摘　　要

R 語言為當今統計分析、資料探勘、資料視覺化領域中，不論使用者人數、易用性、社群支援性均極高的開源語言之一。本研討會為 RStudio 公司為推廣 R 語言及軟體，每年固定舉辦之開發者交流研討會，研討會內容包括技術工作坊、新知及應用成果分享、社群交流等主題，內容亦涵蓋人工智慧等新興熱門領域。藉由參與本次研討會及工作坊，實作學習地理資訊處理及視覺化技巧，吸收最新 R 語言技術資訊，觀摩及交流不同領域之應用經驗，除精進個人資料加值分析能力，相關經驗並可供本署後續持續開發 R 語言資料加值分析平臺之參考。

目　　次

壹、本文

一、目的

疾病管制署自 107 年起，逐步建構以 R 語言為核心之資料加值分析平臺，除透過此平台建立多項視覺化面板、自動化產出業務所需之資料外，並提供同仁於資料伺服器執行資料分析工作，執行業務所需之資料加值應用。

為充分運用本署 R 伺服器資源，使用者需具備一定程度之 R 語言程式碼識讀及撰寫能力，藉由參與本次研討會及工作坊，實作學習地理資訊處理及視覺化技巧，吸收最新 R 語言技術資訊，觀摩及交流不同領域之應用經驗，除精進個人資料加值分析能力，相關經驗並可供本署後續持續開發 R 語言資料加值分析平臺之參考。

二、過程

R 語言為當今統計分析、資料探勘、資料視覺化領域中，不論使用者人數、易用性、社群支援性均極高的開源語言之一，其原因之一即為其主流之程式開發環境(Integrated Development Environment，IDE)－RStudio 之操作直觀便捷，R 語言並擁有大量由軟體公司或使用者自行開發之擴充套件，套件功能廣泛且涵蓋各學術分野，大幅降低程式開發門檻，學習週期也較傳統 C、Java 等程式語言短上許多。

本研討會由 RStudio 公司為推廣 R 語言、相關軟體及服務，105 年起每年例行性舉辦之開發者研討會，本年已邁入第 5 屆，共 2,242 人參與，參與人數逐年成長。本年度會議共 4 天，包括 2 日技術工作坊及 2 日演講(議程如附錄)，技術工作坊共有 19 種，從最基礎的 R 語言入門及基本套件練習，到各種專門領域套件教學、應用開發等主題課程，學員可擇一參加；後 2 日演講除主要 keynote speech 外，同時有 2-4 個主題演講平行進行，分享議題廣泛，包括程式開發、作業流程規劃、套件新知分享、程式教育等，並闢財經、醫藥等領域之應用經驗分享，亦涵蓋人工智慧等新興熱門領域。由於資料科學為一門應用科學，參與者大多非資訊工

程、電腦工程背景，而單純為 R 語言使用者或小型套件開發者，多數演講著重分享個人應用經驗，尤其如何學習、熟悉 R 語言，最終如何將 R 語言導入工作流程中或甚至藉以維生；會議期間亦安排多場 Birds of a feather 交流活動，促成同領域參加者互動。

就本次研討會參加之工作坊、演講節錄重點內容如下：

(一) 地理資訊分析及視覺化工作坊

（Modern Geospatial Data Analysis with R Workshop）

本工作坊講師 Zev Ross 為 ZevRoss Spatial Analysis 公司執行長，具 15 年以上資料分析及視覺化經驗，擅長使用 R 語言開發互動式地圖介面及相關應用軟體，曾執行多項地理資訊視覺化商業專案，亦於 DataCamp 等線上教學網站開授地理資訊視覺化相關課程。

由於空間資訊的處理資料經常涉及座標系統選擇、附加資訊欄位處理，另有多種資料型態及檔案格式，操作上較一般資料更為複雜。此工作坊選擇 sf、raster、tmap、mapview 四個套件建立資料處理 pipeline，完成整個空間資料匯入、處置及視覺化工作，為期 2 天的課程採穿插講課、程式碼示範、學員實機操作三部分，期望工作坊結束後能讓學員熟悉空間資料處理及視覺化步驟、套件選用及常見問題的排解方式，除了每段課程中的練習實作外，最後半天每位學員均使用紐約市地圖圖資、空汙資料、路網資料、普查資料，依步驟實際完成一項迷你分析專案，過程中除講師外亦有 4 位助教從旁協助，可隨時解決學員從環境設定到程式排錯各種疑難雜症。重要 take home message 摘錄如下：

1. 空間資料型態：空間資料分為向量資料(vector)、網格資料(raster)兩大類，向量資料常用檔案格式包含 shapefile、geojson、topojson 等，可使用 sf 套件讀取及處理；網格資料常用檔案格式包含 TIFF/geoTIFF、img、HDF4/5 等，分為單層網格及多層網格資料，可使用 raster 套件讀取及處理；另可使用 rasterize()、rasterToPoints()/rasterToPolygons()函數轉換量種類型的檔案。

2. 座標系統(Coordinate Reference System, CRS)：由於地球為一球體，經緯度等地理座標需透過特定投影方法轉換為二維空間的投影座標，可使用 st_crs()、crs()函式分別檢視或指派向量及網格資料的座標系統，並可使用 st_transform()、projectRaster()函式將地理座標系統轉換為指定的投影座標系統，講師特別強調雖然 tmap 套件對座標系統的容錯度較大，資料就算座標系統不同也可能成功疊加呈現，但使用者仍應保持維護圖資座標系統的好習慣。

3. 圖層繪製：plot()為 R 原生的繪圖函數，設定參數 add = TRUE 即可疊加不同圖層物件，多用來快速檢視匯入資料；tmap 則為一套強大的圖層繪製套件，提供大量圖層設定參數，其撰寫語法及邏輯類似 ggplots2 套件，可繪製靜態及互動式地圖，其附屬的 tmaptools 套件亦提供底圖載入、色彩配置等設定；另 mapview 套件亦為一快速產出互動式地圖之選擇。前述產出之地圖物件均可嵌入 R markdown 文件或 shiny 面板呈現。

4. 其他空間資訊視覺化套件： ggspatial、leaflet、concaveman、cartography、ggmap、tidycensus、rayshader、rgrass7、stars、geogrid、arcgisbridge 等。

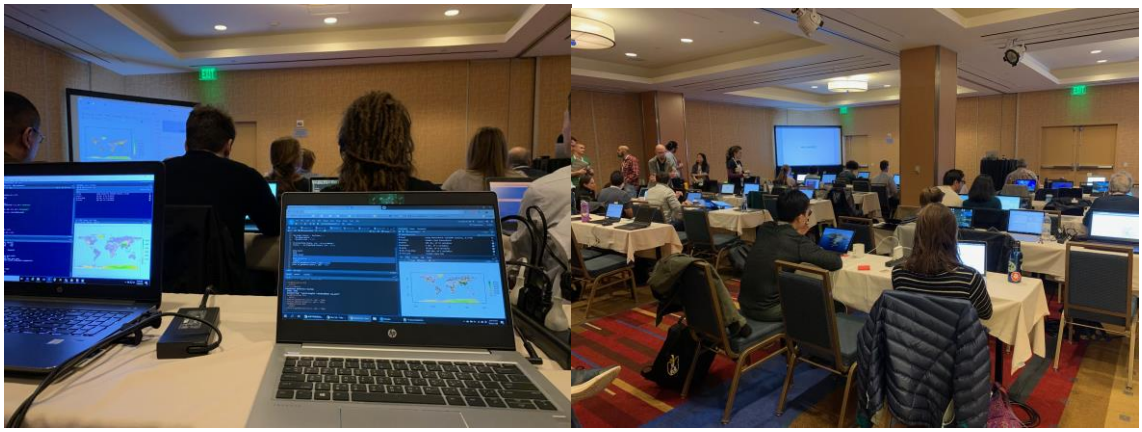

圖 1、工作坊上課情形。學員登入講師提供之伺服器後，使用講師自行開發實機操作套件，練習課程講授之資料處理及圖層產出步驟。

(二) 資料處理

資料處理套件－Tidyverse 新知

Tidyverse 套件為 Rstudio 公司開發的大型套件組，集合眾多 R 語言資料處理熱門套件，從資料載入的 readr，資料處理的 dplyr、tibble、tidyr，視覺化的 ggplot2 等，致力成為一全方位的資料科學套件。2019 年套件新增{{}}運算子(curly-curly operator)功能，取代原本使用 enquo()函數的撰寫方式，使程式碼可更加簡潔；另 vctrs、vroom、tidymodels 套件亦持續更新；2020 年將推出 dplyr 1.0.0 版本，並將發布可整合 google 表單資料的 googlesheets4 套件，可延伸整合 google 相關線上資源。

(三) 資料視覺化

1. 圖表視覺化包含顏色、字型及排版三大元素，講者針對圖表設計提出三項建議：(1)移除非必要的格線及顏色，(2)盡量避免旋轉標示文字，以免影響閱讀動線，(3)妥善設計圖例，圖例位置應依視覺動線放置於左上角，但講者偏好直接在圖表上直接標示，或在標題文字中直接以顏色標示關聯文字(如下圖)。

圖 2、圖例優化示意。圖例應儘量安排於視覺動線前段位置，最好能和原標題或圖表說明文字結合。

2. 3D 圖表視覺化套件－rayshader 簡介

3D 視覺化雖然在目前資訊溝通中較少使用，但隨著圖表媒介逐漸轉變為互動式電子平臺，另 2D 視覺化不足以呈現部分高維度資料，因此 3D 圖表繪製功能仍有其需求性。rayshader 套件除了擅長繪製 2D、3D 地圖外，亦可非常容易的將 ggplot 圖表物件 3D 化，轉換為立體圖表，並可輸出為

gif 檔進行動態展示。

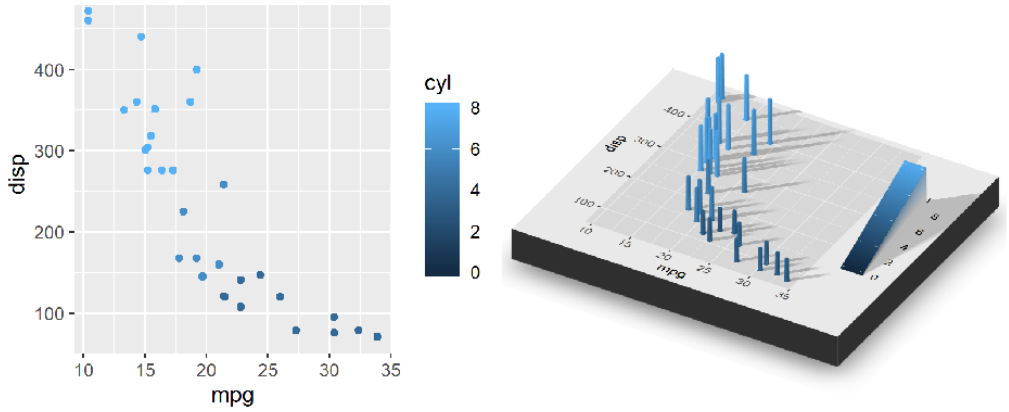

圖 3、使用 rayshader 套件中的 plot_gg()函數，對 ggplot 物件進行 3D 化。

3. 產出動態物件套件－plotly 簡介

plotly 套件可在 R 語言中整合 D3.js 繪圖功能，使用 ggplotly()函數可將靜態的 ggplot 物件轉換為可互動的 web-based 物件，完成後可嵌入 RMarkdown 文件或 Shiny 儀表板發佈，雖然 plotly 是一套眾所皆知的套件，但講者使用該套件開發的制式報表完整性之高，視覺化設計嚴謹，令人十分驚艷。

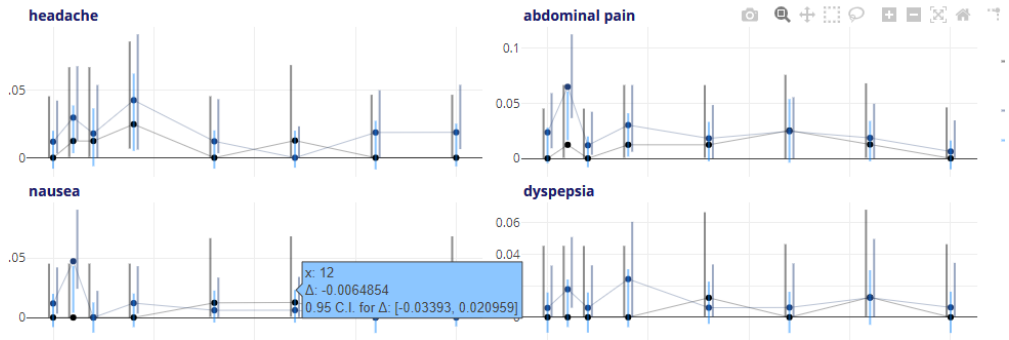

圖 4、講者使用 plotly 繪製圖表物件後嵌入 html 文件，圖表可具備拉近、顯示懸浮視窗等簡易互動功能。

4. ggplot 說明文字彈性調整

ggplot2 套件中對標題、座標軸、圖例等文字設定參數不多，可調整幅度有限，講者建議可於 theme()函數中使用 element_markdonw()函數，讀取預先寫好帶有 html tag 的文字，即可自由調整部分文字顏色、粗斜體、加入

圖片等功能，大幅度擴充圖表客製化彈性。

(四) 人工智慧

Data, visualization, and designing with AI 演講中，Google People + AI Research (PAIR)團隊以「Debug your data first, not your program」為開場，強調資料品質調校及驗證工作在人工智慧研究中的重要性，並分享團隊開發的資料探索工具：What-If、Facets、Embedding Projector，三者均運用非常精巧的視覺化方式，使用者不需撰寫任何程式碼，即可透過瀏覽器進行資料探索及分析，以利切入後續模型開發工作。

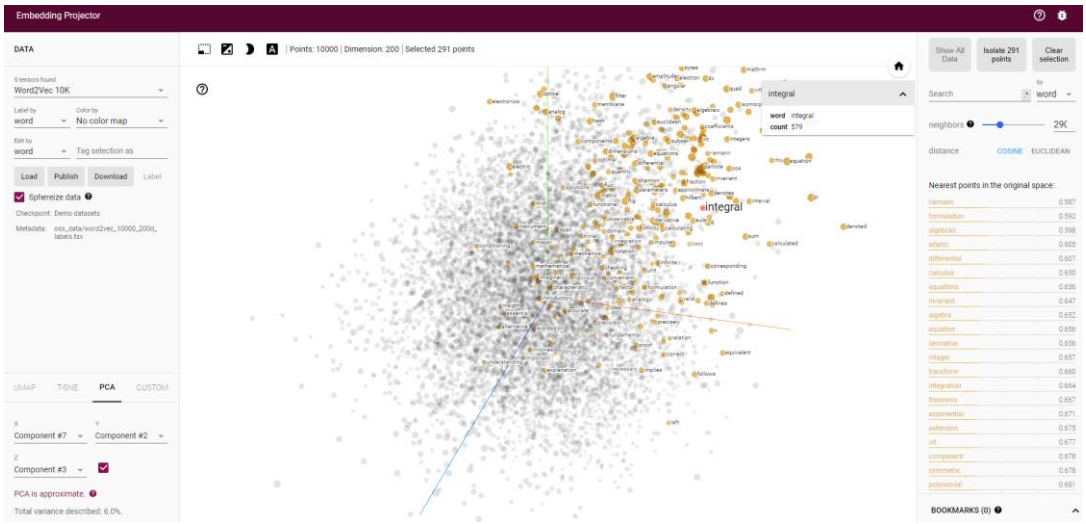

圖 5、Embedding Projector 操作示範，該工具可呈現高維度資料，適合用於影像、文字、音源等分析使用之資料。

(五) 其他應用分享

1. 應用 xaringan 套件製作 ggplot2 課程投影片教材

   程式碼教學投影片通常為了呈現每行指令的效果，需要製作大量投影片。xaringan 套件可使用 R markdown 生成 markdown 檔案，產出 HTML 格式的投影片，以便於投影片中嵌入 R 的程式碼及執行結果，講者分享自製的 ggplot2 線上教材(https://evamaerey.github.io/ggplot_flipbook/ggplot_flipbook_xaringan.html)，充分善用此套工具優勢，可逐步展現 ggplot2 語法對應的視覺化效果，除教材本身值得

參考外，教材的製作方式亦可作為未來本署製作 R 語言教材參考。

圖 6、講者分享自己製作的 ggplot2 flipbook 教材

2. 趣味是學習的最佳調味料

講者建議尋找自己有興趣的主題開發 side project，藉以學習各種可能與工作無關但某天可能會派上用場的套件。以講者為例，因為熱愛電影侏儸紀公園，使用 R 語言中 gganimate 套件練習繪製了電影中主角的移動軌跡。

圖 7、講者分享自己數個有趣的 side project，鼓勵聽眾用自己喜歡的方式，廣泛摸索各種 R 語言套件

三、心得及建議

本次會議是開始學習程式語言後，首次參加的開發者研討會，由於這類應用科學性質的研討會參與者組成複雜、所學專業歧異度極大，和過往參加傳染病領域的研討會經驗不太相同，參與者之間常有頗大的溝通隔閡，但卻是最有趣的地方－不同的專業領域、不同知識背景、使用不同的程式套件，最後試圖解決的核心問題卻十分相近，如同在地理資訊視覺化工作坊結束後，學員們紛紛開始拿出自己的圖層資料，套用到課程程式碼中，試著現學現賣了一番，並互相交流著疾病防治、植物分布、交通控制、糧食缺乏等各式各樣的問題。此外，由於社群支援一直是開源語言開發者的重要資源之一，社群交流的活動安排也是本次會議令人印象深刻的一環，主辦方在會場準備了非常多種類的別針，方便參與者別在識別證上，展示自己的工作類型、專業領域、擅長的套件或技術，以便與會者開啟話題及交流，活動期間亦安排多場午間會面、休息時間會面活動，和幾位與會者閒聊工作上遇到的困難後，他們還很熱心的協助在 twitter 上求救，順便分享如何有效率的使用 hashtag，才能讓最多人看到自己的求救文。

參加本次研討會的收穫，除了認識很多新的套件、聽到了很多程式更漂亮的寫法、吸收了很多資訊科學新知、突然解決了幾個過去無法突破的問題外，更重要的是看著每位熱情洋溢的講者，在分享經驗的同時，引導著聽眾思考：「我們還能做些什麼？我們能用程式語言做什麼？我們為了達成目的還要再學會什麼？」。

隨著各種疫情資訊及相關衍伸資訊量的擴張、資料型態的複雜化，資料科學在傳染病防治工作中扮演的角色日趨關鍵，應用面向也日益寬廣，疾病管制署疫情中心近年亦陸續投注資源於培育署內相關人才、建構 R 語言/Python 語言分析平臺，但實際業務經手的資料大多比課程使用的資料更加原始及複雜，資料處理過程亦更加細緻，初出茅廬的程式新手們很難立刻將所學套用到現實世界中，以致很難在工作中達到做中學、學中做的良性循

環，建議本署未來可在程式基礎教學後，安排小型業務導向的實作工作坊，邀集有類似業務需求的同仁，嘗試撰寫程式語言逐步將工作自動化，另亦建議本署 R 伺服器資料分析平臺可開發專用資料分析套件，將呼叫 ODBC、撰寫 SQL 指令等步驟打包成較友善操作的函式，並針對該套件撰寫一般使用者接受度較高的操作文件，以降低同仁使用該平臺進行資料分析處理之門檻。

圖 8、Birds of a feather 交流活動，會場準備了非常多種類的別針，方便參與者識別其他人的工作、專業領域、擅長的套件或技術

貳、附錄

一、議程

January 28, 2020

09:00 - 17:00 Workshop

January 28, 2020

09:00 - 17:00 Workshop

January 29, 2020

09:00-09:05	Welcome to rstudio::conf 2020			
09:05-10:00	Open Source Software for Data Science			
10:00-11:00	Data, visualization, and designing with AI			
11:00-11:30	Break			
	Case Study	Education	Production	Programming
11:30-11:52	Case Studies in Customer Success	Meet You Where You R	Deploying End-To-End Data Science with Shiny, Plumber, and Pins	Simplified Data Quality Monitoring of Dynamic Longitudinal Data: A Functional Programming Approach
11:53-12:15	How Vibrant Emotional Health Connected Siloed Data Sources and Streamlined Reporting Using R	Data Science Education in 2022	We’re hitting R a million times a day so we made a talk about it	vctrs: Creating custom vector classes with the vctrs package
12:16-12:38	Building a new data science pipeline for the FT with RStudio Connect	Data science education as an economic and public health intervention in East Baltimore	Growth Hacking with R- Product Analytics at Scale using R and RStudio	Asynchronous programming in R
12:39-12:59	How to win an AI Hackathon, without using AI	Of Teacups, Giraffes, & R Markdown	Practical Plumber Patterns	Azure Pipelines and GitHub Actions
13:00-14:15	Lunch Break			

	Community	Finance	Interface	Shiny
14:15-14:37	If you build it, they will come...but then what? Facilitating communities of practice in R	15 Years of R in Quantitative Finance	Accelerating Analytics with Apache Arrow	Production-grade Shiny Apps with golem
14:38-15:00	Embracing R in the Geospatial Community	Deep Learning Extraction for Counterparty Risk Signals from a Corpus of Millions of Documents	Updates on Spark, MLflow, and the broader ML ecosystem	Making the Shiny Contest
15:01-15:23	The development of "datos" package for the R4DS Spanish translation	Rpanda trading simulation- from an idea to a multi-user shiny app	What's new in TensorFlow for R	Styling Shiny apps with Sass and Bootstrap 4
15:24-15:44	R: Then and Now	The good, the bad and the ugly: What I learned while consulting as a data scientist	Deep Learning with R	Reproducible Shiny apps with shinymeta
15:45-16:00	Break			
	Case Study	Learning and Using R	Pharma	Programming
16:00-16:22	Journalism with RStudio, R, and the tidyverse	Flipbooks	Approaches to Assay Processing Package Validation	Getting things logged
16:23-16:45	Putting the Fun in Functional Data: A tidy pipeline to identify routes in NFL tracking data	Learning R with humorous side projects	Building a native iPad dashboard using plumber and RStudio Connect in Pharma	Technical debt is a social problem
16:46-17:08	R + Tidyverse in Sports	Toward a grammar of psychological experiments	FlatironKitchen: How we overhauled a Frankensteinian SQL workflow with the tidyverse	Parallel computing with R using foreach, future, and other packages
17:09-17:29	Making better spaghetti (plots): Exploring longitudinal data with the brolgar package	R for Graphical Clinical Trial Reporting	Using R to Create Reproducible Engineering Test Reports	Future: Simple Async, Parallel & Distributed Processing in R- What's Next?

January 30, 2020

09:00-10:00	Object of type ‘closure’ is not subsettable			
10:00-10:30	Break			
	Communication	**Medicine**	**Visualization**	**Workflow**
10:30-10:52	Branding and Packaging Reports with R Markdown	Building a Medical Device with R	The Glamour of Graphics	RMarkdown Driven Development
10:53-11:15	Don’t repeat yourself, talk to yourself! Repeated reporting in the R universe.	Development of a web-based clinical decision support application for platelet transfusion management	3D ggplots with rayshader	renv: Project Environments to R
11:16-11:38	How Rmarkdown changed my life	Forecasting Platelet Blood Bag Demand to Reduce Inventory Wastage at the Stanford Blood Center	Designing Effective Visualizations	RStudio 1.3 Sneak Preview
11:39-11:59	One R Markdown Document, Fourteen Demos	Shiny New Things: Using R to Bridge the Gap in EMR Reporting	Tidyverse 2019-2020	Using Jupyter with RStudio Server Pro
12:00-13:00	Lunch Break			
	Modeling	**Organizational Thinking**	**Programming**	**ggplot2**
13:00-13:22	MLOps for R with Azure Machine Learning	Small Team, Big Value: Using R to Design Visualizations	Auto-magic package development: Building an R API for building Vega-Lite Specs	Best practices for programming with ggplot2
13:23-13:45	Totally Tidy Tuning Techniques	UnicoRns are real	Bridging the gap between SQL and R: Introducing queryparser and tidyquery	Spruce up your ggplot2 visualizations with formatted text
13:46-14:08	Neural Networks for Longitudinal Data Analysis	Data Science in Meatspace	List-columns in data.table: Reducing the cognitive & computational burden of complex data	The little package that could: taking visualizations to the next level with the scales package
14:09-14:29	Stochastic Block Models with R: Statistically rigerous clusting with rigorous code	Value in Data Science Beyond Models in Production	Advances in tidyeval	Extending your ability to extend ggplot2

14:30-14:45	Break	
	Lightning Talks 1	Lightning Talks 2
14:45-14:50	Making a tidy dress	`livecode`: broadcast your live coding sessions from and to RStudio
14:50-14:55	A high school student’ s journey to bring R into the classroom.	Datasets in Reproducible Research with 'pins'
14:55-15:00	Course Material Creation in the R Ecosystem	Becoming an R blogger
15:00-15:05	Data Science for Software Engineers: busting software myths with R	Mexican electoral quick count night with R
15:05-15:10	Learn to teach, for goodness sake.	Rproject templates to automate and standardize your workflow
15:10-15:15	Learning by Teaching: Mentoring at the R4DS Online Learning Community	Sound annotation with Shiny and wavesurfer
15:15-15:20	Every voice matters: An analysis of @WeAreRLadies	Peer review in data science courses
15:20-15:25	The Five Principles of Data Science Education	Lessons about R I learned from my cat
15:25-15:30	-	TidyBlocks: using the language of the tidyverse in a blocks-based interface
16:00-17:00	NSSD Episode 100	
17:00-17:05	Wrap up	

二、線上資源

(一) 工作坊課程教材

http://files.zevross.com/workshops/spatial/slides/html/0-deck-list.html

(二) 研討會演講直播影片

https://resources.rstudio.com/rstudio-conf-2020

(三) 研討會演講投影片(部分)

https://github.com/EmilHvitfeldt/RStudioConf2020Slides