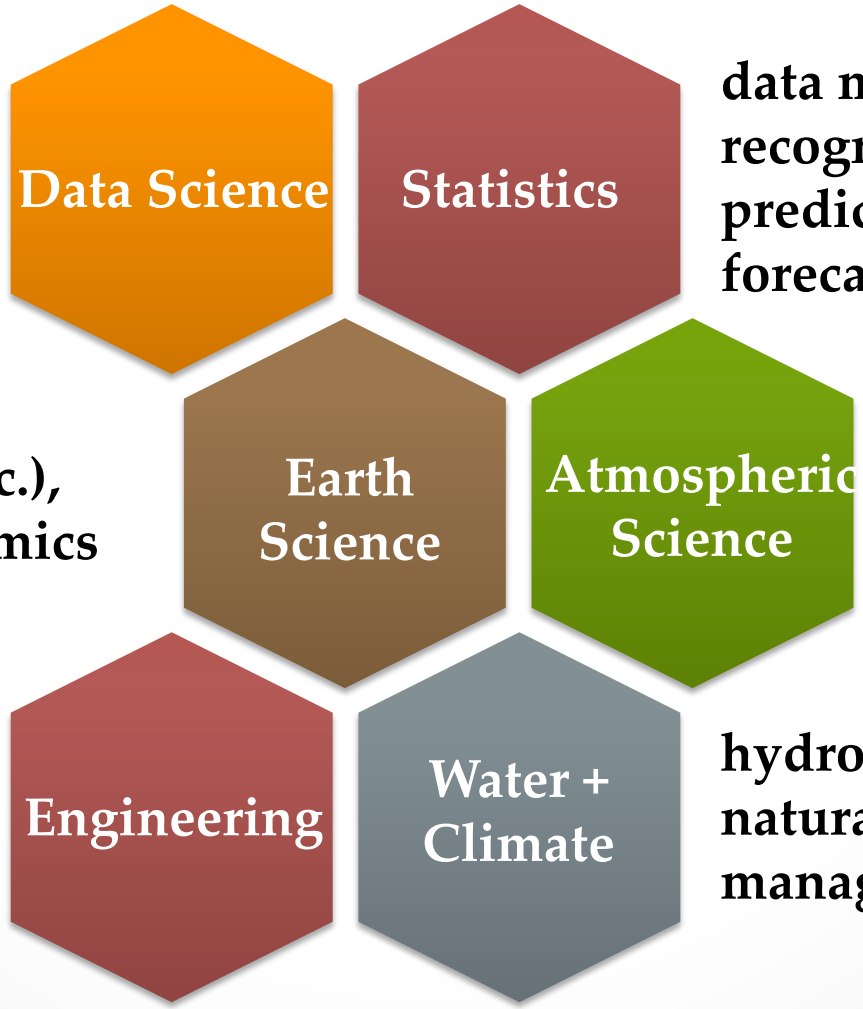
EXPLORATORY DATA ANALYSIS AND VISUALIZATION

Mengqian LU

WHO AM I?



data mining, pattern recognition, prediction and forecasts

severe weather (floods, tornado etc.), atmospheric dynamics

hydroclimatology, natural resources management, policy

WHO ARE YOU?

What do you already know?

What is your expectation?

Survey

(*http://goo.gl/forms/GpBjYuXuoy*)

COMMUNICATION

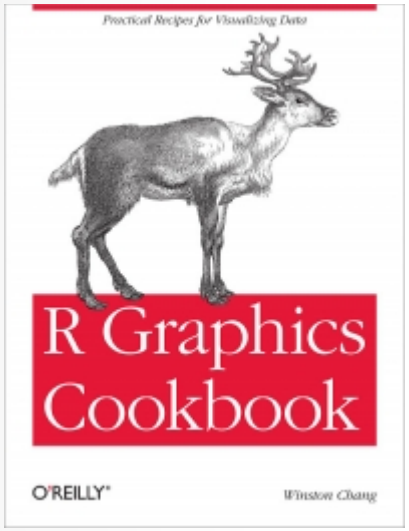
COURSEWORKS@COLUMBIA
Piazza@COURSEWORKS

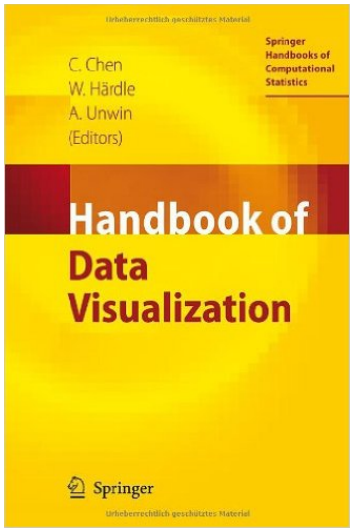
*GITHUB (https://github.com/**MRandomMax**/EDAV)*

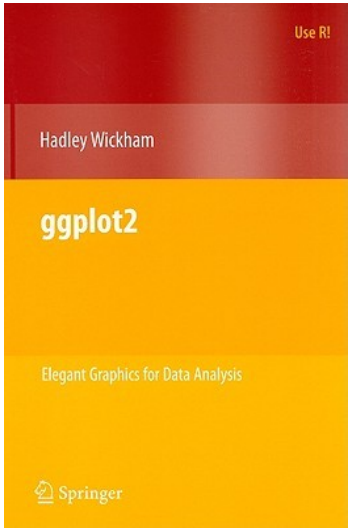
Office Hours: Friday 11AM - 12PM (noon)
EMAILS

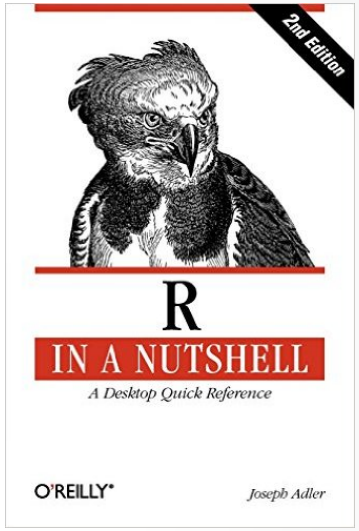
TA: Susanna Makela
Office Hours: TBA 4hours/wk, 2hrs on each day

TEXTS AND WEBSITES









Quick-R: http://www.statmethods.net/
R-bloggers: http://www.r-bloggers.com/
Github: http://github.com/

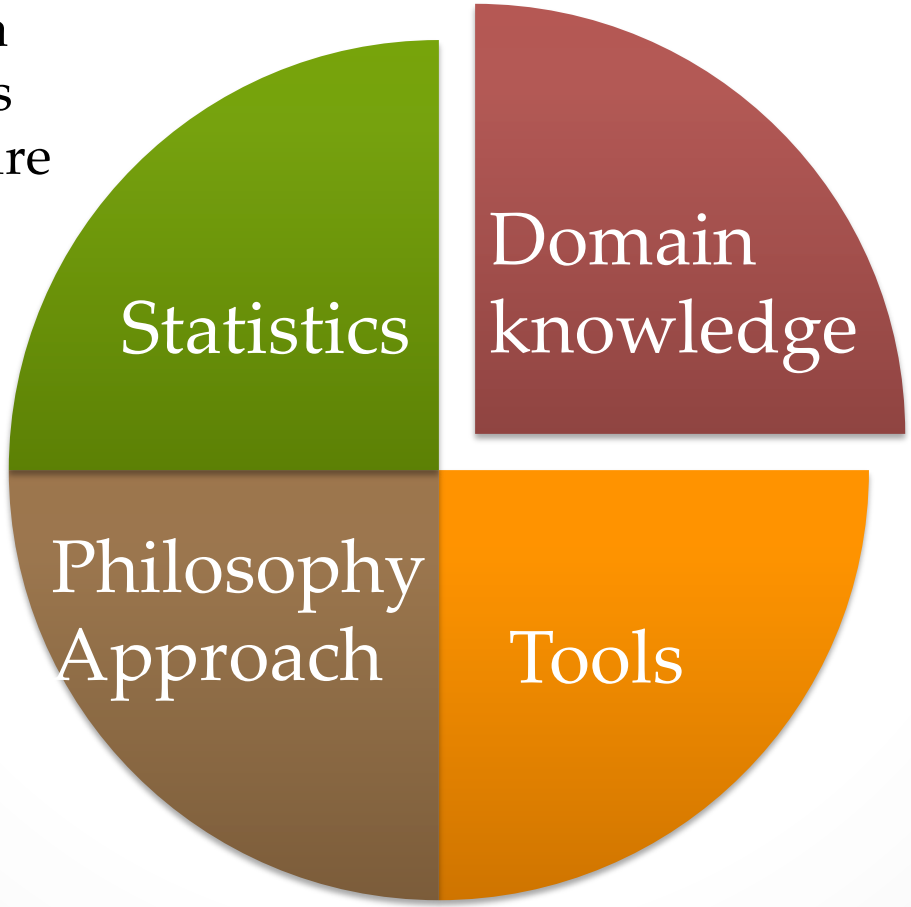
- Upshot: http://github.com/theupshot

GOALS

Work with real data
- Different sources
- Complex structure
- High dimension
- Big

Practical data visualization

Collaborating with others



THE PLAN

- Fundamentals

- Assignments
- Final project

- Themes & topics
- Theory & techniques
- Case studies

MOTIVATIONS

- Students' work in the past

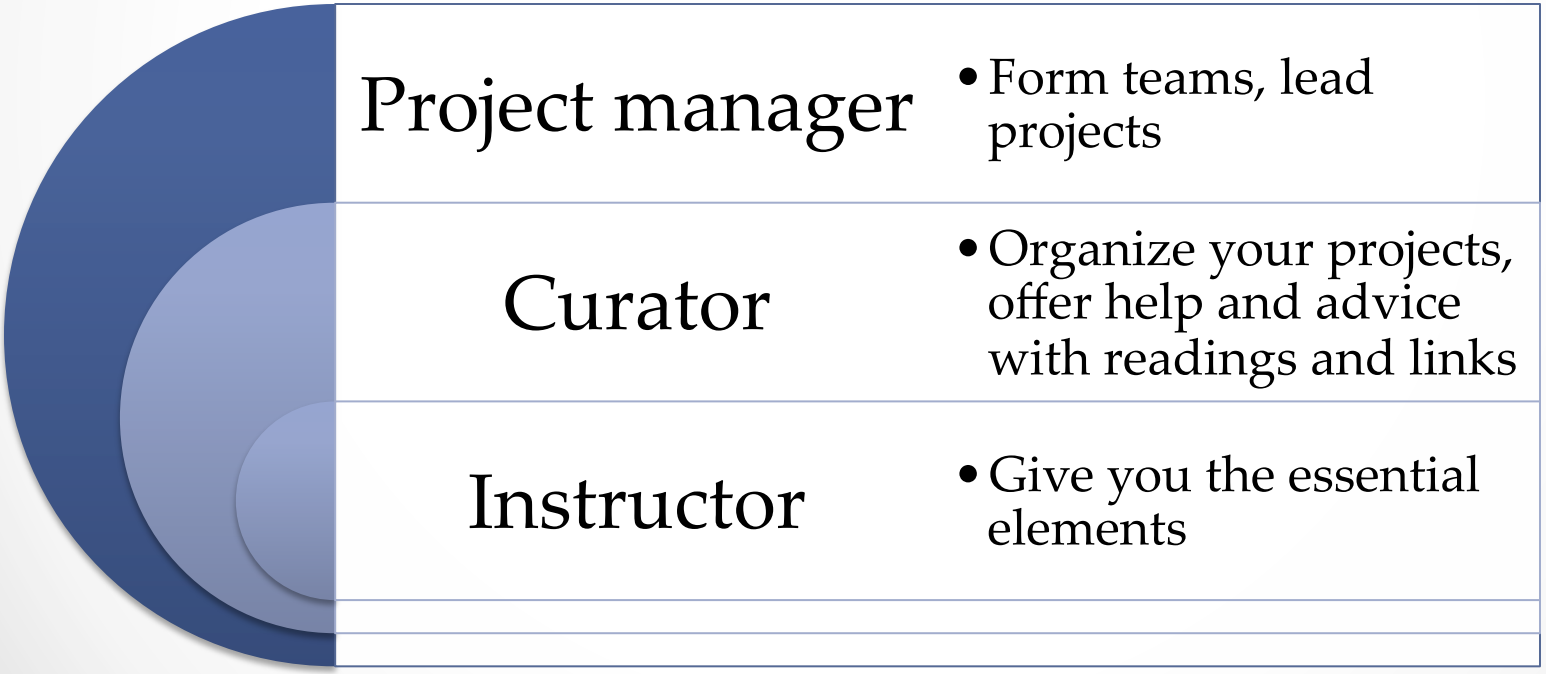
Upcoming events

- Data visualization week in April, organized by Science and Engineering Libraries, Data Science Institute, Columbia University
- Industrial data science demo day, Data Science Institute, Columbia University and companies
- Statistics Club @ Columbia – Hackathon & DataFest

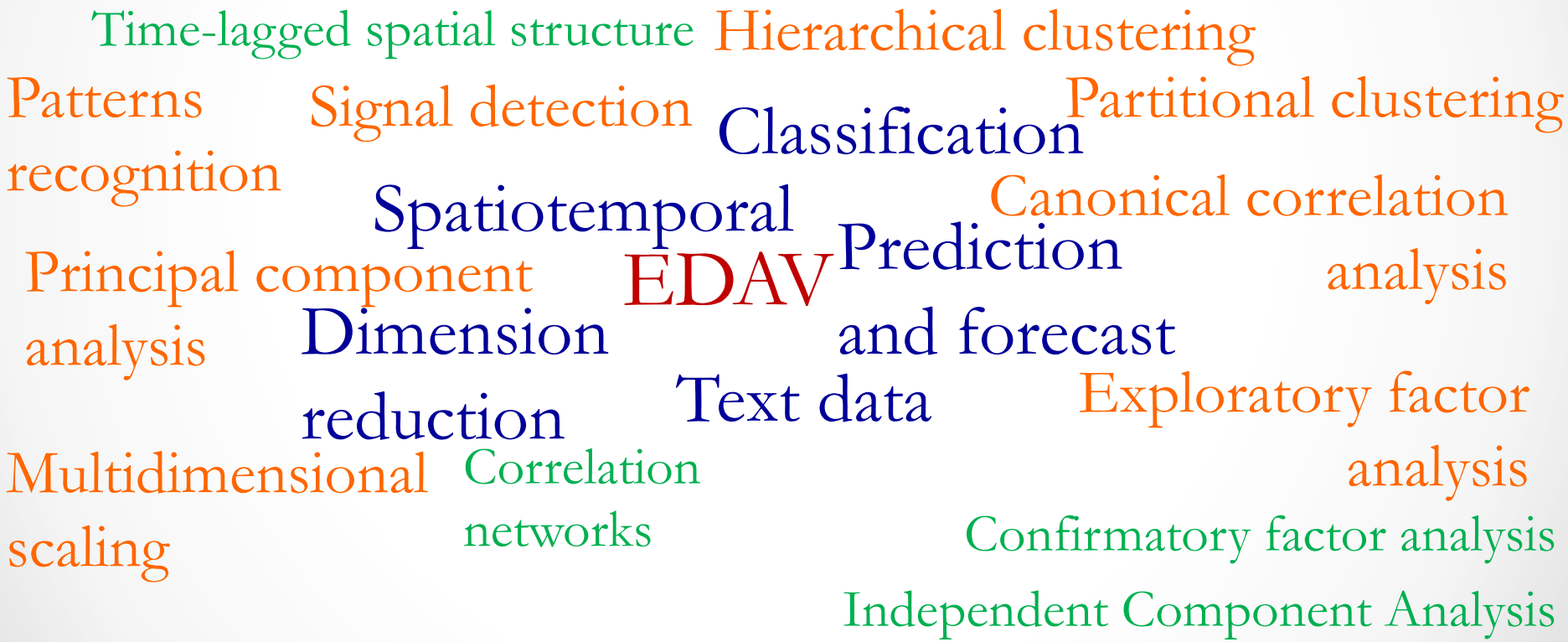
THE EVALUATION

Course Evaluation:	
Participation:	5%
Peer reviewed assignments:	45%
Peer evaluation on team presentations:	15%
Final project (presentation, progress report, final report) evaluated by Instructor and TA:	35%

MY ROLES


Project manager
• Form teams, lead projects
Curator
• Organize your projects, offer help and advice with readings and links
Instructor
• Give you the essential elements

WORD CLOUD


Time-lagged spatial structure
Hierarchical clustering
Patterns recognition
Signal detection
Classification
Partitional clustering
Spatiotemporal
Canonical correlation analysis
Principal component analysis
EDAV
Prediction and forecast
Dimension reduction
Text data
Exploratory factor analysis
Multidimensional scaling
Correlation networks
Confirmatory factor analysis
Independent Component Analysis

What is EDA?

- EDA is an approach or a philosophy for data analysis
  - Not a set of techniques, but an attitude about how you want to carry out analysis
  - Not identical to statistical graphics

- EDA techniques are graphical in nature always with some quantitative techniques
  - Graphs are the key role players but not the only ones
  - Graphs give one unparalleled power to reveal structural secrets, and find information unexpected

DATA ANALYSIS APPROACHES

1. Classical

Question(s) → Data → Model → Analysis → Conclusions

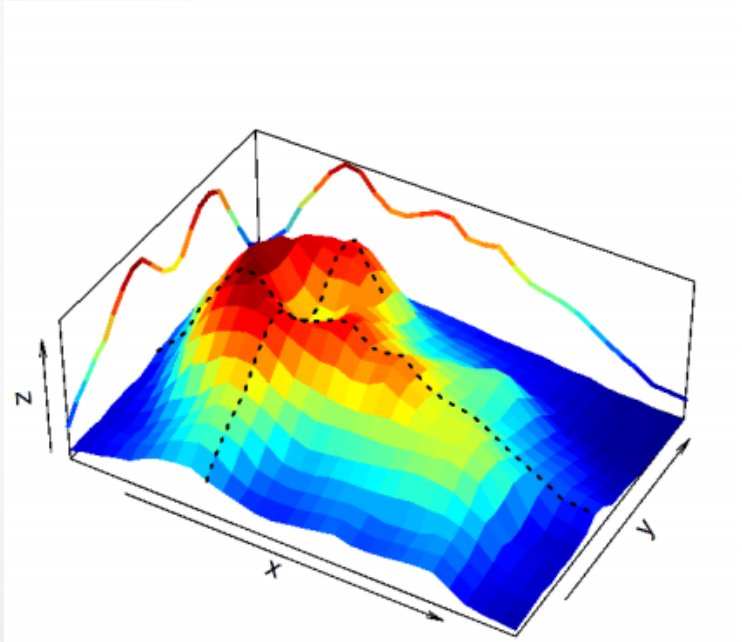
2. Exploratory (EDA)

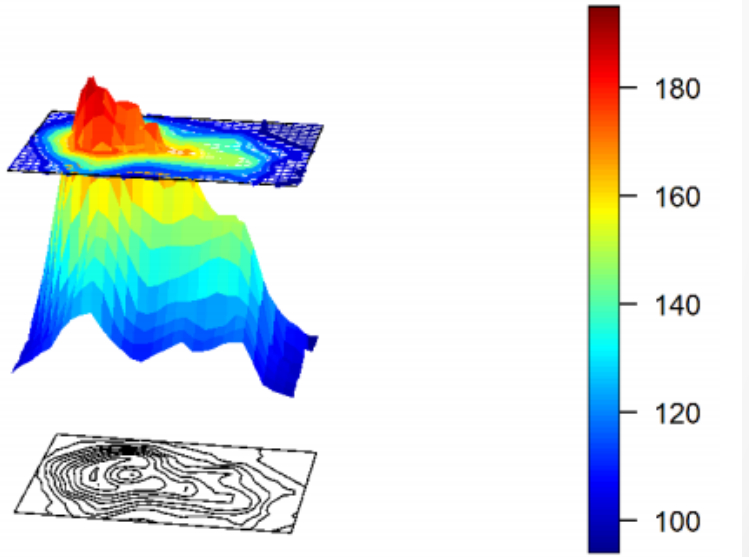
Question(s) → Data → Analysis → Model → Conclusions

3. Bayesian

Question(s) → Data → Model → Prior Distribution → Analysis → Conclusions

THE KEY TO EDA IS VISUALIZATION

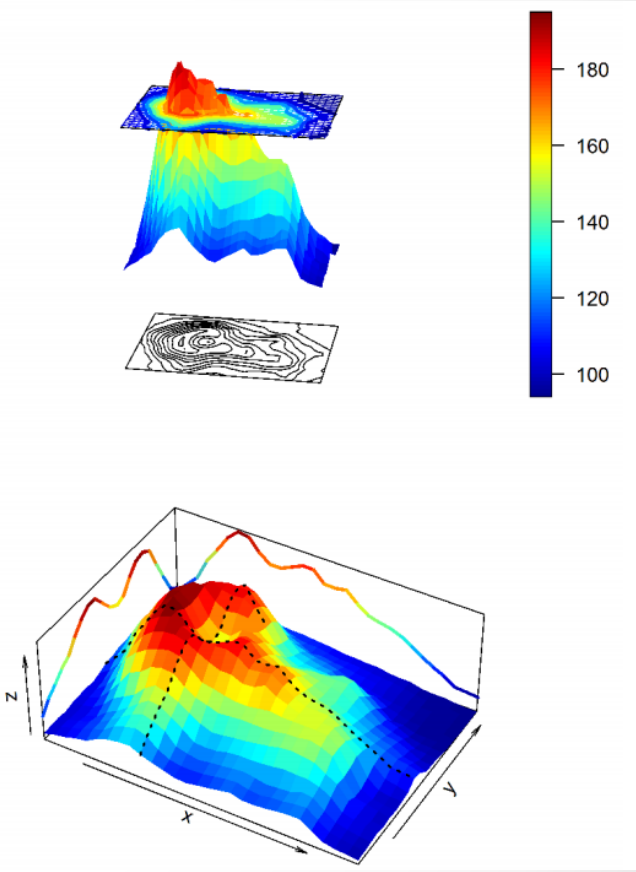




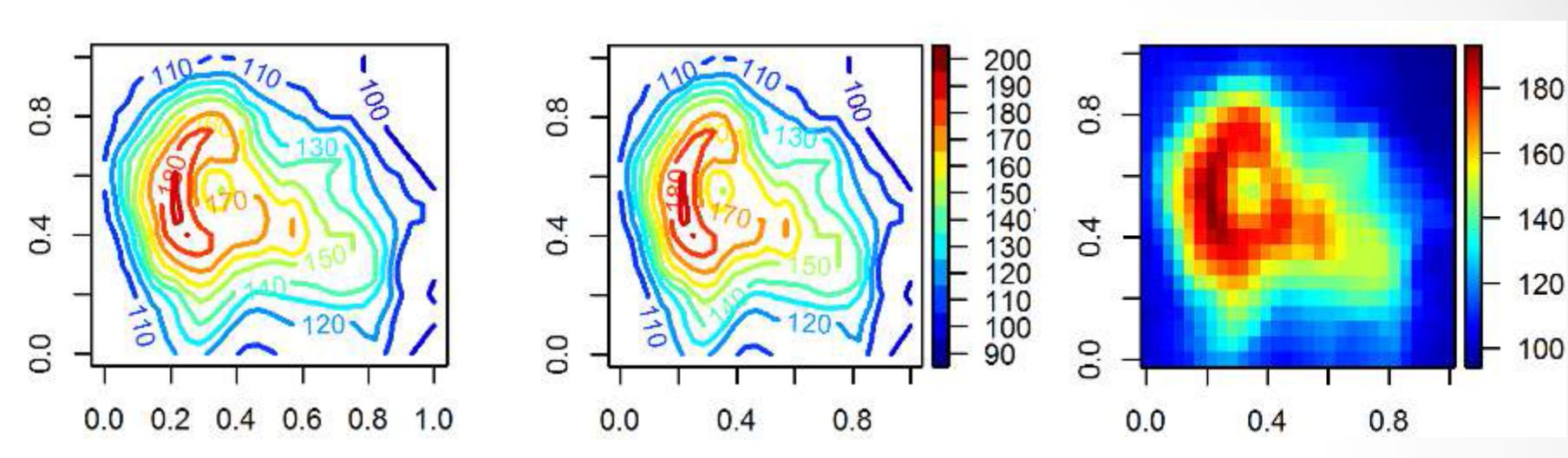
Gridded data

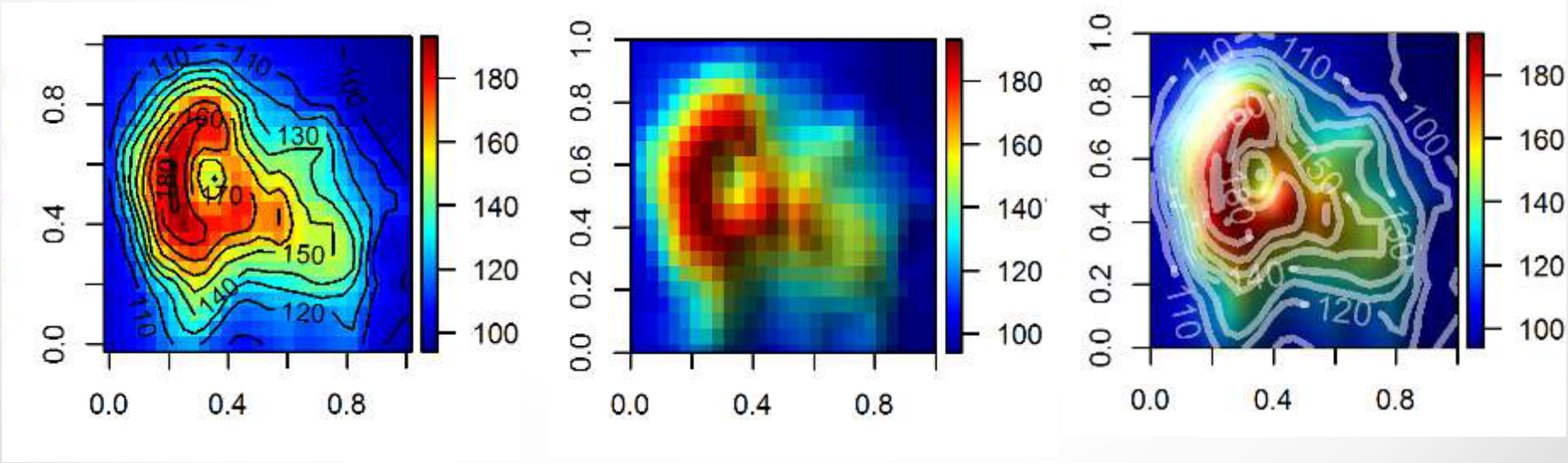
Volcano

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
1	100	101	101	100	102	102	103	101	104	107	108	110
2	103	104	104	103	104	105	106	107	111	118	120	124
3	105	107	107	105	106	109	114	120	123	129	140	142
4	108	110	110	108	113	120	127	136	141	150	158	157
5	110	115	114	117	124	133	150	155	161	165	169	174
6	116	118	121	123	130	147	160	170	179	181	183	187
7	120	126	128	130	136	152	167	178	186	191	193	191
8	122	130	135	139	147	161	172	182	190	189	184	182
9	123	133	140	146	154	164	175	183	185	177	167	164
10	118	129	137	145	151	163	173	180	180	169	158	153
11	114	120	131	138	146	154	164	174	179	169	157	149
12	111	114	120	130	139	147	155	168	177	174	166	161
13	108	112	117	121	132	144	153	164	178	179	176	170
14	107	112	115	120	128	140	150	164	174	179	176	166
15	109	113	117	121	129	141	148	159	166	168	164	159
16	111	115	118	124	131	142	148	160	168	168	160	153
17	113	117	120	125	132	142	150	166	170	170	163	155
18	115	118	121	125	134	142	152	159	162	160	157	150
19	112	115	119	126	136	143	150	155	155	152	148	145
20	112	114	117	127	139	145	150	150	150	149	142	140
21	113	116	118	129	140	146	150	150	150	147	139	136

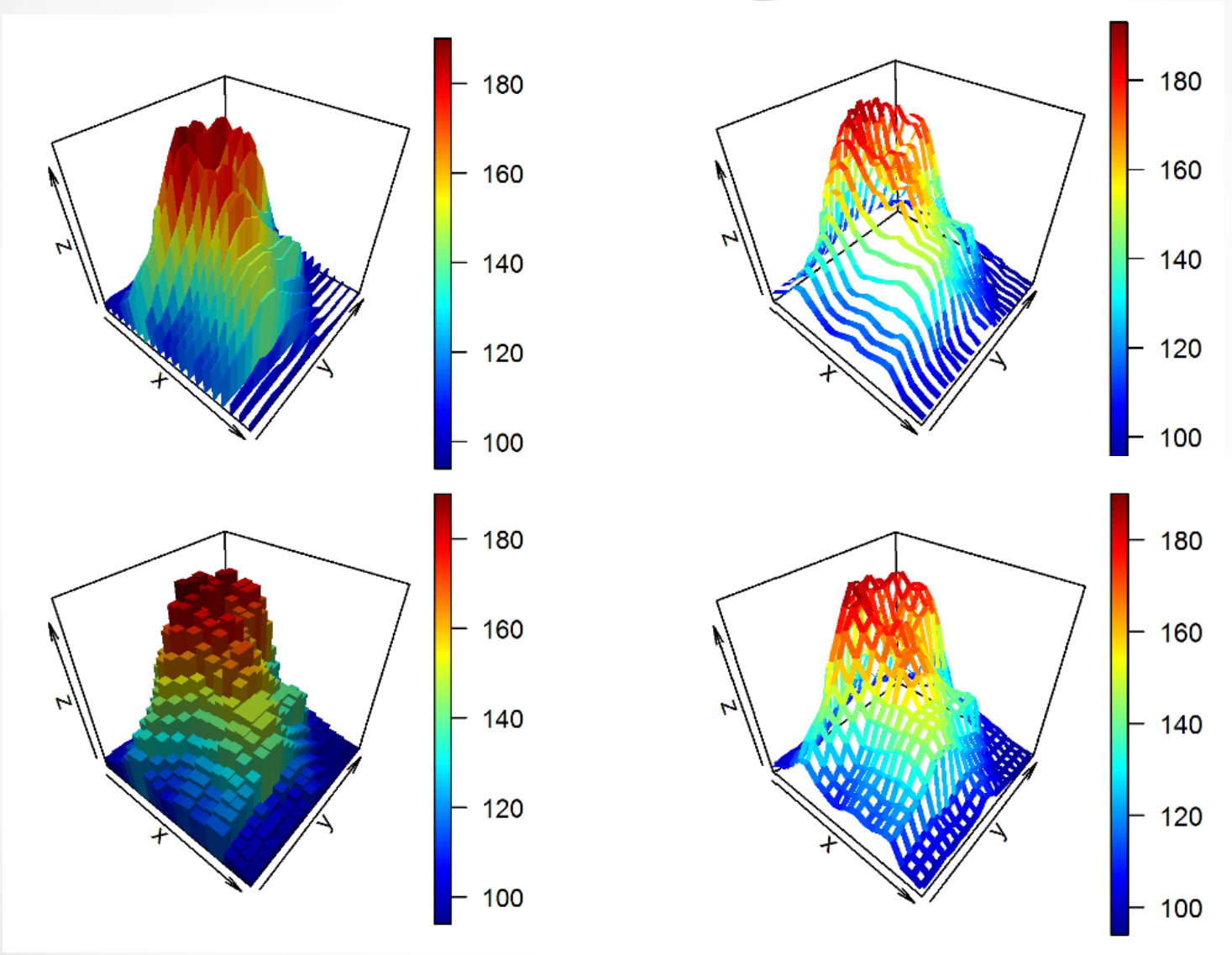


Numerous ways…





Numerous ways…

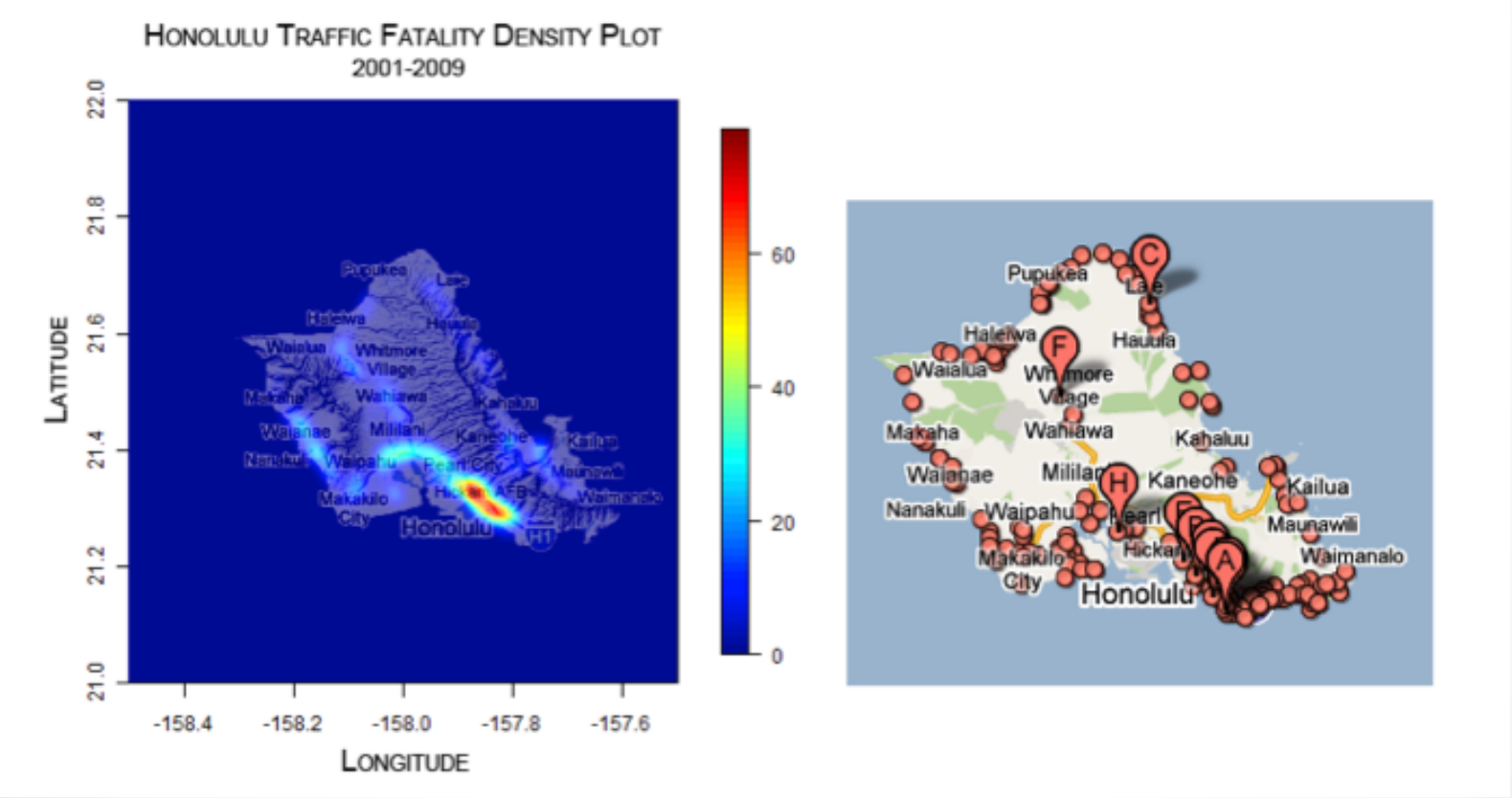


THE TAKEAWAY

- Make the plot based on data
- Design the layout that maximize the information
- Present the key information, avoid overwhelming your viewers

Location, date and time

Example: Traffic fatalities in Honolulu 2001 – 2009



Locations (start, finish), date and time

Volcano | NYCflights.R | MUCflights.R | airports | flight.info

6344 observations of 11 variables

	AirportID	Name	City	Country	IATA	ICAO	Latitude	Longitude	A
1	1	Goroka	Goroka	Papua New Guinea	GKA	AYGA	-6.081689	145.391881	5
2	2	Madang	Madang	Papua New Guinea	MAG	AYMD	-5.207083	145.788700	2
3	3	Mount Hagen	Mount Hagen	Papua New Guinea	HGU	AYMH	-5.826789	144.295861	5
4	4	Nadzab	Nadzab	Papua New Guinea	LAE	AYNZ	-6.569828	146.726242	2
5	5	Port Moresby Jacksons Intl	Port Moresby	Papua New Guinea	POM	AYPY	-9.443383	147.220050	1
6	6	Wewak Intl	Wewak	Papua New Guinea	WWK	AYWK	-3.583828	143.669186	1
7	7	Narsarsuaq	Narssarssuaq	Greenland	UAK	BGBW	61.160517	-45.425978	1
8	8	Nuuk	Godthaab	Greenland	GOH	BGGH	64.190922	-51.678064	2
9	9	Sondre Stromfjord	Sondrestrom	Greenland	SFJ	BGSF	67.016969	-50.689325	1
10	10	Thule Air Base	Thule	Greenland	THU	BGTL	76.531203	-68.703161	2
11	11	Akureyri	Akureyri	Iceland	AEY	BIAR	65.659994	-18.072703	6
12	12	Egilsstadir	Egilsstadir	Iceland	EGS	BIEG	65.283333	-14.401389	7
13	13	Hornafjordur	Hofn	Iceland	HFN	BIHN	64.295556	-15.227222	2
14	14	Husavik	Husavik	Iceland	HZK	BIHU	65.952328	-17.425978	4
15	15	Isafjordur	Isafjordur	Iceland	IFJ	BIIS	66.058056	-23.135278	8
16	16	Keflavik Nas	Keflavik	Iceland	KEF	BIKF	63.985000	-22.605556	1
17	17	Patreksfjordur	Patreksfjordur	Iceland	PFJ	BIPA	65.555833	-23.965000	1

Displayed 1000 rows of 6344 (5344 omitted)

Console ~/

```
      lsk     fnr                          lvg hal ha2 ha3            haf          hafen
52126   L SQ  328 Singapore Airlines          SIN                Singapur      Singapore
52127   L LH 2557 Lufthansa                    TBS                 Tiflis        Tbilisi
52128   L LH 1789 Lufthansa                    YEI                  Bursa         Bursa
52129   L LH  765 Lufthansa                    BOM         Mumbai (Bombay) Mumbai (Bombay)
52130   L QR  009 Qatar Airways                DOH                   Doha          Doha
52131   L LH 2541 Lufthansa                    LED           St.Petersburg   St Petersburg
                     stt                 ett                   lde                len ter ber
52126 2011-01-07 05:10:00 2011-01-07 05:10:00             Singapur           Singapur   2   H
52127 2011-01-07 05:55:00 2011-01-07 05:40:00             Georgien           Georgien   2   H
52128 2011-01-07 05:55:00 2011-01-07 05:35:00               T}rkei             T}rkei   2   H
52129 2011-01-07 06:00:00 2011-01-07 07:50:00               Indien             Indien   2   H
52130 2011-01-07 06:45:00 2011-01-07 06:35:00                Katar              Katar   2   H
52131 2011-01-07 06:50:00 2011-01-07 08:20:00 Russische F|deration Russische F|deration   2   H
       typ ver saa gat
52126 B77W      49 I32
52127 A319 100  49 I30
52128 CRJ9      49
52129 A343      49 I08
52130 A332      49 I38
52131 A320 200  49 I20
> View(airports)
>
```

Environment | History

Import Dataset | Clear | List

Global Environment

Data

airports	6344 obs. of 11 variables
flight.info	774 obs. of 18 variables
flights	0 obs. of 0 variables
myflights	9 obs. of 18 variables
myroutes	1553 obs. of 35 variables

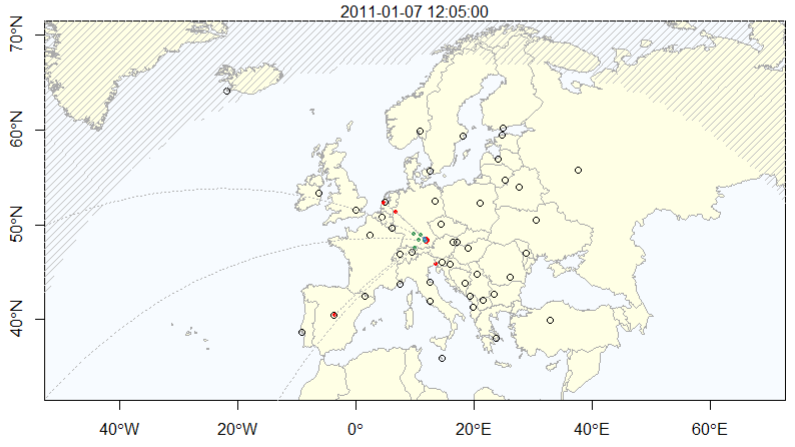
Values

all	List of 0
missing	int [1:12] 1 2 3 4 5 6 7 8 9 10 ...
months	int [1:12] 1 2 3 4 5 6 7 8 9 10 ...
needed	chr [1:12] "2013-1.csv" "2013-2.csv" "2013-3.csv" "2013-4.cs...

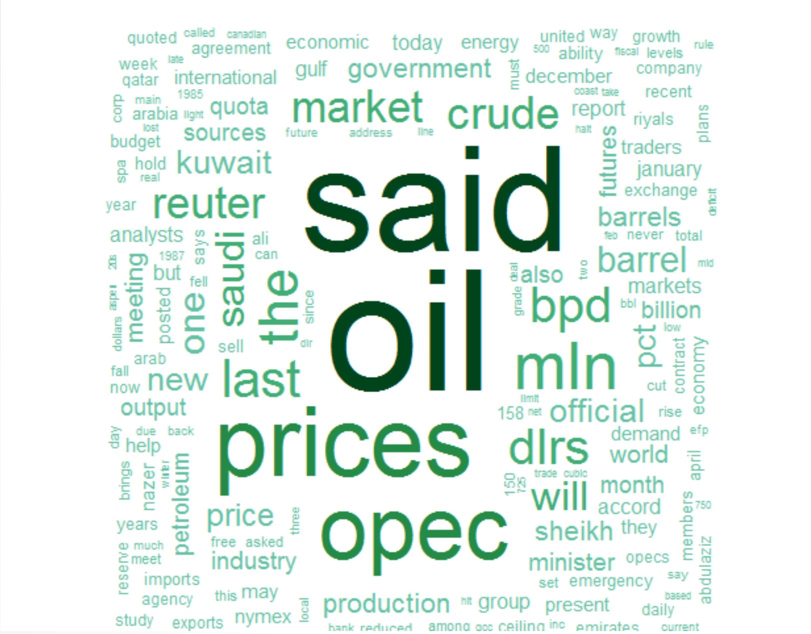
Functions

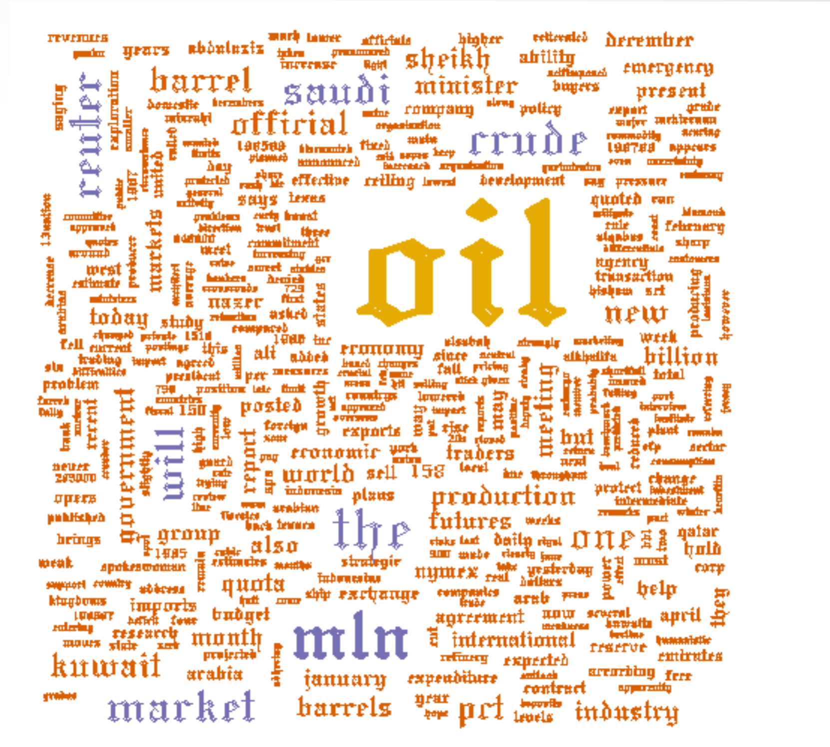
Files | Plots | Packages | Help | Viewer

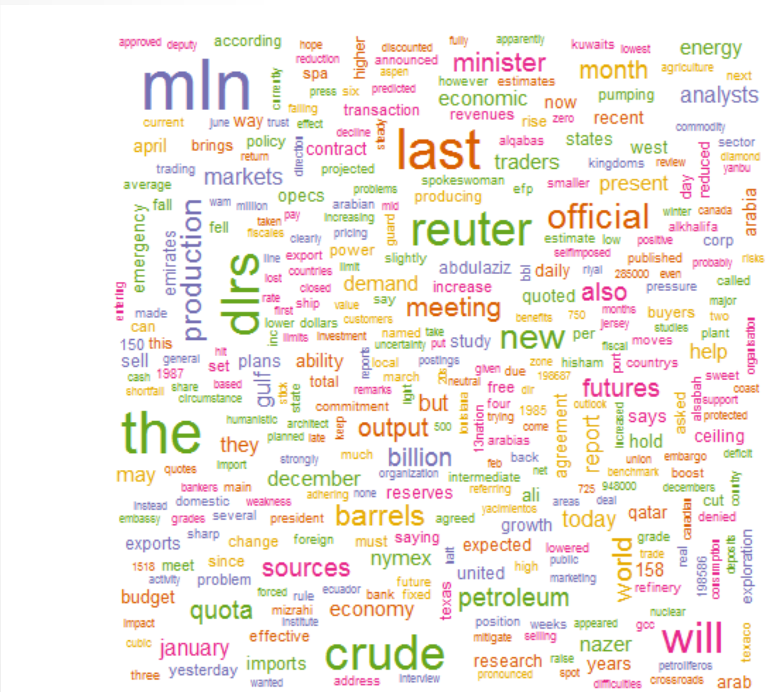
Zoom | Export | Clear All



Example: Flights information from Munich Airport







Text Data

Example: Reuters' articles about crude oil

★STATE OF THE UNION 2015★

AFTER CLASS

1. Complete the survey
2. Get a Github account – learn to fork the repo
   a. *git – the simple guide*
   b. *Getting your project on GitHub*
3. Install R and Rstudio

And About Waiting List…