Data Science in Practice

Saghir Bashir

{i} ilustat
www.ilustat.com

Definition: “Data Science”

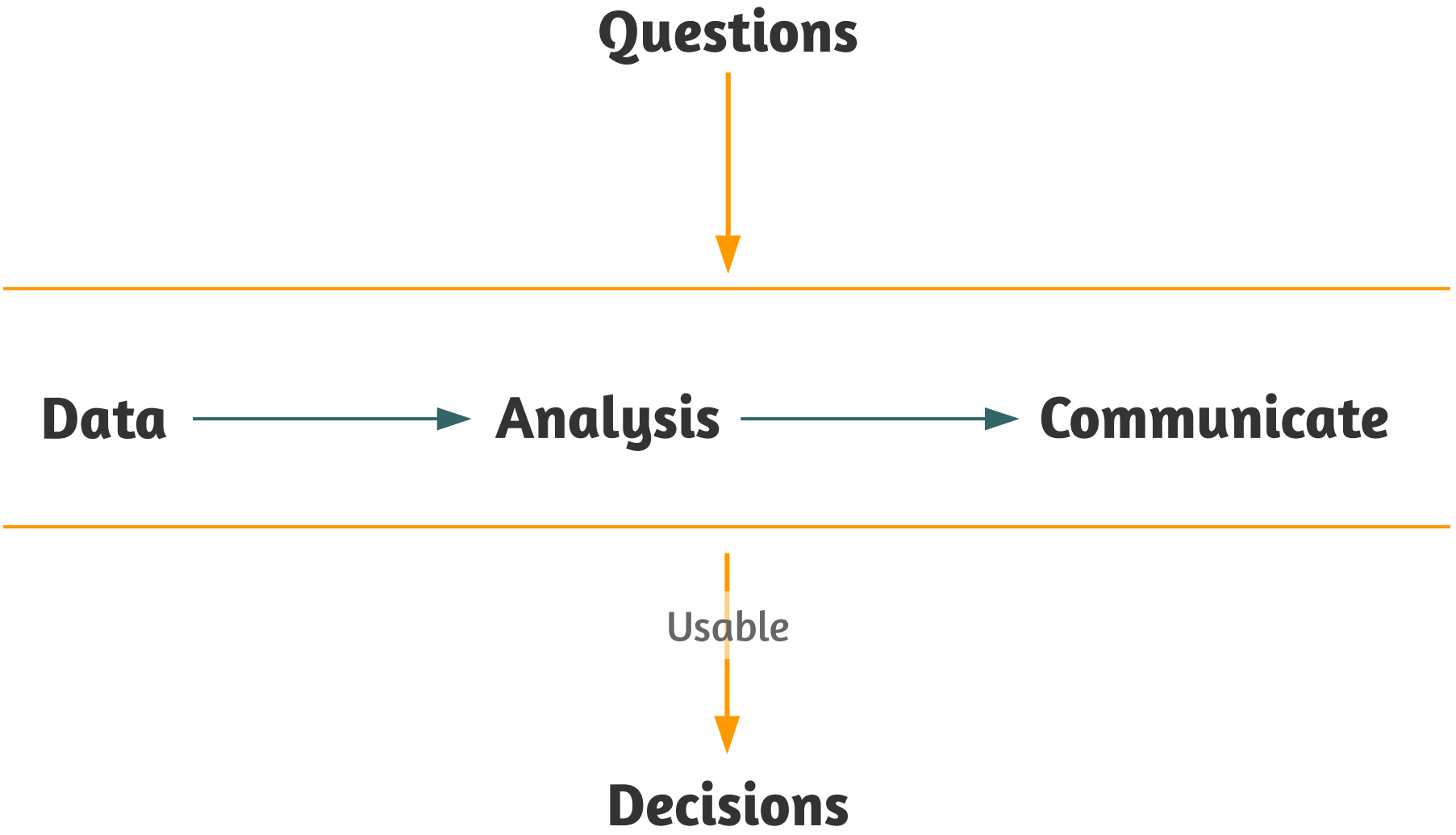
Generally accepted definition *does not exist!*

Presentation definition:

- **Subject Matter Expertise**
- **Statistics**
- **Statistical Programming**

For some Data Science is “Applied Statistics”

Data Science / Applied Statistics



Machine Learning is NOT Data Science!
Data Mining is NOT Data Science!

Machine Learning and Data Mining are types of analyses that you might perform as part of doing Data Science

Outline

Objectives & Background

Real Life Question

Real Life Data

Data Related Challenges

Summary

Objectives

My objectives are to:

- **Show you a real life example**
- **Encourage Data Science thinking**
- **Prepare you for real life Data Science**
  - **Help you embrace your vulnerabilities**
  - **Build your strength from them**

Background

My First Major Data Science Project

- **EU Cancer Mortality Predictions**
- **Years 2000 to 2010**
- **15 European Union Countries**
- **20 Cancer Sites**
- **Data varied between countries (up to 1996)**

Cancer in the EU

Homepage

Homepage
Introduction
Data
Methods
Downloads
Links
Thanks!
Help
Comments

Please select the site and sex for which you would like to see a summary results table of cancer mortality projections in the European Union (EU):

Bucal cavity	Males / Females	**Ovary**	Females
Oesophagus	Males / Females	**Prostate**	Males
Stomach	Males / Females	**Bladder**	Males / Females
Colorectal	Males / Females	**Kidney**	Males / Females
Pancreas	Males / Females	**Brain & CNS**	Males / Females
Larynx	Males / Females	**Non-Hodgkin**	Males / Females
Lung	Males / Females	**Hodgkin's Disease**	Males / Females
Melanoma	Males / Females	**Leukaemia**	Males / Females
Breast	Females	**Remaining Sites**	Males / Females
Cervix uteri	Females	**All Cancers**	Males / Females
Corpus uteri	Females		

Cancer in the EU

Lung cancer, females. Projections (Mortality)

Projected
Observed

Summary
Rate by age
Num by age

Males
Females

Bucal cavity
Oesophagus
Stomach
Colorectal
Pancreas
Larynx
Lung
Melanoma
Breast
Cervix uteri

Lung cancer, females. Projected total number of deaths, projected crude rate and the projected ASR (Age Standardised Rate using the World standard population) for the European Union countries between 1968 and 2012.

Country		Period of Death								
		1968-1972	1973-1977	1978-1982	1983-1987	1988-1992	1993-1997	1998-2002	2003-2007	2008-2012
Austria	**Total**	2360	2716	3094	3463	3867	4294	4931	5833	7086
	Crude	*12.0*	*13.7*	*15.6*	*17.4*	*19.2*	*20.9*	*23.9*	*28.1*	*34.2*
	ASR	*6.2*	*6.8*	*7.5*	*8.2*	*9.1*	*10.3*	*11.8*	*13.7*	*16.0*
Belgium	**Total**	2069	2428	2846	3390	4050	4831	5758	6774	7811
	Crude	*8.4*	*9.7*	*11.3*	*13.4*	*15.9*	*18.7*	*22.1*	*25.8*	*29.7*
	ASR	*4.8*	*5.4*	*6.2*	*7.1*	*8.1*	*9.2*	*10.2*	*11.2*	*12.0*
Denmark	**Total**	1714	2357	3317	4483	5484	6413	7337	8154	8791
	Crude	*13.8*	*18.5*	*25.6*	*34.5*	*42.0*	*48.9*	*55.9*	*62.2*	*67.4*
	ASR	*8.6*	*11.2*	*15.2*	*20.0*	*23.6*	*26.5*	*28.4*	*29.1*	*28.5*
Finland	**Total**	678	895	1166	1442	1704	1953	2197	2416	2626
	Crude	*5.7*	*7.4*	*9.4*	*11.4*	*13.3*	*14.9*	*16.5*	*18.0*	*19.3*
Portugal	**Total**	.	1036	1299	1674	2021	2385	2769	3178	3601
	Crude	.	*4.2*	*5.0*	*6.4*	*7.8*	*9.4*	*10.9*	*12.6*	*14.3*
	ASR	.	*3.1*	*3.5*	*3.9*	*4.4*	*4.9*	*5.3*	*5.7*	*6.2*

Some Thoughts

- Time management *"guess-timates"*:
  - ~80% Data processing, cleaning, understanding, ...
  - ~10% Analysis & ~10% Communication
- Main analysis was a Bayesian Model
  - Each full run took ~4 days non stop on a PC
- I loved working on this project
  - High Pressure & STRESS but lots of learning!
  - Software: STATA, WinBugs, LaTeX, HTML, R, Linux, ...

Workshop Plan

- Focus on Data processing, cleaning, understanding, ...
- Analysis & Communication for another day
- Look at a real life problems
  - There are no right answers!
  - You have to find the best compromises
  - Most importantly you must be able to defend your choices!
- Have fun!

Objectives & Background

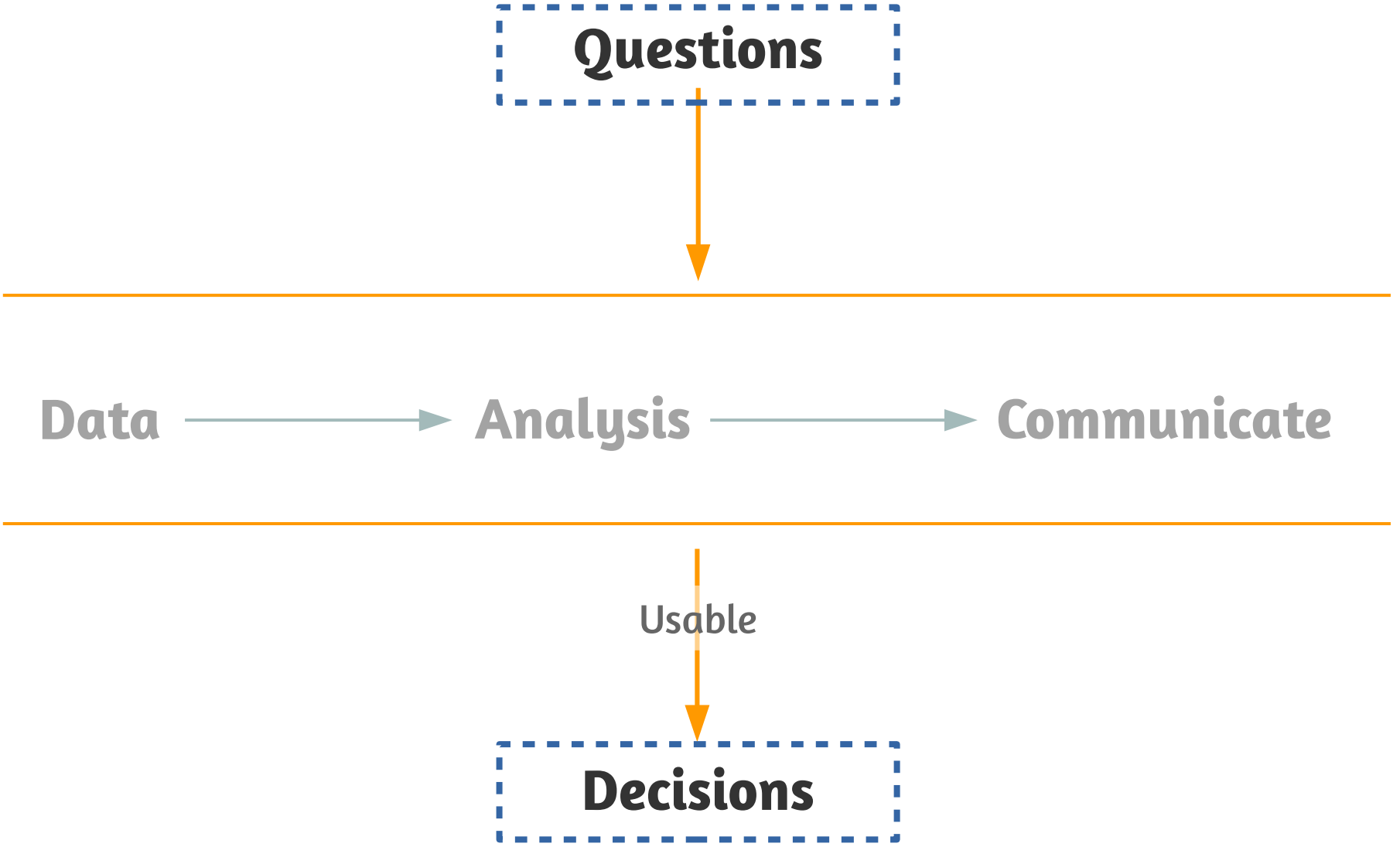
Real Life Questions

Real Life Data

Data Related Challenges

Summary

Data Science In Practice



Questions

> What are the trends and predictions for cancer mortality in Portugal?

> How does Portugal compare to other countries?

Exercise:

- → Discuss what these questions mean to you?
- → What needs to be defined?
- → What actions and/or decisions could be taken?

Questions

- **What are the trends and predictions for cancer mortality in Portugal?**
  - → **How is cancer defined? By site? All combined?**
  - → **What Trends? Current trends starting when?**
  - → **What would you like to predict? Deaths in 10 or 20 years?**
  - → **You need a subject matter expert**
- **How does Portugal compare to other countries?**
  - → **Which countries? European? Asian? Rich? Poor?**

Decisions & Actions

- **Introduce interventions to reduce cancer mortality**
  - **Assess existing interventions**
- **Plan healthcare, social care & other services**
  - **Doctors, nurses, clinics, hospices, medications, ...**
  - **Financial & logistical (e.g. specialised hospitals)**
- **Anticipate future risks and potential changes**

Objectives & Background

Real Life Question

Real Life Data

Data Related Challenges

Summary

WHO Mortality Database

World Health Organisation Mortality Database

- **Data reported by countries**
  - **Civilian registration systems**
- **Compilation of mortality data by:**
  - **Age, sex, year and cause of death**
  - **International Classification of Diseases (ICD)**
- **Available from:**

http://www.who.int/healthinfo/mortality_data/en/

Health statistics and information systems

Health statistics and information systems
Topics
Classifications and indicators
Data collection tools
Data analysis tools
Statistics
Country monitoring and evaluation
Monitoring universal health coverage
Publications

WHO Mortality Database

The WHO Mortality Database is a compilation of mortality data by age, sex and cause of death, as reported annually by Member States from their civil registration systems.

– Access the online database
Number of deaths and age-standardized death rates by country, year, cause, sex and age are presented in a user-friendly application. Cause-of-death data coded according to the ICD-9 and ICD-10 are provided since 1979 to date. Population and live births are provided.

– Query the online database
Cause of Death Query Online (CoDQL) is a user-friendly tool that allows users to extract easily cause-of-death data by country, year, sex and age. Data since 1950 to date as coded according to the ICD-7, 8, 9 and 10 are available. The tool also enables detailed causes of death to be aggregated to form broader cause-category according to the users' need.

– Download raw data files
Basic underlying raw data files, together with the necessary instructions, file structures, code reference tables, etc. These data can be used by

Health statistics and information systems

Health statistics and information systems
Topics
Classifications and indicators
Data collection tools
Data analysis tools
Statistics
Country monitoring and evaluation
Monitoring universal health coverage
Publications

Download the raw data files of the WHO Mortality Database

Data files - Last updated: 01 October 2017

These files do not constitute a user-friendly data collection which the average user can download and access. These are the basic underlying raw data files, together with the necessary instructions, file structures, code reference tables, etc. which can be used by institutions and organizations which need access at this level of detail **mainly for research purposes** and have available the required information technology (IT) resources to use this information. These files will not open in programs like Excel; please refer to the "Documentation.zip" file hereafter for more information on systems requirements.

It should be noted that these data are transmitted on the understanding that no use will be made of them for commercial purposes and that no such permission or right to use may be implied thereby. WHO requests data users to adhere to the guidelines outlined on the next page.

Documentation
zip, 87kb
Contains a Word file with information on the WHO Mortality Database, file specifications and list of causes of death. Last updated: 01 October 2017.

Availability
zip, 224kb
Contains an Excel file with the list of countries-years available for the mortality and population data. Last updated: 01 October 2017.

Country codes
zip, 2kb
Country codes and names. Last updated: 03 November 2014.

Notes
zip, 0kb
Notes pertaining to data for some countries-years. Last updated: 01 October 2017.

Population and live births
zip, 582kb
Reference populations and live births (for regular users, figures are now in units). Last updated: 01 October 2017.

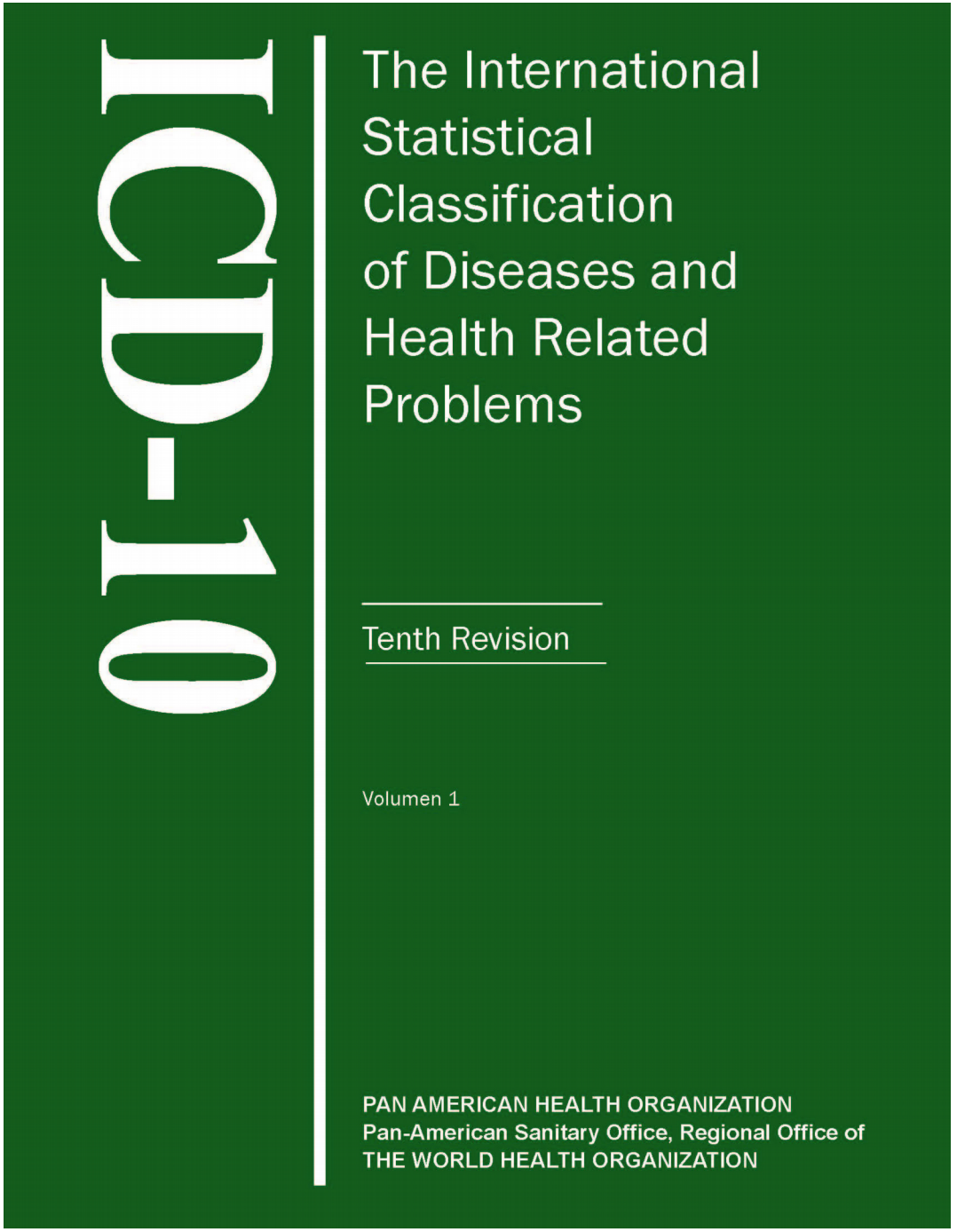
Mortality, ICD-7
zip, 4.88Mb
Data file containing the detailed mortality data for the seventh revision of the ICD (International Classification of Diseases). Last updated: 18 February 2004.

Mortality, ICD-8
zip, 5.45Mb
Data file containing the detailed mortality data for the eighth revision of the ICD (International Classification of Diseases). Last updated: 09 July 2012.

Mortality, ICD-9
zip, 13.51Mb
Data file containing the detailed mortality data for the ninth revision of the ICD (International Classification of Diseases). Last updated: 29 March 2017.

Mortality, ICD-10 (part 1/2)
zip, 15.50Mb
Data file containing the detailed mortality data for the tenth revision of the ICD (International Classification of Diseases). Last updated: 01 October 2017.

Mortality, ICD-10 (part 2/2)
zip, 23.87Mb
Data file containing the detailed mortality data for the tenth revision of the ICD (International Classification of Diseases). Last updated: 01 October 2017.


ICD-10
The International Statistical Classification of Diseases and Health Related Problems
Tenth Revision
Volumen 1
PAN AMERICAN HEALTH ORGANIZATION
Pan-American Sanitary Office, Regional Office of
THE WORLD HEALTH ORGANIZATION

International Classification of Diseases

- **Global health information standard for mortality and morbidity statistics**
- **Defines diseases, disorders, injuries and other related health conditions**
- **Useful for:**
  - ***Storing, retrieving & analysing health information***
  - ***Sharing and comparing health information***

Now what?

- We will work with a subset of the WHO data
  - Issues have been simplified for this workshop
  - They still reflect the reality
- Three areas will be covered
  - Making the raw data usable
  - Handling data difference
  - Understanding & managing “standards” & “definitions”

Countries...

Portugal

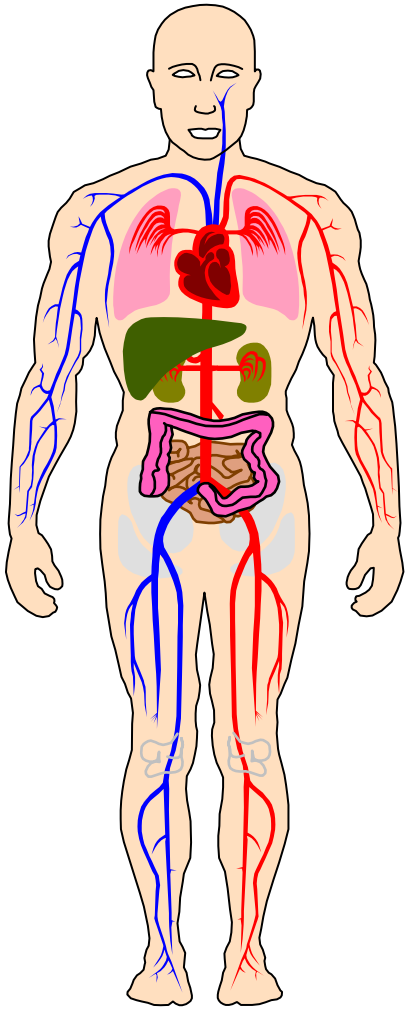
Greece

Hungary

Spain

Sweden

Cancers...



Colon

Leukaemia

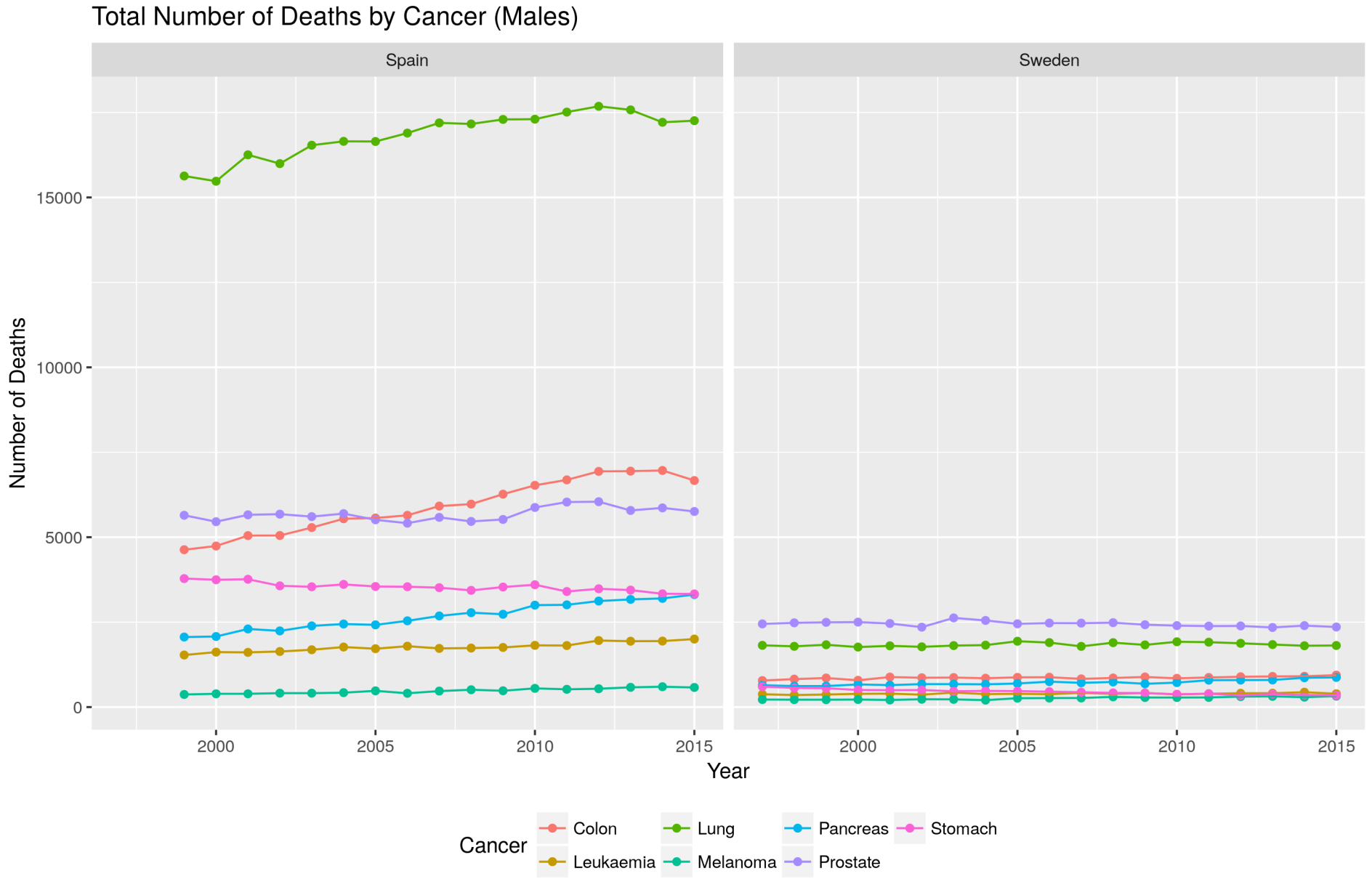
Lung (incl. trachea and bronchus)

Melanoma of skin

Pancreas

Prostate

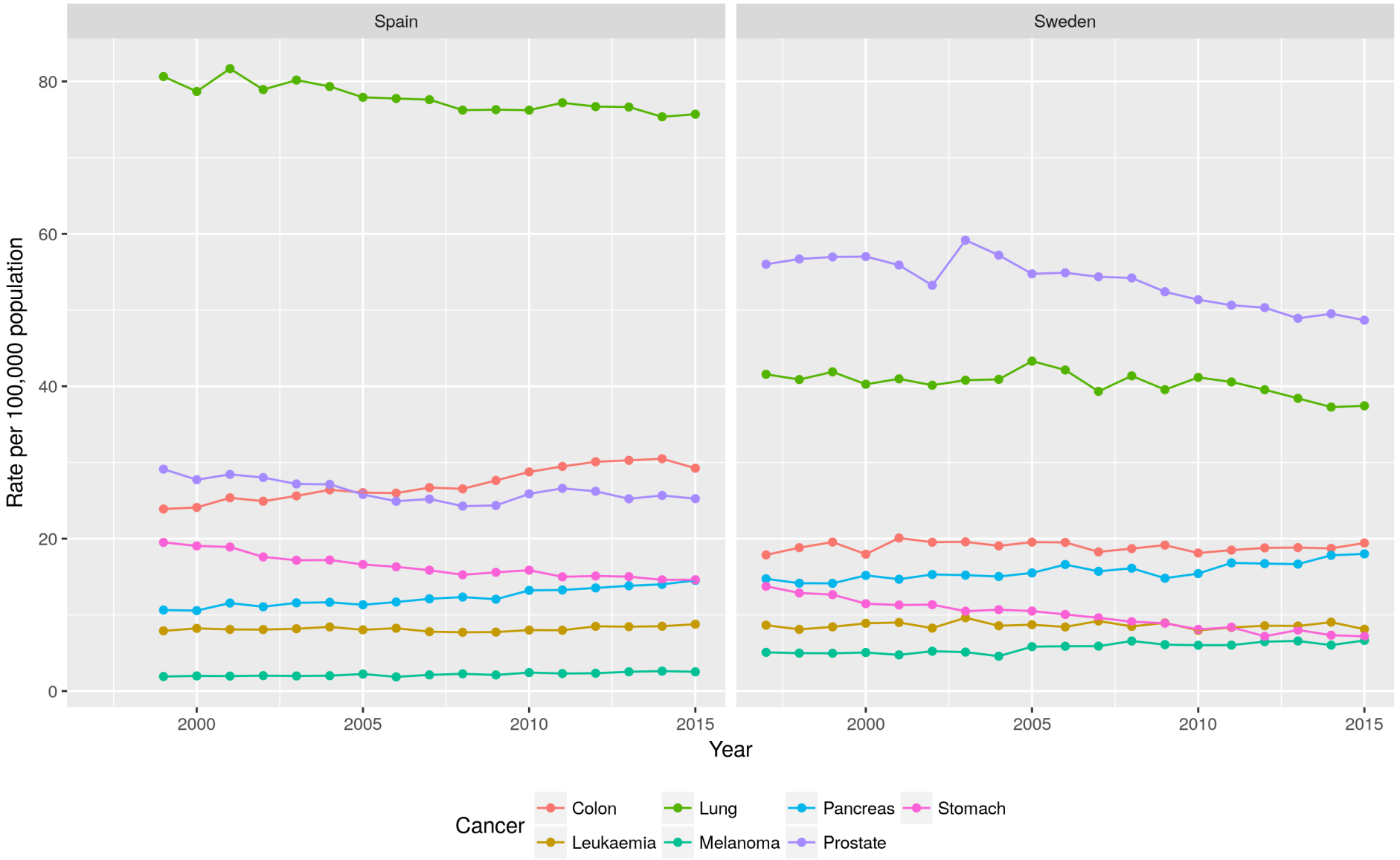
Stomach


Total Number of Deaths by Cancer (Males)
Spain
Sweden
Number of Deaths
15000
10000
5000
0
2000
2005
2010
2015
Year
Cancer
Colon
Lung
Pancreas
Stomach
Leukaemia
Melanoma
Prostate

Rate per 100,000 Populations

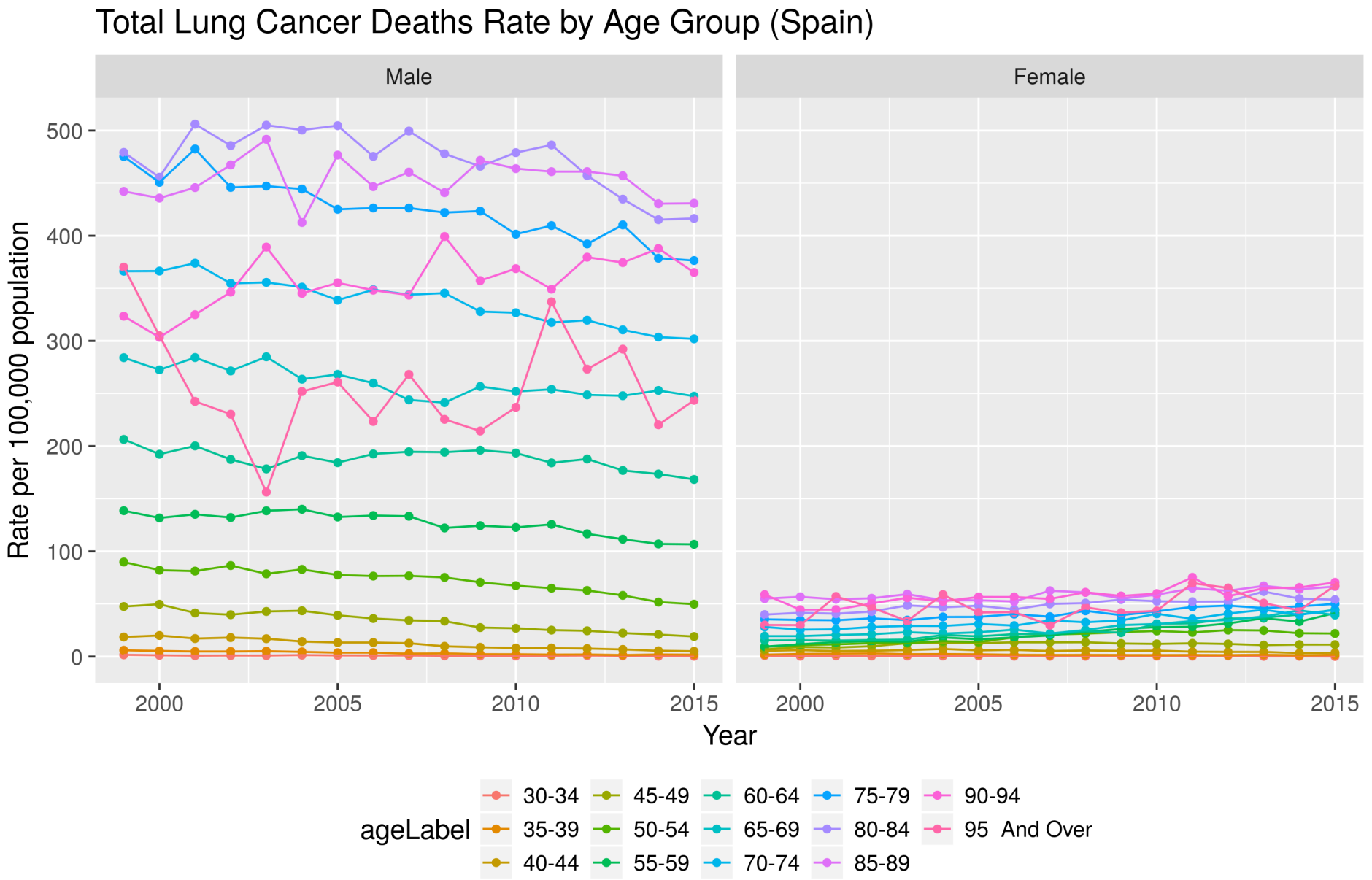
- Population varies by country, time and sex
  - Spain ~46M & Sweden ~10M
  - Better to use Rate rather number of deaths
  - Allows for fairer comparisons
- Rate = 100,000* (Number of Deaths / Population)
  - For 100,000 people how many deaths would there be

Note: I use the UK notation where comma is “thousands” separator not a decimal place.

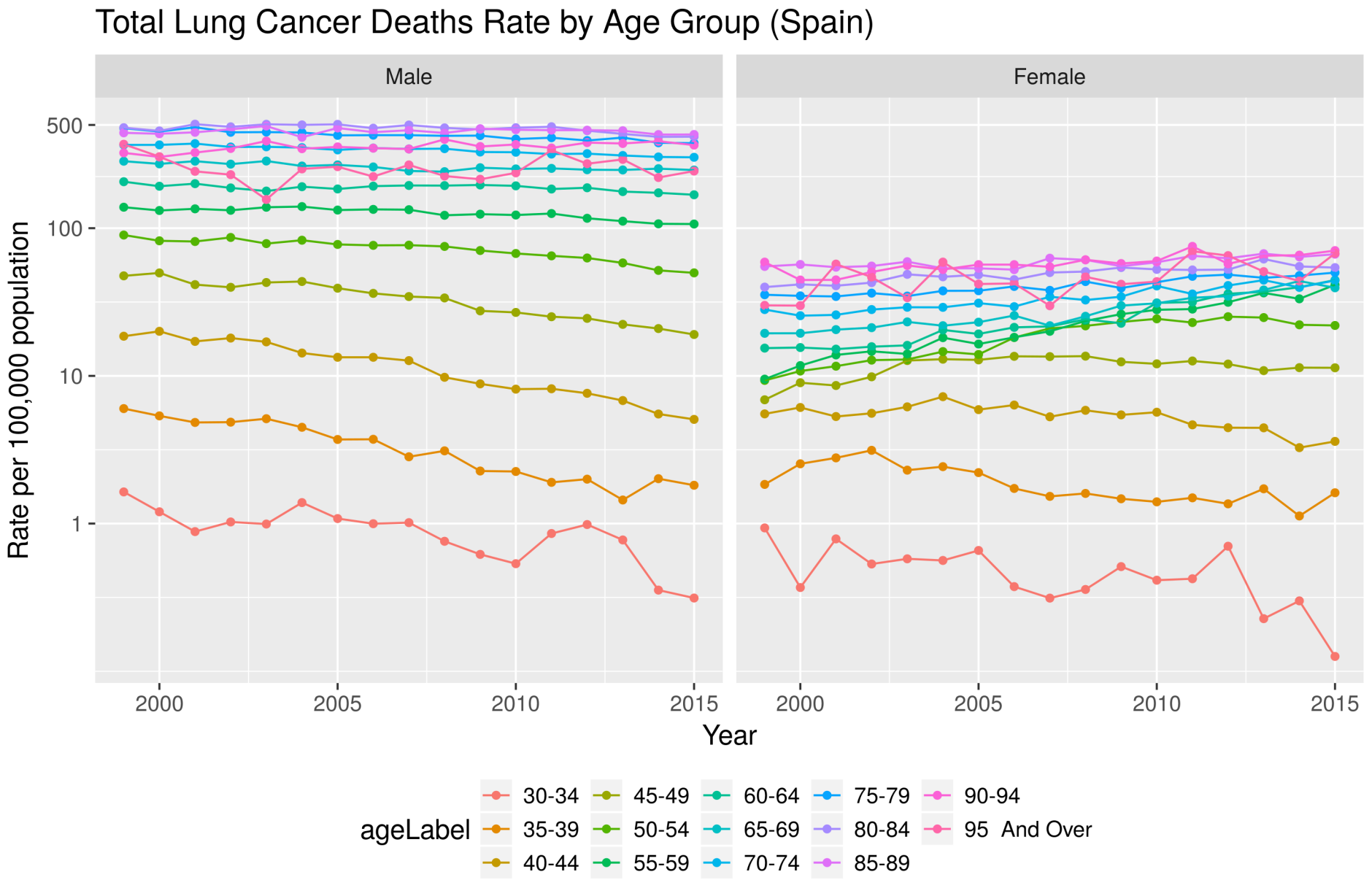

Total Deaths Rate by Cancer (Males)
Spain
Sweden
Rate per 100,000 population
0
20
40
60
80
2000
2005
2010
2015
Year
Cancer
Colon
Lung
Pancreas
Stomach
Leukaemia
Melanoma
Prostate

Deaths Rates / 100,000 Population (M)

Country	Cancer	2000	2005	2010	2015
Spain	Colon	24.1	26.0	28.8	29.2
	Leukaemia	8.2	8.0	8.0	8.8
	Lung	78.7	77.9	76.2	75.7
	Melanoma	2.0	2.2	2.4	2.5
	Pancreas	10.6	11.3	13.2	14.5
	Prostate	27.7	25.8	25.9	25.2
	Stomach	19.0	16.6	15.9	14.6
Sweden	Colon	18.0	19.5	18.1	19.4
	Leukaemia	8.9	8.7	8.0	8.1
	Lung	40.3	43.3	41.2	37.4
	Melanoma	5.1	5.8	6.0	6.6
	Pancreas	15.2	15.5	15.4	18.0
	Prostate	57.0	54.8	51.4	48.7
	Stomach	11.5	10.5	8.1	7.2


Total Lung Cancer Deaths Rate by Age Group (Spain)
Male
Female
Rate per 100,000 population
500
400
300
200
100
0
2000
2005
2010
2015
Year
ageLabel
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85-89
90-94
95 And Over

Log Scale


Total Lung Cancer Deaths Rate by Age Group (Spain)
Male
Female
Rate per 100,000 population
500
100
10
1
2000
2005
2010
2015
Year
ageLabel
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85-89
90-94
95 And Over

Lung Cancer Rate – Spain (M)

Age Group	2000	2005	2010	2015
30-34	1.2	1.1	0.5	0.3
35-39	5.4	3.7	2.3	1.8
40-44	20.0	13.4	8.1	5.1
45-49	49.8	39.2	26.9	19.1
50-54	82.1	77.6	67.4	49.8
55-59	131.8	132.7	122.8	106.7
60-64	192.3	184.3	193.5	168.3
65-69	272.5	268.2	251.9	247.5
70-74	366.4	338.8	326.8	301.9
75-79	450.8	425.1	401.5	376.4
80-84	455.6	504.7	479.0	416.4
85-89	435.8	476.7	463.8	430.8
90-94	303.5	355.2	368.7	365.2
95+	304.9	260.9	236.9	243.6

Exercise

- In the handouts you have a sample of
  - → ICD10 Mortality Data
  - → Population Data
- Using the variables definitions:
  - → What do you understand about the data?
  - → How would you restructure the data to produce the graphs and tables in the previous slides?

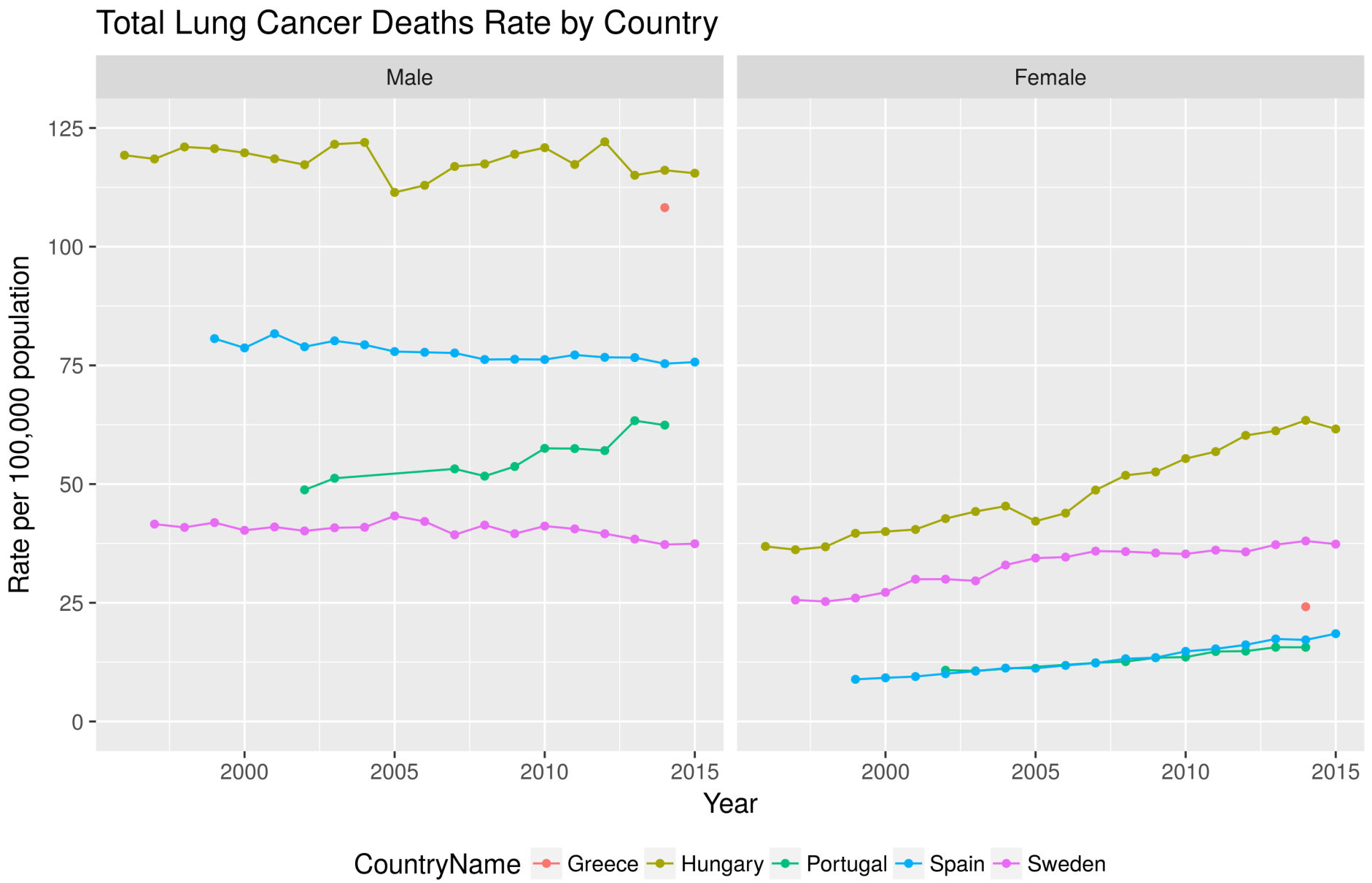
Objectives & Background

Real life Question

Real life Data

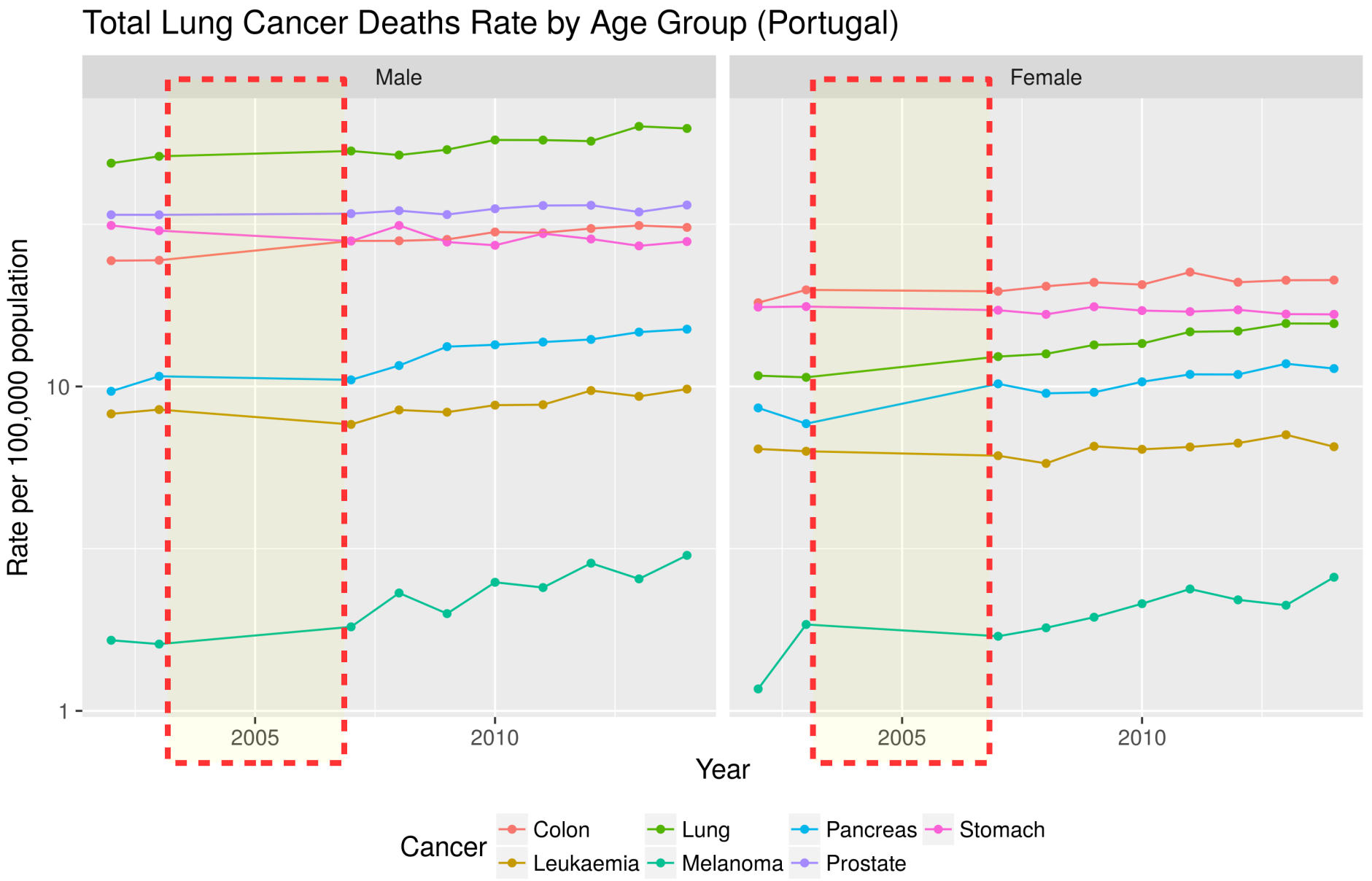
Data Related Challenges

Summary


Total Lung Cancer Deaths Rate by Country
Male
Female
Rate per 100,000 population
125
100
75
50
25
0
2000
2005
2010
2015
Year
CountryName
Greece
Hungary
Portugal
Spain
Sweden

Data Related Challenges (1)

- How do you define a country?
  - Germany was formerly East and West Germany
  - Czech Republic & Slovakia were formerly Czechoslovakia
- How do you define the European Union?
  - Start (1951) 6 countries – Now (2018) 28 countries
  - The UK has voted to leave (2019)
  - How can you fairly compare to the EU average?


Total Lung Cancer Deaths Rate by Age Group (Portugal)
Male
Female
Rate per 100,000 population
10
1
2005
2010
Year
Cancer
Colon
Lung
Pancreas
Stomach
Leukaemia
Melanoma
Prostate

ICD10 Data?

Year	Greece	Hungary	Portugal	Spain	Sweden
1996	-	Y	-	-	-
1997	-	Y	-	-	Y
1998	-	Y	-	-	Y
1999	-	Y	-	Y	Y
2000	-	Y	-	Y	Y
2001	-	Y	-	Y	Y
2002	-	Y	Y	Y	Y
2003	-	Y	Y	Y	Y
2004	-	Y	-	Y	Y
2005	-	Y	-	Y	Y
2006	-	Y	-	Y	Y
2007	-	Y	Y	Y	Y
2008	-	Y	Y	Y	Y
2009	-	Y	Y	Y	Y
2010	-	Y	Y	Y	Y
2011	-	Y	Y	Y	Y
2012	-	Y	Y	Y	Y
2013	-	Y	Y	Y	Y
2014	Y	Y	Y	Y	Y
2015	-	Y	-	Y	Y

ICD9 Data?

Year	Greece	Hungary	Portugal	Spain	Sweden
1994	Y	Y	Y	Y	Y
1995	Y	Y	Y	Y	Y
1996	Y	-	Y	Y	Y
1997	Y	-	Y	Y	-
1998	Y	-	Y	Y	-
1999	Y	-	Y	-	-
2000	Y	-	Y	-	-
2001	Y	-	Y	-	-
2002	Y	-	-	-	-
2003	Y	-	-	-	-
2004	Y	-	-	-	-
2005	Y	-	-	-	-
2006	Y	-	-	-	-
2007	Y	-	-	-	-
2008	Y	-	-	-	-
2009	Y	-	-	-	-
2010	Y	-	-	-	-
2011	Y	-	-	-	-
2012	Y	-	-	-	-
2013	Y	-	-	-	-

Cancer Dictionary

Cancer Site	ICD 9 Code	ICD 10 Code
All cancers	140-208	C00-C97,B21
Colon	153	C18
Leukaemia	204-208	C91-C95
Lung (incl. trachea & bronchus)	162	C33-C34
Melanoma of skin	172	C43
Pancreas	157	C25
Prostate	185	C61
Stomach	151	C16

Data Related Challenges (2)

- How do you manage years where data is missing?
- How do you handle when
  - → Some diseases could be split whilst others could be joined
  - → Countries use ICD9 and ICD10 at different times
- Should you use ICD7 & ICD8 data?
  - → Is data going back to the 1950s comparable?
  - → How do you handle partial coverage of death registrations?

Some Comments

- Data Dictionaries & Standards are commonly used
  - → Food classification, Medication classification, ...
  - → Classifying professions, social economic status, ...
- They are useful for harmonisation
  - → “Fairer” comparison, data sharing & retrieval, ...
  - → Important to understand the details of implementation
- Compromises have to be made for analysis
  - → Document decisions and choices openly and transparently
  - → By being honest you will save yourself a lot of stress later

Objectives & Background

Real life Question

Real life Data

Data Related Challenges

Summary

Summary

We looked at data processing, cleaning & understanding

- **It usually takes the most time**
  - **Making the raw data usable**
  - **Handling data differences & abnormalities**
  - **Understanding & managing “standards” & “definitions”**
- **There are always surprises**
  - **Work closely with a subject matter expert**
  - **Document everything openly and transparently**

Thank you

Saghir Bashir

{i} ilustat
www.ilustat.com