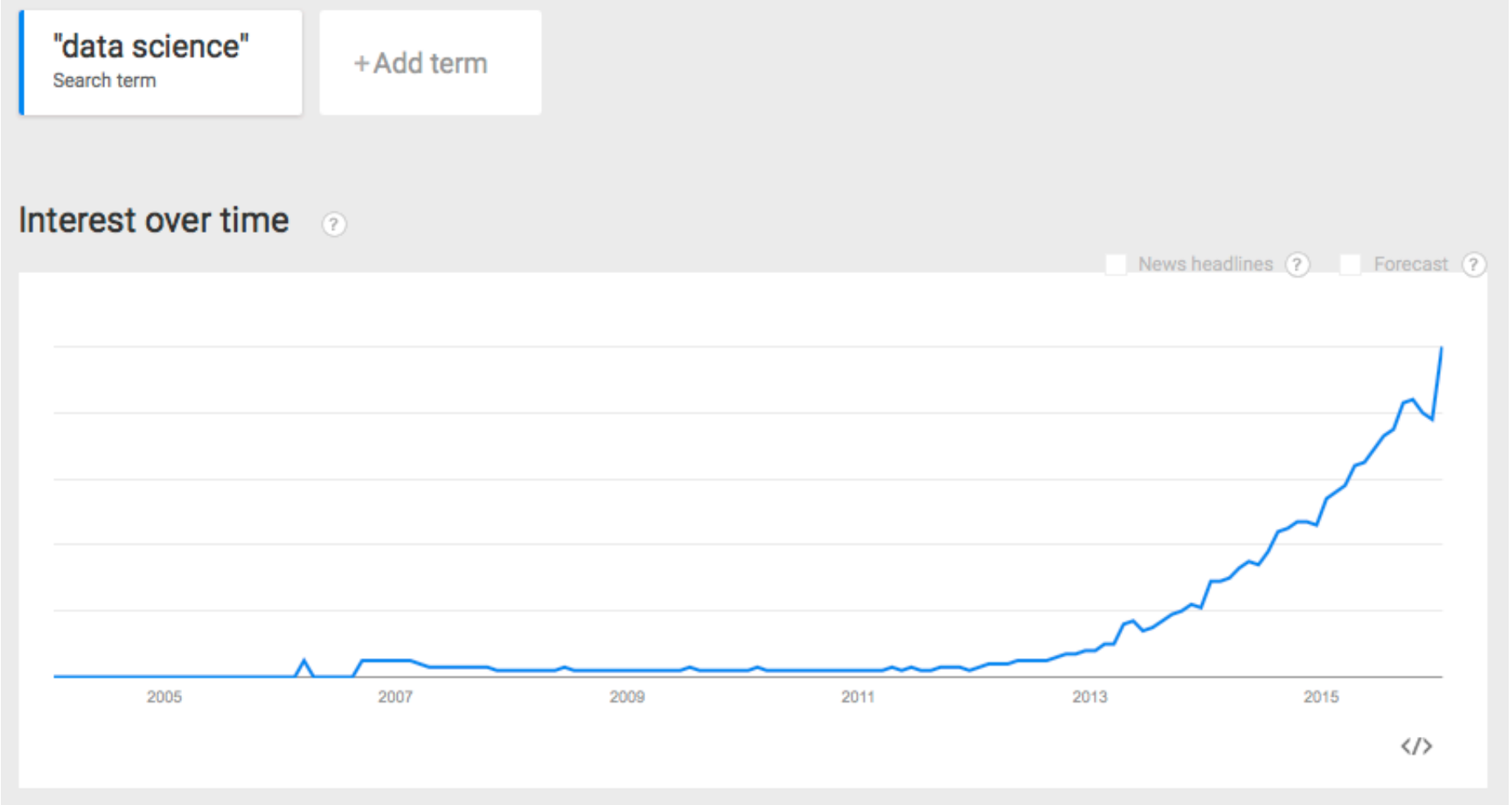
COMP10020
Introduction to Programming II
Getting to Know Your Data - Introducing Data Science & Pandas

Dr. Brian Mac Namee
brian.macnamee@ucd.ie
School of Computer Science
University College Dublin

WHAT IS DATA SCIENCE?

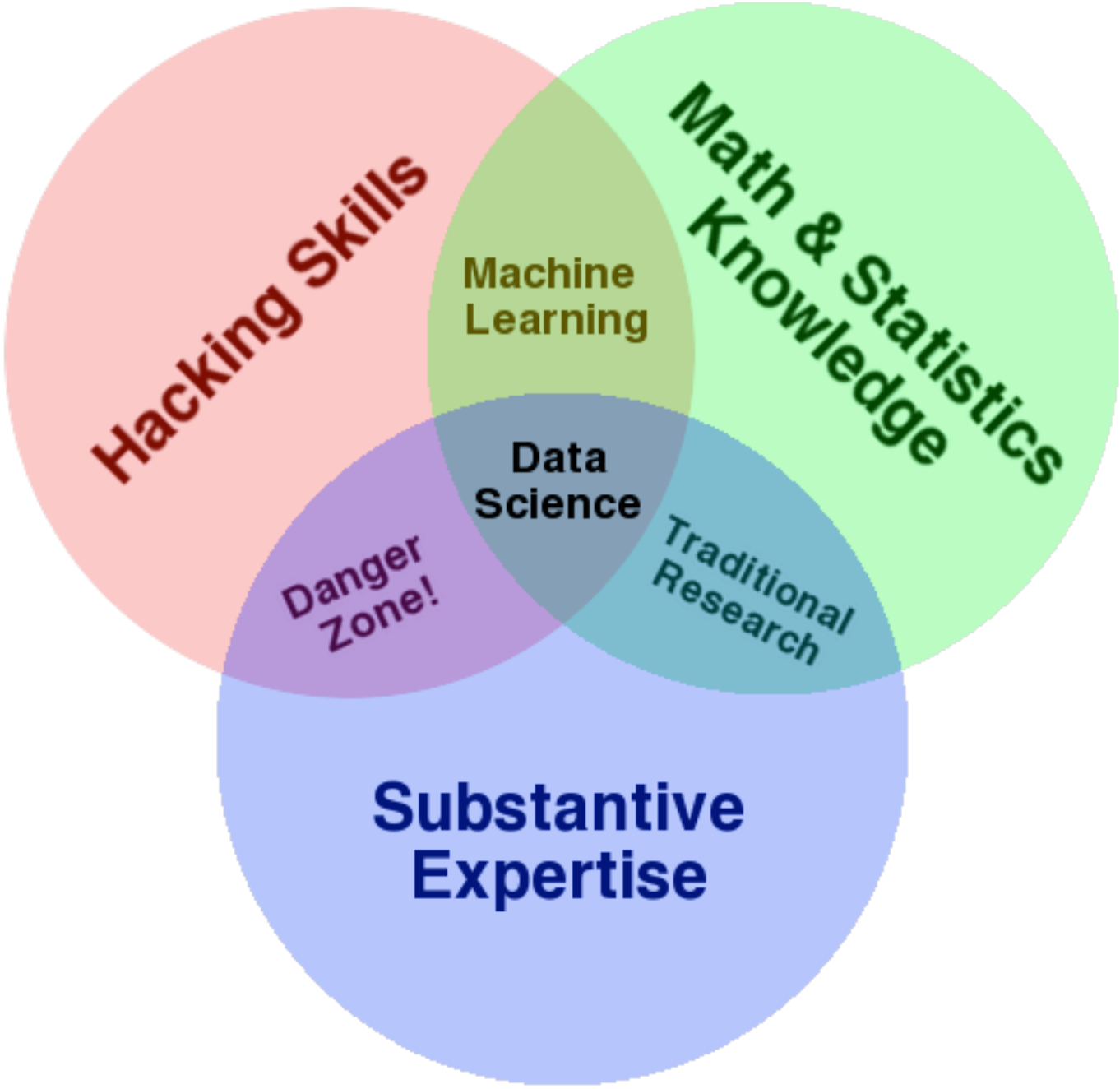
What Is Data Science?



https://www.google.com/trends/explore#q=%22data%20science%22&cmpt=q

What Is Data Science?

At its core Data Science is about developing the infrastructure and processes for dealing with data at scale, recognising and understanding patterns within large, diverse datasets, generating predictions based on these patterns, and creating revealing visualizations and crafting compelling narratives with and about data



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

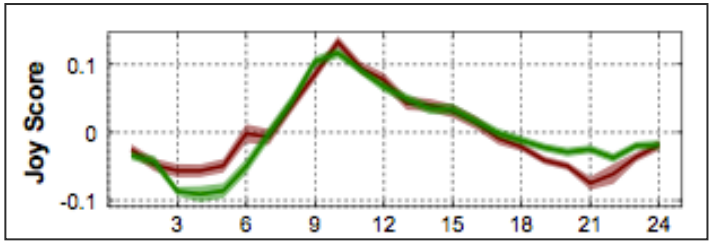
What Is Data Science?

Data Science brings together key ideas from multiple fields

- Computer science (algorithms, representation, visualization, application development)
- Statistics (modelling, analysis, prediction)
- Design (information design, interaction design)
- Psychology and cognitive science (language and perception),
- Humanities and social sciences (storytelling and narrative, social learning)

http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

DATA-DRIVEN EVERYTHING

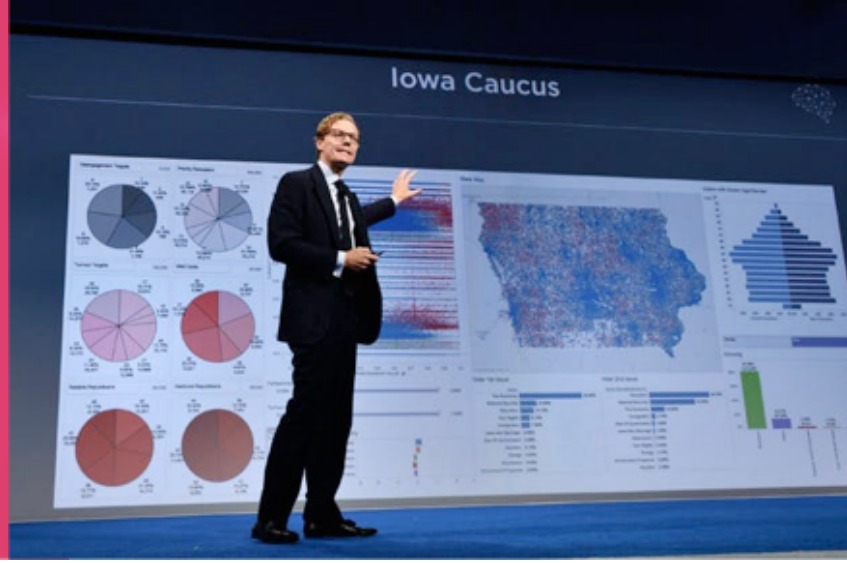


Data, Data Everywhere

A fun Twitter thread in which Dylan Curran illustrates all of the data large organisations like Google and Facebook collect on their users

https://twitter.com/iamdylancurran/status/977559925680467968

Cambridge Analytica



Same demographics, different personalities

Female
25-35 Years old
AMEX User

Openness	———————
Conscientiousness	———
Extraversion	———
Agreeableness	———
Neuroticism	——

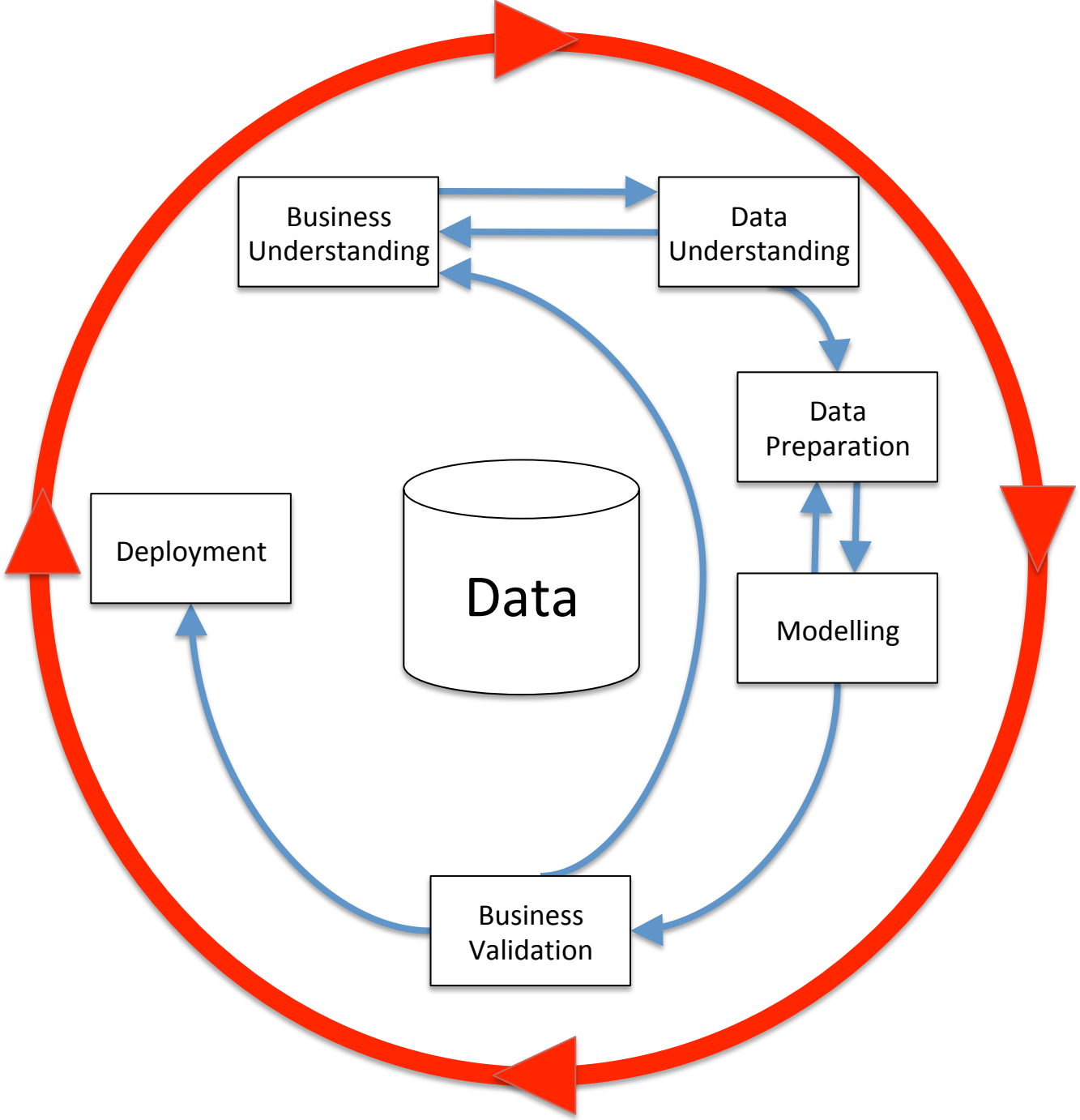
People with high openness and extraversion love new experiences they can share with lots of people.

Female
25-35 Years old
AMEX User

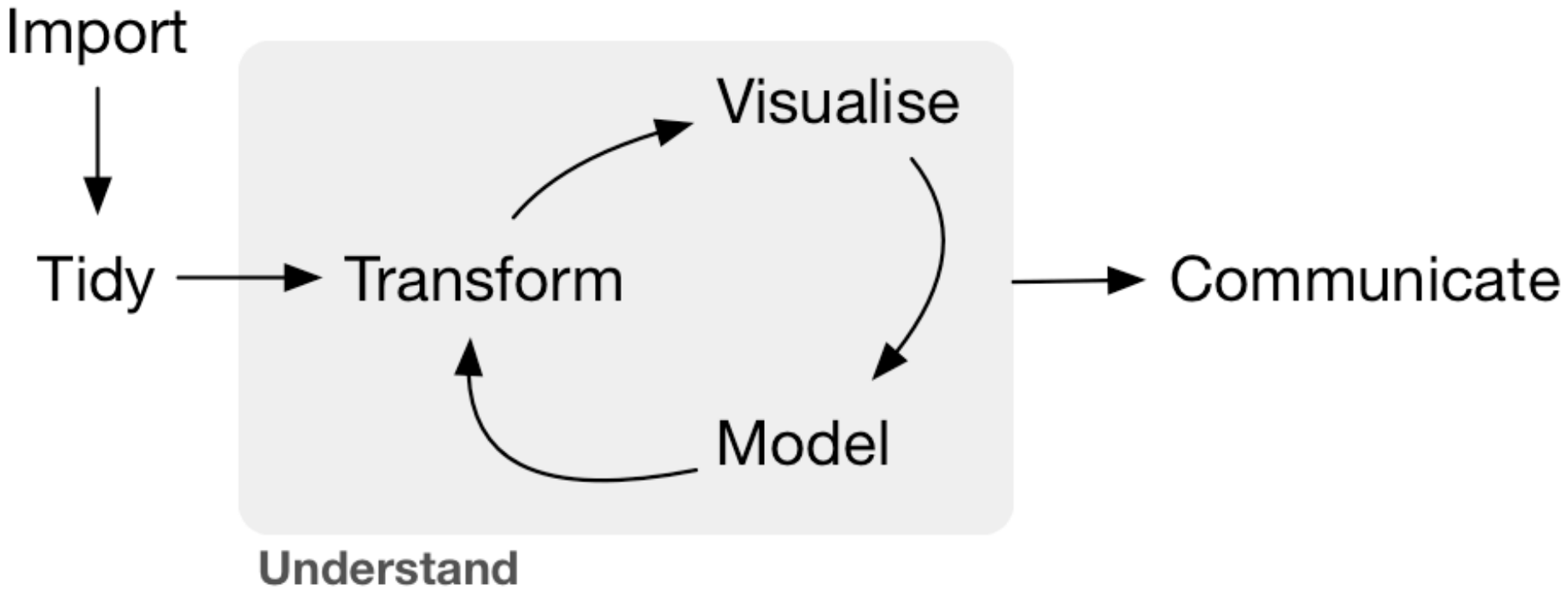
Openness	——
Conscientiousness	———
Extraversion	——
Agreeableness	———
Neuroticism	———

People with low openness and extraversion really value down time spent with their closest friends.

DATA SCIENCE PIPELINE



https://www.the-modeling-agency.com/crisp-dm.pdf

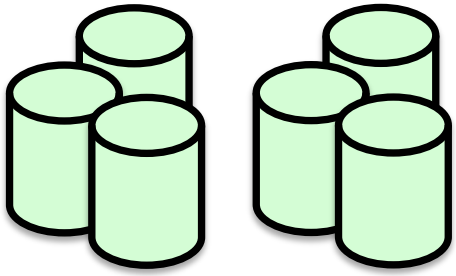


http://r4ds.had.co.nz/intro.html

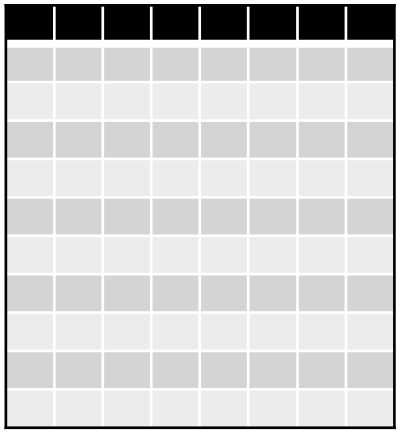
INTRODUCING DATA FRAMES


vodafone
26.22
STRAVA
SMS
BANKLINK
You
Tube
Withings
leap
Bank of
Ireland
Insta
EIGHT
HTML
Wi Fi
eflow
you first
CLUBCARD

Where Does Data Come From?

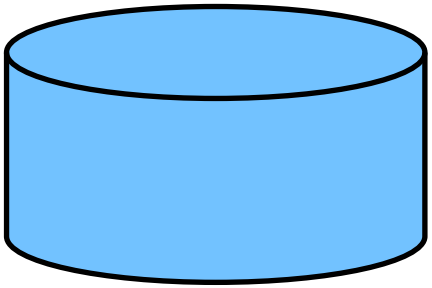


Operational Databases

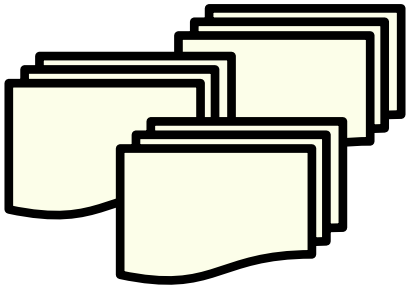


Data Frame

Exotic Data



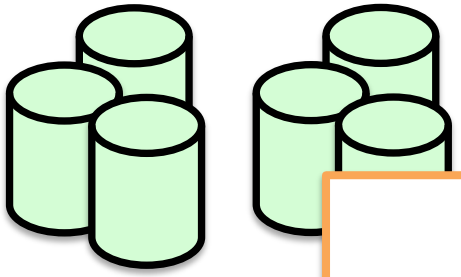
Data Warehouses & Data Marts



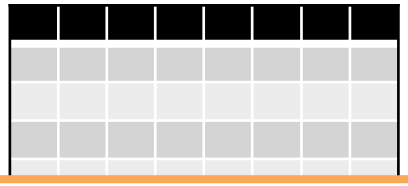
Flat Files

External Data Feeds

Where Does Data Come From?

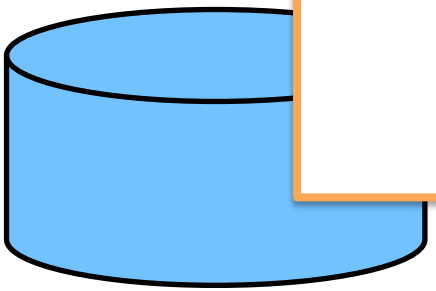


Operation
Database

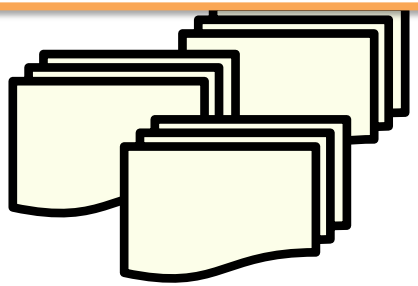


tic Data

To do anaything useful we first need to *get to know our data*



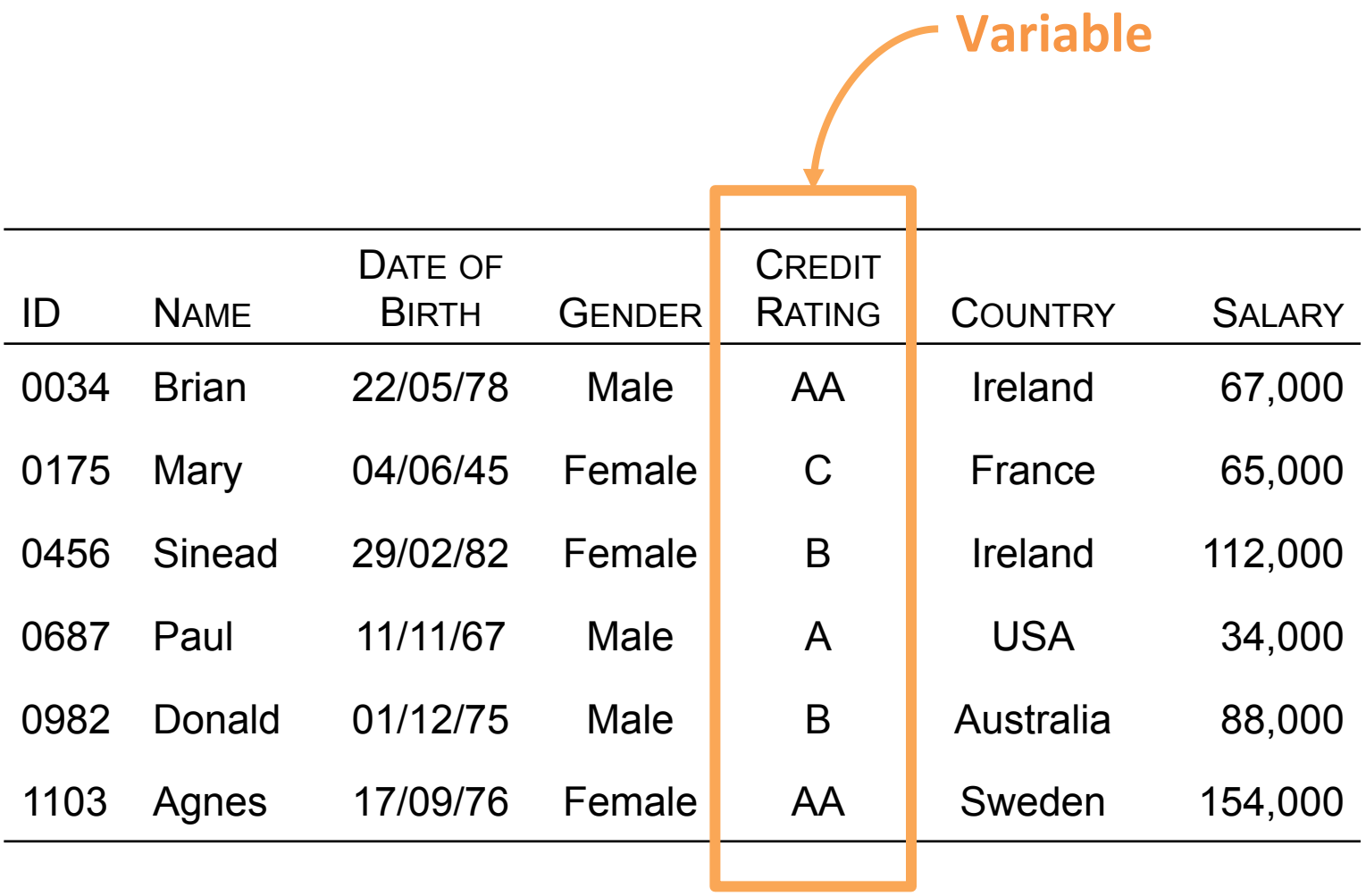
Data Warehouses &
Data Marts



Flat Files

External Data
Feeds

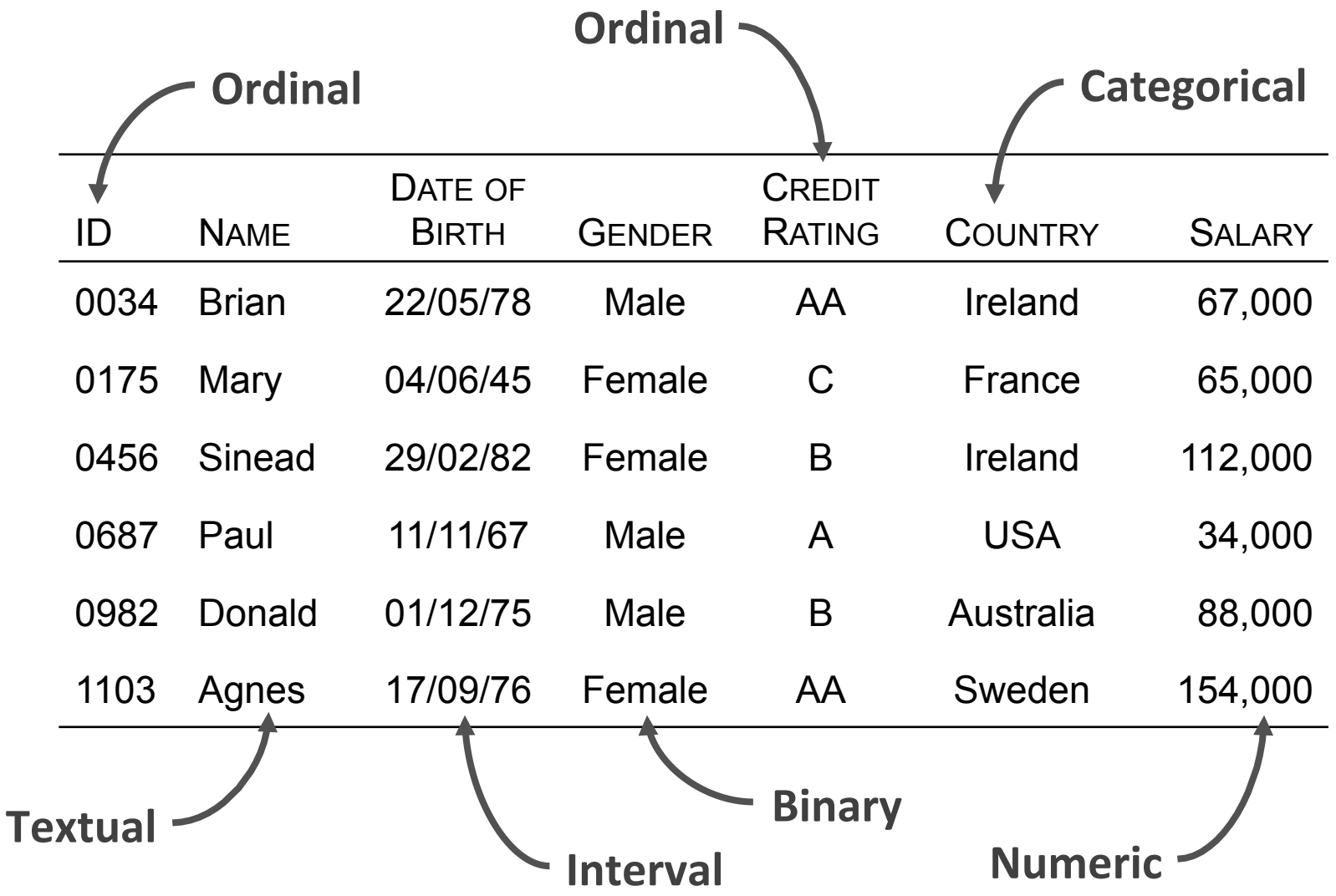
ID	Name	Date of Birth	Gender	Credit Rating	Country	Salary
0034	Brian	22/05/78	Male	AA	Ireland	67,000
0175	Mary	04/06/45	Female	C	France	65,000
0456	Sinead	29/02/82	Female	B	Ireland	112,000
0687	Paul	11/11/67	Male	A	USA	34,000
0982	Donald	01/12/75	Male	B	Australia	88,000
1103	Agnes	17/09/76	Female	AA	Sweden	154,000



ID	Name	Date of Birth	Gender	Credit Rating	Country	Salary
0034	Brian	22/05/78	Male	AA	Ireland	67,000
0175	Mary	04/06/45	Female	C	France	65,000
0456	Sinead	29/02/82	Female	B	Ireland	112,000
0687	Paul	11/11/67	Male	A	USA	34,000
0982	Donald	01/12/75	Male	B	Australia	88,000
1103	Agnes	17/09/76	Female	AA	Sweden	154,000

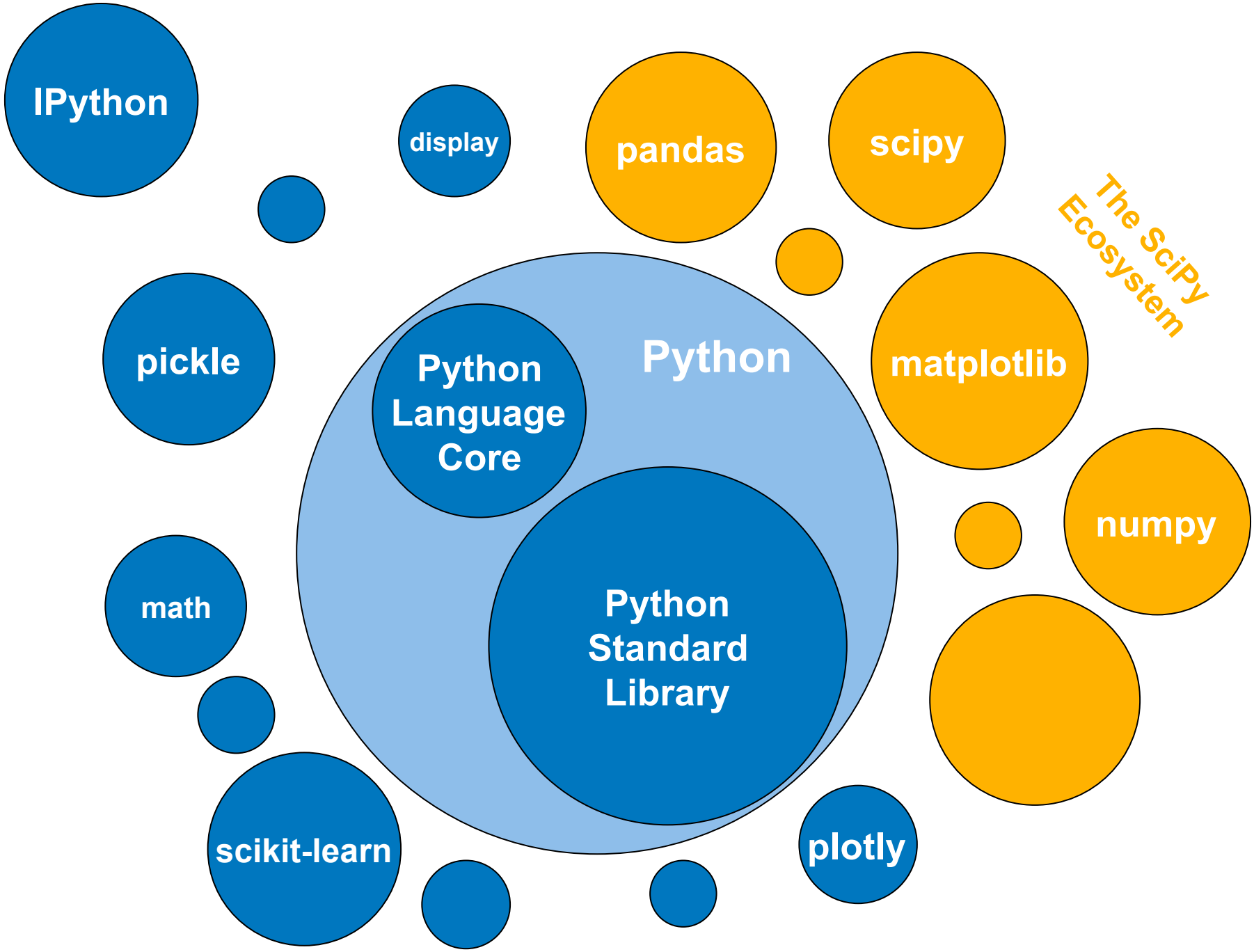
Record

ID	Name	Date of Birth	Gender	Credit Rating	Country	Salary
0034	Brian	22/05/78	Male	AA	Ireland	67,000
0175	Mary	04/06/45	Female	C	France	65,000
0456	Sinead	29/02/82	Female	B	Ireland	112,000
0687	Paul	11/11/67	Male	A	USA	34,000
0982	Donald	01/12/75	Male	B	Australia	88,000
1103	Agnes	17/09/76	Female	AA	Sweden	154,000


Ordinal
Ordinal
Categorical
ID
NAME
DATE OF BIRTH
GENDER
CREDIT RATING
COUNTRY
SALARY
0034 Brian 22/05/78 Male AA Ireland 67,000
0175 Mary 04/06/45 Female C France 65,000
0456 Sinead 29/02/82 Female B Ireland 112,000
0687 Paul 11/11/67 Male A USA 34,000
0982 Donald 01/12/75 Male B Australia 88,000
1103 Agnes 17/09/76 Female AA Sweden 154,000
Textual
Interval
Binary
Numeric

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

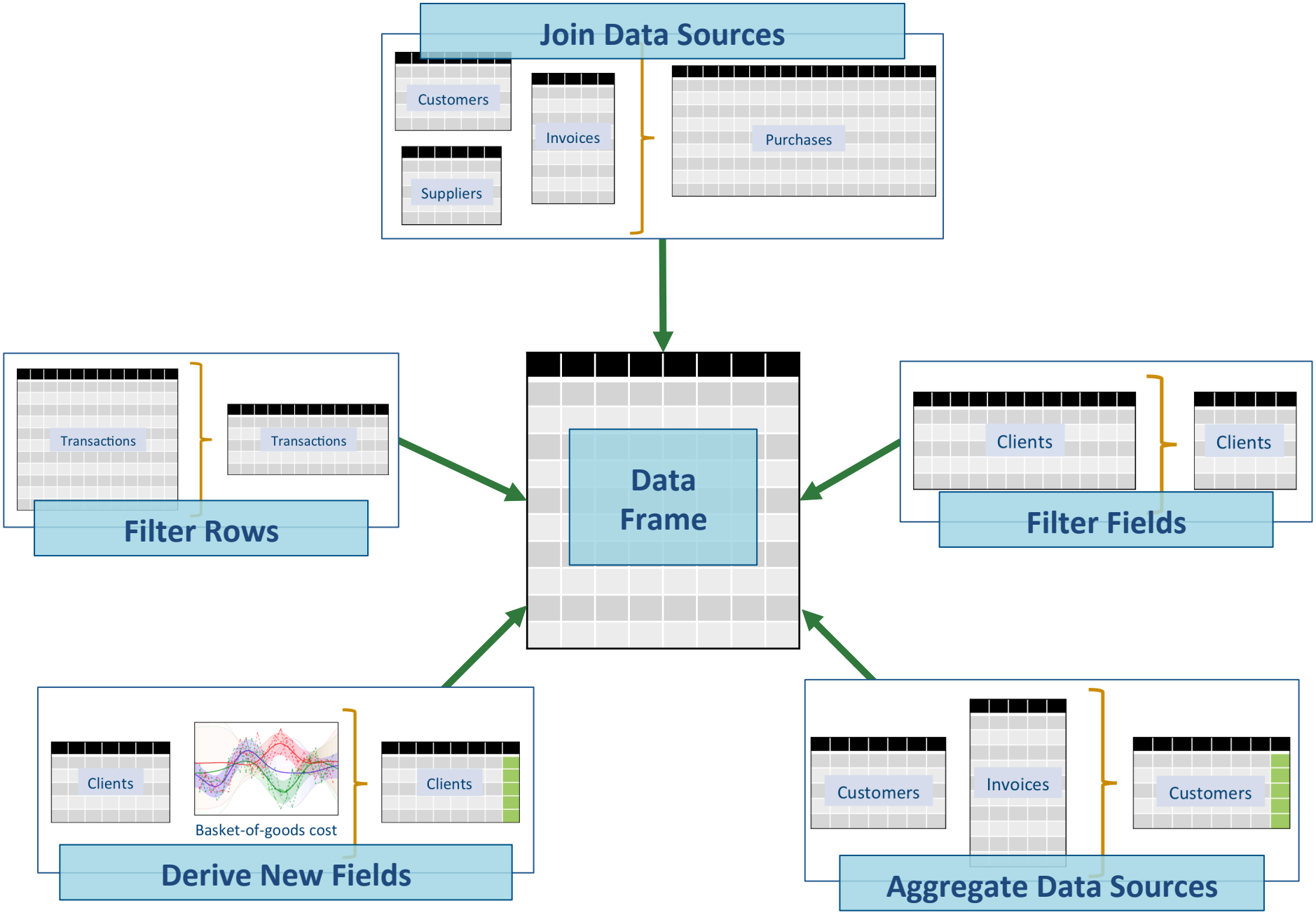
MANIPULATING DATA FRAMES


IPython
display
pandas
scipy
The SciPy Ecosystem
pickle
Python
Python Language Core
matplotlib
numpy
math
Python Standard Library
scikit-learn
plotly

The SciPy Ecosystem

SciPy (pronounced “Sigh Pie”) is a Python-based ecosystem of open-source software for mathematics, science, and engineering containing the following key packages:

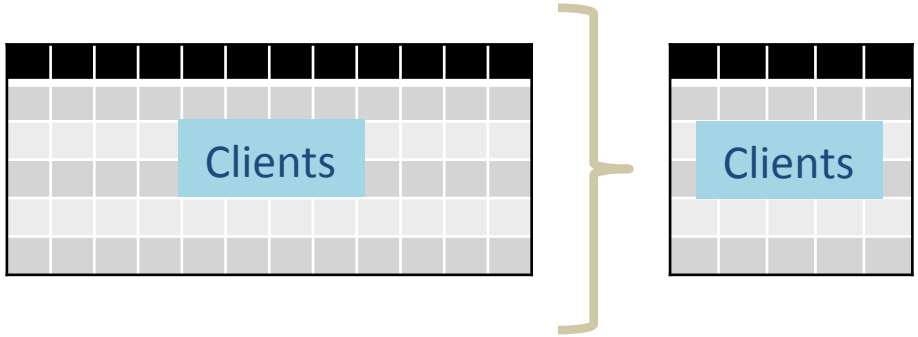
- **scipy** numerical algorithms such as signal processing, optimization and statistics
- **numpy** defines numerical array and matrix types and mathematical operations on them
- **pandas** high performance easy to use data structures and operations on them
- **matplotlib** publication-quality 2D plotting library


Join Data Sources
Customers
Invoices
Purchases
Suppliers
Transactions
Transactions
Filter Rows
Data Frame
Clients
Clients
Filter Fields
Clients
Basket-of-goods cost
Clients
Derive New Fields
Customers
Invoices
Customers
Aggregate Data Sources

Manipulating Data Frames

Filter fields

- Simply remove fields from a data source
- **Example:** Create a simple picture of a client from a more complex dataset



Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

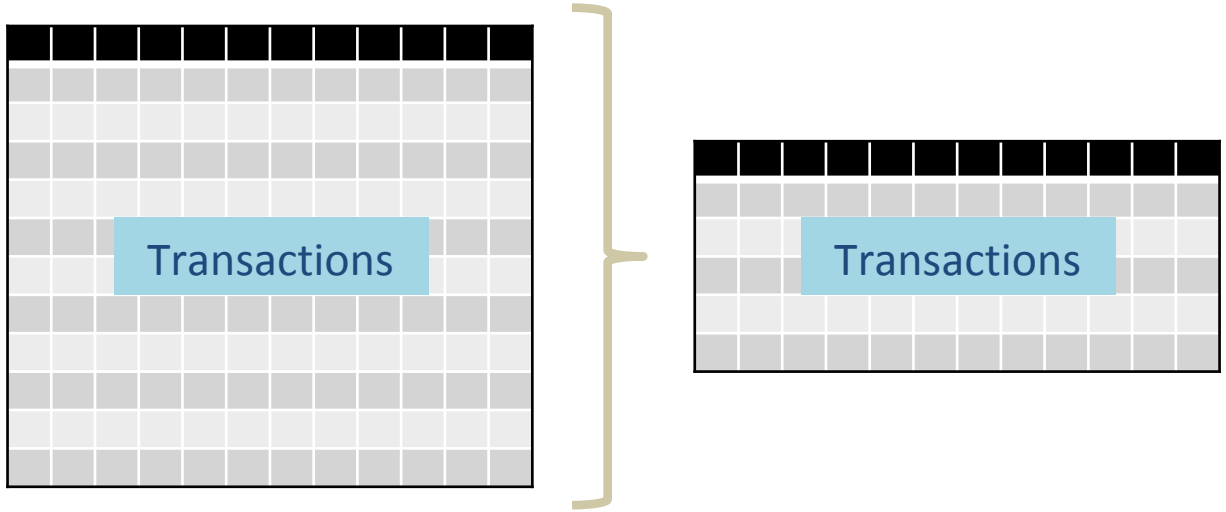
Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

Manipulating Data Frames

Filter rows

- Simply remove records from a data source - often based on selecting data from a particular time period
- **Example:** Build a transaction-based ABT using only the most recent transactions



Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

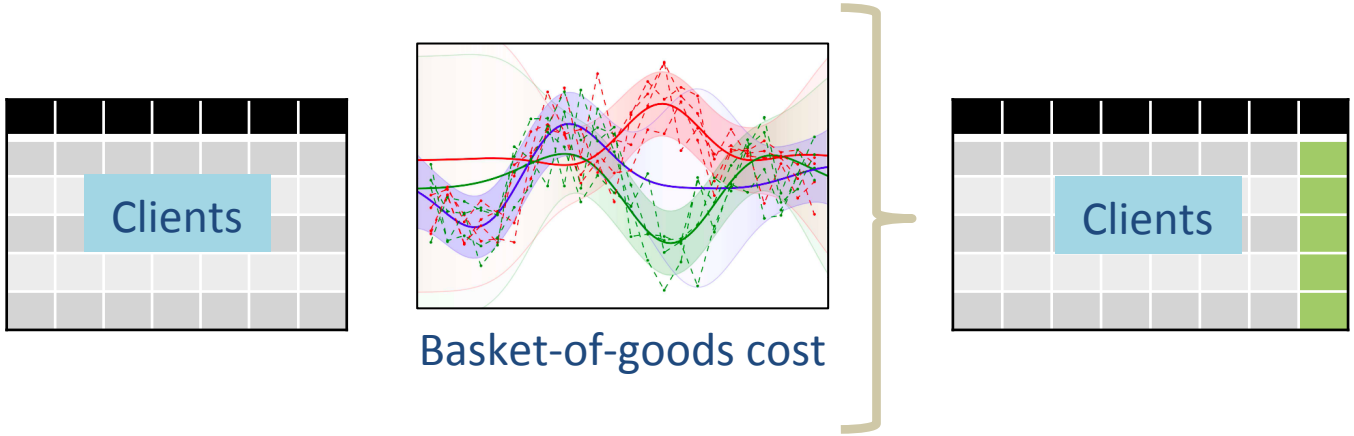
Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

Manipulating Data Frames

Derive new fields

- Derive new fields by combining fields from single or multiple data sources
- **Example:** Derive a new affordability score using a client's salary data and an external *basket-of-goods* cost based on inflation



Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI	High Education
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171	True
Haiti	45	47.67	73.1	0.09	3.4	1.7999	True
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493	True
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622	True
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961	True
China	74.87	29.98	13.7	1.95	6.4	3.6356	True
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741	True
Israel	81.3	28.8	3.6	6.77	12.5	5.8069	True
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357	True
Ireland	80.15	27.23	3.5	0.60	11.5	7.536	True
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751	True
Germany	80.24	22.07	3.5	1.31	12.0	8.0461	True
Canada	80.99	24.79	4.9	1.42	14.2	8.6725	True
Australia	82.09	25.4	4.2	1.86	11.5	8.8442	True
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985	True
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627	True

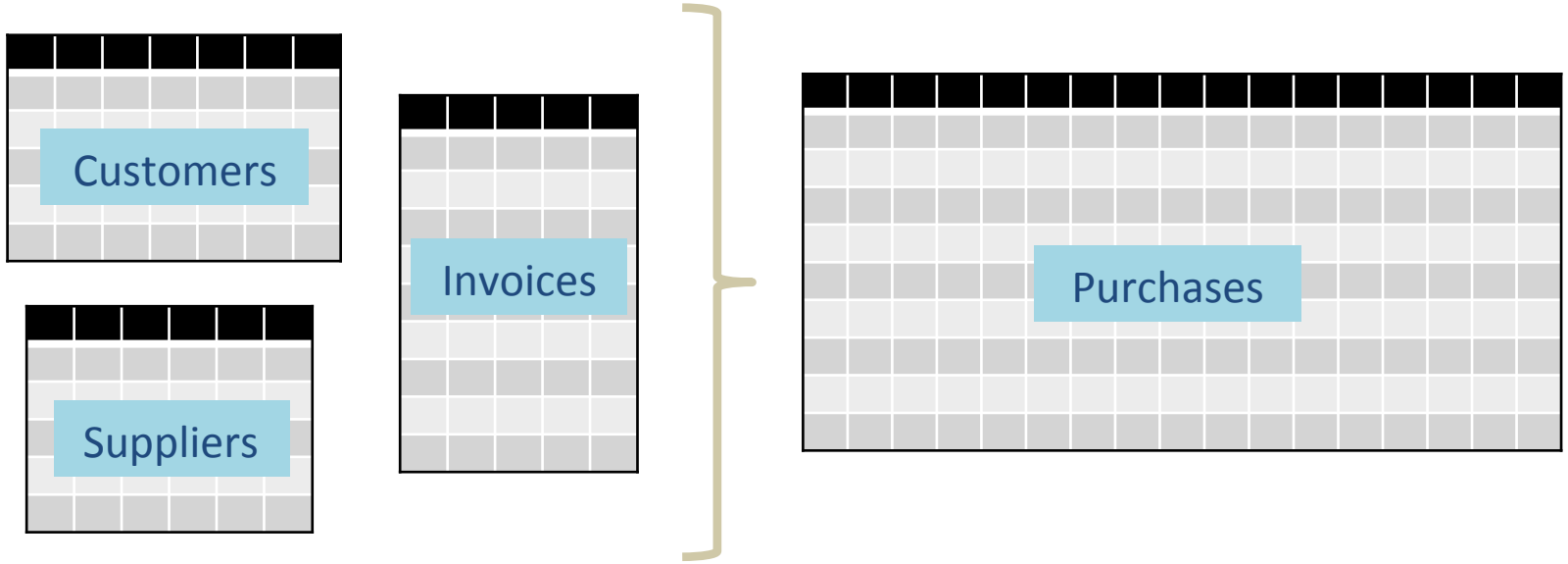
Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI	High Education
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171	False
Haiti	45	47.67	73.1	0.09	3.4	1.7999	False
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493	False
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622	False
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961	True
China	74.87	29.98	13.7	1.95	6.4	3.6356	False
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741	False
Israel	81.3	28.8	3.6	6.77	12.5	5.8069	True
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357	True
Ireland	80.15	27.23	3.5	0.60	11.5	7.536	True
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751	True
Germany	80.24	22.07	3.5	1.31	12.0	8.0461	True
Canada	80.99	24.79	4.9	1.42	14.2	8.6725	True
Australia	82.09	25.4	4.2	1.86	11.5	8.8442	True
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985	True
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627	True

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961

Manipulating Data Frames

Merge data sources

- Combine data sources based on a shared keys
- **Example:** Join an invoices, a customers and a suppliers table into a single table giving complete purchase details



Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.3	28.8	3.6	6.77	12.5	5.8069
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.536
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.4	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

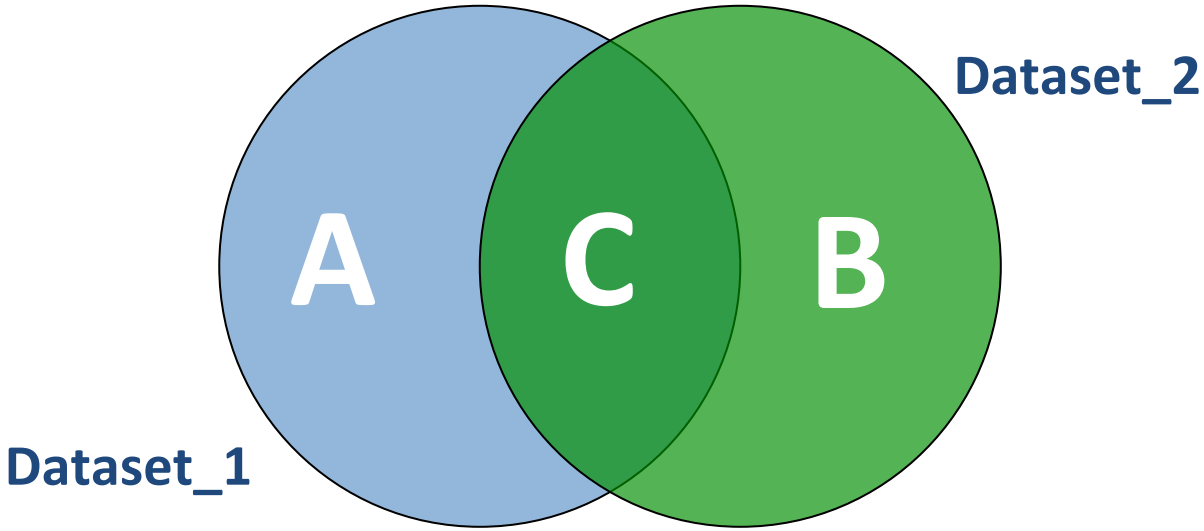
Country ID	Population
Afghanistan	27101365
Haiti	186988000
Nigeria	6198154
Egypt	90369500
Argentina	43590400
China	1357000000
Brazil	200400000
Israel	321068000
U.S.A	4635400
Ireland	65097000
U.K.	81459000
Germany	187820
Canada	35985751
Australia	23992700
Sweden	9845155
New Zealand	4659070

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI	Population
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171	27101365
Haiti	45	47.67	73.1	0.09	3.4	1.7999	186988000
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493	6198154
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622	90369500
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961	43590400
China	74.87	29.98	13.7	1.95	6.4	3.6356	1357000000
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741	200400000
Israel	81.3	28.8	3.6	6.77	12.5	5.8069	321068000
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357	4635400
Ireland	80.15	27.23	3.5	0.60	11.5	7.536	65097000
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751	81459000
Germany	80.24	22.07	3.5	1.31	12.0	8.0461	187820
Canada	80.99	24.79	4.9	1.42	14.2	8.6725	35985751
Australia	82.09	25.4	4.2	1.86	11.5	8.8442	23992700
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985	9845155
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627	4659070

Controlling Merging

The merge function allows us to control the way in which merge matching occurs

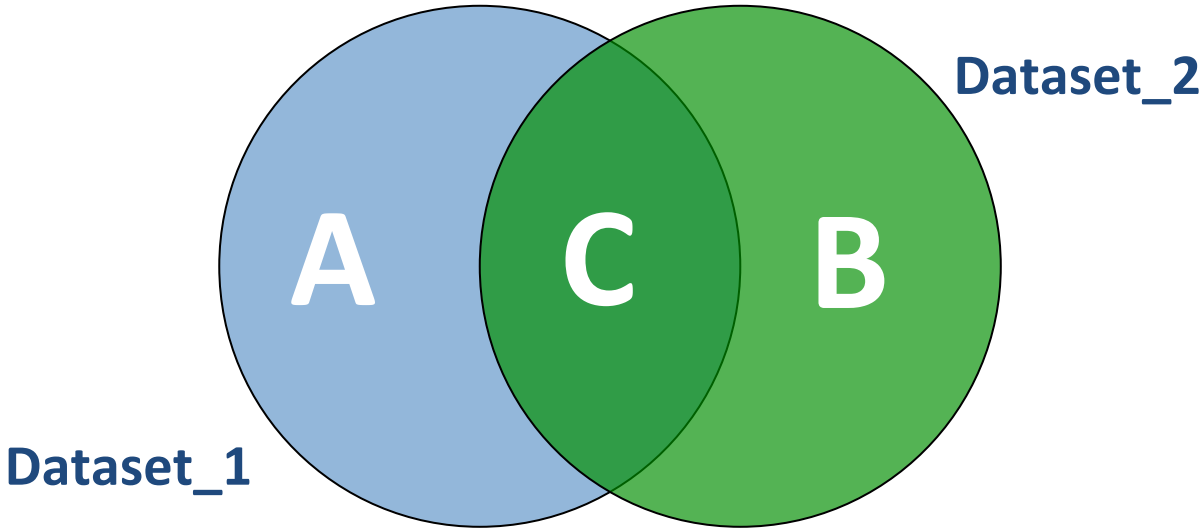
A Venn diagram of the keys in the two datasets to be merged helps understand this



Controlling Merging

There are three regions in this diagram

- **A:** Keys that are in Dataset_1 but not Dataset_2
- **B:** Keys that are in Dataset_2 but not Dataset_1
- **C:** Keys that are both Dataset_1 and Dataset_2

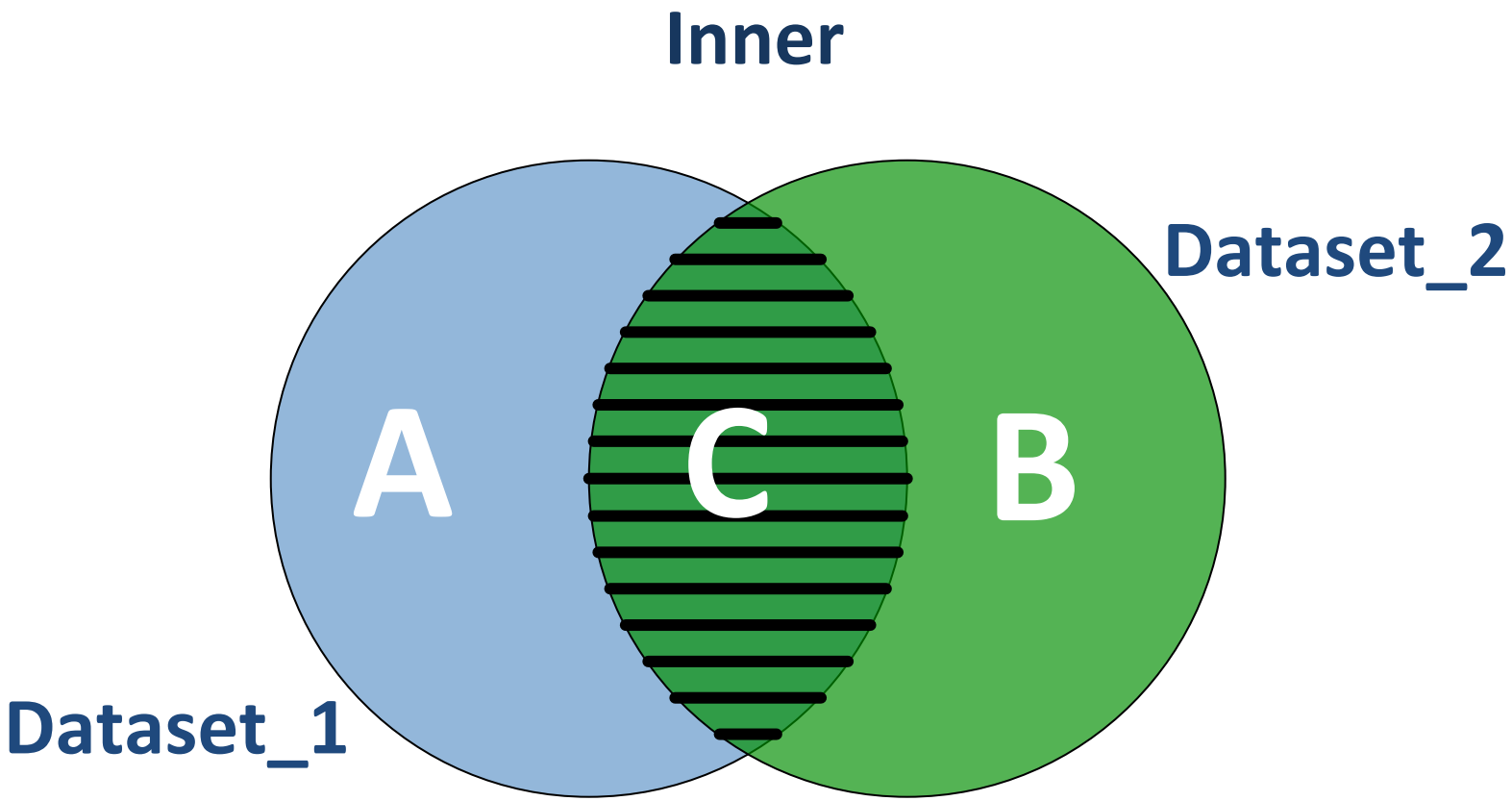


Controlling Merging

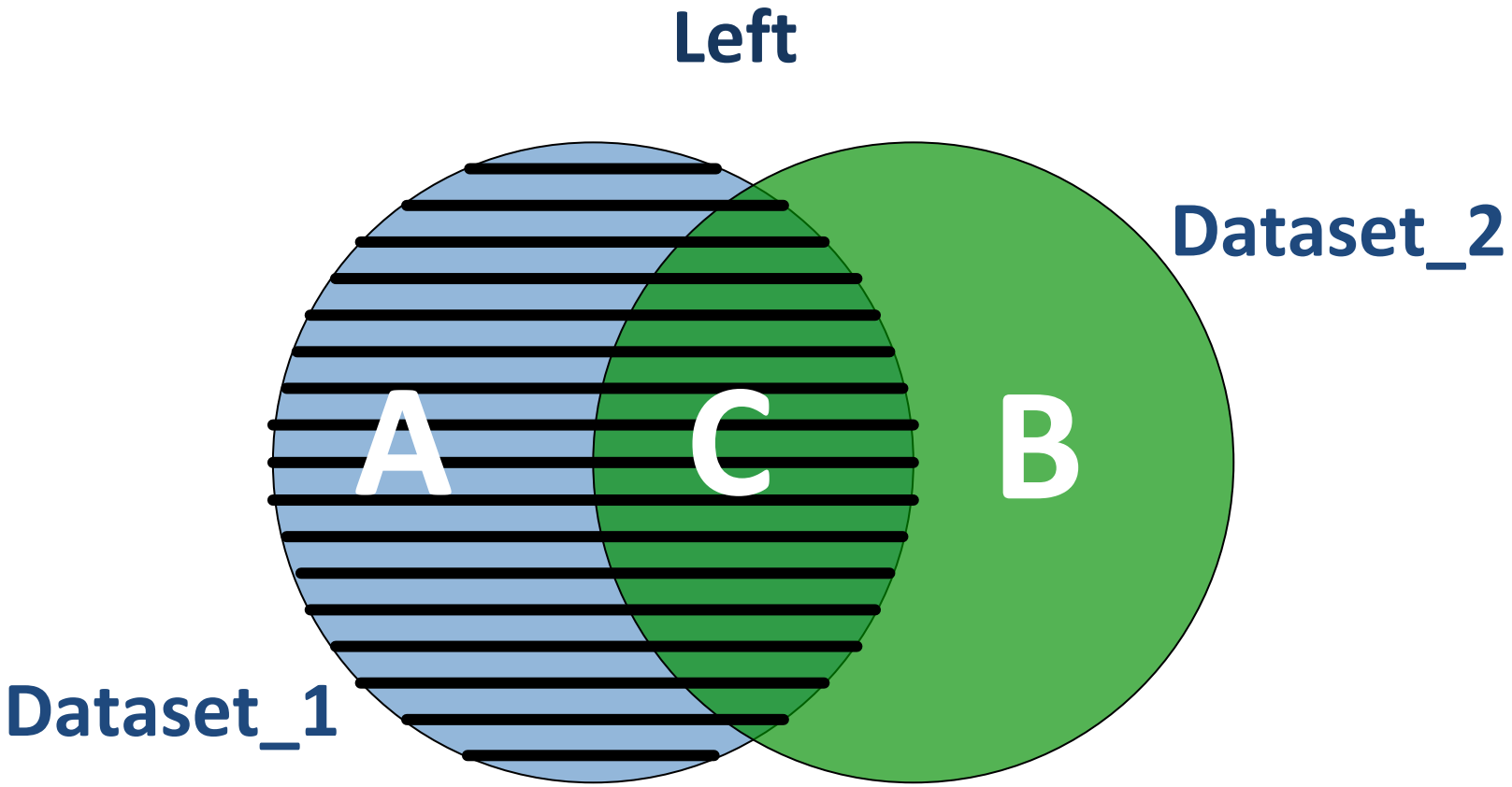
The are four types of merge we can do:

- left
- right
- outer
- inner

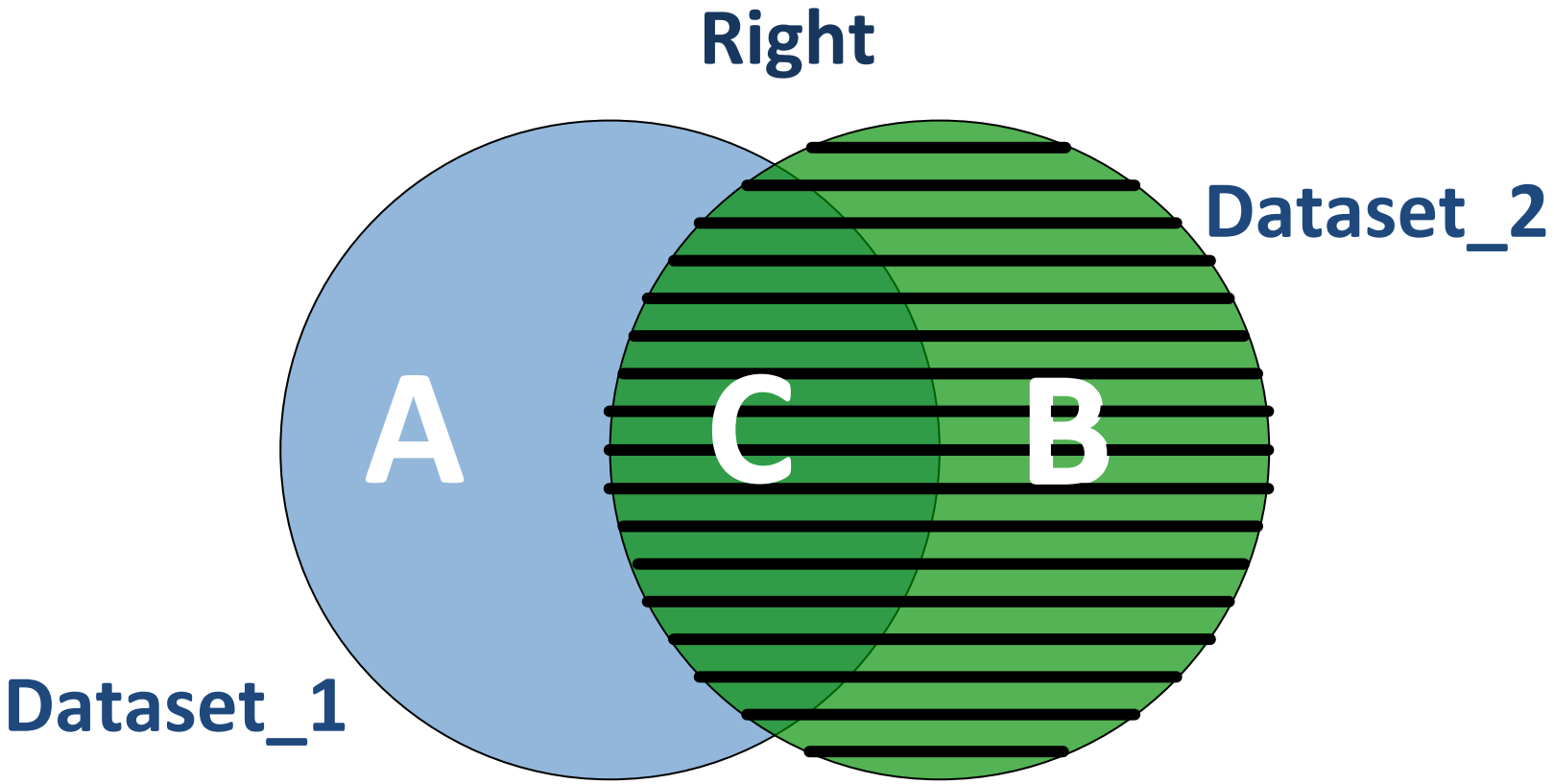
Controlling Merging



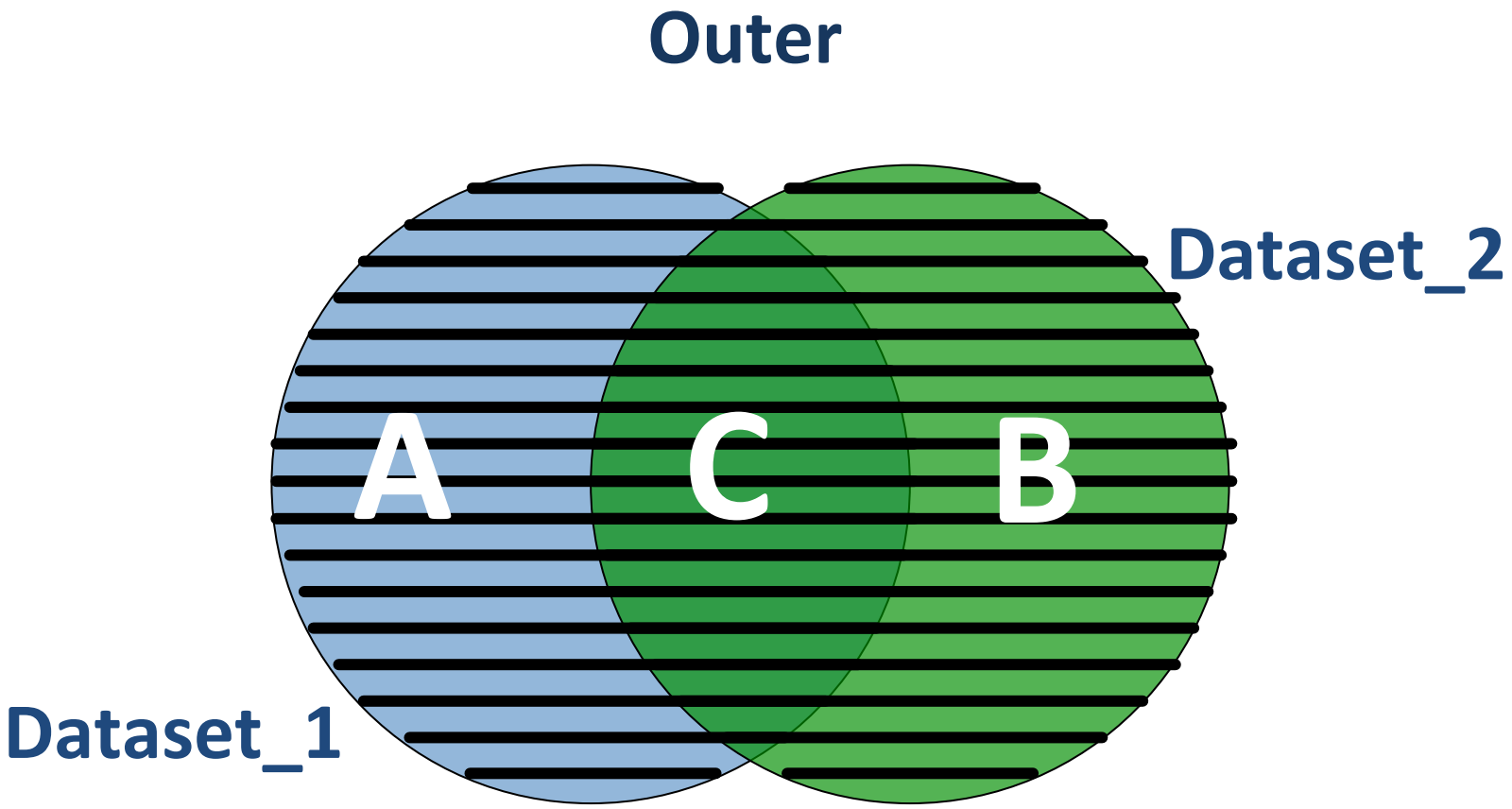
Controlling Merging



Controlling Merging



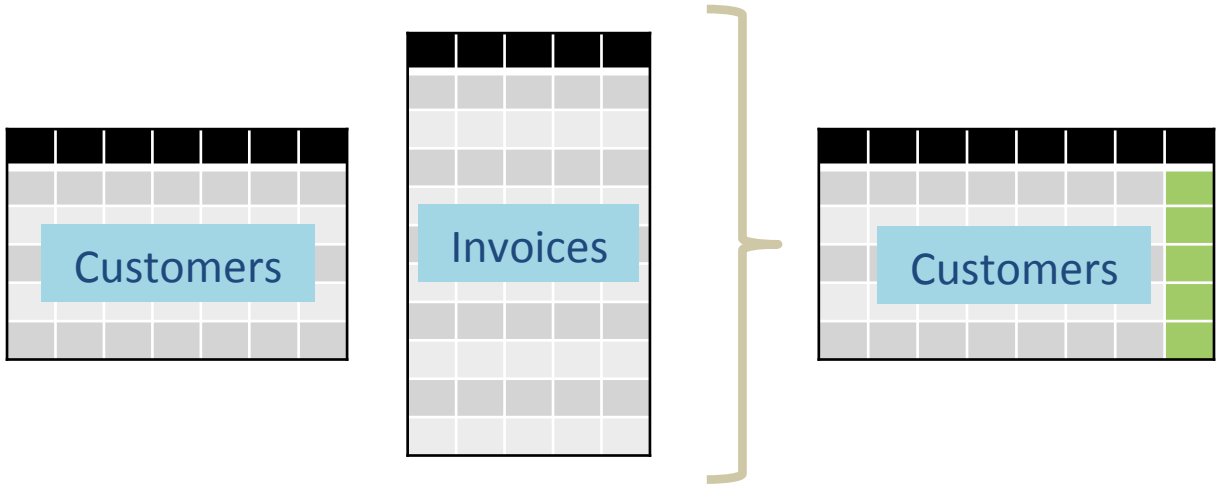
Controlling Merging



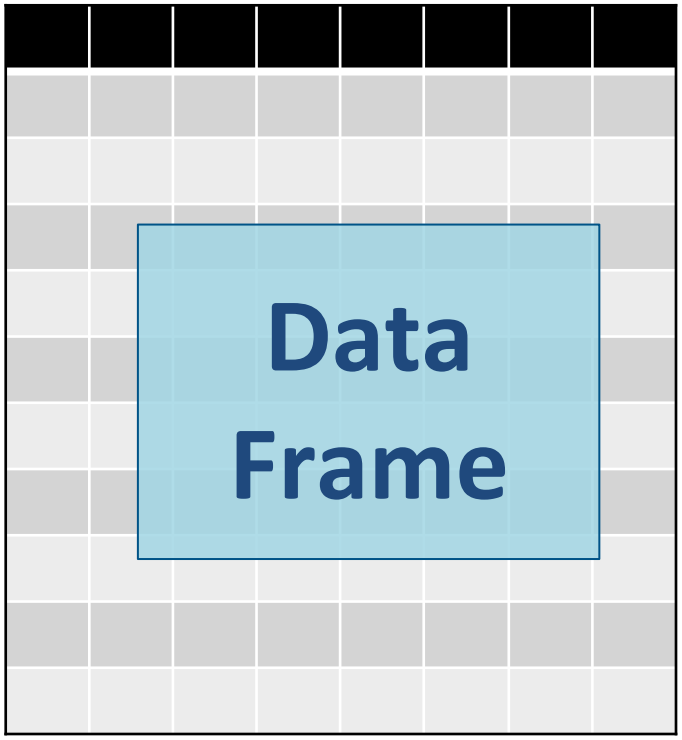
Manipulating Data Frames

Aggregate data sources

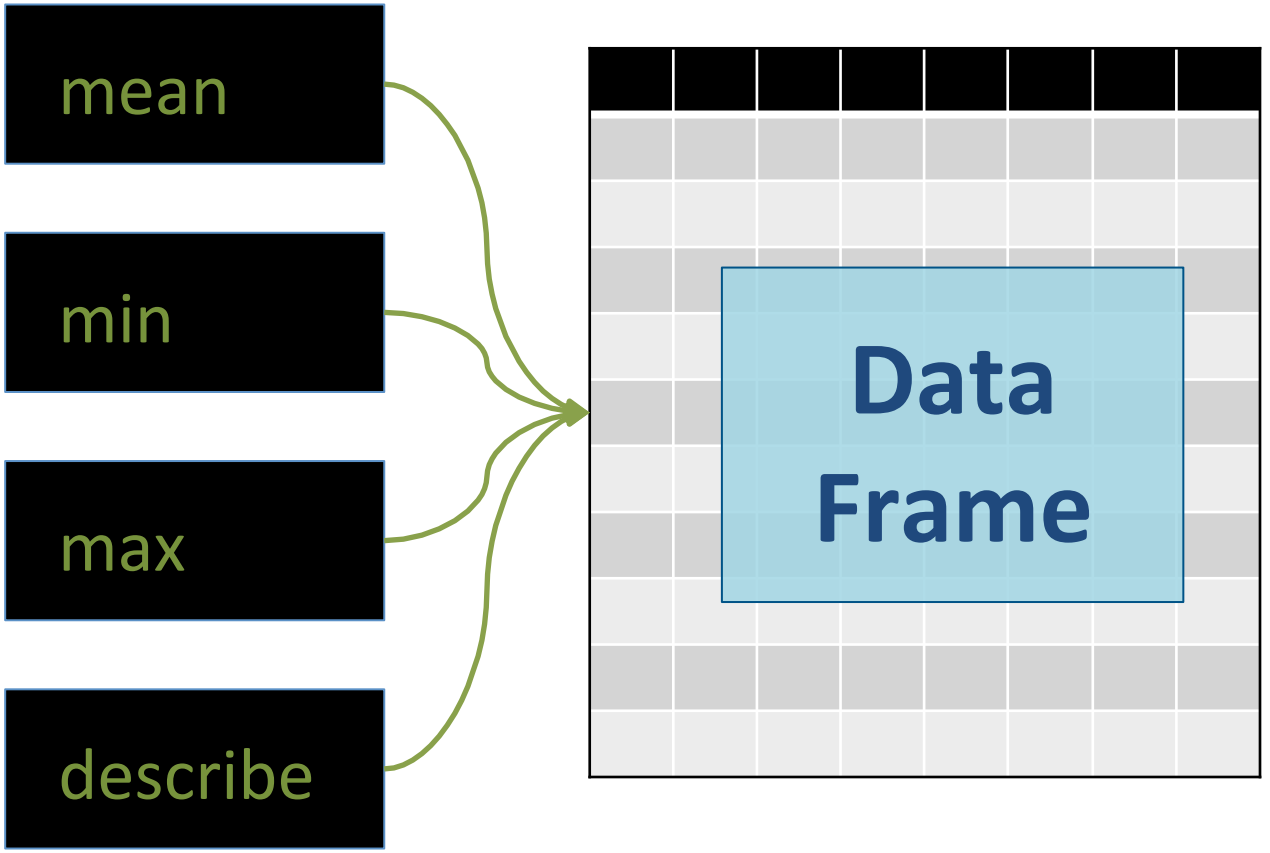
- Combine a set of records from one source into a single field in another source
- **Example:** Combine a customer invoice records into a single purchase count field to add to a customer record



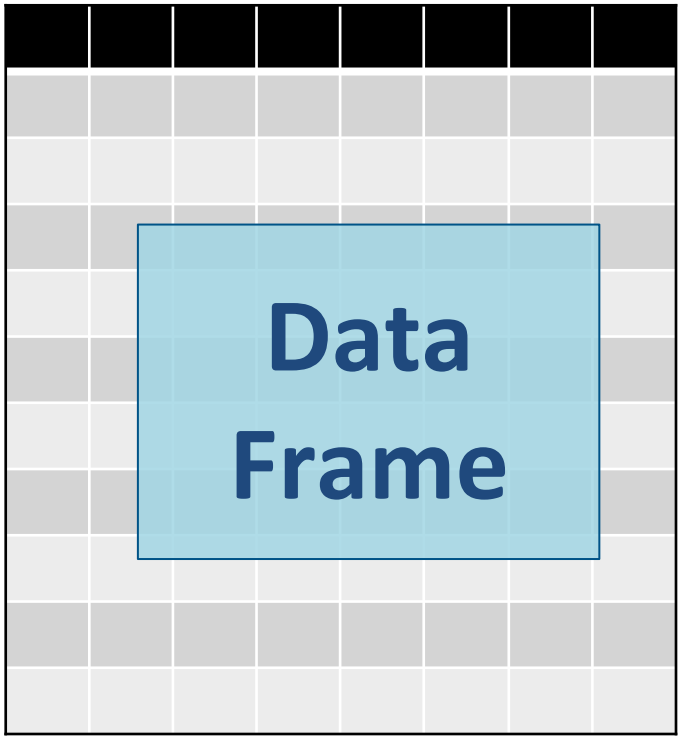
Using A Group By Object in pandas



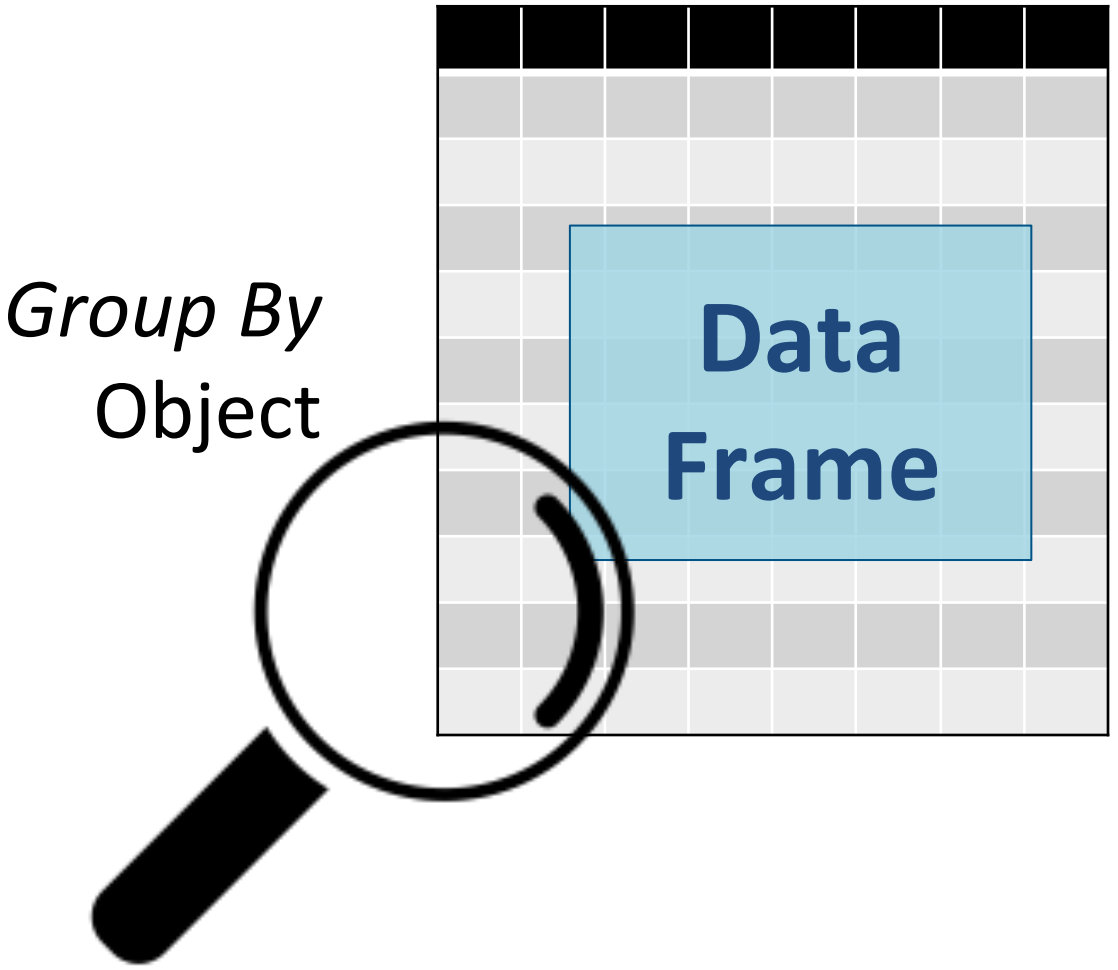
Using A Group By Object in pandas



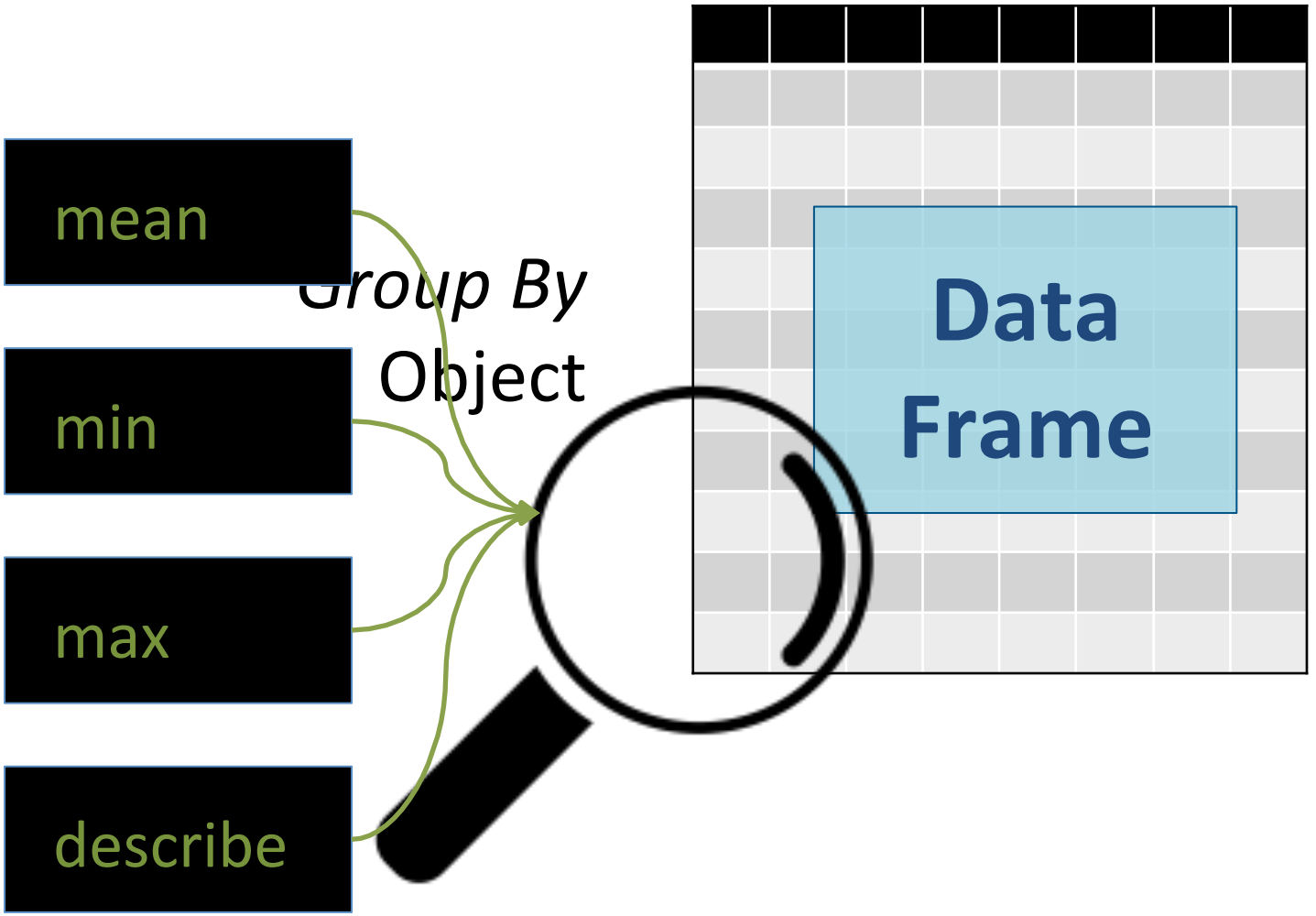
Using A Group By Object in pandas



Using A Group By Object in pandas



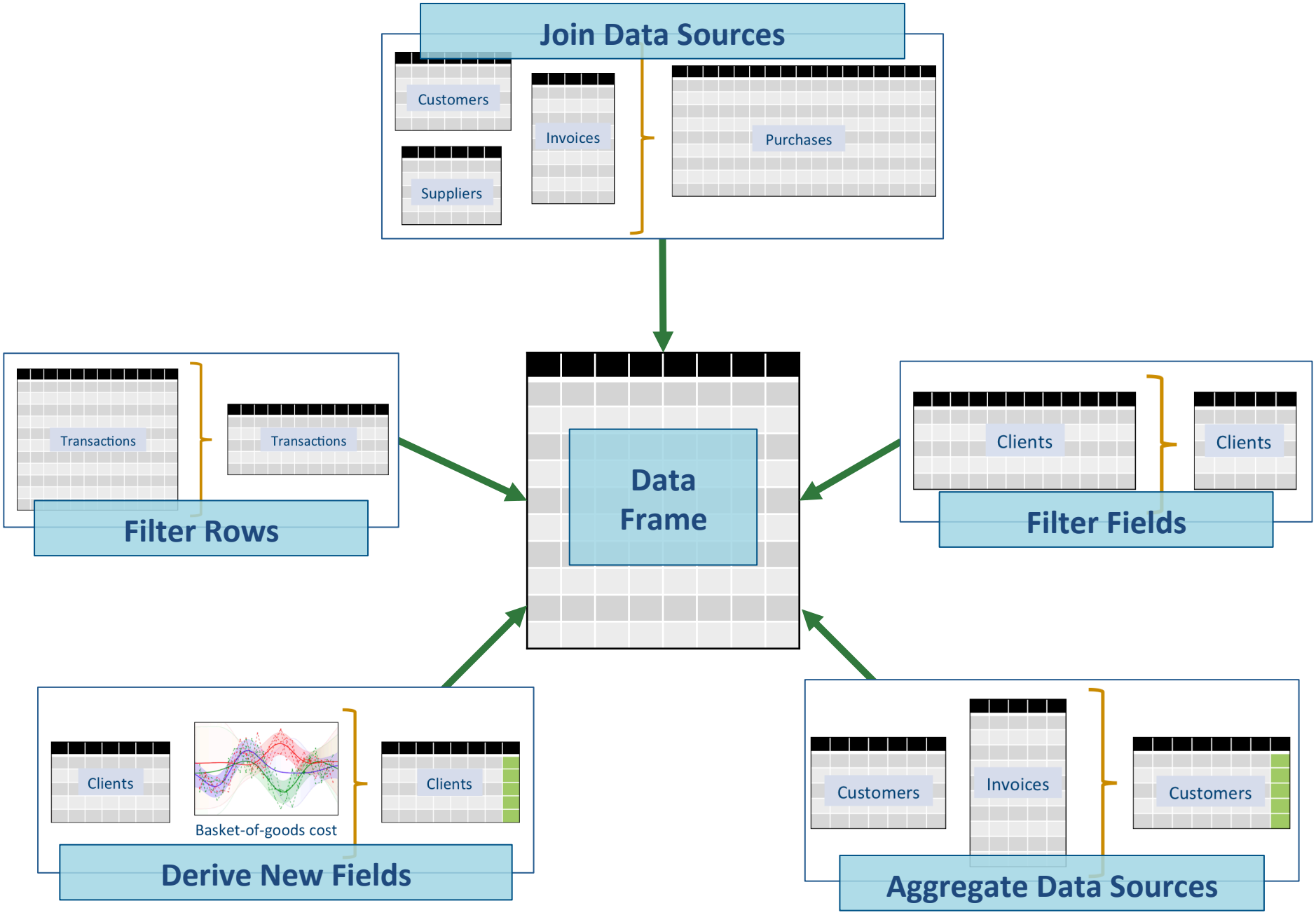
Using A Group By Object in pandas

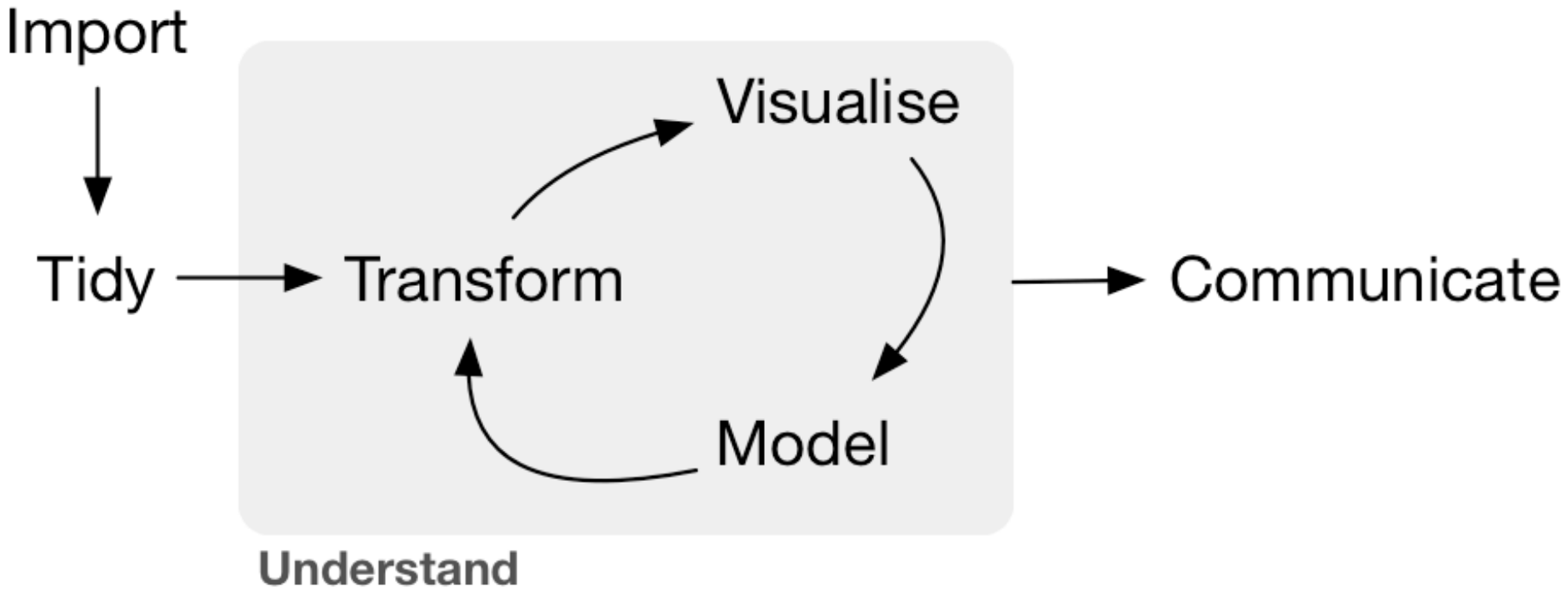


Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI	Continent
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171	Asia
Haiti	45	47.67	73.1	0.09	3.4	1.7999	North America
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493	Africa
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622	Africa
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961	South America
China	74.87	29.98	13.7	1.95	6.4	3.6356	Asia
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741	South America
Israel	81.3	28.8	3.6	6.77	12.5	5.8069	Asia
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357	North America
Ireland	80.15	27.23	3.5	0.60	11.5	7.536	Europe
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751	Europe
Germany	80.24	22.07	3.5	1.31	12.0	8.0461	Europe
Canada	80.99	24.79	4.9	1.42	14.2	8.6725	North America
Australia	82.09	25.4	4.2	1.86	11.5	8.8442	Oceania
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985	Europe
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627	Oceania

Country ID	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI	Continent
Nigeria	51.3	38.23	82.6	1.07	4.1	2.4493	Africa
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622	Africa
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171	Asia
China	74.87	29.98	13.7	1.95	6.4	3.6356	Asia
Israel	81.3	28.8	3.6	6.77	12.5	5.8069	Asia
Ireland	80.15	27.23	3.5	0.6	11.5	7.536	Europe
U.K.	80.09	28.49	4.4	2.59	13	7.7751	Europe
Germany	80.24	22.07	3.5	1.31	12	8.0461	Europe
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985	Europe
Haiti	45	47.67	73.1	0.09	3.4	1.7999	North America
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357	North America
Canada	80.99	24.79	4.9	1.42	14.2	8.6725	North America
Australia	82.09	25.4	4.2	1.86	11.5	8.8442	Oceania
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627	Oceania
Argentina	75.77	32.3	13.3	0.76	10.1	2.9961	South America
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741	South America

Continent	Life Exp.	Top-10 Income	Infant Mort.	Mil. Spend	School Years	CPI
Africa	60.890000	32.4050	51.100000	1.465000	4.700000	2.655750
Asia	71.926667	27.3300	30.533333	4.386667	6.433333	3.653200
Europe	80.477500	24.9925	3.450000	1.442500	12.325000	8.163925
North America	61.755000	38.7600	39.700000	2.405000	8.550000	4.467800
Oceania	81.380000	26.6050	4.550000	1.495000	11.900000	9.153450
South America	75.770000	32.3000	13.300000	0.760000	10.100000	2.996100


Join Data Sources
Customers
Invoices
Purchases
Suppliers
Transactions
Transactions
Filter Rows
Data Frame
Clients
Clients
Filter Fields
Clients
Basket-of-goods cost
Clients
Derive New Fields
Customers
Invoices
Customers
Aggregate Data Sources



https://ismayc.github.io/moderndiver-book/2-intro.html

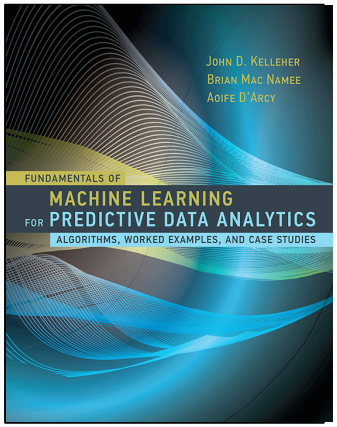
SUMMARY

Summary

Data science is concerned with extracting insight from data to help people make better decisions

The data frame is the key data structure that we use when analyzing datasets

As well as performing analysis we must be able to manipulate data



Fundamentals of Machine Learning for Predictive Data Analytics
John D. Kelleher, Brian Mac Namee, Aoife D'Arcy
MIT Press
www.machinelearningbook.com