Data Wrangling

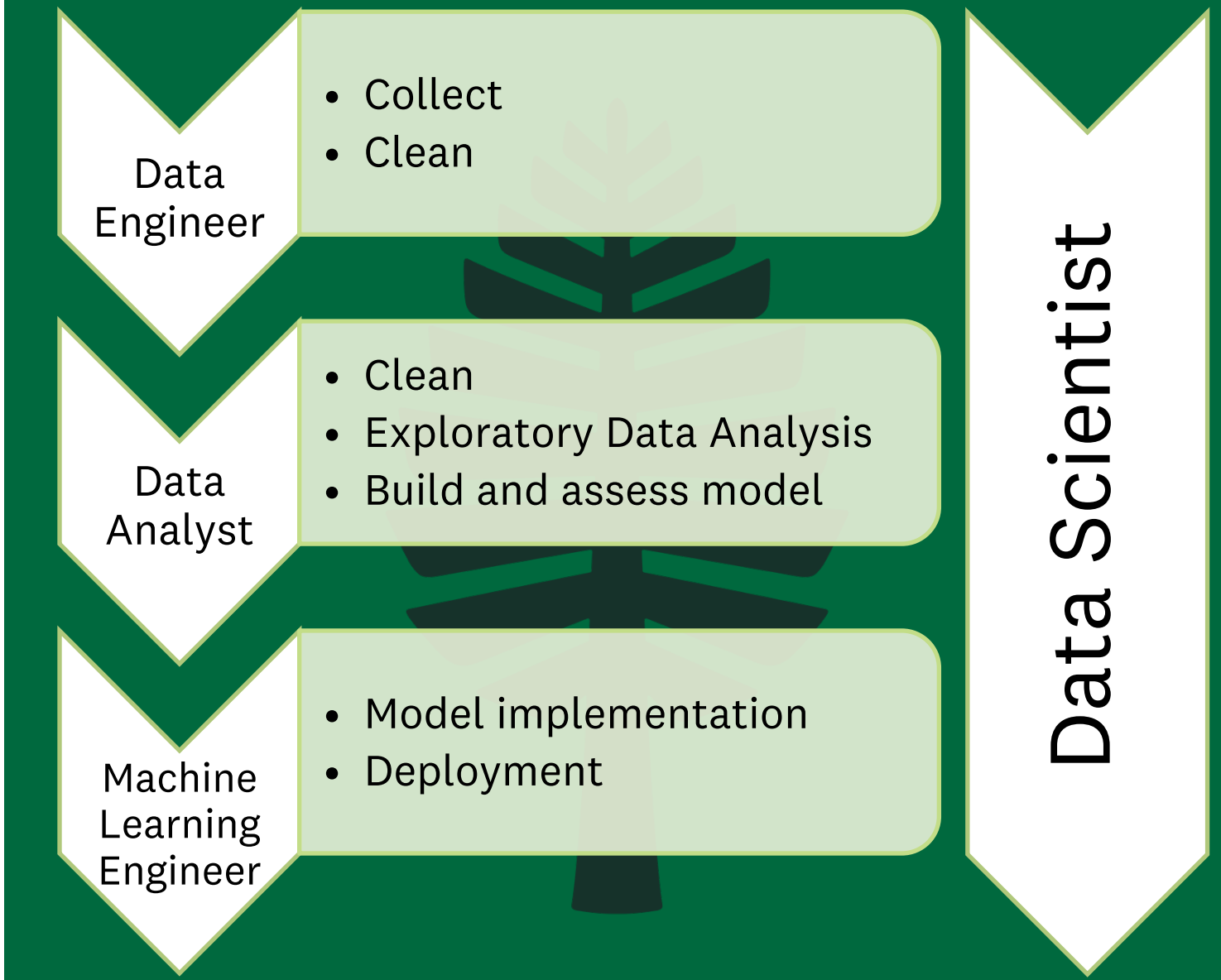
QBS 101.5: Applied Data Science

Simon Stone
Research Data Services
Dartmouth College Library

Introduction
Data Science

Data Science is OSEMN!*

- 📥 Obtain
- 🧽 Scrub
- 🔎 Explore
- 🤖 Model
- ⁉️ iNterpret



*pronounced “awesome” - /ˈɔ.səm/
https://www.datascience-pm.com/osemn

Why is my data dirty?!

- Data processing needs **consistently structured data**

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate	Birthrate	Deathrate	Agriculture	Industry	Service
23	Bhutan	ASIA (EX. NEAR EAST)	2279723	47000	48,5	0,00	0	100,44	1300.0	42,2	14,3	3,09	0,43	96,48	2	33,65	12,7	0,258	0,379	0,363
107	Kenya	SUB-SAHARAN AFRICA	34707817	582650	59,6	0,09	-0,1	61,47	1000.0	85,1	8,1	8,08	0,98	90,94	1,5	39,72	14,02	0,163	0,188	0,651
42	China	ASIA (EX. NEAR EAST)	1313973713	9596960	136,9	0,15	-0,4	24,18	5000.0	90,9	266,7	15,4	1,25	83,35	1,5	13,25	6,97	0,125	0,473	0,403
167	Reunion	SUB-SAHARAN AFRICA	787584	2517	312,9	8,22	0	7,78	5800.0	88,9	380,9	13,6	1,2	85,2	2	18,9	5,49	0,08	0,19	0,73
189	South Africa	SUB-SAHARAN AFRICA	44187637	1219912	36,2	0,23	-0,29	61,81	10700.0	86,4	107,0	12,08	0,79	87,13	1	18,2	22	0,025	0,303	0,671
6	Anguilla	LATIN AMER. & CARIB	13477	102	132,1	59,80	10,76	21,03	8600.0	95,0	460,0	0	0	100	2	14,17	5,34	0,04	0,18	0,78
132	Mauritania	SUB-SAHARAN AFRICA	3177388	1030700	3,1	0,07	0	70,89	1800.0	41,7	12,9	0,48	0,01	99,51	1	40,99	12,16	0,25	0,29	0,46
41	Chile	LATIN AMER. & CARIB	16134219	756950	21,3	0,85	0	8,8	9900.0	96,2	213,0	2,65	0,42	96,93	3	15,23	5,81	0,06	0,493	0,447
33	Burundi	SUB-SAHARAN AFRICA	8090068	27830	290,7	0,00	-0,06	69,29	600.0	51,6	3,4	35,05	14,02	50,93	2	42,22	13,46	0,463	0,203	0,334
116	Lesotho	SUB-SAHARAN AFRICA	2022331	30355	66,6	0,00	-0,74	84,23	3000.0	84,8	23,7	10,87	0,13	89	3	24,75	28,71	0,163	0,443	0,394

Excerpt from: https://www.kaggle.com/datasets/fernandol/countries-of-the-world (CC0)

Why is my data dirty?!

- Data processing needs **consistently structured data**
- **Published datasets** are (usually) **tidy**
- But data "**in the wild**" can be **messy**:
  - Different **sources** with different **formats**
  - **Missing** or incomplete data
  - **Inconveniently** formatted and/or structured

What you will learn in this session

- Importing and consolidating data from **inconsistent sources**
- **Cleaning** up inconsistent or inconveniently formatted entries
- Dealing with **missing values**

Data Wrangling,

Data Munging

What we will work with in this session

- Our **example data** was scraped from the product catalog on **www.lego.com**
- We will use **Python** and **Pandas** to clean it up
- Techniques are **universal**!

python™

pandas

We better get started...

Photo by Rick Mason on Unsplash