Week 1
What is Data Engineering

ISTA 322 - Data Engineering

What is Data Engineering?

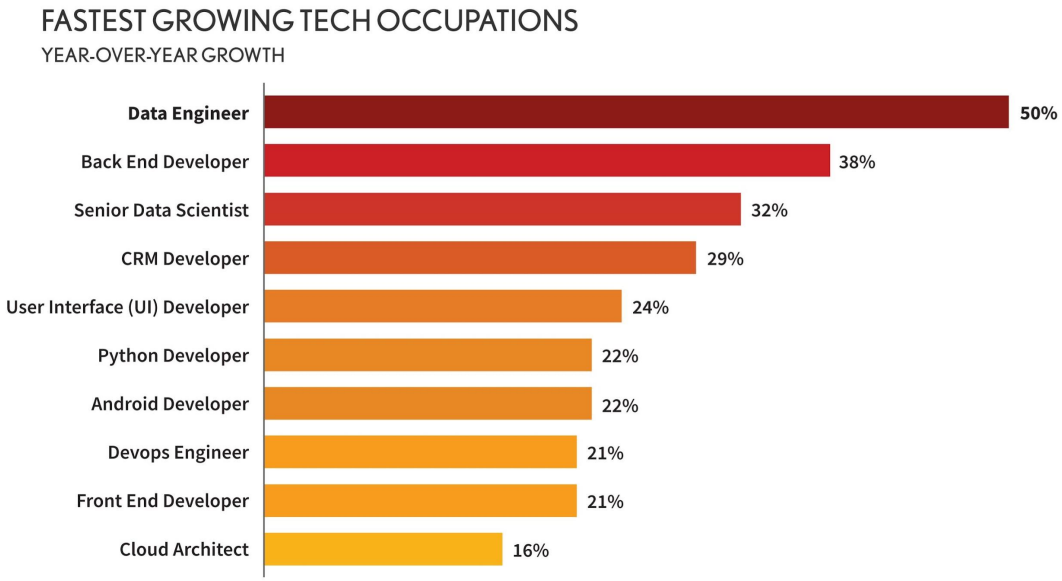
- “Data” engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state. ref
- data engineers are concerned with the production readiness of that data and all that comes with it: formats, scaling, resilience, security, and more. ref
- they build pipelines that transform that data into formats that data scientists can use. ref

What is Data Engineering?

- “Data” engineers design and build pipelines that transform and transport data into a format wherein, by the time other end users, it
- data engi uction readiness with it: formats, d more. ref
- they build pipelines that transform that data into formats that data scientists can use. ref

General definition - Building pipelines to make data useful for data scientists and analysts

DE jobs are booming



[Ref - Dice Tech Jobs Report]

#8 33% annual growth

Data Engineer

What you should know:
Data has quickly become every company's most valuable resource, and they need savvy engineers that can build infrastructure to keep it organized. The hiring growth rate of professionals in this emerging job has increased by nearly 35% since 2015, and industries from Retail to Automotive are snapping up this hard-to-hire talent. Interestingly, Amazon Web Services has emerged as one of the top skills held by Data Engineers, something that didn't show up in our analysis of professionals who held this role in 2015.

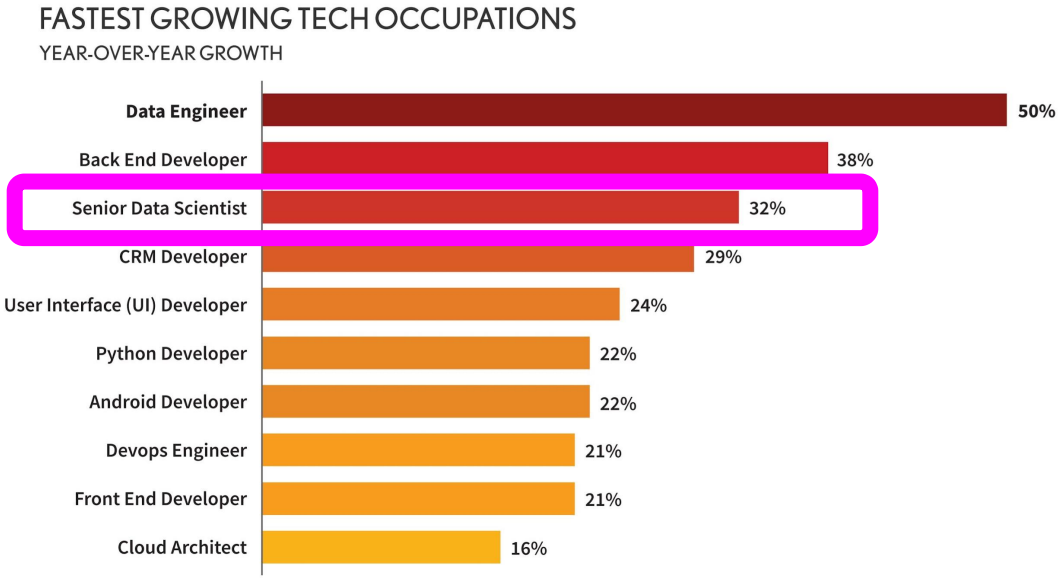
Skills unique to the job:
Apache Spark, Hadoop, Python, Extract/Transform/Load (ETL), Amazon Web Services

Where the jobs are:
San Francisco Bay Area, New York, Seattle, Boston, Chicago

Top industries hiring this talent:
Information Technology & Services, Internet, Computer Software, Financial Services, Hospital & Health Care

[Ref - LinkedIn Jobs Report]

But why though?



[Ref - Dice Tech Jobs Report]

#3 37% annual growth

Data Scientist

What you should know:
Data science is another field that has topped the Emerging Jobs list for three years running. It's a specialty that's continuing to grow significantly across all industries. Our data indicates some of this growth can likely be attributed to the evolution of previously existing jobs, like Statisticians, and increased emphasis on data in academic research.

Skills unique to the job:
Machine Learning, Data Science, Python, R, Apache Spark

Where the jobs are:
San Francisco Bay Area, New York, Washington, D.C., Seattle, Boston

Top industries hiring this talent:
Information Technology & Services, Computer Software, Internet, Financial Services, Higher Education

[Ref - LinkedIn Jobs Report]

But why though?

DATA

Data Scientist: The Sexiest Job of the 21st Century

Prof's Note - Lol

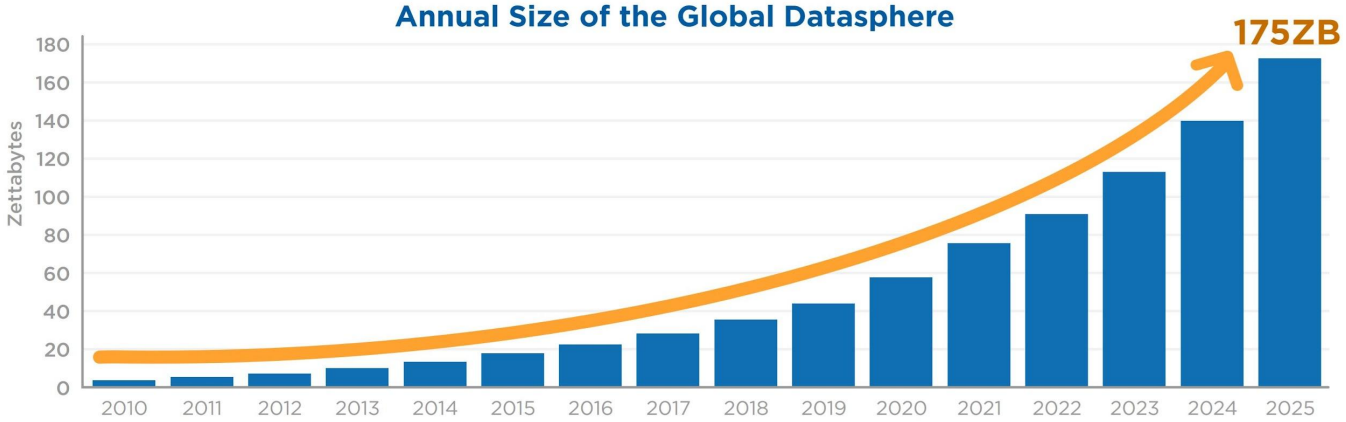
by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

Ref

Stepping back to big data

- “Sexy data scientists” came about in 2012
  - Although the job without title has been around for a while
- But the field of data science promised big things
- And companies had lots of data

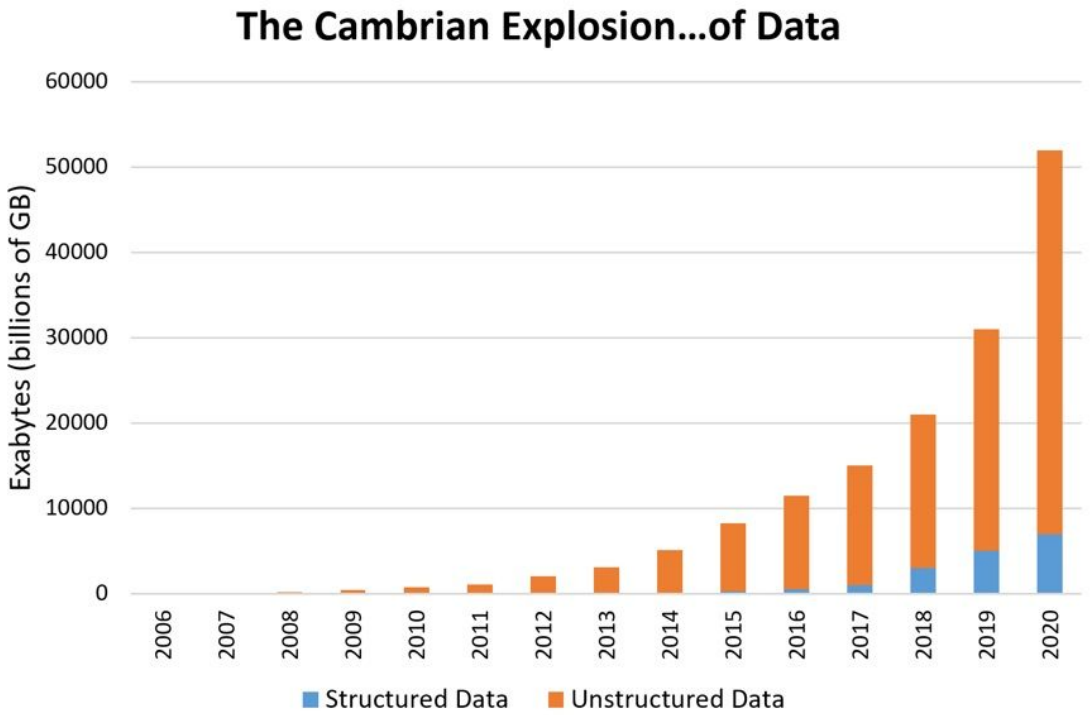


1 Zetabyte is
- 350 trillion songs
- 100k copies of wikipedia

[Ref - Seagate DataAge 2020 Report]

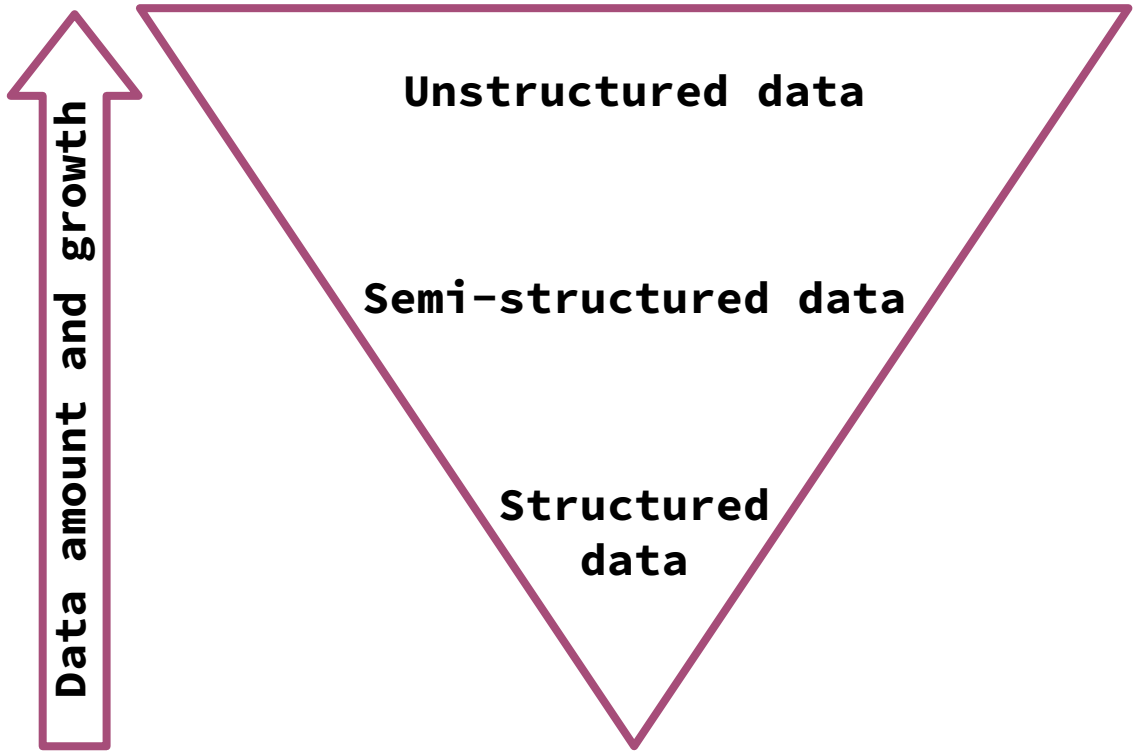
Stepping back to big data

- The boom is in **Unstructured data**
- What's the difference?



Pyramid of data organization

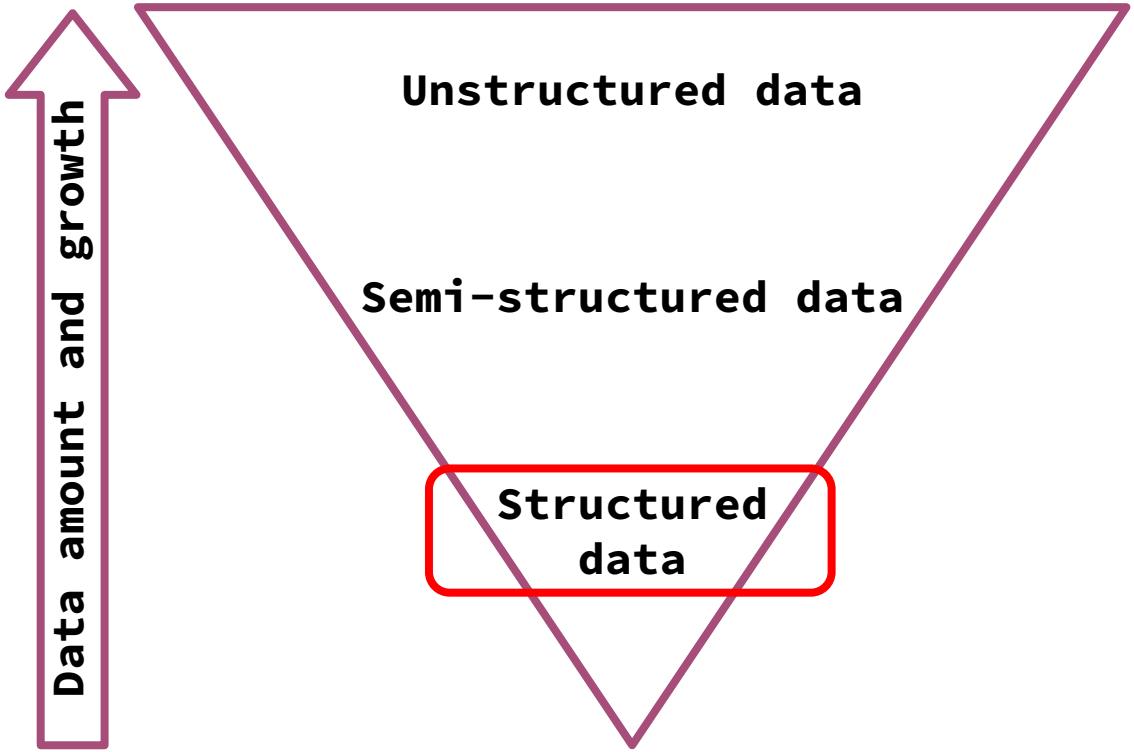
- Better to think of structure as being continuous
- Less and less data is **Structured**
- More and more is **Unstructured**
- Still room in between



Pyramid of data organization

Structured data

- Fits into a database nicely
- ‘square’ or ‘cube’ formats
- Don’t need to do work to analyze or query



Pyramid of data organization

Structured data

- SQL DB with two tables
- Simple query to get total sales for AZ stores
- No data processing needed to do this

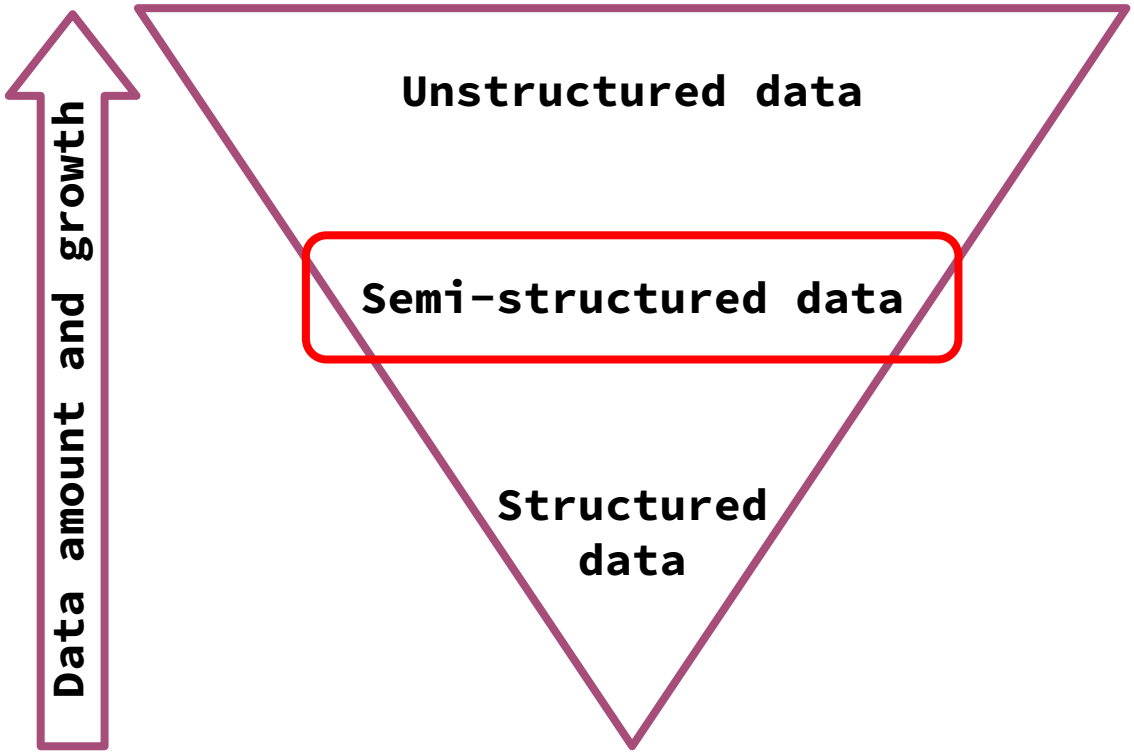
TABLE ID: STORE		
store_id	**store_state**	**country**
az_23	AZ	USA
az_45	AZ	USA
ca_12	CA	USA
to_39	Ontario	Canada

TABLE ID: TRANSACTIONS			
transact_id	**store_id**	**UPC**	**price**
x88943	az_23	49914	2.57
x88943	az_23	99371	1.99
a85921	to_39	95831	8.99
a85921	to_39	99492	5.49
a85921	to_39	27482	4.49
z88930	az_45	33491	0.99

Pyramid of data organization

Semi-structured data

- JSON, XML, csv, tsv
- Has some consistent format
- Minimal work to get into useable format



Pyramid of data organization

Semi-structured data

- Excel, csv, tsv
- May need to clean
- Need to join a bunch
- Aggregate
- May take time, but relatively simple

Sample AirBNB data from CSV file

A	B	C	D	E	F	G	H	I	J
id	name	host_id	host_name	neighbourhood_g	neighbourhood	latitude	longitude	room_type	price
2539	Clean & quiet apt	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149
2595	Skylit Midtown Ca	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225
3647	THE VILLAGE O	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150
3831	Cozy Entire Floor	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89
5022	Entire Apt: Spaci	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80
5099	Large Cozy 1 BR	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/apt	200
5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesa	40.68688	-73.95596	Private room	60
5178	Large Furnished	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79
5203	Cozy Clean Gues	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79
5238	Cute & Cozy Low	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150
5295	Beautiful 1br on U	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Entire home/apt	135
5441	Central Manhatta	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.98867	Private room	85
5803	Lovely Room 1, G	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779	Private room	89
6021	Wonderful Guest	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.96113	Private room	85
6090	West Village Nes	11975	Alina	Manhattan	West Village	40.7353	-74.00525	Entire home/apt	120
6848	Only 2 stops to M	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352	Entire home/apt	140
7097	Perfect for Your P	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.97185	Entire home/apt	215
7322	Chelsea Perfect	18946	Doti	Manhattan	Chelsea	40.74192	-73.99501	Private room	140
7726	Hip Historic Brow	20950	Adam And Charit	Brooklyn	Crown Heights	40.67592	-73.94694	Entire home/apt	99
7750	Huge 2 BR Upper	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190
7801	Sweet and Spaci	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.95718	Entire home/apt	299
8024	CBG CtyBGd He	22486	Lisel	Brooklyn	Park Slope	40.68069	-73.97706	Private room	130
8025	CBG Helps Haiti	22486	Lisel	Brooklyn	Park Slope	40.67989	-73.97798	Private room	80
8110	CBG Helps Haiti	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Private room	110

Pyramid of data organization

Semi-structured data

- JSON, xml
- Has an overall schema/organization
- Need to parse and organize to make useable
- May take time, but relatively simple

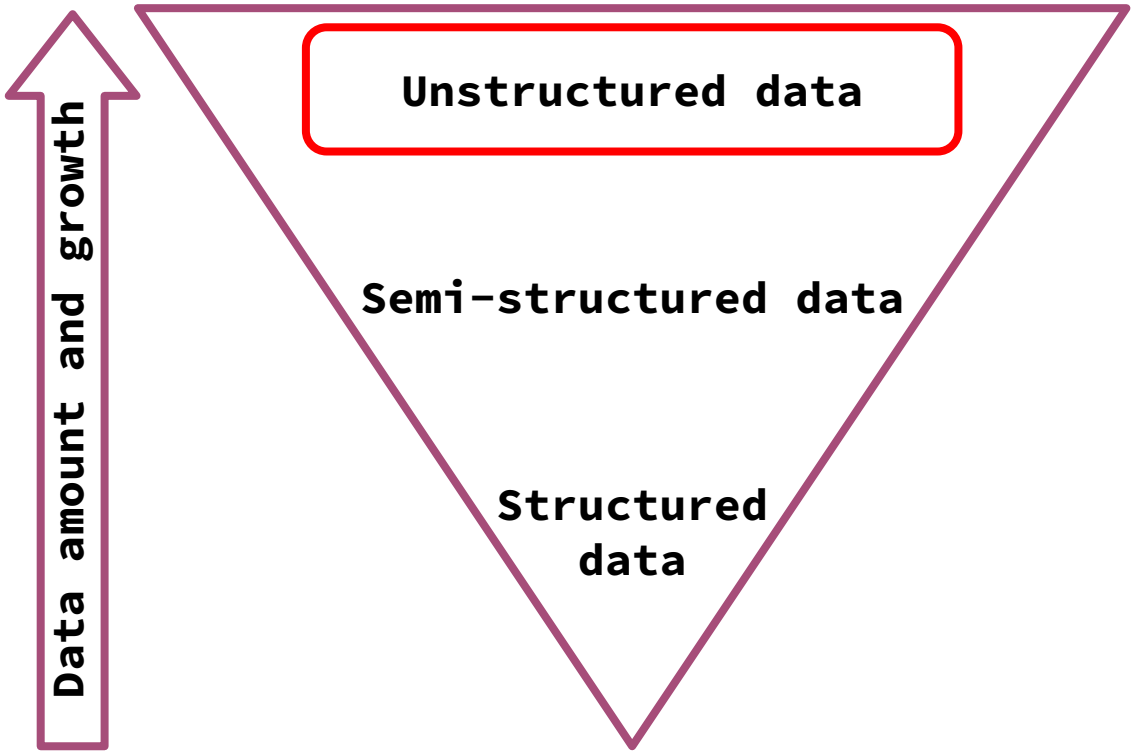
Sample Twitter data in JSON format


```
{
    "contributors": null,
    "coordinates": null,
    "created_at": "Fri Jun 28 07:31:35 +0000 2019",
    "display_text_range": [
        0,
        1
    ],
    "entities": {
        "hashtags": [],
        "symbols": [],
        "urls": [
            {
                "display_url": "twitter.com/polhomeeditor/\u2026",
                "expanded_url": "https://twitter.com/polhomeeditor/status/1144289510739587073",
                "indices": [
                    2,
                    25
                ],
                "url": "https://t.co/0fgkUFjCaB"
            }
        ],
        "user_mentions": []
    },
    "favorite_count": 3,
    "favorited": false,
    "full_text": "? https://t.co/0fgkUFjCaB",
    "geo": null,
    "id": 1144508626738044929,
    "id_str": "1144508626738044929",
```

Pyramid of data organization

Unstructured data

- Text, pdf, images, video
- Zero structure
- Lots of work to extract useful data from



Pyramid of data organization

Unstructured data

- Raw text - tweets, facebook, reviews
- How do you get something useful?
- Lots of processing
- Need to understand language
- Specific to use case

Sample of Twitter text

```
## [1] "If you are feeling the impacts of climate change in your daily
life, you're not alone. What do you want to protect from #climatechange?
Tell us below. #AdaptOurWorld https://t.co/RFyNFRLHGB"
## [2] "Tree 🌲 planting is great but it cannot become a PR machine for
fake climate initiatives by the governments. We will not push
#climatechange back by planting trees. Governments must listen and act.
Real-meaningful actions. Not PR!"
## [3] "Cigarette smokers &amp;amp; vapers are exhaling additional CO2 into
the air. I call on all politicians to stop smoking and ban cigarettes and
vaping for the earth. If not, then shut up about #climatechange and
CO2.\n\nLet’s see how much politicians care about the things they talk
about. https://t.co/0v60pNd38q"
## [4] "@ThemeParkReview @Starbucks Make your own coffee. You can even
buy the Starbucks brand in a grocery store, if it’s that important. This
really is a ridiculous tantrum over something with many solutions. There
are better things to worry about like the #ClimateChange that keeps
causing these hurricanes."
## [5] "@GaryCMeleJr @DebraMessing The two party system isn’t working
for the people &amp;amp; #Democrats need to do better because my independent
vote is on loan to them. But it’s the #GOP breaking constitutional norms,
refusing to protect our elections, shoving church into state, denying
#ClimateChange, caging kids etc"
## [6] "Why are #hurricanes getting bigger and moving slower?
🌪️\n\n#HurricaneDorian \n#ClimateChange #ClimateChangeIsReal
\n#ThereIsNoPlanetB 🌎 https://t.co/2z1JlpVnUg"
```

Pyramid of data organization

Unstructured data

- pdf files
- Very long format
- Might want to synthesize 1000's of medical papers to determine effect
- Formats vary across journals

Sample of journal text

PLOS ONE

RESEARCH ARTICLE

Who needs 'lazy' workers? Inactive workers act as a 'reserve' labor force replacing active workers, but inactive workers are not replaced when they are removed

Daniel Charbonneau[1]*, Takao Sasaki[2], Anna Dornhaus[3]

1 Graduate Interdisciplinary Program in Entomology & Insect Science, University of Arizona, Biological Sciences West, 1041 East Lowell, Tucson, AZ, United States of America, **2** Department of Zoology, University of Oxford, Oxford, United Kingdom, **3** Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, United States of America

* charbonneau.daniel@gmail.com

Check for updates

OPEN ACCESS

Citation: Charbonneau D, Sasaki T, Dornhaus A (2017) Who needs 'lazy' workers? Inactive workers act as a 'reserve' labor force replacing active workers, but inactive workers are not replaced when they are removed. PLoS ONE 12(9): e0184074. https://doi.org/10.1371/journal.pone.0184074

Editor: James A. R. Marshall, University of Sheffield, UNITED KINGDOM

Received: September 1, 2016

Accepted: August 17, 2017

Published: September 6, 2017

Data Availability Statement: All data files are available from the Dryad database (Provisional DOI: doi:10.5061/dryad.77110).

Funding: Research supported through the GIDP-EIS and EEB Department at University of Arizona, as well as NSF grants no. IOS-1045239, IOS-0841756, and DBI-1262292 (to A.D.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Social insect colonies are highly successful, self-organized complex systems. Surprisingly however, most social insect colonies contain large numbers of highly inactive workers. Although this may seem inefficient, it may be that inactive workers actually contribute to colony function. Indeed, the most commonly proposed explanation for inactive workers is that they form a 'reserve' labor force that becomes active when needed, thus helping mitigate the effects of colony workload fluctuations or worker loss. Thus, it may be that inactive workers facilitate colony flexibility and resilience. However, this idea has not been empirically confirmed. Here we test whether colonies of *Temnothorax rugatulus* ants replace highly active (spending large proportions of time on specific tasks) or highly inactive (spending large proportions of time completely immobile) workers when they are experimentally removed. We show that colonies maintained pre-removal activity levels even after active workers were removed, and that previously inactive workers became active subsequent to the removal of active workers. Conversely, when inactive workers were removed, inactivity levels decreased and remained lower post-removal. Thus, colonies seem to have mechanisms for maintaining a certain number of active workers, but not a set number of inactive workers. The rapid replacement (within 1 week) of active workers suggests that the tasks they perform, mainly foraging and brood care, are necessary for colony function on short timescales. Conversely, the lack of replacement of inactive workers even 2 weeks after their removal suggests that any potential functions they have, including being a 'reserve', are less important, or auxiliary, and do not need immediate recovery. Thus, inactive workers act as a reserve labor force and may still play a role as food stores for the colony, but a role in facilitating colony-wide communication is unlikely. Our results are consistent with the often cited, but never yet empirically supported hypothesis that inactive workers act as a pool of 'reserve' labor that may allow colonies to quickly take advantage of novel resources and to mitigate worker loss.

Pyramid of data organization

Unstructured data

- Images
- Obviously we can determine elements in a picture
- Lots of work to get a computer to figure that out
- What items are in it
  - Dan, people, motorcycle, helmet, sunglasses, road, mountain, desert, etc.

Image of me and on a motorcycle ride that I posted on Instagram

Why is this talk about data structure relevant?

- Let's think about what a data analyst/scientist does
- **Data Analyst:** Processes and analyzes data with the goal of supporting decision making.
- **Data Science:** Extracting useful information from data using computational methods.
- Rough definitions, but both are using data to create understanding. Analysts do more whole data aggregation, summaries, basic stats. Data Scientists often delve deeper into stats and machine learning.
- Either way, let's consider the structure of a common model used by both.

Extracting info from data

- Let’s say you want to predict if someone shopping on amazon will buy a iPhone
- This is a common classification problem.
  - Target - Will they purchase
  - Features - Age, income level, browser, searched for iPhone, etc
- This is an easy model to fit in R/Python and created a prediction with.

```
P(buy_iPhone) ~ age + income_level + browser + search_iPhone
```

- But, what format do the data need to be in to run this model?

Extracting info from data

`P(buy_iPhone) ~ age + income_level + browser + search_iPhone`

buy	age	income	browser	search_iPhone		
yes	32	123000	safari	yes		
no	56	56000	chrome	no		
no	47	75000	firefox	no		
yes	21	36000	safari	yes		

- Not hard to get data in this format through SQL queries
- But what if these are scattered across messy databases
- Or say information is in JSON files

Extracting info from data

P(buy_iPhone) ~ age + income_level + browser + search_iPhone + apple_reviews + apple_sent

buy	age	income	browser	search_iPhone	apple_revs	apple_sentiment
yes	32	123000	safari	yes	3	4
no	56	56000	chrome	no	0	0
no	47	75000	firefox	no	1	2
yes	21	36000	safari	yes	5	5

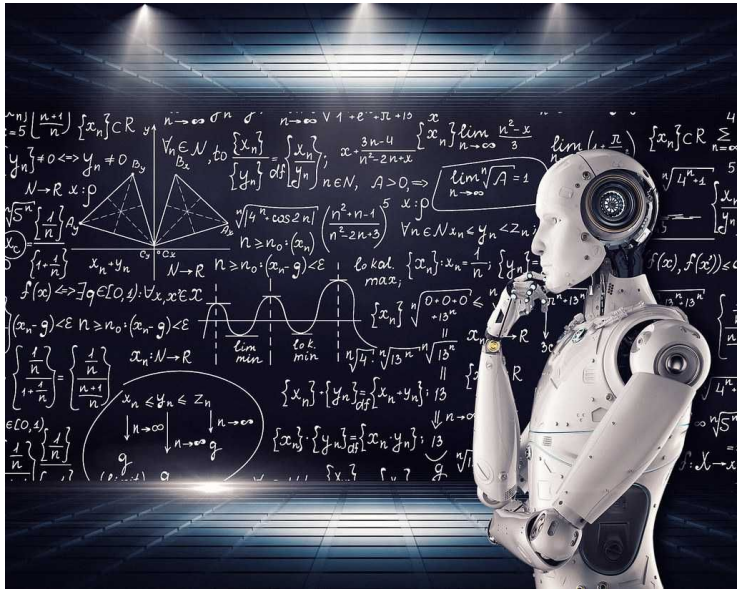
- Not hard to get data in this format through SQL queries
- But what if these are scattered across messy databases
- Or say information is in JSON files
- Or you want features that are from unstructured data sources
- Or your data has lots of errors

Data science promised big things using these methods

- Predict things - supervised
  - Click an ad, is a purchase fraud, who you should friend, driving time, shipping time, what posts you should see, etc etc etc
- Uncover hidden insights - unsupervised
  - Customer segmentation, anomaly detection, market basket, etc
- Essentially, all things that would (and very much do) make money. Lots of money.
- Very cool models and tools to do all this
  - Regression, logistic regression, SVM, XGBoost, Decision Trees, naive Bayes, knn, k-means, dbscan, hierarchical clustering, PCA, t-SNE, etc
  - Tableau, PowerBI, Looker, Qlikview

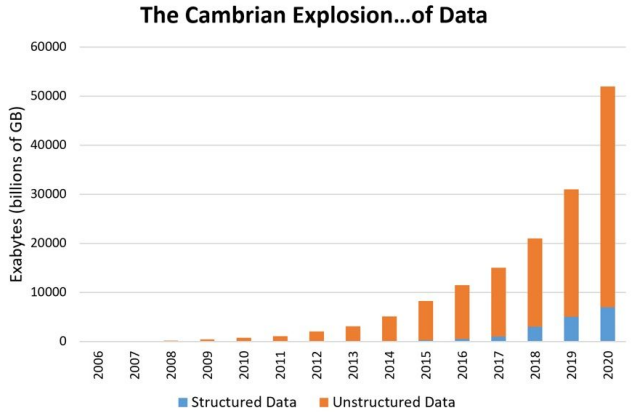
Lots of hype was had

- Tons of media about data being the future implying it can do anything
- Sexiest job title
- Terrible graphics - ohhhhhh
- ‘Oh so you do code?’
  - Quote - my aunt
- Data science boomed, and everyone wanted these results
- But…



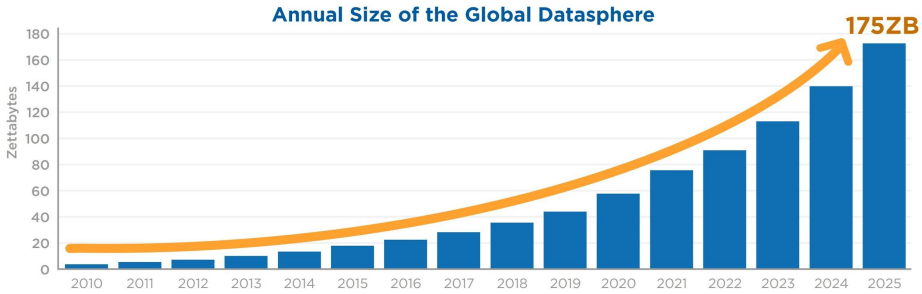
When hype meets reality

- So what's the problem with implementing these models or using those tools?
- Virtually all of the models follow this structure
  - Target ~ feature_1 + feature_2 + feature_3 + … + feature_n
  - This is just a n x m matrix of your target and features
  - And the tools need a structured DB format
  - Easy if your data is already structured



When hype meets reality

- So what's the problem with implementing these models or using those tools?
- Virtually all of the models follow this structure
  - Target ~ feature_1 + feature_2 + feature_3 + … + feature_n
  - This is just a n x m matrix of your target and features
  - And the tools need a structured DB format
  - Easy if your data is already structured
- Oh, and there's TONS of it
  - Too much to even process locally

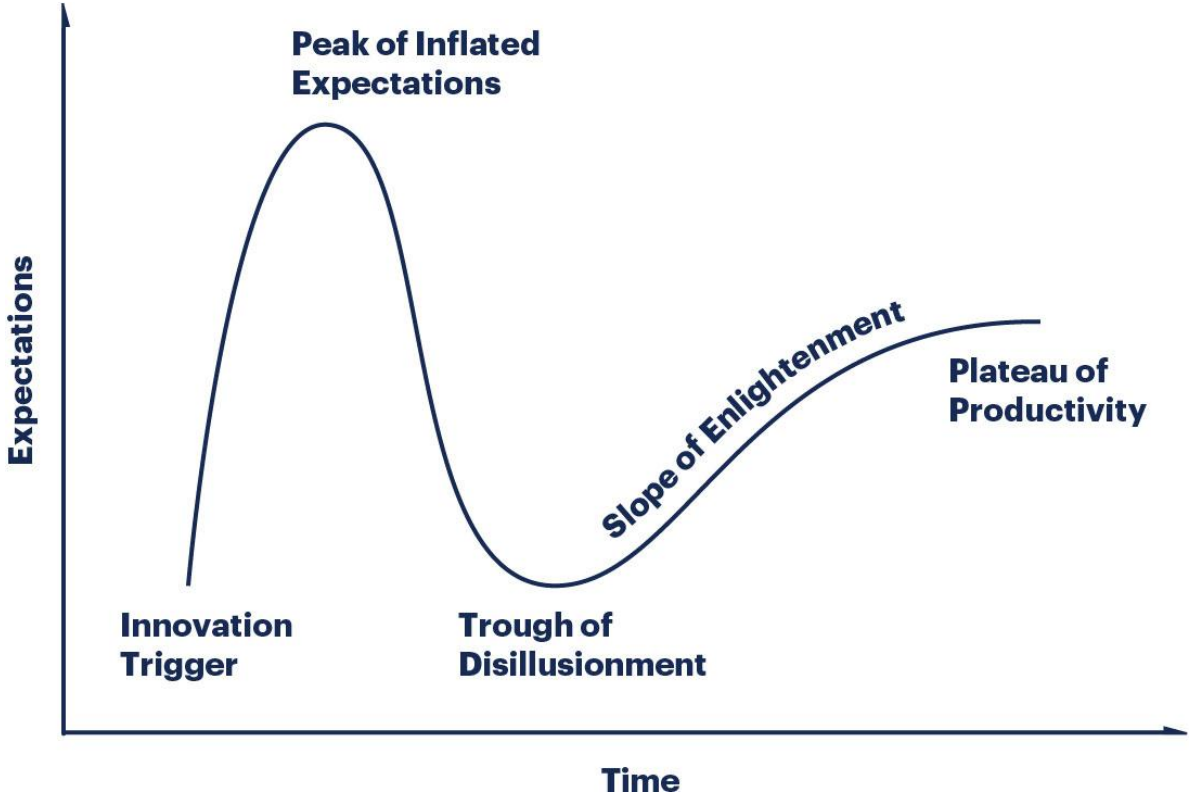


When hype meets reality

- Companies hired people that could use the data
  - But those people didn’t have access to it!
  - It was messy, undocumented, in various databases, unstructured, missing, etc etc etc
- Early data scientists wound up also having to work really hard to get the data
  - Extract it from various sources
  - Clean, manipulate, aggregate, etc… transform it into something useful
  - Load it somewhere to be used by those fancy fancy models

When hype meets reality

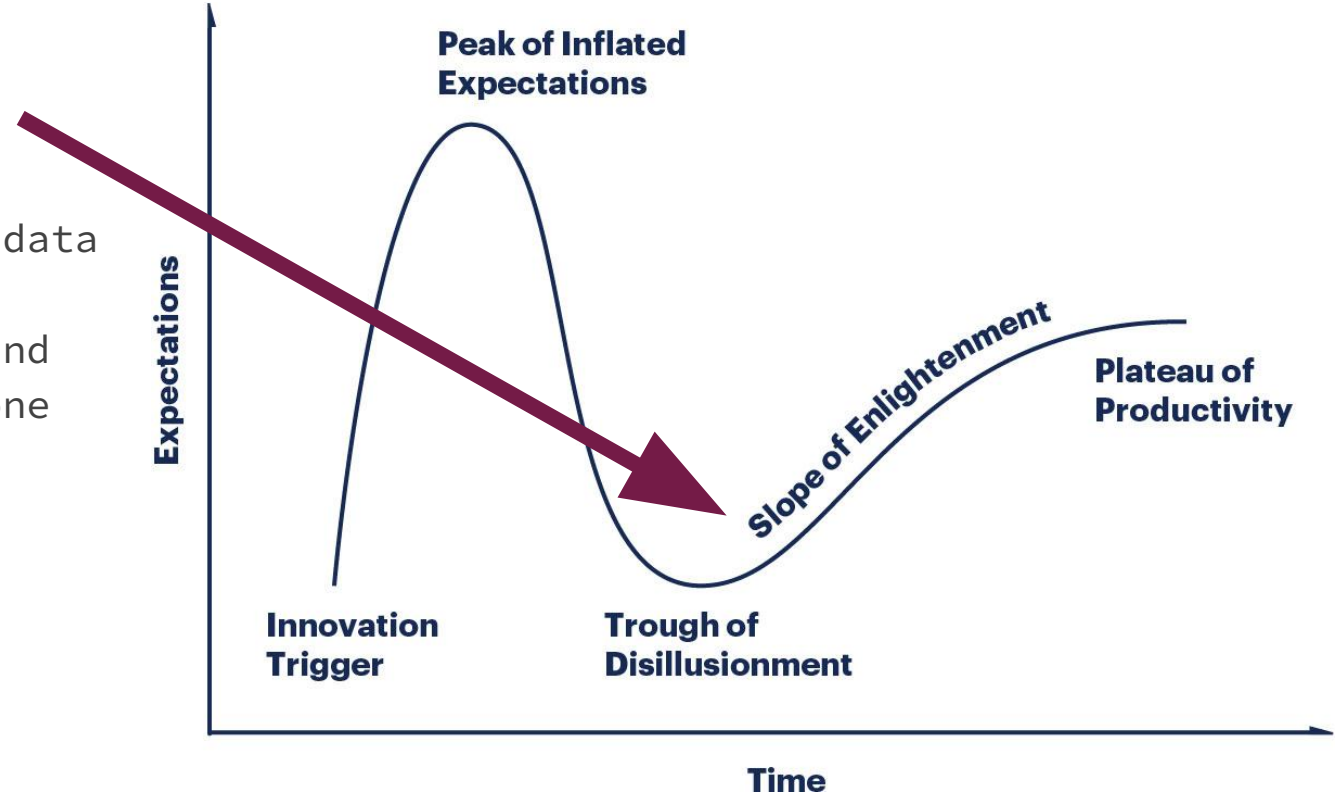
- This caused problems
- ‘We hired you, where are my results?’
- ‘Well, it was 1yr of DE work before I could run a regression’
- ‘Well, need to spend lots of money to even be able to process our raw data’



[Ref - Gartner hype cycle]

The reports of my death are greatly exaggerated

- In come the data engineers!
- Hiring people explicitly to do data engineering
- Data science by and large can’t be done without data engineering



[Ref - Gartner hype cycle]

Pyramid of needs

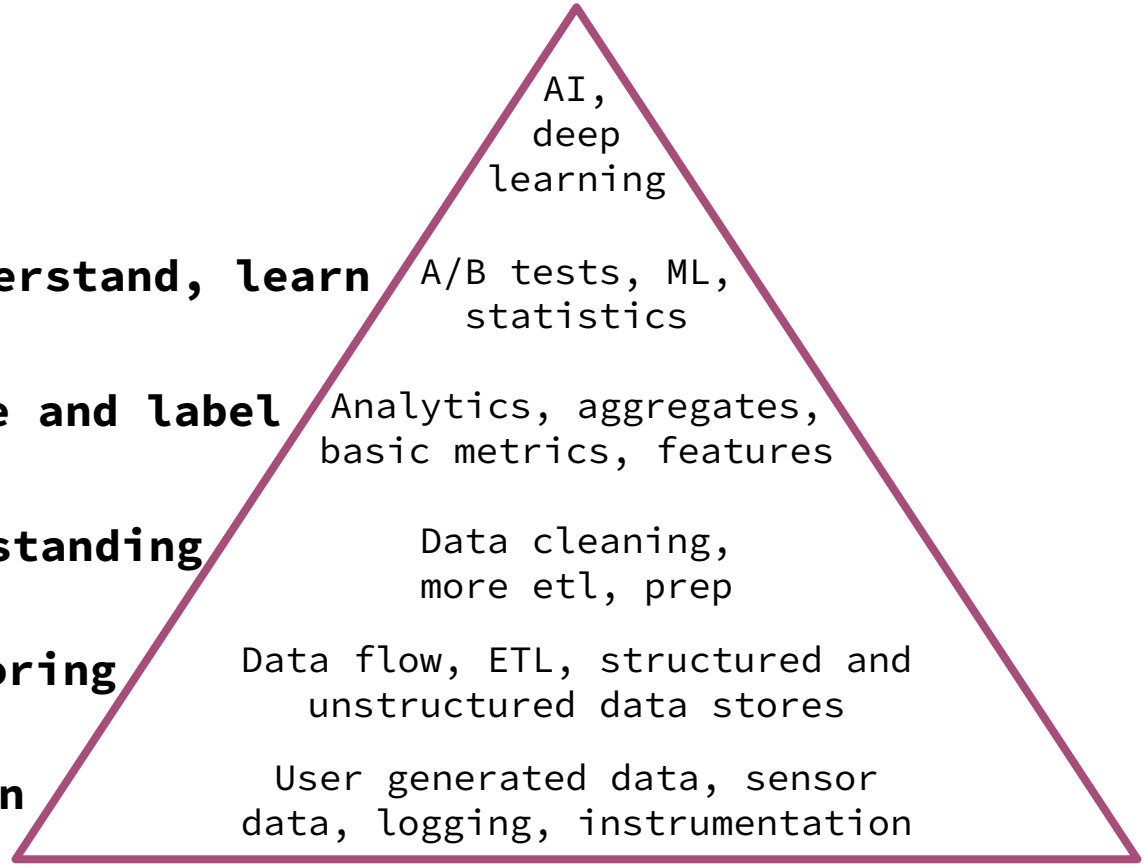


Fig adapted from - Ref

Pyramid of needs

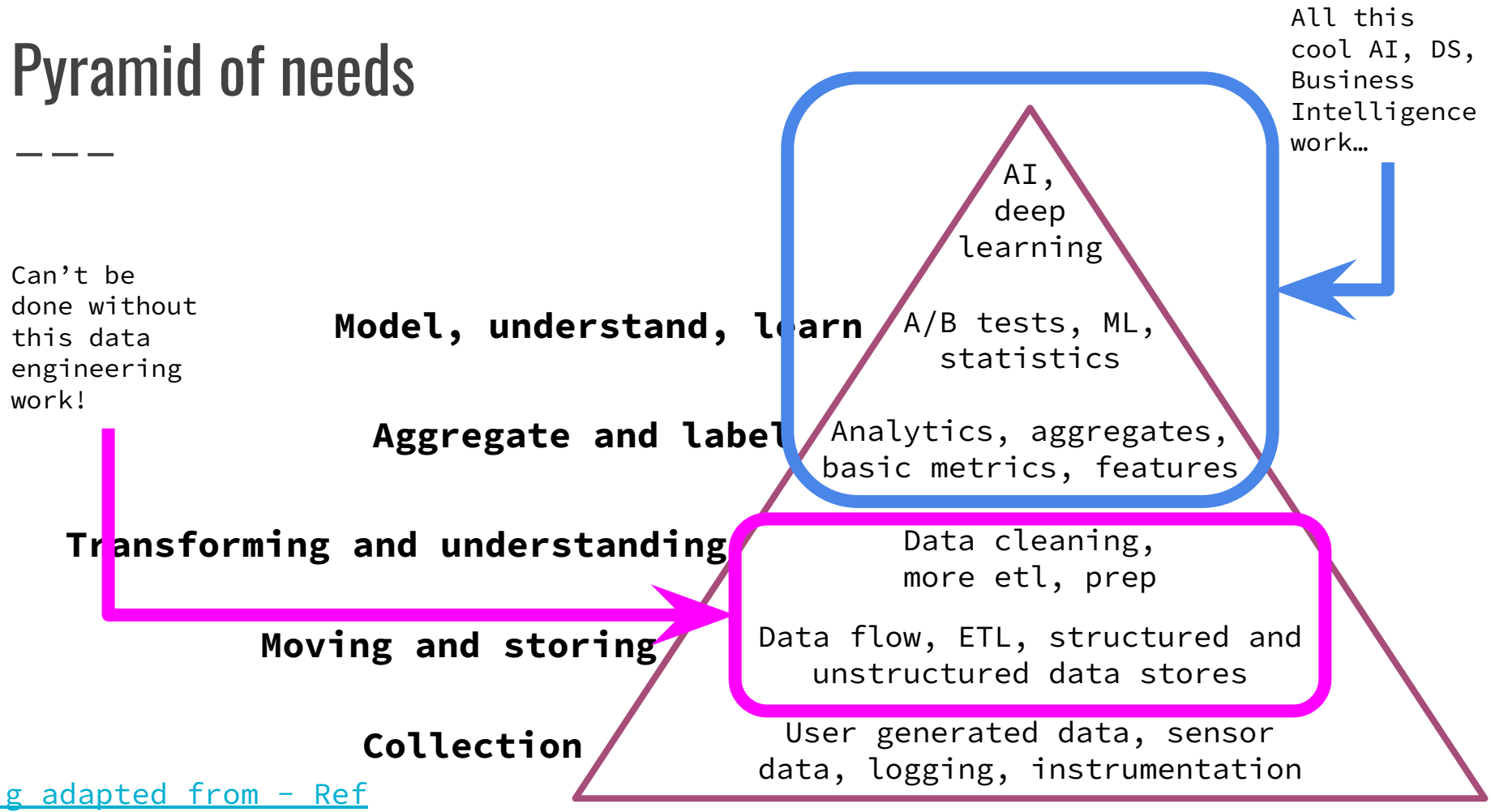


Fig adapted from - Ref

Data engineering is foundational to data science

- Data engineers take the raw data that’s stored and coming in from all over the place and transform it to a useful format
- Data scientists take this data and make inference from it
- Ideally they work hand and hand
  - DS knows what the models need, DE knows how to get those data
- Industry hiring is following the pyramid
  - Linkedin DS job search in Phoenix on 08/17/2020 = 204 jobs
  - Linkedin DE job search in Phoenix on 08/17/2020 = 347 jobs

So what's a data engineering again?

- Should be clear why we need data engineering
- Next lecture dive deeper into
  - Where data comes from - bottom of the pyramid
  - What tasks DEs do to make it useful - next two levels of pyramid
- Read Hammerbacher 2009 - Rise of the Data Scientist
  - Great story of a DS/DE who helped build the early data infrastructure of Facebook
- Read Rogati 2017 - The AI Hierarchy of Needs
  - Short but good blog post on why DE is fundamental