La Serena School for Data Science

Intro to DS 2024

Federica Bianco
University of Delaware
Rubin Observatory
Special thanks to A. Bayo

access this presentation at

https://slides.com/federicabianco/lsdss24_intro

1/5

what data science is (not)

What is data science?

Join at

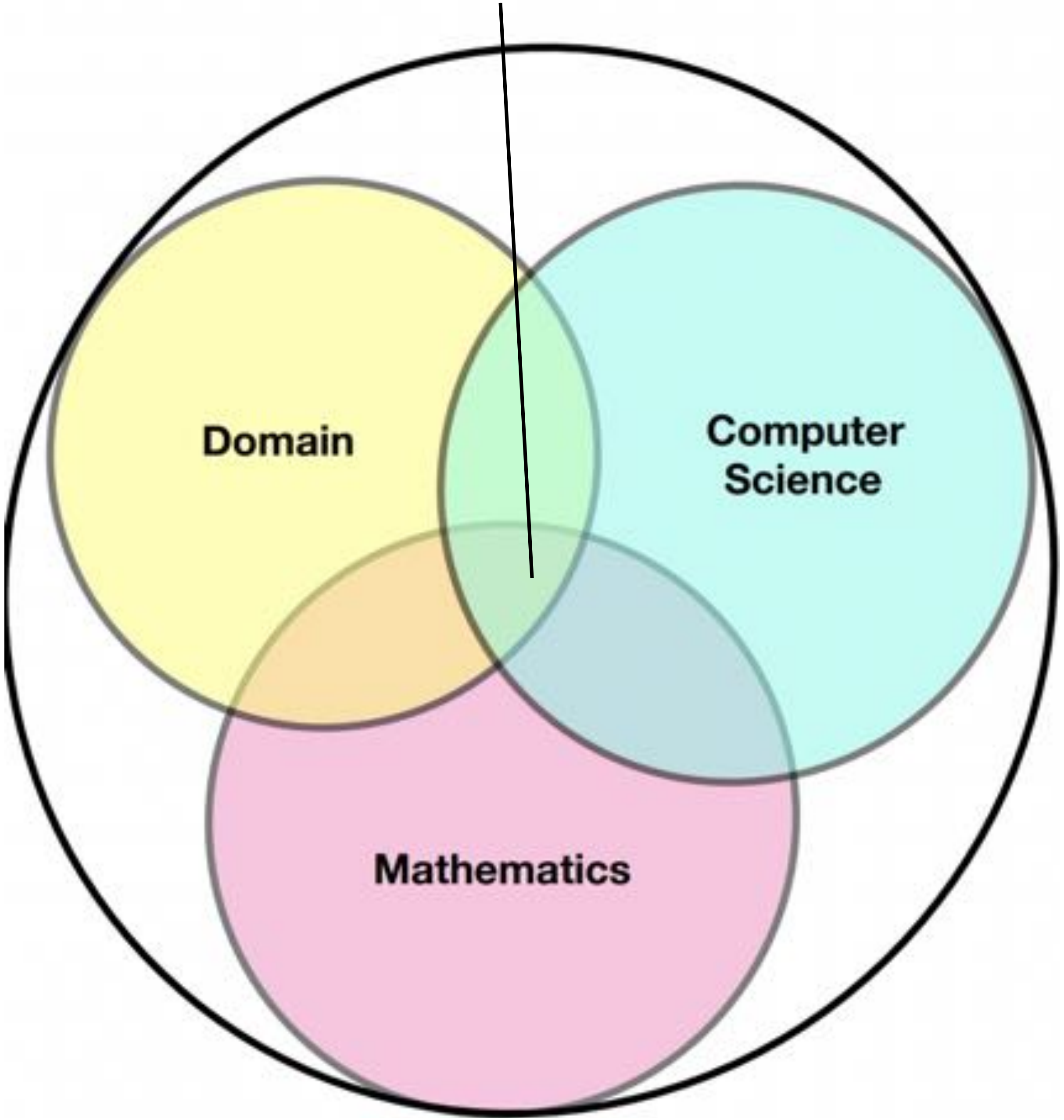
slido.com

#2546 631

What is data science?

https://wall.sli.do/event/xitEfdRvwXv4f8ysYnTqsq?section=c76bf790-f879-45de-b6ee-96df010539bf

Data Science



DS

DS:

The discipline that deals with extraction of information from data

DS ≠ ML

DS:

The discipline that deals with extraction of information from data

ML:

Machine Learning is the domain that develops, interprets, and applies mathematical model with *parameters* that are learned from data.

What are the necessary skills for a data scientist?

Join at

slido.com

#2546 631

What are the necessary skills for a data scientist?

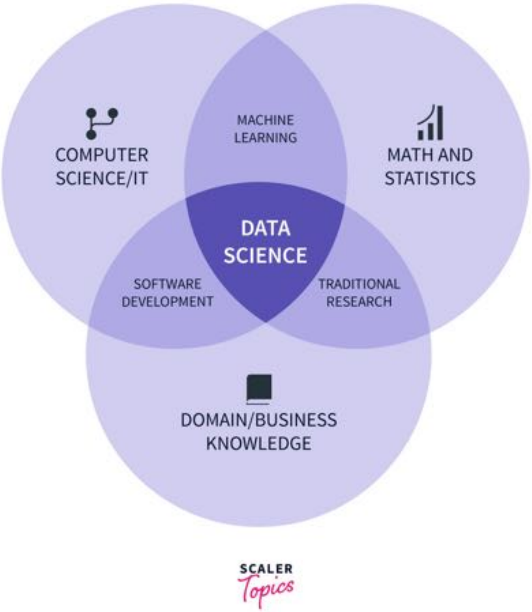
https://wall.sli.do/event/xitEfdRvwXv4f8ysYnTqsq?section=c76bf790-f879-45de-b6ee-96df010539bf

What are your strengths and assets?

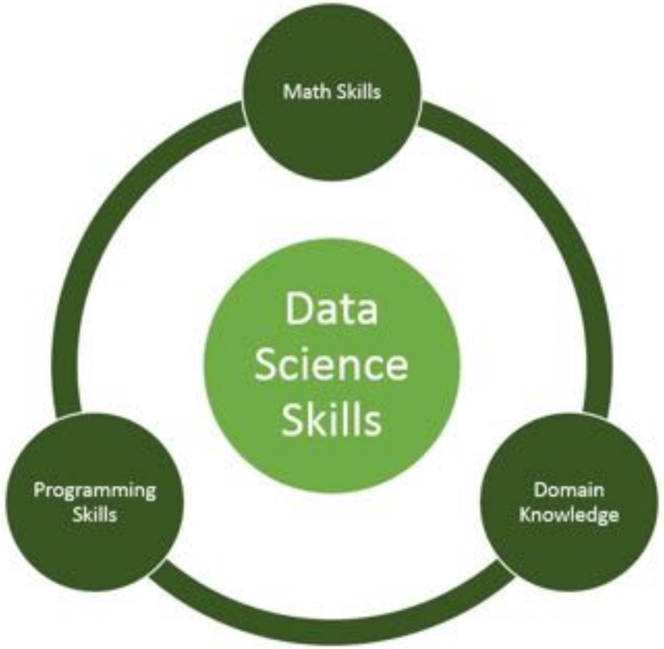
Join at
slido.com
#2546 631

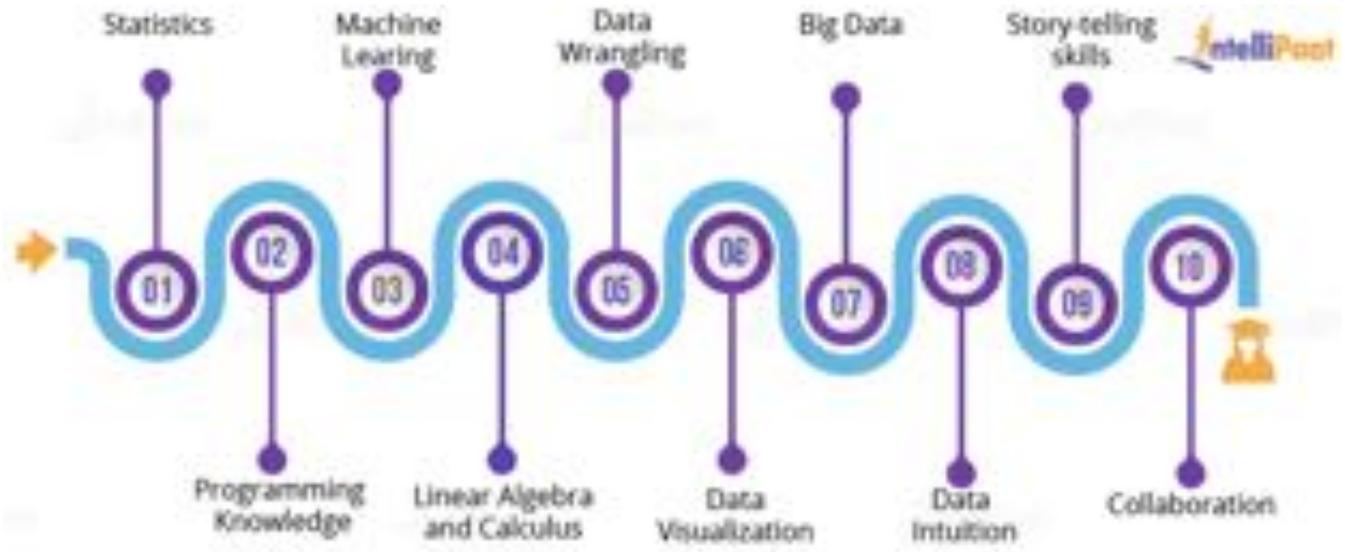
What are your strengths and assets?

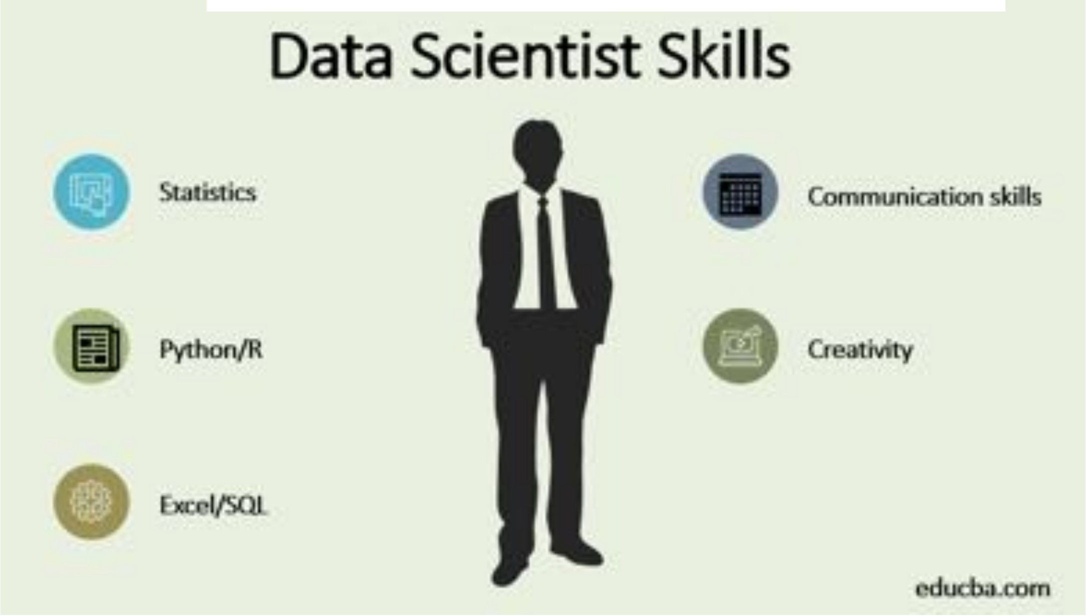
https://wall.sli.do/event/xitEfdRvwXv4f8ysYnTqsq?section=c76bf790-f879-45de-b6ee-96df010539bf

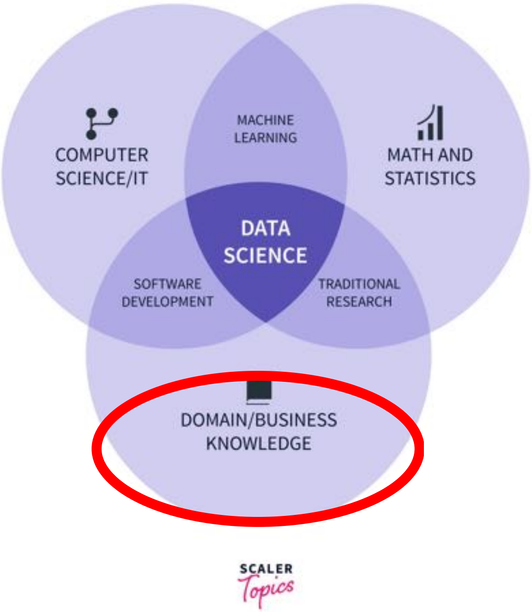

COMPUTER SCIENCE/IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAIN/BUSINESS KNOWLEDGE
SCALER Topics


Data Flair
Data Science
Visualization
Statistical Modeling
Statistical Computing
Data Technology
Data Research
Data Consulting
Real-world Application
Scientific Methods


Math Skills
Data Science Skills
Programming Skills
Domain Knowledge

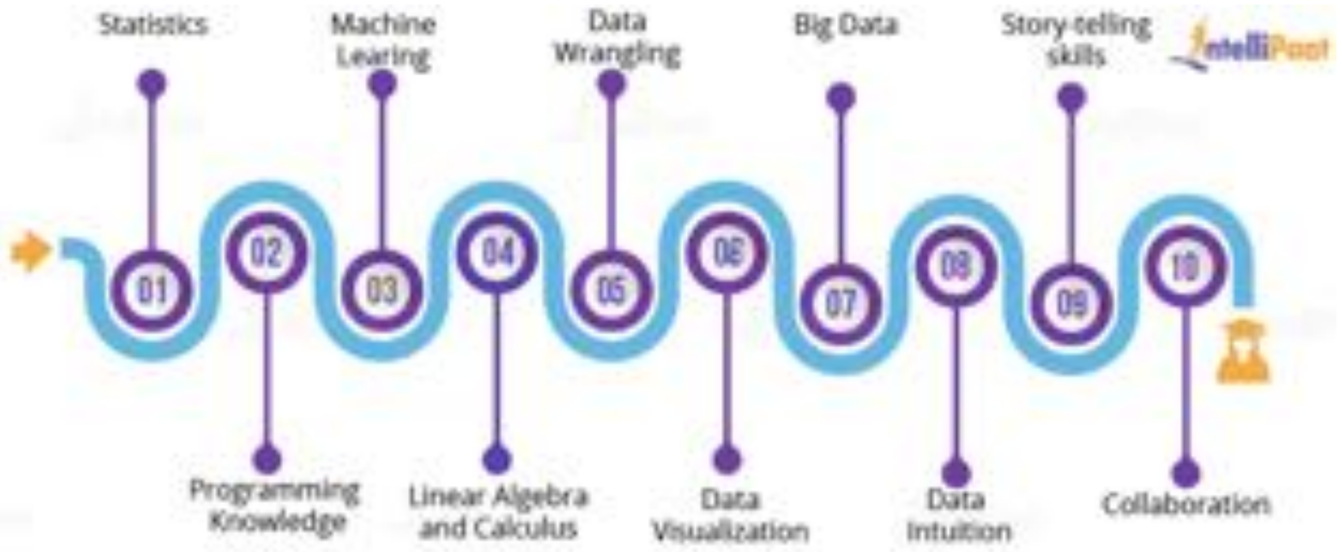

Statistics
Machine Learing
Data Wrangling
Big Data
Story-telling skills
Programming Knowledge
Linear Algebra and Calculus
Data Visualization
Data Intuition
Collaboration

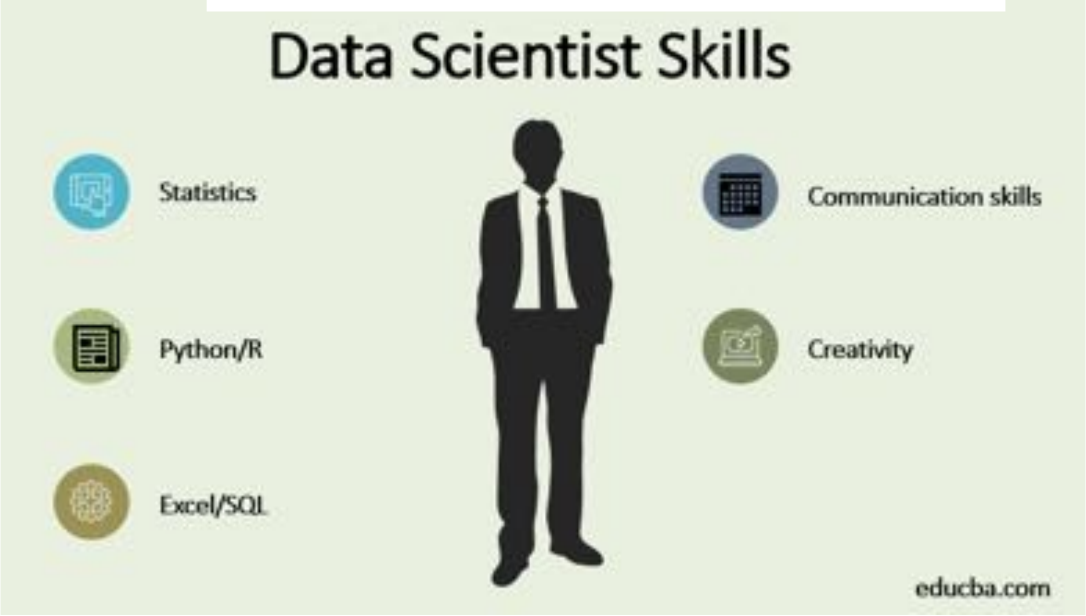

Data Scientist Skills
Statistics
Python/R
Excel/SQL
Communication skills
Creativity
educba.com

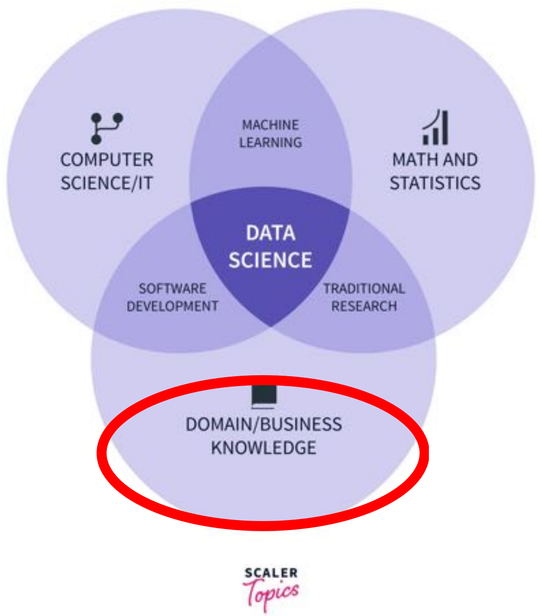

COMPUTER SCIENCE/IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAIN/BUSINESS KNOWLEDGE
SCALER Topics


Data Flair
Data Science
Visualization
Statistical Modeling
Statistical Computing
Data Technology
Data Research
Data Consulting
Real-world Application
Scientific Methods


Math Skills
Data Science Skills
Programming Skills
Domain Knowledge

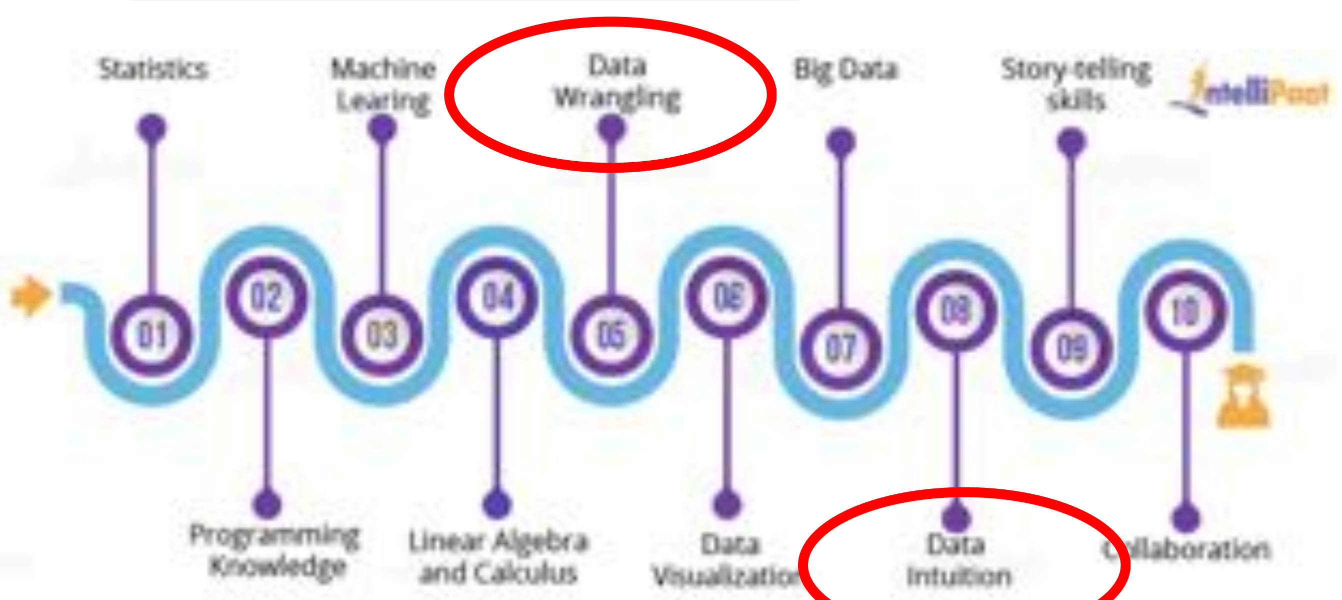

Statistics
Machine Learing
Data Wrangling
Big Data
Story-telling skills
01
02
03
04
05
06
07
08
09
10
Programming Knowledge
Linear Algebra and Calculus
Data Visualization
Data Intuition
Collaboration


Data Scientist Skills
Statistics
Python/R
Excel/SQL
Communication skills
Creativity
educba.com

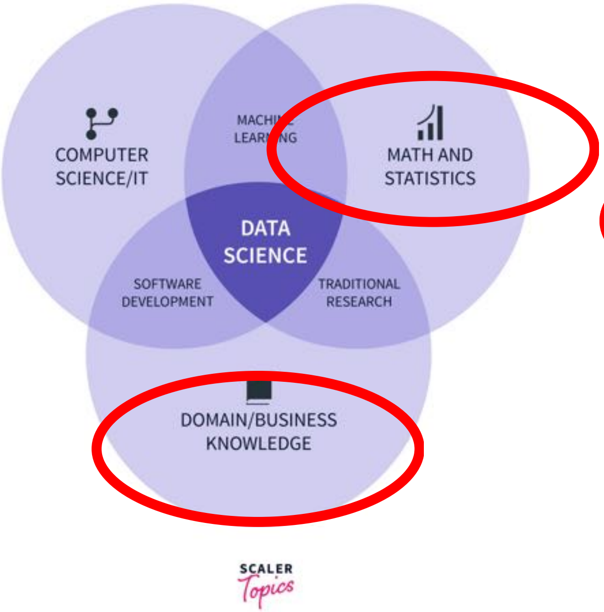

COMPUTER SCIENCE/IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAIN/BUSINESS KNOWLEDGE
SCALER Topics


Data Flair
Visualization
Scientific Methods
Statistical Modeling
Data Science
Real-world Application
Statistical Computing
Data Consulting
Data Technology
Data Research

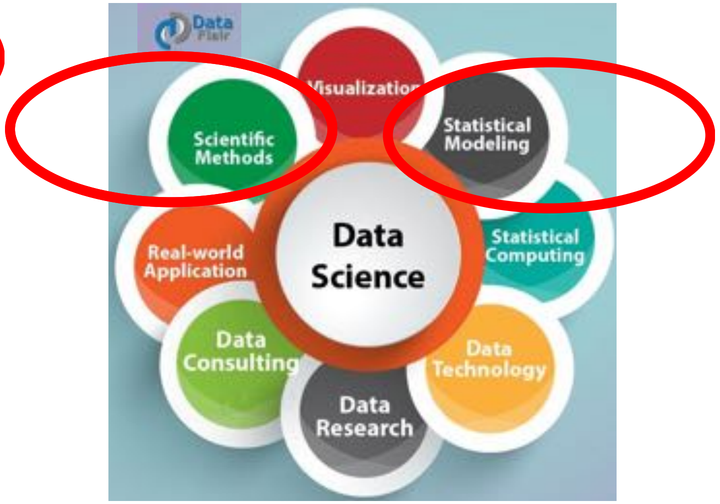

Math Skills
Data Science Skills
Programming Skills
Domain Knowledge


Statistics
Machine Learing
Data Wrangling
Big Data
Story-telling skills
01
02
03
04
05
06
07
08
09
10
Programming Knowledge
Linear Algebra and Calculus
Data Visualization
Data Intuition
Collaboration

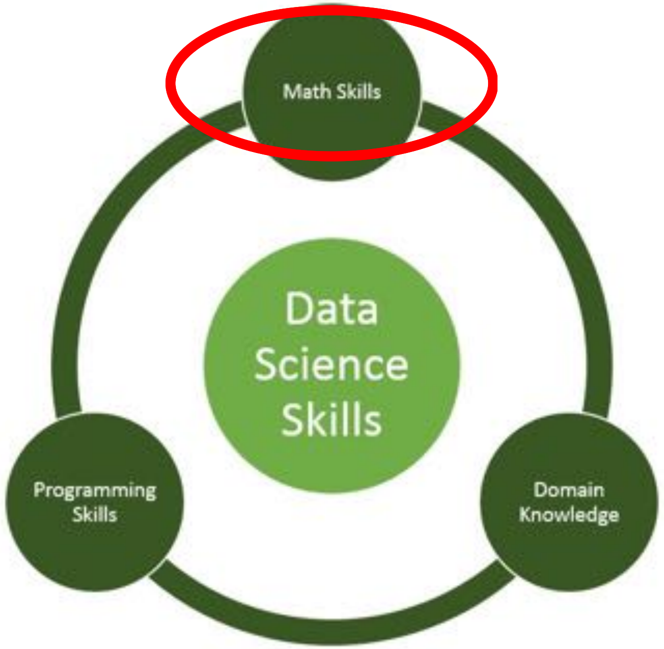

Data Scientist Skills
Statistics
Python/R
Excel/SQL
Communication skills
Creativity
educba.com



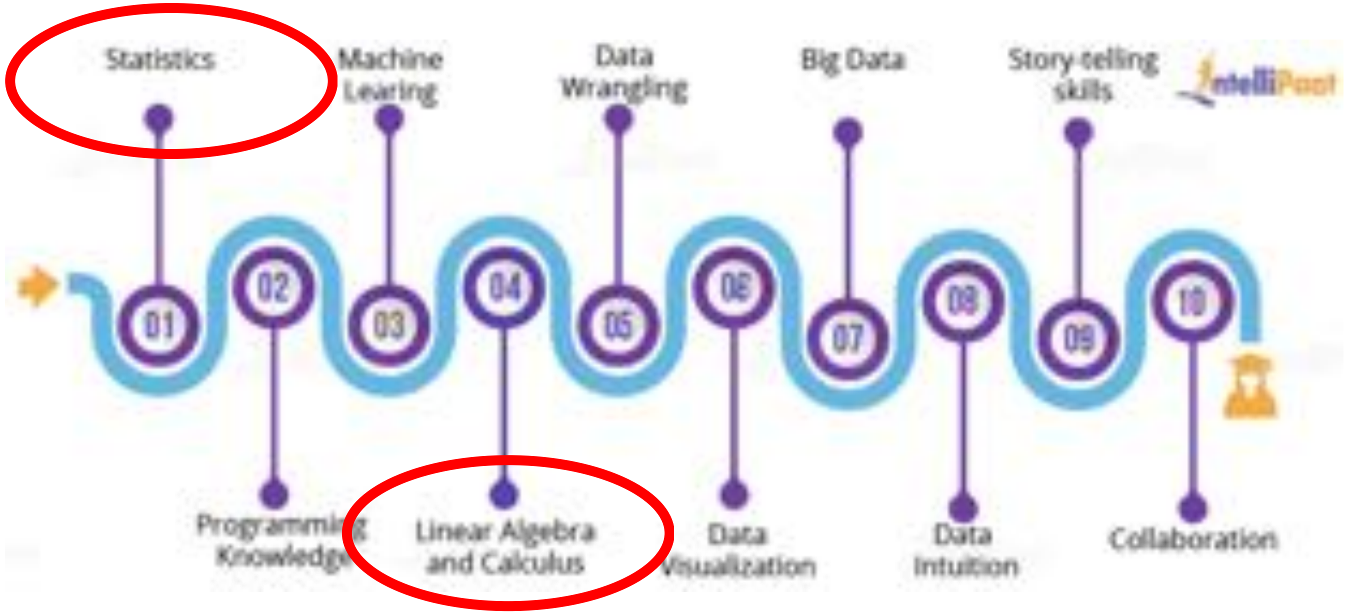
COMPUTER SCIENCE/IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAIN/BUSINESS KNOWLEDGE
SCALER Topics



Data Flair
Visualization
Scientific Methods
Statistical Modeling
Data Science
Real-world Application
Statistical Computing
Data Consulting
Data Technology
Data Research

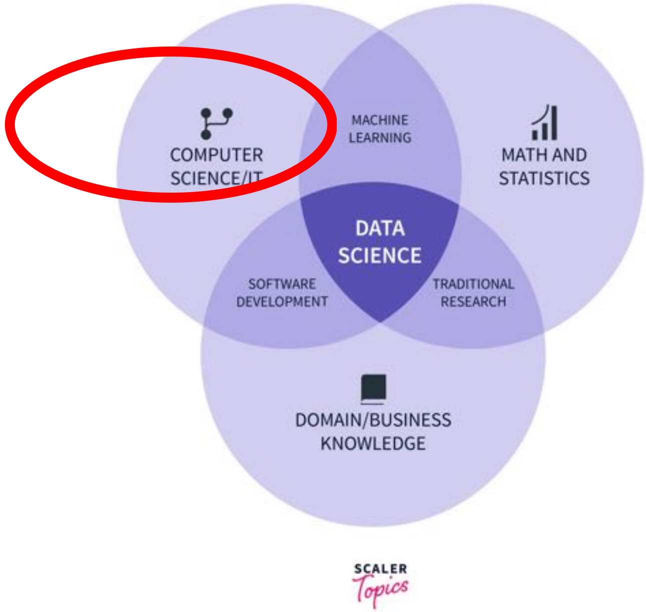


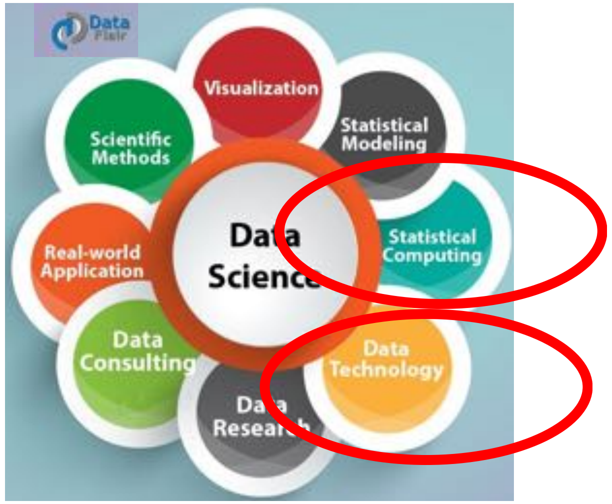
Math Skills
Data Science Skills
Programming Skills
Domain Knowledge

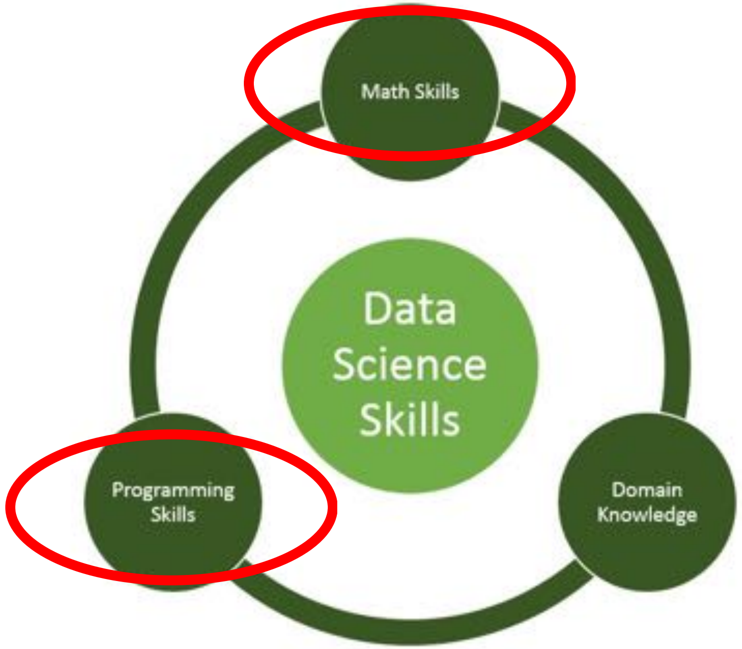


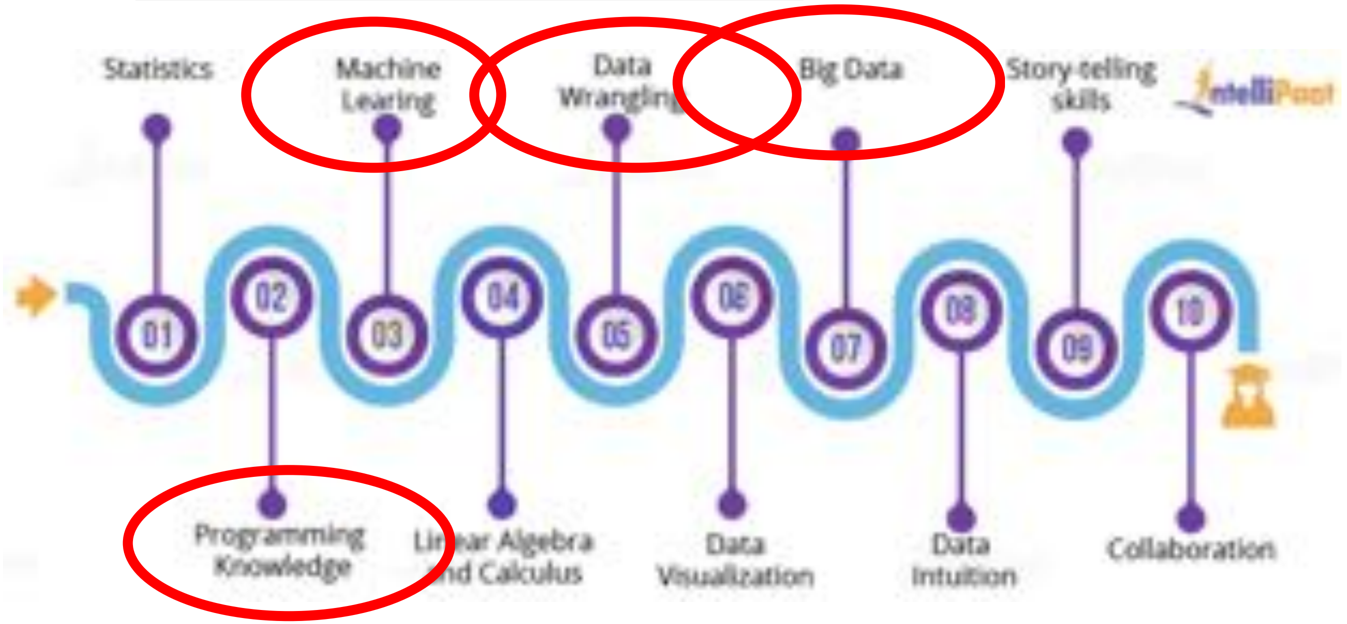
Statistics
Machine Learing
Data Wrangling
Big Data
Story-telling skills
IntelliPaat
01
02
03
04
05
06
07
08
09
10
Programmi
Knowledg
Linear Algebra and Calculus
Data Visualization
Data Intuition
Collaboration

Data Scientist Skills
Statistics
Python/R
Excel/SQL
Communication skills
Creativity
educba.com

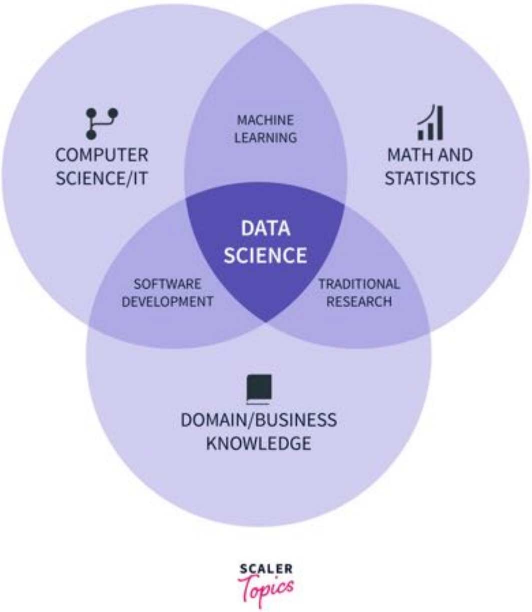

COMPUTER SCIENCE/IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAIN/BUSINESS KNOWLEDGE
SCALER Topics

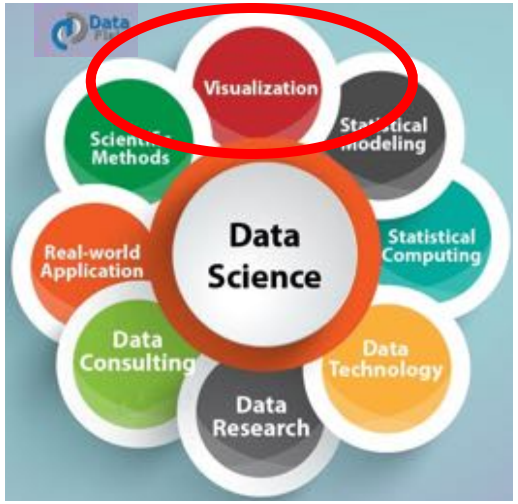

Data Flair
Visualization
Statistical Modeling
Scientific Methods
Data Science
Statistical Computing
Real-world Application
Data Consulting
Data Technology
Data Research

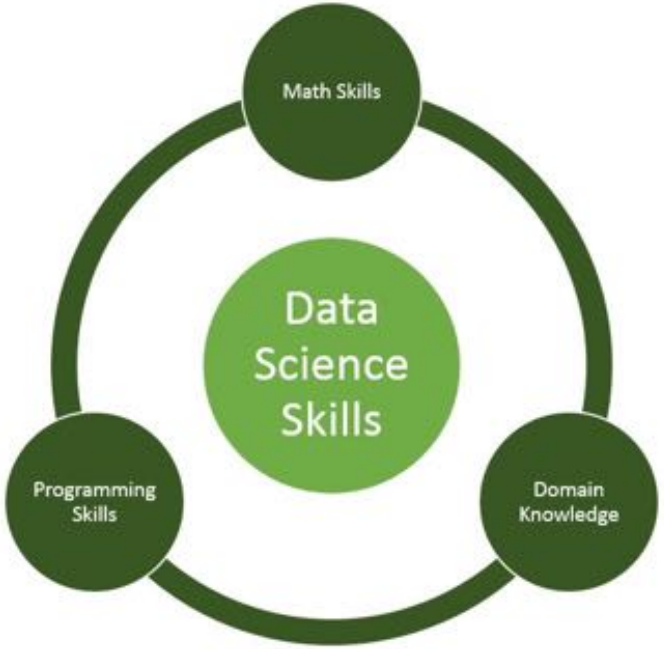

Math Skills
Data Science Skills
Programming Skills
Domain Knowledge

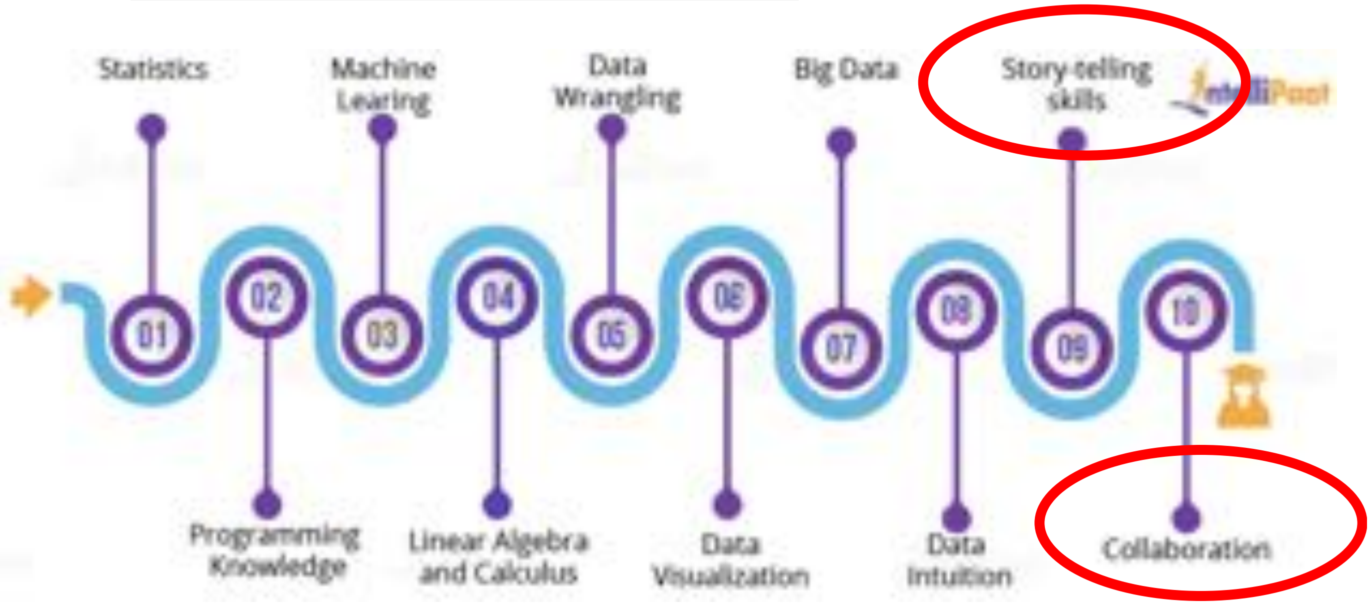

Statistics
Machine Learing
Data Wrangling
Big Data
Story-telling skills
IntelliPaat
01
02
03
04
05
06
07
08
09
10
Programming Knowledge
Linear Algebra and Calculus
Data Visualization
Data Intuition
Collaboration

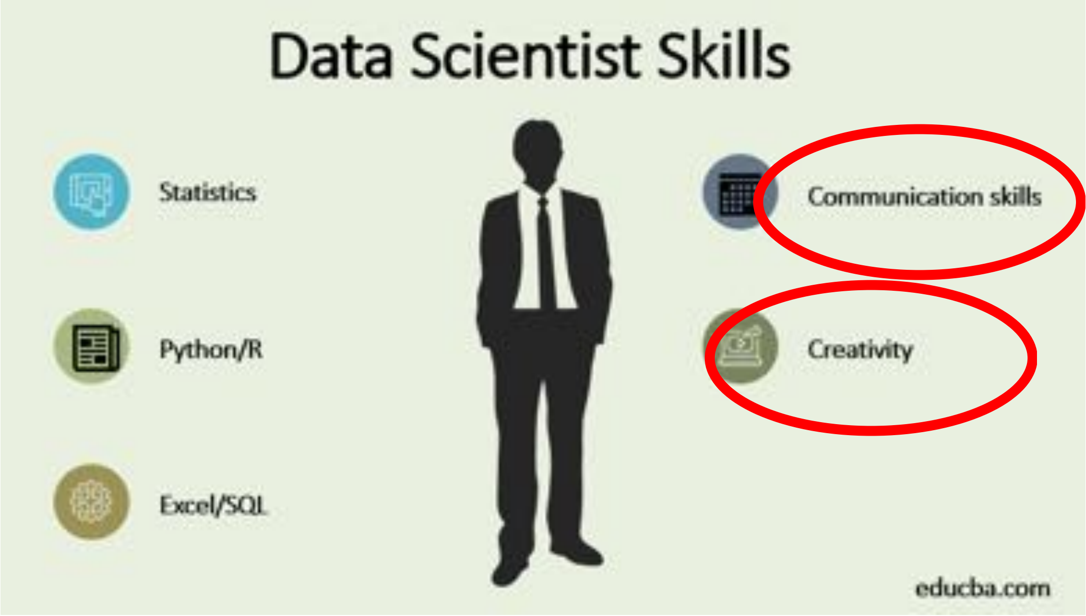

Data Scientist Skills
Statistics
Python/R
Excel/SQL
Communication skills
Creativity
educba.com


COMPUTER SCIENCE/IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAIN/BUSINESS KNOWLEDGE
SCALER Topics

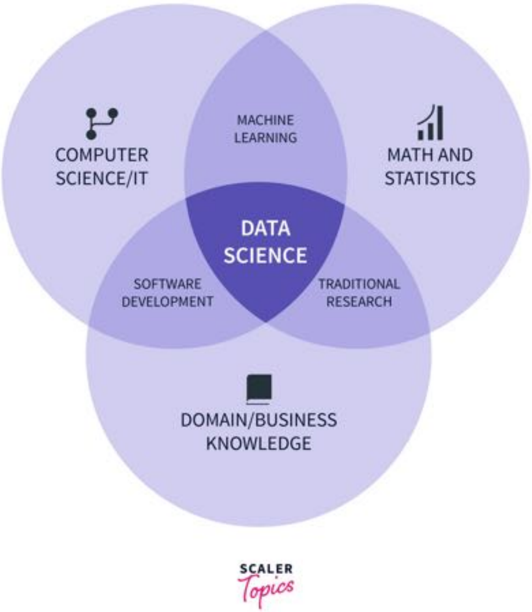

Visualization
Data Science
Statistical Computing
Data Technology
Data Research
Data Consulting
Real-world Application

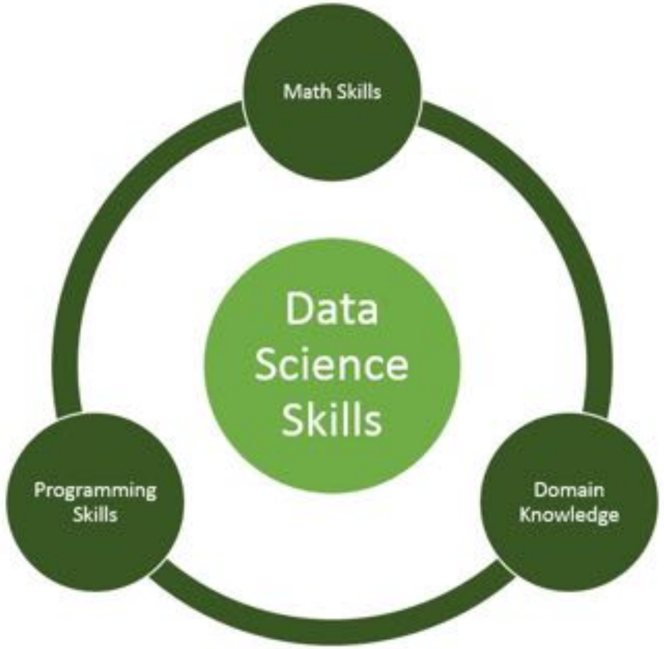

Math Skills
Data Science Skills
Programming Skills
Domain Knowledge

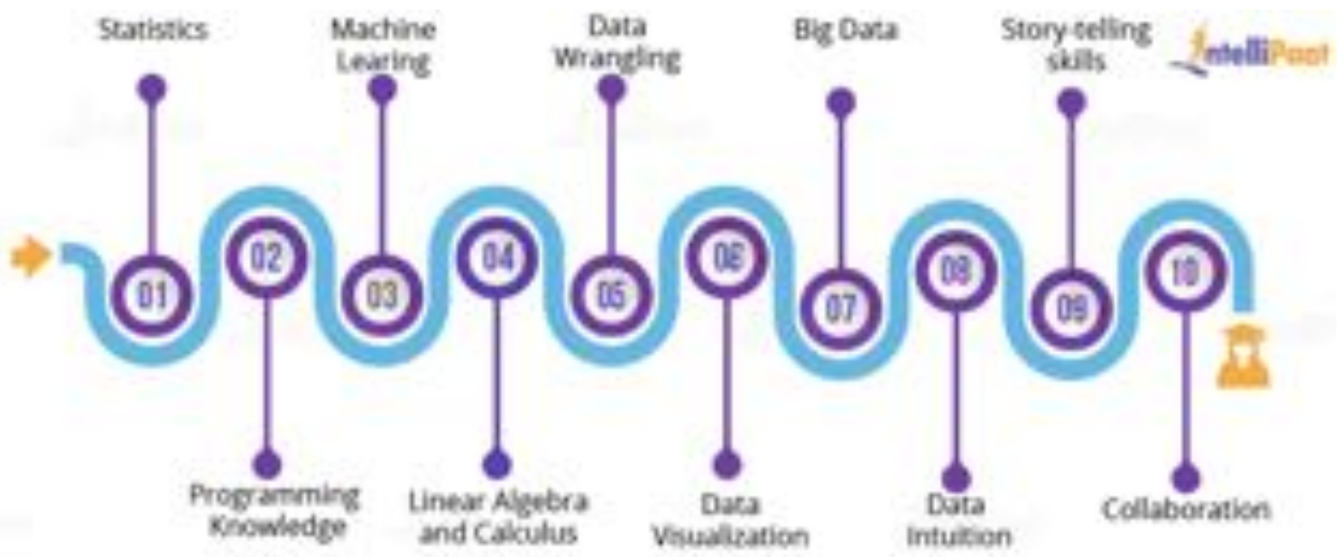

Statistics
Machine Learing
Data Wrangling
Big Data
Story-telling skills
01
02
03
04
05
06
07
08
09
10
Programming Knowledge
Linear Algebra and Calculus
Data Visualization
Data Intuition
Collaboration

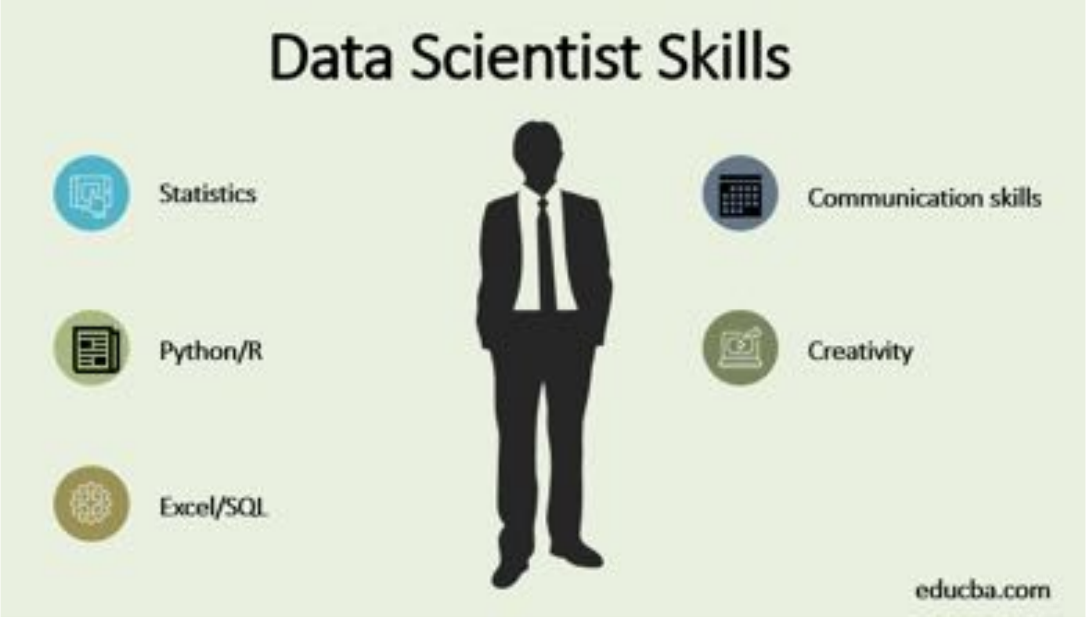

Data Scientist Skills
Statistics
Python/R
Excel/SQL
Communication skills
Creativity
educba.com

What is missing??









Life Cycle of a DS project

scientific question

Life Cycle of a DS project

scientific question

Data collection

Proposal writing

Instrument Building

Calibration

Deployment

Collection

Life Cycle of a DS project

scientific question

Data collection

Proposal writing

Instrument Building

Calibration

Deployment

Collection

...Search web for data...

Life Cycle of a DS project

scientific question

Data collection

Data exploration

Data Engineering

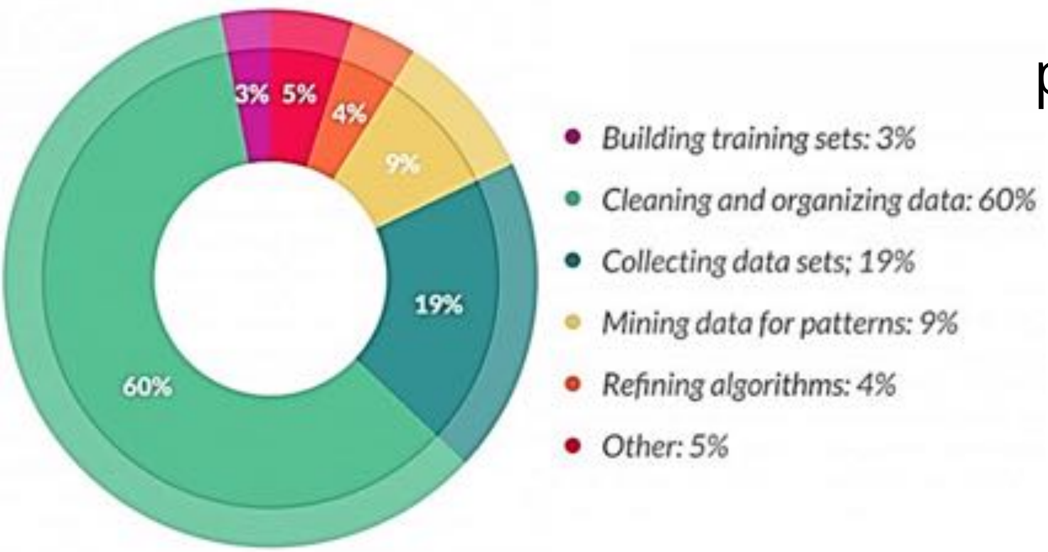
statistical analysis and extraction of statistical properties

Data Wrangling

Data cleaning

Feature extraction

Feature engineering



Data preparation and preprocessing for broadcast systems monitoring in PHM framework

- DOI: 10.1109/CoDIT.2019.8820370

DS ≠ ML

DS:

The discipline that deals with extraction of information from data, including all phases of data driven inference **from data collection through modeling and communication, and its interpretation in a domain context**

ML:

Machine Learning is the domain that develops, interprets, and applies mathematical model with *parameters* that are learned from data.

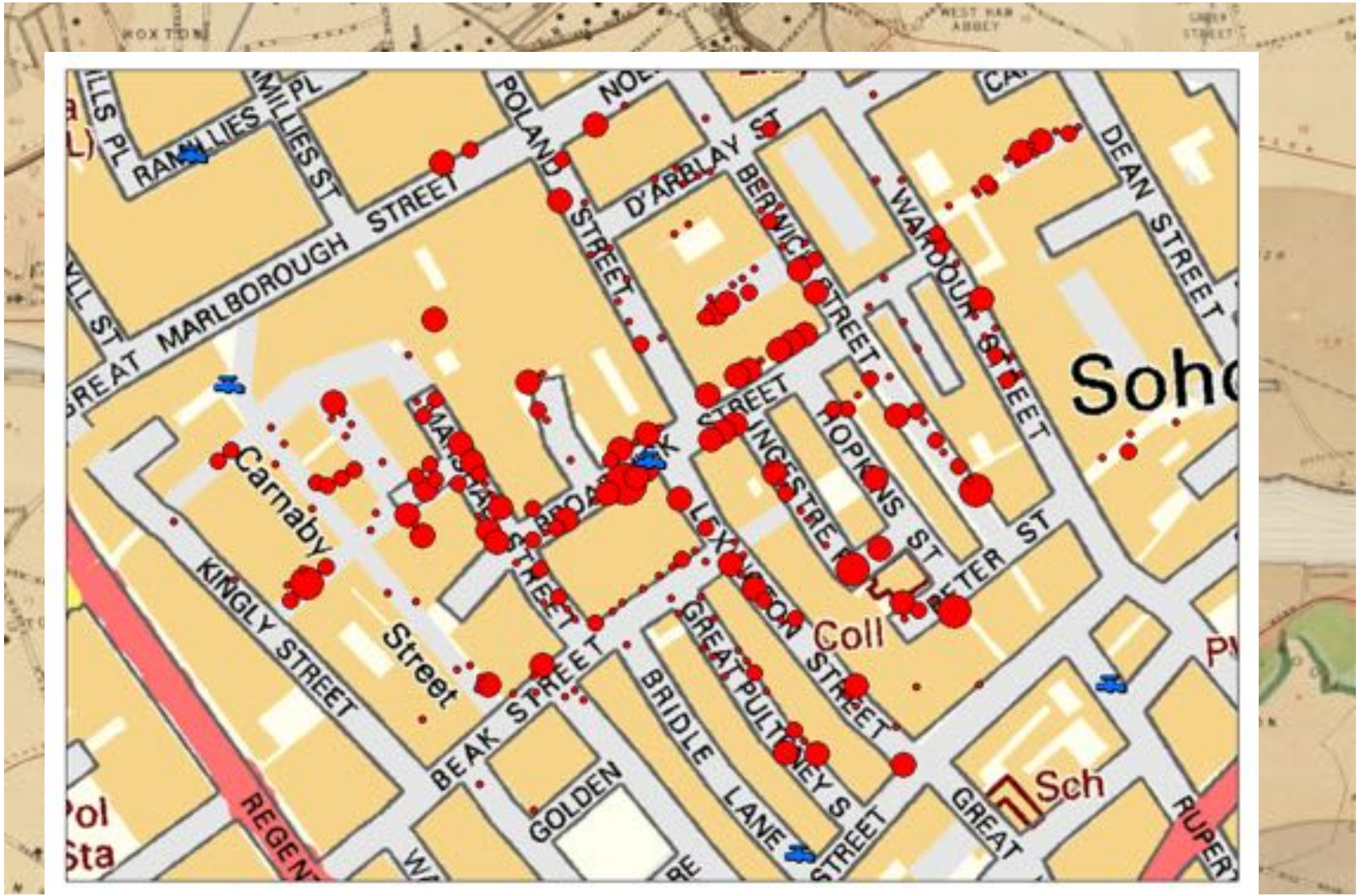
The fist data science project: John Show map of cholera

Idea driven by domain knowledge (he was a doctor)

Data collection

Data exploration

The fist data science project:
John Show map of cholera



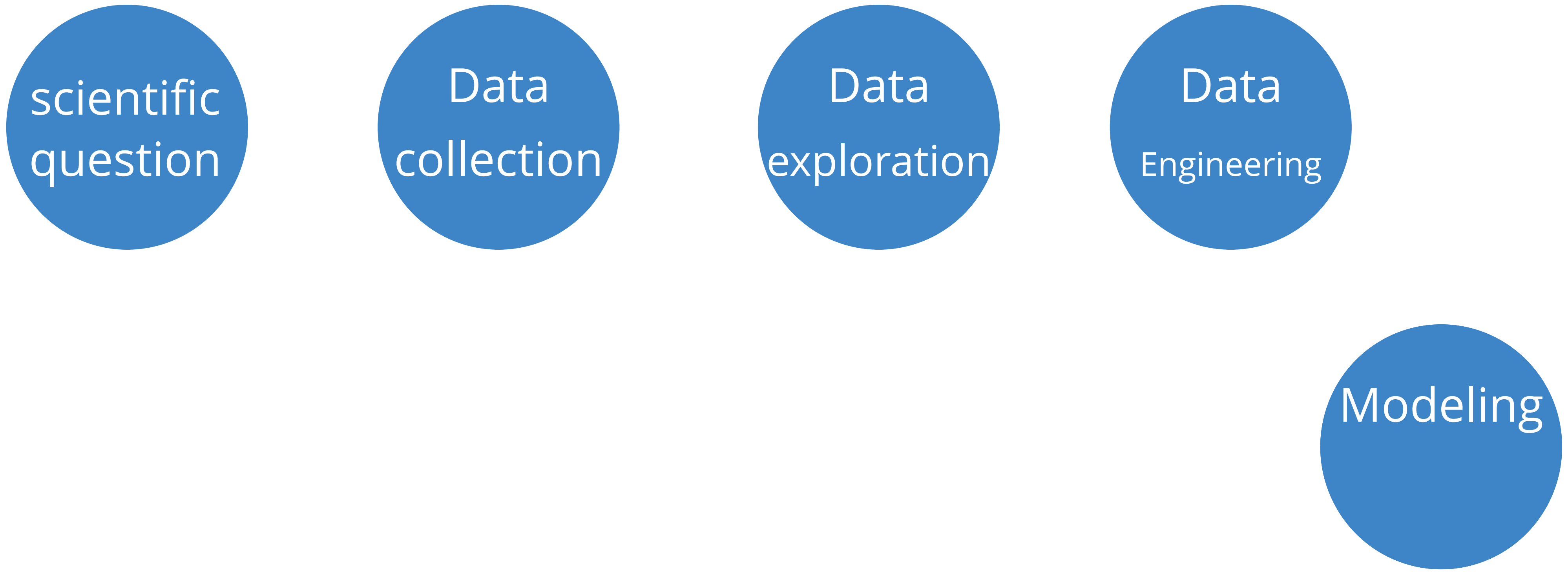
Idea driven by domain knowledge (he was a doctor)

Data collection

Data exploration

digitized data accessibel here https://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/

Life Cycle of a DS project



Apply ML models

Adapt ML model

Create ML model

Life Cycle of a DS project

scientific question

Data collection

Data exploration

Data Engineering

Modeling

Communication of results

Visualize data
Visualize results
Tell the story
Paper
Report
Presentation
Blog post...

The fist data science project: John Show map of cholera

Idea driven by domain knowledge (he was a doctor)

Data collection

Data exploration

Communication

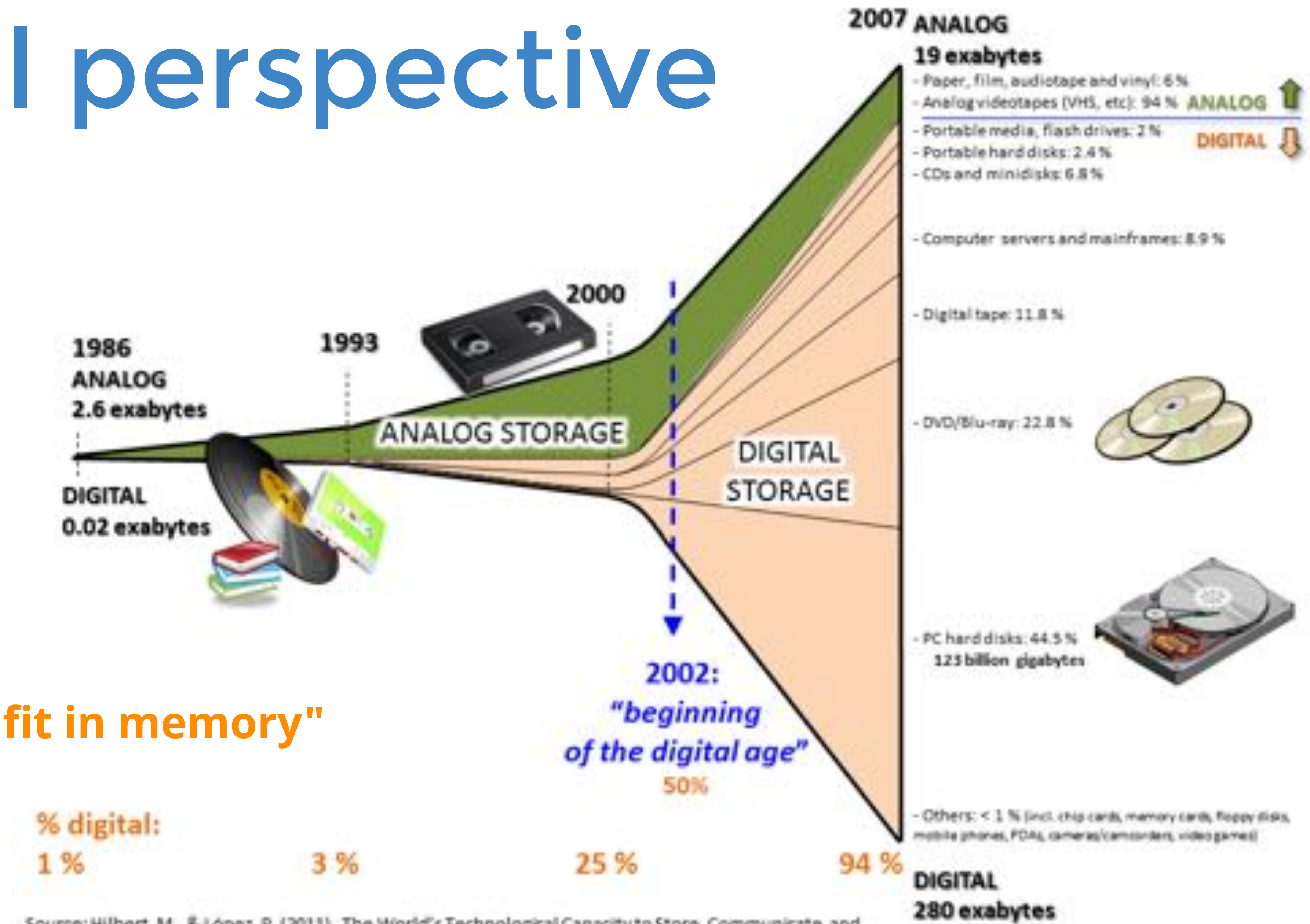
What are your strengths and assets?

https://wall.sli.do/event/xitEfdRvwXv4f8ysYnTqsq?section=c76bf790-f879-45de-b6ee-96df010539bf

2/5

what's astronomy got to do with it

Historical perspective



"Data that does not fit in memory"

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 -65. http://www.martinhilbert.net/WorldInfoCapacity.html

@fedhere

astronomical data production

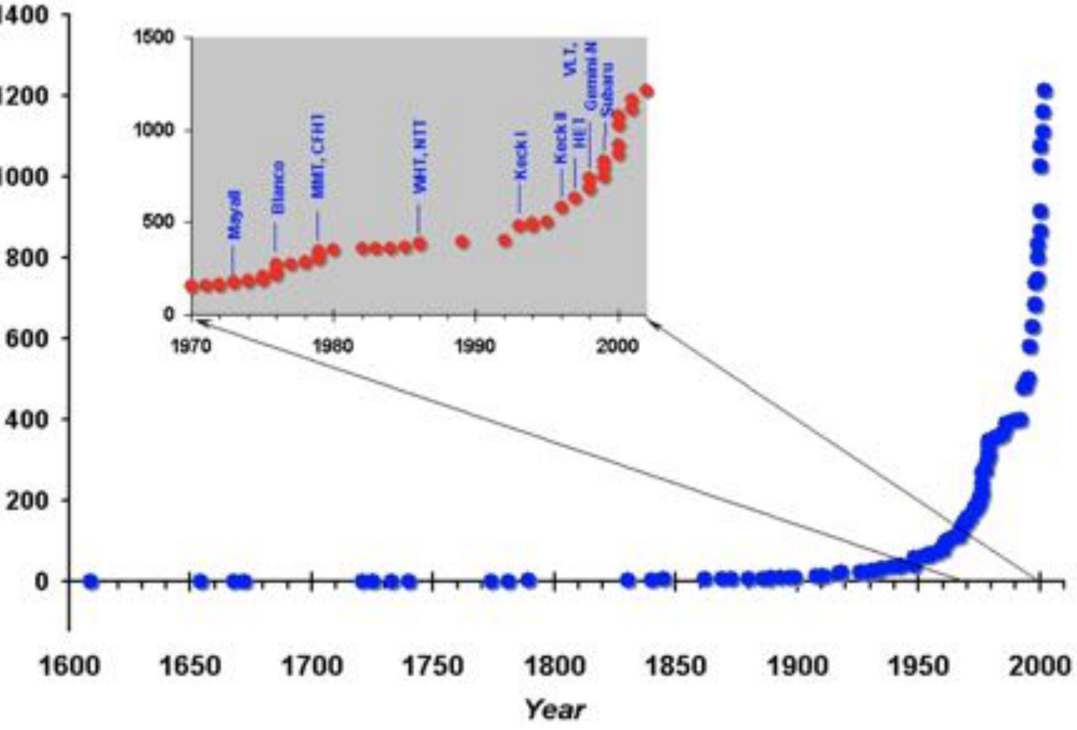
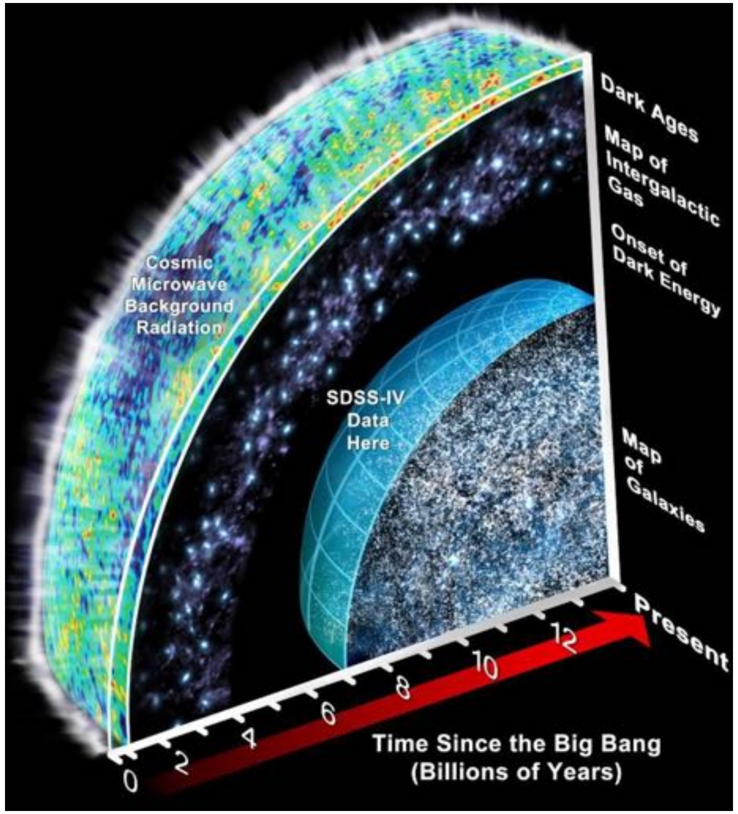


Figure 1 The growth in cumulative telescope collecting area over the past 400 years, with each point representing a completed ground-based telescope. The combination of the ability to manufacture and support large mirrors combined with adaptive optics has given the new generation of large telescopes tremendous scientific gains over the previous 4-m telescopes. For example, an 8-m telescope delivering images of 0.1 arcsec can observe point-like objects at least 20 times fainter than a conventional 4-m telescope delivering 1.0 arcsec images. Will we see such gains in the next generation of telescopes? MMT, Multiple Mirror Telescope; CFHT, Canada-France-Hawaii Telescope; WHT, William Herschel Telescope; NTT, New Technology Telescope; HET, Hobby-Eberly Telescope; VLT, Very Large Telescope; Gem-N, Gemini North Telescope.

Area vs Volume



Both data volumes and data rates grow exponentially, with a doubling time ~ 1.5 years

It is also estimated that everyone has access to 50% of the existing data!

@fedhere

Big Data: Astronomical or Genomical?

Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson

Published: July 7, 2015 • https://doi.org/10.1371/journal.pbio.1002195

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

@fedhere

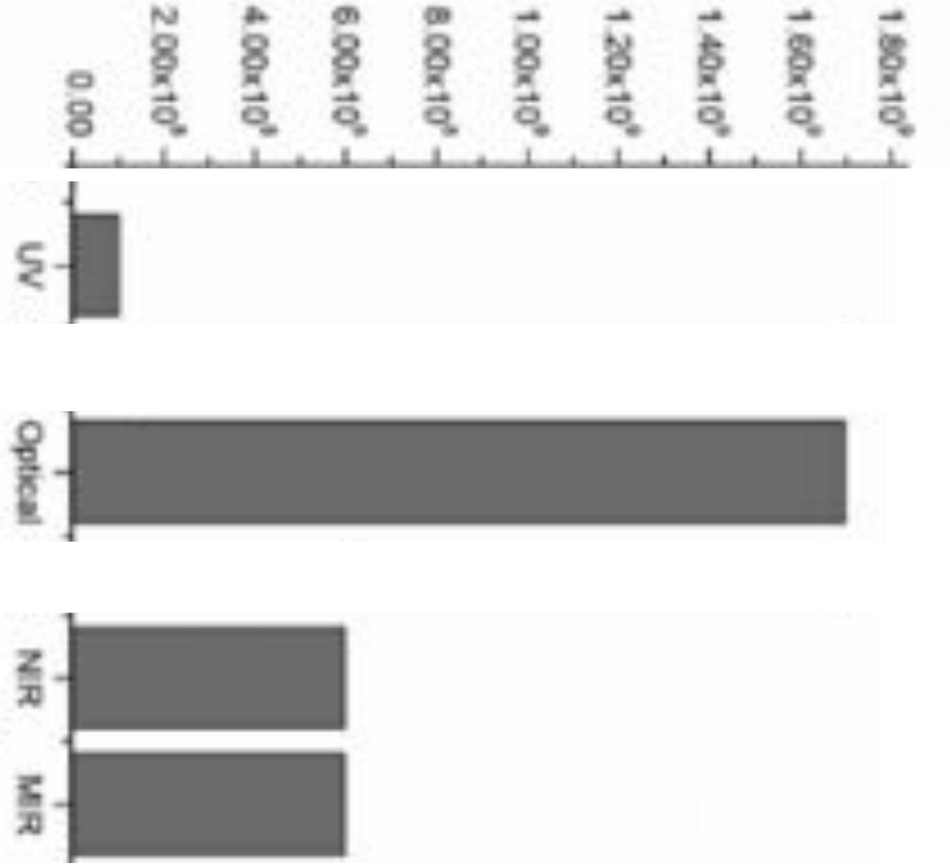
astronomical data volume

number of sources

https://www.degruyter.com/view/journals/astro/25/1/article-p75.xml?language=en

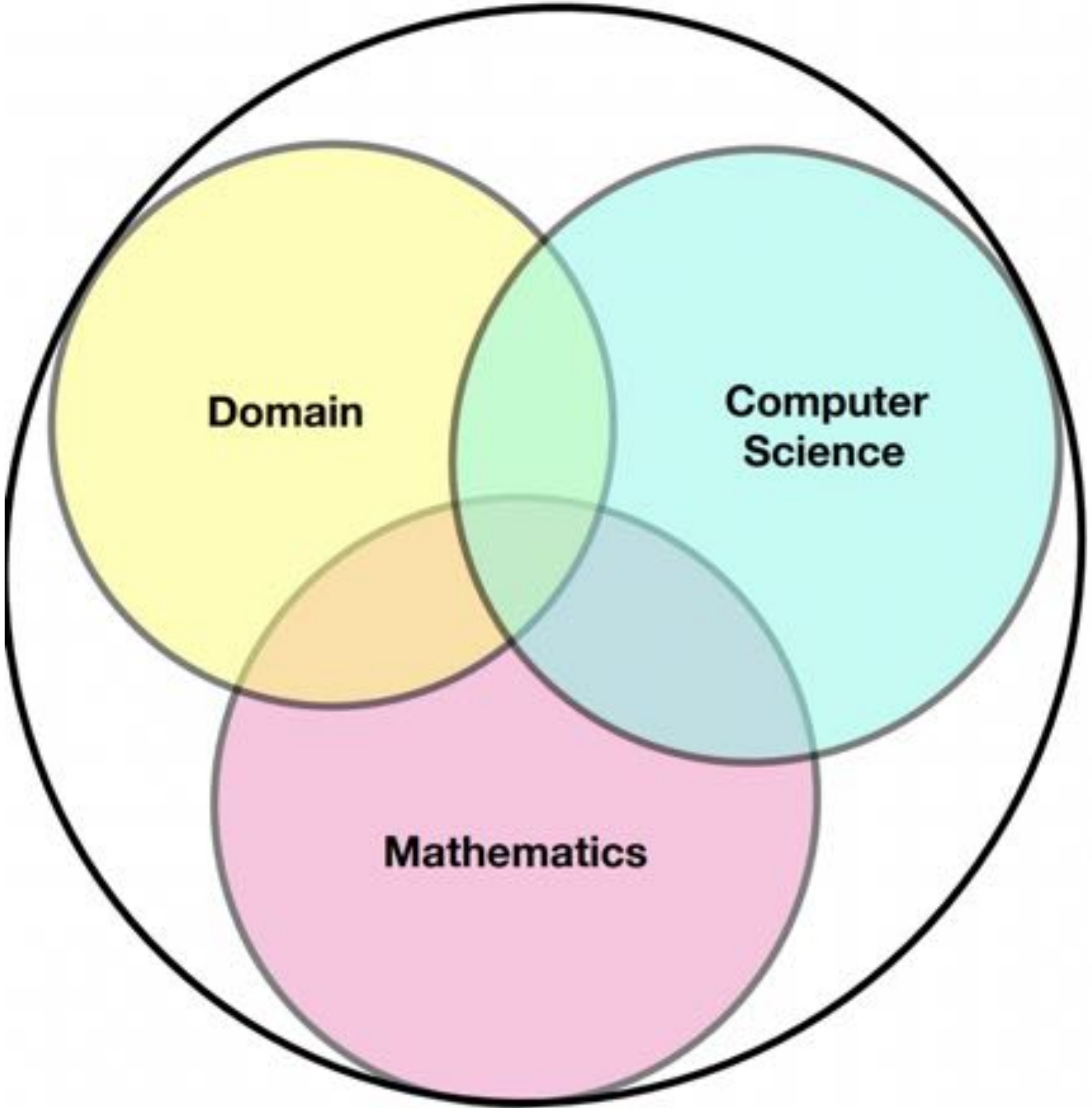
Table 1. Main data for the most important all-sky and large-area astronomical surveys providing multi-wavelength photometric data. Catalogues are given in the order of increasing wavelengths.

Survey, catalogue	Years	Spectral range	Sky area (deg^2)	Sensitivity (mag/mJy)	Number of sources	Density (obj/deg^2)
Fermi-GLAST	2008–2014	10 MeV–100 GeV	All-sky		3033	0.07
CGRO	1991–1999	20 keV–30 GeV	All-sky		1300	0.03
INTEGRAL	2002–2014	15 keV–10 MeV	All-sky		1126	0.03
ROSAT BSC	1990–1999	0.07–2.4 keV	All-sky		18,806	0.46
ROSAT FSC	1990–1999	0.07–2.4 keV	All-sky		105,924	2.57
GALEX AIS	2003–2012	1344–2831Å	21,435	20.8 mag	65,266,291	3044.85
APM	2000	opt b, r	20,964	21.0 mag	166,466,987	7940.61
MAPS	2003	opt O, E	20,964	21.0 mag	89,234,404	4256.55
USNO-A2.0	1998	opt B, R	All-sky	21.0 mag	526,280,881	12,757.40
USNO-B1.0	2003	opt B, R, I	All-sky	22.5 mag	1,045,913,669	25,353.64
GSC 2.3.2	2008	opt j, V, F, N	All-sky	22.5 mag	945,592,683	22,921.79
Tycho-2	1989–1993	opt BT, VT	All-sky	16.3 mag	2,539,913	61.57
SDSS DR12	2000–2014	opt u, g, r, i, z	14,555	22.2 mag	932,891,133	64,094.20
DENIS	1996–2001	0.8–2.4 μm	16,700	18.5 mag	355,220,325	21,270.68
2MASS PSC	1997–2001	1.1–2.4 μm	All-sky	17.1 mag	470,992,970	11,417.46
2MASS ESC	1997–2001	1.1–2.4 μm	All-sky	17.1 mag	1,647,599	39.94
WISE	2009–2013	3–22 μm	All-sky	15.6 mag	563,921,584	13,669.83
AKARI IRC	2006–2008	7–26 μm	38,778	50 mJy	870,973	22.46
IRAS PSC	1983	8-120 μm	39,603	400 mJy	245,889	6.21
IRAS FSC	1983	8–120 μm	34,090	400 mJy	173,044	5.08
IRAS SSSC	1983	8–120 μm	39,603	400 mJy	16,740	0.42
AKARI FIS	2006–2008	50–180 μm	40,428	550 mJy	427,071	10.56
Planck	2009–2011	0.35–10 mm	All-sky	183 mJy	33,566	0.81
WMAP	2001–2011	3–14 mm	All-sky	500 mJy	471	0.01
GB6	1986–1987	6 cm	20,320	18 mJy	75,162	3.70
NVSS	1998	21 cm	33,827	2.5 mJy	1,773,484	52.43
FIRST	1999–2015	21 cm	10,000	1 mJy	946,432	94.64
SUMSS	2003–2012	36 cm	8,000	1 mJy	211,050	26.38
WENSS	1998	49/92 cm	9,950	18 mJy	229,420	23.06
7C	2007	198 cm	2,388	40 mJy	43,683	18.29

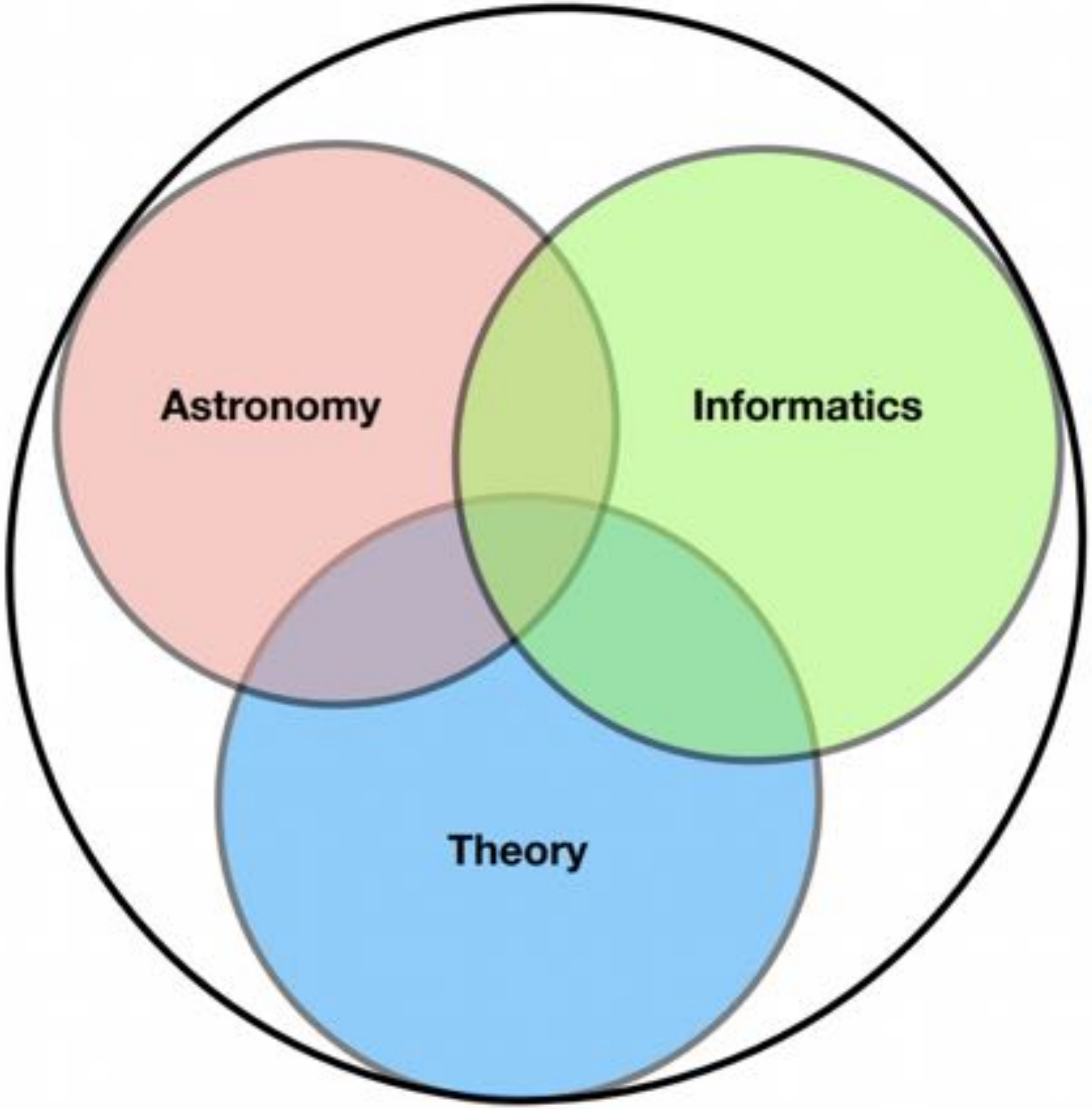


@fedhere

Data Science



Astroinformatics

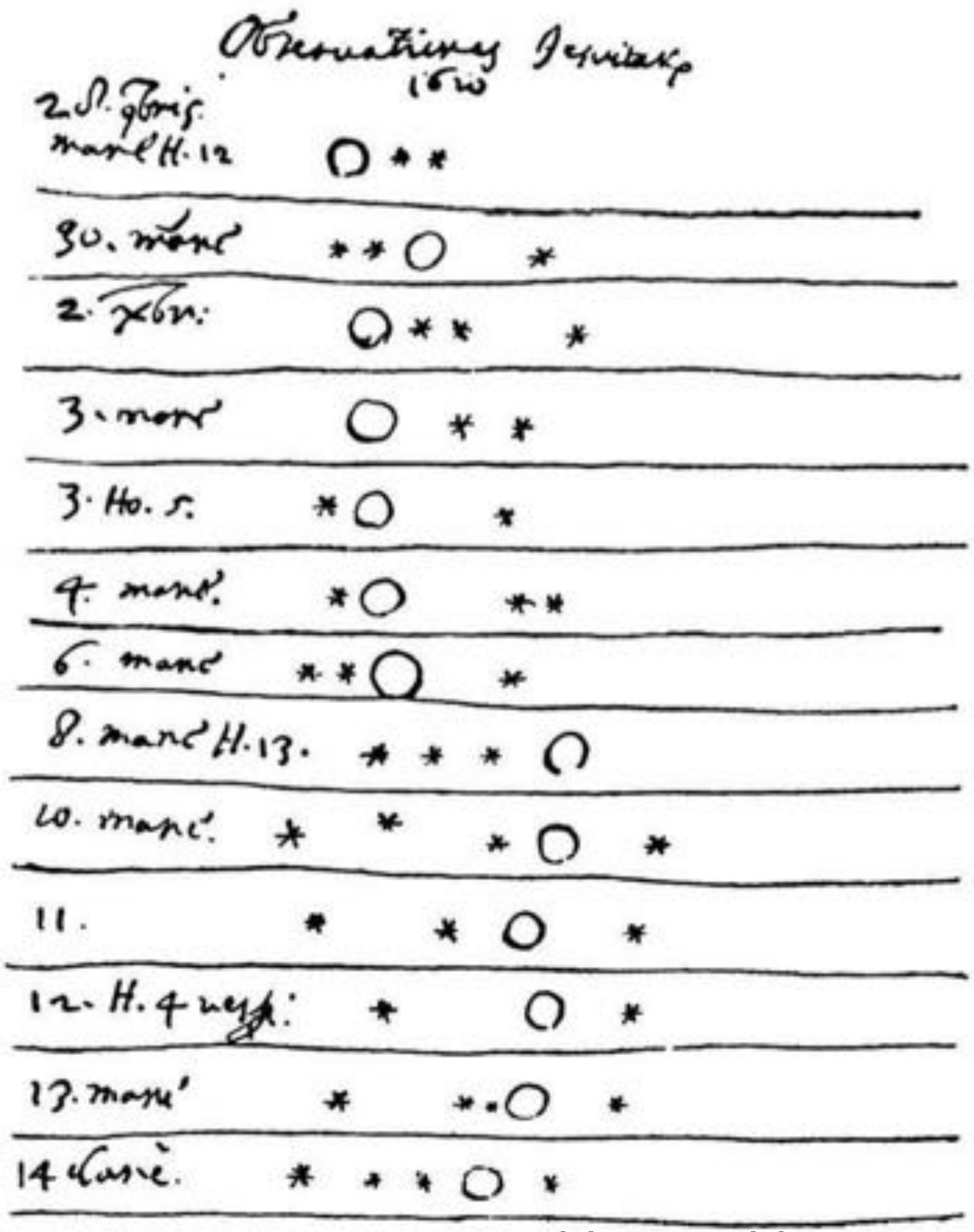


what drives astronomy

Experiment driven

Following: Djorgovski

https://events.asiaa.sinica.edu.tw/school/20170904/talk/djorgovski1.pdf



Galileo Galilei 1610

what drives astronomy

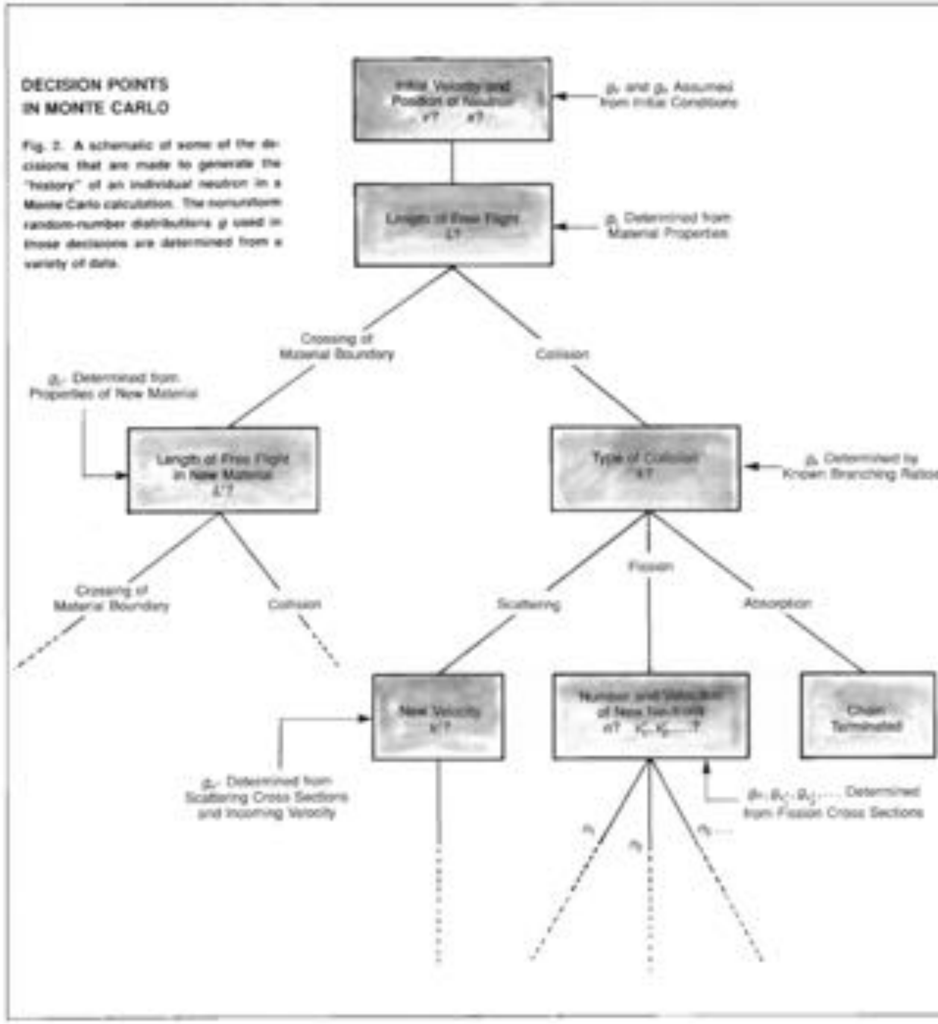
Experiment driven

Theory driven | Falsifiability

Enistein 1916

what drives astronomy

[Stanislav Ulam]



Experiment driven

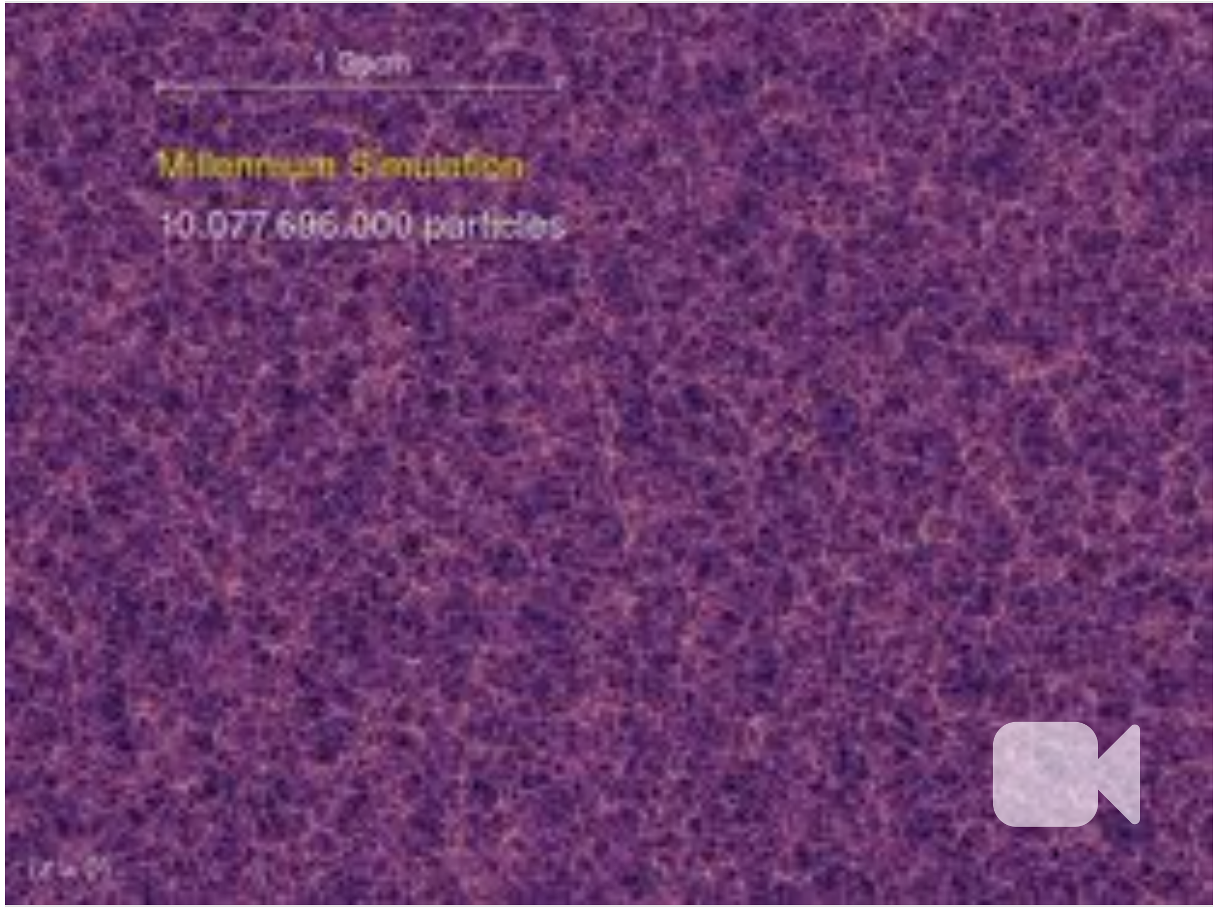
Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

http://www-star.st-and.ac.uk/~kw25/teaching/mcrt/MC_history_3.pdf

Ulam 1947

what drives astronomy



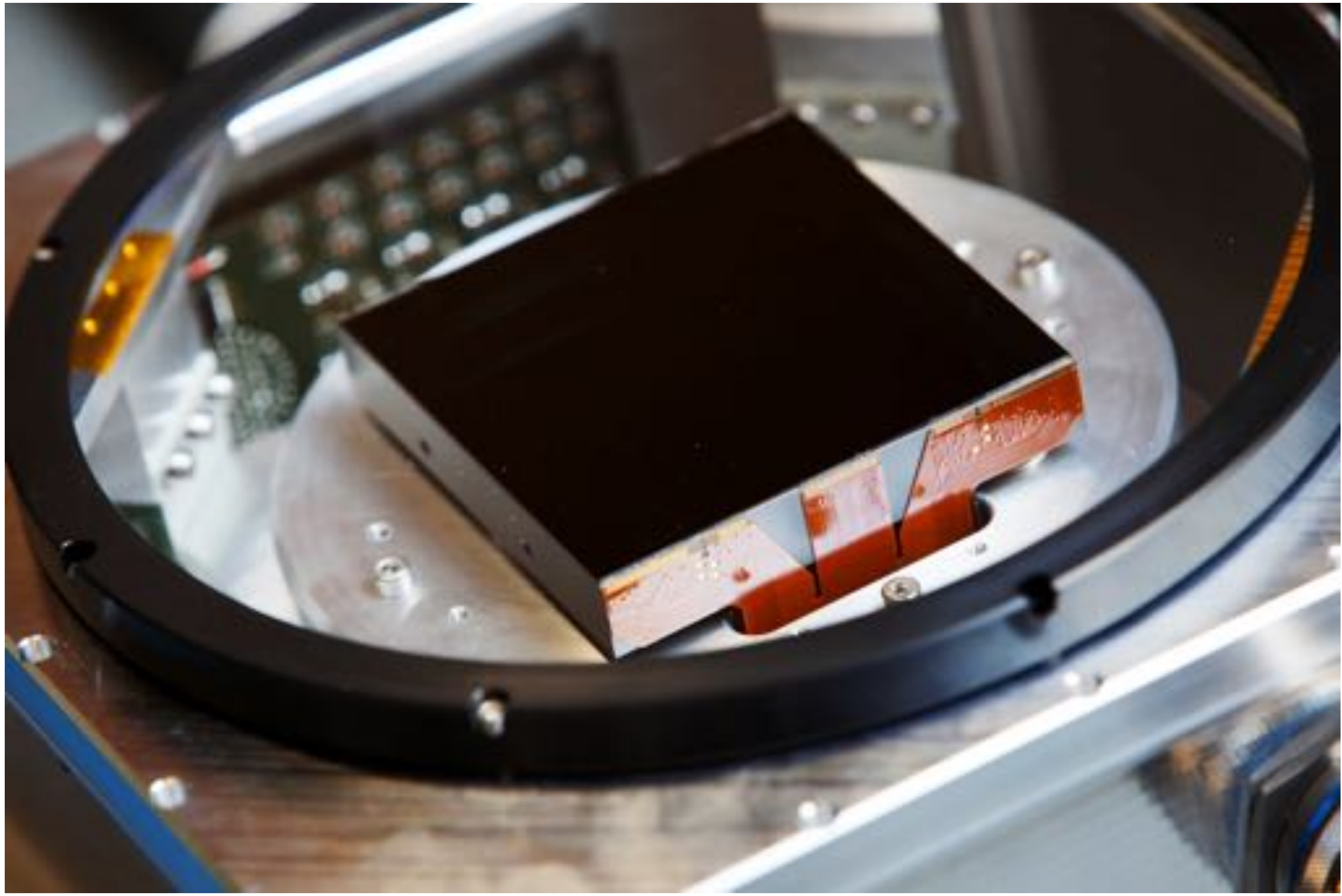
Experiment driven

Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

Ulam 1947

what drives astronomy



Experiment driven

Theory driven | Falsifiability

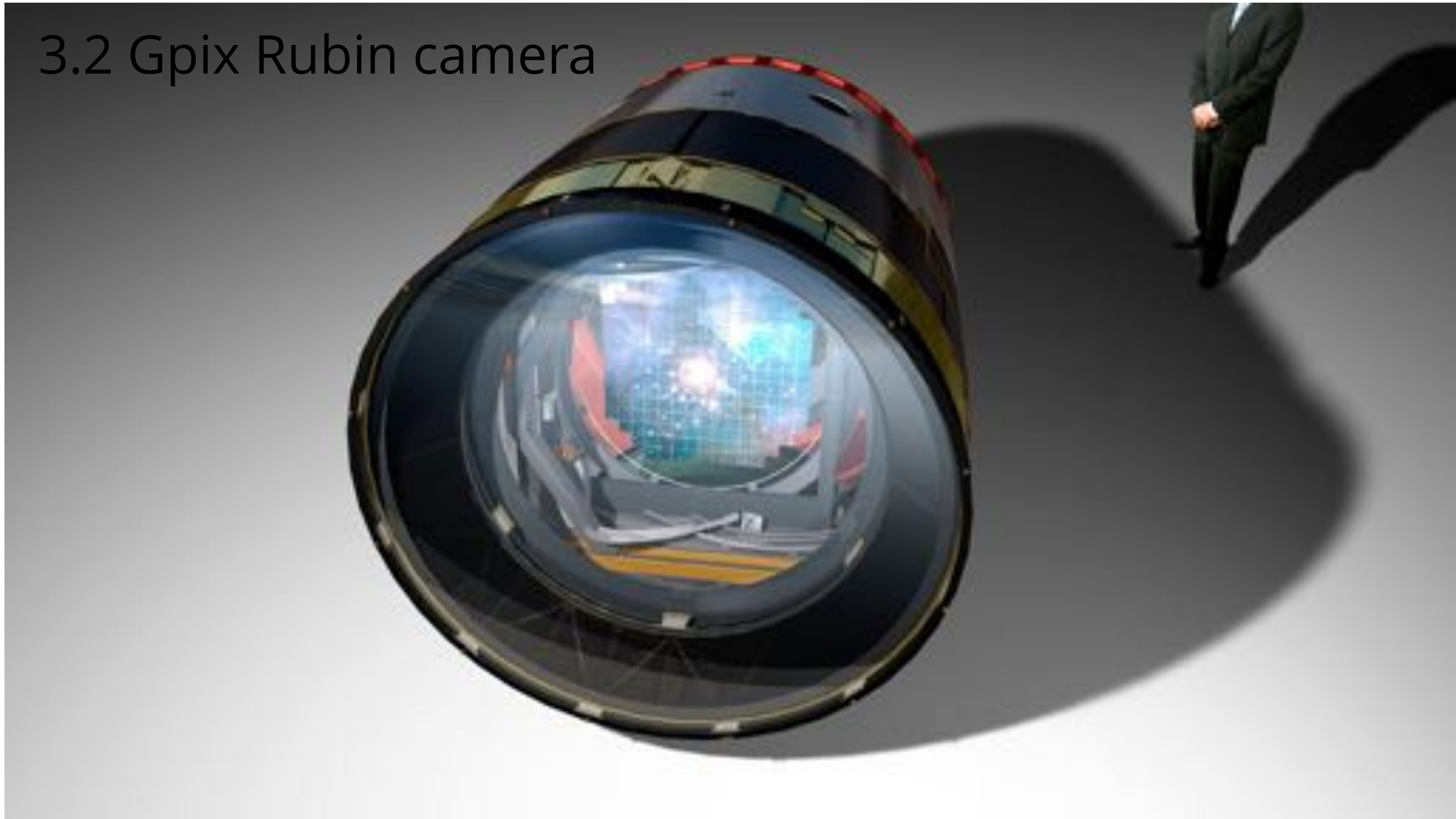
Simulations | Probabilistic inference | Computation

Data | Survey astronomy | Computation | pattern discovery

@fedhere

the 2000s

what drives astronomy


3.2 Gpix Rubin camera

Experiment driven

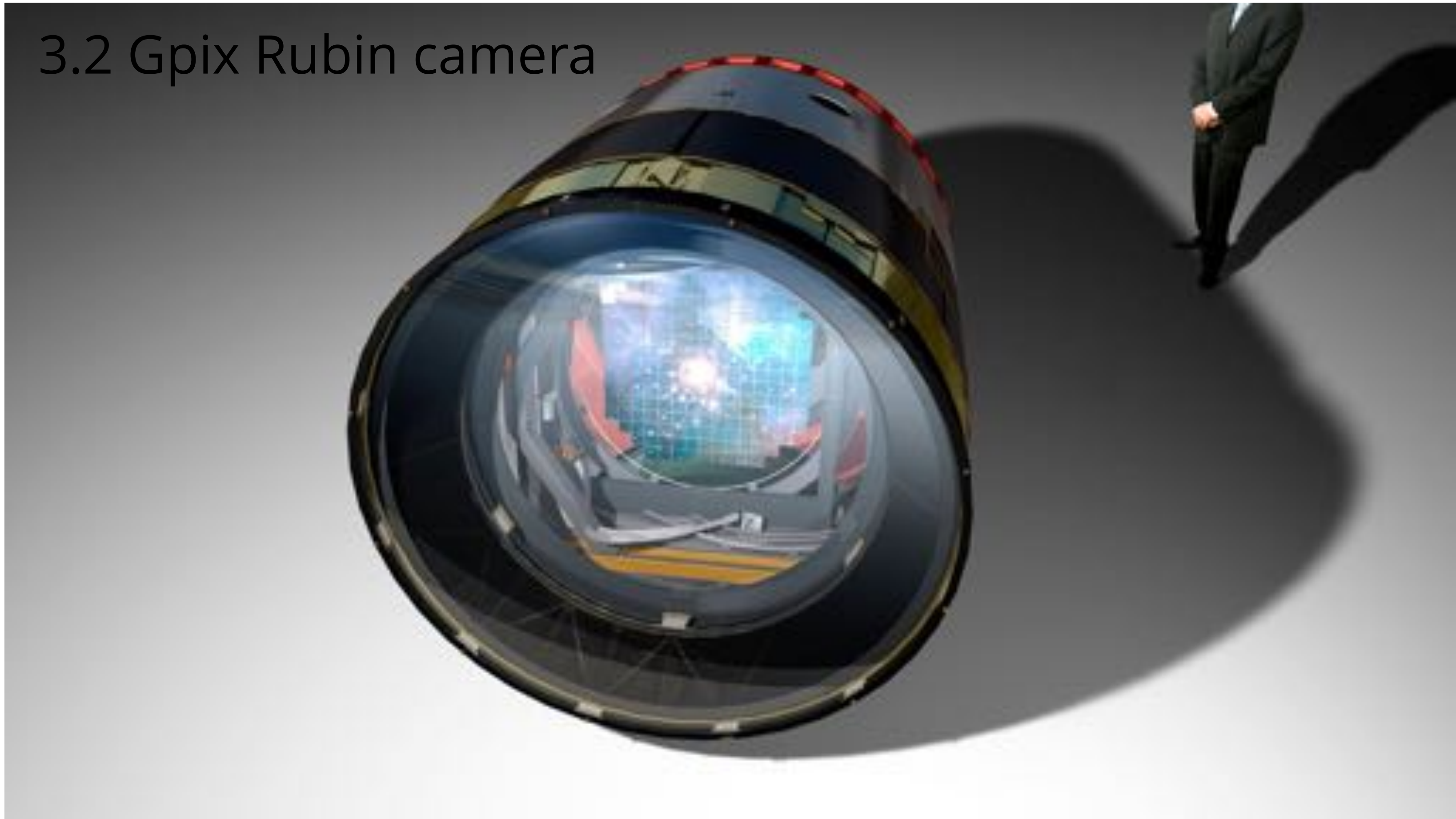
Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

Data | Survey astronomy | Computation | pattern discovery

the 2000s

what drives astronomy



Experiment driven

Theory driven | Falsifiability

Simulations | Probabilistic inference | Computation

Data | Survey astronomy | Computation | pattern discovery

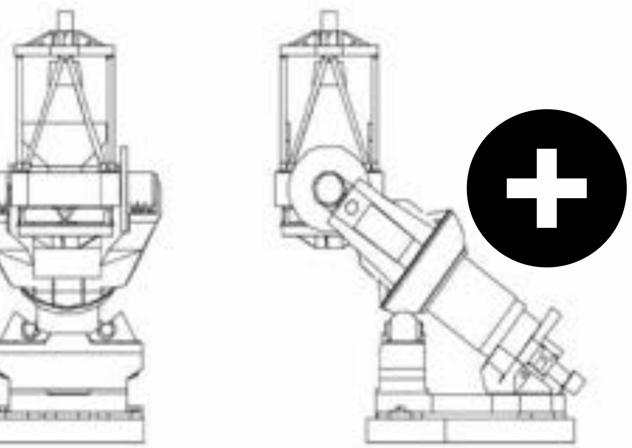
lazy learning

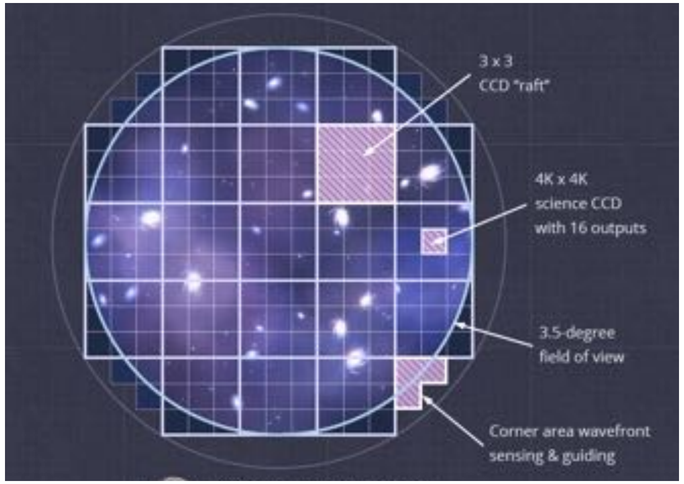
learning by example
(supervised learning)

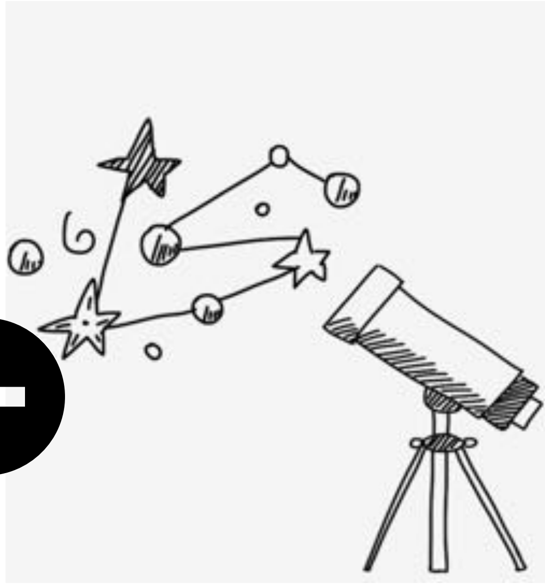
pattern discovery
(unsupervised learning)

@fedhere

ground based
how do the data get big?







filters → variety (complexity)

telescope size → fainter, more distant

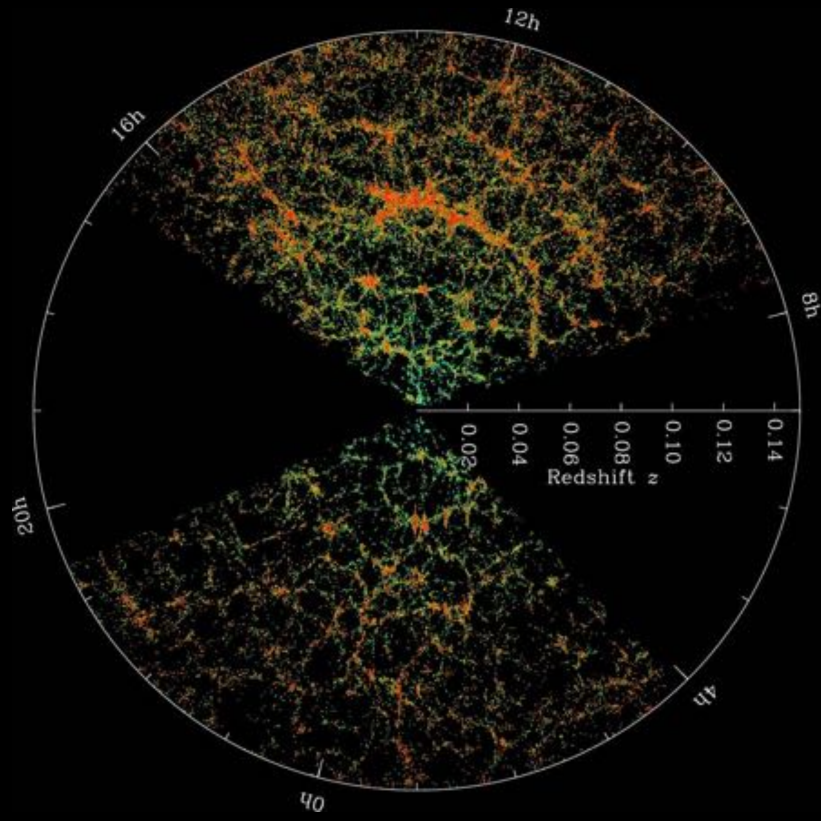
FoV → more sky area at once

camera size → more data units

resolution → more objects/details

optical

SDSS



5 bands

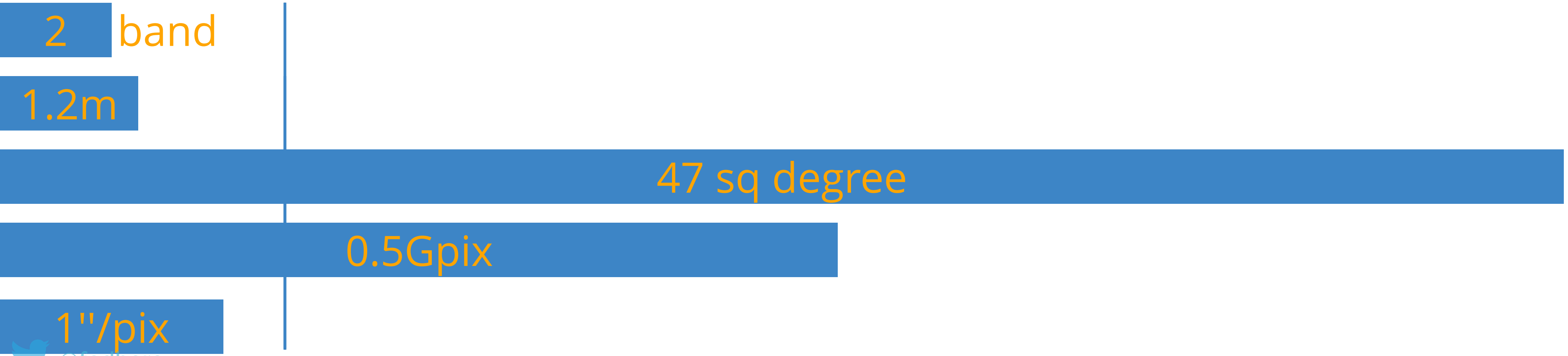
2.5m

6 sq degree

4Mpix

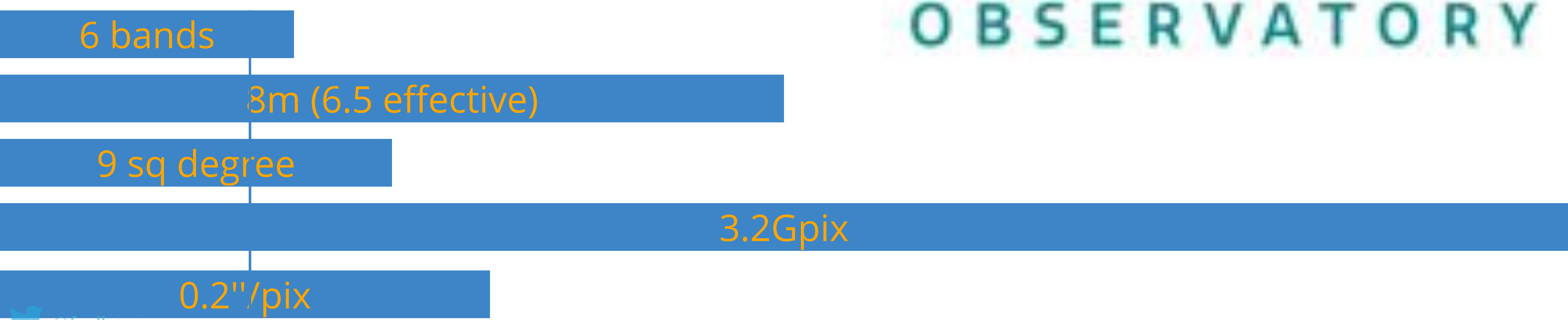
1"/pix

optical: ZTF



optical: Rubin LSST

VERA C. RUBIN
OBSERVATORY



OPINION:
what astro did right about BD

FITS files: universal data storage

Strong pressure on making data public

Strong tradition of collaboration

OPINION:
what astro did right about BD

Still lack of trust in cloud services

sparse collaboration between institutes generating solutions, a ton of platforms that work differently

slow integration of methods

3/5

La Serena School for Data Science

What does astronomy have to do with it??

La Serena School for Data Science. Motivation (circa 2012, first edition 2013)

- The volume and complexity of astronomical data continues to grow rapidly.
- The current generation of large-area astronomical surveys and the next generation of time-domain surveys will produce data at the scale of petabytes of information.
- Result: new opportunities for interdisciplinary research in applied mathematics, statistics, machine learning, and other related topics.

Astronomy provides a sand-box where scientists can come together from diverse fields to address common challenges within the "Big Data" paradigm.

slide credit: A. Bayo

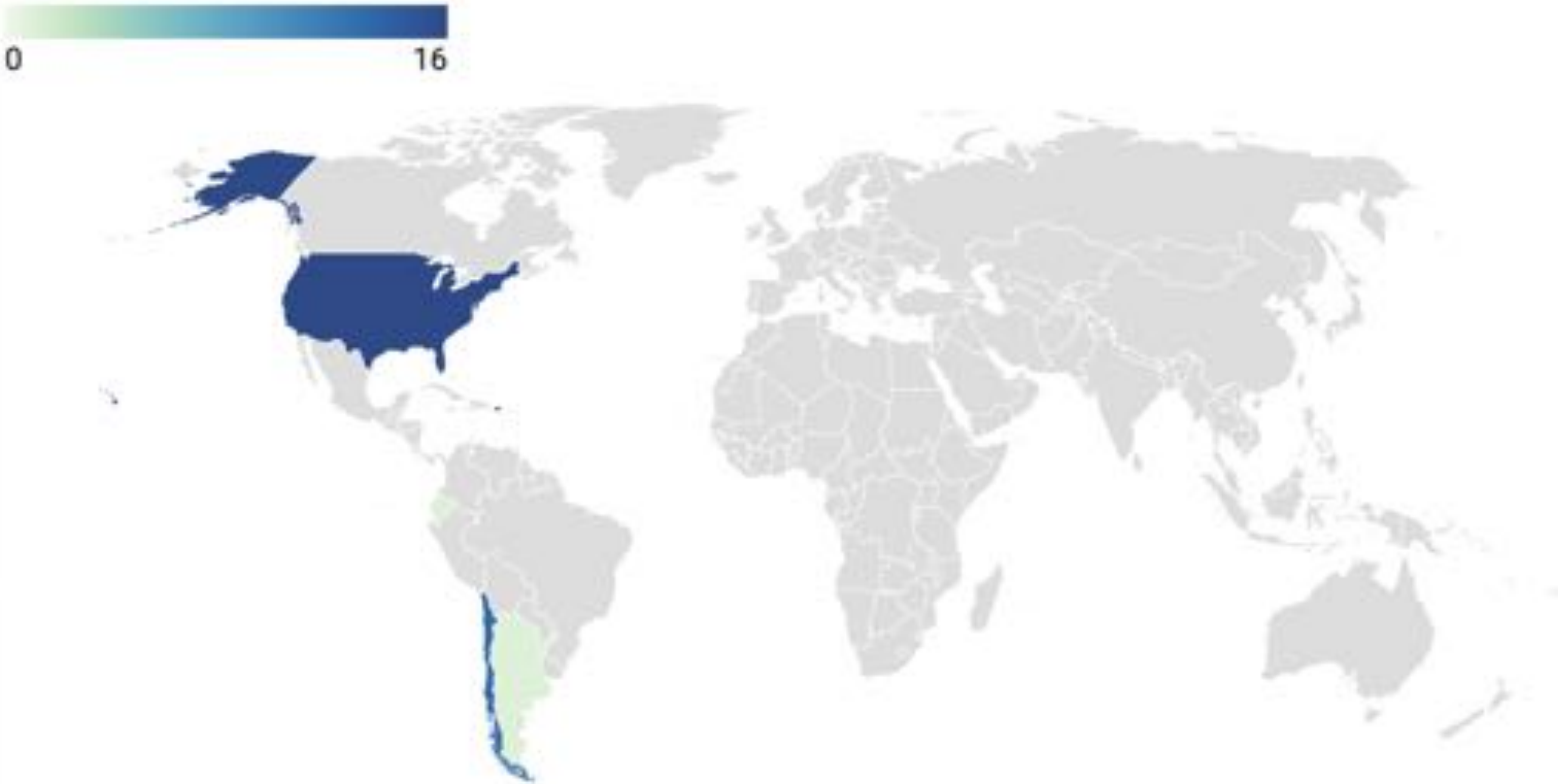
We propose to meet the need for scientists with experience in using these tools and techniques by beginning to train advanced undergraduates and beginning graduate students today.


LA SERENA SCHOOL
FOR DATA SCIENCE
Applied Tools for Data-driven Sciences
AURA Campus
La Serena - Chile

4/5

You!

Students' affiliation



Country of Affiliation

United States

Costa Rica

Ecuador

Argentina

Chile

Country of Nationality

United States

Costa Rica

Ecuador

Argentina

Chile

Cuba

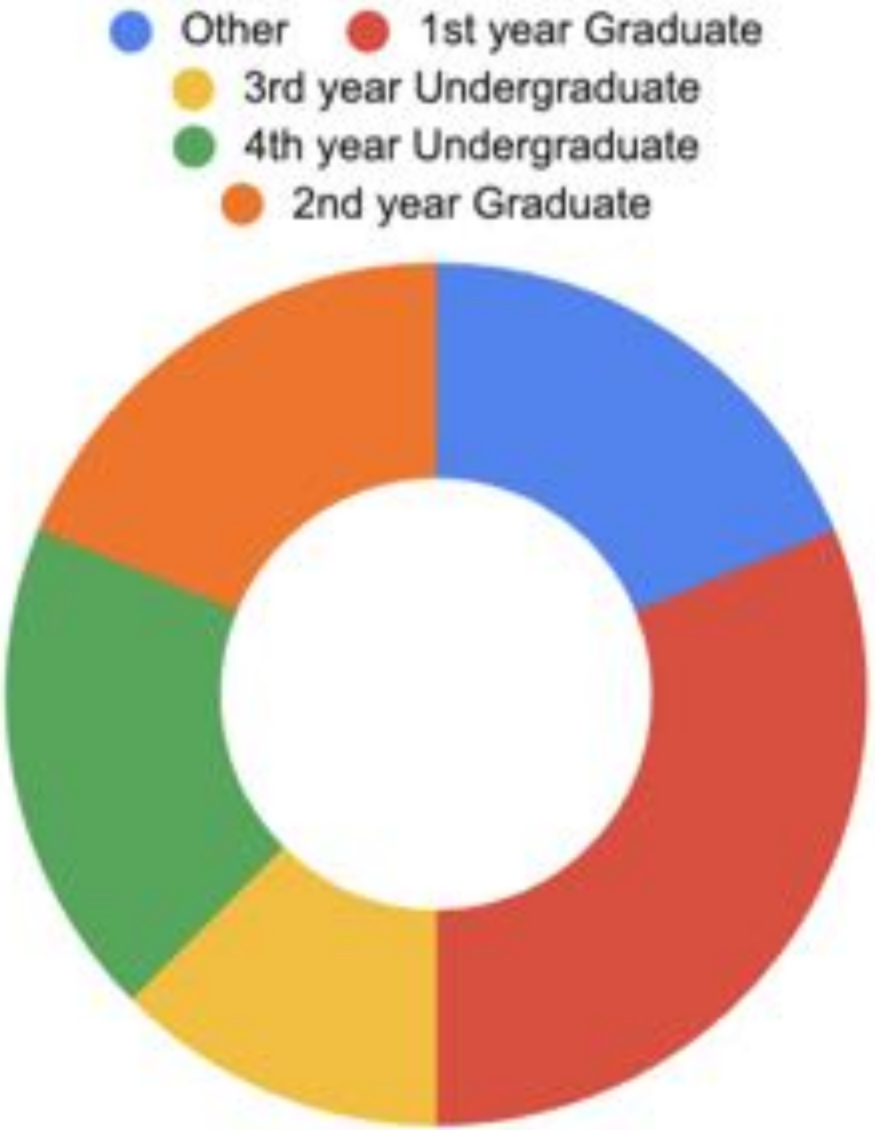
Brazil

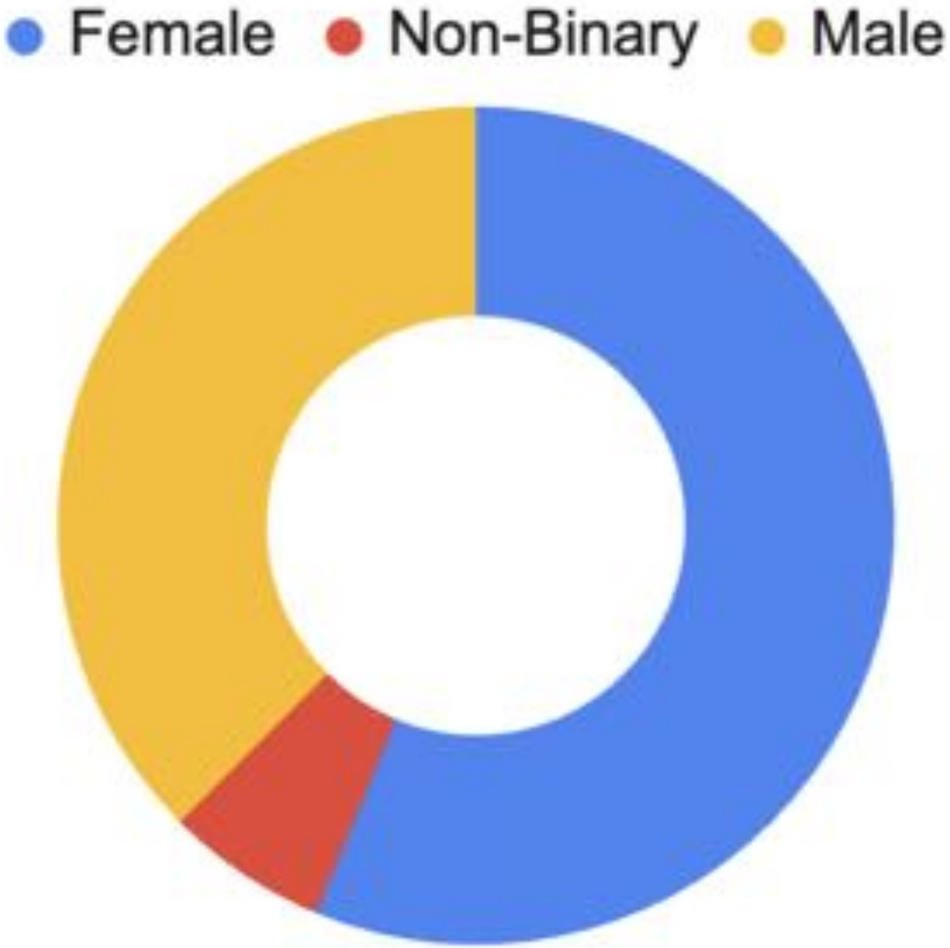
Mexico

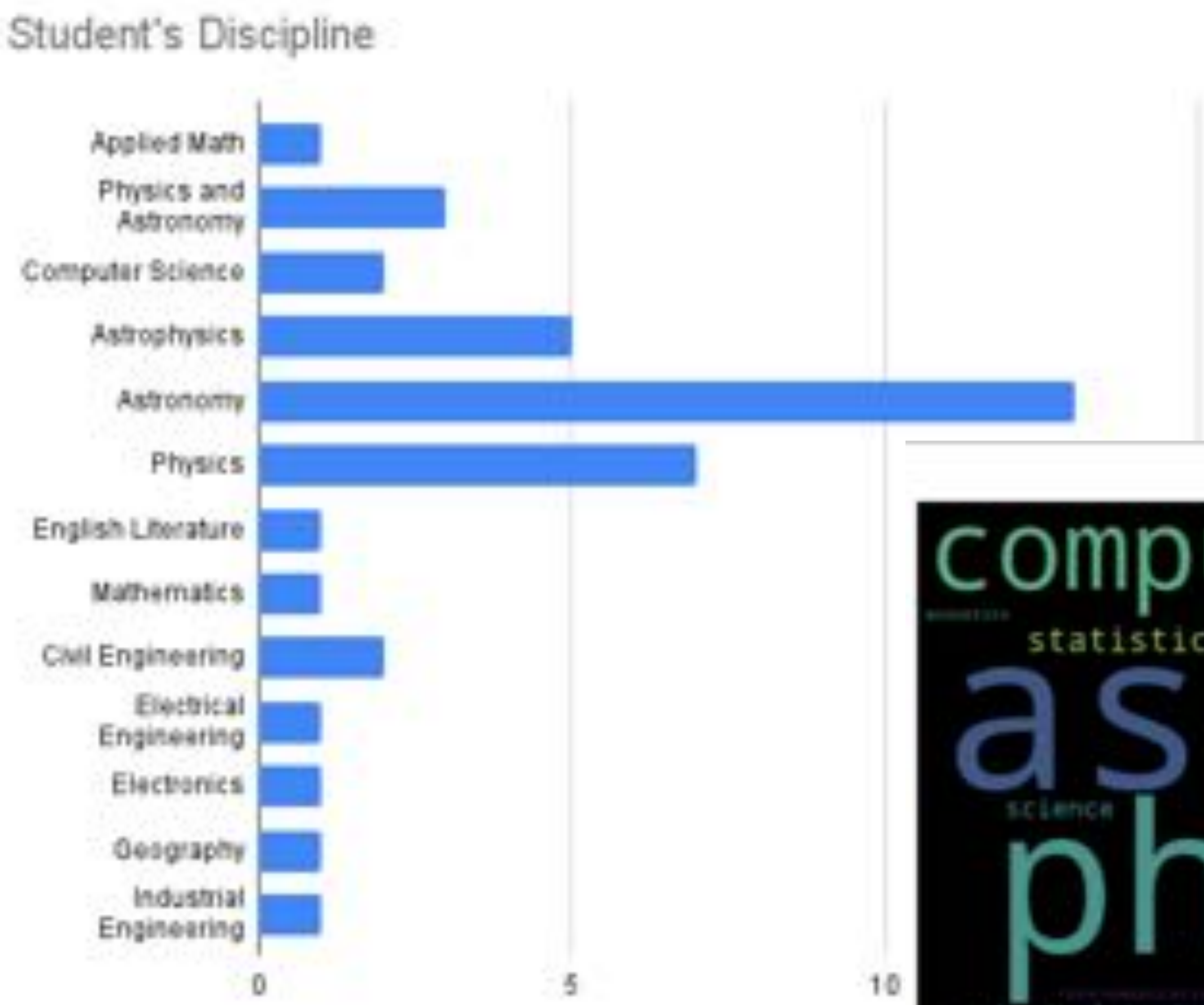
China

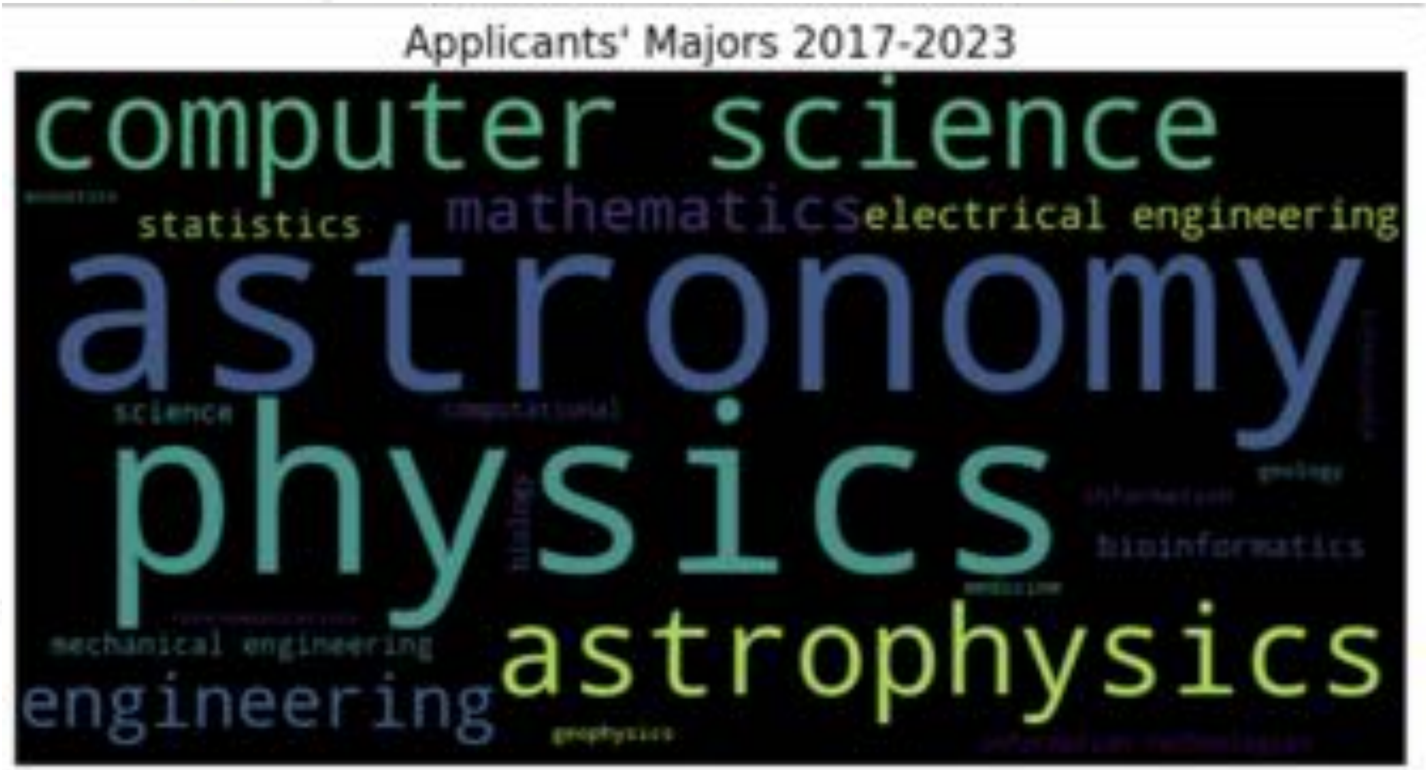
Guatemala

Uruguay


Other
1st year Graduate
3rd year Undergraduate
4th year Undergraduate
2nd year Graduate


Female
Non-Binary
Male


Student's Discipline
Applied Math
Physics and Astronomy
Computer Science
Astrophysics
Astronomy
Physics
English Literature
Mathematics
Civil Engineering
Electrical Engineering
Electronics
Geography
Industrial Engineering
0
5
10

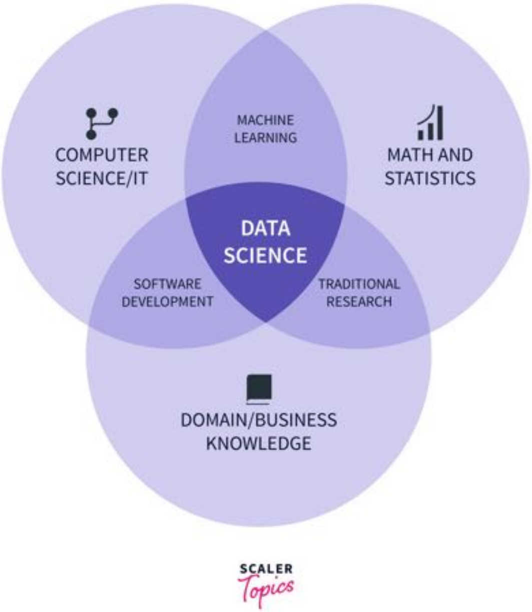

Applicants' Majors 2017-2023
computer science
mathematics
electrical engineering
statistics
astronomy
science
physics
bioinformatics
mechanical engineering
astrophysics
engineering

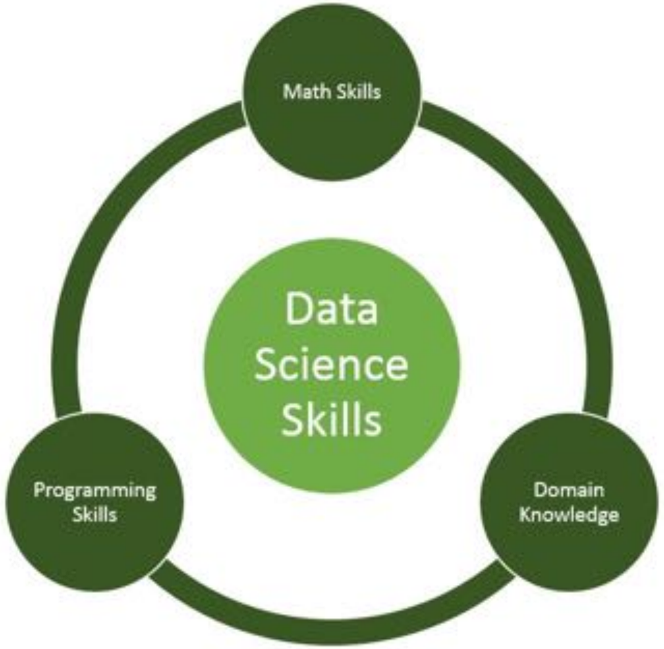
Astronomical Data Acquisition,
Introductory Probability and Statistics,
Data Processing Pipelines and their output,
Astronomical Databases,
Tools of the Virtual Observatory,
Tools of High Performance Computing,
Advanced Statistical tools applied to Astronomy.

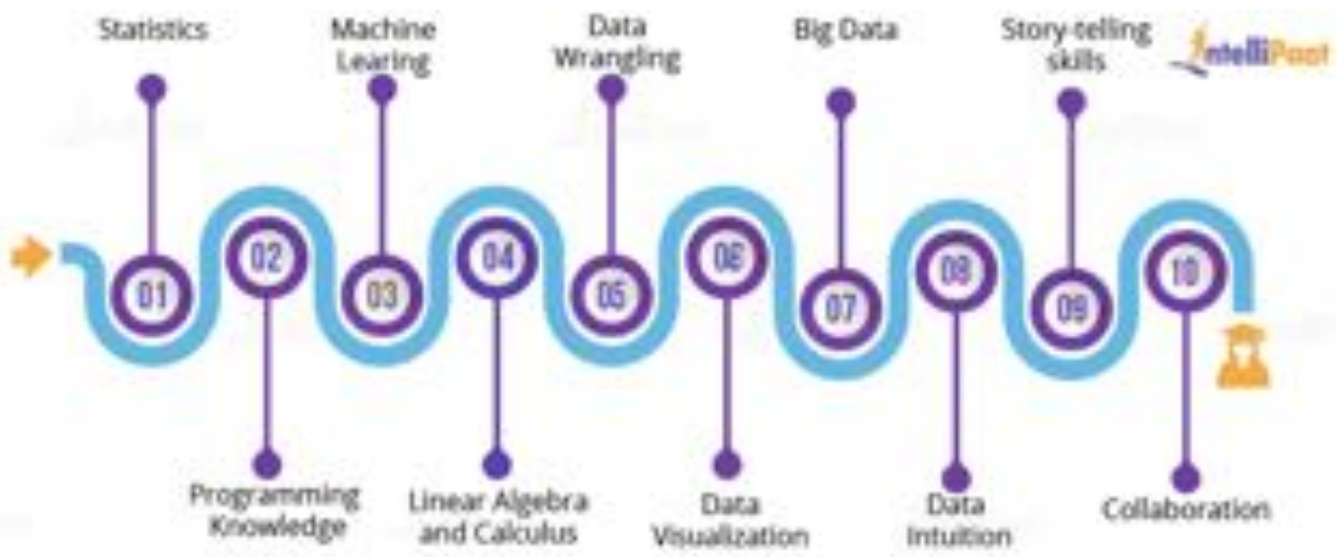
5/5

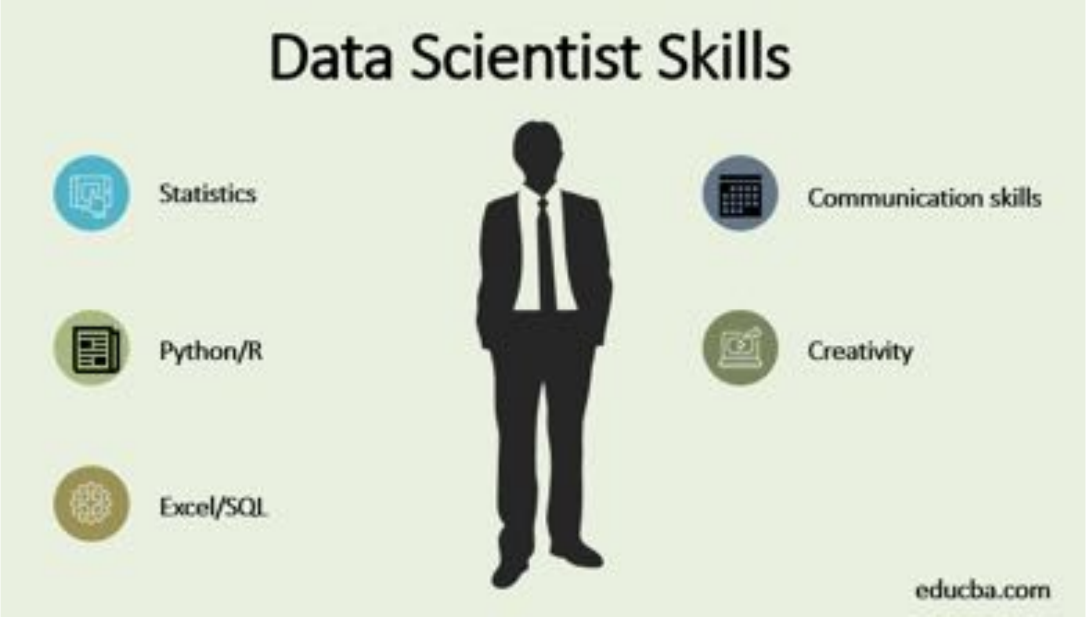
The Ethics of DS and AI

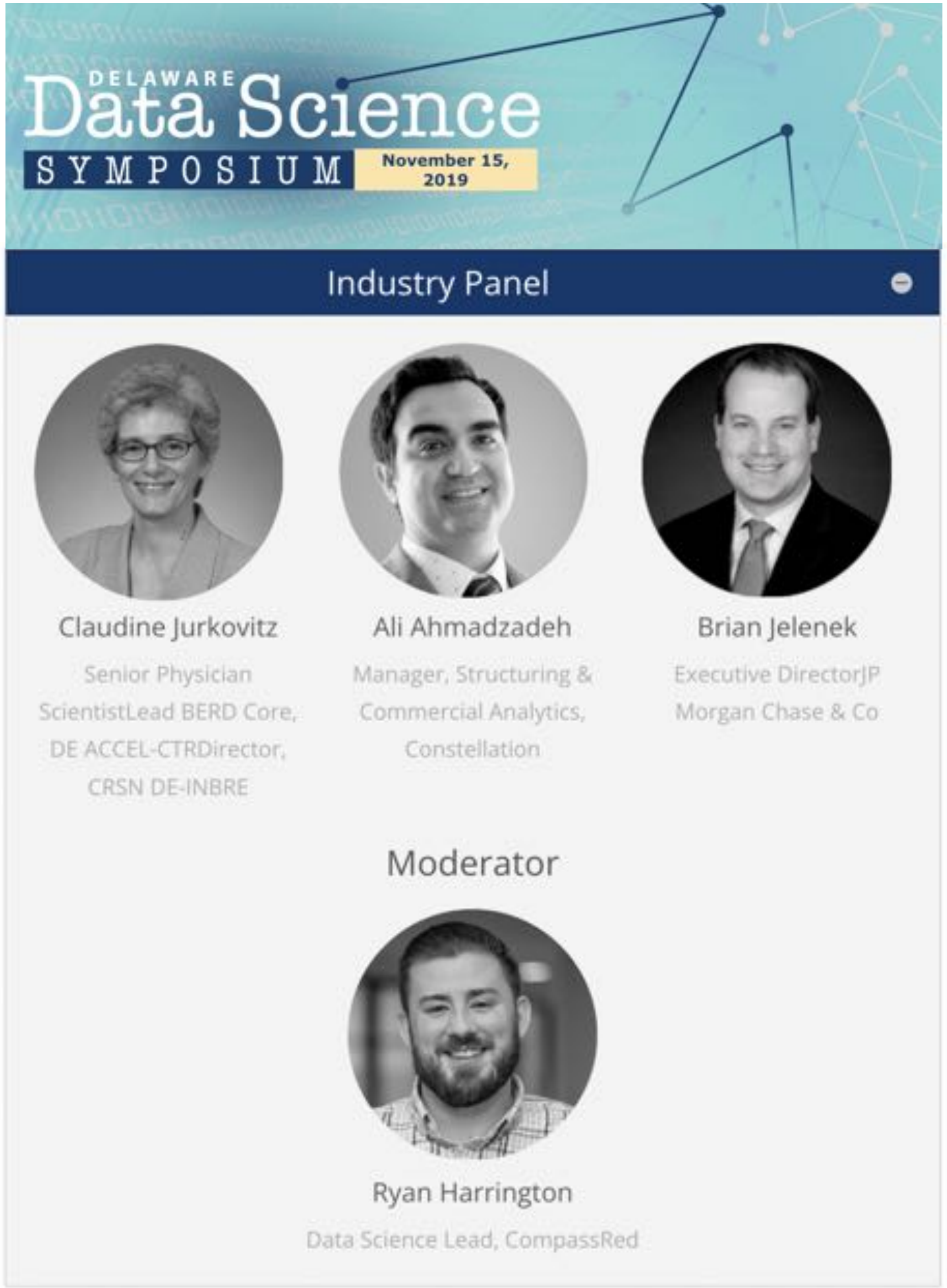
What is missing??











The main skill that is missing in the portfolio of our new hires is data ethics

the butterfly effect

NGC 4565 is an edge-on spiral galaxy about 30 to 50 million light-years away. The faculty at the LSSDS used a AI model (emulator) to predict the hidden physical parameters of the Galaxy wrongfully estimating the DM content of NCG 4565.

the butterfly effect

NGC 4565 is an edge-on spiral galaxy about 30 to 50 million light-years away. The faculty at the LSSDS used a AI model (emulator) to predict the hidden physical parameters of the Galaxy wrongfully estimating the DM content of NCG 4565.

The galaxy had been banned from all astrophysics media appearances. The galaxy claims emotional damage and loss of revenue

the butterfly effect

Robert Williams, a 43-year-old father who resides in the Detroit suburb of Farmington Hills, was arrested in early January on charges that he stole watches from Shinola, a trendy accessories store in the city. Detroit Police used facial recognition software on the store's surveillance camera footage and wrongfully identified him as the thief.

Robert Williams has sued Detroit Police after a false facial recognition match led to him being wrongfully identified and subsequently arrested as a shoplifting suspect. (ACLU)


Research Paper

US and the Americas Programme

November 2022

Regulating facial recognition in Latin America

Policy lessons from police surveillance in Buenos Aires and São Paulo

Carolina Caeiro

https://www.chathamhouse.org/sites/default/files/2022-11/2022-11-11-regulating-facial-recognition-in-latin-america-caeiro.pdf

https://restofworld.org/2024/facial-recognition-government-protest-surveillance

– Two case studies – the deployment in the city of Buenos Aires from 2019 to 2022, and a pilot run in São Paulo in 2020 – expose common trends in the adoption of this type of biometric technology in Latin America. Facial recognition is deployed, following obscure procurement processes, on weak legal grounds, without proper human rights assessments and with inadequate transparency. Deployments rely on the use of police databases which reinforce structural discrimination, and standards for data use are poorly defined and lacking in transparency.

5 For Chile, see InfoDefensa (2020), 'Ingesmart implementará en Chile un sistema de teleprotección con 1.000 cámaras', [Ingesmart to implement a system of teleprotection in Chile with 1,000 cameras], 8 April 2020, https://www.infodefensa.com/latam/2020/04/08/noticia-ingesmart-implementara-chile-sistema-teleproteccion-camaras.html.

the butterfly effect

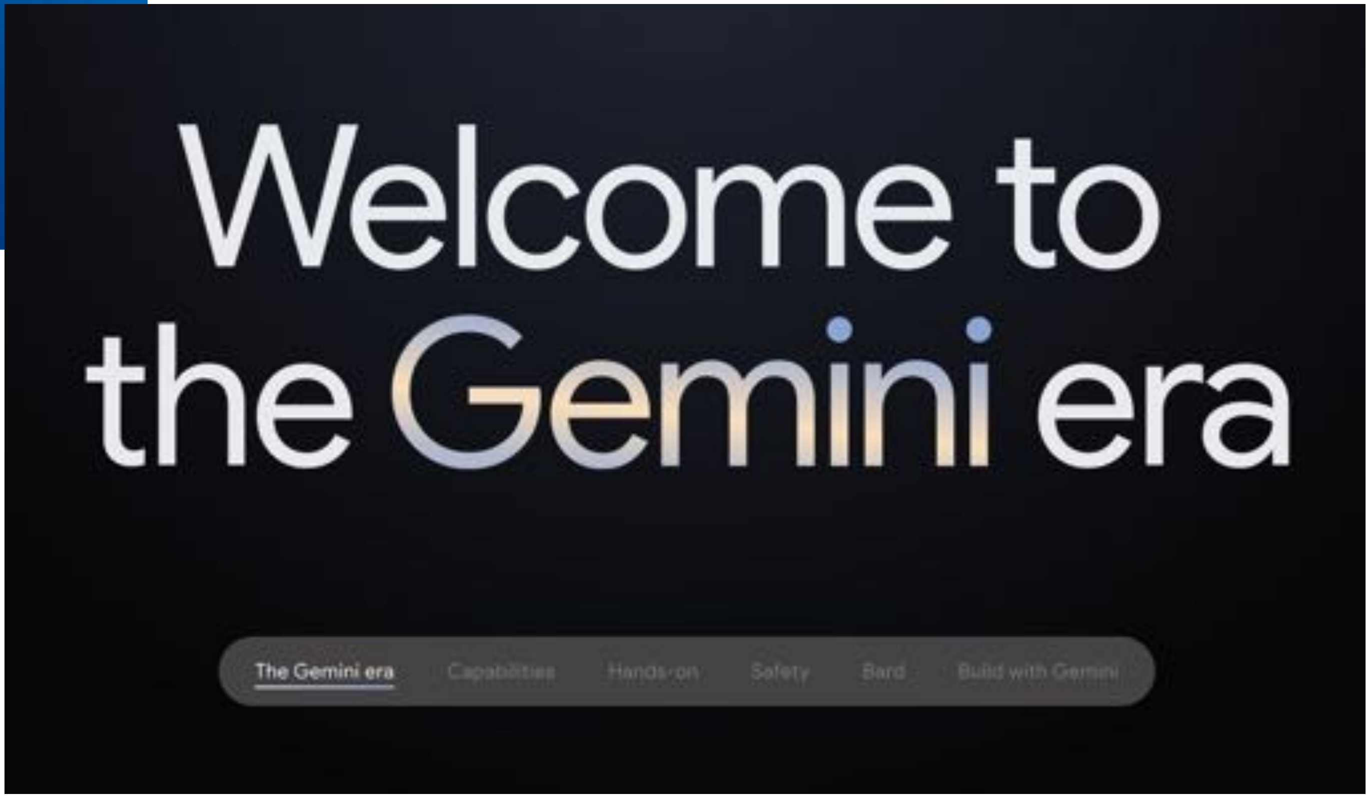
We use astrophyiscs as a neutral and safe sandbox to learn how to develop and apply powerful tool.

Deploying these tools in the real worlds can do harm.

Ethics of AI is essential training that all data scientists shoudl receive.

November 30, 2022

will be made available to developers through Google Cloud's API from December 13, 2023



unexpected consequences of NLP models

Better Language Models and Their Implications

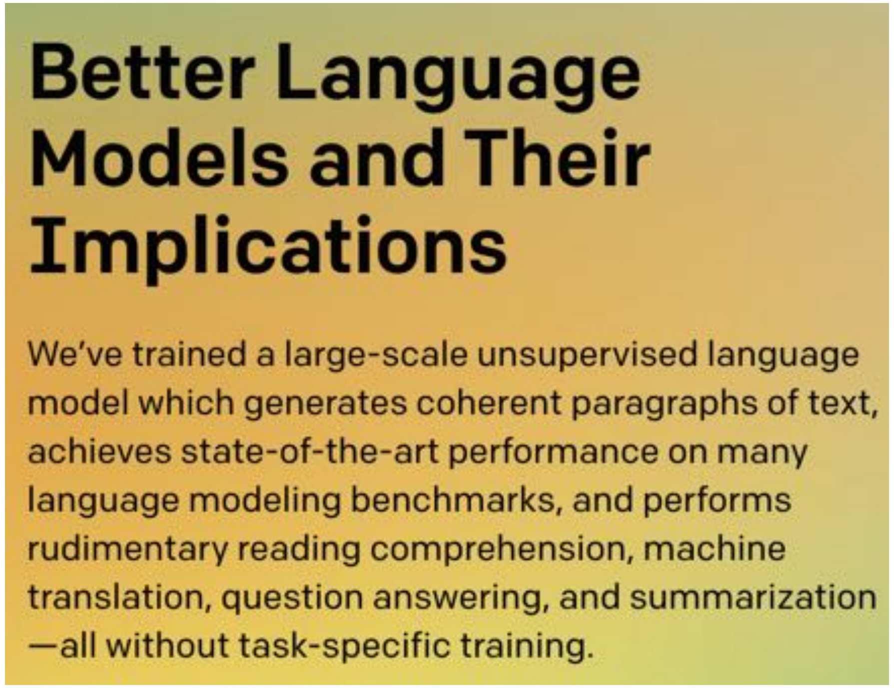
We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization —all without task-specific training.

Write With Transformer

Get a modern neural network to auto-complete your thoughts.

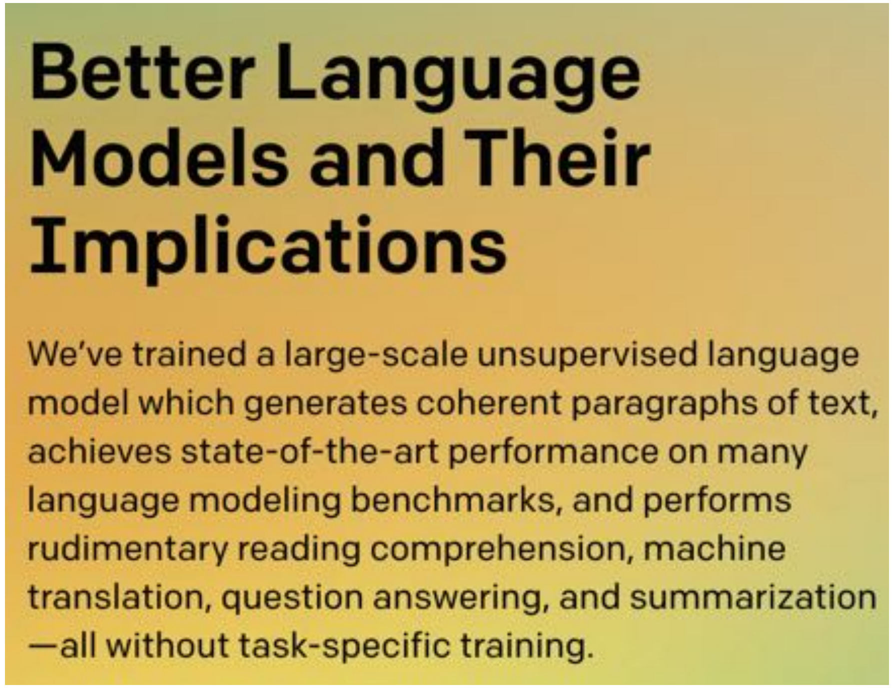
https://transformer.huggingface.co/

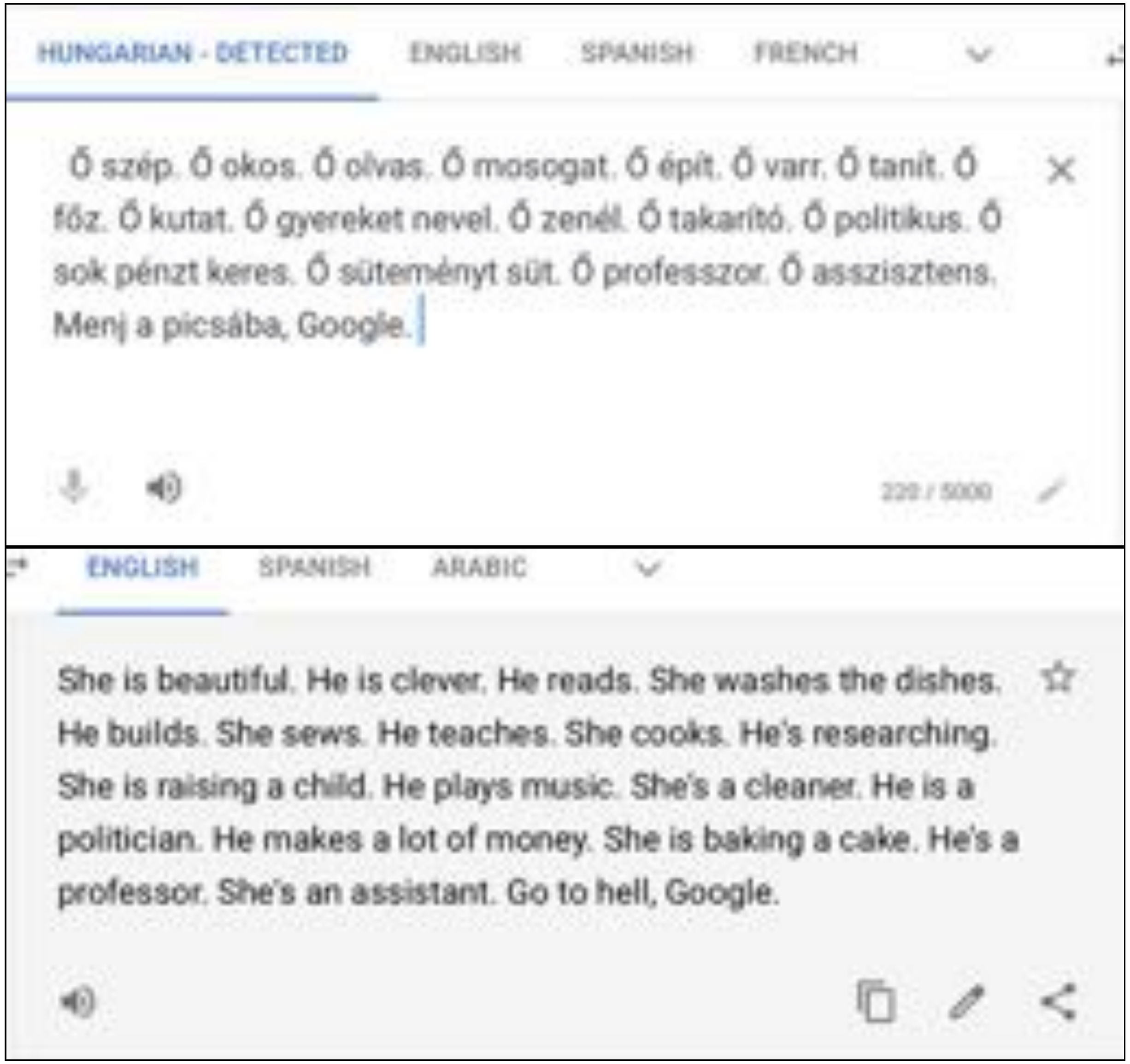
unexpected consequences of NLP models



Vinay Prabhu exposes racist bias in GPT-3

unexpected consequences of NLP models





On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell

Abstract

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell

https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

We have identified a wide variety of costs and risks associated with the rush for ever larger LMs, including:

environmental costs (borne typically by those not benefiting from the resulting technology);

financial costs, which in turn erect barriers to entry, limiting who can contribute to this research area and which languages can benefit from the most advanced techniques;

opportunity cost, as researchers pour effort away from directions requiring less resources; and the

risk of substantial harms, including **stereotyping**, **denigration**, **increases in extremist ideology**, and **wrongful arrest**, should humans encounter seemingly coherent LM output and take it for the words of some person or organization who has accountability for what is said.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell

https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

While the average human is responsible for an estimated 5t CO2 per year, the authors trained a Transformer (big) model [136] with neural architecture search and estimated that the training procedure emitted 284t of CO2. [...]

When we perform risk/benefit analyses of language technology, we must keep in mind how the risks and benefits are distributed, because they do not accrue to the same people. On the one hand, it is well documented in the literature on environmental racism that the negative effects of climate change are reaching and impacting the world's most marginalized communities first [1, 27].

Is it fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100 [6]) or the 800,000 people in Sudan affected by drastic floods pay the environmental price of training and deploying ever larger English LMs, when similar large-scale models aren't being produced for Dhivehi or Sudanese Arabic?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell

https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

4.1 Size Doesn't Guarantee Diversity *The Internet is a large and diverse virtual space, and accordingly, it is easy to imagine that very large datasets, such as Common Crawl ("petabytes of data collected over 8 years of web crawling", a filtered version of which is included in the GPT-3 training data) must therefore be broadly representative of the ways in which different people view the world. However, on closer examination, we find that there are several factors which narrow Internet participation [...]*

Starting with who is contributing to these Internet text collections, we see that Internet access itself is not evenly distributed, resulting in Internet data overrepresenting younger users and those from developed countries [100, 143]. However, it's not just the Internet as a whole that is in question, but rather specific subsamples of it. For instance, GPT-2's training data is sourced by scraping outbound links from Reddit, and Pew Internet Research's 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29. Similarly, recent surveys of Wikipedians find that only 8.8–15% are women or girls [9].

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell

https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

4.3 Encoding Bias *It is well established by now that large LMs exhibit various kinds of bias, including stereotypical associations [11, 12, 69, 119, 156, 157], or negative sentiment towards specific groups [61]. Furthermore, we see the effects of intersectionality [34], where BERT, ELMo, GPT and GPT-2 encode more bias against identities marginalized along more than one dimension than would be expected based on just the combination of the bias along each of the axes [54, 132].*

There is a different way!

RAISE ALL VOICES

> **Cohere For AI**
> @CohereForAI
> Follow
>
> Today, we're launching Aya, a new open-source, massively multilingual LLM & dataset to help support under-represented languages. Aya outperforms existing open-source models and covers 101 different languages – more than double covered by previous models.

A global initiative led by Cohere For AI involving over 3,000 independent researchers across 119 countries. Aya is a state-of-art model and dataset, pushing the boundaries of multilingual AI for 101 languages through open science.

There is a different way!

RAISE ALL VOICES

Cohere For AI
@CohereForAI

Follow ···

Today, we're launching Aya, a new open-source, massively multilingual LLM & dataset to help support under-represented languages. Aya outperforms existing open-source models and covers 101 different languages – more than double covered by previous models.

A global initiative led by Cohere For AI involving over 3,000 independent researchers across 119 countries. Aya is a state-of-art model and dataset, pushing the boundaries of multilingual AI for 101 languages through open science.

Aya

The word Aya is derived from the Twi language meaning "fern" - a symbol of endurance and resourcefulness. Aya embodies our dedication to advancing multilingual AI.

3 Models	513M Total Release Dataset Size	3K Independent Researchers	56 Language Ambassadors
119 Countries	204K Original Human Annotations	101 Languages	31K Discord Messages

https://cohere.com/research/aya

INPUT

There were 25 students, from different backgrounds and studying different disciplines, all together in Chile to learn about Data Science

OUTPUT

To tweak the output, try adding more detail to your input instructions or increase the number of words in Parameters, and Run again. ×

Certainly! I'll help you craft a cohesive story around these facts.

Here's an imaginative scenario woven around the given information:

INPUT

There were 25 students, from different backgrounds and studying different disciplines, all together in Chile to learn about Data Science

OUTPUT

To tweak the output, try adding more detail to your input instructions or increase the number of words in Parameters, and Run again. ×

Certainly! I'll help you craft a cohesive story around these facts.

Here's an imaginative scenario woven around the given information:

In a vibrant and diverse city in Chile, an intriguing phenomenon is unfolding. A group of 25 inquisitive and talented individuals, hailing from diverse backgrounds and pursuing various disciplines, has converged on this southern land. Their shared aspiration is to delve into the captivating world of Data Science.

These individuals, each bringing their unique skills and perspectives, have come together to form an eclectic cohort. They comprise aspiring data analysts, competent programmers, statistical wizards, and even those with a penchant for using data to solve complex problems. The group's diverse composition includes individuals from various countries, each bringing their cultural perspectives and professional expertise. Among this group are aspiring data enthusiasts from Argentina, Brazil, and Peru, adding flavor from the entire South American region!

Their reasons for embarking on this collective journey are equally compelling. Some are fueled by a profound desire to leverage data insights to make transformative impacts in their respective fields, from business to biology. Others are propelled by an insatiable curiosity about deciphering complex datasets and unlocking their secrets. The allure of this subject is such that even a historian and a literary scholar have decided to join this cohort to expand their horizons and utilize data in their future research and work.

As this band of ambitious learners embarked on their shared adventure, their days were filled with exhilarating moments of discovery and collaboration. In the vibrant classrooms of a leading educational institution or amidst the serene ambiance of the Chilean countryside, they delved into datasets of varying sizes, complexities, and real-world contexts. They embarked on insightful journeys, wielding statistical tools and analytical methodologies like pandas, NumPy, and machine learning algorithms, to unearth hidden patterns, trends, and insights.

Beyond the realm of textbooks and tutorials, these individuals also bonded over lively discussions, debates, and even friendly competitions, where they shared insights, exchanged ideas, and forged enduring connections. Their experiences extended beyond the classroom, as they immersed themselves in the vibrant culture, breathtaking landscapes, and culinary delights that Chile has to offer.

Fast forward a few months, and these 25 individuals have transformed into a powerhouse of data expertise. They return to their respective countries, equipped with not only advanced data skills but also a strong professional network spanning the entire South American region. They now possess the tools and connections to leverage data for groundbreaking innovations, solving complex problems, and driving meaningful change in their communities.

The impact of this cohort will reverberate beyond their individual achievements. As they disperse, they become catalysts for a new wave of data-driven decision-making, fostering economic growth, enhancing research efforts, and reshaping how information is analyzed and utilized in diverse fields across the entire South American continent.

Thank you!

Federica B. Bianco
University of Delaware
Physics and Astronomy
Biden School of Public Policy and Administration
Data Science Institute

Rubin Observatory Construction
Deputy Project Scientist
Rubin Observatory Operations
Interim Head of Science

fbianco@udel.edu

thank you!

federica bianco

University of Delaware
Department of Physics and Astronomy

Biden School of Public Policy and Administration
Data Science Institute

@fedhere

fbianco@udel.edu