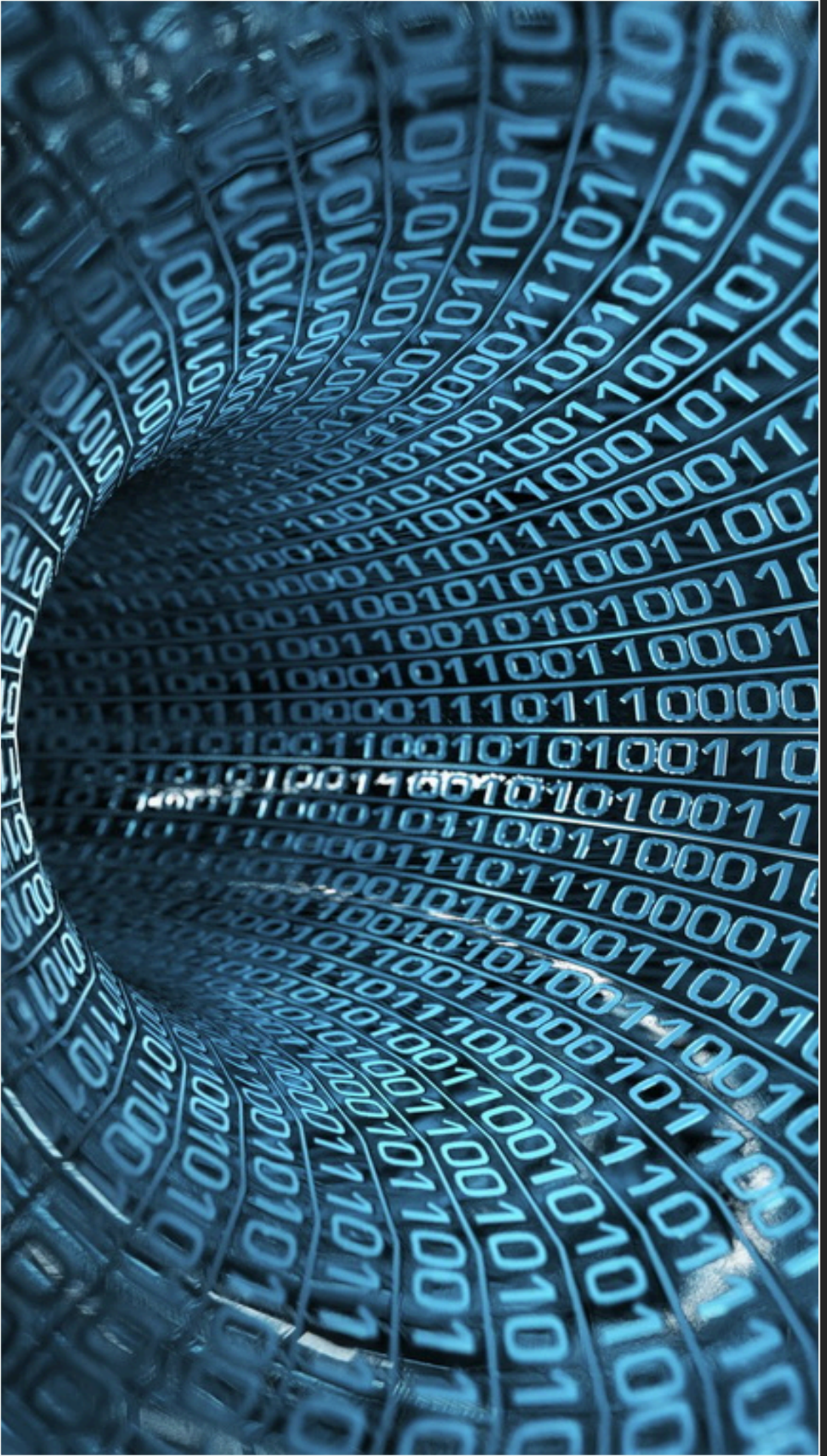


DOING SCIENCE WITH

BIG DATA

Nicholas Giangreco
Systems Biology PhD Candidate
http://tatonettilab.org @ Columbia University

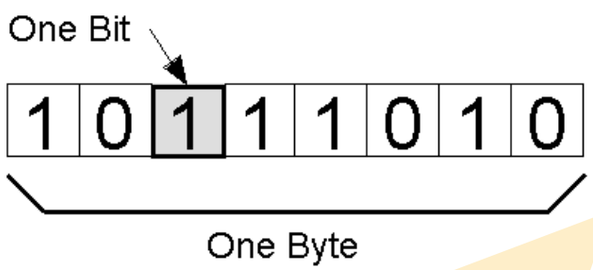
OUTLINE

- How big are we talking?
- Scientific process with data
- What I'm working on
- What the lab is working on DEMO
  - Open data, data-bases and interactive data
- Hardware that supports our software

MAKING SENSE OF BIG DATA?

DATA HAVE SIZE

MAKING SENSE OF BIG DATA?



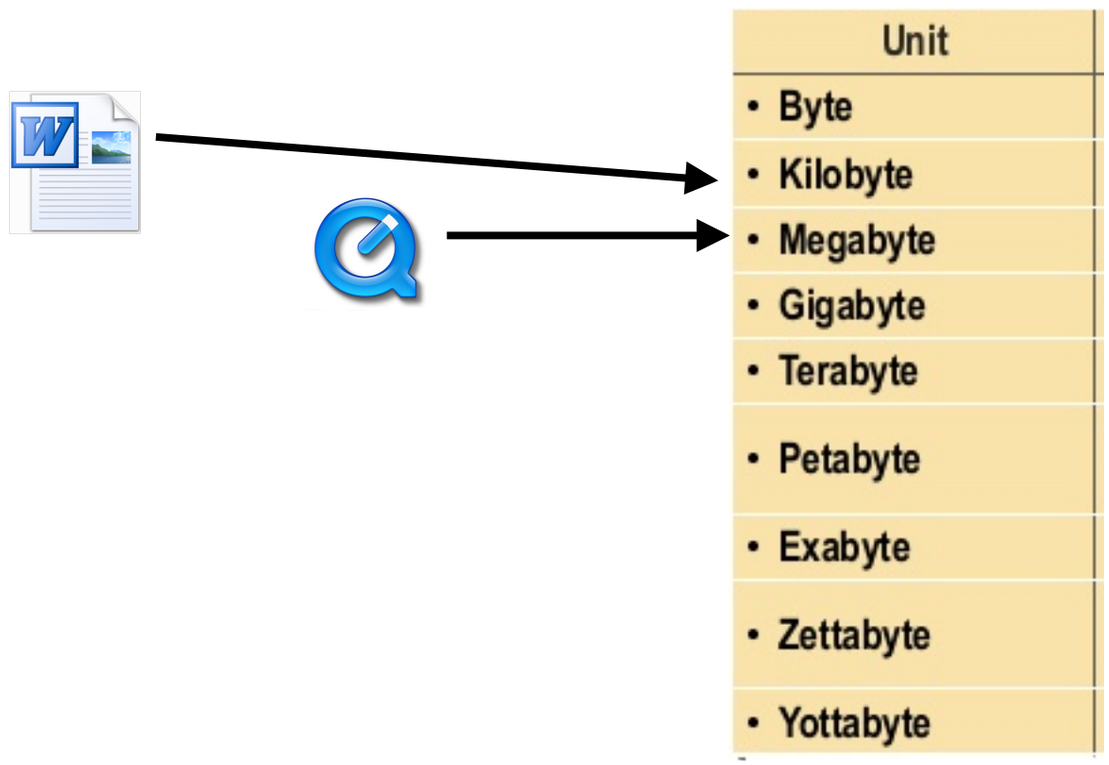
Unit	Size Indications
• Byte	A single letter **e.g. 'A'**
• Kilobyte	One page of typed text is 2KB
• Megabyte	A typical pop song is about 4MB
• Gigabyte	A 2 hour film can be compressed into 1 - 2GB
• Terabyte	Big enough to hold all the x_-ray files in a modern hospital
• Petabyte	Big enough to hold 13 years' worth of high-definition TV content. Google processes 1 PB every hour
• Exabyte	Equivalent to 10 billion copies of the Economist
• Zettabyte	If we're able to record every human word that has ever been spoken they would fill up about 42 zettabytes worth of memory
• Yottabyte	Too big to imagine

https://www.slideshare.net/barryooi/mustbig-datapresentation-29-jan-2015

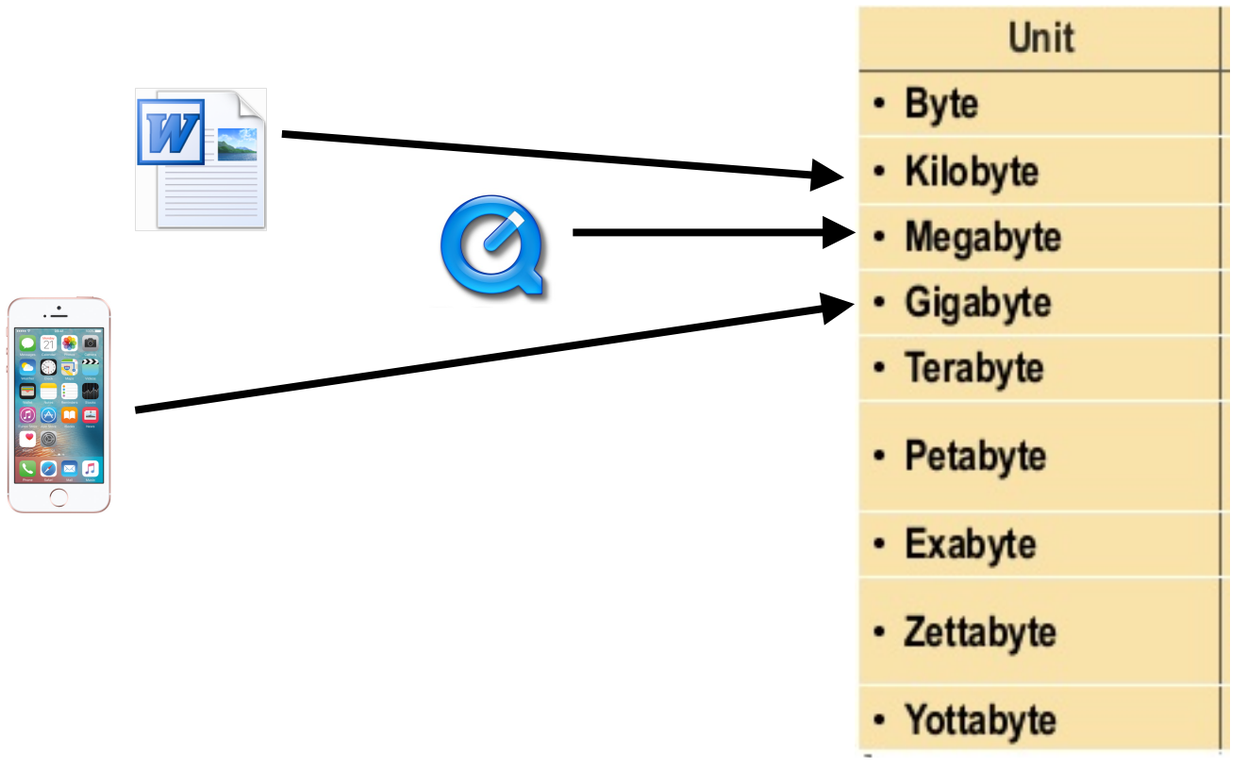
MAKING SENSE OF BIG DATA?

Unit
• Byte
• Kilobyte
• Megabyte
• Gigabyte
• Terabyte
• Petabyte
• Exabyte
• Zettabyte
• Yottabyte

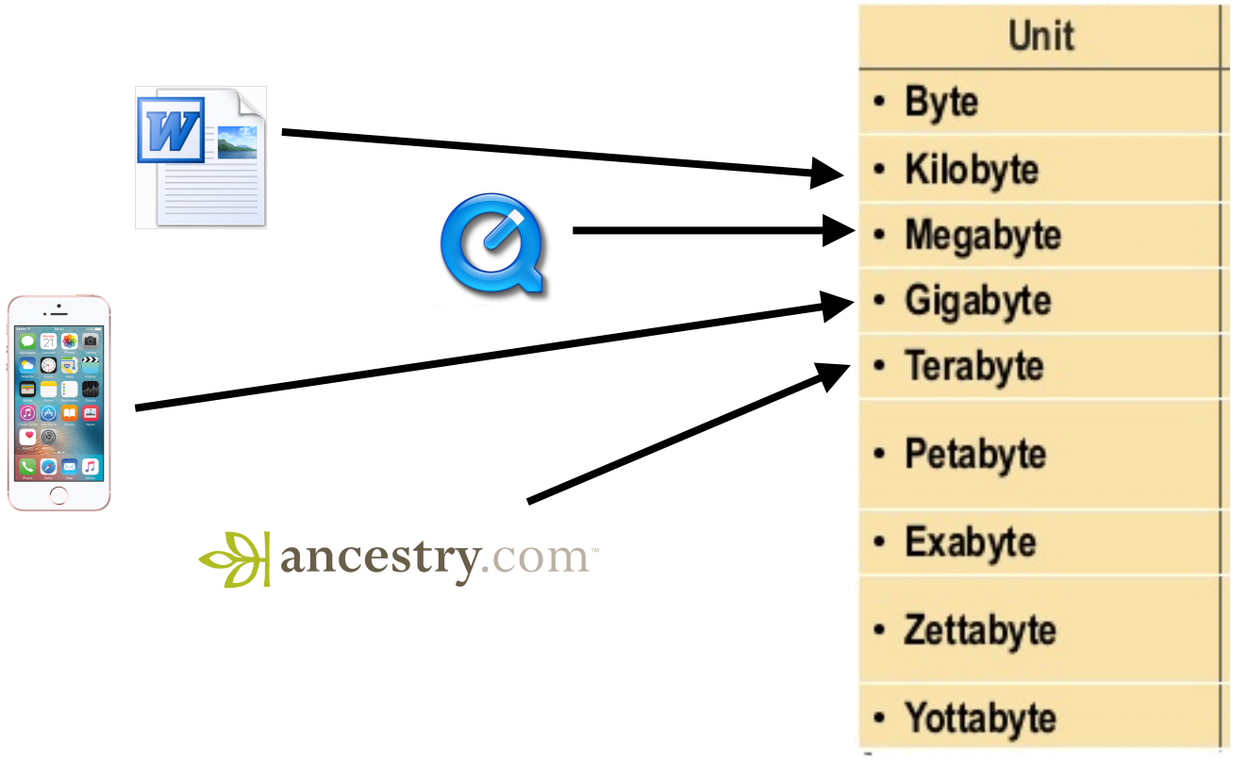
MAKING SENSE OF BIG DATA?



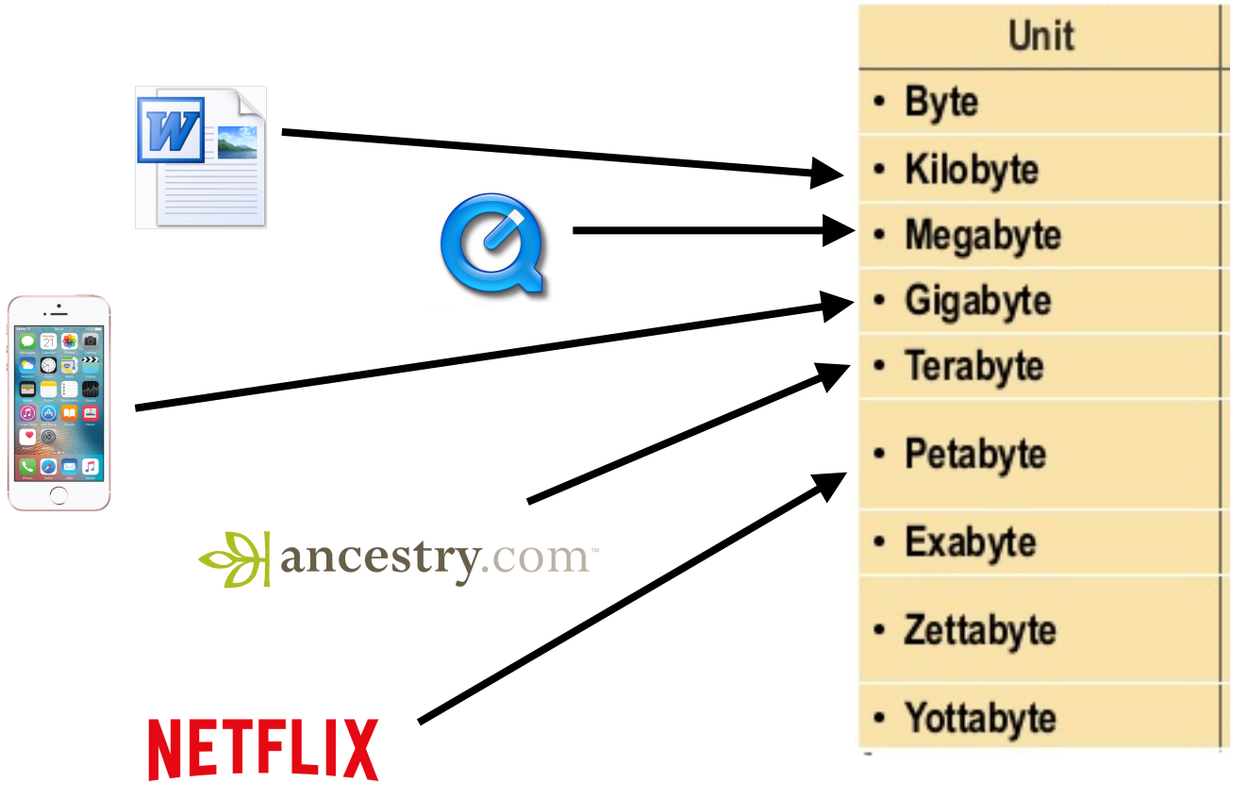
MAKING SENSE OF BIG DATA?



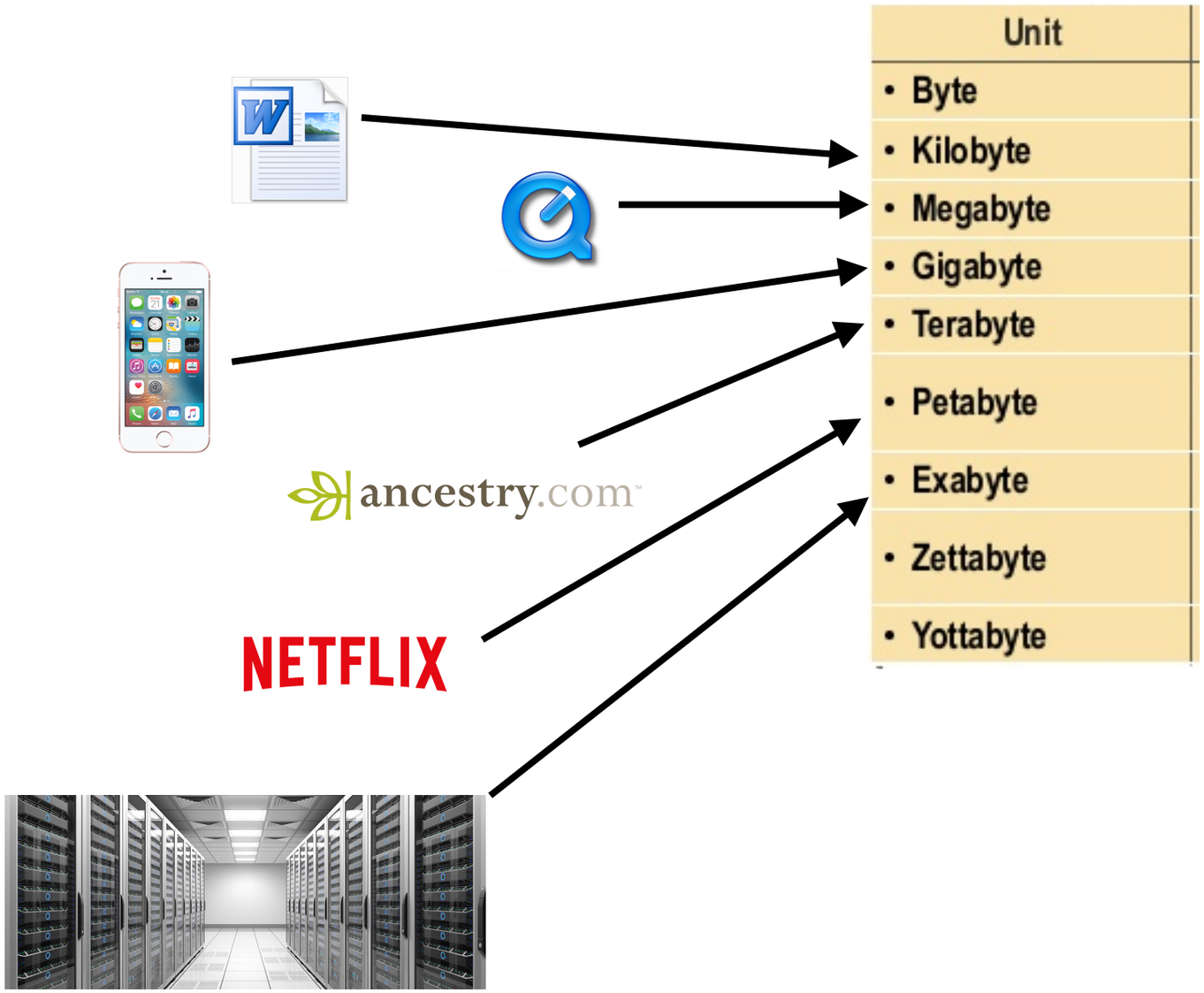
MAKING SENSE OF BIG DATA?



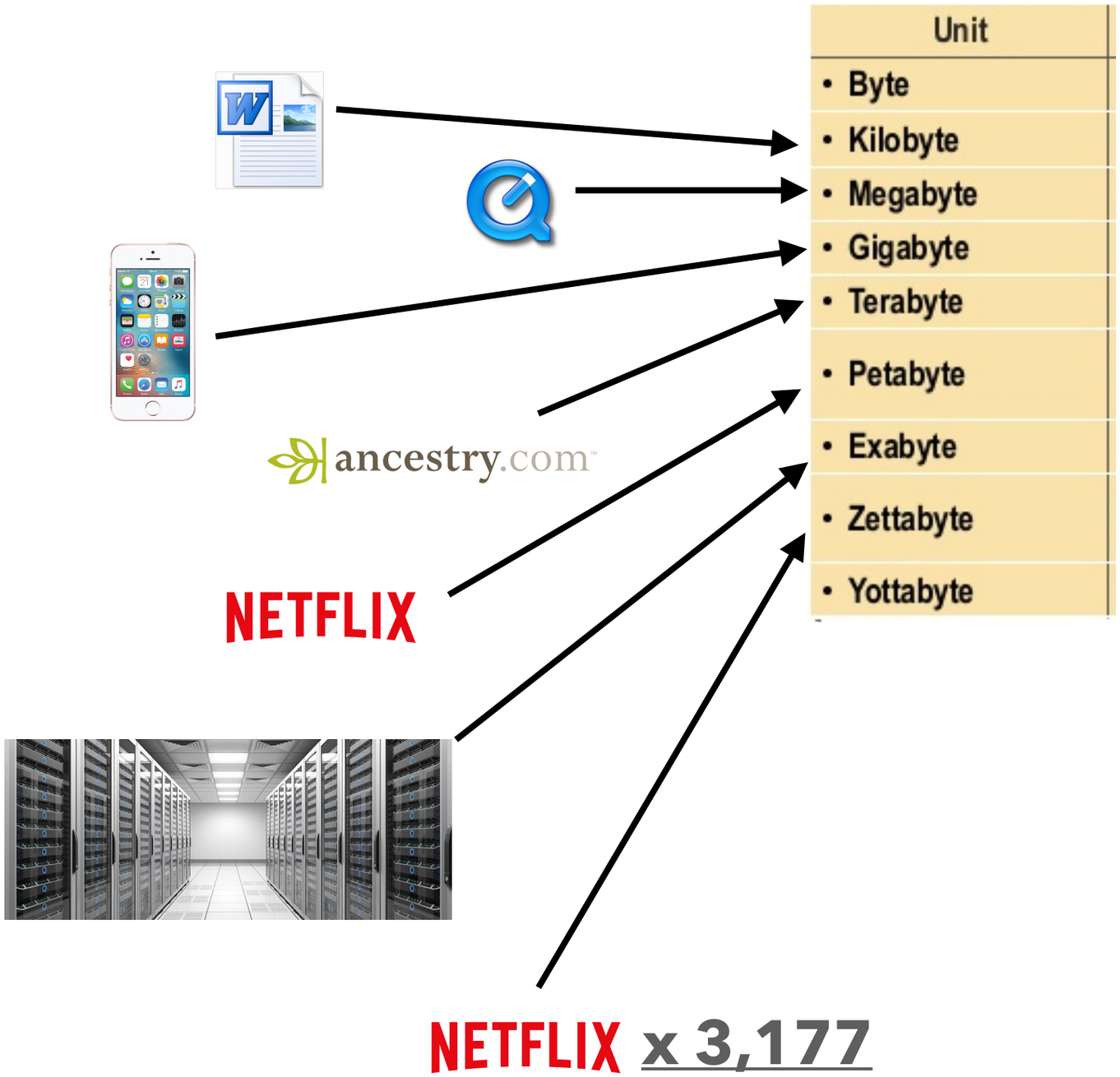
MAKING SENSE OF BIG DATA?



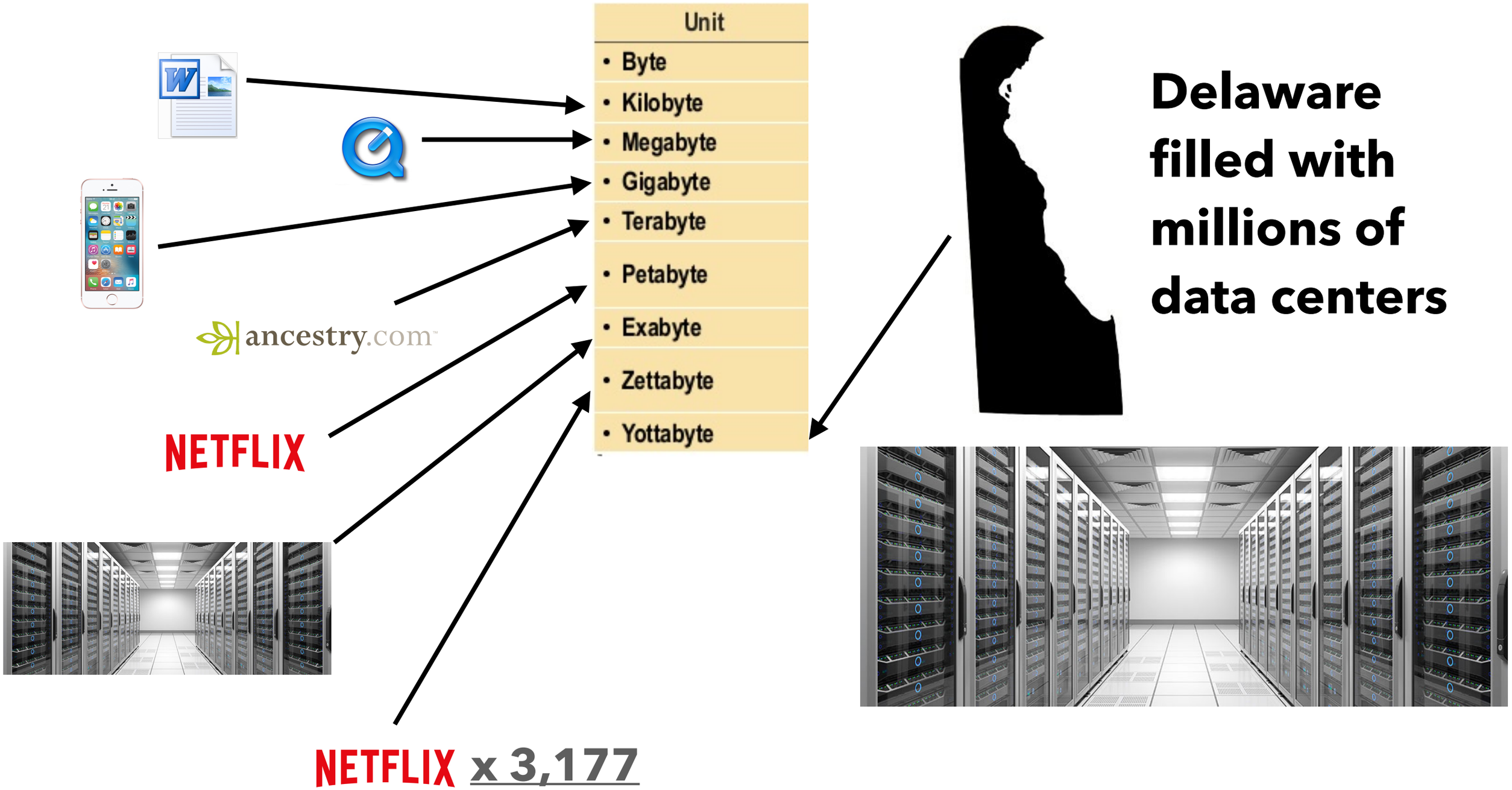
MAKING SENSE OF BIG DATA?



MAKING SENSE OF BIG DATA?



MAKING SENSE OF BIG DATA?



HEALTH DATA

Genomic and healthcare databases

- Gigabyte
- Terabyte
- Petabyte

HEALTH DATA

Genomic and healthcare databases

- Gigabyte
- Terabyte
- Petabyte

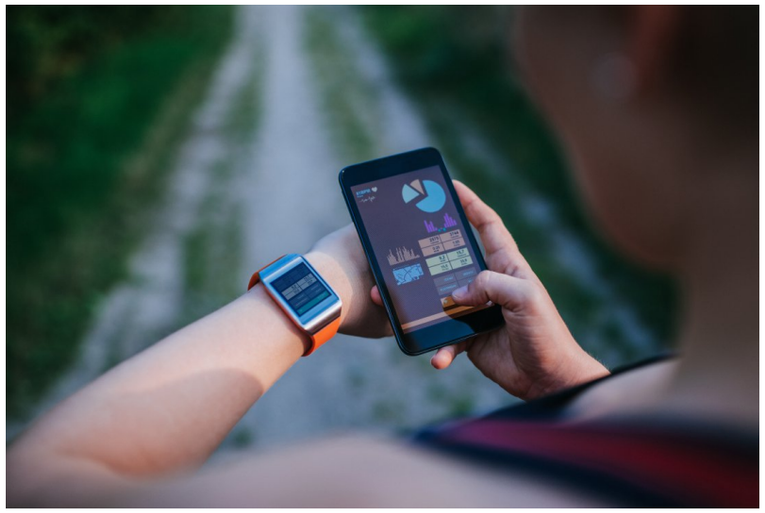
Database growth over many years…

HEALTH DATA

Genomic and healthcare databases

- Gigabyte
- Terabyte
- Petabyte

Database growth over many years…



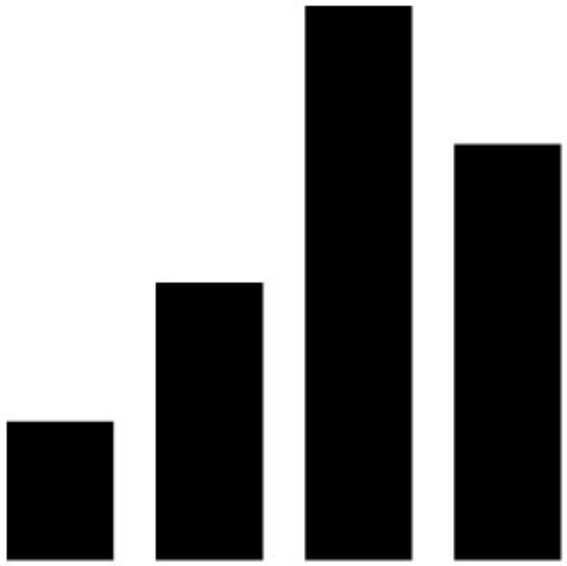
Challenges:

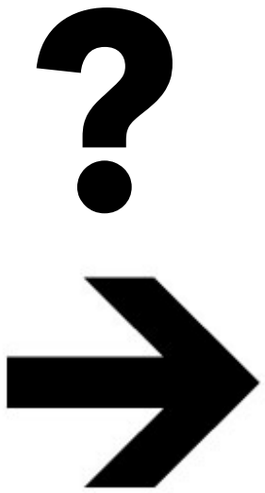
HEALTH DATA

Challenges:

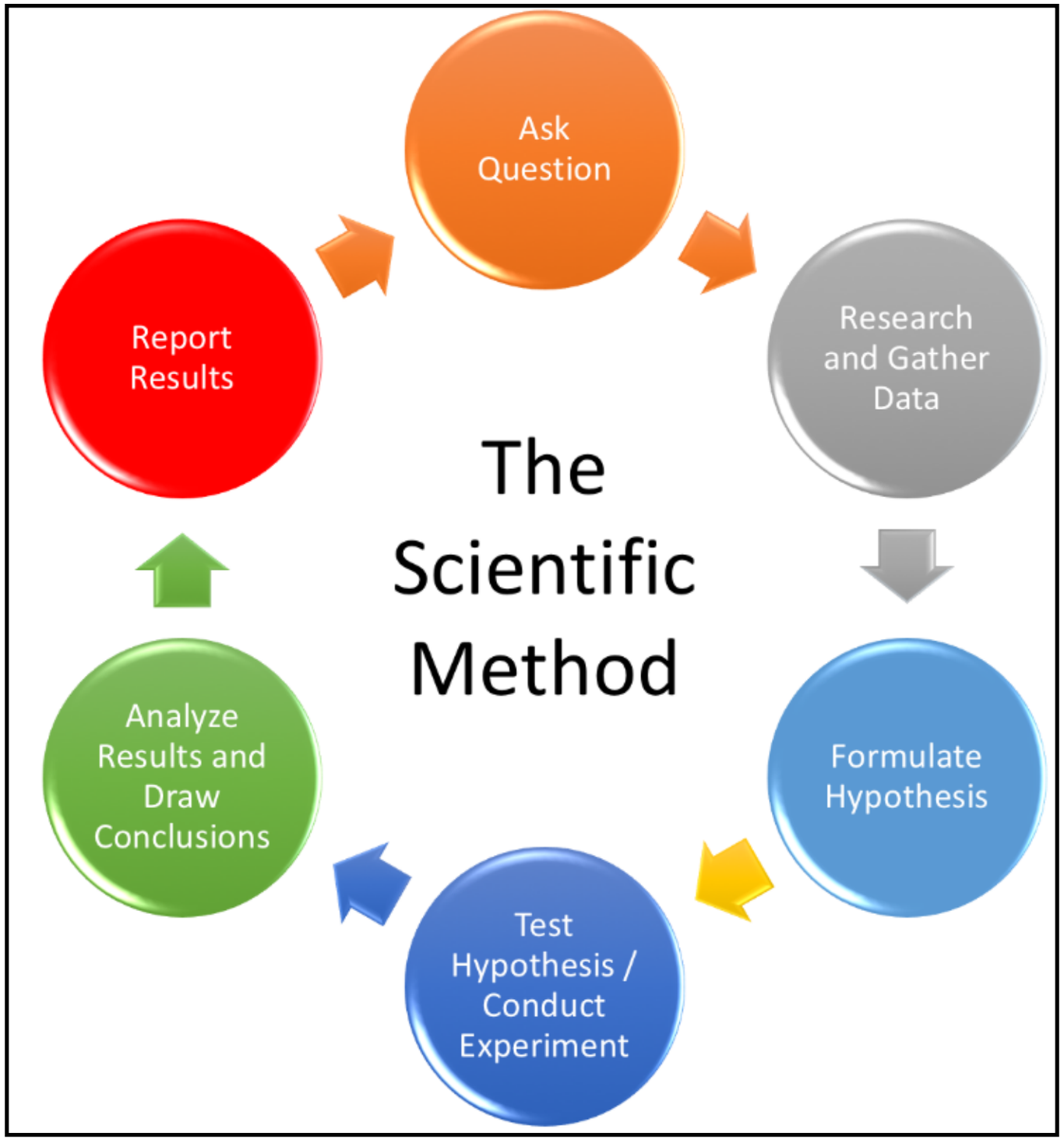
HEALTH DATA

Challenges:

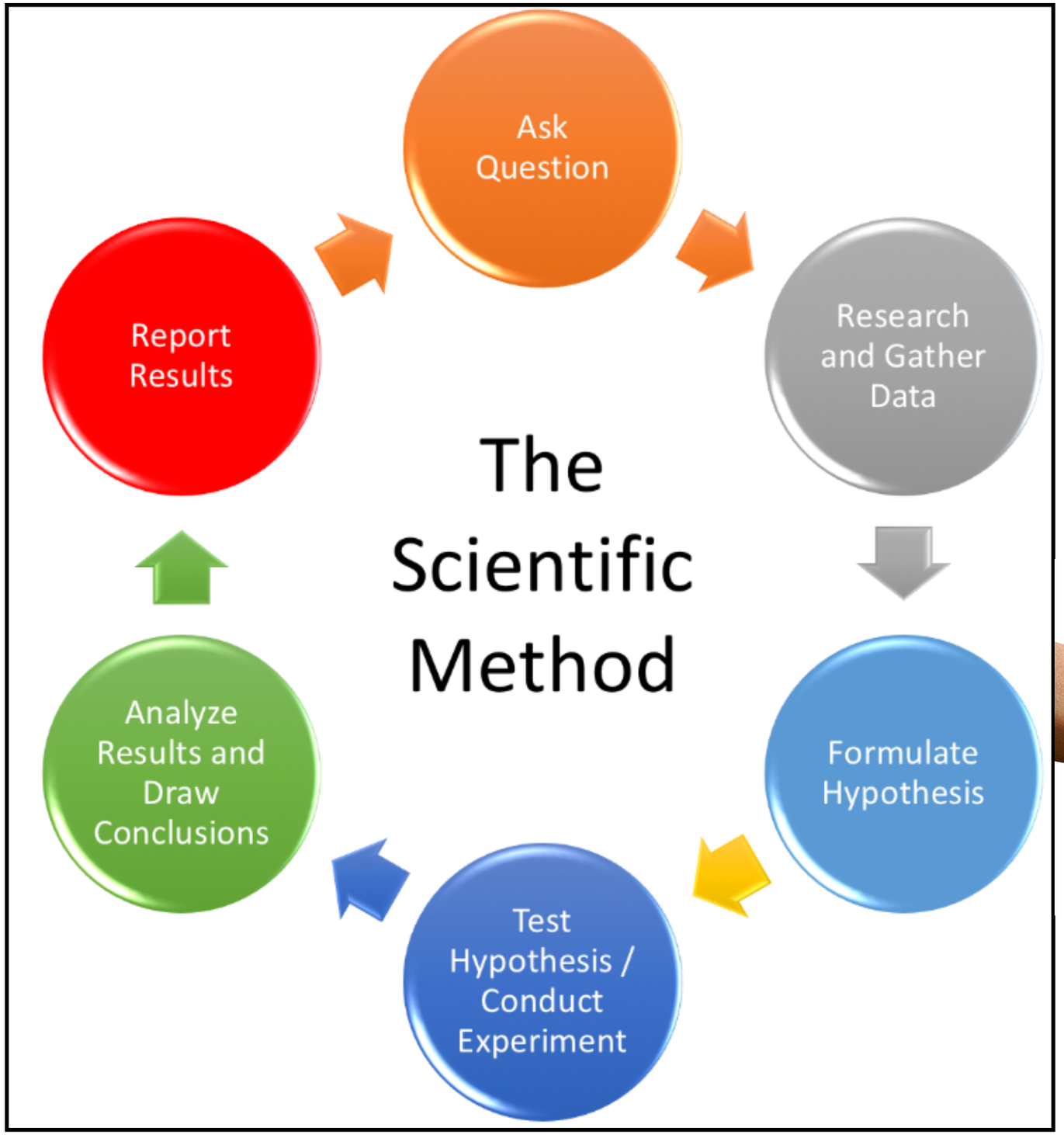




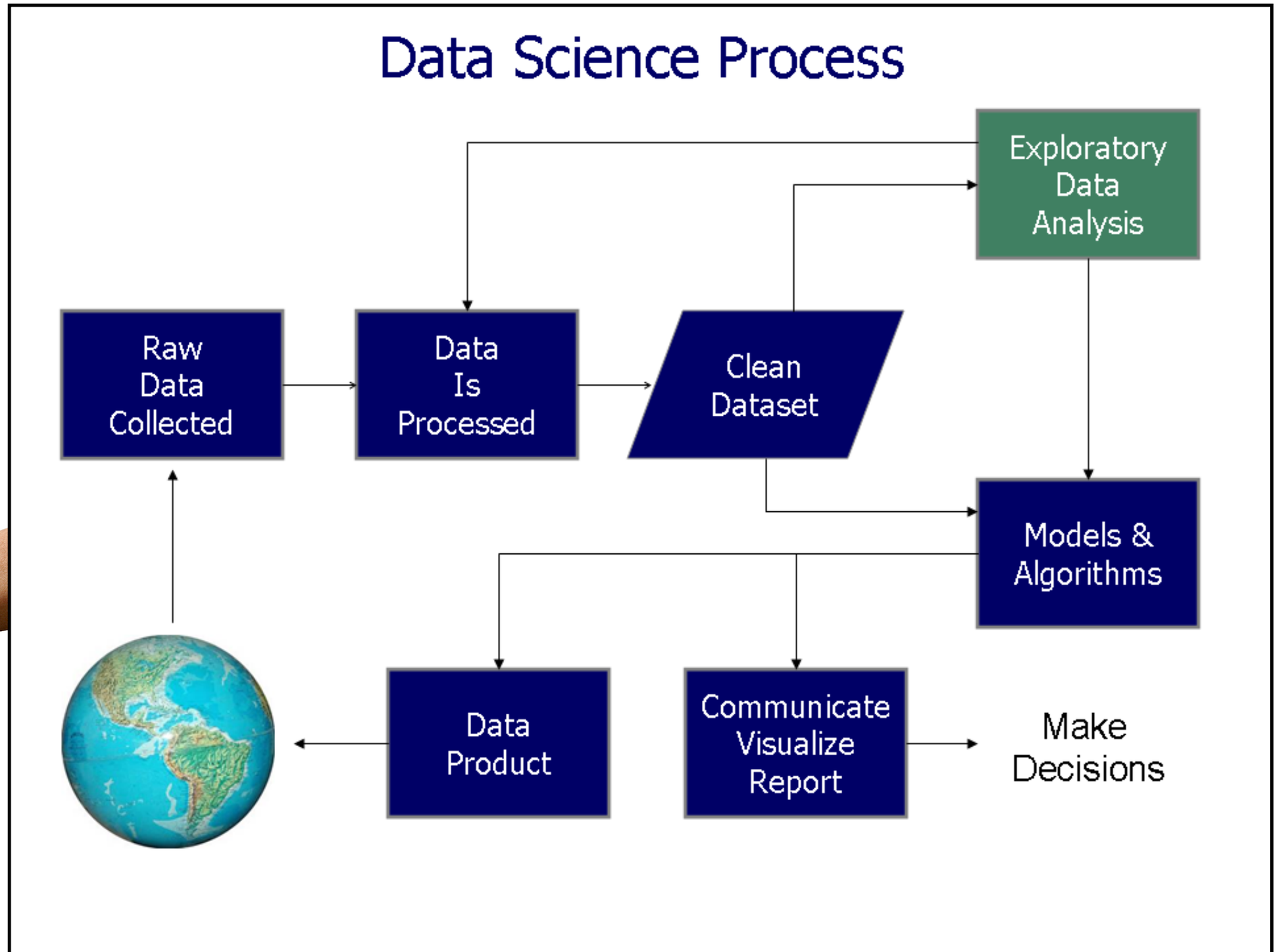
SCIENTIFIC PROCESS WITH DATA



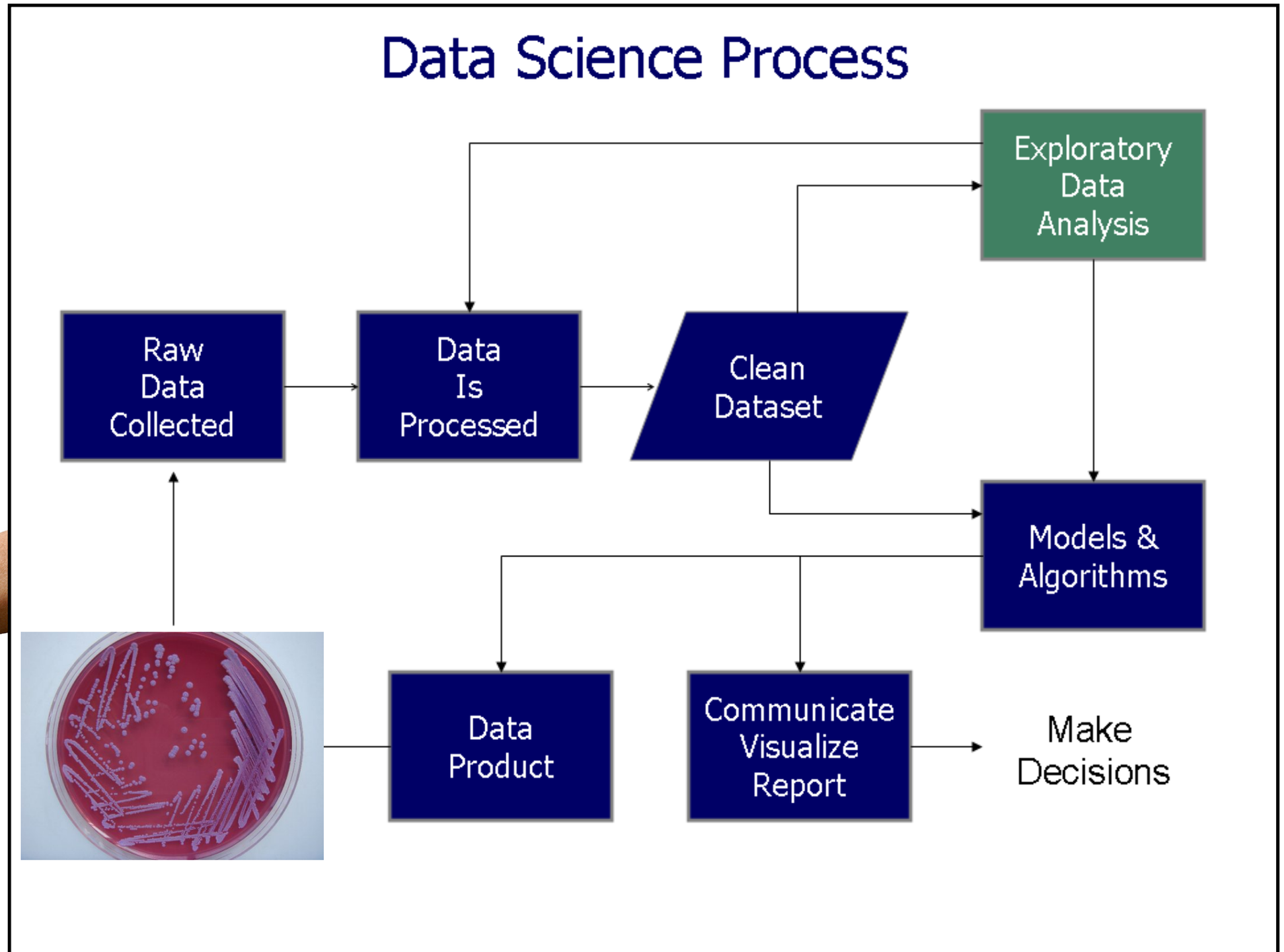
SCIENTIFIC PROCESS WITH DATA



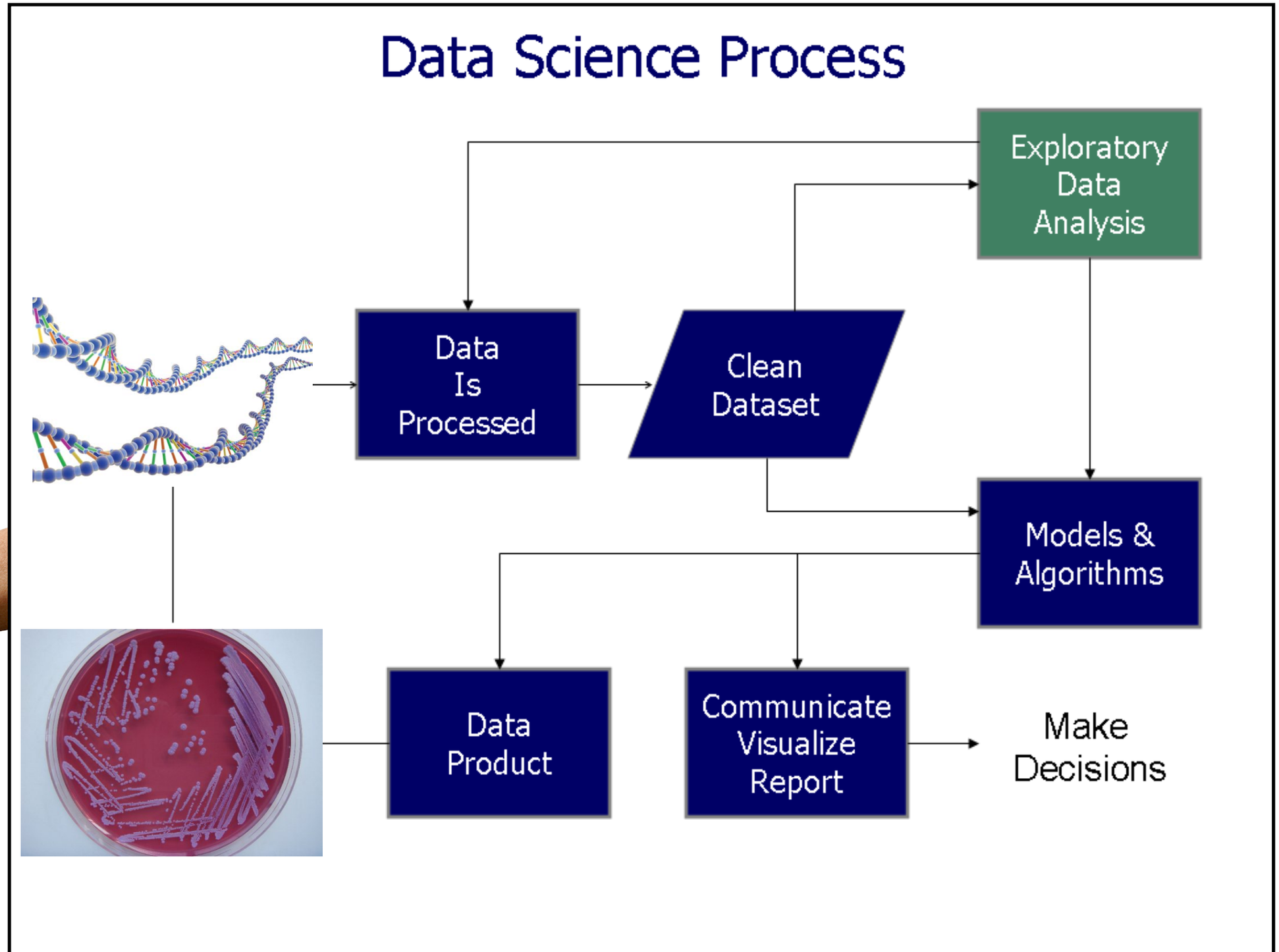
SCIENTIFIC PROCESS WITH DATA



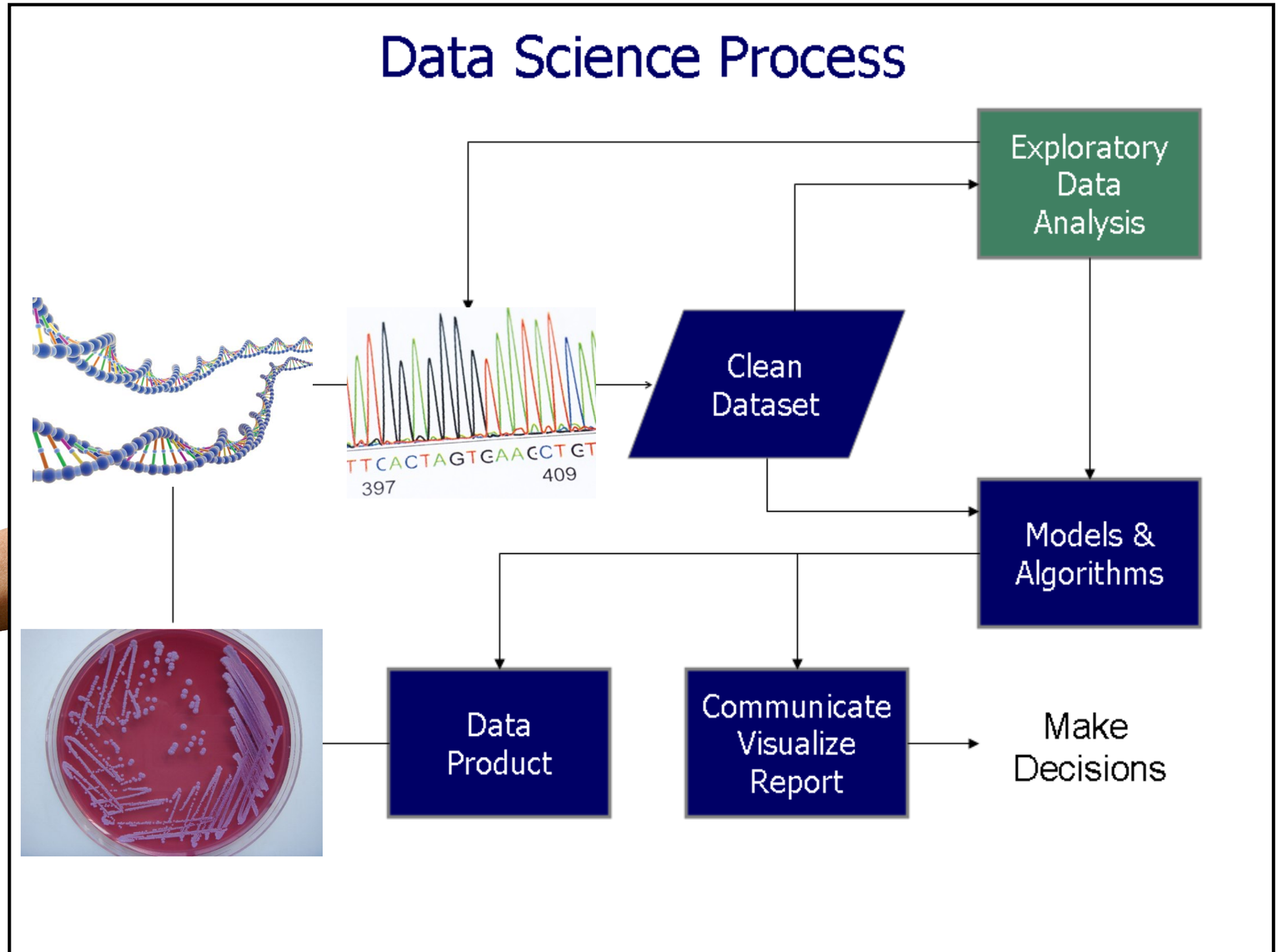
SCIENTIFIC PROCESS WITH DATA



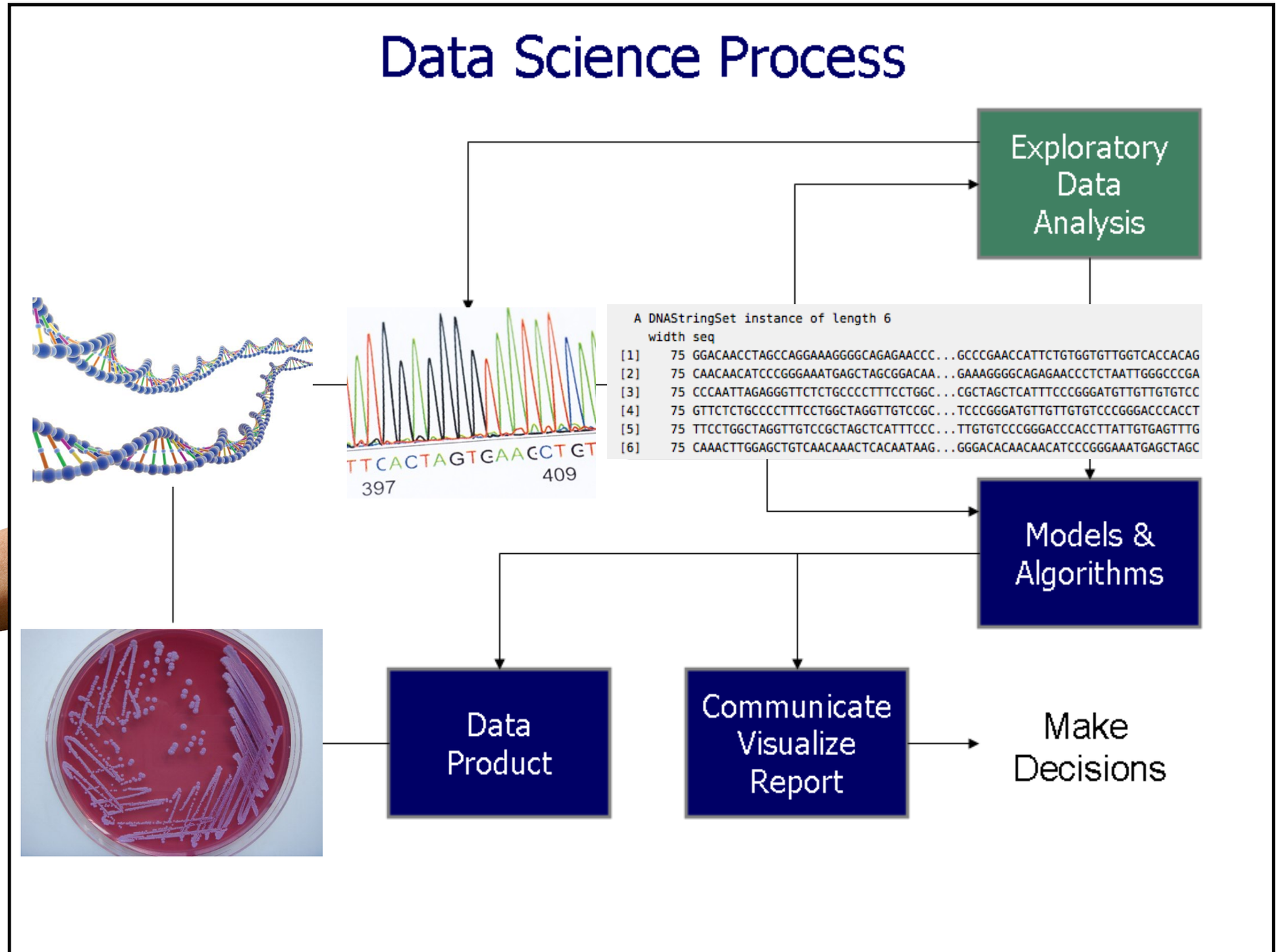
SCIENTIFIC PROCESS WITH DATA



SCIENTIFIC PROCESS WITH DATA

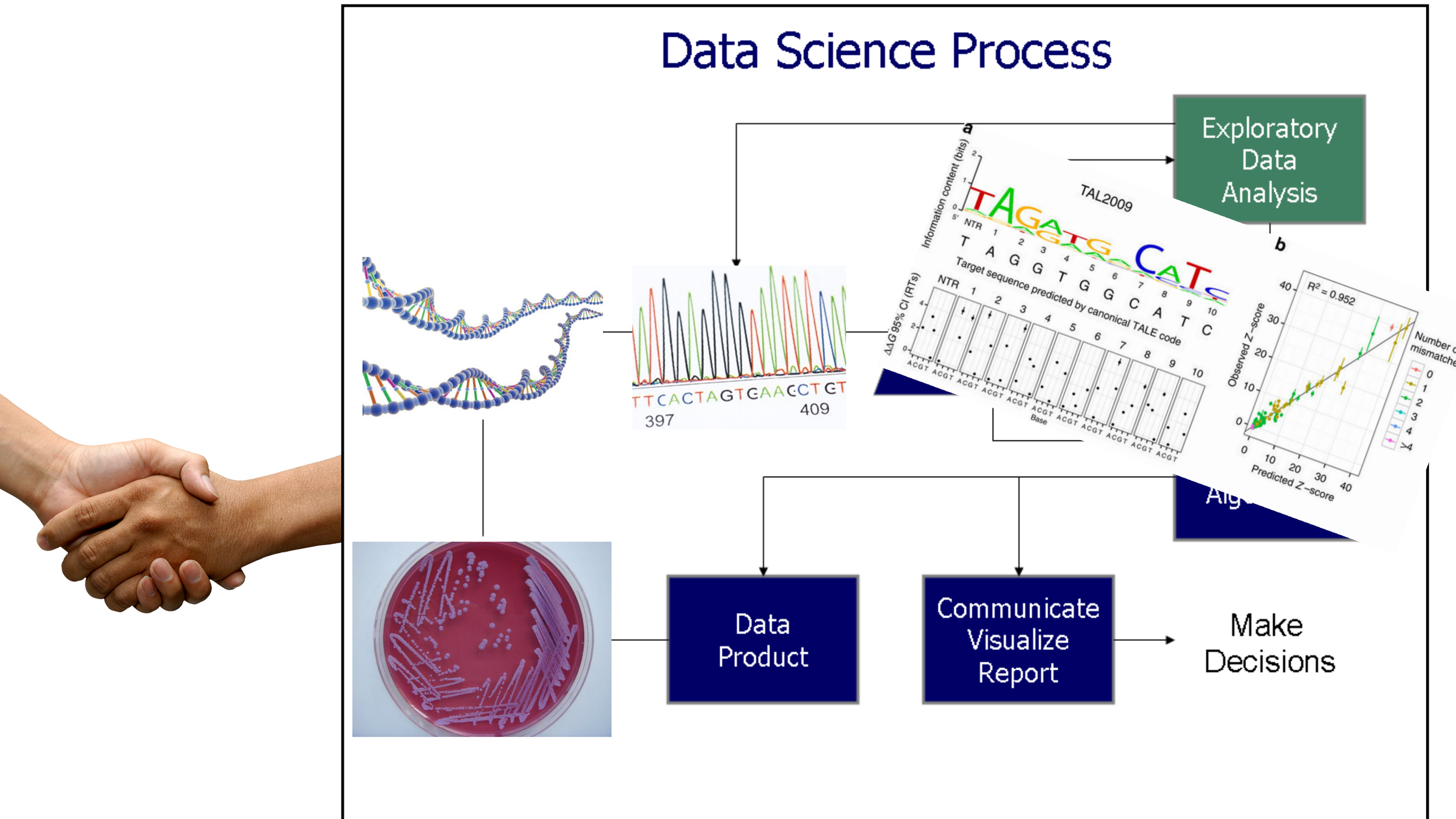


SCIENTIFIC PROCESS WITH DATA



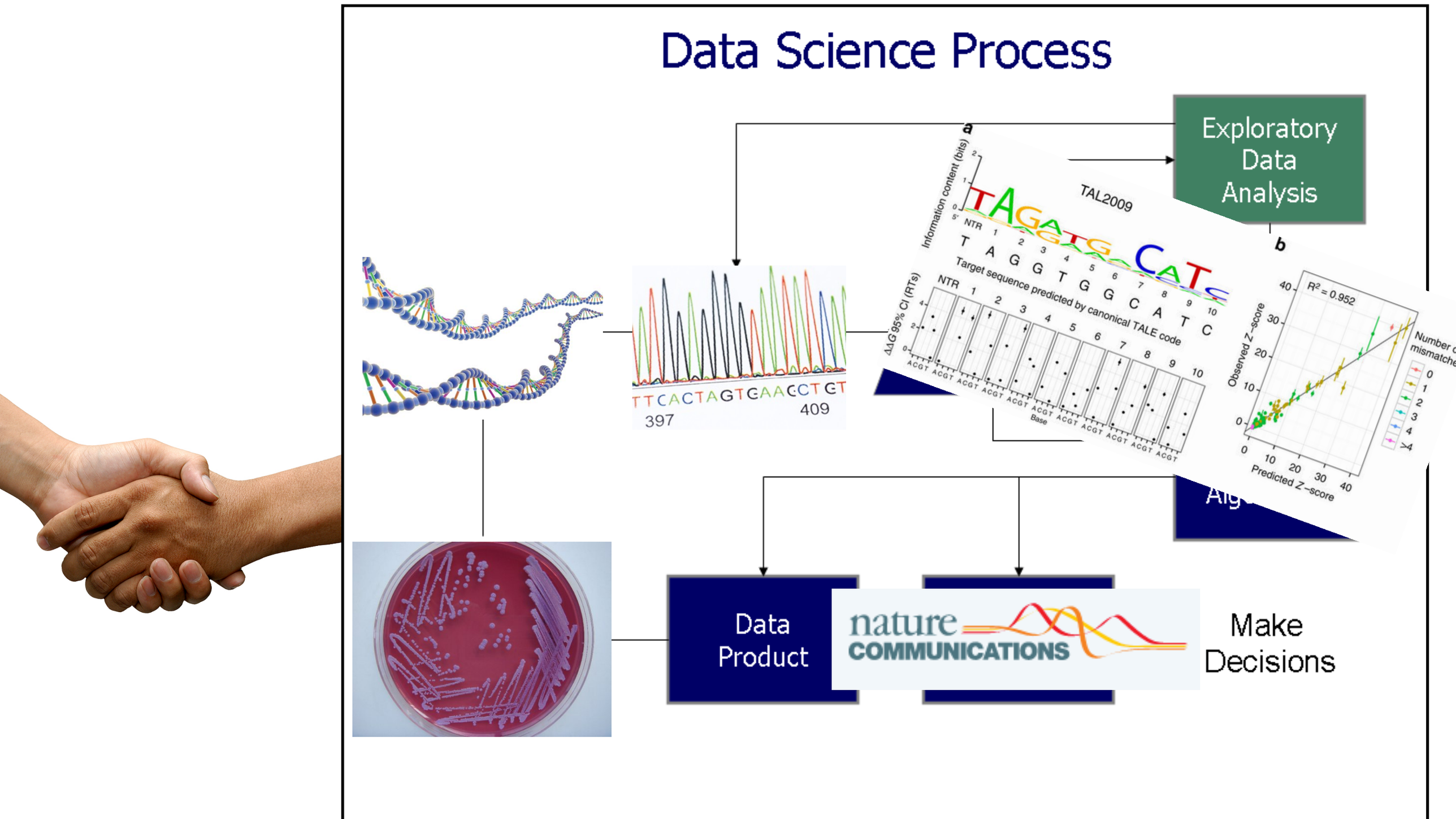
SCIENTIFIC PROCESS WITH DATA

Rogers et al. Nature 2015



SCIENTIFIC PROCESS WITH DATA

Rogers et al. Nature 2015



WHAT I'VE BEEN WORKING ON...

My current projects

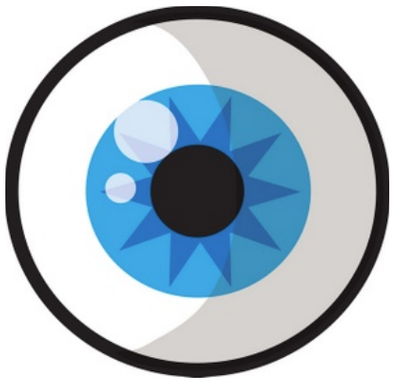
My current projects

My current projects

PROTEIN BIOMARKERS IN HEART FAILURE

GENE EXPRESSION IN HEART FAILURE

GENE EXPRESSION DURING



EYE DEVELOPMENT

PHARMACOLOGY OF EPILEPSY

My current projects

EPILEPSY

- Seizures that occur more than once.

EPILEPSY

- Seizures that occur more than once.
- Early onset (child) epilepsy is largely genetic.[1]

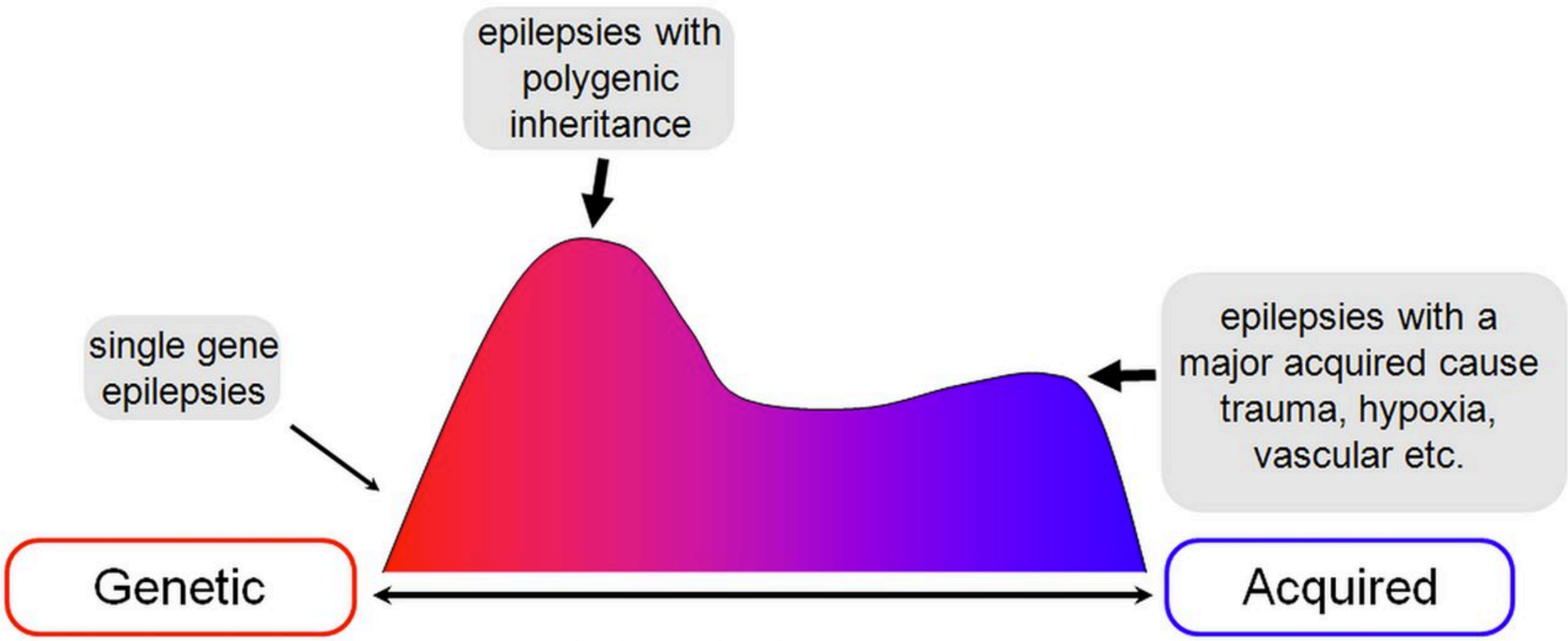
1. World Health Organization

EPILEPSY

- Seizures that occur more than once.
- Early onset (child) epilepsy is largely genetic.[1]
- Late onset (adult) epilepsy is largely acquired.[1]

1. World Health Organization

EPILEPSY



Journal of Medical Genetics 2012

ADVERSE DRUG REACTIONS

- Drugs are supposed to treat your disease
  - Anti-epileptic drugs treat epilepsy

ADVERSE DRUG REACTIONS

- Drugs are supposed to treat your disease
  - Anti-epileptic drugs treat epilepsy
- But, not all drugs work the same

ADVERSE DRUG REACTIONS

- Drugs are supposed to treat your disease
  - Anti-epileptic drugs treat epilepsy
- But, not all drugs work the same
- May cause an **adverse reaction**,
  - fever, nausea, bloating, dizziness/confusion
  - heart attack, liver damage, kidney failure, seizures

ADVERSE DRUG REACTIONS

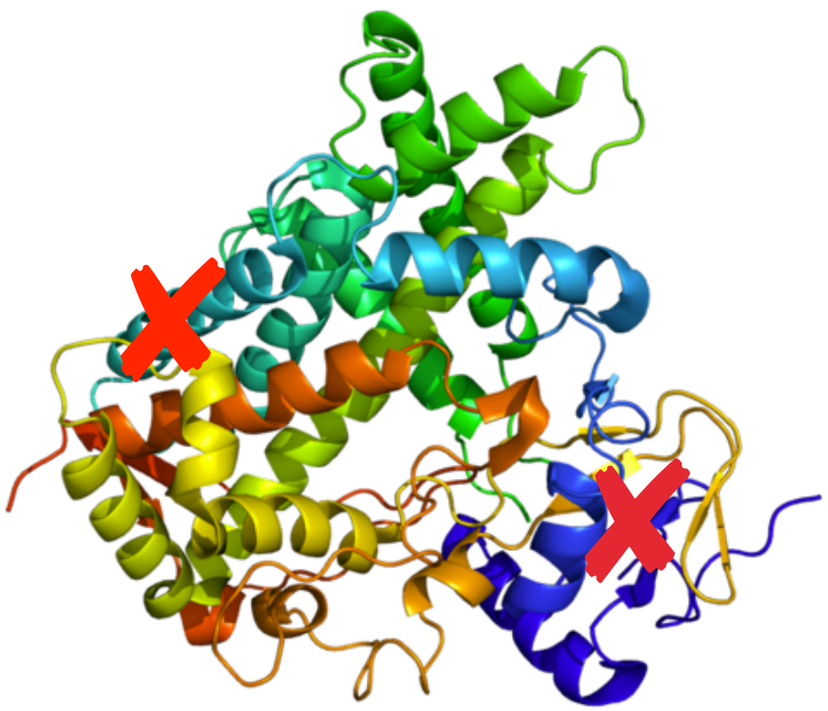
- Why?

ADVERSE DRUG REACTIONS

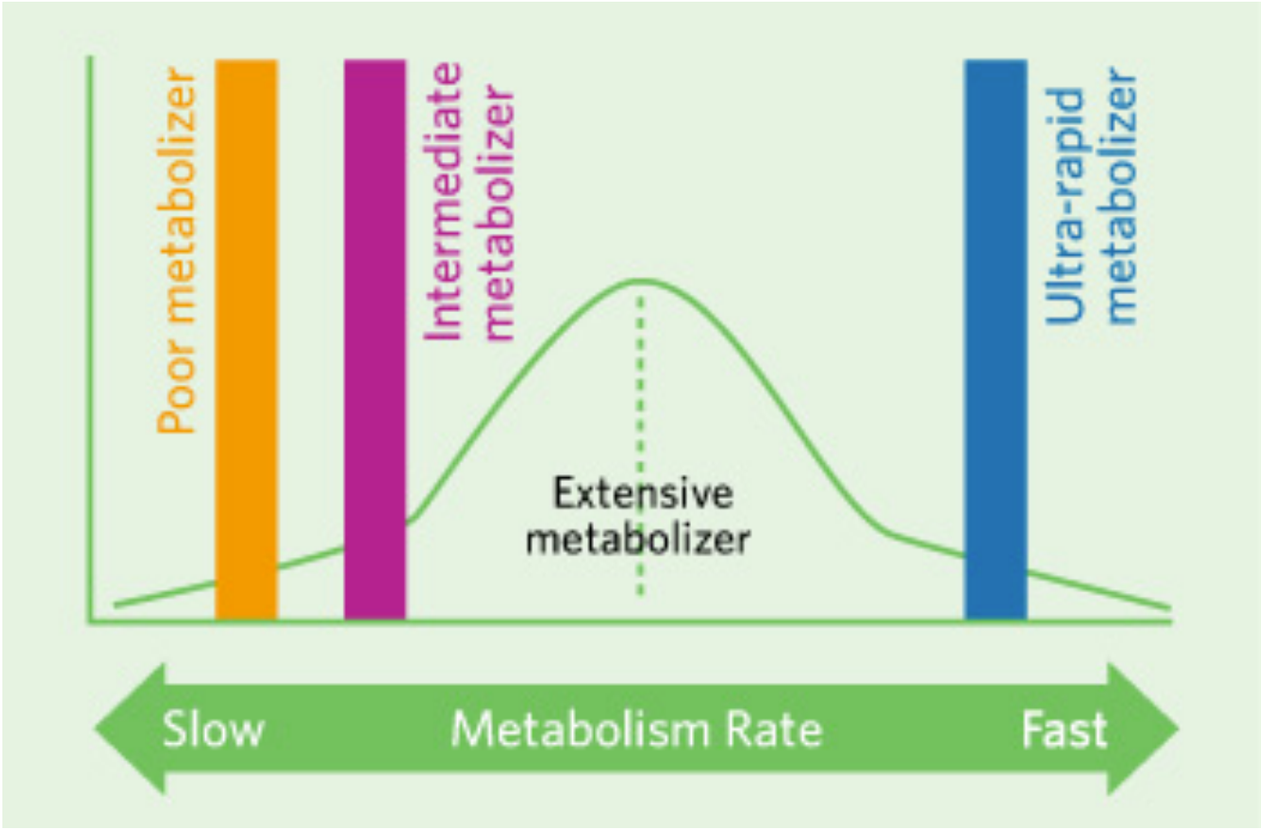
- Why?
  - Genetics could be a reason

ADVERSE DRUG REACTIONS

- Why?
  - Genetics could be a reason:



ADVERSE DRUG REACTIONS



http://www.questdiagnostics.com/home/physicians/testing-services/by-test-name/prescription-drug-monitoring/genetic-testing

ADVERSE DRUG REACTIONS IN CHILDREN WITH EPILEPSY

- Do children respond differently to anti-epileptic drugs?

ADVERSE DRUG REACTIONS IN CHILDREN WITH EPILEPSY

- Do children respond differently to anti-epileptic drugs?
  - Are they more or less frequent?

ADVERSE DRUG REACTIONS IN CHILDREN WITH EPILEPSY

- Do children respond differently to anti-epileptic drugs?
  - Are they more or less frequent?
  - Is the type of adverse reactions different?

DATA

DATA

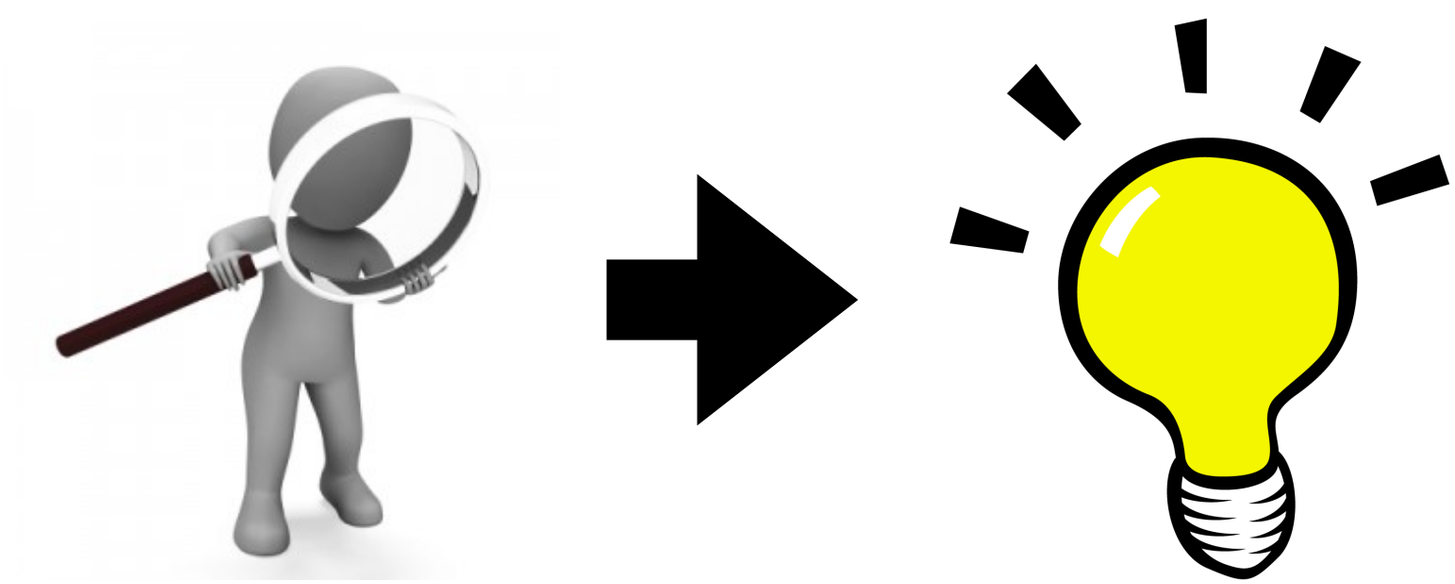
- Reports submitted indicate:
  - Drug(s) a patient was taking
  - Adverse reactions the patient experienced
  - Age of the patient

FDA
FAERS | FDA Adverse Event Reporting System

>8 million reports from 2004-2015

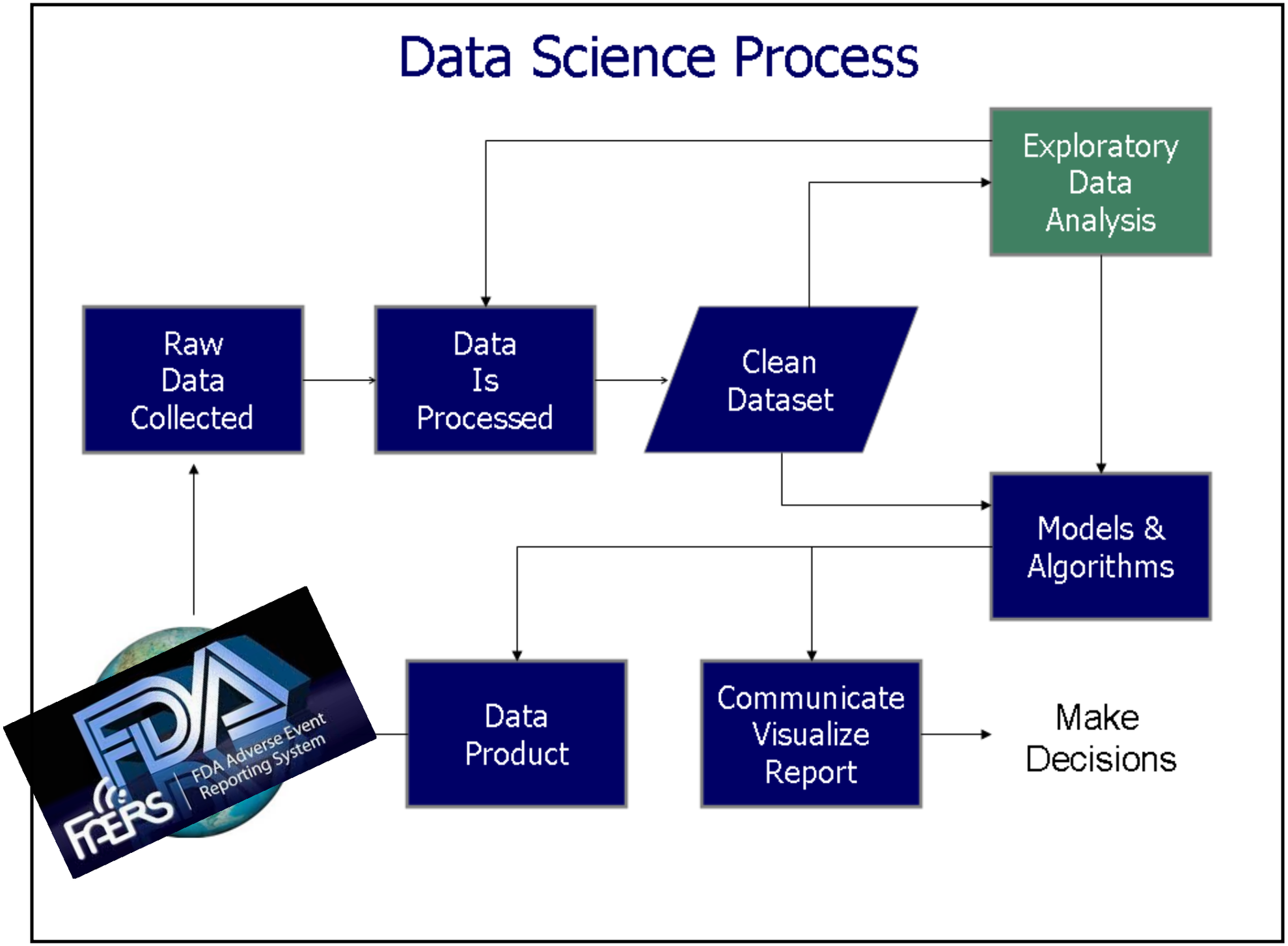
DATA

- Reports submitted indicate:
  - Drug(s) a patient was taking
  - Adverse reactions the patient experienced
  - Age of the patient
- Objective:



>8 million reports from 2004-2015

APPROACH:


Data Science Process
Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Data Product
Communicate Visualize Report
Make Decisions
FDA
FDA Adverse Event Reporting System
FAERS

Data is

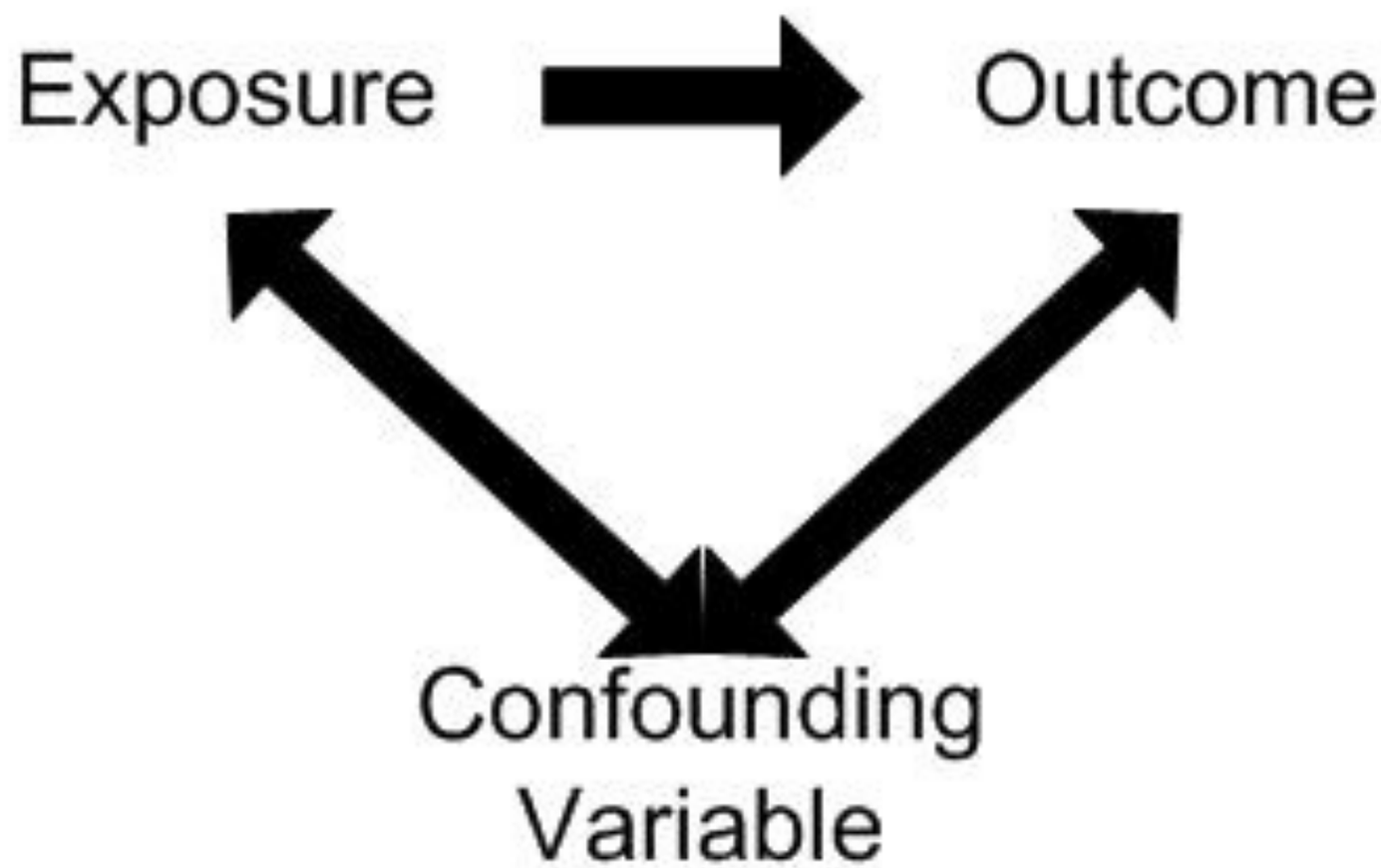
- Messy

Data is

- Messy
- Error-prone

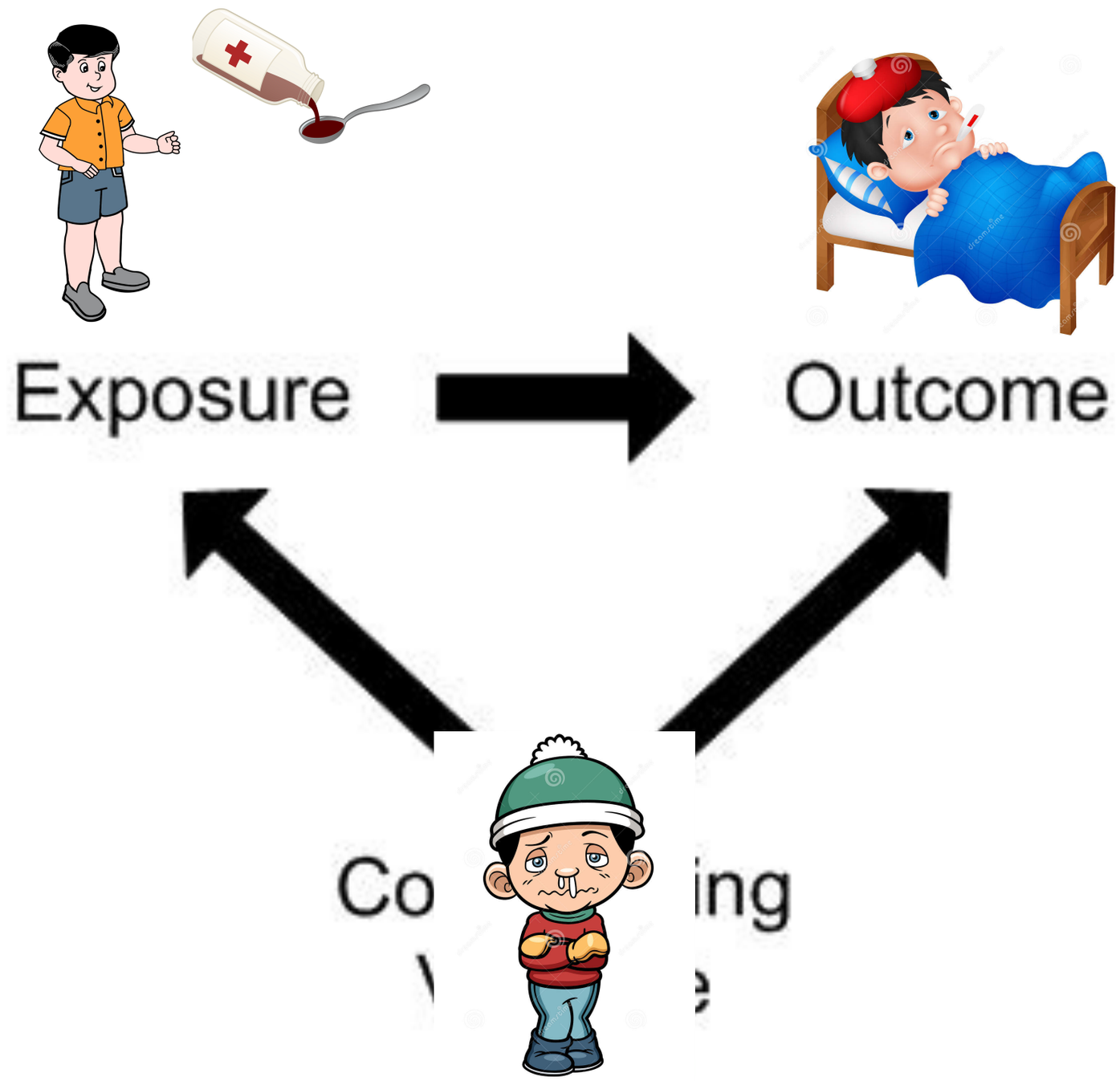
Data is

- Messy
- Error-prone
- Biased



Data is

- Messy
- Error-prone
- Biased

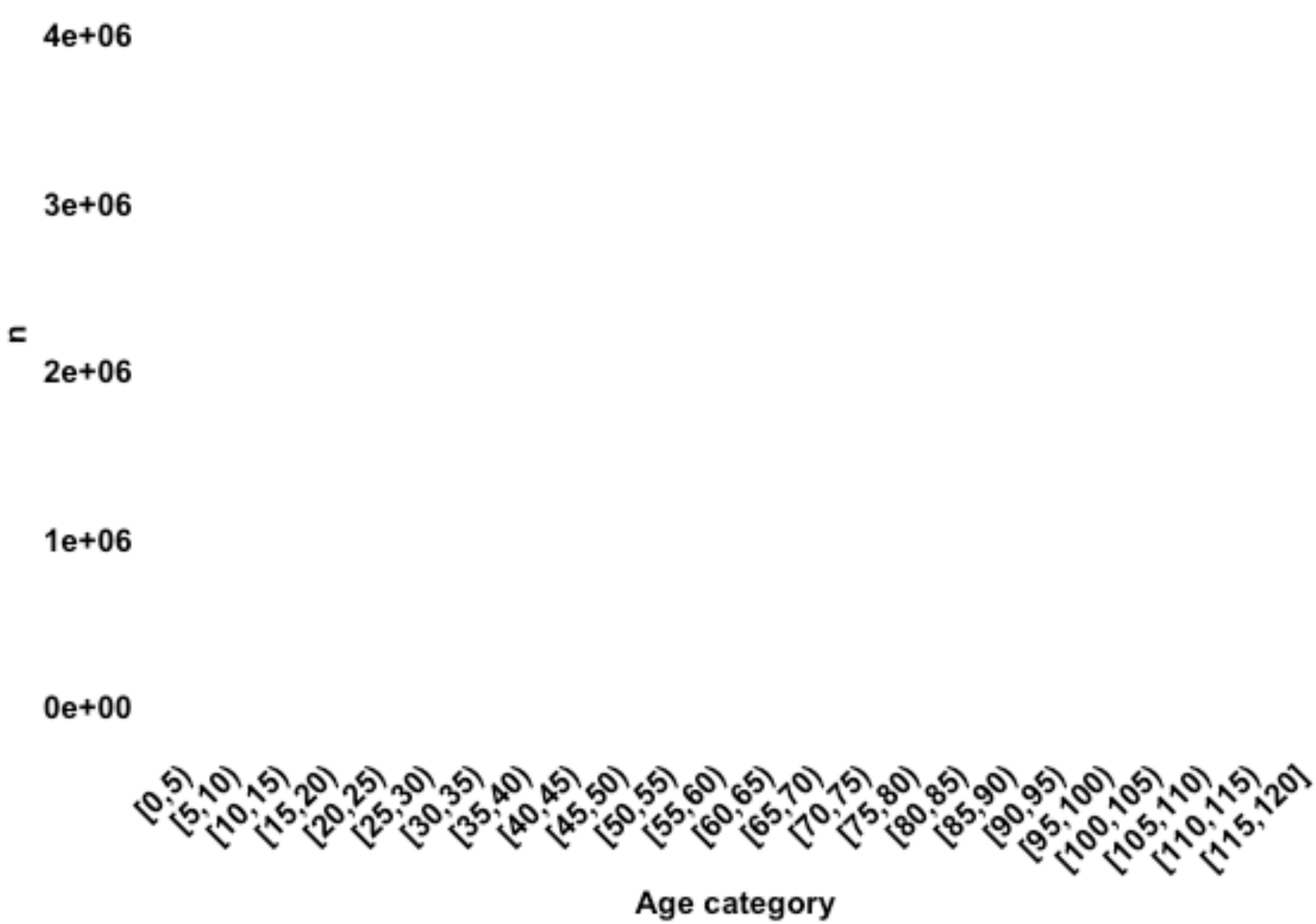


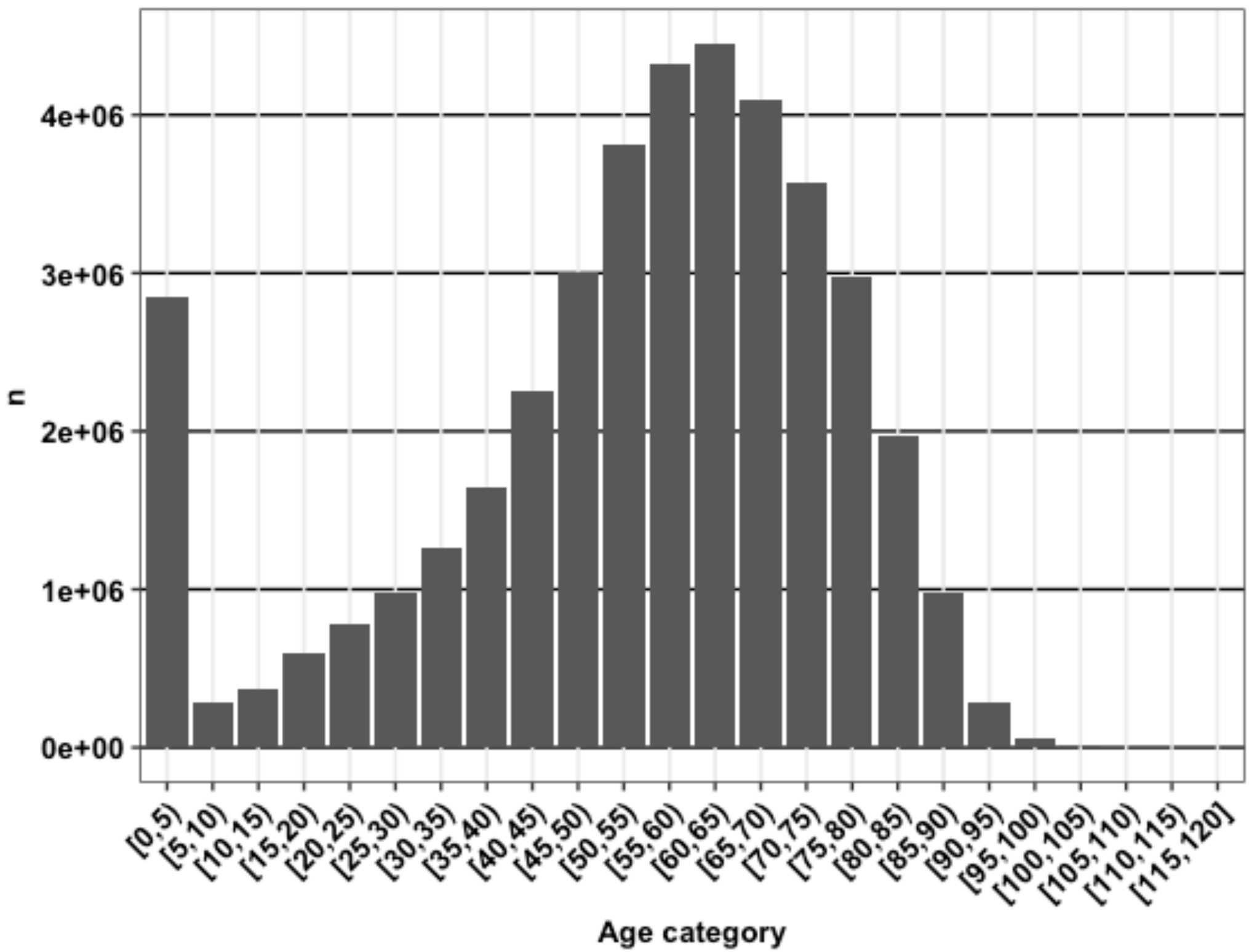
To *understand* the data,

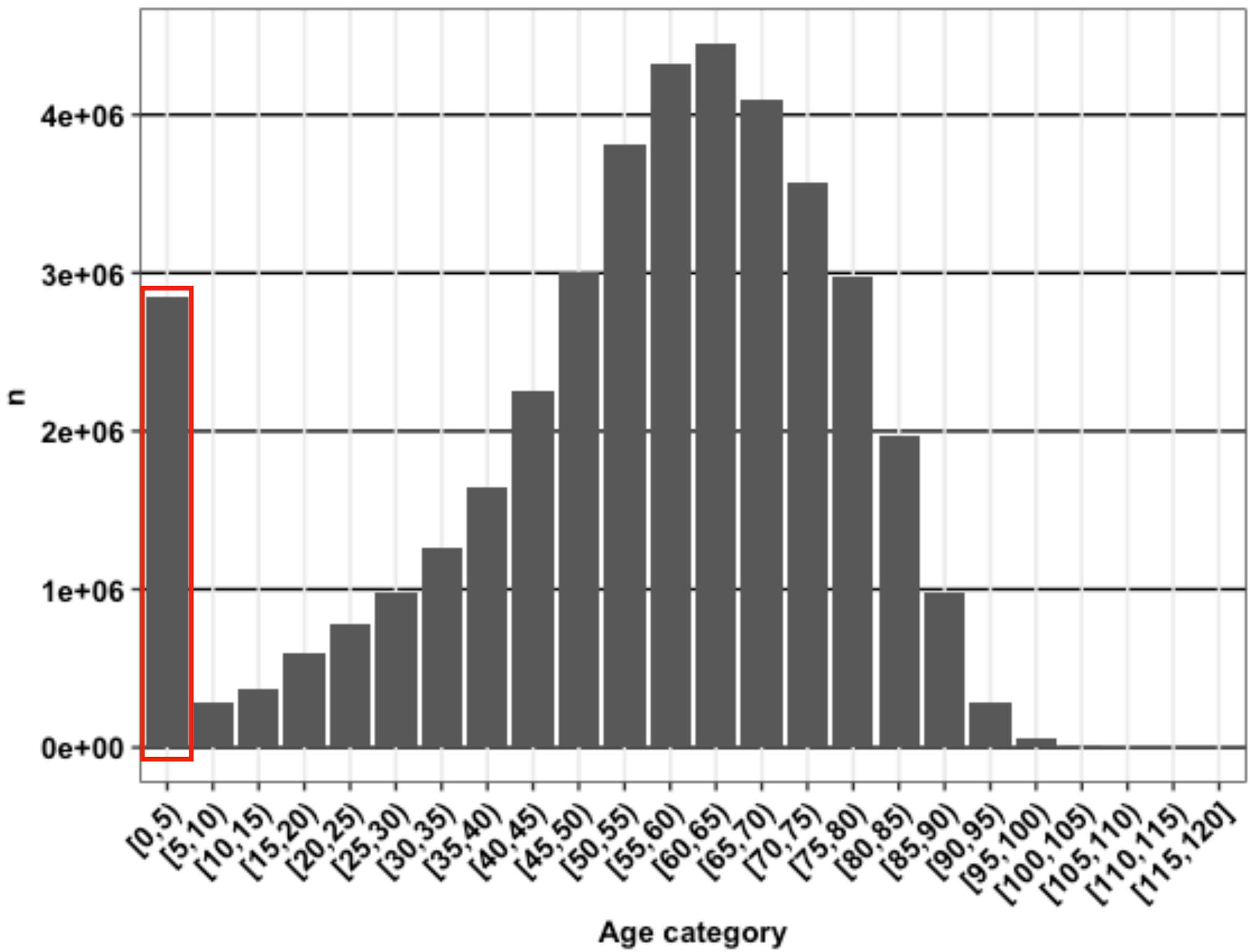
we need to *investigate* the data

[0,5)
[5,10)
[10,15)
[15,20)
[20,25)
[25,30)
[30,35)
[35,40)
[40,45)
[45,50)
[50,55)
[55,60)
[60,65)
[65,70)
[70,75)
[75,80)
[80,85)
[85,90)
[90,95)
[95,100)
[100,105)
[105,110)
[110,115)
[115,120]

Age category


4e+06
3e+06
2e+06
1e+06
0e+00
n
[0,5)
[5,10)
[10,15)
[15,20)
[20,25)
[25,30)
[30,35)
[35,40)
[40,45)
[45,50)
[50,55)
[55,60)
[60,65)
[65,70)
[70,75)
[75,80)
[80,85)
[85,90)
[90,95)
[95,100)
[100,105)
[105,110)
[110,115)
[115,120]
Age category

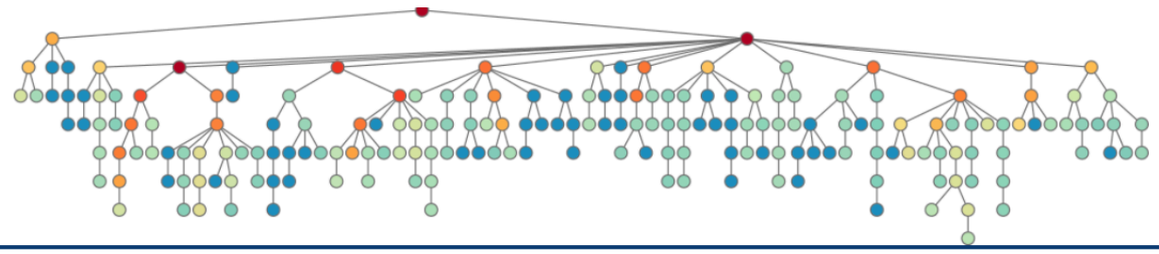

4e+06
3e+06
2e+06
1e+06
0e+00
n
[0,5)
[5,10)
[10,15)
[15,20)
[20,25)
[25,30)
[30,35)
[35,40)
[40,45)
[45,50)
[50,55)
[55,60)
[60,65)
[65,70)
[70,75)
[75,80)
[80,85)
[85,90)
[90,95)
[95,100)
[100,105)
[105,110)
[110,115)
[115,120]
Age category


4e+06
3e+06
2e+06
1e+06
0e+00
n
[0,5)
[5,10)
[10,15)
[15,20)
[20,25)
[25,30)
[30,35)
[35,40)
[40,45)
[45,50)
[50,55)
[55,60)
[60,65)
[65,70)
[70,75)
[75,80)
[80,85)
[85,90)
[90,95)
[95,100)
[100,105)
[105,110)
[110,115)
[115,120]
Age category

TATONETTI LAB RESEARCH AND RESOURCES

Tatonetti Lab

at Columbia University Medical Center



Research Courses Resources Positions Publications People ∞

Translational Medicine in the Age of Data

Billions of clinical measurements are recorded every day and stored in electronic health systems around the world. Each one of these *experiments* is a window into the human system, creating the most comprehensive and diverse medical data set ever imagined. Unfortunately, traditional statistical techniques were not developed to handle such diversity, instead they excel at analyzing homogenous data sets with first order effects. Because of this, these techniques are simply unable to untangle the sophisticated web of biological pathways and genetic interactions governing the human system.

With enormous data come enormous opportunity

Data Science is a new field dedicated to developing the methods, algorithms, and tools to unravel the complexities of enormous data. In our lab we advance data science by designing rigorous computational and mathematical methods that address the fundamental challenges of health data science. Foremost, we integrate our **medical observations** with **systems and chemical biology models** to not only explain drug effects, but also further our understanding of basic biology and human disease.

One particular area of interest is the integration of high-throughput data capture technologies, such as next-generation genome and transcriptome sequencing, metabolomics, and proteomics, with the electronic medical record to study the complex interplay between genetics, environment, and disease.

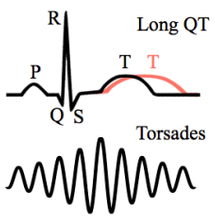
For a more in-depth information on our research areas of interest see our reviews in WIREs System Biology and Medicine, Science Translational Medicine, and Clinical Pharmacology & Therapeutics.

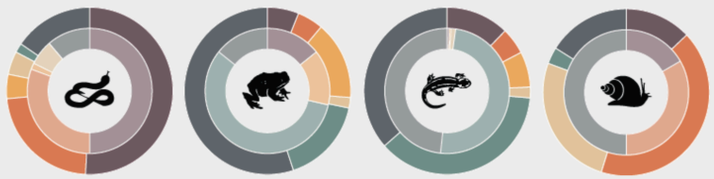
nSides

A comprehensive database of drug-drug(s)-effect relationships

ΔQT Database

ΔQTDb (delta QT Database) is a resource for exploring the effects of one or more drugs on the QT interval. It has been built from a deidentified subset of electrocardiogram and drug exposure data at Columbia University Medical Center.





Welcome to VenomKB

This is the home page for VenomKB, a centralized resource for discovering therapeutic uses for animal venoms and venom compounds

Browse Data Download

WHERE DOES THE MAGIC HAPPEN?

COMPUTING INFRASTRUCTURE

- 392 CPU cluster
- Over 5 TB RAM
- 450 TB storage
- GPU machines
- Data Platforms: Linux, Apache, MySQL, PHP…
- Programming languages: Java, C, Python, R, MATLAB, SAS…
- Server farms in NYC and NJ
- Human-Computer Interface, NLP, Text Mining, Clinical and Bioinformatics…

Our most important equipment...

YOURS TRULY, THE T-LAB!

- Rami Vanguri
- Kayla Quinnies
- Theresa Kolek
- Victor Nwankwo
- Yun Hao
- Joe Romano
- Phyllis Thangaraj
- Alexandre Yahi
- Fernanda Polubriaginof
- Julie Prost
- Jenna Kefeli
- Deidre Gregory

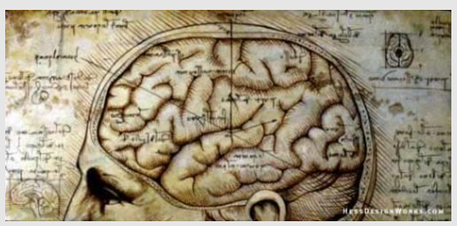
A night of good old-fashioned science

Featuring:

Debanjana Chatterjee
Bridging the world through science diplomacy

Amanda Buch
Music and Parkinson's Disease: A Novel Approach to Therapy

Maria Augusta Carrera-Haro
Immunopoetry: poems, prose and pictures from a stylin' immunologist



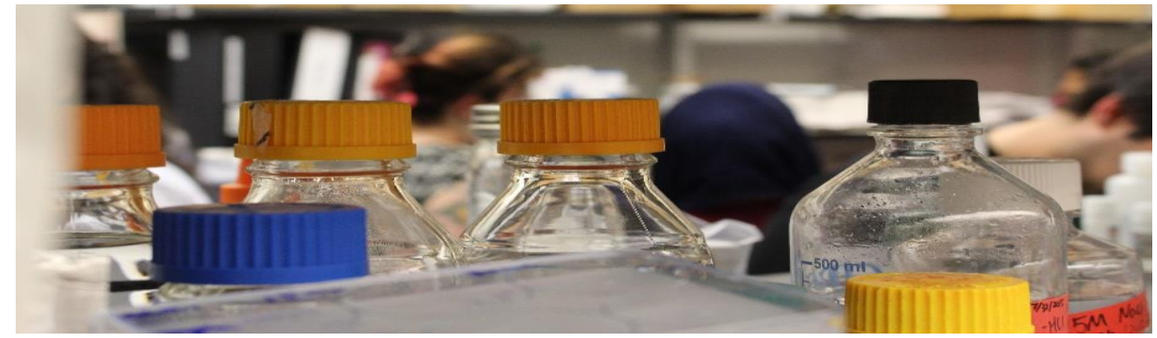
Joao Carlos Goncalves
The human brain: why is it different?

Ajit Muley
Lymphatic vessels: More than just drain-pipes

Open to all
Location:
Vagelos Education Center VEC 1203
104 Haven Ave, New York, NY 10032

Date:
December 11th
6pm to 8pm

Thanks to Columbia University Neuroscience Outreach (CUNO) and YALE CIENCA ACADEMY (YCA)



CIENCIA EN ESPAÑOL:

¿QUÉ OCURRE EN EL CEREBRO CUANDO SENTIMOS DOLOR?

Dr. Anita Burgos

Lugar: Jerome L. Greene Science Center, 3227 Broadway (Entre la 129th y 130th) Room L8.084

Día: Martes, 12 de Diciembre

Hora: 6:00 PM – 8:00 PM

¡Tour del Laboratorio Incluido!

Bebidas y aperitivo incluidos

¡Evento para todos los públicos, de cualquier edad!

Regístrese GRATIS en Eventbrite: "Ciencia en Español: ¿Qué ocurre en el cerebro cuando sentimos dolor?"

Si tiene alguna duda, puede contactar con Elena Carazo
ec2949@columbia.edu

Columbia University Medical Center

Columbia University Neuroscience Outreach