INTRO TO DATA SCIENCE FOR INNOVATION

NEXT FRONTIERS FOR SUSTAINABLE INNOVATION

AIKO YAMASHITA, PHD.
SENIOR DATA SCIENTIST - ADVANCED ANALYTICS, DNB
PROFESSOR II - OSLO METROPOLITAN UNIVERSITY

MODULE 1:

WHAT IS DATA SCIENCE AND WHAT PROBLEMS CAN SOLVE?

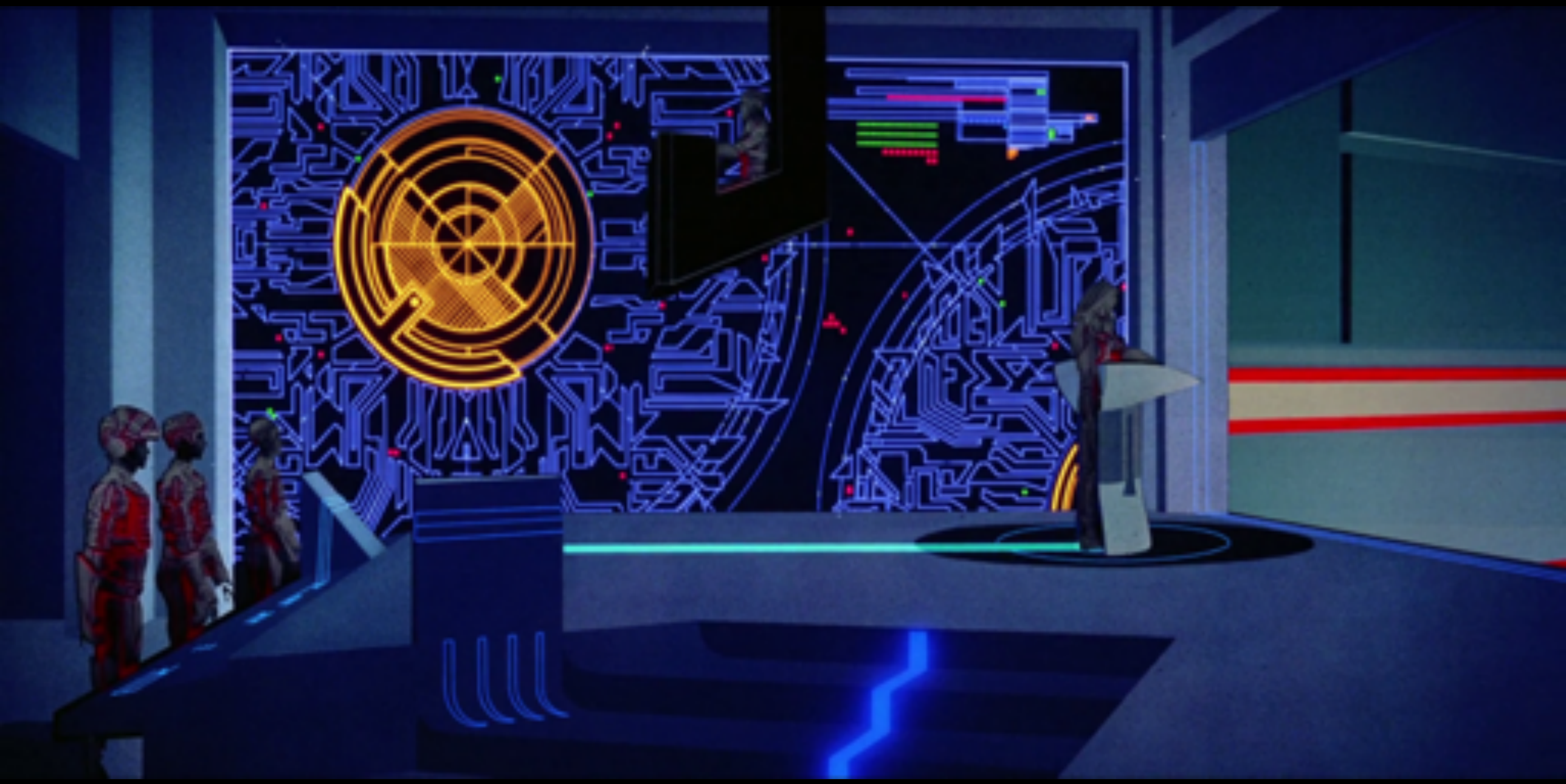
DATA SCIENCE, OR DATA-DRIVEN SCIENCE IS AN INTERDISCIPLINARY FIELD ABOUT SCIENTIFIC METHODS, PROCESSES, AND SYSTEMS TO EXTRACT **KNOWLEDGE** AND/OR **INSIGHTS** FROM **DATA**, WHICH MAY COME IN A STRUCTURED OR UNSTRUCTURED FORM.

IT CAN INVOLVE MANY FIELDS: STATISTICS, MACHINE LEARNING, INFORMATION SCIENCE, DATA VISUALIZATION AND MINING.
...BUT WHY IT HAS BECOME SO IMPORTANT?

Data is Everywhere

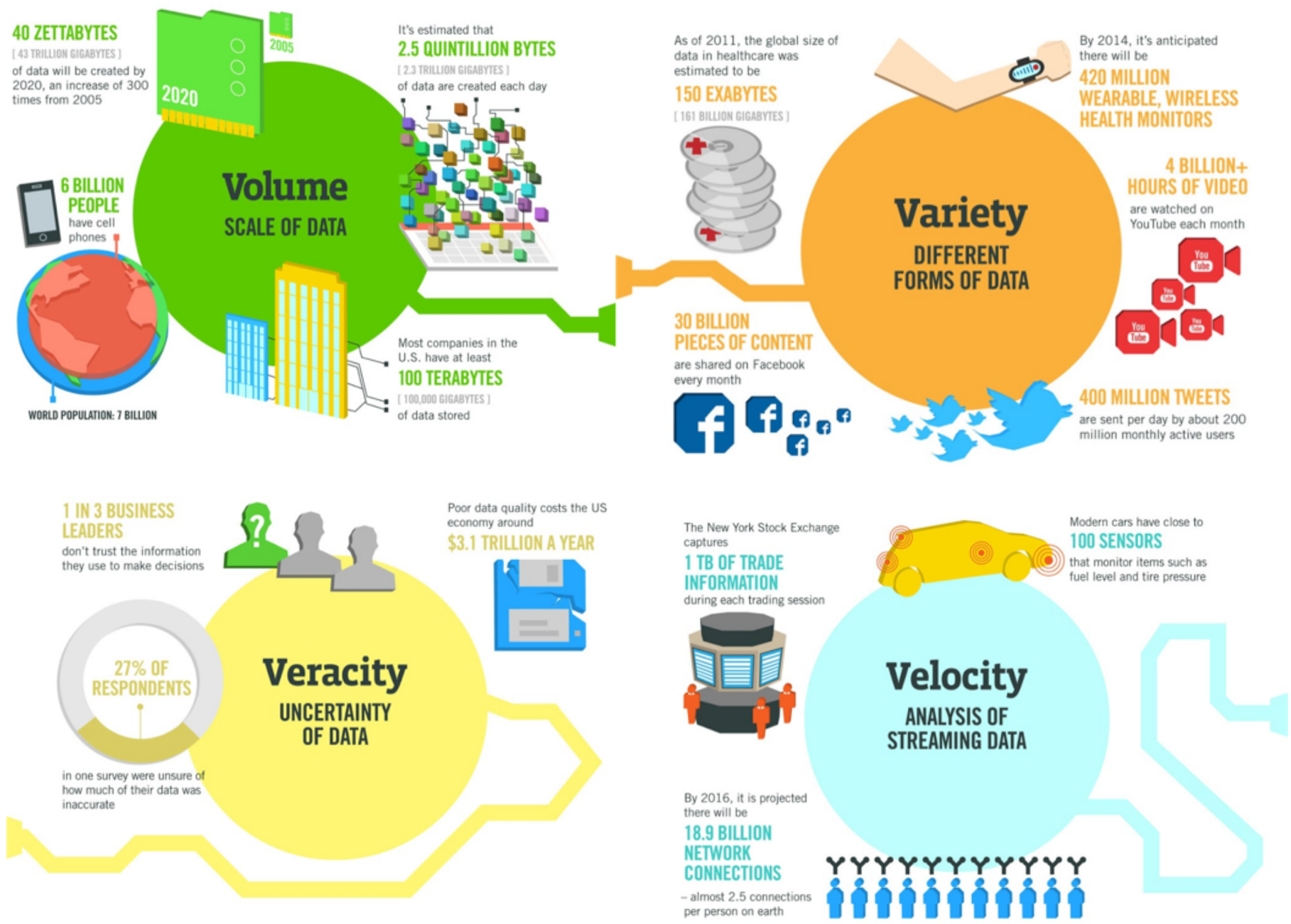
TRON (1982)

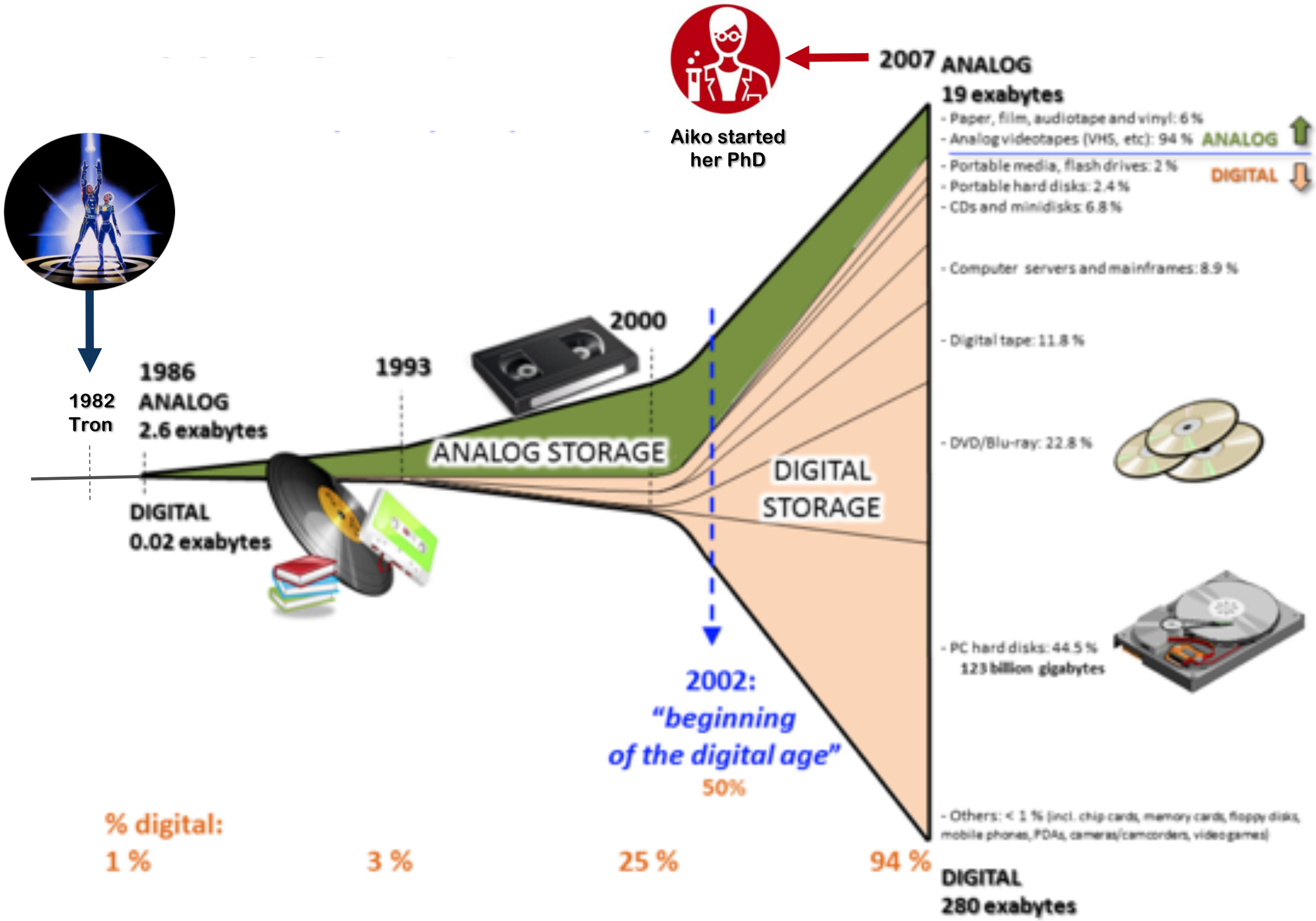
http://www.imdb.com/title/tt0084827/



This is how we visualized “Cyber-Cities” in the 80’s

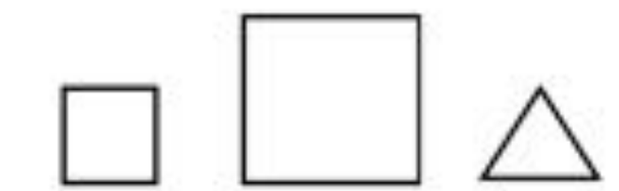
WHAT HAS CHANGED IN THE LAST 30 YEARS?



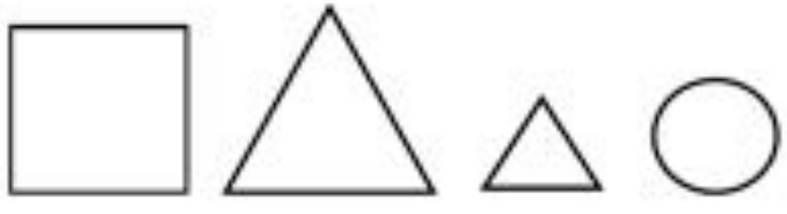

Aiko started her PhD
2007 ANALOG
19 exabytes
- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 %
ANALOG
DIGITAL
- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %
- PC hard disks: 44.5 %
123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)
DIGITAL
280 exabytes
2000
1993
1986
ANALOG
2.6 exabytes
1982
Tron
DIGITAL
0.02 exabytes
ANALOG STORAGE
DIGITAL STORAGE
2002:
"beginning of the digital age"
50%
% digital:
1 %
3 %
25 %
94 %

1.

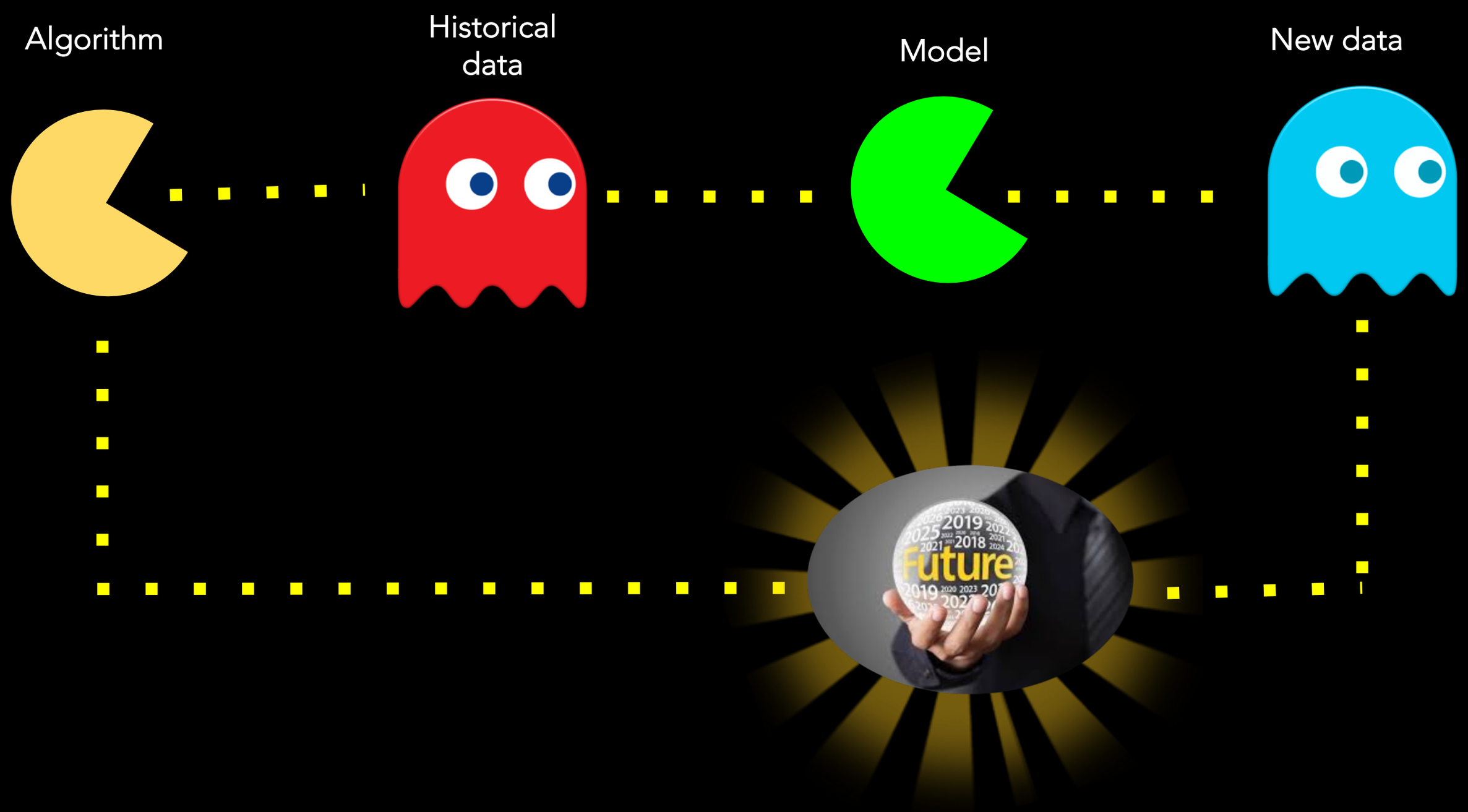
2.



3.

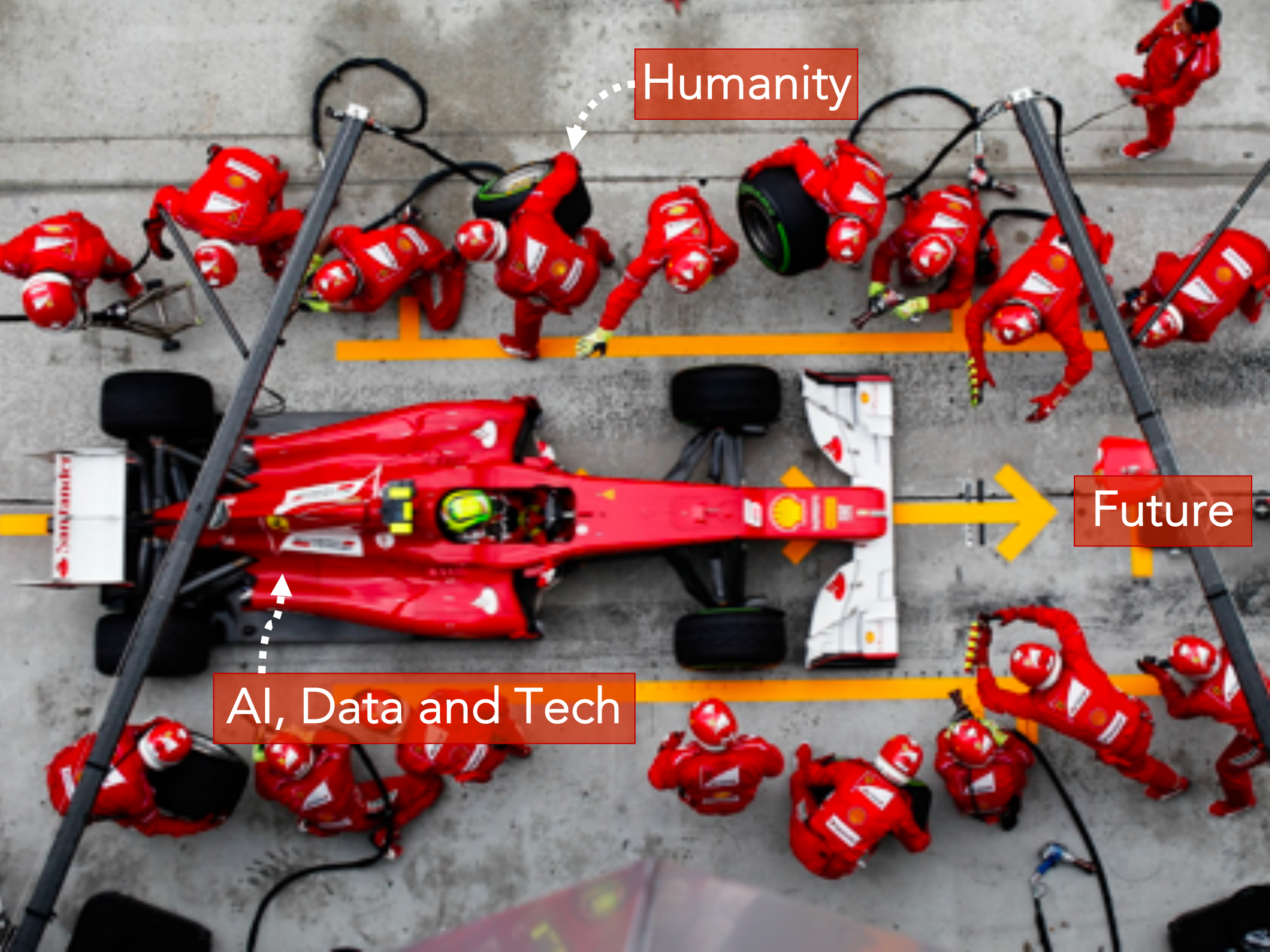


YOU 'TRAIN' YOUR ALGORITHMS WITH **DATA**, SO THEY CAN PREDICT BASED ON NEW VALUES..



ALPHAGO
02:08:18

NOVO 500
NOVO 500

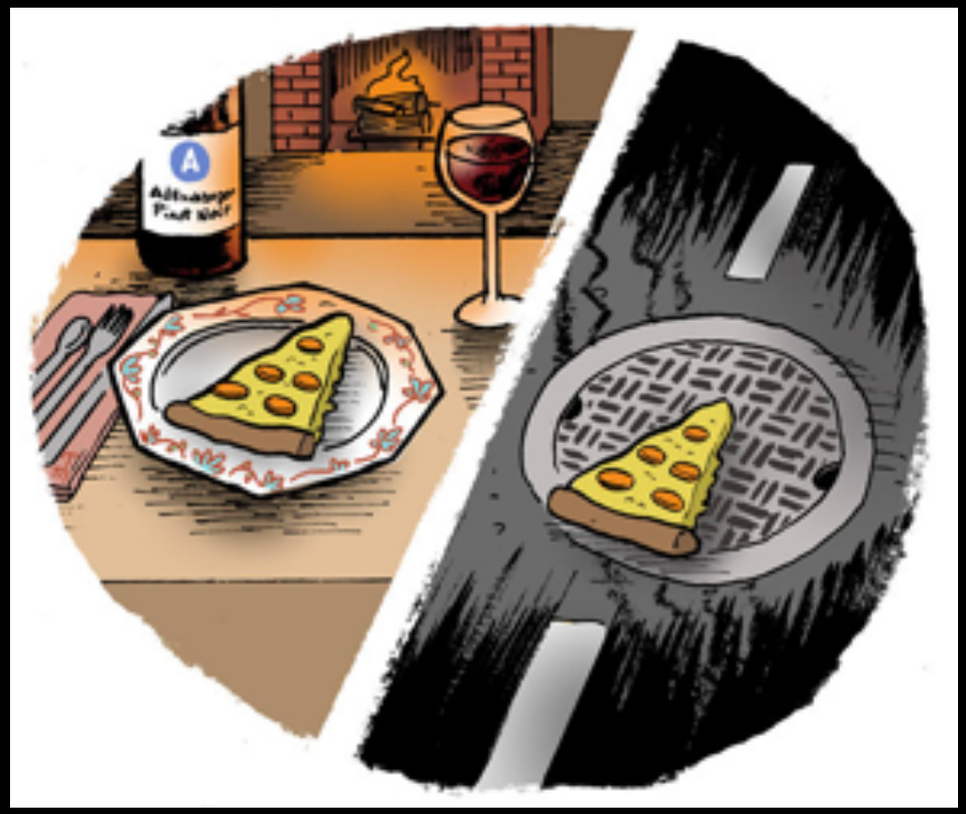

Humanity
Future
AI, Data and Tech

WARNING

WHAT IS THE ROLE OF HUMAN BEINGS IN THE NEXT ECONOMY?

Empaty - Compassion

Creativity



Understanding the context

Determining direction / purpose

Intended effects..

theguardian

The Observer

The great British Brexit robbery: how our democracy was hijacked

A shadowy global operation involving big data, billionaire friends of Trump and the disparate forces of the Leave campaign influenced the result of the EU

Zuckerberg put on back foot as House grills Facebook CEO over user tracking

In his second day on Capitol Hill, Zuckerberg says his own data was handed to Cambridge Analytica

Five things we learned from Mark Zuckerberg's Facebook hearing

▲ Zuckerberg faces Congress: the biggest highlights from day two – video

626 MILLION CAMERAS
1.4 BILLION PEOPLE
1 PER 2 PEOPLE RATIO

NO MATTEL, MY CHILDREN ARE NOT "ASSETS"

MATTEL: "WE USE THE MOST ADVANCED AI-DRIVEN TECHNOLOGY TO MAKE IT EASIER FOR PARENTS TO PROTECT, DEVELOP, AND NURTURE THE MOST IMPORTANT ASSET IN THEIR HOME — THEIR CHILDREN."

A RACE FOR HUMAN ATTENTION

In 2017, People tapped, swiped and clicked a whopping 2,617 times each day, on average

Unintended effects..

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler and supporting Donald Trump

▲ Tay uses a combination of artificial intelligence and editorial written by a team including improvisional comedians. Photograph: Twitter

A bit rounded.. Perhaps..
Heh.. Let's move on...

Uhm...

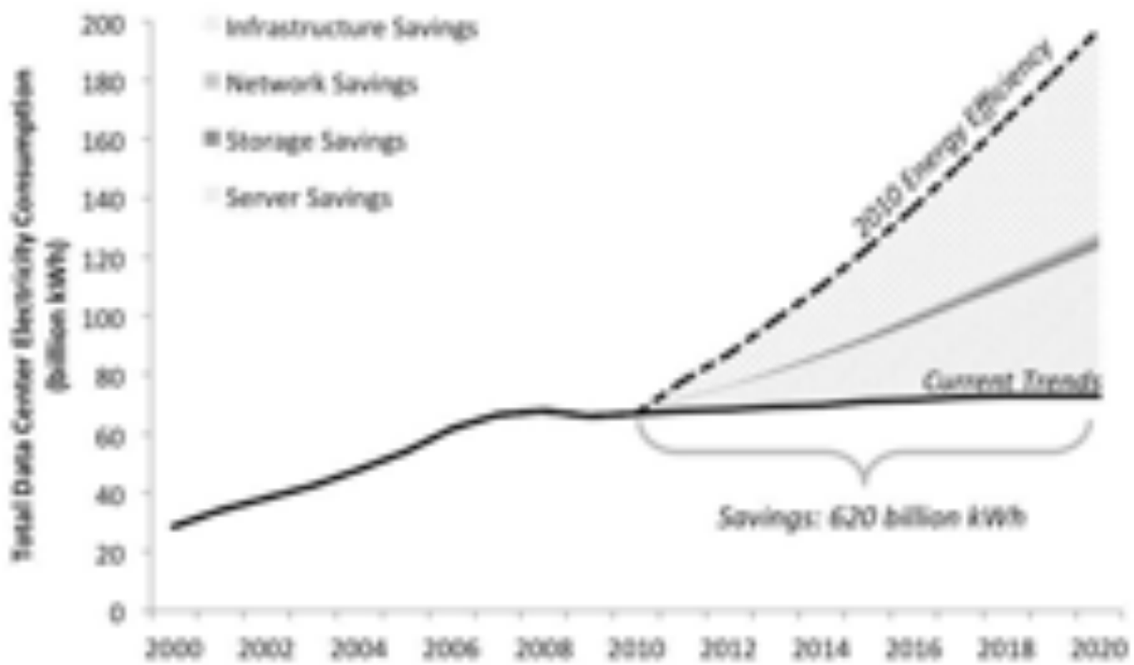
?!?!
Back to the drawing board...

Algorithms are stupid..

JUNE 9, 2019 WEBLOG

Researchers show glare of energy consumption in the name of deep learning

by Nancy Cohen , Tech Xplore



TECHNOLOGY

The Environmental Toll of a Netflix Binge

The data centers that support the Internet use a huge amount of energy.

INGRID BURRINGTON DEC 16, 2015

And resource consuming!

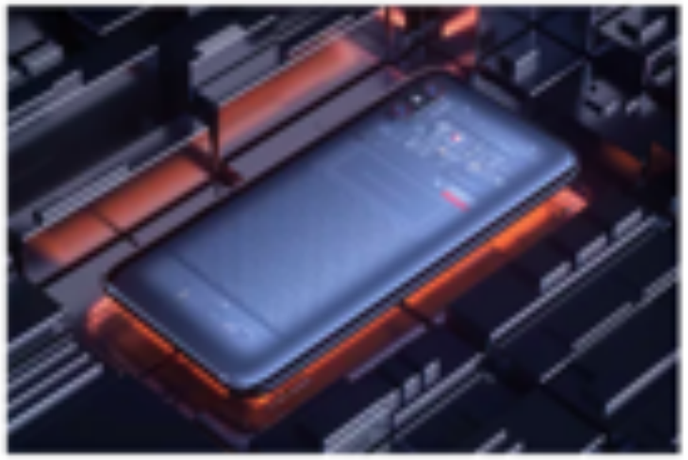
What if we don’t take the wheel?

New survey finds Americans' trust in Facebook continues to decline

By Shannon Liao | @Shannon_Liao | Mar 26, 2018, 10:58am EDT

SHARE

MOST READ



Xiaomi announces Mi 8 flagship phone, transparent Explorer Edition, and smaller SE model

NORWAY IS NOT AN EXCEPTION

Nyheter

Nye personvernregler

Skeptiske forbrukere kan gi trøbbel for norske bedrifter

Fire av ti nordmenn svarer i en undersøkelse at de ikke ønsker å gi fra seg persondata til ulike

skritt på å trykke «enig» til uforståelige vilkår med liten skrift.

«Personvernet er klart viktigst»

Det kan gi norske bedrifter en utfordring. Ifølge en ferisk undersøkelse gjennomført

PwCs personvern-undersøkelse

Alle spørsmålene i undersøkelsen er stilt under forutsetninger om at personene er kjent

Blant respondentene i undersøkelsen som er mer positive til å utlevere personlige data, oppgir 61 prosent å være villige til å gi fra seg stedsdata. Svaret få sier derimot seg villig til å gi fra seg data ved nettbesøk, søkt data fra sletvisningsmåtere som treningsapper

negativt. En virksomhetens kunder ikke ønsker å dele personlig data.

– En mulig konsekvens kan være at virksomheten mister kundeforhold og taper omsetning, fordi man ikke vil gi ut de opplysninger man ønsker, sier Sæ.

TRUST IS A CURRENCY

THIS IS FINE.

How to set the direction?

Set People First

Think Diversity

Think Sustainable

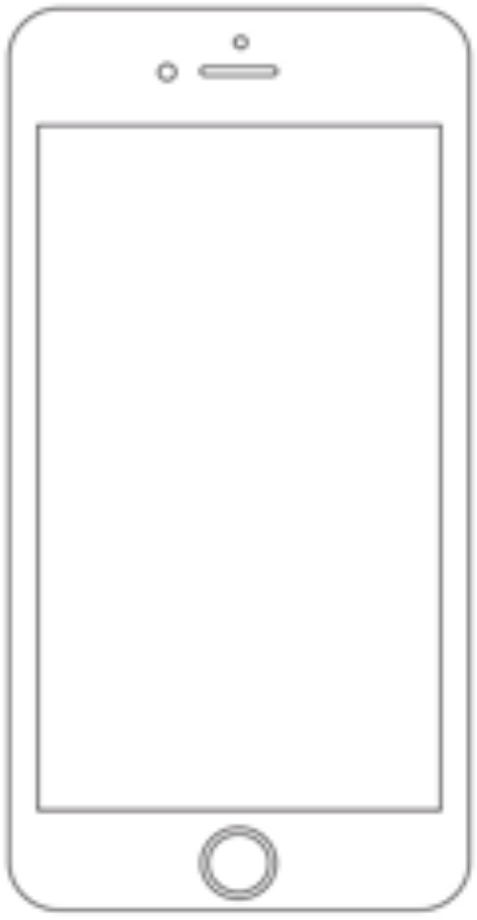
Think Global, Test Locally

Set People First

What is best for People?

With every design choice, our products should be about what's best for people, not what's best for grabbing eyeballs.

[Center for Humane Technology]



Devices

- What if we designed devices to help us disconnect without missing something important?
- What if we designed devices to help us fall asleep on our schedule?
- What if we designed devices for quick in-and-out uses, not endless interactions?

Home Screens & Browsers

- What if we designed home screens to promote off-screen choices, not just on-screen choices?

Social Media

- What if we designed social media to reduce loneliness–and made it easier to coordinate with others?

News Feeds

- What if we designed news feeds to reward the most in-depth reporting, not most clicks?

Think Diversity (at All Levels)

The more diverse our trained thinking patterns are, the more accurately we will be able to interact with information around us.

FEATURED STORIES

It's Not What You Know, It's How You Think

Understanding our pattern-seeking mind is the first step toward real wisdom

Zat Rana Follow
Sep 15, 2018 · 6 min read ★

Photo: Gerd Altmann/Pixabay

Data comes with biases

Social data definitively comes with biases

Products and Algorithms behind them and can also come with biases inherent in those who are making the design decisions!

DATA FOR BLACK LIVES CONFERENCE

"…designers get to define the **objective** and the "neutral".

And that definition never has to be articulated, just understood.

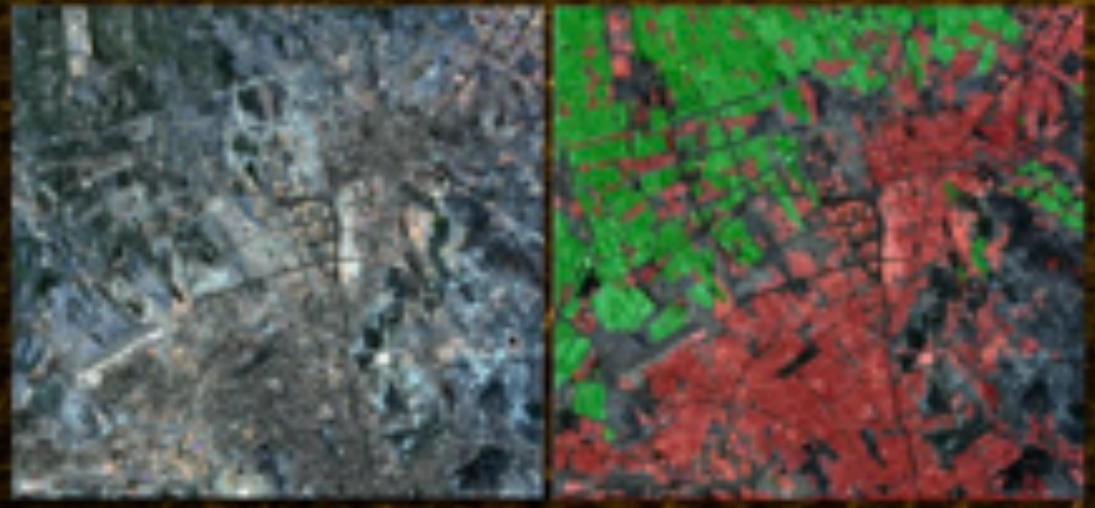
And we have come to understand **objective** to be associated with white and male…"

PREGNANT WOMAN SUFFERS NECK FRACTURE.
[Age: 34 years | Speed: 60 km/h | Accident: Side Collision]

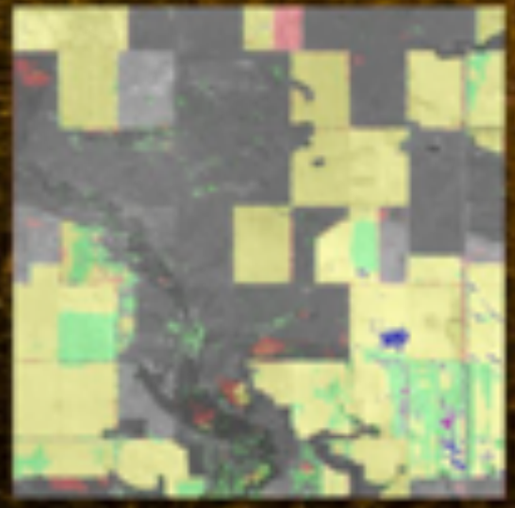
Partly because male crash test
dummies are the standard.

Think Sustainable

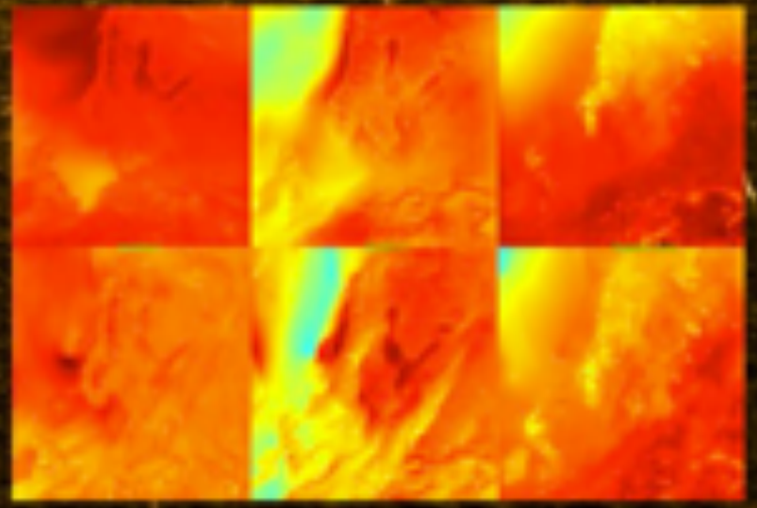
AI HAS THE POTENTIAL TO INCREASE YIELDS FOR 540MILLION SMALL FARMS, POTENTIALLY IMPACTING THE GLOBAL ECONOMY



Farm land segmentation



Crop type identification



Weather prediction

50% drop in error margin comp to current models

Can greatly improve Ignitia's model

Think Global, Test Locally

Where you want to go?

WORKSHOP 1.1: LET'S DEFINE A PROBLEM THAT CAN BE SOLVED BY DATA SCIENCE

YOU HAVE TO ASK THE RIGHT QUESTIONS!

From a marketing perspective, you may ask..

1. **Who** are the customers?
2. **Why** are they buying our product?
3. **How do we predict** if a customer is going to buy our product?
4. What is **different** from segments who are performing well and those that are performing below expectations?
5. **How much money** will we lose if we don't actively sell the product to these groups?

YOU HAVE TO ASK THE RIGHT QUESTIONS!

Let's make it more general and relevant..

1. **Who** are the implicated subjects/objects?
   - What is their behavior/situation?
2. **Why** they have such behavior/situation?
3. **How can we predict** if certain subjects/objects are going to have such behavior/situation?
4. What are the **differences** between groups of subjects/objects having different behavior/situation?
5. What are the **implications/impact** of a change in their behavior/situation?

WORKSHOP 1.2: LET'S BE CRITICAL ON POSSIBLE POSITIVE/NEGATIVE CONSEQUENCES OF AI

MODULE 2:
DATA TYPES AND SOURCES

WHAT TYPES OF DATA (VARIABLES) EXIST?

Categorical

Nominal:

- Gender: Female/Male
- Drug is effective: Yes/No

Ordinal:

- Size: Small, Medium, Large, XL
- Credit: Bronze, Silver, Gold

Numerical (Quantitative)

Absolute Scale: The count the number of elements in the entity set.

- Weight
- Distance
- Money
- Anything you can count :)

WE CAN SIMPLIFY IT..

- **Categorical** ("Labels" under which a value may fall)
  - Nominal
  - Ordinal

- **Continuous** (Quantifiable types)
  - Interval
  - Ratio
  - Absolute

WHAT DATA SOURCES EXIST?

- **Structured**: Relational Databases
- **Semi-structured**: File-based with a certain amount of structure (CSV, XML, JSON, NoSQL)
- **Unstructured**: text, multimedia

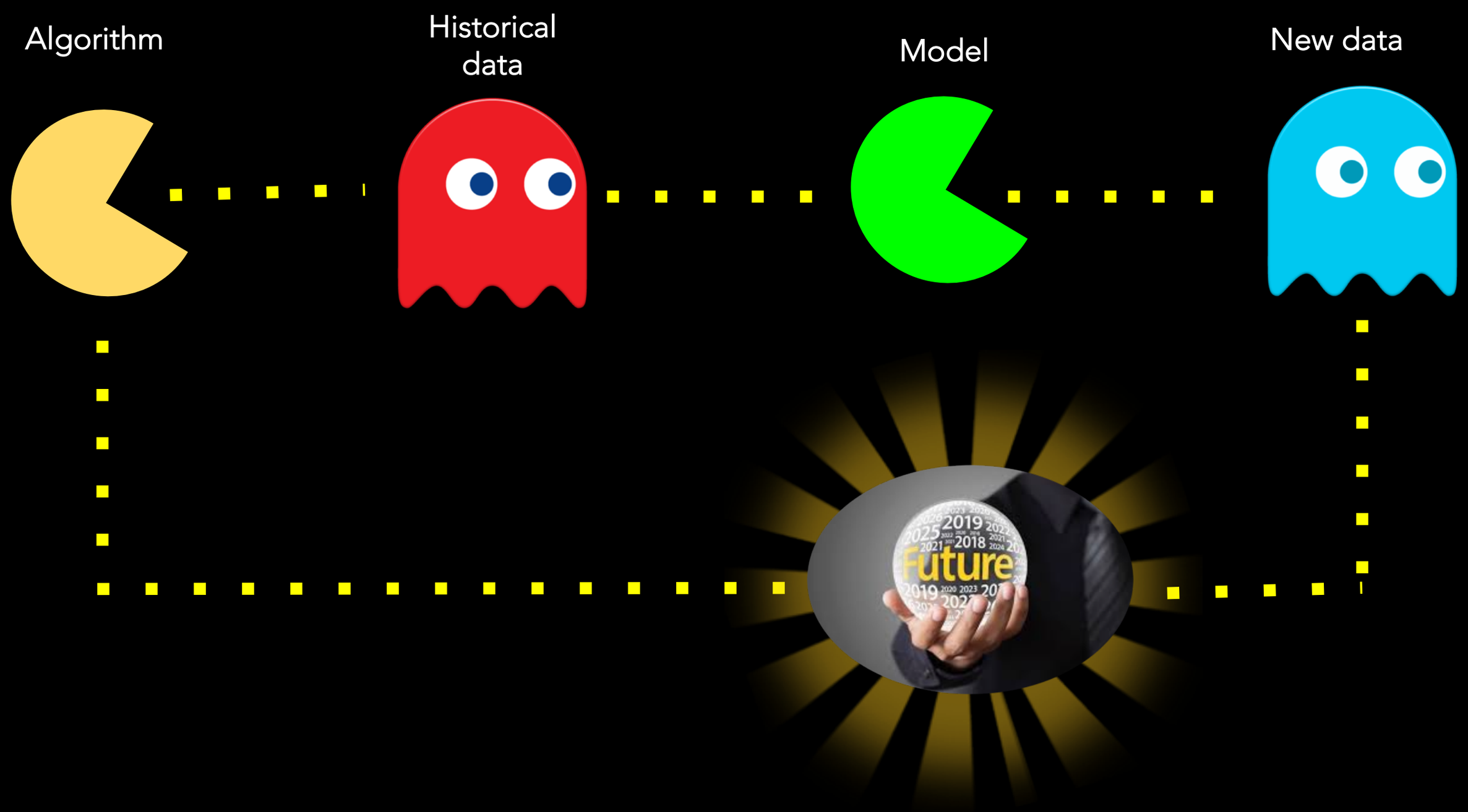
WORKSHOP 2: LOOKING AT DATA WITH PYTHON

https://tinyurl.com/dnb-grad

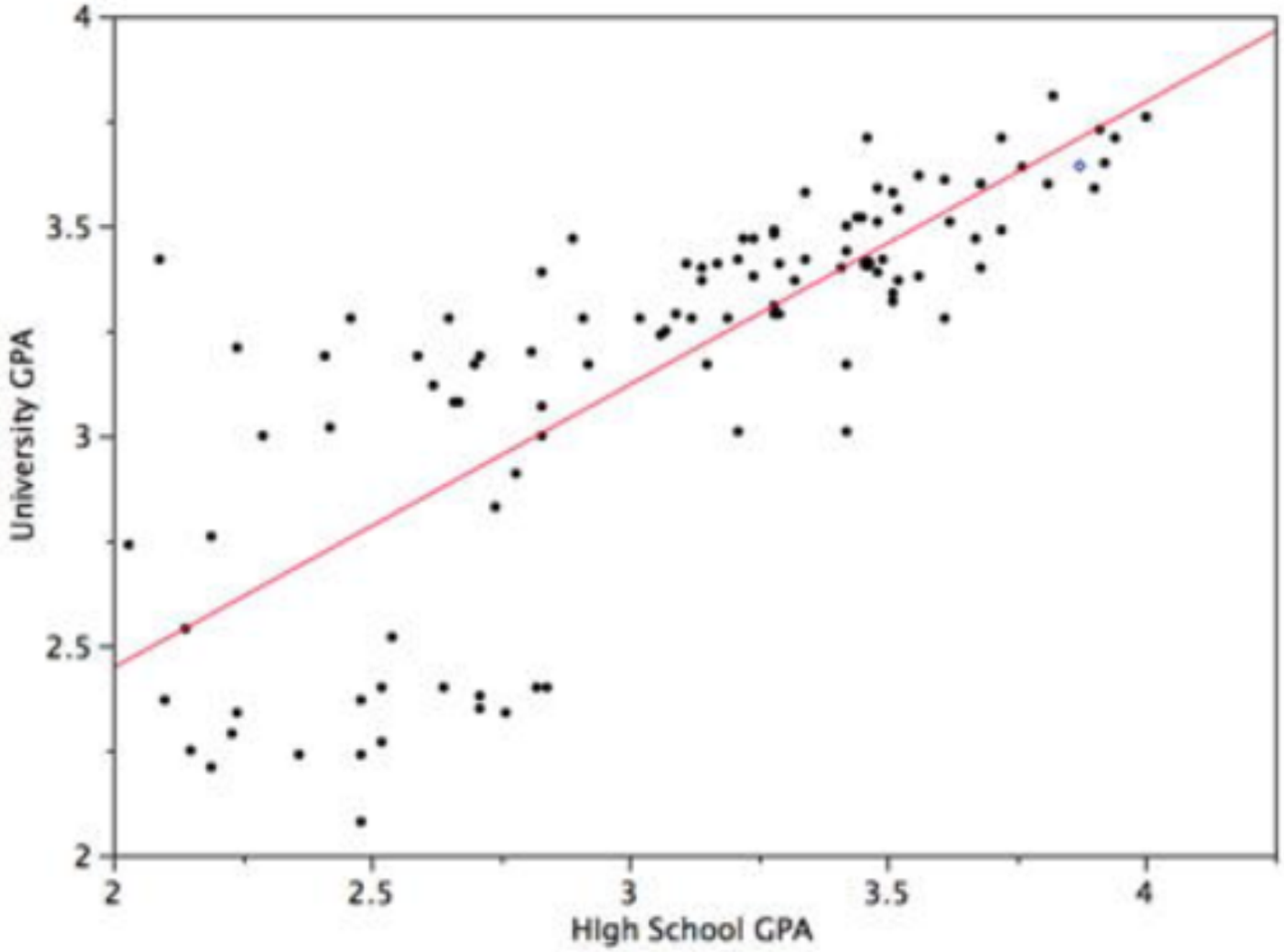
```
In [ ]: import statsmodels.api as sm

In [ ]: from sklearn import datasets ## imports datasets from scikit-learn
        data = datasets.load_boston() ## loads Boston dataset from datasets library

In [ ]: print(data.DESCR)
```

MODULE 3:
TO ML ALGORITHMS: HOW DO THEY WORK?

YOU 'TRAIN' YOUR ALGORITHMS WITH **DATA**, SO THEY CAN PREDICT BASED ON NEW VALUES..



ALGORITHMS IDENTIFY PATTERNS FOR DOING PREDICTIONS..



DIFFERENT ML ALGORITHMS ARE USED DEPENDING OF THE TYPE OF DATA YOU WANT TO PREDICT

Predicting numerical values → Linear Regression

Predicting Categories → Logistic Regression

ALGORITHMS GALORE...

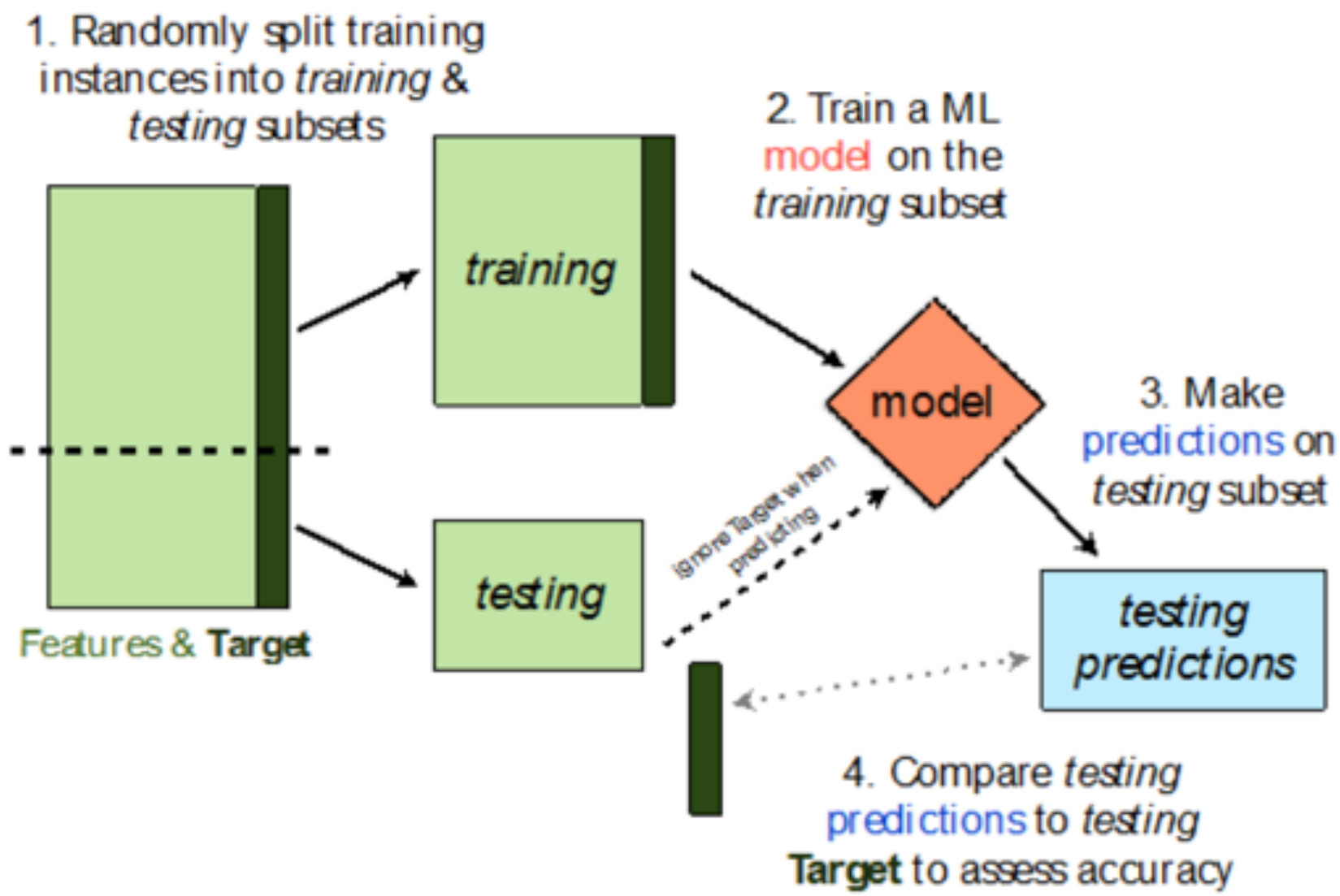
WHEN YOU ARE PREDICTING SOMETHING, YOU NEED TO HAVE VARIABLES THAT ARE PREDICTORS, BUT MOST IMPORTANTLY..

YOU NEED THE PREVIOUS VALUES OF WHAT YOU WANT TO PREDICT!

Target

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q

YOU NORMALLY SPLIT THE DATA IN TWO: ONE FOR TRAINING AND ONE FOR <u>TESTING</u> YOUR MODEL...



WORKSHOP 3.1: HOW TO TRAIN AN ALGORITHM TO CREATE A MODEL

https://tinyurl.com/dnb-grad

Let's define the predictors

In [ ]:
```
boston_predictors = pd.DataFrame(data.data, columns=data.feature_names)
boston_predictors.head()
```

Let's define the target

In [ ]:
```
# Put the target (housing value -- MEDV) in another DataFrame
boston_target = pd.DataFrame(data.target, columns=["MEDV"])
boston_target.head()
```

Let's start with one predictor..

In [ ]:
```
one_predictor = boston_predictors[['LSTAT']]
boston_target = boston_target['MEDV']
```

Let's split into training and testing sets

In [ ]:
```
# Training set: everything except the last 20
one_predictor_train = one_predictor[:-20]
boston_target_train = boston_target[:-20]

# Testing set: the last 20 for testing our model
one_predictor_test = one_predictor[-20:]
boston_target_test = boston_target[-20:]
```

In [ ]:
```
# Create linear regression object
lm = linear_model.LinearRegression()

# Train the model using the training sets
model = lm.fit(one_predictor_train, boston_target_train)
```

In [ ]:
```
# This is thre R-squared of the model
lm.score(one_predictor_train, boston_target_train)
```

WORKSHOP 3.2: HOW TO MAKE A PREDICTION BY USING A CREATED MODEL

https://tinyurl.com/dnb-grad

```
In [ ]: predictions = lm.predict(one_predictor_test)

In [ ]: pred_vs_actual = pd.DataFrame({'prediction':predictions, 'actual':boston_target_test, 'difference': abs(predic

In [ ]: pred_vs_actual

In [ ]: x_axis = one_predictor_test[['LSTAT']].values.flatten()

In [ ]: # Plot data, regression line
        plt.figure(2)
        plt.scatter(x_axis, predictions, color='r')
        plt.scatter(x_axis, boston_target_test, color='b')
        plt.show()
```

MODULE 4:

HOW GOOD ARE OUR PREDICTIVE MODELS?

CONFUSION MATRIX

n=192	Predicted: 0	Predicted: 1
Actual: 0	118	12
Actual: 1	47	15

TERMINOLOGY RELATED TO CONFUSION MATRIX

- **True Positives** (TP): we correctly predicted that they do have diabetes
- **True Negatives** (TN): we correctly predicted that they don't have diabetes
- **False Positives** (FP): we incorrectly predicted that they do have diabetes (a "Type I error") -> "**false alarm**"
- **False Negatives** (FN): we incorrectly predicted that they don't have diabetes (a "Type II error") -> "**alarm failed**"

WORKSHOP 4: LET'S COMPARE THE RESULTS FROM OUR MODEL TO REAL DATA

https://tinyurl.com/dnb-grad

Predicting probabilities of something happening: Logistic Regression

In [ ]:
```
import matplotlib
import numpy as np
import pylab as plab
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
from sklearn import datasets, linear_model
```

In [ ]:
```
titanic_train = pd.read_csv("titanic_train.csv")
titanic_train.head()
```

In [ ]:
```
#Clean the data for missing values (NaN) in Cabin
char_cabin = titanic_train["Cabin"].astype(str)     # Convert cabin to str
new_Cabin = np.array([cabin[0] for cabin in char_cabin]) # Take first letter
titanic_train["Cabin"] = pd.Categorical(new_Cabin)  # Save the new cabin var
```

In [ ]:
```
# Impute median Age (28) for missing (NA) Age values
new_age_var = np.where(titanic_train["Age"].isnull(), # Logical check
                       28,                            # Value if check is true
                       titanic_train["Age"])          # Value if check is false

titanic_train["Age"] = new_age_var
```

Use the Logistic Regression Algorithm to train the model..

In [ ]:
```
from sklearn import linear_model
from sklearn import preprocessing

# Initialize label encoder
label_encoder = preprocessing.LabelEncoder()

# Convert Sex variable to numeric
encoded_sex = label_encoder.fit_transform(titanic_train["Sex"])
```

In [ ]:
```
# Initialize logistic regression model
log_model = linear_model.LogisticRegression()

# Train the model
log_model.fit(X = pd.DataFrame(encoded_sex),
              y = titanic_train["Survived"])
```

In [ ]:
```
predictions = pd.Series(log_model.predict(X=pd.DataFrame(encoded_sex)))
predictions.name = "Predicted"

# Generate table of predictions vs actual
pd.crosstab(predictions,titanic_train["Survived"])
```

In [ ]:
```
# Generate table of predictions vs Sex
pd.crosstab(titanic_train["Sex"], predictions)
```

In [ ]:
```
# Convert more variables to numeric
encoded_class = label_encoder.fit_transform(titanic_train["Pclass"])
encoded_cabin = label_encoder.fit_transform(titanic_train["Cabin"])
```

In [ ]:
```
train_features = pd.DataFrame([encoded_class,
                               encoded_cabin,
                               encoded_sex,
                               titanic_train["Age"]]).T

# Initialize logistic regression model
log_model = linear_model.LogisticRegression()

# Train the model
log_model.fit(X = train_features ,
              y = titanic_train["Survived"])
```

In [ ]:
```
# Make predictions
preds = pd.Series(log_model.predict(X=train_features))
preds.name = "Predicted"

# Generate table of predictions vs actual
pd.crosstab(preds,titanic_train["Survived"])
```

In [ ]:
```
# Generate table of predictions vs Sex
pd.crosstab(titanic_train["Sex"], preds)
```

In [ ]:
```
# Generate table of predictions vs Sex
pd.crosstab(titanic_train["Sex"], titanic_train["Survived"])
```

FURTHER READING

Here are some resources:

- Crash course on statistics for research
- How to get started in Python
- Intro to Python for Data Science
- YouTube channel: https://tinyurl.com/datascience-iknow
- https://medium.com/towards-data-science