

Big Data for SAS programmers – Hadoop, NoSQL database and Machine Learning

Kevin Lee
Director of Data Science

The Agenda

- The Data Trend
- Big Data Storage Solution - Hadoop and NoSQL Database
- Machine Learning – Supervised, Unsupervised and Deep Neural Network
- Questions and Discussion



Q1 : World's Largest Transportation Company?



In December 2015, there were about **162,037** “active drivers”.

The number of new drivers signing up has **doubled every six months** for the past two years.

Q2: World's largest accommodation provider?



- Inc's 2014 Company of the Year
- In 2014 alone, 10 million people used airbnb. It has more the 800,000 listings worldwide.

Common Characteristics of Exponential Organization

Exponential &
Scalable Growth

➤ Data (Big)



➤ Algorithm



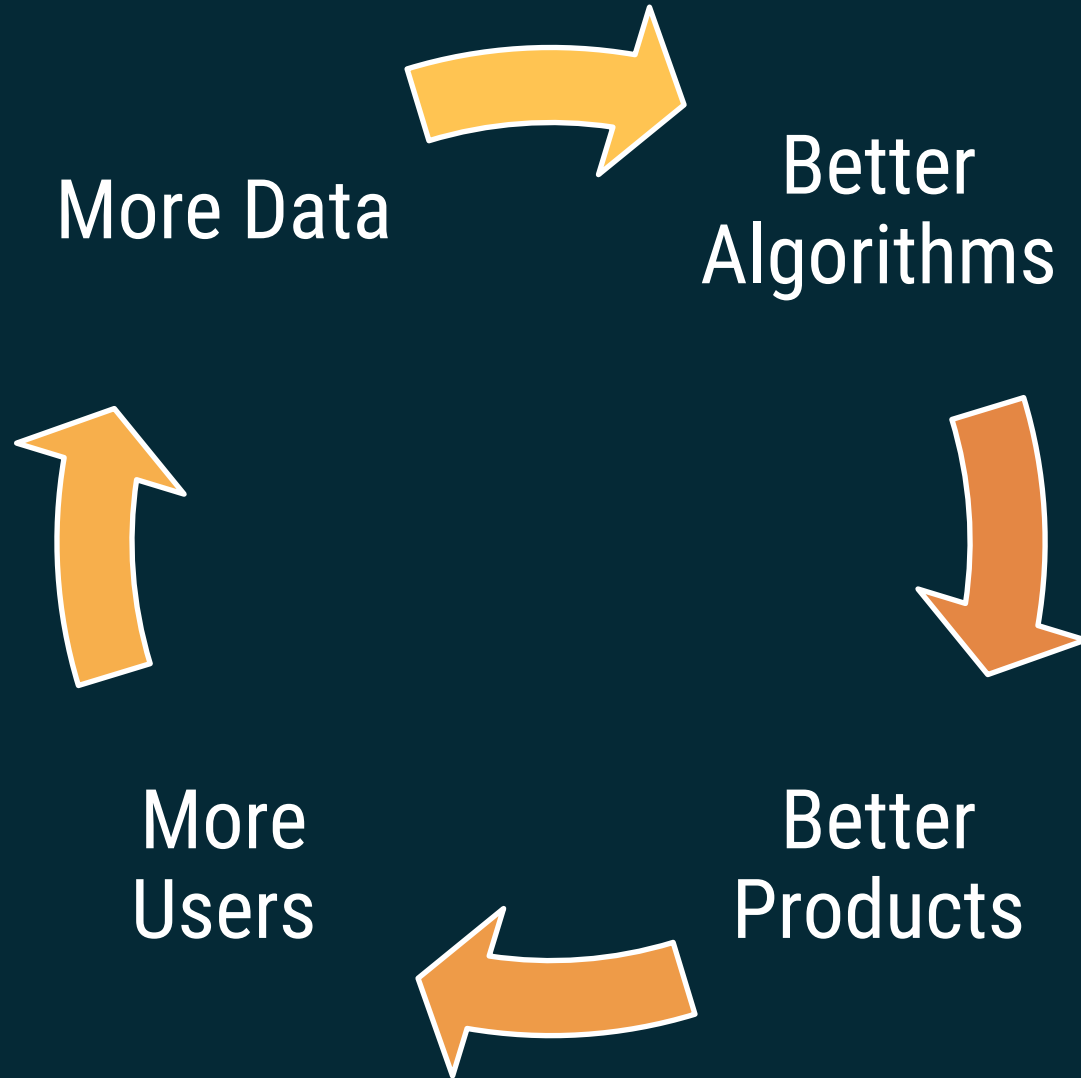
How Big Data will **help** the industry?



- When a customer is ready to buy.
- Engine needs service.
- A person is at a risk of a disease.

How will Big Data **help**?

7





The Economist - The world's most valuable resource is no longer oil, but DATA.

VOLUME



3 V's of current Data Trend

Volume

Velocity

Variety



Big Data Storage Solution

- Hadoop
- NoSQL



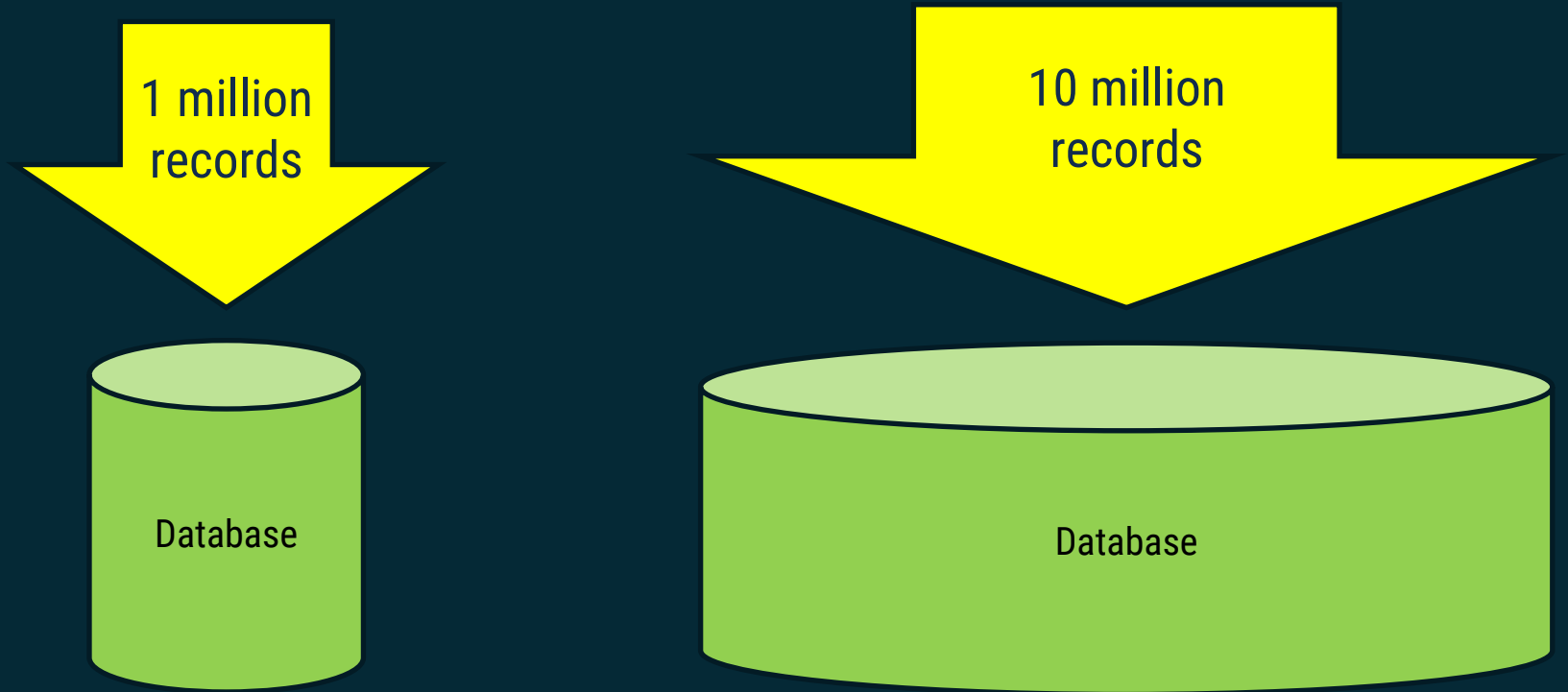
Apache Hadoop



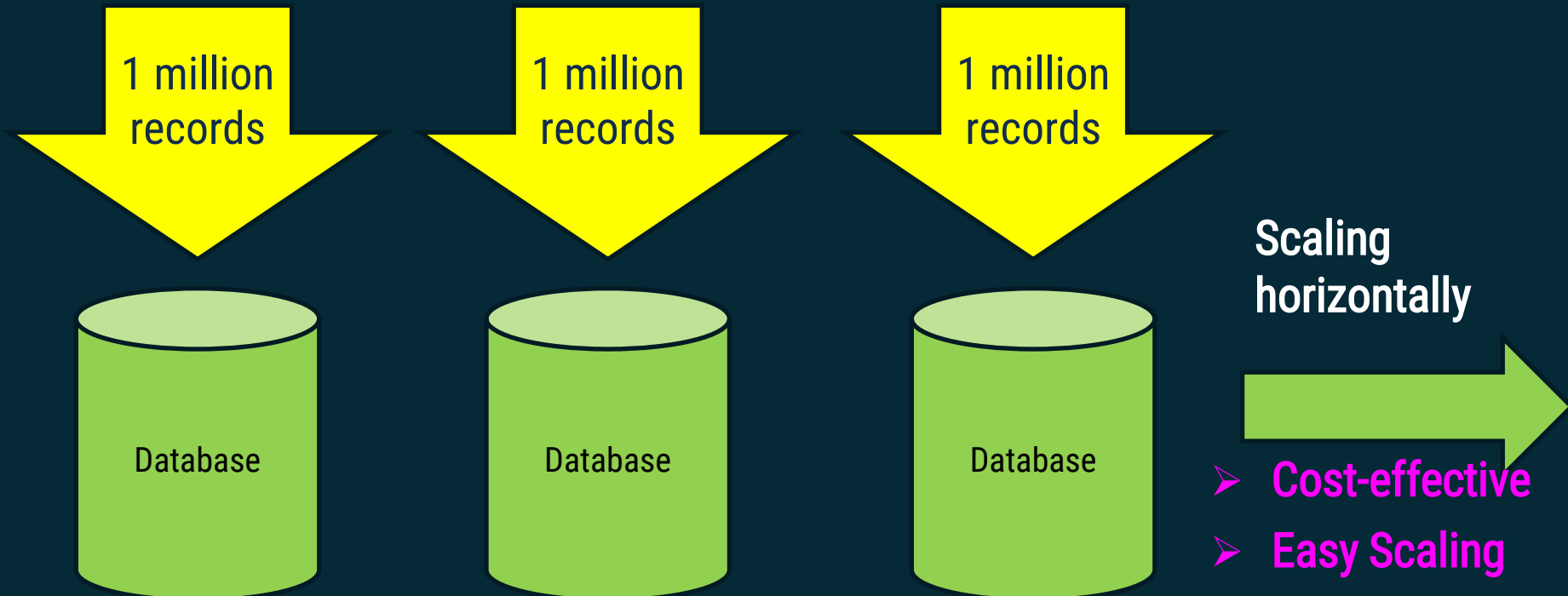
- Hadoop Distributed File Systems (HDFS)
- Map/Reduce

- Open source of distributed storage system with parallel-processing approach.
- Provide the storage / computing power for Big Data.

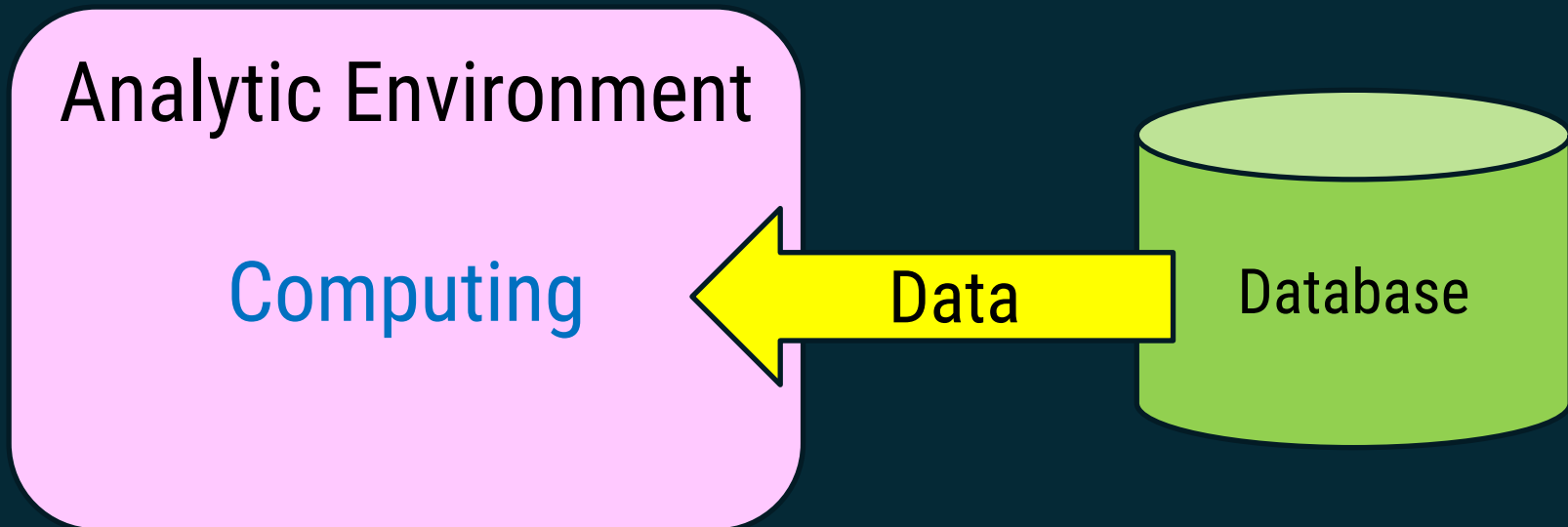
Typical database scaling



HDFS (Hadoop Distributed File System)

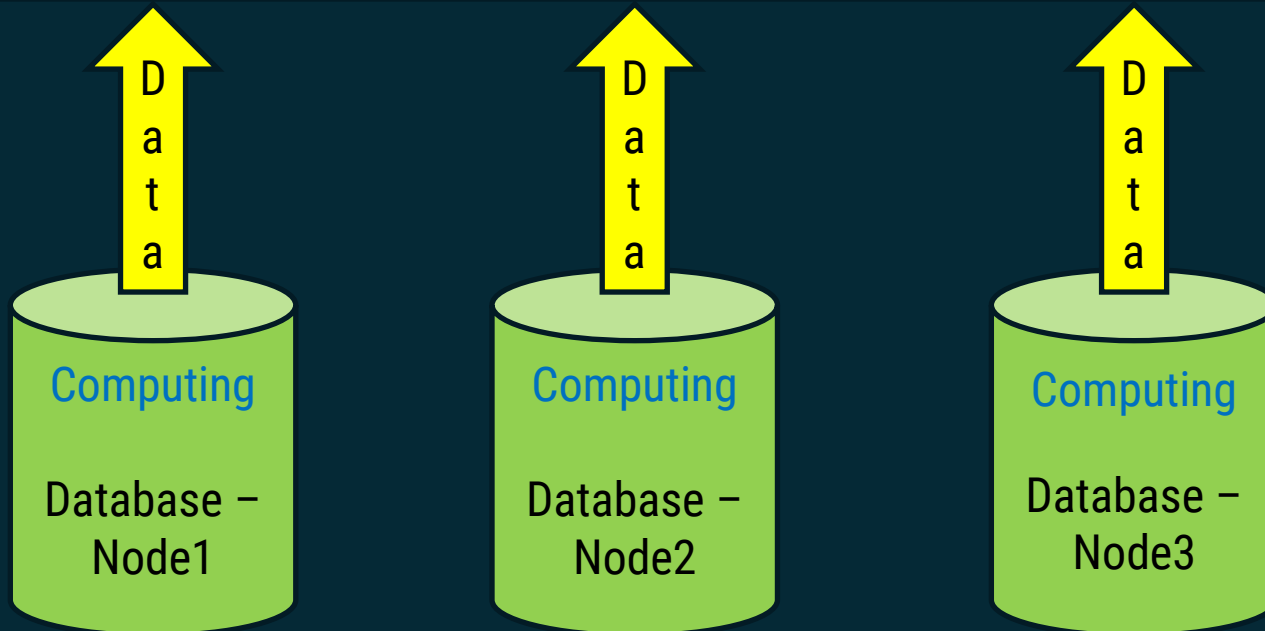


Typical analysis using database



Parallel computing in Hadoop – Map/Reduce

Analytic Environment
Computing

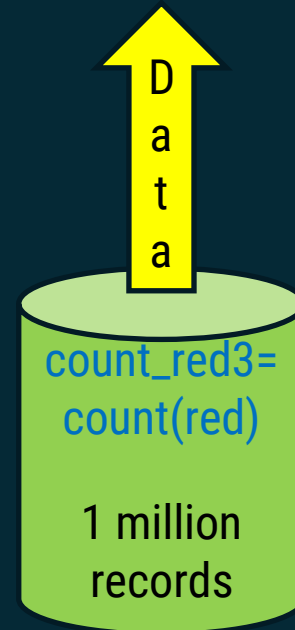
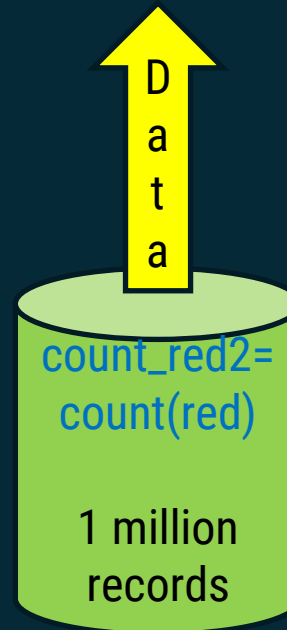
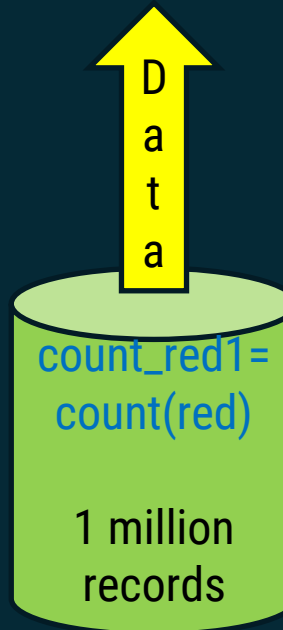
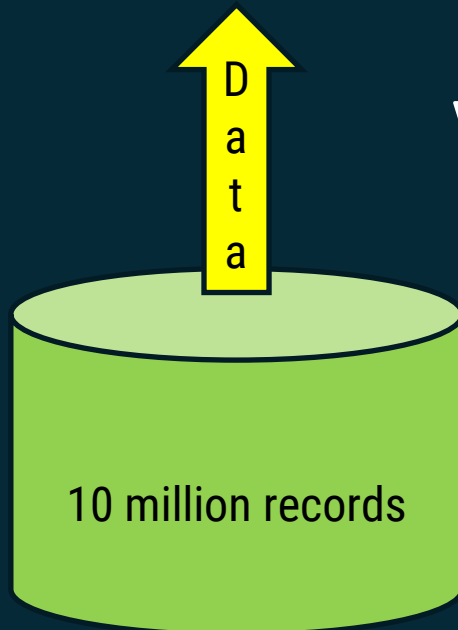


Parallel computing in Hadoop – Use Case

```
count_red =  
count(red)
```

```
count_red = count(count_red1, count_red2,  
count_red3,,,, count_red10)
```

Vs.



10
nodes

SAS integration to Hadoop

➤ Proc Hadoop

```
PROC HADOOP USERNAME="phuse_sde_bigdata";  
  HDFS COPYFROMLOCAL='c:/study01/sdtm/dm.xpt'  
  OUT='/study01/sdtm/dm.xpt';  
RUN;
```

➤ Libname to HDFS

```
LIBNAME mydata HADOOP  
SERVER='server.phuse_sde_bigdata.sas.com' SUBPROTOCOL=hive2  
PORT=1000 USER="phuse" PW="phuse";  
PROC MEANS data=mydata.dm;  
  VAR age;  
RUN;
```

SAS integration to Hadoop

➤ SAS/ACCESS to Hadoop Hive

```
PROC SQL;
```

```
    CONNECT TO HADOOP (
```

```
        SERVER='server.phuse_sde_bigdata.sas.com'
```

```
        SUBPROTOCOL=hive2 PORT=1000
```

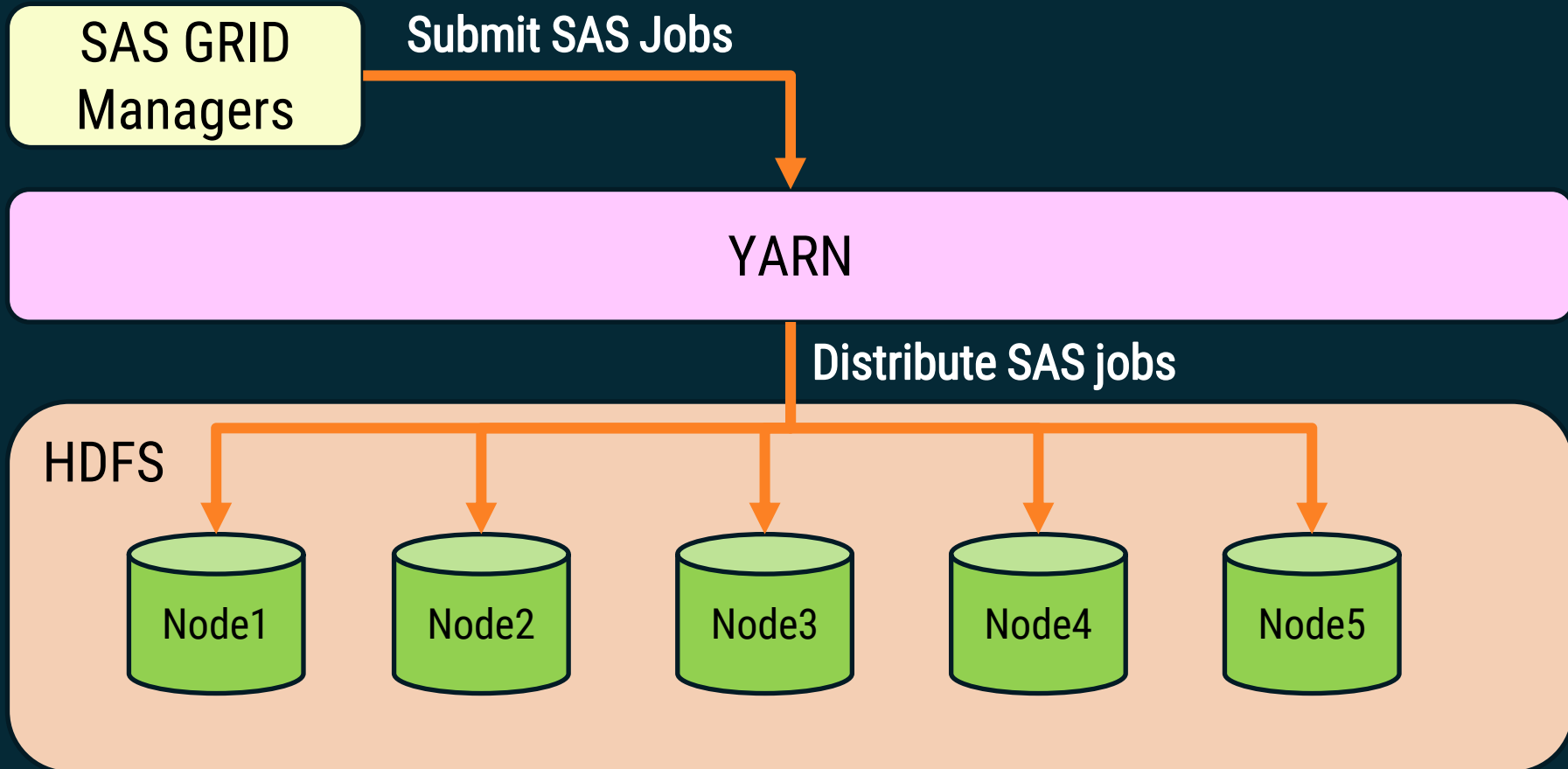
```
        USER="phuse" PW="phuse");
```

```
    EXECUTE (create table....)
```

```
    DISCONNECT FROM HADOOP;
```

```
QUIT;
```

SAS integration to Hadoop



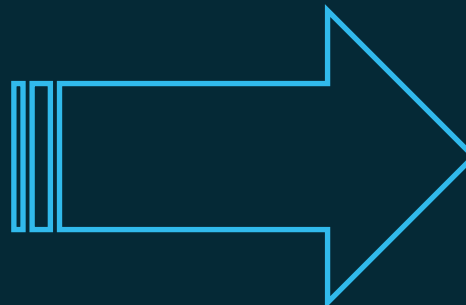
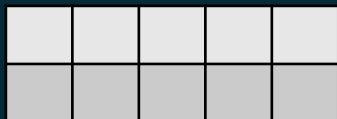
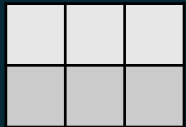
NoSQL Database

Non-only SQL database that can store and retrieve data that do not fit nicely in relational database.



NoSQL Database – Schema Agnostic

Any Data
structures



NoSQL
Database

- Minimal Data Modeling
- Minimal/No ETL
- No pre-defined Schema necessary

Relational vs NoSQL Data - Examples

DM		
SUBJID	SEX	RACE
001	M	WHITE
002	F	ASIAN

```
<DM>
  <ROW>
    <SUBJID>001</SUBJID>
    <SEX>M</SEX>
    <RACE>WHITE</RACE>
  </ROW>
  <ROW>
    <SUBJID>002</SUBJID>
    <SEX>F</SEX>
    <RACE>ASIAN</RACE>
  </ROW>
</DM>
```

NoSQL data usage – multiple RACE

DM		
SUBJID	SEX	RACE
001	M	WHITE
002	F	MULTIPLE

SUPPDM		
SUBJID	QNAM	RACE
002	RACE1	WHITE
002	RACE2	ASIAN

```

<DM>
  <ROW>
    <SUBJID>001</SUBJID>
    <SEX>M</SEX>
    <RACE>WHITE</RACE>
  </ROW>
  <ROW>
    <SUBJID>002</SUBJID>
    <SEX>F</SEX>
    <RACE>WHITE</RACE>
    <RACE>ASIAN</RACE>
  </ROW>
</DM>

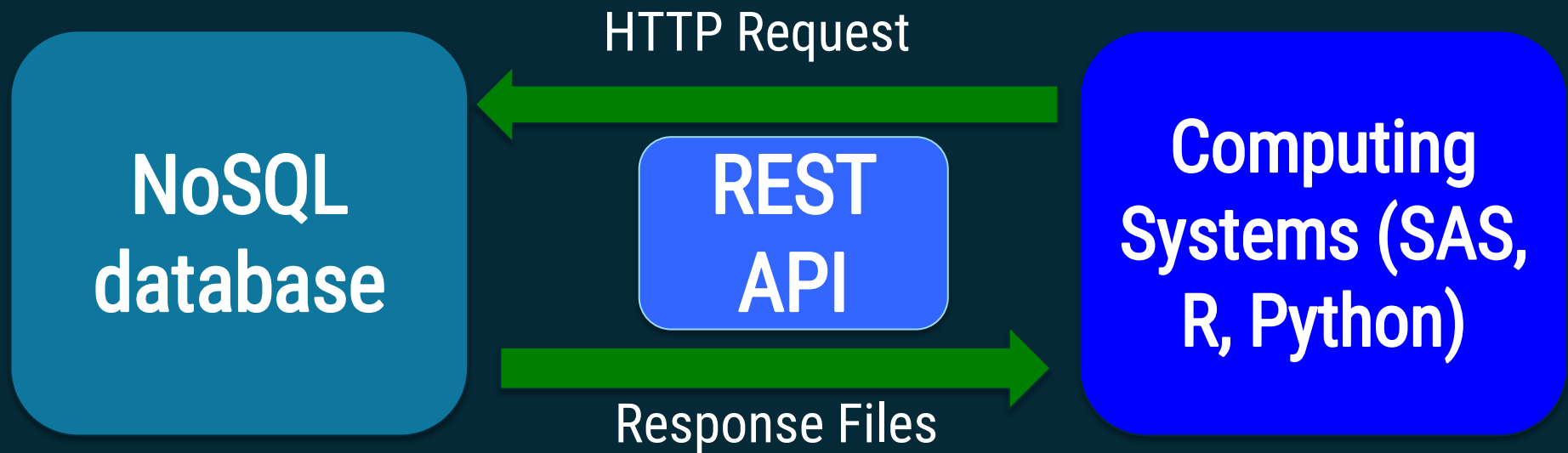
```

WHAT IS **REST** **API?**

- Representational State Transfer (REST)
- A simple data exchange format which is platform-, system- and language-independent and communicates through the internet
- It uses HTTP and the response files come ready to be used
- Popular data communication tools for NoSQL Database

API Architecture

25



Use Case – New York Times NoSQL Database

- New York Times NoSQL Database – Mango DB
- REST API (developer.nytimes.com)
- Obtain Developer API key
- Find NYT REST API:Book Reviews in http://developer.nytimes.com/docs/books_api/Books_API_Book_Reviews
- Find parameter – version, api-key, response-format, isbn, title, author

Use Case 1 – Using SAS for REST API

```
*** file name that will receive document;  
filename review "C:\KL\BookReview\isbn-  
9780062409850.xml";
```

```
**** call resp api using proc http;
```

```
proc http
```

```
  out=review
```

```
  url="http://api.nytimes.com/svc/books/v3/reviews.  
xml?isbn=9780062409850&api-key=xxxxx"
```

```
  method="GET" ;
```

```
run;
```

Use Case 1 – Results Data

"C:\KL\BookReview\isbn-9780062409850.xml"

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<result_set>
```

```
  <status>OK</status>
```

```
  <copyright>Copyright (c) 2016 The New York Times Company. All Rights Reserved.</copyright>
```

```
  <num_results>1</num_results>
```

```
  <results>
```

```
    <result>
```

```
      <url>http://www.nytimes.com/2015/07/14/books/review/harper-lees-go-set-a-watchman.html</url>
```

```
      <publication_dt>2015-07-14</publication_dt>
```

```
      <byline>RANDALL KENNEDY</byline>
```

```
      <book_title>Go Set a Watchman</book_title>
```

```
      <book_author>Harper Lee</book_author>
```

```
      <summary>"Go Set a Watchman" demands that its readers abandon the immature  
        sentimentality ingrained by middle school and the film adaptation of "To Kill a  
        Mockingbird."
```

```
      </summary>
```

```
    <isbn13>
```

```
      <isbn13_item>9780062409850</isbn13_item>
```

```
      <isbn13_item>9780062409874</isbn13_item>
```

```
      <isbn13_item>9780062409881</isbn13_item>
```

```
    </isbn13>
```

```
  </result>
```

```
</results>
```

Use Case 2 – Using R for REST API

```
install.package('RCurl', 'XML')  
library('RCurl')  
library('XML')
```

```
## character files
```

```
file1<-
```

```
getURL("http://api.nytimes.com/svc/books/v3/reviews.xml?isbn=97  
80062409850&api-key=xxx")
```

```
## list files
```

```
file2 <-
```

```
xmlParse("http://api.nytimes.com/svc/books/v3/reviews.xml?isbn=  
9780062409850&api-key=xxx")
```

SAS Codes converting XML Documents to SAS Datasets

```
**** response xml files;  
filename resp " C:\KL\BookReview\isbn-9780062409850.xml ";
```

```
**** Create response xml map file;  
filename respmap " C:\KL\BookReview\response.map ";  
libname resp xmlv2 xmlmap=respmap automap=replace;
```

```
**** Convert response xml files to SAS temporary dataset in  
work area;  
proc copy in=resp out=work;  
run;
```

SAS Datasets from XML Documents

31

- 5 SAS datasets: isbn13, isbn13_item, result, result_set and results
- result SAS dataset

results_ORDINAL	result_ORDINAL	url	Publication_dt	byline	Book_title	Book_author	summary
1	1	http://www.nytimes.com/2015/07/14/books/review/harper-lees-go-set-a-watchman.html	2015-07-14	RANDALL KENNEDY	Go Set a Watchman	Harper Lee	"Go Set a Watchman" demands that its readers abandon the immature sentimentality ingrained by middle school and the film adaptation of "To Kill a Mockingbird."

<result>

<url><http://www.nytimes.com/2015/07/14/books/review/harper-lees-go-set-a-watchman.html></url>

<publication_dt>2015-07-14</publication_dt>

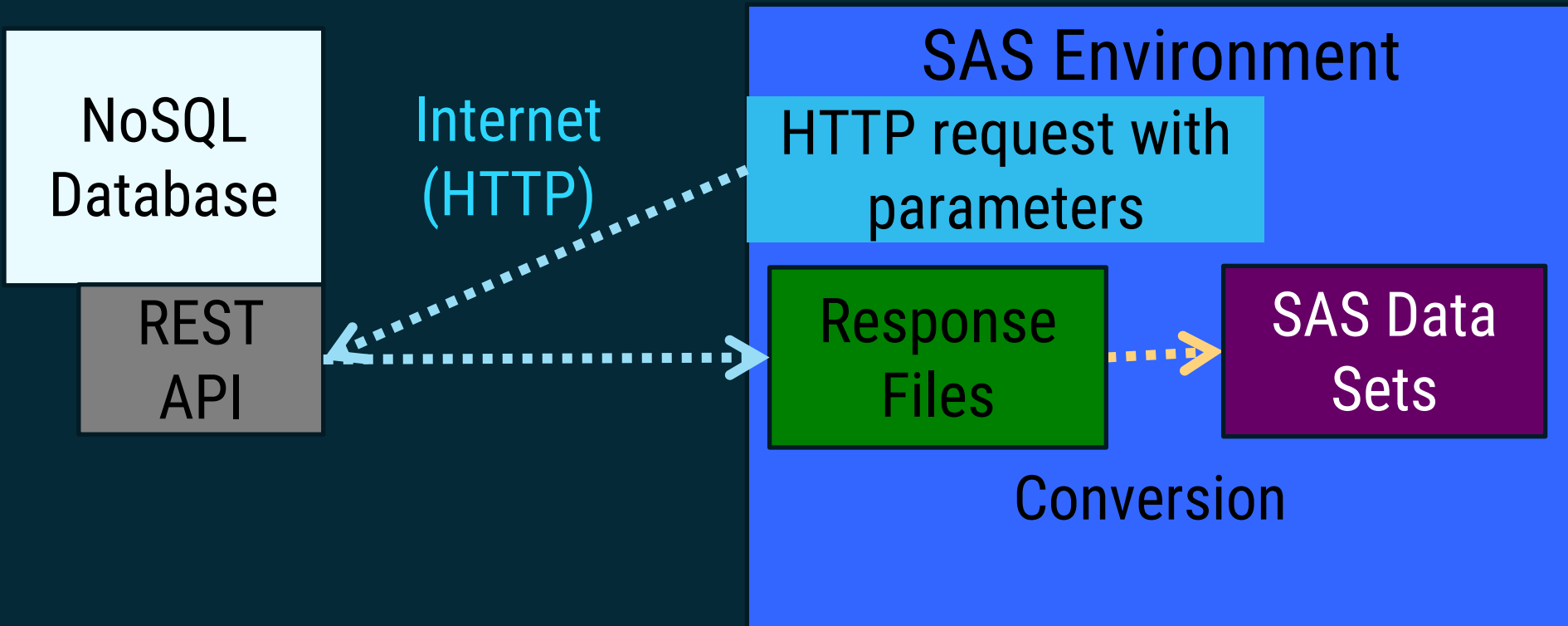
<byline>RANDALL KENNEDY</byline>

.....

</result>

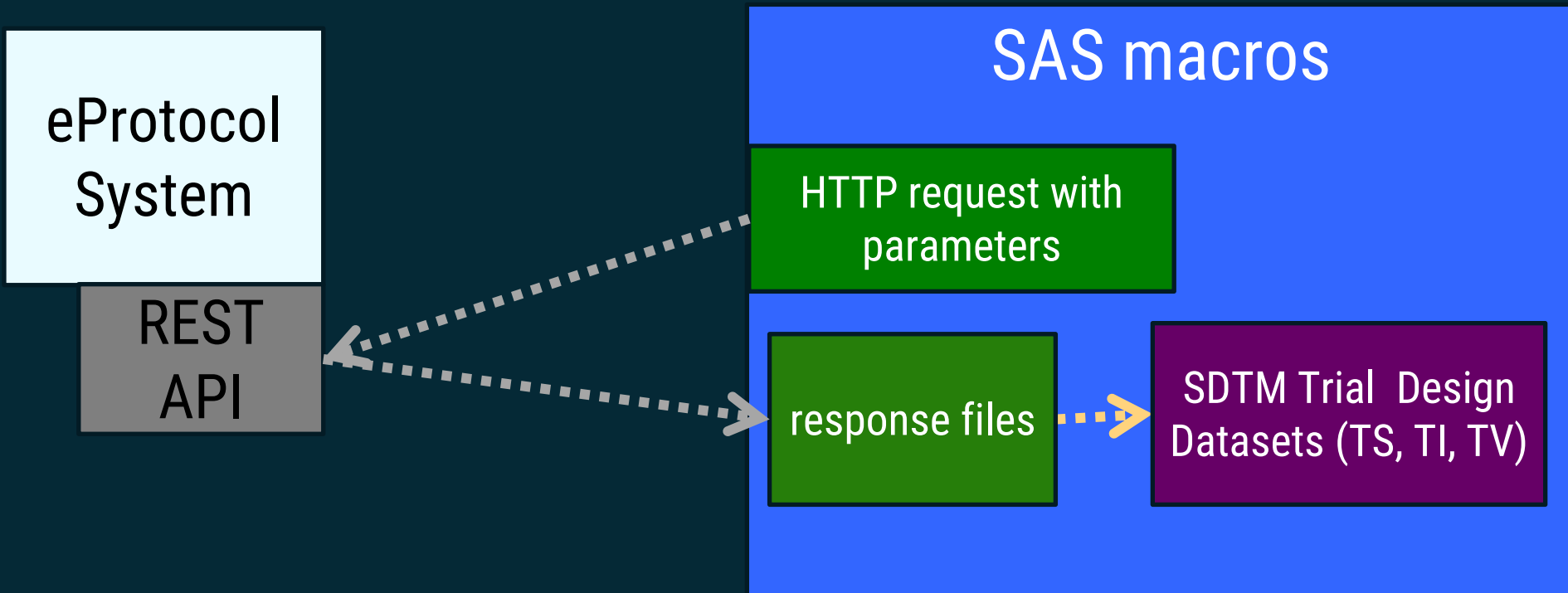
Architecture Design of Integration between SAS and NoSQL Database

32



Use Case – SDTM Trial Design Domains

33



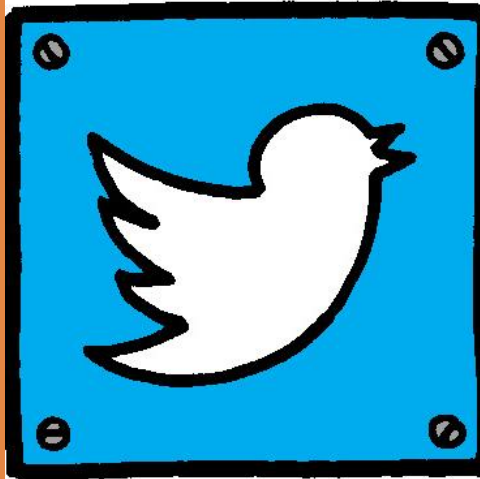
Google
(Bigtable,
Level DB)



LinkedIn
(Voldemort)



Facebook
(Cassandra)



Twitter
(Hadoop/Hbase,
FlockDB,
Cassandra)

Netflix (SimpleDB,
Hadoop/Hbase,
Cassandra)



CERN (CouchDB)



What's
Next?



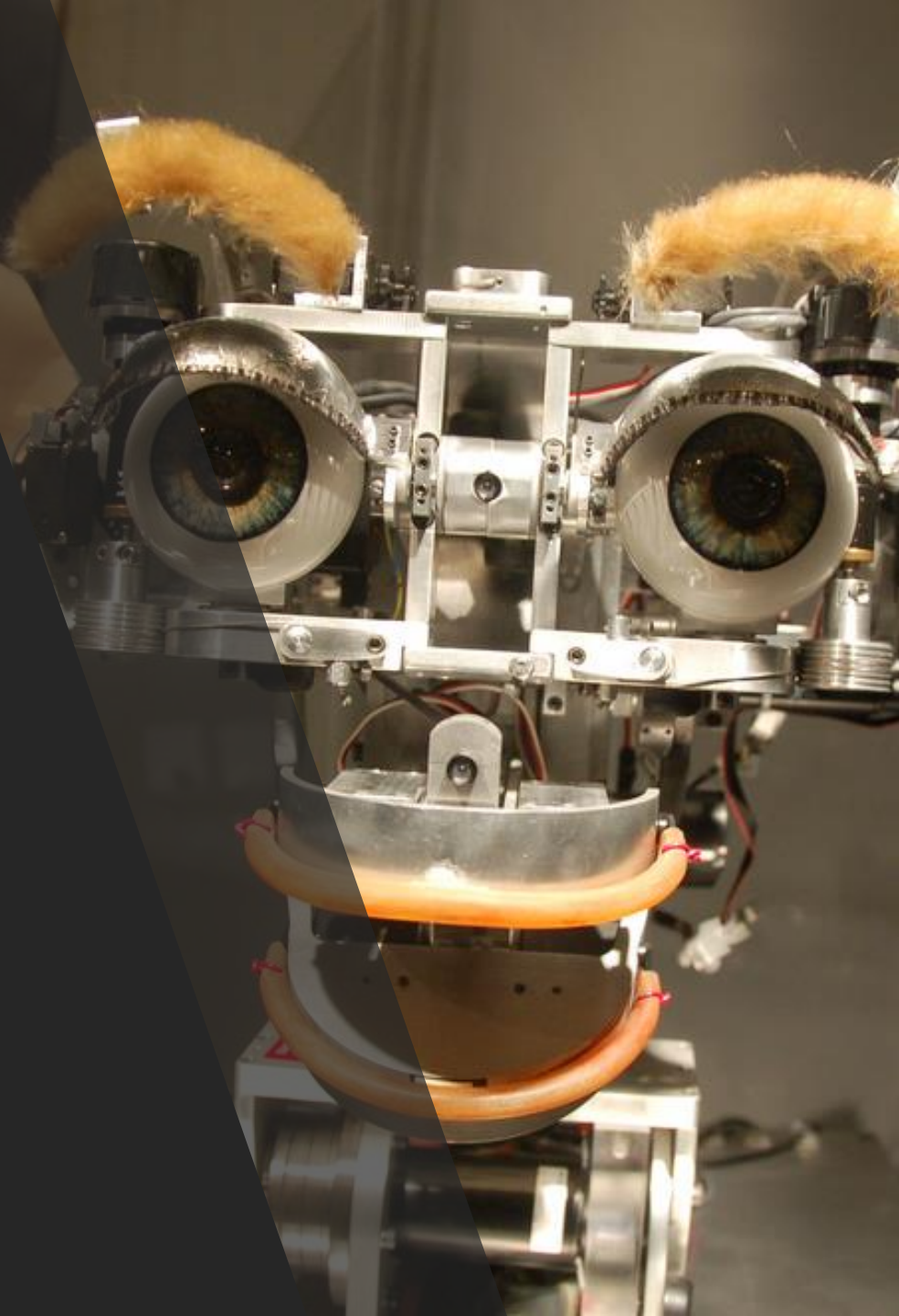
Honey, do you
know about
**Machine
Learning?**

Why did my wife ask / expect me if I know about **Machine Learning**?

- Programming
- Statistics / modeling
- Working with data all the times

What is ML?

An application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience **without being explicitly programmed.**



```

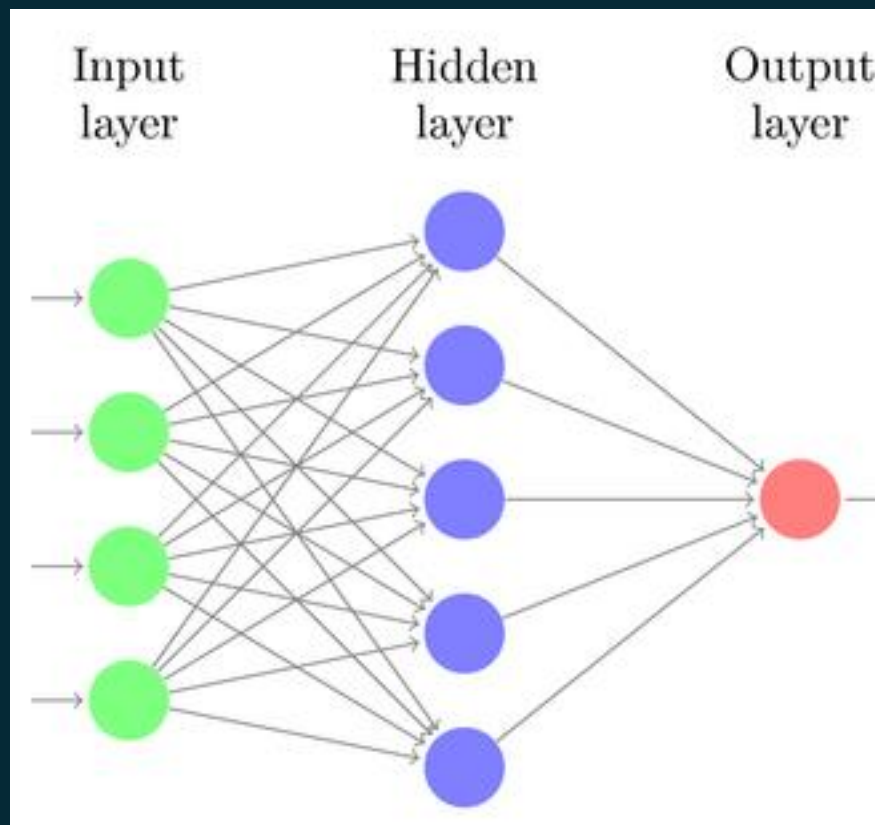
* Derivation of SEX variables **;
  if upcase(_sex)='M' then do; sex='Male';    sexn=1;
else if upcase(_sex)='F' then do; sex='Female'; sexn=2;
else if missing(_sex)      then do; sex='Missing'; sexn=3;

* Derivation of ETHNIC variables **;
if not missing(ethnic) then do;
  if upcase(ethnic)='HISPANIC OR LATINO' then th
  else if upcase(ethnic)='NOT HISPANIC OR LATINO' th

  ** Convert the ETHNIC text to lowercase with first let
  ethnic=propcase(ethnic);
  if indexw(upcase(ethnic), 'OR')>0 then do;
    substr(ethnic, indexw(ethnic, "Or"), 2)='or';
  end;
end;
else if missing(ethnic) then do; ethnic='Missing'; ethn

* Derivation of RACE variables **;
if not missing(race) then do;
  if upcase(race) = 'AMERICAN INDIAN' then r
  else if upcase(race) = 'CANADIAN ABORIGINAL' then r
  else if upcase(race) = 'BLACK' then d
  else if upcase(race) = 'NATIVE HAWAIIAN' then r

```



Explicit programing

Automatically learn and improve



How does Human Learn?
- Experience

How does Machine Learn?



Input
Data

Algorithm

How ML works

Labeled Input data

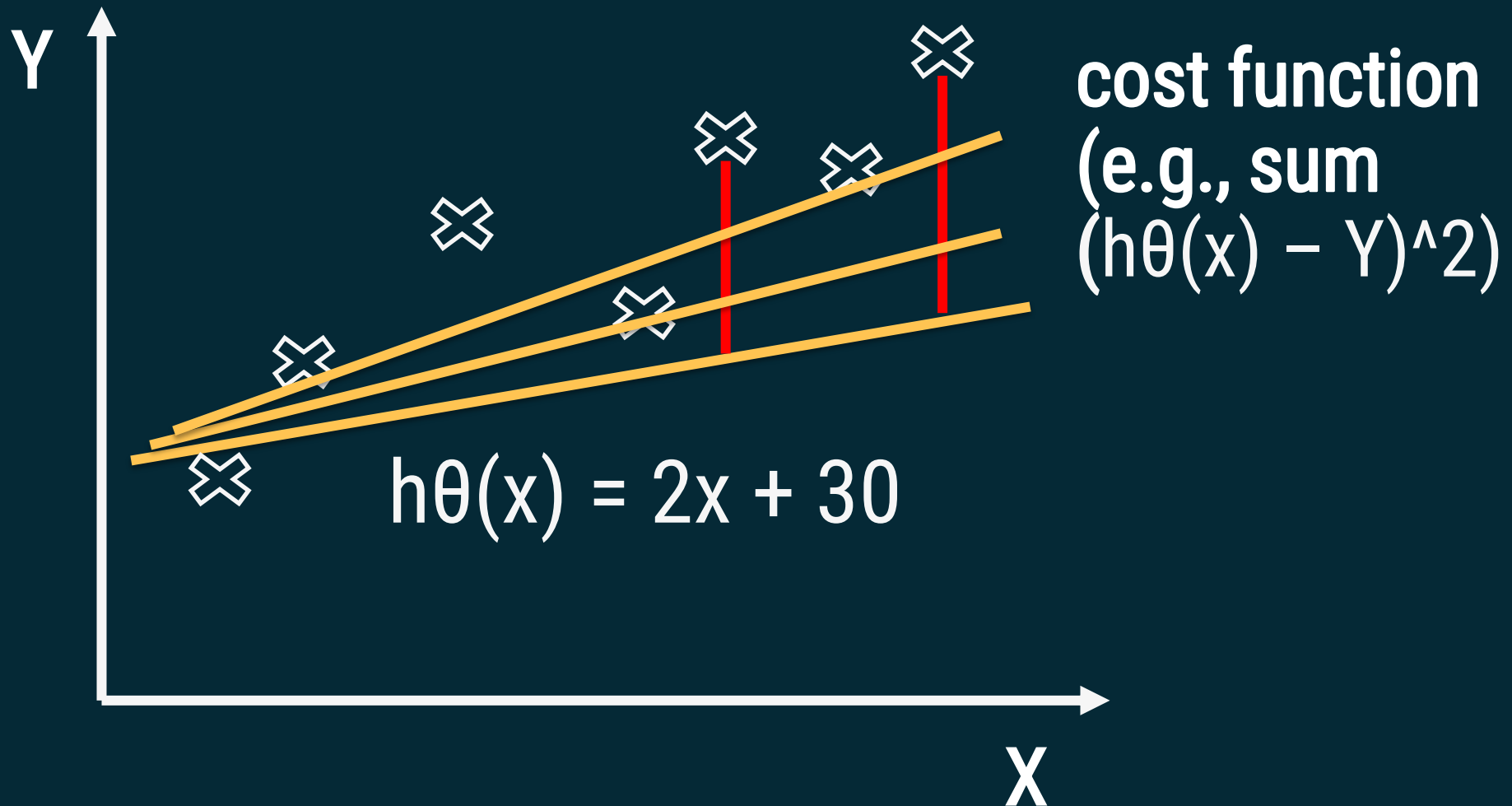
X0	X1	X2	...	Xn	Y

- Data builds models
- More data, the better models

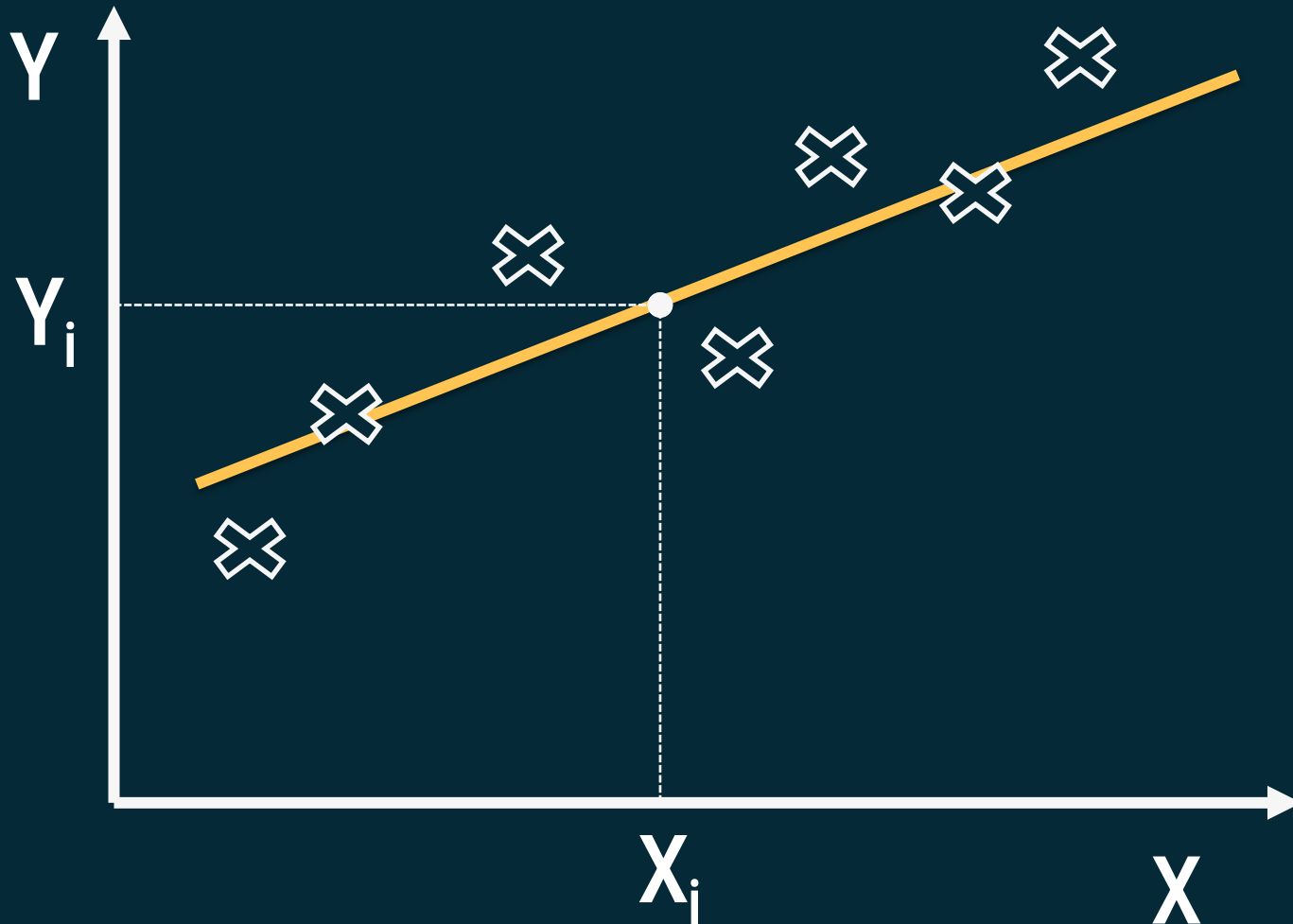
Algorithm

- Hypothesis Function - $h_{\theta}(x) = \theta x + b$
- Minimize Cost Function - $J(\theta) = h_{\theta}(x) - Y$

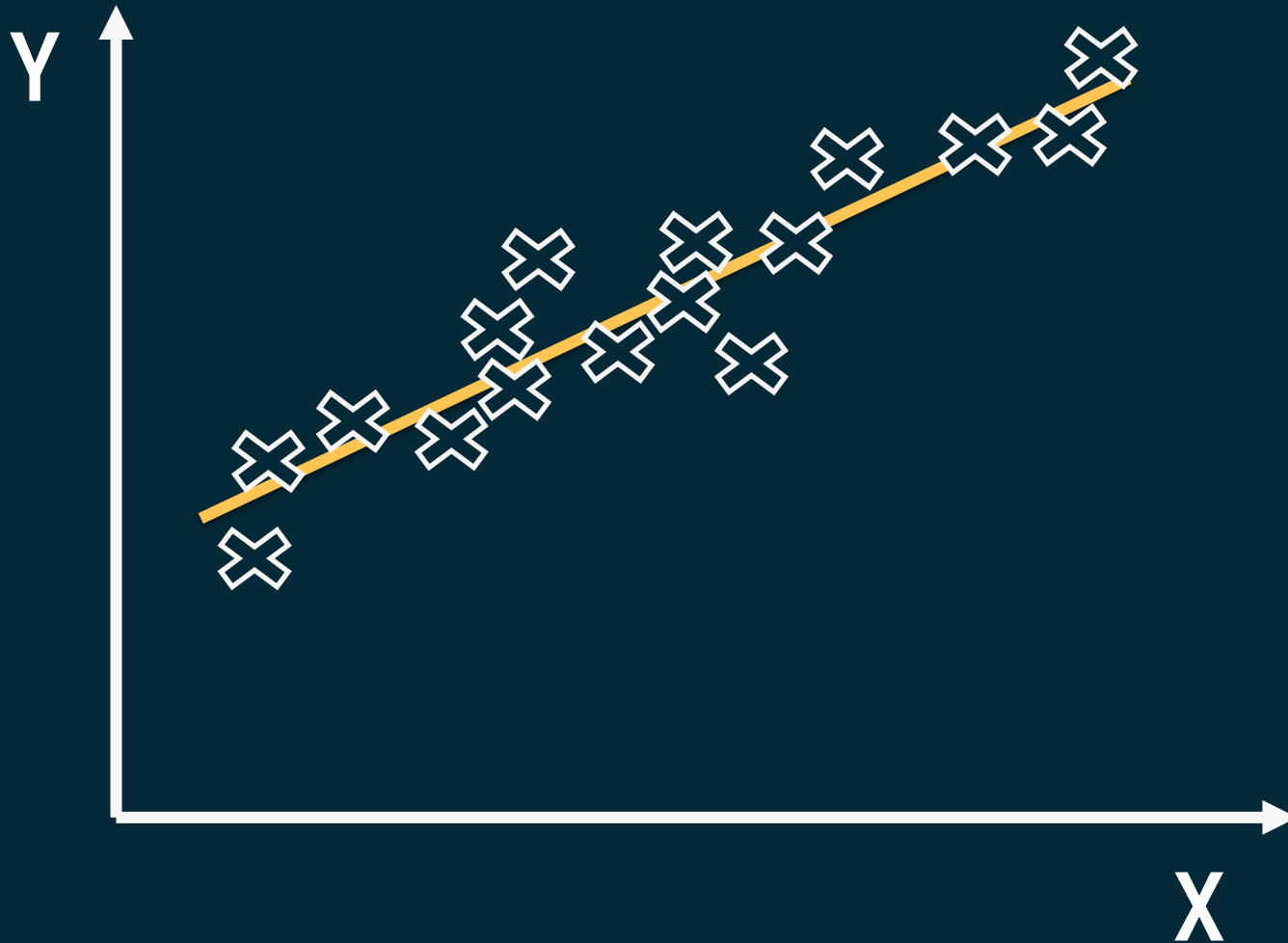
How machine builds better models



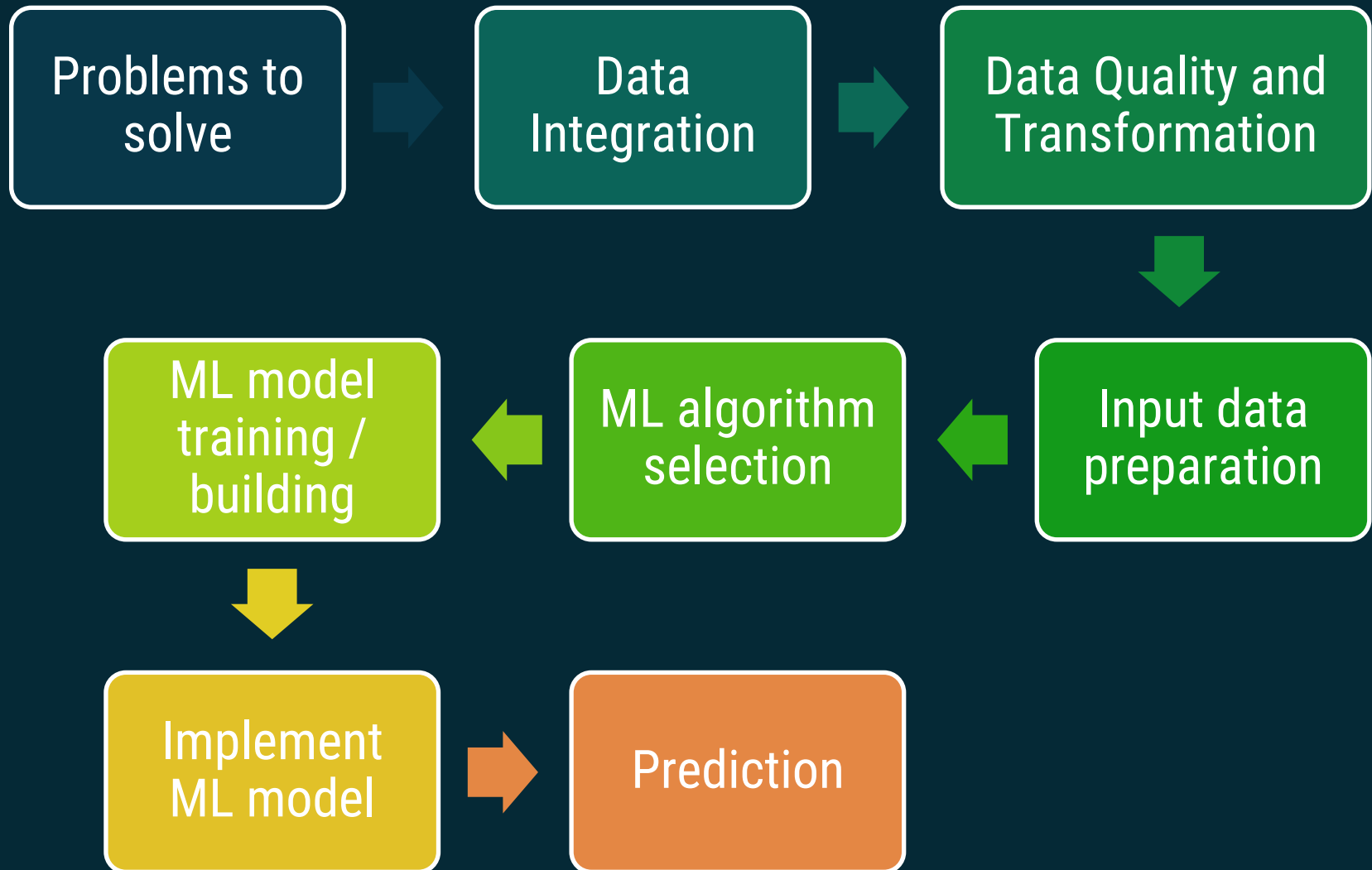
How machine builds best models



More data, the better model



Typical Machine Learning Workflow



Supervised Machine Learning

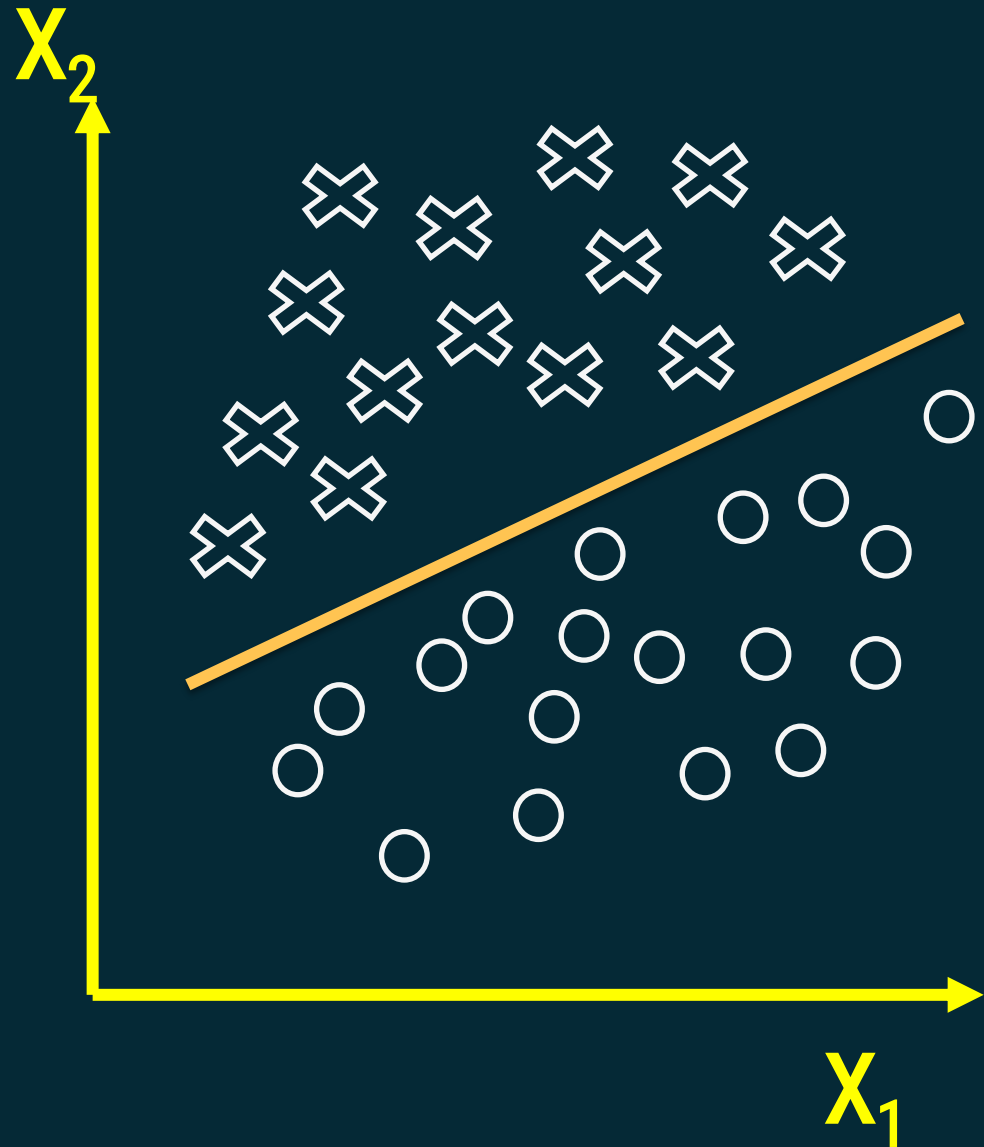
- Input data labeled – has a target

X0	X1	X2	...	Xn	Y

- Specific purpose
- Types
 - Classification
 - Regression

Classification

- Categorical target
- Often binary
- Example : Yes/No, 0 to 9
- Algorithms: Logistic Regression, SVM



Python codes for ML Logistic Regression

```
#import ML algorithm  
from sklearn.linear_model import LogisticRegression
```

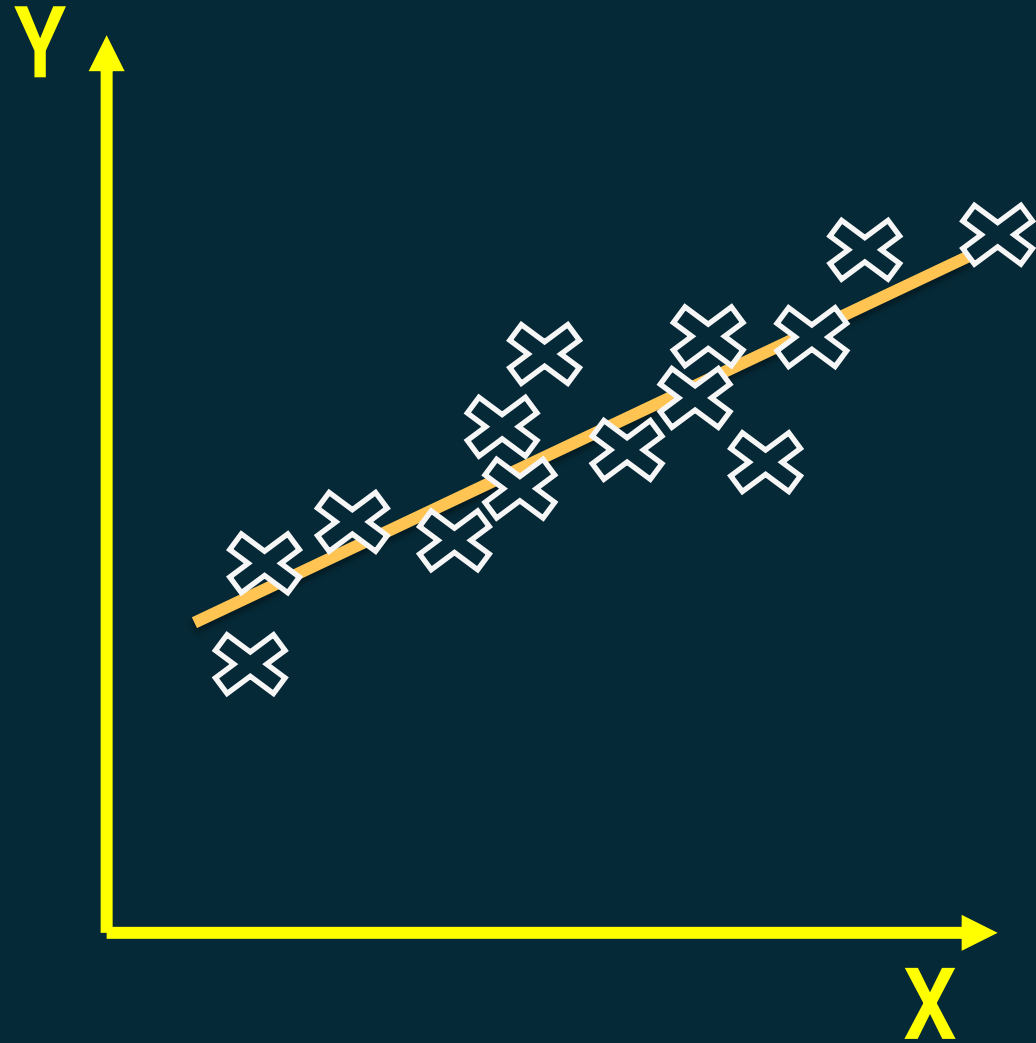
```
#prepare train and test datasets  
x_train = ...  
y_train = ....  
x_test = ....
```

```
#select and train model  
Log_Reg = LogisticRegression()  
Log_Reg.fit(x_train, y_train)
```

```
#predict output  
predicted = Log_Reg.predict(x_test)
```

Regression

- Numeric target
- Continuous variables
- Example : predicting house price per sqft
- Algorithms: Linear Regression, Polynomial Regression



Python codes for ML Linear Regression

51

```
#import ML algorithm  
from sklearn import linear_model
```

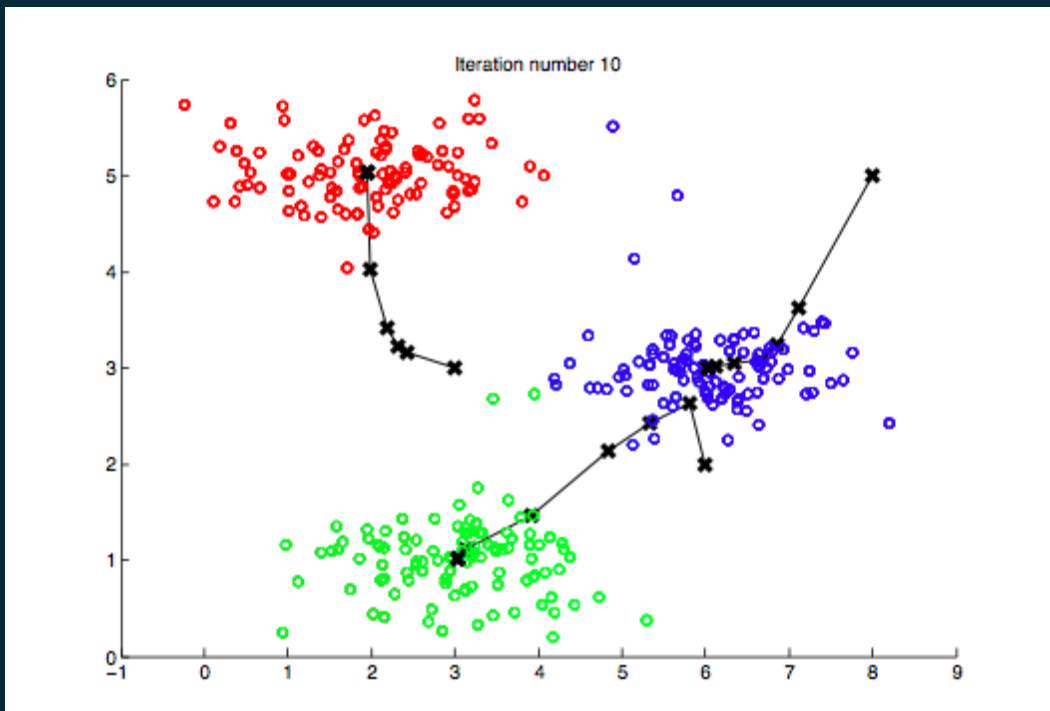
```
#prepare train and test datasets  
x_train = ...  
y_train = ....  
x_test = ....
```

```
#select and train model  
linear = linear_model.LinearRegression()  
linear.fit(x_train, y_train)
```

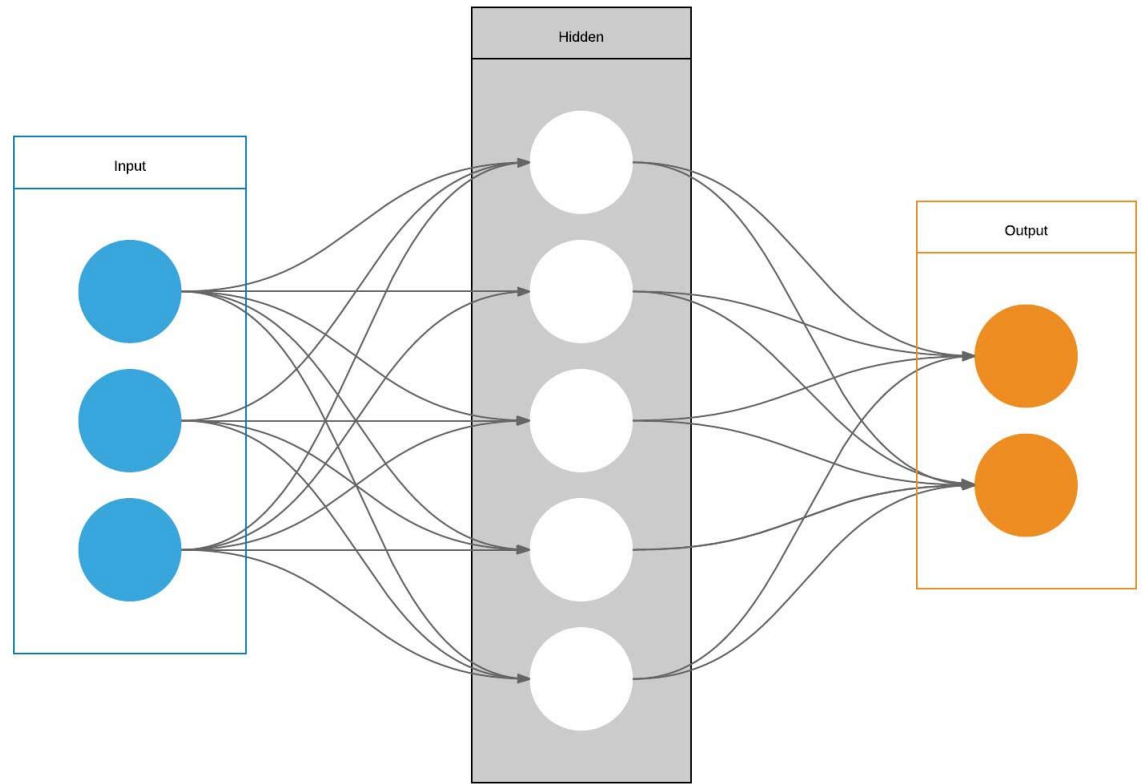
```
#predict output  
predicted = linear.predict(x_test)
```

Unsupervised Machine Learning

- Input data not-labeled – no target
- Exploratory
- Type
 - Clustering – the assignment of set of observations into subsets (clusters)

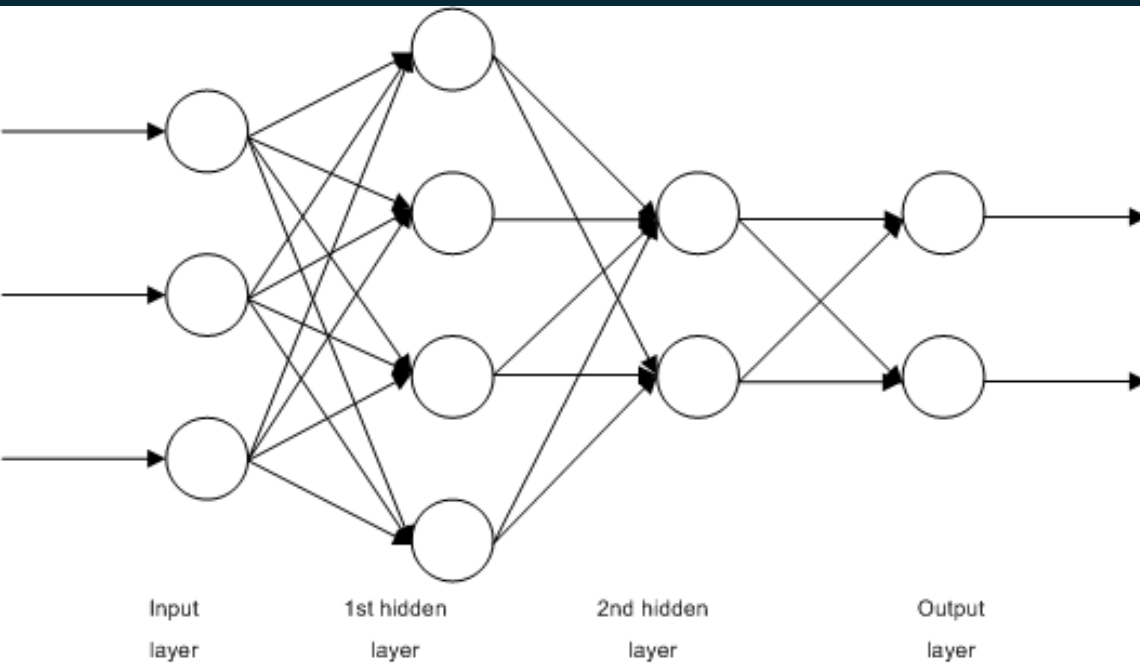
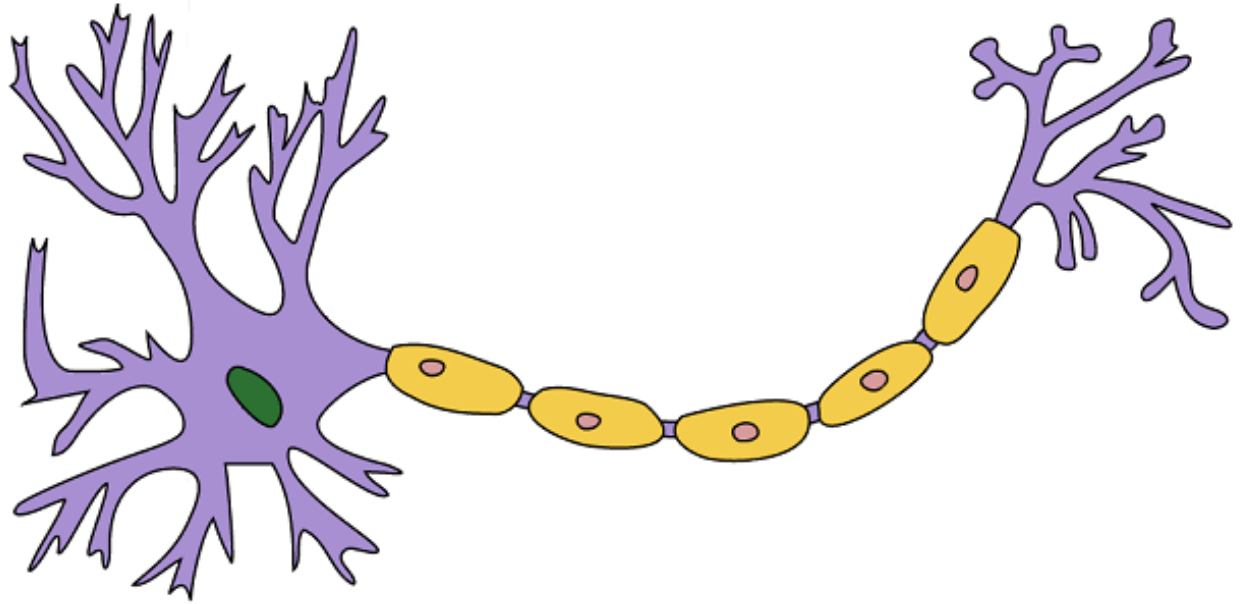


Artificial Neural Network (ANN)



- Most powerful ML algorithm
- Game Changer
- Works very much like human brain – Neural network

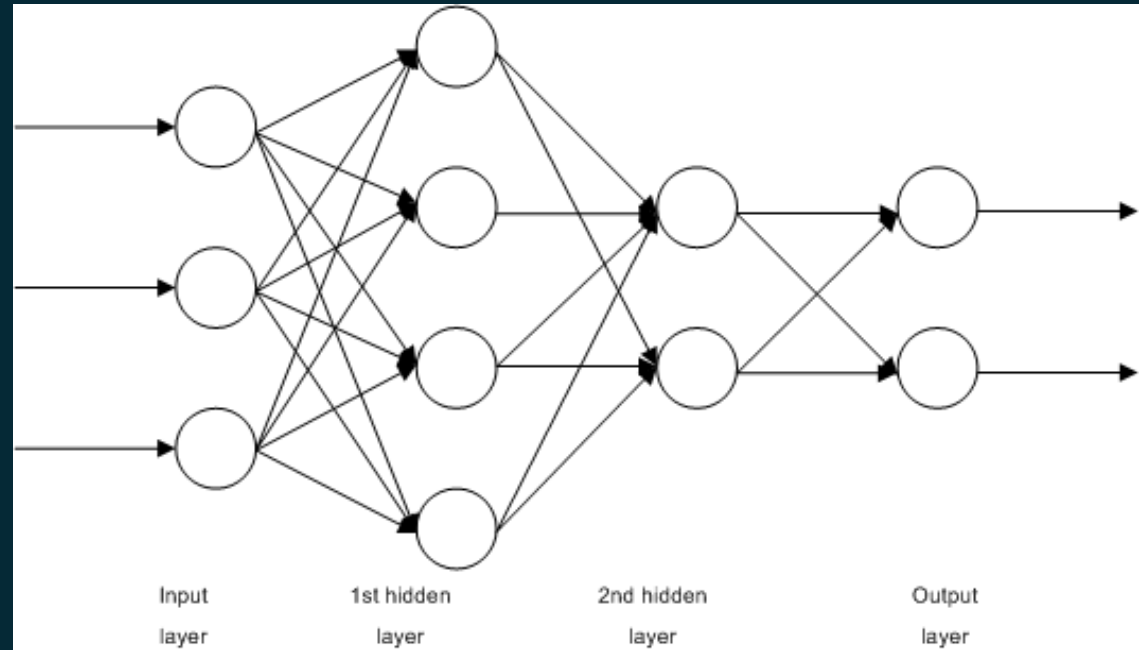
Human Neuron



Neural Network

ANN Architecture

- Input layer
 - 3 features (variables)
- Hidden layer
 - Hidden layer1
 - - 4 neurons
 - Hidden layer2
 - - 2 neurons
 - Other parameters – weight, activation function, learning rate
- Output layer



Python codes for DNN

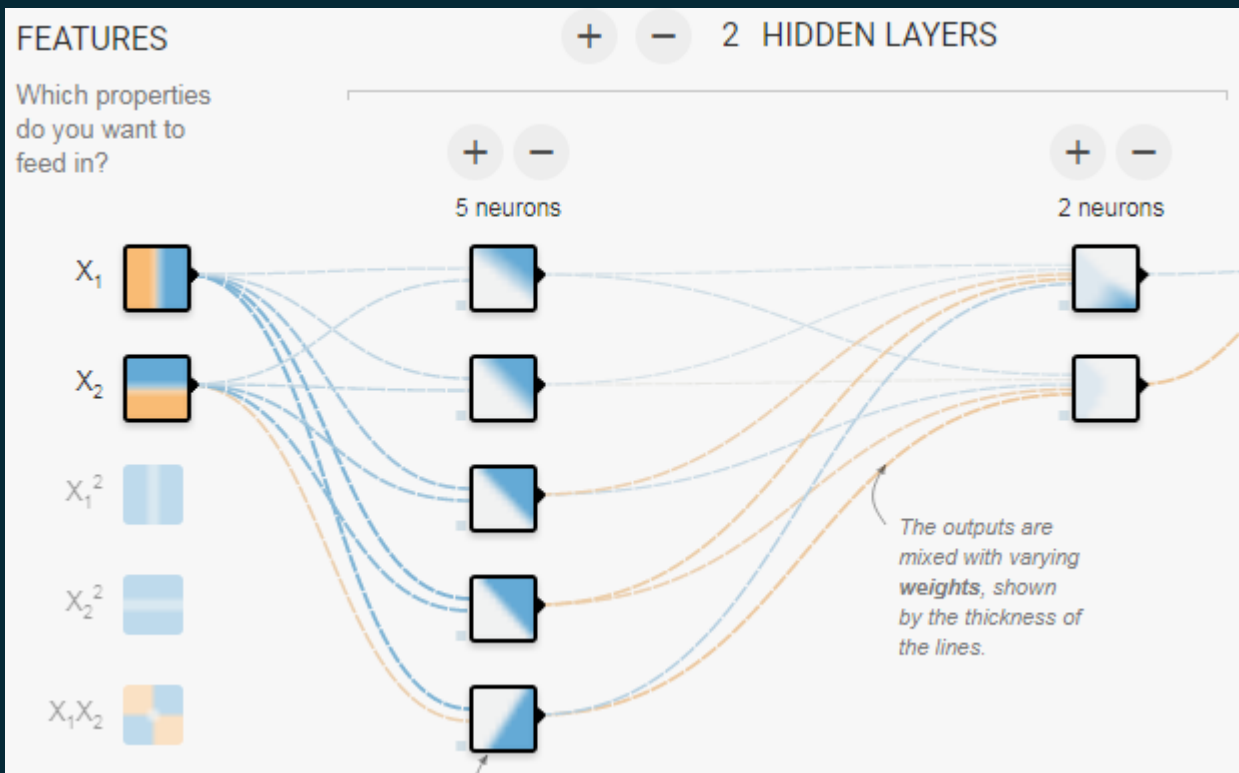
```
#import ANN - TensorFlow  
Import tensorflow as tf
```

```
X = tf.placeholder(..)  
Y = tf.placeholder(..)  
hidden1 = tf.layer.dense(X, 4, activation=tf.nn.relu)  
hidden2 = tf.layer.dense(hidden1, 2, activation=tf.nn.relu)  
logits = neuron_layer(hidden2, 2)  
....  
loss = tf.reduce_mean(...)  
optimizer = tf.train.GradientDescentOptimizer(0.1)  
training_op = optimizer.minimizer(loss)  
  
tf.Session.run(training_op, feed_dict={X:x_train, Y:y_train})
```

Tensor Flow Demo

57

<http://playground.tensorflow.org>





Where is SAS in ML?

SAS Visual Data Mining and ML

- Linear Regression
- Logistic Regression
- Support Vector Machine
- Deep Neural Networks (limited layers)

Why is AI(ML) so popular now?

- **Cost effective**
 - Automate a lot of works
 - Can replace human labors
 - “Pretty much anything that a normal person can do in <1 sec, we can now automate with AI”
Andrew Ng
- **Accurate**
 - Better than humans
- **Can solve a lot of complex business problems**

Healthcare AI market

- US - 320 million in 2016
- Europe – 270 million in 2016
- 40% annual rate
- 10 billion in 2024
- AI Start up driven growth

ML application in Pharma R&D

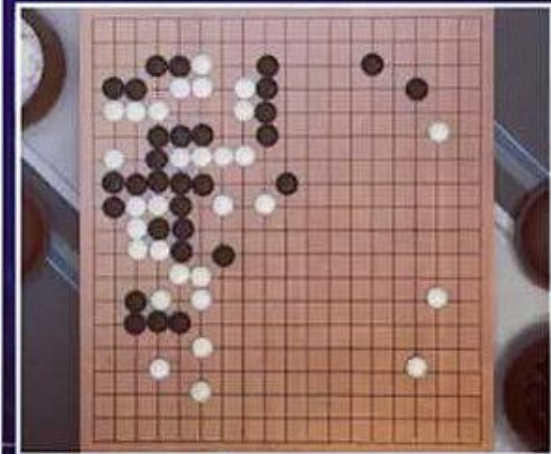
- Drug discovery
- Drug candidate selection
- Clinical system optimization
- Medical image recognition
- Medical diagnosis
- Optimum site selection / recruitment
- Data anomaly

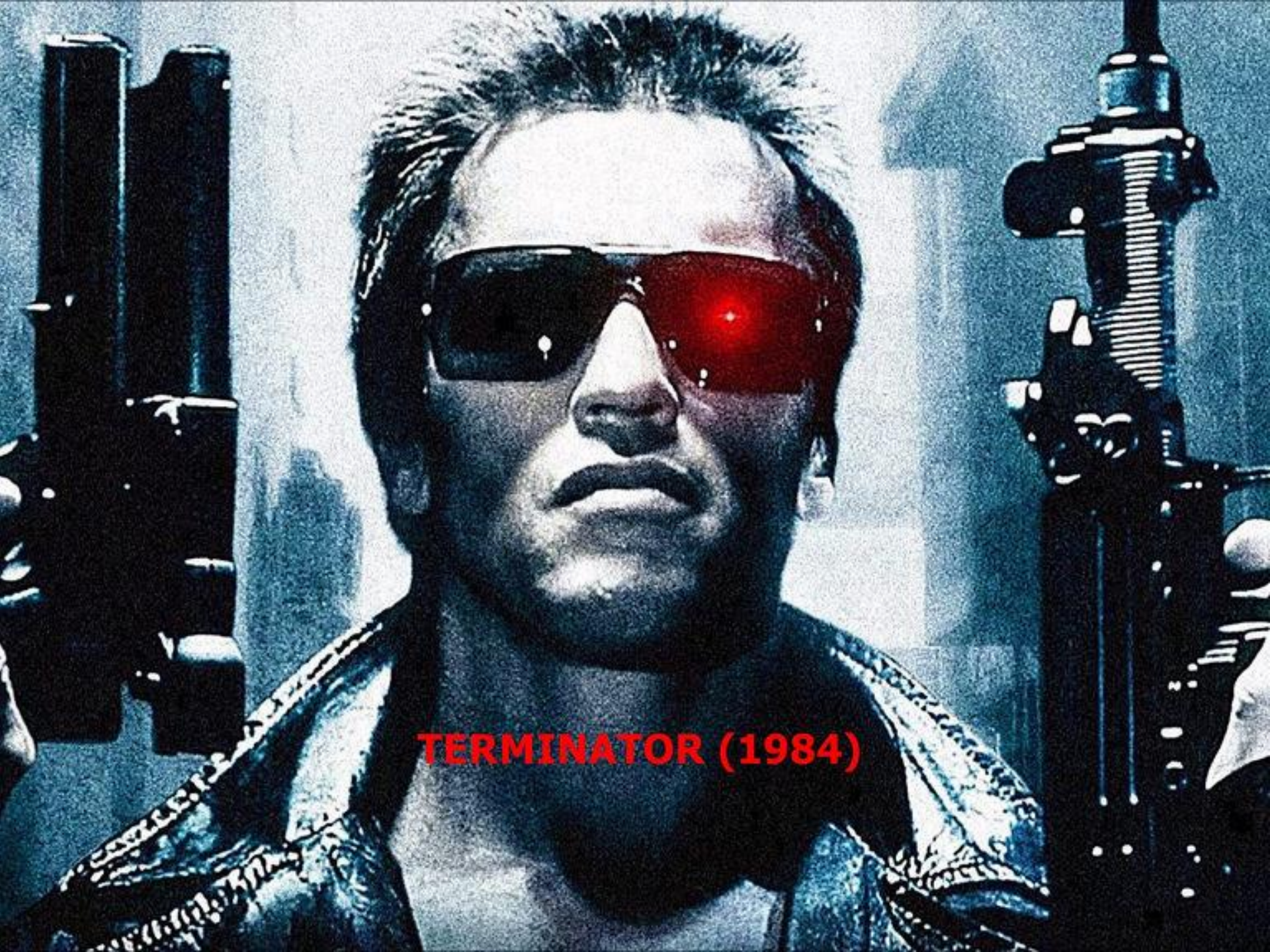
AlphaGO



● ALPHAGO
01:27:15

● LEE SEDOL
00:45:18





TERMINATOR (1984)

Kevin, do you
know about
**Machine
Learning or Big
Data?**



Contact Us!

Contact Clindata Insight to learn more about Big Data and Machine Learning.

Email us at
klee@clindatainsight.com
consulting@clindatainsight.com
<https://www.linkedin.com/in/HelloKevinLee>



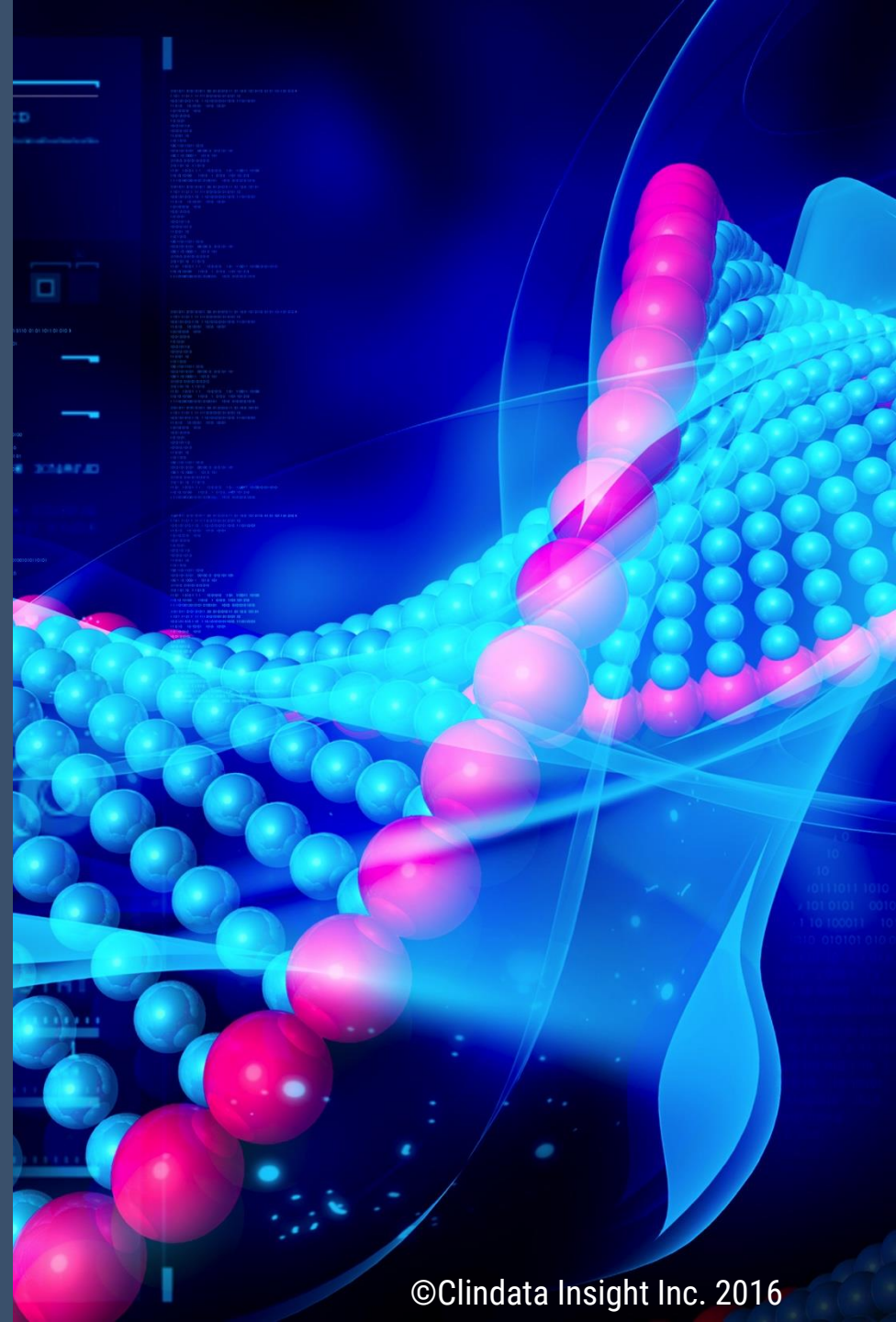
Like us on Facebook @
[Facebook.com/clindatainsight](https://www.facebook.com/clindatainsight)



Twitter @clindatainsight



WeChat @clindatainsight



THAT'S ALL



THANKS

Kevin Lee
klee@clindatainsight.com