FACULTEIT INGENIEURSWETENSCHAPPEN
EN ARCHITECTUUR

IDLAB

TRAIN OCCUPANCY PREDICTION

dieter.dewitte@ugent.be

Lyon, February 6, 2018

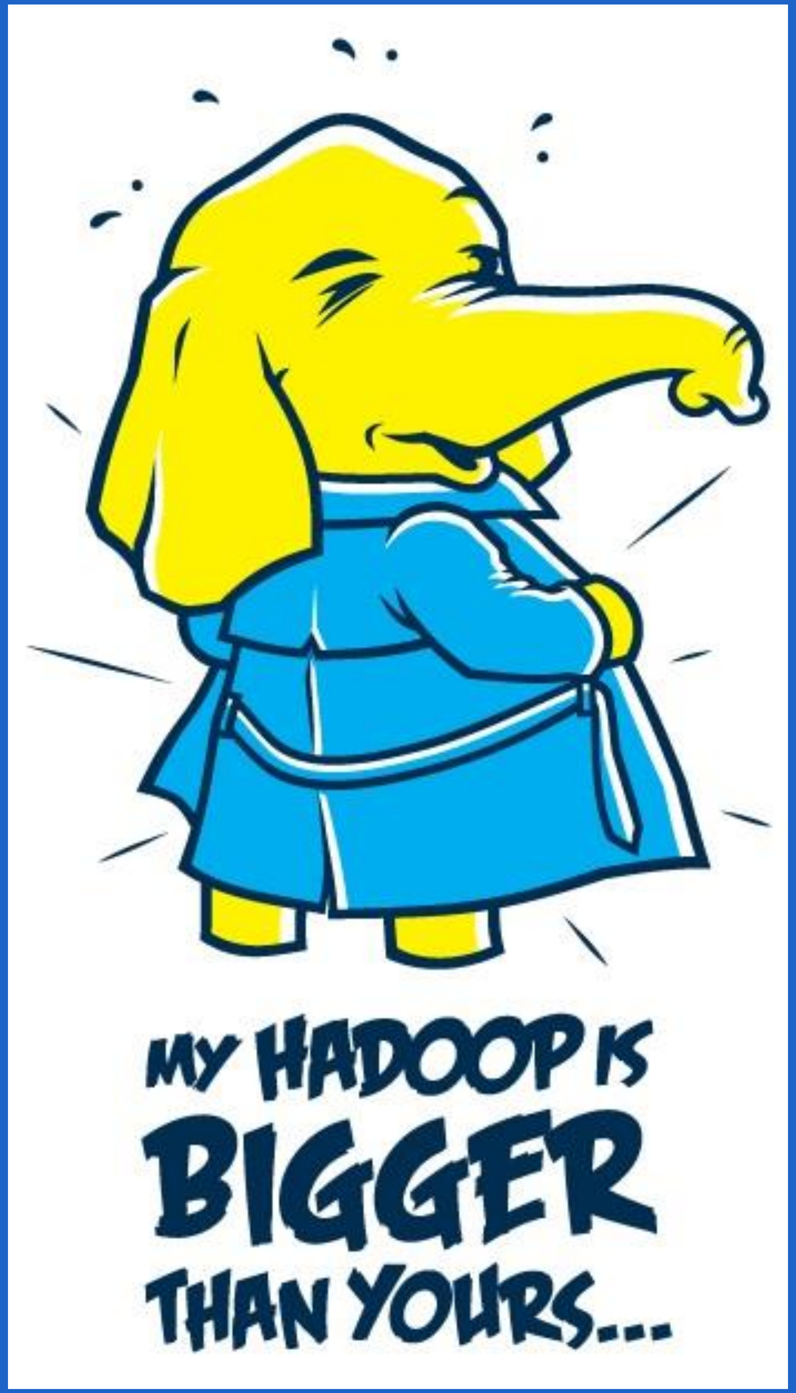
UNIVERSITEIT GENT

imec

TUESDAY MORNING PROGRAM

- 9:30 Intro / iRail / Data Prep & EDA
- 9:50 Setup
- 10:00 Hands-on: Exploratory Data Analysis
- 11:00 Recap & Break
- 11:10 Machine Learning Quick Overview
- 11:30 Hands-on: Building your first model(s)
- 12:00 Recap & Tuning
- 12:05 Final sprint: Who can build the best model?
- 12:30 Lunch Break

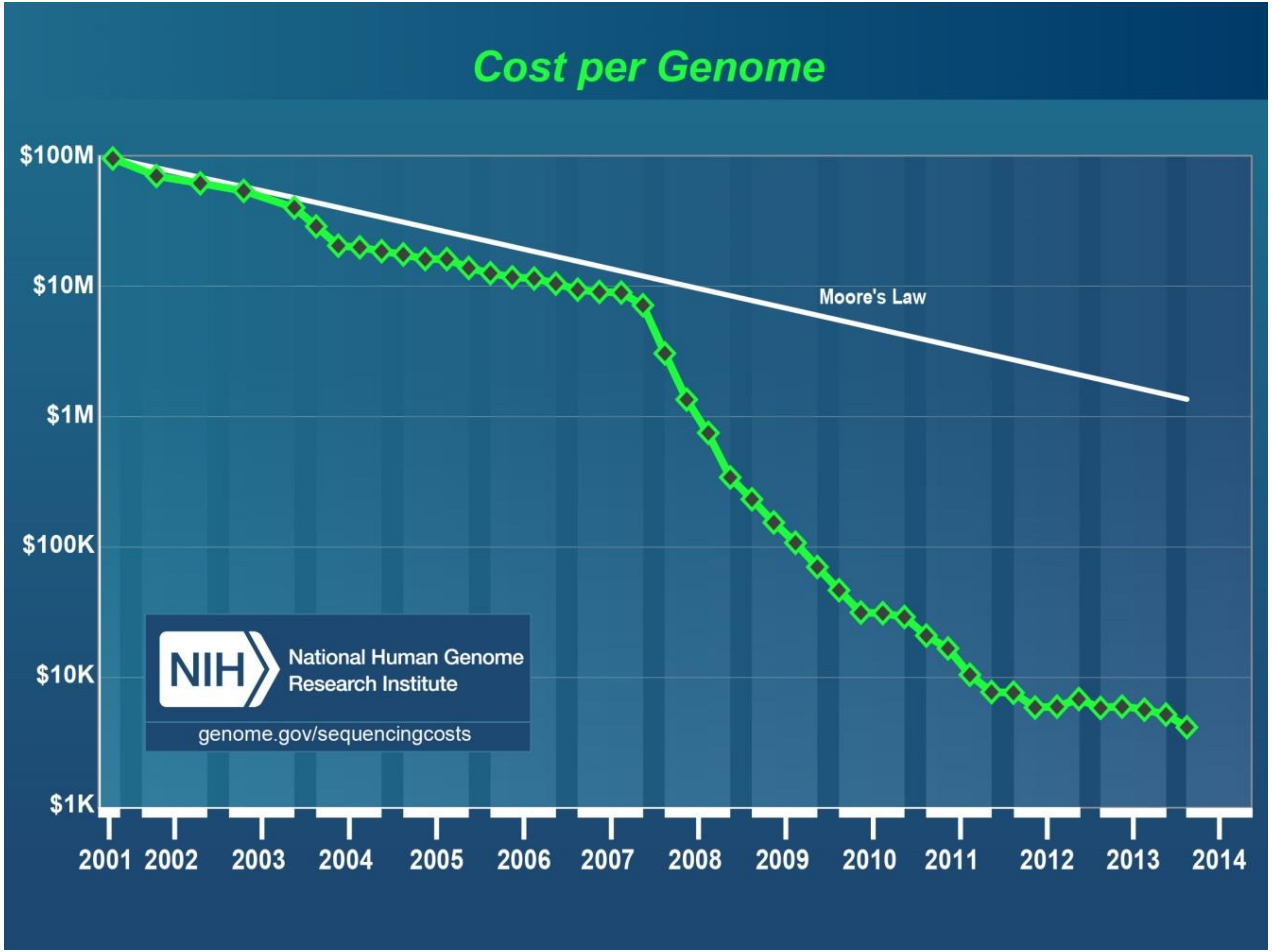
BIG DATA INTRODUCTION



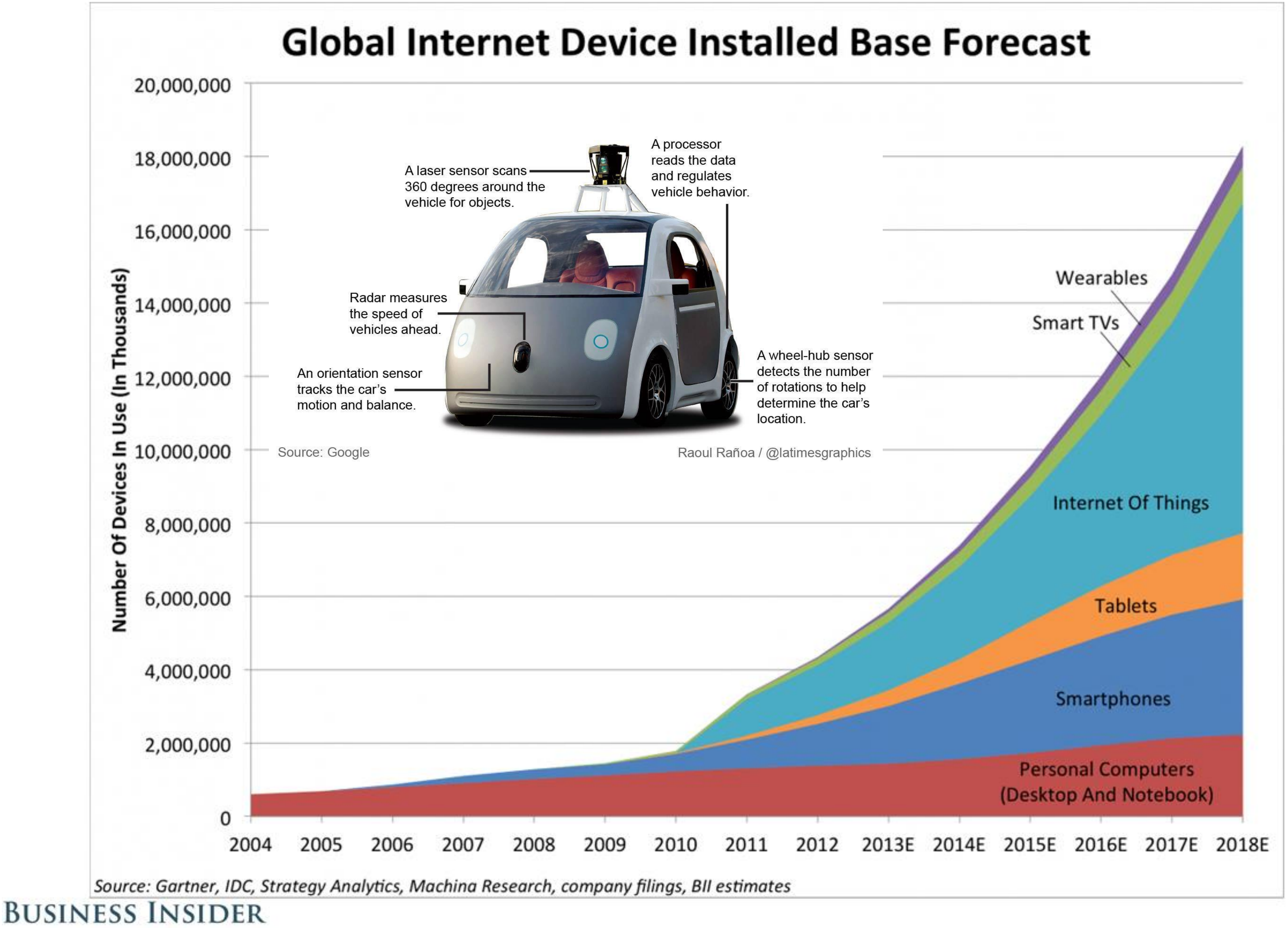
IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban DICTIONARY
1,600+ READS ON Scribd.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
You Tube
25+ HOURS TOTAL DURATION
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
GO-Globe.com
web technologies

USER BEHAVIOR = BIG DATA

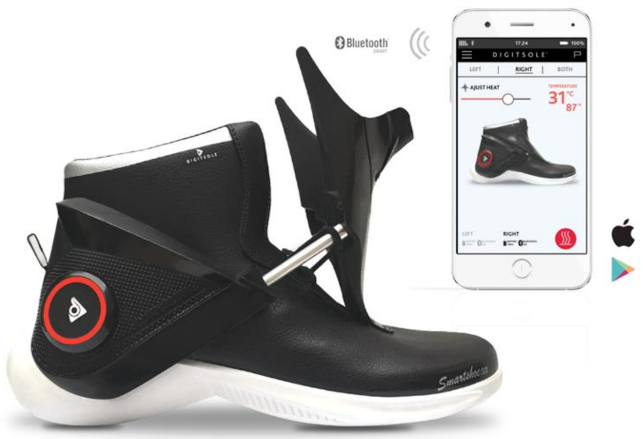
USERS THEMSELVES ARE DATA SOURCES…

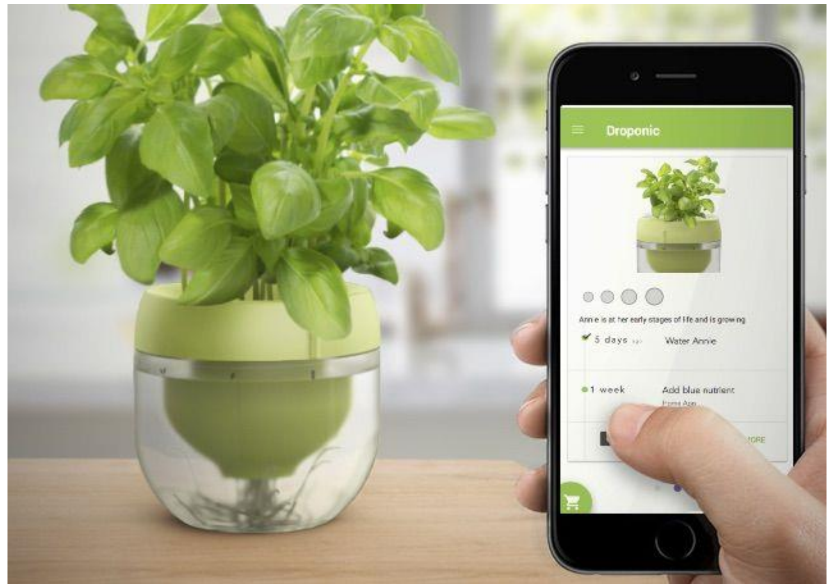


DEVICES PUSH DATA GENERATION

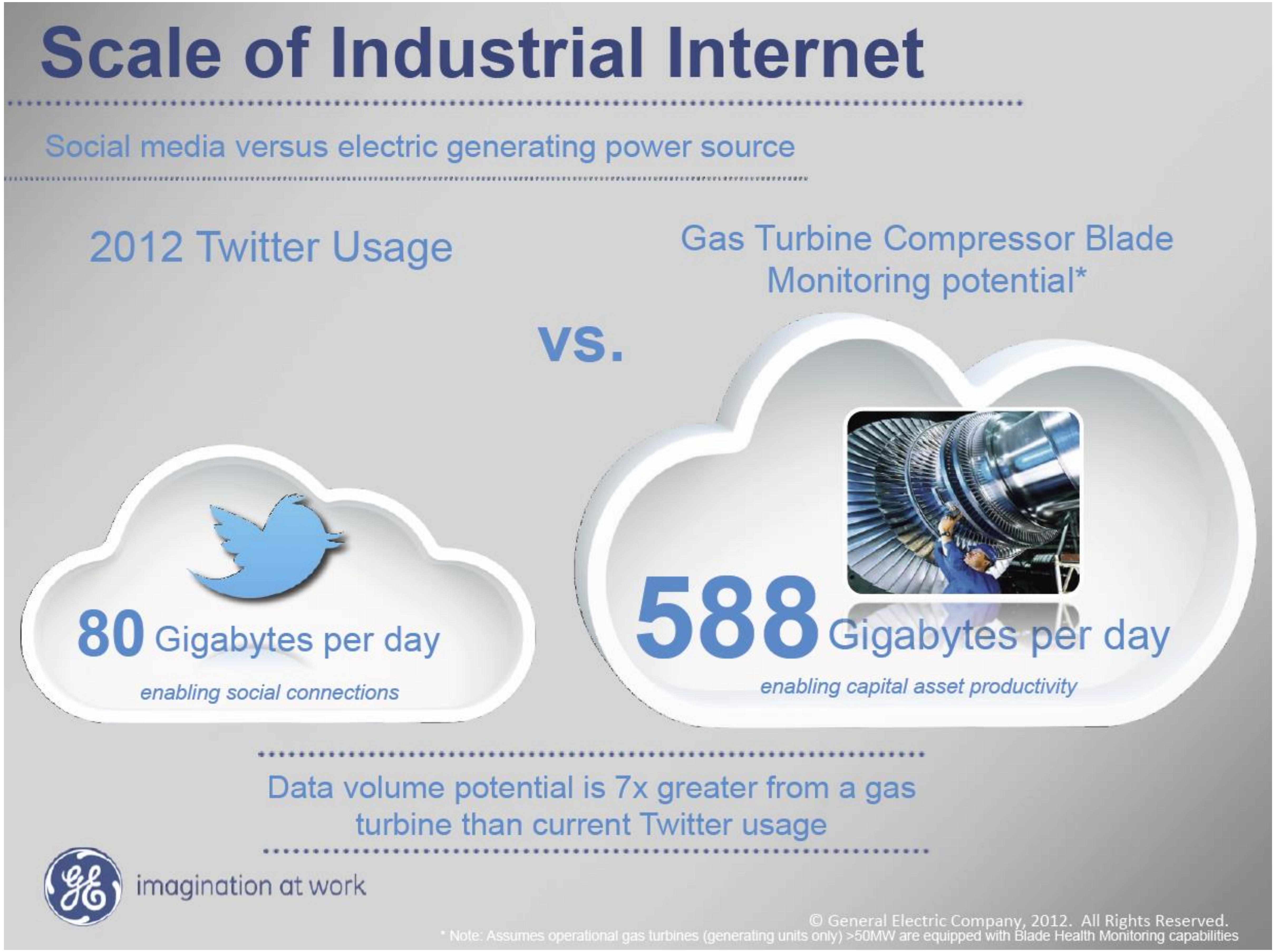


IN CASE YOU AREN'T CONVINCED YET…

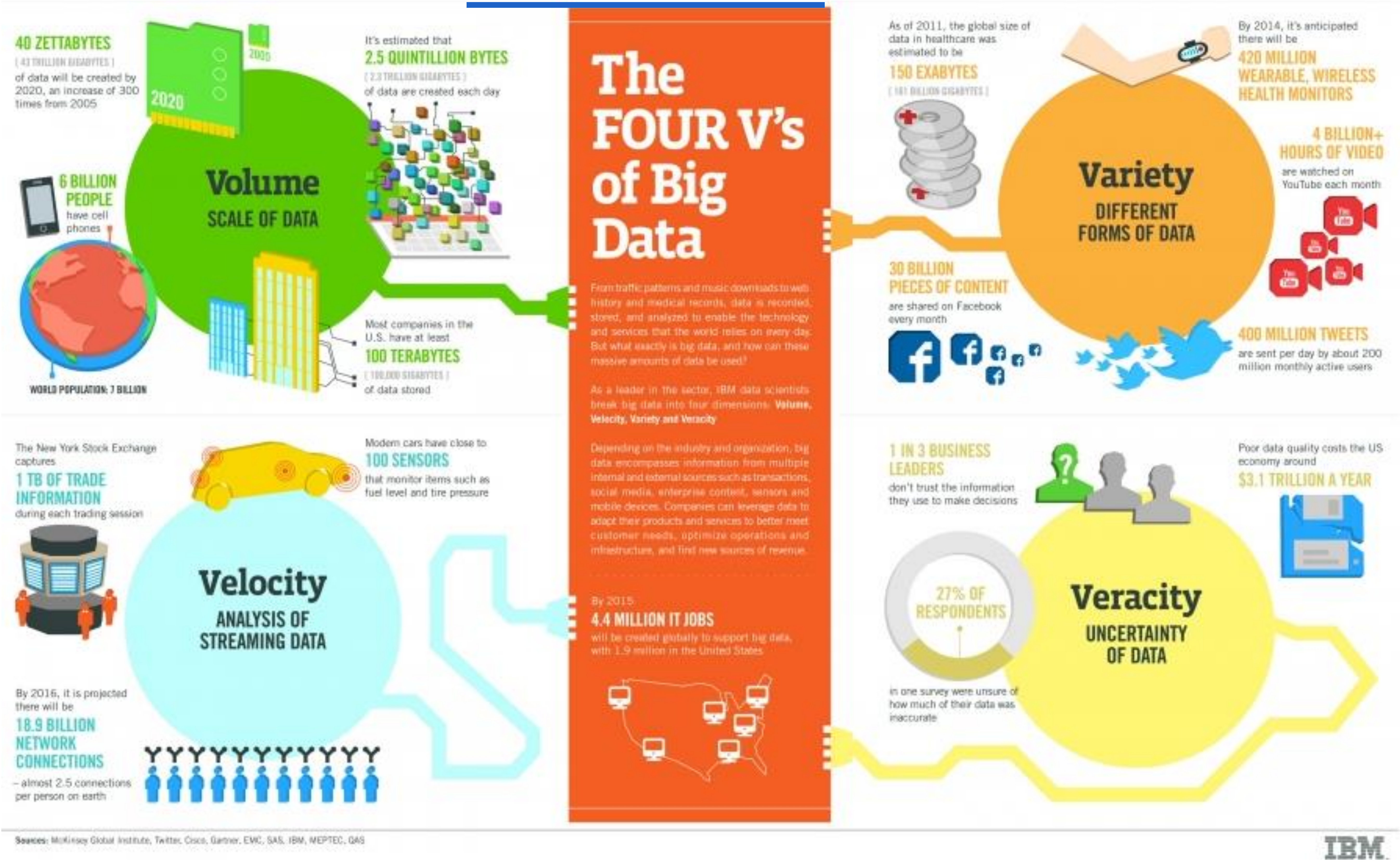




INDUSTRY 4.0


Scale of Industrial Internet
Social media versus electric generating power source
2012 Twitter Usage
Gas Turbine Compressor Blade Monitoring potential*
vs.
80 Gigabytes per day
enabling social connections
588 Gigabytes per day
enabling capital asset productivity
Data volume potential is 7x greater from a gas turbine than current Twitter usage
imagination at work
© General Electric Company, 2012. All Rights Reserved.
* Note: Assumes operational gas turbines (generating units only) >50MW are equipped with Blade Health Monitoring capabilities

NOT JUST ‘MY DATA IS BIGGER THAN YOURS’…



TRAIN OCCUPANCY PREDICTION

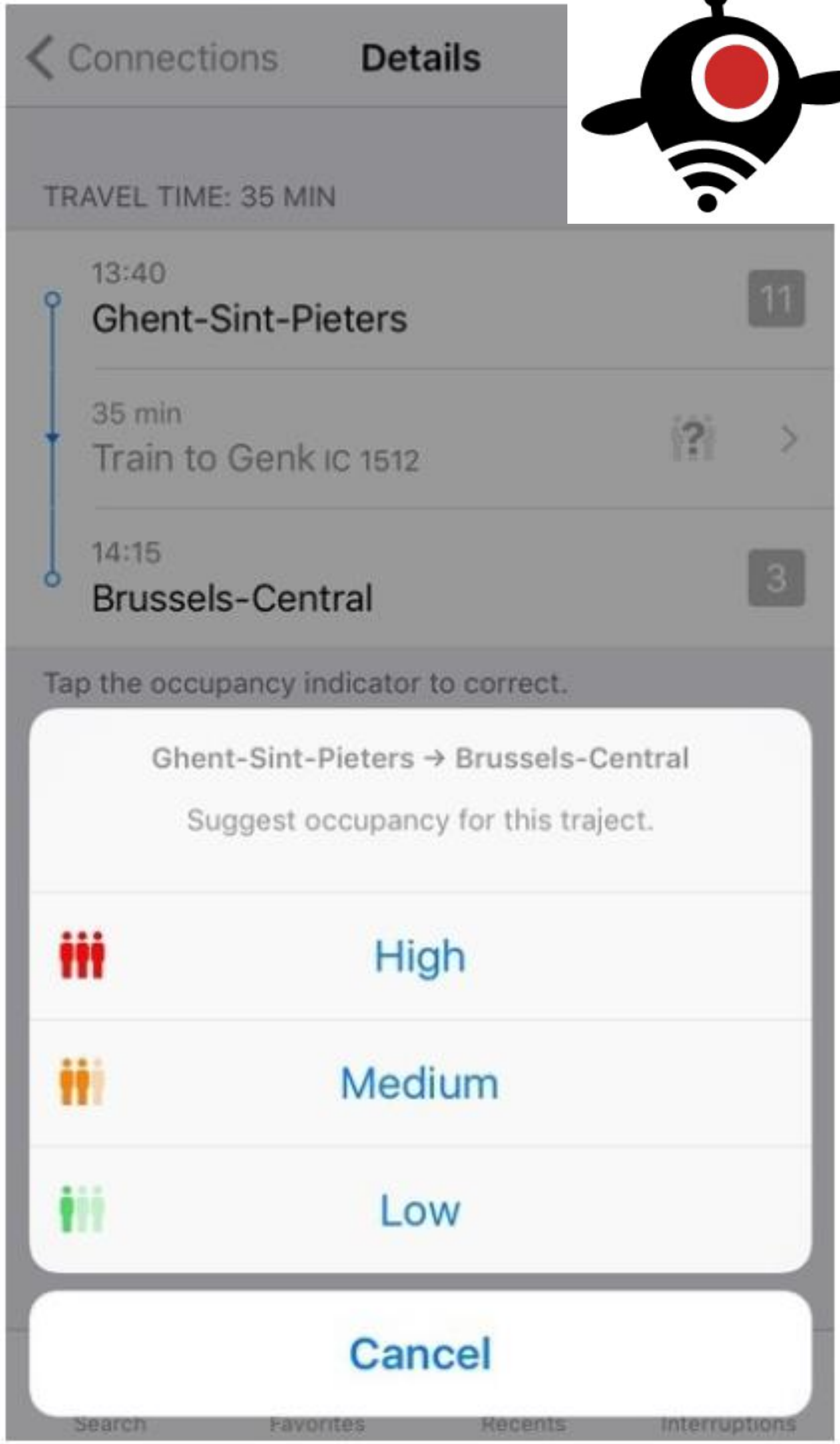
SPITSGIDS

A crowd-funded project in 2016 led by iRail and TreinTramBus

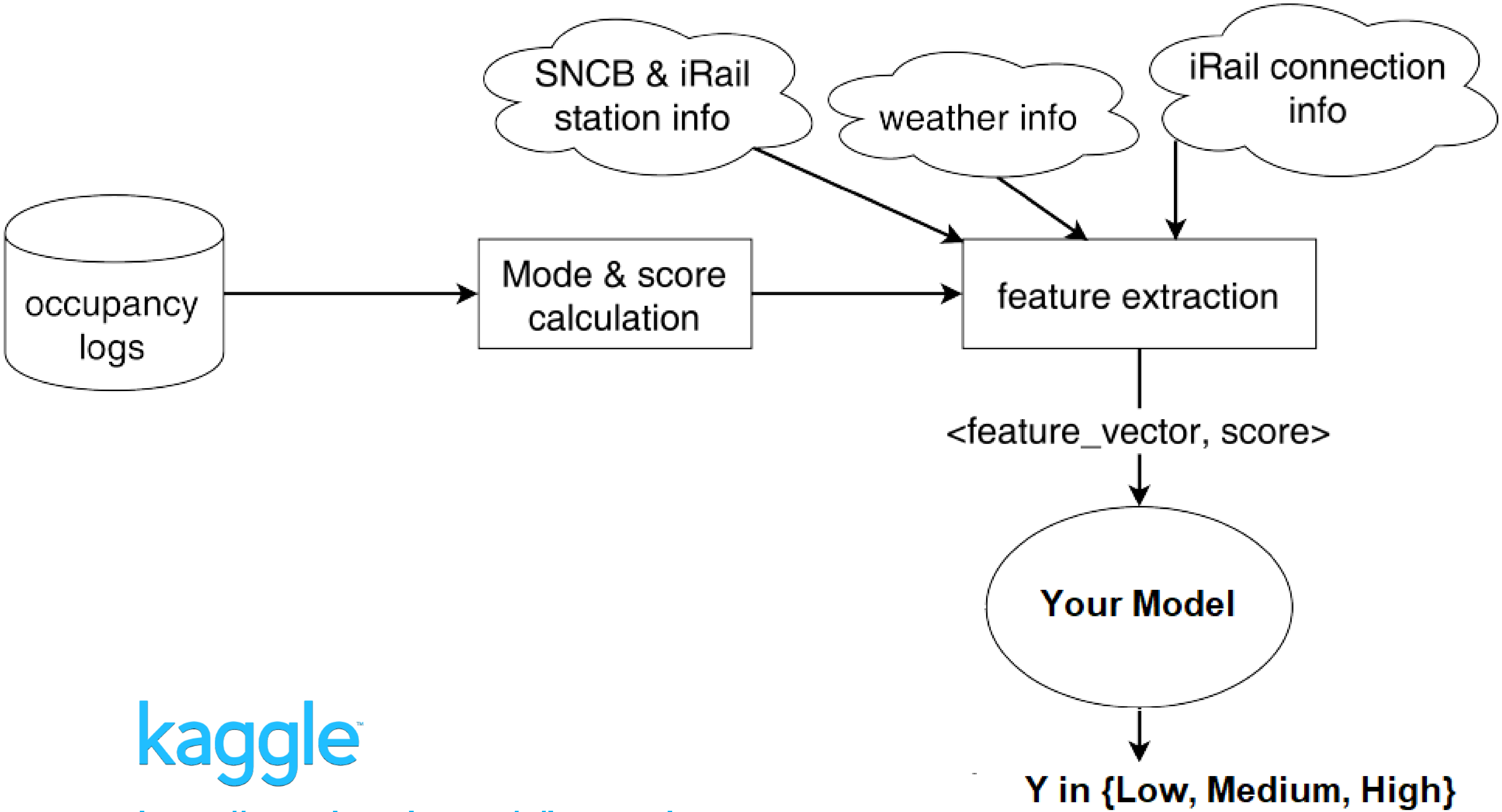
– **Question:**
“How can we avoid busy trains?”

  – **Idea:** Motivate people to take another train by informing them about train occupancy beforehand.

  – **Solution:** Add a module to iRail to measure train occupancy



YOUR MISSION: BUILD A WINNING MODEL

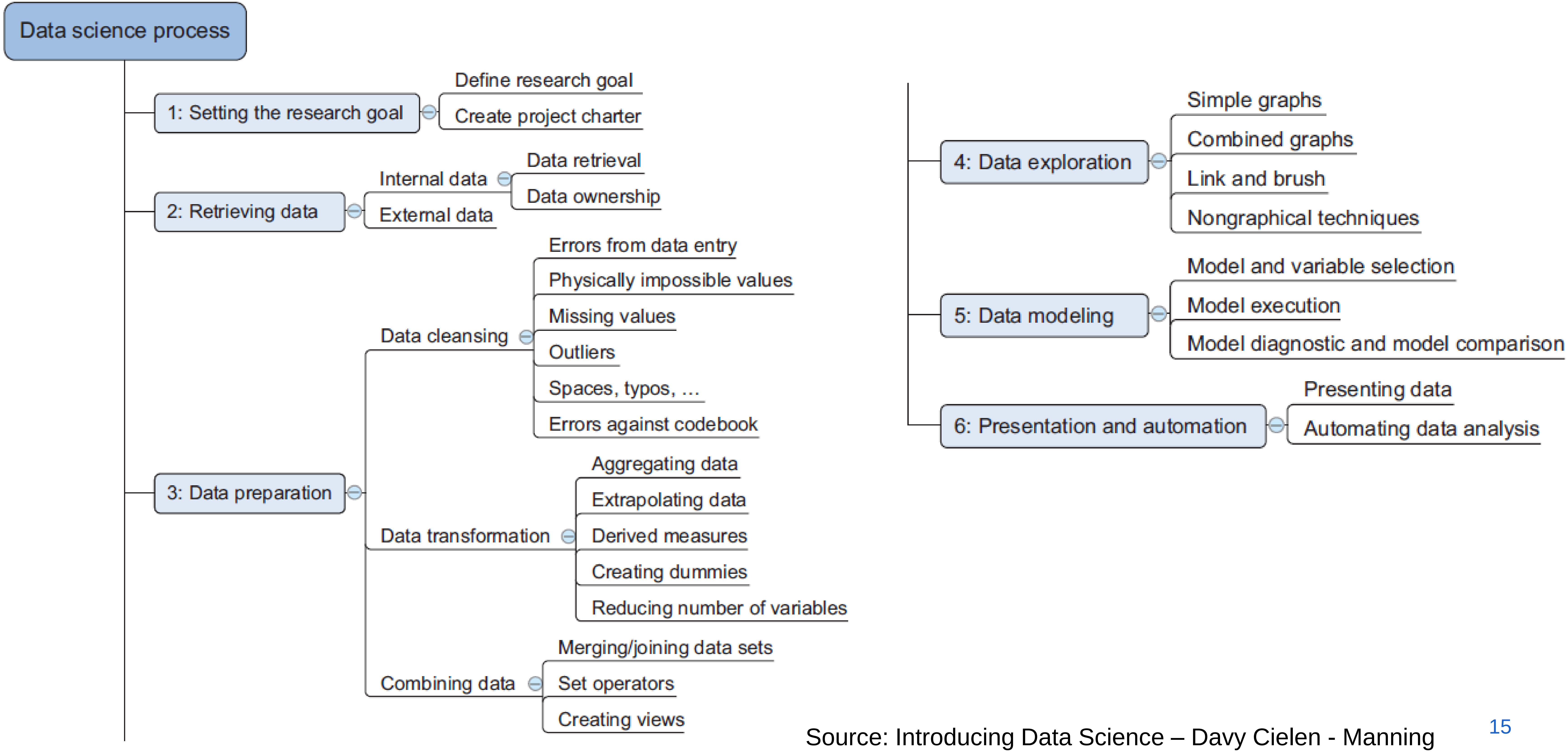


kaggle

https://www.kaggle.com/c/lyon-train-occupancy

DATA PREPARATION & EXPLORATORY DATA ANALYSIS

THE DATA SCIENCE PROCESS (ITERATION)

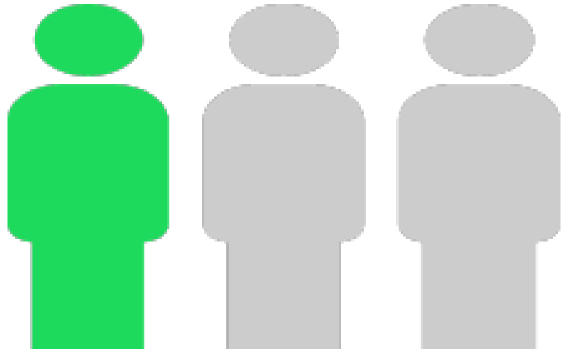


Source: Introducing Data Science – Davy Cielen - Manning

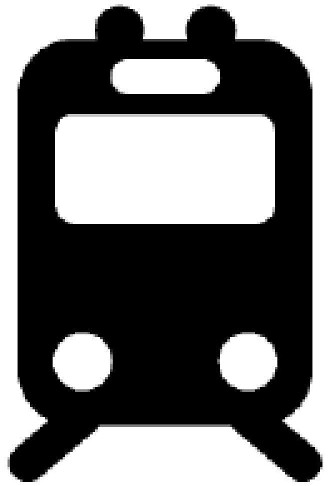
UNDERSTANDING THE DATA DOMAIN (1)

– What is contained in the Query Logs?

Query time



Occupancy



Vehicle ID
(structural)

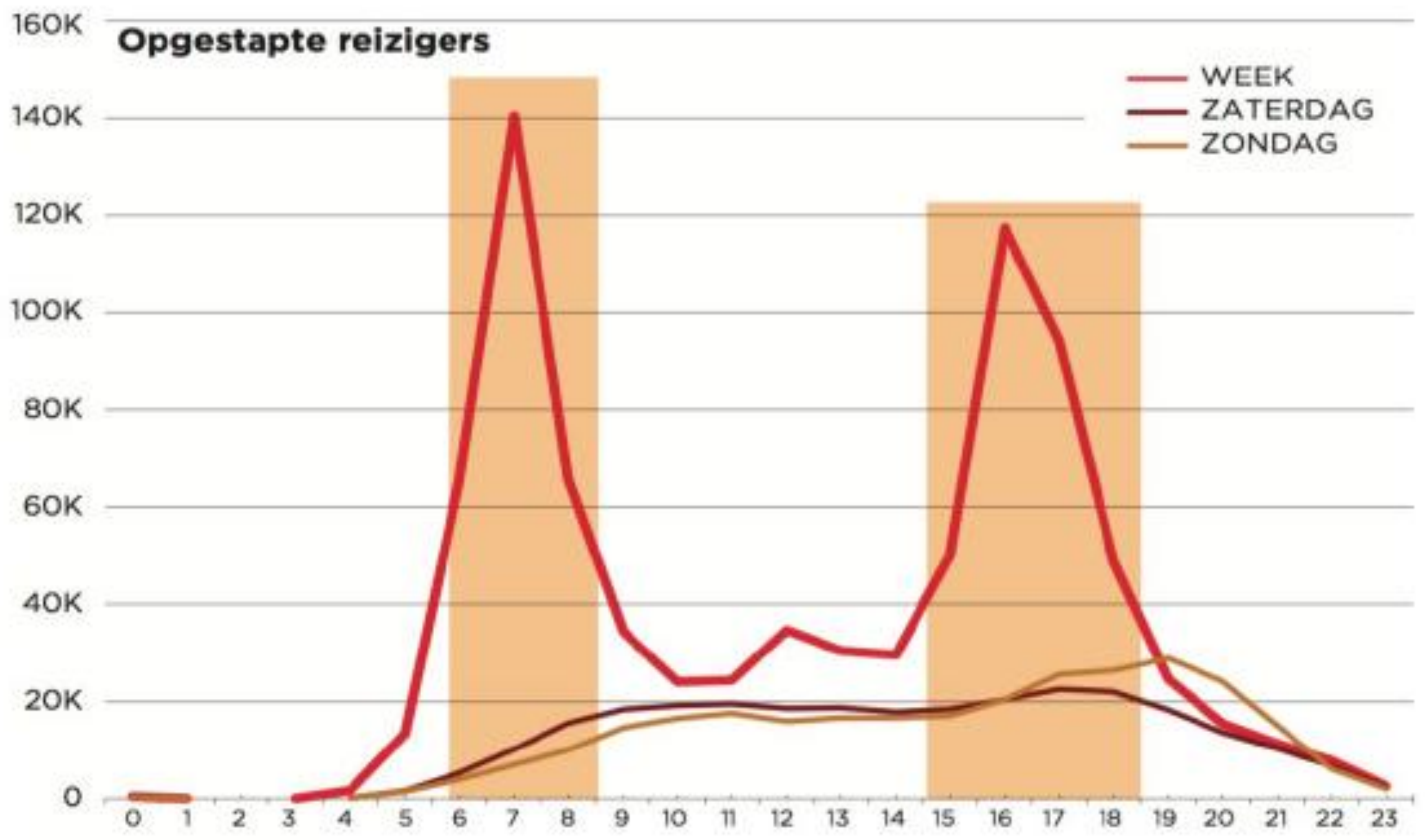
Departure station

Arrival station

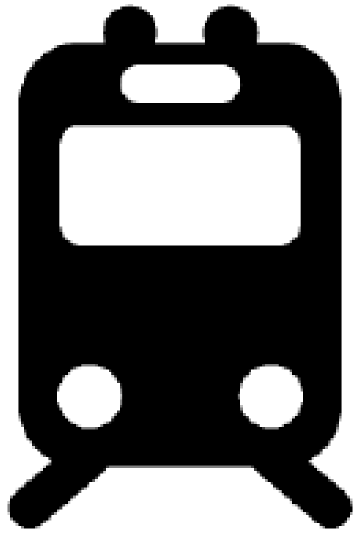
UNDERSTANDING THE DATA DOMAIN (2)

– What ‘features’ can we extract from the query time?

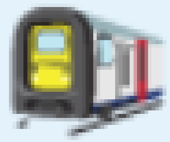
Query time



– What does the Vehicle ID say?



Vehicle ID (structural)



IC 507 03/02/2018 From: Oostende

To: Eupen

- Vehicle Types: L, IC, P,… (IC = InterCity)
- 500 Train Series (Oostende -> Eupen)
- 7: is the sequence number (7:32 → 10:43)
- Seq nr. +25: Eupen -> Oostende!

http://www.belgianrail.be/en/

https://nl.wikipedia.org/wiki/Treincategorie%C3%ABn_Belgi%C3%AB

DATA PREPARATION: WHY & HOW?

- Machine Learning Algorithms require:
  - A **dense** dataset in **m x n matrix** form X
    - m Rows = data objects
    - n Columns = different attributes
    - Column values = **numerical** (!)
  - Supervised ML algorithms require a label vector Y
- How to?
  - How to make the data **tabular**?
    - Preprocessing of non-tabular data formats, (ex.: JSON)
  - Make all attribute values **numerical**?
    - Categorical Data Transformation
      - Label Encoding: {Yes, No} → {0,1}
      - One-Hot-Encoding {Single, Married, Divorced} => {[1,0,0],[0,1,0],[0,0,1]}
  - Improve matrix **quality**
    - Missing data imputation: NaN => average or drop
    - Removing near-duplicate data: averaging
    - Numerical Scaling to improve learning algorithm performance: {125k, 100k, 75k} => {1, 0,8, 0,6} (min-max scaling)
    - Are columns independent? Can we remove some?

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

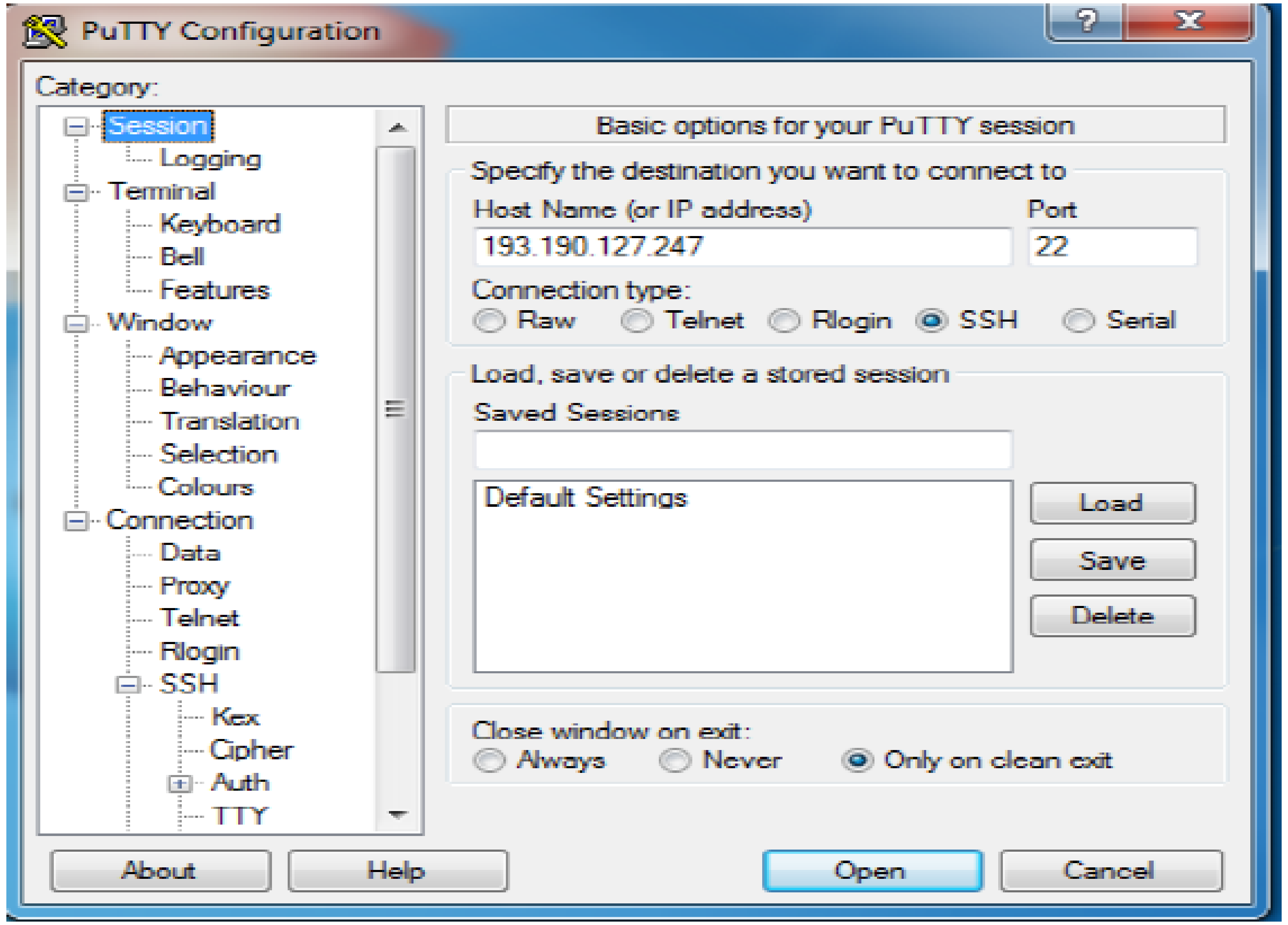
UNDERSTANDING THE DATA WITH EDA

- Do we really need a Machine learning algorithm!?
  - “Seeing is understanding”
    - Humans are quite capable of performing visual pattern recognition!
- Methods for Exploratory Data Analysis:
  - Summary Statistics: mean, median, range, variance,..
  - Visualization: histograms, boxplots, scatter plots, correlation plots

HANDS-ON: SETUP & EDA

SETUP YOUR LOCAL (BIG DATA) ENVIRONMENT

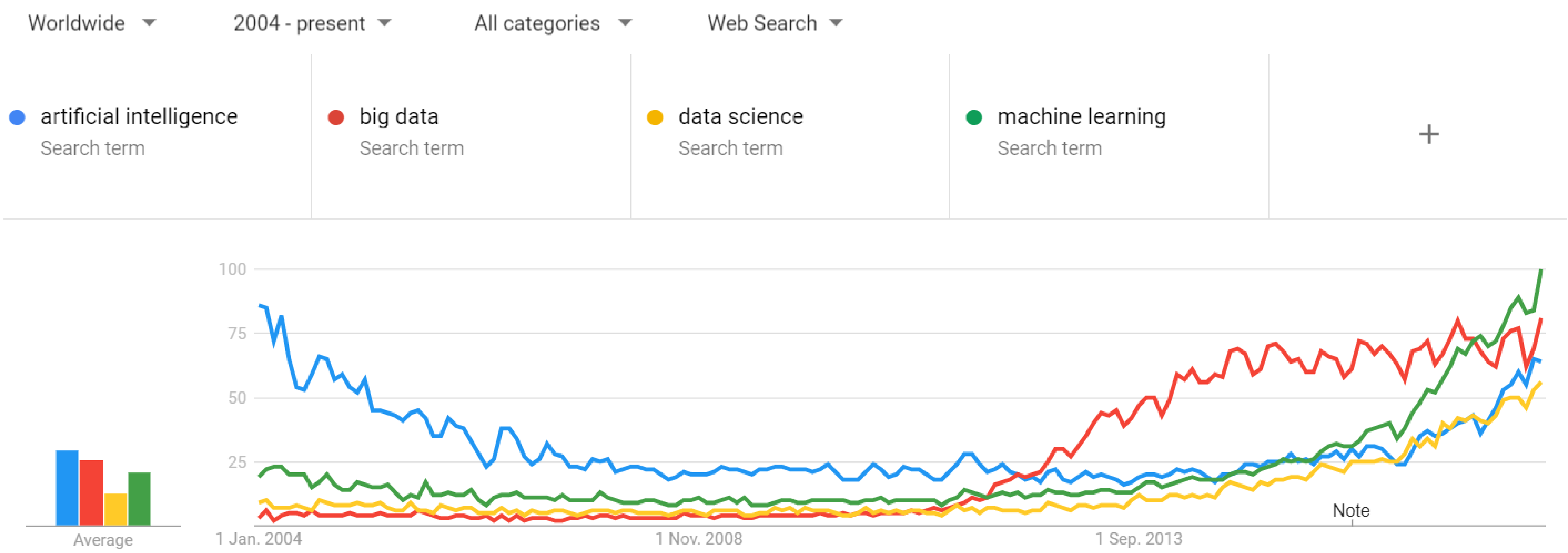
1. ssh-client (www.putty.org)
2. Connect to imec cloud
   – Per 2 you'll get Ubuntu VM
     – Pre-installed with docker
     – Connect via putty to <IP, pwd>
3. We'll use a docker image:
   – https://github.com/jupyter/docker-stacks
   – Learn more about docker:
     – https://github.com/wsargent/docker-cheat-sheet
4. All course material on Github: drdwitte / lyon
   – Lyon_Installation.docx
   – Download/Upload data from Kaggle site
   – Notebooks are numbered *.ipynb files
   – **NOTE: This slide deck is also in the GitHub Repo**
   – Have a look at the pandas cheat sheet



EDA: RECAP

MACHINE LEARNING: QUICK OVERVIEW

BATTLE OF THE BUZZWORDS



Google Trends

SO WHAT IS MACHINE LEARNING?

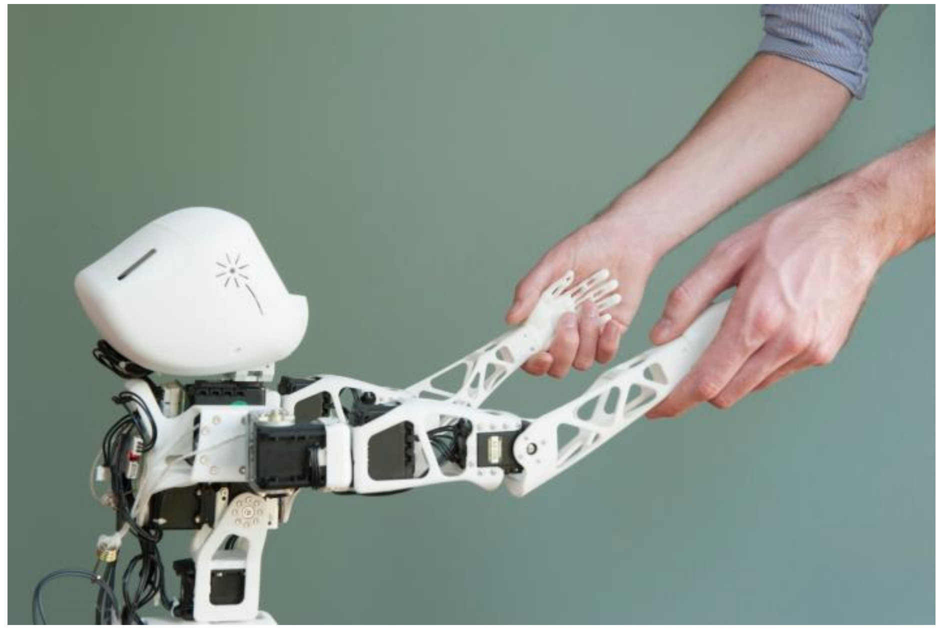
– **Definition:** (Arthur Samuel, 1959)

ML is a field of computer science that gives computers the ability to learn without being explicitly programmed

LEARNING ALGORITHMS

Supervised Learning

(= learning from labeled examples)



dog

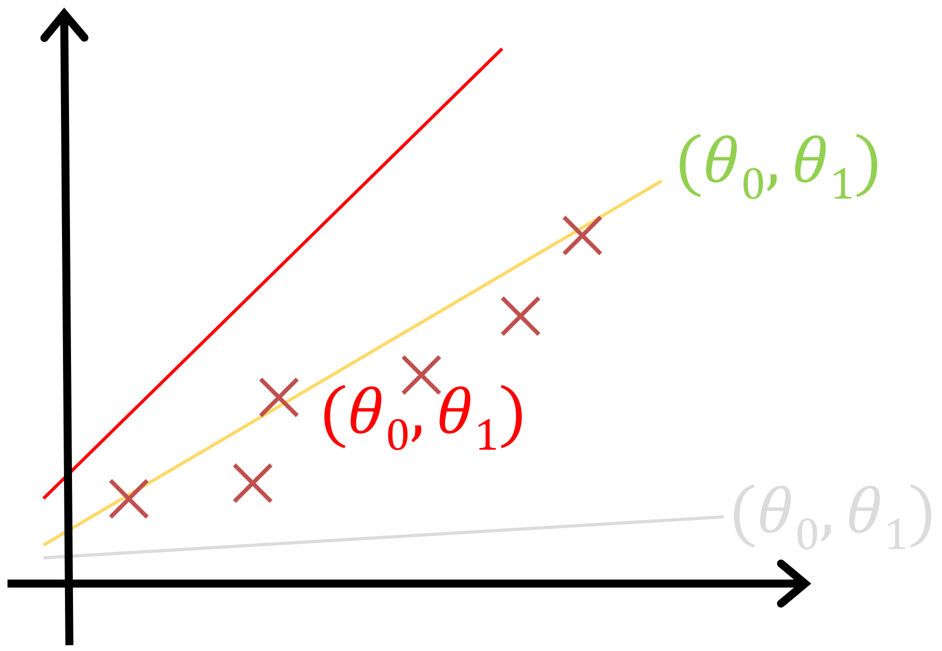
cat

Unsupervised Learning

(= learning from non-labeled examples)

WHAT IS LEARNING?

- Parameters of a machine learning algorithm are often found my minimizing a cost function
  - This minimization process = learning process
  - Minimization often via gradient descent

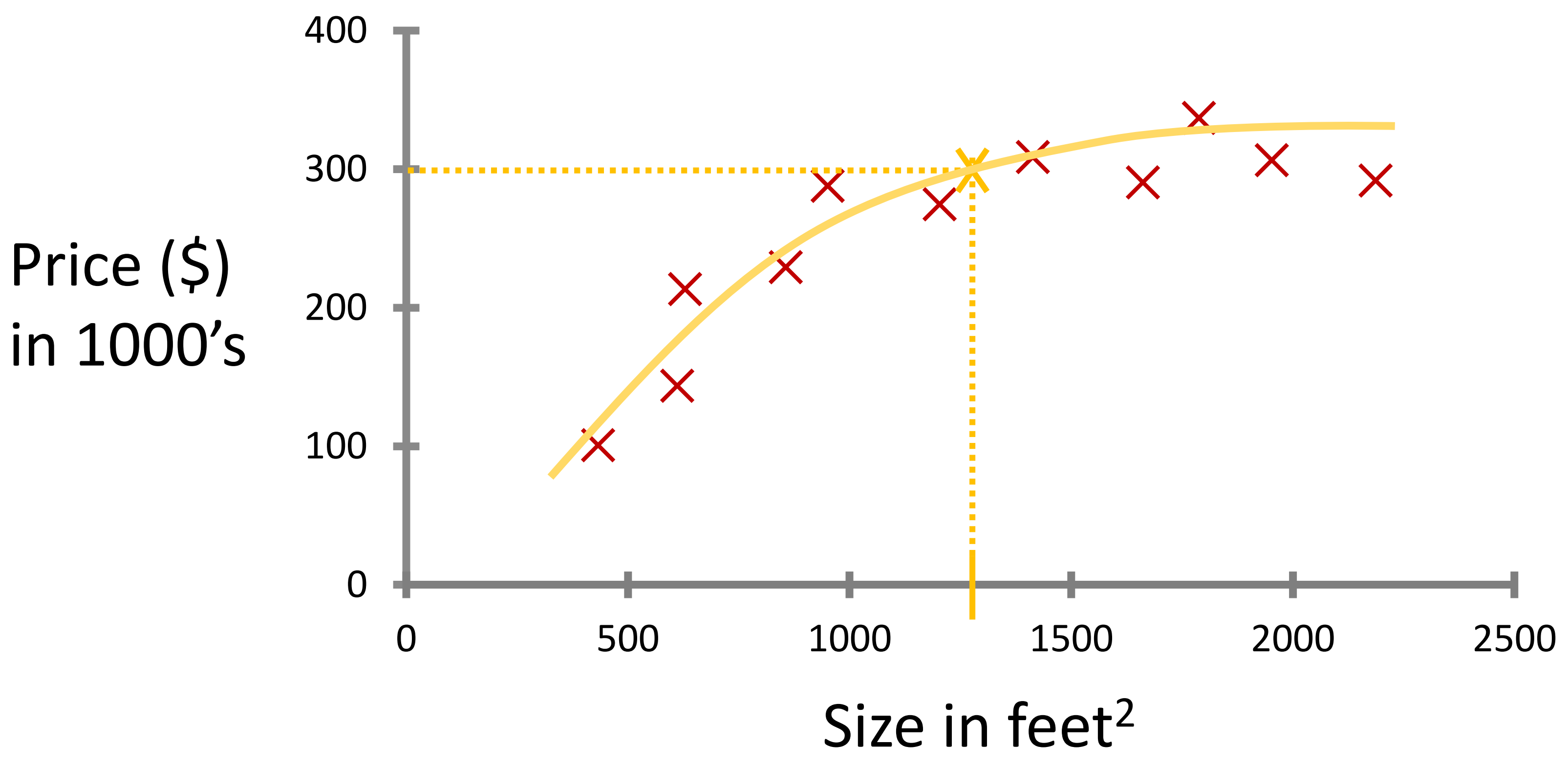


- Example Univariate Linear Regression:
  - Model: $\boldsymbol{h_\theta(x) = \theta_0 + \theta_1 x}$
  - Cost function $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{MSE}{2}$
  - Gradient Descent
    - Stepwise procedure to find minimum of J

SUPERVISED LEARNING

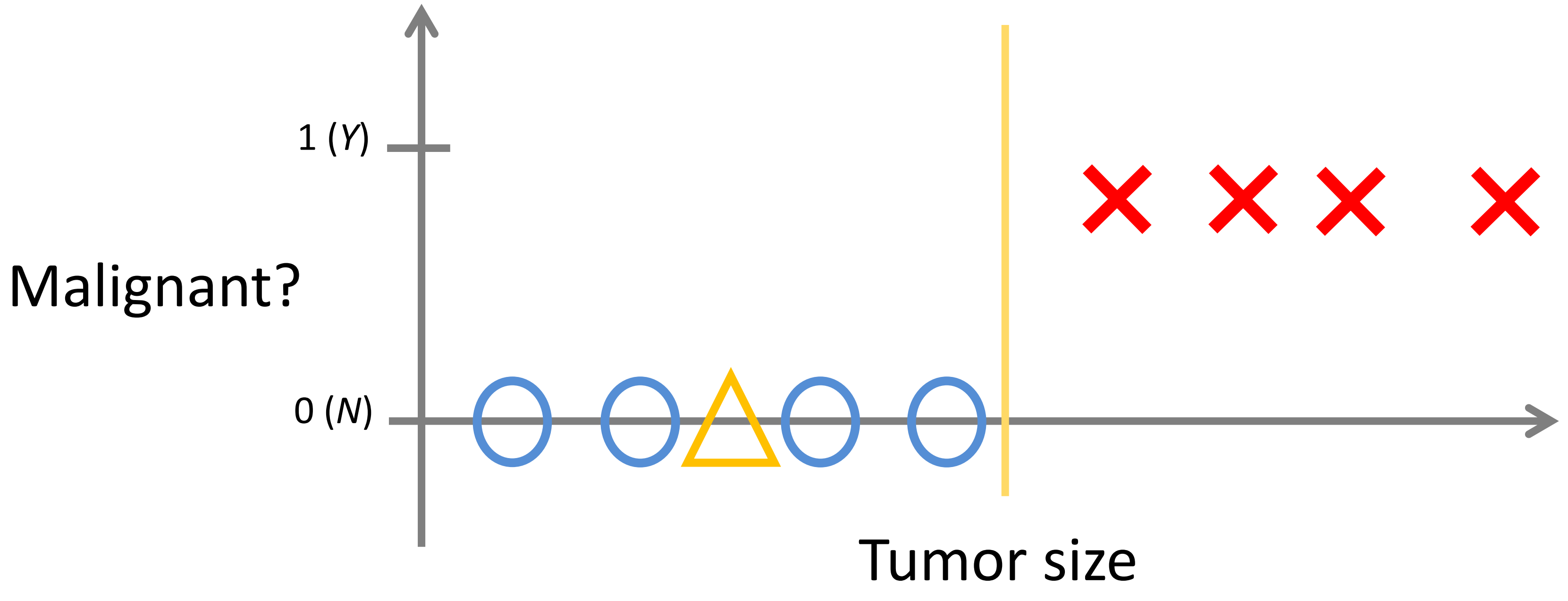
Regression

- Y is continuous
- Ex.: predicting housing prices



Classification

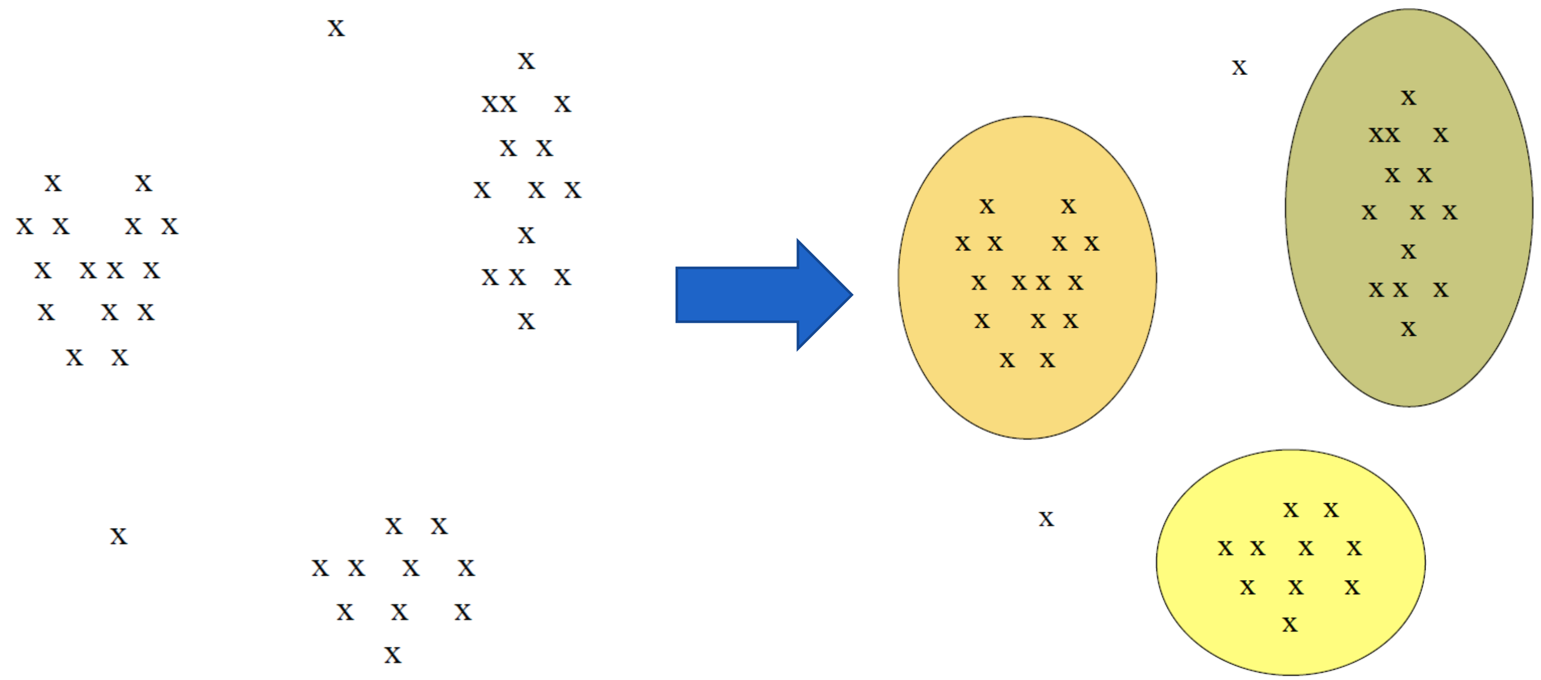
- Y is categorical
- Ex.: tumor is benign/malignant



UNSUPERVISED LEARNING

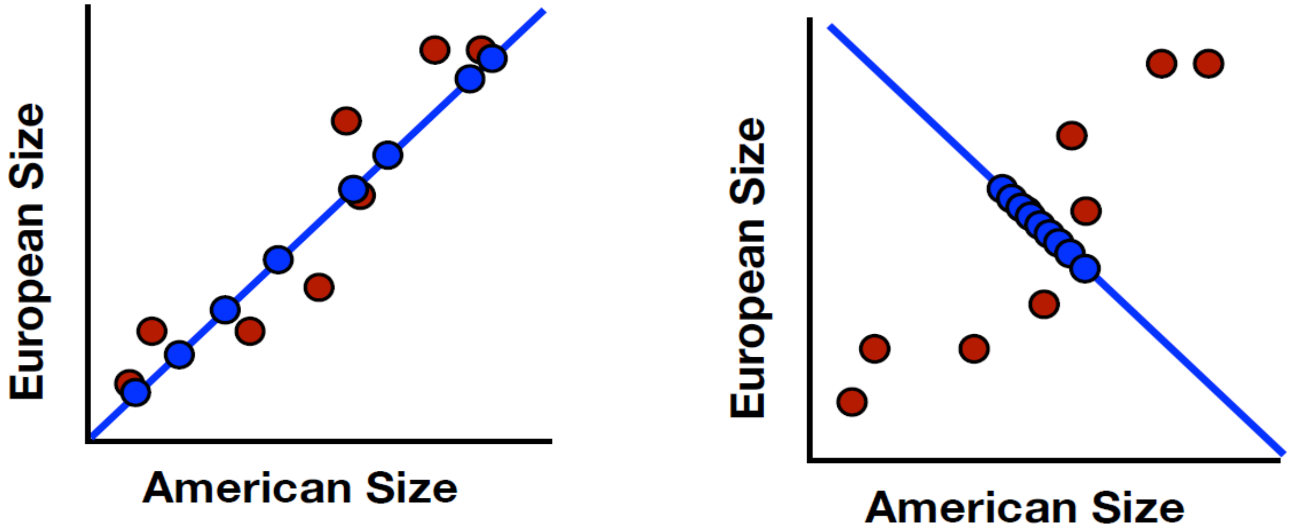
Clustering

- Distance metric d
- Cluster:
  - p, q in same cluster d(p,q) small
  - p, q different cluster d(p,q) large



Dimensionality Reduction

- Curse of dimensionality:
  - in high dimensional spaces almost all p,q are at the same d
- Projection:
  - on a lower dimension with minimal information loss

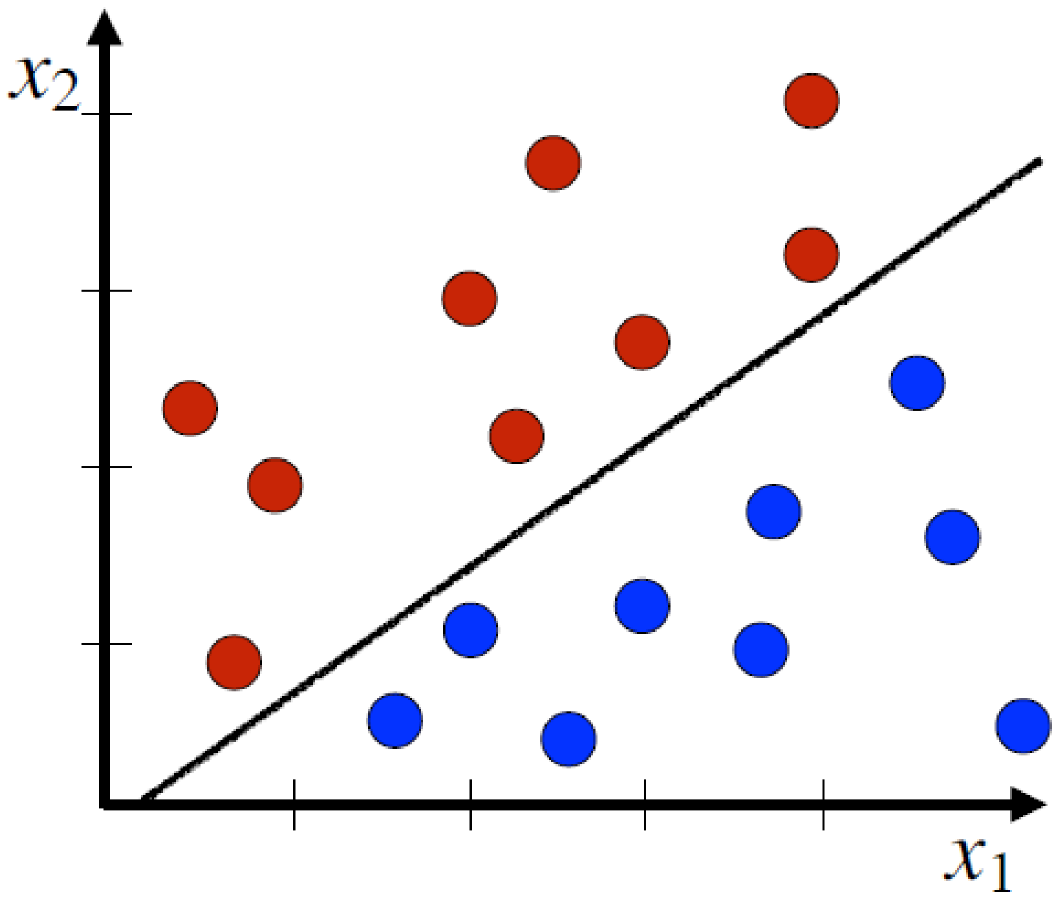


Which of these Machine Learning types do we need for occupancy prediction?

CLASSIFICATION ZOO: LOGISTIC REGRESSION

– Find the parameters w which..

$$\hat{y} = \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$$



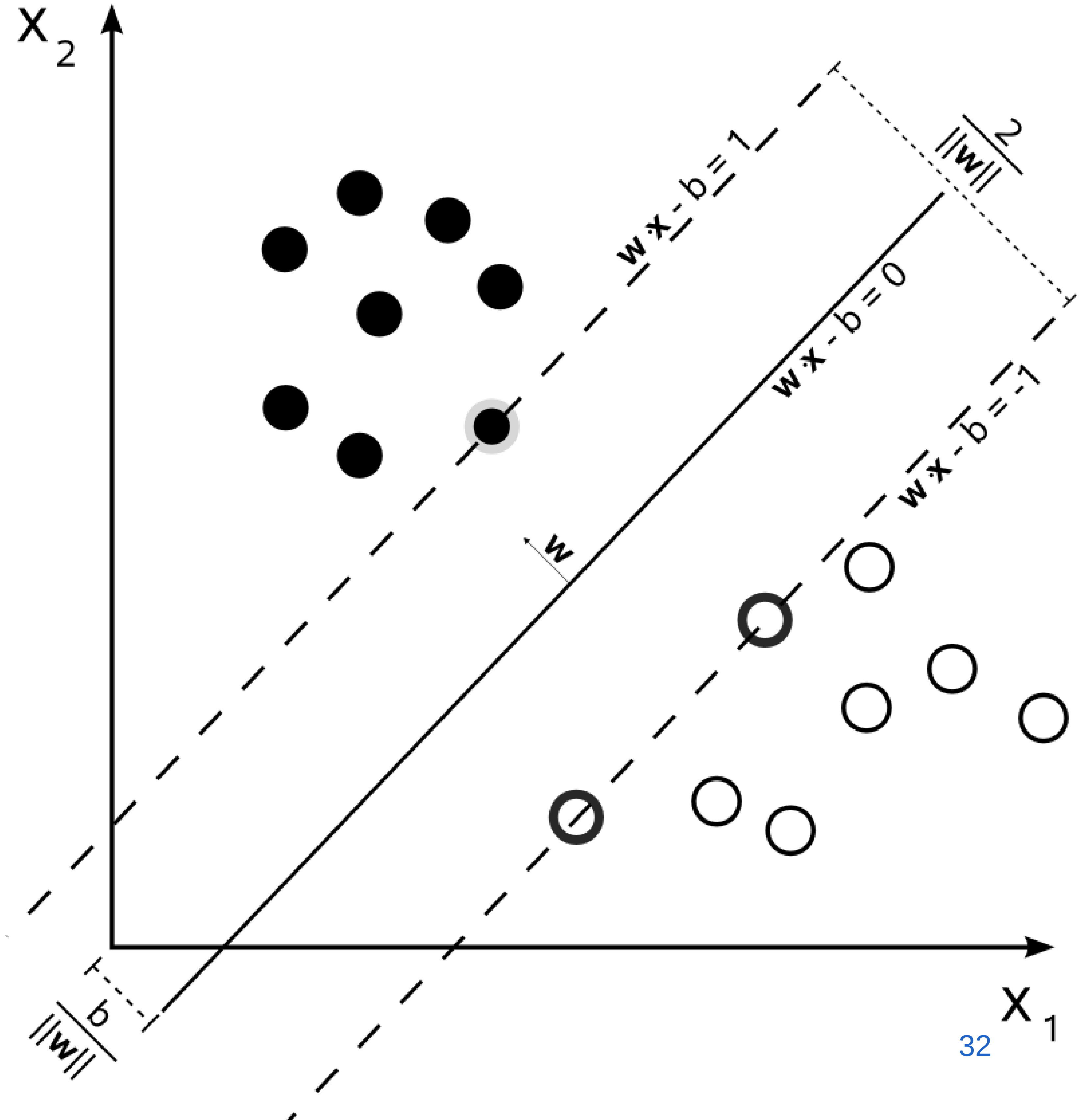
– … separates 2 classes

– Decision boundary: $\mathbf{w}^\top \mathbf{x}$

– Returns class probabilities: P[y=1 | x] = sigmoid($\mathbf{w}^\top \mathbf{x}$)

ZOO: SVM

- Support Vector Machine
  - is a learning algorithm that tries to find the optimal decision boundary
- Optimal:
  - maximize the width of the 'gutter' b (support vectors)
- Distance = kernel(x,y)
  - Often Gaussian, but kernels for many cases exist:
    - String kernels
    - Graph kernels
    - …

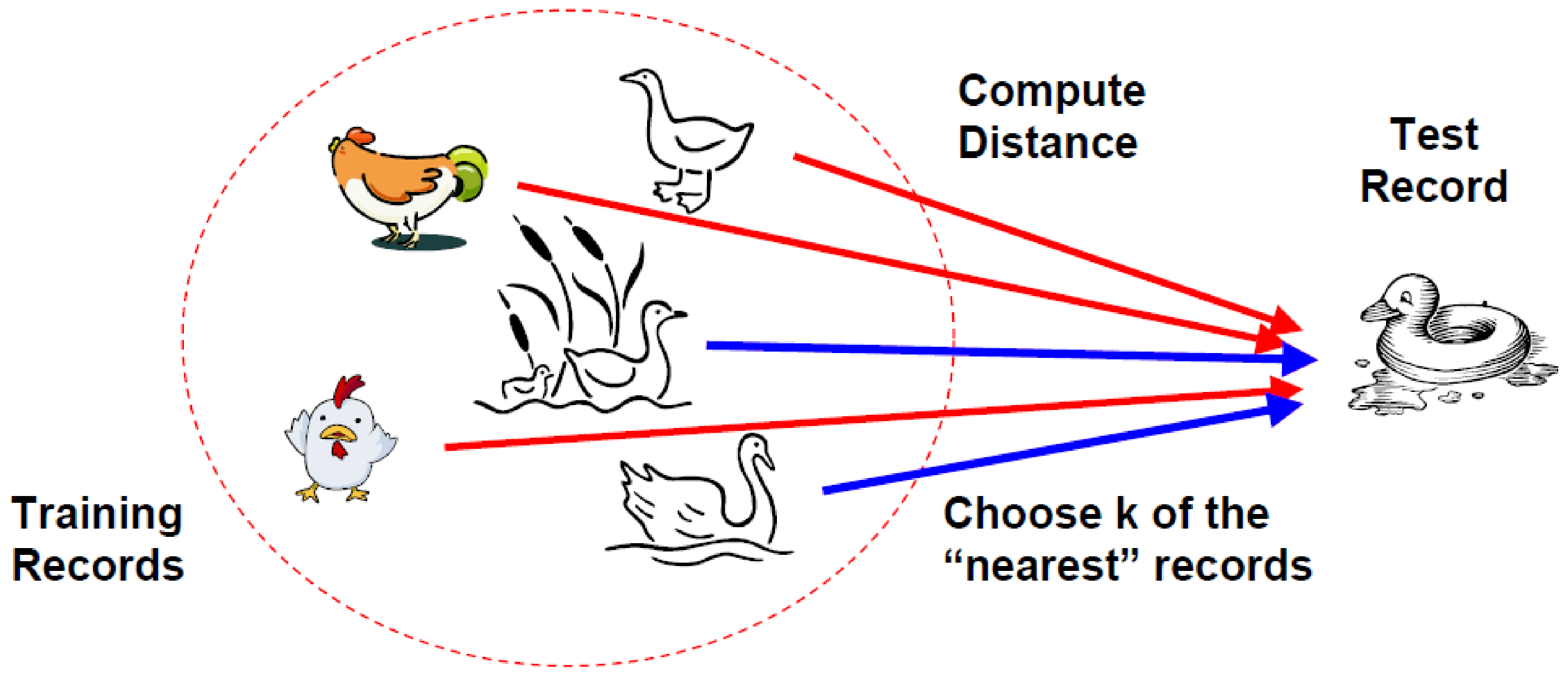


ZOO: INSTANCE-BASED LEARNING

- Classify by memorizing training data and use k most ‘similar’ items to vote on class label (k nearest neighbours)

- “If it walks like a duck, quacks like a duck, then it’s probably a duck”

- Drawback:
  - Curse of dimensionality
  - Many distances to calculate!



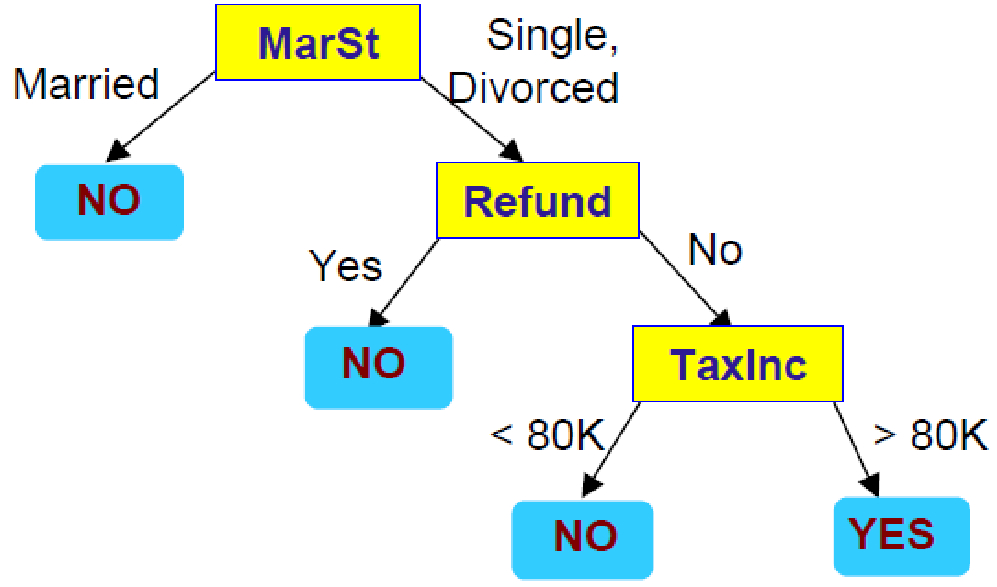
ZOO: DECISION TREES

- Greedy (top-down) algorithm that tries to split the data based on an attribute test in order to optimize a purity criterion

- Impurity:
  - Gini index
  - Entropy
  - misclassification error

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

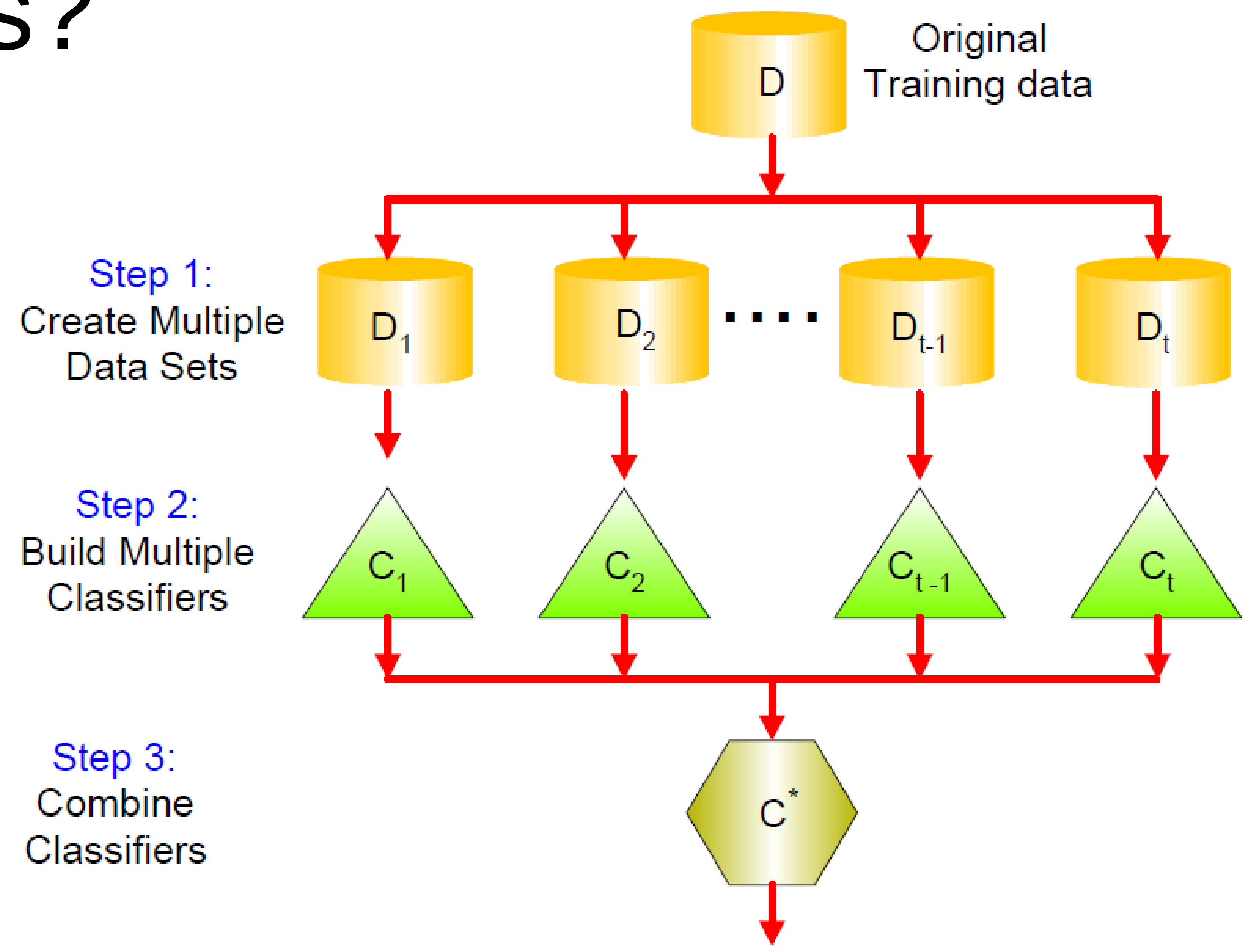


There could be more than one tree that fits the same data!

- ++White box model!
- -- Overfitting issues

ZOO: ENSEMBLE METHODS – RANDOM FOREST

- Decision Trees are sensitive to overfitting (greedy strategy)
- What about training multiple trees?
  - Each tree has
    - a subset of the data
    - a sub selection of the features
- Idea
  - Suppose we have 25 independent trees with error rate 0.35
  - Esemble error rate: $\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$



NAIVE BAYES

- Which class C has the highest probability for a record with features A1,… An?

- How to estimate P(C | A1, A2,.. An)?

  Bayes Theorem!

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

- Naïve? P(A1, A2,…An | C) = P(A1 | C) * P(A2 | C) * …

BUILD YOUR FIRST MODEL(S)

– Have a look at sklearn cheat sheet

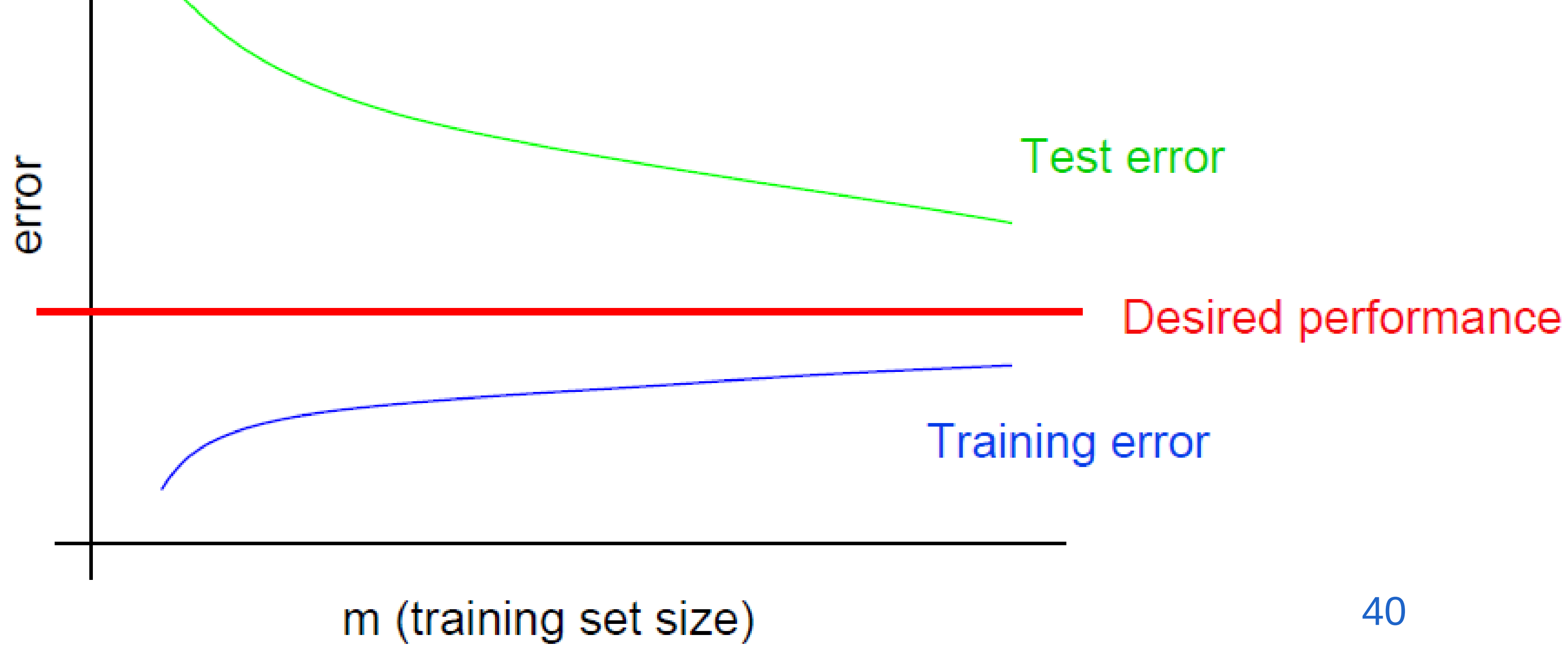
RECAP & TUNING

DIAGNOSTIC I: BIAS VS VARIANCE

– Learning curves shows the evolution of the train and test error as a function of training set size

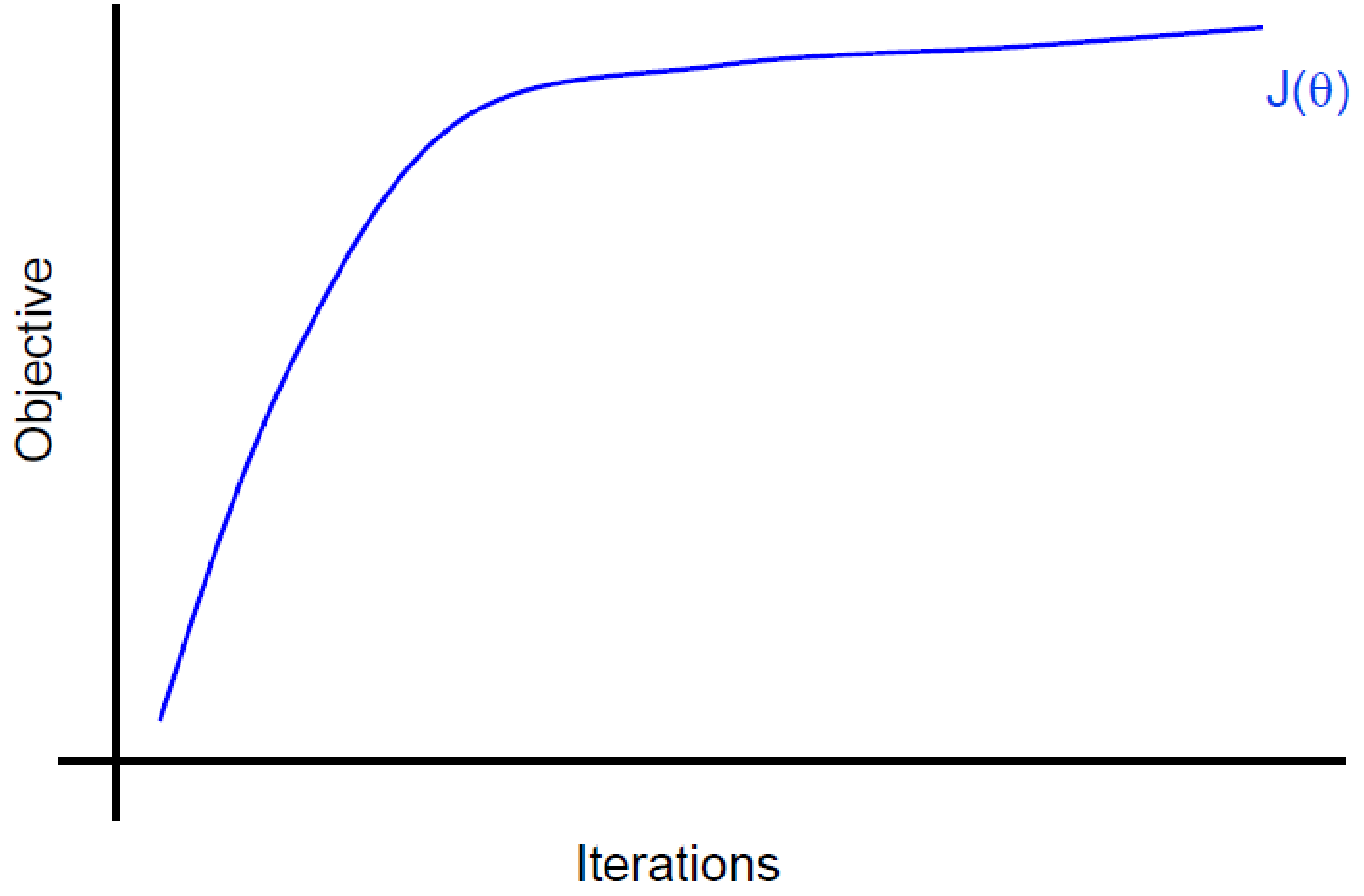
BIAS: model is too simple

VARIANCE: model is overfitting => bigger m or less features



DIAGNOSTIC II: MODEL CONVERGENCE

– Plot the cost function as a function of the number of iterations of the learning algorithm



DIAGNOSTIC III: ABLATIVE ANALYSIS

- For complex ML pipelines with many sub-components we may want to investigate how much does each part add to the overall performance?

- Ablative analysis: remove one component at a time and measure effect on system accuracy

DIAGNOSTIC IV: CONFUSION MATRIX

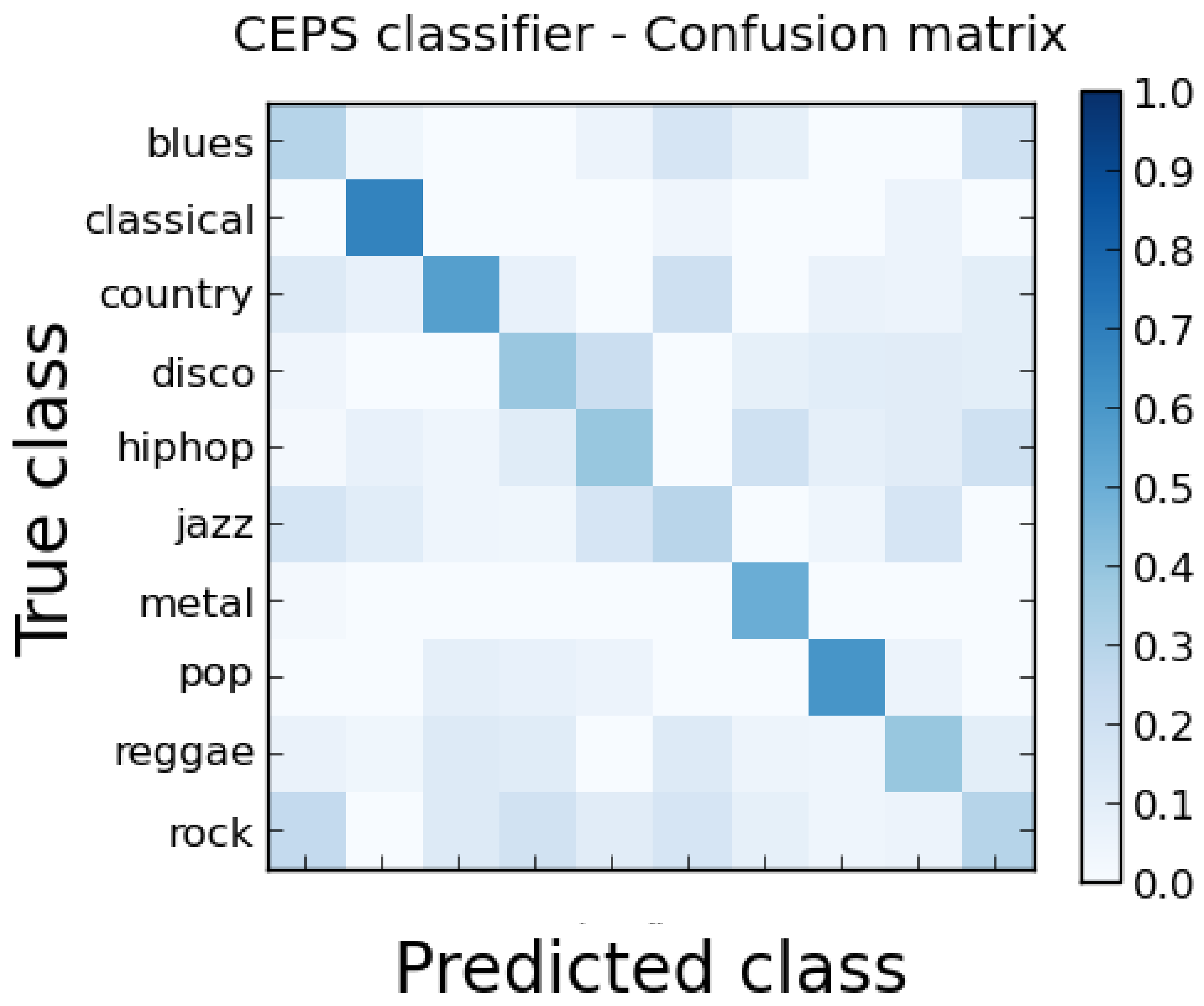
– Comparing predictions to actual class labels

– Where did my algorithm get confused?

$\Rightarrow$ improvement!

– Figure on the right:

  – Jazz is hard to identify
  – Rock and blues are ofte
    mixed up



FINAL SPRINT