DATA SCIENCE IN ACTION FOR EVALUATION

Dharana Rijal, Data Scientist, DECAT

IEG Learning Days

March 2021

IEG Independent Evaluation Group | WORLD BANK GROUP World Bank • IFC • MIGA

Data in Action Toolkit

Key Concepts in Data Science

Bridging theory and application

What is Data Science?

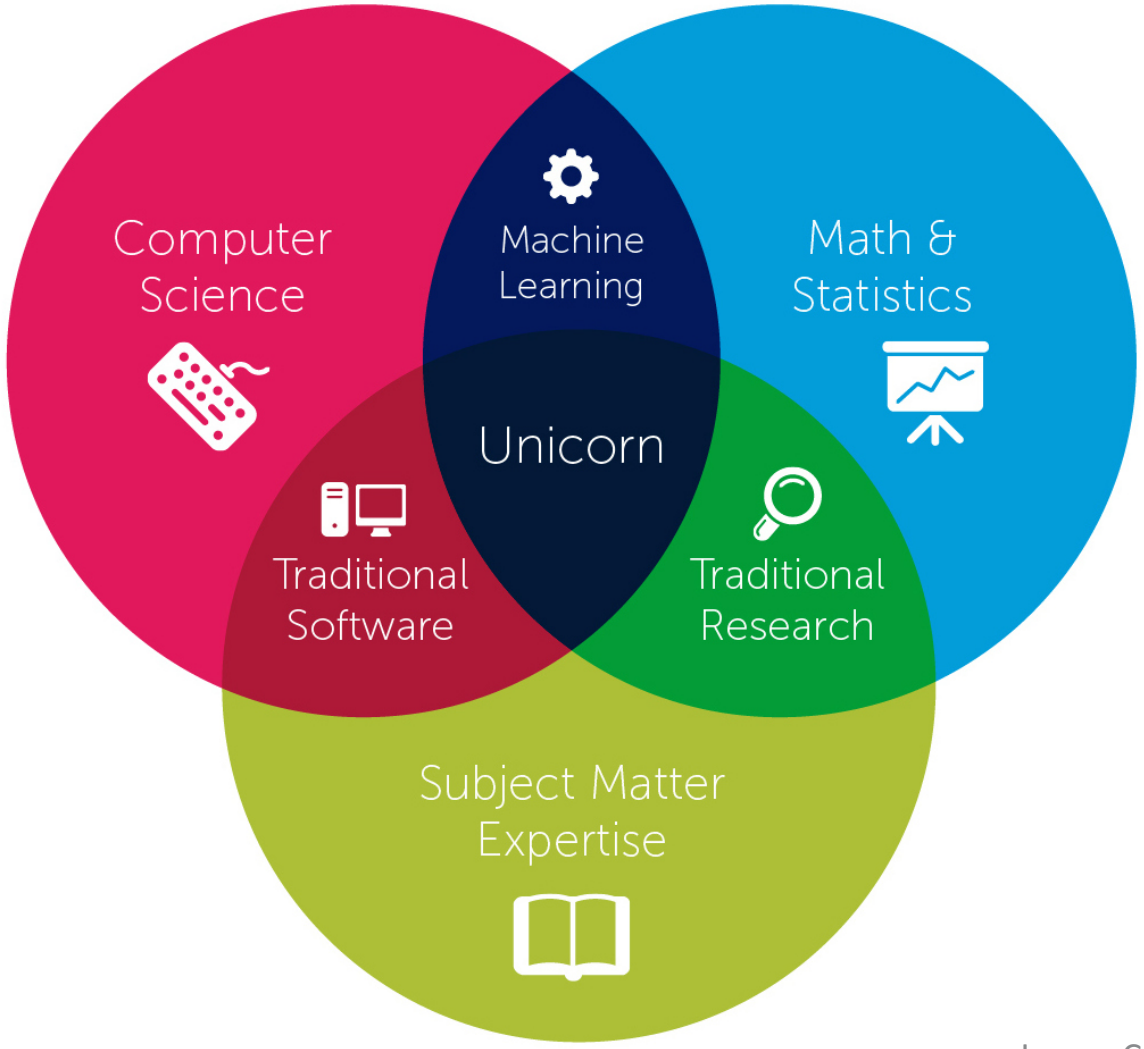
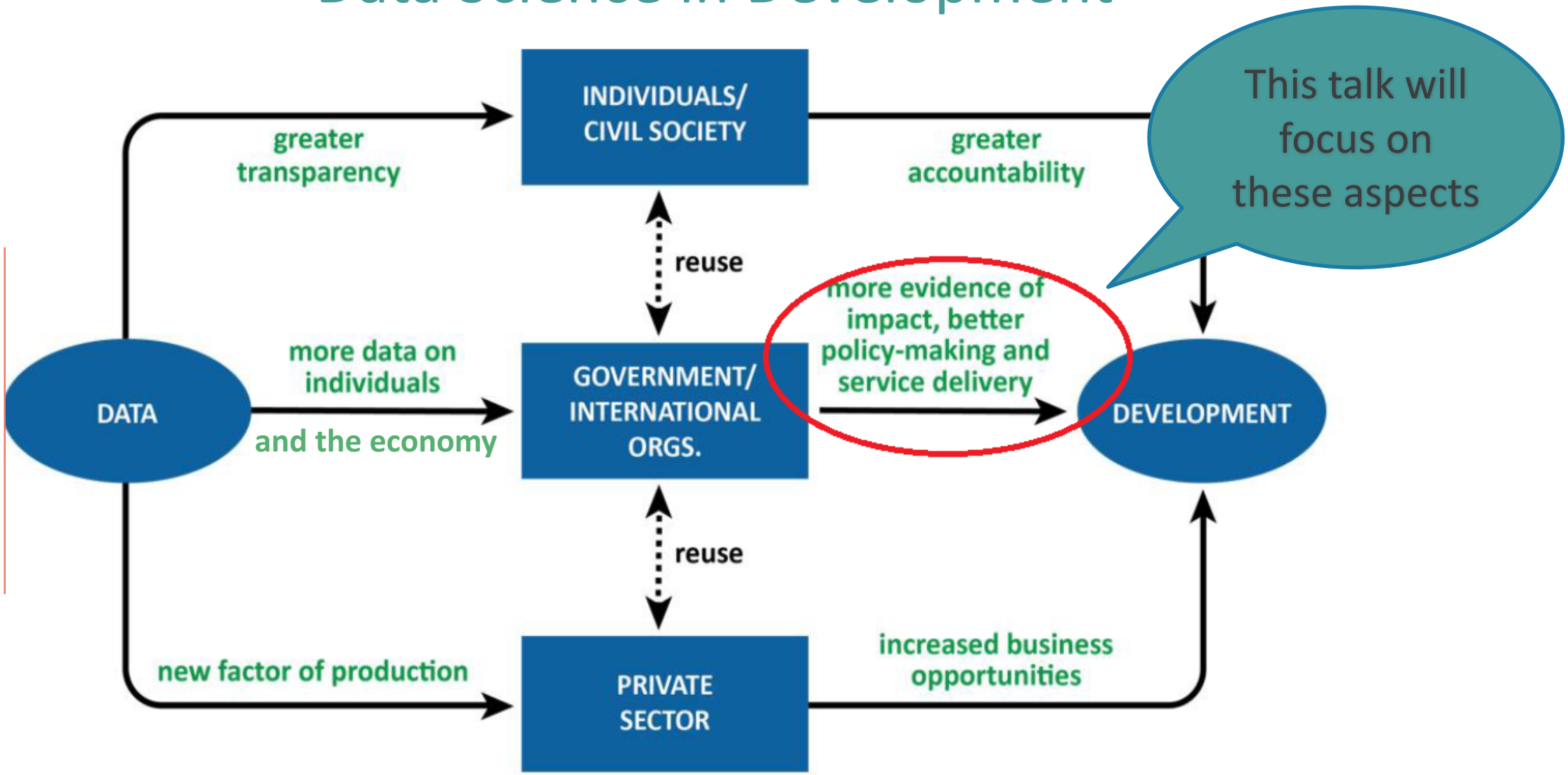


Image Credit: Steven Geringer Raleigh, NC

Data Science in Development



Source: World Development Report 2021: Data for Better Lives: Concept Note

BIG DATA IN ACTION
THEORY OF CHANGE

START WITH A QUESTION

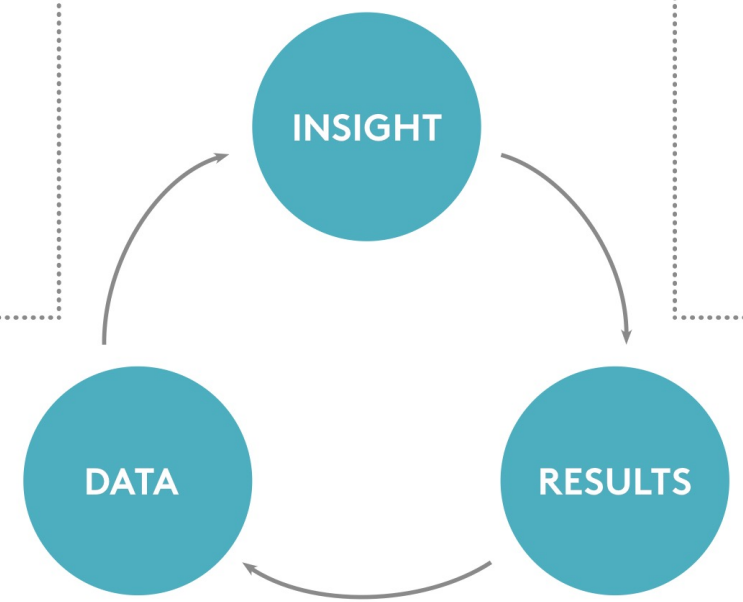
1

DATA INSIGHTS
WHAT INSIGHTS DO WE EXPECT TO FIND?

3

ACTION IMPLEMENTATION
WHAT ACTIONS CAN BE TAKEN, BASED ON INSIGHTS?

4



DATA GENERATION
UNIT, TRIGGER, MEDIA

2

RESULTS
WHAT OUTCOMES OR BEHAVIORS DO WE WANT TO INFLUENCE?

1

STEP # 1

START
WITH A QUESTION

Problem Definition:
What are we trying to do?

Description

Detection

Prediction

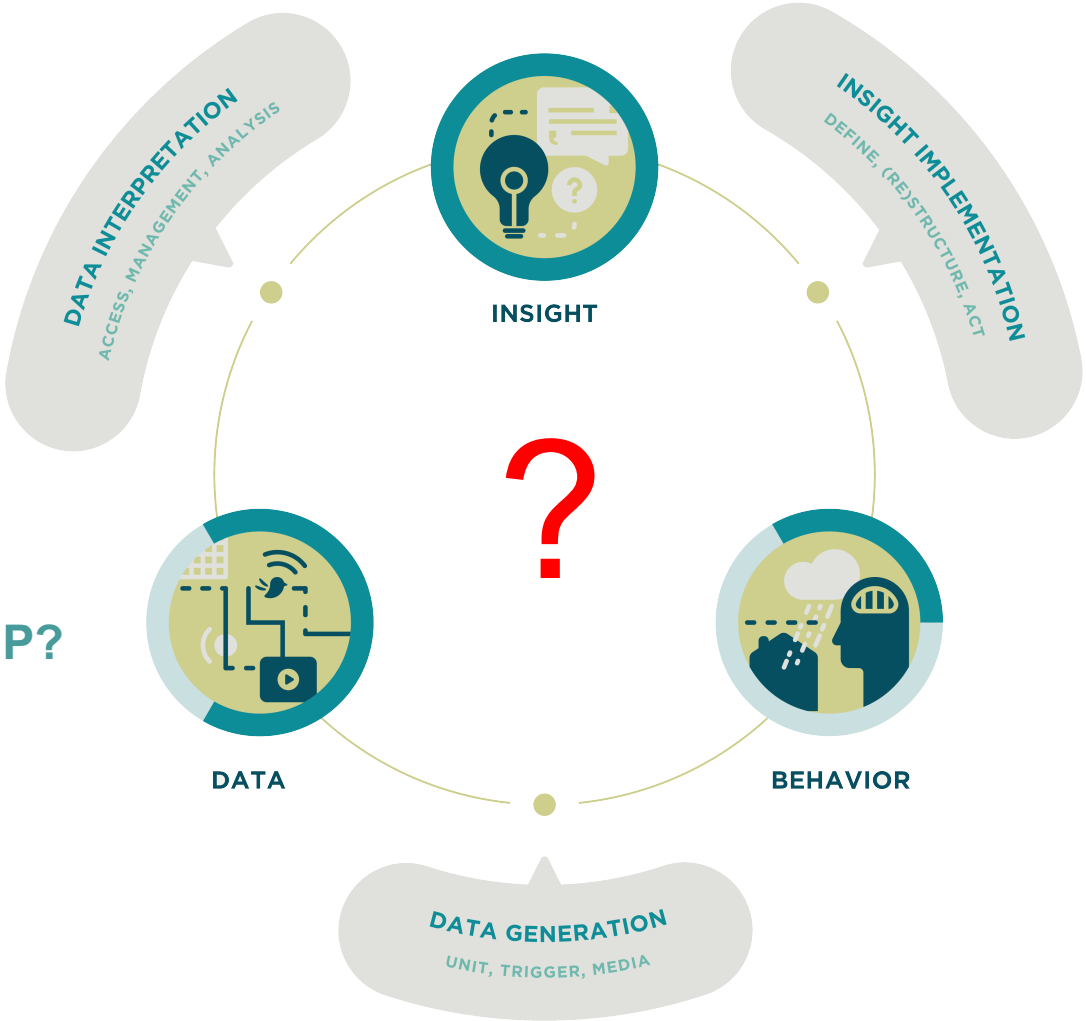
Causal Inference

What are some questions guiding IEG projects?

Data in Action for Evaluation

Have outcomes improved in area X?

How effective was project P?



Can changes in Y be attributed to the WBG's work?

How impactful is WBG's work in area Z?

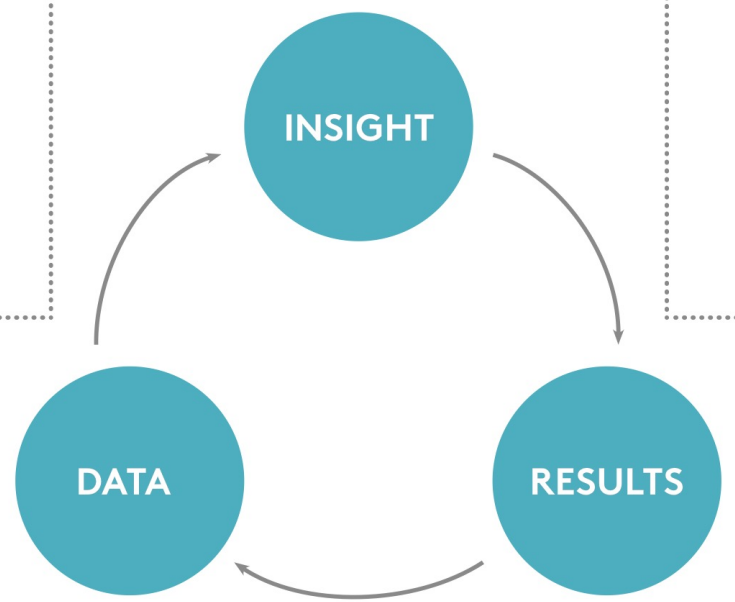
BIG DATA IN ACTION
THEORY OF CHANGE

START WITH A QUESTION

1

DATA INSIGHTS
WHAT INSIGHTS DO WE EXPECT TO FIND?

ACTION IMPLEMENTATION
WHAT ACTIONS CAN BE TAKEN, BASED ON INSIGHTS?



DATA GENERATION
UNIT, TRIGGER, MEDIA

2

RESULTS
WHAT OUTCOMES OR BEHAVIORS DO WE WANT TO INFLUENCE?

1

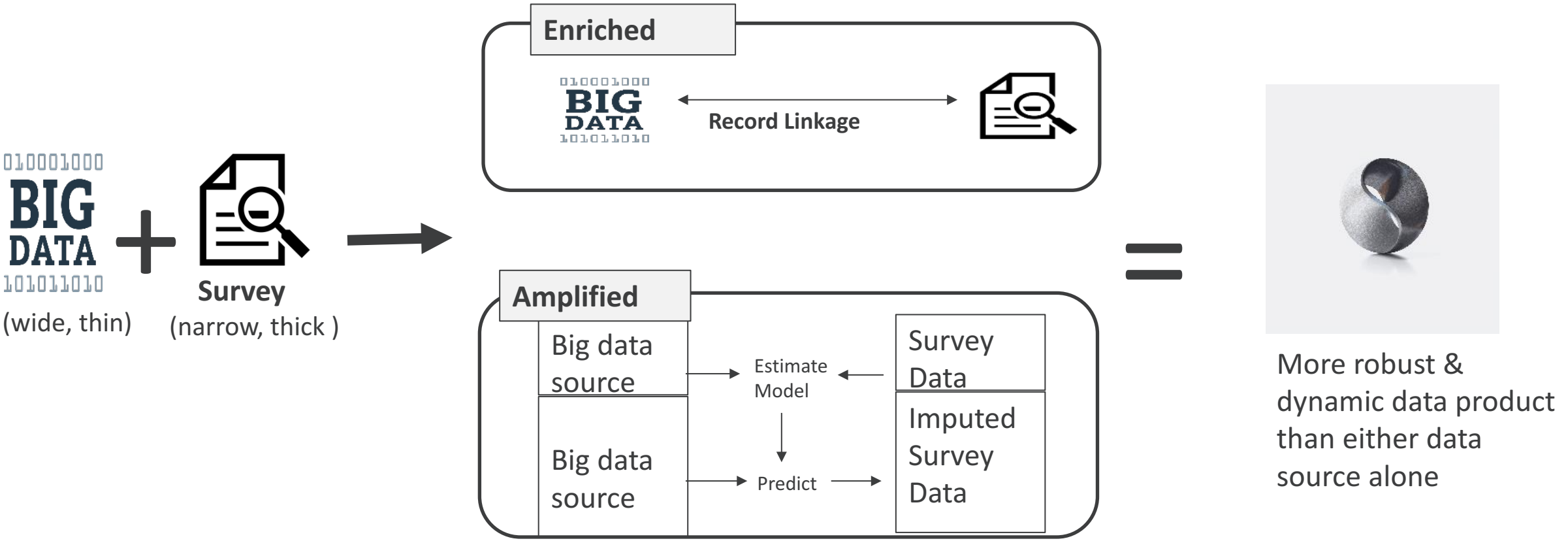
STEP # 2

IDENTIFY DATASETS THAT CAN HELP ANSWER THE QUESTION

“Big data increases the value of survey data”

-Mathew Salganik, Bit by Bit

Common approaches for data integration and fusion



SURVEY DATA VS BIG DATA

SURVEY DATA

The Good:
- Representative
- Standard errors known
- Fit for purpose

The Bad:
- Costly
- Gaps & Lags in Coverage
- Usually not as granular as is needed for some decisions

BIG DATA

The Good:
- Strength in numbers
- Always on
- High temporal & spatial resolution

The Bad:
- Non-representative
- Drifting
- Incomplete

Traditional data	Remote Sensed	Mobile Phone Data	Text based	Other (Crowd-sourced / Internet of Things)
Survey/Census Administrative transaction/ service records	Satellite imagery: nighttime lights; spectrum Drone imagery	Call Detail Records GPS data Telemetry location & mobility	Search Web forums Social networks News aggregators Documents	Sensors/ wearables/Devices Industrial systems Crowd-sourced Open Street Map Audio/Video content

TOP BIG DATA SOURCES USED IN DEVELOPMENT

*adapted from Andrew Whitby

Data Fusion Benefits for Evaluation

- Higher temporal and spatial resolutions
- Longer term assessments
- Compare out of sample (external validity)
- Control for covariates
- Sub-group analysis (heterogeneous effects)
- Efficiency
- Breadth

BIG DATA IN ACTION
THEORY OF CHANGE

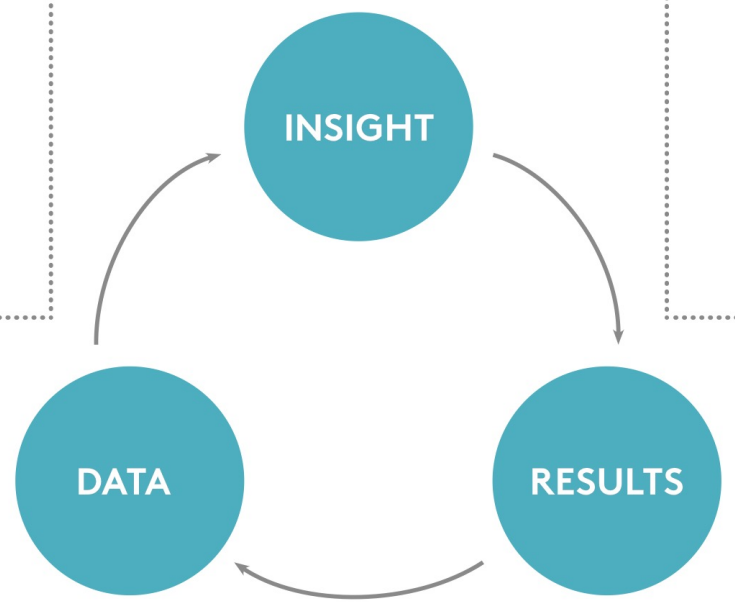
START WITH A QUESTION

1

DATA INSIGHTS
WHAT INSIGHTS DO WE EXPECT TO FIND?

3

ACTION IMPLEMENTATION
WHAT ACTIONS CAN BE TAKEN, BASED ON INSIGHTS?



DATA GENERATION
UNIT, TRIGGER, MEDIA

2

RESULTS
WHAT OUTCOMES OR BEHAVIORS DO WE WANT TO INFLUENCE?

1

STEP # 3

ANALYSIS
FOR RELEVANT INSIGHTS

Machine Learning

Question	Data	Model training	Calibrate	Predict	Evaluate	Results
Question of interest identified.	Relevant Data Obtained.	Data cleaned Feature Engineering Split Train/Test Dataset Model Trained	Model parameters tuned.	Predictions made on test set.	Predictions validated.	Results communicated.

git/GitHub for reusability and smooth collaborative workflows

Ethics, Privacy, Confidentiality, and Security

"More data beats clever algorithms
but better data beats more data."
- Peter Norvig

Machine Learning

Simple definition:
How machines learn rules from examples.

When fitting machine learning algorithms, it is common to split data into training and test sets

Split dataset

Build model on the training set

Evaluate model on the test set

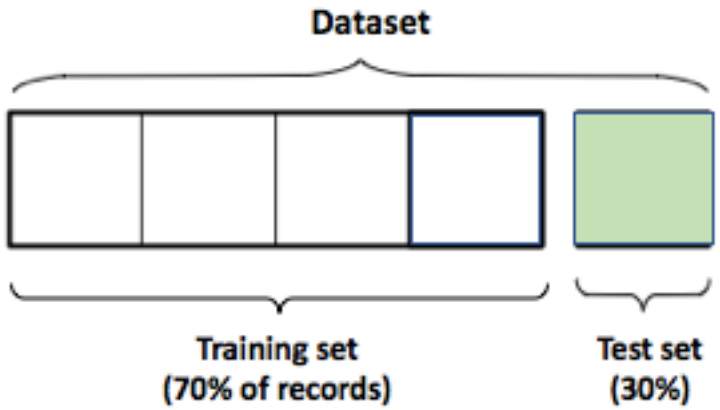


Image credit: D. Ziganto "Standard Deviations" blog

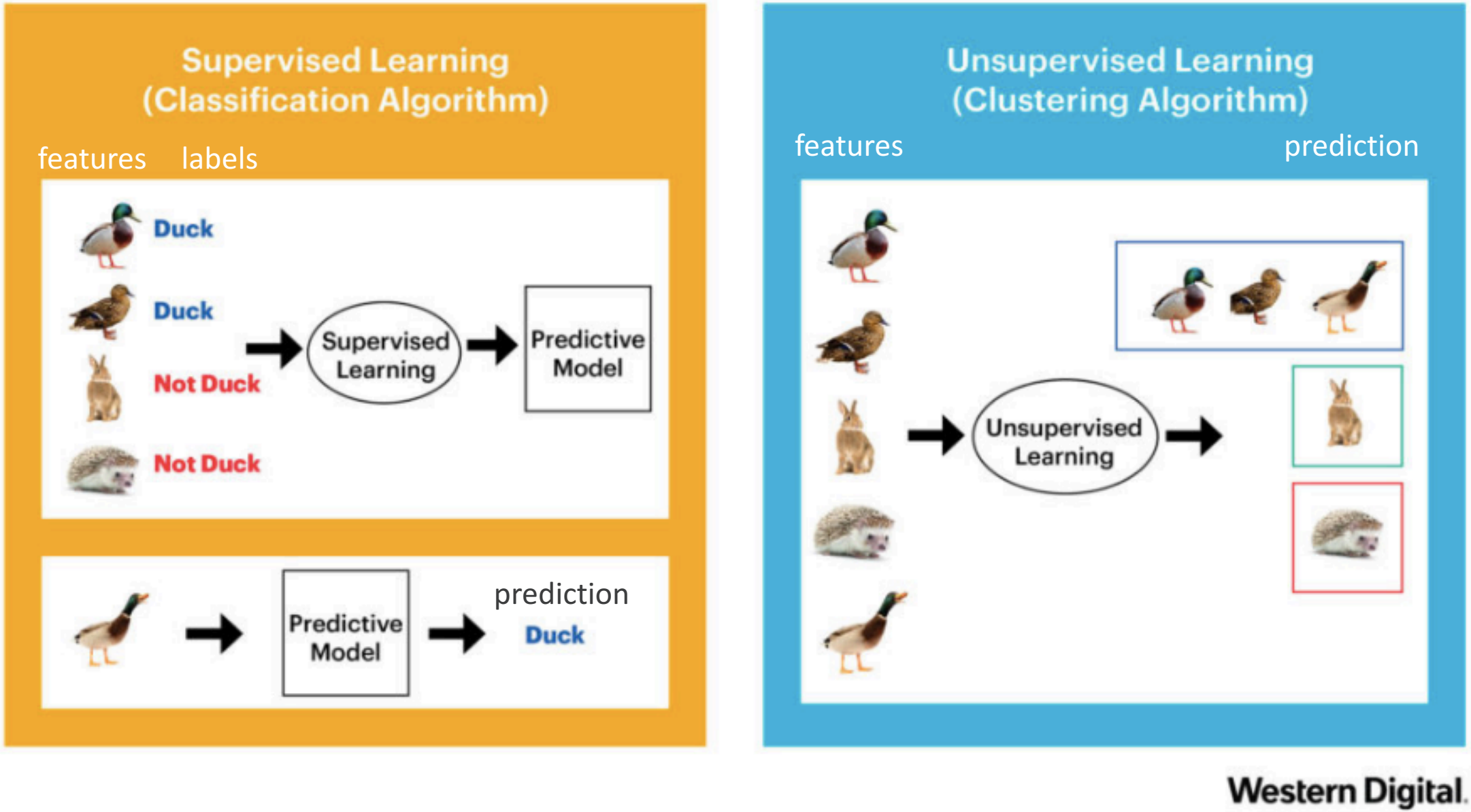
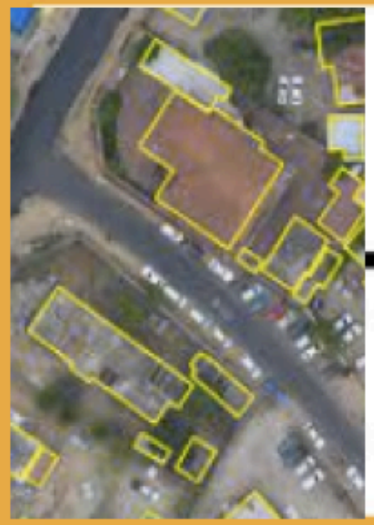


Image Credit: Western Digital (https://blog.westerndigital.com/machine-learning-pipeline-object-storage/)

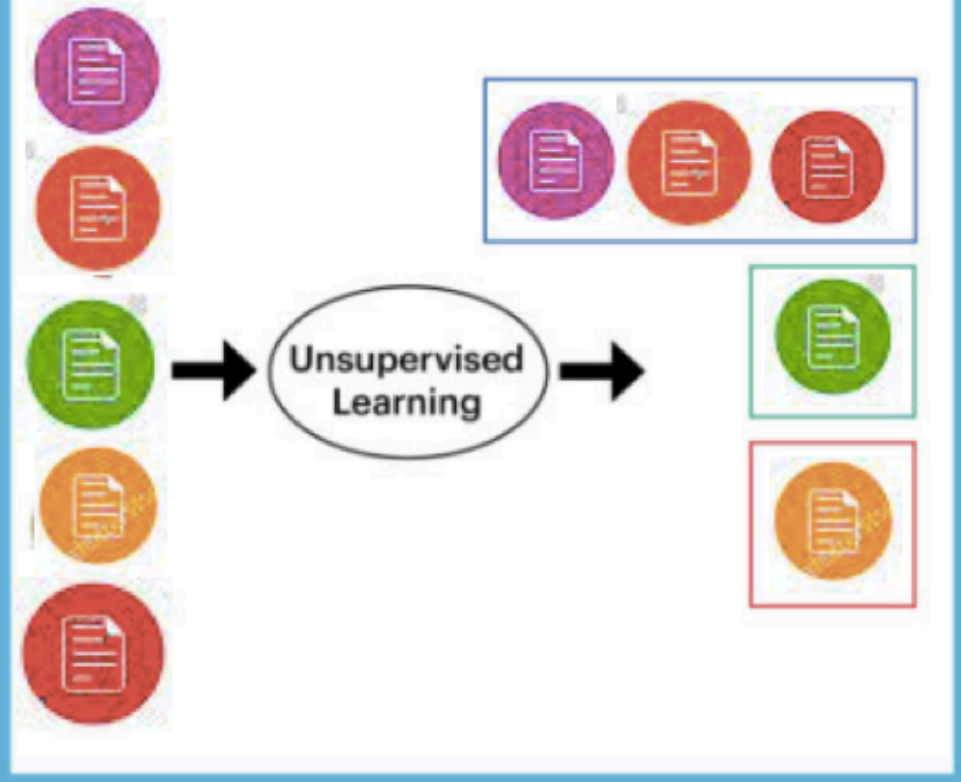
Supervised Learning (Classification Algorithm)



Supervised Learning → Predictive Model

Predictive Model → Correctly classify built environment

Unsupervised Learning (Clustering Algorithm)



Machine Learning

Formal definition:

A program is said to learn from experience *E* with regard to task *T* and performance measure *P*, if its performance on task T improves with experience.

Machine Learning

Formal definition:

A program is said to learn from experience *E* with regard to task *T* and performance measure *P*, if its performance on task T improves with experience.

#	Task	Experience	Performance Measure
1	Understand sentiment from twitter feeds	?	?
2	Analyze wider economic benefits of transport project	?	?
3	Identify documents on similar topics	?	?

Machine Learning

Formal definition:

A program is said to learn from experience *E* with regard to task *T* and performance measure *P*, if its performance on task T improves with experience.

#	Task	Experience	Performance Measure
1	Understand sentiment from twitter feeds	Model trained on words labeled as positive or negative	% tweets correctly identified as positive
2	Analyze wider economic benefits of transport project		
3	Identify documents on similar topics		

Machine Learning

Formal definition:

A program is said to learn from experience *E* with regard to task *T* and performance measure *P*, if its performance on task T improves with experience.

#	Task	Experience	Performance Measure
1	Understand sentiment from twitter feeds	Model trained on words labeled as positive or negative	% tweets correctly identified as positive
2	Analyze wider economic benefits of transport project	Model trained on manually-tagged land use categories	% changes to built environment correctly identified
3	Identify documents on similar topics		

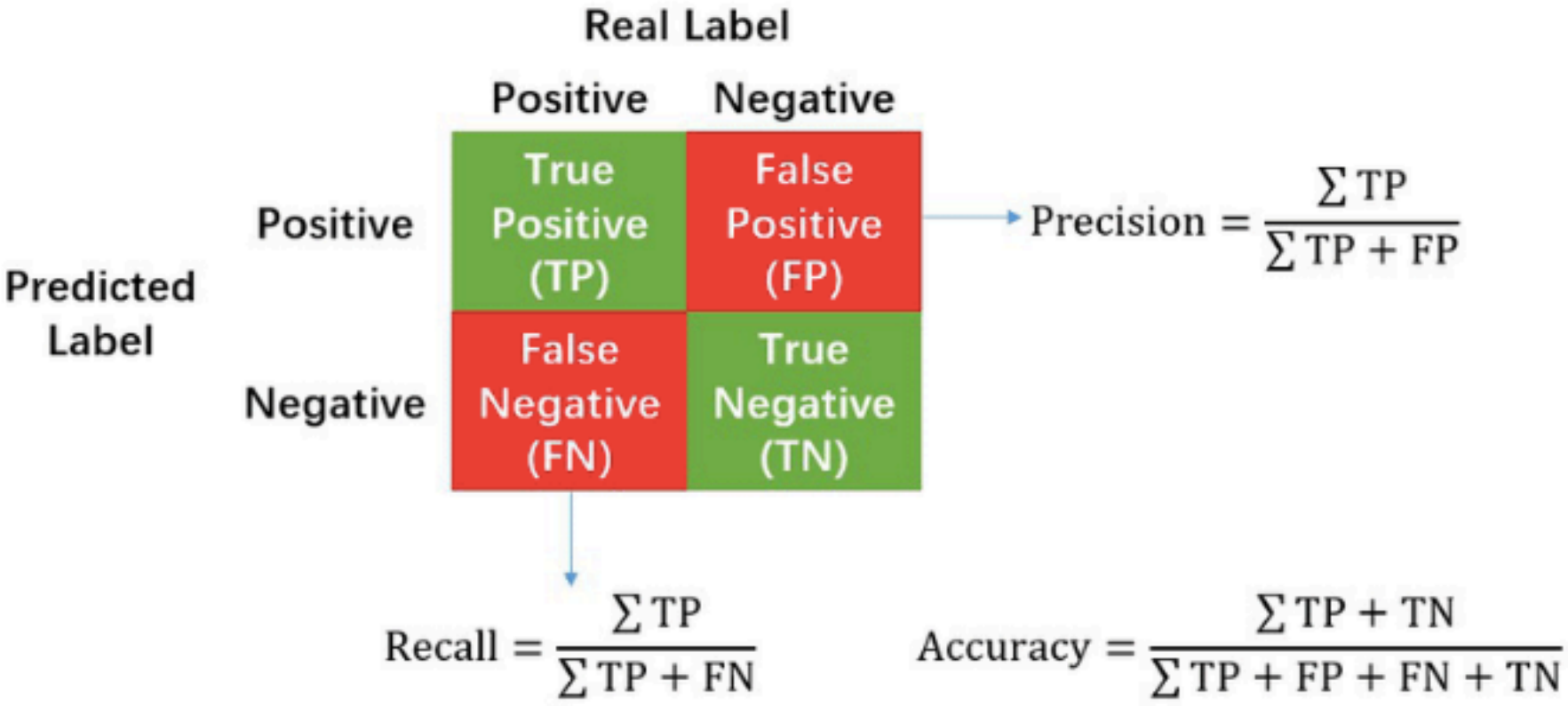
Machine Learning

Formal definition:

A program is said to learn from experience *E* with regard to task *T* and performance measure *P*, if its performance on task T improves with experience.

#	Task	Experience	Performance Measure
1	Understand sentiment from twitter feeds	Model trained on words labeled as positive or negative	% tweets correctly identified as positive
2	Analyze wider economic benefits of transport project	Model trained on manually-tagged land use categories	% changes to built environment correctly identified
3	Identify documents on similar topics	Existing documents clustered based on description or content of document	% new set of documents correctly assigned to relevant clusters

Evaluation Metrics



Source: https://www.researchgate.net/figure/Calculation-of-Precision-Recall-and-Accuracy-in-the-confusion-matrix_fig3_336402347

Model Selection

- Performance vs overfitting

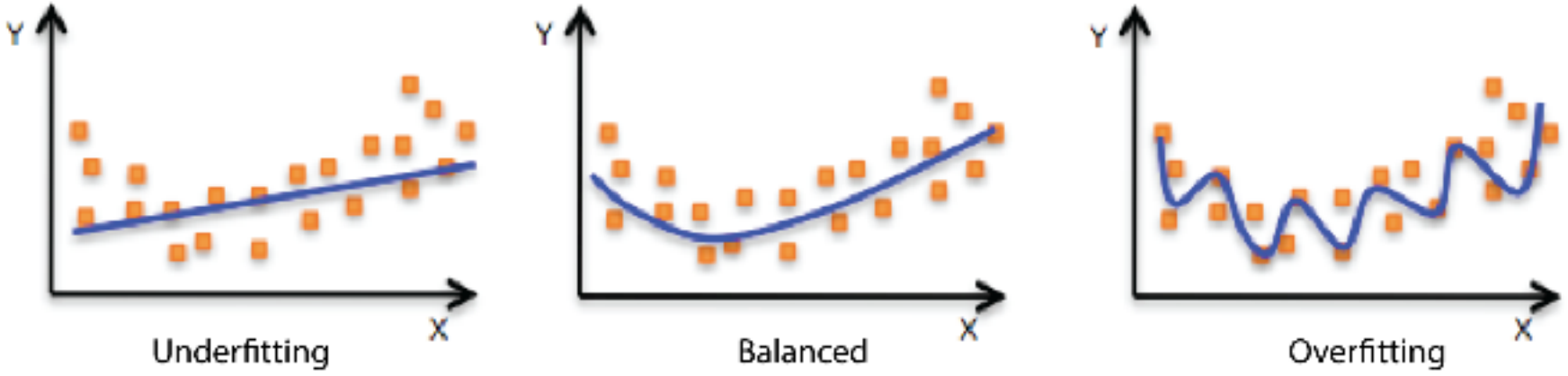
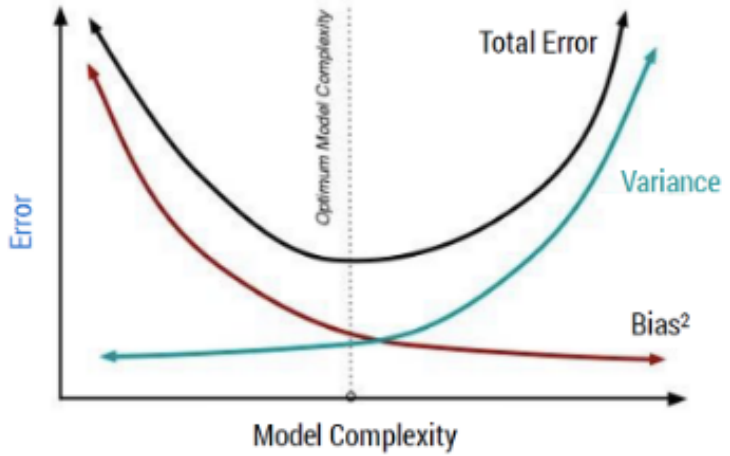


Image Credit: Amazon Machine Learning, Developer Guide

- Bias-variance tradeoff



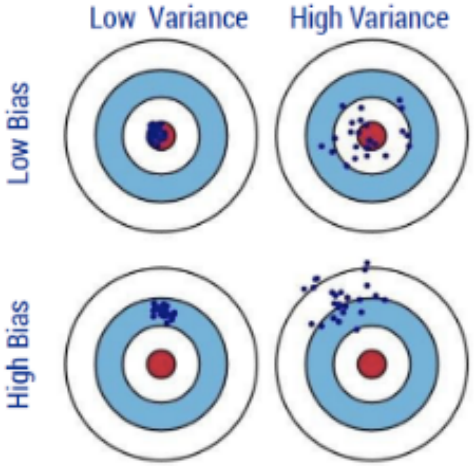


Image Credit: Understanding the Bias-Variance Tradeoff, by Scott Fortmann-Roe via KDnuggets

Considerations

- Evaluation Metrics
- Model complexity (parameters)
- Iterative process / Incremental training
- Training time / speed
- Interpretation

BIG DATA IN ACTION
THEORY OF CHANGE

START WITH A QUESTION

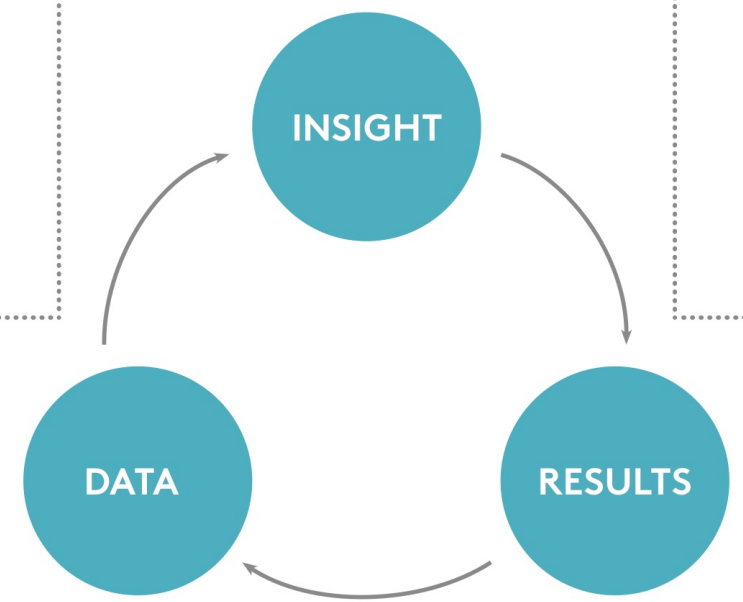
1

DATA INSIGHTS
WHAT INSIGHTS DO WE EXPECT TO FIND?

3

ACTION IMPLEMENTATION
WHAT ACTIONS CAN BE TAKEN, BASED ON INSIGHTS?

4



DATA GENERATION
UNIT, TRIGGER, MEDIA

2

RESULTS
WHAT OUTCOMES OR BEHAVIORS DO WE WANT TO INFLUENCE?

1

STEP # 4

ACTION
IMPLEMENTATION

Action Implementation / Management Aspect I

- Big data and ML => not quick-fix solutions!
- Team members, including TTLs and managers leading a data science projects should have a good understanding of:
  - The Data:
    - representativeness/gaps in training set (bias)
    - assumptions and steps taken in data processing
    - Remember: garbage in; garbage out!
  - The Model:
    - parameters
    - evaluation metrics
    - validation methods

Action Implementation / Management Aspect II

- Ethics and Privacy concerns relating to the project
    - data security/confidentiality/anonymization
    - Transparency: how and why were certain decisions made?
    - bias / discrimination / equity
    - accountability: who will this affect? what is their voice in this?
- Communication/ Visualization (tools: ggplot, matplotlib, seaborn, bokeh, R Shiny etc.)
- Proper documentation of code, data, and products
    - adopt git/GitHub for reusability and smooth collaborative workflows

DATA SCIENCE STARTER KIT

GENERAL PURPOSE	RESOURCES | TOOLS | PRACTICES
Open Data Science	Python | R |Gitub | GitLab | | Jupyter | Anaconda | Kaggle | Cookie Cutter Data Science | Docker | BDAS | BITTS | PySal |
Interoperability	Tidy Data | Slippy | GeoStat | World Pop – GRID3 | NIST | Analysis Ready Data | IPUMS | SDMX| DDI | DublinCore | Schemas.org | ONS-ML
Ethics & Privacy Preserving Methods	UN Handbook of Privacy Preserving Methods | Differential Privacy | UNGP Risk Assessment Tool | Deon DrivenData Checklist
Rapid Data Collection, Tasking, Labeling	MTURK | AWS Sagemaker Ground Truth | Figure8 | Hive Data | Samasource |Qualtrics Prolific | Premise | Native | Snorkle | Kobo | Qfield |RapidPro SurveyCTO| Survey Solutions | Geo-referencing data for ML
Data Catalogs	WB Data Catalog | Google Data Search | Open Street Map | Enigma | AWS Open Data | WorldPop | Kaggle| Awesome Satellite Data

STEP	SATELLITE	MOBILE	TEXT
COLLECTION - High Frequency Data Collection; Data Labelling;	Label Maker | Cumulus | MTURK |Hive Data	Flowkit | UTSDC | OPAL | Positium | FB MTK API	Google API | Twitter decahose stream | FB Marketing API, LinkedInDevAPI | GDELT | Factiva Content API |
DISCOVERY – Training Data Sets; Rich Context Data Search; Knowledge Products | Primitives	Earth on AWS | GeoNet | Nasa Earth Science | Maxar Open Data | Planet Explorer | GEOSTAT | Global Change Master Directory | Radiant ML Hub| Carto Observatory |WSF | ELA| Sentinel-Hub | PopGrid | STAC | COG | ASL SpaceNet | UCI	Cubiq | Mapbox Telemetry | SmartGraph (M-SDK) | Orbital Insights Go | OpenMobile | Mirage | Uber-D | OMF | OpenCellID |	Coleridge Rich Context Scholarly articles API | Kaggle-Text | Awesome-NLP
INTEGRATION & ANALYTICS Analytic tools; privacy preserving methods; integration frameworks	Grid3 | GEE | eo-learn | Solaris | Raster Vision | Pangeo | RoboSat | Sat StatePlay | G-DIF | GeoPandas | Orbital Insights Go | PDAL | Sentinal Toolbox | Orfeo | mosaicJSON | GeoNode | SNAP | RasterFoundry | PPovRepo | Spfeas	Bandicoot | Flowkit | UT-CDR Analysis Kit | Mobile Privacy Model | Movesense | Grid3 | Scikit-Mobility	BERT | GLUE | GPT-2 | Microsoft Turing | Pytorch | StanfordNLP_Stanza | NLTK | | OpenNLP

Credit: Trevor Monroe, WBG

THANK YOU

"I liked it better before big data and metadata when we just had good old regular data."