Data Science Bootcamp

Learning the art of building data-driven products

Bootcamp @ The Fifth Elephant

Amit Kapoor amitkaps.com

Bargava Subramanian bargava.com

Anand Chitpothu anandology.com

What is Data Science?

See the world through a data lens

"Data is just a clue to the end truth"

— Josh Smith

"Science is knowledge which we understand so well that we can teach it to a computer. Everything else is art"

— Donald Knuth

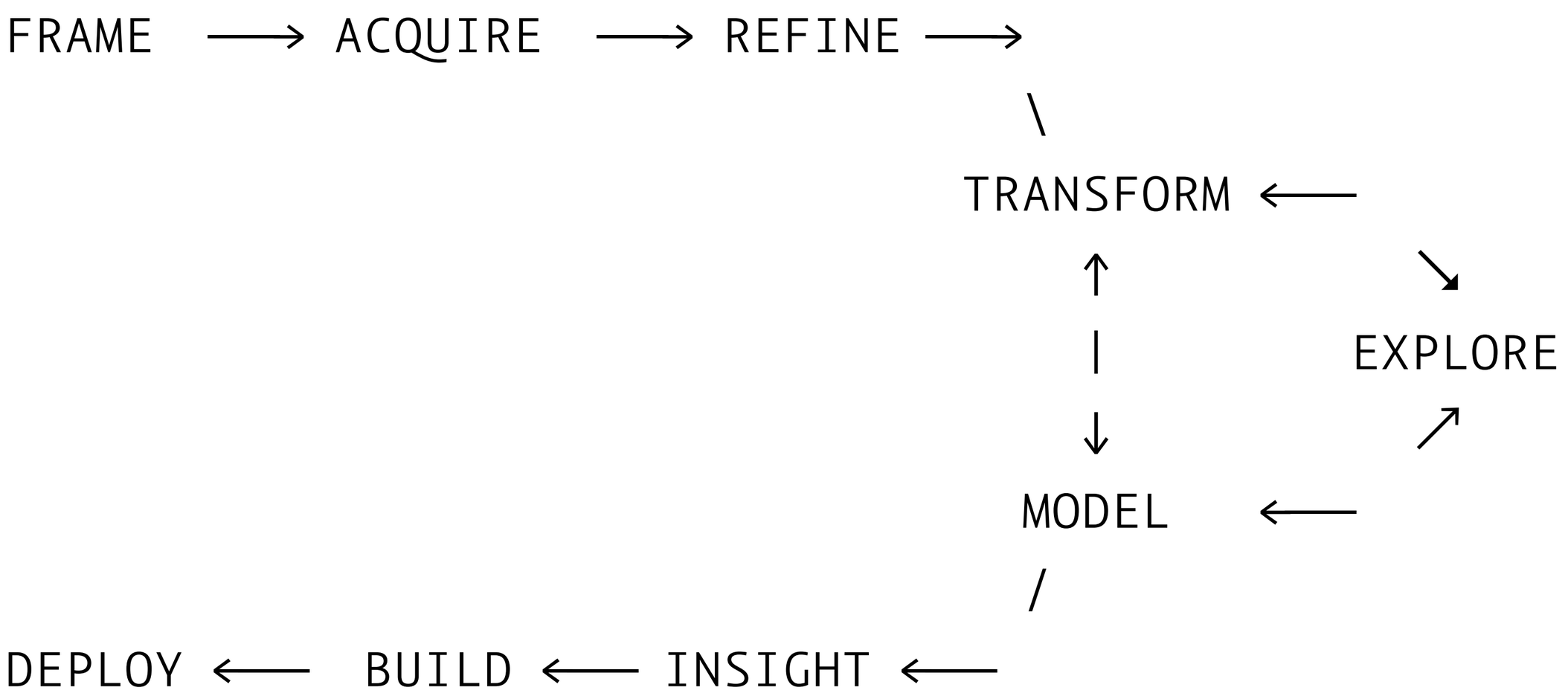
Data Science is an Art

- **Intent**: How can I solve the problem?
- **Process**: What steps should I take?
- **Knowledge**: What building blocks do I need to know?
- **Tools**: Which tools should I use?

Intent for the Bootcamp

- Understand the end-to-end Data Science process
- Learn the Data Science fundamental building blocks
- Build your Data Science Portfolio

Data Science Process



data scientist: the people who are building products from data

What is required to know?

- *Domain Knowledge*
- *Data Management*
- *Modelling & Prototyping*
- *Product Design*
- *Data Engineering*

"Jack of all trades, master of none, though oft times better than master of one."

The Unicorn Skillset

- *Domain Knowledge*: business / social landscape, knowledge
- *Data Management*: data ingestion & wrangling
- *Modelling & Prototyping*: statistics, visualisation, machine learning
- *Product Design*: data narrative, dashboards, applications
- *Data Engineering*: data pipelines, cloud infrastructure

Fundamental Building Blocks

1. The Art of Data Science
2. Data Visualisation
3. HackerMath for Machine Learning
4. Applied Machine Learning
5. Full Stack Data Science

Data Science Portfolio

- Elapsed time to grok concepts
- Practice to reinforce learning
- Deliberate Feedback
- Different problem types for portfolio

Tools

- Understand DS tools landscape
- Learn the **Python** data science stack
- Use **Github** for showcasing your portfolio

Getting Started

- Download the Repo: https://github.com/amitkaps/art-data-science
- Finish installation
- Run jupyter notebook in the console

Outline - Day 1

Session 1: **Introduction and Concepts**

- Overview of Data Science
- Data Science Process
- Jupyter Notebook, Data Structures in Python

Session 2: **Acquire, Refine, Transform**

- Case 1: Peeling the Onion
- Acquire the Data
- Refine the Data
- Transform the Data

Outline - Day 1 (contd.)

Session 3: **Model & Communicate**

- Model the Problem: Descriptive
- Model the Problem: Predictive
- Communicating the Insights

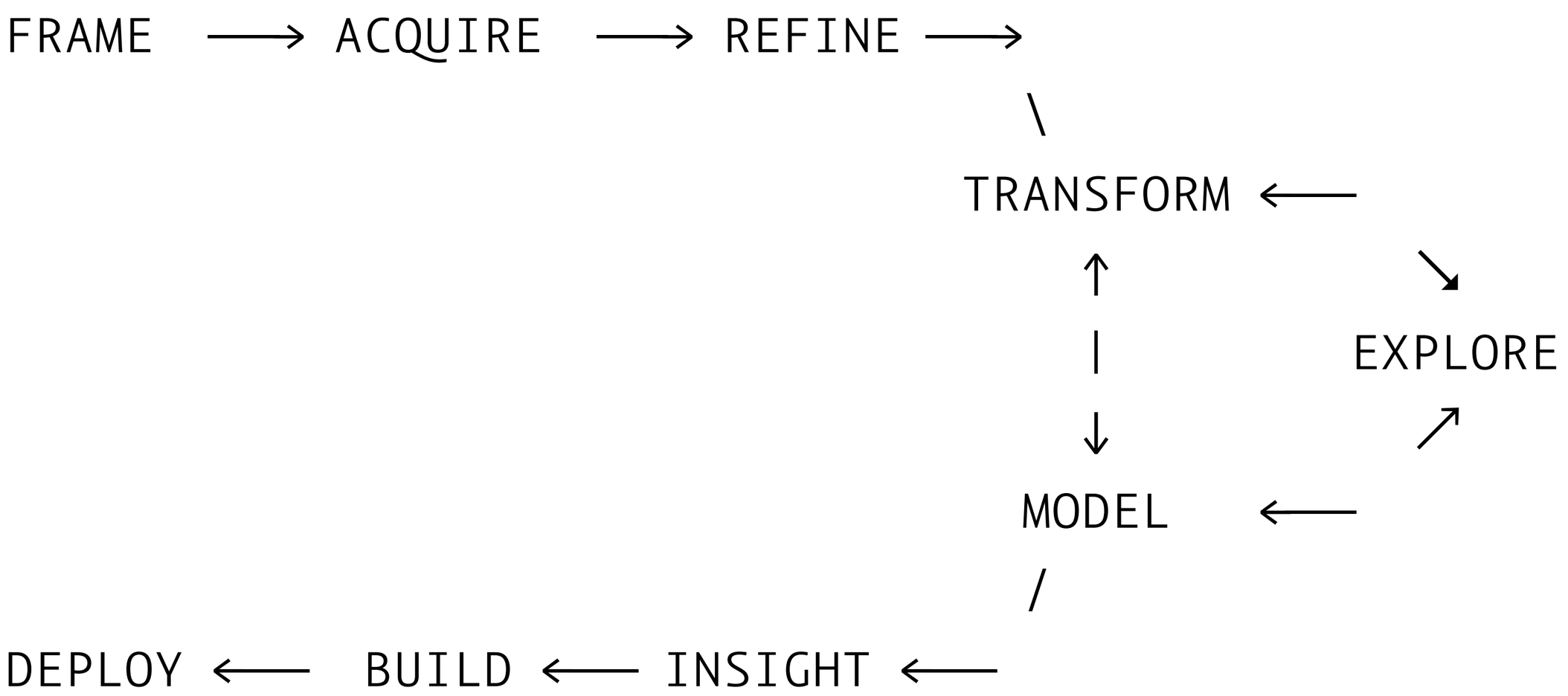
Session 4: **Build & Deploy, Next Steps**

- Communicating the Insights (contd.)
- Build & Deploy
- Next Steps: DS Portfolio
- Wrap-up and Feedback

Schedule

09:30 to 10:00 : *Installation*
10:00 to 11:30 : **Session 1**
11:30 to 11:45 : *Tea break*
11:45 to 13:30 : **Session 2**
13:30 to 14:30 : *Lunch break*
14:30 to 16:00 : **Session 3**
16:00 to 16:15 : *Tea break*
16:15 to 18:00 : **Session 4**
18:00 to 19:00 : *Office Hours (optional)*

Data Science Process



Data Science Process

- *Frame*: Problem definition
- *Acquire*: Data ingestion
- *Refine*: Data wrangling
- *Transform*: Feature creation
- *Explore*: Feature selection
- *Model*: Model creation & selection
- *Insight*: Decision Making
- *Deploy*: Model deployment
- *Build*: Application building

Metaphor

— The price of Onions have been going up and down.

— In 2010, the great Indian onion crisis happened.

— You are planning to adopt a data-driven lens to this problem?

What are the **type of questions** you can ask?

Type of Questions

- What is the average price for Onion across a year in Bangalore?
- How is the price on onion correlated with volume of onion?
- What is the price of onion likely to be next month?
- Does the change in production of onion have an impact on the onion prices?

Type of Questions

- Descriptive
- Inquisitive
- Predictive
- Causal

Data-driven Analytics

- **Descriptive**: Understand Pattern, Trends, Outlier
- **Inquisitive**: Conduct Hypothesis Testing
- **Predictive**: Make a prediction
- **Causal**: Establish a causal link

Frame

"An approximate answer to the right problem is worth a good deal"

Hypothesis Driven Approach

- Toy Problems
- Simple Problems
- Complex Problems
- Business Problems
- Research Problems

Data Types

— What are the types of data on which we are learning?

— Can you give example of say measuring temperature?

Data Types e.g. Temperature

- **Categorical**
  - *Nominal*: Burned, Not Burned
  - *Ordinal*: Hot, Warm, Cold
- **Continuous**
  - *Interval*: 30 °C, 40 °C, 80 °C
  - *Ratio*: 30 K, 40 K, 50 K

Data Types - Operations

- **Categorical**
  - *Nominal*: = , !=
  - *Ordinal*: =, !=, >, <
- **Continuous**
  - *Interval*: =, !=, >, <, -, % of diff
  - *Ratio*: =, !=, >, <, -, +, %

Data Types

- **Categorical**
  - *Nominal*: home owner [rent, own, mortgage]
  - *Ordinal*: credit grade [A > B > C > D > E]
- **Continuous**
  - *Interval*: approval date [20/04/16, 19/11/15]
  - *Ratio*: loan amount [3000, 10000]

Acquire

"80% perspiration, 10% great idea, 10% great output"

- Scraping (structured, unstructured)
- Files (csv, xls, json, xml, pdf, images)
- Database (sqlite, MySql, HDFS)
- APIs
- Streaming

Ways to acquire data

Typical data source

- Download from an internal system
- Obtained from client, or other 3rd party
- Extracted from a web-based API
- Scraped from a website
- Extracted from a PDF file
- Gathered manually and recorded

Refine

"All data is messy."

— **Remove** e.g. remove redundant data

— **Derive** e.g. State and City from the market field

— **Parse** e.g. extract date from year and month column

Other stuff you may need to do to refine are...

— **Missing** e.g. Check for missing or incomplete data

— **Quality** e.g. Check for duplicates, accuracy, unusual

Transform

"A rough diamond is cut and shaped into a beautiful gem"

- **Convert** e.g. free text to coded value
- **Calculate** e.g. percentages, proportion
- **Merge** e.g. first and surname for full name
- **Aggregate** e.g. rollup by year, cluster by area
- **Filter** e.g. exclude based on location
- **Sample** e.g. extract a representative data
- **Summary** e.g. show summary stats like mean

Transform

- **Data Transformations**
- **Encodings** e.g.
  - One Hot Encoding
  - Label Encoding
- **Feature Transformation** e.g.
  - Log Transform
  - Sqrt Transform

Explore

"I don't know, what I don't know."

— Single & Dual Dimension Vis

— Multi Dimensional Vis

— Geographic Vis

— Large Data Vis (Bin - Summarise - Smooth)

— Interactive Vis

Prediction Challenge

It's tough to make predictions, especially about the future.

— Yogi Berra

How to make a Prediction?

- **Human Learning**: Make a *Judgement*
- **Machine Programmed**: Create explicit *Rules*
- **Machine Learning**: Learn from *Data*

Machine Learning (ML)

[Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

— *Arthur Samuel*

Machine learning is the study of computer algorithm that improve automatically through experience

— *Tom Mitchell*

Machine Learning: Essense

- A pattern exists
- It cannot be pinned down mathematically
- Have data on it to learn from

"Use a set of observations (data) to uncover an underlying process"

ML Problems

— “Is this cancer?”

— “What is the market value of this house?”

— “Which of these people are friends?”

— “Will this person like this movie?”

— “Who is this?”

— “What did you say?”

— “How do you fly this thing?”.

ML in use Everyday

- Search
- Photo Tagging
- Spam Filtering
- Recommendation
- ...

Broad ML Application

- Database Mining e.g. Clickstream data, Business data
- Automating e.g. Handwriting, Natural Language Processing, Computer Vision
- Self Customising Program e.g. Recommendations

Model

"All models are wrong, but some are useful"

- Supervised Learning
  - *Continuous: Regression*
  - *Discrete: Classification*
- Unsupervised Learning
  - *Cluster Analysis*
  - *Dimensionality Reduction*
- Reinforcement Learning

Model Creation

Types of ML Model

- Linear
- Tree-Based
- Manifold
- Neural Network

Choosing a Model

1. Interpretability
2. Run-time
3. Model complexity
4. Scalability

ML Terminology

Features: $\mathbf{x}$

- age, income, years, ownership, grade, amount

Target: y

- default

Training Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_n, y_n)$

- historical records

ML Paradigm: Supervised

Given a set of **feature x**, to predict the value of **target** y

Learning Paradigm: **Supervised**

— If y is *continuous* - **Regression**

— If y is *categorical* - **Classification**

Insight, Build & Deploy

"The goal is to turn data into insight"

- Narrative Visualisation
- Dashboard Visualisation
- Decision Making Tools
- Automated Decision Tools

"Doing data analyis requires quite a bit of thinking and we believe that when you've completed a good data analysis, you've spent more time thinking than doing."

— Roger Peng

If you torture the data enough, it will confess.

— Ronald Case

Challenges

- Data Snooping
- Selection Bias
- Survivor Bias
- Omitted Variable Bias
- Black-box model Vs White-Box model
- Adherence to regulations

Data Science Bootcamp

Learning the art of building data-driven products

Bootcamp @ The Fifth Elephant

Amit Kapoor amitkaps.com

Bargava Subramanian bargava.com

Anand Chitpothu anandology.com