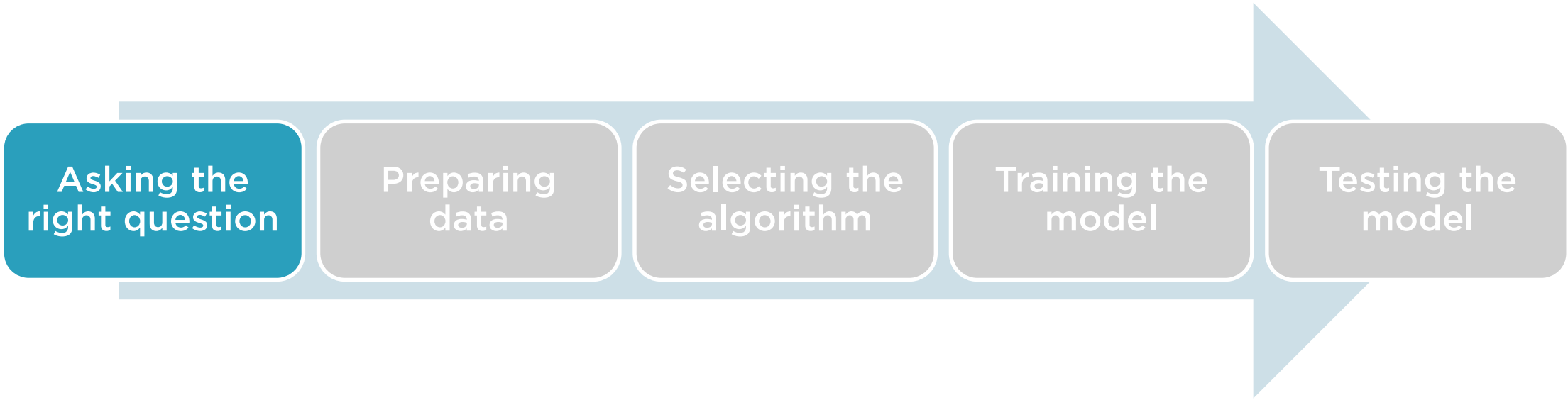
Preparing Your Data

Jerry Kurata
CONSULTANT

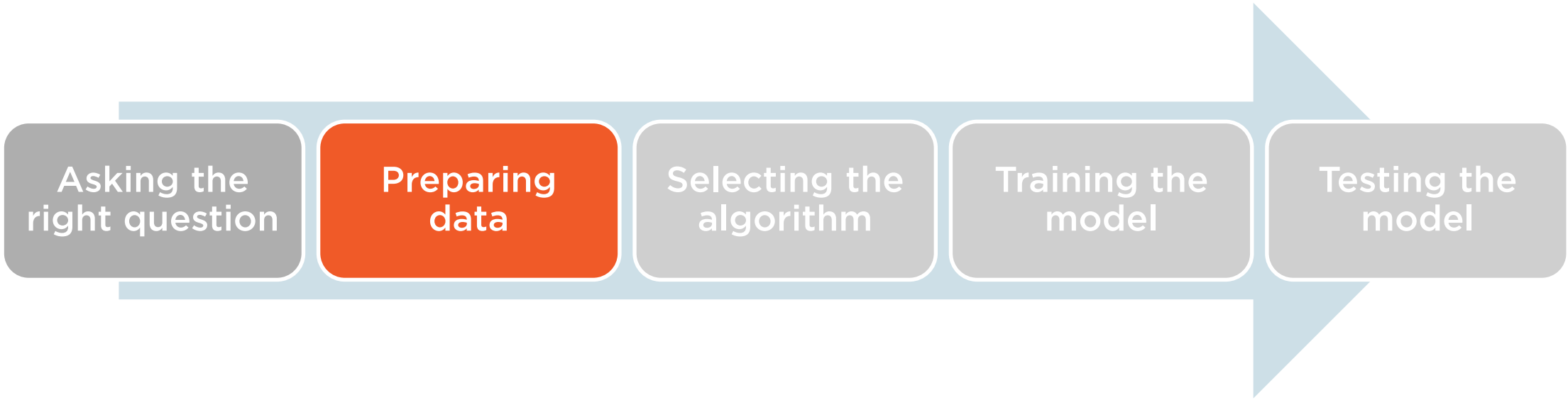
@jerrykur www.insteptech.com

Machine Learning Workflow

Machine Learning Workflow



Machine Learning Workflow



Overview

Find the data we need

Inspect and clean the data

Explore the data

Mold the data to Tidy data

Demos in R and R Studio

Tidy Data

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure:

each variable is a column,

each observation is a row,

each type of observational unit is a table.

Hadley Wickham

50-80% of a ML project
is spent
getting, cleaning, and
organizing data

Getting Data

Google

Government databases

Professional or company data sources

Your company

Your department

All of the above

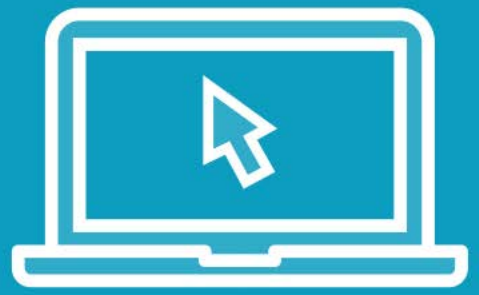
Flight	Origin	Dest	Depart Time	Arrival Time
324	ALT	LAX	1645	1755
232	NYC	ATL	0930	1059
127	LAX	SFO	1920	2100
857	SFO	LAX	2200	2325
776	PHX	ATL	1650	2100

On-time data available

DOT collects on-time data

Data is extractable

Demo



Getting DOT On-time Flight Data

http://bit.ly/DOT_O

nTime

Data Rule #1

Closer the data is to what
you are predicting,
the better

Data Rule #2

Data will never be in the format you need

Columns to Eliminate

Not used

No values

Duplicates

Correlated Columns

Same information in a different format

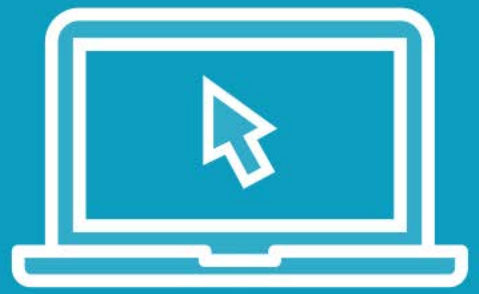
- ID and value associated with ID

Add little information

Can cause algorithms to get confused

- Price = x * Area(sq ft)
  + y * Area(sq m)
  + z * # of rooms

Demo



Loading Data

Exploring Data

Cleaning Data

Molding Data

Dropping rows

Adjusting data types

Creating new columns, if required

Fixing Arr_Del15

Arr_Del15 = 1 if 15 minute delay

Value we are trying to predict

Must be 0 or 1

May contain NA

May contain “”

Remove rows with NA or “”

- Arr_Del15 and Dep_Del15

Data Rule #3

Accurately predicting rare events is difficult

Data Rule #4

Track how you manipulate data

Summary

Reviewed data source

Downloaded data from DOT site

Used R to load CSV file

Cleaned up data

Molded data

Discussed data rules