INTRO TO PROJECT 3

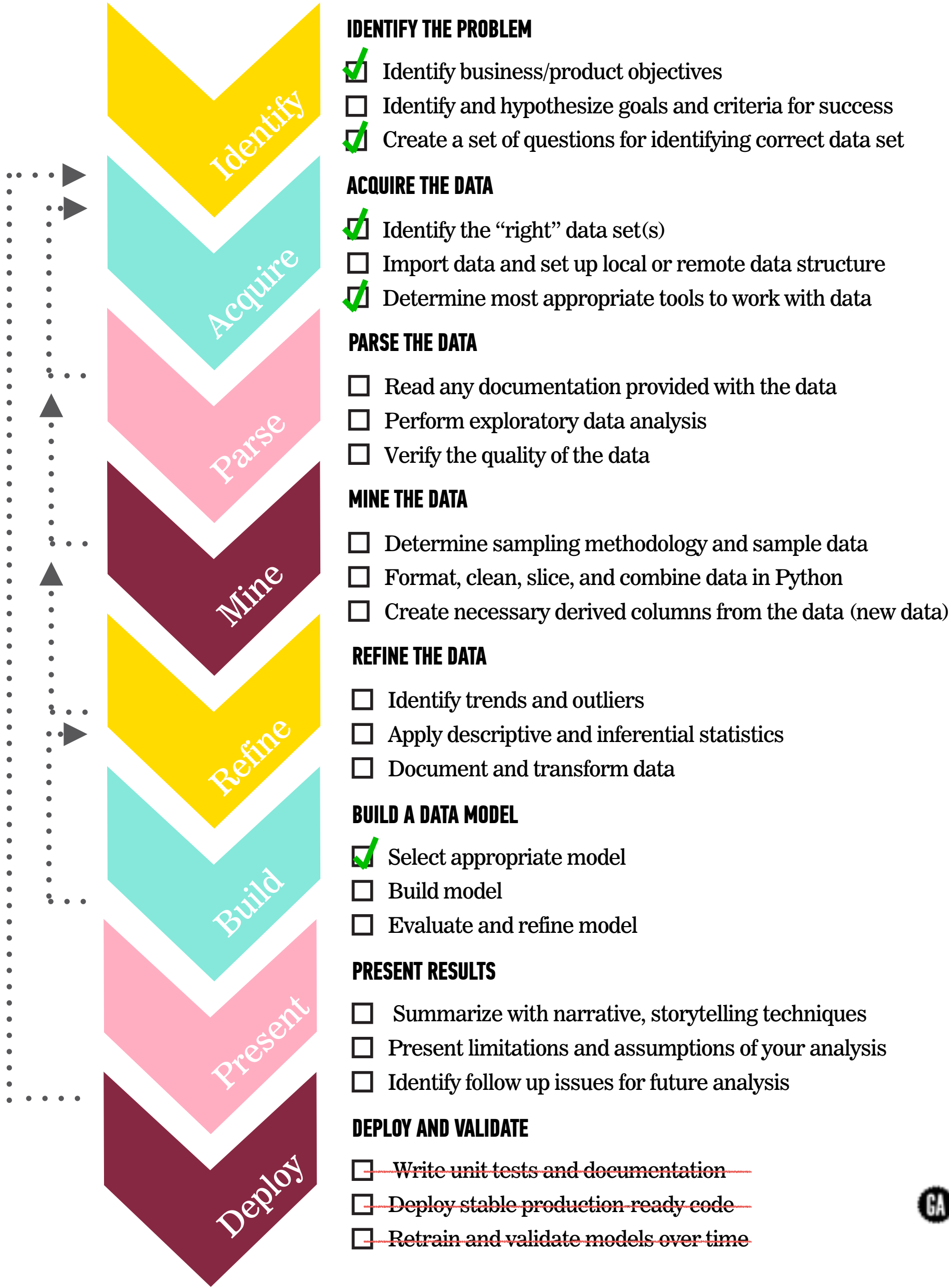## SKILLS YOU'LL DRAW ON

- LINEAR MODELING
- CROSS VALIDATION
- HYPOTHESIS TESTING
- DATA CLEANING AND EXPLORATION IN PANDAS
- PLOTTING
- PRESENTING TO STAKEHOLDERS

DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

PRESENT RESULTS

- ☐ Summarize with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up issues for future analysis

DEPLOY AND VALIDATE

- ☐ Write unit tests and documentation
- ☐ Deploy stable production-ready code
- ☐ Retrain and validate models over time

Identify · Acquire · Parse · Mine · Refine · Build · Present · Deploy

GA

DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- ☑ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☑ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- ☑ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☑ Determine most appropriate tools to work with data

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL

- ☑ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

PRESENT RESULTS

- ☐ Summarize with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up issues for future analysis

DEPLOY AND VALIDATE

- ☐ ~~Write unit tests and documentation~~
- ☐ ~~Deploy stable production ready code~~
- ☐ ~~Retrain and validate models over time~~

Identify · Acquire · Parse · Mine · Refine · Build · Present · Deploy

DATASET

IOWA LIQUOR SALES, 2015-16

	Date	Store Number	City	Zip Code	County Number	County	Category Name	Item Description	Bottle Volume (ml)	State Bottle Cost	State Bottle Retail	Bottles Sold	Sale (Dollars]
0	11/04/2015	3717	SUMNER	50674	9.0	Bremer	APRICOT BRANDIES	Mr. Boston Apricot Brandy	750	$4.50	$6.75	12	$81.00
1	03/02/2016	2614	DAVENPORT	52807	82.0	Scott	BLENDED WHISKIES	Tin Cup	750	$13.75	$20.63	2	$41.26
2	02/11/2016	2106	CEDAR FALLS	50613	7.0	Black Hawk	STRAIGHT BOURBON WHISKIES	Jim Beam	1000	$12.59	$18.89	24	$453.36
3	02/03/2016	2501	AMES	50010	85.0	Story	AMERICAN COCKTAILS	1800 Ultimate Margarita	1750	$9.50	$14.25	6	$85.50
4	08/18/2015	3654	BELMOND	50421	99.0	Wright	VODKA 80 PROOF	Five O'clock Vodka	1750	$7.20	$10.80	12	$129.60

DATASET

IOWA LIQUOR SALES, 2015-16

	Date	Store Number	City	Zip Code	County Number	County	Category Name	Item Description	Bottle Volume (ml)	State Bottle Cost	State Bottle Retail	Bottles Sold	Sale (Dollars)
0	11/04/2015	3717	SUMNER	50674	9.0	Bremer	APRICOT BRANDIES	Mr. Boston Apricot Brandy	750	$4.50	$6.75	12	$81.00
1	03/02/2016	2614	DAVENPORT	52807	82.0	Scott	BLENDED WHISKIES	Tin Cup	750	$13.75	$20.63	2	$41.26
2	02/11/2016	2106	CEDAR FALLS	50613	7.0	Black Hawk	STRAIGHT BOURBON WHISKIES	Jim Beam	1000	$12.59	$18.89	24	$453.36
3	02/03/2016	2501	AMES	50010	85.0	Story	AMERICAN COCKTAILS	1800 Ultimate Margarita	1750	$9.50	$14.25	6	$85.50
4	08/18/2015	3654	BELMOND	50421	99.0	Wright	VODKA 80 PROOF	Five O'clock Vodka	1750	$7.20	$10.80	12	$129.60

TWO PROBLEM STATEMENTS

SCENARIO 1:
STATE TAX BOARD

You are a data scientist in residence at the Iowa State tax board. The Iowa State legislature is considering changes in the liquor tax rates and wants a report of current liquor sales by county and projections for the rest of the year.

SCENARIO 2:
MARKET RESEARCH

A liquor store owner in Iowa is looking to expand to new locations and has hired you to investigate the market data for potential new locations. The business owner is interested in the details of the best model you can fit to the data so that his team can evaluate potential locations for a new storefront.

PRESENTING YOUR ANALYSIS

ON FRIDAY AFTERNOON YOU WILL PRESENT A MINIMUM VIABLE PRODUCT

‣ 5 minute presentations

‣ Your basic findings for 1 of the 2 scenarios

‣ Include your problem statement, model, and supporting statistics

ON FRIDAY AFTERNOON YOU WILL PRESENT A MINIMUM VIABLE PRODUCT

- 5 minute presentations
- Your basic findings for 1 of the 2 scenarios
- Include your problem statement, model, and supporting statistics

BOTH SCENARIOS DUE ON GITHUB NEXT THURSDAY, 6/30