Introduction to the course (Machine Learning)

Nazgul Rakhimzhanova

IITU 2021

OUTLINE

- **Course** Rules
- **Syllabus** and Evaluation criteria
- **Intro** to DS and ML
- **Tools** for DS
- **Readings**

COURSE RULES

- 1ST: Enjoy
- 2nd: Read -> Think -> Implement
- 3rd: Ask questions…But **before, think** once again
- 4th: Be **responsible** for your **actions**
- 5th: Cheating == Instant karma

COURSE RULES

- **Your** attendance == your grades
- **Listen on Lectures** – implement on Labs ask questions on Practice Classes
- **No re-explanations** at Labs

COURSE RULES

- **You** should/have/must
  read more than we ask
  code more than we ask
  analyze more than we ask
  comment more than we ask

COURSE CALENDAR

week	Mid Term (weeks 01-07)	End Term (weeks 08-14)	week
01	Intro: Data Science Area and open source tools for Data Science	Statistics: Distribution – Lognormal, Exponential	08
02	NumPy package for data science	Sampling and Estimation	09
03	Pandas package for data science	Correlation and Covariance	10
04	Visualization with matplotlib	Hypothesis testing	11
05	Statistics: Distribution – Normal	Decision Tree	12
06	Exploratory Data Analysis (EDA)	Linear Regression	13
07	**Summary for 6 weeks QA session**	**Summary for 6 weeks QA session**	**14**
15	**Course summary**		

SYLLABUS: Resume

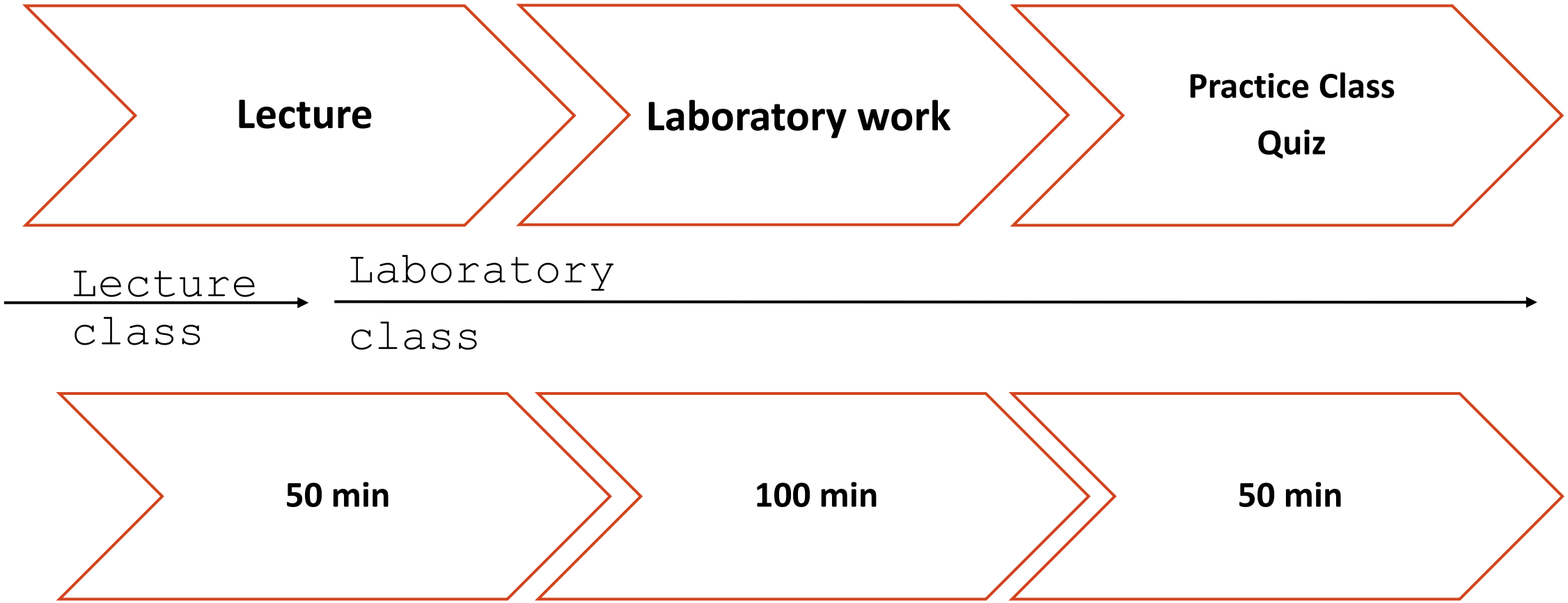
- **Mid Term Week01-Week07:**
  - 5 graded lab works 01-05 (10 points each)
  - 5 graded Quizzes 01-05 (4 points each)
  - Mid Term task (30 points)

  - **Note!** Laboratory work 00 and Quiz 00 are ungraded.

SYLLABUS: Resume

- **End Term Week08 – Week14:**
  - 5 graded lab works 06-10 (10 points each)
  - 5 graded Quizzes 06-10 (4 points each)
  - End Term task (30 points)

WEEK STRUCTURE



- What **questions** do you have any about your Evaluation criteria's?

INTRO TO
Data Science
and
Machine Learning

INTRO TO DS and ML

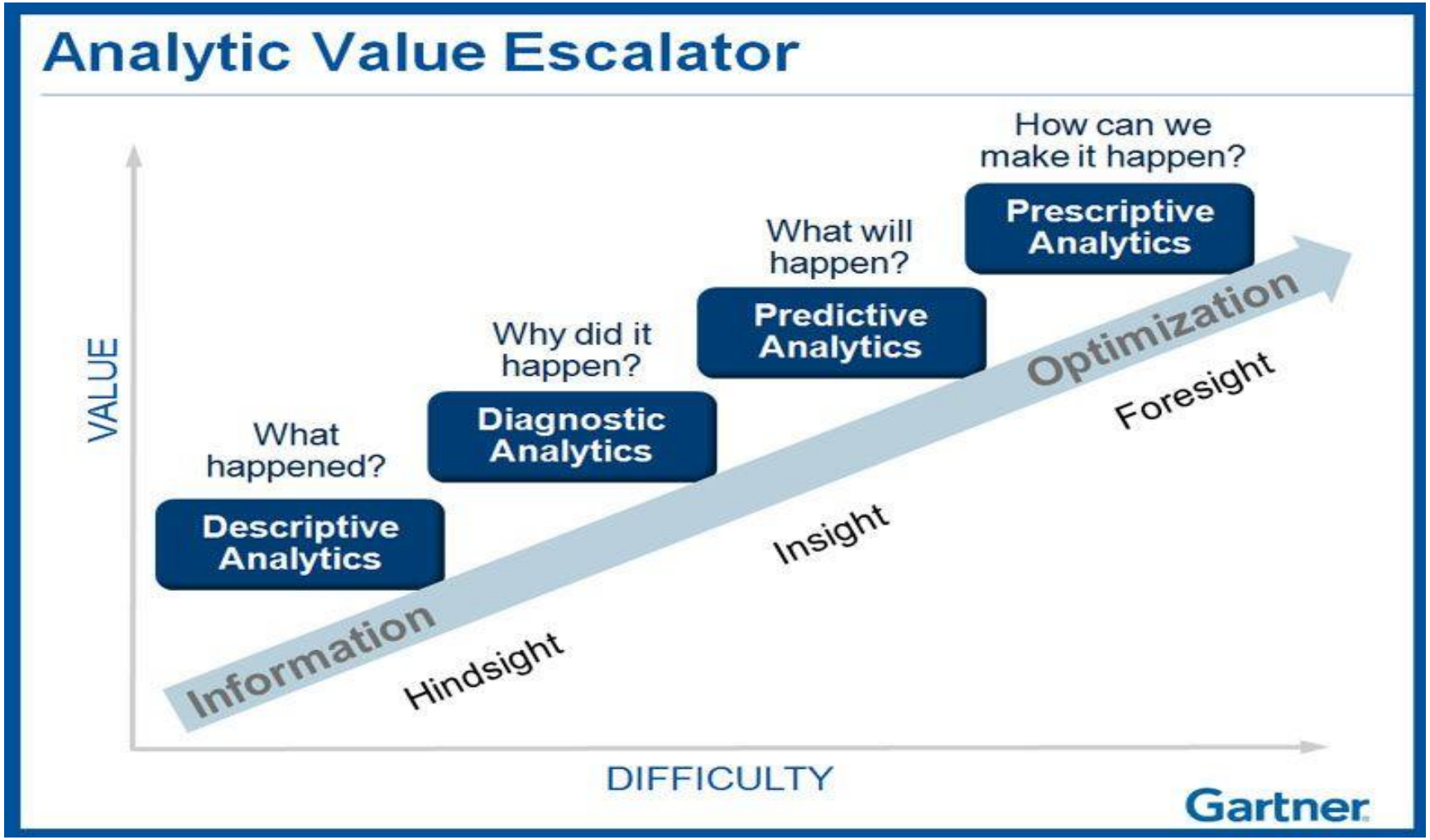
- **Data science** - Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.
- **Data science** is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems“

https://en.wikipedia.org/wiki/Data_science

INTRO TO DS and ML

- **Data Science is an art.**

INTRO TO DS and ML



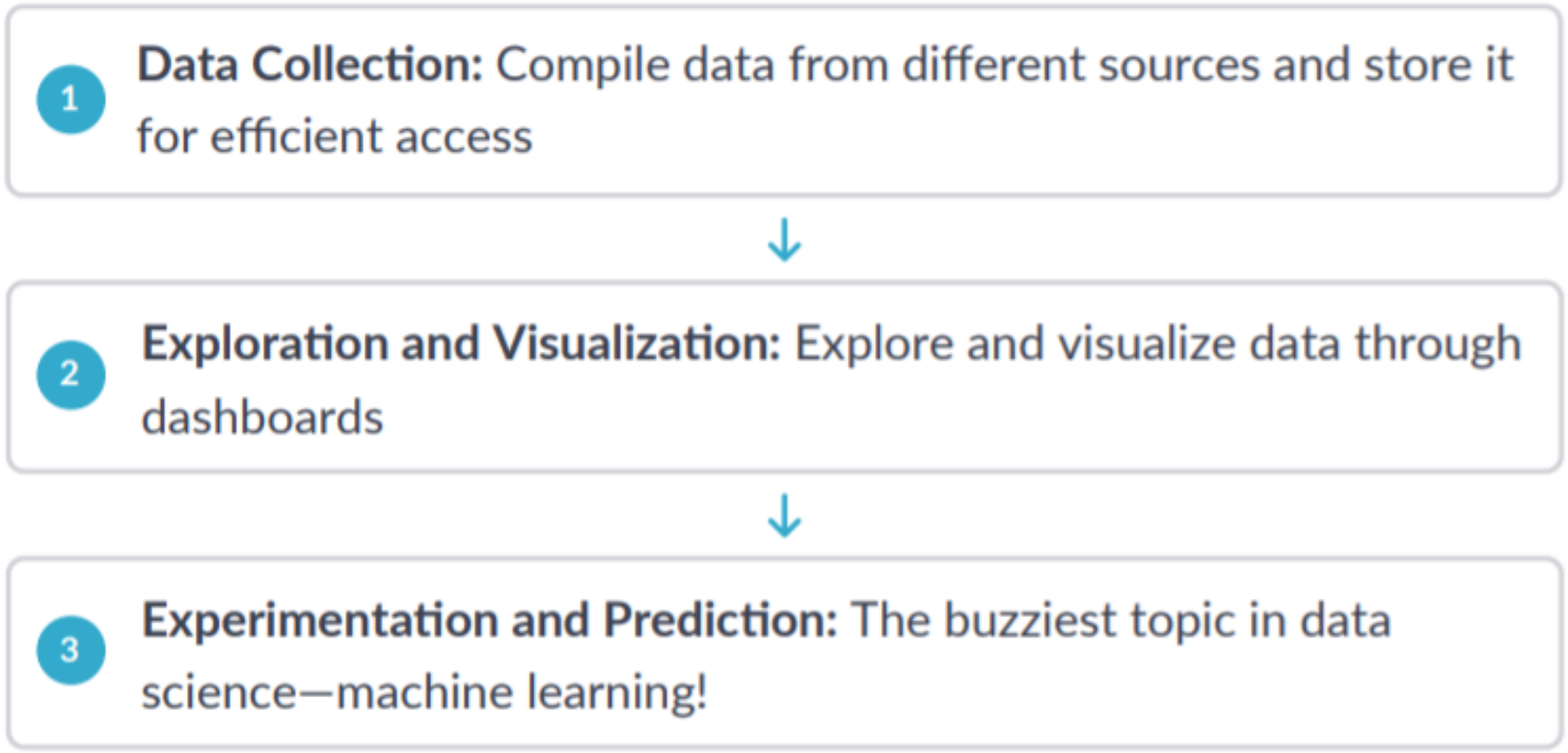
INTRO TO DS AND ML

Types of Data Science:

- **Descriptive Analytics** (Business Intelligence): Get useful data in front of the right people in the form of dashboards, reports, and emails
- **Predictive Analytics** (Machine Learning): Put data science models continuously into production
- **Prescriptive Analytics** (Decision Science): Use data to help a company make decisions

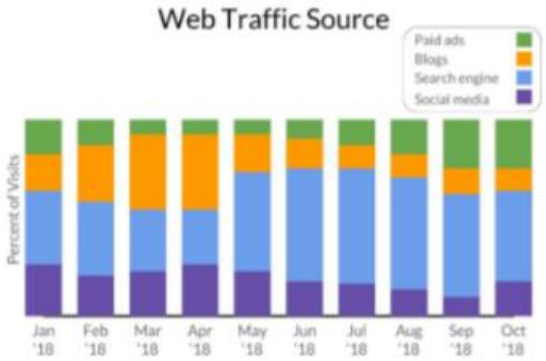
INTRO TO DS AND ML

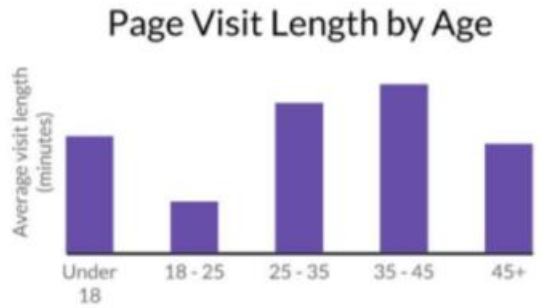
- The standard Data Science Workflow:



INTRO TO DS AND ML

- Exploration and Visualization





Popular Dashboard Tools

Spreadsheets	BI Tools	Customized Tools
Excel	Power BI	R Shiny
Sheets	Tableau	d3.js
	Looker	

INTRO TO DS AND ML

- Experimentation and Prediction

	Supervised Machine Learning	Unsupervised Machine Learning
Purpose	Makes predictions from data with labels and features	Makes predictions by clustering data with no labels into categories
Example	Recommendation systems, email subject optimization, churn prediction	Image segmentation, customer segmentation

INTRO TO DS and ML

Data Engineer
Store and maintain data
SQL/Java/Scala/Python

Data Analyst
Visualize and describe data
SQL + BI tools

Data Scientist
Build custom models to drive business decisions
Python/SQL/R

Machine Learning Engineer
Write production-level code to predict with data
Python/Java/R

Problems DA can solve

- **Prediction**: traffic, flood, disease, earthquakes, election outcome etc.
- **Detection(classification)**: fraud, illegal immigrants, suspicious individuals etc.

Problems DA can solve

- Give me some examples for prediction tasks and classification tasks.

TOOLS for DS

- Python / Java / R / C++ / Scala
- SQL / NoSQL
- IDEs – notebooks: Jupyter, Polynote and, Google Collab, Kaggle etc.
- BI tools (Power BI, Tableau)

TOOLS for this course

- Jupyter Notebook https://jupyter.org/about

or

- https://colab.research.google.com/
- https://www.kaggle.com/notebooks/welcome

Install on your personal laptops

- Python 3.8 (you can encounter some issues with 3.8 python)
- Jupyter Lab, Jupyter Notebook (or Conda)
- pip

What will you learn?

- **Jupyter** notebook + Python for DS
- **Basics of** statistics for machine learning
- **Visualization** tools
- **Basics** of classic machine learning

Readings

- https://jupyter.org/documentation
- **Practical Statistics** for Data Science, book by Peter Bruce, Andy Bruce, O'Reilly
- **Python Data** Science Handbook: Essential Tools for Working with Data, Book by Jacob T. VanderPlas
- **Data Science from Scratch**, Book by Joel Grus

Additional resources

- Khan academy