INSEAD

The Business School for the World®

Prof. Anton Ovchinnikov

Prof. Spyros Zoumpoulis

DSB Classes 11-12, February 13, 2018

• The Data Science Process

Structure of the course

INSEAD
The Business School for the World®

- SESSIONS 1-2 (AO): Data analytics process; from Excel to R
  - Tutorial 1: Getting comfortable with R
- SESSIONS 3-4 (AO): Time Series Models
- SESSIONS 5-6 (AO): Intro to classification, logistic regression and machine learning
  - Tutorial 2: Midterm R help / classification
- SESSIONS 7-8 (SZ): Advanced Classification; From .R to Notebooks; Dimensionality reduction
- SESSIONS 9-10 (SZ): Dimensionality Reduction; Clustering and Segmentation
  - Tutorial 3: Q&A on R for three main modules
- **SESSIONS 11-12 (SZ): The Data Science Process; Guest speaker**
  - Tutorials 4, 5: Hands-on help on projects
- SESSIONS 13-14 (AO+SZ): Project presentations

Plan for the day
Learning objectives

INSEAD
The Business School for the World®

- The process of a data science project
  - Guest speaker: Elias Baltassis, Director Europe, Data & Analytics, The Boston Consulting Group
  - In-class assignment: reflect on the analytics process
  - Share, get feedback, extract learnings about the data science process

Elias Baltassis

INSEAD
The Business School for the World®

- Director Europe, Data & Analytics, The Boston Consulting Group
- Industry expertise: financial services and insurance
- Before: Partner with Opera Solutions, partner with Bain
- INSEAD alum

In-class assignment

INSEAD
The Business School for the World®

- Assignment: You are leading a team of data scientists and consultants in an organization.
  1. What three data analytics projects do you kick-start? With what criteria would you hierarchize them?
  2. Pick one idea of the three. How would you implement it? Discuss the process.

- Break-out Rooms: 315-325

(A) Process for Data Science Projects

INSEAD
The Business School for the World®

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

This is an ITERATIVE process!

Step 1. Business understanding

INSEAD
The Business School for the World®

- Describe in detail the problem you want to solve
  - Define the issue with metrics
- Specify expected benefits in business terms
  - Is there a current practice in place? What is a benchmark performance?
- Identify key individuals in the organization
  - Who manages the issue now and how?
  - Who needs to be involved to activate the solution?

Step 2. Data understanding

INSEAD
The Business School for the World®

- What data is available?
  - Are there any relevant external sources?
  - Which variables should be used?
- How much history is required?
  - Have there been some major changes in the business/industry recently?
- What is the right level of aggregation/granularity?
  - Individual, household, or zip code level? Daily or weekly?
- What data quality issues do we have?
  - Do missing values indicate something?
  - How do we handle non-numeric data?
- Simple hypotheses generation
  - How do we expect specific variables to affect the solution?

Step 3. Data preparation

INSEAD
The Business School for the World®

- Merge all data relevant sources
  - Ensure time or any other alignment
- Deal with data quality issues
  - Handle non-numeric issues
  - Handle missing values
  - Handle data errors
  - Understand outliers
- Feature engineering
  - Derive new (simple) features
- Split data in training, validation, and testing
  - How will the solution be used in practice? Can we simulate it?

Step 4. Modeling

INSEAD
The Business School for the World®

- Start with simple analyses
  - Descriptive statistics and visualization
- Identify sub-problems fitting with analytic tools
  - Can we group variables that are highly correlated? (Factor analysis)
  - Do we need to develop different solutions for different segments? (Clustering)
  - Do we predict binary outcomes? (Classification)
- Estimate and assess model parameters
  - Are they statistically valid?
  - Do they make sense?

Step 5. Evaluation

INSEAD
The Business School for the World®

- Measure various performance metrics
  - Classification: do false positives or false negatives matter most? ROC curve, lift curve, profit curve
- Rank the candidate models
- Is there overfitting?
- Are the results easy to explain?
  - Highlight particularly novel or unique findings
- Do the analyses, our judgment, and our business criteria all agree?

Step 6. Deployment

INSEAD
The Business School for the World®

- Who needs to be involved in deployment?
  - Change management
- What is the data pipeline?
  - How are data sources and IT integrated?
  - How are data failures handled?
- How to test the solution before full deployment?
  - A/B testing setup
- How do we know our solution/model expired?
  - What metrics do we monitor?

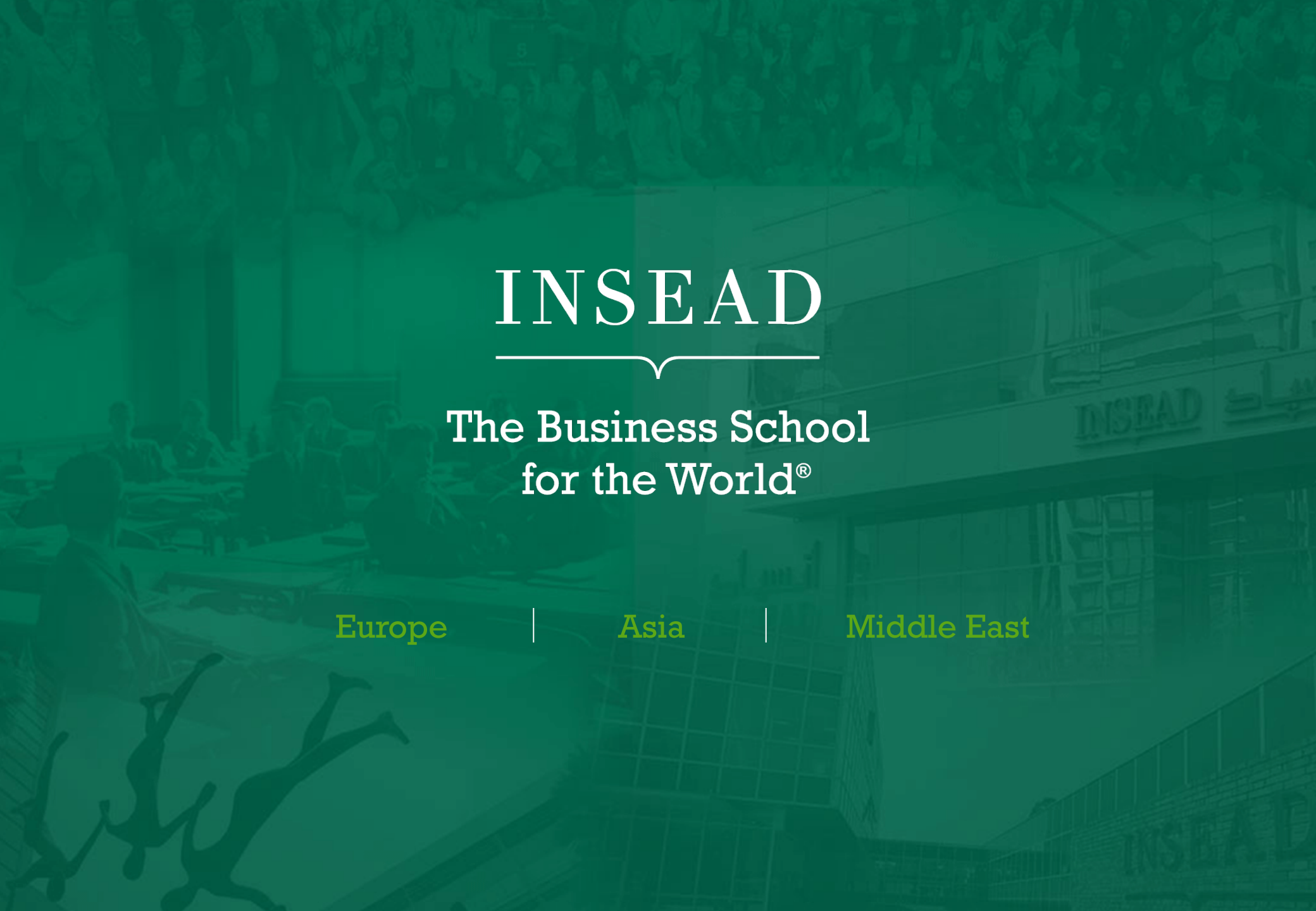
Next…

INSEAD
The Business School for the World®

- Proposal for final project (due Feb 14)

- Tutorials 4 & 5 [Wed & Thu Feb 14 & 15, 7.15 pm]
  - Hands-on help with final projects ideas and implementation
  - Tutorial 4: dplyr, ggplot packages

- Sessions 13-14 [Tue Feb 20, Amphi 105].
  - Final project (due Feb 20)
  - Prepare to present


INSEAD
The Business School
for the World®
Europe | Asia | Middle East