

Advanced Statistics for Data Science

Spring 2022

Lecture 1: Introduction, Course Overview, Exploratory Data Analysis

Dr. Alon Kipnis

March 1st 2022

Outline of first lecture

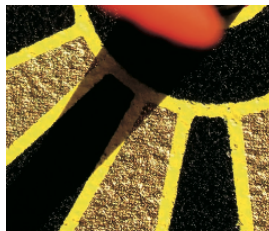
1. Overview
2. Course outline and organizational matters
3. Break
4. Notebook: Examples
5. Introduction to Linear Regression
6. Notebook: Exploratory Data Analysis

**Why should you take this
course?**

- The **Information Age** –
 - Data availability – communication, storage, sensing devices
 - Data analysis – computing power, **algorithms**

- The **Information Age** –
 - Data availability – communication, storage, sensing devices
 - Data analysis – computing power, **algorithms**
- The **Data Age** –
 - More data-driven business, healthcare, government **decisions** based on massive and ever-increasing datasets

- The **Information Age** –
 - Data availability – communication, storage, sensing devices
 - Data analysis – computing power, **algorithms**
- The **Data Age** –
 - More data-driven business, healthcare, government **decisions** based on massive and ever-increasing datasets
 - **Successful Applications:**
 - Web search engine
 - Voice recognition systems
 - Targeted advertising
 - Recommendation systems
 - Challenges are at the intersections of **hardware**, **software**, and **statistics**



The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner’s article “The Unreasonable Effectiveness of Mathematics in the Natural Sciences”¹ examines why so much of physics can be neatly explained with simple mathematical formulas

such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary par-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The

Example – Predicting Housing Prices

BsmtSF	Heating	HeatingQC	CentralAir	Electrical	1stFlrSF	2ndFlrSF	LowGrLivArea	BsmtL	BsmtHt	FullBath	HalfBath	Bedroom	Kitchen	Kitchen	SalePrice	
856	GasA	Ex	Y	SBrkr	856	854	0	1710	1	0	2	1	3	1	Gd	208500
1262	GasA	Ex	Y	SBrkr	1262	0	0	1262	0	1	2	0	3	1	TA	181500
920	GasA	Ex	Y	SBrkr	920	866	0	1786	1	0	2	1	3	1	Gd	223500
756	GasA	Gd	Y	SBrkr	961	756	0	1717	1	0	1	0	3	1	Gd	140000
1145	GasA	Ex	Y	SBrkr	1145	1053	0	2198	1	0	2	1	4	1	Gd	250000
796	GasA	Ex	Y	SBrkr	796	566	0	1362	1	0	1	1	1	1	TA	143000
1686	GasA	Ex	Y	SBrkr	1694	0	0	1694	1	0	2	0	3	1	Gd	307000
1107	GasA	Ex	Y	SBrkr	1107	983	0	2090	1	0	2	1	3	1	TA	200000
952	GasA	Gd	Y	FuseF	1022	752	0	1774	0	0	2	0	2	2	TA	129900
991	GasA	Ex	Y	SBrkr	1077	0	0	1077	1	0	1	0	2	2	TA	118000
1040	GasA	Ex	Y	SBrkr	1040	0	0	1040	1	0	1	0	3	1	TA	129500
1175	GasA	Ex	Y	SBrkr	1182	1142	0	2324	1	0	3	0	4	1	Ex	345000

- $x = (\text{sqm}, \#Bd, \#windows, \dots, \text{CrimeRate})$
- $y = \text{SalePrice}$

THE WALL STREET JOURNAL.

MARKETS

The Future of Housing Rises in Phoenix

High-tech flippers such as Zillow are using algorithms to reshape the housing market

By [Ryan Dezember](#) and [Peter Rudegeair](#) / Photographs by Benjamin Hoste
for *The Wall Street Journal*

June 19, 2019 11:10 am ET

The Data Age: Fail I

THE WALL STREET JOURNAL.

MARKETS

The Future of Housing Rises in Phoenix

High-tech flippers such as Zillow are using algorithms to reshape the housing market

By [Ryan Dezembe](#)
for The Wall Street Journal
June 19, 2019 11:10 am EDT

The New York Times

Daily Business Briefing >

Zillow, facing big losses, quits flipping houses and will lay off a quarter of its staff.

The real estate website had been relying on its algorithm that estimates home values to buy and resell homes. That part of its business lost about \$420 million in three months.

Zillow is sitting on thousands of houses worth less than what the company paid for them. Caitlin O'Hara for The New York Times

By **Stephen Gandel**

Nov. 2, 2021

The Data Age: Fail II

The New York Times

Science

WORLD

U.S.

N.Y. / REGION

BUSINESS

TECHNOLOGY

SCIENCE

HEALTH


SPORTS

OPINION

ENVIRONMENT

SPACE & COSMOS

Computer Wins on 'Jeopardy!': Trivial, It's Not



Carol Kaelson/Jeopardy Productions Inc., via Associated Press

Two "Jeopardy!" champions, Ken Jennings, left, and Brad Rutter, competed against a computer named Watson, which proved adept at buzzing in quickly.

By JOHN MARKOFF

Published: February 16, 2011

The Data Age: Fail II

The New York Times

Science

WORLD	U.S.	N.Y. / REGION	BUSINESS	TECHNOLOGY	SCIENCE	HEALTH	SPORTS	OPINION
					ENVIRONMENT	SPACE & COSMOS		


Computer Wins on 'Jeopardy!': Trivial, It's Not


Forbes

YORKTOWN HEIGHTS, NY - JANUARY 13: A general ...

Feb 8, 2013, 02:22pm EST

IBM's Watson Gets Its First Piece Of Business In Healthcare

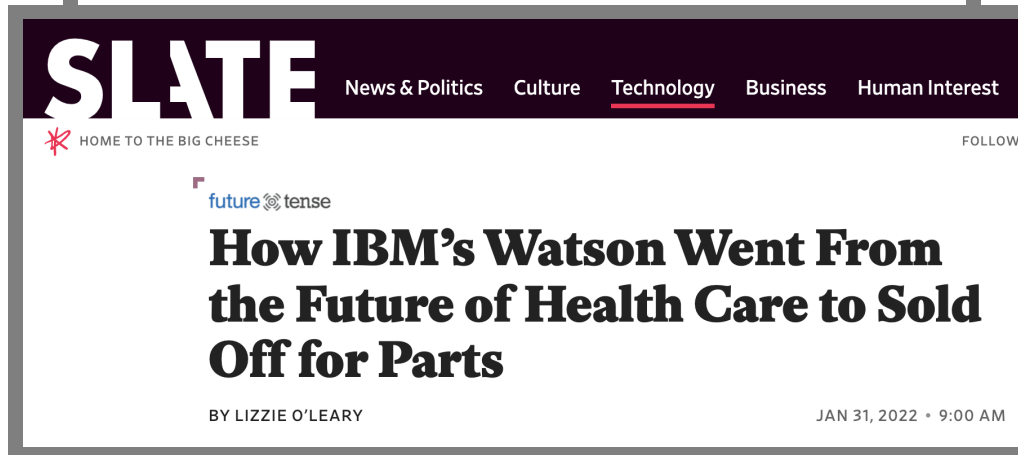


Bruce Upbin Former Contributor 
Tech
I manage our technology coverage.

Follow

clinical research it can get its hard drives on. Today Watson has analyzed 605,000 pieces of medical evidence, 2 million pages of text, 25,000 training cases and had the assist of 14,700 clinician hours fine-tuning its decision accuracy. Six "instances" of Watson have already been installed in the last 12 months.

The Data Age: Fail II



This Class...

- ...is about making decisions based on data using **models** (see next slide)
- ... focuses on **connecting** methods to problems correctly (challenges are more philosophical than technical)
- ...is mostly about the **linear model**, through which we will also develop the concepts of
 - **Hypothesis testing**
 - **Model selection**
 - **Variable/feature Selection**

The Two-Cultures

- According to Leo Brieman (2001), there are “two cultures in the use of statistical modeling to reach conclusions from data”:

- **Data Modeling Culture:**

$$x \rightarrow \text{model} \rightarrow y$$

here the statistician decides on a **model**, learns its **parameters**, and assesses its fit

- **Algorithmic Modeling Culture:**

$$x \rightarrow \text{unknown} \rightarrow y$$

here the statistician applies an **algorithm** and assesses its ability to predict unseen y -s given new x -s.

- This course is mostly model based

- Tibshirani & Efron (1993):
***Statistics** is the science of **learning** from experience.*
- Wikipedia (2021):
***Statistics** is the discipline that concerns the **collection**, **organization**, **analysis**, **interpretation**, and **presentation** of **data**.*
- Wikipedia (2021):
***Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge** and **insights** from noisy, structured and unstructured **data**, and **apply** knowledge and actionable insights from data across a broad range of application domains.*

Course outline and organizational matters

Organizational matters

- **Instructor:** Dr. Alon Kipnis
- **Lectures:** Tue. 18:30 - 21:00
- **Teaching Assistant:** Mr. Ben Galili
- **Course Staff Email Address:** alon.kipnis@idc.ac.il
- **Office Hours:** Monday 14:00 - 15:00
- **TA Office Hours:** will be posted on course website

1. Lecture material (slides, sample code, homework etc.) on **Moodle**
(<https://moodle.idc.ac.il/2022/course/>)
2. Other course-related announcements on **Moodle**
3. Discussions on **Piazza**
(<https://piazza.com/class/kz5imoo7xi991>)
4. Home assignments and grades will be posted on **Moodle**

This class is new

Cons:

- Expect more typos and errors in material than usual

Pros:

- Teaching stuff is more attentive to requests and suggestions: let us know if you have suggestions on how to improve your learning experience
- **We are here to help.** We look forward to seeing you in our office hours

Tips for succeeding in the class

- Review previous lecture **before** the beginning of the current one
- Discuss home assignments with peers and instructors; solve **individually**
- Attend office hours **after** reviewing relevant class material

- Lectures will be recorded. They will be available on **Moodle**.
- I strongly encourage you to attend the class live.

- **Israel** time (usually UTC+02:00)
- If you are currently not in Israel, please let us know what time zone you're in.

Prerequisites:

- Calculus and linear algebra
- Introductory course in probability/statistics
- Familiarity with Python and basic packages (numpy, scipy, pandas)

Textbooks

- The class **does not** follow one textbook in particular
- Here is a non-exhaustive list of **relevant books and notes**:
 - Cosma Shalizi, “The Truth About Linear Regression”,
<https://www.stat.cmu.edu/~cshalizi/TALR/>
 - Jonathan Taylor, “Stanford’s STATS 203 lecture notes: Introduction to Regression and Analysis of Variance.” 2005
 - Emanuel Candes, “Stanford’s 300C lecture notes: Theory of Statistics”, 2019
 - “Regression: Linear Models in Statistics”, by Bingham and Fry, 2010.
- Related classes:
 - Art Owen, Stanford STATS 305A: “Applied Statistics”
 - Cosma Sahlizi, CMU 36-401: “Modern Regression”
 - Rob Tibshirani and Trevor Hastie, Stanford STATS 315: “Introduction to Modern Applied Statistics”

Assessment and grading:

- **Grading:** 60% regular homework assignments, 40% exam.
- Exam:
 - About **3 hour** time-limit
 - Ideology: those who solved all home assignments **individually** will receive above 85% of exam's credit

Homeworks

- Constitute **60%** of the final grade.
- Mix of **theoretical** (pen and paper) and **coding** exercises.
- Will be posted about every **two weeks**.
- Due **before** the weekly lecture
- **Late submissions**: 10% penalty for every 24 hours beyond the submission deadline, up to 72 hours after which the submission is no longer accepted.
- **Regrade requests** must be submitted within **one week** after grading has been published

- We encourage discussions between classmates, either on Piazza or elsewhere
- Please send us interesting related dataset and articles so we can share with everyone

Interacting with the Instructors

- Interacting with your instructors is a great way of **promoting your career**
- Several ways of doing so **effectively**:
 - Participate in class discussions
 - Attend office hours
 - Ask/comment on Piazza

Tentative List of Topics

List of Topics

- The linear model (intro to linear regression, ordinary least squares)
- Math and probability review (distribution, multivariate normal distribution, F-distribution, goodness-of-fit, quadratic forms)
- The linear model (continued) (distributional properties of least squares solution, applications)
- Hypothesis testing (basics, one-sample, two-samples, A/B testing, controlled vs. uncontrolled)
- ANOVA (fixed and random effects)
- More linear regression (model-order selection, confidence and prediction bands, multiple regression)
- Other linear response models (logistic/probit, Poisson regression)
- Multiple Testing (FDR, methods of combining P-values)
- Variable selection
- Validation (cross validation) and permutation tests
- Quantile regression