Intro to Machine Learning for the Social Sciences

I. Exploring how Social Science produces knowledge

II. Thinking through the benefits of adding Machine Learning to our toolkit

Lecture Overview

- Review: Knowledge generating process for Social Science
- Introduce you to Machine Learning
- Think through some benefits of ML generally, and for social scientists in particular

The Nature of Science

- Most people associate science with the natural sciences-
  - Chemistry, physics, and biology
- Probably picture scientist in a white lab coat surrounded by test tubes and microscopes
- 76% of Americans think biology is “very scientific”
- Only 8.9% say the same of social science disciplines such as sociology

The Aim of Science

- “Science” from Latin “Scientia”, which means “knowledge”

- Aim of science is:
  - To produce knowledge and…
  - To understand and explain some aspect of the world around us

- What characteristics does social science share with natural science?
- What makes them BOTH sciences?

Science as a Product

- All scientific knowledge shares certain defining characteristics
    - First = types of questions that may be addressed
    - Let’s differentiate between some scientific and nonscientific questions

Scientific Versus Nonscientific Questions

- Scientific questions require answers that are observable.

  - Scientists assume that:
    - The world exists,
    - Empirically verifiable knowledge is possible,
    - That we can know the world through our senses,
    - And that there is order in the world.

  - Philosophical Qs about morality, essence, existence, outside the realm of science.

Knowledge as Description

- First step in producing knowledge is description.
  - Must describe objects and events before we can understand/explain relationships among them
    - Each discipline defines new concepts
    - Definition of Concepts:
      - Abstractions communicated by words or other signs that refer to common properties among phenomena.
      - Example: “Weight” as a concept.
        - Symbolizes a conception of a common property of all physical objects
    - B/C scientists care about building knowledge through observation…
      - Concepts are defined in terms of precise, reliable observations
- When concepts are no longer effective at explaining reality, they can be discarded

Knowledge as Explanation and Prediction

- Scientific knowledge attempts to
  - Explain the past and the present AND
  - To predict the future

- Ideal in science is to
  - Develop the most general understanding:
    - To establish propositions capable of explaining and predicting the widest possible outcomes.

Knowledge as Explanation and Prediction

- Lots of terms used to denote propositions:
  - May be called **empirical generalizations**
    - When they are derived from observations
  - Or **hypotheses** when they have been proposed but not tested
  - If propositions have been repeatedly verified AND are widely accepted, they become known as **laws**.
- As scientific propositions they
  - Describe, explain, and predict phenomena.

Knowledge as Explanation and Prediction

- To explain empirical generalizations and laws, science introduces theories…
- Theory, however, is one of the more misunderstood terms in science.
  - A scientific theory consists of a set of interconnected propositions that have the same form as laws but are more general and abstract
  - Theories must be logically consistent and empirically verifiable
- Many theories might explain evidence for the same phenomena.
  - Ultimately, the best theory will
    - Involve the fewest statements and assumptions
    - Explain the broadest range of phenomena, and
    - Make the most accurate predictions

Knowledge as Explanation and Prediction

- So, to summarize,
    - Scientific concepts describe ***what*** is being studied
    - Scientific hypotheses and theories explain *how* and *why* patterned events occur

Knowledge as Understanding

- The overarching goal of this process is to create a theory that:
  - Describes the causal process connecting events
  - In other words, accurately contributes an understanding of how patterned events occur
- When science produces the causal process it has contributed to our accurate understanding of reality.

Knowledge as Understanding

- Scientific hypotheses and theories provide a sense of understanding by describing the underlying causes of phenomena

Tentative Knowledge

- B.F. Skinner, “science is unique in showing a cumulative progress,” which enables each succeeding generation of scientists to begin a little further along
- Part of the tentative nature of science
  - Answers lead to new questions
  - New fact, law, or theory presents new problems…
  - Such that there is always more to know
- Also, science based on observable evidence
  - Conclusions always open to change with new evidence

Therefore, Knowledge = Best understanding produced thus far

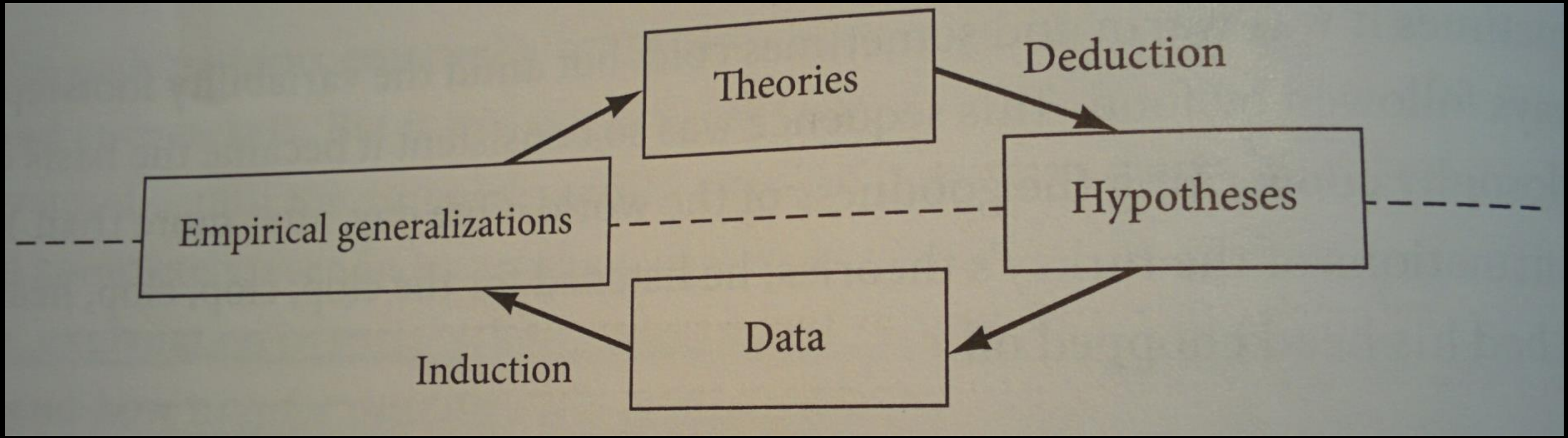
Science as a Process

- Process defined:
    - Signifies a series of operations or actions that bring about an end result.

Science as a Process

- What is the scientific process?
  - At some point scientists…
    - Collect data and record facts
    - Then they try to describe and explain what they see
    - Then they make predictions on the basis of their theories
    - Which they check against their observations (i.e.-facts or data) again

  - In the reported words of Einstein,
    - Science “must start with the facts and end with the facts, no matter what theoretical structures it builds in between”



- Essence of science is a cyclical process involving continuous interplay between theory and research
- Development of theory is the ultimate goal
  - (everything above the dotted line separates the world of theory from the world of research)
- Research supports theory building through
  - Systematic observation that generates data
  - Data used to infer theories and to test theories

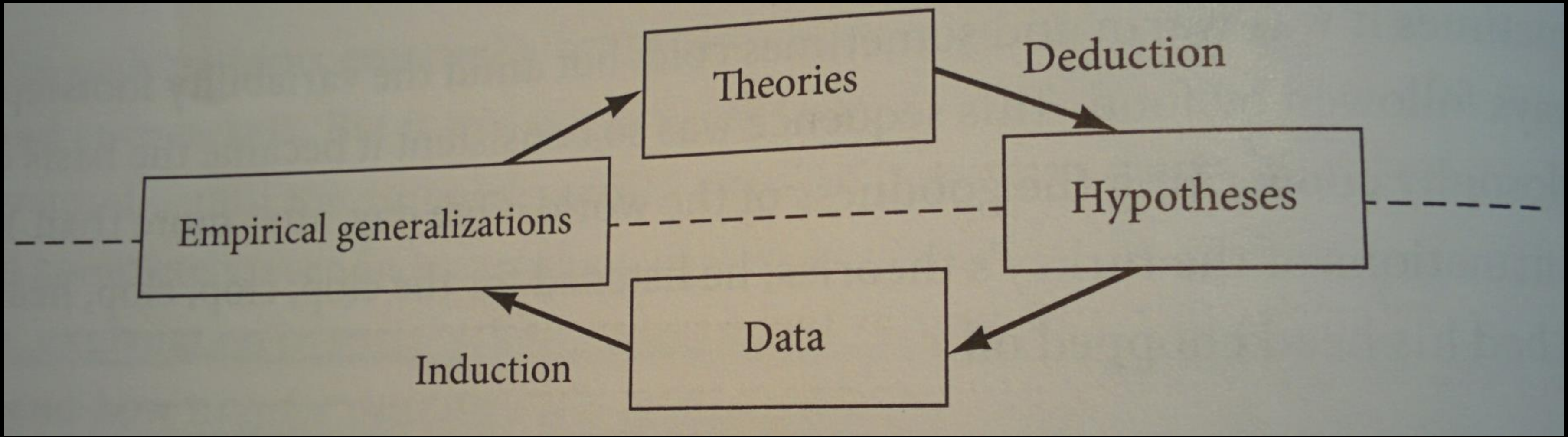
Example: Durkheim’s Study of Suicide

- Suicide one of most studied social problems of 19th century
  - Durkheim’s study of suicide (first published in 1897) good example of interplay between theory and research
    - Regarded as model of social research
- Began by considering existing theories
  - Theories were based on nonsocial factors
    - Insanity
    - Alcoholism
    - Climate
- Argued against theory on purely logical grounds
- Then, he presented data to test his reasoning

Example: Durkheim's Study of Suicide

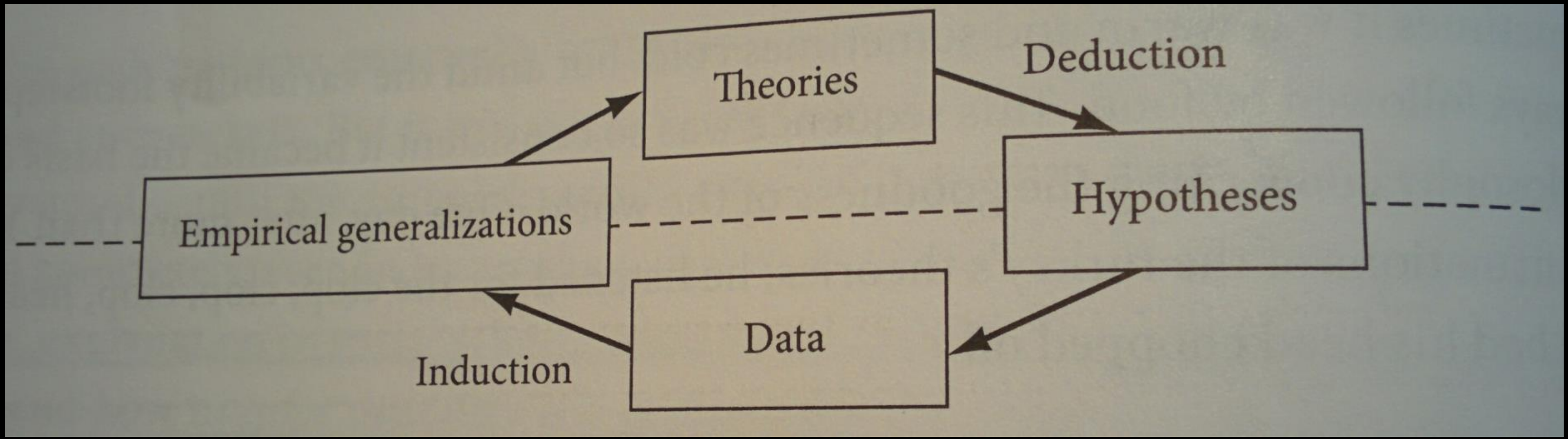
- Arguments against theory that insanity leads to suicide
  - If suicide results from insanity, same groups with high rates of insanity ought to have high rates of suicide

  - Data:
    - Women outnumbered men in insane asylums, but men far more likely to commit suicide
    - Jews had high rates of insanity, but low rates of suicide
    - No correlation between rates of insanity and suicide among 10 European countries
- Moving from theories to hypotheses to data (as seen in figure)



Moving from theories to hypotheses to data (as seen in figure)

- After rejecting factors like insanity that did not fit the data, he collected new data and developed new empirical generalizations…
- Durkheim turned to social factors that may explain increased suicide rates
  - Noted that Catholic countries had lower rates of suicide than Protestant countries
  - Wasn't caused by industrialization, b/c same pattern occurred for catholics and protestant populations within industrialized countries
- Durkheim's empirical generalizations–
  - Protestantism allows more individual thought with less common beliefs and practices
  - Leads to Protestants feeling less of a bond with others
  - And weaker social integration of religious community means individuals will be more likely to commit suicide



- **End result of this scientific study is a theory based in fact or data**
- Durkheim called this the theory of “egoistic suicide”

- This study was part of a continuous scientific cycle-
  - It built on work of Wagner, Morselli, and others
  - Many studies since Durkheim have questioned, refined, and tested his insights

Logical Reasoning

- Scientists rely on deductive and inductive reasoning in this process

- <u>Inductive reasoning </u>is a bottom up process
  - Hubert, Walter, and Joan, who are union members, are Democrats
  - Therefore, all union members are Democrats

  - Sometimes said that induction moves from specific instances to general principles
- Scientists use inductive reasoning when they infer empirical generalizations from data

Logical Reasoning

- Deductive reasoning is a top down process
  - All union members are Democrats
  - Joan belongs to the union
  - Therefore, Joan is a Democrat

  - Sometimes said that deduction moves from general principles to the specific observations or facts
- Scientists use deductive reasoning when they show how a hypothesis explains or predicts specific facts

Logical Reasoning

- Scientists use deductive reasoning when they show how a hypothesis explains or predicts specific facts
  - Durkheim example:
  - If one group is more socially integrated than another, then its suicide rates will be lower. (HYPOTHESIS)
  - Catholics are more socially integrated than Protestants.
  - Therefore, the suicide rate is lower among Catholics than among Protestants (FACT)
- Deductions are either valid or invalid…
  - Allows science to test theoretical expectations.

Principles for gathering evidence

- 1) Empiricism
  - Only study problems/issues that can be resolved by making observations
- Empiricism defined:
  - Way of knowing or understanding the world that relies directly or indirectly on what we experience through our senses
  - Data acceptable insofar as they can be observed
- Statements from authorities not accepted to be true without evidence
  - Or b/c tradition or common sense leads to a way of thinking

Principles for gathering evidence

- 2) Objectivity
  - Typically defined as
    - Observation free from emotion, conjecture, or personal bias.
  - Observation free from bias is assumed to be practically impossible.
  - Instead scientists assign more useful meaning to the word

- Scientific definition:
  - Must be possible to reach **intersubjective testability**
    - Two or more independent observers working under same conditions must be able to agree that they are observing the same thing or event

Principles for gathering evidence

- 2) Objectivity
    - Forces scientists to describe research in detail
    - Outline methods and logic
    - Such that others can retest their findings

- Allows scientific community to judge if subjectivity of research has entered in to empirical generalizations
- Thus objectivity is determined by scientific community

Principles for gathering evidence

- 3) Control
  - Research open to variety of interpretations…
  - Idea of control is to employ procedures that rule out all explanations except one research is interested in.

The Ideal and the Reality of (Social) Scientific Inquiry

- Thus far we have discussed a somewhat idealized view of science
- There are caveats we need to point out that apply to social science
  - First,
    - Social scientific theories tend to be stated less formally than the logical deductions we discussed
    - Certainly they are often stated less formally than the mathematical equations often found in the natural sciences
  - We have defined theory as a set of interrelated propositions from which testable hypotheses can be deduced.

The Ideal and the Reality of (Social) Scientific Inquiry

- Theory has a much looser meaning in social sciences
- May refer to all sorts of speculative ideas offered as explanations for phenomena
- Common to see the terms “theory” and “hypothesis” used interchangeably

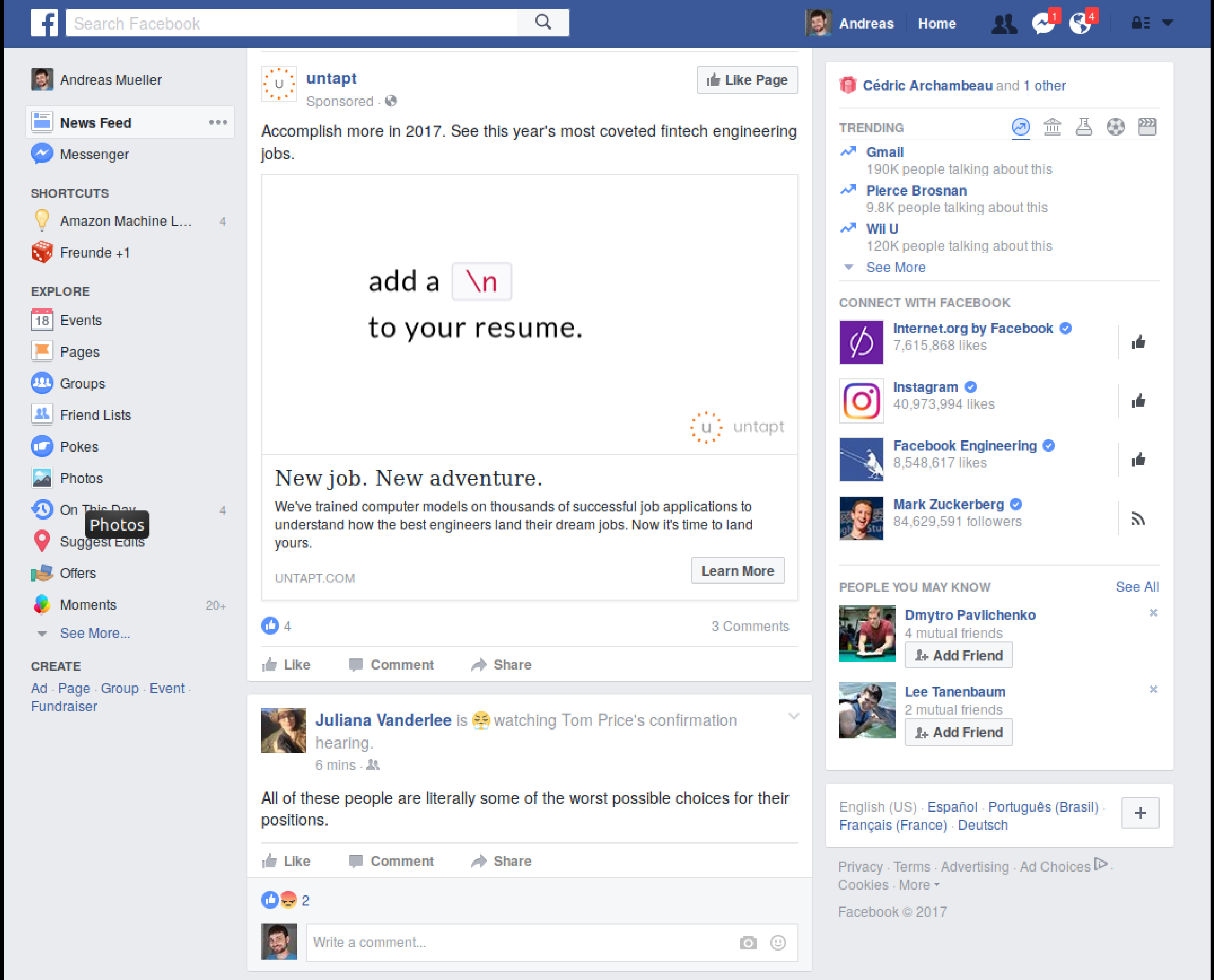
- We study complex social behavior, which leads to a messier process than science in the natural processes.

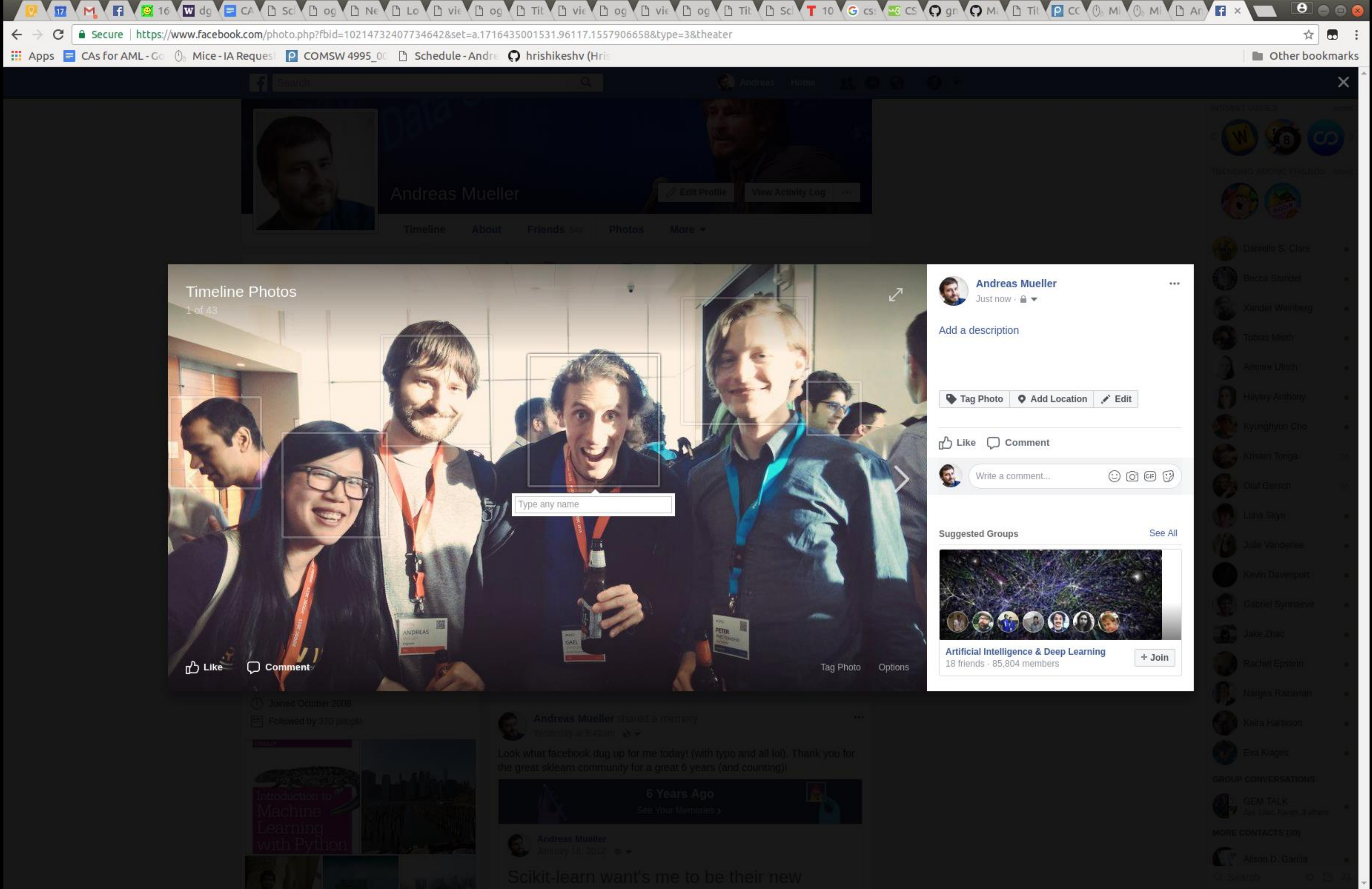
The Ideal and the Reality of (Social) Scientific Inquiry

- The course of social science inquiry tends to be irregular and circuitous
- Sociologist Walter Wallace, process of scientific inquiry may occur:
  - Sometimes quickly, sometimes slowly
  - Sometimes with a high degree of formalization and rigor, sometimes quite informally, unconsciously, intuitively
  - Sometimes through interaction of several scientists in distinct roles
    - (of say theorist, research director, interviewer, methodologist, sampling expert, statistician,etc.)
  - Sometimes through the efforts of a single scientist
  - And sometimes only in the scientist's imagination, sometimes in actual fact

What and Why of Machine Learning

What is machine learning?


Search Facebook
Andreas
Home
Andreas Mueller
News Feed
Messenger
SHORTCUTS
Amazon Machine L...
Freunde +1
EXPLORE
Events
Pages
Groups
Friend Lists
Pokes
Photos
On This Day
Photos
Suggest Edits
Offers
Moments
See More...
CREATE
Ad · Page · Group · Event · Fundraiser
untapt
Sponsored
Like Page
Accomplish more in 2017. See this year's most coveted fintech engineering jobs.
add a \n
to your resume.
untapt
New job. New adventure.
We've trained computer models on thousands of successful job applications to understand how the best engineers land their dream jobs. Now it's time to land yours.
UNTAPT.COM
Learn More
4
3 Comments
Like
Comment
Share
Juliana Vanderlee is watching Tom Price's confirmation hearing.
6 mins
All of these people are literally some of the worst possible choices for their positions.
Like
Comment
Share
2
Write a comment...
Cédric Archambeau and 1 other
TRENDING
Gmail
190K people talking about this
Pierce Brosnan
9.8K people talking about this
Wii U
120K people talking about this
See More
CONNECT WITH FACEBOOK
Internet.org by Facebook
7,615,868 likes
Instagram
40,973,994 likes
Facebook Engineering
8,548,617 likes
Mark Zuckerberg
84,629,591 followers
PEOPLE YOU MAY KNOW
See All
Dmytro Pavlichenko
4 mutual friends
Add Friend
Lee Tanenbaum
2 mutual friends
Add Friend
English (US) · Español · Português (Brasil) · Français (France) · Deutsch
Privacy · Terms · Advertising · Ad Choices · Cookies · More
Facebook © 2017


Secure | https://www.facebook.com/photo.php?fbid=10214732407734642&set=a.1716435001531.96117.1557906658&type=3&theater
Apps
CAs for AML - Go
Mice - IA Request
COMSW 4995_00
Schedule - Andre
hrishikeshv (Hri
Other bookmarks
Andreas Mueller
Edit Profile
View Activity Log
Timeline
About
Friends
Photos
More
Timeline Photos
1 of 43
Type any name
Like
Comment
Tag Photo
Options
Andreas Mueller
Just now
Add a description
Tag Photo
Add Location
Edit
Like
Comment
Write a comment...
Suggested Groups
See All
Artificial Intelligence & Deep Learning
18 friends · 85,804 members
+ Join
Andreas Mueller shared a memory.
Look what facebook dug up for me today! (with typo and all lol). Thank you for the great sklearn community for a great 6 years (and counting)!
6 Years Ago
See Your Memories
Introduction to Machine Learning with Python
Scikit-learn want's me to be their new

Search for people, places and things

Andreas Home

Andreas Mueller added 8 new photos.
June 6 at 5:07pm ·

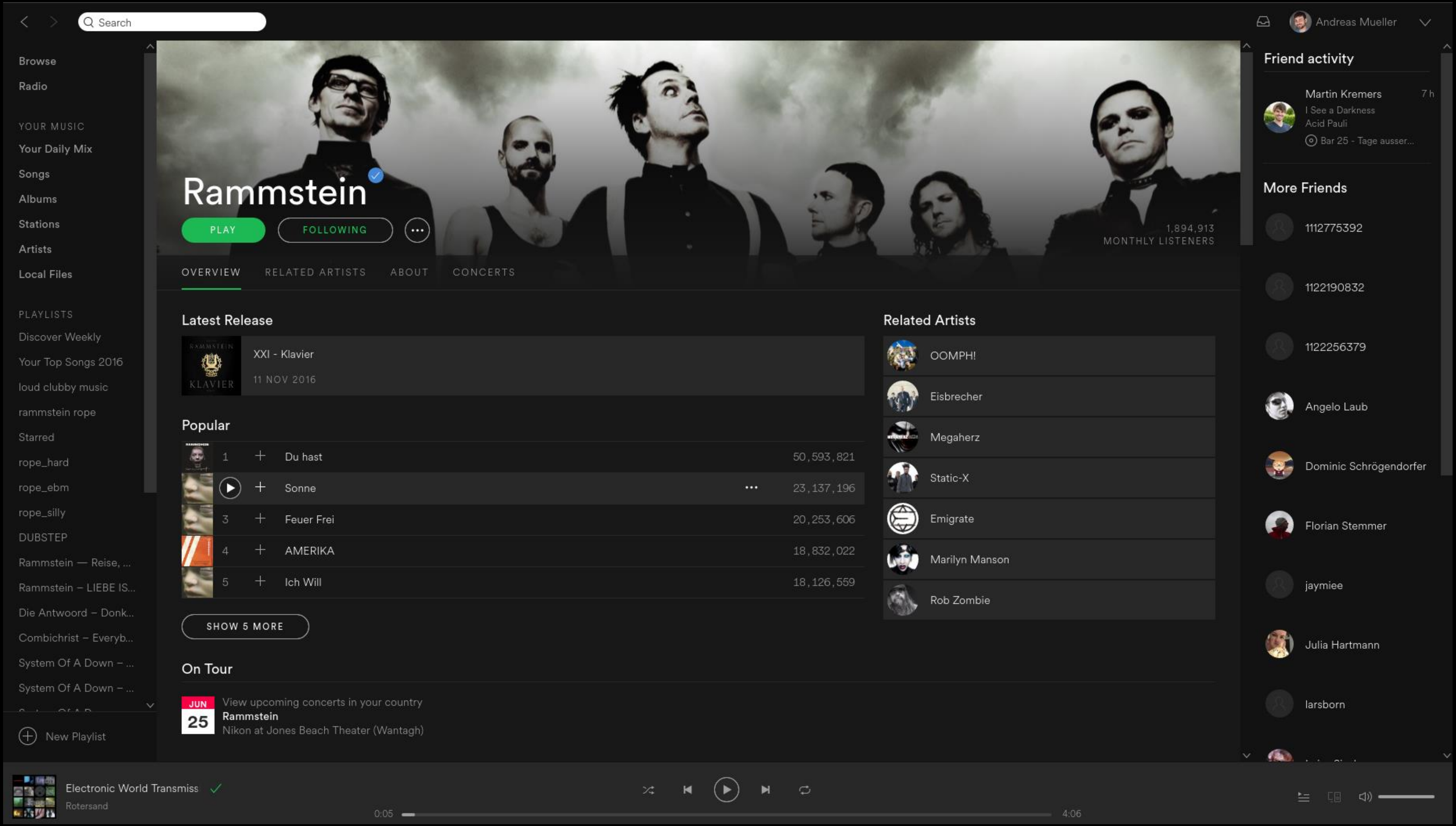
Bye bye hong kong

Like · Comment · Share

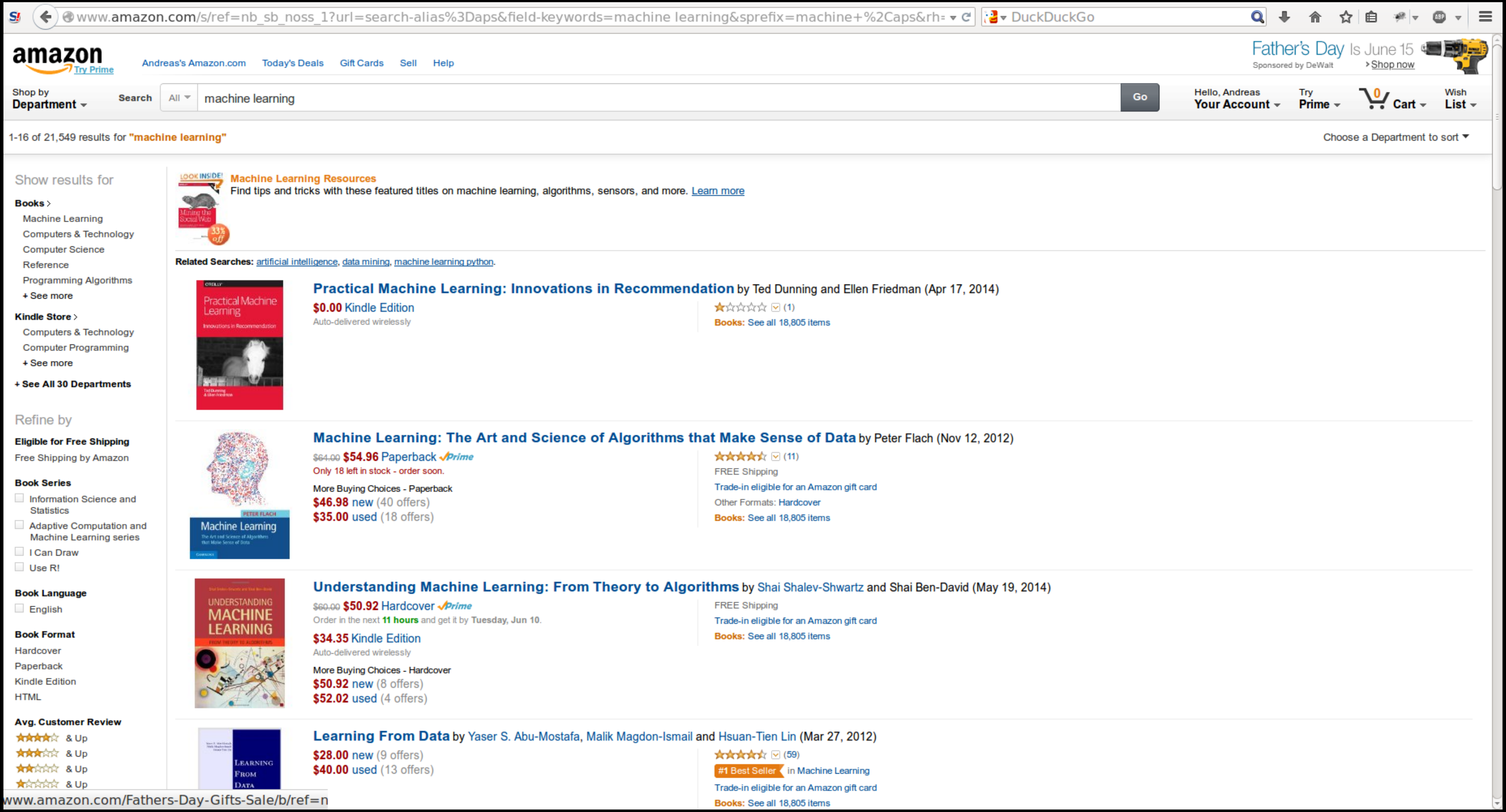
and 6 others like this.

Looks like such a vibrant city! I bet it was awesome 🙂
June 6 at 7:39pm · Like

Write a comment...



Search
Andreas Mueller
Browse
Radio
YOUR MUSIC
Your Daily Mix
Songs
Albums
Stations
Artists
Local Files
PLAYLISTS
Discover Weekly
Your Top Songs 2016
loud clubby music
rammstein rope
Starred
rope_hard
rope_ebm
rope_silly
DUBSTEP
Rammstein — Reise, ...
Rammstein – LIEBE IS...
Die Antwoord – Donk...
Combichrist – Everyb...
System Of A Down – ...
System Of A Down – ...
New Playlist
Rammstein
PLAY
FOLLOWING
1,894,913
MONTHLY LISTENERS
OVERVIEW
RELATED ARTISTS
ABOUT
CONCERTS
Latest Release
XXI - Klavier
11 NOV 2016
Popular
1 Du hast 50,593,821
Sonne 23,137,196
3 Feuer Frei 20,253,606
4 AMERIKA 18,832,022
5 Ich Will 18,126,559
SHOW 5 MORE
On Tour
JUN
25
View upcoming concerts in your country
Rammstein
Nikon at Jones Beach Theater (Wantagh)
Related Artists
OOMPH!
Eisbrecher
Megaherz
Static-X
Emigrate
Marilyn Manson
Rob Zombie
Friend activity
Martin Kremers 7 h
I See a Darkness
Acid Pauli
Bar 25 - Tage ausser...
More Friends
1112775392
1122190832
1122256379
Angelo Laub
Dominic Schrögendorfer
Florian Stemmer
jaymiee
Julia Hartmann
larsborn
Electronic World Transmiss
Rotersand
0:05
4:06

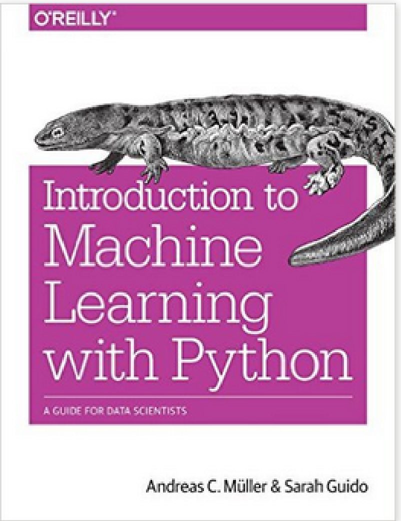

www.amazon.com/s/ref=nb_sb_noss_1?url=search-alias%3Daps&field-keywords=machine learning&sprefix=machine+%2Caps&rh=
DuckDuckGo
amazon
Try Prime
Andreas's Amazon.com
Today's Deals
Gift Cards
Sell
Help
Father's Day Is June 15
Sponsored by DeWalt
Shop now
Shop by Department
Search
All
machine learning
Go
Hello, Andreas Your Account
Try Prime
0 Cart
Wish List
1-16 of 21,549 results for "machine learning"
Choose a Department to sort
Show results for
Books
Machine Learning
Computers & Technology
Computer Science
Reference
Programming Algorithms
+ See more
Kindle Store
Computers & Technology
Computer Programming
+ See more
+ See All 30 Departments
Refine by
Eligible for Free Shipping
Free Shipping by Amazon
Book Series
Information Science and Statistics
Adaptive Computation and Machine Learning series
I Can Draw
Use R!
Book Language
English
Book Format
Hardcover
Paperback
Kindle Edition
HTML
Avg. Customer Review
& Up
& Up
& Up
& Up
Machine Learning Resources
Find tips and tricks with these featured titles on machine learning, algorithms, sensors, and more. Learn more
Related Searches: artificial intelligence, data mining, machine learning python.
Practical Machine Learning: Innovations in Recommendation by Ted Dunning and Ellen Friedman (Apr 17, 2014)
$0.00 Kindle Edition
Auto-delivered wirelessly
(1)
Books: See all 18,805 items
Machine Learning: The Art and Science of Algorithms that Make Sense of Data by Peter Flach (Nov 12, 2012)
$64.00 $54.96 Paperback Prime
Only 18 left in stock - order soon.
More Buying Choices - Paperback
$46.98 new (40 offers)
$35.00 used (18 offers)
(11)
FREE Shipping
Trade-in eligible for an Amazon gift card
Other Formats: Hardcover
Books: See all 18,805 items
Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David (May 19, 2014)
$60.00 $50.92 Hardcover Prime
Order in the next 11 hours and get it by Tuesday, Jun 10.
$34.35 Kindle Edition
Auto-delivered wirelessly
More Buying Choices - Hardcover
$50.92 new (8 offers)
$52.02 used (4 offers)
FREE Shipping
Trade-in eligible for an Amazon gift card
Books: See all 18,805 items
Learning From Data by Yaser S. Abu-Mostafa, Malik Magdon-Ismail and Hsuan-Tien Lin (Mar 27, 2012)
$28.00 new (9 offers)
$40.00 used (13 offers)
(59)
#1 Best Seller in Machine Learning
Trade-in eligible for an Amazon gift card
Books: See all 18,805 items
www.amazon.com/Fathers-Day-Gifts-Sale/b/ref=n

Introduction to Machine Learning with Python: A Guide for Data Scientists 1st Edition

by Andreas C. Müller (Author), Sarah Guido (Author)

9 customer reviews

Look inside



ISBN-13: 978-1449369415
ISBN-10: 1449369413
Why is ISBN important?

Have one to sell? Sell on Amazon

Add to List

Share

Kindle	Paperback	Other Sellers
$24.99	$41.23	from $26.53

Buy new — Prime $41.23

In Stock.

Ships from and sold by Amazon.com. Gift-wrap available.

Prime

List Price: ~~$49.99~~ Save: $8.76 (18%)

36 New from $26.53

Qty: 1

Want it tomorrow, Jan. 19? Order within **5 hrs 20 mins** and choose **One-Day Shipping** at checkout. Details

Add to Cart

Turn on 1-Click ordering

Ship to:

Andreas C Mueller- New York - 10009

More Buying Choices

51 used & new from $26.53

36 New from $26.53 | 15 Used from $29.09

See All Buying Options

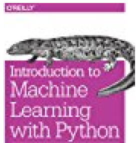
Machine learning has become an integral part of many commercial applications and research projects, but this field is not exclusive to large companies with extensive research teams. If you use Python, even as a beginner, this book will teach you practical ways to build your own machine learning solutions. With all the data available today, machine learning applications are limited only by your imagination.

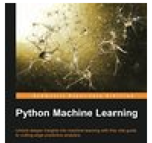
You'll learn the steps necessary to create a successful machine-learning application with Python and the

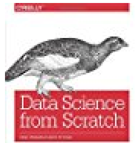
Read more

Report incorrect product information.

Frequently Bought Together



+


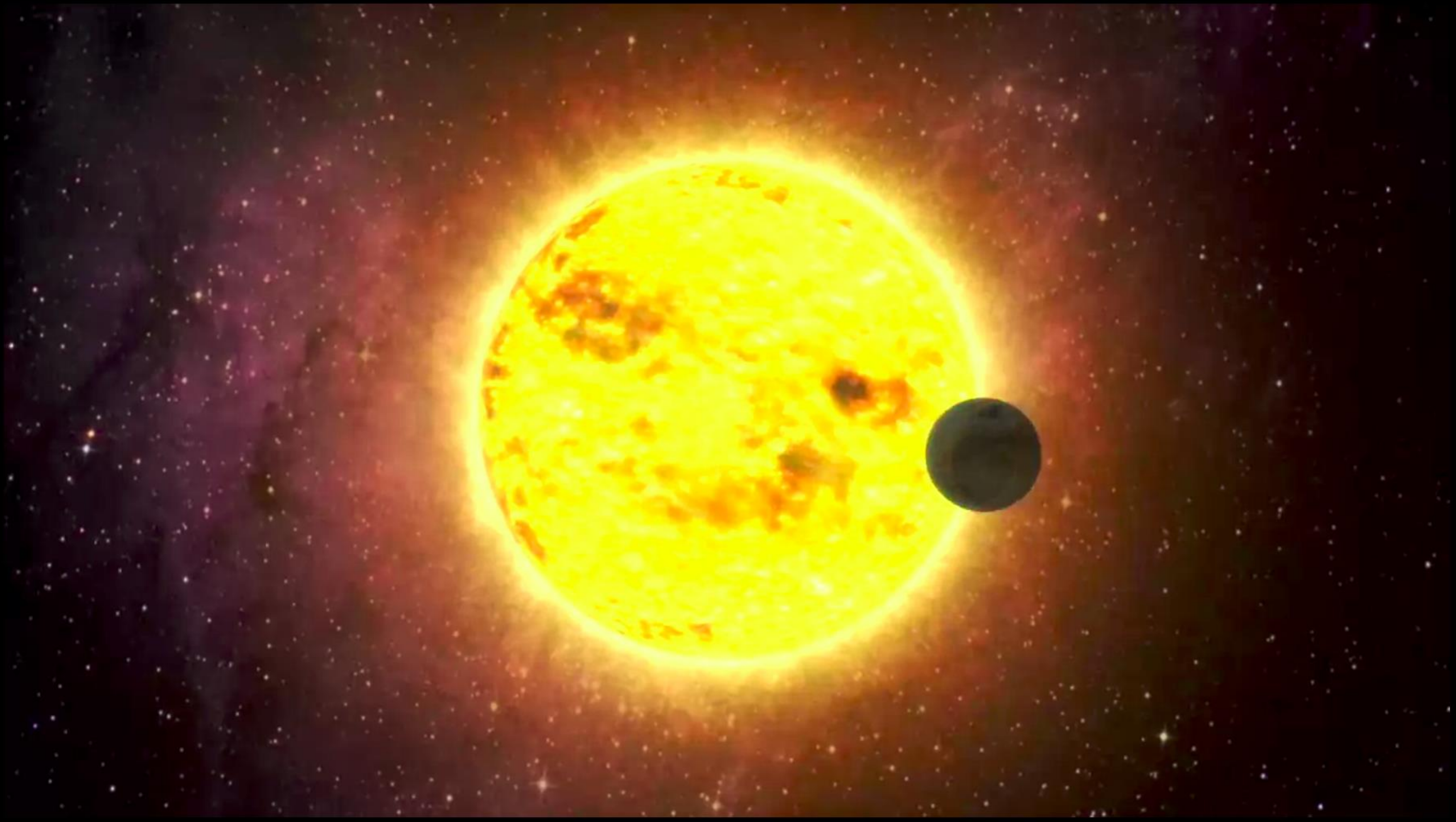
+


Total price: $108.29

Add all three to Cart

Add all three to List

Science!



Types of Machine Learning

Types of Machine Learning

Supervised

Unsupervised

Reinforcement

Supervised Learning

$$(x_i, y_i) \propto p(x, y) \text{ i.i.d.}$$

$$x_i \in \mathbb{R}^p$$

$$y_i \in \mathbb{R}$$

$$f(x_i) \approx y_i$$

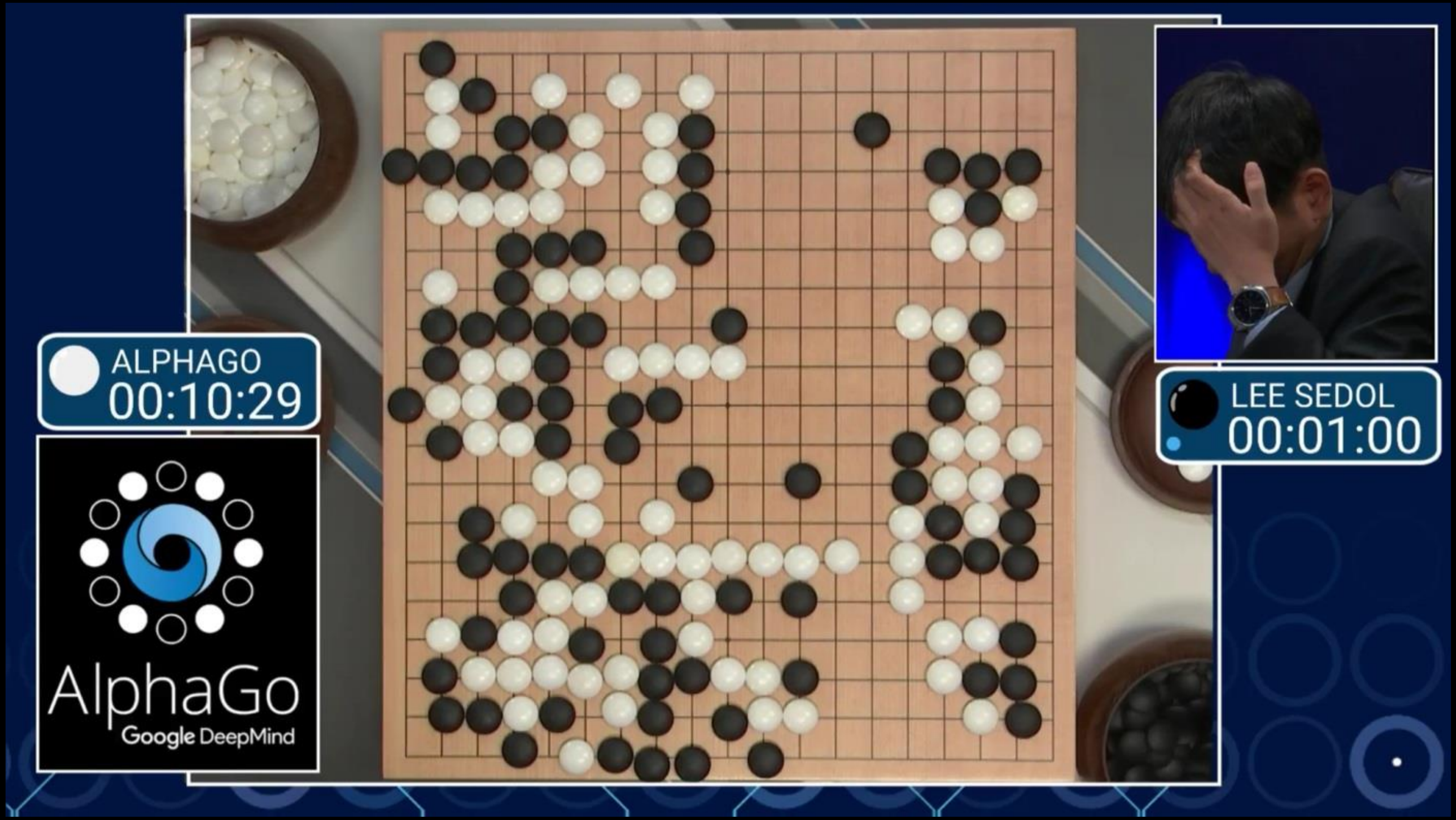
Some examples of Supervised Learning

Unsupervised Learning

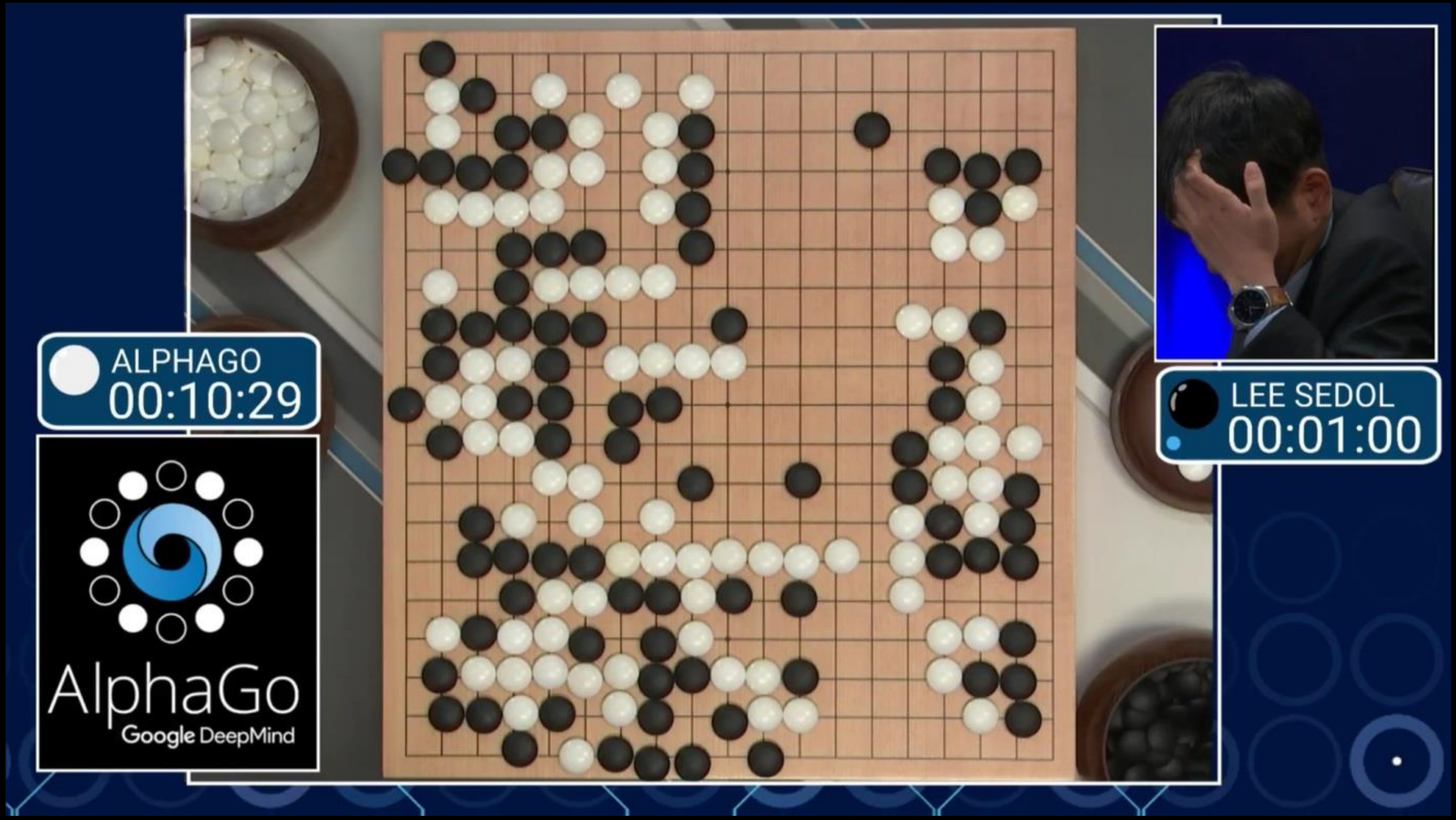
$$x_i \propto p(x) \text{ i.i.d.}$$

Learn about p.

Reinforcement Learning



Reinforcement Learning



Other kinds of learning

Classification and Regression

Classification

- target y discrete
- Will you pass?

Regression

- target y continuous
- How many points will you get in the exam?

Generalization

Not only $f(x_i) \approx y_i$,

also for new data: $f(x) \approx y$

Relationship to Social Science approach

- Statistics focus
  - Causation
  - Inference Making
  - Real world insights
  - Model coefficients represent potential causal factors

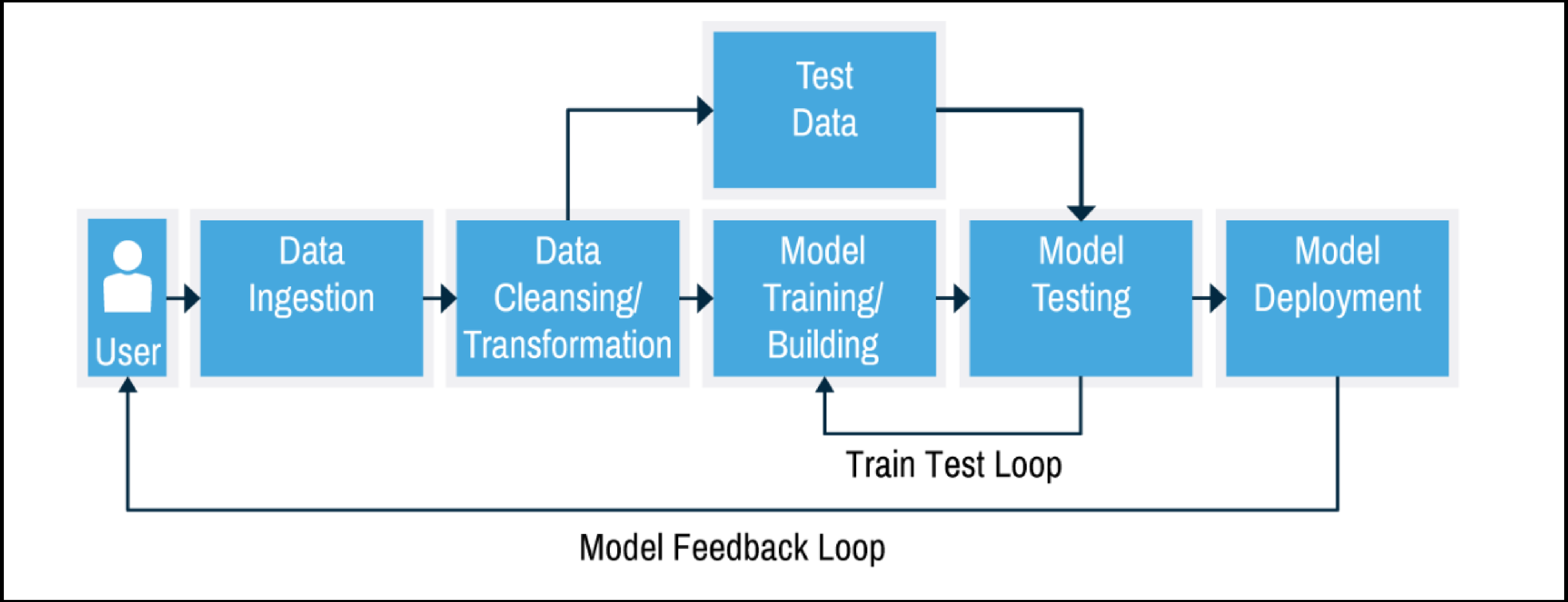
- Machine Learning focus
  - Prediction
    - Elevates importance of training data
  - Real world insights secondary to accuracy of prediction

Relationship to Social Science approach

- Statistics focus
  - Causation
  - Inference Making
  - Real world insights
  - Model coefficients represent potential causal factors

- Machine Learning focus
  - Prediction
    - Elevates importance of training data
  - Real world insights secondary to accuracy of prediction

The Machine Learning Work-Flow



Overview of the course

Infrastructure and basic tools

- Python, Jupyter
- Numpy,
- matplotlib,
- Pandas
- (Extra) git, github

Getting started w/ Supervised Learning

- Generalization in practice
- Nearest Neighbors, Nearest Centroid
- Linear Classification and Regression
- Penalties and Regularization

Data Preparation

- Preprocessing
- Feature engineering
- Dealing with missing values
- Feature Selection

Non-linear machine learning models

- Support Vector Machines
- Decision Trees
- Random Forests
- Gradient Boosting
- Model Calibration

Model evaluation and imbalanced data

- Metrics for binary and multi-class classification
- Metrics for regression
- Analyzing predictions
- Handling imbalanced datasets for classification

Decomposition Methods

- PCA
- Discriminant Analysis
- Manifold Learning
- Non-negative Matrix Factorization

Clustering

- K-Means
- DBScan
- Agglomerative Clustering
- Spectral Clustering
- Supervised evaluation metrics
- Unsupervised evaluation metrics

Outlier Detection

- One Class SVM
- Robust Covariance Estimates
- Isolation Forests

Neural Networks

- Backpropagation
- Tensorflow
- Keras
- Learning Algorithms
- Image data and convolutional neural networks
- Best practices for neural networks

What will you take away from the course?

1. Powerful new tools for building datasets
2. Overview of cutting edge modelling techniques
3. Python coding skills for stats/data wrangling/ ML implementation
4. Understanding of main ML algorithms
5. Hopefully, will also help you come up with innovative new ideas for your work

Questions?