[80245013 Machine Learning, Fall, 2019]

Machine Learning

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://ml.cs.tsinghua.edu.cn/~jun
Institute for AI
Tsinghua University

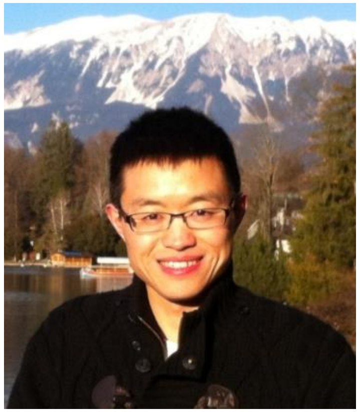
September 10, 2019

A bit about Jun…

- Jun Zhu, Professor, Depart. of Computer Science. I received Ph.D. in 2009. My research interest includes machine learning, Bayesian methods, and data mining
- I did post-doc at the Machine Learning Department in CMU with Prof. Eric P. Xing. Before that I was invited to visit CMU for twice. I was also invited to visit Stanford for joint research (with Prof. Li Fei-Fei)
- 2015-2018: Adjunct Associate Professor at CMU
- Published 100+ papers on the top-tier ML conferences and journals, including JMLR, TPAMI, ICML, NIPS, etc.
- Served as Area Chairs for ICML, NIPS, UAI, AAAI, IJCAI; Associate Editor-in-Chief for PAMI, AI Journal
- Research is supported by National 973, NSFC, "Tsinghua 221 Basic Research Plan for Young Talents".
- IEEE AI's 10 to Watch; MIT TR35 China (pioneers)
- Homepage: http://ml.cs.tsinghua.edu.cn/~jun

A bit about Jie…

- Jie Tang, Professor, Department of Computer Science of Tsinghua University. My research interests include social network, data mining, and machine learning.
- I have been visiting scholar at Cornell U. (working with John Hopcroft, Jon Kleinberg), UIUC (working with Jiawei Han), CUHK (with Jeffrey Yu), and HKUST (with Qiong Luo).



- I was awarded with the CCF Young Scientist Award, NSFC Excellent Young Scholar, Newton Advanced Fellowships Award, IBM Innovation Faculty Award, and New Star of Beijing S&T.
- Have published more than 200 paper on major international conf/journals, including KDD (19), IJCAI/AAAI (16), IEEE Trans. (21), ICML, Machine Learning
- #Citation: 7,962 and H-index: 46
- Have a notable system, AMiner.org for academic researcher network analysis. The system has attracted 8.32 million users from 220 countries/regions.
- **Homepage:** http://keg.cs.tsinghua.edu.cn/jietang/

Contact Information

- Jun Zhu
  - Institute for AI, Department of Computer Science, Tsinghua U.

  - Office: Rm 4-513, FIT Building
  - E-mail: dcszj@tsinghua.edu.cn
  - Phone: 62772322, 18810502646
  - Office hours: Thursday afternoon 3:30pm-5:00pm
    - Better to make an appointment in advance

Contact Information

- Jie Tang
  - Software Division, Department of Computer Science, Tsinghua U.
  - Office: Rm 1-308, FIT Building
  - E-mail: jietang@tsinghua.edu.cn
  - Phone: 62788788-20, 13911215746
  - Open hours: Tuesday afternoon 2:00pm-5:00pm

Teaching Assistants

- Tianyu Pang
  - E-mail: pty17@mails.tsinghua.edu.cn
  - Phone: 62795869, 13661150589
  - Bayesian methods, Deep learning
  - Publish at ICML, NeurIPS, CVPR

TA from Jie’s group

- Qibin Chen
  - PhD student
  - Publish at KDD, ACL, EMNLP
  - https://www.qibin.ink/

Resources

- Mainly class slides/notes
- Recommended text books
  - Christopher M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2007.
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman. Elements of Statistical Learning. 2nd Edition, Springer, 2009.
  - Yoshua Bengio, Ian J. Goodfellow, and Aaron Courville. Deep Learning. 2016.

- Further readings:
  - Conferences:
    - Theory: ICML, NIPS, UAI, COLT, AISTATS, AAAI, IJCAI
    - App: KDD, SIGIR, WWW, ACL
  - Journals:
    - JMLR, PAMI, MLJ

Prerequisites

- Knowledge of probability, linear algebra, statistics and algorithms
  - Calculus:
    - Derivative, integral of multivariate functions
  - Linear Algebra
    - Matrix inversion, eigen-decomposition, …
  - Basic Probability and Statistics
    - Probability distributions, Mean, Variance, Conditional probabilities, Bayes rule, …

- Knowledge of programming languages, e.g., C/C++, Java, matlab, Python

- **Homework 0**: take the Self-Evaluation
  - Minimum & modest background tests (available at course webpage)

Potential achievements

- Able to understand the underlying principles of classical ML algorithms

- Able to apply right ML algorithms to the applications at your hand

- Able to design effective ML algorithms to solve new problems

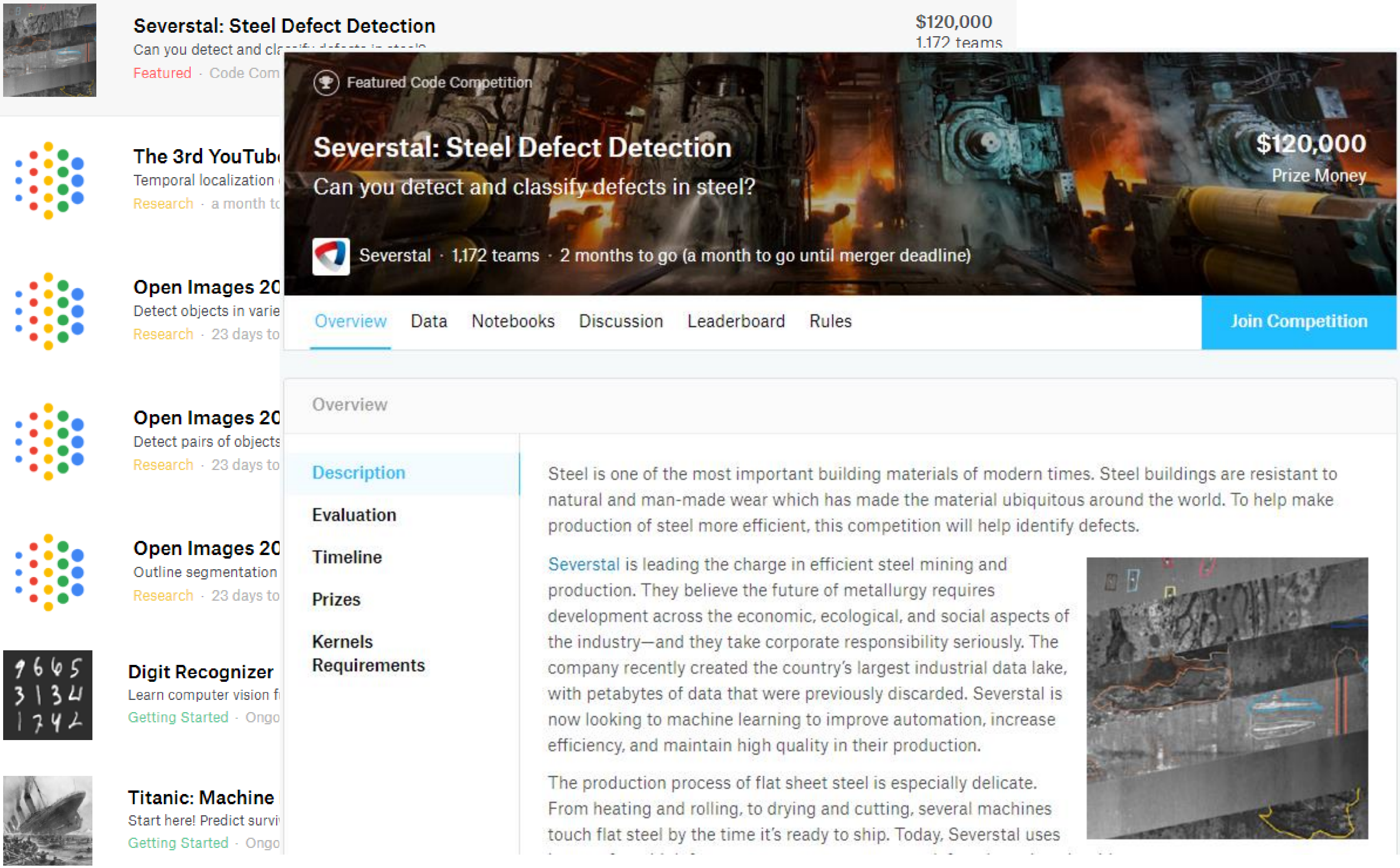
Overview of Class

◈ Introduction	3 units	
◈ Unsupervised learning	6 units	HW1 out
◈ Supervised learning	6 units	
◈ Reinforcement Learning	6 units	HW1 due HW2 out
◈ Convolutional neural network	3 units	
◈ Auto-Encoders	3 units	HW2 due HW3 out
◈ Recurrent neural network	3 units	
◈ Representation Learning	6 units	HW3 due HW4 out
◈ GAN and AutoML	6 units	
		HW4 due in 2 weeks

Grading

- Participation (10%)
  - 1 mid-term quiz (10 points)
- Homeworks (40%)
  - 4 homeworks (10 points each time)
- Project (50%)
  - <=2 students to form a team
  - Apply machine learning to solve a real problem
    - Choose one task at Kaggle (http://www.kaggle.com/competitions)
  - Submit materials:
    - a proposal (6th week), a mid-term report (9th week), a final report (18th week), and the implementation code (18th week)
  - All reports should be in NIPS format, written in English: (http://nips.cc/Conferences/2014/PaperInformation/StyleFiles)
  - Poster presentation (16th week)

Some example Kaggle tasks



Severstal: Steel Defect Detection
Can you detect and cl
Featured · Code Com
$120,000
1,172 teams

The 3rd YouTub
Temporal localization
Research · a month t

Open Images 20
Detect objects in varie
Research · 23 days to

Open Images 20
Detect pairs of objects
Research · 23 days to

Open Images 20
Outline segmentation
Research · 23 days to

Digit Recognizer
Learn computer vision f
Getting Started · Ongo

Titanic: Machine
Start here! Predict survi
Getting Started · Ongo

Featured Code Competition

Severstal: Steel Defect Detection

Can you detect and classify defects in steel?

$120,000
Prize Money

Severstal · 1,172 teams · 2 months to go (a month to go until merger deadline)

Overview · Data · Notebooks · Discussion · Leaderboard · Rules · Join Competition

Overview

- Description
- Evaluation
- Timeline
- Prizes
- Kernels Requirements

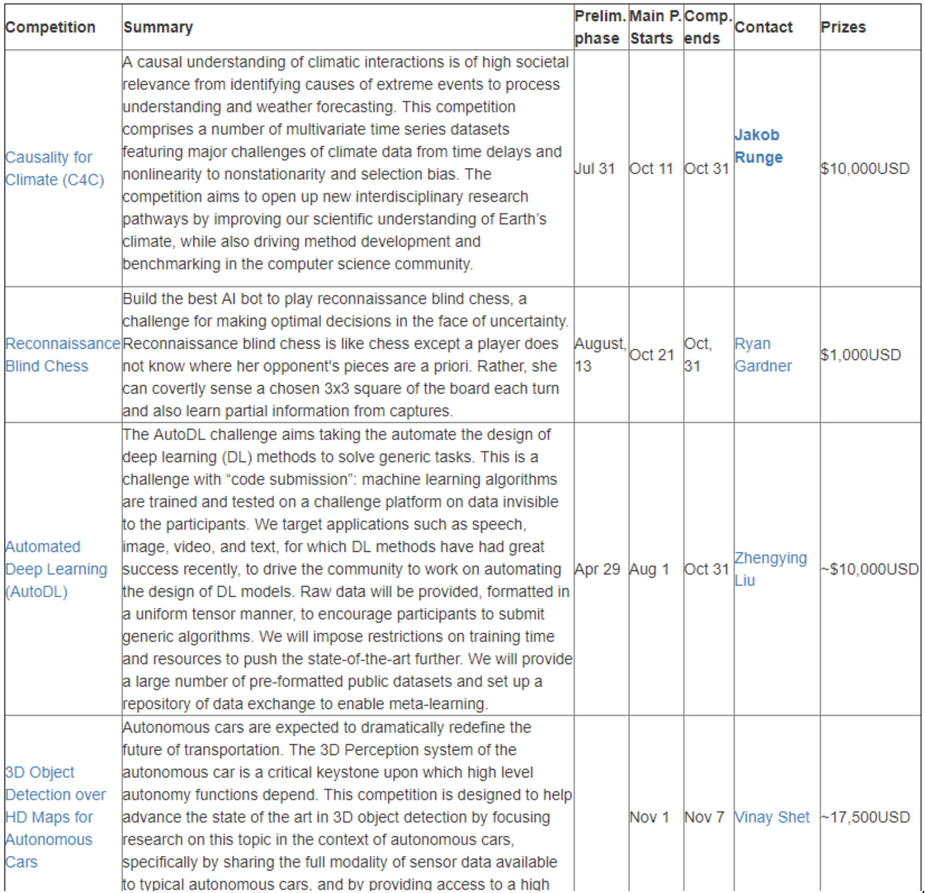
Steel is one of the most important building materials of modern times. Steel buildings are resistant to natural and man-made wear which has made the material ubiquitous around the world. To help make production of steel more efficient, this competition will help identify defects.

Severstal is leading the charge in efficient steel mining and production. They believe the future of metallurgy requires development across the economic, ecological, and social aspects of the industry—and they take corporate responsibility seriously. The company recently created the country's largest industrial data lake, with petabytes of data that were previously discarded. Severstal is now looking to machine learning to improve automation, increase efficiency, and maintain high quality in their production.

The production process of flat sheet steel is especially delicate. From heating and rolling, to drying and cutting, several machines touch flat steel by the time it's ready to ship. Today, Severstal uses

NeurIPS Competitions

- Website: https://neurips.cc/Conferences/2019/CallForCompetitions
- Many are research oriented
- Early due dates
- Datasets can be used



Competition	Summary	Prelim. phase	Main P. Starts	Comp. ends	Contact	Prizes
Causality for Climate (C4C)	A causal understanding of climatic interactions is of high societal relevance from identifying causes of extreme events to process understanding and weather forecasting. This competition comprises a number of multivariate time series datasets featuring major challenges of climate data from time delays and nonlinearity to nonstationarity and selection bias. The competition aims to open up new interdisciplinary research pathways by improving our scientific understanding of Earth's climate, while also driving method development and benchmarking in the computer science community.	Jul 31	Oct 11	Oct 31	**Jakob Runge**	$10,000USD
Reconnaissance Blind Chess	Build the best AI bot to play reconnaissance blind chess, a challenge for making optimal decisions in the face of uncertainty. Reconnaissance blind chess is like chess except a player does not know where her opponent's pieces are a priori. Rather, she can covertly sense a chosen 3x3 square of the board each turn and also learn partial information from captures.	August, 13	Oct 21	Oct, 31	Ryan Gardner	$1,000USD
Automated Deep Learning (AutoDL)	The AutoDL challenge aims taking the automate the design of deep learning (DL) methods to solve generic tasks. This is a challenge with "code submission": machine learning algorithms are trained and tested on a challenge platform on data invisible to the participants. We target applications such as speech, image, video, and text, for which DL methods have had great success recently, to drive the community to work on automating the design of DL models. Raw data will be provided, formatted in a uniform tensor manner, to encourage participants to submit generic algorithms. We will impose restrictions on training time and resources to push the state-of-the-art further. We will provide a large number of pre-formatted public datasets and set up a repository of data exchange to enable meta-learning.	Apr 29	Aug 1	Oct 31	Zhengying Liu	~$10,000USD
3D Object Detection over HD Maps for Autonomous Cars	Autonomous cars are expected to dramatically redefine the future of transportation. The 3D Perception system of the autonomous car is a critical keystone upon which high level autonomy functions depend. This competition is designed to help advance the state of the art in 3D object detection by focusing research on this topic in the context of autonomous cars, specifically by sharing the full modality of sensor data available to typical autonomous cars, and by providing access to a high		Nov 1	Nov 7	Vinay Shet	~17,500USD

- If the end date is later than the end of this semester, report the position in the leaderboard;
- Otherwise, follow the standard partition or ask TAs to define a train/test split and compare your methods with 1 or 2 baselines.

Other Projects

- Self-defined topics
  - Need to propose as early as possible to filter out improper ones

- Other candidates
  - Chinese handwritten characters generation and recognition
  - Adversarial attacks and defense of deep learning
  - Deepfake detection challenge
  - Reinforcement learning
  - More to come

About final report

- We expect to see
  - Problems (what?)
  - Motivations (why?)
  - Techniques (how?)
  - Results & Analysis (did you verify what you claimed above?)
  - Conclusions

- The final report should look like a NeurIPS technical paper
  - Style file: https://neurips.cc/Conferences/2019/PaperInformation/StyleFiles

Questions?