Topics on ML Systems

Fall 2020
Babak Falsafi
Martin Jaggi
Anne-Marie Kermarrec

parsa.epfl.ch/course-info/cs723
Where are we?

### Class intro
- Logistics
- Grades
- Topical intro

<table>
<thead>
<tr>
<th>M</th>
<th>T</th>
<th>W</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>14-Sep</td>
<td>15-Sep</td>
<td>17-Sep</td>
<td>18-Sep</td>
<td></td>
</tr>
<tr>
<td>21-Sep</td>
<td>22-Sep</td>
<td>23-Sep</td>
<td>24-Sep</td>
<td>25-Sep</td>
</tr>
<tr>
<td>28-Sep</td>
<td>29-Sep</td>
<td>30-Sep</td>
<td>01-Oct</td>
<td>02-Oct</td>
</tr>
<tr>
<td>05-Oct</td>
<td>06-Oct</td>
<td>07-Oct</td>
<td>08-Oct</td>
<td>09-Oct</td>
</tr>
<tr>
<td>02-Nov</td>
<td>03-Nov</td>
<td>04-Nov</td>
<td>05-Nov</td>
<td>06-Nov</td>
</tr>
<tr>
<td>09-Nov</td>
<td>10-Nov</td>
<td>11-Nov</td>
<td>12-Nov</td>
<td>13-Nov</td>
</tr>
<tr>
<td>16-Nov</td>
<td>17-Nov</td>
<td>18-Nov</td>
<td>19-Nov</td>
<td>20-Nov</td>
</tr>
<tr>
<td>23-Nov</td>
<td>24-Nov</td>
<td>25-Nov</td>
<td>26-Nov</td>
<td>27-Nov</td>
</tr>
<tr>
<td>30-Nov</td>
<td>01-Dec</td>
<td>02-Dec</td>
<td>03-Dec</td>
<td>04-Dec</td>
</tr>
<tr>
<td>07-Dec</td>
<td>08-Dec</td>
<td>09-Dec</td>
<td>10-Dec</td>
<td>11-Dec</td>
</tr>
<tr>
<td>14-Dec</td>
<td>15-Dec</td>
<td>16-Dec</td>
<td>17-Dec</td>
<td>18-Dec</td>
</tr>
</tbody>
</table>
Who should take CS723?

Graduate Students (MS/PhD)
1. Computer system designers
2. Big data folks

Required knowledge
• Machine learning
• Computer systems

About the Course
• Discussion-oriented
• Emphasis on reading/understanding cutting-edge issues

Feedback
• Individual feedback upon request
Where do I find info about CS723?

Info about the class:

- **Web:** parsa.epfl.ch/course-info/cs723
  - Reading list
  - Presentation schedule
- **Slack:** cs723-2020.slack.com
- **Zoom:** link available via email & slack
Logistics for the course

Class times
  • Lectures: Thursdays 10:15-12:00, online

Profs
  • Babak Falsafi
  • Martin Jaggi
  • Anne-Marie Kermarrec

Postdoc (will help)
  • Yunho Oh
CS 723: A Transversal Skill Course

Read & understand technical work
  • Super useful in research & industry

Communicate:
  • Write about what you read & understood
  • Present it

Why is communication key?
  • Need to explain your work (helps you understand it)
  • Need to explain others’ work (e.g., in management)
CS 723: Components

Readings
- Two readings per week

Write-ups
- Simple (one or two paragraphs) answer to each of six questions

Presentations
- First, an elevator pitch to warm up the audience
- Then, present your answers to six questions
- Finally, in-class discussion on the six questions
Readings (on the web)

Two papers per week

This week (only):
- Task of the Referee by Alan Smith
Write-ups

One write-up (for each paper)

Answer the following questions

1. What is the problem? / How important is it?
2. What are the insights?
3. What is the solution? / Is it feasible?
4. What is the takeaway message?
5. Will this paper win the test of time award?
6. Name one reason why this paper should have not appeared in MLSYS, NeurIPS, ICML, OSDI, ASPLOS, etc.?

Email to yunho.oh@epfl.ch (in PDF) before noon Monday
Presentations

Not your usual conference presentation

First, a two-minute elevator pitch
• Concise, clear, powerful overview of the paper

Second, present answers to write-up questions
• One slide for each answer
• Provoke discussions

Sign up for presentations/preferred papers now
• Contact Yunho (E-mail or Slack)
Grading/Regrading

Grades (curved)

- Presentations
  - Based on Patterson's ten commandments
- Weekly paper assignments

Check your grades

- Please send us an email to find out where you stand
Technical Roadmap for the Term

- Benchmarks
- Systems & ML
- Federated learning
- Decentralized learning
- Deep learning with low-precision computation
- Training with low-precision gradients
- Distributed Imagenet & Transformers training
- Training with model parallelism
- Neural architecture search
- Domain-specific languages for ML
- ML inference at scale
- Hardware accelerators for deep learning
- Security
This week

How to Give a Bad Talk: Ten Commandments

Good/Bad presentations for Task of the Referee

2nd week:
  • MLPerf training benchmark
  • MLPerf inference benchmark
How to Give a Bad Talk: Ten Commandments
by David Patterson, UC Berkeley
SARCASM

it is a form of art
Thou shalt not be neat

Why waste research time preparing slides?

Ignore spelling, grammar and legibility

Who cares what 50 people think?
Simple spelling examples

Many will be caught/corrected by the software

Others are harder to catch:
• “lead” vs. “led”
• “their” vs. “they’re”
• “where” vs. “were”
Thou shalt not waste space

Historical reference:
• Too much in plastic transparencies/slides

Today:
• PowerPoint files with images are large (this file ~8MB)
• Save space by cramming info into a slide
• If you can cut 30% of your slides, you can GBs!
• Just think of MBs of traffic
1 Consistent, Synchronous Timing of State Meta Data Collection

Use the heartbeat/beacon function (send to / get from devices, platforms...) as a subset publish subscribe e.g., CSG Horizons Fusion. Use the intrinsic millisecond - 90 minute timing function of the heartbeat to enable consistent, synchronized collection of raw state meta data (e.g. location, moving, fault, IP address, origin and organization if BIER/ONE transfer to queues, SAHS, device... prior to data fusion improving filtering / data intel fusion.

Heartbeat Protocol: low level state meta data browsler. Use to conserve bandwidth before handshaking. HP Enterprise implemented a similar mechanism that often uses the heartbeat mechanism.

The Heartbeat / Beacon as subset publish subscribe mechanism

2 Maneuver the Network / Spontaneous Organization

Use state meta data collected during step 1 to enable systems & network management of assets Management Information Base (MIB) supporting Internet and physical state meta data for spontaneous (no organization) splits, join, add, via broadcasts, P2P, multicast... of heartbeat /Heartbeat state meta data (i.e., location, status, moving, etc.).

3 Unified Alerts / Event Trigger

Workflow
- Efficient XML XML DUAL COMMUNICATION...
- child schemas for CVD, profiles, etc.
- specific area encodings, input points
- data tables for special needs
- data files for fast movers

4 Millisecond Transmissions via Beacon Tech

Millisecond exchanges via 30 second screen matrix while increasing / decreasing radius of disaster / event / threat radius represented by multicast zones corresponding to US / EU / Latin American systems.

Instigate National Command Authority HCA check-in - workflows over multicast / anycast / IP using "true code" data to be used, e.g., Toundeds with AgileBible Efficient XML module embedded synchronized across 8 complex systems achieving synchronized, interoperable collaboration based on synchronized event timing and common symbols via the Common Alert Protocol as a universal event / alert trigger.
A point about images/graphs/figures

Do not need full resolution
• Make sure it’s visible on screen (not hazy)

If using info from a paper
• Try your best not to cut and paste from the paper
• Reproduce a simplified version of a graph (not all info that appears in a graph in a paper)
• Do not include captions
Thou shalt not covet brevity

It’s well known that engineers can not write!

Prove otherwise:

• Use complete sentences, never just key words
• If possible, use whole paragraphs and read every word
• At a minimum, make sure your bullets wrap around at least once!
Comment threads and how to disagree in comments

- In general, comment to another person on the higher level of polite and/or where more the Be Nice, Be Res intimidation other people.

- Disagreement and on the page more helpful. In and polite, and as long as the person you are commenting on or Quora and/or don’t know as hostile or disrespectful?

- In multi-comment thread, a person should stop that they are harassing.

Given that we require a show one another. A key goal of rage or
to making the respectful, on the part of I am new to your comment should be no
people, a able impression
Thou shalt cover thy naked slides

Historical reference:
- You needed the suspense!
- Overlays were too flashy

Today: Animating bullets (show one at the time)
- You still need the suspense!
- Animate every word, keep their attention focused
- Don’t let the audience read ahead
Thou shalt not write large

Be humble – use a small font

Important people sit in front (w/ COVID distance preserved)

Who cares about the riff-raff?
Make text large enough to read

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ZS</td>
<td>HC</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>CP1</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>HSKR</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CHK</td>
<td>RVD</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>HNSDCV</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>OKHDNRCS</td>
<td>24</td>
<td>Minimum font size</td>
</tr>
<tr>
<td>7</td>
<td>VHDNKUOSRC</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>NAGULWEDNA</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>BDCLZWVYIHROA</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>AUREYOREYESSOREYET</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>THISISWAYTOOSMALL</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>ATTHISPOINTYOUAREBEINGSILLY</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>FONTSIZEUSEDINHOAQUARTERLYREVIEW</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thou shalt not use color

Flagrant use of color indicates uncareful research

It's also unfair to emphasize some words over others
Keep It Simple (Text)

- Too many colours
- Too Many Fonts and Styles
- The 6 x 6 rule
  - No more than 6 lines per slide
  - No more than 6 words per line
Thou shalt not illustrate

Confucius says “A picture = 10K words”
but Dijkstra says “Pictures are for weak minds”

Who are you going to believe?

Wisdom from the ages or the person who first counted goto's?
Motor Car, any self-propelled vehicle with more than two wheels and a passenger compartment, capable of being steered by the operator for use on roads. The term is used more specifically to denote any such vehicle designed to carry a maximum of seven people.

The primary components of a car are the power plant, the power transmission, the running gear, and the control system. These constitute the chassis, on which the body is mounted. The power plant includes the engine and its fuel, the carburettor, ignition, lubrication, and cooling systems, and the starter motor.
Thou shalt not make eye contact

You should avert eyes to show respect

Blocking screen can also add mystery
Thou shalt not skip slides in a long talk

You prepared the slides; people came for your whole talk; so just talk faster

Skip your summary and conclusions if necessary
Thou shalt not practice

Why waste research time practicing a talk?

It could take several hours out of your two years of research

How can you appear spontaneous if you practice?

If you do practice, argue with any suggestions you get and make sure your talk is longer than the time you have to present it
Conclusions

1. Thou shalt not be neat
2. Thou shalt not waste space
3. Thou shalt not covet brevity
4. Thou shalt cover thy naked slides
5. Thou shalt not write large
6. Thou shalt not use color
7. Thou shalt not illustrate
8. Thou shalt not make eye contact
9. Thou shalt not skip slides in a long talk
10. Thou shalt not practice

Commandment 10 is most important. *Even if you break the other nine, this one can save you.*
The Task of the Referee

Alan Jay Smith
Peer reviewing

Refereeing is a public service, one of the obligations of a scientist
Why should we care?

**Bad Referee**
- Mislead
- Waste time
- Damage careers

**Good Referee**
- Enlighten
- Help progress
- Give credit

This paper: The task of a “good” referee
Task of the “good” referee (1/2)

Decide whether the paper should be published or not

Publishable if it makes *sufficient contribution*
- New interesting research results
- Insightful synthesis of existing results

NOT publishable if
- Repetition of other papers
- Good idea but expressed badly
Task of the “good” referee (2/2)

Write a referee report (review) about the paper

A recommendation for or against publication
  • Provide guidance to authors and editors
  • Justify recommendations with discussion

A list of necessary and recommended changes
  • Improve final version or next submission
... and don’t forget

A referee who says
- always YES, encourages poor research

A referee who says
- always NO, blocks/delays good research from publication
Supplementary Material to Watch: Bengio’s Interview

https://www.youtube.com/watch?time_continue=863&v=JymNsYC3ZPk&feature=emb_title
Answers to write-up questions
What is the problem?

Refereeing is a public service
  • one of a scientist’s obligations

Referees, learn to produce reports w/o formal instruction
  • By practice, feedback from editors, seeing others

→ Problem: How to evaluate a research paper?
What are the key insights?

**Paper evaluation:**
- A paper is publishable if it makes a *sufficient contribution*
- A referee should provide an opinion as to whether the paper makes a *sufficient contribution*

**Referee report:**
- Should be a recommendation for or against publication
- Should provide necessary/recommended changes to paper
What is the solution?

Provide guidance to referees on:
• How to evaluate a paper
• How to write a referee report
What is the takeaway message?

A referee should have a middle ground view
- Insufficiently critical referee encourages poor research
- Overly critical referee blocks/delays good research
Will this paper win the test of time award?

Consider the venue and decide
  • Is it an influential paper?
  • Will the contributions impact the next 10-15 years?

This paper was written more than 25 years ago!
  • We can assume that it got the award

For recent examples check:
  • SIGOPS Hall of Fame
  • SIGCOMM test of time awards
Why should this paper not have appeared in MLSys, NeurIPS, OSDI, ASPLOS?

Is there something wrong with
- Idea?
- Solution?
- Evaluation?
- Methodology?
- Venue?

Name and explain only the most important one
- Remember the task of the “good” referee