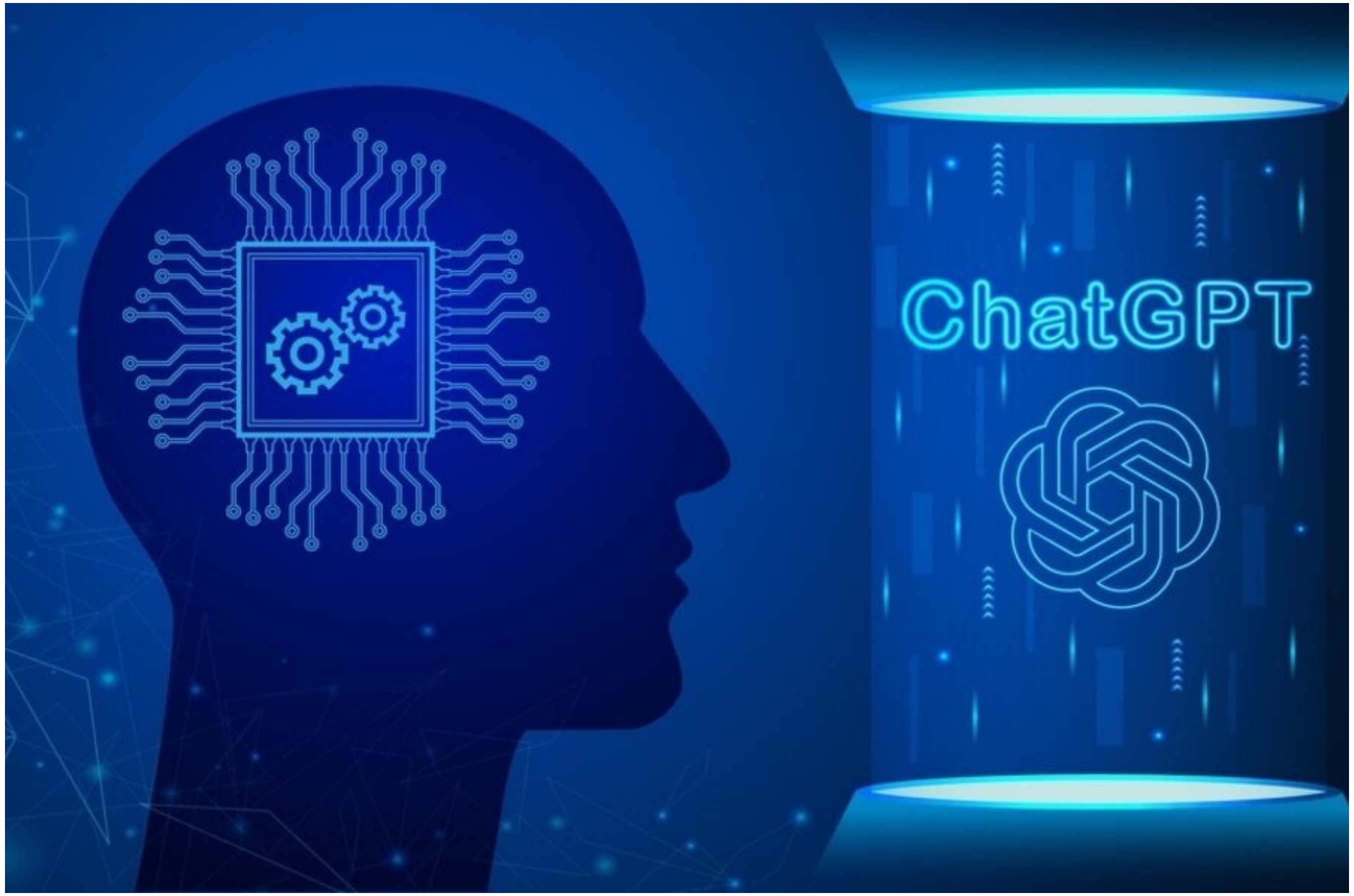
CS 723

Topics on ML Systems

Spring 2023
Babak Falsafi
Martin Jaggi
Anne-Marie Kermarrec



parsa.epfl.ch/course-info/cs723

Where are we?

M	T	W	T	F
➡	21-Feb	22-Feb	23-Feb	24-Feb
27-Feb	28-Feb	1-Mar	2-Mar	3-Mar
6-Mar	7-Mar	8-Mar	9-Mar	10-Mar
13-Mar	14-Mar	15-Mar	16-Mar	17-Mar
20-Mar	21-Mar	22-Mar	23-Mar	24-Mar
27-Mar	28-Mar	29-Mar	30-Mar	31-Mar
3-Apr	4-Apr	5-Apr	6-Apr	7-Apr
10-Apr	11-Apr	12-Apr	13-Apr	14-Apr
17-Apr	18-Apr	19-Apr	20-Apr	21-Apr
24-Apr	25-Apr	26-Apr	27-Apr	28-Apr
1-May	2-May	3-May	4-May	5-May
8-May	9-May	10-May	11-May	12-May
15-May	16-May	17-May	18-May	19-May
22-May	23-May	24-May	25-May	26-May
29-May	30-May	31-May	1-Jun	2-Jun

Class intro

- Logistics
- Grades
- Topical intro

Who should take CS723?

Graduate Students (MS/PhD)

1. Computer system designers
2. ML 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: cs723-2023.slack.com

Logistics for the course

Class times

- Tuesdays 2:15pm-4:00pm

Profs

- Babak Falsafi
- Martin Jaggi
- Anne-Marie Kermarrec

Help

- Martijn de Vos
- Simla Harma

Anne-Marie

Babak

Martin

Martijn

Simla

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)

Roughly two papers per week

This week (only):

- Task of the Referee by Alan Smith

The Task of the Referee

Alan Jay Smith
University of California at Berkeley

There is an endless stream of research papers submitted to conferences, journals, newsletters, anthologies, annuals, trade journals, newspapers, and other periodicals. Many such publications use impartial, external experts to evaluate papers. This approach is often called peer review, and the reviewers are called referees. Refereeing is a public service, one of the professional obligations of a computer science and engineering professional. Unfortunately, referees typically learn to produce referee reports without any formal instruction; they learn by practice, by feedback from editors, by seeing referee reports for their own papers, and by reading referee reports written by others.

Computer researchers have a professional obligation to referee the work of others. This article tells you how to evaluate a paper and write a report using common standards and procedures.

This article tells you how to evaluate a paper, write a referee report, and apply common standards and procedures. It is intended to replace Forscher's rules,[1] which are distributed by some editors but do not reflect the procedures used in computer science and engineering. This article focuses on research papers in applied areas of computer science and engineering, such as systems, architecture, hardware, communications, and performance evaluation, but most of the discussion is generally applicable; separate sections consider research proposals and survey and tutorial papers. Authors might find this material useful for preparing papers for publication. Another recent paper discusses refereeing in theoretical computer sciences; there are some differences between theory and the applied areas considered here.

The referee's task

Your role as referee is to decide whether a paper makes a sufficient contribution to the field. The contribution can be new and interesting research results, a new and insightful synthesis of existing results, a useful survey of or tutorial on a field, or a combination of those types. To quote a referee for this article:

> Small results which are surprising and might spark new research should be published; papers which are mostly repetitions of other papers should not; papers which have good ideas badly expressed should not be published but the authors should be encouraged to rewrite them in a better, more comprehensible fashion.

MS/PhD Students Interested in Empirical Science

Highly recommended reading (not required)

- Strong Inference = Hypothesis+Insight+Test
- Helps you with clarity for effective research
- Saves you a lot of time
- For ML, saves LOTS of emissions for planet
- By John R. Platt, Science, 1964

16 October 1964, Volume 146, Number 3642 **SCIENCE**

Strong Inference

Certain systematic methods of scientific thinking may produce much more rapid progress than others.

John R. Platt

Scientists these days tend to keep up a polite fiction that all science is equal. Except for the work of the misguided opponent whose arguments we happen to be refuting at the time, we speak as though every scientist's field and methods of study are as good as every other scientist's, and perhaps a little better. This keeps us all cordial when it comes to recommending each other for government grants.

But I think anyone who looks at the matter closely will agree that some fields of science are moving forward very much faster than others, perhaps by an order of magnitude, if numbers could be put on such estimates. The discoveries leap from the headlines—and they are real advances in complex and difficult subjects, like molecular biology and high-energy physics. As Alvin Weinberg says (*1*), "Hardly a month goes by without a stunning success in molecular biology being re-

in scientific advance is an intellectual one. These rapidly moving fields are fields where a particular method of doing scientific research is systematically used and taught, an accumulative method of inductive inference that is so effective that I think it should be given the name of "strong inference." I believe it is important to examine this method, its use and history and rationale, and to see whether other groups and individuals might learn to adopt it profitably in their own scientific and intellectual work.

In its separate elements, strong inference is just the simple and old-fashioned method of inductive inference that goes back to Francis Bacon. The steps are familiar to every college student and are practiced, off and on, by every scientist. The difference comes in their systematic application. Strong inference consists of applying the following steps to every problem in sci-

"nature" or the experimental outcome chooses—to go to the right branch or the left; at the next fork, to go left or right; and so on. There are similar branch points in a "conditional computer program," where the next move depends on the result of the last calculation. And there is a "conditional inductive tree" or "logical tree" of this kind written out in detail in many first-year chemistry books, in the table of steps for qualitative analysis of an unknown sample, where the student is led through a real problem of consecutive inference: Add reagent A; if you get a red precipitate, it is subgroup alpha and you filter and add reagent B; if not, you add the other reagent, B′; and so on.

On any new problem, of course, inductive inference is not as simple and certain as deduction, because it involves reaching out into the unknown. Steps 1 and 2 require intellectual inventions, which must be cleverly chosen so that hypothesis, experiment, outcome, and exclusion will be related in a rigorous syllogism; and the question of how to generate such inventions is one which has been extensively discussed elsewhere (*2*, *3*). What the formal schema reminds us to do is to try to make these inventions, to take the next step, to proceed to the next fork, without dawdling or getting tied up in irrelevancies.

It is clear why this makes for rapid and powerful progress. For exploring the unknown, there is no faster method; this is the minimum sequence of

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 martijn.devos@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 Martijn (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

- Generic topics
  - Benchmarks
  - ML Inference at Scale
  - Large Language Models
  - Sustainability
- Platforms
  - Systems & ML
  - Deep learning with Low-Precision Encoding
  - Hardware Accelerators for Deep Learning
  - Sparsity in Deep Neural Networks
  - Domain-Specific Languages for ML
- Federated & Distributed Learning
  - New Training Paradigms
  - Federated Learning
  - Asynchronous Federated Learning
  - Decentralized Learning in Heterogeneous Environments

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

Warning! Sarcasm is a form of art

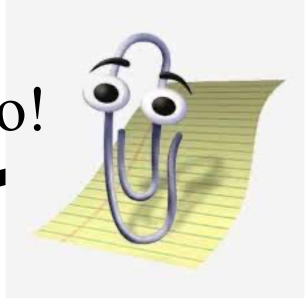
Thou shalt not be neat

Why waste research time preparing slides?

Ignore spelling, grammar and legibility

Who cares what 50 people think?

Simple spelling examples



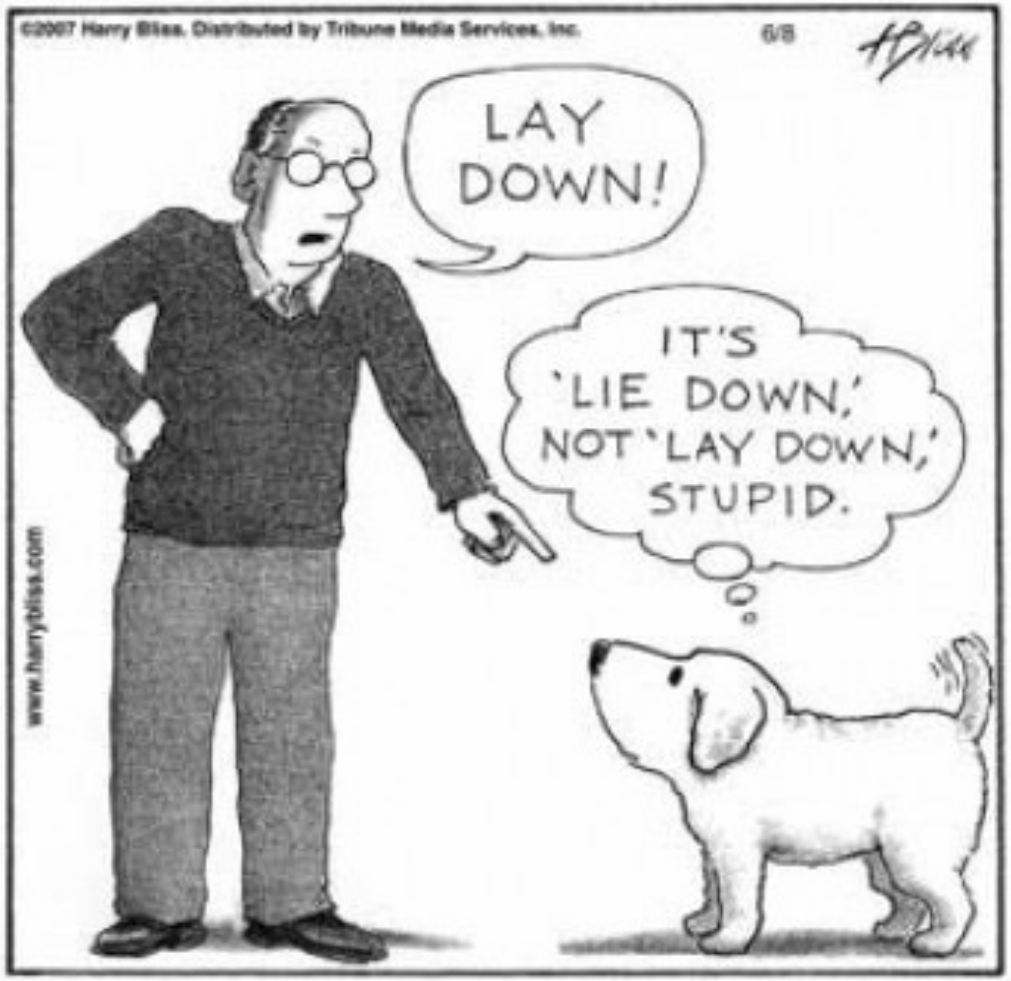
Typo!

Many will be caught/corrected by the saftware

Others are harder to catch:

- “lead” vs. “led” vs. “lead”
- “their” vs. “they’re” vs. “there”
- “where” vs. “were” vs. “wear”
- “tear” vs. “tear” vs. “tear” (“wear” and “pear” only have one pronunciation)

And a cartoon to explain it all!


©2007 Harry Bliss. Distributed by Tribune Media Services, Inc.
6/8
HBliss
LAY DOWN!
IT'S 'LIE DOWN,' NOT 'LAY DOWN,' STUPID.
www.harrybliss.com

Thou shalt not waste space

Historical reference:

- We used plastic transparencies

Today:

- PowerPoint files with images are large (this file ~12.5MB)
- Save space by cramming (six slides of) info into one slide
- People don’t have time to go through a large slide deck

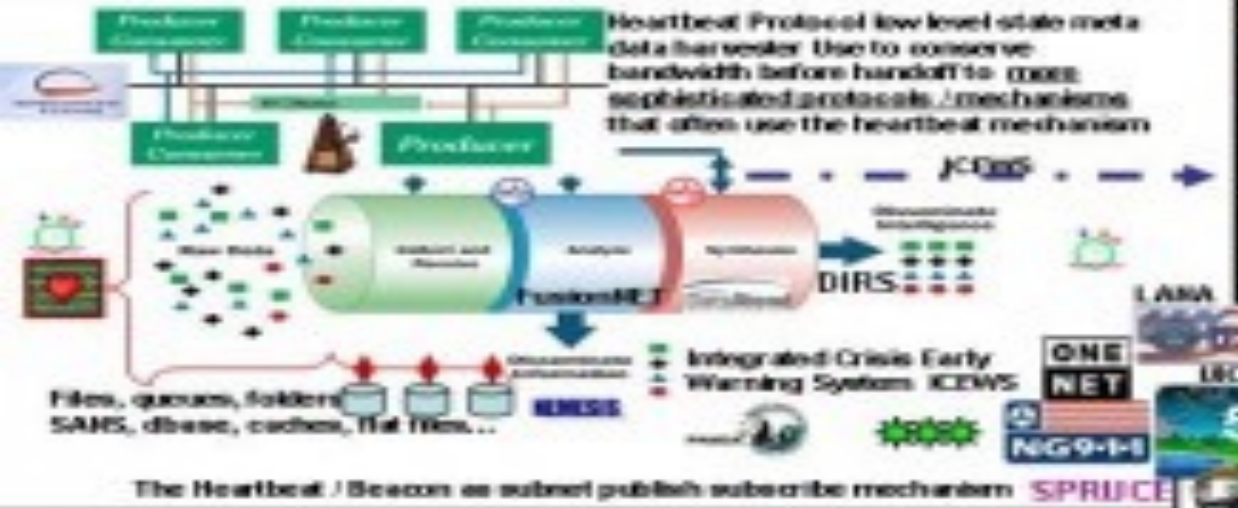
Synchronized Event / Alerts I

Millisecond Exchanges II

COMMON SYMBOLS III

1 Consistent, Synchronous Timing of State Meta Data Collection

Use the heartbeat / beacon function (send to / get from devices, platforms...) as a subnet publish-subscribe e.g., OSD Horizonal Fusion. Use the intrinsic millisecond - 99 minute timing function of the heartbeat to enable consistent, synchronized collection of raw-state meta data (geo location, moving, halt, IP address, unit / organization ID) BEFORE transfer to queues, SANS, dbase... prior to data fusion improving filtering / data intel fusion



Heartbeat Protocol low level state meta data harvester Use to conserve bandwidth before handoff to more sophisticated protocols / mechanisms that often use the heartbeat mechanism

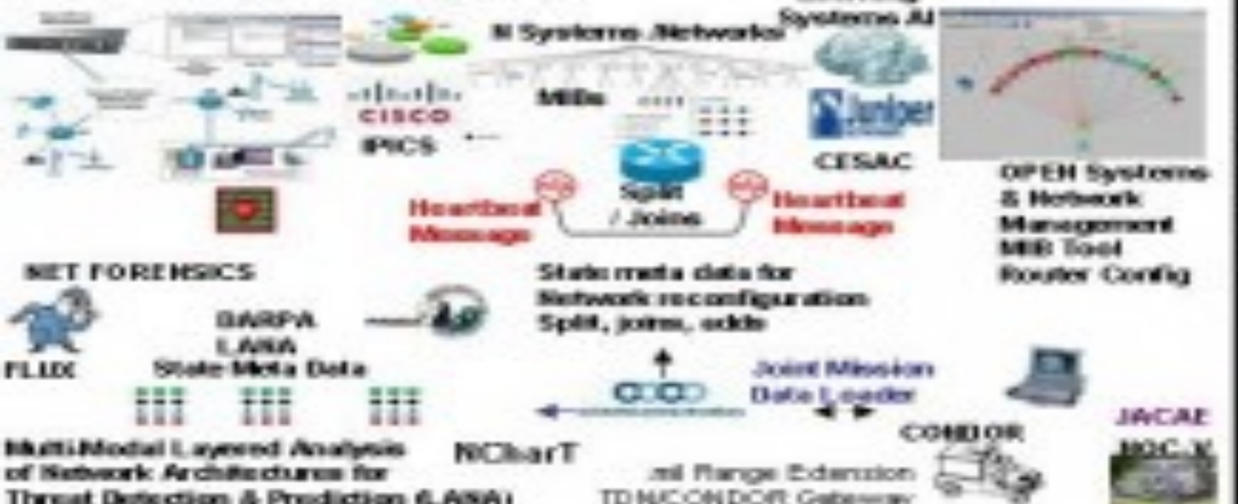
Integrated Crisis Early Warning System ICEWS

Files, queues, folders SANS, dbase, caches, flat files...

The Heartbeat / Beacon as subnet publish subscribe mechanism

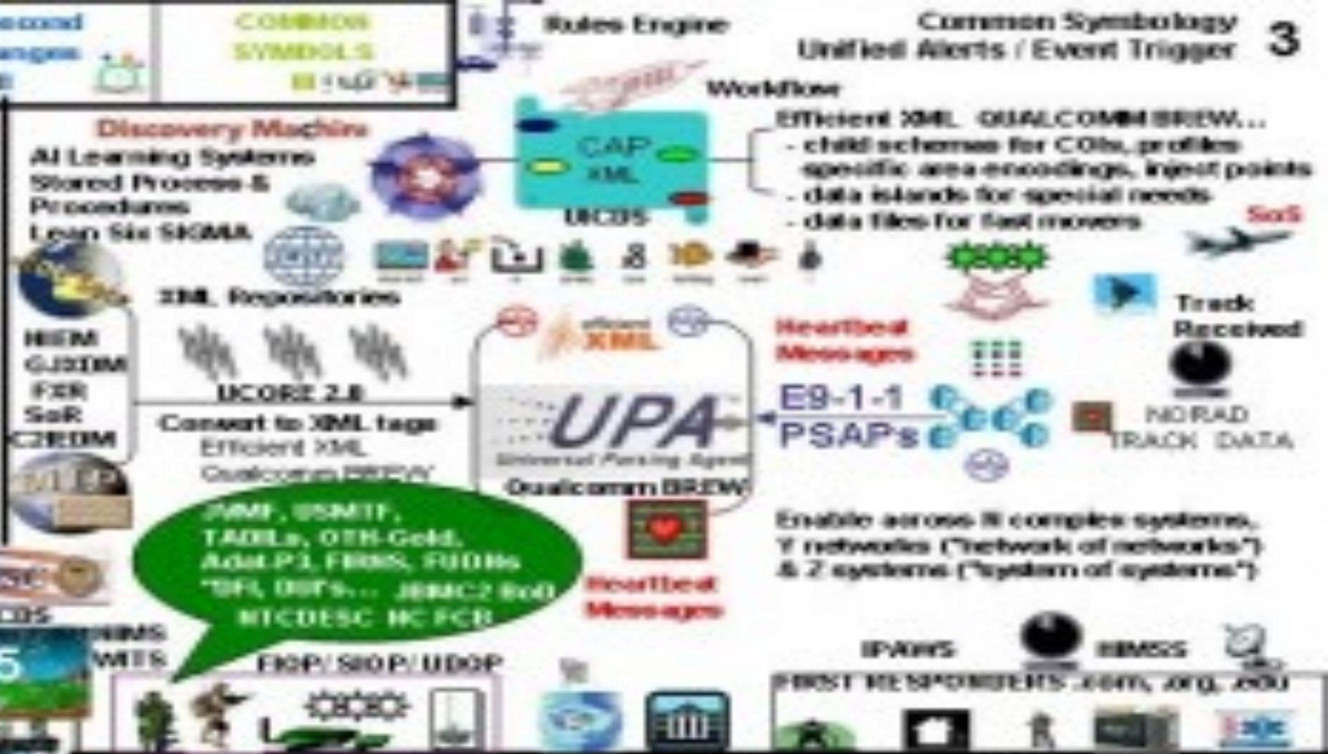
2 "Maneuver the network" "Spontaneous Organization"

Use state meta data collected during step 1 to enable systems & network management of router Management Information Bases MIBs supporting Router/switch MIB dbase files for spontaneous (re) organization-split, join, adds via broadcasts, P2P, multicast.... of heartbeat harvested state meta data i.e., location, status: moving, halt...)



Multi-Modal Layered Analysis of Network Architectures for Threat Detection & Prediction (LANA)

Common Symbology Unified Alerts / Event Trigger 3

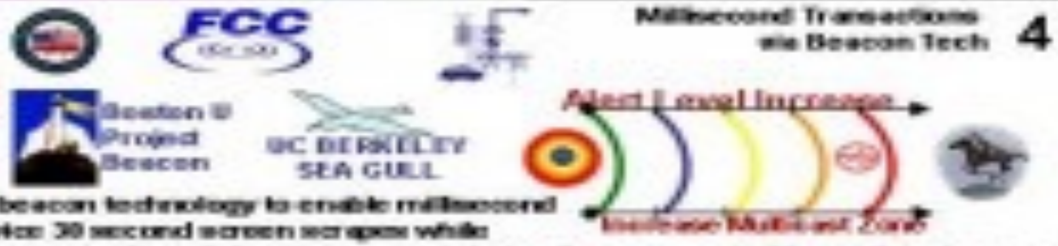


Efficient XML QUALCOMM BREW...

- child schemas for COIs, profiles specific area encodings, inject points
- data islands for special needs
- data files for fast movers

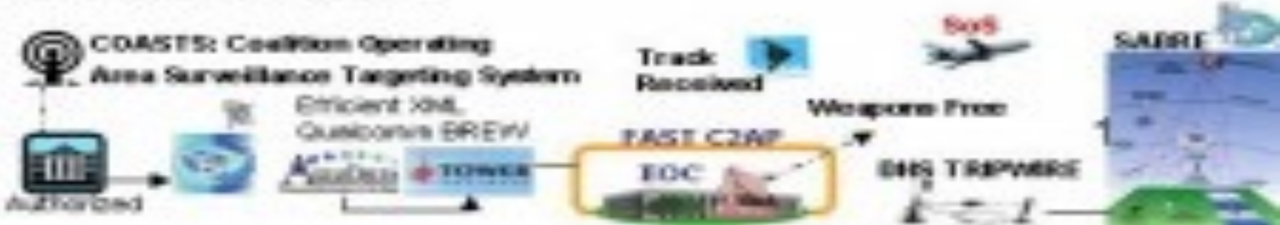
Enable across X complex systems, Y networks ("network of networks") & Z systems ("system of systems")

Millisecond Transactions via Beacon Tech 4



Apply broadcast beacon technology to enable millisecond data exchanges vice 30 second screen scrapes while increasing / decreasing radius of disaster / event / alert radius represented by multicast zones corresponding to US / UK... five level advisory systems

Instantiate National Command Authority NCA chopchain - workflows over multicast / anycast IP using "true cots" tools in use e.g., Towersoft w/AgileDelta Efficient XML module embedded synchronized across N 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

◆ If it’s hazy, the audience can squint

If using info from a paper

◆ Try your best to cut and paste from the paper

◆ Make sure you include captions from paper (e.g., “Figure 12:…”)

Keep all the lines (in graphs) and bars (in barcharts)

◆ You ran 52 experiments

◆ Showcase all results in the graph (why drop info)

Excellent example

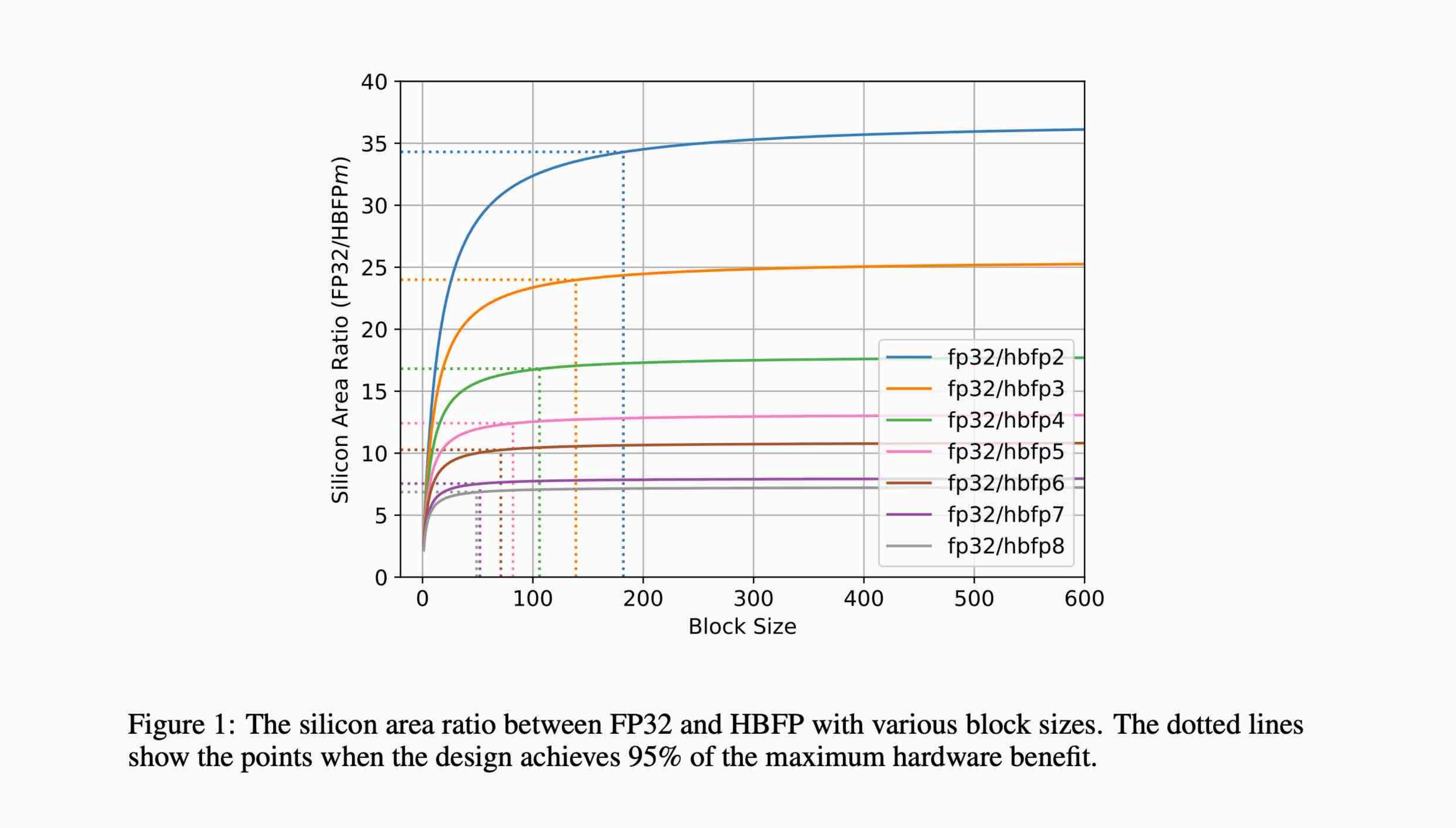


Figure 1: The silicon area ratio between FP32 and HBFP with various block sizes. The dotted lines show the points when the design achieves 95% of the maximum hardware benefit.

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!

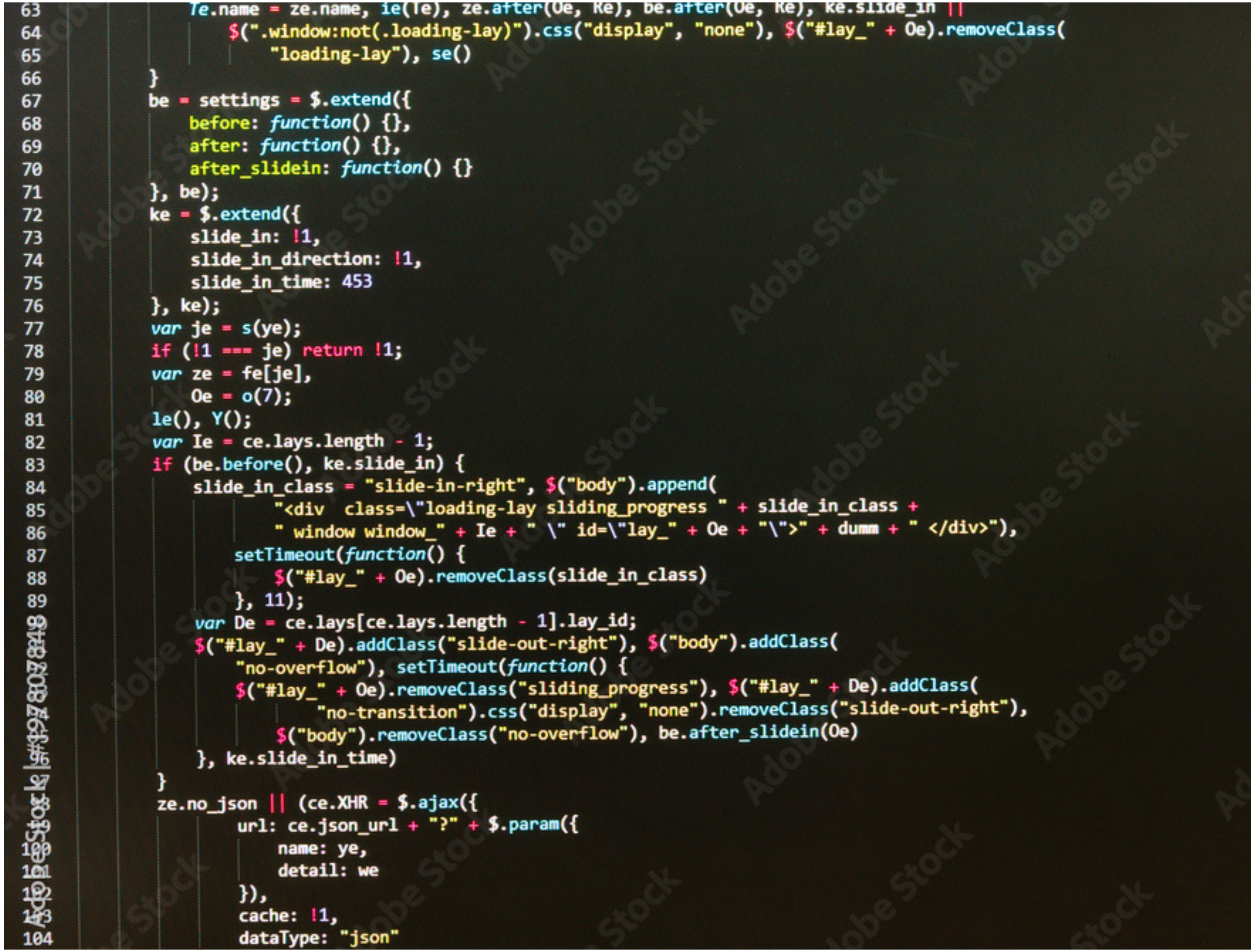
Comment threads and how to disagree in comments

- In general, comment threads on Quora are interactions among strangers. Given that another person on the site may be new to Quora and/or doesn't know you, we require a higher level of politeness than other interactive platforms where users know one another and/or where more adversarial social norms are established and tolerated. A key goal of the Be Nice, Be Respectful policy is ensure that comments do not discourage or intimidate other people on Quora

- Disagreement and debate on Quora is encouraged and is often important to making the page more helpful. It is OK to disagree as long as your comments are civil, respectful, and polite, and as long as you give the impression of assuming good will on the part of the person you are disagreeing with. A good way of framing this test is: "If I am new to Quora and/or don't know you, would it be reasonable for me to perceive your comment as hostile or disrespectful toward me or what I've written?" The answer should be no

- In multi-comment threads where there is significant disagreement among people, a person should stop commenting on the thread before creating the reasonable impression that they are harassing, attacking, and/or bullying another person

A great example with code!



```
    Te.name = ze.name, ie(Te), ze.after(Oe, Ke), be.after(Oe, Ke), ke.slide_in ||
        $(".window:not(.loading-lay)").css("display", "none"), $("#lay_" + Oe).removeClass(
            "loading-lay"), se()
}
be = settings = $.extend({
    before: function() {},
    after: function() {},
    after_slidein: function() {}
}, be);
ke = $.extend({
    slide_in: !1,
    slide_in_direction: !1,
    slide_in_time: 453
}, ke);
var je = s(ye);
if (!1 === je) return !1;
var ze = fe[je],
    Oe = o(7);
le(), Y();
var Ie = ce.lays.length - 1;
if (be.before(), ke.slide_in) {
    slide_in_class = "slide-in-right", $("body").append(
            "<div  class=\"loading-lay sliding_progress " + slide_in_class +
            " window window_" + Ie + " \" id=\"lay_" + Oe + "\">" + dumm + " </div>"),
        setTimeout(function() {
            $("#lay_" + Oe).removeClass(slide_in_class)
        }, 11);
    var De = ce.lays[ce.lays.length - 1].lay_id;
    $("#lay_" + De).addClass("slide-out-right"), $("body").addClass(
        "no-overflow"), setTimeout(function() {
        $("#lay_" + Oe).removeClass("sliding_progress"), $("#lay_" + De).addClass(
                "no-transition").css("display", "none").removeClass("slide-out-right"),
            $("body").removeClass("no-overflow"), be.after_slidein(Oe)
    }, ke.slide_in_time)
}
ze.no_json || (ce.XHR = $.ajax({
        url: ce.json_url + "?" + $.param({
            name: ye,
            detail: we
        }),
        cache: !1,
        dataType: "json"
```

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
  - That way you can also read the words one at a time
- Don’t let the audience read ahead

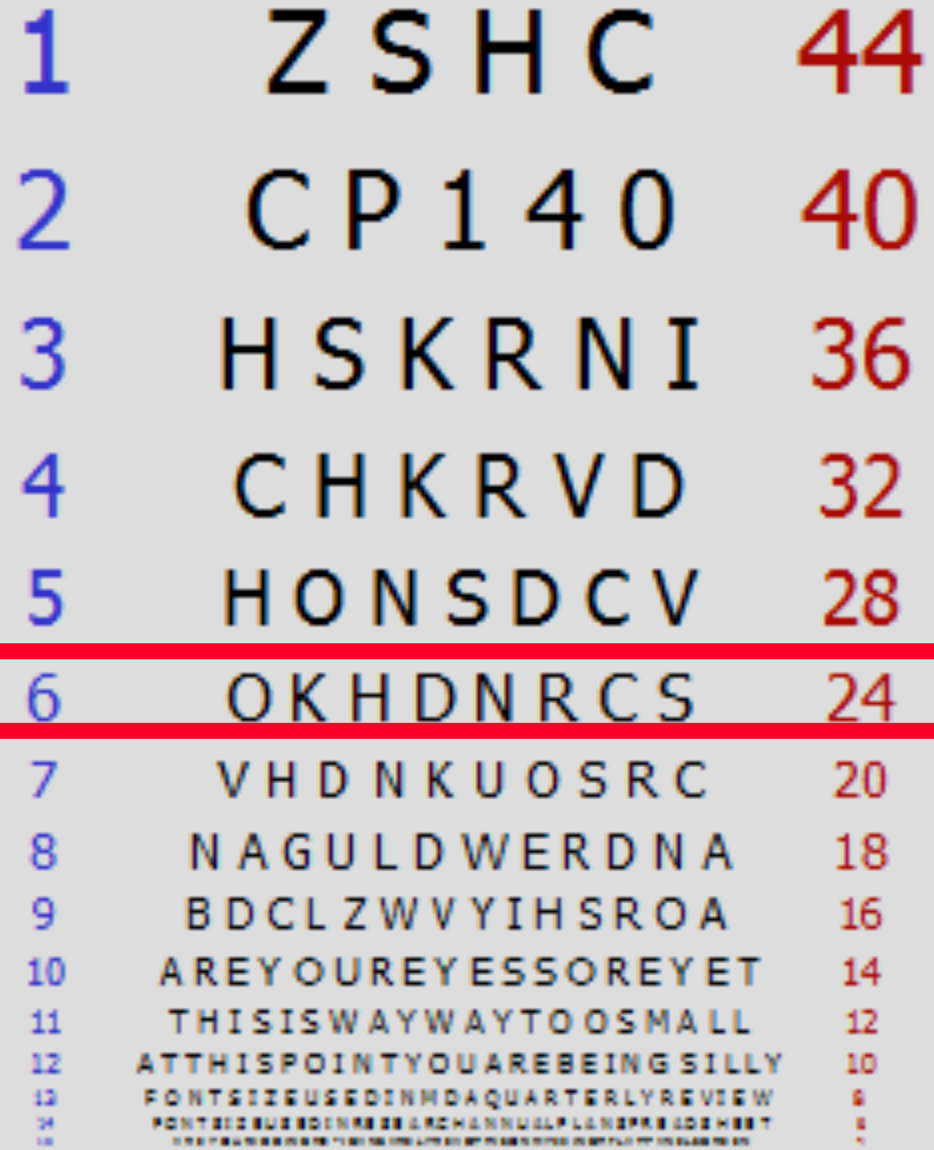
Thou shalt not write large

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



1	Z S H C	44
2	C P 1 4 0	40
3	H S K R N I	36
4	C H K R V D	32
5	H O N S D C V	28
6	O K H D N R C S	24
7	V H D N K U O S R C	20
8	N A G U L D W E R D N A	18
9	B D C L Z W V Y I H S R O A	16
10	AREYOUREYESSOREYET	14
11	THISISWAYWAYTOOSMALL	12
12	ATTHISPOINTYOUAREBEINGSILLY	10
13	FONTSIZEUSEDINMDAQUARTERLYREVIEW	8
14	FONTSIZEUSEDINRESEARCHANNUALPLANSPREADSHEET	6
[illegible]	[illegible]	[illegible]

Minimum font size

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

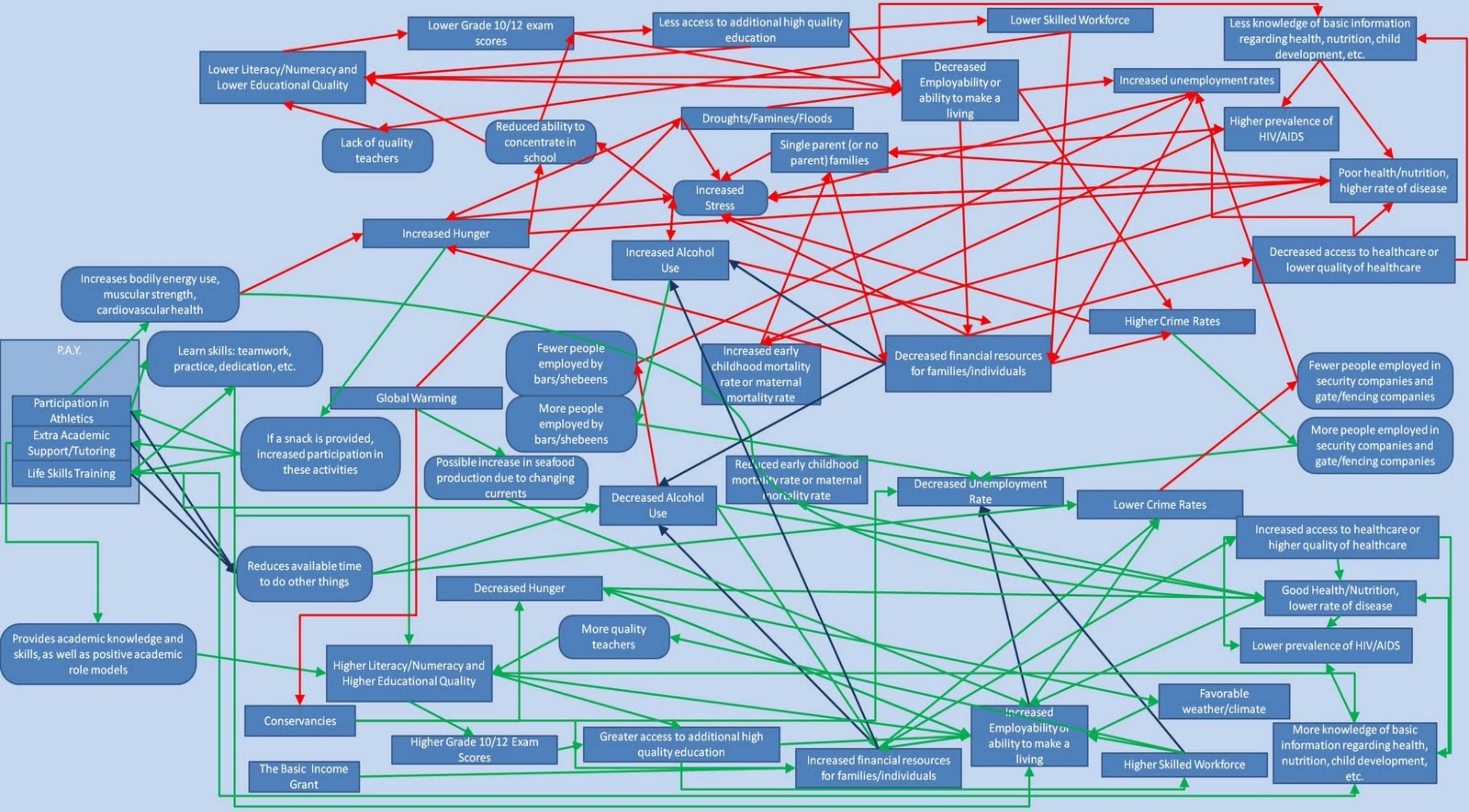
Thou shalt not illustrate

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?


Lower Grade 10/12 exam scores
Less access to additional high quality education
Lower Skilled Workforce
Less knowledge of basic information regarding health, nutrition, child development, etc.
Lower Literacy/Numeracy and Lower Educational Quality
Decreased Employability or ability to make a living
Increased unemployment rates
Droughts/Famines/Floods
Lack of quality teachers
Reduced ability to concentrate in school
Higher prevalence of HIV/AIDS
Single parent (or no parent) families
Poor health/nutrition, higher rate of disease
Increased Stress
Increased Hunger
Increased Alcohol Use
Decreased access to healthcare or lower quality of healthcare
Increases bodily energy use, muscular strength, cardiovascular health
Higher Crime Rates
P.A.Y.
Learn skills: teamwork, practice, dedication, etc.
Fewer people employed by bars/shebeens
Increased early childhood mortality rate or maternal mortality rate
Decreased financial resources for families/individuals
Fewer people employed in security companies and gate/fencing companies
Participation in Athletics
Global Warming
More people employed by bars/shebeens
More people employed in security companies and gate/fencing companies
Extra Academic Support/Tutoring
If a snack is provided, increased participation in these activities
Life Skills Training
Possible increase in seafood production due to changing currents
Reduced early childhood mortality rate or maternal mortality rate
Decreased Unemployment Rate
Decreased Alcohol Use
Lower Crime Rates
Increased access to healthcare or higher quality of healthcare
Reduces available time to do other things
Decreased Hunger
Good Health/Nutrition, lower rate of disease
More quality teachers
Lower prevalence of HIV/AIDS
Provides academic knowledge and skills, as well as positive academic role models
Higher Literacy/Numeracy and Higher Educational Quality
Favorable weather/climate
Conservancies
Increased Employability or ability to make a living
Higher Grade 10/12 Exam Scores
Greater access to additional high quality education
Higher Skilled Workforce
More knowledge of basic information regarding health, nutrition, child development, etc.
The Basic Income Grant
Increased financial resources for families/individuals

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; 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 care?

Bad Referee

Mislead

Waste time

Damage careers

Good Referee

Enlighten

Help progress

Give credit

Our reading: 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

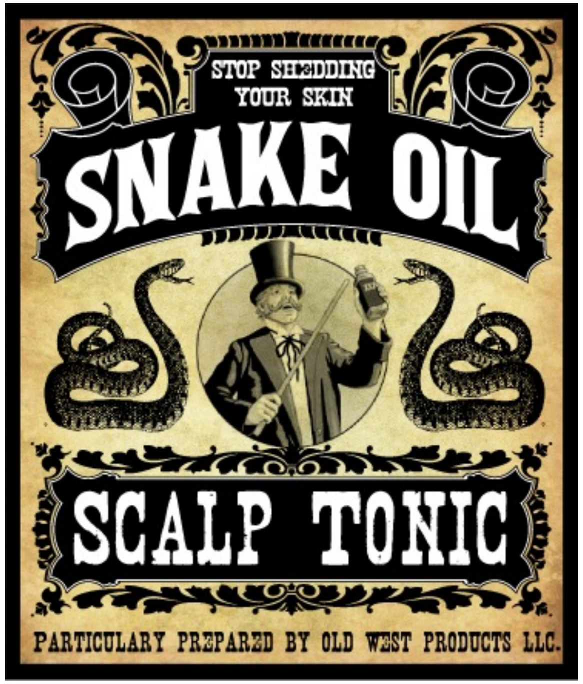
A list of necessary and recommended changes

- Improve final version or next submission

… and don't forget

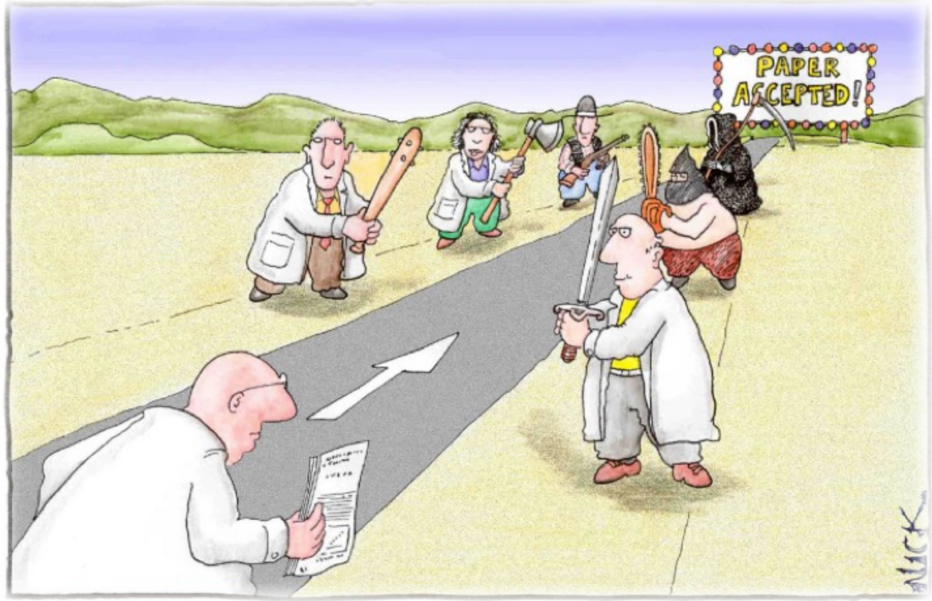
A referee who always says YES

→ encourages poor research



A referee who always says 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

- A scientist’s obligation

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