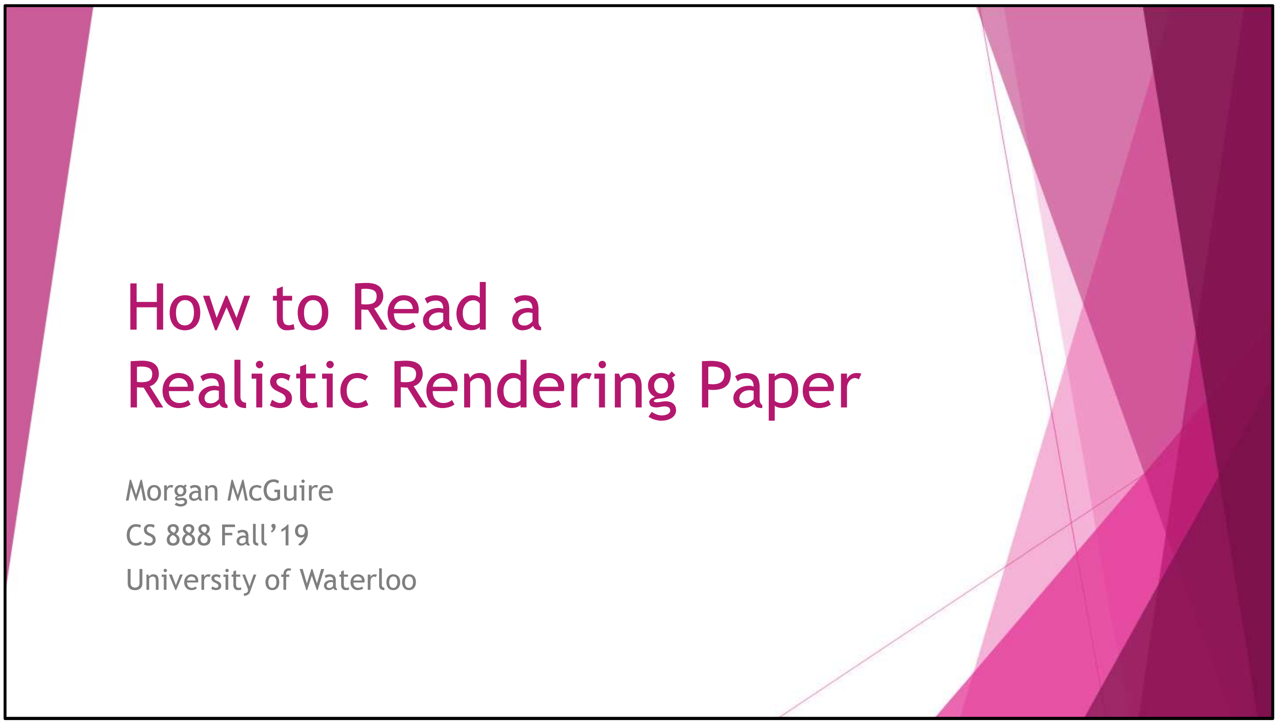
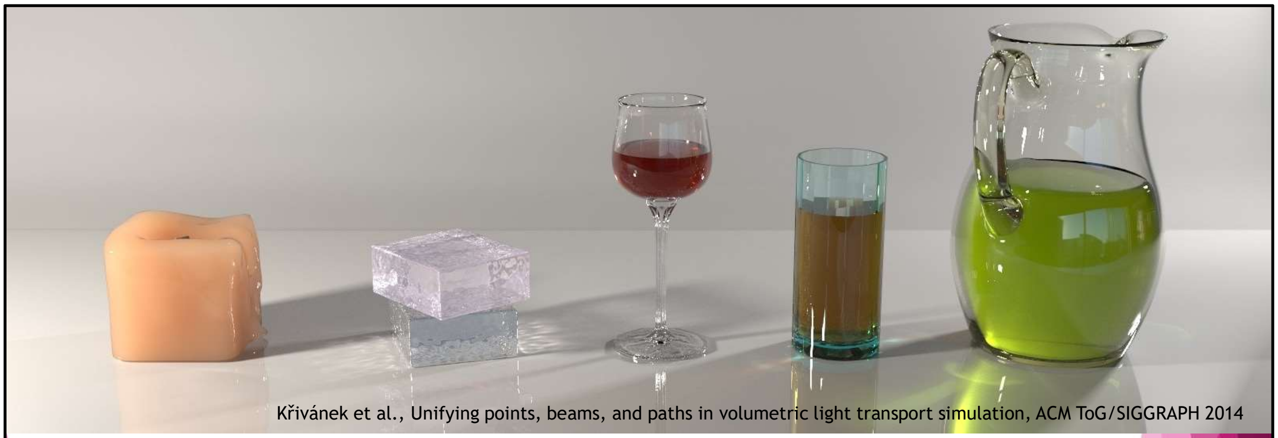


How to Read a Realistic Rendering Paper

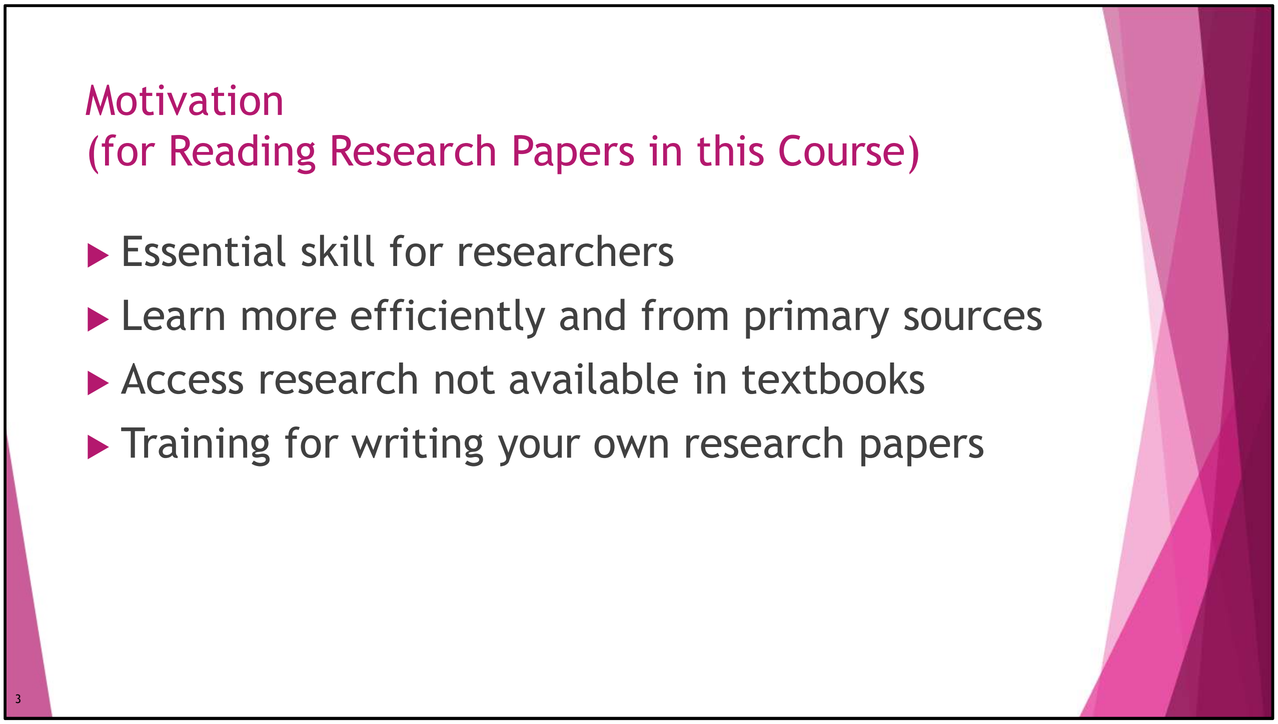
Morgan McGuire
CS 888 Fall’19
University of Waterloo

[These are lecture notes from a graduate course, which were prepared with feedback from co-teachers Stephen Mann and Craig Kaplan and colleagues Peter Shirley, Eric Haines, and Ward Lopes]



Křivánek et al., Unifying points, beams, and paths in volumetric light transport simulation, ACM ToG/SIGGRAPH 2014

Overview



Motivation
(for Reading Research Papers in this Course)

- Essential skill for researchers
- Learn more efficiently and from primary sources
- Access research not available in textbooks
- Training for writing your own research papers

3

A major part of this course is teaching you how to learn from research papers and evaluate them critically.

There are several reasons:

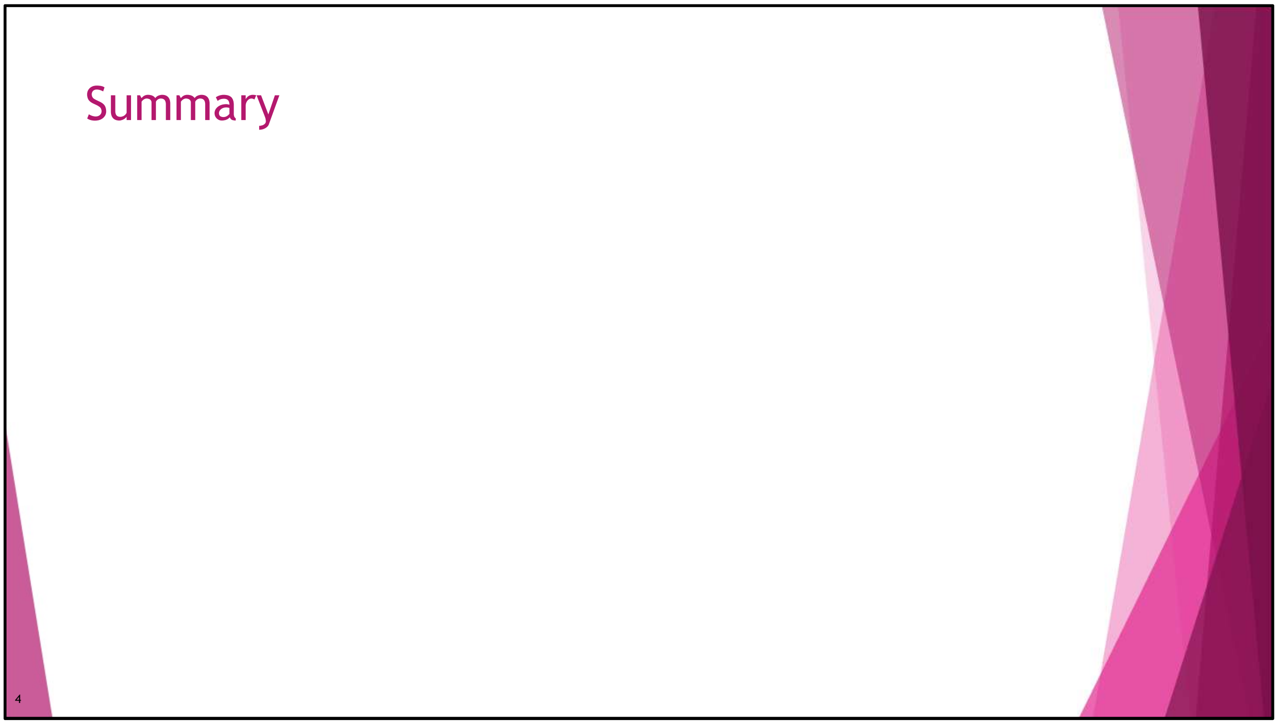
Most important, it is expected that any scientist or senior graduate student is able to read research papers and will do so continually (without them being assigned in a course) to stay abreast of their field. Reading them from start to finish is not a very efficient way of understanding the material, so you need to learn the skill of reading them more strategically.

Learn more efficiently and from primary sources, without the reinterpretation and shifted emphasis of textbooks.

This will give you access to work that is cutting edge, good ideas lost from the main stream, and in other fields. None of those are well-represented in textbooks. (For better and for worse: it is a service that textbooks remove outdated ideas and disproven results; but if you want to master an area, you may need to know that history and be able to perform pruning for yourself on newer work)

It will also train you to perform research, write that work up for publication, and referee and edit the work of others.

With regards to *performing* research, the “story” told in a paper of the authors’ motivation, path to discovering a solution, and conclusions is often not how it really happened…but is often how it should have happened in an ideal world and a good model to aspire to.



Summary

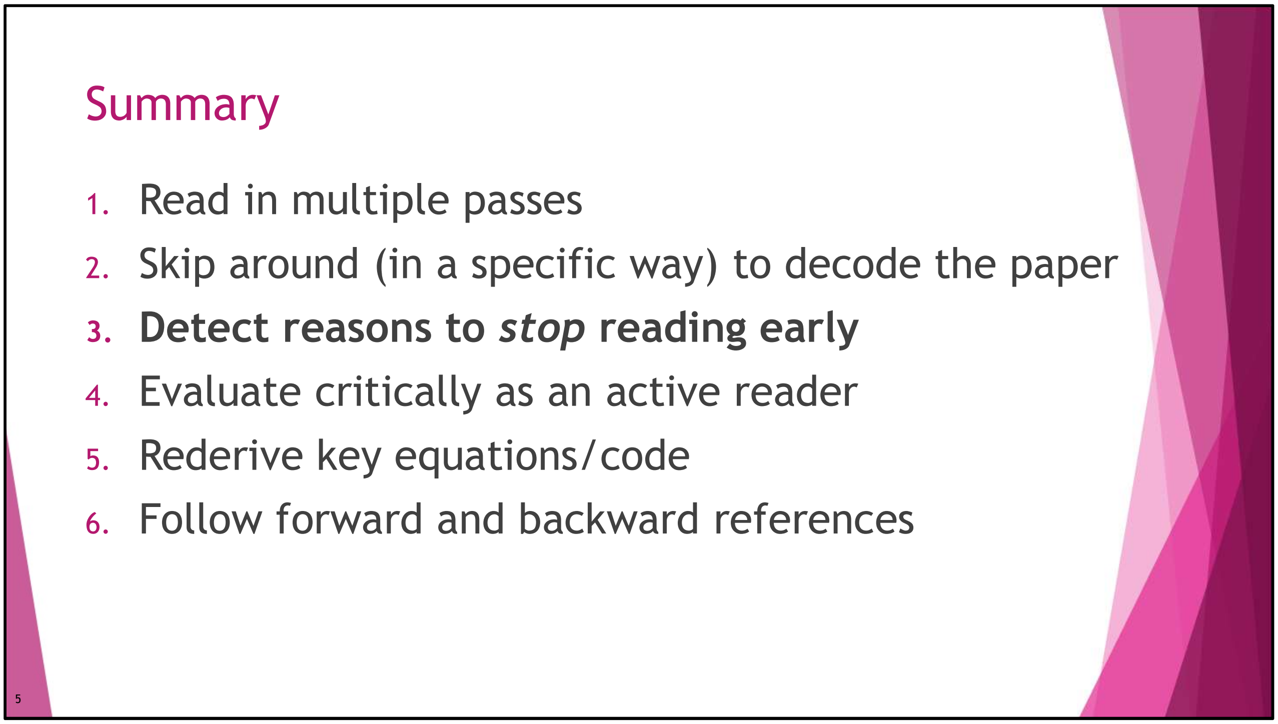
Most of today’s presentation is context for understanding research papers. I’ll address the topic of how to read them relatively quickly at the end once you’re familiar with the parts of papers. Here’s the punchline for that part.

The goal of a textbook is to help you learn something deeply and with context.

Papers are not textbooks. They are highly constrained and thus rely on significant reader knowledge to interpret.
They also serve more audiences than yourself. In fact, teaching you is typically *not* among the main goals of the paper.

So, you have to work much harder to learn a topic from a paper than a textbook, and need specific skills. Those skills will make you an active reader, and a bit of a research paper archeologist, following this process:

…



Summary

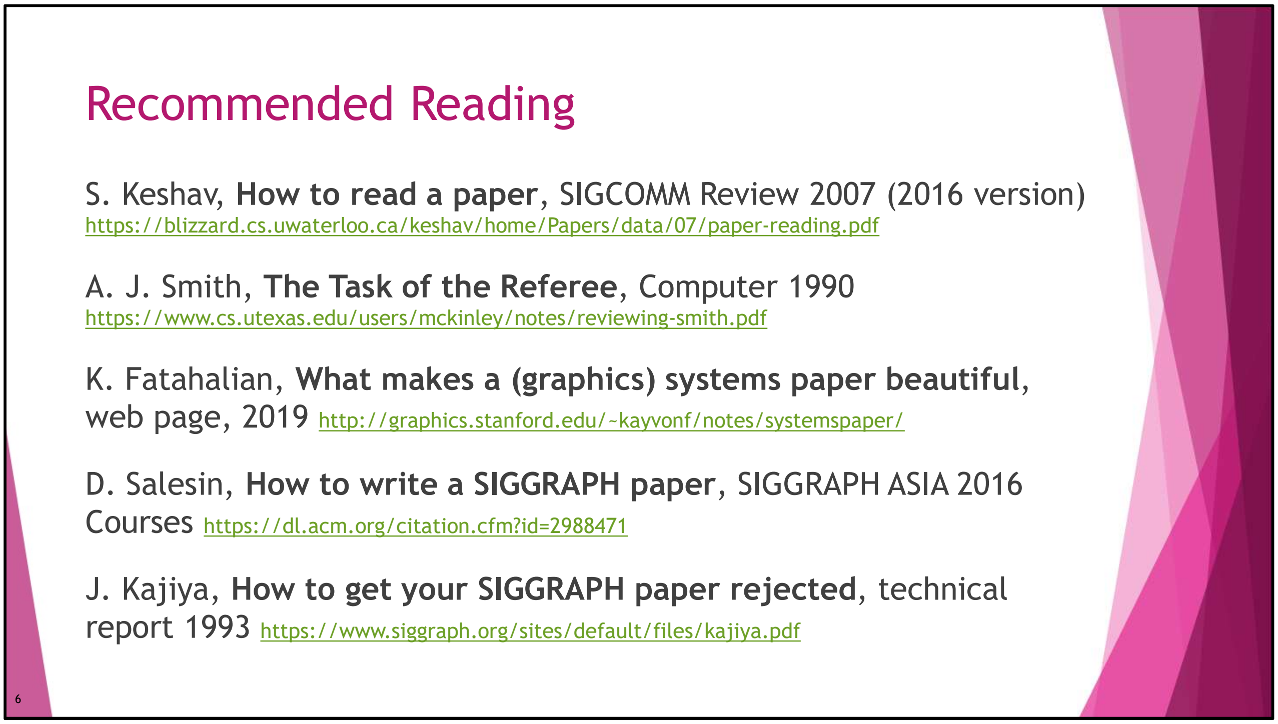
1. Read in multiple passes
2. Skip around (in a specific way) to decode the paper
3. **Detect reasons to *stop* reading early**
4. Evaluate critically as an active reader
5. Rederive key equations/code
6. Follow forward and backward references

5

Note that one reason you don’t read research papers straight through from start to finish is that you’re usually trying to figure out if this is even the right paper to read!

It takes a long time to unpack a paper and fully understand it. If what you’re looking for isn’t there, you’d like to know that as soon as possible so that you can invest that time reading a *different* paper. So, we read in multiple passes, looking for specific kinds of information, and continuously evaluate if we have extracted enough information from the work.

You might also be able to get the level of understanding that you need after five or ten minutes. Maybe you’re not trying to deeply understand the result from a particular paper, but just get context for how another paper differs from it. You can probably get what you need without reading 80% of the text in that case.



Recommended Reading

S. Keshav, **How to read a paper**, SIGCOMM Review 2007 (2016 version)
https://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.pdf

A. J. Smith, **The Task of the Referee**, Computer 1990
https://www.cs.utexas.edu/users/mckinley/notes/reviewing-smith.pdf

K. Fatahalian, **What makes a (graphics) systems paper beautiful**, web page, 2019 http://graphics.stanford.edu/~kayvonf/notes/systemspaper/

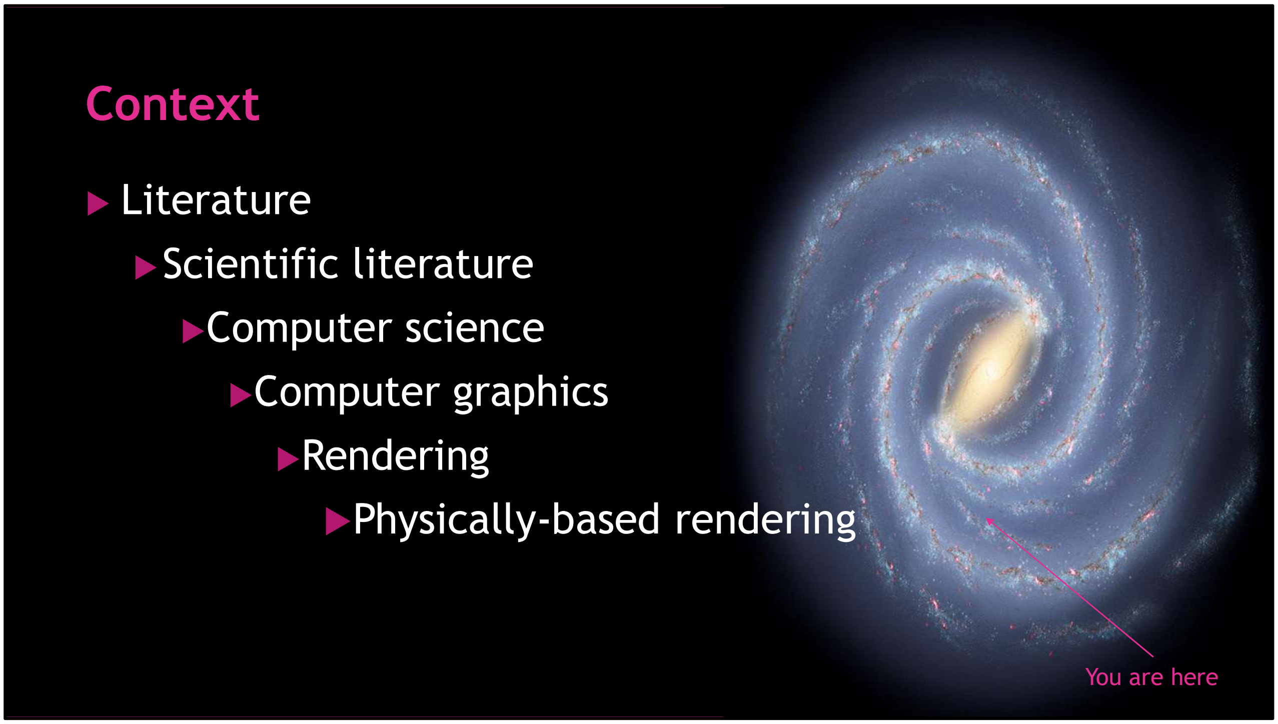
D. Salesin, **How to write a SIGGRAPH paper**, SIGGRAPH ASIA 2016 Courses https://dl.acm.org/citation.cfm?id=2988471

J. Kajiya, **How to get your SIGGRAPH paper rejected**, technical report 1993 https://www.siggraph.org/sites/default/files/kajiya.pdf

6

There are many good articles, primarily by other faculty and written for courses such as this, available on the web discussing how to read research papers.

I recommend also reading articles about how to *write* and *review* a research paper, which will give insight into the structure and motivation for sections of a paper, as well as train you for participating in the peer review process.



Context

- Literature
  - Scientific literature
    - Computer science
      - Computer graphics
        - Rendering
          - Physically-based rendering

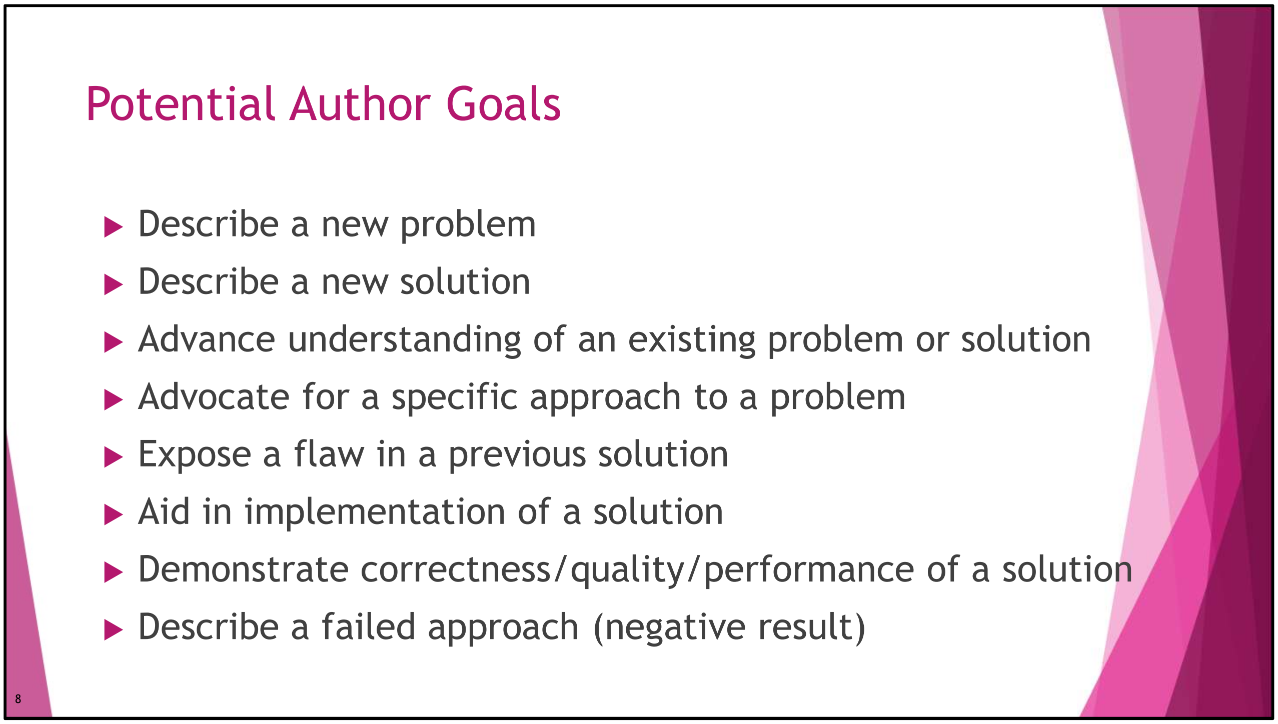
You are here

The advice I’m presenting is tailored for the material that you’ll be reading in this course on physically-based rendering via advanced ray tracing techniques.

A lot of what we’ll discuss today generalizes, and I’ll try to specifically note cases that don’t generalize

But you should be aware that we’re down here in a specific tiny region of a vast hierarchy of written material.

You should also be aware that a ”modern” graphics paper written after the mid 1990s is a little different than a “classic” paper written in the 70s or 80s because the field and formats have matured. I’m going to focus today on modern papers even though we will also read some classic papers in this course.

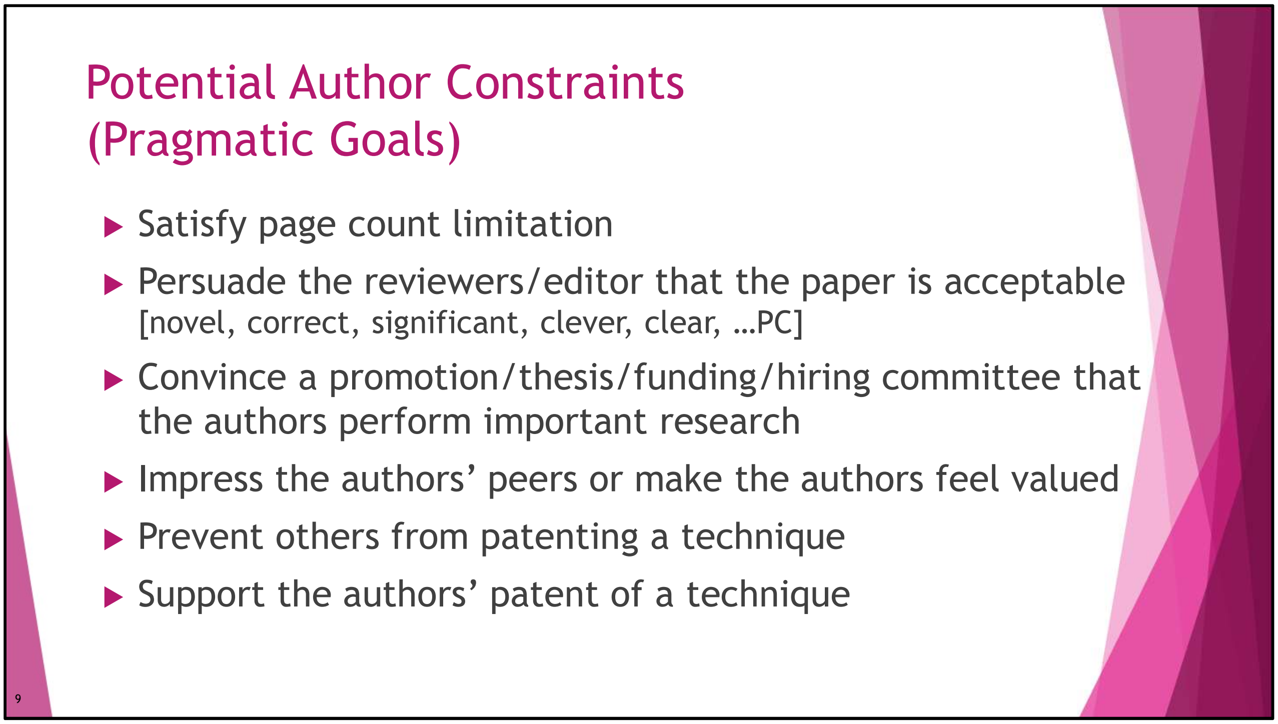


Potential Author Goals

- Describe a new problem
- Describe a new solution
- Advance understanding of an existing problem or solution
- Advocate for a specific approach to a problem
- Expose a flaw in a previous solution
- Aid in implementation of a solution
- Demonstrate correctness/quality/performance of a solution
- Describe a failed approach (negative result)

8

There are many reasons to write a paper, and authors have many (and sometimes conflicting) goals within the paper. The main reasons that someone “should” write a paper are on this page. It helps you when reading a paper to determine what the goal of the paper is.



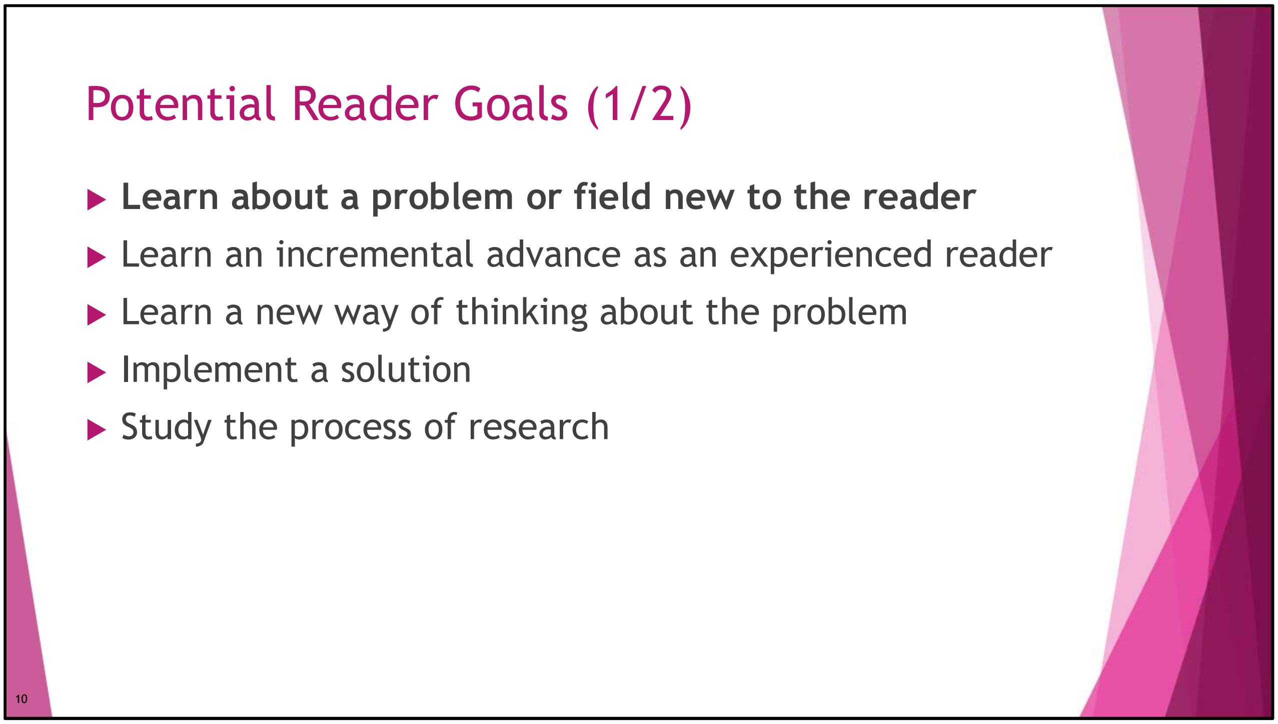
Potential Author Constraints (Pragmatic Goals)

- Satisfy page count limitation
- Persuade the reviewers/editor that the paper is acceptable [novel, correct, significant, clever, clear, …PC]
- Convince a promotion/thesis/funding/hiring committee that the authors perform important research
- Impress the authors' peers or make the authors feel valued
- Prevent others from patenting a technique
- Support the authors' patent of a technique

9

You need to appreciate that the authors also have some less idealized goals, and realize that some of the paper is structured in a way that does not help you especially because a goal other than your enlightenment was in play. You shouldn't get too cynical about this—almost all research papers really are motivated by the goal of advancing the field and finding truth.

But if there's a citation that doesn't seem relevant, an overinflated claim, a really telegraphic paragraph, or exotic terminology and notation, don't feel like you've failed or expend too much effort. That might be an artifact of the pressures the author was under instead of a deep technical insight you're missing. These constraints are often lifted for *presentations* and blog posts, so check if the same author has discussed this material in a different format elsewhere.



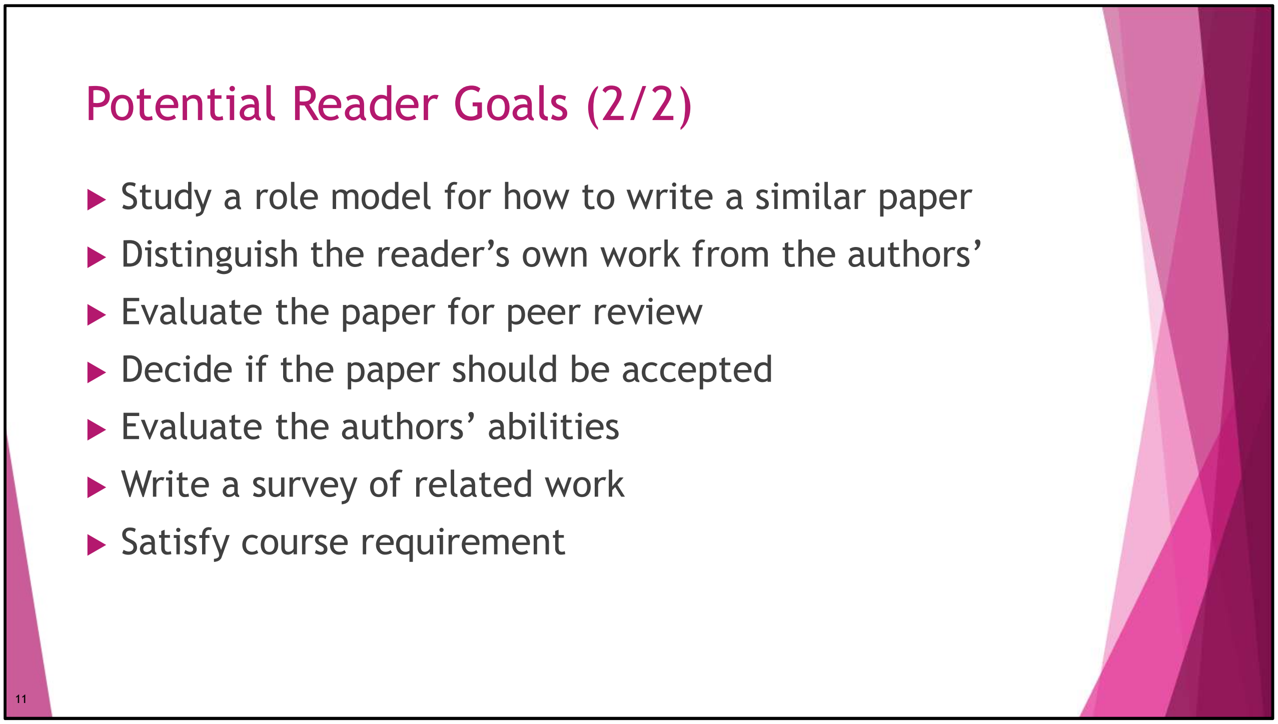
Potential Reader Goals (1/2)

- **Learn about a problem or field new to the reader**
- Learn an incremental advance as an experienced reader
- Learn a new way of thinking about the problem
- Implement a solution
- Study the process of research

10

There are a lot of reasons someone might read a research paper.

Your goal in this course will primarily be the first one on this slide, although all of the ones on this slide will all matter at various points.



Potential Reader Goals (2/2)

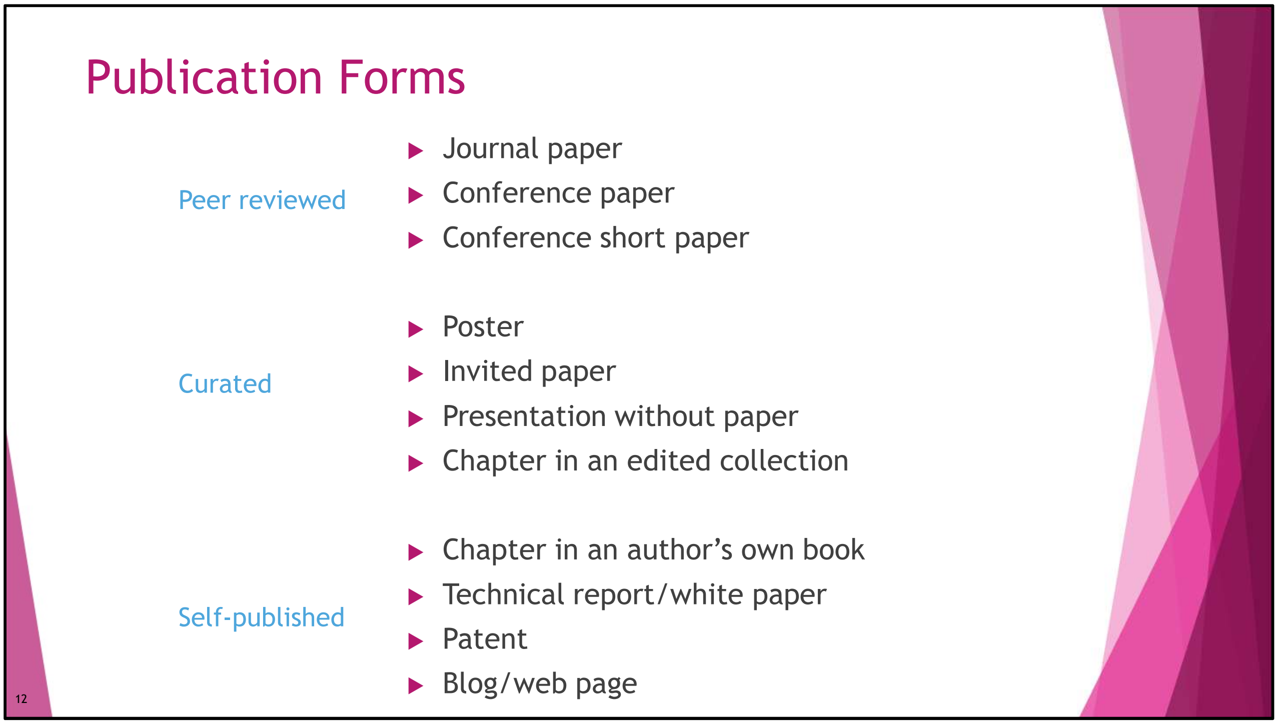
- Study a role model for how to write a similar paper
- Distinguish the reader’s own work from the authors’
- Evaluate the paper for peer review
- Decide if the paper should be accepted
- Evaluate the authors’ abilities
- Write a survey of related work
- Satisfy course requirement

11

There are other reasons that people read papers.
I hope you never read anything solely because I assigned it, without the intention of actually learning from the work.

But I do expect that you’ll read many papers for reasons beyond learning about the topics, because you should. As you reach senior years as a graduate student, or if you continue in the field, you will be asked to review and edit papers.

When writing a paper, I often look to my favorite papers in computer science as role models for a good tone and structure.



Publication Forms

Peer reviewed

- Journal paper
- Conference paper
- Conference short paper

Curated

- Poster
- Invited paper
- Presentation without paper
- Chapter in an edited collection

Self-published

- Chapter in an author's own book
- Technical report/white paper
- Patent
- Blog/web page

12

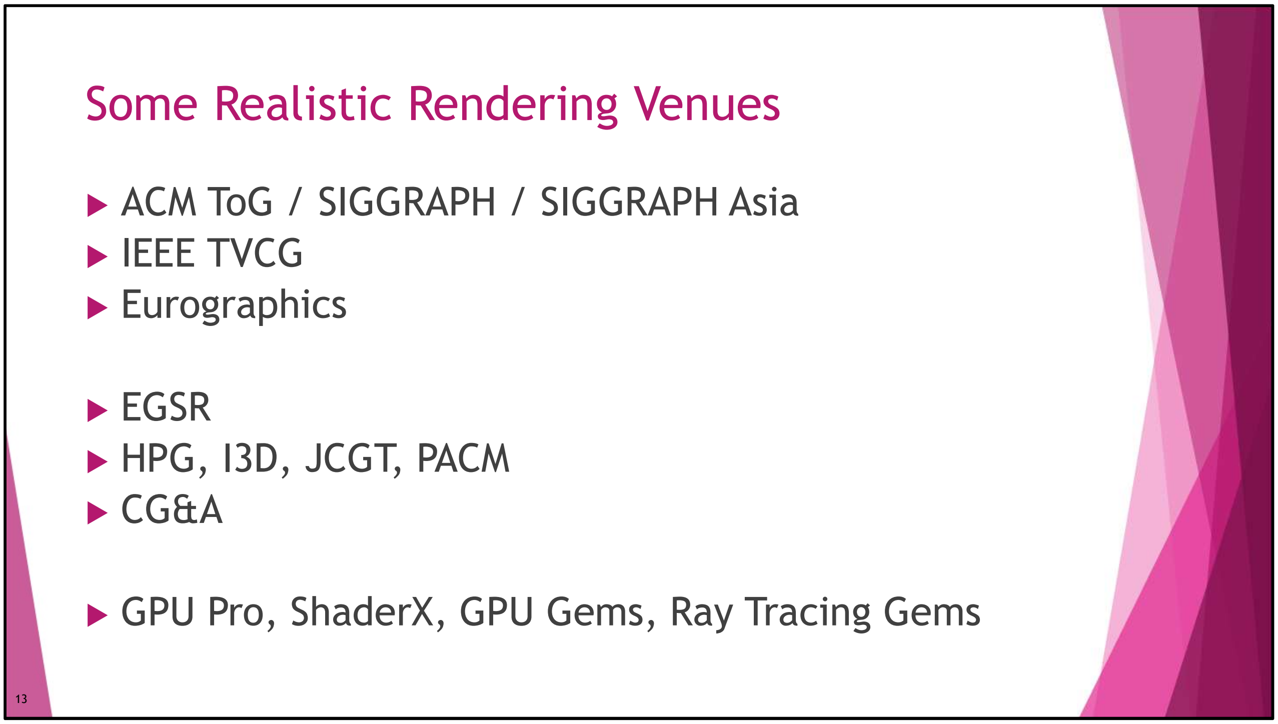
Three major categories, based on peer review.
Fully **peer reviewed** work has been scrutinized by typically 3-5 experts in the area. They've checked it for correctness, writing quality, and reproducibility. The quality of this checking varies depending on the standards and process of the venue and the peers—they could be 20 year veterans, students reviewing their first paper, someone going line by line through one paper, or someone with 30 other papers to review that week.

Peer review isn't an assurance of quality, but it indicates that the scientific community accepts this work as worth reading and adhering to current standards. This tends to force work a bit towards the mean—the very best work can be blunted by this process, as well as less-useful texts being improved or outright excluded by it.

"**Curated**" work is "**Lightly reviewed**". The subject matter, author, and outline of the material have been approved by peers (maybe a single editor or chairperson), but the actual content has not been verified in most cases. Frequently, only a proposal was seen by the reviewer and not the final version. The largest quality signal here is the reputations of the venue and the authors. Expect higher variation in correctness and writing standard than peer review…

In rendering, game developers and some offline rendering experts working in the industry prefer publishing in these forms because it is more efficient from the authors’ perspective. Many influential works appeared as book chapters and Game Developers Conference talks instead of peer reviewed papers.

Self-published work is unreviewed. It has the highest variance. This is a very good way for authors to present new results and readers to stay on top of a fast-moving field, where best practices can change every few months. It is also important for a field driven by applied and industry research, where authors/inventors may not have institutional support for participating in the peer-review process. When One is fully dependent on the authors’ reputation. Again, some very influential rendering work has appeared in these forms, such as the PBRT book and Stephen Hill’s blog, which won Academy and SIGGRAPH awards.



Some Realistic Rendering Venues

- ACM ToG / SIGGRAPH / SIGGRAPH Asia
- IEEE TVCG
- Eurographics

- EGSR
- HPG, I3D, JCGT, PACM
- CG&A

- GPU Pro, ShaderX, GPU Gems, Ray Tracing Gems

13

You can tell a lot about a paper and whether it will have what you're looking for just by where it was published.

The big venues favor academic research; they are always impressive and good science but have varying levels of short-term applicability and are often a bit conservative.

ACM Transactions on Graphics, "ToG", is a journal that has become essentially the same as the SIGGRAPH annual conference and its newer SIGGRAPH Asia counterpart for your purposes: These all feature big results, usually with polished exposition and figures and many motivating examples. You'll notice that most of the papers for this course come from these venues.

Transactions on Visualization and Computer Graphics is essentially IEEE's equivalent of ACM ToG.

The Eurographics annual conference is another big venue.

Smaller conferences tend towards more risky and aggressive ideas, as well as ones

closer to industry practice. There’s a lot more variance in significance and presentation quality, but many of the most important ideas—especially for *real time* rendering– appeared first in small conferences instead of SIGGRAPH.
These are often good papers to reproduce as student projects.

The Eurographics Symposium on Rendering (EGSR) favors more theoretical work in rendering. EGSR papers also usually emphasize offline rendering.

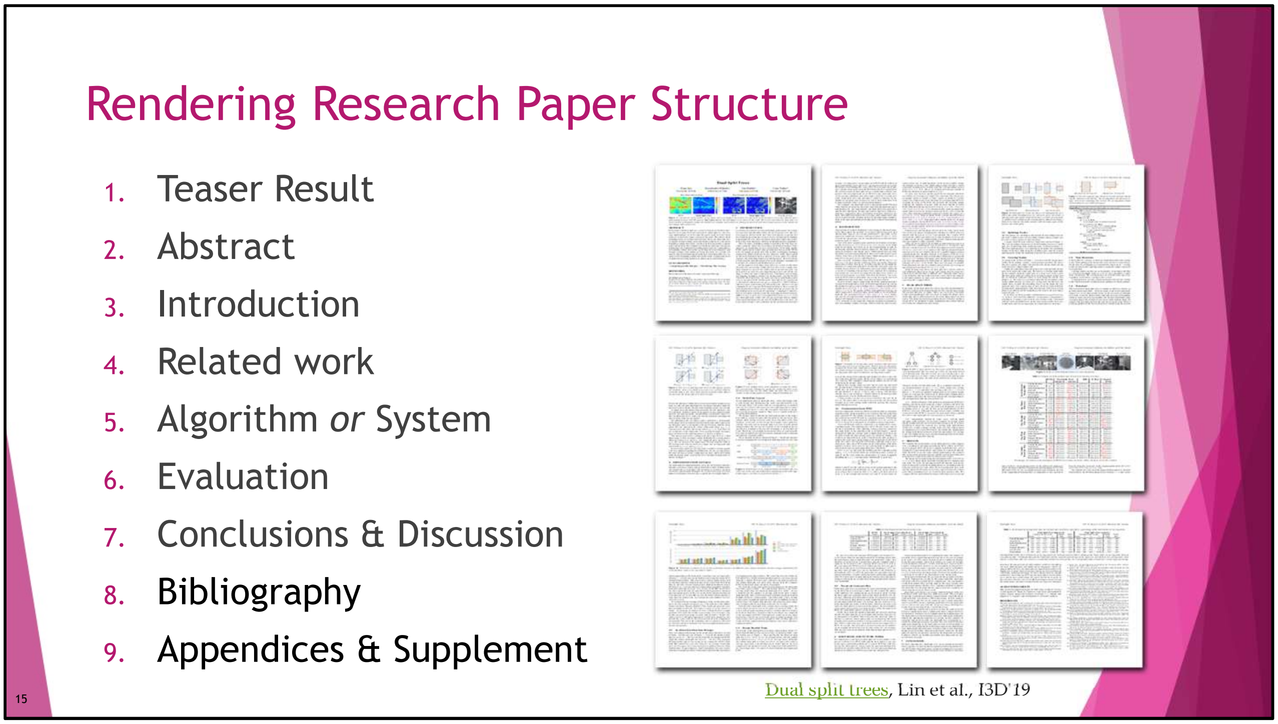
High Performance Graphics (HPG) and Interactive 3D Graphics and Games (I3D) are conferences that emphasize real-time rendering and practical, industry-facing solutions. They are closely related to the Journal of Computer Graphics Techniques and Proceedings of the ACM journals.

IEEE Computer Graphics & Applications is an outlier as a more informal journal tied to practice and implications.

The books on the last line are examples of journal-like edited collections favored by game developers. They aren’t peer-reviewed in the traditional sense and tend to be less academic, but are in many ways similar to JCGT as extremely practical advice on specific problems. These tend to have drop-in solutions for state-of-the-art renderers, especially real-time ones.

Parker et al., GPU Ray Tracing, Comm ACM 2013

Structure of a Paper



Rendering Research Paper Structure

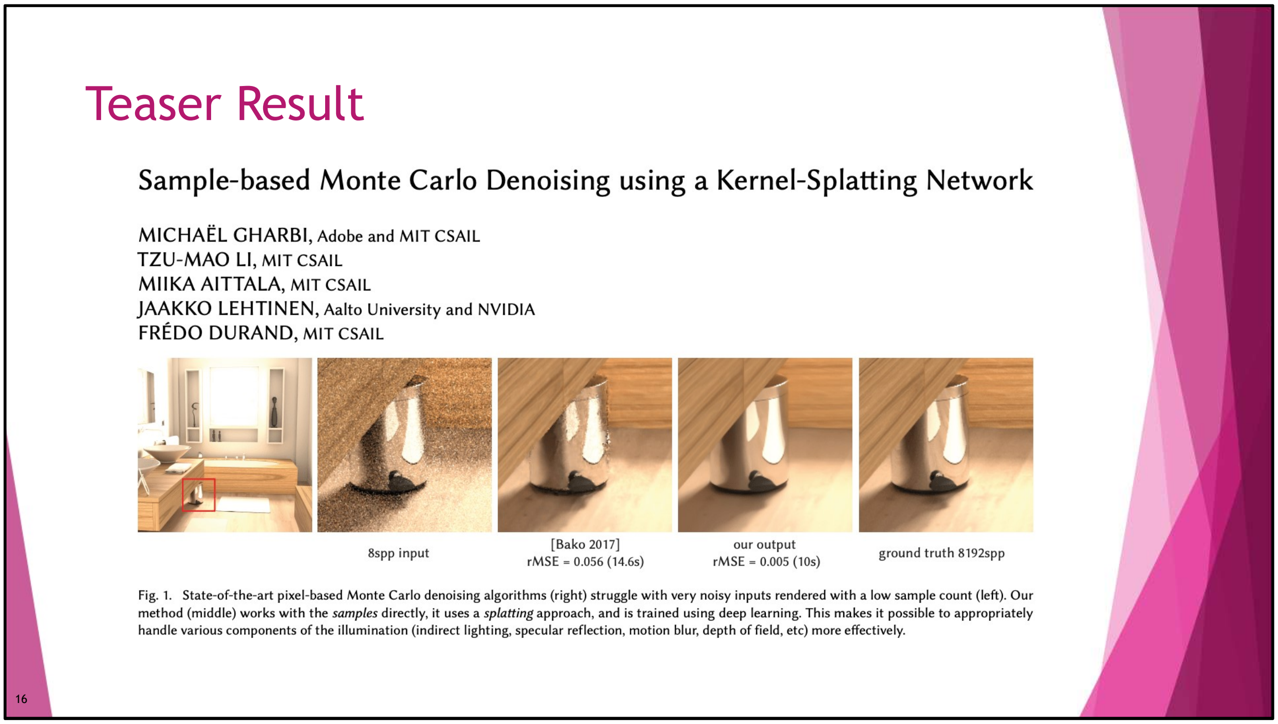
1. Teaser Result
2. Abstract
3. Introduction
4. Related work
5. Algorithm *or* System
6. Evaluation
7. Conclusions & Discussion
8. Bibliography
9. Appendices & Supplement

Dual split trees, Lin et al., I3D'19

15

Rendering papers tend to follow this main structure.

Most scientific writing is similar, although the emphasis shifts by field. For example, “related work” is usually a full-page survey in rendering but in psychology or economics is often a handful of citations in passing. In lab sciences, there is often a significant methodology section, and in mathematics there is rarely an experimental evaluation section.



Teaser Result

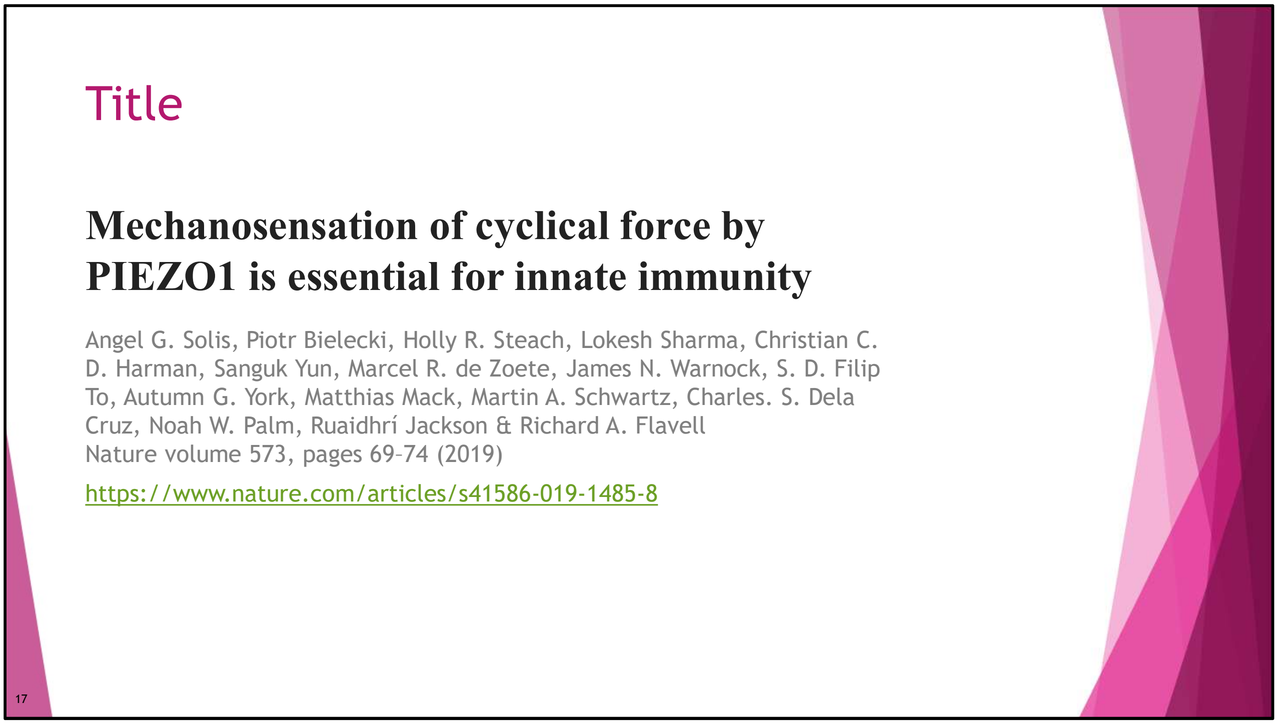
Sample-based Monte Carlo Denoising using a Kernel-Splatting Network

MICHAËL GHARBI, Adobe and MIT CSAIL
TZU-MAO LI, MIT CSAIL
MIIKA AITTALA, MIT CSAIL
JAAKKO LEHTINEN, Aalto University and NVIDIA
FRÉDO DURAND, MIT CSAIL

Fig. 1. State-of-the-art pixel-based Monte Carlo denoising algorithms (right) struggle with very noisy inputs rendered with a low sample count (left). Our method (middle) works with the *samples* directly, it uses a *splatting* approach, and is trained using deep learning. This makes it possible to appropriately handle various components of the illumination (indirect lighting, specular reflection, motion blur, depth of field, etc) more effectively.

16

Modern computer graphics papers almost always begin with a "teaser" image that shows the main result of the work and compares it to alternative methods.
This is a terrific practice which allows you to immediately understand what problem the paper is solving and how well its solution work in the best case.
It also provides a nice visual reference—there are many papers that I remember, e.g., as "the one with the egg pictures" or "the ring caustic in grass", because of some iconic image from the paper.



Title

Mechanosensation of cyclical force by PIEZO1 is essential for innate immunity

Angel G. Solis, Piotr Bielecki, Holly R. Steach, Lokesh Sharma, Christian C. D. Harman, Sanguk Yun, Marcel R. de Zoete, James N. Warnock, S. D. Filip To, Autumn G. York, Matthias Mack, Martin A. Schwartz, Charles. S. Dela Cruz, Noah W. Palm, Ruaidhrí Jackson & Richard A. Flavell
Nature volume 573, pages 69–74 (2019)

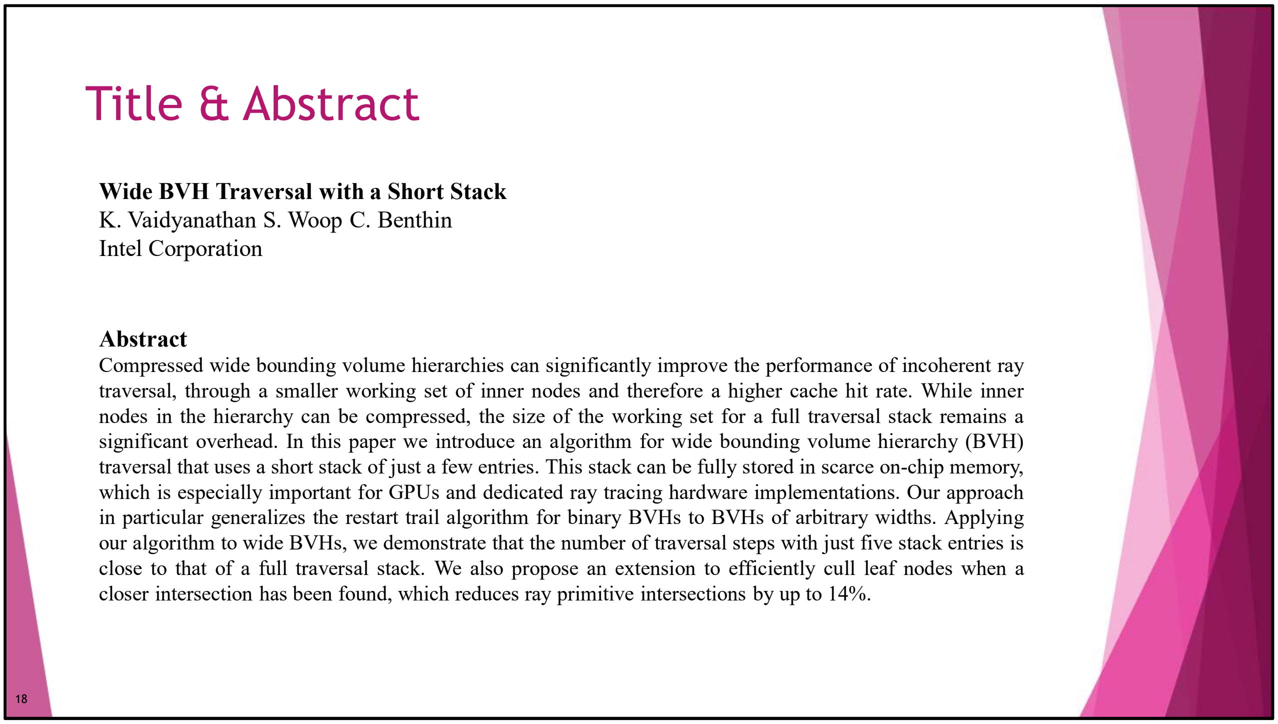
https://www.nature.com/articles/s41586-019-1485-8

17

For a really good scientific paper, the title is the main point of the paper, e.g.

This is easier to accomplish in some fields than others. In rendering, the title will usually give a hint as to the problem and main technique used to approach it, because we're largely evaluating our own algorithm designs rather than measuring or discovering something about the natural world.

Some graphics papers have really memorable titles ”Building Rome in a Day” or “Style Machines” that indicate the problem and will help you find the paper later, but tell you little about the technique or the result.



Title & Abstract

Wide BVH Traversal with a Short Stack
K. Vaidyanathan S. Woop C. Benthin
Intel Corporation

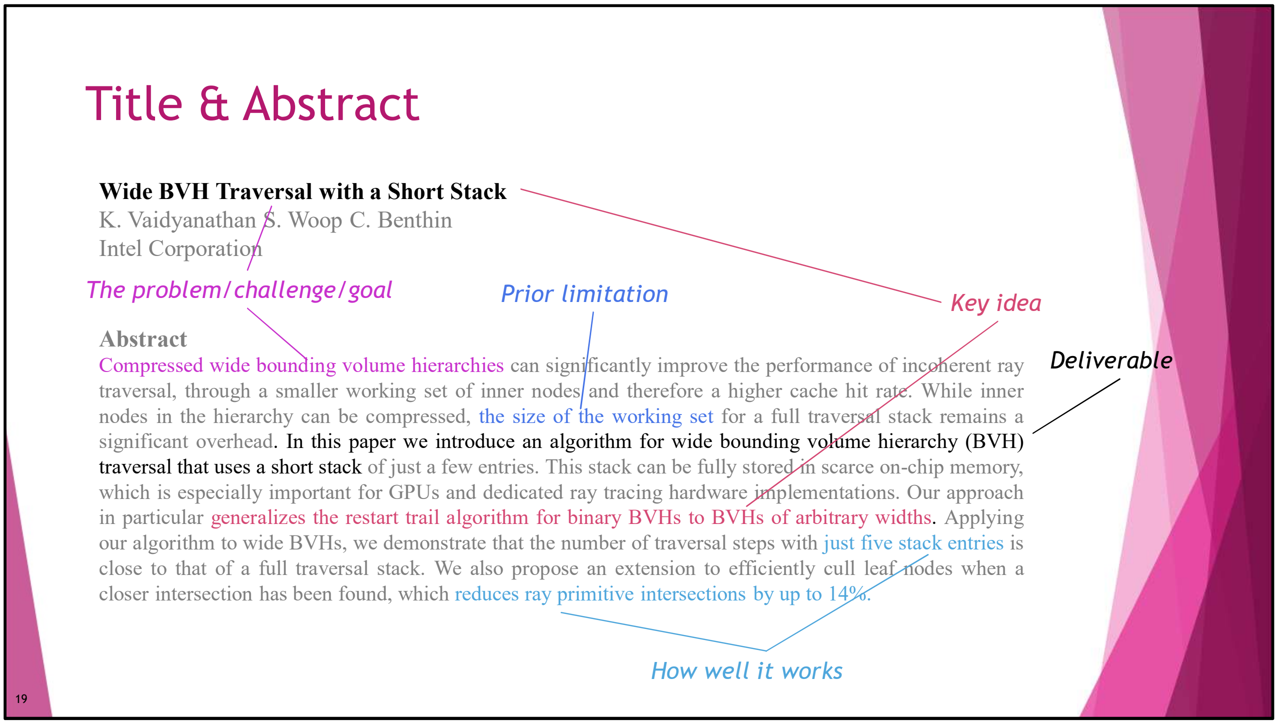
Abstract
Compressed wide bounding volume hierarchies can significantly improve the performance of incoherent ray traversal, through a smaller working set of inner nodes and therefore a higher cache hit rate. While inner nodes in the hierarchy can be compressed, the size of the working set for a full traversal stack remains a significant overhead. In this paper we introduce an algorithm for wide bounding volume hierarchy (BVH) traversal that uses a short stack of just a few entries. This stack can be fully stored in scarce on-chip memory, which is especially important for GPUs and dedicated ray tracing hardware implementations. Our approach in particular generalizes the restart trail algorithm for binary BVHs to BVHs of arbitrary widths. Applying our algorithm to wide BVHs, we demonstrate that the number of traversal steps with just five stack entries is close to that of a full traversal stack. We also propose an extension to efficiently cull leaf nodes when a closer intersection has been found, which reduces ray primitive intersections by up to 14%.

For a really good scientific paper, the title is the main point of the paper, e.g. Mechanosensation of cyclical force by PIEZO1 is essential for innate immunity

This is easier to accomplish in some fields than others.

An abstract tells you:

1. The problem that the paper is solving
2. Why previous solutions are imperfect
3. The key idea for the new solution or insight into how to look at the problem
4. How well the new solution works



Title & Abstract

Wide BVH Traversal with a Short Stack
K. Vaidyanathan S. Woop C. Benthin
Intel Corporation

The problem/challenge/goal

Prior limitation

Key idea

Deliverable

Abstract
Compressed wide bounding volume hierarchies can significantly improve the performance of incoherent ray traversal, through a smaller working set of inner nodes and therefore a higher cache hit rate. While inner nodes in the hierarchy can be compressed, the size of the working set for a full traversal stack remains a significant overhead. In this paper we introduce an algorithm for wide bounding volume hierarchy (BVH) traversal that uses a short stack of just a few entries. This stack can be fully stored in scarce on-chip memory, which is especially important for GPUs and dedicated ray tracing hardware implementations. Our approach in particular generalizes the restart trail algorithm for binary BVHs to BVHs of arbitrary widths. Applying our algorithm to wide BVHs, we demonstrate that the number of traversal steps with just five stack entries is close to that of a full traversal stack. We also propose an extension to efficiently cull leaf nodes when a closer intersection has been found, which reduces ray primitive intersections by up to 14%.

How well it works

In this particular example, I like that:

- The abstract gives quantitative results
- The title makes the problem (wide BVH traversal) and solution (short stack) clear
- The abstract covers all of the areas that it should
- The abstract is short!

For what it’s worth, this could be improved by:

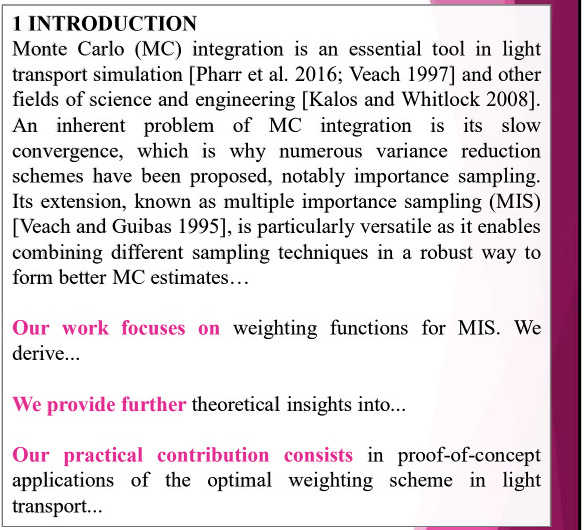
- Quantifying the speedup for GPU BVH traversal using this stack, which is the real payoff
- Moving the “In this paper…” sentence to the *front*, so that I immediately know what it is delivering: an algorithm
- Explaining *why* keeping the stack in registers (not just “on chip”) gives such a huge speedup for GPUs
- Removing some of the repetition, “short stack…just a few entries”, “just five stack entries” to make room for the above

…but *writing* a paper isn’t our goal today, and I want to leave you with the high-order impression that this is a good abstract and that the highlighted words are what you're trying to get from it as a reader.

Introduction

- Very readable, high-level
- Background & problem statement
- Seminal related work citations
- Description of constraints
- The paper's approach to the problem
- **Contributions of this paper**

This is what you need for your 1st pass



1 INTRODUCTION

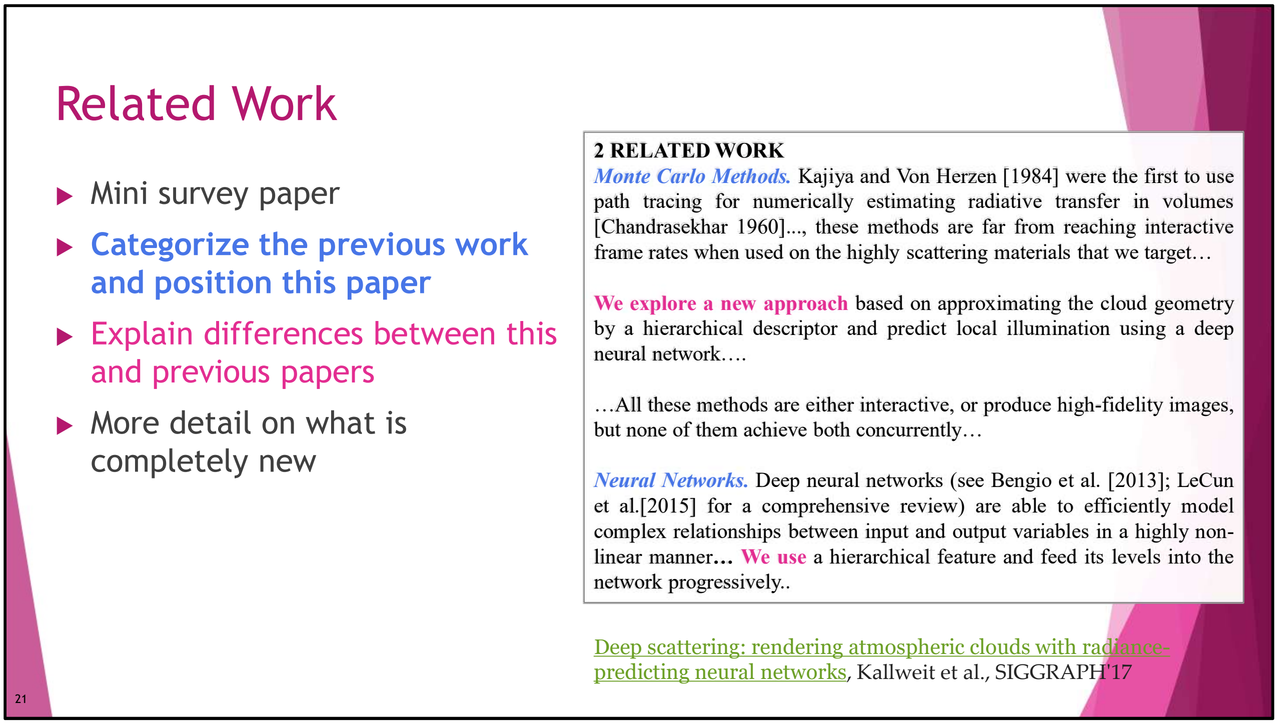
Monte Carlo (MC) integration is an essential tool in light transport simulation [Pharr et al. 2016; Veach 1997] and other fields of science and engineering [Kalos and Whitlock 2008]. An inherent problem of MC integration is its slow convergence, which is why numerous variance reduction schemes have been proposed, notably importance sampling. Its extension, known as multiple importance sampling (MIS) [Veach and Guibas 1995], is particularly versatile as it enables combining different sampling techniques in a robust way to form better MC estimates…

Our work focuses on weighting functions for MIS. We derive...

We provide further theoretical insights into...

Our practical contribution consists in proof-of-concept applications of the optimal weighting scheme in light transport...

Optimal multiple importance sampling, Kondapaneni et al., SIGGRAPH'19



Related Work

- Mini survey paper
- Categorize the previous work and position this paper
- Explain differences between this and previous papers
- More detail on what is completely new

2 RELATED WORK

Monte Carlo Methods. Kajiya and Von Herzen [1984] were the first to use path tracing for numerically estimating radiative transfer in volumes [Chandrasekhar 1960]..., these methods are far from reaching interactive frame rates when used on the highly scattering materials that we target…

We explore a new approach based on approximating the cloud geometry by a hierarchical descriptor and predict local illumination using a deep neural network….

…All these methods are either interactive, or produce high-fidelity images, but none of them achieve both concurrently…

Neural Networks. Deep neural networks (see Bengio et al. [2013]; LeCun et al.[2015] for a comprehensive review) are able to efficiently model complex relationships between input and output variables in a highly non-linear manner**…** **We use** a hierarchical feature and feed its levels into the network progressively..

Deep scattering: rendering atmospheric clouds with radiance-predicting neural networks, Kallweit et al., SIGGRAPH'17

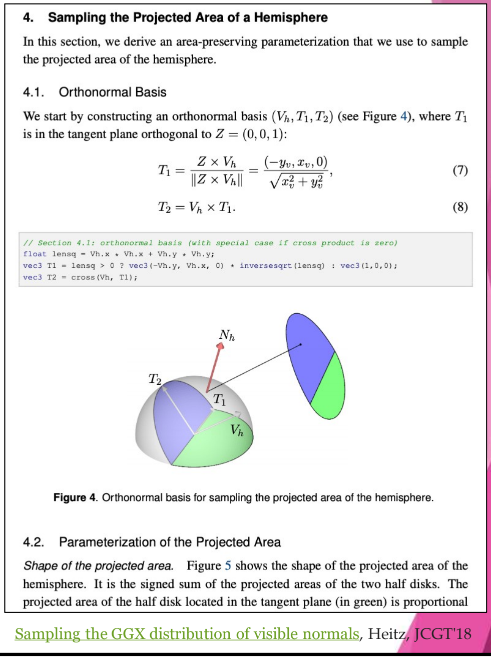
What you're trying to get on your first pass through the paper is an understanding of "how is this different from other approaches", which is what I'm highlighting in pink.

If you don't know much about this paper's particular area, then the blue categorization will give you some context, however you should just read the subsection/paragraph titles at first and skip most of the text.

Algorithm or System

- Main body of the new technique
- Includes both derivation and solution
- May span multiple sections

- Only a small part of this is the new contribution you're looking for!
- …and it may depend on pieces that *only* appear in other work, not even here



4. Sampling the Projected Area of a Hemisphere

In this section, we derive an area-preserving parameterization that we use to sample the projected area of the hemisphere.

4.1. Orthonormal Basis

We start by constructing an orthonormal basis (V_h, T_1, T_2) (see Figure 4), where T_1 is in the tangent plane orthogonal to $Z = (0, 0, 1)$:

$$T_1 = \frac{Z \times V_h}{\|Z \times V_h\|} = \frac{(-y_v, x_v, 0)}{\sqrt{x_v^2 + y_v^2}}, \tag{7}$$

$$T_2 = V_h \times T_1. \tag{8}$$

```
// Section 4.1: orthonormal basis (with special case if cross product is zero)
float lensq = Vh.x * Vh.x + Vh.y * Vh.y;
vec3 T1 = lensq > 0 ? vec3(-Vh.y, Vh.x, 0) * inversesqrt(lensq) : vec3(1,0,0);
vec3 T2 = cross(Vh, T1);
```

Figure 4. Orthonormal basis for sampling the projected area of the hemisphere.

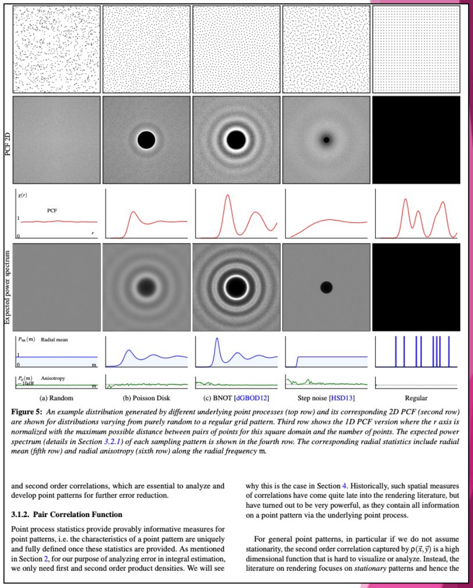
4.2. Parameterization of the Projected Area

Shape of the projected area. Figure 5 shows the shape of the projected area of the hemisphere. It is the signed sum of the projected areas of the two half disks. The projected area of the half disk located in the tangent plane (in green) is proportional

Sampling the GGX distribution of visible normals, Heitz, JCGT'18

Evaluation (a.k.a. Results)

- Quantitative evaluation
- Comparison to previous work and “ground truth”
- Measured performance or asymptotic analysis
- Look at images & read captions on first reading pass
- Beware of measurement differences between papers


PCF 2D
g(r)
PCF
1
0
r
Expected power spectrum
$P_{\mathrm{m}}(\mathrm{m})$ Radial mean
1
0
m
$P_{a}(\mathrm{m})$ Anisotropy
−10dB
m

(a) Random (b) Poisson Disk (c) BNOT [dGBOD12] Step noise [HSD13] Regular

Figure 5: *An example distribution generated by different underlying point processes (top row) and its corresponding 2D PCF (second row) are shown for distributions varying from purely random to a regular grid pattern. Third row shows the 1D PCF version where the r axis is normalized with the maximum possible distance between pairs of points for this square domain and the number of points. The expected power spectrum (details in Section 3.2.1) of each sampling pattern is shown in the fourth row. The corresponding radial statistics include radial mean (fifth row) and radial anisotropy (sixth row) along the radial frequency* m.

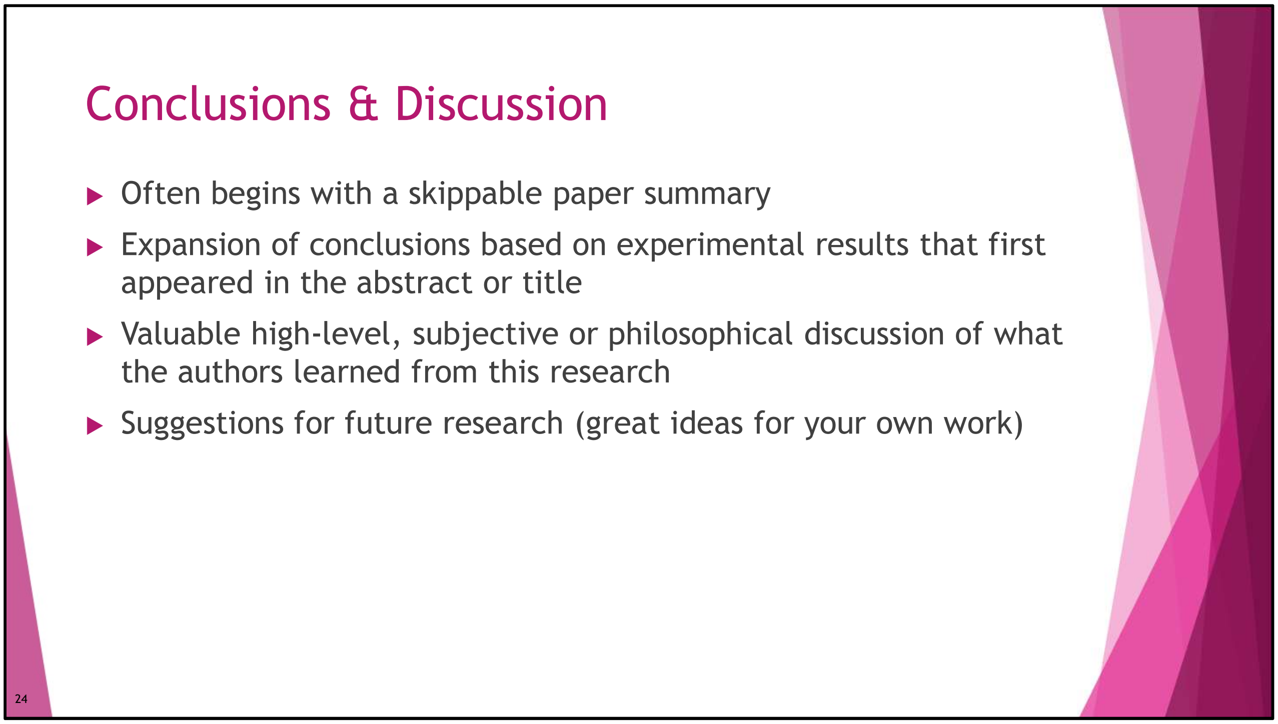
and second order correlations, which are essential to analyze and develop point patterns for further error reduction.

3.1.2. Pair Correlation Function

Point process statistics provide provably informative measures for point patterns, i.e. the characteristics of a point pattern are uniquely and fully defined once these statistics are provided. As mentioned in Section 2, for our purpose of analyzing error in integral estimation, we only need first and second order product densities. We will see why this is the case in Section 4. Historically, such spatial measures of correlations have come quite late into the rendering literature, but have turned out to be very powerful, as they contain all information on a point pattern via the underlying point process.

For general point patterns, in particular if we do not assume stationarity, the second order correlation captured by $\rho(\vec{x},\vec{y})$ is a high dimensional function that is hard to visualize or analyze. Instead, the literature on rendering focuses on *stationary* patterns and hence the

Analysis of Sample Correlations for Monte Carlo Rendering,
Singh et al., Eurographics'19



Conclusions & Discussion

- Often begins with a skippable paper summary
- Expansion of conclusions based on experimental results that first appeared in the abstract or title
- Valuable high-level, subjective or philosophical discussion of what the authors learned from this research
- Suggestions for future research (great ideas for your own work)

Burley et al., The Design and Evolution of Disney's Hyperion Renderer, ACM ToG 2018

© 2018 DISNEY

Reading Process

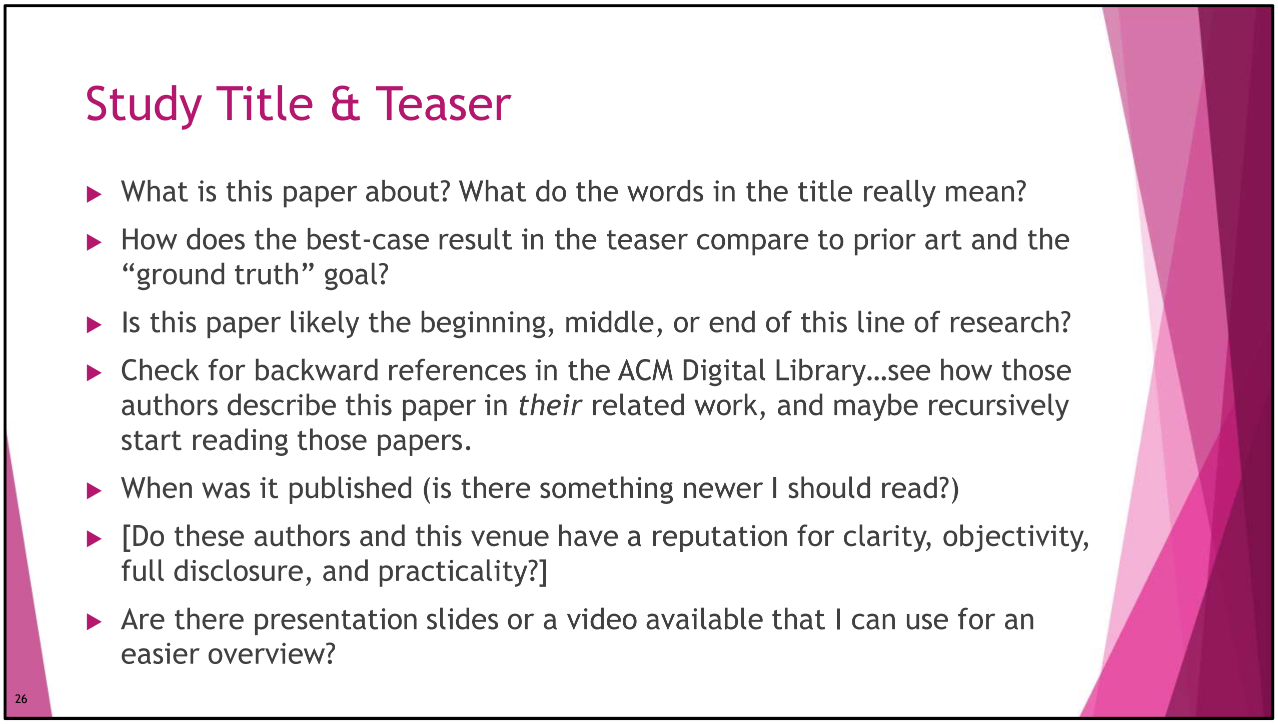
About 10% of a paper actually describes the new information that you’re trying to learn.

The rest is context and evaluation. Those are valuable, of course, but you need to first identify what the new piece to understand *why* the paper matters, whether you’re reading the *right* paper, and how important that new piece actually is to you (Does it produce good results? Does it execute efficiently? Does it have unacceptable constraints or limitations?) These are of course the same questions a reviewer is trying to answer!

Here’s how I read a realistic rendering paper. Some of this is rendering-specific (or Morgan-specific!), but all scientists read in multiple passes and jump around within each pass roughly in this way.

As a running example, I’ll use the Kajiya ’86 paper that was assigned reading for today. I’m choosing that because it is the key paper from which everything else in this course follows, not because it is the most representative paper in its structure or the best-written paper (although it has an awful lot to like).

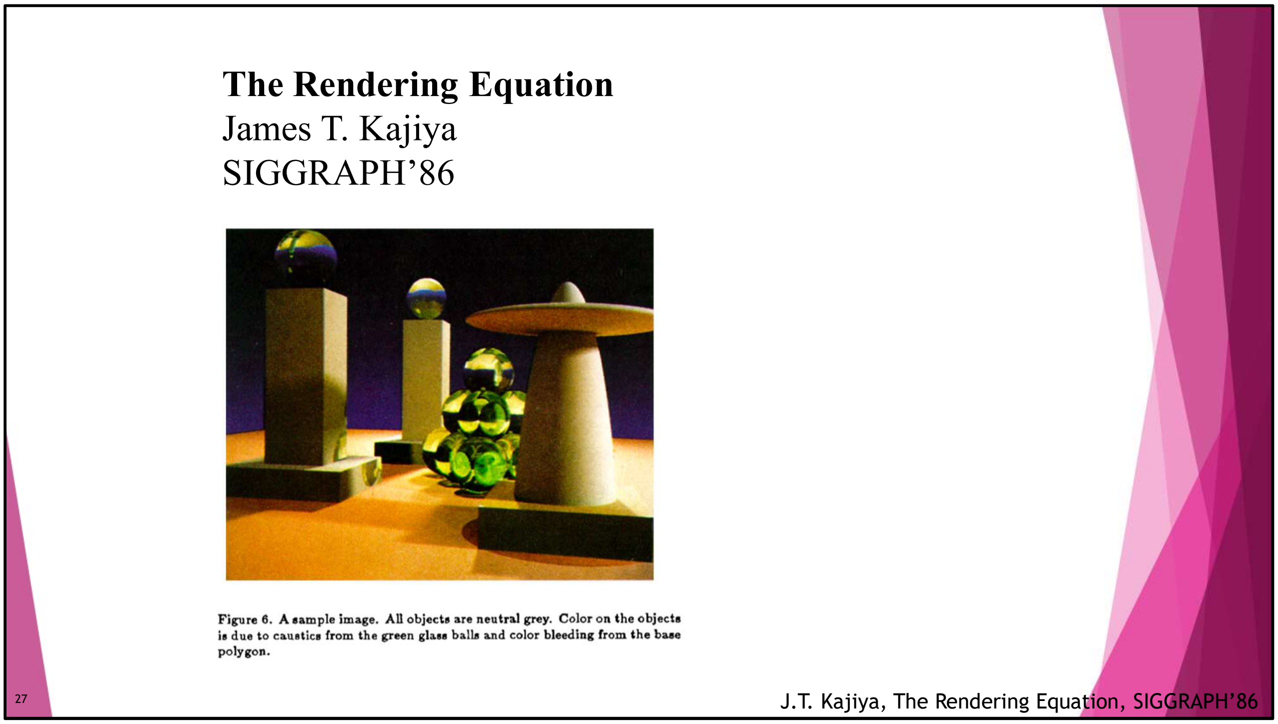
I’m going to show you how to *read* it right now, not how to *present* it. Please don’t structure your in-class presentations like this! The next lecture covers how to present a paper, and then the one after that shows the same content structured as an undergraduate lecture instead of a research paper or seminar.



Study Title & Teaser

- What is this paper about? What do the words in the title really mean?
- How does the best-case result in the teaser compare to prior art and the "ground truth" goal?
- Is this paper likely the beginning, middle, or end of this line of research?
- Check for backward references in the ACM Digital Library…see how those authors describe this paper in *their* related work, and maybe recursively start reading those papers.
- When was it published (is there something newer I should read?)
- [Do these authors and this venue have a reputation for clarity, objectivity, full disclosure, and practicality?]
- Are there presentation slides or a video available that I can use for an easier overview?

I spend several minutes trying to figure the paper out from just the title and teaser image (plus caption). The helps me to understand how important this paper will be (maybe I should spend my time reading something else first) and how to approach it.



The Rendering Equation
James T. Kajiya
SIGGRAPH’86

Figure 6. A sample image. All objects are neutral grey. Color on the objects is due to caustics from the green glass balls and color bleeding from the base polygon.

27

J.T. Kajiya, The Rendering Equation, SIGGRAPH’86

I’m going to use this paper as a running example. It predates teasers, but here’s the final image from the paper, which *would* have been a teaser in the modern format.

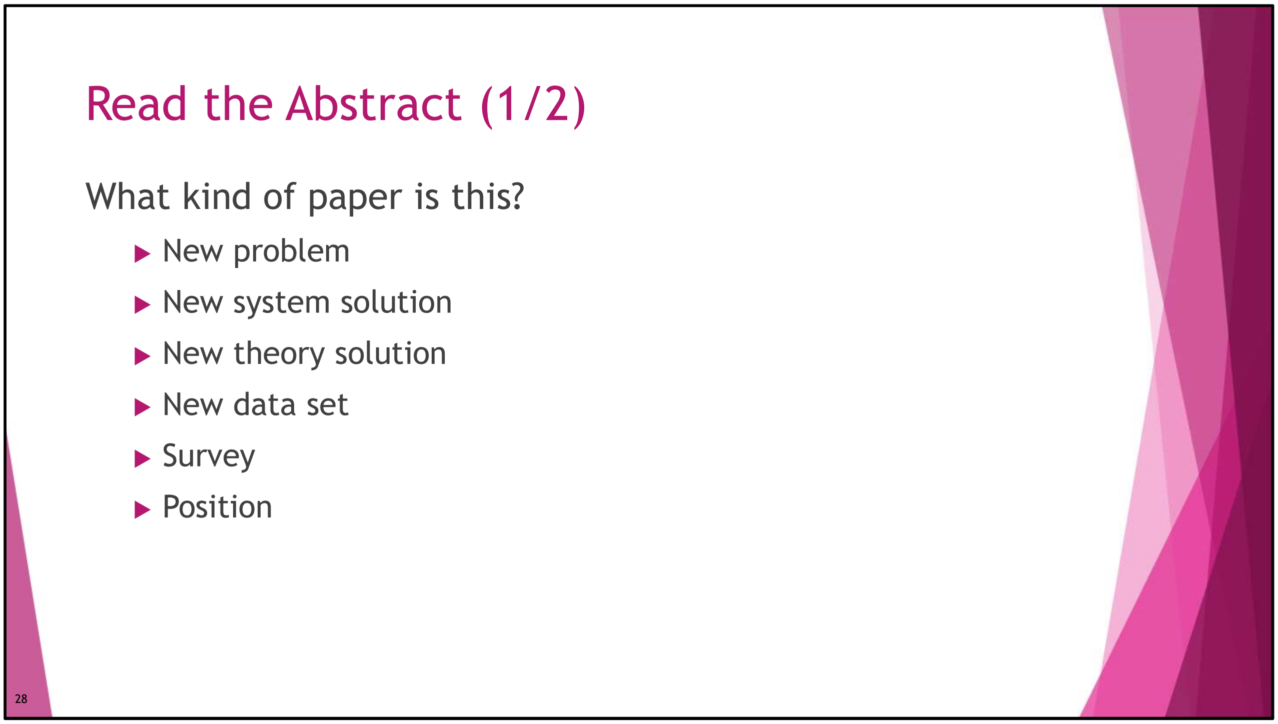
The title is “*The* Rendering Equation”. That’s awfully presumptuous. But it is clearly about introducing a formal framework for *all* rendering. So, I can expect to get some big picture perspective from this paper.

There are also pictures, so clearly I can also expect to get a rendering *algorithm* for solving that equation. As indicated by the caption, the image has very complicated lighting effects, so this algorithm must handle a pretty general case well.

This paper is from 1986 and has been cited 3,117 times according to the ACM. So, it must be a very important paper (100 citations for a rendering paper would be a lot). The algorithm as originally presented also must be somewhat dated, considering that there are at least three thousand more recent advances on this topic.

Kajiya was a professor at CalTech at the time and had coauthored about 30 previous papers, including with folks like Turner Whitted (the recursive “ray tracing” inventor) and Bill Dally (NVIDIA Chief Scientist), so he’s probably a credible and important person.

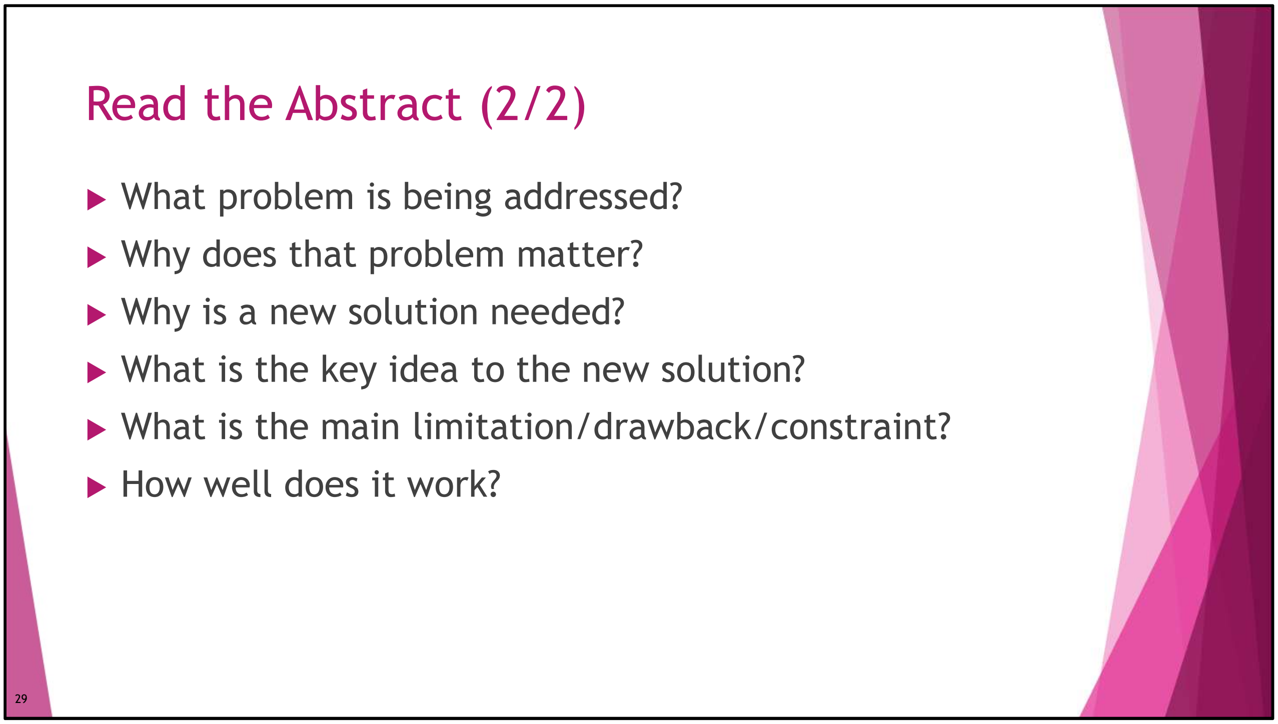
There’s no presentation by Kajiya online, but there are a lot of textbooks and surveys that describe this work, as well as many lecture notes that explain it. That won’t be the case for most papers, so let’s skip those for the moment and move on…



Read the Abstract (1/2)

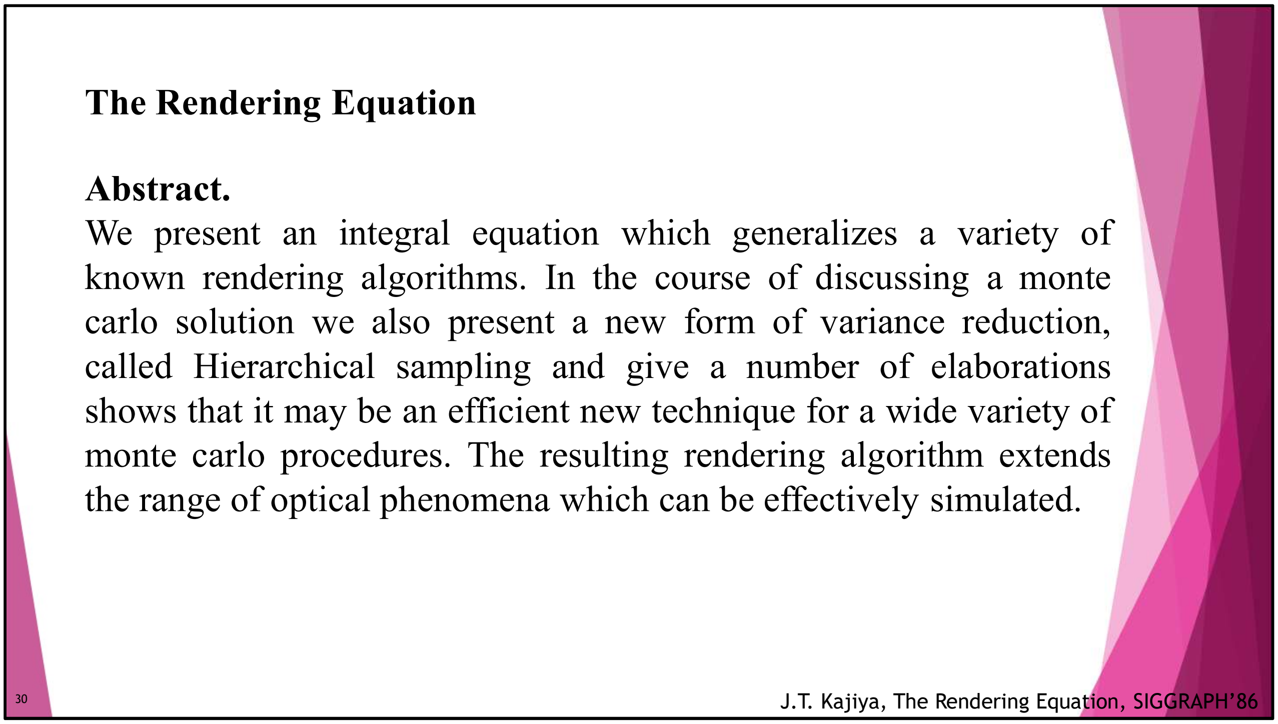
What kind of paper is this?

- New problem
- New system solution
- New theory solution
- New data set
- Survey
- Position



Read the Abstract (2/2)

- What problem is being addressed?
- Why does that problem matter?
- Why is a new solution needed?
- What is the key idea to the new solution?
- What is the main limitation/drawback/constraint?
- How well does it work?

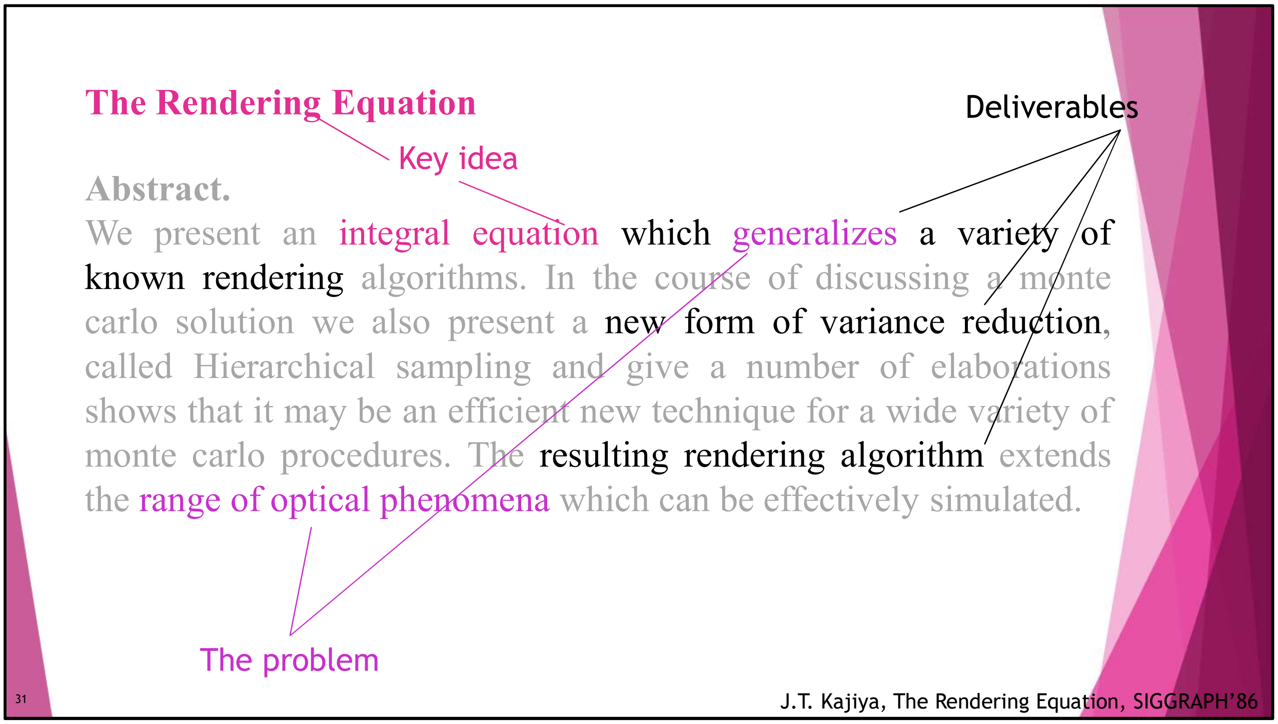


The Rendering Equation

Abstract.
We present an integral equation which generalizes a variety of known rendering algorithms. In the course of discussing a monte carlo solution we also present a new form of variance reduction, called Hierarchical sampling and give a number of elaborations shows that it may be an efficient new technique for a wide variety of monte carlo procedures. The resulting rendering algorithm extends the range of optical phenomena which can be effectively simulated.

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

(there's a missing comma after "sampling" and "shows" should probably be "that show"…this is verbatim from Kajiya's abstract)



The Rendering Equation

Key idea

Deliverables

Abstract.

We present an integral equation which generalizes a variety of known rendering algorithms. In the course of discussing a monte carlo solution we also present a new form of variance reduction, called Hierarchical sampling and give a number of elaborations shows that it may be an efficient new technique for a wide variety of monte carlo procedures. The resulting rendering algorithm extends the range of optical phenomena which can be effectively simulated.

The problem

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

The problem this paper addresses generalizing rendering so that it can address more optical phenomena. It is also seeking to reduce the variance (aliasing/noise) in Monte Carlo sampling.

The key idea is a new way of thinking about the problem. The titular "rendering equation" is going to be an integral equation that generalizes previous approaches.

There are three deliverables:

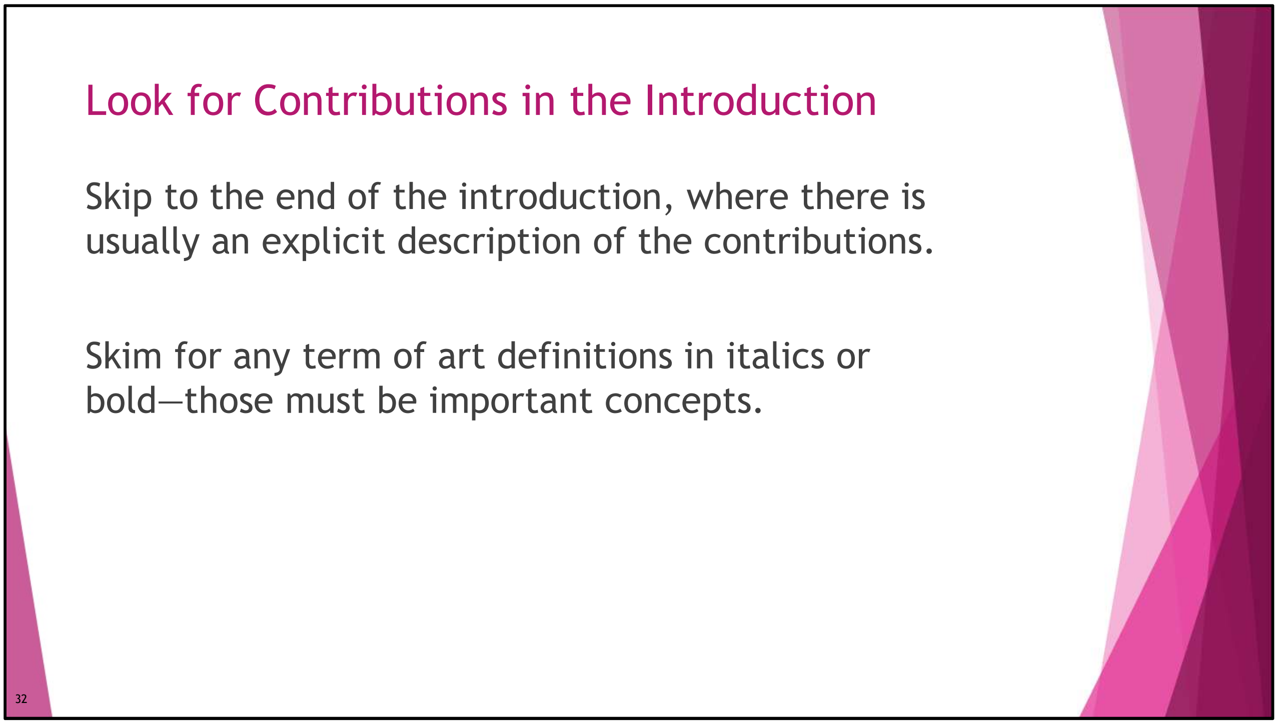
- A general way of thinking about rendering (The Rendering Equation; not claimed to be entirely new)
- A new algorithm for variance reduction (Hierarchical Sampling)
- A new rendering algorithm (Path Tracing). This is buried a bit: it is "a monte carlo solution" + "the resulting rendering algorithm"

(The hierarchical sampling turns out to not be very important as applied in this paper. It is mostly an antialiasing method for primary rays which was soon replaced by the quasi-monte carlo methods that we'll study in this course, and in its other applications by adaptive sampling and denoising. I point this out because recognizing what is NOT important to read right now is a key skill in working with research papers. There's nothing in this abstract that indicates the Hierarchical Sampling is not the important part; rather, it actually seems like *the* most important part as it is the only named algorithm in the abstract. You need outside knowledge to recognize that it isn't essential.)

The limitation of previous work is implicit: it cannot simulate a large range of phenomena

Note carefully the word choices here, which you might miss if only skimming. "extends the range"…Kajiya is telling you that he improves the state of the art, but doesn't think that he's solved for *all* optical phenomena. He generalizes "known" algorithms, not all possible rendering algorithms. "May be …efficient"; and not just limited to his one motivating application but a "wide variety" of others.

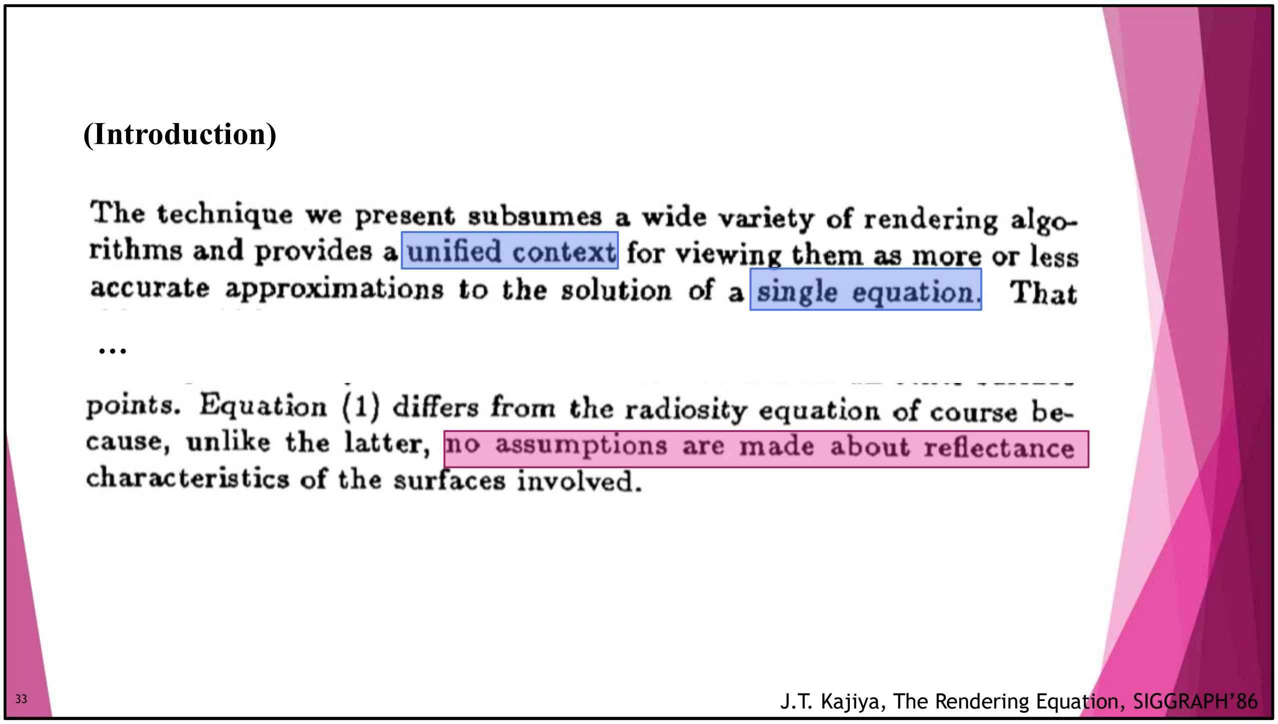
Two one aspects I'd personally change about this abstract to improve it are to call out path tracing by name and to be more explicit about how well the provided algorithms perform. E.g., What's the asymptotic performance? ($O(2^k)$ → $O(k)$ for k scattering events, converges like sqrt(n) for n path samples per pixel). Wall clock time? (20 hours at 512x512) What some of these newly-simulated optical phenomena (caustics, glossy reflection, diffuse interreflection, natural extension to spectral and polarization rendering)? Do these methods fail to perform well in certain cases, such as narrow caustics, tiny apertures, or shiny surfaces under large lights? (hint: yes, path tracing has problems there! Ten years later Veach and Guibas note those problems in Kajiya's work and write another masterpiece paper, on Metropolis Light Transport, that we'll also read in this course)



Look for Contributions in the Introduction

Skip to the end of the introduction, where there is usually an explicit description of the contributions.

Skim for any term of art definitions in italics or bold—those must be important concepts.



(Introduction)

The technique we present subsumes a wide variety of rendering algorithms and provides a unified context for viewing them as more or less accurate approximations to the solution of a single equation. That

…

points. Equation (1) differs from the radiosity equation of course because, unlike the latter, no assumptions are made about reflectance characteristics of the surfaces involved.

33

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

The contributions are:

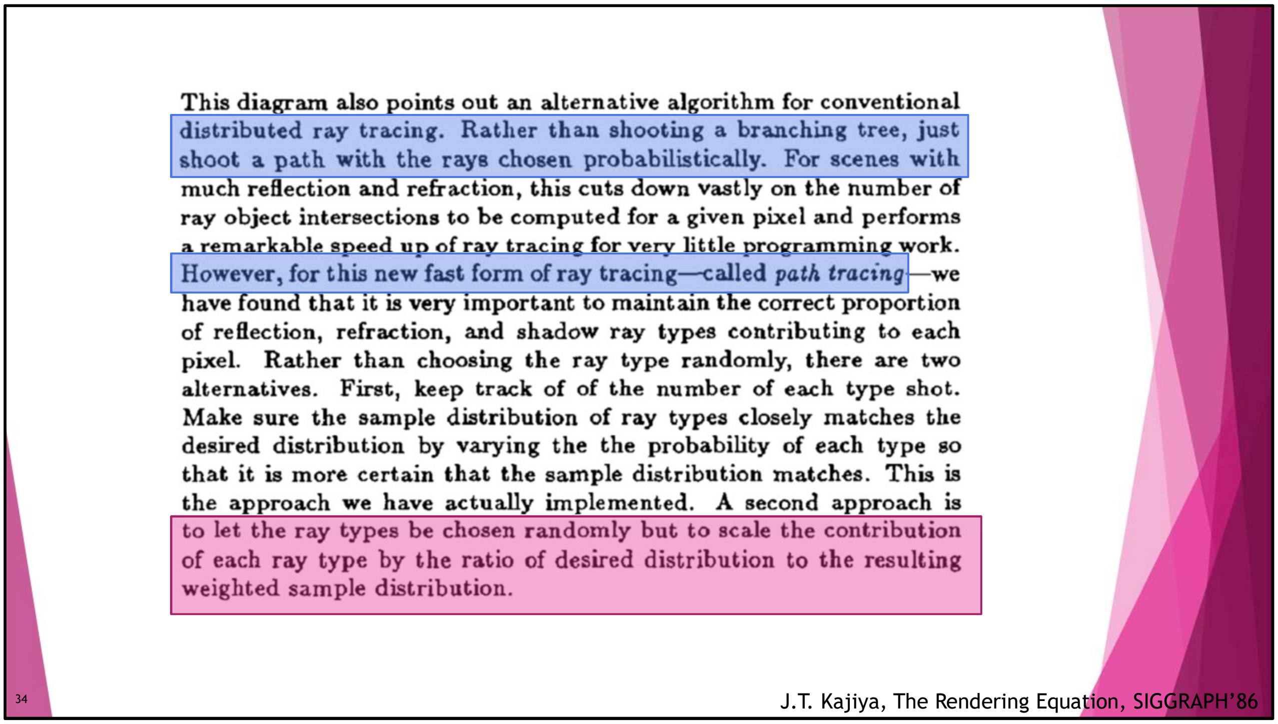
1. the generality of the unified approach
2. no assumptions about reflectance

…

and one more contribution that is not mentioned in the introduction or title.

This is a major weakness of the paper's exposition. This other unmentioned contribution turns out to be *the most important aspect of the paper* with the benefit of historical hindsight.

What is this contribution? Well, much later, at the very end of the algorithm section, Kajiya mentions:



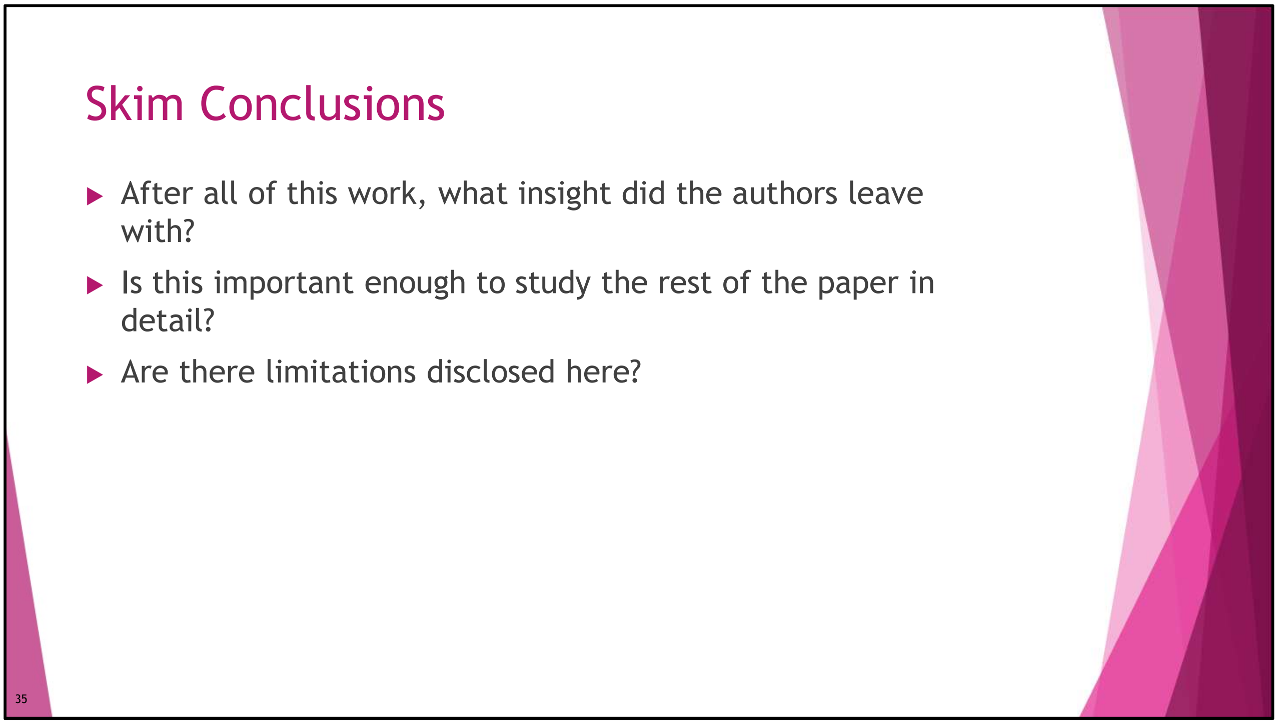
This diagram also points out an alternative algorithm for conventional distributed ray tracing. Rather than shooting a branching tree, just shoot a path with the rays chosen probabilistically. For scenes with much reflection and refraction, this cuts down vastly on the number of ray object intersections to be computed for a given pixel and performs a remarkable speed up of ray tracing for very little programming work. However, for this new fast form of ray tracing—called *path tracing*—we have found that it is very important to maintain the correct proportion of reflection, refraction, and shadow ray types contributing to each pixel. Rather than choosing the ray type randomly, there are two alternatives. First, keep track of of the number of each type shot. Make sure the sample distribution of ray types closely matches the desired distribution by varying the the probability of each type so that it is more certain that the sample distribution matches. This is the approach we have actually implemented. A second approach is to let the ray types be chosen randomly but to scale the contribution of each ray type by the ratio of desired distribution to the resulting weighted sample distribution.

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

This is *the path tracing algorithm*, which is the basis for this entire course. That's why the paper was cited over 3000 times. This isn't unusual, though…often when writing a paper, we don't know which parts will stand the test of time. And to be fair, Kajiya's insight of path tracing grew largely from thinking about the rendering equation (which was already known to others, as he points out, and is expressed better in the modern solid angle form by a simultaneously-published paper, Immel et al.'86!) in a new way, so he is right that the key insight is thinking about the problem differently and the algorithm "falls out" of that thinking.

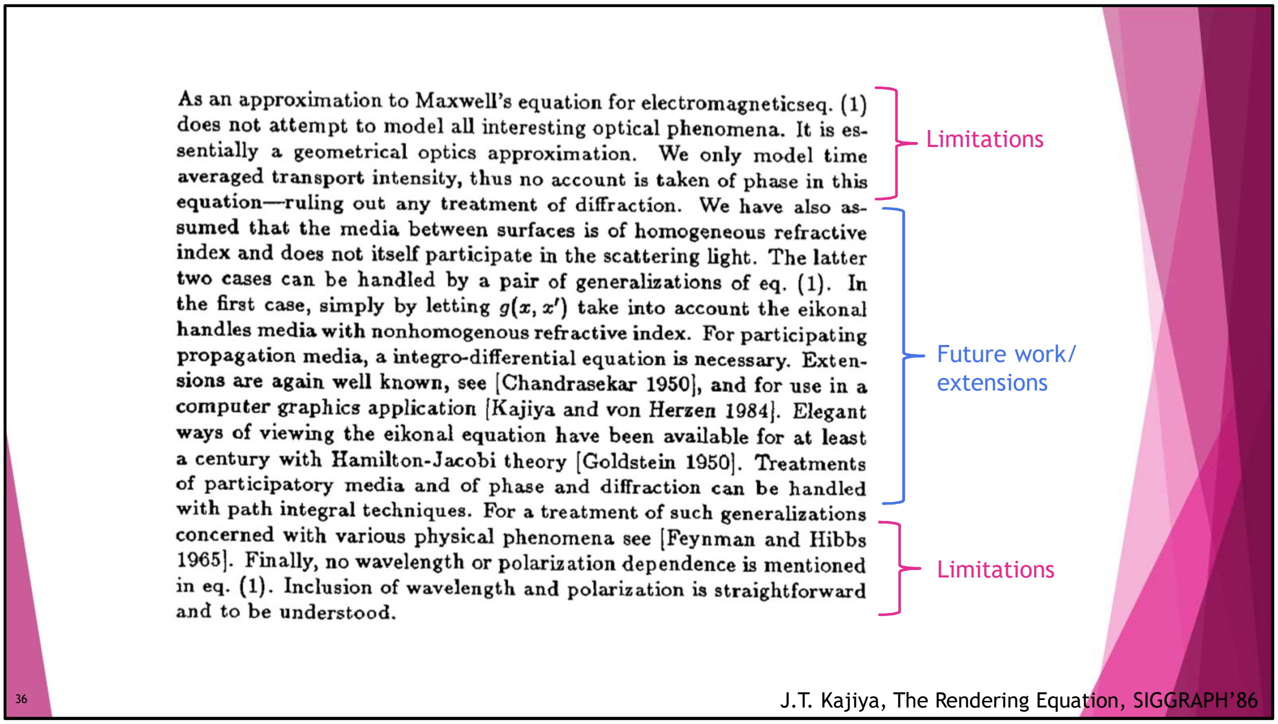
To add more irony, this very last sentence, *which describes a method that the author didn't even bother to implement*, is how path tracers are actually written today. It is a genius idea. But even geniuses can't always recognize which of their own ideas are going to be important in the long run.

Having read all of this, we now wish the paper had been titled "the path tracing algorithm as a monte carlo solution to the rendering equation" and that "the path tracing algorithm" had been explicitly mentioned in the introduction as one of the contributions. Let's give Kajiya credit, however, both for inventing the algorithm and for recognizing that thinking about the rendering problem as monte carlo estimators for a general rendering equation—both of which are core ideas to modern rendering and come from a single paper.



Skim Conclusions

- After all of this work, what insight did the authors leave with?
- Is this important enough to study the rest of the paper in detail?
- Are there limitations disclosed here?



As an approximation to Maxwell's equation for electromagneticseq. (1) does not attempt to model all interesting optical phenomena. It is essentially a geometrical optics approximation. We only model time averaged transport intensity, thus no account is taken of phase in this equation—ruling out any treatment of diffraction. We have also assumed that the media between surfaces is of homogeneous refractive index and does not itself participate in the scattering light. The latter two cases can be handled by a pair of generalizations of eq. (1). In the first case, simply by letting $g(x, x')$ take into account the eikonal handles media with nonhomogenous refractive index. For participating propagation media, a integro-differential equation is necessary. Extensions are again well known, see [Chandrasekar 1950], and for use in a computer graphics application [Kajiya and von Herzen 1984]. Elegant ways of viewing the eikonal equation have been available for at least a century with Hamilton-Jacobi theory [Goldstein 1950]. Treatments of participatory media and of phase and diffraction can be handled with path integral techniques. For a treatment of such generalizations concerned with various physical phenomena see [Feynman and Hibbs 1965]. Finally, no wavelength or polarization dependence is mentioned in eq. (1). Inclusion of wavelength and polarization is straightforward and to be understood.

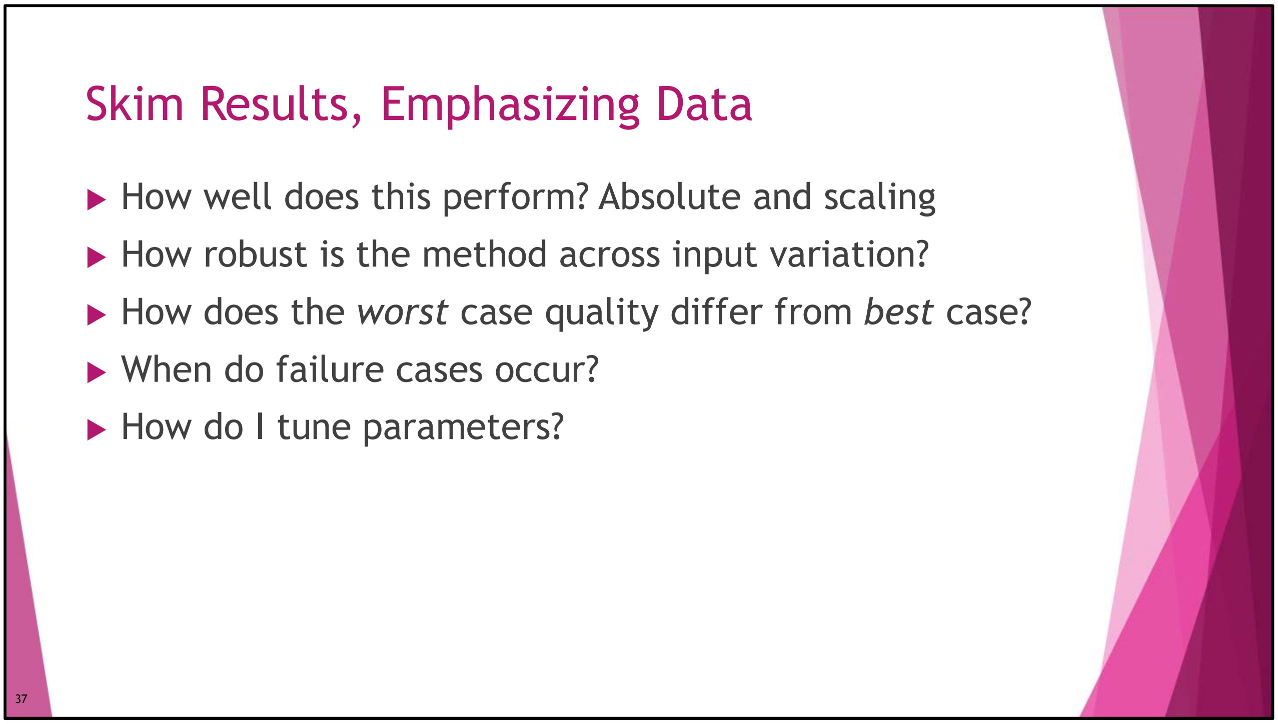
36

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

Kajiya oddly does not have any conclusions in this paper. That's mostly because he front-loaded the philosophy and insights (older papers tend to do this). This is the end of the introduction section, which we'd expect to find at the end of the entire paper today.

He already opened the entire paper with the big conclusion: thinking about rendering as a Monte Carlo Markov Chain solution for a stripped-down expression of Maxwell's equations.

This section of the introduction gives a list of limitations and future work that might otherwise appear at the end of the paper. For what it is worth, I appreciate papers that put the limitations in the abstract or introduction like this, so that I don't have to hunt for them.



Skim Results, Emphasizing Data

- How well does this perform? Absolute and scaling
- How robust is the method across input variation?
- How does the *worst* case quality differ from *best* case?
- When do failure cases occur?
- How do I tune parameters?

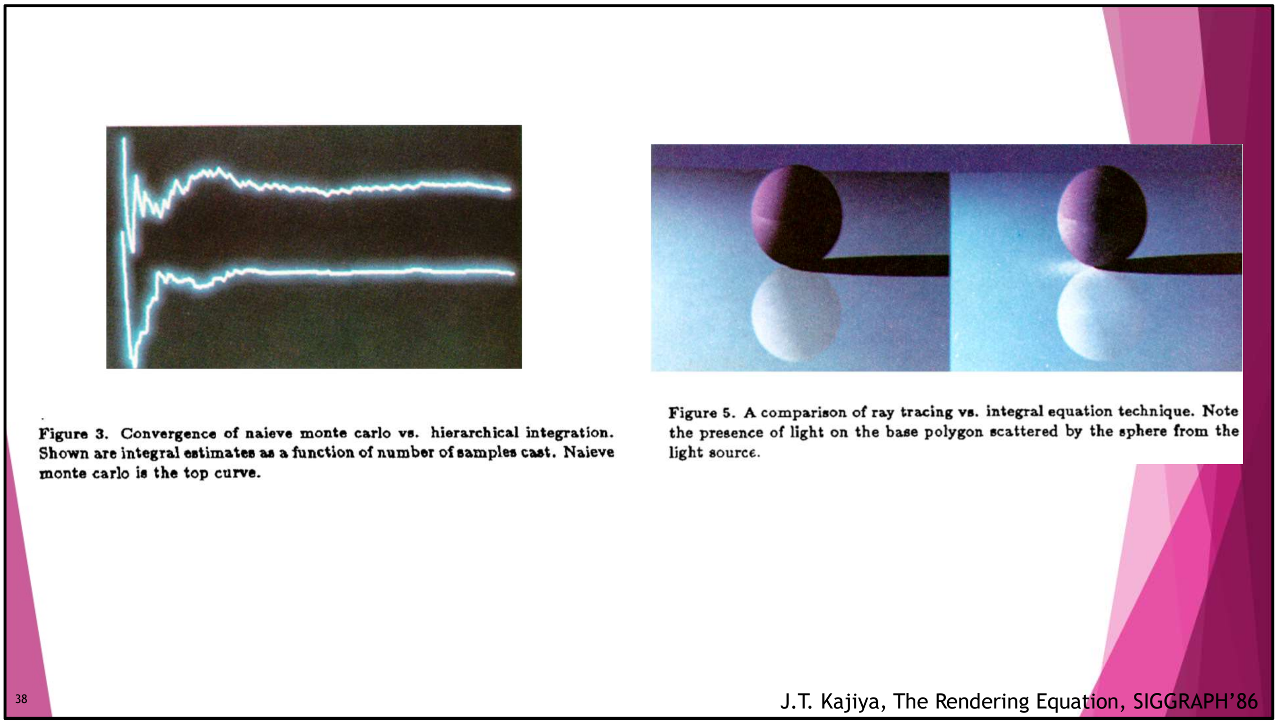


Figure 3. Convergence of naieve monte carlo vs. hierarchical integration. Shown are integral estimates as a function of number of samples cast. Naieve monte carlo is the top curve.

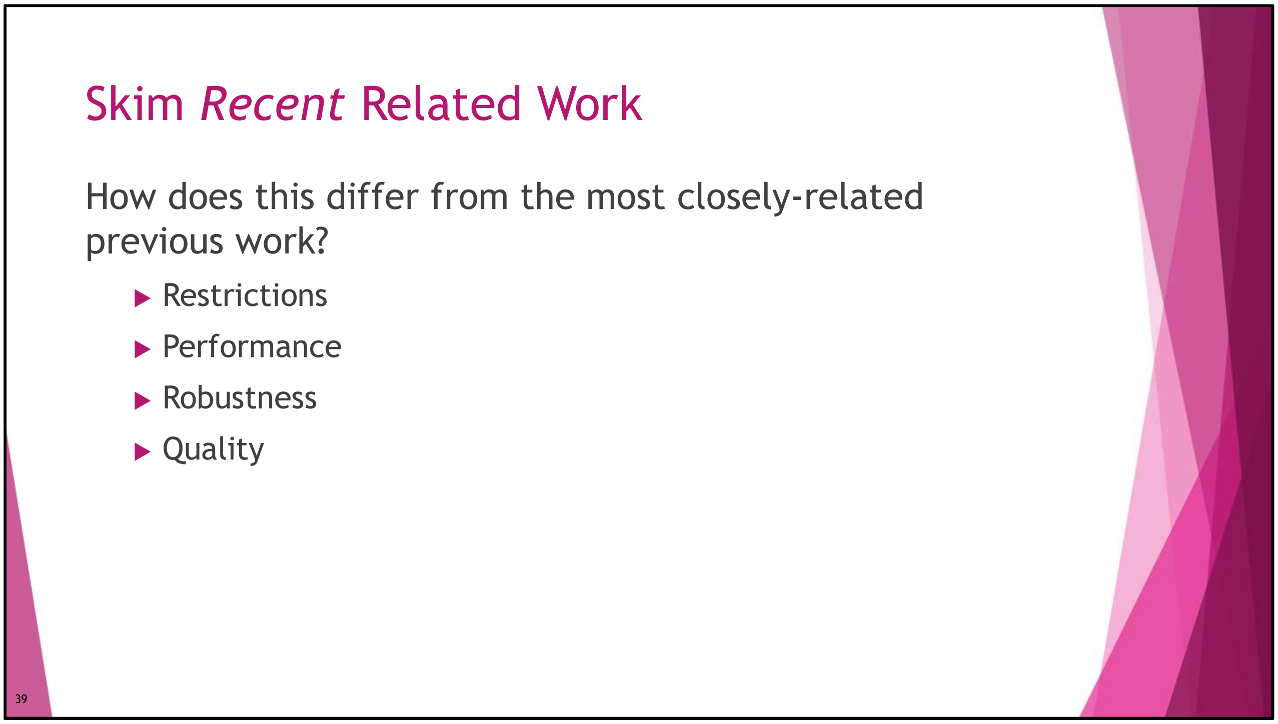
Figure 5. A comparison of ray tracing vs. integral equation technique. Note the presence of light on the base polygon scattered by the sphere from the light source.

38

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

This figure on the left shows two plots of a sequence converging, as value vs. time. The top one is the typical method. The bottom one is the new method. The new method appears to have less wiggle, so that looks good. Especially as it seems to hit the asymptote much earlier in time.

The figure on the right shows a rendered image before and after the technique. The sphere has more realistic lighting in the “after” picture on the right, as well as in the reflection. That’s the more important result from this paper and looks good. There is a clear improvement.

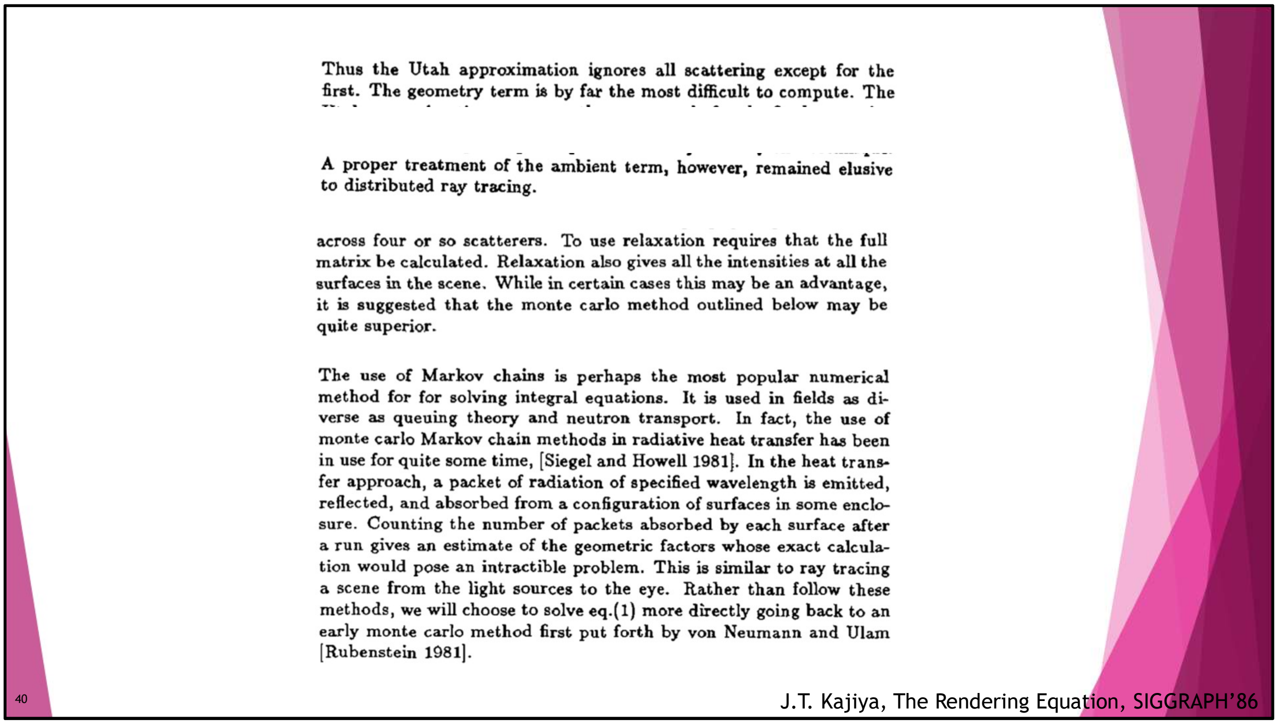


Skim *Recent* Related Work

How does this differ from the most closely-related previous work?

- Restrictions
- Performance
- Robustness
- Quality

39



Thus the Utah approximation ignores all scattering except for the first. The geometry term is by far the most difficult to compute. The

A proper treatment of the ambient term, however, remained elusive to distributed ray tracing.

across four or so scatterers. To use relaxation requires that the full matrix be calculated. Relaxation also gives all the intensities at all the surfaces in the scene. While in certain cases this may be an advantage, it is suggested that the monte carlo method outlined below may be quite superior.

The use of Markov chains is perhaps the most popular numerical method for for solving integral equations. It is used in fields as diverse as queuing theory and neutron transport. In fact, the use of monte carlo Markov chain methods in radiative heat transfer has been in use for quite some time, [Siegel and Howell 1981]. In the heat transfer approach, a packet of radiation of specified wavelength is emitted, reflected, and absorbed from a configuration of surfaces in some enclosure. Counting the number of packets absorbed by each surface after a run gives an estimate of the geometric factors whose exact calculation would pose an intractible problem. This is similar to ray tracing a scene from the light sources to the eye. Rather than follow these methods, we will choose to solve eq.(1) more directly going back to an early monte carlo method first put forth by von Neumann and Ulam [Rubenstein 1981].

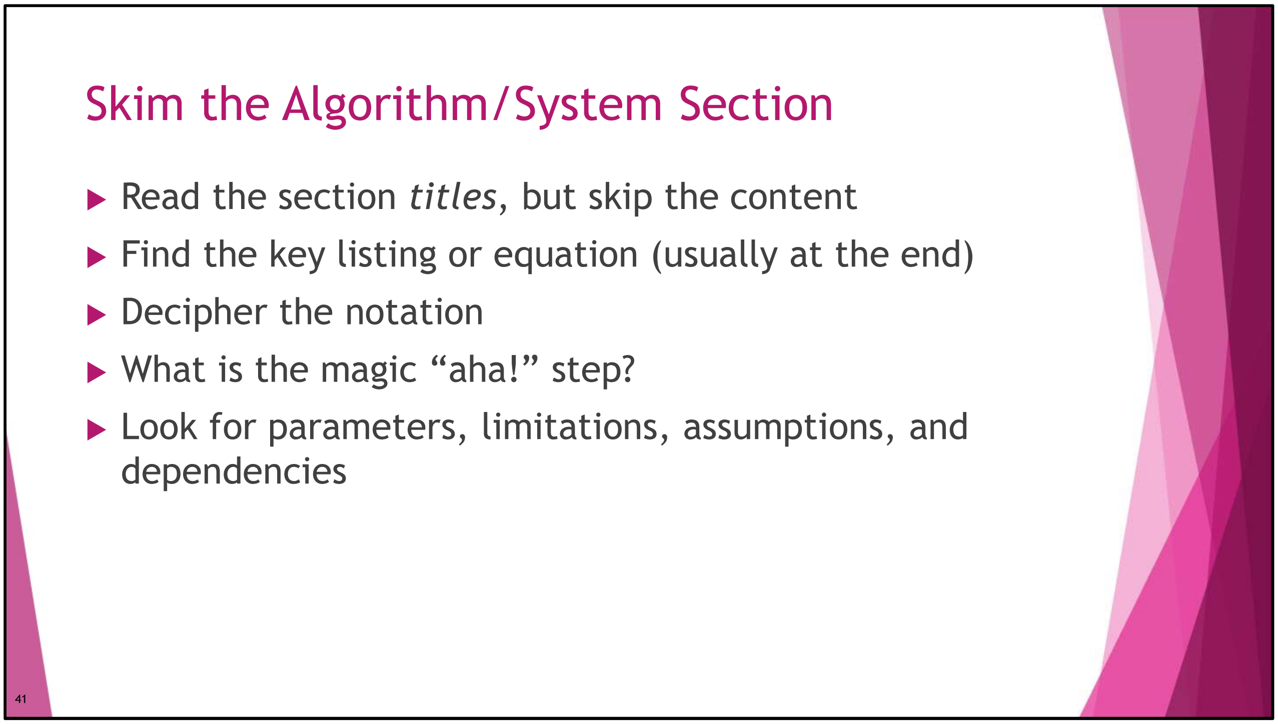
J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

The main related work section is titled “3. Methods for approximate solution” in this paper. The first sentence begins “In this section we shall review…”, which lets you know that.

Raster direct illumination approaches (“Utah”) are direct illumination only. Distribution ray tracing computes a small number of recursive illumination paths, but assumes a hand-tuned “ambient” term for the infinite and diffuse reflections. Radiosity only supports perfectly Lambertian reflectors and is slow.

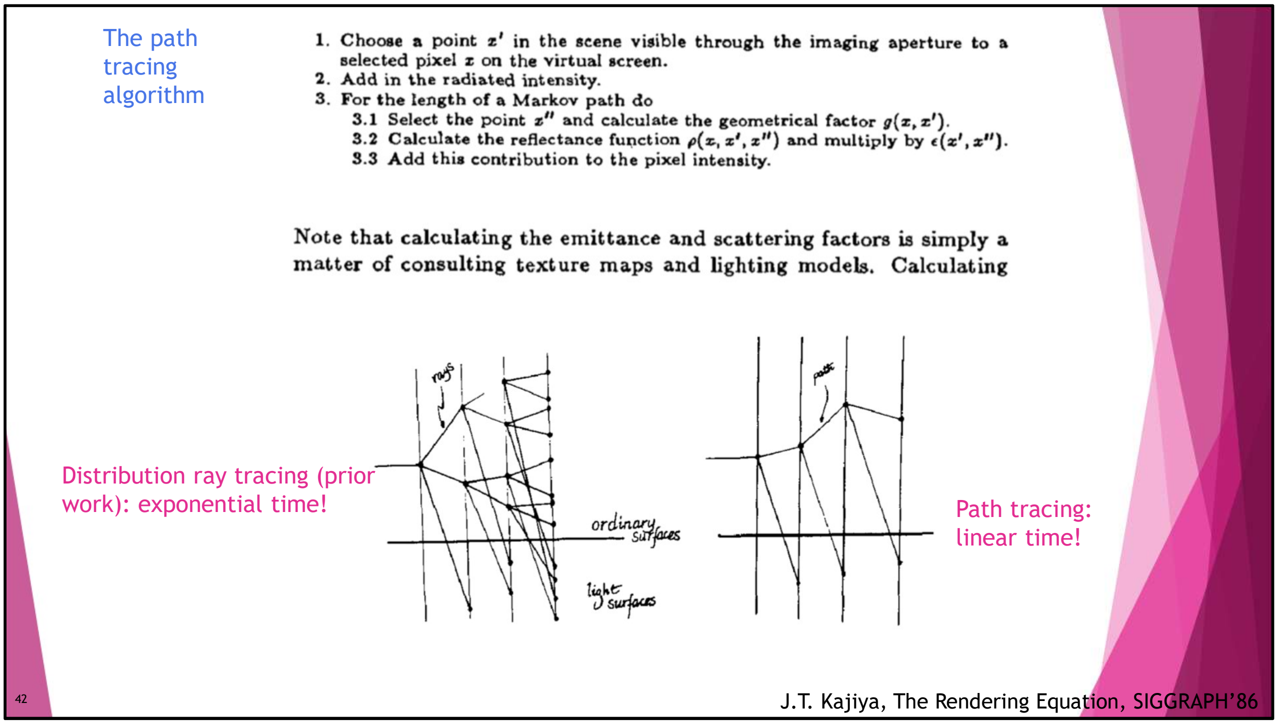
Monte carlo simulation of light paths has been done in other fields, but starts at the light sources. I’m going to more efficiently go from the eye back to the light.

So: general solution, which follows closely on distribution ray tracing but attacks the “diffuse” interreflection term by using a direct monte carlo simulation of light paths from the eye. Not much new to the algorithm, but a new way of thinking about the problem will help it to scale better.



Skim the Algorithm/System Section

- Read the section *titles*, but skip the content
- Find the key listing or equation (usually at the end)
- Decipher the notation
- What is the magic “aha!” step?
- Look for parameters, limitations, assumptions, and dependencies



The path tracing algorithm

```
1. Choose a point $x'$ in the scene visible through the imaging aperture to a
   selected pixel $x$ on the virtual screen.
2. Add in the radiated intensity.
3. For the length of a Markov path do
    3.1 Select the point $x''$ and calculate the geometrical factor $g(x, x')$.
    3.2 Calculate the reflectance function $\rho(x, x', x'')$ and multiply by $\epsilon(x', x'')$.
    3.3 Add this contribution to the pixel intensity.
```

Note that calculating the emittance and scattering factors is simply a matter of consulting texture maps and lighting models. Calculating

Distribution ray tracing (prior work): exponential time!

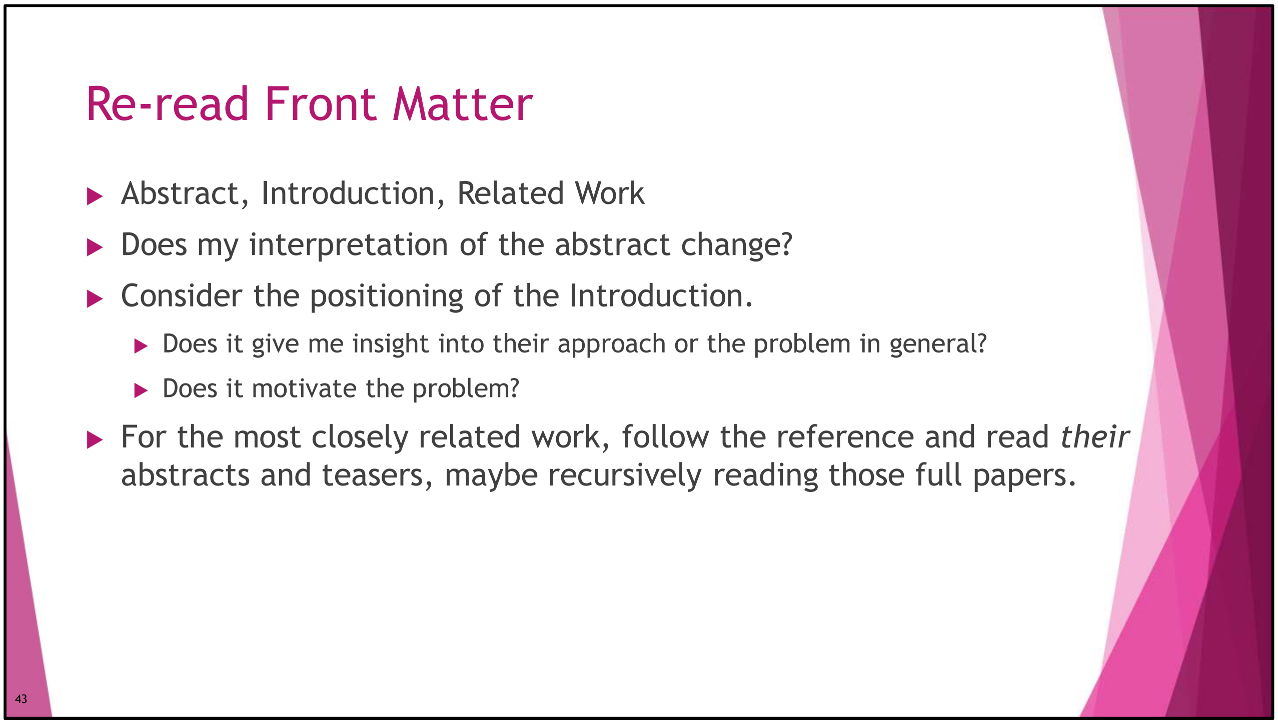
Path tracing: linear time!

J.T. Kajiya, The Rendering Equation, SIGGRAPH'86

The key to understanding this paper is this (hand drawn!) diagram. What it is showing is light transport depicted as a tree search. The horizontal axis is recursive iterations (scattering event depth). The vertical axis is if you took all of the points in the scene and just sorted them into a line, with the lights on the bottom of the line. The diagonals drawn in are a trace of which points the algorithm considers as it iterates horizontally from left (starting at the eye) to right.

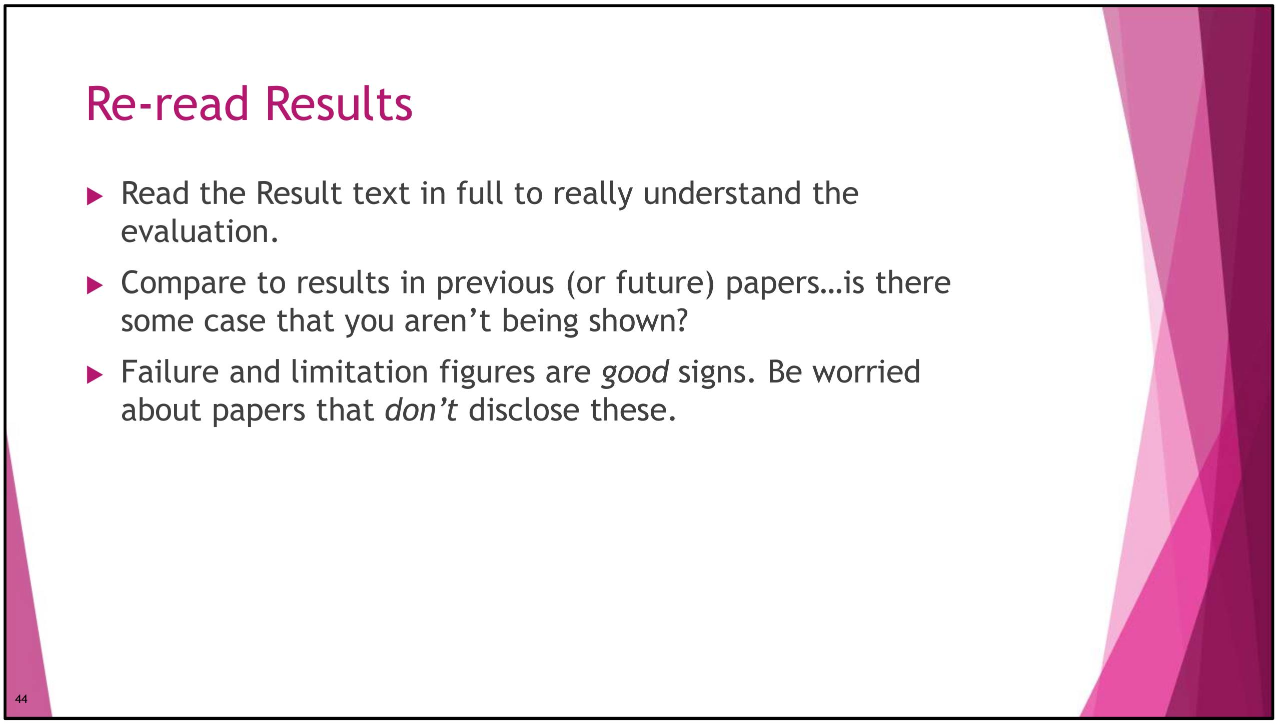
The algorithm at the top describes the same thing. All of the diagrams and equations in section 2 are complicated because Kajiya choose to express the integrals over points instead of angles. You can largely ignore them by just choosing the better (and modern) reference frame/parameterization for the integrals.

This is a brilliant piece of exposition. In one listing and one diagram he presents the core algorithmic result. You just have to stare at them for a week to catch all of the implications—most of the text isn't as important. The core theoretical result was equation 1 on the first page of the paper. That's quite the economy of writing!



Re-read Front Matter

- Abstract, Introduction, Related Work
- Does my interpretation of the abstract change?
- Consider the positioning of the Introduction.
  - Does it give me insight into their approach or the problem in general?
  - Does it motivate the problem?
- For the most closely related work, follow the reference and read *their* abstracts and teasers, maybe recursively reading those full papers.



Re-read Results

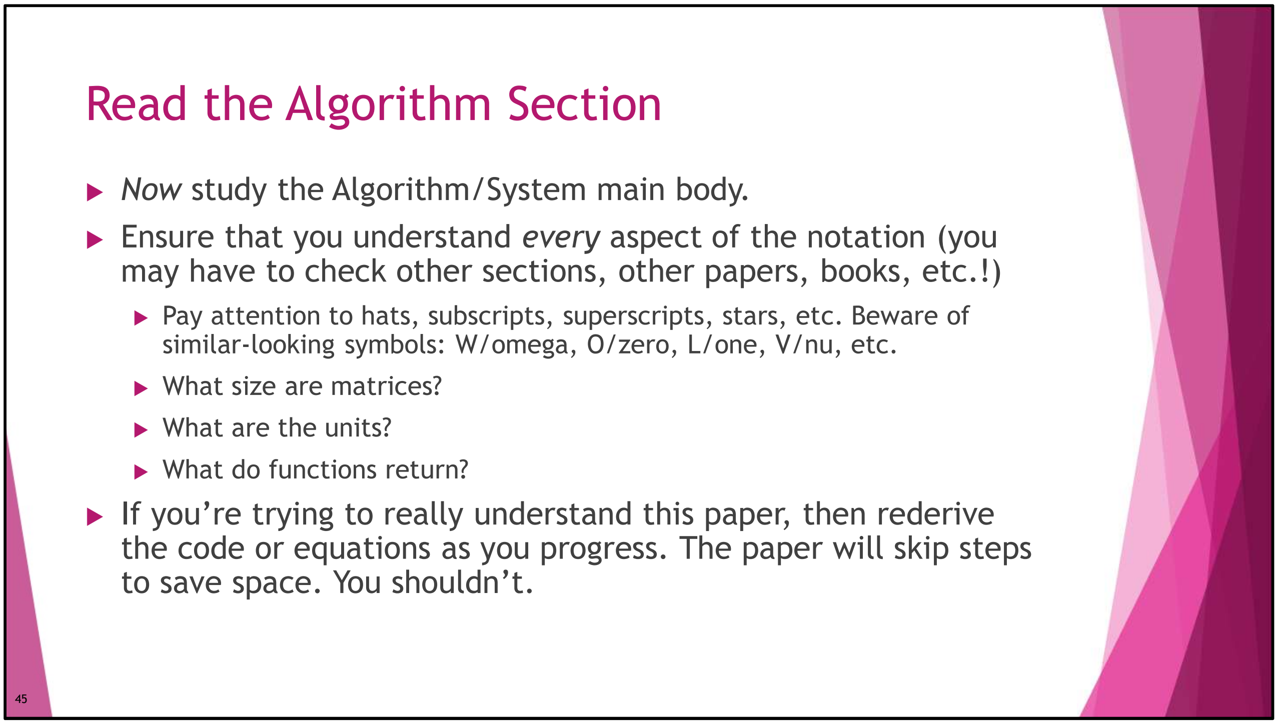
- Read the Result text in full to really understand the evaluation.
- Compare to results in previous (or future) papers…is there some case that you aren’t being shown?
- Failure and limitation figures are *good* signs. Be worried about papers that *don’t* disclose these.

44

Note that Kajiya was very up front about the limitations; on page 1 he told us all of the cases that the algorithm couldn’t handle. The result figures that we already considered showed nice comparisons to previous work. In the results section he also describes the performance in wall-clock time on an IBM 3081 computer, and he showed us graphically that the new solution is linear instead of exponential in path length.

He didn’t discuss or render any failure cases, but it took him 20 hours to render the two 512x512 result pictures which he had to photograph with a film camera because screenshots and digital typesetting didn’t yet exist, so you can probably understand why he stopped at two images.

Yet, that is a shortcoming of the paper…the difficulty of handling complicated chains of specular reflections from the light ending in a diffuse reflection into the eye (caustics) and very narrow apertures (such as a keyhole) is what motivates the great bidirectional path tracing, metropolis light transport, and photon mapping papers a decade later. Perhaps those solutions would have been invented sooner had Kajiya disclosed this problem.



Read the Algorithm Section

- *Now* study the Algorithm/System main body.
- Ensure that you understand *every* aspect of the notation (you may have to check other sections, other papers, books, etc.!)
  - Pay attention to hats, subscripts, superscripts, stars, etc. Beware of similar-looking symbols: W/omega, O/zero, L/one, V/nu, etc.
  - What size are matrices?
  - What are the units?
  - What do functions return?
- If you're trying to really understand this paper, then rederive the code or equations as you progress. The paper will skip steps to save space. You shouldn't.

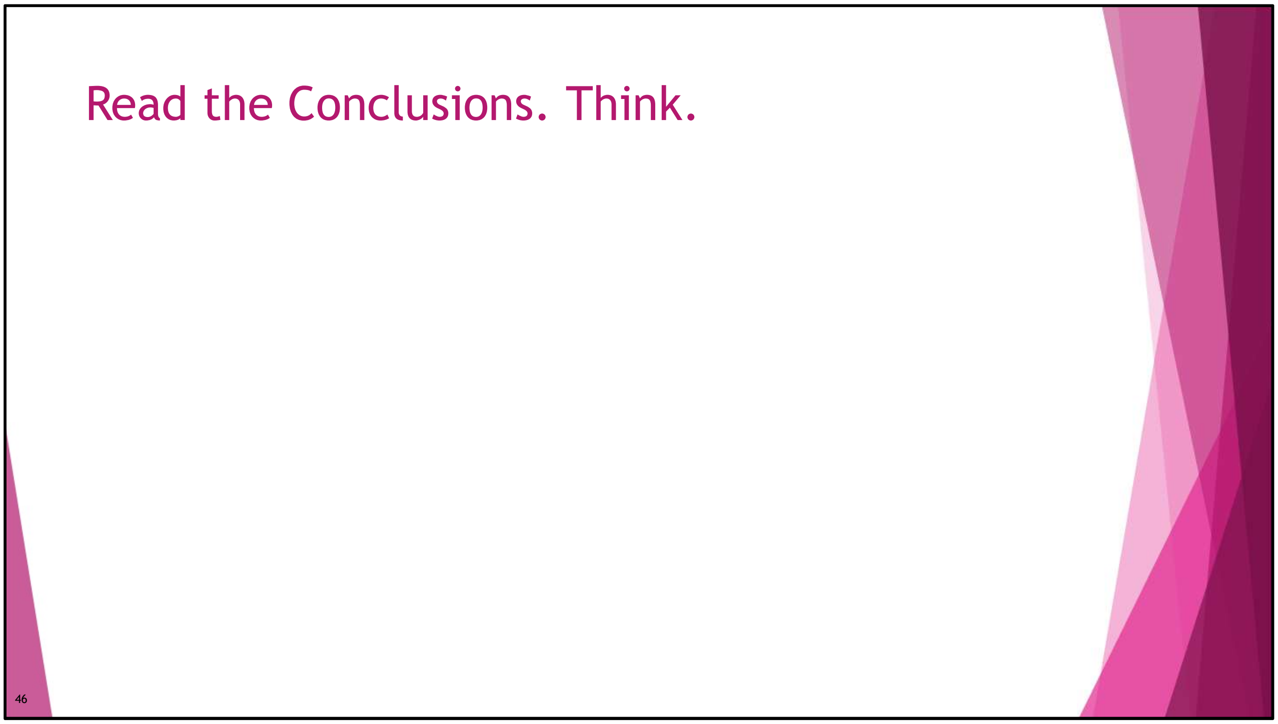
If you are going to implement the paper, publish work that directly follows from it, or teach a detailed lecture on it, then you need to rederive the key results.

If you're presenting the paper at a high level (as we do in this course), surveying it as tangential related work, or just learn about the big ideas, then you can skip that step.

I often find that by rederiving results I discover hidden gems. Maybe the paper (especially if it is a bit older) describes a contribution that is not directly relevant to the problem I'm working on…but a clever technique that they used to build their solution might be a great and reusable tool.

Recall that the invention of the z-buffer—*the* dominant real-time visibility method--was described in one paragraph in the *second* appendix of Sutherland et al.'s "A Characterization of Ten Hidden-Surface Algorithms". They didn't even count it among the ten algorithms. Ed Catmull also described and named in his PhD thesis, published shortly thereafter. His thesis was about curves and the one and a half pages he spends on the z-buffer are mostly apologizing for the fact that it uses too much space and produces aliasing. Authors don't always know what is going to stand the test of time or be a major contribution. The z-buffer of course is no longer a hidden gem.

But, I wonder what is.



Read the Conclusions. Think.

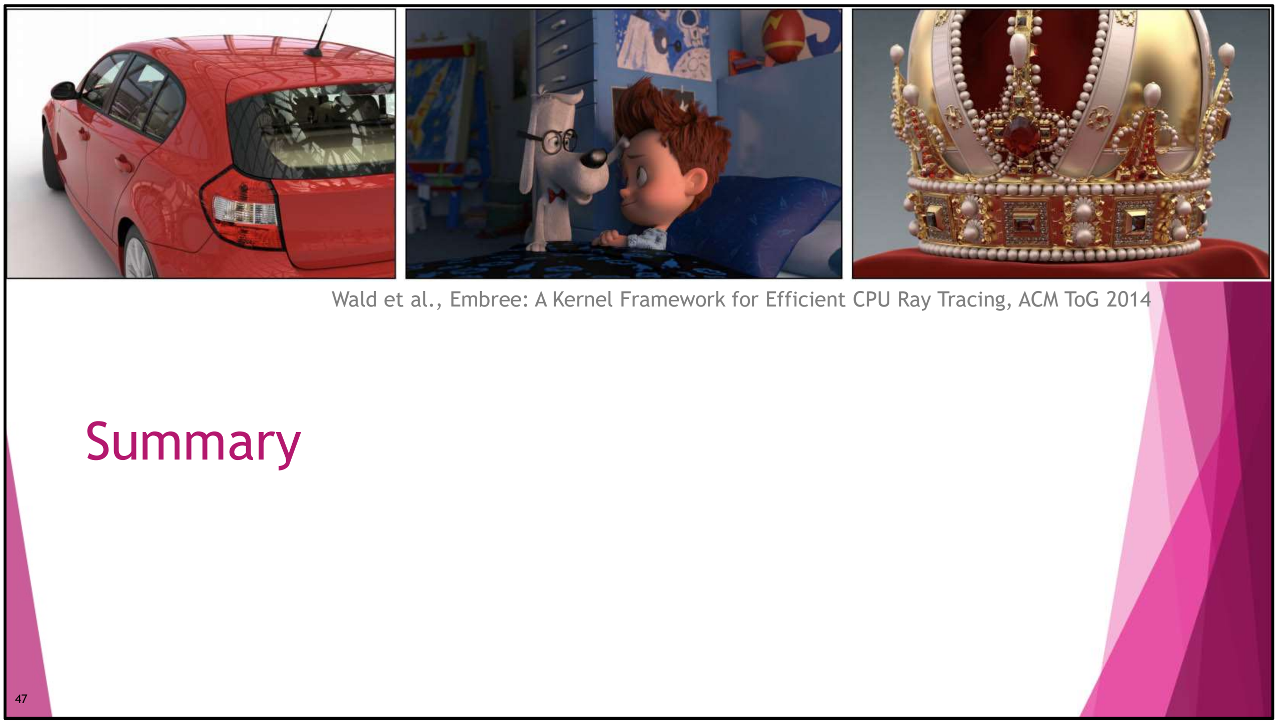
46

Once you really understand a paper, the introduction and discussion sections can become the most important parts.

Unless this is provably the best possible result, the paper is likely describing a solution that will some day be obsolete.

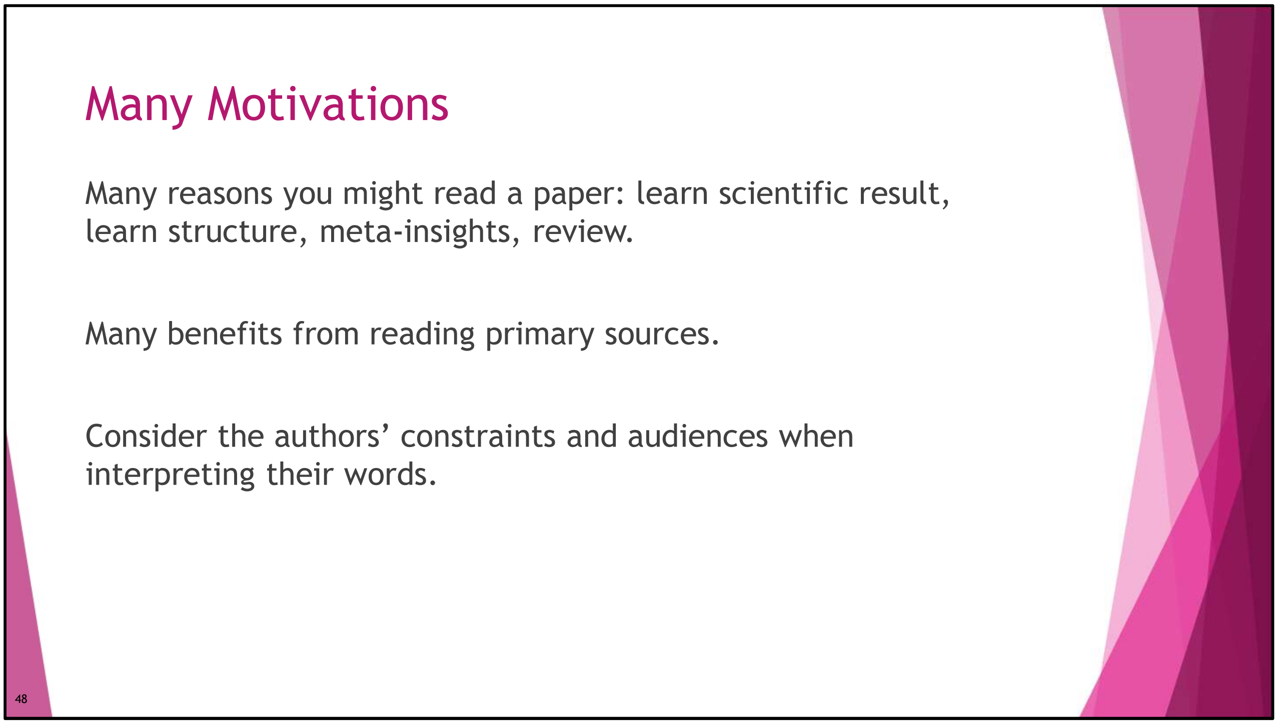
But the approach that the authors took, the techniques that they leveraged, and the insights that they gained may be perpetually useful.

For the very best papers, the core algorithm and results that previously seemed so important may only be a vehicle for reaching a newly enlightened way of looking at the problem and the field. For rendering, I count Distributed Ray Tracing, The Rendering Equation, and Metropolis Light Transport papers among those where the big idea remains essential and beautiful even though the original algorithm and results are now obsolete.



Wald et al., Embree: A Kernel Framework for Efficient CPU Ray Tracing, ACM ToG 2014

Summary

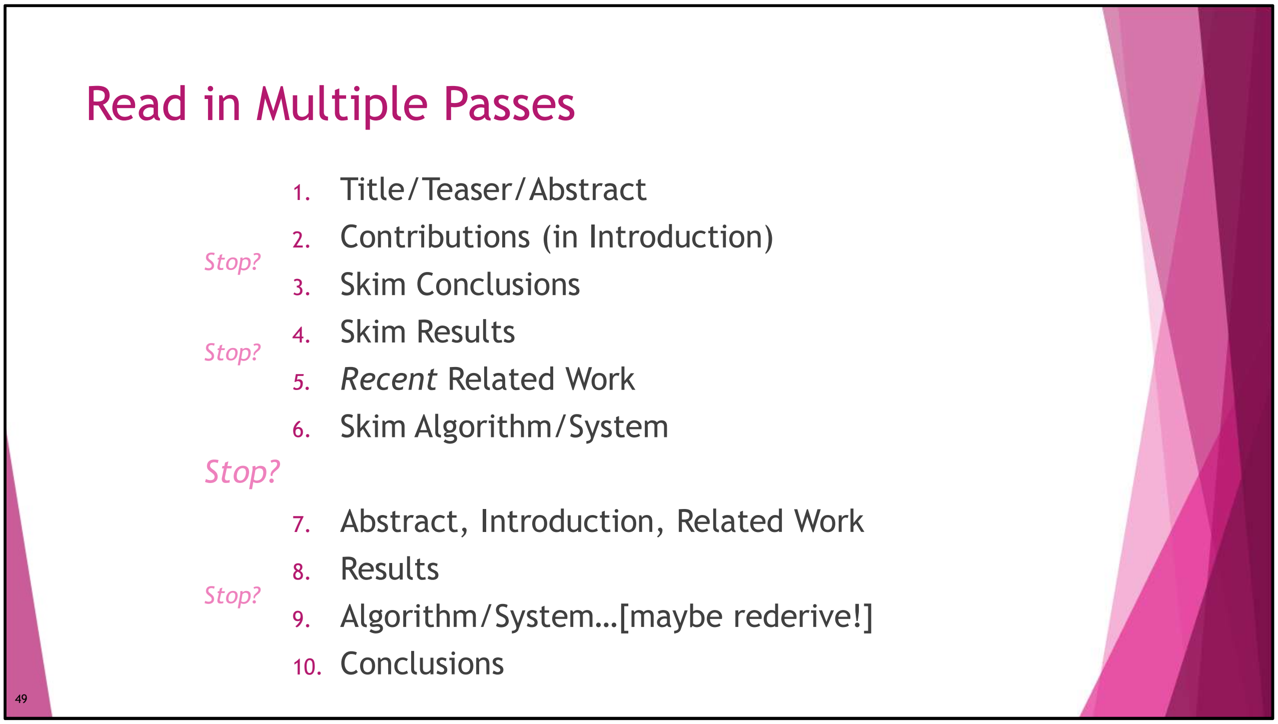


Many Motivations

Many reasons you might read a paper: learn scientific result, learn structure, meta-insights, review.

Many benefits from reading primary sources.

Consider the authors' constraints and audiences when interpreting their words.



Read in Multiple Passes

1. Title/Teaser/Abstract
2. Contributions (in Introduction)

Stop?

3. Skim Conclusions
4. Skim Results

Stop?

5. *Recent* Related Work
6. Skim Algorithm/System

Stop?

7. Abstract, Introduction, Related Work
8. Results

Stop?

9. Algorithm/System…[maybe rederive!]
10. Conclusions

49

Here’s the iterative process of how I read a research paper all on one slide.

Remember that I’m periodically deciding whether to continue.
I most frequently stop after step 6, having determined that this isn’t the paper I need to read right now.