Physically Based Shading

Naty Hoffman
Vice President of Technology, 2K

Hi. Thanks for coming to my talk. During the next hour, I'll talk about physically based shading.

Easier Realism

Robustness

There are two main advantages to physically based shading models. One is the ease of achieving realism; the parameters of the material models fit more clsoely to the properties of real materials so it's easier for artists to author real-looking materials. The other is robustness; if an object looks good in one lighting environment, it will look good in others and less tweaking needs to be done.

Myth: Physically Based Shading Restricts the Art Style

A common misconception about physically based shading is that it restricts the visual style and all movies / games using it will look similar. It's true that if you're going for a painterly or hand-drawn style, physically-based shading is not very useful. But it's not nearly as constricting as you might think.

The thing is – live action movies have physically based shading, right? They can’t change their shading models, they are stuck with reality. But nobody would say that all live action movies look the same! Whether it’s by set & costume design, lighting, film stock, post processing, or all of the above, directors bend “reality” to their distinct creative visions.

Images from the films Apocalypse Now, 300, The Third Man, Amélie, Hero, Heaven’s Gate, The Wizard of Oz, Sin City, The Godfather used for scholarly commentary under fair use.

It’s true that PBS enables filmmakers to produce extremely photorealistic imagery…

Image from the film The Avengers used for scholarly commentary under fair use.

…the two guys on the right are CG, but they fit in just fine next to the actors. ILM was one of the first visual effects companies to embrace physically based shading, having used it since 2007.

But physically based shading can be used to create cartoony looks…

Image from the film Cloudy with a Chance of Meatballs used for scholarly commentary under fair use.

…and has been used this way by various CG animation studios.

Image from the film Wreck-It Ralph used for scholarly commentary under fair use.

In fact, with physically based shading you can mix and match gritty realism…

Image from the film WALL-E used for scholarly commentary under fair use.

…with cartoony stylization, in the same movie…

Image from the film WALL-E used for scholarly commentary under fair use.

…even in the same scene.

Image from the film WALL-E used for scholarly commentary under fair use.

An interesting trend in CG animation is toward combining extremely stylized shapes with very realistic looking materials and lighting…

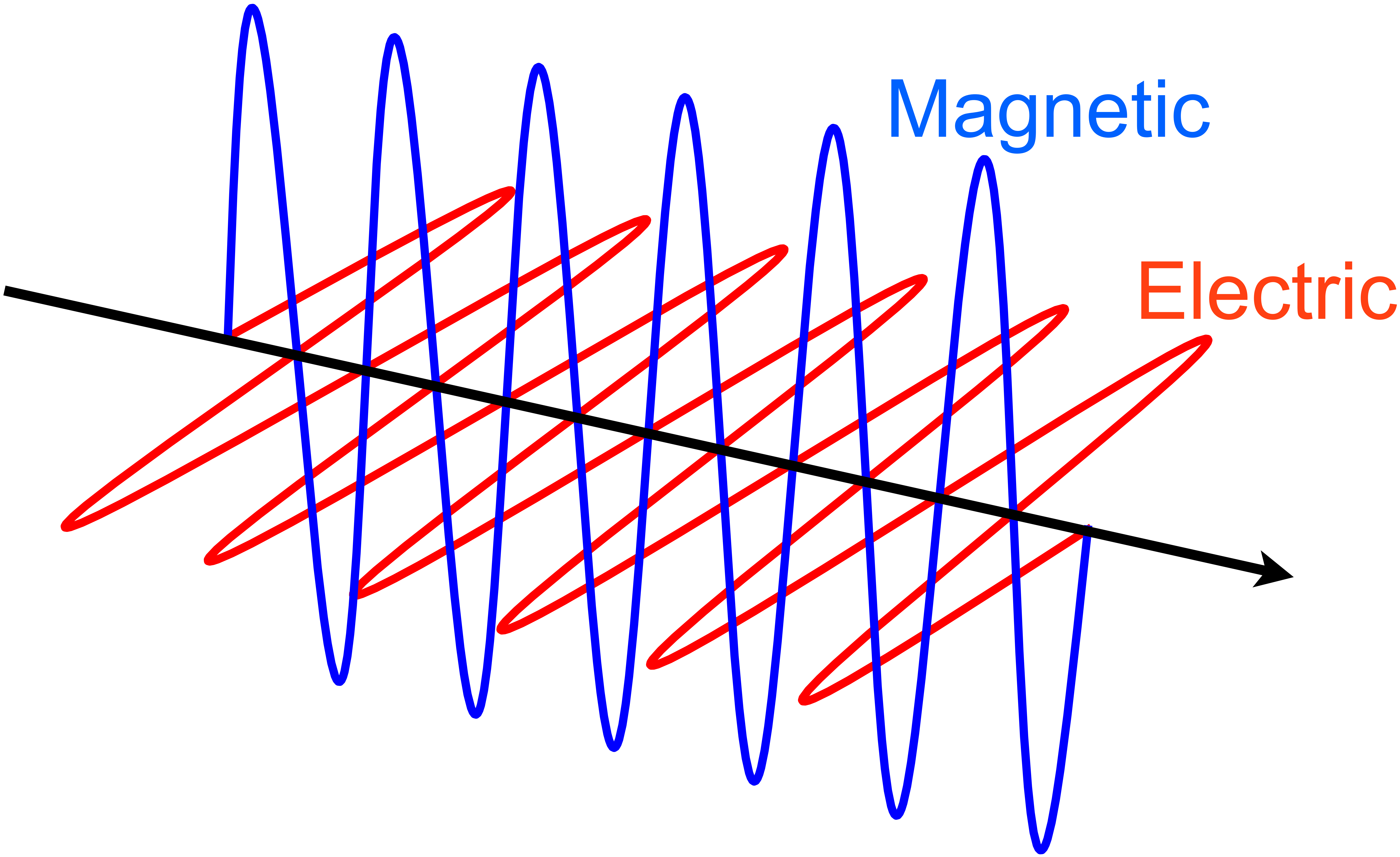
Image from the film Rango used for scholarly commentary under fair use.

…and this seems to be where most of the animation industry is going.

Image from the film Monsters University used for scholarly commentary under fair use.

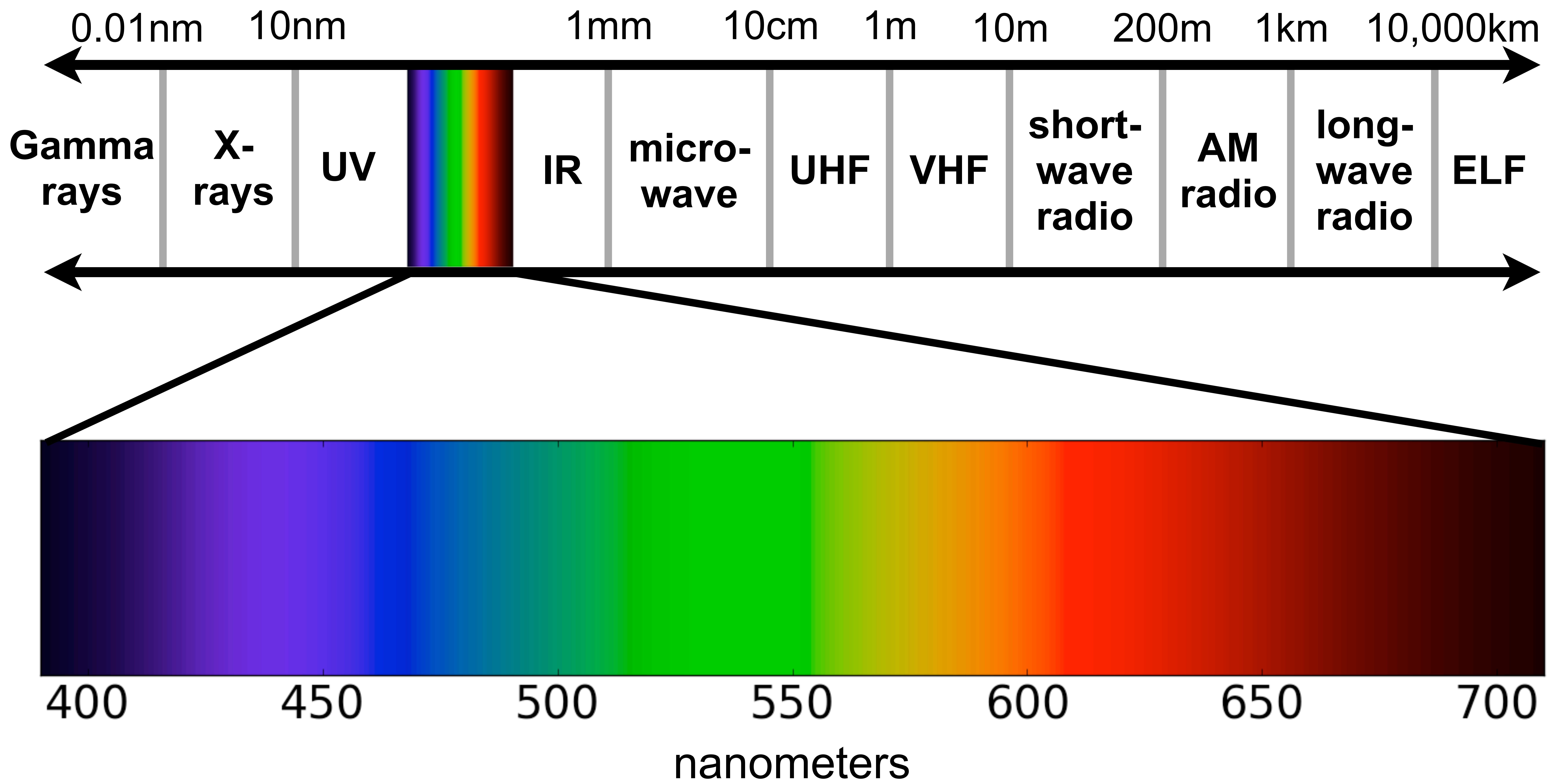
Physics and Math of Shading

For the rest of the talk, I'll be going from the physics underlying shading to the math used to describe it.



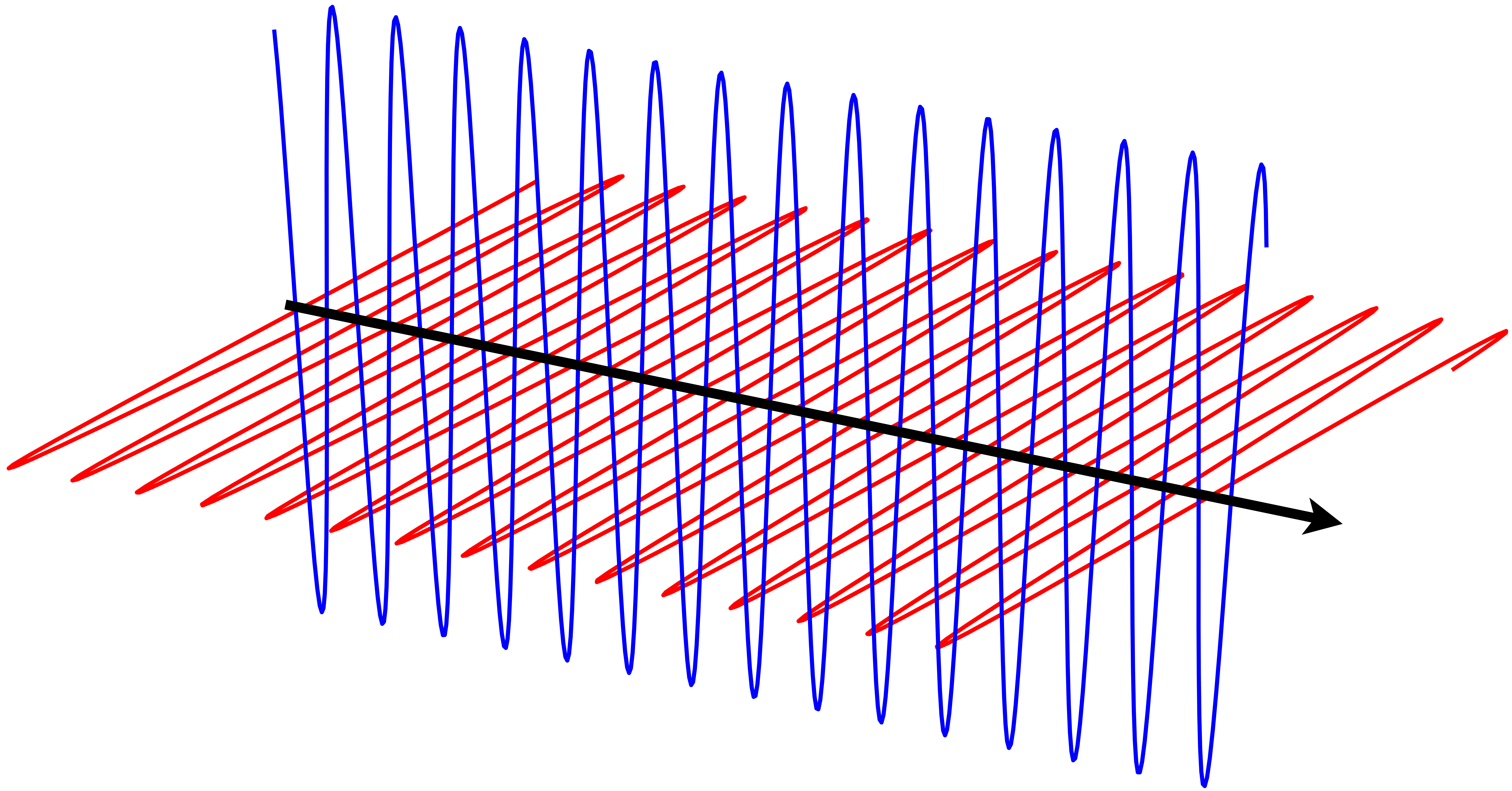
So what IS light, from a physics standpoint? It's technically an electromagnetic transverse wave, which sounds very fancy but actually means that the electromagnetic field wiggles sideways as the energy propagates forwards. This wiggling in the electromagnetic field can be seen as two coupled fields, electric and magnetic, wiggling at 90 degrees to each other.

Electromagnetic waves can be characterized by frequency (the number of wiggles they do in a second) or wavelength (the distance between two wave peaks).



Engineers in various disciplines have to deal with electromagnetic wavelengths that range from gamma waves less than a hundredth of a nanometer in length, to extreme low frequency radio waves tens of thousands of kilometers long—and everything in between. But the range that we can actually see with our eyes is a tiny, tiny subset of that range, extending from 400 nanometers for violet light to 700 nanometers for red light.

To give you a bit of intuition for what 400 to 700 nanometers actually is (the physical size of these waves will become relevant later in the talk), on the left you can see visible light wavelengths relative to this grey cylinder. That's a single strand of spider silk, which is about 1 micron in width. And on the right, to give some extra context, you see that same strand of spider silk relative to the width of a human hair. So visible light wavelengths are very tiny but not unimaginably so; you can see a strand of spider silk with the naked eye and visible light wavelengths are around 1/2 to 1/3 of that width.



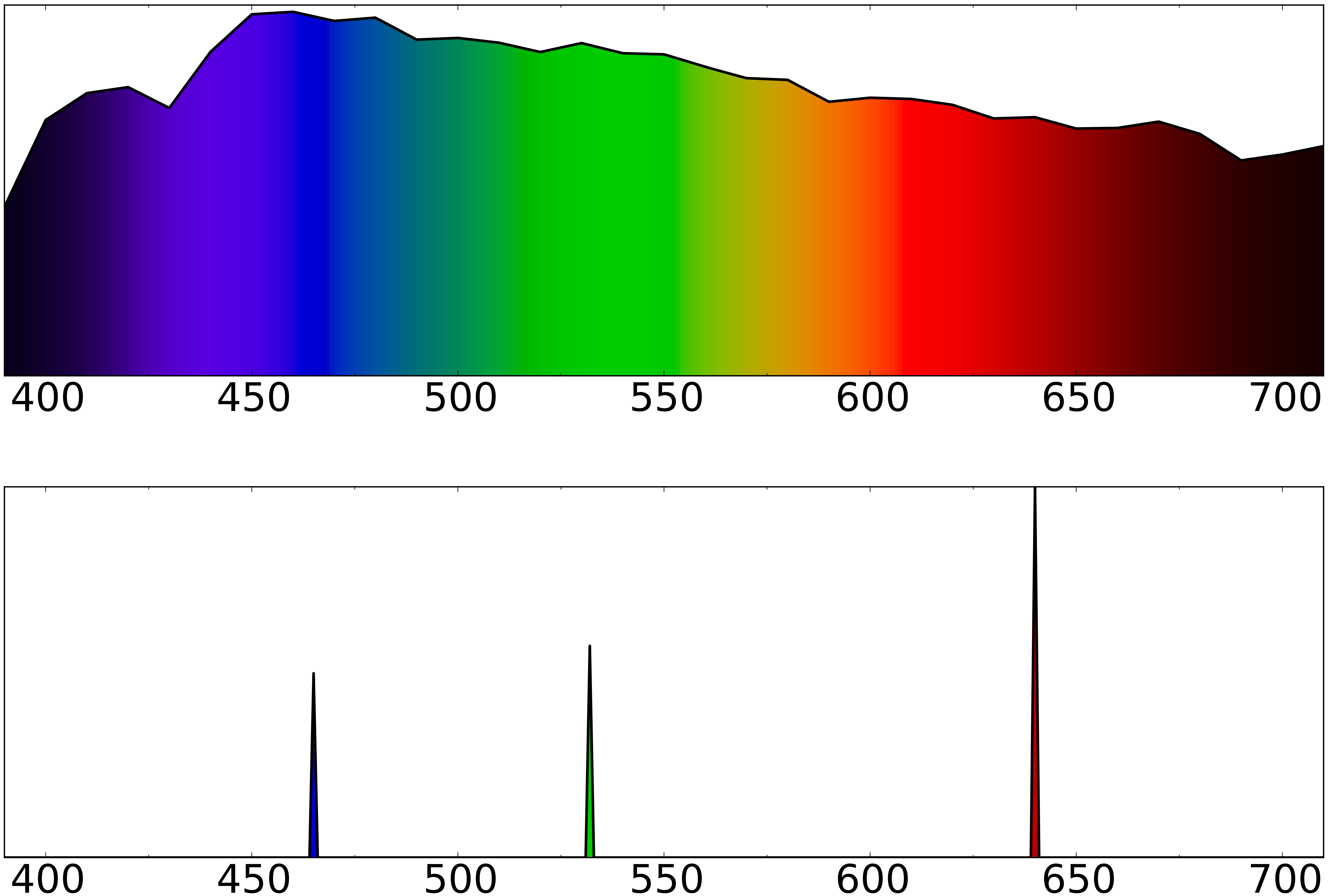
So far what I’ve shown you are simple sine waves that have a single unique wavelength. This is the simplest possible type of light wave, but it is far from the most common.

Most light waves contain many different wavelengths, with a different amount of energy in each. This is typically visualized as a spectral power distribution (SPD for short), as seen in the upper left. It shows that this wave's energy is all in a single wavelength, in this case in the green part of the spectrum. This is typical of light emitted by a laser; it's SPD is sort of a Dirac delta function (in actuality, lasers have a little bit of bandwidth, but they are extremely narrow).

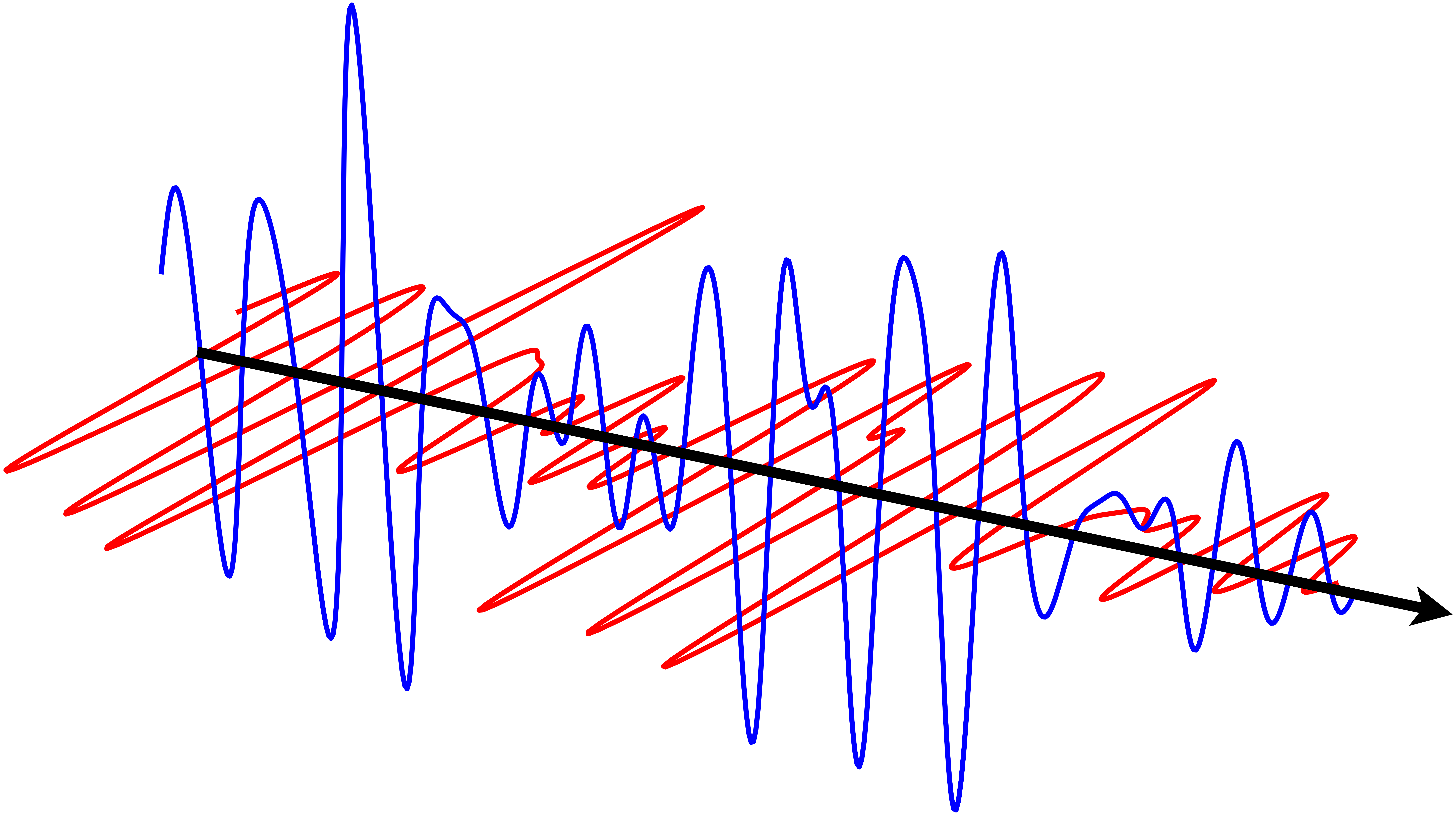
Here we see the SPDs for red, green and blue lasers that are each multiplied by a factor and added together to produce the SPD on the right. This kind of spectral power distribution is similar to what you would see light from a laser projector. Laser projectors are starting to show up in higher-end movie theaters; they offer high contrast and a high color gamut. This is an extremely spiky spectrum.

If we look at the resulting waveform, we can see that after adding these three simple sine waves we end up with something that maybe looks a little more complicated than a single sine wave, but not that much more. There's a little more detail—an underlying frequency, and some larger beats.

But most light you will see in nature doesn't look like that; the SPDs are more like this one. This is the SPD for D65, a standard spectrum for white outdoors light. And its waveform is very complex. In general, the broader the SPD is, the more chaotic-looking the waveform will be, and the less it will resemble a simple sine wave.

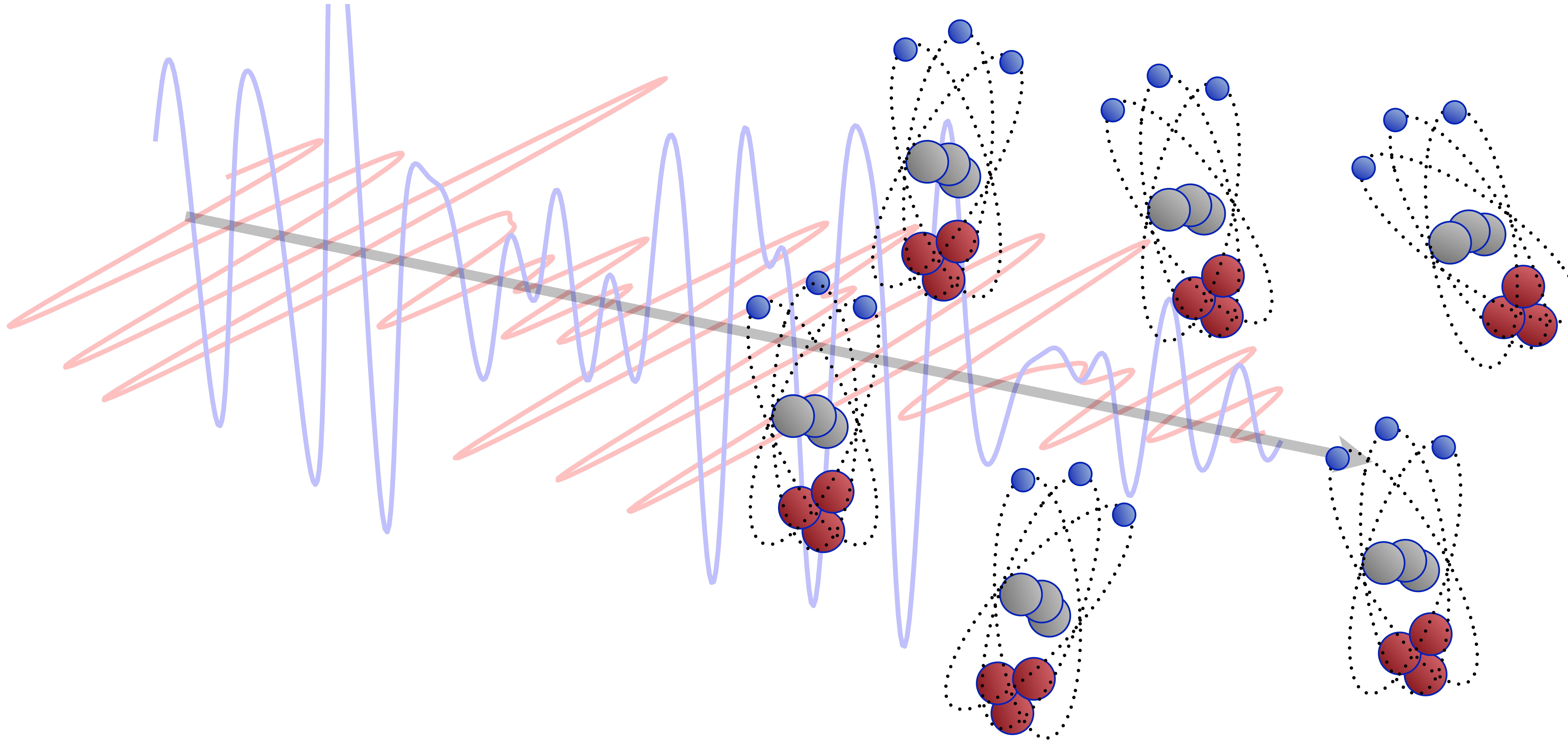


Now the interesting thing is that although these two SPDs couldn't be more different—one is smooth and broad, the other is three delta functions—they have the exact same color appearance to humans (note that the y-axis isn't to the same scale). The fact that human vision cannot distinguish between these two very different signals, tells us that human color vision is <u>incredibly</u> lossy. It maps the infinite-dimensional SPD down to a mere 3-dimensional perceptual space.

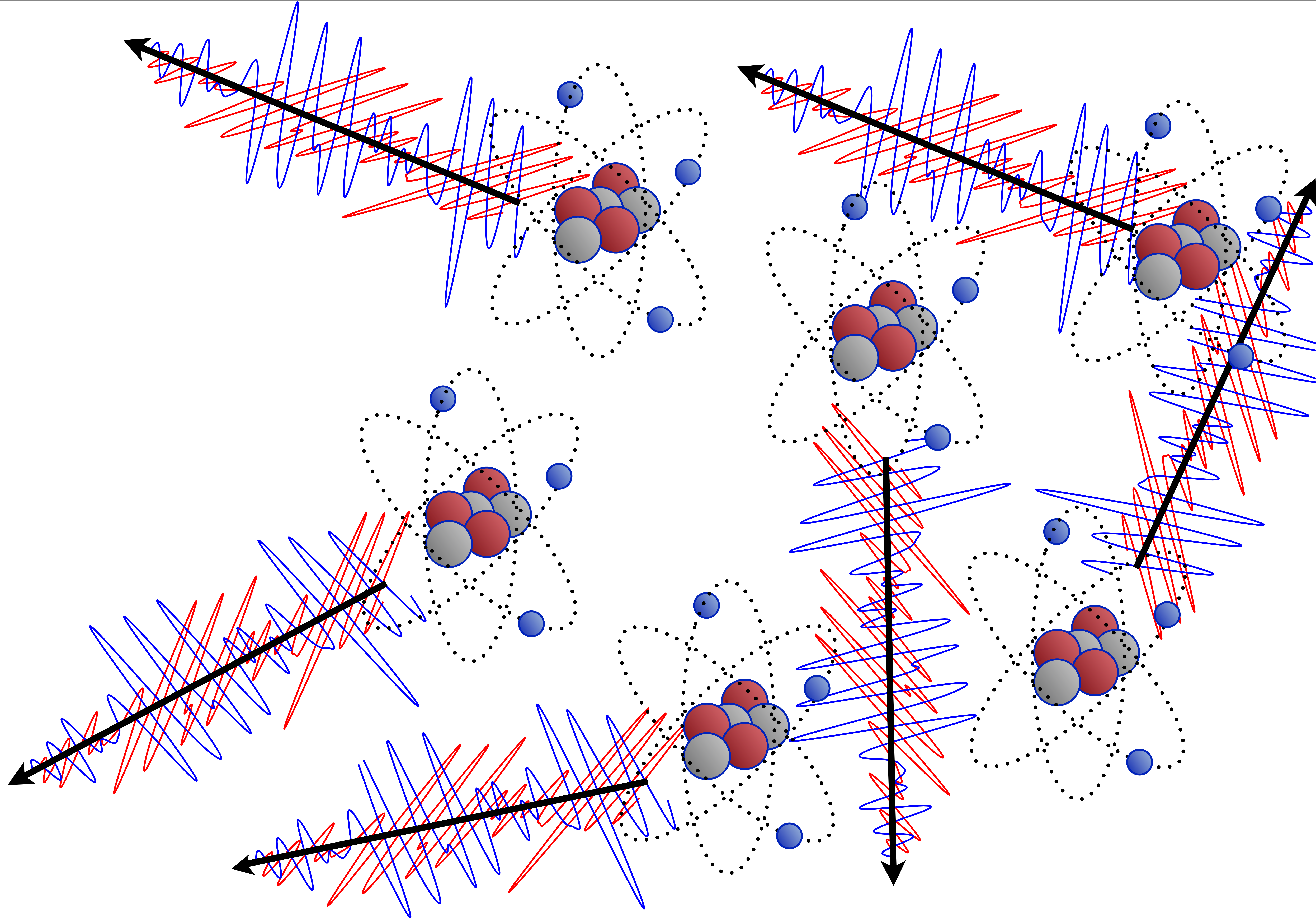


If we look at this waveform in vacuum, it will just continue to propagate forever; the electric and magnetic waves will reinforce each other and it will just keep on going and going. But for rendering what we care about is what happens when this light interacts with matter.

What happens when an electromagnetic wave hits a bunch of atoms or molecules, is that it *polarizes* them.



That means that it stretches and separates the molecules' positive and negative charges, forming *dipoles*. This process absorbs energy from the incoming wave...

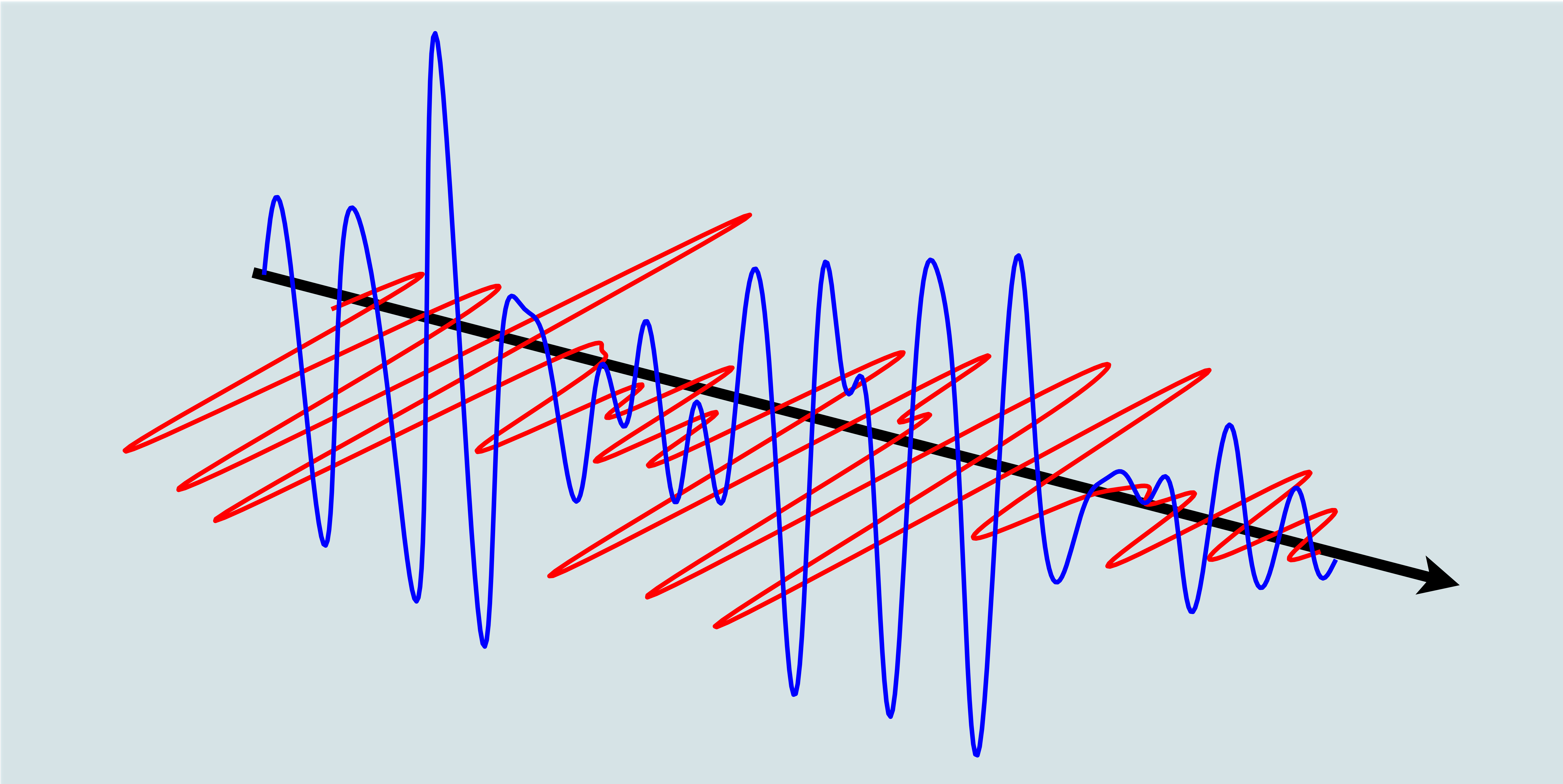


...and this energy is radiated back outwards as the stretched molecules “snap back” like tiny springs (some of the energy may be lost to heat in the process). The result is a bunch of new waves radiating in all directions. In a thin gas, the molecules are far apart and can be treated individually; in this case there are relatively simple physical and mathematical formulations that can be used to understand what’s happening. In other cases (dense gases, liquids, solids) the dipoles interact with each other, and the waves emitted by them also interfere with each other. The whole thing is—in the general case—much too complex to accurately simulate.

Physical (Wave) Optics

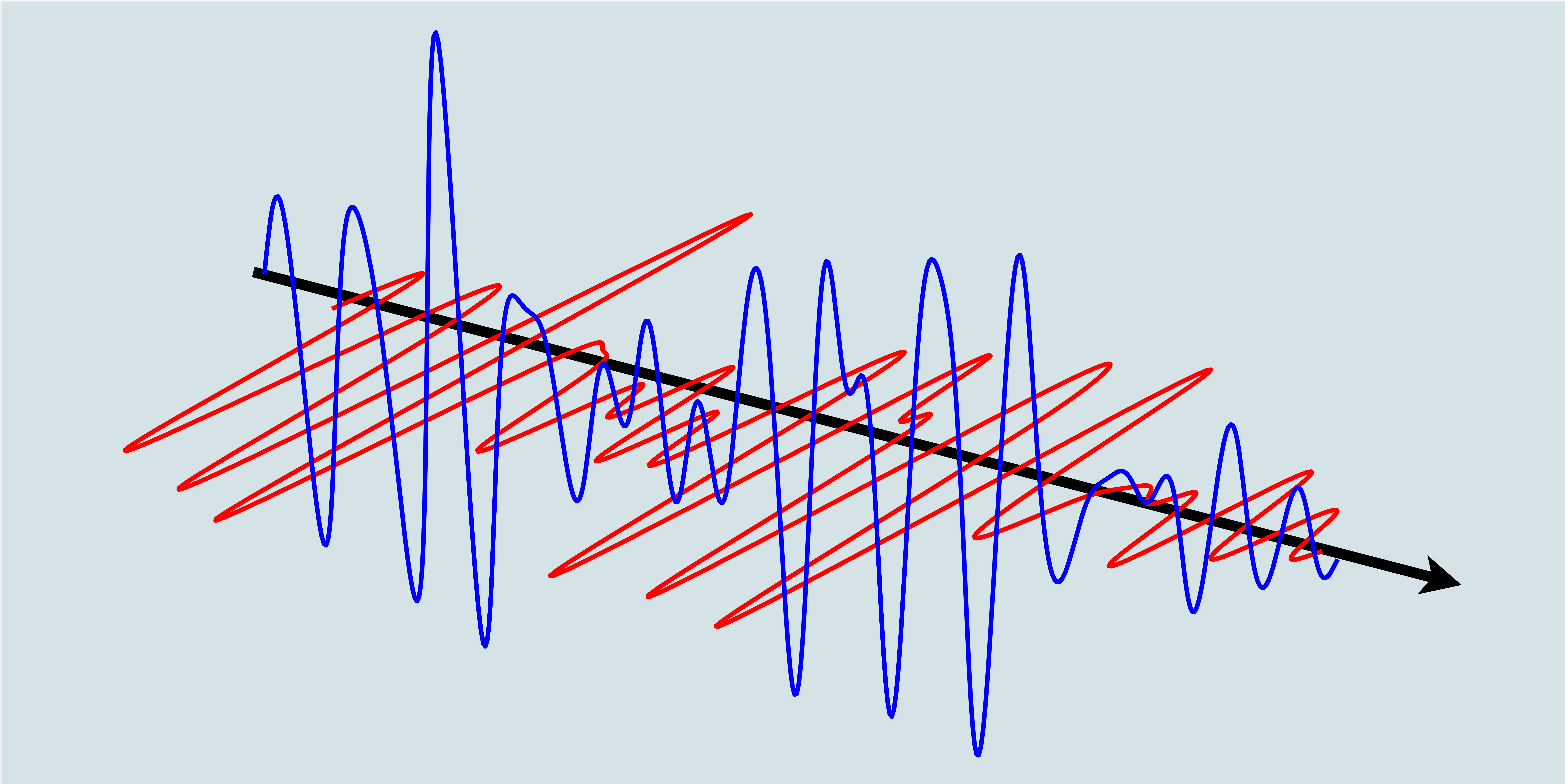
The science of optics (in this case physical, or wave, optics), in order to tame this chaotic situation, adopts certain abstractions, simplifications and approximations.

Homogeneous Medium



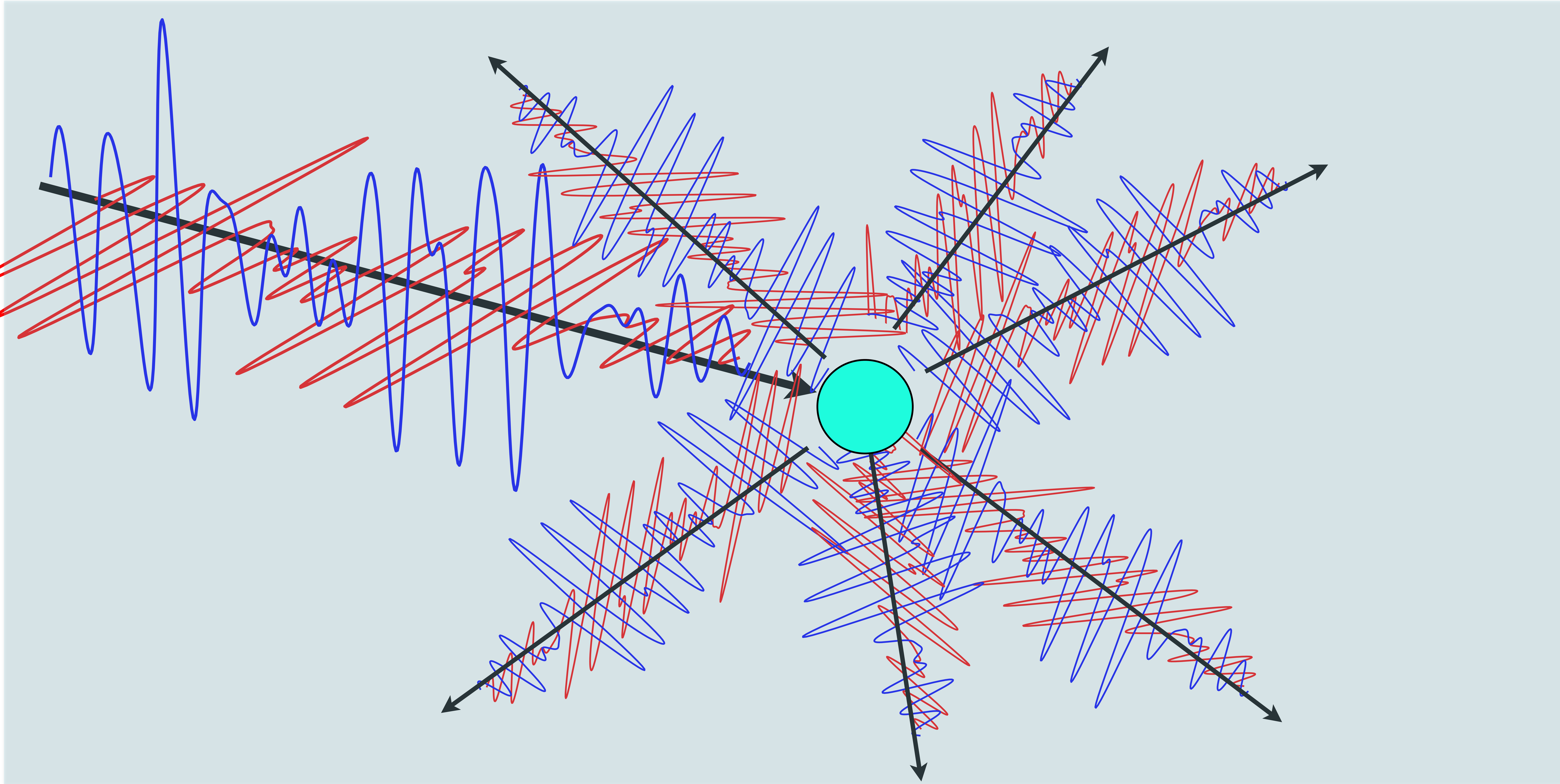
For example, optics has the concept of a *homogeneous medium*, through which light travels in a straight line. This is an abstraction; matter composed of atoms can never be truly homogeneous at all scales. But in practice it works well for materials with uniform density and composition.

Index of Refraction

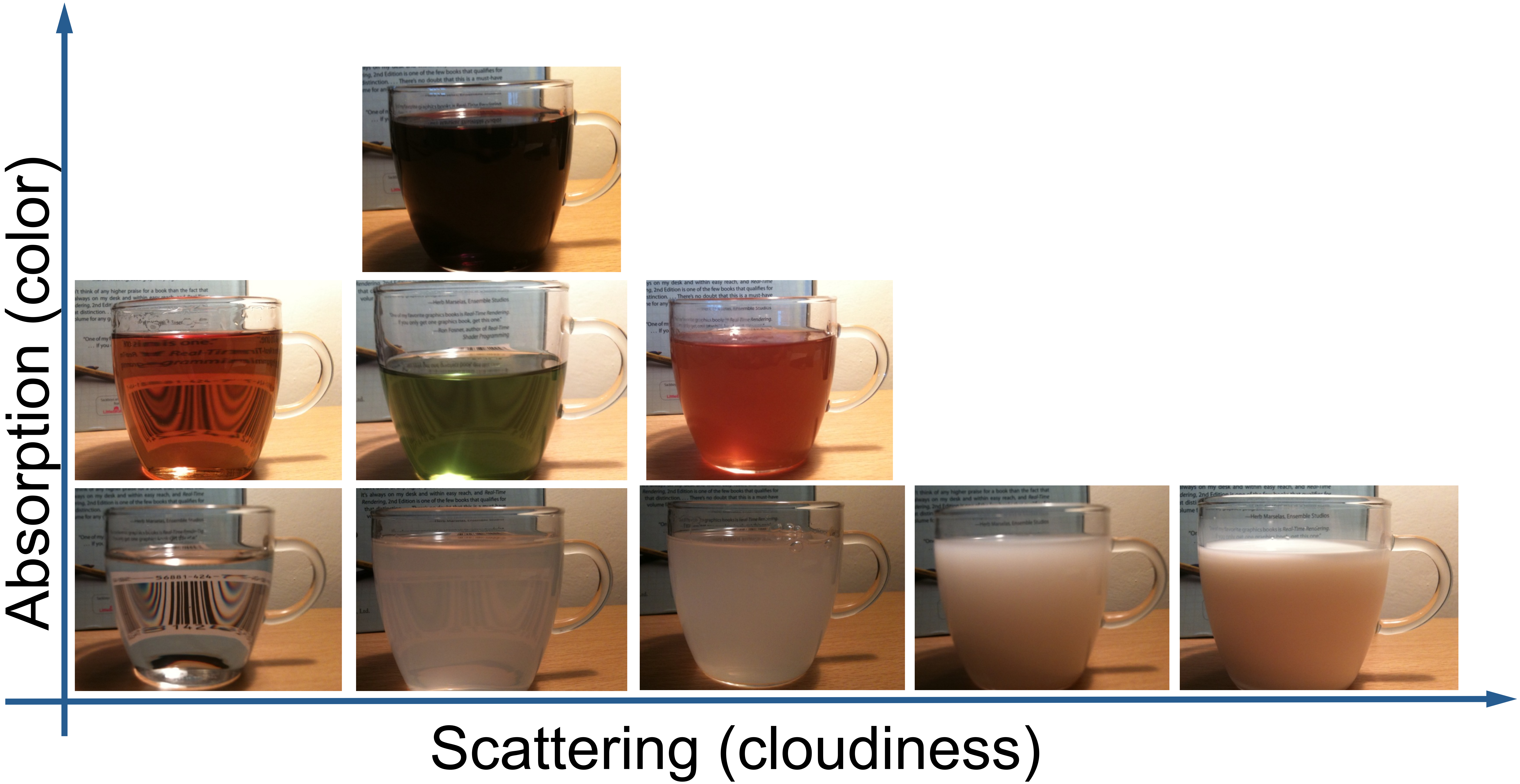


The optical properties of an homogeneous medium is described by its *index of refraction* (IOR for short). This is a complex (in other words, two-part) number. One part of the IOR describes the speed of light through the medium, and the other part describes how much light is absorbed by the medium. There are many media that are completely non-absorbent, and for them the second number will be zero).

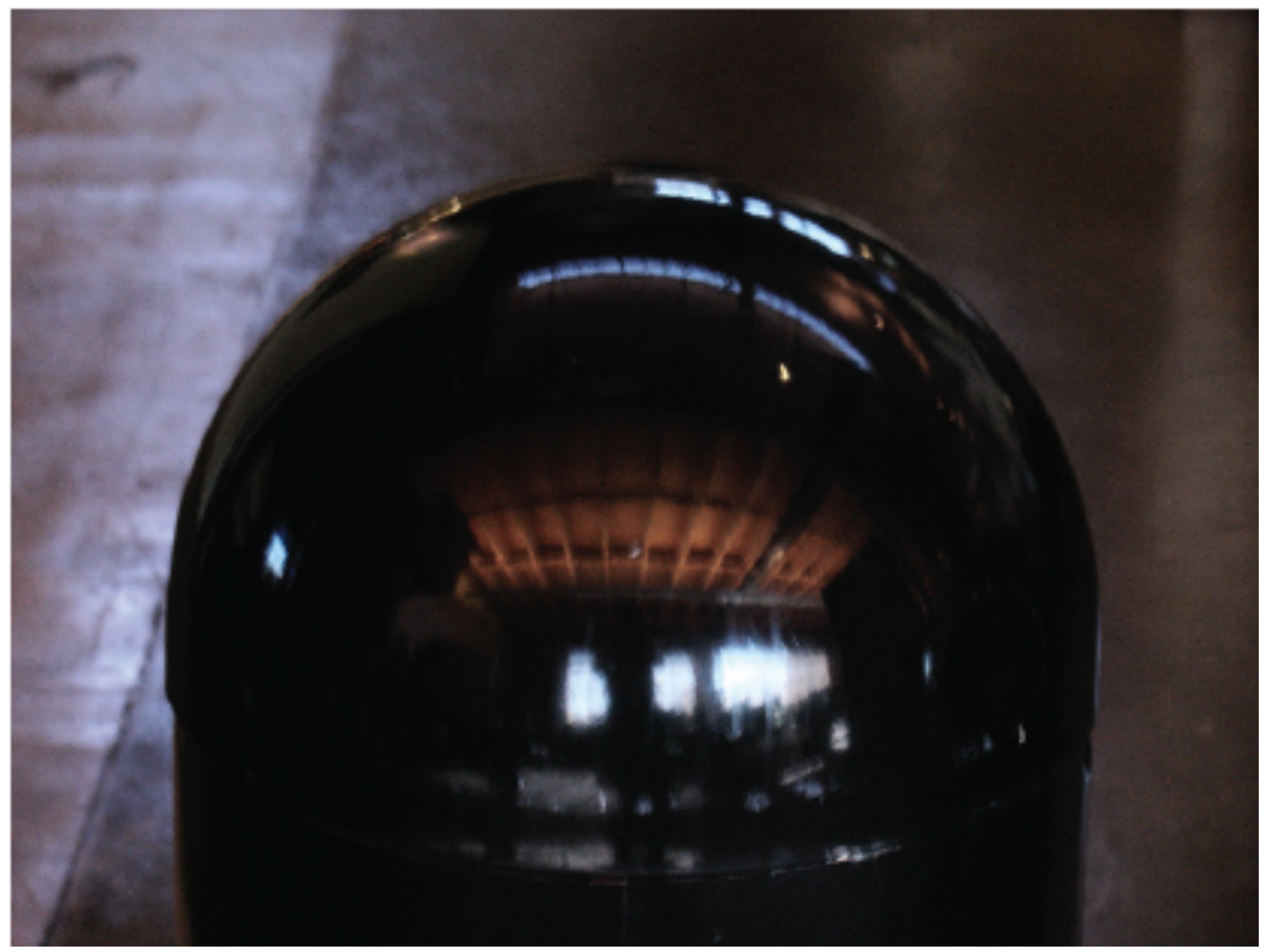
Scattering Particle



Localized inhomogeneities in the medium are modeled as *particles*; We can't only model perfectly homogeneous media; we also need to handle media that have localized inhomogeneities, which in optics are modeled as particles. The assumption is that we have these abrupt IOR discontinuities, and they *scatter* the incoming light over various directions. This is similar to the individual molecule polarization discussed earlier, but these particles can be composed of many molecules. There are formalisms to handle this situation—like the concept of an homogeneous media, the concept of a scattering particle is a bit of an abstraction over what's actually happening.

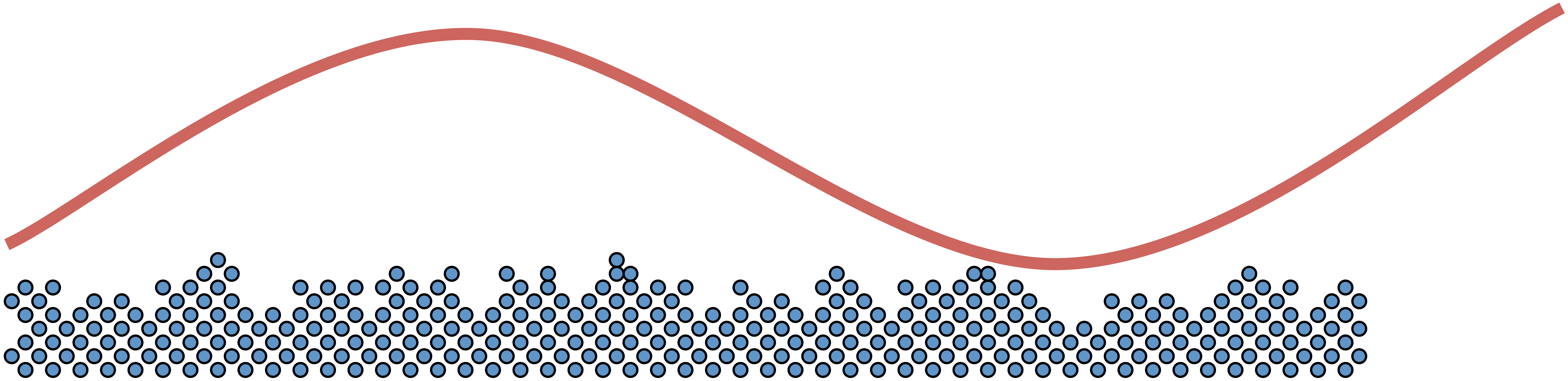


The overall appearance of a medium is determined by the combination of its absorption and scattering properties. These pictures of liquids show absorption and scattering as two independent axes. For example, a white appearance (like the whole milk in the lower right corner) is caused by high scattering and low absorption. If a liquid is colored, that means that it absorbs more light in some wavelengths than in others; the absorption is spectrally selective.



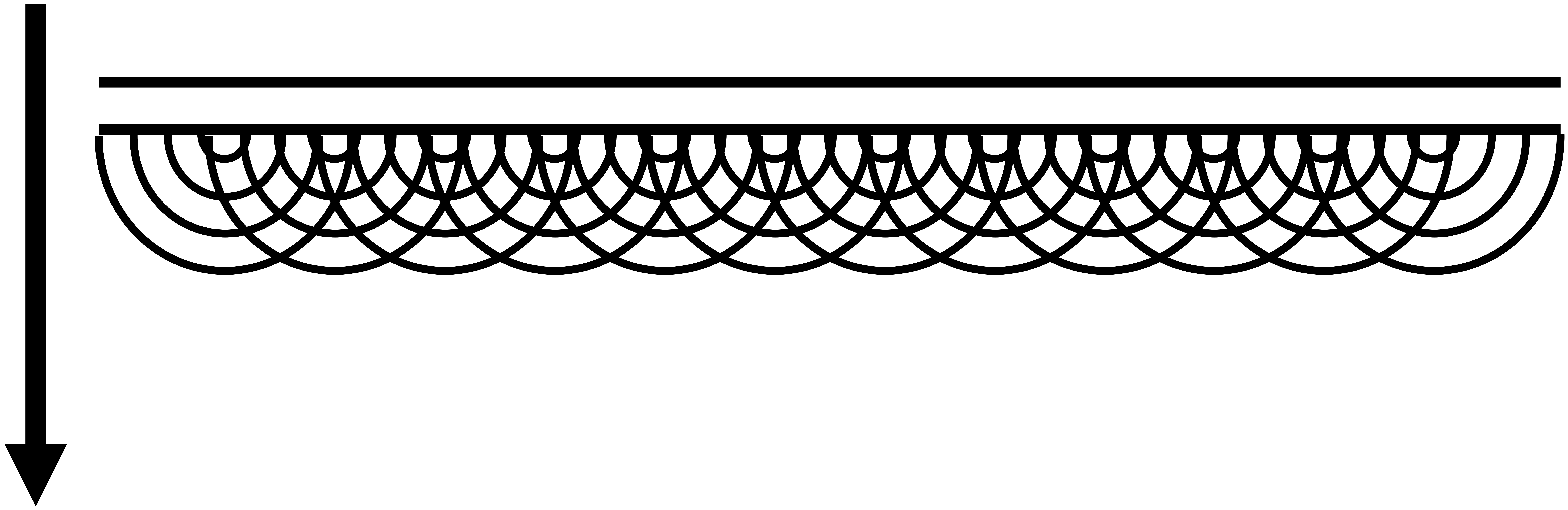
Although we've briefly touched on participating media, the rest of this talk will focus on object surfaces, which are a more common and basic case in rendering.

Nanogeometry



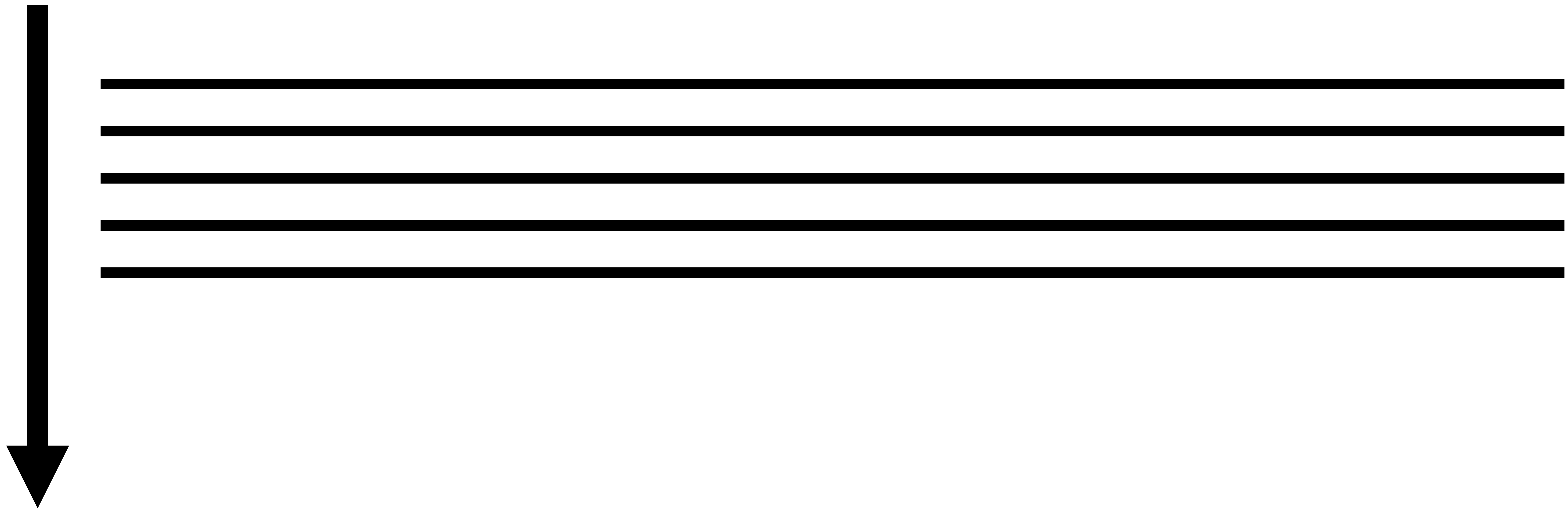
From an optical perspective, the most important thing about a surface is its roughness. No surface can be perfectly flat; at the very least you will have irregularities at the atomic level. Irregularities that are of similar size or smaller than the light wavelength (which we will refer to as *nanogeometry*) cause a phenomenon called *diffraction*.

Huygens-Fresnel Principle



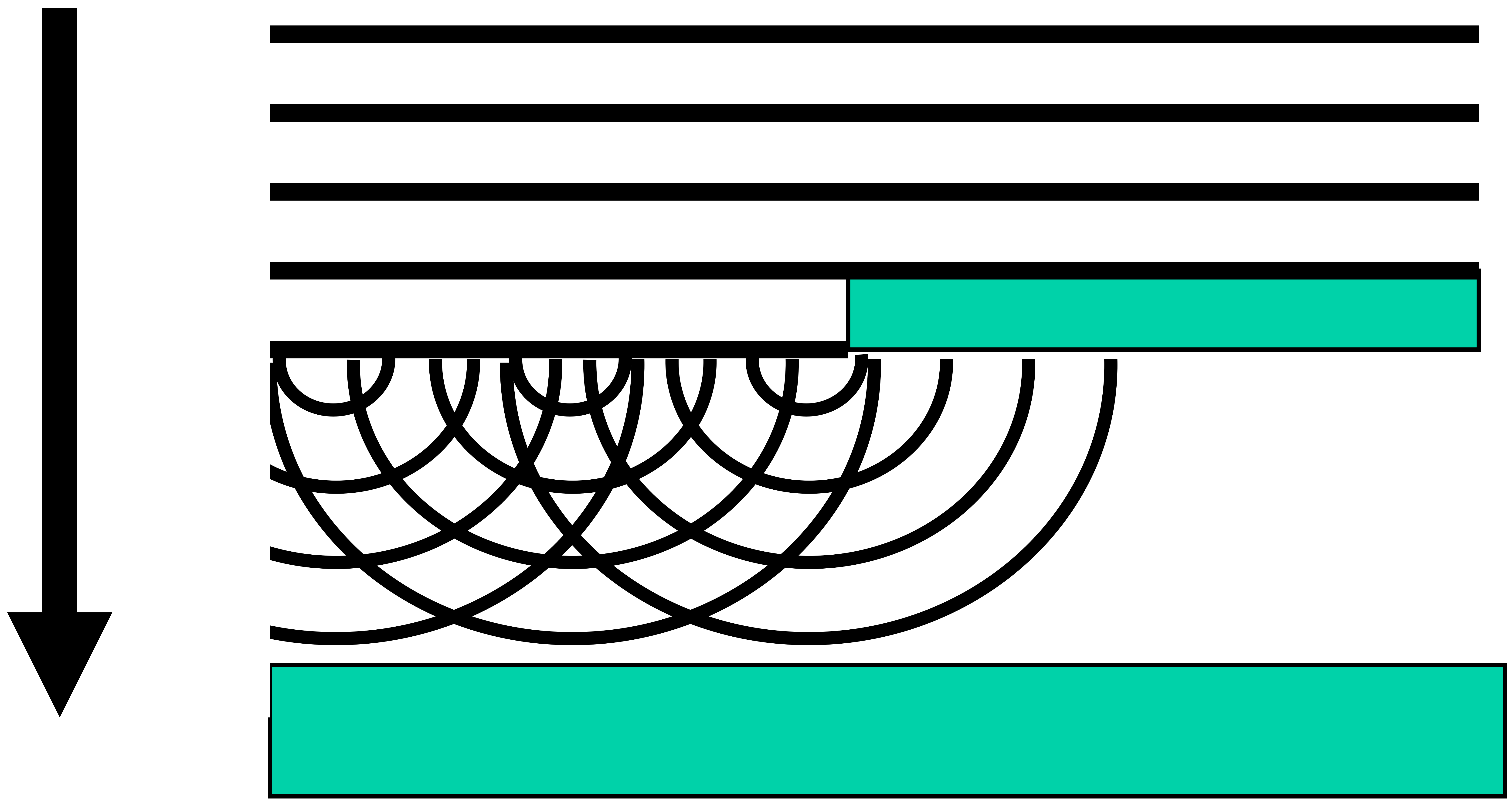
There is something called the Huygens-Fresnel principle that can help understand diffraction in an intuitive way. It states that each point on a planar light wave can be seen as the center of a new spherical wave; these spherical waves interfere with each other...

Huygens-Fresnel Principle



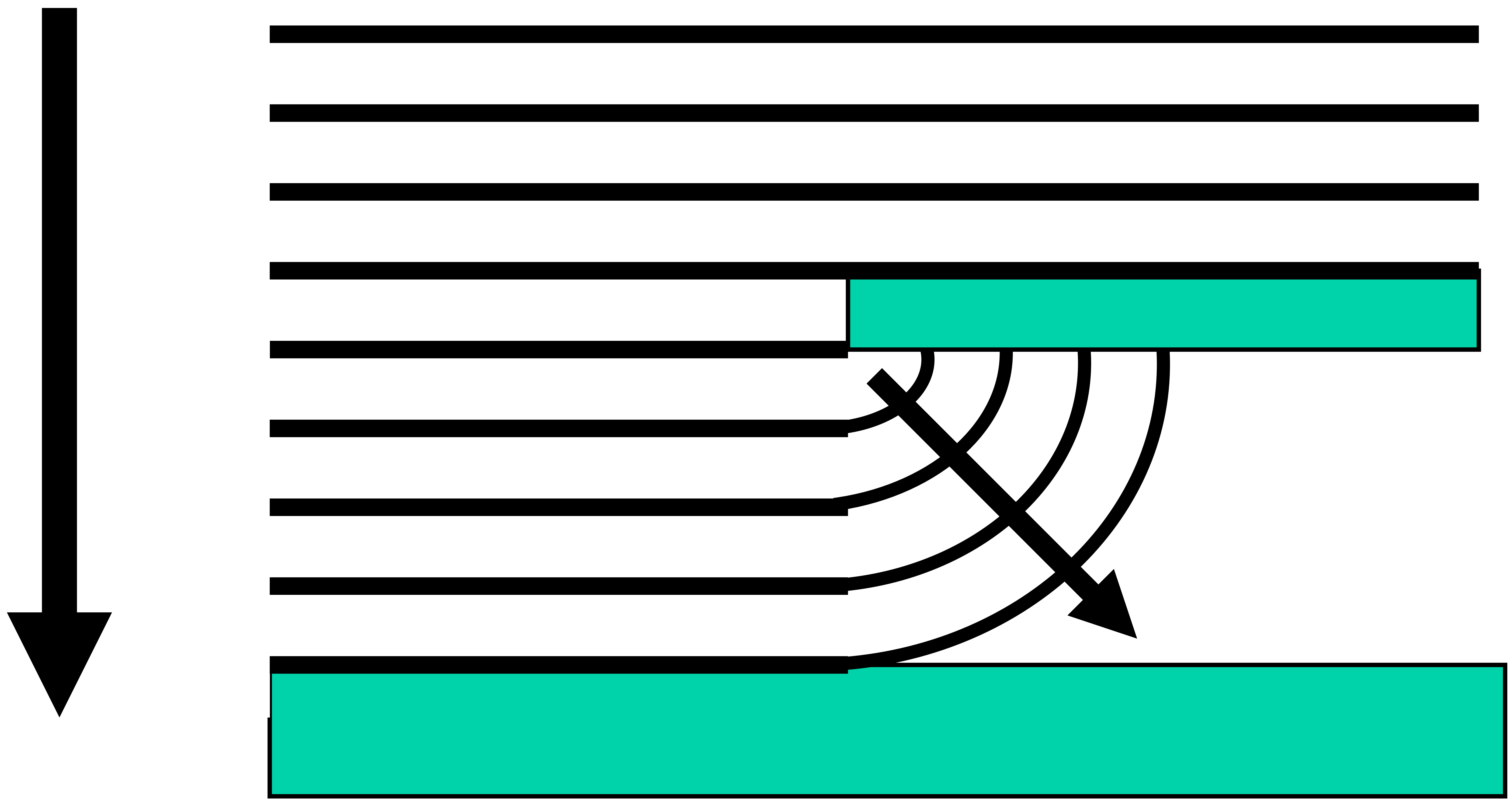
...to produce the resulting planar wave. So far this hasn't gained us anything in terms of intuition—we went from a plane wave, through a bunch of sphere waves, back to a plane wave...

Diffraction



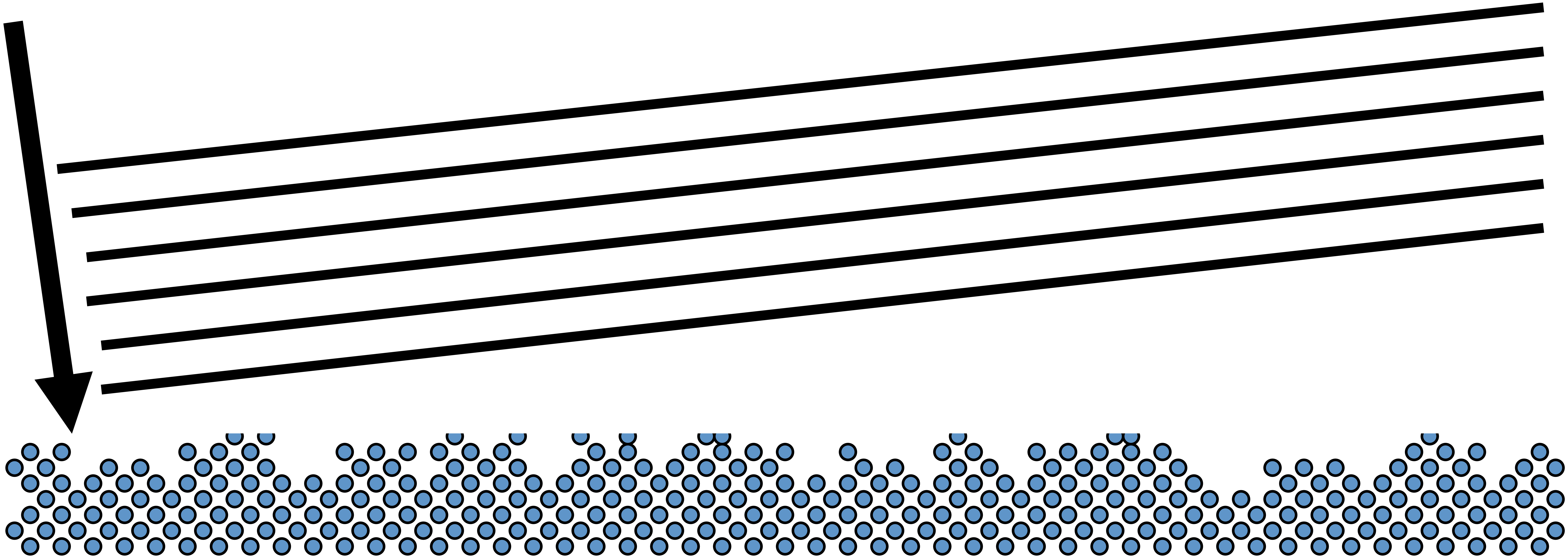
...but where this sort of mental picture helps, is when the light hits an obstacle, and that's when you get diffraction. The sphere wave on the edge of the obstacle is not fully canceled out because there are no waves to the right of it, so you can see...

Diffraction



... that it effectively bends around the corner; the wave no longer behaves strictly as a plane wave moving in a straight line. That's something that you don't get with standard geometric (non-wave) optics. Among other phenomena, this causes a slight softening of shadows, even from very small light sources.

Diffraction from Optically-Smooth Surface



What's more interesting for reflectance (which is what this course is all about) is diffraction on a surface. For this case we're going to look at an *optically-smooth* surface, defined as a surface where all irregularities are in the nanogeometry category, or smaller than a light wavelength. It's actually not that hard to polish a surface to that degree; commercial glass is commonly polished to a tolerance far smaller than a visible light wavelength. If we look at this plane wave hitting this surface, that's irregular on a scale of tens or hundreds of nanometers, then we can apply the same Huygens-Fresnel principle.

Diffraction from Optically-Smooth Surface

Every point on the surface is emitting its own spherical wave. This looks chaotic, because it is: some of these surface points are higher and some are lower, due to the nanogeometry.

Diffraction from Optically-Smooth Surface

The combination of constructive and destructive interference between all these sphere waves will result in a complexly-structured wavefront (this sketch doesn't really do it justice), and some amount of light will be scattered in various directions. The lower the nanogeometry bumps are, less light will be diffracted and more light *reflected* in the "regular" specular way that I'll talk about later. In other words, the amount of diffraction is a function of the height of the bumps. Even at the limit, for a surface polished so smoothly that the only remaining irregularity is caused by the atoms making up the surface (such surfaces are referred to in the optics industry as "super polished"), there is a small but measurable amount of diffraction - around half a percent of incident light. So there is no such thing as a perfect surface.

Geometric (Ray) Optics

We are now taking a break from wave optics and moving to geometric or ray optics, which is a more simplified model and it's the one used in computer graphics - with a handful of exceptions - for the last 30 years. One simplification we will make is to ignore nanogeometry and diffraction. Basically as far as geometric optics is concenred, if it's smaller than a light wavelength it doesn't exist. Optically smooth surfaces are treated as mathematically perfect flat surfaces.

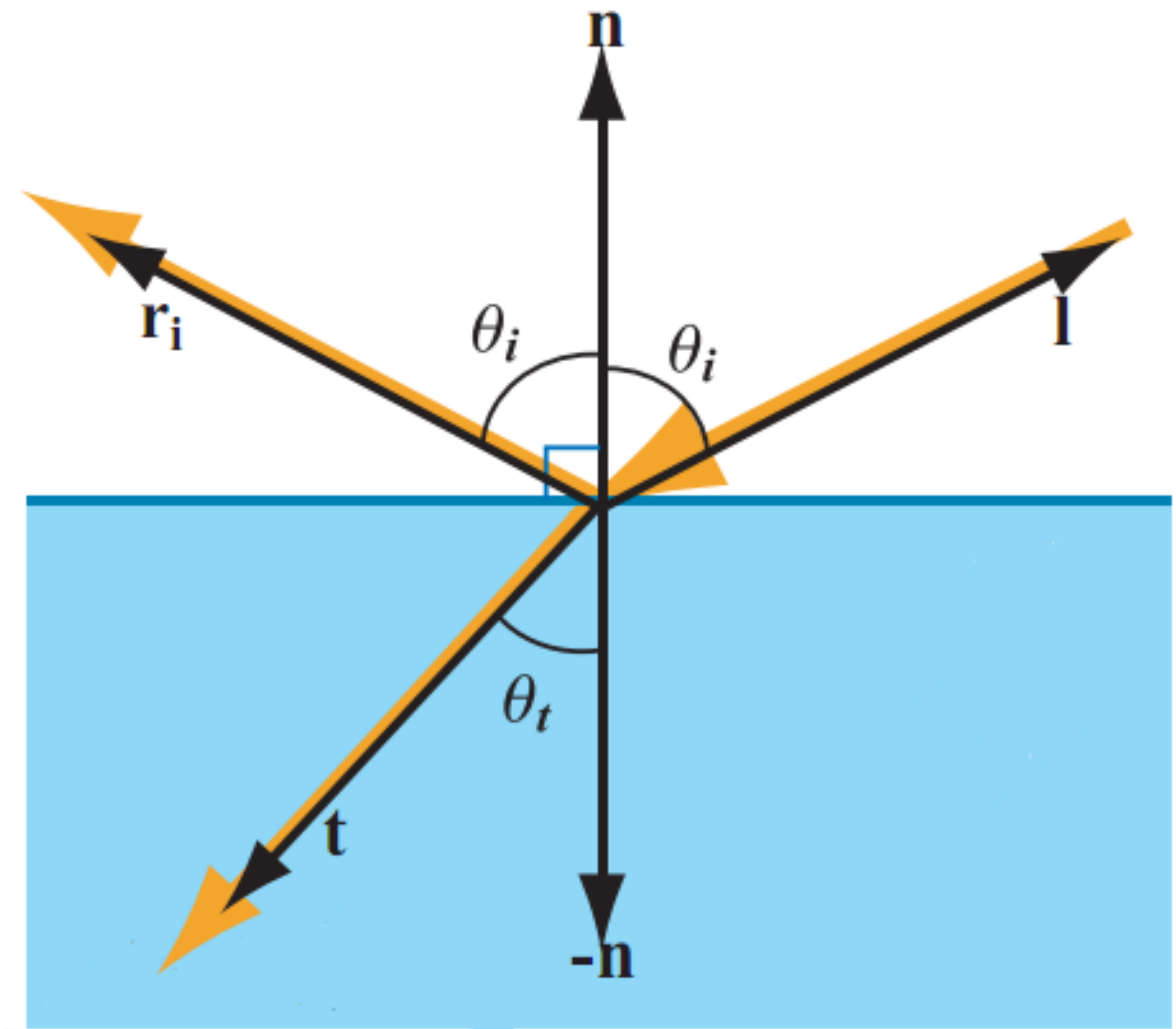


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

It can be shown from the equations governing electromagnetic waves, that such a perfectly flat surface will split light into exactly two directions: reflection and refraction.

Microgeometry

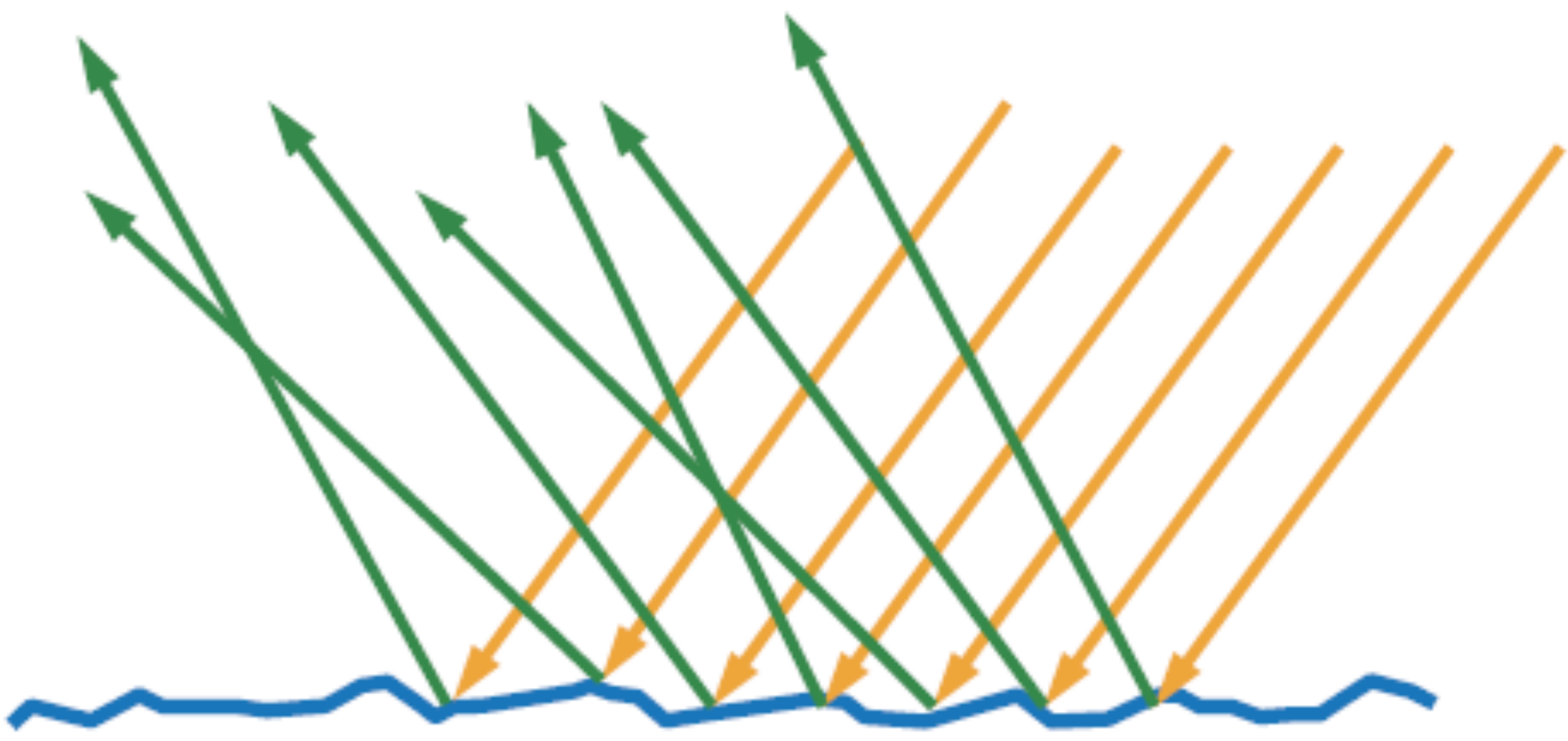
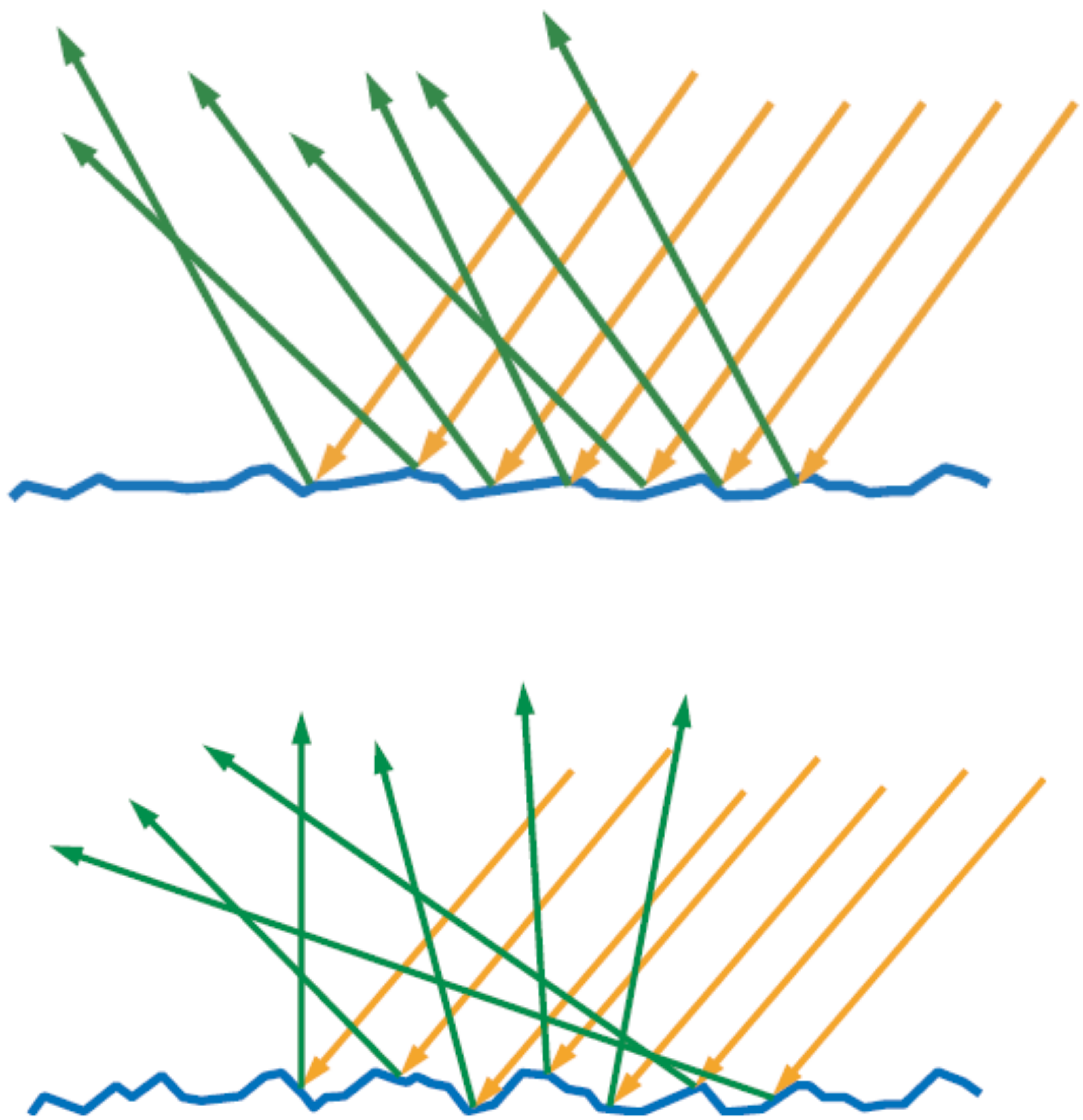


Image from "Real-Time Rendering, 3rd Edition", A K Peters 2008

Most real-world surfaces are not optically smooth but possess irregularities at a scale much larger than a light wavelength, but smaller than a pixel. This *microgeometry* variation causes each surface point to reflect (and refract) light in a different direction: the appearance is the aggregate result of these reflection and refraction directions.

Rougher = Blurrier Reflections



Images from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

These surfaces seem equally smooth but differ at the microscopic scale. The top surface is only a little rough; incoming light rays hit surface points that are angled slightly differently and get reflected to somewhat different outgoing directions, causing slightly blurred reflections. The bottom surface is much rougher, causing significantly blurrier reflections.

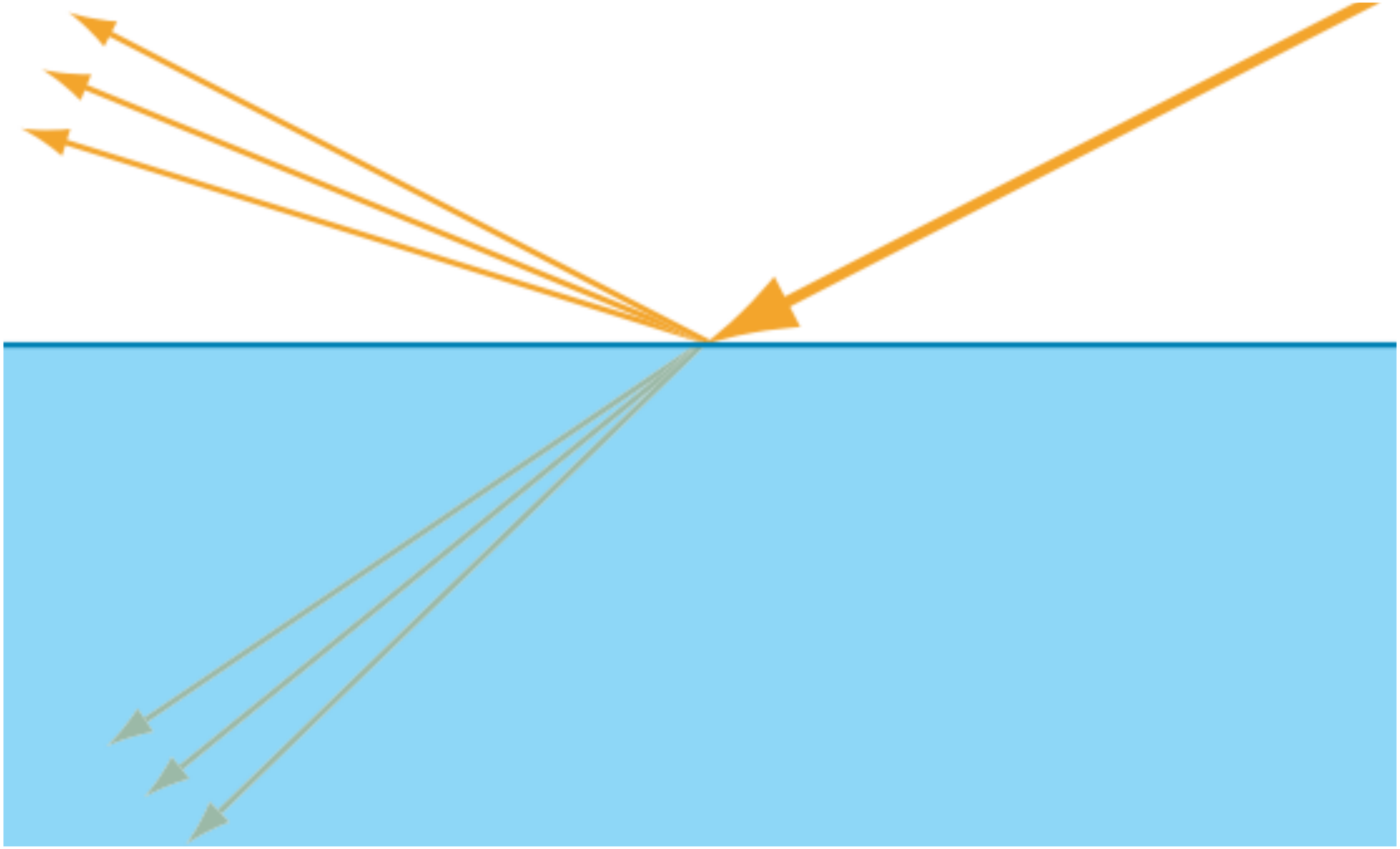
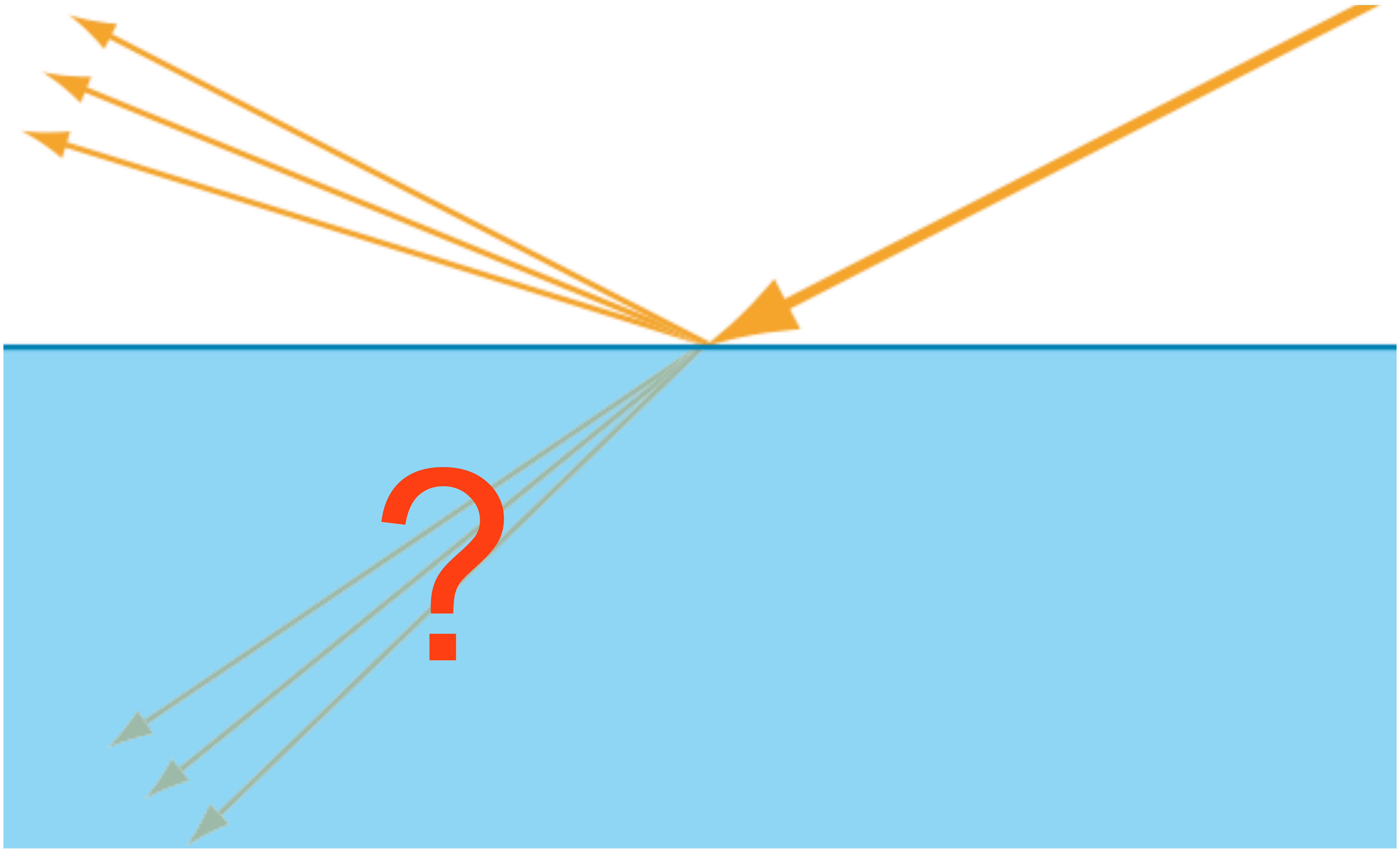


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

In the macroscopic view, we treat the microgeometry statistically and view the surface as reflecting (and refracting) light in multiple directions. The rougher the surface, the wider the cones of reflected and refracted directions will be.



What happens to the refracted light? It depends on what kind of material the object is made of.

Metals (Conductors)

Dielectrics (Insulators)

Semiconductors

Light is composed of electromagnetic waves. So the optical properties of a substance are closely linked to its electric properties. Materials can be grouped into three main optical categories: metals (or conductors), dielectrics (or insulators), and semiconductors.

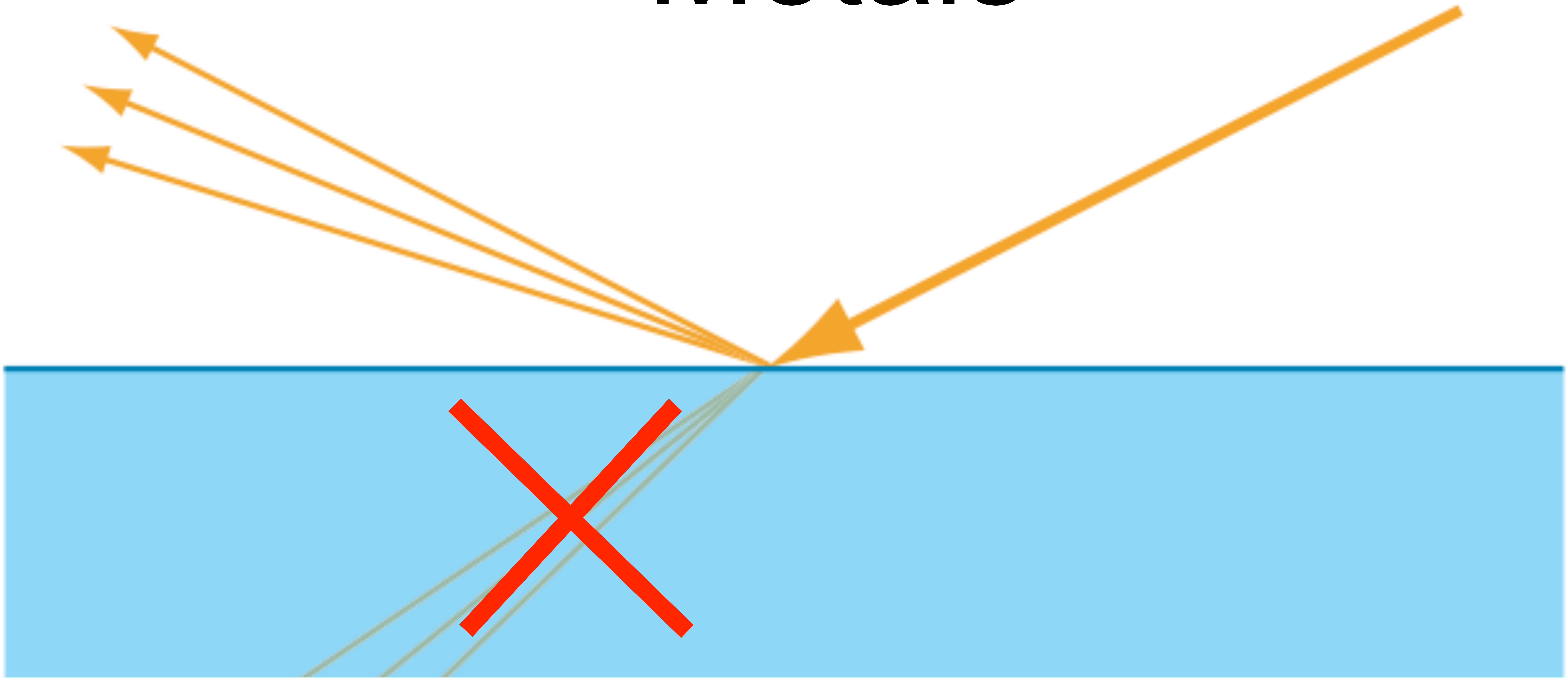
Metals

Non-Metals

~~Semiconductors~~

Since visible object surfaces are rarely semiconductors, for practical purposes we can do a simpler grouping, into metals and non-metals.

Metals



Metals immediately absorb all refracted light.

Non-Metals

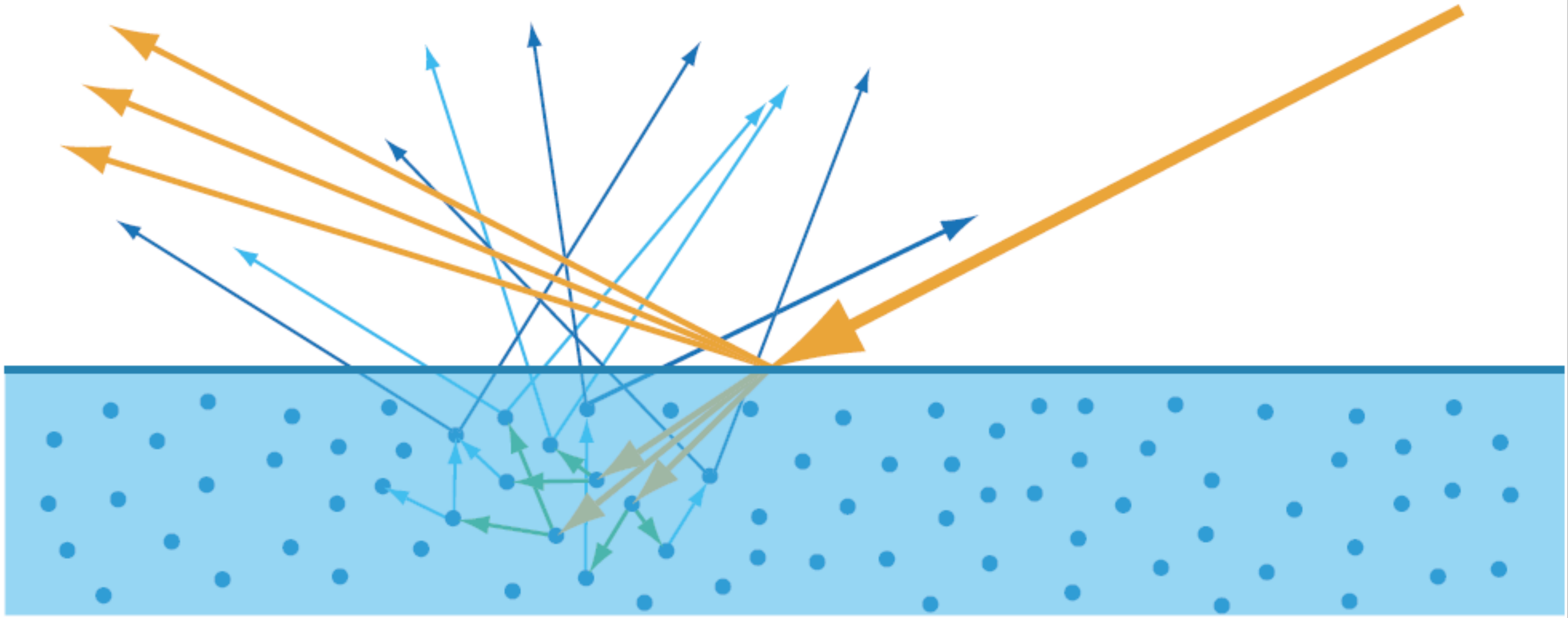
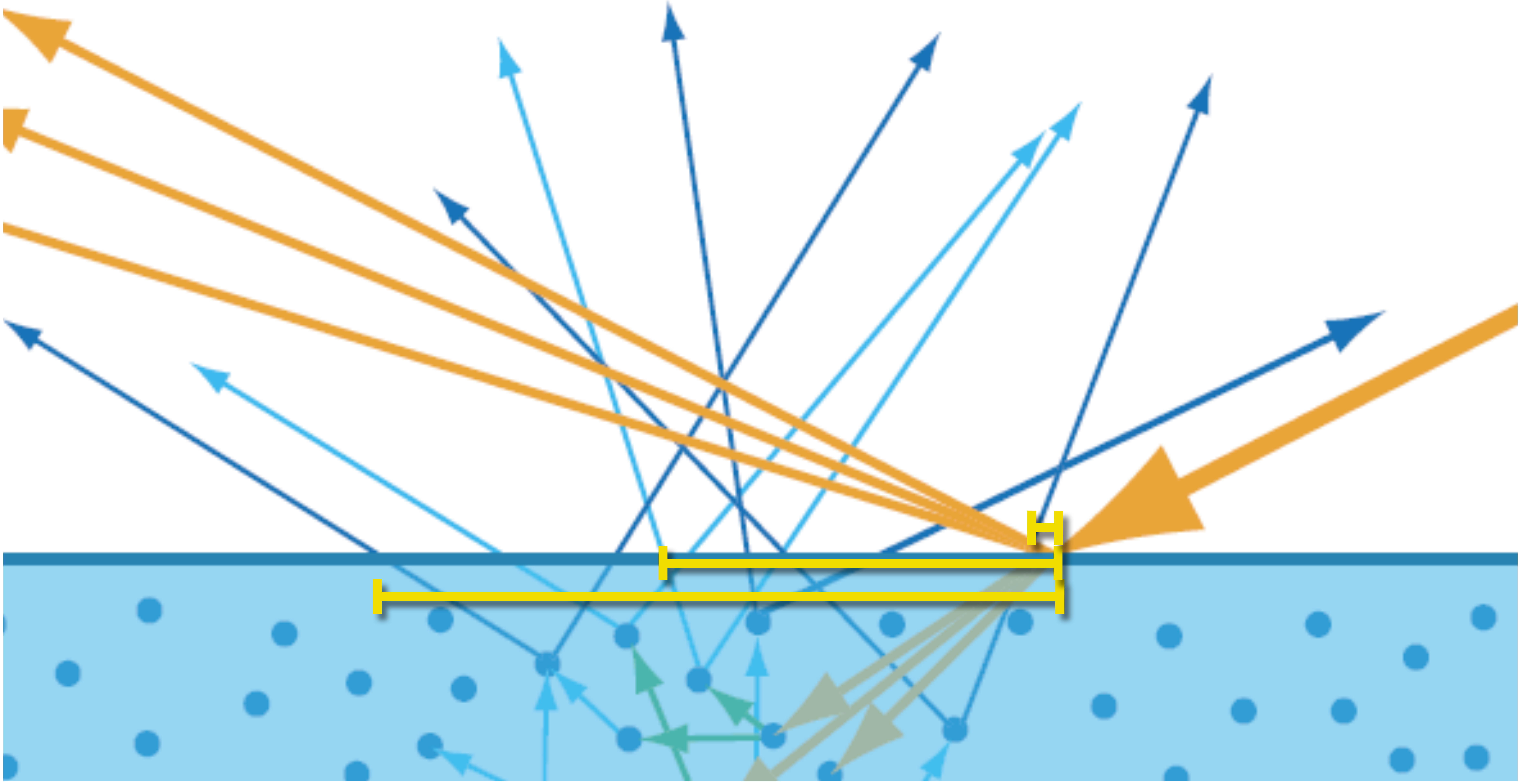


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

Non-metals behave like those cups of liquid we saw earlier: refracted light is scattered and/or absorbed to some degree. Unless the object is made out of a clear substance like glass or crystal, there will be enough scattering that some of the refracted light is scattered back out of the surface: these are the blue arrows you see coming out of the surface in various directions.



The re-emitted light comes out at varying distances (shown by the yellow bars) from the entry point. The distribution of distances depends on the density and properties of the scattering particles.

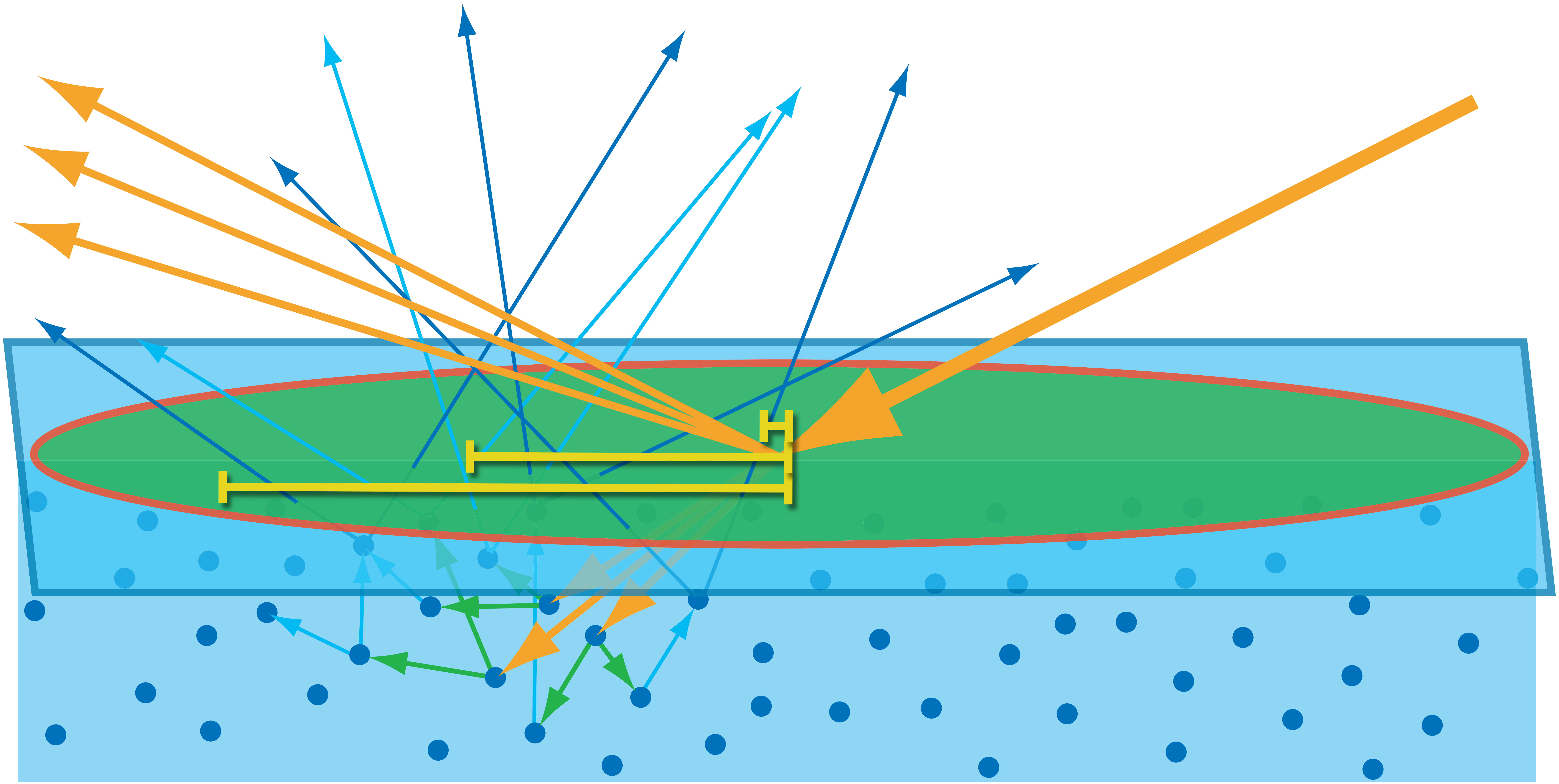


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

If the pixel size (or shading sample area) is large (like the red-bordered green circle) compared to the entry-exit distances, we can assume that the distances are effectively zero for shading purposes.

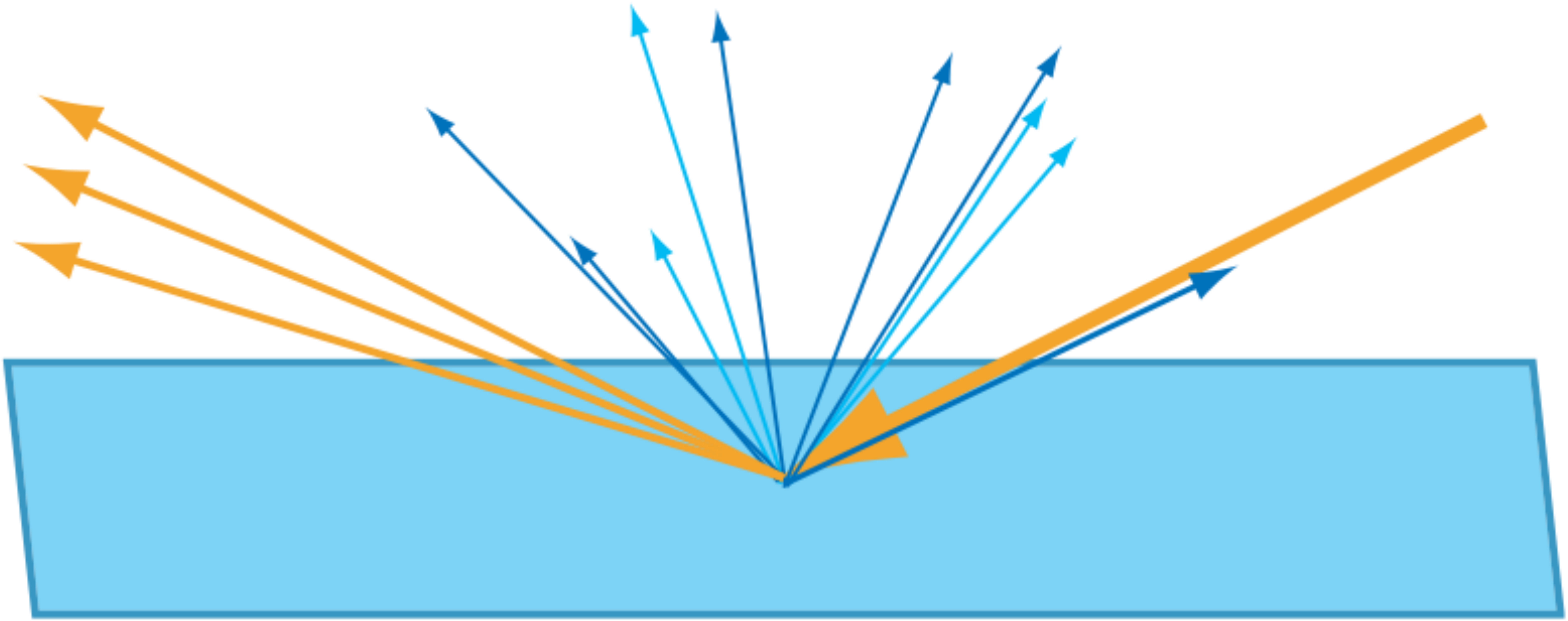


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

By ignoring the entry-to-exit distance, we can then compute all shading locally at a single point. The shaded color is only affected by light hitting that surface point.

It is convenient to split these two very different light-material interactions into different shading terms. We call the surface reflection term “specular” and the term resulting from refraction, absorption, scattering, and re-refraction we call “diffuse”.

If the pixel is small compared to the entry-exit distances (like the red-bordered green circle), then special “subsurface scattering” rendering techniques are needed. Even regular diffuse shading is a result of subsurface scattering: the difference is the shading resolution compared to the scattering distance. For example, plastic displays noticeable diffusion in extreme close-up shots (e.g. of small toys).

Physics → Math

So far we've discussed the physics of light/matter interactions. To turn these physics into mathematical models that can be used for shading, the first step is to quantify light as a number.

Radiance

Radiometry is the measurement of light. Of the various radiometric quantities, we'll use *radiance*...

Radiance

Single Ray

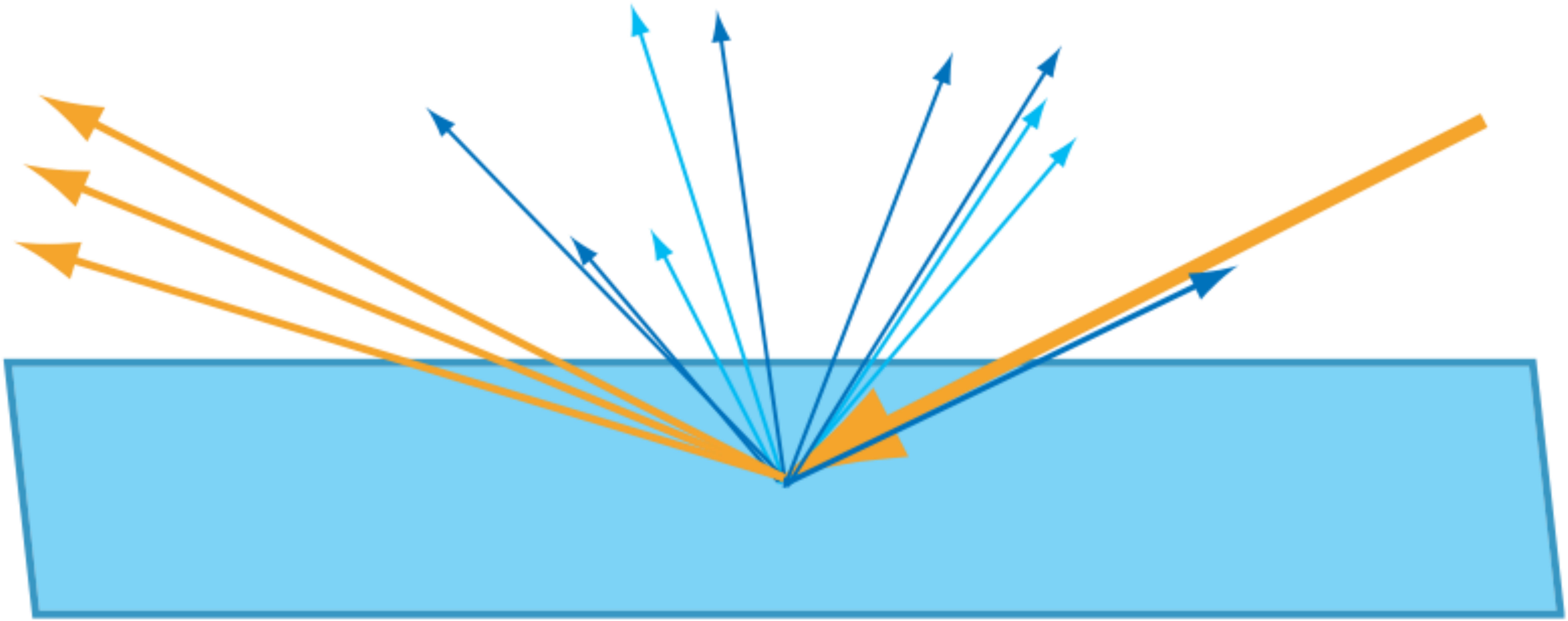
...which measures the intensity of light along a single ray...

Radiance

Single Ray

Spectral/RGB

...and is spectrally-varying. Radiance values are properly expressed as SPDs, like the ones I showed earlier. However, for the rest of this talk I'll follow traditional film and game usage, using RGB for spectrally varying quantities like radiance. The units of radiance are Watts per steradian per square meter.



Given the assumption that shading can be handled locally, light response at a surface point only depends on the light and view directions.

Bidirectional **R**eflectance **D**istribution **F**unction

$$f(\mathbf{l}, \mathbf{v})$$

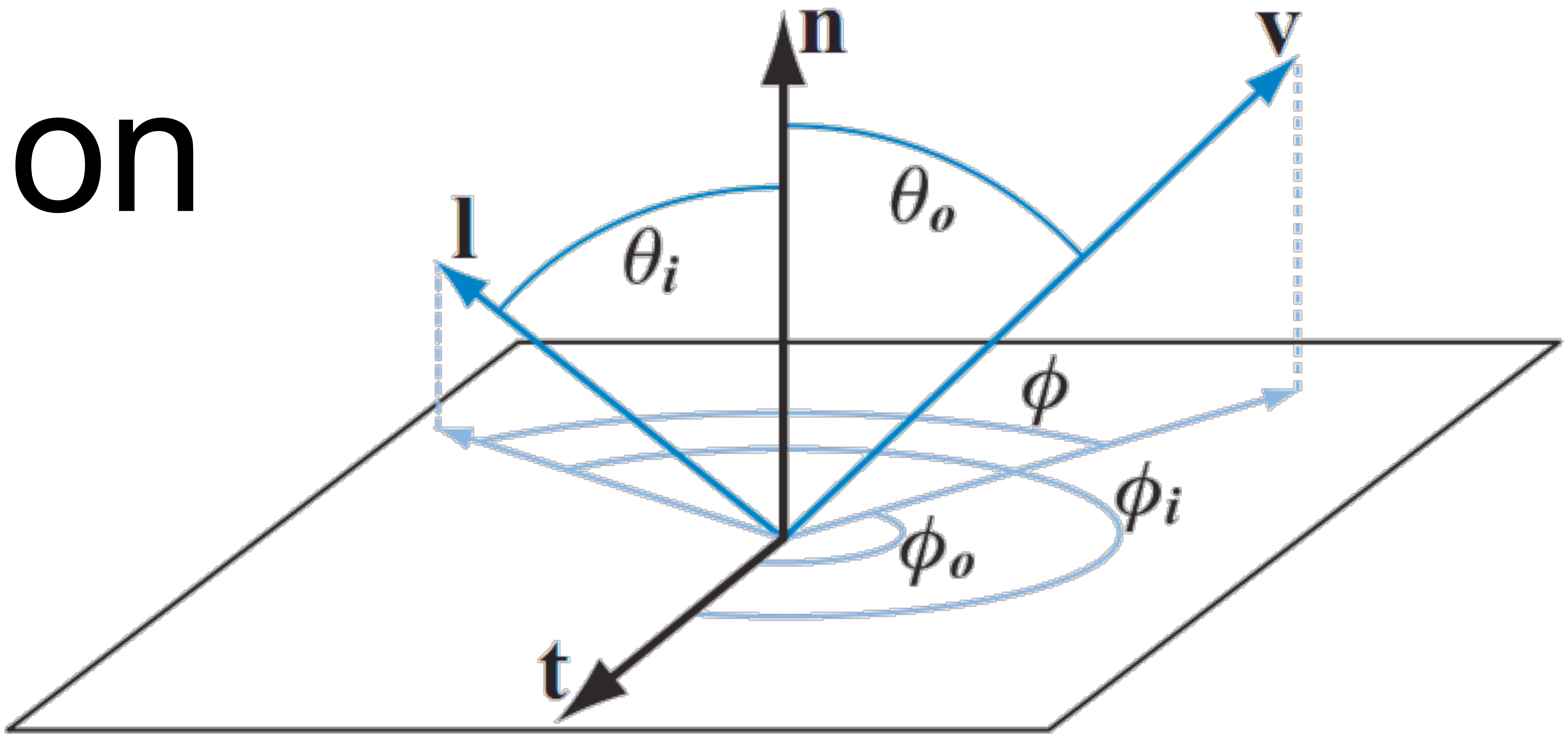


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

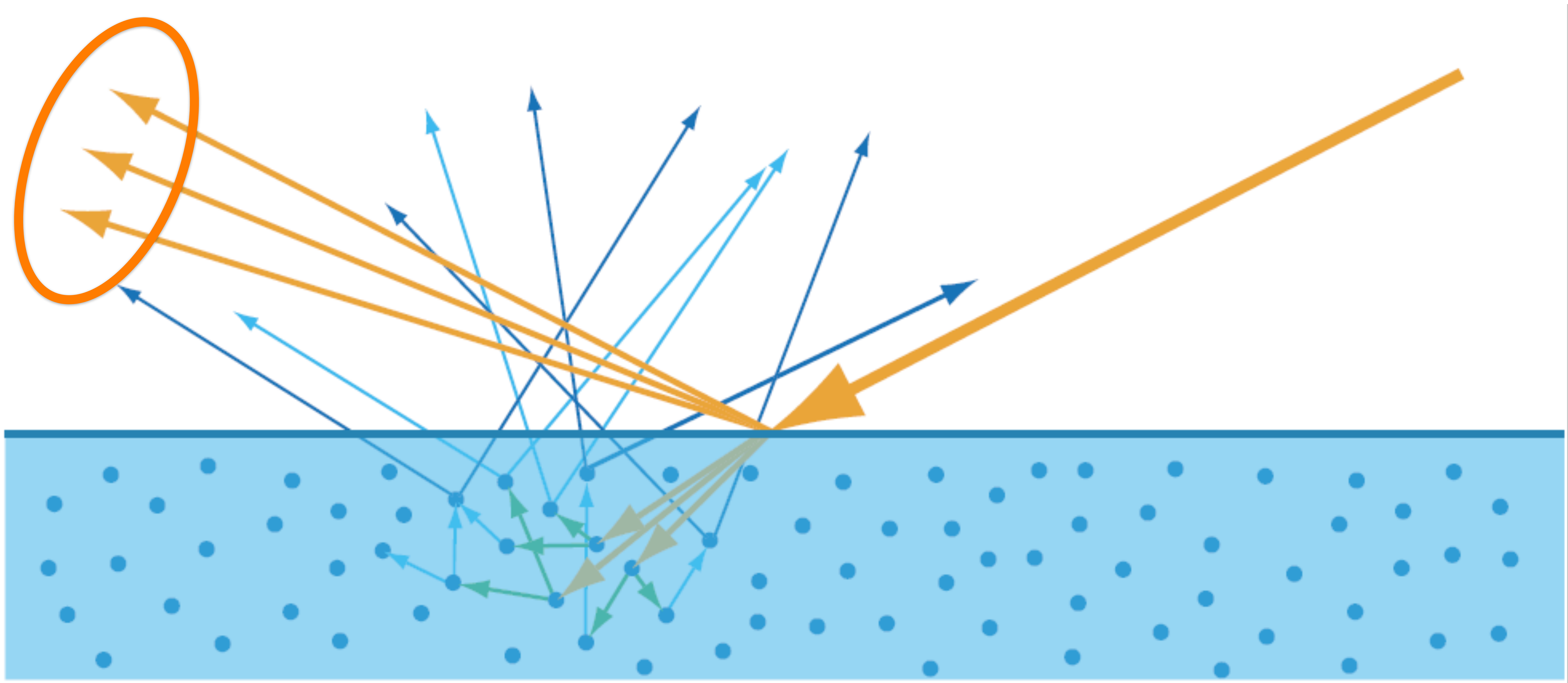
We represent this variation with the *BRDF*, a function of light direction $\mathbf{l}$ and view direction $\mathbf{v}$. In principle, the BRDF is a function of the 3 or 4 angles shown in the figure. In practice, BRDF models use varying numbers of angles. Note that the BRDF is only defined for light and view vectors above the macroscopic surface.

The Reflectance Equation

$$L_o(\mathbf{v}) = \int_\Omega f(\mathbf{l},\mathbf{v}) \otimes L_i(\mathbf{l})(\mathbf{n}\cdot\mathbf{l})d\omega_i$$

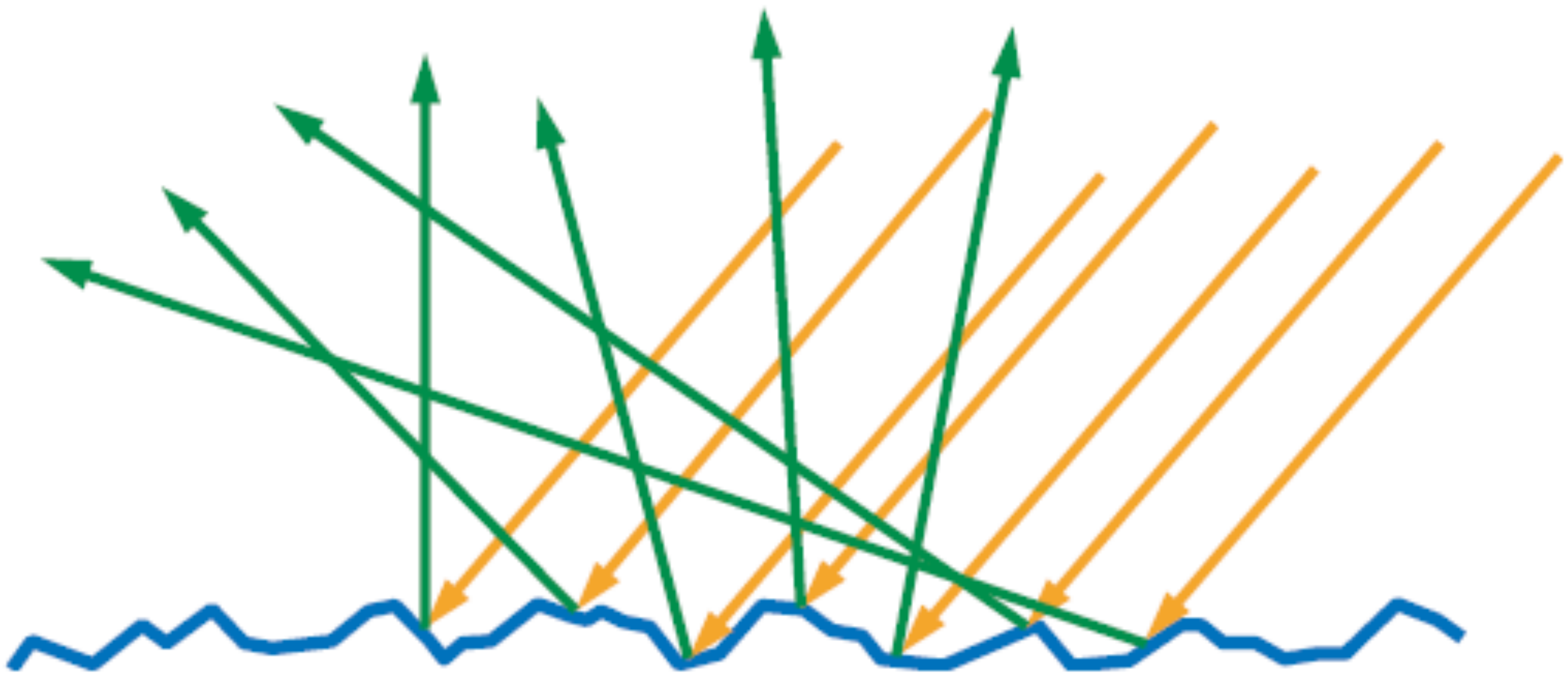
This scary-looking equation just says that outgoing radiance from a point equals the integral of incoming radiance times BRDF times a cosine factor, over the hemisphere of incoming directions. If you're not familiar with integrals you can think of this as a sort of weighted average over all incoming directions. The "X in circle" notation is from the *Real-Time Rendering* book: it means component-wise RGB multiplication.

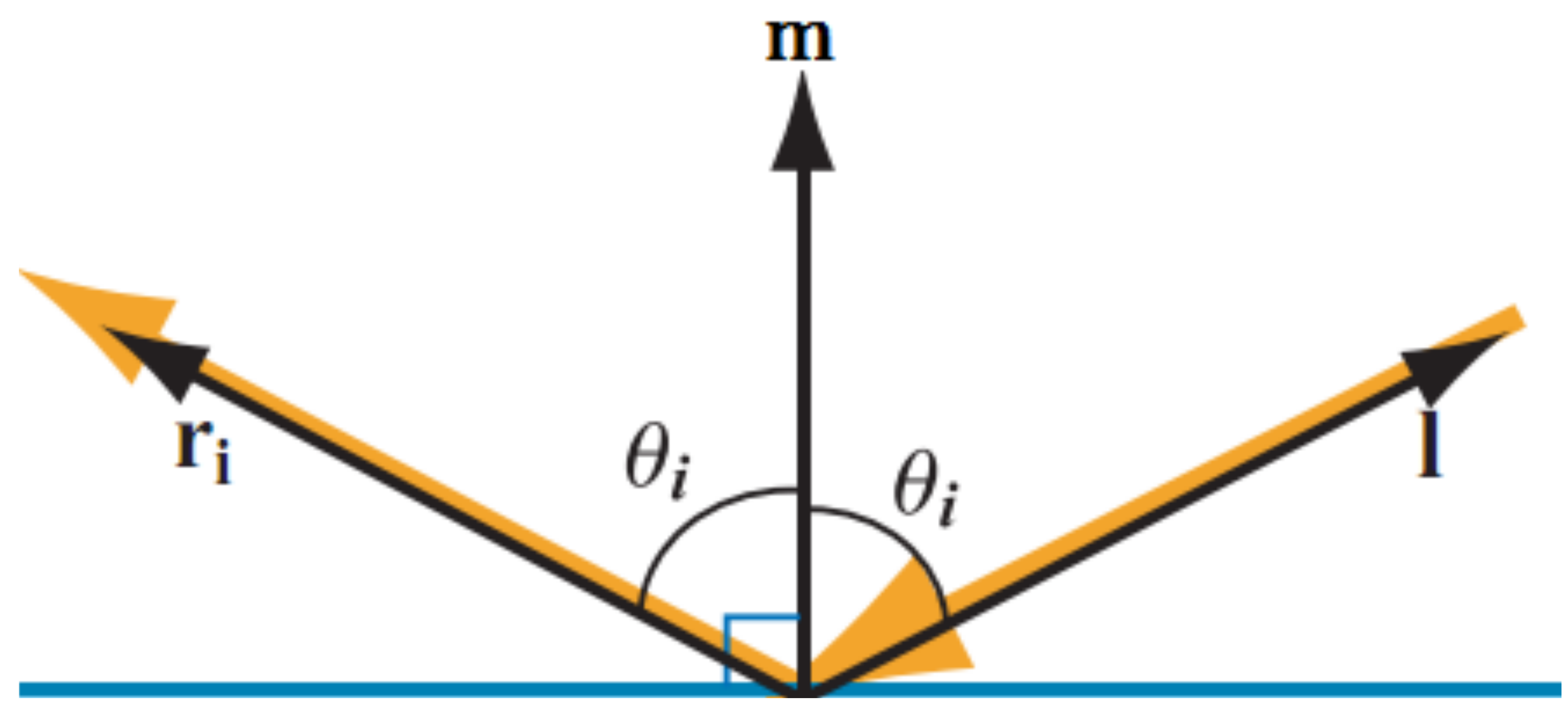
Surface Reflection (Specular Term)



We'll start by looking at the surface, or specular term. In this figure, it is denoted by the orange arrows reflecting back from the surface.

Microfacet Theory





Microfacet theory is a way to derive BRDFs for surface reflection from non-optically flat surfaces. The assumption behind it is a surface with detail that is small compared to the scale of observation but large compared to a light wavelength. Each point is locally a perfect mirror, reflecting each incoming ray of light into one outgoing direction, which depends on the light direction $\mathbf{l}$ and the microfacet normal $\mathbf{m}$.

The Half Vector

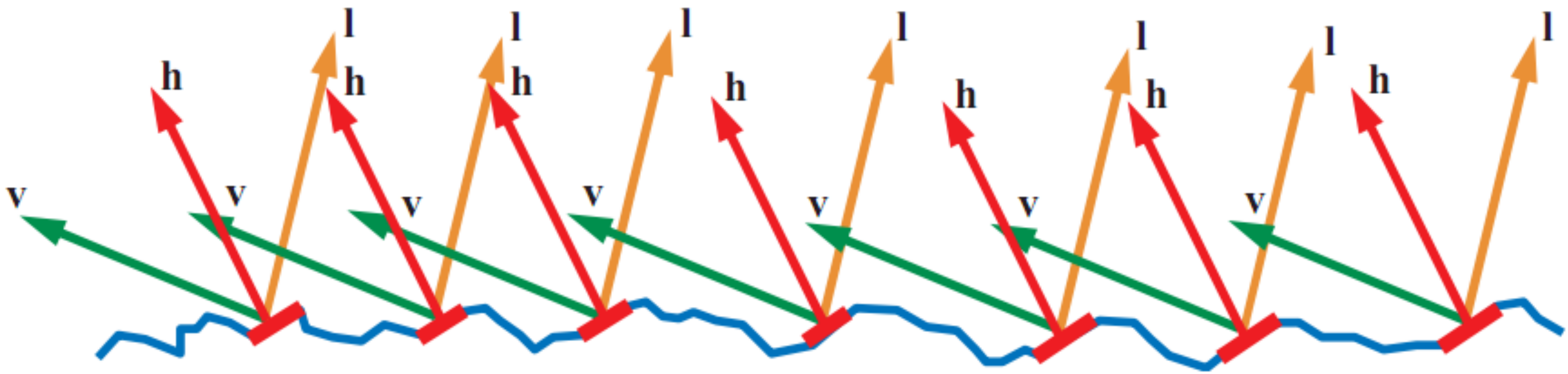
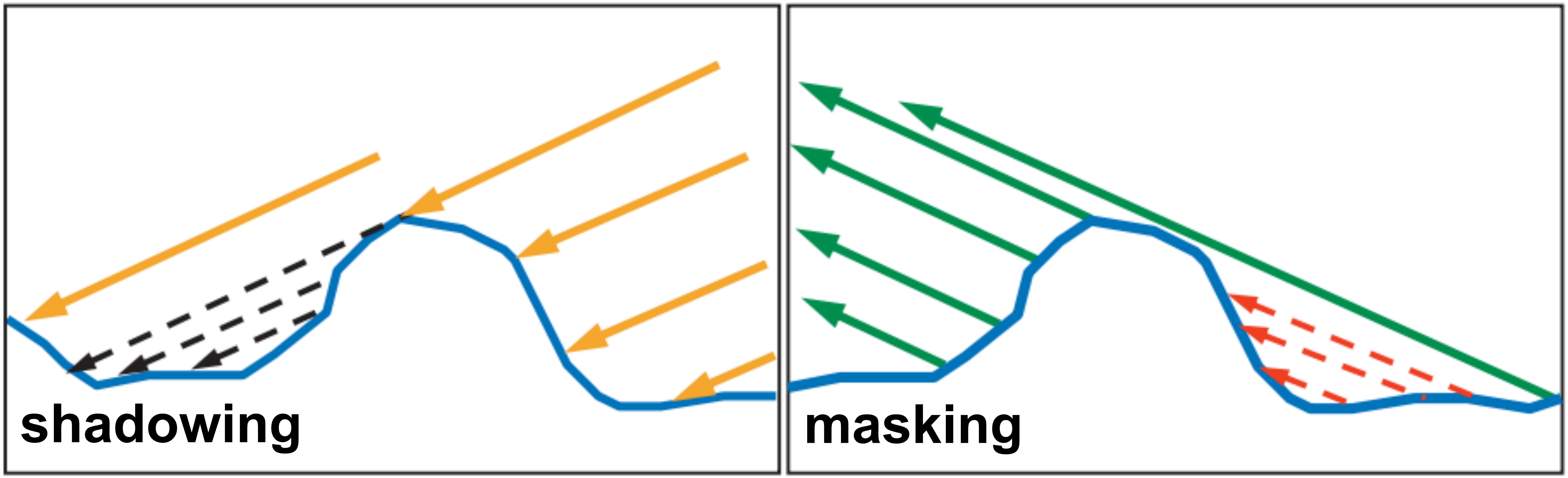


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

Only those microfacets that happen to have their surface normal **m** oriented exactly halfway between **l** and **v** will reflect any visible light: this direction is the half-vector **h**.

Shadowing and Masking



Images from "Real-Time Rendering, 3rd Edition", A K Peters 2008

Not all microfacets with $\mathbf{m} = \mathbf{h}$ will contribute: some will be blocked by other microfacets from either the light direction (*shadowing*) or the view direction (*masking*).

Multiple Surface Bounces

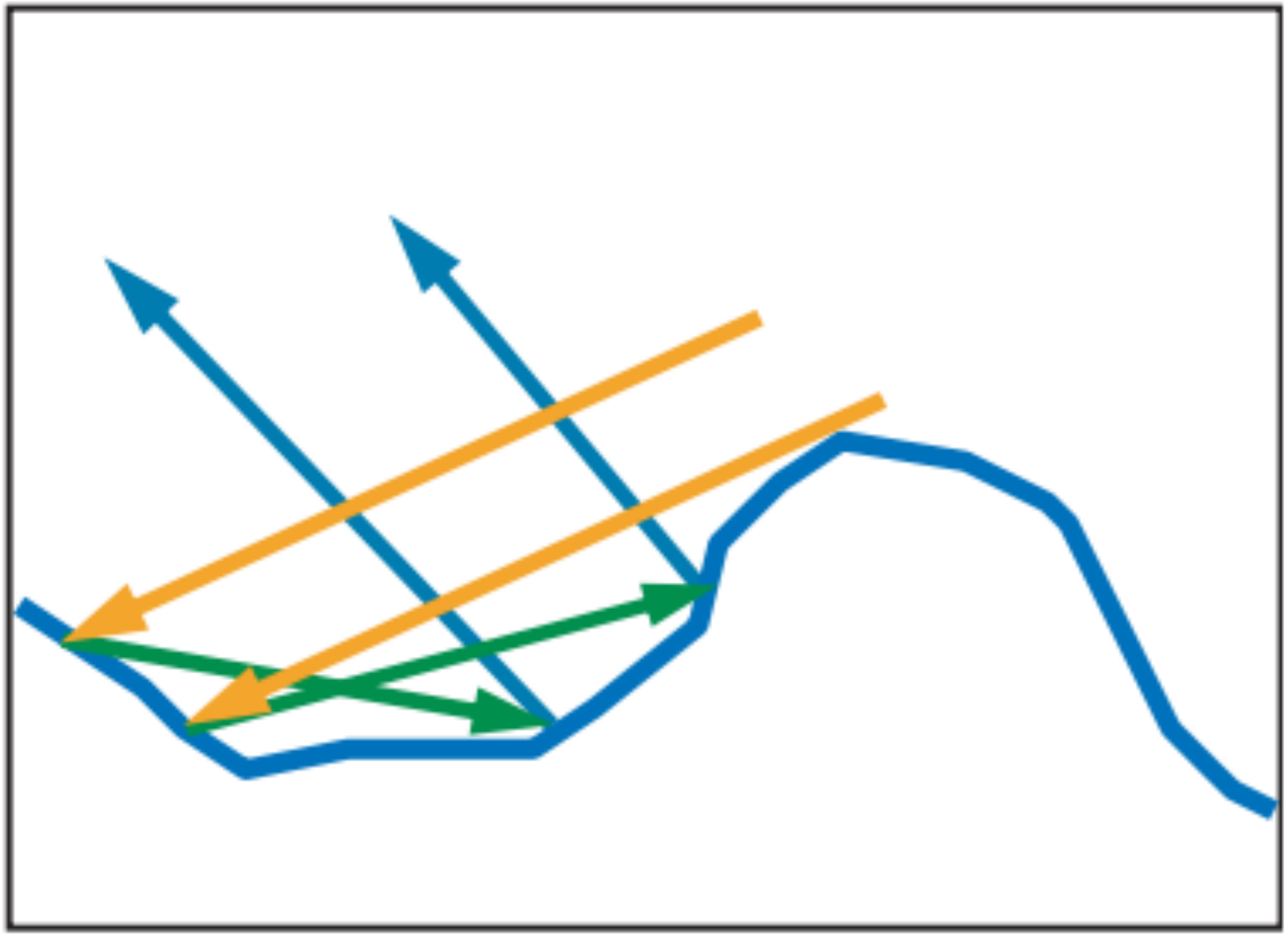


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

In reality, blocked light continues to bounce; some will eventually contribute to the BRDF. Microfacet BRDFs ignore this, so effectively they assume all blocked light is lost.

Microfacet Specular BRDF

$$f(\mathbf{l},\mathbf{v}) = \frac{F(\mathbf{l},\mathbf{h})G(\mathbf{l},\mathbf{v},\mathbf{h})D(\mathbf{h})}{4(\mathbf{n}\cdot\mathbf{l})(\mathbf{n}\cdot\mathbf{v})}$$

This is a general microfacet specular BRDF. I'll go over its various parts, explaining each.

Fresnel Reflectance

$$f(\mathbf{l}, \mathbf{v}) = \frac{F(\mathbf{l}, \mathbf{h})G(\mathbf{l}, \mathbf{v}, \mathbf{h})D(\mathbf{h})}{4(\mathbf{n} \cdot \mathbf{l})(\mathbf{n} \cdot \mathbf{v})}$$

The Fresnel reflectance is the fraction of incoming light that is reflected (as opposed to refracted) from an optically flat surface of a given substance. It varies based on the light direction and the surface (in this case microfacet) normal. Fresnel reflectance tells us how much of the light hitting the relevant microfacets (the ones facing in the half-angle direction) is reflected.

Fresnel Reflectance

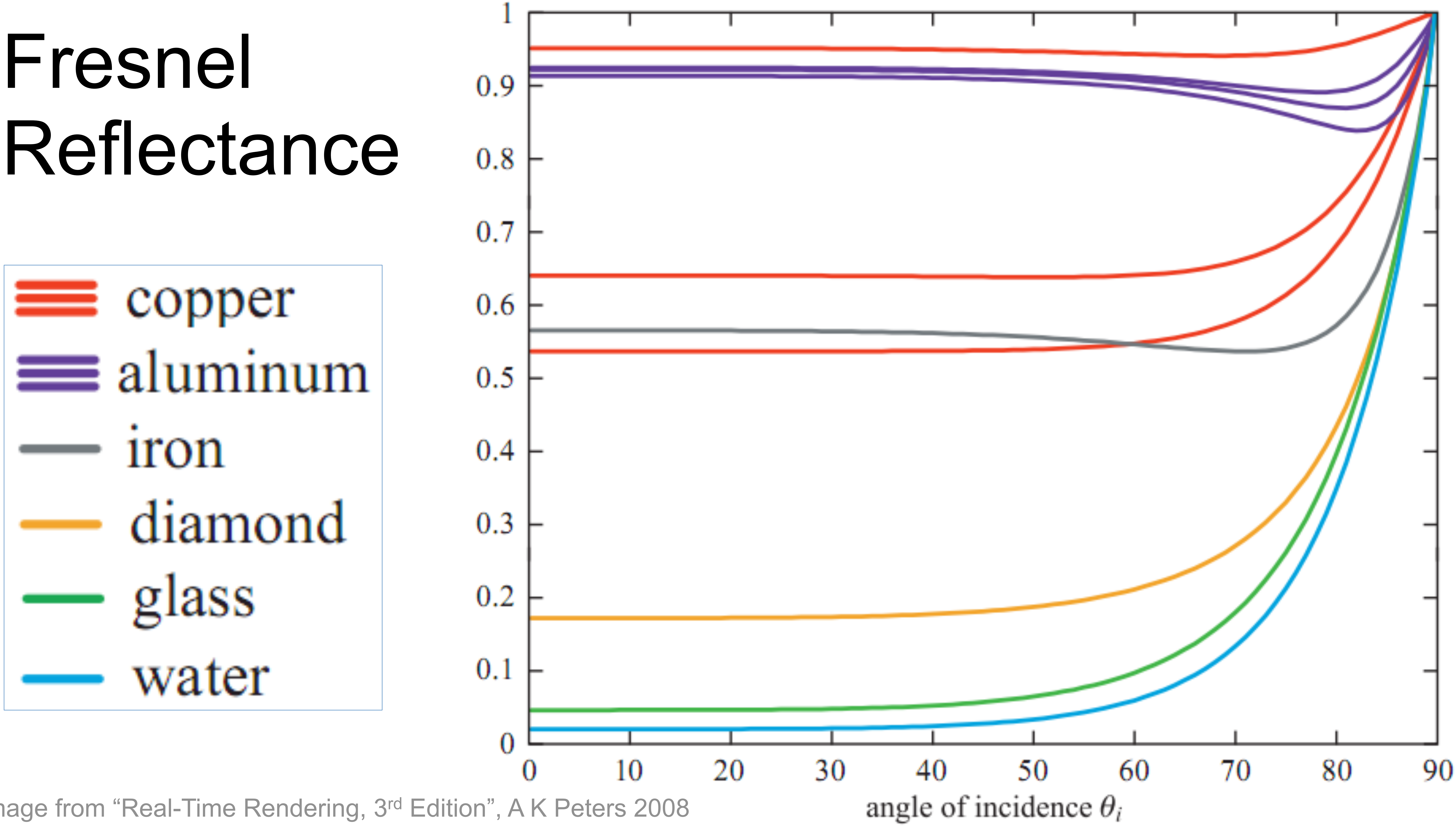


Image from “Real-Time Rendering, 3rd Edition”, A K Peters 2008

Fresnel reflectance (on the y-axis in this graph) depends on refraction index (in other words, what the object’s made of) and the incoming light angle (which is plotted here on the x-axis). In this graph, substances with three lines (copper & aluminum) have colored reflectance, which is plotted separately for the R, G and B channels—the other substances, with one line, have uncolored reflectance.

With an optically flat surface, the relevant angle for Fresnel reflectance is the one between the view and normal vectors. This image shows the Fresnel reflectance of glass (the green curve from the previous slide) over a 3D shape. See how the dark reflectance color in the center brightens to white at the edges.

Fresnel Reflectance

Image from "Real-Time Rendering, 3rd Edition", A K Peters 2008

As the angle increases, the Fresnel reflectance barely changes for the first 45 degrees (the green area on the graph); afterwards it starts changing, first slowly (the yellow area, up to about 75 degrees) and then for very glancing angles (the red zone) it rapidly goes to 100% at all wavelengths.

Here’s a visualization of the same zone colors over a 3D object. We can see that the vast majority of visible pixels are in the areas where the reflectance changes barely at all (green) or only slightly (yellow).

Fresnel Reflectance

$F_0 = F(0°)$

Is the surface's characteristic specular color

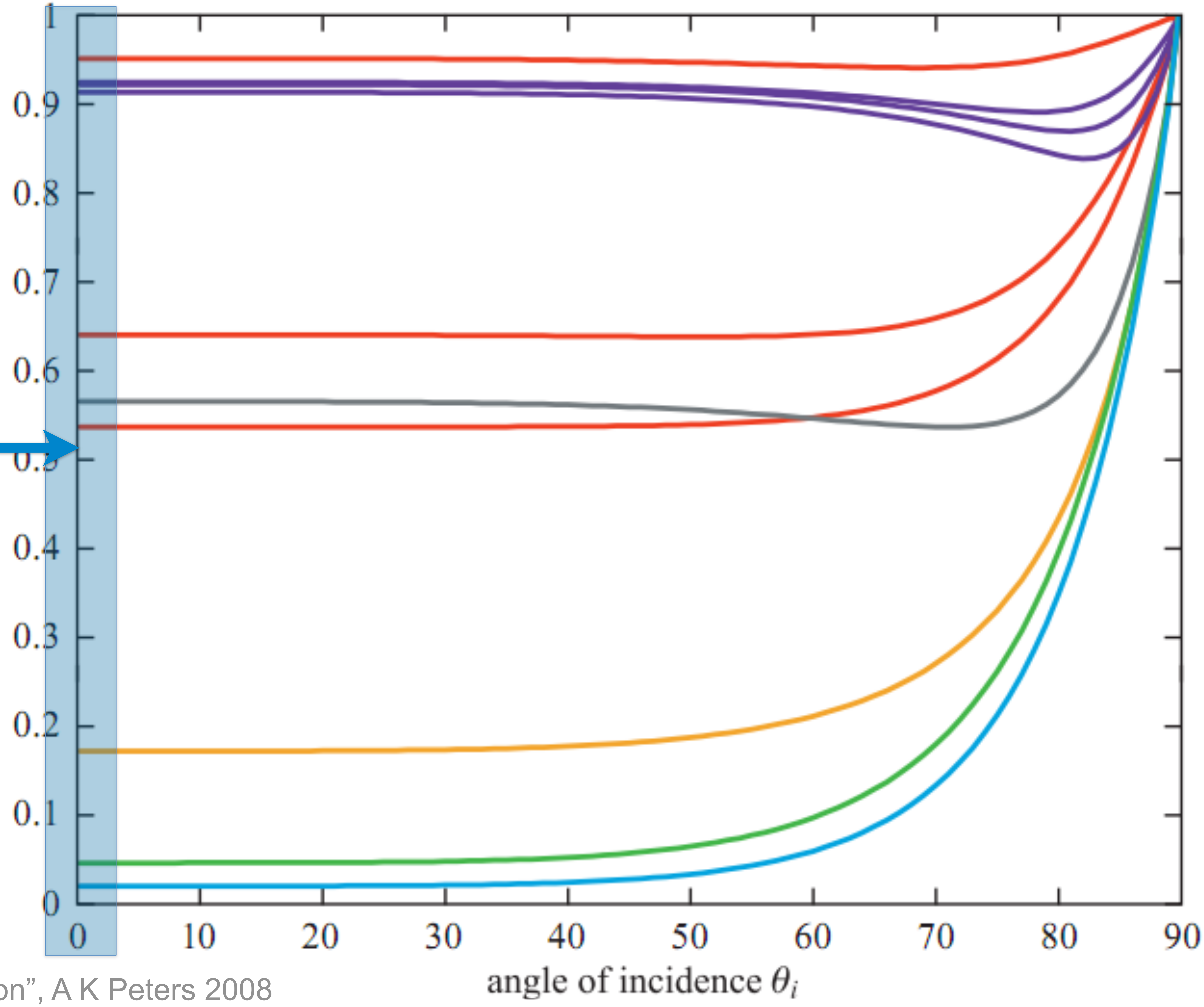


Image from "Real-Time Rendering, 3rd Edition", A K Peters 2008

Since over most of the visible surface the Fresnel reflectance value is similar to the value for 0 degrees, we can treat this value as the surface's characteristic specular color.

Metal	F_0 (Linear, Float)	F_0 (sRGB, U8)	Color
Titanium	0.542,0.497,0.449	194,187,179	
Chromium	0.549,0.556,0.554	196,197,196	
Iron	0.562,0.565,0.578	198,198,200	
Nickel	0.660,0.609,0.526	212,205,192	
Platinum	0.673,0.637,0.585	214,209,201	
Copper	0.955,0.638,0.538	250,209,194	
Palladium	0.733,0.697,0.652	222,217,211	
Zinc	0.664,0.824,0.850	213,234,237	
Gold	1.022,0.782,0.344	255,229,158	
Aluminum	0.913,0.922,0.924	245,246,246	
Silver	0.972,0.960,0.915	252,250,245	

As noted earlier, it’s useful to divide substances into metals, dielectrics and semiconductors. Metals have bright specular; with one exception (gold blue channel), the linear values in this table never go far below 0.5 and most are much higher. Besides linear values, we also give 8-bit sRGB values for texture authoring. Since they lack subsurface scattering, metals get their color from surface reflection.

Metal	F_0 (Linear, Float)	F_0 (sRGB, U8)	Color
Titanium	0.542,0.497,0.449	194,187,179	
Chromium	0.549,0.556,0.554	196,197,196	
Iron	0.562,0.565,0.578	198,198,200	
Nickel	0.660,0.609,0.526	212,205,192	
Platinum	0.673,0.637,0.585	214,209,201	
Copper	0.955,0.638,0.538	250,209,194	
Palladium	0.733,0.697,0.652	222,217,211	
Zinc	0.664,0.824,0.850	213,234,237	
Gold	1.022,0.782,0.344	255,229,158	
Aluminum	0.913,0.922,0.924	245,246,246	
Silver	0.972,0.960,0.915	252,250,245	

Some metals are strongly colored, especially gold; besides an unusually low blue channel value, its red channel value is greater than 1 (it’s outside sRGB gamut). The fact that gold is so strongly colored probably contributes to its unique cultural and economic significance. Despite its low blue value, gold is also one of the brightest metals—this table is ordered by lightness (CIE Y coordinate) of specular color.

F_0 Values for Dielectrics

Dielectric	F_0 (Linear, Float)	F_0 (sRGB, U8)	Color
Water	0.020	39	
Plastic, Glass	0.040 – 0.045	56 – 60	
Crystalware, Gems	0.050 – 0.080	63 – 80	
Diamond-like	0.100 – 0.200	90 – 124	

On the other hand, dielectrics have dark specular colors which are achromatic, which is why this table gives single values instead of RGB triples. They also typically have a diffuse color in addition to the specular color shown in this table, so unlike metals, this is not the only source of surface color.

F_0 Values for Dielectrics

Dielectric	F_0 (Linear, Float)	F_0 (sRGB, U8)	Color
Water	0.020	39	
Plastic, Glass	0.040 – 0.045	56 – 60	
Crystalware, Gems	0.050 – 0.080	63 – 80	
Diamond-like	0.100 – 0.200	90 – 124	

Here we group common dielectrics into categories of increasing F_0 value, from water at 2% through common plastics and glass, then decorative substances, and finally diamonds and diamond simulants. Since the vast majority of dielectrics have F_0 values in the "Plastic and Glass" range, it's not uncommon to cover all dielectrics with a constant representative value such as 4%.

F_0 Values for Semiconductors

Substance	F_0 (Linear, Float)	F_0 (sRGB, U8)	Color
Diamond-like	0.100 – 0.200	90 – 124	
Crystalline Silicon	0.345,0.369,0.426	159,164,174	
Titanium	0.542,0.497,0.449	194,187,179	

What about semiconductors? As you would expect, they tend to have values in between the brightest dielectrics and the darkest metals, as we can see here using silicon as an example. Typically you will never see semiconductor surfaces in production scenes, so for practical purposes the range of F_0 values between 20 and 45 percent is a “forbidden zone” which should be avoided for realistic surfaces.

Fresnel Reflectance

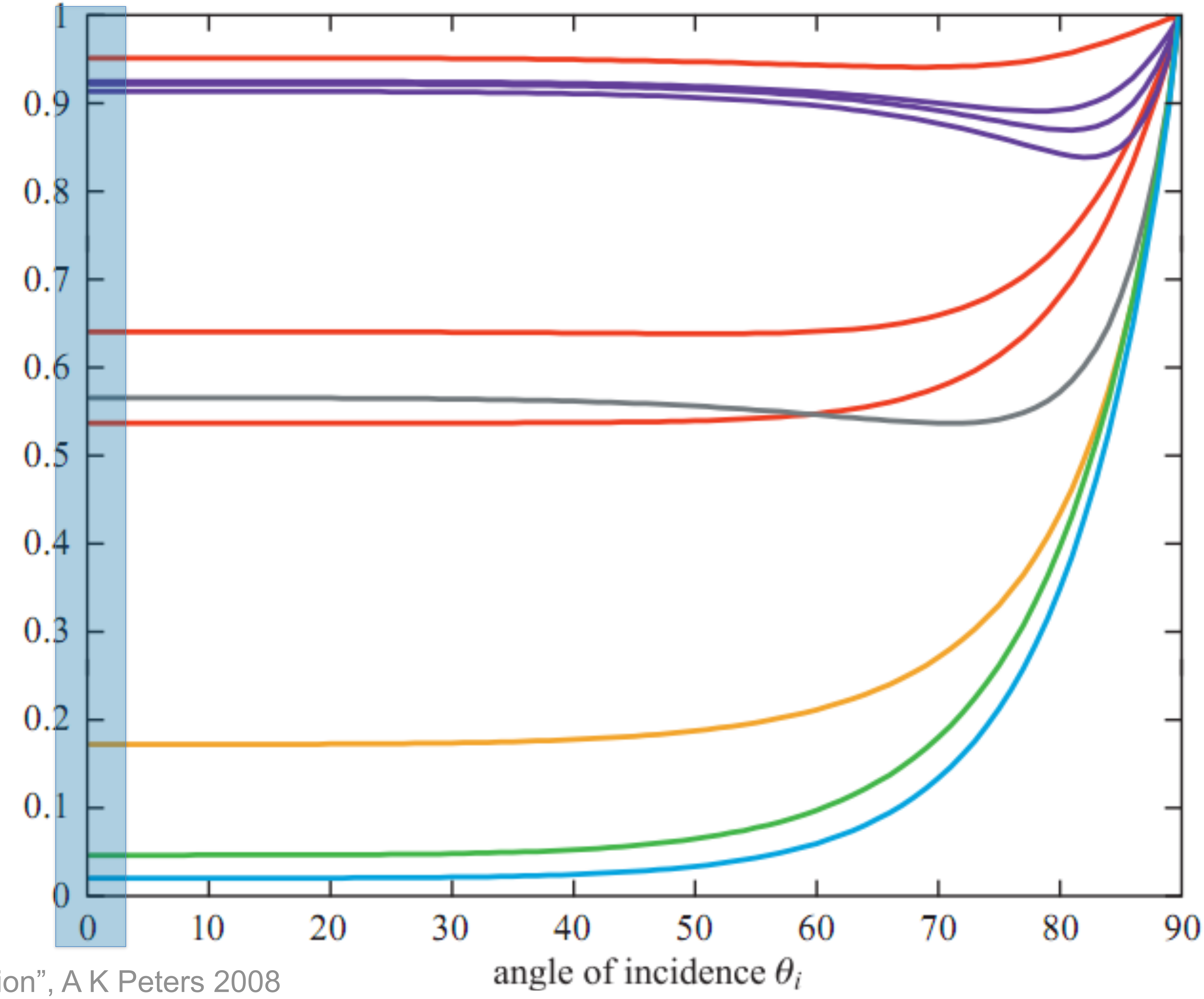


Image from "Real-Time Rendering, 3rd Edition", A K Peters 2008

We've talked about how to get the value for 0 degrees. But what about the angular variation?

The Schlick Approximation to Fresnel

- Fairly accurate, cheap, parameterized by F_0

$$F_{\text{Schlick}}(F_0, \mathbf{l}, \mathbf{n}) = F_0 + (1 - F_0)(1 - (\mathbf{l} \cdot \mathbf{n}))^5$$

- For microfacet BRDFs ($\mathbf{m} = \mathbf{h}$):

$$F_{\text{Schlick}}(F_0, \mathbf{l}, \mathbf{h}) = F_0 + (1 - F_0)(1 - (\mathbf{l} \cdot \mathbf{h}))^5$$

In production, the Schlick approximation to Fresnel is commonly used. It is cheap and reasonably accurate; more importantly, it is parameterized by specular color. As we saw previously, when using it in microfacet BRDFs, the $\mathbf{h}$ vector is used in place of the normal.

Normal Distribution Function

$$f(\mathbf{l}, \mathbf{v}) = \frac{F(\mathbf{l}, \mathbf{h}) G(\mathbf{l}, \mathbf{v}, \mathbf{h}) D(\mathbf{h})}{4(\mathbf{n} \cdot \mathbf{l})(\mathbf{n} \cdot \mathbf{v})}$$

The next part of the microfacet BRDF we will discuss is the microfacet normal distribution function, or NDF. The NDF gives the concentration of microfacet normals pointing in a given direction (in this case, the half-angle direction), relative to surface area. The NDF determines the size and shape of the highlight.

$$D_p(\mathbf{m}) = \frac{\alpha_p + 2}{2\pi}(\mathbf{n} \cdot \mathbf{m})^{\alpha_p}$$

$$D_{uabc}(\mathbf{m}) = \frac{1}{\left(1 + \alpha_{abc1}\left(1 - (\mathbf{n} \cdot \mathbf{m})\right)\right)^{\alpha_{abc2}}}$$

$$D_{tr}(\mathbf{m}) = \frac{\alpha_{tr}^2}{\pi\left((\mathbf{n} \cdot \mathbf{m})^2\left(\alpha_{tr}^2 - 1\right) + 1\right)^2}$$

$$D_b(\mathbf{m}) = \frac{1}{\pi\alpha_b^2(\mathbf{n} \cdot \mathbf{m})^4} e^{-\left(\frac{1 - (\mathbf{n} \cdot \mathbf{m})^2}{\alpha_b^2(\mathbf{n} \cdot \mathbf{m})^2}\right)}$$

$$D_{\text{sgd}}(\mathbf{m}) = \frac{\text{p22}\left[\frac{1-(\mathbf{n}\cdot\mathbf{m})^2}{(\mathbf{n}\cdot\mathbf{m})^2}\right]}{\pi(\mathbf{n} \cdot \mathbf{m})^4}$$

The computer graphics literature details various options for NDFs.

Some NDFs are Gaussian with “blobby” highlights...

Others have a more “spiky” shape with long tails, leading to sharp highlights with “halos” around them.

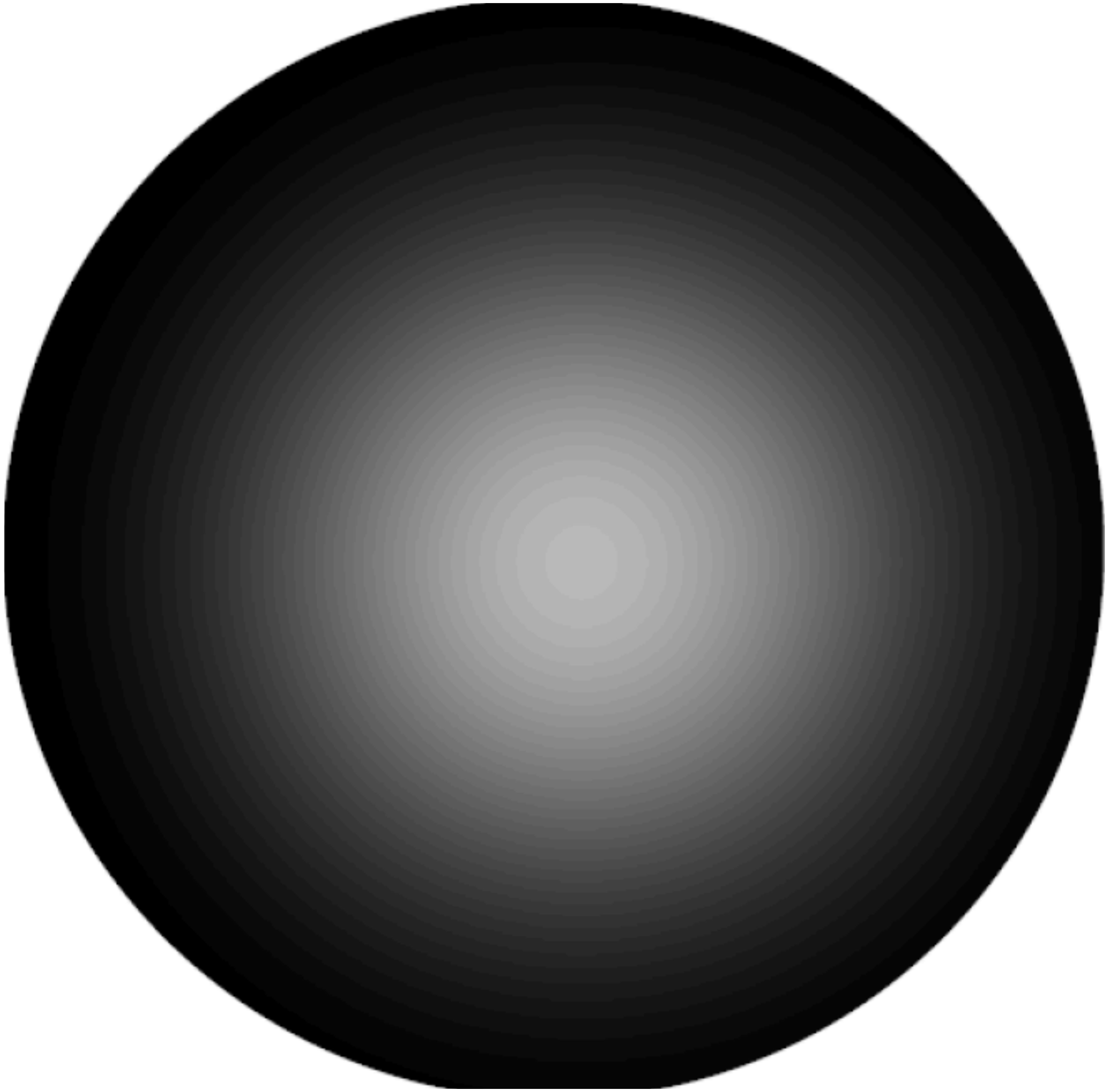


Image from “Rendering Glints on High-Resolution Normal-Mapped Specular Surfaces”, Yan et al., SIGGRAPH 2014

Many surfaces are not well represented by such smooth functions, as I’ll show with some images from the SIGGRAPH 2014 glint rendering paper by Yan et al. Production BRDFs and normal map filtering techniques use smooth lobes that are either isotropic...

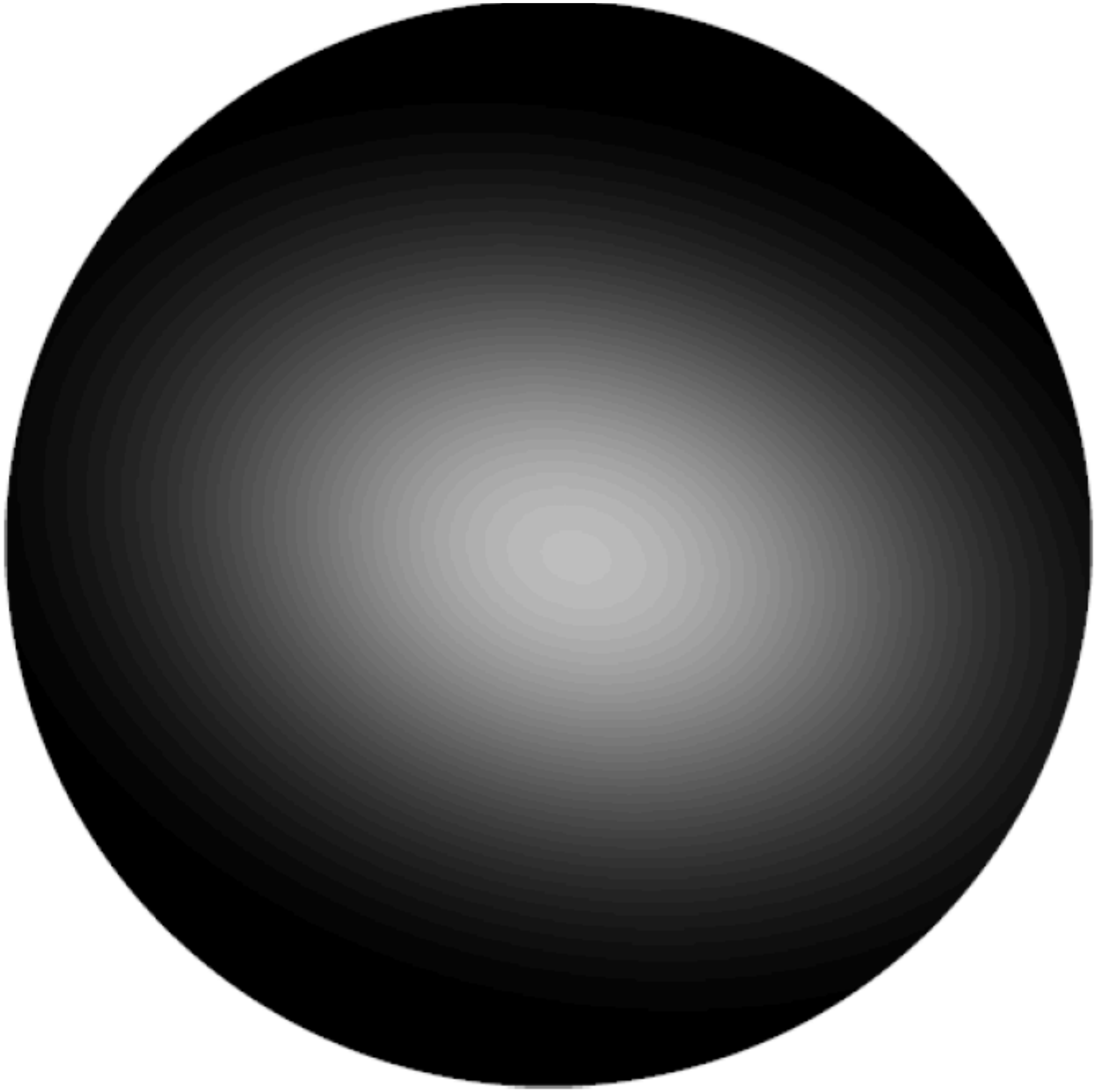


Image from “Rendering Glints on High-Resolution Normal-Mapped Specular Surfaces”, Yan et al., SIGGRAPH 2014

...or anisotropic. However, many surfaces have relatively coarse microgeometry, leading to NDFs that look like...

Image from “Rendering Glints on High-Resolution Normal-Mapped Specular Surfaces”, Yan et al., SIGGRAPH 2014

...this, causing a “glinty” appearance. Though last year’s “glint rendering” paper offered a solution that could be used for film production, it’s too costly for game use. Games use more ad-hoc methods; the snow sparkle talk from the SIGGRAPH 2015 “Advances in Real-Time Rendering” course is a good example of the current state of the art (a more principled real-time solution, from an I3D 2016 paper by Zirr & Kaplanyan, will be presented in this room tomorrow).

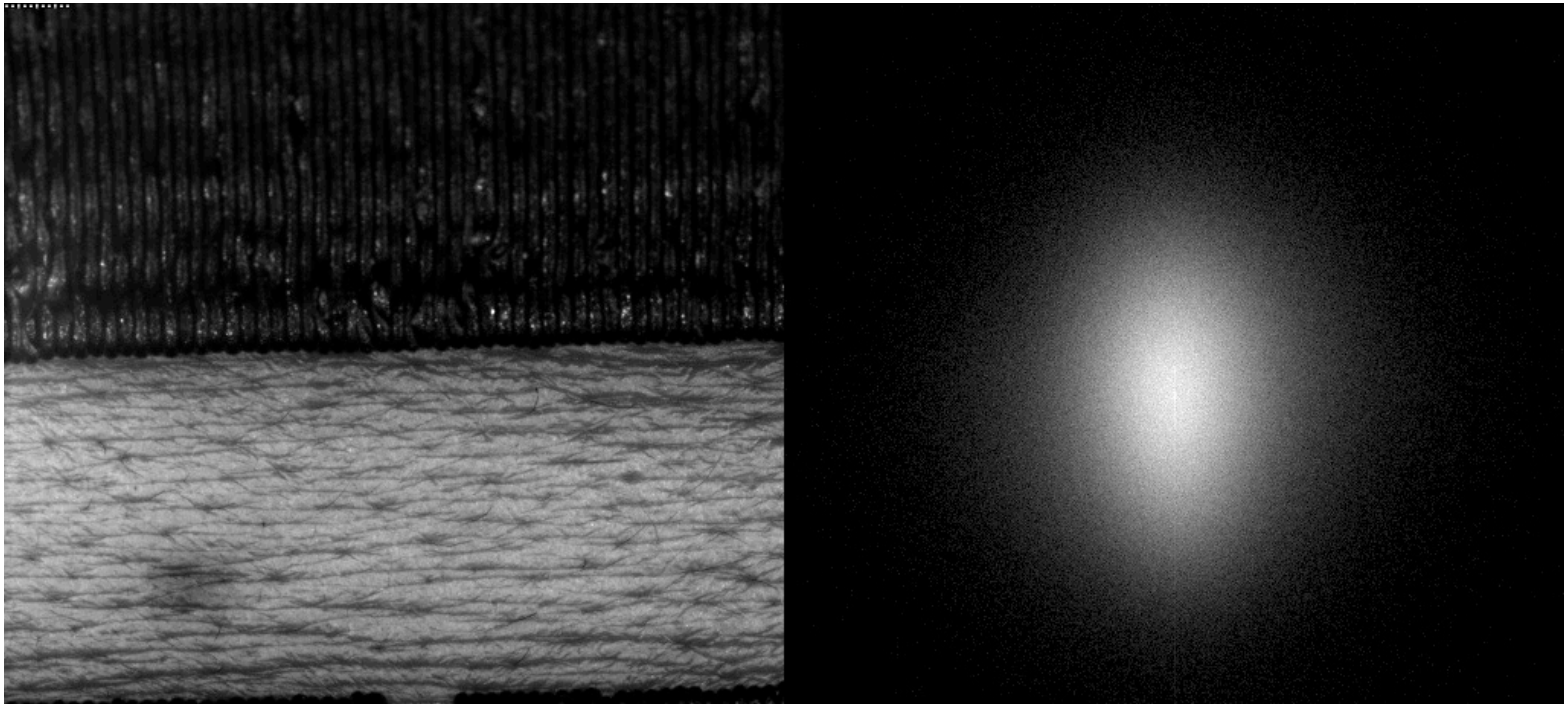


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

It’s important to account for the effect of surface deformation on NDFs. The following images are from the SIGGRAPH 2015 “Skin Microstructure Deformation” paper by Nagano et al. The left side shows a patch of skin under varying amounts of compression & stretch; its NDF is on the right.

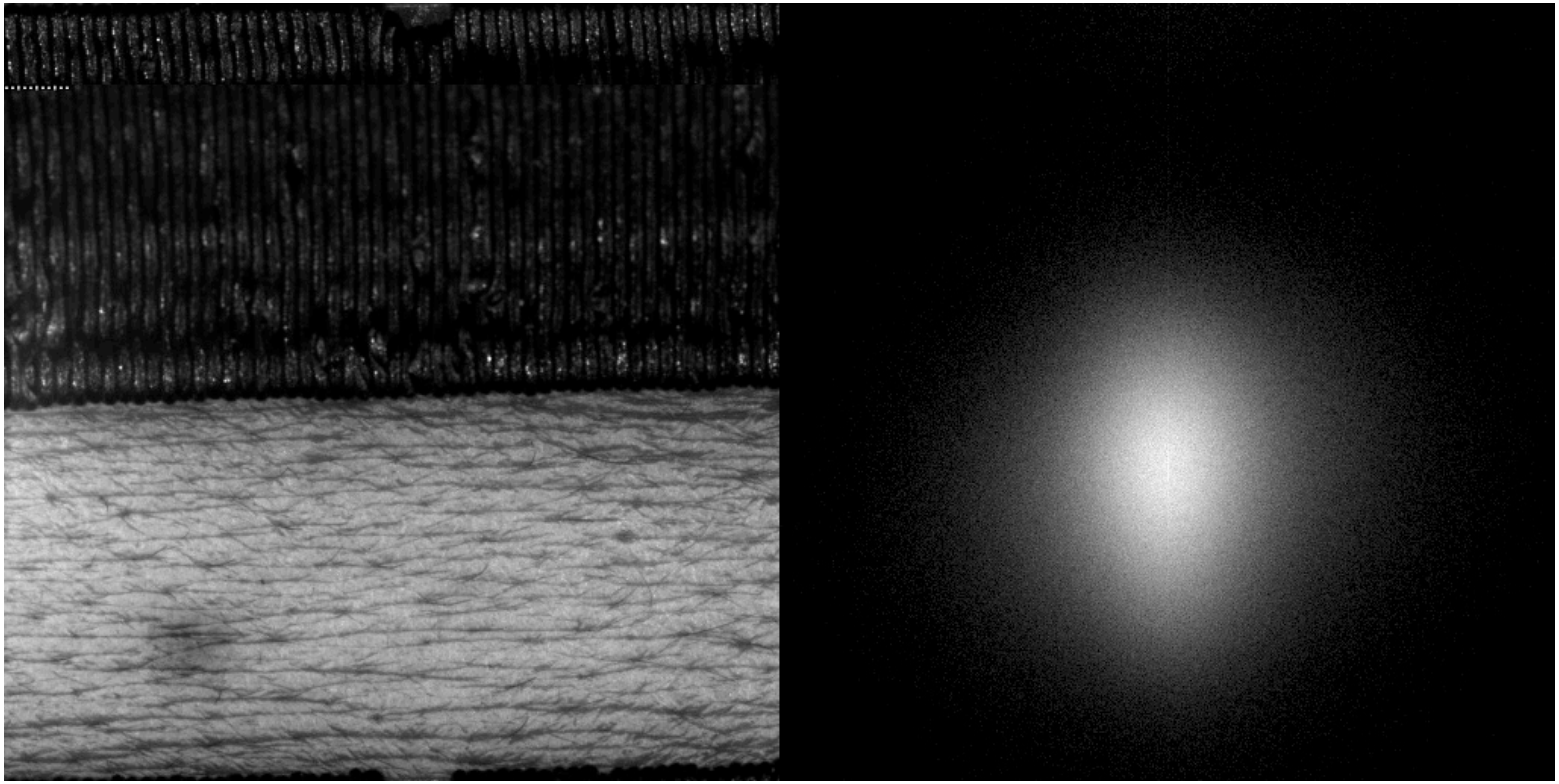


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

We can see that as the patch of skin changes from being compressed...

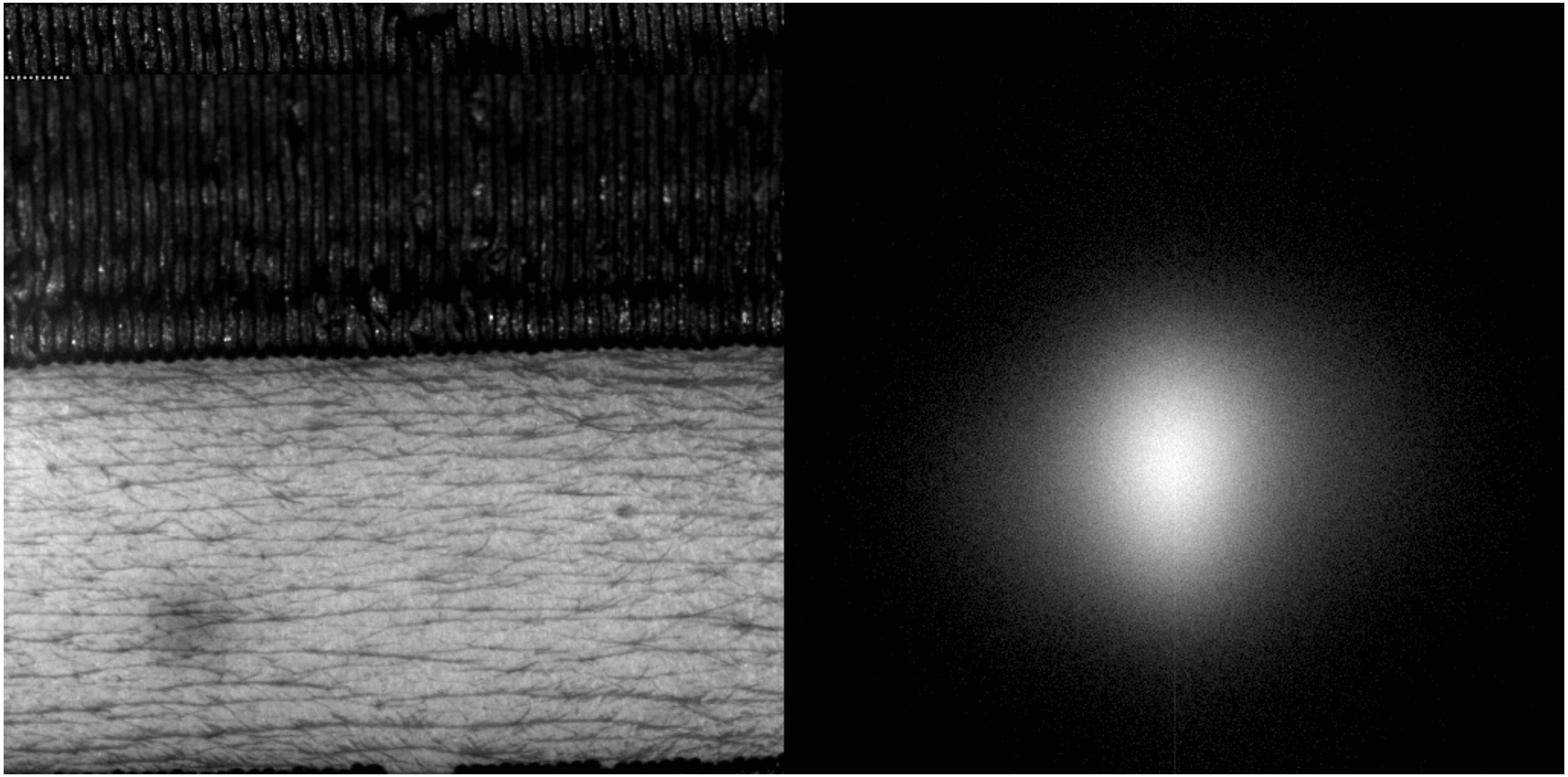


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

…

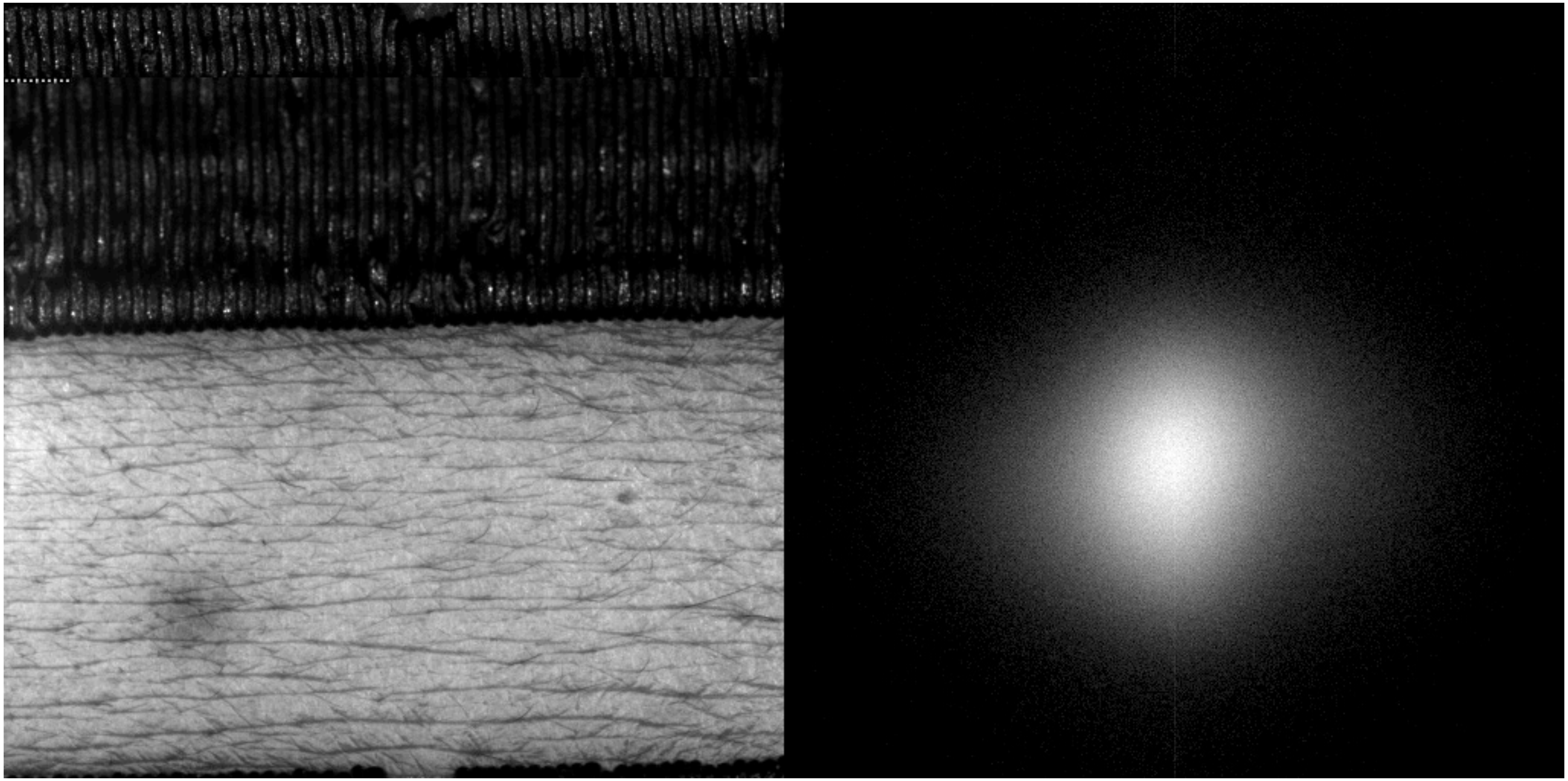


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

…

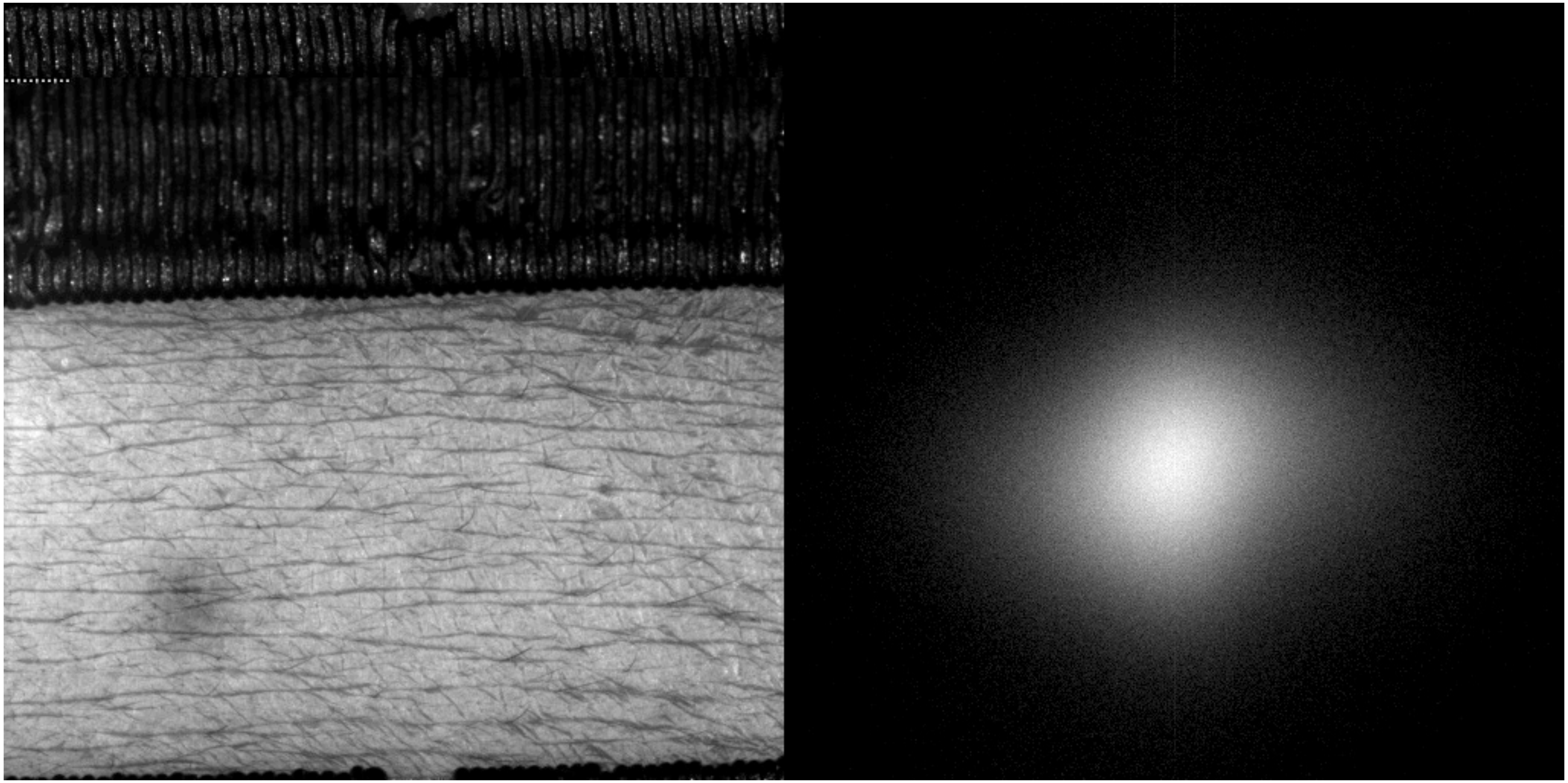


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

...to being stretched, its NDF...

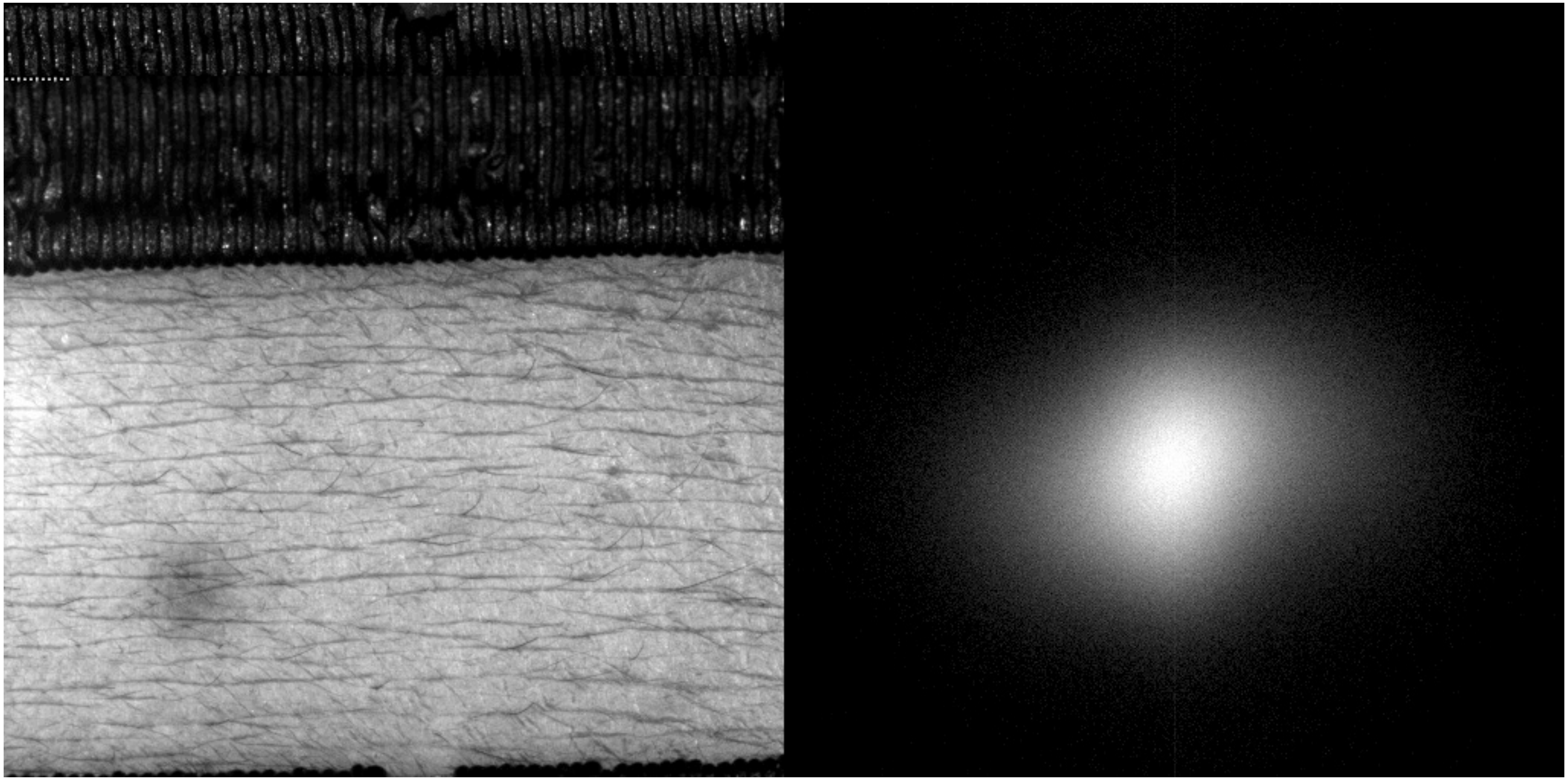


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

...changes...

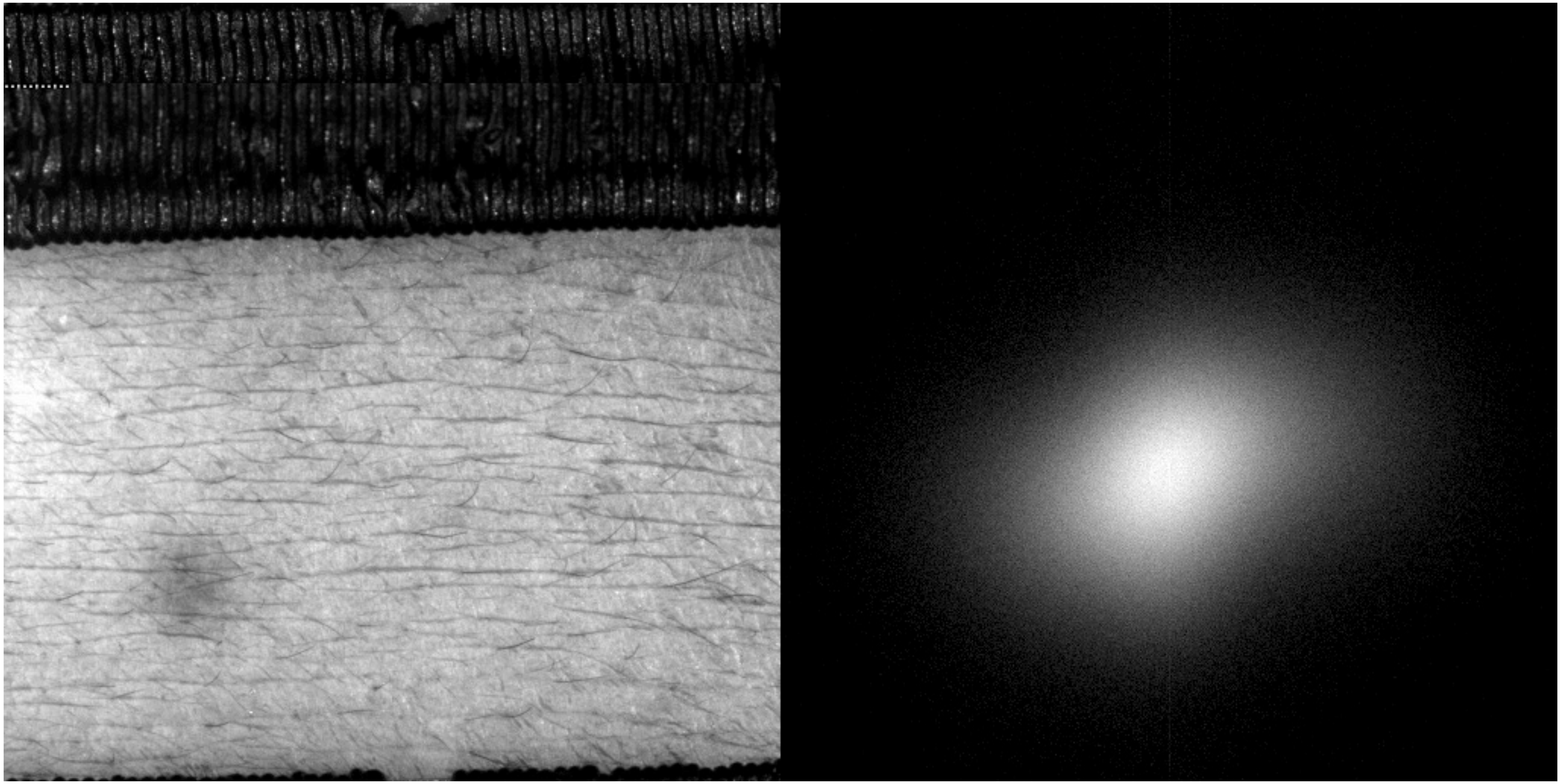


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

...accordingly.

<one more image>

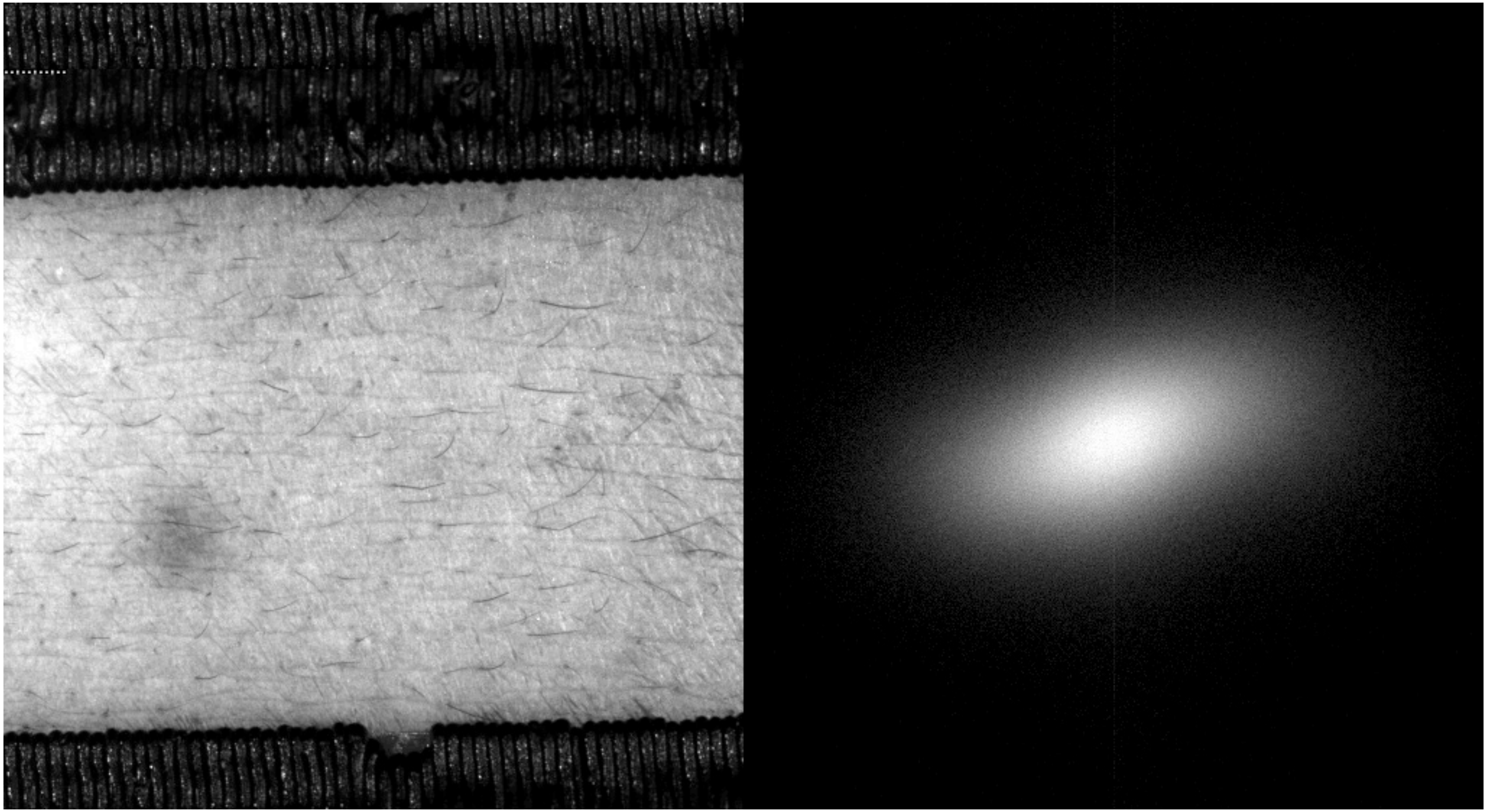


Image from “Skin Microstructure Deformation with Displacement Map Convolution”, Nagano et al., SIGGRAPH 2015

Although this paper was about skin, this type of behavior will occur with any flexible surface material .

Geometry Function

$$f(\mathbf{l}, \mathbf{v}) = \frac{F(\mathbf{l}, \mathbf{h})G(\mathbf{l}, \mathbf{v}, \mathbf{h})D(\mathbf{h})}{4(\mathbf{n} \cdot \mathbf{l})(\mathbf{n} \cdot \mathbf{v})}$$

The geometry or shadowing-masking function gives the chance that a microfacet with a given orientation (again, the half-angle direction is the relevant one) is lit and visible (in other words, not shadowed and/or masked) from the given light and view directions.

$$G_{\mathrm{ct}}(\mathbf{l},\mathbf{v},\mathbf{h}) = \min\left(1, \frac{2(\mathbf{n}\cdot\mathbf{h})(\mathbf{n}\cdot\mathbf{v})}{(\mathbf{v}\cdot\mathbf{h})}, \frac{2(\mathbf{n}\cdot\mathbf{h})(\mathbf{n}\cdot\mathbf{l})}{(\mathbf{v}\cdot\mathbf{h})}\right)$$

$$\frac{G_{\mathrm{ct}}(\mathbf{l},\mathbf{v},\mathbf{h})}{(\mathbf{n}\cdot\mathbf{l})(\mathbf{n}\cdot\mathbf{v})} \approx \frac{1}{(\mathbf{l}\cdot\mathbf{h})^2}$$

$$G_{\mathrm{implicit}}(\mathbf{l},\mathbf{v},\mathbf{m}) = (\mathbf{n}\cdot\mathbf{l})(\mathbf{n}\cdot\mathbf{v})$$

$$G_{\mathrm{s}}(\mathbf{l},\mathbf{v},\mathbf{h}) = G_{\mathrm{s1}}(\mathbf{l},\mathbf{h})G_{\mathrm{s1}}(\mathbf{v},\mathbf{h})$$

The literature has various options for the geometry function. However, Eric Heitz has shown (in a thorough analysis which I recommend reading), that only...

$$G_{\mathrm{ct}}(\mathbf{l},\mathbf{v},\mathbf{h}) = \min\left(1, \frac{2(\mathbf{n}\cdot\mathbf{h})(\mathbf{n}\cdot\mathbf{v})}{(\mathbf{v}\cdot\mathbf{h})}, \frac{2(\mathbf{n}\cdot\mathbf{h})(\mathbf{n}\cdot\mathbf{l})}{(\mathbf{v}\cdot\mathbf{h})}\right)$$

$$\frac{G_{\mathrm{ct}}(\mathbf{l},\mathbf{v},\mathbf{h})}{(\mathbf{n}\cdot\mathbf{l})(\mathbf{n}\cdot\mathbf{v})} \approx \frac{1}{(\mathbf{l}\cdot\mathbf{h})^2}$$

$$G_{\mathrm{implicit}}(\mathbf{l},\mathbf{v},\mathbf{m}) = (\mathbf{n}\cdot\mathbf{l})(\mathbf{n}\cdot\mathbf{v})$$

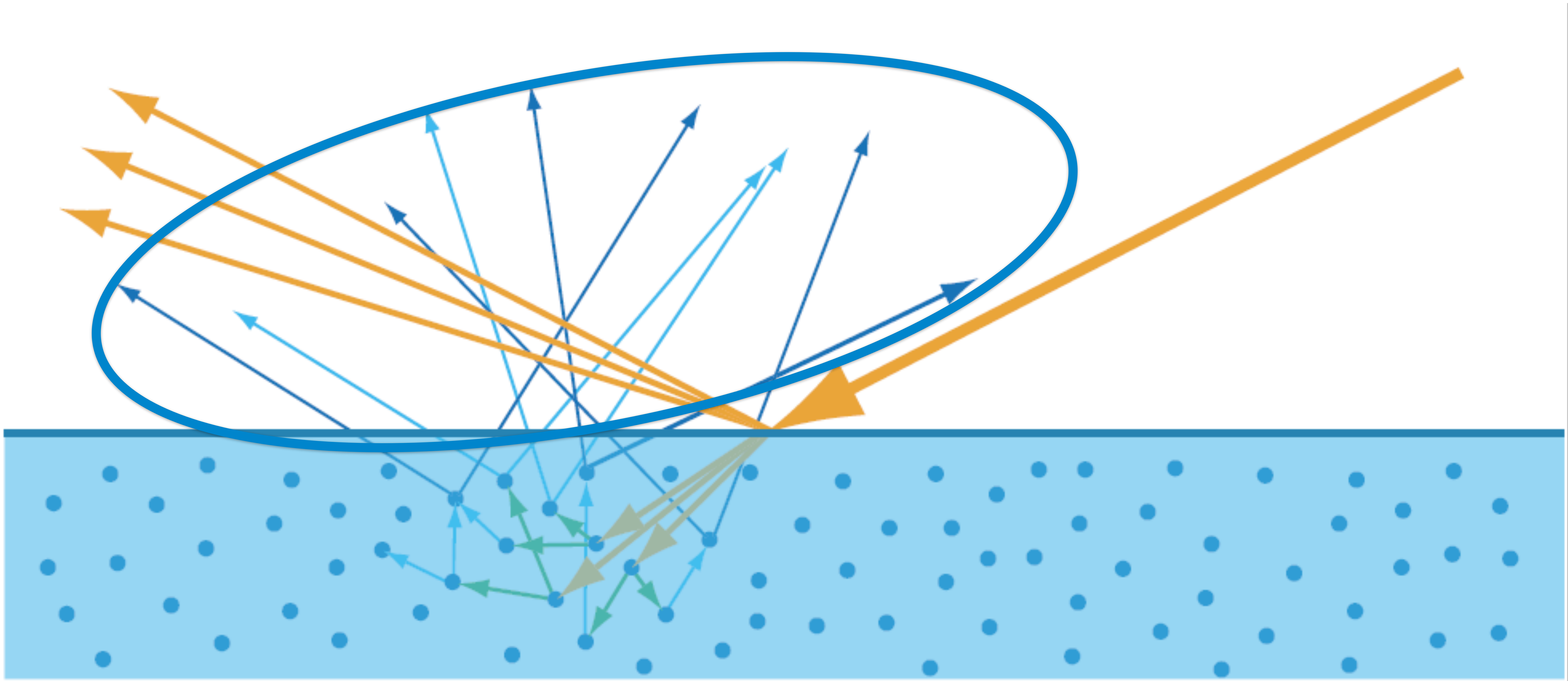
$$G_{\mathrm{s}}(\mathbf{l},\mathbf{v},\mathbf{h}) = G_{\mathrm{s1}}(\mathbf{l},\mathbf{h})G_{\mathrm{s1}}(\mathbf{v},\mathbf{h})$$

...the Smith function (the uncorrelated form of which is shown here) is both mathematically valid and physically realistic. Further details (including various correlated forms of Smith) can be found in Heitz' paper.

$$f(\mathbf{l}, \mathbf{v}) = \frac{F(\mathbf{l}, \mathbf{h})G(\mathbf{l}, \mathbf{v}, \mathbf{h})D(\mathbf{h})}{4(\mathbf{n} \cdot \mathbf{l})(\mathbf{n} \cdot \mathbf{v})}$$

Putting it all together, we see that the BRDF is proportional to the concentration of active microfacets (the ones with normals aligned with h) times their visibility times their Fresnel reflectance. The rest of the BRDF (in the denominator) consists of correction factors relating to the various frames involved (light frame, view frame, local surface frame).

Subsurface Reflection (Diffuse Term)



Until now we've been focusing on the specular—or surface—reflection term. Next, we'll take a quick look at the diffuse (or subsurface) term.

Lambert

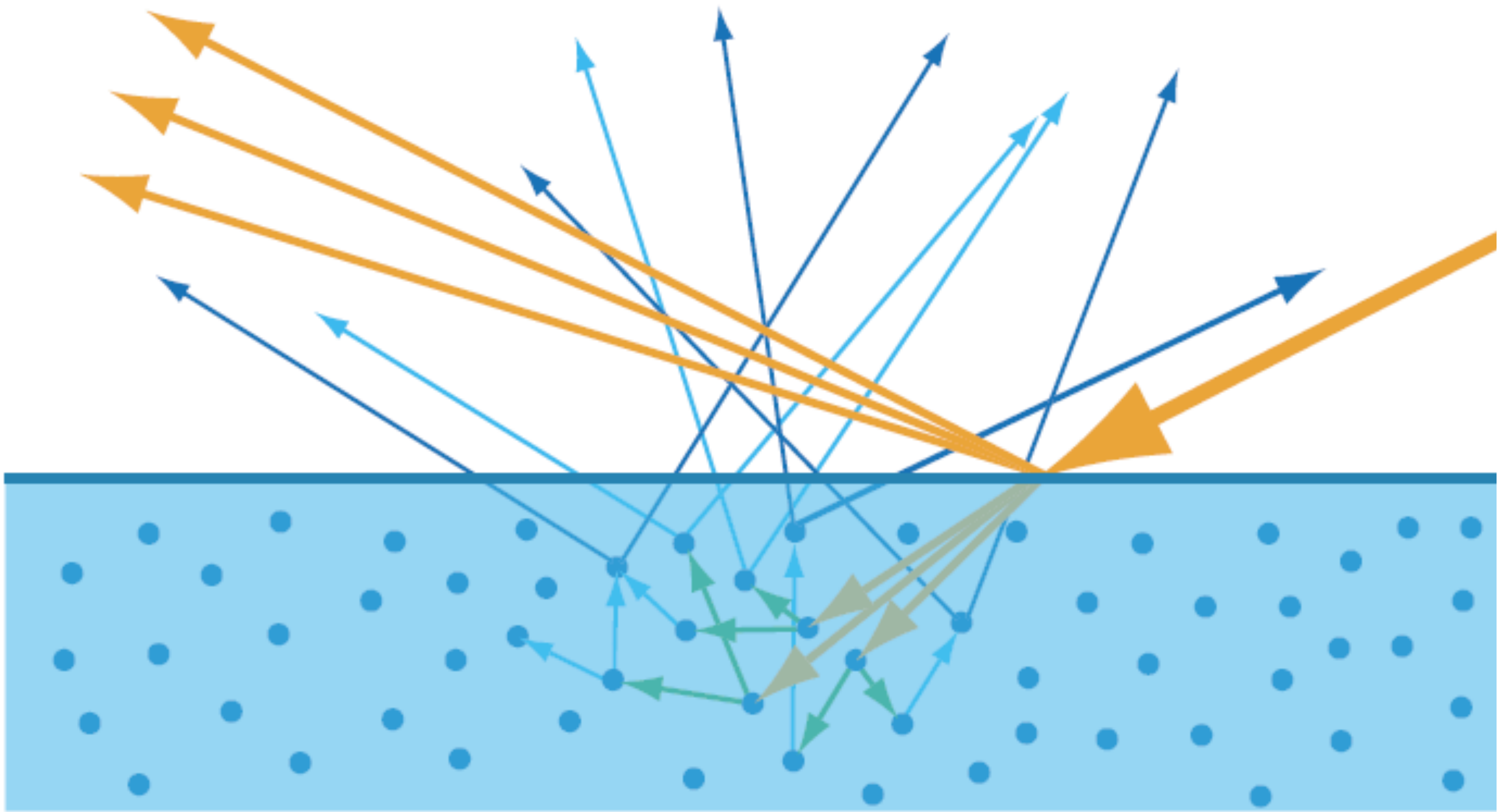
- Constant value ($\mathbf{n} \bullet \mathbf{l}$ is part of reflectance equation):

$$f_{\text{Lambert}}(\mathbf{l}, \mathbf{v}) = \frac{\mathbf{c}_{\text{diff}}}{\pi}$$

- $\mathbf{c}_{\text{diff}}$: fraction of light reflected, or diffuse color

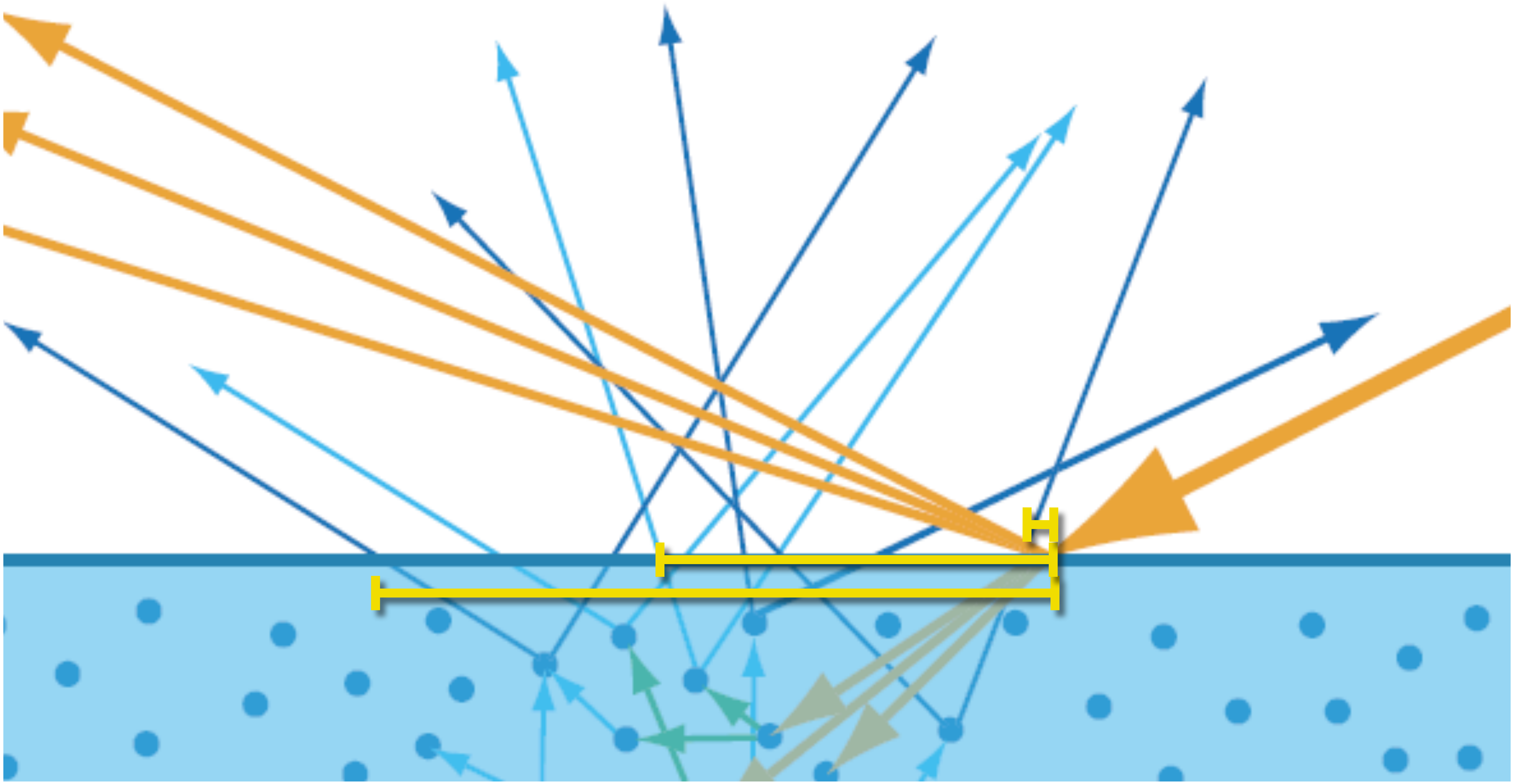
The Lambert model is the most common diffuse term used in game and film production. By itself, it's the simplest possible BRDF: a constant value. The well-known cosine factor is part of the reflectance equation, not the BRDF.

Beyond Lambert: Diffuse-Specular Tradeoff



There are a few important physical phenomena that Lambert doesn't account for. Diffuse comes from refracted light. Since the specular term comes from surface reflection, in a sense it gets "dibs" on the incoming light and diffuse gets the leftovers. Since surface reflection goes to 100% at glancing angles, it follows that diffuse should go to 0%. The literature discusses a few ways to model this.

Beyond Lambert: Surface Roughness



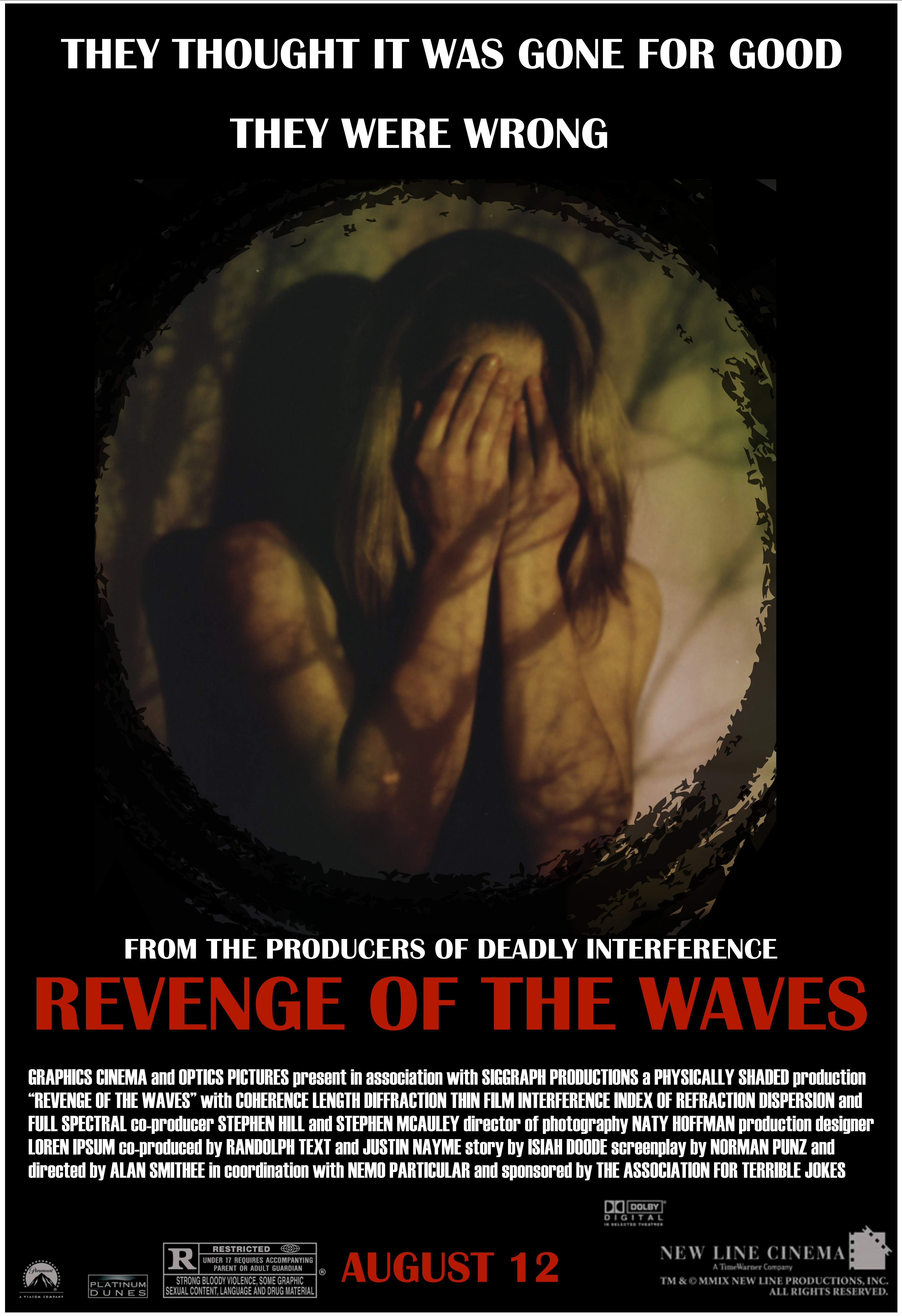
Lambert also doesn't account for surface roughness. In most cases, microscopic roughness only affects specular; diffuse reflectance at a point comes from incoming light over an area, which tends to average out any microgeometry variations. But some surfaces have microgeometry larger than the scattering distance, and these do affect diffuse reflectance. That's when you need models like Oren-Nayar.

Diffuse Roughness = Specular Roughness

It's become common to use rough diffuse models like Oren-Nayar or "Disney diffuse" for all surfaces, and to plug the specular roughness into them. But I want to take this opportunity to point out a problem with this approach.

Diffuse R ughness Specula Roughness

It’s been known for a while that diffuse response effectively smooths out small bumps, as can be seen from LightStage’s separate diffuse and specular normal maps. But this applies even more strongly to roughness. Ideally you should use a separate roughness value for these models; otherwise use them sparingly, only for materials where you know the microgeometry is larger than the scattering distance.



I'll briefly go back to wave optics, which I'm sure you thought I had forgotten about after I abandoned it earlier in the talk.

Image by flickr user 55Laney69; licensed CC-Attribution (https://creativecommons.org/licenses/by/2.0/)

Diffraction from Optically-Smooth Surface

With few exceptions, the computer graphics community has either ignored the effects of nanogeometry diffraction or has asserted their insignificance. However, in a recent white paper, Holzschuch and Pacanowski showed convincing evidence that part of visible BRDF behavior (the “long tail” of the highlights in particular) was due to this phenomenon.

Microgeometry & Nanogeometry

Microgeometry	**Nanogeometry**
Lobe shape determined by surface statistics (micro-scale NDF)	Lobe shape determined by surface statistics (nano-scale SPD)
No wavelength dependence	Strong wavelength dependence
Incidence angle may affect surface statistics via visibility	Incidence angle may affect surface statistics via foreshortening

It appears that in many materials, reflectance is affected by roughness on both the micro- and nano- scales. I'll go over some high-level differences between the two; for more detail see Holzschuch & Pacanowski's white paper. The Nano-scale lobe shape is controlled by the surface SPD (similar to the SPDs we saw earlier, but with respect to 2D surface spatial frequency rather than 1D wave temporal frequency). *<read rest>*

Math → Rendering

Once you have the math, the next step is to implement it in a film or game renderer. This is beyond the scope of this talk, but many talks at GDC and SIGGRAPH (especially in the SIGGRAPH physically based shading course) have additional details.

Acknowledgements

- Steve Hill: assistance with course notes & slides, WebGL framework used for Fresnel visualization
- Brent Burley, Paul Edelstein, Yoshiharu Gotanda, Eric Heitz, Christophe Hery, Sébastien Lagarde, Dimitar Lazarov, Cedric Perthuis, Brian Smits: inspirational discussions on physically based shading models
- A K Peters, ACM, authors of “Rendering Glints on High-Resolution Normal-Mapped Specular Surfaces” and “Skin Microstructure Deformation with Displacement Map Convolution” papers: permission to use images

I’d like to thank some people who helped me with this talk...

...and end by noting that 2K is hiring: there are open positions at many of our studios, and my central tech department is looking for top-notch rendering & engine programmers.