slide1:

Welcome, everyone. Today we begin Lecture Eleven, which is all about image quality assessment. This is really a landmark lecture in our course. After today, we'll move on to imaging modalities themselves. By then, the green textbook you have will serve as your main reference, and I'll continue using my slides to bring in a consistent story and sometimes involving state-of-the-art developments.

So why is quality assessment so important? Think about it this way: once an imaging system produces an image, the very first question we must ask is, is this image good enough? Is it clear, reliable, and suitable for diagnosis? That is what quality assessment is all about — it's the final piece of the foundation we need before diving into the modalities.

So, with that background, let's begin our journey into image quality assessment.

slide2:

We are right on schedule in our journey through this material.

If you've had a chance to look over the reading materials for today's lecture, that's great — it will make it easier to connect the ideas we cover. If not, that's fine too. I encourage you to follow along closely and take time afterward to review the main points. Consistently reinforcing concepts as you go will help you develop a stronger and more intuitive grasp, especially as the material becomes more mathematically detailed

slide3:

Here you can see the textbook's outline of Chapter Five, which I have actually moved earlier in our course. The reason is simple: once you understand the basics of image quality assessment, you can apply these ideas to almost any imaging modality.

No matter whether we're talking about X-ray, CT, MRI, ultrasound, or optical imaging, the general principles remain the same. Concepts such as resolution, signal-to-noise ratio, contrast-to-noise ratio, and artifacts are universal. They may appear in different forms, but the underlying logic applies everywhere.

The good news is that this chapter is relatively easy reading compared to some of the heavier mathematical topics we've already tackled. The ideas are straightforward. If you follow carefully and think through the examples, you'll see how each concept connects to real images.

For today's lecture, I will guide you mainly through three key aspects of image quality assessment. We'll start with general mathematical measures, then move to system-specific characteristics, and finally consider task-specific and human-observer models. This will give you a well-rounded view of what it means to evaluate image quality in practice.

slide4:

Here's the outline for today's lecture. We'll approach image quality assessment from three perspectives: general measures, system-specific measures, and task-specific measures.

First, we'll talk about general measures. Imagine you have an image and you also know the ground truth — the ideal image you want to reconstruct. The goal is to measure how close your reconstructed image is to the ground truth. The simplest way is to calculate the difference between the two. In mathematics, we often describe this using the idea of a distance in a vector space. The most familiar example is the Euclidean distance, which leads directly to the mean squared error, or MSE.

There are other distances too. One example I'll briefly mention is the Kullback–Leibler divergence, or KL distance. This is a bit trickier, so I won't test you on it, but it's good to be aware of. And then there is a very practical and powerful concept called structural similarity, or SSIM. This has become extremely influential — the original paper has been cited tens of thousands of times — because it captures how humans actually perceive similarity between images. We'll look at why this idea is so effective.

Next, we move to system-specific measures. Think of the imaging system as a kind of camera — it could be an X-ray detector, an ultrasound probe, or an MRI scanner. Every camera has its own specifications. Here we'll talk about how noisy the images are, how well the system can distinguish a signal from background noise, and what kind of resolution it provides. Resolution itself has several dimensions: spatial resolution, contrast resolution, temporal resolution, and spectral resolution. We'll also talk about artifacts — those misleading structures that appear in images even though they don't exist in reality. Understanding noise, resolution, and artifacts is key to judging the performance of any imaging system.

Finally, we'll cover task-specific measures. This perspective is slightly different. Instead of asking "How good is my camera?", we ask "How well does this imaging system perform a specific clinical task?" For example, can it reliably detect whether a bone is fractured? Can it identify a tumor in the lung? The most important thing in medicine is whether the imaging system helps doctors and patients make correct decisions. Even if a system isn't perfect in a technical sense, if it consistently allows us to succeed in the clinical task, then it's doing its job.

So these three aspects — general, system-specific, and task-specific — are connected, but they are not identical. Together they give us a complete picture of how to assess image quality. We'll begin with the first: general measures.

slide5:

Let's begin with one of the most basic and widely used quality measures: the Mean Squared Error, or MSE.

Suppose we have two images: the true image, which we'll call y, and the reconstructed image, which we'll call y-hat. Each image is made up of many pixels, indexed by i. If the image is 512 by 512, then the total number of pixels, n, is over 260,000.

The formula for MSE is simple:MSE equals one over n, times the sum from i equals one to n, of the difference between y-sub-i and y-hat-sub-i, squared.

In words, this means we compare the two images pixel by pixel. At each pixel, we take the difference, square it so that positive and negative errors don't cancel out, and then average over all pixels. That gives us the mean squared error.

Now, let's go a bit deeper. When we estimate a parameter — say theta — we often write the estimate as theta-hat. The error between theta-hat and the true value theta can be analyzed in expectation, meaning

averaged over many trials. When you expand the algebra, you find that the mean squared error naturally splits into two parts.

The first part is the variance. This tells us how much our estimates fluctuate around their average value. You can think of variance as a measure of random scatter.

The second part is the bias squared. This measures the difference between the average of our estimates and the true value. If our method consistently overshoots or undershoots, that's bias.

So, in summary: MSE equals variance plus bias squared. Variance captures random error, bias captures systematic error, and together they define the total error.

This decomposition is very useful. It reminds us that an algorithm might have low variance but high bias, or vice versa. Understanding both helps us judge the quality of an estimator or an image reconstruction method.

slide6:

Now, the mean squared error is not the only way to measure differences. There are several variants, each with slightly different properties.

The first, which we've already discussed, is the Mean Squared Error, or MSE. This is the average of the squared differences between prediction and truth.

A closely related measure is the Root Mean Squared Error, or RMSE. Here we simply take the square root of the mean squared error. Why? Because this brings the units back to the same scale as the original measurement. For example, if we are measuring pixel intensities, RMSE will be expressed in the same units as those intensities, which makes it easier to interpret.

One important property of squaring is that it emphasizes larger errors much more strongly. If a difference is 100, squaring turns it into 10,000. That means MSE and RMSE heavily penalize large deviations.

Sometimes we want a measure that treats all errors more equally. That's where the Mean Absolute Error, or MAE, comes in. Instead of squaring, we take the absolute value of the difference at each pixel, then average. This is also called the L1 norm, while MSE is associated with the L2 norm. The L1 norm is less sensitive to outliers compared to the L2 norm.

Finally, we have the Mean Absolute Percentage Error, or MAPE. This is the mean absolute error expressed as a percentage of the true value. In other words, it's MAE divided by the ground truth at each point, multiplied by 100 percent. This can be useful when we want to understand an error in relative terms — for example, saying "the error is 5 percent" rather than giving a raw number.

So, these different distance measures — MSE, RMSE, MAE, and MAPE — give us different perspectives on error. The choice depends on the problem: do we want to penalize large errors more, or do we care more about relative error percentages?

slide7:

So far, these measures seem very reasonable. Think about it this way: you have one signal or one image, and you also have a standard — the ground truth.

By comparing them pixel by pixel, we're essentially measuring the difference between two curves, or between two surfaces, or even between two volumes in three dimensions.

The formula reduces to something very intuitive: it's about the area between the two curves. The yellow region you see here represents those differences. The larger the area, the greater the error.

This is why MSE and related measures are so widely used — they give us a direct and interpretable way to say how close or how far two images are. It's simple, mathematically neat, and visually intuitive.

But here's an important point: while this is a good first step, it is not the whole story. Measuring differences point by point tells us something, but not everything. In medical imaging, we also care about structural similarity, system behavior, and clinical tasks. So, as we continue, you'll see that this is only the beginning.

slide8:

Now, let me briefly mention another type of distance, called information divergence. This is where probability theory comes in.

Suppose you don't just have two images, but instead you have two probability distributions — for example, two different histograms of pixel values. The question becomes: how do we measure the difference between these two distributions?

One option is to use Euclidean distance, just as before. But there is a more meaningful way in the context of information theory. This is called the Kullback–Leibler divergence, or KL distance for short.

The formula looks a bit unusual:KL divergence equals the sum over x of p of x, multiplied by the logarithm of p of x divided by q of x.

You don't need to worry too much about the details — this is beyond the scope of our lecture — but the idea is important. The KL divergence is always greater than or equal to zero, and it becomes exactly zero if and only if the two distributions are identical.

One interesting property is that the KL divergence is not symmetric. In other words, the distance from P to Q is not the same as the distance from Q to P. That may sound strange, but it has a good analogy. Think of climbing a mountain: going uphill is much harder than going back downhill, even though it's the same physical path. In the same way, KL divergence measures directionality in information.

So, while we won't use this directly in our course, it's good to be aware that such information-based distances exist. They play a big role in areas like machine learning and statistical signal processing.

slide9:

Now, just for your broader knowledge, let me connect this to another important concept in information theory: mutual information.

Mutual information is a way to measure how much knowing one random variable tells us about another. For example, suppose we have two variables, X and Y. If they are completely independent, then measuring X

tells us nothing about Y. In that case, their mutual information is zero. On the other hand, if X and Y are perfectly dependent — meaning that once you know X, you completely know Y — then their mutual information is very high. Most real-world situations fall somewhere in between.

Mathematically, mutual information can actually be expressed in terms of the KL divergence. Specifically, it's the KL divergence between the joint distribution of X and Y, and the product of their marginal distributions. Don't worry about the details of the formula — the key idea is that it quantifies how much information one variable provides about the other.

You can also think of it this way: when you measure one variable, how much does your uncertainty about the other variable decrease? That decrease in uncertainty is exactly what mutual information captures.

In practice, this idea is very useful in areas like image registration, where we align two images. Instead of just matching pixel intensities, we can maximize the mutual information between the two images. That way, even if the images look very different in terms of brightness or contrast, we can still measure how well they correspond.

So, mutual information is essentially a generalization of correlation, but in the language of information theory. It goes beyond simple linear relationships and captures any kind of statistical dependence.

slide10:

Mutual information, which we just discussed, is actually defined in terms of another very fundamental concept: entropy.

So, what is entropy in this context? Think of it as a measure of uncertainty. If you have a probability distribution that is very spread out and uniform, then you have a lot of uncertainty — you don't really know what the outcome will be. That means the entropy is high.

On the other hand, if the distribution is sharply peaked — like a delta function, where one outcome is guaranteed — then there is no uncertainty. In that case, the entropy is very low, even zero.

So entropy tells us how much information, or how much unpredictability, is contained in a random variable. In information theory, this is a central concept because information itself is really about reducing uncertainty.

Now, mutual information can be written as the difference between two entropies: the entropy of Y by itself, minus the entropy of Y given X. In other words, it measures how much uncertainty about Y is reduced when you know X. That's exactly what we mean by "how much does X tell us about Y."

Again, you don't need to get bogged down in the formulas here. The key point is: entropy captures uncertainty, and mutual information measures how two variables share or reduce that uncertainty.

For our purposes, I just want you to know that these information-theoretic measures — KL divergence, mutual information, and entropy — are very powerful, but they are more advanced than what we need right now. Think of them as tools in the background, which complement the simpler measures like mean squared error.

And with that, we will soon return to the more practical measures that are directly used in image quality assessment.

slide11:

Now let's see why mean squared error, or MSE, is not always good enough.

At the top left, we have the original image — the best version, taken under ideal conditions. Below it, you see five different degraded versions of the same image. Some look noisy, some are blurry, and some have other distortions. Clearly, to the human eye, these images do not look equally good.

But here's the problem: when we compute the mean squared error between the original image and each of these five degraded ones, the result is the same — two hundred and twenty-five. Mathematically, MSE tells us they are equally different from the original.

Visually, though, that's obviously not true. Some versions look much closer to the original, while others look far worse. Our eyes immediately pick up those differences, but MSE cannot.

And this is the key point: MSE does not reflect human perception very well. It measures pixel-wise differences, but it cannot capture whether the overall structure of the image is preserved.

This leads us to the next important idea — we need a measure that better matches what humans actually see. That's where structural similarity, or SSIM, comes in.

slide12:

Now let's think about this from the perspective of the human visual system, or HVS.

The human eye does not look at images pixel by pixel, the way mean squared error does. MSE simply compares each pixel individually, treats every error the same, and then adds them up. That's a bottom-up approach — starting from the smallest units and working upward.

But our visual system works very differently. We are much more sensitive to structural information — the patterns, edges, and relationships that give an image its overall form. For example, even a small change in the background, or a shift in texture, is something we can notice right away. Our brains are highly adapted to detect these kinds of contextual changes.

In the classical view, the focus was on error visibility: if you see a discrepancy, count it as an error. In the newer view, the focus shifts to structural distortion: what matters is whether the structure of the image has been preserved.

This also connects to the way our vision system interprets images. Rather than starting with tiny details, many researchers argue that we first process the global structure — sometimes described in terms of topological features. These are high-level properties, such as whether objects are connected, how many distinct regions there are, or whether certain shapes remain intact. Importantly, these properties don't change if the image is stretched, rotated, or rescaled.

So here's the philosophical shift: instead of measuring differences pixel by pixel, we want to measure how much of the structural information has been preserved. That is what structural similarity is all about, and it's why it is far more aligned with how humans actually see images.

slide13:

Here we arrive at what I like to call an instant classic. This is the landmark 2004 paper by Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli, titled Image Quality Assessment: From Error Visibility to Structural Similarity.

Before this work, most image quality measures focused on error visibility — essentially counting differences pixel by pixel, the way mean squared error does. What this paper introduced was a completely different perspective: instead of measuring errors, we should measure structural similarity.

The idea is built on the assumption that the human visual system is highly tuned to extract structural information from images. So rather than asking, “How many errors can we see?”, the SSIM framework asks, “How well is the structure of the image preserved?”

This shift in thinking turned out to be incredibly powerful. The paper has been cited tens of thousands of times, and SSIM quickly became a standard tool not only in image processing research but also in practical applications like video compression, image restoration, and medical imaging.

So, from this point forward, when we talk about perceptual image quality, we are really building on the foundation laid by this work.

slide14:

Now let’s look at an example to see how the Structural Similarity Index, or SSIM, actually works in practice.

On the top left, we have the reference image — the original, undistorted version. If we compare this image with itself, the SSIM score is exactly 1. That makes sense, because they are identical.

Now, compare the original with the two images next to it. On the top row, both look quite similar to the original. One produces an SSIM value of 0.949, the other 0.989. These numbers are very close to one, reflecting the fact that the images are almost identical to our eyes.

On the bottom row, however, we see very different results. One image has been heavily pixelated, another blurred, and another corrupted with noise. When compared to the original, their SSIM values drop significantly — around 0.67, 0.69, and about 0.72. That means they retain only about two-thirds of the structural similarity.

The important thing here is that SSIM values line up with what we visually perceive. Images that look good to us score close to one. Images that look distorted or degraded score much lower. That is the power of SSIM — it bridges the gap between mathematical measurement and human vision.

slide15:

Now let’s look at how the Structural Similarity Index, or SSIM, is actually computed.

Suppose we have two images, which we’ll call signal X and signal Y. The first step is to look at their luminance, which simply means the average brightness level. We measure the mean intensity of each image and make sure we are comparing them on the same scale.

Next, we look at the contrast. This is captured by the standard deviation — how much the pixel values vary around the mean. A high standard deviation means strong contrast, while a low standard deviation means the image looks flat or washed out. So we compute the standard deviation for each image to quantify its contrast.

After normalizing for luminance and contrast, we focus on the most important part: the structure. This step captures how patterns of pixels in one image relate to patterns in the other — whether the edges, textures, and fine details line up.

Finally, SSIM combines all three comparisons — luminance, contrast, and structure — into a single value between zero and one. A score of one means the images are structurally identical. A lower score means important structural information has been lost.

So, in summary: SSIM works by checking three things — do the images have the same brightness, the same contrast, and the same structural patterns? When all three align, we say the images are very similar.

slide16:

As I mentioned, the Structural Similarity Index is built on three components: luminance, contrast, and structure — or L, C, and S for short.

Let's start with luminance. Suppose we have an image X with N pixels. To find its average brightness, or mean, we simply add up all the pixel values and divide by N. This gives us the mean value, which we call mu of X.

Next, we remove the mean by subtracting the mean of X from every pixel. This is the first step of normalization — it centers the image so that the average brightness is zero.

Once we have the mean, we can also compute the standard deviation. This measures how much the pixel values vary around the mean. Notice that the formula here uses 1 divided by N minus 1, not just N. That detail comes from statistics — it gives us an unbiased estimate of the standard deviation.

Finally, we can normalize the image even further by dividing each pixel by the standard deviation. After this second step of normalization, the transformed image will have a mean of zero and a standard deviation of one. In other words, it has been rescaled so that brightness and contrast are standardized.

We repeat this process for both images, X and Y. Only then do we compare them — first in terms of luminance, then contrast, and finally their structural relationship.

So, these are the basic operations behind SSIM. They are simple statistical tools, but together they allow us to capture how similar two images are in terms of brightness, contrast, and structure.

slide17:

When we construct a similarity measure, the goal is to capture not just how different two images are, but how similar they are. To do this properly, we want our measure to satisfy three basic conditions, or postulates.

First, symmetry. The similarity between X and Y should be the same as the similarity between Y and X. In other words, order doesn't matter. If you compare image A with image B, you should get the same result as comparing image B with image A.

Second, boundedness. The similarity value should always be between zero and one. A score of one means the images are perfectly identical. A score closer to zero means they are very different.

Third, a unique maximum. The similarity should only reach the maximum value of one if the two images are exactly the same, pixel for pixel. That way, the measure has a clear and meaningful interpretation.

These three conditions — symmetry, boundedness, and unique maximum — make sure that a similarity measure behaves logically and reliably. SSIM is designed to satisfy all of these postulates.

slide18:

So let's begin with the first component of SSIM: the luminance comparison.

Remember, luminance simply means the average brightness of an image. If two images have the same mean intensity, they should score highly on luminance similarity. If one image is much brighter or darker than the other, the similarity should be lower.

Mathematically, the luminance comparison is written like this:

L of X and Y equals two times mu-X times mu-Y, plus a constant C1, all divided by mu-X squared plus mu-Y squared, plus that same constant C1.

Here, mu-X and mu-Y represent the average brightness of images X and Y. The constant C1 is added to make the formula stable — otherwise, if both means are very close to zero, the denominator becomes tiny and the ratio unstable.

To choose C1, we take the square of K1 times L, where L is the dynamic range of pixel values. For an 8-bit grayscale image, L is 255. K1 is just a very small number, much less than one.

Now, notice a few important properties:

If the two means are equal, then the numerator and denominator are the same, so the luminance comparison equals 1, meaning perfect similarity.

If the two means are very different, the numerator becomes small relative to the denominator, so the luminance comparison approaches 0, meaning poor similarity.

This simple formula ensures that luminance similarity satisfies the three postulates we discussed earlier: it is symmetric, bounded between 0 and 1, and it reaches the maximum value of 1 only when the two images have identical mean brightness.

So luminance is the first step in measuring structural similarity. Next, we'll look at contrast comparison.

slide19:

Let's take a closer look at the luminance term and see why it behaves the way we want.

Here, we let mu-X, the mean of image X, be represented by the letter A. And we let mu-Y, the mean of image Y, be represented by the variable X. Now, the question is: as X changes, what value of X will maximize the luminance comparison?

To answer this, we take the formula for the luminance term and compute its first derivative with respect to X. By setting that derivative equal to zero, we can find the point where the function reaches its maximum.

When we go through the algebra, the conclusion is very clear: the maximum occurs when X equals A, in other words, when the mean brightness of image Y is equal to the mean brightness of image X. At that point, the luminance comparison reaches its maximum possible value, which is one.

This confirms the intuition: two images will only be perfectly similar in luminance if their average brightness levels are the same. If one image is brighter or darker than the other, the luminance comparison will drop accordingly.

So this mathematical exercise is really just a proof of what we already expect: the luminance term is maximized when the two images have the same mean brightness.

That's the first aspect of structural similarity. Next, we'll move on to the second aspect: contrast comparison.

slide20:

The second component of SSIM is the contrast comparison.

Even if two images have the same average brightness, they may still look very different if their contrast is not the same. One might look sharp and vivid, while the other looks flat or washed out. To capture this, we use the standard deviation of pixel values — sigma-X for image X and sigma-Y for image Y.

The formula for contrast comparison looks very similar to the one we used for luminance. It is written as:

C of X and Y equals two times sigma-X times sigma-Y, plus a constant C2, divided by sigma-X squared plus sigma-Y squared, plus that same constant C2.

Here again, the constant is added to prevent instability when values are close to zero. C2 is defined as K2 times L squared, where L is the dynamic range of the image.

Now, notice how this works:

If the two images have the same standard deviation, meaning their contrasts are the same, then the numerator and denominator match, and the contrast comparison equals 1.

If one image has much higher or lower contrast than the other, then the ratio becomes smaller, and the similarity drops.

This design is also consistent with how the human visual system works. For example, in a very dark room, small changes in brightness can be quite noticeable. But in a very bright, high-contrast scene, the same small changes are much harder to see. The formula reflects this effect, sometimes called contrast masking.

So contrast is the second pillar of structural similarity — it ensures that two images not only have the same average brightness but also the same level of variation around that average.

Next, we'll look at the third pillar: the structural comparison.

slide21:

Now let's analyze the contrast term a bit more carefully.

The key idea here is that the human visual system does not respond to absolute changes in brightness, but rather to relative changes. This is exactly what Weber's law describes. It states that the smallest noticeable change in a stimulus is proportional to the background level of that stimulus.

For example, if you are in a very dark room, even a tiny spot of light is noticeable right away. But if you are in a brightly lit room, the very same tiny light would be invisible. The background intensity is so high that the small change is masked.

You've probably experienced this outdoors: at night, under a dark sky, you can see countless stars. But during the day, those same stars are still there — yet you can't see them, because the bright background light overwhelms them.

The same principle applies to hearing. In a quiet room, even a whisper can be heard. But in a noisy party, you need to shout for anyone to notice.

Mathematically, this means that the sensitivity of the contrast measure depends on the ratio of the change, delta-X, to the baseline, X. If X is small, even a small delta makes a big difference. If X is large, the same delta hardly matters.

So the contrast term in SSIM naturally captures this idea — it reflects the fact that our visual system is tuned to relative, not absolute, differences. This is why the measure agrees so well with human perception.

slide22:

Here's a simple example of a changeover background.

Look at the first column on the left. The top image has about 10 dots, while the bottom has about 20 dots. The difference is 10. When you compare them, it's very obvious that the bottom one has more dots than the top one.

Now look at the second column. The top image has about 110 dots, and the bottom has about 120 dots. Again, the difference is 10. But this time, it's much harder to notice the difference.

Why? Because the background level — the total number of dots — is so much higher. A change of 10 is a large fraction when the background is only 10, but it's a tiny fraction when the background is already 110.

This illustrates the principle we just discussed: what our visual system detects is not the absolute change, but the relative change — delta divided by the background.

So, when we design a similarity measure, we want it to reflect this property of human vision. And that's exactly what the contrast comparison in SSIM does. It behaves in a way that is consistent with how we see.

slide23:

The third and final component of SSIM is the structural comparison.

Here's the idea. Once we normalize each image — removing its mean brightness and adjusting its contrast — what's left is essentially its structure. This is where we ask: do the patterns, textures, and shapes in the two images align?

Mathematically, this is captured using the cross-correlation between the two images. We call it sigma-X-Y, which measures how the variations in image X line up with the variations in image Y. To make the measure stable, we again introduce a constant, just like we did with luminance and contrast.

Another way to think about this is in terms of vectors in high-dimensional space. Imagine each image as a vector, with each pixel value being one coordinate. The structural similarity is then like taking the inner product of these two vectors, normalized by their lengths. Geometrically, this is just the cosine of the angle between the two vectors.

If the two vectors point in exactly the same direction — meaning the images are identical up to scaling — the angle is zero, and the similarity is maximized.

If they are completely uncorrelated, the inner product is small, and the similarity is low.

This is the same intuition we used when studying Fourier analysis: a Fourier coefficient is computed by projecting a signal onto a basis function. Here, we are projecting one image onto another and measuring how well they align.

So, structural comparison boils down to asking: once brightness and contrast are accounted for, do the fine details and patterns of the two images still match?

That completes the three pillars of SSIM: luminance, contrast, and structure. Next, we'll see how they combine into the full similarity measure.

slide24:

Now let's connect this to a very important mathematical principle: the Cauchy–Schwarz inequality.

You may remember this from linear algebra. It tells us that the inner product of two vectors is always less than or equal to the product of their lengths. Equality holds only when the two vectors are linearly related — in other words, when one is just a scaled version of the other.

Why is this important here? Well, recall that in structural comparison, we represent each image as a vector in a high-dimensional space. The structural similarity is essentially the normalized inner product between those two vectors.

The Cauchy–Schwarz inequality guarantees that this similarity value will always lie between zero and one, and it will reach one only when the two images are structurally identical, differing at most by a scaling factor.

So, this inequality is the mathematical reason why the structural comparison term works. It's not just a rule chosen arbitrarily — it's grounded in solid geometry.

And as engineers and scientists, it's always valuable to know not just how something works, but why. Understanding the principle behind the rule allows you to be creative and adapt these ideas in new situations.

slide25:

And now, we can finally put everything together.

We've defined three components: luminance comparison, contrast comparison, and structural comparison. Each one captures a different aspect of how two images may look alike or differ. The Structural Similarity Index, or SSIM, is built by combining all three.

Mathematically, the SSIM of X and Y is written as:

luminance term raised to the power alpha, multiplied by the contrast term raised to the power beta, multiplied by the structure term raised to the power gamma.

The parameters alpha, beta, and gamma simply allow us to adjust the relative importance of the three components. In practice, we usually set all three equal to one, giving equal weight to luminance, contrast, and structure.

To keep the formula stable, we also include constants — C1, C2, and C3. These prevent divisions by very small numbers, which could make the result unstable. For convenience, researchers often set C3 equal to half of C2. These are empirical design choices, not strict theory, but they work well in practice.

With these simplifications, we arrive at the familiar form of SSIM that is widely used today. This measure is often referred to as the Universal Quality Index with stabilizing constants added.

The remarkable thing is how well SSIM works. Despite its simplicity, it has become one of the most widely used metrics for image quality, both in research and in industry. Even today, when we use deep learning and advanced algorithms for image processing, SSIM and mean squared error remain the standard reference metrics.

So, this is the moment where SSIM is born — a simple but powerful measure that aligns closely with human perception and has stood the test of time.

slide26:

Here's another example that demonstrates the power of SSIM.

The image in the top left, marked (a), is the reference image — the ground truth. The other five images are distorted versions, each with a different type of degradation: contrast stretching, mean shift, JPEG compression, blurring, and salt-and-pepper noise.

If we were to use mean squared error, all of these distorted images would end up with roughly the same error value. But visually, they do not look equally bad. Some are close to the original, while others are clearly worse.

Now look at the SSIM scores shown here. The contrast-stretched image has an SSIM close to 0.92, which makes sense because it looks fairly good. The mean-shifted version scores about 0.90, also quite close. But

the JPEG-compressed version drops down to around 0.69, and the blurred version to 0.70. The salt-and-pepper noisy image scores about 0.77.

If you compare these numbers with what your eyes tell you, the agreement is clear. Images that look closer to the original score higher; images that look worse score lower.

This is why SSIM is so widely used. It captures the essential aspects of the human visual system — brightness, contrast, and structure — in a single number. And in practice, this means we can use SSIM to guide optimization. For example, when designing a communication channel or an imaging system, we can adjust the system parameters to maximize SSIM, ensuring that the images produced look good to the human eye.

So, SSIM is not just a theoretical measure — it is a practical tool that connects mathematics directly to human perception.

slide27:

Because SSIM works so well, researchers have developed many extensions of the method.

The version we've been discussing applies to grayscale images, where each pixel has only one intensity value. But in reality, most images are in color. So naturally, the first extension was to adapt SSIM for color image quality assessment. This involves measuring similarity not only in brightness and contrast, but also in the relationships between color channels.

Another extension is to video quality assessment. Here, SSIM is applied not just frame by frame, but also across time, because our eyes are sensitive to temporal consistency. This has become very important in areas like video compression and streaming.

There is also multi-scale SSIM, which evaluates images at different levels of resolution. This is especially useful because human vision itself operates on a multi-scale level — we notice both fine details and large structures, depending on how we view an image.

And finally, there is complex wavelet SSIM, which can handle images that have complex values, such as those produced in MRI. This is a powerful extension that broadens SSIM to applications in medical imaging and beyond.

The core idea of SSIM — comparing luminance, contrast, and structure — has proven to be so flexible that it has been adapted to various domains.

Now, let me pause here and use an analogy that connects to your own experience. When we design an exam, we want the scores to spread out enough to show meaningful differences among students. If everyone scores 90 or higher, we can't distinguish performance very well. Ideally, the mean should be somewhere in the middle, say around 50, with variation above and below. That way, the test reveals the true distribution of understanding.

It's the same idea with SSIM: we normalize by the mean and by the variation, so we can focus on the meaningful structural differences between signals.

So, whether we're evaluating images or student performance, the principle is the same: measure differences relative to expectations and variability, not in absolute terms.

slide28:

Now we move into the second part of our lecture: system-specific measures.

So far, we've discussed general mathematical ways of comparing images — measures like mean squared error, KL divergence, and SSIM. These are important, but they don't tell the whole story, because image quality also depends heavily on the imaging system itself.

Every imaging system — whether it's an X-ray detector, an MRI scanner, or an ultrasound probe — has its own characteristics and limitations. To assess the quality of images produced by a system, we need to look at metrics that reflect the system's performance.

This includes measures such as:

Noise, and the related quantities signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR).

Resolution, which can be spatial, contrast, temporal, or spectral, depending on what aspect of the system we are evaluating.

And finally, artifacts, which are false or misleading structures that appear in the image because of imperfections in the imaging process.

Together, these system-specific measures tell us how good a given imaging device is at capturing and representing reality.

So let's begin this section with one of the most important system-specific measures: the signal-to-noise ratio, or SNR.

slide29:

In engineering, one of the most familiar terms you'll encounter is the Signal-to-Noise Ratio, or SNR.

Whenever we take a measurement — whether it's a photograph, an MRI scan, or an ultrasound image — we always have some amount of noise. Noise is unavoidable. It arises from the fundamental randomness of physical processes, and ultimately from quantum mechanics itself. Nothing is ever perfectly certain.

On top of this background noise, we have the signal — the part of the measurement that actually carries meaningful information.

SNR is a simple but powerful way of expressing how strong the signal is relative to the noise. Mathematically, it is defined as the ratio of signal power to noise power. Since power is proportional to the square of amplitude, you can also write SNR as the square of the signal amplitude divided by the noise amplitude.

Here's the intuition:

If the SNR is around 1, the signal varies in the same range as the noise. That means the signal is barely visible.

If the SNR is 5 or 10 or higher, the signal stands out clearly above the noise, and it can be easily detected.

A related measure is the Contrast-to-Noise Ratio, or CNR. Instead of comparing one signal to background noise, we compare the difference between two signals — for example, a feature versus its background — divided by the noise level. This measure is very common in medical imaging because it directly reflects how well we can distinguish one structure from another.

So, SNR tells us whether a signal can rise above the noise at all, while CNR tells us whether two signals can be distinguished against that noisy background.

And next, we'll move on to another fundamental aspect of image quality: resolution.

slide30:

Next, let's talk about spatial resolution, which is one of the most important measures of image quality.

The simplest way to think about spatial resolution is in terms of how well an imaging system can distinguish two objects that are close together. For example, imagine two very small, bright points — like two tiny tumors. Mathematically, we describe each of those points as a delta function, which represents a perfect single dot.

But in reality, no imaging system can reproduce a perfect dot. Instead, each point becomes blurred into a small disk, usually shaped like a Gaussian curve. This blur is called the point spread function, or PSF. It tells us how the system responds to a single point source.

Now, suppose we have two points. If they are far apart, even though each one is blurry, you can still clearly see two separate spots. But as they move closer together, the two blurred shapes begin to overlap. Eventually, when the separation between them is too small, the two spots merge and appear as one.

The critical threshold is defined by the full width at half maximum, or FWHM, of the point spread function. In other words, when the distance between the two points is equal to the width of the blur at half its height, that's about the limit of what the imaging system can resolve. Any closer, and the system can no longer distinguish the two points.

This is why we say spatial resolution is the minimum separation at which two objects can still be seen as distinct. Often, when people talk about "resolution" in imaging, this is what they mean: the ability to resolve fine detail.

Another way to analyze resolution — which we'll get to shortly — is through the modulation transfer function, or MTF, which uses Fourier analysis to quantify how well different spatial frequencies are preserved in the image.

slide31:

Another way to describe spatial resolution is through the Modulation Transfer Function, or MTF.

Here's the idea: every image can be broken down into sinusoidal components using Fourier analysis. These components can be low frequency — representing smooth, gradual changes — or high frequency — representing fine details like sharp edges or thin lines.

If we feed a sinusoidal pattern into an imaging system, the system does not pass all frequencies equally well. Low-frequency components pass through almost perfectly. But as the frequency increases, the system begins to attenuate, or weaken, those components. At very high frequencies, the system may blur them so much that they become invisible.

The MTF curve shows this behavior. On the left side, at low frequencies, the modulation is close to 100 percent. As frequency increases, the response gradually drops. Eventually, at very high frequencies, the response falls to zero — meaning the system can no longer reproduce those details.

A practical way to think about this is with a line-pair phantom, where black and white bars alternate like a test pattern. At low frequencies, the bars are wide, and the system can reproduce them clearly. At high frequencies, the bars get narrower and closer together. At some point, the system can no longer distinguish the bars — they blur into a uniform gray. That point defines the system's resolution limit.

So MTF gives us a frequency-domain way of measuring resolution. It tells us not just whether two points can be separated, but how well the system preserves detail across different scales.

This measure is widely used in both medical imaging and general optics. It provides a more complete picture of resolution compared to the single-number definition from the point spread function.

slide32:

Next, let's look at contrast resolution, sometimes called low-contrast resolution.

This is different from the high-contrast resolution we just discussed. High-contrast resolution deals with distinguishing small, bright details — for example, two tiny dots very close together.

Contrast resolution, on the other hand, is about the ability of an imaging system to detect subtle differences in intensity. The structures may not be small — they may even be large — but if their contrast relative to the background is very low, they may be difficult to see.

In the example shown here, notice how in one image the faint circular structures stand out more clearly, while in the other they are barely visible against the noisy background. This difference reflects the system's contrast resolution.

Clinically, this is very important. For example, in CT imaging, a tumor may look very similar in intensity to surrounding soft tissue, because both are made of similar biological material. A system with good contrast resolution allows the tumor to be detected clearly, while a noisy or lower-quality system might completely obscure it.

Beyond contrast resolution, there are two more types of resolution we should briefly mention.

Temporal resolution refers to how quickly an imaging system can capture a snapshot. A high-speed system can freeze motion — like capturing a sharp image of a beating heart. But if the acquisition is too slow, moving objects blur together. This is why, when you take a photo of a moving car with your phone, you sometimes see motion blur. In medical imaging, temporal resolution is critical for dynamic studies, such as cardiac CT or MRI.

Spectral resolution refers to the ability of a system to distinguish between different frequencies or energies. In vision, this corresponds to color perception — someone with color blindness has poor spectral resolution. In imaging systems, spectral resolution allows us to distinguish X-rays of different energies or ultrasound

waves of different frequencies. Better spectral resolution means finer discrimination of subtle differences in material properties.

Finally, let's talk about artifacts. Artifacts are false structures that appear in images but are not really there. They are "ghosts" created by limitations or mismatches in the imaging process. For example, in CT, the motion of the heart can create blurry or duplicated structures. In ultrasound, echoes may bounce back and forth and appear as multiple spots even though only one structure exists. Recognizing and understanding artifacts is critical, because they can mislead interpretation.

slide33:

Now let's look at a specific example of artifacts — in this case, metal artifacts in CT imaging.

Suppose a patient has a hip fracture and receives metal implants. When we take X-rays or CT scans, the metal is so dense that it blocks or severely distorts the X-ray beams. The reconstruction algorithm, however, assumes it has complete and accurate data along every path. It doesn't "know" that some of the information is missing or corrupted.

As a result, the system produces streaks and bands radiating from the metal. These bright and dark lines are not real anatomical structures — they are purely computational artifacts caused by missing or distorted data.

In the figure here, the first column shows uncorrected CT images with severe streak artifacts. The last column shows the ground truth, what the anatomy should look like.

Researchers have developed various ways to reduce these artifacts. One common method is NMAR, or normalized metal artifact reduction, which improves the image somewhat. More recently, deep learning methods, such as convolutional neural networks, have been applied to further clean up the image and recover structures hidden by artifacts.

You can see that the CNN results are closer to the ground truth, with clearer anatomy and fewer streaks.

The key point here is that artifacts are not real structures, but they can easily confuse interpretation. They depend on the imaging modality, the reconstruction algorithm, and the clinical situation. That's why recognizing artifacts — and knowing how to reduce them — is so important in medical imaging.

slide34:

So far, we've covered general measures like MSE, KL distance, and SSIM, and then system-specific measures such as noise, resolution, and artifacts.

Now we move to the third and most important part: task-specific measures.

When manufacturers design an imaging system, they often advertise its specifications — resolution, noise levels, dose reduction, acquisition speed, and so on. These system specs are useful, but in practice, they don't directly answer the most important question: Can the system help us make the right clinical decision?

Think about it. As patients or physicians, we don't actually care whether an image has 0.5-millimeter resolution or 1-millimeter resolution, or whether the noise level is 2 percent or 5 percent. What we really

care about is whether the image allows us to detect a tumor, diagnose a fracture, or rule out disease with confidence.

That's why task-based assessment is so important. It focuses on the end goal: whether the imaging system can successfully support the clinical task.

So in this last part of the lecture, we'll discuss task-specific measures such as sensitivity and specificity, ROC and AUC analysis, human and mathematical observers, and even modern approaches like neural networks and radiomics.

These measures shift our attention from the imaging device itself to the diagnostic performance of the entire workflow, which is ultimately what matters most in medical imaging.

slide35:

Here we see a typical scene from a radiology department. A radiologist, with years of training and experience, carefully reviews CT scans and X-rays. Their task is not to admire image sharpness or pixel resolution — it is to answer very specific clinical questions:

Does this smoker have a lung tumor?

Are the coronary arteries narrowed?

Is there evidence of disease that requires treatment?

This highlights why task-specific measures are so important.

System-specific measures — like spatial resolution, SNR, or temporal resolution — are useful, but they are not the end goal. They only tell us about the technical performance of the imaging system. What really matters is whether the system can support the clinical task: detecting, diagnosing, and guiding treatment decisions.

So in this part of the lecture, we'll look at task-specific measures such as sensitivity, specificity, ROC curves, and observer studies, which are all designed to assess performance in clinical terms rather than engineering terms.

This shift in perspective is critical: we move from asking "How sharp is this image?" to asking "Can this image help us detect disease reliably?"

slide36:

Let's make this idea of task-specific measures more concrete with a simple example.

Imagine you are doing edge detection on an image — for instance, outlining the Eiffel Tower. Each pixel can belong to one of two classes: edge or non-edge. Similarly, in medical imaging, each case can be classified as disease present or disease absent.

This gives us four possible outcomes:

True Positive: The patient has a tumor, and the system or doctor correctly identifies it.

True Negative: The patient is healthy, and the report correctly says no tumor.

False Positive: The patient is healthy, but the system says there is a tumor. This can cause unnecessary anxiety, follow-up scans, or even invasive procedures.

False Negative: The patient does have a tumor, but the system or doctor misses it. This is often the most serious error, because it can delay diagnosis and treatment.

In clinical practice, both types of errors matter. False positives burden patients emotionally and financially, while false negatives can cost lives by missing opportunities for early treatment.

That's why doctors and imaging researchers often talk about sensitivity and specificity. Sensitivity reflects how good a system is at detecting true positives — catching disease when it is really there. Specificity reflects how good a system is at avoiding false alarms — correctly identifying when disease is not present.

These terms can be confusing at first, but they are central to evaluating diagnostic imaging systems. They shift the focus away from just image sharpness or resolution and toward what really matters: whether the imaging system can support accurate medical decisions.

slide37:

Now that we've defined the four possible outcomes — true positives, true negatives, false positives, and false negatives — we can formally define sensitivity and specificity.

Sensitivity is the proportion of true positives among all actual positive cases. In formula form, it is:

Sensitivity equals TP over TP plus FN.

In other words, sensitivity answers the question: "If the patient really has the disease, how likely are we to detect it?"

For example, suppose 100 patients truly have lung cancer. If our system correctly detects 50 of them, but misses the other 50, then the sensitivity is 50%. Sensitivity is essentially a measure of how confident we are when we say "Yes, disease is present."

Now let's move to specificity.

Specificity is the proportion of true negatives among all actual negative cases. The formula is:

Specificity equals TN over TN plus FP.

This answers the question: "If the patient really does not have the disease, how likely are we to correctly say no?"

For example, if 100 healthy people are scanned, and 95 are correctly reported as negative while 5 are mistakenly labeled as positive, the specificity is 95%. Specificity is essentially a measure of how confident we are when we say "No, there is no disease."

So, in summary:

Sensitivity tells us how good we are at catching disease cases.

Specificity tells us how good we are at avoiding false alarms.

Both are critical in clinical imaging. High sensitivity ensures we don't miss disease, while high specificity ensures we don't overwhelm patients with unnecessary worry or treatment.

slide38:

Now, to make things even more interesting, we introduce two more important concepts: positive predictive value (PPV) and negative predictive value (NPV).

Positive predictive value, or PPV, tells us: "If the imaging study says a patient is abnormal, how likely is it that they really have the disease?"

The formula is:

PPV equals True Positives over (True Positives plus False Positives).

So, among all the patients labeled as abnormal, PPV is the fraction who actually do have the disease.

On the other hand, negative predictive value, or NPV, tells us: "If the imaging study says a patient is normal, how likely is it that they really are disease-free?"

The formula is:

NPV equals True Negatives over (True Negatives plus False Negatives).

So, among all the patients labeled as normal, NPV is the fraction who truly have no disease.

Let's put this in clinical terms. Suppose we are using CT to detect a tumor. If the scan flags a suspicious spot, PPV tells us how likely it is that this really is a tumor and not just a false alarm. If the scan comes back clear, NPV tells us how confident we can be that the patient truly has no tumor.

So, sensitivity and specificity tell us how the test performs in an abstract sense, relative to ground truth. PPV and NPV tell us what the test result actually means for the patient sitting in front of us.

slide39:

Here's a concrete example using tuberculosis screening with chest X-ray. In this study, nearly 2,000 people were tested, but only 30 actually had tuberculosis.

Let's break this down.

Among the 30 true TB patients, 22 were correctly reported as positive, while 8 were missed.

Among the 1,790 healthy individuals, 51 were incorrectly reported as positive, but the majority — 1,739 — were correctly classified as negative.

Now, from this table, we can calculate the standard measures:

Sensitivity: 22 out of 30 true TB cases were detected → about 73%. This means the system caught almost three-quarters of real cases.

Specificity: 1,739 out of 1,790 healthy cases were correctly identified → about 97%. This shows the system rarely gave false alarms.

Positive Predictive Value (PPV): 22 out of 73 reported positives were real TB cases → only 30%. This is much lower, meaning that when the system flagged a case as positive, it was wrong more often than it was right.

Negative Predictive Value (NPV): 1,739 out of 1,747 reported negatives were truly healthy → nearly 99.5%. This means a negative result was highly reliable.

Diagnostic Accuracy: The proportion of all correct results, positive and negative, was about 97%.

Prevalence: Only 30 out of 1,820 people had TB → around 1.6%.

Notice something important here: although accuracy and NPV are very high, the PPV is quite low. That's because the disease is rare in this population — when prevalence is low, even a small number of false positives can outweigh the true positives.

This example shows why we need to interpret these metrics carefully, especially in screening programs. Sensitivity, specificity, PPV, NPV, and prevalence all interact, and each tells us something different about how the test performs in real-world conditions.

slide40:

Now let me introduce one of the most powerful tools in diagnostic performance evaluation: the receiver operating characteristic curve, or ROC curve.

Here's the key idea: sensitivity and specificity are not fixed properties of a test. They depend on the threshold we use to decide between "disease" and "no disease."

For example, if I set the threshold very low, I'll call almost everything abnormal. That means I'll catch nearly every true case, so sensitivity will be very high. But I'll also generate many false alarms, so specificity will drop.

On the other hand, if I set the threshold very high, I'll call almost everything normal. That means I'll have very few false alarms — so specificity will be excellent. But I'll also miss many real cases, so sensitivity will be low.

The ROC curve captures this trade-off. On the x-axis, we plot 1 – specificity, which is the false positive rate. On the y-axis, we plot sensitivity, the true positive rate. By sweeping through all possible thresholds, we trace out the curve.

The ROC curve shows the overall performance of the test across all possible decision boundaries. And importantly, the area under the ROC curve, or AUC, provides a single number to summarize diagnostic accuracy. An AUC of 1.0 means perfect classification. An AUC of 0.5 — a diagonal line — means the test is no better than random guessing.

So the ROC curve is widely used in medical imaging research, because it provides a clear, quantitative way to compare diagnostic systems and observer performance.

slide41:

Here's another way to look at the ROC curve — in terms of true positive fraction and false positive fraction.

The true positive fraction is simply the sensitivity: among all real disease cases, what fraction we detect correctly.The false positive fraction is equal to one minus specificity: among all healthy cases, what fraction we mistakenly call positive.

When we plot sensitivity on the vertical axis against 1 – specificity on the horizontal axis, we get the ROC curve.

If the curve lies along the diagonal dashed line, it means the test is performing no better than random guessing — the true positive fraction increases at the same rate as the false positive fraction.

But if the curve bends upward toward the top-left corner, that indicates good diagnostic performance: we are capturing many true positives while keeping false positives relatively low.

So in practice, the more the ROC curve bulges toward the upper left, the better the test or imaging system is at separating disease from non-disease.

This is exactly why the ROC framework is so valuable. It captures the trade-off between sensitivity and specificity in a single curve, and it gives us a clear visual and quantitative way to compare different diagnostic systems.

slide42:

Let’s imagine the ideal case of diagnosis.

Here we have two groups: on the left, the non-diseased population, shown in red; on the right, the diseased population, shown in green.

Suppose we measure a certain feature — for example, the diameter of a vessel or the size of a tumor. In an ideal world, the distributions of healthy and diseased cases do not overlap at all. Healthy cases always fall into the red range, and diseased cases always fall into the green range.

With this perfect separation, we can place a threshold right between the two distributions. Anything to the right is declared diseased, and anything to the left is declared healthy.

In this situation, the diagnostic test is flawless. There are no false positives and no false negatives. Sensitivity and specificity are both 100%. The ROC curve would go straight up to the top-left corner and across the top — a perfect AUC of 1.0.

Of course, this is rarely the case in real clinical imaging. Features usually overlap between healthy and diseased groups, which makes things more complicated. But this ideal case provides a useful reference point for understanding what we’re aiming for.

slide43:

In reality, things are rarely so clear-cut.

Most of the time, the distributions of healthy and diseased patients overlap. That means the same measured value could belong either to a healthy individual or to someone with a disease.

For example, suppose we run a blood test or measure a tumor-related biomarker. Healthy patients tend to cluster around one range, while diseased patients cluster around another. But because of biological variation, measurement noise, and overlapping physiology, there is no perfect separation.

So when the two distributions overlap, a threshold placed in the middle will inevitably create two types of errors. Some diseased patients will fall below the threshold and be misclassified as healthy — these are false negatives. Some healthy patients will fall above the threshold and be misclassified as diseased — these are false positives.

This overlap is exactly what forces us to think carefully about where to place the decision threshold. And it's what makes tools like the ROC curve so valuable, because they let us analyze the trade-off between sensitivity and specificity across all possible thresholds.

slide44:

Now, let's take our overlapping case and place a conservative threshold, shown by the blue line here.

With this choice, notice what happens. For diseased patients, only those who fall far to the right of the threshold are called positive. That means many truly diseased patients on the left side are missed. The sensitivity — or true positive rate — drops to around 50%.

On the other hand, for healthy patients, almost all are correctly classified. Only a very small slice on the right is misclassified as diseased. That means the false positive fraction is low, and specificity is high.

So in ROC space, where we plot sensitivity against 1 minus specificity, this operating point appears toward the lower-left corner. It reflects a cautious decision-making style: you're very reluctant to call a disease, so you minimize false alarms. But the trade-off is that you miss a lot of true cases.

This illustrates the key idea: where you set the threshold directly determines the balance between sensitivity and specificity.

slide45:

Now let's look at a moderate threshold setting.

This time, we shift the decision boundary slightly to the left. What happens? More of the diseased distribution now falls to the right of the threshold. That means a greater fraction of patients with the disease are correctly identified. In other words, the sensitivity increases.

But the trade-off is clear. By moving the line left, we also capture more of the non-diseased distribution in the red region. That means more healthy patients are incorrectly flagged as diseased. The false positive fraction increases.

On the ROC curve, you can see this move as a step upward — higher sensitivity — but also a step to the right — higher false positive fraction. In other words, we've shifted from the black cross to the yellow cross on this plot.

This "moderate" threshold represents a more balanced decision-making strategy: better at catching true cases, but at the cost of more false alarms.

slide46:

Now let’s look at an even more aggressive threshold.

By moving the cutoff further to the left, we dramatically reduce the chance of missing diseased patients. Almost everyone with the disease is now flagged as positive. That means our sensitivity is very high — close to 100%.

But the price is steep. With the threshold this low, about half of the non-diseased population also falls into the positive range. In other words, the false positive fraction climbs to nearly 50%. Many healthy patients would be told they might have a problem, leading to unnecessary worry and follow-up tests.

On the ROC curve, this decision strategy places us near the top-right portion — very sensitive, but much less specific.

So, if you imagine sliding this threshold back and forth, you’re really tracing out the ROC curve. Different thresholds give you different balances between catching true positives and avoiding false alarms. And as I’ll show you next, the quality of the diagnostic test itself is reflected in how far this ROC curve bends toward the upper-left corner.

slide47:

Now you see how the entire ROC curve is formed.

At one extreme, if I simply call every patient “positive,” I would achieve 100% sensitivity — I would never miss a diseased case. But at the same time, I would also reach 100% false positive fraction, meaning every healthy patient is incorrectly flagged.

At the other extreme, if I call everyone “negative,” I would achieve 100% specificity, but sensitivity would drop to zero.

By moving the decision threshold step by step, we trace the ROC curve from the lower-left corner, where both sensitivity and false positives are low, toward the upper-right corner.

An ideal diagnostic system bends sharply upward and hugs the left and top axes, giving an area under the curve, or AUC, close to 1. A poor test, on the other hand, falls close to the diagonal line, which is equivalent to flipping a coin — nothing more than random guessing.

So, the ROC curve captures the complete balance between sensitivity and specificity across all thresholds. And the area under this curve provides a single, powerful number that summarizes diagnostic performance. That’s why ROC analysis has become such a central tool in medical imaging, in machine learning, and in clinical decision-making.

slide48:

Now, let’s step back and think about what diagnostic performance really means.

Not every doctor reads images with the same skill. Some physicians, with years of training and experience, can clearly separate diseased from non-diseased cases. For them, the two distributions hardly overlap, and their ROC curve bends sharply toward the upper-left corner — that's excellent diagnostic power.

Others, perhaps junior residents still in training, may not yet recognize subtle features. Their separation is weaker, the overlap is greater, and the ROC curve lies closer to the diagonal. In the extreme case, if you cannot tell disease from normal at all, your decision is no better than flipping a coin — that's the chance line.

So diagnostic performance is shaped by two things: the reader's skill and the technology's power. Better imaging systems reduce overlap, but equally important is the ability of the radiologist to interpret the image correctly.

That's why, in practice, the same scan might lead to different conclusions depending on who reads it. A highly trained radiologist can extract subtle discriminating features that others may miss. This is exactly what the ROC curve captures — the combined effect of technology and human expertise.

slide49:

Now let's quantify what we've been discussing. The receiver operating characteristic, or ROC curve, gives us a full picture of diagnostic performance across different decision thresholds. But how do we summarize the curve with a single number?

That's where the area under the ROC curve, or AUC, comes in.

If the curve falls along the diagonal, the area is 0.5. That means the test has no predictive value — you're essentially flipping a coin. A perfect test, one that never misses disease and never gives false alarms, would trace along the top and left borders, with an AUC equal to 1. In practice, no system achieves that ideal, because there's always some chance of error.

So most real diagnostic tests fall somewhere in between. The higher the area under the curve, the better the test is at distinguishing diseased from non-diseased cases.

This is why AUC has become such a standard benchmark. It condenses all those trade-offs between sensitivity and specificity into a single number. And just like in teaching or training, performance varies — some students, or some doctors, do exceptionally well; others struggle. The ROC and its AUC make that difference visible in a very clear, quantitative way.

slide50:

Here is a real-world example that makes these abstract concepts much more concrete.

This plot shows the diagnostic performance of 108 radiologists in the United States, taken from a study by Beam and colleagues. On the vertical axis, you see the true positive fraction — in other words, sensitivity. On the horizontal axis, you see the false positive fraction — that is, one minus specificity.

Each point here represents a single radiologist's performance. And what stands out immediately is how widely they are spread. Some radiologists operate near the upper-left corner, meaning they consistently

detect disease with very few false alarms. Others are closer to the diagonal, where their decisions are only slightly better than random guessing.

So even though all of these doctors are trained professionals looking at the same kinds of images, their diagnostic accuracy varies enormously. That variation has practical consequences. Finding a skilled radiologist can make the difference between catching an early tumor and missing it altogether.

It's very similar to what we see in education and research. A strong student working with the right mentor can achieve exceptional results, while another, equally intelligent student might struggle under weaker guidance. In both cases, performance isn't just about the system — it's also about the human observer.

slide51:

Here's another striking example, this time from a classic chest film study by Dr. E. James Potchen in 1999.

What you're looking at are ROC curves comparing three groups: the top 20 radiologists, the bottom 20 radiologists, and a group of 71 radiology residents.

Notice the clear separation. The top 20 radiologists perform extremely well — their ROC curve stays high, close to the upper-left corner. That means they consistently detect abnormalities with very few false alarms.

In contrast, the bottom 20 radiologists struggle. Their ROC curve lies much closer to the diagonal, which means their decisions are only a little better than chance.

And then there are the residents. They fall somewhere in between — they're still training, still developing the skills to interpret subtle features in chest films.

What this tells us is that diagnostic performance is not determined by the technology alone. The human factor — training, experience, and even natural ability — plays a huge role. The same X-ray image can be interpreted very differently depending on who is reading it.

This is why task-specific measures are so important. At the end of the day, what really matters is not just whether the imaging system produces a sharp picture, but whether that picture supports accurate clinical decisions.

slide52:

Now that we've seen how radiologists vary in their diagnostic performance, let's ask an important question: can we systematically model this decision-making process?

This brings us to the concept of model observers. A model observer is essentially a mathematical framework or a computational algorithm that simulates the way a human might read an image and make a decision.

The idea is motivated by a practical challenge. Human reader studies are expensive, time-consuming, and often inconsistent — one doctor may interpret an image differently than another. If we want to optimize imaging systems or evaluate new reconstruction methods, relying on large-scale human studies is not always feasible.

So instead, we can build a numerical observer that "looks" at the images and performs the task — whether that's detecting a small tumor, classifying an abnormality, or distinguishing between signal and noise.

In this review paper by Xin He and Subok Park from the FDA, the authors summarize the foundations of model observers in medical imaging research. They describe how these observers can be based on rigorous statistical decision theory, how they can be tailored to mimic human visual performance, and how they're applied in practice — from system optimization to regulatory science.

In the next few slides, I'll introduce you to some of the key types of model observers, starting from the ideal observer, and then moving toward more practical approximations like the Hotelling observer and modern approaches involving machine learning and neural networks.

slide53:

Before we define model observers in detail, let's step back and look at how we can mathematically describe the imaging process itself.

In most medical imaging systems, the process of acquiring an image can be represented by a very simple but powerful equation:

g = Hf plus n.

Here, f represents the object we are trying to image — for example, the radioactivity distribution in nuclear medicine or the internal structure of a patient in CT.

The imaging system is described by H, which acts like a mapping or an operator. You can think of it as a large matrix that tells us how each point in the object contributes to the measured image. For instance, in CT, H would represent all the X-ray projections and how they are combined.

Then comes n, the noise. Every real imaging system is contaminated by noise — whether it's photon noise, electronic noise, or other random fluctuations. That uncertainty is inherent to the physics of measurement.

Finally, g is what we actually record — the image data. So what you see on the screen is not the object itself, but rather the result of the object being transformed by the system response and corrupted by noise.

This linear system model is the foundation for both image reconstruction and image analysis. It's also where model observers come in: they take this same mathematical description and use it to predict how well a system can perform a diagnostic task.

slide54:

Now let's take the imaging model and apply it to a very common clinical task — binary classification.

In medicine, this usually means asking a simple yes-or-no question. For example, does this patient have a tumor, or not?

Mathematically, we describe this with two competing hypotheses.

The first hypothesis says the image only contains the background — in other words, normal anatomy — plus noise.

The second hypothesis says the image contains both the background and an additional signal, such as a tumor, again with noise added.

Here, we think of the background as the normal structures in the body, and the signal as the diagnostic feature we're trying to detect. Depending on the situation, we might know exactly what that signal looks like, or only know its general statistical properties. For example, in phantom studies, we know exactly what pattern is inserted, but in real patients, tumors can vary in shape, size, and contrast.

To simplify the discussion, we often imagine a "clean" background image without noise, and a "clean" signal image without noise. The actual measurement we record is just those two components, combined with noise from the imaging process.

So in essence, the classification task is: given the measured image, do we believe it came from the background-only case, or from the background-plus-signal case?

This very simple framework — background versus background plus signal — forms the basis of many observer models in medical imaging.

slide55:

Now, let's introduce the idea of the ideal observer.

In the context of binary classification, the ideal observer is defined as the one who makes use of all the statistical information available in order to maximize task performance. In other words, it's the absolute gold standard — the best decision-maker you could have, whether human or machine.

How does it work? Imagine you have an image, which we'll call g. For this image, there are two possible explanations:

Hypothesis zero says it comes from a normal case, with no abnormality.

Hypothesis one says it comes from an abnormal case, where a signal, such as a tumor, is present.

The ideal observer looks at both possibilities and asks: given the data I see, how likely is it that it came from the normal case, and how likely is it that it came from the abnormal case?

The decision is then based on whichever likelihood is greater. If the abnormal case is more likely, the observer calls it abnormal; if the normal case is more likely, the observer calls it normal.

This simple but powerful rule — often called the likelihood ratio test — provides an upper bound on performance. No other observer, whether it's another model or even an expert human reader, can systematically do better.

That's why the ideal observer always produces the highest possible ROC curve for the task. It gives us a theoretical benchmark — a way to measure how close or how far real systems and real doctors are from perfection.

slide56:

The ideal observer, as we just discussed, requires complete statistical knowledge of the problem. But in practice, we almost never have that luxury. So, instead, we turn to simplified models.

One of the most important of these is the Hotelling Observer.

The idea is straightforward: rather than using all the complex statistics of the image, we take the image data and apply a linear template to it. In other words, we reduce the entire image into a single number — a test statistic. That number is then used to decide whether the case is more likely normal or abnormal.

The Hotelling Observer is designed to be the best possible linear observer. What does “best” mean here? It means the observer chooses its template in a way that maximizes the signal-to-noise ratio — the separation between the diseased and non-diseased cases. The larger the separation, the easier the decision becomes.

This is why the Hotelling Observer is so widely studied. It strikes a balance: not as powerful as the ideal observer, but still mathematically optimal within the class of linear decision rules.

And if we don’t even want to use all the image data, we can go a step further. We might only use selected features or specific channels of the image, instead of the full dataset. That gives us the Channelized Hotelling Observer — a practical simplification that still works well in many real-world medical imaging problems.

slide57:

Now, let’s talk about what is called the channelized observer. The idea here is to simplify the problem of analyzing very high-dimensional image data.

Instead of working with every single pixel value, which can be overwhelming, we break the image down into a set of channels. Each channel acts like a filter or a template that extracts a particular aspect of the image. Think of them as different ‘views’ or ‘features’ of the same data.

When we apply these channels to the image, we get a series of scalar responses — simple numbers that summarize how the image looks under each channel. We then collect these responses into a much smaller vector, which is far easier to work with than the original full image data.

This is what’s known as a channelized observer. Reducing the dimensionality allows us to perform statistical analysis more efficiently, while still retaining the essential information needed to judge image quality or detect a signal.

So in short, channelized observers strike a balance: they simplify the computational task but still capture the critical diagnostic features

slide58:

To make the idea of channelized observers more concrete, let me give you an example.

Here you see four channels illustrated both in the frequency domain, across the top row, and in the spatial domain, across the bottom row. Each channel acts like a filter. In the frequency domain, these filters look like concentric rings that select specific frequency ranges. When we transform them back into the spatial domain, they appear as blurred or oscillatory patterns.

Why do we do this? Because instead of trying to analyze the entire frequency spectrum all at once, we divide it into pieces. Each channel captures information from one part of the spectrum. When we apply these filters to an image, we can see how strongly the image responds to each band of frequencies.

This is very powerful because medical images contain structures at many different scales — some large and smooth, others fine and detailed. Using channels allows us to separate these scales and study them systematically.

So, these four channels are just one example, but in practice, you could design more channels depending on how much detail you want to capture."

slide59:

Now let's move into the topic of radiomics, which is an exciting area in modern medical imaging.

Radiomics is all about turning images into data. Instead of relying only on what a radiologist can see with the naked eye, we extract hundreds or even thousands of quantitative features from medical images.

For example, once we outline a tumor on an image, we can study different categories of features:

Shape features tell us about the geometry — is the tumor round, irregular, or elongated?

Intensity features describe how bright or dark the pixels are, and how their distribution looks in a histogram.

Texture features capture the patterns inside the tumor, such as whether it looks smooth, coarse, or heterogeneous. These can be computed using statistical methods like gray-level co-occurrence matrices or run-length matrices.

After we collect these features, we don't stop there. The next step is to feed them into a prediction model. This model may perform tasks like selecting the most important features, classifying patients into groups, or predicting outcomes such as how well a patient will respond to treatment.

So radiomics really bridges imaging with data science. It takes us beyond the visual impression and allows us to mine the hidden information in images, often leading to insights that are not visible to the human eye.

slide60:

Up to this point, we have mostly talked about linear observers, such as the Hotelling observer and its channelized version. But in reality, many problems in medical imaging are nonlinear.

Take the XOR problem as a classic example. If you try to separate the inputs using a simple straight line, you quickly find it cannot be done. Linear observers fail in such situations because the data are not linearly separable.

To overcome this, we use nonlinear observers, often built using neural networks. Here, you see a simple network with two input variables, two hidden units, and one output. By combining the information in a nonlinear way, the network can correctly separate the classes.

This idea is powerful for medical imaging tasks. Images are complex — disease and non-disease patterns may overlap, and linear models are not enough to make accurate decisions. Neural networks, however, can learn complicated nonlinear boundaries, allowing them to detect subtle patterns and interactions that traditional methods miss.

So nonlinear observers represent the next step: moving from simple linear discrimination toward flexible, data-driven models that can capture the true complexity of biological signals.

slide61:

Supervised learning builds on this idea by training models directly on labeled data. For example, in the XOR problem, linear classifiers cannot draw a single straight line to separate the classes. But by introducing hidden layers and nonlinear activation functions, a neural network can learn to separate the classes correctly.

The figure shows how the inputs are transformed step by step through hidden units and weights, leading to correct outputs for all training examples. This principle extends far beyond toy problems—it is the basis of modern deep learning, where very large networks can learn hierarchical features from medical images and achieve state-of-the-art performance.

slide62:

Here we see a fuzzier, more realistic version of the XOR problem. Instead of perfectly separable points, the data is noisy and overlaps. This is much closer to what we encounter in medical imaging, where disease and non-disease cases are not cleanly separated but instead distributed with overlaps.

The nonlinear classifier can still learn a curved boundary that separates the classes reasonably well. This illustrates why nonlinear observers and machine learning are crucial: they can adapt to complex, messy data, rather than relying on oversimplified linear rules.

slide63:

Deep radiomics takes things one step further. Instead of manually extracting features, we use convolutional neural networks (CNNs) to learn them directly from the image data.

In this workflow, random patches from images are used to train a CNN to recognize tumor regions. Multiple layers of convolution and pooling extract increasingly abstract features, from simple edges to complex patterns. The responses from the deep layers can then be aggregated across the whole image, forming high-dimensional feature maps.

These learned features are then encoded and passed to classifiers for prediction tasks, such as distinguishing between tumor subtypes or predicting patient outcomes. Deep radiomics represents the cutting edge—combining the statistical rigor of radiomics with the power of deep learning to capture complex image biomarkers.

slide64:

Finally, here is your homework. You will use the MATLAB code available online to compute the Structural Similarity Index, or SSIM, between two images. SSIM is a perceptual metric that compares luminance, contrast, and structure.

Your task is two-fold. First, compute SSIM for a pair of images—either the ones shown here or another pair of your choice. Second, design an example where sensitivity is 90% and specificity is 80%, and show your calculations.

This exercise is meant to reinforce your understanding of both image quality metrics and diagnostic performance measures. The due date is one week later, so please manage your time accordingly.