[Slide 2]

All right, I'm Charles Huyck, Executive Vice President of ImageCat and I'm going to be introducing the topic of building exposure. I'm talking about it in general, in the context of loss estimation and cat modelling, and also focusing on the work that was done as a part of project METEOR. So the purpose here, this section will introduce key concepts to a general audience such as data resolution, difference between aggregate and building-specific Data, and what is key for assessing Vulnerability.

[Slide 3]

As for takeaways, at the end of the presentation, you'll hopefully have a solid grasp of what exposure data is and how it is used in loss estimation process, a basic process of developing exposure Data, the value of Earth Observation data in the context of building exposure data, and how to check your exposure data and ensure that it's fit.

[Slide 4]

All right, so drilling down into each of these, what is exposure data and how is it used in the loss estimation process? I'll be covering- what is exposure data, levels of exposure Data. That's something that we've come up with on the METEOR project- I'll be discussing those in detail, an introduction to the idea of spatial resolution, building vulnerability attributes, talking about replacement costs vintage, as well as talking about some challenges and expectations people often have with them, of building exposure data.

[Slide 5]

So what is building exposure data? Often when we talk about building exposure data we go in and start to talk to people about risk projects right and off the bat they think of this type of Data- which is I think a virtual 3d representation of buildings in a very urbanized setting. This is not what we're talking about.

[Slide 6]

This is not the type of building exposure that we're talking about. That type of building exposure data is used for rendering and could be used for planning Purposes, but essentially the focus here is capturing what buildings look like and although it does have some sort of aerial representation it doesn't have the type of square footage replacement cost structure type information that we would use.

[Slide 7]

So instead what we're talking about really is the art of distributing inventories of people into estimates of buildings throughout a given area. It's a process that's very specific to loss estimation itself and answers the question- given a number of people, how many households are there going to be, how much dwelling area is associated with those households, and then given that what is the replacement cost structure type that can be used to assess damage from multiple hazards, and then how does this change throughout a given area that you're looking at?

[Slide 8]

This is a representation that came out of the METEOR project, hosted by GEM, of what that it looks like for the LDCs in Africa. I guess it's all in Africa here in this presentation. So what this is a gridded representation for each of these grids. You have an estimate that you can click on and reveal the replacement cost in thousands of Dollars, in this case mapped throughout the area, and for each of these cells, how that replacement cost is distributed in terms of these codes. These codes here map directly to a vulnerability assignment that you can cross-reference With, say ground motion or depth of flooding, to get an estimate of damage. In this case, these are specific for earthquake and that's really what we're talking about when we're talking about building exposure data in this context. It is different than an inventory. An inventory of buildings might be, if you have thousands of buildings in your town that you know are in a particularly vulnerable area and you want to get details of each of those buildings for the purpose of looking at mitigation options or retrofitting or what have you, could say that that's a building exposure at the at the end of the day. That would be as part of what we would call a level five building exposure, but it's really different, tends to be different purposes that would be for sort of a portfolio analysis or municipal management purposes. What we're talking about in terms of building exposure is typically an inference process over a wider area, so when people talk about building exposure they're generally not talking about a building inventory, although it they could be interchangeable in that sense. But a building exposure database that we're talking about is definitely not a building inventory although the other way around might work, Arguably.

[Slide 9]

And here's a here's another example of 480. You can see in this case the scale is such that you can see the individual grids and you can see that it's largely an estimation process. This is a global product that we've estimated here, so this will be referred to as a level one product, and you can see in certain areas in the right by the bay there, I'm sorry, by the internal lake, there are areas where you can see that there's building stock that weren't able to be captured by the process. So, there is air involved in the process. It is an intelligent smearing, if you Will, and it's important to understand how that's done, where that comes from, in order to use the data Appropriately, which will be some items that we will touch on later in the presentation.

[Slide 10]

And here's another example for Tanzania. It just has the raw count of the buildings throughout the area and you can see the urban areas such as Dar es Salaam might be quite intense in terms of the number of buildings, and then it's distributed all throughout the country throughout the various rural areas.

[Slide 11]

And in this case, in both cases actually, it's a 15 arc second, which equates approximately to 500 meters depending where you are in the world. So there's different ways of putting together, many different ways of putting together building exposure data and at the end of the day, you receive this building exposure data and in many cases you're not quite sure what you have or how it was developed or what the process was or how dependable it is, and people try to use that, have to use that to make decisions often without understanding the limitations of the data. So as part of the METEOR project, one of the things that we tried to do and tried to reinforce is appropriate use of and understanding the scale of the data so, we came up with the idea of scales and levels, depending on borrowing a lot from the GEM levels that were already in place, but also the literature and so forth. So that there would be a way to refer to the actual process that was used to put together the data in a way that you could communicate to others what has happened. So we came up with five different levels and here they are represented for Los Angeles County, which was used to represent uncertainty on the project as well. This is why it's in Los Angeles, but here we have five levels of data represented that we collected for Los Angeles. So the first level would be global scale data, so that's data that would be collected using global data sets, and made it make assumptions from there. Level two is where you collect national data and use that to put together in an exposure

Database. Level three is where you augment that with local information. Level four is aggregated up from building-specific data and level five is essentially building-specific data, so if you look at each of these you can see that there might be, there's quite a bit of difference in terms of the spatial representation of, in this case, the value of buildings. As you go into level 3,4,and 5 you can see what's actually the Wilshire Corridor coming up from the coast. There this line of purple gets sort of more distinct, more or less distinct depending on which data sources are used, and it becomes crisper, almost like something is brought into focus and that is because of the scale. that's of the data that's used so, if as you get to level five you can really start to see the changes in density throughout the suburban areas of Los Angeles whereas that is a much less clear in types of one and two so and likewise you can see between two and three there's quite a bit of difference in terms of how that weighting and value is represented as you move out from the center of the city so, we'll talk about talk about these levels in a little bit more detail As I go along here but an important thing to re to remember is even though you've got a representation that goes from quite arguably quite crude at the global level to building specific data you don't necessarily get a better understanding of what risk is resolution of data is not accuracy of data and we did some interesting experiments with the data sets that are represented here and what we found is just that which is an important lesson I think for the community to come out of the project.

[Slide 12]

All right now I'm going to go through those levels in a little bit more detail so when we talk about level one level three level five data you'll have an understanding of what that means. Level one data set as i said is global data. It's typically global but it can be continental or regional. It's data that was not collected at the country specific level. but essentially, it's it tends to be an aggregate of aggregates. It can be based on global population data sets with some assumptions with just a few things that are that are tweaked from a specific for a specific area but the purpose is essentially to do a level one, to get something that can be used very crudely to assess risk, and figure out where you might want to take a closer look.

[Slide 13]

Level 2 is country-specific exposure data. A good example here is displayed on the right. That's actually the HAZUS data. The HAZUS U.S. database behind the FEMA product. It's when you've collected all of the data, the best data, that's available at each point in the exposure development process at the country level and bake that into the best level two exposure database that you can. Typically, if there's any regional adjustments- and things such as a household size number of people per household and so forth that will either come from the census or based on just a real limited change that an expert might make. Obviously if you were you release level two and a specific municipality is trying to make decisions, the losses that you get and the results that you get from the level two analysis may or may not be appropriate.

Typically with these national packages they might be run at the state level or run to come up with national reports, but if a local entity wants to use it to come up with their own hazard mitigation plan or what have you they would take that and have experts augment that data to come up with a level three four or five database, and we'll discuss that as we move along here. So there are certain key pieces of information that have to be brought in to make this exposure database that are done at the national level that are worth noting. Structural type distributions are typically quite crude on national level with minimal vocal adjustments. A number of people per household there's a key statistic and coming up as we'll talk about throughout the results here. Number of people per household is key and getting to an estimate of actual square footage distribution of buildings as well as the replacement cost. The number of people per household in a level two will would hopefully come from census data and if not the assumptions that are made at the nationwide level can really cause a lot of error. Again also household size- that's a tougher one not typically inventoried by census but where it is, that's that can be used and makes the data quite a bit more accurate. But if not if you say the average house is 1500 square feet or whatever that will cause error as you move geographically out through different climate regions and economic conditions and so forth where that assumption may or may not hold. Finally, the building replacement cost per square meter is also a key determinant of the final results which when you have that estimate at the country level that's prone to error.

[Slide 14]

Now level three is when you basically have that same type of national data that you use as a base but you've gone in and made significant changes based on looking at those key variables and looking at how they might vary throughout a specific region. Here's an example, I believe this is for Addis Ababa in Ethiopia, where you take a closer look at things like construction patterns how that might vary throughout a given country, what the difference in fragility would be in a given state that might have done quite a bit of work to look at things like hazard mitigation and so forth, identifying major urban areas and enhancing building counts, structural mapping schemes in those areas particularly in high hazard areas and so forth. So it's essentially taking those key variables that I was talking about above- the vulnerability assumptions, people per household, and the replacement cost and trying to figure out how these can be fine-tuned or modified on a state-by-state level to come up with results that are more accurate. That can be as simple as integrating more census data, or sometimes it's more difficult where you have to really drill down and have structural engineers look at areas or get cost reports that give you a better handle on what's going on in specific areas throughout a country.

[Slide 15]

Now level four, we sort of change paradigms. Level one two and three started from GIS data that has already been aggregated. Information that was collected locally, largely through census data, and essentially distributed to gridded or more localized area. With level four data, you're starting with building specific data and aggregating that up to make a decision. That tends to be- not always, but tends to be a more regional analysis. It might be a county as we did for Los Angeles or it might be a tiny area of a county. Or you might have a significant area with EO data and building extraction. You might have data for all of California for example, and you'll be able to aggregate that up with some of the Microsoft building footprint data or OSM data. Then that becomes more reasonable to do if there's a complete survey of an area but you have to be very careful. Now, why would you… if you have building specific data why would you create this aggregated and not just use data at the building level data? There's a lot of reasons. Typically if you do have building specific data it's not going to have a lot of the information that you need to perform a risk assessment. It's not going to have detailed structural information. It might not have information on the number of stories, for example, or might have replacement costs that are collected for a different purpose, so by aggregating that data up from the detailed level to an aggregated level you're able to overlay, if you will, assumptions about modifiers or the distribution of attributes that aren't collected at the building specific level that you can then use for a loss estimation project. So, say for example, if you have OSM data that's complete for a given region say this region which is Dar es Salaam, you could use that building footprint data to come up with an estimate of the number buildings and the square footage. But the OSM data might not have complete information about the number of stories likely it won't and it might not have any information on the vulnerability, so you could take that OSM data, aggregate it to that gridded level, and then on top of that perhaps come up with a structural distribution of the of the vulnerability attributes that are that are available on a gridded basis on or on a neighborhood basis, and then apply those assumptions to the grid cells. Likewise, if you don't have a number of stories, you might divide the city into different classifications and then come up with a distribution that you are then able to apply for those grid cells which would provide a multiplier to extrapolate the total square footage of buildings from just the square footage of the building footprint. So the level four builds on building specific data. Arguably, you would think it would be more accurate than a level three data because it does have that distribution of buildings on a site specific level. But still, depending on the data that you're starting with, it is not necessarily more accurate which is something we need to keep in mind.

[Slide 16]

Level five. Site specific data. This is when all the data is provided at a site level. So this arguably is a building inventory, although it's not necessarily an inventory for a risk assessment project. For example if you do have, again going back to the OSM example for Dar es Salaam, if you do decide if you're looking at flooding results and you really want to get that site-specific answer in terms of where things are spatially, you don't want to aggregate up to a larger level because you're worried about the way that that would intersect with the flood inundation polygons so you decide to use that site level. What you can do is take those attributes that you would have at the gridded level and sort of push those down to the site-specific data level using a randomized process. If you understand that that's what you've done and the limitations in that, that can be very powerful for looking at a very localized hazards such as landslides and flooding. The risk there however is if that data set is used by somebody else then there might bring up a building and without knowing that you've randomly assigned building heights or without bothering to read the metadata they might say “this is a one-story, not a five-story building” and that has a reputation hit. Or they've decided to plug that into their tax assessor data. All sorts of things can happen when the data leaves the office. There's a risk there in that the data could be misinterpreted and misused later down the line. But a level five data can be something where the data has a much tighter relationship between the attributes and the level of analysis that the loss estimation analysis where you're able to use that data without too much augmentation. Or maybe you just have an uplift factor for example if you have a valuation used for tax assessment purposes and it's a few years old and you want to adjust for inflation. You can do that uplift without too much of a risk. So each of these levels are meant to communicate the resolution and the methods of coming up with a building exposure database, but it doesn't cover those details. Those details still have to be documented and communicated to the end-users in a way that they can consider it in terms of both understanding the risk analysis that's done specifically with the data and what the appropriate uses are for the data sets if they're trying to repurpose it for other projects. It does not communicate the accuracy we're going to get into that a little bit more later. You can have a level five data set that is less accurate than a level one data set for a specific purpose so I'll be talking a little bit more about that later

[Slide 17]

All right, so we've already talked quite a bit about the inherent resolution here, about data, without talking about what spatial resolution is. We need to make sure that we leave you have a concept of that. Spatial resolution is the unit at which the data is posted. So in this case on the left here, this is Los Angeles City. These are, I believe, census tracts, and the spatial resolution of that data is the census tract level. You would say it's not an even unit geographically speaking because some of these are quite tiny in areas that have dense population, and then some of them are quite large. In this case these are in mountainous regions. So it's not even geographically. But the spatial resolution we would say is at the track level. This is all for the City of Los Angeles. If all of that data was one point at the city you would say that that resolution was at the city level. You can have the county level, the state level, and what have you these are associated essentially with administrative districts. Typically although you can even have uneven polygons that are not associated with administrative districts but that spatial resolution goes down to a unit of polygon that's typically named and the way that is you typically see it would be a census tract level, the city level, county level, or state level. But it's important to remember that sort of a uneven vector level resolution that is typically associated with GIS data sets and the one in the middle here is a gridded data set. In this case this is not exposure data this is hazard data that's posted on an even grid. I can't remember what the resolution of this grid from the Northridge earthquake. In this case there's contours around these PGA contours around those grids so you can see the difference between the contours and the grids and that that resolution would be expressed as the actual dimensions of that grid itself. Typically they're either in a latitude longitude units in which case it's going to be uneven in terms of what you've got vertically and horizontally so it's not a square on the earth and you can see that here these grids are oblong, rectangles. Actually, instead of squares or he might have a 100 meter or 500 meter grid and so on. So typically what we would have is a expressed in units of arc seconds 15 arc seconds 30 arc seconds and so on which correlate roughly with a kilometer and a 500 meter grid cell depending on where you are in the Earth. But it's a gridded data set and of course how accurate that is really depends on the process that needs to be understood. You can have a 15 arc second database that's based on a global product or you can have a 15 arc second database that's based on aggregating up from OSM data that you had for a specific project very different in terms of the resolution the accuracy of the number of buildings at that specific location. So it's very important to understand the difference between spatial resolution and the accuracy. You can post data pretty much any resolution that you want that's not going to make it more accurate. And to the right here we just have a an illustration of if you've got administrative data and grid of data and you try to cross-reference those two you'll have a disconnect between- well not a disconnect- you'll have assumptions that have to be made where some grids have some administrative units and have many grids within and some grids have many administration districts within them. In this case census tracts within them so just because again just because you've posted results at a very fine level in this case the census tract level doesn't mean that the data behind that is has meaningful variation of observations throughout that level so in this case these results were generated using this a gridded product. In areas with very small census tracts that variation that you see in terms of the results are only because of variations in the exposure not because of variations in the hazard. Whereas these larger polygons had many different observations of the hazard that went into it. You have to be very careful to understand where the data came from and not take it at face value in terms of the spatial resolution with respect to the accuracy.

[Slide 18]

All right next discussing building vulnerability attributes. I’m not going to go into too much detail here because the attributes of buildings that are required for loss estimation vary greatly by hazard and by vulnerability functions that are used in a given risk assessment product, but essentially first we want to introduce the idea of what they are. If you look here in terms of the upper right here what we have looks like six damage functions related with flooding. And what this is here is if you look on the axis is this is the water depth estimated in meters and this is the percent of damage that you would expect the damage factor which can easily be translated into a percent that you would expect given that depth of flooding. So what it says here is if you look at the level one there's about 90 85 90 percent damage is expected if there's one meter of water, whereas this one on the bottom in purple you'd only expect about 30 percent. Now I don't know what types of buildings they are but just without knowing anything I know that the building materials associated with one are much less resilient to flooding than they are at level four so level four might be a brick building or a steel building whereas a level one might be an informal building. You'd have a very different damage factor if you were looking for those same types of buildings. If you were looking at earthquake for example. If you've got a brick building might be much more hazardous than a wood frame building, which would be more damaged in a flooding event. For each type of these hazards you have many different attributes that can go into the vulnerability that have to be represented or that can be represented. Right off the bat number of stories is very important both to look at the exposure that's on the bottom floor of floods or the spectra that it's most vulnerable to for earthquakes. A first-floor elevation is something that we often look at in terms of floor of flooding because that tells you where the water is going to start to get into the first floor where it can cause damage. The structural materials that it's been made of whether it's informal materials indigenous materials wood light wood brick or what have you. Lateral force resisting system. This is a much more complicated in terms of earthquakes looking at what the structure of the building is made up and how that responds to shaking. Whether or not there's been retrofits is something that's quite key in some risk projects. Nail density that one is quite important for looking at roofs for hurricanes. The more dense the nails are the less likely those shingles are to fall off and the exterior of the building compromised so that the wind can go in and then that starts to do quite a bit of damage. So that one factor that's very specific, it's very difficult to inventory but it can make a big difference. Even the distance between buildings is an example. So I've taken just these sort of overview of the types of things you're going to look at for a specific exposure development product is going to depend on the hazard that's relevant to that area in the world and what your purposes are. The reason I've chosen some things like nail density retrofits that you can't see is to make the point that a lot of these things are going to have to be inferred and they might be inferred from things like when structures were built or when that neighborhood was built out which as either assumptions that you can make from satellite data or you can make based on detailed tax assessors data. A lot of you can go out and inventory structures and try to make an assumption from there. There's a lot of different things that you can do we'll be talking about some of them a little bit more in detail later when we talk about putting the databases together. And Vitor in his presentations will be talking about vulnerability as a concept in quite a bit more detail as well.

[Slide 19]

Now moving on to replacement costs. The first question that you have often with replacement cost is: "what is it?" A replacement cost typically, is referred to as the price it costs to rebuild a structure. As you can see the damage functions are often applied directly to that number, to have that damage factor. To get that 85 percent that 85 percent gets multiplied by the replacement cost to yield what the damage is. But there's a lot of debate as to whether that replacement cost value actually, includes the cost of demolition. Whether it includes architectural costs whether it includes a depreciated value, in terms of what a building is worth now that it's say 100 years old rather than if it's a structure that was built just this year. There's no easy answer for this the damage functions don't typically take that into account. So, what we tend to say is be very clear in terms of what the number that you're using represents to you, for that specific project and document that so that others will know when they when they use it for a project. That's really the most important thing to say. The damage functions are not refined enough, and the results that come out of these loss estimates are not really refined enough to have a different curve for architectural costs for example for demolition. What we assume is if there's 85 percent cost in there, damage, for example which is less than a hundred percent, that the other costs that you're going to have that are going to go in are going to scale proportionally. So we just put in one replacement cost we don't have different damage curves but it's clear that you need to make be very specific about what you're putting in there so that others know you're looking at typical replacement costs.

What does it cost to build a wood frame structure for example. You're not looking at exceptional, you're not trying to capture the range of buildings. It's really what the typical costs are for that type of structure. If you have information on how that varies by census track by all means uses it, but you're not typically looking at building specific values or trying to look at architecturally tricky buildings and so forth. Tricky to develop. It can be very tricky to develop and it's not always meaningful. Here in this lower right-hand side I've got a picture of this hut. I just love this picture of this hut. It's amazing to think how that was put together it's just ingenious, but the replacement cost by some metrics would be zero, right? And that would be zero because whoever built the hut probably wasn't paid. It's probably their own home. The materials that they used to build it out of they've probably collected themselves so. There's been no exchange of money for labor, or, for argument's sake, there has been no purchasing the materials from a source that can be monetized. However, obviously there's a lot of work there and the disruption of that the damage of that from a flood, for example, needs to be accounted for. It can be very tricky in terms of trying to evaluate on traditional structures in terms of money, and as long as you understood that that's what you've done, and you can communicate that, then that's all right. But it's important to make bare your assumptions.

You can correlate with income. This is something that is, I think, really important to understand. People generally build the best home that they can. If they've got more money, there's more expense that are that's involved there. So, there's a human factor that also needs to be captured. One of these loss estimates in terms of population exposed. But having that replacement cost vary by economic activity can really refine your results.

In terms of sources. There can be all sorts of sources, but you have to be careful in how you apply them of course. Here's three types of ways to come up with replacement costs that that we use in various circumstances. One would be the one in the middle here for Senegal would be resettlement action plans. In this case they were developing, well I guess it's not in Englis. They were developing a dam, or building, a highway, or doing something where they had to purchase buildings to get them out of the way. They've negotiated a fair price with the locals there or what they decided was a fair price for this project and in doing so provided evaluation per structure that we can use to come up with the square footage the value, if there's enough detail in that report. We find those very powerful. Other sources- this JRC report here that has global depth damage functions also has valuation of structures based on GDP. Quite a clever method here that we've used in various circumstances and then just looking at one of the regular pricing surveys. But generally speaking, I would apply to towards only the most robust construction.

[Slide 20]

Vintage can be very important. Here we have a pop grid viewer that looks at different population data sets that are available globally, and in this case they're providing the vintage of a single data set GPW global population of the world version 4.11 and it's telling you how old the data sets are. These red ones are 2001 to 2005. The yellow ones are 2006 to 2010. The blue ones are the only ones that are since 2010. In this case they only go up to 2015. So in all those circumstances those data sets are going to be old. Those population data sets are going to be old and if you're basing subsequent extrapolation of the number of buildings the square footage of buildings and replacement cost on those old populations, you're going to introduce error in terms of 1) where either where construction has been or where there's been development within a country. You might not get some cities. Some cities might be less represented than others, and 2) as a whole you're going to be off. So if there's a 10 per cent wth in a given country since 2006, for example, you're not going to be able to capture that. It may not be a big deal if you're looking at countries where there hasn't been a significant amount of development, but generally, if you're trying to use the numbers that come out of here at face value, you want to at least scale based on estimated population statistics which can be obtained from groups like the UN or you can even find things on Wikipedia. But vintage is a key determinant.

[Slide 21]

Challenges. You've had a whirlwind tour of what exposure data is and what some of the components are. I've introduced some of the challenges along the way but these are things that you need to think about when you're putting together exposure data, or that you need to consider that others might have done. First, is data availability. When you're trying to get these data sets it's sometimes very hard to put your hands on them. If you even find them, you might not get permission to use them for a given purpose or there might be a charge that actually makes things prohibitively expensive in some cases. So those can be challenges. Processing challenges. Obviously taking all these various pieces and putting them together with GIS and remote sensing software into a product can be a challenge. It takes a lot of different skill sets. You need to have a lot of GIS work, but you also have to have structural engineers involved as well to help make the decisions in terms of vulnerability. So that's a skill set that not a lot of folks have that can present quite a challenge. Bias. When you're putting together these data sets from multiple sources there's likely to be bias in those data sets. As I mentioned, the population just even that if the vintage is old that's going to be a bias away from areas where there's been more growth. But bias in terms of census. Sometimes census isn't as thorough in some areas as others and all sorts of bias. One thing I think I've mentioned aggregating up OSM data which can be very powerful but also introduces bias because you then sort of looked at only the areas where there's been OSM projects or people have decided to contribute to OSM. What you get you know might be towards areas where there's universities. For example, and perhaps less in rural areas not just depending on the country. Human error. A lot of steps to put all these pieces together and it's very easy to make mistakes. That's why it's very important to look at what you've done and make sure that it makes sense. Even if you think you've done every single step independently right when you look at the final product, you might find things that stick out and having folks that were not involved in the development process. Take a look so that they can see us well. Data, gaps, no matter what you're doing, there's going to be something that's not available. That's something that you don't know where there's going to have to be some sort of inference and you're going to have to guess. You're, going to end getting the right people that are appropriate to make. That guess is ideal, but sometimes those people aren't available and are not answering your emails, and you have to make that guess yourself. So that might be something like the height distribution of buildings in a certain area, or it might be something like the number of nails in roofs and how that varies by your build, for example. There's a lot of assumptions that need to be made, and this is why documentation of the process is so important. One that I face quite a bit is misperception of the data. You know it's very easy to look at those grids and see the count of buildings in a GIS map and say: "Oh well, look at that. That looks great. It seems to vary where the population with the population and it must be- or it must be perfect", but not understanding the assumptions that have gone into it. Where the data gaps were what the limitations are is a challenge and one that we've tried to address here by first introducing the levels, but also really focusing on the documentation and the metadata, that's provided along with exposure products. Having people take a look at that is a good chance to dispel some of those misperceptions. Accuracy. It's a tricky thing to characterize, and the validation is a tricky thing to characterize because you have so many pieces and so many data gaps that are part of the process. For example, it's hard to say, well, the results are accurate to within 50 percent unless you have some completely independent way of getting that and checking that at a building level and with the losses you know. If you're talking about the, how accurate are the losses, the losses that will result and the risk that results from using in one of these exposure databases when cross-referenced with a hazard data and vulnerability where there's quite a bit of uncertainty on those components as well, becomes extremely difficult to characterize. Indeed, that has to be considered when using the data. It's really something where you have to sort of understand the process, rather than can be given something plus or minus about 50 percent, for example, and that leads right into the next challenge is false precision and that is taking those results that essentially been extrapolated at a local level and using that. For example, to make a decision as to where to fit a specific building rather than a community or type of building. For example, you don't want to definitely when making site-specific decisions, take into account what you should see with your very own eyes to augment the results from a loss estimation product. Explaining the data with clarity, so I've hit on a lot of different topics, even in this just beginning introductory section that illustrates some of the complexity here, it's important to be able to describe what you've done with some clarity. We'll talk about a little bit more about that in our metadata section. Inappropriate legacies. I've referred to this a few times in this presentation. This would be if you've collected data for one purpose and it's reused for another purpose, and this can be. It was reused in 10 years and it's no suitable for a given community or it can be. The planning department got a hold of it and is using it to decide how many people to send out to do surveys. For example, it's really the repurposing of it. Turnover. A lot of times people who've put together these data sets and put a lot of effort into coming up with these exposure data sets and over a few years, go on and get other jobs. If you've at the national level. If there's turnover in terms of exposure development projects, the legacy of being able to understand how that has happened and to be able to do it again can often be quite a challenge. This last one is sort of a nice problem to have is advancing technologies. There's a lot of changes in artificial intelligence, UVAs additional sensors coming up started is now a lot more prevalent. Being able to understand how to use those sources to intelligently augment the development process. Exposure development process is a challenge, but it's a nice challenge. So AI, for example. A lot of folks are running analysis on street view, data and being able to do things like infer the number of stories or whether you've got soft stories or whether you've got a residential or commercial district. How to use that data. That's site specific, along with other data, sets to develop data regionally, for example, is a challenge that comes up with some amazing representations of buildings. How to use that to augment building inventories or building exposures is a challenge.

[Slide 22]

This gets right into the expectations. You cannot typically expect accurate number of buildings at the cell level. The count is approximate. Now, if you've aggregated up that with a level four data set, that might be an exact number. But typically that's not something that you can expect will be there. You can expect more accuracy than the original base data sets right. You're not going to have a census that was taken 20 years ago. Try to do your best to come up with a building exposure database and do a better job than the original census data, even if you've scaled it. It's just not going to be able to capture that distribution. You cannot always expect to capture small unmapped rural areas, even with remote sensing data. This is a continuing challenge and if you've got a project where that is important for example, you're looking at landslides, then you're going to have to really do some focused reconnaissance or some manual work. There, challenges with remote sensing data sets will impact the results. There are some areas that are almost always under clouds very difficult to see. There can be tree cannon canopies over structures at all points in time, and you won't be able to see those from remote sensing processing. Lighting. The night light data VIIRS data- is really great for being able to capture rural areas and a lot of cases, but some areas do not have any lighting at night and they can be detected from space. So that makes it even more of a challenge. Also when you've got situations where buildings are hard to distinguish from material, for example. If you think of a hut- very different. It's difficult to distinguish that from a bale of hay, for example. It's just that the building materials is what and the difference between man-made building materials and natural building natural materials is really what allows us to detect that some areas that becomes very hard. If you don't have that's going to start roofing material, it allows you to detect these things and again. I’ve made this point a couple time- you cannot repurpose the data for civic purposes such as address specific information for tax purposes. This comes up quite a bit. The sustainable development goals into poverty, sustainable cities and communities, climate action, all these SDGs and the Sendai Framework answering these questions at the national level, where you have to make tough decisions and in terms of how to put forth these numbers at them at the national level. That's really the type of thing that done these building exposure databases can help with.

[Slide 23]

I'll be talking about the basic process of developing exposure data. Developing exposure data is quite complicated, although you can do some work right off the bat like any discipline. Looking at vulnerability or hazard, it can get quite complicated, so you'll by no means be experts. Here we won't be going into the structural engineering details. We won't be going into the GIS processing steps, remote sensing processing, steps and so forth, but I'll give you an idea of the various components and what to watch out for. And I think, most importantly, it will give you an idea of how sausage is made, as we like to say. What the ingredients are and how that will impact the decision making process with regard to ultimate use.

[Slide, 24]

So, how is exposure data developed? Here's the general steps. Again we're sort of focusing on those levels, one two and three, not so much looking at if you've got building specific data available. But first collect your census data. Interpolate density. You can secondly estimate the building attributes for vulnerability, refine the spatial distribution, estimate number of buildings and estimate the replacement value. So those are five major steps.

[Slide 25]

First, collecting census data population census data and housing census data if you can get a hold of it. Get the most recent data you can it's worth trying to get the updates. That might be more frequent than a kind of decadal type of census and extrapolate those to current values. If you can- and you generally want to get the highest resolution- data that you can for the country, you're looking at a census, tract type of analysis. And, if you can get micro census data that that is beneficial as well.

[Slide 26]

Next estimate building attributes. These are the type of attributes that are on the legend here and the GEM application in the upper right. These map directly to vulnerability. It's generally good to use, a statistical or mapping scheme approach and because you're not dealing with an inventory here. You don't have information about every building, so taking a statistical or what we call mapping scheme approach. Mapping scheme is short-hand for a mapping, a certain occupancy or a certain development pattern to a distribution of structural types. So, for example, if you were to say residential structures, you might have more wood frame buildings in some countries, or you might have more apartment buildings in some countries. Maybe people tend to live in URMs, whereas you have commercial districts in residential neighborhoods. They might look very similar, but in downtown urban cores they might have a very different structural distribution. So finding those groups and refining those distributions is what's required here. Lots of ways that you can do it, and obviously you have a different approach for level one than you would for something along the lines of a level four. It can be as simple as what we did for the level one data that we produced on the METEOR project, was to do things like look at the world housing encyclopedia. We do a lot of web-based reconnaissance, look at as much street view and photos taken insitu, and just regular satellite data and aerial photograph data that we can; and then use that to essentially make assumptions. That's what we would call the global level. For level three, you want to get much more accurate data. You might do a full literature review on each country to look at the predominant building construction types, do a much more detailed interpretation of satellite data where you might manually extract certain areas like downtown business, cores and military areas, soliciting expert opinion. You can get structural engineers. You will be able to recognize some things, but a lot of times you can't see a whole lot from the outside of a building given the cladding and so forth. So having experts that are in country who understand the construction and maybe how things change by era, how you might be able to use what you can see on the outside to interpret what's going on underneath is very useful. Virtual reconnaissance. That's the street view type approach, site surveys. This would be the type of thing that we did for our level three data. We'll have a session where we discuss that in much more detail, where you actually send people out and do a stratified sample. So, within a given region, a development pattern that you're able to identify and say the remote sensing data you drop random points and for each point you go and collect, maybe 10 buildings nearest to that point. The structural characteristics, and sometimes the occupancy distribution. If you have that type of information that you can map it to, and then that becomes the basis for extrapolating for like regions throughout the country. Of course, it's more accurate. If you can collect that data in several different places in the country that might have very different patterns in mountainous regions than in the valleys or in the north of a country where it's cold, rather than in the west, or what have you. So being able to identify those patterns and collect site surveys at multiple points is ideal. And of course, the closer those site surveys are to where the risk assessment is being done the more accurate it's going to be. So, for example, we collected data in Dar es Salaam and Kathmandu for our level 3 analysis. We were able to determine that there wasn't significant structural differences between Dar es Salaam and the rest of the country that would have impacted those mapping schemes, but we did go through the process of looking at as much information as we could to adjust those as appropriate. For example, Zanzibar had a very different mapping scheme,

[Slide 27]

In the next few slides we have some examples of some of those methods of collecting, attribute information. You can have just paper forms that people fill out for a given area where they're very simple, whether they circle a given type of structure and the details and the error of construction, for example, number of stories. Or you can have very sophisticated tools where you've got phone applications or desktop applications that tie a mapping, mapped point to observations. And, of course, there's a little bit of an adjustment. The device records where you are and you have to track it associated with the building. So you have to click on the building and make sure to fill those points in or else you'll have the recording in the wrong place, which can make a difference. A word of warning during this process. You know you do a random a sample, but sometimes you get outliers in your random sample if you're doing an assessment and a mapping distribution- and you happen to hit a 15-story hotel, for example, and there's only five of them in the country and those get replicated in urban cores, based on that one sample, you have to make the judicious decision to eliminate that point. So really you're trying to characterize what the most common structural patterns are. So field tools, there's an IDCT tool which I believe is still distributed by GEM, which is a very handy tool. We've used quite extensively. OSM has their tools that were used in this project in Kathmandu, and we did that survey as well as some surveys in Tanzania. A field notes pro is a common application that does a pretty good job of allowing you to take notes in the field very simple, with a lot of flexibility. Paper surveys with just geotagged photos allow you to make an awful lot of observations. If you can just collect a lot of those without having to send engineers in the field, that can be quite nice and again it's ideal to link that back to the footprints. So you have those observations tied to building areas, a number of stories too, so that helps you.

[Slide 28]

Here's some examples I believe from a project we did in Indonesia on the bottom. You see four different photos of buildings. These types of observations were tied, two different commercial districts. And in the end, you have essentially a distribution of the total square meters and the total number of buildings which allow you to make your remote sensing. You can't see as much from the sky, but you can see a lot more from the ground.

[Slide 29]

So, for example, China has got a lot of very regular development of residential structures that are- well, I think these ones are six stories high, but sometimes they have ones that are 14 stories high and they look very similar and you can spot them in all urban cores and map them quite effectively with satellite data. So you can figure out what those are made of. You can match that with photos of construction. You can talk to Chinese engineers and so on. But being able to have that additional satellite perspective allows you to figure out exactly where they are and use that, in your risk, analysis and kind of join those two pieces of information is quite effective.

[Slide 30]

A sort of word of warning: don't go overboard. There's a lot of different attributes of structures that affect the vulnerability and a lot of data out there about how that happens. But if you're doing a study and you're saying well for each site that we identify, we want to figure out what the occupancy is and given the occupancy, we want to figure out what the structure type is and then from there, whether it's a reinforced or unreinforced or confined masonry or if we want to figure out the number of stories and if we assign the era, you very quickly get to sort of branches in this in in the tree of options that will be more frequent than or have a greater number than the frequency of buildings that you're trying to distribute those distribute the them to. So what you're really trying to do is get the solid picture of what's going on in an area and when you can have those additional attributes that affect vulnerability. Sometimes it's good to statistically represent those on the side for those doing the risk analysis to be able to assess. But if you're, just using that sample data to distribute buildings in a Monte Carlo simulation, you really want to go for the most common types and construction types in the buildings and in that in within that regard too. That allows you to sort of avoid the situation where whoops- and this is an actual case that happened in Indonesia, where we randomly sampled a 15-story hotel and that ended up first, the first run of the analysis ended up popping up, so you really want to go for the most common.

[Slide 31]

And next refine spatial distribution. This is starting to drill down into some of the benefits of earth observation data, remote sensing data. So in this sample we have administrative units, I believe these are in Nepal. I'm, not sure how large they are, but these dots are about 500 meters apart, so the buildings that are referenced in that administrative unit bin or the maybe they're, not buildings. Maybe the populations represent a distribution throughout this entire area. So, if you're looking at earthquake risk that may or may not matter, distributing them evenly throughout this zone, but certainly it's more accurate, and particularly, if you're starting to talk about things like a landslide or flood to be able to localize them by these 500 meter grid cells based on where you think they are. That is the process of which earth observation really helps, is localizing. That distribution and you can see in some of these areas, you're starting to characterize the valleys or the small towns that are within these administrative districts. So that's the benefit of remote sensing in that process. This term that started to pop up quite a bit daysmetric mapping, but really what you're doing is inferring the density of buildings, of your estimated buildings within that polygon and distributing them accordingly. That's the meat of what we've done here for the level. One analysis, for example, for the globe that census data that we have in that case might be quite large and you're, basically interpolating it to a finer surface. With EO based weighting now several products that do this globally with population data that you can use I'll talk about some of those later. If you don't have the expertise to do that, distribution house, you can use some of that stuff. Land Scan, World Pop, for example, GPW Pop Grid, is an effort that does a good job of cataloging a lot of those and advising people when to use one over the other. But when you're really looking at buildings can be quite different from different in different development patterns. You know rural versus city and so forth, and it's you want to probably go that extra step that the population mapping doesn't do to look at the different construction patterns that you're able to characterize with that sampling,

[Slide, 32]

Next, estimate number of buildings in square foot footage per building. This can really impact results quite a bit and impact the maximum amount of your losses. The more detailed information you have about the number of essentially people per building ultimately. You typically start with information about how many people are where. Then you have to infer from that of the number of housing units. Or maybe you have a distribution of the number of housing units which gives you that, then you have to figure out what the area of those housing units are and then you have to apply the replacement cost, which is the next step. So there's a lot of different sources that you can. You can get to be able to estimate the building space, ultimately per person. Sometimes the housing census will do better than the population census. It will have the number of buildings sometimes and they might even have things such as the number of rooms if they have the number of rooms, and you can use that typically to estimate an area in meters quite effectively. If not, you have to drill down to another level of detail, to try to infer that OSM building footprint data can really help out here. In that case you have to really review it for areas to figure out what that relationship is between the population as estimated by the census and the number of buildings. And if you're going to use that as your parameter, you have to review that OSM data to make sure that it's complete for that area of interest and you have to make sure that the vintage of your census data is pretty recent or that area of that you're using to make that adjustment has not had rapid growth in the past few years. To do that, you can really just go back to the google earth platform and drill back and look at the previous aerial imagery. But if you essentially get some samples throughout the country in lieu of any other data that will help you out. Height profiles, this can be very important if you're trying to use that OSM data and you're trying to extrapolate height out of that. Micro census data- if you can get your hands on it depending on how much data a country has collected, can really be very valuable here. We used an IPUMS database, distributed here from the University of Minnesota, to be able to make those assumptions regionally and what we couldn't have for a specific we couldn't find for a specific country. We were able to basically chunk out into regions. Of course, we do provide the documentation about how we use those regions in the metadata. The last idea here, population density can also correlate with square footage in a way that can be characterized.

[Slide 33]

Next, estimate the replacement value. I talked about this quite a bit earlie. Lots of different sources for this- things, like blue book, the JRC method of trying to extrapolate based on GDP. One of my favourite things is to drill down into the resettlemnts settlement action plans. But you have building value by meter by building type or by occupancy. Different sources, break it out in different ways. You need to figure out what's the best for your project. Sometimes you can use one method in one area of the country, another method in another part of the country. Building construction manuals can be useful, but they really tend to particularly in developing countries, focus on the most rigorous construction in the city capitals and may not be relevant as you start to go outside those capital areas. So you have to be intelligent about how you use them and when you use them and how you might downscale values. You can use them to essentially pin a value for the most rigorous construction and characterize temporary construction or semi-permanent construction as a percentage- that can work very effectively. We've done that quite a bit in Africa. Expert opinion is good, but expert opinion can vary quite a bit. So again it gets down to how you ask the question, the expertise of your expert. It can be best if used as a data point to triangulate in on an answer. If you have other sources as well, again, the GDP by median income is an effective method. We've used that and updated it from the JRC type of approach. I encourage you to take a look at that scaling by building durability. I was just talking about with the temporary and it can be. It can be quite difficult to estimate the replacement costs in developing countries. I use the example of a hut,

[Slide, 34]

where it may or may not be meaningful. You just have to keep that in mind and you have to be able to communicate to people in those circumstances. Sometimes the number of buildings will be a more meaningful metric than the replacement value alone that are affected by giving ground shaking. Again I mentioned these resettlement action plans. Here's one for Tanzania, it is 2012, so that's almost eight years ago. So that means these numbers have to be inflated based on changes in construction value, but it provides very detailed information. Pictures of buildings that were purchased, a description of the building, as well as a replacement cost. All that stuff is really gold when you're trying to do this type of work.

[Slide, 35]

Next, the value of earth observation: this is an earth observation project.

[Slide, 36]

We'll be talking about the role of EO data and how we use that. Global population data sets are really very valuable, but you really have to take a close look at them in the country or the region or the city that you're looking at to make sure that it's accurate in that location and that it doesn't need to be adjusted- and we'll show some examples of the variance. A global urban rural or urban intensity data sets are something that are being generated more and more frequently. An urban intensity data set will just give you the percentage of built up in an area whether it's a hundred percent, fifty percent, ten percent twenty percent and so forth. That can be used to infer the number of buildings throughout an area where you're spreading them from the census tract or the province or whatever level you've got the census data at quite effectively. Just basically a weighted distribution there. Again those don't always work great. Sometimes they go haywire in mountainous areas, but a very nice addition to the interpreted EO data sets, I would call them. Also that that urban intensity can correlate quite nicely with structural patterns segmentation of development patterns. I'll talk about that in a little bit more detail. Building footprint extraction; average building size. Again getting that down into that information about housing units and how much area to associate with a person. When you're spreading this data and then I'll talk about challenges and some emerging research.

[Slide, 37]

Global population data sets. Here's an example from pop grid, which I mentioned out of CIESIN NASA center in in New York associated with Columbia University. They did some great work here. Actually they put this together specifically for this use case, trying to evaluate the applicability of population data sets for hazard purposes, where you're able to go into a specific area and see multiple data sets of that area and assess what how accurate that data set is for you. As you can see here, I don't have the legends on here, but there's quite a bit of difference between these data sets. It's essentially how they've done their own interpolation from the data set that they had, which might be associated with or if they've done it with an administrative unit to a more localized region. So if you look up in the upper right-hand corner here for Kathmandu, you can see the admin zones, gridded, that the data was provided by. In the upper right hand of the upper right hand, for example in the mountainous areas, you can see and that's because in this dataset they don't do any interpolation based on the remote sensing data. That's just an aggregated global product that they distribute so that others can use it, and that's why others can use that to essentially come up with an interpolated product. And all these other ones, the GHS, World, Pop and landscan are different, represent different methods of doing that. That interpolation, you can see in some places where there hasn't been any indicators of human development, they've, essentially interpolated, that very evenly to a low number. Or you can use the word "smeared" rather than interpolated, and that's why you have these sort of large orange blocks on blocks surrounded by these white zones. Which is, by their opinion, areas that are unpopulated. You can see, there's a much stronger density along roads and then in the city of Kathmandu itself. And if you go down to World Pop, it looks a little more like the GPW data. You can still see the administrative units a little bit. There is a sort of a floor in the interpolation and there are no white spots and the lower right landscape is essentially somewhere between the two does not distribute as much population along roads. So, in the process of coming up with the level 3 database for Nepal, we had to take all of these data sets and figure out what we thought was the most appropriate methods and where. And we came up with our own interpolation algorithms to do this; but if you're using one of these right off the shelf, you have to look at this (and OSM data is really good here, too, where it's complete) and figure out: what is what is the most area accurate for your purposes.

[Slide 38]

Drilling down into that global urban rural urban intensity data sets, here's two examples again from Nepal. Here's southern Nepal right there on the border of India, a really great area to look at this is in the valley. These mountains are up here. This is just a long river I believe, an area where you have quite a bit of distribution of population and then just a real sort of speckle of farms and villages outside of that that major core development. A nice challenge to be able to identify those smaller communities and I think we were able to do it. This urban rural intensity, I think this is the GHS data, didn't always do a great job, but sometimes, if you're, trying to infer where you've got different development patterns. For example, these areas that are black speckled throughout here a much the small settlements are very likely to be very rural. These settlements have a little a smaller, a different type of structural development, so that you can use that to essentially develop your patterns. Here's what it looks like for Kathmandu. I'm, sorry, it looks like it's got a different color scheme here, but you can see quite clearly where you've got development and where you don't.

[Slide, 39]

So now, I go through an example of the process of using EO segmentation of development patterns. This is a use case for Beijing, China, where you might have data by these very large zones. That might be the finest population data that you're able to have- and your question might be "I know where my exposure is with respect to those zones, but how does it fit with respect to these flooding zones?" right? And obviously the data that you have that's aggregated up at that higher level doesn't really answer the question of water patterns.

[Slide, 40]

So. Here's an example of just how you would use that segmentation to solve that type of problem a little bit different than the risk problem that we have but very similar same tool set. So here's another just a LandSat image of Beijing. You can see we've circled some industrial areas,

[Slide 41]

and, you can use training data sets to basically run an interpretation algorithm over the whole country- it's called supervised classification- to identify the development patterns.

[Slide 42]

Typically there's going to be a lot of speckle, a lot of noise, so it becomes more accurate if you, aggregate it up a bit. We've done that here. This is probably the 15 arc second level, and you can see these patterns throughout the city. So, as you get to the older city core of Beijing, you have more single family and mid-rise structures. The higher rise are actually in the newer areas and pockets outside of the city and that's interspersed around following that.

[Slide 43]

And, then the idea here is that you take those development patterns and you're able to do some sampling of this section from another country, but you're able to do some sampling within there to figure out what percentage of buildings are and that's the structural classification for the vulnerability functions.

[Slide 44]

Building footprint extraction. This is a more standard, typical use of remote sensing data, it's very powerful in one sense, but not so powerful in another sense. There's loads of information- there's loads of papers and so forth that you can find online about what algorithms to use with this regard. Feature extraction modules, I think for ARCGIS, or there used to be any way. So a sort of canned applications just geared specifically towards taking EO data and doing building extraction. And if you have high resolution data at your disposal and your area is small- it's definitely worth considering this type of operation. But it can be challenging and I think these two images here illustrate some of the of those challenges. So the image on the left here is a SAR data set, and this is actually an elevation data set for Los Angeles. And you can really see the development patterns quite well. These large blocks here in the upper left are actually studios movie studios and, as you have this lighter speckle here. these are actually the residential areas.

So it does a okay job at this resolution of extracting out those large buildings; the other ones not so much, but it gives you an idea of what the area are and certainly where they are for development pattern purposes. Here's a similar method, that's done from segmentation, not elevation, but segmentation. Which is just a supervised classification that's done in any standard remote sensing package with optical data. I believe that in this case, looking at a building damage in Port-au-Prince, Haiti following the event. But regardless, these orange boxes, here represent the footprints of buildings. If you were to square these off, sometimes there's algorithms to square these off, to interpolate through the green areas which, in these cases are trees. It gives you a little bit more accurate of an estimate. You can put a lot of work into doing this type of extraction, and if your area is small, it could be worth it certainly to try to get the building area-to-person proportion for different samples. It can be worth it, but what this is good at doing is telling you the total area of the building footprint. By footprint I mean the area as shown on the ground of the first floor, the square footage associated with the first floor. It doesn't tell you how tall they are so there's that adjustment that still has to be made can be made from sampling or expert opinion as well or just based on the development patterns of what you know of an area. But it's not very good at telling you the count of buildings; and if you look carefully, you can see how a lot of these orange areas sort of smash into each other. Some of that might be due to damage in some areas. It works quite well, they're, very distinct and others, they bleed into each other in a way that you wouldn't be able to get a good count of the buildings. Now I would argue that square footage is more important than count when you're talking about losses, but it also tends to work better for some roof types than others. A lot of roofs look like cement. So if you look carefully there's some of these square red zones, which is probably characterized of cement type material, that you also see crisscross and some of these parks and some of these shadows, what we've got here this large building on the right with the courtyard in front of it, is actually the palace, the Port-au-Prince palace. It could be some damage there, but it's got these sort of sort of red areas which is interpreted as a different material, and if you were to count up all the orange speckles there, you would have much more than one that we could see with our eyes is the building, so in general, for loss estimation purposes it can be powerful to have some of this data to give you a sense of the people per building area equation. But it's not something that you generally want to run over very large areas, and it's not something that you typically want to use verbatim as a loss result for level five. It's something that you want to use to make your assumptions for a level one, two, three or three type analysis.

[Slide 45]

I would argue that the better way to do it is to basically define your development patterns that you're, seeing in a country that you can start to classify with EO data and then just have someone go in and digitize those. It's not too much work, gives you a better answer, gives you a firm count and allows you to spent a lot less time wrestling with finicky algorithms.

[Slide 46]

Some of the challenges and emerging research that come in the EO. Low light. I think I've mentioned a lot of these before- low lighting, particularly in large parts of Africa, where the villages won't have lights you won't be able to detect them. It can make it more challenging. That doesn't impact the optical sensors, that are, the standard, luminance optical, visible bands which can still be used to detect these things. But, depending on the material, it can still be quite complicated. Under tree canopies is pretty much impossible. Cloud cover can be a real drag. Something that you have to account for and look at mountainous regions if you're looking at those visible bands a lot of times the patterns that are made by mountains, the real bright exposed rock and the sharp edges look like what AI determines is probably a man-made material. You'll see that in a lot of the global products- they've, they can account for those with a DEM and just sort of strip them out, but then that also strips out settlements in the mountainous region. So you have to really account for that and look carefully and that's something that we did quite a bit of in Nepal for the level three data. Indigenous materials. Sometimes, in a Bam Iran, for example, there was a lot of the construction that was made out of the local material. And from the sky, it was the same color. It was the same color as the streets and the same colors the yards. So if color is not an indicator, then you have to basically use patterns. Pattern recognition type stuff. AI- can be useful there, but it's again something you have to roll up your sleeves and do. Emerging research, new sensors. A lot more SAR data coming online. A lot more high resolution data is coming online. Figuring out how to use that effectively. A lot of building footprint data is being released by Microsoft and others, and the OSM data all of that stuff can really help. If it's used correctly. You have to be very careful about how you do that. It can overestimate exposure if you're not careful or underestimated, but it's definitely a tool that that you can use in your toolbox, as well as data from street view and UAVS. AI algorithms process these images and video streams to detect building types, for example, is something that a lot of people do. But I would argue that it's very difficult to do that for very widespread areas and it's best to be used almost in a sampling capacity.

[Slide 47]

All right now for the next section, we've gone through an introduction to exposure data. So hopefully you have an understanding of what it is. What we're talking about; what we're not talking about; a very basic idea of the elements that that go into putting together an exposure data set; and now trying to answer the question: is it any good? Is it useful for my purpose, not an easy question to answer. I think you need the background in terms of the first two sections that we talked about here, to be able to understand, truly understand exposure data, but I'll be trying to give you some more tools to do that.

[Slide, 48]

Again, going through the levels of exposure data in a little more detail. Talking about the resolution and scale of the of the data and the hazard. With respect to the hazard, I'm harping on the key parameters a little bit more and talking about metadata. We've got a whole other section on metadata that our analysts have put together the format for that in this project. We'll be doing talking about validation just a little bit and understanding the limitations.

[Slide 49]

I want to go back to some research that we did on this project, which was for the Los Angeles area. Again, here are the five levels of exposure data. We had an opportunity because of our history here in Los Angeles, since the Northridge earthquake, to get our hands on a lot of different exposure, databases that have been put together for various purposes and that allowed us to do essentially a la all the way from a level one to a level five full probabilistic analysis of what the losses would be through time, given the different data sets. And that's because we do a lot of work here in California, so we had the engine to do that, so to speak. A seismic program here, and we were able to wrestle the data into those formats and look at the difference. What we found was quite illuminating. We didn’t want to spend too much time in putting these data together and correcting for bias and so on. We just wanted to sort of get it in there and see what the results were. If we were had taken a little bit more care, there might have been a little bit more convergence here in the results. But this was a thought exercise. It would have been ideal to do that if we had the different data sets in in Nepal, where we do there's significant earthquake risk, and maybe we can look into doing something like that later.

[Slide 50]

So level, one data for Los Angeles, we actually downloaded from the GEM open quake site. It's essentially was a prepared, I believe with a Kishor Jaiswal as part of a FEMA p366, essentially mapping their data into GEM taxonomy.

[Slide, 51]

And here's what it's looks like, I believe it was 60 arc seconds rather than 30 arc seconds, so maybe 1 kilometer data, and here it is mapped.

[Slide 52]

Exposure level two data. We just took HAZUS data right out of the package. It's census data. We wanted to do trend for spatial resolution, so what we did was we ran everything at 30 arc second level. So that means that we sort of we took that level one data cut into four pieces there, so that would be the 15 arc second level. For level two data, we took that and just that basic area weighted distribution to chop it up into those squares, so they would be a 30 arc second grid.

[Slide, 53]

Here's what it looks like you can see these large sort of flat areas where things are homogenous.

[Slide 54]

Level three. We use work that we had done essentially looking at building exposure data sets, developing a ITV insurance devaluation. Basically using EO data to do a smarter inference of what the replacement cost would be. So we did that as a level three, and we also used in that context interpreted development patterns throughout the Los Angeles area to adjust the mapping schemes or structural classes of the buildings just a bit too.

[Slide 55]

And, here's the level three data. So you can see it doesn't look too different than the level 2 data, but a little bit a little more differentiation, particularly in the residential areas. These sort of orange swaths on the southern portion of the county.

[Slide, 56]

So level four data was interesting. We got building footprint data that was released from Microsoft Bing. Essentially building footprints that have been released for the whole country. Now, they're really looking at trying to do this type of work for the world. They need more training data to do that, but this is something that we can expect to be developed in probably the next five years. You'll start to see these building footprint data sets all over. So one would assume having that footprint, would give you a better idea of the square footage at least of how that building is represented on the ground, but then that has to be extruded to come up with square footage of an entire area.

[Slide 57]

What you can see this looks quite different than the level two or level three. There's higher highs and lower lows. I think that goes down to and that's it just does a better job at, in this case, capturing at the cell based level where you've got a green areas speckled throughout Los Angeles, where you've got areas where there's more building footprints, a higher density of building footprints throughout Los Angeles. So and if you're familiar with the valley here, you can start to see major corridors, major development around particularly large boulevards and so forth, and that yes, this is the same area that was in that SAR photograph or Image I showed earlier. You can start to see some of the movie studios and so forth in some of these areas. So in that sense, there's a lot more differentiation.

[Slide, 58]

For level five. We used some research that we had done maybe 14 years ago, mapping the structural classes; well, not structural classes. The year built the occupancy and the fire codes that are in the tax assessor data for Los Angeles into vulnerability classes that could be used for seismicat- essentially for our program. Now, seismicat is very much the way that the vulnerability is dependent on a lot of year of development or your construction, so that was a was able to use that data, which was quite useful. So we actually when we, but when we ran it, we did aggregate to a 500 meter, area.

[Slide, 59]

And this is that it looks like. It's similar to the Microsoft office level four data.

[Slide 60]

Here you have them all side by side again. The level one data for the entire area, not as much differentiation a lot of low areas. The HAZUS data which you're able to see things like the Wilshire corridor and more areas of concentration in places like Pasadena. Level, three, which attenuates that a bit looks like down in most cases. Four, where you start to really bring things into focus. Of course, that's because it's aggregated up from building specific footprints. And then level five, too, which is an interesting, almost mix between the level two and the level. Four, at least that's what it looks like visually.

[Slide, 61]

So I want to take a step back now that I've kind of gone through those five levels and discussed them within the context of an actual project where we put together all those data sets for a concerted area. I want to talk about uncertainty because we're really one of the things that we were doing here of running these exposure data sets through the probabilistic analysis, was because we really wanted to figure out what the uncertainty was associated with them. Those data sets, if you put more effort into developing a level five. Is it going to pay off? That's really the question and how much difference is it going to make and if I'm worried about figuring out where I want to do something or where I need to focus? There's certain questions where it might not justify a huge effort, or maybe you don't have the funds to do a huge effort. How much is that going to impact the analysis that you're doing so a lot of people really when you start to talk to them about exposure data?

Talk about the law of large numbers and the law of large numbers is if I throw one dice. There's two dice here, but throw one die hundreds of times the results will converge on a mean eventually. So here you've got the observed average, which will vary. But as you look here, it's the you get to your 200 300th 400th row it really starts to converge along a mean and based on one of the more times that you throw it. So the idea here is, if you've got if you sample a thousand buildings you're going to represent what happens on average, pretty frequently if you've got 200 well, you've done. You've got pretty close if you're only going to sample 5. Well, you might have the answer: 5.6.

[Slide, 62]

But what happens when you've got bias in your results? That theoretical mean is nowhere near what you're dealing with in reality. And, unfortunately, this type of bias happens all throughout the exposure development process. For example, if I’ve got the construction manuals from the turner document, there I go through that and I figure out what it costs to build. A fancy, hotel and Dar es Salam and I apply that to residential construction, I’m, going to be way off and there's all sorts of problems like that bias, the different components that are in the exposure development process.

[Slide, 63]

And, so some of the ones that are right off the bat I’ve talked about them several times already- replacement cost. Results are very sensitive to replacement cost. People per household, average dwelling size, exchange rates can really be one. Obviously we didn't run into that in Los Angeles, but we've run into that in Africa before. And buildings are seldom actually sampled to get these results, and that can make a big difference if you've got a theoretical way that you're thinking the buildings are going to be, rather than sending people out in the field to look at them and then use that to as a with a stratified sampling approach and the census, data itself often has bias. So there's all these sort of problems that you can have.

[Slide 64]

The question is, if you're mapping what your loss is going to be over a return period. This is just a theoretical curve. What's going to be the difference between this level 1 and this level 5? Is it going to be a factor of four? This would be a factor of eight factor of two we really didn't know. And we felt as if no one had really studied this problem and significantly gone through the entire probabilistic analysis, with various exposure data sets to see what those results would be and to see how much error actually can be attributed to the exposure data itself. Simply because in any given exposure project you use the best data. You have right, of course, if I’ve got level five data at my disposal and enough budget I’m going to process that and use it rather than the stuff, that's off the shelf to look at my hazard, so we thought at least for this one hazard- earthquake, we would de-trend spatially and analyze the stuff at the 30 arc second level and then run this to see what how much error that we got. That was just associated with those…

[Slide 65]

…five different ways of putting the data.

[Slide 66]

Here are the results that we got. So these are the losses in U.S. dollars on the y-axis here in the return period and the level five here is on the bottom and the level four here is on the top, and then you have got the level two under level four and level three in the middle and then level one under level three. So it's not monotonically increasing it's not converging, as you have more spatially resolution high resolution data incorporated. So this was really quite a finding. The distribution I think alone is meaningful.

[Slide, 67]

What we decided to do was take a random sample of buildings here in Los Angeles, and I think we had something like 600 buildings- and run them through as if they were records in these various exposure data sets to see what was really the result of what was really impacting the results most. What we found was the area associated with buildings and the replacement cost associated with buildings. You see this level. Five here is down on the bottom. That's the tax assessor data in Los Angeles tax assessor data. In Los, Angeles doesn't have every building and they don't have every hallway. They don't have every garage there's a lot of a building area that they don't include. There are some systematic reasons why, in Los Angeles the assessed value of homes tends to be lower than one would expect. This level four here at the at the top turns out- that's the Microsoft office building footprints. They squared off the buildings and when we extrapolated that, based on height that intended to give us a way more square footage than we were finding when we went in and looked more carefully at buildings turns out. Buildings aren't cubes the second floors and third floors often have less building area and the buildings tended to sort of overestimate, particularly when they started to look at garage areas. So when we detrended for that, what we found is level two three and four were right on top of each other, which makes sense because the damage. The vulnerability there was assessed almost completely from the HAZUS database, with just a little bit of adjustments made by our engineers based on the remote sensing data. And we see level two is a little bit above level, three and four level; three and four having the adjustments that we made. But they're all basically right on top of each other and the level one and level five have different vulnerability assessments that have a different curve, essentially.

[Slide 68]

What that tells us is that there's really very key things that you need to really focus on when you're developing these exposure data sets- and persons per household living area per household, rebuilding costs and exchange rates are all things where will exactly proportionally impact your results. And so, if you think that there's one person per household and it turns out that there's two people per household you'll be high by a hundred percent. It will affect your results by a hundred percent. If you've got the persons per household varying throughout your country substantially based on the way people live, so that it's nine people per household in the north part of the country and three people per household in the southern part of the country, your results are going to be very off if you take a mean, for example, and the same as for with living area per household. If people are have a lot more living area in rural areas, where they're, less cramped and that's not reflected in the data and your estimates, that's going to impact your results proportionally and rebuilding costs and exchange rates.

[Slide 69]

This is some results for exchange rates to the dollar. For a country we were working in Africa and you can see it went to being tied for 100 units per dollar in 2008 to 700 and none in 2000. I think, 16 or 17. It was so that's a factor, seven if you've got the wrong exchange rate you're going to be off by a factor of 7. If you chose 2011 rather than 2016., so you have to very watch for these things and adjust for them.

[Slide 70]

All right, more understanding, exposure data all right. So let's take a look at the resolution and scale of impact, given the hazard. Once again so I think I introduced this problem already. This is the I, don't know if it's a province or administrative unit in Beijing and the problem of how does that correspond to flood in Beijing

[Slide 71]

So. You go back to the original data. If..

[Slide 72]

you want to try to figure out what the exposure that you have in a given area correlates with those flood hazards, it's not going to you're going to need to figure out a way to essentially localize that that data.

[Slide 73]

And you've got here are the two data sets that I had introduced the classified image and the classified

[Slide 74]

aggregated up. So that that data, that's aggregated up, is going to be more accurate in terms of the number of building estimates, the distribution of buildings. Because, obviously, if you're trying to distribute buildings at this tiny, speckled level, you're going to barely have a building to distribute, for example, and because of the essentially the law of large numbers, it's going to give you a higher accurate for this disaggregated area. But the image on the right is going to have higher resolution some examples here:

[Slide, 75]

If you look at some of these mixed pixel effects and this middle circle- and this is a big large train section on the left. Throughout this area, you can see just based on the shadows alone, these large apartment blocks and where they are throughout Beijing, where they've cropped up on the outer areas of town,

[Slide, 76]

And. If you go to a classification of this data, you'll get something like this and you can see where it's not typical development, for example, where those train tracks are, it might not come up with an answer or might come up with a wrong answer. It does an okay job of figuring out where your tall buildings are and identifying those in this deep red color in some areas you have just nothing. That's classified it couldn't figure out essentially what it was, but in some areas you've got these speckles of, yellow and so forth. So in the trend, it seems to be okay and if you add it up.

[Slide, 77]

And here we have the aggregated results for that. Now, if you're trying to count up buildings and assign vulnerability, this is the one that you're going to want to go with. If your purpose is to figure out where you could have flood impacts, that's a very different scenario. So if this is a river, you can see there's a river running through, and if you're trying to characterize the exposure, you might do a better job by having that higher resolution information and trying to you use the amount of area that intersects your flooded area estimate and not use the gridded data that would be more accurate. But then you have the problem of what's the legacy of that data? Are people going to miss understand it as you start to distribute it? So there's a fine sort of trade-off between accuracy and resolution where, if you're using sort of the same data, the higher resolution could be less accurate, but in some cases, it'd be what you anyway.

[Slide, 78]

So all of this is getting us to the key point that it's important to understand the metadata. Here, it's important to understand the source of the data, the processing steps, the vintage of the data, key contacts who you can talk to about how the data was put together. If you've got any questions, particularly if this, if data that you're using was produced five years ago on another project, and you got it and you've got a name that you can call and say: “hey, you did this work in Tanzania five years ago?”. We would like to use some elements of this. What is it that you did it's a very handy to be able to do a resolution of the data not only of the final product that was developed, but the resolution of the input data into it because you know you can take a census data set and interpolate it to a finer grid and use that for flood analysis, and you probably should, if you're trying to find out. You know what might happen in a different flood event, but passing that on people are going to mistake that a level 4 data set, for example, so trying to understand the metadata. That's why we've introduced this concept of the five different levels, for example, and understand bake right in there into the iso formatted metadata exactly what the methods were to put a given data set together so that five years later, you can take it and perhaps adjust it for inflation and population and use it again or you'll know not to use it for building specific decisions if it was put together at the global level, for example. So we have we're going to have another entire session on understanding metadata and Paul Amyx will be leading that that charge.

[Slide 79]

So, really what we've tried to do is this. The scope of this project is illuminate the process of how the data sets are put together so that that can be used in the context of decision making and acknowledge the uncertainty and really explore the uncertainty and try to at least start to answer the question-“How good is it?”

[Slide, 80]

Now validation. There's lots of things that you can do for validation and they're, all pretty time intensive to me. They all take some expert analysis. You can go through and use a street view data to do your own independent assessment of the structural types. For example, you can do your own validation, send people out in the field to double check the mapping schemes and structural classification estimates. You can do as we did and make certain assumptions and then come up with a different methods and run multiple scenarios that that's what's been done here on the right as well. These are losses associated with a single building with many different assumptions in terms of what the structural classes are to see, how much error that that actually introduces- and you can and should independently validate things like building area and replacement costs and people per household estimates that might really impact results. You can get confusion matrices and so on and so forth and calculate the error.

[Slide 81]

But I would argue that the most important thing to do for validation is to look at the data and understand where it came from. That's going to give you- because of the complexity of the analysis required to come up with these exposure data sets, and the infeasibility of coming up with an error for every part of the chain and determining what the error is for the whole process, without going through something as rigorous as some as we did for Los Angeles. Looking at the data figuring out where it works, where it doesn't, how that might impact some of the results that are coming out of it and understanding the entire process, putting together. Or maybe you have some expert that that within your unit that can do all of the deep dive and then explain to the rest of you what's happened, is more probably more feasible. But really look at the data understand what it is that you're dealing with what the appropriate and use that to determine what the appropriate decisions are. That can be made.

[Slide 82]

The limitations. This is an application. The pop grid viewer was put together specifically to understand the limitations of the data by giving you multiple views of estimates of population. Just for the purpose of understanding that uncertainty. Being able to visualize this uncertainty is going to be far more illustrative to people than having just a single number associated with error associated with the entire analysis. For example, because of the larger law of large numbers, all four of these data sets, if you ask how well how accurate they are at the country level, you're going to get the answer, probably that they're accurate as the Nepalese census. Because they probably all use the Nepal census as a starting point. How much does that vary on a cell by cell level? You know you can do the statistics and it might value by vary by as much as 200 percent, so understanding the base data. What the potential problems are that that might be associated with base data is the best way to understand the limitations of that.

[Slide 83]

And yeah…trying to step back and figure out why you're doing what you're doing and now you're going to use that answer. So, for example, if you're running a loss estimate for this region in this country, the replacement cost value itself may not mean a whole lot. But being able to understand that you have potentially 5 000 structures that would be severely impacted by a probable flood or earthquake, and that would result in displacement of X number of people- would probably be much more relevant information. So really keeping your attention on the question that you're trying to answer and how you can use these tools to answer your question rather than letting the tools, the results of the tools dictate how you think about the problem, is important.

[Slide 84]

So. How do we price disruption, because, after all, what we're really trying to do here is characterize the risk associated with various natural hazards and how we might be able to mitigate that risk and how we might be able to prepare for that risk and ultimately reduce that risk. Putting a price on it can help in terms of making decisions such as should we spend government funds or international funds to strengthen these buildings here rather than there, or should we pay to strengthen the buildings rather than ensure food security in the near to term future.

[Slide 85]

But, ultimately, just looking at the price itself doesn't always illuminate the question that you're trying to answer. So that's really what's important is to focus on what these events do to people, how it disrupts people's lives, and how can you use these risk results in order to prioritize doing something about it? Now exactly what you're planning to do about it that doesn't come from the tools itself. These tools will tell you perhaps where you should focus doing it, but you have to come up with the question. First, the tools will help that it's not a means to an end.

[Slide 86]

Okay. So here are some examples of some things that we were able to do with our EO-based building exposure after the earthquake in in Nepal very rapidly we weren't working in Nepal. This did not have to do with the level three data sets. This was essentially a level one data set that we were able to put together quite rapidly. We estimated that the fatalities ranged from 9 000 to 22 000, which my understanding is very close. We were able to estimate the number of collapsed buildings. We were able to use remote sensing data to find damage…

[Slide 87]

and use that to adjust the hazard, which at that point was not realistic. The estimate of the hazard was not realistic for the event and, ultimately, we were able to estimate where the displaced people were likely to be throughout the area, and we provided this information to UNICEF and that was used for outreach purposes.

[Slide 88]

So I would argue, these are the kind of things that can be done with these data sets effectively without even drilling down to that monetary level. Now, after, if, after that event trying to place a dollar figure for reconstruction, that is important. That's when you start to use that number, but you have to really be able to understand what those dollar figures mean in some of those areas where the construction is a little less formal. So, yes, questions that it does help answer: Is it cost effective to retrofit certain types of buildings? Regionally, where should we focus our retrofitting efforts? Are building codes, cost effective and if so, where? What might happen after 100-year flood volcano earthquake or there's just been an earthquake? What are the likely impacts and where are people affected the most and where should we deploy resources? Those are the type of questions that we can answer with exposure, a good exposure data and a proper risk process. If it's been put together, things that are unlikely to be able to answer with these type of exposure. Regional exposure, databases is a cost effective to retrofit. This specific building which buildings fell down, which homes are flooded exactly how many buildings fell down. It's all based on an estimate dependent on the rule of large numbers. We try to correct for bias as much as we can during that process. I might tell you whether it's cost effective to retrofit this building. If this building is a very typical, for example, in an area, that's high hazard and you've decided to have a program to mitigate or retrofit all of those types of buildings, but whether or not it's going to allow it's not going to allow you to make the determination in terms of what a building is without noting what type of building it is. But looking at it, if you don't have the structural type, it's not going to give you that additional information and that's not what has gone into a lot of these exposure development process. If you're, for example, if you think of that the tree of all the possible structural classifications in a given grid cell, if you sample very diligently and bring all them to the table, it's more there's more branches than there are buildings should distribute right. So when you're you have a building in an area if that building is not something that you've identified as some high-risk region, it's not it's not appropriate to assume the structure. Based on that aggregate data, you have to look at specific buildings to make all right.

[Slide 89]

So, that's my presentation. Thank! You very much. You can reach me at ckh@imagecatinc.com. If you have questions. I'd love to hear from you. Thanks very much.