MITOCW | Lec 19 | MIT 2.830J Control of Manufacturing Processes, S08

DUANE BONING: Good morning, everyone, and good evening in Singapore. Where is everyone?

AUDIENCE: I think they are on the way. We'll be here in a minute.

DUANE BONING: OK, I'm going to go ahead and start. What I wanted to do is, I sent a message around about these team guidelines, so-- our team project guidelines. So I'm not going to go into great detail through these slides. Hopefully, everybody is well on the way towards identifying a team. But I wanted to go quickly through this and then answer any questions you may have on the team projects.

Quick reminder that the team projects are about 20% of the grade. So they are an important part of this. The basic expectation is, this is where you get a chance to dive deep into the analysis and exercising of some of the kinds of tools and approaches we've been talking about with a set of data. Our hope is that you can find data yourself.

Through the various members of the class, I think, in one of the earlier problem sets, we asked for your experiences, perhaps with some past data. I think there's plenty of data around. If you are not able to tap into some past data you or colleagues have generated in the lab, in a previous job that you can share, or whatever, you can also tap into the literature. So if you find data in the literature, that is perfectly fine as well.

And then, in a few cases, we actually have some data from our own experimentation and research. So if you're really stuck, let me know or let Hayden know. And we can try to connect you up to something.

I've listed a few topic areas. I don't want these to be limiting. But I mean, these are the main kinds of topics we've been dealing with in the class, things like process diagnosis-- looking at the data to detect problems; ways of optimizing or improving the process through design of experiments, optimization, and so on; and then some other sort of advanced applications, things like nested variants, some of the spatial yield modeling, that sort of thing.

So I think it works well to have three or even four members in the team. So the goal is have a fairly rich data set, and have multiple members of the team applying different kinds of analyses to the problem, OK? So I think it does work where you're able to identify contributions from each of the team members. But the whole point of this is actually to be able to pick over the data, talk about the data, talk about your approach to analysis, and looking at the problem with your teammates.

Thursday, the project proposal is due. This is just one or two pages describing the basic data, the basic problem in your plan for attack, what kinds of analyzes you hope to do. And of course, list who your team members are and what their roles are. The goal then is, over the coming week, and maybe even starting, perhaps, even after class for folks in Singapore, on Thursday, and certainly on Tuesday, to have a chance to meet with me or Dave, if he's available, to basically go over your plan and your analysis so far, if you're partway through it.

So we'll need to do that by video conference with folks in Singapore. If we do it right after class, we can probably try to arrange this facility. Otherwise, we'll use the conference room near the SMA-- your SMA offices.

Then the expectations are two parts. There will be a group presentation. By the way, not necessarily every single member of your team needs to present part of the presentation. Trying to have 14 members in a 15-minute chunk often can be kind of a coordination mess. So as a team, you need to put together a presentation, and have one or maybe two people, most likely, presenting some of those results.

And then at the end of that week, on Friday, May 16, which is the end of the term given that we do not have a final exam, a single shared group report needs to be submitted. These are just some example topics. I'm not sure if those have been drawn from past semesters or not. Here's some additional kinds of suggestions that I mentioned-- kinds of analyses that might be of interest.

In the report-- these are some guidelines on what we'd like to see in the report. One note here, and I haven't posted it yet, is-- whoops-- is we'd like to use a common template just for the formatting and guidelines. And basically, the IEEE two-column format has worked really nicely. We have Word templates. I think we even have LaTeX templates if there's anybody who still uses LaTeX. I will post that template on the web.

What's nice about that is we can actually then put together all of the written reports from the class into almost a little booklet. And it almost looks like a little mini journal, or a little mini conference proceeding, if you will, with these reports. It also is a nice, professional looking report. So we'll distribute that as well.

So finally, I just grabbed, from the past two or three years, some of the titles from some of the process-- or team report and presentations, just to give you a little bit of a sense of the variety of kinds of things that we saw. So those might give you some idea. By the way, back in 2005, we were more focused on, like, two-person projects or one-person projects. And these were great reports. But we did decided that sharing and interacting made things a better learning experience for all.

So with that, let me open it up. Anybody have questions on how the team reports, team projects, team presentations need to go? I'll try to break-- does pretty much everybody have a team in formation? Are there people who kind of are still looking for possible partners? OK, one or two. OK, I'll try to break maybe five minutes earlier-- which may mean actually break on time-- at the end of class, so could be a little bit of discussion while people are still face-to-face on that. Any other questions on that?

AUDIENCE: Professor?

DUANE BONING: Yes.

AUDIENCE: Just now, you mentioned having a rich set of data. So from my understanding, we are supposed to do the experiments ourselves. So I would like to know how big a set of data are you expecting?

DUANE BONING: Well, my expectation, or minimal expectation I would say, is more that you might be able to tap into data that has already previously been generated. It's ambitious, and I would applaud people wanting to go and generate additional data. So we could certainly talk, and depending on the data you have in mind.

But what I mean by rich set of data is something that will support enough of the kinds of analyses that we've been talking about through the term. So if it's a DOE, it needs to be, you know, probably several parameters of input, several parameters of output. Maybe you have some additional twist to it that's kind of interesting or non-standard that included some thorny issues in the data. Maybe it's an unbalanced design, or an imperfect design, and you have to try to accommodate that.

All I mean is enough data to do what you would like to propose to do and that we can talk about finding acceptable. So were you actually looking at going in and generating some data on one of the processes, or tools, or equipment there?

AUDIENCE: My group haven't had a discussion yet. At first, we talk taught this as a requirement.

DUANE BONING: OK, yeah, so we can follow up with Hayden on that. I think bonding and that sort of thing would be great. So we won't be too ambitious. It will be doable. OK, other questions? Yeah?

AUDIENCE: I think maybe as an example of all kind of manufacturing processes, but I have a passed out data for kind of material engineering. Is that OK if I can use material engineering data?

DUANE BONING: Sure, so the question was is material engineering data, mechanical processes, mechanical properties, is that OK? And the answer is absolutely. I mean, if you look at some of these past projects, you've got everything from some things that look kind of like semiconductor stuff-- laser diode, it's not transistors. But you have other things of bending printed circuit boards, sheet metal blanket, surface quality in milling, all kinds. This is sort of general manufacturing and general processes. So that would be no problem at all. Yeah. Yeah?

AUDIENCE: Prof, can you post more of the reports on Stata so we can take a look?

DUANE BONING: Yeah, I think what we've done in the past is put up some example-- some of the prior year reports. We haven't posted those yet. But I'll try to do that today.

AUDIENCE: Great, thank you.

DUANE BONING: OK, any other questions? All right, let me switch here. So again, what we're doing now is sort of switching to some case studies. There's still, perhaps in a few of the lectures, some new material, if you will, coming along. In this case study-- the first case study-- I don't think there's actually that much new in terms of methodologies, and so on. I just think this is a nice example, though a little dated now, of a fairly straightforward but thorough application of design of experiments and response surface modeling.

So I also, by the way, posted the paper for this case study on the site as well. It's down in the reading section that goes-- readings that goes with the lecture. So you'll have to go and pop off this-- what is it-- eight or nine page paper as well. I should mention, some of the other case studies will also have papers from the literature for you to grab and read.

So this case study comes from semiconductor manufacturing. It's a little dated now. It goes back to 1991. But this was actually kind of just the time when things like statistical process control, more thorough design of experiments, response surface modeling was really finally hitting the semiconductor industry, and really starting to be more widespread in adoption. So I think this is a nice case. It's a nice case study because they actually show us the experimental conditions, the actual data, and the actual models that they generated.

Some of the things are a little dated. For example, they talk about 125 millimeter wafers. So this was 5-inch wafer processing, which is kind of a weird size. But it also-- we're up to 300 millimeter, 12-inch wafers. So it shows the march of-- progress of technology. However, CVD and CVD of blanket tungsten is still heavily used and very integral. So it's still a relevant process.

So what I want to do in looking at this case is go through, talk a little bit about the background of tungsten CVD, things that aren't in the paper but may give you a little bit of a feel on-- more of a feel on the process; talk a little bit about the preliminary work that they did looking back at the prior paper that they cite; and then we'll go through the experimental design that they performed to try to optimize this tungsten CVD process, to understand it, model it, and then optimize it.

So we'll look at the data. We'll do some RSM analysis, or mostly looking at their analysis in the paper. But assuming I have time at the end, and I think I will, I'll switch over. And we'll use-- play around with JUMP a little bit with that same data set, and show you a few of the kinds of analyses that are possible with that tool, both for regression-- some of the stepwise regression that's referred to in the paper-- and then some of the neat tools that actually in the interactive capabilities of JUMP for optimization.

Now throughout this, think a little bit about where you see weaknesses, or alternative ideas, or things like that. Because at the end, I might ask, what are some of your ideas of things you think might have been interesting, or important to try or to improve on this. So have that in mind. Where do you think they did a good job? Where do you think they did a weaker job? What might you suggest? And I'll ask you that question at the end, OK?

So let me talk a little bit about tungsten, and in particular, tungsten plugs. So tungsten, here, is generally used as a conductor in semiconductor fabrication. It's not a great conductor, meaning it's often a little bit higher resistivity than other metals, like aluminum or, now, copper. However, it continues to be heavily used for contacts and plugs.

So contacts are the holes that are drilled between-- into an insulating layer from the first level of metal down to the underlying silicon. Plugs are the same idea, but at other levels of interconnect. So you would have a plug from metal one to metal two, or metal two to metal three, and so on. Terminology is not completely uniform. But that's the typical terminology.

The examples drawn from this paper, they talk about tungsten plugs in conjunction with aluminum interconnect. And they're actually mostly focused, I think, on tungsten between-- or tungsten contacts, with the idea that in older technology, in fact, you didn't need to drill and form separate plugs between layers. You could just allow, say, an aluminum deposition to go in and fill that layer, and then etch off where you didn't want it.

With scaling, it became pretty much essential to be able to form the plugs separately from the interconnect. So then, tungsten was used to fill those plugs between metal layers, as well. As we've gone to copper interconnect, tungsten still appears in most advanced technologies, now primarily just as the first level of contact from a copper interconnect down to the silicon. So it is still used, and extensions where it may not purely be tungsten plugs, but it may also be, in some cases, some local interconnect. So you can, down very close to the silicon, form little straps or other sorts of local tungsten-based short-- very short-run interconnect to form electrical connections.

Now, the important part of tungsten-- why it's used-- is that it also can be formally filled, and can, especially with chemical vapor deposition, or plasma processing, can fill fairly small holes. So the basic idea or progression in a tungsten deposition is you primarily get conformal deposition. So in time, the gases that are present in the plasma chamber and reactor interface with the surface form reactions, and basically build up in a conformal way to fill the plug.

Now, you want to do this in a way that does not result in seams or voids in the middle. So that can be a little bit tricky. But that can be done in such a way that one avoids those plugs.

Now, the way this paper uses tungsten plugs is with a process they refer to as plasma etchback. So if I do that conformal deposition, finally, to this the surface, now I've got the tungsten not only in the plug, but I've also got all this excess tungsten out here in the neighboring regions that I want to remove. And the idea of etchback was that through that deposition process, I filled the surface, but I leave, essentially, a nearly planar surface that now, if I just etch uniformly back that whole surface, I can remove down to the underlying interface metal or interface layer, or the oxide, and leave the metal just in the plug where I wanted it.

In the paper, they refer to using this as-- using plasma etchback to do this uniform etchback. In modern processing, this step has pretty much been replaced with CMP, because that has a much higher controllability, and gets around an important limitation that's kind of pictured here in this diagram having to do with very tight constraints on uniformity with a plasma etchback.

The basic issue with uniformity is if I, in fact, have deposition that results in different thicknesses-- either in different locations within the chip or in different chips across the wafer-- and I uniformly etchback, first off, if I only etchback this amount in both regions, then I have not completely removed what I needed to in the other regions. So I have to make sure I do enough over-etch to at least clear off and make sure I don't have any shorts anywhere on the wafer. And when you do that, to make sure that I'm down to the surface in the thickest area, you get over-etch lossage in the thinnest regions.

So you have a fairly tight constraint then on what the thickness uniformity is, as well as the etch uniformity, in order to achieve that, OK? So the paper does talk a bit-- and in fact, we'll see later, explicitly models the wafer uniformity as one of the outputs in order to try to understand, as they play with some of the process conditions, how uniformity is going to be affected by their choice of these conditions.

There's a couple of other interesting aspects of plasma etchback that are important, as well, in particular, help inform the modeling that this paper undertakes. And that has to do with surface roughness. So not only will these sorts of features, this topography like that little dimple there-- but also, if the surface of the film-- the deposited film-- is rough, and I do a uniform plasma etchback, sort of a vertical plasma etchback, that surface roughness can in fact then-- this is rough. That, in truth, may be transmitted to the final surface of the plug.

So another important parameter that they will be explicitly modeling is tungsten-- the deposited tungsten roughness. And generally, less rough is better. It's going to give a better surface quality, less prone to defects, more uniform in resistance, and so on.

That, by the way, is another reason for switching to CMP. CMP planarizes-- can deal with not only this kind of surface topography on a large scale, but it also smooths the surface on sort of the atomic scale. And so you tend to get much, much higher quality final plugs. OK, any questions on basic tungsten-- a little lesson on tungsten plugs.

So let's talk a little bit about the process equipment and the process parameters that are going to be in this case study. This is a schematic cross section of an Applied Materials Precision 5000 reactor configured for tungsten CVD. By the way, this is a classic tool. The P5000-- AMaP P5000, there's one of these in the Smithsonian Institute. It was the tool that made Applied Materials what it is today. It was one of the first single-wafer processing tools that really gave good control over the atmospheric conditions for deposition, for etching, these sorts of things. So we're seeing a glimpse back into history here even looking at this cross section.

By the way P5000's are still in use all over the place. We have some-- I guess just one in MTL, I think mostly for etching. But even in some FABs I still run into them. They've been scaled up, and improved, and configured-- changed in configuration, but they are still around.

So the basic idea here is that the wafer will sit here on a susceptor in this plasma reactor-- this low pressure plasma reactor. In fact, there are usually pins-- I think three or four pins-- that hold the wafer. The robot loads the wafer in onto these pins. And then the pins-- these lift fingers here, I guess-- recede. And the wafer then sits on the susceptor.

This susceptor is optically heated. So there are these heat lamps. I think there's something like 1,000 watt heat lamps in this configuration that heat up and hold the temperature of the whole platen, the whole susceptor. There's then gas flow coming in from a gas feed. There's a little mixing chamber in here so that you can mix different combinations of different gas flows. And then it goes through a shower head.

This thing right here is typically referred to as a shower head. And you can imagine, it's basically like a shower head in your shower. It's got holes in it that basically allow the gases to flow through, hopefully in a fairly uniform distribution across the surface of the wafer. And then there's a vacuum control, and a number of different vacuum outlets for the gases to flow.

So some of the key inputs-- there's lots and lots of knobs to control on these equipment. Some of the key inputs that are used in this study-- in the paper that we're looking at-- are gap space temperature, H2 pressure, and WF6 pressure. So the gap space refers to, I believe, essentially the height of this susceptor. One can play with the distance between the gas inflow and the wafer, moving it between, say, 2 to 4 or 5 inches. And that changes a number of the plasma characteristics. It changes perhaps some of the gas flow characteristics. It may also change some of the thermal characteristics. So it is a control parameter that is not entirely obvious up front how it affects the growth of-- or deposition of this tungsten.

So that's going to be one of the parameters that needs to be empirically studied. Certainly, temperature is going to have a very strong effect in this process on things like growth rate, perhaps also the surface morphology or the roughness of the film, the amount of incorporation of tungsten, and the ultimate resistivity of the film-- those sorts of parameters. And then, both hydrogen and tungsten hexafluoride gas flow pressures will also be clearly of influence in rate resistivity and so on.

So this is kind of a neat paper, because they actually are measuring about eight different parameters and modeling each of them. So it's kind of a four input, eight output kind of DOE. So just to go through some of these parameters, many of them are pretty obvious, but a couple of them I want to explain a little bit more so that you have a good feel for what they are. Deposition rate, I think that's pretty obvious. Generally, you would like to have a fairly high rate. It's going to make the process time as short as possible, efficiency in the tool use, and so on. The paper mentions wanting to have about 500 nanometers a minute-type rates with these processes.

The resistivity value-- or rho here-- generally, you kind of want a low resistance, or a low resistivity film. An interesting twist here is, depending on the application, the actual resistivity you want and whether it's all that low actually changes. So for example, if you're using it as a short run metal wire, pretty much all the time, low resistance is important.

But it turns out, interestingly, when you use tungsten plugs in conjunction with aluminum layers, you may actually want the tungsten plug to be a little bit higher resistance than the aluminum itself, so that, because it's a very small, constrained spatial area, the higher resistance causes the current flow to spread out a little bit more uniformly, and actually improves the reliability of these tungsten plugs. So you'd actually like to understand what the resistance is and be able to tune that in. So that's going to motivate some of the modeling that that's done here. It's not always purely a lower resistance is better.

RS uniformity refers to sheet resistance uniformity. This is the ohms per square of sheet resistance, but then a uniformity number calculated for that measured sheet resistance across the entire wafer. And so you want a very good uniformity.

The 3% non-uniformity number is the deviation-- percentage deviation of the thickness. Sort of one standard deviation over mean is what they're referring to in here, I believe, for this uniformity. And again, one of the motivations there is that uniformity in the plasma etchback, you want to avoid recess of these plugs in some region which can make it difficult for the subsequent aluminum to reach down in and actually contact those recessed plugs.

Film stress is another important parameter. Let me see if I've got other slides on that. No, not much. Essentially, there is a volumetric mismatch to the deposited tungsten compared to the underlying silicon substrate wafer. And in fact, what you'd typically find is that the wafer will undergo-- or be in tensile stress. So basically, there's more volume in the deposited film. And it's causing a thick-- a layer, depending on the thickness of the layer-- in fact, in the design, they're talking about two different thicknesses that they explore-- that causes the wafer to bow.

So I've got a film that wants to spread out a little bit more, especially at lower temperature compared to the temperature of processing. And that wafer bow is a nice indicator of intrinsic internal stress in those films. In general, you don't want stress mismatches between different layers on your wafer, because that means there's sort of this excess energy around that would like to release itself. And one way it releases itself is by having those layers delaminate, OK?

So a reliability concern is to try to avoid very high stresses in order to support subsequent processing and long-term reliability of the material on the wafer. So they are making some measurements of film stress. In fact, the way they make this measurement, I believe-- if I remember right-- is they measure the wafer bow. And then, based on assumed numbers for the Young's modulus and perhaps the thickness-- they may measure the thickness separately-- they basically calculate or estimate what the film stress is. Yeah?

AUDIENCE: And that's originally CTE mismatch?

DUANE BONING: I believe that it is CTE mismatch, that in some sense, at whatever temperature you're doing, if you could deposit the film and keep the wafer at 300 degrees, or 100 and whatever it is, at that point in time, I don't believe the wafer undergoes or experiences much stress during this deposition. But then, it's a temperature expansion difference.

I'm not entirely sure of that. It's possible that the simple incorporation at the time of deposition also has some stress. But really, we're talking about intrinsic stress that-- or extrinsic stress that remains at the end of processing. You know any more on that, Hayden?

AUDIENCE: Sounds reasonable. I mean, there are situations where you're depositing an epitaxial layer of, say, a semiconductor, that has a very clearly defined lattice, whose picture is mismatched with the substrate underneath. In that case, there's a-- I guess a definite built-in stress that's going to result in that deposition but isn't connected with thermal contraction. But I would have thought, in this situation, it's mostly to do with the presence of impurity is changing the CTE of the deposited layer that stretched.

DUANE BONING: Good. The fifth parameter that they measure and model is step coverage. And so here, that refers to essentially how good the deposition process is-- I know you've read here-- in filling these features. Now, if all of your features are exactly the same size, then that's a little bit easier. So if all you're doing is filling contacts of exactly the same size, then you've got a nice, well-defined number for step coverage. And that's really what they're doing in this application.

But I should mention that these kind of conformal depositions can actually be a very strong challenge for dealing with patterns or features of different sizes. So you can imagine, for example, having some contacts here, having some structures that might in fact be lines, and then having some other structures that are in fact very wide lines. And now, if I can formally fill, I might be able to fill the narrow structures, but the wider ones will not necessarily fill completely. I won't have excellent step coverage.

Now, I believe what they're doing in this paper is really just looking locally after the deposition, and looking to see sort of what percentage of volume is missing in that little indentation resulting from the conformal deposition, and looking for fairly high-- near 100%-- step coverage, so that there's not a missing percentage of the feature compared to, say, one of these large structures that might have a very high missing volume. Generally, they're able to achieve fairly high step coverages. So that's not one of the critical output parameters in this process.

Now, also shown on this picture, and maybe a little bit hard for you to see-- it's also in the paper. But on the right, you see a top down photograph, or SEM micrograph, of the surface structure. And it gives you a sense, a little bit, of two things. One is there's a lot of surface morphology going on, surface roughness. But the other basic thing that happens is, as you're depositing these materials, they are lining up in sort of a polycrystalline kind of way. You do get some local self-organization.

And you can see that also over here on the left side. If you kind of see, there's little crystals, or little threads. And in here, you can see a little bit of the polycrystalline structure. And that's actually one of the things that gives that surface roughness. It's not surface roughness on the atom by atom level. It's really these preferential growth directions in some of these grains, if you will, as they're growing.

Only two or three more parameters here. Another one-- and this is kind of neat, I think, because it's a nice parameter that contrasts with the other ones. The other ones are-- let's see. Most of these-- things like step coverage, film stress, uniformity, resistance uniformity-- those are parameters of the final wafer or the intermediate wafer state that you want to achieve. Deposition rate and WF6 conversion are both more manufacturing parameters. In some sense, they don't really matter to the final wafer product. But they can matter a great deal for the cost efficiency or the time efficiency of the process.

So WF6 conversion is what percentage of the incoming gas is actually resulting in the deposited film. So you might have only 1% of the tungsten going onto the wafer and 99% going out the exhaust pipe. That's terrible waste. It may have some environmental implications, but it also has a lot of cost implications. Because tungsten hexafluoride is not a cheap gas-- very, very highly purified. So I like this paper in terms of they're actually trying to track, and keep track of, and model the WF6 conversion rate, and try to optimize that parameter also.

Then finally, a couple more parameters here. Reflectance is basically an optical reflectance, percentage of light reflected kind of measurement of the surface topography. So that's how they're trying to estimate or correlate to the roughness of the surface film. And then finally, another manufacturing parameter is reproducibility, not just wafer uniformity one wafer at a time, but wafer to wafer repeatability or reproducibility of the process.

So they are actually making, in this case, I think, the assumption that pretty much all of these processes are more or less similarly repeatable. But they want to estimate how repeatable the process is. And we'll see that in their DOE. They're going to basically do some replicate at one of the design points, and use that to estimate reproducibility. So they're not going in and trying to estimate where the noise or the variance is minimum across their operating space. They're actually assuming that that's similar, OK?

So here's what they do in the paper. Their goals are to build response surface models-- just classic response surface models. And then, as I said, the reproducibility is really focused just on the center point designs and applying that throughout the entire space.

And then, they want to use these response surface models. They want to qualitatively understand some of the trends-- very good engineering use, trying to understand from an engineering point. And then, they also want to use these models in an optimization kind of loop to try to drive towards satisfaction of multiple limits or goals on those-- what was it-- seven or eight output parameters.

Now, the paper does refer to some prior work on screening experiments. In fact, these Precision 5000 tools and most FAB tools probably have, literally in some cases, dozens of control settings, some subset of which, maybe only 15, are easily settable electronically. And another 15 might be more mechanical, or other more permanent adjustments. But there's a huge space of possible parameters.

The paper just refers to prior screening experiments to narrow down to these four parameters. So I don't want you to think that they knew automatically, up front, what four parameters they wanted to play with. There is a reference if you want to go try to track it down for their prior work. It's actually in some obscure conference thing, so it's hard to find. So the details exactly of what they did in their screening are not entirely clear.

But the point is, they probably did something like a fractional factorial-- a highly fractionalized factorial kind of design to explore the space of just trying to see what parameters are significant, what influence-- what input parameters influence the outputs that they were of concern, and then narrow that down to just these four parameters that they wanted to do a second set of experiments-- which is what's in this paper-- that are focused on response surface modeling, OK?

There's another interesting, subtle point that pops up in the description of the paper. Some of these parameters, things like sheet resistance-- sheet resistance is resistivity-- is it divided by per square-- the pure resistivity divided by the thickness of the film, so that the thickness of the deposited film will change the RS measure on the wafer. What that means is, to do a fair comparison just of truly the resistivity of the film, or to estimate that, you actually want to grow all of your films to about the same thickness.

Well, if the deposition rate is a parameter that you don't necessarily know, and you're trying to model, how do you do that? So in other words, if I run for a fixed period of time-- maybe I run for 60 seconds on each of these design points-- but I have deposition rates that are different by a factor of 2, 5, 10, I'm going to get widely varying thicknesses. And therefore, when I make some of these simple sheet resistance measurements, that's going to be very, very tricky.

Furthermore, if my thickness is varying it makes a basis for comparison of many of these other parameters really hard. How about film stress? Well, is it a film stress that I'm seeing because of the thickness difference, or is it truly intrinsic film stress in the layer? So what they've done is been-- and I think this is very cool-- they were very careful. And they say that they are doing this to where all films are grown to about 1 micron thickness in this DOE.

What's sneaky is, I think to actually do this-- which is also good, because that is really the thickness that's close to their target application. What's a little bit sneaky is to do that, I think they basically had to do prior runs, maybe at each individual DOE point, to do the run, measure the thickness, estimate the growth rate, and then adjust the time for the real DOE run on each of the design points. So I think that there's a whole extra duplicate set of experiments that they're not talking about that were done in order to be able to normalize and get a very nice set of experiments in here that they could talk about and be able to use to make the right kinds of optimization and judgment.

So it's kind of an interesting lesson. I think there may often be additional experimentation around your main DOE that you have to do in order for your main DOE to really be a good, clean DOE. Now, it's possible that maybe their screening experiment gave them some information to be able to do a rough estimate-- a good enough estimate of deposition rate. So it's possible they were able to get that benefit from their prior screening DOE.

OK, here's their experimental design. What they're going to do is a central composite with five levels at each design point, which is pretty cool. I think that's better than the three levels. What's one reason that you might like five levels rather than three, especially if you want to build, say, quadratic response surface models?

We haven't explicitly done this before, but for example, on quiz two, you were going from a linear model to a quadratic model and asking questions about lack of fit. In order to say whether a linear model had lack of fit, I needed a third design point so that I could say, well, does that third design point give me indication that the linear model is not sufficient? If I do three levels and build a quadratic model, and I want to ask the question, is a quadratic model sufficient, or is there a lack of fit in the quadratic model, can I do that with just three data points?

No, right? You can fit a quadratic to three data points. So you won't detect if there's some other structure going on. So in order to be able to do a lack-of-fit analysis up to seeing if a quadratic model exhibits, then you need at least additional levels implicitly somewhere in the design. So that's kind of a neat thing that they do here.

So what they've done is normalized all of their factors. There are four factors here-- space, temperature, and so on. Again, not just two-- in their case, not just two of them, minus 1 plus 1. But they use, with their five levels, a minus 2 to plus 2 normalization. But essentially. They're doing up to five levels on each of their parameters. And so they're using coded variables that we'll see in the later tables just like we're used to.

Now, of course, the other benefit of doing a screening experiment is that also helps give you some information about reasonable design space for picking the range that you want to explore these parameters in. Otherwise, they probably would have had to just go by prior experience. But that's another advantage of doing sort of this two-stage experimentation with screening followed by a more extensive-- either a new DOE, or adding to the screening DOE.

So they're going to be building second order polynomial models, that is to say quadratic models with both interactions and then square terms-- just square terms in individual parameters. And they also mention that they're not going to do any transformations of the output variables, such as a log or an inverse. OK, so we've already talked, I think, about those two points.

So in the paper, they show the variable test level is coded. Here's, again, our four inputs and the trials. And if you look, you can start to get a very nice sense of what this central composite looks like. What they're essentially doing here-- breaking it up into, let's call it, three different segments of design points. What are these design points? What do those correspond to?

There's a term. We refer to those kind of design points as? Where does one of those design points lie? Take trial number one, right? It's way out, minus 2 on the space axis and holding the others at their median or their nominal value, right? So that's out on the end of one of the axes. Then the plus 2 minus 2, that's the other end of this.

So those are our axial points in a central composite design. They're sort of the classic hold everything else constant, vary one parameter at a time. So they are giving you unambiguous sense of what that one parameter is doing, but not able to capture interactions, and so on. They're also exploring sort of the outer limits of the design. How about these points?

How many of them are there? Looks like there are 16 of them. Corner points, exactly. So those are all corner combinations of four parameters-- 4 squared, 16 different corners. And then finally, these are easy. What are these? Center points, right. And so I don't know whether they have seven replicates.

So the other points are unreplicated, by the way. So you might be a little bit susceptible to outliers in some of the other points. But they have a fairly reasonable number-- seven replicates at the center. So that'll allow some more or less reasonable estimate of some of the reproducibility-- some of the inherent variances in the process. They say, by the way, that those replicates generally result in about 1 1/2% to 5% 1 sigma range in the output parameters.

The paper says they did a randomized run order. Why would you do that? Why not just run-- you've got trial 1, et cetera, et cetera. Why not just run all the way down to trial number 31 in this order?

AUDIENCE: To be independent. To be independent

DUANE BONING: To be independent. Independent of what?

AUDIENCE: Different pressure, different pressure.

DUANE BONING: I think you're right. But as you changes parameters, we're only playing with these four parameters. But there might still be noise-- other noise in the process, or drifts like we talked about. And it turns out, in many of these deposition processes, you actually have to worry about possible drifts, in that you may get film buildup on the inside of the reactor that, over time, might change a little bit the deposition process. So by randomizing the run order, you're randomizing the output, and converting-- if there is a trend, you basically convert that systematic trend into noise.

It may still perturb your values, but now, it does it kind of distributed in a random fashion across all of your output. And it will add to the noise-- change the reproducibility values. But at least then, it's not fooling you in a systematic way. If there is that kind of other drift, it will not confuse you or alias with some other parameter.

I have one minor annoyance in the paper. And so if you're writing papers and you do a randomized run, tell me what the run order actually was. If you're telling me the data, I would love to actually know the run order. Because then, when I do residual analysis and look at the residuals as a function of time, I can actually go in and see if there was perhaps a systematic time drift. And it really doesn't cost much to do that in ink space in the paper.

So for example, here they had trial. Some papers, you will actually see, they go in and might put a number sort of right beside it in parentheses or something like that, or another column in the table or whatever, that tells me run order, not just trial number. So do that if you write papers with data in it.

They also say they did some outlier analysis. They don't really say what that was. But they do note it in the data. And by the way-- oh, I may not have put this on the site yet. I will also add the spreadsheet with this basic data. So if you want to grab this later and play with the data, and do some of your own response surface model fits with JUMP or some of the other tools, you can do that.

Here's the output-- trial number, and then each of the seven different outputs-- so these are the seven different outputs-- and basically just the raw, measured data. They also may have asterisked or footnoted a few of these, which is a good practice. They said, you know, weird things like "indicates entries determined to be outliers." Or if they knew something weird happened with the measurement of that particular data point or that entire row, they would note that. That can be really critical when you're interpreting the data.

And then, what they've done down here at the bottom is simply calculated things like the standard deviation across all of those. I think that's across just the replicate runs. You can see that underneath here. Yeah, OK, so it gives you a sense of the percentage variance-- 1 sigma percentage variance in each of the output parameters. Yes, question?

AUDIENCE: One of the entries has an issue. Doesn't that mean the whole row has an issue? So shouldn't that whole row be discarded. How can you, for example in round six, have an issue in, let's say, 8.1, and say this looks weird? And they're not going to use it in the row.

DUANE BONING: Yeah, so the question for folks in Singapore is, how can you have just an issue with one element and not an issue with the whole row? Basically, I think it comes down to some engineering knowledge on what the issue might actually be. So in particular, if you're worried that there's something wrong with the measurement, then that does not necessarily contaminate your measurements of other columns.

If you think it's really an indicator that the whole run, or the whole trial, was messed up somehow, then I agree. You would probably want to omit that entire run from your whole table, or maybe repeat that run. In this case, with single-wafer processing, it's probably an easy thing to do.

AUDIENCE: I guess the second. If we were experiments that we heard that's an issue. One of the numbers looks funky. The fact-- and we want to eliminate that whole row. What does that mean for our experiment if there's, for example, like, we only-- we didn't replicate the center point, and now we don't have a center point.

DUANE BONING: Yeah, so the other question is if you do have a bad data row, how does that mess up your data analysis? And the answer is depending on the kind of DOE that you have, and the kinds of analysis that you apply, you can be more or less robust to those kinds of missing data points. One of the nice things about a central composite design, and especially one where they've got kind of these five levels, is if you do that in conjunction with response surface modeling-- regression modeling in particular, the numerical regression modeling-- you're fairly well off if you have a few missing data points.

It would throw you for a loop and some of the shorthand formation of contrasts that we talked about. Because then, your typical averaging that's implicit in some of those are wrong. So you've got to be a little-- but if you do regression modeling in conjunction with a fairly rich-- sort of extra data points are lurking in there, extra degrees of freedom are lurking in there, that'd still allow you to reliably fit models.

So the paper says the ANOVA is performed but it is not shown. I want to see the ANOVA table! So, well, we can basically generate those on our own. And so in this case, if the paper has the raw data in it, I think it's OK that they didn't show the ANOVA, because I can regenerate that.

Papers I hate are they don't show you the data and they don't show you the ANOVA, and so you don't know really what the noise was, you don't know significances of parameters, you don't know which parameters-- often, they won't tell you, really, the significance level that was used to accept a model, all of those sorts of things. So in this case, I'm happy with the paper, because I've got the raw data.

They claim that each of the model is significant at 99.9% significance level. That's just their p value on the F test for the whole model. So these are very-- at least there is some parameter that is significant and influencing the output. They report also r squared, which is nice. And for most of the models, they are really quite high-- in the 0.88 0.97 range. The lowest one was resistivity at about 0.79.

They also report some lack of fit. I mentioned a little bit of evidence of lack of fit for uniformity and reflectance. This lack of fit for sheet resistance uniformity is really interesting. And in fact, so interesting that I think, a week from today-- if Dave is talking on Thursday-- I will spend a whole lecture on modeling of uniformity response surface modeling approaches, spatial modeling approaches appropriate for uniformity.

The preview is, think of the calculations that go into a wafer level uniformity-- you know, the calculation of that. You might measure-- I think they're measuring 49 points across the wafer. They're forming sums of squared deviations and then taking a square root, and then perhaps normalizing by some mean. Does that sound like a set of linear operations?

No, there's a lot of implicit non-linearity lurking just in the calculation of that metric. And so it's not surprising that there's quadratic lack of fit, or inability of a quadratic model to deal with a uniformity metric-- a spatial uniformity metric. And we'll talk more about that next week.

But, so they do mention this at least. They also mention some potential lack of fit in reflectance. I don't have an explanation for that. They conjecture it's due to a very small pure error term, so that you're able to detect, with pretty fine resolution, if there's any subtle non-quadratic trends. But they're not concerned with it. They still feel that they have very good models.

It's also the case that there may be transformations that could be applied to improve these models. And that also, for this sheet resistivity, or sheet resistance measurement, there might be natural transformations that you might think that apply. In fact, if this is related to things like resistance divided by a thickness, and the thickness grows linearly with-- maybe you have a linear dependence on growth rate or something like that, there may be inherent in there some one over kinds of transformations that might actually be a better thing to do, to take that sheet resistance, invert it, and model the inverse of sheet resistance as a function of the input parameters.

So you could play around with that. And actually, that's a fun thing to play around with some of these models. They didn't do that, but you could. And then they do show the regression models.

Here's the output. These are the regression coefficients-- all of the linear terms, and the cross terms, and the square terms. And what they've done in this table is report the fit for just the significant terms. If the coefficient was determined in their stepwise regression to be insignificant, they've left it off. I don't think they actually told us what their criteria for inclusion-- what their significance level cut off was in here. So that would be a good thing to do as well.

But you can see, lots of models, lots of second order terms. These are interactions-- b1,1; b,2,2; b3,3; b-- those are our pure square terms. And then these are interactions. So there's a fair amount of square activity going on as well as interactions.

They also then use those fit models to start to form contour plots. And this is how they start to make some engineering judgment about different trends. So they would plot-- realize, they've got four inputs and seven outputs. You can't possibly have a-- whatever-- four-dimensional dimensional input space is very hard to visualize. So what you resort to are typically sort of two-dimensional plots where you might plot one output or two outputs as a contour.

So in this particular plot, the solid line, I believe, is growth rate. And the dashed line here is WF6 conversion. And you can start to get a feel for as, for example, the WF6 pressure and the temperature change what trends you get in those two parameters and what trade offs you get in those two parameters. So things like as the temperature increases-- so temperature increasing this way-- the growth rate kind of increases. So the growth rate is the solid line. So pretty directly with temperature, that looks like a fairly strong direct impact, as you might expect with a chemically dependent kind of process.

So they can start to get a feel for what some of these trends are, and look at what, for example, has the biggest impact on growth rate of these parameters, and which have either insignificant or small in magnitude kinds of effects. Now, they can also start to do things like overlay requirements or conditions on these to start to understand the feasible operating space for process design and optimization. So for example, if your growth rate had to be 500 nanometers per minute or higher, you might then have-- kind of shading on this plot, you would say, OK, I really need to be above that 500 nanometer per minute contour. And that tells me where I need to be, at least with respect to those two parameters in terms of feasible operating space.

So already, it tells you a lot just looking at these plots of where you might need to be in order to achieve your multiple goals. It may not give you a numerical optimum point. But all by itself, lots of looking at the data and plotting of the data is really valuable. And they do that for not only growth rate in WF6, but things like resistivity, reflectance. They start to look and see is the overall design going to be feasible, what has the biggest effect on morphology, those sorts of things.

We found if you look at their contour plots, most of these contour plots are reasonably-- these are nice, semilinear kinds of plots. Their uniformity plot, these look bizarre-- very complex surfaces. And again, that's a reflection that this uniformity metric, the spatial uniformity metric, is a very complex function of the input parameters.

What they then do, finally, to round out the paper, is talk a little bit about optimization. So what they will have is, some of those limits-- like I said, for example, on the growth rate, I guess it wasn't a 500 nanometer cut off. They needed at least 300 nanometer. But they would like to get that-- higher is better. And they would basically have different lower or upper bounds on each of the other parameters, as well as desires to sort of hit either targets or make those bigger or smaller.

Now, they don't really describe very much what their optimization procedure is. All they do is report that using the model, they were able to achieve the following best condition, which they report and are very satisfied with-- so things like very high, 99% step coverage, improved or fairly good conversion rate, and so on. So that's actually one of the fun things one could play with, I think, with the actual data, is apply some of the hill climbing or other kinds of optimization algorithms to this data.

They do note in the paper that some of these constraints or limits do not greatly reduce their factor space. They do not constrain where they need to operate. And that gives them lots of latitude to be able to-- that's a good thing. It gives them latitude to play with those parameters to achieve perhaps some other goals.

OK, so what I want to do in the last six minutes is switch over here. So are we ready to do a switch? Because what I want to do is--

AUDIENCE: Why don't you drop a pin there?

DUANE BONING: OK. OK, what's that number? So what I'm going to be doing here is basically showing you some JUMP analysis of this data. Oh, dear. Maybe I'm not online.

So what I've basically done is just sucked in the raw data and their input conditions and their output conditions. And we'll play with two things. I just want to give you a little bit of a glimpse here of--

AUDIENCE: We can't hear.

DUANE BONING: OK, excellent. OK, can you see that in Singapore? Excellent. All right, I'm psyched. OK, so what I've done here is use JUMP-- JUMP 7 in fact, which is a more recent version than I'm used to. So some of the controls are a little bit different.

By the way, I don't know if you played with JUMP before. As you're looking at your projects, there is a 30-day free trial license of the full version of JUMP. So you can download the trial version if you want-- if you want to grab it and play with it, especially for the project. I believe there's also an academic version that's, like, $50 for a year license or something, which is pretty nice. Because the commercial version is hundreds of dollars. So it's a nice package.

So all I did is I sucked in the inputs. These are our four input conditions here, and then our growth rate and output conditions. And I would always encourage you, as a first step, to explore your data. Just look at it. See if it makes sense. Do univariate plot of outputs as a function of inputs.

And by the way, there's some very cool things like scatterplot 3D, where you might take three of these parameters and plot them against each other. And so you get cool things like this that give you a picture of the scatterplot as a function of those three inputs. Or I guess this is a scatterplot of three outputs versus each other. See if there's some correlation between these outputs. You can do it one output as a function of two others, and so on.

So one of the nice things I like about JUMP-- and you should look for it in a good interactive package-- is that interaction. You can actually interact with the data, one of the things that Excel is not good at. So other tools like MATLAB are good at this as well.

OK, so having done that, another thing I like about JUMP is it allows you to highlight what are bad data points. So I don't-- I can deal with missing data points in JUMP. So I've pulled those from the paper as well.

And then I can start to do things like build regression models. So there are graph-- whoops, I'm sorry-- analyze fit model. So here, I might say, OK, I want to build a model for-- I'll just pick one for the moment-- growth rate as a function of these four input parameters. And I can do things like, say, I really am doing a response surface model up to polynomial degree 2. And it builds for me the template, if you will, for all of the interaction terms, the square terms, all of those things that if you were doing this in Excel, you would have to yourself build those fictitious input parameters. But here, I'm basically telling it directly what the structure of the model I want to fit is.

And one thing that we haven't talked a lot about that I wanted to highlight is how to go about building the model. I could go ahead and do a full ANOVA-- a full regression in ANOVA-- and look at each of the terms and decide what's significant and what's not. There is a stepwise version of that same process where, essentially, what we do is, one parameter at a time, add--

AUDIENCE: Hi, Prof, sorry. Yes We just lost the connection. We can't see a screen anymore.

DUANE BONING: I wonder if I lost my-- can you see it just-- maybe we'll just transmit. I'll try-- it's not accepting the net meeting.

AUDIENCE: I can put it through the Codex so they can see it one their screen.

DUANE BONING: Yeah, it won't look as good. But let's at least do that since we only have a couple more minutes.

AUDIENCE: I can't do anything about your net component.

DUANE BONING: Yeah, that's fine. So we're just going to project so you can see kind of crudely on the screen. But the basic idea of this stepwise regression is that you can set the significance level for accepting a model parameter. And then, one step at a time-- so for example, if I do this step, it's pulling in additional coefficients that are of acceptable significance. And in fact, I think the probability, to enter, it needs to be set to 0.25. This is 75% confidence level needs to be achieved in order to accept the model.

And I can keep stepping through. And it iteratively is basically making determinations based on my setting of significance values to pull in or discard model terms. And if it discards a model term, then it takes those degrees of freedom, throws them back into the error term, not purely as a function of the replicate error, but it does this kind of stepwise multiple regression in order to build the overall model.

So here you can see, for example, for this particular output of growth rate, you can see which terms end up being significant. You've got things like temperature squared, and temperature H2 pressure interactions. OK, doing that, you can then accept one of these models and build another column in JUMP that is, in fact, the prediction formula for one of these parameters. So for example, this might be the prediction formula for growth rate. It now has that polynomial model in it.

And one of the cool things about JUMP, just so that it's there and then you can play with it, are you can also now start to go in and do things like profiler or contour profiler for optimization and exploration of that design space. So doing a simple one-parameter profiler, I might say, OK, I want to profile growth rate and row. I'll just do those two parameters. And what I've got is this prediction profiler that has the two outputs. Growth rate is the first row, and-- whatever it was. Was it-- I can't remember what the second one was. It's a little hard to see there-- formula row. OK, so the resistance and the growth rate as a function of each of the individual parameters.

So this is the univariate dependence depending on my current operating point. And then I can move that operating point and see how the other parameters change. So as I change temperature, how does that affect my outputs? OK, so it's kind of fun. You can play around and see how those things happen.

You can also do, essentially, close to what they did in the paper, and do contour plot explorations. So I might, for example, do those same parameters-- yeah, that's fine-- and actually, now, plot the current operating point, the response of growth rate, row, resistance, uniformity in the design space so that I can see how those things change as I change my gap spacing or my temperature. And I can also apply things like the limits. Maybe I have a low limit of 500 for the growth rate. I need, again, 500 nanometers. And what it does, it'll start to shade out unallowable regions to achieve those kinds of goals.

So this is just a nice, interactive tool that lets you do that kind of dynamic exploration of the optimal space. So I'll put that data up on the web as well. And you can play around with it. But this is kind of where I'll leave it for you to explore more if you want to with JUMP, either on this data, or perhaps more efficiently, on your team project data. But wanted to get this case study so some of the capabilities.

So I think with that, I'm going to break. And we can chat a little bit if people have questions about team projects and whatnot.