COGNITIVE COMPUTING FOR BIG DATA

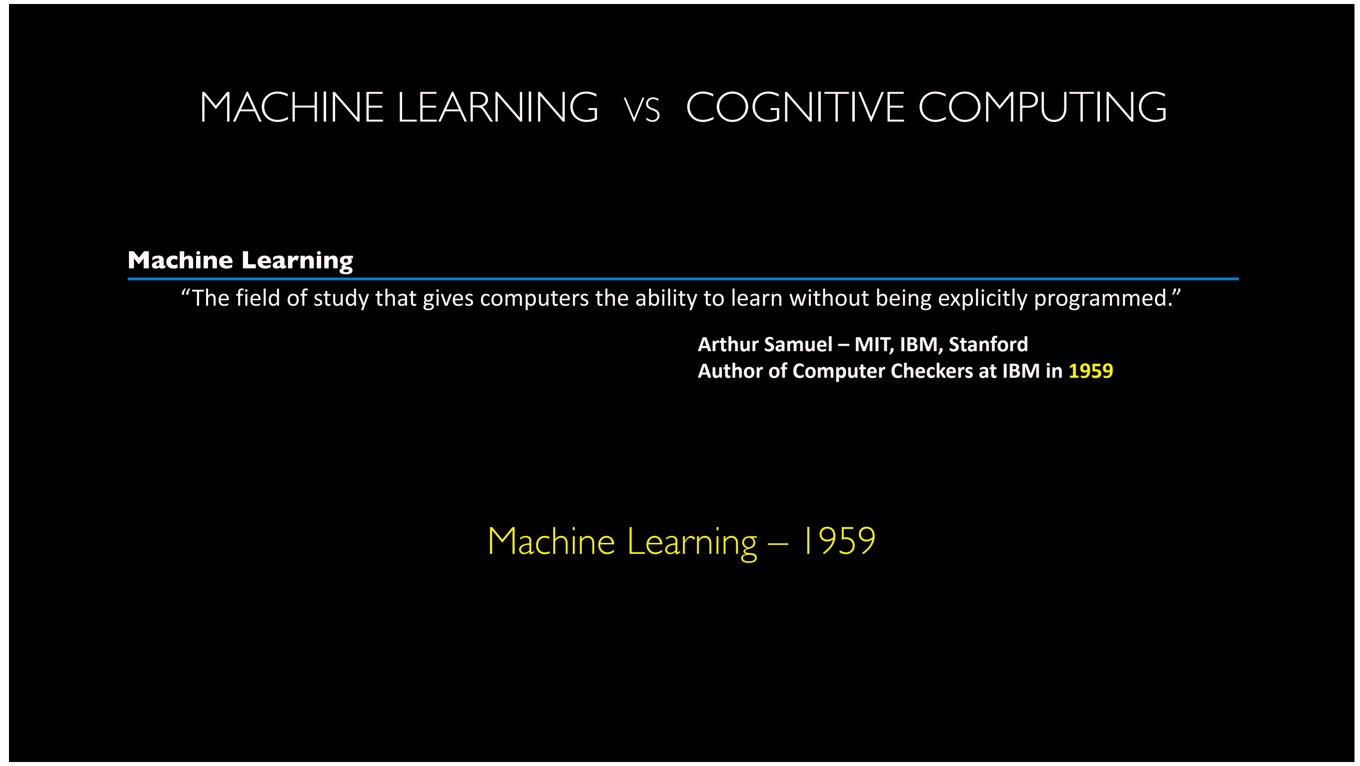
SCIANTA ANALYTICS
DEEP INSIGHT™

Thanks for inviting me to give this talk! It's great to be here. This should take 30m, depending on questions.

Full disclosure, I'm part of a company that does this cognitive computing stuff.

If you'd like to exchange money for software, let's discuss that later.

Machine Learning, Cognitive Computing, and Artificial Intelligence are heavily overloaded terms, and often treated as if they could be a "silver bullet" remedy to any problem in Big Data. We're going to dig into those definitions, and what they can and can't do.



I'm not really going to talk about machine learning. Well, just a little. What is ML?

Train a system with a bunch of high quality labeled data, it finds the statistical outliers.

This is great if you assume that outliers are bad, but you know what they say about assumptions.

Still, ML techniques are widely available now, and they're pretty useful.

"The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans."

Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013

SCIANTA ANALYTICS
DEEP INSIGHT™

What's the next step? How can we make systems that are a little more useful?

Cognitive Computing is a natural evolution of Machine Learning, which comes from applying the wealth of ML techniques recursively to themselves. Just as a person can observe reactions to their behavior and learn to moderate that behavior.

"Cognitive computing refers to systems that learn at scale, reason with purpose and interact with humans naturally. Rather than being explicitly programmed, they learn and reason from their interactions with us and from their experiences with their environment."

Dr. John E. Kelly III
SVP, IBM Research and Solutions Portfolio
October 2015

SCIANTA ANALYTICS
DEEP INSIGHT™

IBM's Watson is an example of Cognitive Computing. They're going really big.

Tackling massive problem domains, and speech recognition, and NLP at the same time.

We're here to talk about Splunk, so that's a more constrained problem space.

Just typing SPL into a search bar instead of speaking English or German into a microphone makes the problem easier.

ARTIFICIAL INTELLIGENCE vs COGNITIVE COMPUTING

FUNDAMENTALLY THE TWO ARE QUITE SIMILAR
THE DIFFERENCE IS

INTENT

ARTIFICIAL INTELLIGENCE
Systems Make Intelligent Decisions for Humans

COGNITIVE COMPUTING
Systems Give Humans Insight to Make Intelligent Decisions

SCIANTA ANALYTICS
DEEP INSIGHT™

In fact, when you talk about a machine learning and reasoning like a person, that sounds a lot like Artificial Intelligence.

I do want to be clear that we're not trying to build a depressed elevator or a murderous spaceship.

“What it’s really about is involvement of a human in the loop… Cognitive Computing is ‘augmented intelligence’ rather than ‘artificial intelligence.’”

Rob High
CTO, IBM Watson
June 2017

SCIANTA ANALYTICS
DEEP INSIGHT™

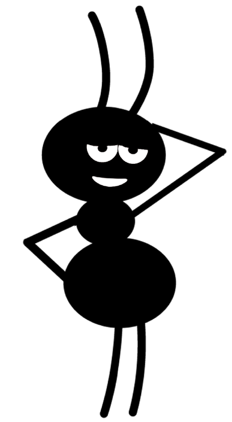
We’re thinking more along the lines of supporting tools for already intelligent Data Scientists.

There’s a lot of brain power on this planet, we think we can help focus that.

We’re building navigators, more than autopilots.

MACHINE ASSISTED UNDERSTANDING

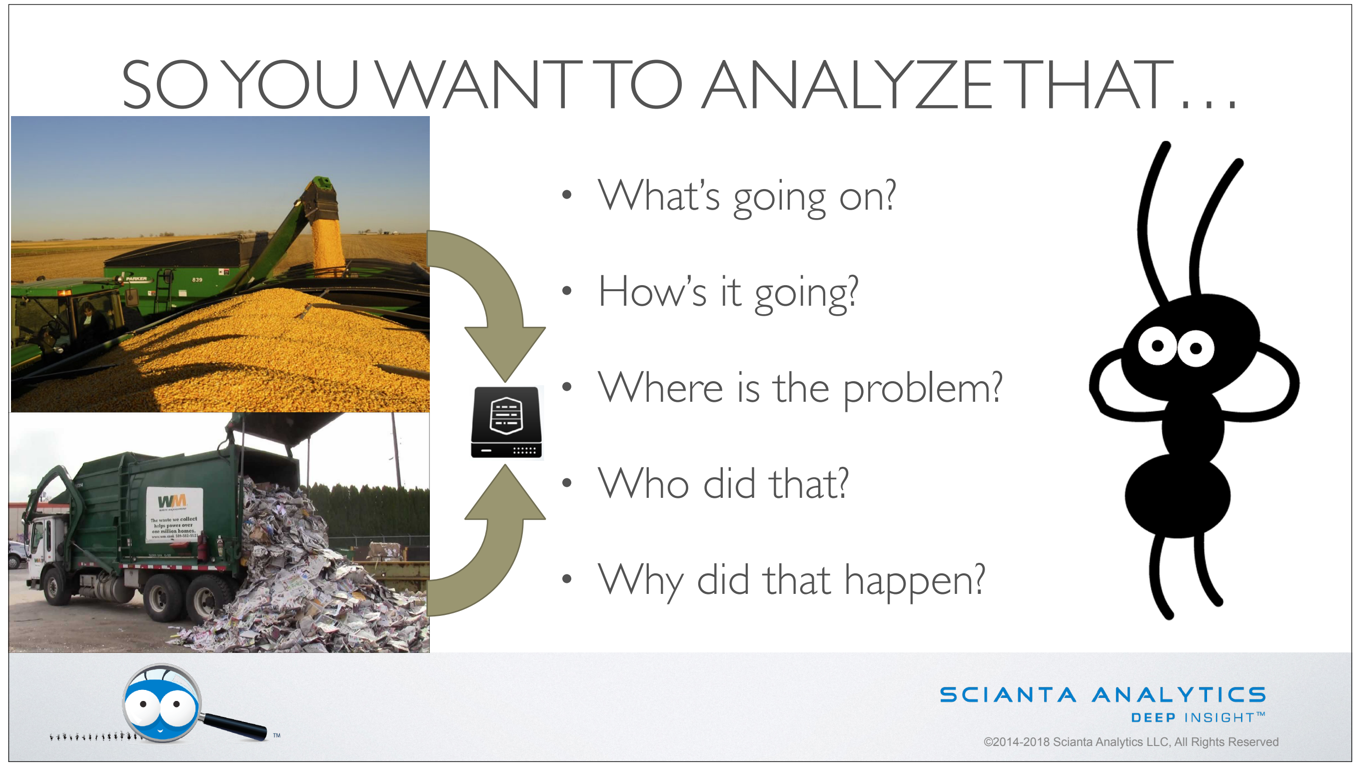
- **Qualitative over Quantitative**: Express rules and results in terms that are easy for humans to understand
- **Pattern Recognition & Anomaly Detection**: Automate the simpler edges of what a subject matter expert does
- **Recursive Quality Review**: Evaluate how well the system is working, adjust accordingly



SCIANTA ANALYTICS
DEEP INSIGHT™

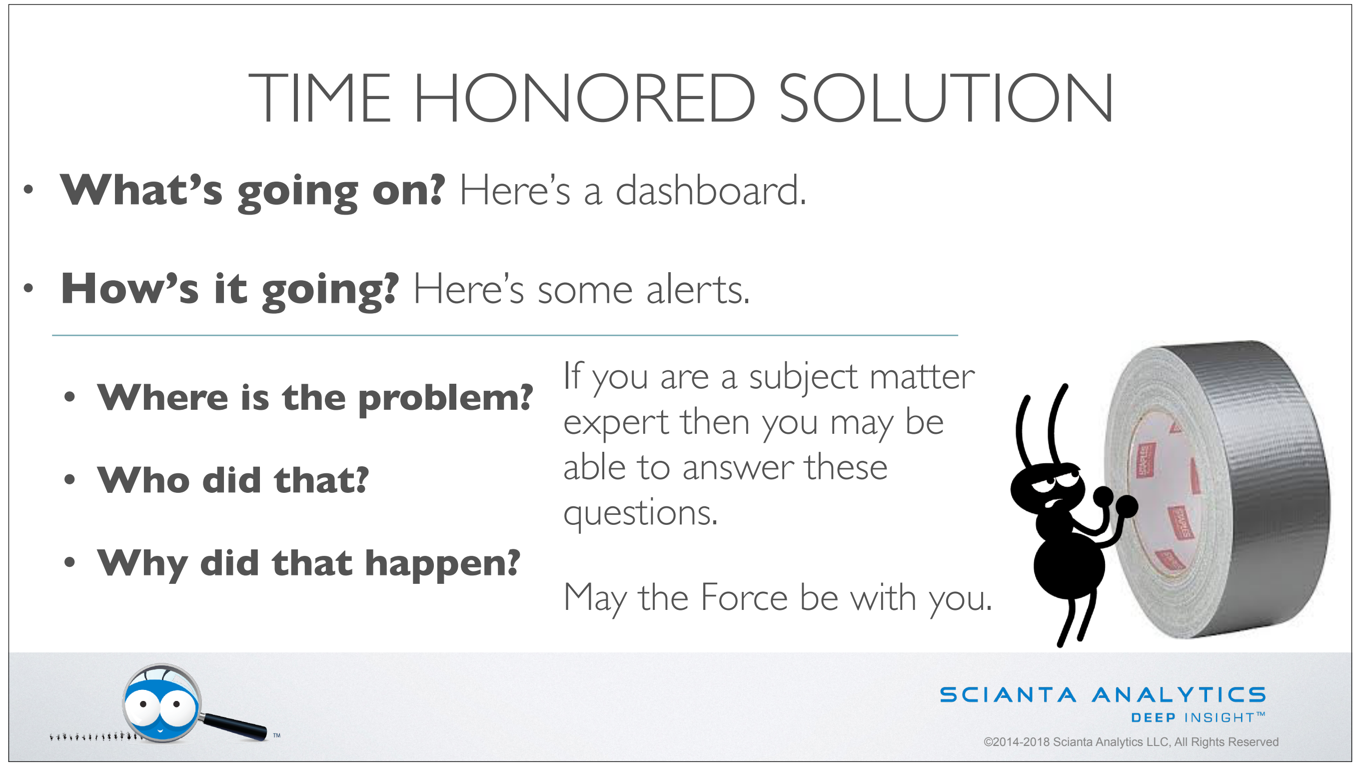
So what we’re seeing is an exciting evolution within the big data ecosystem, tools that give the user more assistance with understanding.

Here’s three concepts that we’ve proven as viable on our chosen machine data platform, Splunk.



Viable doesn't mean easy, so let's make sure we're on the same page about the problem. I like to think in terms of user stories, so: "As a data analyst, I want to throw good and bad data together and achieve good results."

We can generalize those results, too, into five basic questions.



Those are hardly new questions, so how are they being handled today?

General purpose analytics tools that rely on you knowing more than them.

These are great tools for doing your work, if you know how to do it already.

Anyone can solve their own problem, but it's hard to solve everyone else's problem, problem that you're not familiar with.

IS THERE AN APP FOR THAT?

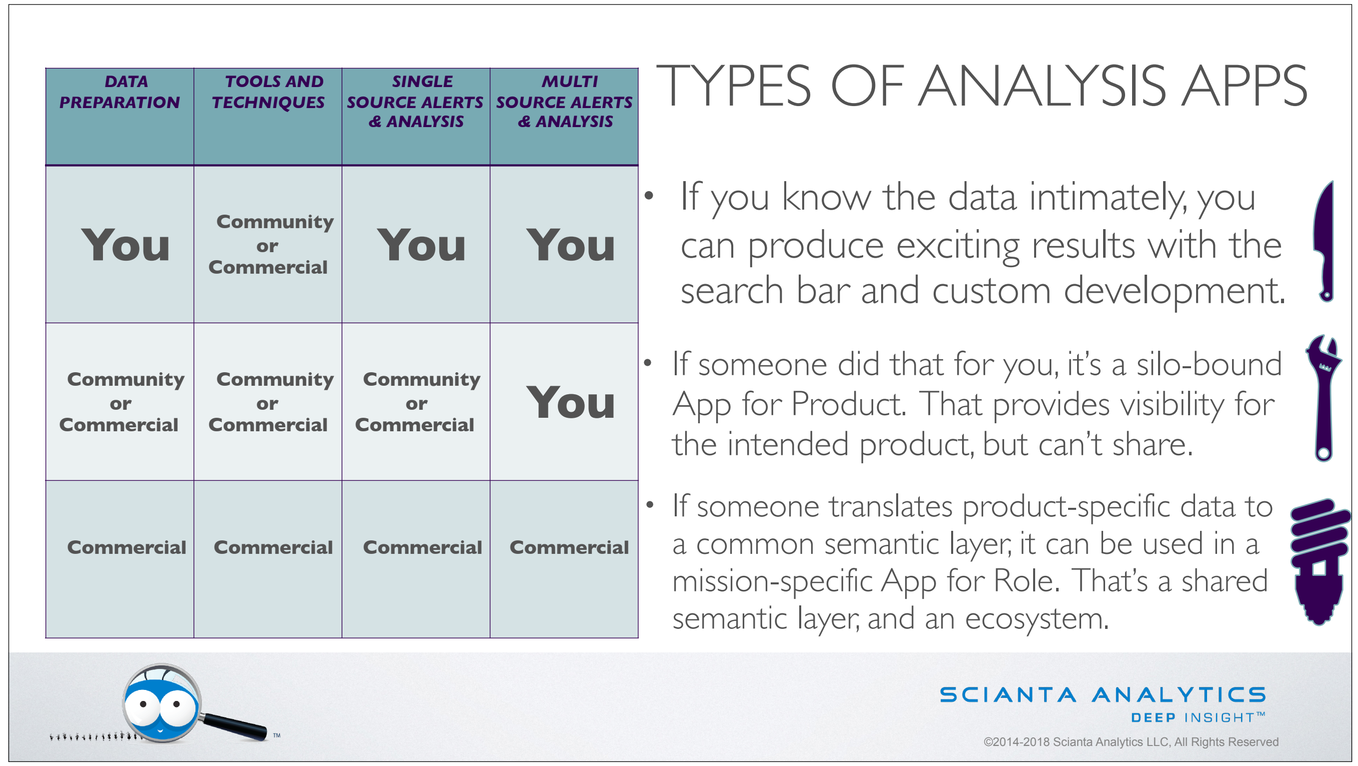
SCIANTA ANALYTICS
DEEP INSIGHT™

Why does that matter?

Maybe you want to make a build vs buy decision.

Custom solutions are better than no solution, but buying off the shelf is better still.

Let's take a second to think about what types of analysis apps are even possible to build.



The farther down you go, the harder it gets, and the less open the solutions are.

Maybe there's a common information model, but is it good? Only if the data collection and use case agree with that model to make a cohesive data system.

Common semantic layers aren't easy, and someone's working hard somewhere to make results that are any good.

There is no I in TEAM, but there is a YOU in "YOU'RE GOING TO DO A LOT OF WORK".

HELP ANALYSTS USE THEIR TIME MORE EFFICIENTLY

SCIANTA ANALYTICS
DEEP INSIGHT™

So there are some Apps are out there, and they do help.

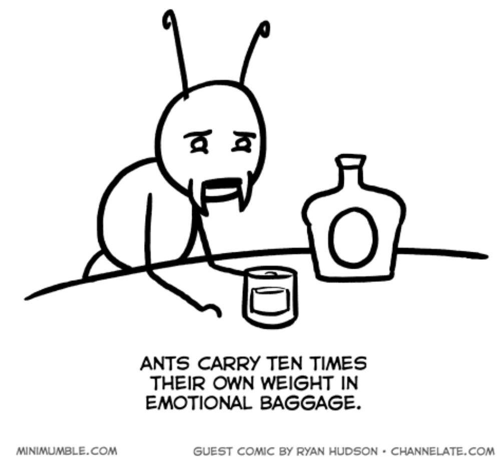
You should probably expect to pay for them; at least in time, but probably in dollars too.

Algorithms and techniques are a growing part of that.

But it can be tough to understand what they actually can and cannot do.

THE BIG PROBLEMS WITH MACHINE LEARNING

- Holiday Problem
- OCP (Black Swans)
- PEBCAK
- Pareidolia (Faces in Things)
- Set And Forget!



SCIANTA ANALYTICS
DEEP INSIGHT™

Computers are force multipliers, they can’t help when you’re doing it wrong.

Instead they make new problems that people have to find and clean up.

Let’s look at the worst of those problems.

HOLIDAY PROBLEM

You're measuring "normal" on a daily basis, but some days swing radically and you can't predict why.

CHOOSE, MORTAL:

- Make the learning window short and accept false anomalies
- Make the learning window long and miss real anomalies
- Simply obtain and maintain an accurate list of the holidays that affect your organization!

IT activities are reduced because users aren't at work
Customer activities are increased because they're shopping and traveling
Manufacturing activities are increased by extra shifts before the day off, then stopped on the day off

Solar year: May Day, Independence Day, Bastille Day
Lunar year: Chinese New Year, Easter, Rosh Hashanah
Short cycles: Fiscal quarters, Academic semesters
Irregular or difficult to predict cycles: Japan's Silver Week, local European saint's days

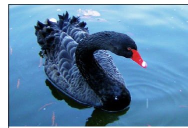
SCIANTA ANALYTICS
DEEP INSIGHT™

Let's start with the simplest problem… what's going on today?

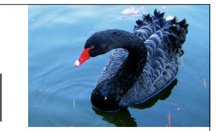
Does this affect your inbound activity, your outbound productivity, or both?

Is this abnormality publicly known?

Are you more likely to see an attack during the distraction?



OUTSIDE CONTEXT PROBLEM



You're measuring within the context you understand, but the situation changes in an unpredictable way.

- Humans have to stay involved
- Machines will never have complete context, so they'll produce weird alerts that need interpreting

Activity is high! Because the company grew.
Metric X dropped to zero! Because that service was migrated to a SaaS
We're selling the hell out of bottled water! Because of a natural disaster
Every actor is behaving anomalously! It's our IPO day.
There's tens of thousands of unexpected people here! Pokemon Go festival
We're expecting tens of thousands who aren't here! Pokemon Go festivals

SCIANTA ANALYTICS
DEEP INSIGHT™

This is why we can't have nice things.

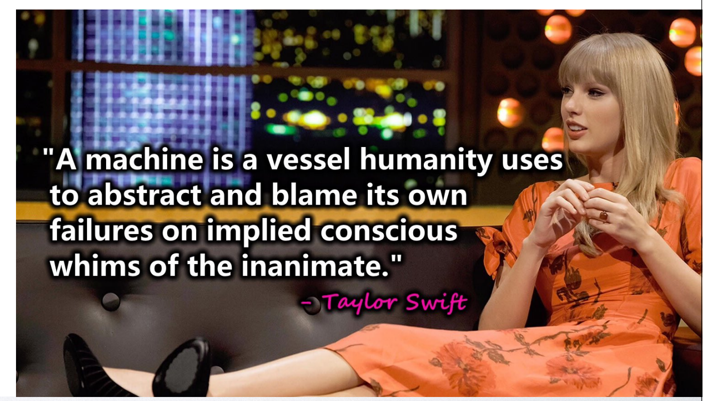
Uber surge pricing is the classic example of this problem. The algorithm increases price when lots of passengers indicate interest, which demonstrates elasticity just like an Econ 101 textbook.

Good for sporting events, bad for terrorist attacks.

PROBLEM EXISTS BETWEEN CHAIR AND KEYBOARD

You're relying on human input to teach the machine, but the humans aren't helpful

- Incidents get handled incorrectly (e.g. ignored)
- Signals get suppressed incorrectly
- Rules get tuned incorrectly
- Data ingest gets handled incorrectly

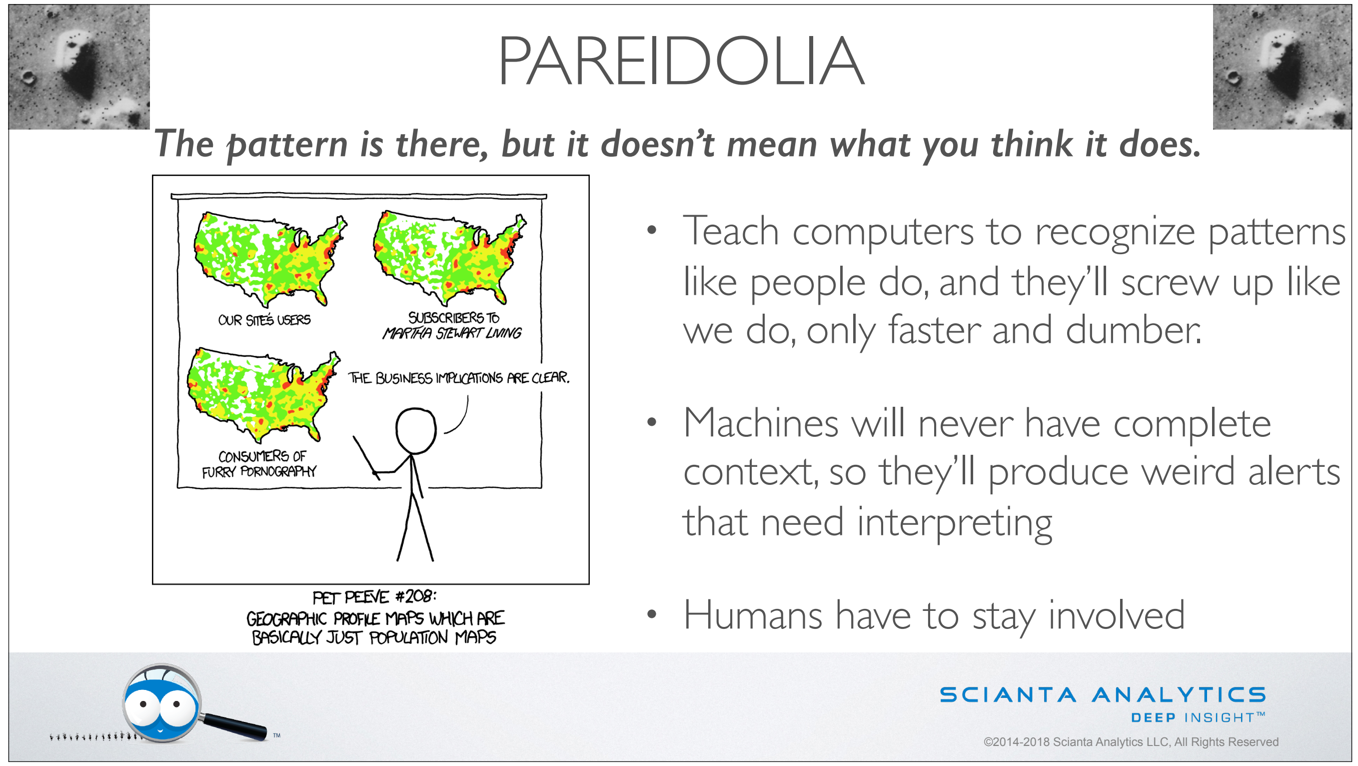


Systems can be misused by a dedicated opponent or ignored by an apathetic partner.

Dumb anomaly detection is prone to this because it false-positives all the time.

So what if I did an unusual thing on Monday? Humans get to do that.

Who trains your watch dog how to recognize a real problem?

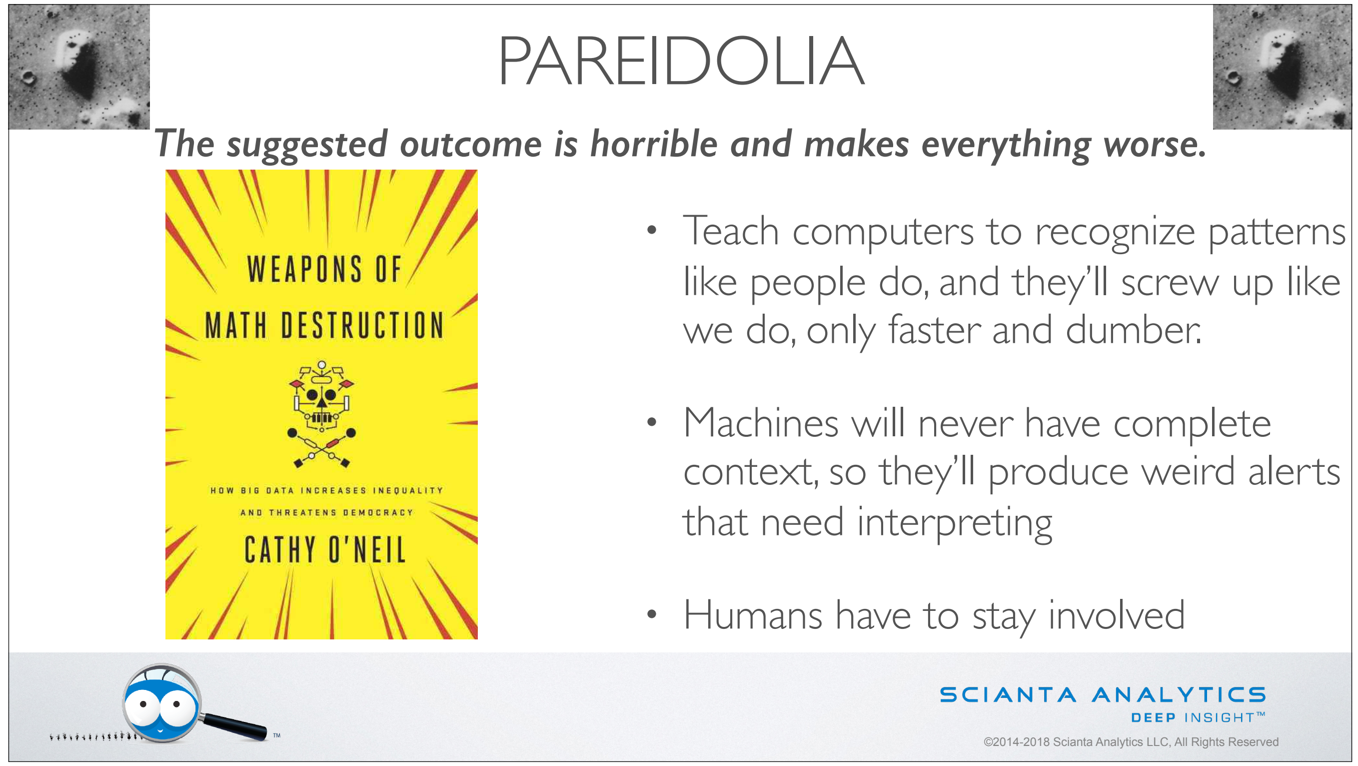


Classic example here: US population maps correlate with a million irrelevant factors. Correlation isn't causation and if you have enough data, you can find correlations easily.

I don't think we're going to make machines that are smarter than people.

People have to be trained to realize they're failing on these problems — & they can be.

So can cognitive computing systems.



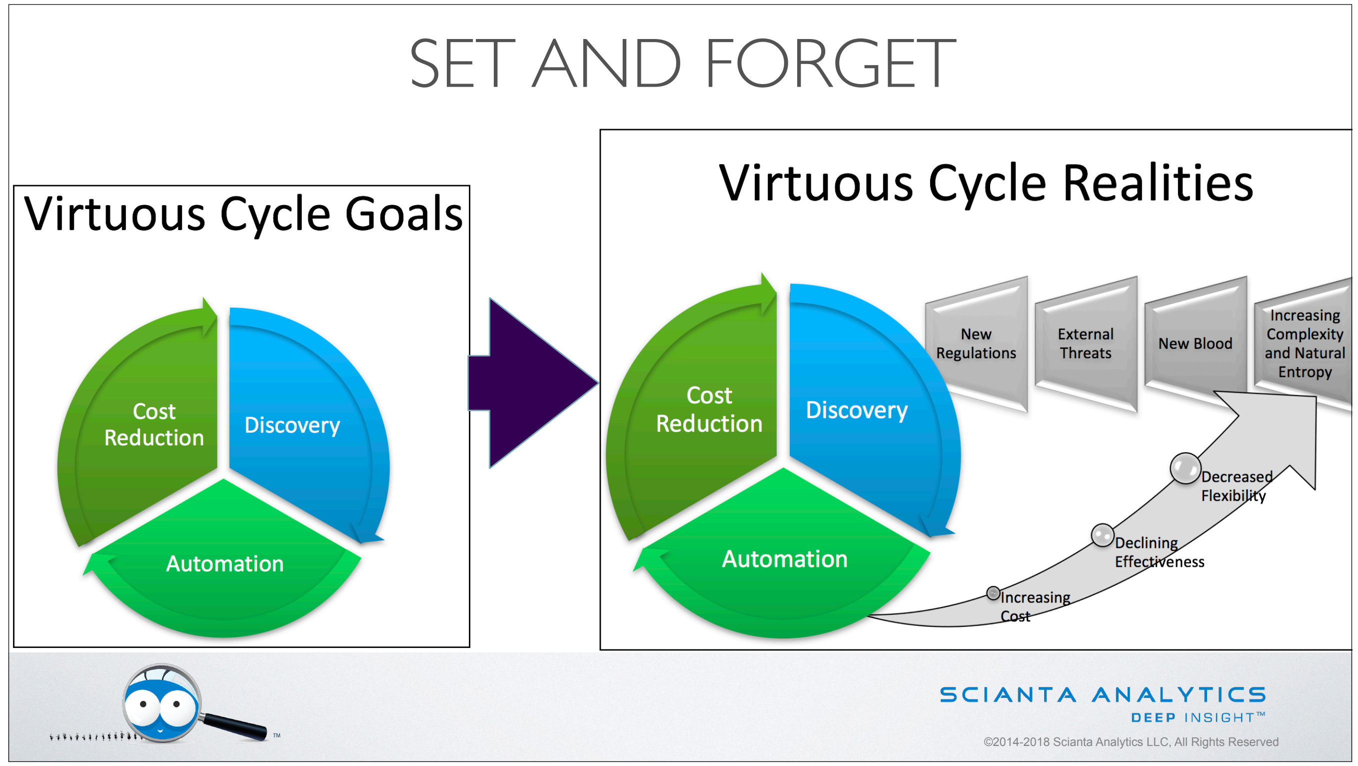
This problem can get a lot worse than just missing an alert.

Who here has seen more than one false accusation against an employee?

How about red-lining customers by race and gender?

Encoded bias, reinforcement of bad norms… this is stuff that impacts real people.

Please, don't let your automation run free.



"After a brief training period this robot will perform perfectly forever!"

Learn bias, add outside context problems, review with inattentive humans…

The model is not useful any more.

It has learned wrongness and needs to be thrown away.

What we need from a cognitive computing system is understanding.

NOW I'M SAD

Don't be sad! Sure, there's problems.

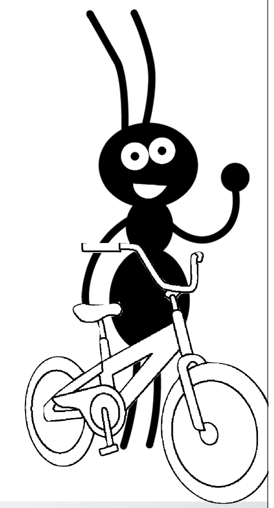
But people and software can work together effectively.

We just need smarter software.

That's what Cognitive Computing is trying to solve.

MACHINE ASSISTED UNDERSTANDING

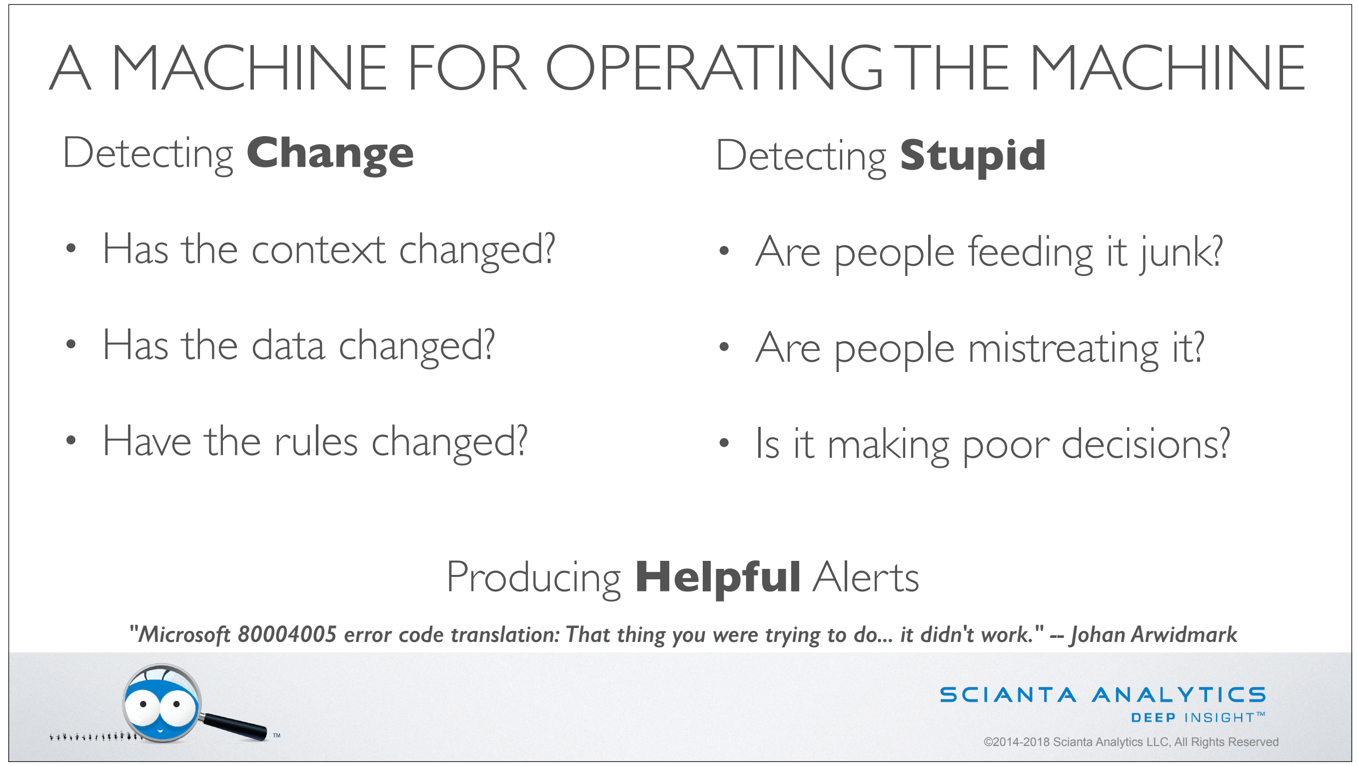
- **Qualitative over Quantitative**: Helps you make context-aware decisions from numeric data
- **Pattern Recognition & Anomaly Detection**: Use that context to find all the signals in the data and make them available
- **Recursive Quality Review**: Apply qualitative measurement, sensitivity to context, and objective awareness of informational density to present results sensibly



SCIANTA ANALYTICS
DEEP INSIGHT™

A Cognitive Computing system is a machine that helps us understand the data, not a machine that does it for us.

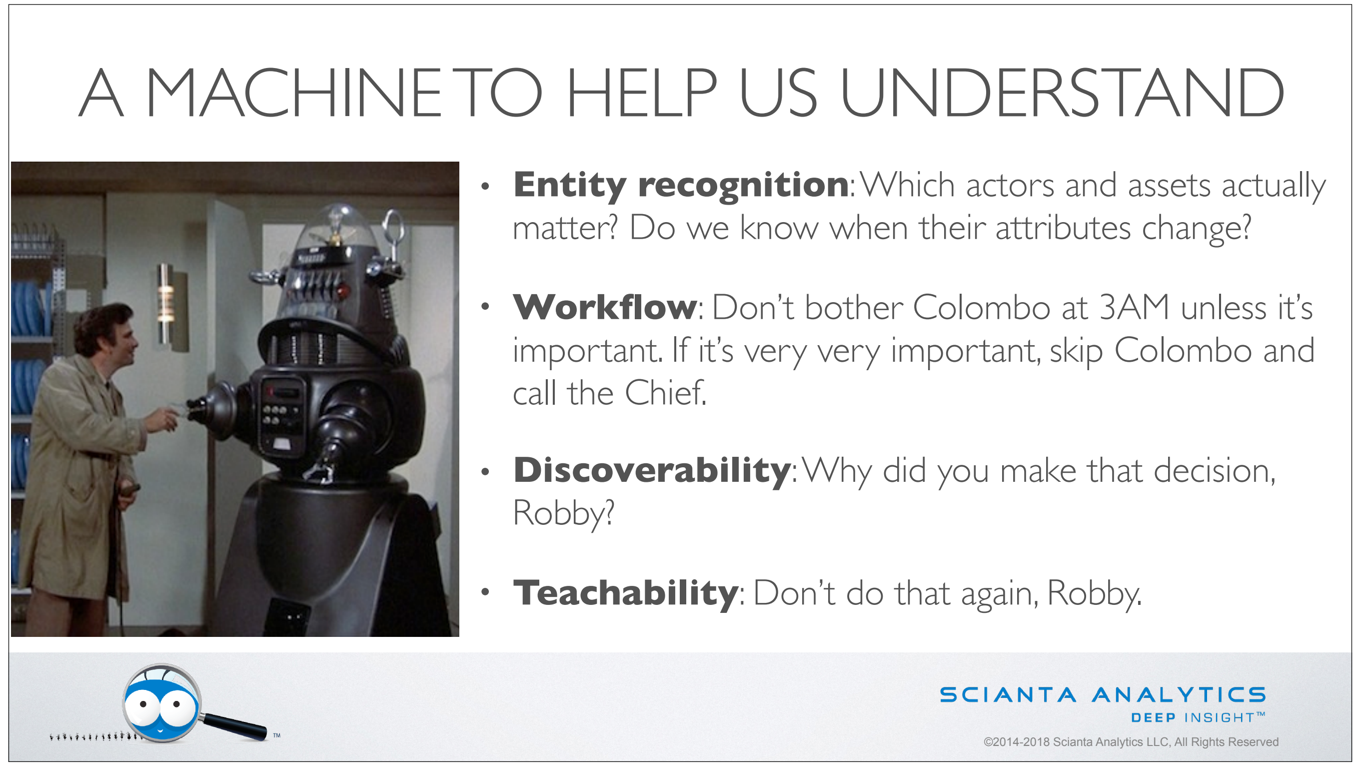
What if we take these three techniques and apply them to training the model to evaluate itself?



That gives us some achievable goals.

A cognitive computing system needs to be smart enough to see these problems and alert on them.

That's a huge step forward.



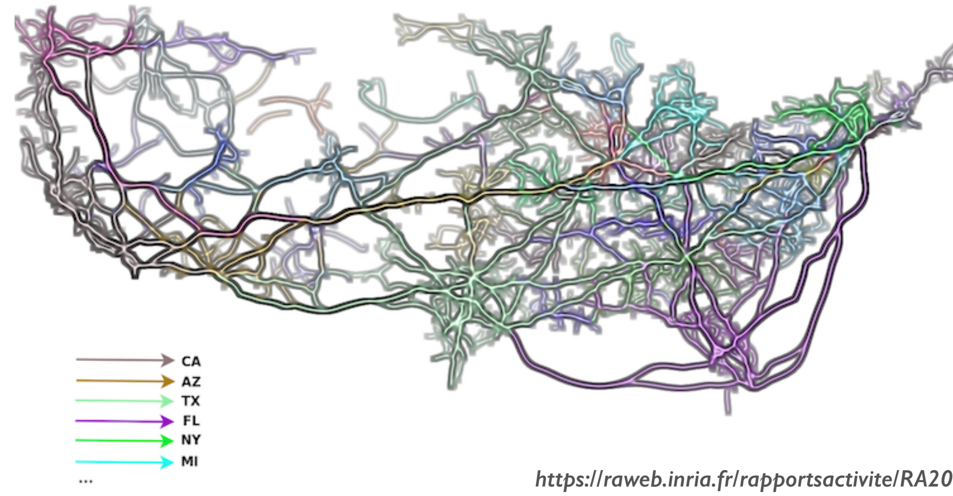
Humans automate so that we can stop being machines and start operating them.

We're going way past the capabilities of most big data platforms, and looking at features from other types of enterprise software: ITSM, BPM, workflow orchestration.

I highly recommend taking advantage of those when designing alerting and analysis.

If you can, they're often blocked by OSI layers 8 and 9.

BEHAVIOR IS BETTER



https://raweb.inria.fr/rapportsactivite/RA2010/gravite/uid55.html

SCIANTA ANALYTICS
DEEP INSIGHT™

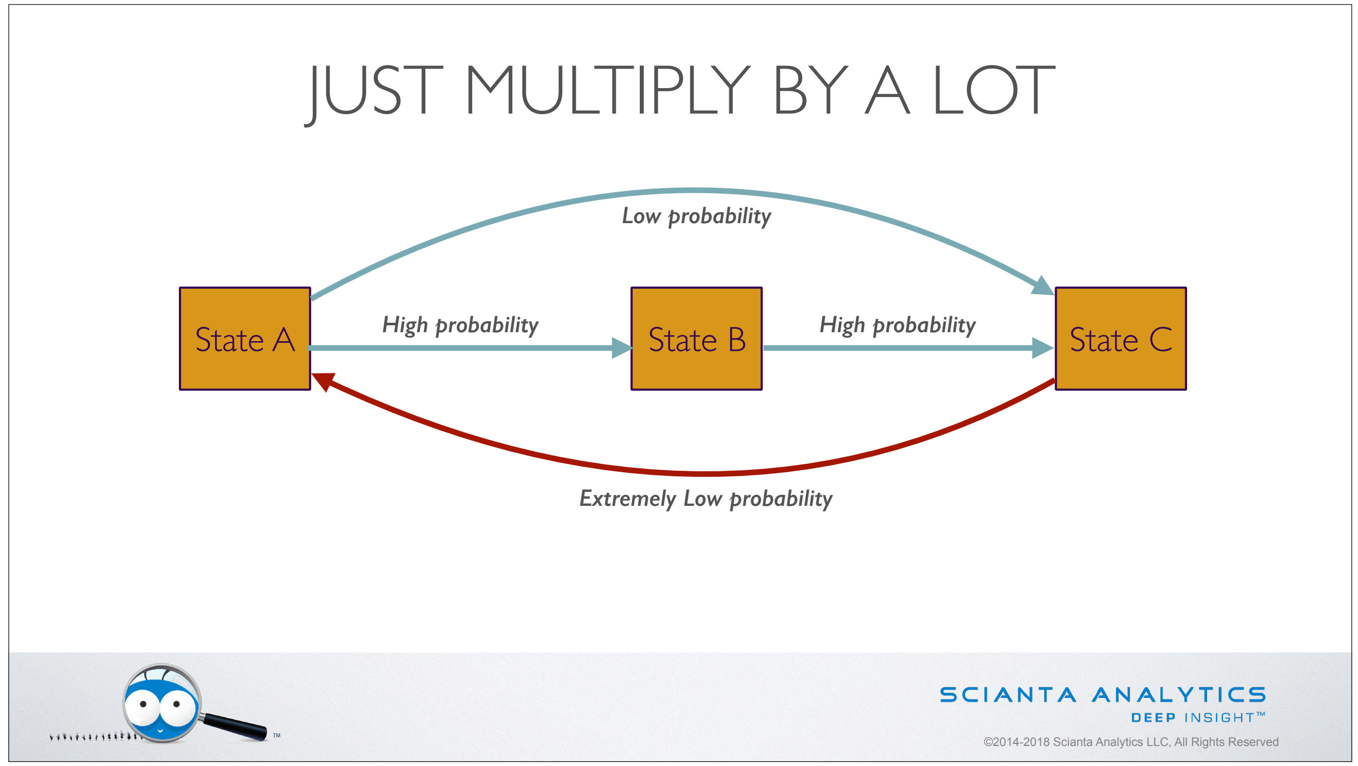
And the reason we want those features is to understand behavior.

Behavior analysis is pretty exciting as a way to derive more meaningful signals.

Simple triggers from known bad events are cheap and easy, so use them when you can.

Normalcy testing is great too when it's right for the data source.

But both can be used to review behavioral sequences and understand an actor's journey, which is super powerful.



This is a simple graph, showing movements between states.

Each one of these movements can be thought of as an abnormality measurement, and then considered on its own or with the larger transactional session.

So there's your signal when reality does something abnormal.

This gets really exciting, especially when each node is itself a multi-event transaction.

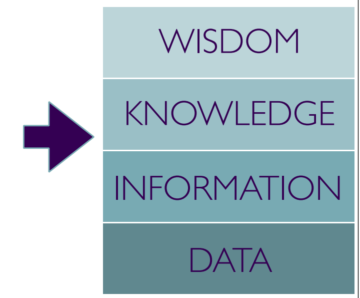
NEW ANALYSES FROM BEHAVIOR

Sequence analysis

- How normal is this pattern of events for this actor in this time frame?

Peer analysis

- How normal is this actor's behavior compared to others in the same group?

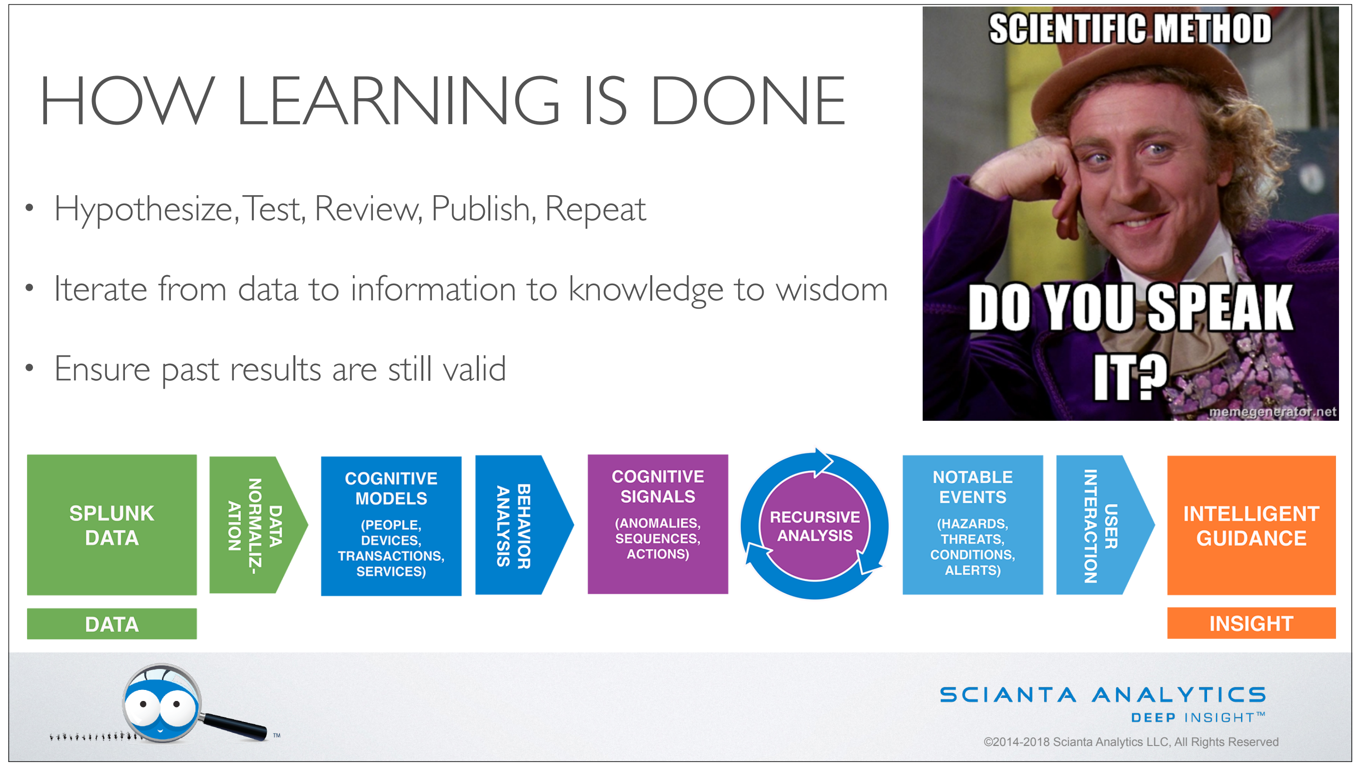


SCIANTA ANALYTICS
DEEP INSIGHT™

Understanding behavior increases the power of our tools and enables Sequence and Peer analysis, for better prediction and alerting.

And if those signals are coming from our own data analysis, and the behavior of analysts, we're measuring the health of our data system.

So if we teach the machine to learn from itself… there's a big step into Cognitive Computing.

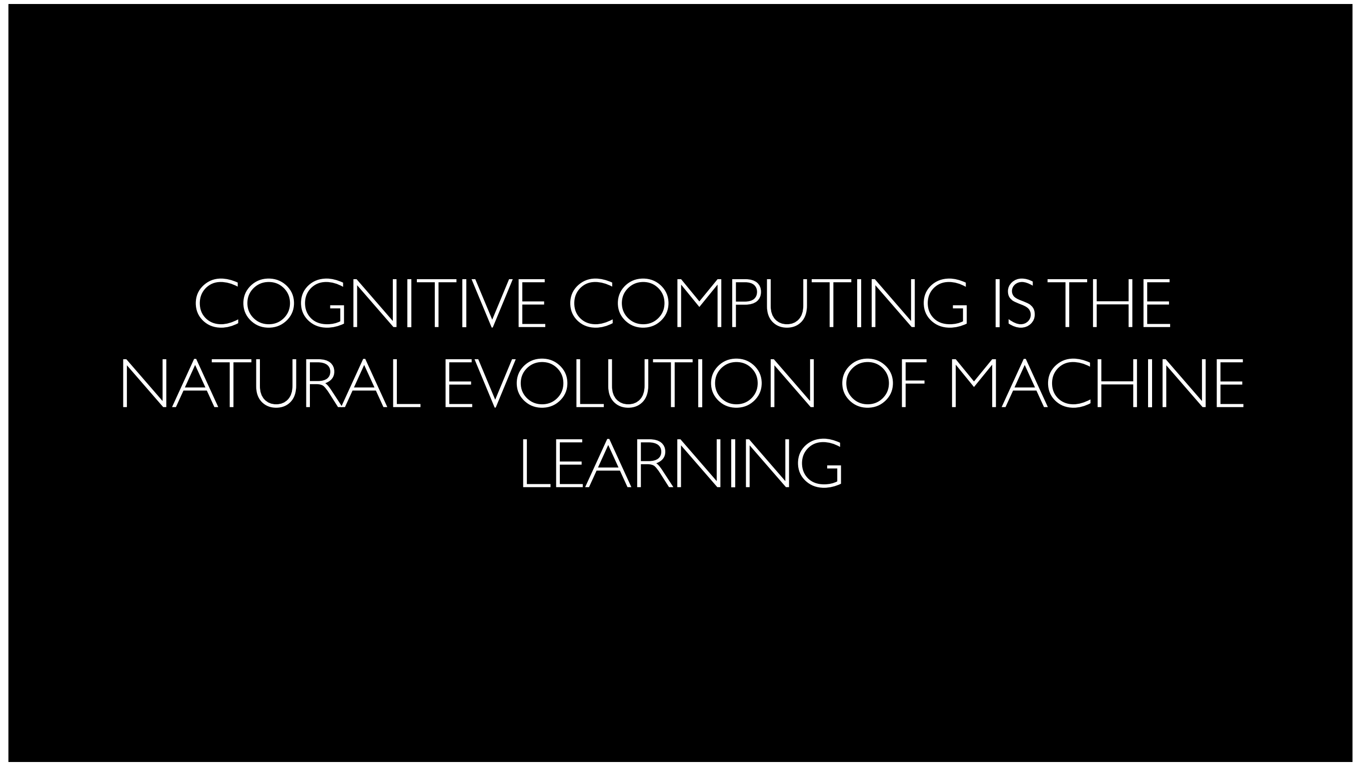


The ideal answer for teaching machines starts looking a lot like Scientific Method. We groom our data: recognizing entities, finding transactions, labeling events.

We form a hypothesis, which is the model and the probabilities of sequences.

We perform multiple passes over data from multiple sources, then make decisions.

Then we recursively evaluate our signals for their fit against raw data and the hypothesis, adjust their weights, and notify the analyst.



So I like the phrase Machine Assisted Understanding to describe where we’re at now. We’ve found a greater level of fitness which is more successful at making qualitative decisions from quantitative data.
It’s not perfect, but it can help us be better at our roles.

Thank you again, and let’s do questions!