Many industries have a few big players and a long tail of smaller companies. Naturally, this leads to a long tail of document variety... and to a key document challenge

The Long Tail in Documents: How GLYNT Tackles This Key Challenge

Our guide to a key challenge to document automation

GLYNT.AI

CHAPTER ONE

AI for Documents: Learn the First Key Question

It was only a few years ago that innovators started including AI-powered features in enterprise software, but it's been just long enough for patterns to emerge.

One pattern is a bit disturbing: AI often has Diseconomies of Scale. Yikes! This conclusion applies to many AI systems that transform documents into data.

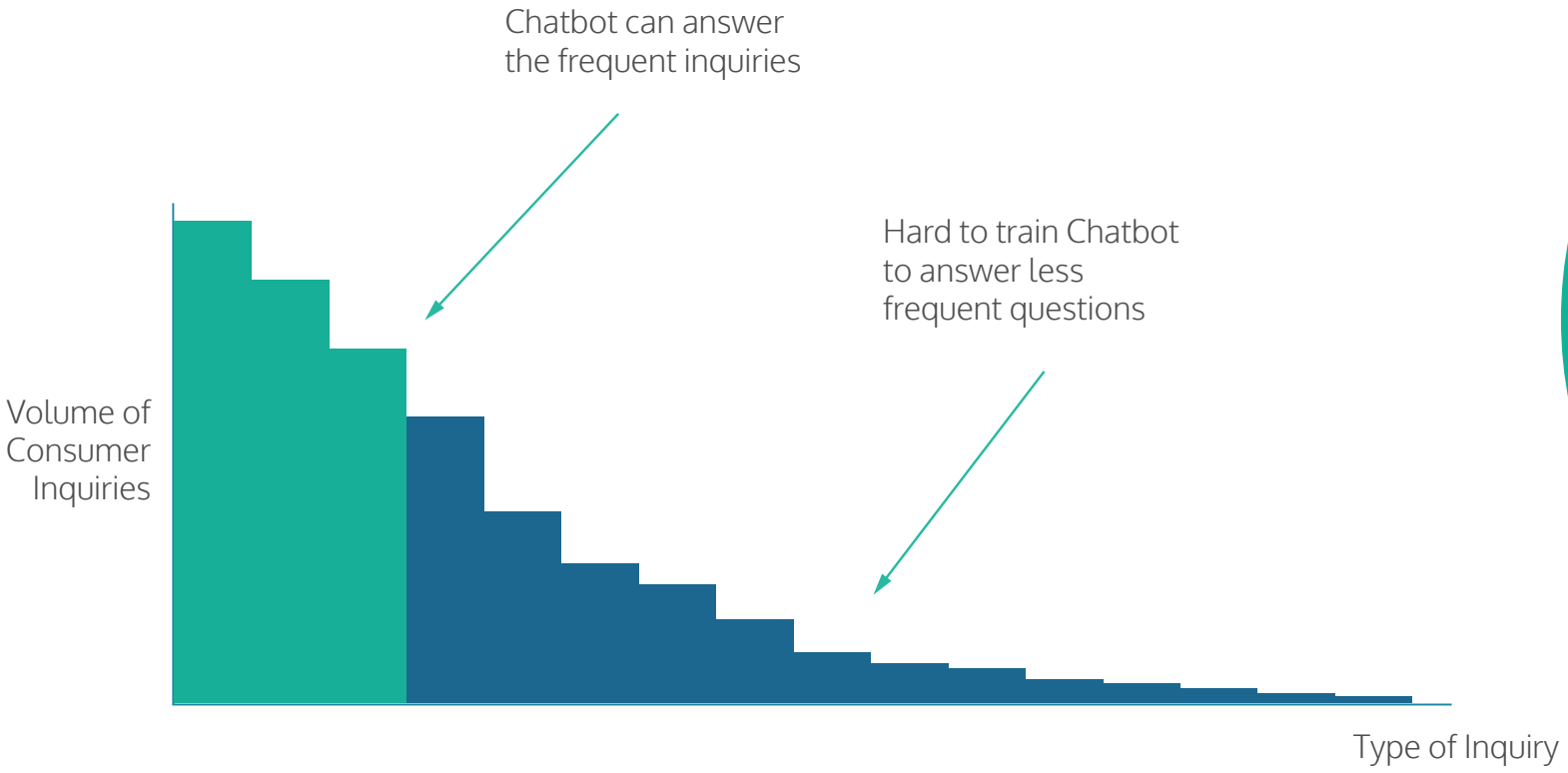
In this white paper, we explain the challenge of the long tail facing AI today, and then look at different AI systems and how they fare at tackling this headache. Part I outlines the significance of top-down vs. bottom-up AI, and how the wrong choice leads to Diseconomies of Scale. Part II looks at the task of categorizing data and how the right AI can produce reliable, ready-to-use data.

For AI-powered document-to-data solutions, differences in the AI system can impact your ROI. Read on to learn what key questions you should keep in mind while searching for a winning solution that meets all of your needs.

The Challenge of the Long Tail

To understand why AI might have Diseconomies of Scale, let's take a look at a familiar AI application, chatbots. Consumers pose thousands of questions. How many replies can be automated? Here is what venture capitalists are reporting: About 20%

The problem is that there is a long tail of infrequently asked questions. It is expensive to catalogue them, get enough to train the AI, organize the content for AI learning, and maintain the AI as questions drift to new types over time. Soon the AI system manager is facing increasing costs, in that it costs 2-3X more to train the AI to do a new question than it did to set up the initial frequent questions.



Cost Tradeoff:

Some questions are more cheaply answered by AI, and some are more cheaply answered by humans. That's how the 20% is set.

Top-Down AI

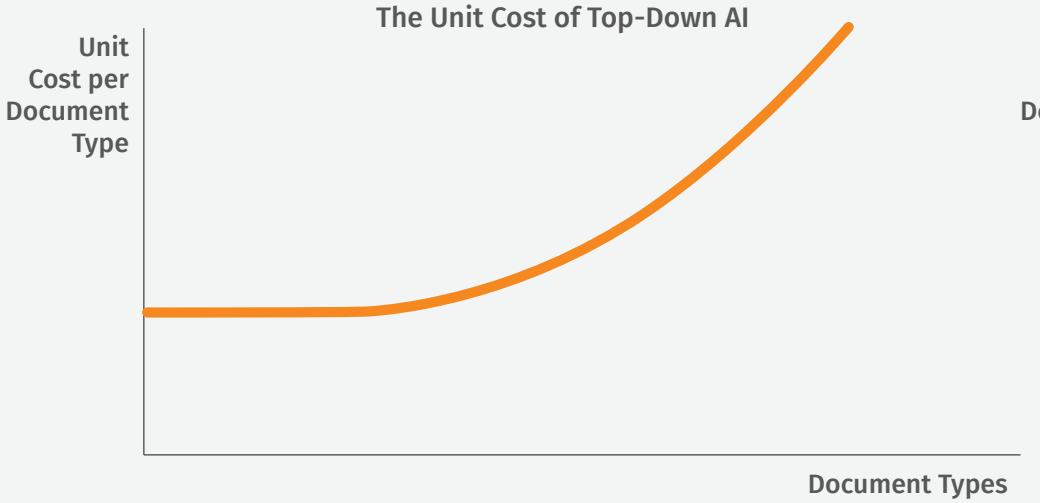
Documents also have a long tail. Every company knows that some documents are frequent and regular, W-2s for example. But even a simple document such as a driver's license will have a lot of variations: At least one version for every state, then there are auto, motorcycle and truck licenses, and then… the long tail emerges.

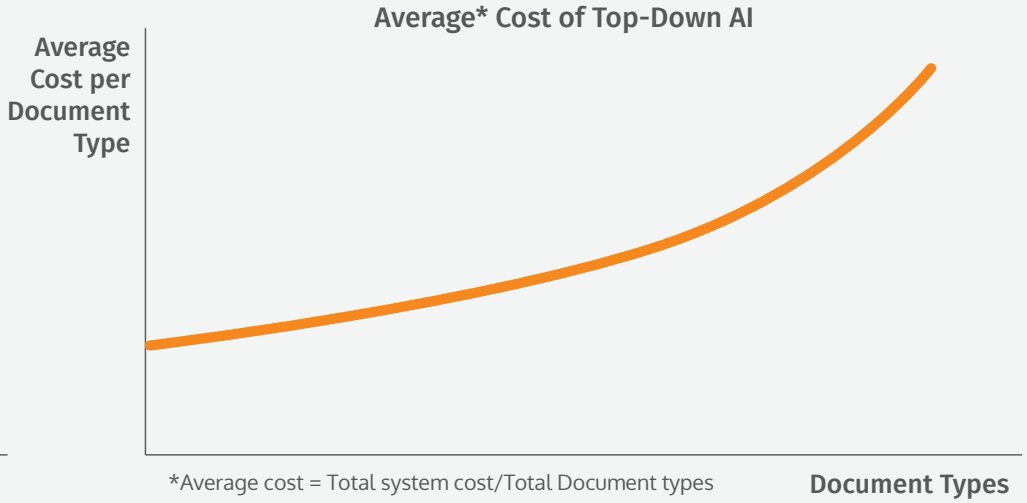
The AI system most frequently used for chatbots and documents is built on a huge library of examples, typically 200,000 or more. For documents, this means the fields are selected, the correct data is marked, and the AI is trained. The system is called Large-Corpus AI (reflecting the size of the training set) or Top-Down AI (reflecting the pre-selection of fixed fields).

Economies of Scale:
unit costs fall as more units are produced.

Diseconomies of Scale:
unit costs increase as more units are produced.

Top-Down AI has Diseconomies of Scale





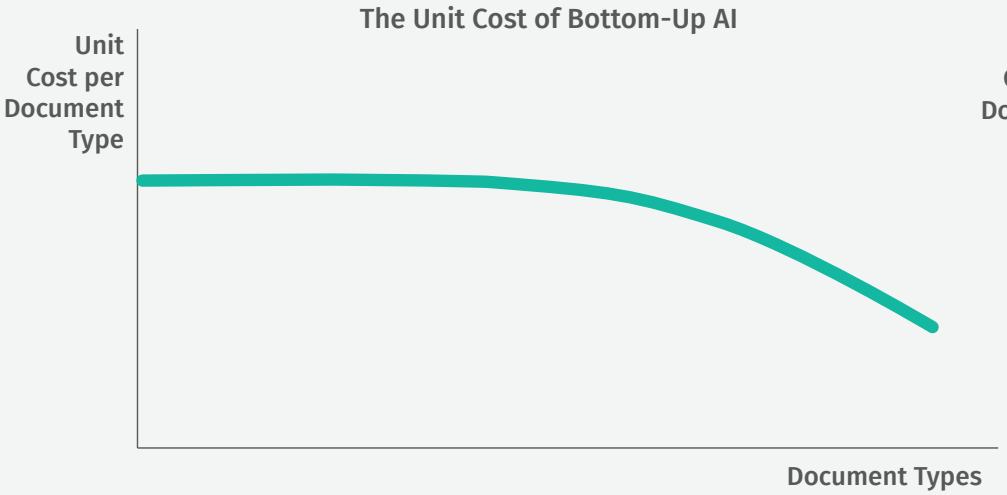
Bottom-Up AI

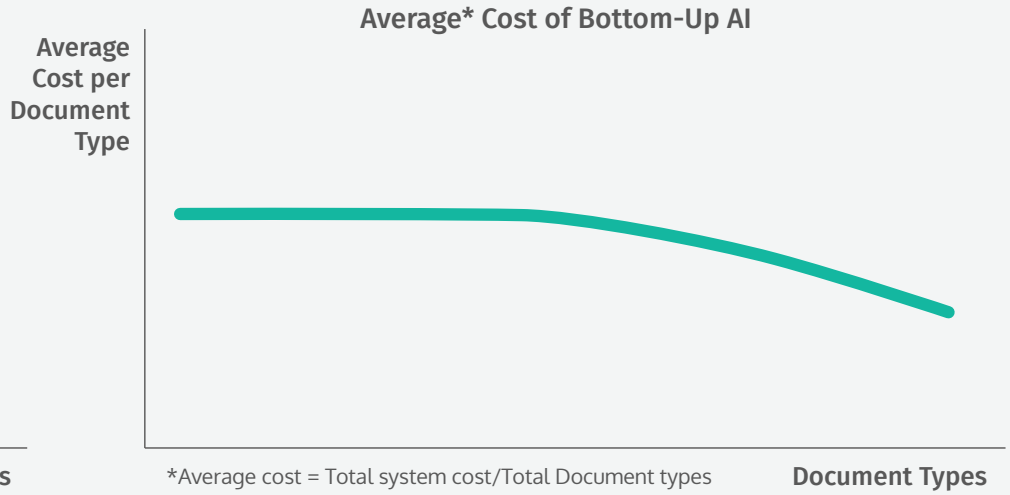
Now consider GLYNT's Few Shot machine learning system. It trains on just a few example documents. This means only 3-7 documents are needed in a training set, and any field can be selected and trained. Importantly, the costs of training the first document group are the same as the 100th or 10,000th document group. GLYNT has flat unit costs.

Plus, GLYNT is building out libraries, enabling faster training because we've seen the document type before. The libraries are of document math, not data, preserving the privacy of every tenant on our system. Few Shot plus Libraries delivers decreasing unit costs to add a new document type, eg Economies of Scale. Because GLYNT uses so few training documents and has field selection flexibility, it is known as a **Bottom-Up AI**.

Bottom-Up AI has Economies of Scale

Thanks to our Few Shot machine learning system, GLYNT is Bottom-Up AI





*Average cost = Total system cost/Total Document types

The Key Questions to Ask

Obviously we think you should ask a potential vendor if they are **Top-Down AI** or **Bottom-Up**. What might not be as obvious is all the business implications of that single question.

BUSINESS QUESTION	OTHERS: TOP-DOWN AI	GLYNT: BOTTOM-UP AI
Can you automate our long tail of document variety?	No. We persistently route a lot of documents to human teams.	Yes. We're set up to automate nearly every one of your incoming documents
What is your accuracy rate?	Very high if the document is frequent. Very low if the document is infrequent. Blended rate of accuracy is middling	Same very high accuracy for every document type.
What if there is an error?	File a ticket. If we get enough of the same request, we may be able to add the document group to our training system. Wait months.	Re-train GLYNT in minutes.
Does your AI "learn?"	Sort of...If it is a frequently seen document and we go back and update our very large training set, and we have enough examples...Yes, our system learns.	Yes. You can teach GLYNT with just 1-3 example documents.
Do we get your best AI, the one that aggregates learning from all of your customers?	Yes. We'll do transfer learning and federated learning to bring you the advantages of seeing all the documents. But it is costly to manage such a complex system, so expect higher prices.	Yes. GLYNT shares the math models of documents we've seen across customers. This speeds up your training experience, so our efficiency lowers your cost.

The Bottom Line

If you're looking for a documents solution, ask your vendors the single key question: **Is your AI Top-Down or Bottom-Up?** If they look confused, ask them the list of questions above. You'll be able to decipher the answers.

As you look for a documents solution, remember every aspect of the AI system impacts your ROI:

More AI coverage of the long tail
→ **Lower costs**

Higher accuracy
→ **Lower costs**

Faster error fixes and training
→ **Easier to maintain and lower costs**

An AI that learns quickly
→ **Easier to maintain and lower costs**

In the next chapter, we'll cover how AI-powered documents solutions grapple with data categorization, and what to look out for if you want verified data ready to use in minutes.

CHAPTER TWO

AI for Documents: Learn the Second Key Question

Imagine a chocolate factory where you can build the chocolate bar of your dreams: Pick your cocoa beans and toppings, and the factory processes it into one delicious chocolate bar ready for you to enjoy.

Everyone who uses GLYNT wants to select lots and lots of fields and capture all the data items they have wanted for years, just like kids in a chocolate factory. Since GLYNT is built on 'Few Shot' machine learning with on-demand training, getting to that chocolate factory experience is easy. But…

The Challenge

Once GLYNT delivered all of the requested data, our early users realized that lots of documents leads to lots of document variety which leads to incredibly jumbled up data feeds. Even simple documents, such as healthcare insurance cards and paystubs have state-level regulations to meet, leading to a huge variation in how fields are laid and what they are called.

Integration engineers were brought to their knees. How could the data be reliable? Every field seemed to have a different format and meaning. Business users screamed in pain. Automation wasn't working, every document got routed back to the manual data entry team for a cleanup. A confusing tangle of data. Sound familiar?

The Semantic and Transformation Layer

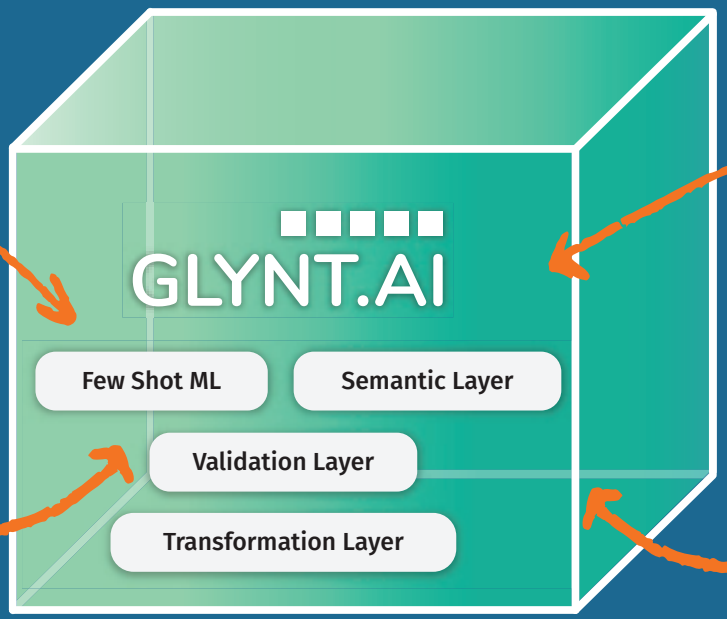
To tackle the problem of document variation, we added the Semantic Layer to GLYNT, a tool that automatically normalizes the variations of what's printed on documents.

And then, based on customer feedback, we added the Transformation Layer, which maps the document data into a structured, ready-to-consume format. We think of this as gold foil chocolate bars. Data ready to eat!

We started with picking the cocoa beans and toppings (lots of data!), then went through a refining process (no more semantic mess!), and finally arrived at gold foil chocolate bars (yummy!). Machine learning alone is not enough.

"Give me lots of data! Now that data is easy and cheap, I want it all!"

"Don't ask our team to do complex data mapping. Give us data ready to slip right into our systems."



"Remove complexity! Take out the varied data schemas in documents. Make it easy to map data into our systems."

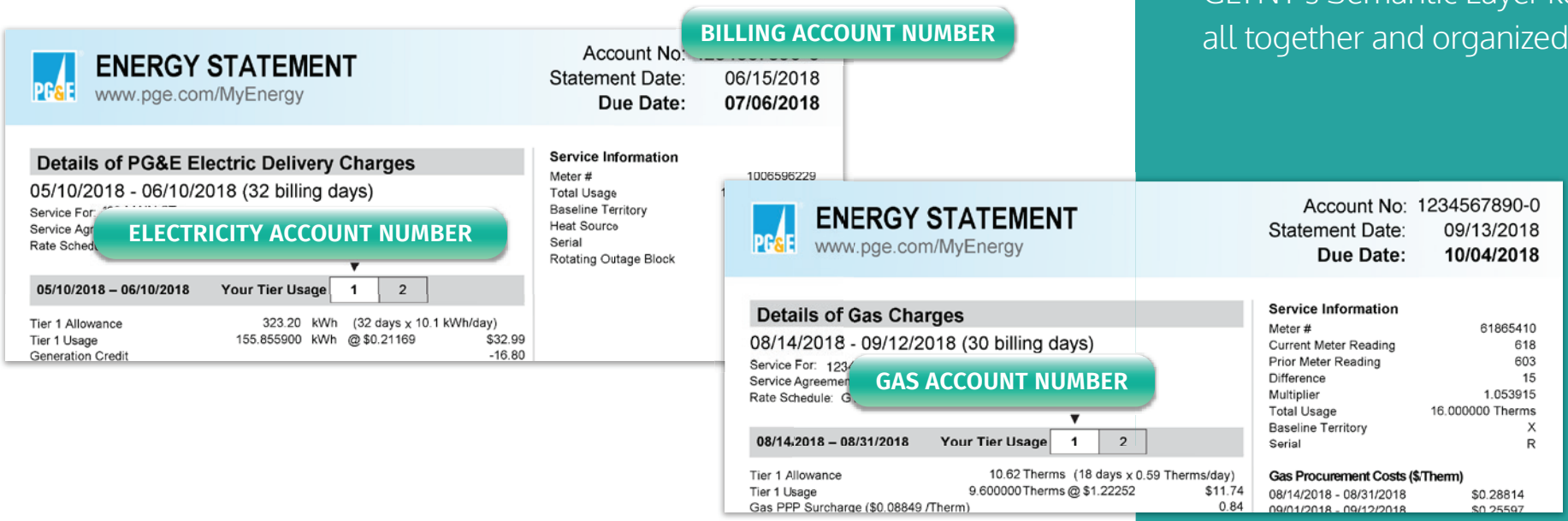
"Automate the tasks we do to clean, normalize and validate data. Send alerts when something doesn't look right."

An Example of Semantics: Categorizing Data

Here is a use case for our Semantic Document Layer that comes up all the time. GLYNT processes utility bills and they often have multiple services on the same bill.

For example, in Northern California, PG&E delivers gas and electricity to businesses and homes. So the simple question "What is your account number?" has three answers: the billing account number, the gas account number and the electricity account number.

GLYNT uses meta tags to help the user categorize the data we extract. The tags are applied to the data items.



Rinse and Repeat

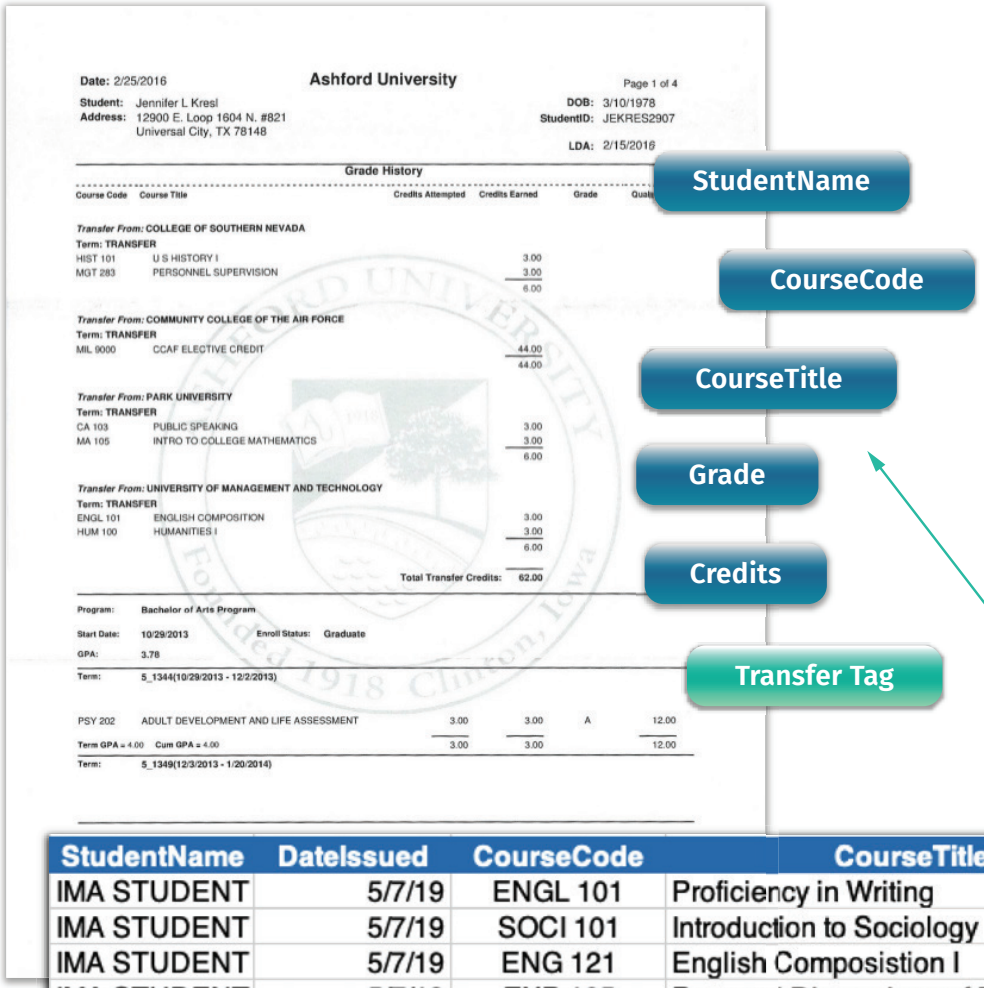
GLYNT has developed data identification and categorization tags based on our years of experience in the energy industry. And we've found that the same system works on other data-intense invoices, such as complex invoices, healthcare records and government documents too.

All have lots of data, contained in nested tables, and with data context spilling across pages. GLYNT's Semantic Layer keeps it all together and organized.

Delivering Data with Context

In our view, the Semantic Layer for a document type is complete when the user has complete context for the selected data items.

To make this clear, let's look at school transcripts. See how this transcript is transformed into reliable, structured data.



The course data for Spring Semester in Year 1 is great stuff, but one needs the information in the header area to provide context, such as Which student? Which year? And so on.

A system of field names, tags and data transformation recipes

StudentName	DateIssued	CourseCode	CourseTitle	Grade	Credits	Transfer Tag	DateCompleted
IMA STUDENT	5/7/19	ENGL 101	Proficiency in Writing	Not Printed	3	Transfer	Not Printed
IMA STUDENT	5/7/19	SOCI 101	Introduction to Sociology	Not Printed	3	Transfer	Not Printed
IMA STUDENT	5/7/19	ENG 121	English Composistion I	B+	3	InResidence	2/5/18
IMA STUDENT	5/7/19	EXP 105	Personal Dimensions of Education	C+	3	InResidence	4/2/18

The Key Questions to Ask

The one key question to ask of your documents to data vendor is:

"Will the data provided be immediately consumable?"

Anything but an unqualified "yes" in reply will raise your overall cost. Here are a few follow-up questions to help you check the boxes:

BUSINESS QUESTION	OTHERS: TOP-DOWN AI	GLYNT: BOTTOM-UP AI
Can we tell you what fields we want? Or add a field?	No. Fields are set. If you want an additional field we can't help.	Yes. Just select any field you want.
Can you get data items on any page?	No. We only get the fields we set up from the start.	Sure!
Can we choose the file names?	No. We have a list, or we use what is printed on the document.	Sure!
Can you provide context for data on a table?	Maybe. If it is a standard table on a document we've trained on.	Sure!
Can you de-nest table data?	No.	Sure!
Can you do sub-tables of tables?	Maybe. Only if it is a standard sub-table on a document we've trained on.	Sure!
How many fields do you get per document?	Typically 12-25.	As many as you want.
Can you customize the data output to my structured data schema?	No.	Sure!

The Bottom Line

Powerful machine learning is not enough. A big shoutout to our customers who went on a journey with us. Through the chocolate factory phase all the way to where we are today: gold foil covered chocolate bricks. Structured data, ready to consume. Delicious! The suite of tools has proven to be incredibly useful to our customers and work well across industries.

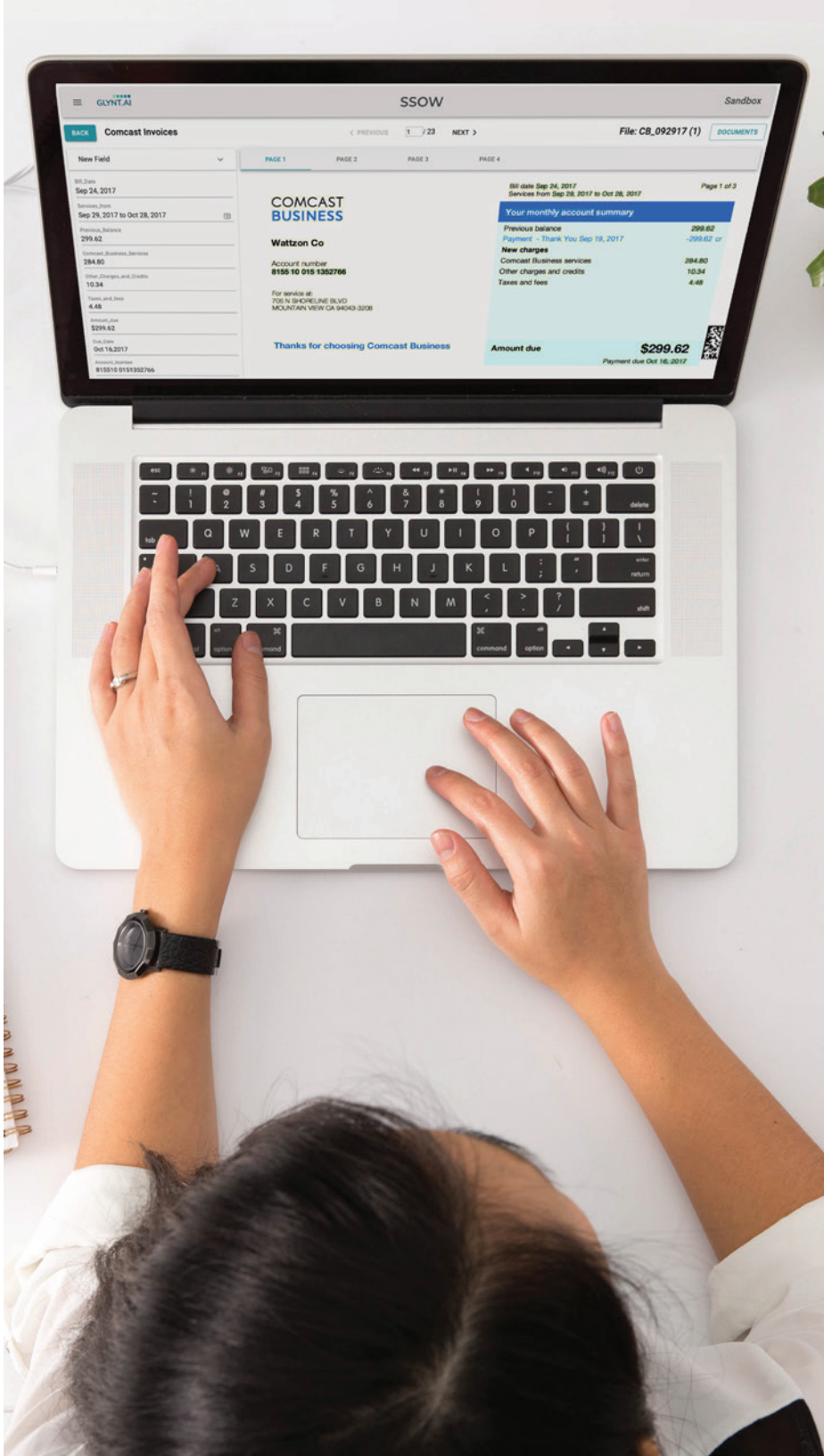
The bottom line is that anything but immediately consumable and reliable data from documents is a big headache and a source of unending costs. Documents have huge variation, even within an industry or document type. Get a solution that knows how to conquer document variety.

Semi-automated and manual data entry systems are slow and difficult to manage. Scalable solutions remove the chokepoint to revenue growth. An easy-to-use no-code product lets your team thrive and scale

READY TO TRY GLYNT?

Contact us for a customized demo on your documents! You send us the documents you see every day, and some tricky edge cases. We'll show you how GLYNT tackles the tough challenges. Let's have some fun!

GET IN TOUCH



GLYNT.AI
530 Showers Dr, #7416, Mountain View, CA 94040