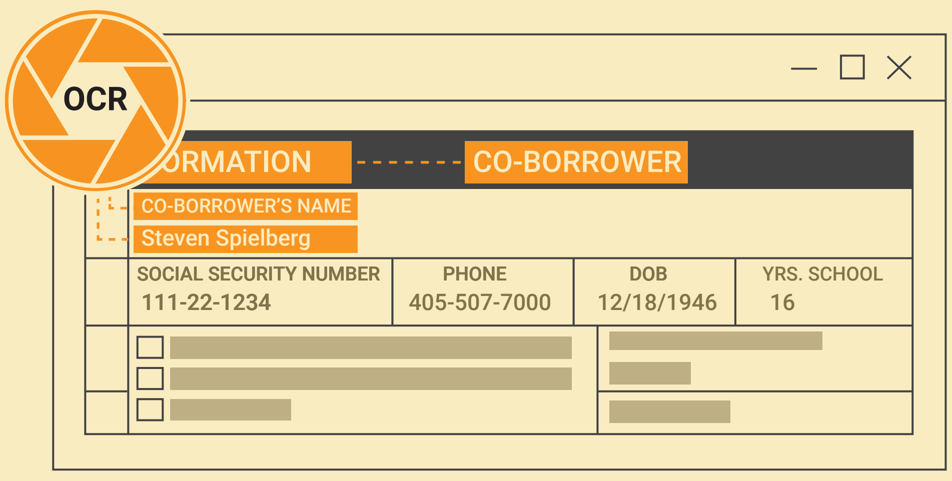
Grooper®

INTELLIGENT DOCUMENT PROCESSING

A framework for data integration

Humans read and understand documents based on the data and structure of words on a page. A computer “reads” using Optical Character Recognition (OCR) to produce a sequence of characters. OCR is just a mechanical process that doesn’t recognize the meaning of data. **Creating synthetic understanding of data is a difficult task and we’ve done it.**



ORMATION ------- CO-BORROWER

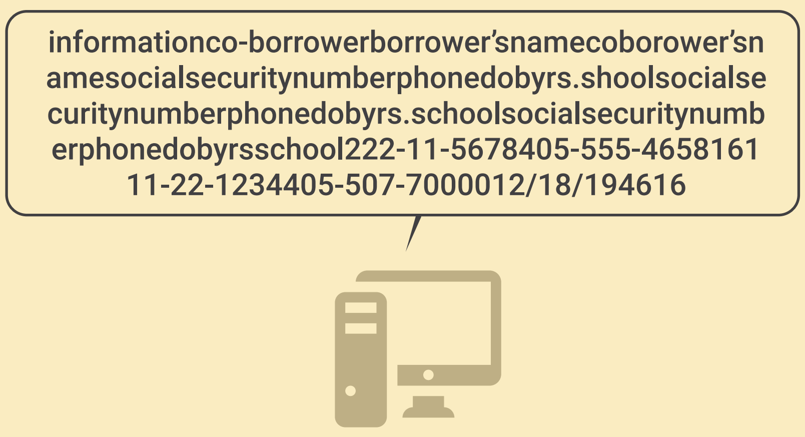
CO-BORROWER’S NAME
Steven Spielberg

SOCIAL SECURITY NUMBER	PHONE	DOB	YRS. SCHOOL
111-22-1234	405-507-7000	12/18/1946	16

HUMAN RESPONSE:

I DIDN’T KNOW THAT STEVEN SPIELBERG AND I SHARED A BIRTHDAY!

COMPUTER RESPONSE:


informationco-borrowerborrower’snamecoborower’snamesocialsecuritynumberphonedobyrs.shoolsocialsecuritynumberphonedobyrs.schoolsocialsecuritynumberphonedobyrsschool222-11-5678405-555-4658161 11-22-1234405-507-7000012/18/194616

THE ANSWER USE FEATURES AND DATA TO CREATE INFORMATION AND UNDERSTANDING

STEP 1 IDENTIFY PAGES

WHAT IS A PAGE?

- A physical document page (paper, micrographics, etc.)
- A scanned image of a physical page
- An electronic document or file
- Structured / unstructured
- Changing / inconsistent

STEP 2 IDENTIFY FEATURES

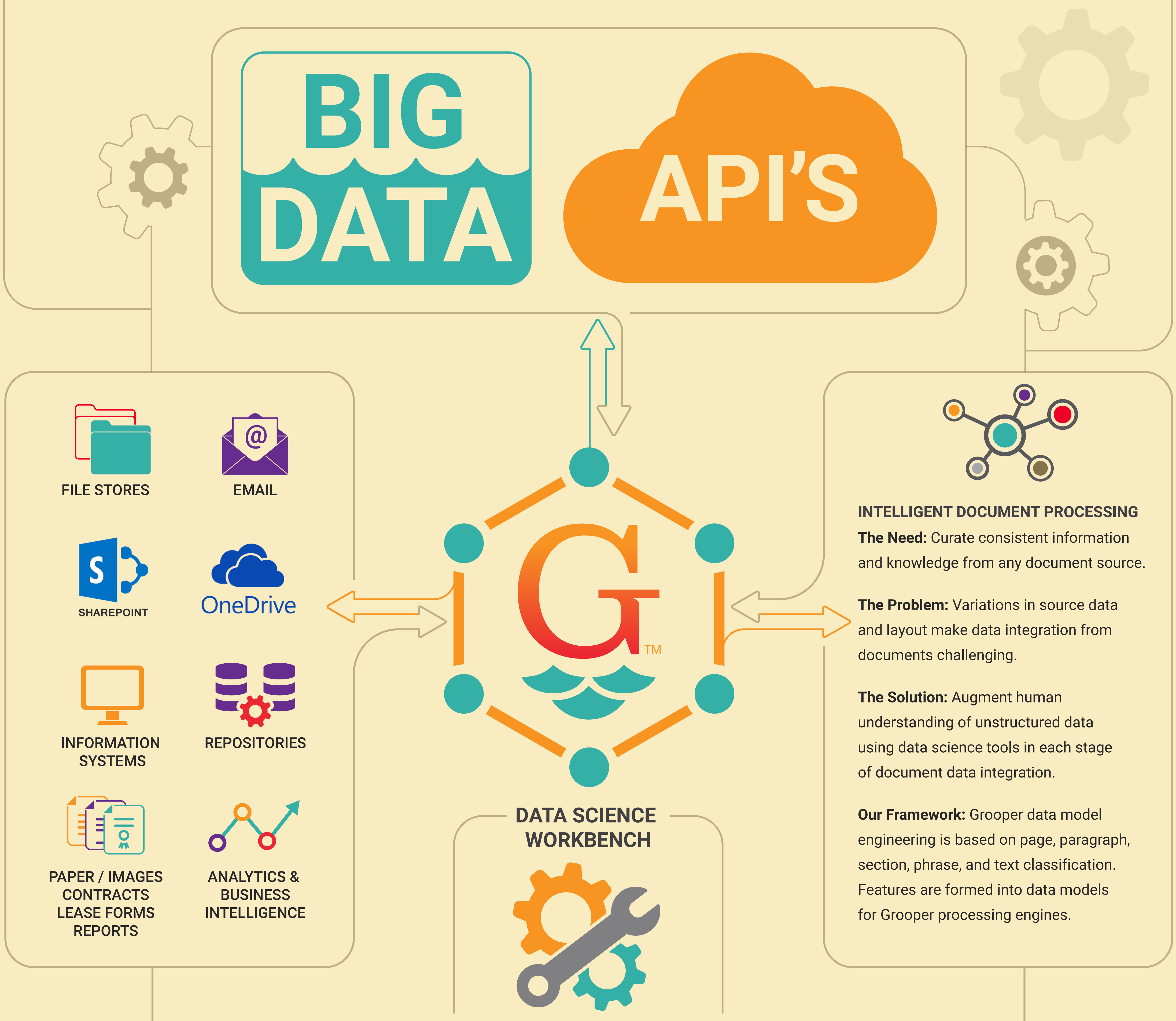
WHAT IS A FEATURE?

- Context
- Language
- Format
- Specific data elements
- Content tagging attribute
- Named entities
- N-grams
- Structure elements
- Visual elements

STEP 3 CURATE DATA

HOW?

Use features to make data models that produce synthetic understanding of data. Stream valuable data into workflows and downstream software applications.



INTELLIGENT DOCUMENT PROCESSING

The Need: Curate consistent information and knowledge from any document source.

The Problem: Variations in source data and layout make data integration from documents challenging.

The Solution: Augment human understanding of unstructured data using data science tools in each stage of document data integration.

Our Framework: Grooper data model engineering is based on page, paragraph, section, phrase, and text classification. Features are formed into data models for Grooper processing engines.

The working environment where decisions get made. Relationships between internal and external data elements are formed to engineer understanding and context.

NATURAL LANGUAGE PROCESSING
Allows data collection from free-form documents in which data can exist anywhere on a page.

SIGNATURE EXTRACTIONS
Determines the presence or absence of a signature with great precision by dropping out lines and other elements near the signature.

INDUSTRY-SPECIFIC LEXICONS
Matching and smart lookups on fields containing known values.

FUZZY REGULAR EXPRESSIONS
Matches data correctly despite OCR misreads using transparent weighting algorithms.

IMAGE PROCESSING
Use over 70 built-in image processing commands to create two document images. One for high-accuracy OCR and the other for pristine archival images.

TABLE EXTRACTION
Enables collection of full rows of data by utilizing fuzzy matching on individual columns.

LAYERED OCR
Many documents contain varying fonts, unaligned text, and handwriting. Collect more data with higher accuracy without the limitations of traditional OCR.

CLASSIFICATION
Lexical, rules-based, and visual classification options for transparent trainable document classification.

www.bisok.com

800-662-0777