

How to extract tabular data from scanned PDFs?

Proof-of-Concept made by Railsware



Agenda

Project overview

PDF challenges

Processing pipeline

Bonus: Heroku deployment

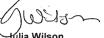


3i plc
Statement of financial position
as at 31 March 2017

	Notes	2017	2016
		£'000	£'000
Assets			
Non-current assets			
Investments in subsidiaries		10,000	10,000
Property, plant and equipment	8	3,397	4,068
Retirement benefit surplus	6	121,270	132,379
Amounts due from group undertakings	13	385,874	317,679
Total non-current assets		520,541	464,126
Current assets			
Trade and other receivables	11	10,978	26,196
Amounts due from group undertakings	13	900	2,084
Cash and cash equivalents		4,393	3,382
Total current assets		16,271	31,662
Total assets		536,812	495,788
Liabilities			
Non-current liabilities			
Trade and other payables	12	(113,341)	(82,286)
Amounts due to group undertakings	14	(114,739)	(117,332)
Provisions	15	(1,801)	(868)
Total non-current liabilities		(229,881)	(200,466)
Current liabilities			
Trade and other payables	12	(66,441)	(59,218)
Amounts due to group undertakings	14	(15,691)	(6,046)
Provisions	15	(2,925)	(5,734)
Total current liabilities		(84,857)	(70,998)
Total liabilities		(314,538)	(271,484)
Net assets		222,274	224,304
Equity			
Issued capital	9	110,000	110,000
Share based payment reserve		23,683	23,217
Retained earnings		88,591	91,087
Total equity		222,274	224,304

The notes on pages 13 to 24 form an integral part of these financial statements.

The financial statements have been approved and authorised for issue by the Board of Directors.


Julia Wilson
Director
Date: 16/5/17

	Notes	2017	2016
		£'000	£'000
Assets			
Non-current assets			
Investments in subsidiaries		10,000	10,000
Property, plant and equipment	8	3,397	4,068
Retirement benefit surplus	6	121,270	132,379
Amounts due from group undertakings	13	385,874	317,679
Total non-current assets		520,541	464,126
Current assets			
Trade and other receivables	11	10,978	26,196
Amounts due from group undertakings	13	900	2,084
Cash and cash equivalents		4,393	3,382
Total current assets		16,271	31,662
Total assets		536,812	495,788

	2017	2016
Assets	£'000	£'000
Total assets	536,812	495,788

Seems simple, right?



Well... it's not that
simple...





Who was the client?



Fin-tech company

UK-based

Project overview

More fin than tech

Looks for automation instead of
hiring people

Cares about trends on the market



What did the client want?



Ideal flow

Project overview

Excel

Company number



Financial data points
extracted for last two years

Excel



What was the goal?



Semi-automated extraction

Proof of concept

3 weeks

27 3 financial data points

500 100 reports



What did we do?



Research & MVP

Rapid prototyping with Rails

Project overview

Sidekiq-based processing pipeline

Cloud storage

3rd-party services



Let's talk about PDF



What is a PDF?

move cursor, draw, text box
PDF Reference, iText RUPS

```
((in%)) Tj
/CS0 cs
0.894 0.11 0.224 scn
/GS0 gs
/T1_2 1 Tf
6.3 0 0 6.3 135.13 690.3 Tm
[France, -3925.4, Eurozone, -2460.4,
Kingdom] TJ
15.932 1.159 Td
(United) Tj
7.173 -1.159 Td
(States) Tj
-0.206 1.159 Td
(United) Tj
```



Where are the tables then?



Tables are not there

...per se, but text elements are

Journal of example text

Page 2

As each word in the PDF documents is separated into its own text box, and there is no information available on which words together are intended to form sentences together, or which sentences would continue on the next line of text, this has to be done by the algorithms processing

which sentences would continue on the next line of text, this has to be done by the algorithms processing the elements on the page.

Data set	Precision	Recall	F-Score
1	80%	80%	80%
2	70%	75%	73%



Table in PDF is not
a duck - if it looks like
a table, it definitely
isn't one



How do we know where a table is?

Image processing

edge detection, position analysis
Tabula, Nurminen algorithm

PDF challenges

Consolidated statement of financial position

for the year ended 31 January



	State	2016 £m	2015 £m
Non-current assets			
Goodwill and other intangible assets	9	129.8	125.7
Property, plant and equipment	11	122	100
Investments in associates, joint ventures and associates	12	10	8
Deferred tax assets	7	52	31
Other receivables	14	—	21
		1,382	1,368
Current assets			
Inventories	13	3	3
Trade and other receivables	14	122	107
Contract receivable	—	—	1
Financial instruments, equivalents	15	106	89.2
		353	106
Total assets		1,725	1,368
Current liabilities			
Trade and other payables	16	(50.8)	(49.8)
Current tax payable	17	(7)	—
Provisions	18	(3)	(3)
		(53.8)	(50.8)
Non-current liabilities			
Borrowings and loans	17	(2,020)	(2,041)
Pension and other obligations	18	(21)	(15)
Defined benefit other post-employment scheme net liabilities	19	(2.6)	(2.4)
Provisions	20	(7)	(2)
Interest-free technical provisions	20	(1)	(1)
		(2,048)	(2,077)
Total liabilities		(5,781)	(5,013)
Equity			
Share capital	21	1	1
Share premium	22	52.9	20.0
Own shares	23	(22)	—
Share capital/valuation reserve	24	(1)	(2)
Shareholders' hedge reserve	25	(50)	(12)
Retained earnings	26	(20.3)	(26.6)
Total equity attributable to equity holders of the parent		(1,056)	(2,055)

Signed for and on behalf of the Board on 4 April 2016 by



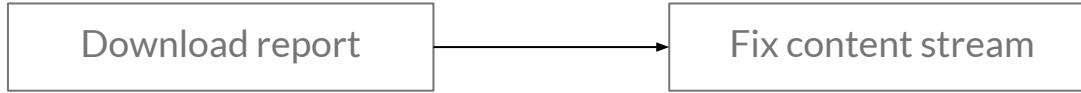


How difficult can it be?



Download report

Processing pipeline



Processing pipeline

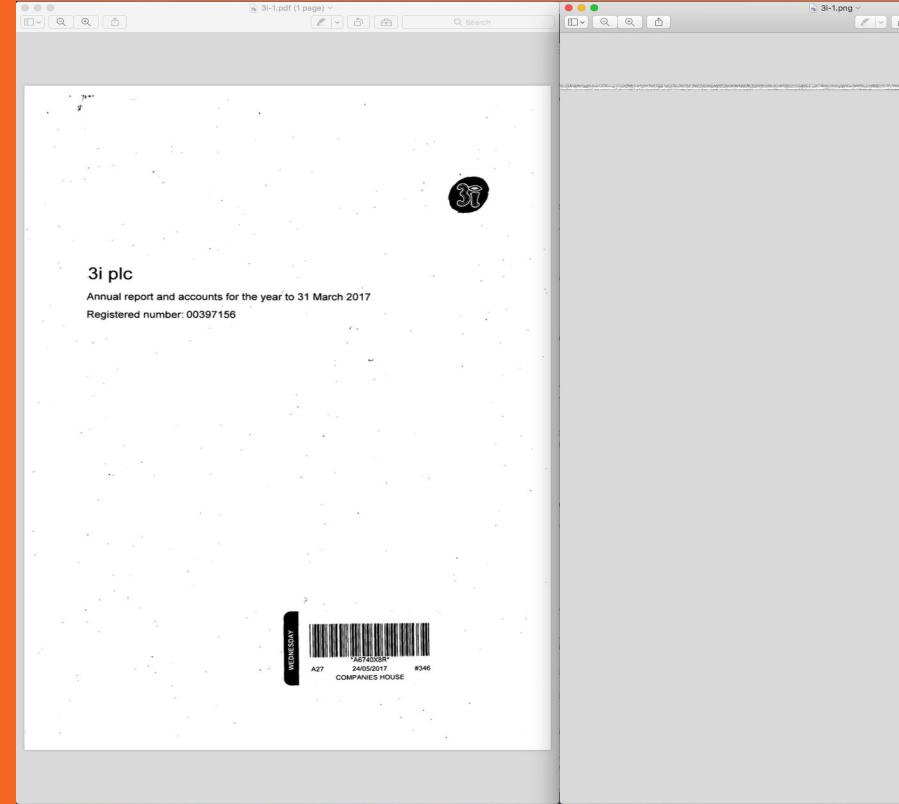


Why would you ever
have to fix a PDF?



Content stream ended up unexpectedly

Output may be incorrect





How to fix the broken files?



Split pages

iText

Processing pipeline

Fast

Reliable

Quite expensive volume-based
license



Maybe something else
would work, though



or maybe not...

PDF Reader

Grim

Docsplit

pdftk

Apache PDF Box



How do we cut price on iText, then?



Fix PDFs before splitting

iText + Apache PDF Box

Processing pipeline

Reduce file-based license cost by 80%

Use free software to split PDFs



Processing pipeline



How do we know which pages are relevant?



We care only about balance sheets

Processing pipeline

Close to the end of a report

$\frac{1}{3}$ 40% of pages



**Seems quite a lot, how can
we reduce that number?**



Keyword analysis of unstructured text

[Google Vision API](#)

Processing pipeline

\$0.0015 per page

Every single page can be processed

40% 8% pages get qualified



Why don't we use Google Vision API for more?



Google Vision API

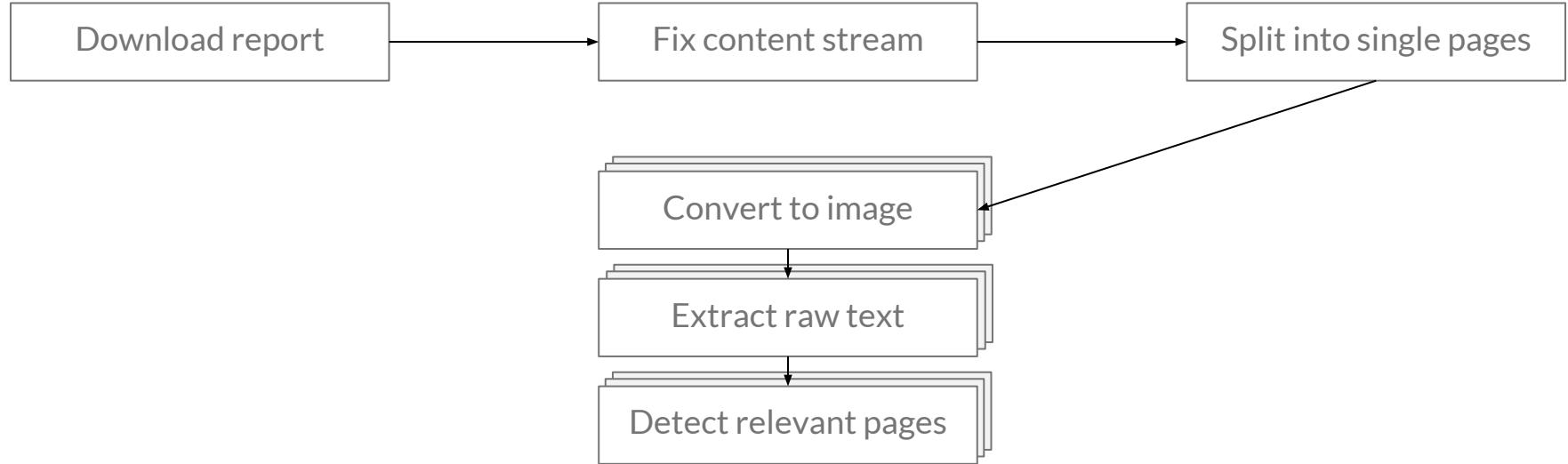
OCR != layout detection

Processing pipeline

Highly specialized

OCR

Labels



Processing pipeline



OK, but how to
extract tables?



Tables with text from images

[Abbyy OCR SDK](#)

Processing pipeline

OCR is never 100% correct

RTF format contains tables detected

\$0.03 per page - a bit pricy...



**It's cheaper to use
Google Vision API to
get keywords and pass
only the most relevant
pages to Abbyy**



How good is Abbyy?



~30 reports assessed manually

Missing headers,
Missing tables,
text distortion

Processing pipeline

Page Id	Page link	Score	Comments
22	report_pages/22	10	
501	report_pages/501	10	
649	report_pages/649	10	
3812	report_pages/3812	10	unit to be taken from year
6174	report_pages/6174	10	requires table splitting
6070	report_pages/6070	10	diverted layout - years in rows; assumptions in code
1858	report_pages/1858	9	missing decimal point - just space - for 2015
327	report_pages/327	8	years outside of table; to be restored based on context
1476	report_pages/1476	8	countries outside of table; UK and US; thankfully
2040	report_pages/2040	8	discount rate appears twice in the table; to be discounted
5057	report_pages/5057	8	header row and two rows with values: past and future
5760	report_pages/5760	8	page relevant; table detected; years outside of table
2749	report_pages/2749	7	columns to be grouped by unit; two discount rates
2483	report_pages/2483	7	countries outside of table; UK, US, other; thankful
4353	report_pages/4353	7	countries outside of table; UK, US, other; thankful
5967	report_pages/5967	6	remove column with garbage (apart from header)
856	report_pages/856	5	years outside of table; columns to be grouped by unit
2982	report_pages/2982	5	split countries in headers, merge countries under
1165	report_pages/1165	4	two cells (for 2016 and for 2015) combined together
4273	report_pages/4273	3	page relevant; table detected; UK group impossible
3301	report_pages/3301	2	page relevant; table not detected
3625	report_pages/3625	2	page relevant; table not detected
5464	report_pages/5464	2	page relevant; table not detected
5504	report_pages/5504	2	page relevant; table not detected
4124	report_pages/4124	2	page relevant; table not detected
		0	where is it in the document?
		0	where is it in the document?
		0	where is it in the document?



What if tables are missing?

angloamerican-149-searchable.pdf (1 page)

angloamerican-149.rtf

FINANCIAL STATEMENTS AND OTHER FINANCIAL INFORMATION NOTES TO THE FINANCIAL STATEMENTS

EMPLOYEE REMUNERATION

26. EMPLOYEE NUMBERS AND COSTS

The average number of employees, excluding contractors and associates' and joint ventures' employees, and including a proportionate share of employees within joint operations by segment was:

	2016	2015
De Beers	9	11
Platinum	45	48
Copper	4	5
Nickel	2	2
Nickel and Phosphate ¹⁰	2	2
Iron Ore and Manganese	7	10
Coal	10	11
Corporate and other	1	2
Total	80	91

¹⁰ Nickel and Phosphate was sold on 30 September 2016 (see note 30).

The average number of employees, excluding contractors and associates' and joint ventures' employees, and including a proportionate share of employees within joint operations, by principal location of employment was:

	2016	2015
South Africa	61	65
Other Africa	4	4
South America	9	10
North America	1	2
Australia and Asia	3	4
Europe	2	1
Total	80	91

Payroll costs in respect of the employees included in the tables above were:

	2016	2015
US\$ million		
Wages and salaries	3,107	3,798
Social security costs	110	132
Post employment benefits ¹¹	285	332
Share-based payments (note 26)	236	209
Total payroll costs	3,738	4,411

Recapitalisation:

	2016	2015
Less: employee costs capitalised	(256)	(310)
Less: employee costs included within special items	(144)	(202)
Employee costs included in operating costs	3,336	3,955

¹¹ Includes contributions to defined contribution pension and medical plans, current and past service costs related to defined benefit pension and medical plans and other benefits provided to certain employees during retirement (see note 27).

Key management

Key management personnel are those persons having authority and responsibility for planning, directing and controlling the activities of the Group, directly or indirectly, including any director (executive and non-executive) of the Group. Key management comprises members of the Board and the Group Management Committee.

Compensation for key management was as follows:

	2016	2015
US\$ million		
Salaries and short term employee benefits	15	22
Social security costs	3	4
Termination benefits	5	2
Long term employee benefits	3	3
Share-based payments	17	13
Total	47	44

Disclosure of directors' emoluments, pension entitlements, share options and long term incentive plan awards required by the Companies Act 2006 and those specified for audit by Regulation 11 and Schedule 8 of the Large and Medium-Sized Companies and Groups (Accounts and Reports) Regulations 2008 are included in the Remuneration report.

Tables are there... most of the time



How to read missing tables?



Convert RTF to HTML

online-convert.com

Processing pipeline

Most of the tools can handle single type of tables

Abbyy uses more than one



How to fix other problems with tables?



Pre-analysis cleanup

Processing pipeline

Try to correct common OCR errors

Insert missing headers with years

Split tables by header detection



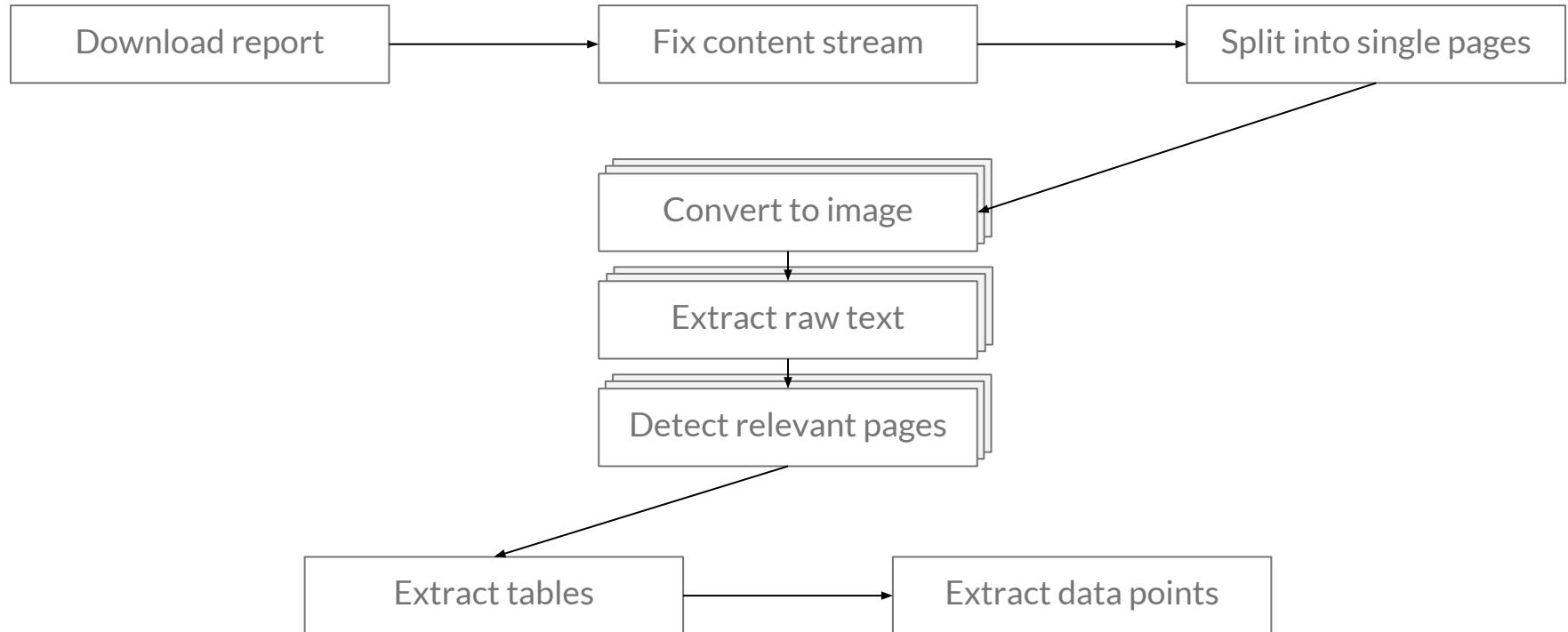
Where is the value though?



Extraction based on a heat map

Total assets, UK, 2017

Balance sheet	2017	2017	2016	2016
£'000	UK	US	UK	US
Total assets	3,456	6,543	2,345	5,432
Total liabilities	(2,109)	(3,210)	(1,098)	(2,109)



Processing pipeline



How reliable is the extracted value?



Probability assessment

3 models tested

Processing pipeline

Weighted keywords

Value formatting and range

Other similarly good results



Wait, so it might be wrong...



Probability is calculated
to bring attention to
potential errors as
extraction might be
wrong



What's the error rate?

60%

Error rate



Meaningful data
extraction from
images is hard for
computers



But it's easy
for humans...



3i plc

No. 00397156

Statement of financial position

as at 31 March 2017

	Notes	2017 £'000	2016 £'000
Assets			
Non-current assets			
Investments in subsidiaries		10,000	10,000
Property, plant and equipment	8	3,397	4,068
Retirement benefit surplus	6	121,270	132,379
Amounts due from group undertakings	13	385,874	317,679
Total non-current assets		520,541	464,126
Current assets			
Trade and other receivables	11	10,978	26,196
Amounts due from group undertakings	13	900	2,084
Cash and cash equivalents		4,193	3,382
Total current assets		16,071	31,662
Total assets		536,812	495,788
Liabilities			
Non-current liabilities			
Trade and other payables	12	(113,341)	(82,286)
Amounts due to group undertakings	14	(114,739)	(117,332)
Provisions	15	(1,801)	(868)
Total non-current liabilities		(229,881)	(200,486)
Current liabilities			
Trade and other payables	12	(66,441)	(59,218)
Amounts due to group undertakings	14	(15,691)	(6,046)
Provisions	15	(2,525)	(5,734)



How to automate this?



Mechanical Turk

Scalable workforce from Amazon

Comparable cost per page to Abbyy

Pay only for approved answers

Automated cross-check policy



How does it work?

You must accept this Requester's HIT before working on it. Learn more

iframe with HTML question or external question

Type the text, very carefully.

Warning

Your HIT won't be approved if you don't stick to the following instructions

Instructions

- If a € symbol is present and not on your keyboard, copy/paste it from this line: €
- Other special characters that you may need: % & * ° < >
- Be careful, "ABCD" and "abcd" are different.
- If you cannot read a text. NEVER EVER write "I can't read". If you do write it, your HIT will be rejected and we may block you.
Two options:
 - A checkbox I cannot read this text is available: just tick it!
 - The checkbox is not available: leave the text input blank, it won't be penalized.
- When there are 1, 2 or more spaces, type just 1 space.
- If no image is loading, do NOT accept this HIT.

More instructions here

SENGO CORSE X18 333 5,25 EUR 1

PERFECT E CHAT SEC S 5,95 EUR 3

TOP BUDGET CREME FR. 1,07 EUR 1

Submit



What's the error rate?



60%

Error rate for
Automated extraction

18%

Error rate for
Mechanical Turk



Where do we go from
here?



Next steps

RW Labs

More reliable probability model

Limit relevant pages

Add connector for Mechanical Turk

...

AI?



What have we learned?



Lessons learned

Build measure learn

Automation is a part of the picture

Human supervision is necessary

Mechanical Turk - valid alternative

Start simple and be pragmatic

Cost estimation is important for development



BONUS: How to configure Heroku for all of that?



Buildpacks

[Buildpack API](#)

Heroku deployment

Ruby

Node.js

Java

Google Cloud SDK



Why do you need Java?



java -jar ...

iText, Apache PDF Box

Heroku deployment

Requires pom.xml

```
<project NAMESPACES_HERE>
<modelVersion>4.0.0</modelVersion>
<groupId>GROUP_HERE</groupId>
<artifactId>ARTIFACT_HERE</artifactId>
<version>VERSION_HERE</version>
</project>
```

And some memory...



What about
Google Cloud SDK?



Google Cloud SDK

[heroku-google-cloud-buildpack](#)

Heroku deployment

Download and install SDK

Build JSON with credentials from
ENV variables

Activate service account using
profile.d script



Questions?



Thank you