Building an image processing pipeline in Python

Franck Chastagnol, PyCon 2013

Agenda

- Introduction
- Architecture
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling / re-processing
- Q&A

Introduction

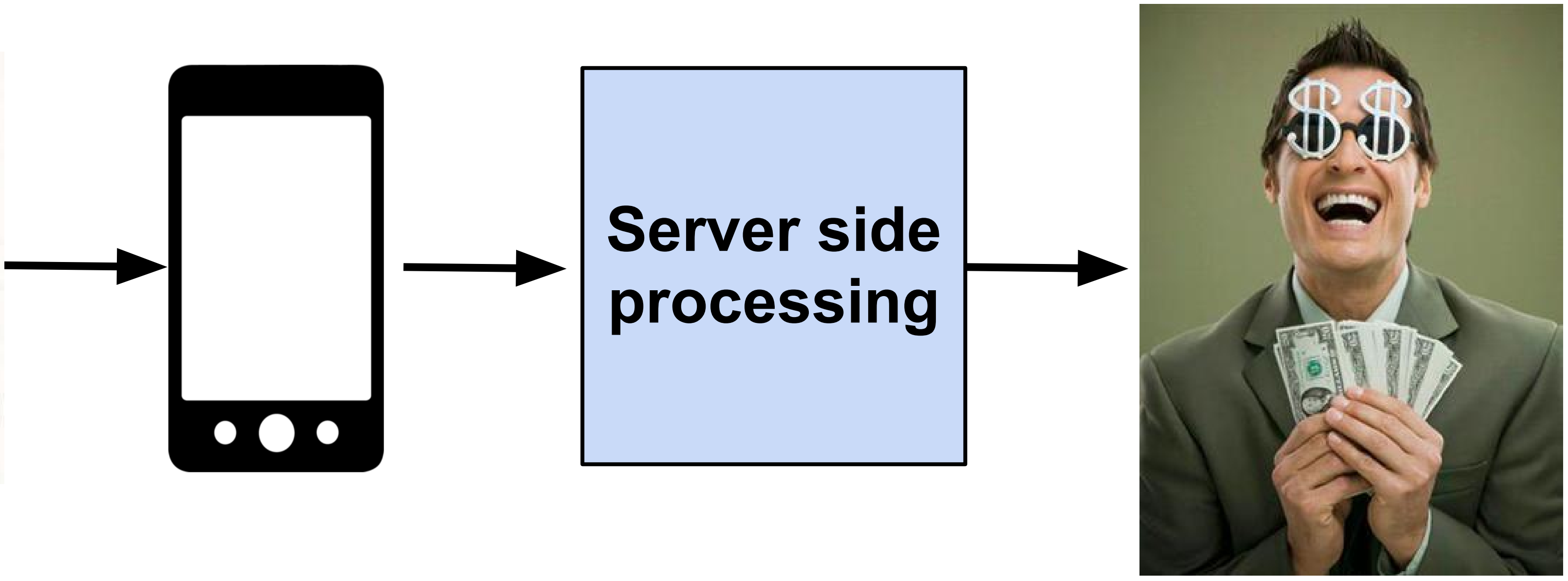
- Background

- Today's case study
  - Image processing pipeline built for Endorse.com

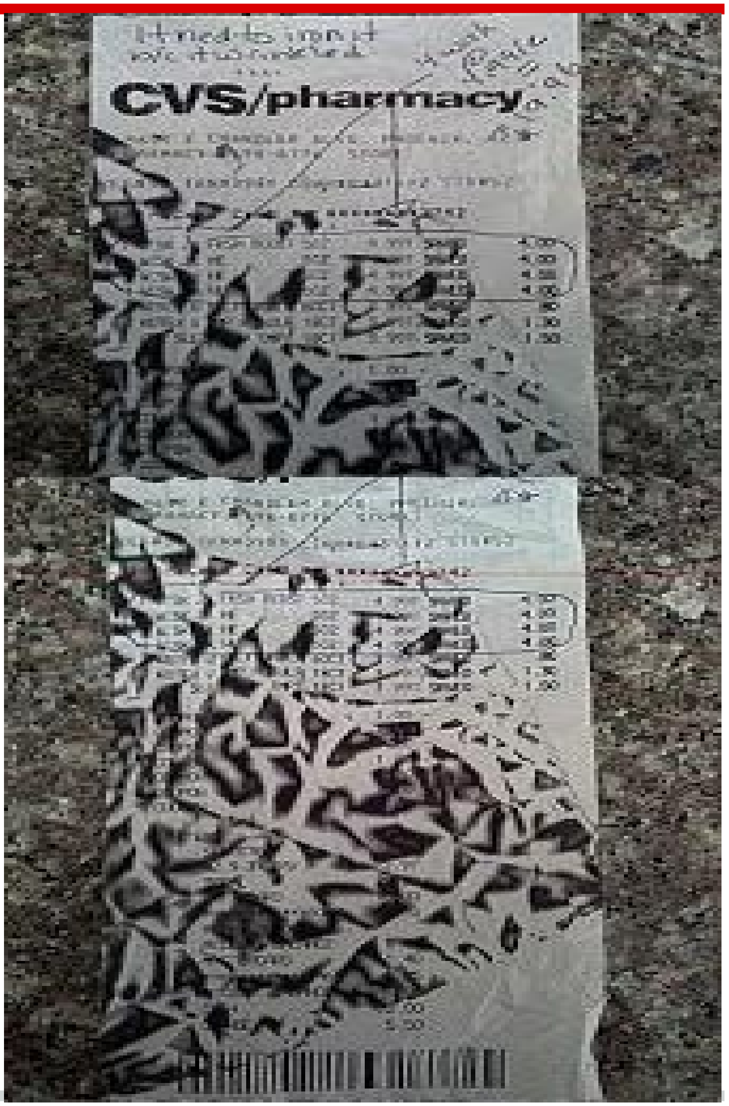
Endorse.com mobile app

- Reward for buying specific brand products
- Shop anywhere, upload pic of receipt, get $$

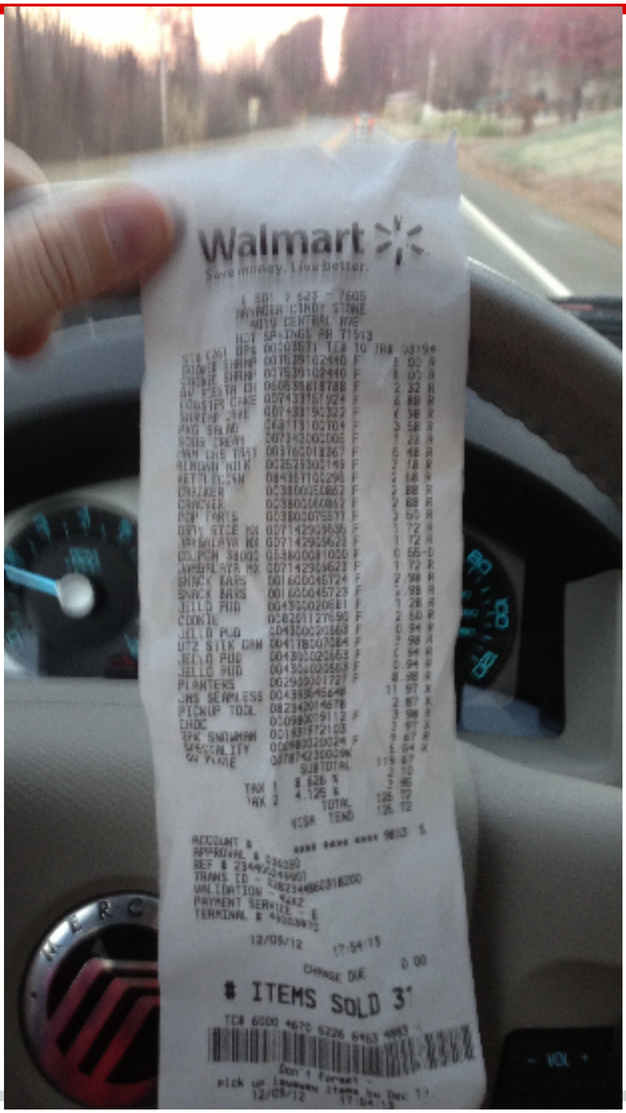
```
WAL*MART
ALWAYS LOW PRICES.
Always

WE SELL FOR LESS
MANAGER DAVID KINMAN
( 816 ) 224 - 4800
ST# 4068 OP# 00002825 TE# 39 TR# 03366
HAMBCOMBO     007874235066K      3.17 X
SLC OF CHEES 007874223757K       0.20 X
 ** VOIDED ENTRY **
HAMBCOMBO     007874235055K      3.17-X
 ** VOIDED ENTRY **
SLC OF CHEES 007874223757K       0.20-X
                    SUBTOTAL     0.00
CHEESEBURGER 007874223303K       2.17 X
MCC DRINK     007874223281K      0.98 X
                    SUBTOTAL     3.15
         TAX 1   3.000 %         0.09
         TAX 3   3.475 %         0.11
                       TOTAL     3.35
                   CASH  TEND   20.35
                  CHANGE DUE    17.00

4-21 ORDER #666

TC# 3680 8362 6229 1132 7028

Ready For the Weekend?
Check Out Lawn and Garden
04/21/05       12:26:25
```



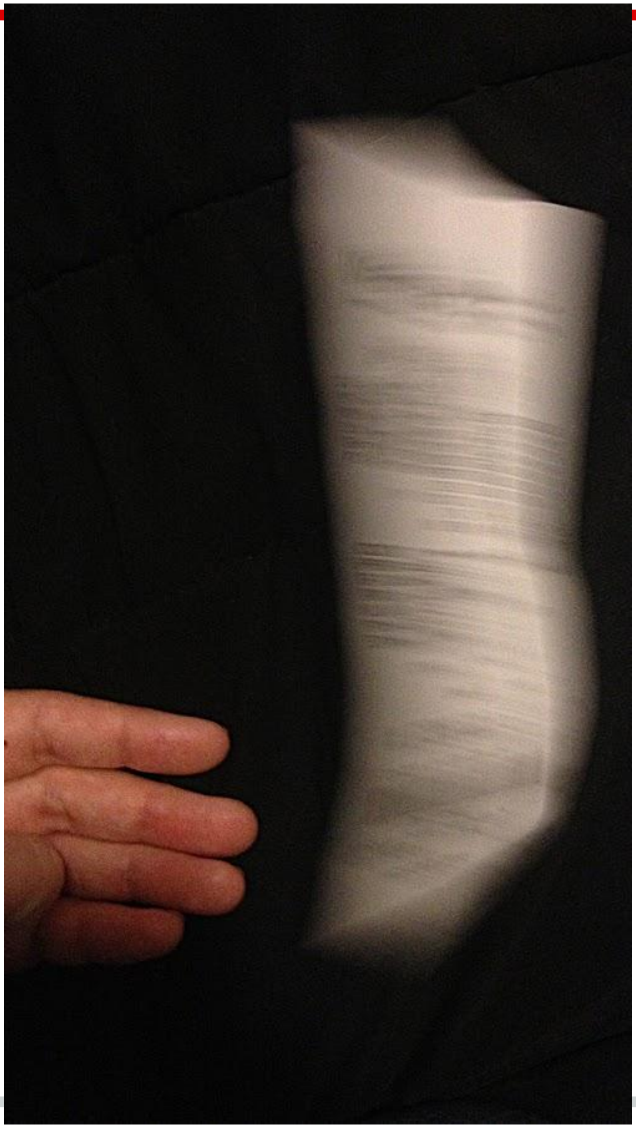
Pics of receipts are... fun ! (1)



Pics of receipts are... fun ! (2)



```
Walmart
Save money. Live better.

      [illegible] 623 - 7605
     MANAGER CINDY STONE
       4019 CENTRAL AVE
     HOT SPRINGS AR 71913
ST# [illegible] OP# 00[illegible]3571 TE# 10 TR# [illegible]194
[illegible] SHAMP  007639102440 F        [illegible] 00 R
[illegible] SHAMP  007539102440 F        [illegible] 05 R
[illegible] FIESTA [illegible]  060538818788 F        2 32 R
LOBSTER CAKE  007433151924 F        6 88 R
SHRIMP CAKE  007433190322 F        [illegible] 98 R
PKG SALAD    [illegible]  F        3 58 R
SOUR CREAM   007342002005 F        [illegible] 22 R
[illegible]  003160018367 F        6 48 R
ALMOND MILK  002529300145 F        [illegible] 18 R
KETTLE [illegible]  084351100295 F        [illegible] 58 R
CRACKER      003800050862 F        2 88 R
CRACKER      003800060862 F        2 88 R
POP TARTS    003800075511 F        3 50 R
[illegible] RICE MX  007142909536 F        1 72 R
JAMBALAYA MX  007142909623 F        1 72 R
COUPON 38000  053800081000 F        0 55-D
JAMBALAYA MX  007142909623 F        1 72 R
SNACK BARS   001600045724 F        2 98 R
SNACK BARS   001600045723 F        [illegible] 98 R
JELLO PUD    004300020681 F        1 28 R
COOKIE       008201127690 F        2 50 R
JELLO PUD    004300020663 F        [illegible] 94 R
UTZ STIK CAN  004178007084 F        [illegible] 98 R
JELLO PUD    004300020653 F        [illegible] 94 R
JELLO PUD    004300020563 F        [illegible] 94 R
PLANTERS     002900001727 F        8 98 R
JMS SEAMLESS  004339046640          11 97 X
PICKUP TOOL  082342014678           2 87 X
CHOC         000980019112 F        3 98 X
3PK SNOWMAN  001981972103           [illegible] 97 X
SPECIALITY   000980020024 F        [illegible] 67 X
GV [illegible]  007874230029K          [illegible] 64 X
                  SUBTOTAL      115 67
      TAX 1   8.626 %             2 10
      TAX 2   4.125 %             7 95
                  TOTAL         125 72
              VISA  TEND        125 72

ACCOUNT #      **** **** **** 9833  S
APPROVAL # 536280
REF # 234400249901
TRANS ID - [illegible]
VALIDATION - [illegible]
PAYMENT SERVICE - E
TERMINAL # [illegible]

       12/09/12        17:54:15

               CHANGE DUE      0 00

   # ITEMS SOLD 3[illegible]

   TC# 6000 4670 6226 6453 4883

          Don't forget -
  pick up layaway items by Dec 11
       12/09/12        17:54:15
```



Pics of shopping receipts are... challenging to process !

- Taken in various environment, lighting
- Resolution varies depending on device
- Quality of receipt printers varies greatly
- It is not english
- Diff. format, no universal UPC / shortnames

Agenda

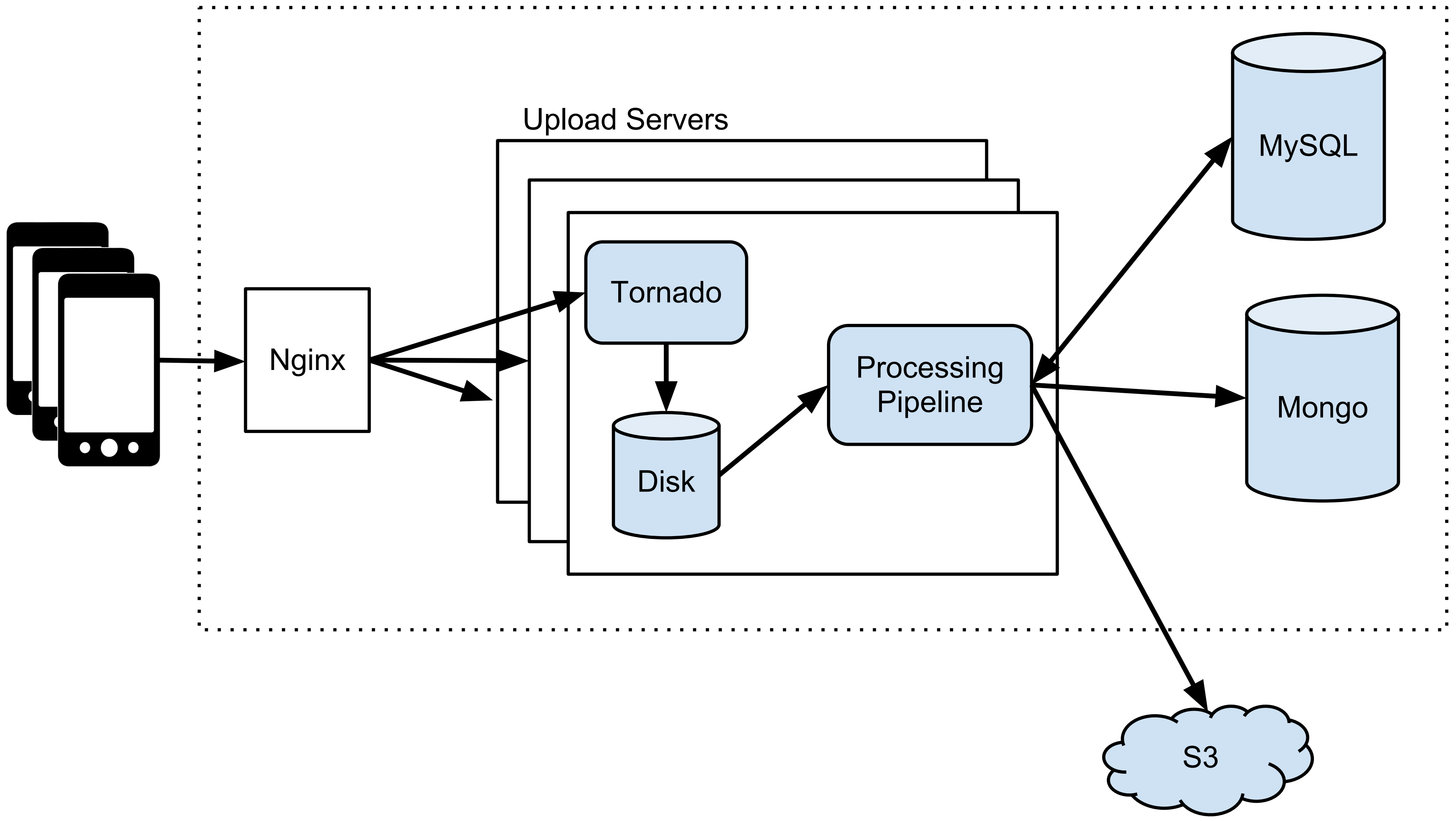
- Introduction
- Architecture
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling / re-processing
- Q&A

Technologies

- Common
  - Server Central cloud
  - Linux (ubuntu)
  - Nginx load balancer
  - Tornado app server
  - Python 2.7
  - Redis
  - S3 storage
- Web
  - Mako templates
  - MySQL

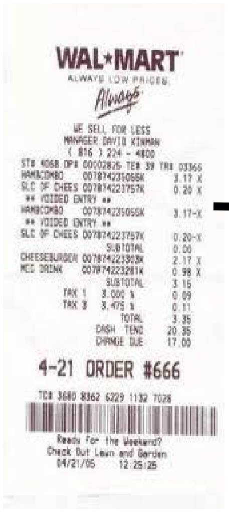
- Receipt processing
  - OpenCV
  - NumPy
  - IMagick
  - Tesseract OCR
- Data mining
  - MongoDB
  - Hadoop

System diagram



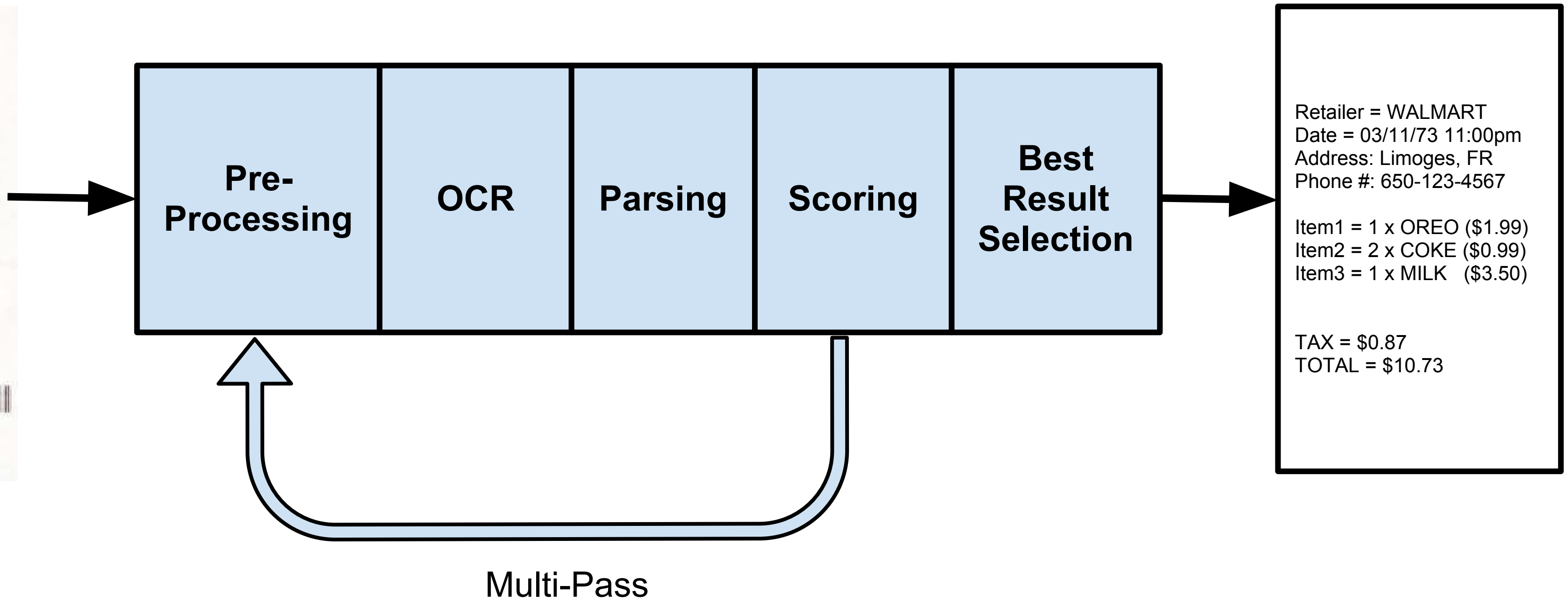
Pipeline

Receipt Image



```
WAL*MART
ALWAYS LOW PRICES
Always
WE SELL FOR LESS
MANAGER DAVID KINMAN
( 816 ) 224 - 4800
ST# 4068 OP# 00002825 TE# 39 TR# 03366
HAMBCOMBO     007874235055K     3.17 X
SLC OF CHEES 007874223757K      0.20 X
 ** VOIDED ENTRY **
HAMBCOMBO     007874235055K     3.17-X
 ** VOIDED ENTRY **
SLC OF CHEES 007874223757K      0.20-X
                   SUBTOTAL     0.00
CHEESEBURGER 007874223303K      2.17 X
MED DRINK     007874223281K     0.98 X
                   SUBTOTAL     3.15
        TAX 1    3.000 %        0.09
        TAX 3    3.475 %        0.11
                      TOTAL     3.35
                 CASH  TEND    20.35
                 CHANGE DUE    17.00

  4-21 ORDER #666

  TC# 3680 8362 6229 1132 7028
      Ready for the Weekend?
    Check Out Lawn and Garden
      04/21/05      12:26:25
```

Structured Doc



Retailer = WALMART
Date = 03/11/73 11:00pm
Address: Limoges, FR
Phone #: 650-123-4567

Item1 = 1 x OREO ($1.99)
Item2 = 2 x COKE ($0.99)
Item3 = 1 x MILK ($3.50)

TAX = $0.87
TOTAL = $10.73

Agenda

- Introduction
- Architecture
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling / re-processing
- Q&A

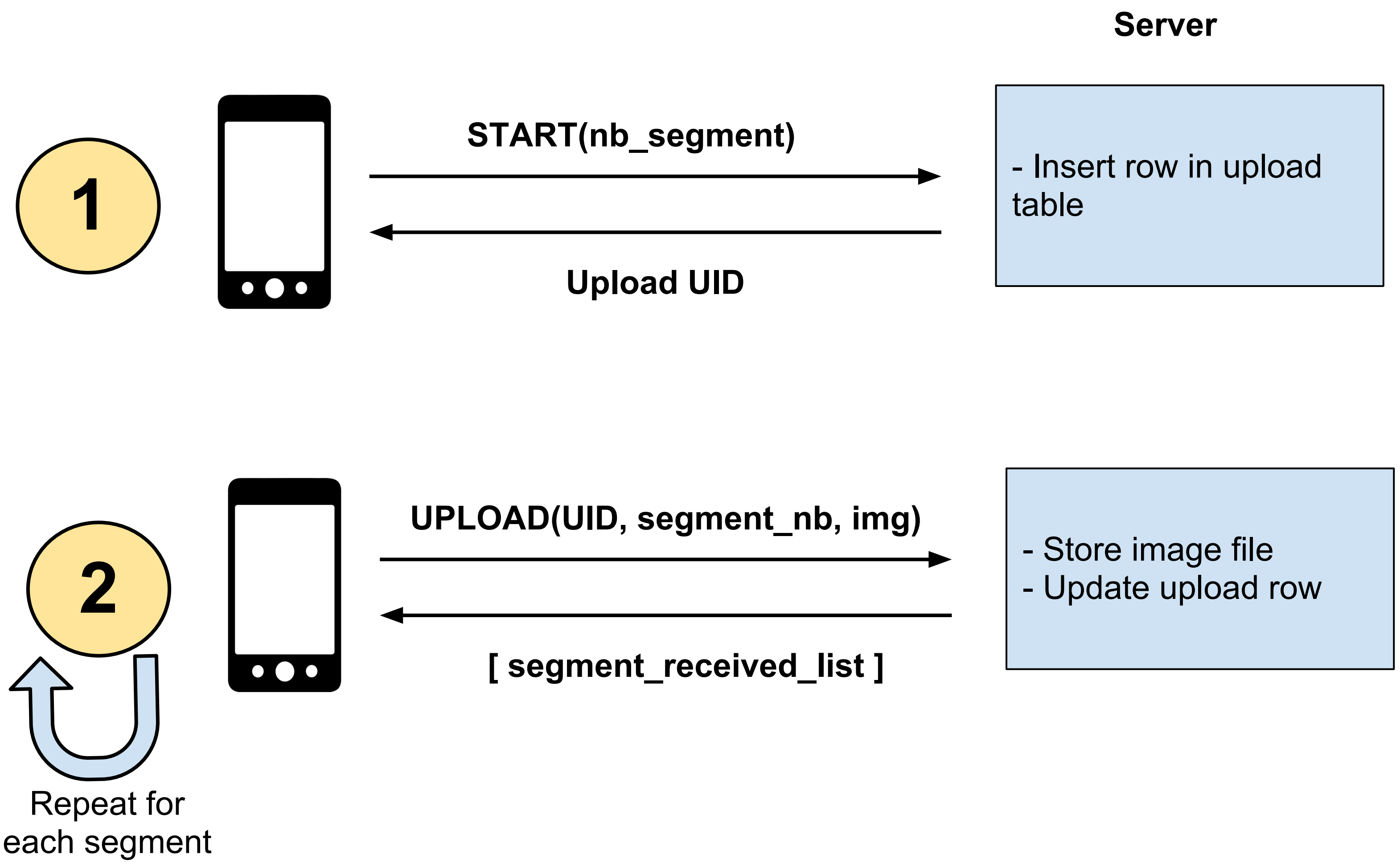
Mobile uploads

- Images are not small: ~1MB per segment

- Mobile data connection
  - can be spotty
  - upload bandwidth varies

- Ensuring high upload success rate:
  - App capable of re-trying in background
  - Simple and resumable APIs

Upload workflow



Upload - scalability

- Nginx
  - sticky session module

- Tornado writes img files to local disk

- Job picks up img files once upload finished
  - Store originals in S3
  - Run pipeline

Agenda

- Introduction
- Architecture
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling / re-processing
- Q&A

But why ??

- OCR is a solved problem... for book scans
- Clean b&w 300 dpi images of book pages scanned under perfect conditions
  => recognition rate = 95% to 99%
- Wrinkled paper, bad quality print, inconsistent lighting, noise, angle, etc...
  => recognition rate = ~25% or less

Pre-processing steps

- From color to b&w
  - unblur / sharpen filters
  - un-highlight color regions
  - adaptive thresholding

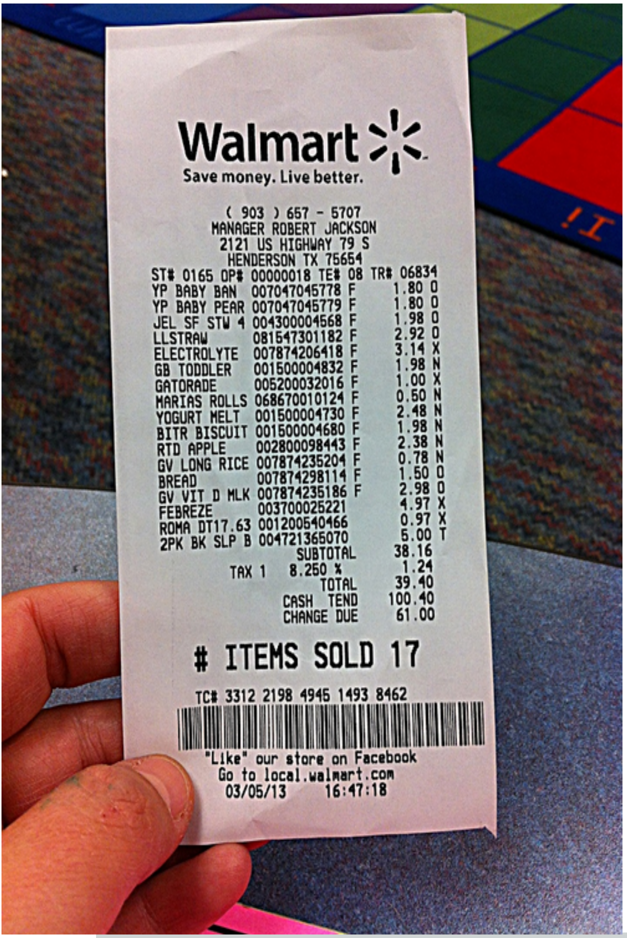
- Cropping
  - The carpet problem

- Extracting lines
  - OCR does poorly on non-straight lines
  - Lines recognition

=> OpenCV + Numpy is great

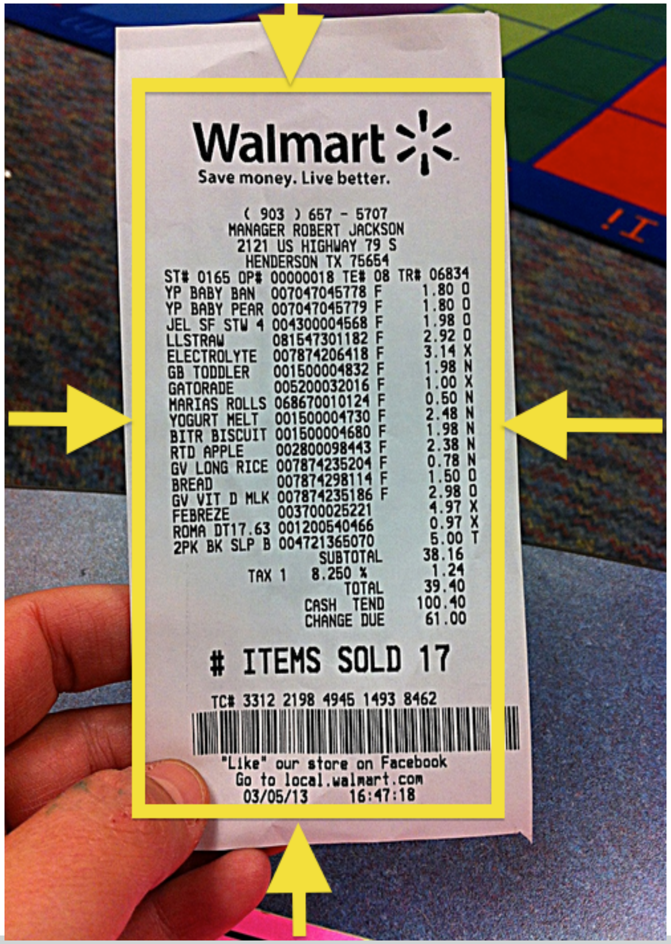
Image pre-processing example

Original

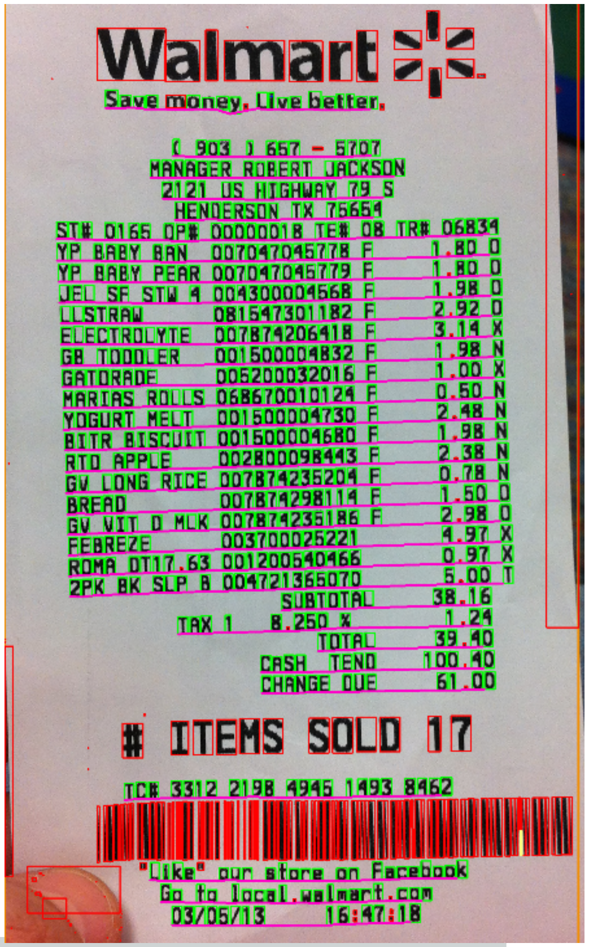


```
Walmart
Save money. Live better.

( 903 ) 657 - 5707
MANAGER ROBERT JACKSON
2121 US HIGHWAY 79 S
HENDERSON TX 75654
ST# 0165 OP# 00000018 TE# 08 TR# 06834
YP BABY BAN  007047045778 F     1.80 O
YP BABY PEAR 007047045779 F     1.80 O
JEL SF STW 4 004300004568 F     1.98 O
LLSTRAW      081547301182 F     2.92 O
ELECTROLYTE  007874206418 F     3.14 X
GB TODDLER   001500004832 F     1.98 N
GATORADE     005200032016 F     1.00 X
MARIAS ROLLS 068670010124 F     0.50 N
YOGURT MELT  001500004730 F     2.48 N
BITR BISCUIT 001500004680 F     1.98 N
RTD APPLE    002800098443 F     2.38 N
GV LONG RICE 007874235204 F     0.78 N
BREAD        007874298114 F     1.50 O
GV VIT D MLK 007874235186 F     2.98 O
FEBREZE      003700025221       4.97 X
ROMA DT17.63 001200540466       0.97 X
2PK BK SLP B 004721365070       5.00 T
                   SUBTOTAL    38.16
      TAX 1   8.250 %           1.24
                      TOTAL    39.40
             CASH  TEND       100.40
             CHANGE DUE        61.00

# ITEMS SOLD 17

TC# 3312 2198 4945 1493 8462
"Like" our store on Facebook
Go to local.walmart.com
03/05/13        16:47:18
```

Cropping



Lines extract.



Agenda

- Introduction
- Architecture
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling / re-processing
- Q&A

Tesseract

- Tesseract
  - Open source
  - Started at HP in the 90s
  - Google uses it for Book scan project
  - C++ core engine, APIs
  - Python bindings

OCR Training

- Shopping receipt fonts are not standard !
  - Training process is no fun
    - scanned various receipt types
    - extracted each letter from alphabet
    - generated synthetic receipts used for training
- Shopping receipts are not english !
  - OCR uses dictionaries to improve its output quality:
    - words dictionary with frequency in language
    - word pairs probability
    - punctuations / non alpha character rules

Agenda

- Introduction
- Architecture
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling / re-processing
- Q&A

You got text, now what ?

```
( 903 ) 657 - 5707
MANAGER R0BERT JACKSON
2121 US HIGHWAY 79 S
HENDERSON TX 75654
ST# 0165 DP# 00000018 TE# 08 TR# 06834
ELECTROLYTE 007874206418 F 3.14 X
GATORADE 005200032016 F 1.00 X
YOGURT MELT 001500004730 F 2.48 N
RTD APPLE 002800098443 F 2.38 N
BREAD 007874298114 F 1.50 0
FFBRFZE 003700025221 4.97 X
2PK BK SLP B 004721365070 5.00 T
SVBT0TAL 38. 16
TAX1 8.250 X 1.24
TOTAL 39 .40
CASH TEND 100.40
CH8NGE DVE 61.00
TC# 3312 2198 4945 1493 8462
03/05/13 16:47.18
```

- Parser
  - In: Text
  - Out: Structured doc
- Receipt
  - Store
  - List
    - Items (UPC, price)
  - SubTotal
  - Taxes
  - Total

Regex = headache

- Wide variety of mistakes in OCR output makes using regex hard / impossible

- Levenshtein distance is your friend
    - Similarity score between 2 strings (e.g. nb edits)
    - Pure Python implementation is slow. C lib + Python bindings faster

- "fuzzy matcher"
    - **Pattern**: "%s TAX (%d.d%%) = $%d.%d ON $%d.%d"
    - **Input**: "CA T8X (8.0%) = $4.00 ON $50.00
    - **Output**: Score = 1 (e.g. 1 edit)

Extracting + storing structured data

- Shopping receipts come in a variety of format
  - Specific parsers for most common formats
  - Generic parser for others
  - Store document in Mongo

- Mongo DB benefits
  - schemaless
  - map-reduce capabilities makes it a scalable data-mining solution

Agenda

- Introduction
- Workflow
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling/re-processing
- Q&A

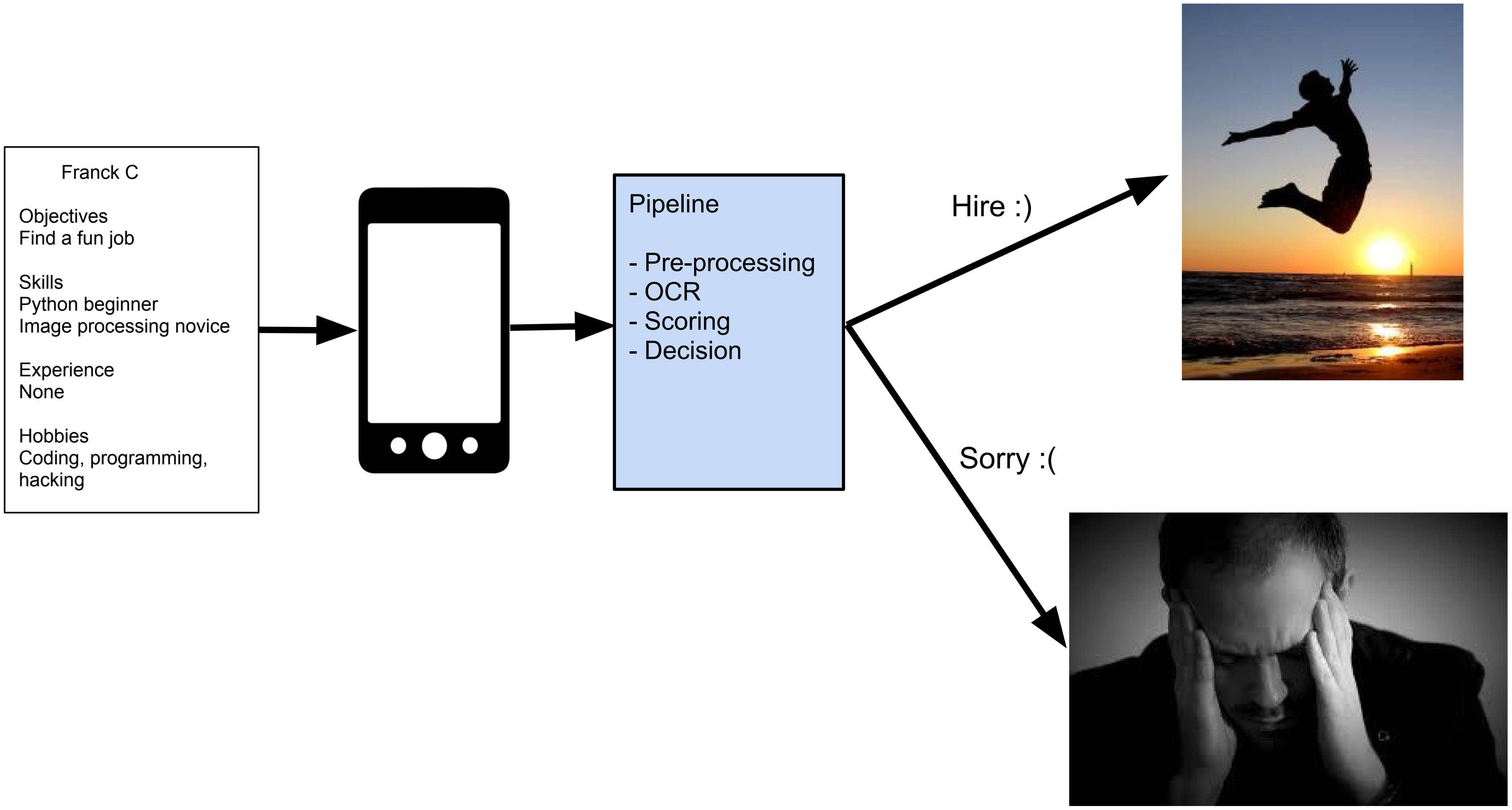
Breakage will happen

- You are a great coder, but...
  - Your co-workers ? interns ?
  - Pipeline will crash, servers will die
- How to get some good sleep at night ?
  - Good strategy for storing originals
  - Support re-runs

Agenda

- Introduction
- Workflow
- Upload
- Image pre-processing
- OCR
- Structured data extraction
- Error handling/re-processing
- Q&A

Hiring pipeline (in Python)


Franck C

Objectives
Find a fun job

Skills
Python beginner
Image processing novice

Experience
None

Hobbies
Coding, programming, hacking

Pipeline

- Pre-processing
- OCR
- Scoring
- Decision

Questions &
(hopefully some) Answers