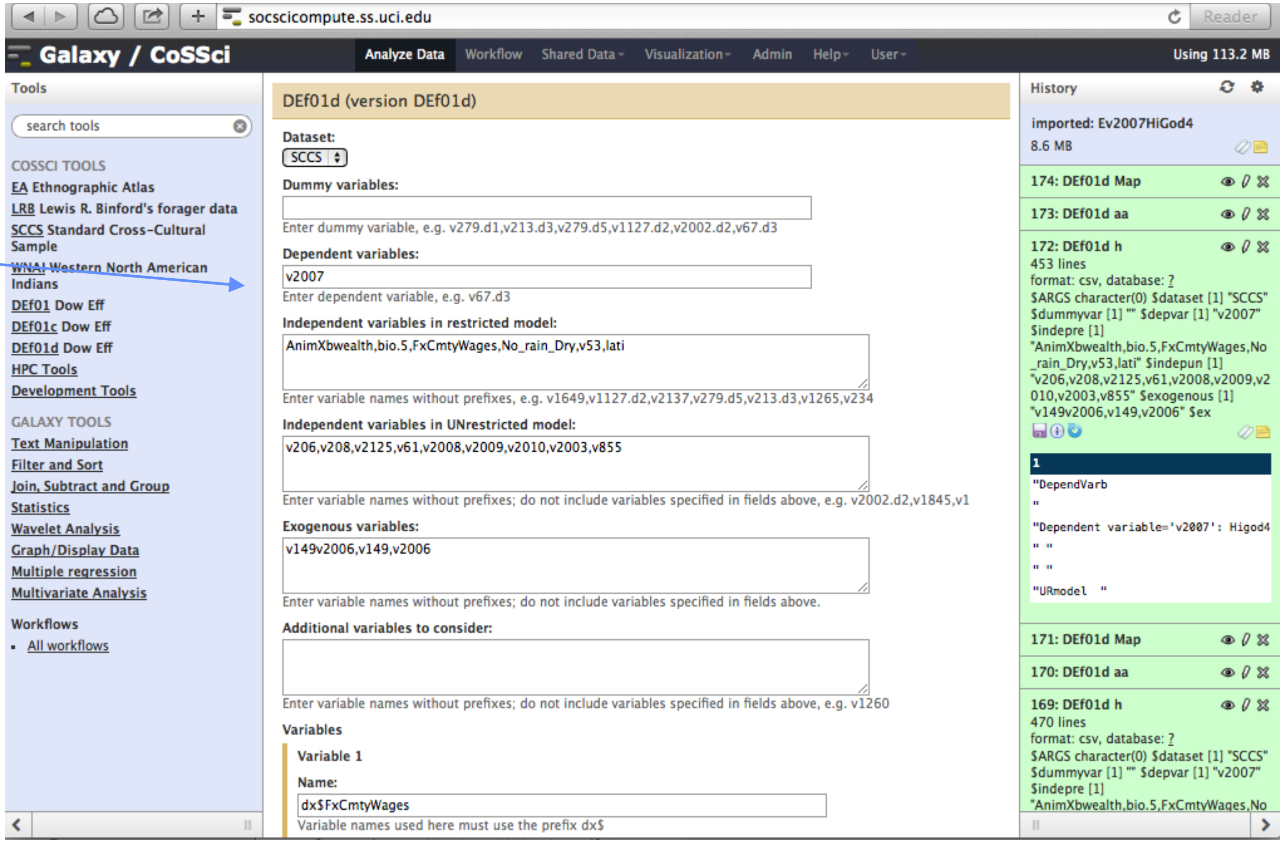
Other ML related activites

Complex Social Science Gateway – a tool for cross-cultural analysis in R

Select dataset,
Select varialbes,
Submit analysis

http://socscicompute.ss.uci.edu/

(but moving soon)



DEf01d (version DEf01d)

Dataset: SCCS

Dummy variables:

Enter dummy variable, e.g. v279.d1,v213.d3,v279.d5,v1127.d2,v2002.d2,v67.d3

Dependent variables: v2007

Enter dependent variable, e.g. v67.d3

Independent variables in restricted model: AnimXbwealth,bio.5,FxCmtyWages,No_rain_Dry,v53,lati

Enter variable names without prefixes, e.g. v1649,v1127.d2,v2137,v279.d5,v213.d3,v1265,v234

Independent variables in UNrestricted model: v206,v208,v2125,v61,v2008,v2009,v2010,v2003,v855

Enter variable names without prefixes; do not include variables specified in fields above, e.g. v2002.d2,v1845,v1

Exogenous variables: v149v2006,v149,v2006

Enter variable names without prefixes; do not include variables specified in fields above.

Additional variables to consider:

Enter variable names without prefixes; do not include variables specified in fields above, e.g. v1260

Variables

Variable 1

Name: dx$FxCmtyWages

Variable names used here must use the prefix dx$

imported: Ev2007HiGod4
8.6 MB

174: DEf01d Map

173: DEf01d aa

172: DEf01d h
453 lines
format: csv, database: ?
$ARGS character(0) $dataset [1] "SCCS" $dummyvar [1] "" $depvar [1] "v2007" $indepre [1] "AnimXbwealth,bio.5,FxCmtyWages,No_rain_Dry,v53,lati" $indepun [1] "v206,v208,v2125,v61,v2008,v2009,v2010,v2003,v855" $exogenous [1] "v149v2006,v149,v2006" $ex

```
1
"DependVarb
"
"Dependent variable='v2007': Higod4
" "
" "
"URmodel  "
```

171: DEf01d Map

170: DEf01d aa

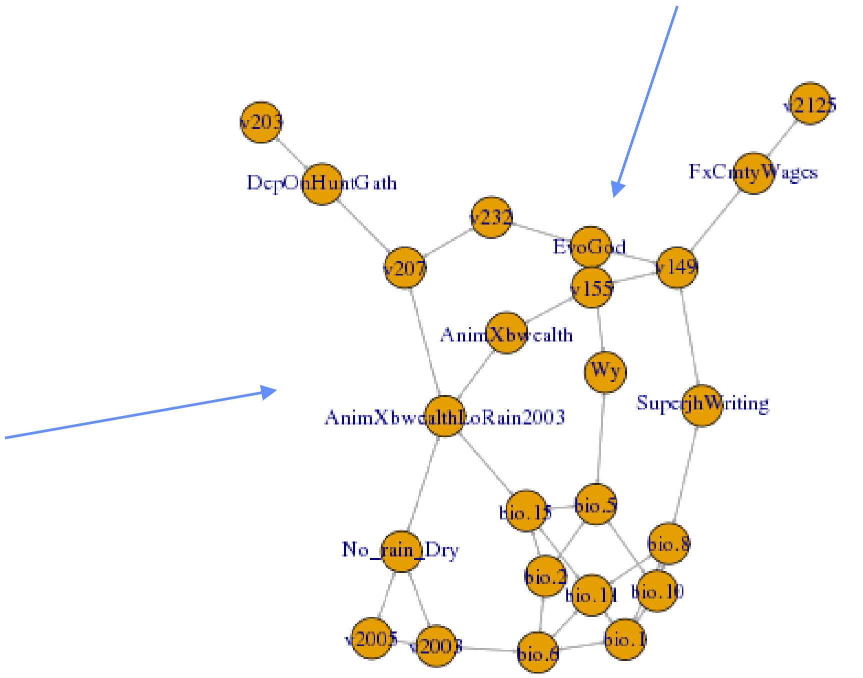
169: DEf01d h
470 lines
format: csv, database: ?
$ARGS character(0) $dataset [1] "SCCS" $dummyvar [1] "" $depvar [1] "v2007" $indepre [1] "AnimXbwealth,bio.5,FxCmtyWages,No

R Analysis options

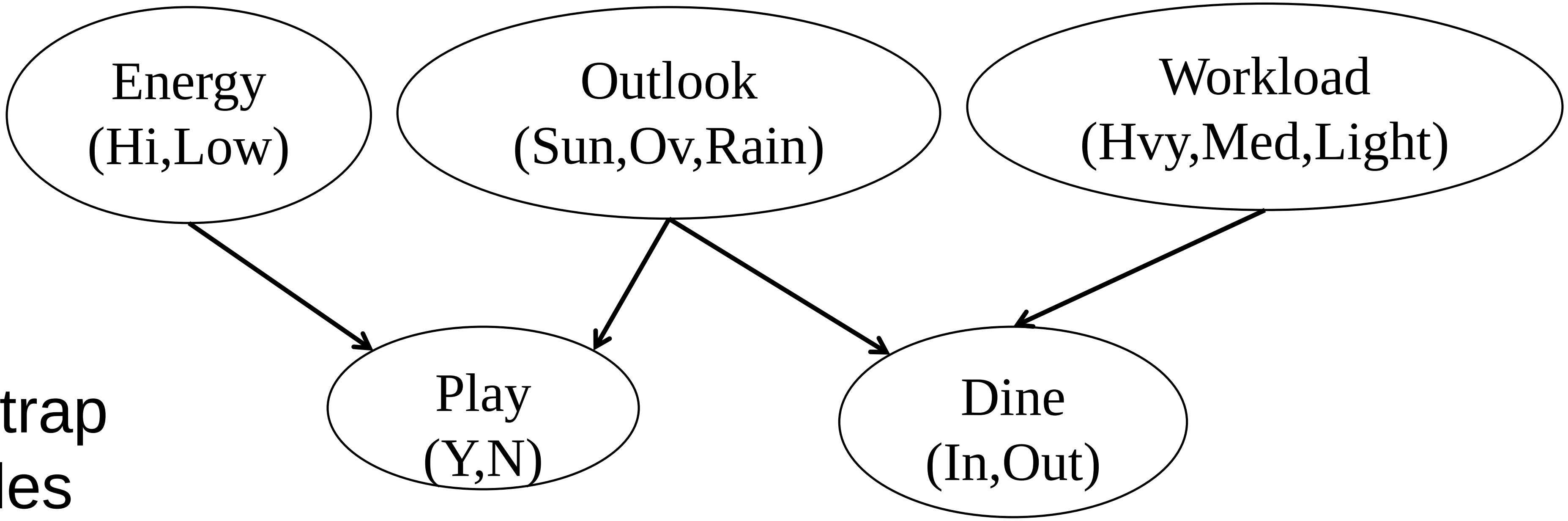
- **Two-stage least squares to handle spatially correlated errors (OLS, logit, multinomial logit)**

- **Bootstrap sampling of Bayesian network (package bnlearn) to confirm OLS effects, or suggest other moderating/mediating effects**

Depend. var

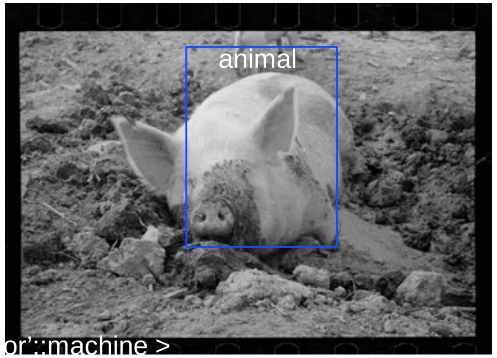


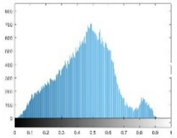
In a nutshell: Bayesian Network captures probabilistic dependencies between variables

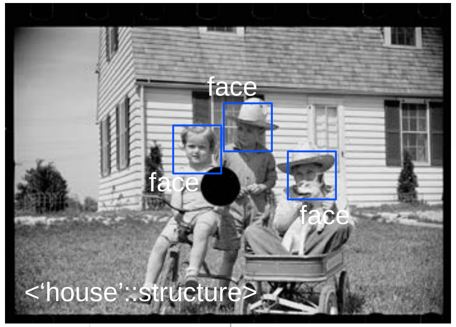


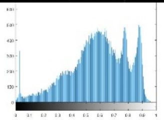
AND:
Bootstrap samples determine most likely edges

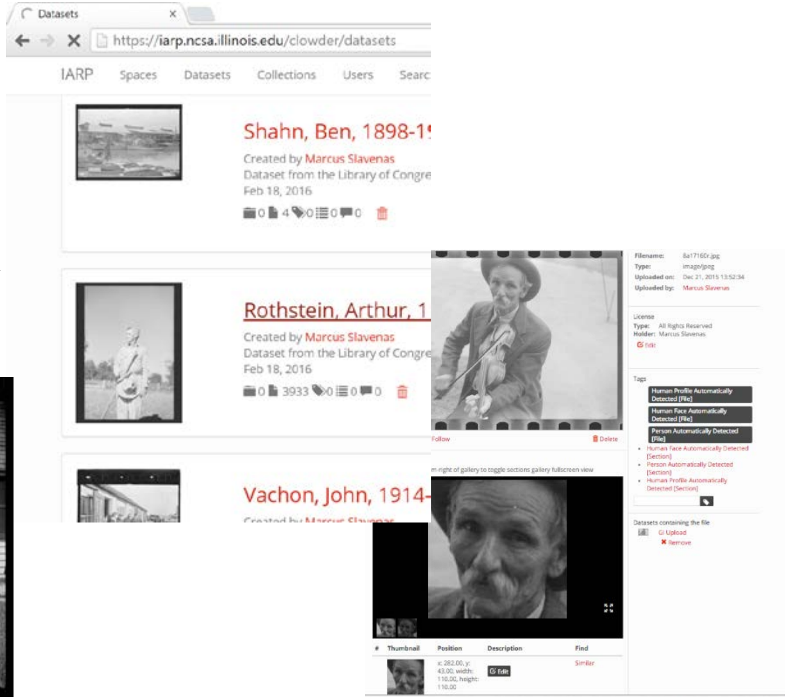
Image Analysis of Rural Photography
175K war and depression era photos
extracting features for datamining











Shahn, Ben, 1898-1!
Created by Marcus Slavenas
Dataset from the Library of Congre
Feb 18, 2016
0 4 0 0 0

Rothstein, Arthur, 1
Created by Marcus Slavenas
Dataset from the Library of Congre
Feb 18, 2016
0 3933 0 0 0

Vachon, John, 1914-
Created by Marcus Slavenas

Filename: 8a17160r.jpg
Type: image/jpeg
Uploaded on: Dec 21, 2015 13:52:34
Uploaded by: Marcus Slavenas

License
Type: All Rights Reserved
Holder: Marcus Slavenas
Edit

Tags
Human Profile Automatically Detected [File]
Human Face Automatically Detected [File]
Person Automatically Detected [File]
- Human Face Automatically Detected [Section]
- Person Automatically Detected [Section]
- Human Profile Automatically Detected [Section]

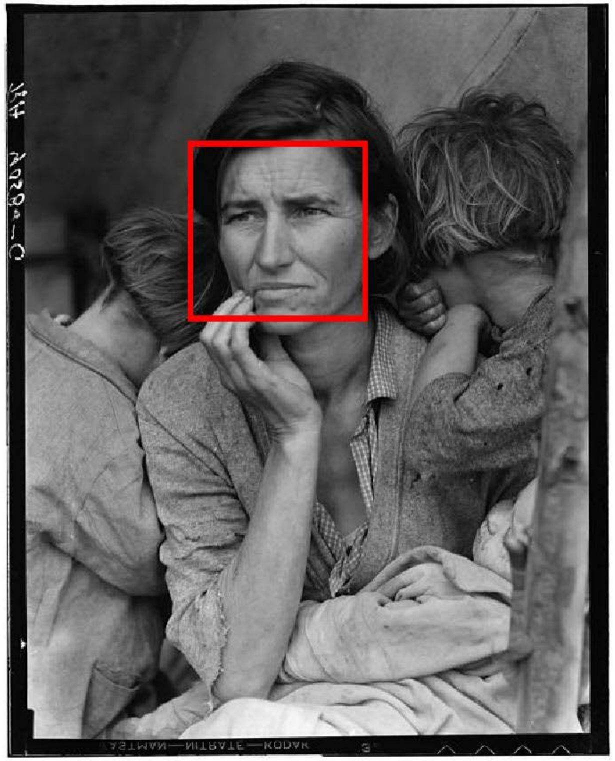
Datasets containing the file
Upload
✖ Remove

Follow
Delete
m right of gallery to toggle sections gallery fullscreen view

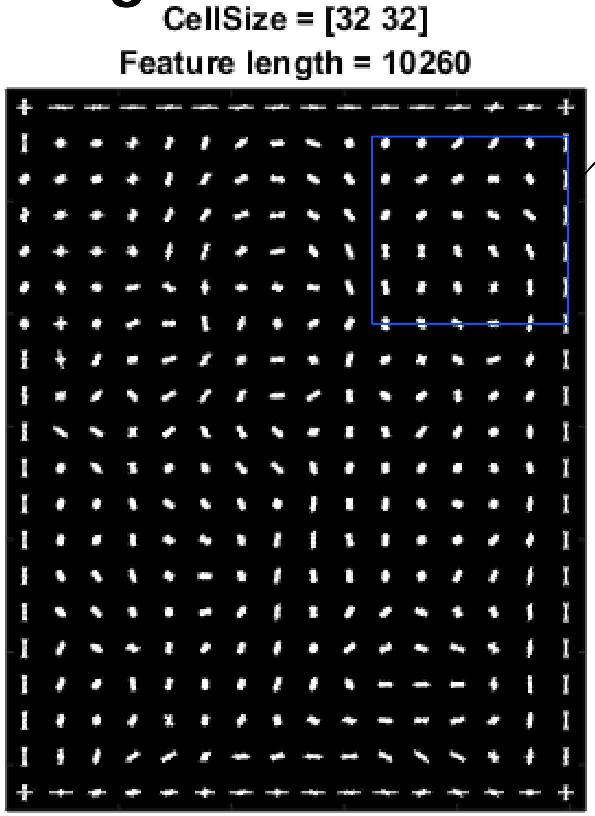
#	Thumbnail	Position	Description	Find
		x: 282.00, y: 43.00, width: 110.00, height: 110.00	Edit	Similar

Title:

”Destitute pea pickers in California. Mother of seven children.”



Histogram of Gradients



For each pixel in a cell, take filters:

[-1 0 1]

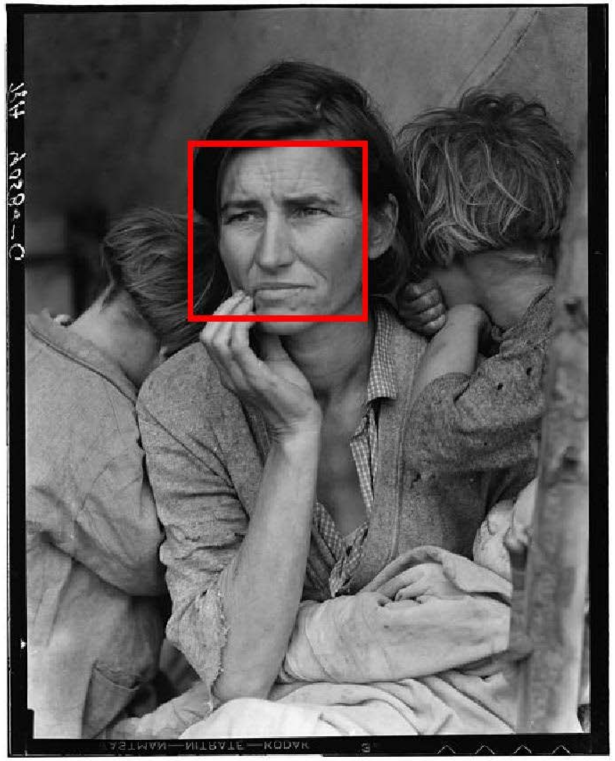
And

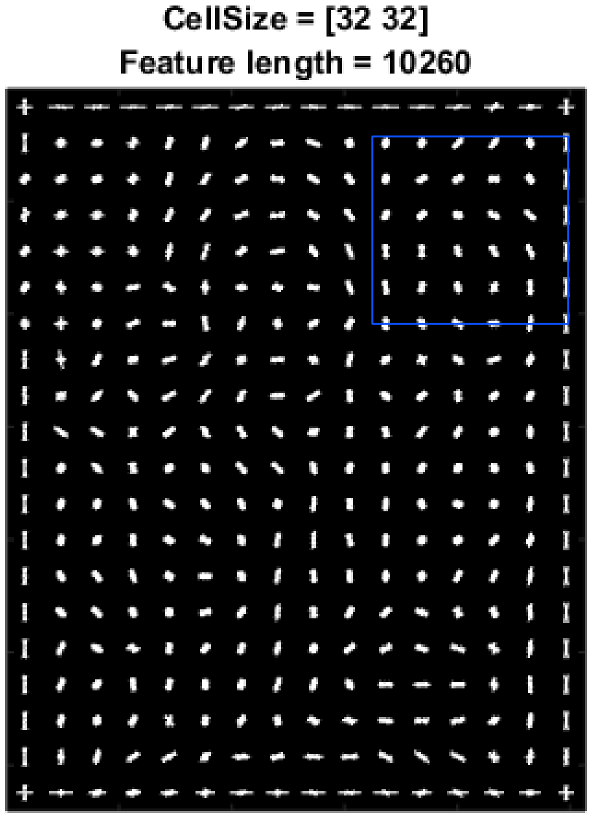
[-1
0
1]

Take weighted average and bin into 9 orientations; the bin frequency is like magnitude

Title:

”Destitute pea pickers in California.
Mother of seven children.”





Take all orientations, at different scales, as 1 big vector, and feed into classifier trained to recognize Face.

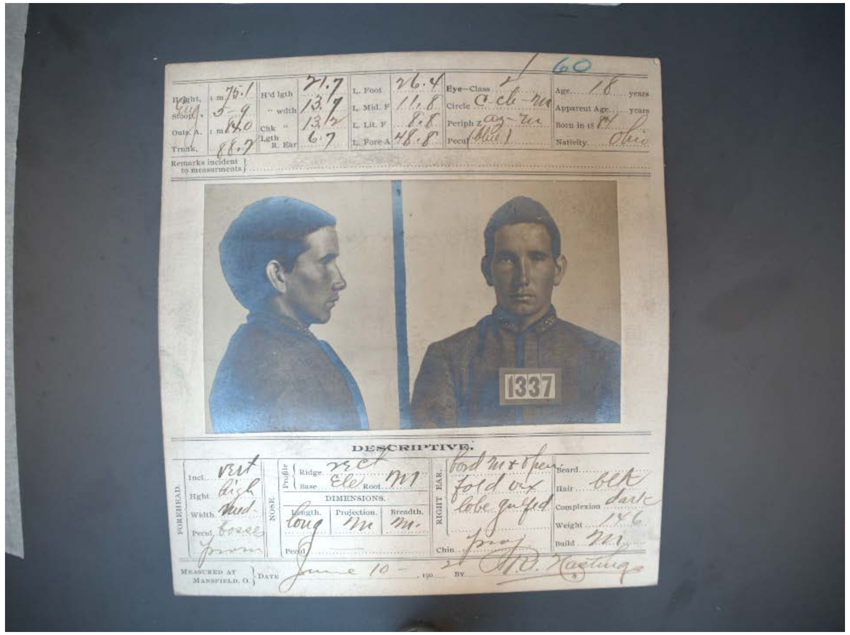
Title:

”Destitute pea pickers in California. Mother of seven children.” By D. Lange, 1936, California, [metadata]

Metadata processing:

- Parse and tag speech (using Stanford NLP tools, word ontologies, in Python NLP toolkit)
- Several words identify ‘person’
- SQL: give me all pictures by Lange with possible ‘person’ and num_faces > 0

Early 20th century, ~15k prison Bertillon id cards extracting information



1 60

Height,	1 m 75.1	H'd lgth	19.7	L. Foot	26.4	Eye—Class	2	Age	18 years
Stoop,	5-9	" wdth	13.7	L. Mid. F	11.8	Circle	c-cl-me	Apparent Age	years
Outs. A.	1 m 84.0	Chk "	13.2	L. Lit. F	8.8	Periph z	az-me	Born in 18	84
Trunk,	88.7	Lgth R. Ear	6.7	L. Fore A	48.8	Pecul	(blue)	Nativity	Ohio

Remarks incident to measurments

DESCRIPTIVE.

FOREHEAD. Incl. vert; Hght high; Width med.; Pecul. bosses frown

NOSE. Profile: Ridge rect; Base Ele; Root M

DIMENSIONS. Length. long; Projection. M; Breadth. M.

Pecul.

RIGHT EAR. bord m + open. fold ox. lobe grafted

Chin proj

Beard

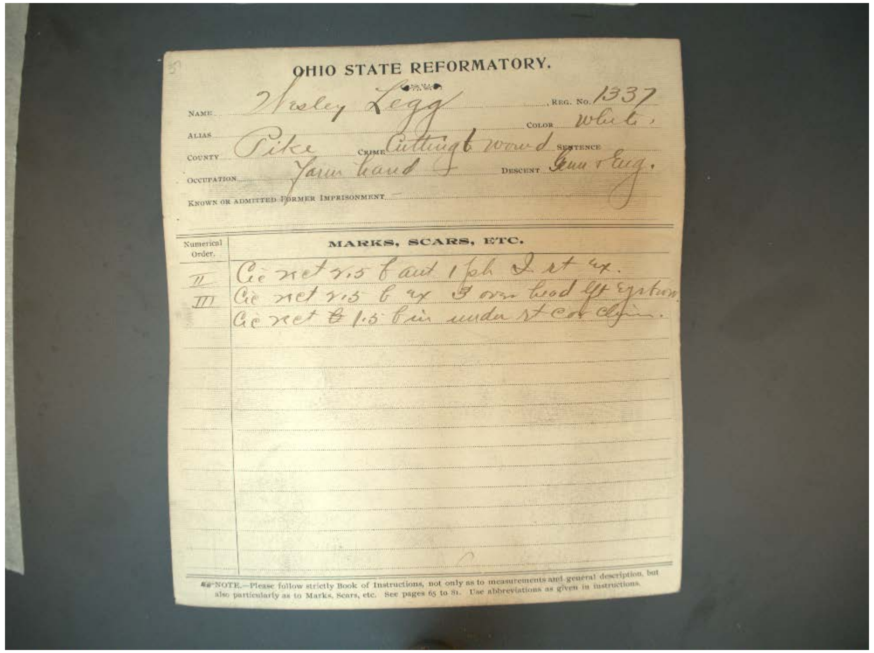
Hair blk

Complexion dark

Weight 146

Build M

MEASURED AT MANSFIELD, O. DATE June 10 - 190 1 BY A.D. Hastings



OHIO STATE REFORMATORY.

NAME Wesley Legg, REG. NO. 1337

ALIAS COLOR white.

COUNTY Pike CRIME Cutting to wound SENTENCE

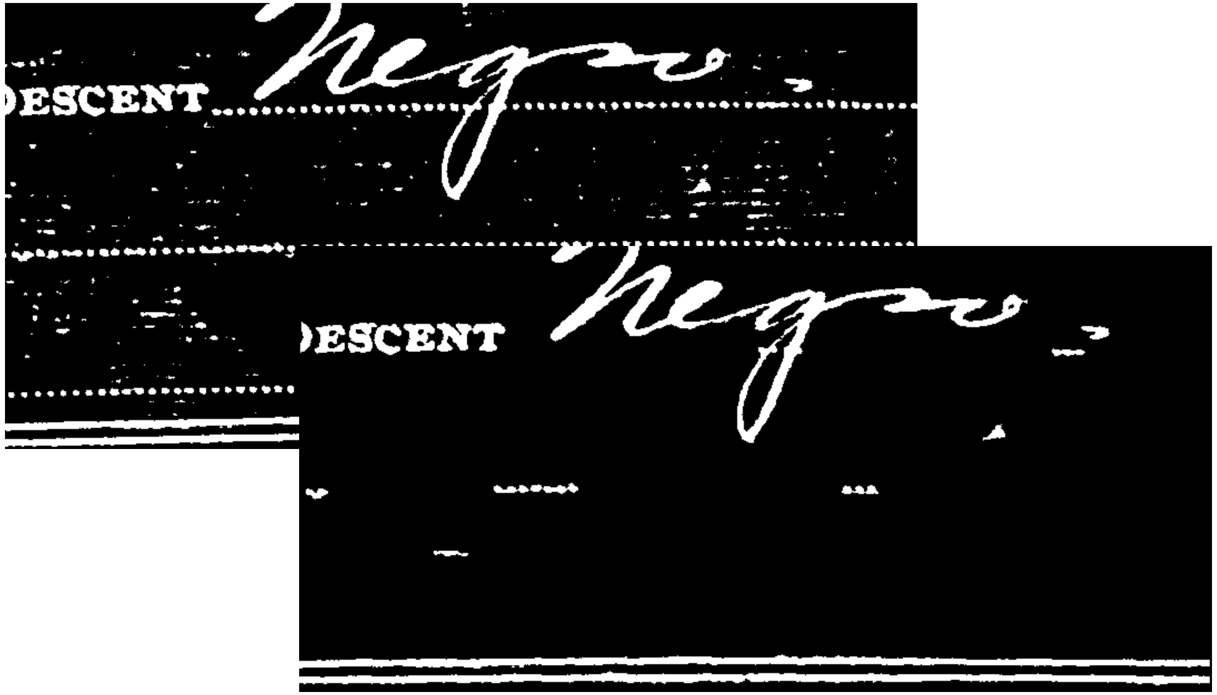
OCCUPATION Farm hand DESCENT Ger & Eng.

KNOWN OR ADMITTED FORMER IMPRISONMENT —

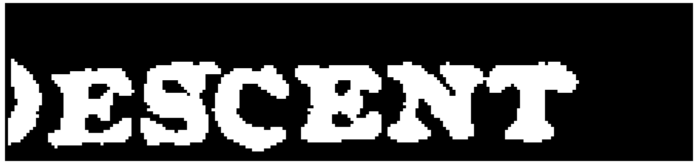
Numerical Order.	MARKS, SCARS, ETC.
II	Cic rect 2.5 b aut 1 ph I at ex.
III	Cic rect 2.5 b ex 3 ovr head lft eyebrow.
	Cic rect & 1.5 b in under st ce chin.

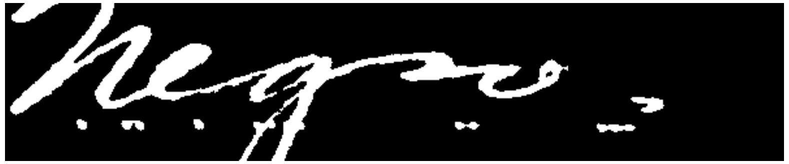
NOTE.—Please follow strictly Book of Instructions, not only as to measurements and general description, but also particularly as to Marks, Scars, etc. See pages 65 to 81. Use abbreviations as given in instructions.

Segment,
binarize,
denoise

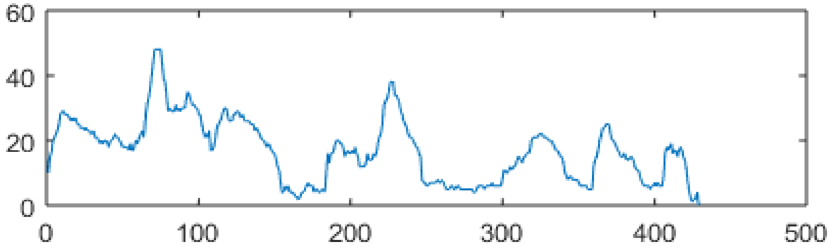

ESCENT
negro
ESCENT
negro

extract field and cell





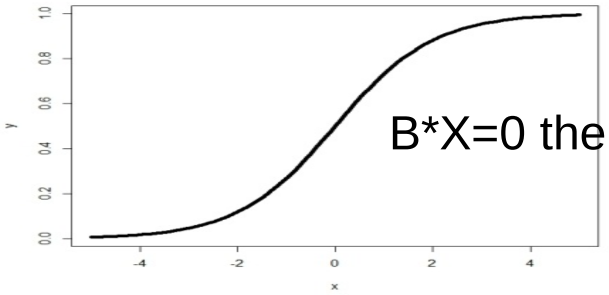
(word spotting) Get profile and compare to known templates



Linear to Logistic to Neural Network model

- *yi*= $bo * 1 + b_1 * xi_1$ **+** $b_2 * xi_2$ **… = B*X**

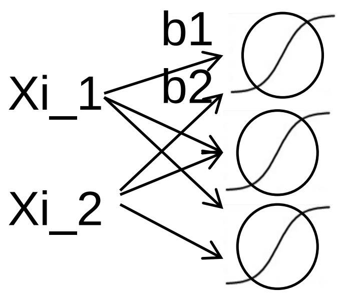
- **Squash** $bo * 1 + b_1 * xi_1$ **to 0,1 range using Logistic Function:**



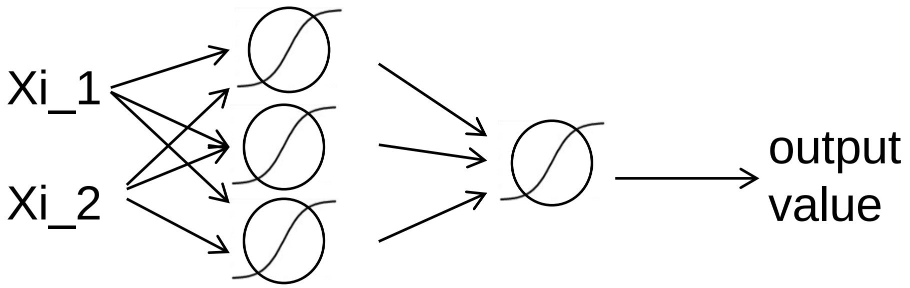
B*X=0 then y=0.5

Logistic Regression to Neural Networks

- **Use several squash functions (hidden layer)**



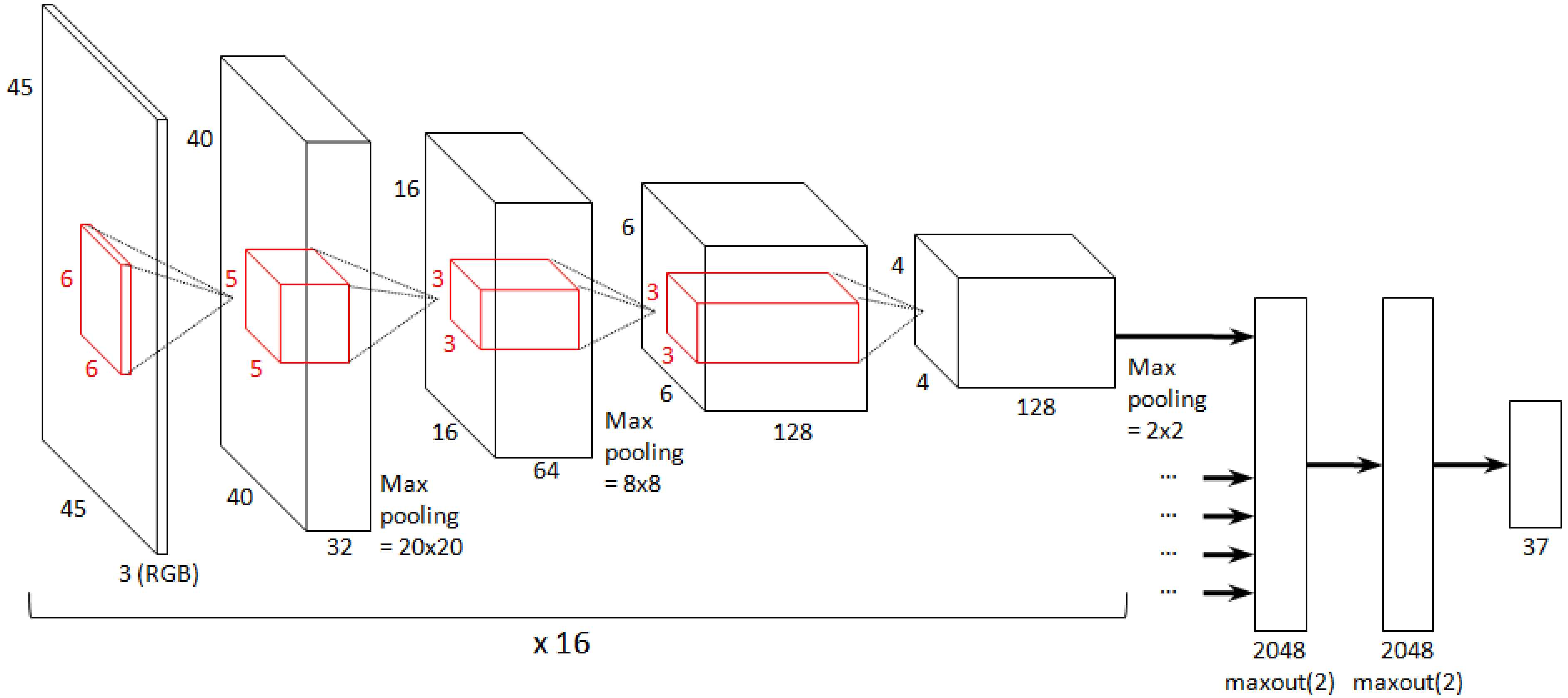
- **Take further combinations (output layer)**



- **More powerful but more complex**

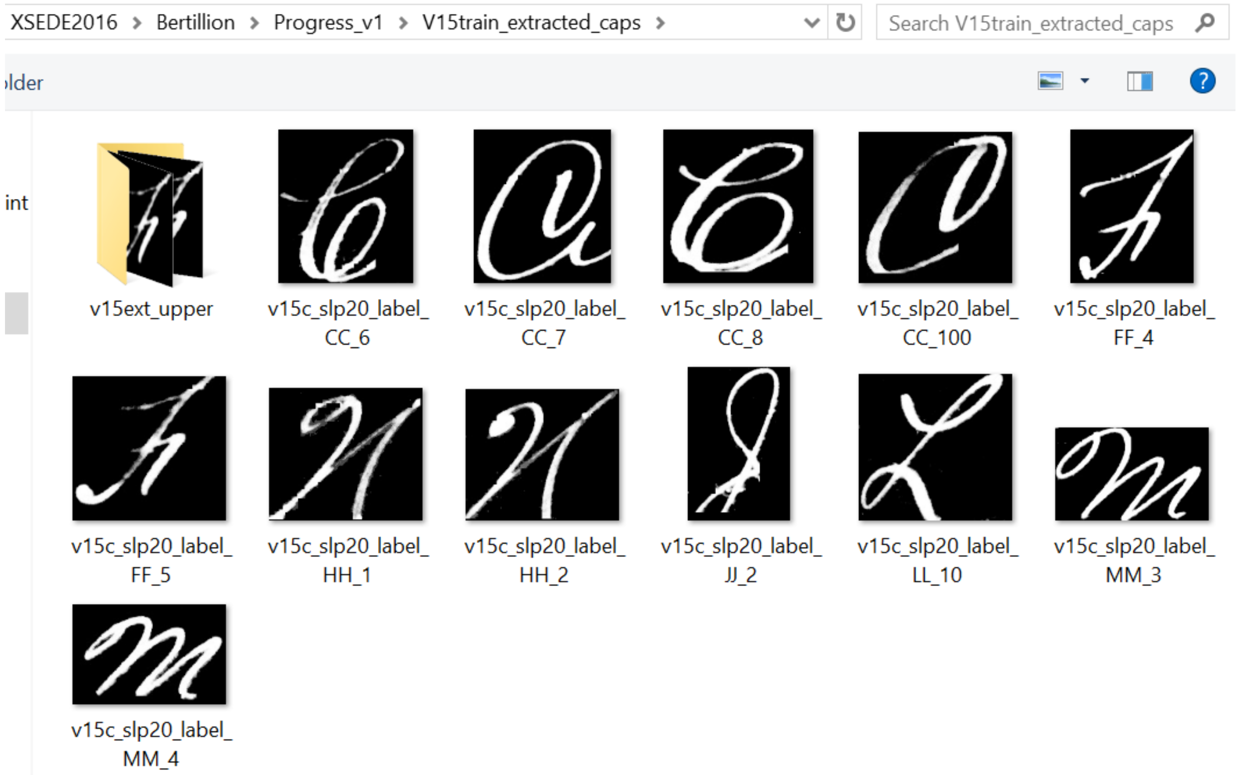
  many parameters, many options, needs more training

organize connections into cells, add layers (deepen), add special pooling operations at some layers - you get a convolution network



SciKit python package has a convolution neural network

```
nn2 = Classifier(
    layers=[
     Convolution("Rectifier", channels=numch, kernel_shape=(10,10),pool_shape=(2,2)),
     Convolution("Rectifier", channels=numch, kernel_shape=(6,6),pool_shape=(4,4)),
     Layer("Sigmoid",units=numalpha2do*4),
     Layer("Sigmoid",units=numalpha2do*2)
     ],
     verbose=False,
  learning_rate=0.001,valid_set=(Xtrain,Ytrain),
  n_iter=myiter)
  nn2.fit(Xtrain,Ytrain)
```



XSEDE2016 > Bertillion > Progress_v1 > V15train_extracted_caps >

Search V15train_extracted_caps

lder

int

v15ext_upper

v15c_slp20_label_CC_6

v15c_slp20_label_CC_7

v15c_slp20_label_CC_8

v15c_slp20_label_CC_100

v15c_slp20_label_FF_4

v15c_slp20_label_FF_5

v15c_slp20_label_HH_1

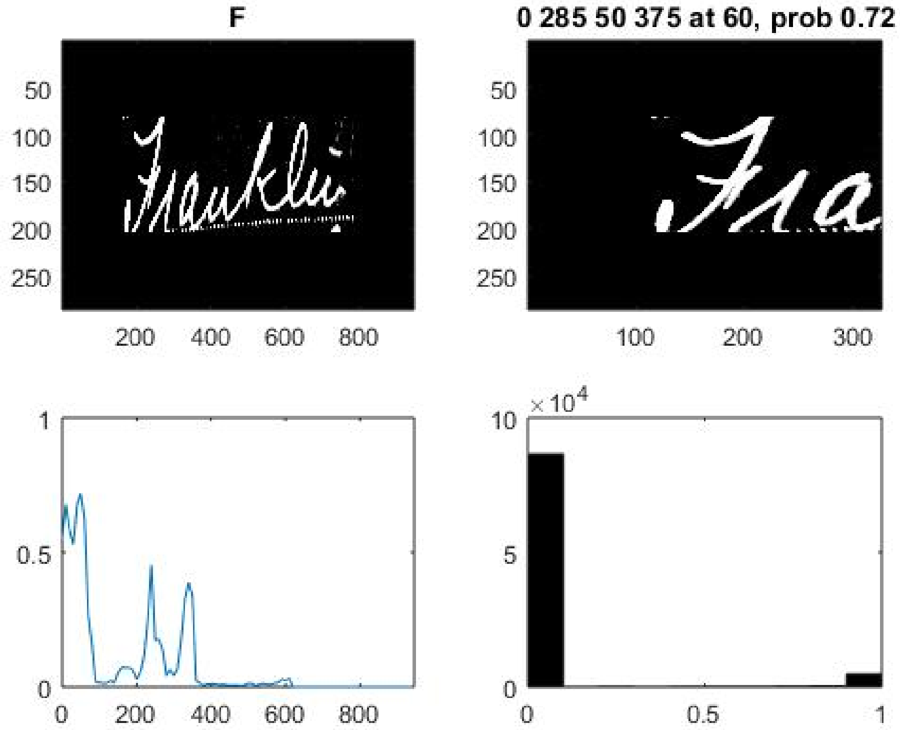
v15c_slp20_label_HH_2

v15c_slp20_label_JJ_2

v15c_slp20_label_LL_10

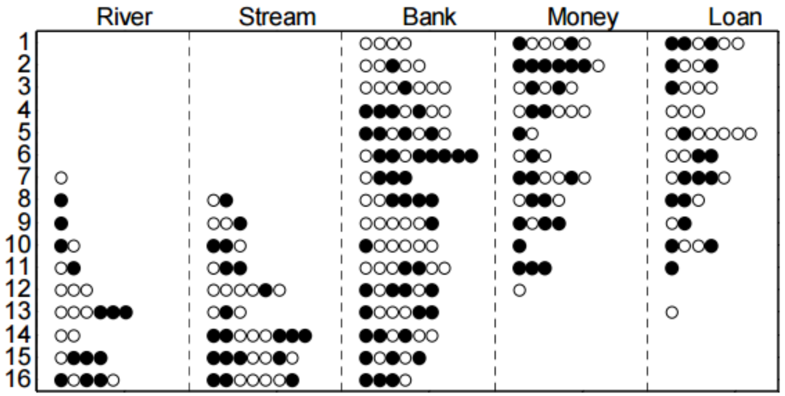
v15c_slp20_label_MM_3

v15c_slp20_label_MM_4

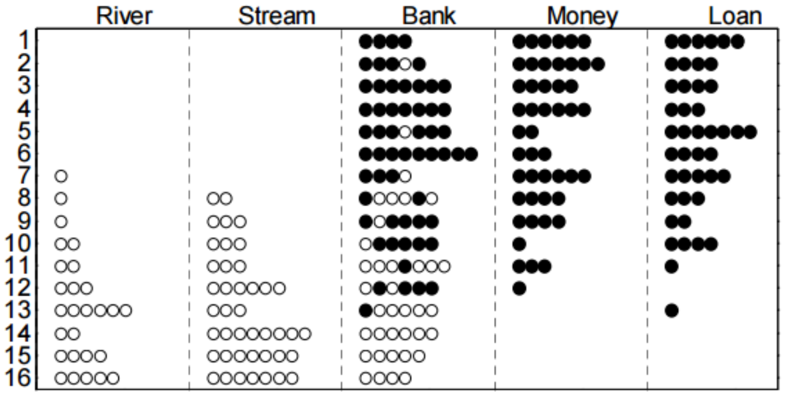


Topic Modelling with Latent Dirichlet Allocation

- Each circle is 1 word occurrence
- 2 topics (filled/empty circles), 15 docun
- Initially random assignments



- After learning, topics are well formed



LDA optimization

- **Start with initial guess of topic=*t*, and parameters**
- **Repeat:**

Compute the expected value of word=*w*

Compute the parameters that maximize likelihood L of *t given w*

Parameters are estimated from word/topic counts

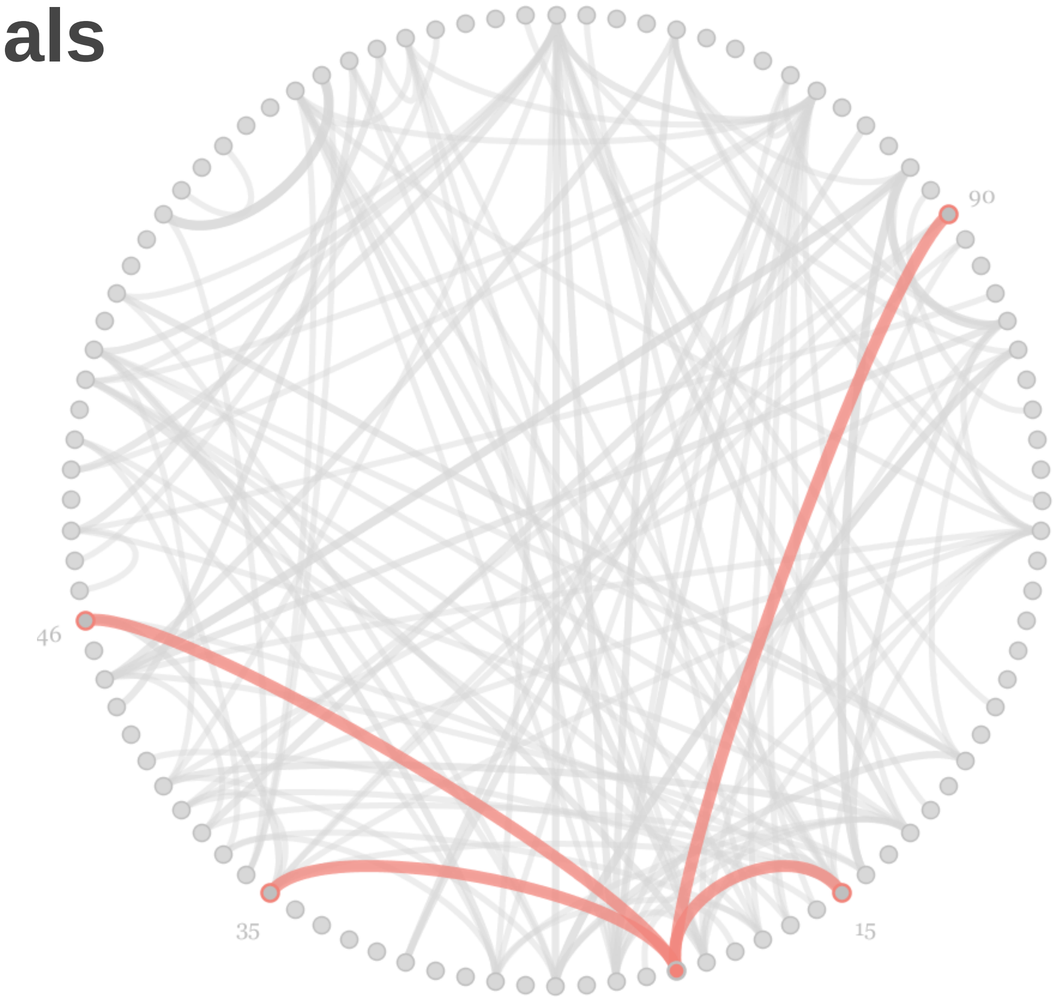
With each iteration, objective function L goes up

Topic Modelling with Latent Dirichlet Allocation on HPC

- R LDA package: wraps C programs for Gibbs sampling or EM
- Mallet: Gibbs sampling   multicore, java code
- Spark LDA: EM
- Asymptotic Distributed LDA : MPI based, no bells&whistles

- **Example Case:**

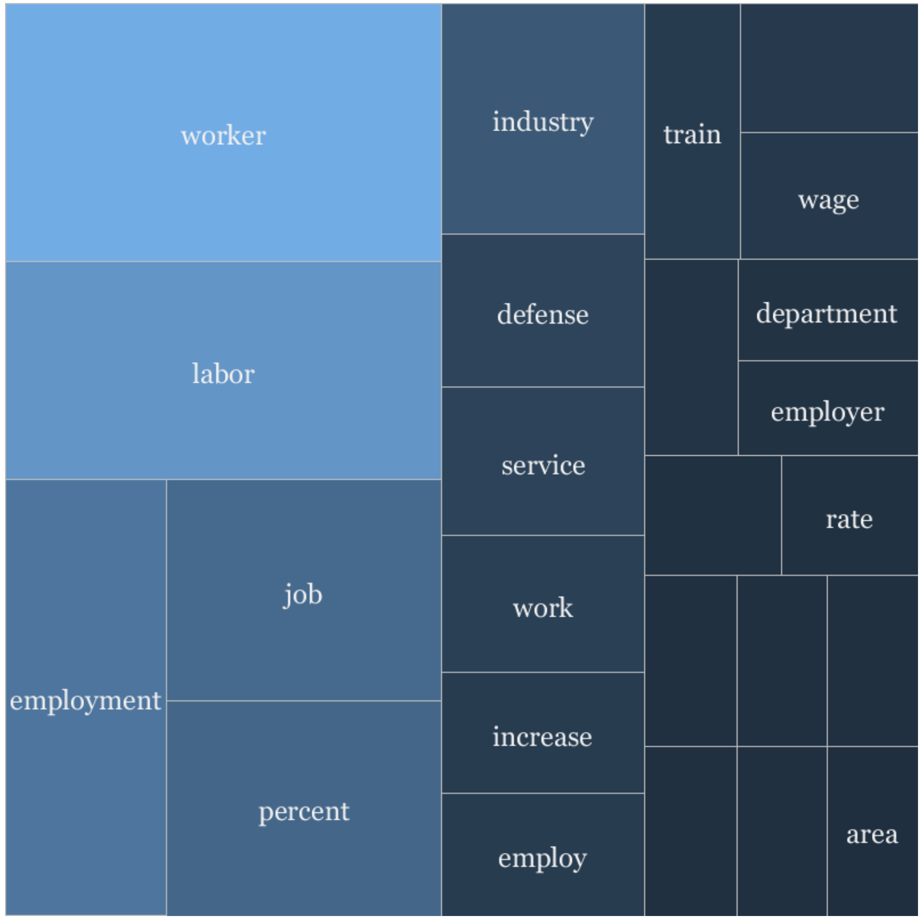
articles from post WWII journals



topic21	topic35	topic46	topic90	topic15
worker	county	farm	house	work
labor	city	land	family	day
employment	state	farmer	area	labor
job	town	acre	home	time
percent	york	crop	city	pay
industry	public	agricultural	build	hour
defense	district	county	live	week
service	mayor	family	room	wage
work	relief	agriculture	unit	make
increase	local	cotton	community	month
employ	population	labor	project	condition
train	person	state	income	year
unemployment	citizen	area	rend	employ
wage	community	migrant	move	find
employee	place	rural	neighborhood	case
department	residence	year	facility	money
employer	resident	tenant	urban	service
occupation	board	migration	low	receive
rate	settlement	make	occupy	leave
number	large	large	resident	care
industrial	welfare	grower	lodge	employer
production	number	small	condition	order
earnings	part	california	apartment	good
woman	aid	camp	neighborhoods	require

- **Example Case:**

Sample topic plot (tree map)


worker
labor
employment
job
percent
industry
defense
service
work
increase
employ
train
wage
department
employer
rate
area