Towards Better Identification of Fact Tables in Statistical Spreadsheets

Paul Kronlund-Drouault

Supervisor: Ioana Manolescu

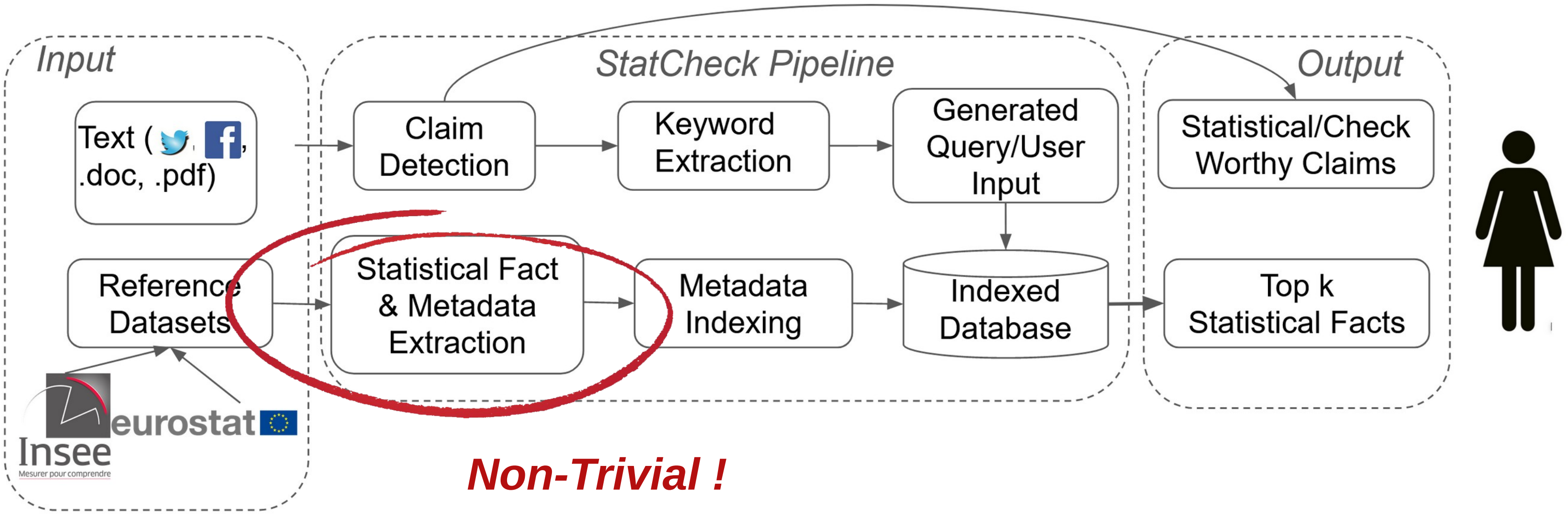
INSTITUT POLYTECHNIQUE DE PARIS

ENS DE LYON

Inria

Abstract & Motivation

1) Fact-checking is more relevant than ever
2) Computers can help us to fact-check efficiently with systems like ***StatCheck***
3) To fact-check we need facts



Non-Trivial !

Abstract & Motivation

1. Fact-checking is more relevant than ever
2. Computers can help us to fact-check efficiently with systems like *StatCheck*
3. To fact-check we need facts
4. Trusted institutions like *INSEE* or *Eurostat* produce usable facts
5. For *StatCheck* we need to extract those facts into a structured position-aware data format
6. The Sheets were the facts are stored are made for humans and thus can be very messy

INSEE statistical spreadsheets have very heterogeneous layouts, making it hard for heuristic-based methods to parse

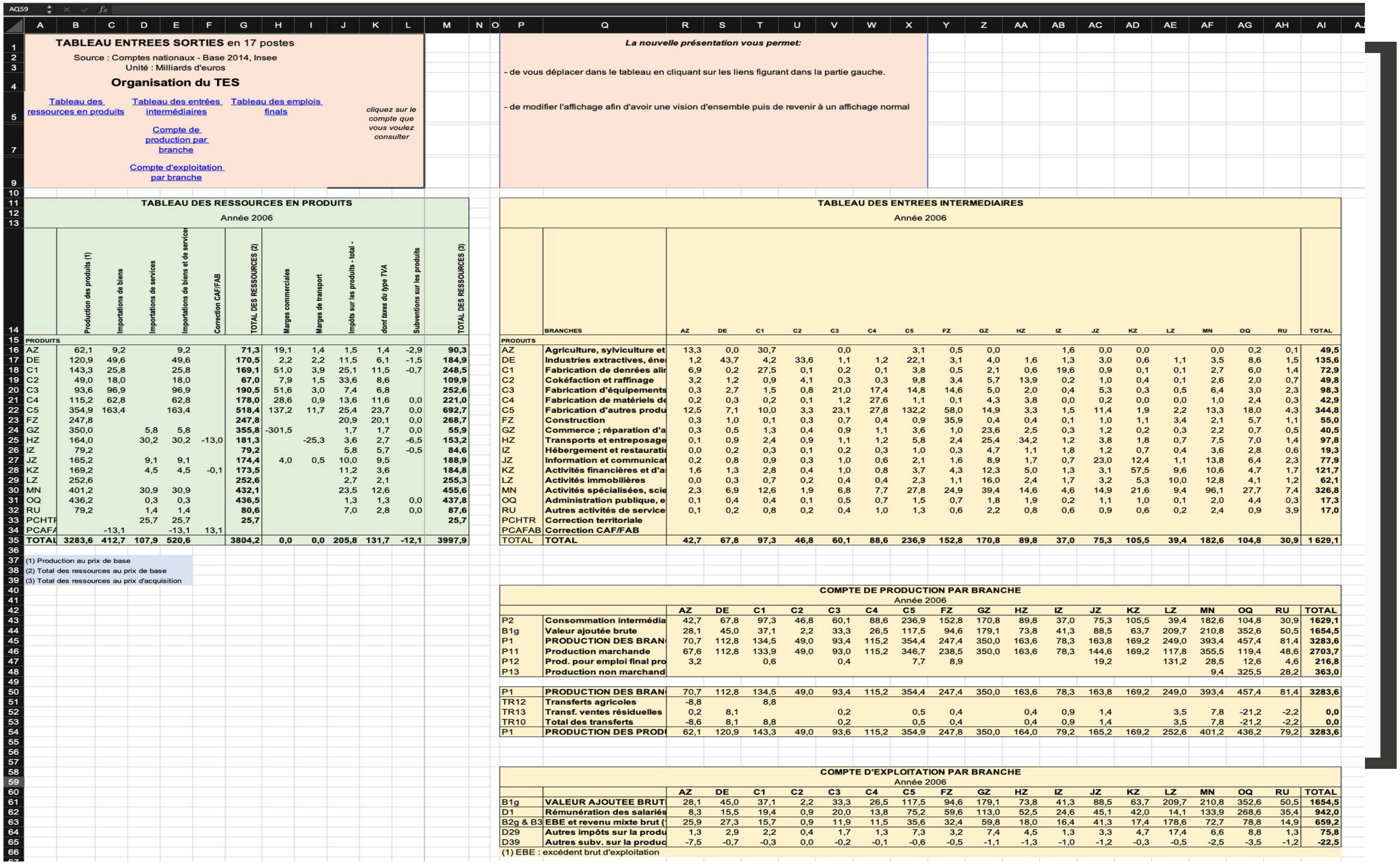


TABLEAU ENTREES SORTIES en 17 postes

Source : Comptes nationaux - Base 2014, Insee
Unité : Milliards d'euros

Organisation du TES

Tableau des ressources en produits — Tableau des entrées intermédiaires — Tableau des emplois finals

Compte de production par branche

Compte d'exploitation par branche

cliquez sur le compte que vous voulez consulter

La nouvelle présentation vous permet:

- de vous déplacer dans le tableau en cliquant sur les liens figurant dans la partie gauche.

- de modifier l'affichage afin d'avoir une vision d'ensemble puis de revenir à un affichage normal

TABLEAU DES RESSOURCES EN PRODUITS
Année 2006

	Production des produits (1)	Importations de biens	Importations de services	Importations de biens et de services	Correction CAF/FAB	TOTAL DES RESSOURCES (2)	Marges commerciales	Marges de transport	Impôts sur les produits - total -	*dont taxes du type TVA*	Subventions sur les produits	TOTAL DES RESSOURCES (3)
PRODUITS												
AZ	62,1	9,2		9,2		**71,3**	19,1	1,4	1,5	1,4	-2,9	**90,3**
DE	120,9	49,6		49,6		**170,5**	2,2	2,2	11,5	6,1	-1,5	**184,9**
C1	143,3	25,8		25,8		**169,1**	51,0	3,9	25,1	11,5	-0,7	**248,5**
C2	49,0	18,0		18,0		**67,0**	7,9	1,5	33,6	8,6		**109,9**
C3	93,6	96,9		96,9		**190,5**	51,6	3,0	7,4	6,8		**252,6**
C4	115,2	62,8		62,8		**178,0**	28,6	0,9	13,6	11,6	0,0	**221,0**
C5	354,9	163,4		163,4		**518,4**	137,2	11,7	25,4	23,7	0,0	**692,7**
FZ	247,8					**247,8**			20,9	20,1	0,0	**268,7**
GZ	350,0		5,8	5,8		**355,8**	-301,5		1,7	1,7	0,0	**55,9**
HZ	164,0		30,2	30,2	-13,0	**181,3**		-25,3	3,6	2,7	-6,5	**153,2**
IZ	79,2					**79,2**			5,8	5,7	-0,5	**84,6**
JZ	165,2		9,1	9,1		**174,4**	4,0	0,5	10,0	9,5		**188,9**
KZ	169,2		4,5	4,5	-0,1	**173,5**			11,2	3,6		**184,8**
LZ	252,6					**252,6**			2,7	2,1		**255,3**
MN	401,2		30,9	30,9		**432,1**			23,5	12,6		**455,6**
OQ	436,2		0,3	0,3		**436,5**			1,3	1,3	0,0	**437,8**
RU	79,2		1,4	1,4		**80,6**			7,0	2,8	0,0	**87,6**
PCHTF			25,7	25,7		**25,7**						**25,7**
PCAFA		-13,1		-13,1	13,1							
TOTAL	**3283,6**	**412,7**	**107,9**	**520,6**		**3804,2**	**0,0**	**0,0**	**205,8**	**131,7**	**-12,1**	**3997,9**

(1) Production au prix de base
(2) Total des ressources au prix de base
(3) Total des ressources au prix d'acquisition

TABLEAU DES ENTREES INTERMEDIAIRES
Année 2006

	BRANCHES	AZ	DE	C1	C2	C3	C4	C5	FZ	GZ	HZ	IZ	JZ	KZ	LZ	MN	OQ	RU	TOTAL
PRODUITS																			
AZ	**Agriculture, sylviculture et**	13,3	0,0	30,7		0,0		3,1	0,5	0,0		1,6	0,0	0,0		0,0	0,2	0,1	**49,5**
DE	**Industries extractives, éne**	1,2	43,7	4,2	33,6	1,1	1,2	22,1	3,1	4,0	1,6	1,3	3,0	0,6	1,1	3,5	8,6	1,5	**135,6**
C1	**Fabrication de denrées alir**	6,9	0,2	27,5	0,1	0,2	0,1	3,8	0,5	2,1	0,6	19,6	0,9	0,1	0,1	2,7	6,0	1,4	**72,9**
C2	**Cokéfaction et raffinage**	3,2	1,2	0,9	4,1	0,3	0,3	9,8	3,4	5,7	13,9	0,2	1,0	0,4	0,1	2,6	2,0	0,7	**49,8**
C3	**Fabrication d'équipements**	0,3	2,7	1,5	0,8	21,0	17,4	14,8	14,6	5,0	2,0	0,4	5,3	0,3	0,5	6,4	3,0	2,3	**98,3**
C4	**Fabrication de matériels de**	0,2	0,3	0,2	0,1	1,2	27,6	1,1	0,1	4,3	3,8	0,0	0,2	0,0	0,0	1,0	2,4	0,3	**42,9**
C5	**Fabrication d'autres produ**	12,5	7,1	10,0	3,3	23,1	27,8	132,2	58,0	14,9	3,3	1,5	11,4	1,9	2,2	13,3	18,0	4,3	**344,8**
FZ	**Construction**	0,3	1,0	0,1	0,3	0,7	0,4	0,9	35,9	0,4	0,4	0,1	1,0	1,1	3,4	2,1	5,7	1,1	**55,0**
GZ	**Commerce ; réparation d'a**	0,3	0,5	1,3	0,4	0,9	1,1	3,6	1,0	23,6	2,5	0,3	1,2	0,2	0,3	2,2	0,7	0,5	**40,5**
HZ	**Transports et entreposage**	0,1	0,9	2,4	0,9	1,1	1,2	5,8	2,4	25,4	34,2	1,2	3,8	1,8	0,7	7,5	7,0	1,4	**97,8**
IZ	**Hébergement et restauratio**	0,0	0,2	0,3	0,1	0,2	0,3	1,0	0,3	4,7	1,1	1,8	1,2	0,7	0,4	3,6	2,8	0,6	**19,3**
JZ	**Information et communicat**	0,2	0,8	0,9	0,3	1,0	0,6	2,1	1,6	8,9	1,7	0,7	23,0	12,4	1,1	13,8	6,4	2,3	**77,9**
KZ	**Activités financières et d'a**	1,6	1,3	2,8	0,4	1,0	0,8	3,7	4,3	12,3	5,0	1,3	3,1	57,5	9,6	10,6	4,7	1,7	**121,7**
LZ	**Activités immobilières**	0,0	0,3	0,7	0,2	0,4	0,4	2,3	1,1	16,0	2,4	1,7	3,2	5,3	10,0	12,8	4,1	1,2	**62,1**
MN	**Activités spécialisées, scie**	2,3	6,9	12,6	1,9	6,8	7,7	27,8	24,9	39,4	14,6	4,6	14,9	21,6	9,4	96,1	27,7	7,4	**326,8**
OQ	**Administration publique, e**	0,1	0,4	0,4	0,1	0,5	0,7	1,5	0,7	1,8	1,9	0,2	1,1	1,0	0,1	2,0	4,4	0,3	**17,3**
RU	**Autres activités de service**	0,1	0,2	0,8	0,2	0,4	1,0	1,3	0,6	2,2	0,8	0,6	0,9	0,6	0,2	2,4	0,9	3,9	**17,0**
PCHTR	**Correction territoriale**																		
PCAFAB	**Correction CAF/FAB**																		
TOTAL	**TOTAL**	**42,7**	**67,8**	**97,3**	**46,8**	**60,1**	**88,6**	**236,9**	**152,8**	**170,8**	**89,8**	**37,0**	**75,3**	**105,5**	**39,4**	**182,6**	**104,8**	**30,9**	**1 629,1**

COMPTE DE PRODUCTION PAR BRANCHE
Année 2006

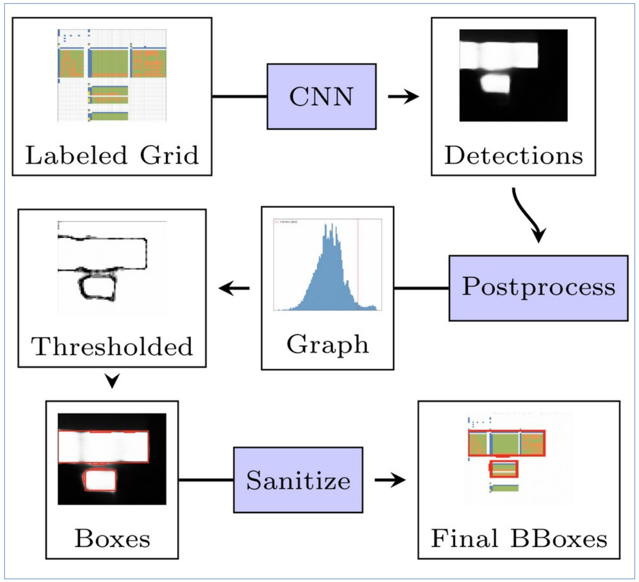
		AZ	DE	C1	C2	C3	C4	C5	FZ	GZ	HZ	IZ	JZ	KZ	LZ	MN	OQ	RU	TOTAL
P2	**Consommation intermédia**	42,7	67,8	97,3	46,8	60,1	88,6	236,9	152,8	170,8	89,8	37,0	75,3	105,5	39,4	182,6	104,8	30,9	**1629,1**
B1g	**Valeur ajoutée brute**	28,1	45,0	37,1	2,2	33,3	26,5	117,5	94,6	179,1	73,8	41,3	88,5	63,7	209,7	210,8	352,6	50,5	**1654,5**
P1	**PRODUCTION DES BRAN**	70,7	112,8	134,5	49,0	93,4	115,2	354,4	247,4	350,0	163,6	78,3	163,8	169,2	249,0	393,4	457,4	81,4	**3283,6**
P11	**Production marchande**	67,6	112,8	133,9	49,0	93,0	115,2	346,7	238,5	350,0	163,6	78,3	144,6	169,2	117,8	355,5	119,4	48,6	**2703,7**
P12	**Prod. pour emploi final pro**	3,2		0,6		0,4		7,7	8,9				19,2		131,2	28,5	12,6	4,6	**216,8**
P13	**Production non marchand**															9,4	325,5	28,2	**363,0**
P1	**PRODUCTION DES BRAN**	70,7	112,8	134,5	49,0	93,4	115,2	354,4	247,4	350,0	163,6	78,3	163,8	169,2	249,0	393,4	457,4	81,4	**3283,6**
TR12	**Transferts agricoles**	-8,8		8,8															
TR13	**Transf. ventes résiduelles**	0,2	8,1			0,2		0,5	0,4		0,4	0,9	1,4		3,5	7,8	-21,2	-2,2	**0,0**
TR10	**Total des transferts**	-8,6	8,1	8,8		0,2		0,5	0,4		0,4	0,9	1,4		3,5	7,8	-21,2	-2,2	**0,0**
P1	**PRODUCTION DES PROD**	62,1	120,9	143,3	49,0	93,6	115,2	354,9	247,8	350,0	164,0	79,2	165,2	169,2	252,6	401,2	436,2	79,2	**3283,6**

COMPTE D'EXPLOITATION PAR BRANCHE
Année 2006

		AZ	DE	C1	C2	C3	C4	C5	FZ	GZ	HZ	IZ	JZ	KZ	LZ	MN	OQ	RU	TOTAL
B1g	**VALEUR AJOUTEE BRUT**	28,1	45,0	37,1	2,2	33,3	26,5	117,5	94,6	179,1	73,8	41,3	88,5	63,7	209,7	210,8	352,6	50,5	**1654,5**
D1	**Rémunération des salariés**	8,3	15,5	19,4	0,9	20,0	13,8	75,2	59,6	113,0	52,5	24,6	45,1	42,0	14,1	133,9	268,6	35,4	**942,0**
B2g & B3	**EBE et revenu mixte brut (**	25,9	27,3	15,7	0,9	11,9	11,5	35,6	32,4	59,8	18,0	16,4	41,3	17,4	178,6	72,7	78,8	14,9	**659,2**
D29	**Autres impôts sur la produ**	1,3	2,9	2,2	0,4	1,7	1,3	7,3	3,2	7,4	4,5	1,3	3,3	4,7	17,4	6,6	8,8	1,3	**75,8**
D39	**Autres subv. sur la produc**	-7,5	-0,7	-0,3	0,0	-0,2	-0,1	-0,6	-0,5	-1,1	-1,3	-1,0	-1,2	-0,3	-0,5	-2,5	-3,5	-1,2	**-22,5**

(1) EBE : excédent brut d'exploitation

Grid-based table detection



Labeled Grid

CNN

Detections

Postprocess

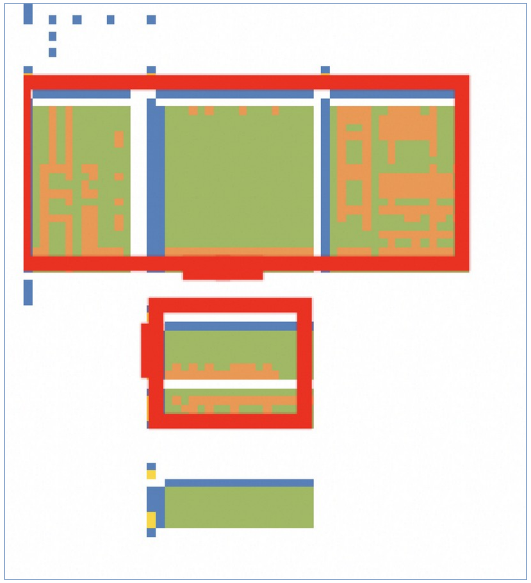
Graph

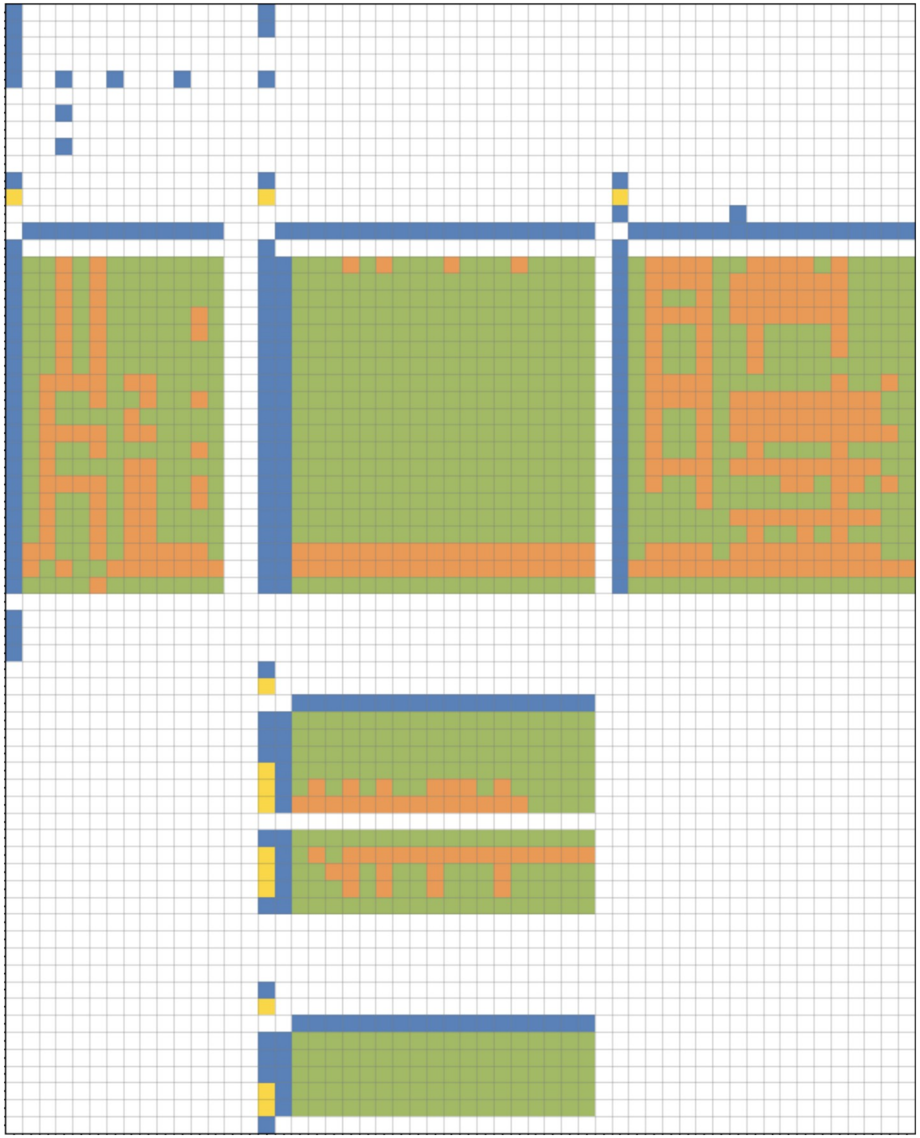
Thresholded

Boxes

Sanitize

Final BBoxes





Structure Classification

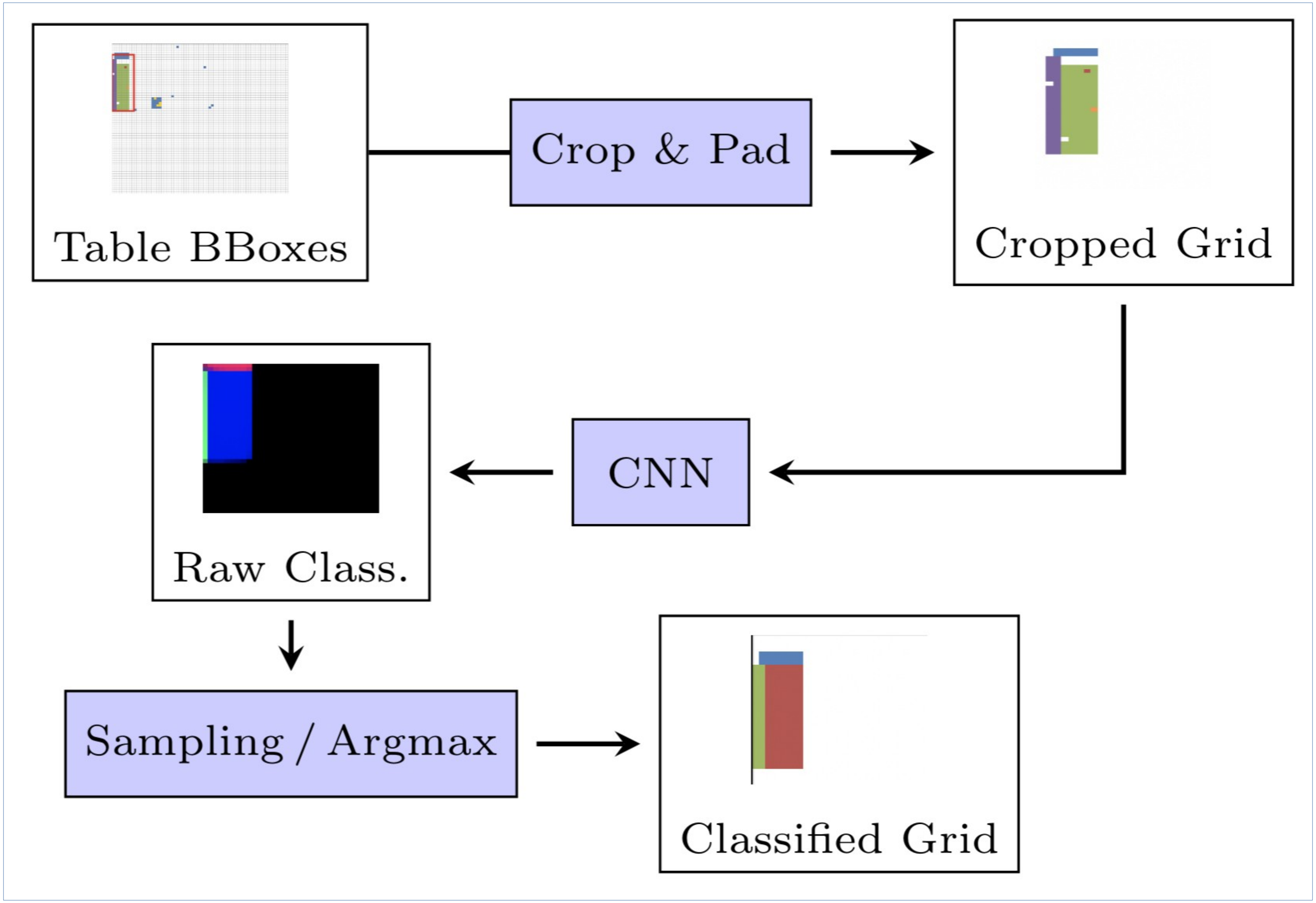


Table BBoxes
Crop & Pad
Cropped Grid
CNN
Raw Class.
Sampling / Argmax
Classified Grid