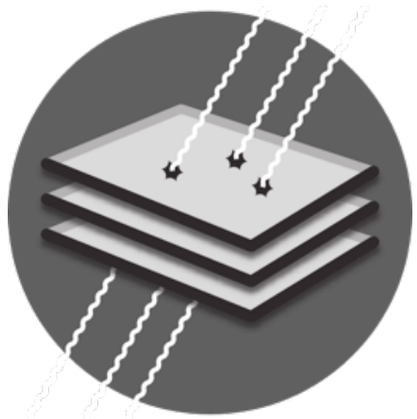
DAMIC-M meeting,
Paris 11-13 June 2018



Data Analysis and Simulations: from DAMIC to DAMIC-M

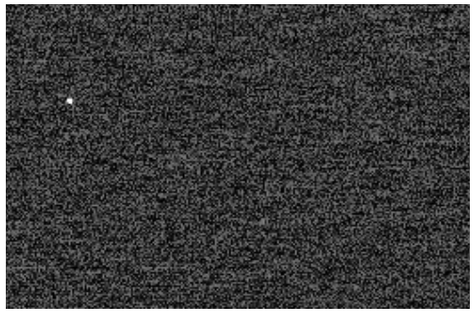
Mariangela Settimo
SUBATECH, CNRS-IN2P3, Nantes (France)

Outline

- Data processing / analysis
- Data monitoring
- Simulations
- Data storage
- Databases
- Computing Resources

From DAMIC ...

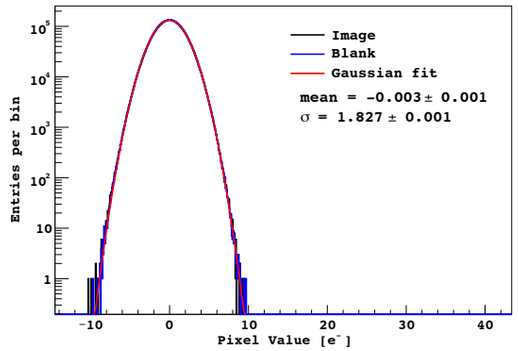
DAMIC: From raw data to DM search



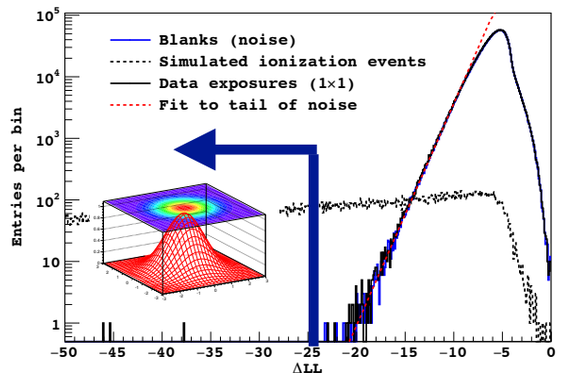
RAW images

Simulated Images

Image pre-processing
(pedestal, mask, CSN)



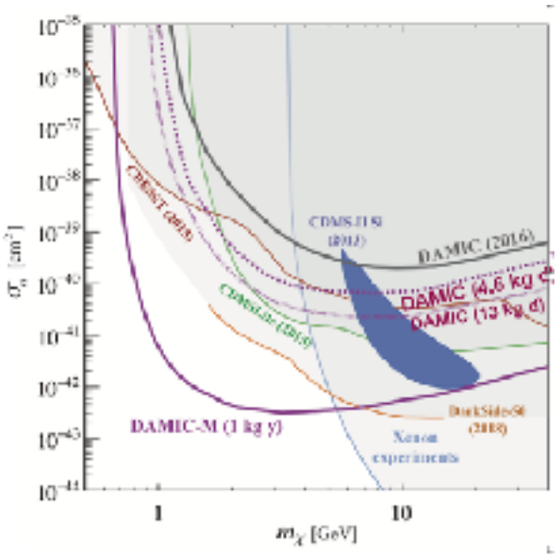
Reconstruction
event selection



Calibrat.

Energy
spectrum

E resolution,
exposure



results (physics, bkg, calib.)

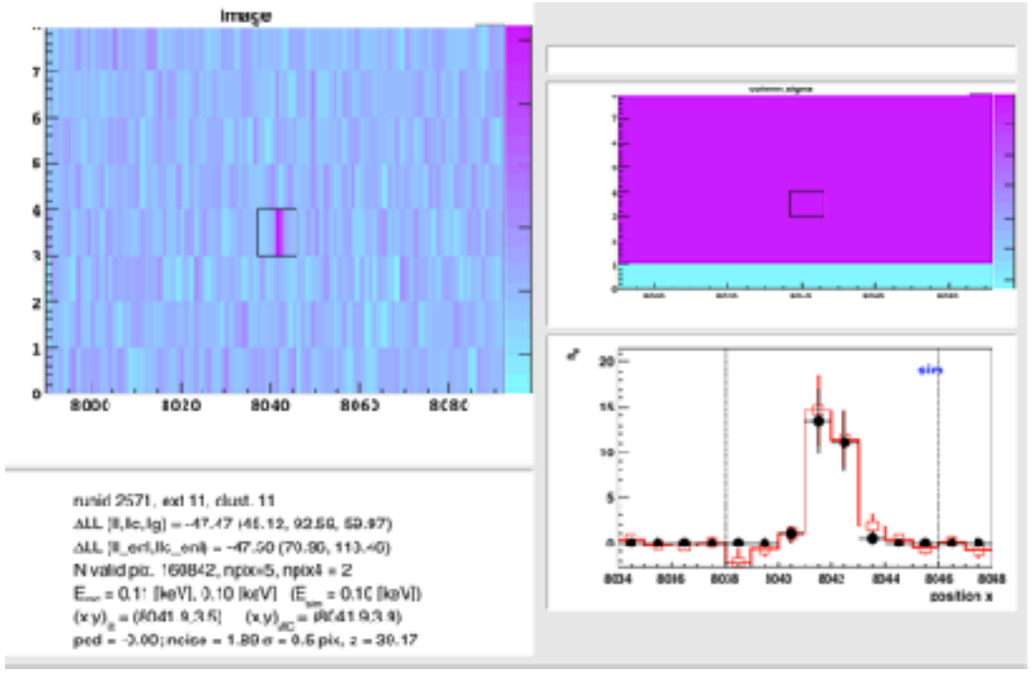
Bkg model
(G4sims)

Summary of analysis tools in DAMIC

- Analysis details (see A. Chavarria's talk yesterday, backup slides)

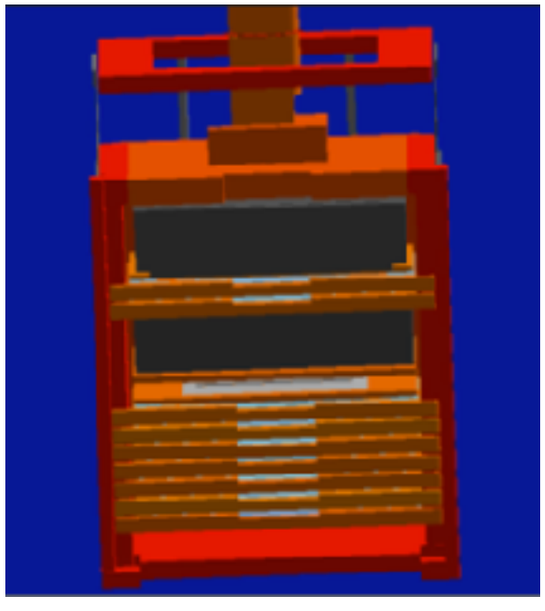
Image processing and reconstruction

- set of ``independent" programs/ macros in C/ROOT and python
- EventDisplay (ROOT-based)
- User's code in ROOT / python



Simulation

- Fast sims in C + ROOT
- Geant4 for background studies (Joao's talk)



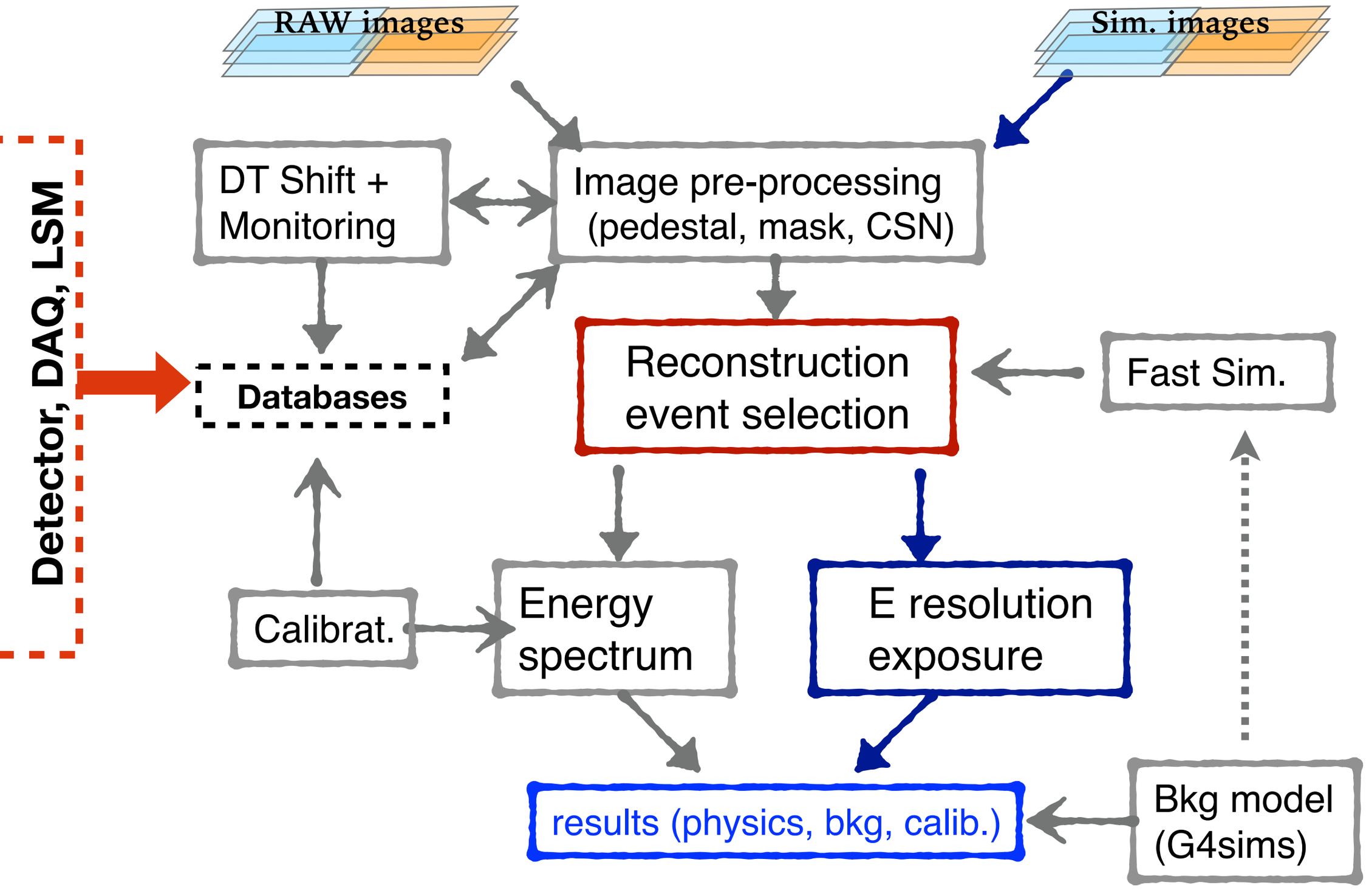
- Proposal to have compact code (fast-sims + reconstruction)

Monitoring System for DAMIC@Snolab

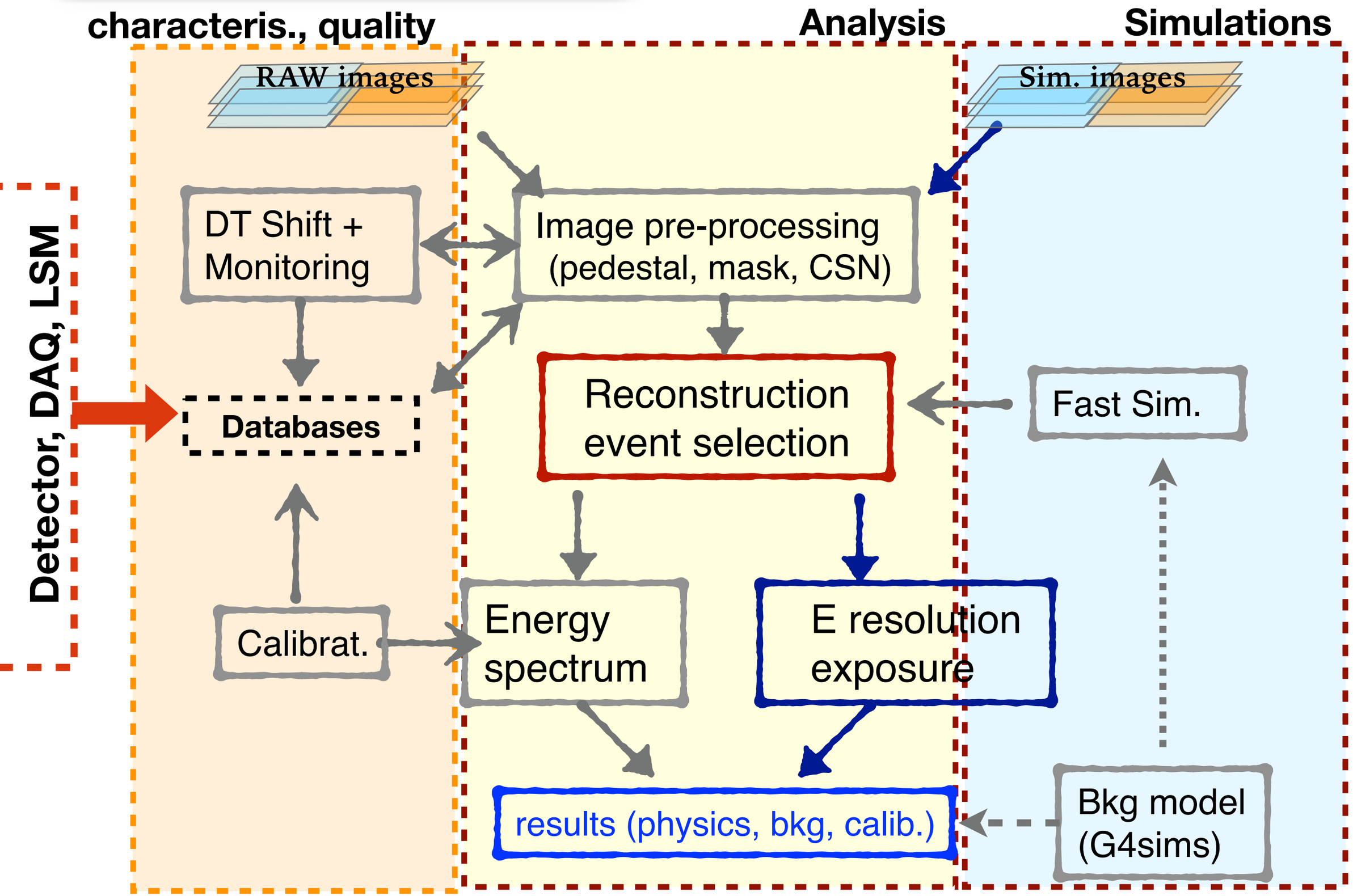
- Grafana system (open platform)
- Monitoring data processed locally (Snolab) & simple analyses in python
- Alarms set in Snolab/Fermilab and partly in CCIN2P3
- Images after crashes or bad-DAQ conditions “manually excluded“ from analysis

... To DAMIC-M

Sub-tasks and interfaces



Sub-tasks and interfaces



Detector and Data Quality Monitoring

- A ``**real-time**'' and a **redundant** system (with alarms) is a must
  - Detector conditions (temperature, pressure, …), Pedestal stability and noise
  - LSM infrastructure (UPS, network,…)

- Other **high-level variables** "quasi" real-time
  - Overscan and Img/Overscan variables,
  - Leakage Current, Number of clusters (seed),
  - Number of saturated pixels (or above given thresholds)

- **DQM** highly integrated with DAQ, data taking shifts and logbook

- On-line and "quasi-online" results filled in a DB

Image processing and analysis

- Analysis steps mostly unchanged (in the concept)
  - Details for processing and reconstruction **depends on RO/DAQ**
  - Runs for **physics, calibration** and **background**
  - Start to develop new reconstruction/analyses asap we get a CCD working or on simulated events

- Some requirements :
  - **compatible** with different RO/run modes **(skipper, CDS-equivalent, background, calibration)**
  - **Light / fast** code (scalability problems), different **tools** (ROOT/python/R/ML,…)
  - **“Real-time” processing** (a test running for DAMIC at Lyon)
  - **Access to DB** (monitoring, detector status, calibration, …)

- Analysis output / strategy:
  - Simple and light (avoid replications, external libraries dependencies, add a 2nd detailed stream if needed)
  - **Blinding procedure**

Analysis/Reconstruction

- A framework for data and simulations
  - A **unique program** for data and sims (preserving high **modularity**)
  - **New reconstruction algorithms** (LL definition, seeding, …)
  - Improve **flexibility** and **performances** of the current program
  - Reduce external dependences and keep easy the implementation of new codes

- Some possible optimisations
  - Parallelize master bias, mask and pedestal subtraction (large images)
  - Use the real-time preprocessing for **monitoring/shift purposes** (DQM, warning message can be delivered in case of unexpected-behaviour)

  - **DB to simplify access** to run informations and files

Simulations

1) **Fast simulations** (energy deposit)
   - Cross/check or validation of the analytical diffusion model
   - Integrate in the analysis framework

2) **Background simulations (Geant4)**
   - many inputs will come from the results of DAMIC (e.g. contaminant studies, Si32, DL, see Joao's talk)

3) **Design studies (Geant4 - high priority)**
   - for **shielding and bkg contamination**
   - **muon-induced background** (resume studies)
     - Split simulations in stages to optimise simulations
   - **calibration** studies

Simulations++

1) Some requirements for Geant4

- interface tool to import Mechanical design in Geant4 (e.g. CAD->G4)
- Improved modularity/flexibility (w.r.t to current code)
- cross-check with MCNP (on a specific part of the detector)

2) Beyond DAMIC-M

- Current Physics List (especially EM part) is validated down to 100 eV, decently fine down to 50eV.

- Proposition: involve Geant4 developers to develop custom-made physics lists at < 50 eV and use DAMIC-M CCDs for validation

Databases for DAMIC-M

- **Detector** : includes info on the material / elements of the detector (screening results, …).
  Filled at detector construction and in case of modifications

- **DetectorConfig** : slow control parameters (N_{ccd}, CCD position, connectors, Vbias, …., RO mode / parameters, set temperature, …)
  Filled only when run or detector settings change (or in case of alarm)

- **DetectorStatus** : parameters to be monitored continuously (detector status, temp, pressure, radon level, …).
  Automatically filled every xx min

- **LSM status** : parameters related to LSM (UPS, network, radon-free system status, …) *Automatically every yy min (or in case of changes, when the monitoring system identify a change in the system/alarm)*

- **DQM :** based on image pre-processing (noise, DarkCurrent, Nclusters, …)
  Automatically filled every xx min (or at the end of image acquisition)

Databases for DAMIC-M (II)

- **DB a solution to keep trace of simulations** :
  - DB to store info on simulations, useful to retrieve simulations, and at submission level for scheduler.

  - Simulations access DB to get the detector configuration (useful if sims are written in flexible/modular way)

- **Technical considerations :**
  - MySQL / PostgreSQL good options for relational DB with size < 1TB
  - both easily available on servers

  - Not a strong preference; similar performance and stability
    - mySQL well known, easier to get support
    - PostgreSQL more flexible for customize var_types (not necessarily our case)

Summary (I) in key-words

Monitoring:

- Improve integration with detector/shifts/data (DB)
- real-time availability, complementarity and redundancy

Analysis:

- New analyses/reconstruction and processing steps (related to RO/DAQ, CCD performances): start working asap as we get some working CCD
- Improve flexibility, performance (large datasets), automatisation and accessibility (DB)

Simulations:

- Improve flexibility, integration with detector configuration,
- New validations, performance, accessibility (DB)

Computing resources

Disk space consumption (Simple Scaling)

	DAMIC	DAMIC100	DAMIC1kg
n. CCDs in DAQ	6	12	50
size (1x100)	8000 x 2000	8000 x 4000	(2x) 6000 x 6000
Size (MB/exp) (*) 1x100 (1x1)	2 (15)	17 (300)	(2x) 150 (8GB) (+)
Size (/day) (x)	~12 (50) MB	~70 (900) MB	1.5 GB (80 GB)
Size (MB/year) (**)	4 GB (400 GB)	18 GB (~1.8 TB)	~1 (30) TB

(*) x100 for acquisition in 1x1 mode

Img size: DAMIC100 = 4.3 x DAMIC

(x) 30ks exposure

(+) assuming 2 param as output of continuous RO

1x100 (1x1) binning, includes overscan and header

8 GB = 4 ampl.Img or 4ampl.(Img&noise)

- Assuming continuous RO (1ms/pix RO time) with 4 amplifiers
- Improve (lossless) data compression level (or Zeros-suppression?)

Disk space consumption (Simulations)

Fast simulations :

- same disk space per image as in data but we normally "paste" many clusters on the same image (increase statistics at low cost)
- ~ 1TB for DAMIC simulations (2016 paper, 1x100 and 1x1)
- For DAMIC-M : depends on DAQ/RO mode, many optimisation possible (on the output format, x2 improvement at least): < **5 TB** sufficient

Geant4 sims :

- Some examples from DAMIC
  - 500 MB for 100k Si32, 60 MB for 100k Pb210 decays
  - Simulated clusters are then paste on images (?)
- DAMIC-M : larger image size, many simulations needed for design studies. BTW some output optimization possible: 100 TB (extrapolating from Joao's talk) …. **50 TB** reachable estimate (?)

Memory and CPU - time estimates

1x1 case

	CPU time/ ext	Mem./ext (Mb)
Equalis.	2.5 min	**400**
master biases + Mask	**8 min (65 blanks)**	560
cluster search	**25 sec**	500

1x100 case

	CPU time/ ext	Mem./ext (Mb)
Equalis.	1 min	**50**
master biases + Mask	**3 min (65 blanks)**	70
cluster search	**15 sec**	500

M.Settimo, DAMIC-1k meeting, 2017

Analysis (extrapolated): CPU time ~ 12 min, Mem: 800 Mb

Geant4 Simulations estimates

- for DAMIC (few min for Si32 and Pb210)
- muons sims needed for DAMIC-M (not for DAMIC) are longer (~10x)
- from Joao's talk: ~ 100kcore.h (DAMIC) —> add DAMIC-M design studies

Available computing centers

4 computing centers available

- CCIN2P3 (France)
- Chicago & Midway U. (USA)
- IFCA (Spain)
- SDU (Denmark)

	CCIN2P3	Chicago	IFCA/Altamira	SDU
CPU time	**1.0 - 2.0 MHS06.h**	**2x16 cores + 2Mcore.h**	**158x16 cores**	**240kCPU.h**
Space disk	**8.5 TB (2018) + 6TB every yr**	**20 TB (Kavli) + 1.5 TB (UMidway)**	**3 PB (infrastructure)**	**7TB**
Grant Access	OK (several members already using it)	@Kavli : Internal @UMidway possible?	OK	Ok
Notes	allocated for DAMIC (2018-2022) increase possible for DAMIC-M	Resources shared with Auger. 1Mcore.h can be requested in addition	Dedicated DAMIC quotas/access ?	
Other services	GRID, DB, Web space/mailing-list svn/git-in2p3 available	Svn available	GRID	

Data storage and access

- Goals:
  - **Safe storage** and mirroring/backup
  - **Easy access** (and transfer)
  - Optimized data access performance (at the lowest cost and to limit lost CPU time while waiting for IO)

- The CCIN2P3 as example ：
  - Several storage systems available (for data and sims):
    - **sps** (semi-permanent space): used as a standard mounted disk, not automatically backuped
    - **HPSS** (store large data volumes on cartrige, only for big-files)
    - Data access on HPSS through xROOTd

  - **Data access through virtualized storage systems (iRODS)**

Data storage and access (II)

1) **Redundancy:**
   - data (Img & monitoring) HW backup on local DAQ server (temporary)
   - synch in CCIN2P3 + at least another site
   - relatively small amount of data to transfer
     —> sync directly from LSM, more mirroring possible from servers

2) Depending on the data format:
   - **Store on HPSS the raw data** and have smaller **pre-processed files on /sps/** available for DQM, shifts, analyses.

   - **iRODS to access the files: both simulations and data** (physical location/storage type is transparent for the user)

     - Possibility of **iRODS federation** (ok for CCIN2P3, check in other sites)

3) **Do we need GRID**?
   - too much effort for the effective gain (amount of data and simulations of DAMIC-M)?

Summary

- Main changes for DAMIC-M **analysis** related to the **new RO and data format**
  - strategy conceptually similar, develop software **as soon as we have inputs**
  - Some ideas for framework, output format, strategy presented

- **Simulations for design studies** (bkg studies on DAMIC can give many inputs)

- Reinforced **integration with detector/DAQ :**
  - **inputs for analysis and complementary monitoring**

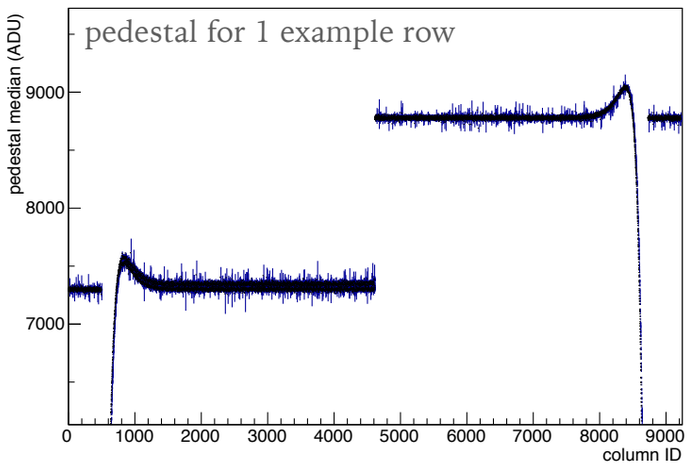
- **DB as interface** between analysis/detector/simulations

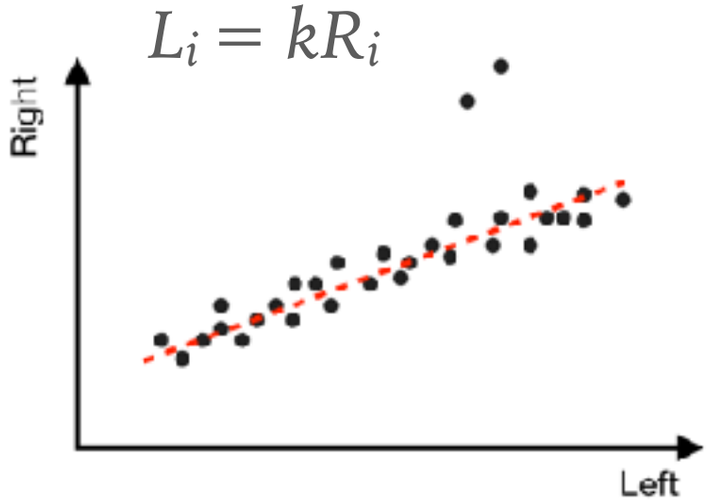
- **Computing resources adequate but better estimate depends on the DAQ**
  - Data storage redundancy/accessibility strategy
  - CPU time : better estimates needed

Discussion/questions ?

BACKUP

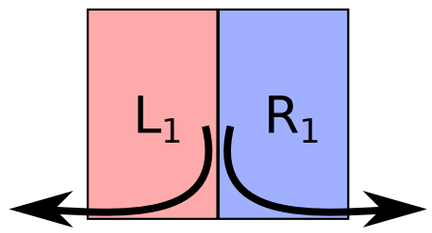
(1) Preprocessing and (2) Masking

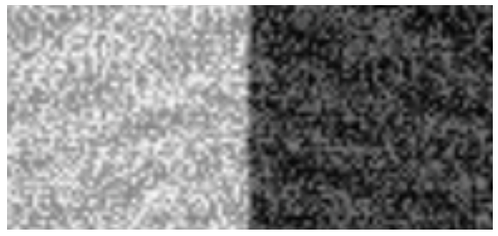




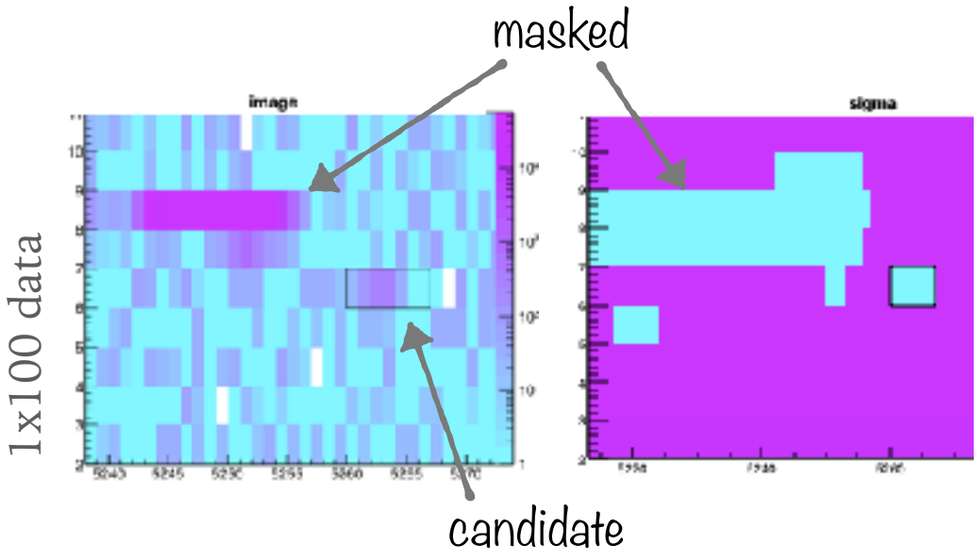
- median subtraction by col. & row
- Correlated noise subtracted

readout





- **mask (I)**: left-side of the image + "hot pixs": $|S_i - \langle S_i \rangle| > 2\sigma_i$ in more than 50% of the images
- **mask (II)**: (after the cluster search): clusters > 10 keV

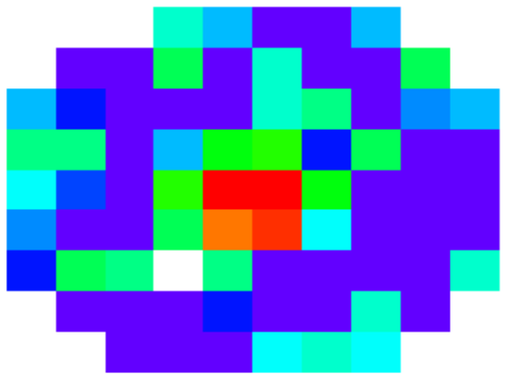


- 4x50 pixs masked around (1x1)
- 200 pixs on left side (1x100)

~ 5-8% of the image masked

3. cluster search

- Scan in $N_x \times N_y$ windows over the image
  - i) **likelihood $\mathbf{L_n}$** : white noise only
  - ii) **likelihood $\mathbf{L_G}$** : 2D-Gaussian+white noise
    - *Note: for the 1×100 mode, 1D-Gaussian used*

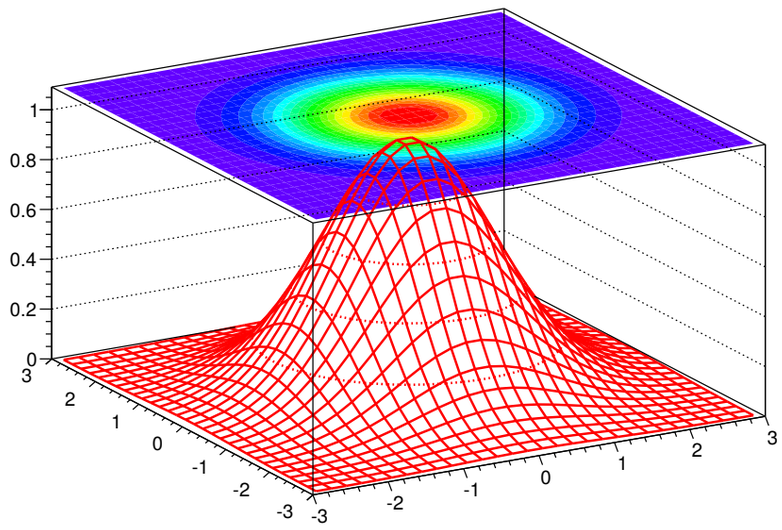


$$N_e(E) \times \texttt{Gaus}(x, y, \mu_x, \mu_y, \sigma(z))$$

Number of ionized electrons

Best estimate for mean of energy deposition

Lateral spread



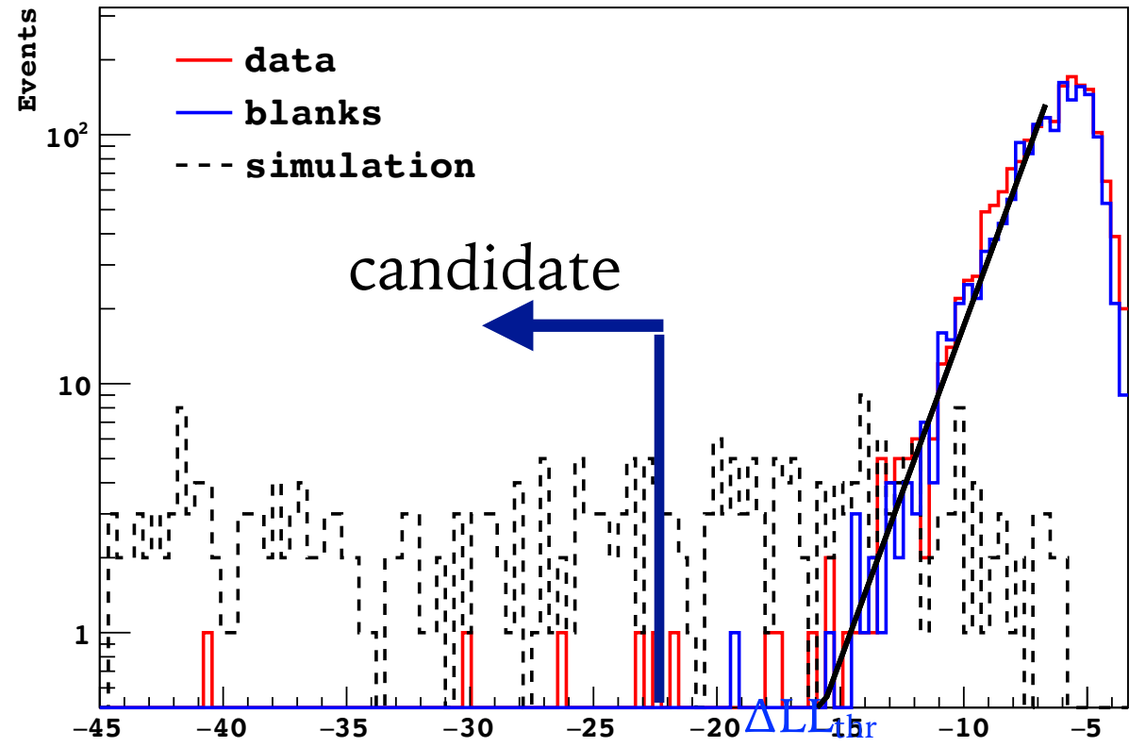
Params of the Gaussian fit:

E, x, y, σ_{xy} ($\propto$z)

4. Candidate selection

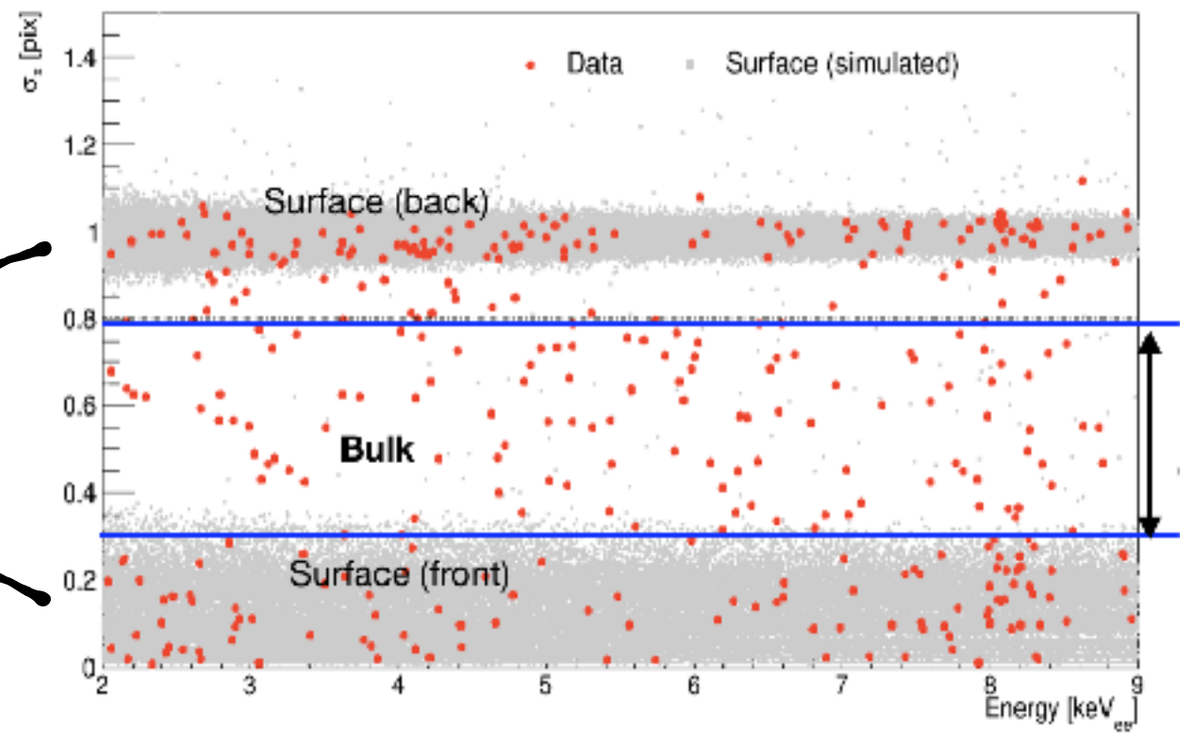
- **Test statistic:** $\Delta LL = -\ln\left[\frac{\max(\mathcal{L}_G)}{\mathcal{L}_n}\right]$
- **Candidates** : $\Delta LL < \Delta LL_{thr}$

  ΔLL_{thr} : < 0.01 bkg events (exp. from the ΔLL tail distribution)



- Surface events rejection
  - Cut on σ_{xy}

Front/back surface events from simulations



5. Simulations and exposure

- **Energy deposit** uniformly distributed in the CCD volume **+ diffusion model**
- Image processing and cluster search as for real data
- Efficiency of event reconstruction and resolutions/energy bias